Grupo EAD655 - Métodos Estatísticos

Douglas Cardoso Eduardo Salis Débora Kono Denise Del Col
Henry Hideki Joanne Araujo

6/11/2021

Exercícios

1. Considere o arquivo do EXCEL denominado TRABALHO_04062021 referente a uma amostra de 100 empresas clientes de uma grande empresa HATCO que é fornecedora no setor industrial (HATCO – Hair, Anderson and Tatham Company).

Estas empresas clientes avaliaram atributos do atendimento da HATCO em uma escala de 0 a 10. Variáveis da pesquisa:

- X_1 = satisfação com a rapidez na entrega do produto
- X_2 = satisfação com o nível de preço
- X_3 = satisfação com a flexibilidade de preço
- X_4 = satisfação com a imagem do fornecedor
- X_5 = satisfação com o serviço como um todo
- X_6 = satisfação com a imagem da força de vendas
- X_7 = satisfação com a qualidade do produto

As demais variáveis do banco de dados têm os seguintes nomes e códigos:

- X_8 = porte das empresas clientes: 0 = pequeno; 1 = grande
- X_9 = nível de uso: percentual mensal de compras de produtos da HATCO
- X_{10} = nível geral de satisfação: notas de 0 a 10
- X_{11} = estrutura do setor de compras da empresa cliente: 0 descentralizada, 1 = centralizada
- X_{12} = tipo de indústria: 0 = tipo 0, 1 = tipo 1

Classificar as variáveis X_1 até X_{12}, segundo o tipo de escala: nominal, ordinal, razão;

X_1 = Razão

X_2 = Razão

X_3 = Razão

X_4 = Razão

X_5 = Razão

X_6 = Razão

X_7 = Razão

X_8 = Ordinal

X_9 = Razão

X_{10} = Razão

X_{11} = Nominal

X_{12} = Nominal

2. *Considerando X1 até X12, selecione um par de variáveis, no qual seria razoável supor uma relação de causa e efeito; dizer qual das duas variáveis selecionadas será Y (efeito) e qual será X (causa);*

É razoável supor uma relação de causa e efeito com a variável X_{10} e X_9, o nível geral de satisfação (causa) influencia o percentual mensal de compras de produtos da HATCO (efeito), assim, X_{10} = X e X_9 = Y.

3. *Usando o par de variáveis selecionadas, obter a expressão da reta de regressão de Y sobre X. Como a amostra é grande (n = 100), os cálculos das somatórias poderão ser feitos no EXCEL; porém, os resultados de cada somatória deverão ser transportados para o Word.*

```
dados <- readxl::read_excel('dados.xlsx')

# selecionando apenas as features X9 e X_10
dados <- dados[c('X9', 'X10')]

# mudando o nome da coluna
colnames(dados) <- c('Y', 'X')

# criando as colunas do cálculo
dados <- dados %>%
  mutate(
    'X^2' = map2_dbl(.x = X, .y = X, ~ .x * .y),
    'Y^2' = map2_dbl(.x = Y, .y = Y, ~ .x * .y),
    'XY' = map2_dbl(.x = X, .y = Y, ~ .x * .y)
  )

dados %>%
  head() %>%
  knitr::kable()
```

Y	X	X^2	Y^2	XY
32	4.2	17.64	1024	134.4
43	4.3	18.49	1849	184.9
48	5.2	27.04	2304	249.6
32	3.9	15.21	1024	124.8
58	6.8	46.24	3364	394.4
45	4.4	19.36	2025	198.0

```
# somando as colunas
soma <- dados %>% map_dbl(sum)
print(soma)
```

```
##         Y         X       X^2       Y^2        XY
##   4610.00    477.10   2348.71 220520.00  22535.40
```

```
# médias
X_mean <- mean(dados$X)
Y_mean <- mean(dados$Y)

# std
X_std <- sd(dados$X)
```

```r
# n length
n = nrow(dados)

# prints
cat('Média de X:', X_mean)
```

```
## Média de X: 4.771
```

```r
cat('Média de Y:', Y_mean)
```

```
## Média de Y: 46.1
```

```r
cat('Desvio padrão de X:', X_std)
```

```
## Desvio padrão de X: 0.8555576
```

