# BIG DATA

Tema da aula
**Regressão Linear**

# FIA Business School



**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

FUNDAÇÃO INSTITUTO DE ADMINISTRAÇÃO

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

# LABDATA FIA



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**.

Profª Drª Alessandra Montini

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)



LABDATA | FIA FUNDAÇÃO INSTITUTO DE ADMINISTRAÇÃO BUSINESS SCHOOL

# CONTEÚDO PROGRAMÁTICO

ANALYTICS

MACHINE LEARNING

DEEP LEARNING

ANÁLISE EXPLORATÓRIA

TÉCNICAS DE PROJEÇÃO

TÉCNICAS DE CLASSIFICAÇÃO

TÉCNICAS DE SEGMENTAÇÃO

TÉCNICAS DE ANALYTICS

LINGUAGEM

PYTHON

R

PROJETO

PROJETO ANALYTICS

# Conteúdo da Aula



1. Introdução
2. Coeficiente de Correlação
3. Regressão Linear Simples
4. Regressão Linear Múltipla
   i. Multicolinearidade
   ii. Variáveis explicativas qualitativas
5. Código em R

# 1. Introdução

# Case: Limite de Cartão de Crédito



**Exemplo**
Predizer o valor do limite do cartão de crédito em função da renda do cliente.

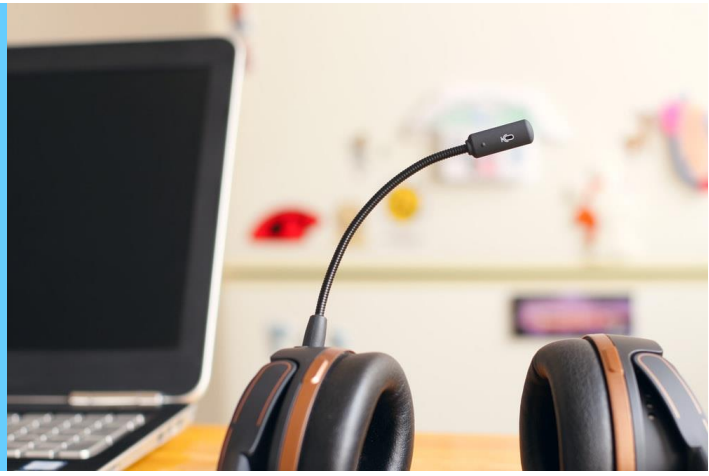
**Aplicação**
Área de Crédito do Segmento Bancário (Emissores de cartão de crédito).

# Case: SAC em Empresas de Serviço



**Exemplo**
Predizer o valor a ser investido em uma central de atendimento (SAC) de uma empresa de serviços com base na quantidade de clientes.

**Aplicação**
Área de Ouvidoria de empresas de serviços (Telecom, Bancos, Seguradoras, etc.)



# Case: Educação



**Exemplo**
Predizer o percentual de rematrículas em uma escola de Idiomas com base nas notas dos alunos do ano anterior.

**Aplicação**
Áreas de Marketing e Vendas de Instituição de Ensino.

# Case: Venda de Seguros



**Exemplo**
Predizer o faturamento de um time Comercial com base na quantidade de vendedores ativos.

**Aplicação**
Área de Planejamento Comercial.

# Case: Venda de Eletrônicos pela Internet



**Exemplo**
Predizer o volume (R$) de vendas em eletrônicos em função do investimento (R$) em Mídia Digital (Facebook, Instagram, Mídia Programática, *Search*).

**Aplicação**
Área de Mídias Digitais.

# Case: Covid-19



**Exemplo**
Predizer o valor gasto pelas prefeituras para o tratamento dos infectados de COVID-19 numa certa região com base no tamanho da população.

**Aplicação**
Área de Saúde Pública.

# 2. Coeficiente de correlação

# Existe relação linear entre as variáveis?



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?

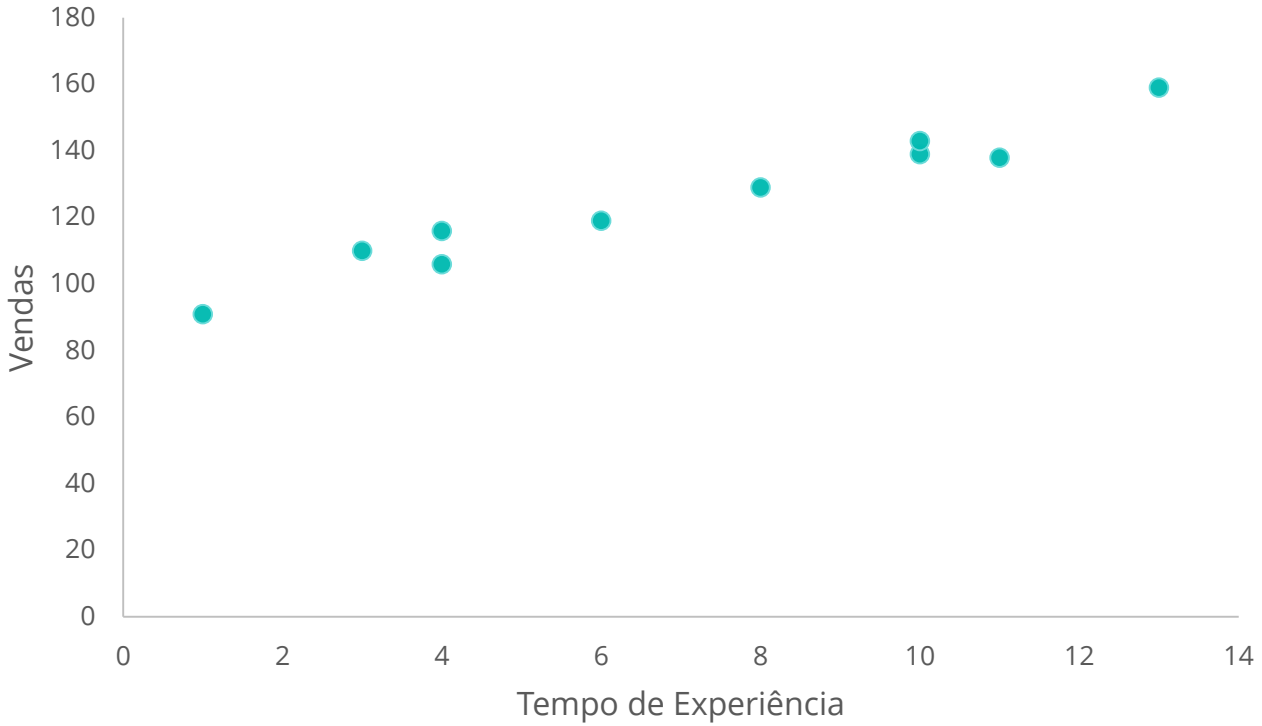
| Tempo de Experiência (Anos) | Vendas Mensais (Mil R$) |
|---|---|
| 1 | 91 |
| 3 | 110 |
| 4 | 106 |
| 4 | 116 |
| 6 | 119 |
| 8 | 129 |
| 10 | 139 |
| 10 | 143 |
| 11 | 138 |
| 13 | 159 |

Arquivo “Regressão linear simples.xlsx”, aba “cia_aerea”

# Existe relação linear entre as variáveis?



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?



# Coeficiente de Correlação



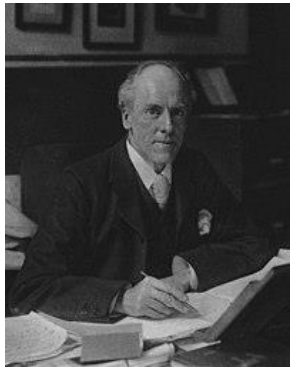
Mede a **relação linear entre duas variáveis** quantitativas.

O coeficiente **r** varia entre -1 e 1, sendo:

- valores próximos a **1**: forte correlação linear positiva (diretamente proporcional).
- valores próximos a **-1**: forte correlação linear negativa (inversamente proporcional).
- valores próximos a **0**: não existe relação linear entre as variáveis.

O coeficiente também é conhecido como CORRELAÇÃO DE PEARSON

O coeficiente de correlação linear foi a primeira medida de relação introduzida na Estatística

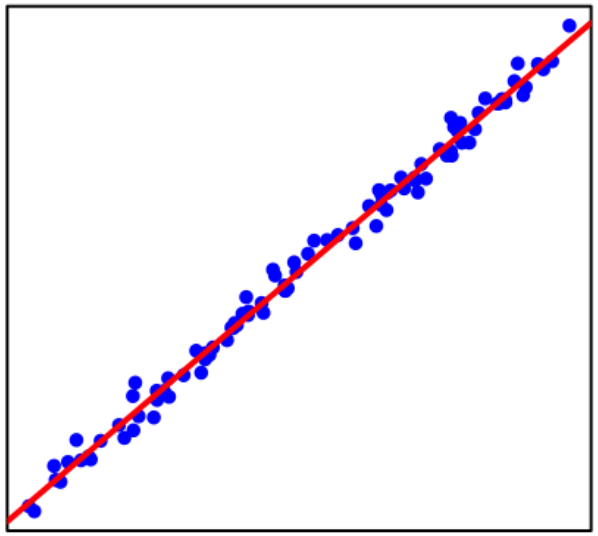


**Karl Pearson**
(Londres, 1857-1936)

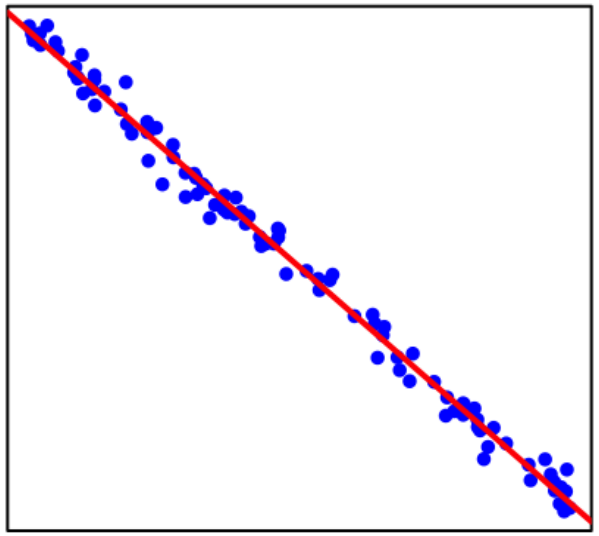
# Interpretação dos valores de correlação



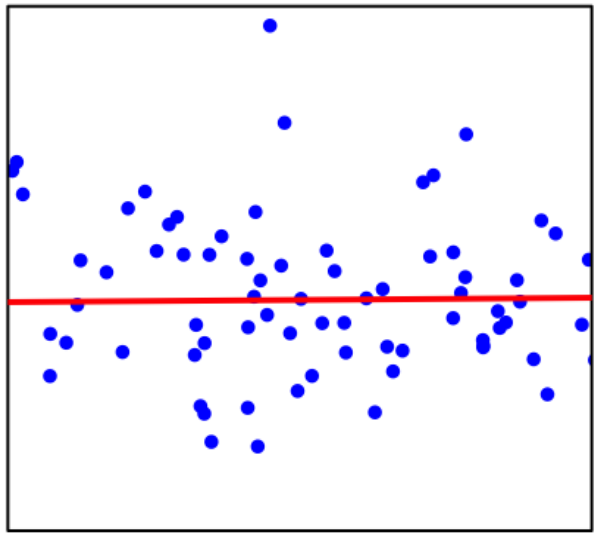
**Correlação linear positiva forte**
(valor de r próximo a 1)



**Correlação linear negativa forte**
(valor de r próximo a -1)



**Sem correlação linear**
(valor de r próximo a 0)



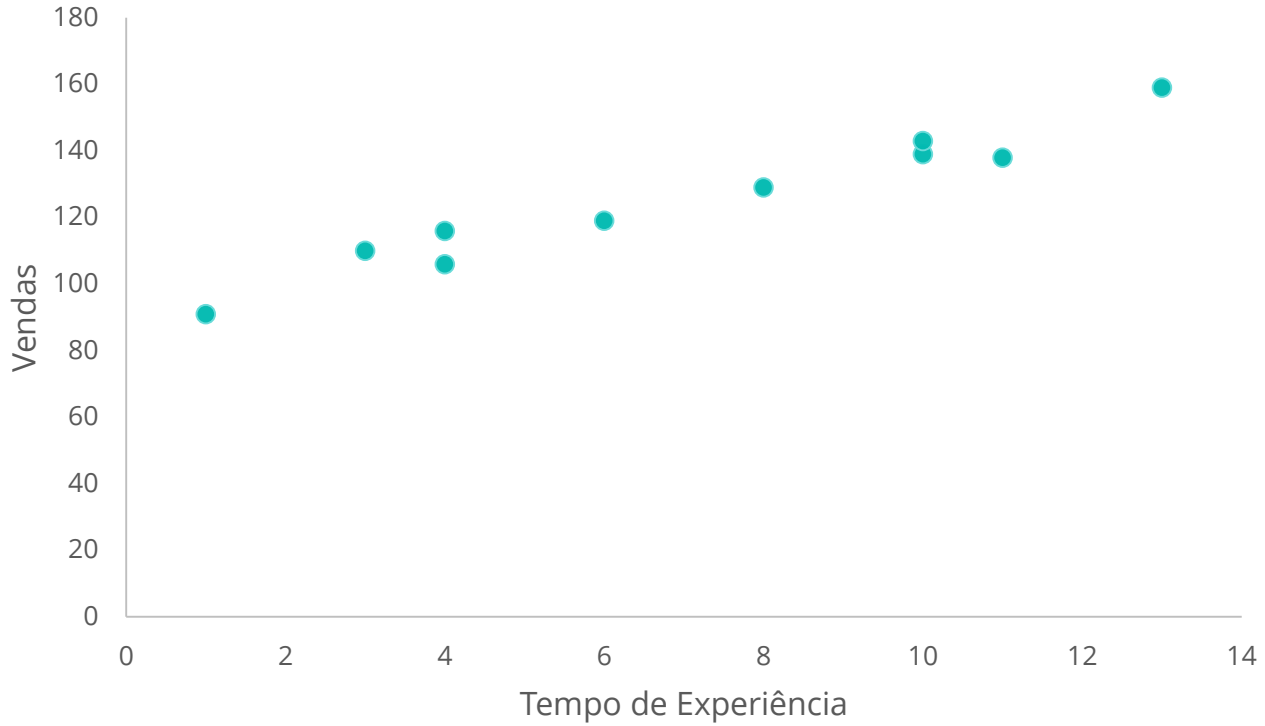
Fonte: Department of Earth Sciences - Freie Universität Berlin

Na prática, consideramos valores acima de |r|>0,7 como alta/forte correlação linear.

