3 - REGRESSÃO

Para a análise de regressão presume-se que pelo menos duas observações são feitas sobre cada elemento da amostra. A amostra consistirá, então, de pares de valores, um valor para cada uma das variáveis, designadas X e Y. Um indivíduo “i” qualquer apresenta o par de valores (X_i; Y_i).

O objetivo visado quando se registra pares de valores (observações) de uma amostra é o estudo das relações entre as variáveis X e Y.

Para a análise de regressão interessam, principalmente, os casos em que a variação de uma variável é sensivelmente dependente de outra variável.

O problema consiste em estabelecer a função matemática que melhor represente a relação existente entre as duas variáveis. Simbolicamente a relação é expressa por uma equação de regressão e graficamente por uma curva de regressão.

3.1 - REGRESSÃO LINEAR SIMPLES

A análise de regressão linear tem por objetivo descrever através de um modelo matemático, a relação existente entre duas variáveis, a partir de n observações dessas variáveis, através do seguinte modelo:

$$Y = \alpha + \beta x + \varepsilon i,$$

onde “α” e “β” são os parâmetros da reta teórica da nuvem de pontos.

Estimaremos os parâmetros α e β da reta por meio dos valores estimados “a” e “b” fornecidos pela amostra, logo:

$$\hat{y} = a + b.x + \varepsilon_i$$

será a fórmula geral da equação de regressão, onde:

- a é o coeficiente linear, ponto onde a reta corta o eixo da variável y;
- b é o coeficiente angular, tangente do ângulo que a reta forma com a paralela do eixo da variável x;
- ε_i é o erro aleatório.

Pressuposições básicas para realizar a regressão:

a) A relação entre X e Y é linear (os acréscimos em X produzem acréscimos proporcionais em Y e a razão de crescimento é constante).
b) Os valores de X são fixados arbitrariamente, ou seja, X não é uma variável aleatória (v.a.).
c) Y é uma v.a. que depende, entre outras coisas, dos valores de X.

d) ε_i é o erro aleatório, portanto uma v.a. com distribuição normal, de média zero e variância σ^2, [$\varepsilon_i \cong N(0, \sigma^2)$]. ε_i representa a variação de Y que não é explicada pela variável independente X.
e) Os erros são considerados independentes.

Com isto temos o objetivo de:

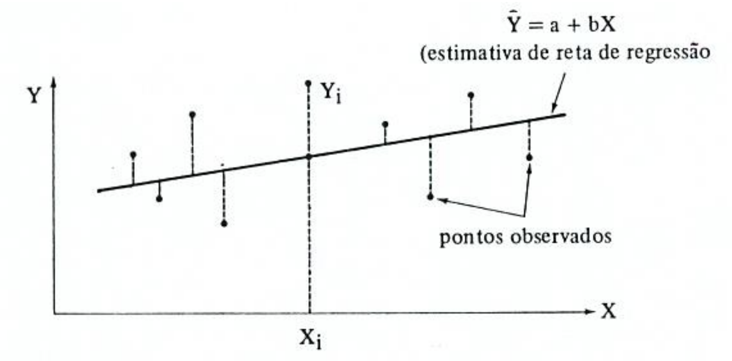
⇨ estimar valores de uma variável, com base em valores conhecidos da amostra;
⇨ explicar valores de uma variável em termos da amostra.

Para estimarmos "Y" a partir de "X", expressamos Y como uma função linear de X, interpolando a nuvem de pontos em uma reta, sendo que a reta que forneceu melhor "ajustamento" deve ser escolhida.

A escolha dessa reta obedece ao critério dos Mínimos Quadrados e a reta de regressão tem a propriedade de sempre passar pelo ponto $(\overline{X}, \overline{Y})$.

3.2 - MÉTODO DOS MÍNIMOS QUADRADOS (M.M.Q.)

