Inferência Estatística - Modelos e MLE

Renato Martins Assunção

UFMG

2013

Um modelo para os dados

- Suponha que $y_1, \ldots, y_n$ são os dados da amostra.
- São instâncias das v.a.'s $Y_1, \ldots, Y_n$.
- Precisamos assumir um modelo de probabilidade para a distribuição conjunta dessas v.a.'s
- Aprendemos até agora:
  - $Y_1, \ldots, Y_n$ são i.i.d.: Exemplo: todas são $N(\mu, \sigma^2)$.
  - $Y_1, \ldots, Y_n$ são independentes mas não são i.d. Exemplo: modelo de regressão linear ou logística
  - $Y_1, \ldots, Y_n$ não são independentes: Exemplo: só vimos um até agora: $\boldsymbol{Y} \sim N_n(\boldsymbol{\mu}, \Sigma)$
- Vamos assumir que a distribuição conjunta possa ser indexada por um vetor de parâmetros $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$.
- Isto leva ao conceito de modelo estatístico para uma amostra de dados.

Modelos Estatísticos para Amostras de Dados

- **Modelo estatístico:** um conjunto de hipóteses que se supõem válidas para a distribuição de probabilidades das variáveis aleatórias medidas na amostra.
- Estas hipóteses serão satisfeitas de forma *aproximada*.
- A qualidade de um modelo para descrever a distribuição de uma população será dada pelo grau de aproximação que tenham as conseqüências teóricas do modelo com a distribuição real, como observada nos dados.
- Um modelo contínuo (distribuição normal) pode ser usado para a distribuição de variáveis discretas.
- Por exemplo, para a colheita num conjunto de 5000 lotes de uma fazenda, a distribuição de Y, a colheita no lote, é normal con média μ e variância σ^2.

Modelos paramétricos

- A distribuição $F(y)$ de cada variável Y medida na população pertence a uma família de distribuição $F(y)$ que depende de um número finito de parâmetros reais. Por exemplo:
  - $F(y)$ pertence à família $N(\mu, \sigma^2)$.
  - $(Y|x)$ tem $F(y)$ na família $N(\beta_0 + \beta_1 x, \sigma^2)$.
  - $F(y)$ pertence à família $\text{Bin}(n, \theta)$
  - $(Y|x)$ tem $F(y)$ na família $\text{Bin}(n, 1/(1 + \exp(\beta_0 + \beta_1 x)))$
  - $F(y)$ pertence à família $\text{Poisson}(\lambda)$.
  - $(Y|x)$ tem $F(y)$ na família $\text{Poisson}(e^{\beta_0 + \beta_1 x})$.

Modelos paramétricos

- Em geral, um modelo paramétrico para a distribuição $F(y)$ de uma única variável Y terá a seguinte forma:
  - $F(y)$ é especificada para Y ou para $(Y|x)$ (condicional a x).
  - Vamos considerar apenas o caso mais geral $(Y|x)$. O caso para Y significa condicionar em nada.
  - $F(y|x)$ pertence à uma família $F(y|x, \boldsymbol{\theta})$
  - onde $\boldsymbol{\theta} = (\theta_1, ..., \theta_k)$ é o vetor de parâmetros
  - $\boldsymbol{\theta}$ toma valores em um conjunto $\Theta \subset \mathbb{R}^k$.
  - Θ é chamado de *espaço paramétrico*.

Modelos

- Isto significa que existe algum valor $\boldsymbol{\theta} \in \Theta$, digamos $\boldsymbol{\theta}_0$, tal que $F(y|x, \boldsymbol{\theta}_0)$ coincide com a verdadeira distribuição $F(y|x)$ dos dados.
- Não esperamos que exista uma coincidência perfeita
- Esperamos apenas que a verdadeira distribuição dos dados $F(y|x)$ e uma das distribuições da família escolhida, $F(y|x, \boldsymbol{\theta}_0)$, sejam parecidas.
- A similaridade deve ser tal que conclusões baseadas no modelo aproximado $F(y|x, \boldsymbol{\theta}_0)$ sejam aproximadamente verdadeiras para a distribuição real $F(y)$.

Exemplos de modelos para uma v.a.

- Y é a colheita agrícola num lote. Podemos assumir $Y \sim N(\mu, \sigma^2)$. Neste caso, $\boldsymbol{\theta} = (\mu, \sigma^2)$ e $\Theta = \mathbb{R} \times (0, \infty)$.
- Y é o tempo de espera até a ocorrência de um evento (a próxima view de um vídeo no YouTube). Um possível modelo é assumir $Y \sim \exp(\lambda)$ sendo que $\Theta = \{\lambda \in \mathbb{R}; \quad \lambda > 0\} = (0, \infty)$
- Y é binária significando SPAM versus NÃO-SPAM. $\mathbb{P}(Y = \text{SPAM})$ depende de features coletadas no vetor $\boldsymbol{x} = (1, x_1, \ldots, x_k)$ e relacionadas ao conteúdo da mensagem (palavras-chave no subject ou no corpo da msg), ao endereço IP de envio, e características do receptor.

  Podemos assumir então um modelo logístico:

  $$\mathbb{P}(Y = \text{SPAM}) = \frac{1}{1 + \exp(-\boldsymbol{x}'\,\boldsymbol{\theta})}$$

  onde $\boldsymbol{\theta} = (\beta_0, \beta_1, \ldots, \beta_k)$. Como não existe restrição nos β's, o espaço paramétrico é $\Theta = \{\boldsymbol{\theta} \in \mathbb{R}^{k+1}\}$.

Logística para SPAM

Partitioned Logistic Regression for Spam Filtering

Ming-wei Chang
University of Illinois
201 N Goodwin Ave
Urbana, IL, USA
mchang21@uiuc.edu

Wen-tau Yih
Microsoft Research
One Microsoft Way
Redmond, WA, USA
scottyih@microsoft.com

Christopher Meek
Microsoft Research
One Microsoft Way
Redmond, WA, USA
meek@microsoft.com

Figura: Paper recente: KDD 2008 (Knowledge Discovery and Data Mining).

Exemplos de modelos para UMA v.a.

