Statistical Foundations of Data Science

a view towards applications

volume II

Renato Assunção

PUBLISHED BY PUBLISHER

BOOK-WEBSITE.COM

First printing, March 2020

Contents

16 Linear Regression 9

16.1 Introdução 9

16.2 Modelos de Regressão 14

17 Logistic Regression 19

17.1 Introdução 19

17.1.1 A função logística 21

17.1.2 Estimando β_0 e β_1 22

17.2 $L(\theta) = L(\beta_0, \beta_1)$: a função de verossimilhança 25

17.3 Obtendo o MLE 27

17.4 Regressão Logística Múltipla 30

17.4.1 O preditor linear $\eta = \mathbf{x}'\beta$ 30

17.5 Outros exemplos usando regressão logística 32

17.6 E se tivermos um efeito não-monótono? 33

17.7 Feature Engineering 34

17.8 Regressão logística como classificador 35

18 Maximum Likelihood Estimator 39

18.1 Introdução 39

18.1.1 Estimando o tempo médio de sobrevida 40

18.1.2 Quais valores do parâmetro são verossímeis? 42

18.1.3 Verossimilhança Relativa e máxima 45

18.1.4 Resumo informal da máxima verossimilhança 46

18.1.5 Por quê a máxima verossimilhança? 47

18.2 Modelos paramétricos **47**
18.3 Estimativa de máxima verossmilhança: MLE **49**
18.4 Obtendo o MLE **52**
18.5 MLE: soluções analíticas **54**
18.5.1 Verossimilhança Bernoulli 54
18.5.2 Verossimilhança Binomial 56
18.5.3 Verossmilhança Poisson 56
18.5.4 Verossimilhança de Poisson truncada 57
18.5.5 Dados em forma de Tabelas 58
18.6 MLE: soluções numéricas **61**
18.6.1 Caso unidimensional: $\theta \in \mathbb{R}$ 61
18.6.2 Exemplos de MLE 1-dim por métodos numéricos 63
18.6.3 Caso multidimensional: $\theta \in \mathbb{R}^p$ 66
18.6.4 De volta à regressão logística 67

19 Theory of Point Estimation 73
19.1 Outros métodos de estimação **73**
19.1.1 Método de Momentos 73
19.1.2 Um método sem nome 75
19.1.3 Mais um método sem nome 76
19.2 Estimadores são variáveis aleatórias **78**
19.3 Estimação Pontual **80**
19.4 Critérios para escolher estimadores **84**
19.5 Escolhendo um estimador **84**
19.6 Decompostion of the MSE: Bias and variance **85**
19.7 Consistency of estimators **87**
19.7.1 Nearest Neighbor estimators are consistent 87
19.7.2 Classification trees may be consistent 87
19.7.3 Boosting is consistent 87

20 MLE Optimality 89
20.1 MLE Optimality for uni-dimensional θ **89**
20.2 Multivariado **98**

21 Confidence Intervals and Hypothesis Tests 107
21.1 Introdução **107**
21.2 Intervalos de Confiança baseados no EMV **109**
21.3 Caso Multivariado **113**
21.3.1 Estimação de Intervalos baseados no EMV para o caso multivariado 115
21.3.2 Caso da Normal Multivariada 115
21.4 Intervalos de Confiança no Modelo de Regressão Linear **116**
21.5 Método Delta Univariado e Multivariado **117**
21.6 Quando $[I^{-1}]_{ii} \neq 1/I_{ii}$ **117**
21.7 Informacao observada e esperada **118**

22 Mixture Models . . . 123
22.1 Misturas de distribuições: introdução 123
22.2 Misturas de distribuições: formalismo 125
22.2.1 Misturas de normais multivariadas . . . 127

23 EM Algorithm . . . 131
23.1 Estimando uma distribuição de mistura 131
23.2 Dados e rótulos 131
23.2.1 A verossimilhança completa . . . 133
23.2.2 A verossimilhança incompleta . . . 134
23.3 O algoritmo EM 134
23.3.1 A distribuição de $\mathbf{Z}|\mathbf{Y}=\mathbf{y}$ . . . 135
23.3.2 Escolhendo $\theta^{(0)}$ . . . 135
23.3.3 $\mathbb{E}(Z_i|\mathbf{Y}=\mathbf{y},\theta^{(0)})$ . . . 136
23.3.4 Os passos do algoritmo EM . . . 136
23.3.5 Resumo do algoritmo EM . . . 137
23.4 Exemplos de uso do algoritmo EM 137
23.5 Convergência d algoritmo EM 137
23.5.1 Definições preliminares . . . 137
23.5.2 Desigualdade de Jensen . . . 138

24 Sufficiency and Exponential Family . . . 141
24.1 Introduction 141
24.2 Definition of sufficient statistics 142
24.3 Neyman–Fisher Factorization Theorem 145
24.4 Multidimensional Case 148
24.5 Sufficiency in Linear Regression 148
24.6 Sufficiency in Logistic Regression 151
24.7 Existence and non-uniqueness of sufficient statistics 152
24.7.1 Optional reading: Another not-so-useful sufficient statistics . . . 153
24.8 Theorem of Darmois-Koopman-Pitman 154
24.9 MLE is a function of sufficient statistics 155
24.10 MLE is approximately sufficient 156
24.11 Missing 157

25 Generalized Linear Models (GLM) . . . 159
25.1 Introdução 159
25.2 Definição 159
25.3 Exemplos 160
25.4 GLM em um só algoritmo 160

26 Stochastic Approximation and SGD 161
26.1 Introduction 161
26.2 Target of the Robbins-Monro Algorithm in MLE: Role of θ and θ^* 164
26.3 NOvo - NOVO - NOVO - NOVO 166
26.4 Stochastic Approximation and the Robbins-Monro Algorithm 166
26.4.1 MLE as a Root-Finding Problem 166
26.4.2 Example 1: Normal Distribution with Known Variance 166
26.4.3 Example 2: Bernoulli Distribution (Unknown θ) 167
26.4.4 Mini-Batch Robbins-Monro on Fixed Dataset 167
26.5 Robbins-Monro Estimation and the Role of the Population Score Equation 168
26.6 Stochastic Approximation - Fukunaga Text 169
26.7 Introduction to Stochastic Gradient Descent for MLE 171
26.8 MLE Estimation of the Rate Parameter λ of an Exponential Distribution via Robbins-Monro 171
26.9 Robbins-Monro Convergence Theorem 173
26.10 Practical Considerations 174
26.10.1 Convergence of the Robbins-Monro Algorithm for Multivariate MLE Estimation 174
26.11 Robbins-Monro Method for Linear Regression 176
26.12 Robbins-Monro Method for Logistic Regression 177
26.13 Two Conceptual Questions on the Robbins-Monro Method 178

27 Refined Error Decomposition 181

28 Model Selection 183
28.1 Introduction 183
28.2 Entropia de um evento 183
28.3 Introdução 199
28.3.1 The dependence of two random vectors may be quantified by mutual information. 199
28.4 Kullback-Leibler distance and Fisher information 200
28.5 Kullback-Leibler 201
28.5.1 KL e MLE 202
28.6 Wasserstein Distance 202
28.6.1 Univariate Case: $x \in \mathbb{R}$ 202
28.6.2 Multivariate Case: $\mathbf{x} \in \mathbb{R}^n$ 203
28.7 Kullback-Leibler and Fisher Information 204
28.7.1 The dependence of two random vectors may be quantified by mutual information. 204
28.8 Kullback-Leibler distance and Fisher information 208
28.9 AIC 208

bibliography 211
Books 211
Articles 211

Index 213

16. Linear Regression

16.1 Introdução

Neste capitulo, apresentar apenas o MODELO de regressao linear com regressors tipo x^2 e categoricas. Dizer que UM metodo de estimacao dos parametros é o metodo de minimos quadrados: projecao linear. Coincide com o EMV, a ser explicado nos capitulos 18 e 19. Por ora, apenas saiba que este metodo está embutido na funcao glm (ou lm) do R. Como usar lm. Como interpretar os parametros.

Predição de preços imobiliários Qual o valor de um imóvel? Existem softwares para fazer esta predição de forma automática a partir de várias características do imóvel. Menos subjetivo, mais rápido, primeira avaliação. Como um software desses pode ser construído?

Preços de imóveis Coletamos preços de 1500 imóveis a venda no mercado de BH. Alguns são caros, outros são baratos. O que faz com que os preços dos imóveis variem? As três coisas mais importantes que afetam o valor de um imóvel são: em primeiro lugar, sua localização. Em segundo lugar, a localização e, em terceiro lugar, a localização. Depois disso, vêm os demais fatores.

Localização Existem maneira sofisiticadas de introduzir a localização num modelo de regressão linear mas nós vamos adotar uma estratégia simples aqui. Localização é status socio-econômico; Status é mensurado por renda. Renda é medida pelo IBGE em 2000 pequenas áreas da cidade. Renda do "chefe do domícilio". Então: "localização" = renda média da região onde está o imóvel.

Outras características do imóvel

- Ano da construção
- Área total do imóvel
- Número de quartos
- Número de suítes
- Quantos aptos por andar?
- Possui salão de festas? 0 ou 1
- Possui piscina? 0 ou 1
- ETC...

Ao todo, usamos 30 características numéricas para cada um dos 1500 imóveis.

Visão matricial Organizamos os dados como vetores e matrizes. Preços: um vetor Y de dimensão

1500. As características: uma matriz 1500×30

- Cada linha = um imóvel
- $1^{a.}$ coluna = renda média da região
- $2^{a.}$ coluna = ano da construção
- $3^{a.}$ coluna = área total
- Etc.

Visão matricial

Preços de 1500 imóveis (vetor de dimensão 1500)

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_{1499} \\ y_{1500} \end{pmatrix}$$

$$X = \begin{pmatrix} \text{renda}_1 & \text{área}_1 & \cdots & \text{salão}_1 \\ \text{renda}_2 & \text{área}_2 & \cdots & \text{salão}_2 \\ \vdots & \vdots & \vdots & \\ \text{renda}_{1499} & \text{área}_{1499} & \cdots & \text{salão}_{1499} \\ \text{renda}_{1500} & \text{área}_{1500} & \cdots & \text{salão}_{1500} \end{pmatrix}$$

30 características de 1500 imóveis (Matriz X de dimensao 1500×30)

Preço é uma soma ponderada

Procuramos um modelo matemático simples que possa explicar, a partir das características, porque alguns imóveis são caros e outros são baratos. Área total: quanto maior o imóvel, maior o preço.

Influência de área

Vamos fazer uma primeira aproximação, talvez muito grosseira e sujeita a revisões. Mas será um ponto de partida. Vamos imaginar que, APROXIMADAMENTE, o preço aumenta linearmente com a área do imóvel . Isto é, que o preço $Y \approx a + b * \text{área}$.

Um gráfico com 150 imóveis

Cada ponto é um imóvel. O eixo vertical tem os preços (em milhares de reais) O eixo horizontal tem as áreas (em metros quadrados)

Parece que o preço é, grosseiramente, uma função linear da área. Isto é, $Y \approx a + b * \text{área}$.

Um gráfico com 150 imóveis

Reta no gráfico corresponde a esta equação: Preço $Y \approx 50 + 2 * \text{área}$. **Área não é tudo**

- Dois imóveis com praticamente a mesma área possuem preços diferentes.
- O que causa a diferença?
- Idade do imóvel?
- Dois imóveis, com áreas iguais: se um for mais velho, provavelmente será mais barato.

Ampliando o modelo inicial

Podemos então imaginar que a idade traz um impacto adicional ao nosso modelo de preço. Neste momento, temos $Y \approx a + b * \text{área}$. Já vimos até mesmo que $a \approx 50$ e $b \approx 2$ Podemos agora acrescentar o impacto de idade imaginando que: $Y \approx a + b * \text{área} + c * \text{idade}$.

Como maior idade reduz o preço, devemos ter $c < 0$.

Um modelo ainda mais complexo

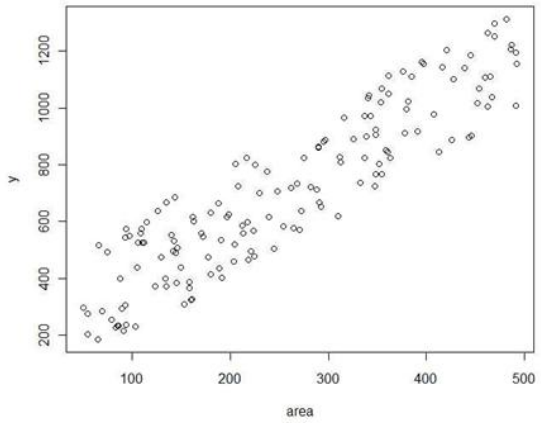


Figure 16.1: Caption

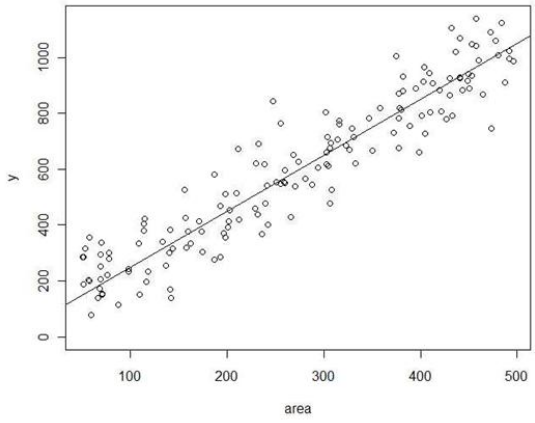


Figure 16.2: Caption

Mas o preço não depende apenas de área e idade. Dois imóveis com mesma área e mesma idade podem ter preços bem diferentes dependendo de:

- Sua localização (renda da sua região)
- Número de suítes
- Número de vagas na garagem
- Etc.

Cada fator pode ser acrescido ao modelo inicial de forma linear.

Modelo mais complexo

Vamos considerar um modelo que, a partir das 30 características do imóvel, fornece uma predição do preço da seguinte forma:

Y é aproximadamente igual a

$$a + b * \text{área} + c * \text{idade} + d * \text{localização} + \text{ETC}\ldots$$

O problema é:

como encontrar os valores de $a, b, c, etc.$ que tornem a aproximação a melhor possível?

O problema de forma matemática

Queremos que cada um desses 1500 valores seja aproximadamente igual a uma combinação linear das 30 características (mais a constante a)

$$\begin{aligned}
y_1 &\approx a + b * \text{área}_1 + c * \text{idade}_1 + \ldots \\
y_2 &\approx a + b * \text{área}_2 + c * \text{idade}_2 + \ldots \\
&\vdots \\
y_{1500} &\approx a + b * \text{área}_{1500} + c * \text{idade}_{1500} + \ldots
\end{aligned}$$

Podemos escrever isto de forma matricial.

O problema de forma matemática

Para facilitar a notação no futuro, vamos escrever os pesos que multiplicam cada característica como b_0 (para a constante), b_1 (para área), b_2 (para idade), ..., b_{30} para a presença ou não de salão de festas

$$\begin{aligned}
y_1 &\approx b_0 + b_1 * \text{área}_1 + b_2 * \text{idade}_1 + \ldots b_{30} * \text{salão}_1 \\
y_2 &\approx b_0 + b_1 * \text{área}_2 + b_2 * \text{idade}_2 + \ldots b_{30} * \text{salão}_2 \\
&\vdots \\
y_{1500} &\approx b_0 + b_1 * \text{área}_{1500} + b_2 * \text{idade}_{1500} + \ldots b_{30} * \text{salão}_{1500}
\end{aligned}$$

O problema de forma matricial

$$\begin{aligned}
y_1 &\approx b_0 + b_1 * \text{área}_1 + b_2 * \text{idade}_1 + \ldots b_{30} * \text{salão}_1 \\
y_2 &\approx b_0 + b_1 * \text{área}_2 + b_2 * \text{idade}_2 + \ldots b_{30} * \text{salão}_2 \\
&\vdots \\
y_{1500} &\approx \underbrace{b_0 + b_1 * \text{área}_{1500} + b_2 * \text{idade}_{1500} + \ldots b_{30} * \text{salão}_{1500}}
\end{aligned}$$

$$b_0 \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \\ 1 \end{pmatrix} + b_1 \begin{pmatrix} \text{área}_1 \\ \text{área}_2 \\ \vdots \\ \text{área}_{1499} \\ \text{área}_{1500} \end{pmatrix} + b_2 \begin{pmatrix} \text{idade}_1 \\ \text{idade}_2 \\ \vdots \\ \text{idade}_{1499} \\ \text{idade}_{1500} \end{pmatrix} + \ldots + b_{30} \begin{pmatrix} \text{salão}_1 \\ \text{salão}_2 \\ \vdots \\ \text{salão}_{1499} \\ \text{salão}_{1500} \end{pmatrix}$$

Os valores $y_1, y_2, \ldots, y_{1500}$ devem ser colocados em um vetor de dimensão 1500.

Forma vetorial

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_{1499} \\ y_{1500} \end{pmatrix} \approx b_0 \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \\ 1 \end{pmatrix} + b_1 \begin{pmatrix} \text{área}_1 \\ \text{área}_2 \\ \vdots \\ \text{área}_{1499} \\ \text{área}_{1500} \end{pmatrix} + b_2 \begin{pmatrix} \text{idade}_1 \\ \text{idade}_2 \\ \vdots \\ \text{idade}_{1499} \\ \text{idade}_{1500} \end{pmatrix} + \ldots + b_{30} \begin{pmatrix} \text{salão}_1 \\ \text{salão}_2 \\ \vdots \\ \text{salão}_{1499} \\ \text{salão}_{1500} \end{pmatrix}$$

Y é um vetor de dimensão 1500 escrito como combinação linear de 31 vetores, cada um deles de dimensão 1500. Problema: encontrar os coeficientes $b_0, b_1, \ldots, b_{30}$ que tornem a aproximação acima a melhor possível.

A solução do problema

Veremos com detalhes mais tarde no curso como resolver este problema. Neste momento, basta dizer que nosso problema fica reduzido a um sistema de equações lineares. Ou ainda, a um problema de inverter uma certa matriz quadrada.

A matriz de desenho X

Seja X a matriz 1500×31 abaixo (note que ela tem uma coluna composta apenas de 1's):

$$X = \begin{pmatrix} 1 & \text{renda}_1 & \text{área}_1 & \cdots & \text{salão}_1 \\ 1 & \text{renda}_2 & \text{área}_2 & \cdots & \text{salão}_2 \\ \vdots & \vdots & \vdots & \vdots & \\ 1 & \text{renda}_{1499} & \text{área}_{1499} & \cdots & \text{salão}_{1499} \\ 1 & \text{renda}_{1500} & \text{área}_{1500} & \cdots & \text{salão}_{1500} \end{pmatrix}$$

Vetores próximos Nosso problema é encontrar os coeficientes $b_0, b_1, \ldots b_{30}$ tais que

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_{1499} \\ y_{1500} \end{pmatrix} \approx b_0 \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \\ 1 \end{pmatrix} + b_1 \begin{pmatrix} \text{área}_1 \\ \text{área}_2 \\ \vdots \\ \text{área}_{1499} \\ \text{área}_{1500} \end{pmatrix} + b_2 \begin{pmatrix} \text{idade}_1 \\ \text{idade}_2 \\ \vdots \\ \text{idade}_{1499} \\ \text{idade}_{1500} \end{pmatrix} + \ldots + b_{30} \begin{pmatrix} \text{salão}_1 \\ \text{salão}_2 \\ \vdots \\ \text{salão}_{1499} \\ \text{salão}_{1500} \end{pmatrix}$$

Ou seja, encontrar $b_0, b_1, \ldots b_{30}$ tais que

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_{1498} \\ y_{1499} \\ y_{1500} \end{pmatrix} \approx \begin{pmatrix} 1 & \text{renda}_1 & \text{área}_1 & \cdots & \text{salão}_1 \\ 1 & \text{renda}_2 & \text{área}_2 & \cdots & \text{salão}_2 \\ \vdots & \vdots & \vdots & \vdots & \\ 1 & \text{renda}_{1499} & \text{área}_{1499} & \cdots & \text{salão}_{1499} \\ 1 & \text{renda}_{1500} & \text{área}_{1500} & \cdots & \text{salão}_{1500} \end{pmatrix} \begin{pmatrix} b_0 \\ b_1 \\ \vdots \\ b_{30} \end{pmatrix} = Xb$$

onde $b = (b_0, \ldots b_{30})^t$.

Isto é, queremos $Xb \approx Y$. Como resolver isto?

Solução: um sisteam linear

Queremos encontrar b para resolver o "sistema" linear $Y \approx Xb$

X é uma matriz 1500×31 e Y é um vetor de 1500 posições.

Como X não é uma matriz quadrada, não é um sistema linear usual: não tem soluç ao, em geral.

Solução: um sistema linear

Um truque para resolver este "sistema" linear: multiple dos dois lados pela matriz X^t (como se fosse uma constante) e troque $\approx$ por $=$

$$\underbrace{(X^tX)}_{A}b \approx \underbrace{X^tY}_{c}$$

Assim, terminamos com um sistema linear legítimo do tipo $Ab = c$
onde $A = X^tX$ é matriz quadrada 31×31 e $c = X^tY$ é vetor com 31 posições.

Solução: um sistema linear

A solução $\mathbf{b} = (b_0, b_1, \ldots, b_{30})^t$ de nosso problema é dada pelo vetor 31×1 que é a solução desta equação matricial:

$$X^tX\mathbf{b} = X^tY$$

Ou ainda, $\mathbf{b} = (X^tX)^{-1}X^tY$. A matriz X^tX é de dimensão 31×1. Inversão via eliminação gaussiana ou, mais profissionalmente, decomposição QR.

16.2 Modelos de Regressão

Exemplo de preço de apto

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_{1499} \\ y_{1500} \end{pmatrix} \approx b_0 \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \\ 1 \end{pmatrix} + b_1 \begin{pmatrix} \text{área}_1 \\ \text{área}_2 \\ \vdots \\ \text{área}_{1499} \\ \text{área}_{1500} \end{pmatrix} + b_2 \begin{pmatrix} \text{idade}_1 \\ \text{idade}_2 \\ \vdots \\ \text{idade}_{1499} \\ \text{idade}_{1500} \end{pmatrix} + \ldots + b_{30} \begin{pmatrix} \text{salão}_1 \\ \text{salão}_2 \\ \vdots \\ \text{salão}_{1499} \\ \text{salão}_{1500} \end{pmatrix}$$

Y é um vetor de dimensão 1500 escrito como combinação linear de 31 vetores, cada um deles de dimensão 1500. Problema: encontrar os coeficientes $b_0, b_1, \ldots, b_{30}$ que tornem a aproximação acima a melhor possível.

A matriz de desenho X Seja X a matriz 1500×31 abaixo (note que ela tem uma coluna composta apenas de 1's):

$$X = \begin{pmatrix} 1 & \text{renda}_1 & \text{área}_1 & \cdots & \text{salão}_1 \\ 1 & \text{renda}_2 & \text{área}_2 & \cdots & \text{salão}_2 \\ \vdots & \vdots & \vdots & \vdots & \\ 1 & \text{renda}_{1499} & \text{área}_{1499} & \cdots & \text{salão}_{1499} \\ 1 & \text{renda}_{1500} & \text{área}_{1500} & \cdots & \text{salão}_{1500} \end{pmatrix}$$

Vetores próximos Nosso problema é encontrar os coeficientes $b_0, b_1, \ldots b_{30}$ tais que

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_{1499} \\ y_{1500} \end{pmatrix} \approx b_0 \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \\ 1 \end{pmatrix} + b_1 \begin{pmatrix} \text{área}_1 \\ \text{área}_2 \\ \vdots \\ \text{área}_{1499} \\ \text{área}_{1500} \end{pmatrix} + b_2 \begin{pmatrix} \text{idade}_1 \\ \text{idade}_2 \\ \vdots \\ \text{idade}_{1499} \\ \text{idade}_{1500} \end{pmatrix} + \ldots + b_{30} \begin{pmatrix} \text{salão}_1 \\ \text{salão}_2 \\ \vdots \\ \text{salão}_{1499} \\ \text{salão}_{1500} \end{pmatrix}$$

Ou seja, encontrar $b_0, b_1, \ldots b_{30}$ tais que

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_{1498} \\ y_{1499} \\ y_{1500} \end{pmatrix} \approx \begin{pmatrix} 1 & \text{renda}_1 & \text{área}_1 & \cdots & \text{salão}_1 \\ 1 & \text{renda}_2 & \text{área}_2 & \cdots & \text{salão}_2 \\ \vdots & \vdots & \vdots & \vdots & \\ 1 & \text{renda}_{1499} & \text{área}_{1499} & \cdots & \text{salão}_{1499} \\ 1 & \text{renda}_{1500} & \text{área}_{1500} & \cdots & \text{salão}_{1500} \end{pmatrix} \begin{pmatrix} b_0 \\ b_1 \\ \vdots \\ b_{30} \end{pmatrix} = Xb$$

onde $b = (b_0, \ldots b_{30})^t$. Isto é, queremos $Y \approx Xb$. Como resolver isto?

Solução: projeção ortogonal

X é uma matriz 1500×31, Y e Xb são vetores 1500-dim. Além disso, Xb é uma combinação linear das colunas da matriz X.

Queremos encontrar b tal que o vetor Xb seja o mais próximo possível do vetor Y.

Queremos $\widehat{b} = \arg\min_b ||Y - Xb||^2$

Seja $\mathfrak{M}(X)$ o sub-espaço vetorial de $\mathbb{R}^{1500}$ formado pelas combinações lineares das colunas de X.

Se as colunas de X são linearmente independentes, então $\mathfrak{M}(X)$ é um espaço de dimensão igual ao número de colunas de X (que é 31, no nosso exemplo).

Solução: Xb é a projeção ortogonal de Y em $\mathfrak{M}(X)$.

Projeção ortogonal

Seja W um sub-espaço vetorial de $\mathbb{R}^n$. A projeção ortogonal de um vetor $Y \in \mathbb{R}^n$ em W é o vetor w tal que $w \perp Y - w$. Matriz de projeção ortogonal em $\mathfrak{M}(X)$ é $H = X(X'X)^{-1}X'$ Para qualquer vetor $Y \in \mathbb{R}^{1500}$, o vetor $HY \in \mathfrak{M}(X)$ Aleém disso, HY é a projeção ortogonal de Y em $\mathfrak{M}(X)$. De fato,

$$HY \,.\, (Y - HY) = Y^tHY - Y^tH^tHY = Y^tHY - Y^tHHY = 0$$

pois $H^tH = HH = H$ (verifique isto você mesmo)

Solução de mínimos quadrados

$\widehat{b} = \arg\min_b ||Y - Xb||^2$ Matriz de projeção ortogonal em $\mathfrak{M}(X)$ é $H = X(X'X)^{-1}X'$ Para qualquer vetor $Y \in \mathbb{R}^{1500}$, o vetor $HY \in \mathfrak{M}(X)$ HY é a projeção ortogonal de Y em $\mathfrak{M}(X)$. Assim, a solução Xb é $HY = X(X'X)^{-1}X'Y$ Isto é, $b = (X'X)^{-1}X'Y$

Uma visão estocástica

Uma abordagem mais probabilística vai permitir estudar melhor as propriedades do estimador de mínimos quadrados. Vamos assumir que o vetor Y é composto de n v.a.'s *independentes.* Como estas v.a.'s assumem seus valores altos e baixos? Porque alguns aptos tem preços altos e outros possuem preços baixos? Vamos explicar como esta variação ocorre quebrando suas causas em dois componentes:

Causas determinadas pelos atributos na matriz X Outras causas. **Um modelo para a distribuição de** Y

Vamos considerar o i-ésimo apto com atributos representado pelo vetor-linha p-dim

$$x_i = (1, x_{i1}, x_{i2}, \ldots, x_{i,p-1})$$

Procuramos ver o preçco Y_i deste apto aproximadamente como uma função linear dos atributos:

$$Y_i \approx \beta_0 + \beta_1 x_{i1} + \ldots + x_{i,p-1}\beta_p$$

uma combinação dos atributos com pesos FIXOS para todo i. Vamos agora pensar em Y_i como uma variável aleatória. Vamos escrever

$$Y_i = \beta_0 + \beta_1 x_{i1} + \ldots + x_{i,p-1}\beta_p + \varepsilon_i$$

A quantidade aleatória ε_i é o "erro" aletório de aproximação de Y_i pela função linear dos atributos.

O erro ε_i

Considere o "erro" aletório

$$\varepsilon_i = Y_i - (\beta_0 + \beta_1 x_{i1} + \ldots + x_{i,p-1}\beta_p)$$

Podemos SEMPRE assumir que $\mathbb{E}(\varepsilon_i) = 0$. Para ver isto, suponha que $\mathbb{E}(\varepsilon_i) = \alpha \neq 0$. Defina um novo erro aleatório ε_i^* da seguinte forma:

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 x_{i1} + \ldots + x_{i,p-1}\beta_p + \varepsilon_i \\ &= \beta_0 + \beta_1 x_{i1} + \ldots + x_{i,p-1}\beta_p + \varepsilon_i - \alpha + \alpha \\ &= (\beta_0 - \alpha) + \beta_1 x_{i1} + \ldots + x_{i,p-1}\beta_p + (\varepsilon_i - \alpha) \\ &= \beta^* + \beta_1 x_{i1} + \ldots + x_{i,p-1}\beta_p + \varepsilon_i^* \end{aligned}$$

O 1o. termo do lado direito é uma combinação linear dos atributos O novo erro tem

$$\mathbb{E}(\varepsilon_i^*) = \mathbb{E}(\varepsilon_i - \alpha) = \mathbb{E}(\varepsilon_i) - \alpha = \alpha - \alpha = 0$$

Modelo para Y_i

Assim, temos

$$\begin{aligned} Y_i &= \mathbb{E}(Y_i) + \varepsilon_i \\ &= \beta_0 + \beta_1 x_{i1} + \ldots + x_{i,p-1}\beta_p + \varepsilon_i \\ &= x_i'\beta + \varepsilon_i \end{aligned}$$

Supomos que os atributos em x_i NÃO SÃO aleatórios. Mas, num apto escolhido ao acaso, como supor que o némero de quartos não é uma v.a.? Nós assumimos um modelo DISCRIMINATIVO: Condicionamos nos valores dos atributos em x_i. DADAS AS CARACTERÍSTICAS do apto selecionado em x_i, nós supomos que seu preço é

$$Y_i = x_i'\beta + \varepsilon_i$$

Modelos Discriminativos

A partir dos n dados, um modelo **discriminativo** aprende a distribuição de probabilidade **condicional** de uma das v.a.'s dadas as demais. Vamos isolar uma das variáveis, denotada por Y, a variável resposta. CONDICIONAMOS em valores específicos e fixos $x' = (x_1, \ldots, y_{p-1})$ para as demais variáveis Vamos bolar um modelo para a distribuição de $Y|x$.

Modelo de regressão linear

Temos $Y_1, Y_2, \ldots, Y_n$ v.a.'s independentes mas não i.d. Como a distribuição muda com i? Muda em função dos atributos em x_i', a linha i da matriz X. Até agora temos

$$(Y_i|x_i') = x_i'\beta + \varepsilon_i$$

Como nós condicionamos nos valores do vetor de atributos x_i, o termo $x_i'\beta$ é um valor não-aleatório. Assim, toda a aleatoriedade de Y_i deriva daquela de ε_i:

$$\begin{aligned} (Y_i|x_i') &= \beta_0 + \beta_1 x_{i1} + \ldots + x_{i,p-1}\beta_p + \varepsilon_i \\ &= x_i'\beta + \varepsilon_i \\ &= \mathbb{E}(Y_i|x_i') + \varepsilon_i \\ &= \text{Termo não-aleatório} + \text{Termo aleatório} \end{aligned}$$

Modelo de regressão linear

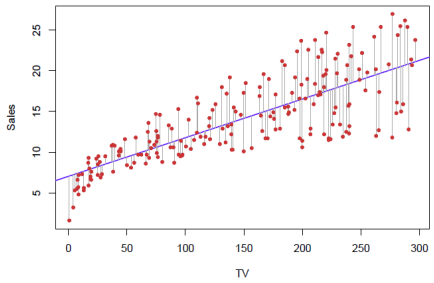


Figure 16.3: Modelo de Regressão linear simples: reta "verdadeira" $\beta_0 + \beta_1 x_1$ e dados $(x_{i1}, y_i$ onde $y_i = \beta_0 + \beta_1 x_{i1} + \varepsilon_i$.

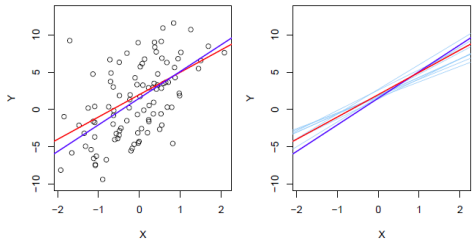


FIGURE 3.3. *A simulated data set.* Left: *The red line represents the true relationship, $f(X) = 2 + 3X$, which is known as the population regression line. The blue line is the least squares line; it is the least squares estimate for $f(X)$ based on the observed data, shown in black.* Right: *The population regression line is again shown in red, and the least squares line in dark blue. In light blue, 10 least squares lines are shown, each computed on the basis of a separate random set of observations. Each least squares line is different, but on average, the least squares lines are quite close to the population regression line.*

Figure 16.4: Esquerda: Reta "verdadeira" $\beta_0 + \beta_1 x_1$ e reta estimada $\widehat{\beta}_0 + \widehat{\beta}_1 x_1$ com os dados. Note que $\beta_0 \neq \widehat{\beta}_0$ e que $\beta_1 \neq \widehat{\beta}_1$. Direita: 10 retas estimadas usando 10 diferentes conjuntos de dados, todos gerados independentemente pelo modelo verdadeiro.

Temos $Y_1, Y_2, \ldots, Y_n$ v.a.'s independentes mas não i.d.

$$(Y_i | x_i') = x_i'\beta + \varepsilon_i = \mathbb{E}(Y_i | x_i') + \varepsilon_i$$

Já aprendemos que ε_i é uma v.a. com $\mathbb{E}(\varepsilon_i)$. Como Y_i é uma v.a. contínua, então ε_i também será uma v.a. contínua Vamos assumir que $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ são i.i.d. $N(0, \sigma^2)$.

Modelo de regressão linear

Seja $p(y|x)$ a densidade de probabilidade da v.a. Y dados os valores em x. Os modelos de regressão caem nesta classe de modelos condicionais. Queremos um modelo para Y (preço do imóvel) quando são conhecidos os valores de área (x_1) e número de quartos (x_2). Fazemos $x = (x_1, x_2)$. Modelo: $(Y | x = (x_1, x_2)) \sim N(\beta_0 + \beta_1 x_1 + \beta_2 x_2, \sigma^2)$

Propriedades do estimador de minimos quadrados

Estimamos β por $\widehat{\beta} = (X'X)^{-1}X'Y$. Note que $\widehat{\beta} = AY$ onde $A = (X'X)^{-1}X'$ é uma matriz $p \times n$ com constantes (isto é, uma matriz não-aleatória). O vetor $\widehat{\beta}$ é um vetor aleatório porque o vetor Y é um vetor aleatório. Como $Y \sim^d N_n(X\beta, \sigma^2 I_n)$ e como $\widehat{\beta} = AY$ temos

$$\widehat{\beta} = AY \sim^d N_p\left(A\mathbb{E}(Y), A\Sigma_Y A'\right)$$

(usando as propriedades da normal multivariada) Substituindo $A = (X'X)^{-1}X'$ na expressão anterior, nós encontramos

$$\widehat{\beta} = AY \sim^d N_p\left(\beta, \sigma^2 (X'X)^{-1}\right)$$

Propriedades do estimador de minimos quadrados

Como

$$\widehat{\beta} \sim^{d} N_p\left(\beta, \sigma^2 (X'X)^{-1}\right) ,$$

temos como consequência que

$$\mathbb{E}\left(\widehat{\beta}\right) = \beta$$

Dizemos que $\widehat{\beta}$ é não-viciado para estimar β.

17. Logistic Regression

17.1 Introdução

O modelo de regressão logística é uma extensão de regressão linear para o caso em que a variável resposta Y possui distribuição binomial. Temos n exemplos ou instâncias de dados. Eles são independentes mas não são i.d. A variável resposta Y_i é *binária*, com apenas dois valores possíveis: 0 ou 1. Temos uma ou mais variáveis explicativas (ou features) no vetor $\mathbf{x}_i$. A regressão logística é modelo estatístico para descrever como a *distribuição* de Y varia com as features no vetor $\mathbf{x}$. Isto é, descrever $Y|\mathbf{x}$, a distribuição de Y condicionada nos valores de $\mathbf{x}$. Como Y é binária, existe apenas um elemento que determina esta distribuição, a probabilidade de observarmos $Y = 1$. Assim, a regressão logística descreve como a probabilidade $\mathbb{P}(Y = 1|\mathbf{x})$ varia como função de $\mathbf{x}$.

Vamos introduzir a regressão logística fazendo uso de um exemplo. Os testes de desenvolvimento de Denver são usados para a detecção precoce de problemas em crianças `https://en.wikipedia.org/wiki/Denver_Developmental_Screening_Tests`. Eles são aplicados em crianças entre o nascimento e os seis anos de idade, para confirmação de suspeitas na avaliação subjetiva do desenvolvimento feitas em consultas rotineiras a pediatras. Composto por 125 testes separados, eles são subdivididos em quatro domínios de funções: pessoal-social, motor-adaptativo, linguagem e motor grosseiro. Os itens incluem a coordenação do olho e da mão, a manipulação de pequenos objetos, a produção de som (balbuciar mamãe e papai), capacidade de reconhecer faces e objets, entender e usar a linguagem, ter controle do corpo para sentar, caminhar ou pular. Os testes são muito variados e cada um deles é apropriado apenas dentro de uma certa faixa de idade. Por exemplo, um teste para verificar se uma criança é capaz de pular não é aplicado a crianças que ainda nem conseguem engatinhar. O fato é que, dos 125 testes, apenas alguns poucos são aplicados num dado exame, em função da idade da criança.

Como mostrado na Figura 17.1, cada um dos 125 itens está representado por uma barra horizontal. O eixo horizontal mostra a idade das crianças em meses. A barra de um dado teste começa na idade em que 25% saudáveis são capazes de executar a tarefa. A barra termina na idade em que 90% das crianças saudáveis executam a tarefa. A barra costuma ser dividida num ponto em que 50% das crianças executam a tarefa. Uma criança de idade x realiza os poucos testes indicados para sua faixa etária. Suponha que a criança sob exame não tenha executado uma

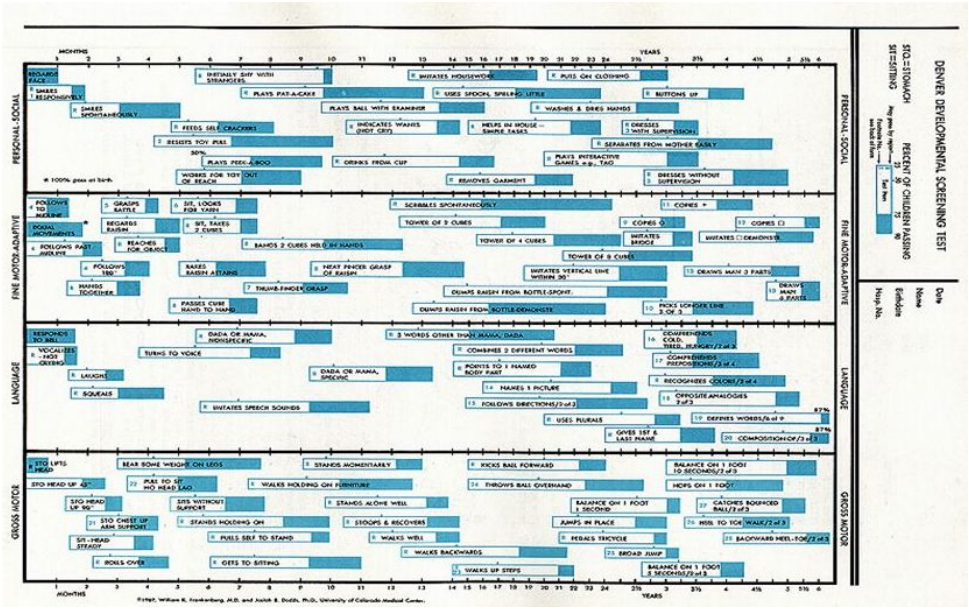


Figure 17.1: Testes de Desenvolvimento de Denver. Eixo horizontal representa a idade. Cada teste é representado por uma caixa retangular. A faixa de idade determinada pela caixa especifica quais ciranças devem realizar aquela tarefa. Os testes estão separados em quatro grupos de acordo com os seus tipos.

dessas tarefas-testes. O pediatra sabe que, dentre crianças saudáveis com aquela idade x, uma certa porcentagem $p(x)$ executa corretamente uma das tarefas-testes. Se $p(x) = 25\%$, não há motivos para preocupação. Mas se $p(x) = 90\%$, um sinal de alerta é aceso e exames mais minuciosos são conduzidos. Como estas porcentagens para cada idade x são determinadas para uso rotineiro nos consultórios? Com regressão logística.

Uma amostra de n crianças de diversas idades e *sabidamente sem problemas de desenvolvimento* procuram executar a tarefa de um dos testes. As crianças são consideradas saudáveis pois elas submetidas a uma exame bastante minuciosos e a um acompanhamento de longo prazo. $Y_i = 1$ denota o sucesso e $Y_i = 0$, o fracasso da i-ésima criança na execução da tarefa. No caso de crianças sem problemas de desenvolvimento, mais cedo ou mais tarde, todas elas acabam executando a tarefa sem erros. O sucesso ou fracasso depende principalmente da idade x da criança. Sendo velha o suficiente, a criança vai executar a tarefa. Por outro lado, se ainda for muito nova, é quase impossível que ela execute a tarefa. As faixas etárias apropriadas variam com o tipo de tarefa.

A Figura 17.2 mostra os dados de 173 crianças com a idade x_i em meses no eixo horizontal e a variável binária y_i no eixo vertical. Os dados y_i estão ligeiramente deslocados por uma pequena quantidade aleatória para melhor visualização. Consideramos que os resultados Y_i das crianças são v.a.'s independentes pois o sucesso ou fracasso de uma das crianças não afeta o desempenho das demais.

Sejam $Y_1, \ldots, Y_n$ ensaios de Bernoulli independentes

$$Y_i = \begin{cases} 1, & \text{com probabilidade } \sigma_i \\ 0, & \text{com probabilidade } 1 - \sigma_i \end{cases}$$

Daqui a pouco vai ficar clara a razão para usarmos o símbolo σ_i para representar a probabilidade de sucesso da i-ésima criança. Esta probabilidade σ_i vai variar de criança para criança e por isto está indexada pelo índice i. Queremos que esta variação individual ocorra em função de sua idade x_i. Isto é, queremos escrever $\sigma_i = g(x_i)$, onde g é uma função matemática adequada. Além disso, queremos que g tenha algumas propriedades:

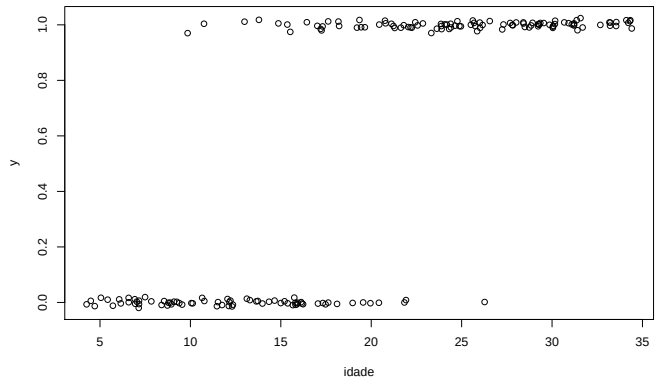


Figure 17.2: Dados de 173 crianças com informação de idade x_i em meses e variável y_i indicando sucesso ou fracasso na realização de uma tarefa. Dados estão deslocados de uma pequena quantidade aleatória para melhor visualização.

- queremos que $g(x)$ tenha seus valores entre 0 e 1 para qualquer idade x pois ela representa uma probabilidade;
- queremos que $g(x)$ seja crescente com x: quanto maior a idade, maior a probabilidade de executar a tarefa;
- queremos que $g(x) \to 1$ quando x cresce;
- e que $g(x) \to 0$ quando x diminui.

17.1.1 A função logística

Existem algumas escolhas populares para a função $g(x)$: logística, probit e log-log complementar. A mais usada é a função logística. Ela possui poucos parâmetros; é simples de entender; é flexível, ajustando-se facilmente a diferentes situações. Voltaremos às outras opções (probit e log-log complementar) mais tarde, no contexto mais geral de modelos lineares generalizados (GLM) no capítulo 25.

A função logística é definida da seguinte forma:

$$\sigma(x) = \frac{1}{1+\exp(-(\beta_0+\beta_1 x))}\,. \tag{17.1}$$

A Figura 17.3 mostra diferentes curvas logísticas que podem ser obtidas variando os dois parâmetros β_0 e β_1. Do lado direito as curvas possuem um comportamento crescente com x enquanto do lado esquerdo temos curvas que decrescem com o aumento de x. A função representada no lado direito possui as propriedades desejadas que listamos anteriormente para representar como a probabilidade de sucesso na realização da tarefa varia com a idade x. O seu gráfico é uma curva na forma de um "S" ao longo do eixo x com assintótas em $y = 0$ e $y = 1$. A Figura 17.3 mostra diferentes logísticas que podem ser obtidas variando os dois parâmetros β_0 e β_1. O parâmetro β_0 está associado com o valor da probabilidade quando a idade for zero. Ele controla onde a curva logística vai se posicionar ao longo do eixo horizontal. Realmente, quando $x = 0$ teremos $\sigma(0) = 1/(1+e^{-\beta_0+\beta_1*0}) = 1/(1+e^{-\beta_0})$. Assim, alteramos a altura da curva no ponto $x = 0$ mexendo em β_0 (e apenas em β_0, deixando β_1 com o seu valor fixo). Isto equivale a mover *rigidamente* a curva ao longo do eixo horizontal, deslocando a curva inteira para a direita ou para a esquerda. Assim, o parâmetro β_0 serve para localizar a curva em alguma região do eixo x.

O valor de β_1 está associado com a forma do S, mais alongado ou com uma subida mais

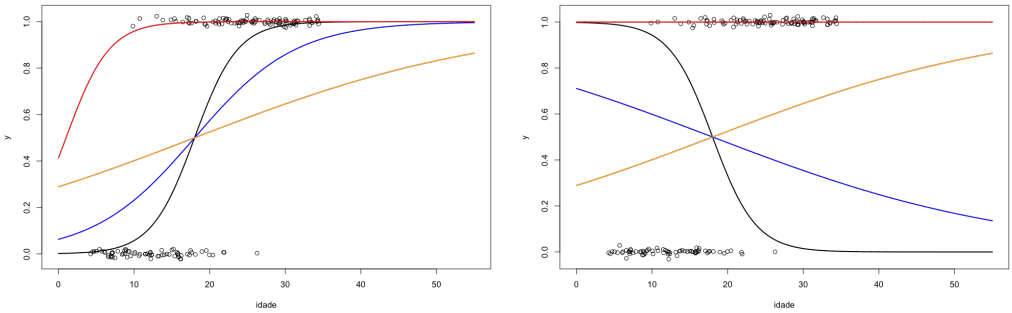


Figure 17.3: Algumas curvas logísticas. As curvas do laesquerdo possuem $\beta_1 > 0$. As do lado direito possuem $\beta_1 < 0$.

íngreme. Para verificar isso, primeiro olhamos o ponto $\hat{x}$ em que a curva $\sigma(x)$ atinge a altura $1/2$:

$$1/2 = \sigma(x) = 1/(1+e^{-(\beta_0+\beta_1 x)} \rightarrow \hat{x} = -\beta_0/\beta_1\,.$$

Em seguida, podemos calcular a derivada $\sigma'(x)$, que mede a inclinação da reta tangente em x:

$$\sigma'(x) = \frac{-\beta_1 e^{-(\beta_0+\beta_1 x)}}{(1+e^{-(\beta_0+\beta_1 x)})^2}$$

Visualmente, olhando a Figura 17.3, é possível intuir que o ponto $\hat{x} = -\beta_0/\beta_1$ é aquele em que a inclinação da reta tangente é máxima. Nos extremos, com x muito pequeno ou x muito grande, a inclinação da reta tangente é praticamente zero. No centro, temos a inclinação máxima. Isto ocorre no ponto ponto $\hat{x} = -\beta_0/\beta_1$ (para verificar, tome a derivada segunda $\sigma''(x)$ e iguale a zero para encontrar o ponto de máximo da derivada primeira). Neste ponto $\hat{x} = -\beta_0/\beta_1$ de máximo da derivada primeira temos

$$\sigma'(\hat{x}) = \sigma'(-\beta_0/\beta_1) = -\beta_1/2\,.$$

Encontramos uma interpretação simples para o parâmetro β_1. Ele mede a inclinação da reta tangente à curva logística no ponto $\hat{x}$ em que $\sigma(\hat{x}) = 1/2$. Isto é, ele mede quão íngreme é a subida (ou descida) da curva no seu ponto central. Ele é a (metade da) inclinação da curva no ponto central, quando a probabilidade de sucesso fica igual a 1/2. Se $\beta_1 > 0$, a curva "S"tem o formato crescente com x, aquele das curvas do lado esquerdo da Figura 17.3. Se $\beta_1 < 0$, a curva fica espelhada ao contrário, decrescendo com x. Um valor β_1 muito positivo leva a uma curva com uma mudança muito íngreme dos valores $\sigma(x) \approx 0$ para os valores $\sigma(x) \approx 1$. A medida em que β_1 vai ficando próximo de zero, a curva vai ficando achatada, mudando de valor muito lentamente (como a curva laranja na Figura 17.3). Mudar apenas β_1 significa acelerar ou retardar a passagem do estágio de não execucção quase certa para o estágio de execução quase certa da tarefa.

É a forma de s da curva que nos leva a usar o símbolo $\sigma(x)$ para a probabilidade de sucesso. A letra grega σ produz o mesmo som que "s" em português e outras línguas latinas. Cuidado para não confundir esta letra σ usada aqui com a mesma letra σ usada para representar o desvio-padrão (ou σ^2 para representar a variância). São conceitos completamente diferentes e apenas usam o mesmo símbolo σ.

17.1.2 Estimando β_0 e β_1

Nosso interesse é estimar os valores de β_0 e β_1 para traçar o gráfico de x versus $\sigma(x)$ para que o pediatra possa tomar as decisões adequadas. A curva logística $\sigma(x)$ deve ser compatível com

os dados da amostra. A Figura 17.3 mostra algumas curvas claramente incompatíveis com os dados observados. A curva vermelha do lado esquerdo, por exemplo, prediz probabilidades de sucesso $\sigma(x)$ muito altas a partir da idade $x = 5$ meses. Deveríamos então ter quase todas as crianças com $Y = 1$. Não é isto que vemos nos dados, onde aproximadamente apenas metade das crianças tiveram $Y = 1$. A curva laranja representa outra escolha de valores para os parâmetros β_0 e β_1 e ela também não é consistente com os dados observados. Ela tem um comportamento crescente com x mas este crescimento é muito lento. Considerando os dados, vemos que para $x > 20$, praticamente todas as crianças tem sucesso. No entanto, pela curva laranja, a probabilidade de sucesso sobe aproxidamente de 0.50 para 0.70 entre $x = 20$ e $x = 40$. Outras escolhas para β_0 e β_1 produzem curvas logísticas que são muito razoáveis como modelos para gerar os dados que realmente observamos. Este é o caso da curva preta no lado esquerdo que parece adequada para representar a probabilidade de sucesso $\sigma(x)$. Vamos derivar um critério objetivo para estimar os parâmetros β_0 e β_1. Este critério é a função de verossimilhança, um tópico fundamental em ciência dos dados e assunto central neste livro a partir do capítulo 18.

O modelo estatístico para *uma única* observação é dado por

$$Y_i = \begin{cases} 1, & \text{com probabilidade } \sigma_i = \sigma(x_i) = 1/(1+\exp(-(\beta_0+\beta_1 x_i))) \\ 0, & \text{com probabilidade } 1-\sigma_i \end{cases}$$

Vamos adotar a função logística para modelar a probabilidade como função da idade x_i da i-ésima criança:

$$\mathbb{P}(Y_i = 1|x_i) = \sigma_i = \sigma(x_i) = \frac{1}{1+\exp(-(\beta_0+\beta_1 x_i))}$$

A probabilidade de fracasso pode ser escrita de várias formas similares e vamos fazer uso ora de uma delas, ora de outra dependendo da conveniência:

$$\begin{aligned} 1-\sigma(x_i) &= \mathbb{P}(Y_i = 0|x_i) = 1-\mathbb{P}(Y_i = 1|x_i) \\ &= 1-\frac{1}{1+e^{-(\beta_0+\beta_1 x_i)}} \\ &= \frac{e^{-(\beta_0+\beta_1 x_i)}}{1+e^{-(\beta_0+\beta_1 x_i)}} \\ &= \frac{1}{1+e^{(\beta_0+\beta_1 x_i)}} \quad \text{multiplic. num e den por } e^{(\beta_0+\beta_1 x_i)} \end{aligned}$$

Não existe erro de digitação aqui. A última expressão é mesmo muito parecida com aquela de σx_i, a diferença residindo no sinal do expoente no denominador.

Um truque importante é que podemos escrever a probabilidade de Y ser 0 ou 1 usando apenas uma única linha:

$$\mathbb{P}(Y_i = y_i) = \sigma_i^{y_i}(1-\sigma_i)^{1-y_i}. \tag{17.2}$$

Usando a expressão em (17.2), se $y_i = 0$, temos

$$\mathbb{P}(Y_i = 0) = \sigma_i^0(1-\sigma_i)^{1-0} = 1-\sigma_i;$$

e se $y_i = 1$, temos

$$\mathbb{P}(Y_i = 1) = \sigma_i^1(1-\sigma_i)^{1-1} = \sigma_i(1-\sigma_i)^0 = \sigma_i.$$

Assim, obtemos os dois valores possíveis para as probabilidades da variável binária Y_i com uma única linha de código. Estepequeno truque parece trivial mas será importante para criar uma função objetivo (função de verossimilhança) que poderá ser maximizada sem dificuldade.

Vamos denotar por $\theta = (\beta_0, \beta_1)$ o vetor de parâmetros envolvidos na função logística. Suponha que tenhamos apenas $n = 5$ crianças com idades

$$(x_1, x_2, x_3, x_4, x_5) = (5, 12, 22, 25, 30)\,.$$

e com os seguintes resultados do teste: $(0,1,0,1,1)$. Este é o evento observado para o vetor aletório $\mathbf{Y}$ e podemos escrever este evento como $[\mathbf{Y} = (y_1, \ldots, y_5) = (0,1,0,1,1)]$. Pela independência das v.a.'s, a probabilidade de observarmos o evento $[\mathbf{Y} = (0,1,0,1,1)]$ (dadas as idades x_i) é igual a

$$\begin{aligned}
\mathbb{P}(\mathbf{Y} = (0,1,0,1,1) \mid x_1 = 5, \ldots, x_5 = 30) &= \mathbb{P}(Y_1 = 0, \ldots, Y_5 = 1 | x_1 = 5, \ldots, x_5 = 30) \\
\text{(pela indep)} &= \mathbb{P}(Y_1 = 0 | x_1 = 5)\mathbb{P}(Y_2 = 1 | x_2 = 12) \ldots \mathbb{P}(Y_5 = 1 | x_5 = 30) \\
&= (1 - \sigma_1)\sigma_2(1 - \sigma_3)\sigma_4\sigma_5
\end{aligned}$$

Assumindo de maneira bem arbitrária que $\theta = (\beta_0, \beta_1) = (-6.3, 0.35)$, vamos obter a probabilidade de gerar os dados observados $[\mathbf{Y} = (0,1,0,1,1)]$ com o modelo logístico (e idades x_i associadas). Para enfatizar que a probabilidade é calculada com estes valores arbitrários para os parâmetros, vamos carregar um pouquinho mais a notação deixando isto explícito:

$$\sigma(x) = \mathbb{P}(Y = 1 | \theta = (-6.3, 0.35), x) = \frac{1}{1 + e^{-(-6.3 + 0.35x)}}\,.$$

Vamos denotar a probabilidade $\mathbb{P}(\mathbf{Y} = (0,1,0,1,1) | \theta = (-6.3, 0.35), x_1, \ldots, x_5)$ simplesmente por L e calcular:

$$\begin{aligned}
L &= (1 - \sigma(5))\ \sigma(12)\ (1 - \sigma(22))\ \sigma(25)\ \sigma(30) \\
&= \frac{1}{1 + e^{(-6.3 + 0.35 * 5)}}\ \frac{1}{1 + e^{-(-6.3 + 0.35 * 12)}}\ \frac{1}{1 + e^{(-6.3 + 0.35 * 22)}}\ \frac{1}{1 + e^{-(-6.3 + 0.35 * 25)}}\ \frac{1}{1 + e^{-(-6.3 + 0.35 * 30)}} \\
&= 0.01936855
\end{aligned}$$

Vamos refazer este cálculo obtendo essa probabilidade L com diferentes valores de $\theta = (\beta_0, \beta_1)$. A probabilidade L será uma função do valor fixado para θ. Por isto, vamos escrever $L(\theta) = L(\beta_0, \beta_1)$. Acima, nós calculamos o valor de $L(-6.3, 0.35)$. Para um vetor θ genérico, teremos:

$$\begin{aligned}
L(\theta) &= L(\beta_0, \beta_1) \\
&= (1 - \sigma(5))\ \sigma(12)\ (1 - \sigma(22))\ \sigma(25)\ \sigma(30) \\
&= \frac{1}{1 + e^{\beta_0 + \beta_1 * 5)}}\ \frac{1}{1 + e^{-(\beta_0 + \beta_1 * 12)}}\ \frac{1}{1 + e^{(\beta_0 + \beta_1 * 22)}}\ \frac{1}{1 + e^{-(\beta_0 + \beta_1 * 25)}}\ \frac{1}{1 + e^{-(\beta_0 + \beta_1 * 30)}}
\end{aligned}$$

A função $L(\theta) = L(\beta_0, \beta_1)$ depende dos parâmetros β_0 e β_1 de forma complicada, difícil de antecipar simplesmente olhando para a expressão matemática acima.

Parece que estamos criando uma grande quantidade de notação sem necessidade. Entretanto, acredite, ela é importante. A probabilidade de ocorrer o que temos em mãos (isto é, de ocorrerem os dados $[\mathbf{Y} = (0,1,0,1,1)]$) depende do valor que fixamos para o vetor de parâmetros $\theta = (\beta_0, \beta_1)$. O evento para o qual estamos calculando a probabilidade está fixado, é $[\mathbf{Y} = (0,1,0,1,1)]$, aquele que foi realmente observado. O que estamos avaliando é como muda a chance de sua ocorrência em função dos diferentes valores que $\theta = (\beta_0, \beta_1)$ pode ter. No cálculo de probabilidades, nós fixamos os parâmetros β_0 e β_1 (por exemplo, com os valores $\beta_0 = -6.3$ e $\beta_1 = 0.35$) e calculamos a probabilidade de diversos eventos ocorrerem, tais como a probabilidade $\mathbb{P}(\mathbf{Y} = (0,0,0,0,0))$ de termos todas as respostas negativas ou a probabilidade de termos pelo menos dois sucessos nos cinco testes realizados. Em estatística, e neste problema de regressão logítica, nós invertemos o problema. Fixamos um evento particular, uma instância dos dados. Fixamos o evento realmente observado, $\mathbf{Y} = (0,1,0,1,1)$, com as cinco idades $x_1, \ldots, x_5$ correspondentes. A seguir, nos perguntamos quais

os valores de $\theta = (\beta_0, \beta_1)$ que são "compatíveis" com este evento que ocorreu. Por exemplo, se fixarmos $\beta_0 = -9.0$ e $\beta_1 = 1.8$, a probabilidade $L(\theta)$ para θ igual a $(-9.0, 1.8)$ é $L(-9.0, 1.8) = 2.6 \times 10^{-14}$, uma probabilidade ridiculamente pequena. Com outros valores de β_0 e β_1 teremos uma probabilidade bem mais alta de gerar o evento $\mathbf{Y} = (0,1,0,1,1)$. Quais são os valores dos parâmetros que poderiam gerar o evento $\mathbf{Y} = (0,1,0,1,1)$ com boa probabilidade? Qual o valor $\hat{\theta}$ que torna máxima a probabilidade $L(\theta)$? Este ponto de máximo é um bom candidato para produzir uma curva logística que se ajusta bem aos dados.

17.2 $L(\theta) = L(\beta_0, \beta_1)$: a função de verossimilhança

A probabilidade $\mathbb{P}(\mathbf{Y} = (0,1,0,1,1) | \theta, x_1, \ldots, x_5)$ do evento observado $\mathbf{Y} = (0,1,0,1,1)$ vista como uma função dos parâmetros β_0 e β_1 é chamada de *função de verossimilhança*. Para enfatizar que vamos olhar para esta expressão como uma função de θ nós a escrevemos como $L(\theta) = L(\beta_0, \beta_1)$. O motivo para usar um termo e notação diferente da probabilidade usual é porque, no cálculo de probabilidades, nós fixamos um modelo probabilístico e avaliamos as probabilidades da ocorrência de diversos eventos que podem ocorrer. Aqui entretanto, o evento já ocorreu e foi $\mathbf{Y} = (0,1,0,1,1)$. Nós não queremos calcular probabilidades de outros eventos potenciais. O que nós queremos é descobrir qual é o modelo probabilístico que pode ter gerado este evento que acabou de ocorrer. Descobrir o modelo é determinar o valor de $\theta) = (\beta_0, \beta_1)$.

Alguns modelos de probabilidade (isto é, alguns valores de θ)) são inverossímeis pois a probabilidade da ocorrência de $\mathbf{Y} = (0,1,0,1,1)$ é absurdamente pequena. A palavra *verossímil* é formada pela conjunção de *vero* (verdadeiro, real, autêntico) e *símil* (semelhante, similar). Algo é verossímil se *parece* verdadeiro, se é plausível, se é semelhante à verdade, se é coerente o suficiente para se passar por verdade. Ao dizer que algo é verossímil, não dizemos que é verdadeiro. Algo é verossímil se parece verdadeiro pois está de acordo com todas as evidências disponíveis.

Assim, alguns valores de θ são inverossímeis, outros são mais verossímeis. Deve existir um ou mais que serão os mais verossímeis. A função de verossmilhança $L(\theta) = L(\beta_0, \beta_1)$ nos permite saber quais valores θ são verossímeis e quais são inverossímeis. A Figura 17.4 mostra, no lado esquerdo, o gráfico da superfície produzida pela função de verossmilhança $L(\theta) = L(\beta_0, \beta_1)$ com β_0 variando em $[-10, 4]$ e β_1 variando em $[-0.2, 0.8]$. Nós estamos usando o exemplo fictício com os cinco dados $\mathbf{Y} = (0,1,0,1,1)$ com idades $(5, 12, 22, 25, 30)$. A função $L(\theta)$ neste gráfico possui um único ponto de máximo. Numa grande região do plano (β_0, β_1), a função $L(\theta)$ é praticamente igual a zero. Esta região contém os valores de θ que são inverossímeis. Eles não são plausíveis. Se de fato o verdadeiro valor de θ fosse um dos pontos desta região em que $L(\theta) \approx 0$ nós estaríamos testemunhando a ocorrência de um evento extremamente raro. Não é impossível mas claramente existem valores para θ que são muito mais verossímeis. Estes valores mais verossímeis estão localizados numa faixa estreita que atravessa o plano (β_0, β_1), onde a função $L(\theta)$ sobe rapidamente para atingir seu máximo. O gráfico do lado direito na Figura 17.4 permite visualizar melhor qual é o ponto de máximo. Ele mostra as curvas de nível dessa superfície e fica claro que o ponto de máximo está aproximadamente em torno da posição $(-2.0, 0.1)$.

Definition 17.2.1 — Maximum Likelihood Estimate. Vamos representar o ponto de máximo da função de verossmilhança por $\hat{\theta}$. Isto é, $\hat{\theta}$ é o valor do argumento θ que maximiza a função $L(\theta)$:

$$\hat{\theta} = \arg\max_{\theta} L(\theta)$$

Ele é chamado de *Estimativa de Máxima Verossimilhança* do parâmetro θ Vamos denotá-lo por MLE a partir do seu nome em inglês, *maximum likelihood estimate*.

Voltando ao caso geral da regressão logística, temos o vetor aleatório $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)$

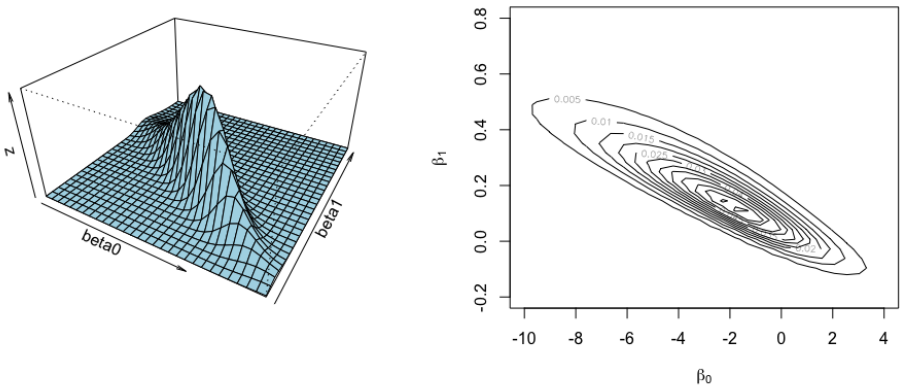


Figure 17.4: Função $L(\theta)$.

assumindo o valor específico $\mathbf{y} = (y_1, \ldots, y_n)$, observado na amostra particular com n crianças. Cada um dos valores y_i é igual a 0 ou 1. A diferença de notação representa uma diferença de conceitos: a letra maiúscula representa uma v.a. (isto é, duas listas, uma de valores possíveis e outra de probabilidades associadas) e a letra minúscula representa apenas um valor numérico fixo (ou 0 ou 1). Sejam $(x_1, x_2, \ldots, x_n)$ as idades correspondentes. A função de verossimilhança é dada por

$$
\begin{aligned}
L(\theta) &= L(\beta_0, \beta_1) = \mathbb{P}(\mathbf{Y} = (y_1, y_2, \ldots, y_n) \mid \theta, x_1, \ldots, x_n) && (17.3)\\
&= = \prod_{i=1}^{n} \sigma_i^{y_i} (1 - \sigma_i)^{1-y_i} && (17.4)\\
&= \prod_{i=1}^{n} \left(\frac{1}{1 + e^{-(\beta_0 + \beta_1 x_i)}} \right)^{y_i} \left(\frac{e^{-(\beta_0 + \beta_1 x_i)}}{1 + e^{-(\beta_0 + \beta_1 x_i)}} \right)^{1-y_i} && (17.5)
\end{aligned}
$$

onde $\sigma_i = 1/(1 + \exp(-(\beta_0 + \beta_1 x_i)))$. Queremos encontrar o estimador de máxima verossimilhança (MLE) de $\hat{\beta}_0$ e $\hat{\beta}_1$ maximizando $L(\beta_0, \beta_1)$.

De volta às $n = 173$ crianças: $L(\beta_0, \beta_1)$ é um produto de $n = 173$ fatores, um para cada criança. Se a i-ésima crianca teve um sucesso (e portanto $y_i = 1$), ela contribui para $L(\beta_0, \beta_1)$ com um fator multiplicativo igual a $\sigma_i = 1/(1 + \exp(-(\beta_0 + \beta_1 x_i)))$. Se a i-ésima crianca teve um fracasso (e portanto $y_i = 0$) então ela contribui um fator igual a

$$
1 - \sigma_i = 1 - \frac{1}{1 + \exp(-(\beta_0 + \beta_1 x_i))} = \frac{\exp(-(\beta_0 + \beta_1 x_i))}{1 + \exp(-(\beta_0 + \beta_1 x_i))}
$$

para $L(\beta_0, \beta_1)$.

A Figura 17.5 mostra a contribuição de cada um de 8 pontos amostrais (x_i, y_i) para a função de verossimilhança $L(\theta)$. Mostramos duas curvas logísticas, ou seja, dois valores para o parâmetro θ. A contribuição de cada ponto amostral é representada pela linha vermelha vertical. Se $y_i = 1$, a linha vermelha tem comprimento σ_i (a altura da curva na idade x_i). Se $y_i = 0$, a linha vermelha possui comprimento $1 - \sigma_i$ e é a distância da curva até o valor 1. Alguns pontos possuem um segmento vermelho tão pequeno, com comprimento tão próximo de zero, que ele acaba não sendo visível. Para cada curva, a sua função de verossimilhança $L(\theta)$ é o produto do comprimento de todos os seus segmentos verticais vermelhos. Claramente, a curva da esquerda é mais compatível com os dados que a curva da direita. De fato, o produto dos comprimentos de seus segmentos vermelhos é bem maior que o produto dos comprimentos dos segmentos vermelhos da direita, a

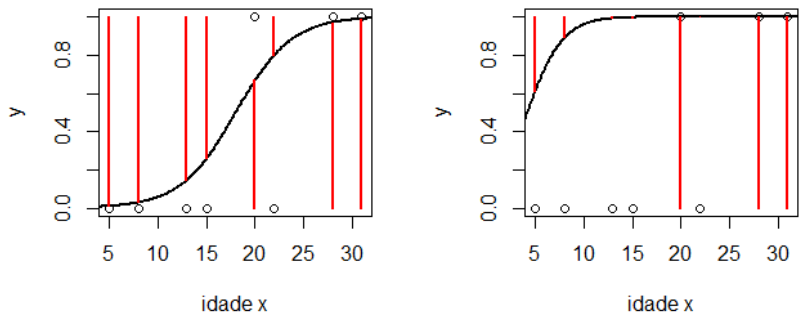


Figure 17.5: Gráfico de duas curvas logísticas com a contribuição de cada ponto da amostra para a função de verossimilhança.

maioria dos quais é praticamente zero. O MLE $\hat{\theta}$ é aquele valor de θ cuja curva logística possuir um valor máximo para esse produto do comprimento dos segmentos vermelhos.

17.3 Obtendo o MLE

Encontrar o MLE no caso da regressão logística requer um método numérico e a escolha universal recai no método de otimização de Newton (ou Newton-Raphson), que será discutido em detalhes no capítulo 18. Neste capítulo vamos aprender apenas os comandos em R necessários para obter o MLE $\hat{\theta}$ sem olhar os algoritmos que encontram este ponto de máximo. No capítulo 18 voltaremos ao problema de encontrar o MLE mostrando os algoritmos usados no caso da regressão logística e de qualquer outro modelo estatístico.

A função `glm` do R calcula o MLE $\hat{\theta}$ e vários outros aspectos relevantes do ajuste logístico. Basta usar o comando `glm(y ~ x, "binomial")$coef` que retorna os coeficientes estimados $\hat{\beta}_0$ e $\hat{\beta}_1$. Vamos gerar alguns dados similares aos dados das crianças do teste de Denver e verificar a capacidade do MLE em recuperar os valores verdadeiros de β_0 e β_1. Primeiro, geramos dados que seguem exatamente o modelo logístico com $\theta = (\beta_0, \beta_1) = (-6.3, 0.35)$:

```
set.seed(12)
x <- runif(173, 4, 35)
px = 1/(1 + exp(- (-6.3 + 0.35*x)))
y = rbinom(173, 1, px)
```

Os dados gerados podem ser vistos na Figura 17.2. A função de verossmilhança com as suas curvas de nível são obtidos com os comandos abaixo e o resultado está na Figura 17.6.

```
beta0 = seq(-10, -4, length=30)
beta1 = seq(0.2, 0.6, length=30)
z = matrix(0, nrow=length(beta0), ncol=length(beta1))

for(i in 1:length(beta0)){
 for(j in 1:length(beta1)){
   pr = 1/(1 + exp(-(beta0[i] + beta1[j]*x)))
   z[i,j] = exp(sum( y*log(pr) + (1-y)*log(1-pr) ))
   }
```

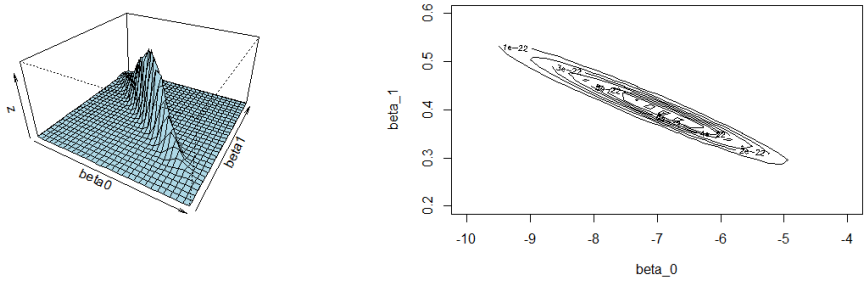


Figure 17.6: Esquerda: Gráfico da função de verossimilhança $L(\beta_0, \beta)$ avaliada nos pontos de um reticulado regular 30×30 no intervalo $(-10, -4)$ (para β_0) e em $(0.2, 0.6)$ (para β_1). Direita: Gráfico de curvas de nível da função de verossimilhança $L(\beta_0, \beta_1)$.

```
}

persp(beta0, beta1, z, theta = 30, phi = 30, expand = 0.5, col = "lightblue")
contour(beta0, beta1, z, xlab=expression(beta_0), ylab=expression(beta_1))
```

Estes gráficos mostram que este ponto de máximo deve ser aproximadamente igual a $\hat{\beta}_0 \approx -7$ e $\hat{\beta}_1 \approx 0.4$. A função `glm` do R calcula este ponto de máximo de forma precisa:

```
glm(y ~ x, "binomial")$coef
(Intercept)           x
 -6.9144513   0.3946116
```

Assim, enquanto o verdadeiro valor de θ usado para gerar os dados foi $(\beta_0, \beta_1) = (-6.3, 0.35)$, o MLE ficou igual a $\hat{\theta} = (\hat{\beta}_0, \hat{\beta}_1) = (-6.91, 0.39)$. O MLE é bem próximo do valor verdadeiro mostrando que maximizar a verossmilhança conseguiu recuperar aproximadamente θ. Estimamos a probabilidade de sucesso com idade x usando $\hat{\beta}_0 = -6.91$ e $\hat{\beta}_1 = 0.39$ na expressão que modela como varia a probabilidade de sucesso com a idade da criança:

$$\widehat{\mathbb{P}}(Y = 1|x) = \frac{1}{1 + \exp(-(-6.91 + 0.39x))} \tag{17.6}$$

A probabilidade de gerarmos a sequência de 173 dados 0's e 1's que realmente observamos é a maior possível usando $\hat{\theta}$. ***Veja que o valor mais verossímil*** $\hat{\theta} = (-6.91, 0.39)$ ***não é igual ao valor*** $\theta = (-6.3, 0.35)$ ***usado para gerar os dados.*** O lado esquerdo da Figura 17.7 mostra as duas curvas logísticas, aquela usada para gerar as 173 instâncias de dados observados (em azul) e a curva em (17.6), usando o MLE (em vermelho). Maximizar uma função objetivo (neste caso, a função de verossiilhança $L(\theta)$) não garante que vamos recuperar exatamente o valor verdadeiro que foi usado para gerar os dados. De fato, mantendo o mesmo θ e gerando várias amostras distintas de v.a.'s binárias, todas com 173 dados, obtemos o MLE $\hat{\theta}$ para cada uma das amostras. Vamos gerar $S = 100$ amostras de 173 v.a.'s $\mathbf{Y} = (Y_1, \ldots, Y_{173})$ mantendo $\theta = (-6.3, 0.35)$. Além disso, mantivemos tambéM inalterado o conjunto de 173 idades usadas para gerar os dados do lado esquerdo da Figura 17.7. O código em R está abaixo e o resultado é mostrado no lado direito da Figura 17.7. Existem 100 diferentes curvas logísticas representando as diferentes estimativas obtidas pelo MLE $\hat{\theta}$. Todas estas amostras foram obtidas nas mesmas condições, com o mesmo tamanho de amostra, o mesmo modelo (mesmo $\theta = (-6.3, 0.35)$), as cirnaças tinham o mesmo

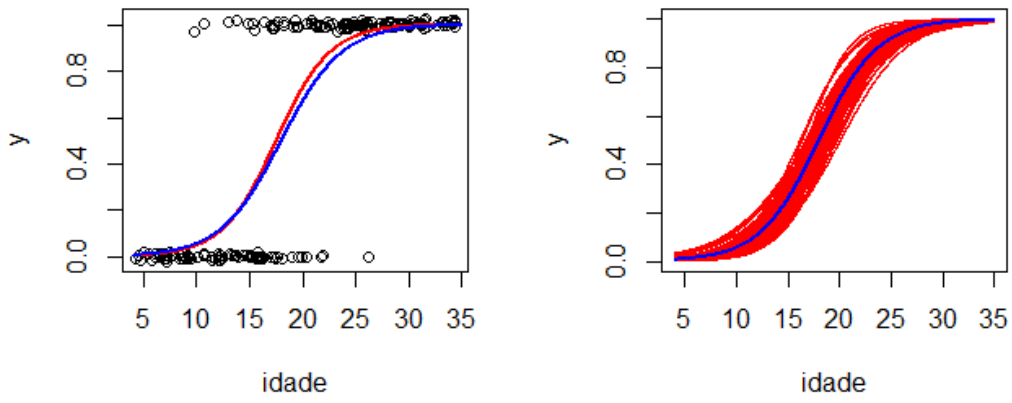


Figure 17.7: Esquerda: Dados, curva logística usada para gerar os dados (azul) e curva logística em (17.6), usando o MLE (vermelho). Direita: Curvas obtidas ao estimar o MLE em 100 simulações do processo gerador dos dados $\mathbf{Y} = (Y_1, \ldots, Y_{173}$ com o mesmo parâmetro $\theta = (-6.3, 0.35)$ e o mesmo conjunto de idades x.

conjunto de idades. Assim, as diferentes curvas refletem a variabilidade que o MLE possui, reflexo da natureza aleatória dos dados $\mathbf{Y} = (Y_1, \ldots, Y_{173})$. Na prática, numa análise real, apenas uma dessas curvas vermelhas é obtida pois teremos apenas uma amostra de dados. Qualquer uma das curvas vermelhas mostradas poderia ser a curva estimada num conjunto de dados reais. VÊ-se que algumas das curvas vermelhas afstam-se da curva real (em azul) mais do outras. Nunca conseguiremos saber se a única amostra que temos em mãos é uma amostra "boa" ou uma amostra "ruim" (isto é, levando a estimativa $\hat{\theta}$ bem próxima do valor verdadeiro $\theta = (-6.3, 0.35)$ ou mais afstada desse valor). A discussão mais aprofundada deste importante assunto será feita no capítulo 19. Por ora, basta sabermos que, em geral, teremos $\hat{\theta}$ com um certo erro de estimação em relação ao verdadeiro valor do parâmetro (caso exista um).

```
set.seed(12)
x <- runif(173, 4, 35)
px = 1/(1 + exp(-0.35*(x-18)))
y = rbinom(173, 1, px)
aux = glm(y ~ x, "binomial")$coef
par(mfrow=c(1,2))
plot(x, y+rnorm(173, 0, 0.01), xlab="idade", ylab="y")
xx = seq(4, 35, by=0.01)
yy = 1/(1+exp(-aux[2]*(xx+aux[1]/aux[2])))
lines(xx, yy, lwd=2, col="red")
yr = 1/(1+exp(-(-6.3 + 0.35*xx)))
lines(xx, yr, lwd=2, col="blue")

plot(xx, yr, type="n", lwd=2, col="blue", xlab="idade", ylab="y", ylim=c(0,1))
Nsim = 100
for(i in 1:Nsim){
  y = rbinom(173, 1, px)
```

```
  aux = glm(y ~ x, "binomial")$coef
  yy = 1/(1+exp(-aux[2]*(xx+aux[1]/aux[2])))
  lines(xx, yy, lwd=0.5, col="red")
}
lines(xx, yr, lwd=2, col="blue")
```

Os valores $\hat{\beta}_0 = -6.91$ e $\hat{\beta}_1 = 0.39$ são os mais verossímeis para os parâmetros tendo em vista os dados observados. Por causa do erro de estimação inerente ao processo de aprendizagem de θ, não são apenas estes dois valores $\hat{\beta}_0 = -6.91$ e $\hat{\beta}_1 = 0.39$ que são razoáveis como estimativas de β_0 e β_1. Outros valores ligeiramente diferentes destes dois também são compatíveis com os dados. Obtemos esta faixa de valores razoáveis calculando os intervalos de confiança para β_0 e β_1, assunto a ser visto no capítulo 18, depois de estudarmos o MLE em geral.

17.4 Regressão Logística Múltipla

Como na regressão linear, podemos ter mais de uma variável independente. A probabilidade de sucesso na execução da tarefa poderia depender de outras variáveis (features) além da idade. Por exemplo, podemos ter a probabilidade de sucesso da i-ésima criança dependendo de:

- x_{i1}: sua idade em meses.
- x_{i2}: escolaridade da mãe (medida em anos de estudo).
- x_{i3}: sexo da criança (0 = menina e 1 = menino).

Podemos incorporar estas variáeis facilmente no modelo de regressão logística. Imaginamos que a probabilidade de sucesso $\sigma_i = \mathbb{P}(Y_i = 1)$ varia com as variáveis x_{i1}, x_{i2} e x_{i3}. Seja $bsx_i = (x_{i1}, x_{i2}, x_{i3})$ o vetor de features que queremos incorporar no modelo de regressão logística. Para deixar mais explícita a dependência, vamos escrever

$$\sigma_i = \mathbb{P}(Y_i = 1 | \mathbf{x}_i = (x_{i1}, x_{i2}, x_{i3}))$$

Esta é a probabilidade de sucesso do i-ésimo item da amostra condicionada nos seus valores para as variáveis independentes ou features. Imaginamos que σ_i cresce com o aumento da idade x_{i1}. Podemos também conjecturar que, para algumas tarefas, σ_i também cresceria com o aumento da escolaridade x_{i2}. Entretanto, dependendo da tarefa poderíamos ter σ_i descrecendo com o aumento x_{i2}. Em alguns casos podemos também ter um efeito nulo desta variável x_{i2}: a probabilidade de sucesso não mudaria se variarmos escolaridade. Para a variável independente x_{i3}, sexo da criança i, podemos ter uma probabilidade de sucesso maior ou menor dependendo do sexo. Podemos também não ter efeito de sexo na probabilidade de modo que, neste caso, esta variável poderia ser eliminada do conjunto de features sem prejuízo. Entretanto, só saberemos isto se, inicialmente, a usarmos junto com todas as demais.

