Diagnóstico, Remediação e Validação

Nossos modelos lineares são baseados em uma série de pressuposições, a lembrar:

1. Os valores/níveis de X são medidos sem erro

Diagnóstico, Remediação e Validação

Nossos modelos lineares são baseados em uma série de pressuposições, a lembrar:

1. Os valores/níveis de X são medidos sem erro
2. Existe uma relação linear entre X e Y (só para regressão)

Diagnóstico, Remediação e Validação

Nossos modelos lineares são baseados em uma série de pressuposições, a lembrar:

1. Os valores/níveis de X são medidos sem erro
2. Existe uma relação linear entre X e Y (só para regressão)
3. Os erros ϵ (e por consequência, Y) tem variância constante (σ^2)

Diagnóstico, Remediação e Validação

Nossos modelos lineares são baseados em uma série de pressuposições, a lembrar:

1. Os valores/níveis de X são medidos sem erro
2. Existe uma relação linear entre X e Y (só para regressão)
3. Os erros ϵ (e por consequência, Y) tem variância constante (σ^2)
4. Os erros ϵ (e por consequência, Y) são independentes

Diagnóstico, Remediação e Validação

Nossos modelos lineares são baseados em uma série de pressuposições, a lembrar:

1. Os valores/níveis de X são medidos sem erro
2. Existe uma relação linear entre X e Y (só para regressão)
3. Os erros ϵ (e por consequência, Y) tem variância constante (σ^2)
4. Os erros ϵ (e por consequência, Y) são independentes
5. Os erros ϵ (e por consequência, Y) são normalmente distribuídos:
    - $\epsilon \sim N(0, \sigma^2)$
    - $Y \sim N(\beta_0 + \beta_1 X, \sigma^2)$

Diagnóstico e Remediação

Muitas vezes, estas pressuposições não correspondem à realidade. Por esta razão, todo processo de modelagem inclui etapas de diagnóstico, remediação, e validação:

Diagnóstico e Remediação

Muitas vezes, estas pressuposições não correspondem à realidade. Por esta razão, todo processo de modelagem inclui etapas de diagnóstico, remediação, e validação:

Diagnóstico: processo de avaliação da adequação dos dados e resultados às pressuposições do modelo.

Diagnóstico e Remediação

Muitas vezes, estas pressuposições não correspondem à realidade. Por esta razão, todo processo de modelagem inclui etapas de diagnóstico, remediação, e validação:

Diagnóstico: processo de avaliação da adequação dos dados e resultados às pressuposições do modelo.

Remediação: processo de melhoria da adequação dos dados e resultados às pressuposições do modelo.

Diagnóstico e Remediação

Muitas vezes, estas pressuposições não correspondem à realidade. Por esta razão, todo processo de modelagem inclui etapas de diagnóstico, remediação, e validação:

Diagnóstico: processo de avaliação da adequação dos dados e resultados às pressuposições do modelo.

Remediação: processo de melhoria da adequação dos dados e resultados às pressuposições do modelo.

Validação: processo de verificação da performance do modelo na explicação/previsão do fenômeno de interesse

Análise de resíduos

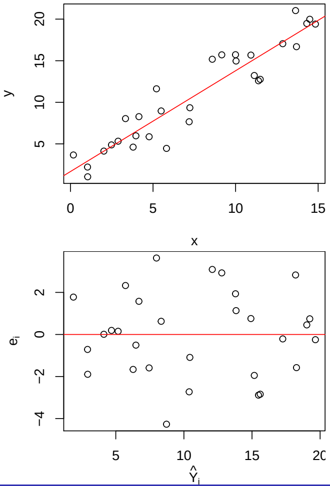
Uma das princpais análises diagnósticas é um scatterplot dos resíduos vs. valores estimados ($\hat{Y}$)

Análise de resíduos

Uma das princpais análises diagnósticas é um scatterplot dos resíduos vs. valores estimados ($\hat{Y}$)

Vejamos um exemplo de um modelo de regressão apropriado:

- Resíduos aletoriamente distribuidos ao redor de zero
- Variação constante ao longo de Y_i

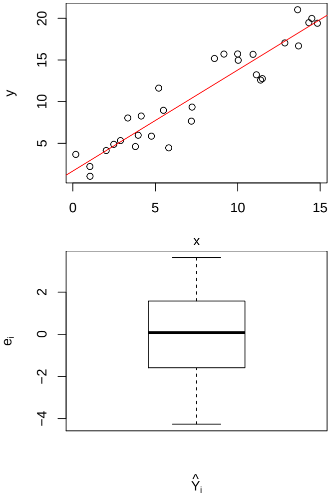


Análise de resíduos

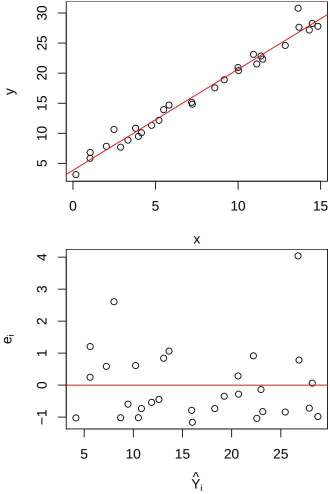
Uma das princpais análises diagnósticas é um scatterplot dos resíduos vs. valores estimados ($\hat{Y}$)

Vejamos um exemplo de um modelo de regressão apropriado:

- Resíduos aletoriamente distribuidos ao redor de zero
- Variação constante ao longo de Y_i



Análise de resíduos



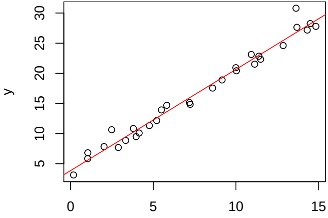
Dados com resíduos não-normais:

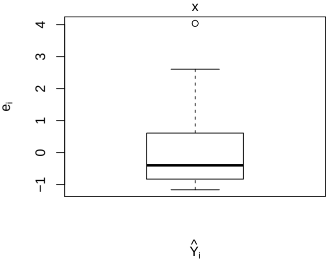
- Resíduos positivos maiores do que resíduos negativos
- Distribuição Assimétrica

Análise de resíduos

Dados com resíduos não-normais:

- Resíduos positivos maiores do que resíduos negativos
- Distribuição Assimétrica





Análise de resíduos

Resíduos não-normais: **Q-Q plot**

- Ferramenta gráfica bastante utilizada para avaliação de aderência à normalidade

Análise de resíduos

Resíduos não-normais: **Q-Q plot**

- Ferramenta gráfica bastante utilizada para avaliação de aderência à normalidade
- Plota os quantis dos dados contra os quantis correspondentes de uma distribuição normal com os mesmos parâmetros ($\bar{X}$, s^2)

Análise de resíduos

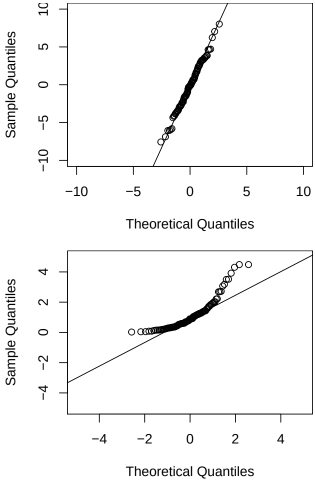
Resíduos não-normais: **Q-Q plot**

- Ferramenta gráfica bastante utilizada para avaliação de aderência à normalidade
- Plota os quantis dos dados contra os quantis correspondentes de uma distribuição normal com os mesmos parâmetros ($\bar{X}$, s^2)
- Quanto mais normal a distribuição, mais "iguais" serão os quantis

Análise de resíduos

Resíduos não-normais: **Q-Q plot**

- Ferramenta gráfica bastante utilizada para avaliação de aderência à normalidade
- Plota os quantis dos dados contra os quantis correspondentes de uma distribuição normal com os mesmos parâmetros ($\bar{X}$, s^2)
- Quanto mais normal a distribuição, mais "iguais" serão os quantis



Análise de resíduos

Dados com resíduos heteroscedásticos:

Análise de resíduos

Dados com resíduos heteroscedásticos:

homoscedástico = variância constante

heteroscedástico = variância inconstante

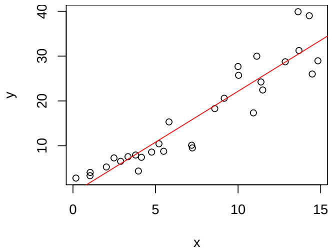
Análise de resíduos

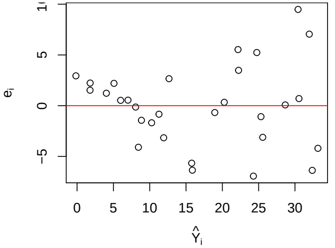
Dados com resíduos heteroscedásticos:

homoscedástico = variância constante

heteroscedástico = variância inconstante

- Resíduos aletoriamente distribuidos ao redor de zero
- Variância dos resíduos aumenta (ou diminui) ao longo de $\hat{Y}$





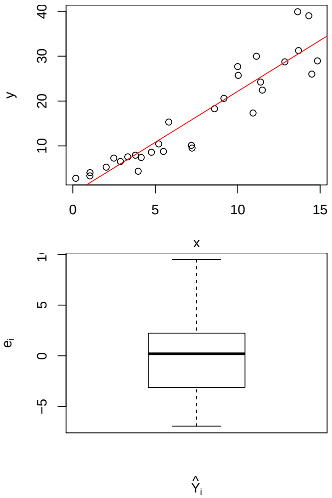
Análise de resíduos

Dados com resíduos heteroscedásticos:

homoscedástico = variância constante

heteroscedástico = variância inconstante

- Resíduos aletoriamente distribuidos ao redor de zero
- Variância dos resíduos aumenta (ou diminui) ao longo de $\hat{Y}$



