AULA 6: MODELOS LINEARES GERAIS III

Análise Estatística e Modelagem de Dados Ecológicos

Thiago S. F. Silva - tsfsilva@rc.unesp.br

30 de Março de 2015

Programa de Pós Graduação em Ecologia e Biodiversidade - UNESP

OUTLINE

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)

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

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:
   - $\varepsilon \sim N(0, \sigma^2)$
   - $Y \sim N(\beta_0 + \beta_1 X, \sigma^2)$

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

ANÁLISE DE RESÍDUOS

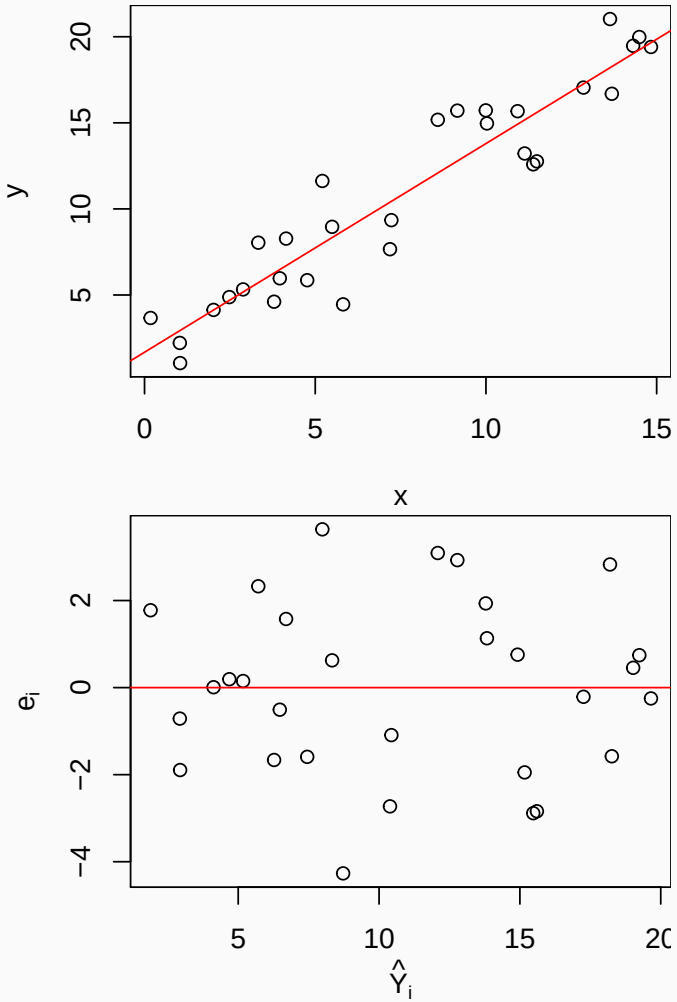
Uma das princpais análises diagnósticas é um scatterplot dos resíduos vs. valores estimados ($\hat{Y}$)

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

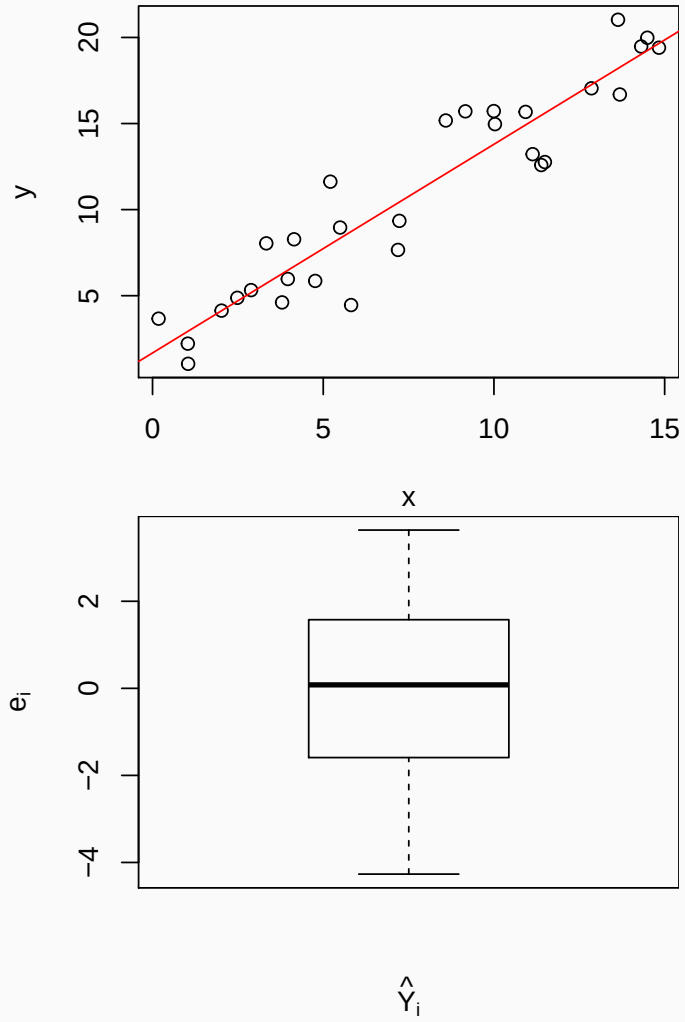


ANÁLISE DE RESÍDUOS

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

Vejamos um exemplo de um modelo de regressão apropriado:

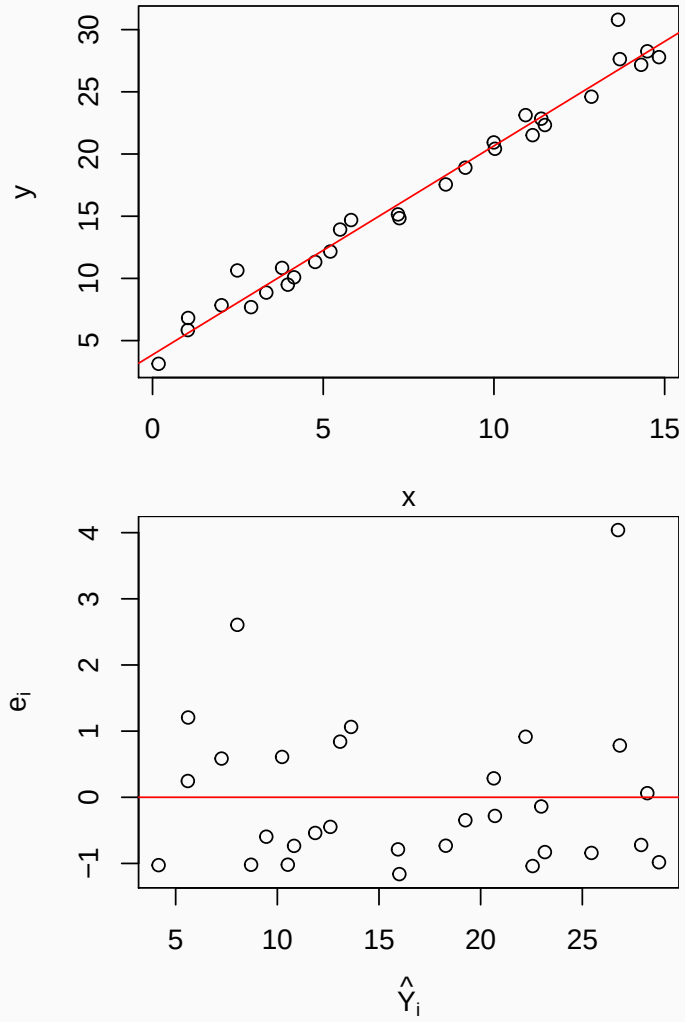
- Resíduos aletoriamente distribuidos ao redor de zero
- Variação constante ao longo de $\hat{Y}_i$



ANÁLISE DE RESÍDUOS

Dados com resíduos não-normais:

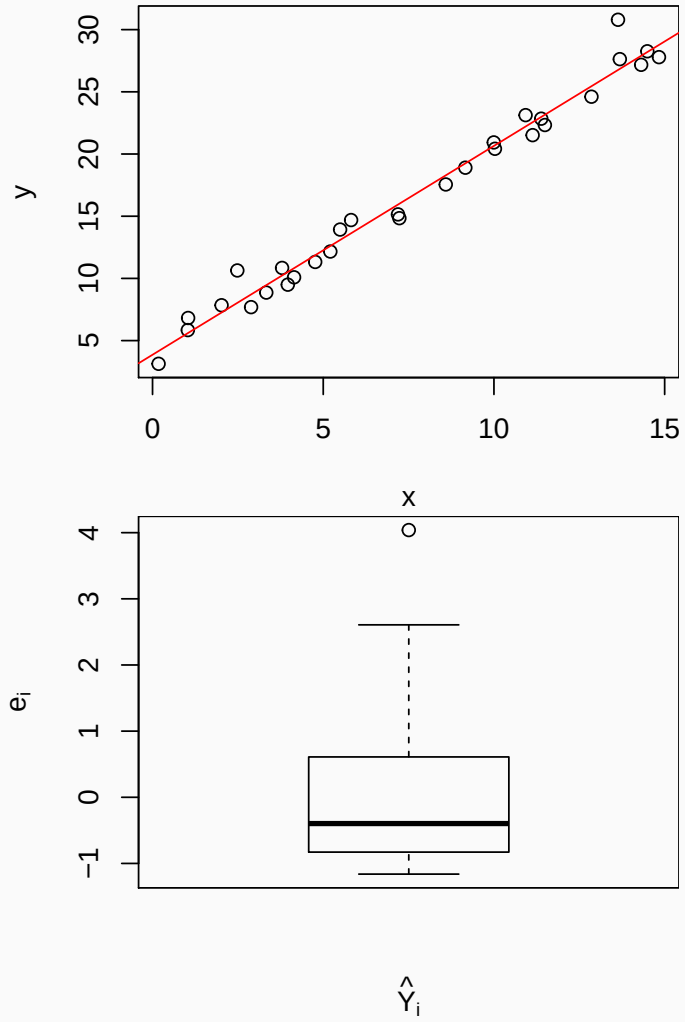
- Resíduos positivos maiores do que resíduos negativos
- Distribuição Assimétrica



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

Resíduos não-normais: **Q-Q plot**

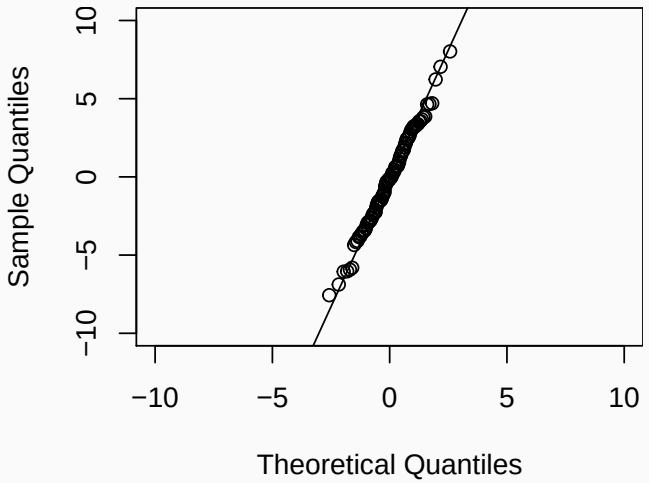
- Ferramenta gráfica bastante utilizada para avaliação de aderência à normalidade
- Plota os quantis dos dados contra os quantis correspondentes de uma distribuição normal com os mesmos parâmetros: $N(0, s^2)$

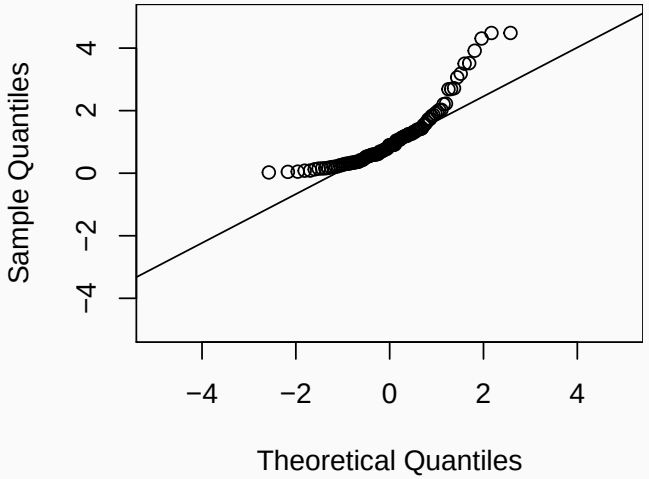
Resíduos não-normais: **Q-Q plot**

- 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: $N(0, s^2)$
- Quanto mais normal a distribuição, mais "iguais" serão os quantis

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: $N(0, s^2)$
- Quanto mais normal a distribuição, mais "iguais" serão os quantis





ANÁLISE DE RESÍDUOS

Dados com resíduos heteroscedásticos:

ANÁLISE DE RESÍDUOS

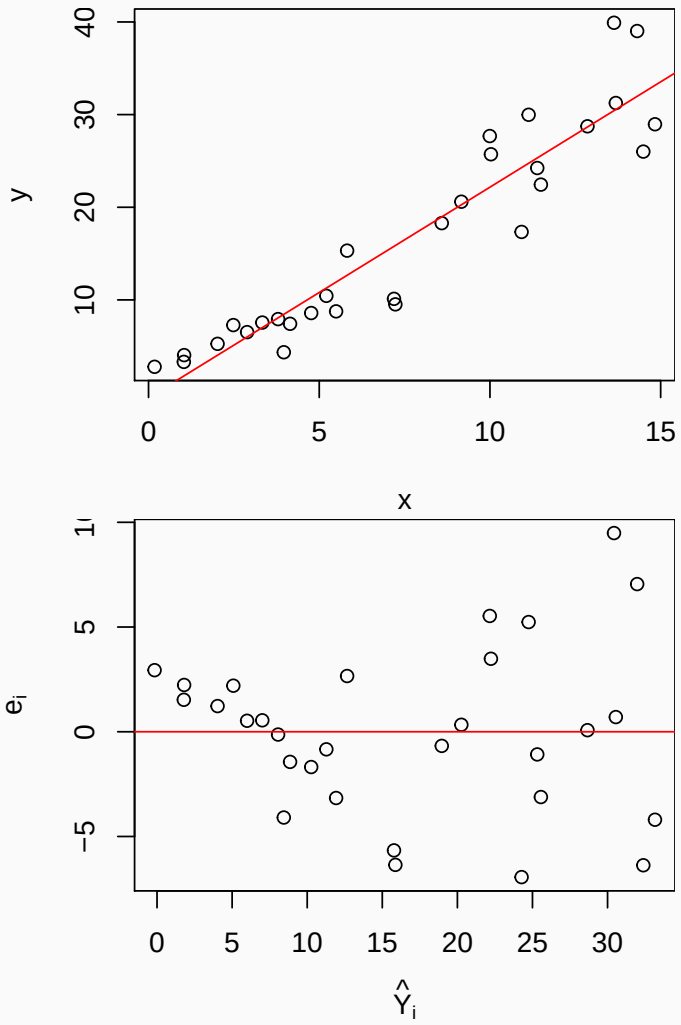
Dados com resíduos heteroscedásticos:

- **homoscedástico** = variância constante
- **heteroscedástico** = variância inconstante

ANÁLISE DE RESÍDUOS

Dados com resíduos heteroscedásticos:

- **homoscedástico** = variância constante
- **heteroscedástico** = variância inconstante
- Resíduos aleatoriamente distribuídos ao redor de zero
- Variância dos resíduos aumenta (ou diminui) ao longo de $\hat{Y}$



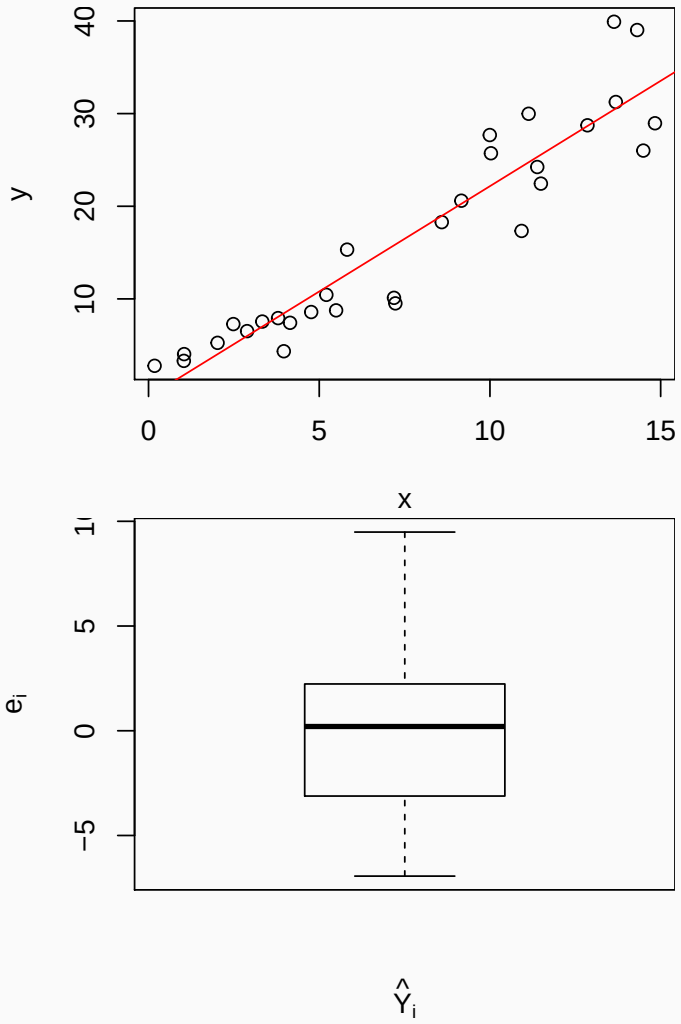
ANÁLISE DE RESÍDUOS

Dados com resíduos heteroscedásticos:

ANÁLISE DE RESÍDUOS

Dados com resíduos heteroscedásticos:

- **homoscedástico** = variância constante
- **heteroscedástico** = variância inconstante
- Resíduos aleatoriamente distribuídos ao redor de zero
- Variância dos resíduos aumenta (ou diminui) ao longo de $\hat{Y}$



ANÁLISE DE RESÍDUOS

Dados com resíduos não-independentes:

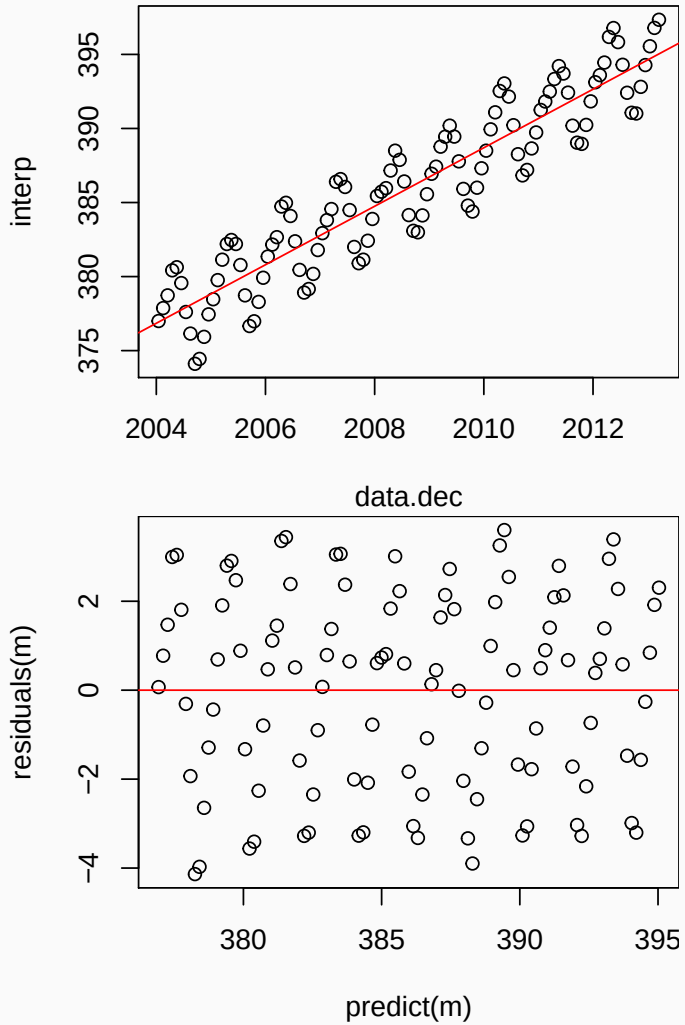
Curva de Keeling

ANÁLISE DE RESÍDUOS

Dados com resíduos não-independentes:

Curva de Keeling
Concentração de CO_2 atmosférico medido em Mauna Loa, Hawaii

- Resíduos distribuidos sistematicamente ao redor de zero

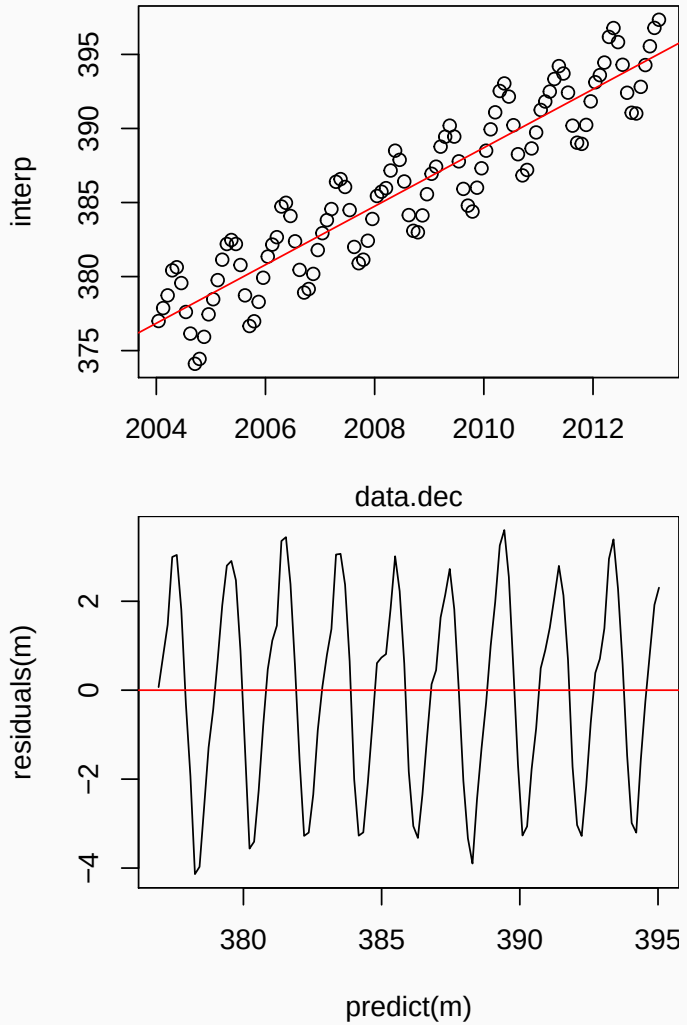


ANÁLISE DE RESÍDUOS

Dados com resíduos não-independentes:

Curva de Keeling
Concentração de CO_2 atmosférico medido em Mauna Loa, Hawaii

- Resíduos distribuídos sistematicamente ao redor de zero



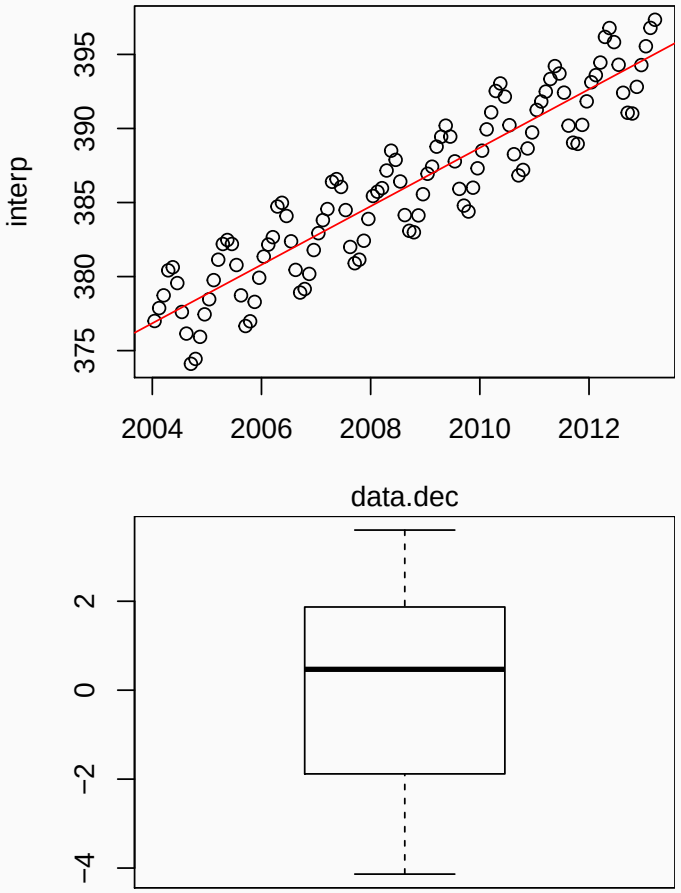
ANÁLISE DE RESÍDUOS

Dados com resíduos não-independentes:

Curva de Keeling

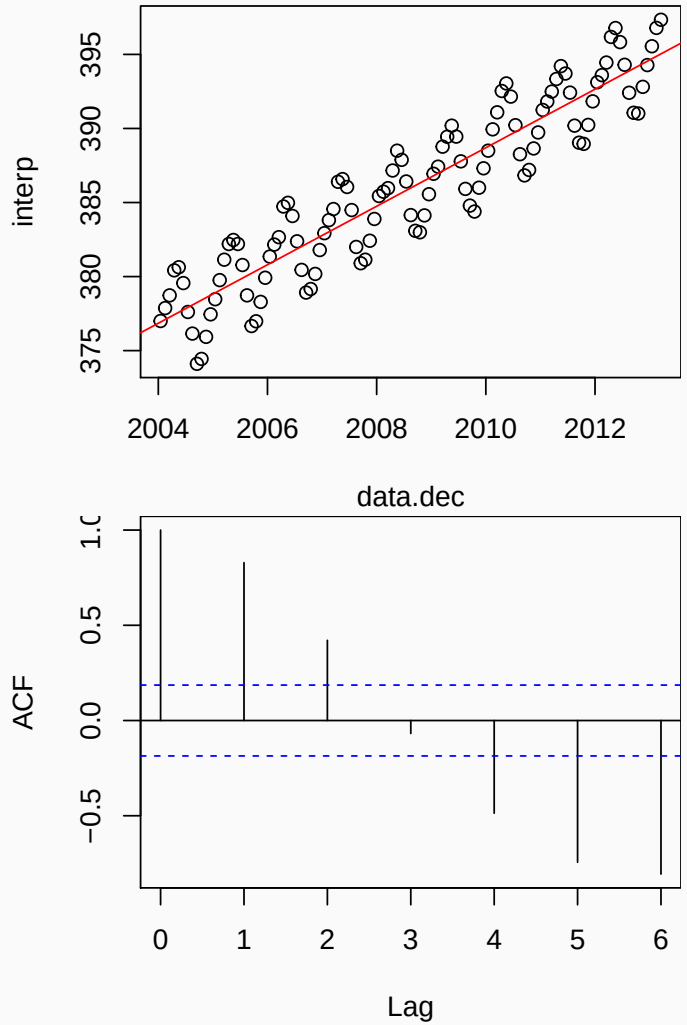
Concetração de CO2 atmosférico medido em Mauna Loa, Hawaii

- Resíduos distribuidos sistematicamente ao redor de zero



Função de Autocorrelação

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



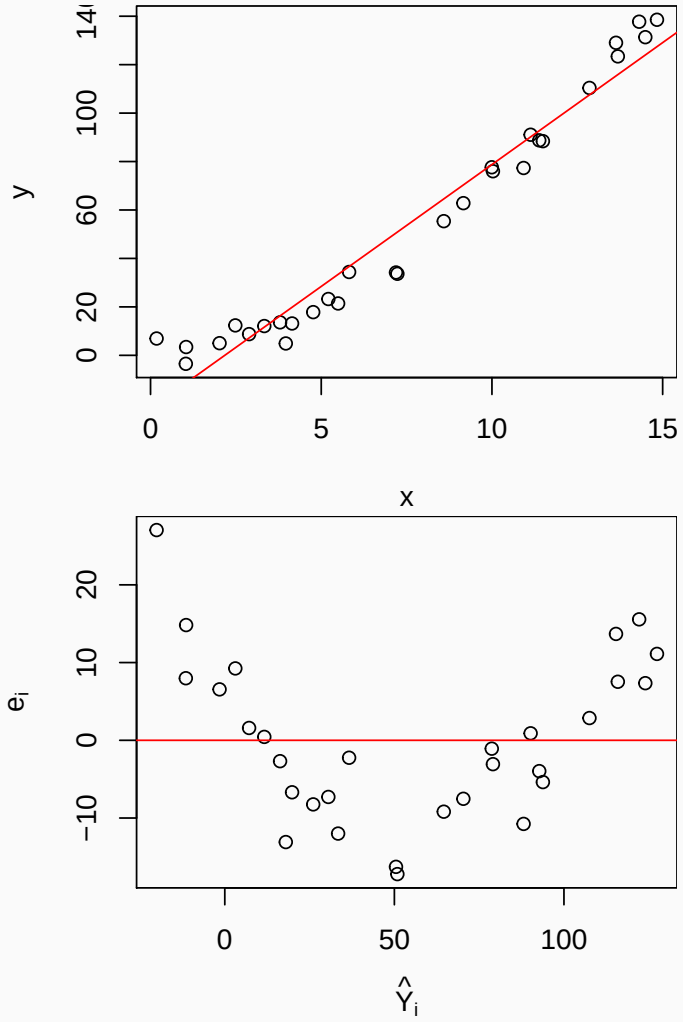
ANÁLISE DE RESÍDUOS

Relação entre X e Y não é linear

ANÁLISE DE RESÍDUOS

Relação entre X e Y não é linear

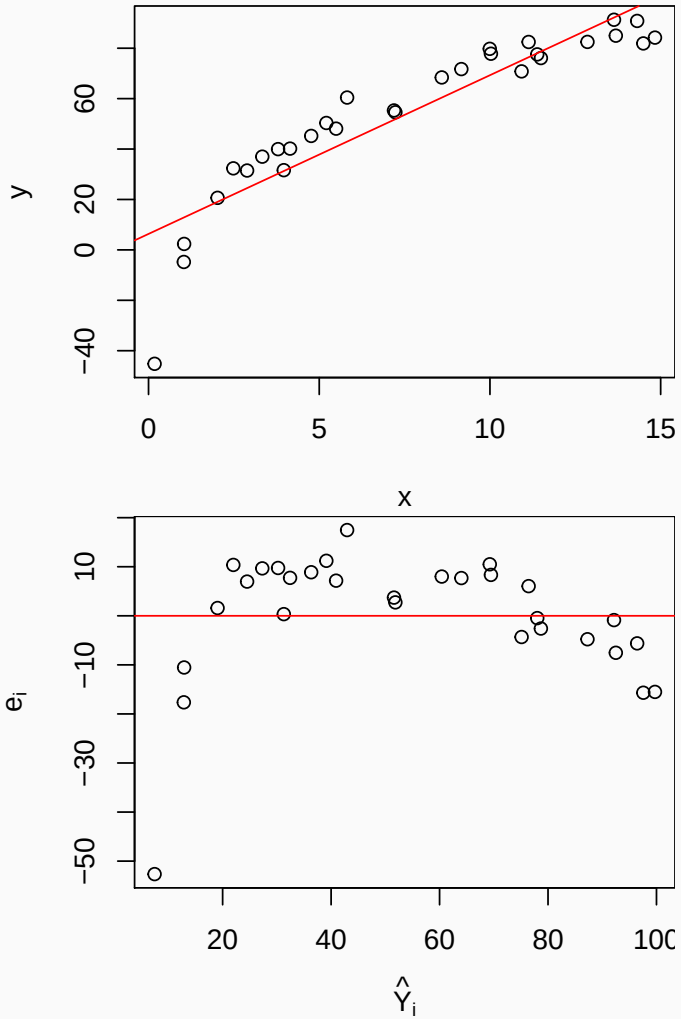
- Resíduos distribuidos segundo um padrão
- O padrão sugere o tipo de relação



