

UNIVERSIDADE DE PASSO FUNDO FACULDADE DE CIÊNCIAS ECONÔMICAS, ADMINISTRATIVAS E CONTÁBEIS CENTRO DE PESQUISA E EXTENSÃO DA FEAC (www.upf.br/cepeac)

Texto para discussão

Texto para discussão Nº 08/2021

MANUAL BÁSICO DE USO DO R PARA RLS E RLM

Caroline Lorensi da Siva Samuel Supptitz

Passo Fundo - RS - Brasil

MANUAL BÁSICO DE USO DO R PARA RLS E RLM

Resumo

Este manual elaborado pelos alunos do mestrado do Programa de Pós Graduação em Administração da Universidade de Passo Fundo (PPGAdm/UPF) tem por objetivo auxiliar na análise de dados estatísticos de modelos de RLS e RLM no software R

Caroline Lorensi da Siva/Samuel Supptitz 190606@upf.br/154267@upf.br

Sumário

1.	Utilizando o R	3
2.	Comandos básicos	4
	2.1 Operações básicas	4
	2.2 Potência	4
	2.3 Raiz quadrada de 16	4
	2.4 Atribuições	4
3.	Estudo de Caso	5
	3.1 Etapa 1: Importação de dados	5
	3.2 Etapa 2: Analisando o modelo – Estatística Descritiva	6
	3.2.1 Valor médio	6
	3.2.2 Mediana	7
	3.2.3 Moda	7
	3.2.4 Amplitude	7
	3.2.5 Quartil	7
	3.2.6 Variância	8
	3.2.7 Desvio padrão	8
	3.2.8 Coeficiente de correlação	8
	3.2.9 Dados de amplitude, quartil, mediana e média – estatística descritiva da base	8
	3.2.10 Teste para NORMALIDADE SHAPIRO WILK (valores de p > 0,05 indicam dados normais)	
		8
	3.3 Etapa 3: Analisando o modelo – Regressões RLS e RLM	9
	3.3.1 Primeiro modelo RLS- (tempo em função da distância)	9
	3.3.2 Segundo Modelo RLM – (tempo em função da distância + quantidades de semáforos1)12	1
	3.3.3 Terceiro Modelo RLM – (tempo em função da distância + quantidades de semáforos2) 13	3
	3.3.4 Quarto Modelo RLM – (tempo em função da distância + quantidades de semáforos3)15	5
4.	Gráficos16	5
	4.1 Amplitude interquartil	7
	4.2 Relação entre as variáveis	8
	4.3 Análise visual para homogeneidade dos resíduos (visualmente eles devem se distribuir igualmente abaixo e acima da linha)	9
	4.4 Distribuição dos resíduos	0
	4.5 Normalidade dos resíduos21	1
	4.6 Detecção de valores alavanca e significativos	1

4.7 Exportar csv	
Referências	

1. Utilizando o R

Descreveremos a seguir como executar e interpretar um modelo estatístico, extraindo dados de estatística descritiva, regressão linear simples, regressão linear múltipla e correlações, utilizando o programa de *software* R Studio. Este *software* pode ser utilizado e obtido de forma gratuita, acessando o o site <u>https://www.rstudio.com/products/rstudio/download/</u>. Além disso, em Resources, você encontrará informações, livros, vídeos, e outros recursos que lhe auxiliarão a utilizar o programa. Inicialmente orientaremos sobre o uso de comando e operações básicas e posteriormente utilizaremos um estudo de caso para auxiliar o entendimento e análise dos resultados.

2. Comandos básicos

2.1 Operações básicas 5+5 10-6 10*2 10/2

2.2 Potência 5**2

2.3 Raiz quadrada de 16 Sqrt (16)

2.4 Atribuições

Atribuições são muito importantes para podermos utilizar o valor de uma variável novamente. É como se disséssemos para o R "realize essa operação e guarde em um objeto chamado x". Para atribuir algo a um valor usamos o "nome da variável" seguido de <- seguido da "operação". A atribuição será bastante utilizada para desenvolvimento de operações, exportação de dados e construção de gráficos.

os parênteses são usados para estabelecer uma ordem de operações, igual aprendemos na matemática

5*(50+10)

3. Estudo de Caso

Para demonstrar como aplicar na prática os recursos do R utilizaremos um modelo que possui informações de tempo mínimo do percurso, distância do percurso e 3 diferentes valores de quantidade de cruzamentos existentes no percurso.

