Modelos Estatísticos I
Modelos para desfecho contínuo

Leo Bastos – leonardo.bastos@fiocruz.br

PROCC – Fundação Oswaldo Cruz

https://github.com/lsbastos/eae2

Modelo Linear

Na aula de hoje vamos rever:

- O modelo linear
- Análise de resíduos

Modelo linear

- Seja $Y_1, Y_2, \ldots, Y_n$ variáveis aleatórias independentes
- Assume-se que
$$Y_i \sim N(\mu_i, \sigma^2)$$
onde $\mu_i = \boldsymbol{x}_i^T \boldsymbol{\beta} = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi}$
- E o modelo é usualmente escrito como
$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}, \qquad \boldsymbol{e} \sim N_n(0, \sigma^2 I_n)$$
onde
$$\boldsymbol{y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}, \quad \boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_1^T \\ \vdots \\ \boldsymbol{x}_n^T \end{bmatrix}, \quad \boldsymbol{e} = \begin{bmatrix} e_1 \\ \vdots \\ e_n \end{bmatrix}$$

Inferência para os parâmetros β

- $\hat{\boldsymbol{\beta}}$ pode ser encontrado por máxima verossimilhança e mínimos quadrados
$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$$
desde que $\boldsymbol{X}^T\boldsymbol{X}$ seja singular.
- $\hat{\boldsymbol{\beta}}$ é não viciado, i.e.
$$\mathbb{E}[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}.$$
- a variância de $\hat{\boldsymbol{\beta}}$ é dada por
$$\mathbb{V}[\hat{\boldsymbol{\beta}}] = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$$
- E, finalmente, σ^2 é estimado usando
$$\hat{\sigma}^2 = \frac{1}{N-p}(\boldsymbol{y} - \boldsymbol{X}\hat{\beta})^T(\boldsymbol{y} - \boldsymbol{X}\hat{\beta})$$

1. O IC de 95% para o coeficiente β_i é dado por:

$$\left[\hat{\beta}_i - 1.96\sqrt{sd(\hat{\beta}_i)}; \quad \hat{\beta}_i + 1.96\sqrt{sd(\hat{\beta}_i)}\right]$$

   onde o desvio padrão de $\hat{\beta}_i$ é i-ésimo valor da diagonal da matriz $\hat{\sigma}^2 = \boldsymbol{X}^T\boldsymbol{X}$.

2. Se quisermos testar as hipóteses H_0: $\beta_i = b$ versus H_1: $\beta_i \neq b$, basta calcular a estatística

$$Z = \frac{\hat{\beta}_i - b}{sd(\hat{\beta})}$$

   e verificar onde esse valor ocorre em uma distribuição normal padrão. Calculando, por exemplo, o valor-p (ou p-value). (Teste de Wald)

FIOCRUZ

Deviance

- Nelder & Wedderburn (1972): $D = 2[l(\hat{\boldsymbol{\beta}}_{max}) - l(\hat{\boldsymbol{\beta}})]$
- Modelo normal:

$$D = \frac{1}{\sigma^2}(\boldsymbol{y} - \boldsymbol{X}\hat{\beta})^T(\boldsymbol{y} - \boldsymbol{X}\hat{\beta})$$

- Analise de variância de modelos consiste em um teste F, uma vez que a normal tem um parâmetro extra σ^2.
- No R:
  anova(modelo, test="F")
  (teste default se modelo for um objeto *lm*)

