3.7 - INTERVALOS DE CONFIANÇA

Intervalo de confiança para o coeficiente angular β

$$P\left(b - t_{\alpha/2} \cdot \frac{S}{\sqrt{S_{xx}}} \le \beta \le b + t_{\alpha/2} \cdot \frac{S}{\sqrt{S_{xx}}} \right) = 1 - \alpha$$, "t" de Student, com n–2 graus de liberdade.

Intervalo de confiança para o coeficiente linear α

$$P\left(a - t_{\alpha/2}\, S . \left(\frac{1}{n} + \frac{\overline{X}^2}{S_{xx}} \right)^{\frac{1}{2}} \le \alpha \le a + t_{\alpha/2}\, S . \left(\frac{1}{n} + \frac{\overline{X}^2}{S_{xx}} \right)^{\frac{1}{2}} \right) = 1 - \alpha$$, "t" de Student com n–2 graus de liberdade.

Intervalo de confiança para a previsão f(x) ou E [Y/X_i]

Neste caso estamos interessados no I.C. para f(x) ou a média de Y dado um particular valor de X_i, isto é, I.C. para E [Y/X_i].

$$P\left(\hat{y}_i - t_{\alpha/2} . S \left(\frac{1}{n} + \frac{(X_i - \overline{X})^2}{S_{xx}} \right)^{\frac{1}{2}} \le f(x_i) \le \hat{y}_i + t_{\alpha/2} . S \left(\frac{1}{n} + \frac{(X_i - \overline{X})^2}{S_{xx}} \right)^{\frac{1}{2}} \right) = 1 - \alpha$$, "t" de Student, com n–2 graus de liberdade.

Intervalo de confiança para a previsão de um valor individual de Y (Y_i) dado um X_i

$$P\left(\hat{y} - t_{\alpha/2} . S \sqrt{1 + \frac{1}{n} + \frac{(X - \overline{X})^2}{S_{xx}}} \le y \le \hat{y} + t_{\alpha/2} . S \sqrt{1 + \frac{1}{n} + \frac{(X - \overline{X})^2}{S_{xx}}} \right) = 1 - \alpha$$, "t" de Student, com n–2 graus de liberdade.

Fatores que afetam a amplitude dos intervalos de previsão

1°) **TAMANHO DA AMOSTRA (n):** Quanto maior o tamanho de n, menor será o valor da raiz quadrada nas fórmulas dos intervalos de confiança, logo será menor a amplitude do I.C.

2°) **DESVIO PADRÃO (s):** Quanto menor for S, em geral menores serão os intervalos. A magnitude de S é inerente ao problema e não depende do tamanho da amostra.

3°) **DESVIO DE X_i EM RELAÇÃO A $\overline{X}$:** Quanto mais X_i se afastar de $\overline{X}$, maior será o valor de $(X_i - \overline{X})^2$, logo mais amplo será o do intervalo de X_i.

4°) **VARIABILIDADE DOS X_i OBSERVADOS:** Quanto maior a variabilidade dos X_i observados, maior será o valor de S_{xx} e menor o valor da raiz quadrada, logo o intervalo será de amplitude mais estreita.

Exemplo 3.3: Vamos supor que exista uma relação linear entre as variáveis X = despesas com propaganda e Y = vendas de um certo produto. Considerando os dados abaixo, iremos determinar a reta de mínimos quadrados, os testes, os I.C. e o coeficiente de explicação.

X	1,5	5,5	10,0	3,0	7,5	5,0	13,0	4,0	9,0	12,5	15,5
Y	120	190	240	140	180	150	280	110	210	220	310

$\Sigma X = 86,0$ e $\Sigma Y = 2150,0$, X representa milhões de reais e Y milhões de unidades.

	X	**Y**	**XY**	$\mathbf{X^2}$	$\mathbf{Y^2}$
	1,5	120	180	2,25	14400
	5,5	190	1045	30,25	36100
	10,0	240	2400	100,0	57600
	3,0	140	420	9,00	19600
	7,5	180	1350	56,25	32400
	5,0	150	750	25,00	22500
	13,0	280	3640	169,0	78400
	4,0	110	440	16,0	12100
	9,0	210	1890	81,0	44100
	12,5	220	2750	156,25	48400
	15,0	310	4650	225,00	96100
Σ	86,0	2150	19515	870,0	461700

$$\overline{Y} = \frac{\sum y}{n} = \frac{2150}{11} = 195,45 \qquad \overline{X} = \frac{\sum x}{n} = \frac{86}{11} = 7,82$$

$$S_{xy} = \sum xy - \frac{\sum x . \sum y}{n} = 19515 - \frac{86 . 2150}{11} = 2.705,91$$

$$S_{xx} = \sum x^2 - \frac{(\sum x)^2}{n} = 870 - \frac{86^2}{11} = 197,64$$

$$S_{yy} = \sum y^2 - \frac{(\sum y)^2}{n} = 461700 - \frac{2150^2}{11} = 41472,73$$

logo: $b = \frac{S_{xy}}{S_{xx}} = \frac{2705,91}{196,64} = 13,69$

$a = \overline{Y} - b\overline{X}$ ⇨ 195,45 – (13,69) (7,82) = 88,39

$\hat{Y} = a + bx \Rightarrow \hat{Y} = 88{,}39 + 13{,}69X$

Teste t

$H_0: \beta = 0$ $\quad \alpha = 5\%$ $\quad t = \dfrac{b}{\dfrac{S}{\sqrt{S_{xx}}}}$