# Existe relação linear entre as variáveis?



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?

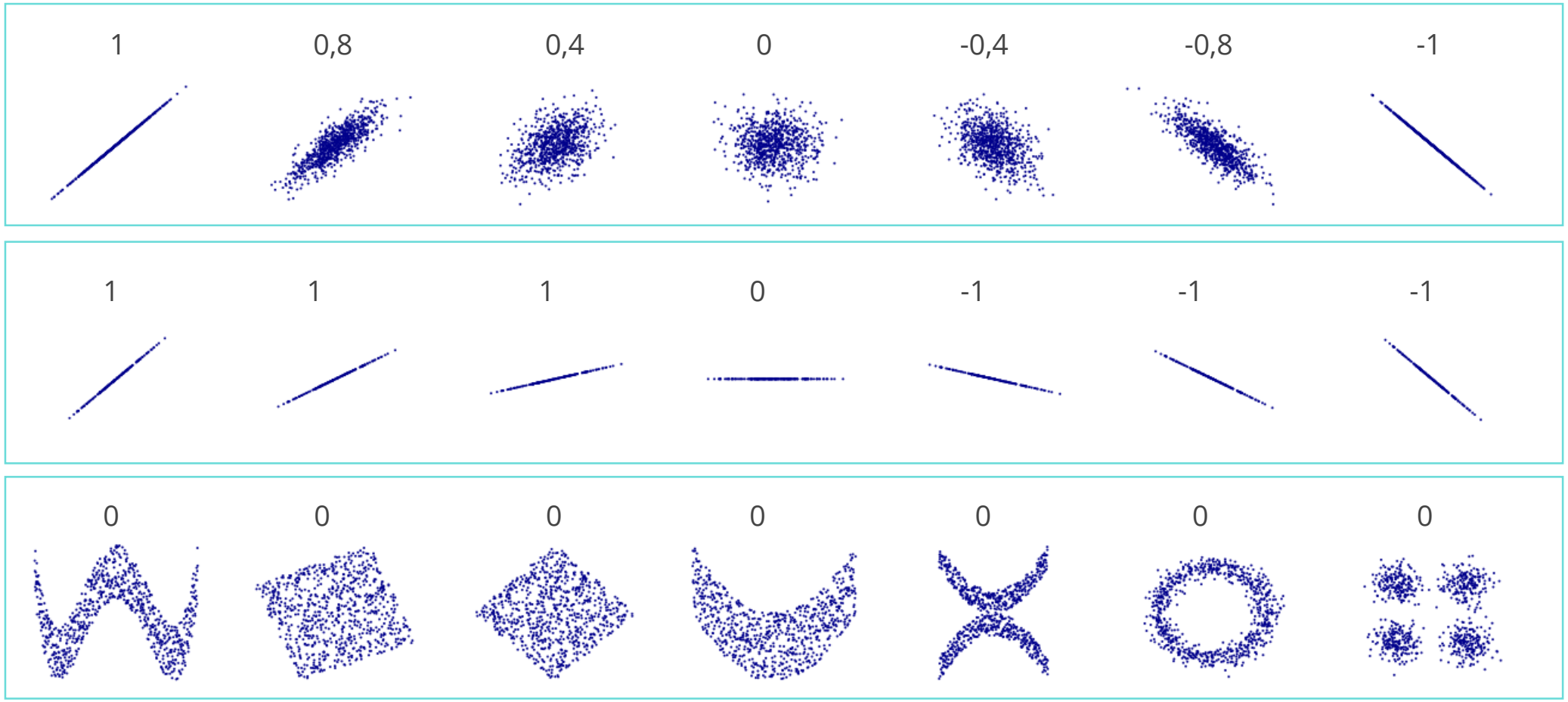



Existe uma forte correlação (r=0,98) entre as duas variáveis, ou seja, quanto maior o tempo de experiência, maior o volume de vendas mensais.

- Excel: CORREL(col1, col2)
- R: cor(var1, var2)

# Correlação de Pearson: somente relação LINEAR





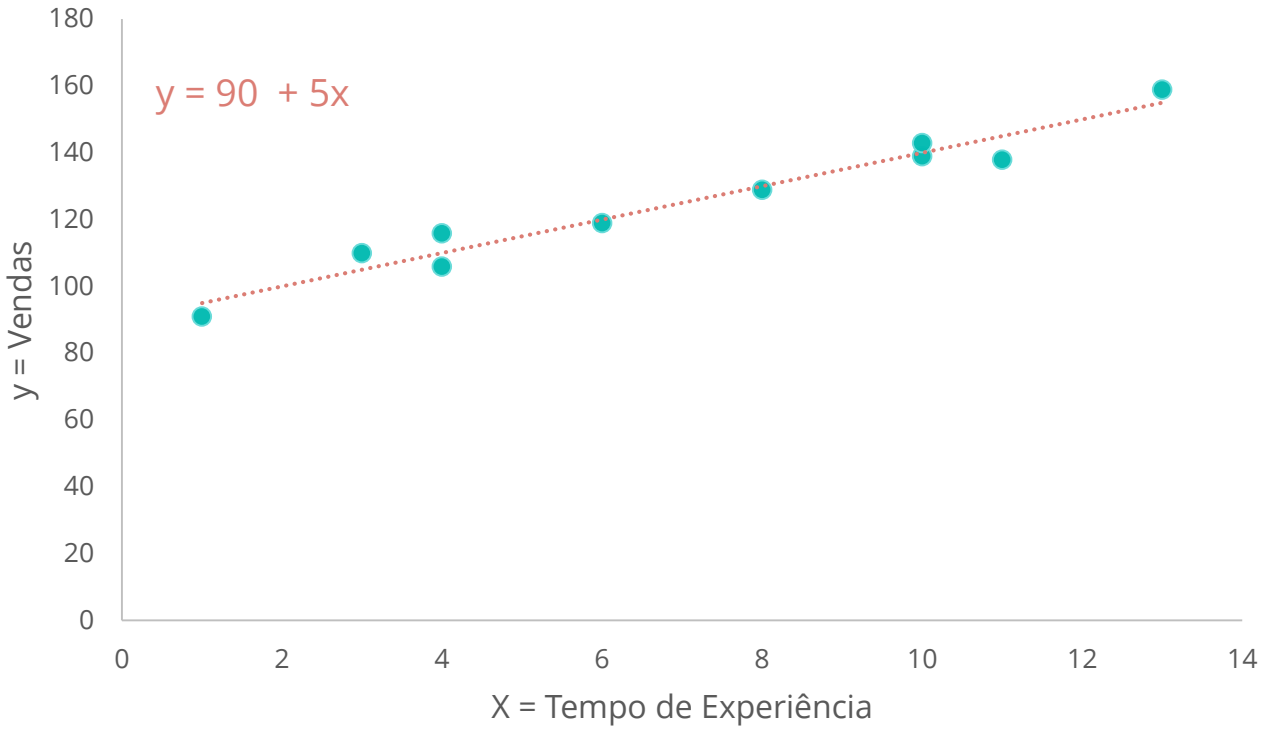

https://en.wikipedia.org/wiki/Correlation_and_dependence#/media/File:Correlation_examples2.svg

# 2. Regressão Linear Simples

# É possível expressar essa relação por meio de um modelo estatístico?



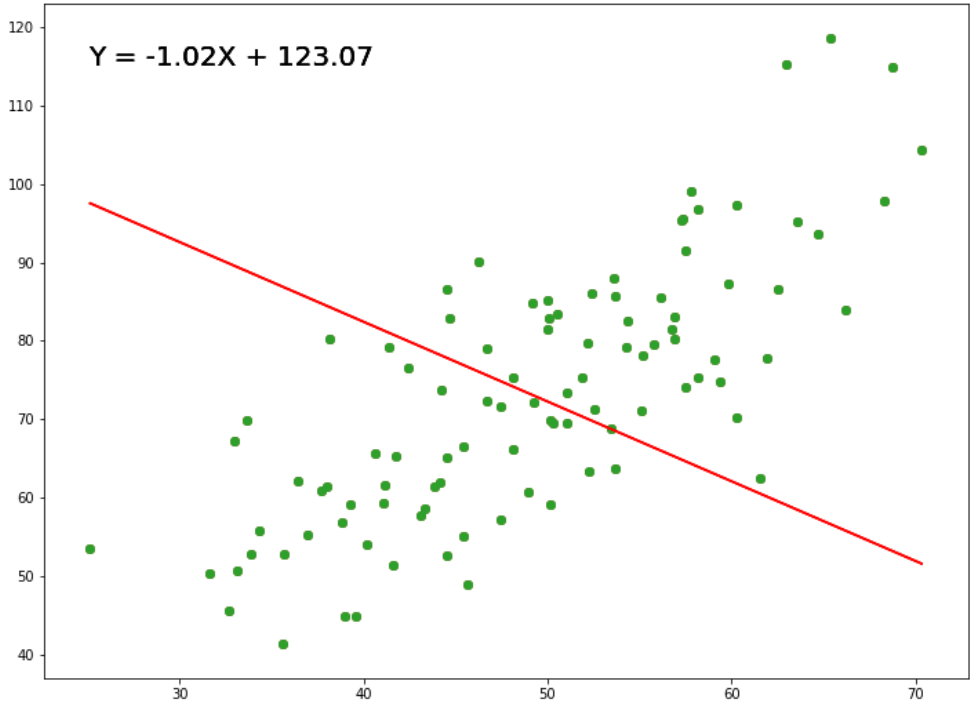
Quando se realiza uma análise de dados, busca-se de alguma forma um padrão ou modelo presente nas observações.




**PERGUNTA DE NEGÓCIO:**
Se aumentar em 1 ano o tempo de experiência, em quanto aumentam as vendas mensais (R$)?

# Animação: Valores de intercepto e inclinação






https://towardsdatascience.com/linear-regression-using-least-squares-a4c3456e8570

# É possível expressar essa relação por meio de um modelo estatístico?



Quando se realiza uma análise de dados, busca-se de alguma forma um padrão ou modelo presente nas observações.

**PERGUNTA DE NEGÓCIO:**
Se aumentar em 1 ano o tempo de experiência, em quanto aumentam as vendas mensais (R$)?

**Vendas = 90 + 5 * 0**

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 1 ano no tempo de experiência do vendedor

# É possível expressar essa relação por meio de um modelo estatístico?



Quando se realiza uma análise de dados, busca-se de alguma forma um padrão ou modelo presente nas observações.

**PERGUNTA DE NEGÓCIO:**
Se aumentar em 1 ano o tempo de experiência, em quanto aumentam as vendas mensais (R$)?

**Vendas = 90 + 5 * (anos de experiência)**

Qual é o valor da venda mensal estimada para um vendedor com 6 anos de experiência?

# É possível expressar essa relação por meio de um modelo estatístico?



Quando se realiza uma análise de dados, busca-se de alguma forma um padrão ou modelo presente nas observações.

**PERGUNTA DE NEGÓCIO:**
Se aumentar em 1 ano o tempo de experiência, em quanto aumentam as vendas mensais (R$)?

**Vendas = 90 + 5 * (anos de experiência)**

Qual é o valor da venda mensal estimada para um vendedor com 6 anos de experiência?

O valor da venda estimada é de **90+5*(6) = 120** mil reais

# É possível expressar essa relação por meio de um modelo estatístico?



Quando se realiza uma análise de dados, busca-se de alguma forma um padrão ou modelo presente nas observações.

**PERGUNTA DE NEGÓCIO:**
Se aumentar em 1 ano o tempo de experiência, em quanto aumentam as vendas mensais (R$)?

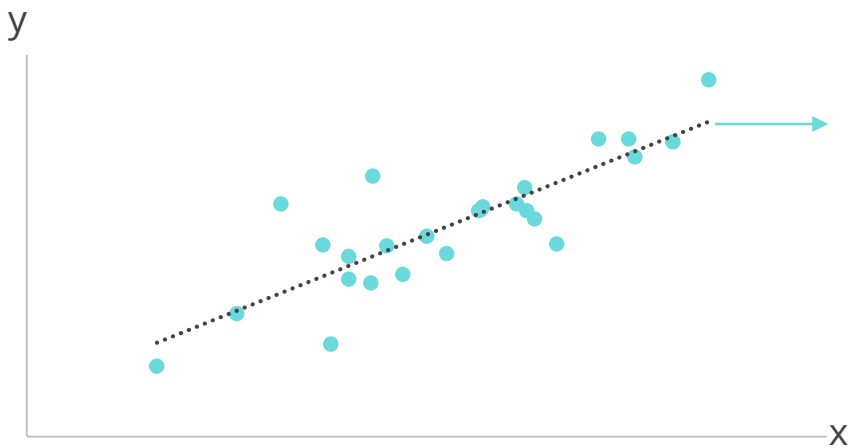
| Tempo de Experiência (Anos) | Valor da venda mensal (mil R$) | **Valor da venda mensal estimado pelo modelo** | Erro |
|---|---|---|---|
| 1 | 91 | **95** | -4 |
| 3 | 110 | **105** | 5 |
| 4 | 106 | **110** | -4 |
| 4 | 116 | **110** | 6 |
| 6 | 119 | **120** | -1 |
| 8 | 129 | **130** | -1 |
| 10 | 139 | **140** | -1 |
| 10 | 143 | **140** | 3 |
| 11 | 138 | **145** | -7 |
| 13 | 159 | **155** | 4 |

e: **erro** = **resíduo** = venda observada – venda ajustada

# Modelo estimado



A equação de regressão linear simples **estimada** é dada por: $\hat{y} = b_0 + b_1 x$




$\hat{y} = b_0 + b_1 x$

Em que :

- ✓ $b_0$ e $b_1$ são chamados de **parâmetros <u>estimados</u> do modelo**.
- ✓ $\hat{y}$ é a resposta estimada.