O MMQ é aquele que torna mínima a soma dos quadrados das distâncias da reta aos pontos experimentais, medidas no sentido da variação aleatória, ou seja, devemos procurar uma reta que minimiza $\Sigma(Y - \hat{Y})$, que são os erros.



O M.M.Q. consiste em adotar como estimativa dos parâmetros os valores que minimizem a soma dos quadrados dos desvios.

Como a reta a ser determinada será utilizada para fins de previsão, é necessário determinar a equação que forneça os menores erros de previsão, ou seja, a menor diferença entre o valor real e o previsto: $(Y - \hat{Y})$.

Para não usarmos relações lineares distorcidas, minimizar a soma dos erros não é o bastante e sim devemos minimizar a soma dos quadrados dos erros $(Y - \hat{Y})^2$ na obtenção de "a" e "b". Assim:

$$\sum \text{desvio}^2 = \sum e^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$z = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} [y_i - a - bx_i]^2 \quad \text{ou} \quad f(a, b)$$

Essa soma, função de “a” e de “b”, terá mínimo quando suas derivadas parciais em relação a “a” e “b” forem nulas. Para facilitar, considera-se $\sum\limits_{i=1}^{n} = \sum$

$$\begin{cases} \dfrac{\gamma z}{\gamma a} = \sum 2\left[y_i - a - bx_i\right](-1) = 0 \\ \dfrac{\gamma z}{\gamma b} = \sum 2\left[y_i - a - bx_i\right](-x_i) = 0 \end{cases}$$

$$\begin{cases} \sum\left[y_i - a - bx_i\right] = 0 \\ \sum\left[x_i y_i - ax_i - bx_i^2\right] = 0 \end{cases}$$

$$\begin{cases} \sum y_i - na - b\sum x_i = 0 \\ \sum x_i y_i - a\sum x_i - b\sum x_i^2 = 0 \end{cases}$$

Daí obtém-se as equações normais da reta.

$$\begin{cases} na + b\sum x_i = \sum y_i & (1) \\ a\sum x_i + b\sum x_i^2 = \sum x_i y_i & (2) \end{cases}$$

Resolvendo-se esse sistema obtém-se:

$$\frac{na}{n} + b\frac{\sum x_i}{n} = \frac{\sum y_i}{n} \quad \Rightarrow \quad a + b\bar{x} = \bar{y} \quad \Rightarrow \quad \boxed{a = \bar{y} - b\bar{x}} \qquad (3)$$

Substituindo 3 em 2, temos:

$$(\bar{y} - b\bar{x})\sum x_i + b\sum x_i^2 = \sum x_i y_i$$

$$\left(\frac{\sum y_i}{n} - b\frac{\sum x_i}{n}\right)\sum x_i + b\sum x_i^2 = \sum x_i y_i$$

$$\frac{(\sum y_i).(\sum x_i)}{n} - b\frac{(\sum x_i).(\sum x_i)}{n} + b\sum x_i^2 = \sum x_i y_i$$

$$\frac{(\sum y_i).(\sum x_i)}{n} - b\frac{(\sum x_i)^2}{n} + b\sum x_i^2 = \sum x_i y_i$$

$$\frac{(\sum y_i).(\sum x_i)}{n} + b\left[-\frac{\sum(x_i)^2}{n} + \sum x_i^2\right] = \sum x_i y_i$$

$$b\left[-\frac{\sum(x_i)^2}{n} + \sum x_i^2\right] = \sum x_i y_i - \frac{(\sum y_i).(\sum x_i)}{n}$$

$$b = \frac{\sum x_i y_i - \frac{(\sum y_i).(\sum x_i)}{n}}{\sum x_i^2 - \frac{(\sum x_i)^2}{n}} = \frac{S_{xy}}{S_{xx}}$$

onde: $$S_{xy} = \sum x_i y_i - \frac{\sum x_i \sum y_i}{n}$$

$$S_{xx} = \sum x_i^2 - \frac{(\sum x_i)^2}{n}$$

$$S_{yy} = \sum y_i^2 - \frac{(\sum y_i)^2}{n}$$

Interpretação do Coeficiente angular "b":

1º) Se "b" for positivo, indica que acréscimos da variável independente corresponderão a acréscimos na variável dependente, assim a regressão é direta;

2º) Se "b" for negativo, aos acréscimos da variável independente corresponderão decréscimos da variável dependente, ou seja, temos uma relação inversa;

3º) Se "b" for nulo, não há relação entre X e Y e a reta será paralela ao eixo X.