A fórmula para calcular a reta de regressão é

$$Y = a + bX$$

Sabe-se que

$$b = \frac{\sum XY - n\bar{X}\bar{Y}}{(n-1)(Sx)^2}$$
$$a = \bar{Y} - b\bar{X}$$

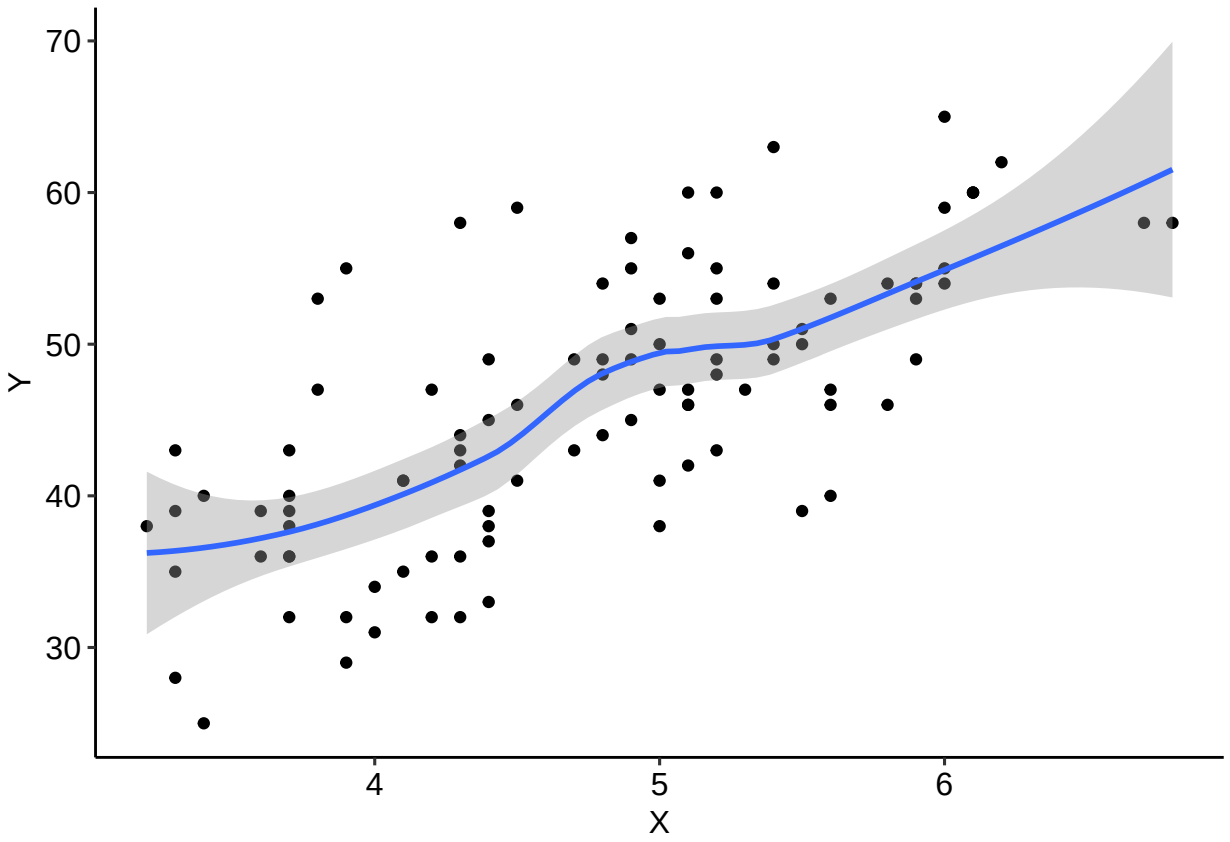
Portanto,

$$b = \frac{22535,4 - 100 \times 4,771 \times 46,1}{(100-1)(0,8555576)^2}$$
$$b = 7.466822$$
$$a = 46,1 - b \times 4,771$$
$$a = 10.47579$$
$$Y = 10.47579 + 7.466822X$$

