
FIAP
MBA

MBA em DATA SCIENCE & ARTIFICIAL INTELLIGENCE

APPLIED STATISTICS

profregina.bernal@fiap.com.br
reginabernal@terra.com.br

Dra. Regina Tomie Ivata Bernal
Cientista de Dados na área da Saúde

Formação Acadêmica:

- Estatístico - UFSCar
- Mestre em Saúde Pública – FSP/USP
- Doutor em Ciências – Epidemiologia - FSP/USP

Atividades Profissionais:

- Professora de pós-graduação na FIAP
- Consultora externa da SVS/MS
- Cientista de Dados em Saúde

REGRESSÃO LINEAR

Regressão Linear

OBJETIVO

Unir de forma paramétrica os dados históricos, buscando sua relação de dependência entre períodos de tempo e na relação de causa e efeito entre variáveis

Regressão Linear

UTILIZAÇÃO

As técnicas quantitativas são aplicadas nas condições:

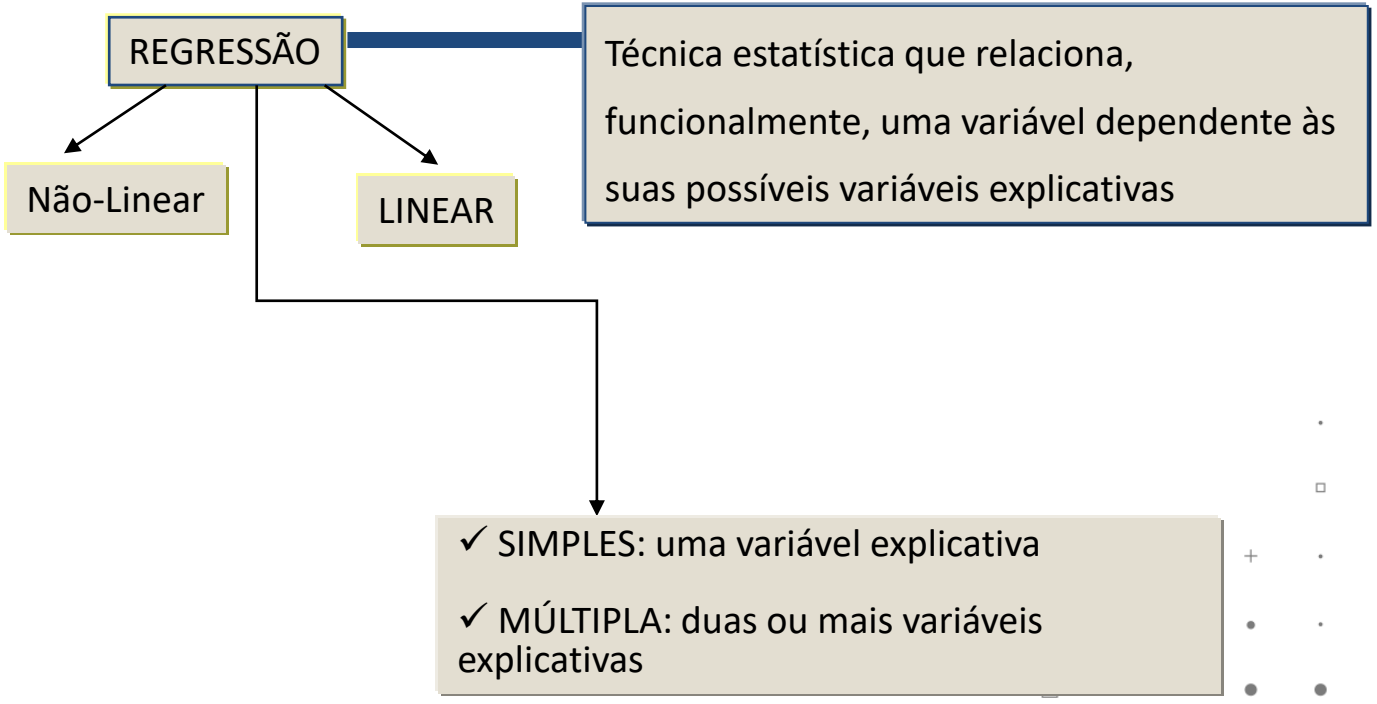
- ✓ Informações do passado disponíveis;
- ✓ Informações quantificáveis em forma numérica;
- ✓ Assumir a hipótese de que algo dos padrões do passado irá se repetir no futuro (hipótese de continuidade).

Regressão Linear

O Modelo Causal permite:

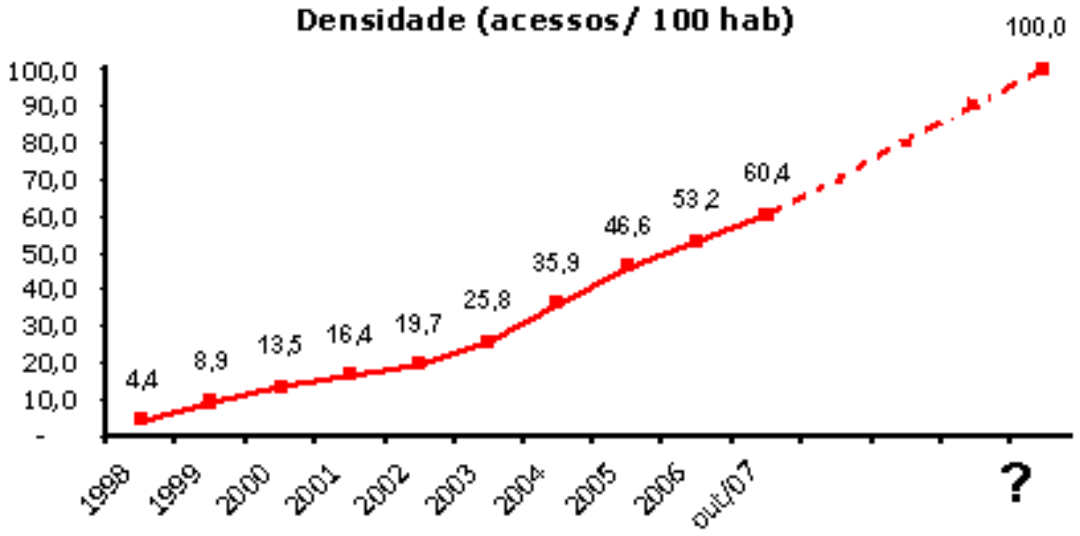
- ✓ Expressar as relações de Causa-Efeito entre variáveis;
- ✓ Entender melhor os mecanismos geradores do fato em estudo;
- ✓ Simular situações de forma a se avaliar o seu impacto na previsão;
- ✓ Analisar situações independentes do tempo.

Modelos de Regressão



Regressão Linear

✓ Case 2: Quando o Brasil vai ter 100 celulares para cada 100 habitantes?



Fonte: http://www.teleco.com.br/comentario/com237.asp

Conceito

MODELO PROBABILÍSTICO

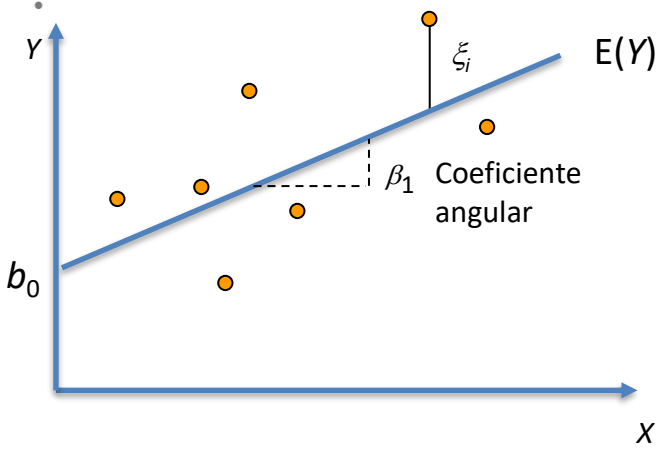
y = Componente Determinístico + Erro Aleatório

onde y é a variável dependente

Escrever a equação linear envolve dois parâmetros:

- ✓ O Intercepto de y
- ✓ A inclinação da reta

Regressão Linear Simples



$$E(Y) = \beta_0 + \beta_1 X$$

Inclinação populacional

Intercepto populacional

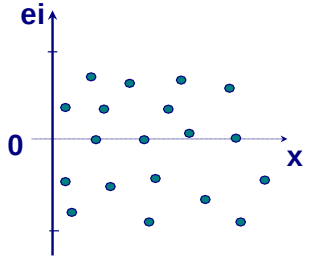
$$Y_i = \beta_0 + \beta_1 X_i + \xi_i$$

Análise de Resíduos

Forma de avaliar se as suposições colocadas no desenvolvimento do modelo não foram violadas

$$\hat{e}_i = y_i - \hat{y}_i$$

Pelo gráfico de dispersão, visualizamos o comportamento dos resíduos



Análise de Resíduos

RESÍDUOS PADRONIZADOS:

$$\frac{e_i}{se}$$

RESÍDUOS STUDENTIZADOS:

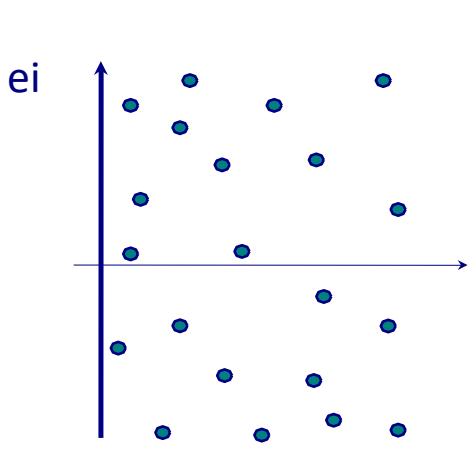
$$\frac{e_i}{se\sqrt{1 - v_{ii}}}$$

onde $v_{ii} = \frac{1}{n} + \frac{(xi - \bar{x})^2}{\sum (xi - \bar{x})^2}$

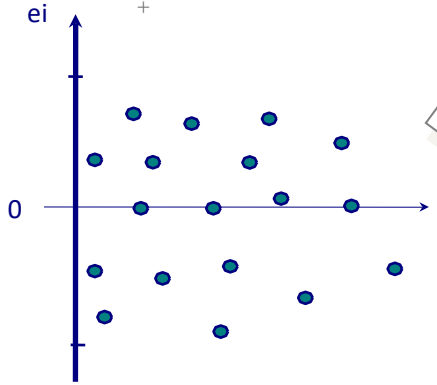
Análise de Resíduos

LINEARIDADE

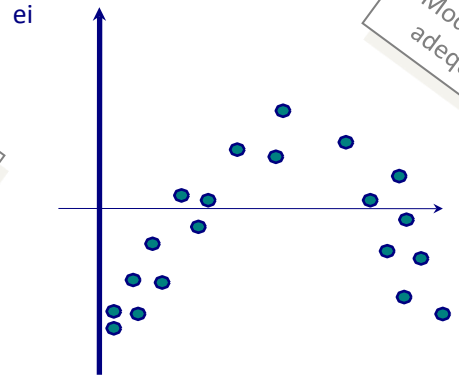
Gráfico de dispersão dos valores preditos (y) e o resíduo ⇨ os resíduos devem estar distribuídos aleatoriamente.



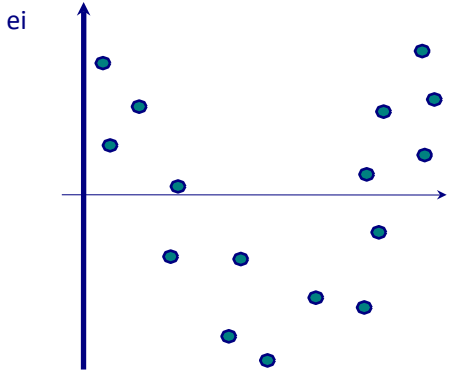
Análise de Resíduos



Modelo adequado



Modelo não adequado



Modelo não adequado

o comportamento não aleatório dos resíduos podem ser eliminados ajustando no modelo um termo quadrático.

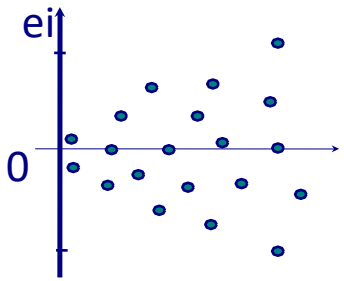
Análise de Resíduos

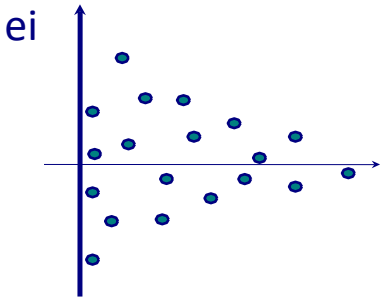
IGUALDADE DE VARIÂNCIA

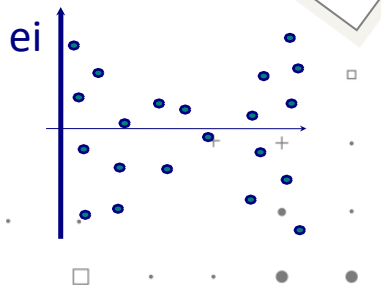
Quando o gráfico de dispersão dos Resíduos Studentizados, contra o valor predito, indica que a extensão dos resíduos aumentam com a magnitude dos valores preditos:

Então a suposição de igualdade da variância está violada

Modelo não adequado







Análise de Resíduos

NORMALIDADE

Pelo histograma dos resíduos padronizados pode-se analisar a suposição de normalidade.