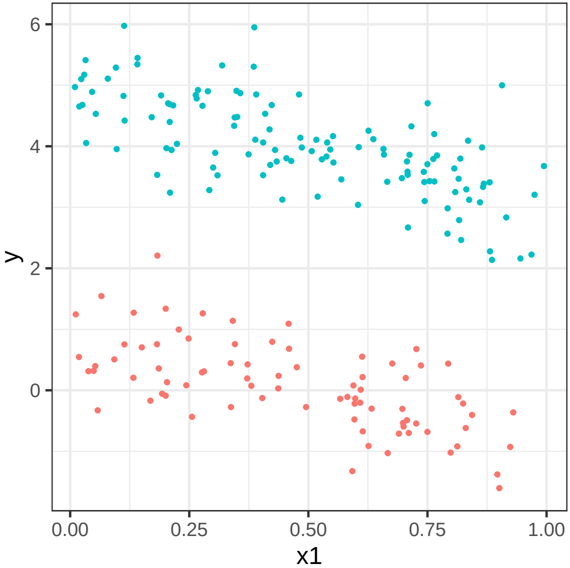
FIOCRUZ

Exemplo: Normal

```
> n = 200
> p = 3
> X1 <- runif(n)
> X2 <- rbinom(n, 1, 0.5)
> dados <- data.frame(x1 = X1,
+                     x2 = X2,
+                     y = 1 - 2*X1 + 4*X2 + rnorm(n,0,.5)
+                     )
> dados$y[200] = 5
> # Modelo com a funcao lm
> lm.ex = lm(y ~ x1 + x2, data = dados)
> # Modelo com a funcao glm
> glm.ex = glm(y ~ x1 + x2, family=gaussian(), data = dados)
```

FIOCRUZ

Exemplo: Normal



Saída R: lm()

```
Call:
lm(formula = y ~ x1 + x2, data = dados)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.29593 -0.36207 -0.01761  0.35400  1.76316 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.94768    0.09344   10.14   <2e-16 ***
x1          -1.86365    0.14724  -12.66   <2e-16 ***
x2           3.97858    0.08186   48.60   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5665 on 197 degrees of freedom
Multiple R-squared:  0.9261,	Adjusted R-squared:  0.9253 
F-statistic:  1234 on 2 and 197 DF,  p-value: < 2.2e-16
```

Saída R: glm()

```
Call:
glm(formula = y ~ x1 + x2, family = gaussian(), data = dados)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.94768    0.09344   10.14   <2e-16 ***
x1          -1.86365    0.14724  -12.66   <2e-16 ***
x2           3.97858    0.08186   48.60   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 0.3208974)

    Null deviance: 855.326  on 199  degrees of freedom
Residual deviance:  63.217  on 197  degrees of freedom
AIC: 345.23

Number of Fisher Scoring iterations: 2
```

Análise de variância no R

```
> anova(lm.ex)
Analysis of Variance Table

Response: y
           Df Sum Sq Mean Sq F value    Pr(>F)
x1          1  34.10   34.10  106.28 < 2.2e-16 ***
x2          1 758.01  758.01 2362.14 < 2.2e-16 ***
Residuals 197  63.22    0.32
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Análise da Deviance no R

```
> anova(glm.ex, test="F")
Analysis of Deviance Table

Model: gaussian, link: identity

Response: y

Terms added sequentially (first to last)


     Df Deviance Resid. Df Resid. Dev       F    Pr(>F)
NULL                   199     855.33
x1    1    34.10       198     821.22  106.28 < 2.2e-16 ***
x2    1   758.01       197      63.22 2362.14 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

FIOCRUZ

Suposições

- Suposições para a inferência
  1. Independência
  2. Normalidade
  3. Homocedasticidade (variância constante)
  4. Linearidade
- Como podemos verificar essas suposições?
  - Vamos estudar os resíduos, ou erros, do modelo

