

# BIG DATA

Tema da aula  
**Regressão Linear: Análise de Resíduos**



## BUSINESS SCHOOL

Graduação, pós-graduação, MBA, Pós-MBA, Mestrado Profissional, Curso In Company e EAD



## CONSULTING

Consultoria personalizada que oferece soluções baseada em seu problema de negócio



## RESEARCH

Atualização dos conhecimentos e do material didático oferecidos nas atividades de ensino



Líder em Educação Executiva, referência de ensino nos cursos de graduação, pós-graduação e MBA, tendo excelência nos programas de educação. Uma das principais **escolas de negócio do mundo**, possuindo convênios internacionais com Universidades nos EUA, Europa e Ásia. +8.000 **projetos de consultorias** em organizações públicas e privadas.



Único curso de graduação em administração a receber as notas máximas



A primeira escola brasileira a ser finalista da maior competição de MBA do mundo



Única *Business School* brasileira a figurar no *ranking* LATAM



Signatária do Pacto Global da ONU



Membro fundador da ANAMBA - Associação Nacional MBAs



Credenciada pela AMBA - Association of MBAs



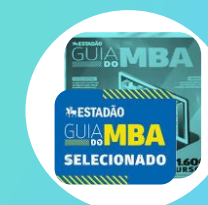
Credenciada ao Executive MBA Council



Filiada a AACSB - Association to Advance Collegiate Schools of Business



Filiada a EFMD - European Foundation for Management Development



Referência em cursos de MBA nas principais mídias de circulação

O **Laboratório de Análise de Dados** – LABDATA é um Centro de Excelência que atua nas áreas de ensino, pesquisa e consultoria em análise de informação utilizando técnicas de **Big Data, Analytics** e **Inteligência Artificial**.



O LABDATA é um dos pioneiros no lançamento dos cursos de *Big Data* e *Analytics* no Brasil

Os diretores foram professores de grandes especialistas do mercado

+10 anos de atuação

+1000 alunos formados

## Docentes

- Sólida formação acadêmica: doutores e mestres em sua maioria
- Larga experiência de mercado na resolução de *cases*
- Participação em Congressos Nacionais e Internacionais
- Professor assistente que acompanha o aluno durante todo o curso

## Estrutura

- 100% das aulas realizadas em laboratórios
- Computadores para uso individual durante as aulas
- 5 laboratórios de alta qualidade (investimento +R\$2MM)
- 2 Unidades próximas a estação de metrô (com estacionamento)

# CONTEÚDO PROGRAMÁTICO



ANÁLISE  
EXPLORATÓRIA

TÉCNICAS DE  
PROJEÇÃO

TÉCNICAS DE  
CLASSIFICAÇÃO

TÉCNICAS DE  
SEGMENTAÇÃO

TÉCNICAS DE  
ANALYTICS

LINGUAGEM



PYTHON



R

PROJETO



# Conteúdo da Aula

5

- 1. Suposições do Modelo
  - i. Normalidade
  - ii. Variabilidade constante
- 2. Métricas de qualidade do modelo
  - i. MAPE
  - ii. SSE
- 3. Código R



# 1. Suposições do modelo





# Análise dos resíduos

REGRESSÃO LINEAR | SUPOSIÇÕES DO MODELO

7

- Para verificar se o modelo ajustado é adequado, deve-se investigar se as suposições do modelo teórico adotado estão satisfeitas.
- Existem várias técnicas formais para conduzir essa análise, e nesta aula aprenderemos a realizá-la por métodos gráficos.

A Regressão Linear Múltipla (modelo teórico) é dada por:

$$Y = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p + \varepsilon \quad \text{com } \varepsilon \sim N(0, \sigma^2)$$

Y: variável dependente.

$X_1, \dots, X_p$ : variáveis independentes.

$\varepsilon$ : erro aleatório associado ao modelo.

## Suposições do modelo

1. A **média** dos **resíduos** é **zero**.
2. Os resíduos seguem uma **distribuição normal**.
3. Os **resíduos** têm a **variabilidade constante** em torno de x.
4.  $\varepsilon_i$  e  $\varepsilon_j$  são **não correlacionados**, para todo  $i \neq j$ .



# Suposições 1 e 2: Distribuição normal dos resíduos

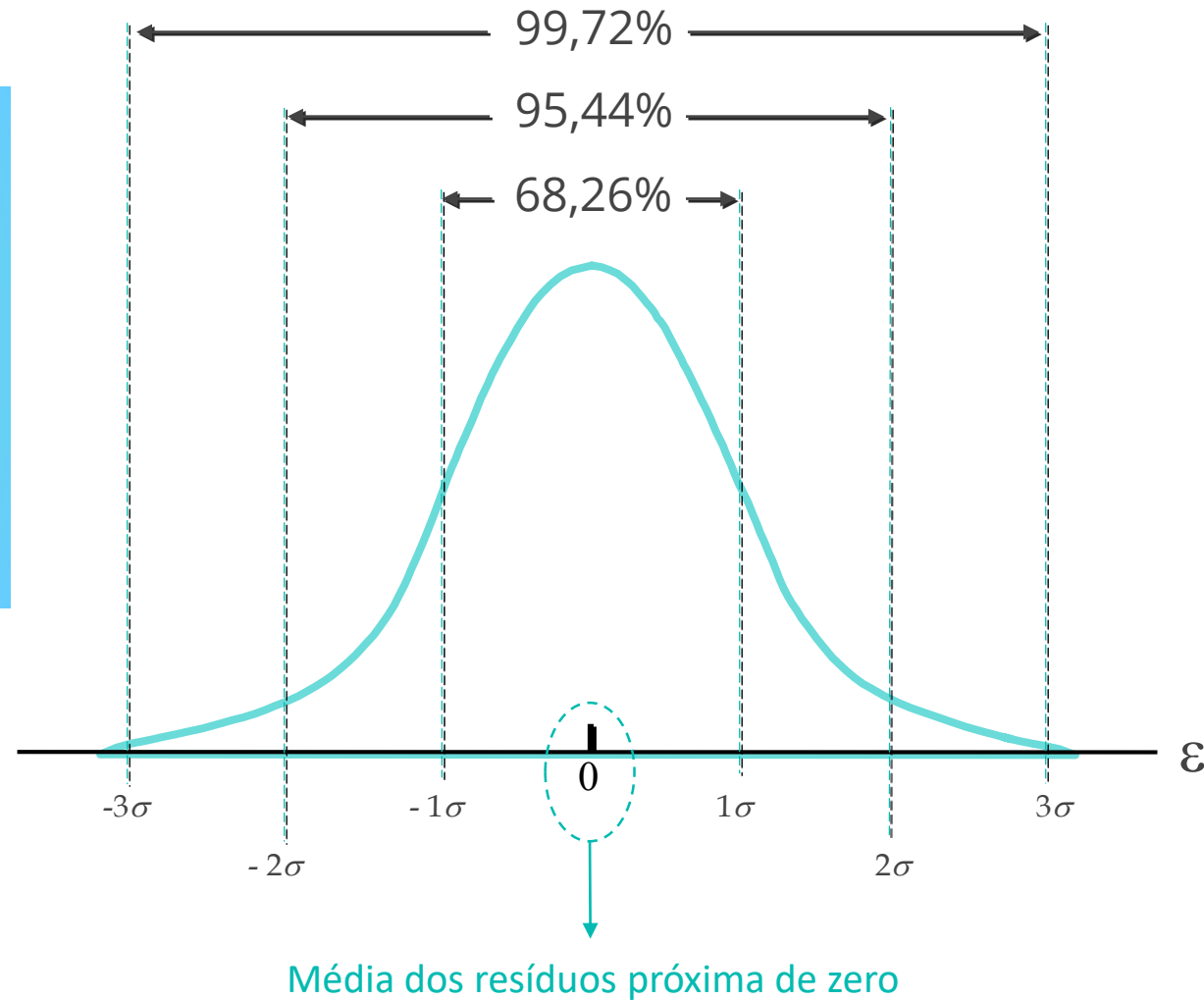
ANÁLISE DE RESÍDUOS | REGRESSÃO LINEAR

8

## Distribuição Normal (Gaussiana) dos resíduos

Fazer um histograma dos resíduos e verificar:

- Simetria
- Distribuição dos dados na proporção ao lado e ao redor de **zero**.

