Lista 1

Regressão

Caio Gomes Alves

1 Questão 1

1.1 Pergunta

Sejam $\hat{\epsilon}_i = \left(Y_i - \hat{Y}_i\right)$. Mostre que:

- **(a)** $\sum_{i=1}^{n} \hat{\epsilon}_i = 0$.
- **(b)** $\sum_{i=1}^{n} \hat{\epsilon}_i x_i = 0$.

1.2 Resposta

1.2.1 a)

$$\begin{aligned}
\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right) &= \sum_{i=1}^{n}\left(Y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_i\right)\right) \\
&= \sum_{i=1}^{n}(Y_i) - \sum_{i=1}^{n}\left(\hat{\beta}_0 + \hat{\beta}_1 x_i\right) \\
&= n\bar{Y} - n\hat{\beta}_0 - \hat{\beta}_1 \sum_{i=1}^{n} x_i \\
&= n\bar{Y} - n\hat{\beta}_0 - n\hat{\beta}_1 \bar{x} \\
&= n\left(\bar{Y} - \hat{\beta}_1 \bar{x} - \hat{\beta}_0\right) \\
&= n\left(\hat{\beta}_0 - \hat{\beta}_0\right) = 0
\end{aligned}$$

1.2.2 b)

Temos que:

$$\begin{aligned}
\sum_{i=1}^{n}\left(Y_i-\hat{Y}_i\right)x_i &= \sum_{i=1}^{n}\left(Y_i-\left(\hat{\beta}_0+\hat{\beta}_1 x_i\right)\right)x_i \\
&= \sum_{i=1}^{n}\left(Y_i-\bar{Y}+\hat{\beta}_1\bar{x}-\hat{\beta}_1 x_i\right)x_i \\
&= \sum_{i=1}^{n}\left[\left(Y_i-\bar{Y}\right)x_i+\hat{\beta}_1 x_i\left(\bar{x}-x_i\right)\right] \\
&= \sum_{i=1}^{n}\left[\left(Y_i-\bar{Y}\right)x_i\right]+\hat{\beta}_1\sum_{i=1}^{n}\left[\left(\bar{x}-x_i\right)x_i\right] \qquad (1)
\end{aligned}$$

Demonstremos agora as partes que compõe a equação (1). Primeiramente:

$$\begin{aligned}
\sum_{i=1}^{n}\left[\left(Y_i-\bar{Y}\right)x_i\right] &= \sum_{i=1}^{n}\left(Y_i x_i-\bar{Y}x_i\right) \\
&= \sum_{i=1}^{n}\left(Y_i x_i\right)-n\bar{Y}\bar{x} \\
&= \sum_{i=1}^{n}\left[\left(Y_i-\bar{Y}\right)\left(x_i-\bar{x}\right)\right] \qquad (2)
\end{aligned}$$

E a segunda parte:

$$\begin{aligned}
\hat{\beta}_1\sum_{i=1}^{n}\left[\left(\bar{x}-x_i\right)x_i\right] &= \frac{\sum_{i=1}^{n}\left(Y_i-\bar{Y}\right)\left(x_i-\bar{x}\right)}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sum_{i=1}^{n}\left[\left(\bar{x}-x_i\right)x_i\right] \\
&= \frac{\sum_{i=1}^{n}\left(Y_i-\bar{Y}\right)\left(x_i-\bar{x}\right)}{\sum_{i=1}^{n}(x_i-\bar{x})x_i-\bar{x}\sum_{i=1}^{n}(x_i-\bar{x})}\sum_{i=1}^{n}\left[\left(\bar{x}-x_i\right)x_i\right] \\
&= \frac{\sum_{i=1}^{n}\left(Y_i-\bar{Y}\right)\left(x_i-\bar{x}\right)}{-\sum_{i=1}^{n}(\bar{x}-x_i)x_i}\sum_{i=1}^{n}\left[\left(\bar{x}-x_i\right)x_i\right] \\
&= -\sum_{i=1}^{n}\left(Y_i-\bar{Y}\right)\left(x_i-\bar{x}\right) \qquad (3)
\end{aligned}$$

