Estatística Básica e Introdução ao R

Profª. Dra. Natalia Giordani

4. Análise de Regressão

- Técnica estatística utilizada para modelar a relação entre uma variável dependente (ou resposta ou Y) e uma ou mais variáveis independentes (ou preditoras ou covariáveis ou X)
- Exemplos
  - Preço de uma casa pode ser predito utilizando a relação entre preço e o n. de quartos, n. de banheiros, área (m^2), localização (região)...
  - Custo médico anual pode ser predito por uma seguradora considerando a relação entre custo e idade do segurado, IMC, histórico de doenças, n. de visitas ao médico nos últimos 3 meses...

4.1 Regressão Linear Simples

- O que é?
  - Uma única variável preditora (X) é utilizada para predizer a variável resposta (ou desfecho ou Y) **contínua** de interesse
- História...
  - Método desenvolvido por Francis Galton no fim do século 19
  - Em estudo da relação entre alturas de pais e filhos notou que a altura dos filhos tanto de pais altos quanto de pais baixos tendiam a regredir para a média do grupo
  - Desenvolveu uma descrição matemática dessa tendência de regressão – precursor dos modelos de regressão que usamos
  - Regressão: termo utilizado para descrever relações estatísticas entre variáveis

4.1 Regressão Linear Simples

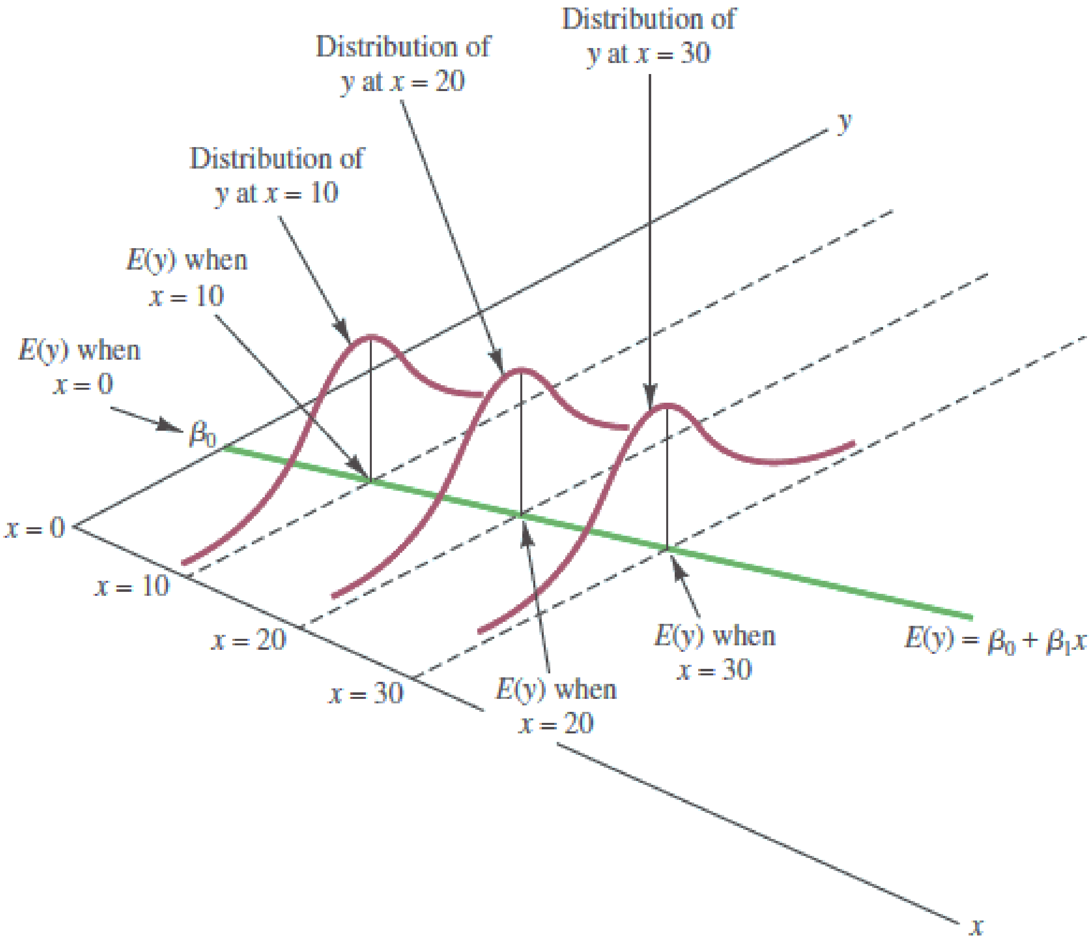
- Conceitos Básicos
  - Um modelo de regressão é composto por **dois ingredientes essenciais** de uma relação estatística
    - Tendência da variável resposta **Y variar de acordo** com a variável preditora **X** de forma sistemática
    - **Dispersão** dos pontos **em torno da curva da relação estatística**
  - Esses dois ingredientes são incorporados em um modelo de regressão ao postular que
    - Há uma distribuição de probabilidades de Y para cada nível de X
    - As médias dessas distribuições de probabilidades variam de forma sistemática com X
  - Declaração formal do modelo
    - $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$

4.1 Regressão Linear Simples

- Conceitos Básicos
  - Um modelo de regressão é composto por dois ingredientes essenciais de uma relação estatística
    - Tendência da variável resposta Y variar de acordo com a variável preditora X de forma sistemática
    - Dispersão dos pontos em torno da curva da relação estatística
  - Esses dois ingredientes são incorporados em um modelo de regressão ao postular que
    - Há uma distribuição de probabilidades de Y para cada nível de X
    - As médias dessas distribuições de probabilidades variam de forma sistemática com X
  - Declaração formal do modelo
    - $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad \Rightarrow \quad E(Y_i) = \beta_0 + \beta_1 X_i$

4.1 Regressão Linear Simples

- Representação gráfica do modelo



- O valor esperado para a variável resposta, E(y), muda de acordo com o valor da variável preditora, x.

- Independente do valor de x, a distribuição de probabilidade do erro e, consequentemente a distribuição de probabilidade de y, seguem uma distribuição normal com a mesma variância.

- O valor do erro em qualquer ponto pode ser positivo ou negativo – vai depender do valor observado de y ser maior ou menor que o valor esperado E(y).

4.1 Regressão Linear Simples

- Conceitos Básicos
  - Modelo de regressão **linear** simples
    - $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$
  - E se...
    - $Y_i = \beta_0 + \exp(\beta_1 X_i) + \varepsilon_i$ ⟶ Modelo **não linear**
    - $Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \varepsilon_i$ ⟶ Modelo **linear** de regressão **polinomial**

4.1 Regressão Linear Simples

- Conceitos Básicos
  - Estimação dos coeficientes: método de mínimos quadrados
  - Modelo ajustado... necessário avaliar a qualidade do ajuste

    1. **Coeficiente de determinação (R^2)**

       - $R^2 = 1 - \frac{SQRes}{SQTotal} = 1 - \frac{\sum(y-\hat{y})^2}{\sum(y-\bar{y})^2}$
       - Mede a porcentagem da variabilidade de Y que é explicada pelo modelo de regressão

4.1 Regressão Linear Simples

- Conceitos Básicos
  - Modelo ajustado... necessário avaliar a qualidade do ajuste

2. **Gráficos de resíduos e resíduos padronizados**

$$Resíduo = y - \hat{y} \sim \text{Normal } (0, \sigma^2)$$

3. **Gráficos da distância de Cook**
   - Mede a mudança nos valores preditos pelo modelo quando eliminamos uma das observações
   - Destaca pontos (influentes ou alavanca) que podem afetar de forma relevante as estimativas dos parâmetros

