EMV Multivariado

Renato Martins Assunção

DCC - UFMG

2018

MLE multivariado

- Como fazer quando $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$ é um vetor?
- Princípio é o mesmo: procurar $\boldsymbol{\theta} \in \Theta$ que maximize a verossimilhança:

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

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

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

MLE via derivadas parciais

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

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

- Defina o vetor gradiente

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

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

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

- Uma solução é um *ponto crítico*.

Ponto crítico é ponto de máximo?

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

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

- Verificamos se ela é definida negativa.
- Outros métodos são necessários se máximo ocorre na fronteira do espaço paramétrico ou quando um dos parâmetros é restrito a um conjunto discreto de valores.

Exemplo com v.a.'s contínuas: normal

- $Y_1, Y_2, \ldots, Y_n$ são i.i.d. $N(\mu, \sigma^2)$.
- A verossimilhança $L(\mu, \sigma^2)$ é a densidade conjunta das v.a.'s avaliada no ponto $\boldsymbol{y} = (y_1, \ldots, y_n)$ realmente observado:

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

- Aqui $\boldsymbol{\theta} = (\mu, \sigma^2)$
- A log-verossimilhança se torna

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

Caso gaussiano ou normal

- Por exemplo, imagine $n = 3$ e que $\boldsymbol{y} = (10.57, 11.45, 8.98)$
- Então

$$\begin{aligned} L(\mu, \sigma^2) &= f(\boldsymbol{y}\,|\mu, \sigma^2) = f(10.57, 11.45, 8.98|\mu, \sigma^2) \\ &= \left[\frac{1}{\sqrt{2\pi\sigma^2}}\right]^3 \exp\left\{-\frac{1}{2\sigma^2}\left((10.57-\mu)^2 + (11.45-\mu)^2 + (8.98-\mu)^2\right)\right\} \end{aligned}$$

- A log-verossimilhança fica igual a

$$\ell(\mu, \sigma^2) = -\frac{3}{2}\log 2\pi - \frac{3}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\left((10.57-\mu)^2 + (11.45-\mu)^2 + (8.98-\mu)^2\right)$$

As derivadas parciais

- As derivadas parciais de primeira ordem são:

$$\begin{aligned}
\frac{\partial \ell(\boldsymbol{\theta})}{\partial \mu} &= -\frac{1}{2\sigma^2}\sum_{i=1}^{n} 2(y_i-\mu)(-1) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i-\mu) \\
\frac{\partial \ell(\boldsymbol{\theta})}{\partial \sigma^2} &= -\frac{n}{2}\frac{1}{\sigma^2} - \frac{(-1)}{(\sigma^2)^2}\cdot\frac{1}{2}\sum_{i=1}^{n}(y_i-\mu)^2
\end{aligned}$$

- A estimativa de máxima verossimilhança $\widehat{\boldsymbol{\theta}} = (\widehat{\mu}, \widehat{\sigma^2})$ satisfaz as duas equações de log-verossimilhança

$$\frac{\partial \ell(\boldsymbol{\theta})}{\partial \mu} = 0 \qquad \text{e} \qquad \frac{\partial \ell(\boldsymbol{\theta})}{\partial \sigma^2} = 0$$

Equação de log-verossimilhança

- Assim, devemos ter

$$\begin{cases} \frac{1}{\widehat{\sigma^2}} \sum_i (y_i - \widehat{\mu}) & = & 0 \\ -\frac{n}{2}\frac{1}{\widehat{\sigma^2}} + \frac{\sum_{i=1}^{n}(y_i - \widehat{\mu})^2}{2(\widehat{\sigma^2})^2} & = & 0 \end{cases} \Rightarrow \begin{cases} \sum y_i - n\widehat{\mu} & = & 0 \\ -n + \frac{\sum_{i=1}^{n}(y_i - \widehat{\mu})^2}{\widehat{\sigma^2}} & = & 0 \end{cases}$$

- A primeira equação nos dá $\widehat{\mu} = \frac{\sum y_i}{n} = \overline{y}$, a média aritmética das observações.
- A segunda equação, ao substituirmos $\widehat{\mu}$ pelo valor $\overline{y}$, resulta em $\widehat{\sigma^2} = \frac{\sum (y_i - \overline{y})^2}{n}$.

MLE no caso gaussiano

- Quando observamos $\boldsymbol{y} = (10.57, 11.45, 8.98)$ estimaremos $\widehat{\boldsymbol{\theta}} = (\widehat{\mu}, \widehat{\sigma^2})$ por

$$\widehat{\boldsymbol{\theta}}(\boldsymbol{y}) = (\widehat{\mu}, \widehat{\sigma^2}) = (\overline{y}, \frac{\sum(y_i - \overline{y})^2}{3}) = (10.33, 1.04)$$

- O vetor $\widehat{\boldsymbol{\theta}}(\boldsymbol{y}) = (10.33, 1.04)$ é uma **estimativa** de $\boldsymbol{\theta}$ baseada na amostra particular $\boldsymbol{y} = (10.57, 11.45, 8.98)$.
- A **estimativa** $\widehat{\boldsymbol{\theta}}(\boldsymbol{y}) = (10.33, 1.04)$ é o valor observado do **estimador**

$$\widehat{\theta}(\boldsymbol{Y}) = (\overline{Y}, \frac{\sum(Y_i - \overline{Y})^2}{3})$$

- O **estimador** é um vetor aleatório: tem lista de valores possíveis e probabilidades associadas.

Máximo ou mínimo?

- Um exame da matriz derivada segunda mostra que $\widehat{\theta}(\boldsymbol{y})$ realmente maximiza $\ell(\boldsymbol{\theta})$.

$$D^2\ell(\boldsymbol{\theta}) = \begin{bmatrix} \frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \mu^2} & \frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \mu \partial \sigma^2} \\ \frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \sigma^2 \partial \mu} & \frac{\partial^2 \ell(\boldsymbol{\theta})}{(\partial \sigma^2)^2} \end{bmatrix} = \begin{bmatrix} \frac{-n}{\sigma^2} & \frac{-1}{\sigma^4}\sum_{i=1}^{n}(y_i - \mu) \\ \frac{-1}{\sigma^4}\sum_{i=1}^{n}(y_i - \mu) & \frac{n}{\sigma^4} - \frac{1}{\sigma^6}\sum_{i=1}^{n}(y_i - \mu)^2 \end{bmatrix}$$

- É função de $\boldsymbol{\theta} = (\mu, \sigma^2)$.
- Substituindo $\boldsymbol{\theta} = (\mu, \sigma^2)$ por $\widehat{\boldsymbol{\theta}}(\boldsymbol{y}) = (\overline{y}, \frac{\sum(y_i - \overline{y})^2}{n})$ avaliamos a matriz de derivada segunda no ponto $\widehat{\boldsymbol{\theta}}(\boldsymbol{y})$.

Máximo ou mínimo?

- Obtemos

$$D^2\ell(\boldsymbol{\theta}) \mid_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}} = -\begin{bmatrix} \frac{n}{\widehat{\sigma}^2} & 0 \\ 0 & \frac{n}{2(\widehat{\sigma}^2)^2} \end{bmatrix}$$

- Como $\widehat{\sigma}^2 > 0$, a matriz é definida negativa
- Isto implica que $\widehat{\boldsymbol{\theta}}(\boldsymbol{y})$ é ponto de máximo de $\ell(\boldsymbol{\theta})$.

Exemplo discreto: ovos de insetos

- O número de ovos deixado por um inseto segue uma distribuição de Poisson(λ).
- Uma vez deixado, cada ovo tem uma chance desconhecida p de gerar um inseto
- A geração de um ovo é independente da geração dos outros.
- Um entomologista estuda um conjunto de 10 destes insetos observando o número de ovos deixados e o número de ovos que vingaram para cada ninho.

Exemplo discreto: ovos de insetos

- Os pares seguintes correspondem aos valores do par (ovos deixados, ovos vingados) para os n pares.

número do inseto	1	2	3	4	5	6	7	8	9	10
ovos deixados	8	9	6	4	1	5	2	12	9	7
ovos vingados	5	6	4	3	0	5	2	9	8	6

- Estime $\boldsymbol{\theta} = (\lambda, p)$ por máxima verossimilhança.
- Precisamos da distribuição conjunta das v.a.'s avaliada na instância realmente observada
- As v.a.'s são discretas: Seja (O, N) onde $O =$ número de ovos deixados e $N =$ número de ovos que vingaram.