Resolvendo pelas ferramentas do R

```r
# grafico entre os dois
ggplot(dados, aes(x = X, y = Y)) +
  geom_point() +
  stat_smooth()
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```




```
# correlação
cor(dados$X, dados$Y)
```

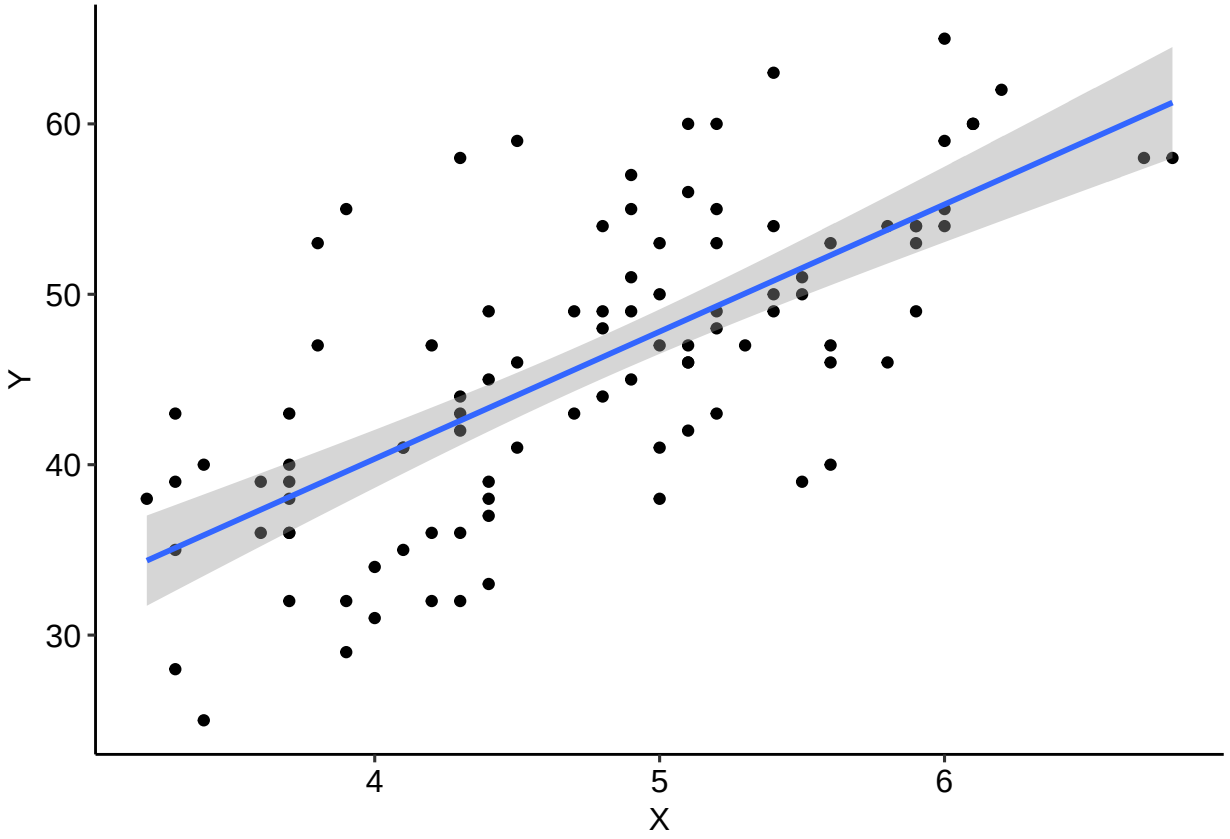
```
## [1] 0.7106975
```

```
# modelo linear
model <- lm(Y ~ X, data = dados)
model
```

```
##
## Call:
## lm(formula = Y ~ X, data = dados)
##
## Coefficients:
## (Intercept)            X
##      10.476        7.467
```

```
# regression line
ggplot(dados, aes(X, Y)) +
  geom_point() +  stat_smooth(method = lm)
```