Análise de resíduos

Dados com resíduos não-independentes:

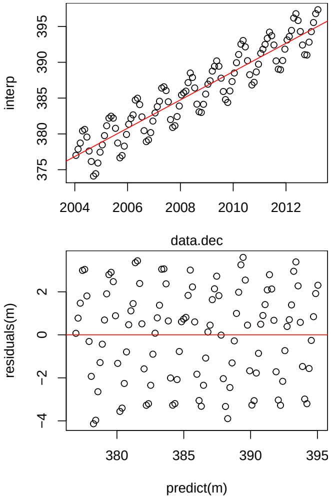
Curva de Keeling

Análise de resíduos

Dados com resíduos não-independentes:

Curva de Keeling
Concentração de CO2 atmosférico medido em Mauna Loa, Hawaii

- Resíduos distribuidos sistematicamente ao redor de zero

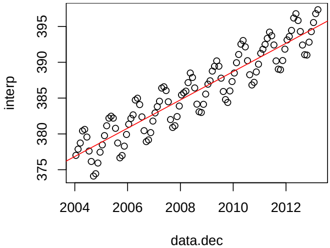


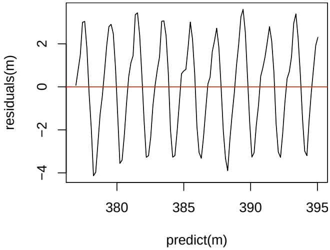
Análise de resíduos

Dados com resíduos não-independentes:

Curva de Keeling
Concentração de CO2 atmosférico medido em Mauna Loa, Hawaii

- Resíduos distribuidos sistematicamente ao redor de zero



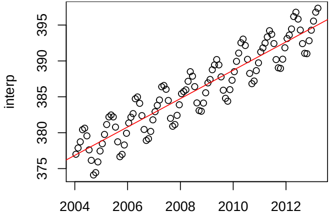


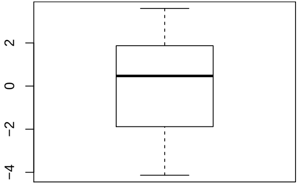
Análise de resíduos

Dados com resíduos não-independentes:

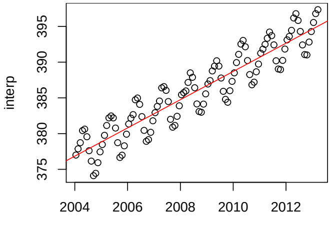
Curva de Keeling
Concetração de CO2 atmosférico medido em Mauna Loa, Hawaii

- Resíduos distribuidos sistematicamente ao redor de zero



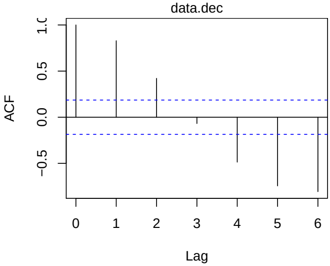


Análise de resíduos



Função de Autocorrelação

Plota a correlação entre X e seus próprios valores, com diferentes lags.



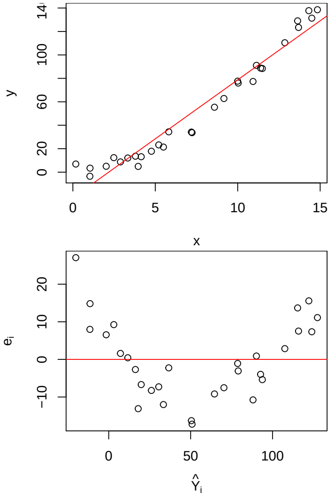
Análise de resíduos

Relação entre X e Y não é linear

Análise de resíduos

Relação entre X e Y não é linear

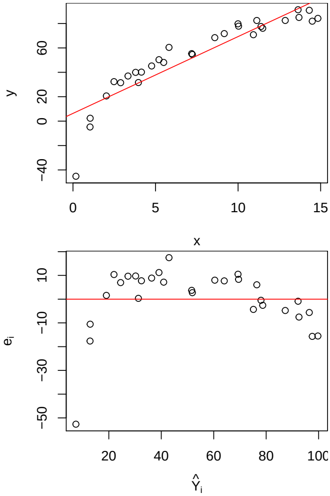
- Resíduos distribuidos segundo um padrão
- O padrão sugere o tipo de relação



Análise de resíduos

Relação entre X e Y não é linear

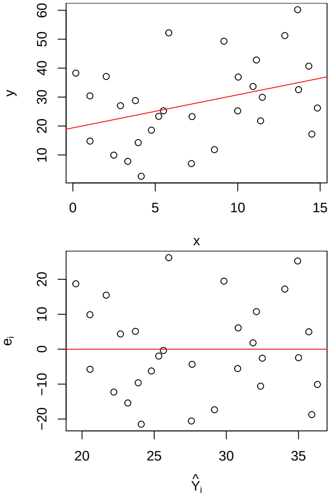
- Resíduos distribuidos segundo um padrão
- O padrão sugere o tipo de relação



Análise de resíduos

Ausência de uma variável explicativa

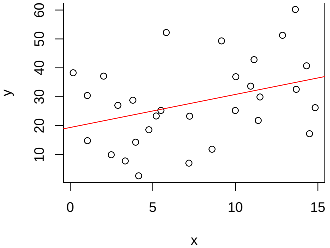
- Grande parte da variância não explicada por X_1
- Resíduos nem sempre revelam um padrão
- Mas pode ocorrer forte relação entre os resíduos e a varável omitida

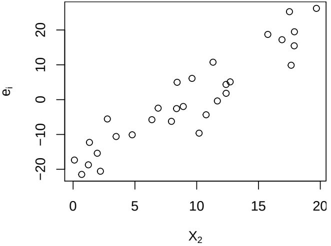


Análise de resíduos

Ausência de uma variável explicativa

- Grande parte da variância não explicada por X_1
- Resíduos nem sempre revelam um padrão
- Mas pode ocorrer forte relação entre os resíduos e a varável omitida





Pontos Influentes

Outro diagnóstico importante é a avaliação do efeito de observações isoladas sobre o ajuste final do modelo linear:

Pontos Influentes

Outro diagnóstico importante é a avaliação do efeito de observações isoladas sobre o ajuste final do modelo linear:

Leverage: Mede o efeito de valores extremos de X. *Leverage* vem de *lever* (alavanca). Valores extremos de X podem alavancar a reta de regressão, que se "equilibra" em $\bar{X}$

Pontos Influentes

Outro diagnóstico importante é a avaliação do efeito de observações isoladas sobre o ajuste final do modelo linear:

Leverage: Mede o efeito de valores extremos de X. *Leverage* vem de *lever* (alavanca). Valores extremos de X podem alavancar a reta de regressão, que se "equilibra" em $\bar{X}$

Distância: Mede o efeito de valores extremos de Y (resíduos extremos).

Pontos Influentes

Outro diagnóstico importante é a avaliação do efeito de observações isoladas sobre o ajuste final do modelo linear:

Leverage: Mede o efeito de valores extremos de X. *Leverage* vem de *lever* (alavanca). Valores extremos de X podem alavancar a reta de regressão, que se "equilibra" em $\bar{X}$

Distância: Mede o efeito de valores extremos de Y (resíduos extremos).

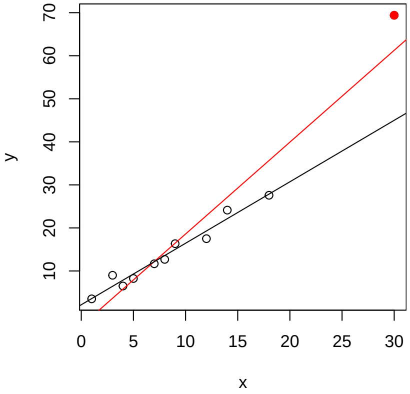
Influência: Combinação de distância e *leverage*, captura efeito total de um *outlier* sobre a reta de regressão

Pontos Influentes

É possível ter um ponto com *leverage* alto, e influência baixa?

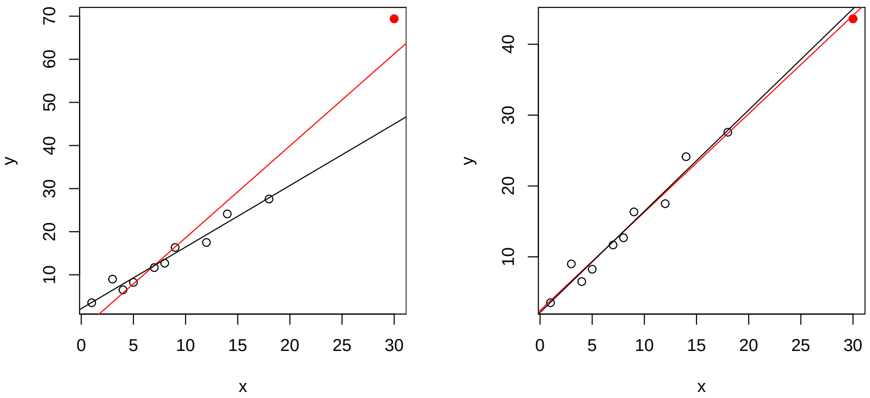
Pontos Influentes

É possível ter um ponto com *leverage* alto, e influência baixa?