Distribuição de (O, N)

- Qual é a $\mathbb{P}(O = k, N = j)$ para UM ÚNICO INSETO?

$$\begin{aligned}
\mathbb{P}(O = k, N = j) &= \mathbb{P}(O = k) \times \mathbb{P}(N = j | O = k) \\
&= \frac{\lambda^k e^{-\lambda}}{k!} \frac{k!}{j!(k-j)!} p^j (1-p)^{k-j} \\
&= \frac{(\lambda(1-p))^k e^{-\lambda}}{j!(k-j)!} \left(\frac{p}{1-p}\right)^j
\end{aligned}$$

- com $(k, j) \in \mathbb{N}^2$ e $k \geq j$.
- Pela independência dos insetos, a conjunta é o produto das marginais (CORRIGIR):

$$\mathbb{P}(O_1 = k_1, N_1 = j_1, \ldots, O_n = k_n, N_n = j_n) = \prod_{i=1}^{n} \mathbb{P}(O_i = k_i, N_i = j_i)$$

Log-verossimilhança

- Assim, a conjunta $\mathbb{P}(O_1 = k_1, N_1 = j_1, \ldots, O_n = k_n, N_n = j_n)$ é dada por

$$\frac{e^{-n\lambda}(\lambda(1-p))^{\sum_i k_i}}{\prod_i (j_i!(k_i - j_i)!)} \left(\frac{p}{1-p}\right)^{\sum_i j_i}$$

- A função log-verossimilhança é $\ell(\boldsymbol{\theta}) = \ell(\lambda, p)$ dada por

$$-n\lambda + \sum_i k_i log(\lambda(1-p)) + (\sum_i j_i) \log(p/(1-p)) + \text{cte}$$

- onde $\text{cte} = -\sum_i \log(j_i!(k_i - j_i)!)$ não depende de $\boldsymbol{\theta} = (\lambda, p)$.
- Então

$$\begin{cases} \frac{\partial \ell}{\partial \lambda} &= -n + \frac{\sum k_i}{\lambda} \\ \frac{\partial \ell}{\partial p} &= \frac{\sum j_i}{p} + \frac{\sum j_i}{1-p} - \frac{\sum k_i}{1-p} = \frac{\sum j_i}{p} + \frac{\sum j_i - \sum k_i}{1-p} \end{cases}$$

MLE

- $\widehat{\boldsymbol{\theta}} = (\widehat{\lambda}, \widehat{p})$ é a solução das equações

$$\begin{cases} -n + \frac{\sum k_i}{\widehat{\lambda}} & = & 0 \\ \frac{\sum j_i}{\widehat{p}} + \frac{\sum j_i - \sum k_i}{1-\widehat{p}} & = & 0 \end{cases}$$

- Isto dá $\widehat{\lambda} = \frac{\sum k_i}{n}$ e $\widehat{p} = \frac{\sum j_i}{\sum k_i}$.
- Isto é, $\widehat{\boldsymbol{\theta}} = (\widehat{\lambda}, \widehat{p}) = \left(\frac{\sum k_i}{n}, \frac{\sum j_i}{\sum k_i} \right)$
- Com os dados da tabela, obtemos a estimativa $\widehat{\boldsymbol{\theta}} = (6.3, 0.76)$ para $\boldsymbol{\theta} = (\lambda, p)$.

MLE

- As estimativas $\widehat{\lambda}$ e $\widehat{p}$ são bem intuitivas.
- $\widehat{\lambda} = \frac{\sum k_i}{n}$ é média aritmética de ovos deixados por todos os insetos
- $\widehat{p} = \frac{\sum j_i}{\sum k_i}$ é a proporção de ovos vingados dentre todos os ovos deixados, agregando sobre os ovos de todos os insetos.
- Uma alternativa também intuitiva para estimar p poderia ser o seguinte estimador:

$$\widehat{\widehat{p}} = \frac{1}{n}\left(\frac{j_1}{k_1} + \frac{j_2}{k_2} + \ldots + \frac{j_n}{k_n}\right)$$

- Embora $\widehat{p}$ e $\widehat{\widehat{p}}$ sejam ambos intuitivos, o MLE $\widehat{p}$ é melhor. Num sentido que faremos preciso mais tarde, ele é o melhor estimador possível de p.

Máximo?

- O ponto crítico $\widehat{\boldsymbol{\theta}} = (\widehat{\lambda}, \widehat{p}) = \left(\frac{\sum k_i}{n}, \frac{\sum j_i}{\sum k_i} \right)$ corresponde realmente a um ponto de máximo?
- A matriz de derivada segunda avaliada no ponto $\underset{\sim}{\widehat{\theta}}$ é definida negativa:

$$D^2\ell(\boldsymbol{\theta}) = \begin{bmatrix} \frac{\partial^2 \ell}{\partial \lambda^2} & \frac{\partial^2 \ell}{\partial \lambda \partial p} \\ \frac{\partial^2 \ell}{\partial p \partial \lambda} & \frac{\partial^2 \ell}{\partial p^2} \end{bmatrix} = \begin{bmatrix} -\frac{\sum k_i}{\lambda^2} & 0 \\ 0 & -\frac{\sum j_i}{p^2} + \frac{\sum j_i - \sum k_i}{(1-p)^2} \end{bmatrix}$$

Máximo?

- Avaliando $D^2\ell(\boldsymbol{\theta})$ no ponto $\widehat{\boldsymbol{\theta}}$ temos

$$D^2\ell(\boldsymbol{\theta}) = \begin{bmatrix} -\frac{n^2}{(\sum k_i)} & 0 \\ 0 & -\frac{(\sum k_i)^2}{\sum j_i} + \frac{(\sum k_i)^2}{\sum j_i - \sum k_i} + \end{bmatrix}$$

- Como $j_i \leq k_i \Rightarrow \sum j_i - \sum k_i < 0$ e portanto a matriz diagonal acima tem todas suas entradas negativas e portanto é definida negativa

EMV Multinomial

- Considere variáveis aleatórias $X_1, ..., X_n$ independentes tais que cada uma delas toma valores em $\{1, 2, ..., J\}$.
- Por exemplo, nós podemos estar classificando filmes em J categorias tais como:
  - $1 \rightarrow$ terror
  - $2 \rightarrow$ drama
  - ...
  - J $\rightarrow$ ação
- Nós queremos estimar as probabilidades $\theta_1, ..., \theta_J$ de obter os resultados $1, 2, ..J$.
- Suponha que nas n classificações observamos n_1 resultados na classe 1, n_2 resultados 2, ..., n_J resultados iguais a J.
- Isto é, temos o vetor aleatório $(n_1, n_2, \ldots, n_J)$ com distribuição multinomial $\mathcal{M}(n; \theta_1, ..., \theta_J)$.

Verossimilhança multinomial

- O espaço paramétrico é dado por

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

- O chute óbvio para estimar cada θ_k é usar $\hat{\theta}_k = \frac{n_k}{n}$. Este é o EMV:

$$P_{\boldsymbol{\theta}}(N_1 = n_1, ..., N_J = n_J) = \frac{n!}{n_1! \ldots n_J!} \theta_1^{n_1} \theta_2^{n_2} \ldots \theta_J^{n_J}$$

- Assim

$$L(\boldsymbol{\theta}) = c + n_1 \log \theta_1 + n_2 \log \theta_2 + \ldots + n_J \log \theta_J$$

onde $c = \log(n!/(n_1! \ldots n_J!)$ é uma constante em termos dos parâmetros.

Equação de verossimilhança

- Queremos maximizar em $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_J)$ a expressão

$$L(\boldsymbol{\theta}) = c + n_1 \log \theta_1 + n_2 \log \theta_2 + \ldots + n_J \log \theta_J$$

- Temos uma restrição:

$$\theta_1 + \theta_2 + \ldots + \theta_J = 1$$

exigindo o uso de multiplicadores de Lagrange.