ANÁLISE DE RESÍDUOS

Relação entre X e Y não é linear

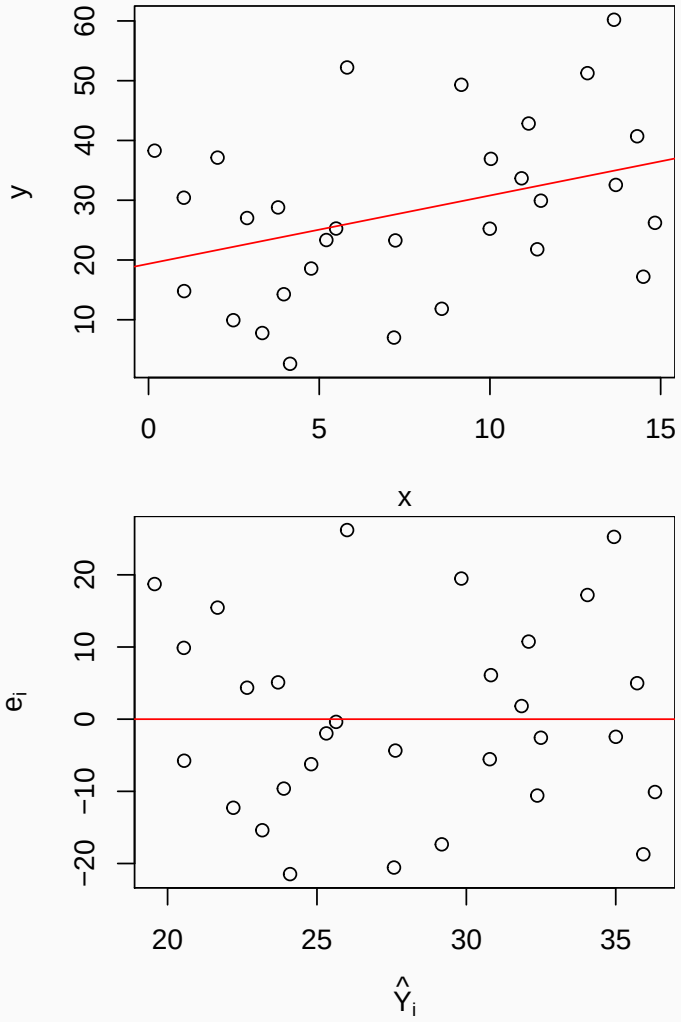
- Resíduos distribuidos segundo um padrão
- O padrão sugere o tipo de relação



ANÁLISE DE RESÍDUOS

Ausência de uma variável explicativa

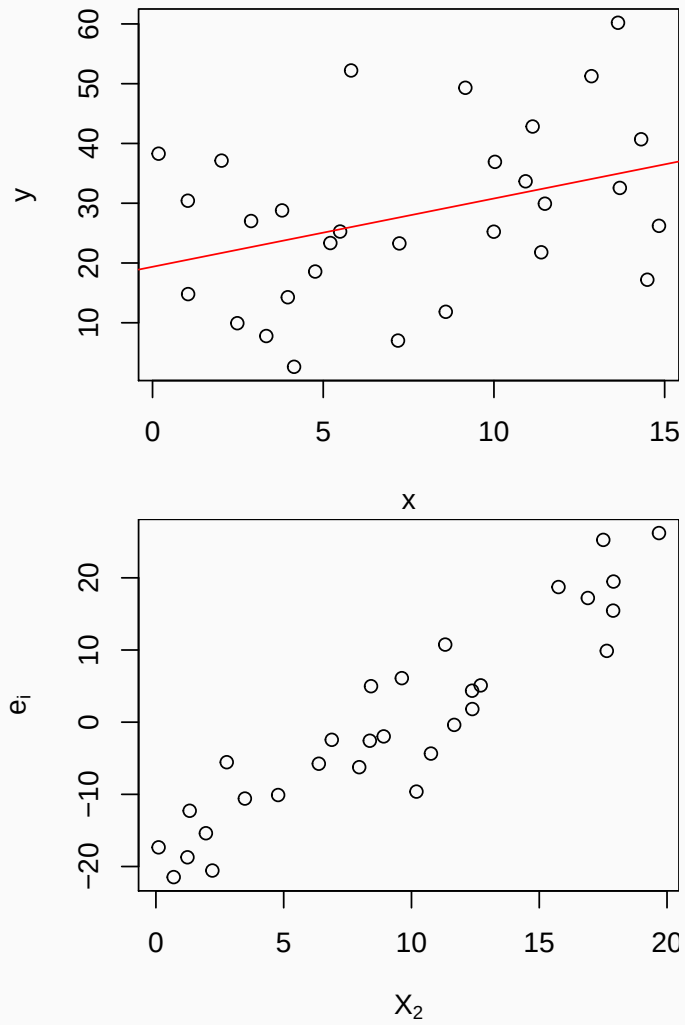
- Grande parte da variância não explicada por X_1
- Resíduos nem sempre revelam um padrão
- Mas pode ocorrer forte relação entre os resíduos e a varável omitida



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



TIPOS DE RESÍDUOS

Resíduos Brutos (*Raw*): Os resíduos originais do modelo, na escala de valores original de Y. Se oriundos de um modelo múltiplo, podem assumir valores ”estranhos”.

Resíduos Normalizados (*Standardized*): Resíduos originais, divididos pelo erro padrão dos resíduos: $e_i/\sqrt{MQ_{Res}}$. Contudo, apesar da pressuposição de variância constante dos **erros (ε)**, os resíduos tendem a ser menos confiáveis quanto mais distante de $(\bar{X}, \bar{Y})$.

TIPOS DE RESÍDUOS

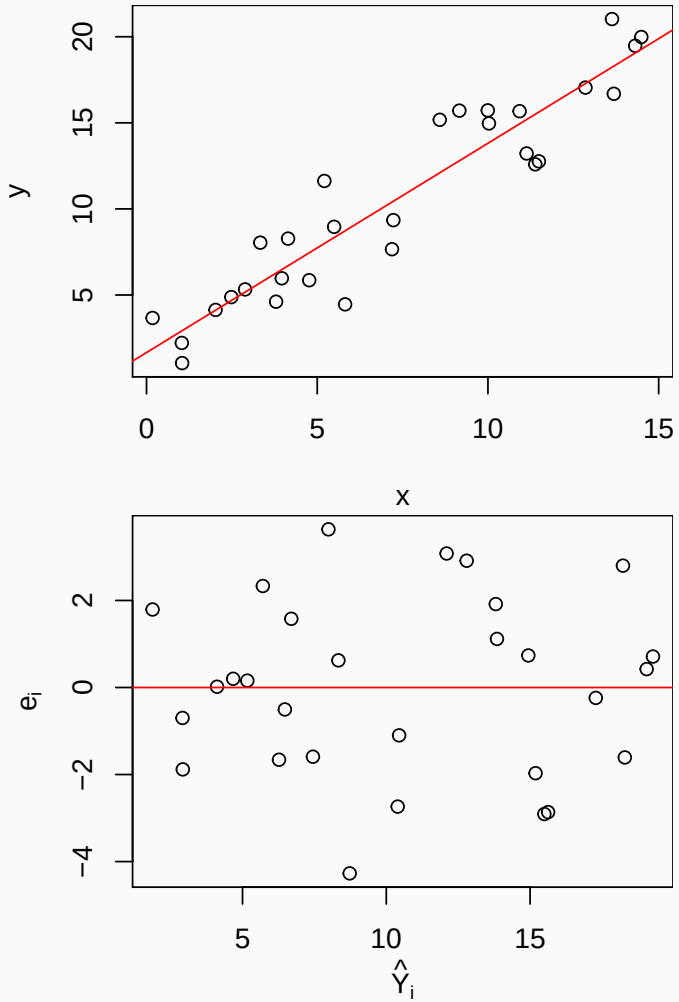
Resíduos Brutos (*Raw*): Os resíduos originais do modelo, na escala de valores original de Y. Se oriundos de um modelo múltiplo, podem assumir valores "estranhos".

Resíduos Normalizados (*Standardized*): Resíduos originais, divididos pelo erro padrão dos resíduos: $e_i/\sqrt{MQ_{Res}}$. Contudo, apesar da pressuposição de variância constante dos **erros (ε)**, os resíduos tendem a ser menos confiáveis quanto mais distante de $(\bar{X}, \bar{Y})$.

Resíduos Estudantizados (*Studentized*): Normalização dos resíduos que leva em conta esse efeito de distância de $\bar{X}$, através da fórmula $e_i/MQ_{Res} \times \sqrt{(1-h_{ii})}$. h_{ii} é um elemento da matriz desenho do modelo, que captura esse efeito.

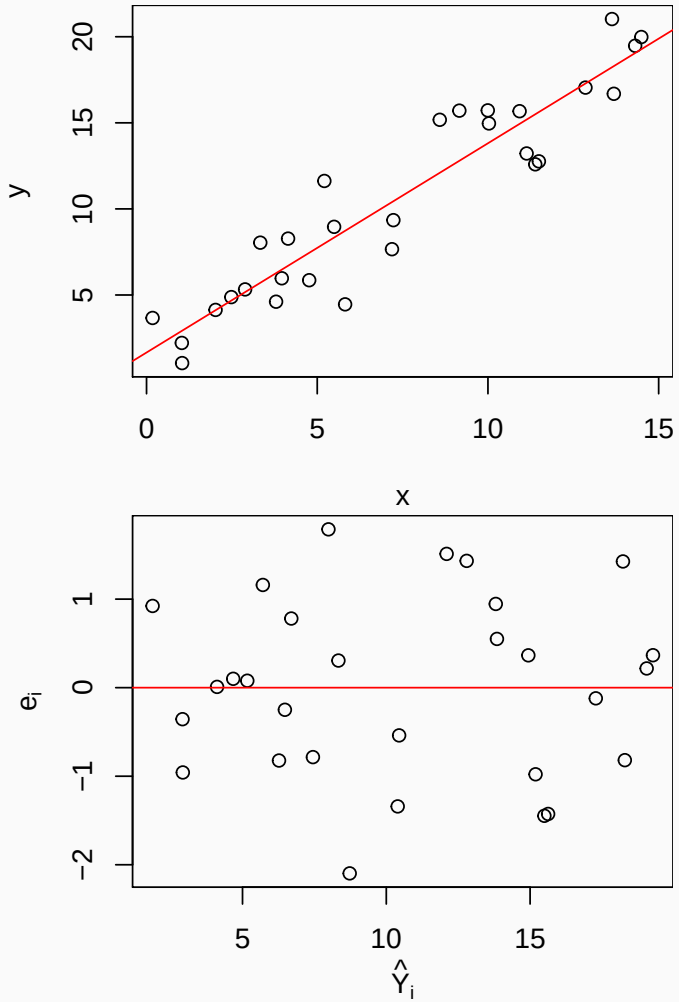
ANÁLISE DE RESÍDUOS

- Resíduos Brutos



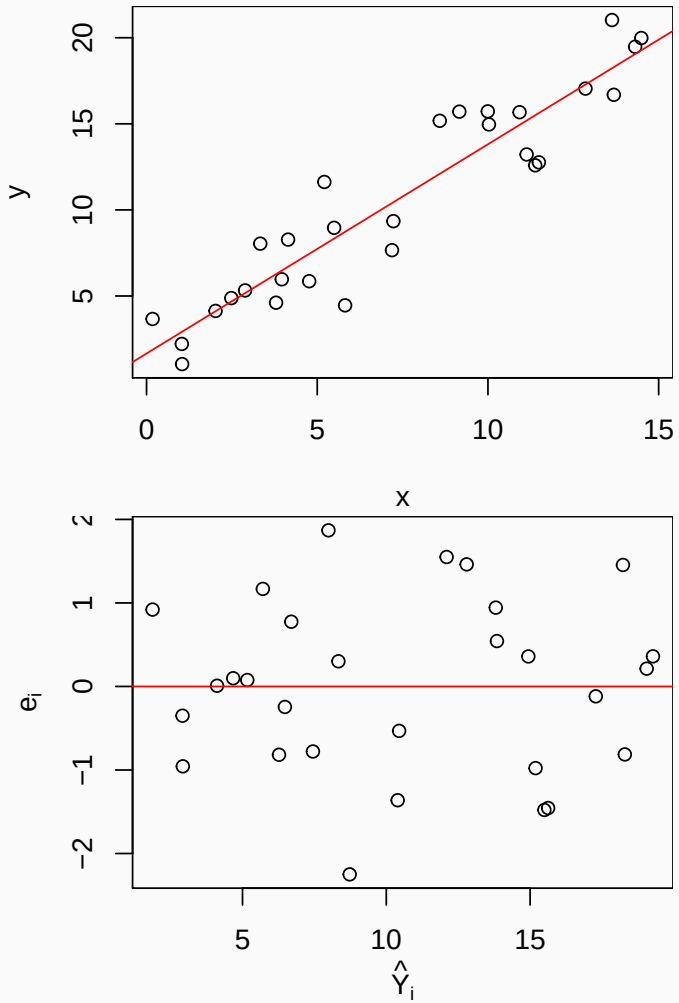
ANÁLISE DE RESÍDUOS

- Resíduos Brutos
- **Resíduos Normalizados**



ANÁLISE DE RESÍDUOS

- Resíduos Brutos
- Resíduos Normalizados
- **Resíduos Estudantizados**



PONTOS INFLUENTES

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

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 $(\bar{X}, \bar{Y})$

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 $(\bar{X}, \bar{Y})$

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

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

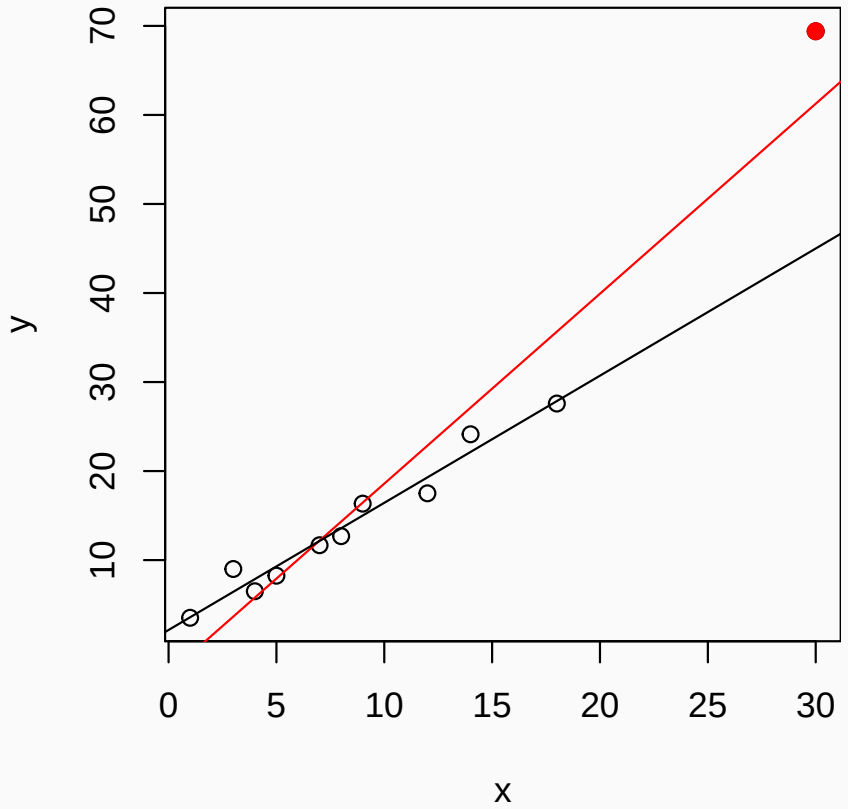
Leverage: Mede o efeito de valores extremos de X. *Leverage* vem de *lever* (alavanca). Valores extremos de X podem alavancar a reta de regressão, que se "equilibra" em $(\bar{X}, \bar{Y})$

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

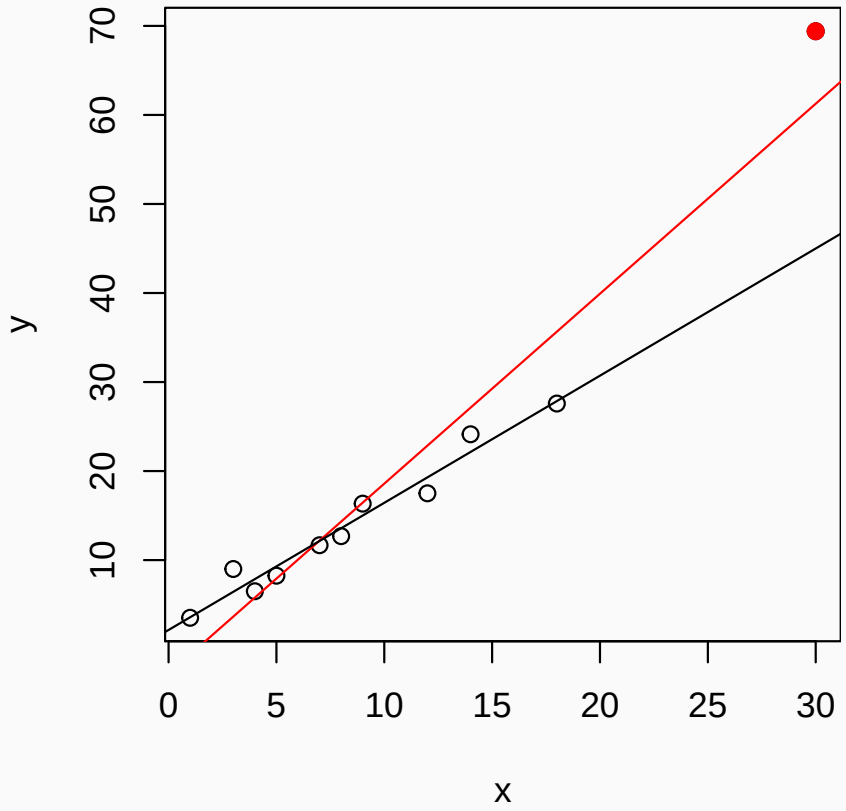
Influência: Combinação de distância e *leverage*, captura efeito total de um *outlier* sobre a reta de regressão

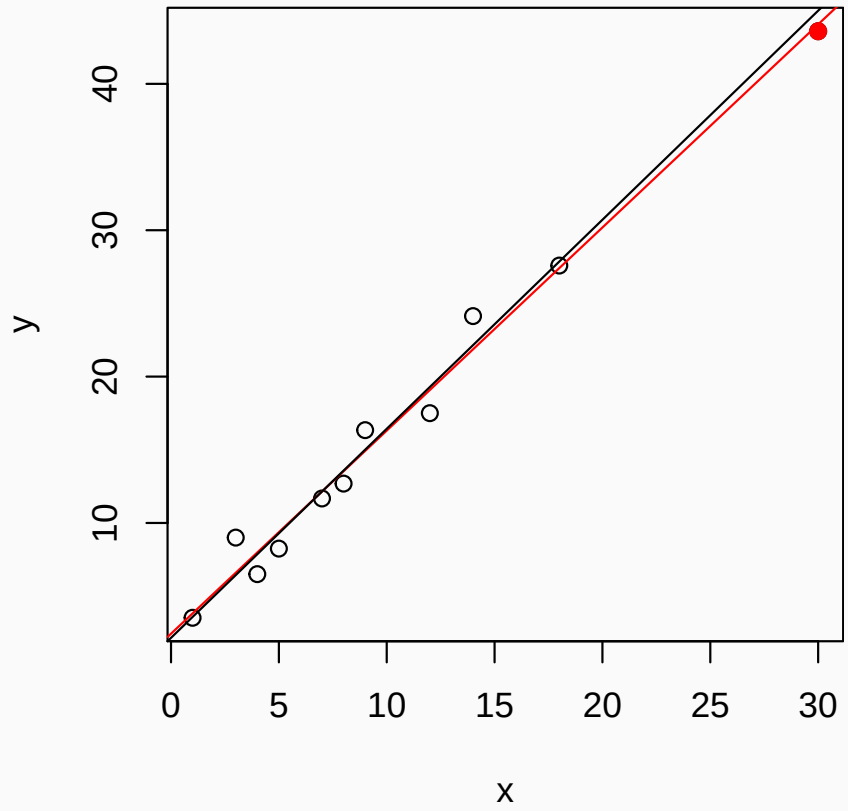
É possível ter um ponto com *leverage* alto, e influência baixa?

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

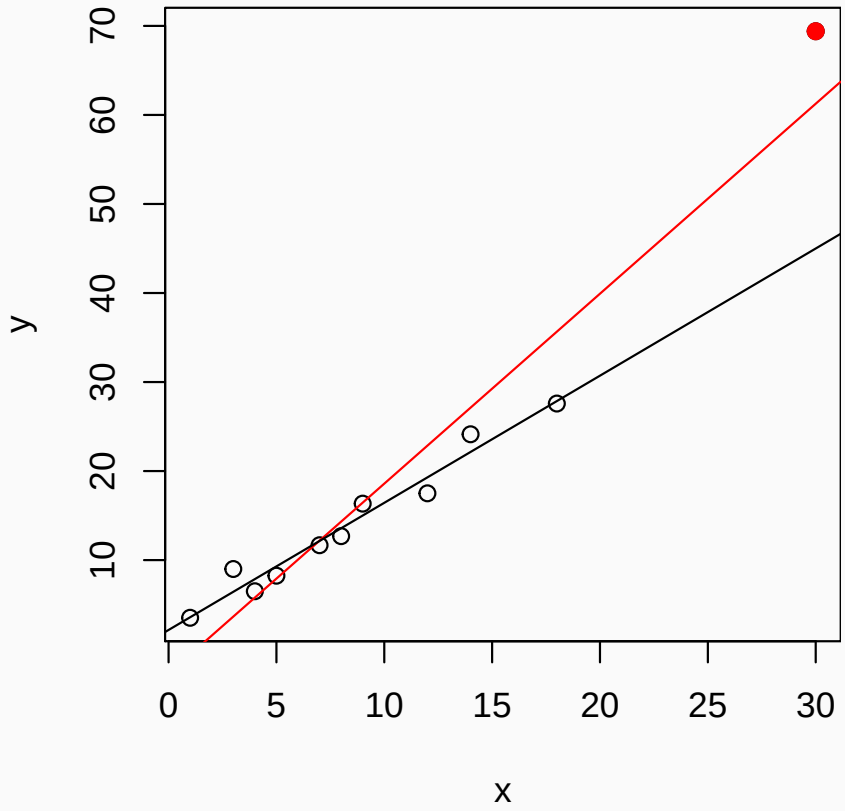


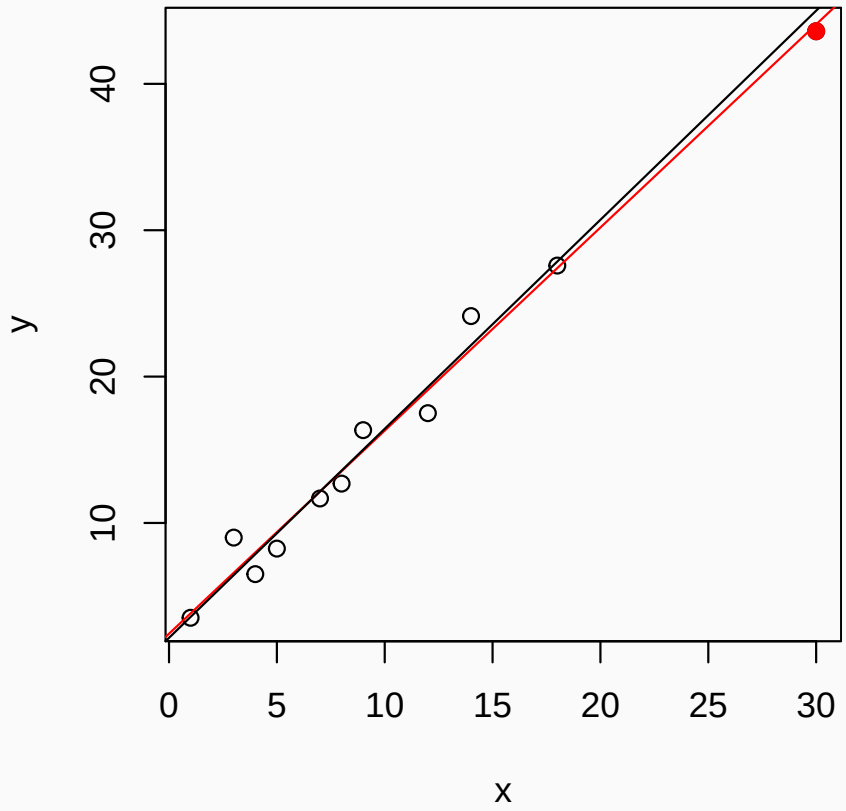
É possível ter um ponto com *leverage* alto, e influência baixa?





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

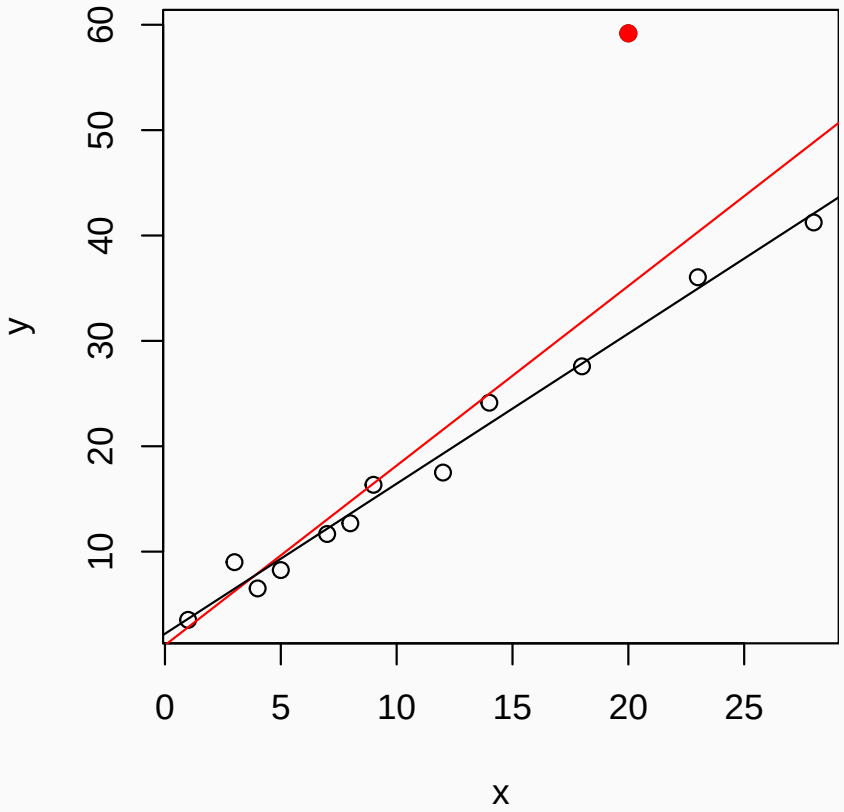




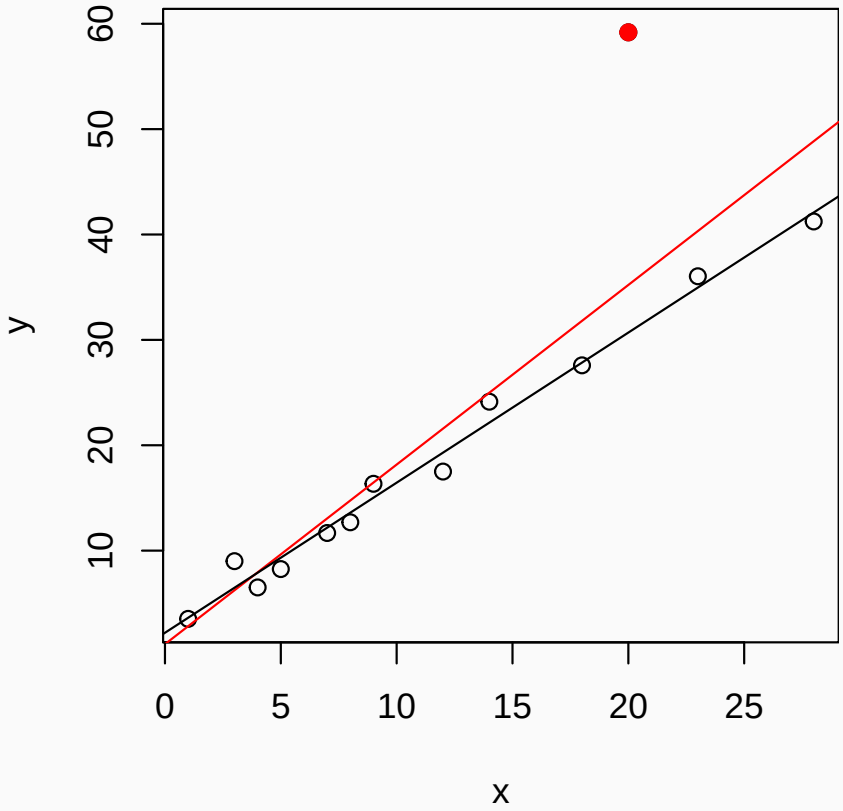
Sim, se a *distância* for baixa.