```
## `geom_smooth()` using formula 'y ~ x'
```




```
# model summary
summary(model)
```

```
##
## Call:
## lm(formula = Y ~ X, data = dados)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.5433  -4.5699  -0.4167   3.8970  15.4169
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  10.4758     3.6184   2.895  0.00467 **
## X             7.4668     0.7466  10.001  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.356 on 98 degrees of freedom
## Multiple R-squared:  0.5051, Adjusted R-squared:    0.5
## F-statistic:   100 on 1 and 98 DF,  p-value: < 2.2e-16
```

4. *Com base no par de variáveis selecionado, calcular e interpretar o desvio (ou resíduo) dos três últimos casos da amostra.*

```
residuals(model)[c(n, n-1, n-2)]
```

```
##         100          99          98
## -10.3298089  -0.5831267  -0.1030333
```

Todos os três desvios são negativos, o que dá para intuir que o modelo superestimou a nossa variável *target* (Y). Pode-se ver que na última linha de nossos dados ($n = 100$) o desvio negativo é bem maior que os outros dois (n_{99} E n_{98}). Analisando os desvios médianos, podemos ter a certeza que, de fato, na mediana, nosso modelo está superestimando o percentual mensal de compras de produtos da HATCO.

II

1. *Calcule os coeficientes de correlação de Spearman e de Pearson para as variáveis: volume de chuvas versus produtividade média por hectare de manga e para as variáveis volume de chuvas versus produtividade média por hectare de abacate.*

```
frutas <- tibble(
  Chuvas = c(42,18,25,18,35,10,25,28,17,18),
  Manga = c(154,76,100,90,154,70,90,65,84,65),
  Abacate = c(146,96,132,105,146,86,140,132,100,94)
)

frutas %>%
  knitr::kable(align = "ccc")
```

Chuvas	Manga	Abacate
42	154	146
18	76	96
25	100	132
18	90	105
35	154	146
10	70	86
25	90	140
28	65	132
17	84	100
18	65	94

(i) Spearman

$$r_s = 1 - \frac{6 \sum D^2}{n(n^2 - 1)}$$

(ii) Pearson

Na fórmula abaixo,

- x e y são dois vetores de tamanho n
- m_x e m_y correspodem as médias de x e y, respectivamente

$$r = \frac{\sum (x - m_x)(y - m_y)}{\sqrt{\sum (x - m_x)^2 \sum (y - m_y)^2}}$$

O p-value da correlação pode ser determinado de dois modos:

- 1. usando a tabela de coeficientes de correlação entre para graus de liberdade: $d_f = n - 2$, onde n é o número de observação de x e variáveis y;
- 2. ou calculando o ***t-value*** como:

$$t = \frac{r}{\sqrt{1 - r^2}}\sqrt{n - 2}$$

Para esse caso (2) o *p-value* correspondente é determinado utilizando a tabela de distribuição t. **Se o *p-value* for menor que 0.5, então a correlação entre x e y é relevante.**

Para volume de chuvas versus produtividade média por hectare de manga

```
cor_spearman_mg = cor(x = frutas$Chuvas, y = frutas$Manga, method =  "spearman")
cat('Spearman:', cor_spearman_mg)
```

```
## Spearman: 0.5621226
```

```
cor_pearson_mg = cor(frutas$Chuvas, frutas$Manga, method = 'pearson')
cat('Pearson:', cor_pearson_mg)
```

```
## Pearson: 0.822712
```

Para as variáveis volume de chuvas versus produtividade média por hectare de abacate

```
cor_spearman_ab = cor(x = frutas$Chuvas, y = frutas$Abacate, method =  "spearman")
cat('Spearman:', cor_spearman_ab)
```

```
## Spearman: 0.9071599
```

```
cor_pearson_ab = cor(frutas$Chuvas, frutas$Abacate, method = 'pearson')
cat('Pearson:', cor_pearson_ab)
```

