analise de regressão multipla

Amauri dos Santos Lima Neto

2022-12-11

Bibliotecas usadas

if(!require(pacman)) install.packages("pacman")

**##** Carregando pacotes exigidos**:** pacman

library (pacman)

pacman::p\_load(dplyr, car, rstatix, lmtest, ggpubr,

Quant Psyc, psych, scatterplot3d)

Banco de dados

dados <- read.csv2('Banco de Dados 12.csv', stringsAsFactors

#View *(*dados*)*

glimpse (dados)

**##** Rows: 200

## Columns: 3

**##** $ Notas

=

T) # Carregamento *do* arquivo csv

# Visualiza??o dos dados em janela separada

<dbl> 8.91, 3.24, 9.72, 7.29, 5.94, 4.59, 1.89, 5.67, 5.40, 8.10,~ **##** $ Tempo\_Rev <int> 43, 28, 35, 33, 44, 19, 20, 22, 21, 40, 32, 20, 24, 38, 24,~ **##** $ Tempo\_Sono <int> 10, 7, 7, 7, 5, 5, 1, 9, 7, 7, 7, 2, 8, 6, 7, 5, 5, 8, 8, 8~

No banco de dadoss acima, são listados as notas de 200 alunos e será analizado se o tempo gasto em revisão e o tempo de sono impactam nas notas desses alunos.

Construindo o modelo

mod <- 1m (Notas

~

Tempo\_Rev Tempo\_Sono, dados)

Analise gráfica

par (mfrow=c(2,2))

plot (mod)

1

√Standardized residualsl

0.0

1.0

Residuals

0

9-

4

Residuals vs Fitted

goo

680

1640

2 3 4 5 6 7 8 9

Fitted values

Scale-Location

1640

2 3 4 5 6 7 8 9

Fitted values

Standardized residuals

-3 0 2

Standardized residuals

-3 0 2

-3

-2

Normal Q-Q

-1 0 1 2 3

Theoretical Quantiles

Residuals vs Leverage

6100

Coobfdistance

0.00 0.02 0.04

0.06 0.08

Leverage

par (mfrow=c(1,1))

No primeiro gráfico são apresentados os resíduos no eixo y e os valores ajustados no eixo x, e podemos ter uma avaliação da linearidade, pois se existir, a linha apresentada em vermelho se apresenta aproximadamente vertical.

No segundo gráfico, podemos verificar se os residuos apresentam distribuição normal, tendo como eixos os residuos padronizados e os residuos teoricos caso os dados realmente seguisse uma distribuição normal, a verificação é feita analizando se todos os valores tem o comportamento da diagonal apresentada.

No terceiro gráfico, é capaz de mostrar se existe homocedasticidade, caso exista, os dados tem uma dispersão aproximadamente retangular, confirmando a homocedasticidade dos nossos residuos.

No ultimo gráfico, é possivel verificar se existe presença de outliers ou pontos influentes, caso exista, aparecerá uma linha pontilhada e esses valores estarão ultrapassando esse limite.

Testando os pressupostos

Normalidade dos resíduos:

shapiro.test(mod$residuals)

**##**

## Shapiro-Wilk normality test

**##**

**##** data: mod$residuals

**##** W

=

0.99555, p-value

= 0.8269

2

Nesse teste temos como H0: distribuição dos dados

=

Normal se p> 0,05 Como como nosso p foi de 0,8269

consideramos que os residuos tem uma distribuição aproximadamente normal.

Existência de outliers

summary(rstandard (mod))

##

Min**.**

1st Qu.

Median

Mean

3rd Qu. 0.6220131

Max**.**

2.6209702

**##** -2.9693035 -0.7338720 -0.0079448 -0.0009045

nesse caso, basta olhar se existe algum valor fora do intervalo [-3, 3], e como não existe nenhum, descartamos a presença de outliers.

independência dos resíduos

durbinWatsonTest (mod)

## lag Autocorrelation D-W Statistic p-value

**##**

1

**##**

0.08871924

1.811357 0.176

Alternative hypothesis: rho != 0

Nesse caso temos como H0: não exista autocorrelaçao. como nosso p>0,05 aceitamos que não há autocorrelação entre os resíduos.