4. **Observações devem ser independentes**

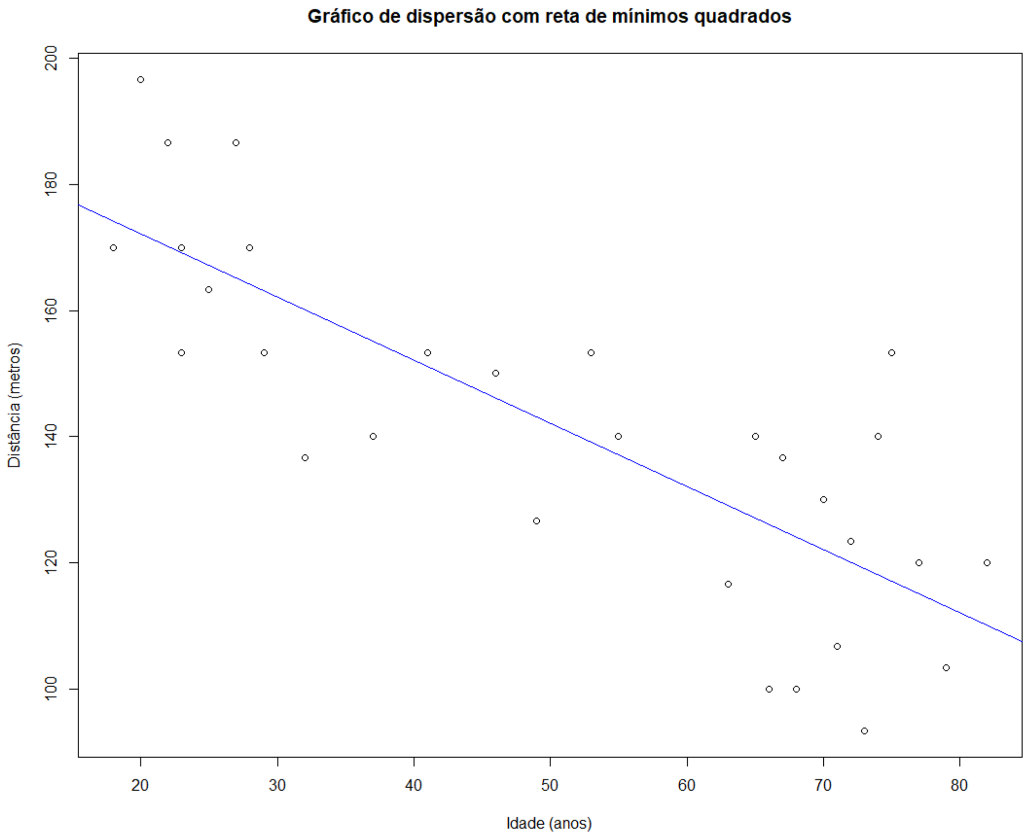
4.1 Regressão Linear Simples

- **Exemplo didático 1:** foi realizado um estudo com o objetivo de avaliar como a distância com que os motoristas conseguem distinguir determinado objeto varia com a idade.

Indivíduo	Idade (anos)	Distância (m)
1	18	170
2	20	197
3	22	187
4	23	170
5	23	153
...	...	...
30	82	120

4.1 Regressão Linear Simples

- **Exemplo didático 1:** foi realizado um estudo com o objetivo de avaliar como a distância com que os motoristas conseguem distinguir determinado objeto varia com a idade.



- $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$
- $distância_i = \beta_0 + \beta_1 Idade_i + \varepsilon_i$

4.1 Regressão Linear Simples

- **Exemplo didático 1:** foi realizado um estudo com o objetivo de avaliar como a distância com que os motoristas conseguem distinguir determinado objeto varia com a idade.

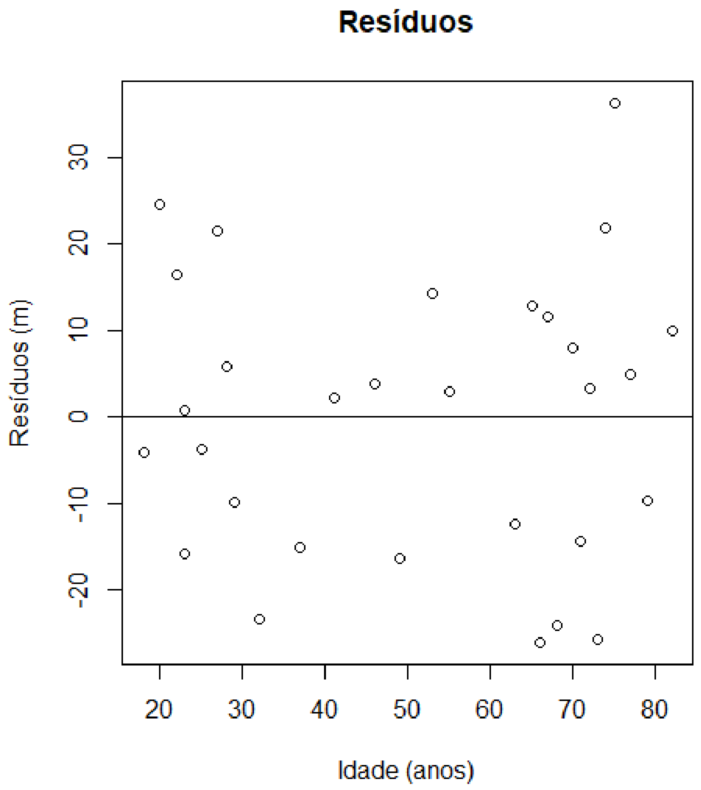
```
> ajuste2 <- stats::lm(distancia ~ idade, data = dados)
> summary(ajuste2)

Call:
stats::lm(formula = distancia ~ idade, data = dados)

Residuals:
    Min      1Q  Median      3Q     Max 
-26.077 -13.903   2.549  11.184  36.277 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 192.2273     7.8236  24.570  < 2e-16 ***
idade        -1.0023     0.1414  -7.086 1.04e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 16.59 on 28 degrees of freedom
Multiple R-squared:  0.642,	Adjusted R-squared:  0.6292 
F-statistic: 50.21 on 1 and 28 DF,  p-value: 1.041e-07
```

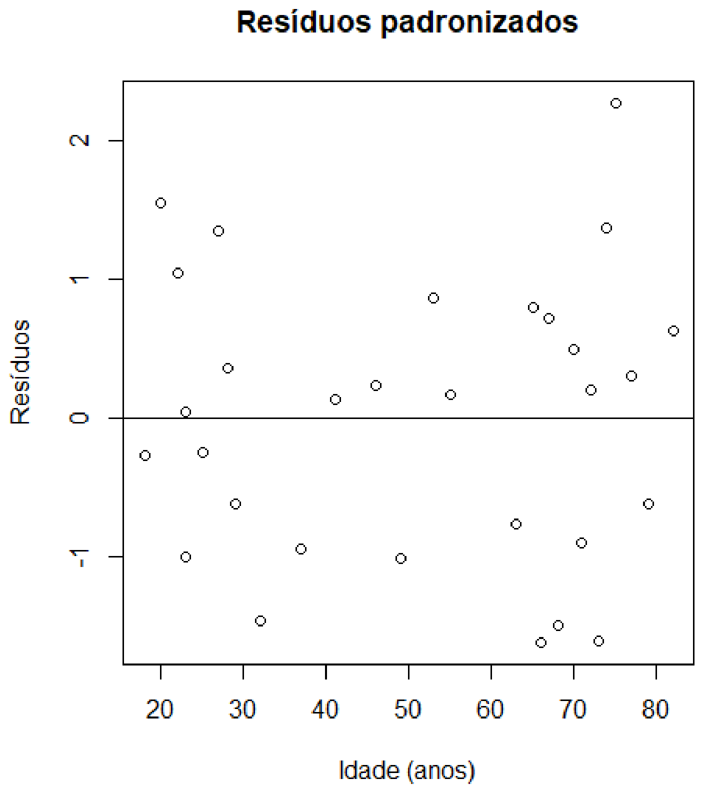
- $distância_i = 192{,}2273 \; - 1{,}0023 \; Idade_i + \varepsilon_i$

- 192,2273 = distância esperada para que motorista de 0 anos consigam distinguir determinado objeto
- -1,0023 = redução da distância esperada para cada ano adicional na idade
- 64,2% da variação total da distância é explicada pelo modelo de regressão

4.1 Regressão Linear Simples

- **Exemplo didático 1:** foi realizado um estudo com o objetivo de avaliar como a distância com que os motoristas conseguem distinguir determinado objeto varia com a idade.