Agora, juntando (2) e (3) em (1), temos:

$$\begin{aligned}
\sum_{i=1}^{n}\left(Y_i-\hat{Y}\right)x_i &= \sum_{i=1}^{n}\left[\left(Y_i-\bar{Y}\right)\left(x_i-\bar{x}\right)\right]-\sum_{i=1}^{n}\left[\left(Y_i-\bar{Y}\right)\left(x_i-\bar{x}\right)\right] \\
&= 0
\end{aligned}$$

2 Questão 2

2.1 Pergunta

Mostre que:

$$\sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2 = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2 + \sum_{i=1}^{n} \left(\hat{Y}_i - \bar{Y}\right)^2$$

2.2 Resposta

Vejamos que:

$$\begin{aligned}
\sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2 &= \sum_{i=1}^{n} \left[\left(Y_i - \hat{Y}_i\right) + \left(\hat{Y}_i - \bar{Y}\right)\right]^2 \\
&= \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2 + \sum_{i=1}^{n} \left(\hat{Y}_i - \bar{Y}\right)^2 + 2\sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)\left(\hat{Y}_i - \bar{Y}\right)
\end{aligned}$$

Portanto, precisamos mostrar que $\sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)\left(\hat{Y}_i - \bar{Y}\right) = 0$:

$$\begin{aligned}
\sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)\left(\hat{Y}_i - \bar{Y}\right) &= \sum_{i=1}^{n} \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)\left(\hat{\beta}_0 + \hat{\beta}_1 x_i - \bar{Y}\right) \\
&= \sum_{i=1}^{n} \left(Y_i - \bar{Y} + \hat{\beta}_1 \bar{x} - \hat{\beta}_1 x_i\right)\left(\bar{Y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 x_i - \bar{Y}\right) \\
&= \sum_{i=1}^{n} \left(\left(Y_i - \bar{Y}\right) + \hat{\beta}_1(\bar{x} - x_i)\right)\hat{\beta}_1(x_i - \bar{x}) \\
&= \sum_{i=1}^{n} \left(\left(Y_i - \bar{Y}\right)\hat{\beta}_1(x_i - \bar{x})\right) + \sum_{i=1}^{n} \left(\hat{\beta}_1(\bar{x} - x_i)\hat{\beta}_1(x_i - \bar{x})\right) \\
&= \hat{\beta}_1 \sum_{i=1}^{n} \left(\left(Y_i - \bar{Y}\right)(x_i - \bar{x})\right) - \hat{\beta}_1^2 \sum_{i=1}^{n} (x_i - \bar{x})^2 \\
&= \hat{\beta}_1 \hat{\beta}_1 \sum_{i=1}^{n} (x_i - \bar{x})^2 - \hat{\beta}_1^2 \sum_{i=1}^{n} (x_i - \bar{x})^2 \\
&= 0
\end{aligned}$$

3 Questão 3

3.1 Pergunta

Encontre $\mathbb{E}(\hat{\beta}_0)$.

3.2 Resposta

$$\begin{aligned}
\mathbb{E}\left(\hat{\beta}_0\right) &= \mathbb{E}\left(\bar{Y} - \hat{\beta}_1 \bar{x}\right) \\
&= \mathbb{E}\left(\bar{Y}\right) - \bar{x}\mathbb{E}\left(\hat{\beta}_1\right) \\
&= \frac{1}{n}\mathbb{E}\left(\sum_{i=1}^{n} y_i\right) - \bar{x}\beta_1 \\
&= \frac{1}{n}\sum_{i=1}^{n}(\beta_0 + \beta_1 x_i) - \bar{x}\beta_1 \\
&= \beta_0 + \beta_1 \bar{x} - \bar{x}\beta_1 \\
&= \beta_0
\end{aligned}$$

4 Questão 4

4.1 Pergunta

Encontre $\mathrm{Var}(\hat{\beta}_1)$.

4.2 Resposta

$$\begin{aligned}
\mathrm{Var}\left(\hat{\beta}_1\right) &= \mathrm{Var}\left[\frac{1}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\left(\sum_{i=1}^{n} Y_i(x_i - \bar{x}) - \sum_{i=1}^{n} \bar{Y}(x_i - \bar{x})\right)\right] \\
&= \frac{1}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\mathrm{Var}\left(\sum_{i=1}^{n} Y_i(x_i - \bar{x}) - \sum_{i=1}^{n} \bar{Y}(x_i - \bar{x})\right) \\
&= \frac{1}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\mathrm{Var}\left(\sum_{i=1}^{n} Y_i(x_i - \bar{x})\right) \\
&= \frac{1}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sum_{i=1}^{n}(x_i - \bar{x})\mathrm{Var}\left(Y_i\right) \\
&= \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})}
\end{aligned}$$

5 Questão 5

5.1 Pergunta

Mostre que a reta de regressão obtida pelo método de mínimos quadrados passa pelo ponto $(\bar{x}, \bar{Y})$.