Teste de Shapiro verifica a hipótese nula que os resíduos do modelo ajustado segue uma distribuição Normal:

H_0 : Distribuição = Normal

H_1: Distribuição ≠ Normal

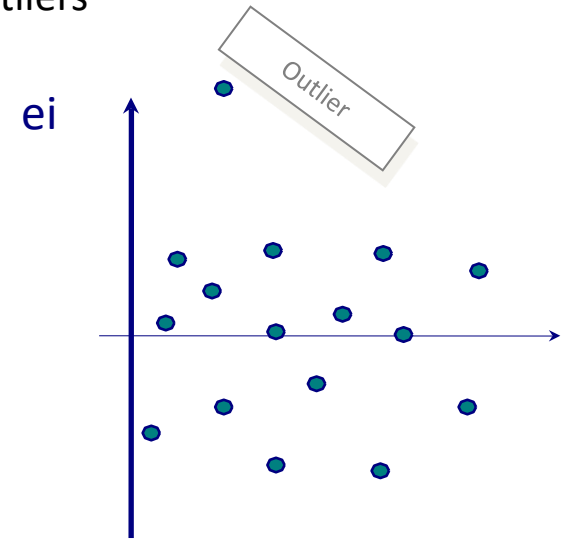
Critério de decisão:

Se p-valor < 0.05 então rejeito H0,
Se p_valor >= 0.05 então não rejeito H0

Análise de Resíduos

LOCALIZANDO OS OUTLIERS:

Em geral, resíduos padronizados com valores maiores que 2 são considerados outliers



Medidas de desempenho dos modelos

Erro Médio (Mean error-ME): $ME=\frac{\sum_{i=1}^{n} y_i-\hat{y}_i}{n}$

Erro Médio Absoluto (Mean Absolut Error-MAE): $MAE=\frac{\sum_{i=1}^{n} |y_i-\hat{y}_i|}{n}$

Raiz do Erro Quadrático Médio (Root Mean Squared Error-RMSE): $RMSE=\sqrt{\frac{\sum_{i=}^{n} (y_i-\hat{y}_i)^2}{n}}$

Erro Percentual Médio (Mean Percent Error-MPE): $$MPE = \frac{\sum_{i=1}^{n} \frac{y_i - \hat{y}_i}{y_i * 100}}{n}$$

Erro Percentual Absoluto Médio (Mean Absolut Percent Error-MAPE): $$MAPE = \frac{\sum_{i=1}^{n} \frac{|y_i - \hat{y}_i|}{y_i * 100}}{n}$$

Sendo:

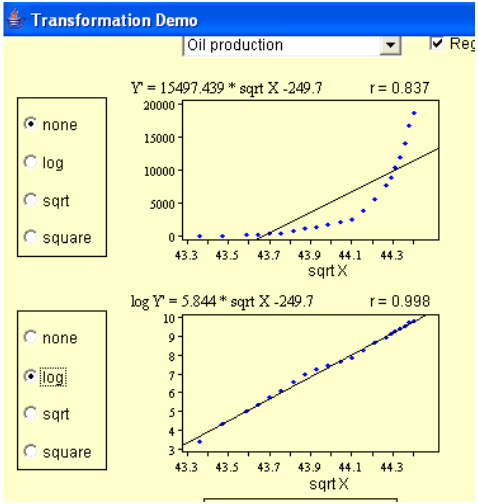
y_i = variável resposta

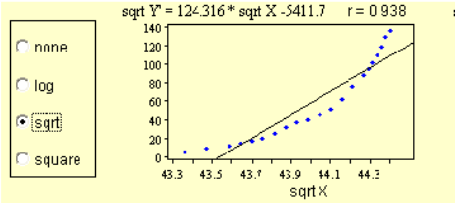
$\hat{y}_i$ = previsão do modelo

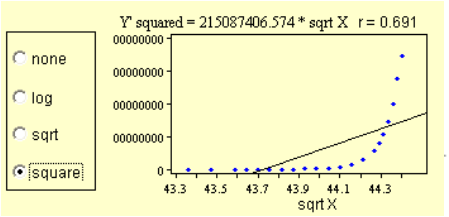
Transformação de Variável

Quando o modelo não é conhecido, pode-se escolher a transformação examinando o gráfico x e y.

Exemplos:

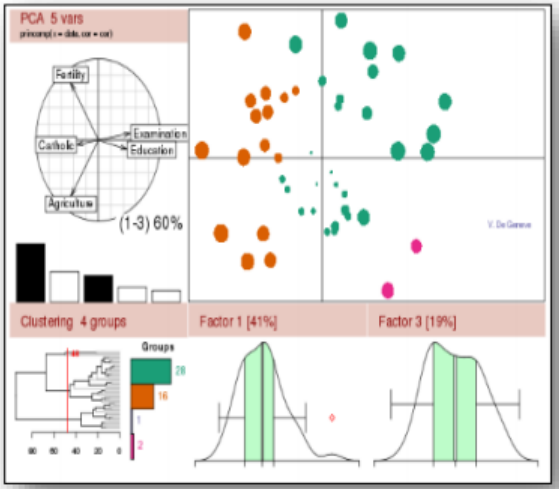






Fonte: http://onlinestatbook.com/stat_sim/transformations/index.html

DISTRIBUIÇÃO DE REGRESSÃO LINEAR MULTIPLA



Exercícios

REGRESSÃO LINEAR MÚLTIPLA

Regressão Linear Múltipla

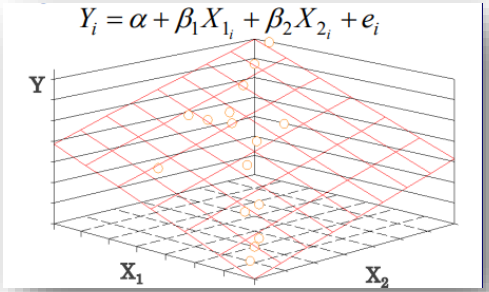
- Modelo Linear Múltiplo: $Y=B_o+B_1X_1+B_2X_2+B_3X_3+...+B_nX_n+ e$

$X_1,X_2,X_3,....,X_n$ = variáveis independentes

Y = variável dependente

B_o = constante

$B_1,B_2,B_3,....,B_n$ = coeficientes de regressão associados às n variáveis



Regressão Linear

Métodos de seleção de variáveis preditoras

Instrumento para selecionar variáveis(atributos) significativos

BACKWARD
FORWARD
STEPWISE

- Backward Selection : Procedimento constrói **adicionando todas as variáveis** e vai eliminando iterativamente uma a uma até que não haja mais variáveis .
- Forward Selection: Procedimento constrói iterativamente **adicionando variáveis uma a uma** até que não haja mais variáveis preditoras
- Stepwise: Combinação de Forward Selection e Backward elimination. Procedimento constrói iterativamente uma seqüência de modelos pela adição ou remoção de variáveis em cada etapa.

Regressão Linear

Com os dados de uma amostra podemos calcular as estimativas dos parâmetros (B) não conhecidos. Usando para isso o ajuste pelo método dos mínimos quadrados.

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_n x_n$$

Minimizar o Soma de quadrados dos erros (Sum Square Erro -SSE):

$$SS_E = \sum (y_i - \bar{y})^2$$

Análise de Variância

A variabilidade total observada na variável dependente está dividido em componentes:

$$\sum (y_i - \bar{y})^2 = \sum (y_i - \hat{y}_i)^2 + \sum (\hat{y}_i - \bar{y})^2$$

Soma de quadrados total	Soma de quadrados residual	Soma de quadrados regressão
SQTot	SQRes	SQReg

Análise de Variância

Podemos resumir todas essas informações numa única tabela anova:

Fonte	gl	SQ	QM	F
Regressão	p - 1	SQReg	$QMReg = \frac{SQReg}{p-1}$	$\frac{QMReg}{Se^2}$
Resíduo	n - p	SQRes	$Se^2 = \frac{SQRes}{n-p}$	
Total	n - 1	SQTOT	$S^2 = \frac{SQTot}{n-1}$	

Análise de Variância

Coeficiente de determinação (R^2): Multiple R-squared

$$R^2 = \frac{SQReg}{SQTot}$$

Coeficiente de determinação ajustado (R^2): Adjusted R-squared

$$R_a{}^2 = 1 - \frac{n-1}{n-(p+1)}(1-R^2)$$

Onde: n = número de observações
p = número de variáveis preditoras

```
> modelo <- lm(df$Vendas ~ df$Budget_Advertising)
> summary(modelo)

Call:
lm(formula = df$Vendas ~ df$Budget_Advertising)

Residuals:
    Min      1Q  Median      3Q     Max
-655330 -256271  -30444  234875  743028

Coefficients:
                          Estimate  Std. Error t value        Pr(>|t|)
(Intercept)           1060550.396  151771.308   6.988 0.0000000463123 ***
df$Budget_Advertising       4.964       0.524   9.473 0.0000000000458 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 349600 on 34 degrees of freedom
Multiple R-squared:  0.7252,    Adjusted R-squared:  0.7172
F-statistic: 89.75 on 1 and 34 DF,  p-value: 0.00000000004575
```

Resultado da ANOVA

$$\sum (y_i - \bar{y})^2 = \sum (y_i - \hat{y}_i)^2 + \sum (\hat{y}_i - \bar{y})^2$$

Soma de quadrados total	Soma de quadrados residual	Soma de quadrados regressão
SQTot	SQRes	SQReg

Exemplo: y=Vendas e x =budget

```
# Análise de variância (ANOVA)
df$predito = 1060550.396 + 4.964*df$Budget_Advertising
df$residuo = df$Vendas - df$predito
df$residuo2 = df$residuo*df$residuo

df$reg = (df$predito - ybarra)*(df$predito - ybarra)

# Soma dos quadrados
sqreg = sum(df$reg) ; sqreg
sqres = sum(df$residuo2) ; sqres
sqtotal = sqreg+sqres

rquadrado = sqreg/sqtotal ; rquadrado
```

```
> # Soma dos quadrados
> sqreg = sum(df$reg) ; sqreg
[1] 10966753022839
> sqres = sum(df$residuo2) ; sqres
[1] 4154940166057
> sqtotal = sqreg+sqres
>
> rquadrado = sqreg/sqtotal ; rquadrado
[1] 0.7252331
>
```

```
> modelo <- lm(df$Vendas ~ df$Budget_Advertising)
> summary(modelo)

Call:
lm(formula = df$Vendas ~ df$Budget_Advertising)

Residuals:
    Min      1Q  Median      3Q     Max
-655330 -256271  -30444  234875  743028

Coefficients:
                          Estimate  Std. Error t value        Pr(>|t|)
(Intercept)           1060550.396  151771.308   6.988 0.0000000463123 ***
df$Budget_Advertising       4.964       0.524   9.473 0.0000000000458 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 349600 on 34 degrees of freedom
Multiple R-squared:  0.7252,    Adjusted R-squared:  0.7172
F-statistic: 89.75 on 1 and 34 DF,  p-value: 0.00000000004575
```

Resultado da ANOVA

Teste de Hipóteses

TESTANDO OS PARÂMETROS B'S

H_o: $B_i = 0$

H_1: $B_i \neq 0$

$$t = \frac{B_i}{erro\_padrao(B_i)}$$ com gl = n - p

Quando $t > t_{\alpha/2} \Rightarrow$ região de rejeição

$$IC : \overline{b}_i + t_{\alpha/2} Sb_i$$

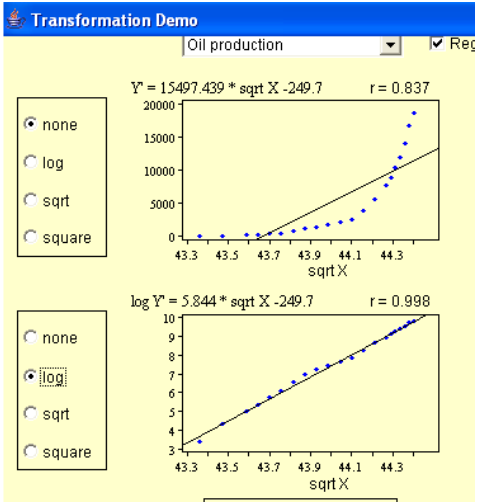
Regressão Linear

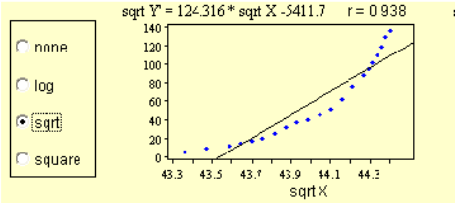
Pontos de atenção	Análise	O que fazer
Colinearidade: Correlação entre duas variáveis preditoras do modelo	Correlação de Pearson	- Escolher uma das variáveis ou - Criar a variável de iteração
Multicolinearidade: Qualquer variável preditora é correlacionada com um conjunto de outras variáveis preditoras	Fator de inflação da variância (VIF)	- Escolher uma das variáveis ou - Criar fatores usando a análise de componentes principais
Relação não linear entre a variável resposta e a preditora	Gráfico de dispersão	- Transformar a variável
Outliers	Resíduos padronizados	- Excluir os outliers da base de dados a cada nova rodada

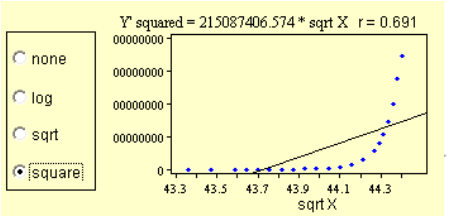
Transformação de Variável

Quando o modelo não é conhecido, pode-se escolher a transformação examinando o gráfico x e y.

