Estatística Aplicada
EAD Ao Vivo

Tema da aula
Regressão Linear

LABDATA | FIA FUNDAÇÃO INSTITUTO DE ADMINISTRAÇÃO BUSINESS SCHOOL

02/09/2020

ESTATÍSTICA APLICADA EAD AO VIVO

Professora:
Drª Karin Ayumi Tamura

Coordenadores:
Profª Drª Alessandra de Álvila Montini
Profª Dr. Adolpho Walter Pimazoni Canton

Currículo - Prof.ª Drª Karin Ayumi Tamura

FORMAÇÃO ACADÊMICA | EXPERIÊNCIA PROFISSIONAL

Profª Dra.
Karin Ayumi Tamura

Contato: karin.tamura@fia.com.br

- **FORMAÇÃO ACADÊMICA:** Pós-doutora (2015), Doutora (2012), mestre (2007) e bacharel (2003) em Estatística pelo Instituto de Matemática e Estatística da USP, tendo como área de pesquisa modelos de regressão, análise multivariada de dados e algoritmos de *machine learning*.
- **ATUAÇÃO PROFISSINAL:** Foi *Head* de *Analytics* por 14 anos, e atualmente é Conselheira Executiva e *Head* de Inovação na *Marketdata Solutions*, uma empresa do grupo WPP, e Professora Doutora no LABDATA FIA.
- **HISTÓRICO:** Atuação no mercado por 17 anos, com experiência profissional no segmento bancário (Bradesco) e consultoria *(Marketdata Solutions)*. Atuou como docente em cursos de pós-graduação (2010-16) no LABDATA FIA e ABEMD. Especialista em Estatística e *Advanced Analytics* trabalhando em projetos de diversos segmentos do mercado. Participante de congressos nacionais e internacionais voltados a área de Estatística, Dados e Algoritmos de *Machine Learning*.

"Tenho duas paixões no meu trabalho: dados e pessoas. Voltar a lecionar no LABDATA FIA está sendo a realização de um sonho planejado desde a minha época de aluna de pós-graduação. Meu objetivo como professora é integrar a visão do mercado com as técnicas e tecnologias de análise de dados, por meio de uma atuação humanista no ensino aos alunos"

Projetos atendidos

FIA Business School

NOSSOS DIFERENCIAIS | QUEM SOMOS

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

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)

Conteúdo Programático do Curso

21 AULAS AO VIVO COM PROFA. KARIN | 27 PLANTÕES AO VIVO COM PROF. STEPHAN, 7 LISTAS DE EXERCÍCIOS E EAD VIDEO AULA EM PYTHON

Dia	Mês	Aula	EAD Ao Vivo	Plantão Prof. Stephan
5	Agosto	Introdução ao Curso e Análise Exploratória de Dados	Aula Prof. Karin	06/ago
12	Agosto	Análise Exploratória de Dados	Aula Prof. Karin	13/ago
19	Agosto	Análise Exploratória de Dados - Introdução ao R	Aula Prof. Karin	20/ago
26	Agosto	Lista de Exercícios em Sala de Aula (19hs-23hs - com presença obrigatória)	-	27/ago
2	Setembro	Regressão Linear Simples	Aula Prof. Karin	03/set
9	Setembro	Regressão Linear Simples e Múltipla	Aula Prof. Karin	10/set
16	Setembro	Regressão Linear Simples e Múltipla	Aula Prof. Karin	17/set
23	Setembro	Lista de Exercícios em Sala de Aula (19hs-23hs - com presença obrigatória)	-	24/set
30	Setembro	Análise de Cluster	Aula Prof. Karin	01/out
7	Outubro	Análise de Cluster	Aula Prof. Karin	08/out
14	Outubro	Lista de Exercícios em Sala de Aula (19hs-23hs - com presença obrigatória)	-	15/out
21	Outubro	Arvore de Decisão	Aula Prof. Karin	22/out
28	Outubro	Lista de Exercícios em Sala de Aula (19hs-23hs - com presença obrigatória)	-	29/out
4	Novembro	Regressão Logística	Aula Prof. Karin	05/nov
11	Novembro	Regressão Logística	Aula Prof. Karin	11/nov
18	Novembro	Lista de Exercícios em Sala de Aula (19hs-23hs - com presença obrigatória)	-	19/nov
25	Novembro	estudo de caso	Aula Prof. Karin	26/nov
2	Novembro	estudo de caso	Aula Prof. Karin	30/dez
9	Dezembro	estudo de caso	Aula Prof. Karin	10/dez
16	Dezembro	Análise de Série Temporal - modelo auto regressivo	Aula Prof. Karin	17/dez
23	Dezembro	Lista de Exercícios em Sala de Aula (Frequência Liberada - véspera Natal)	-	-
Recesso Escolar		EAD - INTRODUÇÃO AO PYTHON EAD - INTRODUÇÃO AO PYTHON	EAD Video Aula (8 horas)	- -
6	Janeiro	Modelos estatísticos em Python	Aula Prof. Karin	07/jan
13	Janeiro	Modelos estatísticos em Python	Aula Prof. Karin	14/jan
20	Janeiro	Modelos estatísticos em Python	Aula Prof. Karin	20/jan
27	Janeiro	Introdução a Big Data - Aplicações de Machine Learning e Deep Learning	Aula Prof. Karin	28/jan
3	Fevereiro	Aplicações de Machine Learning	Aula Prof. Karin	04/fev
10	Fevereiro	Aplicações de Machine Learning	Aula Prof. Karin	11/fev
17	Fevereiro	Lista de Exercícios (Frequência Liberada - quarta de cinzas)	-	18/fev
24	Fevereiro	EXERCICIOS DE REVISÃO - EAD (19hs e 23hs - com presença obrigatória)	-	24/fev
3	Março	Prova (Plataforma On Line: 19hs e 23hs)	-	

Conteúdo da Aula

1. Introdução
2. Coeficiente de correlação
3. Regressão Linear Simples
4. Regressão Linear Múltipla
   i. Multicolinearidade
   ii. Variáveis explicativas qualitativas
5. Exercícios para casa
   i. CASE: Limite de Crédito
   ii. CASE: Predição de valor de imóveis

1. Introdução

Case Limite de Cartão de Crédito

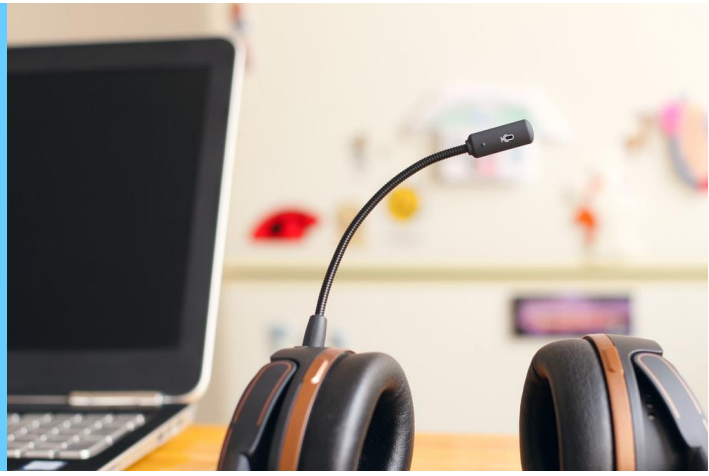
Exemplo
Predizer o valor do limite do cartão de crédito em função da renda do cliente.

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

Case SAC em Empresas de Serviço

Exemplo
Predizer a quantidade de chamados finalizados com sucesso no SAC de uma empresa de serviços com base na quantidade de atendentes contratados.

Aplicação
Área de Ouvidoria de empresas de serviços (p.e. Telcom, 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 a quantidade de vendas de Seguros do time Comercial com base na quantidade de corretores 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.