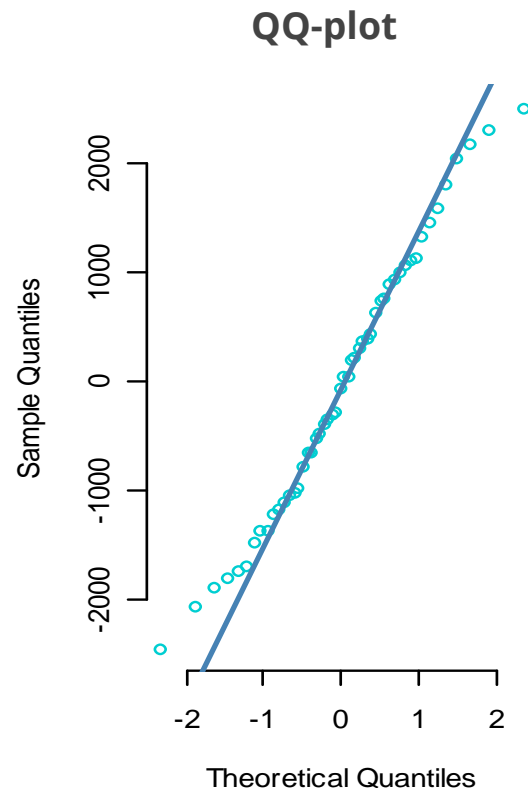
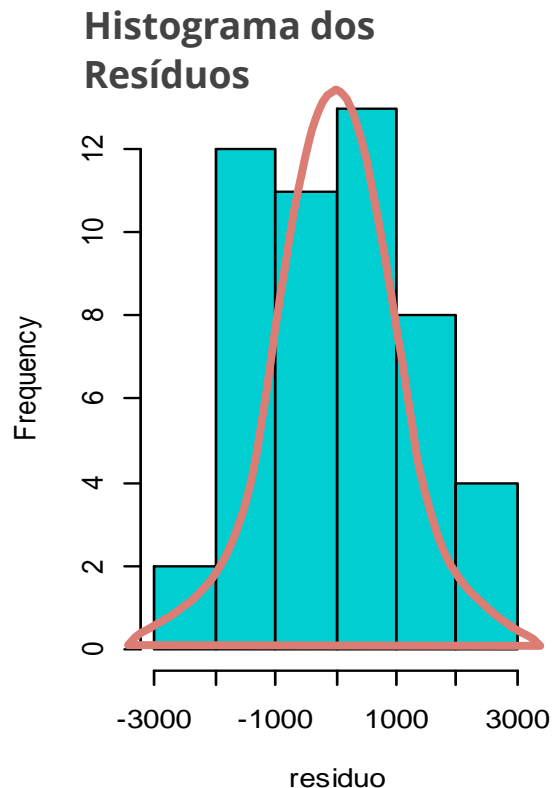




# Empiricamente: Distribuição normal dos resíduos

ANÁLISE DE RESÍDUOS | REGRESSÃO LINEAR

9



- **Histograma dos Resíduos:** mostra a distribuição dos resíduos. Espera-se valores centralizados e simétricos ao redor do **zero**.
- **QQ-plot (gráfico quantil x quantil):** no eixo X, estão dispostos os quantis teóricos da distribuição Normal Padrão; já no eixo y, os valores dos resíduos do modelo. Se os resíduos observados seguirem uma distribuição normal, os pontos devem se dispor ao longo da reta azul.

**Interpretação:** nos ajustes de modelos com dados reais, pode ocorrer uma leve fuga da normalidade. No caso acima, podemos concluir que os resíduos não violam a suposição de normalidade de forma grave.

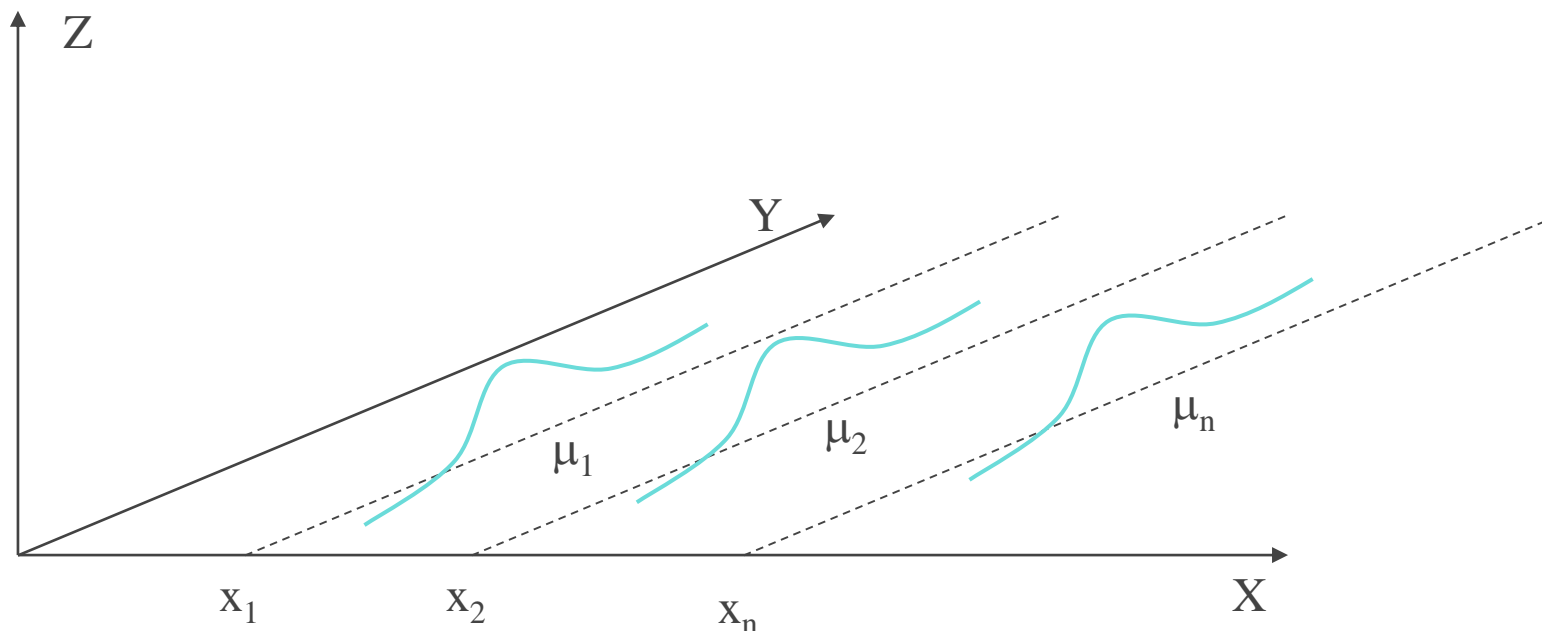


# Suposição 3: Variabilidade constante

ANÁLISE DE RESÍDUOS | REGRESSÃO LINEAR

10

A variância é constante ao longo dos possíveis valores da variável independente.



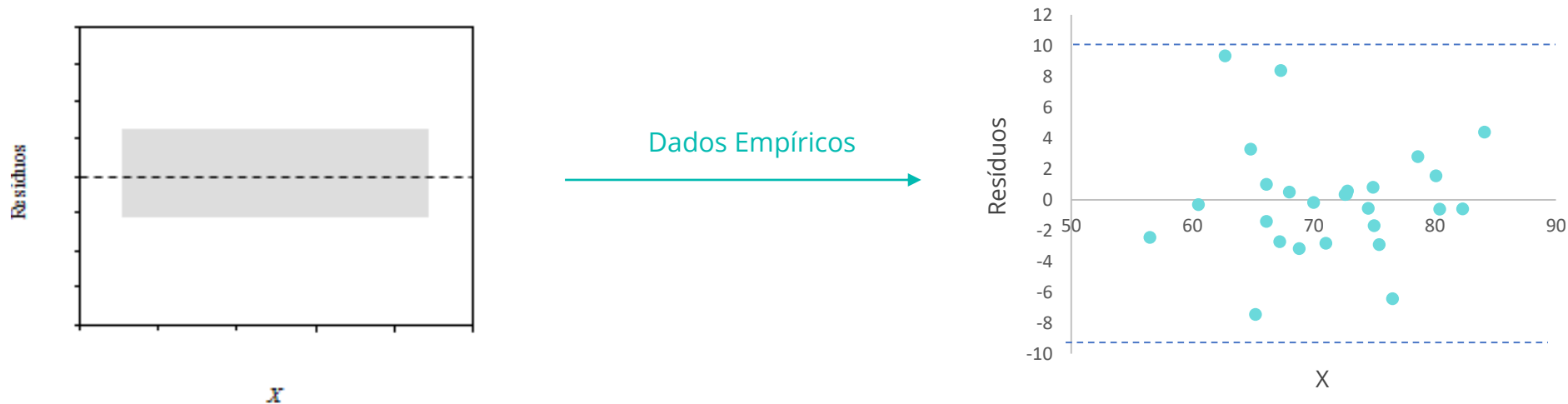
# Variabilidade constante: Suposições 3 e 4

