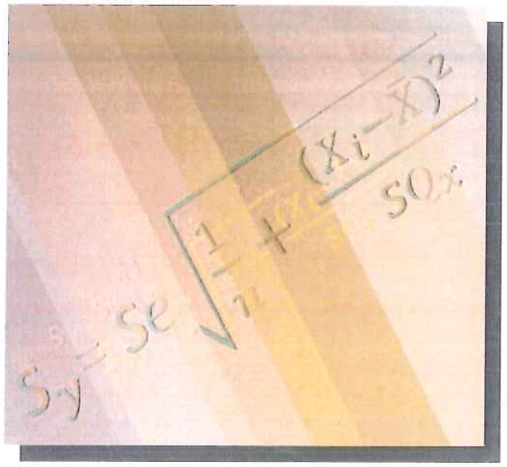
UTILIZAÇÃO DOS MODELOS DE REGRESSÃO (MÉTODO DOS MÍNIMOS QUADRADOS) COM BANDAS DE CONFIANÇA



Eduardo Sá e Silva

Docente do ISCAP

Os modelos de regressão (através dos métodos dos mínimos quadrados) possibilitam que se possam identificar e quantificar relações funcionais entre duas ou mais variáveis. Nestes modelos, tem-se uma variável independente e pode-se ter uma ou mais variáveis dependentes. Um uso frequente destes modelos é para prever. A habilidade de fazer estimativas e de prever eventos e tendências futuras aumenta significativamente as probabilidades de sucesso.

Uma das utilizações são as previsões baseadas nas séries temporais que relacionam o tempo (variável independente) com uma variável dependente, como, por exemplo, o volume de negócios. No entanto outros modelos de previsão podem relacionar quaisquer variáveis, como, por exemplo, determinar o impacto que um determinado gasto tem sobre o volume de negócios ou o nº clientes.

É objetivo do autor apresentar o desenvolvimento de forma simplificada de um exemplo (sem recorrer a aspetos matemáticos complexos) e no fim apresentar o mesmo resultado através da Ferramenta do Excel – ANÀLISE DE DADOS – REGRESSÃO.

No exemplo que é apresentado a seguir vai-se relacionar os gastos com publicidade com o nº clientes e averiguar se existe uma relação direta entre os gastos deste tipo (variável independente) e o número de clientes (variável dependente). Para esse efeito, procedeu-se à recolha de 15 observações (últimos meses).

	Publicidade	Número clientes			
observ.	X	Y	X Y	X^2	Y^2
1	11	16	176	121	256
2	13	18	234	169	324
3	9	14	126	81	196
4	18	21	378	324	441
5	11	17	187	121	289
6	16	22	352	256	484
7	11	15	165	121	225
8	13	21	273	169	441
9	18	23	414	324	529
10	10	17	170	100	289

	Publicidade	Número clientes			
observ.	X	Y	X Y	X^2	Y^2
11	12	15	180	144	225
12	14	19	266	196	361
13	17	24	408	289	576
14	11	16	176	121	256
15	12	14	168	144	196
Total	196	272	3673	2680	5088
Média	13,07	18,13			

A seguir calcularam-se a soma dos quadrados de X, a soma dos quadrados de Y e a soma dos produtos cruzados de XY
Soma dos quadrados de X

$$SQ_x = \sum(X_i - \bar{X})^2 = \sum X^2 - \frac{(\sum X^2)}{n} = 2680{,}00 - 2561{,}07 = 118{,}93$$

Soma dos quadrados de Y

$$SQ_y = \sum(Y_i - \bar{Y})^2 = \sum X^2 - \frac{(\sum Y^2)}{n} = 5088{,}00 - 4932{,}27 = 155{,}73$$

Soma dos produtos cruzados de XY

$$SQ_{xy} = \sum(X_i - \bar{X})(Y_i - \bar{Y}) = \sum XY - \frac{\sum X \sum Y}{n} = 3673{,}00 - 3554{,}13 = 118{,}87$$