Exemplo 3.1: Os registros a seguir correspondem ao número de filhos de uma família (Y) e a escolaridade da mãe (em anos de estudo).
Pede-se: a) construir o diagrama de dispersão; b) verificar se existe correlação linear entre essas variáveis; c) ajustar uma equação de regressão linear; d) interpretar os coeficientes

	X_i	Y_i	X_iY_i	X_i^2	Y_i^2
	0	3	0	0	9
	1	3	3	1	9
	2	5	10	4	25
	3	4	12	9	16
	5	6	30	25	36
	6	7	42	36	49
Total	17	28	97	75	144

Sxx =

Sxy =

Syy =

b =

a =

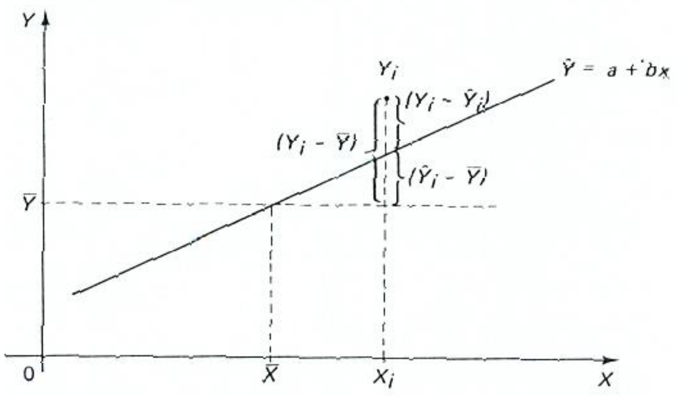
$$\hat{Y} = 2{,}8 + 0{,}66X$$

3.3 - DECOMPOSIÇÃO DA SOMA DE QUADRADOS

A variação de Y compreende três tipos de desvios, que podem ser observados na figura:

a) Desvios totais $\left(Y_i - \overline{Y}\right)$;
b) Desvios explicados ou de regressão $\left(\hat{Y}_i - \overline{Y}\right)$;
c) Desvios não explicados ou resíduos $\left(Y_i - \hat{Y}_i\right)$

Gráfico: Representação da decomposição das Somas de Quadrados



É fácil verificar que os desvios totais resultam da soma dos desvios explicados com a dos desvios não explicados, isto é:

A dispersão da variação aleatória de Y pode ser medida por meio da soma dos quadrados dos desvios em relação a sua média. Essa soma de quadrados será denominada Soma dos Quadrados Total (SQ_{Total}).

$$SQ_{Total} = \Sigma(Y_i - \overline{Y})^2$$

A SQ_{Total} pode ser decomposta da seguinte forma:

$$SQ_{Total} = SQ_{Regressão} + SQ_{Resíduo}$$

$$\Sigma(Y_i - \overline{Y})^2 = \Sigma(\hat{Y} - \overline{Y})^2 + \Sigma(Y_i - \hat{Y}_i)^2$$

Assim, a soma dos quadrados dos desvios em torno da média é igual a soma dos quadrados dos desvios da linha de regressão em torno da média mais a soma dos quadrados dos desvios em torno da reta de regressão.

Essa relação mostra que a variação dos valores de Y em torno de sua média pode ser dividida em duas partes: uma que é explicada pela regressão (variação explicada) e outra, devido ao fato de que nem todos os pontos estão sobre a reta de regressão, que é a parte “não explicada” pela regressão (variação residual).