5.2 Resposta

Como a reta de regressão é definida para todo $x \in \mathbb{R}$, e como sabemos que $\bar{x} \in \mathbb{R}$, o valor de $\bar{x}$ pertencerá à reta de regresão.

Para encontrar o $\hat{Y}$ associado a $\bar{x}$, substituemos na equação da reta estimada:

$$\begin{aligned}
\hat{Y} &= \hat{\beta}_0 + \hat{\beta}_1 \bar{x} \\
\hat{Y} &= \bar{Y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 \bar{x} \\
\hat{Y} &= \bar{Y}
\end{aligned}$$

Assim, para o valor de $x = \bar{x}$, temos que o $\hat{Y}$ é igual a $\bar{Y}$. Assim, o ponto $(\bar{x}, \bar{Y})$ pertence à reta de regressão.

6 Questão 6

6.1 Pergunta

Mostre que os estimadores de máxima verossimilhança e mínimos quadrados para β_0 e β_1 são equivalentes.

6.2 Resposta

Os estimadores de mínimos quadrados para β_0 e β_1 são dados por meio da minimização do quadrado do erro estimado, ou seja:

$$\left(\hat{\beta}_0, \hat{\beta}_1\right) = \underset{\beta_0, \beta_1}{\operatorname{argmin}} \sum_{i=1}^{n} \left(Y_1 - \beta_0 - \beta_1 x_i\right)^2$$

Para a estimação por máxima verossimilhança, temos que encontrar o valor que maximiza o valor da verossimilhança, dada por:

$$\begin{aligned}
\mathcal{L}(\beta_0, \beta 1 | Y_i) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(Y_i - \beta_0 - \beta_1)^2}{2\sigma^2}\right\} \\
&= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left\{-\frac{\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1)^2}{2\sigma^2}\right\}
\end{aligned}$$

É fácil ver que o termo $\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n$ não depende de β_0 nem de β_1, por isso no processo de maximização ele será apenas uma constante. Usualmente trabalha-se com o logaritmo natural da verossimilhança, que não altera o valor que maximiza a função, e que trata as multiplicações para que se tornem somas, como segue:

$$l(\mathcal{L}(\beta_0, \beta_1, \sigma^2 | Y_i)) = n \log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \frac{\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1)^2}{2\sigma^2} \tag{4}$$

Novamente, o denominador $2\sigma^2$ não depende de β_0 nem de β_1, então para encontrar o valor que maximiza a log-verossimilhança, derivamos com relação à esses parâmetros, e encontramos que o estimador é dado por:

$$\left(\hat{\beta}_0, \hat{\beta}_1\right) = \underset{\beta_0, \beta_1}{\operatorname{argmax}} \left(-\sum_{i=1}^{n} \left(Y_1 - \beta_0 - \beta_1 x_i\right)^2\right) = \underset{\beta_0, \beta_1}{\operatorname{argmin}} \sum_{i=1}^{n} \left(Y_1 - \beta_0 - \beta_1 x_i\right)^2$$

7 Questão 7

7.1 Pergunta

Encontre o estimador de máxima verossimilhança para σ^2.

7.2 Resposta

Vejamos que, em (4), o termo $\frac{1}{\sqrt{2\pi}}$ não altera o ponto de maximização para σ^2, de modo que podemos reescrever a log-verossimilhança como:

$$l(\mathcal{L}(\beta_0, \beta_1, \sigma^2 | Y_i)) = -\frac{n}{2}\log(\sigma^2) - \frac{\sum_{i=1}^n (Y_i - \beta_0 - \beta_1)^2}{2\sigma^2}$$

Agora, podemos derivar com relação a σ^2 para encontrar o ponto de máximo:

$$\frac{\partial}{\partial\sigma^2} l(\mathcal{L}(\beta_0, \beta_1, \sigma^2 | Y_i)) = -\frac{n}{2\sigma^2} + \frac{\sum_{i=1}^n (Y_i - \beta_0 - \beta_1)^2}{2\left(\sigma^2\right)^2}$$

