R para Ciência de Dados

Regressão Linear Simples e Múltipla

Profa Carolina Paraíba e Prof Gilberto Sassi

Departamento de Estatística
Instituto de Matemática e Estatística
Universidade Federal da Bahia

Novembro de 2026

Curso R para Ciência de Dados: Regressão Linear Simples e Múltipla

Regressão Linear Simples e Múltipla

- Introdução à Modelagem Estatística
- Regressão Linear Simples
- Modelo de Regressão Linear Simples
- Estimação
- Teste de Hipóteses (Teste de Significância do Modelo)
- Estimação de σ^2
- Intervalos de Confiança
- Análise de Resíduos e Diagnóstico
- Exemplos

Introdução à Modelagem Estatística

Problema de Análise de Regressão: estabelecer e determinar uma função que descreva a relação entre uma variável, chamada de variável resposta e denotada por Y, e um conjunto de variáveis observáveis, chamadas de variáveis preditoras, explicativas ou covariáveis e denotadas por $X_1, X_2, \ldots, X_p$.

Uma vez estabelecida e determinada a relação funcional entre a variável resposta e as covariáveis, a Análise de Regressão pode explorar esta relação para obter informações sobre Y a partir do conhecimento de $X_1, X_2, \ldots, X_p$. Os modelos de regressão podem, então, serem usados para predição, estimação, testes de hipótese e para modelar relações casuais.

Exemplo 1. Usar a informações sobre a altura, o sexo e a raça (e muitas outras mais) de uma pessoa para prever o peso da pessoa. No cenário mais simples, estudamos como uma variável específica de interesse é afetada por uma única variável. Por exemplo, usar a altura de uma pessoa para prever o seu peso. ▪

Para dados experimentais ou de pesquisa, os *modelos matemáticos*, onde as relações entre entradas e saídas são puramente determinística, não são adequados. Isso porque, quando a relação é determinística, a equação do modelo descreve exatamente a relação das variáveis. Assim, para dados experimentais ou de pesquisa, estamos interessados em estabelecer *modelos estatísticos*, nos quais a relação entre as variáveis não é perfeita. Por exemplo, se considerarmos as variáveis altura e peso, então a medida que a altura aumenta, espera-se que o peso também aumente, mas não perfeitamente.

Regressão Linear Simples

Problema de regressão linear simples: estabelecer e determinar uma função linear que descreva a relação entre uma variável, chamada de *variável resposta* ou dependente e denotada por Y, e uma variável chamada de *variável preditora* ou explicativa e denotada por X.

Quando consideramos um modelo de regressão linear simples, queremos ver o que acontece com a variável resposta quando o valor da variável preditora se modifica. Se o valor da variável preditora aumentar, a resposta tende a aumentar, diminuir ou permanecer constante?

Regressão Linear Simples

Por exemplo, podemos ter interesse em investigar a relação (linear?) entre: alturas e pesos; médias de notas do ensino médio e médias de notas da faculdade; velocidade e quilometragem; temperatura exterior e taxa de evaporação.

Lembre-se que a equação de uma reta tem a seguinte forma:

$$y = a \cdot x + b,$$

onde a é a inclinação da reta e b é o intercepto com o eixo y.

Dados dois pontos numa reta, (x_1, y_1) e (x_2, y_2), a inclinação é calculada por

$$a = \frac{y_2 - y_1}{x_2 - x_1} = \frac{\text{mudança em } y}{\text{mudança em } x}.$$

Regressão Linear Simples

A inclinação de uma reta nos diz muito sobre a relação linear entre duas variáveis.

Se a inclinação é positiva, então há uma relação linear diretamente proporcional, isto é, à medida que uma aumenta, a outra aumenta.

Se a inclinação é negativa, então há uma relação linear inversamente proporcional, ou seja, à medida que uma aumenta a outra variável diminui.

Se a inclinação é zero, à medida que uma aumenta (ou diminui), a outra permanece constante.

Regressão Linear Simples

Em modelagem estatística, quando procuramos relações lineares entre duas variáveis, é muito pouco provável que as coordenadas forneçam exatamente uma linha reta: haverá algum erro. E está tudo bem!

Porém, antes de pensar em utilizar um modelo de regressão linear simples para representar e explicar a relação entre duas variáveis, devemos examinar os dados para verificar se faz sentido considerar uma relação linear entre essas duas variáveis. Fazemos isso com um gráfico de dispersão e o coeficiente de correlação linear.

Gráfico de dispersão:

Um *gráfico de dispersão* é uma representação gráfica de duas variáveis quantitativas onde a variável explicativa está no eixo x e a variável resposta está no eixo y e cada par de valores (x, y) é representado por um ponto. Ao analisar um gráfico de dispersão, buscamos responder as seguintes questões:

- Qual é a direção da relação?
- A relação é linear ou não linear?
- A relação é fraca, moderada ou forte?
- Existem valores atípicos ou extremos?

Descrevemos a direção da relação entre as variáveis como *positiva* ou *negativa*. Uma relação positiva indica que à medida que o valor da variável explicativa aumenta, o valor da variável resposta tende a aumentar. Uma relação negativa indica que à medida que o valor da variável explicativa aumenta, o valor da variável resposta tende a diminuir.

Leitura de arquivos no formato `xlsx` ou `xls`:

- **Pacote:** `readxl` do `tidyverse` (instale com o comando `install.packages('readxl')`)
- Parêmetros das funções `read_xls` (para ler arquivos `.xls`) e `read_xlsx` (para ler arquivos `.xlsx`):
  - `path`: caminho até o arquivo.
  - `sheet`: especifica a planilha do arquivo que será lida.
  - `range`: especifica uma área de uma planilha para leitura. Por exemplo: `B3:E15`.
  - `col_names`: Argumento lógico com valor padrão igual a `TRUE`. Indica se a primeira linha tem o nome das variáveis.
- Para mais detalhes, consulte a documentação oficial do *tidyverse*: documentação de `read_xl`.

Leitura de arquivos no formato `xlsx` ou `xls`:

arvores: este conjunto de dados apresenta medições de diâmetro (em polegadas), altura (em pés) e volume (em pés cúbicos) de madeira de 31 árvores da espécie cerejeira-negra que foram derrubadas. As medições foram realizadas a uma altura de 4 pés e 6 polegadas (aprox. 1,37 m) do solo.

```
library(readxl)
library(tidyverse)

dados <- read_xlsx("../dados/arvores.xlsx")
```

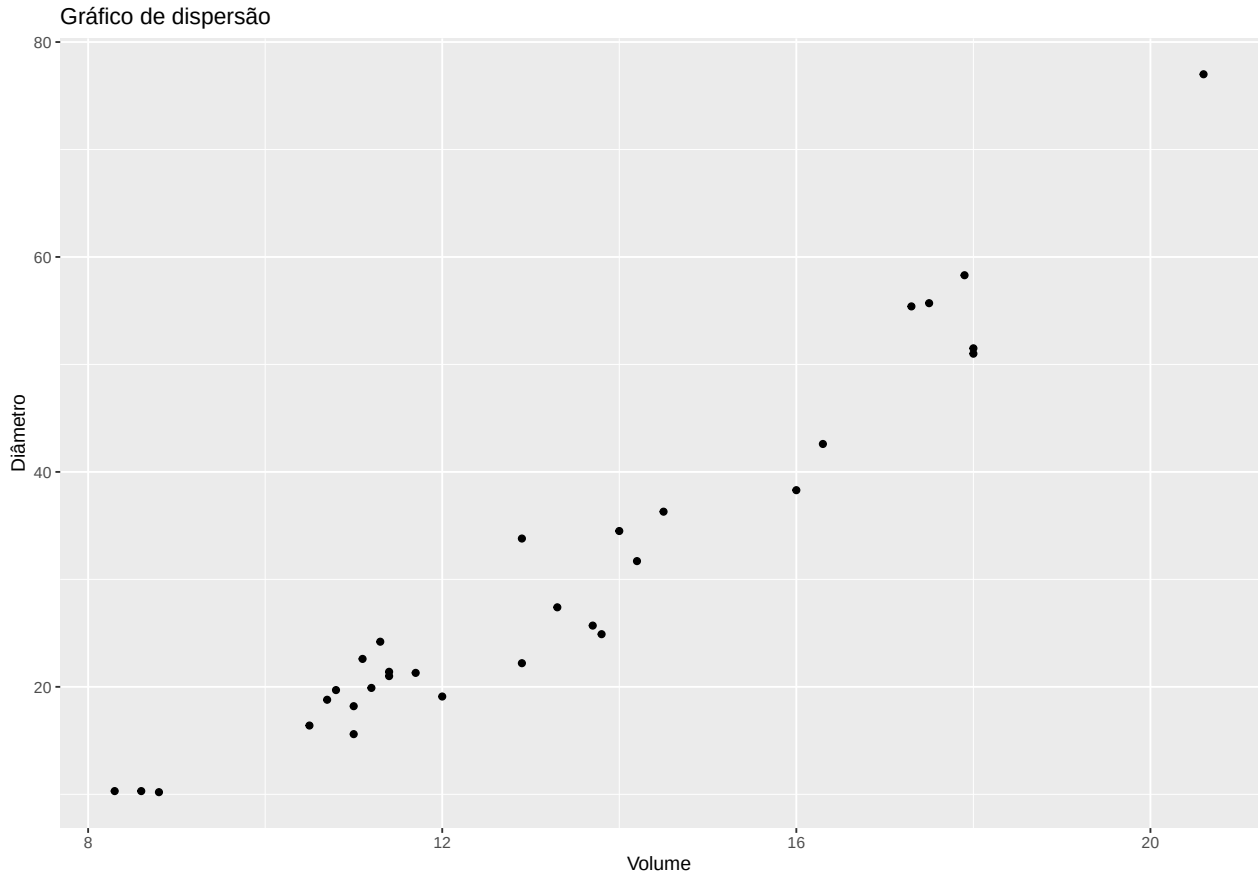
Gráfico de dispersão:

```
library(readxl)
library(tidyverse)

ggplot(data = dados) +
  geom_point(aes(x = diametro, y = volume)) +
  labs(x = "Diâmetro",
       y = "Volume",
       title = "Gráfico de dispersão")
```

Regressão Linear Simples

Gráfico de dispersão:



Coeficiente de correlação linear de Pearson:

O *coeficiente de correlação linear de Pearson* é uma medida numérica da força de associação linear entre duas variáveis quantitativas.