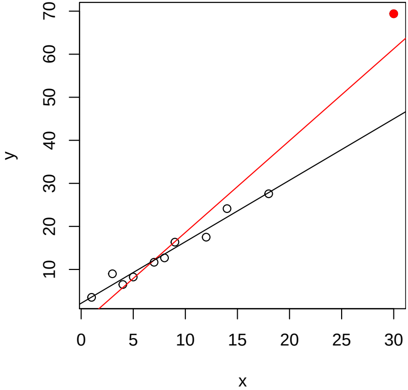
Pontos Influentes

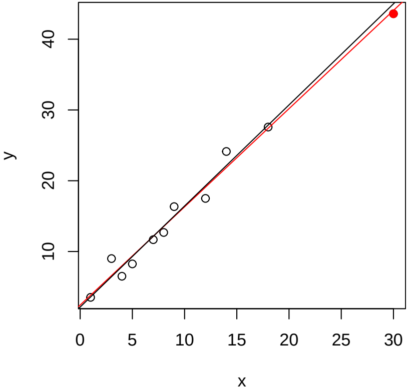
É possível ter um ponto com *leverage* alto, e influência baixa?



Pontos Influentes

É possível ter um ponto com *leverage* alto, e influência baixa?





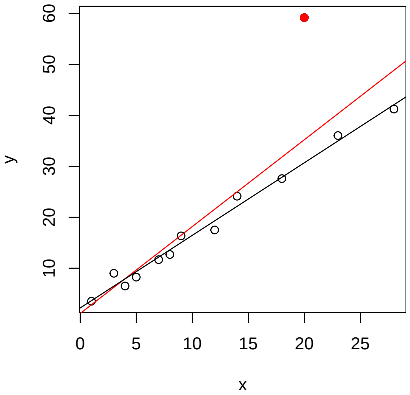
Sim, se a *distância* for baixa.

Pontos Influentes

É possível ter um ponto com *distância* alta, e influência baixa?

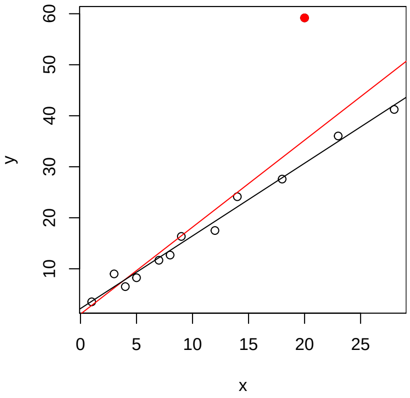
Pontos Influentes

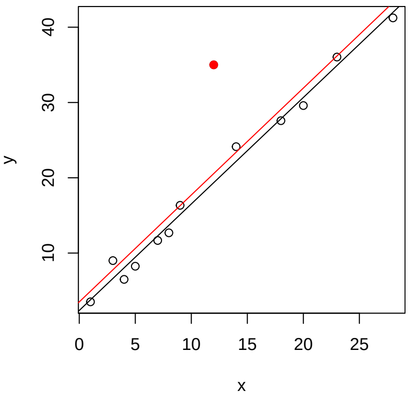
É possível ter um ponto com *distância* alta, e influência baixa?



Pontos Influentes

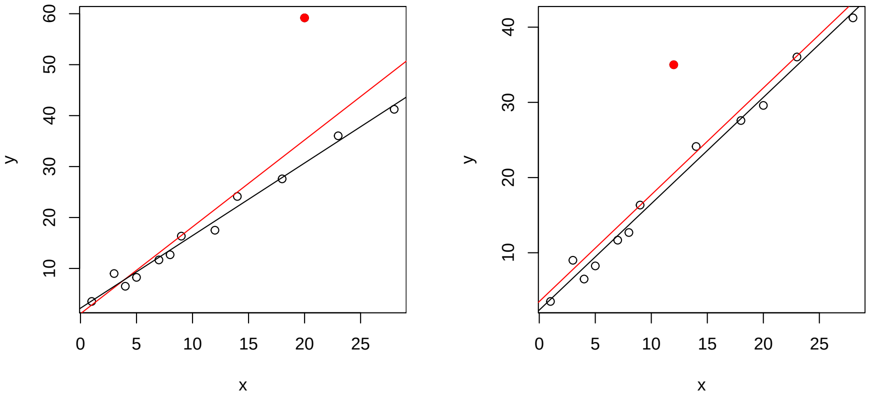
É possível ter um ponto com *distância* alta, e influência baixa?





Pontos Influentes

É possível ter um ponto com *distância* alta, e influência baixa?



Sim, se o *leverage* for baixo.

Pontos Influentes

Como identificar pontos influentes?

Pontos Influentes

Como identificar pontos influentes?

DFFITS: Diferença normalizada entre o valor de $\hat{Y}_i$ no modelo completo, e o valor da mesma estimativa no modelo onde o ponto X_i, Y_i é removido, $\hat{Y}_{i(-i)}$. Identifica pontos com influência sobre estimativas de Y isoladas.

Pontos Influentes

Como identificar pontos influentes?

DFFITS: Diferença normalizada entre o valor de $\hat{Y}_i$ no modelo completo, e o valor da mesma estimativa no modelo onde o ponto X_i, Y_i é removido, $\hat{Y}_{i(-i)}$. Identifica pontos com influência sobre estimativas de Y isoladas.

Distância de Cook: Similar a DFFITS, mas ao invés de avaliar a diferença em um único ponto, avalia a soma dos quadrados das diferenças de todos os $\hat{Y}$. Identifica pontos com influência sobre todas as estimativas de Y.

Pontos Influentes

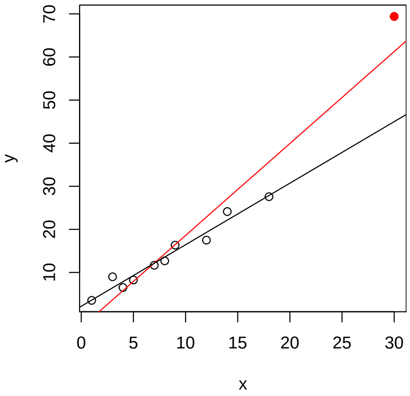
Como identificar pontos influentes?

DFFITS: Diferença normalizada entre o valor de $\hat{Y}_i$ no modelo completo, e o valor da mesma estimativa no modelo onde o ponto X_i, Y_i é removido, $\hat{Y}_{i(-i)}$. Identifica pontos com influência sobre estimativas de Y isoladas.

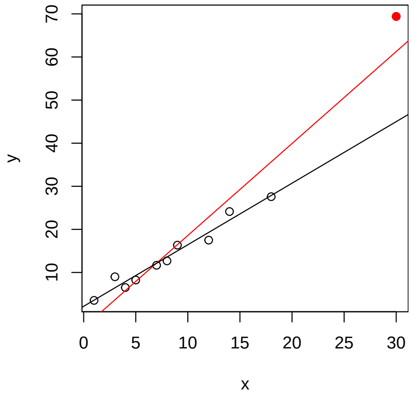
Distância de Cook: Similar a DFFITS, mas ao invés de avaliar a diferença em um único ponto, avalia a soma dos quadrados das diferenças de todos os $\hat{Y}$. Identifica pontos com influência sobre todas as estimativas de Y.

DFBETAS: Diferença normalizada entre o valor de $\_1$ no modelo completo, e o valor da mesma estimativa no modelo onde o ponto X_i, Y_i é removido, $b_{i(-i)}$. Identifica pontos com influência sobre a inclinação da reta.

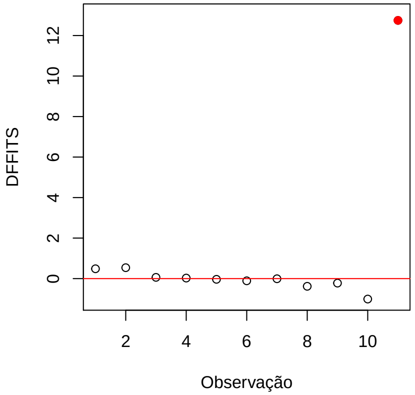
↑ Leverage e ↑ Distance



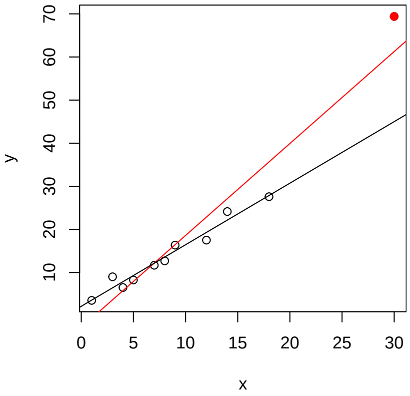
$\uparrow$ Leverage e $\uparrow$ Distance