Veja pela tabela que o Tempo é nossa variável dependente (explicada), a Distância e Quantidade de Cruzamentos nossas variáveis independentes (explicativas).

Tempo (min) (Y)	Distância (km) (X1)	Quantidade de Cruzamantos (X2)	Quantidade de Cruzamantos (X2)	Quantidade de Cruzamantos (X2)
15	8	16	32	12
20	6	12	24	20
20	15	30	60	25
40	20	40	39	37
50	25	50	100	32
25	11	22	44	17
10	5	10	20	9
55	32	64	128	60
35	28	56	112	12
30	20	40	80	17

Para o estudo, os dados foram dispostos no formato da tabela abaixo:

3.1 Etapa 1: Importação de dados

Acesse o RStudio > Quadrante superior direito > Import Dataset > From "formato do arquivo a ser importado"



Selecione o arquivo clicando em Browse... > Arquivo > Open > Import

	nport Excel Data							
	ile/URL:							
	C:/Users/caroli	ine.lorensi/OneDrive	- Metadados Asse	ssoria e Sistemas/F	Projeto/Métodos Quar	titativos/dados_tempo	distandavitx	
	Data Descince							
	Jata Pleview:		Ourselidede	O contridu da	Ourstidada			
	Тетро	Distância	de	de	de			
	(min) (Y)	(km) (X1)	Cruzamantos (X2)3	Cruzamantos (X2)4	Cruzamantos (X2)5			
	(numeric)	(numeric)	(numeric)	(numeric)	* (numeric)	~		
	15	1 8		16	32	12		
	20	5 6		12	24	20		
	20	1 15		30	60	25		
a a a a a a a a a a a a a a a a a a a a a a a a a a a a a a b a a a a b a a a a b a a a a b a a a a b a a a a b a a a a b a a a a b a a a a b a a a a a a a a a a a a b a a a a a b a a a a a b a a a	43	1 20		40	39	37		
a 1 2 4 17 a 1 2 4 17 a 1 1 10 10 a 1 10 10 10 10 a 1 10 <td< td=""><td>50</td><td>i 25</td><td></td><td>50</td><td>100</td><td>32</td><td></td><td></td></td<>	50	i 25		50	100	32		
iii i i iii iii iii iii iii iii iii iiii iiii iii iii iiii iiiii iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii	25	i 11		22	44	17		
	10	/ 5		10	20	9		
31 33 90 10 10 32 30 40 10 10	55	i 32		64	128	60		
a) a) a) a) a) a) a) a) a) a) a)	35	/ 28		56	112	12		
Stating for 50 stmin. Inf Open Service Stating for 50 stmin. Stating for 50 stmin. <t< td=""><td>30</td><td>J 23</td><td></td><td>40</td><td>80</td><td>17</td><td></td><td></td></t<>	30	J 23		40	80	17		
Heing for 25 arms. er Colors er Colors for Source Color Prever Name: Solor (Starter) & Name Color (Starter) & Starter) & Starter) Marker Solor (Starter) & Starter) & Starter) Marker (Starter) & Starter) & Starter) & Starter) Marker								
Standig ford Standia. of Openor Standia. Standia. <								
Heing fot 25 arms. en Colores Color Prever Name: Solor (1990 Color Color Color Prever Name: Solor (1990 Color Color Color Prever Name: Solor (1990 Color Color Color Color Color (1990 Color Color (1990 Color Color (1990								
Stating for 50 series. of Opener Station (Station (S								
Hang Star Starms. et Option: Cone Parlam: Name: Solo (terps) (Star Ca) & Star Star Star Star Star Star Star Star								
Name: decision: Code Preview: Code P								
Henry for D Arms. er Options Conference Stateward Arms (Stateward Arms) (
Name: Society Status St								
Interpret to Bierres. or Option: Bander: Code Preserc Interpret (redsk1)								
