



UNIDADE IX – REGRESSÃO

PROF.: RÔMULO MÔRA

CUIABÁ, MT
2015/2

1. INTRODUÇÃO

A análise de regressão consiste na realização de uma análise estatística com o objetivo de verificar a existência de uma relação funcional entre uma variável dependente com uma ou mais variáveis independentes. Em outras palavras, consiste na obtenção de uma equação que tenta explicar a variação da variável dependente pela variação dos níveis da (s) variável (is) independente (s).

As variáveis independentes são classificadas como quantitativas, cujos níveis representam diferentes quantidades de um mesmo fator.

2. ESCOLHA DO MODELO

✓ Definição das variáveis

- variável independente – variável x
- variável dependente – variável y

✓ comportamento da relação

- linear;
- quadrático;
- cúbico;
- exponencial;
- outros.

Contudo, pode-se verificar que os pontos do diagrama de dispersão, não vão se ajustar perfeitamente a curva do modelo matemático proposto.

Haverá na maior parte dos pontos, uma distância entre os pontos do diagrama e a curva do modelo matemático. Isto acontece, devido ao fato do fenômeno que está em estudo, não ser um fenômeno matemático e sim um fenômeno que está sujeito a influências que acontecem ao acaso.

Assim, o objetivo da regressão é obter um modelo matemático que melhor se ajuste aos valores observados de Y em função da variação dos níveis da variável X.

3 MÉTODO PARA OBTER A EQUAÇÃO AJUSTADA

3.1 – Regressão linear simples

O modelo de regressão definido como um modelo estatístico, difere em conceito de um modelo matemático por apresentar um termo denominado erro aleatório.

Modelo Matemático

$$Y = \beta_0 + \beta_1 X$$

Modelo de regressão

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

⇒ O método de mínimos quadrados ordinários

Este é um método de estimação muito empregado para estimar parâmetros de modelos de regressão. Seu objetivo principal consiste em minimizar a soma de quadrados dos erros, tal como será demonstrado a seguir:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad (1)$$

$$\varepsilon_i = Y_i - (\beta_0 + \beta_1 X_i) \quad (2)$$

$$\sum_{i=1}^n \varepsilon_i^2 = \sum_{i=1}^n (Y_i - \beta_0 - \beta_1 X_i)^2 \quad (3)$$

$$\frac{\partial \sum_{i=1}^n \varepsilon_i^2}{\partial \beta_0} = 2 \sum_{i=1}^n (Y_i - \beta_0 - \beta_1 X_i)(-1) = 0 \quad (4)$$

$$\frac{\partial \sum_{i=1}^n \varepsilon_i^2}{\partial \beta_1} = 2 \sum_{i=1}^n (Y_i - \beta_0 - \beta_1 X_i)(-X_i) = 0 \quad (5)$$

Dividindo-se (4) e (5) por -2, tem-se:

$$\begin{cases} \sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = 0 \\ \sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) X_i = 0 \end{cases} \quad \begin{cases} \sum_{i=1}^n Y_i - n\hat{\beta}_0 - \hat{\beta}_1 \sum_{i=1}^n X_i = 0 \\ \sum_{i=1}^n X_i Y_i - \hat{\beta}_0 \sum_{i=1}^n X_i - \hat{\beta}_1 \sum_{i=1}^n X_i^2 = 0 \end{cases}$$

$$\sum_{i=1}^n Y_i - n\hat{\beta}_0 - \hat{\beta}_1 \sum_{i=1}^n X_i = 0 \quad (6)$$

$$\sum_{i=1}^n X_i Y_i - \hat{\beta}_0 \sum_{i=1}^n X_i - \hat{\beta}_1 \sum_{i=1}^n X_i^2 = 0 \quad (7)$$

$$\hat{\beta}_0 = \frac{\sum_{i=1}^n Y_i}{n} - \hat{\beta}_1 \frac{\sum_{i=1}^n X_i}{n}$$

$$\boxed{\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}} \quad (8)$$

Substituindo (8) em (7) e desenvolvendo a expressão, tem-se:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n X_i Y_i - \frac{\sum_{i=1}^n X_i \sum_{i=1}^n Y_i}{n}}{\sum_{i=1}^n X_i^2 - \frac{\left(\sum_{i=1}^n X_i\right)^2}{n}} \quad (9)$$

EQUAÇÃO DE REGRESSÃO AJUSTADA

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X \quad (10)$$

ANÁLISE DE VARIÂNCIA DA REGRESSÃO (ANOVA)

FV	GL	SQ	QM	FCalc.
Regressão	p	$\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$	$SQReg/GLReg (V_1)$	V_1/V_2
Resíduo	$n - p - 1$	$\sum_{i=1}^n (Y_i - \hat{Y}_i)^2$	$SQRes/GLRes (V_2)$	
Total	$n - 1$	$\sum_{i=1}^n (Y_i - \bar{Y})^2$		

em que:

n = número total de dados;

p = número de variáveis independentes do modelo;

Y_i = valores observados para Y ;

$\hat{Y}$ = valores estimados de Y pela regressão;

$\bar{Y}$ = valor médio de Y .

$$SQRe g = \hat{\beta}_1 \left(\sum_{i=1}^n X_i Y_i - \frac{\left(\sum_{i=1}^n X_i\right) \left(\sum_{i=1}^n Y_i\right)}{n} \right)$$

- Soma de Quadrados do Residuo (SQRes)

$$SQRes = SQTot - SQReg$$

As hipóteses estatísticas testadas pelo teste F

$$\begin{cases} H_0 : \beta_1 = 0 \\ H_1 : \beta_1 \neq 0 \end{cases}$$

Se o valor de F calculado for menor do que o valor tabelado obtido em uma tabela de F a um determinado nível de probabilidade α com $(1; n - 2)$ graus de liberdade, não se rejeita H_0 .

Em termos práticos, se F calculado $> F$ tabelado, então a regressão existe, ou seja, variações ocorridas em Y podem ser explicadas pelas variações ocorridas em X de acordo com a equação ajustada em um nível α de significância.

⇒ **Medidas de precisão da equação ajustada:**

- **Coeficiente de Determinação (R^2):** Informa a percentagem da variação dos dados observados em torno da média que está sendo explicada pela equação ajustada.

$$R^2(\%) = \frac{SQReg}{SQTotal} 100$$

Exemplo de aplicação:
Considere o seguinte conjunto de diâmetros e alturas

Árv	D(cm) - X	H(m) - Y
1	8,0	9,7
2	27,7	27,6
3	23,2	26,5
4	17,7	17,4
5	13,8	12,9
6	17,0	16,5
7	18,8	20,3
8	8,0	11,6
9	15,0	16,7
10	21,6	21,2
11	11,0	12,8
12	24,2	24,7
Soma	206	217,9

FIM