Exemplo
Predizer a quantidade de infectados de COVID-19 numa certa região com base nos dados de mobilidade.

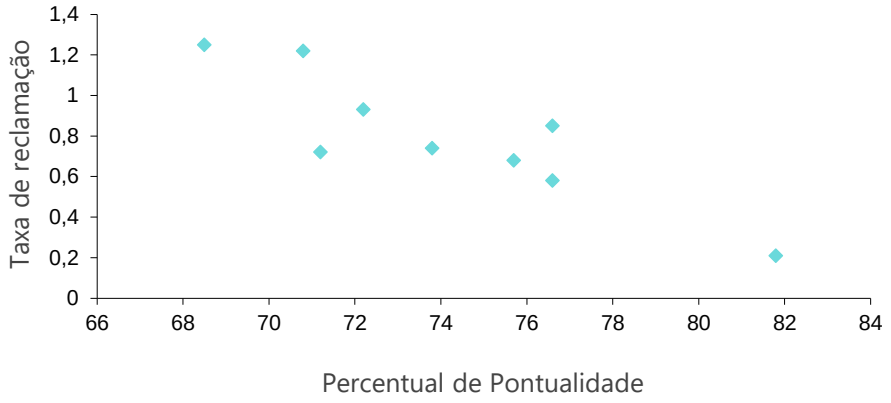
Aplicação
Área de Saúde Pública.

2. Coeficiente de correlação

Existe relação entre pontualidade dos voos e reclamação dos passageiros?

Uma pesquisa deseja estimar a taxa de reclamação em função do percentual de pontualidade das saídas dos voos de companhias aéreas. Existe relação entre as duas informações?

Linha aérea	Percentual de Pontualidade	Taxa de reclamação
1	81,8	0,21
2	76,6	0,58
3	76,6	0,85
4	75,7	0,68
5	73,8	0,74
6	72,2	0,93
7	71,2	0,72
8	70,8	1,22
9	68,5	1,25



Quanto maior o percentual de pontualidade, menor a taxa de reclamação.

Coeficiente de Correlação

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

O coeficiente r varia entre -1 a 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 associação linear entre as variáveis.

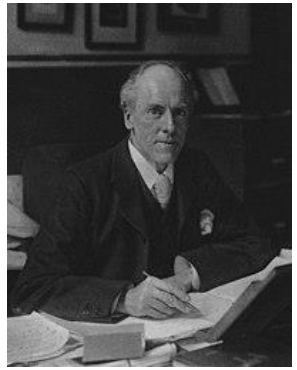
Covariância (x,y)

$$r = r_{xy} = \frac{cov(x,y)}{dp(x)dp(y)} = \frac{\sum(x-\bar{x})\,(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2 \sum(y-\bar{y})^2}}$$

Desvio padrão (x)

O coeficiente também é conhecido como CORRELAÇÃO DE PEARSON

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



Karl Pearson
(Londres, 1857-1936)

Vamos calcular a correlação linear?

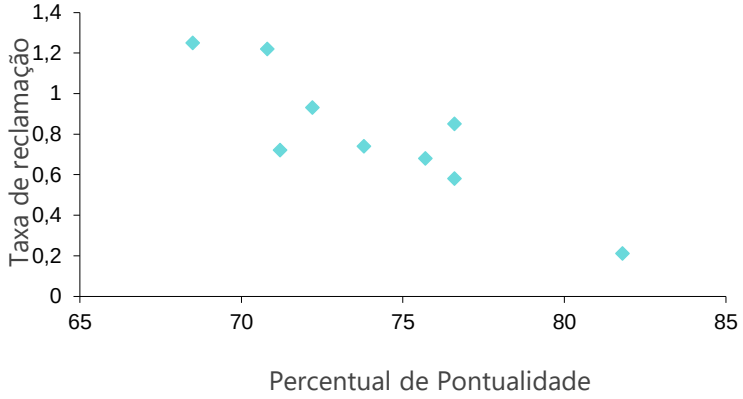
Uma pesquisa deseja estimar a taxa de reclamação em função do percentual de pontualidade das saídas dos voos de companhias aéreas. Existe relação entre as duas informações?

Vamos fazer juntos?

RStudio

Excel

Linha aérea	Percentual de Pontualidade	Taxa de reclamação
1	81,8	0,21
2	76,6	0,58
3	76,6	0,85
4	75,7	0,68
5	73,8	0,74
6	72,2	0,93
7	71,2	0,72
8	70,8	1,22
9	68,5	1,25

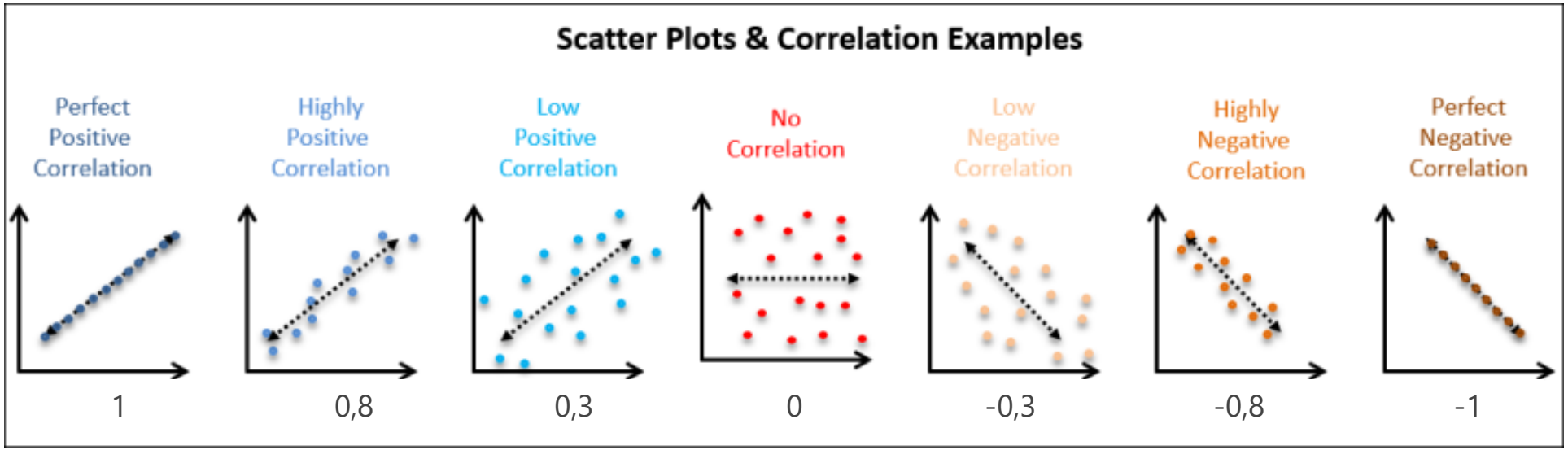


Existe uma forte associação NEGATIVA (r=-0,88) entre as duas variáveis, ou seja, quanto maior o percentual de pontualidade, menor a taxa de reclamação.