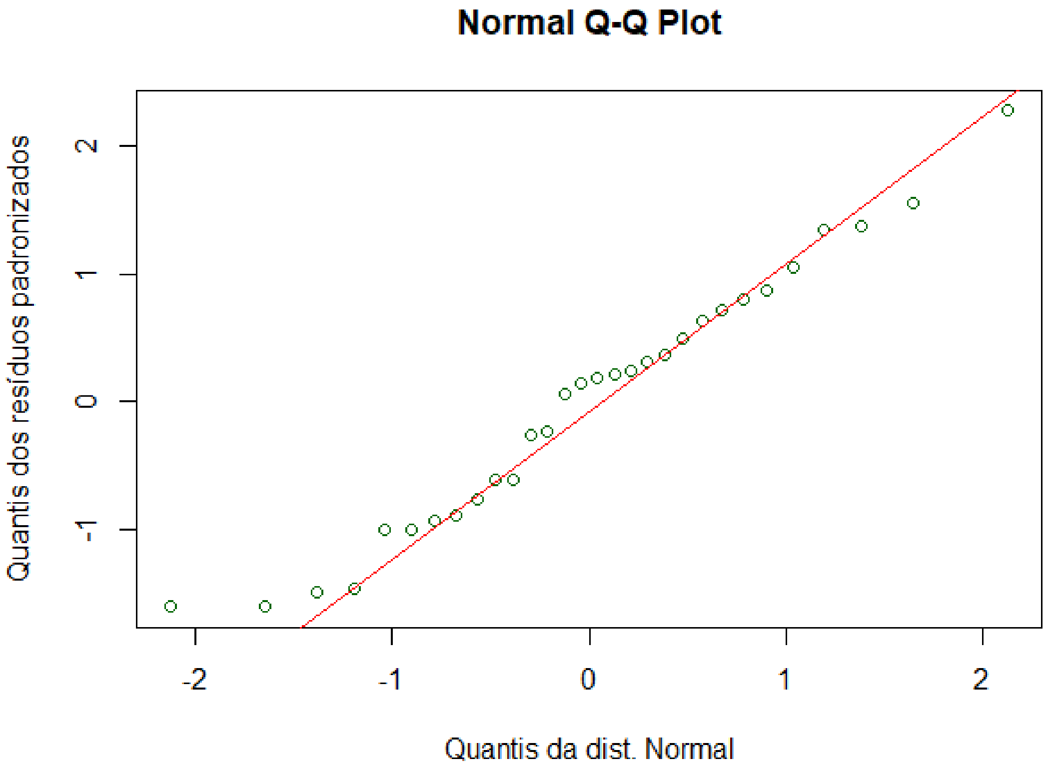




- Distribuídos sem padrão sistemático
- Variabilidade razoavelmente uniforme ao longo dos diferentes valores de x
  - Sugestivo de que suposição de homocedasticidade está atendida

4.1 Regressão Linear Simples

- **Exemplo didático 1:** foi realizado um estudo com o objetivo de avaliar como a distância com que os motoristas conseguem distinguir determinado objeto varia com a idade.



- Suposição de normalidade ✓

4.1 Regressão Linear Simples

- **Exemplo didático 2:** dados de concentração atmosférica do poluente monóxido de carbono, CO, no dia (variável = tempo) i, na cidade de São Paulo, entre 01/jan e 30/abr de 1991.

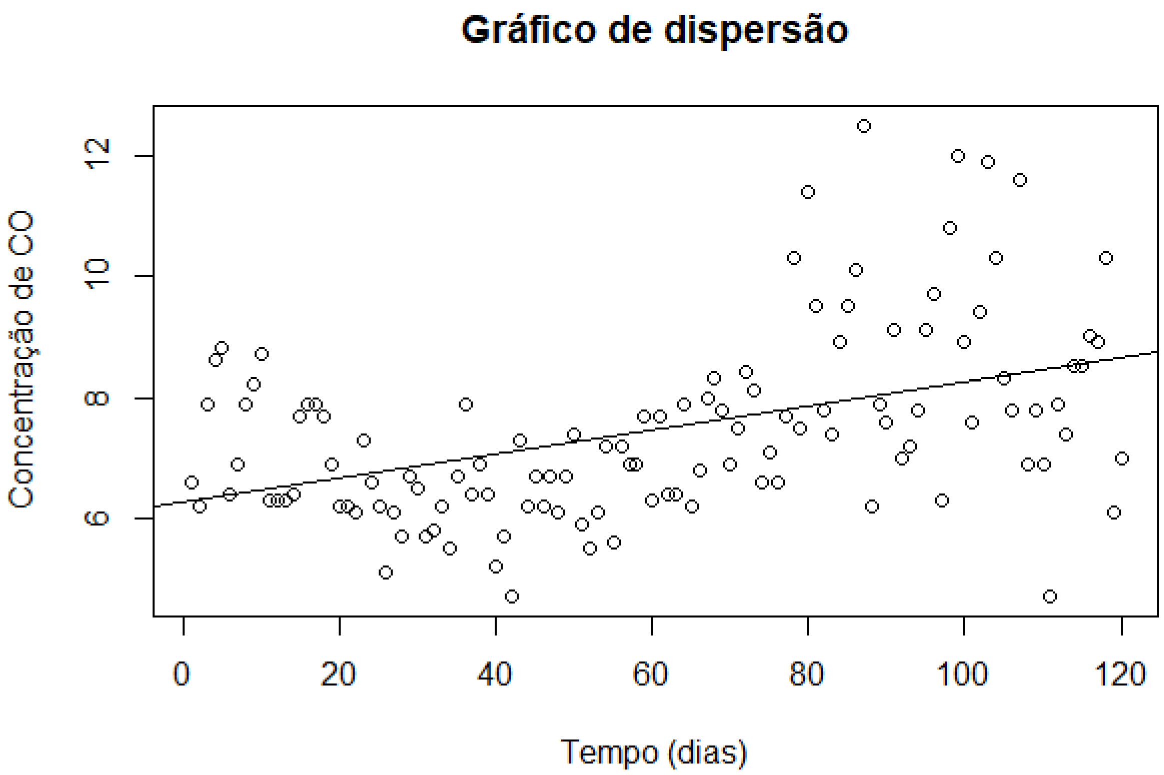
Tempo	CO
1	6,6
...	...
120	7,0

```
> ajuste_ex2 <- stats::lm(CO ~ tempo, data = ex2)
> summary(ajuste_ex2)

Call:
stats::lm(formula = CO ~ tempo, data = ex2)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.7655 -0.9157 -0.1788  0.6613  4.5104 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 6.264608   0.254847  24.582  < 2e-16 ***
tempo       0.019827   0.003656   5.424 3.15e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.387 on 118 degrees of freedom
Multiple R-squared:  0.1996,	Adjusted R-squared:  0.1928 
F-statistic: 29.42 on 1 and 118 DF,  p-value: 3.148e-07
```

Modelo RLS explica muito pouco da variabilidade dos dados!

