IPS ESTB

Regressão Linear Múltipla

Unidade Curricular: Análise e Tratamento de Dados Multivariados

Discentes:

2675 Daniel Marçal

2691 Marcelo Pereira

2814 Bernardo Augusto

Docente: Anabela Cardoso Marques

ÍNDICE

INTRODUÇÃO

Neste trabalho foi-nos solicitado o estudo de, uma equação para se estimar a condicional (valor esperado) de uma variável y, dados os valores de algumas outras variáveis x, regressão linear. Para isso utilizámos a ferramenta da RStudio que é um software livre de ambiente de desenvolvimento integrado para R, uma linguagem de programação para gráficos e cálculos estatísticos.

A análise de regressão é uma técnica estatística para investigar e modelar a relação entre variáveis, sendo uma das mais utilizadas na análise de dados. Pode-se citar inúmeras aplicações de análise de regressão na área da saúde (Freedman et al. 2004, Lyles & Kupper 1997, Chen & Wang 2004).Um dos objetivos da análise de regressão é estimar os parâmetros desconhecidos do modelo. Existem várias técnicas de estimação destes parâmetros, neste relatório foi considerado o método dos mínimos quadrados (R-Square).

Numa primeira parte, começámos por construir uma tabela em Excel que tinha como base os dados fornecidos pelo enunciado que nos foi entregue.Após feita esta tarefa, definimos os nossos objetivos principais. Sendo eles: analisar descritivamente as variáveis; ajustar um modelo de regrassão linear; avaliar criticamente o poder explicativo do modelo; avaliar os pressupostos do modelo por recurso à análise de resíduos; avaliar a significância do modelo e dos seus preditores.

R é um ambiente computacional e uma linguagem de programação que se tem vindo a especializar na manipulação, análise e visualização gráfica de dados. Atualmente é considerado o melhor ambiente computacional para esta finalidade. O programa está disponível para diferentes sistemas operacionais: Unix/Linux, Mac e Windows. Foi criado por Ross Ihaka e por Robert Gentleman no departamento de Estatística da Universidade de Auckland, Nova Zelândia. Posteriormente, desenvolvido pelo esforço colaborativo de pessoas em vários locais do mundo. O nome R provém das iniciais dos criadores (Ross e Robert) como também de um jogo figurado com a linguagem S (da Bell Laboratories, antiga AT&T). O código fonte do R está disponível sob a licença GNU, GPL e as versões binárias pré-compiladas são fornecidas para Windows, Macintosh, e muitos sistemas operacionais Unix/Linux. O ambiente é altamente expansível com o uso de pacotes, pacotes estes que são bibliotecas de dados e funções para diferentes áreas relacionadas com a estatística e áreas afins. Um conjunto básico de pacotes vem embutido na instalação do R, com muito outros disponíveis na rede de distribuição do R (em inglês CRAN).

Caso tenha mais interesse sobre esta linguagem, recomendamos a consulta do link seguinte: https://pt.wikipedia.org/wiki/R_(linguagem_de_programa%C3%A7%C3%A3o).

ANÁLISE DESCRITIVA

NHHD (Necessidades de Aconselhamento psicoemocional em horas-homem / dia):

- Variável quantitativa na escala métrica;

ATD (Número médio de acidentes traumatizantes / dia):

- Variável quantitativa na escala métrica;

DASU (Duração médio no atendimento no serviço de urgência):

- Variável quantitativa na escala métrica;

DAPE (Duração média do primeiro atendimento psicoemocional):

- Variável quantitativa na escala métrica;

NRX (Número de meios de diagnóstico complementares (raio-X) pedidos / dia):

- Variável quantitativa na escala métrica;

MÊS :

- Variável quantitativa na escala ordinal;

- Na figura1 encontram-se resumidos os valores mínimos e máximos, os valores das médias e medianas como também dos 1º e 3º quartis. Pela observação da mesma e tendo em consideração as variáveis NHHD e DAPE, como exemplo, pudemos afirmar que existe uma discrepância entre o número médio, de horas, aconselhado e o número médio, em minutos, da duração da primeira consulta psicoemocional.

```
> summary(Tabela_TRABALHO_ATDM$`NHHD(h)`, digits = 4)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.500   1.800   2.400   2.552   3.200   4.100
> summary(Tabela_TRABALHO_ATDM$`ATD`, digits = 4)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  5.200   5.700   5.800   5.838   6.100   6.400
> summary(Tabela_TRABALHO_ATDM$`DASU (min.)`, digits = 4)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  12.00   18.00   22.00   25.38   30.00   54.00
> summary(Tabela_TRABALHO_ATDM$`DAPE(min)`, digits = 4)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  46.00   50.00   64.00   66.29   75.00  117.00
> summary(Tabela_TRABALHO_ATDM$NRX, digits = 4)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  18.00   35.00   52.00   52.19   64.00   85.00
```