Sendo:

- ✓ $b_0$ é o valor do intercepto (x=0).
- ✓ $b_1$ é coeficiente angular (inclinação da reta).

# Modelo: teórico x ajustado



Regressão Linear Simples
**Modelo teórico**

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

X

Regressão Linear Simples
**Modelo ajustado**

$$\hat{y}_i = b_0 + b_1 x_i$$

# Estrutura do modelo



O modelo de regressão linear simples é dado por:

y: variável resposta, variável dependente ou *target*

$$y = \beta_0 + \beta_1 x + \varepsilon$$

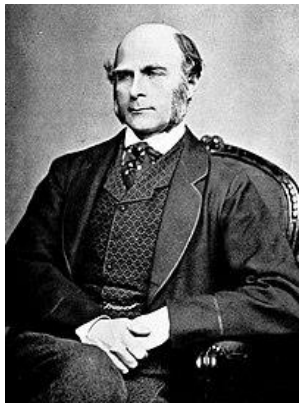
x: variável explicativa, variável auxiliar, variável independente ou covariável

Em que :

- ✓ $\beta_0$ e $\beta_1$ são chamados **parâmetros do modelo**.
- ✓ $\varepsilon$ é uma variável aleatória chamada de **erro** ou **resíduo**.

Explicou pela 1ª vez por meio de um modelo estatístico a relação entre duas variáveis

Ele estudava junto com seu discípulo, Karl Pearson, a relação entre a altura do pai com a altura do filho



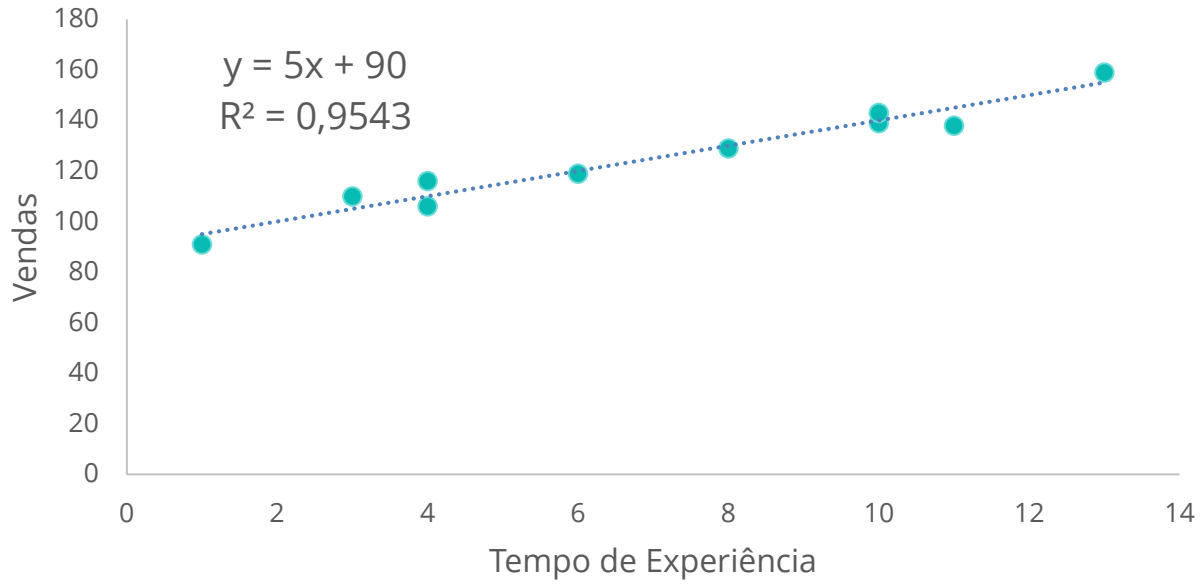
**Francis Galton**
(Londres, 1822-1911)

# Coeficiente de Determinação



No Excel, é possível incluir uma linha de tendência, e ele fornece a estimação dos parâmetros do modelo e o valor $R^2$.

**$R^2$** é o **coeficiente de determinação**, que pode ser calculado pelo **quadrado do coeficiente de correlação**. Quanto maior o valor de $R^2$, mais bem ajustado está o modelo de regressão. Valores de $R^2$ acima de 0,5 já indicam bom ajuste, $0 < R^2 < 1$. Ele pode ser interpretado como o % da variabilidade explicada da variável y pela x.




O modelo explica 95,43% da variabilidade das vendas, por meio do tempo de experiência.

# Case: Captação de alunos



Um gestor de uma instituição de ensino está planejando a abertura de novas vagas para cursos de ensino superior, e gostaria de utilizar os dados de aprovados no ensino médio do ano anterior para estimar o potencial de público alvo que teria para trabalhar com ações de marketing. Para isso, ele analisou os dados disponíveis dos estudantes aprovados, por Estado do Brasil, dos últimos 2 anos (2015 e 2016). Ele gostaria de saber se é possível utilizar os dados do último ano para estimar o percentual de aprovados no ano corrente (2017).

Fonte adaptada: https://seriesestatisticas.ibge.gov.br/series.aspx?no=7&op=2&vcodigo=M13&t=aprovacao-serie-ensino-medio-serie-nova

| Estado | 2015 | 2016 |
|---|---|---|
| Alagoas | 66,1 | 72,4 |
| Amapa | 68,8 | 70,5 |
| Amazonas | 78,6 | 84,7 |
| Bahia | 67,2 | 69,6 |
| Ceara | 80,4 | 82,8 |
| DF | 74,9 | 79,6 |
| Espirito Santo | 70,0 | 74,5 |
| Goiás | 80,1 | 84,7 |
| M. G. do Sul | 64,8 | 73,6 |

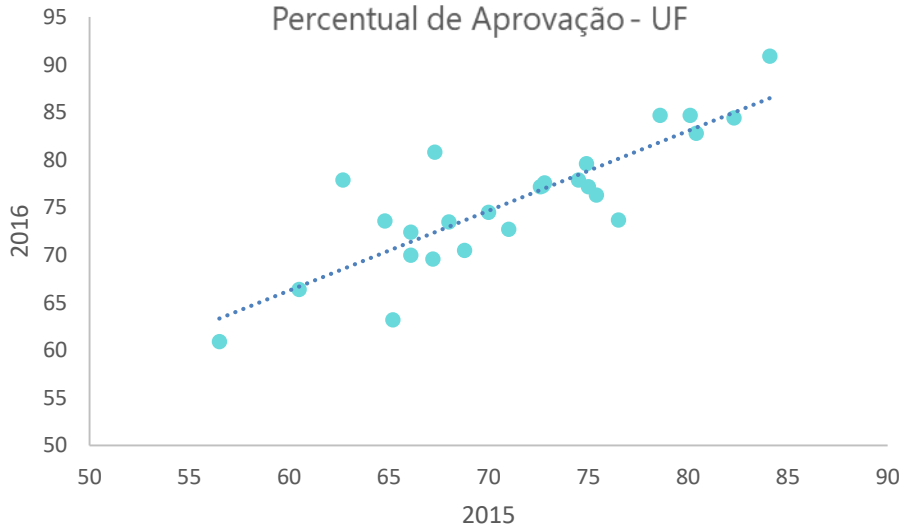
Base “Captacao_Alunos.txt” e código R em “Regressão linear simples.xlsx”, aba “captação_alunos – R”

# Case: Captação de alunos



Um gestor de uma instituição de ensino está planejando a abertura de novas vagas para cursos de ensino superior, e gostaria de utilizar os dados de aprovados no ensino médio do ano anterior para estimar o potencial de público alvo que teria para trabalhar com ações de marketing. Para isso, ele analisou os dados disponíveis dos estudantes aprovados, por Estado do Brasil, dos últimos 2 anos (2015 e 2016). Ele gostaria de saber se é possível utilizar os dados do último ano para estimar o percentual de aprovados no ano corrente (2017).

Fonte adaptada: https://seriesestatisticas.ibge.gov.br/series.aspx?no=7&op=2&vcodigo=M13&t=aprovacao-serie-ensino-medio-serie-nova




Existe uma forte correlação POSITIVA (r=0,84) entre as duas variáveis, ou seja, os estados que apresentam altos % de aprovação em 2015 também tendem a apresentar altos % de aprovação em 2016.

Base “Captacao_Alunos.txt” e código R em “Regressão linear simples.xlsx”, aba “captação_alunos – R”

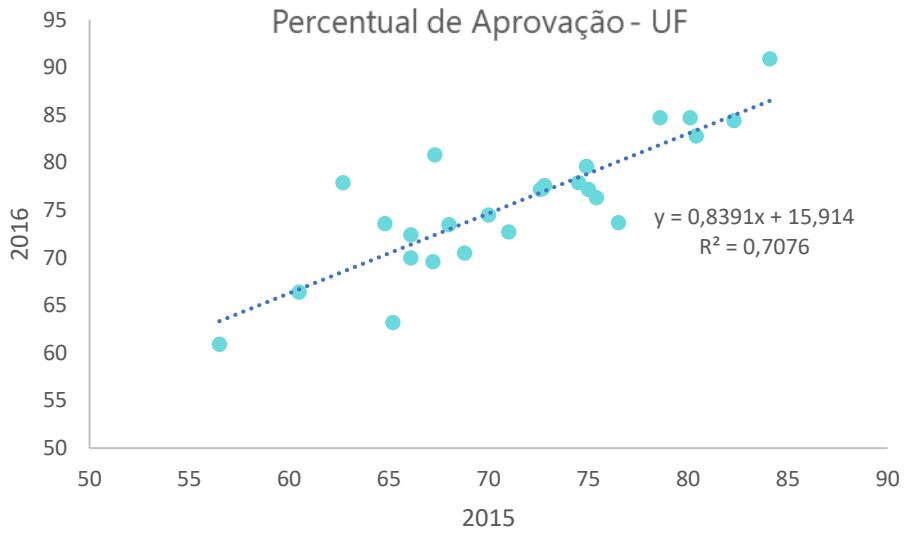
# Interpretação do modelo



No Excel, é possível incluir uma linha de tendência, e ele fornece a estimação dos parâmetros do modelo e o valor $R^2$.

**$R^2$** é o **coeficiente de determinação**, que pode ser calculado pelo **quadrado do coeficiente de correlação**. Ele é interpretado como o % da variabilidade explicada da variável y pela x.

Quanto maior o valor de $R^2$, mais bem ajustado é o modelo de regressão. Valores de $R^2$ acima de 0,5 já indicam bom ajuste, sendo que $0 < R^2 < 1$.




INTERPRETAÇÃO:

✓ **$b_0$ é 15,91**: quando o percentual de aprovados em 2015 é zero, em 2016 é 15,91.

✓ **$b_1$ é 0,84**: quando aumenta 1 p.p. no percentual de aprovação no ano de 2015, aumenta em 0,84 o percentual de aprovação no ano de 2016.

✓ **$R^2$ é 0,71:** 71% da variabilidade do percentual de aprovados de 2016 é explicado pelo percentual de aprovados de 2015, indicando um bom ajuste do modelo aos dados.

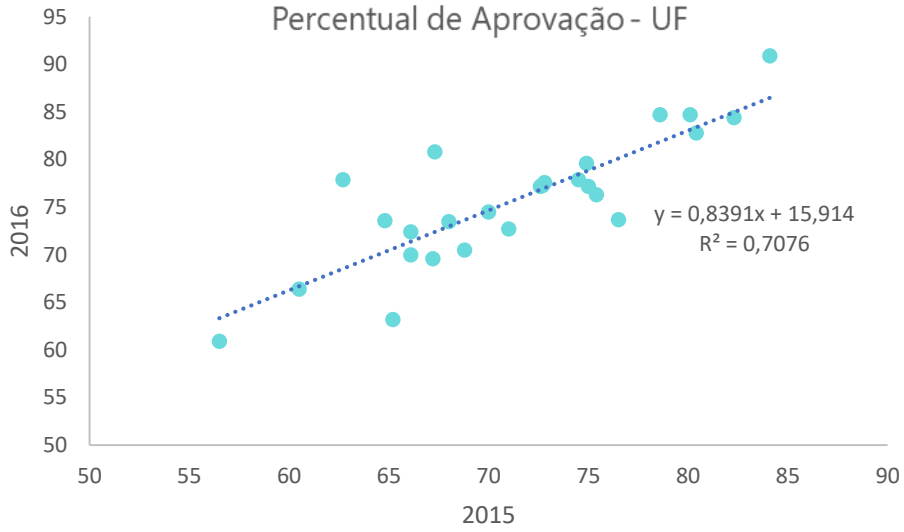
Base "Captacao_Alunos.txt" e código R em "Regressão linear simples.xlsx", aba "captação_alunos – R"

# Exercício: Predição por meio do modelo



Percentual de aprovados em 2016 = 15,914 + 0,8391*Percentual de aprovados em 2015.