E igualando a 0, temos:

$$\begin{aligned}
\frac{-n\hat{\sigma}^2 + \sum_{i=1}^n (Y_i - \beta_0 - \beta_1)^2}{2\left(\hat{\sigma}^2\right)^2} &= 0 \\
n\hat{\sigma}^2 &= \sum_{i=1}^n (Y_i - \beta_0 - \beta_1)^2 \\
\hat{\sigma}^2 &= \frac{\sum_{i=1}^n (Y_i - \beta_0 - \beta_1)^2}{n} \\
\hat{\sigma}^2 &= \frac{\sum_{i=1}^n (Y_i - \hat{Y})^2}{n}
\end{aligned}$$

8 Questão 8

8.1 Pergunta

Considere uma nova observação (x_{n+1}, Y_{n+1}), com $x_{n+1} = \bar{x}$. Sejam $\hat{\beta}_0^{n+1}$ e $\hat{\beta}_1^{n+1}$ os estimadores de mínimos quadrados obtidos utilizando todas as $n+1$ amostras. Mostre que

$$\hat{\beta}_1^{n+1} = \hat{\beta}_1$$

Interprete esse resultado e esboce um gráfico que represente essa propriedade.

8.2 Resposta

Sabemos que o estimador de mínimos quadrados para β_1 é dado por:

$$
\begin{aligned}
\hat{\beta}_1 &= \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\
&= \frac{(y_1 - \bar{y})(x_1 - \bar{x}) + (y_2 - \bar{y})(x_2 - \bar{x}) + \ldots + (y_n - \bar{y})(x_n - \bar{x})}{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2}
\end{aligned}
$$

Assim, o estimador para β_1 incluindo a observação $n+1$ é dada por:

$$
\begin{aligned}
\hat{\beta}_1^{n+1} &= \frac{\sum_{i=1}^{n+1}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\
&= \frac{(y_1 - \bar{y})(x_1 - \bar{x}) + (y_2 - \bar{y})(x_2 - \bar{x}) + \ldots + (y_n - \bar{y})(x_n - \bar{x}) + (y_{n+1} - \bar{y})(\bar{x} - \bar{x})}{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2 + (\bar{x} - \bar{x})^2} \\
&= \frac{(y_1 - \bar{y})(x_1 - \bar{x}) + (y_2 - \bar{y})(x_2 - \bar{x}) + \ldots + (y_n - \bar{y})(x_n - \bar{x}) + 0}{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2 + 0^2} \\
\hat{\beta}_1^{n+1} &= \hat{\beta}_1
\end{aligned}
$$

Assim, podemos ver que o estimador para o β_1 não será alterado, caso a nova observação tenha valor $x_{n+1} = \bar{x}$. Para visualizar esse efeito, consideremos o seguinte exemplo simples:

```r
# Seed para reprodutibilidade:
set.seed(42)

# Considere os valores de x vindos de uma normal N(10, 2):
df <- data.frame(
    x <- rnorm(10, 10, 10)
)

# Os valores de y são gerados a partir de 3*x + N(0, 1):
df$y <- (3 * x + rnorm(10, 0, 1.5))

# Valores de Beta_0 e Beta_1 gerados a partir de mínimos quadrados:
(lm(y ~ x, data = df))
```

```
##
## Call:
## lm(formula = y ~ x, data = df)
##
## Coefficients:
## (Intercept)            x
##       1.453        2.890
```

```r
# Valor da média de x:
(x_barra <- mean(df$x))
```

```
## [1] 15.47297
```

```r
# Considere agora uma nova observação, no ponto (x_barra, 17):
df2 <- data.frame(
    x = c(df$x, x_barra),
    y = c(df$y, 12)
)

# Verifique que a estimativa de Beta_0 foi alterada,
# enquanto que a de Beta_1 se manteve a mesma:
(lm(y ~ x, data = df2))
```

```
##
## Call:
## lm(formula = y ~ x, data = df2)
##
## Coefficients:
## (Intercept)            x
##      -1.653        2.890
```

```r
# Podemos observar graficamente a mudança na estimativa de
# Beta_0, com os graficos a seguir:
```