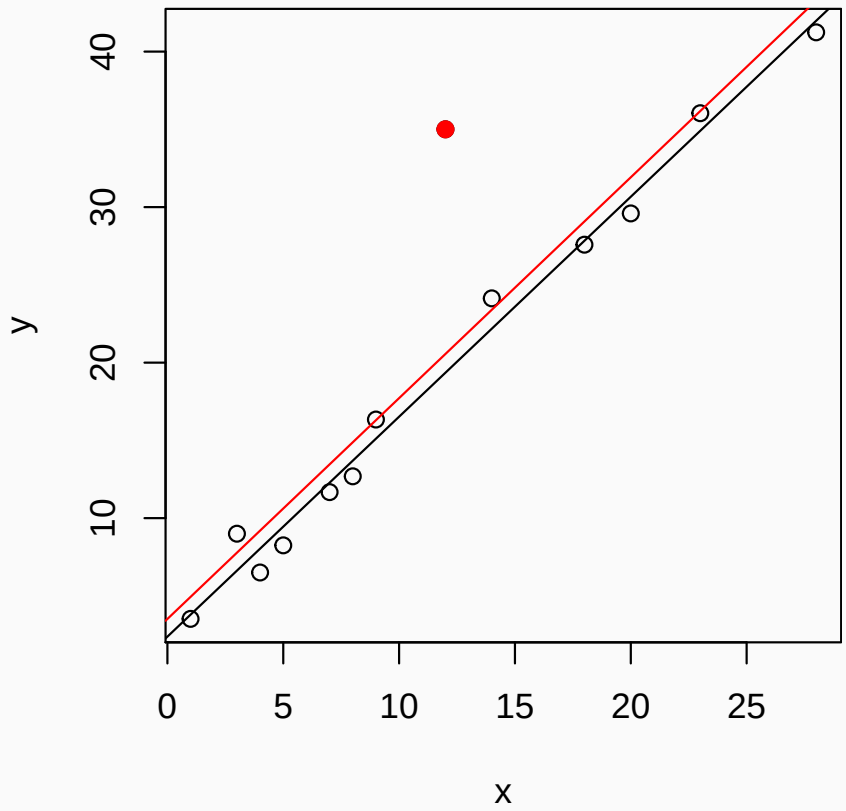
É possível ter um ponto com *distância* alta, e influência baixa?

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

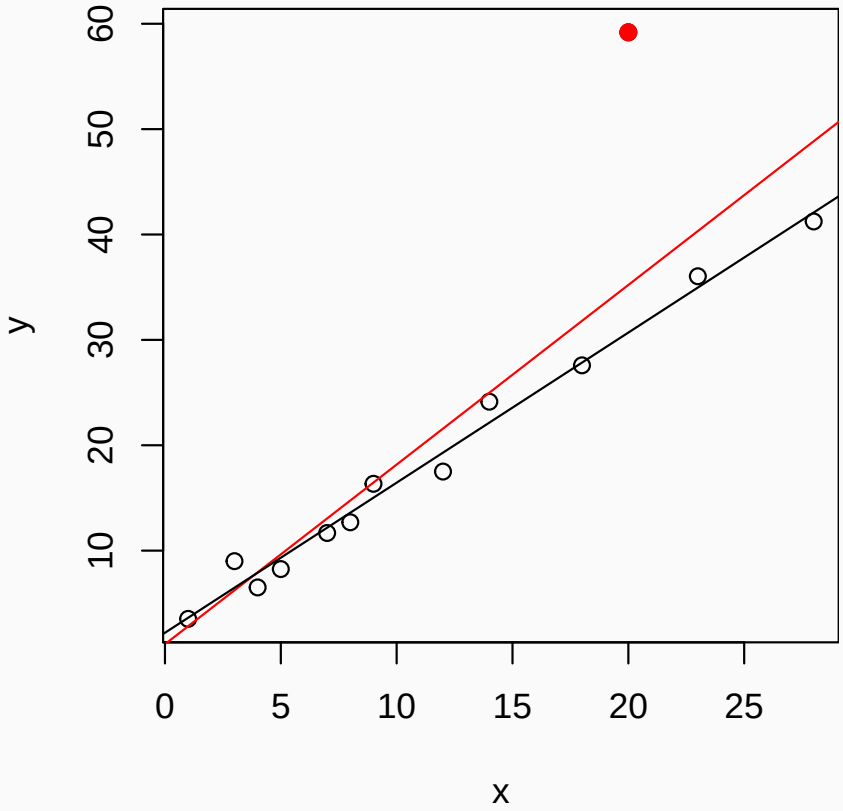


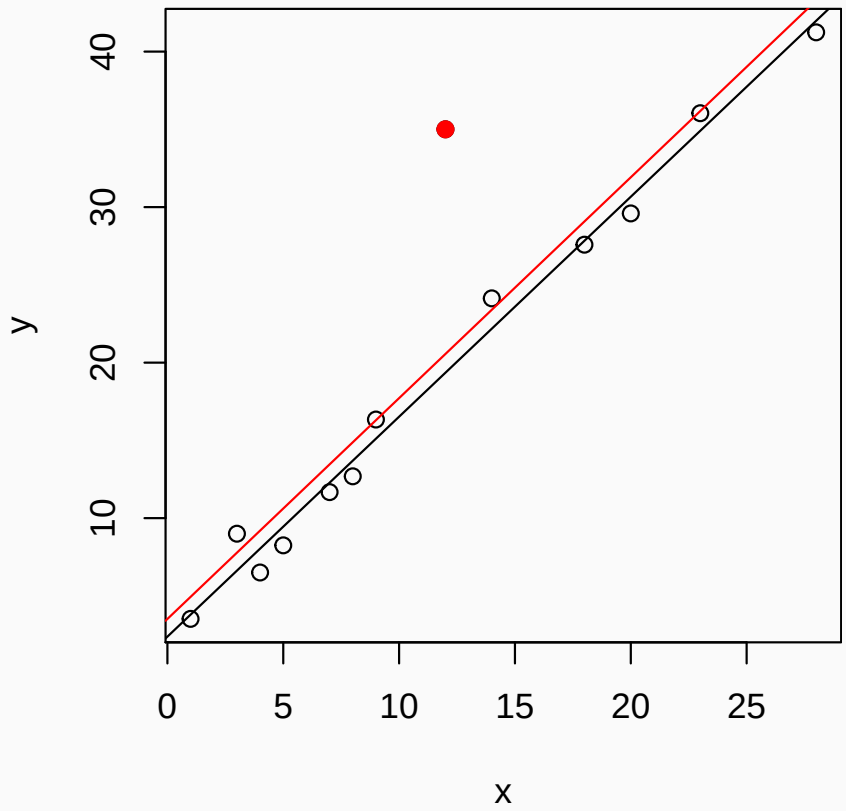
É possível ter um ponto com *distância* alta, e influência baixa?





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.

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.

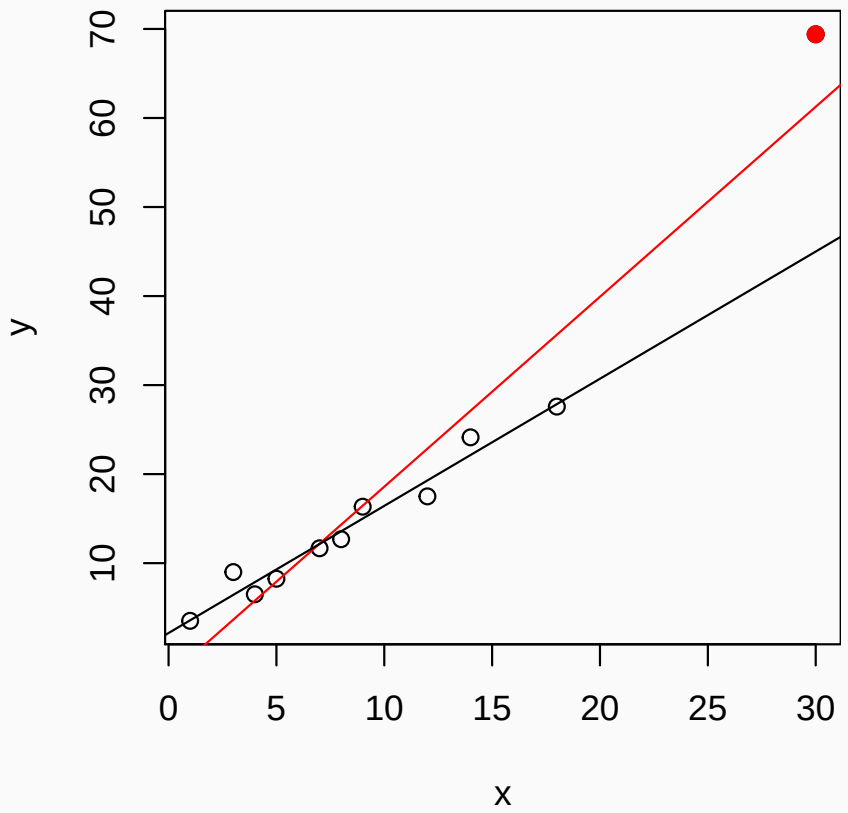
Como identificar pontos influentes?

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.

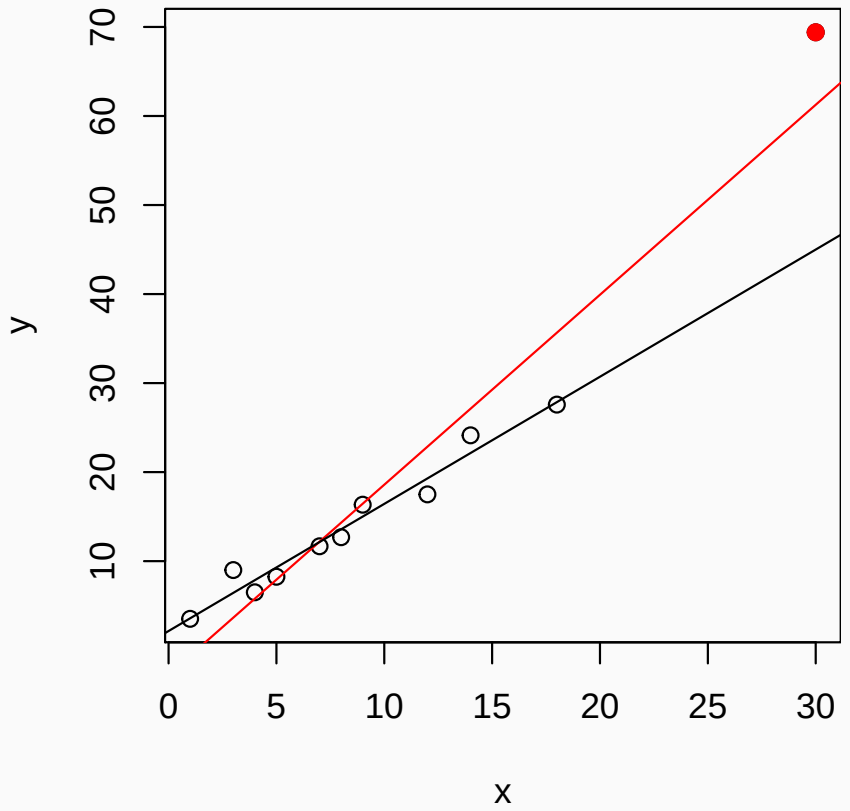
Distância de Cook: Similar a DFFITS, mas ao invés de avaliar a diferença em um único ponto, avalia a soma dos quadrados das diferenças de todos os $\hat{Y}$. Identifica pontos com influência sobre todas as estimativas de Y.

DFBETAS: Diferença normalizada entre os valores dos b 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.