- Y é contínua e sua distribuição depende de covariáveis no vetor $\boldsymbol{x} = (1, x_1, \ldots, x_k)$. Podemos assumir

$$(Y | \boldsymbol{x}) \sim N(\beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k, \sigma^2) = N(\boldsymbol{x}' \boldsymbol{\beta}, \sigma^2)$$

  onde $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^2) = (\beta_0, \beta_1, \ldots, \beta_k, \sigma^2)$.
  O valor de σ^2 deve ser maior que zero mas usualmente β_j não tem restrição.
  Assim, o espaço paramétrico é $\Theta = \{\boldsymbol{\theta} \in \mathbb{R}^{k+1} \times (0, \infty)\}$.
- Suponha que sejam medidas as asas direita (Y_1) e esquerda (Y_2) de um passáro escolhido dentro de certa região. Podemos usar como modelo para a distribuição $F(y_1, y_2)$ do vetor (Y_1, Y_2) a normal bivariada $N_2(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$ e
  $\Theta = \mathbb{R} \times \mathbb{R} \times (0, \infty) \times (0, \infty) \times (-1, 1)$

Definição geral de modelo estatístico

- **Definição:** Um *modelo estatístico paramétrico* para um fenêmeno aleatório gerando variáveis aleatórias $Y_1, ..., Y_n$ com covariáveis FIXAS (não-aleatórias) $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ é constituído de TRÊS elementos:
  - (i) Família $\mathcal{F}$ de distribuições do vetor aleatório $\boldsymbol{Y} = (Y_1, ..., Y_n)$ (possivelmente dependente de covariáveis $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$).
  - (ii) Conjunto $\mathcal{Y} \subset \mathbb{R}^n$ de todos os valores possíveis de $\boldsymbol{Y}$.
  - (iii) Conjunto $\Theta \subseteq \mathbb{R}^k$ chamado *espaço paramétrico* tal que existe bijeção entre $\mathcal{F}$ e Θ. Isto é, Θ é índice de $\mathcal{F}$ de forma que, cada elemento $\boldsymbol{\theta} \in \Theta$ está associado com uma única distribuição em $\mathcal{F}$, e vice-versa.
- Queremos inferir sobre uma função $q(\boldsymbol{\theta})$. A distribuição $\mathcal{F} = \{F_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$ e a função $q(\boldsymbol{\theta})$ dependem do problema.

Modelo de Sequências de Bernoulli i.i.d.

- $Y_1, \ldots, Y_n$: nascimentos sucessivos de não gêmeos numa maternidade.
-
$$Y_i = \begin{cases} 1, & \text{se } i\text{-ésimo nascimento é do sexo masculino} \\ 0, & \text{se se } i\text{-ésimo nascimento é do sexo feminino} \end{cases}$$
- Esta distribuição é especificada dando a probabilidade de sucesso.
- Vamos escrever $P(Y_i = 1) = \theta_i$ onde $\theta_i \in (0, 1)$.
- Assim,
$$P(Y_i = y) = \begin{cases} \theta_i, & \text{se } y = 1. \\ 1 - \theta_i, & \text{se } y = 0. \end{cases} \tag{1}$$

Modelo de Sequências de Bernoulli

- Precisamos especificar a distribuição *conjunta* das n variáveis aleatórias.
- Para isto precisamos responder:
  - as variáveis aleatórias são independentes? Não é possível *provar* isso matemáticamente. Considerando o problema, tudo leva a crer que essa é uma *suposição razoável* acerca desses dados.
  - É possível verificar se esta hipótese é realmente razoável num estágio posterior da análise mas nos momentos iniciais é apenas o conhecimento prévio, ou o puro chute bem informado, que guia o analista.
  - As v.a.'s possuem a MESMA probabilidade de sucesso θ?
- Se a resposta é SIM, o vetor $\boldsymbol{Y}$ é composto de variáveis aleatórias i.i.d

Um truque de notação

- Para $y = 0$ ou $y = 1$, podemos escrever (VERIFIQUE)

$$P(Y_i = y) = \theta^y (1-\theta)^{1-y}$$

- Com isto, se $y_i = 0$ ou 1, temos a conjunta:

$$f_\theta(\boldsymbol{y}) = f_\theta(y_1, ..., y_n) = \prod_{i=1}^{n} \theta^{y_i}(1-\theta)^{1-y_i} = \theta^{\sum_i y_i}(1-\theta)^{n-\sum_i y_i} \quad (2)$$

- Note θ na notação $f_\theta(\boldsymbol{y})$ da densidade. Às vezes: $f(\boldsymbol{y}\,|\theta)$.
- Modelo estatístico: com $y_i = 0$ ou $y_i = 1$, para $i = 1, \ldots, n$, e $\theta \in (0,1)$, temos

$$\mathcal{F}_\theta = \left\{ f_\theta(\boldsymbol{y}) = \theta^{\sum_i y_i}(1-\theta)^{n-\sum_i y_i} \right\}$$

Modelo de erro de medição

- n observações são feitas com erro em um mesmo objeto.
- Represente como v.a.'s $y_i = \mu + \varepsilon_i$
- μ é uma constante desconhecida (o valor real do objeto medido)
- $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ são i.i.d. com distribuição $N(0, \sigma^2)$.
- Note que este modelo implica em supor que:
  - A distribuição dos erros não depende de μ.
  - Um erro em uma medição não afeta o erro em outra medição.
  - O objeto não se altera e dessa forma o valor de μ é constante ao longo das medições.
  - A distribuição do erro é a mesma ao longo de todas as medições.
  - A distribuição dos erros é simétrica em torno de zero e contínua.

Modelo de erro de medição

- $Y_1, .., Y_n$ são i.i.d com $Y_i \sim N(\mu, \sigma^2)$. Seja $\boldsymbol{\theta} = (\mu, \sigma^2)$.
- A distribuição conjunta de $\boldsymbol{y}$ é um membro da família $\mathcal{F} = \{f(\boldsymbol{y} \,|\, \boldsymbol{\theta}) \;;\; \boldsymbol{\theta} \in \Theta\}$ onde conjunta é

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

para $y_i \in \mathbb{R}$. O espaço paramétrico é

$$\Theta = \left\{\boldsymbol{\theta} = (\mu, \sigma^2) \in \mathbb{R}^2 \;:\; \mu \in \mathbb{R}, \sigma^2 \in (0, \infty)\right\}$$

Questões de interesse

- Questões de interesse num modelo de erro de medição:
- Conhecer o tamanho típico do erro de medição.
- Assim, queremos saber o valor de $q(\boldsymbol{\theta}) = \sqrt{\sigma^2} = \sigma$
- Ou então: conhecer o tamanho do erro de medição relativamente ao tamanho do objeto sendo medido.
- Isto é, conhecer o coeficiente de variação $q(\boldsymbol{\theta}) = \sigma/|\mu|$ (que faz sentido apenas se $\mu \neq 0$).

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.
- Modelo natural é a Multinomial $F(y_1, y_2, y_3)$ que pertence à família $M(\theta_1, \theta_2, \theta_3, n)$.
- θ_j é proporção de indivíduos na *população* que estão na categoria j.
- Temos

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

- Uma representação de Θ está no próximo slide.

Espaço paramétrico do modelo multinomial

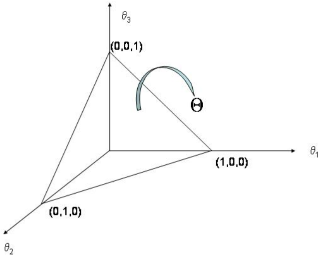


Figura: Espaço paramétrico Θ de um vetor $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3)$ de uma multinomial com três categorias. Θ é representado pelo triângulo de vértices $(1,0,0)$, $(0,1,0)$ e $(0,0,1)$.