Sejam $x_1, x_2, \ldots, x_n$ n valores observados da variável aleatória quantitativa X e sejam $y_1, y_2, \ldots, y_n$ n valores observados da variável aleatória quantitativa Y. A correlação amostral, r, entre X e Y é definida por

$$r = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2}}. \qquad (1)$$

Coeficiente de correlação linear de Pearson:

- O coeficiente de correlação linear é adimensional e é tal que $-1 \leq r \leq 1$.
- Se $r > 0$, temos que as duas variáveis possuem uma relação linear positiva.
- Se $r < 0$, temos que as duas variáveis possuem uma relação linear negativa.
- Quando $r = 0$, temos uma ausência de relação linear entre as duas variáveis.

Coeficiente de correlação linear de Pearson:

```
cor(dados$diametro, dados$volume)
```

```
## [1] 0.9671194
```

Modelo de Regressão Linear Simples

Seja:

- Y a variável resposta;
- $y_1, y_2, \ldots, y_n$, n valores observados da variável resposta Y;
- X a variável preditora;
- $x_1, x_2, \ldots, x_n$, n valores observados da variável preditora.

As observações (x_i, y_i), $i = 1, 2, \ldots, n$ são pares de valores observados de (X, Y)

Modelo de Regressão Linear Simples

É muito pouco provável que as coordenadas $(x_1, y_1), \ldots, (x_n, y_n)$ forneçam exatamente uma linha reta: haverá algum erro que deve ser considerado na construção do modelo. Assim, o modelo de regressão linear simples é descrito por

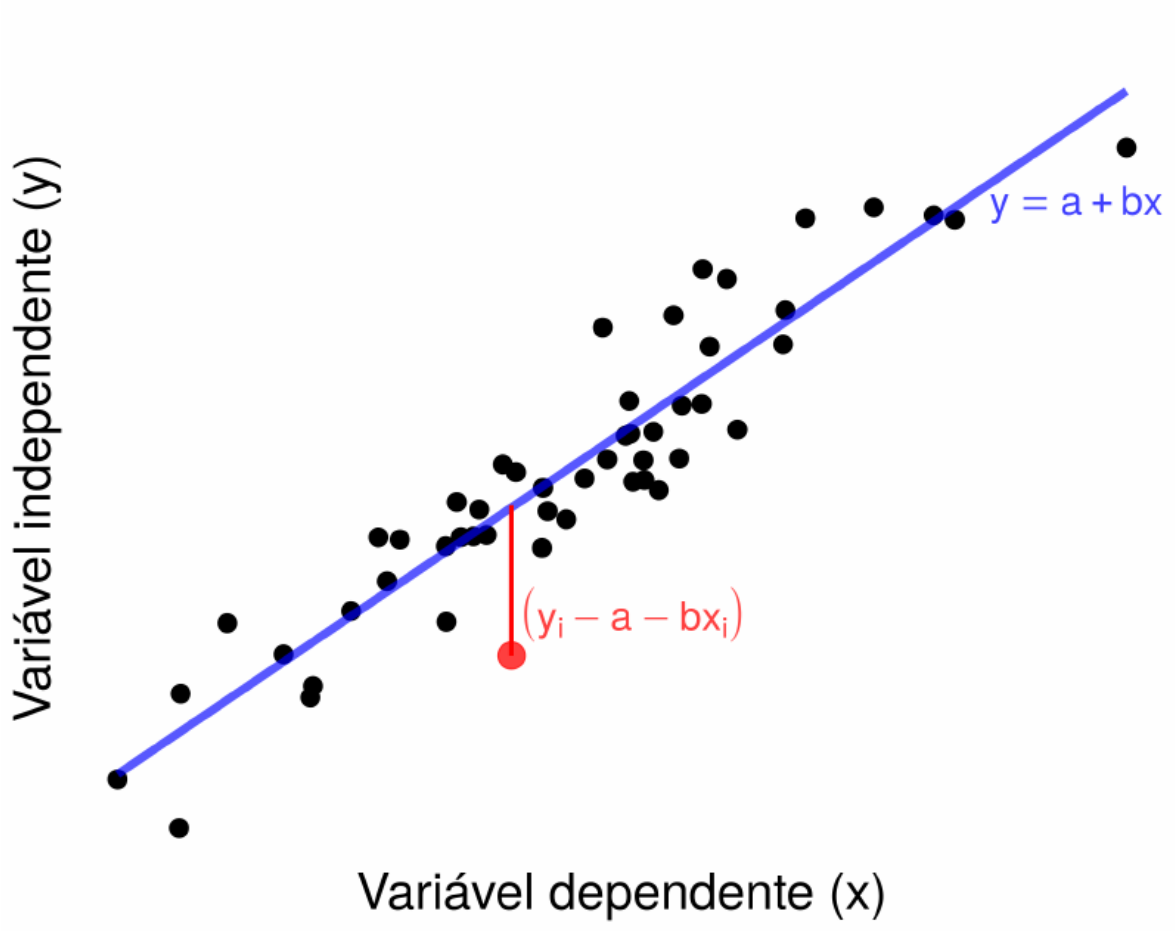
$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \tag{2}$$

onde β_0 é o intercepto, β_1 é o parâmetro de inclinação e ϵ_i é o erro aleatório do valor de y_i com relação à reta $\beta_0 + \beta_1 x_i$, para $i = 1, 2, \ldots, n$.

β_0 e β_1 são parâmetros (populacionais) desconhecidos que devem ser estimados utilizando os métodos de estimação de Inferência Estatística.

Modelo de Regressão Linear Simples

Ilustração dos erros em regressão linear simples:



Modelo de Regressão Linear Simples

Suposições do modelo de regressão linear simples:

No modelo de regressão linear simples usual, os ϵ_i's são variáveis aleatórias sujeitas às seguintes condições:

- O valor esperado de cada erro é zero: $E(\epsilon_i) = 0,\ i = 1, \ldots, n.$
- Os erros têm a mesma variância: $Var(\epsilon_i) = \sigma^2,\ i = 1, \ldots, n.$
- Os erros são não correlacionados:
  $Cov(\epsilon_i, \epsilon_j) = 0,\ \ i \neq j,\ \ i, j = 1, 2, \ldots, n.$

Modelo de Regressão Linear Simples

De uma maneira mais simples, podemos enunciar as suposições do modelo de regressão linear simples como segue:

- Linearidade: a variável resposta Y tem uma relação (aproximadamente) linear com a variável preditora X.
- Homoscedasticidade: para cada valor de X, a distribuição dos erros tem a mesma variância. Isso significa dizer que o nível de erro no modelo é aproximadamente o mesmo independente do valor da variável preditora.
- Independência dos erros: os erros não devem ser correlacionados. Idealmente, não deve ocorrer nenhum padrão entre resíduos consecutivos.

Por último, fazemos uma suposição extra:

- Normalidade: os erros do modelo são normalmente distribuídos.

Estimação

Os parâmetros β_0 e β_1 são desconhecidos e devem ser estimados utilizando os dados amostrais observados.

O método de mínimos quadrados (MMQ) é mais utilizado do que qualquer outro procedimento de estimação em modelos de regressão e fornece os estimadores de β_0 e β_1 tal que a soma de quadrados das diferenças entre as observações y_i's e a linha reta ajustada seja mínima.

Assim, de todos os possíveis valores de β_0 e β_1, os estimadores de mínimos quadrados (EMQ) serão aqueles que minimizam a soma de quadrados dos erros, que é dada por

$$\sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2. \tag{3}$$

Estimação

Usando o MMQ, as estimativas de β_0 e β_1 são, respectivamente

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x} \tag{4}$$

e

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} y_i x_i - n\overline{y}\ \overline{x}}{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2}, \tag{5}$$

onde $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ e $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ são as médias amostrais dos y_i's e x_i's, respectivamente.

Reta ajustada (modelo ajustado):

Uma vez as estimativas de β_0 e β_1 tenham sido obtidas, teremos a reta de regressão linear ajustada.

O modelo de regressão linear simples ajustado é

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i, \ i = 1, \ldots, n, \tag{6}$$

que é a estimativa pontual da média de Y_i para um particular x_i.

Reta ajustada (modelo ajustado):

No R, a função mais comum para ajustar um modelo de regressão linear simples é *lm()*. Os argumentos importantes para essa função são uma fórmula de modelo e um *data.frame* que contém os dados. A fórmula é simbólica.

```
fit <- lm(volume ~ diametro, data = dados)
```

Reta ajustada (modelo ajustado):

```
fit
```

```
##
## Call:
## lm(formula = volume ~ diametro, data = dados)
##
## Coefficients:
## (Intercept)      diametro
##     -36.943         5.066
```

Interpretação dos parâmetros:

Para os dados do problema, a reta de regressão ajustada é

$$\hat{y}_i = -36.943 + 5.066 \cdot x_i, \;\; i = 1, \ldots, n. \tag{7}$$

Mas, o que isso significa?

Isto é, como podemos interpretar essa reta ajustada?

Interpretação dos parâmetros:

A inclinação de uma reta é a mudança na variável y sobre a mudança na variável x. Se a mudança na variável x é um, então a inclinação é interpretada como a mudança em y para um incremento de uma unidade em x. Essa mesma interpretação pode ser aplicada ao parâmetro de inclinação da reta de regressão linear simples ajustada. Assim, temos que:

- $\hat{\beta}_1$ representa o aumento estimado em Y para cada aumento de uma unidade em X. Se o valor de $\hat{\beta}_1$ é negativo, então temos um incremento negativo.
- $\hat{\beta}_0$ é o intercepto da linha de regressão com o eixo-y. Então, quando $X = 0$ é um valor que faz sentido para os dados estudados, $\hat{\beta}_0$ é a estimativa do valor de Y quando $X = 0$.

Interpretação dos parâmetros:

Para os dados de peso e consumo de combustível:

- $\hat{\beta}_1$: Para cada aumento de 1 polegada no diâmetro da árvore, estima-se um acréscimo médio de 5.066 pés cúbicos no volume de madeira comercializável. Contexto Biométrico: Representa a taxa de variação marginal do volume em relação ao crescimento diamétrico.
- $\hat{\beta}_0 = -36.943$: É o valor predito para o volume de madeira quando o diâmetro do tronco é zero. Não possui interpretação biológica direta. Um volume negativo é impossível na realidade. O intercepto atua apenas como uma constante de ajuste matemático para ancorar a reta de mínimos quadrados no plano cartesiano.

Teste de Hipóteses (Teste de Significância do Modelo)

Como determinamos se há ou não uma relação linear (estatisticamente significativa) entre a variável resposta e a variável preditora?