$$r_i = f(y_i, \hat{\mu}_i)$$

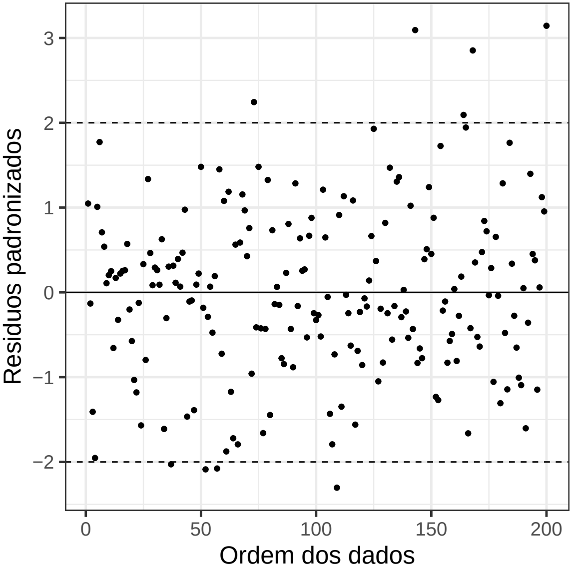
  - Análises gráficas e testes formais.

Alguns possíveis resíduos

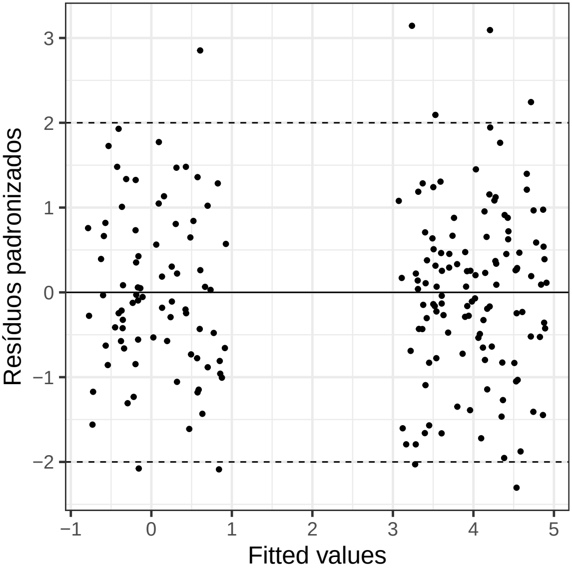
- Resíduo ordinário
$$r_i = y_i - \hat{\mu}_i$$
- Resíduos de Pearson padronizado (resíduo studentizado)
$$r_i = \frac{y_i - \hat{\mu}_i}{\sqrt{\mathbb{V}[\hat{\mu}_i](1 - h_i)}}$$
  onde h_i é o i-ésimo elemento da matriz $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$
  No R: *rstandard(modelo)*
- Esses resíduos podem ser usados para checar:
  1. independência serial
  2. Linearidade
  3. Normalidade
  4. Associação com variáveis não incluídas no modelo