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

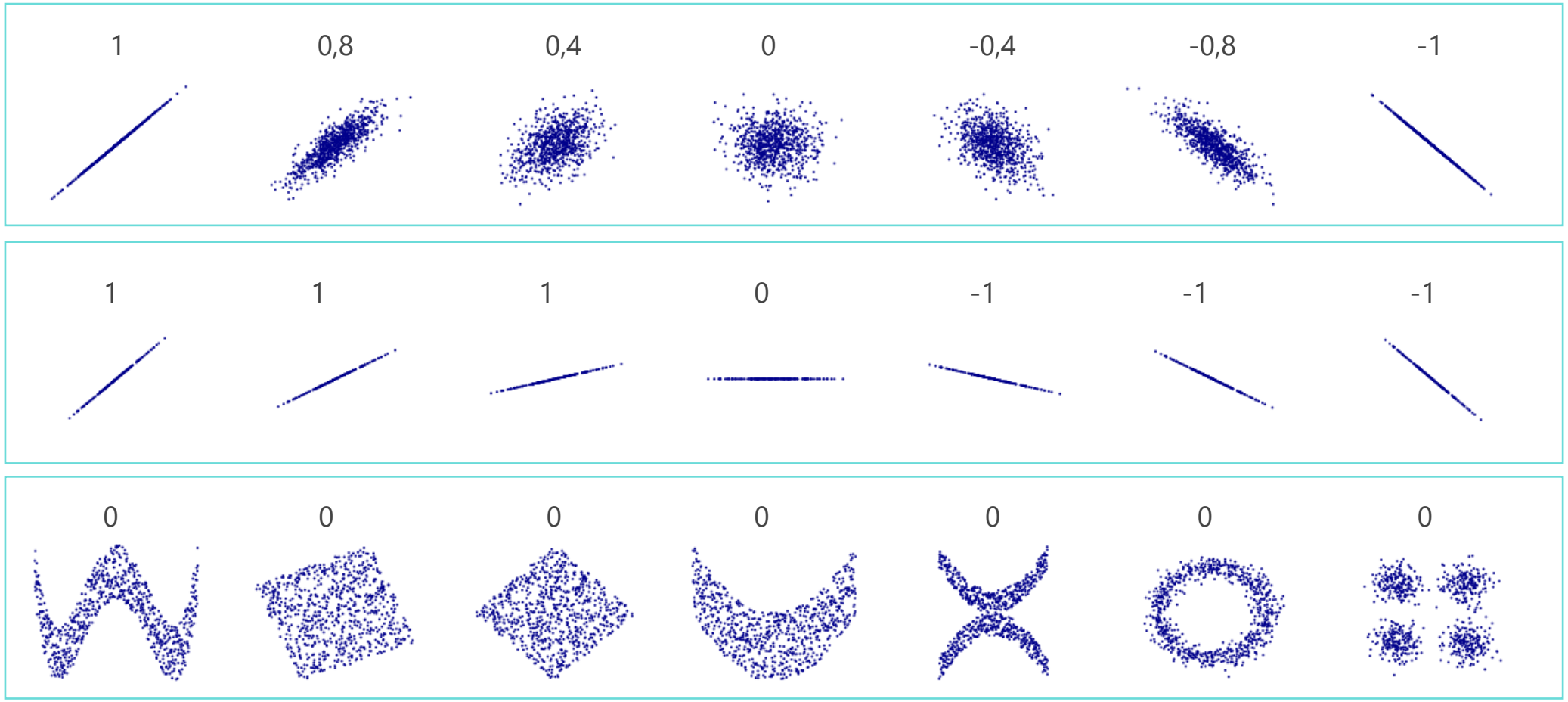
Interpretação dos valores de correlação

Na prática, consideramos valores acima de |r|>0,7 como alta correlação linear e valores abaixo de |r|<0,3 como baixa correlação linear.



https://lytongblog.wordpress.com/2018/12/21/correlation-between-two-variables/

Correlação de Pearson: somente relação LINEAR



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

2. Regressão Linear Simples

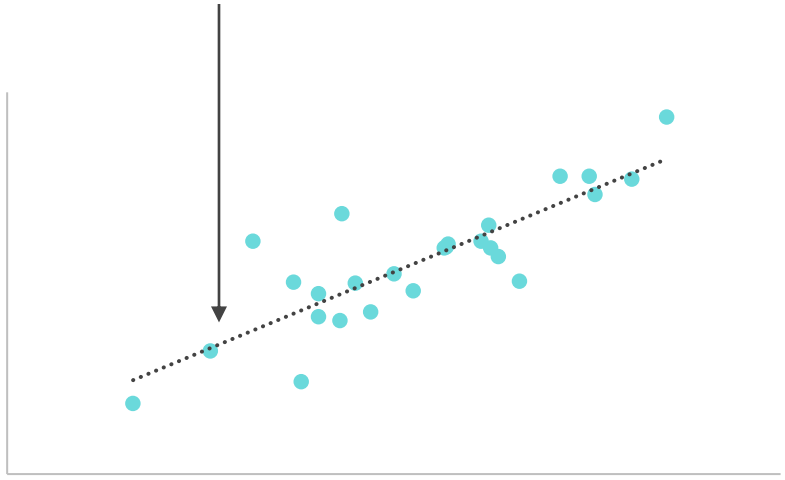
É possível expressar essa relação por meio de um modelo estatístico?

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

PERGUNTA DE NEGÓCIO:
Se aumentar o percentual de voos saídos com pontualidade, em quanto diminui a taxa de reclamação?

Relação entre DADO, MODELO E ERRO:

DADOS = MODELO + ERRO



Estrutura do modelo

O modelo de regressão linear simples é dado por:

y: variável resposta, variável dependente ou *target*

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

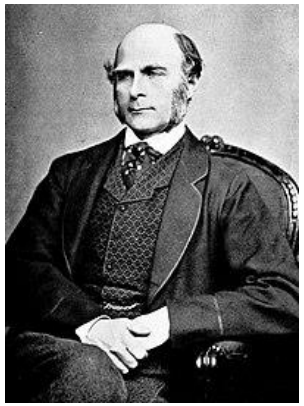
x: variável explicativa, variável auxiliar, variável independente ou covariável

Em que :

- ✓ β_0 e β_1 são chamados **parâmetros do modelo**.
- ✓ ε é uma variável aleatória chamada de erro ou resíduo.

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

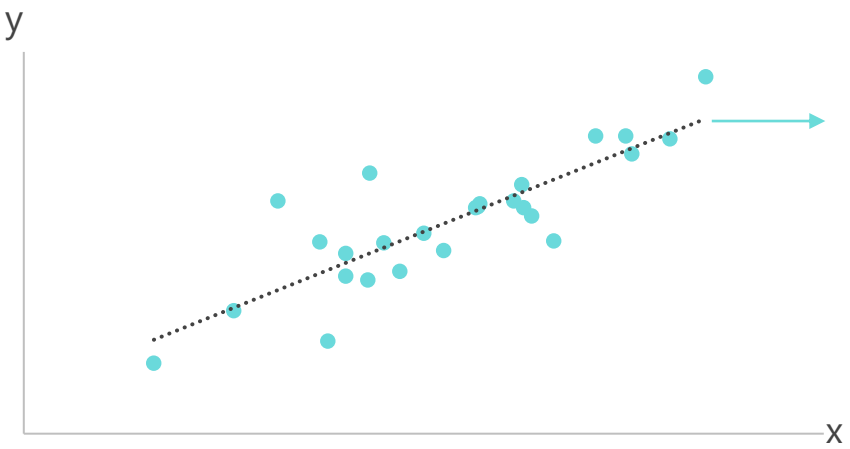
Ele estudava junto com seu discípulo, Karl Pearson, a relação entre a altura do pai com a altura do filho



Francis Galton
(Londres, 1822-1911)

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 estimados do modelo**.

✓ $\hat{y}$ é a resposta estimada.