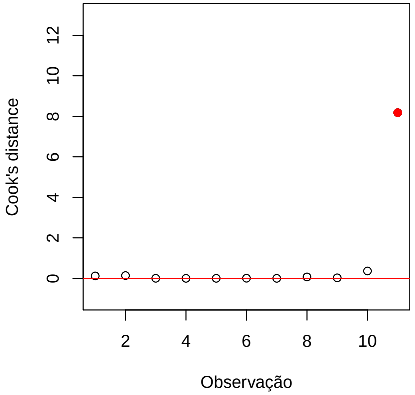

```
fits <- dffits(m)
```



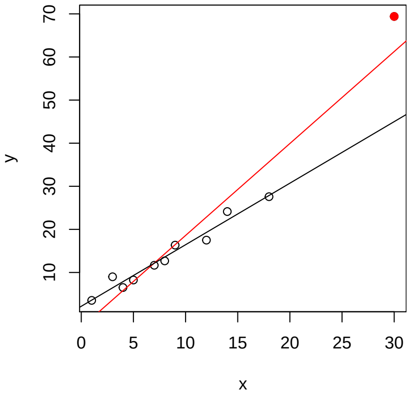
↑ Leverage e ↑ Distance



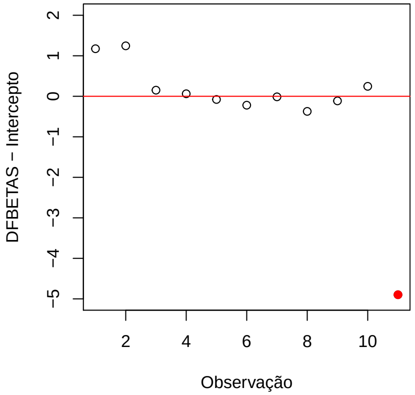

```r
fits <- cooks.distance(m)
```



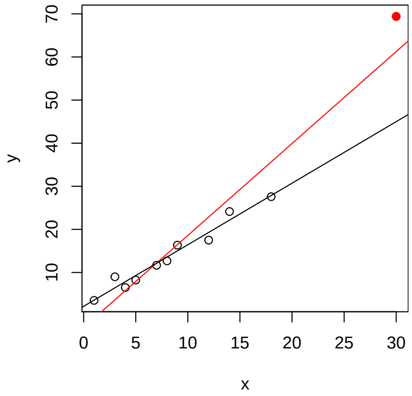
↑ Leverage e ↑ Distance



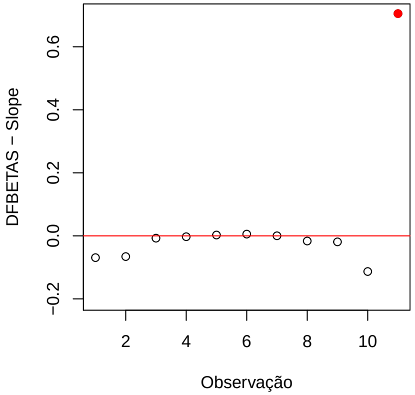

```r
fits <- dfbeta(m)
```



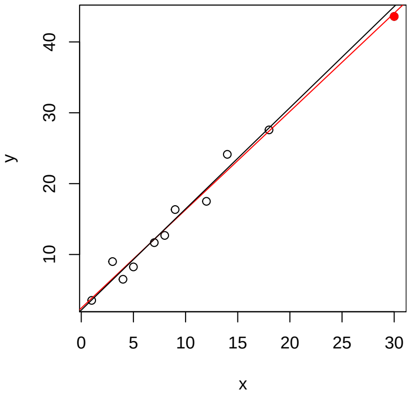
↑ Leverage e ↑ Distance



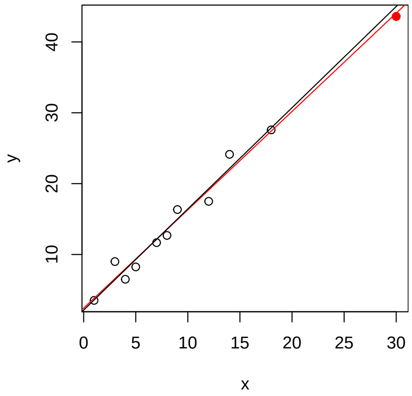

```r
fits <- dfbeta(m)
```



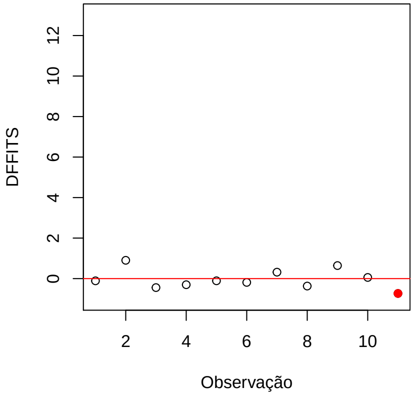
↑ Leverage e ↓ Distance



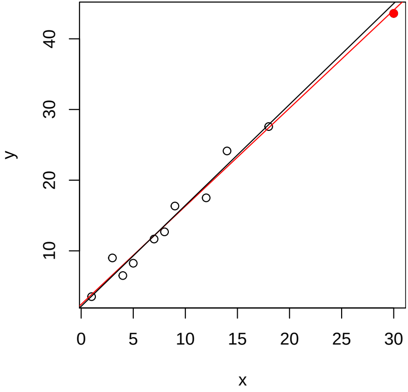
↑ Leverage e ↓ Distance



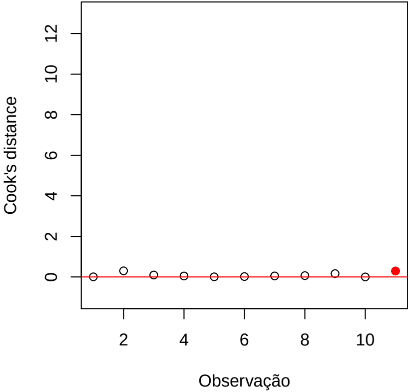

```
fits <- dffits(m)
```



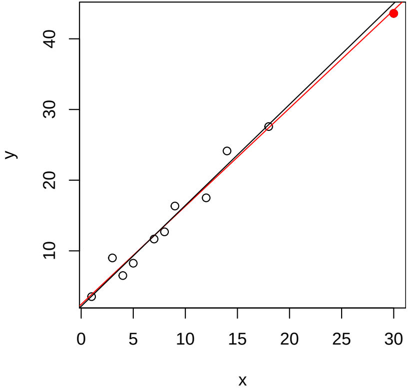
↑ Leverage e ↓ Distance



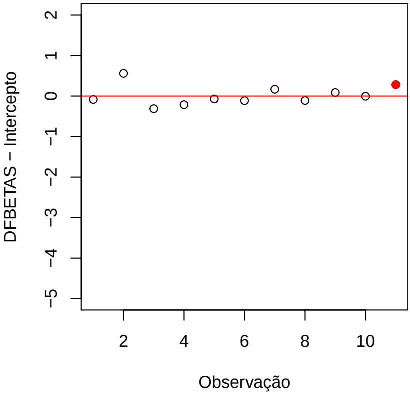

```r
fits <- cooks.distance(m)
```



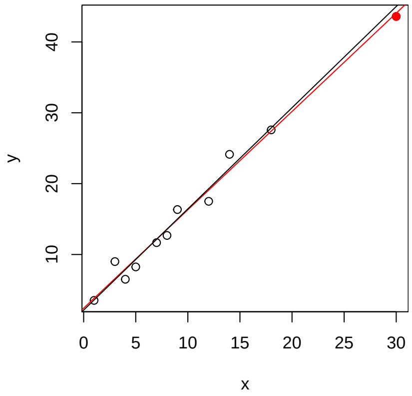
↑ Leverage e ↓ Distance



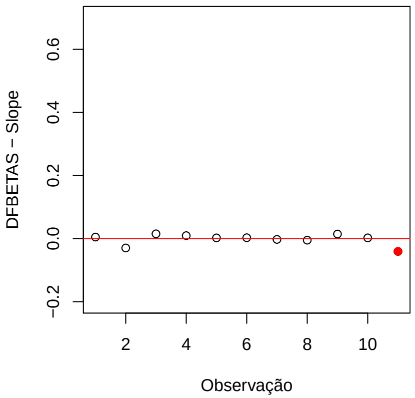

```
fits <- dfbeta(m)
```



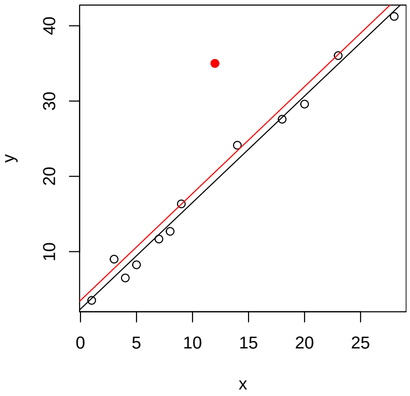
↑ Leverage e ↓ Distance



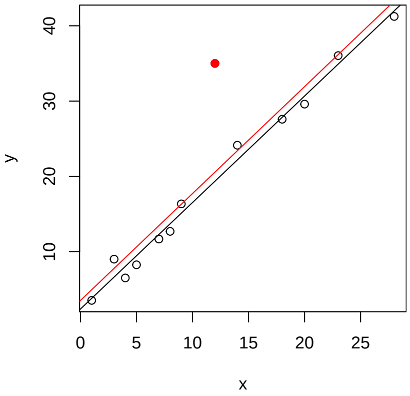

```
fits <- dfbeta(m)
```



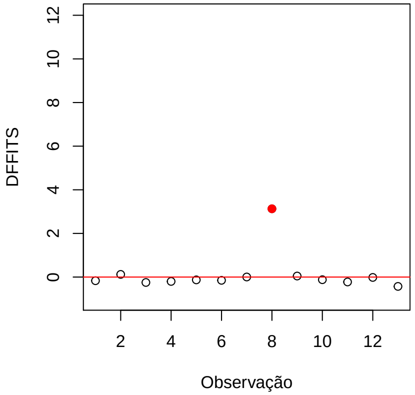
↓ Leverage e ↑ Distance



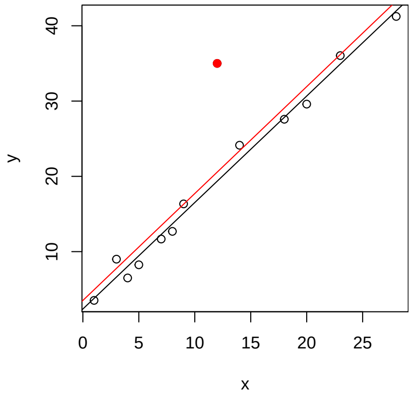
↓ Leverage e ↑ Distance



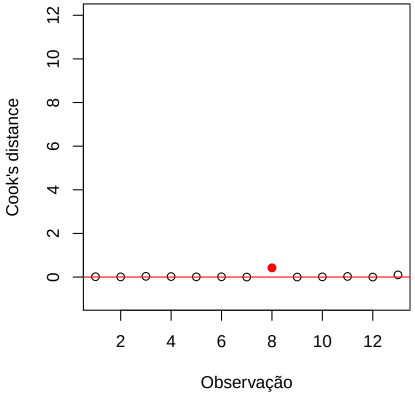

```
fits <- dffits(m)
```



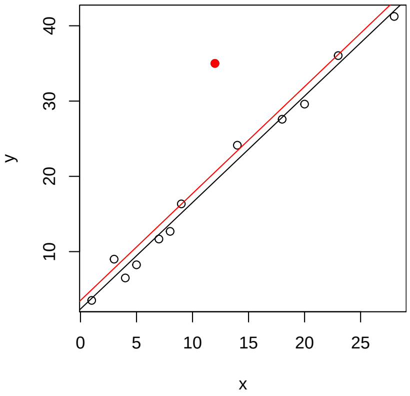
↓ Leverage e ↑ Distance



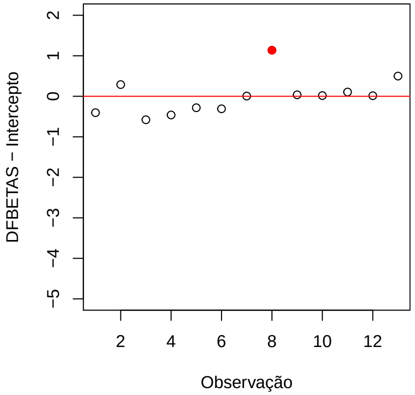

```
fits <- cooks.distance(m)
```



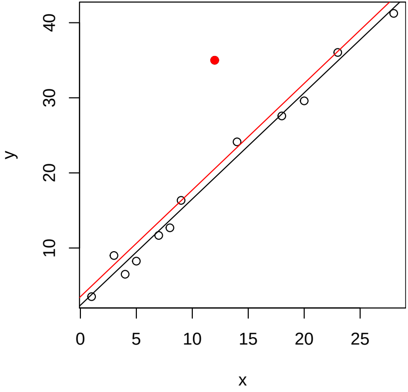
↓ Leverage e ↑ Distance



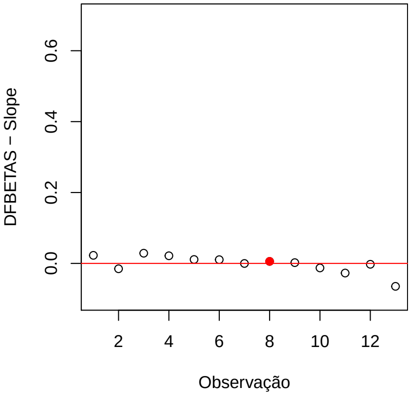

```
fits <- dfbeta(m)
```



↓ Leverage e ↑ Distance

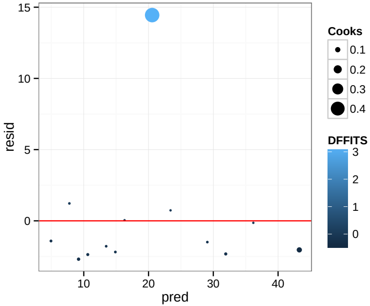


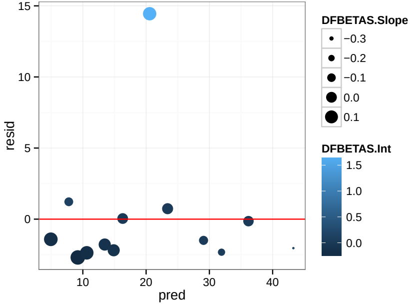

```r
fits <- dfbeta(m)
```



↓ Leverage e ↑ Distance

Visualização + diagnóstico!





Testes Estatísticos para Pressuposições

Os modelos lineares gerais são robustos, e podem tolerar pequenos desvios. Mas se voce realmente quer testar...

Testes Estatísticos para Pressuposições

Os modelos lineares gerais são robustos, e podem tolerar pequenos desvios. Mas se voce realmente quer testar...

- **Normalidade:** Kolmogorov-Smirnov, Shapiro-Wilk, Lilliefors
- **Heterscedasticidade:** Breusch-Pagan, White
- **Independência:** Durbin-Watson, Função de Autocorrelação

Remediação

Após a análise diagnóstica, descobrimos que nosso modelo viola uma ou mais pressuposições. O que fazer?

Remediação

Após a análise diagnóstica, descobrimos que nosso modelo viola uma ou mais pressuposições. O que fazer?

- Transformação de variáveis

Remediação

Após a análise diagnóstica, descobrimos que nosso modelo viola uma ou mais pressuposições. O que fazer?

- Transformação de variáveis
- Métodos Robustos e/ou Não-Paramétricos (outra aula)

Remediação

Após a análise diagnóstica, descobrimos que nosso modelo viola uma ou mais pressuposições. O que fazer?

- Transformação de variáveis
- Métodos Robustos e/ou Não-Paramétricos (outra aula)
- Outros modelos que não Modelos Lineares Gerais (outra aula)

Remediação

Após a análise diagnóstica, descobrimos que nosso modelo viola uma ou mais pressuposições. O que fazer?

- Transformação de variáveis
- Métodos Robustos e/ou Não-Paramétricos (outra aula)
- Outros modelos que não Modelos Lineares Gerais (outra aula)
- Métodos de aleatorização e reamostragem