17.4.1 O preditor linear $\eta = \mathbf{x}'\beta$

Imitando o modelo de regressão linear com resposta gaussiana, vamos criar um função linear das variáveis independentes em $\mathbf{x}_i$:

$$\eta_i = \mathbf{x}_i'\beta = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}$$

Esta é a maneira mais simples de criar um índice baseado na variáveis independentes. Esta combinação linear é chamada de *preditor linear* da probabilidade de sucesso. Precisamos especificar como passar do preditor linear $\eta = \mathbf{x}'\beta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$ para a probabilidade de sucesso $\mathbb{P}(Y = 1|\mathbf{x})$. Imaginamos que $\mathbb{P}(Y = 1|\mathbf{x})$ cresce ou diminui quando variamos η por meio de mudanças nas variáveis independentes. Se as variáveis independentes crescerem muito ou diminuirem muito, o preditor linear vai acabar ficando maior que 1 ou menor que 0, valores impossíveis para

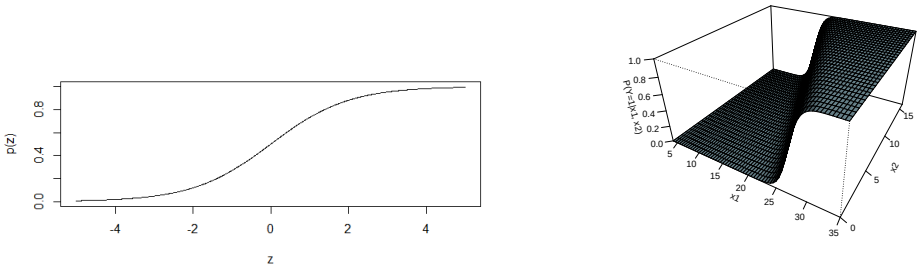


Figure 17.8: Esquerda: Curva logíistica $p(z) = 1/(1+e^{-z})$ onde $z(\mathbf{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$ é o preditor linear. Direita: Gráfico de $\sigma(\mathbf{x}) = \mathbb{P}(Y = 1|\mathbf{x})$ versus idade x_1 e escolaridade x_2 para $x_3 = 0$ (feminino). A probabilidade de sucesso começa a crescer mais cedo para pessoas com mais escolaridade.

uma probabilidade. Precisamos modificar o preditor linear através de uma função de forma que valores muito altos de η levem a $\mathbb{P}(Y = 1|\mathbf{x}) \approx 1$ e valores muito baixos de η levem a $\mathbb{P}(Y = 1|\mathbf{x}) \approx 0$. Além disso, queremos que a probabilidade de sucesso cresça suavemente com o preditor linear η.

O modelo logístico propõe que a probabilidade de sucesso seja a seguinte:

$$\begin{aligned}
\sigma_i &= \mathbb{P}(Y_i = 1|\mathbf{x}_i) \\
&= \frac{1}{1+\exp\left(-(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3})\right)} \\
&= \frac{1}{1+\exp\left(-\mathbf{x}_i'\beta\right)} \\
&= \frac{1}{1+\exp(-\eta_i)}
\end{aligned}$$

Podemos representar graficamente o modelo de duas maneiras. A primeira é imaginar que, para valores fixos de β_0, β_1 e β_2, o predictor linear $\eta(\mathbf{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$ intermedia a propensão ao sucesso $\mathbb{P}(Y = 1|\mathbf{x}) = 1/(1+\exp(-\eta(\mathbf{x}))$. Toma-se o vetor $\mathbf{x}$, calcula-se $\eta(\mathbf{x}) = \mathbf{x}'\beta$ e depois $\mathbb{P}(Y = 1|\mathbf{x})$. Quanto maior o valor de $\eta(\mathbf{x})$, maior a probabilidade de sucesso. (ver gráfico do lado esquerdo na Figura 17.8).

Seja y o vetor de dimensão n com 0's e 1's. Seja x a matriz $n \times k$ com as variáveis independentes como colunas Note que a coluna de 1's para o intercepto é automaticamente adicionada pelo R No R, o valor de β que maximiza $L(\theta)$ no modelo de regressão logística múltipla é obtido com o comando:

```
glm(y ~ x, "binomial")$coef}.
```

Como obter $\mathbb{P}(Y = 1) \approx 1$? A partir da Figura 17.8, vemos que, se $eta \approx 2$, já teremos a probabilidade aproximdamente igual a 0.90 e maior que 0.95 se $\eta > 3$. De forma simétrica, se $eta \approx -2$, a probabilidade de sucesso é aproximadamente 0.10 e menor que 0.05 se $\eta < -3$. Dessa forma, fora da faixa $(-3,3)$, a probabilidade de sucesso é ou bastante pequena ou bastante alta. Itens com configurações das suas variáveis independentes $\mathbf{x}$ levando a eta fora dessa faixa terão probabilidades quase certas de sucesso ou fracasso. Para $\eta \in (-3,3)$, a probabilidade pode variar de 0.05 a 0.95.

Diversas combinações das variáveis independentes no vetor $\mathbf{x}$ levam a $\mathbb{P}(Y = 1|bsx) \approx 1$. Por exemplo, suponha que os coeficientes sejam conhecidos de modo que

$$\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 = -1.2 + 0.6x_1 + 0.1x_2 + 0.7x_3\,.$$

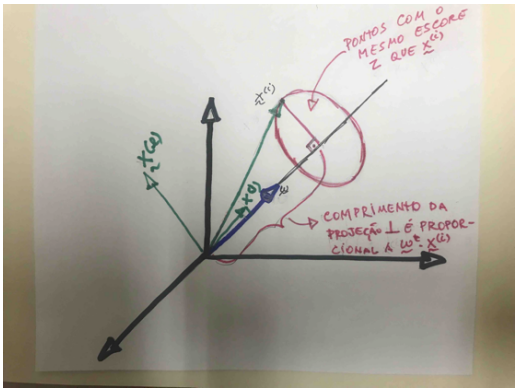


Figure 17.9: O conjunto de pontos amostrais $\mathbf{x}$ que terão o mesmo valor de η e portanto, a mesma probabilidade de sucesso $\sigma(\mathbf{x})$ é o plano formado pelos pontos que têm a mesma projeção ortogonal ao longo do vetor θ (representado aqui como $\mathbf{w}$).

Então tanto $\mathbf{x} = (5.67, 7.98, 0)$ quanto $\mathbf{x} = (3, 17, 1)$ levam a $\eta = 3.0$ e portanto a a uma probabilidade de sucesso aproximadamente igual a 0.95. Infinitas combinações das variáveis em $\mathbf{x}$ levam a um mesmo valor de η. Todas essas crianças com diferentes $\mathbf{x}$ mas o mesmo preditor linear η possuem a mesma probabilidade de sucesso. Variando $\mathbf{x}$ podemos variar η e, em consequência, variar $\mathbb{P}(Y = 1)$. Em suma, é o preditor linear $\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$ que captura todas as formas pelas quais as variáveis independentes $\mathbf{x}$ afetam a probabilidade de sucesso.

Se $\mathbf{x}$ e $\mathbf{x}^*$ tiverem o mesmo valor η, a probabilidade de sucesso será a mesma no dois casos. Para um vetor θ fixo, dois itens com vetor de variáveis independentes $\mathbf{x}$ e $\mathbf{x}^*$ possuem a mesma probabilidade de sucesso se $\mathbf{x}'\theta = \eta = \mathbf{x}^{*\prime}\theta$. Para entender o que isto significa geometricamente, você pode imaginar que o vetor de parâmetros $\theta = (\beta_0, \beta_1, \beta_2, \beta_3)$ determinam uma direção no espaço $\mathbb{R}^4$ das variáveis independentes $\mathbf{x} = (1, x_1, x_2, x_3)$ (estamos redefinindo este vetor colocando a feature constante e igual a 1, associada com β_0, na primeira posição). A projeção ortogonal de $\mathbf{x}$ no vetor θ é dada por

$$\frac{\mathbf{x}'\theta}{||\theta||^2}\theta\,.$$

(ver ?? por referencia desse resultado em capitulos anteriores). Assim, $\mathbf{x}$ e $\mathbf{x}^*$ levam ao mesmo valor de η se, e somente se, a projeção ortogonal deles ao longo de θ é a mesma. A Figura 17.9 mostra esta geometria do preditor linear. Nela, existe um pequena troca de notação: o vetor θ é representado como $\mathbf{w}$.

Outra maneira de visualizar o modelo com várias features é a seguinte: Fixe o valor de todas as variáveis independentes exceto duas delas. No nosso exemplo, vamos fixar sexo, $x_3 = 0$ (feminino), e variar a idade x_1 e a escolaridade da mãe x_2. Por exemplo, tomando $x_{i3} = 0$, vamos supor que o modelo leve a

$$\sigma_i = \mathbb{P}(Y_i = 1|\mathbf{x}_i) = \frac{1}{1+\exp(1.2 - 0.x_{i1} - 0.3x_{i2} - 0.1*0)} = \frac{1}{1+\exp(1.2 - 0.x_{i1} - 0.3x_{i2})}$$

Esta probabilidade é uma função de x_1 e x_2 (já que o sexo x_3 foi fixado). Vamos fazer x_1 variar no intervalo $(5, 35)$ e x_2 no intervalo $(0, 16)$. O resultado está no lado direito da Figura 17.8).

17.5 Outros exemplos usando regressão logística

Regressão logística é um modelo muito popular para modelar dados de resposta binária. Por exemplo, em risco de crédito, seja $Y_i = 0$ ou 1 dependendo do resultado de pagar ou não o

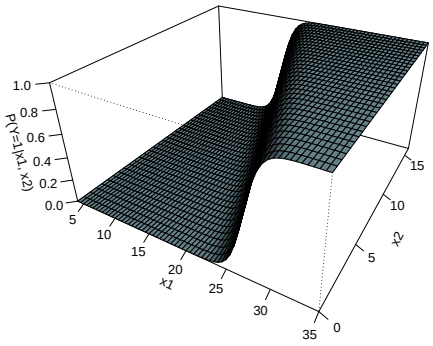


Figure 17.10: Gráfico de $\sigma(\mathbf{x}) = \mathbb{P}(Y = 1|\mathbf{x})$ versus idade x_1 e escolaridade x_2 para $x_3 = 0$ (feminino). A probabilidade de sucesso começa a crescer mais cedo para pessoas com mais escolaridade.

empréstimo Ensaios de Bernoulli: independentes mas não são i.d. O que muda com i? O valor de $p_i = \mathbb{P}(Y_i = 1|\mathbf{x}_i)$ Como esta probabilidade muda? Com variáveis independentes $\mathbf{x}_i$ que afetam a probabilidade de sucesso: idade, sexo, saldo médio na conta, tempo como cliente, etc.

Um outro exemplo clássico são e-mails chegando numa caixa de correio: spam ou não-spam? Qual a probabilidade de ser spam? Depende de vários atributos que podem ser medidos no momento em que o e-mail chega. Abaixo, uma amostra de uma lista de 140 features ou variáveis independentes. Conta-se o número de ocorrências no e-mail.

- número de imagens, número de links, número de caracteres ”, ’?’, ’$’, ’#’, ’(’, ’[’, ’;’
- número de ocorrêencias das palavras “adult”, “bank”, “business”, “free”
- número de palavras em letras maiúsculas (como “TODAY”, “NOW”, etc.)

17.6 E se tivermos um efeito não-monótono?

Suponha que a v.a. binária $(Y|x) \sim Bin(1, p(x))$ tenha uma probabilidade de sucesso que dependa da variável independente x. Vimos como modelar uma situação em que $p(x)$ cresce (ou decresce) monotonicamente com x: usando o modelo logístico. E se $p(x)$ crescer com x até certo ponto, quando então começa a decrescer? Isto é, mais de x é bom para aumentar a probabilidade de sucesso mas só até certo ponto. Depois desse ponto, aumentar x diminui a probabilidade de sucesso $p(x)$.

O lado esquerdo da Figura 17.11 mostra uma possível variação de $p(x)$ seguindo este sobe e desce com x. Esta curva tem a forma de uma logística usando um preditor linear $\eta = \beta_0 + \beta_1 x + \beta_2 x^2$ que envolve a variável x^2. Mais especificamente, nós usamos

$$\begin{aligned}
\sigma(x) &= \mathbb{P}(Y = 1 \mid x) && (17.7)\\
&= 1/(1+\exp(-\eta(x)) && (17.8)\\
&= 1/(1+\exp(-(\beta_0+\beta_1 x+\beta_2 x^2)) && (17.9)\\
&= 1/(1+\exp(-(-2.31+0.48x-0.01x^2)) && (17.10)
\end{aligned}$$

O gráfico no lado direito mostra os dados gerados com esta função $p(x)$. A linha azul representa o ajuste logístico, muito próximo da verdadeira curva que gerou os dados, mostrando que podemos aprender o modelo. O código usado está abaixo.

```
x <- seq(0, 50, by=0.01)
```

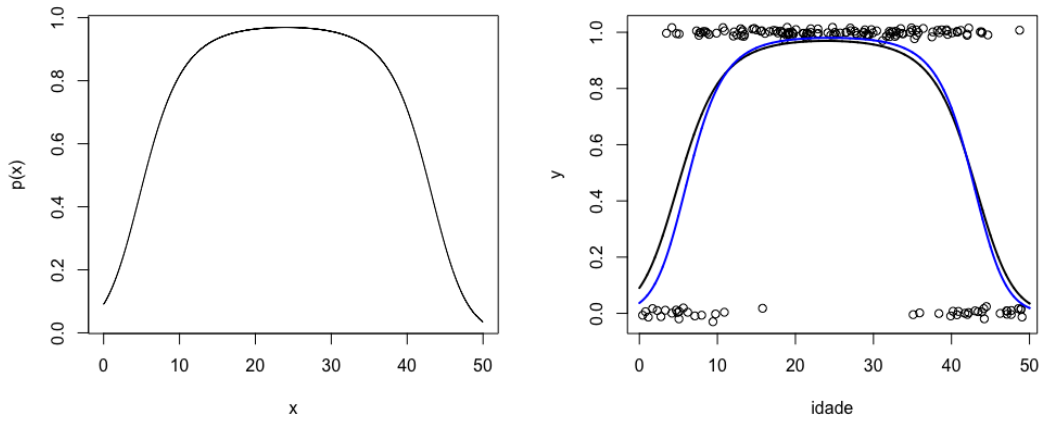


Figure 17.11: Probabilidade $p(x) = 1/(1+\exp(-(\beta_0+\beta_1 x+\beta_2 x^2)))$ com $\beta = (\beta_0,\beta_1,\beta_2) = (-3.31, 0.35, -0.01)$. Dados binomiais gerados com o modelo logístico $p(x) = 1/(1+\exp(-(\beta_0+\beta_1 x+\beta_2 x^2))$ com $\beta = (\beta_0,\beta_1,\beta_2) = (-3.31, 0.35, -0.01)$.

```
b0 <- -2.31; b1 <- 0.48; b2 <- -0.01
px <- 1/(1 + exp(-(b0 + b1*x + b2*x^2)))
par(mfrow=c(1,2))
plot(x, px, type="l", ylab="p(x)")

set.seed(12)
x2 <- runif(173, 0, 50)
px2 <- 1/(1 + exp(-(b0 + b1*x2 + b2*x2^2)))
y <- rbinom(173, 1, px2)
xmat <- cbind(x2, x2^2)
glm(y ~ xmat, family="binomial")$coef
    #  (Intercept)      xmatx2        xmat
    #  -3.26450463  0.58653827 -0.01199062

plot(x2, y+rnorm(173, 0, 0.01), xlab="idade", ylab="y")
lines(x,px, lwd=2)
pfit = 1/(1+exp(-(-3.2645 + 0.5865*x + -0.0120*x^2)))
lines(x, pfit, col="blue", lwd=2)
```

17.7 Feature Engineering

Por este exemplo, vemos nada impede usar polinômios em x no modelo. Eles aparecem como features adicionais, como se fossem outras variáveis independentes. Na verdade, caso seja julgado necessário, qualquer função de x poderia ser usada para ampliar as possibilidades de ajuste do modelo, tal como $\log(x)$ ou $\sqrt{x}$. E podemos ter mais de uma variável independente misturada gerando uma nova feature. Por exemplo, suponha que tenhamos duas variáveis independentes, x_1 e x_2. Muitos efeitos não-lineares dessas variáveis num modelo logístico podem ser introduzidos no modelo. Seja $\sigma = 1/(1+exp(-\eta))$ com as seguinte possibilidades para o preditor linear η:

- $\eta = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i1}^2 + \beta_3 x_{i2} + \beta_4 x_{i2}^2))$
- $\eta = \beta_0 + \beta_1 \log(x_{i1}) + \beta_2 \log(x_{i2}) + \beta_3 x_{i1} x_{i2} + \beta_4 x_{i2}^2$

Em todos os casos, temos um vetor de features expandido usando-se estas funções não-lineares das variáveis básicas:

$$\mathbf{x}_i = (1, x_{i1}, x_{i1}^2, x_{i2}, x_{i2}^2) \quad \text{ou} \quad \mathbf{x}_i = (1, \log(x_{i1}), \log(x_{i1}), x_{i1}x_{i2}, x_{i2}^2)$$

e

$$\theta = (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4)$$

O procedimento de estimação permanece o mesmo de antes. Basta usar a matriz com todas as features dispostas como colunas. Temos de admitir que não é simples encontrar bons argumentos para impor tal modelagem para η. Os exemplos acima são artificiais mas servem para mostrar a flexibilidade do modelo de regressão logística.

Assim, realmnete existe esta flexibilidade da regressão logística para incorporar transformações matemáticas $g(\mathbf{x})$ de um vetor de variáveis básicas como features. O problema prático é que precisamos saber, de antemão, quais são as transformações que queremos incorporar no modelo. Podemos pensar numa solução simples e bruta: simplesmente coloque como features todas as (muitas!) transformações que você pode remotamente desconfiar que podem ser úteis. Por exemplo, suponha que comecemos com $1+k$ variáveis independentes ("1" representa a coluna de 1's). Coloque todas as potências de 2^0 e 3^0 graus dessas variáveis, produzindo uma matriz com $1+k+k+k$ colunas. Em seguida, acrescente todos os produtos $x_i x_j$ de variáveis básicas distintas. Isto soma mais $k(k-1)/2$ colunas. Que tal colocar também o produto de uma variável com o quadrado de outra? São mais $k(k-1)$ colunas. Até aqui, sem usar colunas de features tais como $\log(x)$ ou $\sqrt{(x)}$, já temos $1+3k+3k(k-1)/2$. Se começarmos com $k=10$, estamos falando de uma matriz com 4051 colunas.

E daí, qual o problema? Como veremos no capítulo **??**, usar esta matriz final com as colunas construídas dessa forma, traz dois problemas. O primeiro é uma grande instabilidade numérica na estimação dos coeficientes. Pequenas alterações nos dados levam a grandes mudanças nos coeficientes estimados. O segundo é a necessidade de termos uma amostra muito maior que o número de colunas para que o procedimento numérico possa ser efetuado. Uma tentativa de amenizar estes dois problema é a técnica de regularização, tópico do capítulo **??**.

A regressão logística é comumente considerada um método *linear* pois ela combina linearmente as features especificadas. As features podem ser funções não lineares de outras variáveis básicas. Assim, o termo linear aplica-se aos *coeficientes* θ do modelo. O termo linear aparece porque usamos $\eta = \mathbf{x}'\theta$, uma combinação linear das features em $\mathbf{x}$.

17.8 Regressão logística como classificador

Imagine que temos apenas duas features, x_1 e x_2 e uma resposta binária y. Vamos usar a regressão logística com

$$\sigma = \mathbb{P}(Y=1 \mid \mathbf{x}) = \frac{1}{1+e^{\eta}} = \frac{1}{1+exp(-(\beta_0+\beta_1 x_1+\beta_2 x_2))}$$

Achamos estimativas dos coeficientes β_0, β_1 e β_2 pelo método or máxima verossimilhança.

Podemos usar este modelo estimado para classificar dados futuros. Olhando para (x_1, x_2), calculamos a sua probabilidade de sucesso. Se for maior que um ponto de corte (digamos, maior que $1/2$), classificamos o novo ponto como sucesso (classe 1). Se for menor que $1/2$, classificamos como fracasso (classe 0). Asim, todo ponto do plano pode ser alocado a uma das duas classes com base no modelo logístico estimado.

Considere os pontos (x_1, x_2) do plano tais que a probabilidade de sucesso seja extamente igual a $1/2$. Eles ficam na fronteira entre as duas classes possíveis. Quem são esses pontos do plano (x_1, x_2)? Eles formam uma reta. Para ver isto, considere a equação

$$\frac{1}{1+e^{-\eta}} = \frac{1}{2} \quad \Rightarrow \quad e^{-\eta} = 1 \quad \Rightarrow \quad \eta = 0$$

Assim, os pontos (x_1, x_2) desse conjunto satisfazem a equação

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2 = 0$$

ou, se $\beta_2 \neq 0$,

$$x_2 = -\left(\frac{\beta_0}{\beta_2}\right) - \left(\frac{\beta_1}{\beta_2}\right) x_1$$

Ou seja, satisfazem à equação de uma linha reta no plano (x_1, x_2). Esta reta determina o que chamamos de *fronteira de decisão*: de um lado da reta, a probabilidade de sucesso de sucesso é maior que $1/2$. Do outro lado, ela é menor que $1/2$.

E se houvéssemos escolhido outro ponto de corte diferente de $1/2$? Por exemplo, se tivéssemos decidido classificar como sucesso um novo ponto $\mathbf{x}$ tal que sua probabilidade de sucesso fosse maior que $p \in (01,)$ e como fracasso se fosse menor que p. Então, teríamos outra reta como fronteira de decisão. Esta nova reta seria simplesmente a reta anterior deslocada sem alterar sua inclinaç ao (fica como exercício).

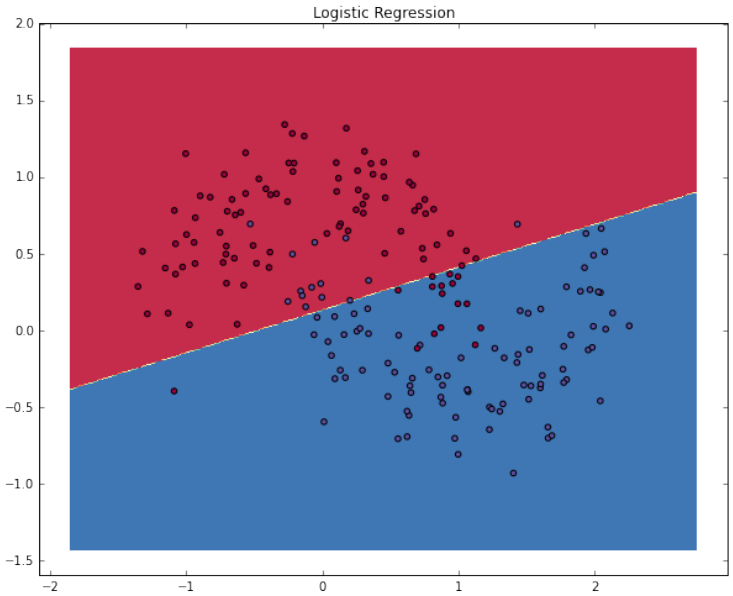


Figure 17.12: Decision Boundary for logistic regression with two features, x_1 and x_2.

Quando existirem mais features, digamos, $\mathbf{x} = (x_1, x_2, \ldots, x_5)$, a fronteira de decisão será um hiperplano no espaço $\mathbb{R}^5$ das features. Uma outra possibilidade é usarmos apenas duas variáveis básicas mas criarmos features derivadas delas. Neste caso, a fronteria de decisão ainda poserá ser representada no plano (x_1, x_2) das variáveis básicas mas já não será uma linha reta. Por exemplo, considere o gráfico do lado esquerdo da Figura 17.13. Nele vemos os pontos (x_1, x_2) de uma amostra rotulados como vermelho e azul de acordo com sua classe. É claro que uma fronteira de

decisão linear não é adequada neste problema. Em matemática, formas quadráticas em x_1 e x_2 permitem formar círculos e elipses, entre outras figuras geométricas. Usando então

$$\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2 + \beta_5 x_1 x_2 \,,$$

ajustamos uma regressão logística. A matriz de features era composta por 6 colunas: a colunas de 1's, a coluna com os valores de x_1 e a coluna com x_2, as colunas com os quadrados das coordenadas x_1 e x_2, e uma última coluna formada pelo produto das colunas x_1 e x_2. O resultado do ajuste gera um preditor linear. O conjunto de pontos do plano que satisfaz $\eta = 0$ é a linha verde do gráfico à direita.

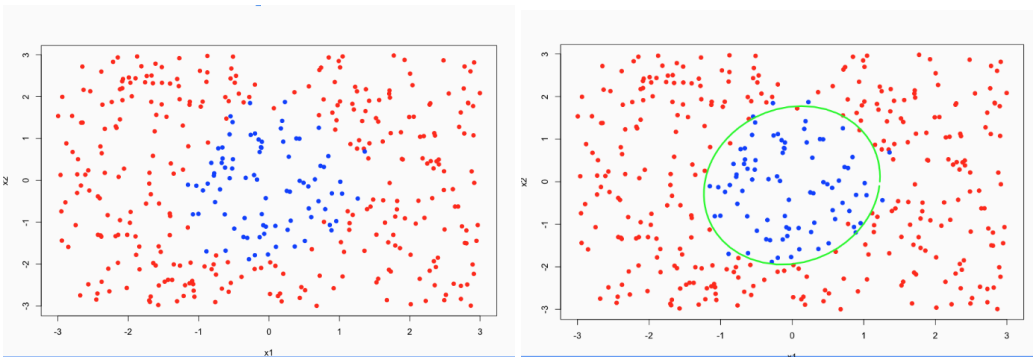


Figure 17.13: Nonlinear Decision Boundary for logistic regression with two features, x_1 and x_2.

(COLOCAR O CODIGO E A SAIDA DO R AQUI).

Dificuldade: como vimos na seção 17.7, é necessário saber de antemão que as features envolvendo os quadrados e produtos das varáveis básicas x_1 e x_2 precisam estar no modelo. Como saber isto? Obviamente, se o problema envolver apenas duas variáveis básicas, um gráfico como o da Figura 17.13 dá a dica necessária se você conhecer um pouco das equações de curvas no plano. Entratnto, a dificuldade é que em problema s práticos teremos um número grande de variáveis básicas e dificlmente o problema vai ser descoberto de tão fácil plotando apenas duas dessas variáveis básicas.

COLOCAR REF sobre feature engineering aqui.

18. Maximum Likelihood Estimator

18.1 Introdução

O método de máxima verossimilhança foi criado por Sir Ronald Fisher, o maior estatístico que já existiu (Figura 18.1). Ele foi uma espécie de Isaac Newton da estatística, responsável pelos principais conceitos e resultados da inferência estatística. Estes conceitos e resultados constituem a maior parte da estrutura da pesquisa científica em estatística e de suas aplicações até os dias de hoje. Como Newton, ele parece ter tido uma epifania que organizou o trabalho das décadas seguintes, completando 100 anos aproximadamente nos dias de hoje. A Figura 18.1 mostra como era *Sir* Ronald Fisher.

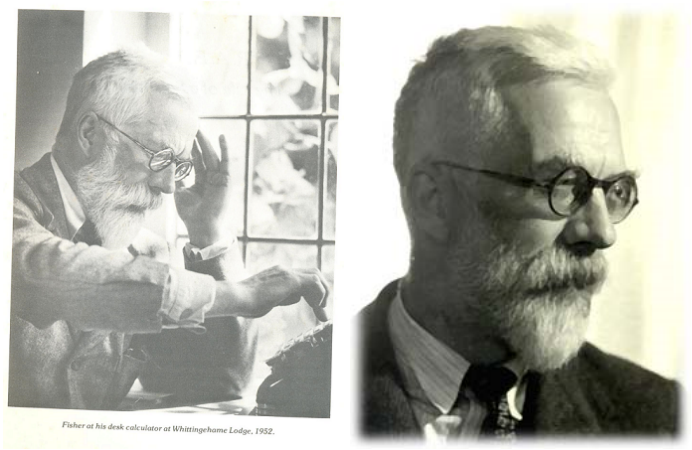


Figure 18.1: Sir Ronald A. Fisher.

Existem muitos *sites* sobre Fisher que, além de estatístico, foi também um dos maiores geneticistas que já existiu. Ao lado de Sewal Wright e Haldane, ele foi responsável por conseguir conciliar a teoria da evoluação de Darwin com a genética de Mendel. Um desses excelentes *sites* é `http://www.economics.soton.ac.uk/staff/aldrich/fisherguide/rafreader.htm` Suas idéias

principais em inferência foram publicadas em dois artigos publicado em 1922 e 1925, entitulados *On the mathematical foundations of theoretical statistics*[3] e *Theory of statistical estimation* [4] (Figura 18.2). Alguns dos principais conceitos (verossimilhança, suficiência e eficiência, por exemplo) e resultados que serão estudados no curso a partir desse capítulo apareceram neste último artigo espetacular, publicado quando ele tinha 35 anos de idade. Uma das idéias mais fundamentais de Fisher é a do estimador de máxima verossimilhança. Para explicar o que é este estimador, vamos apresentá-lo de maneira bastante informal no contexto de um exemplo.

IX. *On the Mathematical Foundations of Theoretical Statistics.*

By R. A. Fisher, M.A., *Fellow of Gonville and Caius College, Cambridge, Chief Statistician, Rothamsted Experimental Station, Harpenden.*

Communicated by Dr. E. J. Russell, F.R.S.

Received June 25,—Read November 17, 1921.

Contents.

Section	Page
1. The Neglect of Theoretical Statistics	310
2. The Purpose of Statistical Methods	311
3. The Problems of Statistics	313
4. Criteria of Estimation	316
5. Examples of the Use of Criterion of Consistency	317
6. Formal Solution of Problems of Estimation	323
7. Satisfaction of the Criterion of Sufficiency	330
8. The Efficiency of the Method of Moments in Fitting Curves of the Pearsonian Type III	332
9. Location and Scaling of Frequency Curves in general	336
10. The Efficiency of the Method of Moments in Fitting Pearsonian Curves	342
11. The Reason for the Efficiency of the Method of Moments in a Small Region surrounding the Normal Curve	355
12. Discontinuous Distributions	356
(1) The Poisson Series	359
(2) Grouped Normal Data	359
(3) Distribution of Observations in a Dilution Series	363
13. Summary	366

Definitions.

Centre of Location.—That abscissa of a frequency curve for which the sampling errors of optimum location are uncorrelated with those of optimum scaling. (6.)

Consistency.—A statistic satisfies the criterion of consistency, if, when it is calculated from the whole population, it is equal to the required parameter. (4.)

Distribution.—Problems of distribution are those in which it is required to calculate the distribution of one, or the simultaneous distribution of a number, of functions of quantities distributed in a known manner. (3.)

Efficiency.—The efficiency of a statistic is the ratio (usually expressed as a percentage) which its intrinsic accuracy bears to that of the most efficient statistic possible. It

Vol. ccxxii.—A 602. 2 x [Published April 19, 1922.

Theory of Statistical Estimation. By Mr R. A. Fisher, Gonville and Caius College.

[*Received* 17 March, *read* 4 May, 1925.]

PREFATORY NOTE.

It has been pointed out to me that some of the statistical ideas employed in the following investigation have never received a strictly logical definition and analysis. The idea of a frequency curve, for example, evidently implies an infinite hypothetical population distributed in a definite manner; but equally evidently the idea of an infinite hypothetical population requires a more precise logical specification than is contained in that phrase. The same may be said of the intimately connected idea of random sampling. These ideas have grown up in the minds of practical statisticians and lie at the basis especially of recent work; there can be no question of their pragmatic value. It was no part of my original intention to deal with the logical bases of these ideas, but some comments which Dr Burnside has kindly made have convinced me that it may be desirable to set out for criticism the manner in which I believe the logical foundations of these ideas may be established.

The idea of an infinite hypothetical population is, I believe, implicit in all statements involving mathematical probability. If, in a Mendelian experiment, we say that the probability is one half that a mouse born of a certain mating shall be white, we must conceive of our mouse as one of an infinite population of mice which might have been produced by that mating. The population must be infinite for in sampling from a finite population the fact of one mouse being white would affect the probability of others being white, and this is not the hypothesis which we wish to consider; moreover, the probability may not always be a rational number. Being infinite the population is clearly hypothetical, for not only must the actual number produced by any parents be finite, but we might wish to consider the possibility that the probability should depend on the age of the parents, or their nutritional conditions. We can, however, imagine an unlimited number of mice produced upon the conditions of our experiment, that is, by similar parents, of the same age, in the same environment. The proportion of white mice in this imaginary population appears to be the actual meaning to be assigned to our statement of probability. Briefly, the hypothetical population is the conceptual resultant of the conditions which we are studying. The probability, like other statistical parameters, is a numerical characteristic of that population.

We only need the conception of an infinite hypothetical population, in connection with random sampling. The ultimate logical elucidation of the one idea implies that of the other. Also, the word infinite is to be taken in its proper mathematical sense as denoting the limiting conditions approached by increasing a finite number indefinitely. I imagine that an exact meaning can be given to all the ideas required by some process such as the following.

Imagine a population of N individuals belonging to s classes, the number in class i being $p_i N$. This population can be arranged in order in $N!$ ways. Let it be so arranged and let us call the first n individuals in each arrangement a sample of n. Neglecting the order within the sample, these samples can be classified into the several possible types of sample according to the number of individuals of each class which appear. Let this be done, and denote the proportion of samples which belong to type j by q_j, the number of types being t. Consider the following proposition.

Given any series of proper fractions $P_1, P_2, \ldots P_s$, such that $S(P_i)=1$,

Figure 18.2: Os artigos fundamentais de Sir Ronald Fisher, de 1921 (esquerda) e de 1925 (direita).

18.1.1 Estimando o tempo médio de sobrevida

Glioblastoma multiforme IV é o mais agressivo tipo de câncer do cérebro. Este é um câncer devastador e o tempo de vida após o diagnóstico é curto. Suponha que, usando o tratamento cirúrgico e terapêutico padrão nestes casos, o tempo médio de sobrevida seja de 12 meses. Uma inovação médica parece promissora mas é muito mais cara. Como não existe a certeza de que o novo tratamento seja realmente melhor que o anterior, existem também restrições éticas quanto a sua adoção indiscriminada. Tanto as seguradoras de saúde quanto os pacientes e médicos envolvidos precisam tomar uma decisão mais bem informada sobre a adoção do novo tratamento em substituição ao antigo.

Suponha que $X_1, \ldots, X_n$ sejam os tempos de vida em meses de n indivíduos após o novo tratamento cirúrgico. Suponha também que elas sejam variáveis aleatórias i.i.d. com distribuição contínua. O interesse é em fazer inferência sobre o valor esperado de X_i. Este é chamado de tempo médio de sobrevida após o diagnóstico. Isto é, queremos fazer inferência sobre $E(X_i) = \mu$ Se μ for maior que 12 meses, o novo procedimento deveria ser considerado atentamente. Para decidir se μ é maior que 12, vamos estimar μ a partir dos dados da amostra usando a sua média aritmética: $\overline{X} = (X_1 + \ldots + X_n)/n$. Se a amostra é grande $\overline{X} \approx \mathbb{E}(X_i) = \mu$. Isto é garantido pelo teorema da Lei dos Grandes Números (LGN).

Qual a natureza de $\overline{X}$? É uma constante? É uma função matemática? É uma v.a.? $X_1, X_2, \ldots, X_n$ são v.a.'s iid com $\mathbb{E}(X_i) = \mu$ e $\mathbb{V}(X_i) = \sigma^2$. Portanto, sua combinação na estatística $\overline{X} = (X_1 + \ldots + X_n)/n$. é também uma v.a. Em cada amostra particular, ela fica instanciada num número específico. Alguns números são mais prováveis que outros. Qual é o valor esperado da v.a. $\overline{X}$? Temos $\mathbb{E}(\overline{X}) = \mu$, o valor esperado populacional. Assim, $\overline{X}$ oscila em torno do mesmo valor esperado que as observações individuais. Às vezes, um pouco acima de μ, à vezes, um pouco abaixo. Quanto de variação temos? Qual a variância da v.a. $\overline{X}$? Temos $\mathbb{V}(\overline{X}) = \sigma^2/n$, a variância

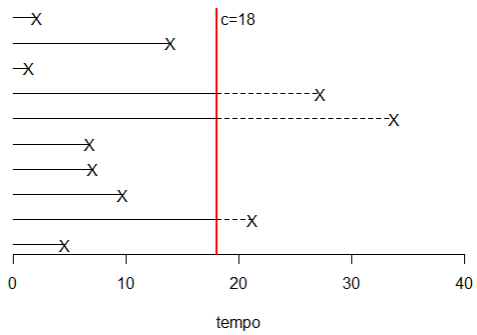


Figure 18.3: Uma amostra particular de $n = 10$ tempos de vida x_i. Três valores são maiores que 18 e portanto não são observados até o fim. Sabe-se apenas que, nestes casos, $x_i \geq 18$.

populacional reduzida pelo divisor n, o tamanho da amostra. Assim, $\overline{X}$ varia muito menos em torno de μ que as observações originais. Mais ainda, pela LGN, sabemos que $\overline{X} \to \mu$.

Assim, checar o valor de $\overline{X}$ dá uma boa base para uma tomada de decisão acerca do valor de μ, principalmente se a amostra é grande. Entretanto, para obter $\overline{X}$ é necessário esperar que todos os indivíduos da amostra faleçam para conhecermos todos os valores $X_1, \ldots, X_n$ e isto pode demorar um longo tempo. Se o novo tratamento não for melhor nem pior que o tratamento padrão, podemos ter, por exemplo, $\mathbb{P}(X_i > 36 \text{ meses}) = 0.10$. Numa amostra de 100 indivíduos, 3 anos após o início dos estudos teríamos aproximadamente 10 pacientes ainda vivos.

Nem sempre é possível esperar tanto tempo. Os diversos interessados na decisão (pacientes, familiares, seguradoras e médicos) precisam tomar decisões, mesmo que sujeitas a revisões posteriores. As decisões precisam ser bem informadas mas não podem esperar tanto tempo pela coleta dos dados. é sempre possível rever decisões errôneas mas, num dado momento, alguma decisão deve ser tomada. E de preferência, ela deve ser uma decisão baseada nas evidências disponíveis no momento.

Uma solução muito comum nos experimentos bioestatísticos é usar a *amostra censurada*. Observamos os pacientes até um tempo limite. Digamos, 18 meses. Para aqueles que viverem mais que o tempo limite, simplesmente anotamos que este evento ocorreu. Isto é, a amostra é composta das variáveis aleatórias $Y_1, \ldots, Y_n$ onde Y_i é igual ao tempo de vida X_i, se $X_i < 18$. Caso ocorra o evento $X_i \geq 18$, anota-se $Y_i = 18$ como um símbolo de que o evento $[X_i \geq 18]$ ocorreu. Em notação matemática, observamos $Y_i = g(X_i)$, uma função matemática de X_i dada por

$$Y_i = \min\{X_i, 18\} = \begin{cases} X_i, & \text{se } X_i < 18 \\ 18, & \text{se } X_i \geq 18 \end{cases}$$

Geralmente, em estudos médicos como estes, as amostras são pequenas. A doença é rara, o paciente precisa ser identificado pelos pesquisadores (por exemplo, ele estar no raio geográfico de busca do estudo), precisa aceitar participar do mesmo, etc. A Figura 18.3 mostra os tempos de vida X_i de uma amostra com apenas $n = 10$ indivíduos. Cada indivíduo é representado por uma linha horizontal que inicia-se no tempo 0 e que prolonga-se até o falecimento do indivíduo. O ponto de censura é de 18 meses. Três indivíduos têm seus tempos de vida censurados. Nestes três casos sabe-se apenas que $X_i \geq 18$ mas não se conhece seu valor exato. Para representar este fato, nós anotamos o valor de Y_i como sendo igual a 18. A Tabela 18.1 apresenta os valores dos tempos de sobrevida x_i de todos os 10 indivíduos, bem como os tempos censurados y_i.

i	1	2	3	4	5	6	7	8	9	10
x_i	4.6	21.2	9.7	7.1	6.8	33.8	27.2	1.4	14.0	2.1
y_i	4.6	18	9.7	7.1	6.8	18	18	1.4	14.0	2.1

Table 18.1: Dados de tempos de sobrevida x_i de uma amostra de 10 indivíduos e os dados censurados y_i que seriam realmente registrados. O tempo de censura é 18 meses.

Nosso problema continua sendo estimar o tempo esperado de sobrevida mas agora usando estes dados censurados. Nenhuma resposta vêm a mente de imediato. A opção anterior, quando os dados não eram censurados, era simplesmente tomar a média aritmética das variáveis aleatórias X_i. O que nós temos agora são as variáveis Y_i que nunca superam o tempo 18. Os maiores tempos de sobrevida (os três valores acima de 18 na Tabela 18.1) são substituídos pelo tempo de censura (18 meses, no exemplo). Portanto, a média aritmética dos tempos censurados y_i será menor que a média aritmética dos tempos não-censurados x_i. Ela tende a subestimar o verdadeiro tempo esperado de sobrevida e por isto ela não é um estimador razoável para μ.

Outra opção poderia ser então ignorar os tempos que foram de fato censurados e tomar a média aritmética apenas dos tempos em que se chegou a observar o falecimento do indivíduo. Esta também não é uma boa idéia. Ela daria um estimador até pior que a média de todos os Y_i pois teríamos apenas os tempos de sobrevida de quem faleceu muito rapidamente.

Fisher fez um raciocínio muito engenhoso que produz um candidato a estimador que parece razoável neste problema. Não só neste problema mas em quase qualquer outro modelo estatístico o mesmo raciocínio pode ser aplicado. Mais surpreendente ainda, este raciocínio gera estimadores imbatíveis num certo sentido. Nenhum estimador pode ser melhor do que aquele que vamos obter aplicando o método criado por Fisher. A bem da verdade, esta afirmação é um tanto exagerada e precisa ser melhor qualificada. Vamos explicar as condições exatas em que a afirmação é válida no Capítulo 19. Entretanto, por enquanto, podemos assumir que, de forma aproximada, os estimadores obtidos através do método de Fisher são os melhores possíveis na maioria das situações práticas e usuais da análise de dados e daí sua popularidade. Isto é mesmo notável: num certo sentido, nada pode ser melhor que o estimador baseado nas ideias de Fisher! O fato é que este raciocínio que vamos expor agora mudou completamente a história da estatística em 1925. Ele transformou o que era um conjunto de ideias e técnicas desconectadas numa ciência, a ciência dos dados.

18.1.2 Quais valores do parâmetro são verossímeis?

Para aplicar o método de Fisher, precisamos de um modelo de probabilidade para os dados observados. Vamos assumir que os tempos de vida são v.a.'s i.i.d. A independência é razoável: se um paciente morre mais cedo que o esperado devido ao seu câncer, isto não aumenta nem diminui as chances de outro paciente também falecer mais cedo. A hipótese de que sejam v.a.'s i.d. é menos razoável. Na prática, a distribuição do tempo de sobrevida X_i depende da idade, sexo, estágio do tumor no diagnóstico, entre outras variáveis. O comum é que um modelo de regressão GLM (capítulo 25) seja usado em problemas como este. Entretanto, para não complicar desnecessariamente nesse momento e para focar apenas no essencial, vamos assumir que os pacientes são idênticos com relação a todas estas características que poderiam afetar a distribuição de X_i. Isto é, vamos supor que eles tenham a mesma idade, mesmo sexo, mesmo estágio de tumor no momento do diagnósico, etc. Então as v.a.'s X_i são i.i.d.

Para explicar o método, assuma que $X_i \sim \exp(\lambda)$. O método da máxima verossimilhança requer a escolha de alguma classe de distribuições para os dados. Esta classe depende de parâmetros desconhecidos que serão estimados pelo método. Os médicos pesquisadores costumam usar a distribuição de Weibull para modelar o tempo de sobrevivência de pacientes em doenças graves como esta (ver, por exemplo, DasGupta et al., 2000). Entretanto, neste momento, vamos adotar

a distribuição exponencial pois ela é muito mais simples para os tempos de sobrevida mas uma aproximação pior que a Weibull. Vamos usá-la apenas para apresentar as principais ideias por trás do estimador de máxima verossimilhança. Na seção vamos voltar a este exemplo usando a distribuição de Weibull. A atratividade da distribuição exponencial para este problema é que ela possui único parâmetro λ e não um vetor multivariado θ. Veremos o méetodo de máxima verossimilhança de Fisher para este modelo particular mas ele funciona do mesmo modo em modelos muito mais sofisticados, como veremos no restante do capítulo.

Em resumo, assumindo que temos uma amostra $X_1, \ldots, X_n$ de v.a.'s i.i.d. com distribuição $\exp(\lambda)$, queremos estimar $\mu = \mathbb{E}(X_i)$. O valor de μ está associado com o parâmetro λ da distribuição exponencial pois $\mu = 1/\lambda$. Então, estimar μ é o mesmo que estimar $\lambda = 1/\mu$.

Considere os 10 dados $y_1, \ldots, y_{10}$ registrados na amostra censurada e exibidos no gráfico da Figura 18.3:

$$4.6, 18^c, 9.7, 7.1, 6.8, 18^c, 18^c, 1.4, 14.0, 2.1$$

Os três tempos de vida censurados são denotados por 18^c para diferenciá-los dos demais. Fisher percebeu que alguns valores de λ não eram compatíveis com os dados observados. Por exemplo, não parece plausível que λ seja igual ou próximo de 100 pois, neste caso, o tempo esperado de sobrevida seria apenas $E(X_i) = 1/100 = 0.01$, um centésimo de mês, menos de um dia. Este é um valor evidentemente muitíssimo menor que os dados observados, mesmo que estes estejam distorcidos pela censura. Do mesmo modo, os dados observados não dão suporte à afirmação de que $\lambda = 0.01$ pois, neste caso, a esperança de X_i seria $\mu = 1/0.01 = 100$ meses, um valor muito acima da maioria dos tempos de sobrevida observados na amostra.

Estes valores extremos para λ são facilmente descartados. Para tentar ser mais preciso sobre os valores de λ que parecem razoáveis tendo em vista a amostra em mãos, Fisher procurou pensar qual seria o princípio lógico que nós estamos usando para descartar esses valores extremos para λ.

Ele imaginou o seguinte: fixado algum valor para o parâmetro desconhecido λ, é possível calcular a chance de observar uma amostra tal como aquela realmente registrada. Por exemplo, o primeiro elemento da amostra foi igual a $y_1 = 4.6$. Supondo que λ é um valor fixo e arbitrário para o parâmetro desconhecido, vamos calcular a probabilidade de Y_1 ser um valor aproximadamente igual ao valor 4.6 realmente observado. Por que não calcular a probabilidade de obter *exatamente* 4.6? A razão é que, para uma v.a. com distribuição exponencial, esta probilidade é zero. De fato, se $X \sim \exp(\lambda)$, então $\mathbb{P}(X = 4.6) = 0$.

Vamos considerar um pequeno intervalo $(4.6 - \Delta/2,\ 4.6 + \Delta/2)$ de comprimento Δ e centrado em 4.6, e que denotaremos por $(4.6 \pm \Delta/2)$. Como 4.6 está longe da região de censura e Δ é pequeno, temos $Y_i = X_i$ e a probabilidade é igual a

$$\mathbb{P}(Y_1 \in (4.6 \pm \Delta/2)) = \mathbb{P}(X_1 \in (4.6 \pm \Delta/2)) \approx \lambda \exp(-4.6\lambda)\Delta,$$

Estamos aproximando a probabilidade pela área do retângulo de base Δ e altura igual a $f_\lambda(4.6)$, a densidade da distribuição exponencial no ponto 4.6. A Figura 18.4 mostra essa aproximação da àrea sob a densidade pela área do retângulo. De maneira análoga, calculamos a probabilidade para todos os outros elementos da amostra em que o valor registrado foi de fato o tempo de sobrevida do paciente.

Passemos agora aos três elementos da amostra que tiveram o valor registrado $y_i = 18^c$. Nós registramos $y_i = 18^c$ se, e somente se, o valor correspondente x_i tiver sido maior que 18, pois o tempo de sobrevida foi superior ao tempo de censura de 18 meses. Neste caso, a probabilidade de registrarmos $y_i = 18$ é dada por

$$\mathbb{P}(Y_i = 18^c) = \mathbb{P}(X_i \geq 18) = \int_{18}^{\infty} \lambda e^{\lambda x}\, dx = \exp(-18\lambda)$$

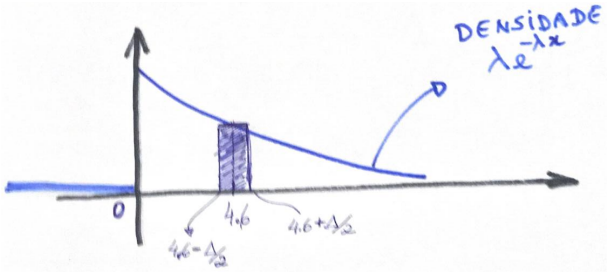


Figure 18.4: Aproximação da àrea sob a densidade $f(x) = \lambda e^{-\lambda x}$ no intervalo $(4.6 \pm \Delta/2)$ pela àrea do retângulo com altura $f(4.6)$.

Como as variáveis $Y_1, \ldots, Y_{10}$ são i.i.d., a probabilidade conjunta de extrairmos uma amostra tal qual a que realmente obtivemos é dada por

$$\mathbb{P}(Y_1 \in (4.6 \pm \Delta/2), Y_2 = 18^c, \ldots, Y_{10} \in (2.1 \pm \Delta/2))$$

que é igual ao produto das probabilidades marginais:

$$\mathbb{P}(Y_1 \in (4.6 \pm \Delta/2))\ \mathbb{P}(Y_2 = 18^c)\ \ldots\ \mathbb{P}(Y_{10} \in (2.1 \pm \Delta/2))$$

e esta por sua vez é igual a

$$\begin{aligned} \lambda \exp(-4.6\lambda)\Delta\ \exp(-18\lambda)\ \ldots\ \lambda \exp(-2.1\lambda)\Delta &= \lambda^7 \exp(-\lambda(4.6+18+\ldots+2.1))\Delta^7 \\ &= \lambda^7 \exp(-99.7\lambda)\Delta^7 \end{aligned}$$

Note que o expoente da exponencial multiplica λ pela soma 99.7 dos 10 valores registrados de y_i, somando tanto os 3 valores censurados quanto os 7 outros não censurados.

Considerando os dados da amostra como números fixos, a expressão acima é uma função de λ e Δ. O valor de Δ é completamente arbitrário e é escolhido pelo usuário sem relação com o verdadeiro valor do parâmetro ou com os valores dos dados na amostra. Já que sua escolha é subjetiva e arbitrária, vamos denotar a probabilidade anterior por $\mathbb{L}(\lambda)\Delta^7$ enfatizando que apenas o primeiro fator é uma função de λ:

$$\begin{aligned} \mathbb{P}(Y_1 \approx 4.6, Y_2 = 18^c, \ldots, Y_{10} \approx 2.1) &\approx \lambda^7 \exp(-\lambda(4.6+18+\ldots+2.1))\Delta^7 \\ &= \lambda^7 \exp(-99.7\lambda)\Delta^7 \\ &\equiv \mathbb{L}(\lambda)\Delta^7 \end{aligned}$$

Isto é,

$$\mathbb{L}(\lambda) = \lambda^7 \exp(-99.7\lambda)\,. \tag{18.1}$$

Note que se variarmos λ, o valor de Δ não se altera. Assim, com respeito a λ, o valor de Δ é uma constante.

Para diferentes valores de λ teremos valores diferentes da probabilidade aproximada $\mathbb{L}(\lambda)\Delta^7$ de obter uma amostra tal como a que realmente obtivemos. Isto fica claro na Figura 18.5. Nela, temos um gráfico de $\mathbb{L}(\lambda)$ versus λ para valores de λ entre 0 e 0.2. O que podemos ver é que, para valores tais como $\lambda > 0.15$, a probabilidade de obter a amostra é praticamente zero.

Observe também *não estamos* dizendo que a probabilidade de λ ser maior que 0.15 é praticamente zero. é uma diferença sutil. Estamos dizendo que: caso λ seja maior que 0.15, os dados que acabamos de receber como amostra constituem um fenônemo raríssimo. Escrevendo de maneira bem informal, estamos calculando

$$\mathbb{P}(\text{ obter os dados realmente observados } | \lambda = 0.15) \approx 0\,,$$

e não

$$\mathbb{P}(\lambda = 0.15 | \text{ os dados realmente observados são tais e tais}) = ??,$$

que não sabemos calcular.

A verossimilhança $\mathbb{L}(\lambda)$ não fornece a probabilidade de λ ser o valor verdadeiro do parâmetro. O parâmetro λ é desconhecido mas não é aleatório. Ele é fixo e desconhecido, pode ser um valor qualquer positivo mas não existem probabilidades associadas com esses diversos valores. O raciocínio correto é o seguinte: caso λ seja algum valor maior que 0.15, a amostra que realmente temos em mãos é um evento com probabilidade muito baixa de acontecer. Isto é, se λ é algum valor maior que 0.15 um evento muito raro (a amostra em mãos) acabou de ocorrer. O mesmo se dá com valores de $\lambda < 0.01$. Estes são também valores de λ para os quais a amostra que temos em mãos tem uma probabilidade muito baixa de ocorrência.

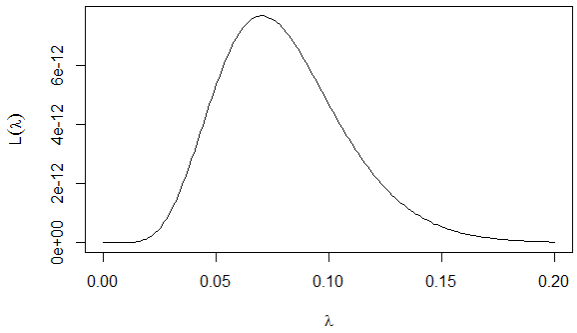


Figure 18.5: Gráfico da função $\mathbb{L}(\lambda) = \lambda^7 \exp(-99.7\lambda)$ versus λ.

Fisher dizia que estes valores tão extremos para λ não são *verossímeis*. A palavra verossímil é formada pelos radicais *vero* (de verdadeiro, real, autêntico) e *símil* (de semelhante, similar). De acordo com os dicionários, algo é verossímil se parece verdadeiro, se não repugna à verdade, se é semelhante à verdade, se é coerente o suficiente para se passar por verdade. Portanto, ao dizer que algo é verossímil, não dizemos que é verdadeiro mas que parece verdadeiro pois está de acordo com todas as evidências disponíveis. A notação $\ell(\theta)$ é devido à palavra *likelihood*, que significa verossimilhança em inglês.

18.1.3 Verossimilhança Relativa e máxima

Fica no ar a questão: devem existir outros valores de λ que, caso sejam os verdadeiros valores do parâmetro, a ocorrência dos dados observados não seja um evento tão improvável. Compare dois valores de λ quanto a sua verossimilhança, quanto a sua suposta veracidade, levando em conta os dados que foram observados. Por exemplo, vamos comparar os valores de $\lambda = 0.06$ e $\lambda = 0.15$ para ver se um deles é mais verossímil que o outro. Para isto, vamos calcular a razão entre $\mathbb{L}(0.06)$ e $\mathbb{L}(0.15)$:

$$\frac{\mathbb{L}(0.06)}{\mathbb{L}(0.15)} = \frac{0.06^7 \exp(-99.7 * 0.06)}{0.15^7 \exp(-99.7 * 0.15)} = 12.92$$

O que isto quer dizer? Supondo $\lambda = 0.06$, a probabilidade de obter uma amostra como a que realmente obtivemos é quase 13 vezes maior que a mesma probabilidade supondo que $\lambda = 0.15$. Neste sentido, o valor $\lambda = 0.06$ é mais verossímil que o valor $\lambda = 0.15$. Ambos podem ser considerados

como valores possíveis para λ mas os dados da amostra podem ocorrer com probabilidade muito maior quando $\lambda = 0.06$ do que quando $\lambda = 0.15$. Se temos que inferir sobre o verdadeiro valor de λ com base nesta amostra, porquê alguém iria preferir $\lambda = 0.15$ a $\lambda = 0.06$? É neste sentido que Fisher dizia que valores de λ maiores que 0.15 ou menores que 0.01 eram inverossímeis. A probabilidade de obter uma nova amostra similar a que temos em mãos e que foi realmente observada é muito maior quando $\lambda \in (0.05, 0.10)$ do que se λ estiver fora desse intervalo. A idéia então é acompanhar os valores $\mathbb{L}(\lambda)$ à medida que λ varre o espaço paramétrico. Quanto maior o valor de $\mathbb{L}(\lambda)$, mais verossímil o valor correspondente de λ. O valor de λ que leva ao valor máximo de $\mathbb{L}(\lambda)$ é chamado de estimativa de *máxima verossimilhança*.

Pela Figura 18.5, vemos que o valor de λ que vai maximizar $\mathbb{L}(\lambda)$ está por volta de 0.075. Podemos obter analiticamente este máximo. Basta derivar $\mathbb{L}(\lambda)$ com respeito a λ e igualar a zero:

$$\frac{d\mathbb{L}(\lambda)}{d\lambda} = \frac{d}{d\lambda}\left(\lambda^7 \exp(-99.7\lambda)\right) = 7\lambda^6 e^{-99.7\lambda} - 99.7\lambda^7 e^{-99.7\lambda} = 0$$

o que implica em

$$7 - 99.7\lambda = 0\,,$$

cuja solução é $\hat{\lambda} = 7/99.7$. Portanto, uma estimativa para o tempo médio de sobrevida é

$$\frac{1}{\widehat{\lambda}} = \frac{99.7}{7} = \frac{10}{7}\,\frac{99.7}{10} = \frac{10}{7}\,\overline{Y}\,.$$

Isto é, a estimativa de máxima verossimilhança do tempo médio pode ser pensada em dois passos: primeiro calcule a média aritmética $\overline{Y}$ de todos os valores da amostra, censurados e não-censurados, obtendo $99.7/10$. Esta estimativa tende a subestimar o valor verdadeiro e deveríamos aumentá-la. Este é o objetivo do fator $10/7 > 1$ que, multiplicando a média anterior, vai aumentar a estimativa mais para perto do valor verdadeiro.

No caso geral, usando k para representar o número de observações censuradas e $\sum_i Y_i$ para denotar a soma aleatória de todos os n valores registrados (k censurados e $n-k$ não censurados), teremos o seguinte estimador de máxima verossimilhança para o tempo médio de sobrevida:

$$\frac{1}{\widehat{\lambda}} = \frac{n}{n-k}\,\frac{\sum_i Y_i}{n} \tag{18.2}$$

Se não houver nenhuma observação censurada, o estimador é a media aritmética simples $\sum_i Y_i/n$ de todas as observações registradas. Se houver alguma observação censurada, a fração $n/(n-k)$ será maior que 1. Como neste caso $\sum_i Y_i/n$ tende a subestimar o verdadeiro tempo médio de sobrevida, o efeito da fração $n/(n-k)$ em (18.2) é dilatar a subestimativa, possivelmente trazendo-a mais para perto do verdadeiro valor que queremos estimar. A fração $n/(n-k)$ aumenta com o número de observações censuradas. Se todas as observações forem censuradas (isto é, se $k = n$), não existe o estimador de máxima verossimilhança.

18.1.4 Resumo informal da máxima verossimilhança

O método de máxima versossimilhança pode ser aplicado em praticamente toda situação de inferência em que os dados aleatórios sigam um modelo estatístico paramétrico, representado por uma família de distribuições de probabilidade para os dados $\mathbf{y}$ indexadas por um parâmetro θ. O método pode ser resumido de maneira informal da seguinte maneira:

- Suponha que $y_1, \ldots, y_n$ sejam os dados da amostra.

- Usando o modelo estatístico, calcule o valor aproximado da probabilidade de observar os dados da amostra e obtenha a *função de verossimilhança* $\mathbb{L}(\theta)$ onde apenas θ pode variar e as v.a.s estão com os seus valores fixados com os dados realmente observados na amostra.
- Obtenha o valor $\widehat{\theta}$ que maximiza $\mathbb{L}(\theta)$. Este valor é a estimativa de máxima verossimilhança.

Em resumo, o método de máxima verossimilhança encontra o valor $\widehat{\theta}$ de θ que é o mais verossímil tendo em vista os dados à mão. O valor $\widehat{\theta}$ é aquele em que, aproximadamente, a probabilidade de observar os dados realmente observados é máxima.

18.1.5 Por quê a máxima verossimilhança?

O método de máxima verossimilhança parece uma idéia engenhosa. Fisher chamou $\mathbb{L}(\theta)$ de função de verossimilhança e propôs que seu ponto de máximo $\hat{\theta}$ fosse usado como estimativa de θ. As razões para isto são as seguintes:

- generalidade: o método usado para obter a estimativa $\widehat{\theta}$ é muito geral e é útil quando a intuição não conseguir sugerir bons estimadores para θ. Ele pode ser aplicado sempre que houver um modelo estatístico paramétrico. Como veremos a seguir, a função de verossimilhança $\mathbb{L}(\theta)$ se reduz à função de densidade conjunta (se os dados forem contínuos) ou à função de probabilidade conjunta (se os dados forem discretos). Assim, é fácil obter $\mathbb{L}(\theta)$ e basta maximizá-la em θ.
- Fisher provou um resultado teórico: se a amostra cresce então a estimativa $\widehat{\theta}$ converge para o verdadeiro valor θ do parâmetro, caso o modelo probabilístico proposto seja o modelo gerador dos dados.
- outro resultado teórico de Fisher: se a amostra cresce então a estimativa $\widehat{\theta}$ é aproximadamente não-viciada para θ. Veremos a definição formal de vício de um estimador no capítulo 19 mas, informalmente, $\hat{\theta}$ não estará sistematicamente subestimando ou superestimando θ.
- mais um resultado teórico de Fisher, e este é sensacional: qualquer estimador não-viciado ou aproximadamente não-viciado terá um erro de estimação médio maior que o estimador de máxima verossimilhança. Assim, de certa forma, ele terá o menor erro médio de estimação possível dentre aqueles que não subestimam ou sobreestimam θ sistematicamente. E isto é válido em praticamente *qualquer* modelo estatístico.
- Finalmente, o estimador de máxima verossimilhança é uma variável aleatória que possui distribuição aproximadamente gaussiana, não importa quão complicada seja a sua fórmula. Este é um fato fundamental para podermos construir intervalos de confiança (ver capítulo 19) e para realizar teste de hipóteses (ver capítulo **??**).

Os resultados teóricos foram apresentados acima de maneira informal, não rigorosa. Vamos definir precisamente os conceitos envolvidos (ser não-viciado, erro médio de estimação, etc) e veremos as condições em que os resultados são válidos no capítulo 19. No entanto, de forma intuitiva e aproximada, o que importa no momento é que vale a pena considerar com carinho e afeto os estimadores de máxima verossimilhança. Numa quantidade enorme de aplicações importantes nada pode ser (muito) melhor que eles, num certo sentido. Podemos até ter estimadores tão bons quanto o estimador de máxima verossmilhança, mas não muito melhores. Isto não será válido sempre mas é válido tão frequentemente que tornou o método de máxima verossimilhança imensamente popular. Mesmo quando ele não for o melhor, ele pode ser um excelente ponto de partida a partir do qual melhorias podem ser acrescentadas.

18.2 Modelos paramétricos

Definition 18.2.1 — Modelo paramétrico. Seja $\mathbf{Y} = (Y_1, \ldots, Y_n)$ um vetor aleatório. Um modelo estatístico paramétrico para este vetor é uma família $\mathscr{P}_\theta = \{f(\mathbf{y}; \theta)\}$ de distribuições de probabilidade conjunta para o vetor $\mathbf{Y}$. As distribuições na família são indexadas por um

vetor de dimensão finita $\theta = (\theta_1, \theta_2, \ldots, \theta_k)$ pertencente a um conjunto $\Theta \subset \mathbb{R}^k$, chamado de espaço paramétrico.

■ **Example 18.1 — Modelo Poisson i.i.d..** Seja $\mathbf{Y} = (Y_1, \ldots, Y_n)$ um vetor aleatório composto de v.a.'s i.i.d. com distribuição Poisson(λ). O parâmetro indexador da família de distribuições é unidimensional: $\theta = \lambda > 0$. O espaço paramétrico é $\Lambda = (0, \infty)$. Assim, o modelo é a família $\mathscr{P}_\lambda = \{f(\mathbf{y}; \lambda)\}$ com distribuição

$$f(\mathbf{y}; \lambda) = \mathbb{P}(Y_1 = y_1, \ldots, Y_n = y_n \mid \lambda) = \prod_{i=1}^{n} \frac{\lambda^{y_i}}{y_i!} e^{-\lambda} = \frac{\lambda^{y_1 + \ldots + y_n}}{y_1! \ldots y_n!} e^{-\lambda n}$$

indexada por $\lambda > 0$. ■

■ **Example 18.2 — Modelo Gaussiano i.i.d..** Seja $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)$ um vetor aleatório composto de v.a.'s i.i.d. com distribuição $N(\mu, \sigma^2)$. Vamos definir $\theta = (\mu, \sigma^2)$, um vetor bi-dimensional com o espaço paramétrico $\Theta = \mathbb{R} \times (0, \infty)$. Então $\mathscr{P}_\theta = \{f(\mathbf{y}; \theta)\}$ com

$$\begin{aligned} f(\mathbf{y}; \theta) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{(y_i - \mu)^2}{2\sigma^2} \right\} \\ &= \left[\frac{1}{\sqrt{2\pi\sigma^2}} \right]^n \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mu)^2 \right\} \\ &= \left(2\pi\sigma^2\right)^{-n/2} \exp\left\{ -\frac{1}{2} \sum_{i=1}^{n} \left(\frac{y_i - \mu}{\sigma} \right)^2 \right\} \end{aligned}$$

■

■ **Example 18.3 — Modelo de regressão linear.** Seja $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)$ um vetor aleatório composto de v.a.'s independentes mas não identicamente distribuídas. Temos $Y_i \sim N(\mu_i, \sigma^2)$. O valor esperado μ_i depende de atributos da i-ésima observação: é uma soma ponderada desses atributos. Mais especificamente, com os atributos reunidos num vetor $\mathbf{x}_i = (1, x_{i1}, \ldots, x_{ip})$ de dimensão $p+1$, temos $\mu_i = \mathbf{x}_i'\beta = \beta_0 + \beta_1 x_{i1} + \ldots \beta_p x_{ip}$. Vamos definir

$$\theta = (\beta, \sigma^2) = (\beta_0, \beta_1, \ldots, \beta_p, \sigma^2),$$

um vetor $(p+2)$-dimensional, com o espaço paramétrico $\Theta = \mathbb{R}^{p+1} \times (0, \infty)$. Assim, o modelo $\mathscr{P}_\theta = \{f(\mathbf{y}; \theta)\}$ é definido por

$$\begin{aligned} f(\mathbf{y}; \theta) &= f(\mathbf{y} | \mu, \sigma^2) = \prod_{i=1}^{n} N(y_i; \mathbf{x}_i'\beta, \sigma^2) \\ &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2} \left(\frac{y_i - (\beta_0 + \beta_1 x_{i1} + \ldots \beta_p x_{ip})}{\sigma} \right)^2 \right) \\ &= \left(2\pi\sigma^2\right)^{-n/2} \exp\left(-\frac{1}{2} \sum_{i=1}^{n} \left(\frac{y_i - (\beta_0 + \beta_1 x_{i1} + \ldots \beta_p x_{ip})}{\sigma} \right)^2 \right) \end{aligned}$$

■

■ **Example 18.4 — Modelo de regressão logística.** Seja $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)$ um vetor aleatório composto de v.a.'s independentes mas não identicamente distribuídas. As v.a.'s Y_i são ensaios de Bernoulii, binárias: $Y_i \sim$ Bernoulli(p_i) onde $p_i = \mathbb{P}(Y_i = 1)$. A probabilidade de sucesso p_i *não* é a mesma para todas as observações. Algumas têm mais chance de ser sucesso do que outras. A probabilidade p_i varia de observação para observação em função de atributos, regressores ou variáveis independentes, medidos em cada exemplo: $\mathbf{x}_i = (1, x_{i1}, \ldots, x_{ip})$, um vetor de dimensão $p+1$

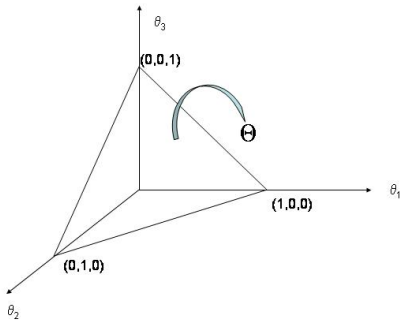


Figure 18.6: Espaço paramétrico Θ de um vetor $\theta = (\theta_1, \theta_2, \theta_3)$ de uma multinomial com três categorias. Θ é representado pelo triângulo de vértices $(1,0,0)$, $(0,1,0)$ e $(0,0,1)$.

No modelo de regressão logística, assumimos uma forma funcional específica para modelar esta dependência de p_i em função dos atributos: a função logística. Pegue um preditor linear do sucesso da i-ésima observacção: $\eta_i = \eta(\mathbf{x}_i) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}$ Transforme este preditor com a função logística para cair no intervalo $(0,1)$, que é a faixa de variação de probabilidades.

$$p_i = \frac{1}{1+e^{-\eta_i}} = \frac{1}{1+e^{-(\beta_0+\beta_1 x_{i1}+\ldots+\beta_p x_{ip})}} = \frac{1}{1+e^{-\mathbf{x}_i^t\theta}}$$

onde $\theta = (\beta_0, \beta_1, \ldots, \beta_p)$ é o vetor $(p+1)$-dimensional de pesos ou parâmetros desconhecidos, com o espaço paramétrico $\Theta = \mathbb{R}^{p+1}$. Assim, as v.a.'s binárias $Y_1, Y_2, \ldots, Y_n$ são independentes e, para cada exemplo, temos o modelo $Y_i \sim$ Bernoulli(p_i). A família $\mathscr{P}_\theta = \{f(\mathbf{y};\theta)\}$ é especificada por

$$\begin{aligned} f(\mathbf{y};\theta) &= \prod_{i=1}^{n} \left(\mathbb{P}(Y_i=1)^{y_i}\, \mathbb{P}(Y_i=0)^{1-y_i}\right) \\ &= \prod_{i=1}^{n} \left(\left(\frac{1}{1+e^{-\mathbf{x}_i^t\theta}}\right)^{y_i} \left(1-\frac{1}{1+e^{-\mathbf{x}_i^t\theta}}\right)^{1-y_i} \right) \end{aligned}$$

Neste exemplo, o vetor θ que indexa as distribuições coincide com o vetor de coeficientes β. ■

■ **Example 18.5 — Modelo multinomial.** Y_1, Y_2, e Y_3 são as contagens do número de pessoas que são católicos (Y_1), protestantes (Y_2), outras religiões ou sem religião (Y_3). Estas contagens foram obtidas a partir de uma amostra de tamanho n. O modelo natural para estes dados é a distribuição multinomial. Assim, $\mathscr{P}_\theta = \{f(\mathbf{y};\theta)\}$ onde $f(\mathbf{y};\theta)$ pertence à família multinomial $\mathscr{M}(\theta;n)$. O espaço paramétrico é

$$\Theta = \left\{\theta = (\theta_1,\theta_2,\theta_3) \in [0,1]^3 \text{ tal que } \theta_1+\theta_2+\theta_3 = 1\right\}$$

A coordenada θ_j é proporção de indivíduos na *população* que pertence á categoria j. Uma representação de Θ está na Figura 18.6.

■

18.3 Estimativa de máxima verossmilhança: MLE

Suponha que o vetor $\mathbf{Y} = (Y_1, \ldots, Y_n)$ seja composto por variáveis aleatórias. Vamos usar a mesma notação tanto para a função de probabilidade conjunta $\mathbb{P}(\mathbf{Y} = \mathbf{y}) = f(\mathbf{y})$ de v.a.'s discretas

quanto para a densidade conjunta $f(\mathbf{y})$ de v.a.'s contínuas. Para evitar ficar distinguindo as duas quando o procedimento for o mesmo para os dois casos, discreto e contínuo, vamos chamá-la simplesmente de densidade conjunta. Como as distribuições que nos interessam no caso do método de máxima verossimilhança pertencem a famílias paramétricas indexada por $\theta \in \Theta$, vamos denotar isto explicitamente escrevendo a densidade conjunta como $f(\mathbf{y};\theta)$.

Definition 18.3.1 — Função de Verossimilhança. Considere a densidade conjunta $f(\mathbf{y},\theta)$ do vetor $\mathbf{Y} = (Y_1,\ldots,Y_n)$ como uma função de θ para $\mathbf{y}$ fixo. Nós chamamos esta função de *função de verossimilhança* do parâmetro θ e vamos denotá-la por $\mathbb{L}(\theta)$.

Se o vetor aleatório $\mathbf{Y}$ é composto por variáveis aleatórias discretas então, a função $\mathbb{L}(\theta) = f(\mathbf{y},\theta)$ é a probabilidade de observar $\mathbf{Y}$ igual ao valor $\mathbf{y}$ *realmente* observado na amostra quando θ é o verdadeiro valor do parâmetro. O valor de $\mathbb{L}(\theta)$ é uma medida de quão verossímil é o valor de θ, verossímil no sentido de ser capaz de gerar produzir o vetor observado $\mathbf{x}$ observado. Se θ_1 e θ_2 são dois valores distintos de θ e se $\mathbb{L}(\theta_1)$ é k vezes maior que $\mathbb{L}(\theta_2)$ então dizemos que θ_1 é k vezes mais verossímil que θ_2.

No caso de v.a.'s conínuas, por causa dos problemas com conjuntos não-enumeráveis, qualquer instanciação $\mathbf{y}$ tem probabilidade zero. Por exemplo, $\mathbb{P}(X_1 = 1.24, X_2 = -2.32) = 0$ se X_1 e X_2 são gaussianas. Probabilidades são áreas sob as curvas densidades e a área associada com um ponto exato é zero. Entretanto, podemos usar intuitivamente o mesmo conceito que no caso discreto. A probabilidade do vetor aleatório $\mathbf{Y}$ ser aproximadamente o vetor $\mathbf{y}$ realmente observadoé

$$\mathbb{P}(\mathbf{Y} = \mathbf{y} \pm \Delta) = \mathbb{P}\left(Y_1 \in (y_1 - \Delta, y_1 + \Delta), \ldots, Y_n \in (y_n - \Delta, y_n + \Delta)\right) = f(\mathbf{y},\theta)(2\Delta)^n = \mathbb{L}(\theta)(2\Delta)^n$$

O fator $(2\Delta)^n$ não envolve o parâmetro θ e portanto pode ser ignorado para maximizar a verossimilhança.

Definition 18.3.2 — Função de Log-Verossimilhança. Chamamos $\ell(\theta) = \log(f(\mathbf{y},\theta))$ de *função de log-verossimilhança* do parâmetro θ.

■ **Example 18.6 — Modelo Poisson i.i.d..** Neste modelo, temos $\theta = \lambda$ e a função de verossimilhança

$$\mathbb{L}(\lambda) = \prod_{i=1}^{n} \frac{\lambda^{y_i}}{y_i!} e^{-\lambda} = \frac{\lambda^{\sum_i y_i}}{\prod_i y_i!} e^{-n\lambda}$$

e a função de log-verossmilhança

$$\ell(\lambda) = (\sum_i y_i)\ \log(\lambda) - n\lambda - \sum_i \log(y_i!)$$

Por exemplo, se $n = 3$ e $y_1 = 1$, $y_2 = 0$ e $y_3 = 1$ então

$$\mathbb{L}(\lambda) = \frac{\lambda^{1+0+1}}{1!0!1!}\, e^{-3\lambda} \quad \text{e} \quad \ell(\lambda) = 2\log(\lambda) - 3\lambda - (\log(1!) + \log(0!) + \log(1!))$$

■

■ **Example 18.7 — Modelo Gaussiano i.i.d..** Neste modelo, temos a função de verossimilhança

$$\begin{aligned}
\mathbb{L}(\theta) &= \mathbb{L}(\mu,\sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\,\sigma^2}} \exp\left(-\frac{1}{2}\left(\frac{y_i - \mu}{\sigma}\right)^2\right) && (18.3)\\
&= \left(\sqrt{2\pi\,\sigma^2}\right)^{-n} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{y_i - \mu}{\sigma}\right)^2\right) && (18.4)\\
&= \left(2\pi\,\sigma^2\right)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2\right) && (18.5)\\
& && (18.6)
\end{aligned}$$

e a função de log-verossmilhança

$$\ell(\theta) = \ell(\mu, \sigma^2) = -\frac{n}{2}\log\left(2\pi\ \sigma^2\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2$$

Por exemplo, se $n = 4$ e $y_1 = 1.11$, $y_2 = 0.32$ e $y_3 = -1.77$ e $y_4 = 0.91$, com $\sum_i y_i = 1.11 + \ldots + 0.91 = 0.57$, então

$$\mathbb{L}(\theta) = \mathbb{L}(\mu, \sigma^2) = \left(\sqrt{2\pi\ \sigma^2}\right)^{-2} \exp\left(-\frac{1}{2\sigma^2}\left((1.11-\mu)^2 + (0.32-\mu)^2 + (-1.77-\mu)^2 + (0.91-\mu)^2\right)\right)$$

e

$$\ell(\theta) = \ell(\mu, \sigma^2) = -2\log\left(2\pi\ \sigma^2\right) - \frac{1}{2\sigma^2}\left((1.11-\mu)^2 + (0.32-\mu)^2 + (-1.77-\mu)^2 + (0.91-\mu)^2\right)$$

■

O vetor $\mathbf{y}$ aparece na expressão da verossimilhança $\mathbb{L}(\theta)$ mas ele é considerado fixo, como nos exemplos acima. Lembre-se que $\mathbf{y}$ significa o conjunto de valores realmente obtidos em um experimento, os valores realizados do vetor aleatório $\mathbf{Y}$.

■ **Example 18.8 — Modelo de regressão linear.** Completar ■

■ **Example 18.9 — Gaussiana bivariada e multivariada.** Completar ■

■ **Example 18.10 — Modelo de regressão logística.** Completar ■

Definition 18.3.3 — Estimativa de máxima verossmilhança. O método de estimação de máxima verossimilhança consiste em encontrar o valor $\widehat{\theta}$ dentro do espaço paramétrico Θ que é o mais verossímil em termos de gerar os dados $\mathbf{y}$ da amostra. Isto é, se observamos $\mathbf{Y} = \mathbf{y}$, nós procuramos $\widehat{\theta}$ que satisfaça

$$\mathbb{L}(\widehat{\theta}) = f(\mathbf{y}, \widehat{\theta}) = \max_{\theta\in\Theta} \mathbb{L}(\theta) = \max_{\theta\in\Theta} f(\mathbf{y}, \theta)$$

■ **Example 18.11 — Modelo Poisson i.i.d..** Suponha que $\mathbf{y} = (y_1, y_2, y_3) = (1, 0, 1)$ no modelo Poisson i.i.d. com $n = 3$. Então pode ser verificado (ver a próxima seção) que tomando $\widehat{\lambda} = (y_1 + y_2 + y_3)/3 = 2/3 \approx 0.667$ maximizamos $\mathbb{L}(\lambda)$:

$$\mathbb{L}(2/3) = \frac{(2/3)^{1+0+1}}{1!0!1!}\ e^{-32/3} \geq \frac{\lambda^{1+0+1}}{1!0!1!}\ e^{-3\lambda} = \mathbb{L}(\lambda) \quad \forall\ \lambda > 0$$

No caso geral de uma amostra com n observações i.i.d Poisson(λ), a estimativa de máxima verossimilhança é $\widehat{\lambda} = (y_1 + y_2 + \ldots + y_n)/n$. ■

■ **Example 18.12 — Modelo Gaussiano i.i.d..** Neste modelo, temos $\theta = (\mu, \sigma^2)$. Se tivermos apenas $n = 4$ observações com $y_1 = 1.11$, $y_2 = 0.32$ e $y_3 = -1.77$ e $y_4 = 0.91$, então pode ser verificado (ver a próxima seção) que tomando $\hat{\mu}$ como a média aritmética,

$$\hat{\mu} = \bar{y} = \sum_i y_i/4 = (1.11 + \ldots + 0.91)/4 = 0.1425$$

e $\widehat{\sigma^2}$ igual à variância amostral

$$\widehat{\sigma^2} = S^2 = \frac{1}{4}\sum_i (y_i - \bar{y})^2 = \frac{1}{4}\sum_i (y_i - 0.1425)^2/4 \approx 1.3036$$

maximizamos $\mathbb{L}(\theta)$:

$$\begin{aligned}\mathbb{L}(\theta) &= \mathbb{L}(\mu,\sigma^2)\\ &\leq \mathbb{L}(\bar{y},S^2) = \mathbb{L}(\widehat{\mu},\widehat{\sigma^2})\\ &= \frac{1}{(2\pi\ 1.3036)^{-4/2}}\ \exp\left(-\frac{(1.11-0.1425)^2+(0.32-0.1425)^2+(-1.77-0.1425)^2+(0.91-0.1425)^2}{2*1.3036}\right.\end{aligned}$$

■

A estimativa de máxima verossimilhança $\widehat{\theta}$ depende dos dados $\mathbf{y}$ da amostra. No caso Poisson, por exemplo, $\widehat{\lambda} = \bar{y} = (y_1 + \ldots + y_n)/n$. No caso gaussiano, $\widehat{\theta} = (\widehat{\mu},\widehat{\sigma^2}) = (\bar{y},S^2)$ onde S^2 é função dos dados $\mathbf{y}$. Assim, para enfatizar este fato, vamos escrever o estimador de máxima verossimilhança como $\widehat{\theta}(\mathbf{y})$.

18.4 Obtendo o MLE

A principal razão para definirmos a função de log-verossimilhança $\ell(\theta)$ é facilitar a obtenção do MLE $\widehat{\theta}$. Na maioria dos problemas, a densidade conjunta Por exemplo, se a amostra é composta de v.a.'a independentes, a densidade conjunta será o produtos das densidades marginais. MEsmo quando a amostra é composta de v.a.'s não independentes, usualmnete a densidade conjunta será expressa como o produto de densidades condicionais. O fato é que, quase sempre, a verossimilhança é o produto de várias funções de θ. A função log transforma produtos em somas e maximizar uma soma de funções é mais simples que maximizar um produto de funções. Como o log é uma função crescente, tomar o log muda a função mas não seu ponto de máximo. Se θ_1 e θ_2 são dois pontos do espaço paramétrico tais que $\mathbb{L}(\theta_1) > \mathbb{L}(\theta_2) > 0$ então $\log(\mathbb{L}(\theta_1)) > \log(\mathbb{L}(\theta_2))$ pois a função log é crescente (se $a > b > 0$ então $\log(a) > \log(b)$). Isto significa que, se existir um ponto de máximo $\widehat{\theta}$ tal que $\mathbb{L}(\widehat{\theta}) \geq \mathbb{L}(\theta)$ para todo $\theta \in \Theta$ então

$$\ell(\widehat{\theta}) = \log\left(\mathbb{L}(\widehat{\theta})\right) \geq \log\left(\mathbb{L}(\theta)\right) = \ell(\theta)$$

Assim, procurar o ponto de máximo de $\mathbb{L}(\theta)$ pode ser reduzida a procurar o ponto de máximo de $\ell(\theta) = \log(\mathbb{L}(\theta)$. A Figura 18.7 ilustra a situação num caso em que θ é uni-dimensional.