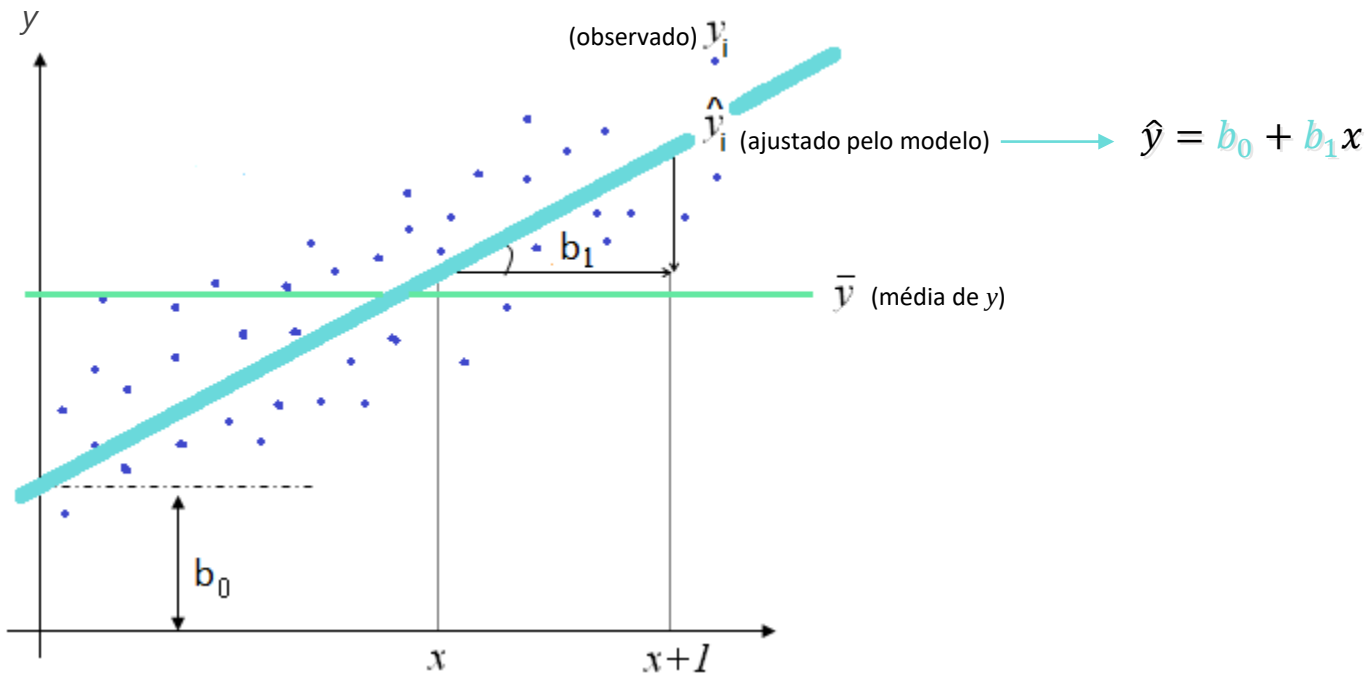
Sendo:

✓ b_0 é o valor do intercepto (x=0).

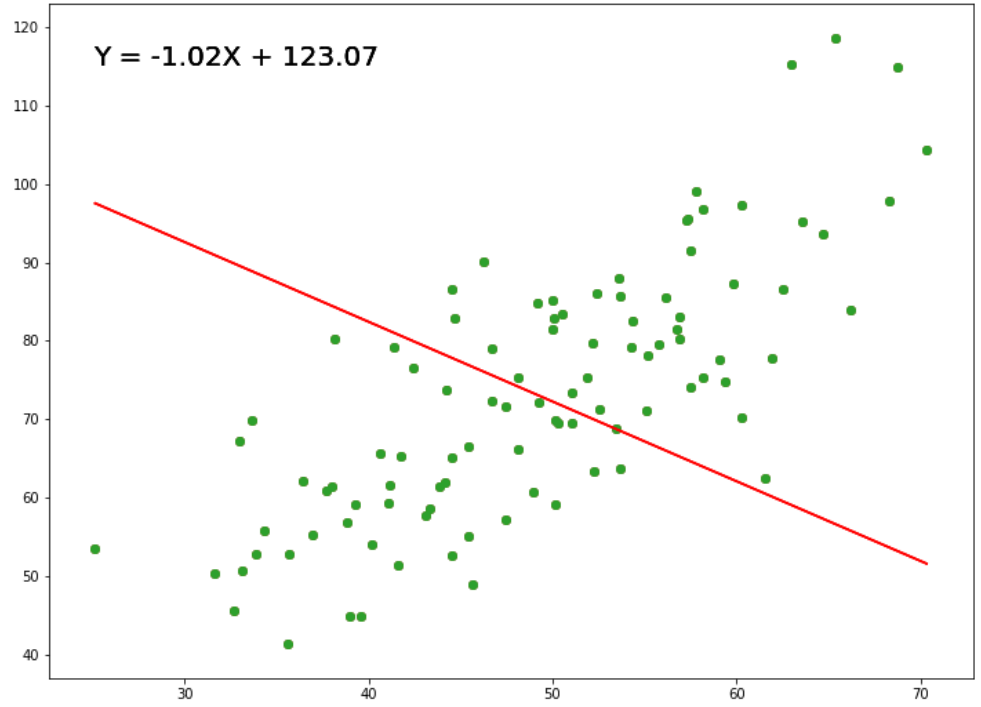
✓ b_1 é coeficiente angular (inclinação da reta).

Interpretação gráfica da equação

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



Animação - Valores de intercepto e inclinação



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

Modelo: teórico x ajustado

Modelo de Regressão Linear Simples
Modelo teórico

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

X

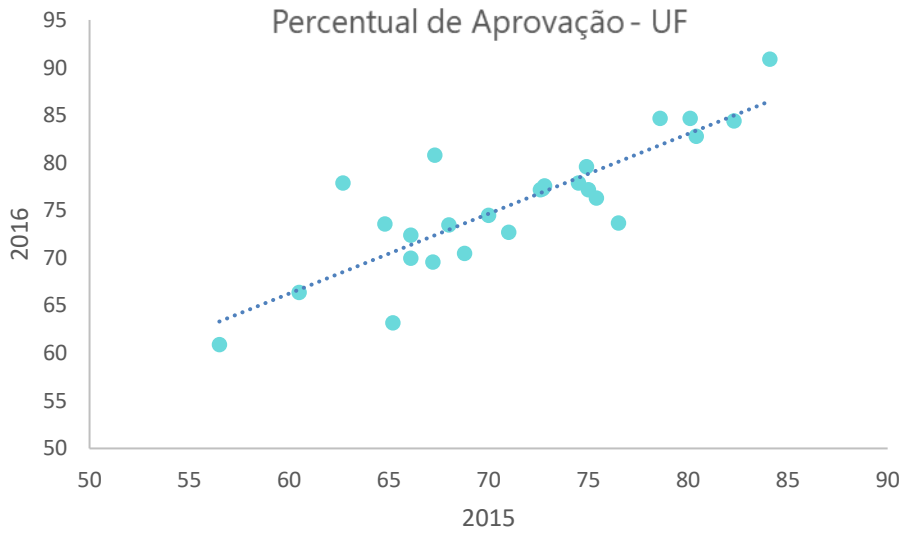
Modelo de Regressão Linear Simples
Modelo ajustado

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

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

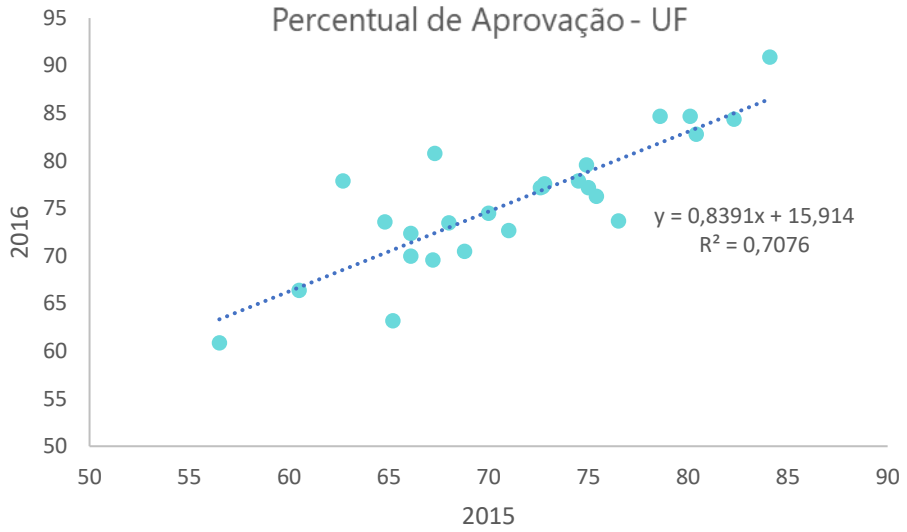


Existe uma forte associação POSITIVA (r=0,84) entre as duas variáveis, ou seja, os estados que apresentaram maior nota em 2015 também apresentaram maior nota em 2016.