Transformação de Variáveis

A alternativa mais simples para violações dos pressupostos é a transformação de variáveis

Transformação de Variáveis

A alternativa mais simples para violações dos pressupostos é a transformação de variáveis

Mas. . . se transformarmos as variáveis originais, não vamos alterar a relação entre elas?

Transformação de Variáveis

A alternativa mais simples para violações dos pressupostos é a transformação de variáveis

Mas. . . se transformarmos as variáveis originais, não vamos alterar a relação entre elas?

As transformações devem ser **monotônicas** (preservam a ordem relativa dos dados): Se $X_i > X_j$, então $f(X_i) > f(X_j$, e vice versa.

Transformação de Variáveis

A alternativa mais simples para violações dos pressupostos é a transformação de variáveis

Mas. . . se transformarmos as variáveis originais, não vamos alterar a relação entre elas?

As transformações devem ser **monotônicas** (preservam a ordem relativa dos dados): Se $X_i > X_j$, então $f(X_i) > f(X_j$, e vice versa.

Se usarmos funções monotônicas, podemos alterar a distância relativa entre os pontos, e assim a variância e a forma da distribuição

Funções de potência

A família de funções de potência oferece flexibilidade, dentro de uma mesma especificação:

Funções de potência

A família de funções de potência oferece flexibilidade, dentro de uma mesma especificação:

$Y' = cY^{\lambda}$ inclui:

Funções de potência

A família de funções de potência oferece flexibilidade, dentro de uma mesma especificação:

$Y' = cY^{\lambda}$ inclui:

- $Y^{-\lambda} = \dfrac{1}{Y^{\lambda}}$

Funções de potência

A família de funções de potência oferece flexibilidade, dentro de uma mesma especificação:

$Y' = cY^{\lambda}$ inclui:

- $Y^{-\lambda} = \dfrac{1}{Y^{\lambda}}$, se $\lambda = 1,\ Y' = Y^{-1} = 1/Y$

Funções de potência

A família de funções de potência oferece flexibilidade, dentro de uma mesma especificação:

$Y' = cY^{\lambda}$ inclui:

- $Y^{-\lambda} = \dfrac{1}{Y^{\lambda}}$, se $\lambda = 1,\ Y' = Y^{-1} = 1/Y$
- $Y^{\frac{1}{\lambda}} = \sqrt[\lambda]{Y}$

Funções de potência

A família de funções de potência oferece flexibilidade, dentro de uma mesma especificação:

$Y' = cY^{\lambda}$ inclui:

- $Y^{-\lambda} = \dfrac{1}{Y^{\lambda}}$, se $\lambda = 1,\ Y' = Y^{-1} = 1/Y$
- $Y^{\frac{1}{\lambda}} = \sqrt[\lambda]{Y}$
- Y^{λ}

Funções de potência

A família de funções de potência oferece flexibilidade, dentro de uma mesma especificação:

$Y' = cY^{\lambda}$ inclui:

- $Y^{-\lambda} = \dfrac{1}{Y^{\lambda}}$, se $\lambda = 1,\ Y' = Y^{-1} = 1/Y$
- $Y^{\frac{1}{\lambda}} = \sqrt[\lambda]{Y}$
- Y^{λ}

c é apenas uma constante de escala

Funções de potência

As relações de potência podem também ser expressas na forma abaixo, conhecida como transformação de Box-Cox

Funções de potência

As relações de potência podem também ser expressas na forma abaixo, conhecida como transformação de Box-Cox

$Y' = \frac{Y^{\lambda}-1}{\lambda}$, para $\lambda \neq 0$

Funções de potência

As relações de potência podem também ser expressas na forma abaixo, conhecida como transformação de Box-Cox

$Y' = \frac{Y^{\lambda}-1}{\lambda}$, para $\lambda \neq 0$

$Y' = log(Y)$, para $\lambda = 0$

Funções de potência

As relações de potência podem também ser expressas na forma abaixo, conhecida como transformação de Box-Cox