Figura1: Output sumários

- Decidimos calcular a correlação entre as variáveis NHHD e DAPE como meio para compreender se o numero de horas aconselhadas estava relacionado com a duração média do primeiro atendimento psicoemocional;
- Com base na figura2 obtemos o valor da correlação de Pearson (0.8741906), o que implica a existência de uma correlação forte. Escolhemos Pearson com base na natureza das variáveis.

```
> cor(Tabela_TRABALHO_ATDM$`NHHD(h)`, Tabela_TRABALHO_ATDM$`DAPE(min)`)
[1] 0.8741906
```

Figura2:Correlação de Pearson entre NHHD e DAPE

- Fizemos também a correlação entre as variáveis DASU e NRX com fim a compreender se a duração das urgências estava relacionado com o número de diagnósticos complementares pedidos;
- Com base na Figura3, comprovamos que existe uma correlação muito fraca (0,0208295). Mais uma vez o coeficiente escolhido (Pearson) teve como base a natureza das variáveis em estudo.

```
> cor(Tabela_TRABALHO_ATDM$`DASU (min.)`, Tabela_TRABALHO_ATDM$NRX)
[1] 0.0208295
```

Figura3:Correlação de Pearson entre DASU e NRX

- Com base na Figura4, obtivemos um boxplot para a variável DAPE onde é possível visualizar um outlier com o valor 117, que se reflete num tempo bastante superior ao tempo médio das consultas. É visível também a assimetria negativa do gráfico.

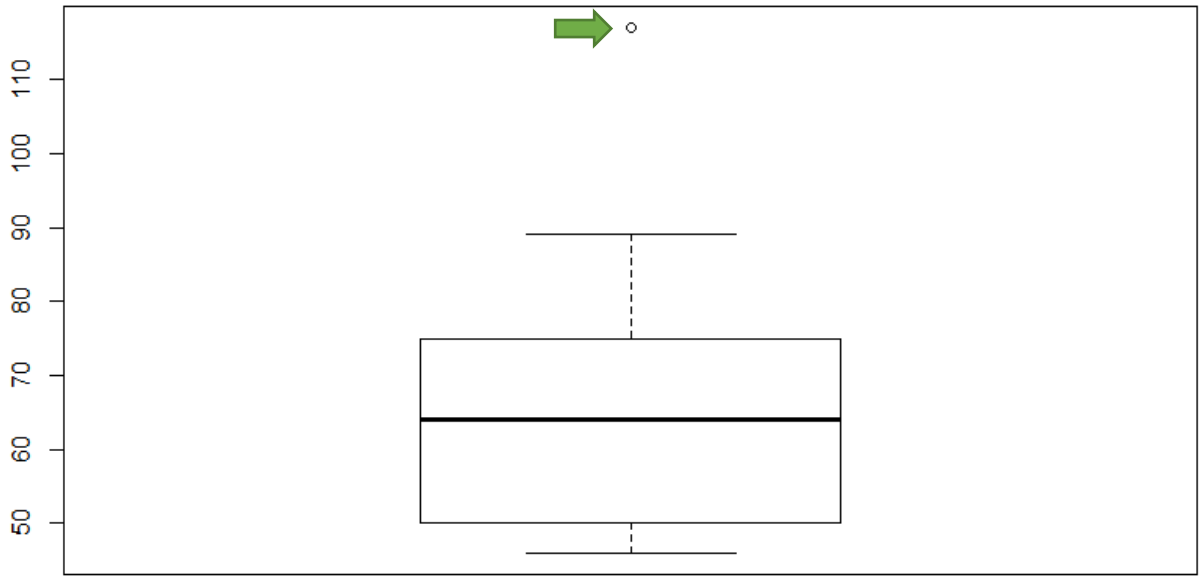

Figura4: boxplot

REGRESSÃO LINEAR

- Começámos por criar uma função para estimar o modelo de regressão linear múltipla (figura5).

```
> lm1 <- lm(Tabela_TRABALHO_ATDM$`NHHD(h)` ~ Tabela_TRABALHO_ATDM$ATD + Tabela_TRABALHO_ATDM$`DASU (min.)`+
 Tabela_TRABALHO_ATDM$`DAPE(min)` + Tabela_TRABALHO_ATDM$NRX + Tabela_TRABALHO_ATDM$MÊS, data= Tabela_TRABAL
HO_ATDM)
```

