Universidade Federal da Bahia
Programa de Pós-Graduação em Ciência de Dados e Big Data
Carlos Magno Santos Ribeiro de Brito

Prof Dr. Paulo Canas Rodrigues

Respostas da terceira lista de exercício

Questão 01) Pretende-se, se possível, modelar através de uma reta de regressão linear simples a quantidade de vidro Y produzido num ponto de reciclagem (Kg), usando como variável independente x o número de dias sem despejar o mesmo. Para tal, registaram-se os seguintes dados.

x_i	2	3	4	5	10	15	20	25
Y_i	100	150	-	320	650	810	1040	1480

O valor de Y para $x = 4$ foi perdido, mas antes foram obtidos os seguintes resultados com base nos dados originais:

$$\sum_{i=1}^{8} x_i = 84 \quad \sum_{i=1}^{8} Y_i = 4800 \quad \sum_{i=1}^{8} x_i^2 = 1404 \quad \sum_{i=1}^{8} Y_i^2 = 4548000 \quad \sum_{i=1}^{8} x_i Y_i = 79700$$

a) Escreva a reta de regressão estimada através do método dos mínimos quadrados.

Precisaremos encontrar os valores de $\bar{x}$ $\overline{Y}$ $\widehat{\beta_1}$ e $\widehat{\beta_0}$ para assim formar a reta de regressão $Y = \widehat{\beta_1}x + \widehat{\beta_0}$:

$$\bar{x} = \frac{1}{8}\sum_{i=1}^{8} x_i = \frac{84}{8} = 10{,}5$$

$$\overline{Y} = \frac{1}{8}\sum_{i=1}^{8} Y_i = \frac{4800}{8} = 600$$

$$\widehat{\beta_1} = \frac{\sum_{i=1}^{8} x_i Y_i - n\bar{x}\overline{Y}}{\sum_{i=1}^{8} x_i^2 - n\bar{x}^2} = \frac{79700 - 8 \times 10{,}5 \times 600}{1404 - 8 \times 10{,}5^2} \approx 56{,}1303$$

$$\widehat{\beta_0} = \overline{Y} - \widehat{\beta_1}\bar{x} = 600 - 56{,}1303 \approx 10{,}6322$$

$$Y = 56{,}1303x + 10{,}6322$$

b) Acha que conseguiu um bom ajuste? Use o coeficiente de determinação.

Tem-se que o coeficiente de determinação é R^2 é dado como:

$$R^2 = \frac{\left(\widehat{\beta_1}^2\right) \sum_{i=1}^{8} x_i^2 - n\bar{x}^2}{\sum_{i=1}^{8} Y_i^2 - n\bar{Y}^2} = \frac{56{,}1303^2 \times (1404 - 8 \times 10{,}5^2)}{4548000 - 8 \times 600^2} \approx 0{,}986$$

Dado o valor de R^2 obtido muito próximo de 1, pode-se afirmar que o ajuste foi bem sucedido.

c) Qual o valor da quantidade de vidro produzida no ponto de reciclagem que prevê ocorrer em 28 dias sem o despejar?

Não é possível extrapolar valores fora do intervalo usado para o ajuste. Assim sendo, para 28 dias, não se pode prever o valor de Y produzido.

d) Teste se o declive da reta de regressão obtida em (a) é zero, usando um nível de significância de 10%. Como interpreta a não rejeição dessa hipótese?

Temos as seguintes hipóteses:

- $H_0 : \beta_1 = 0$
- $H_a : \beta_1 \neq 0$

Cálculo da variância do erro:

$$\widehat{\sigma}^2 = \frac{1}{n-2}\left\{\left(\sum_{i=1}^{8} Y_i^2 - nY^2\right) - \left(\widehat{\beta_1}\right)^2\left(\sum_{i=1}^{8} x_i^2 - n\bar{x}^2\right)\right\}$$
$$\widehat{\sigma}^2 = \frac{1}{6}\{(4548000 - 8 \times 600^2) - 56{,}1303^2 \times (1404 - 8 \times 10{,}5^2)\} \approx 3896{,}8797$$

Cálculo da estatística de teste T:

$$T = \frac{\widehat{\beta_1} - \beta_1}{\sqrt{\frac{\widehat{\sigma}^2}{\sum_{i=1}^{8} x_i^2 - n\bar{x}^2}}} \sim t_{n-2}$$

$$T = \frac{56{,}1303 - 0}{\sqrt{\frac{3896{,}8797}{1404 - 8 \times 10{,}5^2}}} \sim t_6 \approx 20{,}5356$$

Para a região crítica com $gl = 6$ e 10% de significância, temos o valor de $t_0 = 1{,}943$ (tabelado ou no python). Assim sendo, a região de rejeição será:

$$RR =]-\infty; -1{,}943[U]1{,}943; +\infty[$$

Então, como o valor de T obtido é 20,5356, está dentro da região de rejeição, devemos rejeitar a hipótese nula H_0. Isso quer dizer que o valor de β_1 é considerável na inclinação e não pode ser desconsiderado.

e) Qual o erro de previsão quando o ponto de reciclagem não é despejado durante 10 dias?