Inclinação (declive) da reta de regressão

$$b1\ (beta) = \frac{SQxy}{SQx} = 0{,}999439$$

Interseção (constante) da reta de regressão

$$b0\ (alfa) = \bar{Y} - b1\ \bar{X} = 5{,}073991$$

O modelo da regressão é dado pela seguinte expressão:

$$\hat{Y}_i = 5{,}073991 + 0{,}999439\ X_i$$

Isto quer dizer que por cada variação unitária de X (gastos com publicidade) ter-se-á uma variação de 0,999439 de Y (número de clientes).
No quadro seguinte pode-se constar os erros e os erros ao quadrado, bem como a estatística D-W sobre a correlação dos erros

observ.	Y	Y estimado	Erro (e_t)	e_t^2	$(e_t - e_{t-1})^2$
1	16	16,07	-0,07	0,00	
2	18	18,07	-0,07	0,00	0,00
3	14	14,07	-0,07	0,00	0,00
4	21	23,06	-2,06	4,26	3,98
5	17	16,07	0,93	0,87	8,98
6	22	21,07	0,93	0,87	0,00
7	15	16,07	-1,07	1,14	4,01
8	21	18,07	2,93	8,60	16,01
9	23	23,06	-0,06	0,00	8,98
10	17	15,07	1,93	3,73	3,98
11	15	17,07	-2,07	4,27	15,99
12	19	19,07	-0,07	0,00	4,00
13	24	22,06	1,94	3,75	4,01
14	16	16,07	-0,07	0,00	4,01
15	14	17,07	-3,07	9,41	9,00
		272,00	0,00	36,93	82,95
			D-W		2,25

Um dos pressupostos do método dos mínimos quadrados é que os erros são independentes entre si. Se existir autocorrelação isto quer dizer que o erro de um determinado período está relacionado com o erro do período anterior, logo os erros não são independentes. A estatística que é usada para testar se existe autocorrelação é a estatística Durbin-Watson (D-W) e se porventura o seu valor se situar próximo de 2 não existe nenhuma autocorrelação, ou seja, os erros são independentes.

A expressão para a estatística D-W (já referida) é a seguinte com o valor calculado para o caso presente:

$$\text{D-W} = \frac{\sum(e_t - e_{t-1})^2}{\sum e_t^2} = 82.95 / 36,93 = 2,25$$ (não existe autocorrelação dos erros)

Sendo:

$$e_i = Y_i - \hat{Y}_i$$

Relativamente à qualidade de ajustamento, existem, pelo menos, duas medidas:

1) Erro padrão da estimativa (*Se);*
2) Coeficiente de determinação *(r^2)*

O erro padrão da estimativa (*Se*) é uma medida de dispersão dos valores de Y em volta da reta de regressão e é dado pela seguinte expressão. O erro padrão da estimativa é muito similar ao desvio padrão

$$Se = \sqrt{\frac{\sum(Y_i - \hat{Y}_i)^2}{n-2}} = \sqrt{\frac{\sum e_i^2}{n-2}} = \sqrt{\frac{36,93}{15-2}} = 1,6855$$

Nota:

A razão pela qual a soma dos quadrados dos erros é dividida por "*n -2*", tem a ver com o fato de existirem dois parâmetros a estimar (b0 e b1)

A situação do b1 ser positivo indica a direção da relação. Quando aumentam os gastos com publicidade, também aumenta o número de clientes. No entanto, seria útil ter uma medida da intensidade dessa relação. Essa é a função do coeficiente de correlação que regra geral é representado por *r*. O coeficiente Se correlação está compreendido ente -1 e + 1. A expressão para o coeficiente de correlação é a seguinte:

$$r = \frac{SQ_{xy}}{\sqrt{SQ_x SQ_y}} = 0,873409$$

Neste caso, existe uma forte relação positiva entre o número de clientes e a quantia gasta em publicidade.

Uma outra medida é o coeficiente de determinação, r^2, que é o quadrado do *r* e revela qual a percentagem de variação de Y que é explicada pela variação de X.

$$r^2 = 0,762843$$

Isto quer dizer que 76,3% da variação no número de clientes é explicada pelos gastos com publicidade. No entanto o r^2 só tem significado para relações lineares.