FIOCRUZ

Exemplo: Checando independência serial



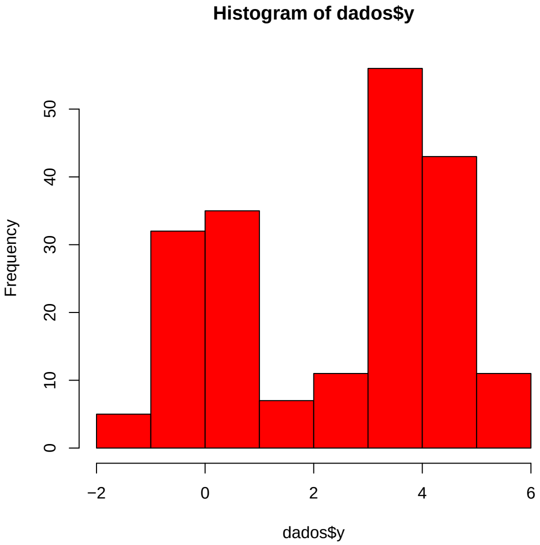
Exemplo: Checando linearidade



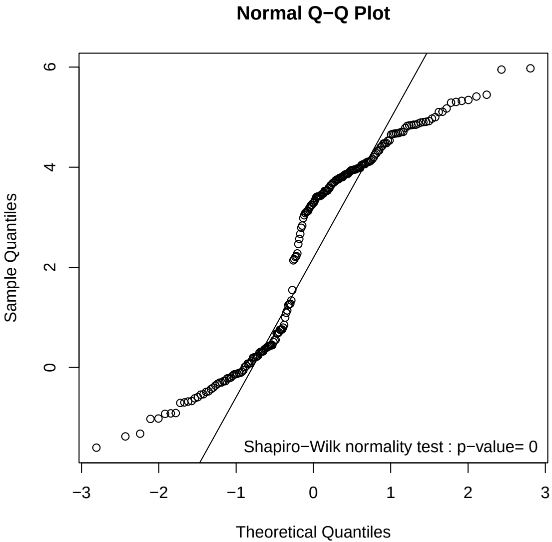
Testando normalidade

- Análise visual:
  - Histograma (buscando distribuição em forma de sino)
  - quantile-quantile plot (QQnorm) (Busca por uma reta com os quantis da normal)
- Teste de hipótese (H_0: Dados segue uma distribuição normal)
  - Teste Shapiro-wilks (*?shapiro.test*)
  - Teste Komolgorov-Smirnov (*?ks.test*)

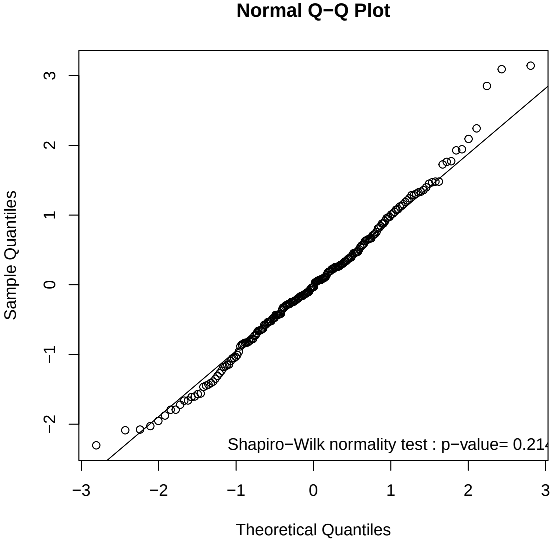
Exemplo: Checando normalidade do desfecho



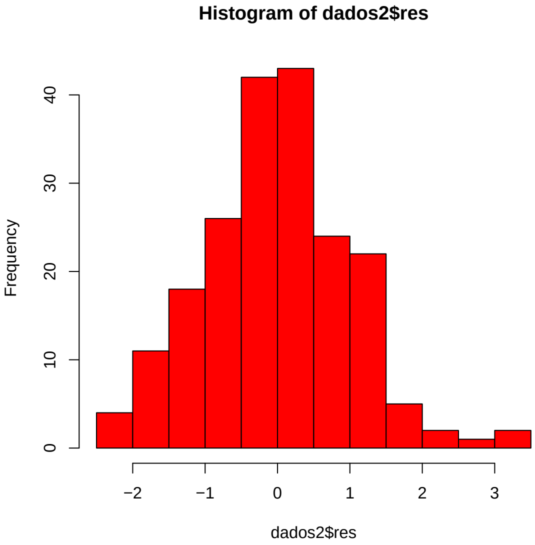
Exemplo: Checando normalidade do desfecho



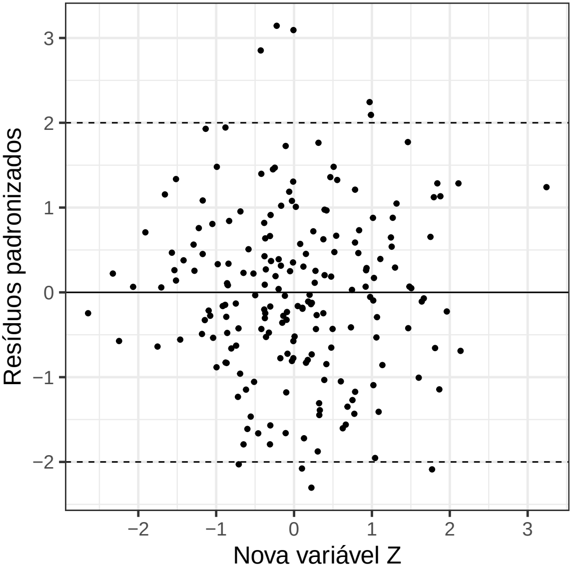
Exemplo: Checando normalidade dos resíduos