O gestor percebeu que o modelo foi ajustado sem o Estado do Acre, uma vez que os dados de 2016 não vieram preenchidos. Seria possível predizer o valor do percentual de aprovados do Estado do Acre para 2016, dado que em 2015 o percentual de aprovação foi de 71,6?




Pelo modelo, a predição para o AC do percentual de aprovação em 2016 é de **75,99**.

Equação da reta

$$Y = b_0 + b_1 X$$

Y= 15,914 + 0,8391* X

Base “Captacao_Alunos.txt” e código R em “Regressão linear simples.xlsx”, aba “captação_alunos – R”

# Predição por meio do modelo



| Estado | 2015 | 2016 | Modelo | Erro |
|---|---|---|---|---|
| Alagoas | 66,1 | 72,4 | 71,4 | 1,0 |
| Amapa | 68,8 | 70,5 | 73,6 | -3,2 |
| Amazonas | 78,6 | 84,7 | 81,9 | 2,8 |
| Bahia | 67,2 | 69,6 | 72,3 | -2,7 |
| Ceara | 80,4 | 82,8 | 83,4 | -0,6 |
| DF | 74,9 | 79,6 | 78,8 | 0,8 |
| Espirito Santo | 70,0 | 74,5 | 74,7 | -0,2 |
| Goias | 80,1 | 84,7 | 83,1 | 1,5 |
| M. G. do Sul | 64,8 | 73,6 | 70,3 | 3,3 |
| Maranhao | 74,5 | 77,9 | 78,4 | -0,6 |
| Mato Grosso | 56,5 | 60,9 | 63,3 | -2,5 |
| Minas Gerais | 75,0 | 77,2 | 78,8 | -1,7 |
| Para | 68,0 | 73,5 | 73,0 | 0,5 |
| Paraiba | 71,0 | 72,7 | 75,5 | -2,8 |
| Parana | 75,4 | 76,3 | 79,2 | -2,9 |
| Pernambuco | 84,1 | 90,9 | 86,5 | 4,4 |
| Piaui | 72,7 | 77,3 | 76,9 | 0,4 |
| R. G. do Norte | 66,1 | 70,0 | 71,4 | -1,4 |
| R. G. do Sul | 65,2 | 63,2 | 70,6 | -7,5 |
| Rio de Janeiro | 76,5 | 73,7 | 80,1 | -6,4 |
| Rondonia | 67,3 | 80,8 | 72,4 | 8,4 |
| Roraima | 72,8 | 77,6 | 77,0 | 0,6 |
| Santa Catarina | 62,7 | 77,9 | 68,5 | 9,3 |
| Sao Paulo | 82,3 | 84,4 | 85,0 | -0,6 |
| Sergipe | 60,5 | 66,4 | 66,7 | -0,3 |
| Tocantins | 72,6 | 77,2 | 76,8 | 0,3 |

A média dos erros é igual a ZERO, pois a reta é ajustada de tal forma que fique centralizada aos dados.

ERRO = Valor observado - Valor ajustado

# Formalização do modelo teórico



$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad , \quad i=1,...,n$$

Em que

$Y_i$ : é o valor associado a i-ésima observação da variável resposta;

$\beta_0$ e $\beta_1$ : são parâmetros;

$X_i$ : é o valor associado a i-ésima observação da variável explicativa;

$\varepsilon_i$ : é o erro (resíduo) aleatório associado a i-ésima observação;

n : número de observações.

**Suposições do modelo:**

Sendo a variável X fixada (não está sujeita a variações aleatórias):

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

$$\varepsilon_i \sim N(0,\sigma^2)$$

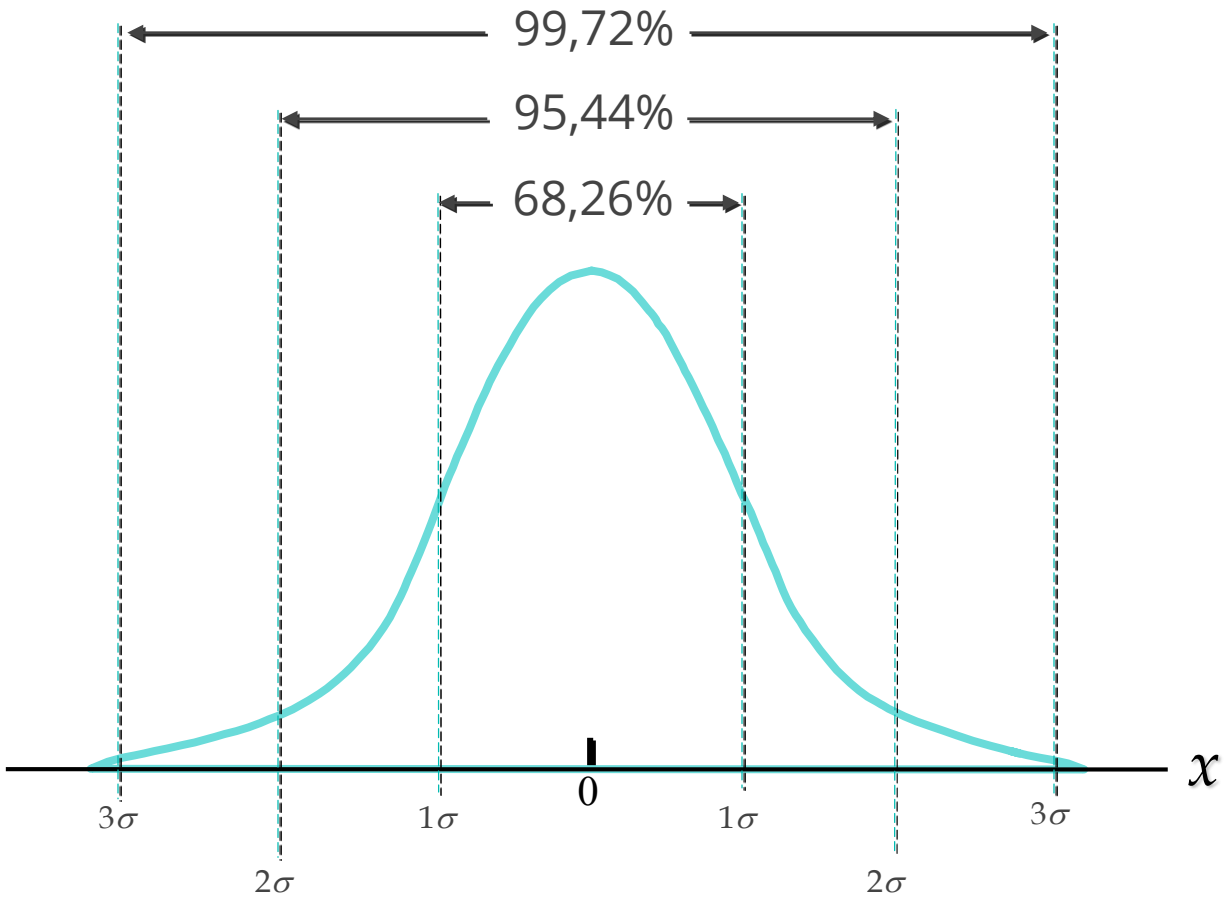
# Distribuição Normal



**Distribuição Normal (Gaussiana) dos resíduos**

Fazer um histograma dos resíduos e verificar:

- Simetria
- Distribuição dos dados na proporção ao lado e ao redor da média.



# Hipóteses sob os parâmetros



Modelo de Regressão Linear Simples

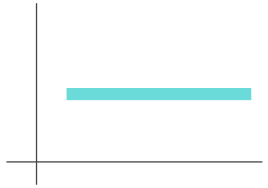
**Modelo teórico**

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

**Hipóteses de interesse sobre $\beta_1$**

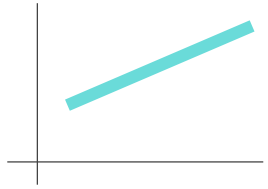
$H_0: \beta_1 = 0$

(não existe relação linear entre as variáveis)



$H_1: \beta_1 \neq 0$

(existe relação linear entre as variáveis)



## Reta de regressão e coeficiente de determinação

### Interpretação do output do R

```
Call:lm(formula = Y2016 ~ X2015, data = dados_rls)

Residuals: Min      1Q     Median      3Q     Max
          -7.4526 -2.2583 -0.2446  0.9462  9.3451

Coefficients: Estimate Std. Error    t     value Pr(>|t|)
(Intercept)   15.9139  7.8860      2.018   0.0549 .
X2015         0.8391   0.1101      7.621 7.36e-08 ***
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.782 on 24 degrees of freedom
Multiple R-squared:  0.7076,         Adjusted R-squared:  0.6954
F-statistic: 58.09 on 1 and 24 DF,  p-value: 7.362e-08
```

Estatísticas descritivas associadas aos resíduos.

0,0000000736

**Testa a hipótese da existência de relação linear entre as variáveis.**

Ao considerar 90 % de confiança:

O p-valor (0,0000000736) < 0,10 indica que existe relação linear entre as variáveis.

Coeficiente de determinação → $R^2$ é 0,71: 71% da variabilidade do percentual de aprovados de 2016 é explicado pelo percentual de aprovados de 2015, indicando um bom ajuste do modelo aos dados.

# Exercício: Faça a predição para o ano de 2017



Faça a predição do percentual de aprovados no Ensino Médio do Mato Grosso e Sergipe para o ano de 2017.

Percentual de aprovados = 15,9139 + 0,8391 * (Percentual de aprovados do ano anterior)

| Estado | 2016 |
|---|---|
| Mato Grosso | 60,9 |
| Sergipe | 66,4 |

# Case: Predição de preço de imóvel



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 as informações sobre o valor do imóvel (R$ mil) por $m^2$, a distância para estação de metrô (km), a quantidade de comércios próximos, e a idade (anos) do imóvel, em um bairro bem localizado de um grande centro urbano. Um cliente à procura de um imóvel faz questão de morar perto do metrô. Explique para o cliente se existe a relação entre preço do imóvel e localização próxima a estação de metrô.

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

| Id_Imovel | Idade_imovel | Distancia_metro_Km | Comercios_proximos | Mil_reais_m2 |
|---|---|---|---|---|
| 1 | 32 | 1,083595131 | 10 | 7,58 |
| 2 | 19,5 | 1,396946429 | 9 | 8,44 |
| 3 | 13,3 | 1,544788954 | 5 | 9,46 |
| 4 | 13,3 | 1,544788954 | 5 | 10,96 |
| 5 | 5 | 1,456009608 | 5 | 8,62 |
| 6 | 7,1 | 1,874980478 | 3 | 6,42 |
| 7 | 34,5 | 1,570122315 | 7 | 8,06 |
| 8 | 20,3 | 1,381344189 | 6 | 9,34 |
| 9 | 31,7 | 2,101860788 | 1 | 3,76 |
| 10 | 17,9 | 1,826514149 | 3 | 4,42 |

RStudio

Arquivo “Regressão linear simples.xlsx”

i. Base de dados na aba “Imobiliario”
ii. Código R na aba “Imobiliario – R”

# Case: Predição de preço de imóvel



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 as informações sobre o valor do imóvel (R$ mil) por $m^2$, a distância para estação de metrô (km), a quantidade de comércios próximos, e a idade (anos) do imóvel, em um bairro bem localizado de um grande centro urbano. Um cliente à procura de um imóvel faz questão de morar perto do metrô. Explique para o cliente se existe a relação entre preço do imóvel e localização próxima a estação de metrô.

Com base nos outputs apresentados, responda às perguntas abaixo:

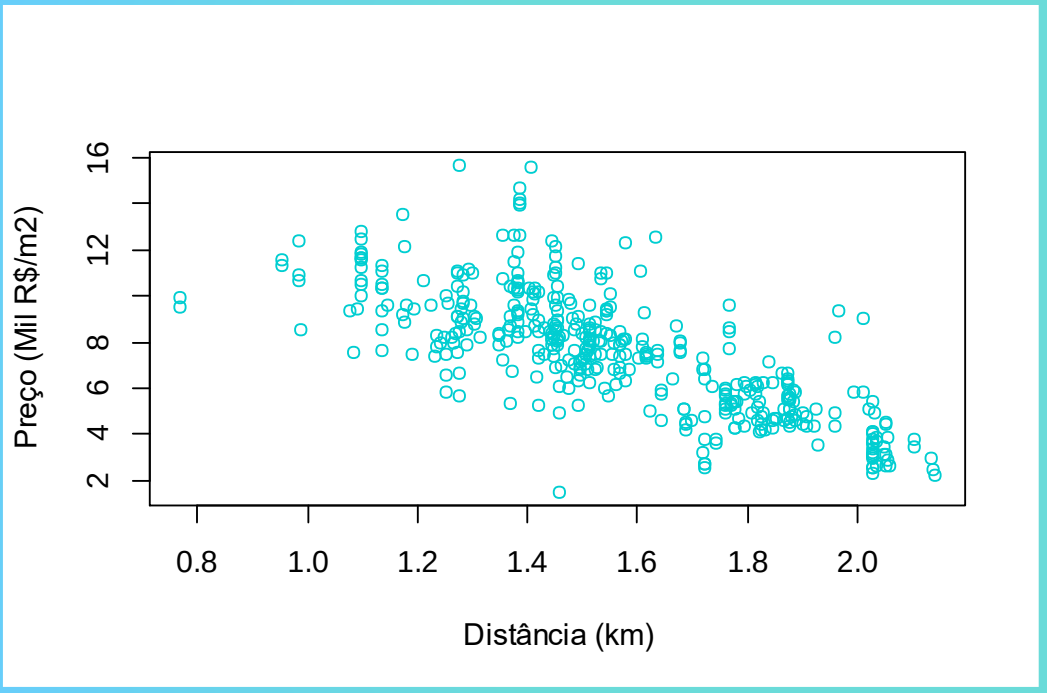
(a) Faça o gráfico de dispersão entre as variáveis Distancia_metro_Km e Mil_reais_m2. Existe relação entre preço do imóvel e distância para o metrô? É positiva ou negativa?
(b) Calcule o coeficiente de correlação linear entre as duas variáveis e interprete o coeficiente.
(c) Por meio do modelo de regressão linear simples, teste se existe relação linear entre as variáveis considerando 90% de confiança.
(d) Interprete os parâmetros do modelo e o coeficiente de determinação.
(e) Apresente a equação do modelo estimada.
(f) Estime o valor do $m^2$ do imóvel caso o cliente desejasse morar a 1 km do metrô.
(g) Para este cliente que deseja morar a 1km do metrô, estime o valor a ser pago em um apartamento de 70 $m^2$.