Explicar bonus-malus em seguro de automóveis

- Sistema que ajusta o prêmio de acordo com a história individual do cliente.
- Tipicamente: entra na seguradora num nível de referência para a sua categoria de idade-sexo-etc
- A partir daí, em cada renovação anual do contrato:
  - Se teve pelo menos um sinistro no ano anterior, sofre incremento de prêmio (malus)
  - Quanto mais sinistros, maior o incremento.
  - se não teve sinistros, tem redução do prêmio (bonus).
- Precisa fazer cálculos fazer estabelecer níveis de bonus e malus.

Regras de Transição em bonus-malus

- Imagine um sistema com cinco classes.
- Segurado entra na CLASSE 3.
- Se passar para a classe 4, seu prêmio aumenta em 130% em relação ao prêmio da classe 3.
- Se passar para a classe 5, aumenta em 160%
- Se passar para a classe 2, o prêmio diminui para 80% do prêmio de referência (classe 3)

Relatividades	Classe	Classe após k sinistros			
		$k=0$	$k=1$	$k=2$	$k \geq 3$
160%	5	4	5	5	5
130%	4	3	5	5	5
100%	3	2	3	4	5
80%	2	1	3	4	5
70%	1	1	3	5	5

Matriz de Probabilidades de Transição

- Coleta-se amostra de 1000 segurados em cada classe (1 a 5).
- Dentro de cada classe, contam-se quantos indivíduos terminaram na classe K.
- Deseja-se saber quais são as probabilidades de migrar da categoria atual para outra categoria.
- Isto é, deseja-se preencher a seguinte tabela:

Classe	Nova Classe				
Atual	5	4	3	2	1
5	$*$	$*$	0	0	0
4	$*$	0	$*$	0	0
3	$*$	$*$	$*$	$*$	0
2	$*$	$*$	$*$	0	$*$
1	$*$	0	$*$	0	$*$

Matriz de Probabilidades de Transição

- Isto é, temos CINCO vetores de dimensão 5 a serem estimados, um vetor para cada classe.
- $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3, \theta_4, \theta_5)$ com $\sum_i \theta_i = 1$.
- Por exemplo, para a classe 4, temos $\boldsymbol{\theta} = (0, 0, \theta_3, 0, \theta_5)$ com $\theta_3 + \theta_5 = 1$.

Classe Atual	Nova Classe 5	4	3	2	1
5	$*$	$*$	0	0	0
4	θ_5	0	θ_3	0	0
3	$*$	$*$	$*$	$*$	0
2	$*$	$*$	$*$	0	$*$
1	$*$	0	$*$	0	$*$

- O vetor para a classe 3 será diferente do vetor para a classe 4, etc.

Regressão Linear Simples

- $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ é composto por variáveis aleatórias independentes
- $Y_i \sim N(\mu_i, \sigma^2) = N(\beta_0 + \beta_1 x_i, \sigma^2)$
- onde $x_1, ..., x_n$ são números fixos e conhecidos.
- Não temos interesse em modelar a variação de x. Serve só para explicar em parte porquê Y_i varia.
- Eles variam porquê sua média μ_i muda (e a média muda apenas se x mudar) e porquê existe um erro aleatório.

Regressão Linear Simples

- Seja $\boldsymbol{\theta} = (\beta_0, \beta_1, \sigma^2)$. Conjunta de $\boldsymbol{y}$ é:
-

$$
\begin{aligned}
f_{\boldsymbol{\theta}}(\boldsymbol{y}) &= \prod_{i=1}^{n} f_{\boldsymbol{\theta}}(y_i) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \prod_{i=1}^{n} \exp\left(-\frac{1}{2}\left(\frac{y_i - \beta_0 - \beta_1 x_i}{\sigma}\right)^2\right) \\
&= \left(2\pi\sigma^2\right)^{-n/2} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{y_i - \beta_0 - \beta_1 x_i}{\sigma}\right)^2\right)
\end{aligned}
$$

- O espaço paramétrico é

$$
\Theta = \left\{\boldsymbol{\theta} = (\beta_0, \beta_1, \sigma^2) \in \mathbb{R}^3 \ : \ \beta_0 \in \mathbb{R}, \beta_1 \in \mathbb{R}, \sigma^2 \in (0, \infty)\right\}
$$

Questões de interesse em regressão

- Saber o valor de $q(\boldsymbol{\theta}) = \beta_1$.
- Algumas vezes, o interesse é apenas saber se β_1 é igual a zero ou não.
- Se $\beta_1 = 0$ então a covariável y não afeta $E(Y)$.
- Assim, o interesse reside em saber o valor da função $q(\boldsymbol{\theta})$ definida da seguinte forma:

$$q(\boldsymbol{\theta}) = \begin{cases} 1, & \text{se } \beta_1 = 0 \\ 0, & \text{caso contrário} \end{cases}$$

- Esta é uma maneira estranha de saber se $\beta_1 = 0$ ou não mas, acredite, ela será útil no contexto de testes de hipóteses.

Regressão Logística Simples

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

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

- $$p_i = p(x_i) = \frac{1}{1 + \exp(-\beta(x_i - \mu))}.$$
- Dependendo dos valores de μ e β nós obtemos diferentes curvas
- Interesse: estimar os valores de μ e β para traçar o gráfico de x versus $p(x)$.

Modelo estatístico

- Temos

$$P(Y = y_i) = p_i^{y_i}(1 - p_i)^{1-y_i} \; .$$

- Vamos escrever $\boldsymbol{\theta} = (\mu, \beta)$.
- indep implica conjunta:

$$\begin{aligned}
p_{\boldsymbol{\theta}}(\boldsymbol{y}) &= \prod_{i=1}^{n} p_i^{y_i} \, (1 - p)^{1-y_i} \\
&= \prod_{i=1}^{n} \left(\frac{1}{1 + e^{-\beta(x_i - \mu)}} \right)^{y_i} \left(\frac{e^{-\beta(x_i - \mu)}}{1 + e^{-\beta(x_i - \mu)}} \right)^{1-y_i} \\
&= \frac{\exp\left(-\beta \sum_i (x_i - \mu)(1 - y_i)\right)}{\prod_{i=1}^{n} (1 + e^{-\beta(x_i - \mu)})} \\
&= \frac{\exp\left(n\mu\beta(1 - \overline{y}) - n\beta(\overline{x} - \overline{xy})\right)}{\prod_{i=1}^{n} (1 + e^{-\beta(x_i - \mu)})}
\end{aligned}$$

onde $\overline{y} = \sum_i y_i / n$, $\overline{x} = \sum_i x_i / n$ e $\overline{xy} = \sum_i x_i y_i / n$.

