1

Etapa 2 - Projeto 3

Eduardo Gebara
Daniel Ruhman

a) $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \cdot x$

$y_i = \hat{y}_i + \varepsilon_i = \hat{\beta}_0 + \hat{\beta}_1 \cdot x_i + \varepsilon_i \quad , \quad i = 1, 2, \ldots, n$

$\varepsilon_i = y_i - \hat{y}_i$

· Soma dos Erros Quadrados:

$$SEQ = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 + \hat{\beta}_1 x_i)^2$$

derivar para minimizar a soma dos erros quadrados:

$$(1) \ \frac{\partial SEQ}{\partial \beta_0} = 0 \quad e \quad (2) \ \frac{\partial SEQ}{\partial \beta_1} = 0$$

$$(1) \ \left.\frac{\partial SEQ}{\partial \beta_0}\right|_{\hat{\beta}_0, \hat{\beta}_1} = -2 \cdot \sum_{i=1}^{n} (y_i - \hat{\beta}_0 + \hat{\beta}_1 x_i) = 0$$

$$\left.\frac{\partial SEQ}{\partial \beta_0}\right|_{\hat{\beta}_0, \hat{\beta}_1} = -2 \left[\sum_{i=1}^{n} y_i - n \cdot \hat{\beta}_0 - \hat{\beta}_1 \cdot \sum_{i=1}^{n} x_i \right] = 0$$

$$\sum_{i=1}^{n} y_i - n \cdot \hat{\beta}_0 - \hat{\beta}_1 \cdot \sum_{i=1}^{n} x_i = 0$$

$$\hat{\beta}_0 = \frac{1}{n} \sum_{i=1}^{n} y_i - \hat{\beta}_1 \cdot \frac{1}{n} \cdot \sum_{i=1}^{n} x_i$$

(2) $\boxed{\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \cdot \bar{x}}$, tal que $\bar{x} = \dfrac{\sum_{i=1}^{n} x_i}{n}$ e $\bar{y} = \dfrac{\sum_{i=1}^{n} y_i}{n}$

(2)

$$\left.\frac{\partial SEQ}{\partial \beta_1}\right|_{\hat{\beta}_0, \hat{\beta}_1} = -2 \cdot \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) x_i = 0$$

$$\left.\frac{\partial SEQ}{\partial \beta_1}\right|_{\hat{\beta}_0, \hat{\beta}_1} = -2 \cdot \left[\sum_{i=1}^{n} x_i y_i - \hat{\beta}_0 \sum_{i=1}^{n} x_i - \hat{\beta}_1 \sum_{i=1}^{n} (x_i)^2 \right] = 0$$

$$\sum_{i=1}^{n} x_i y_i - (\bar{y} - \hat{\beta}_1 \cdot \bar{x}) \cdot \sum_{i=1}^{n} x_i - \hat{\beta}_1 \sum_{i=1}^{n} (x_i)^2 = 0$$

$$\sum_{i=1}^{n} x_i y_i + \hat{\beta}_1 (\bar{x})^2 \cdot n - y \bar{x} \cdot n - \hat{\beta}_1 \sum_{i=1}^{n} (x_i)^2 = 0$$

$$\sum_{i=1}^{n} x_i y_i - \bar{y} \bar{x} \cdot n = \hat{\beta}_1 \cdot \sum_{i=1}^{n} (x_i)^2 - \hat{\beta}_1 \cdot (\bar{x})^2 \cdot n$$

$$\sum_{i=1}^{n} x_i y_i - \bar{y} \cdot \bar{x} \cdot n = \hat{\beta}_1 \left(\sum_{i=1}^{n} (x_i)^2 - (\bar{x})^2 \cdot n \right)$$

$$\boxed{\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - \bar{y} \bar{x} \cdot n}{\sum_{i=1}^{n} (x_i)^2 - (\bar{x})^2 \cdot n}}$$

③

b) A suposição feita é o de que podemos aproximar um conjunto de dados, normalmente representados no eixo cartesiano por um "scatter plot", por uma reta (no caso da regressão linear), usando o método dos mínimos quadrados.

O erro pode ser representado através da soma dos erros quadrados:

$$SEQ = \sum_{i=1}^{n} \underbrace{(y_i - \hat{y}_i)}_{\text{distância entre ponto e valor correspondente na reta gerada}}^2 = \sum_{i=1}^{n} (\varepsilon_i)^2$$

, onde $\hat{y}_i$ é o valor estimado de y_i

e a estimativa da variância σ^2 se dá por:

$$\hat{\sigma}^2 = \frac{SEQ}{n-2}$$

, tal que n é o número de pares de pontos (x_i, y_i)

É importante também incluir o coeficiente de determinação (R^2), que é uma medida de ajustamento do modelo linear em relação aos valores observados.

$$R^2 = 1 - \frac{SEQ}{SQT}$$

, tal que SQT é a soma dos quadrados totais, dado por $SQT = S_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2$.

4

Vale ressaltar que $0 \leq R^2 \leq 1$ e esse valor indica, em porcentagem, o quanto o modelo consegue explicar os valores observados.

c) Teste de Hipóteses

(nula) $H_0 : \beta_1 = \beta_{1,0}$

(alternativa) $H_1 : \beta_1 \neq \beta_{1,0}$, em que $\beta_{1,0}$ é uma constante.

Um caso especial desse teste de hipóteses é quando $\beta_{1,0} = 0$:

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

Não rejeitar H_0 significa dizer que não há relação linear entre as variáveis X e Y. Se H_0 for rejeitada, quer dizer que X explica a variabilidade de Y e portanto há correlação entre as variáveis estudadas.

d) Sim, é possível. A mudança ocorreria na quantidade de variáveis regressoras que são as dependentes de uma variável resposta, e de suas respectivas derivadas parciais. A equação $y_i = \beta_0 + \beta_1 \cdot x_i$ da regressão linear simples precisaria ser modificada para incluir as demais variáveis, o que impacta todo o processo matemático do item a).

5

A equação para m variáveis fica como a seguir:

$$y_i = \beta_0 + \beta_1 \cdot x_{i1} + \beta_2 \cdot x_{i2} + \dots + \beta_m x_{im} + \epsilon_i$$

Em relação a suposição do modelo, quanto maior o número de variáveis, mais perto chegamos de um modelo que se ajusta melhor aos dados.

Por fim, em relação ao teste de hipóteses, quanto maior o número de variáveis, maior será o erro acumulado nas variáveis dependentes $(x_{i1}, x_{i2}, \dots, x_{im})$ e maior a chance de Ho ser rejeitada:

$$H_0 : \beta_1 = \beta_2 = \dots = \beta_m = 0$$

$$H_1 : \text{pelo menos um } \beta = 0$$