ANÁLISE DE RESÍDUOS | REGRESSÃO LINEAR

11

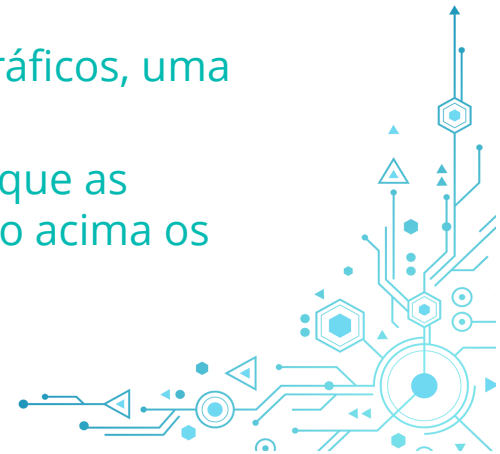
A variância é constante ao longo dos possíveis valores da variável independente.

O gráfico de resíduos *versus* a variável X deve fornecer uma nuvem horizontal de pontos distribuídos **aleatoriamente** ao redor do valor zero.



No caso da Regressão Linear Múltipla, como há  $p$  variáveis independentes, a fim de não gerar  $p$  gráficos, uma alternativa consiste em construir um gráfico cujo eixo x represente o valor predito pelo modelo.

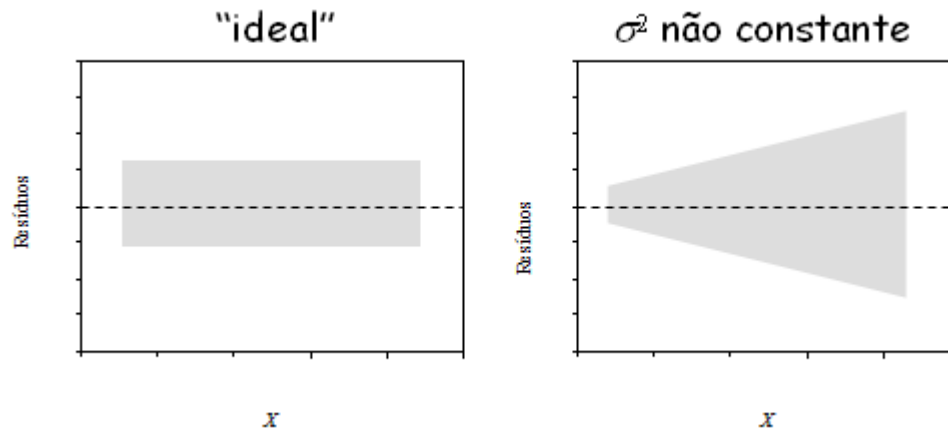
Para garantir a suposição 4, o próprio delineamento do problema deve ser realizado de tal forma que as unidades amostrais (observações analisadas) não tenham dependência entre si, ou seja, no gráfico acima os pontos devem se dispor de forma aleatória.



# Variabilidade constante: Suposições 3 e 4

ANÁLISE DE RESÍDUOS | REGRESSÃO LINEAR

12



## Possíveis transformações:

*raiz(y)*: recomendada quando a variância do erro cresce proporcionalmente a  $x$ .

*ln(y)*: recomendada quando o crescimento da variância do erro é mais acentuado do que o anterior; isto é, a variância cresce proporcionalmente a  $x^2$ .



# CASE: Predição *Startups*

CASE | FAZER ANÁLISE NO R

13

Um investidor deseja estimar o lucro de empresas *startups* de acordo com suas características, com objetivo de tomada de decisão de investimento baseada no lucro projetado da empresa. A base de dados histórica possui informações de investimento por área, gastos administrativos e região das empresas.



Investimento_PeD	Investimento_em_Mkt	Gastos_Administrativos	Estado	Lucro
0	45173,06	116983,8	Rio de Janeiro	14681,4
542,05	0	51743,15	São Paulo	35673,41
0	0	135426,92	Rio de Janeiro	42559,73
1315,46	297114,46	115816,21	São Paulo	49490,75
1000,23	1903,93	124153,04	São Paulo	64926,08
20177,74	28334,72	154806,14	Rio de Janeiro	65200,33

R Studio

Arquivo "Análise de Resíduos.xlsx"

- Base de dados em "Startups"
- Código R em "Startups - R"

@2021 LABDATA FIA. Copyright all rights reserved.



# CASE: Predição *Startups*

CASE | FAZER ANÁLISE NO R

14

Um investidor deseja estimar o lucro de empresas *startups* de acordo com suas características, com objetivo de tomada de decisão de investimento baseada no lucro projetado da empresa. A base de dados histórica possui informações de investimento por área, gastos administrativos e região das empresas.



(a) Após obter o modelo de regressão linear, verifique se os resíduos possuem distribuição normal.



Arquivo "Análise de Resíduos.xlsx"

- i. Base de dados em "Startups"
- ii. Código R em "Startups - R"

@2021 LABDATA FIA. Copyright all rights reserved.





# CASE: Predição *Startups*

SUPOSIÇÕES 1 E 2 | FAZER ANÁLISE NO R

15

## #Ajuste do modelo de Regressão Linear

```
regressao <- lm(data=Startups,  
               Lucro ~  
               Investimento_PeD+Investimento_em_Mkt)
```

## #Fornece os resíduos do modelo

```
residuo<-residuals(regressao)
```

Calcula os resíduos do modelo

Histograma dos resíduos

## #Gráficos para verificar Normalidade resíduos

```
par(mfrow=c(1,2))
```

Matriz de gráficos (1 linha e 2 colunas)

```
hist(residuo, col="darkturquoise")
```

```
qqnorm(residuo, pch = 1,col="darkturquoise", frame = FALSE)
```

```
qqline(residuo, col = "steelblue", lwd = 2)
```