Outra questão relaciona-se com a significância da inclinação (declive) b1. Se a inclinação da reta de regressão dá verdadeira, mas desconhecida, população é zero, não existe relação entre o número de clientes e os gastos com publicidade. Assim, devem-se testar as seguintes hipóteses:

Ho (hipótese nula): $\beta 1 = 0$

H1 (alternativa): $\beta 1 \neq 0$

Isto implica a utilização da estatística *t:*

Teste t para a inclinação de regressão:

$$t_{teste} = \frac{b1 - \beta 1}{s_{b1}}$$ com n-2 graus de liberdade (no caso presente 13=15-2), onde S_{b1} é o erro padrão da distribuição amostral de b1. Diferentes amostras conduzem a diferentes valores de b1. No caso de $\beta 1 = 0$, os valores de b1 devem estar distribuídos em redor de 0. O erro padrão da inclinação (também designada por coeficiente da regressão) é dado pela seguinte expressão:

$$s_{b1} = \frac{Se}{\sqrt{SQ_x}} = \frac{36,93}{\sqrt{118,93}} = 0,154556$$

e o

$$t_{teste} = \frac{0,99439-0}{0,154556} = 6,466525$$

Então para um valor de alfa de 5% (nível de significância de 5% - grau de confiança de 95%) , tem-se que o valor da tabela é igual a $t_{0,05,13} = \pm 2,160$, o que conduz ao seguinte intervalo:

Intervalo para $\beta 1 = b1 \pm t\,(S_{b1})$, o que no caso presente conduz aos seguintes valores

Intervalo para $\beta 1 = 0,999439 \pm 2,160 \times 0,154556$

Com os limites inferior e superior de:

LIMITE MÍNIMO PARA o b1	0,6656
LIMITE MÁXIMO PARA o b1	1,3333

Isto significa que se pode estar 95% certo de que o coeficiente de regressão (inclinação) para a população de todos os valores de X e Y estará compreendido ente 0,6656 e 1,333.

A comprovar igualmente o valor do teste de 6,466525 que é nitidamente superior ao valor da tabela de 2,160, o que nos leva a rejeitar a hipótese nula de $\beta 1 = 0$.

Outro teste que deve ser realizado é o relativo ao coeficiente de correlação para a população (ρ), partindo do coeficiente de correlação amostral de $r = 0,873409$. Assim, tem-se:

Ho (hipótese nula): $\rho = 0$

H1 (alternativa): $\rho \neq 0$

O teste *t* é igualmente utilizado para esse efeito:

t teste para o coeficiente de correlação da população:

$$t_{teste} = \frac{r-\rho}{s_r}$$

onde S_r é o erro padrão do coeficiente de correlação é dado pela seguinte expressão:

$$s_r = \sqrt{\frac{1-r^2}{n-2}} = 6,466525$$

Como $t = 6,466525 > 2,160$, a hipótese nula é rejeitada. Ao nível de 5% de significância, conclui-se que o coeficiente de correlação da população não é zero e que o número de clientes e gastos com publicidade estão relacionados.

Suponhamos, agora, que se queira desenvolver uma estimativa intervalar para média condicional de Y, $\mu_{(y|x)}$. Ela é a média populacional para todos os valores de Y sob a condição de que X é igual a um determinado valor de X. Este intervalo de confiança é construído para o valor da média da população de todos os valores de Y quando X é igual a um dado valor.

A expressão para esta média condicional $\mu_{(y|x)}$ é dada pela seguinte expressão:

Intervalo de confiança para a média condicional $\mu_{(y|x)} = \hat{Y}_i \pm tS_y$

O valor de *t* é igualmente baseado em *n-2* graus de liberdade. Relativamente ao erro padrão da média condicional, ela é dada pela seguinte expressão:

$$S_y = Se\sqrt{\frac{1}{n} + \frac{(X_i-\bar{X})^2}{SQ_x}}$$

Em que:

Se – erro padrão da estimativa

Para o caso de se pretender construir um intervalo de confiança para um único valor de Y (em vez de ser várias vezes) quando X é igual a determinado valor, a expressão anterior é modificada como segue:

$$S_{y'} = Se\sqrt{1 + \frac{1}{n} + \frac{(X_i-\bar{X})^2}{SQ_x}}$$

Os limites inferiores e superiores dessas bandas são indicados a seguir:

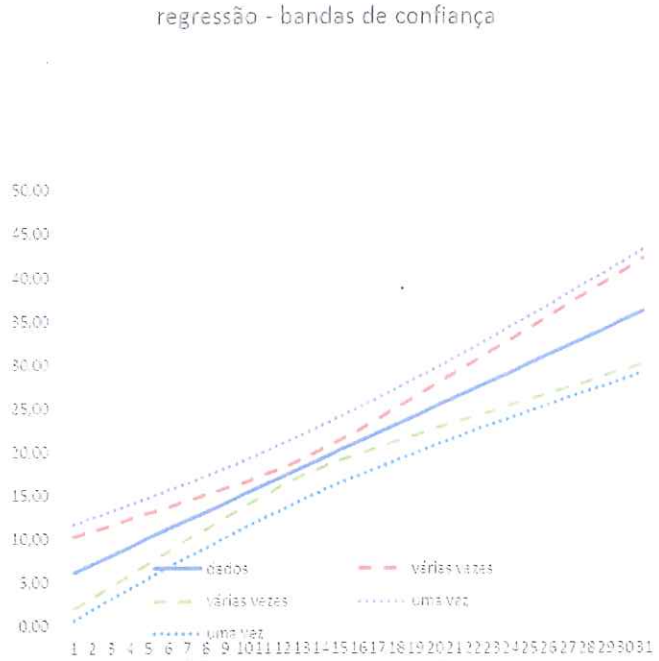
	Se ---->		1,685533						
	1 vez		várias vezes		dados	várias vezes		uma vez	
nº obser.	Sy		Sy'		regressão	superior	inferior	superior	inferior
1	1,513578	2,551187	1,136186	1,91508	6,07	10,21	1,94	11,58	0,56
2	1,447899	2,440483	1,047097	1,764917	7,07	10,89	3,26	12,34	1,80
3	1,385179	2,334766	0,9585	1,615583	8,07	11,56	4,58	13,12	3,03
4	1,325838	2,234745	0,870544	1,467332	9,07	12,24	5,90	13,90	4,24
5	1,27035	2,141217	0,783447	1,320526	10,07	12,92	7,22	14,70	5,45
6	1,21924	2,05507	0,697529	1,175709	11,07	13,61	8,53	15,51	6,63
7	1,173082	1,977269	0,613287	1,033716	12,07	14,30	9,84	16,34	7,80
8	1,13248	1,908833	0,531518	0,895891	13,07	15,00	11,13	17,19	8,95
9	1,098052	1,850803	0,453561	0,764492	14,07	15,72	12,42	18,07	10,07
10	1,070392	1,804182	0,381759	0,643467	15,07	16,46	13,68	18,97	11,17
11	1,050037	1,769873	0,320279	0,539841	16,07	17,23	14,90	19,89	12,24
12	1,037417	1,7486	0,276104	0,465382	17,07	18,07	16,06	20,84	13,29
13	1,032814	1,740842	0,258271	0,435325	18,07	19,01	17,13	21,83	14,31
14	1,036335	1,746778	0,272013	0,458487	19,07	20,06	18,08	22,84	15,29
15	1,0479	1,76627	0,3132	0,527909	20,07	21,21	18,93	23,88	16,25
16	1,067246	1,798878	0,372845	0,628443	21,07	22,42	19,71	24,95	17,18
17	1,09396	1,843906	0,443564	0,747642	22,06	23,68	20,45	26,05	18,08
18	1,12752	1,900472	0,520865	0,877936	23,06	24,96	21,17	27,17	18,96
19	1,167334	1,96758	0,602219	1,015061	24,06	26,26	21,87	28,31	19,81
20	1,212787	2,044193	0,686187	1,15659	25,06	27,56	22,56	29,48	20,65
21	1,26327	2,129284	0,771915	1,301088	26,06	28,87	23,25	30,66	21,46
22	1,318206	2,221881	0,858876	1,447664	27,06	30,19	23,93	31,86	22,26
23	1,377062	2,321084	0,946731	1,595747	28,06	31,51	24,61	33,07	23,05
24	1,439357	2,426084	1,035253	1,744954	29,06	32,83	25,29	34,30	23,82
25	1,504664	2,536161	1,124284	1,895017	30,06	34,15	25,97	35,54	24,58
26	1,572607	2,650682	1,213711	2,04575	31,06	35,48	26,64	36,78	25,33
27	1,642861	2,769096	1,303454	2,197014	32,06	36,80	27,31	38,04	26,08