- Montamos a nova equação

$$g(\boldsymbol{\theta}, \lambda) = L(\boldsymbol{\theta}) - \lambda \left(\sum_{k=1}^{J} \theta_k - 1 \right)$$

- Resolvemos (agora sem restrições) o sistema

$$\begin{cases} \frac{\partial g}{\partial \theta_i} = \frac{\partial L(\boldsymbol{\theta})}{\partial \theta_i} - \lambda \frac{1}{\partial \theta_i} \left(\sum_{k=1}^{n} \theta_k - 1 \right) \\ \frac{\partial g}{\partial \lambda} = \sum_{k=1}^{n} \theta_k - 1 \end{cases} \qquad i = 1, .., J$$

EMV

- Temos então

$$\begin{cases} \frac{n_i}{\theta_i} - \lambda = 0 \\ \sum_{k=1}^{n} \theta_k = 1 \end{cases} \qquad i = 1, .., J$$

- Assim os θ_i maximizadores satisfazem $\lambda = \frac{n_i}{\theta_i}$ ou seja, $\theta_i = n_i/\lambda$
- Somando sobre todos os k e lembrando que $\sum_k \theta_k = 1$, temos que

$$1 = \sum_k \theta_k = \sum_i \frac{n_i}{\lambda} = n/\lambda$$

- Assim, $\lambda = n$ e portanto,

$$\widehat{\theta}(\boldsymbol{x}) = \left(\widehat{\theta}_1, \widehat{\theta}_2, \ldots, \widehat{\theta}_J\right) = \left(\frac{n_1}{n}, ..., \frac{n_J}{n}\right)$$

MLE de Normal Bivariada

- Dados são n pares de observações $(x_1, y_1), ..., (x_n, y_n)$.
- Modelo: i.i.d. normal bivariada
- Parâmetro: $\boldsymbol{\theta} = (\mu_1, \mu_2, \sigma_1, \sigma_2, \rho)$
- Densidade conjunta dos n pares:

$$\left(\frac{1}{\sqrt{2\pi(1-\rho^2)}\sigma_1\sigma_2} \right)^n e^{-\frac{1}{2}Q(\boldsymbol{x},\boldsymbol{y})}$$

- onde

$$Q(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{1-\rho^2} \sum_{k=1}^{n} \left[\left(\frac{x_k - \mu_1}{\sigma_1} \right)^2 - 2\rho \left(\frac{x_k - \mu_1}{\sigma_1} \right) \left(\frac{y_k - \mu_2}{\sigma_2} \right) + \left(\frac{y_k - \mu_2}{\sigma_2} \right)^2 \right]$$

Função log-verossimilhança

- Função log-verossimilhança é

$$\ell(\boldsymbol{\theta}) = -\frac{n}{2}\log\left[2\pi(1-\rho^2)\right] - n\log\sigma_1\sigma_2 - \frac{1}{2}Q(\boldsymbol{x},\boldsymbol{y})$$

- MLE de $\boldsymbol{\theta}$ maximiza $\ell(\boldsymbol{\theta})$: resolvemos as cinco equações simultaneamente

$$\begin{cases} \frac{\partial \ell(\boldsymbol{\theta})}{\partial \mu_1} = 0 \\ \frac{\partial \ell(\boldsymbol{\theta})}{\partial \mu_2} = 0 \\ \frac{\partial \ell(\boldsymbol{\theta})}{\partial \sigma_1} = 0 \\ \frac{\partial \ell(\boldsymbol{\theta})}{\partial \sigma_2} = 0 \\ \frac{\partial \ell(\boldsymbol{\theta})}{\partial \rho} = 0 \end{cases}$$

MLE normal bivariada

- Uma série de manipulações algébricas elementares, descritas nos próximos slides **mas opcionais nesta disciplina introdutória**, mostram que o MLE de $\boldsymbol{\theta} = (\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$ é dado por
-

$$\widehat{\boldsymbol{\theta}} = \left(\overline{x}, \overline{y}, \frac{\sum (x_i - \overline{x})^2}{n}, \frac{\sum (y_i - \overline{y})^2}{n}, \frac{\sum (x_i - \overline{x})((y_i - \overline{y})}{\sqrt{\sum (x_i - \overline{x})^2} \sqrt{\sum (y_i - \overline{y})^2}} \right)$$

MLE normal bivariada

- Temos

$$\begin{aligned}\frac{\partial L(\boldsymbol{\theta})}{\partial \mu_1} &= \frac{\partial Q}{\partial \mu_1} = \frac{1}{1-\rho^2}\left(\sum_{k=1}^{n}\left[\frac{-2}{\sigma_1}\right] - 2\rho\left(\frac{x_k-\mu_2}{\sigma_2}\right)\left(\frac{-1}{\sigma_1}\right)\right) \\ &= \frac{2}{1-\rho^2}\left(\frac{1}{\sigma_1^2}\right)\left(\sum x_k - n\mu_1\right) + \frac{\rho}{\sigma_1\sigma_2}\left(\sum y_k - n\mu_2\right)\end{aligned}$$

- Deste modo, $\frac{\partial L(\boldsymbol{\theta})}{\partial \mu_1} = 0$ implica

$$\frac{1}{\widehat{\sigma_1}}\left(\sum x_k - n\widehat{\mu_1}\right) + \frac{\widehat{\rho}}{\widehat{\sigma_2}}\left(\sum y_k - n\widehat{\mu_2}\right) = 0\,(*)$$

MLE normal bivariada

- Pela simetria de $Q(\boldsymbol{x}, \boldsymbol{y})$, $\frac{\partial L(\boldsymbol{\theta})}{\partial \mu_2} = 0$ implica

$$\frac{1}{\widehat{\sigma_2}} \left(\sum y_k - n\widehat{\mu_2} \right) + \frac{\widehat{\rho}}{\widehat{\sigma_1}} \left(\sum x_k - n\widehat{\mu_1} \right) = 0 \; (**)$$

- Multiplicando $(**)$ por $-\widehat{\rho}$ e somando com $(*)$ obtemos

$$\begin{aligned} \frac{1}{\widehat{\sigma_1}} \left(\sum x_k - n\mu_1 \right) + \frac{\widehat{\rho}^2}{\widehat{\sigma_1}} \left(\sum x_k - n\mu_1 \right) &= 0 \\ \Rightarrow \left(\sum x_k - n\widehat{\mu_1} \right) (1 - \widehat{\rho}^2) &= 0 \Rightarrow \\ \Rightarrow 1 - \widehat{\rho}^2 = 0 \;\; \text{ou} \;\; \sum x_k - n\widehat{\mu_1} &= 0 \end{aligned}$$

MLE normal bivariada

- Mas $1 - \widehat{\rho}^2 = 0 \Rightarrow \widehat{\rho} = \pm 1$ e $\pm 1 \notin \Theta$ já que ρ pertence ao intervalo aberto $(-1, 1)$.
- Assim, temos

$$\sum x_k - n\widehat{\mu_1} = 0 \Rightarrow \widehat{\mu_1} = \overline{x}$$

..

- De maneira análoga, $\widehat{\mu_2} = \overline{y}$
- Substituindo $\widehat{\mu_1} = \overline{x}$ e $\widehat{\mu_2} = \overline{y}$ nas três equações restantes, temos

$$\begin{cases} n + \frac{1}{1-\widehat{\rho}}\left(-\frac{S_{xx}}{\widehat{\sigma}_1^2} + \widehat{\rho}\frac{S_{xy}}{\widehat{\sigma}_1\widehat{\sigma}_2}\right) = 0 \quad (*) \\ n + \frac{1}{1-\widehat{\rho}^2}\left(-\frac{S_{yy}}{\widehat{\sigma}_2^2} + \widehat{\rho}\frac{S_{xy}}{\widehat{\sigma}_1\widehat{\sigma}_2}\right) \quad (**) \\ n\widehat{\rho} - \widehat{\rho}\widehat{Q} + \frac{S_{xy}}{\widehat{\sigma}_1\widehat{\sigma}_2} = 0 \quad (***) \end{cases}$$

- onde $S_{xx} = \sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$ e $\widehat{Q} = \frac{1}{1-\widehat{\rho}}\left(\frac{S_{xx}}{\widehat{\sigma}_1^2} + \frac{2\widehat{\rho}}{\widehat{\sigma}_1\widehat{\sigma}_2}S_{xy} + \frac{S_{yy}}{\widehat{\sigma}_2}\right)$

MLE

- Usando $(*)$ e $(**)$, temos que $\widehat{Q} = 2n$ e portanto $(***)$ fica

$$-n\widehat{\rho} - 2n\widehat{\rho} + \frac{S_{xy}}{\widehat{\sigma}_1\widehat{\sigma}_2} = 0 \Rightarrow \widehat{\rho} = \frac{1}{n}\frac{S_{xy}}{\widehat{\sigma}_1\widehat{\sigma}_2} \quad (***)$$