Name: Code Prevec Code Prevecc Code Preveccc Code Preveccc Code Preveccc Code Preveccc Code Prevecccc <th< td=""><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></th<>								
internet. Code Prenex. Code Pr								
Heining ford () servel. of Cylores Code Preview Stande: Society of Starces Assessor is a Starces Assessor is								
Interpreted to the serve. Code Presect Interpreted to the serve. Interpreted to the serve. Deve. Interpreted to the serve. View(dadds_temp2,distance.s.r/s.r) View(dadds_temp2,distance.s.r) View(dadds_temp2,distance.s.r) View(dadds_temp2,distance.s.r) View(dadds_temp2,distance.s) View(dadds_temp2,distance.s)								
Haing for [2 ansi. er Options Talma: doog terpo_district a for each for e								
inlang fot Di erne. of Oproc. Cole Panec Intravy/reddition Intravy/reddition Intravy/reddition Name Opto Opto Opto Intravy/reddition Intravy/reddition Intravy/reddition Intravy/reddition Intravy/reddition <td< td=""><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></td<>								
Heing fot 25 arms: en Colore Name: Solo Steros Janes Name: Solo Steros Janes Name: Solo Steros Janes Name: Solo Steros Janes Name: Name: N								
inleng fot (2 ems. ent gloros, tempo, districts, for an ent status and the statu								
New State St								
Haveg for 50 error. or 50 prove: Same Set 50								
Heing for 23 ama: or Oprove: Color Peries: Name: Good tempo, distancia Mar. Saue: Prime Reversioners: Color Peries: Color Oprove: - Metadados Assessoria e Sistemas/Projeto Metadade: Supe: D @ Open Data Veries: D @ Open Data Veries: D @ Open Data Veries: Color Data Color Co								
Howay Set 28 January enclosed by CS2 January Rame <u>Geody States</u> <u>Code Panew</u> <u>Code Panew</u> <u>Code</u>								
Cold Prevec Cold Prevec Name: @Ref Rev at Names (Addos_tempo_distance) @Ref Rev at Names	Previewing first 9	0 antries						
or Optime: Code Pareier: Code								
Name Geborg, detando Max Name <pre></pre>	mport Options:						Code Preview:	
Deve:	Name: da	idos tempo distanc	it May Rows		First Row as Names		[liprary(readx]) dados supuno distancia sand avenal/""" (inare/sanalina losansi/Dominiua - Natadados Astantanta - Sistemas/Mendiato	
Deel: Deelad: V Sig: 0 XOPA: Dial Anno: None: Coll.types: Coll.types: Coll.types: Coll.types: Name: Coll.types: Coll.type: Coll.type: Coll.type: Name: Coll.type: Coll.type: Coll.type: Name: Coll.type: Coll.type: Coll.type: View(dado:_teeps_driter(t)) View(dado:_teeps_driter(t)) Nome:	marine. Los	cos_compo_crossre	in an normal				dado_cenp_distancia <-rea_excet_croadistanonia interiorenzionen ine energiana e energiana e anternazionen e energiana e anternazionen e energiana e anternazionen e energiana e e anternazionen e e energiana e e anternazionen e e energiana e e energiana e e e e e e e e e e e e e e e e e e	
Nege (ALDO NA View(dado_temp_distancia) '' teading Conf Res using read/	Sheet: De	fault	 Skip: 	0	Open Data Viewer		col_types = c("numeric", "numeric", "numeric", "numeric", "numeric",	
teading Door fites using read/	Range: A1		NA:				View(dados_tempo_distancia)	
tading Dari Re uling stadt								
	Beading Facel	files using readul						-