Figura5: criação da função

- Depois obtemos um sumário da função criada (figura6):

```
Call:
lm(formula = Tabela_TRABALHO_ATDM$`NHHD(h)` ~ Tabela_TRABALHO_ATDM$ATD +
    Tabela_TRABALHO_ATDM$`DASU (min.)` + Tabela_TRABALHO_ATDM$`DAPE(min)` +
    Tabela_TRABALHO_ATDM$NRX + Tabela_TRABALHO_ATDM$MÊS, data = Tabela_TRABALHO_ATDM)

Residuals:
     Min       1Q   Median       3Q      Max
-0.32904 -0.10365  0.00315  0.10086  0.27057

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                  -2.297933   0.785754  -2.924   0.0105 *
Tabela_TRABALHO_ATDM$ATD      0.206232   0.153380   1.345   0.1987
Tabela_TRABALHO_ATDM$`DASU (min.)` -0.128198   0.021897  -5.854 3.17e-05 ***
Tabela_TRABALHO_ATDM$`DAPE(min)`    0.101606   0.014063   7.225 2.95e-06 ***
Tabela_TRABALHO_ATDM$NRX     -0.003495   0.002469  -1.416   0.1773
Tabela_TRABALHO_ATDM$MÊS      0.058369   0.023347   2.500   0.0245 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2002 on 15 degrees of freedom
Multiple R-squared:  0.9494,    Adjusted R-squared:  0.9326
F-statistic: 56.33 on 5 and 15 DF,  p-value: 3.452e-09
```

Figura6: sumário da função

- Vizualizando este summary podemos observar ,por exemplo, que o número de necessidades de Aconselhamento psicoemocional explicada pelo modelo é de 0,9494.
- O aumento de um número de necessidades conduz a um aumento de 0,206 de número médio de acidentes traumatizantes.
- Neste quadro obtivemos um valor de estatística F (56,33), referente à ANOVA, que tem associado um p-value muito baixo ($3.452*10^{-9}$) que é irrelevante fazendo com que o modelo seja altamente significativo.
- Observando o valor de multiple r-squared (maior que 0,9) podemos aceitá-lo como um indicador de bom ajustamento.
- Também se encontram os valores estimados dos coeficientes, valores estes que nos permitiram construir a reta:
    - NHHD = -2.298 + 0.206ATD - 0.128DASU + 0.102DAPE – 0.0035NRX + 0.058MES (valores arredondados da figura6)

- Depois, procedemos à análise de resíduos por forma a validar os pressupostos do modelo, obtendo o seguinte gráfico (figura7):

