Atividade - ANÁLISE MULTIVARIADA

Charles Barros

18/10/2021

Case 1

> Dados de uma pesquisa de satisfação do paciente em um hospital são mostrados na seguinte tabela: case1.csv. As variáveis do regressor (colunas de dados) são: a idade do paciente, um índice de gravidade da doença (valores maiores indicam maior gravidade), um indicador variável denotando se o paciente é um paciente médico (0) ou um paciente cirúrgico (1), e um índice de ansiedade (valores maiores indicam maior ansiedade), além do indicador de satisfação.

Base de dados

```
db <- read.csv(file = "CS1.csv",header = TRUE,sep = ";");head(db)
```

```
##   observacao idade gravidade segmento ansiedade satisfacao
## 1          1    55        50        0       2.1         68
## 2          2    46        24        1       2.8         77
## 3          3    30        46        1       3.3         96
## 4          4    35        48        1       4.5         80
## 5          5    59        58        0       2.0         43
## 6          6    61        60        0       5.1         44
```

Questoes

(A)

> Ajustar um modelo de regressão linear múltipla, sendo a variável resposta a satisfação, usando idade, gravidade da doença e o índice de ansiedade como os possíveis regressores da sua equação.

(B)

> Estimativa e os erros padrão dos coeficientes de regressão.Comente os resultados.

(C)

Todos os parâmetros do modelo são estimados com a mesma precisão? Por que sim ou por que não?

Resposta

(A)

```
mdl <- lm(satisfacao ~ idade+gravidade+ansiedade,data = db)
```

$Y = \beta_0 + \beta_1 x + \beta_2 x + \epsilon$

Como solicitado foi elaborado um modelo de regressão linear múltipla representado pela fórmula acima, onde temos a variavel dependente representado por *satisfação* é as variaveis indepedentes representado por *Idade,gravidade e ansiedade*.Inlustrando de uma forma mais informal teriamos a equação no seguinte formato : Satisfação = constante + idade * x + gravidade * x + ansiedade * x + erro.

(B)

```
summary(mdl)
```

```
##
## Call:
## lm(formula = satisfacao ~ idade + gravidade + ansiedade, data = db)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18.2812  -3.8635   0.6427   4.5324  11.8734
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 143.8952     5.8975  24.399  < 2e-16 ***
## idade        -1.1135     0.1326  -8.398 3.75e-08 ***
## gravidade    -0.5849     0.1320  -4.430 0.000232 ***
## ansiedade     1.2962     1.0560   1.227 0.233231
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.037 on 21 degrees of freedom
## Multiple R-squared:  0.9035, Adjusted R-squared:  0.8897
## F-statistic: 65.55 on 3 and 21 DF,  p-value: 7.85e-11
```

Ao analisarmos os estimadores obtemos as seguintes equações, Satisfação = 143.89 - 1.1135 * (idade) -0.5849 * (gravidade)+1.2962 * (ansiedade). Sendo assim a cada alteração de valores nas variáveis (*idade e gravidade*) teremos um decréscimo. Enquanto que a variável (*ansiedade*) apresentara um acréscimo. Vale ressaltar que o Intercept será impactado positivamente ou negativamente dependendo do respectivo sinal do estimador de cada variável.

Já o erro padrão é uma medida de variação de uma média amostral em relação à média da população. Sendo assim, é uma medida que ajuda a verificar a confiabilidade da média amostral calculada. O erro padrão representa a distância média em que os valores observados caem da linha de regressão. Convenientemente, ele informa como o modelo de regressão está errado usando as unidades da variável de resposta. Valores menores são melhores porque indicam que as observações estão mais próximas da linha ajustada.

(c)

Vale destacar que o nível de significância utilizado em nosso modelo foi de 5%, sendo assim notamos que apenas as variáveis idade e gravidade são significativas para nosso modelo. Enquanto a variável ansiedade não apresenta significância para nosso modelo pois ao observamos o p-valor da variável, visualizamos o valor de 0.233231(ou aproximadamente 23%), sendo assim o p-valor dessa variável está acima de 5%.

Como explicado anteriormente o fato de a variável ansiedade não ser significativa seria ideal elaborar um novo modelo removendo essa variável, trazendo para uma linguagem mais informal deveríamos basicamente peneirar nosso modelo até que se consiga encontrar um modelo com uma maior precisão, vale ressaltar que antes de remover qualquer variável precisamos verificar o conceito ou interpretação que aquela variável tem em nosso banco de dados.

Quando observado o Coeficiente de determinação (ou Multiple R-squared), apresenta um valor de 0,9035 ou seja isso significa que 90,35% dos dados são explicados através da equação (ou modelo).

Já o valor do R-ajustado (ou ajusted R-squared) foi de 0.8897 (ou 88.97%).