```
## Pearson: 0.9054528
```

2. *Interprete os resultados*

Quando o coeficiente de correlação é o de Spearman, o primeiro par de variáveis (produtividade de manga em função do volume de chuvas) apresenta correlação positiva moderada por estar entre 0,41 e 0,7 ($r_s = 0,573$); já o segundo par de variáveis (produtividade de abacate em função do volume de chuvas) apresenta coeficiente de Spearman ($r_s = 0,909$) com alta correlação positiva por estar acima de 0,7.

Quando tratamos do coeficiente de Pearson, ambos os pares de variáveis apresentam alta correlação positiva pelos resultados estarem acima de 0,7 (o primeiro par com $r = 0,823$ e o segundo com $r = 0,905$).

3. *Qual par de variáveis apresenta a maior correlação de Spearman?*

O par "Volume de chuvas versus produtividade média por hectare de abacate" apresenta maior correlação de Spearman, com o valor 0,909.

4. *Qual par de variáveis apresenta a maior correlação de Pearson?*

O par "Volume de chuvas versus produtividade média por hectare de abacate" também apresenta maior correlação de Pearson, com o valor 0,905.

III

Uma amostra de tamanho 80 produziu a correlação amostral r = 0,54. Fazer os 4 testes de hipótese, a seguir e responder ao item 5.

Teoria

The correlation coefficient, r, tells us about the strength and direction of the linear relationship between x and y. However, the reliability of the linear model also depends on how many observed data points are in the sample. We need to look at both the value of the correlation coefficient r and the sample size n, together. We perform a hypothesis test of the "significance of the correlation coefficient" to decide whether the linear relationship in the sample data is strong enough to use to model the relationship in the population.

The sample data are used to compute r, the correlation coefficient for the sample. If we had data for the entire population, we could find the population correlation coefficient. But because we have only sample data, we cannot calculate the population correlation coefficient. The sample correlation coefficient, r, is our estimate of the unknown population correlation coefficient.

If the p-value is less than the significance level ($\alpha = 0.05$):

- **Decision**: Reject the null hypothesis.
- **Conclusion**: "There is sufficient evidence to conclude that there is a significant linear relationship between x and y because the correlation coefficient is significantly different from zero."

If the p-value is NOT less than the significance level ($\alpha = 0.05$):

- **Decision**: DO NOT REJECT the null hypothesis.
- **Conclusion**: "There is insufficient evidence to conclude that there is a significant linear relationship between x and y because the correlation coefficient is NOT significantly different from zero."

Exercício

Sabe-se que o teste de hipótese de correlação estatística é feito por

$$Z = 1,1513 \times \log\left(\frac{1+r}{1-r}\right)$$

$$Z \sim N(\mu_Z,\ Var_Z)$$

$$\mu_Z = 1,1513 \times \log\left(\frac{1+\ valor\ em\ H0}{1-\ valor\ em\ H0}\right)$$

$$Var_Z = \frac{1}{n-3}$$

$$Z_{obs} = \frac{Z - \mu_Z}{\sigma_Z}$$

Na fórmula acima,

- r é a correlação obtida na amostra
- n é o tamanho da amostra
- Z tem distribuição normal com média μ_Z e variância Var_Z

$$n = 80$$

$$r = 0,54$$

$$Z = 1,1513 \times \log\left(\frac{1+0,54}{1-0,54}\right) = 0,6041595$$

$$\mu_Z = 1,1513 \times \left(\frac{1+0,41}{1-0,41}\right) = 0,435614043$$

$$Var_Z = \frac{1}{80-3} = 0,01298701$$

$$\sigma_Z = \sqrt{Var_Z} = 0,11396058$$

$$Z_{obs} = 1,48$$

Resolvendo no R

```
Z_obs = function(r, n, h0){

  Z = 1.1513 * log((1 + r)/(1 - r), base = 10)

  uz = 1.1513 * log((1 + h0) / (1 - h0), base = 10)

  sigma_z = sqrt(1 / (n - 3))