Se a inclinação da reta de regressão é positiva, então a relação linear é positiva (se uma variável aumenta, a outra também aumenta).

Se a inclinação for negativa, então a relação linear é negativa (se uma variável aumenta, a outra diminui).

Se a inclinação é zero, então enquanto uma variável aumenta, a outra permanece constante: isto é, não há uma *relação preditiva*.

Teste de Hipóteses (Teste de Significância do Modelo)

Para verificar a ausência de relação preditiva entre X e Y, fazemos um teste de hipóteses, conhecido como **teste de significância do modelo**:

$$H_0 : \beta_1 = 0 \quad vs. \quad H_1 : \beta_1 \neq 0. \tag{8}$$

Podemos utilizar a Análise de Variância (ANOVA) para testar a significância da regressão.

A ANOVA é baseada no particionamento da variabilidade total da variável resposta Y.

Teste de Hipóteses (Teste de Significância do Modelo)

Partição da soma de quadrados do modelo:

Soma de quadrados das observações: $\sum_{i=1}^{n}(y_i - \overline{y})^2 = SQT$.

Soma de quadrados dos resíduos: $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = SQR$.

Soma de quadrados do modelo ou da regressão:
$\sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2 = SQM$.

Assim, simbolicamente podemos escrever a equação:
$SQT = SQR + SQM$.

Teste de Hipóteses (Teste de Significância do Modelo)

Graus de liberdade:

A separação dos graus de liberdade é determinado como segue.

A soma de quadrados total tem $n-1$ graus de liberdade.

A soma de quadrados dos resíduos tem $n-2$ graus de liberdade.

A soma de quadrados do modelo tem 1 grau de liberdade.

Os graus de liberdade tem a propriedade de aditividade:
$gl_T = gl_R + gl_M$ ou $n-1 = n-2+1$.

Teste de Hipóteses (Teste de Significância do Modelo)

Teste F de significância do modelo:

Podemos utilizar o teste F da ANOVA para testar a hipótese $H_0 : \beta_1 = 0$.

Considere a estatística teste F_0 definida por

$$F_0 = \frac{SQM/gl_M}{SQR/gl_R} = \frac{SQM/1}{SQR/(n-2)} = \frac{QMM}{QMR}. \tag{9}$$

Segue que F_0 com (1,n-2) graus de liberdade: $F_0 \sim F_{1,n-2}$.

Portanto, para testar a hipótese $H_0 : \beta_1 = 0$, calcula-se a estatística F_0 e rejeita-se H_0 se

$$F_0 > F_{\alpha,1,n-2}. \tag{10}$$

Teste de Hipóteses (Teste de Significância do Modelo)

Tabela ANOVA:

Em resumo, temos a tabela da ANOVA para testar a significância da regressão:

Fonte de Variação	Soma de Quadrados	Graus de Liberdade	Quadrado Médio	F_0
Modelo	SQM	1	QMM	$\frac{QMM}{QMR}$
Resíduo	SQR	$n-2$	QMR	
Total	SQT	$n-1$		

Teste de Hipóteses (Teste de Significância do Modelo)

Teste t de significância do modelo:

Para o modelo de regressão linear simples, o teste F de significância do modelo é equivalentee a um teste t com estatística teste

$$T_0 = \frac{\hat{\beta}_1}{\sqrt{QMR/S_{xx}}}, \tag{11}$$

onde $S_{xx} = \sum_{i=1}^{n}(x_i - \overline{x})^2$ e QMR é o quadrado médio dos resíduos, definido por

$$QMR = \frac{SQR}{n-2} = \frac{\sum_{i=1}^{n} e_i^2}{n-2}. \tag{12}$$

A hipótese nula é rejeitada se $|T_0| > t_{\left(\frac{\alpha}{2}, n-2\right)}$ a um nível de significância $\alpha \in (0,1)$ pré-fixado.

Teste de Hipóteses (Teste de Significância do Modelo)

Usando p-valor para tomar decisão em testes de hipóteses:

O p-valor é o menor nível de significância para o qual rejeita-se H_0 com os dados observados.

Como decidir entre H_0 e H_1 usando o p-valor do teste?

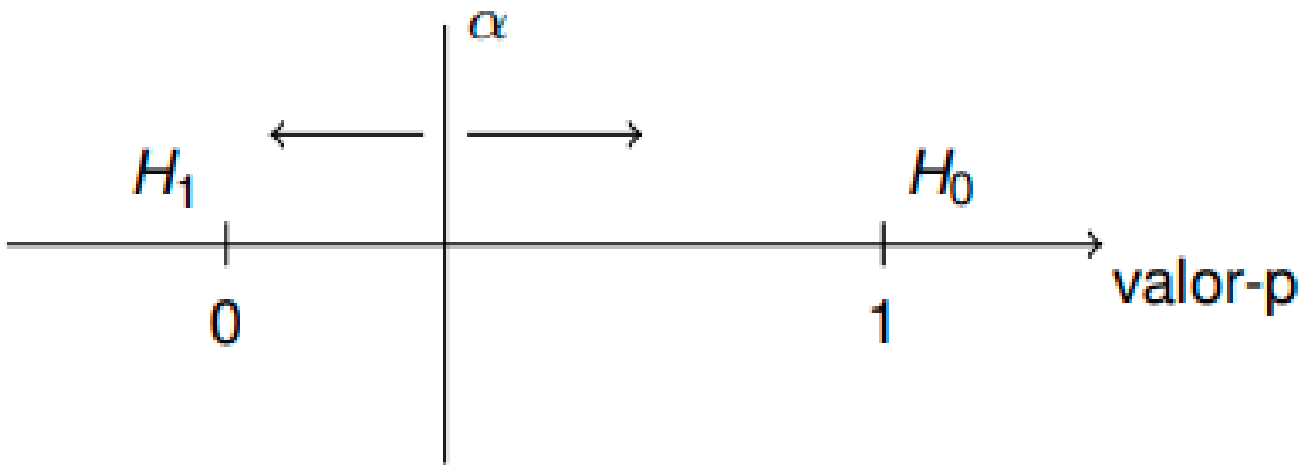
Usamos e interpretamos o p-valor do seguinte modo:

- Se o p-valor for grande, a amostra não fornece evidências suficientes para rejeitar H_0, então podemos decidir por H_0;
- Se o p-valor for pequeno, a amostra fornece evidências suficientes para rejeitar H_0, então podemos decidir por H_1.

Assim, para para um teste de hipóteses com nível de significância $\alpha \in (0, 1)$ pré-fixado, rejeitamos H_0 se o p-valor do teste for menor do que α.

Teste de Hipóteses (Teste de Significância do Modelo)

Ilustração de como decidir entre H_0 e H_1 usando o p-valor do teste:



Teste de Hipóteses (Teste de Significância do Modelo)

```r
fit <- lm(volume ~ diametro, data = dados)

fit_anova <- anova(fit)

fit_sumario <- summary(fit)
```

Teste de Hipóteses (Teste de Significância do Modelo)

```
fit
```

```
##
## Call:
## lm(formula = volume ~ diametro, data = dados)
##
## Coefficients:
## (Intercept)     diametro
##     -36.943        5.066
```

Teste de Hipóteses (Teste de Significância do Modelo)

```
fit_anova
```

```
## Analysis of Variance Table
##
## Response: volume
##           Df Sum Sq Mean Sq F value    Pr(>F)
## diametro   1 7581.8  7581.8  419.36 < 2.2e-16 ***
## Residuals 29  524.3    18.1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.'
```

Teste de Hipóteses (Teste de Significância do Modelo)

```
fit_sumario
```

```
##
## Call:
## lm(formula = volume ~ diametro, data = dados)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -8.065 -3.107  0.152  3.495  9.587
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -36.9435     3.3651  -10.98 7.62e-12 ***
## diametro      5.0659     0.2474   20.48  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.252 on 29 degrees of freedom
## Multiple R-squared:  0.9353, Adjusted R-squared:  0.9331
## F-statistic: 419.4 on 1 and 29 DF,  p-value: < 2.2e-16
```

Estimação de σ^2

A estimativa de σ^2 é necessária para testar hipóteses e construir estimativas intervalares associadas ao modelo de regressão.

Um estimador não viciado de σ^2 é

$$\hat{\sigma}^2 = \frac{SQR}{n-2} = QMR. \tag{13}$$

Intervalos de Confiança

Intervalos de confiança $(1-\alpha)100\%$ **para os parâmetros:**

- Para β_1 o intervalo de confiança $(1-\alpha)100\%$ é dado por:

$$\hat{\beta}_1 \pm t_{(\frac{\alpha}{2}, n-2)} \sqrt{\frac{QMR}{S_{xx}}}. \tag{14}$$

- Para β_0 o intervalo de confiança $(1-\alpha)100\%$ é dado por:

$$\hat{\beta}_0 \pm t_{(\frac{\alpha}{2}, n-2)} \sqrt{QMR \left(\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}} \right)}. \tag{15}$$

Análise de Resíduos e Diagnóstico

Análise de diagnóstico:

Antes de fazer inferência sobre os parâmetros, devemos verificar as suposições do modelo. De fato, não devemos considerar e usar um modelo ajustado antes de verificar as suposições do modelo. Essa *verificação das suposições do modelo* é feita por meio da *Análise de Diagnóstico*.

A Análise de Diagnóstico de um modelo de regressão compreende a *Análise de Resíduos* e o *Diagnóstico de Influência*, fornecendo ferramentas para avaliar se o modelo ajustado está em conformidade com as suposições, verificar a presença de observações discrepantes e avaliar se um modelo representa adequadamente os dados em estudo.

De uma maneira simplista, podemos dizer que a análise de diagnóstico é a etapa da análise que nos dirá se o modelo considerado é de fato um boa representação/descrição da relação entre as variáveis em estudo.

Análise de Resíduos e Diagnóstico

Resíduos:

A diferença entre o valor observado y_i e o valor ajustado correspondente $\hat{y}_i$ é um resíduo. Matematicamente, o $i-$ésimo resíduo é

$$e_i = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i), \ i = 1, \ldots, n. \tag{16}$$

Em geral, a maioria dos métodos de diagnóstico em regressão são baseados, principalmente, no estudo dos resíduos do modelo. Isso porque os resíduos podem ser vistos como realizações dos erros do modelo. Assim, para checar as suposições para os erros de variância constante (homoscedasticidade), ausência de correlação e normalidade, devemos verificar se os resíduos parecem com uma amostra aleatória de uma distribuição com essas propriedades.