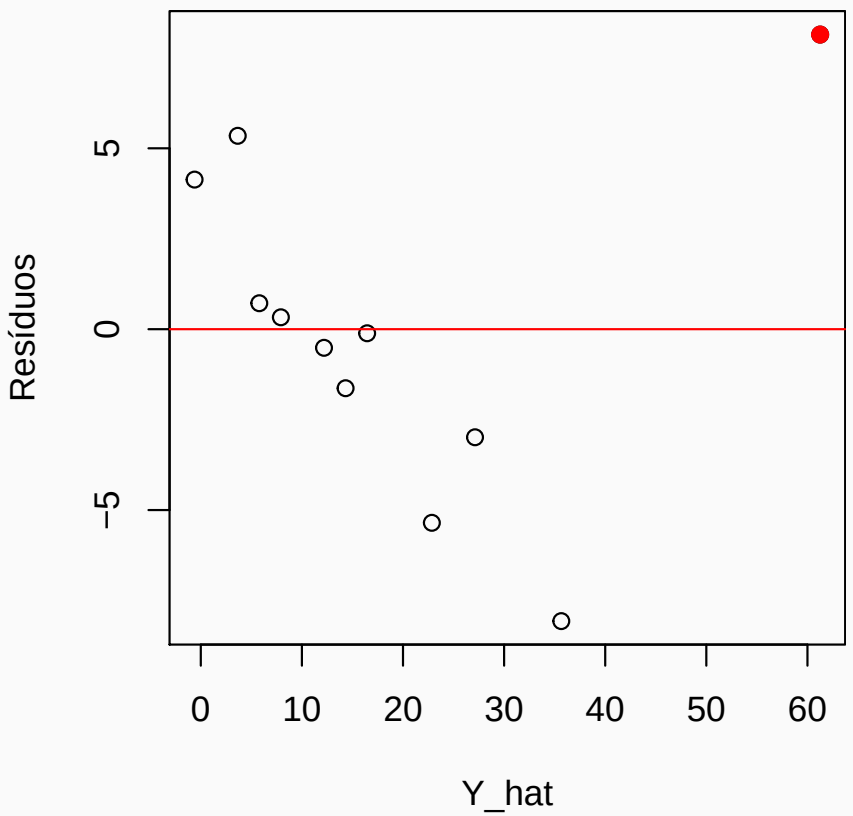
↑ LEVERAGE E ↑ DISTANCE



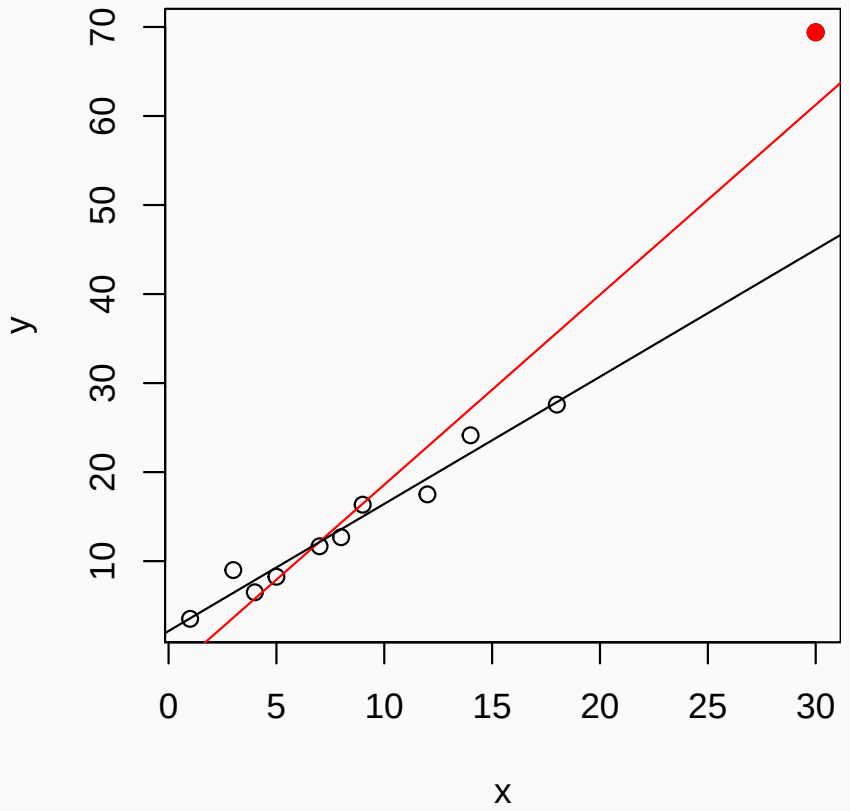
↑ LEVERAGE E ↑ DISTANCE



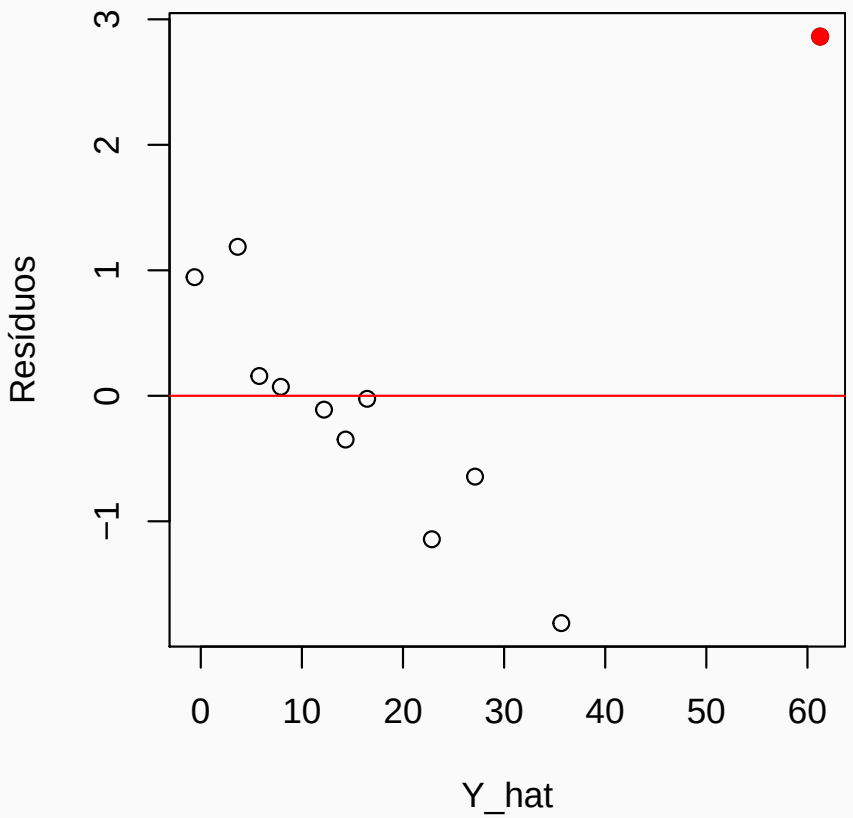

```
res <- residuals(m)
```



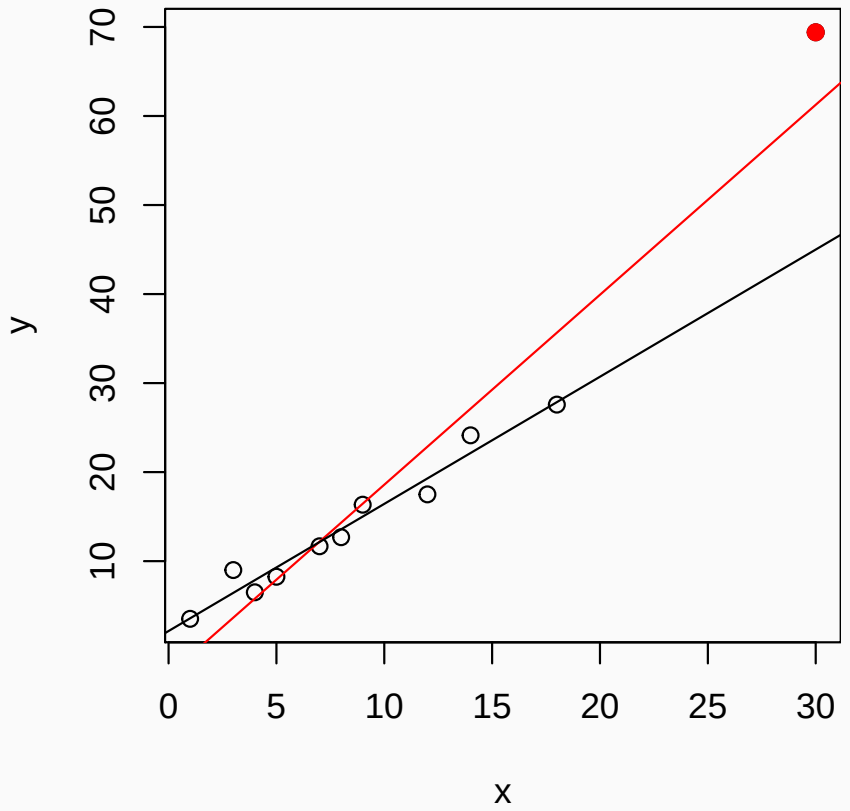
↑ LEVERAGE E ↑ DISTANCE



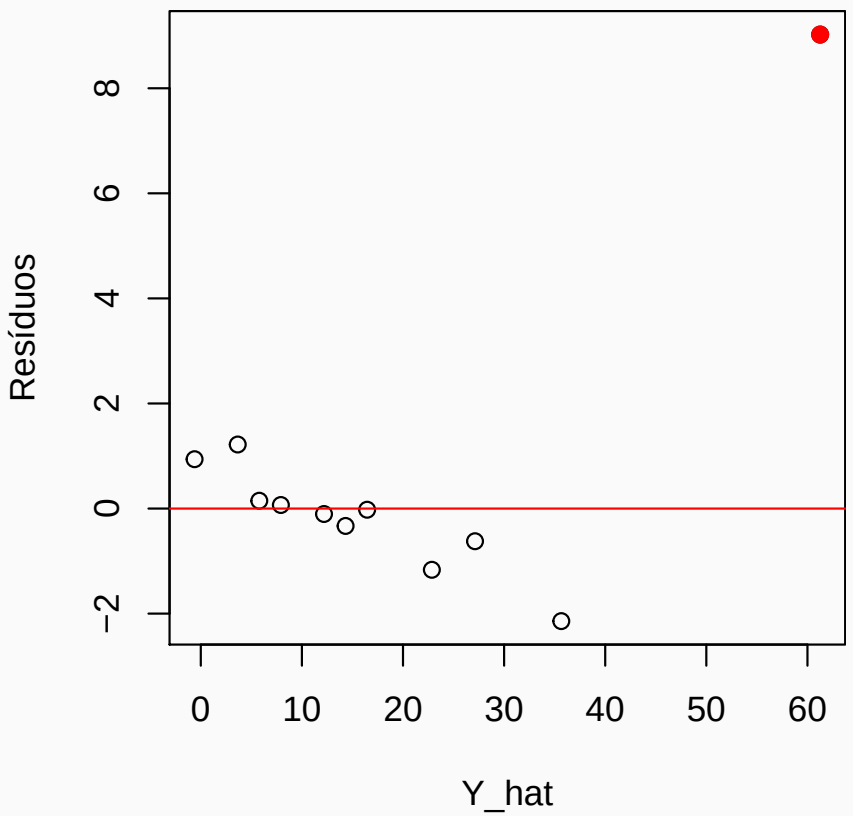

```
res <- rstandard(m)
```



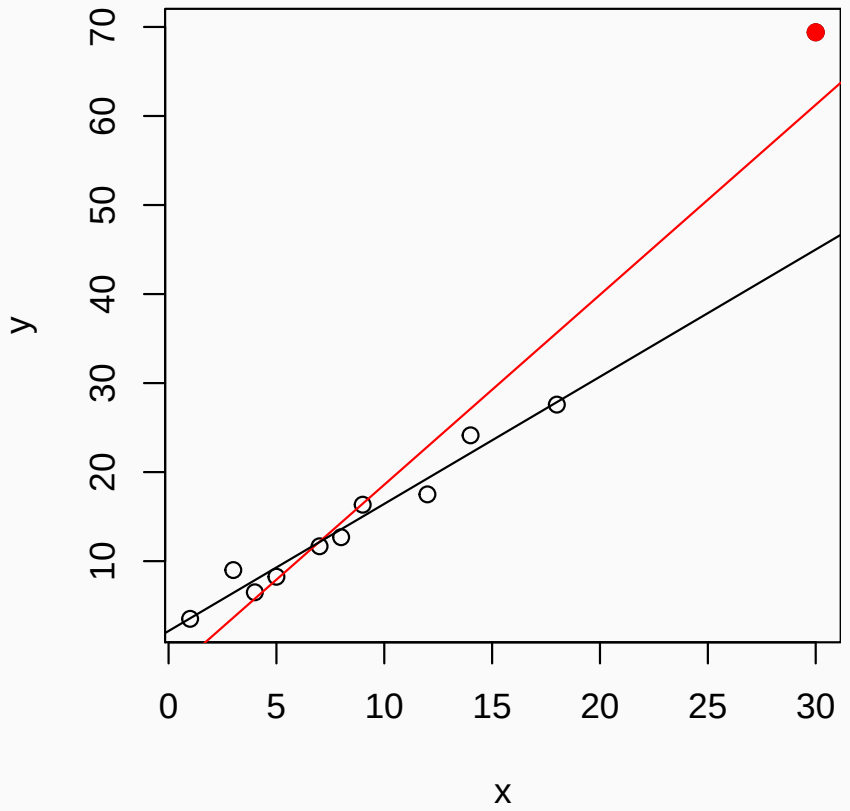
↑ LEVERAGE E ↑ DISTANCE



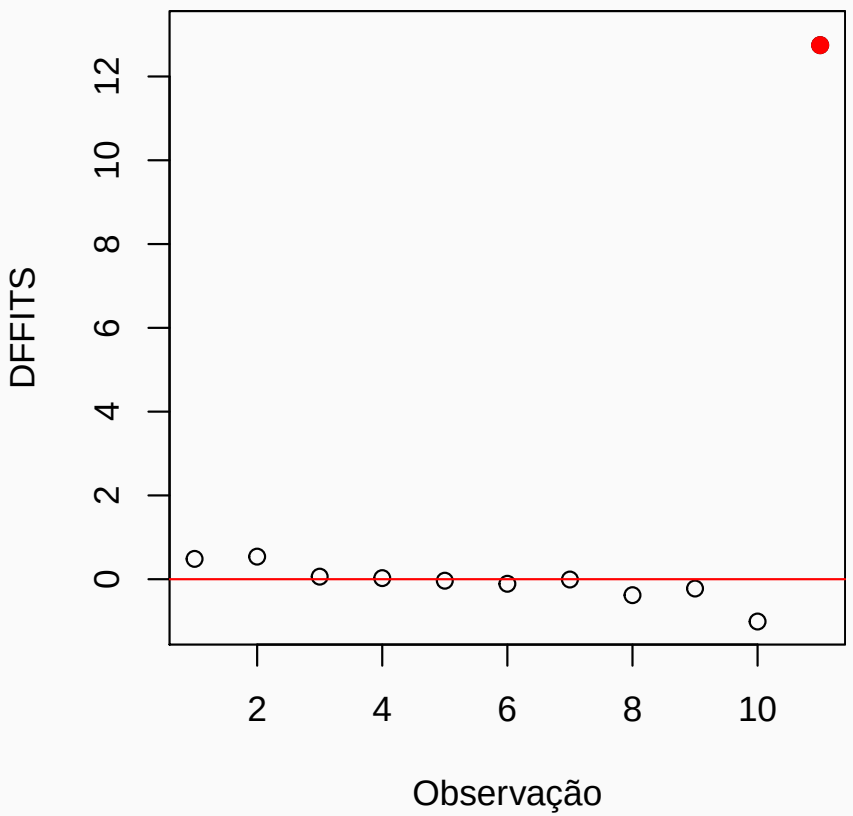

```
res <- rstudent(m)
```



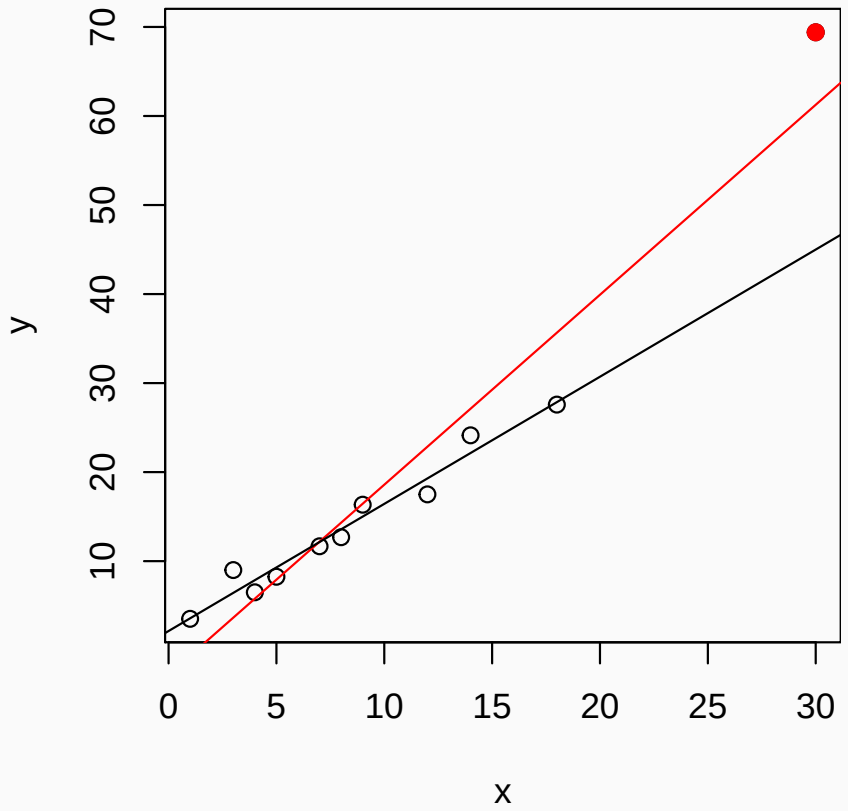
↑ LEVERAGE E ↑ DISTANCE



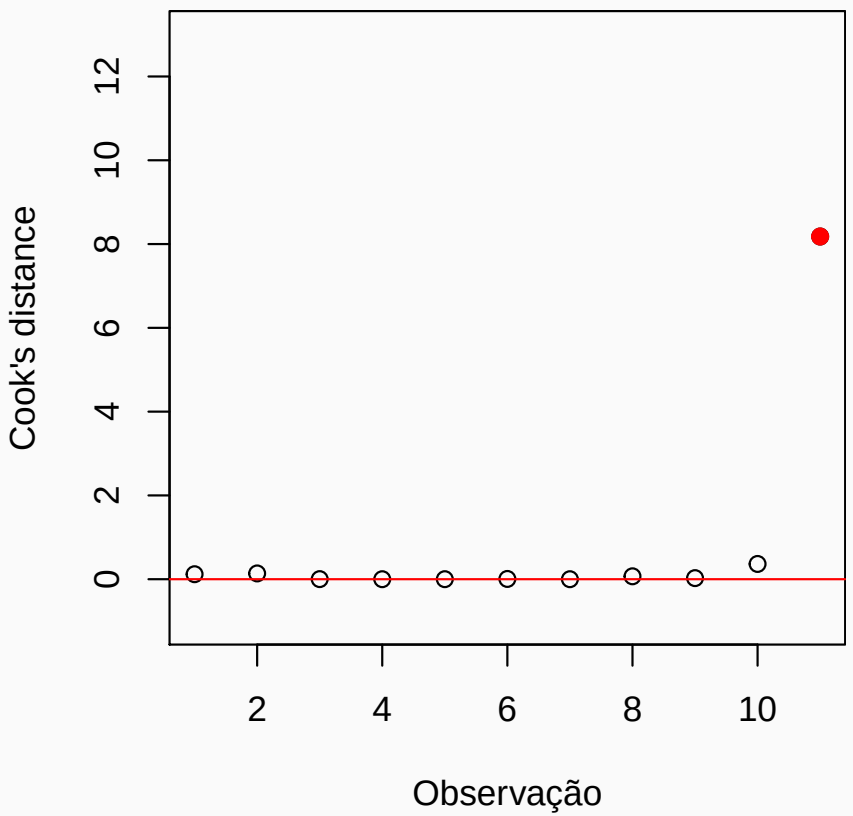

```
fits <- dffits(m)
```



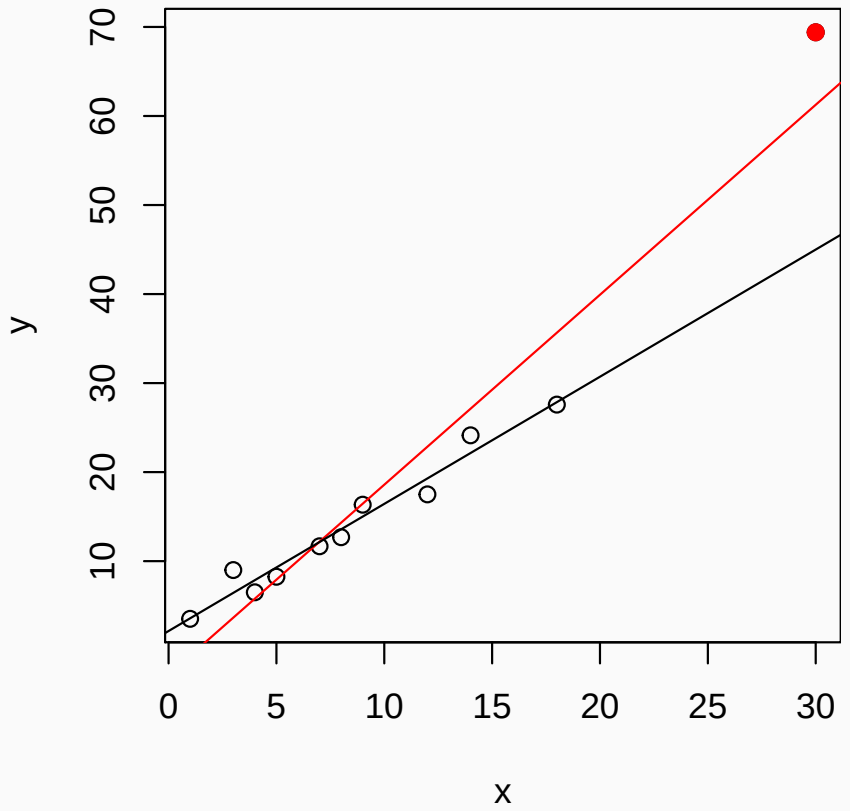
↑ LEVERAGE E ↑ DISTANCE



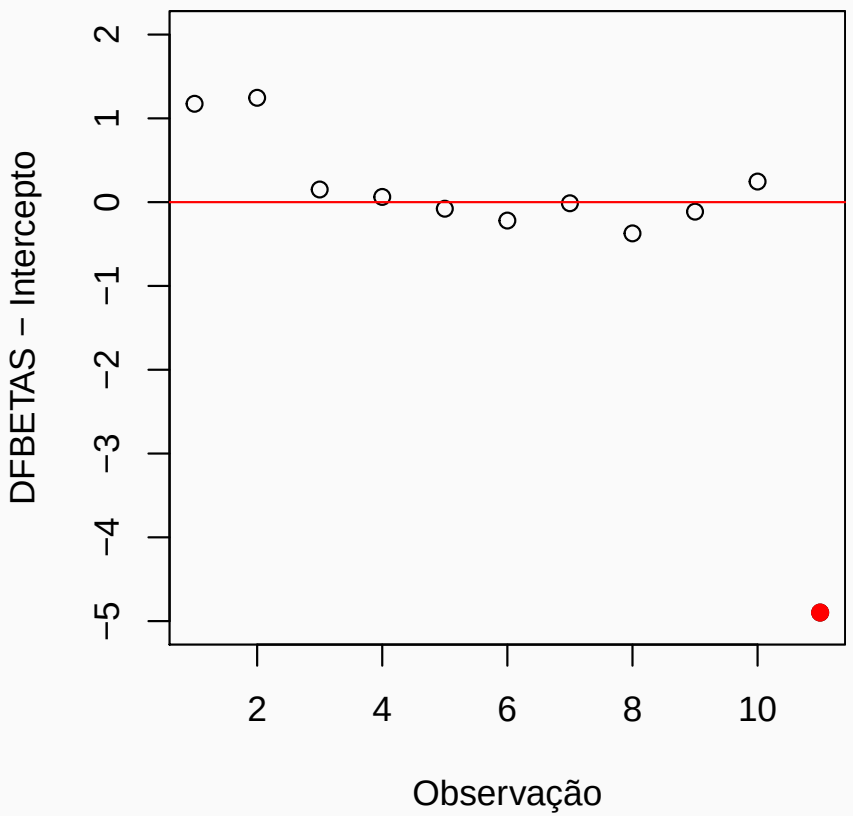

```
fits <- cooks.distance(m)
```



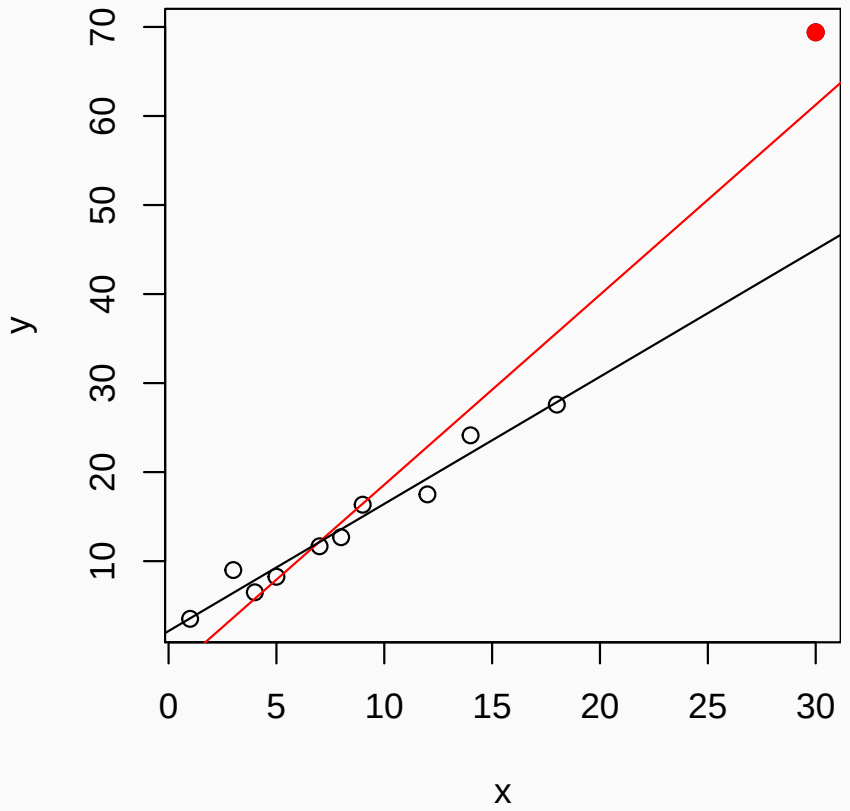
↑ LEVERAGE E ↑ DISTANCE



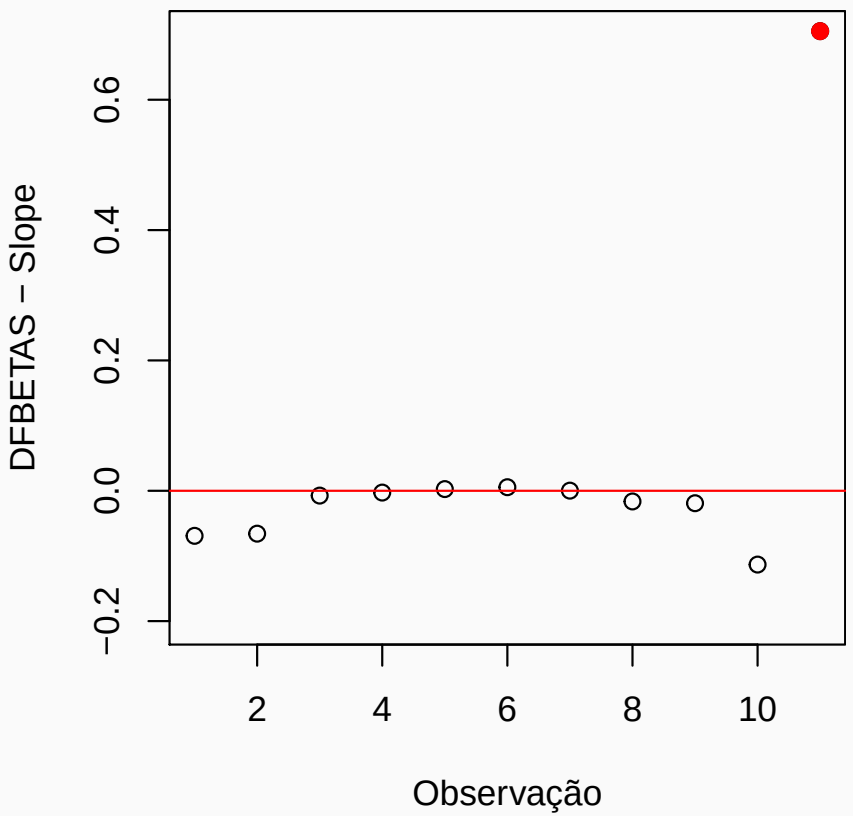

```
fits <- dfbeta(m)
```



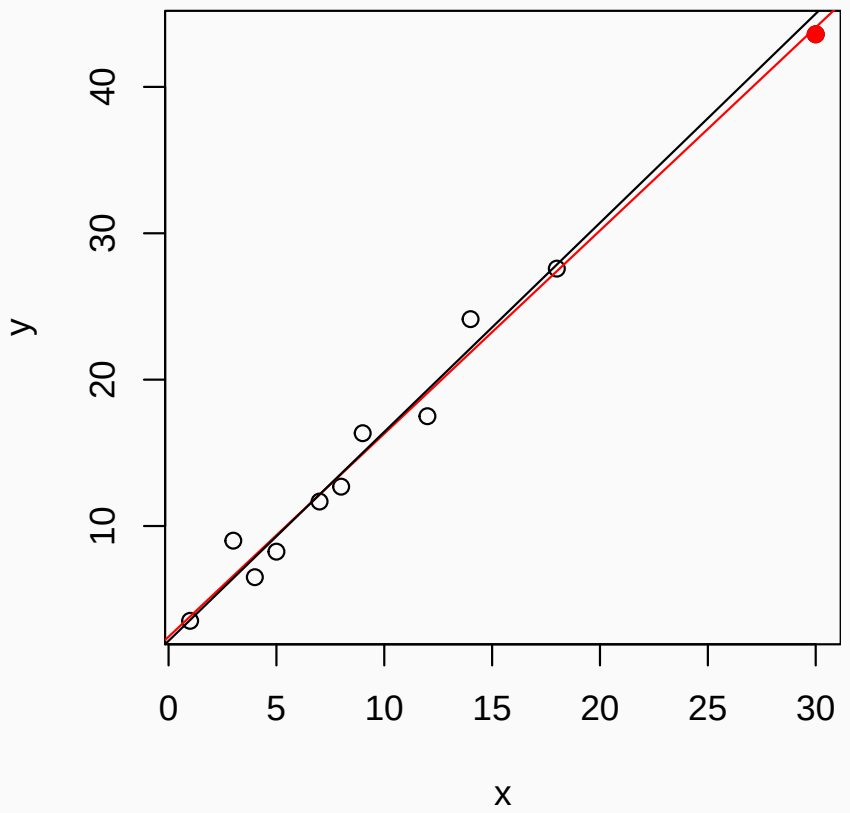
↑ LEVERAGE E ↑ DISTANCE



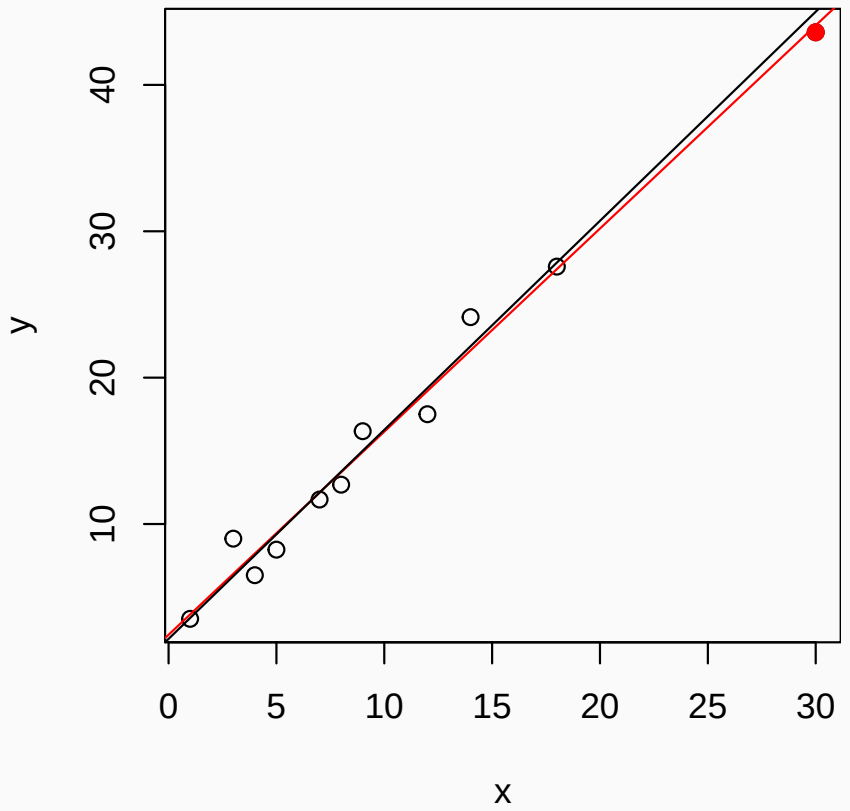

```
fits <- dfbeta(m)
```



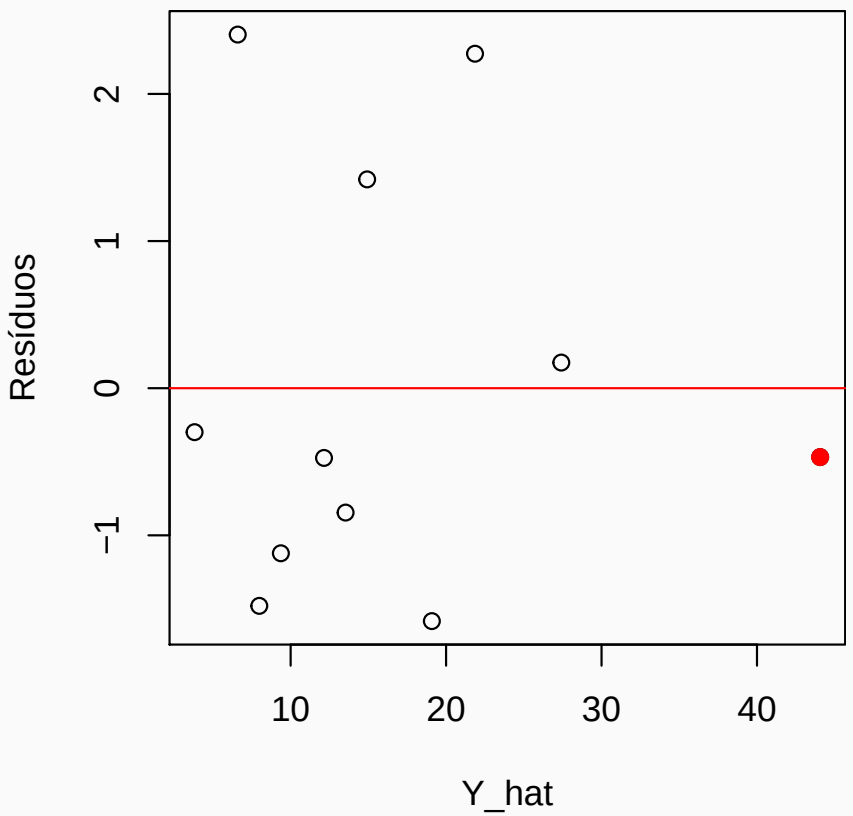
↑ LEVERAGE E ↓ DISTANCE



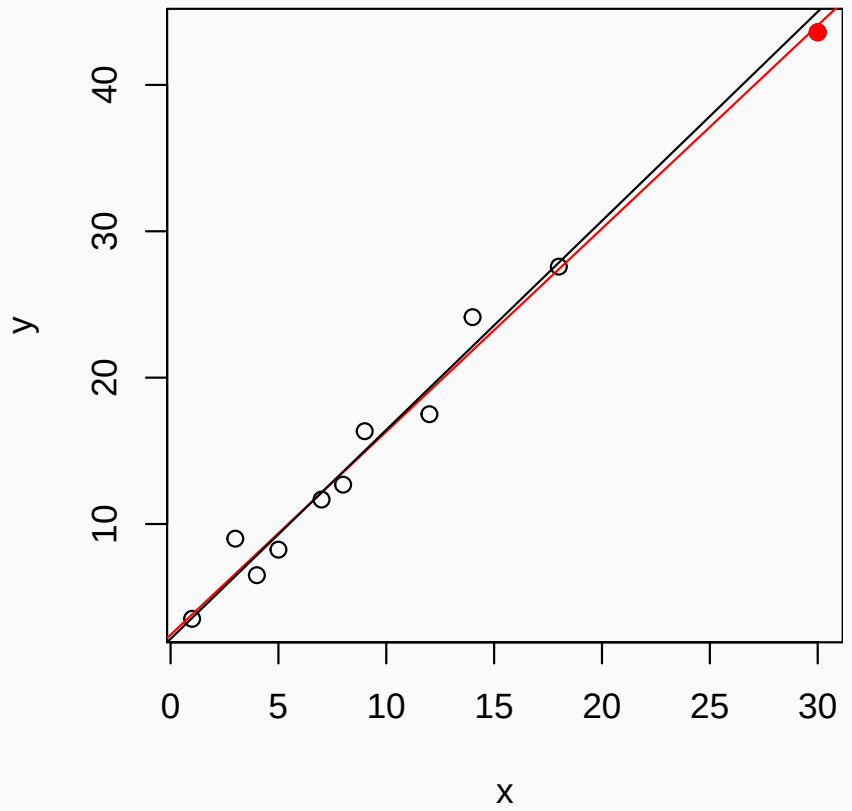
↑ LEVERAGE E ↓ DISTANCE



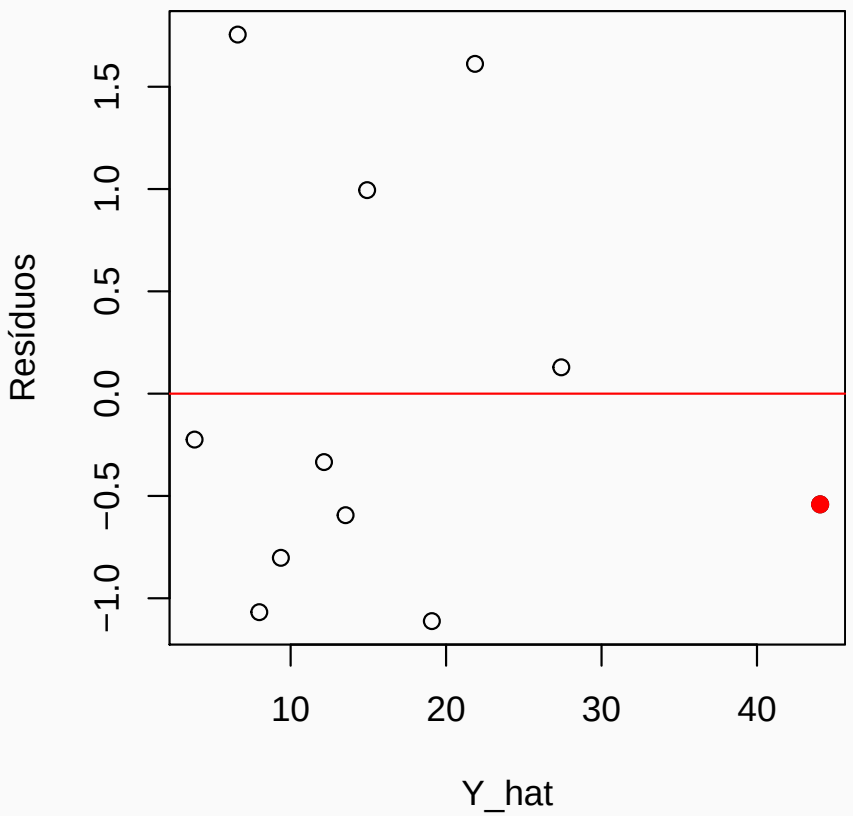

```
res <- residuals(m)
```



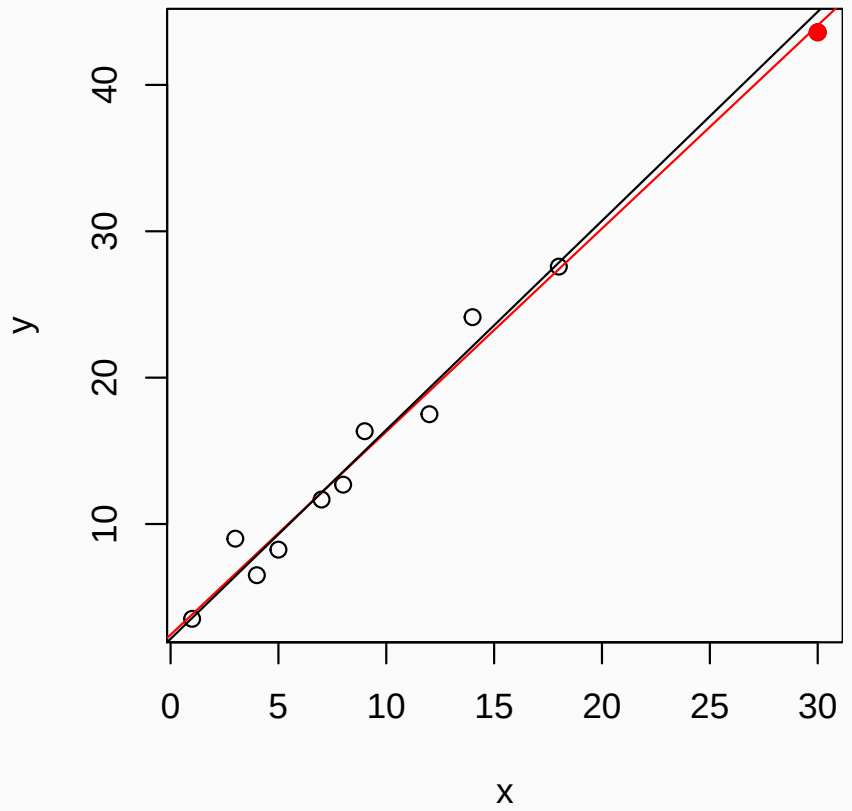
↑ LEVERAGE E ↓ DISTANCE



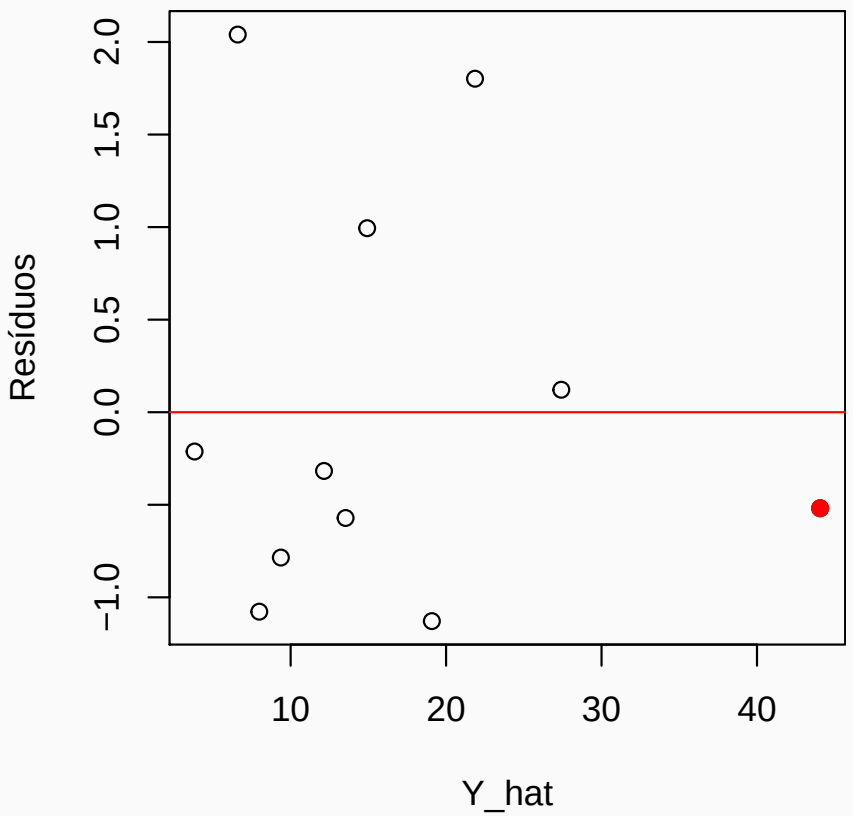

```
res <- rstandard(m)
```



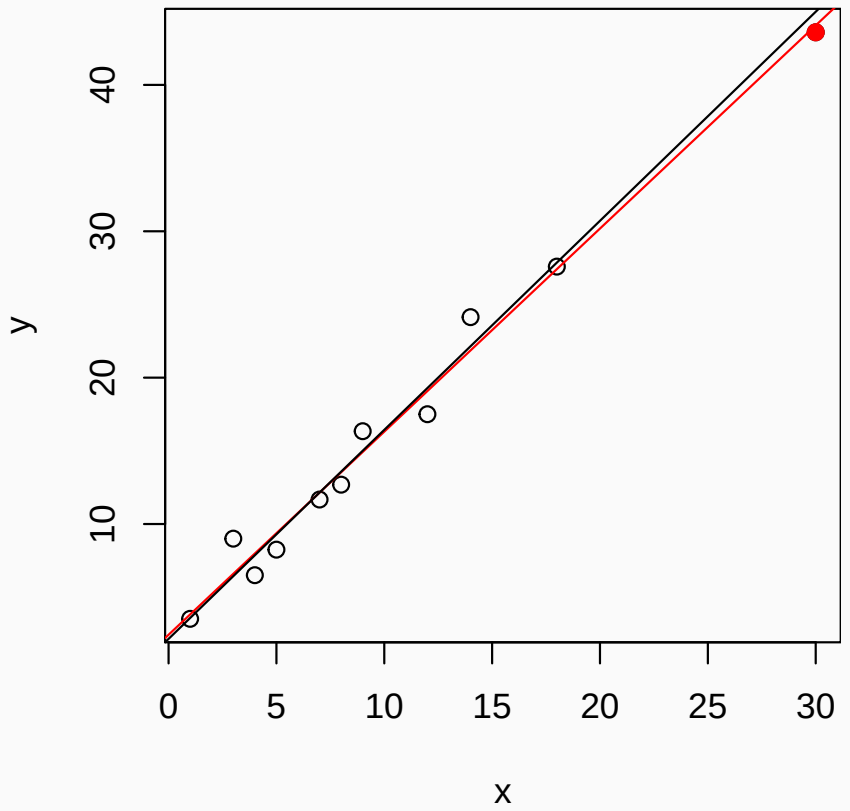
↑ LEVERAGE E ↓ DISTANCE



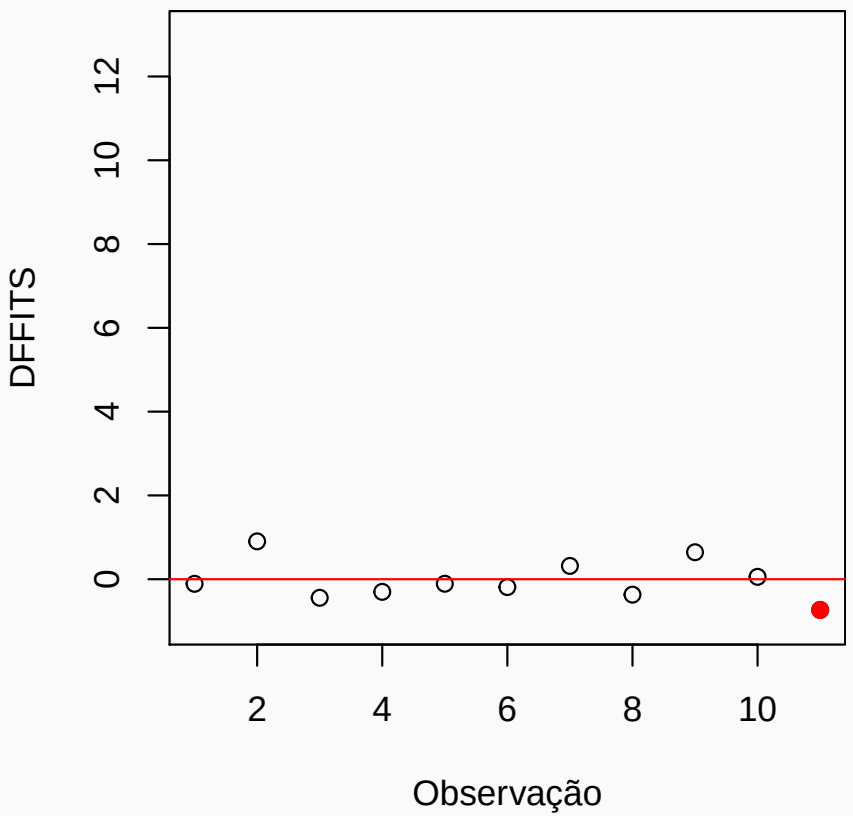

```
res <- rstudent(m)
```



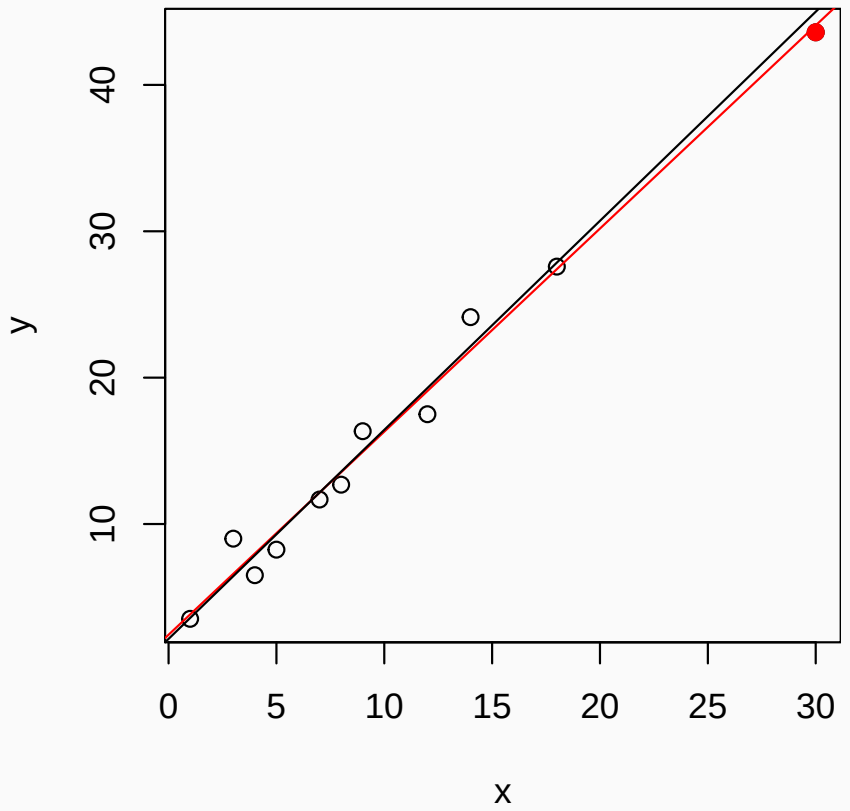
↑ LEVERAGE E ↓ DISTANCE



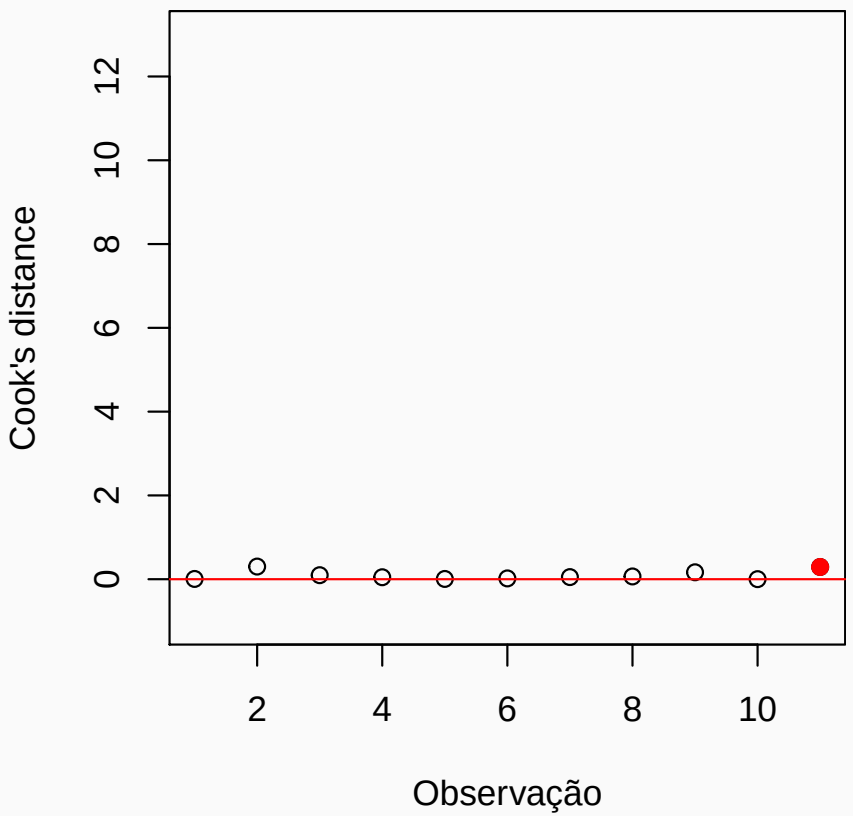

```
fits <- dffits(m)
```



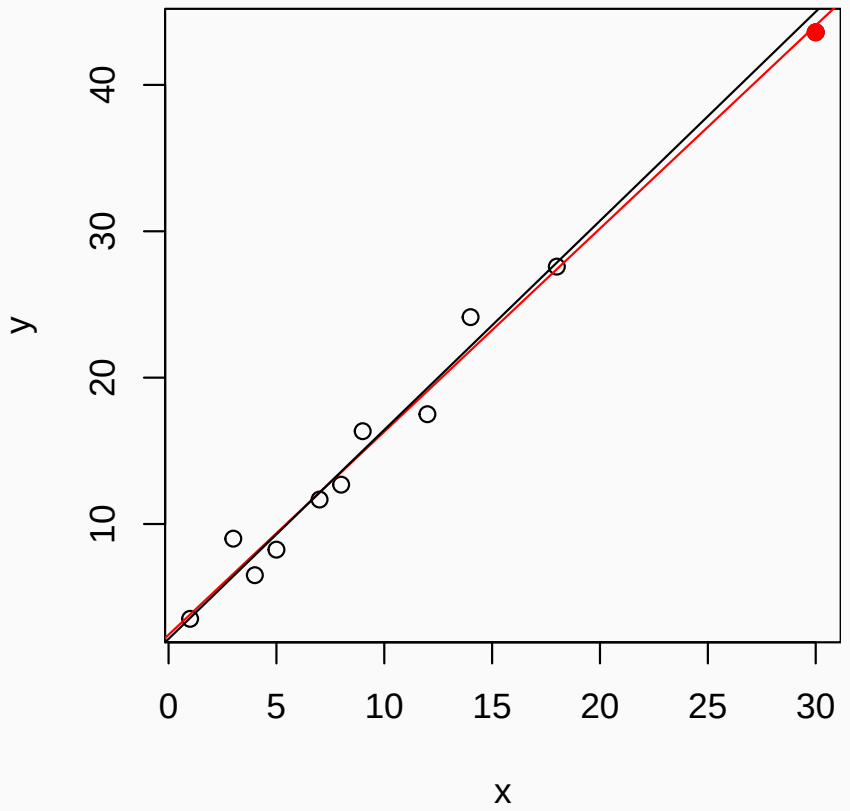