Abra um New File para abrir o programador:

Fil	e	<u>E</u> dit	<u>C</u> ode	<u>V</u> iew	<u>P</u> lots	Session	<u>B</u> uild	<u>D</u> ebu	g <u>P</u> rofile	<u>T</u> ools	<u>H</u> elp
0	•	Q	* • 🔚	81 🖷		Go to file/funct	ion	- 88	Addins +		

Nos exemplos a seguir o nome da base será dados_tempo_distância e as colunas serão:

Tempo	Quantidade	 Quantidade	Quantidade
(min)	de	de	de
(Y)	Cruzamantos	Cruzamantos	Cruzamantos
(X1)	(X2)3	(X2)4	(X2)5

3.2 Etapa 2: Analisando o modelo – Estatística Descritiva

Sempre digite no programados o script, selecione a linha e após selecione a linha digitada e executar na tecla RUN que irá criar o modelo.

👄 Run 🐤 🕞 Source 🖌 🗟

3.2.1 Valor médio

É a soma do total de valores de determinada variável (discreta ou contínua) dividida pelo número total de observações.

É utilizada, por exemplo, para calcular a média de gols em uma partida de futebol; a média de salários dos funcionários de uma empresa; a variação de taxa de câmbio do dólar; dentre outros. A Média é definida por

$$Média = \frac{Soma \ de \ todos \ os \ valores}{Número \ de \ valores \ somados}$$

COMO PROCEDER NO R: Mean(data\$coluna)

3.2.2 Mediana

É uma medida de localização do centro da distribuição de um conjunto de dados ordenados de forma crescente. Seu valor separa a série em duas partes iguais, de modo que 50% dos elementos são menores ou iguais à mediana e os outros 50% são maiores ou iguais à mediana. A Mediana é definida por:

Mediana =
$$\frac{n+1}{2}$$

COMO PROCEDER NO R: median(data\$coluna)

3.2.3 Moda

É o valor do conjunto que mais se repete, este valor pode ser:

- amodal
- bimodal
- trimodal

-....

COMO PROCEDER NO R: mode(data\$coluna)

3.2.4 Amplitude

É a diferença entre os valores extremos do conjunto. É definida como sendo a diferença entre o maior e o menor valor dos dados observados.

COMO PROCEDER NO R: range(data\$coluna)

3.2.5 Quartil

São medidas que dividem os dados em quatro partes iguais. O segundo quartl é exatamente igual a mediana.

COMO PROCEDER NO R: quantile(data\$coluna)

3.2.6 Variância A variância é definida por:

$$s^{2} = \frac{\sum_{i=1}^{n} (x-x)^{2}}{n-1}$$

Como proceder no R:

var(data\$coluna) #para variância de uma coluna var(data) #para variância de toda a tabela

3.2.7 Desvio padrãoO desvio padrão é definido por:

$$s = \sqrt{s^2}$$

COMO PROCEDER NO R: sd(data\$coluna)

3.2.8 Coeficiente de correlação Refere-se ao grau de associação linear entre x e y.

Como proceder no R:

Cor(data\$coluna,data\$coluna) #para correlação de duas variáveis

Cor(data) #para correlação de toda a tabela

3.2.9 Dados de amplitude, quartil, mediana e média – estatística descritiva da base COMO PROCEDER NO R:

summary(data\$coluna) #para dados de uma coluna

summary(data) #para dados de toda a tabela

3.2.10 Teste para NORMALIDADE SHAPIRO WILK (valores de p > 0,05 indicam dados normais)

O Teste de Shapiro-Wilk (1965), também conhecido por Teste W, é um procedimento eficiente para avaliar a suposição de normalidade contra um amplo espectro de alternativas não normal, principalmente se é dado um número relativamente pequeno de observação. O teste W é recomendado para amostras com menos de 2.000 observações, acima de 2.000 observações aconselha-se o Teste K (Kolmogorov Smirnov)

COMO PROCEDER NO R: shapiro.test(rstudent(nomedomodelo))

3.3 Etapa 3: Analisando o modelo - Regressões RLS e RLM

"A análise de regressão diz respeito ao estudo da dependência de uma variável, a variável dependente, em relação a uma ou mais variáveis, as variáveis explanatórias, visando estimar e/ou prever o valor médio (da população) da primeira em termos dos valores conhecidos ou fixados (em amostragens repetidas) das segundas" (GUJARATI; PORTER, 2011, p. 38). Utilizando a base "dados tempo distância", foram realizadas as regressões lineares.

3.3.1 Primeiro modelo RLS- (tempo em função da distância) $y = b_1+b_2 * X_2+e$

Y=variável dependente

 b_1 =coeficiente linear (onde a reta corta o eixo Y)

- b_2 = coeficiente angular (ângulo ou declividade da curva)
- X_2 = variável independente ou explicativa

e = erro da forma funcional (ajuda a analisar se a forma funcional está bem-organizada ou não)

• Reta de regressão do exercício

$$y = b_1 + b_2 * X_2 + e$$
$$T = b_1 + b_2 * D$$

T(tempo)= variável dependente ou resposta

 b_1 = coeficiente linear

 b_2 = coeficiente angular

D= distância (variável explicativa)

Como proceder no R:

#A função para regressão é "lm" e não requer pacote estatístico (variavel resposta ~ variável preditora)

lm(y~x, data = "base de dados") #para regressão linear simples

lm(y~x1+x2, "data=base de dados") #para regressão linear múltipla

#Sumário dos resultados do modelo

summary(nomedomodelo)

Digitar da seguinte forma no programador (tela superior direita):

mood é o nome do modelo, seguido de <-, seguido de **lm**, seguido do "nome da planilha", seguido de \$, seguido do "nome da coluna", seguido de ~, seguido do "nome da planilha", seguido de \$, seguido do nome da "outra coluna analisada", como no exemplo a seguir:

mood <-lm(dados_tempo_distancia\$`Tempo (min) (Y)`~ dados_tempo_distancia\$`Distância
(km) (X1)`)</pre>

Após isso, selecionar a linha digitada e executar na tecla \implies RUN que irá criar o modelo.