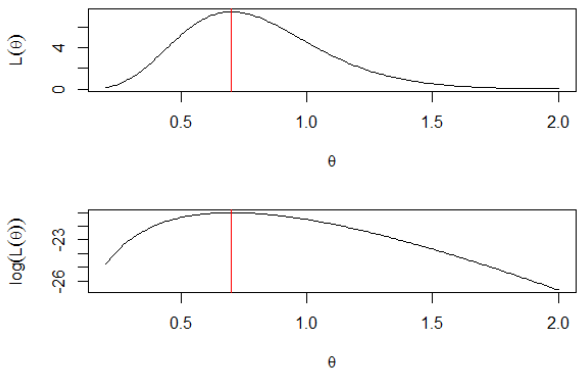


Figure 18.7: Gráfico da função $\mathbb{L}(\theta)$ e $\ell(\theta) = \log(\mathbb{L}(\theta))$. O ponto de máximo $\widehat{\theta}$ é o mesmo nas duas funções embora os valores de $\mathbb{L}(\theta)$ e $\ell(\theta) = \log(\mathbb{L}(\theta))$ sejam completamente diferentes.

Então problema é: como encontrar o ponto de máximo $\widehat{\theta}$ de $\ell(\theta)$? Pontos de máximo e mínimo de funções bem comportadas (com primeira e segunda derivadas contínuas, por exemplo) podem ser

usualmente encontrados buscando os pontos críticos da função. Seja $\theta = (\theta_1, \ldots, \theta_k)$ o parâmetro. Queremos $\theta \in \Theta$ que maximize a log-verossimilhança $\ell(\theta)$:

$$\widehat{\theta} = \left(\widehat{\theta}_1, \widehat{\theta}_2, \ldots, \widehat{\theta}_k\right) = \arg\max_{\theta} \ell(\theta)$$

Temos uma amostra $\mathbf{Y} = (Y_1, \ldots, Y_n)$ composta de variáveis aleatórias discretas ou contínuas com log-verossimilhança $\ell(\theta) = \log f(\mathbf{y}|\theta)$. Usualmente, o MLE $\widehat{\theta}$ é obtido resolvendo-se *simultaneamente* as k equações:

$$\begin{cases} \frac{\partial \ell(\theta)}{\partial \theta_1} = 0 \\ \frac{\partial \ell(\theta)}{\partial \theta_2} = 0 \\ \cdots \\ \frac{\partial \ell(\theta)}{\partial \theta_k} = 0 \end{cases}$$

Representamos o vetor gradiente como um vetor-coluna:

$$\nabla \ell(\theta) = \frac{\partial \ell(\theta)}{\partial \theta} = \left(\frac{\partial \ell(\theta)}{\partial \theta_1}, \frac{\partial \ell(\theta)}{\partial \theta_2}, \ldots, \frac{\partial \ell(\theta)}{\partial \theta_k}\right)^t$$

Assim, o sistema de equações pode ser escrito de forma vetorial:

$$\nabla \ell(\theta) = \frac{\partial \ell(\theta)}{\partial \theta} = (0, 0, \ldots, 0)^t \tag{18.7}$$

Geralmente, este será um sistema de quações não-lineares e vai requerer solução numérica (ver seção ??). Uma solução desse sistema é um *ponto crítico*.

Definition 18.4.1 — Equação de verossmilhança. O sistema de equações $\nabla \ell(\theta) = \mathbf{0}$ é chamado de *equação de verossmilhança*.

Quando um ponto crítico é um ponto de máximo de $\ell(\theta)$? Para responder a isto, nós olhamos para a matriz de derivadas parciais de segunda ordem de $\ell(\theta)$ avaliada no ponto $\widehat{\theta}$:

$$D^2\ell(\widehat{\theta}) = D^2\ell(\theta)\,|_{\theta=\widehat{\theta}} = \begin{bmatrix} \frac{\partial^2 \ell(\theta)}{\partial \theta_1^2} & \frac{\partial^2 \ell(\theta)}{\partial \theta_1 \partial \theta_2} & \cdots & \frac{\partial^2 \ell(\theta)}{\partial \theta_1 \partial \theta_k} \\ \frac{\partial^2 \ell(\theta)}{\partial \theta_2 \partial \theta_1} & \frac{\partial^2 \ell(\theta)}{\partial \theta_2^2} & \cdots & \frac{\partial^2 \ell(\theta)}{\partial \theta_2 \partial \theta_k} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial^2 \ell(\theta)}{\partial \theta_k \partial \theta_1} & \frac{\partial^2 \ell(\theta)}{\partial \theta_k \partial \theta_2} & \cdots & \frac{\partial^2 \ell(\theta)}{\partial \theta_k^2} \end{bmatrix}\Bigg|_{\theta=\widehat{\theta}}$$

A matriz de derivadas parciais de segunda ordem é chamada de *matriz Hessiana*. Avaliando a matriz Hessiana no ponto $\widehat{\theta}$, verificamos se ela é definida negativa. Isto é, se

$$\theta^t \; D^2\ell(\widehat{\theta}) \; \theta < 0$$

para todo vetor $\theta \neq \mathbf{0}$ então $D^2\ell(\widehat{\theta})$ é definida negativa e $\widehat{\theta}$ é um ponto de máximo da log-verossimilhança $\ell(\theta)$.

Em alguns exemplos simples, a equação de verossmilhança (18.7) pode ser resolvida algebricamente resultando numa fórmula para $\widehat{\theta}(\mathbf{y})$. Veremos alguns desses exemplos na seção 18.5. Entretanto, na maioria das situações práticas e mais realistas, é necessário resolver a equação de verossimilhança numericamente usando algum algoritmo de busca de raizes ou de maximização. Veremos alguns exemplos de maximização numérica, incluindo a regressão logística, na seção 18.6.

Existem situações nas quais $\widehat{\theta}$ não pode ser obtido através da equação de verossimilhança (18.7). Nestes casos, outros métodos que não dependem do gradiente $\nabla\ell(\theta)$ são necessários. Se o máximo $\widehat{\theta}$ ocorre na fronteira do espaço paramétrico ou quando um dos parâmetros é restrito a um conjunto discreto de valores, o MLE não pode ser obtido via (18.7). Por exemplo, o máximo global da função $\ell(\theta)$ pode ocorrer na fronteira de Θ e este máximo pode não ser raiz de $\ell'(\theta) = 0$ (ver exemplo ??). Similarmente, se Θ é restrito a um conjunto discreto de valores tais como os inteiros então a equação de verossimilhança não se aplica pois neste caso não poderemos derivar $\ell(\theta)$ com relação a θ (ver exemplo ??). Nestes casos, procedimentos especiais devem ser aplicados caso a caso.

18.5 MLE: soluções analíticas

Nesta seção vamos ver em detalhes alguns casos em que soluções exatas, analíticas, sob a forma de fórmulas matemáticas podem ser obtidas. Embora estes sejam casos muito simplificados, eles servem para mostrar algumas propriedades do MLE.

18.5.1 Verossimilhança Bernoulli

Um experimento de Bernoulli é realizado independentemente 10 vezes e observa-se a sequência de resultados. Seja θ a probabilidade de sucesso em um experimento. O verdadeiro valor do parâmetro θ é desconhecido. Sabemos apenas que o espaço paramétrico Θ é o intervalo $[0,1]$. Suponha que tenha sido observado o evento-sequência

$$E = (\widetilde{C}, C, \widetilde{C}, \widetilde{C}, C, \widetilde{C}, C, \widetilde{C}, C, \widetilde{C})$$

onde C representa sucesso e $\widetilde{C}$ representa fracasso. Podemos representar o experimento através de variáveis aleatórias i.i.d. $X_1, \ldots, X_{10}$ com distribuição de Bernoulli com parâmetro θ. A sequência observada é representada pelos dados $\mathbf{y} = (0,1,0,0,1,0,1,0,1,0)$ onde 1 indica C e 0 indica $\widetilde{C}$. A probabilidade da ocorrência de $\mathbf{y}$, que é também a função de verossimilhança de θ, é dada por :

$$\mathbb{L}(\theta) = \mathbb{P}(\mathbf{X} = \mathbf{x}) = \mathbb{P}(\mathbf{X} = (0,1,0,0,1,0,1,0,1,0)) = \theta^4(1-\theta)^6\,. \tag{18.8}$$

Esta probabilidade depende do valor desconhecido de θ. Antes do experimento sabíamos apenas que $\theta \in [0,1]$. Agora, com o experimento realizado, vemos que alguns valores de θ são muito mais verossímeis que outros. O gráfico da Figura 18.8 mostra a função de verossimilhança $\mathbb{L}(\theta)$ versus θ.

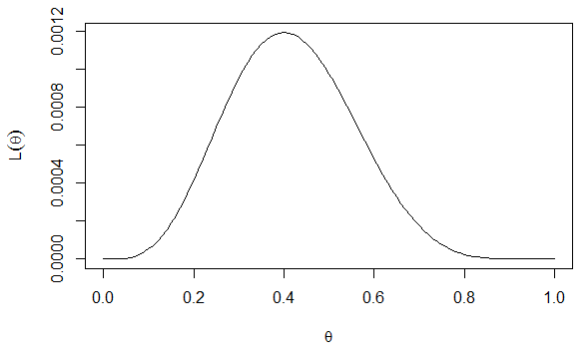


Figure 18.8: Gráfico da função de verossimilhança $\mathbb{L}(\theta) = \theta^4(1-\theta)^6$ versus θ.

Vê-se do gráfico que o valor de θ que maximiza a verossimilhança $\mathbb{L}(\theta)$ é $\widehat{\theta} = 4/10$, a frequência de sucessos nos dez experimentos. Para verificar isto, vamos obter a função log-verossmilhança:

$$\ell(\theta) = \log \mathbb{L}(\theta) = 4\log(\theta) + 6\log(1-\theta)$$

e a equação de verossimilhança é

$$\ell'(\theta) = \frac{\partial \ell}{\partial \theta} = \frac{4}{\theta} - \frac{6}{1-\theta} = 0$$

o que implica na solução $\widehat{\theta} = 4/10 = 0.4$. A partir do gráfico, já sabemos que esta solução é um máximo global.

Não precisamos esperar a realização do experimento para saber quem é a estimativa de máxima verossimilhança. Podemos escrever a estimativa de máxima verossimilhança em função dos valores genéricos de $y_1, \ldots, y_n$ que vão aparecer como reultado da amostragem. Seja $Y_i = 1$ ou $Y_i = 0$ caso o i-ésimo lançamento da moeda seja cara ou coroa, respectivamente. Seja $\mathbf{y} = (y_1, \ldots, y_{10})$ uma realização do experimento. Então $\mathbf{y}$ é uma lista de 1's e 0's e $k = \sum_i y_i$ é igual ao número de caras que ocorreram nesta particular realização do experimento. A probabilidade de ocorrência do evento representado por $\mathbf{y}$ é igual a

$$\mathbb{P}(\mathbf{Y} = \mathbf{y}; \theta) = \theta^k (1-\theta)^{10-k} = \mathbb{L}(\theta)$$

O valor de θ que maximiza a verossimilhança $\mathbb{L}(\theta)$ é encontrado facilmente:

$$\frac{\partial \log \mathbb{L}(\theta)}{\partial \theta} = \frac{\partial}{\partial \theta}\left(k\log(\theta) + (10-k)\log(1-\theta)\right) = \frac{k}{\theta} - \frac{10-k}{1-\theta} = 0$$

o que produz $\widehat{\theta} = k/10$. Este é um ponto de máximo pois $\mathbb{L}(\theta) = 0$ se $\theta = 0$ ou se $\theta = 1$ e obviamente $\mathbb{L}(\theta) > 0$ se $\theta > 0$ e $0 < k < 10$. Portanto, se $0 < k < 10$, tem de existir um ponto de máximo no interior do intervalo. Nos casos extremos, em que $k = 0$ (e portanto $\mathbb{L}(\theta) = (1-\theta)^{10}$) ou em que $k = 10$ (e portanto $\mathbb{L}(\theta) = \theta^{10}$), a solução $\widehat{\theta} = k/10$ ainda é válida.

Assim, em todos os casos, a estimativa de máxima verossimilhança de θ quando o vetor de observações é $\mathbf{y}$ é dado por

$$\widehat{\theta} = \frac{k}{10} = \frac{1}{10}\sum_{i=1}^{10} y_i\,,$$

a média aritmética do vetor de 0's e 1's.

A estimativa $\widehat{\theta}$ obtida acima é uma função do vetor $\mathbf{y}$. Assim, podemos escrever a estimativa de máxima verossimilhança de θ como $\widehat{\theta}(\mathbf{y})$ onde $\mathbf{y}$ é o vetor de observações obtido no experimento.

Este exemplo descreve o método de máxima verossimilhança numa situação simples e, de certo modo, frustrante. Ao final de todo o arrazoado, a estimativa de máxima verossimilhaça é simplesmente o bom e velho $\bar{y}$, uma estimativa que todo mundo já teria proposto sem precisar de um conceito elaborado.

Uma situação muito mais interessante foi apresentada na seção inicial desse capítulo que lidou com o caso de observações de tempo de sobrevivência com dados censurados, quando a característica de interesse não possui um estimador óbvio. No entanto, mesmo no caso em que o estimador de máxima verossimilhança coincide com um estimador intuitivo e simples, será reconfortante saber que este estimador simples é também um estimador ótimo. Esta conclusão virá das propriedades do estimador de máxima verossmilhança que serão estudadas no capítulo 19.

Um comentário importante é sobre o caráter da função de verossimilhança. A função $\mathbb{L}(\theta)$ é derivada a partir de uma função de probabilidade conjunta e, no caso de variáveis aleatórias

discretas, representa uma probabilidade. Entretanto, ela não é a probabilidade de θ ser o verdadeiro valor do parâmetro. Ela fornece uma medida de quão verossímil é cada valor possível do parâmetro θ tendo em vista os dados que foram obtidos. Considerando o exemplo anterior dos sucesso e fracassos de uma moeda, fica mais claro que $\mathbb{L}(\theta)$ não é uma probabilidade nem uma função densidade de probabilidade. A partir de (18.8), sabemos que $\mathbb{L}(\theta) = \theta^4(1-\theta)^6$ para $\theta \in (0,1)$. A integral de $\mathbb{L}(\theta)$ no intervalo $(0,1)$ não é igual a 1. Ela é igual a

$$\int_0^1 \mathbb{L}(\theta)d\theta = \int_0^1 \theta^4(1-\theta)^6 d\theta = \frac{6!4!}{11!},$$

resultado que pode ser obtido facilmente a partir da conhecida constante de integração na expressão da densidade de uma distribuição beta com parâmetros $\alpha = 5$ e $\beta = 7$.

18.5.2 Verossimilhança Binomial

Suponha que desejamos estimar θ, a fração de pessoas que tem tuberculose em uma grande população homogênea. Para isto, nós selecionamos aleatoriamente n indivíduos para um teste e encontramos y deles com a doença. Como a população é grande e homogênea, assunimos que os n indivíduos testados são independentes e cada um deles tem probabilidade θ de ter tuberculose. A verossimilhança de θ é então

$$\mathbb{L}(\theta) = \mathbb{P}(Y = y, \theta) = \binom{n}{y}\theta^y(1-\theta)^{n-y}$$

onde $0 \leq \theta \leq 1$. O espaço paramétrico é $\Theta = [0,1]$. Maximizar $\mathbb{L}(\theta)$ em θ é o mesmo que maximizar $\theta^y(1-\theta)^{n-y}$ e isto foi feito no exemplo anterior da verossimilhança Bernoulli. Portanto, a estimativa de máxima verossimilhança de θ é dada por $\widehat{\theta}(y) = y/n$.

18.5.3 Verossmilhança Poisson

Para o monitoramento de risco ambiental, alguns testes de laboratório são realizados em amostras de água de um rio para determinar se a água é está dentro de limites toleráveis para a saúde humana. De particular interesse é a concentração de certos tipos de bactérias na água. O número destas bactérias é determinado em n amostras de volume unitário da água do rio, dando n contagens observadas $y_1, y_2, \ldots, y_n$. O problema é estimar λ , o número médio de bactérias por unidade de volume no rio.

Nós assumimos que as bactérias estão dispersas na água do rio de modo que as localizações das bactérias são pontos aleatórios no espaço seguindo um processo de Poisson. Então a probabilidade de contar y_i bactérias na i-ésima amostra é dada por uma distribuição de Poisson:

$$\mathbb{P}(Y_i = y_i \mid \lambda) = \frac{\lambda^{y_i}}{y_i!}\exp(-\lambda)$$

onde $y_i = 0, 1, \ldots$ e $0 < \lambda < \infty$. Como volumes disjuntos são independentes, a função de verossimilhança de λ para as contagens observadas $y_1, y_2, \ldots, y_n$ é dada por

$$\mathbb{L}(\lambda) = \mathbb{P}(\mathbf{Y} = \mathbf{y} \mid \lambda) = \prod_{i=1}^{n}\mathbb{P}(Y_i = y_i \mid \lambda) = \prod_{i=1}^{n}\frac{\lambda^{y_i}e^{-\lambda}}{y_i!} = \frac{\lambda^{\Sigma_i y_i}e^{-n\lambda}}{y_1!\ldots y_n!}$$

A função de log-verossimilhança e suas derivadas são as seguintes:

$$\ell(\lambda) = -\log(y_1!,\ldots,y_n!) + \left(\sum_{i=1}^{10} y_i\right)\log\lambda - n\lambda$$

$$\frac{d\ell(\lambda)}{d\lambda} = \frac{1}{\lambda}\sum_{i=1}^{10} y_i - n;$$

$$\frac{d^2\ell(\lambda)}{d\lambda^2} = -\frac{1}{\lambda^2}\sum_{i=1}^{10} y_i \leq 0$$

Se $\sum_i y_i > 0$, a equação de máxima verossimilhança $\ell'(\theta) = 0$ tem uma única solução $\widehat{\lambda}(\mathbf{y}) = \sum_i y_i/n = \bar{y}$. A segunda derivada é negativa neste ponto, indicando que nós temos um máximo relativo. Como $\mathbb{L}(0) = 0$ e $\mathbb{L}(\lambda) \to 0$ quando $\lambda \to \infty$ então $\widehat{\lambda}$ é o máximo global. Se $\sum_i y_i = 0$, temos $\mathbb{L}(\lambda) = e^{-n\lambda}$ e $\ell'(\lambda) = 0$ não tem solução. Fazendo o gráfico da função $\mathbb{L}(\lambda) = e^{-n\lambda}$, vemos que o máximo ocorre na fronteira do espaço paramétrico: $\widehat{\lambda} = 0 = \sum_i y_i/n$. Assim, em qualquer caso, nós temos $\widehat{\lambda}(\mathbf{y}) = \bar{y}$ sendo a estimativa de máxima verossimilhança de λ.

18.5.4 Verossimilhança de Poisson truncada

Usualmente não é possível contar o número Poisson de bactérias em uma amostra de água do rio. Pode-se somente determinar se as bactérias estão ou não presentes na amostra. São incubados e testados n tubos de testes, cada um deles contendo um volume V de água do rio. Um teste negativo mostra que não há bactérias presentes enquanto um teste positivo mostra que existe pelo menos uma bactéria presente no tubo. Se s tubos dentre os n testados dão resultados negativos, qual é a estimativa de máxima verossimilhança de λ?

O resultado do experimento é $\mathbf{y} = (y_1,\ldots,y_n)$ onde $y_i = 1$ ou 0 conforme o teste seja negativo ou positivo, *respectivamente*. Nós não observamos as contagens diretamente, apenas se há ou não bactérias presentes. Apesar disso, de maneira surpreendente, nós ainda podemos ter uma estimativa do número esperado de bactérias mesmo sem observar as contagens diretamente.

Supusemos que a contagem do número de bactérias segue uma distribuição de Poisson com λ bactérias, em média, por unidade de volume. Portanto, a probabilidade de que existam k bactérias em um volume V de água do rio segue uma Poisson com parâmetro λV:

$$p(k) = (\lambda V)^k \frac{e^{-\lambda V}}{k!} \qquad x = 0,1,2,\ldots$$

A probabilidade de uma reação negativa (sem bactérias) é $\pi = p(0) = \exp(-\lambda V)$ e a probabilidade de uma reação positiva (no mínimo uma bactéria) é $1-\pi = 1-\exp(-\lambda V)$. Assumindo que volumes disjuntos são independentes, os n tubos de teste constituem ensaios aleatórios independentes. A probabilidade de observar $s = \sum_i y_i$ reações negativas dentre os n tubos testados é igual a

$$\mathbb{L}(\lambda) = \binom{n}{y}\pi^s(1-\pi)^{n-s} = \binom{n}{s}e^{-s\lambda V}\left(1-e^{-\lambda V}\right)^{n-s} \tag{18.9}$$

onde $0 \leq \lambda < \infty$. Podemos tomar o logaritmo de $\mathbb{L}(\lambda)$ encontrando

$$\mathbb{L}(\lambda) = \text{cte} - s\lambda V + (n-s)\log(1-e^{-\lambda V}) \tag{18.10}$$

onde cte é uma constante em termos de λ. Esta função pode ser derivada e igualada a zero produzindo $\widehat{\lambda} = -\log(s/n)/V$.

Alternativamente, por (18.9), vemos que a função de verossimilhança é um múltiplo da função $\pi^s(1-\pi)^{n-s}$. Vimos no exemplo anterior que esta função atinge seu valor máximo quando $\pi = s/n$. Se $s > 0$, o valor correspondente de λ é obtido resolvendo a equação $\pi = e^{-V\lambda}$, que produz $\lambda = -\log(\pi)/V$. Assim, nós obtemos $\widehat{\lambda} = -\log(s/n)/V = (\log(n) - \log(s))/V$, se $s > 0$. Observe que, se $s = n$, então $\widehat{\lambda} = 0$, no limite do espaço paramétrico.

Caso $s = 0$, a estimativa de máxima verossimilhança não existe. De fato, neste caso, deveríamos ter $0/n = 0 = e^{-V\widehat{\lambda}}$, o que implicaria em $\widehat{\lambda} = \infty$.

Como ilustração numérica, suponha que $n = 40$ tubos de teste, cada um contendo $V = 10$ ml de água do rio estão sob teste. Se $s = 28$ dão testes negativos e $n - s = 12$ dão testes positivos, então

$$\widehat{\lambda} = \frac{\log 40 - \log 28}{10} = 0,0357$$

A concentração de bactérias coliformes por MLE é estimada por 0,0357.

O gráfico da esquerda na Figura 18.9 mostra a função de verossmilhança $\mathbb{L}(\lambda)$ dada em (18.9) para λ entre 0.00 e 0.20. O gráfico da direita mostra mostra a função de log-verossimilhança $\ell(\lambda)$ em (18.10).

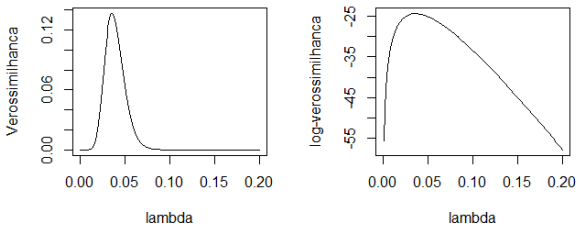


Figure 18.9: Gráfico da função $\mathbb{L}(\lambda)$ (a esquerda) e de $\ell(\lambda)$ (a direita)

Quanto maior a concentração de bactérias no rio, maior a probabilidade de que todos os n tubos dêem resultados positivos. Isto acontece porquê a probabilidade de um tubo ser negativo (contagem igual a zero) é dada por $\exp(-\lambda V)$, uma função que decresce com λ. Isto é relevante porquê os dois casos extremos, $s = 0$ e $s = n$, são indesejáveis. O primeiro porquê produz $\widehat{\lambda} = 0$, uma estimativa baixa demais pois, se $\lambda = 0$, teríamos uma água absolutamente pura, provavelmente uma situação pouco realista. O segundo caso extremo produz $\widehat{\lambda} = \infty$, uma estimativa absurdamente grande. A implicação prática é que devemos escolher um volume V que não seja nem tão pequeno a ponto de gerar apenas tubos negativos (e $s = n$), nem V tão grande que todos os tubos sejam positivos (com $s = 0$). Isto é, se queremos estimar bem o número médio de bactérias por unidade de volume devemos ter tantos tubos positivos quanto tubos negativos na amostra. Isto traz a preocupação prática de escolher um volume V adequado para não terminar com estimativas muito ruins. Usualmente, escolhem-se vários volumes, indo dos volumes maiores, onde sempre há a presença de bactérias, aos volumes menores, onde elas quase nunca estão presentes.

18.5.5 Dados em forma de Tabelas

Dados de n repetições independentes de um experimento são frequentemente resumidos em uma tabela de frequência:

evento ou classe	A_1	A_2	...	A_k	Total
frequência observada	f_1	f_2	...	f_k	n
frequência esperada	np_1	np_2	...	np_k	n

O espaço amostral Ω para uma única repetição do experimento é particionado em k classes ou eventos disjuntos:

$$\Omega = A_1 \cup A_2 \cup \ldots \cup A_k$$

Então f_j é o número de vezes que A_j ocorre em n repetições, sendo que $f_1 + f_2 + \ldots + f_k = n$.

A probabilidade de observar uma tabela de frequência particular é dada pela distribuição multinomial:

$$\mathbb{P}(f_1, \ldots, f_k) = \frac{n!}{f_1!f_2!...f_k!} p_1^{f_1} p_2^{f_2} ... p_k^{f_k}$$

O caso que nos interessa nesta seção é aquele em que o modelo envolve um único parâmetro desconhecido θ e os p_j's são funções deste parâmetro θ. Neste caso, calculamos a estimativa de máxima verossimilhança de θ e, a partir dela, calculamos as freqüências esperadas para efeito de comparação com as freqüências observadas.

Exemplo: Em 200 dias consecutivos de trabalho, dez itens são selecionados diariamente de uma linha de produção e suas imperfeições verificadas. Foram obtidos os seguintes resultados:

número de itens defeituosos	0	1	2	3	≥ 4	Total
freqüência observada	133	52	12	3	0	200

O número de itens defeituosos dentre os 10 observados diariamente segue uma distribuição binomial. Ache a estimativa de máxima verosssimilhança de θ, a probabilidade de que um item é defeituoso e calcule as freqüências esperadas sob o modelo.

De acordo com a distribuição binomial a probabilidade de observar j defeituosos em 10 itens é

$$p_j = \binom{10}{j} \theta^j (1-\theta)^{10-j} \quad j = 0, 1, 2, ..., 10$$

A probabilidade de observar quatro ou mais defeituosos é $p_{4+} = 1 - p_0 - p_1 - p_2 - p_3$. Usando então a distribuição multinomial calculamos a função de verossimilhança de θ:

$$\begin{aligned} \mathbb{L}(\theta) &= k[(1-\theta)^{10}]^{133}[\theta(1-\theta)^9]^{52}[\theta^2(1-\theta)^8]^{12}[\theta^3(1-\theta)^7]^3 \\ &= k\theta^{85}(1-\theta)^{1915} \end{aligned}$$

onde k é a constante

$$k = \frac{200!}{133!52!12!3!0!} \binom{10}{1}^{52} \binom{10}{2}^{12} \binom{10}{3}^{3}.$$

A função de verossimilhança é da mesma forma que a verossimilhança associada com uma simples distribuição binomial com $y = 85$ e com $n = 2000$. Assim, usando os resultados que já obtivemos anteriormente,

$$\widehat{\theta} = \frac{85}{2000} = 0.0425\,.$$

Usando o valor $\theta = 0.0425$, as frequências esperadas podem ser calculadas para $j = 0, 1, 2, 3$. A frequência esperada para a última classe é então obtida por subtração de 200.

número de defeituosos	0	1	2	3	≥ 4	total
frequência observada	133	52	12	3	0	200
frequência esperada	129.54	57.5	11.48	1.36	0.12	200

A concordância entre as freqüências observadas e esperadas parece ser razoavelmente boa. Como os f_js são variáveis aleatórias, é natural que existam algumas diferenças entre as freqüências observadas e esperadas. Um teste de qualidade de ajuste confirma que as diferenças aqui podem facilmente ser atribuidas às variações aleatórias dos f_js e assim o modelo proposto parece satisfatório.

Da fato, agregando as duas últimas classes para não deixar uma classe com contagem esperada menor que um, o teste qui-quadrado fornece:

$$X^2 = \frac{(133-129.54)^2}{129.54} + \frac{(52-57.5)^2}{57.5} + \frac{(12-11.48)^2}{11.48} + \frac{(3-1.48)^2}{1.48} = 2.20$$

que deve ser comparado com a distribuição de referência dada por uma qui-quadrado com 4-1-1 graus de liberdade. O valor 2.20 fornece um p-valor igual a 0.33.

■ **Example 18.13 — nome ??.** Suponha que é retirada uma amostra de uma população em equilíbrio genético com relação a um único gene com dois alelos A e a. Se nós assumimos que os três diferentes genótipos (*AA, Aa, aa*) são identificáveis então somos levados a supor que existem três tipos de indivíduos cujas freqüências relativas esperadas são dadas pelas proporções de Hardy-Weinberg

$$p_{AA} = \theta^2 \quad p_{Aa} = 2\theta(1-\theta) \quad p_{aa} = (1-\theta)^2$$

onde $0 < \theta < 1$ é a proporção da população que possui o alelo A.

Se N_i é o número de indivíduos do tipo i na amostra de tamanho n, então $(N1, N2, N3)$ tem uma distribuição multinomial com parâmetros (n, p_1, p_2, p_3). Queremos estimar θ pelo método de máxima verossimilhança.

Se nós observamos uma amostra de apenas seis indivíduos e obtemos $\mathbf{x} = (AA, Aa, AA, aa, AA, Aa)$, então

$$\mathbb{L}(\theta, \mathbf{x}) = p_{AA}p_{Aa}p_{AA}p_{aa}p_{AA}p_{Aa} = 4\theta^8(1-\theta)^4 .$$

A equação de verossimilhança é

$$\frac{\partial \ell}{\partial \theta} = \frac{\partial}{\partial \theta}(\log 4 + 8\log\theta + 4\log(1-\theta)) = \frac{8}{\theta} - \frac{4}{1-\theta} = 0$$

que tem a única solução $\widehat{\theta} = 2/3$.

Como

$$\frac{\partial^2}{\partial \theta^2}\ell(\theta) = -\frac{8}{\theta^2} - \frac{4}{(1-\theta)^2} < 0$$

para todo $\theta \in (0,1)$, então $2/3$ maximiza $\mathbb{L}(\theta)$.

Em geral, sejam n_{AA}, n_{Aa} e n_{aa} os números de indivíduos dos genótipos *AA, Aa, aa* na amostra. Repetindo os cálculos feitos acima, vemos que, se $2n_1 + n_2$ e $n_2 + n_3$ são ambos positivos, então a estimativa de máxima verossimilhança existe e é dada por

$$\widehat{\theta}(\mathbf{x}) = \frac{2n_1 + n_2}{n} .$$

■

18.6 MLE: soluções numéricas

Na imensa maioria dos casos, o MLE tem de ser encontrado por meio de um método numérico de otimização. A razão é que nem sempre a equação de verossimilhança $\partial\ell(\theta)/\partial\theta = 0$ admite solução analítica. Isto é, não conseguimos isolar $bg\theta$ de um dos lados da equação produzindo assim uma fórmula para o MLE. Nestes casos, precisamos usar um método numérico para encontrar o MLE. Se θ é uni-dimensional ou bi-dimensional, é sempre uma boa idéia fazer um gráfico da função de verossimilhança. A inspeção do gráfico pode revelar situações problemáticas tais como: (a) existem mais de um ponto de máximo local de $\ell(\theta)$; (b) o máximo pode ocorrer na fronteira do espaço paramétrico, o que pode significar que o máximo de $\ell(\theta)$ não se encontra num ponto crítico dessa função.

Vamos começar com o caso unidimensional ($\theta \in \mathbb{R}$) e, em seguida, lidamos com o caso multidimensional ($\theta \in \mathbb{R}^p$).

18.6.1 Caso unidimensional: $\theta \in \mathbb{R}$

Vamos fazer as seguinte suposições:

- O espaço paramétrico Θ é um intervalo $[a,b]$ (o intervalo pode ter comprimento infinito).
- A estimativa de máxima verossimilhança encontra-se no interior de Θ.
- A função log-verossimilhança $l(\theta)$ possui derivadas contínuas até segunda ordem.

Para encontrar o máximo da log-verossimilhança $\ell(\theta)$, basta pesquisar entre as raízes da equação

$$\frac{\partial\ell(\theta)}{\partial\theta} = \ell'(\theta) = 0$$

Vamos ver um dos métodos mais importantes para encontrar as raízes de $\ell'(\theta) = 0$, o método de Newton (ou Newton-Raphson). Denotaremos por $\overline{\theta}$ a raiz procurada. Vamos também considerar uma função genérica $g(\theta)$ para a qual estamos buscando a raiz $\overline{\theta}$. Isto é, $g(\overline{\theta}) = 0$. Logo voltaremos ao caso $g(\theta) = \ell'(\theta)$.

A dificuldade para encontrar a raiz $\overline{\theta}$ é que a função $g(\theta)$ é complicada. Se $g(\theta)$ fosse uma função simples, a tarefa poderia ser trivial. Por exemplo, se $g(\theta)$ for uma função linear de θ, achar a raiz é muito fácil. Por exemplo, achar a raiz de $g(\theta) = 3 + 2\theta = 0$ é imediato: $\overline{\theta} = -3/2$. No caso geral de uma função linear $g(\theta) = a + b\theta$, teríamos $\overline{\theta} = -a/b$. Infelizmente, $g(\theta)$ quase sempre não é simples nem linear. Uma solução grosseira é *forçar* uma aproximação linear para a função $g(\theta)$. Esta é uma das ideias funcdamentais atrás do método de Newton.

Começamos com um valor inicial θ_o. Com sorte, θ_o não estará muito longe de $\overline{\theta}$. Voltaremos ao problema de escolher este valor inicial mais tarde. A Figura 18.10 mostra a função não-linear $g(\theta) = \theta^2 - 5$. Sua raiz é $\overline{\theta} = \sqrt{5} \approx 2.24$. Vamos usar $\theta_o = 1.5$ como valor inicial. O objetivo é encontrar um novo valor θ_1 que esteja mais próximo da raiz $\overline{\theta}$. Para isto, vamos aproximar $g(\theta)$ por uma linha reta. Na prática, nos problemas mais realistas com θ multivariado, não conseguimos ver a função $g(\theta)$ na região de interesse. Precisamos fazer a aproximação linear considerando a função $g(\theta)$ apenas em torno do valor inicial θ_o. Como encontrar a reta que melhor aproxima a curva $g(\theta)$ *em torno do ponto* $(\theta_o, g(\theta_o))$? Ela é a reta tangente à curva $g(\theta)$ passando pelo ponto $(\theta_o, g(\theta_o))$. A equação de uma reta que passa pelo ponto $(\theta_o, g(\theta_o)$ é dada por $g(\theta_o) + b(\theta - \theta_o)$. De fato, esta é a equação de uma reta já que é uma função linear. Além disso, se $\theta = \theta_o$, teremos $g(\theta_o) + b(\theta_o - \theta_o) = g(\theta_o)$. Na Figura 18.10, a reta que passa por $(1.5, g(1.5)) = (1.5, 1.5^2 - 5) = (1.5, -2.75)$ é igual a $-2.75 + b(\theta - 1.5)$.

Falta encontrar a inclinação b. Como a aproximação linear é a reta tangente, temos imediatamente que $b = g'(\theta_o)$. Assim, a reta que melhor aproxima a curva $g(\theta)$ no ponto $(\theta_o, g(\theta_o))$ é dada por $g(\theta_o) + g'(\theta_o)\ (\theta - \theta_o)$. Ao invés de resolver a equação $g(\theta) = 0$, achamos a raiz da reta

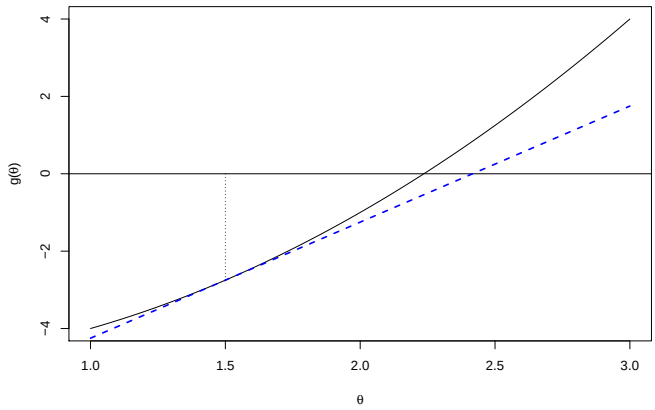


Figure 18.10: Gráfico de uma função $g(\theta)$ e sua aproximação por uma reta que passa pelo ponto $(1.5, g(1.5))$.

tangente. Isto é, achamos θ_1 que soluciona a equação

$$g(\theta_o) + g'(\theta_o)\,(\theta - \theta_o) = 0$$

A resposta é

$$\theta_1 = \theta_o - \frac{g(\theta_o)}{g'(\theta_o)}$$

Atualizamos θ_o dando-lhe o acréscimo $-g(\theta_o)/g'(\theta_o)$. O acréscimo será *positivo* se a função g e a derivada g' no ponto θ_o tiverem sinais trocados. Vamos lembrar que estamos buscando a raiz $\overline{\theta}$ em que $g(\overline{\theta}) = 0$. Assim, por exemplo, se $g(\theta_o) < 0$ e a função estiver crescendo (isto é, $g'(\theta_o) > 0$), então aumentamos θ_o para chegar a um valor θ_1 em que $g(\theta_1) \approx 0$. O tamanho desse acréscimo depende de dois fatores:

- $g(\theta_o)$: mede quão distante nós estamos de $0 = g(\overline{\theta})$. Se $|g(\theta_o)|$ for grande, devemos estar muito distantes do ponto em que g se anula e devemos então fazer acréscimos maiores.
- $g'(\theta_o)$: mede quão rápidamente a função g está mudando de valor. Se $|g'(\theta_o)|$ for muito grande, podemos fazer um acréscimo pequeno.

Considerando a função $g(\theta) = \theta^2 - 5$ da Figura 18.10 com o valor inicial $\theta_o = 1.5$, temos

$$\theta_1 = \theta_o - \frac{g(\theta_o)}{g'(\theta_o)} = \theta_o - \frac{\theta_o^2 - 5}{2\theta_o} = 1.5 - \frac{1.5^2 - 5}{2 \cdot 1.5} = 1.5 + 0.916 = 2.417\,.$$

A ideia de Newton agora é iterar este procedimento. Supondo que θ_1 está mais próximo da raiz que o valor inicial arbitrário, usamos a mesma aproximação tomando θ_1 no lugar de θ_o. Usando o novo ponto θ_1 como valor inicial, repetimos o procedimento acima para encontrar uma nova aproximação θ_2 para a raiz $\overline{\theta}$.

$$\theta_2 = \theta_1 - \frac{g(\theta_1)}{g'(\theta_1)}\,.$$

A Figura 18.11 mostra esta segunda aproximação linear que passa tangenciando a curva no ponto $(\theta_1, g(\theta_1)) = (2.417, g(2.417))$. A raiz desta nova reta é quase idêntica à raiz desejada. Iterando, obtemos θ_3, $\theta_4, \ldots$ sucessivamente que devem estar cada vez mais próximos da raiz desejada.

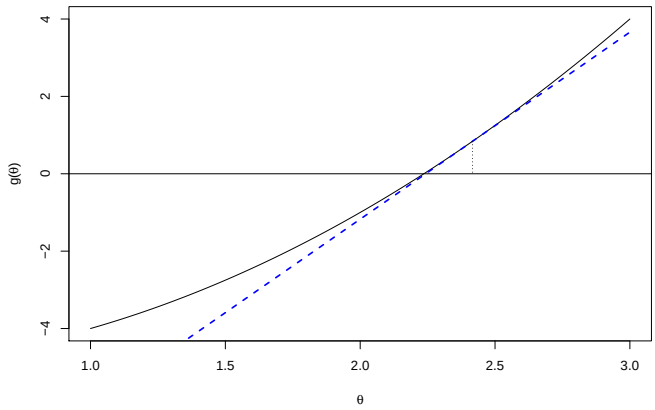


Figure 18.11: Gráfico da função $g(\theta)$ versus θ com a reta tangente que passa pelo ponto $(\theta_1, g(\theta_1)) = (2.417, g(2.417))$. A raiz desta nova reta é quase idêntica à raiz desejada.

Quando a diferença absoluta $|\theta_{n+1} - \theta_n|$ for menor que um pequeno limite ε, interrompemos o procedimento numérico. Usamos o último valor calculado no processo iterativo como sendo a aproximação final para $\overline{\theta}$. Podemos também interromper as iterações quando a diferença relativa $|\theta_{n+1} - \theta_n|/|\theta_n|$ for pequena.

A equação de iteração é

$$\theta_{n+1} = \theta_n - \frac{g(\theta_n)}{g'(\theta_n)}$$

A função $g(\theta)$ de nosso interesse é a derivada da função de log-verossmilhança $g(\theta) = \ell'(\theta)$ Assim, o método de Newton-Raphson para encontrar o máximo da log-verossimilhança fica igual a:

$$\theta_{n+1} = \theta_n - \frac{\ell'(\theta_n)}{\ell''(\theta_n)} \tag{18.11}$$

A convergência, quando ela acontece, costuma ser rápida. Entretanto, podem aparecer dificuldades com o método de Newton-Raphson. Em primeiro lugar, precisamos ter um valor inicial θ_o. Se ele for muito ruim, o método pode demorar a convergir ou pode até não convergir. Em segundo lugar, nem sempre o método de Newton-Raphson funciona. Veja a Figura 18.12.

18.6.2 Exemplos de MLE 1-dim por métodos numéricos

■ **Example 18.14 — Acidentes de trabalho.** Quando há um acidente com um funcionário em uma fábrica, este acontecimento é registrado. Aqui está a lista dos n funcionários acidentados num dado ano com número de acidentes sofridos:

Funcionário	1	2	3	4	5	...	n-1	n
No. de acidentes	1	1	2	1	1	...	1	1

Não se sabe quantos funcionários existem ao todo na fábrica. Isto é, não ficou registrado o número de funcionários que *não se acidentaram no ano*. Suponha que o número de acidentes S que um funcionário sofre num ano segue uma Poisson(λ). Assim, o número esperado de acidentes que um funcionário sofre ao longo de um ano é λ. Vamos supor o mesmo λ para todos os funcionários. Suponha também que os funcionários sofrem acidentes independentemente uns dos outros.

Queremos estimar λ com base apenas nos casos em que pelo menos um acidente foi observado. A dificuldade para obter o MLE de λ é que não observamos X diretamente. Observamos apenas as

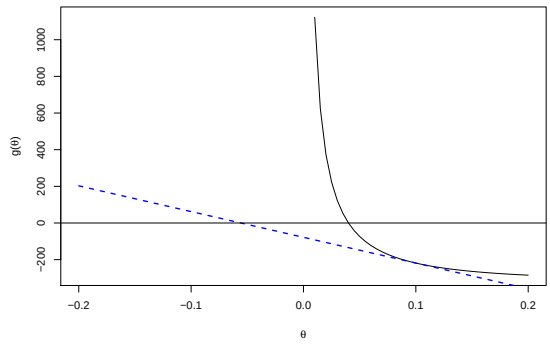


Figure 18.12: Exemplo onde o método de Newton-Raphson não converge se iniciarmos com o valor $\theta_o = 0.1$. Iniciando com $\theta_o < 0.05$, teremos convergência.

v.a.'s X quando $X \geq 1$, chamada *distribuição truncada*. Nunca observamos o evento $[X = 0]$. Não temos como estimar diretamente $\mathbb{P}(X = 0)$ pois não sabemos quantos funcionários existem ao todo na fábrica e quantos deles tiveram 0 acidentes.

Temos $\mathbb{P}(X = k) = \frac{\lambda^k}{k!}e^{-\lambda}$ para $k = 0, 1, 2, \ldots$. A tabela apresenta 11 valores observados independentemente da variável aleatória X truncada em $X = 0$. Isto é, se um funcionário não sofre nenhum acidente isto não é registrado. A variável medida é $Y = (X \mid X > 0)$, o valor de X dado que X é maior que zero. Assim temos na tabela $y_1, \ldots y_n$, os valores observados de $Y_1, \ldots, Y_n$ que são i.i.d. As variáveis aleatórias observadas possuem distribuição dada por

$$\mathbb{P}_\lambda(Y = k) = P_\lambda(X = k \mid X > 0) = \frac{\mathbb{P}_\lambda(X = k)}{\mathbb{P}_\lambda(X > 0)} = \frac{\lambda^k}{k!}\frac{1}{e^\lambda - 1}$$

Assumindo que são i.i.d., a função de probabilidade conjunta é dada por

$$\mathbb{P}(\mathbf{Y} = \mathbf{y}) = \mathbb{P}(Y_1 = y_1)\ldots\mathbb{P}(Y_{11} = y_{11}) = \prod_{i=1}^{n}\left(\frac{e^{-\lambda}}{1 - e^{-\lambda}}\frac{\lambda^{y_i}}{y_i!}\right) = \frac{\lambda^{\sum_{i=1}^{n} y_i}}{(e^\lambda - 1)^n \prod_{i=1}^{n} y_i!}$$

onde $\mathbf{y} = (y_1, y_2, \ldots, y_n)$. A função de log-verossimilhança é igual a

$$\begin{aligned}\ell(\lambda) &= \log\mathbb{P}(\mathbf{Y} = \mathbf{y}) \\ &= -n\log(e^\lambda - 1) + \left(\sum_{i=1}^{n} y_i\right)\log\lambda - \sum_{i=1}^{n}\log(y_i!)\end{aligned}$$

Portanto,

$$\ell'(\lambda) = \frac{-ne^\lambda}{e^\lambda - 1} + \frac{1}{\lambda}\sum_i y_i = \frac{-ne^\lambda}{e^\lambda - 1} + \frac{\sum_i y_i}{\lambda}$$

A estimativa de máxima verossimilhança será a solução $\widehat{\lambda}$ da equação

$$\ell'(\lambda) = \frac{-ne^\lambda}{e^\lambda - 1} + \frac{n\bar{y}}{\lambda} = 0$$

Note que trocamos $\sum_i y_i$ por $n\bar{y}$.

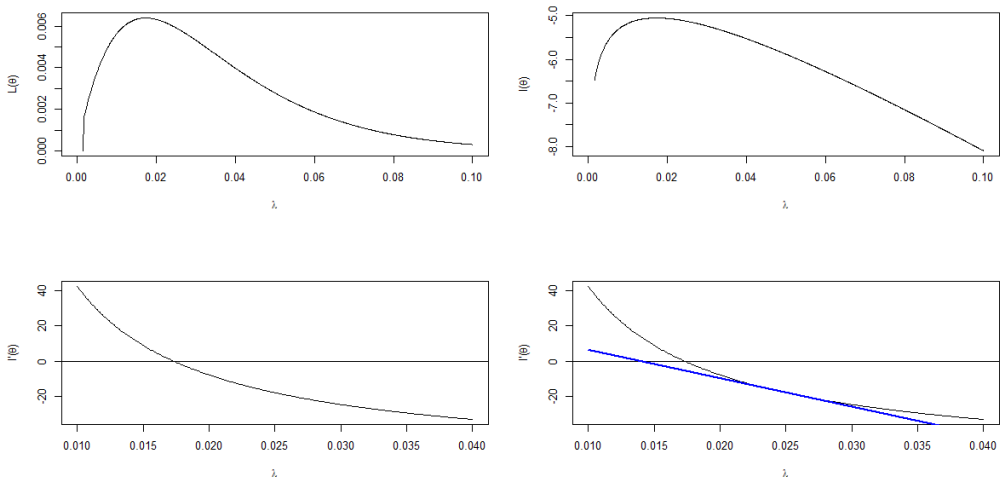


Figure 18.13: Gráfico da função de verossimilhança $L(\lambda)$, da função log-verossimilhança $\ell(\lambda)$ e da função $\ell'(\lambda)$. é mostrado também o primeiro passo do método de Newton-Raphson usando $\lambda_o = 0.025$.

A equação de iteração de Newton-Raphson neste exemplo precisamos das expressões analíticas de $\ell'(\lambda)$ e de $\ell''(\lambda)$. Já temos a primeira delas. A segunda é igual a

$$\ell''(\lambda) = \frac{ne^{-\lambda}}{(1-e^{\lambda})^2} - \frac{n\bar{y}}{\lambda^2}$$

Assim, a fórmula de iteração do método de Newton- Raphson fica

$$\lambda_{k+1} = \lambda_k - \frac{\dfrac{-n}{1-e^{-\lambda_k}} + \dfrac{n\bar{y}}{\lambda_k}}{\dfrac{ne^{-\lambda_k}}{(1-e^{\lambda_k})^2} - \dfrac{n\bar{y}}{\lambda_k^2}}$$

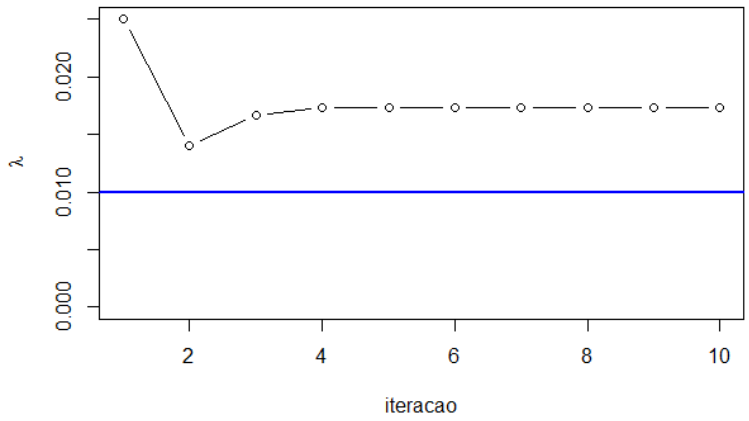
Isto é,

$$\lambda_{k+1} = \lambda_k - \frac{\lambda_k(1-e^{-\lambda_k})(-n\lambda_k + n\bar{y}(1-e^{-\lambda_k}))}{-n\lambda_k^2 e^{-\lambda_k} - n\bar{y}(1-e^{-\lambda_k})^2}$$

Vamos simular dados de uma indústria com 10 mil funcionários sendo o número de acidentes de cada funcionário no ano uma v.a. de Poisson com $\lambda = 0.01$ e independente dos demais funcionários. Vamos usar apenas os $955+56+2 = 1013$ funcionários que tiveram pelo menos um acidente no ano. A média aritmética desses dados é $(955+56\star 2+2\star 3)/1013 = 1.06$, muito maior que o verdadeiro valor $\lambda = 0.01$. A Figura 18.13 mostra o gráfico da função de verossimilhança $L(\lambda)$ e da função log-verossimilhança $\ell(\lambda)$ na parte superior. Na parte inferior, o gráfico da esquerda é o da função $\ell'(\lambda)$. Queremos o ponto λ em que esta função se anula. No lado direito, é mostrado também o primeiro passo do método de Newton-Raphson usando $\lambda_o = 0.025$.

Na Figura 18.14, mostramos os 10 primeiros passos do método de Newton-Raphson. Fica claro que existe uma rápida convergência para o valor 0.01734. O valor verdadeiro de λ usado para gerar os dados é representado pela linha horizontal azul.

■



18.6.3 Caso multidimensional: $\theta \in \mathbb{R}^p$

Como fazer quando $\theta = (\theta_1, \ldots, \theta_k)$ é um vetor? O princípio é o mesmo: procurar $\theta \in \Theta$ que maximize a verossimilhança:

$$\widehat{\theta} = \left(\widehat{\theta}_1, \widehat{\theta}_2, \ldots, \widehat{\theta}_k\right) = \arg\max_{\theta} \ell(\theta)$$

onde $\ell(\theta)$ é a função de log-verossimilhança de θ. Temos uma amostra $\mathbf{y} = (Y_1, \ldots, Y_n)$ composta de variáveis aleatórias discretas ou contínuas com densidade $f(\mathbf{y}|\theta)$. Então

$$\ell(\theta) = \log f(\mathbf{y}|\theta)$$

Usualmente, o máximo MLE $\widehat{\theta}$ é obtido resolvendo *simultaneamente* as k equações baseadas nas derivadas parciais:

$$\begin{cases} \frac{\partial \ell(\theta)}{\partial \theta_1} = 0 \\ \frac{\partial \ell(\theta)}{\partial \theta_2} = 0 \\ \quad \cdots \\ \frac{\partial \ell(\theta)}{\partial \theta_k} = 0 \end{cases}$$

Defina o vetor gradiente

$$\frac{\partial \ell(\theta)}{\partial \theta} = \left(\frac{\partial \ell(\theta)}{\partial \theta_1}, \frac{\partial \ell(\theta)}{\partial \theta_2}, \ldots, \frac{\partial \ell(\theta)}{\partial \theta_k}\right)^t$$

Assim, o sistema de equações pode ser escrito de forma vetorial:

$$\frac{\partial \ell(\theta)}{\partial \theta} = \mathbf{0} = (0, 0, \ldots, 0)^t$$

Uma solução é um *ponto crítico*.

Quando um ponto crítico é um ponto de máximo de $\ell(\theta)$? Olhamos para a matriz de segunda derivada de $\ell(\theta)$ avaliada no ponto $\widehat{\theta}$:

$$D^2 \log L(\theta) \mid_{\theta=\widehat{\theta}} = \begin{bmatrix} \frac{\partial^2 \ell(\theta)}{\partial \theta_1^2} & \frac{\partial^2 \ell(\theta)}{\partial \theta_1 \partial \theta_2} & \cdots & \frac{\partial^2 \ell(\theta)}{\partial \theta_1 \partial \theta_k} \\ \frac{\partial^2 \ell(\theta)}{\partial \theta_2 \partial \theta_1} & \frac{\partial^2 \ell(\theta)}{\partial \theta_2^2} & \cdots & \frac{\partial^2 \ell(\theta)}{\partial \theta_2 \partial \theta_k} \\ & & \cdots & \\ & & \cdots & \\ \frac{\partial^2 \ell(\theta)}{\partial \theta_k \partial \theta_1} & \frac{\partial^2 \ell(\theta)}{\partial \theta_k \partial \theta_2} & \cdots & \frac{\partial^2 \ell(\theta)}{\partial \theta_k^2} \end{bmatrix} \mid_{\theta=\widehat{\theta}}$$

Esta matriz é avaliada no ponto $\widehat{\theta}$. Verificamos se ela é definida negativa. Se for, então o ponto crítico $\widehat{\theta}$ é um ponto de máximo.

MLE é a solução da equação de log-verossimilhança, um sistema de quações não-lineares (às vezes, com restrições):

$$D\ell(\theta) = \begin{bmatrix} \frac{\partial \ell}{\partial \theta_1}(\theta) \\ \vdots \\ \frac{\partial L}{\partial \theta_k}(\theta) \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix} = \mathbf{0}$$

Em geral, este sistema NÃO tem solução fechada (analítica).

Equação iterativa do método de Newton-Raphson no caso univariado:

$$\theta_{n+1} = \theta_n - \frac{\ell'(\theta_n)}{\ell''(\theta_n)}$$

Caso multivariado:

$$\theta_{n+1} = \theta_n - \left[D^2 \ell(\theta_n)\right]^{-1} D\ell(\theta_n)$$

onde $D\ell(\theta_n)$ é o vetor $k \times 1$ de derivadas parciais e $D^2\ell(\theta_n)$ é a matriz $k \times k$ de derivadas parciais de segunda ordem da log-verossimilhança. $D\ell(\theta_n)$ e $D^2\ell(\theta_n)$ são avaliados no valor corrente de θ_n.

18.6.4 De volta à regressão logística

Relembre nosso exemplo de regressão logística: dados são pares de vetores (x_i, y_i) onde x_i é a idade da i-ésima criança $y_i = 1$ ou 0, dependendo do sucesso ou não em executar uma tarefa. Modelo é que as v.a.'s $y_1, \ldots, y_n$ são independentes com distribuição de Bernoulli. A probabilidade de sucesso da criança depende de sua idade. Temos

$$P(Y_i = 1) = p(x_i) = \frac{1}{1 + \exp\left(-(\beta_0 + \beta_1 x_i)\right)}$$

Temos $\theta = (\beta_0, \beta_1)$ e a log-verossimilhança é dada por

$$\begin{aligned} \ell(\theta) &= \log\left(\prod_{i=1}^{n} P(Y_i = y_i)\right) \\ &= \log\left(\prod_{i=1}^{n} p(x_i)^{y_i} (1 - p(x_i))^{1-y_i}\right) \\ &= \beta_0 \sum_{i=1}^{n} y_i + \beta_1 \sum_{i=1}^{n} x_i y_i - \sum_i \log(1 + e^{\beta_0 + \beta_1 x_i}) \end{aligned}$$

Como Y_i é uma variável binária, $\sum_{i=1}^{n} x_i y_i$ e $\sum_{i=1}^{n} y_i$ são subtotais aleatórios das colunas da matriz

$$\begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$$

Os elementos incluídos na soma são aqueles que correspondem à uma resposta do tipo $Y = 1$. O EMV de $\theta = (\beta_0, \beta_1)$ é obtido via Newton-Raphson:

$$\widehat{\theta}_{n+1} = \widehat{\theta}_n - [D^2\ell(\widehat{\theta}_n)]^{-1} D\ell(\widehat{\theta}_n)$$

Com $p(x_i) = p_i = 1/(1+e^{-(\beta_0+\beta_1 x_i)})$ temos

$$D\ell(\theta_n) = \begin{pmatrix} \frac{\partial \log \ell}{\partial \beta_0} \\ \frac{\partial \log \ell}{\partial \beta_1} \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{n} (y_i - p_i) \\ \sum_{i=1}^{n} (x_i y_i - p_i x_i) \end{pmatrix}$$

onde p_i é calculado com o valor corrente $\theta_n = (\beta_{0n}, \beta_{1n})$

Temos

$$D\ell(\theta) = \begin{pmatrix} \frac{\partial^2 \ell}{\partial \beta_0^2} & \frac{\partial^2 \ell}{\partial \beta_0^2 \partial \beta_1^2} \\ \frac{\partial^2 \ell}{\partial \beta_0^2 \partial \beta_1^2} & \frac{\partial^2 \ell}{\partial \beta_2^2} \end{pmatrix} = - \begin{pmatrix} \sum_{i=1}^{n} p_i(1-p_i) & \sum_{i=1}^{n} p_i(1-p_i)x_i \\ \sum_{i=1}^{n} p_i(1-p_i)x_i & \sum_{i=1}^{n} p_i(1-p_i)x_i^2 \end{pmatrix}$$

Como valor inicial, use $\theta_0 = (\log(\bar{y}/(1-\bar{y})), 0)$. Isto corresponde a um modelo sem efeito de idade (pois $\beta_1 = 0$) e portanto com a mesma probabilidade de sucesso para todas as crianças. Neste caso, como $p_i \equiv p$ pode ser estimado pela proporção total de crianças que tiveram sucesso: $\hat{p} = \sum_i y_i/n = \bar{y}$. Então: $\sum_i y_i/n = \hat{p} = 1/(1+\exp(-\widehat{\beta}_0))$ o que implica em $\widehat{\beta}_0 = \log(\bar{y}/(1-\bar{y}))$.

Script R para logística

• Demo

Observamos $Y_1, \ldots, Y_n$ v.a.'s binárias independentes: Ensaios de Bernoulli. A probabilidade de sucesso NÃO é a mesma para todas as observações. Algumas tem mais chance de ser sucesso do que outras. Vamos excrever $p_i = \mathbb{P}(Y_i = 1)$ Como esta chance p_i varia de observação para observação? Varia em função de p atributos medidos em cada exemplo: regressores ou variáveis independentes. Podemos ASSUMIR uma forma funcional específica para modelar esta dependência de p_i em função dos atributos.

Vamos assumir uma função logística. Pegue um preditor linear do sucesso: $\eta = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$ Transforme este preditor para cair no intervalo $(0,1)$, que é a faixa de variação de probabilidades. Fazemos:

$$p = \frac{1}{1+e^{-\eta}} = \frac{1}{1+e^{-(\beta_0+\beta_1 x_1 + \ldots + \beta_p x_p)}} = \frac{1}{1+e^{-\mathbf{x}^t.\theta}}$$

onde $\mathbf{x} = (1, x_1, \ldots, x_p)$ é o vetor de atributos medidos em cada exemplo e $\theta = (\beta_0, \beta_1, \ldots, \beta_p)$ é o vetor de parâmetros desconhecidos.

Seja $\mathbf{X}$ a matriz $n \times (p+1)$ com as variáveis regressoras. A primeira coluna é toda de 1's. As outras colunas são os valores dos regressores para cada um dos exemplos. Crie também o vetor $\mathbf{y}$ de dimensão $n \times 1$ com os valores da variável resposta. Temos o vetor-coluna $\theta = (\beta_0, \beta_1, \ldots, \beta_p)$ de dimensão $(p+1) \times 1$.

Notação matricial

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix} \qquad \theta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} \qquad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

As v.a.'s binárias $y_1, y_2, \dots, y_n$ são independentes e, para cada exemplo, temos o modelo

$$y_i \sim \text{ Bernoulli}(p_i)$$

onde

$$p_i = \frac{1}{1+e^{-\eta_i}} = \frac{1}{1+e^{-(\beta_0+\beta_1 x_{i1}+\ldots+\beta_p x_{ip})}} = \frac{1}{1+e^{-\mathbf{x}_i^t\theta}}$$

onde $\mathbf{x}_i = (1, x_{i1}, \dots, x_{ip})^t$ é a i-ésima LINHA da matriz $\mathbf{X}$ visto como um vetor-coluna e $\theta = (\beta_0, \beta_1, \dots, \beta_p)$ é o vetor-coluna de parâmetros desconhecidos.

Parâmetro que queremos estimar: $\theta = (\beta_0, \beta_1, \dots, \beta_p)$. Verossimilhança: como as v.a.'s y_i são discretas, a veross de certo valor para θ é a probab de observar os dados REALMENTE observados. Isto é

$$\begin{aligned} \ell(\theta) &= \log\left(\prod_{i=1}^{n} P(Y_i=1)^{y_i}\, P(Y_i=0)^{1-y_i}\right) \\ &= \sum_{i=1}^{n} [y_i \log(p_i) + (1-y_i)\log(1-p_i)] \\ &= \sum_{i=1}^{n} \left[y_i\, \mathbf{x}_i^t\theta - \log(1+e^{\mathbf{x}_i^t\theta})\right] \end{aligned}$$

Para maximizar $\ell(\theta)$, tomamos derivadas em relação a cada elemento de $\theta = (\beta_0, \dots, \beta_p)$ e igualamos a zero:

$$\begin{aligned} \frac{\partial \ell(\theta)}{\partial \beta_j} &= \sum_{i=1}^{n} y_i x_{ij} - \sum_{i=1}^{n} \frac{x_{ij} e^{\mathbf{x}_i^t\theta}}{1+e^{\mathbf{x}_i^t\theta}} \\ &= \sum_{i=1}^{n} y_i x_{ij} - \sum_{i=1}^{n} p_i x_{ij} \\ &= \sum_{i=1}^{n} x_{ij}(y_i - p_i) \end{aligned}$$

para todo $j = 0, 1, \dots, p$. Esta é a equação de verossmilhança (na verdade, um sistema de $p+1$ equações não-lineares em θ).

Em forma matricial, nós escrevemos

$$\begin{aligned} \frac{\partial \ell(\theta)}{\partial \theta} &= \sum_{i=1}^{n} \mathbf{x}_i\, (y_i - p_i) \\ &= \mathbf{X}^t(\mathbf{y} - \mathbf{p}) \end{aligned}$$

onde $\mathbf{x}_i$ é a i-ésima LINHA da matriz $\mathbf{X}$ escrita como um vetor-coluna de dimensão $(p+1)\times 1$ e $\mathbf{p} = (p_1, p_2, \dots, p_n)$. Resolver $\partial L(\theta)/\partial\theta = 0$ é equivalente a resolver $\mathbf{X}^t(\mathbf{y}-\mathbf{p}) = \mathbf{0}$, ou seja,

$$\mathbf{X}^t\mathbf{y} = \mathbf{X}^t\mathbf{p}$$

O lado esquerdo é um vetor $(p+1)\times 1$ de constantes conhecidas. O lado direito envolve o parâmetro θ desconhecido: ele está embutido de forma não-linear na expressão para $p_i = 1/(1+e^{\mathbf{x}_i^t\theta})$.

No método de Newton-Raphson, precisamos também da matriz de derivadas parciais de segunda ordem (chamada de matriz Hessiana). O elemento na r-ésima linha e j-ésima coluna da matriz Hessiana é (contando a partir de zero)

$$\begin{aligned}\frac{\partial^2\ell(\theta)}{\partial\theta_r\partial\theta_j} &= -\sum_{i=1}^{n}\frac{(1+e^{\mathbf{x}_i^t\theta})e^{\mathbf{x}_i^t\theta}\,x_{ir}x_{ij}-(e^{\mathbf{x}_i^t\theta})^2\,x_{ir}x_{ij}}{(1+e^{\mathbf{x}_i^t\theta})^2}\\ &= -\sum_{i=1}^{n}x_{ir}x_{ij}p_i - x_{ir}x_{ij}p_i^2\\ &= -\sum_{i=1}^{n}x_{ir}x_{ij}p_i(1-p_i)\end{aligned}$$

Podemos escrever isto numa forma de matriz $(p+1)\times(p+1)$:

$$\frac{\partial^2\ell(\theta)}{\partial\theta\partial\theta^t} = -\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^t\;p_i(1-p_i)$$

Isto é,

$$\frac{\partial^2\ell(\theta)}{\partial\theta\partial\theta^t} = -\mathbf{X}^t\mathbf{W}\mathbf{X}$$

onde $\mathbf{W}$ é uma matriz $n\times n$ diagonal com i-ésimo elemento $p_i(1-p_i)$:

$$\mathbf{W} = \begin{bmatrix} p_1(1-p_1) & 0 & 0 & \dots & 0\\ 0 & p_2(1-p_2) & 0 & \dots & 0\\ 0 & 0 & p_3(1-p_3) & \dots & 0\\ \vdots & \vdots & \vdots & \dots & \vdots\\ 0 & 0 & 0 & \dots & p_n(1-p_n)\end{bmatrix}$$

Comece com um vetor inicial $\theta^{(0)}$ e itere até convergência:

$$\theta^{\text{new}} = \theta^{\text{old}} - \left(\frac{\partial^2 L(\theta)}{\partial\theta\partial\theta^t}\right)^{-1}\frac{\partial L(\theta)}{\partial\theta}$$

onde as derivadas são avaliadas usando-se o valor corrente θ^{old}. Vamos substituir as derivadas pelas expressões matriciais que encontramos anteriormente para a regressão logística.

Como

$$\frac{\partial L(\theta)}{\partial\theta} = \mathbf{X}^t(\mathbf{y}-\mathbf{p})$$

e

$$\frac{\partial^2 L(\theta)}{\partial\theta\partial\theta^t} = -\mathbf{X}^t\mathbf{W}\mathbf{X}$$

Temos então

$$\theta^{\text{new}} = \theta^{\text{old}} + \left(\mathbf{X}^t\mathbf{W}\mathbf{X}\right)^{-1}\mathbf{X}^t(\mathbf{y}-\mathbf{p})$$

Logística no R: `glm`

Comando `glm` implementa a regressão logística (bem como a classe geral Generalized Linear Models, GLM)

Valor de retorno de `glm` é uma lista com vários elementos necessários para a análise de dados:

```
ajuste = glm(y ~ x1 + x2, family=binomial("logit"))

ajuste$coef
(Intercept)        x1        x2
 -0.7681903   0.6820035   0.3665339

names(ajuste)
 [1] "coefficients"      "residuals"         "fitted.values"     "effects"
 [5] "R"                 "rank"              "qr"                "family"
 [9] "linear.predictors" "deviance"          "aic"               "null.deviance"
[13] "iter"              "weights"           "prior.weights"     "df.residual"
[17] "df.null"           "y"                 "converged"         "boundary"
[21] "model"             "call"              "formula"           "terms"
[25] "data"              "offset"            "control"           "method"
[29] "contrasts"         "xlevels"
```

Sem nenhuma consideração por eficiência numérica, vamos considerar uma implementação rudimentar da regressão logística em R usando nossas fórmulas

Ver final do script R do proximo exemplo (diabetes)

IRLS - OPCIONAL

A equação de iteração do MLE pode ser colocada num formato que é geral (servirá para muitas outras distribuições e modelos) e que usa apenas regressão de mínimos quadrados com uma matriz de pesos W. Temos

$$\begin{aligned}\theta^{\text{new}} &= \theta^{\text{old}} + \left(\mathbf{X}^t\mathbf{W}\mathbf{X}\right)^{-1}\mathbf{X}^t(\mathbf{y}-\mathbf{p}) \\ &= \left(\mathbf{X}^t\mathbf{W}\mathbf{X}\right)^{-1}\mathbf{X}^t\mathbf{W}\left(\mathbf{X}\theta^{\text{old}} + \mathbf{W}^{-1}(\mathbf{y}-\mathbf{p})\right) \\ &= \left(\mathbf{X}^t\mathbf{W}\mathbf{X}\right)^{-1}\mathbf{X}^t\mathbf{W}\mathbf{z}\end{aligned}$$

onde $\mathbf{z} = \mathbf{X}\theta^{\text{old}} + \mathbf{W}^{-1}(\mathbf{y}-\mathbf{p})$

Vimos que

$$\theta^{\text{new}} = \left(\mathbf{X}^t\mathbf{W}\mathbf{X}\right)^{-1}\mathbf{X}^t\mathbf{W}\mathbf{z}$$

onde $\mathbf{z} = \mathbf{X}\theta^{\text{old}} + \mathbf{W}^{-1}(\mathbf{y}-\mathbf{p})$ Se $\mathbf{z}$ é visto como um vetor resposta e $\mathbf{X}$ uma matriz de regressores, então θ^{new} é a solução de um problema de mínimos quadrados ponderados:

$$\theta^{\text{new}} \leftarrow \arg\min_{\theta} (\mathbf{z}-\mathbf{X}\theta)^t\,\mathbf{W}\,(\mathbf{z}-\mathbf{X}\theta)$$

OBS: Mínimos quadrados ordinários (não-ponderado) resolve

$$\arg\min_{\theta} (\mathbf{z}-\mathbf{X}\theta)^t\,(\mathbf{z}-\mathbf{X}\theta)$$

Assim, Newton-Raphson reduz-se a uma série iterativa de mínimos quadrados. Este algoritmo é o *Iteratively Reweighted Least Squares* (IRLS).

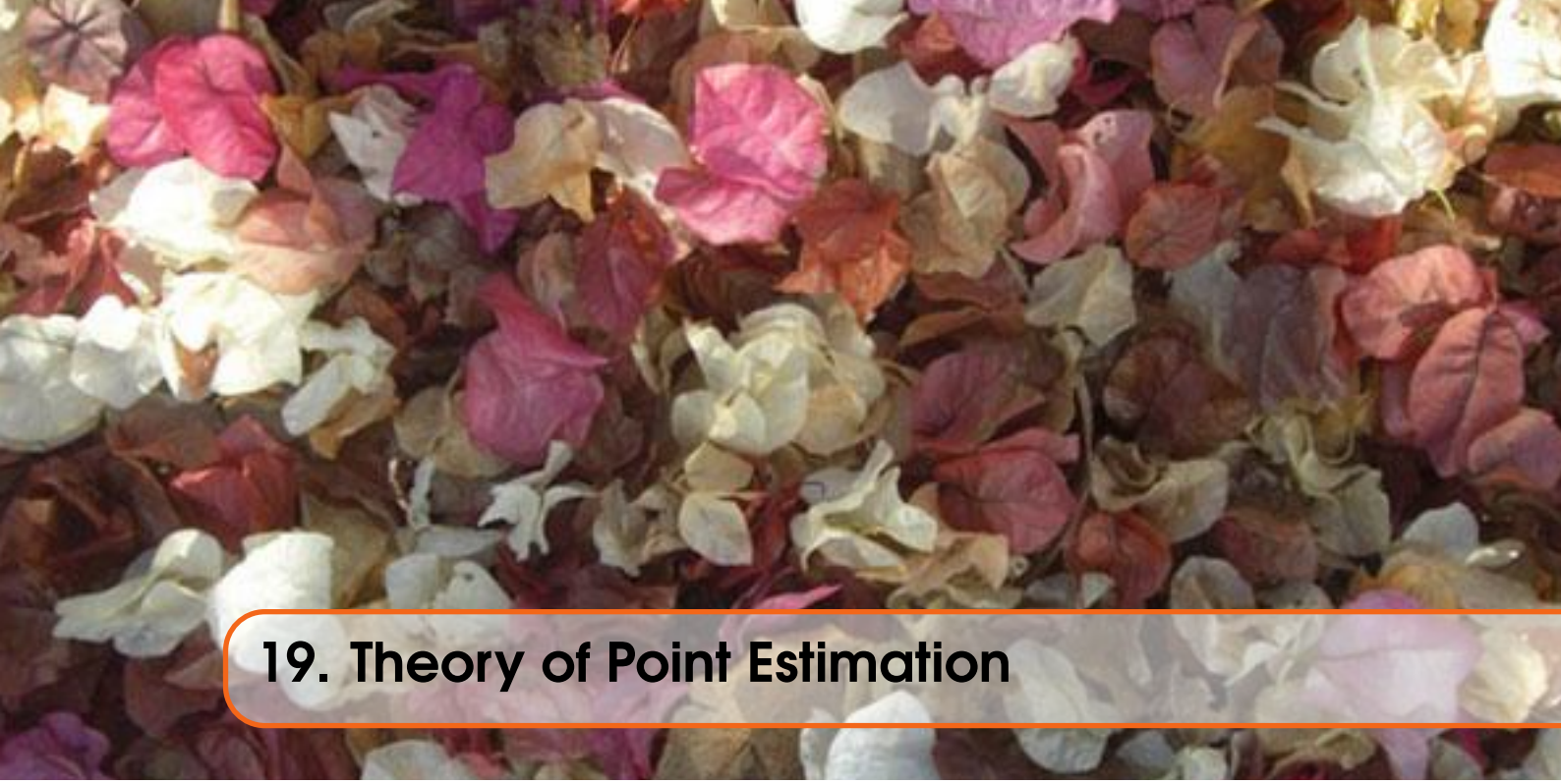
Calcule um vetor inicial $\theta^{(0)} = (\log(\bar{y}/(1-\bar{y})), 0, \ldots, 0)$. Itere até convergência: Calcule o vetor $n \times 1$ de probabilidades $\mathbf{p}$ usando o vetor corrente θ^{old}

$$p_i = p_i(\theta^{\text{old}}) = \frac{1}{1+\exp(-\mathbf{x}_i^t\,\theta^{\text{old}})} \quad i = 1, \ldots, n$$

Calcule a matriz $\widetilde{\mathbf{X}}$ de dimensão $n \times (p+1)$ multiplicando cada linha de $\mathbf{X}$ por $p_i(1-p_i)$:

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1^t \\ \mathbf{x}_2^t \\ \vdots \\ \mathbf{x}_n^t \end{bmatrix} \quad \widetilde{\mathbf{X}} = \begin{bmatrix} p_1(1-p_1)\mathbf{x}_1^t \\ p_2(1-p_2)\mathbf{x}_2^t \\ \vdots \\ p_n(1-p_n)\mathbf{x}_n^t \end{bmatrix}$$

$$\theta \leftarrow \theta + \left(\mathbf{X}^t\widetilde{\mathbf{X}}\right)^{-1} \mathbf{X}^t(\mathbf{y}-\mathbf{p})$$



19. Theory of Point Estimation

19.1 Outros métodos de estimação

Aprendemos um método para estimar o vetor de parâmetros θ de um modelo estatístico para os dados $\mathbf{Y} = (Y_1, \ldots, Y_n)$: , o método de máxima verossimilhança, que produz o estimador de máxima verossimilhança (MLE). Este método é intuitivo e pode ser usado sempre que pudermos escrever a verossimilhança $L(\theta)$. Mas o MLE não é o único método para estimar parâmetros. Existem vários outros métodos, todos tão intuitivos quanto o MLE. Nem sempre eles coincidem nas suas estimativas de θ. Se eles não coincidem, como escolher um deles? Vamos definir propriedades desejáveis que um bom método deveria satisfazer. Em seguida, vamos verificar que o MLE satisfaz estas propriedades desejáveis. Antes de definir estas propriedades, vamos ver exemplos de diferentes métodos de estimação. Eles são úteis para, por exemplo, fornecer valores iniciais para obtero MLE pelo procedimento de Newton.

19.1.1 Método de Momentos

Já usamos este método informalmente várias vezes antes. O caso inicial mais simples é aquele em que $\mathbf{Y} = (Y_1, \ldots, Y_n)$ é formado por variáveis i.i.d. com uma densidade que depende de um único parâmetro. Por exemplo, $\mathbf{Y}$ poderia ser composto por v.a.'s Poisson(θ) ou poderiam ser $N(\theta, 1)$. Como $\mathbb{E}(Y_i) = \theta$, sugerimos usar a média aritmética $\bar{Y} = (Y_1 \ldots + Y_n)/n$ como estimador de $\mathbb{E}(Y_i) = \theta$.

Em geral, o momento teórico $\mathbb{E}(Y_i)$ será uma função matemática $g(\theta)$ do parâmetro θ. Nos casos acima, tivemos $g(\theta) = \theta$ mas podemos ter uma função $g(\theta)$ diferente da função identidade.

■ **Example 19.1 — Pareto e o método de momentos.** Se $\mathbf{Y} = (Y_1, \ldots, Y_n)$ é formado por variáveis i.i.d. com uma densidade de Pareto dependente apenas do parâmetro α e dada por

$$f(y) = \begin{cases} \alpha/y^{\alpha+1} & \text{se } y \geq 1 \\ 0, & \text{se } y < 1 \end{cases}$$

Temos $\mathbb{E}(Y_i) = g(\alpha) = \alpha/(\alpha - 1)$, quando $\alpha > 1$. Se $\alpha < 1$, o método de momentos não pode ser aplicado não existe a esperança de Y_i pois a integral $\int x f(x) dx$ não converge. Igualamos

$\mathbb{E}(Y_i) = g(\alpha) = \alpha/(\alpha-1)$ com a média aritmética $\bar{Y}$ dos dados e invertemos de modo a isolar $\alpha = g^{-1}(\bar{Y})$, obtendo assim o estimador de momentos de α:

$$\alpha/(\alpha-1) = \bar{Y} \Rightarrow \hat{\alpha}_{MM} = \frac{\bar{Y}}{\bar{Y}-1}$$

No caso deste exemplo, o MLE é uma expressão muito diferente: $\hat{\alpha}_{MLE} = \frac{n}{\sum_i \log(Y_i)}$, o inverso da média aritmética dos logaritmos dos dados. Qual dos dois estimadores deveria ser escolhido? E por que não fazer combinações deles, tais como uma média ponderada $\hat{\alpha}_{comb} = 0.7 \times \hat{\alpha}_{MLE} + 0.3\hat{\alpha}_{MM}$ ou $\hat{\alpha}_{comb} = 0.5 \times \hat{\alpha}_{MLE} + 0.5\hat{\alpha}_{MM}$? ∎

Muitas vezes, a distribuição dos dados depende de mais de um parâmetro tais como $N(\mu, \sigma^2)$ ou uma Gama(α, β) onde $\theta = (\mu, \sigma^2)$ e $\theta = (\alpha, \beta)$, respectivamente. A ideia do método de momentos é obter os dois primeiros momentos teóricos, $\mathbb{E}(Y_i)$ e $\mathbb{E}(Y_i^2)$, que serão ambos funções do parâmetro θ. Igualamos estes dois momentos teóricos aos dois primeiros momentos amostrais e resolvemos o sistema de duas equações com duas incógnitas.

■ **Example 19.2 — Gama e o método de momentos.** Se $\mathbf{Y} = (Y_1, \ldots, Y_n)$ é formado por variáveis i.i.d. com uma densidade Gama dependente do parâmetro $\theta = (\alpha, \beta)$ e dada por

$$f(y) = \begin{cases} \text{cte. } y^{\alpha-1}e^{-\beta y} & \text{se } y > 0 \\ 0, & \text{caso contrário} \end{cases}$$

A Figura 19.1 mostra um histograma de amostra de $n = 200$ valores retirados de uma distribuição gama com parâmetros α e β. O problema prático de estimação é: tendo escolhido o modelo gama para os dados, usar apenas os dados da amostra para inferir os valores desconhecidos de α e β. Precisamos dos dois primeiros momentos teóricos: $\mathbb{E}(Y_i) = \alpha/\beta$ e $\mathbb{E}(Y_i^2) = \alpha(\alpha+1)/\beta^2$. Veja que cada um desses momentos teóricos é uma função do vetor de parâmetros $\theta = (\alpha, \beta)$.

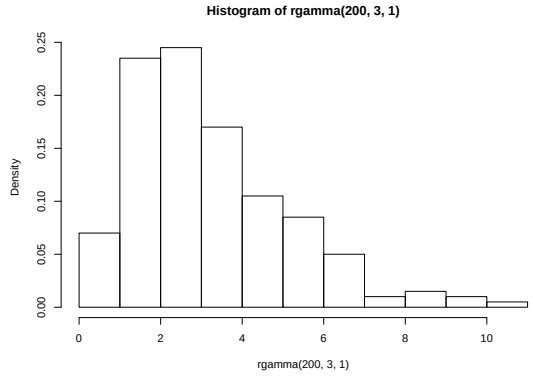


Figure 19.1: Amostra de $n = 200$ valores de Gama$(\alpha = 3, \beta = 1)$

Precisamos também dos dois primeiros momentos amostrais, inteiramente baseados nos dados: $m_1 = \sum_i Y_i/n = \bar{Y}$ e $m_2 = \sum_i Y_i^2/n = \bar{Y^2}$. Igualamos os momentos correspondentes, teóricos e amostrais: $\mathbb{E}(Y) = m_1$ e $\mathbb{E}(Y^2) = m_2$. Isto é:

$$\begin{aligned} \frac{\alpha}{\beta} &= m_1 \\ \frac{\alpha(\alpha+1)}{\beta^2} &= m_2 \end{aligned}$$

Em seguida, resolvemos este sistema de equações não lineares para encontrar a solução $\hat{\alpha}$ e $\hat{\beta}$. Com simples manipulação algébrica, encontramos a solução: $\widehat{\alpha} = m_1^2/(m_2 - m_1^2)$ e $\widehat{\beta} = m_1/(m_2 - m_1^2)$. Estes são os estimadores de momentos de $\theta = (\alpha, \beta)$.