Resíduos estudentizados:

A variância de cada resíduo e_i, $i = 1, \ldots, n$

$$Var(e_i) = \sigma^2(1 - h_{ii}), \tag{17}$$

onde h_{ii} são os elementos da diagonal principal da matriz Chapéu $H = X(X'X)^{-1}X'$, sendo X a matriz de regressão.

Temos que $0 < h_{ii} < 1$, pois $h_{ii} > 0$ e $1 - h_{ii} > 0$.

Um valor alto de h_{ii} faz com que $Var(e_i)$ seja pequena e o valor ajustado $\hat{y}_i$ será próximo de y_i.

Os resíduos estudentizados são dados por

$$r_i = \frac{e_i}{\sqrt{QMR(1 - h_{ii})}}, \quad i = 1, \ldots, n. \tag{18}$$

Análise de Resíduos e Diagnóstico

Gráficos para análise de resíduos:

- Resíduos contra valores ajustados: útil para verificar a suposição de homoscedasticidade, para detectar outliers e para verificar se a forma do modelo (linearidade) está correta.
- Resíduos contra ordem de coleta dos dados: caso os dados sejam coletados sequencialmente, esse gráfico é útil para verificar a presença de correlação serial nos resíduos.
- Resíduos contra variáveis explicativas (incluídas no modelo): em regressão linear simples, esse gráfico tem a mesma utilidade que o gráfico de resíduos contra os valores ajustados.

Análise de Resíduos e Diagnóstico

Gráficos para análise de resíduos:

- Resíduos contra variáveis explicativas (não incluídas no modelo): útil para verificar se alguma covariável que foi omitida na análise deve ser incluída no modelo.
- Gráfico normal probabilístico (Q-Q norm): utilizado para verificar a normalidade dos resíduos.

Análise de Resíduos e Diagnóstico

Gráficos para análise de resíduos:

Em regressão linear simples, para os gráficos dos resíduos contra preditos/ordem/covariáveis, o comportamento que indica que as suposições do modelo são satisfeitas e que o ajuste é razoável é a ausência de qualquer comportamento. Desta forma, esperamos que os gráficos mostrem os pontos dispostos horizontalmente em torno de zero.

No caso do gráfico normal probabilístico (Q-Q norm), o comportamento que indica que as suposições de normalidade é satisfeita é os pontos dispersos segundo uma reta identidade crescente.

Análise de Resíduos e Diagnóstico

Gráficos para análise de resíduos: padrão satisfatório.

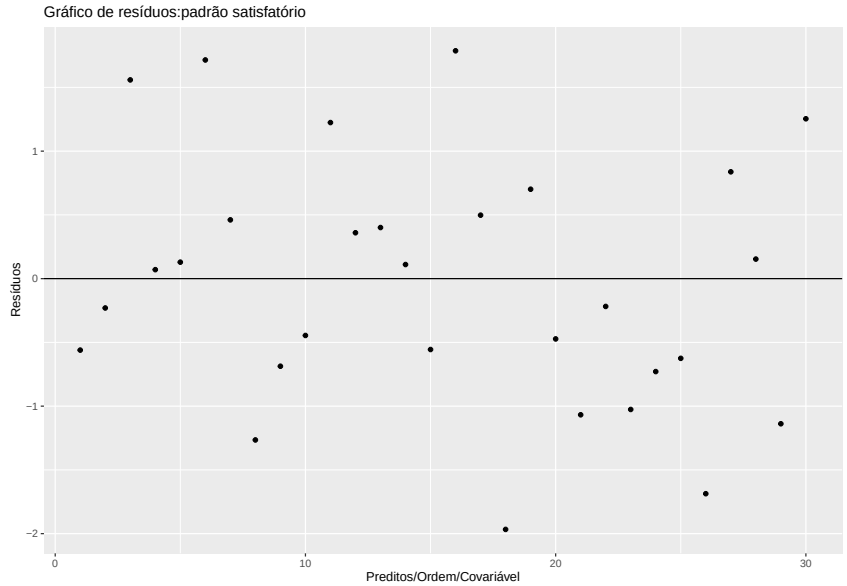


Figura 1: Gráfico de resíduos:padrão satisfatório

Gráficos para análise de resíduos: padrão satisfatório do gráfico normal probabilístico.

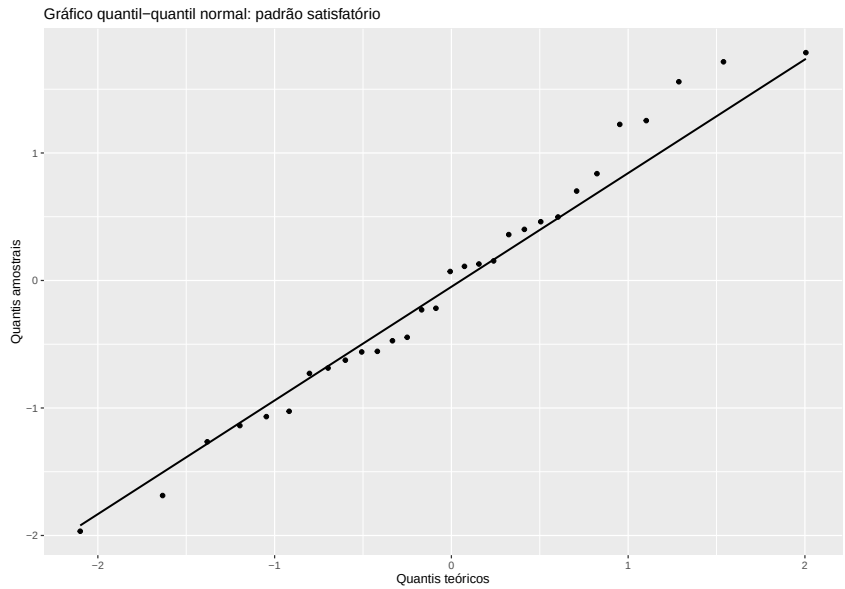
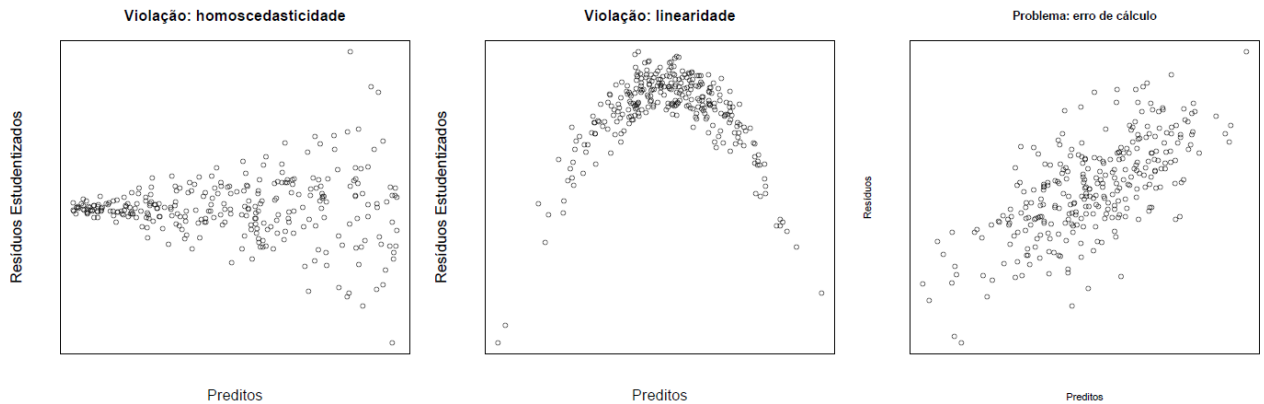


Figura 2: Gráfico quantil-quantil normal: padrão satisfatório

Análise de Resíduos e Diagnóstico

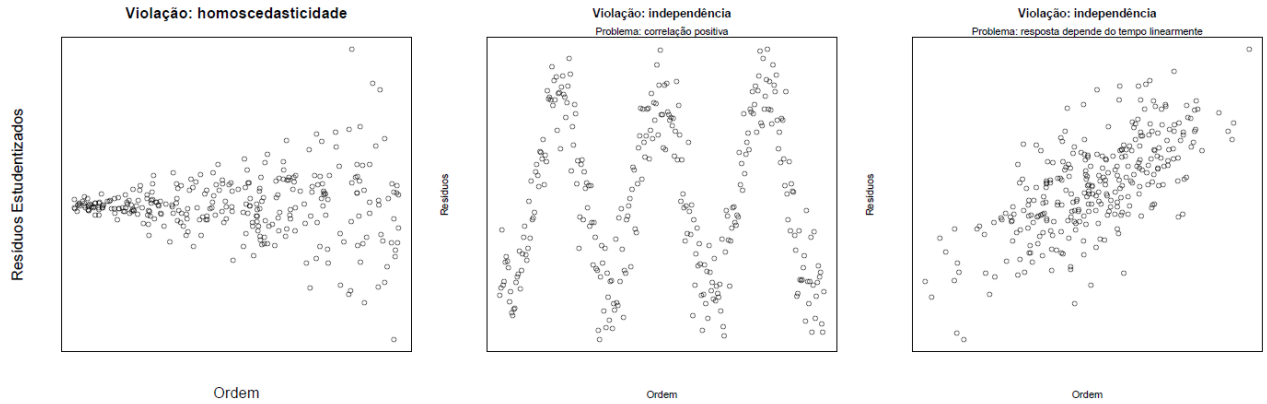
Gráficos para análise de resíduos: padrões que indicam violações das suposições do modelo.

Usualmente, a presença de um padrão nesses gráficos pode ser um indicativo de violação das suposições do modelo.



Análise de Resíduos e Diagnóstico

Gráficos para análise de resíduos: padrões que indicam violações das suposições do modelo.



Alguns comentários sobre as consequências das violações das suposições do modelo:

Quando as suposições sobre os erros são violadas, os estimadores de mínimos quadrados já não estimadores ótimos e os intervalos de confiança e testes de hipóteses para os parâmetros (construídos sob essa suposição) não são apropriados.

Se as suposições do modelo são violadas, será necessário fazer alguns ajustes no modelo ou nos dados. Por exemplo, podemos considerar a inclusão (ou exclusão) de variáveis preditoras no modelo; podemos considerar outro método de estimação dos parâmetros; podemos considerar transformar os dados (variável resposta e/ou variável preditora). E podemos, ainda, considerar um modelo diferente para o problema.