Para analisar o modelo, deve-se programar da seguinte forma:

Escrever summary e entre parêntese o nome do modelo criado "mood" como no exemplo a seguir:

summary(mood)

Após isso, selecionar a linha digitada e executar na tecla \rightarrow RUN que irá apresentar os resultados na tela inferior esquerda.

```
Call:
lm(formula = dados_tempo_distancia$`Tempo (min) (Y)` ~ dados_tempo_distancia$`Distância
 (km) (X1)`)
Residuals:
            1Q Median
    Min
                               3Q
                                        Max
-10.6081 -3.9358 0.6419 5.1351 8.6486
Coefficients:
                                            Estimate Std. Error t value Pr(>|t|)
                                             5.8784 4.5323 1.297 0.230788
1.4189 0.2355 6.025 0.000314
(Intercept)
dados_tempo_distancia$`Distância (km) (X1)
(Intercept)
dados_tempo_distancia$`Distância (km) (X1)` ***
signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 6.719 on 8 degrees of freedom
Multiple R-squared: 0.8194, Adjusted R-squared: 0.7969
F-statistic: 36.3 on 1 and 8 DF, p-value: 0.0003144
```

 $b_1 = 5,8784 \text{ (sig}=0,230788)$

 $b_2 = 1,4189$ (sig=0,000314)

a correlação (calculada nos exemplos anteriores) entre as variáveis tempo e distância é de 0,90522 (correlação positiva e forte). O \mathbb{R}^2 , ou poder do modelo é de 81,94%. Analisando o teste F e o sig da regressão, podemos dizer que os parâmetros são diferentes de 0. Entretanto, ao investigar os parâmetros individualmente, teste t ou sig ou valor p, chega-se à conclusão de que apenas o parâmetro b2 é significativo (sig=0,000314<0,05 ou 5%).

 $T = b_1 + b_2 * D$

T=1,418919*D

Dessa forma podemos dizer que a cada uma unidade em que a distancia aumentar, o tempo irá aumentar em 1,418919 unidades.

3.3.2 Segundo Modelo RLM – (tempo em função da distância + quantidades de semáforos1) $y = b_1 + b_2 * X_2 + b_3 * X_3 + e$

y=variável dependente

 b_1 = coeficiente linear (onde a reta corta o eixo Y)

 b_2 e b_3 = coeficientes angulares (ângulo ou declividade da curva)

 $X_2 e X_3 =$ variáveis independentes ou explicativas

e = erro da forma funcional (ajuda a analisar se a forma funcional está bem-organizada ou não)

- Reta de regressão do exercício:

T(tempo) variável dependente ou resposta = y

 b_1 =coeficiente linear

 $b_2 e b_3$ = coeficientes angulares

 $X_2 = D$ (Distância)

 $X_3 = QC_1$ (quantidade de Cruzamentos 1)

$$y = b_1 + b_2 * X_2 + b_3 * X_3 + e$$
$$y = b_1 + b_2 * D + b_3 * QC_1$$

COMO PROCEDER NO R

Digitar da seguinte forma no programador (tela superior direita):

mood2 é o nome do modelo, seguido de <-, seguido de **lm**, seguido da palavra formula, seguido do "nome da planilha", seguido de \$, seguido do "nome da coluna", seguido de ~, seguido do "nome da planilha", seguido de \$, seguido do nome da "outra coluna analisada", seguido de +, seguido do "nome da planilha", seguido de \$, seguido do nome da "outra coluna analisada", seguido de ", data", seguido do "nome da planilha" como no exemplo a seguir:

mood2 <-lm(formula = dados_tempo_distancia\$`Tempo (min)
(Y)`~dados_tempo_distancia\$`Distância (km) (X1)`+dados_tempo_distancia\$`Quantidade de
Cruzamantos (X2)...3`, data = dados_tempo_distancia)</pre>

Após isso, selecionar a linha digitada e executar na tecla \implies RUN que irá criar o modelo.