O que é o MLE neste caso de v.a.'s iid gama? O MLE não dá o mesmo resultado que o método de momentos. A log-verossimilhança é:

$$\ell(\alpha,\beta) = n\beta\log(\alpha) - n\log(\Gamma(\alpha)) + (\alpha-1)\sum_i \log(y_i) - \beta\sum_i y_i$$

Derivando em relação a α e β, temos:

$$\begin{aligned}\frac{\partial \ell}{\partial \beta} &= \frac{n\alpha}{\beta} - \sum_i y_i \\ \frac{\partial \ell}{\partial \alpha} &= n\log(\beta) - n\frac{\Gamma'(\alpha)}{\Gamma(\alpha)} + \sum_i \log(y_i)\end{aligned}$$

Igualando a zero a derivada em β produz $\widehat{\beta} = \widehat{\alpha}/\overline{y}$. A seguir, igualando a zero a derivada em α produz

$$n\log(\beta) - n\Gamma'(\alpha)/\Gamma(\alpha) + \sum_i \log(y_i) = 0$$

Substituindo $\beta = \alpha/\overline{y}$ teremos

$$n\log(\alpha) - n\log(\overline{y}) - n\Gamma'(\alpha)/\Gamma(\alpha) + \sum_i \log(y_i) = 0$$

Esta é uma equação com uma única incógnita, α. Precisamos encontrar a solução numericamente. A maioria das linguagens possuem implementada a função $\psi(\alpha) = \Gamma'(\alpha)/\Gamma(\alpha)$, chamada de função digama. Pode-se usar o estimador de momentos como valores iniciais num procedimento de Newton. O ponto importante para nós é que o MLE $\hat{\theta}$ difere do estimador de momentos neste problema. Qual dos dois estimadores deveria ser usado?

■

Passando para o caso mais geral, suponha que $\mathbf{Y} = (Y_1,\ldots,Y_n)$ é formado por variáveis i.i.d. cuja distribuição depende do vetor de parâmetros $\theta \in \mathbb{R}^k$. A ideia do método de momentos é igualar os k primeiros momentos teóricos, que serão funções de θ, aos k momentos amostrais:

Momento Teórico	Momento Amostral
$\mathbb{E}(Y) = g_1(\theta)$	$m_1 = \overline{Y} = \sum_{i=1}^n Y_i/n$
$\mathbb{E}(Y^2) = g_2(\theta)$	$m_2 = \sum_i Y_i^2/n$
$\vdots$	$\vdots$
$\mathbb{E}(Y^k) = g_k(\theta)$	$m_k = \sum_i Y_i^k/n$

Resolve-se o sistema de equações não lineares produzindo o estimador de momentos $\hat{\theta}$.

A justificativa do método de momentos é que, pela lei dos grandes números (ver capítulo ??), $m_1 = \sum_i Y_i/n$ converge para $\mathbb{E}(Y)$ quando $n \to \infty$. Assim, se o tamanho n da amostra é grande, podemos esperar $m_1 \approx \mathbb{E}(Y)$. Pela mesma lei dos grandes números, $m_k = \sum_i Y_i^k/n$ converge para $\mathbb{E}(Y^k)$. Assim, $m_2 \approx \mathbb{E}(Y^2)$, $m_3 \approx \mathbb{E}(Y^3)$, etc. Trocamos a aproximação por uma igualdade para obter a estimativa de momentos.

19.1.2 Um método sem nome

Um terceiro método é baseado na estatística de Kolmogorov. Até onde eu saiba, este método está sendo inventado aqui, neste livro, e tem caráter puramente didático. O objetivo é apenas ilustrar que outros métodos intuitivamente razoáveis para estimar parâmetros podem ser inventados facilmente.

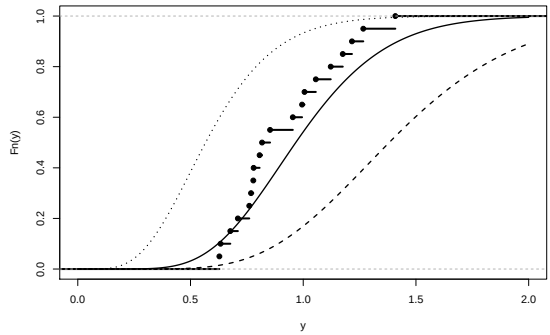


Figure 19.2: Função distribuição acumulada empírica $\mathbb{F}_n(y)$ e três funções de distribuições acumuladas: Gama(10,10) (linha sólida), Gama(10, 7) (linha tracejada) e Gama(6,10) (linha pontilhada).

Considere o modelo de v.a.'s i.i.d. Gama(α,β). Construa a função acumulada empírica com os dados observados. Para cada α e β fixos, podemos calcular a distribuição acumulada teórica de uma Gama(α,β). A ideia então é obter os parâmetros α e β que tornam as duas funções acumuladas, empírica e teórica, o mais próximas possível.

A Figura 19.2 mostra como este método funciona. Nela, vemos um gráfico com a função distribuição acumulada empírica $\mathbb{F}_n(y)$ de uma amostra com $n = 20$ dados (a função escada). Esta função é calculada usando apenas os dados, sem nenhuma referência ou uso de algum modelo teórico. O gráfico mostra também três funções de distribuição acumulada teórica de um modelo gama com diferentes parâmetros α e β: Gama(10,10) (linha sólida), Gama(10, 7) (linha tracejada) e Gama(6,10) (linha pontilhada). É mais ou menos claro que, dentre as três curvas apresentadas, aquela associada com a distribuição Gama(10,10) é a mais próxima da função acumulada empírica. É a melhor possível?

Vamos usar como critério para estimar $\theta = (\alpha,\beta)$ a minimização da distância de Kolmogorov. Para cada valor possível para o parâmetro $\theta = (\alpha,\beta)$, obtenha a função distribuição acumulada teórica $F(y;\theta)$ de uma Gama com parâmetros α e β. A seguir, calcule a distância de Kolmogorov entre esta acumulada teórica $F(y;\theta)$ e a distribuição acumulada empírica $\mathbb{F}_n(y)$:

$$D_n(\mathbf{y},\theta) = \max_y |\mathbb{F}_n(y) - F(y;\theta)|$$

A distância $D_n(\mathbf{y},\theta)$ é função dos dados $\mathbf{y}$ e do valor escolhido para θ. Com os dados $\mathbf{y}$ fixos, ela é função apenas de θ. Obtenha agora o valor de $\theta = (\alpha,\beta)$ que minimiza a distância $D_n(\mathbf{y},\theta)$:

$$\widehat{\theta} = (\widehat{\alpha},\widehat{\beta}) = \arg\min_\theta D_n(\mathbf{y},\theta)$$

Pois bem, este método é intuitivamente razoável também, não? E no entanto, ele vai dar um resultado diferente do método de momentos e do MLE. Como ecolher entre um deles? Ou devemos combiná-los de algum modo tal como, por exemplo, através de uma média ponderada desses três estimadores?

19.1.3 Mais um método sem nome

Lembre-se do modelo de regressão logística com k covariáveis e vetor de coeficientes $\theta = (\beta_0,\beta_1,\ldots,\beta_k)$. O i-ésimo indivíduo tem suas características ou covariáveis agrupadas no vetor $\mathbf{x}_i = (1,x_{i1},\ldots,x_{i,k})'$. O preditor linear da probabilidade de sucesso do i-ésimo indivíduo é a combinação linear $\eta_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} = \mathbf{x}_i'\theta$ das suas características. Este preditor linear

é o que governa a probabilidade de sucesso, dada pelo modelo logístico

$$\begin{aligned}\mathbb{P}(Y_i = 1) &= \frac{1}{1+e^{-\eta_i}} \\ &= \frac{1}{1+\exp(-(\beta_0+\beta_1 x_{i1}+\ldots+\beta_k x_{ik}))} \\ &= p_i(\theta)\end{aligned}$$

Já vimos como obter o MLE de $\theta = (\beta_0, \beta_1, \ldots, \beta_k)$. Um outro método, simples e intuitivo, é o seguinte. Chute um valor qualquer para $\theta = (\beta_0, \beta_1, \ldots, \beta_k)$ e obtenha as probabilidades $p_i(\theta)$ para cada indivíduo. Compare com os dados e ache o valor de θ que mais se ajuste aos dados. Por exemplo, ache o valor de θ que minimize a soma das diferenças (em valor absoluto) entre a resposta binária y_i a probabilidade predita $p_i(\theta)$:

$$\widehat{\theta} = \min_{\theta} D(\mathbf{y}, \theta)$$

onde

$$D(\mathbf{y}, \theta) = \sum_i |y_i - p_i(\theta)| .$$

A função $D(\mathbf{y}, \theta)$ é uma medida da discrepância entre os dados $\mathbf{y}$ e o modelo parametrizado por θ.

Para exemplificar este estimador, considere o caso em que temos uma única covariável, como no caso do teste de desenvolvimento infantil de Denver visto no capítulo 17. Na Figura 19.3, nos dois gráfico, vemos os dados (x_i, y_i) de $n = 173$ crianças onde x_i é a idade da i-ésima criança em meses e y_i é a variável binária indicando o sucesso ou fracasso na dessa criança na realização do teste. A curva logística depende do vetor paramétrico $\theta = (\beta_0, \beta_1)$ e é da forma $p(\theta) = 1/(1+\exp(-\beta_0 - \beta_1 x)$. A Figura 19.3 mostra duas possíveis curvas logísticas: com $\theta = (-6.3, -0.35)$ (à esquerda) e $\theta = (-5.85, -0.39)$ (à direita).

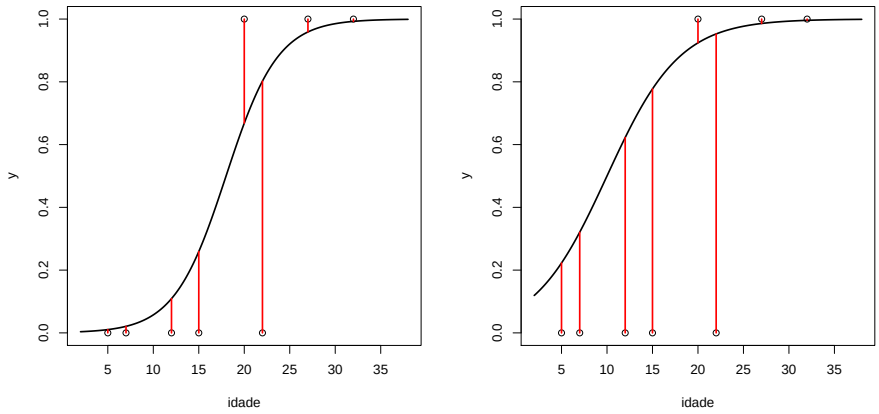


Figure 19.3: Duas possíveis curvas com os valores de $|y_i - p_i(\theta)|$.

A medida de discrepância $D(\mathbf{y}, \theta) = \sum_i |y_i - p_i(\theta)|$ é a soma das barras verticais conectando a resposta y_i de cada indivíduo com a sua probabilidade de sucesso $p_i(\theta)$. Estas barras seráo pequenas quando $y_i = 1$ e $p_i(\theta) \approx 1$ ou então quando $y_i = 0$ e $p_i(\theta) \approx 0$. Nos casos em que existe grande discrepância entre os dados e a predição do modelo com um valor especificado para o parâmetro θ teremos muitos indivíduos com $y_i = 0$ e $p_i(\theta) \approx 1$ ou então com $y_i = 1$ e $p_i(\theta) \approx 0$. O

estimador do parâmetro θ é encontrado obtendo o valor θ que minimiza a medida de discrepância $D(\mathbf{y},\theta)$.

Podemos inventar muitos critérios razoáveis neste problema. Por exemplo:

- Mudando ligeiramente a medida de discrepância para considerar as diferenças ao quadrado: encontre θ que minimize $D(\mathbf{y},\theta) = \sum_i (y_i - p_i(\theta))^2$.
- Dando mais peso aos pontos que possuem probabilidade póxima de 1/2, procuramos θ que minimize $\sum_i (p_i(\theta)(1-p_i(\theta))\ |y_i - p_i(\theta)|)$. Veja que $p(\theta)(1-p(\theta)) \approx 0$ se $p(\theta) \approx 0$ ou ≈ 1. Assim, damos mais peso aos pontos no "centro", aqueles em que $p(\theta) \approx 1/2$.

Podemos misturar estimadores obtidos por métodos diferentes. Suponha que $\widehat{\theta}_{MLE}$ seja o estimador MLE na regressão logística. Seja $\widehat{\theta}_D$ o estimador que minimiza $D(\theta) = \sum_i |y_i - p_i(\theta)|$. Defina então seu estimador preferido como: $\widehat{\theta} = (\widehat{\theta}_{MLE} + \widehat{\theta}_D)/2$, a média aritmética de $\widehat{\theta}_{MLE}$ e $\widehat{\theta}_D$. Podemos preferir usar outra média ponderada qualquer tal como: $\widehat{\theta} = 0.75\ \widehat{\theta}_{MLE} + 0.25\ \widehat{\theta}_D$.

A lição com estes exemplos é que podemos pensar em *infinitos* estimadores, todos aparentemente razoáveis. Como escolher entre eles? Veremos uma teoria que nos ajuda neste sentido a seguir.Mais importante ainda: ela diz que, num certo sentido que vamos tornar preciso, o MLE é o melhor que podemos fazer. Um outro método qualquer pode até ser tão bom quanto o MLE mas não melhor que ele. Esta afirmação é demasiado geral e precisa ser melhor qualificada mas ela explica a imensa popularidade do método de máxima verossimilhança na análise de dados.

19.2 Estimadores são variáveis aleatórias

Para escolher bons estimadores, precisamos de uma teoria que nos guie. Nesta teoria, será *fundamental* ver os estimadores como variáveis ou vetores aleatórios. O que é uma variável ou vetor aleatório? Já aprendemos a entoar o mantra: são duas listas, uma com os valores possíveis, e outra com as probabilidades associadas.

Associada com esta necessidade de diferenciar estimadores como variáveis aleatórias, é importante entender a diferença conceitual entre parãmetros, estimadores e estimativas. Vamos diferenciar μ e $\overline{Y}$. Suponha que $Y_1, Y_2, \ldots, Y_n$ sejam i.i.d. $N(\mu,1)$. Queremos estimar $\mu = \mathbb{E}(Y_i)$. Usamos $\overline{Y} = (Y_1 + \ldots + Y_n)/n$

Qual a diferença entre μ e $\overline{Y}$? Gerei no R cinco v.a.'s $N(0,1)$. Assim, $\mu = 0$. O resultado foi: `-0.962 -0.293 0.259 -1.152 0.196`. Tivemos a estimativa $\overline{y} = -0.390$ para o parâmetro μ. Fazendo uma nova simulação, tivemos `0.030 0.085 1.117 -1.219 1.267` e nova estimativa: $\overline{y} = 0.256$.

Note que μ não muda de valor quando uma nova amostra é retirada. Temos sempre $\mu = 0$ aqui. Entretanto, $\overline{y}$ muda de valor de amostra para amostra. Isto indica que μ e $\overline{y}$ não podem ser as mesmas coisas. De fato, μ é um número real. Ele é desconhecido em geral mas não é uma variável aleatória. O parâmetro μ *não é* composto por duas listas, uma de valores possíveis e e outra com probabilidades associadas.

Para diferenciar entre estimador e estimativa, vamos realizar o experimento de retirar duas amostras de tamanho 5 de $N(0,1)$. Você vai retirar a segunda amostra com nova semente. Antes de fazer o experimento, responda: qual das duas amostras, a 1a. ou a 2a., será a melhor para estimar μ? Isto é, qual das duas amostras vai produzir uma média aritmética $\overline{Y}$ mais próxima de μ? Mesmo sabendo o verdadeiro valor de μ, é impossível responder a isto *antes* de retirarmos as amostras.

A razão é que o estimador $\overline{Y}$ é uma *variável aleatória*. Sendo assim, $\overline{Y}$ é simplesmente a coleção de duas listas: uma com os valores possíveis, e outra com as probabilidades associadas. Quando obtemos uma amostra específica, tal como `-0.962 -0.293 0.259 -1.152 0.196`, calculamos $\overline{y} = (y_1 + \ldots + y_5)/5 = (-0.962 + \ldots + 0.196)/5 = -0.390$. Este número $\overline{y}$ é uma *estimativa*: é a instância específica do estimador $\overline{Y}$, que se materializa numa amostra particular. A estimativa $\overline{y}$ é apenas um número, enquanto $\overline{Y}$ é uma variável aleatória (e portanto, duas listas) Em termos

de notação, a única diferença entre o estimador $\overline{Y}$ e a estimativa $\bar{y}$ é o uso de letras maiúscula e minúsculas.

Alternativamente, poderíamos decidir usar outro estimador, a variável aleatória mediana amostral M. Este estimador é obtido assim:

- ordene a amostra: $Y_{(1)} = \min\{Y_1, \ldots, Y_5\}$, o mínimo da amostra,
- $Y_{(2)}$ é o segundo menor valor da amostra, etc.
- até obtermos $Y_{(5)} = \max\{Y_1, \ldots, Y_5\}$, o máximo da amostra.
- Então tome $M = Y_{(3)}$, o elemento central na lista ordenada, como um estimador de μ

Quem é melhor para estimar μ: a *variável aleatória M* ou a *variável aleatória* $\overline{Y}$? Vamos tentar responder a isto simulando num computador. Com as duas amostras de uma $N(0,1)$ que vimos antes, `-0.962 -0.293 0.259 -1.152 0.196` e `0.030 0.085 1.117 -1.219 1.267`, nós temos:

- Primeira amostra: $\bar{y} = -0.390$ e $m = -0.292$
- Segunda amostra: $\bar{y} = 0.256$ e $m = 0.085$

Note que neste caso simulado, diferentemente do que acontece na análise de dados reais, nós sabemos de antemão que $\mu = 0$. Sendo assim, nem faz sentido prático querer estimar μ se sabemos seu valor verdadeiro, com precisão absoluta. Entretanto, estamos apenas procurando verficar com a simulação qual dos dois estimadores, $\overline{Y}$ ou M, é o melhor estimador.

Nas duas amostras acima, a v.a. M esteve mais próxima de $\mu = 0$ que $\overline{Y}$. Isto talvez seja um indicativo de que M tem um erro de estimação sempre menor que $\overline{Y}$. Isto é *falso*: numa terceira amostra temos os dados `-0.745 -1.131 -0.716 0.253 0.152` com $\bar{y} = -0.437$ e $m = -0.716$. Neste caso, tivemos $\overline{Y}$ mais próximo de μ que M. O fato é que, às vezes, teremos $|\bar{y} - \mu| < |m - \mu|$ mas às vezes teremos o contrário. O que acontece *em geral*? Qual o comportamento *estatístico* das v.a.'s M e $\overline{Y}$?

Vamos fazer um estudo de simulação um pouco mais extenso para estudar o comportamento estatístico dos estimadores M e $\overline{Y}$. Vamos simular 1000 amostras de tamanho 5, calcular M e $\overline{Y}$ em cada delas e verificar o erro que cada estimador cometeu em cada uma das 1000 amostras. O código abaixo mostra o que foi feito. Primeiro 5000 valores i.i.d. de uma $N(0,1)$ foram organizados numa matriz `mat` de dimensão 1000×5. E seguida, calculamos a média aritmética $\overline{Y}$ de cada linha obtendo o vetor `media` de tamanho 1000. Fizemos o mesmo calculando a mediana M e obtendo o vetor `med` de tamanho 1000.

```
mat = matrix(rnorm(5*1000), ncol=5)
media = apply(mat, 1, mean) # media de cada linha
med = apply(mat, 1, median) # mediana de cada linha
aux = range(c(med, media)) # min e max das estimativas
par(mfrow=c(1,2))
plot(media, med, asp=1, ylab="mediana") # media x mediana
abline(0,1)
plot(abs(media), abs(med), asp=1) # |media| x |mediana|
sum(abs(media - 0) > abs(med - 0))
[1] 427
```

A Figura 19.4 mostra um gráfico de dispersão dos 1000 valores de M e $\overline{Y}$. Vemos que eles são positivamente correlacionados. Quando $\bar{y}$ sobrestima $\mu = 0$ (isto é, quando $\bar{y} > 0$), temos a tendência de também observar m acima de $\mu = 0$. Isto é razoável. Quando $M > 0$, temos uma amostra em que 3 ou mais de seus 5 valores estão acima de 0. Neste caso, podemos esperar que a média aritmética $\bar{y}$ desses mesmos 5 valores $\bar{y}$ tenha uma boa chance de também ser positiva. No gráfico à direita mostramos os erros de estimação dos dois estimadores (em valor absoluto) em cada amostra: $|m - \mu| = |m|$ versus $|\bar{y} - \mu| = |\bar{y}|$. Embora seja mais difícil de visualizar, parece

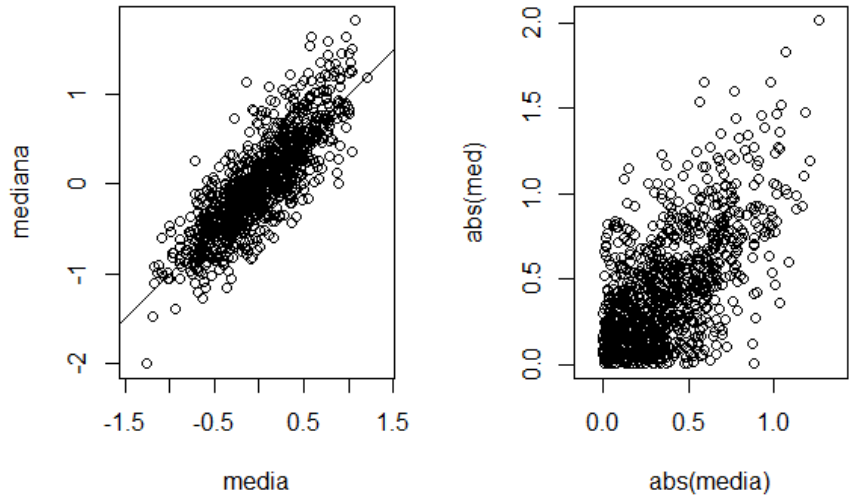


Figure 19.4: Esquerda: Gráfico de dispersão dos 1000 valores de M e $\overline{Y}$, estimadores de $\mu = 0$ em amostras de 5 observações independentes de uma $N(0,1)$. Direita: Gráfico de dispersão dos erros de estimação dos dois estimadores (em valor absoluto) em cada amostra: $|M-\mu| = |M|$ e $|\overline{Y}-\mu| = |\overline{Y}|$.

que $|\bar{y}|$ tende a estar mais próximo de zero que o correspondente valor $|m|$. No último comando do script acima, verificamos que isto é verdade. Calculamos o número de amostras dentre as 1000 em que tivemos o erro de estimação $\overline{Y}$, dado por $|\overline{Y}-\mu| = |\overline{Y}-0|$, maior que o erro de estimação de M, dado por $|M-\mu| = |M-0|$. Obtivemos 427 em 1000. Assim, usando a ideia de frequência relativa, podemos estimar $\mathbb{P}(|\overline{Y}-\mu| > |M-\mu|) \approx 0.427$.

Podemos concluir que o estimador $\overline{Y}$ é melhor que o estimador M sempre? Para todo tamanho de amostra n, e não apenas com $n = 5$ como neste exemplo? Para todo valor de μ, e não apenas para $\mu = 0$? Para todo valor do desvio-padrão σ, e não apenas para $\sigma = 1$? O tamanho de mostras foi adequado? Mil simulações parece um número pequeno para alcançar decisões muito firmes. Note, por exemplo, que a estimativa da probabilidade de ter um erro de estimação menor usando $\overline{Y}$ foi 0.427, muito próximo de 0.5, quando não haveria diferença entre eles. Como concluir de forma geral e definitiva sobre a qualidade de um estimador frente a outro? Vamos dar mais um passo nesta direção na próxima seção.

19.3 Estimação Pontual

Temos uma amostra aleatória $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)$ de v.a.'s. A distribuição conjunta de $\mathbf{Y}$ pertence a uma família (ou modelo) paramétrico com densidade conjunta $f(\mathbf{y}, \theta) = f(y_1, y_2, \ldots, y_n; \theta)$. O parâmetro θ é um vetor de dimensão k pertencente a um conjunto $\Theta \subseteq \mathbb{R}^k$ chamado *espaço paramétrico*. Deseja-se inferir sobre θ. Conhecendo-se θ podemos (em princípio, pelo menos) calcular a probabilidade de qualquer evento ocorrer.

Definition 19.3.1 — Estatística. Uma estatística é uma função matemática $g(\mathbf{Y}) = g(Y_1, \ldots, Y_n)$ que tenha como argumento $\mathbf{Y}$ e que tome valores em $\mathbb{R}^h$ Uma estatística não pode envolver os parâmetros desconhecidos θ.

Definition 19.3.2 — Estimador pontual. Um estimador pontual de θ ou, de forma mais geral, de uma função $q(\theta)$ será qualquer estatística $g(\mathbf{Y}) = g(Y_1, \ldots, Y_n)$. A única diferença entre uma estatística e um estimador é que ao definir um estimador precisamos declarar o quê ele está estimando (declarar $q(\theta)$).

O motivo para lidarmos com a notação mais geral $q(\theta)$ é que, às vezes, podmeos estar interessados apenas num aspecto da distribuição dos dados, sem muito interesse pela distribuição como um todo. Por exemplo, se $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)$ é uma amostra de v.a.'s i.i.d. com distribuição gama com parâmetros $\theta = (\alpha, \beta)$, podemos estar interessado apenas na sua esperança $\mu = q(\alpha, \beta) = \alpha/\beta$. Assim, podemos tentar estimar diretamente μ sem necessariamente precisar estimar ambos, α e β. Entretanto, vamos nos concentrar neste texto no problema de estimar θ, o parâmetro que indexa a família de densidades $f(\mathbf{y}, \theta)$ associada com os dados da amostra.

■ **Example 19.3** Seja uma amostra $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)$ de v.a.'s i.i.d. gaussianas $N(\mu, \sigma^2)$. Isto é, $\mathbb{E}(Y_i) = \mu$ e $\text{Var}(Y_i) = \mathbb{E}(Y_i - \mu)^2 = \sigma^2$. Seja $\overline{Y} = \frac{1}{n}\sum_i Y_i$, a média aritmética das v.a.'s. A média aritmética $\overline{Y}$ é usualmente tomada como um estimador de μ. A variância amostral $S^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})^2$ é usualmente um estimador para σ^2. ■

■ **Example 19.4** Seja uma amostra $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)$ de v.a.'s i.i.d. com distribuição Poisson(θ). No caso da distribuição de Poisson, temos não apenas $\mathbb{E}(Y_i) = \theta$ mas também $\text{Var}(Y_i) = \theta$. Seja $\overline{Y} = \frac{1}{n}\sum_i Y_i$, a média aritmética das v.a.'s. Como $\overline{Y} \approx \mathbb{E}(Y_i)$ quando o tamanho amostral n é grande, usamos $\overline{Y}$ como estimador de θ no caso da Poisson.

O caso curioso aqui é que $\overline{Y}$ é também um estimador da variância das v.a.'s já que $\text{Var}(Y_i)$ é também igual a θ. Mais curioso ainda, como vimos no capítulo **??**, pela Lei dos Grandes Números, a variância amostral $S^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})^2$ converge para a variância populacional se tivermos v.a.'s i.i.d. Assim, $S^2 \approx \text{Var}(Y_i) = \theta$ e portanto S^2 serve como estimador tanto para a média populacional quanto para a variância popuacional no caso de dados Poisson. Na verdade, esperamos que $\overline{Y} \approx S^2$ se n é grande no caso de uma Poisson. Esta ideia de que a razão $S^2/\overline{Y}$ deveria ser aproximadamente igual a 1 é explorada em alguns métodos para testar se os dados seguem de fato uma distribuição de Poisson. ■

■ **Example 19.5** Seja uma amostra $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)$ de v.a.'s i.i.d. sem que especifiquemos uma classe de densidades ou modelo seguido pelso dados. Sejam $\mathbb{E}(Y_i) = \mu$ e $\text{Var}(Y_i) = \sigma^2$. Então $\overline{Y} = \frac{1}{n}\sum_i Y_i$ é um estimador intuitivamente razoável para μ pois, pela Lei dos Grandes Números, esperamos $\overline{Y} \approx \mu$. Pela mesma razão, a variância amostral $S^2 \approx \sigma^2$ e portanto S^2 é um estimador intuitivamente razoável para σ^2. ■

■ **Example 19.6 — Estimador pontual como vetor.** Considere uma amostra $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)$ de v.a.'s i.i.d. gaussianas $N(\mu, \sigma^2)$. Seja que $\theta = (\mu, \sigma^2)$. É comum usarmos o vetor bi-dimensional

$$\mathbf{T} = g(\mathbf{Y}) = \left(\overline{Y}, S^2\right)$$

para fazer inferência sobre θ. θy é uma estatística bi-dimensional já que cada entrada de θy é uma função dos dados $\mathbf{Y}$. A primeira entrada do vetor θy é a média aritmética $\sum_i Y_i/n$ e ela é usada para inferir sobre o valor desconhecido de μ. A segunda entrada é uma medida empírica da variação dos dados em torno de $\overline{Y}_n$ e ela é usada para inferir sobre o valor de $\sigma^2 = \mathbb{E}(Y - \mu)^2$.

Quando algo é um estimador?

Definição de estimador permite que QUALQUER estatística $g(\mathbf{Y})$ seja estimador de θ. Mas então podemos usar $\overline{Y}$ ou $W = \max\{Y_1, \ldots, Y_n\}$ como estimadores de $\sigma^2 = \mathbb{V}(Y_i)$? Podemos mas não devemos. Vamos ver que $\overline{Y}$ e W tem propriedades muito ruins como estimadores de σ^2. Podemos facilmente encontrar estimadores de σ^2, tais como $S^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})^2$, que são muito

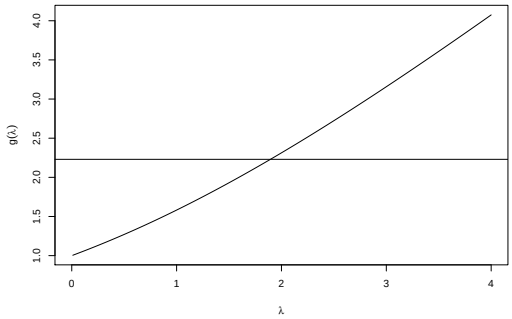


Figure 19.5: Gráfico da função $g(\lambda) = \lambda/(1-e^{-\lambda})$. A linha horizontal corresponde à média aritmética $\bar{y} = 2.23$ dos dados amostrais. O estimador de λ é o valor $\hat{\lambda}$ tal que $g(\lambda) = \bar{y}$. Podemos ver que $\hat{\lambda} \approx 1.95$.

melhores que $\overline{Y}$ ou $W = \max\{Y_1, \ldots, Y_n\}$.

Outros exemplos de estimadores pontuais

$\mathbf{T}(\mathbf{y}) = g(\mathbf{Y}) = (\overline{Y}, (Y_{(n)} - Y_{(1)})/2)$, a média e metade da variação total (range) da amostra $\mathbf{T}(\mathbf{y}) = g(\mathbf{Y}) = \widehat{F}_n(3.2) = \#\{Y_i;\ Y_i \leq 3.2\}/n$ onde $\#A$ é o número de elementos (ou cardinalidade) do conjunto A. Isto é, $\widehat{F}_n(3.2)$ é a proporção de elementos da amostra que são menores ou iguais a 3.2. Poderíamos substituir o ponto $x = 3.2$ por qualquer outro no exemplo acima.

Estimadores não-intuitivos

Alguns estimadores possuem fórmulas matemáticas complicadas e não-intuitivas. Por exemplo, para variáveis aleatórias Y_i positivas (isto é, com $P(Y_i > 0) = 1$) podemos definir a estatística $\mathbf{T}(\mathbf{y}) = g(\mathbf{Y}) = \frac{n}{\sum_{i=1}^{n} \log Y_i}$. Esta estatística estranha é o MLE de um parâmetro θ em um certo modelo estatístico (v.a.'s i.i.d. com distribuição Pareto ou power-law).

Estimadores não-intuitivos

Em outro problema, o método de máxima verossimilhança leva ao seguinte estimador: $T = T(\mathbf{Y})$ é o valor de λ que satisfaz à seguinte restrição:

$$\frac{\lambda}{1-e^{-\lambda}} = \bar{Y}$$

A solução desta equação não-linear deve ser encontrada numericamente. A solução é função dos dados através de $\bar{Y}$ no lado direito.

Estimadores não-intuitivos

Estimadores pontuais em regressão linear

No modelo de regressão linear múltipla, temos v.a.'s $Y_1, \ldots, Y_n$ que são independentes mas não são i.d.

Para a i-ésima observação, temos o modelo:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + \varepsilon_i = \mathbf{x}_i' \beta + \varepsilon_i$$

$\mathbf{x} = (1, x_{i1}, \ldots, x_{ik})^t$ é o vetor $(k+1) \times 1$ com as covariáveis (ou features) associadas com a observação i.

Os erros $\varepsilon_1, \ldots, \varepsilon_n$ são i.i.d. seguindo uma gaussiana $N(0, \sigma^2)$.

Os erros ε_i NÃO SÃO observados diretamente. Observamos apenas Y_i e as covariáveis em $\mathbf{x}_i$.

O modelo implica que $Y_i \sim N(\mathbf{x}_i' \beta, \sigma^2)$ e as v.a.'s são independentes.

Estimadores pontuais em regressão linear

A versão matricial do modelo de regressão linear é

$$\mathbf{Y} = \mathbf{x}\,\beta + \varepsilon$$

onde $\mathbf{Y}$ é vetor $n \times 1$, a matriz de desenho $\mathbf{x}$ é de dimensão $n \times (k+1)$ com k covariáveis e a colunas de 1's. O vetor $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)$ de dimensão $n \times 1$ é composto de v.a.'s i.i.d. $N(0, \sigma^2)$. O parâmetro θ é $\theta = (\beta, \sigma^2) = (\beta_0, \ldots, \beta_p, \sigma^2)$

Queremos estimar β e σ^2.

MLE de β

O MLE $\widehat{\beta}$ de $\beta = (\beta_0, \ldots, \beta_p)$ coincide com o estimador de mínimos quadrados. $\widehat{\beta}$ é uma função do vetor de dados aleatórios $\mathbf{Y}$ e da matriz de regressores $\mathbf{x}$ (que é considerada uma matriz de constantes conhecida). Temos o vetor $(k+1) \times 1$

$$\widehat{\beta} = (\widehat{\beta_0}, \ldots, \widehat{\beta_p})' = (\mathbf{x}^t\mathbf{x})^{-1}\mathbf{x}^t\mathbf{Y}$$

Se escrevermos a matriz $k \times n$ dada por $(\mathbf{x}^t\mathbf{x})^{-1}\mathbf{x}^t$ por A podemos ver que $\widehat{\beta} = A\mathbf{Y}$. Assim, cada elemento de $\widehat{\beta}$ é uma combinação linear dos elementos do vetor $\mathbf{Y}$.

MLE de σ^2

O modelo prediz ou estima o valor de Y_i usando $\widehat{\beta}$. O vetor $n \times 1$ com os valores preditos pelo modelo para os valores realmente observados $\mathbf{Y}$ são dados por

$$\widehat{\mathbf{Y}} = \mathbf{x}\,\beta = \mathbf{x}\,(\mathbf{x}^t\mathbf{x})^{-1}\mathbf{x}^t\mathbf{Y}$$

A diferença entre $\mathbf{Y}$ e a predição $\widehat{\mathbf{Y}}$ forma o vetor de resíduos ou vetor de erros de predição $\mathbf{Y} - \widehat{\mathbf{Y}}$. O MLE de σ^2 é dado pela média dos resíduos ao quadrado:

$$\widehat{\sigma^2} = \frac{1}{n}||\mathbf{Y} - \widehat{\mathbf{Y}}||^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{Y_i})^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \mathbf{x}_i'\,\widehat{\beta})^2$$

Estimadores pontuais em regressão logística

Na regressão logística, onde $p_i = p_i(\theta) = 1/(1+\exp(-\mathbf{x}_i^t\theta))$, o MLE é a solução $\hat{\theta}$ do sistema de equações não-lineares

$$\mathbf{x}\mathbf{Y} = \mathbf{x}\mathbf{p}(\theta)$$

onde $\mathbf{p}(\theta) = (p_1(\theta), p_2(\theta), \ldots, p_n(\theta))$

Embora não exista uma expressão analítica, uma fórmula, para o MLE, podemos ver que a solução vai depender apenas de $\mathbf{x}$ e de $\mathbf{Y}$. Assim, como em regressão múltipla, $\hat{\theta}$ é função dos dados aleatórios binários $\mathbf{Y}$ e da matriz de regressores (ou constantes) $\mathbf{x}$ (embora não possamos escrever explicitamente esta função).

$T(\mathbf{Y})$ é v.a.

Vamos escrever $T(\mathbf{Y})$ ou simplesmente T. Conceito CRUCIAL: $T(\mathbf{Y})$ é uma v.a. Exemplo: $Y_1, \ldots, Y_n$ i.i.d. $N(\mu, \sigma^2)$. Considere $\overline{Y} = (Y_1 + \ldots + Y_n)/n$ é um estimador natural para μ. $\overline{Y}$ é uma v.a!!! Até sabemos qual é a sua distribuição de probabilidade a partir das propriedades de combinação linear de uma normal multivariada:

$$\overline{Y} = \frac{1}{n}\,(1, 1, \ldots 1)^t\mathbf{Y} \sim N(\mu, \sigma^2/n)$$

onde $\mathbf{Y} = (Y_1, \ldots, Y_n)$ é normal multivariada com vetor esperado $(\mu, \mu, \ldots, \mu)$ e matriz de covariância $\sigma^2 I_n$ onde I_n é a matriz identidade.

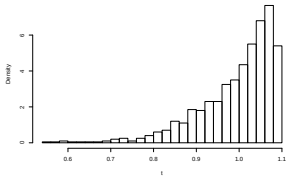


Figure 19.6: Histograma dos 1000 valores de $T = 12/11Y_{(11)}$

$T(\mathbf{Y})$ e $T(\mathbf{Y})$

Devemos distinguir entre a estatística ou estimador $T(\mathbf{Y})$ (que é uma v.a.) e o seu valor observado num conjunto específico de instâncias (que é um número específico). Uma amostra de tamanho 4: $y_1 = 1.6$, $y_2 = 1.8$, $y_3 = 1.5$, $y_4 = 1.8$. Estes são os valores observados das v.a.'s Y_1, Y_2, Y_3, Y_4 nesta amostra particular. O valor observado da v.a. $\overline{Y}$ é o valor $\overline{y} = 1.675$. Note que $\overline{y} = 1.675$ não é uma v.a.: o número 1.675 não possui uma lista de valores possíveis e probabilidades associadas. $\overline{y} = 1.675$ é um dos valores possíveis da v.a. $\overline{Y}$, um valor especial: aquele que calhou de ocorrer na amostra que temos à mão.

$T(\mathbf{Y})$ e $T(\mathbf{Y})$

$\mathbf{Y} = (Y_1, \ldots, Y_{11})$ é vetor com 11 variáveis aleatórias i.i.d com distribuição comum $U[0, \theta]$. Suponha que o verdadeiro valor de θ é 1 mas *isto é desconhecido pelo usuário*. Deseja-se estimar θ. Estimador de θ: $T = 12\max\{Y_1, \ldots, Y_{11}\}/11 = 12Y_{(11)}/11$. Explicação: $\max\{Y_1, \ldots, Y_{11}\}$ deve estar próximo, mas abaixo, do maior valor possível, que é θ. Uma maneira de obter um estimador de θ seria incrementar um pouco o máximo multiplicando por alguma constante maior que 1. A fração $12/11$ faz com que T seja ligeiramente maior que o máximo $Y_{(n)}$. Veremos que isto torna o estimador T não-viciado para estimar θ (lista de exercícios).

$T(\mathbf{Y})$ e $T(\mathbf{Y})$

Uma amostra particular $\mathbf{Y}_1$: obtem $t_1 = g(\mathbf{Y}_1) = 1.0437$. Com outra amostra particular (um segundo dia) $\mathbf{Y}_2$ obtemos $t_2 = g(\mathbf{x}_2) = 0.9845$. Repetindo independentemente 1000 vezes. Geramos mil vetores $\mathbf{Y}_1, \ldots, \mathbf{Y}_{1000}$, cada um deles de tamanho $n = 11$. Em cada uma destas 1000 amostras, calculamos 1000 valores $t_1, t_2, \ldots, t_{1000}$. Isto é, obtemos uma amostra de 1000 valores i.i.d. de T.

$T(\mathbf{Y})$ e $T(\mathbf{Y})$

Com esta amostra de 1000 valores do ESTIMADOR T podemos ter uma idéia da distribuição de probabilidade da v.a. T fazendo um histograma:

Aproximadamente 1% dos valores observados de T caíram abaixo de 0.70: bad days.

O drama da realidade

O drama do usuário é que, na prática, ele provavelmente fará apenas uma única estimação, num único dia específico. Ele fará isto usando uma única amostra de 11 dados nos quais deve se basear para estimar o valor desconhecido de θ. Por isto, ele nunca saberá, nesse dia da estimação, qual o tamanho do erro de estimação que ele está cometendo.

19.4 Critérios para escolher estimadores

19.5 Escolhendo um estimador

Candidatos a estimar θ

Seja $\mathbb{E}(Y) = \mu$ e uma amostra $Y_1, \ldots, Y_n$ Por que não considerar o estimador mediana amostral? Ordene os dados: $X_{(1)} < X_{(2)} < \ldots < X_{(n)}$. Então

$$M = \begin{cases} X_{(k+1)}, & \text{caso } n \text{ seja ímpar, } n = 2k+1 \\ (X_{(k)} + X_{(k+1)})/2, & \text{caso } n \text{ seja par, } n = 2k \end{cases}$$

Ou que sabe a média do primeiro e do terceiro quartil? Como escolher um deles? qual o critério de escolha?

19.6 Decompostion of the MSE: Bias and variance

Mais candidatos a estimar θ

Ou então uma média ponderada entre a média aritmética $\overline{X}_n$ e a mediana:

$$T = wM + (1-w)\overline{X}_n$$

onde $0 \leq w \leq 1$. Como w varia continuamente entre 0 e 1, teremos infinitos possíveis estimadores deste tipo. Ou quem sabe uma média ponderada entre $\overline{X}_n$, a mediana e a média dos quartis? Como escolher um deles? qual o critério de escolha? Obviamente algum critério que faça referência ao tamanho dos erros de estimação, mas qual é esse critério?

Erro de estimação

O erro de estimação que vamos cometer é a *variável aleatória* $T(\mathbf{Y}) - \theta$. Como v.a., ela possui duas coisas: uma "lista" de valores possíveis e uma "lista" de probabilidades associadas. Na prática, como não conhecemos o valor de θ, nunca saberemos o valor do erro de estimação que cometemos em cada caso particular. Apesar disso, podemos conhecer as propriedades estatísticas do erro de estimação. Podemos saber como erraremos em geral, embora não possamos saber como erramos em cada caso particular.

Vício

$T(\mathbf{Y})$ é um estimador vetorial de dimensão k de uma característica vetorial θ de dimensão k da população. Definição: A diferença vetorial $\mathbb{E}(T(\mathbf{Y})) - \theta$ é chamada de vício do estimador (para estimar θ) e é denotada por $b(\theta)$. Quando $b(\theta) = \mathbf{0}$, dizemos que o estimador é não viciado para estimar o parâmetro θ. Um estimador não-viciado tem sua distribuição centrada em torno do parâmetro θ que desejamos estimar. Ocasionalmente ele vai subestimar ou superestimar θ. Mas ele nem subestima nem superestima *sistematicamente*. Um estimador não-viciado é um estimador *acurado*.

Vício:exemplos

Seja $Y_1, \ldots, Y_n$ uma amostra de variáveis i.i.d. com *qualquer* distribuição. Suponha que $\mathbb{E}(Y_i) = \mu$. Então $\overline{Y}_n$ é um estimador não-viciado para estimar μ.

Se os n dados tiverem distribuição i.i.d. normal, a mediana amostral M é não-viciada para estimar μ se n é ímpar e é ligeiramente viciada se n é par.

Se $w \in (0,1)$, então $w\overline{Y} + (1-w)M$ é não-viciado para estimar μ

Vício:exemplos

Seja $Y_1, \ldots, Y_n$ uma amostra de variáveis i.i.d. com *qualquer* distribuição. Suponha que $\text{Var}(Y_i) = \sigma^2$. Então S^2 é um estimador não-viciado para estimar σ^2 onde

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(Y_i - \overline{Y}_n\right)^2$$

A prova é simples e fica para a lista de exercícios. Esta é a razão para ter o denominador $n-1$ no estimador acima: torná-lo não-viciado para estimar σ^2

Vício

O estimador

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \overline{Y}_n\right)^2$$

é viciado

Veja que $\widehat{\sigma}^2 = (n-1)/nS^2$

O vício de $\widehat{\sigma}^2$ é dado por

$$b(\sigma^2) = \mathbb{E}(\widehat{\sigma}^2) - \sigma^2 = \left(\frac{n-1}{n} - 1\right)\sigma^2 = -\frac{\sigma^2}{n}$$

Vício $b(\sigma^2)$ tende a zero quando a amostra cresce.

Vício: exemplo

Tempo de vida $T_1, \ldots, T_n$ são iid com distribuição $\exp(\lambda)$. Observar tempos de vida até um tempo máximo C. Quando $T_i > C$ anotamos simplesmente C. O verdadeiro valor de T_i nestes casos não é observado Dizemos que eles são censurados. Estimar $\mathbb{E}(T_i) = \mu$ usando a média aritmética dos T_i's **registrados** é um estimador viciado (ele subestima μ).

Vício: exemplo

$\mathbf{Y} = (Y_1, \ldots, Y_n)$ são n variáveis aleatórias i.i.d com distribuição comum $U[0,\theta]$. Suponha que o verdadeiro valor de θ é desconhecido e que desejamos estimar θ. Estimador: $T = ((n+1)/n) \times \max\{Y_1, \ldots, Y_n\}$. T é não-viciado para estimar θ pois $\mathbb{E}(T) = \theta$. A constante $(n+1)/n$ é a quantidade perfeita para incrementar o máximo $\max\{Y_1, \ldots, Y_n\}$ e "empurrá-lo" para ficar em torno de θ.

MSE

$T = T(\mathbf{Y})$ é estimador de θ *Definição: MSE* $= \mathbb{E}(T-\theta)^2$ é chamado de erro quadrático médio de estimação (*mean squared error*, em inglês). *Definição:* $\mathbb{E}(|T-\theta|)$ é chamado de erro absoluto médio de estimação. O erro absoluto é mais intuitivo mas temos mais resultados usando o MSE.

Decomposição do MSE

Teorema: Se $T = T(\mathbf{Y})$ é estimador de θ então

$$\mathbb{E}(T-\theta)^2 = \mathrm{Var}(T) + b(\theta)^2 \tag{19.1}$$

Prova: Seja $\mathbb{E}(T) = \mu$. Some e subtraia μ:

$$\begin{aligned}
\mathbb{E}(T-\theta)^2 &= \mathbb{E}(T-\mu+\mu-\theta)^2 \\
&= \mathbb{E}\left[(T-\mu)^2 + 2(T-\mu)(\mu-\theta) + (\mu-\theta)^2\right]^2 \\
&= \mathbb{E}\left[(T-\mu)^2\right] + 2\mathbb{E}\left[(T-\mu)(\mu-\theta)\right] + \mathbb{E}\left[(\mu-\theta)^2\right] \\
&= \mathrm{Var}(T) + 2(\mu-\theta)\mathbb{E}(T-\mu) + (\mu-\theta)^2
\end{aligned}$$

Para terminar: $\mathbb{E}(T-\mu) = 0$ pois $\mathbb{E}(T) = \mu$ e assim

$$\mathbb{E}(T-\theta)^2 = \mathrm{Var}(T) + (\mu-\theta)^2 = \mathrm{Var}(T) + b^2(\theta)\,.$$

Decomposição do MSE

Esta decomposição é de grande utilidade por duas razões. 1) ela permite quebrar o problema de encontrar um bom estimador (um com erro de estimação tipicamente pequeno) em dois sub-problemas. Devemos encontrar um estimador que possua um vício pequeno e, ao mesmo tempo, variância pequena. 2) Por outro lado, ele fornece um critério simples para encontrar bons estimadores. Se o vício de dois estimadores é zero, o *MSE* deles é igual à suas variâncias e assim basta escolher aquele estimador que possui a menor variância. Este estimador não-viciado e de menor variância terá o menor *MSE*.

Ilustração

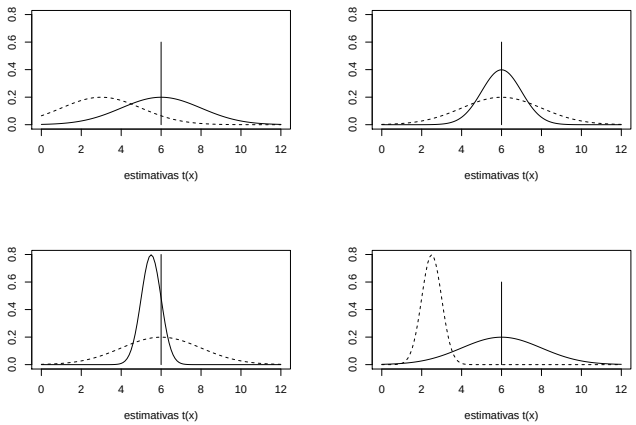


Figure 19.7: Densidades de dois estimadores T_1 e T_2 de um mesmo parâmetro θ. Nestes exemplos, o valor de θ é 6.

19.7 Consistency of estimators

19.7.1 Nearest Neighbor estimators are consistent

19.7.2 Classification trees may be consistent

Standard classification trees, such as those generated by the CART algorithm, are not consistent in general. That is, as the number of training samples $n \to \infty$, the classifier may not converge to the Bayes optimal classifier, nor provide accurate estimates of $\mathbb{P}(Y = 1 \mid X = x)$.

However, under specific regularity conditions, tree-based classification rules can be consistent. The general idea, rooted in nonparametric classification theory, is that if the partition induced by the tree becomes increasingly fine with n, and each region contains more and more data points, then the classifier can approximate the true decision boundary.

A common sufficient condition for consistency is the following:

- The **diameter** of the partition regions (i.e., the leaves) tends to zero as $n \to \infty$.
- The **number of samples per leaf** tends to infinity.

Under these conditions, a tree classifier behaves similarly to a local averaging estimator, such as k-NN, and can consistently estimate the conditional probability.

Moreover, ensemble methods such as Random Forests and Boosted Trees have been shown to be consistent under mild assumptions:

- **Random Forests** were proven to be consistent by Scornet, Biau, and Vert (2015), assuming certain randomness and subsampling conditions.
- **Boosted Trees** can also be consistent if regularization (e.g., shrinkage) and early stopping are applied.

For a general theoretical treatment, see:

- Devroye, Györfi, and Lugosi (1996). *A Probabilistic Theory of Pattern Recognition*. Springer.
- Scornet, Biau, and Vert (2015). *Consistency of Random Forests*. Annals of Statistics, 43(4), 1716–1741.
- Breiman, L. (2001). *Random Forests*. Machine Learning, 45(1), 5–32.

19.7.3 Boosting is consistent

Boosting methods, such as AdaBoost and gradient boosting, can provide consistent estimators of the conditional class probability $\mathbb{P}(Y = 1 \mid X = x)$ under appropriate conditions.

Theoretically, boosting can be interpreted as functional gradient descent in a suitable loss function space, typically minimizing an exponential or logistic loss. When the number of boosting rounds increases with the sample size (but not too quickly), and the base learners are sufficiently expressive, consistency can be achieved.

A typical set of sufficient conditions for the consistency of boosting includes:

- The base classifier is rich enough to approximate the Bayes rule.
- The number of boosting iterations grows with the sample size (commonly with early stopping).
- A small learning rate (shrinkage) is used to avoid overfitting.

Under such conditions, the boosting estimator converges to the Bayes classifier, and probability estimates can be obtained by applying a logistic transformation to the aggregate prediction function:

$$\hat{p}_n(x) = \frac{1}{1 + e^{-2F_n(x)}},$$

where $F_n(x)$ is the ensemble function learned by the boosting procedure.

For theoretical guarantees, see:

- Bartlett, P. L., and Traskin, M. (2007). *Boosting with the Logistic Loss: Consistency and Convergence*. Journal of Machine Learning Research.
- Zhang, T., and Yu, B. (2005). *Boosting with Early Stopping: Convergence and Consistency*. Annals of Statistics.
- Schapire, R. E., and Freund, Y. (2012). *Boosting: Foundations and Algorithms*. MIT Press.

20. MLE Optimality

20.1 MLE Optimality for uni-dimensional θ

- Considere a classe $\mathscr{C}$ de *todos* os estimadores não-viciados de θ.
- Por exemplo, se os dados $Y_1, \ldots, Y_n$ forem uma amostra aleatória de uma $N(\mu, \sigma^2)$ e se $n = 2k+1$ é um ímpar, então:
  - a média amostral $\overline{Y}_n$ é não-viciada para estimar μ e portanto pertence a $\mathscr{C}$
  - a mediana amostral $M = Y_{(r+1)}$ (a estatística de ordem $r+1$) é não-viciada para estimar μ e portanto pertence a $\mathscr{C}$
  - Se $w \in (0,1)$, qualquer combinação linear da forma $w\overline{Y}_n + (1-w)M$ é não-viciada para estimar μ e portanto também pertence a $\mathscr{C}$
  - Existem infinitos outros estimadores não viciados de μ que pertencerão à classe $\mathscr{C}$
- Estratégia: procurar dentre os estimadores não-viciados em $\mathscr{C}$ por um estimador que tenha a variância mínima: estimador *ótimo* para θ na classe dos estimadores não-viciados.
- Como podemos saber que um estimador tem variância mínima?
- Usando a Desigualdade da Informação (Cramér-Rao).
- Fixado o tamanho da amostra n,
  - existe um limite na variância de qualquer $\hat{\theta}$ não-viciado.
  - É a cota inferior de Cramér-Rao.
  - Isto fornece um limite inferior para a precisão (ou MSE) de um estimador não-viaciado de θ.
  - NADA pode ser mais preciso que esta cota de Cramér-Rao (dentre os não–viciados).
- Sejam $Y_1, \ldots, Y_n$ variáveis i.i.d. com densidade conjunta $f(\mathbf{Y}; \theta)$.
- Se $\hat{\theta}$ é qualquer estimador não viciado de θ, então

$$Var(\hat{\theta}) \geq \frac{1}{I(\theta)}.$$

onde $I(\theta)$ é a Informação de Fisher e é dada por

$$I(\theta) = E\left[\frac{\partial}{\partial\theta}\log f(\mathbf{Y};\theta)\right]^2.$$

Exemplo:

- Suponha que $Y_1, \ldots, Y_n$ são i.i.d. Poisson(θ).
- Então

$$p(\mathbf{y};\theta) = \prod_{i=1}^{n} \frac{\theta^{y_i} e^{-\theta}}{y_i!} = \frac{\theta^{\sum_i y_i} e^{-n\theta}}{\prod_{i=1}^n y_i!} .$$

- Portanto

$$\log p(\mathbf{y};\theta) = \left(\sum_{i=1}^{n} y_i\right) \log\theta - n\theta - \log\left(\prod_{i=1}^{n} y_i!\right) .$$

- Derivando com relação a θ temos que

$$\frac{\partial \ell}{\partial \theta} = \frac{\partial \log p(\mathbf{y};\theta)}{\partial \theta} = \frac{\sum_{i=1}^n y_i}{\theta} - n$$

- A quantidade $\dfrac{\partial \ell}{\partial \theta}$ é muito importante e é chamada de função escore (ou *score function*, em inglês).

Exemplo: (continuação)

- No caso de v.a.'s i.i.d. Poisson(θ), a função escore é

$$\frac{\partial \ell}{\partial \theta} = \frac{\sum_{i=1}^n y_i}{\theta} - n$$

- Esta função depende dos dados observados.
- Por exemplo, se $n = 4$ e $\mathbf{Y} = (3,1,0,3)$ então

$$\frac{\partial \ell}{\partial \theta} = \frac{\sum_{i=1}^n y_i}{\theta} - n = \frac{3+1+0+3}{\theta} - 4 = \frac{7}{\theta} - 4$$

- Note que $\partial\ell/\partial\theta$ é uma função de θ.
- É esta função que usamos para obter o MLE ao igualar o escore a zero e resolver para θ:

$$0 = \frac{\partial \ell}{\partial \theta} = \frac{7}{\theta} - 4$$

Exemplo: (continuação)

- Neste exemplo, com $n = 4$ e $\mathbf{Y} = (3,1,0,3)$ o escore

$$\frac{\partial \ell}{\partial \theta} = \frac{\sum_{i=1}^n y_i}{\theta} - n = \frac{7}{\theta} - 4$$

  é uma função matemática de θ, não é uma variável aleatória.
- Os dados $\mathbf{Y} = (3,1,0,3)$ são considerados fixos, são as instâncias observadas no experimento.
- Entretanto, para estudar as propriedades do MLE, vamos transformar este escore numa VARIÁVEL ALEATÓRIA.
- Para isto, vamos substituir o vetor de instâncias $\mathbf{Y} = (3,1,0,3)$ pelas variáveis aleatórias $\mathbf{Y} = (Y_1, Y_2, Y_3, Y_4)$:

$$\frac{\partial \log p(\mathbf{Y};\theta)}{\partial \theta} = \frac{\sum_{i=1}^4 Y_i}{\theta} - 4$$

- O que mudou? O escore $\partial\ell/\partial\theta$ é agora uma v.a.: possui lista de valores possíveis e probabilidades associadas.

Exemplo: (continuação)

- Vamos entender o que é o escore como v.a.

$$\frac{\partial \ell}{\partial \theta} = \frac{\partial \log p(\mathbf{Y};\theta)}{\partial \theta} = \frac{Y_1+Y_2+Y_3+Y_4}{\theta} - 4$$

- O que torna esta expressão uma v.a. é a presença da soma das v.a.'s Y_i no numerador.
- Resultado de probabilidade: Se $Y_1,\ldots Y_n$ são independentes com distribuição Poisson(λ_i) então a sua soma é uma outra v.a. Poisson(λ) com valor esperado $\lambda = \lambda_1 + \ldots \lambda_n$.
- Note a presença da soma $Y_1+Y_2+Y_3+Y_4$ no numerador do escore: esta soma é uma v.a. com distribuição Poisson(4θ).

Exemplo: (continuação)

- Assim, o escore $\partial\ell/\partial\theta$ tem uma distribuição associada com uma Poisson(4θ):

$$\frac{\partial \ell}{\partial \theta} = \frac{Y_1+Y_2+Y_3+Y_4}{\theta} - 4 \sim \frac{\text{Poisson}(4\theta)}{\theta} - 4$$

- Os valores possíveis e probabilidades associadas de $\partial\ell/\partial\theta$ são:

valores	$\frac{0}{\theta}-4$	$\frac{1}{\theta}-4$	$\frac{2}{\theta}-4$	$\frac{3}{\theta}-4$	$\ldots$
probabs	$e^{-4\theta}$	$e^{-4\theta}4\theta$	$e^{-4\theta}(4\theta)^2/2$	$e^{-4\theta}(4\theta)^3/3!$	$\ldots$

- As probabilidades são obtidas a partir da fórmula das probabilidades de uma Poisson(4θ).

Exemplo: (continuação)

- Assim, transformamos o escore numa v.a. substituindo as instâncias $\mathbf{Y}$ pelas v.a.'s $\mathbf{Y}$

$$\frac{\partial \ell}{\partial \theta} = \frac{\partial \log p(\mathbf{Y};\theta)}{\partial \theta} = \frac{Y_1+Y_2+Y_3+Y_4}{\theta} - 4$$

- Sendo agora uma v.a., podemos calcular sua esperança e sua variância.
- Por exemplo, a esperança da função escore:

$$\begin{aligned}
\mathbb{E}\left(\frac{\partial \ell}{\partial \theta}\right) &= \mathbb{E}\left(\frac{Y_1+Y_2+Y_3+Y_4}{\theta} - 4\right) \\
&= \frac{\mathbb{E}(Y_1+Y_2+Y_3+Y_4)}{\theta} - 4 \\
&= \frac{\mathbb{E}(Y_1)+\mathbb{E}(Y_2)+\mathbb{E}(Y_3)+\mathbb{E}(Y_4)}{\theta} - 4 \\
&= \frac{\theta+\theta+\theta+\theta}{\theta} - 4 = 0
\end{aligned}$$

Exemplo: (continuação)

- Mais importante é a variância do escore.
- Para qualquer v.a. X temos $\mathbb{V}(X) = \mathbb{E}(X^2) - [\mathbb{E}(X)]^2$
- Assim, como a esperança do escore é igual a zero, temos

$$\mathbb{V}\left(\frac{\partial \ell}{\partial \theta}\right) = \mathbb{E}\left[\left(\frac{\partial \ell}{\partial \theta}\right)^2\right] + \left[\mathbb{E}\left(\frac{\partial \ell}{\partial \theta}\right)\right]^2 = \mathbb{E}\left[\left(\frac{\partial \ell}{\partial \theta}\right)^2\right]$$

- Assim, usando a definição de $I(\theta)$, temos:

$$\begin{aligned}
I(\theta) &= \mathbb{E}\left[\left(\frac{\partial}{\partial \theta}\log f(\mathbf{Y};\theta)\right)^2\right] = \mathbb{E}\left[\left(\frac{\partial \ell}{\partial \theta}\right)^2\right] \\
&= \mathbb{V}\left(\frac{\partial \ell}{\partial \theta}\right) = \mathbb{V}\left(\frac{\text{Poisson}(4\theta)}{\theta} - 4\right) \\
&= \frac{\mathbb{V}(\text{Poisson}(4\theta))}{\theta^2} = \frac{4\theta}{\theta^2} = \frac{4}{\theta}
\end{aligned}$$

- Pela desigualdade de Cramér-Rao, se $\hat{\theta}$ é não viciado para estimar θ numa amostra de tamanho $n = 4$ de v.a.'s i.i.d. Poisson(θ), então

$$MSE(\hat{\theta}) = \mathbb{V}(\hat{\theta}) \geq \frac{\theta}{n}.$$

- Considere o estimador $\hat{\theta} = \overline{Y}$.
- Faça as contas para verificar que $\overline{Y}$ é não-viciado para θ. Além disso,

$$MSE(\overline{Y}) = \mathbb{V}(\overline{Y}) = \frac{\theta}{n} = \frac{1}{I(\theta)}.$$

- Assim, se as v.a.'s são i.i.d. Poisson(θ), ninguém pode ser melhor do que o bom e velho $\overline{Y}$ para estimar θ na classe dos estimadores não-viciados.
- É muito útil recordar uma fórmula de probabilidade.
- Seja $\mathbf{Y} = (Y_1, \ldots, Y_n)$ um vetor aleatório com densidade de probabilidade $f(\mathbf{Y}) = f(y_1, \ldots, y_n)$.
- Seja $g(\mathbf{Y})$ uma nova v.a. obtida através de uma função matemática qualquer aplicada ao vetor $\mathbf{Y}$.
- Por exemplo, $g(\mathbf{Y})$ poderia ser $g(\mathbf{Y}) = \overline{Y}$ ou $g(\mathbf{Y}) = \sum 1/\log(Y_i) - \pi^2$
- Como calcular a esperança desta nova v.a. $g(\mathbf{Y})$?
- Temos

$$\mathbb{E}(g(\mathbf{Y})) = \int \cdots \int g(\mathbf{Y})\, f(\mathbf{y}) d\mathbf{y}$$

  onde a integral é tomada sobre todos os valores possíveis do vetor $\mathbf{Y}$.
- Por exemplo, se $g(\mathbf{Y}) = \sum 1/\log(Y_i) - \pi^2$ então

$$\begin{aligned}
\mathbb{E}(g(\mathbf{Y})) &= \mathbb{E}\left(\sum_i \frac{1}{\log(Y_i)} - \pi^2\right) \\
&= \int \cdots \int \left(\sum_i \frac{1}{\log(y_i)} - \pi^2\right) f(\mathbf{y}) d\mathbf{y} \\
&= \int \cdots \int \left(\sum_i \frac{1}{\log(y_i)} - \pi^2\right) f(y_1, \ldots, y_n) d\mathbf{y}
\end{aligned}$$

- Não se preocupe. Não teremos de calcular esta integral explicitamente...
- Vamos considerar três lemas auxiliares para provar a desigualdade de Cramér-Rao.
- No caso particular de uma amostra de v.a.'s i.i.d. Poisson, verificamos que a esperança do escore é zero.
- Isto é verdade em qualquer modelo estatístico, não apenas neste exemplo particular.
- Este é o primeiro lema: a esperança da função escore é sempre igual a zero.

$$\mathbb{E}\left[\frac{\partial \ell}{\partial \theta}\right] = \mathbb{E}\left[\frac{\partial}{\partial \theta} \log f(\mathbf{Y}; \theta)\right] = 0$$

Prova:

- Como $f(\mathbf{y}; \theta)$ é uma densidade de probabilidade, sua integral sobre todos os valores possíveis de $\mathbf{Y}$ é igual a 1:

$$1 = \int \cdots \int f(\mathbf{y}; \theta) d\mathbf{y}$$

  Derivando com respeito a θ dos dois lados desta igualdade, temos

$$0 = \frac{\partial(1)}{\partial \theta} = \frac{\partial}{\partial \theta} \int \cdots \int f(\mathbf{y}; \theta) d\mathbf{y} = \int \cdots \int \frac{\partial}{\partial \theta} f(\mathbf{y}; \theta) d\mathbf{y}$$

- Multiplicando e dividindo o integrando por $f(\mathbf{y};\theta)$ obtemos

$$\begin{aligned} 0 &= \int\cdots\int \frac{\partial}{\partial\theta} f(\mathbf{y};\theta)d\mathbf{y} = \int\cdots\int \frac{\frac{\partial}{\partial\theta}f(\mathbf{y};\theta)}{f(\mathbf{y};\theta)} f(\mathbf{y};\theta)d\mathbf{y} \\ &= \int\cdots\int \frac{\partial\ell}{\partial\theta}\, f(\mathbf{y};\theta)d\mathbf{y} \end{aligned}$$

já que

$$\frac{\partial\ell}{\partial\theta} = \frac{\partial}{\partial\theta}\log f(\mathbf{y};\theta) = \frac{\frac{\partial}{\partial\theta}f(\mathbf{y};\theta)}{f(\mathbf{y};\theta)}$$

Pela regra de cálculo de $\mathbb{E}(g(\mathbf{Y}))$ em probabilidade temos que

$$\int\cdots\int \frac{\partial\ell}{\partial\theta}\, f(\mathbf{y};\theta)d\mathbf{y} = \mathbb{E}\left[\frac{\partial\ell}{\partial\theta}\right].$$

- Concluímos assim que

$$\mathbb{E}\left[\frac{\partial\ell}{\partial\theta}\right] = \mathbb{E}\left[\frac{\partial}{\partial\theta}\log f(\mathbf{Y};\theta)\right] = 0\,.$$

- Recordando de probab: Se W e V são v.a.'s então $|\text{Corr}(W,V)| \leq 1$.
- Mas como $\text{Corr}(W,V) = \frac{\text{Cov}(W,V)}{\sqrt{\mathbb{V}(W)}\sqrt{\mathbb{V}(V)}}$, nós temos que

$$\text{Cov}^2(W,V) \leq \mathbb{V}(W)\mathbb{V}(V)$$

Tomando $W = \hat{\theta}$, temos que para qualquer v.a. V,

$$\text{Cov}^2(\hat{\theta},V) \leq \mathbb{V}(\hat{\theta})\mathbb{V}(V)$$

- Rearranjando a ordem, podemos concluir que Para qualquer v.a. V

$$Var(\hat{\theta}) \geq \frac{Cov^2(\hat{\theta},V)}{Var(V)}\,.$$

- Recordando de probab: Se W e V são v.a.'s então $\text{Cov}(W,V) = \mathbb{E}(W\,V) - \mathbb{E}(W)\,\mathbb{E}(V)$.
- Tome $V = \partial\ell/\partial\theta$, o escore e $W = \hat{\theta}$, um estimador QUALQUER de θ (não precisa ser o MLE, é um estimador arbitrário).
- Pelo Lema 1, temos $\mathbb{E}(\partial\ell/\partial\theta) = 0$. Portanto, temos

$$\begin{aligned} \mathbb{V}\left(\frac{\partial\ell}{\partial\theta}\right) &= \mathbb{E}\left[\left(\frac{\partial\ell}{\partial\theta}\right)^2\right] + \left(\mathbb{E}\frac{\partial\ell}{\partial\theta}\right)^2 \\ &= \mathbb{E}\left[\left(\frac{\partial\ell}{\partial\theta}\right)^2\right] + 0 = I(\theta) \end{aligned}$$

e

$$\text{Cov}\left(\hat{\theta},\frac{\partial\ell}{\partial\theta}\right) = \mathbb{E}\left(\hat{\theta}\,\frac{\partial\ell}{\partial\theta}\right) - \mathbb{E}(\hat{\theta})\,E\left(\frac{\partial\ell}{\partial\theta}\right) = \mathbb{E}\left(\hat{\theta}\,\frac{\partial\ell}{\partial\theta}\right)$$

- Pelos três lemas anteriores temos que

$$\mathbb{V}(\hat{\theta}) \geq \frac{\left(\text{Cov}(\hat{\theta},\,\frac{\partial\ell}{\partial\theta})\right)^2}{I(\theta)} = \frac{\left[\mathbb{E}\left(\hat{\theta}\,\frac{\partial\ell}{\partial\theta}\right)\right]^2}{I(\theta)}\,.$$

- Resta apenas mostrar que o numerador é igual a um.
- Como $\hat{\theta}$ é não-viciado para θ temos que

$$\theta = E(\hat{\theta}) = \int \cdots \int \hat{\theta}(\mathbf{y}) f(\mathbf{y};\theta) d\mathbf{y}\,.$$

- Vamos derivar dos dois lados em relação a θ. Pelo lado esquerdo ficamos com

$$\frac{\partial}{\partial \theta}\theta = 1\,.$$

- Pelo lado direito

$$\begin{aligned}
\frac{\partial}{\partial \theta}\int \cdots \int \hat{\theta}(\mathbf{y}) f(\mathbf{y};\theta) d\mathbf{y} &= \int \cdots \int \hat{\theta}(\mathbf{y}) \frac{\partial f(\mathbf{y};\theta)}{\partial \theta} d\mathbf{y} \\
&= \int \cdots \int \hat{\theta}(\mathbf{y}) \frac{\partial f(\mathbf{y};\theta)}{\partial \theta} \frac{f(\mathbf{y};\theta)}{f(\mathbf{y};\theta)}\, d\mathbf{y} \\
&= \int \cdots \int \hat{\theta}(\mathbf{y}) \frac{\partial \log f(\mathbf{y};\theta)}{\partial \theta} f(\mathbf{y};\theta)\, d\mathbf{y} \\
&= \mathbb{E}\left[\hat{\theta}(\mathbf{Y}) \frac{\partial \log f(\mathbf{Y};\theta)}{\partial \theta}\right] \\
&= \mathrm{Cov}\left(\hat{\theta}(\mathbf{Y}), \frac{\partial \log f(\mathbf{Y};\theta)}{\partial \theta}\right)
\end{aligned}$$

- Como os dois lados devem ser iguais, concluímos que o numerador é igual a 1. Isto é, $1 = \mathrm{Cov}\left(\hat{\theta}, \frac{\partial \ell}{\partial \theta}\right)$ e portanto

$$\mathbb{V}(\hat{\theta}) \geq \frac{1}{I(\theta)}\,.$$

Sob condições de regularidade temos que

$$I(\theta) = -\mathbb{E}\left[\frac{\partial^2 \ell}{\partial \theta^2}\right] = -\mathbb{E}\left[\frac{\partial^2}{\partial \theta^2} \log f(\mathbf{Y};\theta)\right]$$

Prova:

$$\begin{aligned}
\frac{\partial^2 \ell}{\partial \theta^2} = \frac{\partial^2}{\partial \theta^2} \log f(\mathbf{y};\theta) &= \frac{\partial}{\partial \theta}\left(\frac{\partial}{\partial \theta} \log f(\mathbf{y};\theta)\right) = \frac{\partial}{\partial \theta}\left(\frac{\frac{\partial}{\partial \theta} f(\mathbf{y};\theta)}{f(\mathbf{y};\theta)}\right) \\
&= \frac{f(\mathbf{y};\theta)\left(\frac{\partial^2}{\partial \theta^2} f(\mathbf{y};\theta)\right) - \left(\frac{\partial}{\partial \theta} f(\mathbf{y};\theta)\right)\left(\frac{\partial}{\partial \theta} f(\mathbf{y};\theta)\right)}{f^2(\mathbf{y};\theta)} \\
&= \frac{f(\mathbf{y};\theta)\frac{\partial^2 f(\mathbf{y};\theta)}{\theta^2} - \left(\frac{\partial}{\partial \theta} f(\mathbf{y};\theta)\right)^2}{f^2(\mathbf{y};\theta)}\,.
\end{aligned}$$

- Tomando a esperança ficamos com

$$-E\left[\frac{\partial^2}{\partial \theta^2} \log f(\mathbf{y};\theta)\right] = -\int \cdots \int \frac{f(\mathbf{y};\theta)\frac{\partial^2}{\partial \theta^2} f(\mathbf{y};\theta) - \left(\frac{\partial}{\partial \theta} f(\mathbf{y};\theta)\right)^2}{f^2(\mathbf{y};\theta)} f(\mathbf{y};\theta) d\mathbf{y}\,.$$

$$= -\int \cdots \int \frac{\partial^2 f(\mathbf{y};\theta)}{\partial \theta^2} d\mathbf{y} + \int \cdots \int \frac{\left(\frac{\partial}{\partial \theta} f(\mathbf{y};\theta)\right)^2}{f(\mathbf{y};\theta)} d\mathbf{y}$$

$$= -\frac{\partial^2}{\partial \theta^2} \int \cdots \int f(\mathbf{y};\theta) d\mathbf{y} + \int \cdots \int \left(\frac{\frac{\partial}{\partial \theta} f(\mathbf{y};\theta)}{f(\mathbf{y};\theta)}\right)^2 f(\mathbf{y};\theta) d\mathbf{y}$$

$$= -\frac{\partial^2}{\partial \theta^2}(1) + \int \cdots \int \left(\frac{\partial}{\partial \theta} \log f(\mathbf{y};\theta)\right)^2 f(\mathbf{y};\theta) d\mathbf{y}$$

$$= 0 + E\left[\frac{\partial}{\partial \theta} \log f(\mathbf{y};\theta)\right]^2 = I(\theta)\,.$$

- Podemos calcular a informação de Fisher com uma única observação $n = 1$.
- Podemos calcular a informação de Fisher com $n > 1$.
- Vamos denotar $I_n(\theta)$ a informação de Fisher com n dados.
- Qual a relação entre $I_n(\theta)$ e $I_1(\theta)$.
- $I_n(\theta)$ aumenta com n? A que taxa?
- Resultado: Se $Y_1, \ldots Y_n$ são i.i.d. então $I_n(\theta) = nI_1(\theta)$.
- A informação sobre θ numa amostra de tamanaho n é igual a n vezes a informação numa amostra de tamanho 1.
- A informação sobre θ aumenta linearmente com n.
- O estimador de máxima de verossimilhança apresenta a seguinte distribuição assintótica:

$$\hat{\theta}_{EMV} \approx N\left(\theta, \frac{1}{I_n(\theta)}\right)$$

- Isso significa que, para n grande e de forma aproximada, o MLE apresenta as seguintes características:
  - Tem distribuição normal;
  - É não viciado;
  - Atinge a cota de Cramér-Rao, ou seja, apresenta a menor variância possível.
- Este resultado é universal, não interessa o modelo de probabilidade para os dados!!
- Vamos lembrar de algumas propriedades importantes:

$$E\left(\frac{\partial \ell}{\partial \theta}\right) = 0 \quad \text{e} \quad E\left(\frac{\partial \ell}{\partial \theta}\right)^2 = -E\left(\frac{\partial^2 \ell}{\partial \theta^2}\right) = I_n(\theta) = nI_1(\theta)$$

- Vamos exemplificar estas propriedades no caso de v.a.'s iid Poisson com parâmetro θ.
- $X_1, X_2, ..., X_n$ v.a.'s iid com ditribuição Poisson(θ).
- Conjunta:

$$p(x_1, x_2, ..., x_n; \theta) = \prod_{i=1}^{n} \frac{\theta^{x_i} e^{-\theta}}{x_i!}$$

- Log-verossimilhança:

$$\ell(\theta) = \sum_{i=1}^{n} [x_i \log(\theta) - \theta - \log(x_i!)]$$

- Denotando $\ell_i = x_i \log(\theta) - \theta - \log(x_i!)$, podemos escrever $l(\theta) = \sum_{i=1}^{n} l_i(\theta)$.
- Derivando em relação a θ podemos obter a função escore

$$\frac{\partial \ell}{\partial \theta} = \sum_{i=1}^{n} \frac{\partial \ell_i}{\partial \theta} = \sum_{i=1}^{n} \left(\frac{x_i}{\theta} - 1\right)$$

- Vista como *variável aleatória*, a função escore é dada por

$$\sum_{i=1}^{n} \left(\frac{X_i}{\theta} - 1\right) = \sum_{i=1}^{n} Y_i$$

- Como $E(X_i) = \theta$, sabemos que

$$E(Y_i) = E\left(\frac{X_i}{\theta} - 1\right) = E\left(\frac{X_i}{\theta}\right) - 1 = 0$$

- Teremos então que

$$Var(Y_i) = E(Y_i^2) = I_1(\theta)$$

- As variáveis aleatórias $Y_1, Y_2, ..., Y_n$ são i.i.d com média zero e variância dada por $I_1(\theta)$.