Diagnóstico de influência:

A análise visual dos resíduos possibilita a identificação de observações outliers, que são observações discrepantes ou incomuns.

Uma observação outlier pode ou não ser influente. Observações influentes são aquelas que, de acordo com vários critérios, aparecem exercendo forte impacto no modelo ajustado.

Uma vez decido que um ponto é outlier, devemos prosseguir com o diagnóstico de influência no sentido de verificar se o ponto outlier é influente no ajuste do modelo.

Diagnóstico de influência:

Um ponto influente é aquele cuja remoção do conjunto de dados poderia causar uma grande mudança no ajuste, sendo que a influência pode ocorrer diretamente no ajuste do modelo ou na estimação dos parâmetros.

Desta maneira, a maioria das métricas de diagnóstico de influência (que não discutiremos aqui nesse momento) são baseadas no princípio de deleção de casos e comparam o que acontece com o ajuste do modelo (ou com a estimação dos parâmetros) na presença e na ausência de uma dada observação.

Diagnóstico de influência: outliers em X.

A determinação de pontos que são outliers em X é baseada no estudo das alavancagem das observações h_{ii}, que são os elementos da diagonal principal da matriz Chapéu $H = X(X'X)^{-1}X'$, onde X é a matriz de regressão.

A alavancagem $h_{ii} = x_i^{'}(X'X)^{-1}x_i$ de uma observação mede a distância das covariáveis com relação às médias de todos os valores das covariáveis. Como $0 < h_{ii} < 1$, um valor alto de h_{ii} fazem com que $Var(e_i)$ seja pequena e o valor ajustado $\hat{Y}_i$ será próximo de Y_i indicando a influência.

Na prática, um caso pode ser considerado outlier se sua alavancagem for maior do que $2(p+1)/n$.

Diagnóstico de influência: outlier em Y.

A forma mais simples de detecção de observações outliers em Y é através dos resíduos. Um valor de resíduo grande é indicativo de outlier em Y, mas um valor de resíduo pequeno não é indicativo de ausência de outliers.

Diagnóstico de influência no ajuste:

Para verificar se uma observação é influente no ajuste, podemos considerar a métrica

$$DFFITS_{(i)} = \frac{\hat{y}_i - \hat{y}_{(i)}}{\sqrt{QMR_{(i)} h_{ii}}},$$

onde $\hat{y}_{(i)}$ é o valor predito considerando o modelo sem a i-ésima observação e $QMR_{(i)}$ é o QMR considerando o modelo sem a i-ésima observação.

Para amostras de tamanho pequeno, dizemos que um ponto outlier é influente se $|DFFITS_{(i)}| > 1$.

Para amostras de tamanho moderado ou grande, um ponto outlier é influente se $|DFFITS_{(i)}| > 2\sqrt{(p+1)/n}$.

Na prática, para qualquer tamanho amostral, podemos considerar que uma observação é influente se $|DFFITS_{(i)}| > 2$.

Diagnóstico de influência na estimação:

Para verificar se uma observação é influente na estimação dos parâmetros, podemos considerar a métrica

$$DFBETA_{j(i)} = \frac{\hat{\beta}_j - \hat{\beta}_{j(i)}}{\sqrt{QMR_{(i)} c_{jj}}},$$

onde c_{jj} é o j-ésimo elemento da diagonal de $(X'X)^{-1}$.

Para amostras de tamanho moderado ou grande, um ponto outlier é influente se $|DFBETA_{j(i)}| > 2\sqrt{(p+1)/n}$.

Na prática, para qualquer tamanho amostral, podemos considerar que uma observação é influente se $|DFBETA_{j(i)}| > 2$.

Diagnóstico de influência na estimação:

A distância de Cook mede a influência da i-ésima observação nos coeficientes da regressão simultaneamente.

$$D_i = \frac{(\hat{\beta} - \hat{\beta}_{(i)})'(X'X)^{-1}(\hat{\beta} - \hat{\beta}_{(i)})}{pQMR} = \frac{(\hat{Y} - \hat{Y}_{(i)})'(X'X)^{-1}(\hat{Y} - \hat{Y}_{(i)})}{pQMR}.$$

Para obter os valores de D_i sem que seja necessário ajustar o modelo sem cada uma das observações fazemos

$$D_i = \frac{e_i^2}{(p+1)QMR}\frac{h_{ii}}{(1-h_{ii})^2}.$$

Uma observação é considerada influente se $D_i > F^{1-\alpha}_{p+1,n-p-1}$.

Na prática, uma observação é considerada influente se $D_i > 1$ ou $D_i > 4/n$.

Modelo de Regressão Linear Simples

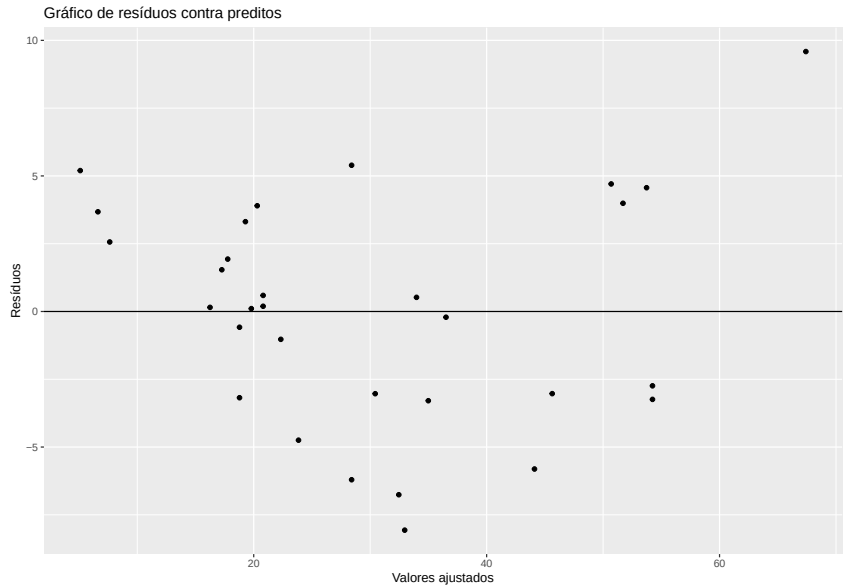
```
library(tidymodels)

df_fit_resid <- augment(fit)

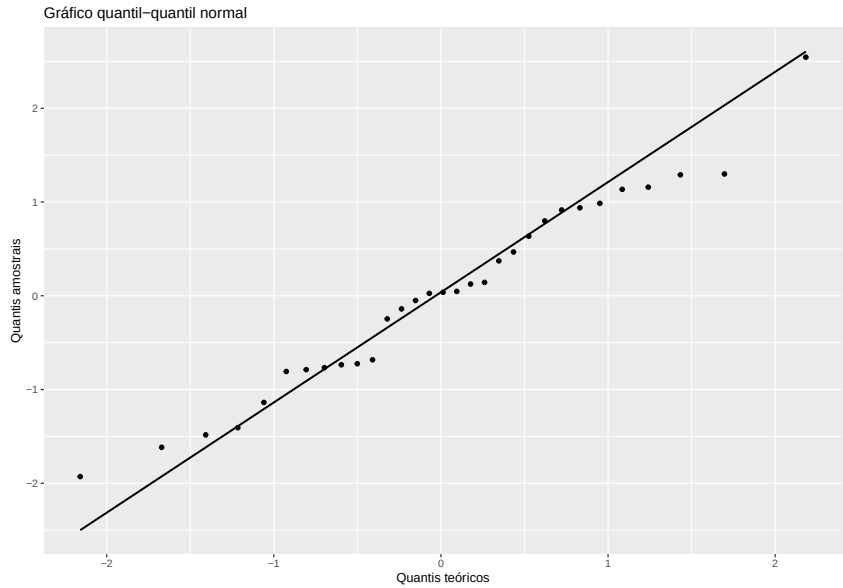
glimpse(df_fit_resid)
```

```
## Rows: 31
## Columns: 8
## $ volume     <dbl> 10.3, 10.3, 10.2, 16.4, 18.8,
## $ diametro   <dbl> 8.3, 8.6, 8.8, 10.5, 10.7, 10
## $ .fitted    <dbl> 5.103149, 6.622906, 7.636077,
## $ .resid     <dbl> 5.1968508, 3.6770939, 2.56392
## $ .hat       <dbl> 0.11514037, 0.10539539, 0.099
## $ .sigma     <dbl> 4.199409, 4.264423, 4.297028,
## $ .cooksd    <dbl> 1.098362e-01, 4.924417e-02, 2
## $ .std.resid <dbl> 1.29930490, 0.91431779, 0.635
```

```r
ggplot(data = df_fit_resid) +
  geom_point(aes(x = .fitted, y = .resid)) +
  geom_hline(yintercept = 0) +
  labs(x = "Valores ajustados", y = "Resíduos",
       title = "Gráfico de resíduos contra preditos")
```



```r
ggplot(data = df_fit_resid, aes(sample = .std.resid)) +
  stat_qq_line() +
  stat_qq_point() +
  labs(x = "Quantis teóricos", y = "Quantis amostrais",
       title = "Gráfico quantil-quantil normal")
```



```r
#teste de normalidade dos resíduos
#H0: normalidade
#H1: não normalidade

shapiro.test(fit$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  fit$residuals
## W = 0.97889, p-value = 0.7811
```

```r
#teste de homoscedasticidade dos resíduos
#H0: resíduos homoscedásticos
#H1: resíduos heteroscedásticos

library(lmtest)

bptest(fit)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  fit
## BP = 5.6197, df = 1, p-value = 0.01776
```

```r
#teste de correlação serial lag 1
#H0: não correlacionados
#H1: correlacionados

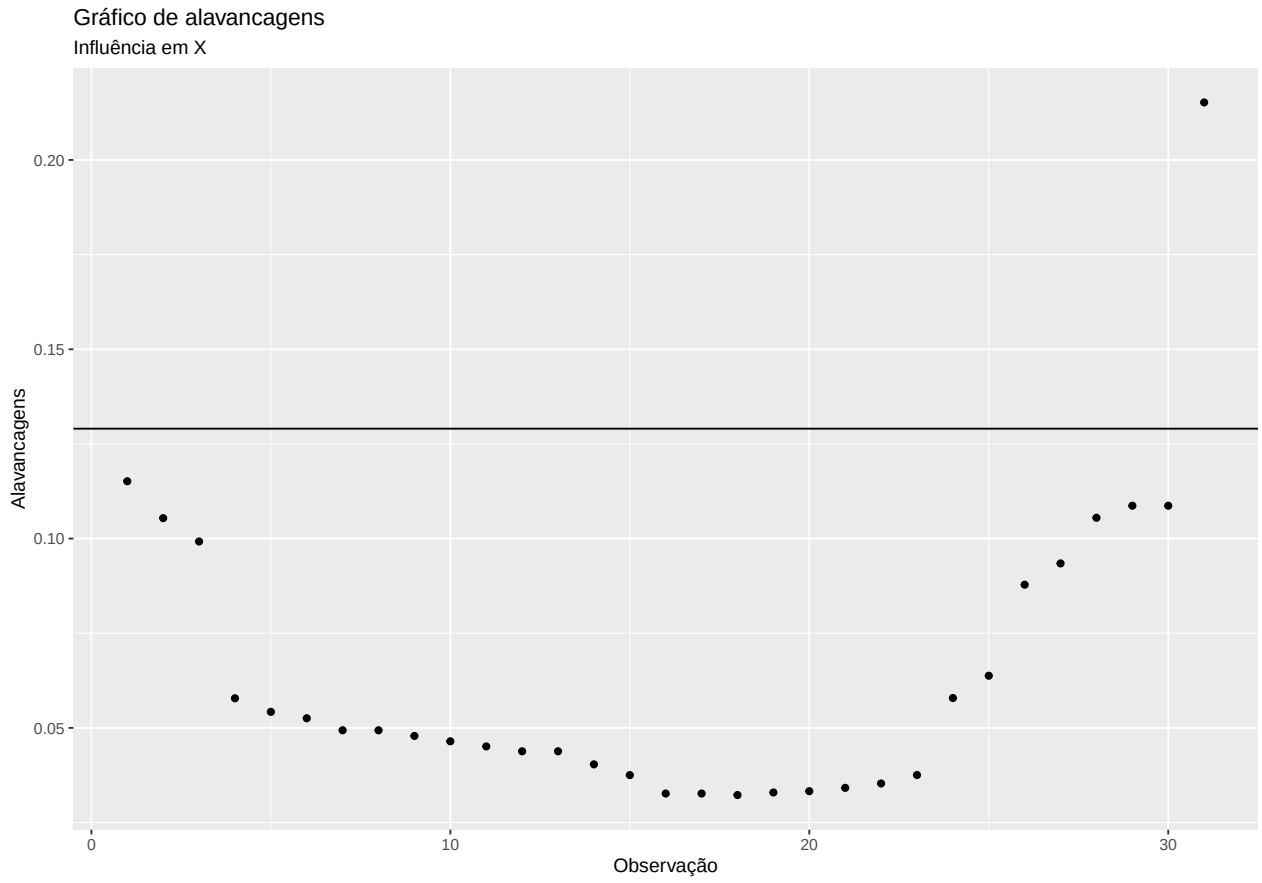
dwtest(fit)
```

```
##
##  Durbin-Watson test
##
## data:  fit
## DW = 1.4372, p-value = 0.03471
## alternative hypothesis: true autocorrelation is greater than 0
```

```r
#influência em X - alavancagens
#limite <- 2 * (p + 1) / n
#p = nro de covariáveis
#n = tamanho amostral

limite <- 2 * (1 + 1) / nrow(dados)

ggplot(data = df_fit_resid) +
  geom_point(aes(x = seq_along(.resid), y = .hat)) +
  geom_hline(yintercept = limite) +
  labs(x = "Observação", y = "Alavancagens",
       title = "Gráfico de alavancagens",
       subtitle = "Influência em X")
```