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**. Quanto maior o valor de R^2, mais bem ajustado é o modelo regressão. Valores de R^2 acima de 0,5 já indicam bom ajuste, $0<R^2<1$. Ele pode ser interpretado com o % da variabilidade explicada da variável y pela x.



INTERPRETAÇÃO:

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

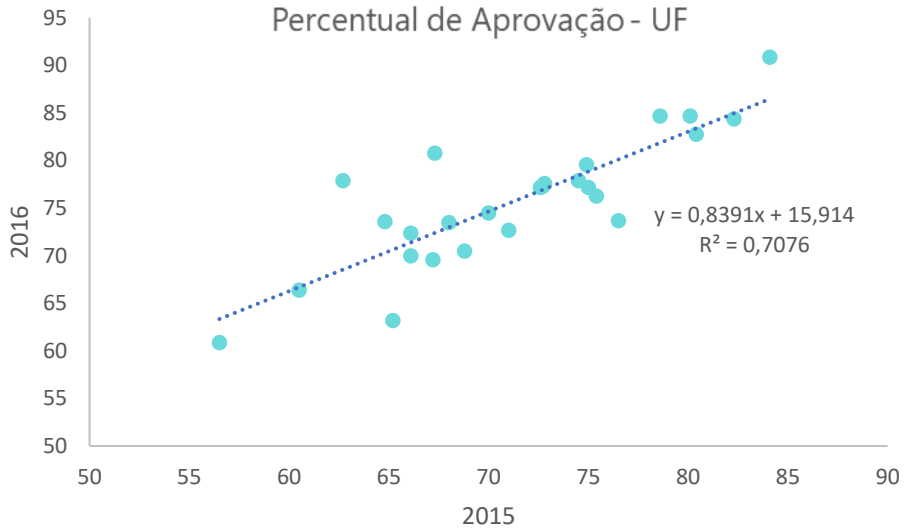
✓ **b_1 é 0,84**: quando aumenta 1 p.p. no percentual de aprovação no ano de 2015, aumenta em 0,84 o percentual de aprovação no ano de 2016.

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

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

Exercício: Faça a predição para o ano de 2017

Faça a predição do percentual de aprovados no Ensino Médio para ano corrente (2017).

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.

DADOS = **MODELO** + ERRO

ERRO = DADOS – **MODELO**

$ERRO_i = y_i$ (observado)- $\mathbf{y_i}$ **(predito)**

i-esimo indivíduo da amotra de dados, onde i=1, ..., n.

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;

β_0 e β_1 : são parâmetros;

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

ε_i : é o erro (resíduo) aleatório associado a i-ésima observação;

n : número de observações.

Suposições do modelo:

Sendo a variável X fixa ou determinística (não está sujeita a variações aleatórias):

1. A **média** dos **resíduos** é **zero.**
2. Os **resíduos** tem a **variabilidade constante** torno de X.
3. ε_i e ε_j são **não correlacionados**, com i ≠ j.
4. Os resíduos seguem uma **distribuição Normal**.

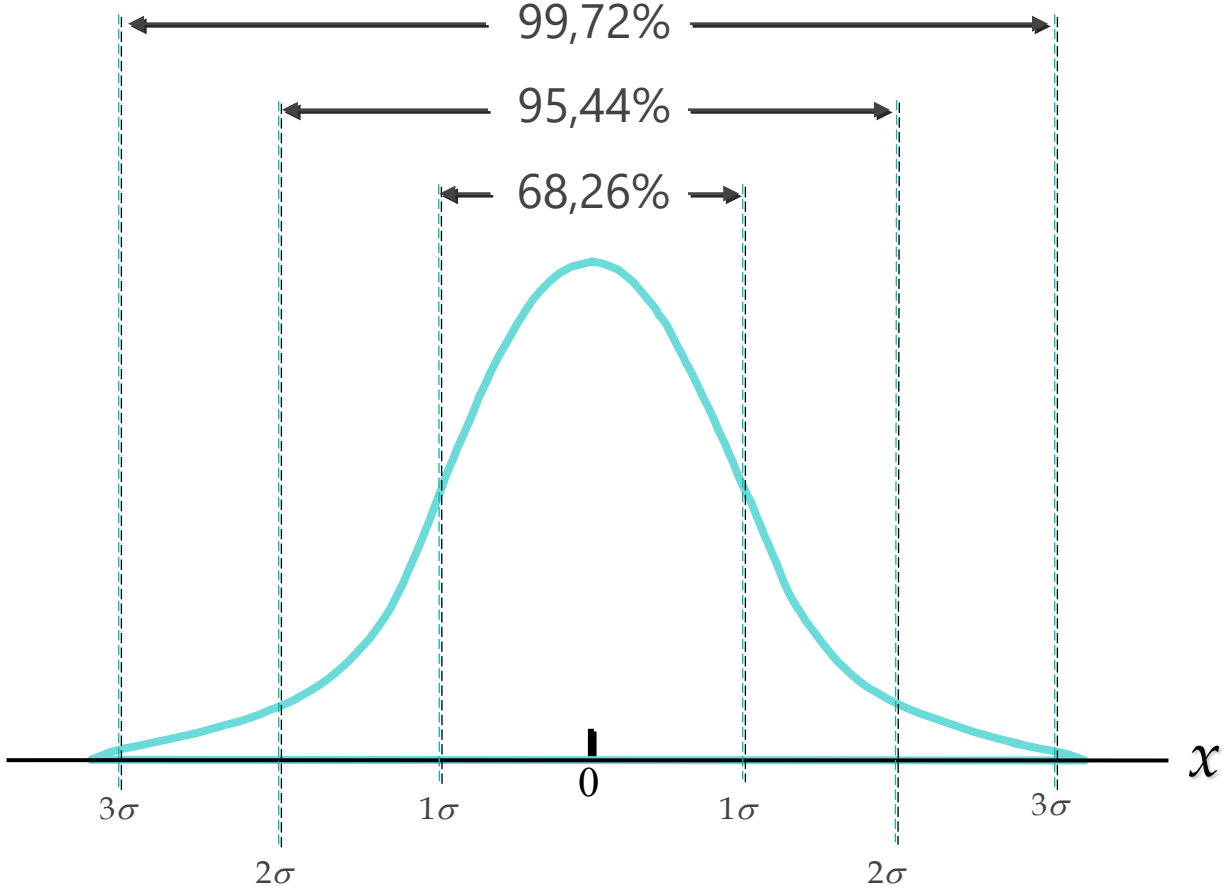
$$\varepsilon_i \sim N(0,\sigma^2)$$

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.