$Y' = \frac{Y^{\lambda}-1}{\lambda}$, para $\lambda \neq 0$

$Y' = log(Y)$, para $\lambda = 0$

A expressão acima é válida pois $\lim_{\lambda \to 0} \frac{Y^{\lambda}-1}{\lambda} = \log_e(X)$

Funções de potência

As relações de potência podem também ser expressas na forma abaixo, conhecida como transformação de Box-Cox

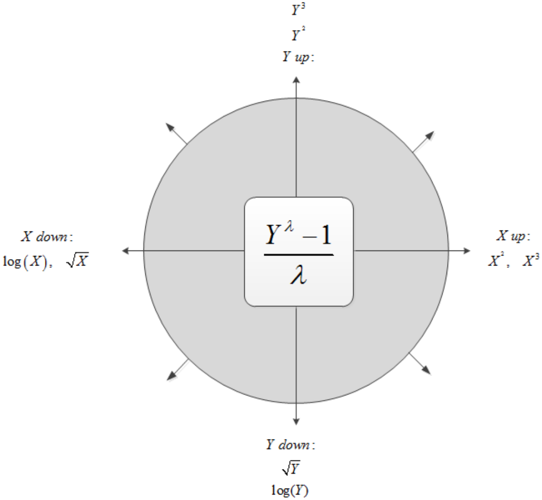
$Y' = \frac{Y^{\lambda}-1}{\lambda}$, para $\lambda \neq 0$

$Y' = log(Y)$, para $\lambda = 0$

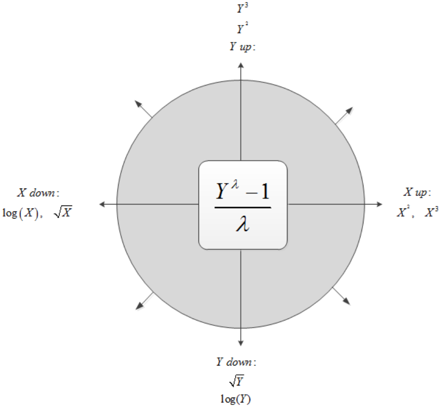
A expressão acima é válida pois $\lim_{\lambda \to 0} \frac{Y^{\lambda}-1}{\lambda} = \log_e(X)$

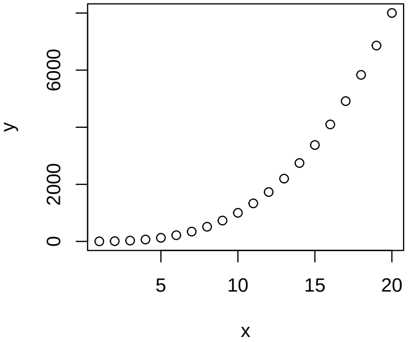
Normalmente se prefere $\log_{10}$ para facilitar a interpretação, pois um aumento de 1 em $\log_{10}(Y)$ é o mesmo que multiplicar Y por 10

Regra da Convexidade de Tukey

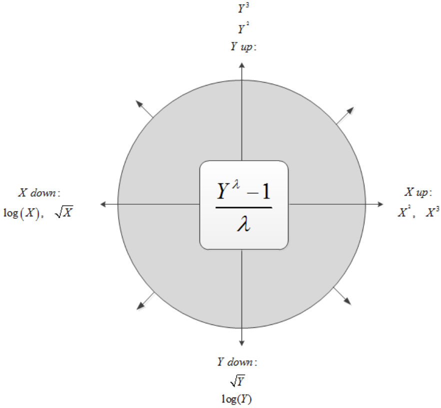


Regra da Convexidade de Tukey

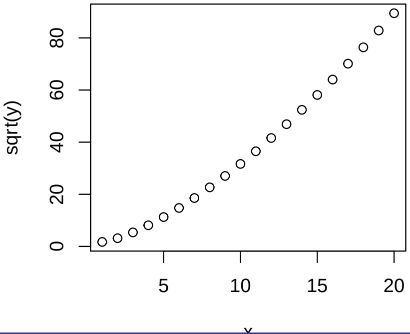




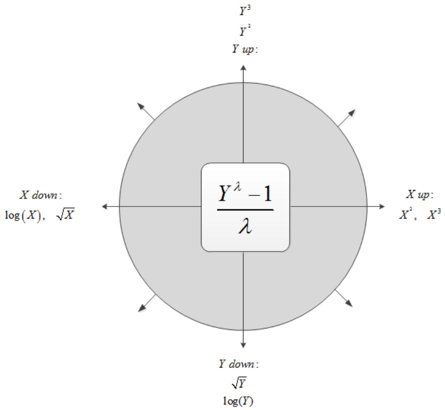
Regra da Convexidade de Tukey



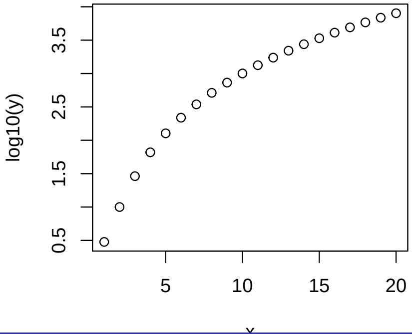

```
plot(x, sqrt(y))
```



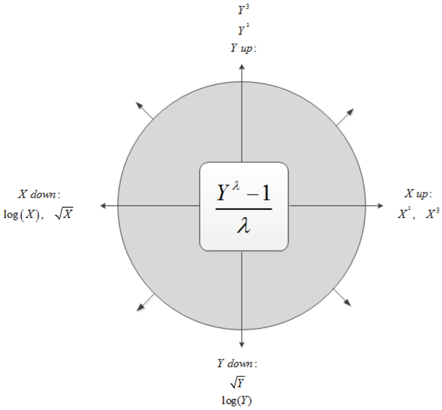
Regra da Convexidade de Tukey



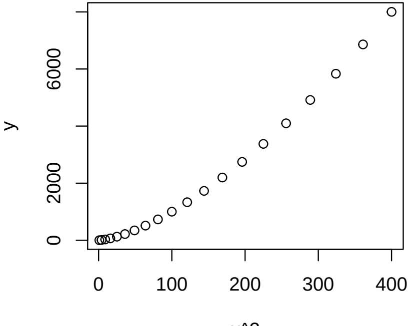

```
plot(x, log10(y))
```



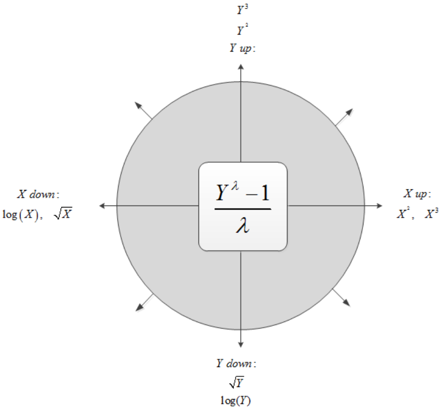
Regra da Convexidade de Tukey



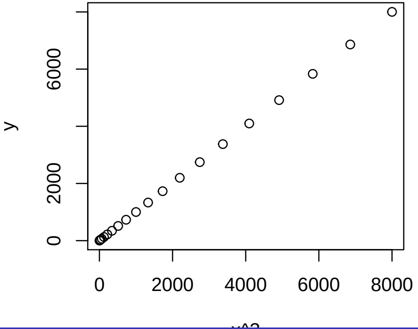

```
plot(x^2, y)
```



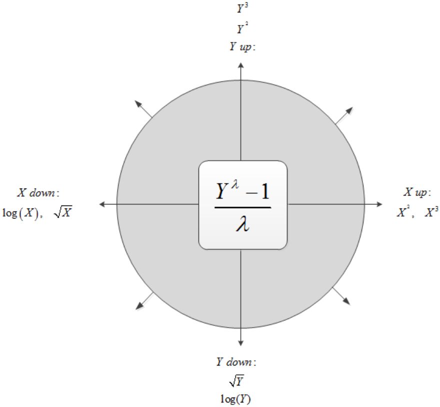
Regra da Convexidade de Tukey

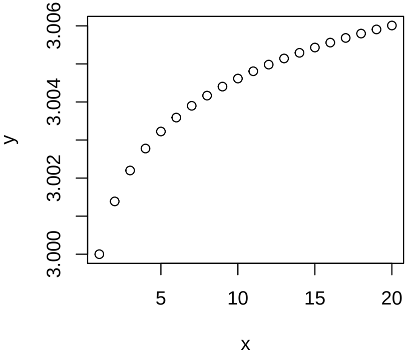


`plot`(x^3, y)