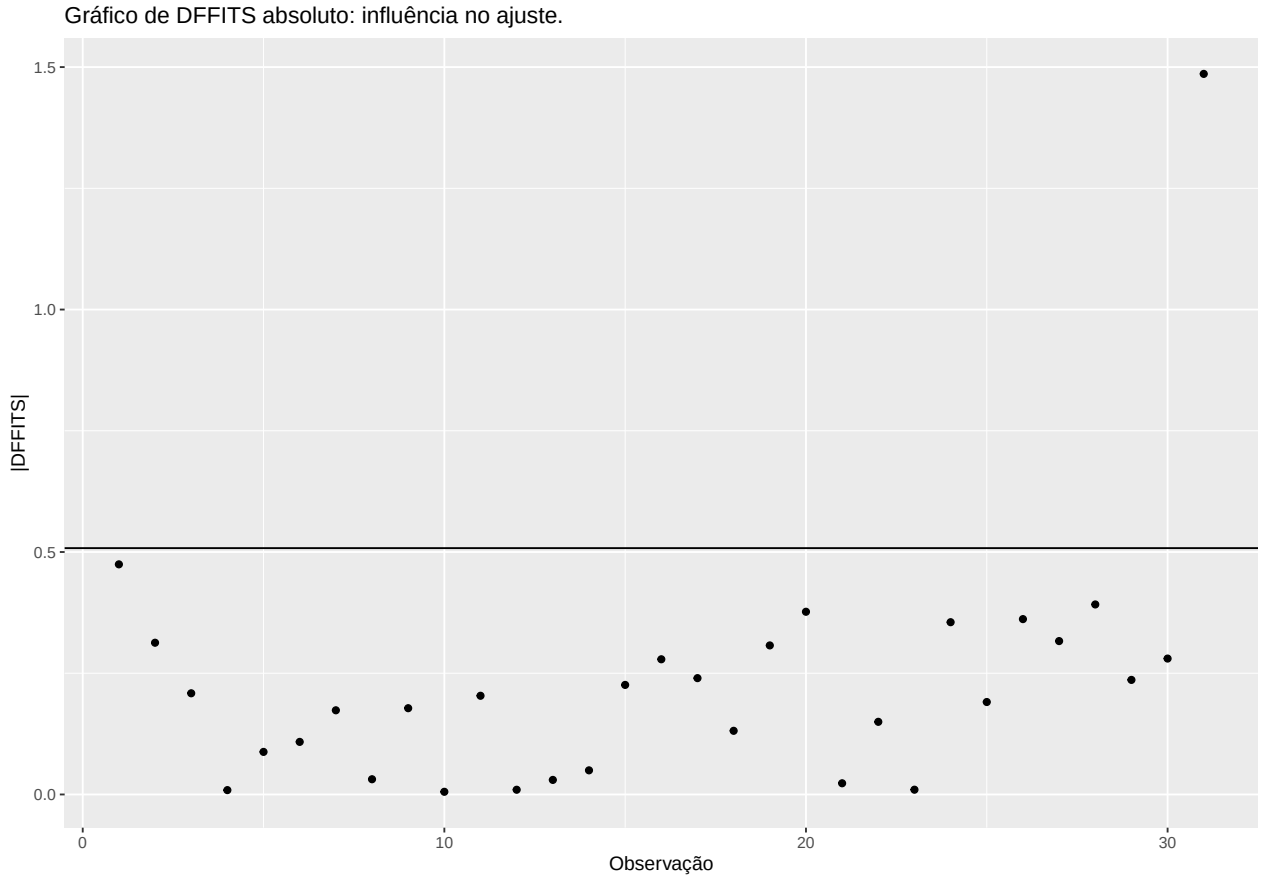

Gráfico de alavancagens
Influência em X
0.20
0.15
0.10
0.05
Alavancagens
0
10
20
30
Observação

```r
#influência no ajuste - DFFITS em módulo
#limite <- 2 * sqrt((p + 1) / n)
#limite <- 2
#p = nro de covariáveis
#n = tamanho amostral

limite <- 2 * sqrt((1 + 1) / nrow(dados))

df_fit_resid <- df_fit_resid |>
  mutate(.dffits = dffits(fit))

ggplot(data = df_fit_resid) +
  geom_point(aes(x = seq_along(.dffits),
                 y = abs(.dffits))) +
  geom_hline(yintercept = limite) +
  labs(x = "Observação", y = "|DFFITS|",
       title = "Gráfico de DFFITS absoluto",
       subtitle = "Influência no ajuste")
```


Gráfico de DFFITS absoluto: influência no ajuste.
1.5
1.0
0.5
0.0
|DFFITS|
0
10
20
30
Observação

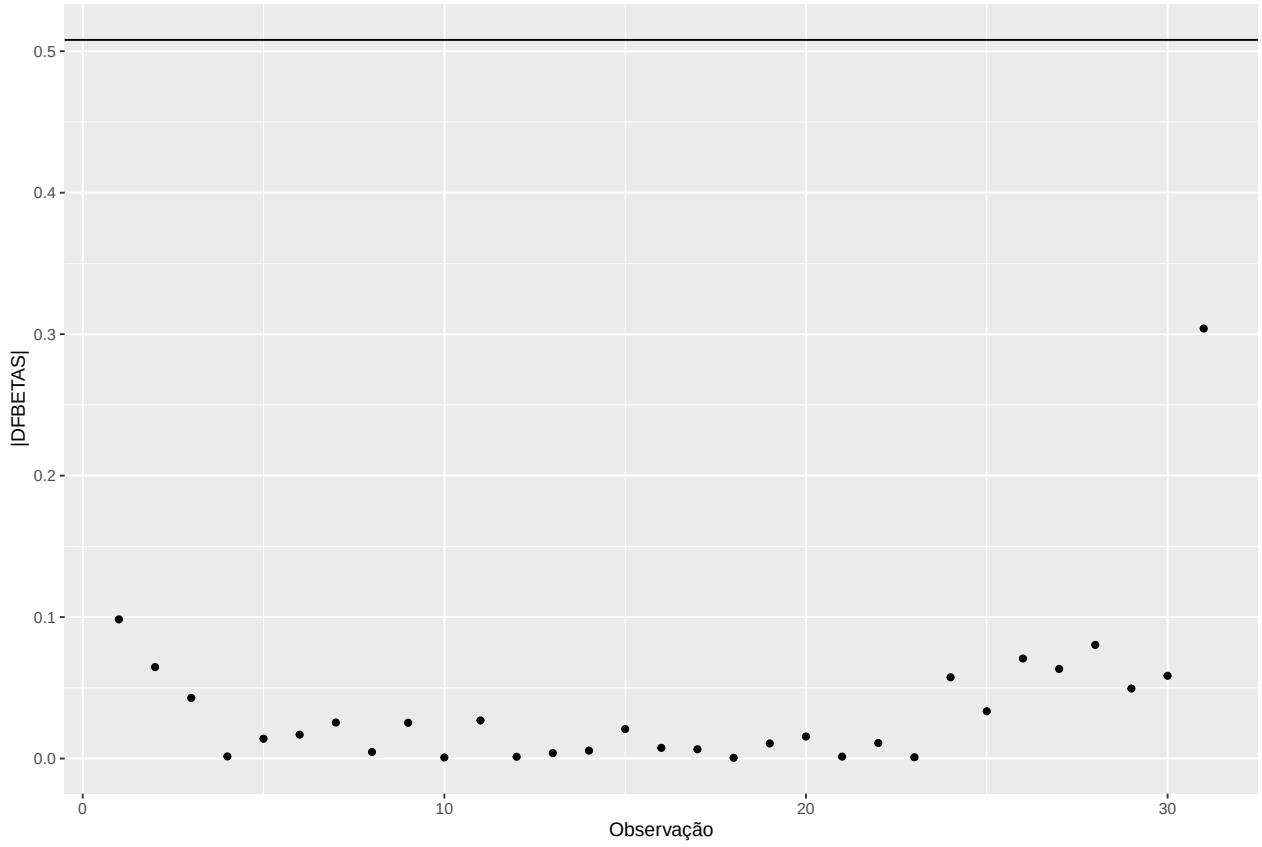
```r
#influência nos parâmetros - DFBETAS em módulo
#limite <- 2 * sqrt((p + 1) / n)
#limite <- 2
#p = nro de covariáveis
#n = tamanho amostral

limite <- 2 * sqrt((1 + 1) / nrow(dados))

dfbetas <- as_tibble(dfbeta(fit))
dfbetas <- dfbetas |>
  rename(intercepto = `(Intercept)`)

ggplot(data = dfbetas) +
  geom_point(aes(x = seq_along(intercepto),
                 y = abs(intercepto))) +
  geom_hline(yintercept = limite) +
  labs(x = "Observação", y = "|DFBETAS|",
       title = "Gráfico de DFBETAS para o intercepto.")
```