Homocedasticidade

bptest (mod)

##

**##**

studentized Breusch-Pagan test

**##**

**##** data: mod

## BP =

1.9771, df

=

2, p-value = 0.3721

Neste teste temos como H0: há homocedasticidade. como nosso p> 0,05 confirmamos que existe homocedas- ticidade.

Ausencia de multicolineariedade

pairs.panels (dados)

3

20

20

40

60

●●

0 2

4 6 8 10

**##**

1.034254 1.034254

Notas

0 10 20

330

410

50 60

0.60

Tempo\_Rev

0.33

0.18

Tempo\_Sono

2 4 6 8 10

6

4

2

8 10

No caso da regressão multipla, não deve existir uma correlação muito alta entre as variaveis independentes. (sugere-se que existe multicolinearidade quando o coeficiente de pearson "r" é maior que 0,8)

vif (mod)

**##** Tempo\_Rev Tempo\_Sono

mod2 < 1m (Notas

~

Tempo\_Rev, dados)

outro meio de se verificar a multicolinearidade é atraves do vif (fator de inflação) e dizemos que existe multicoliaridade se o vif > 10

Comparando modelos:

Gerando um **novo** modelo:

Analizando os modelos:

summary (mod)

**##**

**##** Call:

**##** lm (formula

##

**##** Residuals:

= Notas Tempo\_Rev + Tempo\_Sono, data = dados)

4

0

2 4

6 8 10

##

Min

1Q Median

3Q

Max

4.3810

**##** -4.9348 -1.2260 -0.0133 1.0402

##

**##** Coefficients**:**

**##**

Estimate Std. Error t value Pr(>|t|)

**##** (Intercept) 0.117173

0.614517 0.191

0.849

## Tempo\_Rev

0.099101

0.009905 10.005 <2e-16 \*\*\*

**##** Tempo\_Sono

0.350658

0.087098 4.026 8.09e-05 \*\*\*

**##**

## Signif. codes: 0 \*\*\* ##

0.001 **'**\*\*' 0.01 **\*'** 0.05 .' 0.1

1

**##** Residual standard error: 1.686 on 197 degrees of freedom **##** Multiple R-squared: 0.4075, Adjusted R-squared: 0.4014 **##** F-statistic: 67.73 **on** 2 and 197 DF, p-value: < 2.2e-16

ou seja, analizando o modelo criado anteriormente, vemos que conseguimos a seguinte equação Y= 0,117 + 0,1\* x1 +0,35\*x2 e tem uma capaticidade de explicar a variação dos dados de 40%

summary (mod2)

##

## Call:

**##** 1m (formula

=

Notas Tempo\_Rev, data

=

dados)

**##**

**##** Residuals**:**

##

Min

1Q Median

**##** -4.6885 -1.2931 0.0404 1.0549

**##**

**##** Coefficients:

**##**

**##** (Intercept) 2.29156

3Q

Max

3.9267

## Tempo\_Rev

0.10636

Estimate Std. Error t value Pr(>|t|)

0.30421 7.533 1.73e-12 \*\*\*

0.01011

10.524 <2e-16 \*\*\*

##

## Signif. codes:

0

\*\*\*

1

1

0.001 \*\* 0.01 \* ' 0.05

0.1

1

**##**

## Residual standard error: 1.749 **on** 198 degrees **of** freedom ## Multiple R-squared: 0.3587, Adjusted R-squared**:** 0.3555 **##** F-statistic: 110.7 on 1 and 198 DF, p-value: < 2.2e-16

neste outro modelo, temos uma regressão simples com equação Y = 2,29 +0,10x e com capacidade de explicar a variação dos dados de 35,5%

coeficientes padronizados:

lm.beta (mod)

**##** Tempo\_Rev Tempo\_Sono

**##**

0.5580529 0.2245518

1m.beta (mod2)

## Tempo\_Rev

**##** 0.5989188

os coeficientes padronizados sevem para detectar qual variavel tem maior impacto sobre a variavel independente, neste caso é possível verificar que o tempo de revisão impacta mais na nota que o tempo de sono, o que já era lógico.