- Substituindo este valor de $\widehat{\rho}$ em $(*)$ obtemos:

$$n + \frac{n^2\widehat{\sigma}_1^2\widehat{\sigma}_2^2}{n^2\widehat{\sigma}_1^2\widehat{\sigma}_2^2 - S_{xy}^2}\left(-\frac{S_{xx}}{\widehat{\sigma}_1^2} + \frac{S_{xy}^2}{\widehat{\sigma}_1^2\widehat{\sigma}_2^2}\right) = 0 \text{ isto é ,}$$

$$1 = \left(S_{xx}\widehat{\sigma}_2^2 + \frac{S_{xy}^2}{n}\right)\frac{n}{n^2\widehat{\sigma}_1^2\widehat{\sigma}_2^2 - S_{xy}^2} \text{ donde}$$

$$n^2\widehat{\sigma}_1^2\widehat{\sigma}_2^2 - S_{xy}^2 = nS_{xx}\widehat{\sigma}_2^2 - S_{xy}^2$$

- Isto é,

$$\widehat{\sigma}_1^2 = \frac{S_{xx}}{n} \Rightarrow \widehat{\sigma}_1^2 = \frac{\sum(x_i - \overline{x})^2}{n}$$

MLE normal bivariada

- De maneira análoga,

$$\widehat{\sigma}_2^2 = \frac{\sum(y_i - \overline{y})^2}{n}$$

- Substituindo em $(****)$ temos ,

$$\widehat{\rho} = \frac{\sum(x_i - \overline{x})((y_i - \overline{y})}{\sqrt{\sum(x_i - \overline{x})^2}\sqrt{\sum(y_i - \overline{y})^2}}$$

- Resumindo, a estimativa de máxima verossimilhança de $\boldsymbol{\theta} = (\mu_1, \mu_2, \sigma_1, \sigma_2, \rho)$ é dada por

$$\widehat{\boldsymbol{\theta}} = \left(\overline{x}, \overline{y}, \frac{\sum(x_i - \overline{x})^2}{n}, \frac{\sum(y_i - \overline{y})^2}{n}, \frac{\sum(x_i - \overline{x})((y_i - \overline{y})}{\sqrt{\sum(x_i - \overline{x})^2}\sqrt{\sum(y_i - \overline{y})^2}} \right)$$

Exponencial dupla - OPCIONAL

- EXEMPLO DE LEITURA OPCIONAL
- Até este ponto, o MLE produziu poucas surpresas.
- Se $X \sim N(\mu, \sigma^2)$ então $\mu = EX$ e assim naturalmente estimaríamos μ pela média amostral $\widehat{\mu} = (x_1 + ... + x_n)/n$.
- Mas estas estimativas são naturais somente se estivermos usando a família normal para a distribuição dos dados.
- Sejam $x_1, x_2..., x_n$ i.i.d. com distribuição dupla exponencial com densidade

$$f(x) = \frac{1}{2\beta} e^{-|x-\alpha|/\beta}$$

- onde $\alpha \in \mathbb{R}$ e $\beta > 0$. Faça um esboço dessa densidade para enteder seus parâmetros.
- Parecido com o caso normal, temos $\alpha = EX$ e $\underline{e}ta = \text{Var}(X)/2$.
- Isto poderia levar a estimar α e β^2 pela média amostral e por $s^2/2$.
- Vamos ver que o MLE é diferente dessas estimativas naturais.

MLE de exp dupla - OPCIONAL

- EXEMPLO DE LEITURA OPCIONAL
- A log-verossimilhança de $\boldsymbol{\theta} = (\mu, \beta)$ é

$$\ell(\boldsymbol{\theta}) = -n \log(2\beta) - \frac{1}{\beta} \sum_{k=1}^{n} |x_k - \alpha|$$

- Para qualquer valor de β , o valor de α que maximiza $\ell(\theta)$ é aquele que minimiza

$$S(\alpha) = \sum_{k=1}^{n} |x_k - \alpha|$$

- Isto é, devemos minimizar a soma dos desvios ABSOLUTOS e não dos desvios ao quadrado, como acontece no caso da gaussiana.

MLE de exp dupla - OPCIONAL

- EXEMPLO DE LEITURA OPCIONAL
- Um argumento meio longo mostra que o valor de α que minimiza a soma dos desvios absolutos é a mediana dos dados

$$\widehat{\alpha}(x_1, ..., x_n) = \begin{cases} (\frac{n+1}{2}) \text{ se n é ímpar} \\ \text{qualquer valor em } (x(\frac{n}{2}), x(\frac{n}{2}+1)) \text{ se n é par.} \end{cases}$$

- Como esta estimativa de α, substituimos em $\ell(\boldsymbol{\theta})$ para obter a estimativa de β, que é o desvio absoluto em volta da mediana:

$$\widehat{\beta} = \frac{1}{n}\sum_{k=1}^{n} |x_k - \widehat{\alpha}|$$

MLE ou média? - OPCIONAL

- EXEMPLO DE LEITURA OPCIONAL
- O fato de que a mediana amostral, e não a média amostral, é o estimador de máxima verossimilhança para o parâmetro α na distribuição exponencial dupla não tem utilidade nenhuma a menos que nós saibamos que estimativas de máxima verossimilhança são boas.
- Nós mostraremos que isto é verdadeiro no próximo capítulo.
- O importante é que estimadores que são bons para uma família paramétrica de distribuições são especificamente construídos para aquela família.
- Eles podem ser muito ruins para outras famílias.
- Usar a média amostral na distribuição exponencial dupla dará resultados piores que usar a mediana.

Regressão linear múltipla: preços de imóveis

- Preços de 1500 imóveis (vetor de dimensão 1500)

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

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

- $1 + 30$ características de 1500 imóveis (Matriz X de dimensao $1500 \times (30 + 1)$)

Modelo: Preço é uma soma ponderada

- Um modelo matemático simples que possa explicar, a partir das características, porque alguns imóveis são caros e outros são baratos.
- Preço é uma soma ponderada de fatores (features): achar a melhor combinação de pesos.
- Nosso problema é encontrar os coeficientes $\beta_0, \beta_1, \ldots, \beta_{30}$ tais que

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_{1499} \\ y_{1500} \end{pmatrix} \approx \beta_0 \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \\ 1 \end{pmatrix} + \beta_1 \begin{pmatrix} \text{área}_1 \\ \text{área}_2 \\ \vdots \\ \text{área}_{1499} \\ \text{área}_{1500} \end{pmatrix} + \beta_2 \begin{pmatrix} \text{idade}_1 \\ \text{idade}_2 \\ \vdots \\ \text{idade}_{1499} \\ \text{idade}_{1500} \end{pmatrix} + \ldots + \beta_{30} \begin{pmatrix} \text{salão}_1 \\ \text{salão}_2 \\ \vdots \\ \text{salão}_{1499} \\ \text{salão}_{1500} \end{pmatrix}$$

Forma matricial

- Queremos encontrar $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_{30})$ tais que

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_{1498} \\ y_{1499} \\ y_{1500} \end{pmatrix} \approx \begin{pmatrix} 1 & \text{renda}_1 & \text{área}_1 & \cdots & \text{salão}_1 \\ 1 & \text{renda}_2 & \text{área}_2 & \cdots & \text{salão}_2 \\ \vdots & \vdots & \vdots & \vdots & \\ 1 & \text{renda}_{1499} & \text{área}_{1499} & \cdots & \text{salão}_{1499} \\ 1 & \text{renda}_{1500} & \text{área}_{1500} & \cdots & \text{salão}_{1500} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{30} \end{pmatrix} = X\,\boldsymbol{\beta}$$

- Isto é, queremos $X\,\boldsymbol{\beta} \approx Y$. Como resolver isto? Mínimos quadrados...
- Já aprendemos: $Y \approx \widehat{Y} = X\ \boldsymbol{\beta} = X\ (X'X)^{-1}X'Y$ é a projeção ortogonal de Y no espaço vetorial das combinações lineares das colunas de X.

A visão probabilística

- Esta é uma visão puramente *data-driven* do problema de regressão.
- Podmeos adotar uma outra visão, mais probabilística e centrada num modelo generativo para os dados.
- Ela leva a MESMA SOLUÇÃO numérica anterior.
- A vantagem é que ela é generalizável e ela permite estudar as **propriedades** do estimador de mínimos quadrados
- Abordagem via verossimilhança.