Para analisar o modelo, deve-se programar da seguinte forma:

Escrever summary e entre parêntese o nome do modelo criado "mood2" como no exemplo a seguir:

summary(mood2)

Após isso, selecionar a linha digitada e executar na tecla \rightarrow RUN que irá apresentar os resultados na tela inferior esquerda.

```
Call:
lm(formula = dados_tempo_distancia$`Tempo (min) (Y)` ~ dados_tempo_distancia$`Distância
 (km) (X1)`+
   dados_tempo_distancia$`Quantidade de Cruzamantos (X2)...3`,
   data = dados_tempo_distancia)
Residuals:
    Min 1Q Median 3Q
                                       Max
-10.6081 -3.9358 0.6419 5.1351 8.6486
Coefficients: (1 not defined because of singularities)
                                                        Estimate Std. Error
(Intercept)
                                                          5.8784 4.5323
dados_tempo_distancia$`Distância (km) (X1)`
                                                          1.4189
                                                                    0.2355
dados_tempo_distancia$`Quantidade de Cruzamantos (X2)...3`
                                                             NA
                                                                        NA
                                                         t value Pr(>|t|)
(Intercept)
                                                          1.297 0.230788
                                                           6.025 0.000314 ***
dados_tempo_distancia$`Distância (km) (X1)`
dados_tempo_distancia$`Quantidade de Cruzamantos (X2)...3`
                                                             NA
                                                                      NA
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 6.719 on 8 degrees of freedom
Multiple R-squared: 0.8194, Adjusted R-squared: 0.7969
F-statistic: 36.3 on 1 and 8 DF, p-value: 0.0003144
```

 $b_1 = 5,8784 \text{ (sig} = 0,230788)$

 $b_2 = 1,4189$ (sig=0,000314)

 b_3 = NA (sig=NA) Tal problema deve ser corrigido com o tratamento dos dados, pois pode haver alguma perturbação na coluna analisada.

O R^2 , ou poder do modelo é de 81,94%. Ao investigar os parâmetros individualmente, teste t ou sig ou valor p, chega-se à conclusão de que apenas o parâmetro b2 é significativo (sig=0,000314<0,05 ou 5%).

$$y = b_1 + b_2 * D + b_3 * QC_1$$

T=1,418919*D+QC₁

Dessa forma podemos dizer que a cada uma unidade em que a distância aumentar, o tempo irá aumentar em 1,418919 unidades. Já a quantidade de cruzamentos 1 não possuiu resposta sem testar a normalidade da série.

3.3.3 Terceiro Modelo RLM – (tempo em função da distância + quantidades de semáforos2)

$$y = b_1 + b_2 * X_2 + b_3 * X_3 + e$$

y=variável dependente

 b_1 = coeficiente linear (onde a reta corta o eixo Y)

 b_2 e b_3 = coeficientes angulares (ângulo ou declividade da curva)

 $X_2 e X_3 =$ variáveis independentes ou explicativas

e = erro da forma funcional (ajuda a analisar se a forma funcional está bem-organizada ou não)

- Reta de regressão do exercício:

T(tempo) variável dependente ou resposta = y

 b_1 =coeficiente linear

 $b_2 e b_3 =$ coeficientes angulares

 $X_2 = D$ (Distância)

 $X_3 = QC_2$ (quantidade de Cruzamentos 2)

$$y = b_1 + b_2 * X_2 + b_3 * X_3 + e$$

 $y = b_1 + b_2 * D + b_3 * QC_2$

COMO PROCEDER NO R

Digitar da seguinte forma no programador (tela superior direita):

mood3 é o nome do modelo, seguido de <-, seguido de **lm**, seguido da palavra formula, seguido do "nome da planilha", seguido de \$, seguido do "nome da coluna", seguido de ~, seguido do "nome da planilha", seguido de \$, seguido do nome da "outra coluna analisada", seguido de +, seguido do "nome da planilha", seguido de \$, seguido do nome da "outra coluna analisada", seguido de ", data", seguido do "nome da planilha" como no exemplo a seguir:

mood3 <-lm(formula = dados_tempo_distancia\$`Tempo (min)
(Y)`~dados_tempo_distancia\$`Distância (km) (X1)`+dados_tempo_distancia\$`Quantidade de
Cruzamantos (X2)...4`, data = dados_tempo_distancia)</pre>

Após isso, selecionar a linha digitada e executar na tecla \implies RUN que irá criar o modelo.