Estimativa dos Parâmetros

Busca encontrar o melhor ajuste para um conjunto de dados minimizando a **soma dos quadrados dos resíduos** (das diferenças entre o valor estimado e os valores observados), de forma a maximizar o grau de ajuste do modelo aos dados observados.

MÉTODO DOS MÍNIMOS QUADRADOS

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

$$\varepsilon_i^2 = [Y_i - (\beta_0 + \beta_1 X_i)]^2$$

- Minimizar o erro é equivalente a calcular a derivada da função.
- Igualar o erro a ZERO e encontrar os valores de β_1 e β_0.

$$\beta_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \qquad \beta_0 = \bar{Y} - \beta_1 \bar{X}$$

em que

n : número de observações

$\bar{X}$: média amostral da variável X

$\bar{Y}$: média amostral da variável Y

Propôs os Método dos Mínimos Quadrados

Carl Friedrich Gauss
(Alemanha, 1777-1855)

Hipóteses sob os parâmetros

Modelo de Regressão Linear Simples

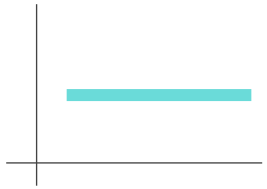
Modelo teórico

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

Hipóteses de Interesse sob β_1

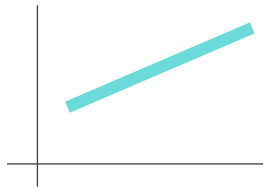
H_0: $\beta_1 = 0$

(não existe relação linear entre as variáveis)



H_1: $\beta_1 \neq 0$

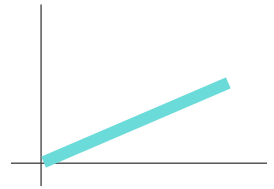
(existe relação linear entre as variáveis)



Hipóteses de Interesse sob β_0

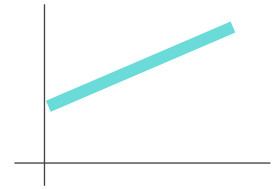
H_0: $\beta_0 = 0$

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



H_1: $\beta_0 \neq 0$

(não passa pela origem (x=0, y= β_0))



Refazer o case no RStudio

Calcule a reta de regressão e o coeficiente de determinação no *software* R e compare os resultados com o Excel.

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
```

Distribuição dos resíduos.

0,0000000736

Testa a hipótese se cada coeficiente é diferente de zero. P-valor <0,10, indica se a estimativa é diferente de zero.

Coeficiente de determinação

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$) por mil 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 a 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.

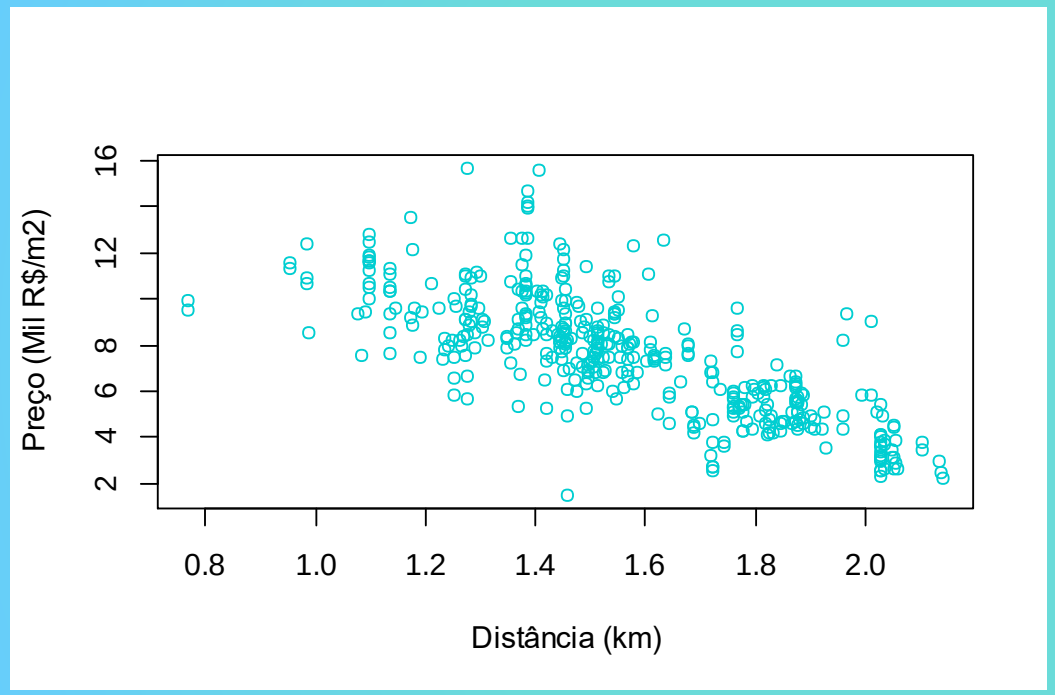
Fonte Adaptada: https://www.kaggle.com/quantbruce/real-estate-price-prediction?select=Real+estate.csv

Siga as seguintes instruções para solução do case:

(a) Existe relação entre preço do imóvel e distância para o metrô? Qual tipo de relação seria?
(b) Calcule a correlação de Pearson entre as duas variáveis e interprete o coeficiente.
(c) Rode o modelo de regressão linear simples. Realize os testes de hipóteses sob os parâmetros, ao nível de 10% de significância.
(d) Interprete os parâmetros do modelo e o coeficiente de determinação.
(e) Apresente a equação do modelo estimada.
(f) Estime o valor do imóvel caso o cliente desejasse morar há 1 km do metrô em um apartamento de 70 m^2.

RStudio

Case: Predição de preço de imóvel



Correlação = -0,76

RStudio

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
```

Interprete o *output* do modelo destacado em verde.

4. Regressão Linear Múltipla

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.

ε : erro aleatório associado ao modelo.

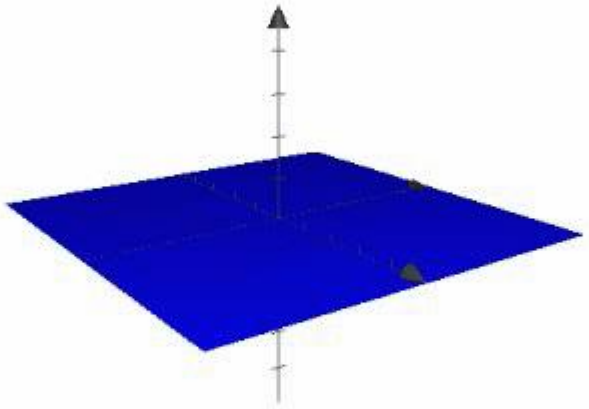
A equação de regressão linear múltipla estimada é dada por:

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

Forma geral do modelo

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 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 disponibilizado uma amostra histórica de clientes com as informações de **Idade**, **Rendimento Total**, **Salário**, **Limite de Crédito Imediato** para investigar se é possível estimar o Limite do Cheque Especial com base nas características disponibilizadas. 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".