Regressão linear, mínimos quadrados e verossimilhança

- Amostra de terinamento de n dados ou exemplos da forma $(\boldsymbol{x}_i, y_i)$ onde $i = 1, \ldots, n$.
- Os regressores $\boldsymbol{x} = (1, x_1, \ldots, x_p)$ são considerados fixos, constantes (MESMO QUE, DE FATO, SEJAM VARIÁVEIS ALEATÓRIAS).
- A razão para isto é que queremos uum modelo para a distribuição de y DADOS OS VALORES EM $\boldsymbol{x}$.
- Modelo de regressão linear: $(y|\,\boldsymbol{x}) \sim \mu(\boldsymbol{x}) +$ "erro"
- com $\mu(\boldsymbol{x}) = \beta_0.1 + \beta_1 x_1 + \ldots \beta_p x_p$.
- Além disso, as observações $y_1, \ldots, y_n$ são v.a.'s independentes.

Duas formas de escrever

- Resultado de probab: Se $\epsilon \sim N(0, \sigma^2)$ e μ é uma constante então $Y = \mu + \epsilon \sim N(\mu, \sigma^2)$.
- Modelo de regressão linear: $(y|\,\boldsymbol{x}) \sim \mu(\boldsymbol{x}) +$ "erro"
- com $\mu(\boldsymbol{x}) = \beta_0.1 + \beta_1 x_1 + \ldots \beta_p x_p$.
- Vamos assumir um "erro" com distribuição $N(0, \sigma^2)$.
- Então $(Y|\,\boldsymbol{x}) = \mu(\boldsymbol{x}) + \epsilon \sim N(\mu(\boldsymbol{x}), \sigma^2)$.
- Os regressores $\boldsymbol{x}$ afetam apenas $\mathbb{E}(Y|\,\boldsymbol{x})$, não afetam a variância σ^2.

Log-Verossimilhança e gaussianas

- Como encontrar BONS estimadores para os parâmetros $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^2) = (\beta_0, \beta_1, \ldots, \beta_p, \sigma^2)$?
- Nossa máquina automática de estimar (BEM) parâmetros: MLE (ou EMV)
- Escreva a densidade conjunta (a "probabilidade") de observar os dados que você de fato observou como uma função de $\boldsymbol{\theta}$ e maximize.
- O que é aleatório? Apenas os y's, que são gaussianos e independentes (mas não são i.d.)

Log-Verossimilhança

- Densidade de UMA gaussiana $N(\mu, \sigma^2)$:

$$f(y|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2\right)$$

- Densidade CONJUNTA de n gaussianas INDEPENDENTES com médias distintas $\boldsymbol{\mu} = \mu_1, \mu_2, \ldots, \mu_n$ e mesma variância σ^2

$$f(\boldsymbol{y} \mid \boldsymbol{\mu}, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\left(\frac{y_i-\mu_i}{\sigma}\right)^2\right)$$

onde $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)$.

Log-Verossimilahnça

- $\ell(\boldsymbol{\theta}) = \log \prod_{i=1}^{n} N(\boldsymbol{x}_i' \boldsymbol{\beta}, \sigma^2)$ onde $\boldsymbol{x}_i' = 1\ \beta_0 + \beta_1 x_1 + \ldots \beta_p x_p$
- Temos
-
$$\ell(\boldsymbol{\theta}) = -\frac{n}{2} \log\left(2\pi\sigma^2\right) - \sum_{i=1}^{n} \frac{1}{2\sigma^2} \left(y_i - \boldsymbol{x}_i' \boldsymbol{\beta}\right)^2$$
- onde $\boldsymbol{x}_i' \boldsymbol{\beta} = \beta_0 + \beta_1 x_{i1} + \ldots \beta_p x_{ip}$

Equação de log-verossimilhança

- Derive $\ell(\boldsymbol{\theta})$ com respeito a cada um dos $p+1$ coeficientes β_j e também com relação a σ^2 e iguale a zero.
- Teremos sistema com $p+2$ equações
- As primeiras $p+1$ equações formam um sistema LINEAR:

$$(X'X)\,\boldsymbol{\beta} = X'Y$$

gerando a solução MLE

$$\widehat{\boldsymbol{\beta}} = (X'X)^{-1}X'Y$$

- Os valores de Y preditos pelo modelo são $\widehat{Y} = X\widehat{\boldsymbol{\beta}} = X(X'X)^{-1}X'Y$
- Com relação a σ^2 encontramos o estimador MLE

$$\widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2$$

Regressão linear em R

```
dados <- read.table("ApsBH.txt", header=T)
head(dados)
y <- as.matrix(dados[, 6], ncol=1)
x <- as.matrix(cbind( rep(1, length(y)), dados[,2:5]), ncol=5)
beta <- solve(t(x) %*% x) %*% (t(x) %*% y)
beta
```

Comando `lm` implementa a regressão linear múltipla usando decompoisção QR
Valor de retorno de `lm` é uma lista com vários elementos necessários para a análise de dados:

```
x = as.matrix(dados[,2:5], ncol=4)
betaR <- lm(as.numeric(y) ~ x)
betaR$coef
```

Regressão com pesos

- Algumas vezes, para estimar β numa regressão linear múltipla, podemos querer dar mais peso a alguns dados que a outros.
- Se alguns dados tiverem erros com maior variância que outros, eles podem ter pesos menores que os de menor variânciaao estimar β.
- Por exemplo, considere o gráfico de preçcos versus área do apto no próximo slide.

Preço versus área

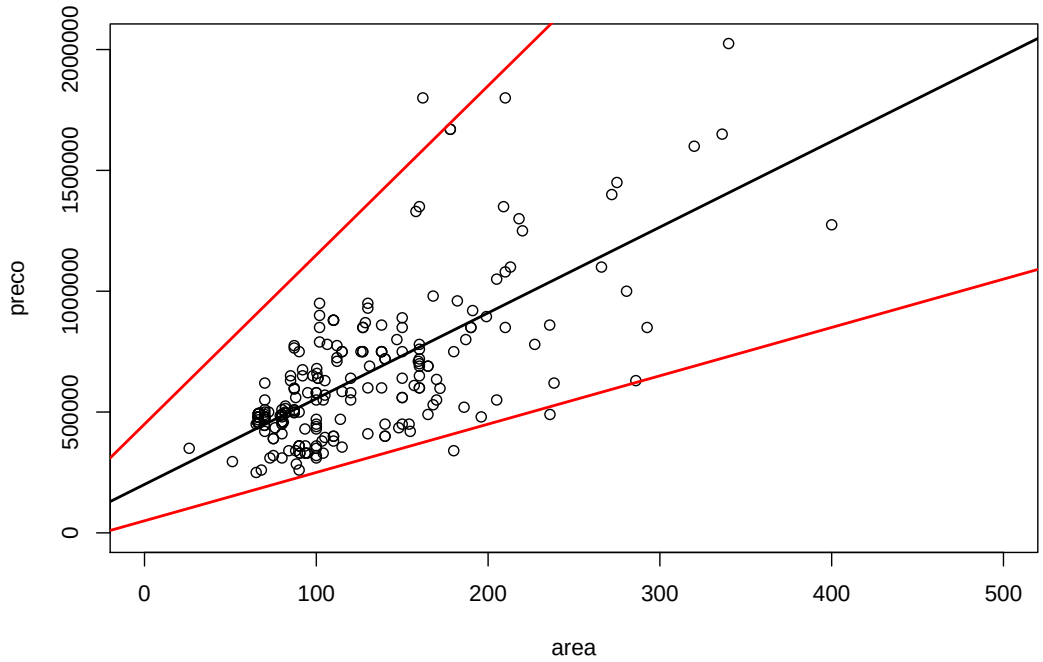

Figura: Eixo vertical e' o preco anunciado. Eixo horizantal é a área do apto. As linhas vermelhas mostram o aumento do desvio padrão σ com a área.

Regressão com pesos

- O desvio padrão parece crescer linearmente com o aumento da área.
- Num caso como este, nosso modelo de regressão pode ser estendido permitindo que não apenas $\mathbb{E}(Y|\boldsymbol{x}) = \mu(\boldsymbol{x})$ seja função das covariáveis.
- Podemos permitir que o desvio padrão também mude de observação para observação com $\boldsymbol{x}$.
- O gráfico anterior sugere adotar o seguinte modelo generativo

$$(Y_i|\boldsymbol{x}) \sim N(\mu(\boldsymbol{x}), \sigma^2 g(\boldsymbol{x})) = N(\boldsymbol{x}'\boldsymbol{\beta}, \sigma^2 x_1^2)$$

onde x_1 é a área do apto (a primeira covariável).