- Derivando pela segunda vez em relação a θ:

$$\frac{\partial^2 \ell}{\partial \theta^2} = \sum_{i=1}^{n} \frac{\partial^2 \ell_i}{\partial \theta^2} = \sum_{i=1}^{n} \frac{-x_i}{\theta^2}$$

- Olhando para essa função como uma variável aleatória temos que

$$E\left(\frac{\partial^2 \ell}{\partial \theta^2}\right) = \sum_{i=1}^{n} E\left(\frac{-X_i}{\theta^2}\right) = -\frac{n}{\theta}$$

- Mas é fácil perceber que

$$E\left(\frac{\partial l}{\partial \theta}\right)^2 = \sum_{i} E\left(\frac{x_i^2}{\theta} - 2\frac{x_i}{\theta} + 1\right) = \frac{n}{\theta}$$

- Ou seja, verificamos o resultado $I_n(\theta) = -E(\frac{\partial^2 l}{\partial \theta^2}) = E(\frac{\partial l}{\partial \theta})^2 = nI_1(\theta)$.
- Vamos mostrar que o MLE é assitoticamente normal: precisamos relembrar dois teoremas importantes.
- **Lei Forte dos Grandes Números** Se $Y_1, Y_2, ..., Y_n$ são variáveis aletórias i.i.d com esperanaça finnita então $\bar{Y}_n \to E(Y)$ quando $n \to \infty$.
- **Teorema Central do Limite**: Se $Y_1, Y_2, ..., Y_n$ são variáveis aletórias i.i.d com $E(Y) = 0$ e $Var(Y) = v$. Então
  - $\sqrt{n}(\bar{Y}_n)$ tende em distribuição para uma $N(0, v)$ ou
  - $\sqrt{n}\frac{\bar{Y}_n}{\sqrt{v}}$ tende em distribuição para uma $N(0,1)$ ou ainda
  - $\frac{1}{\sqrt{n}}(Y_1 + Y_2 + ... + Y_n)$ tende em distribuição para uma $N(0, v)$.
- Vamos fazer a expnsão de $\frac{\partial l}{\partial \theta}$ em torno do verdadeiro valor do parâmetro θ, que denotaremos por θ^*.
- $\frac{\partial l}{\partial \theta}(\theta^*)$ o valor de $\frac{\partial l}{\partial \theta}$ avaliado no ponto θ^*.
- Como o EMV maximiza a função de verossimilhança:

$$\frac{\partial l}{\partial \theta}(\hat{\theta}_{EMV}) = 0$$

- Expandindo a derivada:

$$0 = \frac{\partial l}{\partial \theta}(\hat{\theta}_{EMV}) \approx \frac{\partial l}{\partial \theta}(\theta^*) + \frac{\partial^2 l}{\partial \theta^2}(\theta^*)(\hat{\theta}_{EMV} - \theta^*)$$

- Então

$$\hat{\theta}_{EMV} - \theta^* \approx \left[-\frac{\partial^2 l}{\partial \theta^2}(\theta^*)\right]^{-1} \frac{\partial l}{\partial \theta}(\theta^*)$$

- Logo

$$\begin{aligned} \sqrt{n}(\hat{\theta}_{EMV} - \theta^*) &\approx \sqrt{n}\left[-\frac{\partial^2 l}{\partial \theta^2}(\theta^*)\right]^{-1} \frac{\partial l}{\partial \theta}(\theta^*) \\ &= \left[-\frac{1}{n}\frac{\partial^2 l}{\partial \theta^2}(\theta^*)\right]^{-1} \frac{1}{\sqrt{n}}\frac{\partial l}{\partial \theta}(\theta^*) \end{aligned}$$

- Chegando então a

$$\underbrace{\sqrt{n}(\hat{\theta}_{EMV} - \theta^*)}_{\sqrt{n} * \text{ erro de estimação}} \approx \underbrace{\left[-\frac{1}{n}\frac{\partial^2 l}{\partial \theta^2}(\theta^*)\right]^{-1}}_{A} \underbrace{\frac{1}{\sqrt{n}}\frac{\partial l}{\partial \theta}(\theta^*)}_{B}$$

- Desenvolvendo o termo em (A) temos

$$\left[-\frac{1}{n}\frac{\partial^2 l}{\partial\theta^2}(\theta^*)\right] = -\left(\frac{1}{n}\sum_i \frac{\partial^2 l_i}{\partial\theta^2}(\theta^*)\right)$$

- Olhando para $\frac{\partial^2 l_i}{\partial\theta^2}(\theta^*)$ como uma variável aleatória chega-se a

$$-\left(\frac{1}{n}\sum_i \frac{\partial^2 l_i}{\partial\theta^2}(\theta^*)\right) = -\left(\frac{1}{n}\sum_i Z_i\right)$$

- onde $Z_1, Z_2, ..., Z_n$ são variáveis aleatórias i.i.d. com $E_\theta(Z_i) = -I_1(\theta^*)$
- Pela Lei Forte dos Grandes Números, $\overline{Z}_n \to -I_1(\theta^*)$ quando $n \to \infty$
- Portanto o termo (A) se aproxima de $I_1(\theta^*)^{-1}$ para n suficientemente grande.
- Vamos agora desenvolver o termo em (B)

$$\frac{1}{\sqrt{n}}\frac{\partial l}{\partial\theta}(\theta^*) = \frac{1}{\sqrt{n}}\sum_i \frac{\partial l_i}{\partial\theta}(\theta^*)$$

- Novamente olhando para $\frac{\partial l_i}{\partial\theta^2}(\theta^*)$ como uma variável aleatória temos que

$$\frac{1}{\sqrt{n}}\sum_i \frac{\partial l_i}{\partial\theta}(\theta^*) = \frac{1}{\sqrt{n}}\sum_i W_i$$

- onde $W_1, W_2, ..., W_n$ são variáveis aleatórias i.i.d. com $E_{\theta^*}(W_i) = 0$ e $Var_{\theta^*}(W_i) = E_{\theta^*}(W_i^2) = I_1(\theta^*)$
- Pelo Teorema Central do Limite temos que

$$\frac{1}{\sqrt{n}}(W_1 + W_2 + ... + W_n) \xrightarrow{d} N(0, I_1(\theta^*))$$

  onde a notação d significa tender em distribuição.
- Portanto, como $\sqrt{n}(\hat{\theta}_{EMV} - \theta^*)$ é aproximadamente o termo em (B) multiplicado por $I_1(\theta)$,
- o erro de estimação é normalmente distribuído com

$$E(\sqrt{n}(\hat{\theta}_{EMV} - \theta^*)) = 0 \text{ e } Var(\sqrt{n}(\hat{\theta}_{EMV} - \theta^*)) = \frac{I_1(\theta^*)}{I_1^2(\theta^*)} = \frac{1}{I_1(\theta^*)}$$

- Em outras palavras, o erro de estimação $\sqrt{n}(\hat{\theta}_{EMV} - \theta^*)$ converge em distribuição para uma $N(0, I_1(\theta^*)^{-1})$.
- Logo se n é grande o suficiente o MLE tem distribuição aproximadamente normal com média θ^* e variância $I_1(\theta^*)^{-1}$, ou seja, é aproximadamente não viciado e atinge a cota de Cramer-Rao. Dessa forma, provamos as três principais propriedades do MLE.
- Podemos concluir que

$$\sqrt{n}(\hat{\theta}_{EMV} - \theta^*) \approx N(0, I_1(\theta^*)^{-1})$$

- Isto é,

$$(\hat{\theta}_{EMV} - \theta^*) \approx N\left(0, \frac{1}{nI_1(\theta^*)}\right)$$

- ou ainda

$$\hat{\theta}_{EMV} \approx N\left(\theta^*, \frac{1}{nI_1(\theta^*)}\right)$$

20.2 Multivariado

- Vamos considerar o caso em que o parâmetro $\theta = (\theta_1, \ldots, \theta_k) \in \mathbb{R}^k$ é um vetor com $k > 1$.
- O MLE $\widehat{\theta} = (\widehat{\theta}_1, \ldots, \widehat{\theta}_k)$ é também um vetor de dimensão k.
- O elemento $\widehat{\theta}_i$ é um estimador de θ_i (a entrada i do vetor θ.
- O MLE $\widehat{\theta}$ é obtido resolvendo-se o sistema de equações não-lineares que resulta de igualar o gradiente $\frac{\partial \ell}{\partial \theta}$ da log-verossimilhança $\ell(\theta)$ a zero:

$$\frac{\partial \ell}{\partial \theta} = \begin{bmatrix} \partial \ell / \partial \theta_1 \\ \partial \ell / \partial \theta_2 \\ \vdots \\ \partial \ell / \partial \theta_k \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

- Queremos estudar o comportamento estatístico do MLE.
- Veremos que ele é um estimador ótimo de θ.
- O MLE $\widehat{\theta} = (\widehat{\theta}_1, \ldots, \widehat{\theta}_k)$ é um vetor ALEATÓRIO.
- Ele depende dos dados estocásticos e por isto varia de amostra para amostra.
- Precisamos estudar este comportamento aleatório do MLE.
- Como veremos, o comportamento é similar ao do caso unidimensional.
- Se n não é muito pequeno, de forma aproximada, teremos o seguinte:
  - O MLE $\widehat{\theta}$ é aproximadamente não viciado para estimar o vetor θ. Isto é, $\mathbb{E}(\widehat{\theta}) \approx \theta$.
  - Aproximadamente, a matriz de variância e covariância de $\widehat{\theta}$ é, num certo sentido matemático, a "menor matriz" possível para um estimador não-viciado, a matriz de informação de Fisher $I(\theta)$.
  - O MLE $\widehat{\theta}$ possui distribuição gaussiana multivariada aproximadamente.
- A matriz de variância e covariância de $\widehat{\theta}$ está associada à matriz de informação de Fisher $I(\theta)$.
- Definição: A matriz de informação de Fisher $I(\theta)$ é uma matriz $k \times k$ cujo elemento ij é dado por

$$[I(\theta)]_{ij} = \mathbb{E}\left(\frac{\partial \ell}{\partial \theta_i} \frac{\partial \ell}{\partial \theta_j} \right)$$

  onde $\ell = \ell(\theta)$ é a log-verossimilhança de θ.
- Por exemplo, no caso em que $\theta = (\theta_1, \theta_2, \theta_3) \in \mathbb{R}^3$ teremos

$$I(\theta) = \begin{bmatrix} \mathbb{E}\left(\frac{\partial \ell}{\partial \theta_1}\right)^2 & \mathbb{E}\left(\frac{\partial \ell}{\partial \theta_1}\frac{\partial \ell}{\partial \theta_2}\right) & \mathbb{E}\left(\frac{\partial \ell}{\partial \theta_1}\frac{\partial \ell}{\partial \theta_3}\right) \\ \mathbb{E}\left(\frac{\partial \ell}{\partial \theta_2}\frac{\partial \ell}{\partial \theta_1}\right) & \mathbb{E}\left(\frac{\partial \ell}{\partial \theta_2}\right)^2 & \mathbb{E}\left(\frac{\partial \ell}{\partial \theta_2}\frac{\partial \ell}{\partial \theta_3}\right) \\ \mathbb{E}\left(\frac{\partial \ell}{\partial \theta_3}\frac{\partial \ell}{\partial \theta_1}\right) & \mathbb{E}\left(\frac{\partial \ell}{\partial \theta_3}\frac{\partial \ell}{\partial \theta_2}\right) & \mathbb{E}\left(\frac{\partial \ell}{\partial \theta_3}\right)^2 \end{bmatrix}$$

- Suponha que $Y_1, Y_2, \ldots, Y_n$ sejam i.i.d. $N(\mu, \sigma^2)$.
- Temos $\theta = (\mu, \sigma^2) \in \mathbb{R}^2$
- A log-verossimilhança de θ baseada numa amostra é dada por

$$\ell(\theta) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - -\frac{1}{2\sigma^2}\sum_i (y_i - \mu)^2$$

- O vetor gradiente $k \times 1$ da log-verossimilhança $\ell(\theta)$ é dado por

$$\frac{\partial \ell}{\partial \theta} = \begin{bmatrix} \partial \ell / \partial \mu \\ \partial \ell / \partial \sigma^2 \end{bmatrix} = \begin{bmatrix} \sum_i (y_i - \mu)/\sigma^2 \\ -\frac{n}{2}\frac{1}{\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_i (y_i - \mu)^2 \end{bmatrix}$$

- A partir deste vetor, podemos obter a matriz de informação de Fisher tomando os seus produtos cruzados.
- Temos

$$\begin{aligned} I(\theta) &= \mathbb{E}\left[\begin{array}{cc} \left(\frac{\partial \ell}{\partial \mu}\right)^2 & \left(\frac{\partial \ell}{\partial \mu}\frac{\partial \ell}{\partial \sigma^2}\right) \\ \left(\frac{\partial \ell}{\partial \sigma^2}\frac{\partial \ell}{\partial \mu}\right) & \left(\frac{\partial \ell}{\partial \sigma^2}\right)^2 \end{array} \right] \\ &= \mathbb{E}\left[\begin{array}{cc} \left(\Sigma_i \frac{Y_i-\mu}{\sigma^2}\right)^2 & \left(\Sigma_i \frac{Y_i-\mu}{\sigma^2}\right)\left(-\frac{n}{2}\frac{1}{\sigma^2}+\frac{1}{2}\frac{\Sigma_i (Y_i-\mu)^2}{(\sigma^2)^2}\right) \\ \left(-\frac{n}{2}\frac{1}{\sigma^2}+\frac{1}{2}\frac{\Sigma_i (Y_i-\mu)^2}{(\sigma^2)^2}\right)\left(\Sigma_i \frac{Y_i-\mu}{\sigma^2}\right) & \left(-\frac{n}{2}\frac{1}{\sigma^2}+\frac{1}{2}\frac{\Sigma_i (Y_i-\mu)^2}{(\sigma^2)^2}\right)^2 \end{array} \right] \end{aligned}$$

- Um cálculo de probabilidade laborioso permite obter cada uma das esperanças presentes na matriz $I(\theta)$ resultando no seguinte:

$$I(\theta) = \left[\begin{array}{cc} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{n}{2\sigma^4} \end{array} \right]$$

- De forma similar ao caso unidimensional, podemos calcular a informação de Fisher de forma alternativa, usando a matriz Hessiana (a matriz de derivadas parciais de segunda ordem de $\ell(\theta)$):

$$[I(\theta)]_{ij} = -\mathbb{E}\left(\frac{\partial^2 \ell}{\partial \theta_i \partial \theta_j}\right)$$

- Por exemplo, no caso em que $\theta = (\theta_1, \theta_2, \theta_3) \in \mathbb{R}^3$, por esta fórmula alternativa, temos

$$I(\theta) = -\left[\begin{array}{ccc} \mathbb{E}\left(\frac{\partial^2 \ell}{\partial \theta_1^2}\right) & \mathbb{E}\left(\frac{\partial^2 \ell}{\partial \theta_1 \partial \theta_2}\right) & \mathbb{E}\left(\frac{\partial^2 \ell}{\partial \theta_1 \partial \theta_3}\right) \\ \mathbb{E}\left(\frac{\partial^2 \ell}{\partial \theta_2 \partial \theta_1}\right) & \mathbb{E}\left(\frac{\partial^2 \ell}{\partial \theta_2^2}\right) & \mathbb{E}\left(\frac{\partial^2 \ell}{\partial \theta_2 \partial \theta_3}\right) \\ \mathbb{E}\left(\frac{\partial^2 \ell}{\partial \theta_3 \partial \theta_1}\right) & \mathbb{E}\left(\frac{\partial^2 \ell}{\partial \theta_3 \partial \theta_2}\right) & \mathbb{E}\left(\frac{\partial^2 \ell}{\partial \theta_3^2}\right) \end{array} \right]$$

- Note o sinal negativo na frente da matriz. Além disso, podemos passar o $\mathbb{E}$ para fora da matriz: esperança de matriz é a esperança de cada entrada da matriz.
- Considerando o exemplo anterior de v.a.'s i.i.d. com distribuição $N(\mu, \sigma^2)$, a partir do vetor gradiente obtemos a matriz Hessiana abaixo:

$$\left[\begin{array}{cc} -\frac{n}{\sigma^2} & -\frac{\Sigma_i (y_i-\mu)}{(\sigma^2)^2} \\ -\frac{\Sigma_i (y_i-\mu)}{(\sigma^2)^2} & \frac{n}{2}\frac{1}{(\sigma^2)^2} - \frac{\Sigma_i (y_i-\mu)^2}{(\sigma^2)^3} \end{array} \right]$$

- Para obter $I(\theta)$, substituímos as instâncias $\mathbf{Y}$ pelas v.a.'s $\mathbf{Y}$ e tomamos o negativo da esperança da matriz hessiana.
- Resulta que, neste exemplo, tomar a esperança das expressões da matriz Hessiana é muito simples. $I(\theta)$.
- O resultado é o mesmo de antes:

$$I(\theta) = \left[\begin{array}{cc} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{n}{2\sigma^4} \end{array} \right]$$

- A matriz de informação $I(\theta)$ pode ser definida termo a termo (como fizemos) ou de forma vetorial.
- A primeira fórmula para $I(\theta)$, que especifica o elemento ij da matriz como $[I(\theta)]_{ij} = \mathbb{E}(\partial \ell / \partial \theta_i \cdot \partial \ell / \partial \theta_j)$ pode ser escrita de forma vetorial como a esperança da matriz que

resulta ao multiplicar duas vezes o vetor gradiente de $\ell(\theta)$ de forma transposta:

$$\begin{aligned} I(\theta) &= \mathbb{E}\left(\left[\frac{\partial \ell}{\partial \theta}\right] \cdot \left[\frac{\partial \ell}{\partial \theta}\right]'\right) \\ &= \mathbb{E}\left(\begin{bmatrix} \frac{\partial \ell}{\partial \theta_1} \\ \frac{\partial \ell}{\partial \theta_2} \\ \vdots \\ \frac{\partial \ell}{\partial \theta_k} \end{bmatrix} \cdot \left[\frac{\partial \ell}{\partial \theta_1}, \frac{\partial \ell}{\partial \theta_2}, \ldots, \frac{\partial \ell}{\partial \theta_k}\right]\right) \end{aligned}$$

- Considerando o exemplo de n v.a.'s i.i.d. com distribuição $N(\mu, \sigma^2)$ e $\theta = (\mu, \sigma^2)$ temos

$$\begin{aligned} I(\theta) &= \mathbb{E}\begin{bmatrix} \partial\ell/\partial\mu \\ \partial\ell/\partial\sigma^2 \end{bmatrix} \cdot \left[\partial\ell/\partial\mu, \partial\ell/\partial\sigma^2\right] \\ &= \mathbb{E}\begin{bmatrix} \Sigma_i (y_i-\mu)/\sigma^2 \\ -\frac{n}{2}\frac{1}{\sigma^2} + \frac{1}{2(\sigma^2)^2}\Sigma_i (y_i-\mu)^2 \end{bmatrix} \cdot \left[\sum_i (y_i-\mu)/\sigma^2, -\frac{n}{2}\frac{1}{\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_i (y_i-\mu)^2\right] \end{aligned}$$

- Além da forma vetorial dada por

$$I(\theta) = \mathbb{E}\left(\left[\frac{\partial \ell}{\partial \theta}\right] \cdot \left[\frac{\partial \ell}{\partial \theta}\right]'\right)$$

podemos obter $I(\theta)$ usando uma forma vetorial para expressar a matriz hessiana de derivadas segundas.

- Especificamente, temos

$$I(\theta) = -\mathbb{E}\left[\frac{\partial}{\partial \theta}\frac{\partial \ell}{\partial \theta}\right]$$

- Quando T é um estimador não viciado de $\theta \in \mathbb{R}$ vimos que $\text{MSE}(T) = \mathbb{V}(T) \geq 1/I(\theta)$.
- Temos uma versão multivariada deste resultado para $\theta = (\theta_1, \ldots, \theta_k) \in \mathbb{R}^k$.
- Seja $\theta y = \theta y(\mathbf{Y}) = (T_1, \ldots, T_k) \in \mathbb{R}^k$ um estimador k-dimensional não-viciado de θ (isto é, $\mathbb{E}(\theta y) = \theta$).
- Então, para todo i, temos

$$\text{MSE}(T_i) = \mathbb{V}(T_i) \geq \left[I^{-1}(\theta)\right]_{ii}$$

que é o elemento ii da INVERSA da matriz de informação de Fisher.

- O resultado multivariado que apresentamos é apenas parcial.
- O resultado completo é um pouco mais sutil e útil em alguns problemas estatísticos de estimação. e testes de hipóteses em experimentos planejados.
- Considere qualquer combinação linear dos k elementos do estimador não viciado $\theta y = \theta y(\mathbf{Y})$:

$$W = c_1 T_1 + c_2 T_2 + \ldots + c_k T_k = (c_1, \ldots, c_k)' \cdot (T_1, \ldots, T_k) = \mathbf{c}'\theta y$$

onde $\mathbf{c} = (c_1, c_2, \ldots, c_n)$ são constantes conhecidas.

- É fácil mostrar que W é estimador não-viciado para $c_1\theta_1 + c_2\theta_2 + \ldots + c_k\theta_k = \mathbf{c}'\theta$.
- Então temos

$$\text{MSE}(W) = \mathbb{V}(W) = \mathbb{V}(\mathbf{c}'\theta y) \geq \mathbf{c}'\, I(\theta)\, \mathbf{c}$$

- Suponha que o parâmetro $\theta = (\theta_1, \ldots, \theta_k)$ é um vetor com k componentes
- O MLE de θ é um vetor aleatório $\widehat{\theta}$ de dimensão k.
- O resultado anterior sobre o MLE é válido em multi-dimensões:

$$\widehat{\theta} \approx N_k\left(\theta, I^{-1}(\theta)\right)$$

- Se n não é pequeno, temos aproximadamente:

  - O MLE possui distribuição de probabilidade normal multivariada;
  - é aproximadamente não-viciado (isto é, $\mathbb{E}(\widehat{\theta}) \approx \theta$);
  - A matriz de covariância do MLE é aproximadamente igual ao inverso de $I(\theta)$, que é um limite ótimo para estimadores não-viciados.
- No modelo de regressão linear múltipla, temos v.a.'s $Y_1, \ldots, Y_n$ que são independentes mas não são i.d.: $Y_i \sim N(\mathbf{X}_i' \beta, \sigma^2)$.
- A versão matricial do modelo de regressão linear é

$$\mathbf{Y} = \mathbf{X}\,\beta + \varepsilon$$

- onde $\mathbf{Y}$ é vetor $n \times 1$, a matriz de desenho $\mathbf{X}$ é de dimensão $n \times p$ com $p-1$ variáveis independentes e a coluna de 1's.
- O vetor $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)'$ é composto de v.a.'s i.i.d. $N(0, \sigma^2)$.
- O parâmetro θ é $\theta = (\beta, \sigma^2) = (\beta_0, \ldots, \beta_{p-1}, \sigma^2)$
- Queremos estimar β e σ^2.
- O MLE $\widehat{\beta}$ de $\beta = (\beta_0, \ldots, \beta_p)$ coincide com o estimador de mínimos quadrados.
- Para ver isto, considere a log-verossimilhança $\ell(\theta)$.
- As v.a.'s Y_i são independentes com distribuição $N(\mathbf{X}_i' \beta, \sigma^2)$.
- Então

$$\begin{aligned} \ell(\theta) &= \log\left(\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y_i - \mathbf{X}_i' \beta)^2\right)\right) \\ &= -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_i (y_i - \mathbf{X}_i' \beta)^2 \end{aligned}$$

- É fácil perceber que, para qualquer valor fixo de σ^2, o vetor β que maximiza $\ell(\theta)$ será aquele que minimiza a soma de quadrados $\sum_i (y_i - \mathbf{X}_i' \beta)^2$.
- Assim, o MLE de β é o vetor que minimiza a distância entre o vetor $\mathbf{Y}$ e a combinação linear $\mathbf{X}\beta$:

$$\widehat{\beta} = \arg\min ||\mathbf{Y} - \mathbf{X}\,\beta||^2$$

- Já apredemos que a solucão deste último problema é a combinação que produz a projeção ortogonal de $\mathbf{Y}$ no espaço vetorial das combinações lineares das colunas de $\mathbf{X}$:

$$\widehat{\beta} = (\widehat{\beta_0}, \ldots, \widehat{\beta_p})' = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{Y}$$

- $\widehat{\beta}$ é uma função do vetor de dados aleatórios $\mathbf{Y}$ e da matriz de regressores $\mathbf{X}$ (que é considerada uma matriz de constantes conhecida).
- Se escrevermos a matriz $k \times n$ dada por $(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t$ por A podemos ver que $\widehat{\beta} = A\mathbf{Y}$.
- Assim, cada elemento de $\widehat{\beta}$ é uma combinação linear dos elementos do vetor $\mathbf{Y}$.
- O modelo prediz ou estima o valor de Y_i usando $\widehat{\beta}$.
- O vetor $n \times 1$ com os valores preditos pelo modelo para os valores realmente observados $\mathbf{Y}$ são dados por

$$\widehat{\mathbf{Y}} = \mathbf{X}\,\beta = \mathbf{X}\,(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{Y}$$

- A diferença entre $\mathbf{Y}$ e a predição $\widehat{\mathbf{Y}}$ forma o vetor de resíduos ou vetor de erros de predição $= \mathbf{Y} - \widehat{\mathbf{Y}}$.
- O i-ésimo elemento do vetor de resíduos é dado por

$$r_i = Y_i - \mathbf{X}_i' \widehat{\beta}$$

- Gráficos e análises com os resíduos são uma excelente maneira de identificar falhas nas suposições do modelo de regressão linear
- Seja *RSS* a soma de quadrados dos resíduos (residual sum of squares)

$$RSS = ||\mathbf{Y} - \widehat{\mathbf{Y}}||^2 = \sum_{i=1}^{n} (Y_i - \widehat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - \mathbf{X}_i' \, \widehat{\beta})^2$$

- O MLE de σ^2 é dado pela média dos resíduos ao quadrado:

$$\widehat{\sigma^2} = \frac{RSS}{n}$$

- Para verificar isto, veja que tomando $\beta = \widehat{\beta}$ teremos a log-verossimilhança $\ell(\theta)$ igual a

$$\ell(\widehat{\beta}, \sigma^2) = -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} RSS$$

que é maximizada se tomarmos $\sigma^2 = RSS/n$ (basta derivar e igualar a zero).
- Revisão de probabilidade: A é matriz de constantes e $\mathbf{Y}$ é vetor aleatório de dimensões compatíveis. Então o vetor aleatório $A\mathbf{Y}$ tem um vetor esperado igual a $\mathbb{E}(A\mathbf{Y}) = A\mathbb{E}(\mathbf{Y})$ e a matriz de covariância é $\mathbb{V}(A\mathbf{Y}) = A\mathbb{V}(\mathbf{Y})A^t$.
- O MLE $\widehat{\beta}$ é não-viciado para β pois

$$\begin{aligned} \mathbb{E}(\widehat{\beta}) &= \mathbb{E}((\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{Y}) \\ &= (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t \, \mathbb{E}(\mathbf{Y}) \\ &= (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbb{E}(\mathbf{X}\,\beta + \varepsilon) \\ &= (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t(\mathbf{X}\,\beta + \mathbb{E}(\varepsilon)) \\ &= \beta + \mathbf{0} = \beta \end{aligned}$$

- É possível calcular a matriz de covariância de $\widehat{\beta}$.
- Chame $A = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t$ e então $\widehat{\beta} = A\mathbf{Y}$.
- Como as v.a.'s Y_i são independentes, sua matriz de covariância é $\sigma^2 I_n$, um múltiplo da matriz identidade de ordem n.
- Então, usando o resultado de probab, temos

$$\mathbb{V}(\widehat{\beta}) = A(\sigma^2 I_n)A^t = \sigma^2(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t \, \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1} = \sigma^2(\mathbf{X}^t\mathbf{X})^{-1}$$

- Quanto ao MLE de σ^2 no modelo de regressão linear, podemos mostrar que

$$\mathbb{E}(\widehat{\sigma^2}) = \mathbb{E}(\frac{RSS}{n}) = \frac{n-k-1}{n}\sigma^2$$

- Portanto o MLE $\widehat{\sigma^2}$ é viciado para estimar σ^2 mas seu vício desaparece quando n cresce.
- É comum usarmos um estimador não-viciado para σ^2 dados por

$$S^2 = \frac{RSS}{n-k-1}$$

- O uso deste estimador não-viciado (que não é o MLE) permite fazer vários cálculos exatos de probabiliadde envolvidos em intervalos de confiança e testes de hipótese (mais a frente no curso).
- Para o MLE $\widehat{\sigma^2} = RSS/n$ temos a variância $2\sigma^4(n-k-1)/n^2$.
- Além disso, a covariância entre $\widehat{\beta}$ e $\widehat{\sigma^2}$ é o vetor $\mathbf{0}$.

- Podemos voltar a log-verossmilhança $\ell(\theta)$ e calcular a matriz de informação de Fisher usando vários truques de probabilidade e álgebra linear.
- No final encontramos

$$I(\theta) = \begin{pmatrix} \frac{1}{\sigma^2}\mathbf{X}^t\mathbf{X} & \mathbf{0} \\ \mathbf{0}^t & \frac{n}{2\sigma^4} \end{pmatrix}$$

- A matriz inversa é igual a

$$I^{-1}(\theta) = \begin{pmatrix} \sigma^2(\mathbf{X}^t\mathbf{X})^{-1} & \mathbf{0} \\ \mathbf{0}^t & \frac{2\sigma^4}{n} \end{pmatrix}$$

- A cota-inferior de Cramér-Rao para um estimador nã-viciado de θ é

$$I^{-1}(\theta) = \begin{pmatrix} \sigma^2(\mathbf{X}^t\mathbf{X})^{-1} & \mathbf{0} \\ \mathbf{0}^t & \frac{2\sigma^4}{n} \end{pmatrix}$$

- Compare $I^{-1}(\theta)$ com a matriz de covariância exata do MLE

$$\mathbb{V}(\widehat{\theta}) = \mathbb{V}(\widehat{\beta}, \widehat{\sigma^2}) = \begin{pmatrix} \sigma^2(\mathbf{X}^t\mathbf{X})^{-1} & \mathbf{0} \\ \mathbf{0}^t & \frac{2\sigma^4(n-k-1)}{n^2} \end{pmatrix} \to I^{-1}(\theta)$$

- O MLE do parâmetro β é não-viciado para β e atinge a cota inferior de Cramér-Rao $\sigma^2(\mathbf{X}^t\mathbf{X})^{-1}$
- O MLE de σ^2 tem um vício que vai a zero com n e tem uma variância que aproxima-se rapidamente da cota inferior de Cramér-Rao.
- No caso da regressão logística, o MLE $\widehat{\theta}$ vai satisfazer a equação de verossimilhança não-linear

$$\frac{\partial \ell(\theta)}{\partial \theta} = \mathbf{X}^t\mathbf{Y} - \mathbf{X}^t\mathbf{p} = \mathbf{0}$$

  onde $\mathbf{p} = (p_1, \ldots p_n)$ e $p_i = 1/(1+\exp(-\mathbf{X}_i^t\theta))$.
- O MLE é o resultado da da convergência do algoritmo de Newton-Raphson:

$$\theta^{\text{new}} = \theta^{\text{old}} - \left(\frac{\partial^2 \ell(\theta)}{\partial\theta\partial\theta^t}\right)^{-1} \frac{\partial \ell(\theta)}{\partial \theta}$$

  onde as derivadas são avaliadas usando-se o valor corrente θ^{old}.
- Como valor inicial, podemos usar $\theta^{(0)} = \mathbf{0}$.
- Temos

$$\frac{\partial \ell(\theta)}{\partial \theta} = \mathbf{X}^t\mathbf{Y} - \mathbf{X}^t\mathbf{p}$$

- e

$$\frac{\partial^2 \ell(\theta)}{\partial\theta\partial\theta^t} = -\mathbf{X}^t\mathbf{W}\mathbf{X}$$

  onde

$$\mathbf{W} = \begin{bmatrix} p_1(1-p_1) & 0 & 0 & \ldots & 0 \\ 0 & p_2(1-p_2) & 0 & \ldots & 0 \\ 0 & 0 & p_3(1-p_3) & \ldots & 0 \\ \vdots & \vdots & \vdots & \ldots & \vdots \\ 0 & 0 & 0 & \ldots & p_n(1-p_n) \end{bmatrix}$$

- Temos então

$$\theta^{\text{new}} = \theta^{\text{old}} + \left(\mathbf{X}^t\mathbf{W}\mathbf{X}\right)^{-1}\ \mathbf{X}^t(\mathbf{Y}-\mathbf{p})$$

- Para obter a matriz de informação de Fisher, calculamos a matriz Hessiana (a matriz de derivadas parciais de segunda ordem da log-verossimilhança $\ell(\theta)$):

$$\frac{\partial^2\ell(\theta)}{\partial\theta\partial\theta^t}$$

  e depois calculamos $I(\theta)$ como o negativo do valor esperado de cada elemento desta matriz hessiana.
- A matriz Hessiana é igual a

$$\frac{\partial^2\ell(\theta)}{\partial\theta\partial\theta^t} = -\mathbf{X}^t\mathbf{W}\mathbf{X}$$

- Que sorte!! Esta é uma matriz não-aleatória, somente com constantes.
- Assim, sua esperança é igual a ela mesma!
- Portanto, a matriz de informação de Fisher é

$$I(\theta) = \mathbf{X}^t\mathbf{W}\mathbf{X}$$

- Conclusão:

$$\widehat{\theta} \sim N_{k+1}\left(\theta, (\mathbf{X}^t\mathbf{W}\mathbf{X})^{-1}\right)$$

  com

$$\mathbf{W} = \text{diag}\left(p_1(1-p_1), p_2(1-p_2), \ldots, p_n(1-p_n)\right)$$

- Cada p_i é função das covariáveis e de θ:

$$p_i = 1/(1+\exp(-\mathbf{X}_i^t\theta))$$

- Isto é, o EMV de θ tem uma distribuição aproximadamente normal multivariada de dimensão $k+1$ centrada no verdadeiro valor do parâmetro θ e com matriz de covariância $I^{-1}(\theta) = (\mathbf{X}^t\mathbf{W}\mathbf{X})^{-1}$
- A matrix $\mathbf{W}$ depende do valor desconhecido do parâmetro θ.
- Nós podemos usar nosso (melhor) estimador disponível $\widehat{\theta}$ em $\mathbf{W}$ para obter uma aproximação para $I(\theta)$.
- O MLE $\widehat{\theta} = (\widehat{\theta_1}, \ldots, \widehat{\theta_k})$ é um vetor ALEATÓRIO.
- A sua distribuição conjunta é dada por

$$\widehat{\theta} \sim N_k\left(\theta, \mathbf{I}^{-1}(\theta)\right)$$

- Considerando apenas a i-ésima coordenada, temos

$$\hat{\theta}_i \approx N_1(\theta_i, \left[\mathbf{I}^{-1}(\theta)\right]_{ii})$$

  onde $\left[\mathbf{I}^{-1}(\theta)\right]_{ii}$ é o elemento (i,i) na diagonal da INVERSA da matriz de informação de Fisher.
- Podemos fornecer intervalos de confiança para cada θ_i.

- Considerando apenas a i-ésima coordenada, temos

$$\hat{\theta}_i \approx N_1(\theta_i, \left[\mathbf{I}^{-1}(\theta)\right]_{ii})$$

- Numa gaussiana qualquer, a probabilidade da variável afastar-se de sua esperança por menos que dois desvios-padrão é aproximadamente 0.95.
- Portanto, aproximadamente,

$$\mathbb{P}\left(|\hat{\theta}_i - \theta_i| \leq 2\sqrt{[\mathbf{I}^{-1}(\theta)]_{ii}}\right) \approx 0.95$$

- Desigualdade básica: $|a-b| \leq 2$ se, e somente se, $a-2 \leq b \leq a+2$
- Assim,

$$\mathbb{P}\left(\hat{\theta}_i - 2\sqrt{[\mathbf{I}^{-1}(\theta)]_{ii}} \leq \theta_i \leq \hat{\theta}_i + 2\sqrt{[\mathbf{I}^{-1}(\theta)]_{ii}}\right) \approx 0.95$$

- O intervalo

$$\left[\hat{\theta}_i - 2\sqrt{[\mathbf{I}^{-1}(\theta)]_{ii}}, \leq \hat{\theta}_i + 2\sqrt{[\mathbf{I}^{-1}(\theta)]_{ii}}\right]$$

  é chamado intervalo de confiança de 95% para o parâmetro θ_i.
- A confiança é no método. 95% dos intervalos que você fizer usando o MLE vão cobrir o verdadeiro valor do parâmetro.
- Em geral, $\mathbf{I}^{-1}(\theta)$ depende do parâmetro θ, que é desconhecido.
- Nestes casos, substitua θ por $\hat{\theta}$.

21. Confidence Intervals and Hypothesis Tests

21.1 Introdução

Sejam $X_1, X_2, ..., X_n$ variáveis aleatórias com densidade conjunta $f(\mathbf{x}, \theta)$. Quando estimamos o parâmetro θ, queremos ter uma idéia da precisão dessa estimativa. Isso pode ser obtido por meio de um intervalo de confiança calculado a partir dos dados. Esse intervalo deve ser construído de maneira a cubrir o verdadeiro valor do parâmetro com grande confiança, a qual é expressa em probabilidade. Dessa maneira, grande confiança significaria, por exemplo, uma probabilidade de 0.95 ou 0.99.

Suponha que $S_\mathbf{x} \in \Theta$ é um intervalo que depende apena dos dados $X_1, X_2, ..., X_n$ e de constantes conhecidas, tal como o tamanho da amostra n. Então, se $S_\mathbf{x}$ satisfizer a propriedade

$$P(\theta \in S_\mathbf{x}) = 1 - \alpha$$

o intervalo $S_\mathbf{x}$ é um intervalo de confiança com coeficiente (ou grau) $1 - \alpha$.

Note que, como $S_\mathbf{x}$ depende de variáveis aleatórias, ele é um intervalo aleatório.

Exemplo: Sejam $X_1, X_2, ..., X_n$ variáveis aleatórias i.i.d. com distribuição $N(\theta, 4)$, ou seja, a média é desconhecida, mas a variância é conhecida. Sabemos que

$$\bar{X} \sim N\left(\theta, \frac{4}{n}\right) \text{ e portanto } \frac{\bar{X} - \theta}{2/\sqrt{n}} \sim N(0,1)$$

logo

$$P\left(\left|\frac{\bar{X} - \theta}{2/\sqrt{n}}\right| \leq 1.96\right) = 0.95$$

ou seja, a área do gráfico compreendida entre -1.96 e 1.96 corresponde a uma probabilidade de 0.95, como é representado na Figura 21.1.

Tomando, então, $\alpha = 0.05$ e $S_\mathbf{x} = \left[\bar{x} - 1.96\left(\frac{2}{\sqrt{n}}\right); \bar{x} + 1.96\left(\frac{2}{\sqrt{n}}\right)\right]$ temos que

$$P(\theta \in S_\mathbf{x}) = P\left(\bar{x} - 1.96\left(\frac{2}{\sqrt{n}}\right) \leq \theta \leq \bar{x} + 1.96\left(\frac{2}{\sqrt{n}}\right)\right)$$

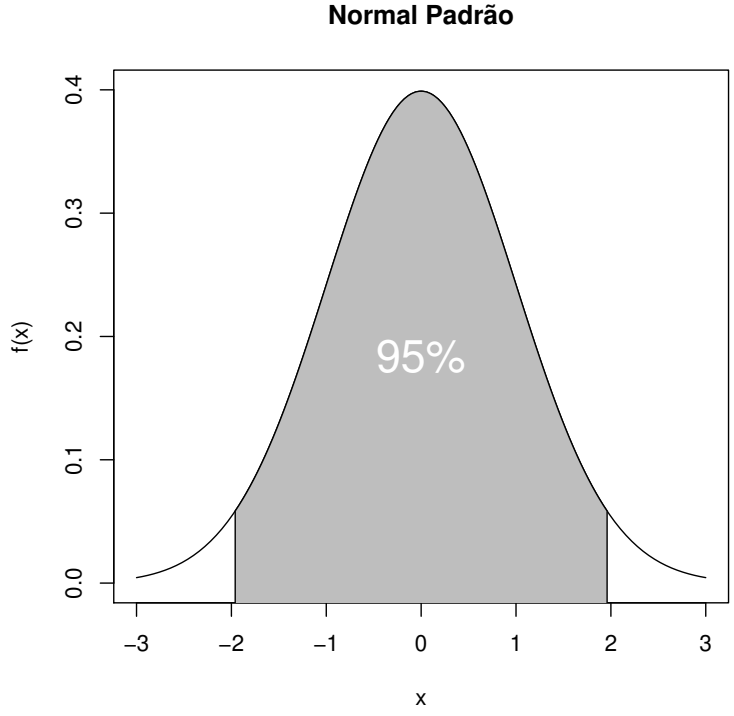


Figure 21.1: Área da curva normal entre -1.96 e 1.96

$$= P\left(-1.96\left(\frac{2}{\sqrt{n}}\right) \leq -\bar{x}+\theta \leq 1.96\left(\frac{2}{\sqrt{n}}\right)\right) = P\left(|\bar{x}-\theta| \leq 1.96\left(\frac{2}{\sqrt{n}}\right)\right) = P\left(\frac{|\bar{x}-\theta|}{2/\sqrt{n}} \leq 1.96\right)$$

mas $\frac{|\bar{x}-\theta|}{2/\sqrt{n}} \sim N(0,1)$ logo, consultando a tabela da Normal Padrão chegamos a

$$P\left(\frac{|\bar{x}-\theta|}{2/\sqrt{n}} \leq 1.96\right) = 0.95 \text{ ou seja } P(\theta \in S_{\mathbf{x}}) = 0.95$$

Vimos então que $S_{\mathbf{x}} = \left[\bar{x} - 1.96\left(\frac{2}{\sqrt{n}}\right); \bar{x} + 1.96\left(\frac{2}{\sqrt{n}}\right)\right]$. Note que:

- $S_{\mathbf{x}}$ está contido no espaço parmétrico, que para esse caso é igual ao reais;
- $S_{\mathbf{x}}$ é um intervalo aleatório, pois seu centro varia de amostra para amostra;
- $S_{\mathbf{x}}$ só depende dos dados **x** e de constantes conhecidas.

Exemplo: Suponha que os seguintes dados constituem de duas amostras aleatórias com $n = 5$ de uma distribuição com média θ e variância 4.

Amostra 1	10.78	6.56	11.40	10.09	9.25	$\bar{x}$=9.62
Amostra 2	13.30	12.65	12.39	14.08	10.66	$\bar{x}$=12.61

Baseado na primeira amostra, temos o seguinte intervalo:

$$\left[9.62 - 1.96\left(\frac{2}{\sqrt{5}}\right); 9.62 + 1.96\left(\frac{2}{\sqrt{5}}\right)\right] = [7.86; 11.37]$$

Já de acordo com a segunda amostra, temos este outro intervalo:

$$\left[12.61 - 1.96\left(\frac{2}{\sqrt{5}}\right); 12.61 + 1.96\left(\frac{2}{\sqrt{5}}\right)\right] = [10.86; 14.37]$$

Nota-se, então, que como os dados foram gerados com $\theta = 10$, apenas o primeiro intervalo cobre o veradeiro valor do parâmetro. Isso está ligado ao conceito de confiança, a qual se refere ao método de construção do intervalo: apenas 5% deles não irão cobrir o verdadeiro valor do parâmetro. Portanto, ao gerarmos duas amostras uma delas caiu no caso dos 5% que não cobrem o verdadeiro valor de θ. Na prática, porém, não sabemos o verdadeiro valor de θ, pois é ele que queremos estimar, portanto não podemos distinguir entre os intervalos bons e ruins.

Vamos agora definir uma quantidade que será importante para compreensão de alguns conceitos posteriormente.

Pivô: Um pivô é uma função contínua $g(\mathbf{x}, \theta)$ cuja distribuição não depende de θ ou de qualquer outro parâmetro da distribuição de $\mathbf{x}$.

O usô de pivôs é necessário para possibilitar obter constantes k_1 e k_2 que não depdendam de θ, de forma que

$$P(k_1 \leq g(\mathbf{x}; \theta) \leq k_2) = 1 - \alpha \text{ para todo } \theta$$

Exemplos:

1. Se $X_1, X_2, ..., X_n$ são variáveis aletórias i.i.d. com distribuição $N(\mu, \sigma^2)$ com $\theta = (\mu, \sigma^2)$. Então
$$g(\mathbf{x}, \theta) = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \sim t_{(n-1)}$$
e, portanto, é um pivô.

2. No modelo de Regressão
$$(\hat{\beta} - \beta)^t \left(\frac{x^t x}{\sigma^2}\right) (\hat{\beta} - \beta) = g(\mathbf{Y}; \theta) \sim \chi^2_2$$
onde $\theta = (\beta_0, \beta_1, \sigma^2)$ e X é uma matriz de constantes conhecidas.

3.
$$\frac{(\mathbf{Y} - X\hat{\beta})^t (\mathbf{Y} - X\hat{\beta})}{(n-2)\sigma^2} \sim \chi^2_{n-2}$$

21.2 Intervalos de Confiança baseados no EMV

Já sabemos que se $\hat{\theta}$ é um estimador de máxima verossimilhança, então

$$\sqrt{I_n(\hat{\theta})}(\hat{\theta}_n - \theta) \xrightarrow{d} N(0,1)$$

Dessa maneira, se o tamanhoda amostra é grande, temos aproximadamente um pivô:

$$\underbrace{\sqrt{n}\sqrt{I_1(\theta)}(\hat{\theta}_n - \theta)}_{g(\mathbf{x};\theta)} \approx \underbrace{N(0,1)}_{\text{não depende de } \theta}$$

É fácil perceber que se $I_1(\theta)$ for constatnte em θ, então é fácil inverter o pivô aproximado para obter um intervalo de confiança. Em alguns casos, mesmo que $I_1(\theta)$ dependa de θ possível isolar o parâmetro para obter o intervalo, como mostra o exemplo a seguir:

Exemplo: Sejam $X_1, X_2, X_3, ..., X_n$ variáveis aleatórias i.i.d. com distribuição de $Poisson(\theta)$. Sabemos que o EMV de θ é dado por $\hat{\theta}_{EMV} = \bar{x}$ e $I_n(\theta) = \frac{n}{\theta}$. Então

$$\sqrt{\frac{n}{\theta}}(\bar{x} - \theta) \xrightarrow{d} N(0,1)$$

e portanto

$$P\left(\sqrt{\frac{n}{\theta}}|\bar{x}-\theta| \leq 1.96\right) \approx 0.95$$

Assim o intervalo

$$S_{\mathbf{x}} = \left\{\theta \text{ tais que } \sqrt{\frac{n}{\theta}}|\bar{x}-\theta| \leq 1.96\right\}$$

é aproximadamente um conjunto de confiança para θ (com coeficiente de 0.95).

Mas percebe-se que pode ser descrever $S_{\mathbf{x}}$ como

$$S_{\mathbf{x}} = \left\{\theta \text{ tais que } (\bar{x}-\theta)^2 \leq |\bar{x}-\theta| \leq (1.96)^2\frac{\theta}{n}\right\} = \left\{\theta \text{ tais que } \theta^2 - \left(2\bar{x}+\frac{(1.96)^2}{n}\right)\theta + \bar{x}^2 \leq 0\right\}$$

que equivale ao seguinte intervalo

$$\left[1/2\left(2\bar{x}+\frac{(1.96)^2}{n} - 1.96\sqrt{4\bar{x}+\frac{(1.96)^2}{n}}\right); 1/2\left(2\bar{x}+\frac{(1.96)^2}{n} + 1.96\sqrt{4\bar{x}+\frac{(1.96)^2}{n}}\right)\right]$$

Como em muitos casos não é fácil isolar o parâmetro como foi feito no exemplo anterior, vamos mostrar agora algumas alternativas tipicamente usadas para estimar $\hat{I}_n$.

A idéia básica de utilização dess estimativa é que se

$\hat{I}_n \approx I_n(\theta)$ eentão $\sqrt{\hat{I}_n}(\hat{\theta}-\theta) \xrightarrow{d} N(0,1)$

dessa maneira, se n é grande

$$P(\sqrt{\hat{I}_n})|(\hat{\theta}-\theta)| \leq 1.96) \approx 0.95 \text{ e assim } P(\hat{\theta} - \frac{1.96}{\sqrt{\hat{I}_n}} \leq \theta \leq \hat{\theta} + \frac{1.96}{\sqrt{\hat{I}_n}})$$

Desse modo,

$$S_{\mathbf{x}} = \left[\hat{\theta} - \frac{1.96}{\sqrt{\hat{I}_n}}; \hat{\theta} + \frac{1.96}{\sqrt{\hat{I}_n}}\right]$$

é aproximadamente um intervalo de confiança para θ.

As alternativas usadas para tal estimação, são as seguintes:

- Alternativa 1

$$\hat{I}_n = I_n(\hat{\theta}_{EMV}) = E\left[-\frac{\partial^2 l}{\partial\theta^2}\right]_{\hat{\theta}_{EMV}}$$

  - Exemplo 1: Sejam $X_1, X_2, ..., X_n$ variáveis aleatórias i.i.d com distribuição Poisson(θ). Então $\hat{\theta}_{EMV} = \bar{x}$ e $I_n(\theta) = \frac{n}{\theta}$, logo $\hat{I} = \frac{n}{\hat{\theta}_{EMV}} = \frac{n}{\bar{x}}$, que não depende de θ. Dessa forma, o intervalo de confinça é construído como

$$\left[\hat{\theta} - 1.96\sqrt{\frac{\bar{x}}{n}}; \hat{\theta} + 1.96\sqrt{\frac{\bar{x}}{n}}\right]$$

    Cabe notar ainda que $\hat{I}/n$ converge em probabilidade para $\frac{1}{\theta} = I_1(\theta)$.

– Exemplo 2: Sejam $X_1, X_2, ..., X_n$ variáveis aleatórias i.i.d com distribuição Binomial(θ). Então $\hat{\theta}_{EMV} = \bar{x}$ e $I_n(\theta) = \frac{n}{\theta(1-\theta)}$, logo $\hat{I} = \frac{n}{(1-\hat{\theta}_{EMV})\hat{\theta}_{EMV}} = \frac{n}{\bar{x}(1-\bar{x})}$, que não depende de θ. Dessa forma, o intervalo confiança pode ser calculado como

$$\left[\hat{\theta} - 1.96\sqrt{\frac{\bar{x}(1-\bar{x})}{n}}; \hat{\theta} + 1.96\sqrt{\frac{\bar{x}(1-\bar{x})}{n}}\right]$$

- Alternativa 2

$$\hat{I}_n = -\frac{\partial^2 l}{\partial \theta^2}|_{\hat{\theta}_{EMV}}$$

– Exemplo: Sejam $X_1, X_2, ..., X_n$ variáveis aleatórias i.i.d com distribuição Poisson(θ). Sabemos então que $\hat{\theta}_{EMV} = \bar{x}$ e $-\frac{\partial^2 l}{\partial \theta^2} = \frac{\sum x_i}{\theta^2}$, então

$$\hat{I}_n = -\frac{\partial^2 l}{\partial \theta^2}|_{\hat{\theta}_{EMV}} = \frac{\sum x_i}{\bar{x}^2} = \frac{n}{\bar{x}}$$

- Alternativa 3

$$\hat{I}_n = \sum_i \left(\frac{\partial l(x_i;\theta)}{\partial \theta}\right)^2 |_{\hat{\theta}_{EMV}} = \sum_i \left(\frac{\partial l_i)}{\partial \theta}\right)^2 |_{\hat{\theta}_{EMV}}$$

A justificativa para isso é a seguinte:

$$\frac{\partial l}{\partial \theta} = \sum_i \frac{\partial l_i}{\partial \theta} \text{ mas } E\left(\frac{\partial l}{\partial \theta}\right) = E\left(\frac{\partial l_i}{\partial \theta}\right) = 0$$

Então

$$I(\theta) = E\left(\frac{\partial l}{\partial l}\right)^2 = Var(\frac{\partial l}{\partial \theta}) = Var\left(\sum_i \frac{\partial l_i}{\partial \theta}\right) = \sum_i Var\left(\frac{\partial l_i}{\partial \theta}\right) = \sum_i E\left(\frac{\partial l_i}{\partial \theta}\right)^2$$

Assim

$$I_n(\theta) = E\left(\sum_i \left(\frac{\partial l_i}{\partial \theta}\right)^2\right) \text{ e espera-se que } I_n(\theta) \approx \sum_i \left(\frac{\partial l_i}{\partial \theta}\right)^2$$

– Exemplo: Sejam $X_1, X_2, ..., X_n$ variáveis aleatórias i.i.d com distribuição Poisson(θ). Sabemos que

$$\left(\frac{\partial l_i}{\partial l}\right)^2 = \left(\frac{x_i}{\theta} - 1\right)^2 \text{ logo } \left(\frac{\partial l_i}{\partial l}\right)_{\hat{\theta}_{EMV}} = \left(\frac{x_i}{\bar{x}} - 1\right)^2$$

Assim

$$\hat{I}_n = \sum_i \left(\frac{x_i}{\bar{x}} - 1\right)^2 = \frac{\sum(x_i - \bar{x})}{\bar{x}^2} = \frac{(n-1)S^2}{\bar{x}^2}$$

Então um Intervalo de Confianaça para θ de 95% é dado por

$$\left[\bar{x} - \frac{1.96}{\sqrt{\hat{I}_n}}; \bar{x} + \frac{1.96}{\sqrt{\hat{I}_n}}\right] = \left[\bar{x} - \frac{1.96\bar{x}}{\sqrt{n-1}S}; \bar{x} + \frac{1.96\bar{x}}{\sqrt{\sqrt{n-1}S}}\right]$$

Cabe notar que

$$\frac{\hat{I}_n}{n} \xrightarrow{d} \frac{Var(X)}{(E(X))^2} = \frac{\theta}{\theta^2} = \frac{1}{\theta} = I_1(\theta)$$

Vamos agora apresentar uma propriedade importante de intervalos de confiança. Se θ é um escalar e $g(\mathbf{x};\theta)$ é uma função contínua e monótona em θ para um $\mathbf{x}$ fixo então

$$P(k_1 \leq g(\mathbf{x};\theta) \leq k_2) = 1-\alpha \leftrightarrow P(c_1(\mathbf{x}) \leq \theta \leq c_2(\mathbf{x})) = 1-\alpha$$

pois para $\mathbf{x}$ fixo, $\{\theta; k_1 \leq g(\mathbf{x};\theta) \leq k_2\}$ é um intervalo, cujos pontos limites dependem de $\mathbf{x}$.

Exemplo Sejam $X_1, X_2, ..., X_n$ vairiáveis aleatórias i.i.d com distribuição $N(\mu, \sigma^2)$, então

$$g(\mathbf{x};\theta) = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1} \text{ para todo } \theta$$

onde $S^2 = \frac{\sum(x_i-\bar{x})^2}{n-1}$

Então

$$P(k_1 \leq g(\mathbf{x};\theta) \leq k_2) = 1-\alpha$$

logo

$$P(\chi^2_{\alpha/2} \leq \frac{(n-1)S^2}{\sigma^2} \leq \chi^2_{1-\alpha/2}) = 1-\alpha$$

então

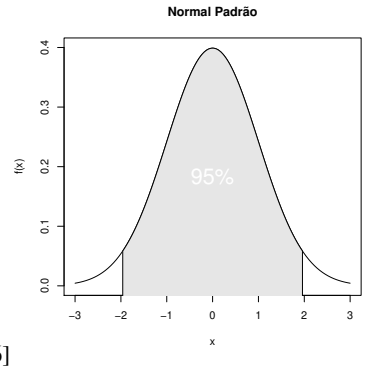
$$P\left(\frac{(n-1)S^2}{\chi^2_{1-\alpha/2}} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi^2_{\alpha/2}}\right) = 1-\alpha$$

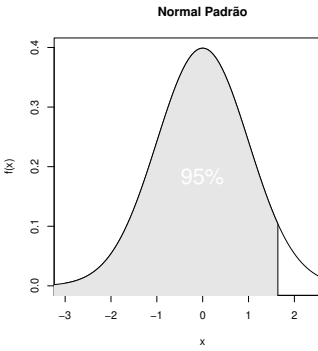
Suponha que $n = 5$ e as observações foram: 7.06, 5.82, 0.44, 11.54 e 8.85. Assim $S^2 = \frac{1}{5-1}\sum_i (x_i - \bar{x})^2 = 17.03$. Assumindo $\alpha = 0.5$ obtemos o seguinte intervalo:

$$[7.18; 95,94]$$

Para finalizar essa seção, uma outra observaçõe importante que devemos fazer a respeito dos intervalos de confiança é que eles não são únicos. Existem várias regiões que combrem a confiança desejada. O que buscamos, então, é aquele intervalo que cobre a região desejada com o menor comprimento possível. Para tanto, devemos manter dentro do intervalo os pontos com mais alta densidade.

A fim de exemplificar esse fato, vamos utilizar a distribuição normal, que, pelo fato de ser simétrica, tem como intervalo de menor comprimento aquele que é centrado na média. Sabemos que se quisermos um intervalo de 95% de uma normal padrão pegaremos a região que está compreendidade entre os percentis -1.96 e 1.96 como mostrado na Figura 21.2. Porém, a área entre os percentis -3.5 e 1.64 cobre a mesma região, de 97%. Por que escolhemos, então, a primeira delas? Pelo fato de nos fornecer um intervalo com menor comprimento e, portanto, mais preciso, como está ilustrado na Figura 21.2.





valo considerando os percentis -1.96 e 1.96] [Intervalo considerando os percentis -3.5 e 1.64]

Figure 21.2: Intervalos de confiança para a Normal Padrão com 95% de confiança

21.3 Caso Multivariado

Para o caso em que θ é um vetor, $S_{\mathbf{x}}$ é a região de confiança do espaço paramétrico do vetor, construída a partir dos dados da amostra $\mathbf{x}$.

Definition 21.3.1 Seja $S_x \in \Theta$ uma região construída a partir dos dados aleatórios $\mathbf{X}$. Se $P(\theta \in S_{\mathbf{x}}) = 1-\alpha$ então $S_{\mathbf{x}}$ é uma região de confiança para θ com coeficiente $1-\alpha$.

A Figura 21.3 apresenta regiões de confiança para o caso em que o parâmetro é um vetor de duas dimensões $\theta = (\theta_1, \theta_2)$ pertence a Θ (espaço paramétrico), que está contido em ^{2}.

A Figura 21.4 ilustra como as regiões de confiança dempendem da amostra observada. Cada uma das elipses corresponde à região de confiança referente a uma amostra distinta.

Exemlo: Sejam $X_1, X_2, ..., X_n$ variáveis aleatórias i.i.d. com dsitribuição $N(\theta, \sigma^2)$ onde σ^2 é desconhecido. Suponha que estamos interessados apenas no parâmetro θ, σ^2 nesse caso é chamado de parâmetro de ruído.

Sabemos que

$$\frac{\bar{x}-\theta}{s/\sqrt{n}} \sim t_{(n-1)}$$

onde s é o desvio padrã amostral.

Então $P\left(\left|\frac{\bar{x}-\theta}{s/\sqrt{n}}\right| \leq t_{(n-1);\alpha/2}\right) = 1-\alpha$.

Suponha que $n = 15$ e considere $\alpha = 0.05$, então, da tabela t, $t_{(n-1);\alpha/2} = t_{14;0.025} = 2.145$. Assim

$$P\left(\left|\frac{\bar{x}-\theta}{s/\sqrt{15}}\right| \leq 2.145\right) = 1-0.05 = 0.95 \text{ isto é} P\left(\left|\frac{\bar{x}-\theta}{s}\right| \leq 0.554\right) = 1-0.05 = 0.95$$

o que implica que

$$P(\bar{x}-0.554 \leq \theta \leq \bar{x}+0.554) = 0.95$$

ou ainda

$$P(\theta \in S_{\mathbf{x}}) = 0.95 \text{ onde } S_{\mathbf{x}}[\bar{x}-0.554s, \bar{x}+0.554s]$$

Antes de prosseguirmos, vamo relembrar propriedades importantes referentes à Algebra Matricial.

Uma A uma matrix simétrica é dita positiva definida se

$$\theta^t A\theta \geq 0 \text{ para todo } \theta \in^2$$

ou, equivalentemente, se seus autovalores são reais positivos.

Um vetor $\vec{V}$ é autovetor de A se

$$A\vec{V} = \lambda\vec{V}$$

sendo que λ é denomidade autovalor de A.

Isso significa que a matriz A, quando aplicada no vetor $\vec{V}$, simplismente o estica (se $\lambda > 1$), encolhe (se $0 < \lambda < 1$) ou inverte sua direção (se $\lambda < 0$).

Considere o ponto $\theta^* = (\theta_1^*, \theta_2^*)$ fixo em $\Theta \in^2$ e seja uma c constante positiva. Considere o conjunto dos pontos $\theta = (\theta_1, \theta_2)$ que satisfazem a desigualdade:

$$\left[\begin{array}{cc} \theta_1-\theta_1^* & \theta_2-\theta_2^* \end{array}\right] A \left[\begin{array}{c} \theta_1-\theta_1^* \\ \theta_2-\theta_2^* \end{array}\right] = \left[\ \theta-\theta^*\ \right]^t A \left[\ \theta-\theta^*\ \right] \leq c$$

Esse conjunto forma uma elipse centrada no ponto θ^*, com os eixos nas direções dos dois autovetores de A, com comprimento de cada um deles proporcional ao seus respecivos autovalores. Essa região está ilustrada na Figura 21.5

Se A é uma matriz de ordem três e $\theta^* = (\theta_1, \theta_2, \theta_3) \in^3$ então a forma quadrática

$$\left[\ \theta-\theta^*\ \right]^t A \left[\ \theta-\theta^*\ \right] \leq c$$

gera um elipsóide em ^{3}, que se encontra ilustrado na Figura 21.6.

21.3.1 Estimação de Intervalos baseados no EMV para o caso multivariado

Se o parâmetro $\theta \in^k$ e $\hat{\theta}$ é o EMV de θ então

$$\hat{\theta} \xrightarrow{d} N(\theta, I_n^{-1}(\theta))$$

onde $N(\theta, I_n^{-1}(\theta))$ é uma distribuição Normal Multivariada.

Então

$$(\hat{\theta} - \theta)^t I_n(\theta)(\hat{\theta} - \theta) \xrightarrow{d} \chi_k^2$$

e, novamente, o lado esquerdo é aproximadamente um pivô.

No caso univariado, às vezes, é possível inverter as desigualdades. Para o caso em que θ é multivariado, isto será impossível.

O que fazemos, então, assim como foi feito no caso univariado, é substituir $I_n(\theta)$ por uma estimativa $\hat{I}_n$. Nesse contexto, as alternativas são as seguintes:

- Alternativa 1

$$\underbrace{\hat{I}_n}_{kxk} = I_{n,kxk}(\theta)|_{\hat{\theta}_{EMV}}$$

- Alternativa 2

$$\underbrace{\hat{I}_n}_{kxk} = \frac{\partial^2 l}{(\partial\theta)^t(\partial\theta)}_{\hat{\theta}_{EMV}}$$

- Alternativa 3

$$\underbrace{\hat{I}_n}_{kxk} = \sum_{i=1}^{n} \frac{\partial l_i}{\partial\theta}\left(\frac{\partial l_i}{\partial\theta}\right)^t_{\hat{\theta}_{EMV}}$$

21.3.2 Caso da Normal Multivariada

Seja $\mathbf{W} = (W_1, W_2)^t$ um vetor de variáveis aleatórias que possuem distribuição Normal Bivariada com vetor de médias $\mu = (\mu_1, \mu_2)^t$ e a seguinte matriz de variância e covariância

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}$$

que é simétrica e positiva definida, o que implica que Σ^{-1} também apresenta tais características.

Percebe-se que o conjunto dos pontos $\mathbf{w} = (w_1, w_2)^t$ tais que

$$(\mathbf{w} - \mu)^t \Sigma^{-1} (\mathbf{w} - \mu) \leq c$$

forma uma elpse centrada em μ, com sua inclinação e grau de circularidade dependendo de ρ, σ_1 e σ_2. Essa região é ilustrada na Figura 21.7.

Considere, agora, a variável aleatória

$$D = (\mathbf{w} - \mu)^t \Sigma^{-1} (\mathbf{w} - \mu)$$

é possível mostrar que $D \sim \chi_2^2$. Para o caso genérico em que

$$\mathbf{W} = \begin{pmatrix} w_1 \\ w_2 \\ . \\ . \\ w_k \end{pmatrix} \text{ temos que } D \sim \chi^2_k$$

Pode-se mostrar ainda que se $\underbrace{\mathbf{w}}_{kx1}$ tem distribuição Normal K-variada com média μ e A é uma matriz mxk então

$$A\mathbf{w} \sim N_m\left(A\mu, A\Sigma A^t\right)$$

21.4 Intervalos de Confiança no Modelo de Regressão Linear

Vamos agora, mostrar como se constrem intervalos de confiança para os parâmetros estimados em um Modelo de Regressão Linear. Em um Modelo de Regressão Linear Simples modelamos a variável resposta da seguinte maneira

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

onde $x_1, x_2, ..., x_n$ são constantes fixas e conhecidas e $\varepsilon_1, \varepsilon_2, ..., \varepsilon_n$ são variáveis aleatórias i.i.d com distribuição $N(0, \sigma^2)$. Isso implica que $Y_1, Y_2, ..., Y_n$ são variáveis aleatórias independetes com a seguinte distribuição

$$Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$$

O que queremos estimar nesse caso é o vetor de parâmetros $\theta = (\beta_0, \beta_1, \sigma^2)$. Isso é feito por Máxima Verossimilhança da seguinte forma:

$$f(y_1, y_2, ..., y_n; \theta) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^2 exp(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2)$$

então o logarítmo da verossimilhança é dado por

$$l(\theta) = k - \frac{n}{2}log(\sigma^2) - \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$$

Essa maximização apresenta a seguinte solução algébrica:

$$\hat{\theta} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\sigma}^2 \end{pmatrix} = \begin{pmatrix} \bar{y} - \frac{s_{xy}}{s_{xx}}\bar{x} \\ \frac{s_{xy}}{s_{xx}} \\ \frac{1}{n}\sum_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \end{pmatrix}$$

onde $s_{xy} = \sum_i (x_i - \bar{x})(y_i - \bar{y})$ e $s_{xx} = \sum_i (x_i - \bar{x})^2$.
Denotando

$$X_{nx2} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ . & . \\ . & . \\ . & . \\ 1 & x_n \end{bmatrix} \quad Y_{nx1} = \begin{bmatrix} Y_1 \\ Y_2 \\ . \\ . \\ . \\ Y_n \end{bmatrix} \quad \varepsilon_{nx1} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ . \\ . \\ . \\ \varepsilon_n \end{bmatrix} \quad \beta_{2x1} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$$

Então o modelo linear pode ser escrito da seguinte forma

$$Y = X\beta + \varepsilon$$

e o Estimador de Máxima Verossimilhança pode ser escrito como

$$\hat{\theta} = \begin{pmatrix} \hat{\beta} \\ \hat{\sigma}^2 \end{pmatrix} = \begin{bmatrix} (X^tX)^{-1}X^tY \\ \frac{1}{n}||\mathbf{Y} - X\hat{\beta}||^2 \end{bmatrix}$$

Percebe-se que, como $\mathbf{Y}$ tem uma distribuição Normal n-variada com média $X\beta$ e variância $\sigma^2 I_n$, então o vetor de estimadores

$$\begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} = \underbrace{(X^tX)^{-1}X^t}_{2xn}\mathbf{Y} = \underbrace{A}_{2xn}\underbrace{\mathbf{Y}}_{nx1}$$

tem distribuição Normal Bivariada, cujo valor esperado é dado por

$$E(A\mathbf{Y}) = AE(\mathbf{Y}) = AX\beta = (X^tX)^{-1}X^tX\beta = \beta$$

e a seguinte matrix se covariância

$$Var(A\mathbf{Y}) = AVar(\mathbf{Y})A^t = A\sigma^2 I_n A^t = (X^tX)^{-1}X^t\sigma^2 I_n X(X^tX)^{-1} = \sigma^2(X^tX)^{-1}$$

ou seja

$$\hat{\beta} \sim N_2(\beta, \sigma^2(X^tX)^{-1})$$

Portanto, no modelo de Regressão Linear

$$D = (\hat{\beta} - \beta)^t \left(\frac{X^tX}{\sigma^2}\right)(\hat{\beta} - \beta) \sim \chi^2_2$$

Consultando uma tabela da Distibuição Qui-Quadrado com dois graus de liberdade, nota-se que o percentil que deixa 5% de áre acima dele é o 5.99, como mostrado na Figura 21.8.

Dessa forma, se σ^2 for desconhecido

$$P\left((\hat{\beta} - \beta)^t \left(\frac{X^tX}{\sigma^2}\right)(\hat{\beta} - \beta) \leq 5.99\right) = 0.95$$

Vamos definir a seguinte região aleatória, que apresenta o formato de uma elipse

$$S_{\mathbf{Y}} = \left\{\mathbf{v} \in^2 \text{ tais que} (\hat{\beta} - \beta)^t \left(\frac{X^tX}{\sigma^2}\right)(\hat{\beta} - \beta) \leq 5.99\right\}$$

Então $S_{\mathbf{Y}}$ não depende de β quando σ^2 é conhecido, ou seja

$$P(\beta \in S_{\mathbf{y}}) = 0.95 \text{ para todo } \beta$$

Isso significa que $S_{\mathbf{y}}$ é a região de confiança de 0.95 para β, quando se sabe o valor de σ^2.

21.5 Método Delta Univariado e Multivariado

def

21.6 Quando $[I^{-1}]_{ii} \neq 1/I_{ii}$

perda de info

21.7 Informacao observada e esperada

Qual e' melhor?

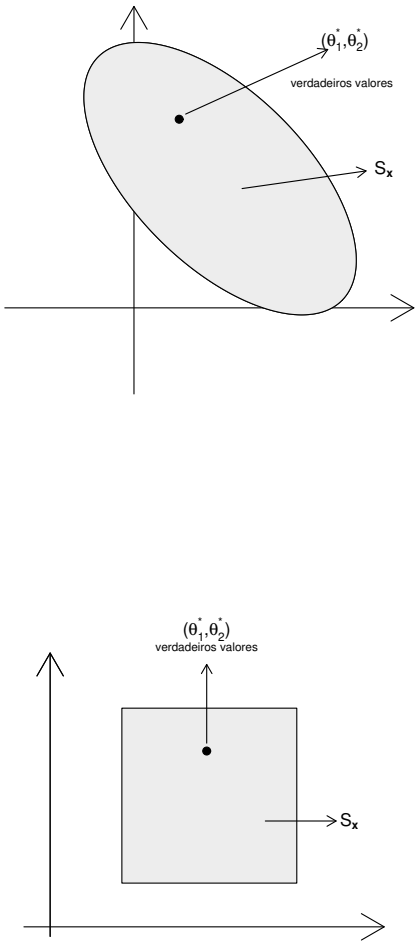


Figure 21.3: Região de Confiança para o caso bivariado

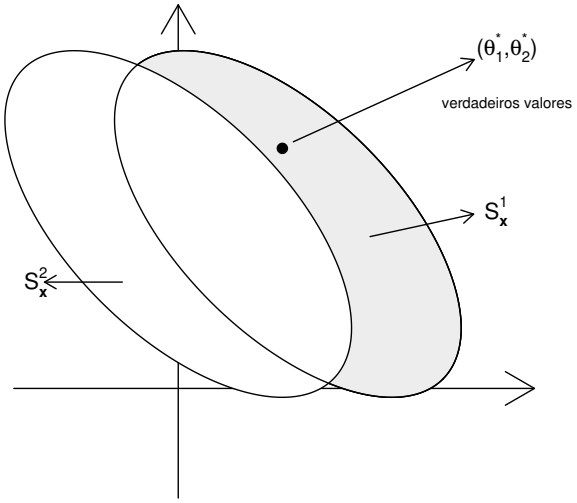


Figure 21.4: Região de Confiança para o caso bivariado considerando-se duas amostras distintas

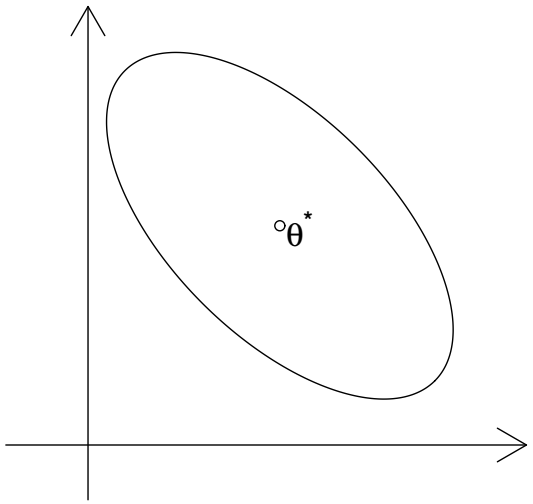


Figure 21.5: Ilustração da região de elipse

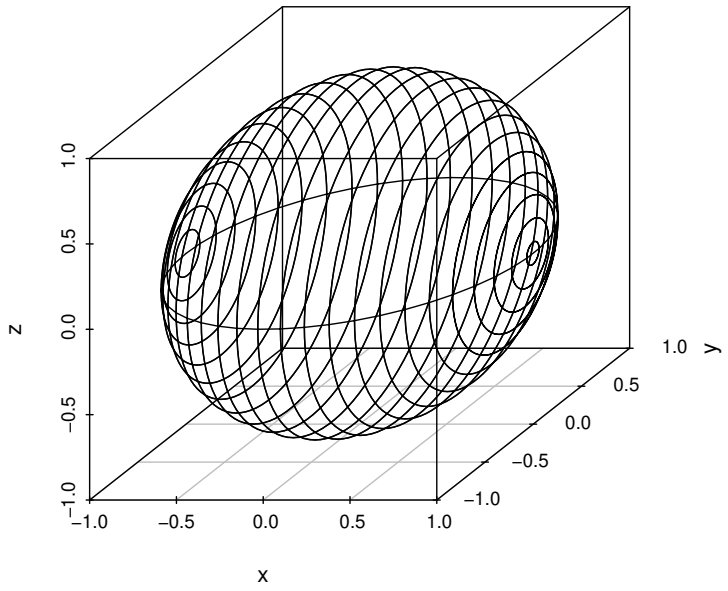


Figure 21.6: Elipsóide

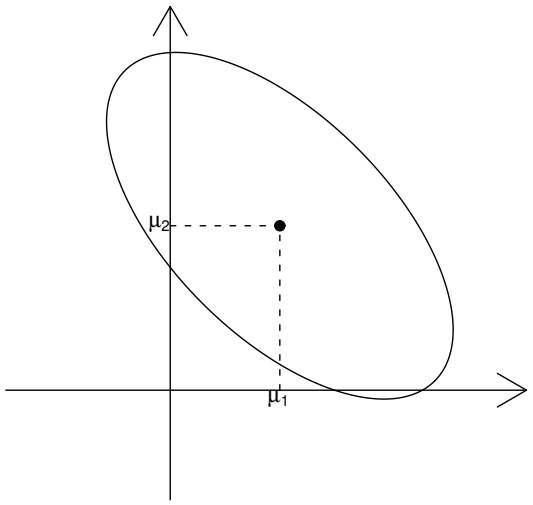


Figure 21.7: Elipse

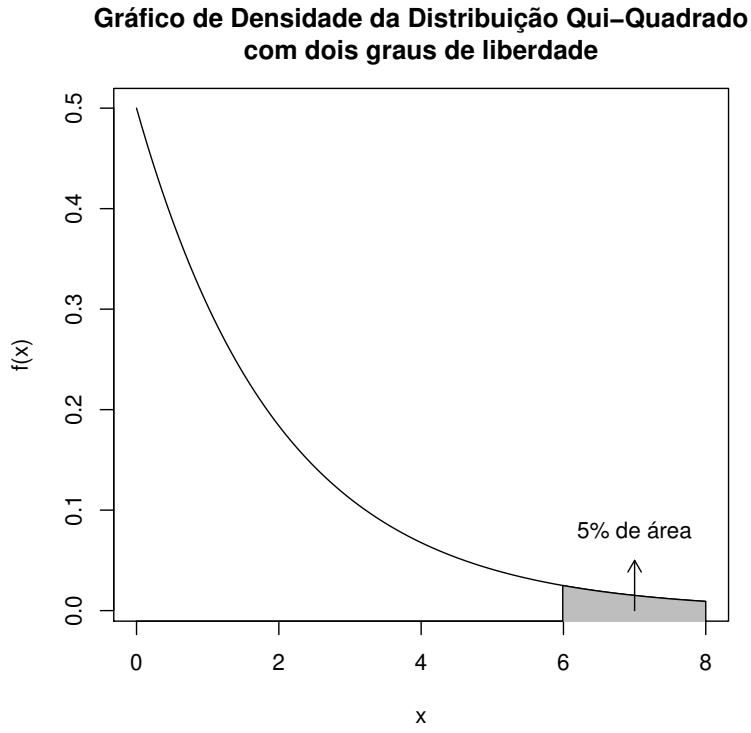


Figure 21.8: Gráfico de Densidade da Qui-Quadrado com dois graus de liberdade

22. Mixture Models

22.1 Misturas de distribuições: introdução

Os modelos básicos de distribuições que dispomos são flexíveis mas não dão conta de tudo que ocorre. Será raro que um conjunto de instâncias seja muito bem modelado por uma das poucas distribuições que aprendemos até agora. Temos duas alternativas:

- Aumentar o nosso "dicionário de distribuições" criando uma looooooooonga lista de distribuições para ajustar aos dados reais.
- Misturar os tipos básicos já definidos criando tipos compostos que ampliam a classe de distribuições disponíveis para análise.

Uma forma de misturar distribuições é construir um modelo de regressão: Cada i-ésimo indivíduo tem uma distribuição Y_i (como a gaussiana, por exemplo) que é modulada pelas suas variáveis independentes ou *features* $\mathbf{x}_i$. Essas variáveis independentes controlam o valor esperado $\mathbb{E}(Y_i) = \mathbf{x}_i'\beta$. Assim, a coleção de v.a.'s $Y_1, \ldots, Y_n$ é uma mistura de diferentes gaussianas, cada uma tendo um valor esperado próprio.

Vai ser comum não termos covariáveis para usarmos como *features* mas ainda assim termos claramente dados vindos de 2 ou mais distribuições. Também será comum nos modelos de fatores latentes que, mesmo usando as *features* para controlar as distribuições, ainda tenhamos que usar mais misturas de distribuições.

Um exemplo clássico em aprendizagem de máquina é o de dados de erupções da geyser Faithful do Parque Yellowstone nos EUA, mostrados na Figura 22.1. Cada ponto representa uma erupção. No eixo horizontal, temos a duração de cada erupção. No eixo vertical, temos o intervalo entre a erupção em questão e a erupção seguinte. Parece que existem duas nuvens elípticas de pontos, indicando que estes dados podem vir de duas normais bivariadas misturadas.

Para entender e dominar as ferramentas para lidar com misturas de distribuições, vamos retornar para o caso de v.a.'s unidimensionais gaussianas. Vamos exemplificar o que é a mistura de três gaussianas. A Figura 22.2 mostra a densidade de probabilidade das gaussianas $N(0,1)$, $N(3,0.5^2)$ e $N(10,1)$ e a densidade de sua mistura nas respectivas proporções 0.6, 0.1 e 0.3. Note como podemos visualizar as densidades originais através dos picos remanescentes na densidade da mistura. Note também que a densidade misturada com a menor proporção, a gaussiana $N(3,0.5^2)$ aparece com o

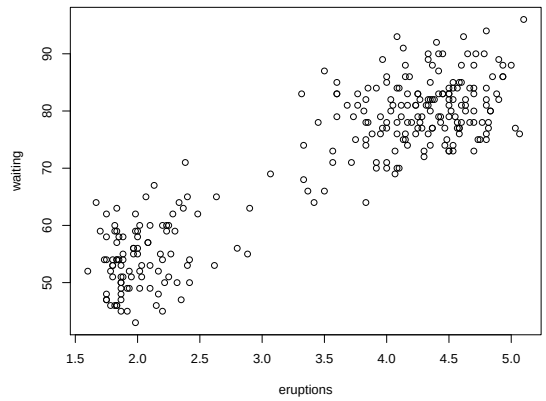


Figure 22.1: Dados de $n = 272$ erupções da geyser Faithful do Parque Yellowstone nos EUA. No eixo horizontal, a duração de cada erupção. No eixo vertical, temos o intervalo entre a erupção em questão e a erupção seguinte. Parece que existem duas normais bivariadas misturadas.

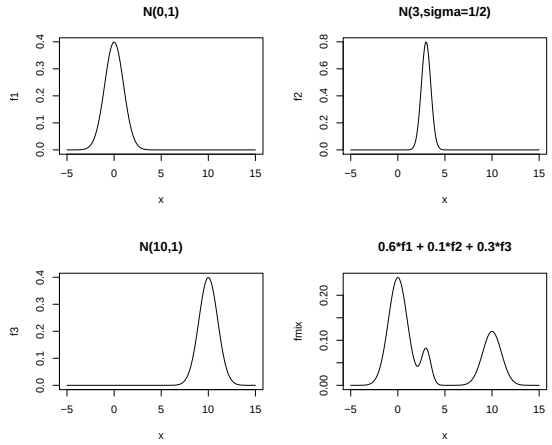


Figure 22.2: Mistura de 3 normais $N(0,1)$, $N(3,0.5^2)$ e $N(10,1)$ nas proporções 0.6, 0.1 e 0.3, respectivamente.

menor pico na densidade da mistura. Antes de cermos como esta densidade da mistura é obtida, vamos mostrar na Figura 22.3 uma amostra de v.a.'s i.i.d vindas dessa distribuição de mistura. Na mesma Figura, mostramos a densidade da mistura sobreposta ao histograma padronizado. Desse modo, quando a msitura fo relativamente simples como nesse caso, poderemos visualmente reconhecer a presença das distribuições componentes.

O caso de uma mistura de v.a.'s discretas é similar. Xiao et al (1999) estudam o período de internação em hospitais para modelagem de custos. Um tipo de financiamento de custos de saúde bastante usado mundo afora é baseado no DRG ?? Explicar ?? Considerando apenas as internações devido a partos...?? É bastante conhecido que partos por cesáriana levam tipicamente a tempos mais longos de internação que partos naturais.

Hospital-maternaidade na Austrália. The total sample size is 5648. The data, based on separation dates, are available from July 1992 to September 1996. Patients' socio-economic characteristics (age, gender, Aboriginality, marital status, country of birth, occupation, employment and insurance status), health provision factors (mode of separation, accommodation status, admission type, source

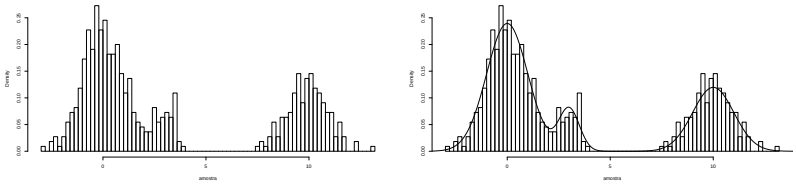


Figure 22.3: Amostra de $n = 550$ dados vindos da densidade mistura de 3 normais. A mesma amostra com a densidade da mistura sobreposta.

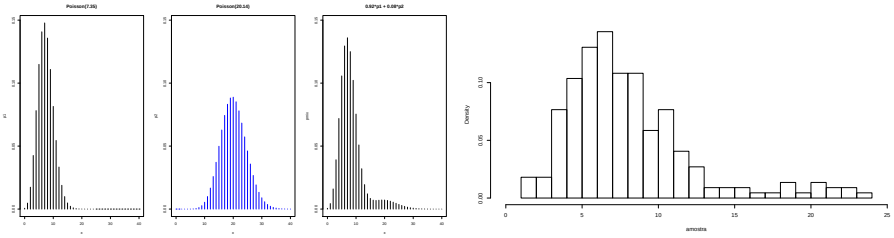


Figure 22.4: Esquerda: Mistura com 92% vindo de uma Poisson($\lambda = 7.35$) e os outros 8% vindo de uma Poisson($\lambda = 20.1$). Amostra de $n = 222$ casos de parto cesáreo com complicações graves. Dados adaptados de Xiao et. al. (1999). Mistura: 92% vêm de uma Poisson($\lambda = 7.35$) e os outros 8% vêm de Poisson($\lambda = 20.1$).

of referral, distance from the hospital and treating medical officer) and other relevant factors (number of diagnoses, number of procedures and number of inpatient theatre attendance) were reviewed and selected from the main database. These were considered as potential factors influencing LOS based on our previous study. Only significant factors were included in the final models.

A Figura 22.4

22.2 Misturas de distribuições: formalismo

Vamos considerar inicialmente o caso contínuo. Estamos olhando um atributo Y. Suponha que temos três sub-populações: 1, 2 e 3. Represente as medições nas diferentes sub-populações como v.a.'s Y_1, Y_2, e Y_3. As sub-populações sao diferentes e isto implica que as v.a.'s têm densidades diferentes. As densidades são: $f_1(y)$, $f_2(y)$ e $f_3(y)$ e as respectivas distribuições acumuladas são $F_1(y)$, $F_2(y)$ e $F_3(y)$. Assim, $F_1'(y) = f_1(y)$, $F_2'(y) = f_2(y)$ e $F_3'(y) = f_3(y)$. Para fixar as ideias, considere mentalmente o caso da Figura 22.2 em que a população 1 seguia uma $N(0,1)$, a densidade 2 era $N(3,1/2^2)$ e a população 3 era uma $N(10,1)$.

A variável realmente observada é representada por Y. Qual a distribuição de probabilidade da v.a. Y? Se o indivíduo vier da população 1, Y terá a mesma distribuição que a v.a. Y_1. Se vier da população 2, $Y \sim Y_2$, e se vier da população 3, $Y \sim Y_3$. O indivíduo da população mistura vem aleatoriamente de *uma* das três populações. Ele vem das 3 populações com as seguintes probabilidades:

- vem da população 1 com probabilidade θ_1;
- vem da população 2 com probabilidade θ_2;
- vem da população 3 com probabilidade θ_3

com $\theta_1 + \theta_2 + \theta_3 = 1$ pois só existem estas três alternativas de onde Y é extraída.