Questão de interesse

- Interesse é descobrir a idade x_{90} tal que 90% das criançcas devem estar executando a tarefa.
- Isto é, queremos encontrar a idade x_{90} que satisfaz a equação $p(x_{90}) = 0.9$.
- Ou seja,
$$0.9 = \frac{1}{1 + \exp(-\beta(x_{90} - \mu))}$$
Manipulando algebraicamente, encontramos
$$x_{90} = \frac{-1}{\beta}\left(\log\left(\frac{0.1}{0.9}\right) + \beta\mu\right) .$$
- Assim, x_{90} é uma função $q(\boldsymbol{\theta})$.

Verossimilhança e MLE

- Suponha que a distribuição conjunta dos dados pertença a um modelo estatístico
- Este modelo será indexado por um vetor θ.
- Por exemplo, no caso da regressão logística, $\theta = (\beta_0, \beta_1)$.
- Usando o modelo estatístico $\mathcal{P}_\theta$, calcule o valor aproximado da probabilidade de observar os dados da amostra.

Verossimilhança e MLE

- Se as y_i's são discretas, calcule a probabilidade de observar os dados da amostra.
- Se as y_i's são contínuas, obtenha a densidade de probabilidade avaliada nos dados da amostra.
- Esta é a *função de verossimilhança* $L(\theta)$ onde apenas θ pode variar.
- Obtenha o valor $\hat{\theta}$ que maximiza $L(\theta)$.
- Este valor é a estimativa de máxima verossimilhança (maximum likelihood estimator, ou MLE).
- O MLE é o valor $\hat{\theta}$ de θ que é o mais verossímil tendo em vista os dados à mão.
- O MLE $\hat{\theta}$ é aquele em que, aproximadamente, é máxima a probabilidade de observar os dados realmente observados.

Função de verossimilhança

- $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ é composto por v.a.'s com função de probabilidade (caso discreto) ou densidade (caso contínuo) conjunta $p(\boldsymbol{y}, \theta)$.
- O parâmetro θ pertence ao conjunto Θ, chamado de espaço paramétrico.
- ASSUMA QUE $\Theta \subseteq \mathbb{R}$ (uni-dimensional).
- Considere $p(\boldsymbol{y}, \theta)$ como uma função de θ para $\boldsymbol{y}$ fixo.
- Nós chamamos esta função de *função de verossimilhança*
- NOTAÇÃO: $L(\theta)$.

EMV

- EMV $\widehat{\theta} = \widehat{\theta}(\boldsymbol{y}) \in \Theta$ é o valor mais verossímil em termos de gerar os dados $\boldsymbol{x}$.
- Isto é, se observamos $\boldsymbol{Y} = \boldsymbol{y}$, nós procuramos $\widehat{\theta}(\boldsymbol{y})$ que satisfaça

$$L(\widehat{\theta}(\boldsymbol{y})) = p(\boldsymbol{y}, \widehat{\theta}(\boldsymbol{x})) = \max_{\theta}\{p(\boldsymbol{x}, \theta) : \theta \in \Theta\} = \max_{\theta}\{L(\theta) : \theta \in \Theta\}$$

- O vetor $\boldsymbol{y}$ aparece na expressão de $L(\theta)$ mas ele é considerado fixo nas instâncias observadas na amostra.
- $\boldsymbol{y}$ significa o conjunto de valores realmente obtidos em um experimento, os valores realizados do vetor aleatório $\boldsymbol{Y}$.
- Se $\widehat{\theta}(\boldsymbol{y})$ é o EMV de θ, então estimamos qualquer função $q(\theta)$ por $q(\widehat{\theta}(\boldsymbol{x}))$.

Log-verossimilhança

- O mais comum é que a função de verossimilhança seja um produto de várias funções envolvendo θ.
- A derivada de produtos de funções é obtida aplicando-se a regra do produto e a equação de verossimilhança pode resultar numa expressão complicada.
- A derivada de somas de funções é geralmente muito mais simples.
- Nós definimos a *função de log-verossimilhança*, denotada por $\ell(\theta)$:

$$\ell(\theta) = \log L(\theta) = \log p(\boldsymbol{y}, \theta)$$

Log-verossimilhança

- Se $\widehat{\theta}$ maximiza $L(\theta) = p(\boldsymbol{y}, \theta)$ então $\widehat{\theta}$ também maximiza $\ell(\theta) = \log p(\boldsymbol{y}, \theta)$.
- Assim, a estimativa de máxima verossimilhança é obtida como a solução da equação

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

- Esta equação é chamada de *equação de log-verossimilhança* ou, de forma mais curta, simplesmente *equação de verossimilhança*.
- Se $\boldsymbol{\theta}$ é um vetor então a equação de verossimilhança é na verdade um sistema de equações, cada uma delas associada com uma derivada parcial. Veja os exemplos de MLE multivariado no próximo bloco de slides.

Experimento de Bernoulli

- Experimento de Bernoulli é realizado independentemente 10 vezes.
- θ a probabilidade de sucesso
- Espaço paramétrico $\Theta = [0, 1]$.
- Observa-se $\boldsymbol{y} = (0, 1, 0, 0, 1, 0, 1, 0, 1, 0)$ onde 1 indica S e 0 indica F.
- Função de verossimilhança de θ:

$$L(\theta) = \mathbb{P}(\boldsymbol{Y} = \boldsymbol{y}) = \mathbb{P}(\boldsymbol{Y} = (0, 1, 0, 0, 1, 0, 1, 0, 1, 0)) = \theta^4 (1-\theta)^6 \,.$$

- Gráfico mostra a função de verossimilhança $L(\theta)$ versus θ.

Experimento de Bernoulli

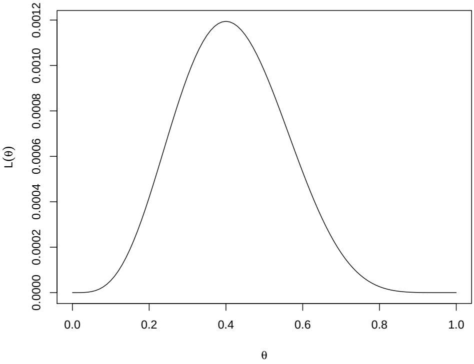


Figura: Função de verossimilhança $L(\theta) = \theta^4(1-\theta)^6$

Experimento de Bernoulli

- Função log-verossmilhança:

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

- A equação de verossimilhança é

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

- Solução $\widehat{\theta} = 0.4$.
- A partir do gráfico, já sabemos que esta solução é um máximo global.