Exemplos:







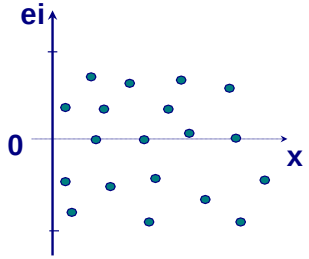
Fonte: http://onlinestatbook.com/stat_sim/transformations/index.html

Análise de Resíduos

Forma de avaliar se as suposições colocadas no desenvolvimento do modelo não foram violadas

$$\hat{e}_i = y_i - \hat{y}_i$$

Pelo gráfico de dispersão, visualizamos o comportamento dos resíduos



Análise de Resíduos

RESÍDUOS PADRONIZADOS:

$$\frac{e_i}{se}$$

RESÍDUOS STUDENTIZADOS:

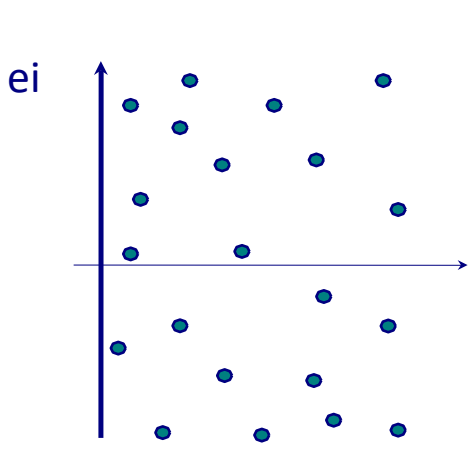
$$\frac{e_i}{se\sqrt{1 - v_{ii}}}$$

onde $v_{ii} = \frac{1}{n} + \frac{(xi - \bar{x})^2}{\sum (xi - \bar{x})^2}$

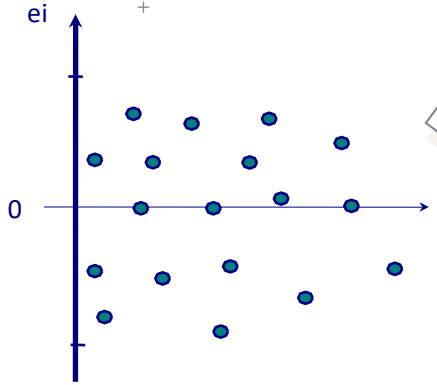
Análise de Resíduos

LINEARIDADE

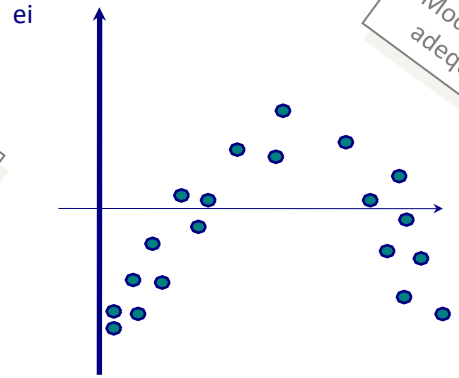
Gráfico de dispersão dos valores preditos (y) e o resíduo ⇨ os resíduos devem estar distribuídos aleatoriamente.



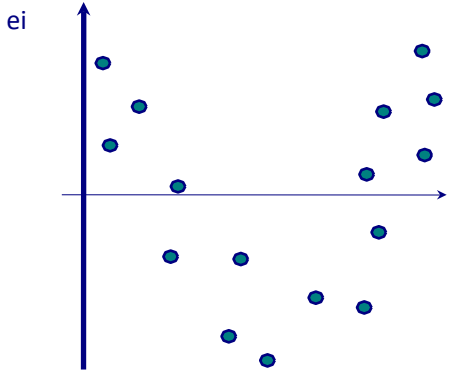
Análise de Resíduos



Modelo adequado



Modelo não adequado



Modelo não adequado

o comportamento não aleatório dos resíduos podem ser eliminados ajustando no modelo um termo quadrático.

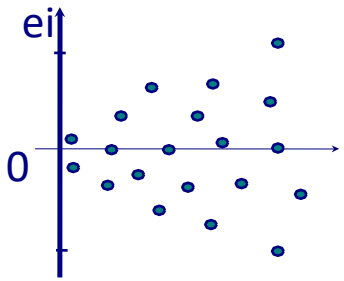
Análise de Resíduos

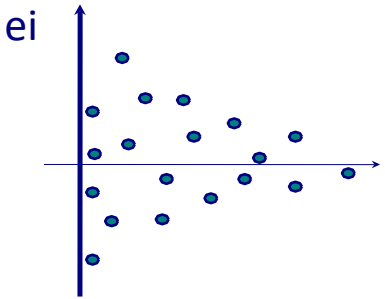
IGUALDADE DE VARIÂNCIA

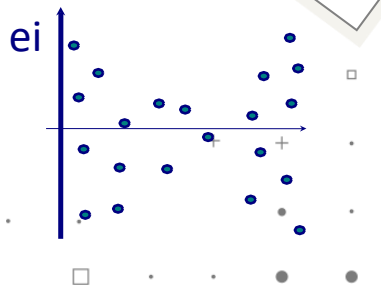
Quando o gráfico de dispersão dos Resíduos Studentizados, contra o valor predito, indica que a extensão dos resíduos aumentam com a magnitude dos valores preditos:

Então a suposição de igualdade da variância está violada

Modelo não adequado







Análise de Resíduos

NORMALIDADE

Pelo histograma dos resíduos padronizados pode-se analisar a suposição de normalidade.

Teste de Shapiro verifica a hipótese nula que os resíduos do modelo ajustado segue uma distribuição Normal:

H_0 : Distribuição = Normal

H_1: Distribuição ≠ Normal

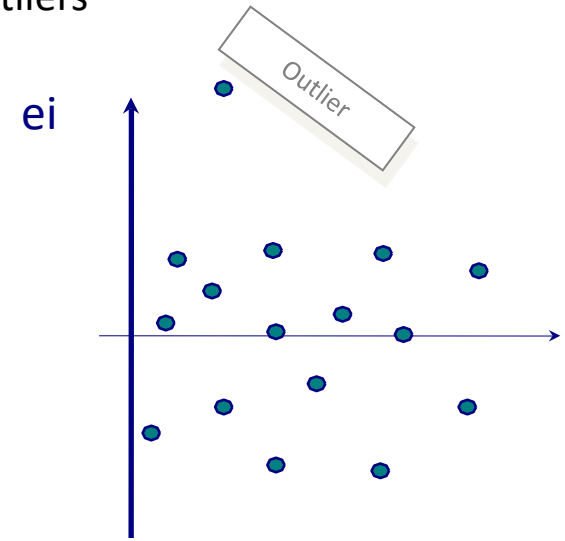
Critério de decisão:

Se p-valor < 0.05 então rejeito H0,
Se p_valor >= 0.05 então não rejeito H0

Análise de Resíduos

LOCALIZANDO OS OUTLIERS:

Em geral, resíduos padronizados com valores maiores que 2 são considerados outliers



Medidas de desempenho dos modelos

Erro Médio (Mean error-ME): $ME=\frac{\sum_{i=1}^{n} y_i-\hat{y}_i}{n}$

Erro Médio Absoluto (Mean Absolut Error-MAE): $MAE=\frac{\sum_{i=1}^{n} |y_i-\hat{y}_i|}{n}$

Raiz do Erro Quadrático Médio (Root Mean Squared Error-RMSE): $RMSE=\sqrt{\frac{\sum_{i=}^{n} (y_i-\hat{y}_i)^2}{n}}$

Erro Percentual Médio (Mean Percent Error-MPE): $$MPE = \frac{\sum_{i=1}^{n} \frac{y_i - \hat{y}_i}{y_i * 100}}{n}$$

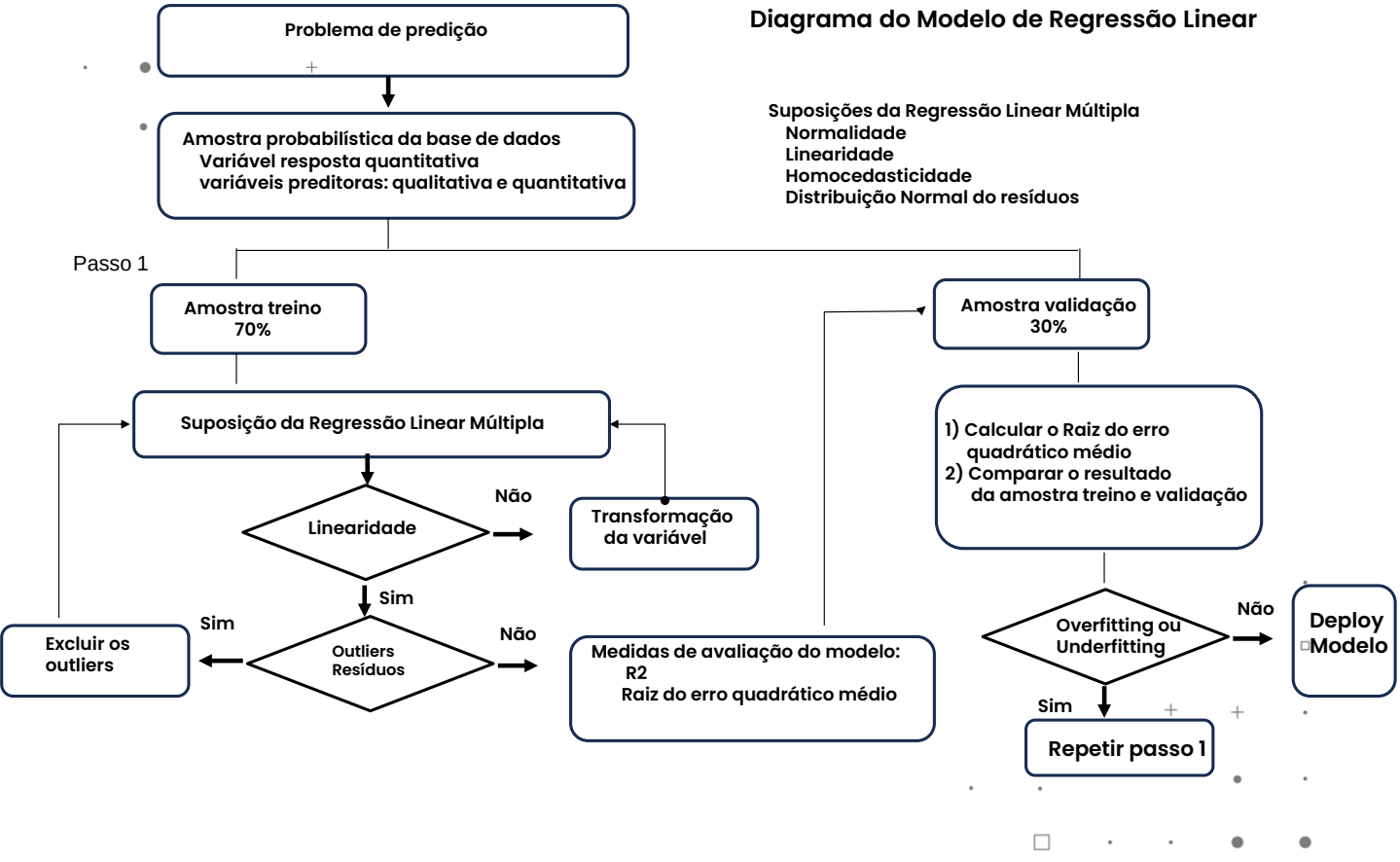
Erro Percentual Absoluto Médio (Mean Absolut Percent Error-MAPE): $$MAPE = \frac{\sum_{i=1}^{n} \frac{|y_i - \hat{y}_i|}{y_i * 100}}{n}$$