↑ LEVERAGE E ↓ DISTANCE



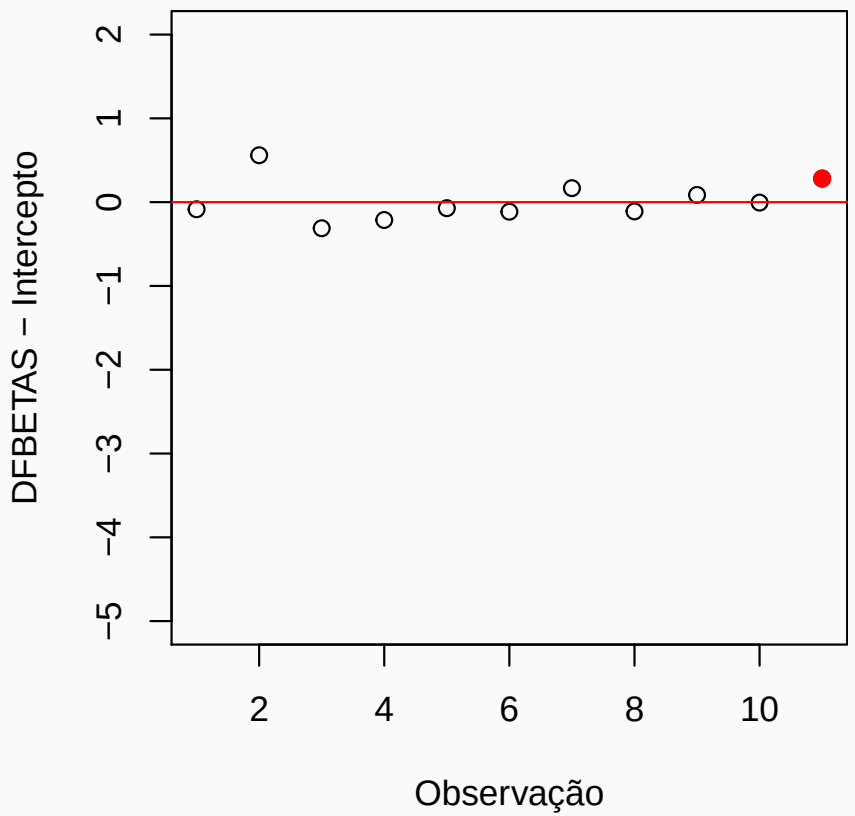

```
fits <- cooks.distance(m)
```



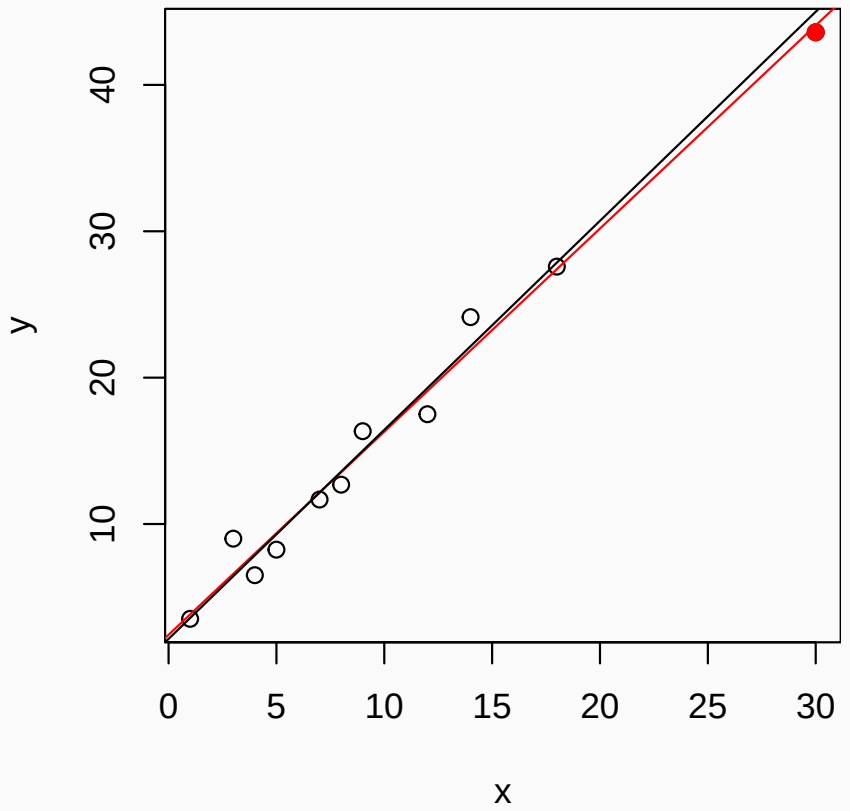
↑ LEVERAGE E ↓ DISTANCE



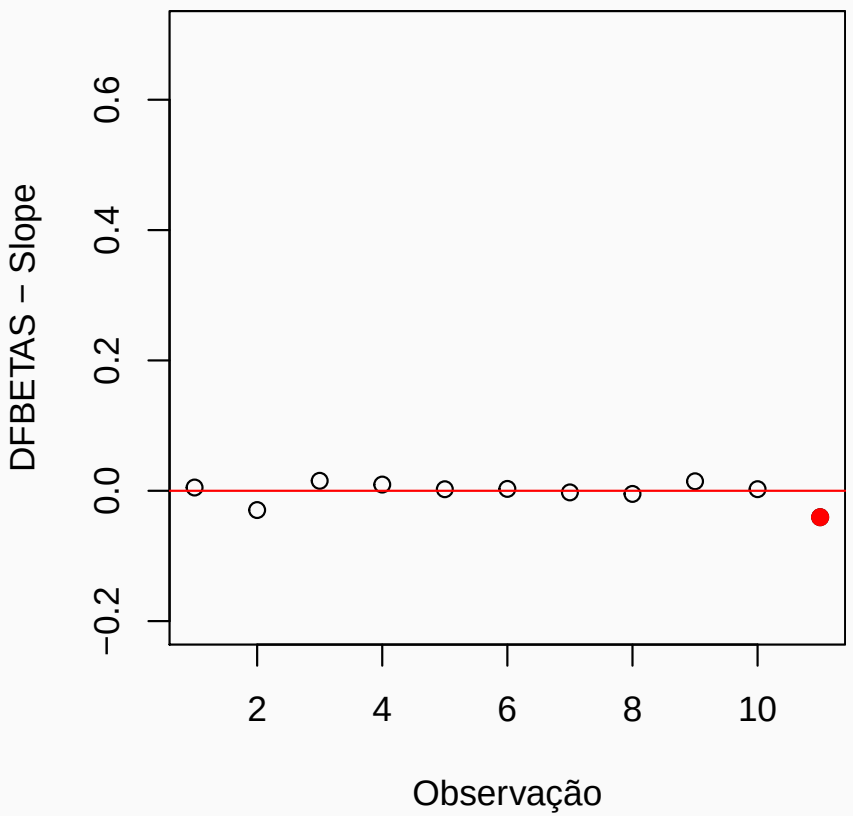

```
fits <- dfbeta(m)
```



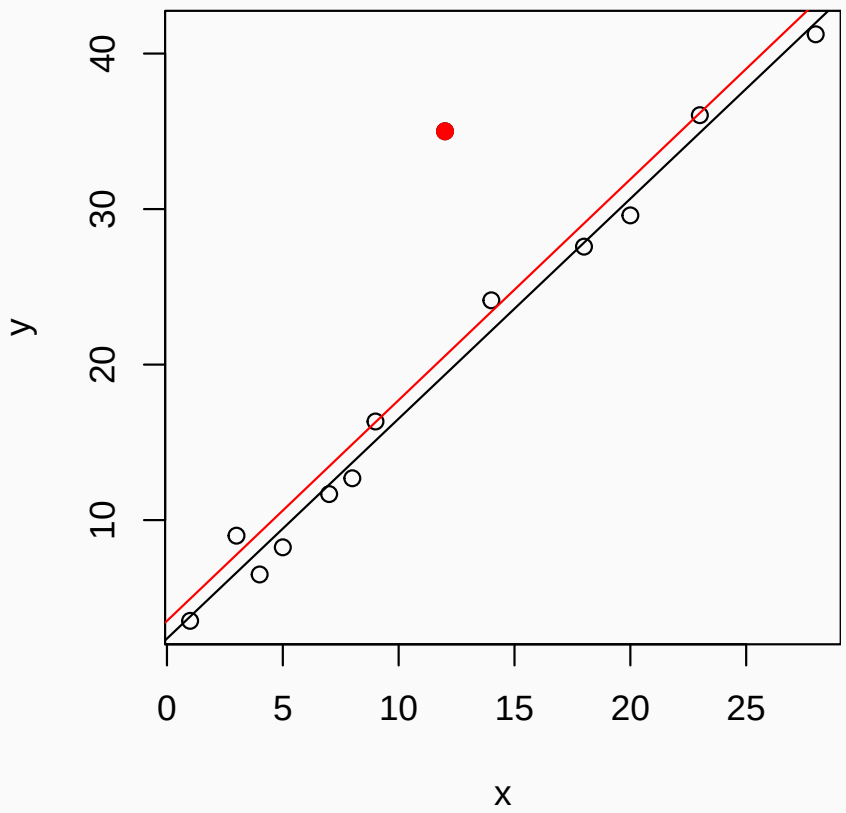
↑ LEVERAGE E ↓ DISTANCE



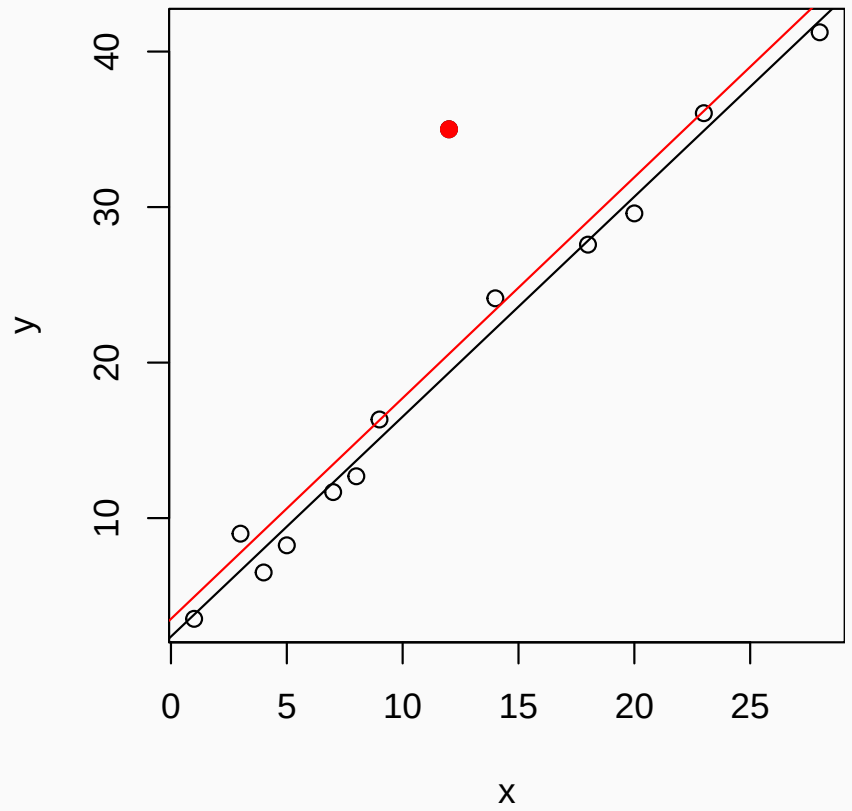

```
fits <- dfbeta(m)
```



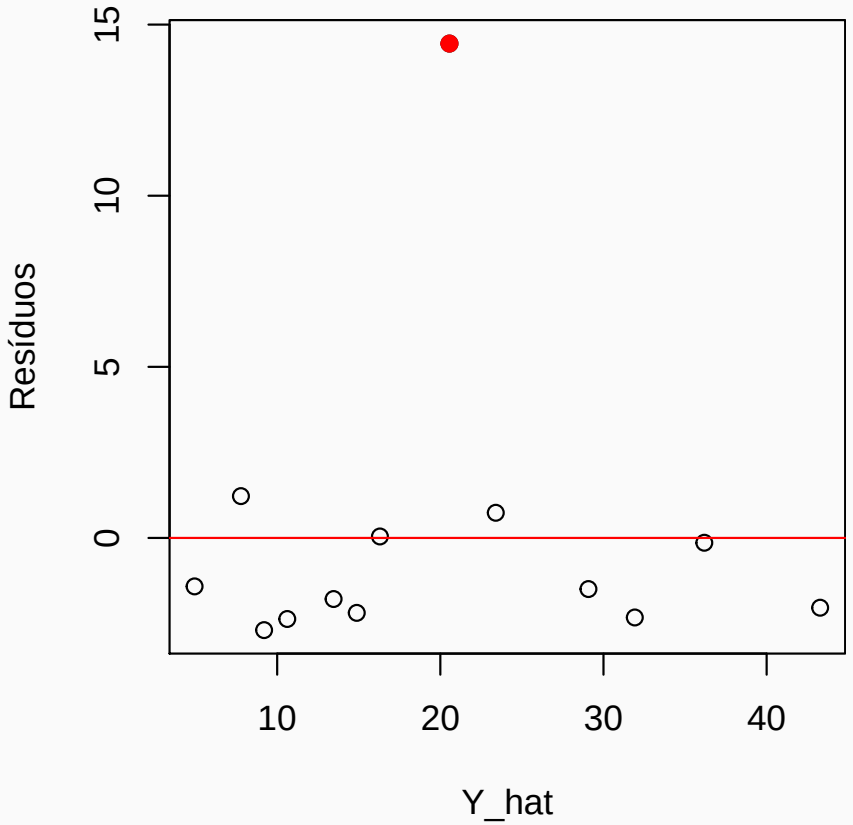
↓ LEVERAGE E ↑ DISTANCE



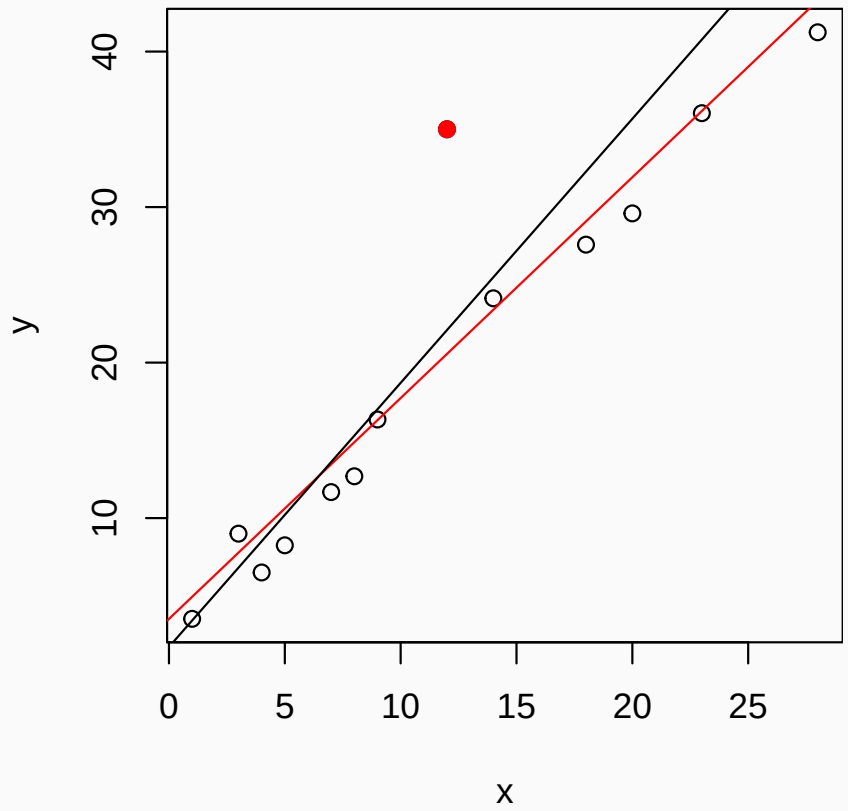
↓ LEVERAGE E ↑ DISTANCE



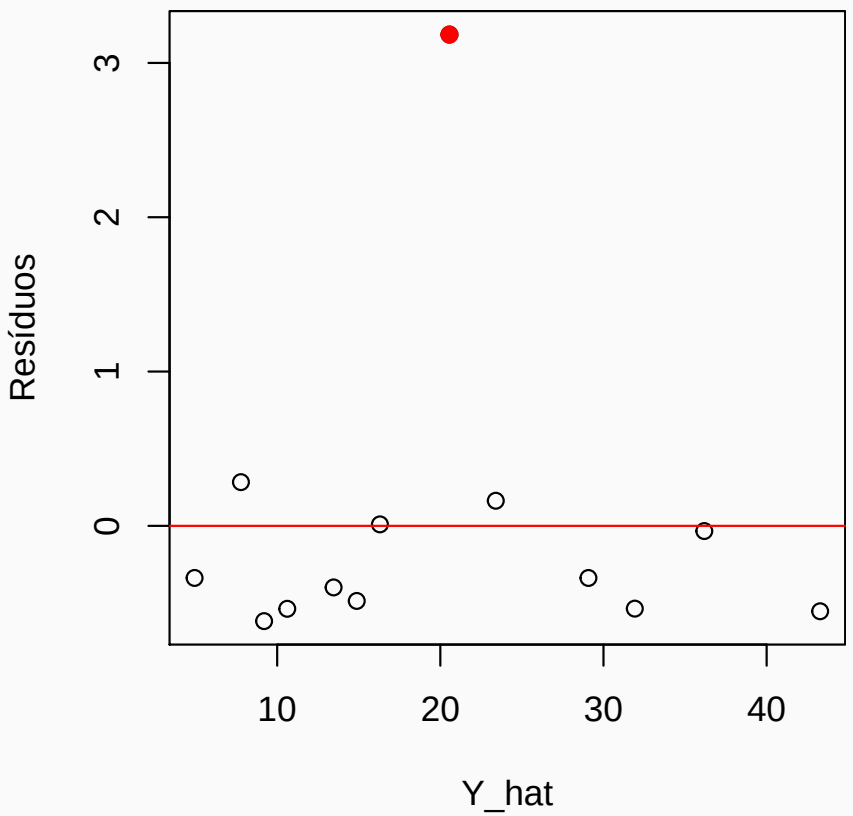

```
res <- residuals(m)
```



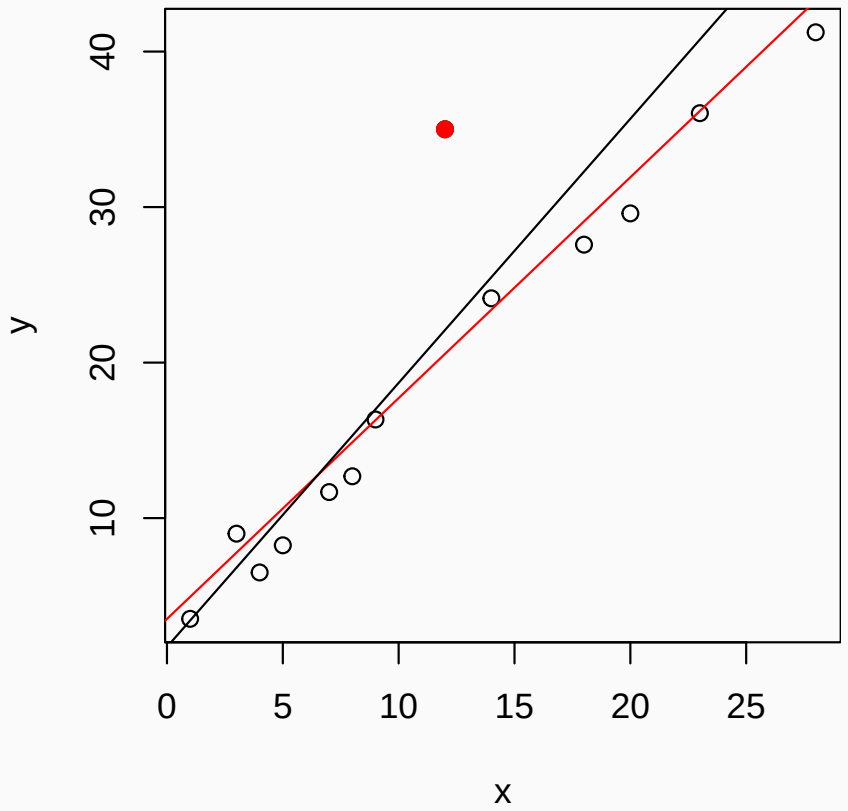
↑ LEVERAGE E ↓ DISTANCE



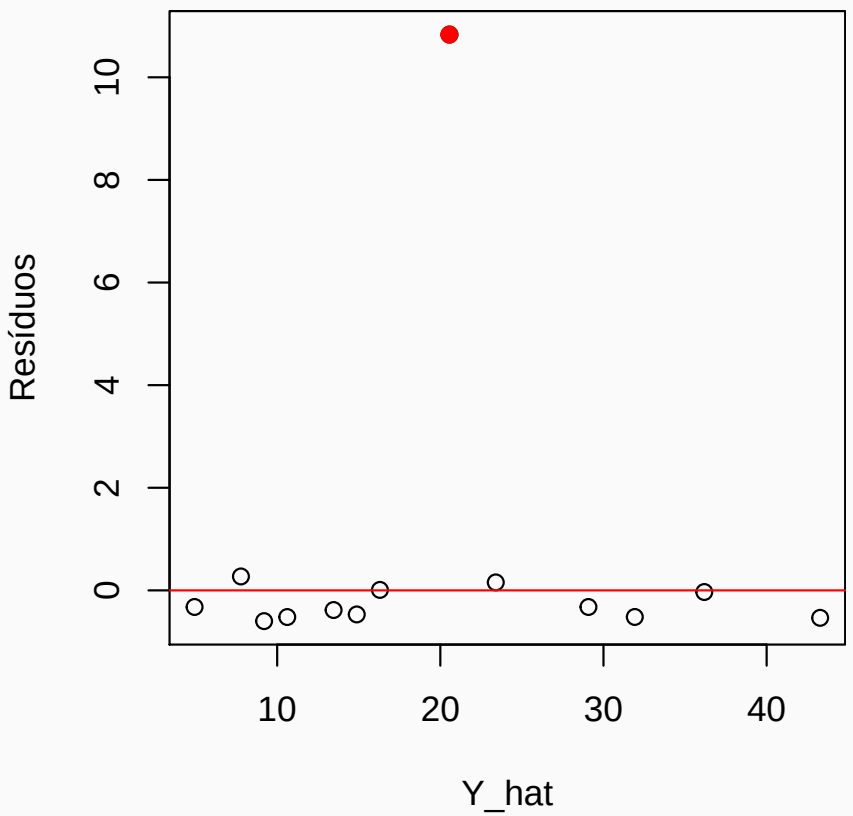

```
res <- rstandard(m)
```



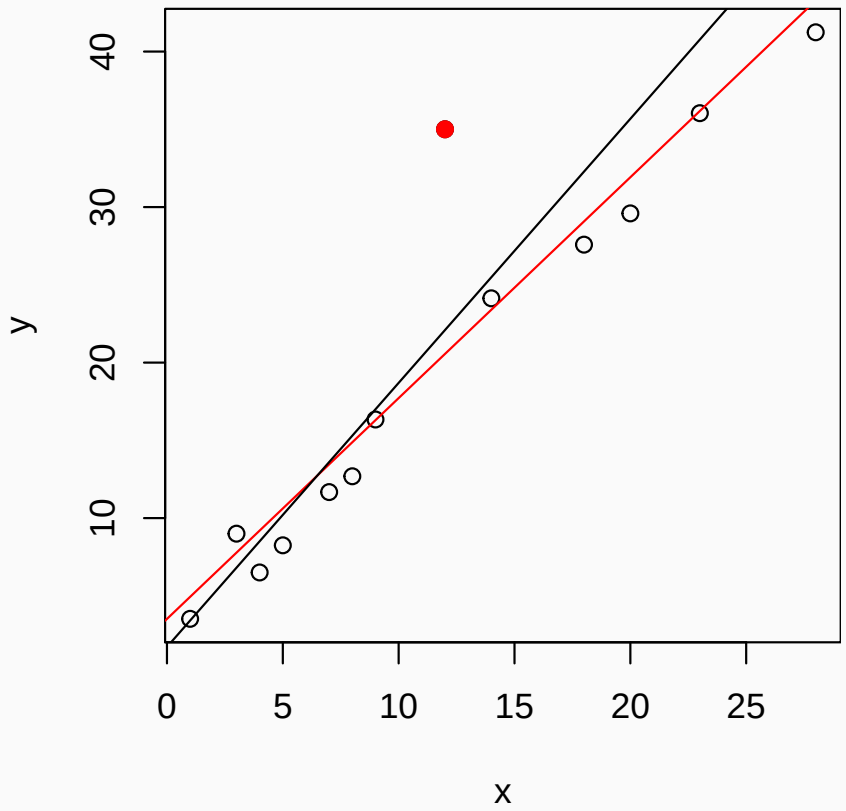
↑ LEVERAGE E ↓ DISTANCE



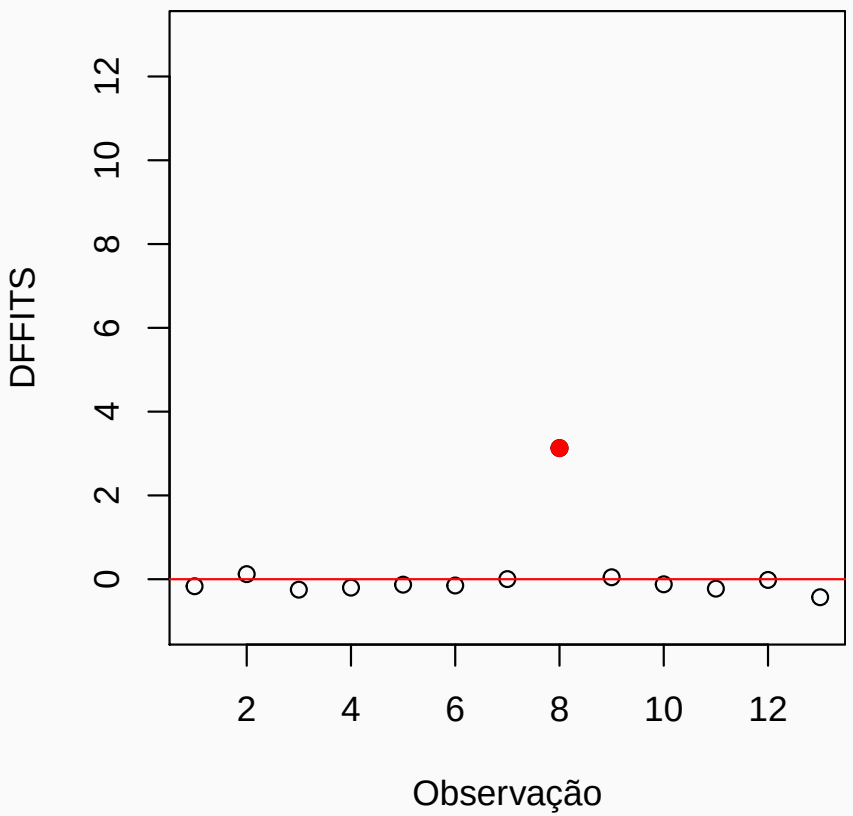

```
res <- rstudent(m)
```



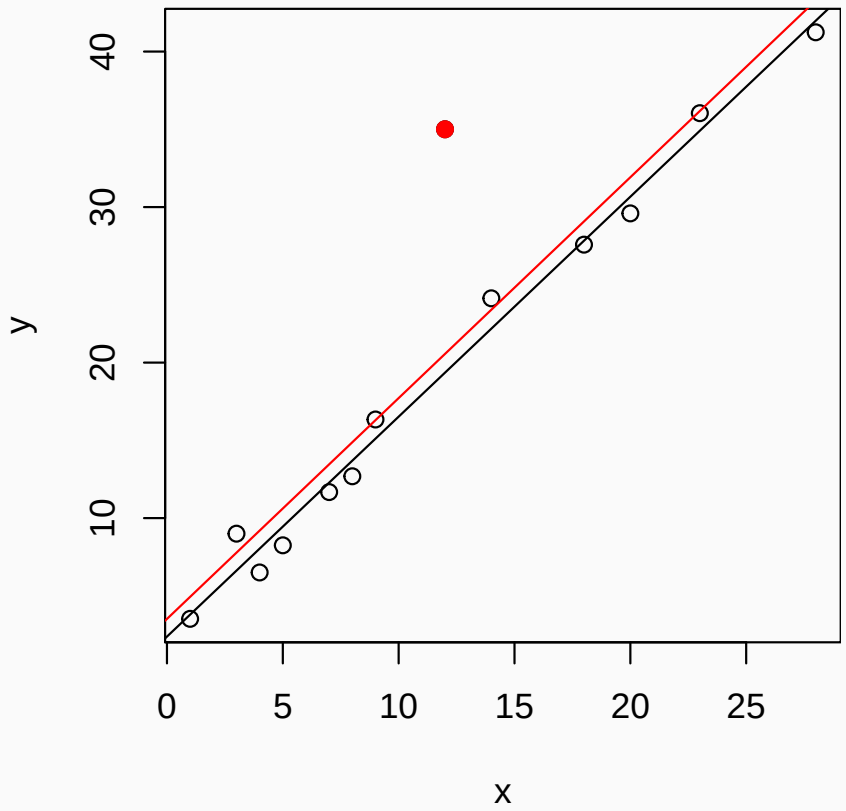
↑ LEVERAGE E ↓ DISTANCE



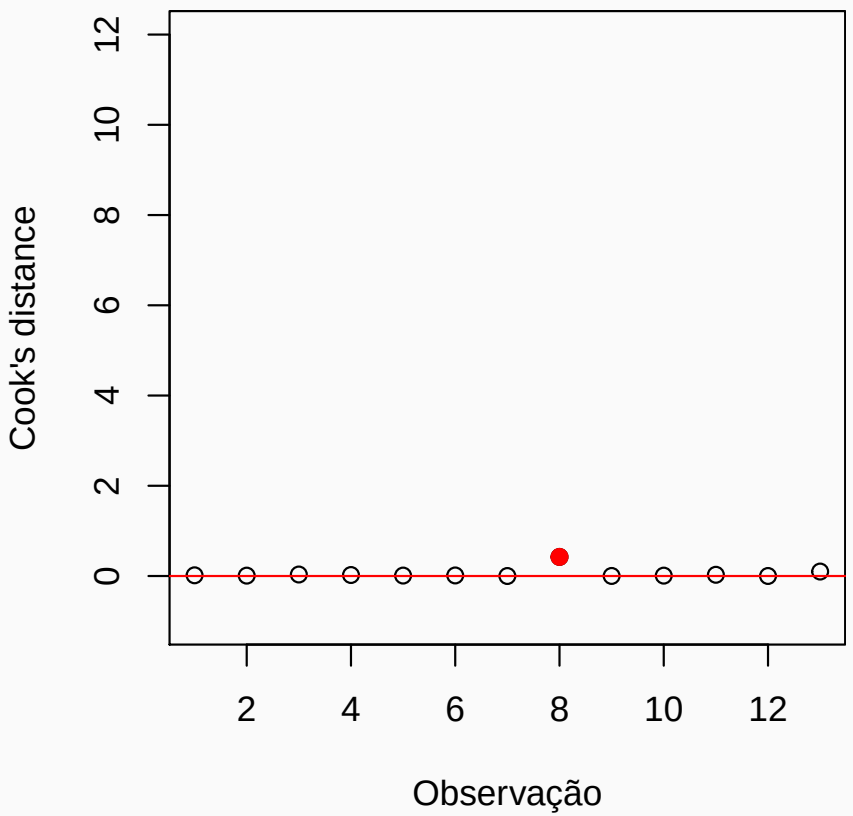

```
fits <- dffits(m)
```



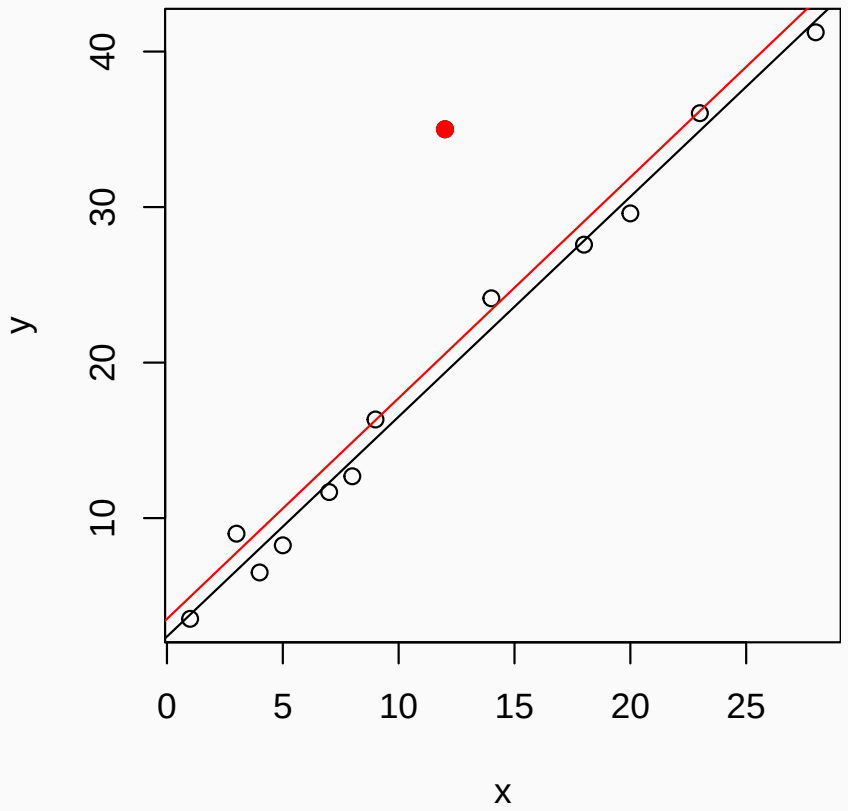
↓ LEVERAGE E ↑ DISTANCE



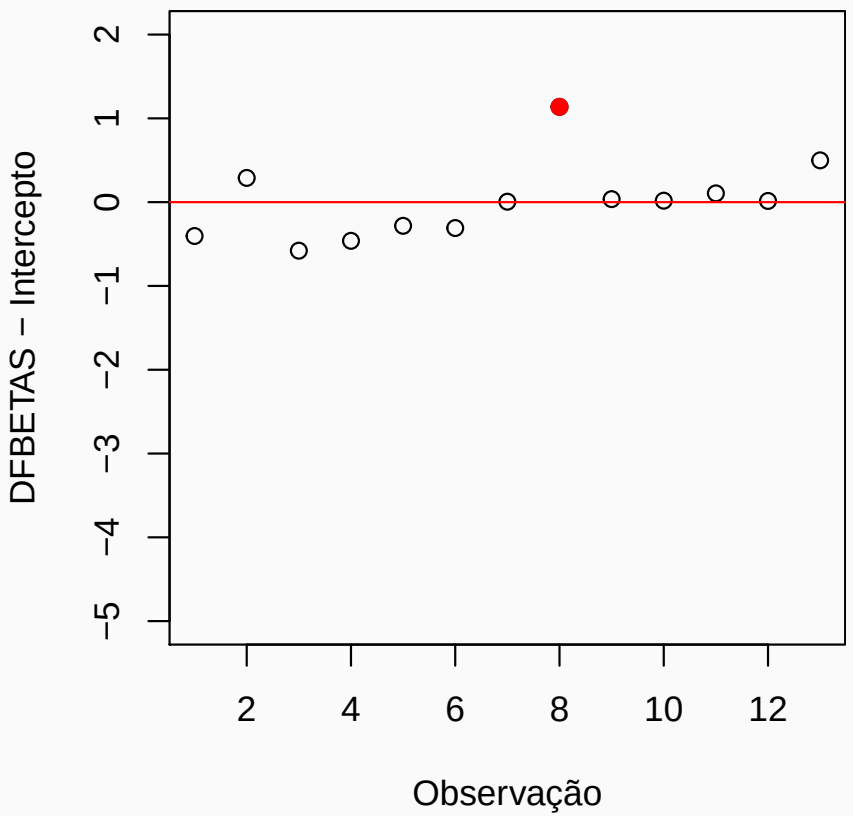

```
fits <- cooks.distance(m)
```



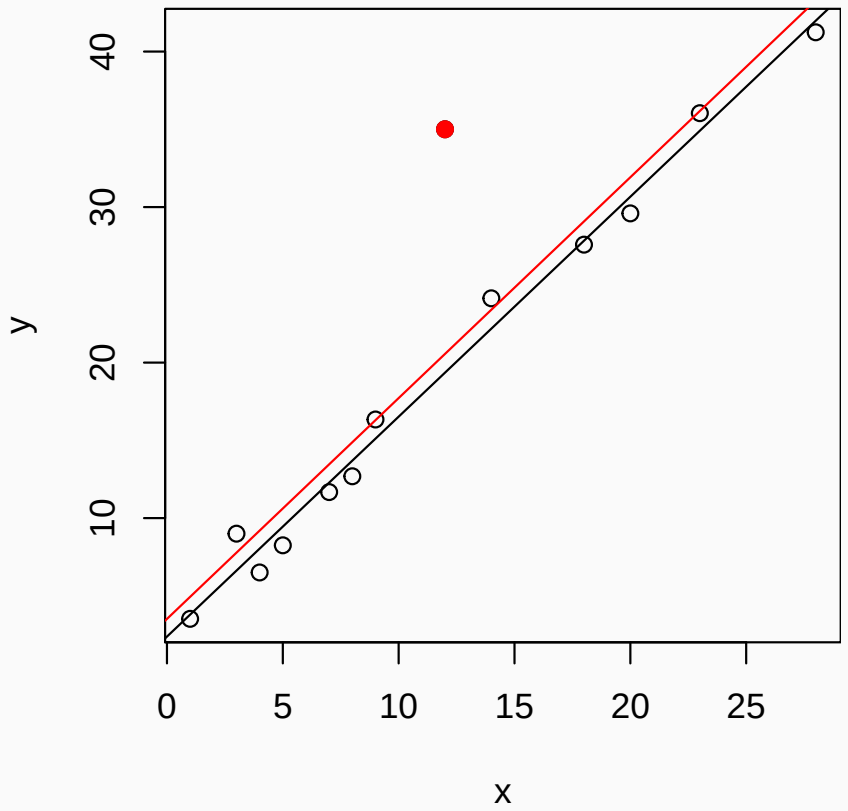
↓ LEVERAGE E ↑ DISTANCE



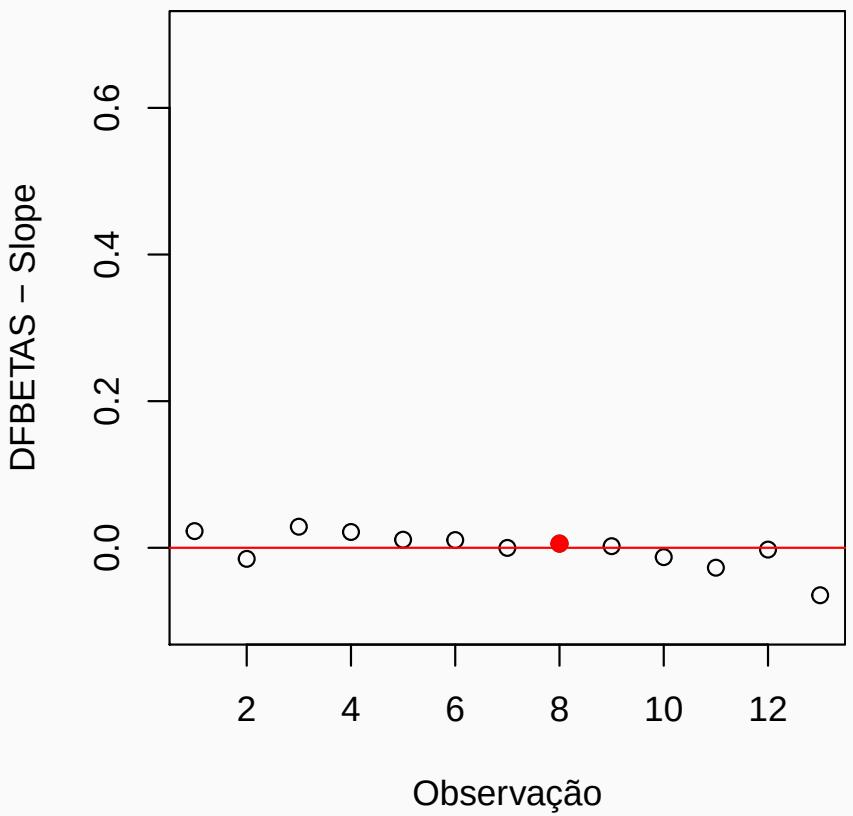

```
fits <- dfbeta(m)
```



↓ LEVERAGE E ↑ DISTANCE

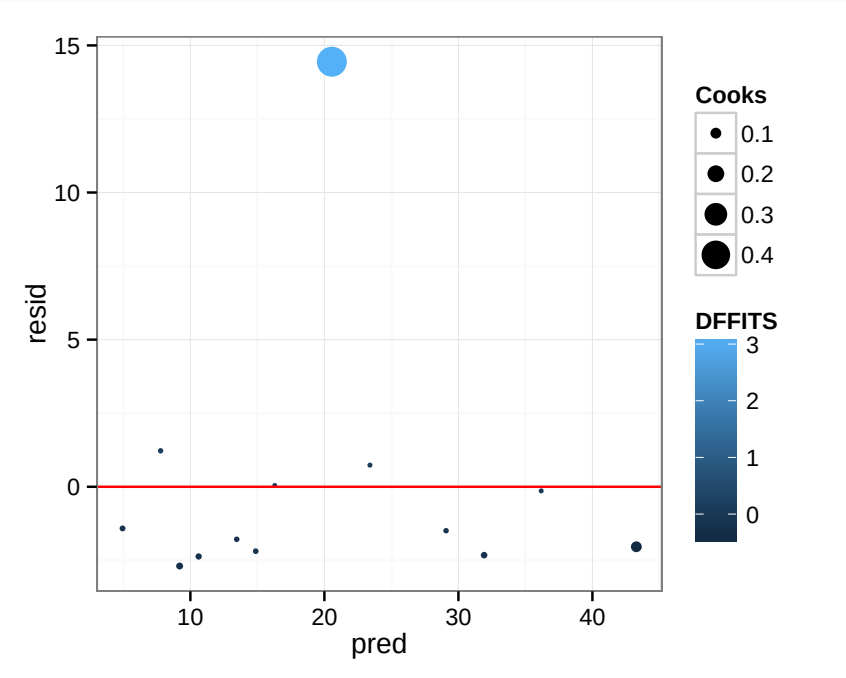


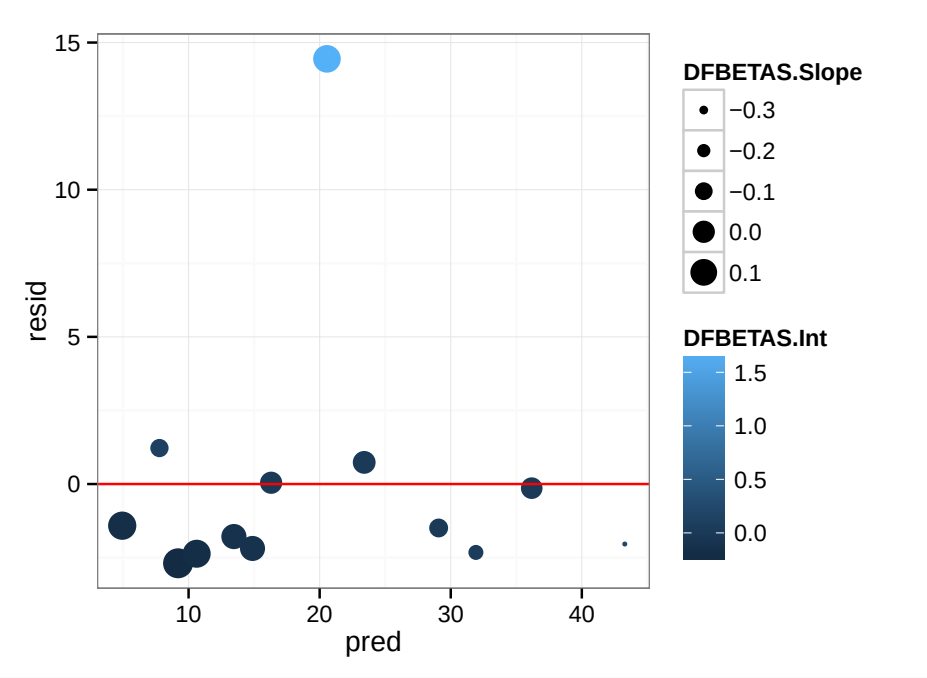

```
fits <- dfbeta(m)
```



↓ LEVERAGE E ↑ DISTANCE

Visualização + diagnóstico!





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

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

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

REMEDIAÇÃO

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.