Experimento de Bernoulli

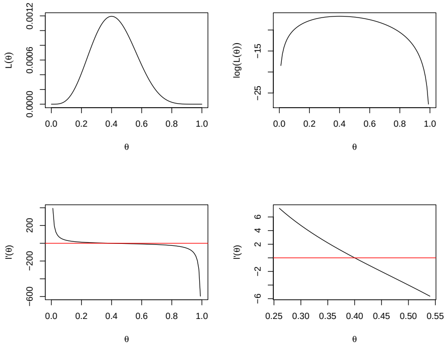


Figura: Função de verossimilhança $L(\theta)$, log-verossimilhança $\ell(\theta)$, Derivada da log-verossimilhança $\ell'(\theta)$ e restrição de $\ell'(\theta)$ no intervalo $\theta \in (0.25, 0.55)$

Bernoulli - Caso Geral

- Podemos obter uma fórmula geral a estimativa de máxima versossimilhança EM FUNÇÃO DO QUE SERÁ OBSERVADO na amostra.
- Seja $\boldsymbol{y} = (y_1, \ldots, y_{10})$ uma realização do experimento, uma lista de 1's e 0's
- Seja $k = \sum_{i=1}^{10} y_i$, o número de caras que ocorreram nesta particular realização do experimento.
- A probabilidade de ocorrer $\boldsymbol{y}$ é igual a

$$\mathbb{P}(\boldsymbol{Y} = \boldsymbol{y} \,|\theta) = \theta^k (1-\theta)^{10-k} = L(\theta)$$

Bernoulli - Caso Geral

- A probabilidade de ocorrer $\boldsymbol{y}$ é igual a

$$\mathbb{P}(\boldsymbol{Y} = \boldsymbol{y}\,|\theta) = \theta^k(1-\theta)^{10-k} = L(\theta)$$

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

$$\begin{aligned} \frac{d}{d\theta}\log\mathbb{P}(\boldsymbol{Y} = \boldsymbol{y}\,|\theta) &= \frac{d}{d\theta}\left(k\log(\theta) + (10-k)\log(1-\theta)\right) \\ &= \frac{k}{\theta} - \frac{10-k}{1-\theta} = 0 \end{aligned}$$

- Isto produz $\widehat{\theta} = k/10 = \sum_i y_i/n$.

Modelo de Contagens de Poisson

- Suponha que $\boldsymbol{Y} = (Y_1, \ldots, Y_4)$ é composto por v.a.'s iid Poisson(λ).
- Observa-se $\boldsymbol{y} = (1, 0, 3, 1)$. A função de verossmilhança $L(\lambda)$ é

$$
\begin{aligned}
L(\lambda) &= \mathbb{P}(\boldsymbol{Y} = \boldsymbol{y} \,|\lambda) \\
&= \mathbb{P}(Y_1 = 1|\lambda) \times \mathbb{P}(Y_2 = 0|\lambda) \times \mathbb{P}(Y_3 = 3|\lambda) \times \mathbb{P}(Y_4 = 1|\lambda) \\
&= \left(\frac{\lambda^1}{1!} e^{-\lambda}\right) \times \left(\frac{\lambda^0}{0!} e^{-\lambda}\right) \times \left(\frac{\lambda^3}{3!} e^{-\lambda}\right) \times \left(\frac{\lambda^1}{1!} e^{-\lambda}\right) \times \\
&= \frac{\lambda^{1+0+3+1}}{1!0!3!1!} \; e^{-4\lambda} \\
&= \frac{\lambda^{y_1+y_2+y_3+y_4}}{y_1!y_2!y_3!y_4!} \; e^{-4\lambda}
\end{aligned}
$$

Modelo de Contagens de Poisson - caso geral

- $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ é composto por v.a.'s iid Poisson(λ).
- Função de verossimilhança de λ:

$$L(\lambda) = p(\boldsymbol{Y} = \boldsymbol{y}, \lambda) = \prod_{i=1}^{n} p(y_i, \lambda) = \prod_{i=1}^{n} \frac{\lambda^{y_i} e^{-\lambda}}{y_i!} = \frac{\lambda^{\sum_i y_i} e^{-n\lambda}}{y_1! \ldots y_n!}$$

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

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

- Veja que, para achar o EMV, podemos IGNORAR o produto $y_1! \ldots y_n!$.

Contagens de Poisson

- Equação de verossimilhança:

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

- Veja que o produto $y_1! \dots y_n!$ NÃO APARECE NESTA EQUAÇÃO.
- EMV é $\widehat{\lambda}(\boldsymbol{y}) = \sum_i y_i / n = \overline{y}$.
- É ponto de máximo se $\sum_i y_i > 0$ pois

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

Contagens de Poisson

- Se $\sum_i y_i = 0$ então $\widehat{\lambda}(\boldsymbol{x}) = \overline{x} = 0$ também é ponto de máximo.
- Veja que, neste caso, $L(\lambda) = \lambda^0 e^{10\lambda}/0! = e^{10\lambda}$.
- Esta função está definida para $\lambda \in \Theta$. Isto é, para $\lambda \geq 0$.
- Seu máximo ocorre na fronteira do espaço paramétrico, quando $\lambda = 0$.
- Este ponto de máximo corresponde a $\overline{y} = \sum_i y_i/10$ mas veja que $l'(\lambda)$ NÃO É igual a zero em $\lambda = 0$.

Caso normal, apenas μ desconhecido

- $Y_1, \ldots, Y_n$ iid $N(\mu, \sigma_0^2)$ onde σ_0^2 é CONHECIDO.
- Densidade conjunta é o produto das densidades marginais:

$$\begin{aligned} L(\mu) &= f(y_1, \ldots, y_n | \mu) \\ &= \prod_{i=1}^{n} \left(\frac{1}{\sqrt{2\pi}\sigma_0} \right) \exp\left(-\frac{1}{2\sigma_0^2}(y_i - \mu)^2 \right) \\ &= \left(\frac{1}{\sqrt{2\pi}\sigma_0} \right)^n \exp\left(-\frac{1}{2\sigma_0^2} \sum_{i=1}^{n} (y_i - \mu)^2 \right) \end{aligned}$$

- Portanto, a log-verossmilhnaça de μ é dada por

$$\ell(\mu) = \log\left(f(y_1, \ldots, y_n | \mu)\right) = -n \log(\sqrt{2\pi}\sigma_0) - \frac{1}{2\sigma_0^2} \sum_{i=1}^{n} (y_i - \mu)^2$$

Caso normal, apenas μ desconhecido

- Logo,

$$
\begin{aligned}
\frac{d}{d\mu}\ell(\mu) &= \frac{d}{d\mu}\left[-n\ln(\sqrt{2\pi}\sigma_0) - \frac{1}{2\sigma_0^2}\sum_{i=1}^{n}(y_i-\mu)^2\right] \\
&= \frac{1}{2\sigma_0^2}\left(\sum_{i=1}^{n} -2(y_i-\mu)\right) \\
&= \frac{1}{\sigma_0^2}\left(\sum_{i=1}^{n} y_i - n\mu\right) \\
&= \frac{n}{\sigma_0^2}\left(\overline{y}-\mu\right)
\end{aligned}
$$

- onde $\overline{y} = \sum_i y_i/n$.
- A equação $\ell'(\mu) = 0$ tem a solução $\widehat{\mu} = \overline{y}$.
- Como $\ell''(\mu) < 0$, este é de fato um ponto de máximo.