Sendo:

y_i = variável resposta

$\hat{y}_i$ = previsão do modelo

Diagrama do Modelo de Regressão Linear



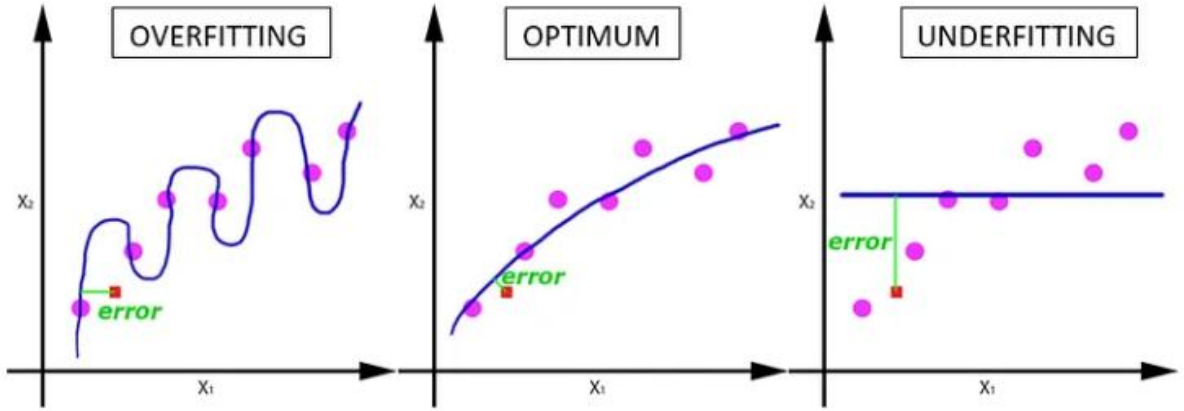
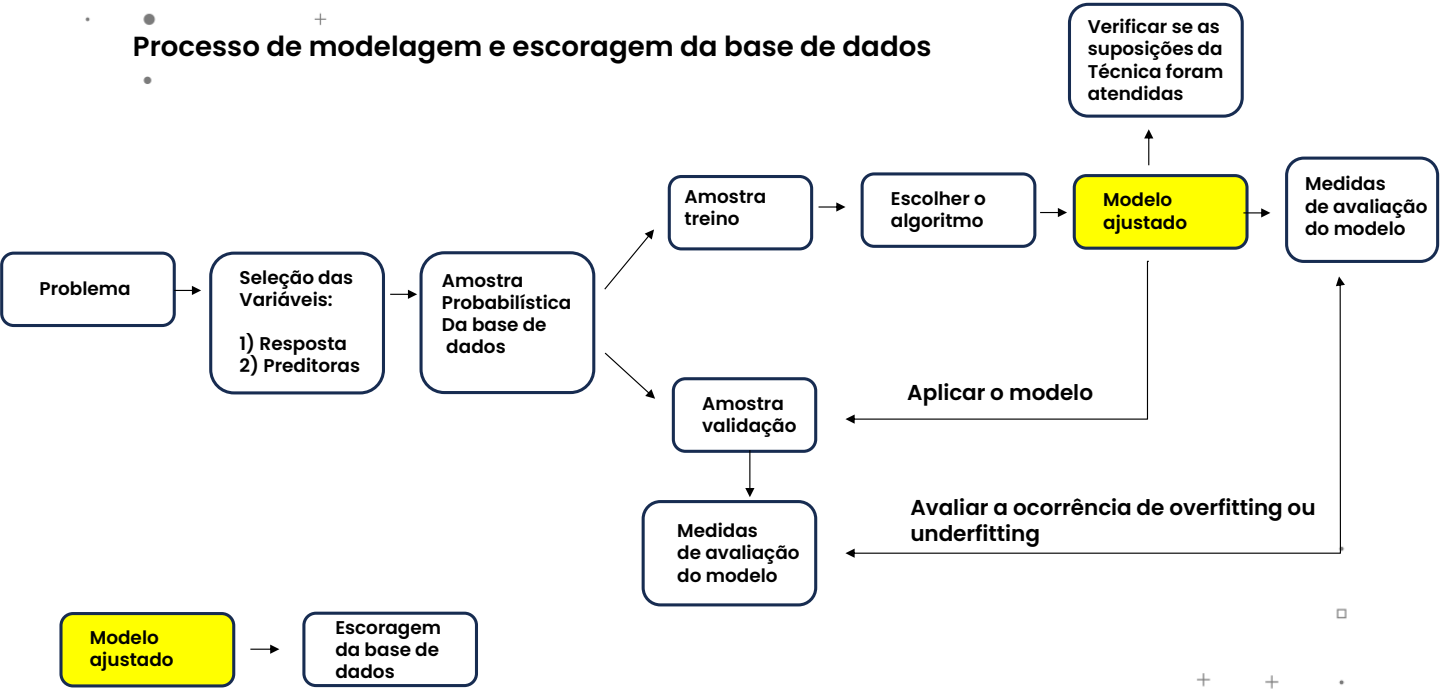


Fig. 2 — Visualização do Overfitting e Underfitting.

Fone: https://medium.com/comunidadeds/o-que-é-overfitting-0b91850d0512

Processo de modelagem e escoragem da base de dados

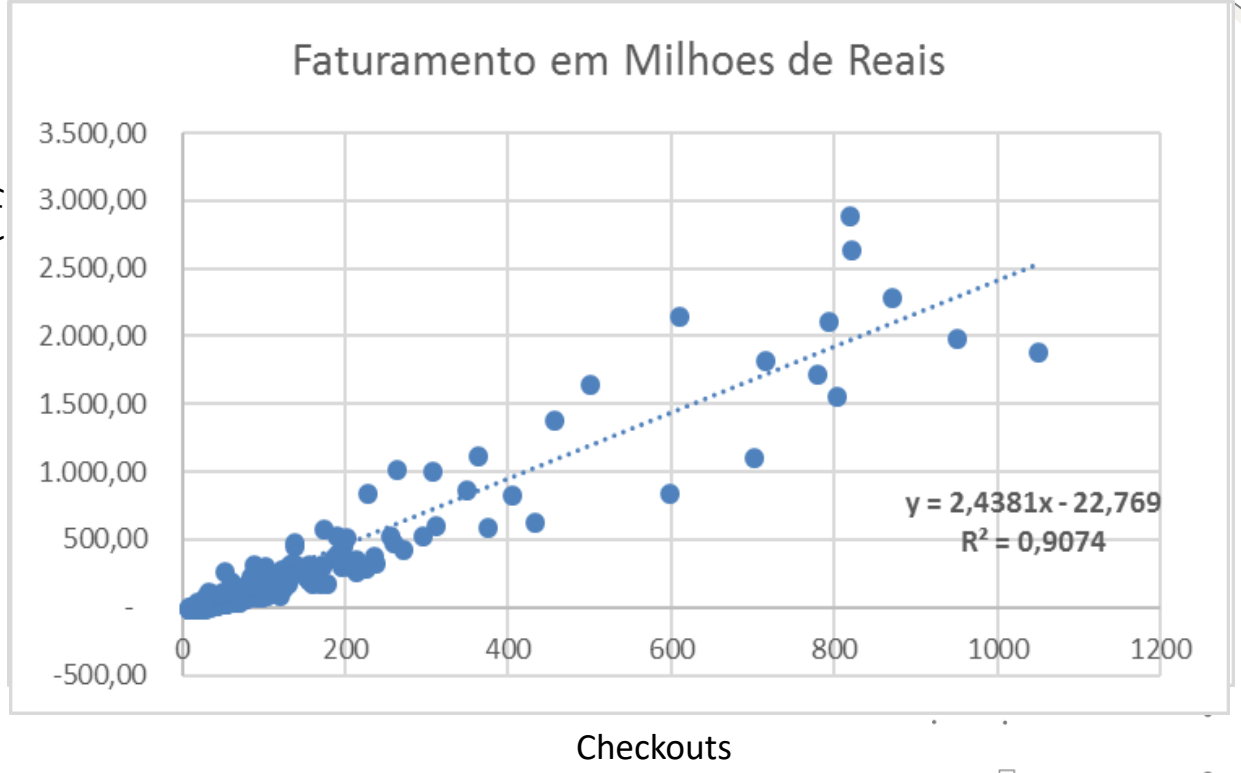


EXEMPLOS DE APLICAÇÕES DO MODELO DE REGRESSÃO LINEAR

Técnica de Regressão: Regressão Linear Simples

Exemplo: Faturamento anual (em milhões de Reais) por número de ckeckouts

Exemplo



Fonte:ABRAS

Exemplo

Predição de Tráfego – Por que?

Pode demorar de 10 a 20 min para ter uma máquina no ar. Dá pra esperar tudo isso?

Evite falsas quedas de tráfego

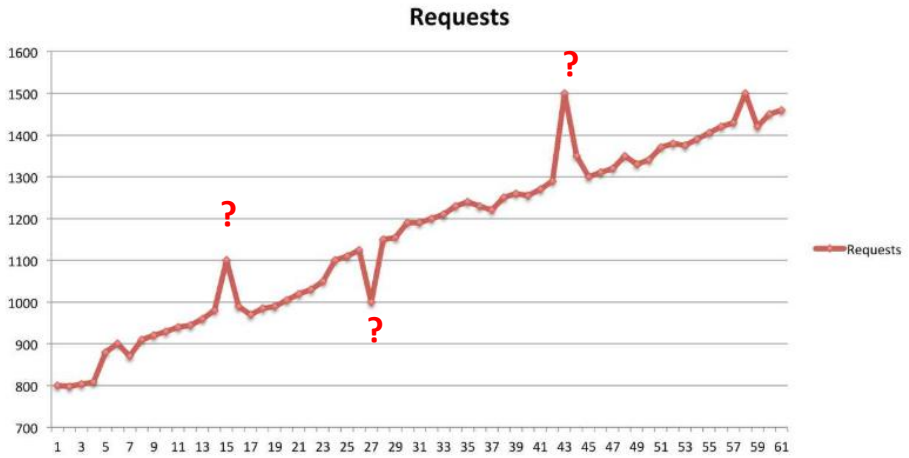
Fonte:https://www.infoq.com/br/presentations/data-science-em-publicidade-digital