4.1 Regressão Linear Simples

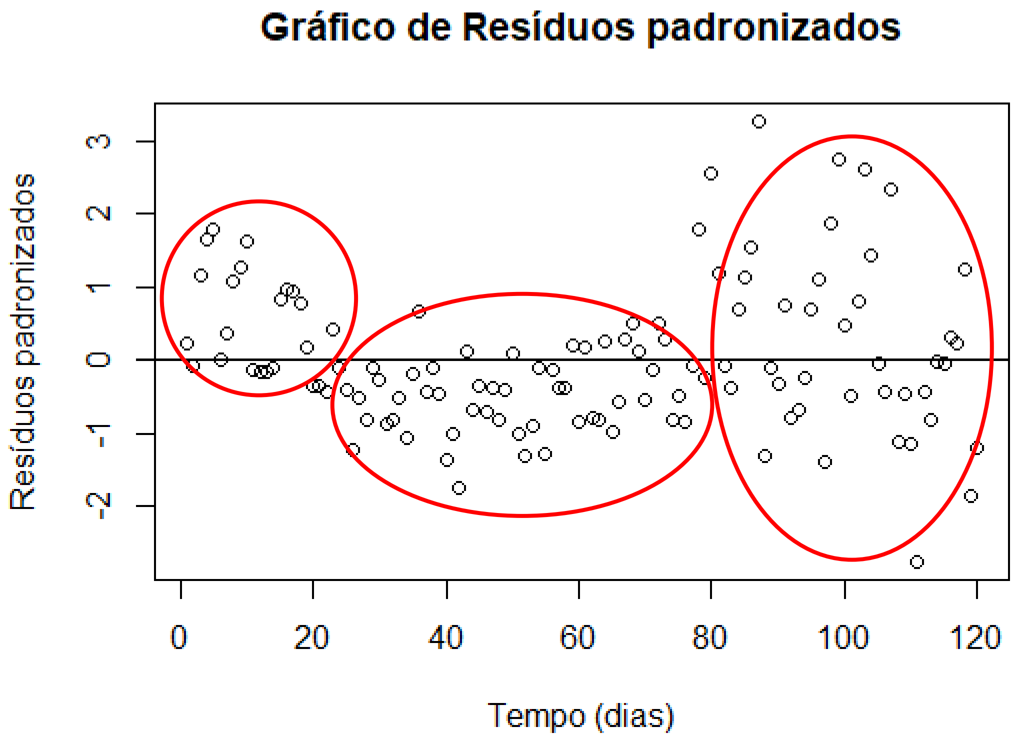
- **Exemplo didático 2:** dados de concentração atmosférica dos poluentes ozônio O3 e monóxido de carbono, além da temperatura média e umidade na cidade de São Paulo entre 01/jan e 30/abr de 1991.



- Relação entre tempo e CO é representada por um reta?

4.1 Regressão Linear Simples

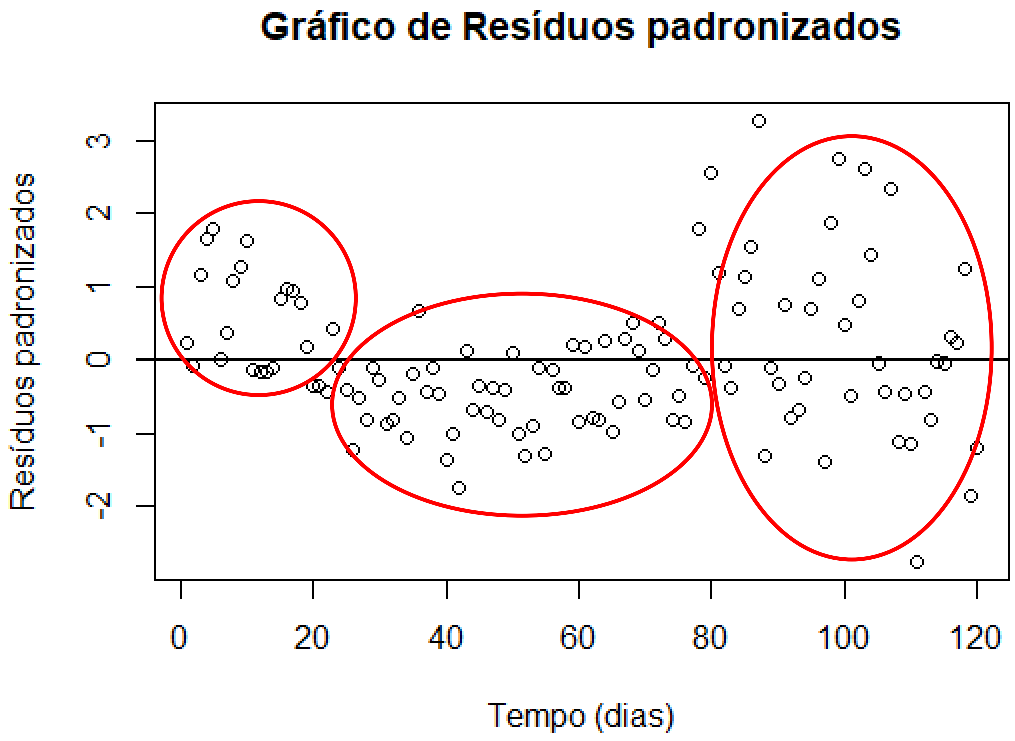
- **Exemplo didático 2:** dados de concentração atmosférica dos poluentes ozônio O3 e monóxido de carbono, além da temperatura média e umidade na cidade de São Paulo entre 01/jan e 30/abr de 1991.



- Padrão na distribuição dos resíduos!
- Dispersão varia com tempo!

4.1 Regressão Linear Simples

- **Exemplo didático 2:** dados de concentração atmosférica dos poluentes ozônio O3 e monóxido de carbono, além da temperatura média e umidade na cidade de São Paulo entre 01/jan e 30/abr de 1991.

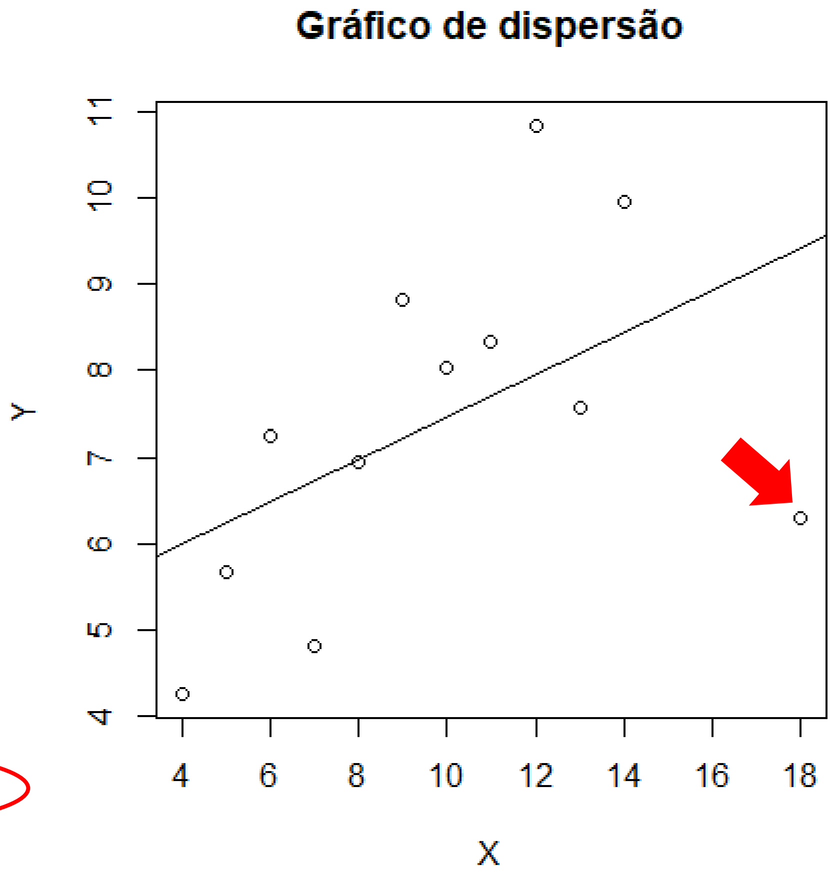


- Padrão na distribuição dos resíduos!
- Dispersão varia com tempo!
- O que isso significa?
  - Esse modelo não é adequado!
  - Abordagem de séries temporais faz mais sentido!