Gráfico de DFBETAS para o intercepto.



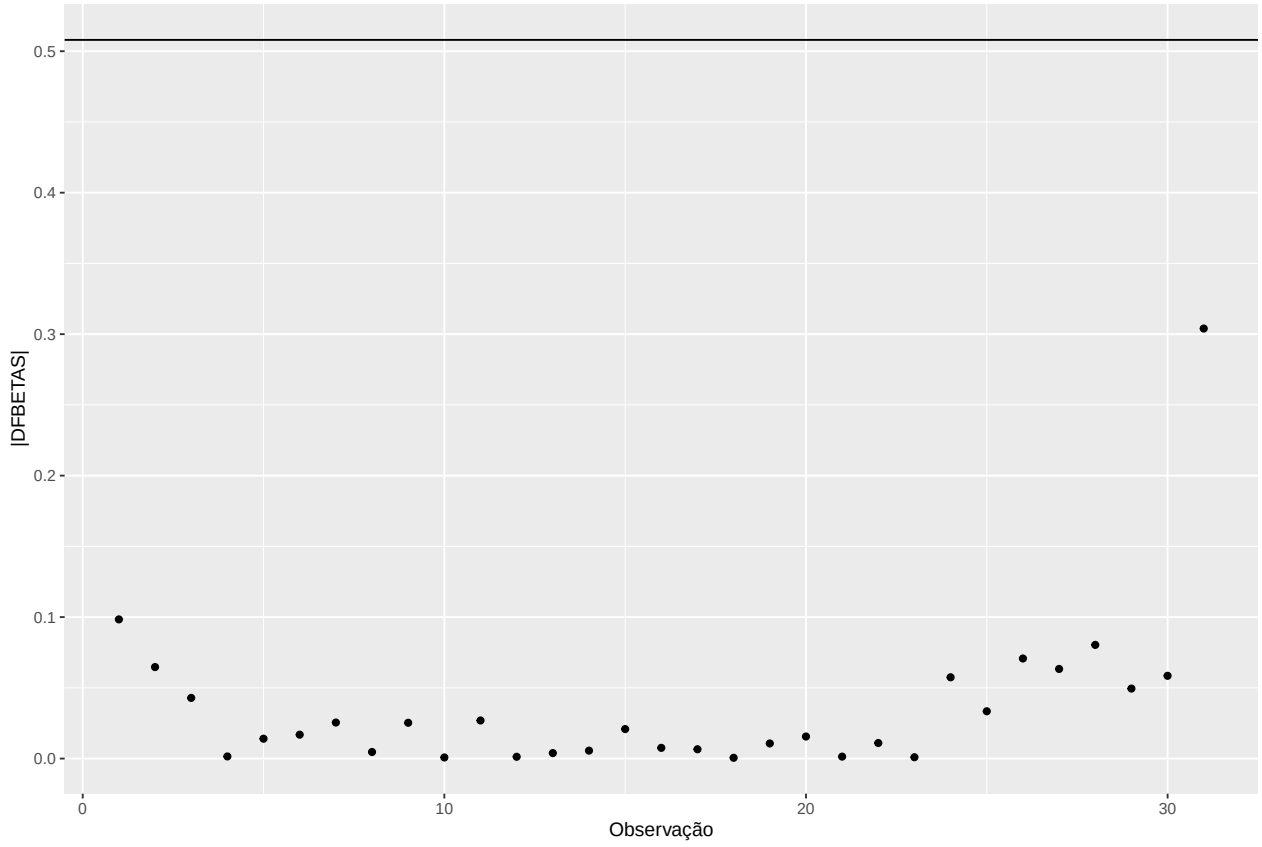
```r
#influência nos parâmetros - DFBETAS em módulo
#limite <- 2 * sqrt((p + 1) / n)
#limite <- 2
#p = nro de covariáveis
#n = tamanho amostral

limite <- 2 * sqrt((1 + 1) / nrow(dados))

dfbetas <- as_tibble(dfbeta(fit))
dfbetas <- dfbetas |>
  rename(intercepto = `(Intercept)`)

ggplot(data = dfbetas) +
  geom_point(aes(x = seq_along(diametro),
                 y = abs(diametro))) +
  geom_hline(yintercept = limite) +
  labs(x = "Observação", y = "|DFBETAS|",
       title = "Gráfico de DFBETAS para a inclinação.")
```

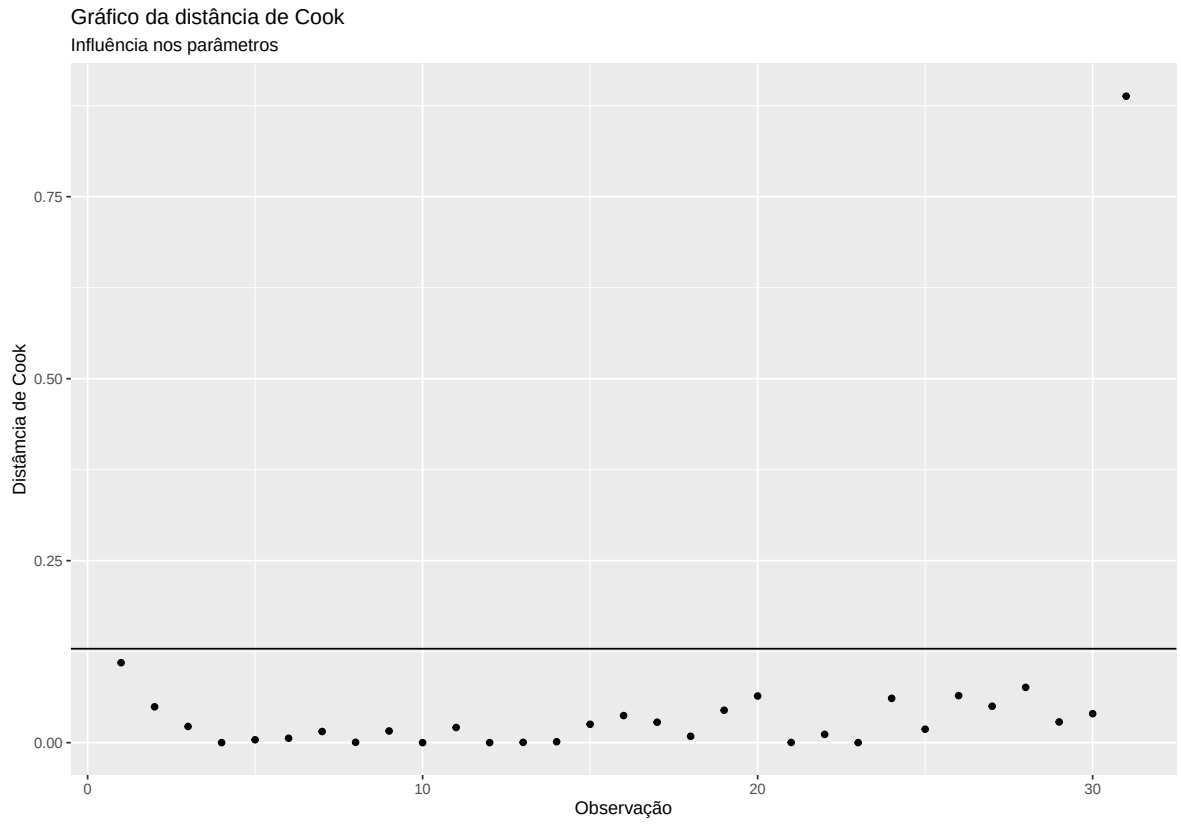
Gráfico de DFBETAS para a inclinação.



```r
#influência nos parâmetros - distância de Cook
#limite <- 1
#limite <- 4 / n
#n = tamanho amostral

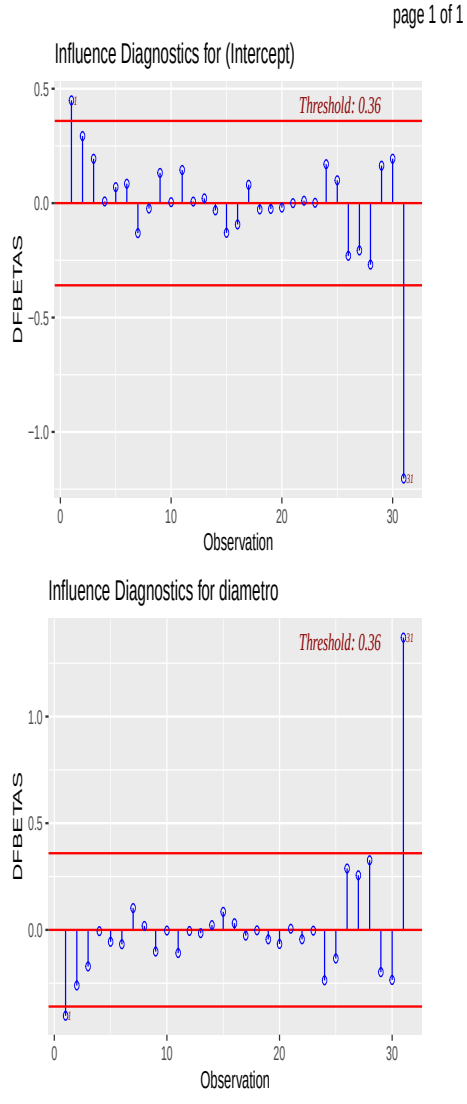
limite <- 4 / nrow(dados)

ggplot(data = df_fit_resid) +
  geom_point(aes(x = seq_along(.cooksd),
                 y = .cooksd)) +
  geom_hline(yintercept = limite) +
  labs(x = "Observação",
       y = "Distâmcia de Cook",
       title = "Gráfico da distância de Cook",
       subtitle = "Influência nos parâmetros")
```

