REGRESSÃO MÚLTIPLA:

uma digressão sobre seus usos.

Autores: **Istvan Karoly Kasznar, *PhD***

Professor Titular da FGV

e Presidente da IBCI

Bento Mario Lages Gonçalves, *MSc*
Consultor Senior da IBCI

IBCI – Institutional Business Consultoria Internacional
Rua da Alfândega, 25, sala 805/806, Centro, Rio de Janeiro, RJ.
Telefax: (21)2263-7017 / 2233-8552 / 2263-0563
e-mail: ibci@ibci.com.br
Home Page: www.ibci.com.br

REGRESSÃO MÚLTIPLA

1- Introdução

A Regressão Múltipla é um dos inúmeros modelos estatísticos explanatórios causais referentes ao tratamento de séries temporais de dados. Sua base estatística advém da Regressão Linear, que se restringe a duas variáveis e a apenas uma equação funcional do primeiro grau ($Y = a + bX$) de ajustamento.

A análise de Regressão Múltipla é uma metodologia estatística de previsão de valores de uma ou mais variáveis de *resposta* (Dependentes) através de um conjunto de variáveis *explicativas* (Independentes)*. Esta metodologia pode ser utilizada também para a avaliação dos efeitos das variáveis explicativas como *previsoras* das variáveis de resposta; isto é, serve para contribuir na obtenção de respostas a perguntas do tipo "Qual é o melhor estimador para ... ?".

Sua aplicação é especialmente importante pois permite que se estime o valor de uma variável com base num conjunto de outras variáveis. Quanto mais significativo for o peso de uma variável isolada, ou de um conjunto de variáveis explicativas, tanto mais se poderá afirmar que alguns fatores afetam mais o comportamento de uma variável de resposta especificamente procurada, do que outros.

Lamentavelmente, o termo *regressão*, cunhado do título do primeiro documento (*paper)* escrito sobre o assunto, e que é de autoria de F. Galton [7], foi desenvolvido a posteriori por Bowerman e O'Connel [3], Neter e Wasserman [13], Draper e Smith [5], Seber [14], e Goldberger [8] que estenderam a sua aplicabilidade e desenvolveram as hipóteses passíveis de regressão múltipla para inúmeras situações diferenciadas.

O formato geral da equação de Regressão Linear Múltipla é :

$$Y = a + b_1X_1 + b_2X_2 + ... + b_kX_k$$

IBCI – Institutional Business Consultoria Internacional
Rua da Alfândega, 25, sala 805/806, Centro, Rio de Janeiro, RJ.
Telefax: (21)2263-7017 / 2233-8552 / 2263-0563
e-mail: ibci@ibci.com.br
Home Page: www.ibci.com.br

Onde :

Y é a Variável Dependente;

a corresponde a um coeficiente técnico fixo, a um valor de base a partir do qual começa *Y*;

b_k corresponde aos coeficientes técnicos atrelados às Variáveis Independentes; e

e X_k as Variáveis Independentes.

As instituições financeiras procuram explicar a evolução dos seus Depósitos Totais a partir da evolução de agregados macroeconômicos como o Produto Interno Bruto – PIB, a População e a Renda *per capita* . A *Tabela A.1* a seguir, apresenta a evolução de tais indicadores no Brasil, ao longo do período de 1970 a 1995 :

Tabela A.1 – Evolução dos Depósitos Totais de Instituição Financeira, PIB a custo de Fatores, População e Renda per capita do Brasil no Período 1970/1995.

Ano	Depósitos Totais (Em US$ Milhões)	PIB (Em US$ Milhões)	População (Nº de Habitantes)	Renda per capita (Em US$/Hab)
1970	312,0	33.027	93.139.037	355
1975	381,5	105.962	105.279.615	1.006
1980	347,4	191.842	119.002.706	1.961
1981	404,2	212.187	121.304.828	1.749
1982	402,1	222.354	124.132.901	1.791
1983	452,0	223.354	126.932.107	1.760
1984	431,7	245.104	129.881.714	1.887
1985	582,3	273.949	130.964.997	2.092
1986	596,6	303.496	132.744.121	2.286
1987	620,8	323.736	135.682.832	2.386
1988	513,6	335.923	138.506.432	2.425
1989	606,9	362.286	141.596.301	2.559
1990	629,0	361.909	146.917.459	2.463
1991	602,7	376.089	147.489.931	2.550
1992	656,7	379.411	150.474.909	2.521
1993	678,5	384.591	153.390.844	2.507
1994	637,6	395.478	155.608.189	2.541
1995	698,2	480.361	158.617.875	3.028

Matematicamente, o relacionamento de tais variáveis pode ser descrito por :

Evolução dos Depósitos Totais = *f* (PIB, População, Renda *per capita*)

Esta equação simplesmente diz que a evolução dos Depósitos para a instituição financeira é uma *função* de, ou depende de três variáveis independentes – PIB, População e Renda *per capita*. Como vimos a equação que exprime as relações lineares e aditivas entre estas variáveis é :

$$Y = a + b_1X_1 + b_2X_2 + b_3X_3$$

Onde :

Y = Evolução dos Depósitos Totais.

a = Coeficiente técnico fixo

b_1 = Coeficiente técnico da variável PIB.