5

Intervalo de confiança com *95*%:

confint (mod)

**##**

2.5% 97.5%

## (Intercept) -1.09470256 1.3290495

**##** Tempo\_Rev

## Tempo\_Sono

confint (mod2)

##

0.07956792 0.1186337

0.17889401 0.5224219

2.5% 97.5%

## (Intercept) 1.69165197 2.891463 ## Tempo\_Rev 0.08642781 0.126288

comparação entre os modelos:

AIC(mod, mod2)

**##**

df

AIC

**##** mod 4 781.4294

**##** mod2 3 795.2431

BIC (mod, mod2)

##

df

BIC

**##** mod 4 794.6226

**##** mod2 3 805.1380

Com esses valores, definimos como melhor modelo, o que apresenta o menor valor entre eles, os valores representam a variância não explicada pelos modelos, nesse caso o primeiro modelo explica melhor os dados.

Anovas:

anova (mod, mod2)

## Analysis of Variance Table

##

**##** Model 1: Notas

**##** Model 2: Notas

~

~

Tempo\_Rev Tempo\_Sono Tempo\_Rev

**##** Res.Df RSS Df Sum of Sq

F

Pr (>F)

## 1

## 2

##

197 559.80

198 605.86 -1

-46.06 16.209 8.089e-05 \*\*\*

**##** Signif. codes: 0 \*\*\* ' 0.001 **'**\*\*' 0.01 \* 0.05 .' 0.1

1

o teste acima faz uma analise de variância, comparando os modelos e atribuindo um valor de p com base numa estatistica F, e h0: igualdade entre os modelos, nesse caso como p < 5% então os modelos são diferentes, e o melhor vai ser o com menor RSS (residual sum of squares), então novamente confirmamos que o primeiro modelo é melhor para explicar os dados.

gráfico de dispersão:

~

graph <- scatterplot3d (dados $Notas dados $Tempo\_Rev + dados $Tempo\_Sono,

pch

=

16, angle

=

=

30, color "steelblue", box

=

FALSE,

6

**##**

Notas

0

2

4

O

xlab="Tempo de revisao", ylab="Tempo de sono", zlab="Notas")

graph$plane3d (mod, col="black", draw\_polygon = TRUE)

8

6

10

*8*

6

4

2

10

20

30 40 50

60 70

Tempo de revisao

Método para seleção de modelo:

pacman::p\_load(MASS)

mod. inicial <- 1m (Notas mod.simples <- lm (Notas 1, data

~

Tempo\_Rev Tempo\_Sono, data = dados)

~

=

dados)

stepAIC (mod. inicial**,** scope = list (upper

## Start:

## Notas

~

AIC=211.85

Tempo\_Rev + Tempo\_Sono

**##**

##

Df Sum of Sq

RSS

AIC

## **<**none>

559.80 211.85

##

-

Tempo\_Sono 1

46.06 605.86 225.67

284.47 844.28 292.03

=

mod.inicial,

lower

=

mod.simples), direction "backward")

=

## Tempo\_Rev 1

**##**

## Call:

## lm (formula = Notas Tempo\_Rev Tempo\_Sono, data = dados)

7

10

Tempo de sono

**##** Coefficients**:**

**##** (Intercept)

##

0.1172

Tempo\_Rev 0.0991

Tempo\_Sono 0.3507

com essa função é possivel definir um modelo inicial e assim ele classifica os respectivos AIC, e como visto anteriormente o melhor é o com menor valor, com essa configuração são adicionadas variaveis independentes até que se ache a melhor formação, nesse caso foi a com todas as variaveis independentes.

conclusão

A regressão linear múltipla mostrou que o tempo de revisão e o tempo de sono tem efeito sobre as notas dos alunos, A cada minuto gasto com revisão acrescenta na nota em média 0,1 pontos, já a cada hora de sono acrescenta em média 0,35 em média na nota final do aluno, e com uma capacidade de predição de 40%.

∞