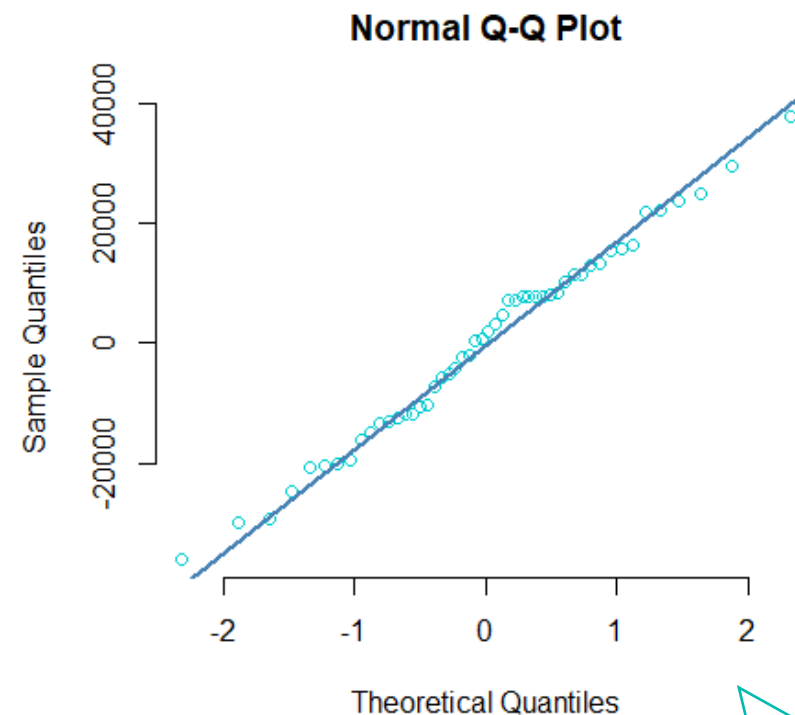
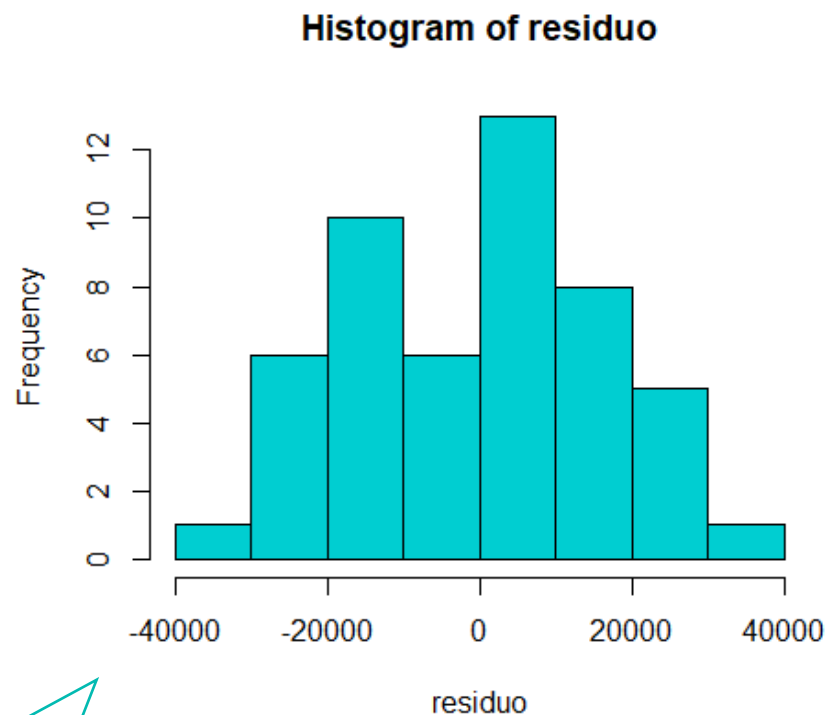
- *qq norm*: recebe os resíduos e compara com os quantis da distribuição normal padrão
- *qq line*: constrói a linha esperada para servir de referência sobre “o que é esperado” caso os resíduos sigam uma distribuição normal padrão.



# CASE: Predição *Startups*

SUPOSIÇÕES 1 E 2 | FAZER ANÁLISE NO R

16



Os resíduos estão distribuídos de forma aproximadamente simétrica em torno do valor zero.

Os quantis da distribuição observada são semelhantes aos quantis da distribuição teórica.

# CASE: Predição *Startups*

CASE | FAZER ANÁLISE NO R

17

Um investidor deseja estimar o lucro de empresas *startups* de acordo com suas características, com objetivo de tomada de decisão de investimento baseada no lucro projetado da empresa. A base de dados histórica possui informações de investimento por área, gastos administrativos e região das empresas.



(b) Após obter o modelo de regressão linear e verificar a normalidade dos resíduos, verifique se sua variância é constante.



Arquivo "Análise de Resíduos.xlsx"

- i. Base de dados em "Startups"
- ii. Código R em "Startups - R"

@2021 LABDATA FIA. Copyright all rights reserved.



# CASE: Predição *Startups*

SUPOSIÇÕES 3 E 4 | FAZER ANÁLISE NO R

18

**#Fornece os valores preditos do modelo**

```
predito<-fitted.values(regressao)
```

Valores ajustados pelo modelo

**#Gráfico para verificar igualdade de variâncias**

```
plot(predito, residuo, main='Resíduos x Ajustado', ylab='Resíduos', col="darkturquoise")
```

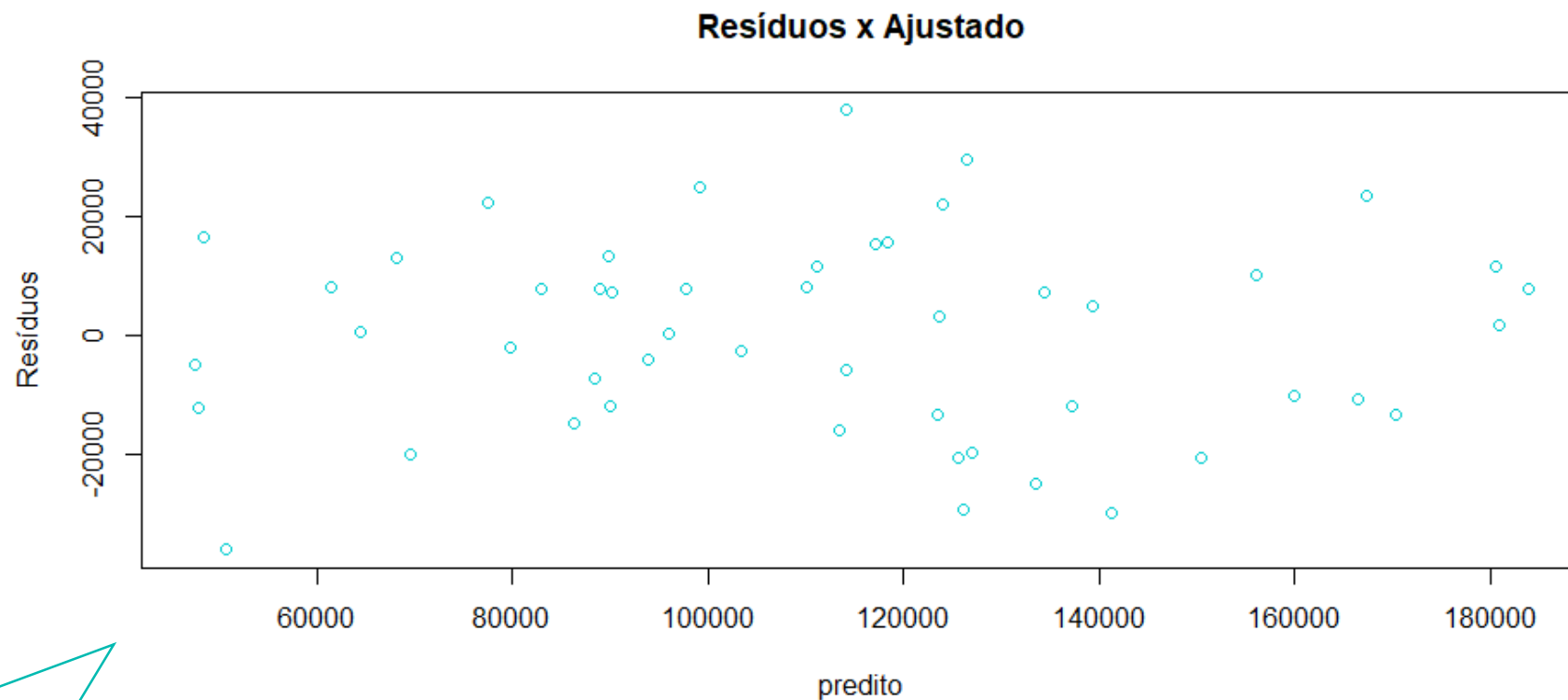
Gráfico de dispersão: Resíduos x Valores ajustados



# CASE: Predição *Startups*

SUPOSIÇÕES 3 E 4 | FAZER ANÁLISE NO R

19



O gráfico fornece uma nuvem horizontal de pontos distribuídos de forma aleatória ao redor do valor zero.



# CASE: Cervejaria

CASE | FAZER ANÁLISE NO R

20

Uma cervejaria deseja iniciar comercialização de uma das marcas de sua cerveja premium em uma cidade no interior de São Paulo. Para isso, ela deseja projetar qual será o consumo de cerveja (em litros) nesta cidade com base em algumas características da região. Cada linha da base de dados é uma cidade.



Temperatura_Media	Precipitacao	População	Renda_Media	Consumo_de_cerveja
27,3	0	38300	7280	14343
27,02	0	51840	9480	14940
24,82	0	50580	9550	16228
23,98	1,2	54180	8220	16748
23,82	0	46570	6810	16956
23,78	12,2	25470	6120	16977
24	0	24840	9790	17075
24,9	48,6	11200	9860	17241
28,2	4,4	36970	5130	17287

R Studio®

Arquivo "Análise de Resíduos.xlsx"

- Base de dados em "Cerveja"
- Código R em "Cerveja - R"





# CASE: Cervejaria

CASE | FAZER ANÁLISE NO R

21

Uma cervejaria deseja iniciar comercialização de uma das marcas de sua cerveja premium em uma cidade no interior de São Paulo. Para isso, ela deseja projetar qual será o consumo de cerveja (em litros) nesta cidade com base em algumas características da região. Cada linha da base de dados é uma cidade.



