Obtenção dos estimadores

Obtenção dos estimadores de β_0 e β_1 a partir do Método dos Mínimos Quadrados, cujo objetivo é encontrar a reta que passa mais próxima ao mesmo tempo de todos os pontos. Neste caso, encontraremos os estimadores $\hat{\beta_0}$ e $\hat{\beta_1}$ que minimizam a soma dos erros ao quadrado.

Temos que o valor estimado de Y dados os x_i observados é dado pela relação:

$$E(Y|x_i) = \beta_0 + \beta_1 x_i$$

Modelamos cada valor y_i

Temos que

E o erro é: $e_i = y_i - \hat{y}_i$

$$y_i = \beta_0 + \beta_1 x_i + e_i$$

$$S(\hat{\beta_0}, \hat{\beta_1}) = \sum_{i=1}^{n} \left(y_i - \hat{\beta_0} - \hat{\beta_1} x_i \right)^2$$

Aplicando a regra da cadeia para a derivada parcial sobre os eixos:

$$\frac{\partial S}{\partial \beta_0} = \frac{\partial S}{\partial \left[\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) \right]} \cdot \frac{\partial \left[(y_i - \beta_0 - \beta_1 x_i) \right]}{\partial \beta_0}$$

Temos que:

$$\frac{\partial S}{\partial \left[\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) \right]} = 2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)$$

e

$$\frac{\partial \left[\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) \right]}{\partial \beta_0} = -1$$

Impondo a condição de valor mínimo do parabolóide para $\hat{\beta_0}$, vamos procurar estimar β_0 e β_1 de maneira a minimizar o erro:

$$\frac{\partial S}{\partial \hat{\beta_0}} = 2 \sum_{i=1}^{n} \left(y_i - \hat{\beta_0} - \hat{\beta_1} x_i \right) (-1) = 0$$

$$\frac{\partial S}{\partial \hat{\beta}_0} = -2\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right) = 0$$

E para $\hat{\beta}_1$:

$$\frac{\partial S}{\partial \hat{\beta}_1} = \frac{\partial S}{\partial\left[\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)\right]} \cdot \frac{\partial\left[\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)\right]}{\partial \hat{\beta}_1}$$

$$\frac{\partial S}{\partial \hat{\beta}_1} = -2\sum_{i=1}^{n} x_i\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right) = 0$$

Dividindo por $2n$ e distribuindo

$$\frac{-2\sum_{i=1}^{n} y_i}{2n} + \frac{2\sum_{i=1}^{n}\hat{\beta}_0}{2n} + \frac{2\sum_{i=1}^{n}\hat{\beta}_1 x_i}{2n} = \frac{0}{2n}$$

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

$$-\bar{y} + \hat{\beta}_0 + \hat{\beta}_1\bar{x} = 0$$

Chegamos à expressão para $\hat{\beta}_0$:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$$

Na expressão acima $\bar{y}$ é a Média Amostral de y e $\bar{x}$ é a Média Amostral de x

Podemos substituir $\hat{\beta}_0$ na relação original:

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

$$\sum_{i=1}^{n}\left[x_i\left(y_i - \bar{y}\right) + x_i\hat{\beta}_1\left(\bar{x} - x_i\right)\right] = 0$$

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

Que nos dá, isolando $\beta 1$:

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

Reescrevendo levando em conta as relações

$$\sum_{i=1}^{n}(x_i - \overline{x})^2 = \sum_{i=0}^{n}(x_i^2 - x_i\overline{x})$$

e

$$\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = \sum_{i=0}^{n}(x_i y_i - y_i\overline{x})$$

desenvolvidas mais abaixo, temos:

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

Relações auxiliares

Temos que

$$\frac{1}{n}\sum_{i=1}^{n} x_i = \overline{x}$$

portanto

$$\sum_{i=1}^{n} x_i = n\overline{x}$$

e temos que

$$\frac{1}{n}\sum_{i=1}^{n} y_i = \overline{y}$$

portanto

$$\sum_{i=1}^{n} y_i = n\overline{y}$$

Relação 1

A igualdade abaixo é importante para entendermos a fórmula dos β :

$$\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = \sum_{i=0}^{n}(x_i y_i - y_i \overline{x})$$