Exemplo: Checando normalidade do resíduo



Exemplo: Checando associação com uma nova variável



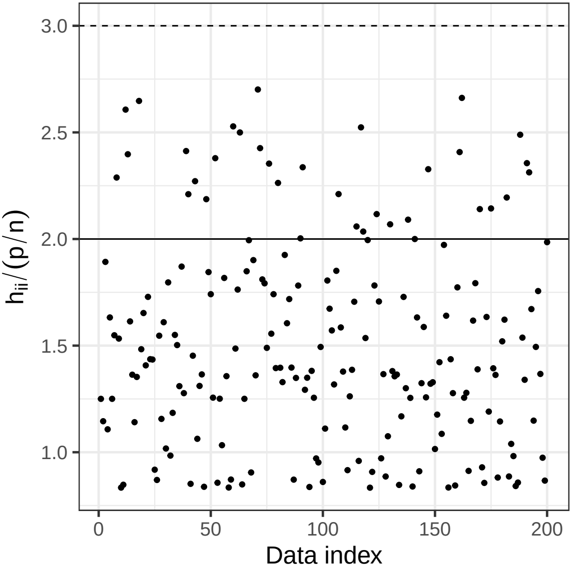
Outliers ou pontos influentes

- Leverage (pontos de alavanca): h_{ii} (*hatvalues(modelo)*)

$$H = X^T (X^T X)^{-1} X^T$$

  Valores h_{ii} maiores que 2 ou 2 vezes p/n merecem uma olhada.
- Leave-one-out measures:
  - DFFIT: Diferença nos ajustes: $\hat{y}_i - \hat{y}_{i(-i)}$
  - DFBETA: Diferença no ajuste de cada coeficiente: $\hat{\beta}_k - \hat{\beta}_{k(-i)}$
  - Distância de Cook: Diferença no ajuste em todos coeficientes

Pontos de alavanca (hatvalues)



Outliers ou pontos influentes: LOO

- DFFITS (*dffits(modelo)*) (Belsley et al. 1980)

$$DFFITS_i = \frac{\hat{y}_i - \hat{y}_i}{s_{(i)\sqrt{h_{ii}}}} = \left(\frac{h_{ii}}{1-h_{ii}}\right)^{1/2}$$

Investigar valores maiores que $2\sqrt{p/n}$

- DFBETAS (*dfbetas(modelo)*) (Belsley et al. 1980)

$$DFBETAS_i = \hat{\hat{\beta}}_k - \hat{\hat{\beta}}_{k(-i)} = \frac{(X^TX)^{-1}x_i^T(y_i - \hat{y}_i)}{1-h_{ii}}$$

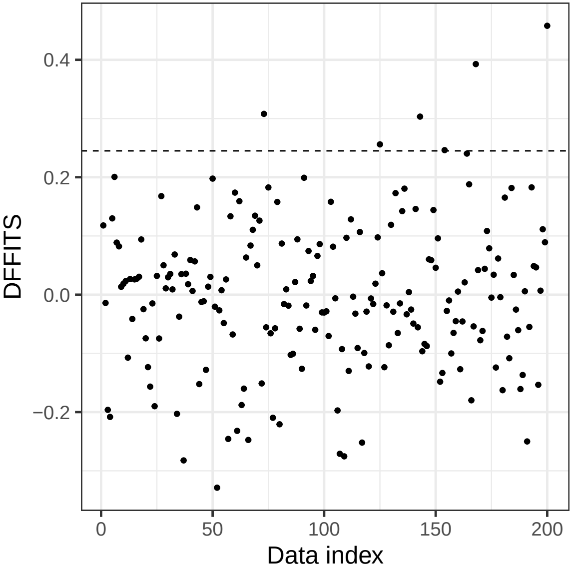
- Distância de Cook (*cooks.distance(modelo)*) (Cook 1977,1979)