4.1 Regressão Linear Simples

- **Exemplo didático 3:**

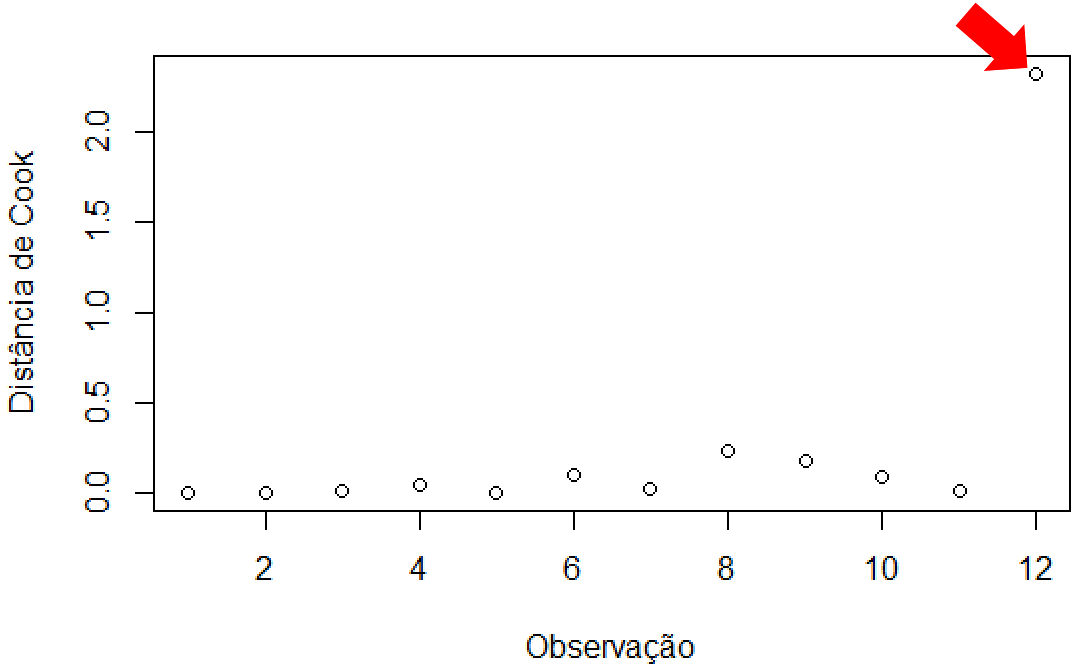
Observação	X	Y
1	4	4,26
2	5	5,68
3	6	7,24
4	7	4,82
5	8	6,95
6	9	8,81
7	10	8,04
8	11	8,33
9	12	10,84
10	13	7,58
11	14	9,96
12	18	6,31



4.1 Regressão Linear Simples

- **Exemplo didático 3:**

Observação	X	Y
1	4	4,26
2	5	5,68
3	6	7,24
4	7	4,82
5	8	6,95
6	9	8,81
7	10	8,04
8	11	8,33
9	12	10,84
10	13	7,58
11	14	9,96
12	18	6,31



4.1 Regressão Linear Simples

- **Exemplo didático 3:**

Sem excluir a observação 12:

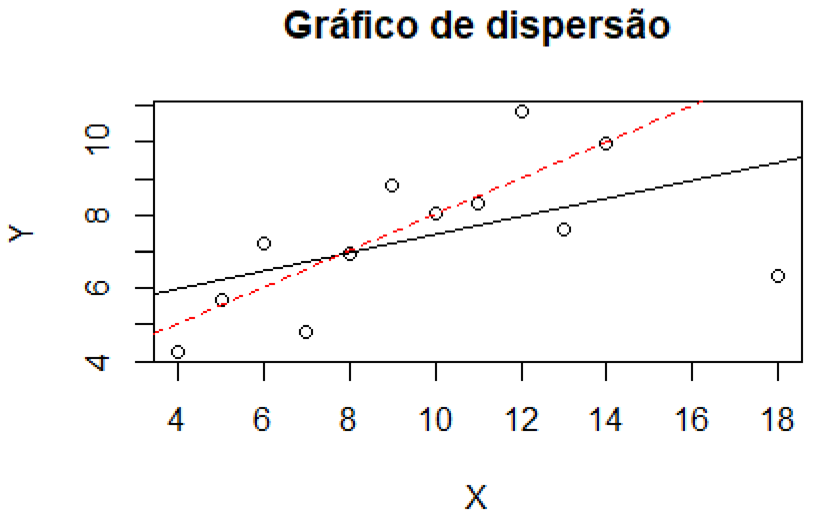
```
> summary(ajuste_ex4)

Call:
stats::lm(formula = y ~ x, data = ex4)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.1149 -0.8969  0.2773  0.9475  2.8866 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)   
(Intercept)   5.0106     1.3738   3.647  0.00448 **
x             0.2452     0.1307   1.876  0.09014 . 
```

Intercepto: 5,01 (1,37)
Inclinação: 0,25 (0,13)

Excluindo a observação 12:

```
> summary(ajuste_ex4_adj)

Call:
stats::lm(formula = y ~ x, data = ex4_adj)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.92127 -0.45577 -0.04136  0.70941  1.83882 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)   
(Intercept)   3.0001     1.1247   2.667  0.02573 * 
x             0.5001     0.1179   4.241  0.00217 **
---
```

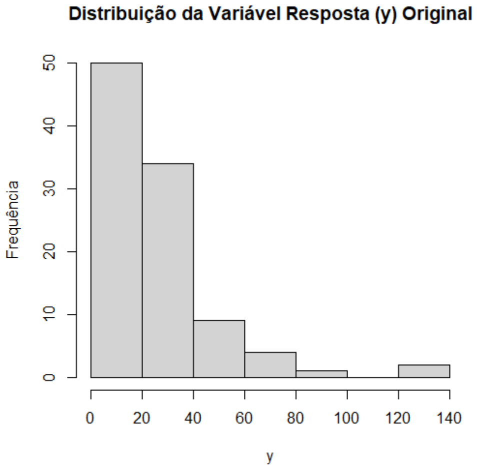
Intercepto: 3,00 (1,12)
Inclinação: 0, 50 (0,12)



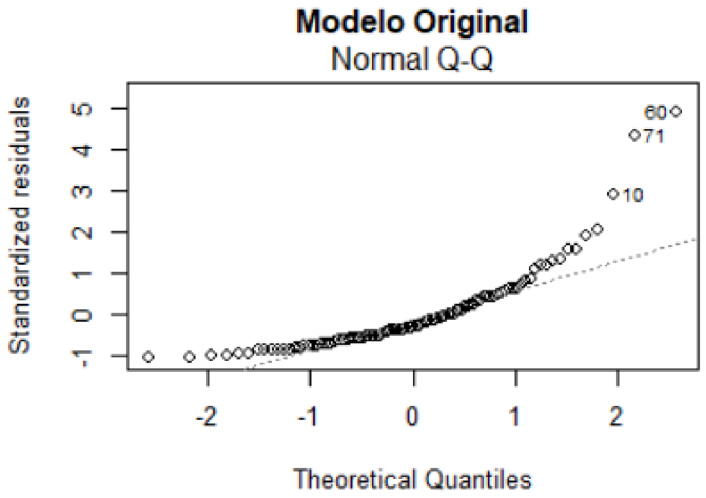
4.1 Regressão Linear Simples

- **Exemplo didático 4:**

Observação	X	Y
1	71,98	11,75
2	63,12	11,42
3	47,35	9,93
...	...	...
100	62,50	15,39