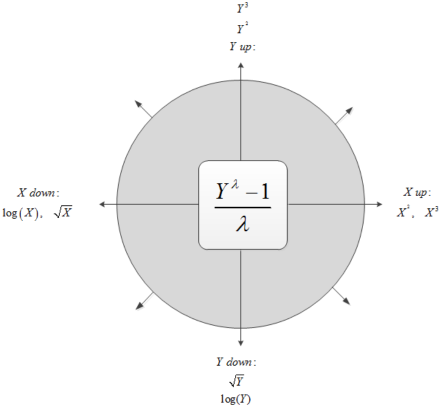


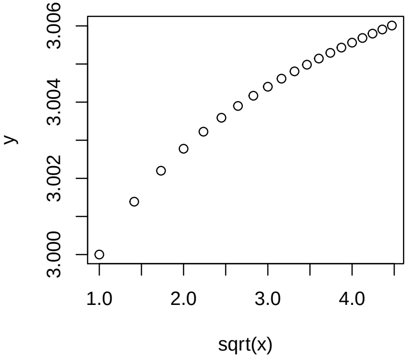
Regra da Convexidade de Tukey



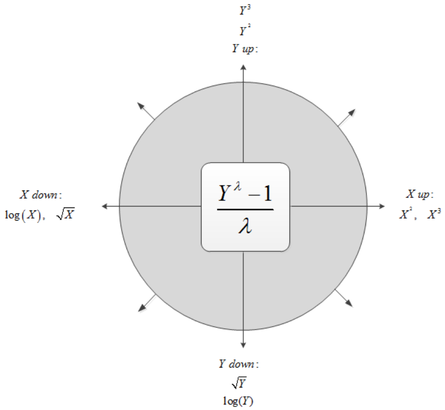


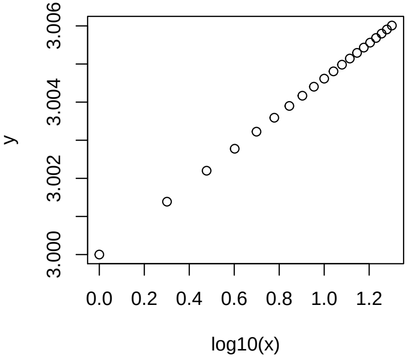
Regra da Convexidade de Tukey



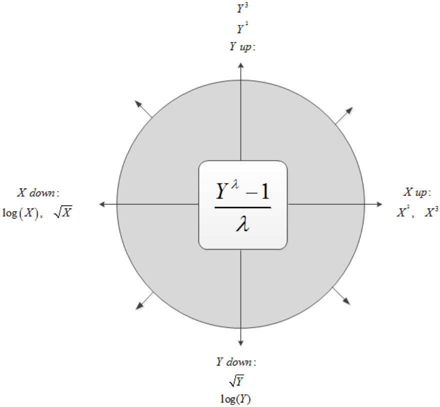


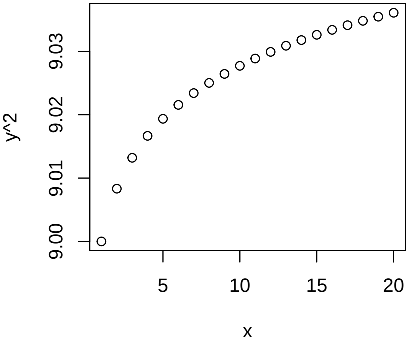
Regra da Convexidade de Tukey



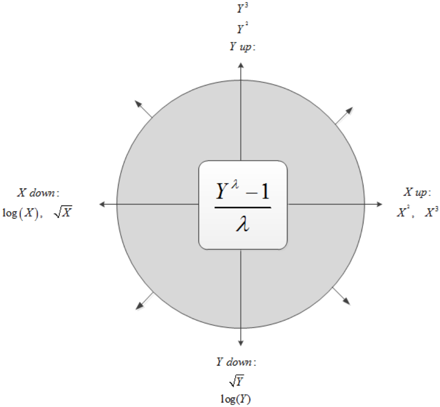


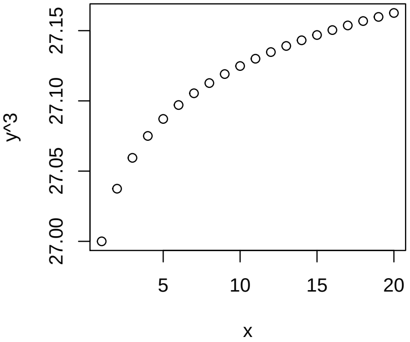
Regra da Convexidade de Tukey





Regra da Convexidade de Tukey

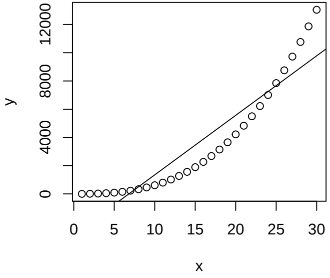




Otimização de λ

Podemos utilizar métodos computacionais para encontrar o melhor valor de λ (médodo de máxima verossimilhança, *(*maximum likelihood)

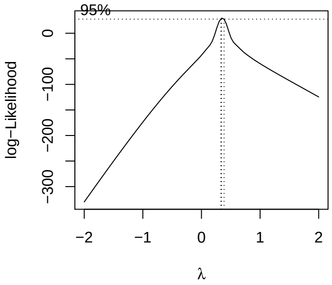
```
x <- c(1:30)
y <- 2 + x^2.786
m <- lm(y ~ x)
plot(x, y)
abline(m)
```



Otimização de λ

Podemos utilizar métodos computacionais para encontrar o melhor valor de λ (médodo de máxima verossimilhança, ou *maximum likelihood*)

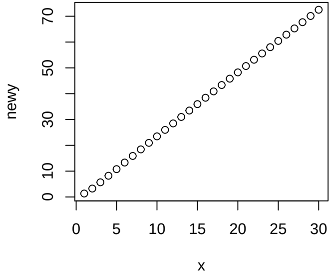
```r
library(MASS)
lambda <- boxcox(m)
```



Otimização de λ

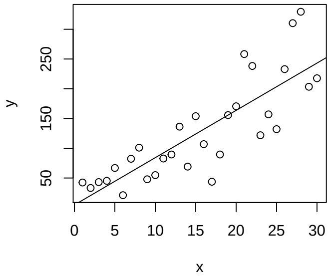
Podemos utilizar métodos computacionais para encontrar o melhor valor de λ (médodo de máxima verossimilhança, *(*maximum likelihood)

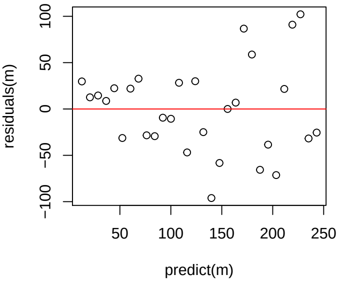
```
which(lambda$y == max(lambda$y))

## [1] 59

lambda$x[59]

## [1] 0.3434

newy <- (y^0.3434343 - 1)/0.3434343
plot(x, newy)
```



Transformações: normalidade e variância

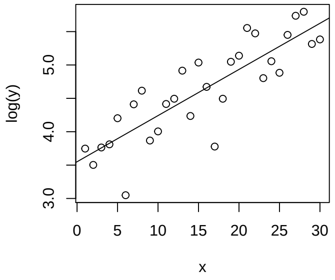
O uso de transformações não se limita à linearização de variáveis, mas também é de grande ajuda na aproximação dos dados para uma distribuição normal e variância constante

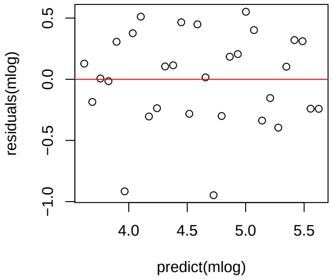




Transformações: normalidade e variância

O uso de transformações não se limita à linearização de variáveis, mas também é de grande ajuda na aproximação dos dados para uma distribuição normal e variância constante





Validação

Já fizemos a análise exploratória, ajustamos o modelo, analisamos os resíduos, resolvemos problemas de violação das pressuposições. Nosso modelo está pronto.

Validação

Já fizemos a análise exploratória, ajustamos o modelo, analisamos os resíduos, resolvemos problemas de violação das pressuposições. Nosso modelo está pronto.

E agora?

Validação

Já fizemos a análise exploratória, ajustamos o modelo, analisamos os resíduos, resolvemos problemas de violação das pressuposições. Nosso modelo está pronto.

E agora?

A última etapa do processo de modelagem consiste na **validação**, isto é, avaliação da "veracidade" do modelo

Validação

Já fizemos a análise exploratória, ajustamos o modelo, analisamos os resíduos, resolvemos problemas de violação das pressuposições. Nosso modelo está pronto.

E agora?

A última etapa do processo de modelagem consiste na **validação**, isto é, avaliação da "veracidade" do modelo

Se o modelo é explicativo, queremos ter certeza sobre nossos coeficientes

Validação

Já fizemos a análise exploratória, ajustamos o modelo, analisamos os resíduos, resolvemos problemas de violação das pressuposições. Nosso modelo está pronto.

E agora?

A última etapa do processo de modelagem consiste na **validação**, isto é, avaliação da "veracidade" do modelo

Se o modelo é explicativo, queremos ter certeza sobre nossos coeficientes

Se o modelo é preditivo, queremos ter certeza sobre as novas predições

Validação

Já fizemos a análise exploratória, ajustamos o modelo, analisamos os resíduos, resolvemos problemas de violação das pressuposições. Nosso modelo está pronto.

E agora?

A última etapa do processo de modelagem consiste na **validação**, isto é, avaliação da "veracidade" do modelo

Se o modelo é explicativo, queremos ter certeza sobre nossos coeficientes

Se o modelo é preditivo, queremos ter certeza sobre as novas predições

Validação

Os coeficientes do nosso modelo explicam a relação entre X e Y, e através desta relação podemos prever novos valores de Y

Validação

Os coeficientes do nosso modelo explicam a relação entre X e Y, e através desta relação podemos prever novos valores de Y

Mas o quanto podemos confiar em b_0 e b_1 como estimativas de β_0 e β_1, e nos valores de $\hat{Y}_{i(novo)}$?

Validação

Os coeficientes do nosso modelo explicam a relação entre X e Y, e através desta relação podemos prever novos valores de Y

Mas o quanto podemos confiar em b_0 e b_1 como estimativas de β_0 e β_1, e nos valores de $\hat{Y}_{i(novo)}$?

1) Intervalos de confiança e testes de hipóteses paramétricos

Validação

Os coeficientes do nosso modelo explicam a relação entre X e Y, e através desta relação podemos prever novos valores de Y

Mas o quanto podemos confiar em b_0 e b_1 como estimativas de β_0 e β_1, e nos valores de $\hat{Y}_{i(novo)}$?