3.4 - CÁLCULO PRÁTICO DAS VARIAÇÕES

a) Variação Total (VT)

$$SQ_{Total} \sum\left(Y - \overline{Y}\right)^2 = \sum Y_i^2 - \frac{\left(\sum Y_i\right)^2}{n} = S_{yy}$$, com (n – 1) graus de liberdade.

b) Variação Explicada (devido à regressão = VE)

$$SQ_{Regressão} = \Sigma(\hat{Y}_i - \overline{Y})^2 = \Sigma(a + bX - \overline{Y})^2 = \Sigma[(\overline{Y} - b\overline{X}) + bX - \overline{Y}]^2$$
$$= \Sigma[b(X - \overline{X})]^2 = b^2\,\Sigma(X - \overline{X})^2 = b^2 S_{xx} = b \cdot S_{xy}$$, com 1 g. l.

c) Variação Residual (VR)

$$SQ_{Total} = SQ_{Regressão} + SQ_{Resíduo}$$

$$\Sigma(Y_i - \overline{Y})^2 = \Sigma(\hat{Y}_i - \overline{Y})^2 + \Sigma(Y_i - \hat{Y}_i)^2.$$

Logo, por $SQ_{Resíduo} = S_{yy} - b\,S_{xy}$, com n – 2 g. l.

3.5 - COEFICIENTE DE DETERMINAÇÃO OU DE EXPLICAÇÃO r^2_{xy} ou R^2

Este coeficiente indica quanto por cento a variação explicada pela regressão representa da variação total.

$$R^2 = \frac{VE}{VT} = \frac{b^2 . S_{xx}}{S_{yy}} = \frac{SQ_{Regressão}}{SQ_{Total}} \quad ou \quad R^2\% = R^2 .100$$, logo:

$$0\% \leq R^2 \leq 100\% \qquad ou \qquad 0 \leq R^2 \leq 1.$$

3.6 - TESTE PARA O COEFICIENTE DE REGRESSÃO LINEAR β

Os testes utilizados para a verificar a existência de regressão ou a significância estatística do coeficiente de regressão linear β_1 são: o teste F (Análise de Variância) ou o teste t de Student (no caso da regressão linear simples os testes são equivalentes).

- **Teste para a existência de regressão – teste F**

Podemos testar a significância da regressão utilizando a análise de variância, ou seja, estudar o comportamento das variações totais, explicadas e residuais. A Soma de Quadrados da Regressão ($SQ_{Regressão}$) segue uma distribuição de χ^2 (qui-quadrado) com 1 grau de liberdade, enquanto que a Soma de Quadrados dos Resíduos ($SQ_{Resíduo}$) segue a mesma distribuição, porém com (n–2) graus de liberdade. Portanto, o quociente:

$$\frac{\frac{SQ_{Regressão}}{1}}{\frac{SQ_{Resíduo}}{n-2}} = \frac{QM_{Regressão}}{QM_{Resíduo}}$$

segue a distribuição F de *Snedecor*, com 1 g.l. no numerador e (n–2) g.l. no denominador [F(m, n)].

Para sintetizar utilizamos o quadro da ANOVA a seguir, sob as seguintes hipóteses:

H_0: $\beta_1 = 0$, não existe regressão linear entre as variáveis X e Y
H_1: $\beta_1 \neq 0$, existe regressão linear entre as variáveis X e Y

ANOVA - Análise de Variância

Causas de Variação	**GL**	**SQ**	**QM**	**F**
Regressão	1 (b)	$SQ_{Regressão}$	$\frac{SQ_{Regressão}}{1}$	$\frac{QM_{Regressão}}{QM_{Resíduo}}$
Resíduo	n – 2 (c)	$SQ_{Resíduo}$	$\frac{SQ_{Resíduo}}{n-2}$	
Total	n – 1 (a)	SQ_{Total}		

onde: QM representa o quadro médio, obtido pela divisão das somas de quadrados pelos respectivos graus de liberdade.