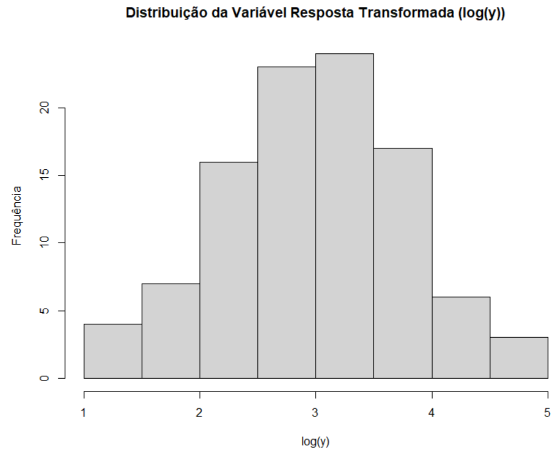

```
> summary(modelo_original)

Call:
lm(formula = y ~ x, data = df)

Residuals:
    Min      1Q  Median      3Q     Max 
-23.783 -14.143  -7.098   7.871 109.817 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)  
(Intercept)  32.7881    12.5531   2.612   0.0104 *
x            -0.1275     0.2411  -0.529   0.5981  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 22.79 on 98 degrees of freedom
Multiple R-squared:  0.002845,	Adjusted R-squared:  -0.00733 
F-statistic: 0.2796 on 1 and 98 DF,  p-value: 0.5981
```



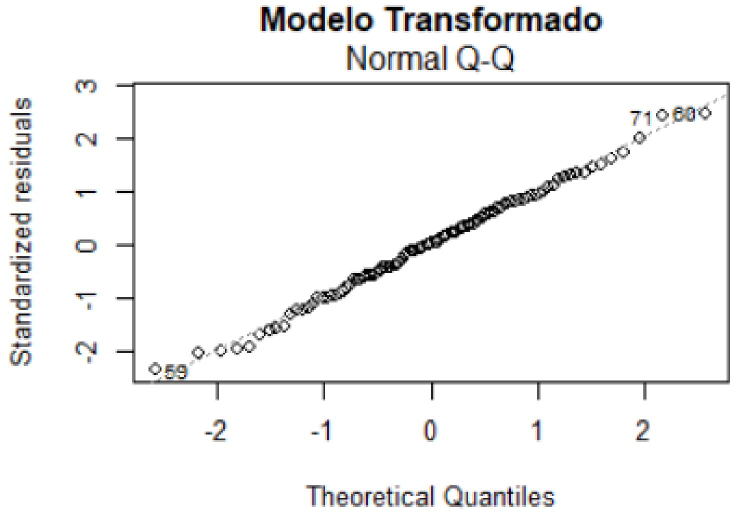
4.1 Regressão Linear Simples

- **Exemplo didático 4:**

Observação	X	Y	log(Y)
1	71,98	11,75	2,46
2	63,12	11,42	2,43
3	47,35	9,93	2,29
...	...	...	...
100	62,50	15,39	2,73




```
> summary(modelo_transformado)

Call:
lm(formula = log_y ~ x, data = df)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.8309 -0.4903  0.0060  0.5559  1.8949 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  3.161259   0.430974   7.335 6.46e-11 ***
x           -0.003680   0.008277  -0.445    0.658    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7823 on 98 degrees of freedom
Multiple R-squared:  0.002013,	Adjusted R-squared:  -0.008171 
F-statistic: 0.1977 on 1 and 98 DF,  p-value: 0.6576
```



4.1 Regressão Linear Simples

- **Exemplo didático 4:**

- **Modelo original**
  - $y_i = 32{,}8 \ - 0{,}13\ x_i + \varepsilon_i$
  - Beta0 = Valor esperado de y quando x é zero
  - Beta1 = Mudança esperada em y para cada unidade de mudança em x
- **Modelo transformado**
  - $\log(y_i) = 3{,}16 \ - 0{,}004\ x_i + \varepsilon_i$
  - Beta0 = Valor esperado de log(y) quando x é zero
  - Beta1 = Mudança esperada em log(y) para cada unidade de mudança em x.
  - <u>Destransformação</u> permite interpretar os coeficientes na escala original
    - $E(y_i) = \exp(3{,}16) \ - \exp(0{,}004)\ x_i$

4.2 Regressão Linear Múltipla

- O que é?
  - *p* variáveis preditoras (X) são utilizadas para predizer a variável resposta contínua (Y)
- Declaração formal do modelo
  - $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + e_i \longrightarrow i = 1, \ldots, n$
  - $e_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$
  - Via estimadores de mínimos quadrados:
    - $\widehat{y_i} = \widehat{\beta_0} + \widehat{\beta_1} x_{i1} + \widehat{\beta_2} x_{i2} + \cdots + \widehat{\beta_p} x_{ip}$
  - Resíduos
    - $e_i = y_i - \widehat{y_i}$

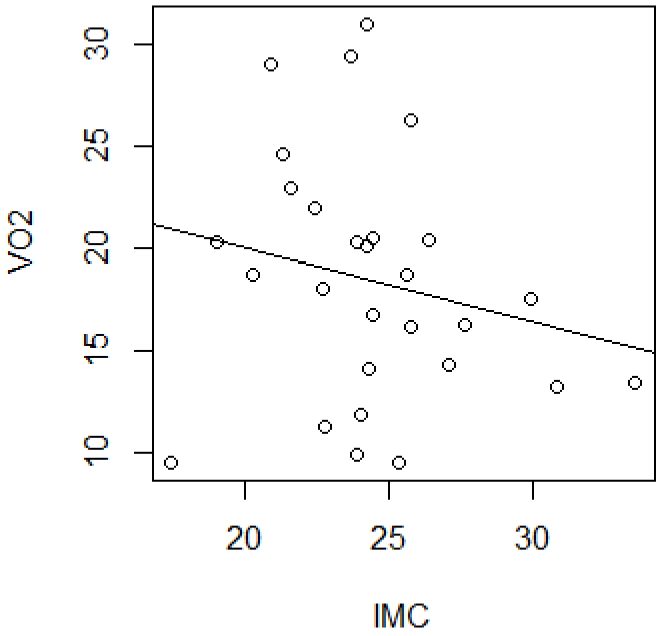
4.2 Regressão Linear Múltipla

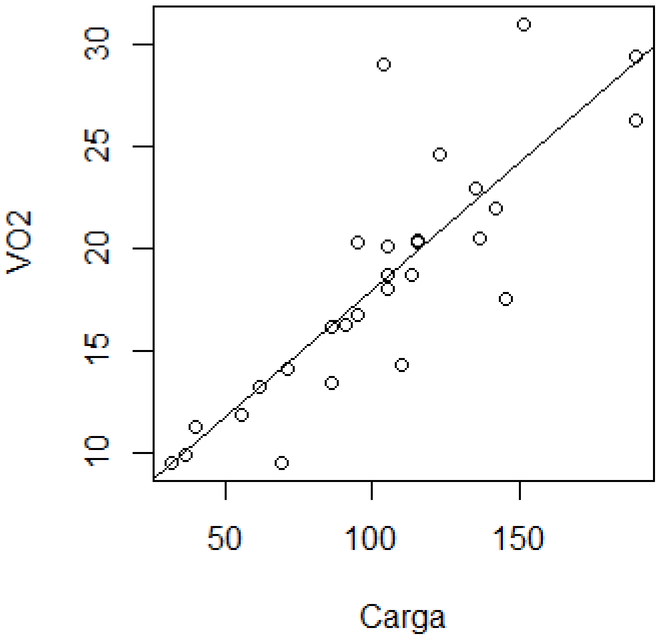
- Medidas de diagnóstico
  - Ver Regressão Linear Simples
- Observação relacionada ao coeficiente de determinação (R^2)
  - $R^2 = SQReg/SQTotal$
  - % da variabilidade de Y explicada pelo modelo
  - $SQReg$ aumenta com a inclusão de mais variáveis explicativas, por isso, para comparação de modelos com números diferentes de covariáveis, utilizar o R^2_{ajust}
- Cuidado em relação a multicolinearidade!