1) Intervalos de confiança e testes de hipóteses paramétricos

2) Validação Independente

Validação

Os coeficientes do nosso modelo explicam a relação entre X e Y, e através desta relação podemos prever novos valores de Y

Mas o quanto podemos confiar em b_0 e b_1 como estimativas de β_0 e β_1, e nos valores de $\hat{Y}_{i(novo)}$?

1) Intervalos de confiança e testes de hipóteses paramétricos

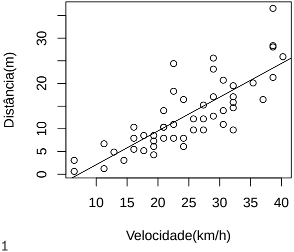
2) Validação Independente

3) Métodos de Aleatorização e Reamostragem

Exemplo

Distância de frenagem pode ser predita pela velocidade do veículo?

```
summary(m)


##
## Call:
## lm(formula = dist ~ speed, data = cars)
##
## Residuals:
##     Min      1Q Median      3Q     Max
## -8.860 -2.903 -0.692  2.809 13.168
##
## Coefficients:
##             Estimate Std. Error t value
## (Intercept)  -5.3581     2.0600   -2.60
## speed         0.7448     0.0787    9.46
##             Pr(>|t|)
## (Intercept)    0.012 *
## speed        1.5e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.69 on 48 degrees of freedom
## Multiple R-squared:  0.651, Adjusted R-squared:  0.644
## F-statistic: 89.6 on 1 and 48 DF,  p-value: 1.49e-12
```



Validação Cruzada com Dados Independentes

A melhor medida da capacidade de predição do modelo é a sua performance em estimar valores não usados no ajuste

Validação Cruzada com Dados Independentes

A melhor medida da capacidade de predição do modelo é a sua performance em estimar valores não usados no ajuste

Mas ao mesmo tempo, queremos usar o máximo de observações possíveis, para ter o melhor ajuste

Validação Cruzada com Dados Independentes

A melhor medida da capacidade de predição do modelo é a sua performance em estimar valores não usados no ajuste

Mas ao mesmo tempo, queremos usar o máximo de observações possíveis, para ter o melhor ajuste

Validação cruzada: dividimos a amostra em 2 partes iguais, e usamos cada metade para validar um modelo ajustado à outra metade

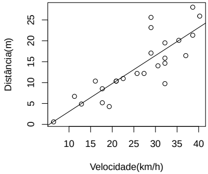
Exemplo

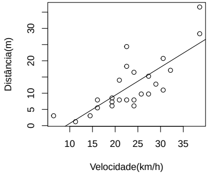
Distância de frenagem pode ser predita pela velocidade do veículo?

```r
dim(cars)
```

```
## [1] 50  2
```

```r
set.seed(89)
samp <- sample(1:50, 25, rep = F)
cars1 <- cars[samp, ]
cars2 <- cars[-samp, ]
m1 <- lm(dist ~ speed, cars1)
m2 <- lm(dist ~ speed, cars2)
```

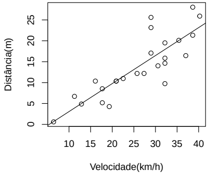


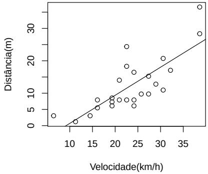


Exemplo

Distância de frenagem pode ser predita pela velocidade do veículo?

```
summary(m1)$coefficients

##             Estimate Std. Error t value
## (Intercept)    -3.63    2.45699  -1.477
## speed           0.67    0.08864   7.559
##              Pr(>|t|)
## (Intercept) 1.531e-01
## speed       1.120e-07

summary(m2)$coefficients

##             Estimate Std. Error t value
## (Intercept)   -7.960     3.4866  -2.283
## speed          0.866     0.1421   6.095
##              Pr(>|t|)
## (Intercept) 3.199e-02
## speed       3.236e-06
```

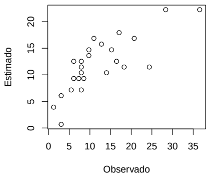


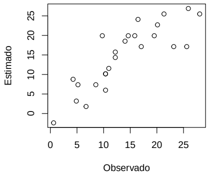


Exemplo

Distância de frenagem pode ser predita pela velocidade do veículo?

```
pr1 <- predict(m1, cars2)
pr2 <- predict(m2, cars1)
rmse1 <- sqrt(mean((cars2$dist - pr1)^2))
rmse2 <- sqrt(mean((cars1$dist - pr2)^2))
rmse1

## [1] 5.311

rmse2

## [1] 4.314
```





Jackknife ou *LOOCV*

Não seria ótimo se pudéssemos usar o máximo possível de observações pra estimar o erro?

Jackknife ou *LOOCV*

Não seria ótimo se pudéssemos usar o máximo possível de observações pra estimar o erro?

E se, ao invés de dividir meio a meio, deixássemos uma observação de fora, e repetíssemos n vezes?

Jackknife ou *LOOCV*

Não seria ótimo se pudéssemos usar o máximo possível de observações pra estimar o erro?

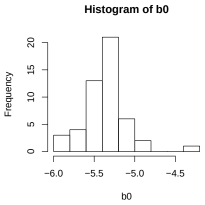
E se, ao invés de dividir meio a meio, deixássemos uma observação de fora, e repetíssemos n vezes?

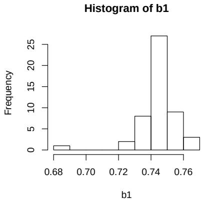
Jacknife* ou *Leave One Out Cross-Validation (LOOCV)

Exemplo

Distância de frenagem pode ser predita pela velocidade do veículo?

```
b0 <- vector()
b1 <- vector()
rq <- vector()
for (i in c(1:50)) {
    m <- lm(dist ~ speed, data = cars[-i,
      ])
    b0 <- c(b0, coefficients(m)[1])
    b1 <- c(b1, coefficients(m)[2])
    rq <- c(rq, summary(m)$r.squared)
}
```

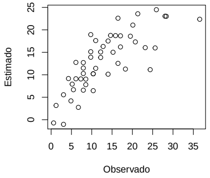


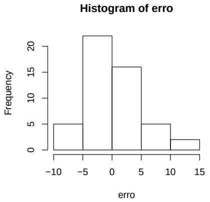


Exemplo

Distância de frenagem pode ser predita pela velocidade do veículo?

```r
pred <- b0 + cars$speed * b1
erro <- cars$dist - pred
rmse <- sqrt(mean(erro^2))
rmse


## [1] 4.785
```





Jackknife ou *LOOCV*

Esse método pode também ser usado para criar intervalos de confiança para β_0, β_1, etc.

Jackknife ou *LOOCV*

Esse método pode também ser usado para criar intervalos de confiança para β_0, β_1, etc.

Conduza a aleatorização, e reporte os percentis $\alpha/2$ e $1 - \alpha/2$

Jackknife ou *LOOCV*

Esse método pode também ser usado para criar intervalos de confiança para β_0, β_1, etc.

Conduza a aleatorização, e reporte os percentis $\alpha/2$ e $1 - \alpha/2$

Problema: poucas observações

Bootstrap

Generalização dos métodos de reamostragem

Bootstrap

Generalização dos métodos de reamostragem

Selecione n novas amostras, com reposição, e recalcule o modelo. Repita **muitas** vezes.

Bootstrap

Generalização dos métodos de reamostragem

Selecione n novas amostras, com reposição, e recalcule o modelo. Repita **muitas** vezes.

Observações mais frequentes vão ser reamostradas mais vezes

Bootstrap

Generalização dos métodos de reamostragem

Selecione n novas amostras, com reposição, e recalcule o modelo. Repita **muitas** vezes.

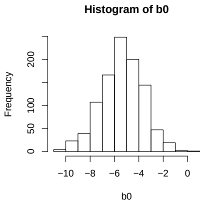
Observações mais frequentes vão ser reamostradas mais vezes

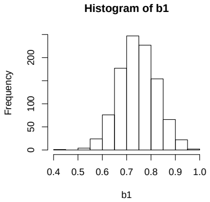
O resultado final aproxima a distribuição original de ϵ e Y (seja ela qual for)

Exemplo

Distância de frenagem pode ser predita pela velocidade do veículo?

```
b0 <- vector()
b1 <- vector()
rq <- vector()
for (i in c(1:1000)) {
    sub <- sample(1:50, 50, replace = T)
    m <- lm(dist ~ speed, data = cars[sub,
        ])
    b0 <- c(b0, coefficients(m)[1])
    b1 <- c(b1, coefficients(m)[2])
    rq <- c(rq, summary(m)$r.squared)
}
```





Conclusão

É importante garantir que os dados satisfaçam às pressuposições dos Modelos Lineares Gerais

Conclusão

É importante garantir que os dados satisfaçam às pressuposições dos Modelos Lineares Gerais

Mas ao mesmo tempo, vale lembrar que estes métodos são bastante robustos, especialmente quando n é grande

Conclusão

É importante garantir que os dados satisfaçam às pressuposições dos Modelos Lineares Gerais

Mas ao mesmo tempo, vale lembrar que estes métodos são bastante robustos, especialmente quando n é grande

Diagnóstico e remediação, mas sem obsessão!

Conclusão

É importante garantir que os dados satisfaçam às pressuposições dos Modelos Lineares Gerais

Mas ao mesmo tempo, vale lembrar que estes métodos são bastante robustos, especialmente quando n é grande

Diagnóstico e remediação, mas sem obsessão!

Próxima Aula: Regressão Múltipla e Anova Multifatorial!