- (a) Após obter o modelo de regressão linear, verifique se os resíduos possuem distribuição normal.
- (b) Após obter o modelo de regressão linear e verificar a normalidade dos resíduos, verifique se sua variância é constante.



Arquivo "Análise de Resíduos.xlsx"

- i. Base de dados em "Cerveja"
- ii. Código R em "Cerveja - R"

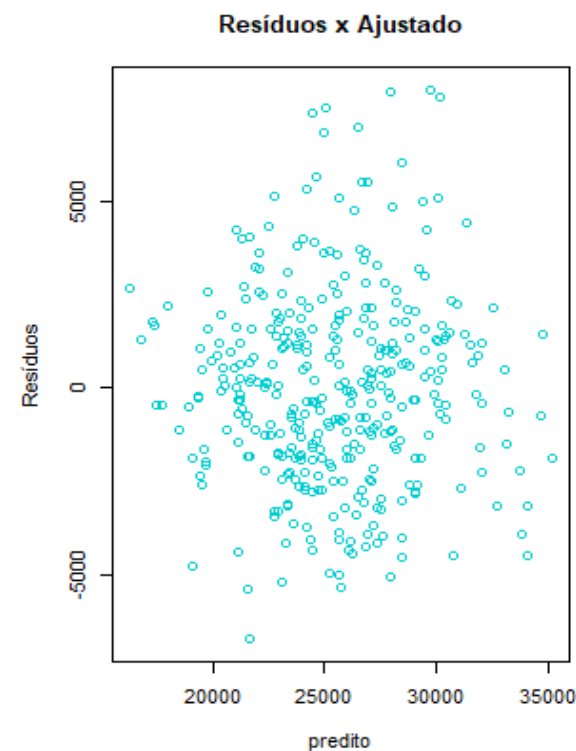
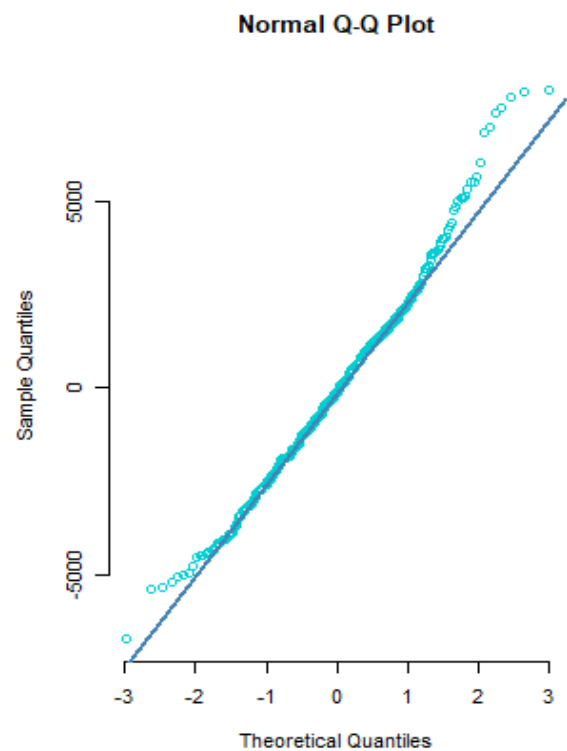
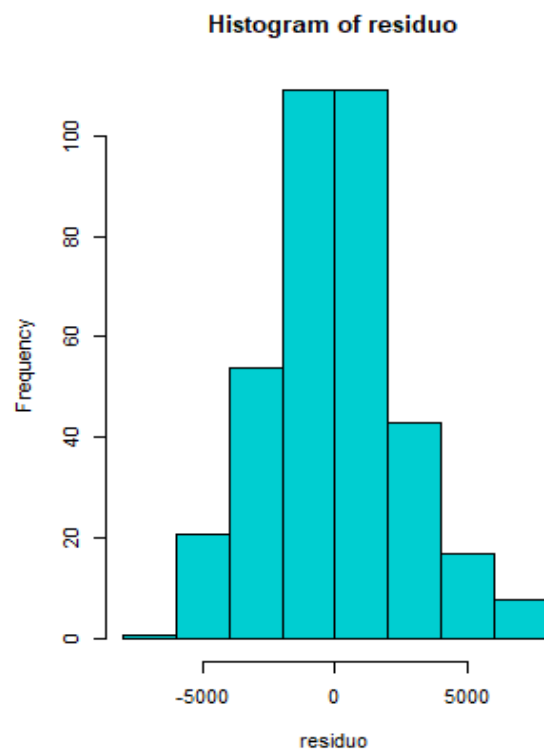
@2021 LABDATA FIA. Copyright all rights reserved.



# CASE: Cervejaria

CASE | FAZER ANÁLISE NO R

22



# CASE: Predição da rentabilidade média de um município

CASE | FAZER ANÁLISE NO R

23

Um município deseja projetar a rentabilidade média da sua população por meio da taxa de desocupação. A hipótese é que, quanto maior a taxa de desocupação (desemprego) da cidade, menores os salários oferecidos pelo mercado de trabalho. Considere a rentabilidade média do município como a soma dos gastos dividida pela soma dos salários.



Taxa de Desocupação	Taxa de rentab. média
21,9	18,5
6,0	33,7
22,8	19,7
18,1	21,0
12,7	35,1
14,5	19,4
20,0	25,3
19,2	17,0
16,0	24,0
6,6	31,4



- Arquivo “Análise de Resíduos.xlsx”
- Base de dados em “Rentabilidade”
  - Código R em “Rentabilidade - R”



# CASE: Predição da rentabilidade média de um município

CASE | FAZER ANÁLISE NO R

24

Um município deseja projetar a rentabilidade média da sua população por meio da taxa de desocupação. A hipótese é que, quanto maior a taxa de desocupação (desemprego) da cidade, menores os salários oferecidos pelo mercado de trabalho. Considere a rentabilidade média do município como a soma dos gastos dividida pela soma dos salários.



- (a) Após obter o modelo de regressão linear, verifique se os resíduos possuem distribuição normal.
- (b) Após obter o modelo de regressão linear e verificar a normalidade dos resíduos, verifique se sua variância é constante.



- Arquivo “Análise de Resíduos.xlsx”
- i. Base de dados em “Rentabilidade”
  - ii. Código R em “Rentabilidade - R”



# CASE: Predição de preço de imóvel

CASE | FAZER ANÁLISE NO R

25

De acordo com a localização de um imóvel, sabe-se que o valor do mesmo pode variar substancialmente. Na base de dados disponibilizada são fornecidas informações sobre o valor do imóvel (R\$) por mil m<sup>2</sup>, a distância para estação de metrô (km), a quantidade comércios próximos, e a idade (anos) do imóvel, em um bairro bem localizado de grande centro urbano. Quais são as características relacionadas ao imóvel que predizem seu valor?

Fonte Adaptada: <https://www.kaggle.com/quantbruce/real-estate-price-prediction?select=Real+estate.csv>



Idade_imovel	Distancia_metro_Km	Comercios_proximos	Mil_reais_m2
32	1,083595131	10	7,58
19,5	1,396946429	9	8,44
13,3	1,544788954	5	9,46
13,3	1,544788954	5	10,96
5	1,456009608	5	8,62
7,1	1,874980478	3	6,42
34,5	1,570122315	7	8,06
20,3	1,381344189	6	9,34
31,7	2,101860788	1	3,76

R Studio

Arquivo "Análise de Resíduos.xlsx"

- Base de dados em "Imobiliário"
- Código R em "Imobiliário - R"



# CASE: Predição de preço de imóvel

CASE | FAZER ANÁLISE NO R

26

De acordo com a localização de um imóvel, sabe-se que o valor do mesmo pode variar substancialmente. Na base de dados disponibilizada são fornecidas informações sobre o valor do imóvel (R\$) por mil m<sup>2</sup>, a distância para estação de metrô (km), a quantidade comércios próximos, e a idade (anos) do imóvel, em um bairro bem localizado de grande centro urbano. Quais são as características relacionadas ao imóvel que predizem seu valor?