De forma equivalente podemos escrever:

ANOVA - Análise de Variância

Fonte de variação	**Soma de Quadrados**	**Graus de liberdade**	**Quadrado Médio**	$\mathbf{F_c}$	$\mathbf{F_\alpha}$
Regressão	1	b^2S_{xx}	b^2S_{xx}	$b^2S_{xx}/\ S_R^2$	$F_{1;n-2;\alpha}$
Residual	n-2	$S_{yy}-b^2S_{xx}$	$S_R^2=S_{yy}-b^2S_{xx}/\ n-2$		
Total	n-1	S_{yy}			

Nota: $b^2S_{xx} = bS_{xy}$

Além das hipóteses e do quadro ANOVA, determinamos o nível de significância α.

Conclusão: Se $F_c < F_{\alpha;m;n}$ aceita-se H_0, ou seja, conclui-se com risco α, que não há regressão linear significativa.

- **Teste para a existência de regressão – teste t**

1°) Enunciar as hipóteses H_0 e H_1, lembrando que $Y_i = \alpha + \beta X_i + \varepsilon_i$, devemos testar o seguinte: H_0: $\beta = 0$ e H_1: $\beta \neq 0$.

2°) Fixar α, que geralmente é 5% ou 10%, com n – 2 graus de liberdade, sendo:

$$b \cong N\left(\beta;\ \frac{\sigma^2}{S_{xx}}\right), \text{ ou seja, } Z = \frac{(b - \beta)}{\left[\dfrac{\sigma}{(S_{xx})^{1/2}}\right]} \approx N(0,1) \text{ que tem distribuição normal padrão.}$$

Entretanto, quando não conhecemos o verdadeiro valor de σ^2, deveremos estimá-lo usando S^2, daí:

$$t_c = \frac{(b - \beta)}{\left[S / \sqrt{S_{xx}}\right]}$$, que tem distribuição “t” de Student, com n–2 graus de liberdade.

3°) Estabelecer a região de rejeição e a de aceitação de H_0.

4°) Conclusão: Se $-t_{\alpha/2} < t_c < t_{\alpha/2}$, aceita-se H_0 e concluímos, com risco α, que não há regressão linear, pois seu coeficiente angular não foi significativo.

Exemplo 3.2: Considerando o exemplo 3.1, realizar o teste F e o teste t para 1%.

$$SQ_{Regressão} = b\left[\sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n}\right]$$

$$SQ_{Regressão} = 0{,}66\left[97 - \frac{(17)(28)}{6}\right] = 11{,}66$$

$$SQ_{Total} = \sum y_i^2 - \frac{\left(\sum y_i\right)^2}{n} = 144 - \frac{(28)^2}{6} = 13{,}33$$

$SQ_{Resíduo}$ = SQT – SQR = 13,33 – 11,66 = 1,67

Análise de Variância

Causas de variação	**GL**	**SQ**	**QM**	**F**
Regressão	1	11,66	11,66	27,76
Resíduo	4	1,67	0,42	
Total	5			

H_0: $\beta = 0$ contra H_1: $\beta \neq 0$.
Como F calculado = 27,76 > 21,20 = $F_{0,01\ (14)}$, conclui-se que a regressão de Y sobre X segundo o modelo $\hat{Y} = 2{,}8 + 0{,}66X$ é significativa ao nível de 1%. Uma vez estabelecida e testada a equação de regressão, a mesma pode ser usada para explicar o relacionamento entre as variáveis e fazer previsões dos valores de Y para valores fixados de X.

Além disso, se quiséssemos estimar o valor de Y para X=4 anos, procederíamos da seguinte maneira: $\hat{Y} = 2{,}8 + 0{,}66(4)$, encontrando $\hat{Y} = 5{,}44$ filhos.