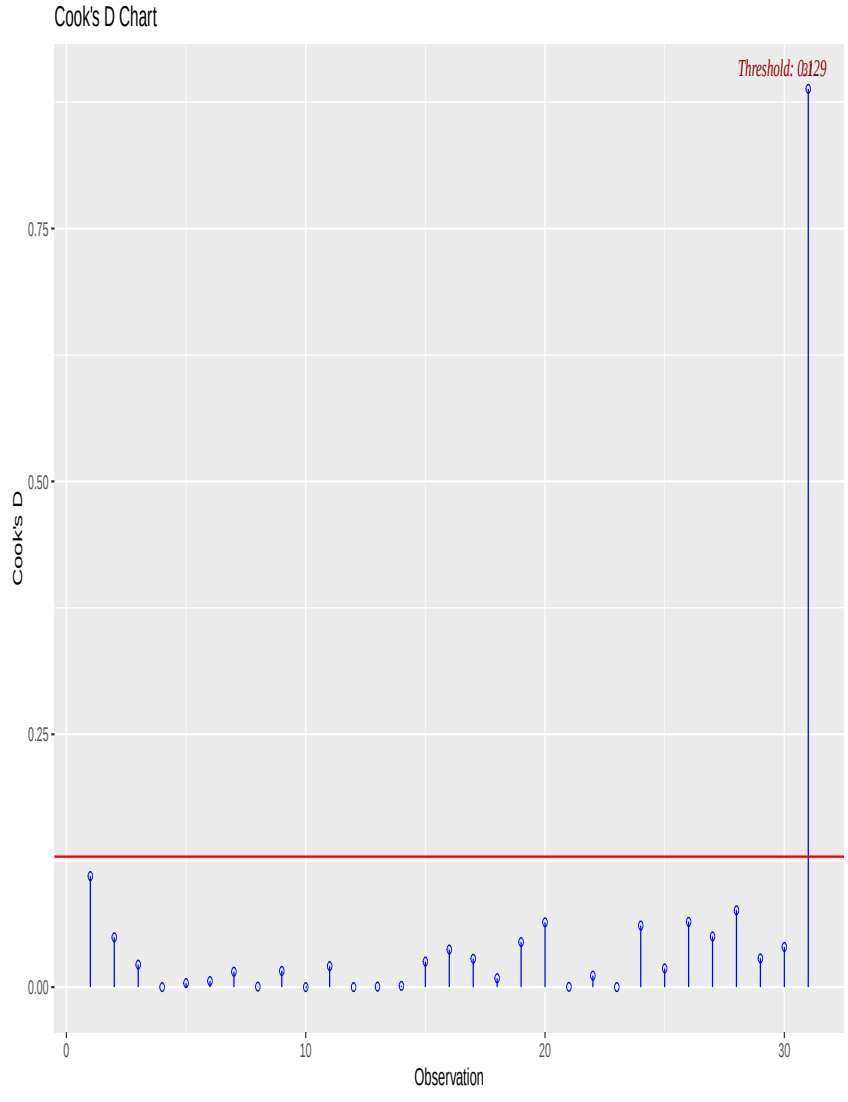

Gráfico da distância de Cook
Influência nos parâmetros
Distâmcia de Cook
0.75
0.50
0.25
0.00
0
10
20
30
Observação

```
library(olsrr)

ols_plot_dfbetas(fit)
```



```
ols_plot_cooksd_chart(fit)
```



```r
#inferência
fit_anova <- anova(fit)

fit_sumario <- summary(fit)

ic_parametros <- confint(fit)

diametro_novo <- tibble(diametro = c(9, 15))

predicao <- predict(fit, diametro_novo)
```

```r
#inferência
glance(fit)
```

```
## # A tibble: 1 x 12
##   r.squared adj.r.squared sigma statistic  p.value    df logLik   AIC   BIC
##       <dbl>         <dbl> <dbl>     <dbl>    <dbl> <dbl>  <dbl> <dbl> <dbl>
## 1     0.935         0.933  4.25      419. 8.64e-19     1  -87.8  182.  186.
## # i 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>
```

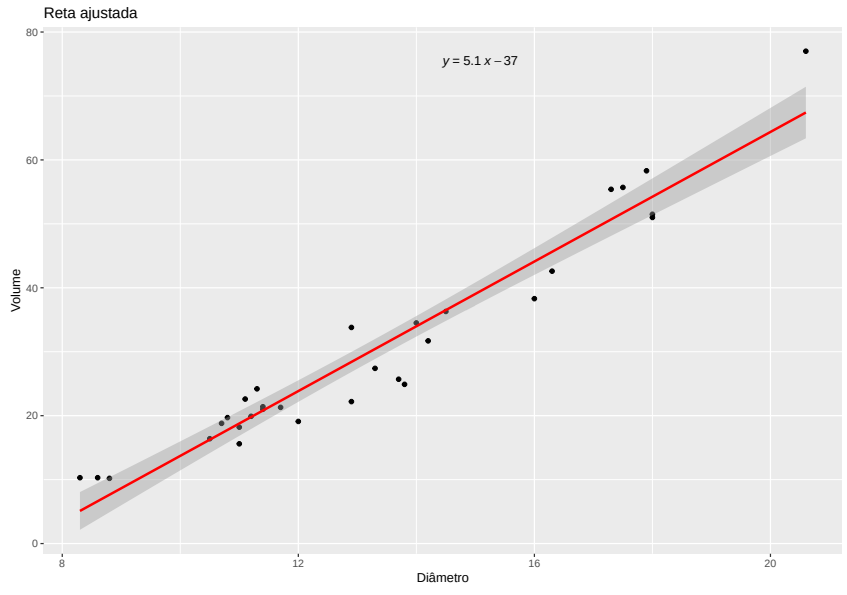
```r
#inferência
tidy(fit)
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)   -36.9      3.37      -11.0 7.62e-12
## 2 diametro        5.07     0.247      20.5 8.64e-19
```

```r
#inferência
tidy(fit_anova)
```

```
## # A tibble: 2 x 6
##   term          df sumsq meansq statistic   p.value
##   <chr>      <int> <dbl>  <dbl>     <dbl>     <dbl>
## 1 diametro       1 7582. 7582.       419.  8.64e-19
## 2 Residuals     29  524.   18.1       NA   NA
```

```r
#gráfico de dispersão com reta ajustada
ggplot(dados, aes(x = diametro, y = volume)) +
  geom_point() +
  stat_smooth(method = "lm", col = "red") +
  stat_regline_equation(label.x.npc = "center") +
  labs(x = "Diâmetro",
       y = "Volume",
       title = "Reta ajustada")
```



Exemplo 2.

Uma Unidade de Saúde realizou um estudo para investigar a relação entre a prática de atividade física e a pressão arterial sistólica de adultos.

Para isso, foi obtido um cum conjunto de dados contendo informações de 40 indivíduos. O objetivo é estudar se existe uma associação linear entre o tempo semanal dedicado à atividade física (variável explicativa, x) e a pressão arterial sistólica (variável resposta, y).

Os dados estão disponíveis na planilha *saude* do arquivo *dados_rls_20262.xlsx*.

▪

Exemplo 3.

Em Ciência Política, uma das hipóteses clássicas de comportamento eleitoral investiga a relação entre os recursos financeiros mobilizados por uma candidatura e o seu desempenho nas urnas.

Neste problema, analisamos o impacto dos gastos de campanha (variável explicativa x, em milhares de reais) sobre o percentual de votos obtidos (variável resposta y, em %) por 30 candidatos em uma eleição proporcional.

Os dados estão disponíveis na planilha *politica* do arquivo *dados_rls_20262.xlsx*.

▪

Exemplo 4.

Na Economia da Educação e na Psicologia Educacional, uma questão fundamental é entender como o tempo investido no aprendizado fora da sala de aula se traduz em desempenho acadêmico.

Neste problema, analisamos o impacto das horas semanais de estudo extraclasse variável independente x) sobre a nota final obtida em um exame padronizado de Matemática variável dependente y, em escala de 0 a 100) por uma amostra de 40 alunos do ensino médio.

Os dados estão disponíveis na planilha *educacao* do arquivo *dados_rls_20262.xlsx*.

▪

Exemplo 5.

Em Ecologia de Comunidades, uma das hipóteses mais investigadas é o gradiente macroclimático de biodiversidade: como a disponibilidade de recursos abióticos — como a água — limita ou favorece a riqueza de espécies em diferentes ecossistemas.

Neste problema, analisamos a relação entre a precipitação média anual (variável explicativa x, em milímetros/ano) e a riqueza de espécies de árvores (variável resposta y, número de espécies identificadas por parcela de 1 hectare) em uma amostragem de 35 parcelas florestais ao longo de um gradiente pluviométrico.

Os dados estão disponíveis na planilha *ecologia* do arquivo *dados_rls_20262.xlsx*.

▪

Exemplo 6.

Na Farmacologia Clínica e na Farmacococinética/Farmacodinâmica (PK/PD), o estudo da relação entre a exposição sistêmica a um medicamento e a magnitude da sua resposta biológica é essencial para definir faixas terapêuticas seguras e eficazes.

Neste problema, analisamos o impacto da concentração plasmática de um fármaco anti-hipertensivo (variável explicativa x, em μg/mL) sobre a redução da pressão arterial sistólica (variável dependente y, em mmHg) mensurada em uma amostra de 35 pacientes em um ensaio clínico.

▪