Assim, a medição Y tem a seguinte estrutura aleatória: Y tem a mesma distribuição que Y_k com probab θ_k ou, de forma mais compacta:

- $Y \sim Y_1$ com probabilidade θ_1

- $Y \sim Y_2$ com probabilidade θ_2
- $Y \sim Y_3$ com probabilidade θ_3

Qual a densidade de Y? Usamos a fórmula da probabilidade total para calcular a função de distribuição acumulada $\mathbb{F}(y) = \mathbb{P}(Y \leq y)$. Vamos condicionar no resultado de qual população Y foi amostrada e a seguir somamos (de forma ponderada) sobre as três possíveis populações. Temos

$$\mathbb{F}(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(Y \leq y \text{ e vem de alguma pop})$$

Temos a igualdade de eventos

$$\begin{aligned}[Y \leq y] &= [Y \leq y \cap \text{ vem de pop 1}] \cup [Y \leq y \cap \text{ vem de pop 2}] \\ &\quad \cup [Y \leq y \cap \text{ vem de pop 3}]\end{aligned}$$

Como os eventos são disjuntos, a probab da sua união é a soma das probabilidades:

$$\begin{aligned}\mathbb{F}(y) &= \mathbb{P}(Y \leq y \cap \text{ vem de pop 1}) + \mathbb{P}(Y \leq y \cap \text{ vem de pop 2}) + \mathbb{P}(Y \leq y \cap \text{ vem de pop 3}) \\ &= \mathbb{P}(Y \leq y|\text{pop 1})\mathbb{P}(\text{pop 1}) + \mathbb{P}(Y \leq y|\text{pop 2})\mathbb{P}(\text{pop 2}) + \mathbb{P}(Y \leq y|\text{pop 3})\mathbb{P}(\text{pop 3}) \\ &= \mathbb{P}_1(Y \leq y)\theta_1 + \mathbb{P}_2(Y \leq y)\theta_2 + \mathbb{P}_3(Y \leq y)\theta_3 \\ &= \mathbb{F}_1(y)\theta_1 + \mathbb{F}_2(y)\theta_2 + \mathbb{F}_3(y)\theta_3\end{aligned}$$

Assim, descobrimos que a função de distribuição acumulada $\mathbb{F}(y)$ é uma média ponderada das dist acumuladas $\mathbb{F}_i(y)$ das componentes da mistura. Outra maneira de dizer isto é: a distribuição acumulada da mistura é a mistura das distribuições acumuladas.

A distribuição acumulada não é muito intuitiva. A densidade é mais interpretável por sua ligação com os histogramas dos dados observados. Se temos a distribuição acumulada, podemos obter a densidade de Y derivando $\mathbb{F}(y)$:

$$\begin{aligned}f(y) &= \mathbb{F}'(y) = \mathbb{F}_1'(y)\theta_1 + \mathbb{F}_2'(y)\theta_2 + \mathbb{F}_3'(y)\theta_3 \\ &= f_1(y)\theta_1 + f_2(y)\theta_2 + f_3(y)\theta_3\end{aligned}$$

Assim, a densidade da mistura Y é a mistura das densidades das componentes Y_1, Y_2 e Y_3.

Reveja a Figura 22.2 para enxergar esta fórmula. O código R para obtê-la é o seguinte:

```
x <- seq(-5, 15, by=0.01)
f1 <- dnorm(x)  # densidade N(0,1) nos pontos de x
f2 <- dnorm(x, 3, 1/2) # densidade de N(mu=3, sigma=1/2)
f3 <- dnorm(x, 10, 1)
fmix <- 0.6*f1 + 0.1*f2 + 0.3*f3
par(mfrow=c(2,2))
plot(x, f1, type="l"); title("N(0,1)")
plot(x, f2, type="l"); title("N(3,sigma=1/2)")
plot(x, f3, type="l"); title("N(10,1)")
plot(x, fmix, type="l"); title("0.6*f1 + 0.1*f2 + 0.3*f3")
```

Se quisermos gerar uma amostra de tamanho $n = 550$ da mistura de três normais, usamos um algoritmo muito simples que reproduz o conceito de mistura:

```
for(i in 1:550){
    Selecione a pop k = 1, 2 ou 3 com probabs p1, p2, p3
    Y = um valor da normal da pop k
}
```

O script R correspondente é o seguinte:

```
## gerando amostra da mistura (n=550)
## 3 subpops normais, probabs = c(0.6, 0.1, 0.3)
## numero de cada subpop
num <- rmultinom(n=1, size=550, prob=c(0.6, 0.1, 0.3))
num # gerou (321, 56, 173)
amostra <- c(rnorm(num[1]), rnorm(num[2], 3, 1/2), rnorm(num[3], 10, 1))
```

No caso de v.a.'s discretas, os resultados são os mesmos do caso contínuo. Suponha que Y seja uma mistura de três v.a.'s discretas: Y_1, Y_2, Y_3 (por exemplo, 3 Poissons) As 3 v.a.'s têm distribuições acumuladas $\mathbb{F}_k(y)$ e função de probabilidade $p_k(y) = \mathbb{P}(Y_k = y)$ para $k = 1,2,3$. Então, a distribuição acumulada da mistura Y é dada por

$$\mathbb{F}(y) = \mathbb{P}(Y \leq y) = \mathbb{F}_1(y)\theta_1 + \mathbb{F}_2(y)\theta_2 + \mathbb{F}_3(y)\theta_3\,,$$

idêntico ao caso contínuo. A função de massa de probabilidade é dada por

$$\begin{aligned} p(y) &= \mathbb{P}(Y = y) \\ &= \mathbb{F}(y) - \mathbb{F}(y-1) \\ &= p_1(y)\theta_1 + p_2(y)\theta_2 + p_3(y)\theta_3 \end{aligned}$$

22.2.1 Misturas de normais multivariadas

Voltemos aos dados de erupção do geyser Faithful mostrados na Figura 22.1. Aparentemente temos duas normais bivariadas misturadas nestes dados. Olhando os dados, podemos chutar grosseiramente os valores dos parâmetros de cada componente. Considerando o componente 1, no canto inferior esquerdo do gráfico, podemos chutar que o vetor de valores esperados é $\mu_1 = (\mu_{11}, \mu_{12}) = (2.1, 52)$ e a matriz de covariância é

$$\Sigma_1 = \begin{bmatrix} \sigma_{11}^2 & \rho_1\sigma_{11}\sigma_{12} \\ \rho_1\sigma_{11}\sigma_{12} & \sigma_{12}^2 \end{bmatrix} = \begin{bmatrix} (0.25)^2 & 0.3\sigma_{11}\sigma_{12} \\ 0.3\sigma_{11}\sigma_{12} & 4^2 \end{bmatrix}$$

Para o componente 2, no canto superior direito do gráfico, temos $\mu_2 = (\mu_{21}, \mu_{22}) = (4.5, 80)$ e a matriz de covariância:

$$\Sigma_2 = \begin{bmatrix} \sigma_{21}^2 & \rho_2\sigma_{21}\sigma_{22} \\ \rho_2\sigma_{21}\sigma_{22} & \sigma_{22}^2 \end{bmatrix} = \begin{bmatrix} (0.35)^2 & 0.7\sigma_{21}\sigma_{22} \\ 0.7\sigma_{21}\sigma_{22} & 5^2 \end{bmatrix}$$

A proporção do componente 1 pode ser estimada grosseiramente em 35% ou $\theta_1 = 0.35$

Cmo no caso univariado, a densidade conjunta do vetor bivariado $\mathbf{Y} = (Y_1, Y_2)$ é uma mistura de duas densidades gaussianas bivariadas:

$$f(\mathbf{y}) = f(y_1, y_2) = \theta_1 f_1(y_1, y_2) + \theta_2 f_2(y_1, y_2)$$

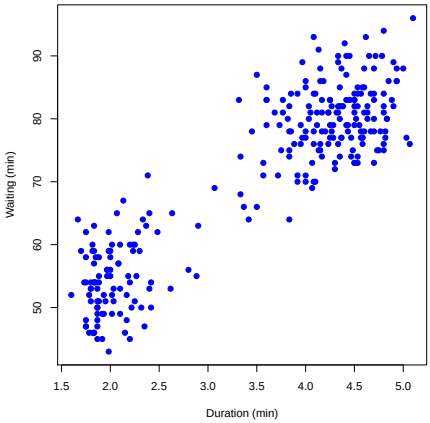
onde $\theta_1 + \theta_2 = 1$ com $\theta_1 \geq 0$ e $\theta_2 \geq 0$ e com $f_1(y_1, y_2)$ sendo a densidade do componente 1 (uma normal bivariada) e $f_2(y_1, y_2)$ sendo a densidade do componente 2 (também uma normal bivariada).

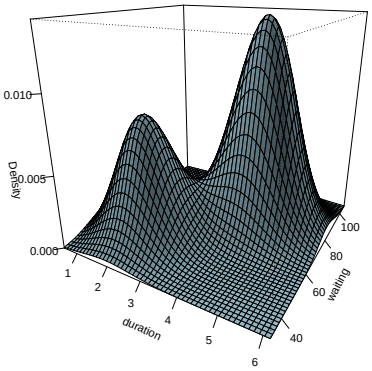
A Figura 22.2.1 mostra a densidade da mistura e a amostra que seria gerada por ela.

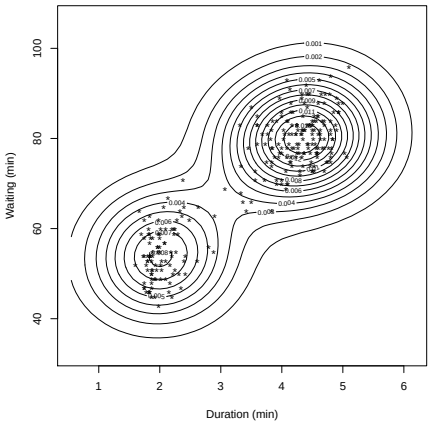
Uma outra visualizacão da densidade da mistura está na Figura 22.2.1.

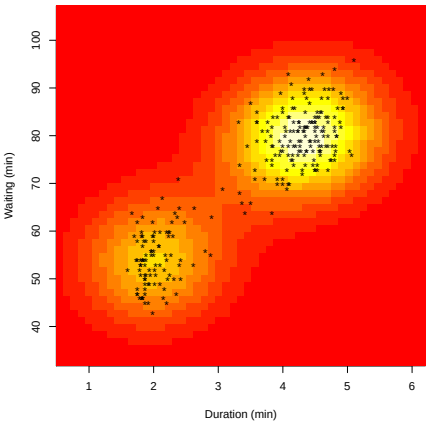
A Figura 22.2.1 mostra a densidade da mistura de 3 normais bivariadas. O código R para esta figura está abaixo.

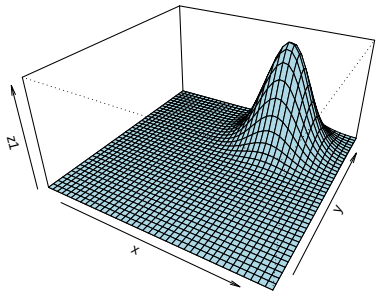
```
par(mfrow=c(2,2), mar=c(0,0,0,0))
x <- seq(-5, 5, length= 40); y <- x
f <- function(x,y) { dnorm(x,2,1)*dnorm(y,3,1) }
z1 <- outer(x, y, f)
```

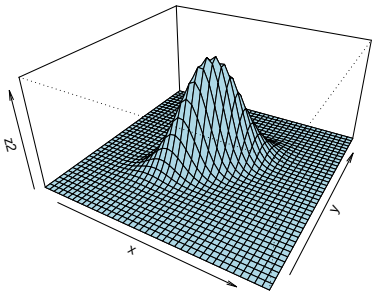

Waiting (min)
Duration (min)

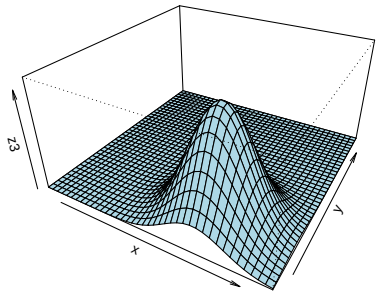

Density
duration
Waiting

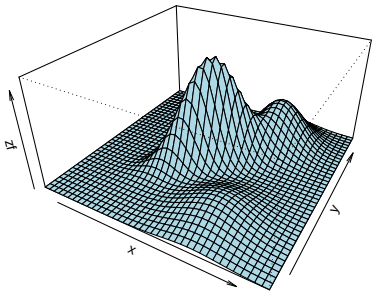

Waiting (min)
Duration (min)


Waiting (min)
Duration (min)


z1
x
y


z2
x
y


z3
x
y


z4
x
y

```
persp(x, y, z1, theta = 30, phi = 30, expand = 0.5, col = "lightblue")

# densidade normal bivariada com rho=0.7
f <- function(x,y, rho=0.7){exp(-(x^2 - 2*rho*x*y + y^2)/(2*(1-rho^2)))/(2*pi*sqrt(1-rho^
z2 <- outer(x, y, f)
persp(x, y, z2, theta = 30, phi = 30, expand = 0.5, col = "lightblue")

f <- function(x,y) { dnorm(x,2,1.2)*dnorm(y,-3,1.2) }
z3 <- outer(x, y, f)
persp(x, y, z3, theta = 30, phi = 30, expand = 0.5, col = "lightblue")

zf <- 0.3*z1 + 0.5*z2 + 0.2*z3
persp(x, y, zf, theta = 30, phi = 30, expand = 0.5, col = "lightblue")
```

23. EM Algorithm

23.1 Estimando uma distribuição de mistura

O algoritmo para gerar dados de uma mistura é simples:

- Input: Número de grupos k
- Input: Densidade de cada grupo: $f_1(\mathbf{y}), f_2(\mathbf{y}), \ldots, f_k(\mathbf{y})$
- Input: proporções de cada grupo: $\theta_1, \theta_2, \ldots, \theta_k$
- Gerar amostra de mistura $\theta_1 f_1(\mathbf{y}) + \ldots \theta_k f_k(\mathbf{y})$ de k componentes: fácil.
- Passo 1: Escolha uma das k componentes ao acaso com probabilidades $\theta_1, \ldots \theta_k$.
- Passo 2: Selecione Y da distribuição $f_i(\mathbf{y})$ da componente i selecionada no passo anterior.

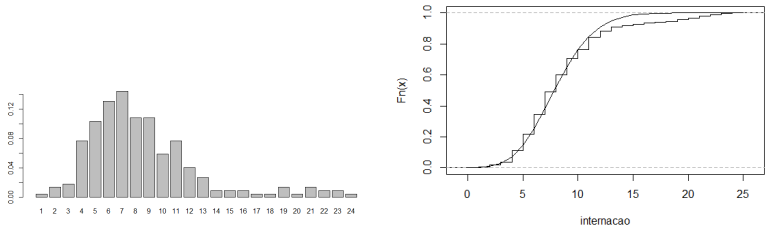
Isto é, dado o mecanismo (o modelo) aleatório) de uma mistura, podemos gerar dados sintéticos. Mas o problema *realmente* relevante é o contrário. Como ajustar um modelo de mistura a dados estatísticos? Isto é, recebemos os dados e queremos inferir qual o modelo que foi usado para gerá-los. Esta tarefa não é tão simples. Primeiro, precisamos saber quantos componentes estão presentes. Vamos discutir isto mais tarde. Neste ponto, vamos supor que o número de componentes é conhecido e é igual K. A outra peça importante é o tipo de distribuição presente na mistura: gaussiana, gama, Poisson, Weibull? Vamos também supor que os possíveis tipos vêm de uma mesma família de distribuições e que esta família é conhecida. Voltaremos a discutir estes dois pontos mais a frente. Assim, se (e este é um grande se)

- soubermos o número K de componentes na mistura (digamos, $K = 3$)
- soubermos a classe da distribuição de probabilidade de cada componente (digamos, normal).

então poderemos usar o algoritmo EM para ajustar o modelo de mistura. A seleção de k é feita via técnicas de escolha de modelos: ajustamos vários modelos com diferentes k e escolhemos o "melhor". Veremos seleção de modelos mais tarde neste curso.

23.2 Dados e rótulos

Nem sempre é fácil obter o MLE devido a dificuldades na otimização da função de verossmilhança. Um problema difícil é quando temos variáveis latentes ou ocultas (hidden or latent states). Por exemplo, nos problemas de mistura que aparecem em diversas áreas como análise de imagens, de



textos, etc. Uma das maneiras de estimar os parâmeros num modelo de análise fatorial é através do algoritmo EM. Vamos começar o estudo do algoritmo EM em problemas simples de misturas.

Suponha que estamos analisando o número de dias de internação de 222 mulheres após um parto cesáreo com complicações. O resultado está no lado esquerdo da Figura 23.2. A cauda estende-se por uma faixa muito longa para vir de uma única Poisson. De fato, se supusermos que uma única Poisson(λ) gerou estes dados, vamos usar a média amostral igual a $\bar{y} = 8.51$ como estimador de λ. Podemos fazer um gráfico da distribuição acumulada empírica $\mathbb{F}_n(y)$ junto com o da distribuição acumulada teórica de uma Poisson($\lambda = 8.51$). O resultado está no lado direito da Figura 23.2. Pela curva teórica deveríamos ter quase toda a amostra abaixo de $y = 15$. Entretanto, a curva empírica mostra que ainda existem vários dados acima desse valor.

```
mean(amostra)
Fn <- ecdf(amostra)
plot(Fn, verticals= T, do.p=F, main="", xlab="internacao")
x <- 0:25
y <- ppois(x, mean(amostra))
lines(x,y)
```

Como a distribuição é discreta, o teste de Kolmogorov não é válido. Usamos então o teste qui-quadrado.Para escolher as classes, devemos ter pelo menos 5 observaçõesem cada classe para que a distribuição assintótica do teste qui-quadrado seja uma boa aproximação.

```
> table(amostra)
amostra
 1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24
 1  3  4 17 23 29 32 24 24 13 17  9  6  2  2  2  1  1  3  1  3  2  2  1
```

Vamos agrupar as contagens iniciais criando uma classe $[Y \leq 3]$, as classes $[Y = 4], \ldots, [Y = 13]$, e as classes $[14 \leq Y19]$ e $[Y \geq 20]$. Assim, terminamos com $m = 13$ classes indexadas por $j = 1, \ldots, 13$ e com as contagens C_j. As probabilidades p_j de uma Poisson($\lambda = 8.51$) em cada uma dessas categorias é facilmente obtida em R e assim as contagens esperadas $E_j = 222 p_j$. Com isto a estatística do teste qui-quadrado

$$X^2 = \sum_j \frac{(C_j - E_j)^2}{E_j} \tag{23.1}$$

é calculada junto com seu p-valor.

```
> x = table(amostra)
> esp = 222*c(ppois(3,8.51), dpois(4:13, 8.51), ppois(19,8.51)-ppois(13, 8.51), 1 - ppois(19, 8.51))
> obs = c(sum(x[1:3]), x[4:13], sum(x[14:19]), sum(x[20:24]))
> xisq =  sum( (obs - esp)^2/esp )
```

```
> xisq
[1] 674.7081
> 1-pchisq(xisq, 13-2)
[1] 0
> round((obs - esp)/sqrt(esp),2)
 0.53  2.31  1.56  1.11  0.62 -1.18 -0.90 -2.33 -0.46 -1.22 -0.95 -0.11 25.59
```

O p-valor é aproximadamente igual a zero e portanto, uma (única) distribuição de probabilidade não é adequada para modelar estes dados. O último comando acima mostra os desvios $(C_j - E_j)/\sqrt{E_j}$ usados (ao quadrado) na estatística qui-quadrado X^2 em (23.1). Observe que, comparado com o esperado sob uma única Poisson, os desvios qui-quadrado mostram que existe um excesso substancial na última categoria, com as contagens mais altas.

Com esta motivação, vamos partir para uma mistura de duas Poissons, uma delas servindo para capturar a pequena parcela de mulheres que ficam um tempo relativamente bastante longo após o parto cesáreo com complicações. Vamos assumir que uma proporção α dos dados vem de uma Poisson com parâmetro λ_a. A proporção $1-\alpha$ restante vem de uma Poisson com parâmetro λ_b. Queremos inferir de maneira automática sobre $\theta = (\lambda_a, \lambda_b, \alpha)$. Como fazer isto?

A partir do gráfico na Figura 23.2, podemos conjecturar que temos uma Poisson($\lambda_a \approx 2$) e uma Poisson($\lambda_b \approx 10$). Queremos o melhor estimador possível, o MLE. Seria muito fácil obter ese MLE se soubéssemos a qual grupo cada observação pertence. Neste caso, bastaria ajustar uma Poisson separadamente a cada um dos dois grupos de dados. O MLE do parâmetro λ de uma Poisson simples é a média aritmética dos dados. Infelizmente não sabemos isto: observamos apenas os dados numéricos, e não sua classe.

Mas como seria no caso em que conhecêssemos os rótulos dos grupos? Se soubéssemos, o vetor de dados com a informação completa, da contagem e do rótulo do grupo, pode ser representado por

$$(\mathbf{y}, \mathbf{z}) = (y_1, \ldots, y_{222}, z_1, \ldots, z_{222})$$

onde y_i é a contagem da mulher i e z_i é o rótulo do seu grupo, com λ_a ou com λ_b. Temos $z_i = 0$ se a i-ésimo mulher for do grupo que se interna menos e portanto a y_i contagem vem de uma Poisson com parâmetro λ_a. Se $z_i = 1$ então a i-ésima mulher é do grupo 2 e $y_i \sim$ Poisson(λ_b). Os dados realmente observados são apenas as contagens $\mathbf{y} = (y_1, \ldots, y_{222})$. As variáveis não-observadas em $\mathbf{z} = (z_1, \ldots, z_{222})$ são chamadas de variáveis latentes ou ocultas (hidden, latent) O vetor de parâmetros é $\theta = (\lambda_a, \lambda_b, \alpha)$.

23.2.1 A verossimilhança completa

O vetor (y_i, z_i) é composto por duas v.a.'s discretas com distribuição conjunta dada por

$$\begin{aligned}
\mathbb{P}(y_i = y, z_i = 0) &= \mathbb{P}(y_i = y | z_i = 0)\mathbb{P}(z_i = 0) = \frac{\lambda_a^y}{y!} e^{-\lambda_a} \cdot (1-\alpha) \\
\mathbb{P}(y_i = y, z_i = 1) &= \mathbb{P}(y_i = y | z_i = 1)\mathbb{P}(z_i = 1) = \frac{\lambda_b^y}{y!} e^{-\lambda_b} \cdot \alpha
\end{aligned}$$

para $y \in \mathbb{N}$. Podemos escrever estas duas expressões com uma única linha. Para $z = 0$ ou $z = 1$, temos

$$\mathbb{P}(y_i = y, z_i = z) = \left[\frac{\lambda_a^y e^{-\lambda_a}}{y!}(1-\alpha)\right]^{1-z} \left[\frac{\lambda_b^y e^{-\lambda_b}}{y!}\alpha\right]^z \tag{23.2}$$

Estamos supondo que as contagens de diferentes mulheres são v.a.'s independentes. Então, a verossimilhança de $\theta = (\lambda_a, \lambda_b, \alpha)$ baseada nos dados completos é

$$L^c(\theta | \mathbf{y}, \mathbf{z}) = \prod_{i=1}^{222} \left(\frac{\lambda_a^{y_i} e^{-\lambda_a}}{y_i!}(1-\alpha)\right)^{1-z_i} \left(\frac{\lambda_b^{y_i} e^{-\lambda_b}}{y_i!}\alpha\right)^{z_i}$$

Tomando logaritmo, temos a log-verossimilhança completa

$$\ell^c(\theta|\mathbf{y},\mathbf{z}) = \sum_{i=1}^{222}(1-z_i)\log\left(\frac{\lambda_a^{y_i}e^{-\lambda_a}}{y_i!}(1-\alpha)\right) + z_i\log\left(\frac{\lambda_b^{y_i}e^{-\lambda_b}}{y_i!}\alpha\right)$$

O MLE de $\theta = (\lambda_a, \lambda_b, \alpha)$ no caso da informação completa $(\mathbf{y},\mathbf{z})$ estar disponível é muito simples (fica como exercício):

$$\hat{\alpha} = \frac{1}{222}\sum_{i=1}^{222} z_i = \text{proporção das pacientes com } z_i = 1 \tag{23.3}$$

$$\hat{\lambda}_a = \frac{\sum_{i=1}^{222} y_i(1-z_i)}{\sum_{i=1}^{222}(1-z_i)} = \text{média das pacientes com } z_i = 0 \tag{23.4}$$

$$\hat{\lambda}_b = \frac{\sum_{i=1}^{222} y_i z_i}{\sum_{i=1}^{222} z_i} = \text{média das pacientes com } z_i = 1 \tag{23.5}$$

$$\tag{23.6}$$

Se pelo menos tivéssemos o vetor completo $(\mathbf{y},\mathbf{z})$..... Mas o que temos é apenas o vetor $\mathbf{y}$ das contagens. Precisamos da versossmilhança de $\alpha, \lambda_a, \lambda_b$ usando *apenas* $\mathbf{y}$.

23.2.2 A verossimilhança incompleta

Vamos obter a verossimilhança marginal dos dados observados $\mathbf{y}$. Como as mulheres são independentes, basta encontrar a distribuição da contagem (y_i) do i-ésimo bloco.

$$\begin{aligned}\mathbb{P}(Y_i = y) &= \mathbb{P}(Y_i = y, Z_i = 0) + \mathbb{P}(Y_i = y, Z_i = 1) \\ &= \alpha\frac{\lambda_a^y e^{-\lambda_a}}{y!} + (1-\alpha)\frac{\lambda_b^y e^{-\lambda_b}}{y!}\end{aligned}$$

Com isto, obtemos a verossimilhança baseada apenas nos dados realmente observados

$$L(\theta|\mathbf{y}) = \prod_{i=1}^{222}\mathbb{P}(Y_i = y_i) = \prod_{i=1}^{222}\left(\frac{\alpha\lambda_a^{y_i}e^{-\lambda_a}}{y_i!} + \frac{(1-\alpha)\lambda_b^{y_i}e^{-\lambda_b}}{y_i!}\right)$$

Esta função já não é tão simples de ser maximizada (na verdade, neste toy example, ela é muito simples). O algoritmo EM vem em nosso socorro, especialmente em problemas mais complicados.

23.3 O algoritmo EM

A primeira coisa a se fazer é obter a distribuição $\mathbb{P}(\mathbf{z}|\mathbf{y},\theta)$ dos dados faltantes $\mathbf{Z}$ condicionados nos valores $\mathbf{y}$ observados. Temos

$$\mathbb{P}(\mathbf{z}|\mathbf{y},\theta) = \prod_{i=1}^{222}\mathbb{P}(z_i|y_i,\theta) = \prod_{i=1}^{222}\frac{\mathbb{P}(y_i,z_i|\theta)}{\mathbb{P}(y_i|\theta)}$$

E agora, o principal truque do algoritmo EM: como não sabemos quem é $\mathbf{z}$, vamos deixá-lo aleatório e tomar o seu valor esperado. Este é o passo 1 do algoritmo EM no problema de mistura: obter a log-verossimilhança ℓ^c baseada nos dados completos mas deixando os dados faltantes como variáveis aleatórias e, a seguir, calcular o seu valor *esperado* $\mathbb{E}(\ell^c)$.

23.3.1 A distribuição de $\mathbf{Z}|\mathbf{Y}=\mathbf{y}$

Mais precisamente, calculamos a log-verossimilhança $\ell^c = \log L^c$ de θ baseada nos dados completos:

$$\begin{aligned}
\ell^c(\theta|\mathbf{y},\mathbf{z}) &= \log L^c(\theta|\mathbf{y},\mathbf{z}) \\
&= \log\left[\prod_{i=1}^{222}\left(\frac{\lambda_a^{y_i}e^{-\lambda_a}}{y_i!}(1-\alpha)\right)^{1-z_i}\left(\frac{\lambda_b^{y_i}e^{-\lambda_b}}{y_i!}\alpha\right)^{z_i}\right] \\
&= \sum_{i=1}^{222}\left[(1-z_i)\log\left(\frac{\lambda_a^{y_i}e^{-\lambda_a}}{y_i!}(1-\alpha)\right)+z_i\log\left(\frac{\lambda_b^{y_i}e^{-\lambda_b}}{y_i!}\alpha\right)\right]
\end{aligned}$$

A seguir, substituímos os valores desconhecidos z_i pelas variáveis aleatórias Z_i fazendo com que $l^c(\theta|\mathbf{y},\mathbf{Z})$ seja a variável aleatória:

$$\ell^c(\theta|\mathbf{y},\mathbf{Z}) = \sum_{i=1}^{222}\left[(1-Z_i)\log\left(\frac{\lambda_a^{y_i}e^{-\lambda_a}}{y_i!}(1-\alpha)\right)+Z_i\log\left(\frac{\lambda_b^{y_i}e^{-\lambda_b}}{y_i!}\alpha\right)\right]$$

Precisamos agora calcular o valor esperado de $\ell^c(\theta|\mathbf{y},\mathbf{Z})$. Observe que, em $\ell^c(\theta|\mathbf{y},\mathbf{Z})$, estamos deixando $\mathbf{y}$ fixado em seus valores observados na amostra. A *única* coisa aleatória em $\ell^c(\theta|\mathbf{y},\mathbf{Z})$ é o vetor $\mathbf{Z}$. Então, ao calcular a esperança de $\ell^c(\theta|\mathbf{y},\mathbf{Z})$ precisamos lembrar que calculamos uma esperança condicionada a $\mathbf{Y}=\mathbf{y}$. Assim,

$$\mathbb{E}[\ell^c(\theta|\mathbf{y},\mathbf{Z})|\mathbf{Y}=\mathbf{y}] = \sum_{i=1}^{222}\left[\log\left(\frac{\lambda_a^{y_i}e^{-\lambda_a}}{y_i!}(1-\alpha)\right)\mathbb{E}(1-Z_i|\mathbf{Y}=\mathbf{y})+\log\left(\frac{\lambda_b^{y_i}e^{-\lambda_b}}{y_i!}\alpha\right)\mathbb{E}(Z_i|\mathbf{Y}=\mathbf{y})\right] \quad (23.7)$$

Existe outra sutileza no nosso caminho. O valor de θ na verossimilhança $\ell^c(\theta|\mathbf{y},\mathbf{Z})$ é um valor θ arbitrário pertencente ao espaço paramétrico Θ_0. Queremos maximizar ℓ^c com respeito a este parâmetro. Mas, ao calcular $\mathbb{E}(\mathbf{Z}_i|\mathbf{Y}=\mathbf{y})$ em (23.7), precisamos usar *algum valor* para o parâmetro θ. Por isto, vamos usar um valor inicial $\theta^{(0)}$ para este parâmetro no cálculo desta esperança. Para deixar tudo bastante explícito, vamos usar uma notação um pouco mais carregada reescrevendo (23.7) como:

$$\mathbb{E}\left[\ell^c(\theta|\mathbf{y},\mathbf{Z})|\mathbf{Y}=\mathbf{y},\theta^{(0)}\right] = \sum_{i=1}^{222}\mathbb{E}\left(1-Z_i|\mathbf{Y}=\mathbf{y},\theta^{(0)}\right)\log\left(\frac{\lambda_a^{y_i}e^{-\lambda_a}}{y_i!}(1-\alpha)\right)+\mathbb{E}\left(Z_i|\mathbf{Y}=\mathbf{y},\theta^{(0)}\right)\log\left(\frac{\lambda_b^{y_i}e^{-\lambda_b}}{y_i!}\alpha\right)$$

Assim, queremos calcular

$$\mathbb{E}\left[\ell^c(\theta|\mathbf{y},\mathbf{Z})|\mathbf{Y}=\mathbf{y},\theta^{(0)}\right] = \sum_{i=1}^{180}\mathbb{E}\left(1-Z_i|\mathbf{Y}=\mathbf{y},\theta^{(0)}\right)\log\left(\frac{\lambda_a^{y_i}e^{-\lambda_a}}{y_i!}(1-\alpha)\right)+\mathbb{E}\left(Z_i|\mathbf{Y}=\mathbf{y},\theta^{(0)}\right)\log\left(\frac{\lambda_b^{y_i}e^{-\lambda_b}}{y_i!}\alpha\right)$$

O vetor $\mathbf{Y}$ está fixado no seu valor observado $\mathbf{y}$. A esperança de Z_i usa um *valor inicial e fixo* $\theta^{(0)}$ para o parâmetro desconhecido. θ é o valor genérico do parâmetro. $\mathbf{Z}$ é o vetor aleatório que torna a função $l^c(\theta|\mathbf{y},\mathbf{Z})$ uma variável aleatória. Vamos denotar $\theta=(\lambda_a,\lambda_b,\alpha)$ e $\theta^{(0)}=(\lambda_a^{(0)},\lambda_b^{(0)},\alpha^{(0)})$

23.3.2 Escolhendo $\theta^{(0)}$

O valor inicial $\theta^{(0)}=(\lambda_a^{(0)},\lambda_b^{(0)},\alpha^{(0)})$ pode ser obtido fazendo uma inspeção grosseira dos dados. Por exemplo, considerando o gráfico de barras para as 222 contagens na Figura 23.2, podemos chutar $\theta^{(0)}=(\lambda_a^{(0)},\lambda_b^{(0)},\alpha^{(0)})=(7,17,0.10)$.

23.3.3 $\mathbb{E}(Z_i|\mathbf{Y}=\mathbf{y},\theta^{(0)})$

Para calcular $\mathbb{E}(Z_i|\mathbf{Y}=\mathbf{y},\theta^{(0)})$ lembramos que Z_i depende apenas de Y_i e que Z_i é uma variável aleatória binária. Portanto,

$$\begin{aligned}
\mathbb{E}(Z_i|\mathbf{Y}=\mathbf{y},\theta^{(0)}) &= \mathbb{P}(Z_i=1|Y_i=y_i,\theta^{(0)}) \\
&= \frac{\mathbb{P}(Z_i=1,Y_i=y_i|\theta^{(0)})}{\mathbb{P}(Z_i=1,Y_i=y_i|\theta^{(0)})+\mathbb{P}(Z_i=0,Y_i=y_i|\theta^{(0)})} \\
&= \frac{\left(\frac{\lambda_b^{(0)y_i}}{y_i!}e^{-\lambda_b^{(0)}}\right)\cdot\alpha_0}{\frac{\lambda_b^{(0)y_i}}{y_i!}e^{-\lambda_b^{(0)}}\cdot\alpha_0+\frac{\lambda_a^{(0)y_i}}{y_i!}e^{-\lambda_a^{(0)}}\cdot(1-\alpha_0)} \\
&= \frac{\lambda_b^{(0)y_i}e^{-\lambda_b^{(0)}}\alpha_0}{\lambda_b^{(0)y_i}e^{-\lambda_b^{(0)}}\alpha_0+\lambda_a^{(0)y_i}e^{-\lambda_a^{(0)}}(1-\alpha_0)}
\end{aligned}$$

onde $\theta^{(0)}=(\lambda_a^{(0)},\lambda_b^{(0)},\alpha_0)$.

Por exemplo, considerando $\theta^{(0)}=(\lambda_a^{(0)},\lambda_b^{(0)},\alpha^{(0)})=(7,17,0.10)$, temos

$$\begin{aligned}
\mathbb{E}(Z_i|\mathbf{Y}=\mathbf{y},\theta^{(0)}) &= \mathbb{P}(Z_i=1|Y_i=y_i,\theta^{(0)}) \\
&= \frac{\lambda_b^{(0)y_i}e^{-\lambda_b^{(0)}}\alpha_0}{\lambda_b^{(0)y_i}e^{-\lambda_b^{(0)}}\alpha_0+\lambda_a^{(0)y_i}e^{-\lambda_a^{(0)}}(1-\alpha_0)} \\
&= \frac{17^{y_i}e^{-17}*0.10}{17^{y_i}e^{-17}*0.10+7^{y_i}e^{-7}*0.90}
\end{aligned}$$

Note que o parâmetro desconhecido θ não está presente nesta expressão. Ele foi substituído por um valor inicial fixo $\theta^{(0)}=(7,17,0.10)$ um tanto arbitrário. O valor $\mathbb{E}(Z_i|\mathbf{Y}=\mathbf{y},\theta^{(0)})$ é computável, é simplesmente um número real.

Tendo o valor de $\mathbb{E}(Z_i|\mathbf{Y}=\mathbf{y},\theta^{(0)})$ podemos então calcular $\mathbb{E}\left[l^c(\theta|\mathbf{y},\mathbf{Z})|\mathbf{Y}=\mathbf{y},\theta^{(0)}\right]$. Este último valor pode ser calculado em pontos arbitrários θ se $\theta^{(0)}$ é fixado. Vamos definir:

$$Q(\theta|\theta^{(0)},\mathbf{y})=\mathbb{E}\left[l^c(\theta|\mathbf{y},\mathbf{Z})|\mathbf{Y}=\mathbf{y},\theta^{(0)}\right] \tag{23.8}$$

Esta expressão é crucial no algoritmo EM. Lembre-se: $\theta^{(0)}$ é um chute inicial e fixo para o parâmetro, θ é um valor arbitrário para o parâmetro e os Z's são os rótulos dos grupos das observações. $\theta^{(0)}$ será atualizado ao longo das iterações, como explicaremos em breve.

Os dados $\mathbf{y}$ estarão fixos ao longo das iterações do algoritmo EM. É importante perceber que $Q(\theta|\theta^{(0)},\mathbf{y})$ é função de dois valores para o parâmetro, um valor genérico e arbitrário, não-especificado, θ e o valor inicial especificado e fixo $\theta^{(0)}$.

23.3.4 Os passos do algoritmo EM

O primeiro passo é chamado E-step: trata-se de obter a esperança da log-verossimilhança, a expressão $Q(\theta|\theta^{(0)},\mathbf{y})$ onde $\theta^{(0)}$ é um valor inicial usado para calcular $E(Z|Y=y)$ e θ é um valor arbitrário θ. O segundo passo do algoritmo EM é chamado M-step. Lembrando que $\theta^{(0)}$ é um valor inicial fixado pelo usuário, no passo M encontramos o valor de θ que maximiza $Q(\theta|\theta^{(0)},\mathbf{y})$.

Isto é, encontramos o valor θ_1 do argumento θ que maximiza $Q(\theta|\theta^{(0)},\mathbf{y})$ mantendo $\theta^{(0)}$ fixo num valor cnhecido:

$$\theta^{(1)} = \arg_{\theta\in\Theta}\max Q(\theta|\theta^{(0)},\mathbf{y})$$

No caso de mistura de Poissons, esta maximização é muito simples. Para simplificar, vamos escrever $\mathbb{E}(Z_i|\mathbf{Y}=\mathbf{y},\theta^{(0)})$ como $\mathbb{E}(Z_i)$. Então

$$\hat{\alpha}^{(1)} = \frac{1}{222}\sum_{i=1}^{222}\mathbb{E}(Z_i) \tag{23.9}$$

$$\hat{\lambda}_a^{(1)} = \frac{\sum_{i=1}^{222} y_i(1-\mathbb{E}(Z_i))}{\sum_{i=1}^{222}(1-\mathbb{E}(Z_i))} \tag{23.10}$$

$$\hat{\lambda}_b^{(1)} = \frac{\sum_{i=1}^{222} y_i\mathbb{E}(Z_i)}{\sum_{i=1}^{222}\mathbb{E}(Z_i)} \tag{23.11}$$

Esta expressões são quase idênticas ao MLE no caso de dados completos. Compare as expressões acima com aquelas de (23.3)-(23.5). Veja que os z_i, conhecidos no caso completo, são substituídos pelo valor corrente de sua estimativa, $\mathbb{E}(Z_i)$, no algoritmo EM. O valor de $\mathbb{E}(Z_i)$ é atualizado tão logo as novas estimativas $\theta^{(1)}$ para o parâmetro são obtidas via (23.3)-(23.5).

23.3.5 Resumo do algoritmo EM

Começamos com um valor de $\theta^{(0)}$ inicial para o parâmetro θ. Calculamos $Q(\theta|\theta^{(0)},\mathbf{y})$ como uma função de θ (com $\theta^{(0)}$ fixo). A seguir, maximizamos $Q(\theta|\theta^{(0)},\mathbf{y})$ com respeito a θ obtendo $\theta^{(1)}$. O processo é iterado:

- calculamos $Q(\theta|\theta^{(j)},\mathbf{y})$ (passo E)
- A seguir, maximizamos em θ para obter $\theta^{(j+1)}$ (passo M)

Este processo iterativo converge para o EMV de θ. A principal desvantagem do algoritmo EM é que esta convergência pode ser lenta. O que muda de problema para problema é a expressão de $Q(\theta|\theta^{(j)},\mathbf{y})$.

Uma grande vantagem adicional do algoritmo EM é que terminamos também com uma estimativa de $\mathbb{E}(Z_i) = \mathbb{P}(Z_i = k)$, a probabilidade de cada observação pertencer ao grupo k.

23.4 Exemplos de uso do algoritmo EM

Terminar EM para o caso Poisson no R Caso normal multivariado: ver wikipedia.

23.5 Convergência d algoritmo EM

Mas por quê o algoritmo EM funciona? Existe uma prova de que o EM converge para um máximo local (ou global) da log-verossimilhança, como veremos neta seção.

23.5.1 Definições preliminares

Definition 23.5.1 — Função convexa. Uma função $g(x)$ é uma *função convexa* se a curva está sempre abaixo da secante. Outra definição equivalente: é convexa se a reta tangente em cada ponto está abaixo da curva. Ou então se a derivada $g'(x)$ é crescente. Ou ainda se a derivada segunda $g''(x)$ é positiva (ou melhor, não-negativa).

O exemplo clássico de função convexa é a parábola $g(x) = x^2$. Cheque cada uma das definições acima para verificar que esta função,de fato, é convexa. Outros exemplos clássico é $g(x) = e^x$.

Para quê tantas caracterizações de função convexa? A razão é que, ao generalizar a definição para funções de várias variáveis, algumas das caracterizações podem ser verificadas mais facilmente.

Definition 23.5.2 — Função convexa multivariada. Uma função $g(\mathbf{x}) : A \subset \mathbb{R}^n \to \mathbb{R}$ é convexa se a matriz de derivadas parciais de segunda ordem $D^2 g$ é uma matriz definida positiva. Isto é, se $\mathbf{x}^t\ D^2 g\ \mathbf{x} > 0$ para todo ponto $\mathbf{x}$.

■ **Example 23.1 — Parabolóide é convexo.** Generalizando o caso da parábola, a função $g(\mathbf{x}) = \beta_1 x_1^2 + \beta_2 x_2^2 + \ldots + \beta_n x_n^2$ com $\beta_j > 0$ é convexa pois

$$D^2 g = \begin{bmatrix} \partial^2 g/\partial x_1^2 & \partial^2 g/\partial x_1 x_2 & \ldots & \partial^2 g/\partial x_1 x_n \\ \partial^2 g/\partial x_2 x_1 & \partial^2 g/\partial x_2^2 & \ldots & \partial^2 g/\partial x_2 x_n \\ \vdots & \vdots & \ldots & \vdots \\ \partial^2 g/\partial x_n x_1 & \partial^2 g/\partial x_n x_2 & \ldots & \partial^2 g/\partial x_n^2 \end{bmatrix} = 2 \begin{bmatrix} \beta_1 & 0 & \ldots & 0 \\ 0 & \beta_2^2 & \ldots & 0 \\ \vdots & \vdots & \ldots & \vdots \\ 0 & 0 & \ldots & \beta_n \end{bmatrix}$$

Agora, como $D^2 g$ é uma matriz diagonal é fácil verificar que, para todo ponto $\mathbf{x} \in \mathbb{R}^n$ e diferente de zero, temos

$$\mathbf{x}^t\ D^2 g\ \mathbf{x} = \sum_i \beta_i x_i^2 .$$

Se $\mathbf{x} \neq \mathbf{0}$ então pelo menos uma de suas coordenadas deve ser estritamente maior que zero e portanto, como $x_i^2 \geq 0$ e $\beta_i > 0$, temos $\mathbf{x}^t\ D^2 g\ \mathbf{x} > 0$. ■

■ **Example 23.2 — Exponencial multivariada é função convexa.** Generalizando o caso da exponencial, a função $g(\mathbf{x}) = \exp(\beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n) = \exp(\mathbf{x}'\beta)$ com $\beta_j > 0$ é convexa pois

$$\nabla g = g(\mathbf{x}) \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}$$

e portanto

$$D^2 g = g(\mathbf{x}) \begin{bmatrix} \beta_1^2 & \beta_1\beta_2 & \ldots & \beta_1\beta_n \\ \beta_2\beta_1 & \beta_2^2 & \ldots & \beta_2\beta_n \\ \vdots & \vdots & \ldots & \vdots \\ \beta_n\beta_1 & \beta_n\beta_2 & \ldots & \beta_n^2 \end{bmatrix}$$

■

Função g é côncava se $-g$ é convexa. No caso côncavo, desigualdade é invertida.

23.5.2 Desigualdade de Jensen

A desigualdade de Jensen é uma desigualdade fundamental em probabilidade. Usualmente, ela aparece nos livros de probabilidade depois da definição de esperança de uma v.a. Entretanto, como seu principal uso neste livro só aparece agora, deixamos para apresentá-la mais tarde, junto com seu primeiro uso no texto.

Seja X uma v.a. qualquer com $E(X) = \mu$. Seja $g(x)$ uma função convexa. Crie uma nova v.a. $Y = g(X)$. Então a esperança dessa nova v.a., $\mathbb{E}(Y) = \mathbb{E}(g(X)) \geq g(\mu) = g(\mathbb{E}(X))$ Exemplo: $\mathbb{E}(g(X)) = \mathbb{E}(X^2) \geq g(\mathbb{E}(X)) = [\mathbb{E}(X)]^2 = \mu^2$

Função LOG é côncava: $E(\log(X)) \leq \log[E(X)]$

Notação

Seja $(\mathbf{y}, \mathbf{z})$ o vetor de dados completos com densidade $f(\mathbf{y}, \mathbf{z}|\theta)$. Vamos também denotar $\ell^c(\theta|\mathbf{y}, \mathbf{z}) = \log f(\mathbf{y}, \mathbf{z}|\theta)$. Seja $f(\mathbf{y}|\theta) = \int_{\mathscr{Z}} f(\mathbf{y}, \mathbf{z}|\theta) d\mathbf{z}$ a densidade marginal de $\mathbf{Y}$. Esta é também a log-verossimilhança de θ baseada apenas nos dados observados $\mathbf{y}$. Isto é, $\ell(\theta|\mathbf{y}) = \log f(\mathbf{y}|\theta)$. Seja

$$k(\mathbf{z}|\mathbf{y}, \theta) = \frac{f(\mathbf{y}, \mathbf{z}|\theta)}{f(\mathbf{y}|\theta)}$$

a densidade condicional de $\mathbf{Z}$ dados as observações $\mathbf{y}$. Vamos usar a letra k para denotar esta densidade condicional e assim evitar mais usos da letra f para densidades.

Escolha d número de componentes.

Como uma rotina numérica para otimização, o algoritmo EM é geralmente estável e confiável. Uma propriedade atraente é que a verossimilhança aumenta a cada iteração sucessiva do algoritmo, resultado do Teorema ??. Assim, a menos que as iterações sucessivas empurrem $\theta^{(n)}$ para infinito, o algoritmo EM vai convergir para algum valor. Mas a convergência para o valor desejado, o MLE, não é garantida: se a função de verossimilhança tem vários modas (ou pontos de máximo local), o algoritmo pode convergir para um desses máximos locais, e não para o máximo global. As condições suficientes para a convergência para o máximo global são dadas por Wu (1983). Apesar do algoritmo EM convergir em geral, a convergência pode ser lenta. Esta é um dos principais pontos fracos desse notável algoritmo.

24. Sufficiency and Exponential Family

24.1 Introduction

We have seen that the maximum likelihood estimator (MLE) possesses desirable properties that make it a strong candidate for solving inference problems. In this chapter, we explore yet another important property of the MLE, one that is based on the concept of *sufficiency*. This concept was also formalized by Fisher in his seminal 1922 paper [3].

The goal of statistical inference is to extract from the data all available information about the underlying probability distribution that generated it. In the parametric setting, this means learning as much as possible about the unknown parameter θ. But how can we formalize this notion of "extracting all the information," and how can we apply it in practice?

Fisher proposed a criterion for determining whether all the information about a parameter θ controlling the $\mathbf{Y}$ data probability distribution is captured by a summary statistic $T(\mathbf{Y})$. This leads to the concept of *sufficiency*, a subtle but powerful idea. To build intuition, we begin with a simple example.

Suppose $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)$ is a vector of independent and identically distributed (i.i.d.) random variables with distribution $N(\theta, 1)$, where $n > 2$. Consider the estimator $\hat{\theta} = (Y_1 + Y_2)/2$, which uses only two observations in the sample. This estimator is clearly suboptimal, as it ignores the remaining $n-2$ observations and thereby fails to use all the information available in the sample.

Now consider a second estimator. Let $Y_{(1)} \leq Y_{(2)} \leq \ldots \leq Y_{(n)}$ denote the order statistics of the sample $\mathbf{Y}$. That is, $Y_{(1)}$ is the smallest observation, $Y_{(n)}$ the largest, and the others are in between. Define the estimator

$$\hat{\theta}^* = \frac{Y_{(1)} + Y_{(n)}}{2},$$

the average of the minimum and maximum values in the sample. Does this estimator use all the information available about θ? What about the observations $Y_{(2)}, \ldots, Y_{(n-1)}$? Do they carry any additional information about θ?

The answer is yes. These middle order statistics contain useful information about θ, and this becomes evident when we compare $\hat{\theta}^*$ with the sample mean $\overline{Y}$, which is the efficient estimator in

this setting. The sample mean achieves the Cramér–Rao lower bound and uses all the data:

$$\begin{aligned}\overline{Y} &= \frac{Y_1 + Y_2 + \ldots + Y_n}{n} = \frac{Y_{(1)} + Y_{(2)} + \ldots + Y_{(n)}}{n} \\ &= \frac{Y_{(1)} + Y_{(n)}}{n} + \frac{Y_{(2)} + \ldots + Y_{(n-1)}}{n} \\ &= \frac{2}{n} \cdot \frac{Y_{(1)} + Y_{(n)}}{2} + \frac{Y_{(2)} + \ldots + Y_{(n-1)}}{n} \\ &= \frac{2\hat{\theta}^*}{n} + \frac{Y_{(2)} + Y_{(3)} + \ldots + Y_{(n-1)}}{n}.\end{aligned}$$

This decomposition makes it clear that the middle order statistics $Y_{(2)}, \ldots, Y_{(n-1)}$ do in fact carry additional information about θ, and that $\hat{\theta}^*$ fails to exploit it.

While this toy example illustrates the idea of not using all available information, formalizing this notion is far more challenging in complex models, such as logistic regression. Suppose we fit such a model and obtain the MLE $\hat{\beta}$ for the coefficients β. Can we be confident that $\hat{\beta}$ has extracted all the information about β from the data? Or could further information still be mined from the data beyond the MLE?

By the end of this chapter, we will see that the answer is no: once the MLE is computed, there is no additional useful information in the data for estimating β in the logistic regression model and in many other models.

24.2 Definition of sufficient statistics

Our goal is to establish a general criterion for determining whether all the information about θ has been captured by a statistic $T(\mathbf{Y})$. Fisher addressed this problem by introducing the concept of a **sufficient statistic**: a function of the data that retains all information relevant to estimating θ. Once we have this summary, the original data can be discarded, as they contain nothing further that aids in inference.

Suppose we have $n = 5$ independent and identically distributed (i.i.d.) random variables $Y_1, Y_2, \ldots, Y_5$, each following a binary Bernoulli distribution with parameter θ. There are 2^5 possible binary sequences, and for each of them, the probability depends on the success probability θ and the number s of successes in the sequence:

$$\mathbb{P}(\mathbf{Y} = \mathbf{Y}) = \theta^s (1-\theta)^{n-s}.$$

For example, by independence,

$$\mathbb{P}(\mathbf{Y} = (1,1,0,1,0);\ \theta) = \theta \cdot \theta \cdot (1-\theta) \cdot \theta \cdot (1-\theta) = \theta^3 (1-\theta)^2.$$

Consider the statistic $T(\mathbf{Y}) = \sum_{i=1}^5 Y_i$, which represents the total number of successes among the five trials. Suppose we are told that the total number of successes is 4. That is, the event $T(\mathbf{Y}) = 4$ has occurred. Given this information, the only uncertainty that remains in the data vector $\mathbf{Y}$ is the specific positions of the four successes and the single failure.

If the statistic $T(\mathbf{Y})$ is truly sufficient for estimating the parameter θ, then the ordering of the successes and failure should be *irrelevant* to the estimation of θ. Indeed, the natural estimator of θ in this setting is $\hat{\theta} = T(\mathbf{Y})/5$, which depends solely on the statistic $T(\mathbf{Y})$.

What is the distribution of the data vector $\mathbf{Y}$ given that $T(\mathbf{Y}) = 4$? Among all 2^5 possible binary sequences of length 5, most now have zero probability of occurring. Specifically, if $\mathbf{y} = (y_1, \ldots, y_5)$ is a binary sequence that does not contain exactly one failure, then

$$\mathbb{P}(\mathbf{Y} = \mathbf{y} \mid T(\mathbf{Y}) = 4;\ \theta) = 0.$$

For instance,

$$\mathbb{P}(\mathbf{Y} = (1,1,0,1,0) \mid T(\mathbf{Y}) = 4;\ \theta) = 0\,.$$

The only sequences with non-zero probability under the condition $T(\mathbf{Y}) = 4$ are the five binary sequences containing exactly one zero:

$$[T(\mathbf{Y}) = 4] = \{(0,1,1,1,1),(1,0,1,1,1),(1,1,0,1,1),(1,1,1,0,1),(1,1,1,1,0)\}\,.$$

Let us compute the conditional probability of one such sequence. For example,

$$\mathbb{P}(\mathbf{Y} = (1,1,0,1,1) \mid T(\mathbf{Y}) = 4;\ \theta) = \frac{\mathbb{P}(\mathbf{Y} = (1,1,0,1,1) \cap T(\mathbf{Y}) = 4;\ \theta)}{\mathbb{P}(T(\mathbf{Y}) = 4;\ \theta)}\,.$$

Since $(1,1,0,1,1)$ belongs to the event $T(\mathbf{Y}) = 4$, the intersection is simply:

$$[\mathbf{Y} = (1,1,0,1,1)] \cap [T(\mathbf{Y}) = 4] = \{(1,1,0,1,1)\}\,,$$

and thus,

$$\mathbb{P}(\mathbf{Y} = (1,1,0,1,1) \cap T(\mathbf{Y}) = 4;\ \theta) = \mathbb{P}(\mathbf{Y} = (1,1,0,1,1);\ \theta) = \theta^4(1-\theta)\,.$$

The denominator is the probability that $T(\mathbf{Y}) = 4$, which follows a Binomial distribution:

$$\mathbb{P}(T(\mathbf{Y}) = 4;\ \theta) = \binom{5}{4}\theta^4(1-\theta)^1 = 5\theta^4(1-\theta)\,.$$

Therefore, the desired conditional probability is

$$\mathbb{P}(\mathbf{Y} = (1,1,0,1,1) \mid T(\mathbf{Y}) = 4;\ \theta) = \frac{\theta^4(1-\theta)}{5\theta^4(1-\theta)} = \frac{1}{5}\,.$$

It is easy to see that this probability is the same for each of the five sequences in the event $[T(\mathbf{Y}) = 4]$. Hence, we conclude that:

$$\mathbb{P}(\mathbf{Y} = \mathbf{y} \mid T(\mathbf{Y}) = 4;\ \theta) = \begin{cases} 1/5, & \text{if } T(\mathbf{Y}) = 4 \\ 0, & \text{otherwise} \end{cases}$$

The parameter θ has vanished from this conditional distribution. This means that although the full distribution of $\mathbf{Y}$ depends on θ, once we know the number of successes, the remaining variation in $\mathbf{Y}$ no longer depends on θ.

For example, consider two different values $\theta_1 = 0.8$ and $\theta_2 = 0.2$. The full probability of the vector $\mathbf{Y}$ varies substantially between these two cases:

$$\mathbb{P}(\mathbf{Y} = (1,1,0,1,1);\ \theta = 0.8) = 0.8^4 \times 0.2 \approx 0.082\,,$$

while

$$\mathbb{P}(\mathbf{Y} = (1,1,0,1,1);\ \theta = 0.2) = 0.2^4 \times 0.8 \approx 0.001\,,$$

with the former being approximately 82 times larger. However, once we condition on $T(\mathbf{Y}) = 4$, the two conditional probabilities become identical:

$$\mathbb{P}(\mathbf{Y} = (1,1,0,1,1) \mid T(\mathbf{Y}) = 4;\ \theta = 0.8) = \frac{1}{5} = \mathbb{P}(\mathbf{Y} = (1,1,0,1,1) \mid T(\mathbf{Y}) = 4;\ \theta = 0.2)\,.$$

We can generalize this example by considering an arbitrary number n of Bernoulli trials and letting $T(\mathbf{Y}) = t$ be the number of successes, where $t \in \{0, 1, \ldots, n\}$. Consider a specific sequence $\mathbf{y}$ such that $T(\mathbf{y}) = t$. Similarly to the example we studied earlier, we obtain:

$$\begin{aligned}\mathbb{P}(\mathbf{Y} = \mathbf{y} \mid T(\mathbf{Y}) = t;\ \theta) &= \frac{\mathbb{P}(\mathbf{Y} = \mathbf{y} \cap T(\mathbf{Y}) = t;\ \theta)}{\mathbb{P}(T(\mathbf{Y}) = t;\ \theta)} \\ &= \frac{\mathbb{P}(\mathbf{Y} = \mathbf{y};\ \theta)\ (t \text{ successes in a determined order})}{\mathbb{P}(T(\mathbf{Y} = t);\ \theta)\ (t \text{ successes in any order})} \\ &= \frac{\prod_{i=1}^{n} \theta^{y_i}(1-\theta)^{1-y_i}}{\binom{n}{t}\theta^t(1-\theta)^{n-t}} = \frac{\theta^t(1-\theta)^{n-t}}{\binom{n}{t}\theta^t(1-\theta)^{n-t}} = \binom{n}{t}^{-1}\end{aligned}$$

Hence,

$$\mathbb{P}(\mathbf{Y} = \mathbf{Y} \mid T(\mathbf{Y}) = t;\ \theta) = \begin{cases} \binom{n}{t}^{-1}, & \text{if } T(\mathbf{y}) = t \\ 0, & \text{otherwise} \end{cases}$$

Think of inference about θ as a two-step process. In the first step, we are told the total number of successes $T(\mathbf{y}) = t$ in the sequence $\mathbf{y}$. This is highly informative about the unknown parameter θ: if $t \approx 0$, we infer that θ is small; if $t \approx n$, we infer that $\theta \approx 1$; and if $t \approx n/2$, then $\theta \approx 1/2$.

What else is in the data? The exact positions where the t successes occur among the n positions in the sequence. But this random allocation of successes (conditioned on their total number) cannot provide additional information about θ, because its distribution does not depend on θ. Any sequence $\mathbf{y}$ with exactly t successes has the same probability of occurring, $1/\binom{n}{t}$, regardless of the value of θ. Therefore, this additional random component carries no information about θ: it is unaffected by θ.

Fisher's brilliant conclusion was that, for estimating θ in this problem, it is **sufficient** to consider only the total number of successes $t(\mathbf{y})$, because the remaining randomness in the data (their order) has no relationship with θ.

Now that we have provided an intuitive idea in a specific example of what a sufficient statistic is, we will proceed to define it formally.

Definition 24.2.1 — Sufficient Statistic. Let $Y_1, Y_2, \ldots, Y_n$ be i.i.d. random variables with joint probability distribution $p(\mathbf{y}; \theta)$. A statistic $T(\mathbf{Y})$ is said to be **sufficient** for the parameter θ if the conditional distribution of $\mathbf{Y}$ given $T(\mathbf{Y})$ does not depend on θ.

■ **Example 24.1** Let $Y_1, Y_2, \ldots, Y_n$ be i.i.d. random variables following a Poisson distribution with parameter θ. The joint probability mass function of the sample is given by:

$$\mathbb{P}(\mathbf{Y} = \mathbf{y};\ \theta) = \prod_{i=1}^{n} \frac{\theta^{y_i} e^{-\theta}}{y_i!} = \frac{\theta^{\sum_i y_i} e^{-n\theta}}{\prod_i y_i!}$$

Consider the statistic $T(\mathbf{Y}) = \sum_{i=1}^{n} Y_i$. The sum of independent Poisson random variables is still a Poisson random variable and hence, $T(\mathbf{Y}) \sim \text{Poisson}(n\theta)$. Because of this, the conditional distribution of $\mathbf{Y}$ given $T(\mathbf{Y}) = t$, for $t = 0, 1, \ldots$, is easy. Assuming a configuration $\mathbf{y}$ for which $T(\mathbf{y}) = \sum_i y_i = t$, we have:

$$\begin{aligned}\mathbb{P}(\mathbf{Y} = \mathbf{y} \mid T(\mathbf{Y}) = t;\ \theta) &= \frac{\mathbb{P}(\mathbf{Y} = \mathbf{y} \cap T(\mathbf{Y}) = t;\ \theta)}{\mathbb{P}(T(\mathbf{Y}) = t;\ \theta)} \\ &= \frac{\mathbb{P}(\mathbf{Y} = \mathbf{y};\ \theta)}{\mathbb{P}(T(\mathbf{Y} = t);\ \theta)} \\ &= \frac{e^{-n\theta}\theta^{\sum y_i}/\prod y_i!}{e^{-n\theta}(n\theta)^t/t!} = \frac{t!}{n^t \prod y_i!}\end{aligned}$$

Considering also the configurations $\mathbf{y}$ that are not compatible with $\sum_i y_i = t$, we can conclude that Hence,

$$\mathbb{P}(\mathbf{Y} = \mathbf{Y} \mid T(\mathbf{Y}) = t;\ \theta) = \begin{cases} \mathrm{t!/(n^t \prod y_i!)}, & \text{if } T(\mathbf{y}) = t \\ 0, & \text{otherwise} \end{cases}$$

This conditional distribution does *not* depend on θ, which confirms that $T(\mathbf{Y}) = \sum Y_i$ is a sufficient statistic for θ. ∎

Unlike the binomial example, the non-zero events in the Poisso example are *not* equally likely. Although the conditional distribution does not depend on θ, the sequences $\mathbf{x}$ that are compatible with $T(\mathbf{X}) = t$ do not have equal probabilities.

For example, with $n = 3$ and $T(\mathbf{X}) = 2$, the possible sequences for the random vector $\mathbf{Y} = (Y_1, Y_2, Y_3)$ include:

$$(2,0,0), \quad (0,2,0), \quad (0,0,2), \quad (1,1,0), \quad (1,0,1), \ldots$$

These sequences have different probabilities:

$$\mathbb{P}(\mathbf{X} = (2,0,0) \mid T(\mathbf{X}) = 2;\ \theta) = \frac{2!}{3^2 \cdot 2! \cdot 0! \cdot 0!} = \frac{1}{9}$$

while

$$\mathbb{P}(\mathbf{X} = (1,1,0) \mid T(\mathbf{X}) = 2;\ \theta) = \frac{2!}{3^2 \cdot 1! \cdot 1! \cdot 0!} = \frac{2}{9}$$

The crucial point in the definition of a sufficient statistic is that the conditional distribution of $\mathbf{Y}$ given that $T(\mathbf{Y}) = t$) **does not depend on** θ. It is not necessary that all compatible events (i.e., all sequences for which $T(\mathbf{Y}) = t$) be equally likely. What matters is that their distribution no longer involves the parameter θ. In other words, once we know the value of the sufficient statistic $T(\mathbf{Y})$, the distribution of the data becomes free of θ. At this point, we can discard the full data and retain only the summary $T(\mathbf{Y})$.

24.3 Neyman–Fisher Factorization Theorem

How can we find a sufficient statistic $T(\mathbf{Y})$? There are two main strategies. The first is to have a (possibly brilliant) insight that suggests a candidate statistic $T(\mathbf{Y})$. Then, we must verify whether the conditional distribution $(\mathbf{Y} \mid T(\mathbf{Y}) = t)$ indeed does not depend on θ. In very simple problems, such as the binomial and Poisson examples, this may be feasible. However, for more sophisticated models this becomes impossible. For instance, in a logistic regression model, what would a sufficient statistic be? If we manage to guess it, we still have to prove that the conditional distribution of the data given this statistic is independent of θ. That's not an easy task.

The second, more systematic option is to use the Neyman–Fisher factorization theorem. This theorem provides a direct, automatic, and often obvious method for identifying sufficient statistics, regardless of how complex the underlying probabilistic model may be.

Theorem 24.3.1 — Neyman–Fisher Factorization Theorem. Let $Y_1, Y_2, \ldots, Y_n$ be random variables with joint probability distribution $f(\mathbf{y}; \theta)$. A statistic $T(\mathbf{Y})$ is **sufficient** for θ if and only if the joint probability density can be factorized as

$$f(\mathbf{y}; \theta) = K_1(T(\mathbf{y}); \theta) \cdot K_2(y_1, y_2, \ldots, y_n),$$

where K_1 is a function of θ and the data *only through* the statistic $T(\mathbf{y})$, and K_2 does *not* depend on θ.

Before presenting the proof of the theorem, let us see how the factorization criterion is applied in practice.

■ **Example 24.2** Suppose once again we have a sample of n i.i.d. random variables following a Poisson distribution with parameter θ. The joint probability mass function of the sample is given by:

$$p(\mathbf{y};\theta) = \prod_{i=1}^{n} \frac{\theta^{y_i} e^{-\theta}}{y_i!} = \frac{e^{-n\theta}\theta^{\sum y_i}}{\prod y_i!} = \underbrace{e^{-n\theta}\theta^{T(\mathbf{y})}}_{K_1(T(\mathbf{y});\theta)} \cdot \underbrace{\frac{1}{\prod y_i!}}_{K_2(y_1,y_2,\ldots,y_n)}$$

where $T(\mathbf{y}) = \sum_{i=1}^{n} y_i$. Therefore, by simple inspection of the joitn distribution, the theorem allows us to see that $T(\mathbf{y}) = \sum_{i=1}^{n} y_i$ is a sufficient statistic to estimate θ. ■

■ **Example 24.3** Consider the case where the random variables follow a normal distribution with known mean equal to 3 and unknown variance σ^2. We are interested in estimating σ^2, and we will show that a sufficient statistic for this parameter is

$$T(\mathbf{y}) = \sum_{i=1}^{n} (y_i - 3)^2$$

The joint probability density function of the sample is:

$$f(\mathbf{y};\sigma^2) = \underbrace{\frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{\sum_{i=1}^{n}(y_i-3)^2}{2\sigma^2}\right)}_{K_1(T(\mathbf{y});\sigma^2)} \cdot \underbrace{1}_{K_2(y_1,y_2,\ldots,y_n)}$$

Therefore, $T(\mathbf{y}) = \sum_{i=1}^{n}(y_i-3)^2$ is a sufficient statistic for estimating σ^2. ■

Proof of the Neyman-Fisher factorization theorem. We will prove the Neyman–Fisher factorization theorem only for discrete random variables. First, we will show that if $T(\mathbf{Y})$ is a sufficient statistic, then the joint probability distribution can be written in the form:

$$\mathbb{P}(\mathbf{Y} = \mathbf{y};\theta) = K_1(T(\mathbf{y});\ \theta) \times K_2(\mathbf{y}).$$

Take an arbitrary configuration $\mathbf{y}$ and let $T(\mathbf{y}) = t$. Then,

$$\mathbb{P}(\mathbf{Y} = \mathbf{y};\theta) = \mathbb{P}([\mathbf{Y} = \mathbf{y}] \cap [T(\mathbf{Y}) = t];\theta) \tag{24.1}$$

because $[\mathbf{Y} = \mathbf{y}] \subseteq [T(\mathbf{Y}) = t]$.

Since $T(\mathbf{y})$ is sufficient, by definition, the conditional probability $\mathbb{P}(\mathbf{Y} = \mathbf{y} \mid T(\mathbf{Y}) = t;\theta)$ does not depend on θ. Hence, it is a function only of the data $\mathbf{y}$ and we name it as K_2:

$$K_2(\mathbf{y}) = \mathbb{P}(\mathbf{Y} = \mathbf{y} \mid T(\mathbf{Y}) = t) \tag{24.2}$$

For any realization $\mathbf{y}$ such that $T(\mathbf{y}) \neq t$, we have $\mathbb{P}(\mathbf{Y} = \mathbf{y} \mid T(\mathbf{Y}) = t) = 0$. For the realizations $\mathbf{y}$ such that $T(\mathbf{y}) = t$, we use (24.1) and (24.2) to write

$$K_2(\mathbf{y}) = \mathbb{P}(\mathbf{Y} = \mathbf{y} \mid T(\mathbf{Y}) = t) = \frac{\mathbb{P}(\mathbf{Y} = \mathbf{y},\ T(\mathbf{Y}) = t;\ \theta)}{\mathbb{P}(T(\mathbf{Y}) = t;\ \theta)} = \frac{\mathbb{P}(\mathbf{Y} = \mathbf{y};\ \theta)}{\mathbb{P}(T(\mathbf{Y}) = t;\ \theta)}$$

and therefore, isolating the numerator, we have

$$\mathbb{P}(\mathbf{Y}=\mathbf{y};\ \theta) = \underbrace{\mathbb{P}(T(\mathbf{Y})=t;\ \theta)}_{K_1(T(\mathbf{y});\ \theta)} K_2(\mathbf{y})$$

We now prove the converse direction. Suppose that the joint distribution factorizes as:

$$p(\mathbf{y};\theta) = K_1(T(\mathbf{y});\theta) \times K_2(\mathbf{y}).$$

For any realization $\mathbf{y}$ such that $T(\mathbf{y}) = t$, the conditional probability is:

$$\mathbb{P}_\theta(\mathbf{Y}=\mathbf{y} \mid T(\mathbf{Y})=t) = \frac{\mathbb{P}_\theta(\mathbf{Y}=\mathbf{y})}{\mathbb{P}_\theta(T(\mathbf{Y})=t)}. \tag{24.3}$$

Moreover, we can express the denominator as:

$$\mathbb{P}_\theta(T(\mathbf{Y})=t) = \sum_{\mathbf{y}\in A_t} p(\mathbf{y};\theta),$$

where A_t is the set of all vectors $\mathbf{y}$ such that $T(\mathbf{y}) = t$.

Substituting the factorized form of $p(\mathbf{y};\theta)$ into equation (24.3), we obtain:

$$\mathbb{P}(\mathbf{Y}=\mathbf{y} \mid T(\mathbf{Y})=t;\theta) = \frac{K_1(T(\mathbf{y});\theta)\cdot K_2(\mathbf{y})}{K_1(T(\mathbf{y});\theta)\cdot \sum_{\mathbf{y}\in A_t} K_2(\mathbf{y})} = \frac{K_2(\mathbf{y})}{\sum_{\mathbf{y}\in A_t} K_2(\mathbf{y})},$$

which does not depend on θ. This proves that $T(\mathbf{Y})$ is a sufficient statistic for θ.

■

▪ **Example 24.4 — Gamma with *known* shape parameter.** Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables with a Gamma distribution of shape parameter 2 and rate parameter θ. The density is:

$$f(x;\theta) = \begin{cases} \theta^2 x e^{-\theta x}, & \text{if } x \geq 0, \\ 0, & \text{otherwise.} \end{cases}$$

The joint density of the sample is then:

$$f(\mathbf{x};\theta) = \begin{cases} \theta^{2n} \left(\prod x_i\right) e^{-\theta \sum x_i}, & \text{if } x_i \geq 0 \text{ for all } i, \\ 0, & \text{otherwise.} \end{cases}$$

We can write this as:

$$K_1\left(\sum x_i;\theta\right) = \theta^{2n} e^{-\theta \sum x_i}, \quad K_2(\mathbf{x}) = \prod_{i=1}^{n} x_i.$$

Therefore, $\sum x_i$ is a sufficient statistic for θ.

▪

▪ **Example 24.5 — General Gamma has no sufficient statistic.** The previous example is quite artificial as it requires that the shape parameter α must have a known value. The common situation is that when neither α and β will be known. We will show that, in this case, there is no simple bi-dimensional sufficient statistic.

Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables with a Gamma distribution of shape parameter α and rate parameter β with $\theta = (\alpha, \beta)$. The density is:

$$f(x;\theta) = \begin{cases} \frac{\theta^\alpha}{\Gamma(\alpha)} x e^{-\theta x}, & \text{if } x \geq 0, \\ 0, & \text{otherwise.} \end{cases}$$

TO COMPLETE....

▪

■ **Example 24.6 — Weibull has no sufficient statistic.** TO COMPLETE....Estah nos slides, ultimos slides.... ■

24.4 Multidimensional Case

Definition 24.4.1 — Sufficiency for multidimensional θ. When the parameter of interest is a vector, that is, $\theta = (\theta_1, \theta_2, \ldots, \theta_k)$, we say that $T(\mathbf{y}) = (T_1(\mathbf{y}), T_2(\mathbf{y}), \ldots, T_m(\mathbf{y}))$ is sufficient for θ if and only if

$$f(\mathbf{y};\theta) = K_1(T(\mathbf{y});\theta) \cdot K_2(\mathbf{y}) .$$

In general, the number of sufficient statistics m is equal to the number of parameters k being estimated ($m = k$), but it is also possible that $m \neq k$.

■ **Example 24.7 — Sufficiency in the Gaussian Case.** Let $Y_1, Y_2, \ldots, Y_n$ be i.i.d. random variables with distribution $N(\mu, \sigma^2)$. In this case, the parameter vector is $\theta = (\mu, \sigma^2)$. The joint density is:

$$\begin{aligned} f(\mathbf{y};\theta) &= \prod_{i=1}^{n} \left[\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y_i - \mu)^2 \right) \right] \\ &= (2\pi\sigma^2)^{-n/2} \cdot \exp\left(-\frac{1}{2\sigma^2} \sum_i (y_i - \mu)^2 \right) \\ &= (2\pi\sigma^2)^{-n/2} \cdot \exp\left(-\frac{\sum_i y_i^2}{2\sigma^2} + \frac{\mu \sum_i y_i}{\sigma^2} - \frac{n\mu^2}{2\sigma^2} \right) \\ &= \left[(2\pi\sigma^2)^{-n/2} \cdot \exp\left(-\frac{n\mu^2}{2\sigma^2} \right) \right] \cdot \exp\left(-\frac{T_1(\mathbf{y})}{2\sigma^2} + \frac{\mu T_2(\mathbf{y})}{\sigma^2} \right) \end{aligned}$$

where $T(\mathbf{y}) = (T_1(\mathbf{y}), T_2(\mathbf{y})) = \left(\sum_i y_i^2, \sum_i y_i\right)$.

In this example, we can write $f(\mathbf{y};\theta) = g(T(\mathbf{y}),\theta) \cdot h(\mathbf{y})$ with $h(\mathbf{y}) = 1$. The first multiplicative factor in $g(T(\mathbf{y}),\theta)$ involves only the parameters μ and σ^2, not the data. The second factor involves both the data and the parameters, but the data appear only through the summaries:

$$T_1(\mathbf{y}) = \sum_i y_i^2 \quad \text{and} \quad T_2(\mathbf{y}) = \sum_i y_i .$$

Thus, $T(\mathbf{y}) = \left(\sum_i y_i^2, \sum_i y_i\right)$ is a sufficient statistic for estimating $\theta = (\mu, \sigma^2)$. In fact, the MLE of θ is a function of this sufficient statistic. We have:

$$\widehat{\mu}_{\text{MLE}} = \overline{Y} = \frac{1}{n}\sum_i Y_i = \frac{T_2(\mathbf{Y})}{n},$$

and

$$\widehat{\sigma^2}_{\text{MLE}} = \frac{1}{n}\sum_i (Y_i - \overline{Y})^2 = \frac{1}{n}\sum_i Y_i^2 - \overline{Y}^2 = \frac{T_1(\mathbf{Y})}{n} - \left(\frac{T_2(\mathbf{Y})}{n}\right)^2 .$$

Therefore, the MLE is obtained directly from the sufficient statistic $\mathbf{T}(\mathbf{Y})$. ■

24.5 Sufficiency in Linear Regression

The Neyman–Fisher factorization theorem allows us to easily identify sufficient statistics in the linear regression setting. We will conclude that both the least squares estimator and the MLE are functions of a sufficient statistic.

Let $Y_1, Y_2, \ldots, Y_n$ be independent but not identically distributed random variables with $Y_i \sim N(\mu_i, \sigma^2)$, where

$$\begin{aligned}\mu_i &= \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} \\ &= \beta_0 + \sum_j x_{ij}\beta_j \\ &= (1, x_{i1}, \ldots, x_{ip})(\beta_0, \beta_1, \ldots, \beta_p)' \\ &= \mathbf{x}_i'\beta.\end{aligned}$$

Here, $\mathbf{x}_i = (1, x_{i1}, \ldots, x_{ip})$ is a $(p+1)$-dimensional vector of constants (not random variables) representing the features of the i-th observation. The parameter vector is $\theta = (\beta, \sigma^2) = (\beta_0, \ldots, \beta_p, \sigma^2)$.

The joint density is similar to the i.i.d. case, except that the expected value μ_i now depends on i rather than being constant:

$$\begin{aligned}f(\mathbf{y};\theta) &= \prod_{i=1}^{n}\left[\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(y_i-\mu_i)^2}{2\sigma^2}\right)\right] \\ &= (2\pi\sigma^2)^{-n/2}\exp\left(-\frac{1}{2\sigma^2}\sum_i (y_i-\mu_i)^2\right) \\ &= (2\pi\sigma^2)^{-n/2}\exp\left(-\frac{\sum_i y_i^2}{2\sigma^2} + \frac{\sum_i \mu_i y_i}{\sigma^2} - \frac{\sum_i \mu_i^2}{2\sigma^2}\right) \\ &= \left[(2\pi\sigma^2)^{-n/2}\exp\left(-\frac{\sum_i \mu_i^2}{2\sigma^2}\right)\right]\exp\left(-\frac{\sum_i y_i^2}{2\sigma^2} + \frac{\sum_i \mu_i y_i}{\sigma^2}\right).\end{aligned}$$

We now want to rewrite the expression in terms of the parameter vector θ by replacing μ_i with $\mathbf{x}_i'\beta$. Define

$$k(\theta) = (2\pi\sigma^2)^{-n/2}\exp\left(-\frac{\sum_i \mu_i^2}{2\sigma^2}\right),$$

which does not depend on the data. Substituting into the expression for $f(\mathbf{y};\theta)$:

$$f(\mathbf{y};\theta) = k(\theta)\exp\left(-\frac{\sum_i y_i^2}{2\sigma^2} + \frac{\sum_i \mu_i y_i}{\sigma^2}\right).$$

Now we expand the term $\sum_i \mu_i y_i$:

$$\begin{aligned}\sum_i \mu_i y_i &= \sum_i y_i(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}) \\ &= \beta_0\sum_i y_i + \beta_1\sum_i y_i x_{i1} + \ldots + \beta_p\sum_i y_i x_{ip}.\end{aligned}$$

Thus,

$$f(\mathbf{y};\theta) = k(\theta)\exp\left(-\frac{\sum_i y_i^2}{2\sigma^2} + \frac{1}{\sigma^2}\left[\beta_0 T_0(\mathbf{y}) + \beta_1 T_1(\mathbf{y}) + \ldots + \beta_p T_p(\mathbf{y})\right]\right),$$

where the sufficient statistic is:

$$\mathbf{T}(\mathbf{y}) = (T_0(\mathbf{y}), T_1(\mathbf{y}), \ldots, T_p(\mathbf{y}), T_{p+1}(\mathbf{y})) = \left(\sum_i y_i, \sum_i y_i x_{i1}, \ldots, \sum_i y_i x_{ip}, \sum_i y_i^2\right).$$

By the Neyman–Fisher factorization theorem, $\mathbf{T}(\mathbf{y})$ is a sufficient statistic for θ.

We can write this statistic more compactly using matrix notation, which is especially useful for generalized linear models. The first $p+1$ components of $\mathbf{T}(\mathbf{y})$ can be written as a column-vector:

$$(T_0(\mathbf{y}), T_1(\mathbf{y}), \ldots, T_p(\mathbf{y}), T_{p+1}(\mathbf{y}))^t = \left(\sum_i y_i, \sum_i y_i x_{i1}, \ldots, \sum_i y_i x_{ip} \right)^t = \mathbf{X}^t \mathbf{y},$$

where

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}, \qquad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}.$$

Another way to see the first $p+1$ components of $\mathbf{T}(\mathbf{y})$ is to recognize that each entry in $\mathbf{T}(\mathbf{y})$ is the inner product between a column of the $\mathbf{X}$ matrix and the response vector $\mathbf{y}$:

$$T_j(\mathbf{y}) = \langle \mathbf{X}_{\cdot,j}, \mathbf{y} \rangle = \mathbf{X}_{\cdot,j}^t \mathbf{y}$$

Recall that the first column of $\mathbf{X}$ is the vector $\mathbf{1} = (1, 1, \ldots, 1)^t$ and hence, the first component is $T_0(\mathbf{y}) = \sum_i y_i = \langle \mathbf{1}, \mathbf{y} \rangle$.

The last component $T_{p+1}(\mathbf{y}) = \sum_i y_i^2$ is simply the squared norm of $\mathbf{y}$:

$$T_{p+1}(\mathbf{y}) = \mathbf{y}^t \mathbf{y}.$$

Hence, the sufficient statistic for linear regression is:

$$\mathbf{T}(\mathbf{y}) = \left(\mathbf{X}^t \mathbf{y}, \mathbf{y}^t \mathbf{y} \right).$$

Note that the MLE of β is a function of this sufficient statistic:

$$\widehat{\beta} = (\mathbf{X}^t \mathbf{X})^{-1} \mathbf{X}^t \mathbf{y},$$

since $\mathbf{X}^t \mathbf{X}$ is a matrix of constants in discriminative models such as linear regression.

The MLE of σ^2 is also a function of the sufficient statistic:

$$\widehat{\sigma^2} = \frac{1}{n} \sum_i (y_i - \widehat{y}_i)^2 = \frac{1}{n} \|\mathbf{y} - \widehat{\mathbf{y}}\|^2,$$

where $\widehat{\mathbf{y}} = \mathbf{X}\widehat{\beta}$ is itself a function of $\mathbf{T}(\mathbf{y})$.

The vector $\widehat{\mathbf{y}}$ is the orthogonal projection of $\mathbf{y}$ onto the vector space spanned by the columns of the design matrix $\mathbf{X}$. That is,

$$\mathbf{y} = \widehat{\mathbf{y}} + \mathbf{r}, \quad \text{with } \mathbf{r} = \mathbf{y} - \widehat{\mathbf{y}} \text{ and } \widehat{\mathbf{y}} \perp \mathbf{r}.$$

By the Pythagorean theorem:

$$\|\mathbf{y}\|^2 = \|\widehat{\mathbf{y}}\|^2 + \|\mathbf{y} - \widehat{\mathbf{y}}\|^2.$$

Therefore, the MLE $\widehat{\sigma^2}$ is a function of the sufficient statistic $\mathbf{T}(\mathbf{y})$:

$$\widehat{\sigma^2} = \frac{1}{n} \|\mathbf{y} - \widehat{\mathbf{y}}\|^2 = \frac{1}{n} \left(\|\mathbf{y}\|^2 - \|\widehat{\mathbf{y}}\|^2 \right) = \frac{1}{n} \left(T_{p+1}(\mathbf{y}) - \|\widehat{\mathbf{y}}\|^2 \right).$$

24.6 Sufficiency in Logistic Regression

Recall the setup: we observe $Y_1, \ldots, Y_n$, which are independent but not identically distributed binary random variables. The probability of success $p_i = \mathbb{P}(Y_i = 1)$ varies across observations. Let $\mathbf{x}_i = (1, x_{i1}, \ldots, x_{ip})$ be the feature vector for observation i. Let $\theta = (\beta_0, \beta_1, \ldots, \beta_p)$ denote the unknown parameter (weight) vector.