Verossimilhança relativa

- Ao comparar diferentes experimentos, será preciso comparar diferentes funções de verossiilhança.
- Para efeito de padronização de escala nos diferentes gráficos de $L(\theta)$, será conveniente selecionar um valor para θ com o qual todos os outros valores de θ possam ser comparados.
- A escolha natural é tomar a estimativa de máxima verossimilhança $\widehat{\theta}$ para ser este valor de referência.
- Definimos a *função de verossimilhança relativa* de θ como

$$R(\theta) = \frac{L(\theta)}{L(\widehat{\theta})} = \frac{L(\theta)}{\max_\theta L(\theta)},$$

- Qualquer constante com respeito a θ que apareça na verossimilhança é cancelada por aparecer no numerador e e no denominador de $R(\theta)$.

Verossimilhança relativa

- Temos
$$0 \leq R(\theta) = \frac{L(\theta)}{L(\widehat{\theta})} = \frac{L(\theta)}{\max_\theta L(\theta)} \leq 1$$
- $R(\theta)$ é a razão de quão verossímil é θ versus o valor mais verossímil $\widehat{\theta}$:
- Seja θ_1 um valor qualquer para o parâmetro.
- Se $R(\theta_1) \leq 0.1$ então θ_1 é um valor do parâmetro mais ou menos implausível porque existem outros valores de θ para os quais os dados que acabamos de observar são 10 vezes mais prováveis de ocorrer.
- Se $R(\theta_1) \geq 0.8$ então θ_1 é um valor do parâmetro razoavelmente verossímil porque a chance dos dados aparecerem está entre 80% e 100% do valor de $L(\widehat{\theta})$, a maior probabilidade possível sob o modelo.

Verossimilhança relativa

- A função de verossimilhança relativa dá uma ordenação a todos os valores do parâmetro de acordo com a verossimilhança de cada um.
- Tome $c \approx 1$.
- O conjunto de valores de $\theta \in \Theta$ tais que $R(\theta) > c$ (isto é, tais que $L(\theta) > cL(\widehat{\theta})$) são também valores verossímeis para θ.
- Em geral, c é escolhido igual a 0.5 ou maior.

Verossimilhança relativa

- Muitas vezes, este conjunto de valores vai formar um intervalo.
- Se este intervalo for pequeno, isto quer dizer que o experimento está conseguindo separar do espaço paramétrico Θ um pequeno intervalo de valores bastante verossímeis para θ.
- Se o intervalo for muito grande, então o experimento não é capaz de diferenciar muito entre valores muito diferentes de θ.
- O experimento está dizendo que valores muito diferentes de θ são igualmente verossímeis.
- Neste sentido, ele não discrimina muito entre os valores possíveis de θ.

Verossimilhança relativa - Poisson

- $Y_1, \ldots, Y_n$ i.i.d. Poisson(λ).
- Verossimilhança: $L(\lambda) = \text{ cte } \lambda^{\sum_i y_i} e^{-n\lambda}$
- EMV $\widehat{\lambda} = \overline{y}$.
- Portanto,

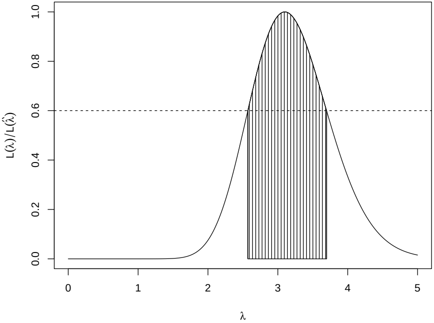
$$R(\lambda) = \frac{L(\lambda)}{L(\widehat{\lambda})} = \frac{\lambda^{\sum_i y_i} e^{-n\lambda}}{\overline{y}^{\sum_i y_i} e^{-n\overline{y}}} = \left(\frac{\lambda}{\overline{y}}\right)^{\sum_i y_i} e^{-n(\lambda - \overline{y})} .$$

- Suponha que $n = 10$ e que as seguintes contagens foram observadas: 1, 5, 4, 3, 5, 3, 2, 0, 5, 3 gerando $\overline{y} = 3.1$.
- Portanto,

$$R(\lambda) = \left(\frac{\lambda}{3.1}\right)^{31} e^{-10(\lambda - 3.1)} .$$

Verossimilhança relativa - Poisson

- $R(\lambda) > 0.6$ implica no intervalo $(2.571, 3.698)$.
- São apenas um pouco menos verossímeis para λ que o EMV $\widehat{\lambda} = \overline{y}$.



Log Verossimilhança relativa

- Às vezes, usamos a LOG-verossimilhança relativa $r(\theta)$:

$$r(\theta) = \log R(\theta) = \log L(\theta) - \log L(\widehat{\theta}) = \ell(\theta) - \ell(\widehat{\theta})$$

- Como $R(\theta)$ está entre 0 e 1, então

$$-\infty < r(\theta) = \log R(\theta) < 0 = r(\widehat{\theta}) = 0 \,.$$

Log Verossimilhança relativa

- $\theta =$ fração de pessoas que tem tuberculose.
- Amostra de n indivíduos, contagem $Y = k$ do número de doentes.
- Verossimilhança de θ:

$$L(\theta) = \binom{n}{k} \theta^k (1-\theta)^{n-k}$$

- EMV $\widehat{\theta} = k/n$.
- A função de verossimilhança relativa é então

$$R(\theta) = \frac{\binom{n}{k} \theta^k (1-\theta)^{n-k}}{\binom{n}{k} \widehat{\theta}^k (1-\widehat{\theta})^{n-k}} = \left(\frac{\theta}{\widehat{\theta}}\right)^k \left(\frac{1-\theta}{1-\widehat{\theta}}\right)^{n-k}$$

Log Verossimilhança relativa

- Dentre 100 pessoas examinadas, 3 tem tuberculose.
- Com base nestas observações que valores de θ são mais verossímeis?
- Compare com os resultados que seriam obtidos se 200 pessoas fossem examinadas e 6 tivessem tuberculose.
- A estimativa de máxima verossmilhança é a mesma nos dois casos $(= 0.03)$ mas baseada em amostras de tamanho bem diferentes.
- Log-verossimilhança para a amostra de tamanho $n = 100$ é igual a

$$\ell(\theta) = 3\log(\theta) + 97\log(1-\theta)$$

- A estimativa de máxima verossimilhança é $\widehat{\theta} = 3/100 = 0.03$.
- O máximo da log-verossimilhança é

$$\ell(\widehat{\theta}) = 3\log(0.03) + 97\log(0.97) = -13.47$$

- A função log-verossimilhança relativa é então

$$r(\theta) = \ell(\theta) - \ell(\widehat{\theta}) = 3\log(\theta) + 97\log(1-\theta) + 13.47$$

Log Verossimilhança relativa

- Se nós observamos 6 doentes em 200 nós teremos

$$\ell(\theta) = 6\log(\theta) + 194\log(1-\theta)$$

- EMV $\widehat{\theta} = 0.03$, exatamente como antes.
- O máximo da log-verossimilhança é agora $\ell(\widehat{\theta}) = -26,95$.
- A figura a seguir mostra a função log-verossimilhança relativa $r(\theta)$ de cada situação.