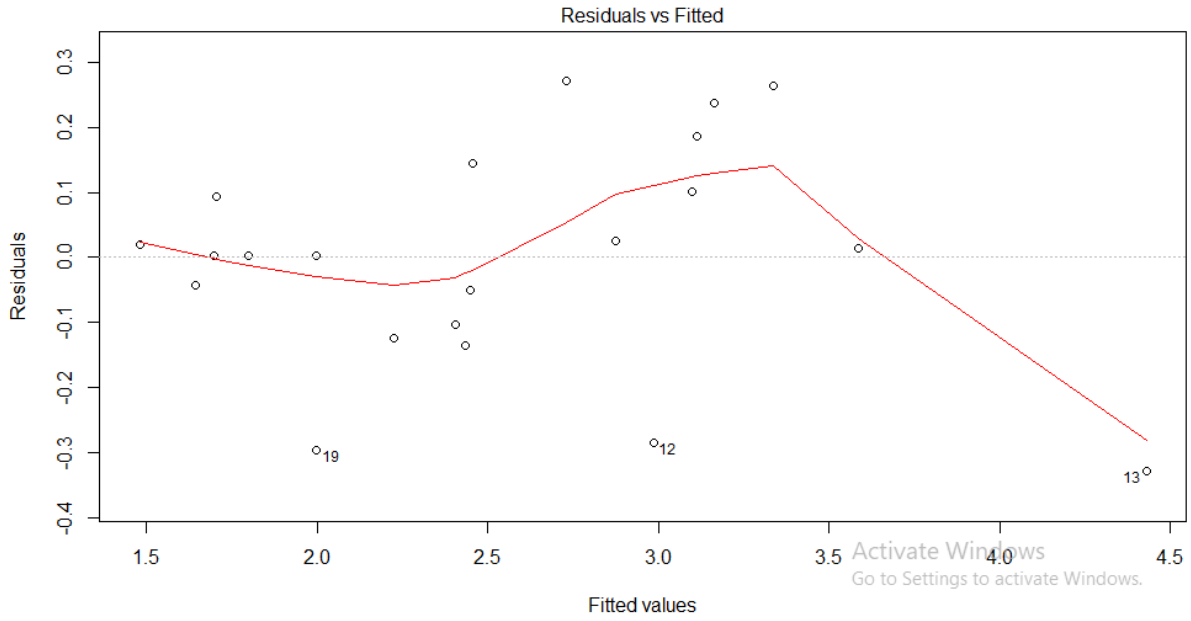

Figura7: Análise de resíduos

- De seguida para determinar o modelo mais simples , usámos o método Backward, Forward e Stepwise;
- Enquanto o método Forward começa sem nenhuma variável no modelo e adiciona variáveis a cada passo, o método Backward faz o caminho oposto, ou seja, incorpora inicialmente todas as variáveis e depois, por etapas, cada uma pode ou não ser eliminada.
- Stepwise é uma modificação do método Forward onde cada passo para todas as variáveis do modelo são previamente verificadas pelas suas estatísticas parciais;
    - Stepwise:

```
Start:  AIC=-62.62
Tabela_TRABALHO_ATDM$`NHHD(h)` ~ Tabela_TRABALHO_ATDM$ATD + Tabela_TRABALHO_ATDM$`DASU (min.)` +
    Tabela_TRABALHO_ATDM$`DAPE(min)` + Tabela_TRABALHO_ATDM$NRX +
    Tabela_TRABALHO_ATDM$MÊS

                                Df Sum of Sq     RSS     AIC
<none>                                       0.60135 -62.615
- Tabela_TRABALHO_ATDM$ATD       1   0.07248 0.67383 -62.225
- Tabela_TRABALHO_ATDM$NRX       1   0.08035 0.68170 -61.982
- Tabela_TRABALHO_ATDM$MÊS       1   0.25057 0.85192 -57.301
- Tabela_TRABALHO_ATDM$`DASU (min.)`  1   1.37406 1.97541 -39.639
- Tabela_TRABALHO_ATDM$`DAPE(min)`    1   2.09291 2.69426 -33.121

Call:
lm(formula = Tabela_TRABALHO_ATDM$`NHHD(h)` ~ Tabela_TRABALHO_ATDM$ATD +
    Tabela_TRABALHO_ATDM$`DASU (min.)` + Tabela_TRABALHO_ATDM$`DAPE(min)` +
    Tabela_TRABALHO_ATDM$NRX + Tabela_TRABALHO_ATDM$MÊS, data = Tabela_TRABALHO_ATDM)

Coefficients:
                    (Intercept)           Tabela_TRABALHO_ATDM$ATD
                      -2.297933                           0.206232
Tabela_TRABALHO_ATDM$`DASU (min.)`    Tabela_TRABALHO_ATDM$`DAPE(min)`
                      -0.128198                           0.101606
          Tabela_TRABALHO_ATDM$NRX           Tabela_TRABALHO_ATDM$MÊS
                      -0.003495                           0.058369
```