$$D_i = \frac{1}{p}\left(\frac{h_{ii}}{1-h_{ii}}\right) r_i^2$$

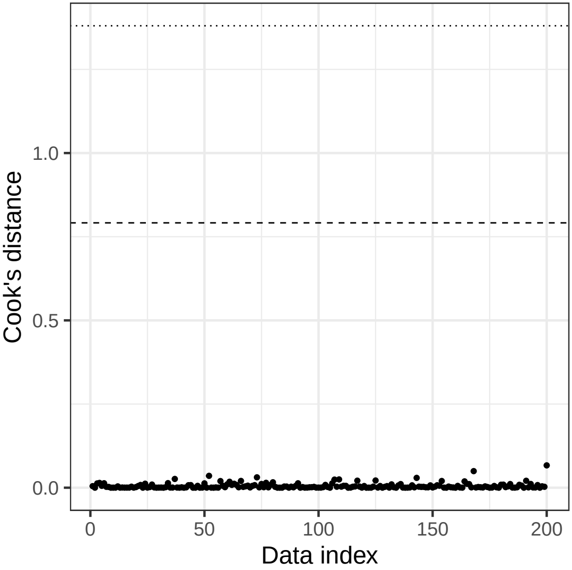
Cook (1979) mostrou que D_i deve ser comparado com uma $F(p, n-p)$

FIOCRUZ

Outliers: DFFITS



Outliers: Distância de Cooks



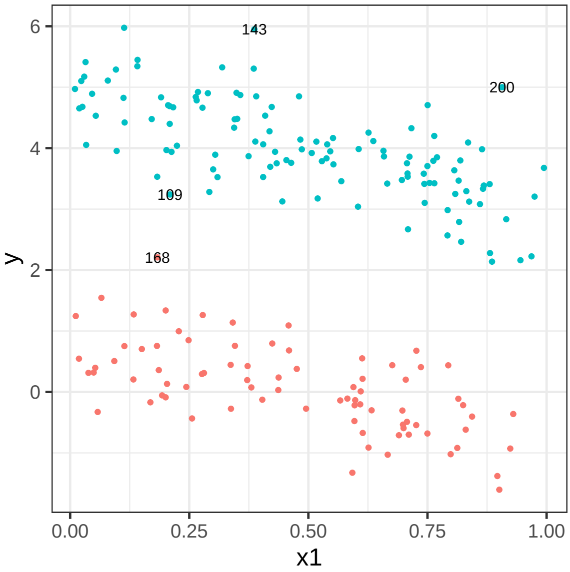
FIOCRUZ

Outliers: Medidas de influência

```
> summary(influence.measures(lm.ex))
Potentially influential observations of
	 lm(formula = y ~ x1 + x2, data = dados) :

    dfb.1_ dfb.x1 dfb.x2 dffit   cov.r   cook.d hat
109 -0.13   0.17  -0.14  -0.28    0.95_*  0.02   0.01
143  0.06  -0.09   0.19   0.30    0.88_*  0.03   0.01
168  0.38  -0.22  -0.24   0.39_*  0.91_*  0.05   0.02
200 -0.26   0.35   0.17   0.46_*  0.89_*  0.07   0.02
```

Potenciais pontos influentes



Resumo: Análise gráfica de resíduos

1. Visão geral. (p.e. Histograma)
2. Resíduos versus sequência observada.
3. Resíduos versus valores ajustados.
4. Resíduos versus variáveis explicativas.
5. Resíduos versus qualquer outra forma particular que seja razoável para checar as suposições do modelo.
6. Busca por medidas influência (outliers ou pontos de alavanca)