Let us recall the matrix notation:

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1p} \\ 1 & x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{np} \end{bmatrix} \qquad \theta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} \qquad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

We assume:

$$\mathbb{P}(Y_i = 1) = p_i = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip})}} = \frac{1}{1 + e^{-\mathbf{x}_i^t \theta}} = \frac{1}{1 + e^{-\eta_i}}$$

where $\mathbf{x}_i$ is the i-th row of the matrix $\mathbf{X}$, seen as a column vector, and θ is the column vector of parameters. We call $\eta_i = \mathbf{x}_i^t \theta$ the linear predictor for success. We aim to estimate the parameter vector $\theta = (\beta_0, \beta_1, \ldots, \beta_p)$.

The joint likelihood is:

$$\begin{aligned} p(\mathbf{y}; \theta) &= \prod_{i=1}^{n} \mathbb{P}(Y_i = 1)^{y_i} \cdot \mathbb{P}(Y_i = 0)^{1 - y_i} \\ &= \prod_{i=1}^{n} \left(\frac{p_i}{1 - p_i} \right)^{y_i} \cdot (1 - p_i) \end{aligned}$$

Using the logistic expression for p_i, we find:

$$\frac{p_i}{1 - p_i} = \exp\left(\mathbf{x}_i^t \theta\right)$$

Thus,

$$\begin{aligned} p(\mathbf{y}; \theta) &= \prod_{i=1}^{n} e^{y_i \mathbf{x}_i^t \theta} \cdot \prod_{i=1}^{n} \left(1 - \frac{1}{1 + e^{-\mathbf{x}_i^t \theta}} \right) \\ &= \exp\left(\sum_{i=1}^{n} y_i \mathbf{x}_i^t \theta \right) \cdot \prod_{i=1}^{n} \left(1 - \frac{1}{1 + e^{-\mathbf{x}_i^t \theta}} \right) \end{aligned}$$

Focusing on the exponential term:

$$\begin{aligned} \sum_{i=1}^{n} y_i \mathbf{x}_i^t \theta &= \sum_{i=1}^{n} y_i (\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}) \\ &= \beta_0 \sum_i y_i + \beta_1 \sum_i y_i x_{i1} + \ldots + \beta_p \sum_i y_i x_{ip} \\ &= \left(\sum_i y_i, \sum_i y_i x_{i1}, \ldots, \sum_i y_i x_{ip} \right)^t \cdot \theta \\ &= \left(\mathbf{X}^t \mathbf{y} \right)^t \theta \end{aligned}$$

Note that $\mathbf{X}^t \mathbf{y}$ is a $(p+1) \times 1$ vector. In conclusion, the joint likelihood is:

$$p(\mathbf{y}; \theta) = \exp\left(\left(\mathbf{X}^t \mathbf{y} \right)^t \theta \right) \cdot \prod_{i=1}^{n} \left(1 - \frac{1}{1 + e^{-\mathbf{x}_i^t \theta}} \right)$$

The second multiplicative term depends only on θ (not on the data $\mathbf{y}$). The exponential term depends on both θ and $\mathbf{y}$, but the data appears only through the vector:

$$\mathbf{T}(\mathbf{y}) = \mathbf{X}^t\mathbf{y} = \left(\sum_i y_i, \sum_i y_i x_{i1}, \ldots, \sum_i y_i x_{ip}\right)$$

Thus, $\mathbf{T}(\mathbf{y}) = \mathbf{X}^t\mathbf{y}$ is a sufficient statistic for estimating θ.

Note that since $y_i \in \{0,1\}$, the j-th entry of $\mathbf{T}(\mathbf{y})$ corresponds to the sum of the j-th covariate values over all observations where $y_i = 1$. That is, $\mathbf{T}(\mathbf{y})$ is a vector of partial sums, considering only the success cases. However, a more useful approach is to recognize that, in the same way as the linear regression model, each entry in $\mathbf{T}(\mathbf{y})$ is the inner product between a column of the $\mathbf{X}$ matrix and the response vector $\mathbf{y}$:

$$T_j(\mathbf{y}) = \langle \mathbf{X}_{\cdot,j}, \mathbf{y} \rangle = \mathbf{X}_{\cdot,j}^t \mathbf{y}$$

Recall the maximum likelihood estimation (MLE) procedure for logistic regression: the MLE $\widehat{\theta}$ is the parameter vector such that the induced vector of success probabilities $\mathbf{p} = (p_1, \ldots, p_n)$ satisfies the likelihood equation:

$$\mathbf{X}^t\mathbf{y} = \mathbf{X}^t\mathbf{p}$$

Therefore, to compute the MLE we need only the design matrix $\mathbf{X}$ and, from the data, only the sufficient statistic $\mathbf{T}(\mathbf{y}) = \mathbf{X}^t\mathbf{y}$. That is, the MLE is a function of the sufficient statistic.

24.7 Existence and non-uniqueness of sufficient statistics

When does a sufficient statistic exist? And if it exists, is it unique? We will address these two questions in this section. The answer to the first question is simple: a sufficient statistic always exists. In any statistical problem, there is at least one sufficient statistic, although it is not be particularly useful. Specifically, the data vector itself, $\mathbf{Y}$, is always a sufficient statistic. That is, consider the identity function $T(\mathbf{Y}) = \mathbf{Y}$. Then, the conditional distribution of $\mathbf{Y}$ given $T(\mathbf{Y}) = \mathbf{Y} = \mathbf{y}$ is simply a point mass at $\mathbf{y}$.

■ **Example 24.8** Let Y_1, Y_2, Y_3 be three i.i.d. continuous random variables with a distribution depending on a parameter θ. Suppose that we observe $\mathbf{Y} = (1.27, 3.28, 5.92)$. What is the distribution of $\mathbf{Y}$ given that we know that it is equal to $(1.27, 3.28, 5.92)$? It is simply

$$\mathbb{P}(\mathbf{Y} = \mathbf{y} | \mathbf{Y} = (1.27, 3.28, 5.92); \theta) = \begin{cases} 1 & \text{if } \mathbf{y} = (1.27, 3.28, 5.92) \\ 0 & \text{if } \mathbf{y} \neq (1.27, 3.28, 5.92) \end{cases}$$

Therefore, this conditional distribution does not depend on the unknown parameter θ, regardless of the original distribution of $\mathbf{Y}$. Thus, $T(\mathbf{Y}) = \mathbf{Y}$ is always a sufficient statistic, no matter which parametric distribution is being considered. ■

This example highlights that there is always at least one sufficient statistic for a given model. Another sufficient statistic that always exists is the vector of order statistics, $\mathbf{Y}_{(n)} = (Y_{(1)}, \ldots, Y_{(n)})$, which is the n-dimensional vector of the sample sorted in increasing order (see the optional reading in Section~**??**subsec order_stats_is_sufficient). While both $T(\mathbf{Y}) = \mathbf{Y}$ and $T(\mathbf{Y}) = \mathbf{Y}_{(n)}$ are always sufficient, they do not reduce the dimensionality of the data or provide interpretable summaries. Therefore, the key challenge is not to find some sufficient statistic, but rather to identify a *low-dimensional* sufficient statistic that still captures all the relevant information about the parameter.

More importantly, we seek not only a low-dimensional sufficient statistic, but one whose dimension is *fixed*. That is, its dimension does not increase with the sample size. Note that both $T(\mathbf{Y}) = \mathbf{Y}$ and $T(\mathbf{Y}) = \mathbf{Y}_{(n)}$ are vectors of the same length as the sample itself.

Thus, the more interesting question is: under what conditions does there exist a sufficient statistic $T(\mathbf{Y})$ that is a vector of **fixed dimension**, independent of the number of observations n? Answer: If and only if $\mathbf{Y}$ belongs to an exponential family of distributions. This is the content of the Koopman–Pitman–Darmois Theorem, independently proven by the three authors and published between 1935 and 1936.

24.7.1 Optional reading: Another not-so-useful sufficient statistics

There is another sufficient statistic in any arbitrary parametric distribution. Let $Y_1, \ldots, Y_n$ be i.i.d. continuous random variables from a distribution with density $f(y;\theta)$, where θ is an unknown parameter. Define the order statistics $Y_{(1)} \leq Y_{(2)} \leq \ldots \leq Y_{(n)}$, and let $\mathbf{Y}_{(n)} = (Y_{(1)}, \ldots, Y_{(n)})$ denote the vector of order statistics. Then these order statistics are a sufficient statistic. Although order statistics play a central role in various areas of statistical analysis, they are generally not very informative as sufficient statistics, since they do not typically provide dimension reduction or meaningful summarization of the data with respect to the parameter of interest.

Theorem 24.7.1 — Sufficiency of order statistics. The order statistics vector $\mathbf{Y}_{(n)}$ is a sufficient statistic for θ, regardless of the specific form of $f(y;\theta)$, provided the variables are i.i.d. and have a continuous distribution.

Proof. The joint density of $\mathbf{Y} = (Y_1, \ldots, Y_n)$ can be written as

$$f(\mathbf{y};\theta) = \prod_{i=1}^{n} f(y_i;\theta)$$

Since the Y_i are i.i.d. with continuous distribution, the joint density of the order statistics $\mathbf{Y}_{(n)}$ is given by

$$f_{\mathbf{Y}_{(n)}}(y_{(1)}, \ldots, y_{(n)};\theta) = n! \cdot \prod_{i=1}^{n} f(y_{(i)};\theta)$$

To prove that the joint density of the **order statistics** $\mathbf{Y}_{(n)} = (Y_{(1)}, \ldots, Y_{(n)})$ is given by

$$f_{\mathbf{Y}_{(n)}}(y_{(1)}, \ldots, y_{(n)};\theta) = n! \cdot \prod_{i=1}^{n} f(y_{(i)};\theta),$$

The order statistics $(Y_{(1)}, \ldots, Y_{(n)})$ is the sorted version of $(Y_1, \ldots, Y_n)$. The transformation from unordered data to order statistics is many-to-one: each ordered vector corresponds to $n!$ permutations of the original i.i.d. vector. That is, the set

$$\{\mathbf{y} = (y_1, \ldots, y_n) : \mathbf{y} \text{ is a sorted version is } (y_{(1)}, \ldots, y_{(n)})\}$$

contains $n!$ elements, each corresponding to a different permutation. The density of the order statistics at a point $(y_{(1)}, \ldots, y_{(n)})$ is the sum of the densities over all permutations of that point:

$$f_{\mathbf{Y}_{(n)}}(y_{(1)}, \ldots, y_{(n)};\theta) = \sum_{\pi \in S_n} f(y_{\pi(1)}, \ldots, y_{\pi(n)};\theta)$$

Since all Y_i are i.i.d., every permutation yields the same product of densities:

$$f(y_{\pi(1)}, \ldots, y_{\pi(n)};\theta) = \prod_{i=1}^{n} f(y_{(i)};\theta) \quad \text{for all } \pi.$$

There are $n!$ permutations in total, so:

$$f_{\mathbf{Y}_{(n)}}(y_{(1)},\ldots,y_{(n)};\theta) = n! \cdot \prod_{i=1}^{n} f(y_{(i)};\theta).$$

This density holds for vectors satisfying $y_{(1)} < y_{(2)} < \cdots < y_{(n)}$. Outside of this domain, the density is zero. Hence,

$$f_{\mathbf{Y}_{(n)}}(y_{(1)},\ldots,y_{(n)};\theta) = \begin{cases} n! \cdot \prod_{i=1}^{n} f(y_{(i)};\theta), & \text{if } y_{(1)} < \cdots < y_{(n)} \\ 0, & \text{otherwise} \end{cases}$$

Now consider the conditional distribution of $\mathbf{Y}$ given $\mathbf{Y}_{(n)}$. Since $\mathbf{Y}$ is just a permutation of its order statistics, this conditional distribution is uniform over all $n!$ permutations of the ordered vector $\mathbf{Y}_{(n)}$ and does not depend on θ. Hence, by the definition of sufficiency, the order statistics $\mathbf{Y}_{(n)}$ are a sufficient statistic for θ. ■

24.8 Theorem of Darmois-Koopman-Pitman

The Darmois-Koopman-Pitman theorem [1, 6, 7] is a fundamental result in mathematical statistics that characterizes the class of probability distributions that allow for finite-dimensional and fixed sufficient statistics. This class is called exponential distribution class. The theorem essentially states that the exponential family is the only class of distributions that consistently allows for a fixed-dimension sufficient statistic for independent and identically distributed (i.i.d.) observations.

The three authors giving name to the theorem established similar versions of it, independently at about the same time, in 1935 and 1936. At that time, the difficulties of communication and relative isolation of the researchers made this a relatively common outcome. Their work was based on a previous result by Fisher [5] who proved the theorem for one-dimensional parametric families.

Theorem 24.8.1 — Darmois-Koopman-Pitman Theorem. Let $f(y;\theta)$ be the density of the random variable Y and let $Y_1,\ldots,Y_n$ be a sample of i.i.d. random variables with the marginal distribution as that of Y. Under certain regularity conditions, a sufficient statistic $\mathbf{T}(\mathbf{Y}) = (T_1(\mathbf{Y}),\ldots,T_k(\mathbf{Y}))$ exists whose dimension k is fixed and does not increase with the sample size n if and only if the family of distributions is of the following form:

$$f(y;\theta) = \exp\left(\sum_{j=1}^{k} \eta_j(\theta) T_j(y) + C(y) + A(\theta)\right)$$

Proof. We will provide only an elementary sketch of the proof for the simpler case in which θ and the sufficient statistic are uni-dimensional (scalar). This proof is based on the lecture notes by Olivier Gaudoin [pp. 18–19]olivierstatistique.

Assume a sufficient statistic $t(\mathbf{y})$ exists. Then, the likelihood can be written as:

$$L(\theta;y_1,\ldots,y_n) = \prod_{i=1}^{n} f(y_i;\theta) = g(t(y_1,\ldots,y_n);\theta)h(y_1,\ldots,y_n)$$

Taking logs:

$$\log L(\theta;y_1,\ldots,y_n) = \sum_{i=1}^{n} \log f(y_i;\theta) = \log g(t(y_1,\ldots,y_n);\theta) + \log h(y_1,\ldots,y_n)$$

Since h does not depend on θ, we differentiate both sides with respect to θ:

$$\sum_{i=1}^{n} \frac{\partial}{\partial\theta} \log f(y_i;\theta) = \frac{\partial}{\partial\theta} \log g(t(y_1,\ldots,y_n);\theta)$$

Now fix any index i and consider:

$$\frac{\partial^2}{\partial\theta\partial y_i}\log L(\theta;y_1,\ldots,y_n)=\frac{\partial^2}{\partial\theta\partial y_i}\log f(y_i;\theta)=\frac{\partial^2}{\partial\theta\partial y_i}\log g(t(y_1,\ldots,y_n);\theta)$$

Applying the chain rule:

$$\frac{\partial^2}{\partial\theta\partial y_i}\log g(t(\mathbf{y});\theta)=\frac{\partial t(\mathbf{y})}{\partial y_i}\cdot\frac{\partial^2}{\partial\theta\partial z}\log g(z;\theta)\bigg|_{z=t(\mathbf{y})}$$

Similarly, for another index $j\neq i$:

$$\frac{\partial^2}{\partial\theta\partial y_j}\log f(y_j;\theta)=\frac{\partial t(\mathbf{y})}{\partial y_j}\cdot\frac{\partial^2}{\partial\theta\partial z}\log g(z;\theta)\bigg|_{z=t(\mathbf{y})}$$

Therefore,

$$\frac{\partial^2}{\partial\theta\partial y_i}\log f(y_i;\theta)\cdot\frac{\partial^2}{\partial\theta\partial y_j}\log f(y_j;\theta)=\frac{\partial t}{\partial y_i}\cdot\frac{\partial t}{\partial y_j}\cdot\left(\frac{\partial^2}{\partial\theta\partial z}\log g(z;\theta)\right)^2$$

The left-hand side depends on y_i, y_j, and θ, but the right-hand side factorizes in such a way that for the equality to hold for all θ, the mixed derivative term must separate:

$$\frac{\partial^2}{\partial\theta\partial y}\log f(y;\theta)=u(y)v(\theta)$$

That is, the mixed derivative must be of the form:

$$\frac{\partial^2}{\partial\theta\partial y}\log f(y;\theta)=u(y)v(\theta)$$

Integrating twice yields:

$$\log f(y;\theta)=a(y)b(\theta)+c(y)+d(\theta)\Rightarrow f(y;\theta)=h(y)\exp(a(y)b(\theta)+d(\theta))$$

which is the exponential family form.

■

24.9 MLE is a function of sufficient statistics

Suppose that the Darmois-Koopman-Pitman applies. The sufficient statistic $T(\mathbf{Y})$ summarizes all the information about θ contained in the data. The distribution of the data $\mathbf{Y}$ conditional on the observed value of $T(\mathbf{Y})$ does not depend on θ. What is the relationship between the MLE $\hat{\theta}$ and the sufficient statistic $T(\mathbf{Y})$? It is summarized in the next theorem.

Theorem 24.9.1 — MLE is a function of the sufficient statistic. If there exists a sufficient statistic $\mathbf{T}(\mathbf{Y})$ of finite dimension, the MLE $\hat{\theta}$ is a function of the sufficient statistic: $\hat{\theta}=g(\mathbf{T}(\mathbf{Y}))$.

Proof. By the factorization theorem, $f(\mathbf{y},\theta)=g(T(\mathbf{y}),\theta)h(\mathbf{y})$. To maximize $f(\mathbf{y},\theta)$ with respect to θ, we must maximize $g(T(\mathbf{y}),\theta)$. Since the data enter only through the summary $T(\mathbf{y})$, the solution $\widehat{\theta}_{\text{MLE}}$ that maximizes $g(T(\mathbf{y}),\theta)$ will depend on the data only through $T(\mathbf{y})$. That is, $\widehat{\theta}_{\text{MLE}}$ is a function of $T(\mathbf{Y})$. ■

■ **Example 24.9** Let $Y_1, \ldots, Y_n$ be i.i.d. Poisson(θ). Then

$$\begin{aligned} f(\mathbf{y}, \theta) &= e^{-n\theta}\theta^{\sum_i y_i} \prod_i (y_i!)^{-1} \\ &= g(T(\mathbf{y}), \theta)\, h(\mathbf{y}), \end{aligned}$$

where $g(t, \theta) = e^{-n\theta}\theta^t$ and $T(\mathbf{y}) = \sum_i y_i$. The maximum of $g(t, \theta)$ occurs at $\widehat{\theta} = t/n$. This maximum depends on the individual counts y_i only through their sum $t = \sum_i y_i$.

Consider two different realizations $\mathbf{y}^1 \neq \mathbf{y}^2$ that yield the same value for the sufficient statistic (i.e., the same total count):

$$T(\mathbf{y}^1) = \sum_i y_i^1 = t = \sum_i y_i^2 = T(\mathbf{y}^2).$$

Then, the MLE based on $\mathbf{y}^1$ is the same as the one based on $\mathbf{y}^2$. ■

24.10 MLE is approximately sufficient

Theorem 24.10.1 — MLE is approximately sufficient. Even when the Darmois-Koopman-Pitman does not apply and we do not have a finite-dimensional sufficient statistic, the MLE $\hat{\theta}_n$ of $\theta \in \mathbb{R}^d$ will be approximately a sufficient statistic if the sample size is not too small. That is:

$$\log(f(\mathbf{y}; \theta)) = G(\hat{\theta}_n; \theta) + H(\mathbf{y}).$$

Proof. We provide an informal sketch of the proof. We factorize the likelihood:

$$\begin{aligned} L_n(\theta) = f(\mathbf{y}, \theta) &= \prod_{i=1}^n f(y_i; \theta) \\ &= \exp\left(\log \prod_{i=1}^n f(y_i; \theta)\right) \\ &= \exp\left(\sum_{i=1}^n \log f(y_i; \theta)\right) \\ &= \exp(\ell_n(\theta)) \end{aligned}$$

Using a second-order Taylor expansion around $\hat{\theta}_n$:

$$\ell_n(\theta) \approx \ell_n(\hat{\theta}_n) + (\theta - \hat{\theta}_n)\ell_n'(\hat{\theta}_n) + \frac{1}{2}(\theta - \hat{\theta}_n)^2 \ell_n''(\hat{\theta}_n)$$

Since $\hat{\theta}_n$ maximizes the log-likelihood, $\ell_n'(\hat{\theta}_n) = 0$ and then:

$$\ell_n(\theta) \approx \ell_n(\hat{\theta}_n) + \frac{1}{2}(\theta - \hat{\theta}_n)^2 \ell_n''(\hat{\theta}_n)$$

Recall that $\ell_n(\theta) = \log f(\mathbf{y}; \theta)$. $\hat{\theta}_n$ is the MLE and therefore DO NOT INVOLVE the unknown parameter θ. Therefore, $\ell_n(\hat{\theta}_n) = \log f(\mathbf{y}; \hat{\theta}_n)$ is not a function of θ it is a function only of the data $\mathbf{y}$. That is: $\ell_n(\hat{\theta}_n) = H(\mathbf{y})$

$$\begin{aligned} \ell_n(\theta) &\approx \overbrace{\ell_n(\hat{\theta}_n)}^{\text{constant in } \theta} + \frac{1}{2}(\theta - \hat{\theta}_n)^2 \ell_n''(\hat{\theta}_n) \\ &= H(\mathbf{y}) + \frac{1}{2}(\theta - \hat{\theta}_n)^2 \ell_n''(\hat{\theta}_n) \end{aligned} \tag{24.4}$$

Now, consider $(1/n)\ell_n''(\theta)$. By the Law of Large Numbers,

$$\frac{1}{n}\ell_n''(\theta) = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2 \log(f(y_i;\theta))}{\partial\theta^2} \to \mathbb{E}_\theta\left(\frac{\partial^2\ell}{\partial\theta^2}\right) = -\mathbb{I}_1(\theta)$$

Recall that $\mathbb{I}_1(\theta)$ is an expected value and therefore contains no random elements. It is a function of the parameter θ, not of the random data $\mathbf{Y}$. In what follows, we evaluate $\mathbb{I}_1(\theta)$ at the maximum likelihood estimator $\hat{\theta}_n$. As a result, $\mathbb{I}_1(\hat{\theta}_n)$ becomes a function of the data $\mathbf{Y}$, but only through the estimator $\hat{\theta}_n$.

Returning to (24.4), we have

$$\begin{aligned}
\log(f(\mathbf{y};\theta)) &\approx H(\mathbf{y}) + \frac{1}{2}(\theta-\hat{\theta}_n)^2\ell_n''(\hat{\theta}_n)\\
&= H(\mathbf{y}) + \frac{1}{2}(\theta-\hat{\theta}_n)^2\frac{n}{n}\ell_n''(\hat{\theta}_n)\\
&\approx H(\mathbf{y}) - \frac{1}{2}(\theta-\hat{\theta}_n)^2 n\mathbb{I}_1(\hat{\theta}_n)\\
&= H(\mathbf{y}) + G(\hat{\theta}_n;\theta)
\end{aligned}$$

Exponentiating both sides of the approximation:

$$f(\mathbf{y};\theta)) \approx e^{H(\mathbf{y})}\cdot e^{G(\hat{\theta}_n,\theta)} = h(\mathbf{y})\; g(\hat{\theta}_n,\theta)$$

Conclusion: MLE $\hat{\theta}_n$ is asymptotically sufficient to estimate θ. ■

24.11 Missing

- Explicar estimadores de Rao-Blackwell
- Seja $L(d(X),\theta)$ uma função de perda convexa qualquer e $R(d(X),\theta) = \mathbb{E}(L(d(X),\theta))$ o risco de estimar θ usando $d(X)$.
- Seja $T(X)$ estatística suficiente para θ e um estimador $\delta(X)$ qualquer.
- Então $\delta(X)$ é dominado por $g(T(X)) = \mathbb{E}(L(\delta(X)|T(X))$ no sentido de que $R(\delta(X),\theta) \geq R(g(T(X)),\theta)$ para todo θ with strict inequality unless $g(T(X)) = \delta(X)$ with probability one.
- Assim, é sempre melhor condicionar na estaística suficiente para estimar θ.
- The need for a non-trivial sufficient statistic obviously restricts the applicability of this theorem in mathematical statistics to exponential families.
- Note que δ não precisa ser não-viciado. Se for, então $g(T(X))$ também será.

25. Generalized Linear Models (GLM)

25.1 Introdução

In exponential family distributions, we have seen that *bsT* é sufficient and dimensão de T não precisa ser igual a de θ.

Ao buscar uma estatística $T(\mathbf{X}) = (T_1(\mathbf{X}), \ldots, T_j(\mathbf{X})$ que seja suficiente para estimar o parâmetro $\theta = (\theta_1, \ldots, \theta_k)$, é importante que o vetor das estatísticas suficiente seja bem menor que o número total n de observações. O melhor é encontrar uma estatística suficiente cuja dimensão vetorial não cresça com o aumento do número de observações. Vamos ver alguns (???, um ???) exemplo onde isto *não* ocorre.

Exemplo: Amostra de tamanho n de Weibull com densidade

$$f(\mathbf{x};\theta) = f(\mathbf{x};\alpha,\beta) = \begin{cases} \alpha\beta x^{\alpha-1} exp(-\beta x^{\alpha}), & \text{se } x > 0 \\ 0, & \text{caso contrário.} \end{cases}$$

Temos então a log-verossimilhança

$$l(\theta) = n\log(\alpha\beta) + (\alpha - 1)\sum_i \log(x_i) - \beta\sum_i x_i^{\alpha}$$

Pelo teorema da fatoração de Fisher-Neyman, a única estatística suficiente é o próprio vetor completo das observações $\mathbf{X} = (X_1, \ldots, X_n)$. Assim, a dimensão da única estatística suficiente, o vetor $\mathbf{T}(\mathbf{x}) = \mathbf{X}$, é igual a n, a dimensão do vetor de dados. Não existe estatística suficiente de dimensão menor que n.

Dito de outra forma, ão existe um pequeno (e fixo) número de funções *apenas dos dados* tais que, condicionada nestas estat´sticas, a distribuição do que resta de aleatoriedade nos dados não dependa do parâmetro desconhecido θ.

Exemplo: Gama? Regressão não-linear ???

25.2 Definição

def

Teorema de Darmois-Pitman-Koopman

25.3 Exemplos

varios

25.4 GLM em um só algoritmo

discussao

Teste, teste,teste

26. Stochastic Approximation and SGD

26.1 Introduction

Stochastic approximation is an algorithm to find the minimum of a regression function (??) involving random variables. Hence, it can be used to find the MLE in parametric families. To motivate the assumptions of this algorithm, suppose we want to estimate $\mu = \mathbb{E}(Y)$ and that we observe a stream of i.i.d. random variables $Y_1, Y_2, Y_3, \ldots$ with the same distribution as Y. The usual estimate of μ based on the first $n-1$ elements in the sequence is the arithmetic mean: $\widehat{\mu}_{n-1} = \frac{1}{n-1}\sum_{i=1}^{n-1} Y_i$. Upon the arrival of Y_n we update the estimate by calculating

$$\begin{aligned}
\widehat{\mu}_n &= \frac{Y_1 + \ldots + Y_{n-1} + Y_n}{n} = \frac{Y_1 + \ldots + Y_{n-1}}{n} + \frac{Y_n}{n} \\
&= \frac{n-1}{n}\left(\frac{Y_1 + \ldots + Y_{n-1}}{n-1}\right) + \frac{1}{n}Y_n \\
&= \frac{n-1}{n}\widehat{\mu}_{n-1} + \frac{1}{n}Y_n \\
&\approx \qquad \widehat{\mu}_{n-1} + \frac{1}{n}Y_n = \widehat{\mu}_{n-1} + a_n Y_n
\end{aligned} \tag{26.1}$$

The previous estimate $\widehat{\mu}_{n-1}$ gets a weight $(n-1)/n$ that is close to 1 when n increases. The new data point Y_n enters the updating with a weight $a_n = 1/n$ that decreases towards 0 as n increases. However, this decrease is not so fast to the point that the cumulative series of the a_n weights. Indeed, this sequence of weights satisfies the following conditions:

$$\lim_{n\to\infty} a_n = 0 \text{ (because } 1/n \to 0 \text{ as } n \to \infty) \tag{26.2}$$

$$\sum_{n=1}^{\infty} a_n = \infty \text{ (the harmonic series } A_N = \sum_{1}^{N} 1/n \text{ diverges to } \infty) \tag{26.3}$$

$$\sum_{n=1}^{\infty} a_n^2 < \infty \text{ (the Basel problem series } B_N = \sum_{1}^{N} 1/n^2 \text{ converges to } \pi^2/6) \tag{26.4}$$

I do not resist pointing out the beautiful history of the Basel problem series that was famously solved by Leonhard Euler in the 18th century (see the delightful book by William Dunham [2]).

The above conditions are important for the convergence of the Robbins-Monro method. In this method, the weights a_n are not necessarily equal to $1/n$ but must satisfy the conditions (26.2), (26.3), and (26.4). We will introduce the Robbins-Monro method as a tool to obtain the MLE. Later, we will see it on a more general view (???).

We assume that $\mathscr{D}_n = \{y_1, \ldots, y_n\}$ are observed random data independently sampled from a parametric probability model $p(y;\theta)$ with $\theta \in \Theta \subseteq \mathbb{R}^d$. We will define a sequence of estimators $\widehat{\theta}_1, \widehat{\theta}_2, \widehat{\theta}_3, \ldots$ with $\widehat{\theta}_n$ based on the first n random variables in $\mathscr{D}_n$. We want to study the convergence of this estimator sequence. For this, we specify a parameter value θ^* that is not tied to a finite sample. As usual in this book, assume that there exists $\theta^* \in \Theta$ that generates the observed data. That is, $Y_i \sim p(y;\theta^*)$.

The maximum likelihood estimator (MLE) is obtained by maximizing the log-likelihood $\ell(\theta)$ (redefined as divided by the constant n) with respect to θ:

$$\widehat{\theta}_{\text{MLE}} = \arg\max_{\theta\in\Theta} \ell(\theta) = \arg\max_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^{n} \log p(y_i;\theta)$$

This is carried out by searching for a root of the score function in the likelihood equation:

$$\frac{\partial \ell(\theta)}{\partial \theta} = \frac{\partial}{\partial \theta}\frac{1}{n}\sum_{i=1}^{n} \log p(y_i;\theta) = 0 \tag{26.5}$$

That is, $\widehat{\theta}_{\text{MLE}}$ satisfies the equation:

$$\frac{\partial}{\partial \theta}\left\{\frac{1}{n}\sum_{i=1}^{n} \log p(y_i;\theta)\right\}\bigg|_{\theta=\hat{\theta}_{\text{MLE}}} = 0$$

Now, substitute the observed values y_i of the sample by their corresponding random variables Y_i. Taking the limit as $n \to \infty$ on the left-hand side of the likelihood equation (26.5), we have the limit of the arithmetic mean of i.i.d. random variables $\partial \log p(y_i;\theta)/\partial\theta$ and the Law of Large Numbers applied to this mean gives us:

$$\lim_{n\to\infty} \frac{\partial}{\partial \theta}\left\{\frac{1}{n}\sum_{i=1}^{n} \log p(Y_i;\theta)\right\} = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \frac{\partial}{\partial \theta} \log p(Y_i;\theta) \to \mathbb{E}_{\theta^*}\left[\frac{\partial}{\partial \theta} \log p(Y;\theta)\right]$$

There are many θ symbols in this expression, as we have already discussed in Section ????. Let us remind their meaning with a simple but long example.

■ **Example 26.1** Suppose $Y_i \sim \exp(\theta^*)$ where θ^* is a fixed and unknown value $\theta^* \in (0,\infty)$and this is the value of the parameter that is generating the data Y_i. The log-likelihood is $\ell(\theta) = \log(\theta^n \exp(-\theta\sum y_i)) = n\log(\theta) - \theta\sum_i y_i$. The likelihood equation is based on the score function $\partial\ell/\partial\theta$. Using the random variables Y_i in the likelihood equation, we have the random expression composed of an arithmetic mean:

$$\frac{\partial \ell}{\partial \theta}(\theta) = \frac{1}{n}\sum_i \left(\frac{1}{\theta} - Y_i\right) \tag{26.6}$$

The θ in the denominator of the derivative function $\partial\ell/\partial\theta$ is simply indicating what variable is the derivative taken. It has no specific value, it is simply indicating the derivative is taken with respect

to this variable. The symbol (θ) is the argument of the score function and it is the value of the parameter at which the derivative is being evaluated. For example,

$$\frac{\partial \ell}{\partial \theta}(3.7) = \frac{1}{n}\sum_i \left(\frac{1}{3.7} - Y_i\right)$$

is the score function evaluated at $\theta = 3.7$. The MLE is the value of θ that makes this score function equal to zero.

Hence, in the expression (26.6), we have a function evaluated at the point (θ, Y_i). The argument θ is an arbitrary point in the parametric space $\Theta = (0, \infty)$. One possible value for the argument θ that we can use to evaluate the empirical score function is the unknown and true value θ^* of the parameter that generates the data:

$$\frac{\partial \ell}{\partial \theta}(\theta^*) = \frac{1}{n}\sum_i \left(\frac{1}{\theta^*} - Y_i\right)$$

Almost always, the empirical score function evaluated at the true value θ^* will not be zero. If this was the case, we would have the MLE exactly equal to the true value of the parameter generating the data, with a null estimation error. This is virtually impossible (???).

The MLE is found searching in Θ for the root of the *empirical* likelihood equation:

$$\overline{Y} = \hat{\theta}_{\text{MLE}} = \arg_{\theta \in \Theta} \left\{ \frac{1}{n} \sum \left(\frac{1}{\theta} - Y_i\right) = 0 \right\}$$

The score function $\partial \ell(\theta)/\partial \theta$ in (26.6) is a random variable due to the presence of the random variables Y_i. It is the arithmetic mean of the i.i.d. random variables $(1/\theta - Y_i)$. Here, θ is an arbitrary value in the parametric space Θ. We can take the expected value of this random score function. For this, we need to specify what is the distribution of the random variables Y_i present in the score function. Well, we know what is this distribution: $Y_i \sim \exp(\theta^*)$ where θ^* is an unknown and fixed value. In the case of the exponential distribution, we have $\mathbb{E}_{\theta^*}(Y_i) = 1/\theta^*$.

Therefore, the score function in (26.6) is the arithmetic mean of i.i.d. random variables $(1/\theta - Y_i)$ with expected value

$$\mathbb{E}_{\theta *}\left[\frac{1}{\theta} - Y_i\right] = \frac{1}{\theta} - \frac{1}{\theta^*}$$

By the Law of Large Numbers, the score function being an arithmetic mean of i.i.d. random variables, will converge to its common expected value:

$$\lim_{n\to\infty} \frac{\partial \ell}{\partial \theta}(\theta) = \lim_{n\to\infty} \frac{1}{n}\sum_i \left(\frac{1}{\theta} - Y_i\right) \to \mathbb{E}_{\theta^*}\left[\frac{1}{\theta} - Y\right] = \frac{1}{\theta} - \frac{1}{\theta^*}$$

The right-hand side limit expression is

$$\frac{1}{\theta} - \frac{1}{\theta^*} = \mathbb{E}_{\theta^*}\left[\frac{1}{\theta} - Y\right] = \mathbb{E}_{\theta^*}\left[\frac{\partial p(Y;\theta)}{\partial \theta}(\theta)\right] = \mathbb{E}_{\theta^*}\left[\nabla_\theta \log p(Y;\theta)\right]$$

is a function of the arbitrary θ and the unknown and fixed θ^*. We can write

$$G(\theta, \theta^*) = \mathbb{E}_{\theta^*}\left[\nabla_\theta \log p(Y;\theta)\right] = \frac{1}{\theta} - \frac{1}{\theta^*} \tag{26.7}$$

The expression in (26.7) can be seen as the theoretical counterpoint of the empirical score function. We call it the theoretical score function. Rather than taking the derivative of the log-likelihood function $\ell(\theta)$ based on the sample of n random variables, we consider the derivative of

the log-likelihood $\log(p(Y;\theta))$ of a single random variable Y evaluated at an arbitrary θ and take its expected value with respect to the true parameter value θ^*. We are exchanging the mean of the n random variables

$$\frac{\partial \log p(Y_1,\theta)}{\partial \theta}(\theta), \frac{\partial \log p(Y_2,\theta)}{\partial \theta}(\theta), \ldots, \frac{\partial \log p(Y_n,\theta)}{\partial \theta}(\theta)$$

by the expected value $\mathbb{E}_{\theta^*}(\cdot)$ of a single random variable

$$\frac{\partial \log p(Y,\theta)}{\partial \theta}(\theta)$$

where $Y \sim p(y,\theta^*)$.

Considering the theoretical score function $G(\theta,\theta^*)$ in (26.7) we can ask for the value of $\theta \in \Theta$ that will make $G(\theta,\theta^*) = 0$. In this particular case, there is only one solution, and it is to take $\theta = \theta^*$. That is, the value θ that solves the theoretical score equation:

$$\mathbb{E}_{\theta^*}\left[\nabla_\theta \log p(X;\theta)\right] = 0,$$

is the true value θ^*

■

Generalizing from this example, We assume that $\mathscr{D}_n = \{y_1,\ldots,y_n\}$ are observed random data independently sampled from a parametric probability model $p(y;\theta)$ with $\theta \in \Theta \subseteq \mathbb{R}^d$. Suppose that the theoretical score equation

$$G(\theta,\theta^*) = \mathbb{E}_{\theta^*}\left[\nabla_\theta \log p(X;\theta)\right] = 0,$$

has only one solution and this solution is to take θ equal to the fixed (and unknown) θ^*. That is, the unique solution is $\theta = \theta^*$.

In this case, we aim for a sequence of estimators $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_n, \ldots$ such that $\hat{\theta}_n$ converges to the single root of the equation

$$\mathbb{E}_{\theta^*}\left[\nabla_\theta \log p(X;\theta)\right] = 0,$$

That is, we aim for a sequence of estimators such that

$$\hat{\theta}_n \to \theta^*$$

as $n \to \infty$.

The Robbins-Monro algorithm provides such a sequence of estimators. More interesting yet, the sequence is extremely simple with a very simple updating of $\hat{\theta}_n$ as a new data point y_{n+1} enters the dataset.

26.2 Target of the Robbins-Monro Algorithm in MLE: Role of θ and θ^*

Let $X \sim p(x;\theta^*)$, where $\theta^* \in \Theta \subseteq \mathbb{R}^d$ is the unknown true parameter value that governs the distribution of the data.

The Robbins-Monro algorithm is applied to find a value θ that solves the following equation:

$$\mathbb{E}_{\theta^*}\left[\nabla_\theta \log p(X;\theta)\right] = 0,$$

where:

- The expectation $\mathbb{E}_{\theta^*}$ is taken over the distribution of $X \sim p(x;\theta^*)$,
- The argument $\theta \in \Theta$ is the variable we are solving for.

This defines a function:

$$m(\theta) := \mathbb{E}_{\theta^*}\left[\nabla_\theta \log p(X;\theta)\right],$$

which depends on the fixed but unknown θ^*, and varies as θ varies over the parameter space.

Interpretation

The Robbins-Monro method treats θ^* as fixed (but unknown), and attempts to find a value of θ that zeroes the population score function $m(\theta)$. That is, we aim to solve:

$$m(\theta) = 0.$$

Solution of the Equation

Under standard regularity conditions (e.g., smoothness, dominated convergence), and assuming the model is *identifiable*, this equation has a unique solution at $\theta = \theta^*$. That is:

$$m(\theta) = 0 \quad \Longleftrightarrow \quad \theta = \theta^*.$$

Conclusion

The Robbins-Monro algorithm estimates the parameter value θ by iteratively solving:

$$\mathbb{E}_{\theta^*}\left[\nabla_\theta \log p(X;\theta)\right] = 0.$$

Although the expectation is taken with respect to the (unknown) true distribution $p(x;\theta^*)$, the argument θ is arbitrary. The goal is to find the value θ that zeroes this function. Under identifiability, this unique solution is the true parameter value:

$$\boxed{\theta = \theta^*}.$$

The root of the regression function $f(\theta)$ is defined by:

$$f(\theta) = E\{z|\theta\} \tag{26.8}$$

In the Robbins-Monro method, the new successive estimate θ_{N+1} based on the present estimate θ_N and a new observation Z_N is given by:

$$\theta_{N+1} = \theta_N - a_N z_N \tag{26.9}$$

Also, a_N is assumed to be a sequence of positive numbers which satisfy the following conditions:

With a sequence of a_N satisfying (8.31) through (8.33), θ_N of (8.30) converges toward θ_0 in the mean-square sense and with probability 1, that is:

$$\lim_{N\to\infty} E\{(\theta_N - \theta_0)^2\} = 0 \tag{26.10}$$

$$\lim_{N\to\infty} Pr\{\theta_N = \theta_0\} = 1 \tag{26.11}$$

More generally, a sequence of the form:

$$a_N = \frac{1}{N^k} \quad 1 \geq k > \frac{1}{2} \tag{26.12}$$

satisfies (8.31) through (8.33).

26.3 NOvo - NOVO - NOVO - NOVO

26.4 Stochastic Approximation and the Robbins-Monro Algorithm

The Robbins-Monro algorithm is a classical method for stochastic root-finding. Suppose we want to find the root θ^* of an unknown function $m(\theta)$, but we can only access noisy observations $Z_n(\theta)$ satisfying

$$\mathbb{E}[Z_n(\theta)] = m(\theta).$$

The Robbins-Monro algorithm recursively updates an estimate of the root as follows:

$$\theta_{n+1} = \theta_n - a_n Z_n(\theta_n),$$

where (a_n) is a sequence of positive step sizes satisfying $\sum_n a_n = \infty$ and $\sum_n a_n^2 < \infty$ (e.g., $a_n = a/n$ or $a/\sqrt{n}$). The algorithm converges under mild regularity conditions.

26.4.1 MLE as a Root-Finding Problem

In maximum likelihood estimation, the MLE $\hat{\theta}$ is the root of the expected score function:

$$\mathbb{E}_\theta \left[\frac{\partial}{\partial \theta} \log f(X;\theta) \right] = 0.$$

If we can simulate or observe samples X_n from $f(X;\theta)$, then the Robbins-Monro method provides a stochastic algorithm to find $\hat{\theta}$.

26.4.2 Example 1: Normal Distribution with Known Variance

Let $X \sim \mathcal{N}(\mu, \sigma^2)$ with known variance $\sigma^2 = 1$. The score function for one observation is:

$$\frac{\partial}{\partial \mu} \log f(X;\mu) = X - \mu.$$

The Robbins-Monro update becomes:

$$\mu_{n+1} = \mu_n + a_n(X_n - \mu_n).$$

R code:

```
set.seed(123)
mu_true <- 5
n_iter <- 1000
a <- 0.8
mu_est <- numeric(n_iter)
mu_est[1] <- 0
for (n in 2:n_iter) {
  X_n <- rnorm(1, mean = mu_true, sd = 1)
  a_n <- a / n
  mu_est[n] <- mu_est[n-1] + a_n * (X_n - mu_est[n-1])
}
```

26.4.3 Example 2: Bernoulli Distribution (Unknown θ)

Let $X \sim \text{Bernoulli}(\theta)$ with $\theta \in (0,1)$. The log-likelihood score for one observation is:

$$\frac{\partial}{\partial \theta} \log f(X;\theta) = \frac{X}{\theta} - \frac{1-X}{1-\theta}.$$

The Robbins-Monro update is:

$$\theta_{n+1} = \theta_n + a_n \left(\frac{X_n}{\theta_n} - \frac{1-X_n}{1-\theta_n} \right).$$

R code:

```
set.seed(123)
theta_true <- 0.7
n_iter <- 1000
a <- 0.05
theta_est <- numeric(n_iter)
theta_est[1] <- 0.5
for (n in 2:n_iter) {
  X_n <- rbinom(1, 1, theta_true)
  a_n <- a / sqrt(n)
  score <- X_n / theta_est[n-1] - (1 - X_n) / (1 - theta_est[n-1])
  theta_est[n] <- theta_est[n-1] + a_n * score
  theta_est[n] <- min(max(theta_est[n], 0.01), 0.99)
}
```

26.4.4 Mini-Batch Robbins-Monro on Fixed Dataset

Suppose we have a fixed dataset $\{X_1, \ldots, X_n\}$. We can simulate stochastic updates by randomly sampling mini-batches at each step. The update becomes:

$$\theta_{t+1} = \theta_t + a_t \cdot \frac{1}{b} \sum_{i \in \mathscr{B}_t} \left(\frac{X_i}{\theta_t} - \frac{1-X_i}{1-\theta_t} \right),$$

where $\mathscr{B}_t$ is a mini-batch of size b drawn randomly from the dataset.

R code:

```
set.seed(123)
theta_true <- 0.7
n_data <- 500
data <- rbinom(n_data, 1, theta_true)
n_iter <- 200
batch_size <- 10
a <- 0.3
theta_est <- numeric(n_iter)
theta_est[1] <- 0.5
for (n in 2:n_iter) {
  idx <- sample(1:n_data, batch_size, replace = TRUE)
  X_batch <- data[idx]
  score <- mean(X_batch / theta_est[n-1] - (1 - X_batch) / (1 - theta_est[n-1]))
  a_n <- a / sqrt(n)
  theta_est[n] <- theta_est[n-1] + a_n * score
  theta_est[n] <- min(max(theta_est[n], 0.01), 0.99)
}
```

This method provides a stochastic approximation to the MLE using only gradients and avoids computing the full likelihood or storing all data in memory.

26.5 Robbins-Monro Estimation and the Role of the Population Score Equation

1. Robbins-Monro Targets the Population Score Equation

The Robbins-Monro stochastic approximation algorithm is used to find the solution θ^* to the equation:

$$\mathbb{E}_{\theta^*}[\nabla_\theta \log p(X;\theta)] = 0,$$

where $X \sim p(x;\theta^*)$ and θ^* is the true (unknown) parameter. This is the *population score equation*.

By contrast, the classical MLE $\hat{\theta}_n$ is the solution to the *empirical score equation*:

$$\frac{1}{n}\sum_{i=1}^{n} \nabla_\theta \log p(X_i;\hat{\theta}_n) = 0.$$

As $n \to \infty$, the empirical score converges to the population score under regularity conditions, and $\hat{\theta}_n \to \theta^*$, which motivates the Robbins-Monro formulation as a direct approach to estimating θ^*.

2. Is the Solution to the Population Score Equation the True Parameter?

Yes. Under model correctness and an *identifiability condition*, the equation

$$\mathbb{E}_{\theta^*}[\nabla_\theta \log p(X;\theta)] = 0$$

has a *unique solution at* $\theta = \theta^*$. This implies that Robbins-Monro converges to the true parameter value when applied to the population score.

3. Reconciling Two Views of MLE Consistency

There are two equivalent views of MLE consistency:

- **Probabilistic view:** The MLE $\hat{\theta}_n \xrightarrow{P} \theta^*$ as $n \to \infty$.
- **Functional view:** The solution to the empirical score equation converges to the solution of the population score equation, which is θ^*.

These two views are compatible: the functional convergence of the empirical score implies the probabilistic convergence of the MLE.

4. Identifiability and Its Role

The identifiability condition ensures that the population score equation has a unique solution:

$$\mathbb{E}_{\theta^*}[\nabla_\theta \log p(X;\theta)] = 0 \quad \Rightarrow \quad \theta = \theta^*.$$

This is typically assumed in theoretical treatments and can be verified in many parametric models, including GLMs. Identifiability ensures that the MLE converges to the true parameter value and not to a spurious local solution.

5. Uniqueness of the MLE in Logistic Regression

In logistic regression, the model is:

$$\mathbb{P}(y = 1 \mid x) = \sigma(x^\top \beta), \quad \text{where } \sigma(z) = \frac{1}{1+e^{-z}}.$$

The score function is:

$$\nabla \ell(\beta) = \sum_{i=1}^{n} x_i(y_i - \sigma(x_i^\top \beta)) = X^\top (y - \mu(\beta)).$$

Key facts:
- The log-likelihood $\ell(\beta)$ is *strictly concave*.
- The Hessian is $\nabla^2 \ell(\beta) = -X^\top W X$, where W is a positive diagonal matrix.
- If the design matrix X has full column rank, then $X^\top W X \succ 0$, ensuring strict concavity.

Conclusion: The MLE exists and is unique if:
- X has full column rank, and
- The data are not completely separable.

Under these conditions, the score equation $\nabla \ell(\beta) = 0$ has a unique finite solution.

6. Summary Table

Aspect	Robbins-Monro	Classical MLE
Solves	$\mathbb{E}_{\theta^*}[\nabla \log p(X;\theta)] = 0$	$\frac{1}{n}\sum \nabla \log p(X_i;\theta) = 0$
Target	Population score root	Finite-sample root
Converges to	θ^* (true parameter)	θ^* (true parameter)
Identifiability required	Yes (unique zero of expectation)	Yes (unique MLE solution)
Score uniqueness in logistic model	Yes, if X full rank and no separation	Same

26.6 Stochastic Approximation - Fukunaga Text

The equation (8.30) is rewritten as:

$$\theta_{N+1} = \theta_N - a_N f(\theta_N) - a_N \gamma_N \tag{26.13}$$

$$\text{where} \quad \gamma_N = z_N - f(\theta_N) \tag{26.14}$$

From the definition of the regression function $f(\theta)$ in (8.29), γ_N is a random variable with zero mean as:

$$E\{\gamma_N | \theta_N\} = E\{z_N | \theta_N\} - f(\theta_N) = 0 \tag{26.15}$$

Also, it is reasonable to assume that the variance of γ_N is bounded, that is:

$$E\{\gamma_N^2\} \leq \sigma^2 \tag{26.16}$$

Next, let us study the difference between θ_0 and θ_N. From (8.37), we have:

$$(\theta_{N+1} - \theta_0) = (\theta_N - \theta_0) - a_N f(\theta_N) - a_N \gamma_N \tag{26.17}$$

Taking the expectation of the square of (8.41):

$$E\{(\theta_{N+1} - \theta_0)^2\} - E\{(\theta_N - \theta_0)^2\} = a_N^2 E\{f^2(\theta_N)\} + a_N^2 E\{\gamma_N^2\} - 2a_N E\{(\theta_N - \theta_0) f(\theta_N)\} \tag{26.18}$$

Therefore, repeating (8.42), we obtain:

$$E\{(\theta_N - \theta_0)^2\} - E\{(\theta_1 - \theta_0)^2\} = \sum_{i=1}^{N-1} a_i^2 [E\{f^2(\theta_i)\} + E\{\gamma_i^2\}] - 2\sum_{i=1}^{N-1} a_i E\{(\theta_i - \theta_0) f(\theta_i)\} \tag{26.19}$$

We assume that the regression function is also bounded in the region of interest as:

$$E\{f^2(\theta_N)\} \leq b \tag{26.20}$$

Then, (8.43) is bounded by:

$$E\{(\theta_N - \theta_0)^2\} - E\{(\theta_1 - \theta_0)^2\} \leq (b + \sigma^2) \sum_{i=1}^{N-1} a_i^2 - 2 \sum_{i=1}^{N-1} a_i E\{(\theta_i - \theta_0) f(\theta_i)\} \tag{26.21}$$

Recall from Fig. 8-3 that the regression function satisfies:

$$\begin{aligned} &f(\theta) > 0 \quad \text{if} \quad (\theta - \theta_0) > 0 \\ &f(\theta) = 0 \quad \text{if} \quad (\theta - \theta_0) = 0 \\ &f(\theta) < 0 \quad \text{if} \quad (\theta - \theta_0) < 0 \end{aligned} \tag{26.22}$$

Therefore,

$$(\theta - \theta_0) f(\theta) \geq 0 \tag{26.23}$$

and

$$E\{(\theta - \theta_0) f(\theta)\} \geq 0 \tag{26.24}$$

Now consider the following proposition:

$$\lim_{i \to \infty} E\{(\theta_i - \theta_0) f(\theta_i)\} = 0 \tag{26.25}$$

Since (8.47) holds for all θ's, (8.49) is equivalent to:

$$\lim_{i \to \infty} Pr\{\theta_i = \theta_0\} = 1 \tag{26.26}$$

The Kiefer-Wolfowitz method modifies θ_N as:

$$\theta_{N+1} = \theta_N - a_N \frac{z(\theta_N + c_N) - z(\theta_N - c_N)}{2c_N} \tag{26.27}$$

Both a_N and c_N are sequences of positive numbers. They must vanish in the limit, that is:

$$\lim_{N \to \infty} a_N = 0 \tag{26.28}$$

$$\lim_{N \to \infty} c_N = 0 \tag{26.29}$$

In order to make sure that we have enough corrective action to avoid stopping short of the minimum point, a_N should satisfy:

$$\sum_{N=1}^{\infty} a_N = \infty \tag{26.30}$$

Also, to cancel the accumulated noise effect we must have:

$$\sum_{N=1}^{\infty} \left[\frac{a_N}{c_N}\right]^2 < \infty \tag{26.31}$$

26.7 Introduction to Stochastic Gradient Descent for MLE

Given observed data $\mathscr{D} = \{x_1, \ldots, x_n\}$ from a parametric model $p(x|\theta)$ with $\theta \in \Theta \subseteq \mathbb{R}^d$, the maximum likelihood estimator (MLE) is obtained by maximizing the log-likelihood:

$$\ell(\theta) = \sum_{i=1}^{n} \log p(x_i|\theta) \tag{26.32}$$

Stochastic Gradient Descent (SGD) Approach

Traditional gradient descent updates parameters via:

$$\theta_{t+1} = \theta_t + \gamma_t \nabla_\theta \ell(\theta_t) \tag{26.33}$$

SGD replaces the full gradient with an unbiased estimate using a random subset $\mathscr{B}_t$ (mini-batch) of size B:

$$\theta_{t+1} = \theta_t + \gamma_t \frac{n}{B} \sum_{x \in \mathscr{B}_t} \nabla_\theta \log p(x|\theta_t) \tag{26.34}$$

Key advantages:

- Computational efficiency (avoids full data scan)
- Natural online learning implementation
- Provable convergence under conditions

26.8 MLE Estimation of the Rate Parameter λ of an Exponential Distribution via Robbins-Monro

Let $X \sim \text{Exponential}(\lambda)$, with density:

$$f(x;\lambda) = \lambda e^{-\lambda x}, \quad x > 0,\ \lambda > 0.$$

We aim to estimate λ by maximizing the log-likelihood using the Robbins-Monro stochastic approximation method.

Maximum Likelihood Estimation

Given a single observation X, the log-likelihood function is:

$$\ell(\lambda;X) = \log \lambda - \lambda X.$$

The score function is:

$$S(\lambda) = \frac{d}{d\lambda} \log f(X;\lambda) = \frac{1}{\lambda} - X.$$

The MLE is the value of λ that satisfies:

$$\mathbb{E}[S(\lambda)] = \mathbb{E}\left[\frac{1}{\lambda} - X\right] = 0.$$

Solving gives:

$$\lambda_{\text{MLE}} = \frac{1}{\mathbb{E}[X]},$$

which, for a fixed dataset, corresponds to the inverse of the sample mean.

Robbins-Monro Formulation

We cast the MLE equation as a root-finding problem:

$$\text{Find } \lambda \text{ such that } \mathbb{E}\left[\frac{1}{\lambda} - X\right] = 0.$$

This fits the Robbins-Monro setup where we define:

$$Z_n(\lambda) = \frac{1}{\lambda} - X_n, \quad \text{with } \mathbb{E}[Z_n(\lambda)] = 0 \text{ when } \lambda = \lambda^*.$$

Thus, the Robbins-Monro update becomes:

$$\lambda_{n+1} = \lambda_n - a_n \left(\frac{1}{\lambda_n} - X_n\right),$$

where:

- $X_n \sim \text{Exponential}(\lambda^*)$ is the n-th observation,
- $a_n > 0$ is a step size (e.g., $a_n = \frac{a}{n}$).

Algorithm Implementation in R

```
set.seed(123)
lambda_true <- 2
n_iter <- 1000
a <- 0.5

lambda_est <- numeric(n_iter)
lambda_est[1] <- 1  # Initial guess

for (n in 2:n_iter) {
  X_n <- rexp(1, rate = lambda_true)
  a_n <- a / n
  lambda_est[n] <- lambda_est[n-1] - a_n * (1 / lambda_est[n-1] - X_n)
}
```

Convergence

Under regular conditions on the step size:

$$\sum_{n=1}^{\infty} a_n = \infty, \quad \sum_{n=1}^{\infty} a_n^2 < \infty,$$

and assuming λ_n remains bounded away from 0 (to avoid instability from $1/\lambda_n$), the Robbins-Monro estimator λ_n converges almost surely to the MLE:

$$\lambda_n \xrightarrow{a.s.} \lambda_{\text{MLE}} = \frac{1}{\mathbb{E}[X]} = \lambda^*.$$

This method provides a simple sequential procedure for MLE estimation that avoids storing or recomputing over all past data, and is particularly useful in online or streaming settings.

26.9 Robbins-Monro Convergence Theorem

The theoretical foundation for SGD was established by Herbert Robbins and Sutton Monro (1951). For MLE estimation:

Theorem (Robbins & Monro, 1951):
Under regularity conditions:

1. The log-likelihood $\ell(\theta)$ is concave
2. Learning rates satisfy:

$$\sum_{t=1}^{\infty} \gamma_t = \infty \quad \text{and} \quad \sum_{t=1}^{\infty} \gamma_t^2 < \infty \tag{26.35}$$

3. The expected gradient is Lipschitz continuous

Then the SGD iterates converge almost surely to the MLE:

$$\theta_t \xrightarrow{a.s.} \theta_{MLE} \quad \text{as} \quad t \to \infty \tag{26.36}$$

Proof

Consider maximizing the log-likelihood $\ell(\theta)$ via SGD updates:

$$\theta_{t+1} = \theta_t + \gamma_t g_t(\theta_t) \tag{26.37}$$

where $g_t(\theta_t) = \frac{1}{|\mathcal{B}_t|} \sum_{x \in \mathcal{B}_t} \nabla_\theta \log p(x|\theta_t)$ is an unbiased stochastic gradient estimate.

Assumptions

1. **Strong convexity**: For some $\mu > 0$,

$$\langle \nabla \ell(\theta) - \nabla \ell(\theta^*), \theta - \theta^* \rangle \leq -\mu \|\theta - \theta^*\|^2 \tag{26.38}$$

2. **Bounded variance**:

$$\mathbb{E}[\|g_t(\theta) - \nabla \ell(\theta)\|^2] \leq \sigma^2 \quad \forall \theta \tag{26.39}$$

3. **Learning rates**:

$$\sum_{t=1}^{\infty} \gamma_t = \infty, \quad \sum_{t=1}^{\infty} \gamma_t^2 < \infty \tag{26.40}$$

Proof Sketch

Define the distance to optimum $D_t = \|\theta_t - \theta^*\|^2$.

$$\begin{aligned} D_{t+1} &= \|\theta_t + \gamma_t g_t(\theta_t) - \theta^*\|^2 \\ &= D_t + 2\gamma_t \langle g_t(\theta_t), \theta_t - \theta^* \rangle + \gamma_t^2 \|g_t(\theta_t)\|^2 \end{aligned}$$

Taking conditional expectations:

$$\begin{aligned} \mathbb{E}[D_{t+1}|\mathcal{F}_t] \leq D_t &+ 2\gamma_t \langle \nabla \ell(\theta_t), \theta_t - \theta^* \rangle \\ &+ \gamma_t^2 \mathbb{E}[\|g_t(\theta_t)\|^2 | \mathcal{F}_t] \end{aligned}$$

Using assumptions:

$$\begin{aligned}\mathbb{E}[D_{t+1}|\mathscr{F}_t] &\leq D_t - 2\gamma_t\mu D_t + \gamma_t^2(L^2 D_t + \sigma^2) \\ &= (1 - 2\gamma_t\mu + \gamma_t^2 L^2)D_t + \gamma_t^2\sigma^2\end{aligned}$$

Taking total expectations and telescoping:

$$\mathbb{E}[D_{t+1}] \leq \prod_{k=1}^{t}(1 - 2\gamma_k\mu + \gamma_k^2 L^2)D_1 + \sigma^2 \sum_{k=1}^{t} \gamma_k^2 \tag{26.41}$$

The product $\to 0$ and sum $\to$ finite by the learning rate conditions, giving:

$$\lim_{t\to\infty} \mathbb{E}[D_t] = 0 \tag{26.42}$$

Under the stated assumptions, SGD converges to the MLE solution θ^* in mean square (and almost surely with additional technical arguments). The key requirements are:

- Properly decreasing learning rates
- Gradient estimator unbiasedness
- Strong convexity (can be relaxed to PL condition)

Key Implications

- Learning rates must decrease but not too quickly (e.g., $\gamma_t = O(1/t)$)
- The theorem guarantees eventual convergence to the true optimum
- Explains why SGD works for large-scale statistical estimation

26.10 Practical Considerations

Modern implementations combine the Robbins-Monro theory with:

- Adaptive learning rates (Adam, RMSprop)
- Variance reduction techniques
- Mini-batch optimization

This theoretical foundation remains crucial for understanding stochastic optimization in machine learning and statistics.

26.10.1 Convergence of the Robbins-Monro Algorithm for Multivariate MLE Estimation

The Robbins-Monro algorithm is a sequential stochastic approximation method used to find the root of an expectation function when only noisy observations are available. In the context of maximum likelihood estimation (MLE), we aim to solve:

$$\mathbb{E}_{\theta^*}[S(\theta)] = \mathbb{E}_{\theta^*}[\nabla_\theta \log f(X;\theta)] = 0,$$

where $\theta \in \mathbb{R}^d$ and θ^* is the true parameter value.

Robbins-Monro Update Rule.

Given an initial guess $\theta_1 \in \mathbb{R}^d$, and a sequence $\{X_n\}_{n\geq 1}$ of i.i.d. observations from the distribution $f(x;\theta^*)$, we define the update:

$$\theta_{n+1} = \theta_n + a_n S_n(\theta_n),$$

where:

- $a_n > 0$ is a sequence of step sizes,
- $S_n(\theta_n)$ is an unbiased estimate of the score, i.e.,

$$\mathbb{E}[S_n(\theta_n) \mid \theta_n] = \mathbb{E}_{\theta^*}[\nabla_\theta \log f(X;\theta_n)].$$

Goal.
We want to show that $\theta_n \to \theta^*$ almost surely under appropriate regularity conditions.

Informal Convergence Sketch (ODE Method)
The Robbins-Monro algorithm can be interpreted as a noisy discretization of the ordinary differential equation (ODE):

$$\frac{d\theta(t)}{dt} = \mathbb{E}[S(\theta(t))],$$

which describes the average dynamics of the algorithm. Under appropriate conditions:

- The ODE has a unique globally attracting equilibrium point at θ^*,
- The stochastic iterates θ_n track the ODE trajectory with decreasing noise,
- The step size sequence a_n controls the tradeoff between learning speed and stability.

This suggests that $\theta_n \to \theta^*$ as $n \to \infty$, in a manner analogous to stochastic gradient descent for convex optimization.

Theorem (Almost Sure Convergence of Robbins-Monro Algorithm)
Let $\{\theta_n\}$ be defined by the recursion:

$$\theta_{n+1} = \theta_n + a_n S_n(\theta_n),$$

and suppose the following assumptions hold:

(A1) (**Step size conditions**) The sequence $\{a_n\}$ satisfies

$$a_n > 0, \quad \sum_{n=1}^{\infty} a_n = \infty, \quad \sum_{n=1}^{\infty} a_n^2 < \infty.$$

(A2) (**Unbiased noisy observations**) For all n,

$$\mathbb{E}[S_n(\theta_n) \mid \mathscr{F}_n] = m(\theta_n), \quad \text{with } m(\theta) = \mathbb{E}_{\theta^*}[\nabla_\theta \log f(X;\theta)],$$

where $\mathscr{F}_n = \sigma(\theta_1, S_1, \ldots, S_{n-1})$.

(A3) (**Mean function**) The function $m(\theta)$ is Lipschitz continuous, and has a unique root at $\theta = \theta^*$.

(A4) (**Noise control**) There exists a constant $C > 0$ such that

$$\mathbb{E}\left[\|S_n(\theta_n) - m(\theta_n)\|^2 \mid \mathscr{F}_n\right] \le C(1 + \|\theta_n\|^2).$$

Then the Robbins-Monro iterates θ_n converge to θ^* almost surely:

$$\theta_n \xrightarrow{a.s.} \theta^*.$$

Sketch of the Proof.
Define the squared error:

$$V_n = \|\theta_n - \theta^*\|^2.$$

Using the update rule:

$$\theta_{n+1} = \theta_n + a_n S_n(\theta_n),$$

we have:

$$\begin{aligned} V_{n+1} &= \|\theta_n + a_n S_n(\theta_n) - \theta^*\|^2 \\ &= \|\theta_n - \theta^*\|^2 + 2a_n\langle \theta_n - \theta^*, S_n(\theta_n)\rangle + a_n^2 \|S_n(\theta_n)\|^2 \\ &= V_n + 2a_n\langle \theta_n - \theta^*, m(\theta_n)\rangle + 2a_n\langle \theta_n - \theta^*, \varepsilon_n\rangle + a_n^2 \|S_n(\theta_n)\|^2, \end{aligned}$$

where $\varepsilon_n = S_n(\theta_n) - m(\theta_n)$ is the noise.

Take conditional expectations and apply (A2)–(A4):

$$\mathbb{E}[V_{n+1} \mid \mathscr{F}_n] \le V_n + 2a_n\langle \theta_n - \theta^*, m(\theta_n)\rangle + a_n^2 C(1 + \|\theta_n\|^2).$$

Since $m(\theta)$ has a unique root at θ^* and is Lipschitz, one can show:

$$\langle \theta_n - \theta^*, m(\theta_n)\rangle \le -\lambda \|\theta_n - \theta^*\|^2,$$

for some $\lambda > 0$, at least when θ_n is near θ^*. Thus:

$$\mathbb{E}[V_{n+1} \mid \mathscr{F}_n] \le V_n - 2a_n\lambda V_n + a_n^2 C(1 + V_n).$$

This is a stochastic version of a decreasing sequence with a small error term. A Robbins-Siegmund-type theorem then gives:

$$V_n \xrightarrow{a.s.} 0 \quad \Rightarrow \quad \theta_n \xrightarrow{a.s.} \theta^*.$$

■

26.11 Robbins-Monro Method for Linear Regression

Consider the classical linear regression model:

$$y_n = x_n^\top \beta + \varepsilon_n, \quad n = 1, 2, \ldots,$$

where $x_n \in \mathbb{R}^p$ is a vector of predictors (including a 1 if an intercept is desired), $\beta \in \mathbb{R}^p$ is the unknown coefficient vector, and ε_n is a noise term with $\mathbb{E}[\varepsilon_n \mid x_n] = 0$.

The goal is to estimate β by minimizing the expected squared loss:

$$L(\beta) = \mathbb{E}[(y - x^\top \beta)^2].$$

Gradient-Based Formulation

The gradient of the loss function is:

$$\nabla_\beta L(\beta) = -2\mathbb{E}[x(y - x^\top \beta)].$$

Setting this gradient to zero gives the normal equation:

$$\mathbb{E}[xx^\top]\beta = \mathbb{E}[xy].$$

Robbins-Monro Iterative Algorithm

We frame this as a root-finding problem for:

$$m(\beta) = \mathbb{E}[x(y - x^\top \beta)] = 0.$$

Since we do not know the full distribution of (x, y), we can only access a sequence of noisy evaluations:

$$Z_n(\beta) = x_n(y_n - x_n^\top \beta).$$

The Robbins-Monro update is:

$$\beta_{n+1} = \beta_n + a_n Z_n(\beta_n) = \beta_n + a_n x_n(y_n - x_n^\top \beta_n),$$

where $a_n > 0$ is a step size satisfying:

$$\sum_{n=1}^{\infty} a_n = \infty, \quad \sum_{n=1}^{\infty} a_n^2 < \infty \quad \text{(e.g., } a_n = \frac{a}{n}\text{)}.$$

This recursion is equivalent to a stochastic gradient descent algorithm applied to the loss:

$$\ell_n(\beta) = (y_n - x_n^\top \beta)^2.$$

Convergence

Under standard assumptions:

- The sequence (x_n, y_n) is i.i.d. with finite second moments,
- $\mathbb{E}[xx^\top]$ is positive definite,
- The step sizes satisfy $\sum a_n = \infty$ and $\sum a_n^2 < \infty$,

the iterates β_n converge almost surely to the minimizer:

$$\beta^* = \left(\mathbb{E}[xx^\top]\right)^{-1} \mathbb{E}[xy],$$

which is the population least squares solution.

Advantages

The Robbins-Monro method:

- Does not require storing or inverting large matrices,
- Can be used in online or streaming contexts,
- Is scalable to high-dimensional problems with sparse features.

26.12 Robbins-Monro Method for Logistic Regression

Consider the binary classification setting where $y_n \in \{0,1\}$, and each observation $(x_n, y_n) \in \mathbb{R}^p \times \{0,1\}$ follows the logistic model:

$$\mathbb{P}(y_n = 1 \mid x_n) = \sigma(x_n^\top \beta), \quad \text{where } \sigma(z) = \frac{1}{1+e^{-z}}.$$

The goal is to estimate $\beta \in \mathbb{R}^p$ from observed data.

Maximum Likelihood Formulation

The log-likelihood for a single observation is:

$$\ell(\beta; x_n, y_n) = y_n \log \sigma(x_n^\top \beta) + (1-y_n)\log(1-\sigma(x_n^\top \beta)).$$

The gradient (score function) is:

$$\nabla_\beta \ell(\beta; x_n, y_n) = x_n(y_n - \sigma(x_n^\top \beta)).$$

The population MLE satisfies:

$$\mathbb{E}[x(y - \sigma(x^\top \beta))] = 0.$$

Robbins-Monro Iterative Algorithm

We define the Robbins-Monro update based on a stochastic approximation of the score:

$$Z_n(\beta) = x_n(y_n - \sigma(x_n^\top \beta)),$$

and iterate:

$$\beta_{n+1} = \beta_n + a_n x_n (y_n - \sigma(x_n^\top \beta_n)),$$

where $a_n > 0$ is a step size sequence such that:

$$\sum_{n=1}^{\infty} a_n = \infty, \quad \sum_{n=1}^{\infty} a_n^2 < \infty.$$

Interpretation

This is a stochastic gradient ascent algorithm on the log-likelihood function. At each step, we update β_n in the direction of the noisy gradient of the log-likelihood based on a single sample (x_n, y_n).

Convergence

Under standard assumptions:

- The data (x_n, y_n) are i.i.d. with bounded $\|x_n\|$,
- The expected Fisher information $\mathbb{E}[\sigma(x^\top\beta)(1-\sigma(x^\top\beta))xx^\top]$ is positive definite,
- The step size satisfies the usual Robbins-Monro conditions,

the iterates β_n converge almost surely to the maximizer of the expected log-likelihood function:

$$\beta^* = \arg\max_\beta \mathbb{E}[\ell(\beta; x, y)].$$

Advantages

The Robbins-Monro method for logistic regression:

- Allows online learning from streaming data,
- Avoids computing or inverting the full Hessian matrix,
- Is suitable for large-scale or sparse datasets.

26.13 Two Conceptual Questions on the Robbins-Monro Method

Question 1: Does Robbins-Monro Solve the Population or Empirical Score Equation?

In classical maximum likelihood estimation (MLE), the estimator $\hat{\theta}_n$ solves the *empirical score equation*:

$$\frac{1}{n}\sum_{i=1}^{n} \nabla_\theta \log f(X_i; \theta) = 0.$$

This solution depends on the specific dataset and converges to the population MLE as $n \to \infty$ under regularity conditions.

By contrast, the Robbins-Monro method is designed to find the root of the *expected score function*:

$$\mathbb{E}_{\theta^*}[\nabla_\theta \log f(X; \theta)] = 0,$$

using only noisy observations of the score function one at a time. That is, Robbins-Monro solves:

$$m(\theta) = 0, \quad \text{where } m(\theta) = \mathbb{E}[Z_n(\theta)], \text{ with } Z_n(\theta) \approx \nabla_\theta \log f(X_n; \theta).$$

Key Distinction:

- The Robbins-Monro algorithm approximates the **population-level MLE** as data accumulates.
- Classical MLE finds the **finite-sample estimate** that exactly satisfies the empirical score equation.

Why Use Robbins-Monro?

Robbins-Monro is particularly useful when:

- The score function cannot be computed exactly, but can be simulated or approximated,
- Data arrives sequentially (online setting),
- The full dataset is too large to store or process at once.

Question 2: Why Does Robbins-Monro Require a Decreasing Learning Rate, Unlike Deep Learning?

In the Robbins-Monro algorithm, the step size a_n must satisfy:

$$\sum_{n=1}^{\infty} a_n = \infty, \quad \sum_{n=1}^{\infty} a_n^2 < \infty,$$

such as $a_n = a/n$ or $a_n = a/n^\gamma$ with $\gamma \in (0.5, 1]$. These conditions ensure:

- **Persistent exploration**: updates do not vanish too quickly,
- **Diminishing noise effect**: update variance decreases over time,
- **Almost sure convergence** to the solution of the population equation.

In contrast, **deep learning** methods like stochastic gradient descent (SGD) often use a *fixed or slowly-decaying* learning rate. This violates the Robbins-Monro conditions but is done intentionally:

- Fixed learning rates allow faster initial exploration,
- Deep networks are highly non-convex — exact convergence is often unnecessary,
- Training benefits from persistent stochasticity, e.g., to escape local minima.

Summary of the Difference:

Aspect	Robbins-Monro	Deep Learning Practice
Step size	Decreasing ($a_n \to 0$)	Fixed or scheduled
Convergence	Almost sure (under assumptions)	Not guaranteed
Target	Population root of $\mathbb{E}[g(\theta)]$	Good local optimum
Noise in gradients	Must vanish	Often beneficial
Objective	Precise root-finding	Empirical generalization

Conclusion:

Robbins-Monro belongs to the realm of statistical root-finding with provable convergence under strong assumptions. Deep learning methods prioritize empirical performance in large-scale, non-convex settings and deliberately relax the strict requirements of Robbins-Monro to gain practical advantages.

27. Refined Error Decomposition

28. Model Selection

28.1 Introduction

28.2 Entropia de um evento

Entropia de uma distribuição de probabilidade: da teoria da informação, é o menor número necessário para codificar os símbolos de uma fonte emissora de sinais. (OBS: número esperado...)

- Vamos usar outra abordagem.
- Seja E um evento num certo espaço amostral.
- O evento E ocorre ou não ocorre em cada realização do experimento aleatório.
- Seja $\mathbb{P}(E) \in [0,1]$ a probabilidade da ocorrência de E.
- A entropia associada com a ocorrência do evento E mede o GRAU DE SURPRESA que a ocorrência de E acarreta.

Entropia de evento é $-\log(p)$

Log em que base?

- Qualquer uma.
- Como $\log_a(x) = \log_a(b)\ \ \log_b(x)$ temos

$$\log_a(x) = c\ \log_b(x)$$

onde $c = \log_a(b)$ é uma constante que depende apenas das duas bases, e não de x.

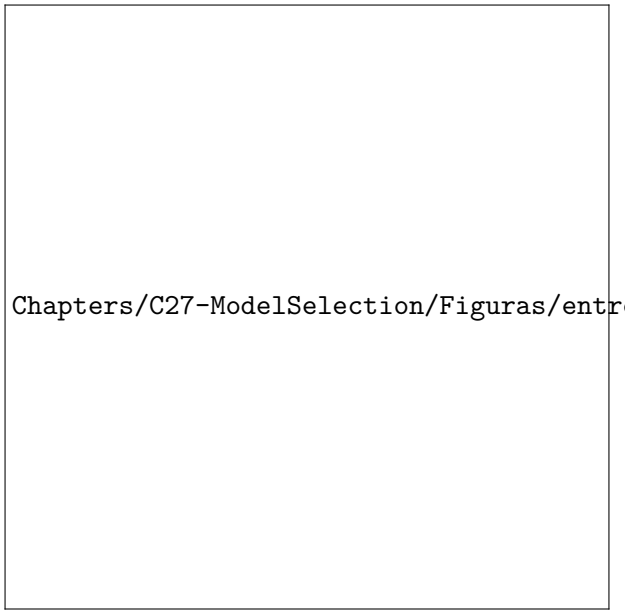
- Isto implica que a diferença absoluta de log's numa base a é igual á diferença na base b vezes uma constante

$$\log_a(x) - \log_a(y) = c\ (\log_b(x) - \log_b(y))$$

- E diferenças relativas são iguais nas duas bases

$$\frac{\log_a(x)}{\log_a(y)} = \frac{\log_b(x)}{\log_b(y)}$$

- A idéia é que muita ou pouca entropia numa base será também muita ou pouca entropia na outra base.



Interpretação de entropia

- Seja a um inteiro entre 1 e 9. Temos

$$0 = \log_{10}(1) \leq \log_{10}(a) \leq \log_{10}(9) \approx 0.95$$

- Seja $p = a.bcdef...10^{-k}$ onde a é um inteiro entre 1 e 9.
- Tome entropia na base 10. Então

$$\begin{aligned} -\log_{10}(p) &= -\log_{10}(a.bcdef\ldots 10^{-k}) \\ &= -\log_{10}(a.bcdef\ldots) - \log_{10}(10^{-k}) \\ &= -\log_{10}(a.bcdef\ldots) + k \\ &\approx k \end{aligned}$$

  já que o primeiro termo é um valor entre 0 e -1 (ver 1a equação).
- Então: ENTROPIA de p é $-\log_{10}(p)$ (aprox) o número de casas decimais antes do primeiro número significativo.

Interpretação de entropia

- Se tomarmos entropia com logs na base 2 (isto é, entropia é $-\log_2(p)$), então a entropia será aprox o número de bits iguais a zero na expansão de p na base 2 antes do primeiro bit significativo.
- Um indivíduo escolhe um número entre $\{0,1,2,\ldots,9\}$
- Uma loteria sorteia um dos números da lista com igual probabilidade.
- A chance de acertar na loteria é $p = 0.1$ com entropia (surpresa) $-\log_{10}(p) = 1$.
- Suponha agora que a lista de números seja $\{0,1,2,\ldots,99\}$.
- A entropia do evento acertar na nova loteria (surpresa) é $-\log_{10}(0.01) = 2$.
- Se a lista de números for $\{0,1,2,\ldots,999\}$.
- a entropia (surpresa) passa a ser $-\log_{10}(0.001) = 3$.

Interpretação de entropia

- Incremento na surpresa é linear com diminuição multiplicativa da probabilidade.
- Incremento de surpresa de ganhar na loteria quando passo de probabilidade 0.1 para 0.01 é $\Delta = 2 - 1 = 1$.
- Incremento ao passar 0.01 para 0.001 é TAMBÉM $\Delta = 3 - 2 = 1$.

- De maneira geral:

$$-\log_{10}\left(\frac{p}{10^k}\right) = -\log_{10}(p) + k$$

- Dividir por 10^k a chance faz aumentar a surpresa em k unidades.
- Surpresa (entropia) cresce linearmente com a ORDEM DE GRANDEZA (ou precisão) de p.

Entropia tem de ser da forma log

- Se entropia (surpresa) funciona desta forma, ela TEM DE SER da forma $-\log(p)$.
- Por quê?
- O que significa "funcionar desta forma"??
- Seja $S : [0,1] \to [0,\infty)$ uma função matemática que visa capturar o sentido de surpresa.
- Queremos que S tenha as seguintes propriedades óbvias:
  1. $S(1) = 0$ (a ocorrência de um evento que tem chance 100% de ocorrência tem surpresa 0, nula).
  2. $S(0) = \infty$ (a ocorrência de um evento impossível traz surpresa infinita)
  3. $S(p)$ é decrescente em p

Propriedade adicional

- Vamos impor uma condição adicional em $S(p)$.
- A função $S(p)$ deverá satisfazer a seguinte propriedade:

$$S(p_2p_1) - S(p_2) = S(p_3p_1) - S(p_3)$$

  para todo p_1, p_2, p_3 em $[0,1]$.
- O que esta propriedade está dizendo?
- Tome o aumento de surpresa $S(p_1p_2) - S(p_2)$ ao passar da ocorrência de um evento com probabilidade p_2 para outro com probabilidade menor p_2p_1.
- Este aumento de supresa é o mesmo se reduzimos a probabilidade p_3 de um evento por p_1 passando então a ter a probabilidade p_3p_1.

Propriedade adicional

- Suponha

$$S(p_2p_1) - S(p_2) = S(p_3p_1) - S(p_3)$$

  para todo p_1, p_2, p_3 em $[0,1]$.
- Por exemplo, ao passar de $p_2 = 0.5$ para $p_2p_1 = 0.5/5 = 0.1$ teremos certo aumento Δ de surpresa.
- Este aumento Δ de surpresa é o mesmo que temos ao passar de $p_3 = 0.0003$ para $p_3p_1 = 0.0003/5 = 0.00006$.
- E isto vale para todo p_1, p_2, p_3.
- Este é o sentido desta propriedade adicional que queremos para a função surpresa.

Teorema

- Uma função S com estas 4 propriedades só pode ser da forma $S(p) = -c\log(p)$, onde c é uma constante positiva qualquer.
- **PROVA:** Tome $p_3 = 1$ na propriedade adicional.
- Como $S(1) = 0$, temos

$$S(p_2p_1) - S(p_2) = S(p_3p_1) - S(p_3) = S(p_1) - S(1) = S(p_1)$$

- e então

$$S(p_2p_1) = S(p_2) + S(p_1)$$

- Isto é, a função $S(p)$ deve transformar produtos em somas.
- A única função com esta propriedade é a função log
- Ver prova disso em livros de análise matemática.

Surpresa acarretada pela ocorrência de v.a. X

- Suponha que X seja uma v.a. discreta com a seguinte distribuição:

Valores Possíveis	x_1	x_2	$\ldots$	x_M
Probabilidades	$p(x_1)$	$p(x_2)$	$\ldots$	$p(x_M)$
Surpresas	$-\log p(x_1)$	$-\log p(x_2)$	$\ldots$	$-\log p(x_M)$

- Um valor aleatório de X é selecionado com as probabilidades acima.
- Suponha que o valor instanciado de X seja x_i
- Se o valor x_i for raro, a surpresa $-\log p(x_i)$ ocasionada por sua ocorrência será grande.
- Se o valor x_i for comum, a surpresa $-\log p(x_i)$ será pequena.
- A surpresa é uma variável aleatória: $-\log p(X)$.
- Qual a surpresa ESPERADA se repetirmos o procedimento de selecionar X com a distribuição acima?