Log Verossimilhança relativa

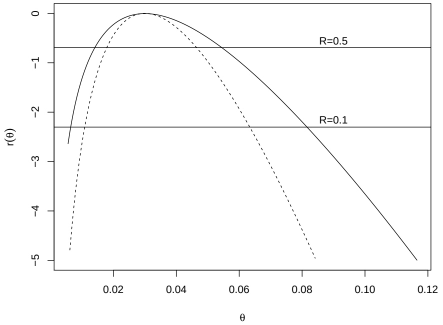


Figura: Gráfico da função log da verossimilhança relativa, $r(\lambda) = \log R(\theta)$, versus θ. A linha contínua é a verossimilhança para 3 casos em amostra de 100 indivíduos e a linha tracejada é para a situação de 6 casos em 200 indivíduos.

Log Verossimilhança relativa

- A função $r(\theta) = \log R(\theta)$ baseada na amostra de 200 pessoas tem uma curvatura maior no ponto de máximo que a função $r(\theta)$ baseada na amostra de 100 pessoas.
- A amostra maior gera intervalos mais curtos que satisfazem $R(\theta) \geq c$.
- O intervalo $(0.011, 0.063)$ satisfaz $R(\theta) \geq 0.1$ para a amostra com $n = 200$ e o intervalo $(0.006, 0.081)$ satisfaz $R(\theta) \geq 0.1$ para a amostra com $n = 100$.
- Em geral, aumentando a quantidade de dados produzirá funções de verossimilhança com maior curvatura e portanto um intervalo mais curto de valores verossímeis para o parâmetro θ.

Introdução

- Nem sempre a equação de verossimilhança $\partial\ell(\theta)/\partial\theta = 0$ admite solução analítica.
- Nestes casos, precisamos usar um método numérico
- É sempre uma boa idéia fazer um gráfico da função de verossimilhança, especialmente se ela tiver apenas um parâmetro θ unidimensional.
- A inspeção do gráfico pode revelar situações problemáticas tais como:
  - máximo não-único com vários máximos locais;
  - máximo na fronteira do espaço paramétrico, o que pode significar que o máximo de $\ell(\theta)$ não se encontra num ponto crítico dessa função.

Suposições

- Suponha que:
  - O espaço paramétrico Θ é um intervalo $[a, b]$.
  - 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 de $L(\theta)$, basta pesquisar entre as raízes da equação

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

- Vamos ver um dos métodos mais importantes para encontrar as raízes de $\ell'(\theta) = 0$, o método de Newton (ou Newton-Raphson).

Gráfico de $\ell(\theta)$

- Pesquisaremos um intervalo I supondo que ele contem apenas uma única raiz de $\ell'(\theta) = 0$.
- Ou seja, vamos supor que conseguimos isolar uma raiz dentro de um intervalo I.
- A maneira mais simples de se achar um tal intervalo dentro de Θ é fazendo um gráfico.
- Basta que se faça um esboço da função $\ell(\theta)$ ou da função $\ell'(\theta)$.
- A seguir, escolha dois pontos do eixo das abcissas entre os quais a função $\ell(\theta)$ tem seu máximo ou a função $\ell'(\theta)$ corta o eixo θ.
- Denotaremos por $\overline{\theta}$ a raiz procurada.

Método de Newton-Raphson

- Raiz $\overline{\theta}$ da equação $g(\theta) = 0$
- $g(\theta)$ é uma função complicada de θ.
- Temos um valor inicial θ_o (com sorte, não está muito longe de $\overline{\theta}$).
- Figura a seguir mostra a função não-linear $g(\theta) = \theta^2 - 5$
- Sua raiz é $\overline{\theta} = \sqrt{5} \approx 2.24$
- Valor inicial é $\theta_o = 1.5$.
- O objetivo é encontrar um novo valor θ_1 que esteja mais próximo da raiz $\overline{\theta}$.

Método de Newton-Raphson

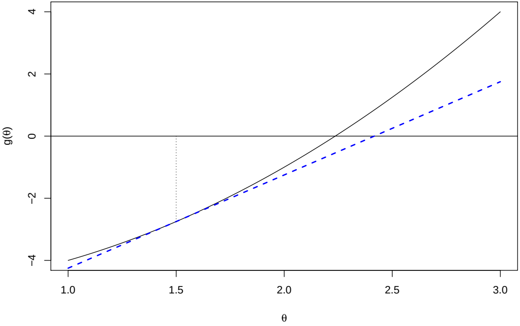


Figura: Gráfico de uma função $g(\theta)$ e sua aproximação por uma reta que passa pelo ponto $(1.5, g(1.5))$.

Método de Newton-Raphson

- Newton-Raphson aproxima a curva complicada $g(\theta)$ por uma linha reta do tipo $a + b\theta$ com intercepto a e inclinação b.
- Ao invés de encontrar a raiz da equação $g(\theta) = 0$, encontramos a raiz da equação $a + b\theta = 0$ que é simplesmente $\theta = -a/b$.
- Esta raiz da reta deve ser um valor mais próximo da raiz desejada $\overline{\theta}$.
- Na figura anterior: reta que é aproximadamente igual à função $g(\theta)$ *em torno do ponto* $(1.5, g(1.5))$.
- A raiz da linha reta é aproximadamente 2.42.
- Este é um valor mais próximo da raiz desejada $\overline{\theta} \approx 2.24$ do que o valor inicial $\theta_o = 1.5$.
- Iteramos até convergência.

Método de Newton-Raphson

- Como encontrar a reta que melhor aproxima a curva $g(\theta)$ *em torno do ponto* $(\theta_o, g(\theta_o))$?
- A reta deve passar pelo ponto $(\theta_o, g(\theta_o))$ que pertence também ao gráfico da função g.
- Se fixarmos o ponto $(\theta_o, g(\theta_o))$ pelo qual passa a reta, basta estabelecermos a inclinação da reta.
- A equação de uma reta que passa pelo ponto $(\theta_o, g(\theta_o)$ é dada por $g(\theta_o) + b(\theta - \theta_o)$.
- Por exemplo, na figura anterior, a reta que passa por $(1.5, g(1.5)) = (1.5, 1.5^2 - 5) = (1.5, -2.75)$ é igual a $-2.75 + b(\theta - 1.5)$.