X_1 = PIB.

b_2 = Coeficiente técnico da variável População.

X_2 = População.

b_3 = Coeficiente técnico da variável Renda *per capita*.

X_3 = Renda *per capita*.

2 – Aplicando a Regressão Múltipla

Para uma melhor compreensão dos conceitos de Regressão Múltipla, a metodologia dos mínimos quadrados será utilizada para obtermos os valores de a, b_1, b_2 e b_3 na equação de regressão. Será assumido que a tarefa principal é prever o comportamento dos Depósitos Totais da Instituição Financeira nos próximos cinco anos (1996 a 2.000), e tais previsões são baseadas no comportamento do PIB (b_1), População (b_2) e Renda *per capita*(b_3). Usando as observações históricas da *Tabela A.1* , deveremos determinar os valores de a, b_1, b_2 e b_3 de forma a minimizar *o Erro Quadrado Médio* da curva de regressão, e então utilizar os estimadores de a, b_1, b_2 e b_3 de forma a montar uma previsão do comportamento dos Depósitos Totais.

Usando os dados históricos, de 1970 a 1995 e uma rotina de Regressão Múltipla computadorizada – como a que se encontra no *STATISTICA Release 5 (1997)*,, os resultados obtidos são :

Tabela A.2 – Resultados da Regressão Múltipla, Análise da Variância, e Matriz de Correlação obtidos a partir dos dados constantes da Tabela A.1

RESULTADOS DA REGRESSÃO MÚLTIPLA		
Var.Dependente : D.Totais	R Múltiplo : 0,93975482	F = 35,26686
	R^2 : 0,88313912	DF = 3,14
Nº de Casos : 18	R^2 Ajustado : 0,85809750	P = 0,000001
	Êrro Padrão da Estimativa : 47,076164896	
Intercepto (a) : 447,01798366	Êrro Padrão : 363,8734	t (14) = 1,2285
		P < 0,2395
b_1 (PIB) = 1,87	b_2 (POP) = -0,22	b_3 (Rpc) = -0,75

ANÁLISE DA VARIÂNCIA – ANOVA					
	Soma dos Quadrados	DF	Média dos Quadrados	F	Nível p
Regressão	234.471,5	3	78.157,18	35,26686	0,00001
Resíduos	31.026,3	14	2.216,17		
Total	265.497,9				

MATRIZ DE CORRELAÇÃO							
	b_K	Correlação Parcial	Correlação Semi-Parcial	Tolerância	R^2	t (14)	Nível p
PIB	1,865174	0,562719	0,232705	0,015566	0,984434	2,54704	0,023251
POP	-0,219347	-0,105515	-0,036273	0,027346	0,972654	-0,39702	0,697341
Rpc	-0,746266	-0,450537	-0,172517	0,053441	0,946559	-1,88826	0,079891

CORRELAÇÃO DOS COEFICIENTES DE REGRESSÃO			
	PIB	POP	Rpc
PIB	1,000000	-0,848786	-0,673549
POP	-0,848786	1,000000	0,200493
Rpc	-0,673549	0,200493	1,000000

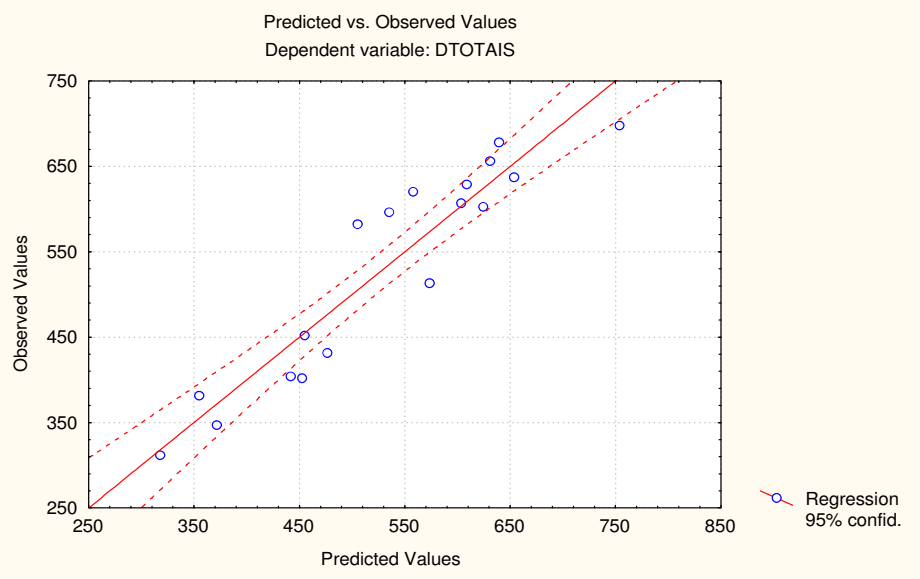
Assim, a equação de previsão para os Depósitos Totais de uma instituição financeira fictícia pode ser definida como :