Entropia de v.a. discreta

- Qual a surpresa ESPERADA que a quantidade aleatória $-\log p(X)$ acarreta?

Valores Possíveis	x_1
Probabilidades	$p(x_1)$
Surpresas	$-\log p(x_1)$

- Esperança é a soma de cada valor possível vezes sua probabilidade de ocorrência

$$\begin{aligned}\mathbb{H}_p(p) &= \sum_{i=1}^{n} -\log(p(x_i))\ p(x_i) \\ &= \mathbb{E}_p\{-\log(p(X))\}\end{aligned}$$

- Esta fórmula é a defini[C]ão de entropia de uma distribuição de probabilidade discreta.

Entropia de v.a. discreta

- Entropia de v.a. discreta:

$$\mathbb{H}_p(p) = \sum_{i=1}^{n} -\log(p(x_i))\ p(x_i) = \mathbb{E}_p\{-\log(p(X))\}$$

- Estude esta última notação.
- Perceba que $p(X)$ é o valor de $p(x_i)$ na tabela escolhido ao acaso como função do valor de X.
- A variável X, por sua vez, é selecionada com as mesmas probabilidades p da tabela.
- O sub-índice p sob o símbolo da função esperança e em $\mathbb{H}_p(p)$ é para enfatizar que o X aleatório no argumento da função possui distrbuição dada por p na tabela acima.
- A notação $\mathbb{H}_p(p)$ parece redundante mas ela será útil quando definirmos a distância de Kullback-Leibler.

Exemplo

- X é v.a. discreta com M valores equiprováveis.
- Isto é, $\mathbb{P}(X = x_i) = 1/M$ para todo valor x_i, com $i = 1, 2, \ldots, M$.
- Então $\mathbb{H}_p(p)$ é dada por

$$\begin{aligned}\mathbb{E}_p\{-\log(p(X))\} &= \sum_{i=1}^{M} -\log\left(\frac{1}{M}\right)\frac{1}{M} \\ &= M\frac{1}{M}\left(-\log M^{-1}\right) \\ &= \log(M)\end{aligned}$$

- Assim, a entropia $\mathbb{H}_p(p)$ de uma uniforme é o logaritmo do número de classes equiprováveis.

Exemplo

- X é v.a. com distribuição Poisson com parâmetro λ.

- Então

$$\begin{aligned}\mathbb{E}_p\{-\log(p(X))\} &= \sum_{i=0}^{\infty} -\log\left(\frac{\lambda^k e^{-\lambda}}{k!}\right)\frac{\lambda^k e^{-\lambda}}{k!} \\ &= \lambda - \sum_{i=0}^{\infty}\log\left(\frac{\lambda^k}{k!}\right)\frac{\lambda^k}{k!}e^{-\lambda}\end{aligned}$$

- A entropia $\mathbb{H}_p(p)$ da distribuição de Poisson não possui uma expressão mais simples que esta.

Caso contínuo

- Se X é uma v.a. contínua com densidade $f(x)$ então

$$\mathbb{H}_f(f) = \int_{\mathbb{R}} -\log(f(x))\, f(x)\, dx = \mathbb{E}_f\left[-\log f(X)\right]$$

  onde o sub-índice f na esperança indica que X é selecionada com densidade f.
- Podemos pensar num procedimento em três etapas:
  - Tome $X \sim f$
  - Tome a altura aleatória $f(X)$ da densidade.
  - Tome a esperança de $-\log(f(X))$.

Exemplo - normal

- Suponha que $X \sim N(\mu, \sigma^2)$.
- Neste exemplo, ao invés de integramos uma função, podemos usar o fato de que a variável aleatória padronizada $Z = (X-\mu)/\sigma$ possui distribuição $N(0,1)$.
- Portanto, $\mathbb{E}(Z) = 0$ e $\mathbb{V}(Z) = \mathbb{E}(Z^2) = 1$.
- Temos

$$\begin{aligned}-\log f(x) &= -\log\left[\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)\right] \\ &= \frac{1}{2}\log(2\pi\sigma^2) + \frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\end{aligned}$$

Exemplo - normal

- Portanto

$$\begin{aligned}\mathbb{H}_f(f) &= -\mathbb{E}_f\left[\log f(X)\right] \\ &= \frac{1}{2}\log(2\pi\sigma^2) + \frac{1}{2}\mathbb{E}_f\left(\frac{X-\mu}{\sigma}\right)^2 \\ &= \frac{1}{2}\log(2\pi\sigma^2) + \frac{1}{2} \\ &= \frac{1}{2}\log(2\pi e\sigma^2)\end{aligned}$$

Exemplo - normal

- Se $X \sim N(\mu, \sigma^2)$ então $\mathbb{H}_f(f) = 0.5\log(2\pi e\sigma^2)$.
- Veja que a entropia depende apenas de σ e não de μ.
- Além disso, a entropia aumenta com σ numa escala logarítimica.
- Outro fato curioso: com v.a.'s contínuas, a entropia pode ser negativa se $\sigma^2 < 1/(2\pi e)$.
- As vezes, vamos escrever simplemenste $\mathbb{H}(X)$ para significar $\mathbb{H}_f(f)$

Suppose there are finitely many possible values of X, say $x_1, \ldots, x_m$, and someone picks one of these values with probabilities given by $f(x_i)$, then we try to guess which value has been picked by asking "yes" or "no" questions (e.g., "Is it greater than x_3?"). In this case the entropy (using $\log_2$, as above) may be interpreted as the minimum average number of yes/no questions that must be asked in order to determine the number, the average being taken over replications of the game.

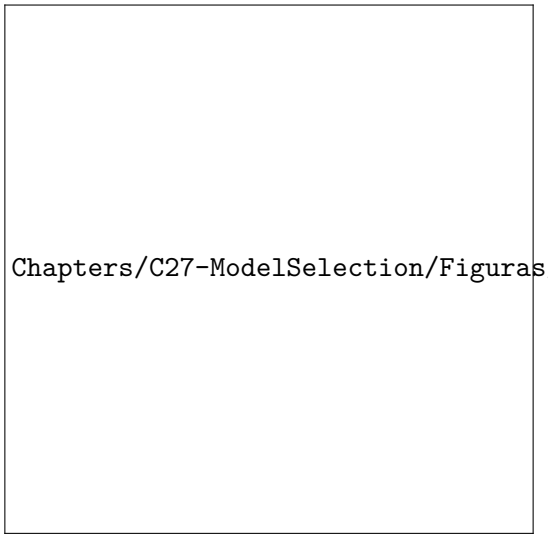


Figure 28.1: Esquerda: $N(0,1)$ e $N(1,1)$ em linha sólida e $N(5,1)$ em linha tracejada. Direita: $N(0,1)$ e $N(1,1)$ em linha sólida e $N(0.5,2^2)$ em linha tracejada.

Entropy may be used to characterize many important probability distributions. The distribution on the set of integers $0,1,2,\ldots,n$ that maximizes entropy subject to having mean μ is the binomial. The distribution on the set of all non-negative integers that maximizes entropy subject to having mean μ is the Poisson. In the continuous case, the distribution on the interval $(0,1)$ having maximal entropy is the uniform distribution. The distribution on the positive real line that maximizes entropy subject to having mean μ is the exponential. The distribution on the positive real line that maximizes entropy subject to having mean μ and variance σ^2 is the gamma. The distribution on the whole real line that maximizes entropy subject to having mean μ and variance σ^2 is the normal.

Medindo distâncias entre distribuições

- Ao comparar duas grandezas físicas A e B (tais como duas massas, duas velocidades ou duas cargas elétricas) sabemos dizer se A e B são aproximadamente iguais ou muito diferentes.
- E ao comparar as *distribuições* de duas variáveis aleatórias X e Y, quando podemos dizer que elas são muito diferentes? Como medir isto?
- Exemplo: $X \sim N(0,1)$, $Y \sim N(1,1)$ e $Z \sim N(5,1)$,
- Esperamos que a distribuição de Y seja próxima daquela de X e afastada da distribuição de Z.
- Mas e se $Z \sim N(0.5,2^2)$? Qual seria mais próxima de X? Y ou Z?

Comparando gaussianas

Medidas de distância entre distribuições

- Existem muitas medidas de distância entre duas distribuições de probabilidade:
  - Kolmogorov: $\sup_x |F_1(x) - F_2(x)|$, onde F_i é a função de distribuição acumulada
  - Hellinger: $\sqrt{0.5 \int (\sqrt{f_1(x)} - \sqrt{f_2(x)})^2 \, dx}$.
  - Variação total: $0.5 \int |f_1(x) - f_2(x)| dx$ Pode-se mostrar quea distância da variação total pode ser escrita como $\sup_A |\int_A (f_1(x) - f_2(x)) dx| = \sup_A |\mathbb{P}_1(A) - \mathbb{P}_2(A)|$: a maior diferença possível entre as probabilidades de A atribuídas por f_1 e f_2.
  - $\int |f_1(x) - f_2(x)| dx$

Medidas de distância entre distribuições

- Qualquer uma dessas medidas é intuitivamente razoável e poderia ser a base de uma teoria de seleção de modelos.
- Entretanto, a distância que gerou mais resultados teóricos e práticos foi a distância de Kullback-Leibler, definida a seguir E DENOTADA POR KL.

Suporte e surpresa

- Suponha que temos duas distribuições f_1 e f_2 competindo como modelos para descrever alguns dados.
- Seja $\mathscr{S}$ o suporte da distribuição f_1. Isto é,
  - No caso discreto: $\mathscr{S}_1 = \{z;\ \mathbb{P}_1(X = z) > 0\}$,
  - No caso contínuo, $\mathscr{S}_1 = \{z;\ f_1(z) > 0\}$
- Vamos assumir que $\mathscr{S}_1 = \mathscr{S}_2 = \mathscr{S}$.
- Para cada $z \in \mathscr{S}$, calculamos a surpresa ocasionada por gerar z sob f_1 e sob f_2.

Surpresa

- Para cada $z \in \mathscr{S}$, vemos a surpresa $-\log(f_i(z))$ de ocorrência sob f_1 e sob f_2.
- Se as duas distribuições são parecidas, esperamos que as surpresas $-\log(f_1(z))$ e $-\log(f_2(z))$ sejam similares para todo z.
- A diferença de surpresa é

$$-\log(f_2(z)) - (-\log(f_1(z))) = \log\left(\frac{f_1(z)}{f_2(z)}\right)$$

- Note que ela é o logaritmo da razão de verossimilhança do modelo 1 versus o modelo 2.
- Se $\log(f_1(z)/f_2(z)) = 0$, o valor z é igualmente provável nas duas distribuições
- Se $\log(f_1(z)/f_2(z)) > 0$, o valor z tem mais chance de ocorrer sob a distribuição 1.

Tirando a média

- Para ter uma idéia global ou um resumo da diferença de supresas considerando todos os possíveis valores z, tiramos uma "média" das diferenças $\log(f_1(z)/f_2(z)) > 0$.
- Queremos uma média ponderada sobre todos os valores possíveis de $z \in \mathscr{S}$.
- Queremos dar mais peso às discrepâncias $\log(f_X(z)/f_Y(z))$ associadas aos valores z que têm mais chance de ocorrer.
- Para isto, precisamos escolher um modelo de probabilidade para os valores $z \in \mathscr{S}$. Temos dois modelos possíveis: $f_1(z)$ ou $f_2(z)$.
- De maneira um tanto arbitrária, vamos escolher $f_1(z)$ para descrever as frequências dos valores $z \in \mathscr{S}$.

Kullback e Leibler (1951)

- A medida de Kullback-Leibler é definida no caso contínuo como

$$KL(f_1, f_2) = 2\int_{\mathscr{S}} \log\left(\frac{f_1(z)}{f_2(z)}\right)\ f_1(z)\,dz$$

- No caso discreto, como:

$$KL(p_1, p_2) = 2\sum_{z_i \in \mathscr{S}} \log\left(\frac{\mathbb{P}_1(X = z_i)}{\mathbb{P}_2(X = z_i)}\right)\ \mathbb{P}_1(X = z_i)$$

- Algumas vezes, a definição não utiliza a constante 2.

KL em passos

- Observe que KL é equivalente ao seguinte procedimento:
  - X é uma v.a. com densidade $f_1(x)$.
  - Ao gerar um valor aleatório X sob f_1, calcule a v.a. $Z = h(X) = 2\log(f_1(X)/f_2(X))$.
  - A seguir, tome esperança de Z (lembrando que X segue a distribuição f_1):

$$\begin{aligned} KL(f_1, f_2) &= \mathbb{E}_1[h(X)] \\ &= 2\,\mathbb{E}_1\left[\log\left(\frac{f_1(X)}{f_2(X)}\right)\right] \\ &= 2\int \log\left(\frac{f_1(x)}{f_2(x)}\right)\ f_1(x)\,dx \end{aligned}$$

Exemplo: Poisson

- $X \sim \text{Poisson}(\mu_1)$ e $Y \sim \text{Poisson}(\mu_2)$.
- Temos

$$\log\left(\frac{p_X(k)}{p_Y(k)}\right) = \log\left(\frac{\mu_1^k \exp(-\mu_1)/k!}{\mu_2^k \exp(-\mu_2)/k!}\right) = k\,\log\left(\frac{\mu_1}{\mu_2}\right) - (\mu_1 - \mu_2)$$

- Por exemplo, se $\mu_1 = 3$ e $\mu_2 = 5$ então $\log(p_X(k)/p_Y(k)) = k\log(3/5) - (3-5) = -0.51\,k + 2$.
- Fazemos k aleatório com a distribuição de X e a medida de Kullback-Leibler é

$$\begin{aligned} KL(p_X, p_Y) &= 2E_X\left[X\,\log\left(\frac{\mu_1}{\mu_2}\right) - (\mu_1 - \mu_2)\right] \\ &= 2\left[E_X(X)\log\left(\frac{\mu_1}{\mu_2}\right) - (\mu_1 - \mu_2)\right] \\ &= 2\left[\mu_1\log\left(\frac{\mu_1}{\mu_2}\right) - (\mu_1 - \mu_2)\right] \end{aligned}$$

Exemplo: Poisson

- Vimos que, se $X \sim \text{Poisson}(\mu_1)$ e $Y \sim \text{Poisson}(\mu_2)$, então

$$KL(p_X, p_Y) = 2\left[\mu_1\log\left(\frac{\mu_1}{\mu_2}\right) - (\mu_1 - \mu_2)\right]$$

- Por exemplo, se $\mu_1 = 3$ e $\mu_2 = 5$, teremos $KL(p_X, p_Y) = 0.9350$.
- Se $\mu_1 = 30$ e $\mu_2 = 50$, teremos $KL(p_X, p_Y) = 9.3504$, o valor anterior multiplicado por 10.
- Se $\mu_1 = 0.3$ e $\mu_2 = 0.5$, teremos $KL(p_X, p_Y) = 0.093504$, o primeiro valor dividido por 10.
- De fato, é bem fácil mostrar que, se μ_1 e μ_2 são multiplicados por uma mesma constante $c > 0$, a distância KL entre duas Poissons também é multiplicada por c (basta olhar a fórmula).

[[t]

- Em geral, $KL(f_1, f_2) \neq KL(f_2, f_1)$.
- Isto é, KL não é simétrica nos seus argumentos.
- Isto implica que a medida de Kullback-Leibler não é, de fato, uma medida de distância no sentido matemático do termo.
- A distância KL de f_1 até f_2 não é igual à distância KL de f_2 até f_1
- Algumas (poucas) pessoas preferem definir uma nova medida (simétrica) dada por

$$\frac{1}{2}\left(KL(f_1, f_2) + KL(f_2, f_1)\right)$$

- Continuaremos a usar $KL(f_1, F_2)$, chamando-a de distância entre as distribuições f_1 e f_2.

Poisson e distância KL

- Com $X \sim \text{Poisson}(\mu_1)$ e $Y \sim \text{Poisson}(\mu_2)$, temos

$$KL(p_X, p_Y) = 2\left[\mu_x\log\left(\frac{\mu_x}{\mu_y}\right) - (\mu_x - \mu_y)\right]$$

- Se $\mu_x = 3$ e $\mu_y = 5$, teremos $KL(p_X, p_Y) = 0.9350$.
- Mas se trocarmos, fazendo $\mu_x = 5$ e $\mu_y = 3$, teremos $KL(p_X, p_Y) = 1.108256$.
- Vamos tentar entender o porquê dessa diferença.

Distância KL é assimétrica

- No caso discreto, KL é definida assim:

$$KL(p_1, p_2) = \mathbb{E}_{Z \sim p_1}\left(\frac{p_1(Z)}{p_2(Z)}\right)$$

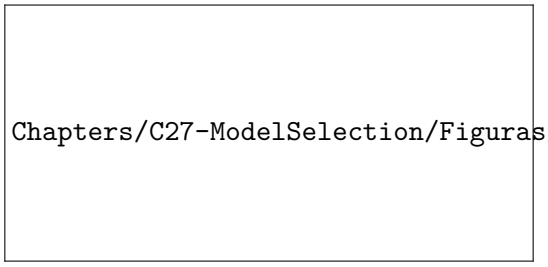


Figure 28.2: Densidades de $f_1(x) \sim U(0,1)$ e $f_2(x) \sim \text{Beta}(20,20)$.

- Geramos um dado aleatório Z por p_1 (aqui está a assimetria)
- Em seguida, olhamos quão mais provável é este Z sob p_1 relativamente à p_2 calculando $p_1(Z)/p_2(Z)$
- Por exemplo, se $Z = z$ e $f_1(z)/f_2(z) = 3$, então o z gerado (por p_1) é 3 vezes mais provável sob p_1 do que sob p_2.
- Se $p_1 \approx p_2$, esperamos que essa razão $p_1(Z)/p_2(Z)$ seja tipicamente próxima de 1.
- Se p_1 for muito distante de p_2, esperamos que essa razão $p_1(Z)/p_2(Z)$ seja em geral bem maior que 1.

Distância KL é assimétrica

- O fato é que olhamos a distância $KL(p_1, P_2)$ numa forma assimétrica.
- Um dado Z é gerado de p_1. Então $KL(p_1, p_2)$ mede o quanto podemos esperar que esse dado aleatório de p_1 pode ter sido gerado por p_2.
- $KL(p_2, p_1)$ mede a situação reversa: um dado Z é gerado por p_2 e nos pergutamos se é razoável que ele tenha vindo de p_1.
- O exemplo a seguir mostra que é razoável que $KL(p_2, p_1) \neq KL(P_1, p_2)$.

Distância KL é assimétrica

- Considere $f_1(x)$ uma uniforme $U(0,1)$ e $f_2(x)$ uma Beta$(20,20)$, bem concentrada em $(0.3, 0.7)$.

Distância KL é assimétrica

- $f_1(x) \sim U(0,1)$ e $f_2(x) \sim \text{Beta}(20,20)$, bem concentrada em $(0.3, 0.7)$.
- Um dado vindo de f_2 estará em geral bem concentrado em torno de $1/2$ e pode facilmente ser considerado como sendo gerado por uma $U(0,1)$. Por exemplo, se $Z = 0.4$ é gerado, poderíamos facilmente tomá-lo como tendo sido gerado por uma $U(0,1)$.
- Veja (a partir do gráfico) que $2\log(f_1(z)/f_2(z))$ fica entre $2\log(1/1) = 0$ e $2\log(5/1) = 3.2$.
- Podemos facilmente obter por simulação $KL(Beta(20,20), U(0,1)) \approx 2.2$.

Distância KL é assimétrica

- Vamos olhar agora a situação reversa: suponha que geramos $Z \sim U(0,1)$.
- Seria razoável esperar que este Z venha de uma Beta$(20,20)$?
- Se $Z \in (0.3, 0.7)$ n´ ao teremos razão para descrtar essa possibilidade.
- Entretanto, $\mathbb{P}(U(0,1) \notin (0.3, 0.7)) = 0.6$.
- Isto é, existe uma chance razoável da $U(0,1)$ gerar um dado fora de $(0.3, 0.7)$ onde a Beta$(20,20)$ está concentrada.
- Um dado fora de $(0.3, 0.7)$ dificilmente seria gerado por uma Beta$(20,20)$.
- De fato, temos $KL(U(0,1), Beta(20,20)) \approx 20$ (por simulação)
- Isto é, $20 \approx KL(U(0,1), Beta(20,20)) >> KL(Beta(20,20), U(0,1)) \approx 2$

Caso normal

- Sejam $\mathbf{X} \sim N_n(\mu_1, \sigma_1^2 I_n)$ e $\mathbf{Y} \sim N_n(\mu_2, \sigma_2^2 I_n)$.

- Neste caso,

$$\begin{aligned}\log\left(\frac{f_{\mathbf{X}}(\mathbf{z})}{f_{\mathbf{Y}}(\mathbf{z})}\right) &= \log\left(\frac{(2\pi\sigma_1^2)^{-n/2}\exp\left(\|\mathbf{z}-\mu_1\|^2/(2\sigma_1^2)\right)}{(2\pi\sigma_2^2)^{-n/2}\exp\left(\|\mathbf{z}-\mu_2\|^2/(2\sigma_2^2)\right)}\right)\\ &= n\log\left(\frac{\sigma_2}{\sigma_1}\right)-\frac{1}{2\sigma_1^2}\|\mathbf{z}-\mu_1\|^2+\frac{1}{2\sigma_2^2}\|\mathbf{z}-\mu_2\|^2\end{aligned}$$

- Usando que $\|\mathbf{X}-\mu_1\|^2/\sigma_1^2 \sim \chi^2(n)$ (qui-quadrado com n graus de liberdade) e que $\|\mathbf{X}-\mu_2\|^2$ é uma qui-quadrado não-central podemos deduzir:

$$KL(f_{\mathbf{X}}, f_{\mathbf{Y}}) = 2\left[n\log\left(\frac{\sigma_2}{\sigma_1}\right)-\frac{n}{2}+\frac{n\sigma_1^2}{2\sigma_2^2}+\frac{\|\mu_1-\mu_2\|^2}{2\sigma_2^2}\right],$$

- É uma função da distância euclidiana entre os vetores de médias e das razões entre as variâncias das distribuições.

Caso normal

- Se $X \sim N(0,1)$, $Y \sim N(1,1)$ e $Z \sim N(5,1)$, então
- $KL(f_X, f_Y) = 1$, $KL(f_X, f_Z) = 25$ e $KL(f_Y, f_Z) = 16$
- Mas se agora tivermos $Z \sim N(0.5, 2^2)$, então
- $KL(f_X, f_Y) = 1$ (como antes) e $KL(f_X, f_Z) = 0.6988$ e $KL(f_Y, f_Z) = 0.6987$.
- Ou seja, Z está igualmente distante de X e Y e mais perto de cada uma delas que a distância entre X e Y.

Relembrando de probab

- Se $Y_1, Y_2, \ldots, Y_n$ são i.i.d. com a mesma distribuição que Y e se $\mu = E(Y)$ então

$$\hat{\mu} = \sum_i Y_i/n \to \mu$$

- Isto é, o estimador $\sum_i Y_i/n$ (média aritmética dos elementos da amostra i.i.d.) converge para a a média populacional (a esperança) de Y.
- Isto vale PARA QUALQUER V.A. QUE POSSUA esperança.
- Repetindo: QUALQUER v.a. Y.
- Isto é muito simples mas muito importante quando acoplado com a idéia de transformar uma v.a. como veremos no próximo slide.

Relembrando de probab

- Se X é uma v.a. e $Y = h(X)$ é uma v.a. obtida como função de X.
- Por exemplo, $Y = X^2$ ou então $Y = \log(1+X^2)$
- Se $X_1, \ldots, X_n$ são i.i.d. com distribuição de X então $Y_1 = h(X_1), \ldots, Y_n = h(X_n)$ são i.i.d. com uma distribuição de $Y = h(X)$.
- Por exemplo: $Y_1 = X_1^2, Y_2 = X_2^2, \ldots, Y_n = X_n^2$ são v.a.'s i.i.d.
- Outro exemplo: $Y_1 = \log(1+X_1^2), Y_2 = \log(1+X_2^2), \ldots, Y_n = \log(1+X_n^2)$ são v.a.'s i.i.d.

Relembrando de probab

- Então $\mu_Y = E(Y) = E(h(X))$ pode ser estimado consistentemente pela média aritmática $\hat{\mu}_Y = \sum_i Y_i/n = \sum_i h(X_i)/n$.
- Por exemplo:

$$\frac{Y_1+Y_2+\ldots+Y_n}{n} = \frac{X_1^2+X_2^2+\ldots+X_n^2}{n} \to \mathbb{E}(X^2) = \mathbb{E}(Y)$$

ou

$$\frac{Y_1+\ldots Y_n}{n} = \frac{\log(1+X_1^2)+\ldots+\log(1+X_n^2)}{n} \to \mathbb{E}(\log(1+X^2))$$

Estimando Kullback-Leibler

- A distância KL é apenas uma esperança de uma v.a. $Y = h(X)$ onde X é selecionada com a distribuição f_1:

$$KL(f_1, f_2) = 2\mathbb{E}_1 [h(X)] = 2\mathbb{E}_1 \left[\log \left(\frac{f_1(X)}{f_2(X)} \right) \right]$$

onde $h(x)$ é a função dada por

$$h(x) = \log \left(\frac{f_1(x)}{f_2(x)} \right)$$

- Se tivermos uma amostra i.i.d. de X obtida da distribuição f_1, podemos transformar cada X com a função h e a seeguir calcular a sua média aritmética.
- Este valor será aproximadamente igual a esperança teórica.

Estimando Kullback-Leibler

- Se $X_1, \ldots, X_n$ são i.i.d. com distribuição de X, podemos estimar

$$KL(f_1, f_2) = 2\mathbb{E}_1 \left[\log \left(\frac{f_1(X)}{f_2(X)} \right) \right]$$

pela média aritmética dos valores $h(X_i)$:

$$\widehat{KL(f_1, f_2)} = 2 \frac{1}{n} \sum_i \log \left(\frac{f_1(X_i)}{f_2(X_i)} \right)$$

Exemplo

- X e Y são v.a.'s Poisson com médias 3 e 5. Então

$$\log \left(\frac{p_X(k)}{p_Y(k)} \right) = k \log \left(\frac{3}{5} \right) - (3-5) = 2 - 0.511\, k$$

- Temos $KL(p_X, p_Y) = 2(2 - 0.511\, E_X(X)) = 2(0.468)$.
- Neste caso, sabemos exatamente qual é o valor de $KL(p_X, p_Y)$ e portanto, nem faz sentido estimá-lo. No entanto, se mesmo assim quiséssemos, podemos fazer o seguinte:
- Os valores observados de uma amostra de tamanho 10 de X com distribuição Poisson de média 3 são os seguintes:

$$5, 3, 4, 1, 2, 5, 2, 1, 8, 2$$

- Então $KL(p_X, p_Y) = 2(0.468)$ pode ser estimado por

$$\widehat{K(p_X, p_Y)} = \frac{2}{10} \sum_i (k_i \log(3/5) + 2) = 2(0.314)$$

Todo modelo é correto

- Em análise de dados com modelos paramétricos, selecionamos uma classe de distribuições $f(\mathbf{Y}, \mathbf{t})$ para o vetor aleatório $\mathbf{Y}$.
- Fizemos uma análise de como o MLE $\widehat{\mathbf{t}}$ se comporta quando um dos elementos desta classe é o modelo gerador dos dados.

- Suponha que $\mathbf{t}_0 \in \Theta$ é o verdadeiro valor do parâmetro.
- Isto é, os dados são gerados pela distribuição $f(\mathbf{Y}, \mathbf{t}_0)$
- Então, o MLE $\widehat{\mathbf{t}}$ baseado numa amostra possui as seguintes propriedades:
  - $\mathbb{E}(\widehat{\mathbf{t}}) \approx \mathbf{t}_0$ (é aproximadamente não-viciado)
  - $\mathbb{V}(\widehat{\mathbf{t}}) \approx \mathbb{I}^{-1}(\mathbf{t}_0)$
  - $\widehat{\mathbf{t}} \approx N(\mathbf{t}_0, \mathbb{I}^{-1}(\mathbf{t}_0)$

Todo modelo é falso

- Uma frase famosa de George Box é: *Todo modelo é falso mas alguns são úteis*.
- Praticamente sempre, nossos modelos são distribuições idealizadas e simplificadoras. Esperamos que eles sejam capazes de descrever de forma aproximada o mecanismo verdadeiro que gera os dados.
- Mas se o modelo verdadeiro que gera os dados não é um membro $f(\mathbf{Y}, \mathbf{t}_0)$ da classe $f(\mathbf{Y}, \mathbf{t})$, então não existe nenhum $\mathbf{t}_0$ verdadeiro.
- Neste caso, o *MLE estará estimando o quê*?
- Ele converge para algum lugar quando a amostra aumenta?
- Se existir um ponto de convergência, ele faz algum sentido prático?
- A discussão a seguir vai supor que os dados sejam i.i.d. mas ela é válida num contexto mais geral de dados independentes mas não i.d. ou até mesmo de dados dependentes.

Minimizando KL

- Suponha que $g(y)$ é a verdadeira densidade que gera os dados observados.
- Propomos um modelo parametrizado como aproximação para g.
- O modelo é uma classe (um conjunto) de densidades $f(y, \mathbf{t})$ indexado pelo parâmetro $\mathbf{t}$.
- Vamos denotar por $f_\mathbf{t}$ a densidade $f(y, \mathbf{t})$.
- Uma estratégia interessante para modelar os dados gerados por g é procurar na classe de densidades $\{f(y, \mathbf{t})\}$ aquela que minimiza a distância KL.
- Isto é, vamos procurar na classe $\{f(y, \mathbf{t})\}$ aquele valor $\mathbf{t}_0$ tal que a densidade $f(y, \mathbf{t}_0)$ seja a mais próxima possível de g.
- Em símbolos,

$$\mathbf{t}_0 = \arg_\mathbf{t} \min KL(g, f_\mathbf{t})$$

Minimizando KL

- Se $\mathbf{t}_0 = \arg_\mathbf{t} \min KL(g, f_\mathbf{t})$ então $f(y, \mathbf{t}_0)$ é o elemento da classe mais próximo de g.
- Se não pudermos encontrar g, o melhor que podemos fazer é usar $f(y, \mathbf{t}_0)$ em seu lugar.
- Temos

$$KL(g, f_\mathbf{t}) = \mathbb{E}_g \log\left(\frac{g(Y)}{f(Y, \mathbf{t})}\right) = \mathbb{E}_g \log\left(g(Y)\right) - \mathbb{E}_g \log\left(f(Y; \mathbf{t})\right)$$

- O primeiro termo, $\mathbb{E}_g(\log g(Y))$, não depende de $\mathbf{t}$.
- Portanto, minimizar $KL(g, f_\mathbf{t})$ é o mesmo que procurar o valor de $\mathbf{t}$ que maximiza o segundo termo:

$$\mathbf{t}_0 = \arg_\mathbf{t} \min KL(g, f_\mathbf{t}) = \arg\max \mathbb{E}_g \log\left(f(Y; \mathbf{t})\right)$$

Minimizando KL

- Como encontrar este elemento

$$\mathbf{t}_0 = \arg_\mathbf{t} \min KL(g, f_\mathbf{t}) = \arg\max \mathbb{E}_g \log\left(f(Y; \mathbf{t})\right) \quad ?$$

- Em alguns casos simples isto é possível, como veremos a seguir.

- g é uma densidade contínua arbitrária e que vai gerar os nossos dados.
- Nosso modelo é a classe de todas as gaussianas: $\{N(\mu,\sigma^2)\}$.
- Aqui $\mathbf{t} = (\mu,\sigma^2)$.
- Como encontrar (se é que existe) uma (única??) gaussiana $N(\mu_0,\sigma_0^2)$ que melhor aproxima uma densidade g arbitrária minimizando a distânca KL?
- Isto é, qual é q $N(\mu,\sigma^2)$ que tem $KL(g,N(\mu,\sigma^2))$ mínima?

Minimizando KL na classe das gaussianas

- Queremos (μ_0,σ_0^2) que minimizem

$$KL(g,N(\mu,\sigma^2)) = 2\mathbb{E}_g \log\left(g(Y)\right) - 2\mathbb{E}_g \log\left(\phi(x;\mu,\sigma^2)\right)$$

  onde $\phi(x;\mu,\sigma^2)$ é a densidade de uma gaussiana com parâmetros (μ,σ^2)
- O primeiro termo não depende de (μ,σ^2) e pode ser ignorado.
- Assim, basta maximizar

$$\begin{aligned}
\mathbb{E}_g \log\left(\phi(x;\mu,\sigma^2)\right) &= \mathbb{E}_g \log\left((2\pi)^{-1/2} - \frac{1}{2}\log\sigma^2 - \frac{1}{2}\left(\frac{X-\mu}{\sigma}\right)^2\right) \\
&= \text{cte.} + \log\sigma^2 + \mathbb{E}_g\left(\frac{X-\mu}{\sigma}\right)^2 \\
&= \text{cte.} + \log\sigma^2 + \frac{1}{\sigma^2}\mathbb{E}_g\left(X-\mu\right)^2
\end{aligned}$$

Minimizando KL na classe das gaussianas

- Queremos (μ_0,σ_0^2) que maximizem

$$\mathbb{E}_g \log\left(\phi(x;\mu,\sigma^2)\right) = \text{cte.} + \log\sigma^2 + \frac{1}{\sigma^2}\mathbb{E}_g\left(X-\mu\right)^2$$

- Para qualquer valor de σ^2 fixo, $\mathbb{E}_g\left(X-\mu\right)^2$ é minimizado se tomarmos $\mu_0 = \mathbb{E}_g(X)$.
- Com este valor μ_0 inserido na expressão acima, temos que achar σ^2 que maximize

$$\text{cte.} + \log\sigma^2 + \frac{1}{\sigma^2}\mathbb{E}_g\left(X-\mu_0\right)^2$$

- Derivando em σ^2 igualando a zero encontramos

$$\sigma_0^2 = \mathbb{E}_g\left(X-\mu_0\right)^2 = \mathbb{V}_g(X)$$

Minimizando KL na classe das gaussianas

- Isto é, dada uma g qualquer, a gaussiana $N(\mu,\sigma^2)$ que tem a distância de Kullback-Leibler mínima é $N(\mu_0,\sigma_0^2)$ onde $\mu_0 = \mathbb{E}_g(X)$ e $\sigma_0^2 = \mathbb{V}_g(X)$.
- Por exemplo, se g é a densidade de uma exponencial dupla (ou distribuição de Laplace, veja na wikipedia), com esperança 1 e variância 1 então a gaussiana que melhor aproxima é a $N(0,1)$.

Minimizando KL na classe das gaussianas

- Outro exemplo: g é a densidade de uma gama com $\alpha = 4$ e $\beta = 1$. Isto implica que g tem esperança 4 e variância 4.
- A gaussiana que melhor aproxima esta gama é a $N(4,4)$.

Quando temos o modelo errado, o MLE estima o quê?

- Suponha que o vetor $\mathbf{Y} = (Y_1,\ldots,Y_n)$ é composto de v.a.'s i.i.d. com uma distribuição desconhecida com densidade $g(\mathbf{Y})$.
- Adotamos um modelo $f(\mathbf{Y},\mathbf{t}) = \prod_i f(y_i,\mathbf{t})$ para os dados i.i.d. e obtemos o MLE maximizando a log-verossimilhança (dividida pela constante n):

$$\widehat{\mathbf{t}} = \arg\max \frac{1}{n}\sum_i \log f(Y_i;\mathbf{t})$$

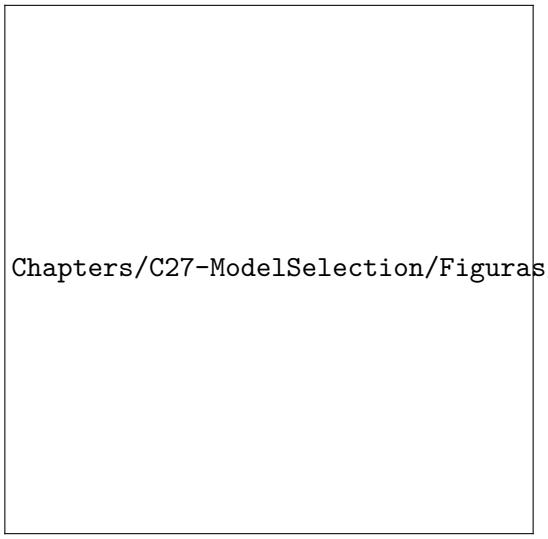


Figure 28.3: Laplace (linha contínua) e gaussiana mais próxima

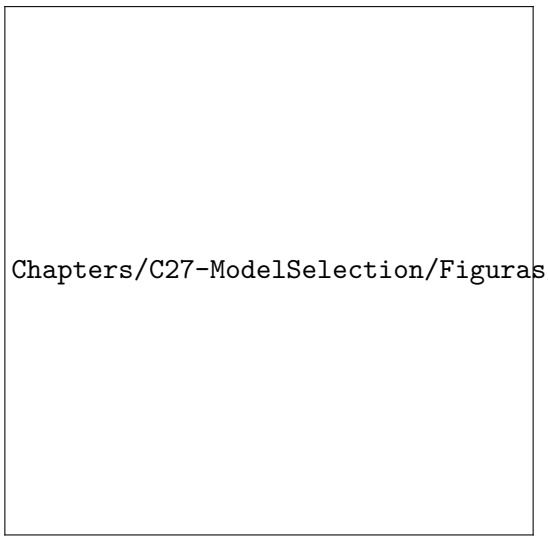


Figure 28.4: Gamma$(4,1)$ (linha contínua) e gaussiana mais próxima $N(4,4)$

O MLE estima o quê?

- Para cada valor $\mathbf{t}$ fixo, considere as v.a.'s $W_1 = \log f(Y_1;\mathbf{t}), \ldots, W_n = \log f(Y_n;\mathbf{t})$
- A média populacional (a esperança) de W é

$$\mathbb{E}_g(W) = \mathbb{E}_g\left(\log f(Y;\mathbf{t})\right)$$

onde Y no lado direito é uma v.a. com densidade $g(y)$.

- Lembre-se de um resultado de probab (a lei dos grandes números): A média aritmética de v.a.'s i.i.d. converge para sua esperança populacional.
- A média aritmética baseada na amostra é

$$\frac{W_1 + \ldots + W_n}{n} = \frac{1}{n}\sum_i \log f(Y_i;\mathbf{t})$$

O MLE estima o quê?

- Pela Lei dos Grandes números, temos

$$\frac{W_1 + \ldots + W_n}{n} \to \mathbb{E}_g(W)$$

- Ou seja,

$$\frac{1}{n}\sum_i \log f(Y_i;\mathbf{t}) \to \mathbb{E}_g\left(\log f(Y;\mathbf{t})\right)$$

para todo $\mathbf{t}$ fixo.

- Por definição o MLE de $\mathbf{t}$ é o argumento em $\mathbf{t}$ que maximiza a log-verossimilhança.
- Em símbolos:

$$\widehat{\mathbf{t}} = \arg_{\mathbf{t}} \max \frac{1}{n}\sum_i \log f(Y_i;\mathbf{t})$$

O MLE estima o quê?

- Vamos definir $\mathbf{t}_0$ como o argumento em $\mathbf{t}$ que maximiza

$$\mathbf{t}_0 = \arg_{\mathbf{t}} \max \mathbb{E}_g\left(\log f(Y;\mathbf{t})\right)$$

- Mas nós acabamos de ver que maximizar $\mathbb{E}_g\left(\log f(Y;\mathbf{t})\right)$ em $\mathbf{t}$ é o mesmo que minimizar $KL(g, f_{\mathbf{t}})$ em $\mathbf{t}$.
- Assim, $\mathbf{t}_0$ definido acima tem o memso significado que antes:

$$\mathbf{t}_0 = \arg_{\mathbf{t}} \max \mathbb{E}_g\left(\log f(Y;\mathbf{t})\right) = \arg_{\mathbf{t}} \min KL(g, f(y;\mathbf{t}))$$

- Vamos agora relacionar o MLE $\widehat{\mathbf{t}}$ e $\mathbf{t}_0$.

O MLE estima o quê?

- O MLE $\widehat{\mathbf{t}}$ é uma v.a. e $\mathbf{t}_0$ é um valor fixo no espaço paramétrico, uma constante.
- Em geral, $\widehat{\mathbf{t}} \neq \mathbf{t}_0$.
- Mas eles estão relacionados. Como:

$$\frac{1}{n}\sum_i \log f(Y_i;\mathbf{t}) \to \mathbb{E}_g\left(\log f(Y;\mathbf{t})\right)$$

podemos esperar que

$$\widehat{\mathbf{t}} = \arg_{\mathbf{t}} \max \frac{1}{n}\sum_i \log f(Y_i;\mathbf{t}) \to \arg_{\mathbf{t}} \max \mathbb{E}_g\left(\log f(Y;\mathbf{t})\right) = \mathbf{t}_0$$

- De fato, podemos demonstrar isto rigorosamente sob certas condições mas não faremos isto neste curso.

O MLE estima o quê?

- Assim, temos

$$\widehat{\mathbf{t}} = \arg_{\mathbf{t}} \max \frac{1}{n} \sum_i \log f(Y_i; \mathbf{t}) \to \arg_{\mathbf{t}} \max \mathbb{E}_g \left(\log f(Y; \mathbf{t}) \right) = \mathbf{t}_0$$

- A medida que n cresce, o MLE $\widehat{\mathbf{t}}$ converge para o valor $\mathbf{t}_0$ que minimiza a distância KL entre o modelo verdadeiro g e a classe $\{f(y, \mathbf{t})\}$.
- Agora entedemos o que o MLE está fazendo.
- Existe um elemento $\mathbf{t}_0$ da classe de distribuições $\{f(y, \mathbf{t})\}$ que forma nosso modelo que é a mais KL-próxima possível da distribuição verdadeira g.
- O MLE $\widehat{\mathbf{t}}$ é um estimador deste valor $\mathbf{t}_0$.

O MLE estima o quê?

- Se g for de fato um elemento $f(y, \mathbf{t}_0)$ da classe especificada no modelo, temos todos os resultados que já vimos neste curso:

$$\widehat{\mathbf{t}} \approx N(\mathbf{t}_0, I^{-1}(\mathbf{t}_0)$$

- No caso mais comum em que g não pertence à classe $\{f(y, \mathbf{t})\}$ do modelo, Akaike demonstrou que, se a amostra não é muito pequena:

$$\widehat{\mathbf{t}} \approx N(\mathbf{t}_0, V)$$

  onde V mistura a informação de Fisher com outra matriz.
- O mais incrível neste processo é que podemos obter o MLE e uma estimativa $f(y, \widehat{\mathbf{t}})$ da melhor aproximação $f(y, \mathbf{t}_0)$ de g sem ter a menor idéia de quem é g.

Comparando modelos

- Suponha que temos dois modelos alternativos, 1 e 2.
- Cada um deles é uma classe de distribuições indexadas por parâmetros.
- Digamos, θ para o modelo 1 e ϕ para o modelo 2.
- Modelo 1: $f(y, \theta)$.
- Modelo 2: $h(y; \phi)$ (vamos usar h para as densidades do modelo 2)
- Os parâmetros θ e ϕ podem ter dimensões e interpretações físicas diferentes.
- Qual deles é o melhor para descrever dados gerados de uma distribuição g desconhecida?

Comparando modelos

- Temos uma medida de distância da entre g e uma classe de distribuições.
- Seja θ_0 o valor de θ que minimiza a distância $KL(g, f(y, \theta))$.
- Isto é,

$$\theta_0 = \arg\min KL(g, f(y, \theta))$$

- A distância mínima entre g e o modelo 1 é $KL(g, f(y, \theta_0))$.
- Do mesmo modo, teremos a distância mínima entre g e o modelo 2 dada por $KL(g, h(y, \phi_0))$ onde

$$\phi_0 = \arg\min KL(g, h(y, \phi))$$

Comparando modelos

- Assim, o natural é comparar $KL(g, f(y, \theta_0))$ e $KL(g, h(y, \phi_0))$.
- O que tiver distância mínima é o escolhido.

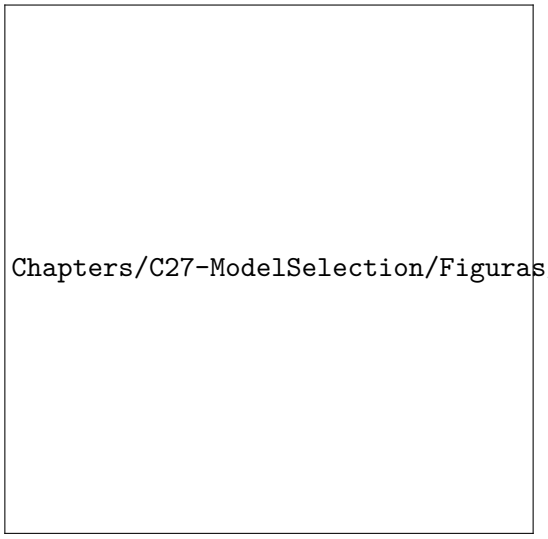


Figure 28.5: kk

- Podemos acrescentar um critério adicional: Se o modelo 2 for muito mais complicado que o modelo 1 e se $KL(g, f(y,\theta_0)) > KL(g, h(y,\phi_0))$ mas a diferença for muito pequena, podemos ficar com o modelo 1, mais simples e que tem praticamente a mesma distância que o modelo 2.
- OK, mas como definir se a distância é pequena?
- E muito mais importante: como calcular $KL(g, f(y,\theta_0))$ e $KL(g, h(y,\phi_0))$ na prática?

Akaike

- Akaike resolveu estes dois problemas para nós.
- Seja k o índice do modelo ($k = 1$ ou $k = 2$, em nosso exemplo, mas podemos ter vários modelos ao mesmo tempo)
- Seja p_k o número de parâmetros livres do modelo k.
- Ele mostrou que, se calcularmos o valor de

$$AIC(k) = -2\log f(\mathbf{Y}, \widehat{\theta}_k) + 2p_k$$

  para cada modelo alternativo, o que tiver o menor $AIC(k)$ deve ser o melhor modelo.
- Para maiores referências, veja o capítulo 13 de Pawitan, In all likelihood.

28.3 Introdução

28.3.1 The dependence of two random vectors may be quantified by mutual information.

It often happensmutual information that the deviation of one distribution from another must be evaluated. Consider two continuous pdfs $f(x)$ and $g(x)$, both being positive on (A,B). The *Kullback-Leibler (KL) discrepancy*Kullback-Leibler (KL) discrepancy is the quantity

$$D_{KL}(f,g) = \mathbb{E}_f\left(\log\frac{f(X)}{g(X)}\right)$$

where the subscript on the expectation $\mathbb{E}_f$ signifies that the random variable X has pdf $f(x)$. In other words, we have

$$D_{KL}(f,g) = \int_A^B f(x)\log\frac{f(x)}{g(x)}dx.$$

The KL discrepancy may also be defined, analogously, for discrete distributions. Note that $D_{KL}(f,g)$ may also be written in the difference form

$$D_{KL}(f,g) = \mathbb{E}_f(\log f(X)) - \mathbb{E}_f(\log g(X)). \tag{28.1}$$

In fact, the KL discrepancy is essentially unique among all discrepancies $D(f,g)$ that satisfy
(i) $D(f,g) = \mathbb{E}_f(\varphi(f(X))) - \mathbb{E}_f(\varphi(g(X)))$ for some differentiable function φ, and
(ii) $D(f,g)$ is minimized over g by $g = f$.

■ **Example 28.1** *Details:* When there are finitely many outcomes (so that sums replace integrals in the definition of $D_{KL}(f,g)$) it may be shown that the form of φ must be logarithmic, i.e., φ must satisfy $\varphi(f(x)) = a + b\log f(x)$ for some a, b, with $b > 0$. See Konishi and Kitagawa (2008, Section 3.1). (Konishi, S. and Kitagawa, G. (2008) *Information Criteria and Statistical Modeling*, Springer.) □ ■

In addition to having the special difference-of-averages property in (28.10), the KL discrepancy takes a simple and intuitive form when applied to normal distributions.

Illustration: Two normal distributions Suppose $f(x)$ and $g(x)$ are the $N(\mu_1,\sigma^2)$ and $N(\mu_2,\sigma^2)$ pdfs. Then, from the formula for the normal pdf we have

$$\log\frac{f(x)}{g(x)} = -\frac{(x-\mu_1)^2-(x-\mu_2)^2}{2\sigma^2} = \frac{2x(\mu_1-\mu_2)-(\mu_1^2-\mu_2^2)}{2\sigma^2}$$

and substituting X for x and taking the expectation (using $\mathbb{E}_X(X) = \mu_1$), we get

$$D_{KL}(f,g) = \frac{2\mu_1^2 - 2\mu_1\mu_2 - \mu_1^2 + \mu_2^2}{2\sigma^2} = \left(\frac{\mu_1-\mu_2}{\sigma}\right)^2.$$

That is, $D_{KL}(f,g)$ is simply the squared standardized difference between the means. This is a highly intuitive notion of how far apart these two normal distributions are. □

The Kullback-Leibler discrepancy may be used to evaluate the association of two random vectors X and Y. We define the *mutual information* of X and Y as

$$I(X,Y) = D_{KL}(f_{(X,Y)}, f_X f_Y) = \mathbb{E}_{(X,Y)}\log\frac{f_{(X,Y)}(X,Y)}{f_X(X)f_Y(Y)}.$$

In other words, the mutual information between X and Y is the Kullback-Leibler discrepancy between their joint distribution and the distribution they would have if they were independent. In this sense, the mutual information measures how far a joint distribution is from independence.

28.4 Kullback-Leibler distance and Fisher information

Given the KL distance between two parametrized probability distributions

$$K\left(\theta,\theta'\right) = \int \log\left(\frac{p(x,\theta)}{p(x,\theta')}\right) p(x,\theta)dx$$

and the Fisher information

$$I(\theta) = \int \left\{\frac{\partial}{\partial\theta}\log p(x,\theta)\right\}^2 p(x,\theta)dx$$

prove that they are related in the following way:

$$\frac{d^2}{d\theta'^2}K\left(\theta,\theta'\right)|_{\theta'=\theta} = I(\theta)$$

Proof:

$$\begin{aligned}
\partial_{\theta'} K(\theta,\theta') &= \int -\frac{1}{p(x,\theta')} \partial_{\theta'} p(x,\theta') p(x,\theta) dx \\
\partial^2_{\theta'} K(\theta,\theta') &= \int \left(\frac{1}{p(x,\theta')^2} (\partial_{\theta'} p(x,\theta'))^2 - \frac{1}{p(x,\theta')} \partial^2_{\theta'} p(x,\theta') \right) p(x,\theta) dx \\
&\to_{\theta' \to \theta} \int (\partial_\theta \log(p(x,\theta)))^2 p(x,\theta) dx - \partial^2_\theta \int p(x,\theta) dx \\
&= \int (\partial_\theta \log(p(x,\theta)))^2 p(x,\theta) dx - 0
\end{aligned}$$

28.5 Kullback-Leibler

$$D_{\mathrm{KL}}(P \parallel Q) = \int_{-\infty}^{\infty} p(x) \log \left(\frac{p(x)}{q(x)} \right) dx$$

$$\begin{aligned}
D_{\mathrm{KL}}(P \parallel Q) &= -\sum_{x \in \mathcal{X}} p(x) \log q(x) + \sum_{x \in \mathcal{X}} p(x) \log p(x) \\
&= \mathrm{H}(P,Q) - \mathrm{H}(P)
\end{aligned}$$

where $\mathrm{H}(P,Q)$ is the cross entropy of P and Q and $H(P)$ is the entropy of P (which is the same as the cross-entropy of P with itself).

The Kullback–Leibler divergence is always non-negative, and it is zero (its minimum) if $P = Q$.

Caso gaussiano: Suppose that we have two multivariate normal distributions, with means μ_0, μ_1 and with (non-singular) covariance matrices Σ_0, Σ_1. If the two distributions have the same dimension, k, then the Kullback–Leibler divergence between the distributions is as follows:

Medida KL não é simétrica e por isto não é uma medida de distância, estrito-senso.

Qual a intuição por trás da falta de simetria? Na verdade, isto é bem razoável dada a definição de KL.

Considere duas gaussianas com a mesma esperança μ mas variâncias muito diferentes (1 e 100). Sejam $P \sim N(0,1)$ e $Q \sim N(0,100)$.

Temos $KL(P||Q) = \mathbb{E}_P[\log(P/Q)] = 1/2(1/100 + 0 - 1 + \log(100/1)) = 1.81$ mas $KL(Q||P) = 47.20$.

EM KL, assumimos que os dados vêm de uma das duas distribuições. Suponhamos que X $P = N(0,1)$. Então, todo e qualquer dado individual gerado a partir de P pode se passar facilmente como tendo sido gerado a partir de Q. As duas densidades possuem alturas diferentes no ponto aleatório X mas a razão $P(X)/Q(X)$ entre as duas não é absurdamente diferente. $X \sim P$ é verossímil sob Q também. Só iríamos distinguir que os dados vem de $P \sim N(0,1)$ (e não de $Q \sim N(0,100)$) a partir de uma grande amostra de P. Para um dado individual, não conseguimos discriminar entre P e Q *caso* $X \sim P$. Entretanto, caso $X \sim Q$, muitas vezes veremos valores X tao afastados da esperança 0 que não poderemos pensar que o dado tenha vindo de $P = N(0,1)$. Neste caso, quando $X \sim Q = N(0,100)$, teremos $P(X)$ tão próxima de zero que a razão $Q(X)/P(X)$ vai explodir. Assim, se $X \sim Q = N(0,100)$, a discrepância entre a altura aleatória $P(X)$ e $Q(X)$ será mais evidente.

28.5.1 KL e MLE

Ao encontrar o MLE estamos minimizando a distância KL entre o modelo e a distribuição real (e desconhecida) que gera os dados.

Seja $\ell(\theta) = \sum_i \log(f(X_i;\theta)$ a log-verossmilhança de uma amostra i.i.d. de v.a.'s. Suponha que θ^* é o verdadeiro valor do parâmetro θ. Defina

$$M_n(\theta) = \frac{1}{n}(\ell(\theta) - \ell(\theta^*)) = \frac{1}{n}\sum_i \log\left(\frac{f(X_i;\theta)}{f(X_i;\theta^*)}\right)$$

Esta é uma média aritmética de v.a.'s i.i.d. Pela L.G.N., ela converge quase certamente para seu valor esperado. Isto é,

$$M_n(\theta) \overset{q.c.}{\to} \mathbb{E}_{\theta^*}\left[\log\frac{f(X;\theta)}{f(X;\theta^*)}\right] = \int_{\mathscr{X}}\left[\log\frac{f(x;\theta)}{f(x;\theta^*)}\right] f(x;\theta)\,dx = KL(f(x;\theta)||f(x;\theta^*))$$

28.6 Wasserstein Distance

It has been proved that Wasserstein distance and Kantorovich distance are equivalent (the same). Other names are Monge and Rubinstein.

28.6.1 Univariate Case: $x \in \mathbb{R}$

We define the Kantorovich metric as

$$d_{Ka}(\mathbb{F},\mathbb{G}) = \int_{-\infty}^{+\infty} |\mathbb{F}(x) - \mathbb{G}(x)|dx \tag{28.2}$$

The two plots in the first column in Figure 28.6 shows what is this measure. It simply evaluates the total area between the two cumulative distribution functions $\mathbb{F}$ and $\mathbb{G}$ as x varies in the straight line. As the expression (28.2) shows, for each x in the real line, we calculate the absolute difference $|\mathbb{F}(x) - \mathbb{G}(x)|$ between the two curves, and integrate it out on the real line.

Chapters/C27-ModelSelection/Figuras/Kantorovich01Graphs

Figure 28.6: Kantorovich distance measure between the cumulative distribution functions $\mathbb{F}$ and $\mathbb{G}$ on the real line. The metric is the area between the two curves and it can be calculated integrating over the horizontal axis or over the vertical axis. See the text for details.

Another way to measure the same area is to vary the vertical scale $t \in (0,1)$. The plots in the second column in Figure 28.6 shows this other way to calculate the area between the curves $\mathbb{F}$ and $\mathbb{G}$. Rather than varying x on the horizontal axis and integrating $|\mathbb{F}(x) - \mathbb{G}(x)|$, the same area can be obtained by finding the difference between the two curves at an arbitrary vertical level t. Considering the top plot, for an arbitrary $t \in (0,1)$, we find the values $\mathbb{F}^{-1}(t)$ and $\mathbb{G}^{-1}(t)$ and the distance between them (that is, find $|\mathbb{F}^{-1}(t) - \mathbb{G}^{-1}(t)|$). After this, we integrate out over $(0,1)$ in the vertical scale. This gives the same area between the curves as (28.2). Therefore, the Kantorovich metric can be calculated in either way:

$$\begin{aligned} d_{Ka}(\mathbb{F},\mathbb{G}) &= \int_{-\infty}^{+\infty} |\mathbb{F}(x) - \mathbb{G}(x)|dx \\ &= \int_0^1 |\mathbb{F}^{-1}(t) - \mathbb{G}^{-1}(t)|dx \end{aligned} \tag{28.3}$$

Even for discrete random variables in which the cumulative distribution function has no inverse, we can define a quantile function in substitution to the inverse of the cumulative functions and the equivalence between the definitions holds valid.

28.6.2 Multivariate Case: $\mathbf{x} \in \mathbb{R}^n$

When x is not on the real line, the definitions in (28.2) and (28.3) can not be used. Assume that $\mathbf{x} \in \mathbb{R}^n$ and let $h(\mathbf{x})$ be a real-valued function. We will require that h satisfies the Lipschitz condition, a concept we will explain next.

We will need to consider functions that do not increase too fast as we vary $\mathbf{x}$ slightly. We will not require that $h(\mathbf{x})$ be a bounded function. It can increase without bounds but we want that it does not do it too quickly. Consider initially $x \in \mathbb{R}$ to simplify the discussion. We will return soon to $\mathbf{x} \in \mathbb{R}^n$. We start assuming that h is smooth and differentiable at every point x with a derivative $h'(x)$. If the absolute value of the derivative is bounded by a certain value, the function will not be able to explode upward or downward too quickly. That is, in this initial attempt, we would require that $|h'(x)| < M < \infty$ for all x where M is some arbitrary positive constant. What this means is that the tangent line at any point x must have a slope (in absolute value) at most equal to M. Consider the left hand plot in Figure 28.7, extracted from vonjd answer in `https://math.stackexchange.com/questions/353276/intuitive-idea-of-the-lipschitz-function`. At $(x, h(x))$, you draw two straight lines with slopes M and $-M$. We require that the derivative at x is smaller that M and hence, the tangent line at x is outside the white cone. This is valid, with the same M, for all x.

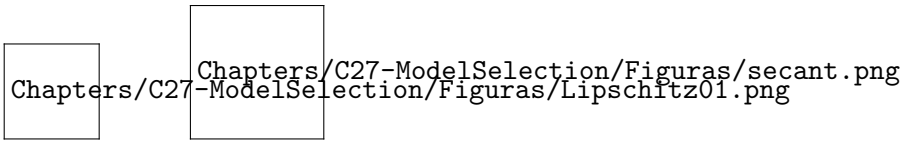


Figure 28.7: Lipschitz

The Lipschitz condition includes this restriction on $|h'(x)|$ when the derivative exists. However, it is more general than this and the function may not even have a derivative. The The real-valued function $h(\mathbf{x})$ satisfies the Lipschitz condition with constant $M > 0$ if

$$|h(x) - h(y)| \leq M|x - y| \tag{28.4}$$

for all pair of points x and y with $x \neq y$. If we take the $|x - y|$ on the right hand side of (28.4) and pass it to the right hand side, we obtain $|h(x) - h(y)|/|x - y|$. This is the slope of the secant line connecting the points $(x, h(x))$ and $(y, h(y))$. The right hand side plot in Figure 28.7 shows a function f and the secant determined by the points a and b.

What the Lipschitz condition with constant M implies for the entire function? Returning to the first hand side plot in Figure 28.7, at each point $(x, h(x))$ draw the lines with slopes M and $-M$. The Lipschitz condition ensures that the function $h(y)$ at all other points $x \neq y$ always remains entirely outside the white cone. This does not make any reference to derivatives, only to the values of the function h itself.

Of course, there is some connection between Lipschitz and derivatives. Indeed, if h is everywhere differentiable, it is Lipschitz if and only if it has bounded first derivative and $M = \sup |h'(x)|$.

Returning to the general case in which $\mathbf{x} \in \mathbb{R}^n$, we say that h is Lipschitz with constant M if

$$|h(\mathbf{x}) - h(\mathbf{y})| \leq M||\mathbf{x} - \mathbf{y}|| \tag{28.5}$$

for all pair of points $\mathbf{x}$ and $\mathbf{y}$. In the right hand side in (28.5), we use the distance between the vectors in $\mathbb{R}^n$ while in the left hand side we are using only the absolute value of the real number $h(\mathbf{x}) - h(\mathbf{y})$.

The Lipschitz condition is important in mathematics and appears in many places. Its initial use was in ordinary differential equations. A Lipschitz function with $M < 1$ has a unique fixed point. This is used to prove the existence of solutions to differential and integral equations. In analysis, Lipschitz functions are uniformly continuous and are differentiable almost everywhere. (REFS??)

We define

$$d_{Ka}(\mathbb{F},\mathbb{G}) = \sup\left\{\left|\int h(\mathbf{x})f(\mathbf{x})d\mathbf{x} - \int h(\mathbf{x})g(\mathbf{x})d\mathbf{x}\right| : ||h||_L \leq 1\right\} \tag{28.6}$$

28.7 Kullback-Leibler and Fisher Information

Let the set $\mathscr{P}$ of probability distributions be parametrized by a k-dimensional vector $\theta = (\theta_1,\ldots,\theta_k)$. We write $f(\mathbf{x},\theta)$ for the density or probability function parametrized by θ. There is a beautiful connection between the Kullback-Leibler distance between two distributions $f(\mathbf{x},\theta)$ and $f(\mathbf{x},\theta+\Delta)$. If $\theta+\Delta$ and θ are sufficiently close, we can use the Taylor expansion to write

$$KL(f(\mathbf{x},\theta+\Delta,f(\mathbf{x},\theta)) = \int f(\mathbf{x},\theta+\Delta)\log\left(\frac{f(\mathbf{x},\theta+\Delta)}{f(\mathbf{x},\theta)}\right)d\mathbf{x} \tag{28.7}$$

$$= \int f(\mathbf{x},\theta+\Delta)\left[\log f(\mathbf{x},\theta+\Delta) - \log f(\mathbf{x},\theta)\right]d\mathbf{x} \tag{28.8}$$

$$\approx \int f(\mathbf{x},\theta+\Delta)\left[\nabla\log(f(\mathbf{x},\theta))'\Delta + \frac{1}{2}\Delta'\left[\frac{\partial^2\log(f(\mathbf{x},\theta))}{\theta_i\theta_j}\right]\Delta\right] \tag{28.9}$$

28.7.1 The dependence of two random vectors may be quantified by mutual information.

It often happensmutual information that the deviation of one distribution from another must be evaluated. Consider two continuous pdfs $f(x)$ and $g(x)$, both being positive on (A,B). The *Kullback-Leibler (KL) discrepancy*Kullback-Leibler (KL) discrepancy is the quantity

$$D_{KL}(f,g) = \mathbb{E}_f\left(\log\frac{f(X)}{g(X)}\right)$$

where the subscript on the expectation $\mathbb{E}_f$ signifies that the random variable X has pdf $f(x)$. In other words, we have

$$D_{KL}(f,g) = \int_A^B f(x)\log\frac{f(x)}{g(x)}dx.$$

The KL discrepancy may also be defined, analogously, for discrete distributions. Note that $D_{KL}(f,g)$ may also be written in the difference form

$$D_{KL}(f,g) = \mathbb{E}_f\left(\log f(X)\right) - \mathbb{E}_f\left(\log g(X)\right). \tag{28.10}$$

In fact, the KL discrepancy is essentially unique among all discrepancies $D(f,g)$ that satisfy
(i) $D(f,g) = \mathbb{E}_f(\varphi(f(X))) - \mathbb{E}_f(\varphi(g(X)))$ for some differentiable function φ, and
(ii) $D(f,g)$ is minimized over g by $g = f$.

■ **Example 28.2** *Details:* When there are finitely many outcomes (so that sums replace integrals in the definition of $D_{KL}(f,g)$) it may be shown that the form of φ must be logarithmic, i.e., φ must satisfy $\varphi(f(x)) = a + b\log f(x)$ for some a,b, with $b > 0$. See Konishi and Kitagawa (2008, Section 3.1). (Konishi, S. and Kitagawa, G. (2008) *Information Criteria and Statistical Modeling*, Springer.) □ ■

In addition to having the special difference-of-averages property in (28.10), the KL discrepancy takes a simple and intuitive form when applied to normal distributions.

Illustration: Two normal distributions Suppose $f(x)$ and $g(x)$ are the $N(\mu_1,\sigma^2)$ and $N(\mu_2,\sigma^2)$ pdfs. Then, from the formula for the normal pdf we have

$$\log\frac{f(x)}{g(x)} = -\frac{(x-\mu_1)^2-(x-\mu_2)^2}{2\sigma^2} = \frac{2x(\mu_1-\mu_2)-(\mu_1^2-\mu_2^2)}{2\sigma^2}$$

and substituting X for x and taking the expectation (using $\mathbb{E}_X(X)=\mu_1$), we get

$$D_{KL}(f,g) = \frac{2\mu_1^2-2\mu_1\mu_2-\mu_1^2+\mu_2^2}{2\sigma^2} = \left(\frac{\mu_1-\mu_2}{\sigma}\right)^2.$$

That is, $D_{KL}(f,g)$ is simply the squared standardized difference between the means. This is a highly intuitive notion of how far apart these two normal distributions are. □

■ **Example 28.3 Auditory-dependent vocal recovery in zebra finches** Auditory-dependent vocal recovery in zebra finches Song learning among zebra finches has been heavily studied. When microlesions are made in the HVC region of an adult finch brain, songs become destabilized but the bird will recover its song within about 1 week. Thompson *et al.* (2007) ablated the output nucleus (LMAN) of the anterior forebrain pathway of zebra finches in order to investigate its role in song recovery. (Thompson, J.A., Wu, W., Bertram, R., and Johnson, F. (2007) Auditory-dependent vocal recovery in adult male zebra finches is facilitated by lesion of a forebrain pathway that includes basal ganglia, *J. Neurosci.*, 27: 12308–12320.) They recorded songs before and after the surgery. The multiple bouts of songs, across 24 hours, were represented as individual notes having a particular spectral composition and duration. The distribution of these notes post-surgery was then compared to the distribution pre-surgery. In one of their analyses, for instance, the authors compared the distribution of pitch and duration before and after surgery. Their method of comparison was to compute the KL discrepancy. Thompson *et al.* found that deafening following song disruption produced a large KL discrepancy whereas LMAN ablation did not. This indicated that the anterior forebrain pathway is not the neural locus of the learning mechanism that uses auditory feedback to guide song recovery. □ ■

The Kullback-Leibler discrepancy may be used to evaluate the association of two random vectors X and Y. We define the *mutual information* of X and Y as

$$I(X,Y) = D_{KL}(f_{(X,Y)}, f_X f_Y) = \mathbb{E}_{(X,Y)}\log\frac{f_{(X,Y)}(X,Y)}{f_X(X)f_Y(Y)}.$$

In other words, the mutual information between X and Y is the Kullback-Leibler discrepancy between their joint distribution and the distribution they would have if they were independent. In this sense, the mutual information measures how far a joint distribution is from independence.

Illustration: Bivariate normal If X and Y are bivariate normal with correlation ρ some calculation following application of the definition of mutual information gives

$$I(X,Y) = -\frac{1}{2}\log(1-\rho^2). \tag{28.11}$$

Thus, when X and Y are independent, $I(X,Y)=0$ and as they become highly correlated (or negatively correlated) $I(X,Y)$ increases indefinitely. □

Theorem For random variables X and Y that are either discrete or jointly continuous having a positive joint pdf, mutual information satisfies (i) $I(X,Y)=I(Y,X)$, (ii) $I(X,Y)\geq 0$, (iii) $I(X,Y)=0$ if and only if X and Y are independent, and (iv) for any one-to-one continuous transformations $f(x)$ and $g(y)$, $I(X,Y)=I(f(X),g(Y))$.

Proof: Omitted. See, e.g, Cover and Thomas (1991). (Cover, T.M. and Thomas, J.Y. (1991) *Elements of Information Theory*, New York: Wiley.) □

Property (iv) makes mutual information quite different from correlation.mutual infomative versus correlation We noted that correlation is a measure of *linear* association and, as we saw in the illustration on page **??**, it is possible to have $Cor(X,X^2) = 0$. In contrast, by property (iv), we may consider mutual information to be a measure of more general forms of association, and for the continuous illustration on page **??** we would have $I(X,X^2) = \infty$.

The use here of the word "information" is important. For emphasis we say, in somewhat imprecise terms, what we think is meant by this word.

> Roughly speaking, informationinformation about a random variable Y is associated with the random variable X if the uncertaintyuncertainty in Y is larger than the uncertainty in $Y|X$.

For example, we might interpret "uncertainty" in terms of variance. If the regression of Y on X is linear, as in (**??**) (which it is if (X,Y) is bivariate normal), we have

$$\sigma^2_{Y|X} = (1-\rho^2)\sigma^2_Y. \tag{28.12}$$

In this case, information about Y is associated with X whenever $|\rho| > 0$ so that $1-\rho^2 < 1$. A slight modification of (28.12) will help us connect it more strongly with mutual information. First, if we redefine "uncertainty" to be standard deviation rather than variance, (28.12) becomes

$$\sigma_{Y|X} = \sqrt{1-\rho^2}\sigma_Y. \tag{28.13}$$

Like Equation (28.12), Equation (28.13) describes a multiplicative (proportional) decrease in uncertainty in Y associated with X. An alternative is to redefine "uncertainty," and rewrite (28.13) in an *additive* form, so that the uncertainty in $Y|X$ is obtained by *subtracting* an appropriate quantity from the uncertainty in Y. To obtain an additive form we define "uncertainty" as the log standard deviation. Assuming $|\rho| < 1$, $\log\sqrt{1-\rho^2}$ is negative and, using $\log\sqrt{1-\rho^2} = \frac{1}{2}\log(1-\rho^2)$, we get

$$\log\sigma_{Y|X} = \log\sigma_Y - \left(-\frac{1}{2}\log(1-\rho^2)\right). \tag{28.14}$$

In words, Equation (28.14) says that $-\frac{1}{2}\log(1-\rho^2)$ is the amount of information associated with X in reducing the uncertainty in Y to that of $Y|X$. If (X,Y) is bivariate normal then, according to (28.11), this amount of information associated with X is the mutual information.

Formula (28.14) may be generalized by quantifying "uncertainty" in terms of *entropy*,entropy which leads to a popular interpretation of mutual information.

Details: We say that the *entropy*entropy of a discrete random variable X is

$$H(X) = -\sum_x f_X(x)\log f_X(x) \tag{28.15}$$

We may also call this the entropy of the distribution of X. In the continuous case the sum is replaced by an integral (though there it is defined only up to a multiplicative constant, and is often called *differential entropy*). The entropy of a distribution was formalized analogously to Gibbs entropy in statistical mechanics by Claude ShannonShannon, Claude in his development of communication theory. As in statistical mechanics, the entropy may be considered a measure of disorder in a distribution. For example, the distribution over a set of values $\{x_1,x_2,\ldots,x_m\}$ having maximal entropy is the uniform distribution (giving equal probability $\frac{1}{m}$ to each value) and, roughly speaking, as a distribution becomes concentrated near a point its entropy decreases.

For ease of interpretation the base of the logarithm is often taken to be 2 so that, in the discrete case,

$$H(X) = -\sum_x f_X(x) \log_2 f_X(x).$$

Suppose there are finitely many possible values of X, say $x_1, \ldots, x_m$, and someone picks one of these values with probabilities given by $f(x_i)$, then we try to guess which value has been picked by asking "yes" or "no" questions (e.g., "Is it greater than x_3?"). In this case the entropy (using $\log_2$, as above) may be interpreted as the minimum average number of yes/no questions that must be asked in order to determine the number, the average being taken over replications of the game.

Entropy may be used to characterize many important probability distributions. The distribution on the set of integers $0, 1, 2, \ldots, n$ that maximizes entropy subject to having mean μ is the binomial. The distribution on the set of all non-negative integers that maximizes entropy subject to having mean μ is the Poisson. In the continuous case, the distribution on the interval $(0, 1)$ having maximal entropy is the uniform distribution. The distribution on the positive real line that maximizes entropy subject to having mean μ is the exponential. The distribution on the positive real line that maximizes entropy subject to having mean μ and variance σ^2 is the gamma. The distribution on the whole real line that maximizes entropy subject to having mean μ and variance σ^2 is the normal.

Now, if Y is another discrete random variable then the entropy in the conditional distribution of $Y|X = x$ may be written

$$H(Y|X = x) = -\sum_y f_{Y|X}(y|x) \log f_{Y|X}(y|x)$$

and if we average this quantity over X, by taking its expectation with respect to $f_X(x)$, we get what is called the *conditional entropy* of Y given X:

$$H(Y|X) = \sum_x \left(-\sum_y f_{Y|X}(y|x) \log f_{Y|X}(y|x) \right) f_X(x).$$

Algebraic manipulation then shows that the mutual information may be written

$$I(X,Y) = H(Y) - H(Y|X).$$

This says that the mutual information is the average amount (over X) by which the entropy of Y decreases given the additional information $X = x$. In the discrete case, working directly from the definition we find that entropy is always non-negative and, furthermore, $H(Y|Y) = 0$. The expression for the mutual information, above, therefore also shows that in the discrete case $I(Y,Y) = H(Y)$. (In the continuous case we get $I(Y,Y) = \infty$.) For an extensive discussion of entropy, mutual information, and communication theory see Cover and Thomas (1991) or MacKay (2003). (Mackay, D. (2003) *Information Theory, Inference, and Learning Algorithms,* Cambridge.)

Mutual information has been used extensively to quantify the information about a stochastic stimulus (Y) associated with a neural response (X). In that context the notation is often changed by setting $Y = S$ for "stimulus" and $X = R$ for neural "response," and the idea is to determine the amount of information about the stimulus that is associated with the neural response.

■ **Example 28.4** In an influential paper, Optican and Richmond (1987) reported responses of single neurons in inferotemporal (IT) cortexinferotemporal cortex of monkeys while the subjects were shown various checkerboard-style grating patterns as visual stimuli. (Optican, L.M. and Richmond, B.J. (1987) Temporal encoding of two-dimensional patterns by single units in primate inferior temporal cortex. III. Information theoretic analysis. *J. Neurophysiol.*, 57: 162–178.) Optican and Richmond computed the mutual information between the 64 randomly-chosen stimuli (the random variable Y here taking 64 equally-likely values) and the neural response (X), represented by firing

rates across multiple time bins. They compared this with the mutual information between the stimuli and a single firing rate across a large time interval and concluded that there was considerably more mutual information in the time-varying signal. Put differently, more information about the stimulus was carried by the time-varying signal than by the overall spike count. □ ∎

In both Example 28.3 and Example **??** the calculations were based on pdfs that were *estimated* from the data. We discuss probability *density estimation* in Chapter **??**.

28.8 Kullback-Leibler distance and Fisher information

Given the KL distance between two parametrized probability distributions

$$K\left(\theta,\theta'\right) = \int \log\left(\frac{p(x,\theta)}{p(x,\theta')}\right) p(x,\theta)dx$$

and the Fisher information

$$I(\theta) = \int \left\{\frac{\partial}{\partial\theta}\log p(x,\theta)\right\}^2 p(x,\theta)dx$$

prove that they are related in the following way:

$$\frac{d^2}{d\theta'^2}K\left(\theta,\theta'\right)|_{\theta'=\theta} = I(\theta)$$

Proof:

$$\begin{aligned}
\partial_{\theta'}K(\theta,\theta') &= \int -\frac{1}{p(x,\theta')}\partial_{\theta'}p(x,\theta')p(x,\theta)dx \\
\partial^2_{\theta'}K(\theta,\theta') &= \int \left(\frac{1}{p(x,\theta')^2}(\partial_{\theta'}p(x,\theta'))^2 - \frac{1}{p(x,\theta')}\partial^2_{\theta'}p(x,\theta')\right)p(x,\theta)dx \\
&\to_{\theta'\to\theta} \int (\partial_\theta \log(p(x,\theta)))^2 p(x,\theta)dx - \partial^2_\theta \int p(x,\theta)dx \\
&= \int (\partial_\theta \log(p(x,\theta)))^2 p(x,\theta)dx - 0
\end{aligned}$$

28.9 AIC

AIC makes assumptions that you: Are using the same data between models Are measuring the same outcome variable between models Have a sample of infinite size

As a rule of thumb, always using AICc is safest, but AICc should especially be used when the ratio of your data points (n) : # of parameters (k) is < 40.

Once the assumptions of AIC (or AICc) have been met, the biggest advantage of using AIC/AICc is that your models do not need to be nested for the analysis to be valid, unlike other single-number measurements of model fit like the likelihood-ratio test. A nested model is a model whose parameters are a subset of the parameters of another model. As a result, vastly different models can be compared mathematically with AIC.

Assume you have calculated the AICs for multiple models and you have a series of AIC scores $(AIC_1, AIC_2, \ldots AIC_n)$. For any given AIC_i, you can calculate the probability that the "i-th" model minimizes the information loss through the formula below, where AIC_{min} is the lowest AIC score in your series of scores.

$$P = exp((AIC_min - AIC_i)/2)$$

Pitfalls of AIC As a reminder, AIC only measures the ** relative ** quality of models. This means that all models tested could still fit poorly. As a result, other measures are necessary to show that your model's results are of an acceptable absolute standard (calculating the MAPE, for example).

Bibliography

Books

[Dun99] William Dunham. *Euler: The master of us all*. 22. Cambridge University Press, 1999 (cited on page 162).

Articles

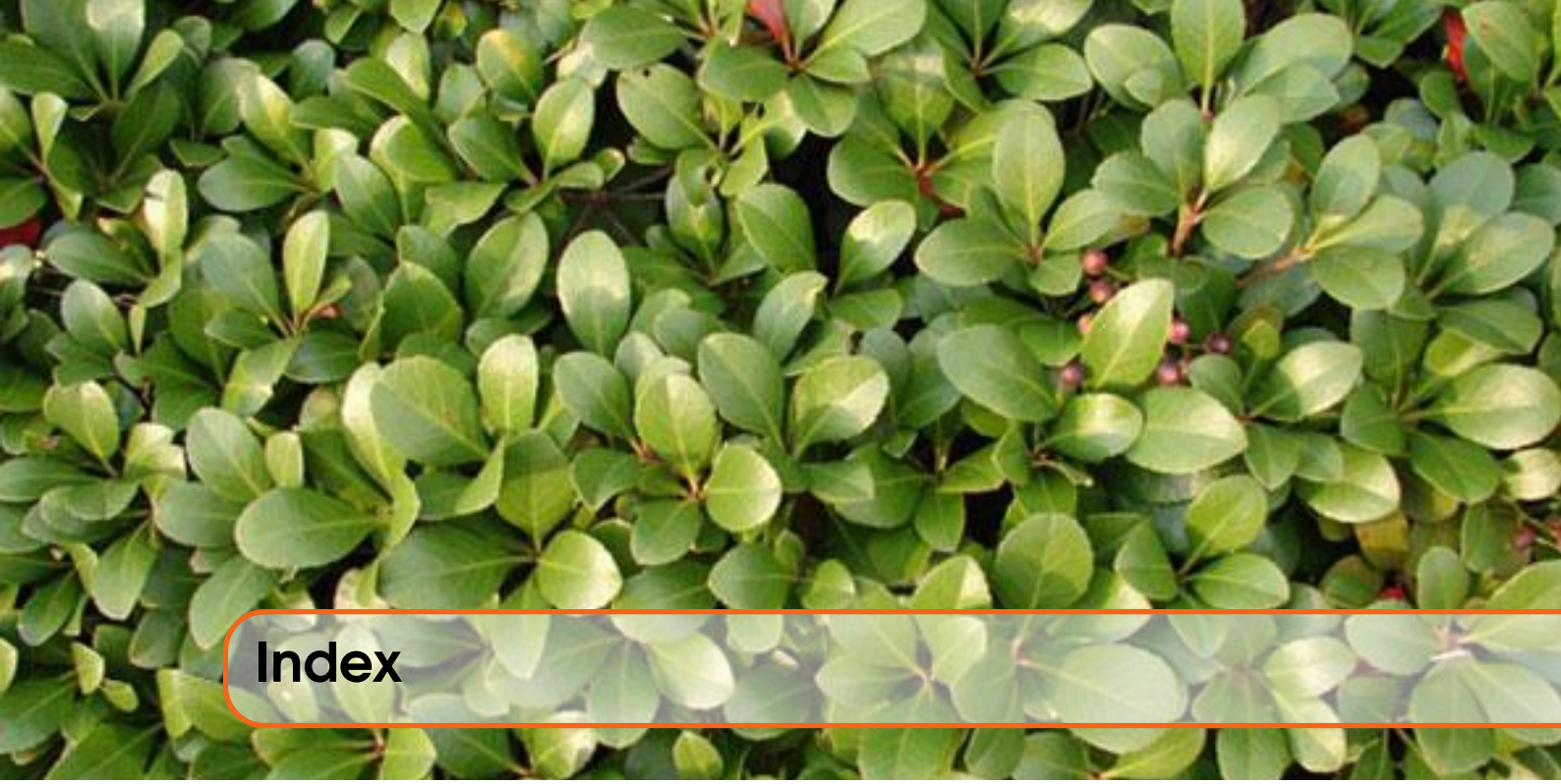
[Dar35] Georges Darmois. “Sur les lois de probabilitéa estimation exhaustive”. In: *CR Acad. Sci. Paris* 260.1265 (1935), page 85 (cited on page 154).

[Fis22] Ronald Aylmer Fisher. “On the mathematical foundations of theoretical statistics”. In: *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character* 222 (1922), pages 309–368 (cited on pages 40, 141).

[Fis25] Ronald Aylmer Fisher. “Theory of statistical estimation”. In: 22.05 (1925), pages 700–725 (cited on page 40).

[Fis34] Ronald Aylmer Fisher. “Two new properties of mathematical likelihood”. In: *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character* 144.852 (1934), pages 285–307 (cited on page 154).

[Koo36] Bernard Osgood Koopman. “On distributions admitting a sufficient statistic”. In: *Transactions of the American Mathematical society* 39.3 (1936), pages 399–409 (cited on page 154).



Index

ex, 205

gen, 199, 204, 206, 207