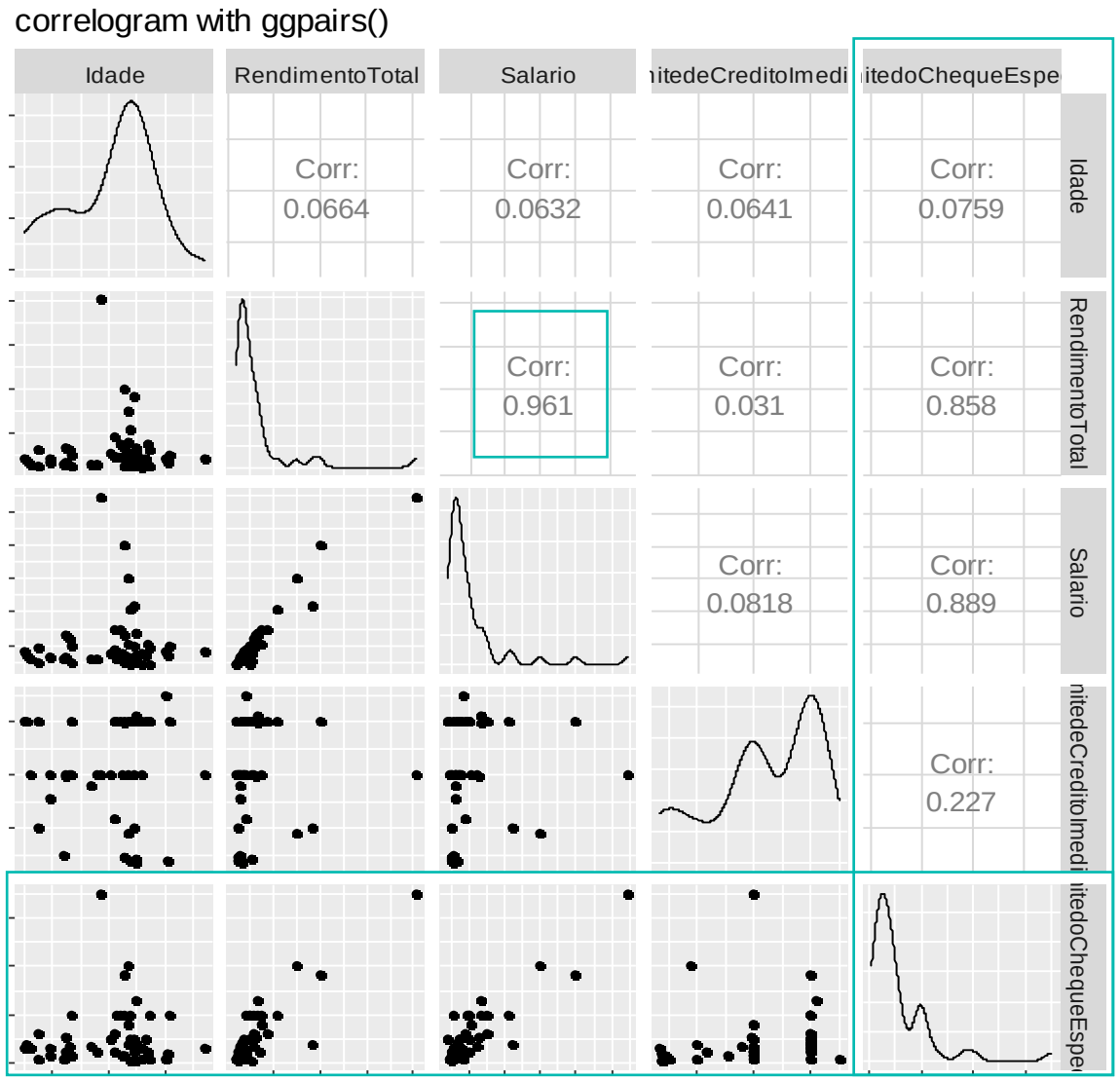
Fonte: Base de dados inspirada em cases reais.

Utilize todas as ferramentas aprendidas até o momento para tirar suas conclusões de negócio.

Vamos fazer juntos?

RStudio

Case: Predição de Limite de Cheque Especial



Antes de partir para o modelo, fazer a Análise Exploratória de Dados (AED).

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 a 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 na 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 na presença da multicolinearidade.
- É considerada alta a correlação entre as variáveis explicativas quando $|r|>0,7$, sendo r o coeficiente de correlação linear de Pearson.

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)             -2.887e+03  1.876e+03  -1.539   0.1311
Idade                    4.320e+00  2.658e+01   0.163   0.8716
Salario                  5.753e-01  4.292e-02  13.402   <2e-16 ***
LimitedeCreditoImediato  1.011e+00  4.270e-01   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: 4.093e-16
```

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

Coeficiente de Determinação Ajustado : usar para RL Múltipla.

Coeficiente de Determinação : usar para RL Simples.

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

Hipótese de Interesse

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

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

R^2-ajustado

Coeficiente de Determinação R^2: proporção da variabilidade da resposta explicada pela equação de regressão múltipla estimada. Em geral, sempre se eleva quando são adicionadas variáveis explicativas no modelo.

Para evitar superestimação do impacto de se adicionar mais uma variável independente no modelo, usamos a seguinte correção do R^2 :

$$R^2\text{-ajustado} = 1 - (1 - R^2)\frac{(n - 1)}{(n-p-1)}$$

Sendo *n* a quantidade de observações e *p* a quantidade de parâmetros estimados pelo modelo.

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)             -2.637e+03  1.061e+03  -2.486   0.0167 *
Salario                  5.757e-01  4.239e-02  13.582   <2e-16 ***
LimitedeCreditoImediato  1.015e+00  4.217e-01   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: < 2.2e-16
```

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

Case: Predição de Limite de Cheque Especial

Interpretação do Modelo Final (escolhido).

Limite do Cheque Especial = -0,002637 + 0,5757*Salário + 1,015*Limite de Crédito Imediato

R^2-ajustado: 0,8057

<u>Interpretação do coeficiente de regressão:</u>

- 0,5757 é o aumento 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 do Limite do Cheque Especial correspondente ao aumento de 1 unidade do Limite de Crédito Imediato, quando o Salário é mantido constante.

<u>Interpretação do R^2-ajustado:</u>

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

Variáveis *dummies*

- Utilizamos até agora no modelo Regressão Linear 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.

  Sexo: feminino e masculino.

  Sexo_M: 0 – feminino e 1 – masculino.

  Estado Civil: solteiro, casado e outros.

  Est_civil_O: 1 – outros e 0 – demais.

  Est_civil_S: 1 – solteiro e 0 – demais.

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

Variáveis *dummies*

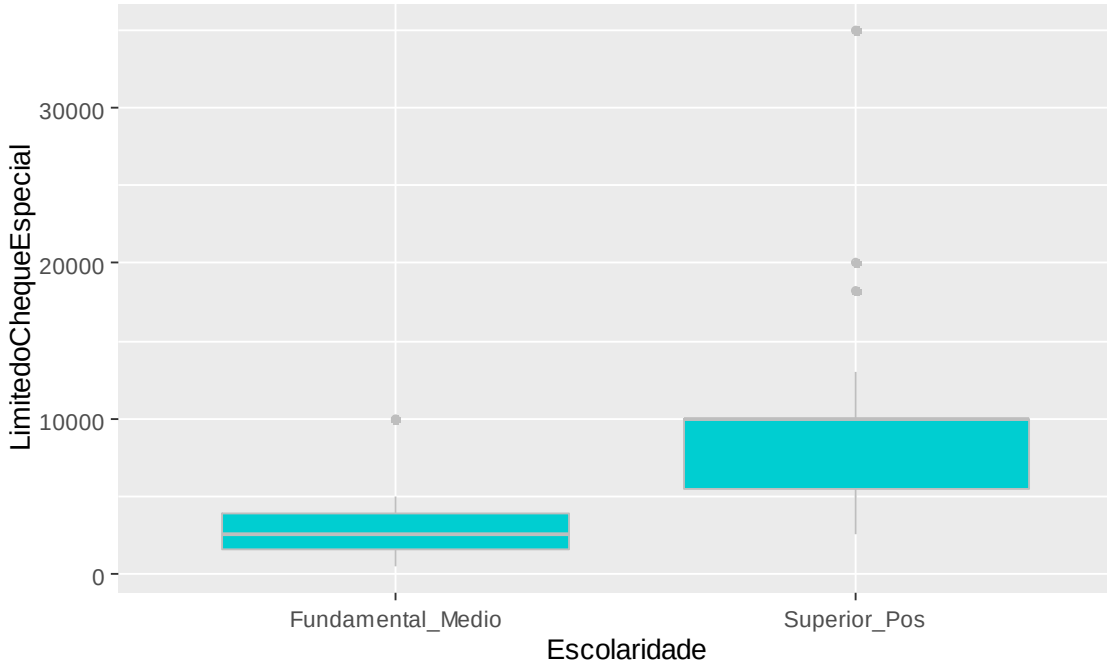
- Codificação das variáveis qualitativas:

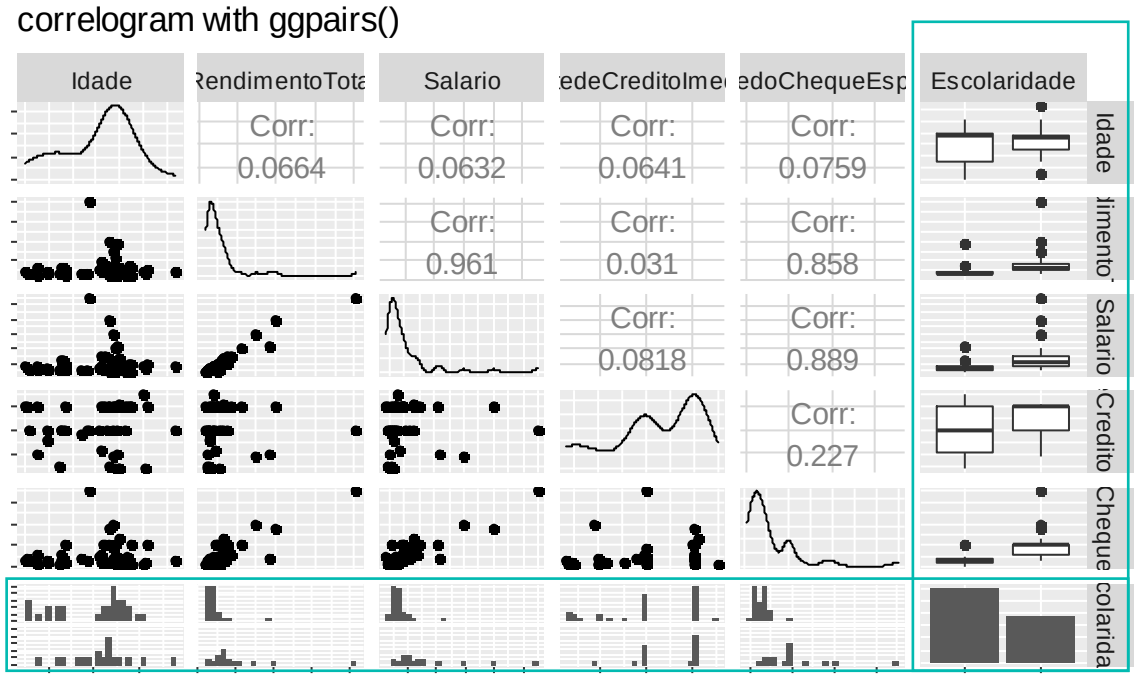
cliente	Sexo	Estado Civil	Sexo_M	Est_civil_O	Est_civil_S
1	feminino	solteiro	0	0	1
2	masculino	casado	1	0	0
3	feminino	outros	0	1	0
4	masculino	solteiro	1	0	1
5	masculino	solteiro	1	0	1
6	masculino	solteiro	1	0	1
7	feminino	casado	0	0	0

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

Case: Predição de Limite de Cheque Especial

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





A variável Escolaridade parece discriminar, sendo os clientes com curso superior ou pós-graduação com valores maiores de limite de cheque especial. Ela também mostra 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 ~ Escolaridade + Salario +
    LimitedeCreditoImediato, data = dados_lim_cred)

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)              -2.322e+03  9.964e+02  -2.330  0.02445 *
EscolaridadeSuperior_Pos  2.445e+03  8.832e+02   2.768  0.00821 **
Salario                   5.171e-01  4.487e-02  11.524  7.05e-15 ***
LimitedeCreditoImediato   7.322e-01  4.066e-01   1.801  0.07860 .
---
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: < 2.2e-16
```

Interpretação do coeficiente associado a Escolaridade:

R$2.445 é o valor atribuído ao Limite de Cheque Especial para os clientes com escolaridade Superior ou Pós, quando as demais covariáveis do modelo são mantidas constantes.

Como interpretar a categoria de Escolaridade Fundamental e Médio?

5. Exercícios para casa

5. Exercícios para casa

DATA DE ENTREGA 04/10/2020 | 2 EXERCÍCIOS-CASE

i. CASE: Limite de Crédito (4,0 ponto)
ii. CASE: Predição de valor de imóvel (6,0 pontos)

Instruções importantes:

- A lista vale nota (0-10) e deve ser entregue até 04/10/2020. Lista entregue até 11/10/2020 valerá 80% da nota. Posteriormente, não será mais aceita a lista para correção. Não serão aceitas listas parciais.
- O exercício será considerado como "realizado", quando tiver, além das análises, a interpretação do resultados.
- Disponibilização apenas do código, tabelas e gráficos mesmo se estiverem corretos, serão considerados na correção como "meio certo", pois o mais importante é a interpretação do resultado.
- Soluções técnicas "elegantes e mais completas" serão consideradas como ponto extra para o aluno (+0,5 na lista geral).
- A lista é individual. No caso de detecção de plágio, lista não será considerada para correção.

BOM ESTUDO ☺

5.i. Case: Predição de Limite de Cheque Especial

BANCO DE DADOS EM .TXT | FAZER ANÁLISE NO R

Uma instituição financeira tem 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 disponibilizado 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 com base nas características disponibilizadas. Avalie a possibilidade de fornecer uma "regra" por meio de um modelo estatístico, interprete como as informações que predizem o evento de interesse e qual a performance desta "regra".

Utilize todas as ferramentas aprendidas até o momento para tirar suas conclusões de negócio.

Siga as seguintes instruções para solução do case:

(a) Construa o gráfico de dispersão entre as variáveis.
(b) Calcule a correlação de Pearson entre elas.
(c) Rode o modelo de regressão linear.
(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$4850.
(g) Faça comentários em relação ao uso dessa "regra de predição" do Limite de Cheque Especial. Você acredita ser um boa regra a ser utilizada pela instituição financeira?

RStudio

5.ii. CASE: Predição de preço de imóvel

BANCO DE DADOS EM .TXT | FAZER ANÁLISE NO R

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 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

Siga as seguintes instruções para solução do case:

(a) Existe relação linear das covariáveis e preço do imóvel (variável resposta)? Qual tipo de relação seria?
(b) Calcule a correlação de Pearson entre todas as variáveis do banco de dados. Interprete os coeficientes.
(c) Realize a análise bidimensional entre as covariáveis existentes e investigue possíveis problemas de multicolinearidade.
(d) Rode o modelo de regressão linear múltipla, considerando um nível de significância de 5% para seleção de variáveis.
(e) Interprete os parâmetros do modelo e o coeficiente de determinação.
(f) Apresente a equação do modelo estimado.
(g) Caso um comprador esteja procurando um imóvel há 1 km do metrô, com 5 comércios próximos e que tenha 5 anos, qual valor ele pagaria em um imóvel de 85 m^2?
(h) De acordo com a medida de qualidade do modelo, ele prediz de forma satisfatória o valor do imóvel? Justifique.

RStudio

Referências

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