	Se ---->			1,685533					
	1 vez		várias vezes		dados	várias vezes		uma vez	
nº obser.		Sy		Sy'	regressão	superior	inferior	superior	inferior
28	1,71514	2,890925	1,39345	2,348707	33,06	38,13	27,99	39,30	26,81
29	1,789199	3,015755	1,483655	2,500751	34,06	39,46	28,66	40,57	27,54
30	1,864827	3,143228	1,574033	2,653084	35,06	40,79	29,33	41,85	28,27
31	1,941839	3,273035	1,664554	2,805661	36,06	42,12	30,00	43,13	28,99

Nota:

Realce-se que a banda para um único valor de Y é mais larga do que quando se tem o valor da média da população de todos os valores de Y.

O que conduz ao seguinte gráfico



O Excel disponibiliza um aplicativo FERAMENTA – ANÁLISE DE DADOS – REGRESSÃO que tem resultados similares aos obtidos anteriormente. A utilização desta ferramenta possibilita que os utilizadores não tenham que efetuar os cálculos atrás descritos. Assim, retomando os dados anteriores tem-se:

SUMÁRIO DOS RESULTADOS

Estatística de regressão	
R múltiplo	0,873409
Quadrado de R	0,762843
Quadrado de R ajustado	0,7446
Erro-padrão	1,685533
Observações	15

ANOVA

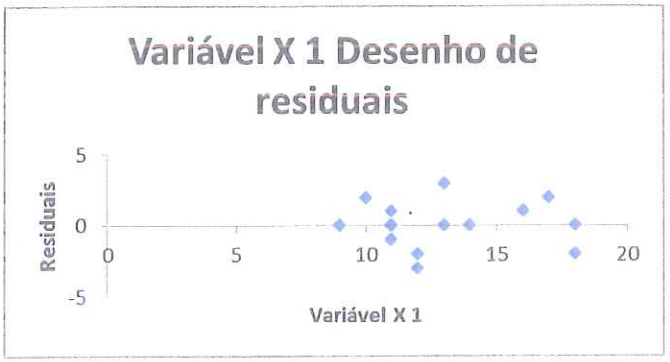
	gl	*SQ*	*MQ*	*F*	*F de significância*
Regressão	1	118,8	118,8	41,81594	2,11E-05
Residual	13	36,9333	2,841023		
Total	14	155,7333			

	Coeficientes	*Erro-padrão*	*Stat t*	*valor P*	*95% inferior*	*95% superior*
Interceptar	5,073991	2,065891	2,456079	0,028881	0,610906	9,537076
Variável X 1	0,999439	0,154556	6,466525	2,11E-05	0,665542	1,333337

RESULTADO RESIDUAL

Observação	*Y previsto*	*Resíduais*
1	16,06783	-0,06783
2	18,0667	-0,0667
3	14,06895	-0,06895
4	23,0639	-2,0639
5	16,06783	0,932175
6	21,06502	0,934978
7	16,06783	-1,06783
8	18,0667	2,933296
9	23,0639	-0,0639
10	15,06839	1,931614
11	17,06726	-2,06726
12	19,06614	-0,06614
13	22,06446	1,935538
14	16,06783	-0,06783
15	17,06726	-3,06726

E o quadro do resíduos



Tentou-se assim dar uma panorâmica da aplicação de uma ferramenta que pode ser útil na previsão. Este exemplo pode ser extensivo a várias outras situações.

Bibliografia

No mercado existe um conjunto alargado de material sobre estatística e métodos econométricos. Para o efeito indicam-se algumas obras que poderão ser utlizadas.

Caiado, J. (2011) *Métodos de Previsão em Gestão*, Edições Sílabo

Levine,D.; Stephen, D. et al (2005) *Estatística – Teoria e Aplicações*, LTC

Guimarães, R e Cabra, J. (2007) Estatística, 2º edição, McGrawHil

Pedrosa, A e Gama, S (2004) *Probabilidade e Estatística*, Porto Editora

Pinto, J. e Curto, J. (2010) Estatística para Economia e Gestão, 2ª edição, Edições Sílabo

Webster, A. (2006) *Estatística aplicada à Administração e Economia*, 3ª edição, McGrawHil