Projeto 3 - Ciência dos Dados

Bruno Gomes e Vitória Rocha
Turma 2B

2ª Etapa: Parte Teórica

A segunda etapa do Projeto 3 consiste no estudo teórico da técnica empregada em estatística denominada regressão, cujo objetivo é analisar como as variáveis explicativas influenciam a variável resposta.

Nesse caso, adotamos como variáveis explicativas o PIB per capita das populações de diferentes países do mundo e os respectivos percentual em relação ao acesso a saneamento básico. E como variável resposta, escolhemos o índice de mortalidade de crianças menores de 5 anos de idade. Todas as variáveis correspondem ao ano de 2007 e foram encontradas a partir do site WWW.gapminder.org.

Análise de Regressão

O tipo mais simples de análise de regressão é a regressão linear simples, a qual envolve a variável explicativa como X e a variável resposta como Y. A equação que representa o modelo de regressão linear simples é a seguinte:

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i = \hat{y}_i + \varepsilon_i,$$

em que

Y_i : valor da variável resposta associada ao i-ésimo elemento da amostra
x_i : valor da variável explicativa associada ao i-ésimo elemento da amostra
β_0 : parâmetro que denota o intercepto da equação
β_1 : parâmetro que denota o coficiente angular da equação
$\hat{y}_i$: é chamada regressão de y em X, como o valor médio de Y dado um determinado x
ε_i : erro estocástico (aleatório)
i : $1, 2, 3, \ldots, n$; e
n : tamanho da amostra

A figura a seguir exemplifica o ajuste da reta que passa mais próxima ao mesmo tempo de todos os pontos, através do método dos mínimos quadrados.

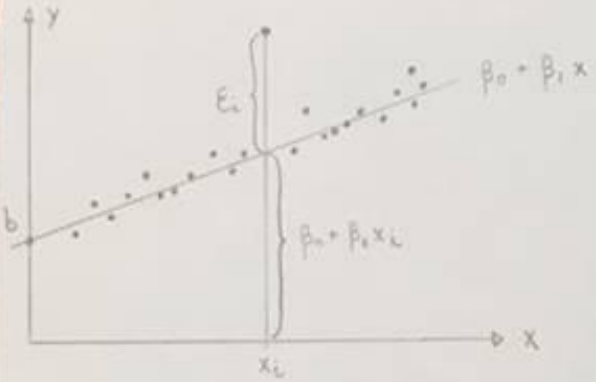

GRÁFICO 1: Representação da Reta de Regressão

a) É através do Método dos Mínimos Quadrados que encontraremos os estimadores β_0 e β_1, da seguinte forma:

$$SQE = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

β_0 :

$$\frac{dSQE}{d\beta_0} = 0 = \sum_{i=1}^{n} 2(y_i - \beta_0 - \beta_1 x_i)(-1)$$

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

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

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

$$-\bar{y} + \beta_0 + \beta_1 \bar{x} = 0$$

$$\boxed{\beta_0 = -\beta_1 \bar{x} + \bar{y}}$$

β_1 :

$$\frac{dSQE}{d\beta_1} = 0 = \sum_{i=1}^{n} 2(y_i - \beta_0 - \beta_1 x_i)(-x_i)$$

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

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

$$\frac{-\cancel{2}\sum x_i (y_i + \beta_1 \bar{x} - \bar{y} - \beta_1 x_i) = 0}{-\cancel{2}}$$

$$\sum_{i=1}^{n} x_i (y_i - \bar{y}) + \beta_1 \sum_{i=1}^{n} x_i (\bar{x} - x_i) = 0$$

$$\boxed{\beta_1 = \frac{\sum_{i=1}^{n} x_i (y_i - \bar{y})}{\sum_{i=1}^{n} x_i (x_i - \bar{x})}}$$

b) Com respeito dos erros, podemos dizer que eles apresentam um modelo de uma distribuição normal, o valor esperado igual a zero e a variância constante. Essas suposições podem ser checadas, na prática, através da probabilidade de termos um erro menor que 50% do valor esperado, da seguinte maneira:

Ao construir a normal, a média é equivalente a 0, pois é o ponto da reta que buscamos descobrir.

A variância é igual a: $Var(\varepsilon_i) = \sigma^2 = E(\varepsilon_i^2) - E^2(\varepsilon_i)$.

E, como podemos notar, a variância é constante ao longo da reta e esse fenômeno é chamado de homocedasticidade.

c) Em relação aos tipos de hipóteses na regressão linear simples, podem ser consideradas duas hipóteses: alternativa e nula.

H_0: $\beta_1 = 0$ → não há relação entre x e y

H_1: $\beta_1 \neq 0$ → há relação entre x e y

Assim, se a H_0 (hipótese nula) for rejeitada, significa que há alguma relação entre x e y e, caso não seja rejeitada, dizemos que não há relação entre x e y.

d) Sim, é possível realizar uma regressão com mais de uma variável explicativa, mas conforme é inserida mais variáveis, a equação torna-se cada vez mais complexa.

As mudanças que ocorreram quando comparada com a regressão simples são as seguintes:

* Ao incluir mais variáveis, a equação assume a seguinte forma:
  $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_m x_m$, sendo m a quantidade de variáveis.
* As suposições continuaram iguais as do item (b)
* Em relação ao teste de hipóteses, permanecerá igual ao do item (c), com o aumento do número de hipóteses, conforme são inseridas novas variáveis.