RStudio

Arquivo “Regressão linear simples.xlsx”
i. Base de dados na aba “Imobiliario”
ii. Código R na aba “Imobiliario – R”

# Case: Predição de preço de imóvel






Correlação = -0,76

# Case: Predição de preço de imóvel



**Output do modelo de Regressão Linear Simples**

```
Call:
lm(formula = Mil_reais_m2 ~ Distancia_metro_Km, data = imobiliario)

Residuals:
    Min      1Q  Median      3Q     Max
-6.7759 -0.9554 -0.1587  0.7327  6.9331

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         18.8154     0.4882   38.54   <2e-16 ***
Distancia_metro_Km  -7.2166     0.3082  -23.41   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.71 on 411 degrees of freedom
Multiple R-squared:  0.5715,         Adjusted R-squared:  0.5705
F-statistic: 548.2 on 1 and 411 DF,  p-value: < 2.2e-16
```

# Case: Predição da rentabilidade média de um município



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 rentabilidade 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 |
| 15,9 | 18,7 |
| 9,2 | 26,8 |

Arquivo “Regressão linear simples.xlsx”

i. Base de dados na aba “Rentabilidade”
ii. Código R na aba “Rentabilidade – R”

# Case: Predição da rentabilidade média de um município



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.


Siga as seguintes instruções para solução do case:

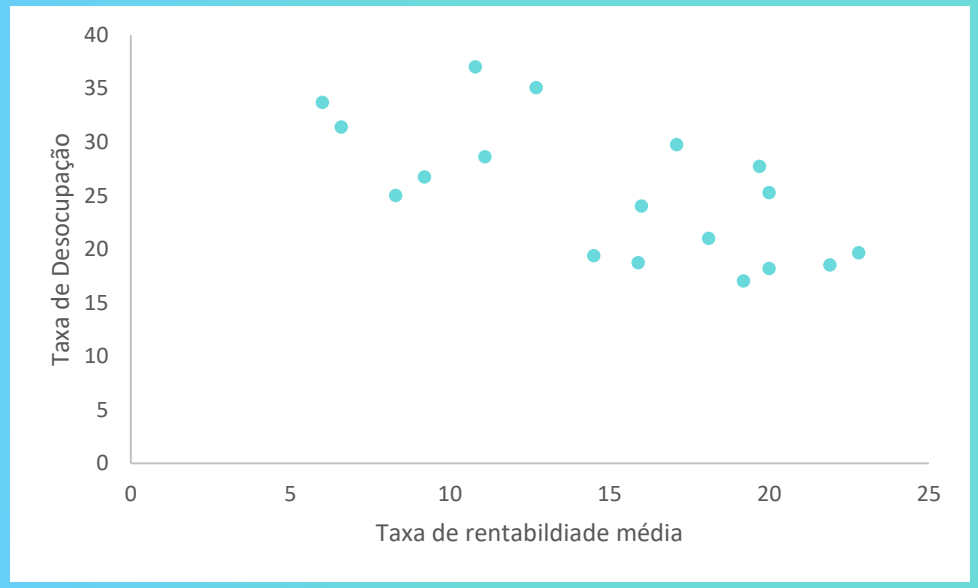
(a) Calcule o coeficiente de correlação linear entre as variáveis taxa de desocupação e taxa de rentabilidade média. A correlação é positiva ou negativa? É uma correlação forte?
(b) Obtenha o modelo de regressão linear simples. Com 90% de confiança, há relação linear entre as variáveis?
(c) Interprete os parâmetros do modelo e o coeficiente de determinação.
(d) Apresente a equação do modelo estimada.
(e) Estime o valor da rentabilidade média da população de um município que possui 15% de taxa de desocupação.

R Studio

Arquivo “Regressão linear simples.xlsx”
i. Base de dados na aba “Rentabilidade”
ii. Código R na aba “Rentabilidade – R”

# Case: Predição da rentabilidade média de um município






Correlação = -0,66

# Case: Predição da rentabilidade média de um município



**Output do modelo de Regressão Linear Simples**

```
Call:
lm(formula = `Taxa de rentabilidade média` ~ `Taxa de Desocupação`,
    data = rentabilidade)

Residuals:
    Min      1Q  Median      3Q     Max
-6.3669 -3.2553 -0.5402  3.1664  8.3702

Coefficients:
                     Estimate Std. Error t value     Pr(>|t|)
(Intercept)           37.0747     3.5277  10.510 0.0000000137 ***
`Taxa de Desocupação`  -0.7792     0.2224  -3.504      0.00294 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.885 on 16 degrees of freedom
Multiple R-squared:  0.4341,         Adjusted R-squared:  0.3988
F-statistic: 12.28 on 1 and 16 DF,  p-value: 0.00294
```

# Case: Predição do faturamento



Uma empresa de e-commerce deseja projetar o faturamento de acordo com o investimento em anúncios na internet. Os dados são de uma pesquisa de empresas do mesmo segmento do e-commerce. O objetivo é entender a variação no faturamento bruto semanal pelo canal de internet, de acordo com o investimento realizado neste canal.

| Investimento (R$) | Faturamento Bruto (R$) |
|---|---|
| 10000 | 41950 |
| 12500 | 52480 |
| 13700 | 57500 |
| 14800 | 62100 |
| 15700 | 65000 |
| 16500 | 69200 |
| 18600 | 78100 |
| 19200 | 80600 |
| 20500 | 86000 |
| 11000 | 46950 |
| 13500 | 57480 |
| 14700 | 62500 |
| 15800 | 67100 |

Arquivo “Regressão linear simples.xlsx”

i. Base de dados na aba “Faturamento”
ii. Código R na aba “Faturamento – R”

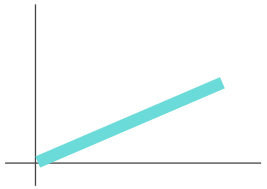
# Case: Predição do faturamento



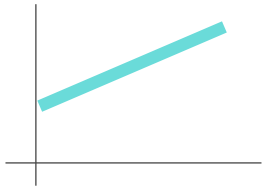
Uma empresa de e-commerce deseja projetar o faturamento de acordo com o investimento em anúncios na internet. Os dados são de uma pesquisa de empresas do mesmo segmento do e-commerce. O objetivo é entender a variação no faturamento bruto semanal pelo canal de internet, de acordo com o investimento realizado neste canal.

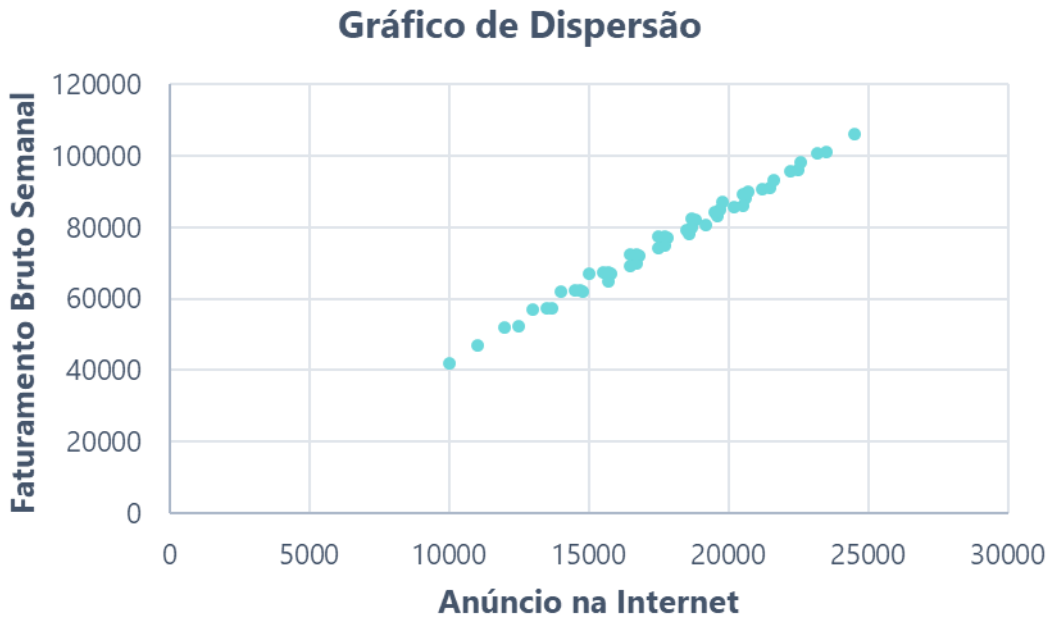
Com base no gráfico de dispersão, você acredita que a relação entre as variáveis segue a opção 1 ou a opção 2?

Opção 1



Opção 2






RStudio

Arquivo “Regressão linear simples.xlsx”

i. Base de dados na aba “Faturamento”
ii. Código R na aba “Faturamento - R”

# Hipóteses sob os parâmetros



Modelo de Regressão Linear Simples

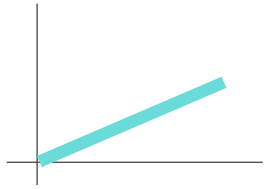
**Modelo teórico**

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

**Hipóteses de Interesse sob $\beta_0$**

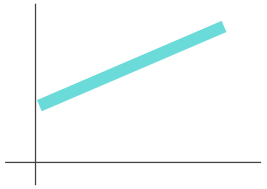
$H_0$: $\beta_0 = 0$

(passa pela origem (x = 0, y = 0))



$H_1$: $\beta_0 \neq 0$

(não passa pela origem (x = 0, y = $\beta_0$))



# Case: Predição do faturamento



Uma empresa de e-commerce deseja projetar o faturamento de acordo com o investimento em anúncios na internet. Os dados são de uma pesquisa de empresas do mesmo segmento do e-commerce. O objetivo é entender a variação no faturamento bruto semanal pelo canal de internet, de acordo com o investimento realizado neste canal.

```
Call:
lm(formula = `Faturamento Bruto Semanal` ~ `Anúncio na Internet`,
    data = faturamento)

Residuals:
     Min       1Q   Median       3Q      Max
-2346.53  -989.90   -24.64   854.51  2635.59

Coefficients:
                      Estimate Std. Error t value             Pr(>|t|)
(Intercept)           -659.66926  955.41039   -0.69                0.493
`Anúncio na Internet`    4.33161    0.05278   82.07 <0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1261 on 48 degrees of freedom
Multiple R-squared:  0.9929,         Adjusted R-squared:  0.9928
F-statistic:  6736 on 1 and 48 DF,  p-value: < 0.00000000000000022
```

Ao considerar 90 % de confiança:

O p-valor (0,493) > 0,10 indica que o intercepto não deve fazer parte do modelo e deve ser retirado.

# Case: Predição do faturamento



Uma empresa de e-commerce deseja projetar o faturamento de acordo com o investimento em anúncios na internet. Os dados são de uma pesquisa de empresas do mesmo segmento do e-commerce. O objetivo é entender a variação no faturamento bruto semanal pelo canal de internet, de acordo com o investimento realizado neste canal.

```
Call:
lm(formula = `Faturamento Bruto Semanal` ~ 0 + `Anúncio na Internet`,
    data = faturamento)

Residuals:
     Min       1Q   Median       3Q      Max
-2444.13 -1028.81   -10.45   926.77  2512.94

Coefficients:
                     Estimate Std. Error t value            Pr(>|t|)
`Anúncio na Internet` 4.295804   0.009798   438.4 <0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1254 on 49 degrees of freedom
Multiple R-squared:  0.9997,    Adjusted R-squared:  0.9997
F-statistic: 1.922e+05 on 1 and 49 DF,  p-value: < 0.00000000000000022
```

Este comando retira o intercepto do modelo

Quando o Beta 0 for igual a 0, deve-se ajustar um novo modelo sem o Beta 0.

**Faturamento = 4,295804 * Anúncio na Internet**

# Case: Predição do faturamento



Uma empresa de e-commerce deseja projetar o faturamento de acordo com o investimento em anúncios na internet. Os dados são de uma pesquisa de empresas do mesmo segmento do e-commerce. O objetivo é entender a variação no faturamento bruto semanal pelo canal de internet, de acordo com o investimento realizado neste canal.

Devemos remover o intercepto do modelo somente quando fizer sentido para o problema de negócio.

# Case: Predição do faturamento



Uma empresa de e-commerce deseja projetar o faturamento de acordo com o investimento em anúncios na internet. Os dados são de uma pesquisa de empresas do mesmo segmento do e-commerce. O objetivo é entender a variação no faturamento bruto semanal pelo canal de internet, de acordo com o investimento realizado neste canal.