Técnica de Regressão: Regressão Linear Simples

FIAP MBA+

Exemplo

Predição de Tráfego

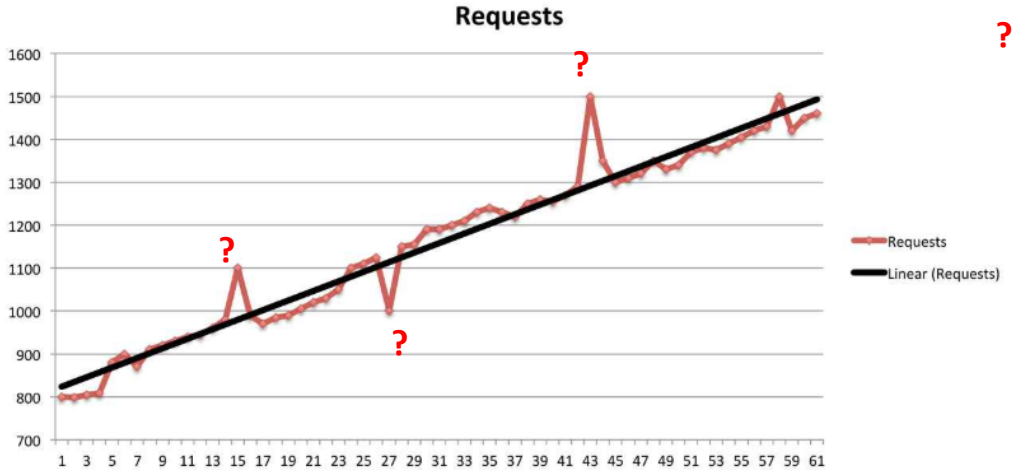


? Tomada de decisão

Fonte:https://www.infoq.com/br/presentations/data-science-em-publicidade-digital

Técnica de Regressão: Regressão Linear Simples

Predição de Tráfego



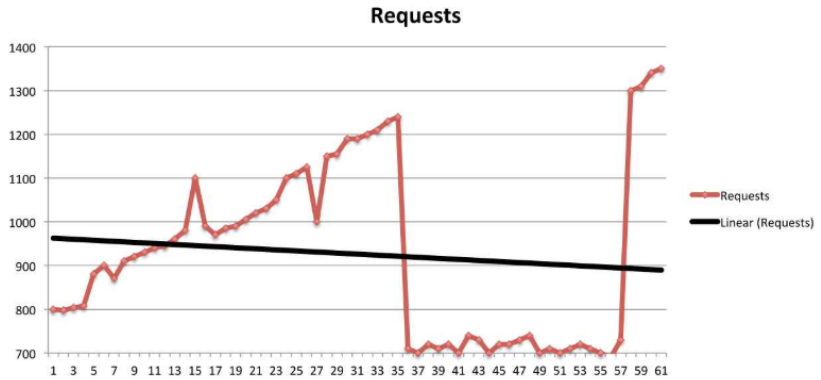
? Tomada de decisão

Fonte:https://www.infoq.com/br/presentations/data-science-em-publicidade-digital

Técnica de Regressão: Regressão Linear Simples

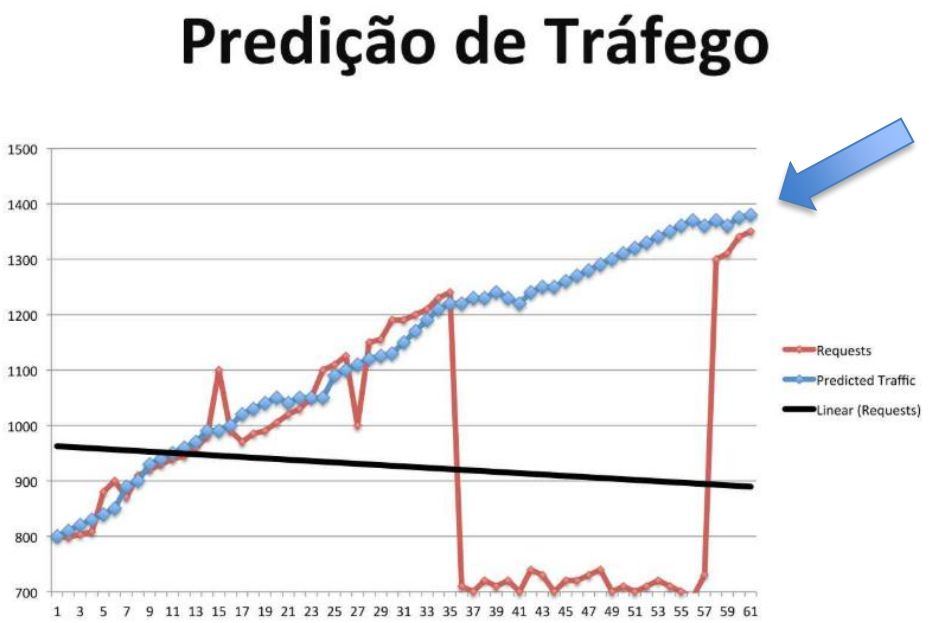
Exemplo

Predição de Tráfego



Fonte:https://www.infoq.com/br/presentations/data-science-em-publicidade-digital

Técnica de Regressão: Regressão Linear Simples



Aplicação da Regressão linear para imputação de dados

Fonte:https://www.infoq.com/br/presentations/data-science-em-publicidade-digital

Campanhas

Variáveis:

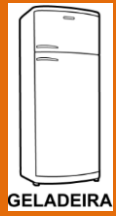
Segmento:
- Pequeno
- Médio
- Grande
- Muito Grande

Dia da semana:
- Domingo
- Segunda-feira
- Terça-feira
- Quarta-feira
- Quinta-feira
- Sexta-feira
- Sábado

Ordem da oferta:
- 1ª
- 2ª
- 3ª
- 4ª
- 5ª

Celebridade:
- Sim
- Não

Marca:
- A
- B
- C



GELADEIRA

Quantidade de vendas

Campanhas

Variáveis:

Segmento: **(X1)** Preditora
- Pequeno (1)
- Médio (2)
- Grande (3)
- Muito Grande (4)

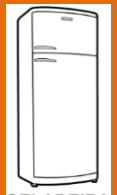
Dia da semana: **(X2)**
- Domingo (1)
- Segunda-feira (2)
- Terça-feira (3)
- Quarta-feira (4)
- Quinta-feira (5)
- Sexta-feira (6)
- Sábado (7)

Ordem da oferta: **(X3)**
- 1ª (1)
- 2ª (2)
- 3ª (3)
- 4ª (4)
- 5ª (5)

Marca: **(X4)**
- A (1)
- B (2)
- C (3)

Celebridade: **(X5)**
- Sim (1)
- Não (2)

Modelo Preditivo



Quantidade de vendas

Y

Resposta

Campanhas

Variáveis:

Segmento: **(X1)**
- Pequeno (1)
- Médio (2)
- Grande (3)
- Muito Grande (4)

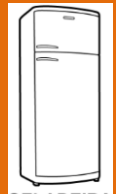
Dia da semana: **(X2)**
- Domingo (1)
- Segunda-feira (2)
- Terça-feira (3)
- Quarta-feira (4)
- Quinta-feira (5)
- Sexta-feira (6)
- Sábado (7)

Ordem da oferta: **(X3)**
- 1ª (1)
- 2ª (2)
- 3ª (3)
- 4ª (4)
- 5ª (5)

Marca: **X4**
- A (1)
- B (2)
- C (3)

Celebridade: **X5**
- Sim (1)
- Não (2)

Analytics



GELADEIRA

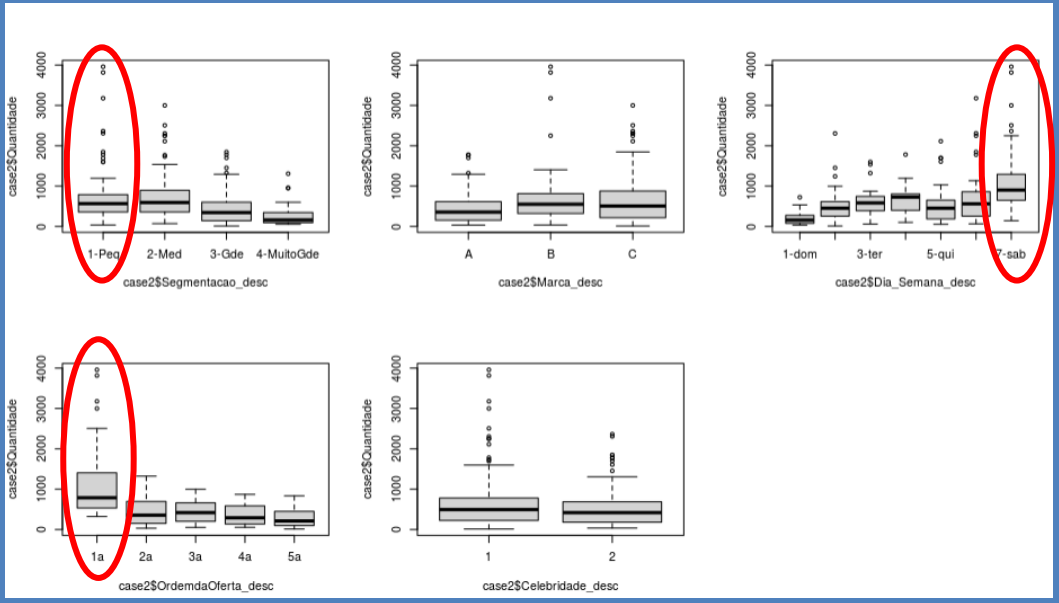
Quantidade de vendas

Y

Modelo de regressão linear múltipla

$$Y = a + b1*X1 + b2*X2+b3*X3+b4*X4+b5*X5$$

Análise bivariada



Variáveis qualitativas → Criar variáveis dicotômicas (booleanas/dummies) para cada categoria

Segmento: **(X1)** Preditora

Pequeno (1)
Médio (2)
Grande (3)
Muito Grande (4)

SegPequeno (X11)	SegMedio(X12)	SegGrande(X13)	SegMuitoGde (X14)
0	0	0	0
1	1	1	1

```
case2$segmentacao1 =case2$Segmentacao_valor
case2$segmentacao1 =ifelse(case2$Segmentacao_valor==1,1,0)
case2$segmentacao2 =case2$Segmentacao_valor
case2$segmentacao2 =ifelse(case2$Segmentacao_valor==2,1,0)
case2$segmentacao3 =case2$Segmentacao_valor
case2$segmentacao3 =ifelse(case2$Segmentacao_valor==3,1,0)
case2$segmentacao4 =case2$Segmentacao_valor
case2$segmentacao4 =ifelse(case2$Segmentacao_valor==4,1,0)
case2$segmentacao5 =case2$Segmentacao_valor
case2$segmentacao5 =ifelse(case2$Segmentacao_valor==5,1,0)
```

Variáveis qualitativas → Criar variáveis dicotômicas (booleanas/dummies) para cada categoria

```
case2$ordem_oferta_valor1 =case2$ordem_oferta_valor
case2$ordem_oferta_valor1 =ifelse(case2$ordem_oferta_valor==1,1,0)
case2$ordem_oferta_valor2 =case2$ordem_oferta_valor
case2$ordem_oferta_valor2 =ifelse(case2$ordem_oferta_valor==2,1,0)
case2$ordem_oferta_valor3 =case2$ordem_oferta_valor
case2$ordem_oferta_valor3 =ifelse(case2$ordem_oferta_valor==3,1,0)
case2$ordem_oferta_valor4 =case2$ordem_oferta_valor
case2$ordem_oferta_valor4 =ifelse(case2$ordem_oferta_valor==4,1,0)
```

```
case2$celebridade_valor1 =case2$celebridade_valor
case2$celebridade_valor1 =ifelse(case2$celebridade_valorr==1,1,0)
case2$celebridade_valor2 =case2$celebridade_valor
case2$celebridade_valor2 =ifelse(case2$celebridade_valor==2,1,0)
```

Variáveis qualitativas → Criar variáveis dicotômicas (booleanas/dummies) para cada categoria

```
case2$Dia_Semana_valor1 =case2$Dia_Semana_valor
case2$ordem_oferta_valor1 =ifelse(case2$Dia_Semana_valor==1,1,0)
case2$Dia_Semana_valor2 =case2$Dia_Semana_valor
case2$ordem_oferta_valor2=ifelse(case2$Dia_Semana_valor==2,1,0)
case2$Dia_Semana_valor3 =case2$Dia_Semana_valor
case2$ordem_oferta_valor3=ifelse(case2$Dia_Semana_valor==3,1,0)
case2$Dia_Semana_valor4=case2$Dia_Semana_valor
case2$ordem_oferta_valor4=ifelse(case2$Dia_Semana_valor==4,1,0)
case2$Dia_Semana_valor5 =case2$Dia_Semana_valor
case2$ordem_oferta_valor5=ifelse(case2$Dia_Semana_valor==5,1,0)
case2$Dia_Semana_valor6 =case2$Dia_Semana_valor
case2$ordem_oferta_valor6=ifelse(case2$Dia_Semana_valor==6,1,0)
case2$Dia_Semana_valor7 =case2$Dia_Semana_valor
case2$ordem_oferta_valor7=ifelse(case2$Dia_Semana_valor==7,1,0)
```

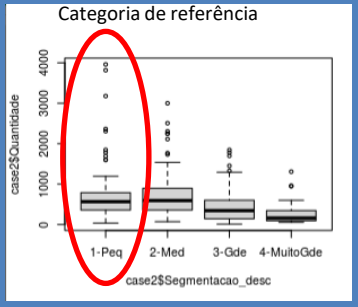
Modelo de regressão linear múltipla

Y = a + b1*X1+b2*X2+b3*X3 ➡ **Quantidade = 738 + 0.87*Seg2-291*Seg3-475*Seg4**

(teórico) **(ajustado)**

```
Call:
lm(formula = Quantidade ~ Segmentacao_valor, data = case2)

Residuals:
   Min     1Q Median     3Q    Max
-703.9 -316.8 -123.9  111.2 3220.1

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         738.8710    56.7257  13.025  < 2e-16 ***
Segmentacao_valor2    0.8706    81.1186   0.011 0.991444
Segmentacao_valor3 -291.0539    82.8690  -3.512 0.000511 ***
Segmentacao_valor4 -475.0199    97.9028  -4.852 1.95e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 547 on 307 degrees of freedom
Multiple R-squared:  0.1044,    Adjusted R-squared:  0.09569
F-statistic: 11.93 on 3 and 307 DF,  p-value: 2.062e-07
```

Categoria de referência



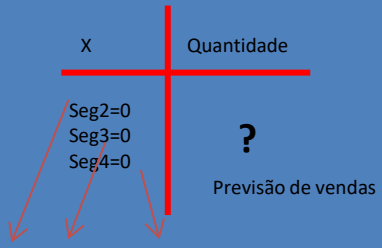
Acerto do modelo:
O segmento do produto explica 9,5% da variação da quantidade de vendas.