$$DT = 447,01 + 1,87\ X_1 - 0,22\ X_2 - 0,75\ X_3 \qquad \text{(IV.I)}$$

Esta equação revela simplesmente que baseado nas observações históricas o melhor modelo de previsão é a equação IV.1, exposta anteriormente. Contudo, um problema escalar ocorre com as variáveis selecionadas, uma vez que a Variável Dependente Depósitos Totais (DT) é expressa em milhões de dólares, e as Variáveis Independentes são expressas em medidas escalares diferentes - o PIB a custo de fatores é expresso em milhões de dólares, a População em valores absolutos e a Renda *per capita*, uma razão entre o PIB e a População, é expressa em Dólares/Habitante. Assim, é incorreto interpretar que a variável PIB é o melhor estimador para os Depósitos Totais simplesmente porque esta apresenta o maior coeficiente técnico (*1,87*). Se a População fosse expressa, por exemplo, em milhões de habitantes, a nova unidade escalar poderia alterar o coeficiente b_1, e tornar o coeficiente b_2 mais "atraente" (maior que b_1).

A interpretação literal da equação IV.I é que quando X_1 , X_2 e X_3 se igualam a zero, os Depósitos Totais da instituição financeira alcançarão a cifra de *US$ 447,01 milhões* (o valor do *intercepto a*); e que quando os Depósitos Totais variarem em *US$ 1 milhão*, o PIB sofrerá uma variação de *US$ 1,87 milhões* (mantendo-se as outras variáveis constantes). Quando a População decrescer em *1 milhão de habitantes* os Depósitos Totais sofrerão um acréscimo de *US$ 200 mil* (de novo, mantendo-se as outras variáveis constantes). Assim, os coeficientes determinados pela Regressão Múltipla simplesmente indicam como alterações unitárias em cada Variável Independente podem influenciar o valor da Variável Dependente, Y .

Uma vez determinados os parâmetros da equação, esta pode ser utilizada para prever os Depósitos Totais da instituição financeira para cada um dos próximos cinco anos. Esta previsão é feita através da substituição simples dos valores de X_1 , X_2 e X_3 na equação IV.I, e os valores encontrados podem ser plotados, conforme o gráfico de dispersão a seguir.



Para introduzirmos uniformidade à série de Variáveis Independentes e a Variável Dependente, devemos homogeneizá-las através do cálculo de suas variações percentuais ao longo do tempo, o que altera a função de relacionamento entre as variáveis para :

$$\Delta\%DT = f\,(\Delta\%PIB,\ \Delta\%População,\ \Delta\%Renda\ per\ capita) \qquad \text{(IV.II)}$$

ou seja, a Variação Percentual dos Depósitos Totais ($\Delta\%DT$) é função da Variação Percentual do PIB (ΔPIB), da Variação Percentual da População ($\Delta\%Pop$) e da Variação Percentual da Renda *per capita* ($\Delta\%Rpc$).

Tabela A.3 – Variação Percentual dos Depósitos Totais de Instituição Financeira, Variação Percentual do PIB a custo de Fatores, Variação Percentual da População e Variação Percentual da Renda per capita do Brasil no Período 1970/1995.

Ano	Δ%DT	Δ%PIB	Δ%Pop	Δ%Rpc
1970	0,00	0,00	0,00	0,00
1975	22,28	220,83	13,03	183,38
1980	-8,94	81,05	13,03	94,93
1981	16,35	10,61	1,93	-10,81
1982	-0,52	4,79	2,33	2,40
1983	12,41	0,45	2,26	-1,73
1984	-4,49	9,74	2,32	7,22
1985	34,89	11,77	0,83	10,86
1986	2,46	10,79	1,36	9,27
1987	4,06	6,67	2,21	4,37
1988	-17,27	3.76	2,08	1,63
1989	18,17	7,85	2,23	5,53
1990	3,64	-0,10	3,76	-3,75
1991	-4,18	3,92	0,39	3,53
1992	8,96	0,88	2,02	-1,14
1993	3,32	1,37	1,94	-0,56
1994	-6,03	2,83	1,45	1,36
1995	9,50	21,46	1,93	19,17

Resolvido o problema de diferenças de escala, através da *Tabela A.3*, podemos buscar quais ou quais Variáveis Independentes explicam melhor a Variável Dependente. Se somente uma variável independente, X_1, X_2 ou X_3, explicar plenamente a evolução dos depósitos desta instituição o problema toma a conotação de uma Regressão Linear Simples.