Verossimilhança

- Suponha que a variância da i-ésima observação é proporcional a um peso conhecido w_i.
- Isto é, $\mathbb{V}(Y_i|\boldsymbol{x}_i) = \sigma^2 w_i$ onde w_i É CONHECIDO.
- Densidade CONJUNTA de n gaussianas INDEPENDENTES com médias e variâncias distintas

$$L(\boldsymbol{\beta}, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2 w_i}} \exp\left(-\frac{1}{2\sigma^2 w_i}\left(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}\right)^2\right)$$

Log-verossimilhança

- Tomando log temos

$$\ell(\boldsymbol{\theta}) = -\frac{n}{2}\log\left(2\pi\sigma^2 w_i\right) - \sum_{i=1}^{n} -\frac{1}{2\sigma^2 w_i}\left(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}\right)^2$$

- onde $\boldsymbol{x}_i'\boldsymbol{\beta} = \beta_0 + \beta_1 x_{i1} + \ldots \beta_p x_{ip}$
- Veja que, para qualquer valor de σ^2, para maximizar $\ell(\boldsymbol{\theta})$ com respeito a $\boldsymbol{\beta}$, basta minimizar a soma de quadrados ponderada

$$\sum_{i=1}^{n} \frac{1}{w_i}(y_i - \boldsymbol{x}_i'\boldsymbol{\beta})^2$$

- Os pesos são os inversos das constantes w_i.
- Lembre-se que os w_i são conhecidos. No nosso exemplo, w_i é a área (ao quadrado) do apto i.

Log-verossimilhança

- Outra maneira de obter o MLE de $\boldsymbol{\theta}$, mais mecânica, é derivando $\ell(\boldsymbol{\theta})$.
- Derivando em relação a cada coeficiente β_j e a σ^2 terminamos com um sistema de equações.
- A parte referente a estimação de $\boldsymbol{\beta}$ é um sistema LINEAR.

Mínimos quadrados ponderados

- MLE de $\boldsymbol{\beta}$ no modelo de regerssão linear (gaussiana) quando a variância $\mathbb{V}(Y_i|\boldsymbol{x}_i) = \sigma^2 w_i$ é a solução de um sistema LINEAR:

$$(X'\Omega^{-1}X)\,\boldsymbol{\beta} = X'\Omega^{-1}Y$$

  onde Ω é uma matriz diagonal com os elementos $(w_1, w_2, \ldots, w_n)$.
- Assim, o MLE de $\boldsymbol{\beta}$ é

$$\widehat{\boldsymbol{\beta}} = \left(X'\Omega^{-1}X\right)^{-1}\; X'\Omega^{-1}Y$$

- O valores preditos pelo modelo são $\widehat{Y} = X\widehat{\boldsymbol{\beta}}$ e o MLE de σ^2 é igual a

$$\widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y_i})^2$$

Mínimos quadrados ponderados no R

- Fazendo no R a regressão usual e a regressão com peso igual a área ao quadrado:
- basta acionar o parâmetro `weights` com o argumento sendo o vetor com os pesos w_i INVERTIDOS.

```
> lm(y ~ x)$coef
(Intercept)       xArea    xQuartos     xSuites       xVaga
-269382.128    1915.898   59637.006  111743.835  191404.12
> lm(y ~ x, weights = 1/x[,1]^2)$coef
(Intercept)       xArea    xQuartos     xSuites       xVaga
-112788.380    1608.312   68863.041   29343.123  156886.44
```

MLE por métodos numéricos

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

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

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

MLE por Newton-Raphson

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

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

- Caso multivariado:

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

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

MLE de 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

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

Verossimilhança

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

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

- 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

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

EMV

- Os elementos incluídos na soma são aqueles que correspondem à uma resposta do tipo $Y = 1$.
- O EMV de $\boldsymbol{\theta} = (\beta_0, \beta_1)$ 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

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

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

A matriz da derivada segunda

- Temos

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

- Como valor inicial, use $\boldsymbol{\theta}_0 = (\log(\bar{y}/(1-\bar{y})), 0)$.
- 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

Regressão logística com múltiplos regressores

- 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.

Função de ligação: log(odds)

- 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:

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

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

Notação matricial

- Seja $\boldsymbol{X}$ a matriz com as variáveis regressoras.
- $\boldsymbol{X}$ é uma matriz $n \times (p+1)$.
- 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 $\boldsymbol{y}$ de dimensão $n \times 1$ com os valores da variável resposta.
- Temos o vetor-coluna $\boldsymbol{\theta} = (\beta_0, \beta_1, \ldots, \beta_p)$ de dimensão $(p+1) \times 1$.

Notação matricial

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

Notação matricial

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

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

- onde

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

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

Verossimilhança em notação matricial

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

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

Maximizando

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

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

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

$\partial \ell(\boldsymbol{\theta})/\partial \boldsymbol{\theta}$ em notação matricial

- Em forma matricial, nós escrevemos

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

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

$$\boldsymbol{X}^t \, \boldsymbol{y} = \boldsymbol{X}^t \, \boldsymbol{p}$$

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

Hessiana

- 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)

