Entrega 2

Ciência dos Dados

Leonardo Lamounier Grotti

Patrick Serrano Wiegerinck

Prof. Maria Kelly Venezuela

2016

São Paulo

1)

Com a minimização da soma dos quadrados dos resíduos, encontra-se β_0 e β_1, que por sua vez trarão a menor diferença entre a previsão de y e o y realmente observado.

$y = \hat{\beta}_{1x} + \hat{\beta}_0$ → Equação da reta da previsão
$y_1 = \hat{\beta}_{1x} + \hat{\beta}_0 + \epsilon$ → Equação da reta realmente observada

$$S(\beta_0 + \beta_1) = \sum_{i=1}^{n} \left(y_i - \hat{\beta}_1 - \beta_0 x_i\right)^2$$

A minimização é feita ao deixar $S(\beta_0; \beta_1)$ em relação a β_0 e β_1 e, então, igualar a 0

$$\frac{dS}{d\beta_1} = \frac{dS}{dx} \times \frac{dx}{d\beta_1} \rightarrow Equação\ 1$$

$$\frac{dS}{d\beta_1} = 2 \sum_{i=1}^{n} \left(y_i - \hat{\beta}_1 - \beta_0 x_i\right)$$

$$\frac{dx}{d\beta_1} = -1$$

$$\frac{dS}{d\beta_1} = -2 \sum_{i=1}^{n} \left(y_i - \hat{\beta}_1 - \beta_0 x_i\right) \rightarrow 0$$

$$\frac{dS}{d\beta_0} = -2 \sum_{i=1}^{n} x_i \left(y_i - \hat{\beta}_1 - \beta_0 x_i\right) \rightarrow 0 \rightarrow Equação\ 2$$

Substituindo na equação 1 e dividindo por 2n:

$$\frac{-2 \sum_{i=1}^{n} y_i}{2n} + \frac{2 \sum_{i=1}^{n} \beta_1}{2n} + \frac{2 \sum_{i=1}^{n} \beta_0 x_i}{2n} = 0$$

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

Sendo $\bar{y}$ a média amostral de y e $\bar{x}$ a média amostral de x.

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

Substituindo esse resultado na equação 2, temos:

$$-2 \sum_{i=1}^{n} x_i \left(y_i - \bar{y} + \hat{\beta}_0 \bar{x} - \beta_0 x_i\right) = 0$$

$$\sum_{i=1}^{n} x_i(y_i - \bar{y}) + \hat{\beta}_0 \sum_{i=1}^{n} x_i(\bar{x} - x_i) = 0$$

Isolando o $\hat{\beta}_0$, chega-se a segunda resposta:

$$\hat{\beta}_0 = \frac{\sum_{i=1}^{n} x_i(\bar{x} - x_i) \times (y_i - \bar{y})}{\sum_{i=1}^{n} (\bar{x} - x_i)^2}$$

2)

Pode-se assumir que em regressões lineares os erros modelos, $\in_i$, são representados por distribuições normais e independentes com μ igual a 0 e variância igual a σ^2 ($\in_i \sim N(0, \sigma^2)$), ou seja, a variância é constante e, portanto, existe homoscedasticidade. Além disso, assume-se que não existe correlação entre os erros ($Corr\left(\in_i + \in_j\right) = 0$).

Para verificar isso basta analisar a curva de probabilidade cumulativa dos resíduos e a da distribuição normal. A semelhança das curvas definirá se o erro é ou não uma distribuição normal. Outros métodos de verificação são: construção de um intervalo de confiança para a média, com o objetivo de verificar a suposição da média, e verificar graficamente se isso se confirma.

3)

Normalmente, uma das hipóteses em análise de regressão é avaliar a significância da regressão, ou seja, os testes de hipóteses verificam a qualidade da regressão para a variável resposta. (No nosso caso Expectativa de vida).

A hipótese nula é: $\hat{\beta}_1 = 0$ e ela diz que não há relação entre x (variável explicativa) e y (variável resposta), por outro lado a hipótese alternativa é beta1 diferente de 0 e nesse caso há relação entre x e y.

Concluindo, caso a hipótese nula seja rejeitada, podemos concluir que há relação entre a variável explicativa e a variável resposta.

4)

Sim, nesse caso estaremos fazendo uma regressão linear múltipla. Para isso acontecer temos que analisar no mínimo 3 variáveis sendo uma a variável resposta e duas ou mais as variáveis explicativas.

Para a equação devemos acrescentar mais termos de acordado com a quantidade de variáveis estudadas ficando assim:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_n x_n + \varepsilon$$

No caso do teste de hipóteses devem ser feitos um teste para cada variável explicativa e eles se comportam exatamente da mesma maneira que na regressão linear simples.

As suposições do modelo continuam iguais, pois, como já foi definido anteriormente, pode-se sempre assumir as suposições do item 2 para regressões lineares.