Desse modo, usando os dados históricos de 1970 a 1995 da *Tabela A.3*, e a mesma rotina de Regressão Múltipla computadorizada do *STATISTICA Release 5 (1997)*, os resultados obtidos são :

Tabela A.4 – Resultados da Regressão Múltipla, Análise da Variância, e Matriz de Correlação obtidos a partir dos dados constantes da Tabela A.3

RESULTADOS DA REGRESSÃO MÚLTIPLA		
Var.Dependente : Δ%DT	R Múltiplo : 0,49037341	F = 1,477452
	R^2 : 0,24046608	DF = 3,14
Nº de Casos : 18	R^2 Ajustado : 0,7770881	P = 0,263508
	Êrro Padrão da Estimativa : 11,962482513	
Intercepto (a) : 10,940557845	Êrro Padrão : 4,682491	T (14) = 2,3365
		P < 0,0348
b_1 (Δ%PIB) = -0,68	b_2 (Δ%POP) = -0,91	b_3 (Δ%Rpc) = 1,63

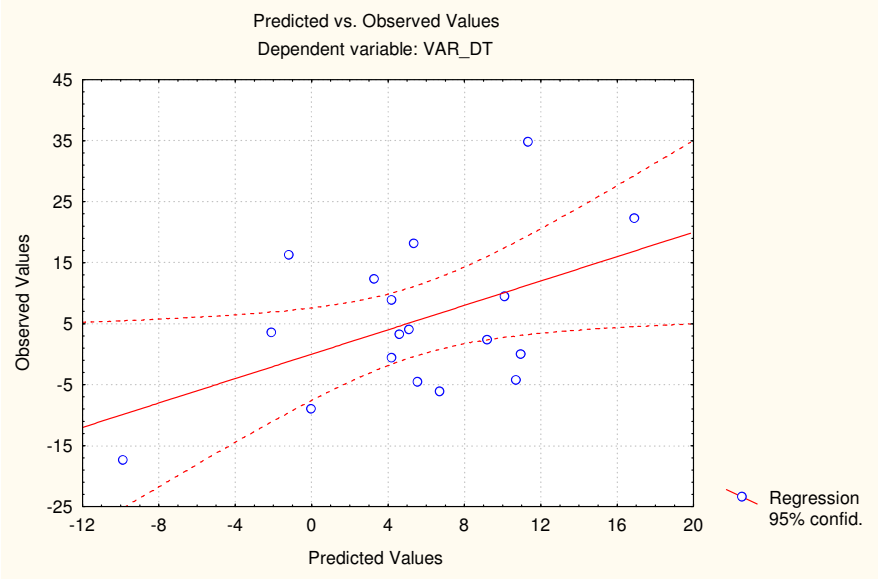
ANÁLISE DA VARIÂNCIA – ANOVA					
	Soma dos Quadrados	DF	Média dos Quadrados	F	Nível p
Regressão	634,275	3	211,4249	1,477452	0,263508
Resíduos	2.003,414	14	143,1010		
Total	2.637,688				

MATRIZ DE CORRELAÇÃO							
	b_K	Correlação Parcial	Correlação Semi-Parcial	Tolerância	R^2	t (14)	Nível p
Δ%PIB	-0,680086	-0,319313	-0,293659	0,186449	0,813551	-1,26076	0,228001
Δ%POP	-0,908551	-0,398694	-0,378882	0,173904	0,826096	-1,62665	0,126101
Δ%Rpc	1,628957	0,485842	0,484434	0,088440	0,911560	-2,07982	0,056397

CORRELAÇÃO DOS COEFICIENTES DE REGRESSÃO			
	Δ%PIB	Δ%POP	Δ%Rpc
Δ%PIB	1,000000	0,135743	-0,707681
Δ%POP	0,135743	1,000000	-0,731026
Δ%Rpc	-0,707681	-0,731026	1,000000

IBCI – Institutional Business Consultoria Internacional
Rua da Alfândega, 25, sala 805/806, Centro, Rio de Janeiro, RJ.
Telefax: (21)2263-7017 / 2233-8552 / 2263-0563
e-mail: ibci@ibci.com.br
Home Page: www.ibci.com.br

E os valores observados e previstos são plotados à seguir :



3 – Correlação Múltipla e Coeficiente de Determinação

É bastante comum numa Regressão Linear que a Variável Dependente (*Y*) se relacione com a Variável Independente (*X*), mas é incorreto afirmar que o valor da primeira depende *em causa e efeito* das alterações no valor da segunda. Neste caso a inter-relação entre as variáveis é demonstrada através da correlação. O coeficiente de correlação, *r* , é a medida de inter-relação entre a Variável Dependente e a Variável Independente. Ele pode variar de 0 (que indica ausência de correlação) a ± 1 (que indica correlação perfeita). Quando o coeficiente de correlação é maior que 0, as duas variáveis são positivamente correlacionadas, em contrapartida quando é menor que 0, as duas variáveis são negativamente correlacionadas. O sinal do coeficiente de correlação numa Regressão Linear é sempre o mesmo sinal do coeficiente de regressão, *b*.