Modelo de regressão linear múltipla

Y = a + b1*X1+b2*X2+b3*X3 ➡ **Quantidade = 738.87 + 0.87 * Seg2 -291.05 * Seg3 -475.02 * Seg4**

(teórico) **(ajustado)**

```
Call:
lm(formula = Quantidade ~ Segmentacao_valor, data = case2)

Residuals:
   Min     1Q Median     3Q    Max
-703.9 -316.8 -123.9  111.2 3220.1

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         738.8710    56.7257  13.025  < 2e-16 ***
Segmentacao_valor2    0.8706    81.1186   0.011 0.991444
Segmentacao_valor3 -291.0539    82.8690  -3.512 0.000511 ***
Segmentacao_valor4 -475.0199    97.9028  -4.852 1.95e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 547 on 307 degrees of freedom
Multiple R-squared:  0.1044,    Adjusted R-squared:  0.09569
F-statistic: 11.93 on 3 and 307 DF,  p-value: 2.062e-07
```



Quantidade = 738.87 + 0.87x0 -291.01x0 – 475.02x0= 738.87

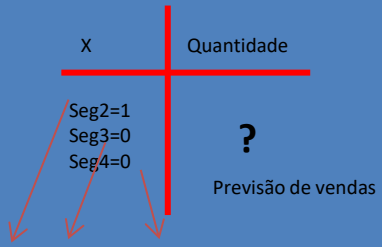
Quantidade = 739

Modelo de regressão linear múltipla

Y = a + b1*X1+b2*X2+b3*X3 ➡ **Quantidade = 738.87 + 0.87 * Seg2 -291.05 * Seg3 -475.02 * Seg4**

(teórico) **(ajustado)**

```
Call:
lm(formula = Quantidade ~ Segmentacao_valor, data = case2)

Residuals:
   Min     1Q Median     3Q    Max
-703.9 -316.8 -123.9  111.2 3220.1

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)         738.8710    56.7257  13.025  < 2e-16 ***
Segmentacao_valor2    0.8706    81.1186   0.011 0.991444
Segmentacao_valor3 -291.0539    82.8690  -3.512 0.000511 ***
Segmentacao_valor4 -475.0199    97.9028  -4.852 1.95e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 547 on 307 degrees of freedom
Multiple R-squared:  0.1044,    Adjusted R-squared:  0.09569
F-statistic: 11.93 on 3 and 307 DF,  p-value: 2.062e-07
```



Quantidade = 738.87 + 0.87x1 -291.01x0 – 475.02x0= 739.84

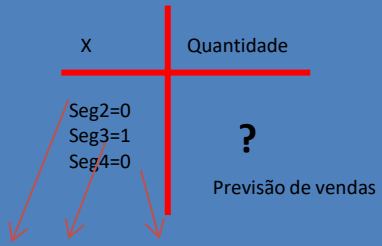
Quantidade = 740

Modelo de regressão linear múltipla

Y = a + b1*X1+b2*X2+b3*X3 ➡ **Quantidade = 738.87 + 0.87 * Seg2 -291.05 * Seg3 -475.02 * Seg4**

(teórico) **(ajustado)**

```
Call:
lm(formula = Quantidade ~ Segmentacao_valor, data = case2)

Residuals:
   Min     1Q Median     3Q    Max
-703.9 -316.8 -123.9  111.2 3220.1

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)         738.8710    56.7257  13.025  < 2e-16 ***
Segmentacao_valor2    0.8706    81.1186   0.011 0.991444
Segmentacao_valor3 -291.0539    82.8690  -3.512 0.000511 ***
Segmentacao_valor4 -475.0199    97.9028  -4.852 1.95e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 547 on 307 degrees of freedom
Multiple R-squared:  0.1044,    Adjusted R-squared:  0.09569
F-statistic: 11.93 on 3 and 307 DF,  p-value: 2.062e-07
```



Quantidade = 738.87 + 0.87x0 -291.01x1 – 475.02x0= 447.86

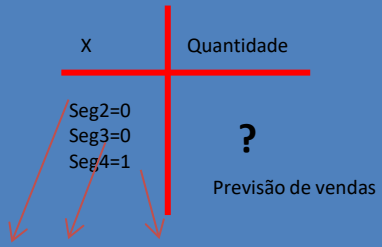
Quantidade = 448

Modelo de regressão linear múltipla

Y = a + b1*X1+b2*X2+b3*X3 ➡ **Quantidade = 738.87 + 0.87 * Seg2 -291.05 * Seg3 -475.02 * Seg4**

(teórico) **(ajustado)**

```
Call:
lm(formula = Quantidade ~ Segmentacao_valor, data = case2)

Residuals:
   Min     1Q Median     3Q    Max
-703.9 -316.8 -123.9  111.2 3220.1

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)         738.8710    56.7257  13.025  < 2e-16 ***
Segmentacao_valor2    0.8706    81.1186   0.011 0.991444
Segmentacao_valor3 -291.0539    82.8690  -3.512 0.000511 ***
Segmentacao_valor4 -475.0199    97.9028  -4.852 1.95e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 547 on 307 degrees of freedom
Multiple R-squared:  0.1044,    Adjusted R-squared:  0.09569
F-statistic: 11.93 on 3 and 307 DF,  p-value: 2.062e-07
```



Quantidade = 738.87 + 0.87x0 -291.01x0 – 475.02x1= 263.85

Quantidade = 264

Modelo de regressão linear múltipla

Y = a + b1*X1+b2*X2+b3*X3 ➡ **Quantidade = 738 + 0.87 * Seg2 -291 * Seg3 -475 * Seg4**

(teórico) (ajustado)

```
Call:
lm(formula = Quantidade ~ Segmentacao_valor, data = case2)

Residuals:
   Min     1Q Median     3Q    Max
-703.9 -316.8 -123.9  111.2 3220.1

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)         738.8710    56.7257  13.025  < 2e-16 ***
Segmentacao_valor2    0.8706    81.1186   0.011 0.991444
Segmentacao_valor3 -291.0539    82.8690  -3.512 0.000511 ***
Segmentacao_valor4 -475.0199    97.9028  -4.852 1.95e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 547 on 307 degrees of freedom
Multiple R-squared:  0.1044,    Adjusted R-squared:  0.09569
F-statistic: 11.93 on 3 and 307 DF,  p-value: 2.062e-07
```

	Data	Quantidade	Segmentacao_valor	predito	residuo	residuop
300	7/12/2018	2112	2	740	1372.26	2.52271
301	8/11/2018	1844	3	448	1396.18	2.56794
302	7/13/2018	2248	2	740	1508.26	2.77273
303	4/7/2018	2249	2	740	1509.26	2.77457
304	7/30/2018	2305	2	740	1565.26	2.87752
305	3/16/2018	2308	1	739	1569.13	2.88393
306	2/10/2018	2363	1	739	1624.13	2.98502
307	3/17/2018	2506	2	740	1766.26	3.24703
308	7/28/2018	2999	2	740	2259.26	4.15335
309	6/8/2018	3177	1	739	2438.13	4.48108
310	3/10/2018	3816	1	739	3077.13	5.65551
311	5/5/2018	3959	1	739	3220.13	5.91833

Modelo de regressão linear múltipla

Y = a + b1*X1+b2*X2+b3*X3+ ... + bpXp

(teórico)

```
lm(formula = Quantidade ~ Segmentacao_valor + marca_valor + ordem_oferta_valor +
    dia_semana_valor + celebridade_valor, data = case2)

Residuals:
    Min      1Q  Median      3Q     Max
-851.84 -203.79  -36.35  131.22 2428.28

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)           790.97      91.02   8.690 2.58e-16 ***
Segmentacao_valor2    -50.56      62.38  -0.810 0.418341
Segmentacao_valor3   -298.83      69.51  -4.299 2.34e-05 ***
Segmentacao_valor4   -387.65      80.12  -4.838 2.12e-06 ***
marca_valor2          -36.97      69.36  -0.533 0.594392
marca_valor3          114.08      56.47   2.020 0.044274 *
ordem_oferta_valor2  -510.69      74.64  -6.842 4.53e-11 ***
ordem_oferta_valor3  -498.71      64.83  -7.693 2.19e-13 ***
ordem_oferta_valor4  -627.77      83.92  -7.481 8.61e-13 ***
ordem_oferta_valor5  -692.40      64.49 -10.737  < 2e-16 ***
dia_semana_valor2     172.14      75.64   2.276 0.023570 *
dia_semana_valor3     309.67     113.32   2.733 0.006664 **
dia_semana_valor4     247.79     104.07   2.381 0.017902 *
dia_semana_valor5     270.77      74.38   3.640 0.000322 ***
dia_semana_valor6     415.35      82.08   5.060 7.40e-07 ***
dia_semana_valor7     716.83      79.67   8.997  < 2e-16 ***
celebridade_valor2     59.89      47.81   1.253 0.211351

Residual standard error: 399.7 on 294 degrees of freedom
Multiple R-squared:  0.542,     Adjusted R-squared:  0.5171
F-statistic: 21.75 on 16 and 294 DF,  p-value: < 2.2e-16
```

Modelo de regressão linear múltipla

Y = a + b1*X1+b2*X2+b3*X3+ ... + bpXp

(teórico)

```
lm(formula = Quantidade ~ Segmentacao_valor + marca_valor + ordem_oferta_valor +
    Dia_Semana_desc + celebridade_valor, data = case2_sout)

Residuals:
    Min      1Q  Median      3Q     Max
-614.46 -164.55  -24.35  128.61 1068.70

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          648.154     60.646  10.688  < 2e-16 ***
Segmentacao_valor2   -19.711     41.910  -0.470 0.638488
Segmentacao_valor3  -246.302     46.370  -5.312 2.20e-07 ***
Segmentacao_valor4  -336.864     52.917  -6.366 7.81e-10 ***
marca_valor2         -43.420     46.165  -0.941 0.347737
marca_valor3          93.464     37.573   2.488 0.013440 *
ordem_oferta_valor2 -335.299     50.080  -6.695 1.16e-10 ***
ordem_oferta_valor3 -330.744     43.651  -7.577 5.06e-13 ***
ordem_oferta_valor4 -436.846     56.191  -7.774 1.42e-13 ***
ordem_oferta_valor5 -508.046     43.635 -11.643  < 2e-16 ***
Dia_Semana_desc2-seg  189.014     49.939   3.785 0.000188 ***
Dia_Semana_desc3-ter  350.749     74.615   4.701 4.05e-06 ***
Dia_Semana_desc4-qua  311.068     68.529   4.539 8.36e-06 ***
Dia_Semana_desc5-qui  251.245     49.317   5.094 6.40e-07 ***
Dia_Semana_desc6-sex  310.554     55.403   5.605 4.93e-08 ***
Dia_Semana_desc7-sab  591.194     53.425  11.066  < 2e-16 ***
celebridade_valor2      4.085     31.972   0.128 0.898433
```

```
Residual standard error: 262.9 on 283 degrees of freedom
Multiple R-squared:  0.6042,    Adjusted R-squared:  0.5818
F-statistic:    27 on 16 and 283 DF,  p-value: < 2.2e-16

>
```

Modelo final
sem resíduos outliers

Seg4
Marca C
Dia da semana 7
Ordem 2
Não tem
celebridade

?

previsão

Modelo de regressão linear múltipla

Y = a + b1*X1+b2*X2+b3*X3+ ... + bpXp

(teórico)

```
lm(formula = Quantidade ~ Segmentacao_valor + marca_valor + ordem_oferta_valor +
    Dia_Semana_desc + celebridade_valor, data = case2_sout)

Residuals:
    Min      1Q  Median      3Q     Max
-614.46 -164.55  -24.35  128.61 1068.70

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          648.154     60.646  10.688  < 2e-16 ***
Segmentacao_valor2   -19.711     41.910  -0.470 0.638488
Segmentacao_valor3  -246.302     46.370  -5.312 2.20e-07 ***
Segmentacao_valor4  -336.864     52.917  -6.366 7.81e-10 ***
marca_valor2         -43.420     46.165  -0.941 0.347737
marca_valor3          93.464     37.573   2.488 0.013440 *
ordem_oferta_valor2 -335.299     50.080  -6.695 1.16e-10 ***
ordem_oferta_valor3 -330.744     43.651  -7.577 5.06e-13 ***
ordem_oferta_valor4 -436.846     56.191  -7.774 1.42e-13 ***
ordem_oferta_valor5 -508.046     43.635 -11.643  < 2e-16 ***
Dia_Semana_desc2-seg  189.014     49.939   3.785 0.000188 ***
Dia_Semana_desc3-ter  350.749     74.615   4.701 4.05e-06 ***
Dia_Semana_desc4-qua  311.068     68.529   4.539 8.36e-06 ***
Dia_Semana_desc5-qui  251.245     49.317   5.094 6.40e-07 ***
Dia_Semana_desc6-sex  310.554     55.403   5.605 4.93e-08 ***
Dia_Semana_desc7-sab  591.194     53.425  11.066  < 2e-16 ***
celebridade_valor2      4.085     31.972   0.128 0.898433
```

```
Residual standard error: 262.9 on 283 degrees of freedom
Multiple R-squared:  0.6042,    Adjusted R-squared:  0.5818
F-statistic:    27 on 16 and 283 DF,  p-value: < 2.2e-16

>
```

Modelo final
sem resíduos outliers

Seg4 → -336.86
Marca 3 (C) → +93.46
Dia da semana 7 → +591.19
Ordem 1 → 0
Não tem celebridade → +4.08

Intercepto → 648.15
Previsão = 1.000 produtos

EXERCÍCIO 2

FAÇA A PREVISÃO DAS VENDAS (R$) MENSAL NO PERÍODO DE 12 MESES DA EMPRESA XYZ A PARTIR DOS DADOS DISPONÍVEIS DE VENDAS (R$) E BUDGET ADVERTISING (R$) DA EMPRESA.