Para o caso, basta achar o valor previsto pelo ajuste e subtrair pelo valor real medido:

- Estimado $\rightarrow Y_{10}^e = 56{,}1303x_{10} + 10{,}6322 = 56{,}1303 \times 10 + 10{,}6322 = 571{,}9352$

- Tabelado $\rightarrow Y_{10}^t = 650$

- Diferença $\rightarrow \|\Delta Y_{10}\| = 571{,}9352 - 650 = 78{,}0648$

Questão 02) Considere o conjunto de dados "Wage" do pacote "ISLR2" do software R. Considere a variável "health_ins" como variável resposta e as variáveis "age", "maritl", "race", "education", "jobclass", "health" e "logwage" como variáveis explicativas. Ajuste uma regressão logística, escreva o modelo final e interprete os coeficientes obtidos.

Questão 03) Numa empresa existem três máquinas para produzir um certo tipo de peça. Foram retiradas amostras aleatórias de dimensão cinco de cada uma das máquinas e foi medido o diâmetro (em mm) de cada uma das peças, resultando nos resultados abaixo:

- Máquina 1: $49, 55, 51, 52, 48$;
- Máquina 2: $53, 54, 58, 49, 56$;
- Máquina 3: $55, 51, 52, 52, 52, 50$;

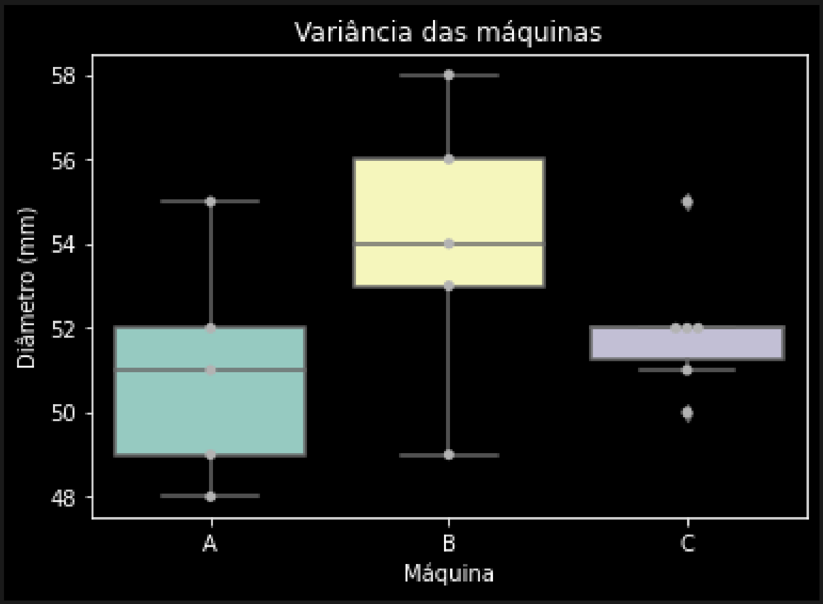
Verifique se existem diferenças entre os diâmetros das peças produzidas por cada uma das máquinas. No caso de haver diferenças, quais os pares de máquinas responsáveis por essas diferenças?

Para o caso em questão, tomemos como hipóteses:

- H_0 : Máquina 1, 2 e 3 têm média de diâmetro iguais;

- H_a : Pelo menos uma das máquinas possui a média de diâmetro diferente;

Temos o seguinte boxplot para esses dados:




```
import pandas as pd
import statsmodels.api as sm
from statsmodels.formula.api import ols
import matplotlib.pyplot as plt
import seaborn as sns
plt.style.use('dark_background')

# dados
data = {"A": {0: 49, 1: 55, 2: 51, 3: 52, 4: 48},
        "B": {0: 53, 1: 54, 2: 58, 3: 49, 4: 56},
        "C": {0: 55, 1: 51, 2: 52, 3: 52, 4: 52, 5: 50}
        }
df = pd.DataFrame(data)

df_melt = pd.melt(df.reset_index(), id_vars=["index"],
value_vars=["A", "B", "C"])

df_melt.columns = ["index", "treatments", "value"]

ax = sns.boxplot(data=df)
ax = sns.swarmplot(data=df, color=".7")
ax.title.set_text("Variancia das maquinas")
ax.set_xlabel("Maquina")
```

```
ax.set_ylabel("Diametro (mm)")

plt.figure()

#Obter tabela com valores da anova
model = ols("value ~ treatments", data=df_melt).fit()
anova_table = sm.stats.anova_lm(model, typ=2)
anova_table
```

Obtém-se a seguinte tabela com os valores da análise de variância:

	sum_sq	df	F	PR(>F)
treatments	23.4375	2.0	1.692708	0.222156
residual	90.0	13.0	NaN	NaN

A partir da tabela com dados gerados, pode-se concluir então que o valor de p (ou $PR(> F)$) obtido a partir da análise ANOVA é, signinicamente maior que 0.05. Não se rejeita a hipótese H_0 de que as médias dos três grupos de dados são iguais.