  (Z - uz) / sigma_z
}
```

1. **H0 : p = 0,41 contra H1 : p > 0,41, ao nível de 0,10 (unicaudal à direita)**

Z_{obs} = 1,48

$Z_{crítico}$ = 1,28

H0 falsa, pois Z_{obs} é maior que $Z_{crítico}$, ou seja, está na área crítica.

2. **H0 : p = 0,41 contra H1 : p < 0,41, ao nível de 0,10 (unicaudal à esquerda)**

Z_{obs} = 1,48

$Z_{crítico}$ = - 1,28

H0 verdadeira, pois Z_{obs} é maior que $Z_{crítico}$, ou seja, está fora da área crítica.

3. **H0 : p = 0,41 contra H1 : p $\neq$ 0,41, ao nível de 0,10 (bicaudal)**

Z_{obs} = 1,48

$Z_{crítico}$ = -1,64 e 1,64

H0 verdadeira, pois Z_{obs} está entre os valores de $Z_{crítico}$, ou seja, está fora das áreas críticas.

4. **H0 : p = 0,41 contra H1 : p < 0,41, ao nível de 0,05 (unicaudal à esquerda)**

Z_{obs} = 1,48

$Z_{crítico}$ = -1,64

H0 verdadeira, pois Z_{obs} é maior que $Z_{crítico}$, ou seja, está fora da área crítica.

5. *De acordo com os resultados apresentados nos 4 itens anteriores, escolher a alternativa correta*

(a) A hipótese de que a correlação populacional é 0,41 é verdadeira ao nível de 0,10, teste unicaudal à direita.

(b) A hipótese de que a correlação populacional é 0,41 é falsa ao nível de 0,10, teste unicaudal à esquerda.

(c) A hipótese de que a correlação populacional é 0,41 é falsa ao nível de 0,10, teste bicaudal.

(d) **A hipótese de que a correlação populacional é 0,41 é verdadeira ao nível de 0,05, teste unicaudal à esquerda.**

(e) Nenhuma das alternativas anteriores.

Reposta correta: D

IV

Para os dados referentes ao tamanho em m2 de filiais de uma rede de lojas de departamento e às suas vendas foi aplicado o modelo de regressão linear, considerando-se nível de significância de 0,05, com os seguintes resultados:

Estatística de regressão	
R-Quadrado	0.73

	Coeficientes	Erro padrão	Stat t	p-value
Interseção	56210,02	2490,913	22,56603	4,96E-07
Tamanho	22,14144	3,099297	7,144018	0,000379

1. *Interprete* R^2

O R^2 de 0,73 significa que o modelo linear está relativamente bem ajustado aos dados da amostra, isto é, está adequado e satisfatório, visto que quanto maior o R^2 mais explicativo está o modelo, pois este informa que percentual da variabilidade na variável "vendas"(Y) é considerado na regressão sobre a variável "tamanho em m2 de filiais"(X). "Relativamente bem" pois explica cerca de 73% dos dados, que não é pouco, mas possui espaço para fine tune, melhorar o modelo dado que está perdendo 27% da explicação.

2. *Quais as hipóteses e qual a decisão do teste de significância para o coeficiente linear?*

H0: coeficiente linear = 0 (é igual de zero)

H1: coeficiente linear $\neq$ 0 (é diferente de zero)

Considerando α = 0,05; pelo segundo método apresentado na aula sabemos que: Se 4,96E-07 $<$ 0,05, temos que H0 é falsa, ou seja, o coeficiente linear é diferente de zero e deve ser considerado na reta de regressão.

3. *Quais as hipóteses e qual a decisão do teste de significância para o coeficiente angular?*

H0: coeficiente angular = 0 (é igual de zero)

H1: coeficiente angular $\neq$ 0 (é diferente de zero)

Considerando α = 0,05; pelo segundo método apresentado na aula sabemos que: Sendo 0,000379 $<$ 0,05, temos que H0 é falsa, ou seja, o coeficiente angular é diferente de zero e deve ser considerado na reta de regressão.

4. *De acordo com os resultados apresentados nos 3 itens anteriores, escolher a alternativa correta*

(a) A variável tamanho é relevante no modelo de regressão, sendo H0 verdadeira.

(b) **O coeficiente linear é relevante no modelo de regressão, sendo H0 falsa.**

(c) O modelo consegue explicar 73% da variância do tamanho das filiais da rede de lojas.

(d) Não valeu a pena incluir a variável vendas no modelo de regressão.

(e) Nenhuma das alternativas anteriores.

Resposta: Alternativa B.