$H_1: \beta \neq 0$ $\quad gl = n - 2 = 11 - 2 = 9$

$S^2 = \dfrac{S_{yy} - b\,S_{xy}}{n-2} = \dfrac{41.472{,}73 - 13{,}69 \text{ x } 2705{,}91}{9} = 492{,}09$, logo: $\quad S = \sqrt{492{,}09} = 22{,}18$

$t = \dfrac{13{,}69}{\dfrac{22{,}18}{\sqrt{197{,}64}}} = 8{,}71$

Teste F

$VE = b \,.\, S_{xy} = 13{,}69 \,.\, 2705{,}91 = 37.043{,}91$

$VT = S_{yy} = 41.472{,}73$

$VR = VT - VE = 41.472{,}73 - 37.043{,}91 = 4.428{,}82$

Fonte de variação	Soma dos Quadrados	GL	Quadrados Médios	F
Regressão	VE = 37.043,91	1	37.043,91	$F = \dfrac{37043{,}91}{492{,}06} = 75{,}28$
Resíduo	VR = 4.428,82	9	$S^2 = \dfrac{4.428{,}82}{9} = 492{,}06$	
Total	VT = 41.472,73	10		

1°) $H_0: \beta = 0$ $\quad$ 2°) $\alpha = 5\%$ $\quad$ gl = 1 ⇨ numerador

$H_1: \beta \neq 0$ $\quad$ gl = 9 ⇨ denominador

$F_{5\%}\,(1,9) = 5{,}12$; $F_c > F_t$; rejeita-se H_0; existe regressão.

Quanto aos Intervalos de Confiança

1°) **<u>Para β</u>**

$$P\left(b - t_{\alpha/2} \,.\, \frac{S}{\sqrt{S_{xx}}} \leq \beta \leq b + t_{\alpha/2} \,.\, \frac{S}{\sqrt{S_{xx}}}\right) = 1 - \alpha$$

$$P\left(13{,}69 - 2{,}2622 \,.\, \frac{22{,}18}{\sqrt{197{,}84}} \leq \beta \leq 13{,}69 + 2{,}2622 \,.\, \frac{22{,}18}{\sqrt{197{,}84}}\right) = 95\%$$

$P\,(10{,}14 \leq \beta \leq 17{,}24) = 95\%$

2°) **Para α**

$$P\left(a - t_{\alpha/2}\, S.\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)^{\frac{1}{2}} \le \alpha \le a + t_{\alpha/2}\, S.\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)^{\frac{1}{2}}\right) = 1 - \alpha$$

$$P\left(88{,}39 - 2{,}2622\,(22{,}18)\sqrt{\frac{1}{11} + \frac{7{,}82^2}{197{,}64}} \le \alpha \le 88{,}39 + 2{,}2622\,(22{,}18)\sqrt{\frac{1}{11} + \frac{7{,}82^2}{197{,}64}}\right) = 95\%$$

P (56,64 ≤ α ≤ 120,14) = 95%

3°) **Para o valor previsto f (x = 7) = E (Y_i/7)**

Logo: $\hat{Y} = 88{,}39 + 13{,}69.X$

$\hat{Y} = 88{,}39 + 13{,}69\ .\ (7) = 184{,}22$

Considerando ainda nível de 95%, teremos:

$$P\left(\hat{y} - t_{\alpha/2}\,.S\sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}}} \le f(x) \le \hat{y} + t_{\alpha/2}\,.S\sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}}}\right) = 1 - \alpha$$

$$P\left(184{,}22 - 2{,}26\,(22{,}18)\sqrt{\frac{1}{11} + \frac{(7 - 7{,}82)^2}{197{,}64}} \le f(X = 7) \le 184{,}22 + 2{,}26\,(22{,}18)\sqrt{\frac{1}{11} + \frac{(7 - 7{,}82)^2}{197{,}64}}\right)$$

P (168,83 ≤ f (X = 7) ≤ 199,61) = 95%

4°) Neste item, calcularemos o I.C. para Y quando X = 7, ou seja, o I.C. para a venda, quando tivermos investido 7 milhões.