4.2 Regressão Linear Múltipla

- **Exemplo didático 5:** estudo foi realizado com o objetivo de avaliar o efeito do IMC e da carga aplicada numa esteira ergométrica no consumo de oxigênio (VO_2) em determinada fase do exercício.

ID	VO2	IMC	Carga
1	14.1	24.32	71
2	16.3	27.68	91
3	9.9	23.93	37
4	9.5	17.50	32
...	...	...	...
28	31.0	24.24	151





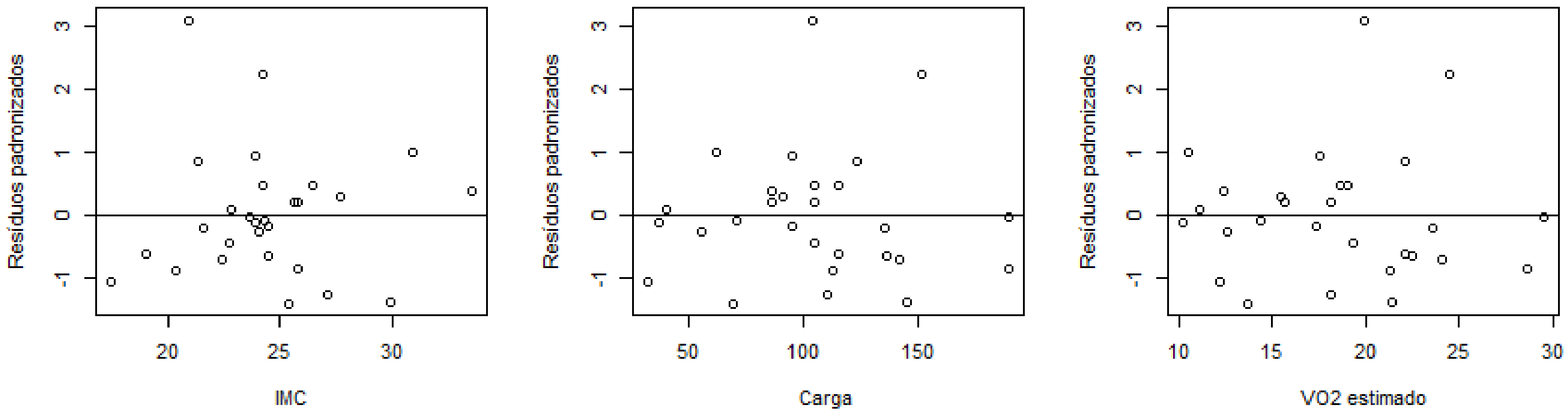
4.2 Regressão Linear Múltipla

- **Exemplo didático 5:** estudo foi realizado com o objetivo de avaliar o efeito do IMC e da carga aplicada numa esteira ergométrica no consumo de oxigênio (VO_2) em determinada fase do exercício.

```
> ajuste_ex5 <- stats::lm(VO2 ~ IMC + carga , data = ex5)
> summary(ajuste_ex5)

Call:
stats::lm(formula = VO2 ~ IMC + carga, data = ex5)

Residuals:
    Min      1Q  Median      3Q     Max
-4.1835 -2.0161 -0.2929  1.0642  9.0869

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 15.44597    4.45469   3.467  0.00192 **
IMC         -0.41311    0.17179  -2.405  0.02391 *
carga        0.12617    0.01465   8.614 5.95e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.058 on 25 degrees of freedom
Multiple R-squared:  0.759,	Adjusted R-squared:  0.7397
F-statistic: 39.36 on 2 and 25 DF,  p-value: 1.889e-08
```

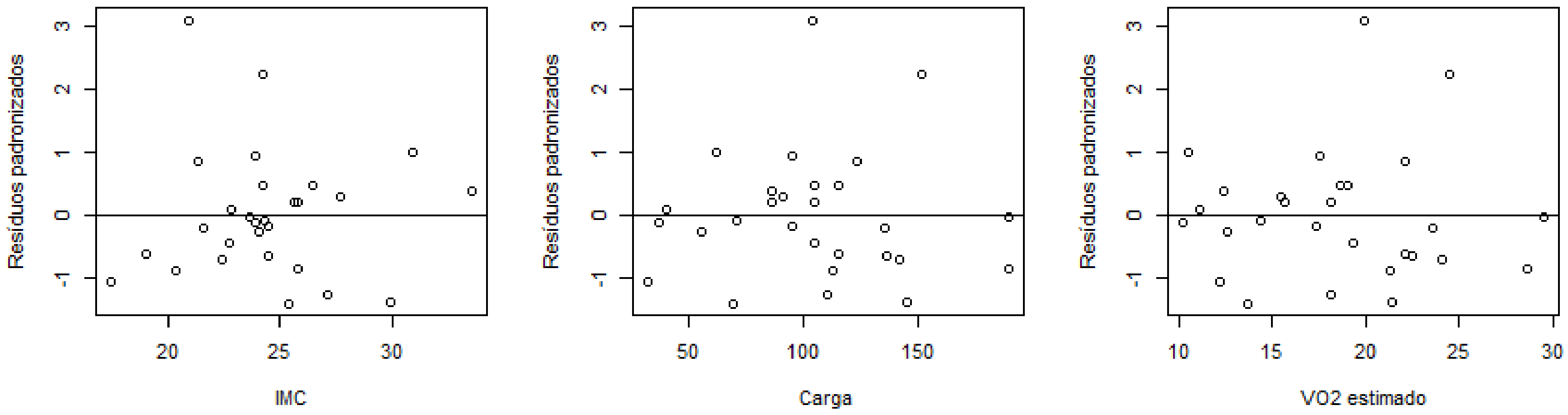
- $\widehat{VO2_i} = 15{,}45\ - 0{,}41\ IMC_i + 0{,}13 carga_i$
- $\widehat{\beta_1}$: Valor esperado de VO_2, mantendo fixa a carga da esteira, reduz em 0,41 unidades para cada aumento de uma unidade de IMC
- $\widehat{\beta_2}$: Valor esperado de VO_2 para indivíduos com o mesmo IMC aumenta em 0,13 unidades para cada aumento de uma unidade da carga na esteira
- 74% da variabilidade de VO_2 é explicada pelo modelo

4.2 Regressão Linear Múltipla



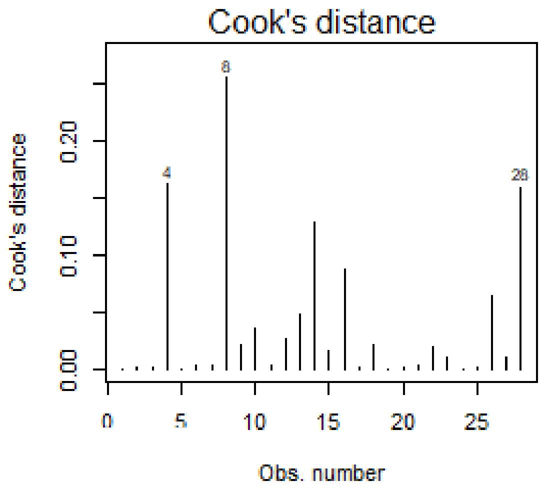
- Gráficos dos resíduos
  - Homocedasticidade ✓