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

Hessiana em notação matricial

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

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

- Isto é,

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

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

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

Equação iterativa de Newton-Raphson

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

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

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

Newton-Raphson em forma matricial

- Como

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

- e

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

- Temos então

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

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.devianc
[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"
```

Logística do zero

- 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

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

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

IRLS - opcional

- Vimos que
$$\theta^{\text{new}} = \left(\boldsymbol{X}^t \boldsymbol{W} \boldsymbol{X}\right)^{-1} \boldsymbol{X}^t \boldsymbol{W} \boldsymbol{z}$$
- onde $\boldsymbol{z} = \boldsymbol{X}\,\theta^{\text{old}} + \boldsymbol{W}^{-1}(\boldsymbol{y} - \boldsymbol{p})$
- Se $\boldsymbol{z}$ é visto como um vetor resposta e X uma matriz de regressores, então θ^{new} é a solução de um problema de mínimos quadrados ponderados:
$$\theta^{\text{new}} \leftarrow \arg\min_{\theta} (\boldsymbol{z} - \boldsymbol{X}\,\theta)^t \boldsymbol{W} (\boldsymbol{z} - \boldsymbol{X}\,\theta)$$
- OBS: Mínimos quadrados ordinários (não-ponderado) resolve
$$\arg\min_{\theta} (\boldsymbol{z} - \boldsymbol{X}\,\theta)^t (\boldsymbol{z} - \boldsymbol{X}\,\theta)$$
- Assim, Newton-Raphson reduz-se a uma série iterativa de mínimos quadrados.
- Este algoritmo é o *Iteratively Reweighted Least Squares* (IRLS).

Algoritmo IRLS - pseudo código - opcional

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

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

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

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

-

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

Logística para classificação

- O modelo de regressão logística pode ser usado para classificação.
- Dados de treinamento (amostra) baseados em dados históricos.
- n exemplos: $(y_1, \boldsymbol{x}_1), (y_2, \boldsymbol{x}_2), \ldots, (y_n, \boldsymbol{x}_n)$
- y é v.a. binária. Sua chance de ser 0 ou 1 não é constante.
- $\mathbb{P}(y = 1)$ é uma função de $\boldsymbol{x}$, depende de $\boldsymbol{x}$.
- Queremos encontrar uma função h que tenha $\boldsymbol{x}$ como argumento e que tenha o rótulo y como resposta: $h(\boldsymbol{x}) = y$).
- Objetivo: usar h em FUTUROS casos para predizer o valor do rótulo y.
- Isto é, no futuro, teremos apenas $\boldsymbol{x}$ e queremos predizer o valor de y.

Exemplos

- Spam versus não-spam (resposta y) com base em atributos da mensagem (os regressores $\boldsymbol{x}$, na forma de um vetor de tamanho fixo calaculado em cima de cada mensagem).
- Classificação automática de imagens de mamografia como produzindo um diagnóstico inicial positivo ou negativo de câncer de mama. Como atributos são usados descritores quantitativos que podem ser automaticamente extraídos de uma imagem de mamografia.
- Mensagens de fóruns/blogs: Análise de sentimentos (positivo ou negativo) em relação a certos produtos. Atributos: frequência de certas palavras no texto.
- Crânios classificados como M ou F com base em medições nos

Diabetes data set

- Descobrir se um paciente tem diabetes ou não (y) com base em um conjunto de atributos.
- Number of times pregnant
- Plasma glucose concentration a 2 hours in an oral glucose tolerance test
- Diastolic blood pressure (mm Hg)
- Triceps skin fold thickness (mm)
- 2-Hour serum insulin (mu U/ml)
- Body mass index (weight in kg/(height in m)^2)
- Diabetes pedigree function
- Age (years)

Exemplos

- From UCI machine learning database repository: `http://archive.ics.uci.edu/ml/datasets/Pima+Indians+Diabetes`
- 768 samples in the dataset
- regressores: 8 quantitative variables
- 2 classes; with or without signs of diabetes
- Missing values? There are zeros in places where they are biologically impossible, such as the blood pressure attribute.
- Por conveniência de visualização vamos primeiro considerar dois indicadores criados a partir dos oito regressores. first two principal components as the new feature variables.
- São os dois primeiros componentes principais: veremos mais a frente no curso.

Classificação e logística

- Deseja-se obter uma função que seja, aproximadamente, a probabilidade de ser diabético ($y = 1$)com base nos atributos x_1 e x_2.
- Queremos isto para predizer o valor da classe y nos casos em que tivermos APENAS os atributos x_1 e x_2.
- Se $\mathbb{P}(y = 1) > 1/2$ vamos predizer (ou classificar) $y = 1$. Se $\mathbb{P}(y = 1) \leq 1/2$, vamos predizer que $y = 0$.
- Vamos modelar $\mathbb{P}(y = 1) = 1/(1 + \exp(-(\beta_0 + \beta_1 x_1 + \beta_2 x_2)))$.
- O que significa $\mathbb{P}(y = 1) > 1/2$ neste modelo?

Classificação com logística: geometria

- Suponha que classificamos uma observação como SUCESSO (1) caso

$$\mathbb{P}(Y = 1|\boldsymbol{x}) = 1/(1 + \exp(-(\beta_0 + \beta_1 x_1 + \beta_2 x_2))) \geq p$$

onde $p \in (0, 1)$. Usualmente, vamos tomar $p = 1/2$.

- Manipulando:

$$1/(1 + \exp(-(\beta_0 + \beta_1 x_1 + \beta_2 x_2))) \geq p$$

se, e somente se,

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2 \geq \log\left(\frac{p}{1-p}\right)$$

- CASO $\beta_2 > 0$, a desigualdade não se altera ao dividir por β_2 e teremos

$$x_2 \geq \frac{1}{\beta_2}\left(\log\left(\frac{p}{1-p}\right) - \beta_0\right) - \frac{\beta_1}{\beta_2} x_1$$

- Por exemplo, se $p = 1/2$, temos $\log\left(\frac{p}{1-p}\right) = 0$ e

Classificação com logística: geometria

- Assim, com $p = 1/2$ e $\beta_2 > 0$, classificamos como SUCESSO aqueles pontos em que os atributos (x_1, x_2) satisfazem

$$x_2 \geq -\frac{\beta_0}{\beta_2} - \frac{\beta_1}{\beta_2} x_1$$

- Se $\beta_2 > 0$, a desiguladade será $\leq$.