Fonte Adaptada: <https://www.kaggle.com/quantbruce/real-estate-price-prediction?select=Real+estate.csv>



- (a) Após obter o modelo de regressão linear, verifique se os resíduos possuem distribuição Normal.
- (b) Após obter o modelo de regressão linear e verificar Normalidade dos resíduos, verifique se sua variância é constante.



Arquivo “Análise de Resíduos.xlsx”

- i. Base de dados em “Imobiliário”
- ii. Código R em “Imobiliário - R”

@2021 LABDATA FIA. Copyright all rights reserved.

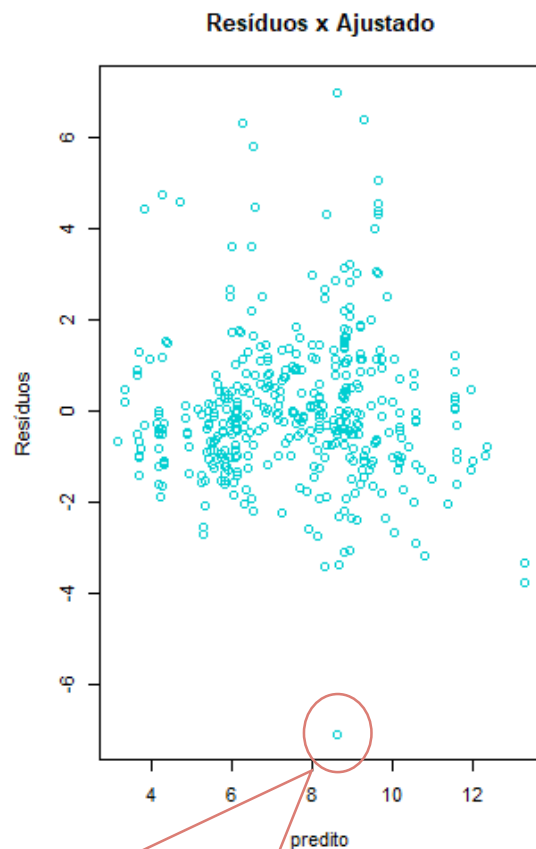




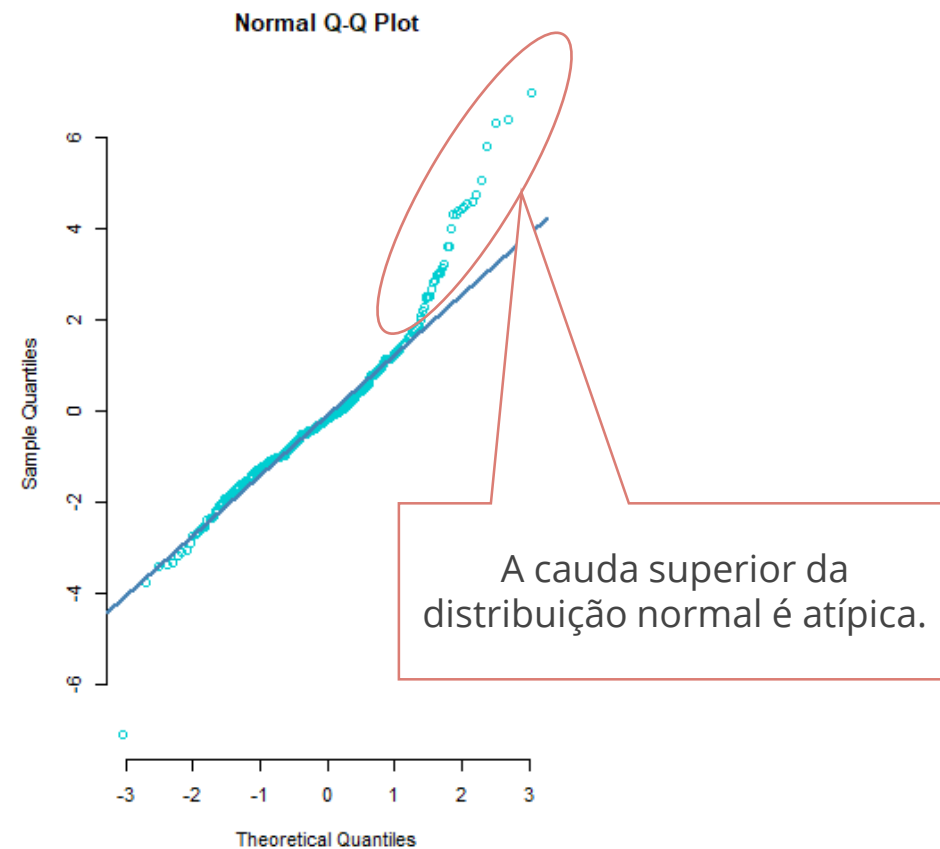
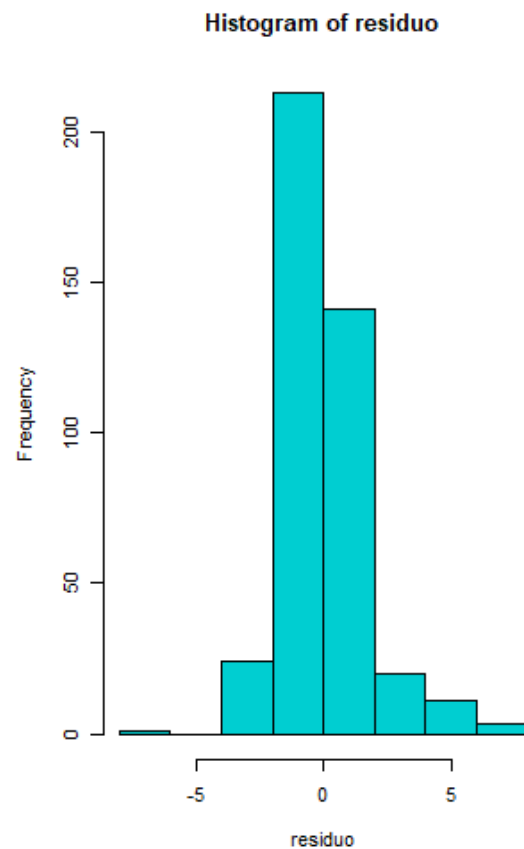
# CASE: Predição de preço de imóvel

CASE | FAZER ANÁLISE NO R

27



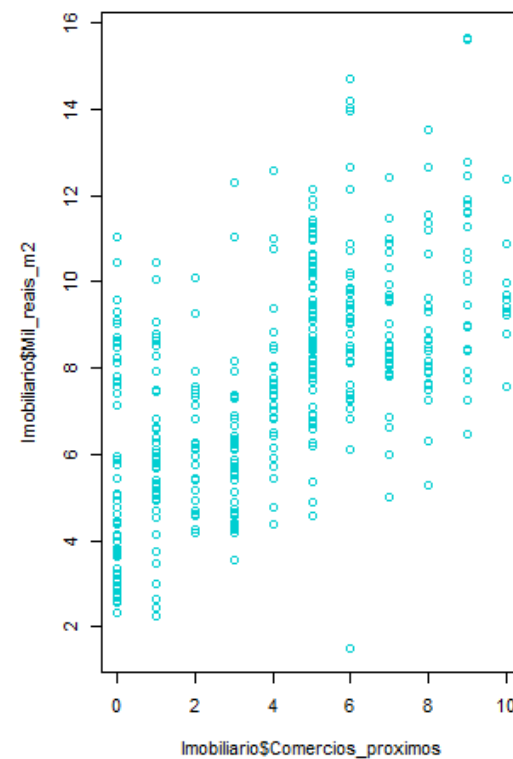
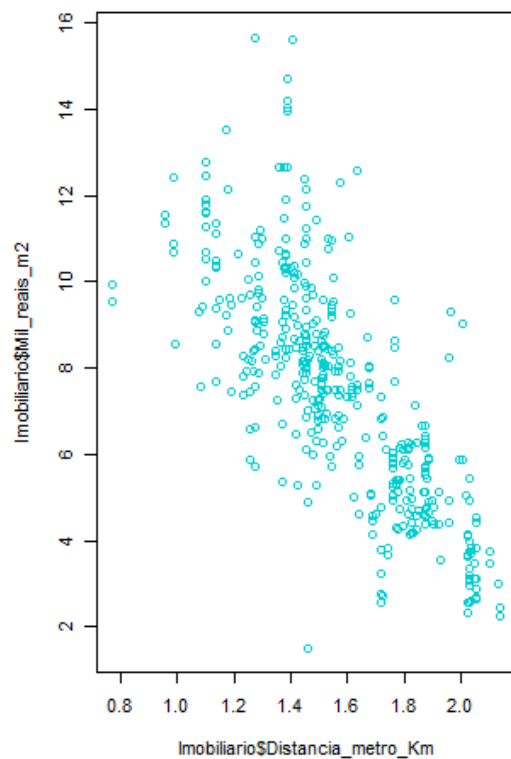
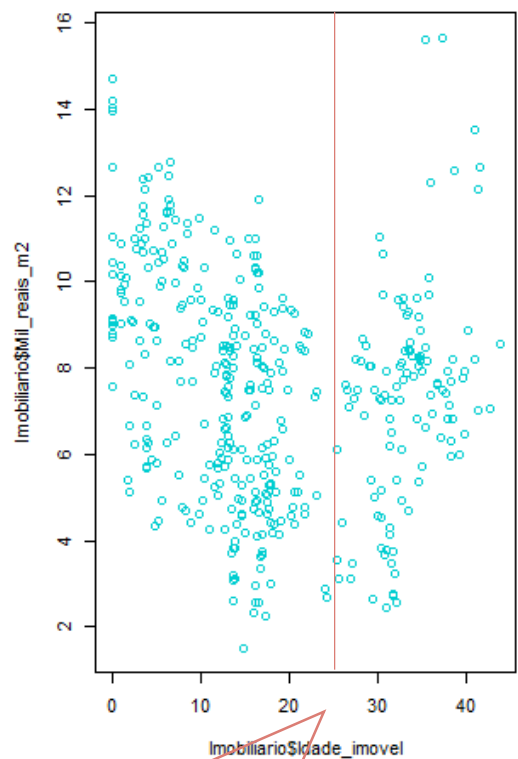
Há 1 ponto atípico com resíduo negativo muito mais alto que os demais.