Faça a previsão das vendas (R$) mensal no período de 12 meses da empresa XYZ a partir dos dados disponíveis de Vendas (R$) e Budget Advertising (R$) da empresa.

	Data	ano	Vendas	Budget_Advertising
1	jan/16	2016	1160081	72800
2	fev/16	2016	1622540	123392
3	mar/16	2016	1597260	135761
4	abr/16	2016	1640675	148064
5	mai/16	2016	1511270	159746
6	jun/16	2016	1634073	183353
7	jul/16	2016	1856971	190722
8	ago/16	2016	1585566	197802
9	set/16	2016	2041672	248891
10	out/16	2016	1933557	256353
11	nov/16	2016	2076910	298805
12	dez/16	2016	1740202	268925

Faça a previsão das vendas (R$) mensal no período de 12 meses da empresa XYZ a partir dos dados disponíveis de Vendas (R$) e Budget Advertising (R$) da empresa.

	Data	ano	Vendas	Budget_Advertising
1	jan/16	2016	1160081	72800
2	fev/16	2016	1622540	123392
3	mar/16	2016	1597260	135761
4	abr/16	2016	1640675	148064
5	mai/16	2016	1511270	159746
6	jun/16	2016	1634073	183353
7	jul/16	2016	1856971	190722
8	ago/16	2016	1585566	197802
9	set/16	2016	2041672	248891
10	out/16	2016	1933557	256353
11	nov/16	2016	2076910	298805
12	dez/16	2016	1740202	268925

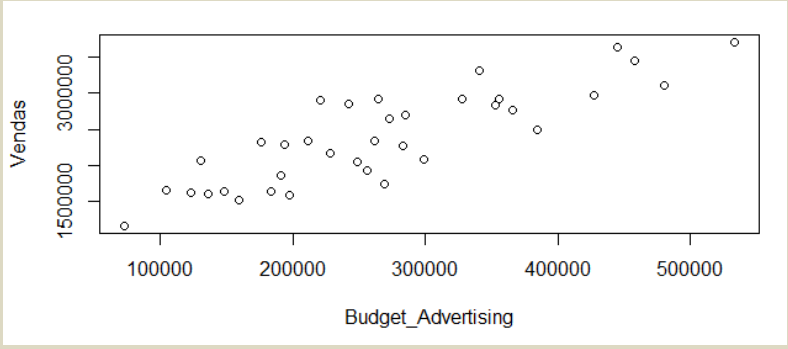
Média Desvio padrão Média Desvio padrão

	ano	`mean(Vendas)`	`sd(Vendas)`	`mean(Budget_Advertis~`	`sd(Budget_Advertis~`
	<fct>	<dbl>	<dbl>	<dbl>	<dbl>
1	2016	1700065.	253028.	190384.	67306.
2	2017	2428664.	361468.	271978	96151.
3	2018	3035830.	451835.	339973	120189.

>

Faça a previsão das vendas (R$) mensal no período de 12 meses da empresa XYZ a partir dos dados disponíveis de Vendas (R$) e Budget Advertising (R$) da empresa.

	Data	ano	Vendas	Budget_Advertising
1	jan/16	2016	1160081	72800
2	fev/16	2016	1622540	123392
3	mar/16	2016	1597260	135761
4	abr/16	2016	1640675	148064
5	mai/16	2016	1511270	159746
6	jun/16	2016	1634073	183353
7	jul/16	2016	1856971	190722
8	ago/16	2016	1585566	197802
9	set/16	2016	2041672	248891
10	out/16	2016	1933557	256353
11	nov/16	2016	2076910	298805
12	dez/16	2016	1740202	268925




```
> mc = cor(dadosquant);mc
                    Vendas Budget_Advertising
Vendas              1.000                0.852
Budget_Advertising  0.852                1.000
```

Faça a previsão das vendas (R$) mensal no período de 12 meses da empresa XYZ a partir dos dados disponíveis de Vendas (R$) e Budget Advertising (R$) da empresa.

	Data	ano	Vendas	Budget_Advertising
1	jan/16	2016	1160081	72800
2	fev/16	2016	1622540	123392
3	mar/16	2016	1597260	135761
4	abr/16	2016	1640675	148064
5	mai/16	2016	1511270	159746
6	jun/16	2016	1634073	183353
7	jul/16	2016	1856971	190722
8	ago/16	2016	1585566	197802
9	set/16	2016	2041672	248891
10	out/16	2016	1933557	256353
11	nov/16	2016	2076910	298805
12	dez/16	2016	1740202	268925

Modelo de regressão linear simples

Y = a + b1*X1 **(teórico)**

Vendas = a + b1*Budget

Faça a previsão das vendas (R$) mensal no período de 12 meses da empresa XYZ a partir dos dados disponíveis de Vendas (R$) e Budget Advertising (R$) da empresa.

	Data	ano	Vendas	Budget_Advertising
1	jan/16	2016	1160081	72800
2	fev/16	2016	1622540	123392
3	mar/16	2016	1597260	135761
4	abr/16	2016	1640675	148064
5	mai/16	2016	1511270	159746
6	jun/16	2016	1634073	183353
7	jul/16	2016	1856971	190722
8	ago/16	2016	1585566	197802
9	set/16	2016	2041672	248891
10	out/16	2016	1933557	256353
11	nov/16	2016	2076910	298805
12	dez/16	2016	1740202	268925

```
> summary(modelo)

Call:
lm(formula = Vendas ~ Budget_Advertising)

Residuals:
    Min      1Q  Median      3Q     Max
-655330 -256271  -30444  234875  743028

Coefficients:
                    Estimate  Std. Error t value       Pr(>|t|)
(Intercept)       1060550.396  151771.308    6.99 0.000000046312 ***
Budget_Advertising       4.964       0.524    9.47 0.000000000046 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 350000 on 34 degrees of freedom
Multiple R-squared:  0.725,     Adjusted R-squared:  0.717
F-statistic: 89.7 on 1 and 34 DF,  p-value: 0.0000000000458
```

Vendas = a + b1*Budget

(modelo ajustado) ➡ **Vendas = 1060550 + 4.964 * Budget**

A variável Budget explica 72,5% da variação das Vendas.

R$100 ➔ Vendas 1060550+4.96*1000 = 1.110.150

Faça a previsão das vendas (R$) mensal no período de 12 meses da empresa XYZ a partir dos dados disponíveis de Vendas (R$) e Budget Advertising (R$) da empresa.

	Data	ano	Vendas	Budget_Advertising
1	jan/16	2016	1160081	72800
2	fev/16	2016	1622540	123392
3	mar/16	2016	1597260	135761
4	abr/16	2016	1640675	148064
5	mai/16	2016	1511270	159746
6	jun/16	2016	1634073	183353
7	jul/16	2016	1856971	190722
8	ago/16	2016	1585566	197802
9	set/16	2016	2041672	248891
10	out/16	2016	1933557	256353
11	nov/16	2016	2076910	298805
12	dez/16	2016	1740202	268925

Modelo de regressão linear simples

```
> summary(modelo)

Call:
lm(formula = Vendas ~ Budget_Advertising)

Residuals:
    Min      1Q  Median      3Q     Max
-655330 -256271  -30444  234875  743028

Coefficients:
                    Estimate  Std. Error t value       Pr(>|t|)
(Intercept)       1060550.396  151771.308    6.99 0.000000046312 ***
Budget_Advertising       4.964       0.524    9.47 0.000000000046 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 350000 on 34 degrees of freedom
Multiple R-squared:  0.725,     Adjusted R-squared:  0.717
F-statistic: 89.7 on 1 and 34 DF,  p-value: 0.0000000000458
```

Faça a previsão das vendas (R$) mensal no período de 12 meses da empresa XYZ a partir dos dados disponíveis de Vendas (R$) e Budget Advertising (R$) da empresa.

Projeção para 2019

		Vendas	Budget
jan/19	2019		91000
fev/19	2019		154240
mar/19	2019		169702
abr/19	2019		185081
mai/19	2019	?	199683
jun/19	2019		229192
jul/19	2019		238403
ago/19	2019		247253
set/19	2019		311114
out/19	2019		320442
nov/19	2019		373507
dez/19	2019		336157

Modelo de regressão linear simples

```
> summary(modelo)

Call:
lm(formula = Vendas ~ Budget_Advertising)

Residuals:
    Min      1Q  Median      3Q     Max
-655330 -256271  -30444  234875  743028

Coefficients:
                    Estimate  Std. Error t value       Pr(>|t|)
(Intercept)       1060550.396  151771.308    6.99 0.000000046312 ***
Budget_Advertising        4.964       0.524    9.47 0.000000000046 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 350000 on 34 degrees of freedom
Multiple R-squared:  0.725,     Adjusted R-squared:  0.717
F-statistic: 89.7 on 1 and 34 DF,  p-value: 0.0000000000458
```

Faça a previsão das vendas (R$) mensal no período de 12 meses da empresa XYZ a partir dos dados disponíveis de Vendas (R$) e Budget Advertising (R$) da empresa.

Projeção para 2019

		Vendas	Budget
jan/19	2019		91000
fev/19	2019		154240
mar/19	2019		169702
abr/19	2019		185081
mai/19	2019		199683
jun/19	2019		229192
jul/19	2019		238403
ago/19	2019		247253
set/19	2019		311114
out/19	2019		320442
nov/19	2019		373507
dez/19	2019		336157

Modelo de regressão linear simples

```
> summary(modelo)

Call:
lm(formula = Vendas ~ Budget_Advertising)

Residuals:
    Min      1Q  Median      3Q     Max
-655330 -256271  -30444  234875  743028

Coefficients:
                    Estimate  Std. Error t value       Pr(>|t|)
(Intercept)       1060550.396  151771.308    6.99 0.000000046312 ***
Budget_Advertising        4.964       0.524    9.47 0.000000000046 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 350000 on 34 degrees of freedom
Multiple R-squared:  0.725,     Adjusted R-squared:  0.717
F-statistic: 89.7 on 1 and 34 DF,  p-value: 0.0000000000458
```

Vendas janeiro/19 = 1060550 + 4.964 * 91000 = 1.512.274

Faça a previsão das vendas (R$) mensal no período de 12 meses da empresa XYZ a partir dos dados disponíveis de Vendas (R$) e Budget Advertising (R$) da empresa.

Projeção para 2019

		Vendas	Budget
jan/19	2019	1.512.27	91000
fev/19	2019	4	154240
mar/19	2019		169702
abr/19	2019		185081
mai/19	2019		199683
jun/19	2019		229192
jul/19	2019		238403
ago/19	2019		247253
set/19	2019		311114
out/19	2019		320442
nov/19	2019		373507
dez/19	2019		336157

Modelo de regressão linear simples

```
> summary(modelo)

Call:
lm(formula = Vendas ~ Budget_Advertising)

Residuals:
    Min      1Q  Median      3Q     Max
-655330 -256271  -30444  234875  743028

Coefficients:
                    Estimate  Std. Error t value       Pr(>|t|)
(Intercept)       1060550.396  151771.308    6.99 0.000000046312 ***
Budget_Advertising        4.964       0.524    9.47 0.000000000046 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 350000 on 34 degrees of freedom
Multiple R-squared:  0.725,     Adjusted R-squared:  0.717
F-statistic: 89.7 on 1 and 34 DF,  p-value: 0.0000000000458
```

Vendas janeiro/19 = 1060550 + 4.964 * 91000 = 1.512.274

INTERPRETAÇÃO DOS RESULTADOS DA REGRESSÃO LINEAR NO PYTHON

EXEMPLO
Saída Python

OLS Regression Results

Dep. Variable:	total	R-squared:	0.834
Model:	OLS	Adj. R-squared:	0.829
Method:	Least Squares	F-statistic:	155.1
Date:	Fri, 28 Jun 2024	Prob (F-statistic):	8.07e-181
Time:	22:15:28	Log-Likelihood:	-4134.9
No. Observations:	511	AIC:	8304.
Df Residuals:	494	BIC:	8376.
Df Model:	16		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	1280.4793	268.313	4.772	0.000	753.303	1807.655
temperatura	5076.9496	364.771	13.918	0.000	4360.255	5793.644
umidade	-1170.1294	337.658	-3.465	0.001	-1833.553	-506.706
vel_vento	-2470.6194	517.814	-4.771	0.000	-3488.009	-1453.230
estacao_2	1067.9020	134.970	7.912	0.000	802.715	1333.089
estacao_3	805.0801	177.537	4.535	0.000	456.259	1153.901
estacao_4	1501.8086	112.956	13.295	0.000	1279.874	1723.743
ano_1	2081.7851	72.871	28.568	0.000	1938.610	2224.961
feriado_1	-497.0951	190.435	-2.610	0.009	-871.257	-122.933
dia_semana_1	-219.3825	89.090	-2.462	0.014	-394.424	-44.341
dia_semana_2	-67.8356	92.962	-0.730	0.466	-250.485	114.814
dia_semana_3	34.1617	96.843	0.353	0.724	-156.113	224.437
dia_semana_4	18.0619	97.111	0.186	0.853	-172.739	208.863
dia_semana_5	64.0076	98.883	0.647	0.518	-130.275	258.290
dia_semana_6	370.8155	132.143	2.806	0.005	111.185	630.446
dia_util_1	326.1083	82.207	3.967	0.000	164.589	487.628
clima_2	-456.3215	94.393	-4.834	0.000	-641.782	-270.861
clima_3	-2069.9972	244.814	-8.455	0.000	-2551.002	-1588.992

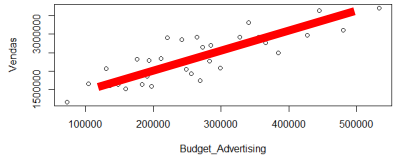
Omnibus:	53.648	Durbin-Watson:	2.126
Prob(Omnibus):	0.000	Jarque-Bera (JB):	116.161
Skew:	-0.587	Prob(JB):	5.97e-26
Kurtosis:	5.019	Cond. No.	5.66e+15

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The smallest eigenvalue is 4.65e-29. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.