- Em qualquer caso, será um semi-plano de (x_1, x_2) determinado pela expressão linear acima.
- Num dos seus lados, classificamos $y = 1$. No outro, classificamos $y = 0$.
- Após obter o MLE de $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)$ podemos determinar estas regiões do plano.

Diabetes Data Set

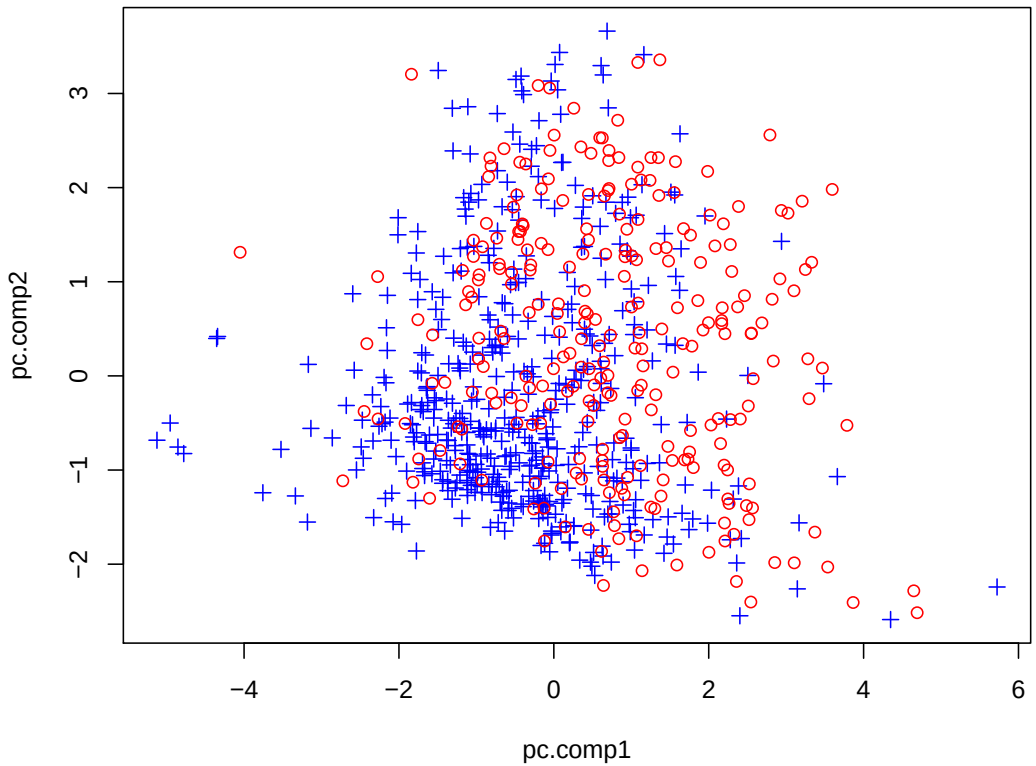

Figura: Diabetes data set. The red circles are Class 1 (with diabetes), and the blue circles are Class 0 (non diabetes).

EMV

- Em R, o comando `glm` implementa a regressão logística.
- `flrm <- glm(y ~ x1 + x2 , family=binomial("logit"))`
- Os valores ajustados de y são as probabilidades

$$\widehat{p}_i = \frac{1}{1 + \exp(-\widehat{\beta}_0 - \widehat{\beta}_1 x_{i1} - \widehat{\beta}_2 x_{i2})}$$

- Classifique o ponto amostral como Classe 1 se seu valor ajustado é > 0.5.
- Caso contrário, classifique como Classe 0.
- O resultado (ver script R) foi $\widehat{\beta} = (-0.77, 0.68, 0.37)$ o que implica em

$$\widehat{p}_i = \frac{1}{1 + \exp(0.77 - 0.68 x_{i1} - 0.37 x_{i2})}$$

EMV

- Assim, se

$$\widehat{p}_i = \frac{1}{1 + \exp(0.77 - 0.68x_{i1} - 0.37x_{i2})} > \frac{1}{2}$$

ou, equivalentemente, se

$$\widehat{\eta}_i = 0.77 - 0.68x_{i1} - 0.37x_{i2} > 0\,,$$

nós classificamos a observação i como sucesso (diabetes).
- Caso contrário, classificamos como fracasso (não-diabetes).
- O gráfico de dispersão a seguir mostra a classificação de todos os pontos pelo ajuste da regressão logística.
- Os círculos vermelhos são os classificados como 1 (diabetes), e as cruzes azuis são aqueles classficados como 0 (não-diabetes).
- Veja a nítida separação linear criada no plano dos regressores x_1 e x_2.
- A fronteira de separação das classes é a reta $0.77 - 0.68x_1 - 0.37x_2 = 0$. Verifique isto graficamente.

Diabetes Data Set

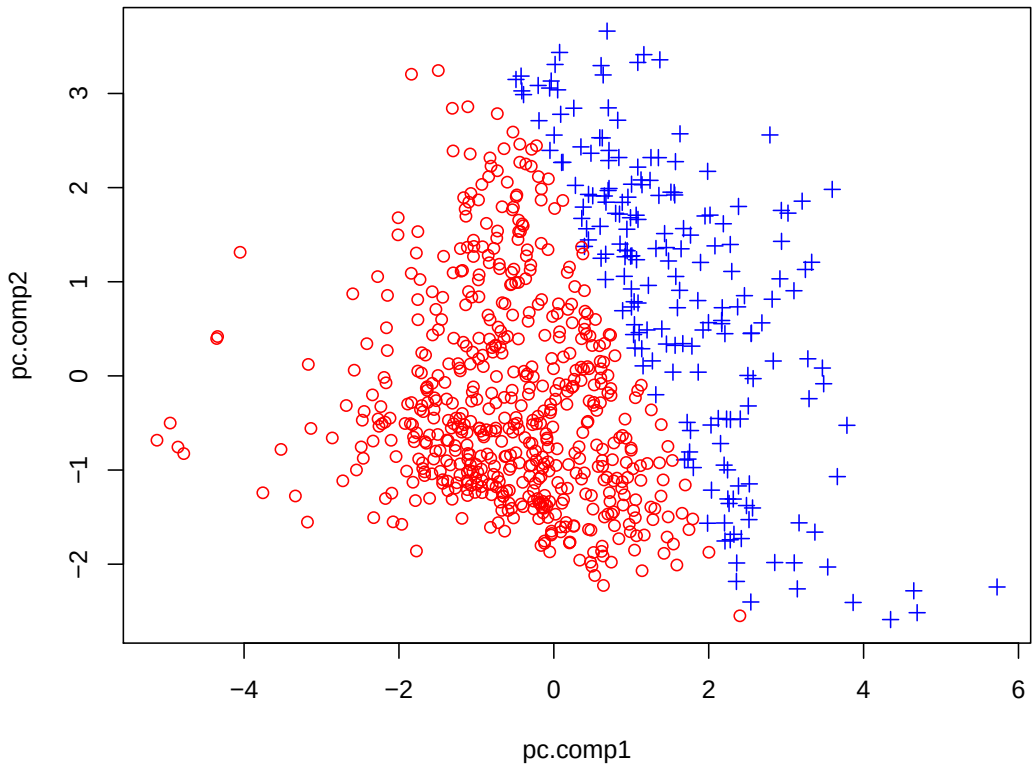

Figura: FITTED classification. The red circles are Class 1 (with diabetes), and the blue crosses are Class 0 (non diabetes).

Diabetes Data Set

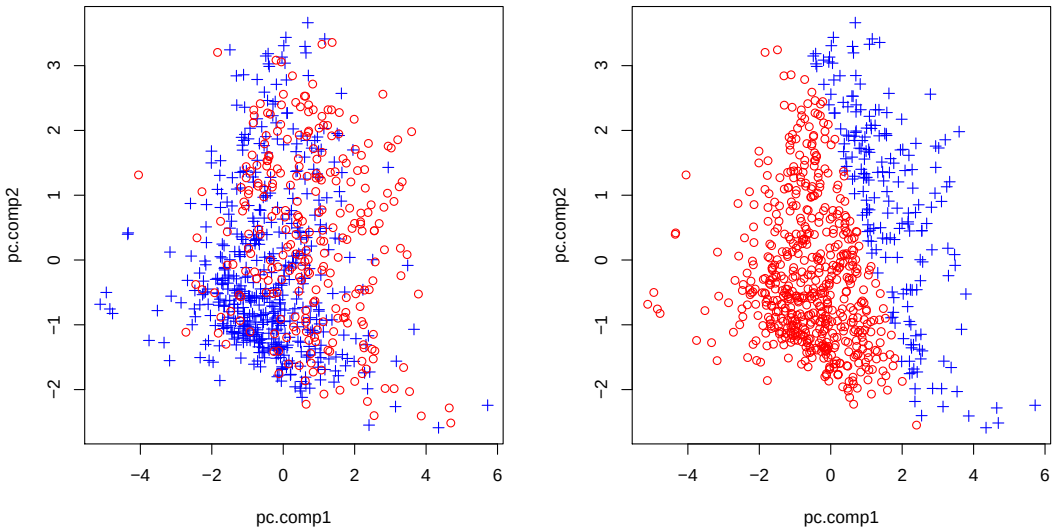

Figura: Diabetes data set: comparing true class and fitted class. The red circles are Class 1 (with diabetes), and the blue crosses are Class 0 (non diabetes).

Métricas para avaliar a regra de classificação

- A classificação feita pela nossa regra de decisão (baseda na regressão logística não é perfeita.
- Ela comete vários erros: indivíduoa que de fato são diabéticos não possuem as características x_1 e x_2 típicas de um diabético.
- Em consequência, a nossa regra de decisão (que olha apenas os regressores em $\boldsymbol{x}$) aloca estes indivíduos à classe 0 (não diabéticos).
- Estes são os *falso-negativos* (o diagnóstico é falsamente negativo).
- Analogamente, vários não-diabéticos possuem características típicas de diabéticos e são então alocados pela regra de decisão logística à categora 1 (diabéticos).
- Estes são os *falso-positivos* (o diagnóstico é falsamente positivo).
- Claro, existe o conceito de *verdadeiro-positivo* e *verdadeiro-negativo*.

Falso-positivos e Falso-negativos

- Idealmente, queremos poucos falso-positivos e poucos falso-negativos (ou muitos verdadeiro-positivos e muitos verdadeiro-negativos).
- Isto será obtido se tivermos uma pequena probabilidade de ter um falso-positivo (FP) e um falso-negativo (FN).

$$\mathbb{P}(FP) = \mathbb{P}(\text{classificado como } + | \text{é -}) = \frac{\mathbb{P}(\text{classif } + \text{ e é -})}{\mathbb{P}(\text{é -})}$$

e

$$\mathbb{P}(FN) = \mathbb{P}(\text{classificado como } - | \text{é } +) = \frac{\mathbb{P}(\text{classif - e é } +)}{\mathbb{P}(\text{é -})}$$

- No caso de verdadeiro-positivos , temos

$$\mathbb{P}(VP) = \mathbb{P}(\text{classificado como } + | \text{é } +) = \frac{\mathbb{P}(\text{classif } + \text{ e é } +)}{\mathbb{P}(\text{é } +)}$$

Recall ou revocação ou sensibilidade

- No caso de verdadeiro-positivos , temos

$$\mathbb{P}(VP) = \mathbb{P}(\text{classificado como } + | \text{é } +) = \frac{\mathbb{P}(\text{classif } + \text{ e é } +)}{\mathbb{P}(\text{é } +)}$$

- Esta probabilidade (estimada) é chamada de RECALL (revocação) em aprendizado de máquina ou de sensibilidade ou sensitividade em estudos epidemiológicos.
- Recall alto significa que o algoritmo retornou a maioria dos resultados relevantes.

Verdadeiro-negativos ou especificidade

- Quanto aos verdadeiro-negativos,

$$\mathbb{P}(VN) = \mathbb{P}(\text{classificado como -}|\text{é -}) = \frac{\mathbb{P}(\text{classif - e é -})}{\mathbb{P}(\text{é -})}$$

- Esta medida é chamada de *especificidade*.
- A idéia é que o algoritmo é específico para o que ele se propõe classficar.
- Se o item não é $+$, ele não retorna $+$.
- Veja que $\mathbb{P}(VN) + \mathbb{P}(FP) = 1$ pois um indivíduo que é negativo, será classificado ou como negativo (corretamente) ou como positivo (falsamente).
- Do mesmo modo, $\mathbb{P}(VP) + \mathbb{P}(FP) = 1$.

Estimando falso-positivos e falso-negativos

- Estimamos estas quantidades a partir dos dados comparando a verdadeira classe dos exemplos com a classe alocada a eles pela regressão logística.

	Diag -	Diag +
é -	429	71
é +	145	123

- Assim, o RECALL é estimado como

$$\mathbb{P}(VP) \approx \frac{123/768}{(145+123)/768} = \frac{123}{145+123} = 0.47$$

- Estamos acertando no diagnóstico de aprox metade dos verdadeiramente diabéticos.
- $\mathbb{P}(VN) \approx 429/(429+71) = 0.86$: acertamos mais frequentemente no diagnóstico dos verdadeiramente não-diabéticos.

Precisão, recall e especificidade

- Em aprendizado de máquina, uma métrica muito comum inverte os eventos usados na definição do RECALL.
- Temos RECALL igual a $\mathbb{P}(VP) = \mathbb{P}(\text{classificado como } +|\text{é } +)$.
- A PRECISÃO de um algoritmo de classificação é dada por

$$\text{Precisão} = \mathbb{P}(\text{é } +|\text{classificado como } +)$$

.

- Alta precisão indica que um algoritmo retornou mais resultados relevantes que irrelevantes.
- A partir da tabela anterior, podemos estimar a precisão como $123/(123 + 71) = 0.63$.
- Mais uma métrica, especificidade ($\mathbb{P}(VN) = \mathbb{P}(\text{classif -}|\text{e' -})$), estimada como $429/(429 + 71) = 0.86$.

Usando os 8 atributos

- Ao invés dos dois componentes principais, vamos usar os oito atributos originais.
- Ver script R