# CASE: Predição de preço de imóvel

CASE | FAZER ANÁLISE NO R

28



Pode-se segmentar as observações, para manter a relação linear entre a idade e a variável resposta.

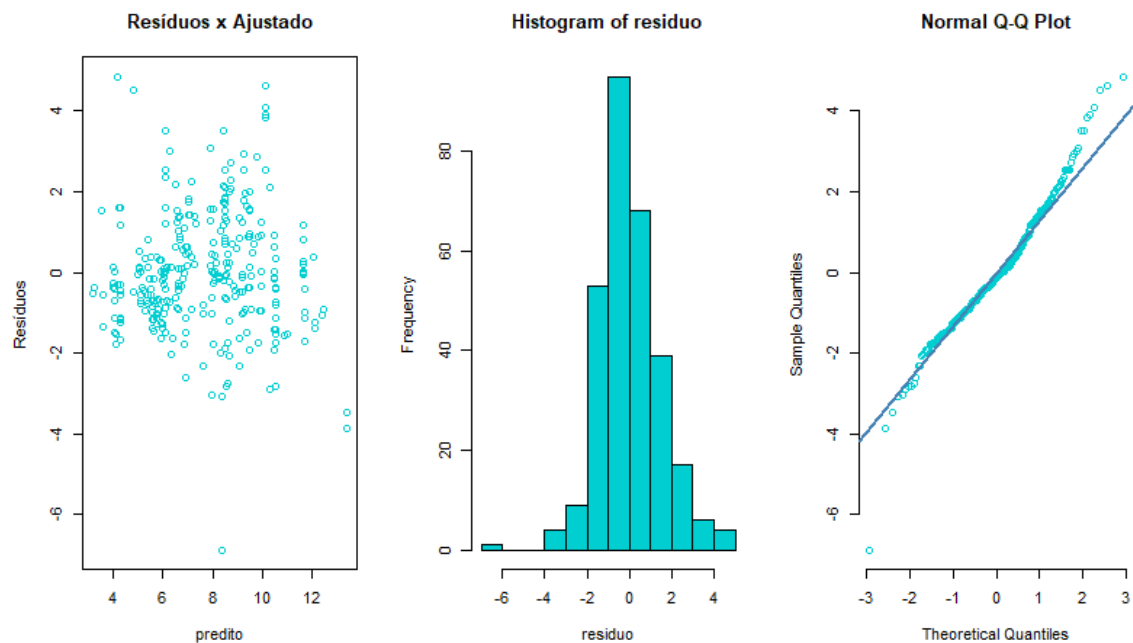


# CASE: Predição de preço de imóvel

CASE | FAZER ANÁLISE NO R

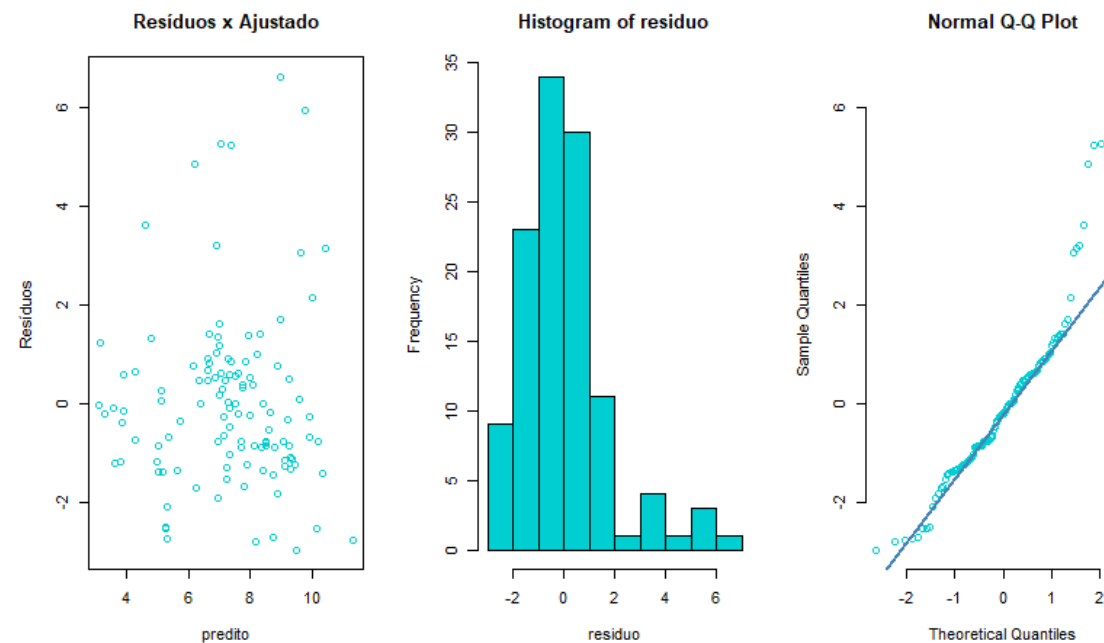
29

## Imóveis mais novos



A normalidade dos resíduos é melhor atendida para os imóveis mais novos.

## Imóveis mais antigos



Já para imóveis mais antigos, o modelo não está adequado. É necessário incluir variáveis para diminuir os resíduos, por exemplo.

## 2. Qualidade do modelo



# Quantificação dos resíduos

REGRESSÃO LINEAR | QUALIDADE DO MODELO

31

O **coeficiente de determinação** foi apresentado inicialmente, como indicador de qualidade de ajuste modelo de regressão linear. De forma complementar, apresentamos duas medidas de qualidade adicionais, o **SSE** e **MAPE**, definidos por:





- ❖ **SSE** (*Sum of Squares Errors*): é dado pela soma dos quadrados dos resíduos.
- ❖ **MAPE** (*Mean Absolute Percentage Error*): avalia a média absoluta dos resíduos, em relação ao valor original da resposta ( $y$ ).

$$\text{SSE} = \sum (y - \hat{y})^2$$

$$\text{MAPE} = \frac{100\%}{n} \sum \left| \frac{y - \hat{y}}{y} \right|$$

**Interpretação:** quanto menores os valores de SSE e MAPE, melhor o ajuste do modelo, pois buscamos resíduos mais próximos de zero.



Estatística	Critério
R-Quadrado (regressão linear simples)	Quanto <b>maior</b> , melhor 
R-Quadrado Ajustado (regressão linear múltipla)	Quanto <b>maior</b> , melhor 
SSE	Quanto <b>menor</b> , melhor 
MAPE	Quanto <b>menor</b> , melhor 





# Case: Companhia aérea

CASE | FAZER ANÁLISE NO R

33

Uma empresa de turismo deseja estimar as vendas mensais (R\$) de passagens aéreas em função do tempo de experiência (anos) dos agentes de viagem. Existe relação linear entre as duas informações?