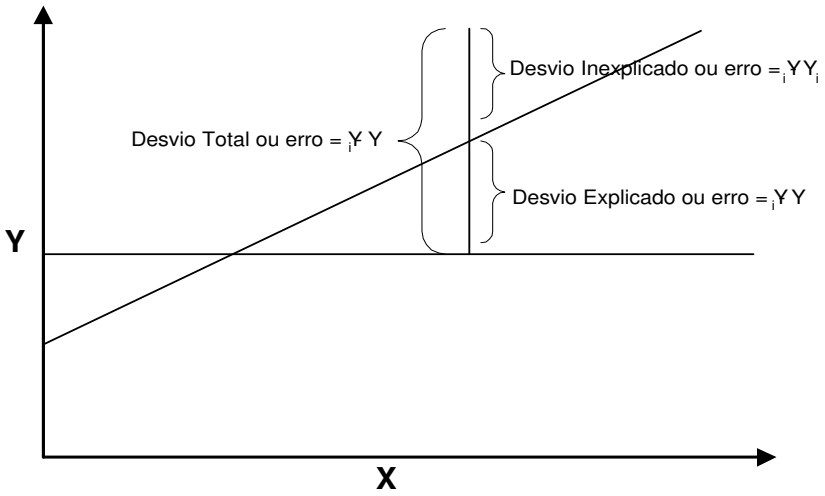
O coeficiente de correlação, *r*, é calculado a partir da seguinte fórmula (aonde *n* é o Número de Observações (pontos) necessários para o ajuste da linha de regressão) :

$$r = \frac{n\sum XY - \sum X \sum Y}{\sqrt{\left(n\sum X^2 - \left(\sum X\right)^2\right)\left(n\sum Y^2 - \left(\sum Y\right)^2\right)}}$$

Dessa expressão, deduz-se o coeficiente de determinação (r^2) que nada mais é que o quociente da variação explicada sobre a variação total; ou seja :

$$r^2 = \frac{\sum (\overleftrightarrow{Y}_i - \bar{y})^2}{\sum (y_i - \bar{y})^2} = \frac{\left(n\sum XY - \sum X \sum Y\right)^2}{\left[n\sum X^2 - \left(\sum X\right)^2\right]\left[n\sum Y^2\right]}$$

Assim, r^2 é a medida de quão bem as observações se ajustam ao longo da linha de regressão. Para ajuste dos pontos Y_i , teríamos graficamente :



No caso de uma Regressão Múltipla o coeficiente de determinação deve computar também o quociente entre a variação explicada e a variação total, porém para todas as Variáveis Independentes. Este coeficiente de determinação,

identificado por R^2, pode assumir valores entre 0 e 1, sendo o último o que representa a situação onde toda a variação é explicada. A equação utilizada para o cálculo do coeficiente de determinação de uma Regressão Múltipla é a seguinte :

$$R^2 = \frac{\sum (\overleftrightarrow{Y}_i - \bar{Y})^2}{\sum (Y_i - \bar{Y})^2}$$

Onde Y_i são as observações esperadas e $\bar{Y}$ a média das observações.

No caso do exemplo da *Tabela A.2* o coeficiente de determinação encontrado foi de *0,88313912* - isto significa que 88,31% da variação dos Depósitos Totais da instituição financeira são explicados pela variação combinada do PIB, População e Renda *per capita*. No exemplo da *Tabela A.4* o coeficiente de determinação encontrado foi de *0,24046608* – o que significa que 24,04% da variação percentual dos Depósitos Totais da instituição financeira são explicados pela variação percentual combinada do PIB, População e Renda *per capita*.o que confere ao segundo conjunto de variáveis, um grau explanatório bastante inferior.

A Matriz de Correlação possui grande significado informativo para a Regressão Múltipla porque estabelece como os pares de variáveis Dependentes(*Y*) e Independentes (X1, X_2,..., X_k)se correlacionam. Esta informação é utilizada na seleção das variáveis que deverão fazer parte da equação de regressão - ou seja, variáveis com correlação elevada e positiva deverão ser incluídas no modelo proposto, enquanto que as variáveis na condição inversa deverão ser descartadas. Este será um dos critérios estatísticos para a seleção de variáveis no modelo do PCPA. A Correlação Múltipla e o Coeficiente de Determinação (R^2) também indicam como a relação expressa através da equação de regressão explica as variações da Variável Dependente (*Y*).