Siga as seguintes instruções para solução do case:

(a) Obtenha o modelo de regressão linear simples. Com 90% de confiança, há relação linear entre as variáveis?
(b) Interprete os parâmetros do modelo e o coeficiente de determinação.
(c) Apresente a equação do modelo estimada.
(d) Estime o valor do faturamento para uma empresa que investe R$12.000 na internet.

RStudio

Arquivo “Regressão linear simples.xlsx”
i. Base de dados na aba “Faturamento”
ii. Código R na aba “Faturamento - R”

# Exercício: *People Analytics* - Salário



Um recrutador deseja estimar o salário de um candidato, a partir da nota média de várias provas realizadas durante o processo seletivo de admissão na empresa. O objetivo é ajudar os gestores a atribuir o salário do candidato dentro do intervalo já estipulado pela política de remuneração da empresa.

| Nota Média | Salario Mensal (R$) |
|---|---|
| 2,60 | 2800 |
| 3,40 | 3100 |
| 3,60 | 3500 |
| 3,20 | 3000 |
| 3,50 | 3400 |
| 2,90 | 3100 |
| 3,60 | 2900 |
| 4,40 | 3200 |
| 4,60 | 3600 |
| 4,20 | 3100 |
| 4,50 | 3500 |

Arquivo “Regressão linear simples.xlsx”

i. Base de dados na aba “Salario”
ii. Código R na aba “Salario - R”

# Exercício: *People Analytics* - Salário



Um recrutador deseja estimar o salário de um candidato, a partir da nota média de várias provas realizadas durante o processo seletivo de admissão na empresa. O objetivo é ajudar os gestores a atribuir o salário do candidato dentro do intervalo já estipulado pela política de remuneração da empresa.

Siga as seguintes instruções para solução do case:

(a) Obtenha o gráfico de dispersão entre as variáveis.
(b) Calcule o coeficiente de correlação entre as variáveis. É uma correlação positiva ou negativa? É uma correlação forte?
(c) Obtenha o modelo de regressão linear simples. Com 90% de confiança, há relação linear entre as variáveis?
(d) Interprete os parâmetros do modelo e o coeficiente de determinação.
(e) Apresente a equação do modelo estimada.
(f) Estime o valor do salário para um candidato que possui a nota média igual a 7.

RStudio

Arquivo "Regressão linear simples.xlsx"
i. Base de dados na aba "Salario"
ii. Código R na aba "Salario - R"

# 4. Regressão Linear Múltipla

# Case: Limite de Cartão de Crédito



**Exemplo**
Predizer o valor do limite do cartão de crédito em função da renda do cliente, do tempo de relacionamento e da idade.

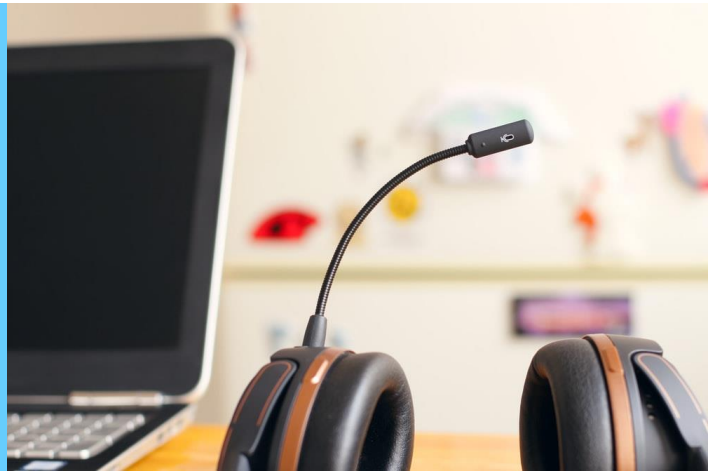
**Aplicação**
Área de Crédito do Segmento Bancário (Emissores de cartão de crédito).

# Case: SAC em Empresas de Serviço



**Exemplo**
Predizer o valor a ser investido em uma central de atendimento - SAC de uma empresa de serviços com base na quantidade de clientes, na quantidade de reclamações e no número de funcionários.

**Aplicação**
Área de Ouvidoria de empresas de serviços (Telecom, Bancos, Seguradoras, etc.)



# Case: Educação



**Exemplo**
Predizer o percentual de rematrículas em uma escola de Idiomas com base nas notas dos alunos do ano anterior, na idade dos alunos e na renda do chefe da família.

**Aplicação**
Áreas de Marketing e Vendas de Instituição de Ensino.

# Case: Venda de Seguros



**Exemplo**
Predizer o faturamento de um time Comercial com base na quantidade de vendedores ativos, tempo médio de experiência, tamanho da carteira de clientes e região.

**Aplicação**
Área de Planejamento Comercial.

# Case: Venda de Eletrônicos pela Internet



**Exemplo**
Predizer o volume (R$) de vendas em eletrônicos em função do investimento (R$) em Mídia Digital (Facebook, Instagram, Mídia Programática, *Search*), diversidade de produtos, região, etc.

**Aplicação**
Área de Mídias Digitais.

# Case: Covid-19



**Exemplo**
Predizer o valor gasto pelas prefeituras para o tratamento dos infectados de COVID-19 numa certa região com base no tamanho da população, no número de pessoas no grupo de risco e na renda da população.

**Aplicação**
Área de Saúde Pública.

# Forma geral do modelo



O modelo de regressão linear múltipla teórica é dado por:

$$Y=\beta_0+ \beta_1X_1+...+ \beta_pX_p+ \varepsilon \qquad \text{com, } \varepsilon \sim N(0,\sigma^2)$$

Y: variável dependente.

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

$\varepsilon$ : erro aleatório associado ao modelo.

A equação de regressão linear múltipla estimada é dada por:

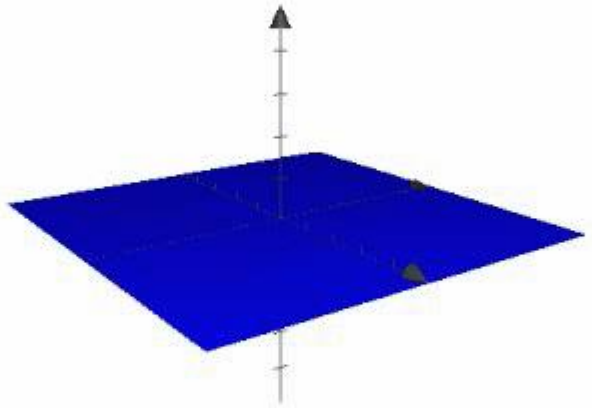
$$\hat{y} = b_0 + b_1x_1 + b_2x_2 + \ldots + b_px_p$$

# Forma geral do modelo



Exemplo: Modelo tridimensional (Y, $X_1$ e $X_2$)

$y=b_0+ b_1X_1+ b_2X_2$



https://commons.wikimedia.org/wiki/File:2d_multiple_linear_regression.gif

# Case: Predição de Limite de Cheque Especial



Uma instituição financeira tem o objetivo de estimar o valor de **Limite de Cheque Especial** para seus novos clientes, com base em informações disponíveis em seu banco de dados. Para o estudo, foi disponibilizada uma amostra histórica de clientes com as informações de **Idade, Rendimento Total, Salário** e **Limite de Crédito Imediato** para investigar se é possível estimar o Limite do Cheque Especial. Avalie a possibilidade de fornecer uma "regra" por meio de um modelo estatístico, interprete como as informações predizem o evento de interesse e qual a performance desta "regra".


Fonte: Base de dados inspirada em cases reais.

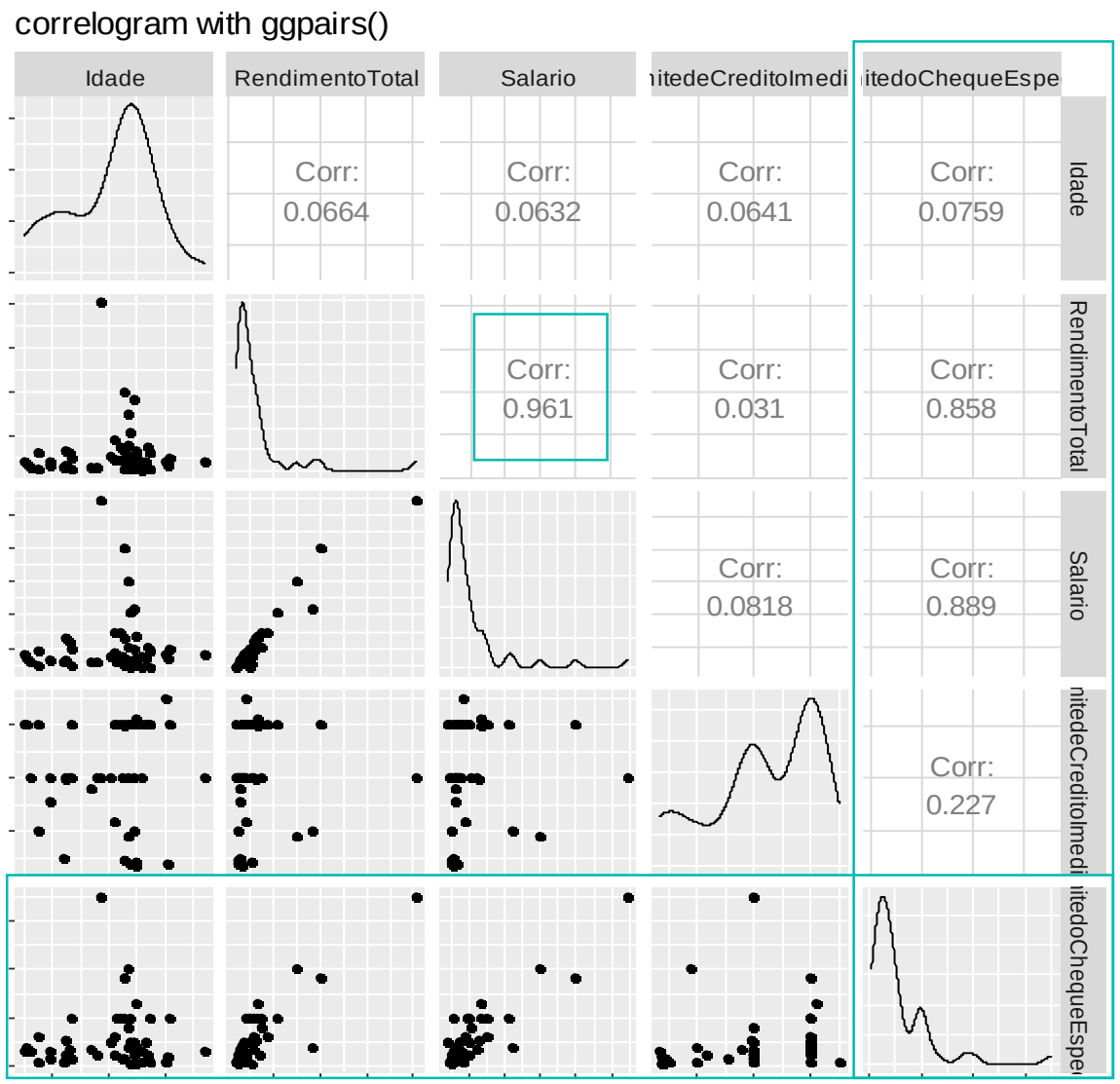
| Idade | RendimentoTotal | Salario | LimitedeCreditoImediato | LimitedoChequeEspecial |
|---|---|---|---|---|
| 72 | 4300 | 4300 | 2000 | 1000 |
| 75 | 4400 | 4400 | 3000 | 1000 |
| 66 | 4800 | 4800 | 440 | 1500 |
| 35 | 5000 | 5000 | 1000 | 1000 |
| 69 | 5000 | 5000 | 2000 | 2500 |
| 47 | 5000 | 5000 | 2000 | 1700 |
| 68 | 5000 | 5000 | 380 | 600 |
| 44 | 5800 | 5800 | 500 | 800 |
| 54 | 6000 | 6000 | 1790 | 3600 |

Arquivo "Regressão linear múltipla.xlsx"

i. Base de dados na aba "Limite_Credito (1)"
ii. Código R na aba "Limite_Credito - R (1)"

# Case: Predição de Limite de Cheque Especial






Antes ajustar o modelo, deve-se percorrer os passos a seguir.

**Passo 1:** Fazer a análise exploratória univariada.

**Passo 2:** Fazer a análise bidimensional (ou bivariada) da resposta vs variável explicativa para investigar as relações lineares ou não, e investigar o quanto as covariáveis auxiliariam na explicação da resposta.

```
library(GGally)
ggpairs(dados_lim_cred, title="correlogram with ggpairs()")
```

**Passo 3:** Fazer a análise bidimensional das covariáveis entre si para identificar correlação entre elas utilizando a correlação de Pearson e gráfico de dispersão.

# Relação entre as variáveis explicativas



- A **multicolinearidade** refere-se à correlação entre as variáveis explicativas do modelo.
- Quando as variáveis explicativas são altamente correlacionadas, não é possível determinar o efeito separado de uma particular variável explicativa sobre o comportamento da variável resposta.
- Quando a multicolinearidade é grave, pode ocorrer troca do sinal de alguns parâmetros do modelo. Neste caso, os coeficientes individuais tornam-se questionáveis.
- Uma forma de avaliar o efeito da multicolinearidade entre duas variáveis é retirar uma das covariáveis do modelo e avaliar a alteração do valor do $R^2$-ajustado. Recomenda-se manter no modelo a covariável que maximizar o $R^2$-ajustado.