Figura8: Método Stepwise

- Backward:

```
Start:  AIC=-62.62
Tabela_TRABALHO_ATDM$`NHHD(h)` ~ Tabela_TRABALHO_ATDM$ATD + Tabela_TRABALHO_ATDM$`DASU (min.)` +
    Tabela_TRABALHO_ATDM$`DAPE(min)` + Tabela_TRABALHO_ATDM$NRX +
    Tabela_TRABALHO_ATDM$MÊS

                                Df Sum of Sq     RSS     AIC
<none>                                       0.60135 -62.615
- Tabela_TRABALHO_ATDM$ATD       1   0.07248 0.67383 -62.225
- Tabela_TRABALHO_ATDM$NRX       1   0.08035 0.68170 -61.982
- Tabela_TRABALHO_ATDM$MÊS       1   0.25057 0.85192 -57.301
- Tabela_TRABALHO_ATDM$`DASU (min.)`  1   1.37406 1.97541 -39.639
- Tabela_TRABALHO_ATDM$`DAPE(min)`    1   2.09291 2.69426 -33.121

Call:
lm(formula = Tabela_TRABALHO_ATDM$`NHHD(h)` ~ Tabela_TRABALHO_ATDM$ATD +
    Tabela_TRABALHO_ATDM$`DASU (min.)` + Tabela_TRABALHO_ATDM$`DAPE(min)` +
    Tabela_TRABALHO_ATDM$NRX + Tabela_TRABALHO_ATDM$MÊS, data = Tabela_TRABALHO_ATDM)

Coefficients:
                      (Intercept)            Tabela_TRABALHO_ATDM$ATD
                        -2.297933                            0.206232
Tabela_TRABALHO_ATDM$`DASU (min.)`    Tabela_TRABALHO_ATDM$`DAPE(min)`
                        -0.128198                            0.101606
          Tabela_TRABALHO_ATDM$NRX            Tabela_TRABALHO_ATDM$MÊS
                        -0.003495                            0.058369
```

Figura9: Método de Backward

- Forward:

```
Start:  AIC=-62.62
Tabela_TRABALHO_ATDM$`NHHD(h)` ~ Tabela_TRABALHO_ATDM$ATD + Tabela_TRABALHO_ATDM$`DASU (min.)` +
    Tabela_TRABALHO_ATDM$`DAPE(min)` + Tabela_TRABALHO_ATDM$NRX +
    Tabela_TRABALHO_ATDM$MÊS


Call:
lm(formula = Tabela_TRABALHO_ATDM$`NHHD(h)` ~ Tabela_TRABALHO_ATDM$ATD +
    Tabela_TRABALHO_ATDM$`DASU (min.)` + Tabela_TRABALHO_ATDM$`DAPE(min)` +
    Tabela_TRABALHO_ATDM$NRX + Tabela_TRABALHO_ATDM$MÊS, data = Tabela_TRABALHO_ATDM)

Coefficients:
                      (Intercept)            Tabela_TRABALHO_ATDM$ATD
                        -2.297933                            0.206232
Tabela_TRABALHO_ATDM$`DASU (min.)`    Tabela_TRABALHO_ATDM$`DAPE(min)`
                        -0.128198                            0.101606
          Tabela_TRABALHO_ATDM$NRX            Tabela_TRABALHO_ATDM$MÊS
                        -0.003495                            0.058369
```

Figura10: Método Forward

CONCLUSÃO

O relatório incidiu no trabalho solicitado em aula, ou seja, obter um modelo de regressão linear múltipla com base numa equação, avaliar a sua significância e analisar o poder explicativo do mesmo.

Após chegar ao objetivo principal do trabalho com sucesso foi-nos permitido obter as seguintes conclusões:

O valor de R-squared (0,9494) era próximo de 1 o que nos levou a concluir que não se justificava a eliminação de alguma das variáveis para melhorar a precisão da reta.

A equação obtida é igual à do exemplo, presente no enunciado, que nos foi fornecido pela docente.

Devido ao grande número de variáveis é recomendado usar métodos de seleção automática para eliminar aquelas com efeitos insignificantes. Como neste trabalho tínhamos 5 variáveis, justifica-se a utilização deste métodos. Como os métodos Backward e Stepwise são bastante parecidos e o método Forward indica que devemos manter as variáveis.

Por fim, este projeto tinha algum nível de dificuldade que ultrapassámos em grupo e juntando o facto do R ser um ambiente intuitivo com uma linguagem simples, quando comparada às outras linguagens por nós já estudadas, permitiu-nos alcançar todos os objetivos requeridos pela docente. Embora o tempo de contacto com R não tenha sido tanto quanto gostaríamos foi enriquecedor para o grupo e reconhecemos a importância que tem na nossa área e pode vir a ter no nosso futuro.

BIBLIOGRAFIA

Freedman et al. 2004, Lyles & Kupper 1997, Chen & Wang 2004, 14-01-2020

Análise com SPSS, João Marôco

https://rpubs.com/pmedeiros/ex1rlm , 14-01-2020

https://www.portalaction.com.br/analise-de-regressao/2723-selecao-stepwise, 21-01-2020

https://www.portalaction.com.br/analise-de-regressao/2722-selecao-backward, 21-01-2020