4 – Testes de Significância Estatística

A significância estatística dos resultados obtidos na Análise de Regressão deve ser estabelecida antes do uso de tais resultados numa previsão. A determinação dos coeficientes técnicos (b_1, b_2,...,b_k) é baseada simplesmente nas observações históricas. O propósito dos testes de significância estatística é determinar a confiança que pode ser depositada nos resultados da regressão e a sua aplicabilidade na população de valores possíveis.

Apesar da existência de inúmeros testes de significância estatística, somente dois dos principais testes serão abordados :

1. O *teste F* ou *estatística F* indica se a equação de regressão é significativa – ou seja, se a relação funcional estabelecida entre a Variável Dependente e os efeitos combinados das Variáveis Independentes são relevantes. O valor do *teste F* é determinado pelo quociente entre a *variância explicada* e a *variância inexplicada*. Esta relação pode ser expressa matematicamente de duas formas equivalentes :

$$F = \frac{\sum \left(\overleftrightarrow{Y}_i - \overline{Y}\right)^2 /(k-1)}{\sum \left(Y_i - \overleftrightarrow{Y}\right)^2 /(n-k)}$$

$$F = \frac{\dfrac{R^2}{k-1}}{\dfrac{1-R^2}{n-k}}$$

ou :

onde R^2 é o coeficiente de determinação.

No caso do exemplo da *Tabela A.2* a *estatística F* encontrada foi de 35,26686 - isto significa que num intervalo de confiança de 95% a equação de regressão tem um nível de significância de 35,26%, o que pode ser considerado baixo. No exemplo da *Tabela A.4* a *estatística F* encontrada foi de 1,477452 o que confere a equação de regressão praticamente nenhuma significância. O resultado do *teste F* não deve ser considerado de forma isolada - isto é, somente seus resultados não devem descartar totalmente uma equação de regressão, uma vez que os coeficientes da regressão podem apresentar correlação significativa.

2. O segundo teste determina a significância (correlação) dos coeficientes da equação de regressão (a, b_1, b_2,...,b_k) individualmente. O questionamento essencial deste teste é se o valor atribuído a cada coeficiente é significativamente diferente de 0 ou se tal valor ocorreu simplesmente ao acaso.

Este teste consiste em calcular a variância de cada coeficiente da regressão e, através de sua raiz quadrada, estabelecer o erro padrão, o que determina se o valor de cada coeficiente é significativamente diferente de 0.

O *teste t* para o exemplo da *Tabela A.2* foi de *1,2285* para *a*, *2,54704* para b_1, *-0,39702* para b_2 e *–1,88826* para b_3 - isto significa que num intervalo de confiança de 95% a equação de regressão tem somente uma Variável Independente, o PIB, positivamente correlacionada, estando as duas outras variáveis negativamente correlacionadas entre si e com a Variável Dependente (Depósitos Totais). Tal fato não implica no simples descarte das variáveis negativamente correlacionadas, mas confere a estas um baixo nível significância.

O *teste t* do intercepto *a* significa que o mesmo encontra-se num nível bem diferente de 0, o que lhe confere significância na equação de regressão, ou seja, numa previsão com os dados históricos apresentados, *a* não deve ser desprezado. No exemplo da *Tabela A.4* o *teste t* foi de *2,3365* para *a*, *-1,26076* para b_1, *-1,62665* para b_2 e *–2,07982* para b_3 - isto significa que num intervalo de confiança de 95%, a equação de regressão tem todas as variáveis negativamente correlacionadas entre si e com a Variável Dependente e diferente de 0, o que torna o redimensionamento da regressão necessário.

Além dos testes de significância estatística, também podem ser construídos em torno da equação de regressão intervalos de confiança. Estes intervalos são baseados no desvio padrão da regressão, traduzindo-se num maior nível de confiança no modelo de regressão.

5 – Os Pressupostos da Análise de Regressão

Para obtenção dos resultados, a análise de regressão baseia-se em quatro pressupostos básicos :

5.1 – Linearidade

Apesar de parecer um pressuposto restritivo matematicamente toda função não-linear pode ser transformada numa função linear através de técnicas logarítmicas, polinomiais e de relações recíprocas. Não nos cabe neste texto discutir as formulações matemáticas de transformação, porém a sua existência é de fundamental importância uma vez que a análise de regressão não pode ser aplicada se a função não puder ser transformada para a forma linear.