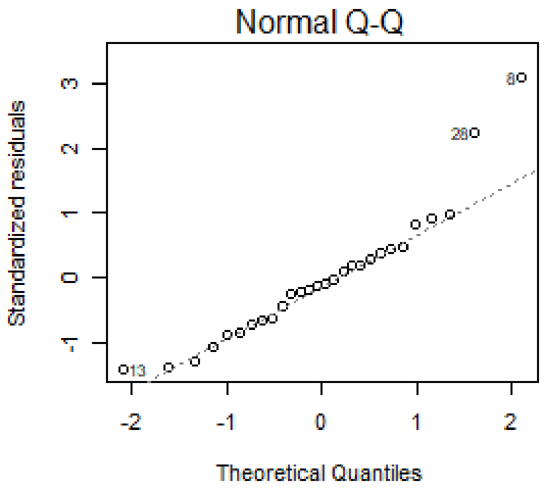
4.2 Regressão Linear Múltipla



- Gráficos dos resíduos
  - Homocedasticidade

DECISÃO: REMOVER 3 PONTOS DE INFLUÊNCIA





- Gráfico da distância de Cook
  - 3 pontos atípicos
  - Resíduos associados a distâncias de Cook > 4/n

- QQPlot
  - 2 pontos deixam em dúvida suposição de normalidade dos resíduos

4.2 Regressão Linear Múltipla

- **Exemplo didático 5:** estudo foi realizado com o objetivo de avaliar o efeito do IMC e da carga aplicada numa esteira ergométrica no consumo de oxigênio (VO_2) em determinada fase do exercício.

```
> ajuste_ex5_adj <- stats::lm(VO2 ~ IMC + carga , data = ex5_adj)
> summary(ajuste_ex5_adj)

Call:
stats::lm(formula = VO2 ~ IMC + carga, data = ex5_adj)

Residuals:
    Min      1Q  Median      3Q     Max 
-4.1642 -0.8579 -0.1169  1.0763  3.4125 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 14.89386    3.47094   4.291 0.000296 ***
IMC         -0.35634    0.12606  -2.827 0.009822 ** 
carga        0.11304    0.01052  10.743 3.23e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.987 on 22 degrees of freedom
Multiple R-squared:  0.8581,	Adjusted R-squared:  0.8452 
F-statistic:  66.5 on 2 and 22 DF,  p-value: 4.708e-10
```

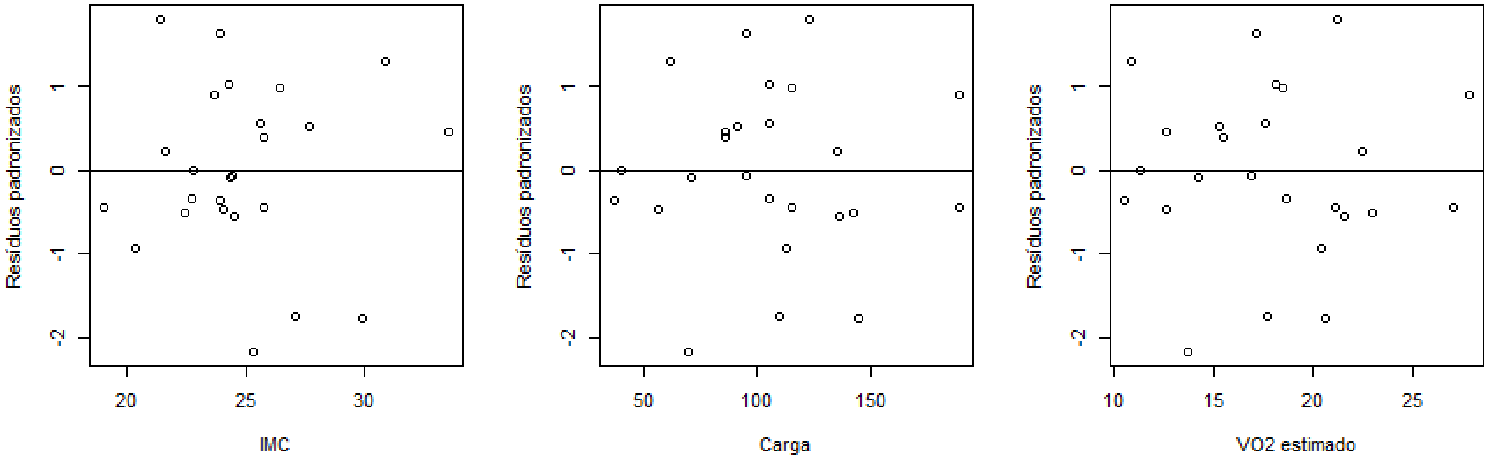
- $\widehat{VO2}_i = 14{,}9 - 0{,}36\, IMC_i + 0{,}11 carga_i$

- $\widehat{\beta_1}$: Valor esperado de VO_2, mantendo fixa a carga da esteira, reduz em 0,36 unidades para cada aumento de uma unidade de IMC

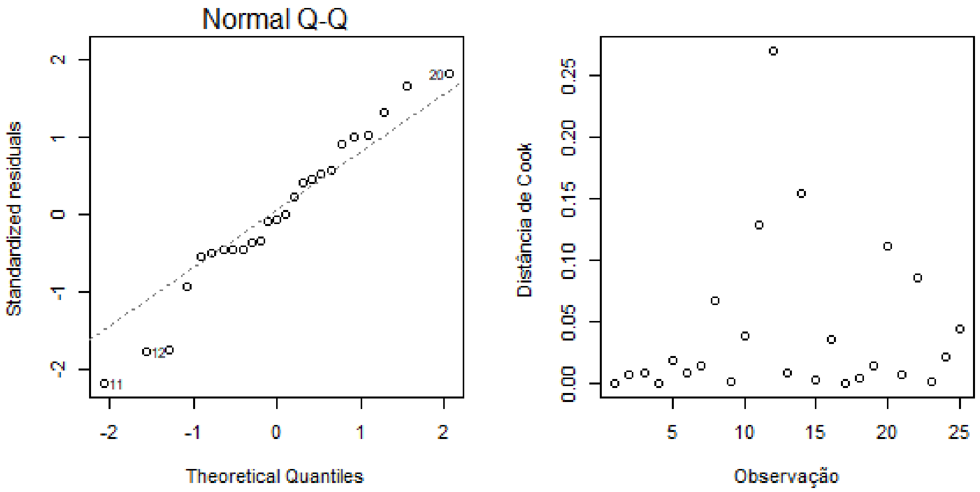
- $\widehat{\beta_2}$: Valor esperado de VO_2 para indivíduos com o mesmo IMC aumenta em 0,11 unidades para cada aumento de uma unidade da carga na esteira

- 84,5% da variabilidade de VO_2 é explicada pelo modelo

4.2 Regressão Linear Múltipla



- Gráficos dos resíduos
  - Homocedasticidade



- QQPlot
  - Suposição de normalidade dos resíduos

4.2 Regressão Linear Múltipla

- **Exemplo didático 5:** estudo foi realizado com o objetivo de avaliar o efeito do IMC e da carga aplicada numa esteira ergométrica no consumo de oxigênio (VO_2) em determinada fase do exercício.
  - Qual seria o consumo de oxigênio esperado para um indivíduo com IMC de 25 submetido a uma carga na esteira de 100?
    - $\widehat{VO2_i} = 14{,}9 \; - 0{,}36 \, IMC_i + 0{,}11 carga_i$
    - $\widehat{VO2_i} = 14{,}9 \; - (0{,}36 * 25) + (0{,}11 * 100) = 17{,}29$

```
> new.dat <- data.frame(IMC = 25, carga = 100)
> new.dat
  IMC carga
1  25   100
> predict(ajuste_ex5_adj, newdata = new.dat, interval = 'confidence')
      fit      lwr      upr
1 17.2897 16.45916 18.12023
```

Vamos praticar!

- **Objetivo inicial:** verificar como peso, potência e deslocamento impactam no consumo de combustível
  - Dados: mtcars
- Conceitos a desenvolver/discutir
  - Ajuste modelo
  - Avaliação ajuste
  - Interpretação
  - Predição

Referências

- Morretin, PA; Singer JDM. Estatística e Ciência de Dados. Rio de Janeiro: LTC Editora, 2022.

- Kutner, M; Wasserman, W; Neter J. Applied Linear Regression Models. McGraw-Hill/Irwin, 2004.