Para analisar o modelo, deve-se programar da seguinte forma:

Escrever summary e entre parêntese o nome do modelo criado "mood3" como no exemplo a seguir:

summary(mood3)

Após isso, selecionar a linha digitada e executar na tecla \rightarrow RUN que irá apresentar os resultados na tela inferior esquerda.

```
call:
lm(formula = dados_tempo_distancia$`Tempo (min) (Y)` ~ dados_tempo_distancia$`Distância
 (km) (X1)`+
    dados_tempo_distancia$`Quantidade de Cruzamantos (X2)...4`,
    data = dados_tempo_distancia)
Residuals:
   Min 1Q Median 3Q Max
-9.700 -3.308 -0.899 4.544 9.485
Coefficients:
                                                          Estimate Std. Error
(Intercept)
                                                            5.6371 4.5975
dados_tempo_distancia$`Distância (km) (X1)`
                                                            2.0254
                                                                       0.7182
dados_tempo_distancia$`Quantidade de Cruzamantos (X2)...4` -0.1576
                                                                       0.1760
                                                          t value Pr(>|t|)
(Intercept)
                                                            1.226 0.2598
dados_tempo_distancia$`Distância (km) (X1)`
                                                            2.820
                                                                    0.0258 *
dados_tempo_distancia$`Quantidade de Cruzamantos (X2)...4` -0.895
                                                                    0.4004
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 6.804 on 7 degrees of freedom
Multiple R-squared: 0.838, Adjusted R-squared: 0.7917
F-statistic: 18.1 on 2 and 7 DF, p-value: 0.001712
b_1 = 5,6371 \text{ (sig}=0,2598)
```

 $b_2 = 2,0254 \text{ (sig=0,0258)}$

 $b_3 = -0,1576$ (sig=0,4004)

O R^2 , ou poder do modelo é de 83,8%. Ao investigar os parâmetros individualmente, teste t ou sig ou valor p, chega-se à conclusão de que apenas o parâmetro b2 é significativo (sig=0,000314<0,05 ou 5%), mas nenhum parâmetro é significativo a 1%.

$$y = b_1 + b_2 * D + b_3 * QC_2$$

T=1,418919*D+QC₂

Dessa forma podemos dizer que a cada uma unidade em que a distância aumentar, o tempo irá aumentar em 1,418919 unidades. Já o coeficiente para quantidade de cruzamentos2 não possuiu significância.

3.3.4 Quarto Modelo RLM – (tempo em função da distância + quantidades de semáforos3) $y = b_1 + b_2 * X_2 + b_3 * X_3 + e$

y=variável dependente

 b_1 = coeficiente linear (onde a reta corta o eixo Y)

 b_2 e b_3 = coeficientes angulares (ângulo ou declividade da curva)

 $X_2 e X_3 =$ variáveis independentes ou explicativas

e = erro da forma funcional (ajuda a analisar se a forma funcional está bem-organizada ou não)

- Reta de regressão do exercício:

T(tempo) variável dependente ou resposta = y

 b_1 =coeficiente linear

 $b_2 e b_3 =$ coeficientes angulares

 $X_2 = D$ (Distância)

 $X_3 = QC_3$ (quantidade de Cruzamentos 3)

$$y = b_1 + b_2 * X_2 + b_3 * X_3 + e$$
$$y = b_1 + b_2 * D + b_3 * QC_3 + e$$

COMO PROCEDER NO R

Digitar da seguinte forma no programador (tela superior direita):

mood4 é o nome do modelo, seguido de <-, seguido de **Im**, seguido da palavra formula, seguido do "nome da planilha", seguido de \$, seguido do "nome da coluna", seguido de ~, seguido do "nome da planilha", seguido de \$, seguido do nome da "outra coluna analisada", seguido de +, seguido do "nome da planilha", seguido de \$, seguido do nome da "outra coluna analisada", seguido de ", data", seguido do "nome da planilha" como no exemplo a seguir:

mood4 <-lm(formula = dados_tempo_distancia\$`Tempo (min) (Y)`~dados_tempo_distancia\$`Distância (km) (X1)`+dados_tempo_distancia\$`Quantidade de Cruzamantos (X2)...5`, data = dados_tempo_distancia)

Após isso, selecionar a linha digitada e executar na tecla \implies RUN que irá criar o modelo.