5.2 – Independência dos resíduos

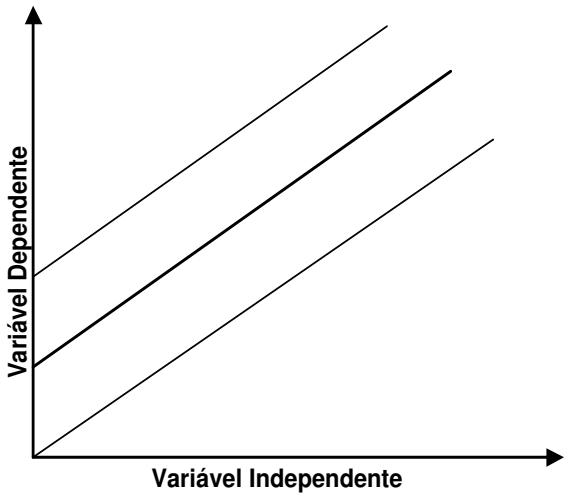
A violação do pressuposto da independência dos resíduos implica na existência de forte correlação (autocorrelação) entre os residuais sucessivos. Isto é, e_t não é independente de e_{t-1}..., e_{t-i+1} ..., e_{t+1}, e_{t+2},..., e_{t+n} . A falta de independência não afeta o valor dos parâmetros estimados, mas afeta diretamente as variâncias estimadas. A falta de independência dos resíduos implica em R^2 e *estatística F* elevados e *teste t* reduzido se a autocorrelação é positiva e todos os testes com resultados elevados se a autocorrelação for negativa.

(A autocorrelação pode ser resultante de 1) especificação incorreta (a inclusão de um número não ótimo de variáveis), o que causa *dependência* entre as Variáveis Independentes, (ou 2) forma funcional incorreta (deficiência de modelagem), (ou 3) forte tendência entre as variáveis. A autocorrelação pode ser visualizada através da plotagem dos resíduos, no entanto, existe um teste estatístico, o *de Durbin-Watson* (*teste D-W*) que pode ser utilizado para verificar a existência de autocorrelação.

5.3 – Homocedasticidade

Se os resíduos não estão distribuídos ao longo da linha de regressão em torno de todo o intervalo de observações, o pressuposto da variância constante, ou homocedasticidade, é violado.

O gráfico a seguir ilustra o significado da variância constante dos resíduos:



IBCI – Institutional Business Consultoria Internacional
Rua da Alfândega, 25, sala 805/806, Centro, Rio de Janeiro, RJ.
Telefax: (21)2263-7017 / 2233-8552 / 2263-0563
e-mail: ibci@ibci.com.br
Home Page: www.ibci.com.br

A ocorrência de variâncias não constantes nos resíduos é chamada de heterocedasticidade. Sua ocorrência pode estar condicionada a especificações incorretas no modelo de regressão, e sua detecção é possível através do estudo residual dos erros.

A teste Durbin-Watson pode indicar heterodasticidade e sua correção esta vinculada à eliminação de algumas variáveis ou a transformação matemática do modelo, trazendo uniformidade dos erros percentuais ao longo da linha de regressão.

5.4 – Normalidade dos resíduos

Esta hipótese também apresenta características pouco restritivas uma vez que os resíduos são resultantes de um sem número de fatores menos importantes no que tange a influência no comportamento da variável dependente (senão deveriam ser incluídos na equação de regressão, perdendo sua característica residual). Na média, sua influência pode ser desprezada, uma vez que o erro médio apresenta um comportamento *"normalizado"*.

Estatisticamente se possuímos um número de observações superior a 30 a previsão de dados assume a *"normalidade"*. Isto porque a distribuição amostral dos estimadores pode ser aproximada a curva normal onde *n* possua amplitude suficiente, o que na maior parte ocorre quando *n* é igual a 30. O *Teorema do Limite Central* da estatística permite esta aproximação e torna possível o uso da curva normal na avaliação da dispersão dos dados, inclusive dos resíduos, da amostra em torno do parâmetro central (média). Assim ao calcularmos sua média e variância, a extensão de possíveis erros pode ser avaliada; o que introduz um intervalo de confiança de 30 observações para a variância.

Quando o pressuposto da normalidade dos resíduos é questionado, não existem testes estatísticos específicos para sua avaliação; todavia os resíduos podem ser plotados com vistas a detecção de sua distribuição próxima a normal e o seu intervalo de variação (o maior menos o menor valor) pode ser medido com vistas a determinação de sua dispersão (se próxima a 6.0 é considerado dentro da distribuição normal).

6 - Multicolinearidade

A multicolinearidade é um problema computacional que se desenvolve quando duas ou mais variáveis independentes possuem forte correlação. O resultado é uma Matriz de Correlação com variabilidade única (próximo de zero) tendo em vista o efeito da divisão de um número por uma variação absoluta

extremamente pequena (próxima de zero). O resultado desta divisão é um número com um número de casas decimais bastante elevado, o que torna a aproximação computacional totalmente ineficaz. A existência de multicolinearidade introduz erros grosseiros no resultado da regressão, produzindo sérios erros na previsão da Variável Dependente. Felizmente a multicolinearidade é de fácil detecção e correção.