Tempo de Experiência (Anos)	Vendas Mensais (Mil R\$)
1	91
3	110
4	106
4	116
6	119
8	129
10	139
10	143
11	138
13	159

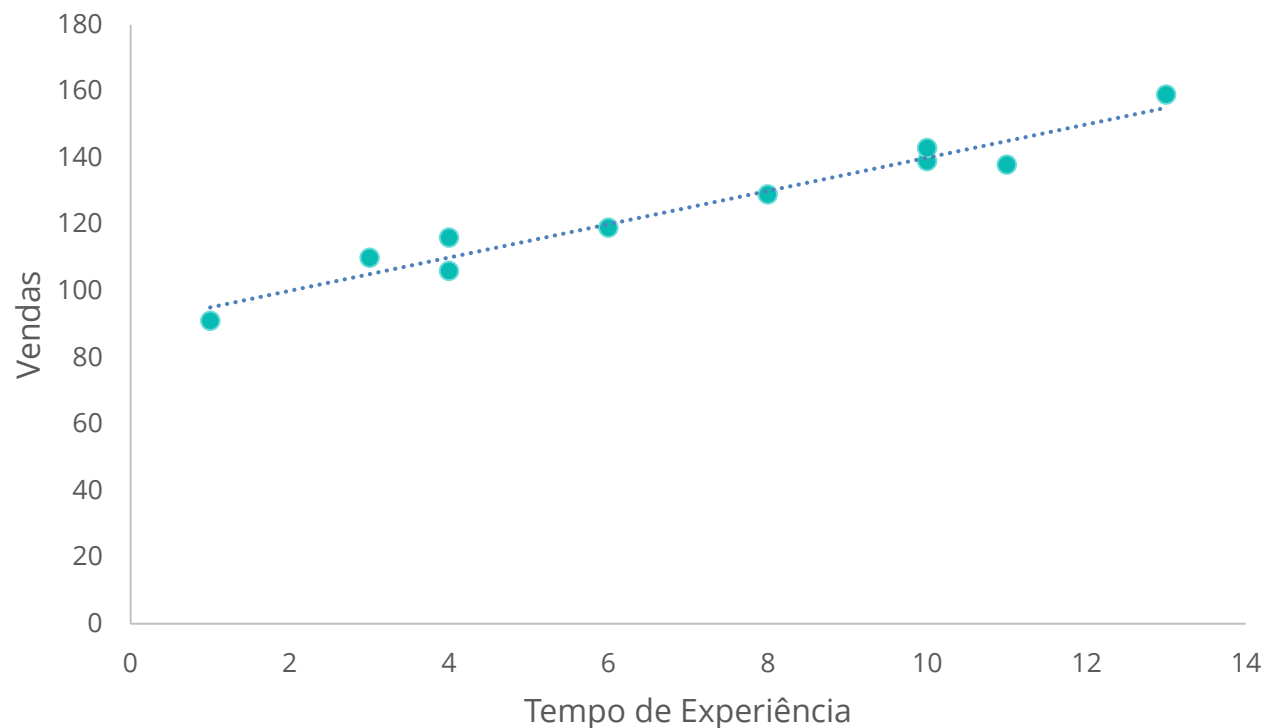


# Case: Companhia aérea

CASE | FAZER ANÁLISE NO R

34

Uma empresa de turismo deseja estimar as vendas mensais (R\$) de passagens aéreas em função do tempo de experiência (anos) dos agentes de viagem. Existe relação linear entre as duas informações?



$$\text{Vendas} = 90 + 5 * (\text{anos de experiência})$$

90: R\$90.000 é o valor esperado da venda mensal para um vendedor que não possui experiência

5: R\$5.000 seria o acréscimo esperado na venda mensal a cada variação de um ano no tempo de experiência do vendedor



# Exemplo: quantificação dos resíduos

REGRESSÃO LINEAR | QUALIDADE DO MODELO

35

Tempo de Experiência (Anos)	Valor da venda mensal	Valor da venda mensal estimado pelo modelo	Erro	Erro Absoluto	Erro Absoluto Percentual	Erro <sup>2</sup>
1	91	95	-4	4	4,40%	16
3	110	105	5	5	4,55%	25
4	106	110	-4	4	3,77%	16
4	116	110	6	6	5,17%	36
6	119	120	-1	1	0,84%	1
8	129	130	-1	1	0,78%	1
10	139	140	-1	1	0,72%	1
10	143	140	3	3	2,10%	9
11	138	145	-7	7	5,07%	49
13	159	155	4	4	2,52%	16

MAPE	SSE
2,99%	170

**SSE** e **MAPE** são muito utilizados para avaliação de “melhoria do ajuste” a cada passo de redução de variáveis até se chegar no modelo final ajustado.

Também é útil para comparar com outras técnicas, utilizando a mesma base de dados.



# CASE: Predição *Startups*

CASE | FAZER ANÁLISE NO R

36

Um investidor deseja estimar o lucro de empresas *startups* de acordo com suas características, com objetivo de tomada de decisão de investimento baseada no lucro projetado da empresa. A base de dados histórica possui informações de investimento por área, gastos administrativos e região das empresas.



- (a) Ajuste o modelo de Regressão Linear Múltipla. A cada passo de redução de variáveis, avalie as medidas MAPE e SSE do modelo.
- (b) Discuta os valores de MAPE e SSE em conjunto com o indicador de  $R^2$ -ajustado.



Arquivo “Análise de Resíduos.xlsx”

- i. Base de dados em “Startups”
- ii. Código R em “Startups - R”

@2021 LABDATA FIA. Copyright all rights reserved.



# CASE: Cervejaria

CASE | FAZER ANÁLISE NO R

37

Uma cervejaria deseja iniciar comercialização de uma das marcas de sua cerveja premium em uma cidade no interior de São Paulo. Para isso, ela deseja projetar qual será o consumo de cerveja (em litros) nesta cidade com base em algumas características da região. Cada linha da base de dados é uma cidade.



- (a) Ajuste o modelo de Regressão Linear Múltipla. A cada passo de redução de variáveis, avalie as medidas MAPE e SSE do modelo.
- (b) Discuta os valores de MAPE e SSE em conjunto com o indicador de  $R^2$ -ajustado.



Arquivo "Análise de Resíduos.xlsx"

- i. Base de dados em "Cerveja"
- ii. Código R em "Cerveja - R"

@2021 LABDATA FIA. Copyright all rights reserved.



## 3. Código em R



# Análise de Resíduos e Qualidade de ajuste

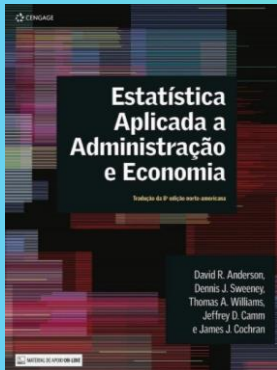
REGRESSÃO LINEAR | EXEMPLO: CASE STARTUP

39

```
regressao <- lm(data = Startups,  
               Lucro ~  
               Investimento_PeD +  
               Investimento_em_Mkt)  
summary(regressao)  
residuo <- residuals(regressao) #fornece os resíduos do modelo  
  
# Gráficos para verificar normalidade dos resíduos  
par(mfrow = c(1,2))  
hist(residuo, col = "darkturquoise")  
qqnorm(residuo, pch = 1,col = "darkturquoise", frame = FALSE)  
qqline(residuo, col = "steelblue", lwd = 2)  
  
#Gráfico para verificar igualdade de variâncias  
predito <- fitted.values(regressao) #fornece os preditos do modelo  
par(mfrow = c(1,1))  
plot(predito, residuo, main = 'Resíduos x Ajustado', ylab = 'Resíduos', col = "darkturquoise")  
  
# Quantificação dos resíduos  
library(Metrics)  
mape(actual = Startups$Lucro, predicted = predito)  
sse(actual = Startups$Lucro, predicted = predito)
```







1. Anderson, R. A., Sweeney, J. D. e Williams, T. A. *Estatística Aplicada à Administração e Economia*. Editora Cengage. 4ª edição, 2019.