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:

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

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

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

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

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

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

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

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

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

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

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

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

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

c é apenas uma constante de escala

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

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

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

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 $\lambda = 0$

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 $\lim_{\lambda \to 0} \frac{Y^\lambda - 1}{\lambda} = \log_e(X)$

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

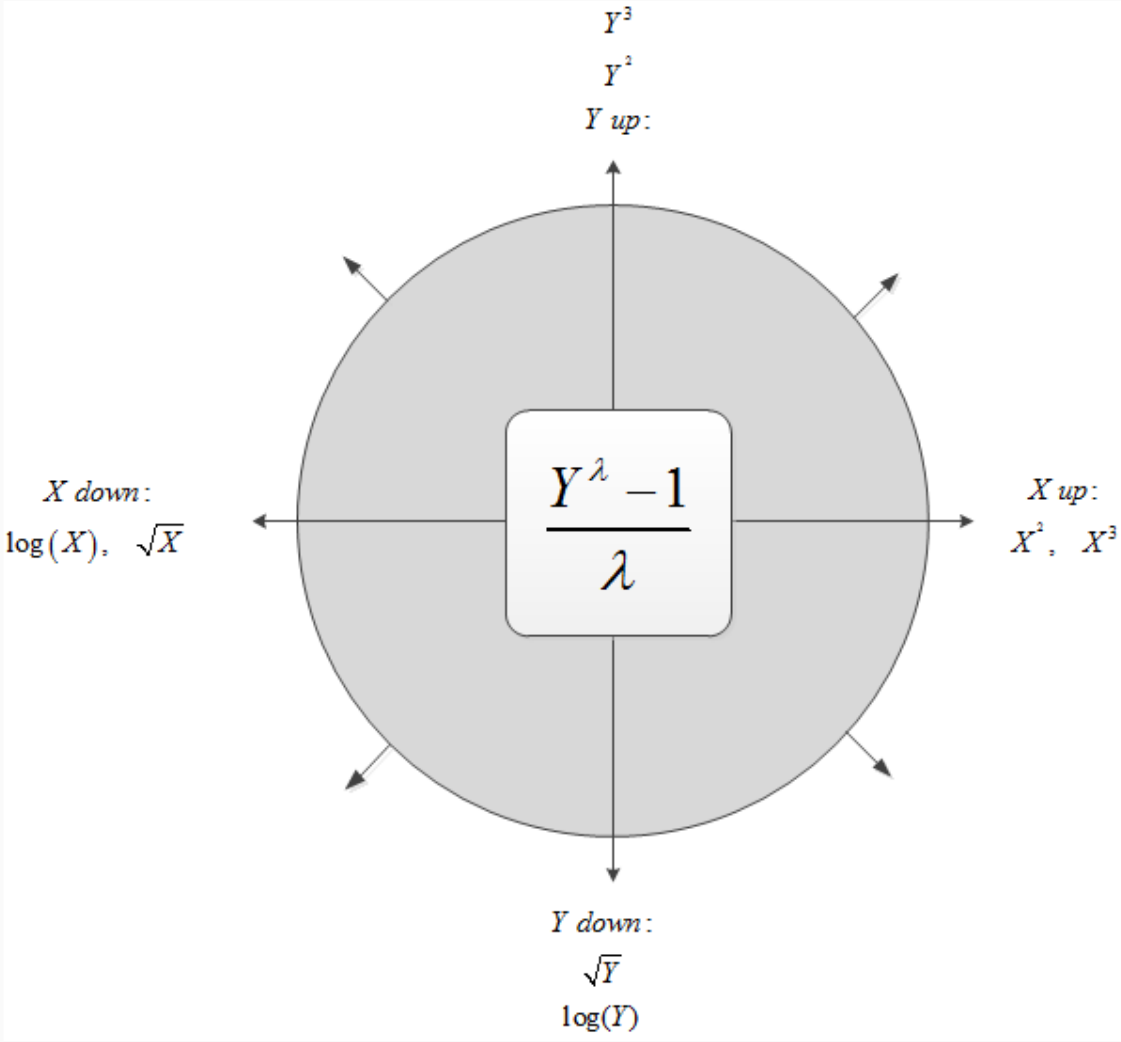
$Y' = \frac{Y^\lambda - 1}{\lambda}$, para $\lambda \neq 0$

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

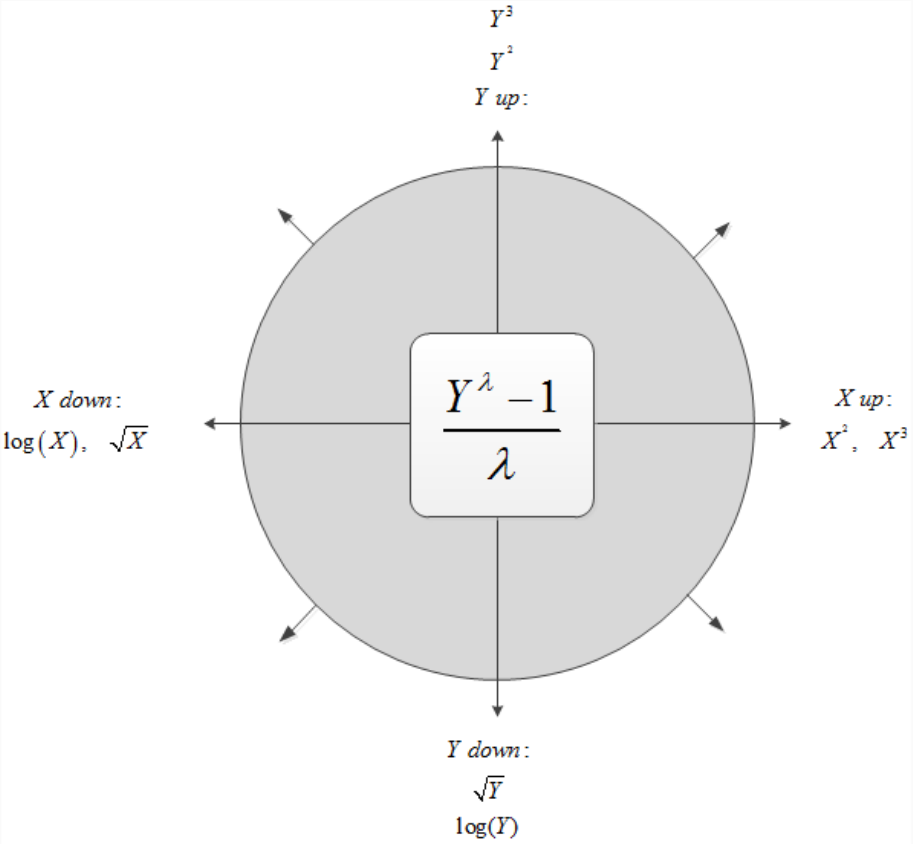
A expressão acima é válida pois $\lim_{\lambda \to 0} \frac{Y^\lambda - 1}{\lambda} = \log_e(X)$

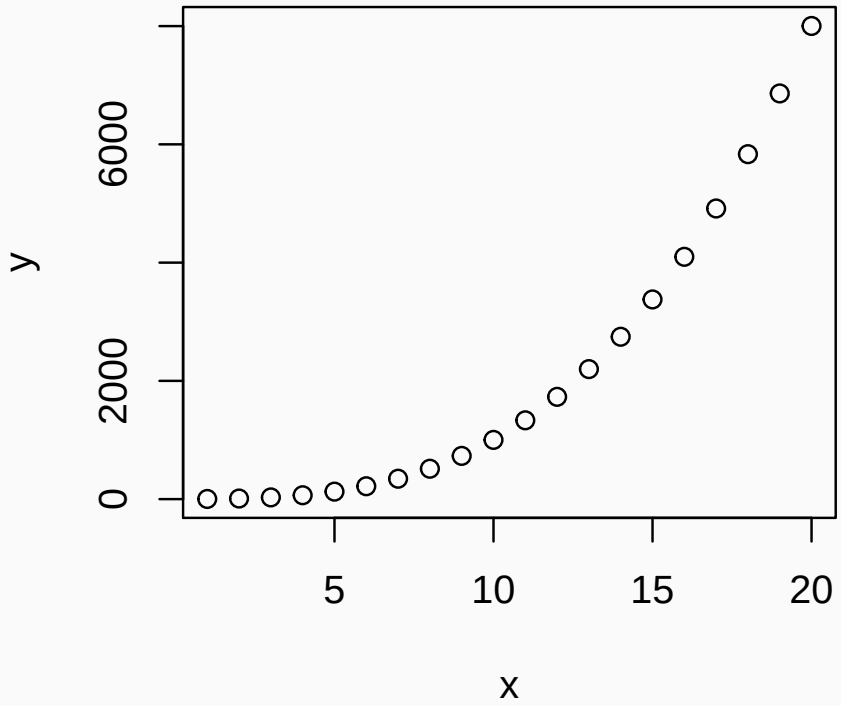
Normalmente se prefere $\log_{10}$ para facilitar a interpretação, pois um aumento de 1 em $\log_{10}(Y)$ é o mesmo que multiplicar Y por 10