A ocorrência de multicolinearidade se dá quando um ou mais dos *testes t* assume resultado(s) muito pequenos (não significativos) e os valores de R^2 ou da *estatística F* são muito grandes. Se este for o caso, a Matriz de Correlação deve ser avaliada e, possivelmente, uma das Variáveis Independentes com forte correlação ser eliminada. (Como regra geral, um coeficiente de correlação superior ou próximo a *0,7* entre duas Variáveis Independentes indica problemas de multicolinearidade).

A multicolinearidade também pode ser detectada se a percentagem da variação explicada por alguma Variável Independente for negativa. Esta condição é verificada nos dois exemplos de regressão desenvolvidos nas *Tabelas A.2* e *A.4* aonde se observam correlações negativas entre as Variáveis Independentes e um nível de R^2 bastante elevado em ambos os casos (o menor valor de R^2 encontrado foi de *0,813551*). Tal fato, se deve a inclusão de uma Variável Independente (a Renda *per capita*) que é resultante do quociente entre as duas outras Variáveis Independentes (o PIB e a População), o que introduz a problemática da multicolinearidade de forma primária, ou seja, na própria formulação do modelo de regressão. Assim, uma das variáveis deve ser excluída do modelo de forma a buscar um conjunto de variáveis de maior valor explanatório para as variações nos Depósitos Totais da instituição financeira.

A multicolinearidade é um problema freqüentemente encontrado nos dados econômicos e de negócios tendo em vista a elevada correlação do tempo entre diferentes agregados como a população,a população economicamente ativa, o PIB, o nível de renda disponível para consumo, vendas, estoques, custos, lucros, etc. A problemática da multicolinearidade em tais casos não deve ser desprezada, uma vez que a elevada correlação existente entre as mesmas pode prejudicar a sua utilização e, consequentemente, a modelagem.

Basicamente podemos encontrar entre os diversos agregados econômicos Variáveis Independentes com elevada correlação entre sí, porém de valor limitado para a regressão múltipla. O que distingue um dado conjunto de Variáveis Independentes como um bom conjunto de estimadores para uma determinada Variável Dependente é o seu conteúdo informacional, ou seja, o seu "valor explicativo". Ë claro que do ponto de vista estatístico as Variáveis Independentes mais "necessárias" são aquelas com um nível médio de correlação, o que torna a sua obtenção condicionada a diversas "rodadas computacionais" com vistas a depuração do modelo.

7 – Bibliografia

1. Anderson, T.W., *An Introduction to Multivariate Statistical Methods* (Second Edition), New York : John Wiley & Sons, 1984.

2. Belsley, D.A., E.Kuh, and R.E.Welsh, *Regression Diagnostics*, New York : John Wiley & Sons, 1980.

3. Bowerman, B.L., and R.T.O'Connel, *Linear Statistical Models : An Applied Aproach* (Second Edition), Boston : PWS-Kent, 1990.

4. Chatterjee, S. and B.Price, *Regression Analysis By Example*, New York : John Wiley & Sons, 1977.

5. Draper, N.R., & H.Smith, *Applied Regression Analysis*, (Second Edition), New York : John Wiley & Sons, 1981.

6. Durbin, J. and G.S.Watson, "Testing for Serial Correlation in Least Squares Regression, II", *Biometrika*, Vol.38 (1951), 159-178.

7. Galton, F., "Regretion Toward Mediocrity in Heredity Stature", *Journal of Anthropological Institute*, Vol.15 (1885), 246-263.

8. Goldberger, A.S., *Econometric Theory*, New York : John Wiley & Sons, 1964.

9. Johnson, R.A. & Wichern, D.W., *Applied Multivariate Statistical Analysis* (Third Edition) , New Jersey : Prentice Hall, 1992.

10. Kasznar, I.K., "Análise da Evolução do Produto Interno Bruto (PIB) e das Dívidas por Estados, 1970-95", Revista de Administração Pública, Vol.30 nº 6 (1996).

11. Kendall, M.G., *Multivariate Analysis*, New York: Hafner Press, 1975.

12. Makidrakis, S. and Wheelright S.C., Forecasting Methods and Applications, New York : John Wiley & Sons, 1978.

13. Neter, J., and W.Wasserman, *Applied Linear Statistical Models*, Homewood III : Richard D.Irwin, 1974.

14. Seber, G.A.F., *Linear Regression Analysis*, New York : John Wiley & Sons, 1977.

15. Sharpe, W.F., *Investments*, New Jersey : Prentice Hall, 1982.

16. Wonnacott T.H. & Wonnacott R.J., *Introdutory Statistics for Business and Economics* (Second Edition), New York : John Wiley & Sons, 1979.