Faça a previsão das vendas (R$) mensal no período de 12 meses da empresa XYZ a partir dos dados disponíveis de Vendas (R$) e Budget Advertising (R$) da empresa.

EXEMPLO
Saída Python

	Data	ano	Vendas	Budget_Advertising
1	jan/16	2016	1160081	72800
2	fev/16	2016	1622540	123392
3	mar/16	2016	1597260	135761
4	abr/16	2016	1640675	148064
5	mai/16	2016	1511270	159746
6	jun/16	2016	1634073	183353
7	jul/16	2016	1856971	190722
8	ago/16	2016	1585566	197802
9	set/16	2016	2041672	248891
10	out/16	2016	1933557	256353
11	nov/16	2016	2076910	298805
12	dez/16	2016	1740202	268925




```
> mc = cor(dadosquant);mc
                    Vendas Budget_Advertising
Vendas              1.000              0.852
Budget_Advertising  0.852              1.000
```

Modelo de regressão linear simples

Y = b0 + b1*X1 **(teórico)**

```
import statsmodels.api as sm
from scipy import stats
from statsmodels.compat import lzip
from scipy.stats.stats import pearsonr

# Ordinary Least Square (OLS)
X_ = sm.add_constant(X)
est = sm.OLS(y, X_)
est2 = est.fit()
print(est2.summary())
```

Adiciona o intercepto no modelo.

OLS Regression Results

Dep. Variable:	Vendas	R-squared:	0.725
Model:	OLS	Adj. R-squared:	0.717
Method:	Least Squares	F-statistic:	89.75
Date:	Mon, 21 Jun 2021	Prob (F-statistic):	4.58e-11
Time:	23:18:47	Log-Likelihood:	-509.57
No. Observations:	36	AIC:	1023.
Df Residuals:	34	BIC:	1026.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	1.061e+06	1.52e+05	6.988	0.000	7.52e+05	1.37e+06
Budget_Advertising	4.9641	0.524	9.473	0.000	3.899	6.029

Omnibus:	1.236	Durbin-Watson:	0.781
Prob(Omnibus):	0.539	Jarque-Bera (JB):	1.112
Skew:	0.256	Prob(JB):	0.574
Kurtosis:	2.309	Cond. No.	7.54e+05

Faça a previsão das vendas (R$) mensal no período de 12 meses da empresa XYZ a partir dos dados disponíveis de Vendas (R$) e Budget Advertising (R$) da empresa.

EXEMPLO
Saída Python

OLS Regression Results

Dep. Variable:	Vendas	R-squared:	0.725
Model:	OLS	Adj. R-squared:	0.717
Method:	Least Squares	F-statistic:	89.75
Date:	Mon, 21 Jun 2021	Prob (F-statistic):	4.58e-11
Time:	23:18:47	Log-Likelihood:	-509.57
No. Observations:	36	AIC:	1023.
Df Residuals:	34	BIC:	1026.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	1.061e+06	1.52e+05	6.988	0.000	7.52e+05	1.37e+06
Budget_Advertising	4.9641	0.524	9.473	0.000	3.899	6.029

Omnibus:	1.236	Durbin-Watson:	0.781
Prob(Omnibus):	0.539	Jarque-Bera (JB):	1.112
Skew:	0.256	Prob(JB):	0.574
Kurtosis:	2.309	Cond. No.	7.54e+05

Adequação do modelo ajustado

Coeficientes do modelo

Análise de resíduos

Faça a previsão das vendas (R$) mensal no período de 12 meses da empresa XYZ a partir dos dados disponíveis de Vendas (R$) e Budget Advertising (R$) da empresa.

EXEMPLO
Saída Python

OLS Regression Results

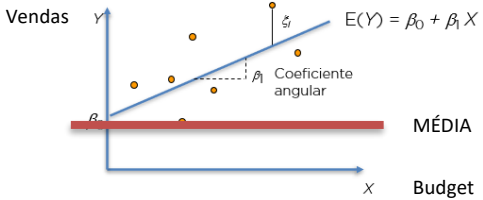
Dep. Variable:	Vendas	R-squared:	0.725
Model:	OLS	Adj. R-squared:	0.717
Method:	Least Squares	F-statistic:	89.75
Date:	Mon, 21 Jun 2021	Prob (F-statistic):	4.58e-11
Time:	23:18:47	Log-Likelihood:	-509.57
No. Observations:	36	AIC:	1023.
Df Residuals:	34	BIC:	1026.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	1.061e+06	1.52e+05	6.988	0.000	7.52e+05	1.37e+06
Budget_Advertising	4.9641	0.524	9.473	0.000	3.899	6.029

Omnibus:	1.236	Durbin-Watson:	0.781
Prob(Omnibus):	0.539	Jarque-Bera (JB):	1.112
Skew:	0.256	Prob(JB):	0.574
Kurtosis:	2.309	Cond. No.	7.54e+05

Acurácia do modelo:
A variável Budget explica 72,5% da variação das Vendas.

Análise de variância (ANOVA)
H0: regressão = média
H1: regressão <> média



Faça a previsão das vendas (R$) mensal no período de 12 meses da empresa XYZ a partir dos dados disponíveis de Vendas (R$) e Budget Advertising (R$) da empresa.

EXEMPLO
Saída Python

Hipótese estatística:

$H_0 : B0 = 0$

$H_1 : B0 \neq 0$

Critério de decisão:

n= 36

gl = n-1

α = 0,05

$t_{0,05} = 2.030$

(Tabela t-Student)

OLS Regression Results

Dep. Variable:	Vendas	R-squared:	0.725
Model:	OLS	Adj. R-squared:	0.717
Method:	Least Squares	F-statistic:	89.75
Date:	Mon, 21 Jun 2021	Prob (F-statistic):	4.58e-11
Time:	23:18:47	Log-Likelihood:	-509.57
No. Observations:	36	AIC:	1023.
Df Residuals:	34	BIC:	1026.
Df Model:	1		
Covariance Type:	nonrobust		

p-value | Intervalo de confiança

	coef	std err	t	P>|t|	[0.025	0.975]
b0 const	1.061e+06	1.52e+05	6.988	0.000	7.52e+05	1.37e+06
b1 Budget_Advertising	4.9641	0.524	9.473	0.000	3.899	6.029

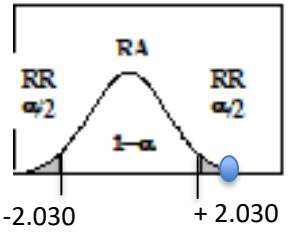
Omnibus:	1.236	Durbin-Watson:	0.781
Prob(Omnibus):	0.539	Jarque-Bera (JB):	1.112
Skew:	0.256	Prob(JB):	0.574
Kurtosis:	2.309	Cond. No.	7.54e+05

$$t_{observado} = 6.988$$

Vendas = b0 + b1 * Budget

(modelo ajustado) ➡ **Vendas = 1060550 + 4.964 * Budget**

Teste Bilateral



Faça a previsão das vendas (R$) mensal no período de 12 meses da empresa XYZ a partir dos dados disponíveis de Vendas (R$) e Budget Advertising (R$) da empresa.

EXEMPLO
Saída Python

OLS Regression Results

Dep. Variable:	Vendas	R-squared:	0.725
Model:	OLS	Adj. R-squared:	0.717
Method:	Least Squares	F-statistic:	89.75
Date:	Mon, 21 Jun 2021	Prob (F-statistic):	4.58e-11
Time:	23:18:47	Log-Likelihood:	-509.57
No. Observations:	36	AIC:	1023.
Df Residuals:	34	BIC:	1026.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	1.061e+06	1.52e+05	6.988	0.000	7.52e+05	1.37e+06
Budget_Advertising	4.9641	0.524	9.473	0.000	3.899	6.029

Omnibus:	1.236	Durbin-Watson:	0.781
Prob(Omnibus):	0.539	Jarque-Bera (JB):	1.112
Skew:	0.256	Prob(JB):	0.574
Kurtosis:	2.309	Cond. No.	7.54e+05

Análise de resíduos

Verificar se há correlação entre os resíduos: Se Durbin = 2 não há correlação

Teste de normalidade dos resíduos:
Se p-valor < 0,05 → Distribuição <> Normal
Se p-valor >= 0,05 → Distribuição = Normal

Faça a previsão das vendas (R$) mensal no período de 12 meses da empresa XYZ a partir dos dados disponíveis de Vendas (R$) e Budget Advertising (R$) da empresa.

EXEMPLO
Saída RStudio

```
modelo <- lm(Vendas ~ Budget_Advertising)
summary(modelo)
```

Projeção para 2019

		Vendas	Budget
jan/19	2019	**1.512.274**	91000
fev/19	2019		154240
mar/19	2019		169702
abr/19	2019		185081
mai/19	2019		199683
jun/19	2019		229192
jul/19	2019		238403
ago/19	2019		247253
set/19	2019		311114
out/19	2019		320442
nov/19	2019		373507
dez/19	2019		336157

Modelo de regressão linear simples

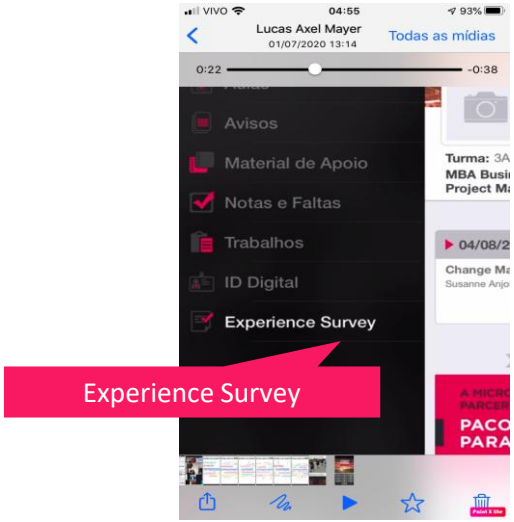
```
> summary(modelo)

Call:
lm(formula = Vendas ~ Budget_Advertising)

Residuals:
    Min      1Q  Median      3Q     Max
-655330 -256271  -30444  234875  743028

Coefficients:
                    Estimate  Std. Error t value       Pr(>|t|)
(Intercept)       1060550.396  151771.308    6.99 0.000000046312 ***
Budget_Advertising        4.964       0.524    9.47 0.000000000046 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 350000 on 34 degrees of freedom
Multiple R-squared:  0.725,     Adjusted R-squared:  0.717
F-statistic: 89.7 on 1 and 34 DF,  p-value: 0.0000000000458
```

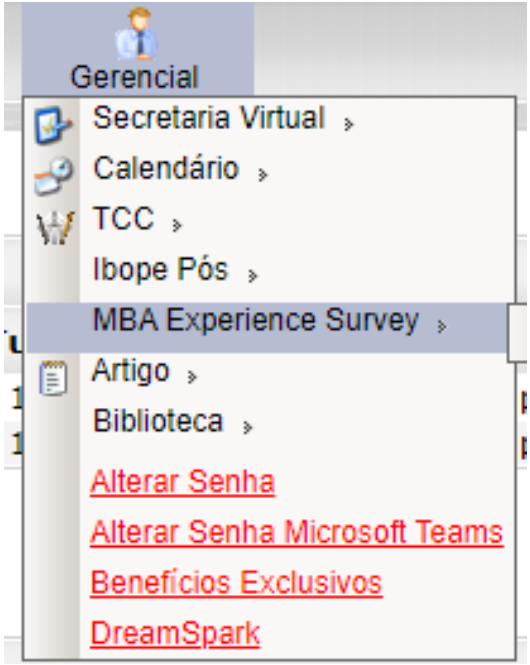
Vendas janeiro/19 = 1060550 + 4.964 * 91000 = 1.512.274

O que você achou da aula de hoje?

Pelo aplicativo da FIAP

(Entrar no FIAPP, e no menu clicar em Experience Survey)





A grande finalidade do conhecimento não é conhecer, mas agir.

T. Huxley

OBRIGADO

 / Regina T. I. Bernal


FIAP