Método de Newton-Raphson

- Falta encontrar b.
- A reta que melhor aproxima uma curva é a reta tangente à curva no ponto e por isto $b = g'(\theta_o)$.
- Isto é, 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 tangente.
- Isto é, achamos θ_1 que soluciona a equação

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

- A resposta é

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

Método de Newton-Raphson

- Observe a iteração:
$$\theta_1 = \theta_o - \frac{g(\theta_o)}{g'(\theta_o)}$$
- 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.
- 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 do acréscimo depende de dois fatores:
  - $g(\theta_o)$: quão distante nós estamos de $0 = g(\bar{\theta})$. Se estivermos muito distantes, devemos fazer acréscimos maiores.
  - $g'(\theta_o)$: quão rápidamente a função g está mudando de valor. Se $g'(\theta_o)$ for muito grande, podemos fazer um acréscimo pequeno.

Método de Newton-Raphson

- Seja $g(\theta) = \theta^2 - 5$
- Valor inicial $\theta_o = 1.5$
- Temos

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

- Agora, basta iterar o método.
- Usando o novo ponto θ_1 como valor inicial, repetimos o procedimento acima para encontrar uma nova aproximação θ_2 para a raiz $\overline{\theta}$.

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

Método de Newton-Raphson

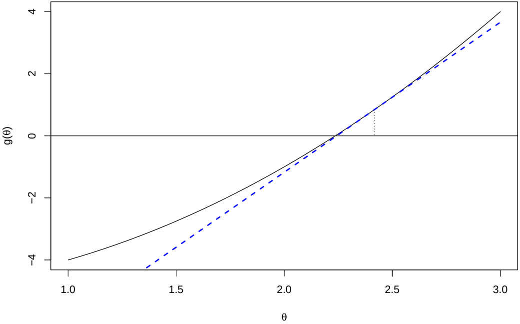


Figura: Gráfico da função $g(\theta)$ versus θ com a reta tangente que passa pelo ponto $(\theta_1, g(\theta_1)) = (2.417, g(2.417))$. A raiz desta NOVA reta é quase idêntica à raiz desejada.

Método de Newton-Raphson

- Iterando, obtemos θ_3, $\theta_4, \ldots$
- A equação recursiva é

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

- 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.

Newton-Raphson e Verossimilhança

- A equação recursiva de Newton-Raphson 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 $g(\theta) = \ell'(\theta)$
- Newton-Raphson fica então:

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

- Convergência costuma ser rápida.

Dificuldade s com o método de Newton-Raphson

- Precisamos ter um valor inicial θ_o. Se ele for muito ruim, o método pode demorar a convergir ou pode até não convergir.
- Nem sempre o método de Newton-Raphson funciona.

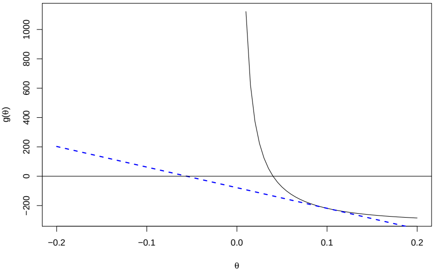


Figura: Exemplo onde o método de Newton-Raphson não converge se iniciarmos com o valor $\theta_o = 0.1$. Iniciando com $\theta_o < 0.05$, teremos convergência.

Exemplo: Acidentes de trabalho

- Quando há um acidente com um funcionário em uma fábrica, este acontecimento é registrado.
- 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*.

Um modelo para acidentes de trabalho

- 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.

EMV de λ

- Qual o EMV de λ?
- Problema: Não observamos X!!
- Observamos apenas as v.a.'s X quando $X \geq 1$. (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.

Variáveis truncadas

- 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 $y_1, ...y_n$, os valores observados de $Y_1, ..., Y_n$ que são i.i.d.
- As variáveis aleatórias observadas possuem distribuição dada por

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

Verossimilhança em acidentes de trabalho

- Assumindo que são i.i.d., a função de probabilidade conjunta $\mathbb{P}(\boldsymbol{Y} = \boldsymbol{y}) = \mathbb{P}(Y_1 = y_1) \ldots \mathbb{P}(Y_{11} = y_{11})$ é dada por

$$\prod_{i=1}^{n} \left(\frac{e^{-\lambda}}{1 - e^{-\lambda}} \frac{\lambda^{y_i}}{y_i!} \right) = \frac{\lambda^{\sum_{i=1}^{n} y_i}}{(e^{\lambda} - 1)^n \prod_{i=1}^{n} y_i!}$$

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

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

Verossimilhança em acidentes de trabalho

- 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 $n\bar{y}$.

Newton-Raphson em acidentes de trabalho

- Precisamos das expressões analíticas de $\ell'(\lambda)$ e de $\ell''(\lambda)$.
- A primeira já temos. A segunda é igual a

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

- Fórmula de recursão 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,

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

Exemplo em acidentes de trabalho

- 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.

```
set.seed(12)
x = rpois(10000, 0.01)
table(x)
##      0    1    2    3
##   8987  955   56    2

x = x[x > 0]
n = length(x)
sx = sum(x)
```

- 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 $\lambda = 0.01$.

Acidentes de trabalho

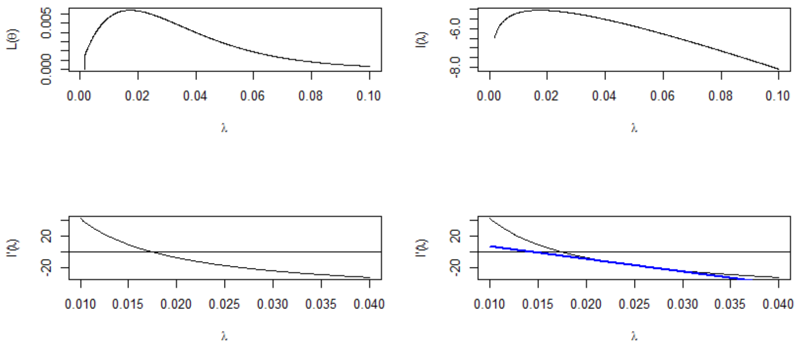
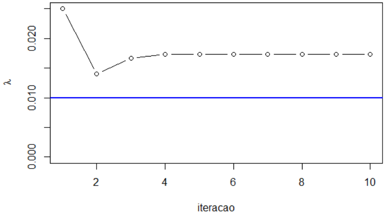


Figura: Gráfico da função de verossimilhança $L(\lambda)$, da função log-verossimilhança $\ell(\lambda)$ e da função $\ell'(\lambda)$. É mostrado também o primeiro passo do método de Newton-Raphson usando $\lambda_o = 0.025$.

Os 10 primeiros passos do Newton-Raphson: $\rightarrow$ 0.01734




```
a = rep(0, 10)
a[1] = 0.025
for(i in 2:10){
  ga = -n/(1-exp(-a[i-1])) + sx/a[i-1]
  gla = n*exp(-a[i-1])/(1-exp(-a[i-1]))^2 - sx/(a[i-1])^2
  a[i] = a[i-1] - ga/gla
}
```