REGRA DA CONVEXIDADE DE TUKEY



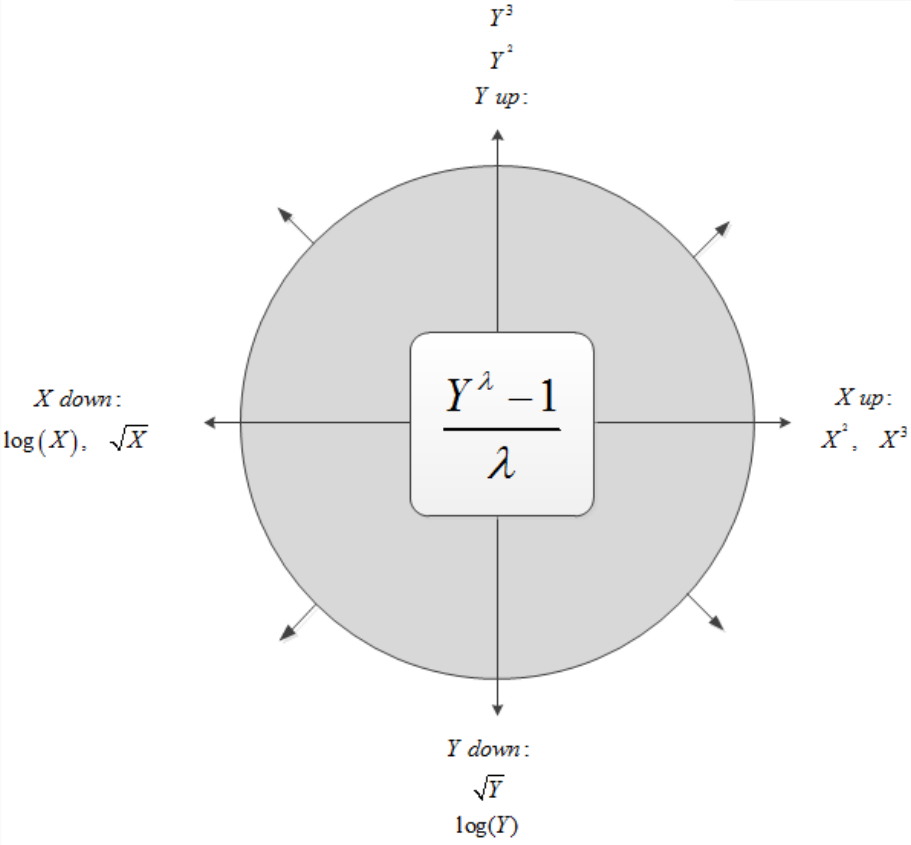
REGRA DA CONVEXIDADE DE TUKEY

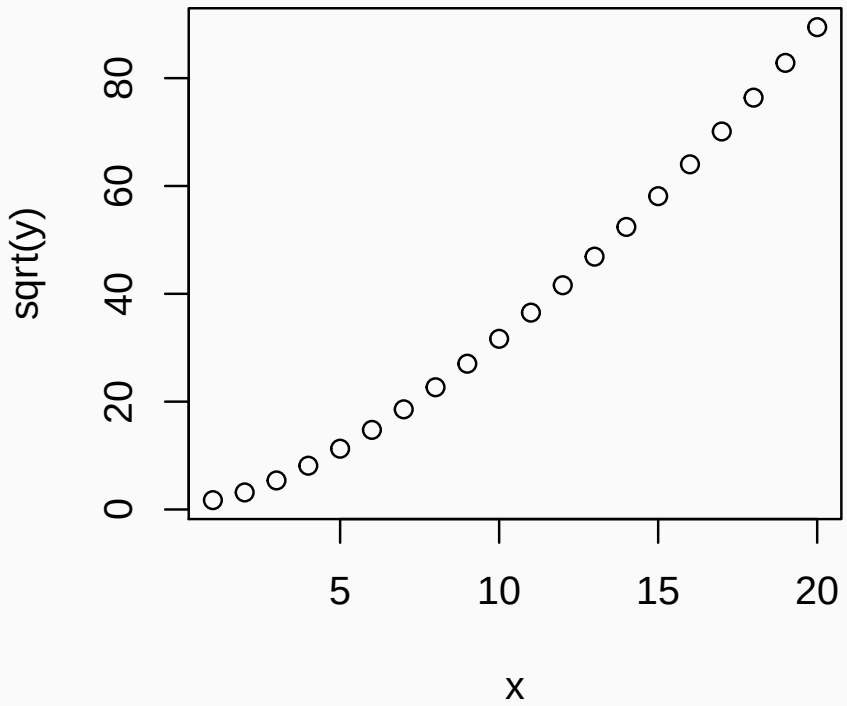




REGRA DA CONVEXIDADE DE TUKEY

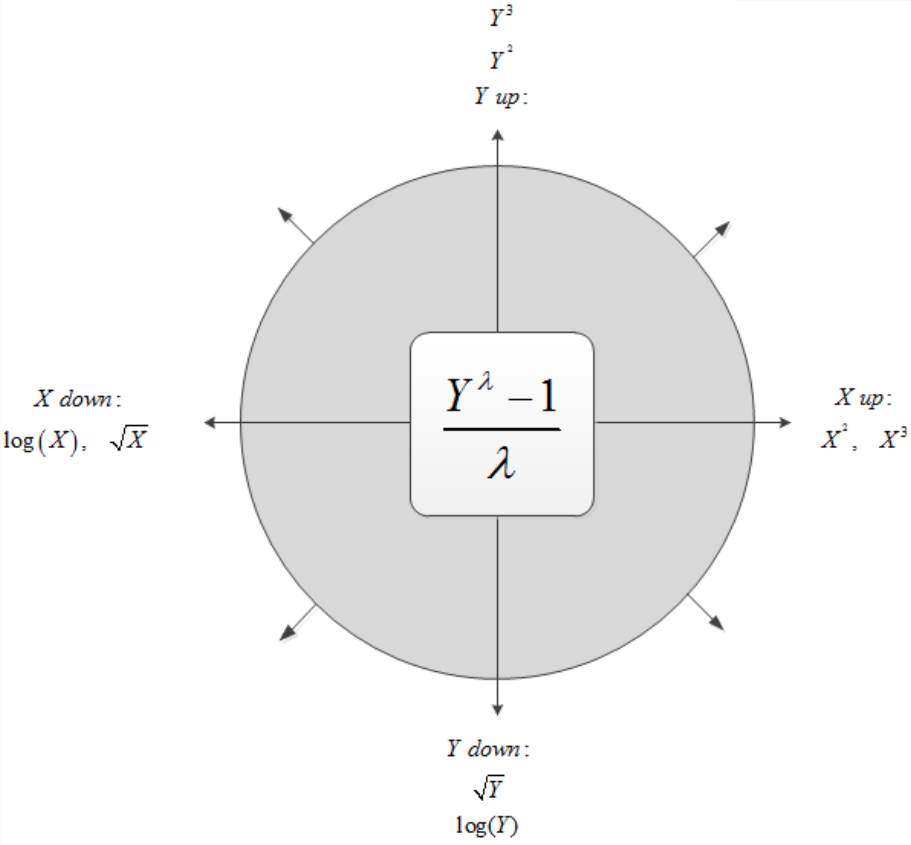
```
plot(x,sqrt(y))
```

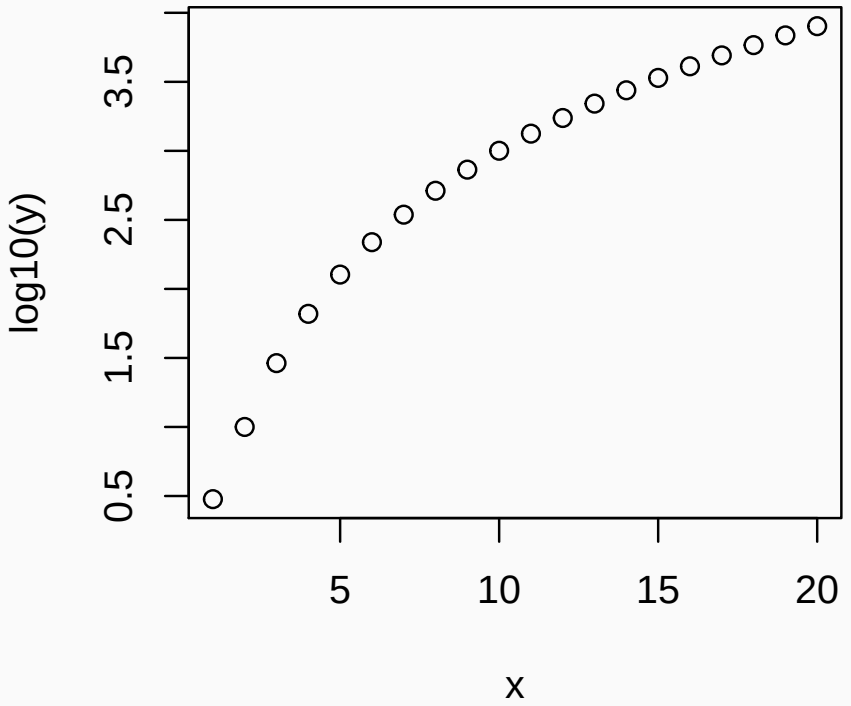




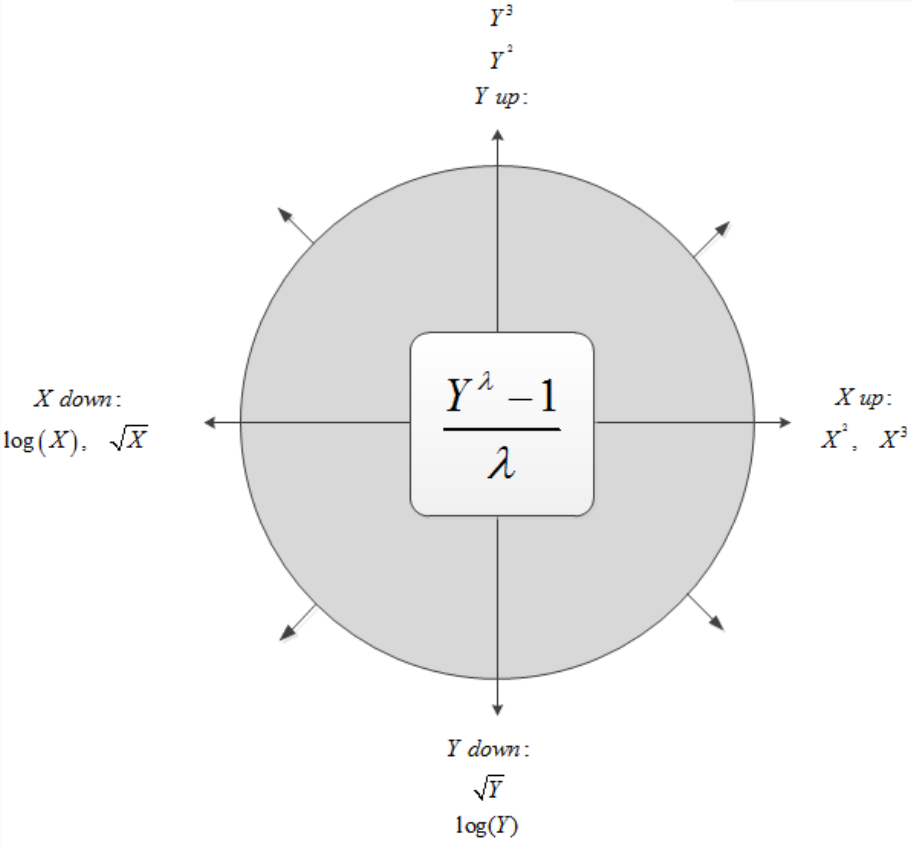
REGRA DA CONVEXIDADE DE TUKEY

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

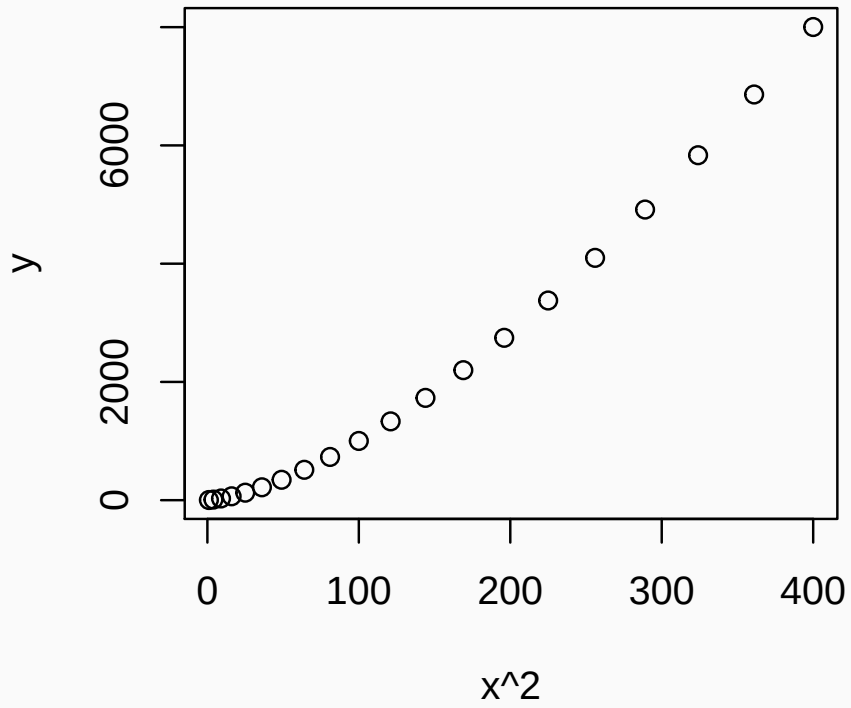




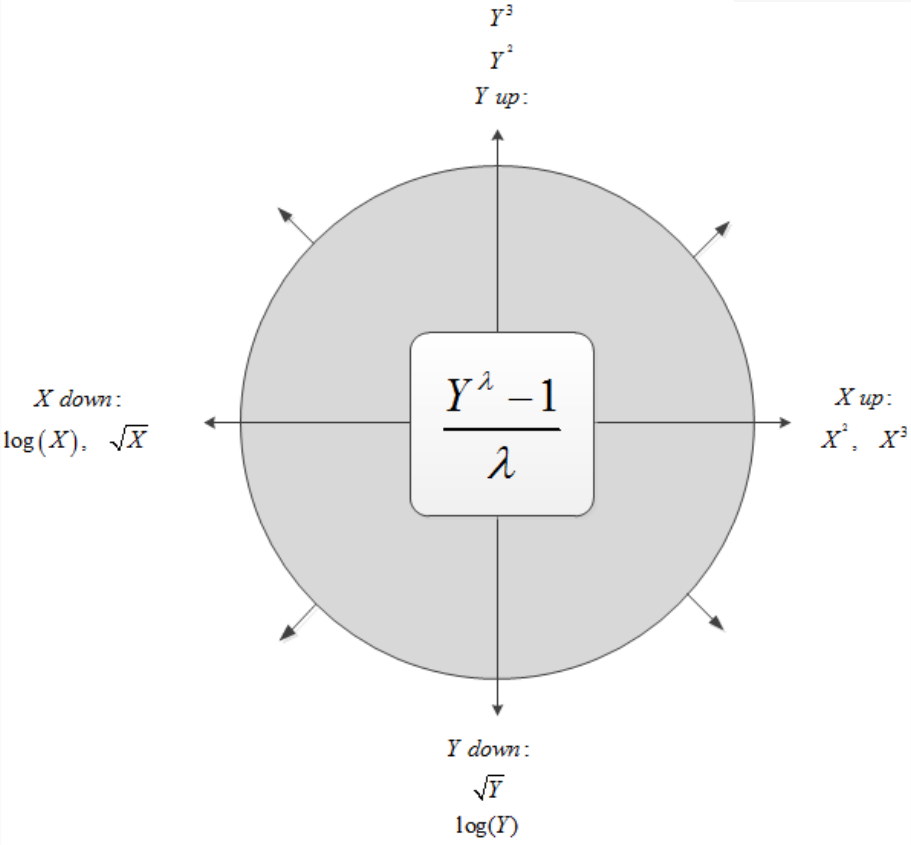
REGRA DA CONVEXIDADE DE TUKEY



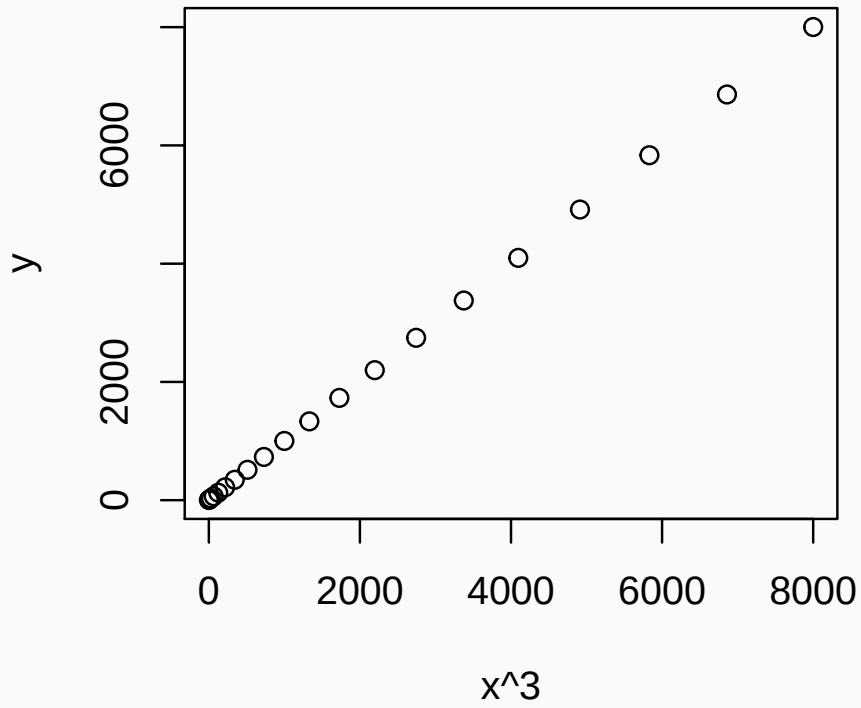

```
plot(x2, y)
```



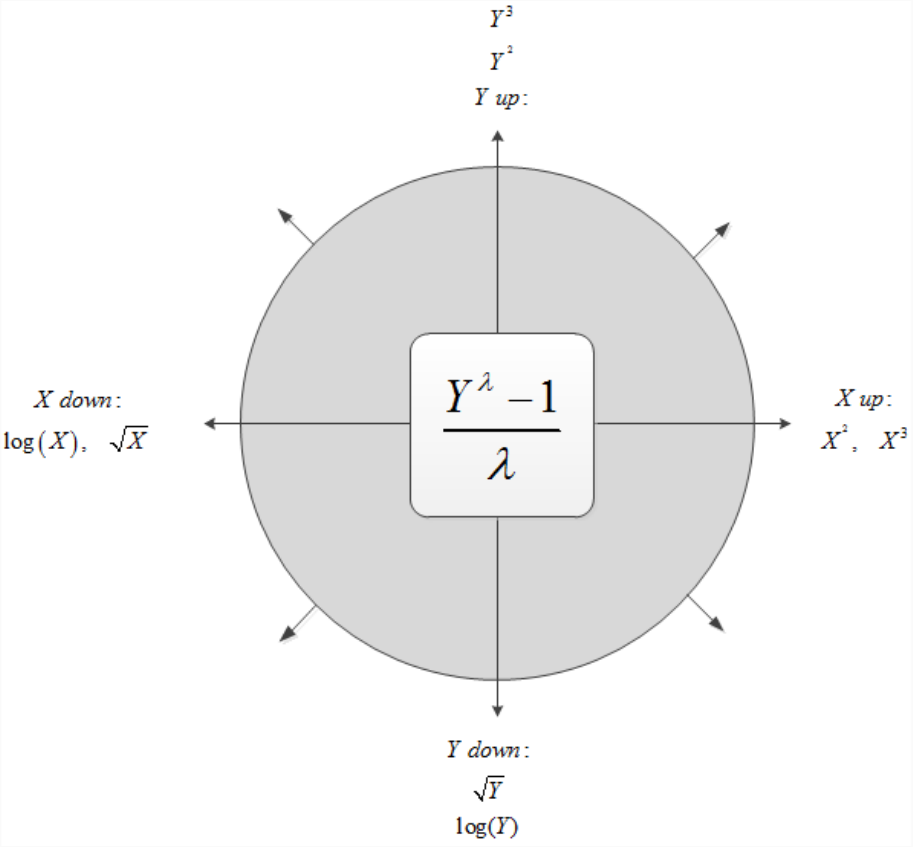
REGRA DA CONVEXIDADE DE TUKEY

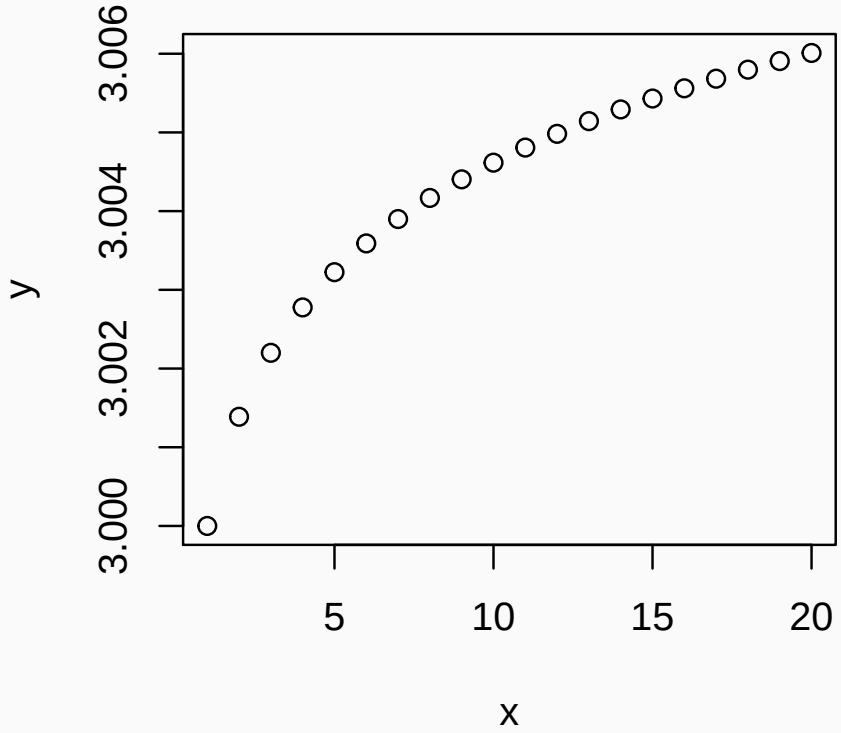



```
plot(x3, y)
```

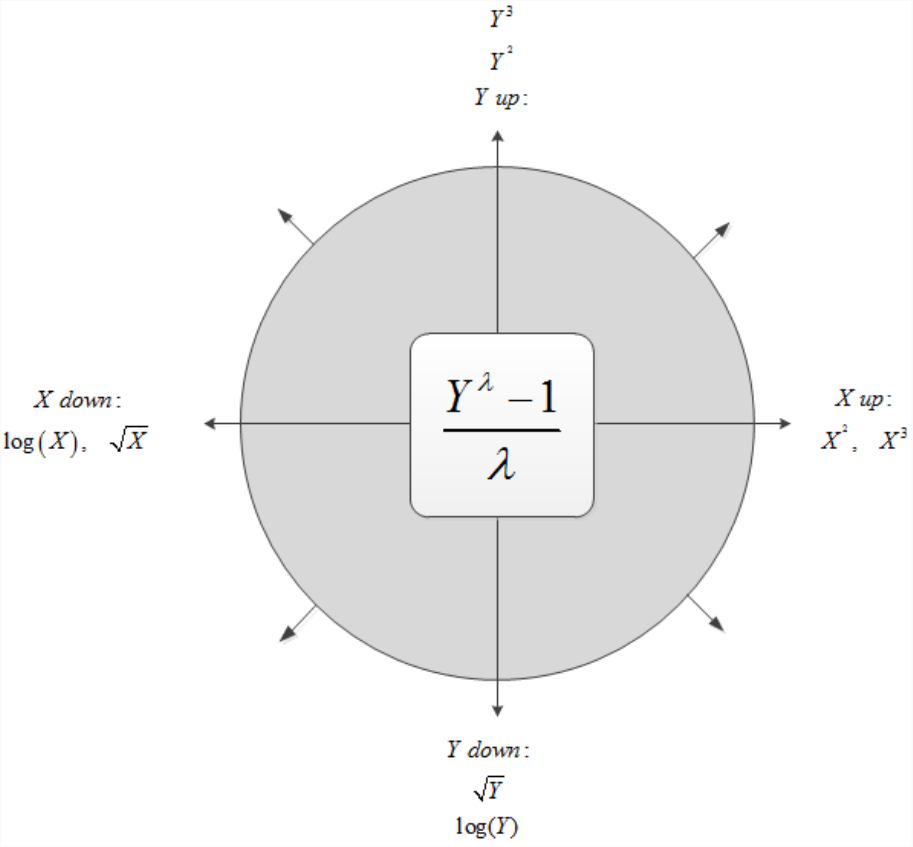


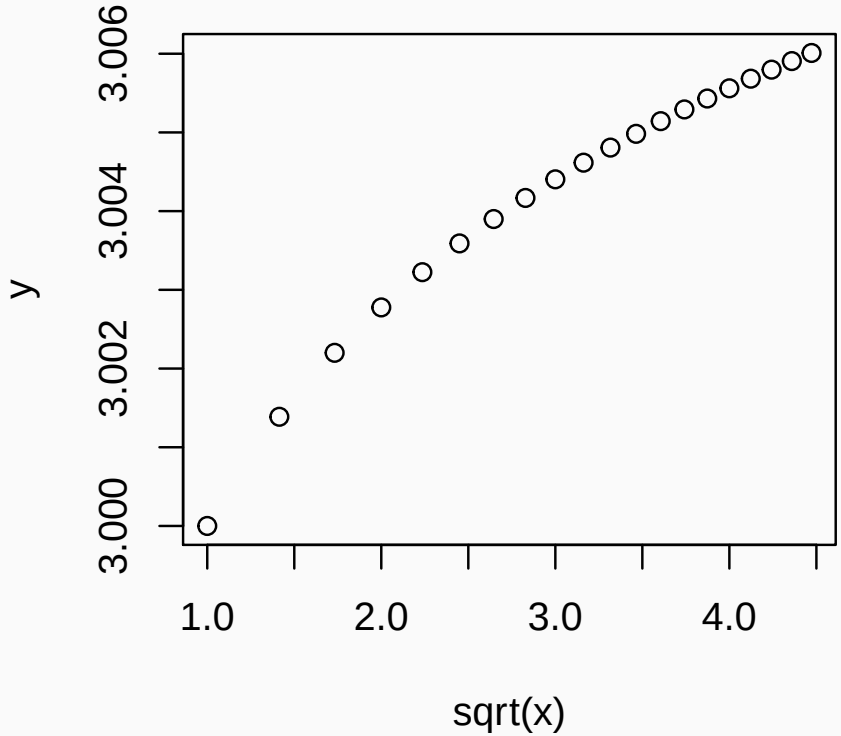
REGRA DA CONVEXIDADE DE TUKEY



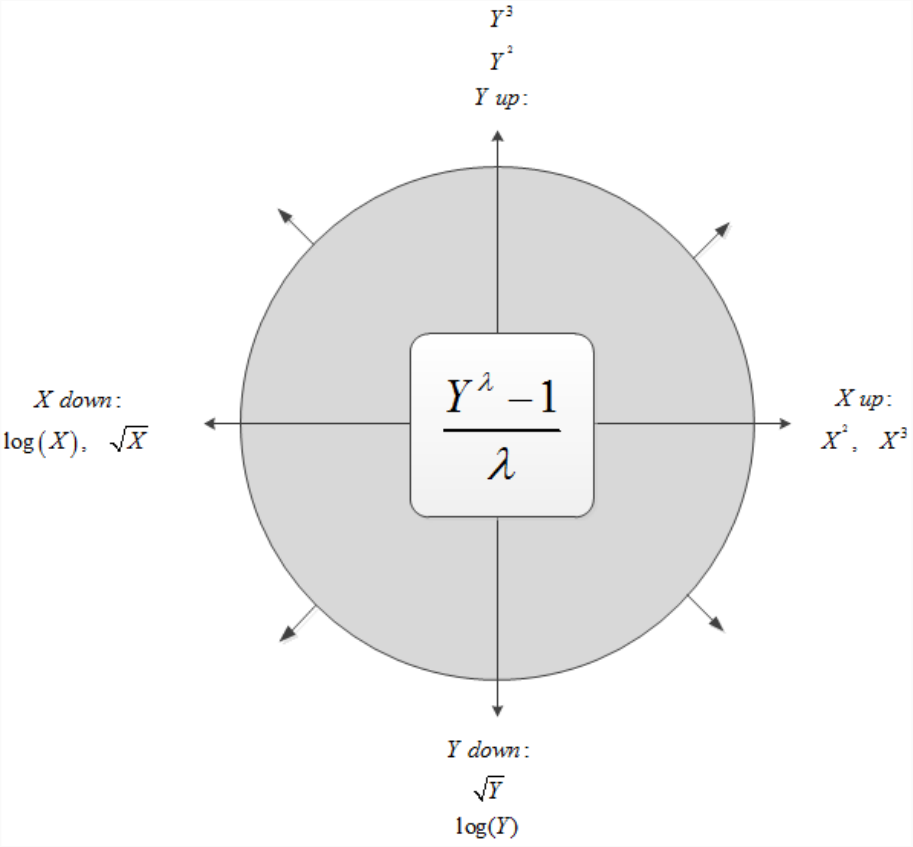


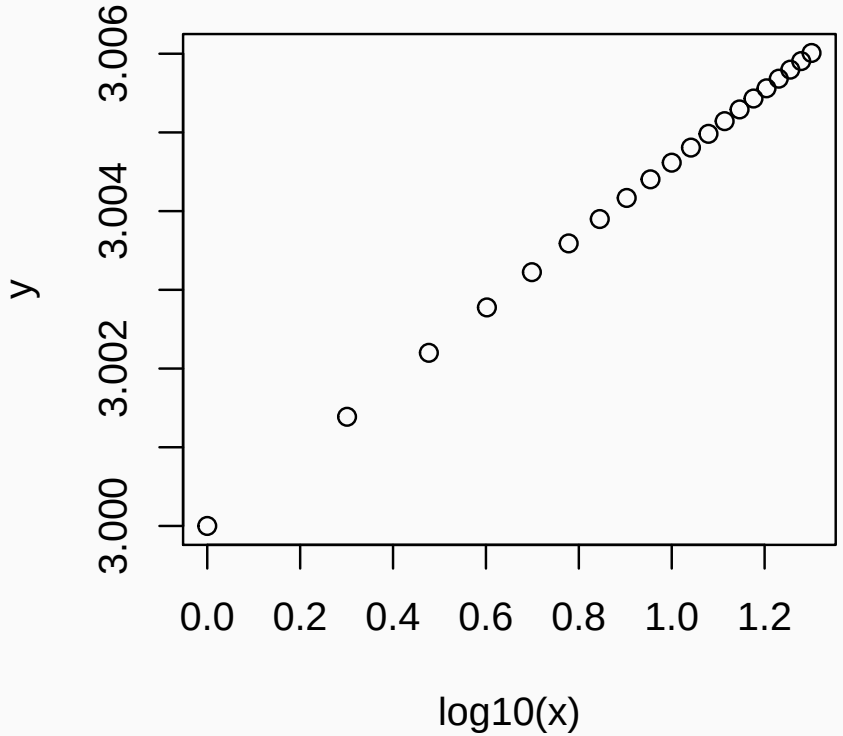
REGRA DA CONVEXIDADE DE TUKEY



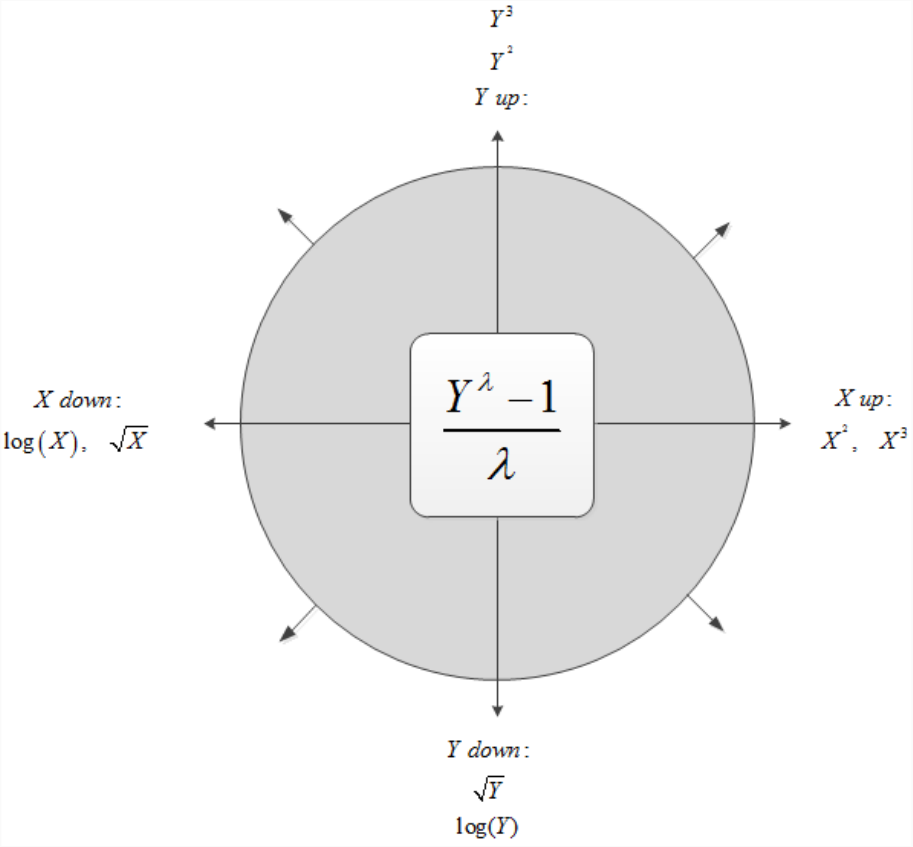


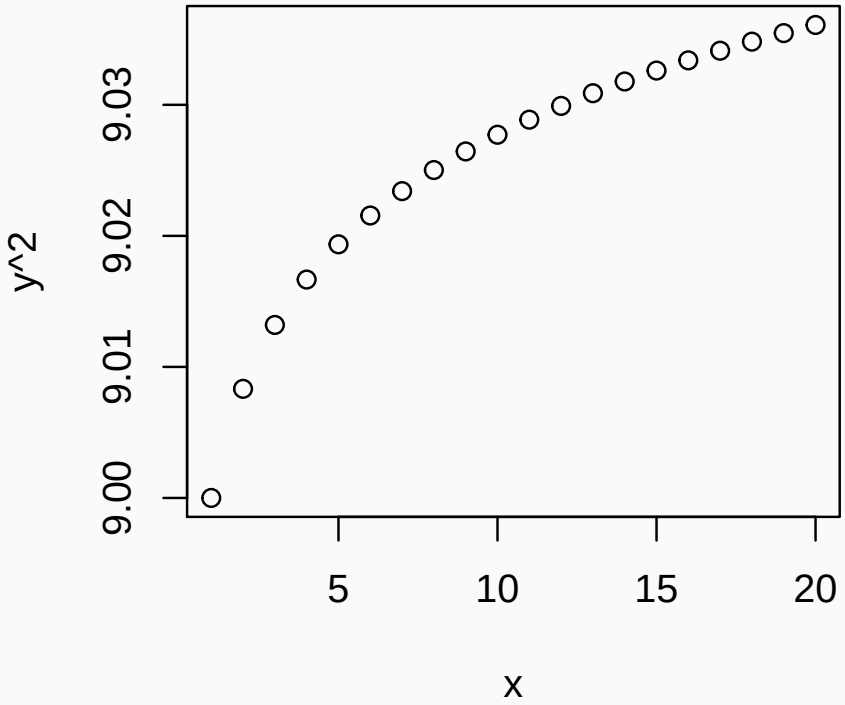
REGRA DA CONVEXIDADE DE TUKEY



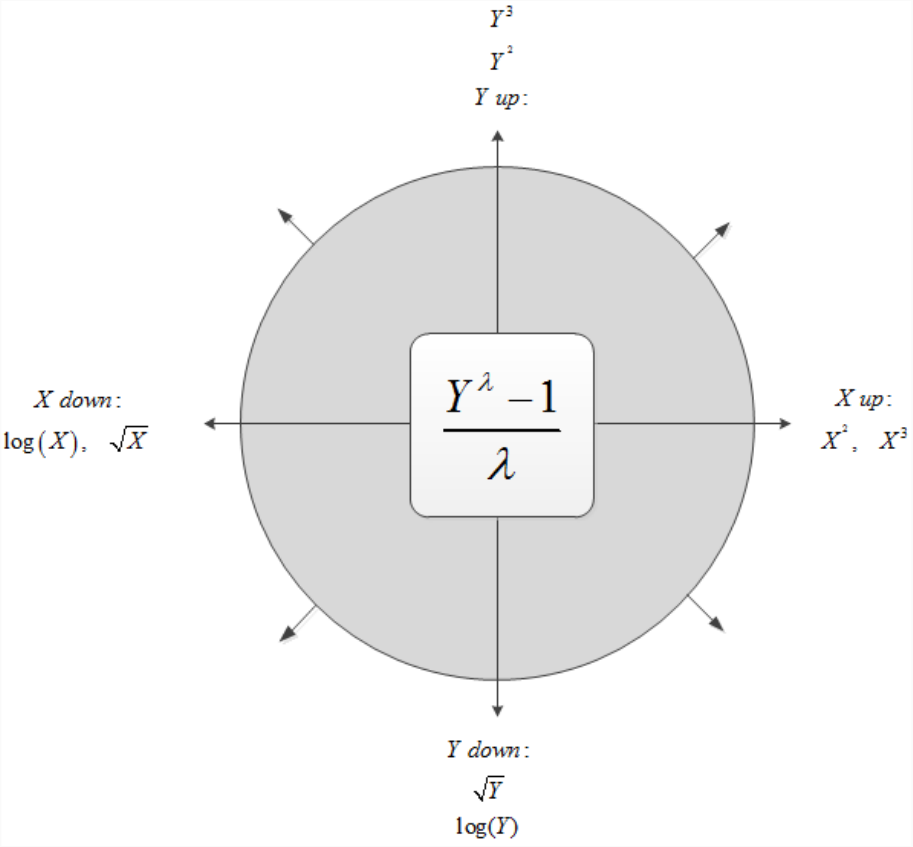


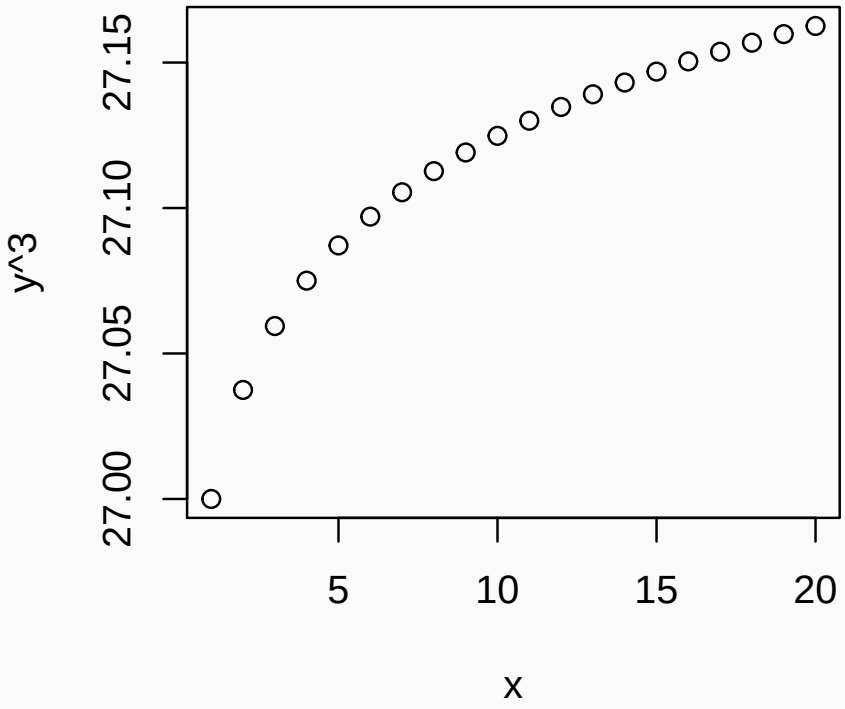
REGRA DA CONVEXIDADE DE TUKEY





REGRA DA CONVEXIDADE DE TUKEY

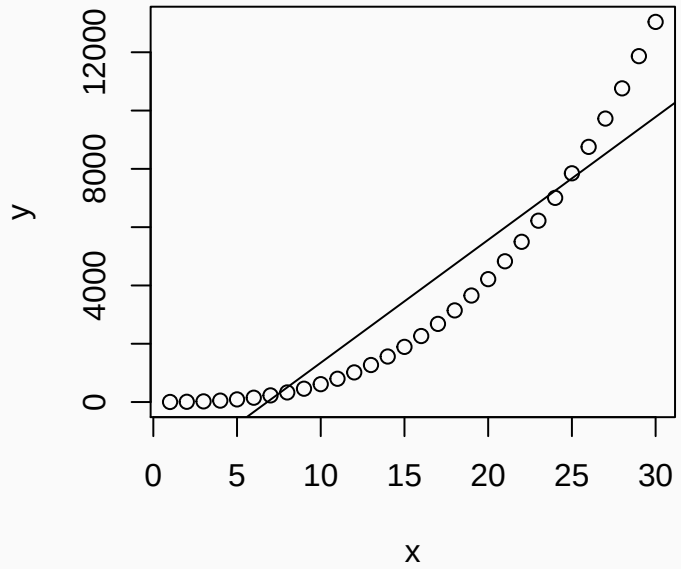




OTIMIZAÇÃO DE λ

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

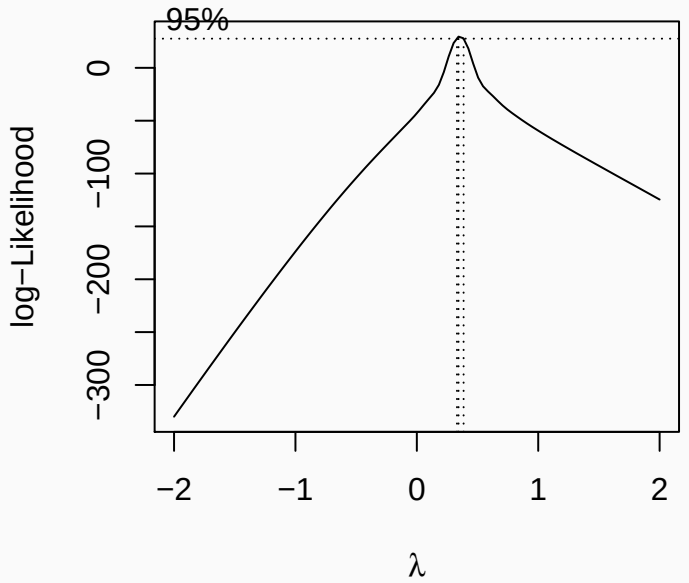
```
x <- c(1:30)
y <- 2 + x^2.786
m <-lm(y~x)
plot(x,y)
abline(m)
```



OTIMIZAÇÃO DE λ

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

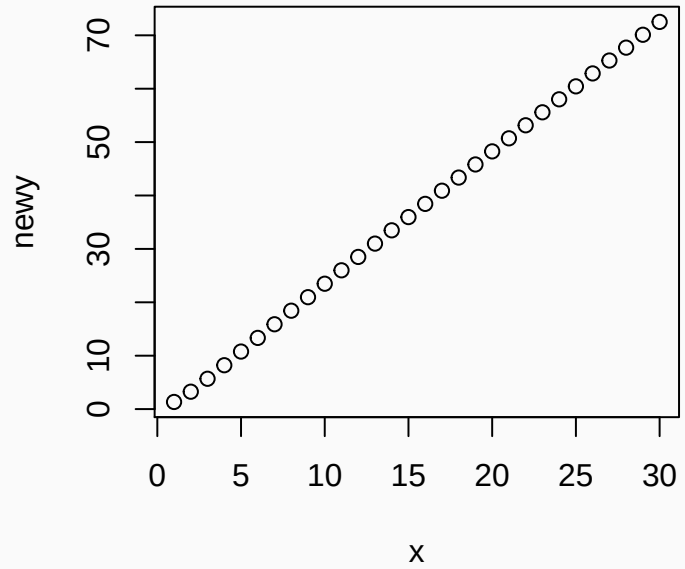
```
library(MASS)
lambda <- boxcox(m)
```



OTIMIZAÇÃO DE λ

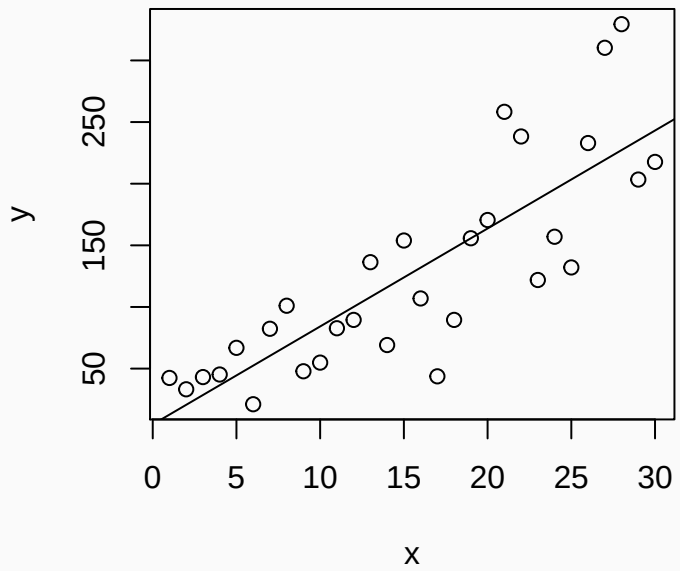
Podemos utilizar métodos computacionais para encontrar o melhor valor de λ (médodo de máxima verossimilhança, (maximum likelihood)

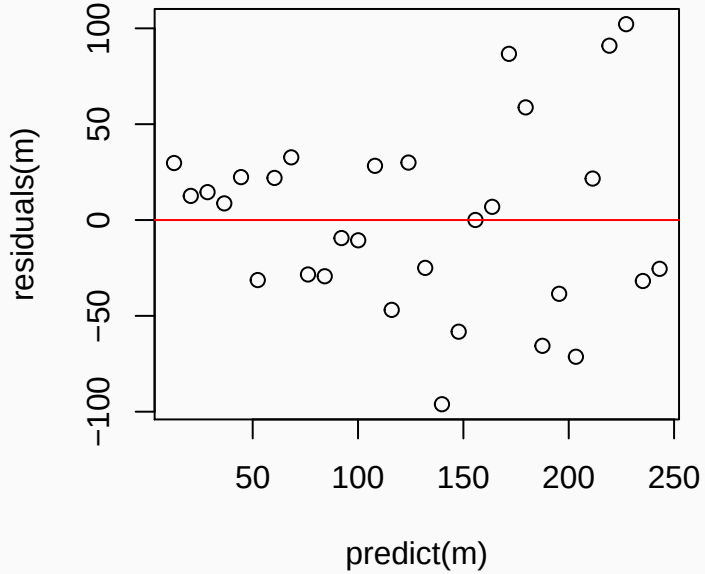
```
which(lambda$y == max(lambda$y))
## [1] 59
lambda$x[59]
## [1] 0.3434343
newy <- (y^0.3434343 -1)/0.3434343
plot(x,newy)
```



TRANSFORMAÇÕES: NORMALIDADE E VARIÂNCIA

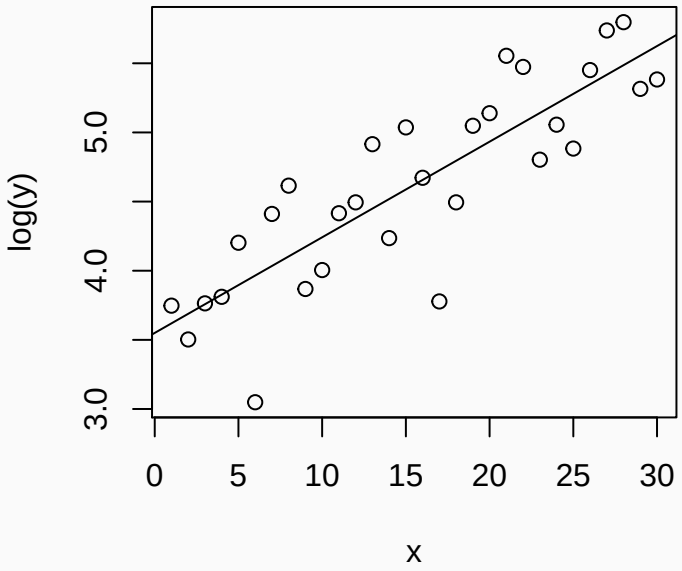
O uso de transformações não se limita à linearização de variáveis, mas também é de grande ajuda na aproximação dos dados para uma distribuição normal e variância constante

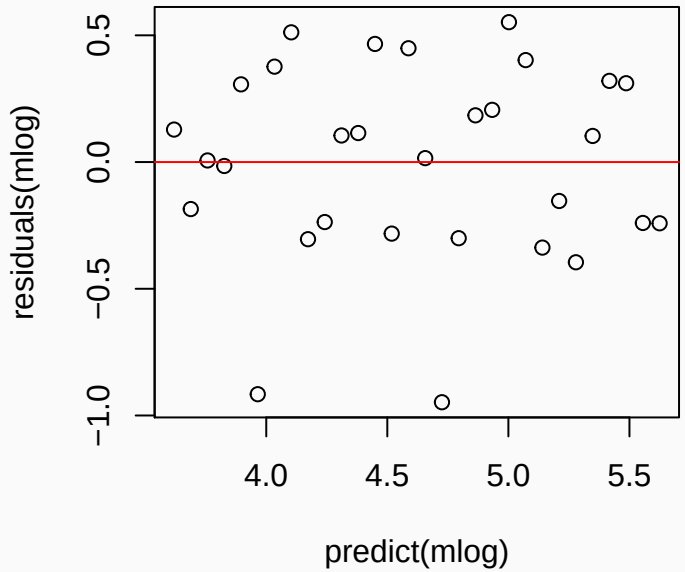




TRANSFORMAÇÕES: NORMALIDADE E VARIÂNCIA

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.

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

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

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

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 $\hat{Y}_{i(novo)}$?

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 paramétricos

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 paramétricos

2) Validação independente

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 paramétricos

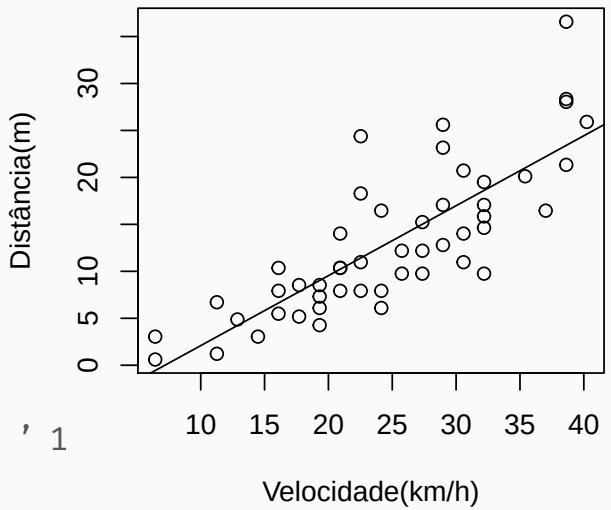
2) Validação independente

3) Métodos de aleatorização e reamostragem

Distância de frenagem pode ser predita pela velocidade do veículo?

```
summary(m0)
```

```
##
## Call:
## lm(formula = dist ~ speed, data = cars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.8603 -2.9033 -0.6925  2.8086 13.1678
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -5.3581     2.0600  -2.601   0.0123 *