Vamos estudar como reescrever a relação:

$$\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$$

Aplicando a propriedade distributiva:

$$\sum_{i=1}^{n}(x_i y_i - x_i \overline{y} - \overline{x} y_i + \overline{xy}) =$$

$$\sum_{i=1}^{n} x_i y_i - \overline{y} \sum_{i=1}^{n} x_i - \overline{x} \sum_{i=1}^{n} y_i + \overline{xy} \sum_{i=1}^{n} 1 =$$

$$\sum_{i=1}^{n} x_i y_i - \overline{y} n \overline{x} - \overline{x} n \overline{y} + \overline{xy} n =$$

$$\sum_{i=1}^{n} x_i y_i - n \overline{xy}$$

A relação acima pode ser escrita como:

$$\sum_{i=1}^{n} x_i y_i - n \overline{xy} = \sum_{i=1}^{n} x_i y_i - \overline{x} \sum_{i=0}^{n} y_i =$$

$$\sum_{i=0}^{n}(x_i y_i - y_i \overline{x})$$

Relação 2

Outra relação importante é a seguinte :

$$\sum_{i=1}^{n}(x_i - \overline{x})^2 = \sum_{i=0}^{n}(x_i^2 - x_i\overline{x})$$

Vamos estudar como reescrever a relação:

$$\sum_{i=1}^{n}(x_i - \overline{x})^2$$

Expandindo o quadrado:

$$\sum_{i=1}^{n}(x_i^2 - 2x_i\overline{x} + \overline{x}^2) =$$

$$\sum_{i=1}^{n} x_i^2 - 2n\overline{x}^2 + \sum_{i=1}^{n} x_i^2\overline{x}^2 =$$

$$\sum_{i=1}^{n} x_i^2 - 2n\overline{x}^2 + nx_i^2\overline{x}^2 =$$

$$\sum_{i=1}^{n} x_i^2 - n\overline{x}^2 =$$

$$\sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_i\overline{x} = \sum_{i=1}^{n} \left(x_i^2 - x_i\overline{x}\right)$$

Resíduos e coeficiente de determinação

Variâncias e covariâncias

Lembrando que:

S_{xx} é a variação total elevada ao quadrado de x em relação à média $\bar{x}$

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2$$

S_{yy} é a variação total elevada ao quadrado de y em relação à média $\bar{y}$

$$S_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

Temos também que as variância σ_X^2 e σ_Y^2 são:

$$\sigma_X^2 = \frac{S_{xx}}{n}$$

$$\sigma_Y^2 = \frac{S_{yy}}{n}$$

S_{xy} é o produto da variação total de cada variável em relação à sua média $\bar{x}$ e $\bar{y}$:

$$SS_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

A covariância $Cov(X, Y)$ é:

$$Cov(X, Y) = \frac{S_{xy}}{n}$$

Regressão simples

Conforme foi demonstrado na entrega 2 do projeto, os resultados para regressão de mínimos quadrados são:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

e

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

Lembrando que $\hat{\beta}_0$ e $\hat{\beta}_1$ são os estimadores encontrados a partir dos dados para os parâmetros β_0 e β_1 do modelo de regressão.

Erros na regressão

Soma dos quadrados dos resíduos

A soma dos quadrados dos resíduos é o quadrado da variação encontrada nos dados que **não é explicada** pelo modelo de regressão. Ou seja, é a diferença entre y_i que está presente nos dados e o valor $\hat{y}_i$ que a reta dá para o x_i correspondente.

Este valor costuma ser chamado de soma dos quadrados dos resíduos ($SQRes$) ou também *error sum of squares* ou SS_E

$$SQRes = SS_E = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}\epsilon_i^2$$

Soma dos quadrados da regressão

É a variabilidade que é explicada pela regressão. Tipicamente é chamada de SQR ou SS_R

$$SQReg = SS_R = (\hat{y}_i - \bar{y})^2$$

Soma dos quadrados totais

É a soma da variabilidade total presente no modelo. Costuma ser chamado de SQT ou de SS_T .

$$SQT = SS_T = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

A soma dos quadrados totais é a soma da porção explicada pela regressão com a parte que não é explicada.

$$SS_T = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y}_i)^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = SS_R + SS_E = SQReg + SQRes$$

Coeficiente de determinação R^2

É uma medida de quão bem uma regressão descreve os dados.

$$R^2 = 1 - \frac{SS_E}{SS_T}$$