# Case: Predição de Limite de Cheque Especial



**Output do Regressão Linear Múltipla – SEM RENDIMENTO TOTAL**

```
Call:
lm(formula = LimitedoChequeEspecial ~ Idade + Salario + LimitedeCreditoImediato,
    data = dados_lim_cred)

Residuals:
    Min      1Q  Median      3Q     Max
-7078.6 -1302.6  -220.6  1047.9  6201.9

Coefficients:
                          Estimate  Std. Error t value            Pr(>|t|)
(Intercept)            -2886.78469  1876.25053  -1.539              0.1311
Idade                      4.32030    26.58314   0.163              0.8716
Salario                    0.57528     0.04292  13.402 <0.0000000000000002 ***
LimitedeCreditoImediato    1.01120     0.42704   2.368              0.0223 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2732 on 44 degrees of freedom
Multiple R-squared:  0.8141,        Adjusted R-squared:  0.8014
F-statistic: 64.23 on 3 and 44 DF,  p-value: 0.0000000000000004093
```

Realizar o processo de redução, iniciando pela variável com maior nível descritivo, e rodar o modelo novamente.

### Hipótese de Interesse

$H_0$: $\beta_1 = \beta_2 = \ldots = \beta_p = 0$

$H_1$: $\beta_1 \neq 0$ **ou** $\beta_2 \neq 0$ **ou .... ou** $\beta_p \neq 0$

Testa a hipótese de que existe relação linear de pelo menos uma variável explicativa com a variável resposta. Quando este valor for < 0,10 concluímos que existe relação linear de pelo menos uma variável explicativa em relação à variável resposta.

# Case: Predição de Limite de Cheque Especial



**Output do Regressão Linear Múltipla – SEM RENDIMENTO TOTAL** → Ajuste o modelo sem a variável rendimento total que apresentou alta correlação com o salário.

```
Call:
lm(formula = LimitedoChequeEspecial ~ Idade + Salario + LimitedeCreditoImediato,
    data = dados_lim_cred)

Residuals:
    Min      1Q  Median      3Q     Max
-7078.6 -1302.6  -220.6  1047.9  6201.9

Coefficients:
                          Estimate  Std. Error t value            Pr(>|t|)
(Intercept)            -2886.78469  1876.25053  -1.539               0.1311
Idade                      4.32030    26.58314   0.163               0.8716
Salario                    0.57528     0.04292  13.402 <0.0000000000000002 ***
LimitedeCreditoImediato    1.01120     0.42704   2.368               0.0223 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2732 on 44 degrees of freedom
Multiple R-squared:  0.8141,         Adjusted R-squared:  0.8014
F-statistic: 64.23 on 3 and 44 DF,  p-value: 0.0000000000000004093
```

Na regressão linear múltipla, pode-se ter várias variáveis com nível descritivo superior a 0,10. Neste caso, deve-se escolher um método para seleção de variáveis.

Neste case, será utilizado o método *Backward* de eliminação de variáveis.

# Case: Predição de Limite de Cheque Especial



**Output do Regressão Linear Múltipla – SEM IDADE**

```
Call:
lm(formula = LimitedoChequeEspecial ~ Salario + LimitedeCreditoImediato,
    data = dados_lim_cred)

Residuals:
    Min      1Q  Median      3Q     Max
-7043.8 -1313.0  -249.4  1028.4  6209.2

Coefficients:
                          Estimate  Std. Error t value            Pr(>|t|)
(Intercept)            -2636.56049  1060.62552  -2.486               0.0167 *
Salario                    0.57568     0.04239  13.582 <0.0000000000000002 ***
LimitedeCreditoImediato    1.01532     0.42165   2.408               0.0202 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2703 on 45 degrees of freedom
Multiple R-squared:  0.814, Adjusted R-squared:  0.8057
F-statistic: 98.47 on 2 and 45 DF,  p-value: < 0.00000000000000022
```

Adotando nível de significância de 0,10, todos os parâmetros são diferentes de zero.

Coeficiente de Determinação Ajustado : usar para RL Múltipla.

# Case: Predição de Limite de Cheque Especial



## Interpretação do Modelo Final (escolhido)

Limite do Cheque Especial = -2.636,56049 + 0,5757*Salário + 1,015*Limite de Crédito Imediato

$R^2$-ajustado: 0,8057

Interpretação do coeficiente de regressão:

- 0,5757 é o aumento esperado do Limite do Cheque Especial correspondente ao aumento de 1 unidade no Salário, quando seu Limite de Crédito Imediato é considerado constante.
- Similarmente, 1,015 é o aumento esperado do Limite do Cheque Especial correspondente ao aumento de 1 unidade do Limite de Crédito Imediato, quando o Salário é mantido constante.

Interpretação do $R^2$-ajustado:

- 81% da variabilidade do Limite do Cheque Especial é explicada pelas variáveis Salário e Limite de Crédito Imediato pela Regressão Linear Múltipla.

# Case: Predição de Limite de Cheque Especial



## Exercício

Limite do Cheque Especial = -2.636,56049 + 0,5757*Salário + 1,015*Limite de Crédito Imediato

Com base no modelo ajustado, realizar a projeção para os 10 casos abaixo, descritos no excel “Regressão linear multipla.xls”, na aba “Limite_Credito - Exercício”.

| Idade | RendimentoTotal | Salario | LimitedeCreditoImediato | Limite Cheque Projetado |
|---|---|---|---|---|
| 63 | 8700 | 5700 | 3000 | |
| 62 | 8784 | 8784 | 1170 | |
| 65 | 8800 | 7300 | 2000 | |
| 47 | 10000 | 10000 | 3000 | |
| 69 | 10000 | 10000 | 3000 | |
| 82 | 10000 | 10000 | 3000 | |
| 68 | 10527 | 4027 | 3000 | |
| 70 | 10736 | 5214 | 400 | |
| 61 | 11000 | 7500 | 2000 | |
| 35 | 12000 | 9000 | 3000 | |

# Variáveis *dummies*



- Utilizamos o modelo de Regressão Linear, até agora, apenas para **covariáveis quantitativas**.
- Quando as covariáveis são **qualitativas**, é necessário transformar as características em variáveis indicadoras (*dummies*), atribuindo a presença ou não da característica.
- Exemplo:

| cliente | Escolaridade | EscolaridadeSuperior_Pos |
|---|---|---|
| 1 | Fundamental ou Medio | 0 |
| 2 | Superior ou Pos | 1 |
| 3 | Fundamental ou Medio | 0 |
| 4 | Superior ou Pos | 1 |

- No R, caso a variável seja qualitativa (*string*), o *software* já "entende" e faz a atribuição da primeira categoria (pela ordem alfabética) como sendo a **categoria de referência (recebe valor zero)**, e a categoria subsequente receberá o valor 1.

# Case: Predição de Limite de Cheque Especial



Uma instituição financeira tem o objetivo de estimar o valor de **Limite de Cheque Especial** para seus novos clientes, com base em informações disponíveis em seu banco de dados. Para o estudo, foi disponibilizada uma amostra histórica de clientes com as informações de **Idade, Rendimento Total, Salário, Limite de Crédito Imediato** e **Escolaridade** para investigar se é possível estimar o Limite do Cheque Especial. Avalie a possibilidade de fornecer uma "regra" por meio de um modelo estatístico, interprete como as informações predizem o evento de interesse e qual a performance desta "regra".


Fonte: Base de dados inspirada em cases reais.

| Idade | RendimentoTotal | Salario | LimitedeCreditoImediato | LimitedoChequeEspecial | Escolaridade |
|---|---|---|---|---|---|
| 70 | 10736 | 5214 | 400 | 500 | Fundamental_Medio |
| 68 | 5000 | 5000 | 380 | 600 | Fundamental_Medio |
| 44 | 5800 | 5800 | 500 | 800 | Fundamental_Medio |
| 72 | 4300 | 4300 | 2000 | 1000 | Fundamental_Medio |
| 75 | 4400 | 4400 | 3000 | 1000 | Fundamental_Medio |
| 35 | 5000 | 5000 | 1000 | 1000 | Fundamental_Medio |
| 80 | 8100 | 8100 | 3500 | 1000 | Fundamental_Medio |
| 66 | 4800 | 4800 | 440 | 1500 | Fundamental_Medio |
| 39 | 6320 | 6320 | 1550 | 1640 | Fundamental_Medio |

Nova variável

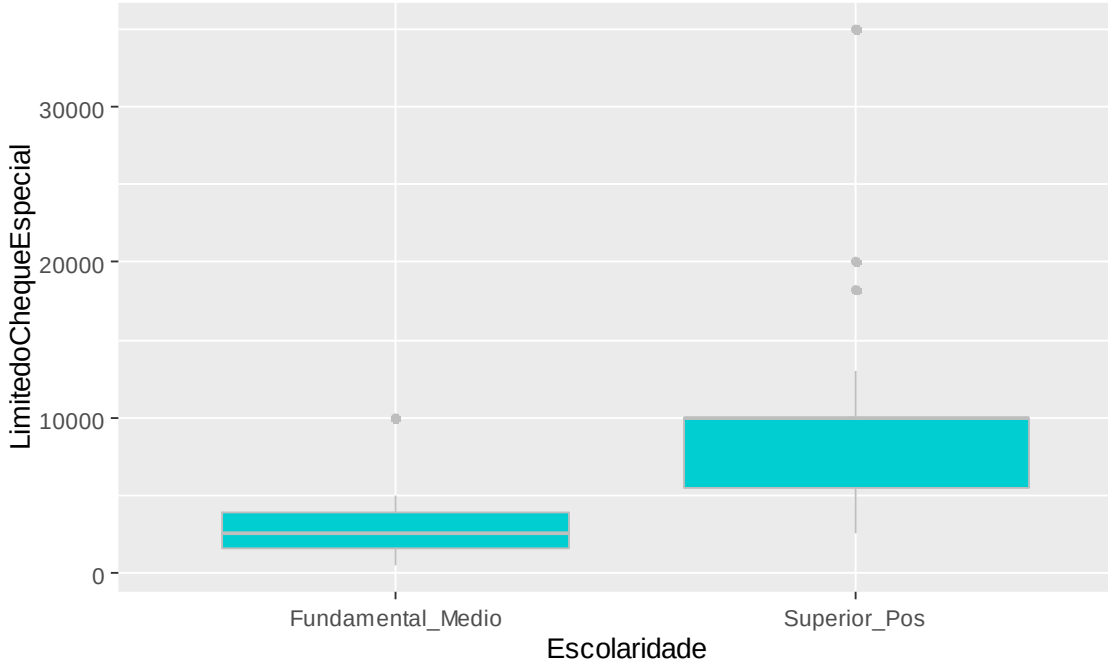
Arquivo "Regressão linear múltipla.xlsx"

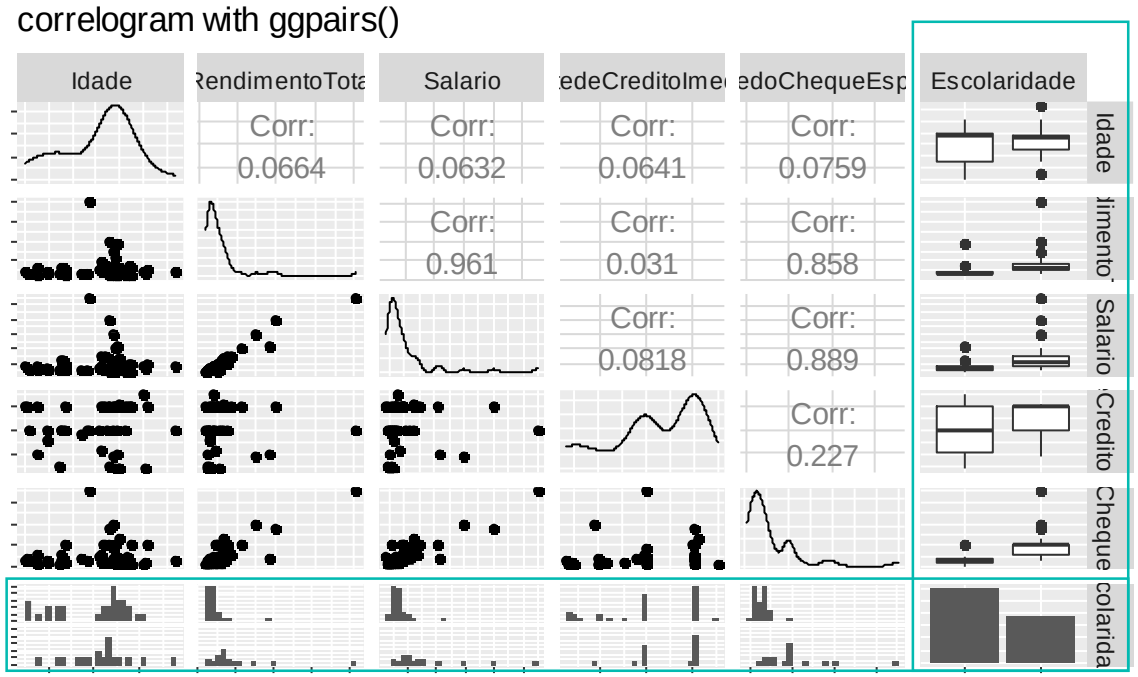
i. Base de dados na aba "Limite_Credito (2)"
ii. Código R na aba "Limite_Credito - R (2)"

# Case: Predição de Limite de Cheque Especial



Nunca esquecer de fazer AED da **variável nova**.




A variável **Escolaridade** parece discriminar, tendo em vista que os clientes com curso superior ou pós-graduação com valores maiores de limite de cheque especial.

Ela também mostra uma leve relação com as variáveis Idade e Limite de cheque imediato.