Para analisar o modelo, deve-se programar da seguinte forma:

Escrever summary e entre parêntese o nome do modelo criado "mood4" como no exemplo a seguir:

summary(mood4)

Após isso, selecionar a linha digitada e executar na tecla \rightarrow RUN que irá apresentar os resultados na tela inferior esquerda.

```
call:
lm(formula = dados_tempo_distancia$`Tempo (min) (Y)` ~ dados_tempo_distancia$`Distância
 (km) (X1)`+
    dados_tempo_distancia$`Quantidade de Cruzamantos (X2)...5`,
    data = dados_tempo_distancia)
Residuals:
             1Q Median
                            3Q
    Min
                                    Мах
-8.2584 -2.0734 -0.9101 2.7025 8.8563
Coefficients:
                                                           Estimate Std. Error
(Intercept)
                                                             3.6635 3.6942
dados_tempo_distancia$`Distância (km) (X1)`
                                                             1.0343
                                                                       0.2448
dados_tempo_distancia$`Quantidade de Cruzamantos (X2)...5`
                                                            0.3632
                                                                        0.1504
                                                           t value Pr(>|t|)
                                                            0.992 0.35439
(Intercept)
dados_tempo_distancia$`Distância (km) (X1)`
                                                             4.224 0.00392 **
dados_tempo_distancia$`Quantidade de Cruzamantos (X2)...5`
                                                            2.415 0.04643 *
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
                                                           '1
Residual standard error: 5.305 on 7 degrees of freedom
Multiple R-squared: 0.9015, Adjusted R-squared: 0.8734
F-statistic: 32.03 on 2 and 7 DF, p-value: 3e-04
b_1 = 3,6635 \text{ (sig}=0,354339)
```

 $b_2 = 1,0343 \text{ (sig=0,00392)}$

 $b_3 = 0,3632 \text{ (sig}=0,04643)$

O R², ou poder do modelo é de 90,15%. Ao investigar os parâmetros individualmente, teste t ou sig ou valor p, chega-se à conclusão que os parâmetros b2 e b3 são significativos (sig=0,000314<0,05 ou 5%).

$$y = b_1 + b_2 * D + b_3 * QC_3$$

 $T=1,0343 * D + 0,3632 * QC_3$

Dessa forma podemos dizer que a cada uma unidade em que a distância aumentar, o tempo irá aumentar em 1,0343 unidades. E a cada unidade que aumentar da quantidade de cruzamentos3, o tempo irá aumentar em 0,3632 unidades.

4. Gráficos

Abra o pacote de gráficos:

library(ggplot2)

library(boxplot)

4.1 Amplitude interquartil

boxplot(d,e) #para plotar no mesmo gráfico (comparação)

boxplot(d); boxplot(e) #para plotar em gráficos diferentes

boxplot(Q1, Q2, Q3, Q4, Q5, col="blue")





4.2 Relação entre as variáveis

O gráfico de Draftman (*Draftman's plot*), também conhecido como *scatterplot matrix* ou gráficos de pares. Com esse plot conseguimos observar os gráficos de dispersão para cada par de variáveis e entender melhor os números que aparecem na matriz de correlação. Porém, esta matriz é redundante ao repetir informação nas diagonais, deixando de lado informações interessantes.

pairs(dados_tempo_distancia, col = 2, pch = 19)



4.3 Análise visual para homogeneidade dos resíduos (visualmente eles devem se distribuir igualmente abaixo e acima da linha) plot(rstudent(resmodelo) ~ fitted(resmodelo), pch = 19)

abline(h = 0, lty = 2)



#Visualização gráfica lty é o tipo da linha 1: linha contínua; 2: linha descontínua

plot(dados_tempo_distancia\$`Tempo (min) (Y)`~dados_tempo_distancia\$`Distância (km) (X1)`+dados_tempo_distancia\$`Quantidade de Cruzamantos (X2)...3`+dados_tempo_distancia\$`Quantidade de Cruzamantos (X2)...4`+dados_tempo_distancia\$`Quantidade de Cruzamantos (X2)...5`, data = dados_tempo_distancia)

resmodelo1<-lm(variavely~variavelx1+variavelx2a)

abline(resmodelo,lty=2)



4.4 Distribuição dos resíduos plot(resmodelo, which = 1)



4.5 Normalidade dos resíduos plot(resmodelo, which = 2)



4.6 Detecção de valores alavanca e significativos plot(resmodelo, which = 5)



4.7 Exportar csv

write.xlsx(data, file = "nomedoarquivo.xlsx")

write.csv(data, file = "nomedoarquivo.csv")

Referências

PACHECO et al. Aprendendo R. Escola Nacional de Saúde Público: Fiocruz, 2017.

GUJARATI, Damodar N. PORTER, Dawn C. **Econometria Básica**. 5. ed. Porto Alegre: AMGH Editora Ltda, 2011.