## speed         0.7448     0.0787   9.464 1.49e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.688 on 48 degrees of freedom
## Multiple R-squared:  0.6511,^^IAdjusted R-squared:  0.6438
## F-statistic: 89.57 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

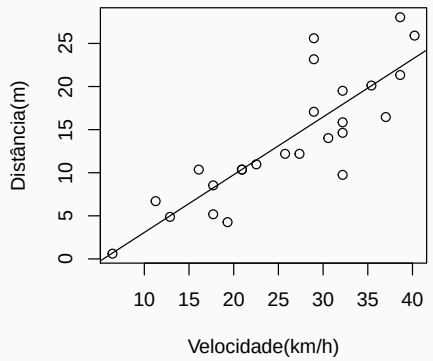
EXEMPLO

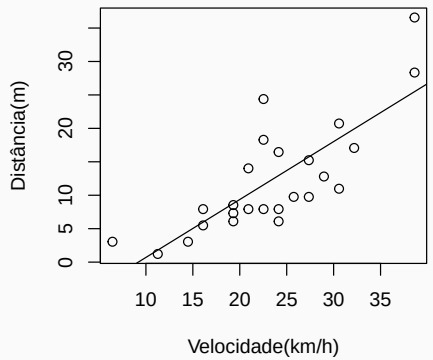
Distância de frenagem pode ser predita pela velocidade do veículo?

```
dim(cars)
```

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

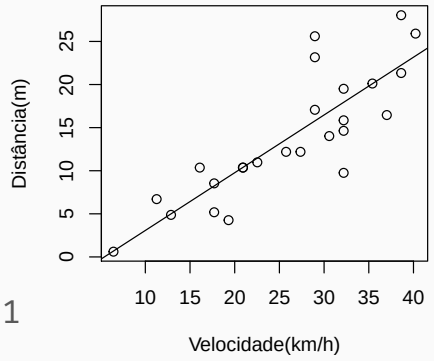
```
set.seed(89)
samp <- sample(1:50,25,rep=F)
cars1 <- cars[samp,]
cars2 <- cars[-samp,]
m1 <- lm(dist ~ speed, cars1)
m2 <- lm(dist ~ speed, cars2)
```

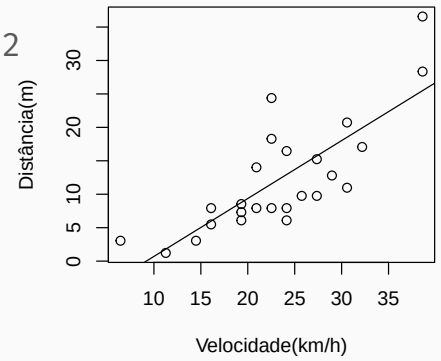




Distância de frenagem pode ser predita pela velocidade do veículo?

```
summary(m1)$coefficients
##             Estimate Std. Error   t value    Pr(>|t|)
## (Intercept) -3.6299189 2.45699491 -1.477382 1.531367e-01
## speed        0.6699674 0.08863644  7.558600 1.119803e-07
summary(m2)$coefficients
##             Estimate Std. Error   t value    Pr(>|t|)
## (Intercept) -7.9600734  3.4865900 -2.283054 3.199371e-02
## speed        0.8660454  0.1421012  6.094569 3.235797e-06
```





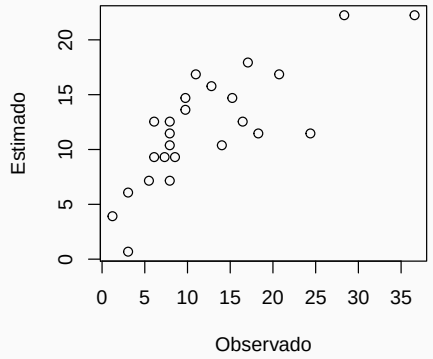
Distância de frenagem pode ser predita pela velocidade do veículo?

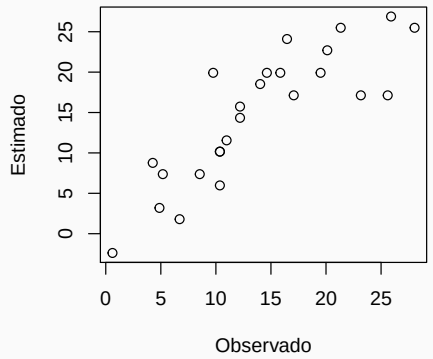
```
pr1 <- predict(m1,cars2)
pr2 <-predict(m2,cars1)
rmse1 <- sqrt(mean((cars2$dist - pr1)^2))
rmse2 <- sqrt(mean((cars1$dist - pr2)^2))
rmse1
```

```
## [1] 5.311186
```

```
rmse2
```

```
## [1] 4.313988
```





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

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?

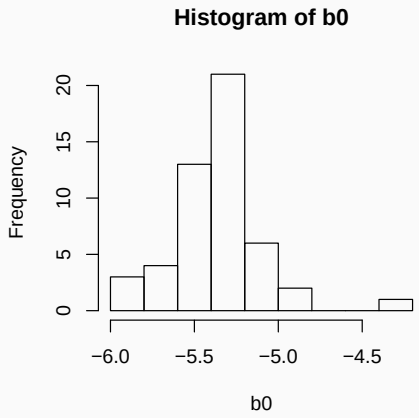
Não seria ótimo se pudéssemos usar o máximo possível de observações pra estimar o erro?

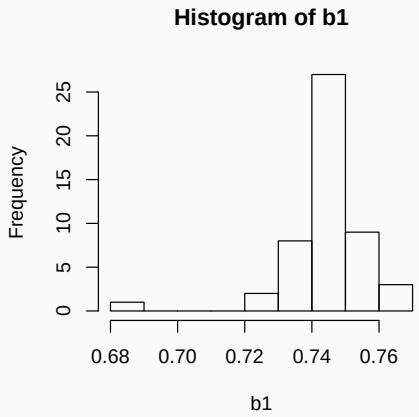
E se, ao invés de dividir meio a meio, deixássemos uma observação de fora, e repetíssemos n vezes?

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

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)
}
```

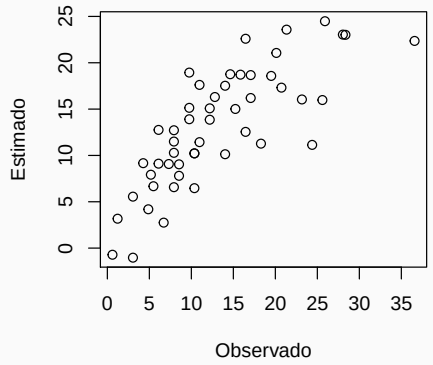


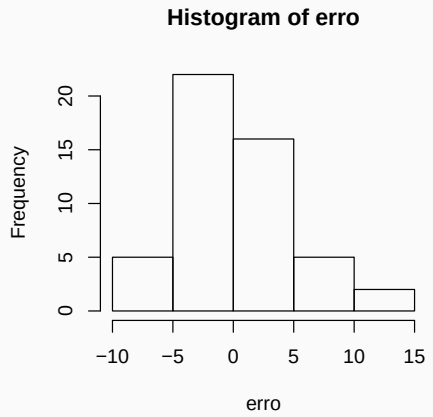


Distância de frenagem pode ser predita pela velocidade do veículo?

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

```
## [1] 4.784539
```





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

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 $1 - \alpha/2$

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

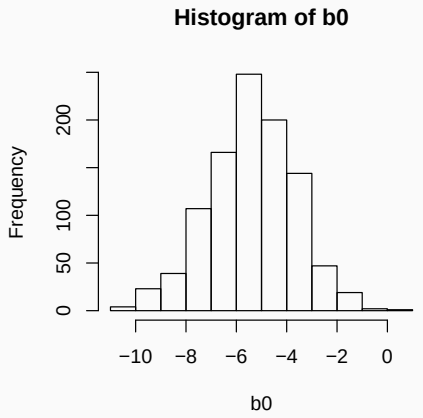
Selecione n novas amostras, com reposição, e recalcule o modelo. Repita **muitas** vezes.

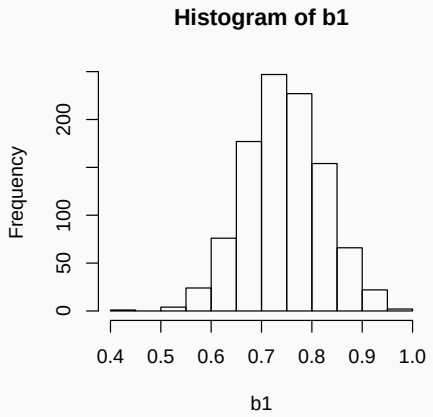
Observações mais frequentes vão ser reamostradas mais vezes

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

Distância de frenagem pode ser predita pela velocidade do veículo?

```r
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)
}
```





EXEMPLO

```r
quantile(b0,probs=c(0.025,0.975))
```

```
##      2.5%     97.5%
## -9.022321 -2.187944
```

```r
quantile(b1,probs=c(0.025,0.975))
```

```
##      2.5%     97.5%
## 0.5945032 0.8997511
```

```r
confint(m0)
```

```
##                  2.5 %     97.5 %
## (Intercept) -9.4999606 -1.2162557
## speed        0.5865477  0.9030047
```

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

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

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!

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!