# Case: Predição de Limite de Cheque Especial



**Output do Regressão Linear Múltipla – COM Escolaridade**

```
Call:
lm(formula = LimitedoChequeEspecial ~ Salario + LimitedeCreditoImediato +
    Escolaridade, data = dados_lim_cred_esc)

Residuals:
    Min      1Q  Median      3Q     Max
-5785.9 -1014.5   -36.8   843.6  7077.9

Coefficients:
                           Estimate  Std. Error t value            Pr(>|t|)
(Intercept)               -2321.65980   996.37901  -2.330             0.02445 *
Salario                       0.51706     0.04487  11.524 0.00000000000000705 ***
LimitedeCreditoImediato       0.73219     0.40660   1.801             0.07860 .
EscolaridadeSuperior_Pos   2445.09099   883.23658   2.768             0.00821 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2522 on 44 degrees of freedom
Multiple R-squared:  0.8416,         Adjusted R-squared:  0.8308
F-statistic: 77.92 on 3 and 44 DF,  p-value: < 0.00000000000000022
```

Interpretação do coeficiente associado a **Escolaridade**: R$2.445 é o acréscimo esperado no valor atribuído ao Limite de Cheque Especial para os clientes com escolaridade Superior ou Pós, quando as demais covariáveis do modelo são mantidas constantes.

# Case: Predição de Limite de Cheque Especial



## Modelo Final (escolhido):

Limite do Cheque Especial = -2321,65980 + 0,51706*Salário + 0,73219*Limite de Crédito Imediato + 2445,09099* EscolaridadeSuperior_Pos

$R^2$-ajustado: 0,8308

# Variáveis *dummies*



- Quando há mais de duas categorias na variável qualitativa:

  Variável original: **Estado Civil**, com os valores: “solteiro”, “casado” e “outros”.

  Categoria de referência: “casado” (por ser a primeira, em ordem alfabética)

  Variáveis dummy criadas: **Est_civil_O**, com os valores 1 (“outros”) ou 0 (demais categorias).

  **Est_civil_S**, com os valores 1 (“solteiro”) ou 0 (demais categorias).

- Note que a quantidade de *dummies* é ‘quantidade de categorias – 1’.

# Variáveis *dummies*



- Exemplo:

| cliente | Sexo | Estado Civil | Sexo_M | Est_civil_O | Est_civil_S |
|---|---|---|---|---|---|
| 1 | feminino | solteiro | 0 | 0 | 1 |
| 2 | masculino | casado | 1 | 0 | 0 |
| 3 | feminino | outros | 0 | 1 | 0 |
| 4 | masculino | solteiro | 1 | 0 | 1 |
| 5 | masculino | solteiro | 1 | 0 | 1 |
| 6 | masculino | solteiro | 1 | 0 | 1 |
| 7 | feminino | casado | 0 | 0 | 0 |

# Case: Predição de Limite de Cheque Especial



Uma instituição financeira tem o objetivo de estimar o valor de **Limite de Cheque Especial** para seus novos clientes, com base em informações disponíveis em seu banco de dados. Para o estudo, foi disponibilizada uma amostra histórica de clientes com as informações de **Idade, Rendimento Total, Salário, Limite de Crédito Imediato**, **Escolaridade, Gênero** e **Região** para investigar se é possível estimar o Limite do Cheque Especial. Avalie a possibilidade de fornecer uma "regra" por meio de um modelo estatístico, interprete como as informações predizem o evento de interesse e qual a performance desta "regra".


Fonte: Base de dados inspirada em cases reais.

Siga as seguintes instruções para solução do case:

(a) Faça o gráfico de dispersão entre as variáveis.
(b) Calcule o coeficiente de correlação entre as variáveis quantitativas. Interprete os coeficientes.
(c) Obtenha o modelo de regressão linear múltipla e faça a seleção de variáveis pelo método *Backward*.
(d) Interprete os parâmetros do modelo e o coeficiente de determinação.
(e) Apresente a equação do modelo estimado.
(f) Estime o valor do limite de cheque especial para um cliente que tem salário de R$4.850, de São Paulo e gênero masculino.

RStudio®

Arquivo "Regressão linear múltipla.xlsx"
i. Base de dados em "Base Limite_Credito (3)"
ii. Código R em Limite_Credito - R (3)

# Case: Predição *Startups*



Um investidor deseja estimar o lucro de startups de acordo com suas características, com objetivo de tomada de decisão de investimento baseada no lucro projetado da empresa. A base de dados possui características de investimento e região das empresas já investidas do histórico do investidor.

| 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 |

RStudio

Arquivo “Regressão linear múltipla.xlsx”

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

# Case: Predição *Startups*



Um investidor deseja estimar o lucro de startups de acordo com suas características, com objetivo de tomada de decisão de investimento baseada no lucro projetado da empresa. A base de dados possui características de investimento e região das empresas já investidas do histórico do investidor.

Siga as seguintes instruções para solução do case:

(a) Calcule o coeficiente de correlação entre as variáveis quantitativas do banco de dados. Interprete os coeficientes.
(b) Realize a análise bidimensional entre as covariáveis existentes e investigue possíveis problemas de multicolinearidade.
(c) Obtenha o modelo de regressão linear múltipla, considerando um nível de significância de 10% para seleção de variáveis.
(d) Interprete os parâmetros do modelo e o coeficiente de determinação.
(e) Apresente a equação do modelo estimado.
(f) Calcule o lucro projetado de uma startup de São Paulo que apresenta R$40.000 com gastos em P&D e R$100.000 com gastos em marketing.

RStudio

Arquivo “Regressão linear múltipla.xlsx”
i. Base de dados em “Startups”
ii. Código R em “Startups - R”

# Case: Predição de consumo de cerveja



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 mensal de cerveja (em litros) nesta cidade, com base em características de outras cidades do Estado. Cada linha da base de dados representa uma cidade.

| Temperatura_Media | Precipitacao | População | Renda_Media | Consumo_de_cerveja |
|---|---|---|---|---|
| 27,3 | 0 | 38.300 | 3.640 | 14.343 |
| 27,0 | 0 | 51.840 | 4.740 | 14.940 |
| 24,8 | 0 | 50.580 | 4.775 | 16.228 |
| 24,0 | 1,2 | 54.180 | 4.110 | 16.748 |
| 23,8 | 0 | 46.570 | 3.405 | 16.956 |
| 23,8 | 12,2 | 25.470 | 3.060 | 16.977 |
| 24,0 | 0 | 24.840 | 4.895 | 17.075 |
| 24,9 | 48,6 | 11.200 | 4.930 | 17.241 |
| 28,2 | 4,4 | 36.970 | 2.565 | 17.287 |
| 26,8 | 0 | 25.400 | 4.130 | 17.399 |

Arquivo “Regressão linear múltipla.xlsx”

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

# Case: Predição de consumo de cerveja



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 mensal de cerveja (em litros) nesta cidade, com base em características de outras cidades do Estado. Cada linha da base de dados representa uma cidade.

Siga as seguintes instruções para solução do case:

(a) Calcule o coeficiente de correlação entre as variáveis quantitativas do banco de dados. Interprete os coeficientes.
(b) Realize a análise bidimensional entre as covariáveis existentes e investigue possíveis problemas de multicolinearidade.
(c) Obtenha o modelo de regressão linear múltipla, considerando um nível de significância de 10% para seleção de variáveis.
(d) Interprete os parâmetros do modelo e o coeficiente de determinação.
(e) Apresente a equação do modelo estimado.
(f) Projete o consumo de cerveja para cidades quem possuem precipitação de 2, população de 5.000 habitantes e renda média de R$1.500.

R Studio

Arquivo “Regressão linear múltipla.xlsx”
i. Base de dados em “Startups”
ii. Código R em “Startups - R”

# Case: Predição de preço de imóvel



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 $m^2$, a distância para estação de metrô (km), a quantidade de comércios próximos e a idade (em 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 |

Arquivo “Regressão linear múltipla.xlsx”

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

# Case: Predição de preço de imóvel



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 $m^2$, a distância para estação de metrô (km), a quantidade de comércios próximos e a idade (em 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

Siga as seguintes instruções para solução do case:

(a) Calcule o coeficiente de correlação entre as variáveis quantitativas do banco de dados. Interprete os coeficientes.
(b) Realize a análise bidimensional entre as covariáveis existentes e investigue possíveis problemas de multicolinearidade.
(c) Obtenha o modelo de regressão linear múltipla, considerando um nível de significância de 5% para seleção de variáveis.
(d) Interprete os parâmetros do modelo e o coeficiente de determinação.
(e) Apresente a equação do modelo estimado.
(f) Caso um comprador esteja procurando um imóvel a 1 km do metrô, com 5 comércios próximos e que tenha 5 anos, qual valor médio ele pagaria em um imóvel de 85 $m^2$?

RStudio

Arquivo “Regressão linear múltipla.xlsx”
i. Base de dados em “Imobiliario”
ii. Código R em “Imobiliario - R”

# 5. Código em R



# Case: Limite de Crédito



```
#utilizamos a função options(scipen=999) para evitar que o R
#imprima valores em notação científica

options(scipen=999)
```

O argumento **scipen=999** evita notações cientificas.

# Case: Limite de Crédito



```
#Carregar a base de dados: Regressão linear múltipla.xlsx
install.packages("readxl")
library(readxl)

dados_lim_cred  <- read_excel("Regressão linear
múltipla.xlsx",
                              sheet="Limite_Credito" )
```

Sempre que rodar um pacote pela primeira vez em seu computador, é necessário instalar o pacote **install.packages("readxl")**.

Após instalar o pacote, precisamos carregar a biblioteca utilizando **library**.

O primeiro argumento da função **read_excel** representa o nome do banco de dados, entre "aspas".

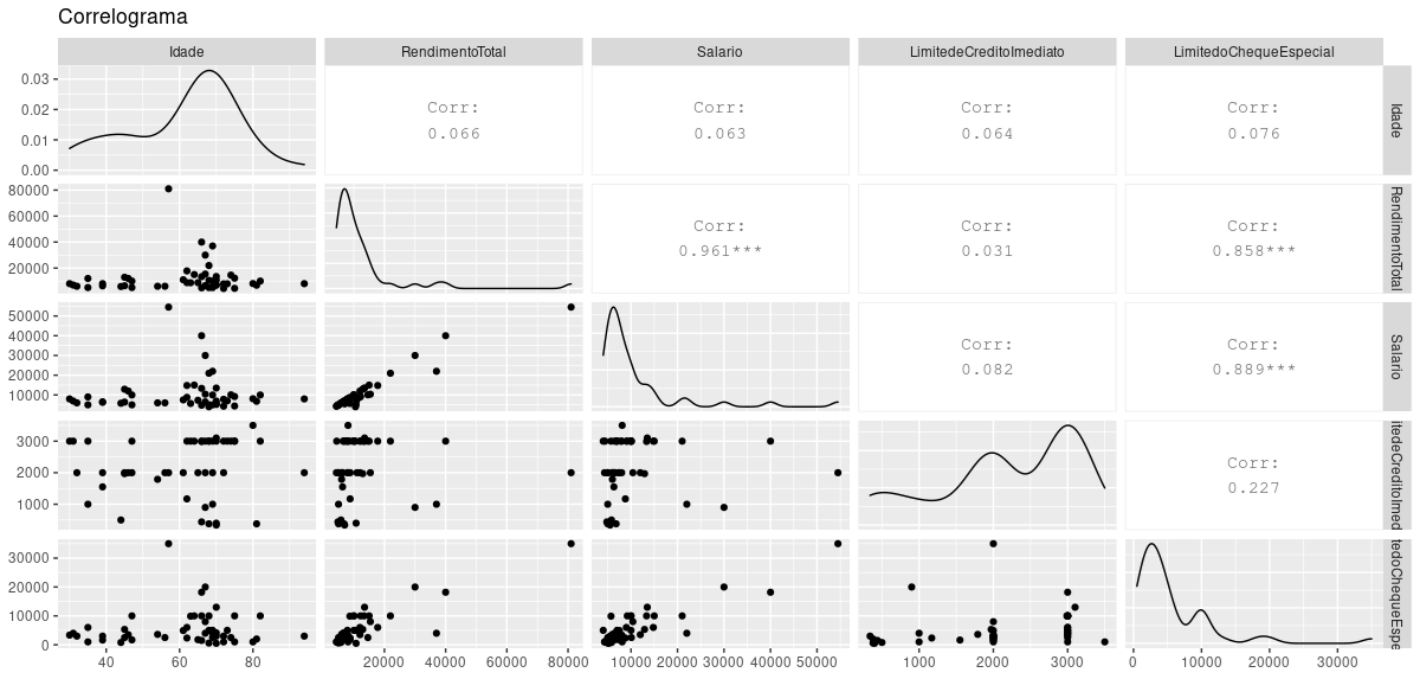
O argumento "**sheet=**" representa o nome da aba do arquivo, nome da planilha entre "aspas".

# Case: Limite de Crédito



```
#Matriz de Gráfico de Dispersão
#Gráfico de Dispersão

library(GGally)
ggpairs(dados_lim_cred, title="Correlograma")
```




A biblioteca **GGally** computa as métricas de correlação.

O R é *case sensitive*, atentar-se às letras maiúsculas e minúsculas.

A função **ggpairs** é responsável por imprimir o gráfico e o argumento **"title="** é utilizado para alterar o título do gráfico.

# Case: Limite de Crédito



```
#Regressão Linear Múltipla


regressao <- lm(data = dados_lim_cred,
                LimitedoChequeEspecial ~
                Idade + Salario +
LimitedeCreditoImediato)

summary(regressao)
```

A função **lm** (de *linear model)* ajusta o modelo de Regressão Linear.

**"data="** corresponde ao objeto contendo a base de dados.

O segundo argumento deve receber a variável resposta separada por "~" das variáveis explicativas. Por exemplo:

y ~ var1+ var2 + var3.

# Referência




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