$$P\left(\hat{y} - t_{\alpha/2}\,.S\sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}}} \le y \le \hat{y} + t_{\alpha/2}\,.S\sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}}}\right) = 1 - \alpha$$

$$P\left(184{,}22 - 2{,}26\,(22{,}18)\sqrt{1 + \frac{1}{11} + \frac{(7 - 7{,}82)^2}{197{,}64}} \le y \le 184{,}22 + 2{,}26\,(22{,}18)\sqrt{1 + \frac{1}{11} + \frac{(7 - 7{,}82)^2}{197{,}64}}\right) = 95\%$$

P (131,78 ≤ y ≤ 236,65) = 95%

O coeficiente de explicação

$$R^2 = \frac{b\,S_{xy}}{S_{yy}} \qquad R^2\% = \frac{13{,}69\,(2705{,}91)}{41.472{,}73}.100 = 89\%$$

Este resultado mostra que nosso modelo explica 89% da variação total de Y.

3.8 - REGRESSÃO LINEAR SIMPLES: NOTAÇÃO MATRICIAL

A equação para o modelo de regressão linear é:

$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ $\quad$ i = 1, ..., n (n pares de dados)

Nós podemos escrever o conjunto de observações como:

$$\begin{aligned} Y_1 &= \beta_0 + \beta_1 X_1 + \varepsilon_1 \\ Y_2 &= \beta_0 + \beta_1 X_2 + \varepsilon_2 \\ Y_3 &= \beta_0 + \beta_1 X_3 + \varepsilon_3 \\ &\cdot \quad\quad \cdot \\ &\cdot \quad\quad \cdot \\ &\cdot \quad\quad \cdot \\ Y_n &= \beta_0 + \beta_1 X_n + \varepsilon_n \end{aligned} \qquad (1)$$

Em notação de vetor e matriz, podem ser escritas como:

$$\underline{Y} = \begin{vmatrix} Y_1 \\ Y_2 \\ Y_3 \\ \cdot \\ \cdot \\ \cdot \\ Y_n \end{vmatrix}_{(nx1)} \quad \underline{X} = \begin{vmatrix} 1 & X_1 \\ 1 & X_2 \\ 1 & X_3 \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \\ 1 & X_n \end{vmatrix}_{(nx2)} \quad \underline{\beta} = \begin{vmatrix} \beta_0 \\ \beta_1 \end{vmatrix}_{(2x1)} \quad \underline{\varepsilon} = \begin{vmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \cdot \\ \cdot \\ \cdot \\ \varepsilon_n \end{vmatrix}_{(nx1)}$$

onde: $\underline{Y}$ é o vetor de observações (variável dependente);
$\underline{X}$ é a matriz das observações independentes;
$\underline{\beta}$ é o vetor de parâmetros desconhecidos;
$\underline{\varepsilon}$ é o vetor de erros (aleatórios).

As equações em (1) podem ser escritas como:

$$\underline{Y} = \underline{X}\underline{\beta} + \underline{\varepsilon}$$

$$\begin{vmatrix} Y_1 \\ Y_2 \\ Y_3 \\ \cdot \\ \cdot \\ \cdot \\ Y_n \end{vmatrix} = \begin{vmatrix} 1 & X_1 \\ 1 & X_2 \\ 1 & X_3 \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \\ 1 & X_n \end{vmatrix} \times \begin{vmatrix} \beta_0 \\ \beta_1 \end{vmatrix} + \begin{vmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \cdot \\ \cdot \\ \cdot \\ \varepsilon_n \end{vmatrix}$$

Estimativa dos parâmetros β_0 e β_1

Para determinar as estimativas dos parâmetros β_0 e β_1 aplicamos o método dos mínimos quadrados (MMQ), de modo que seja mínima a soma:

$$Z = \sum (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} \varepsilon_i^2 = \varepsilon'.\varepsilon$$

$$Y = X\beta + \varepsilon \qquad \varepsilon = Y - X\beta \qquad \varepsilon = Y - E(Y)$$

Multiplicando-se $\varepsilon = Y - X\beta$ por ε', nos dois lados da equação, obtemos:

$$\varepsilon'.\varepsilon = (Y' - \beta'X')(Y - X\beta)$$
$$\varepsilon'.\varepsilon = Y'Y - Y'X\beta - \beta'X'Y + \beta'X'X\beta \qquad \text{como } Y'X\beta = \beta'X'Y$$
$$\varepsilon'.\varepsilon = Y'Y - 2\beta'X'Y + \beta'X'X\beta$$

$Z = \varepsilon'.\varepsilon = f(\beta)$ que apresenta um mínimo para o valor β que torne a diferença identicamente nula. Efetuando a derivada, temos:

$$DZ/d\beta = -2d(\beta')X'Y + d(\beta')X'X\beta + d(\beta)\beta'X'X$$

Onde: $d(\beta')X'X\beta + d(\beta)\beta'X'X$ são matrizes (1x1) e uma é a transposta da outra.

$$DZ/d\beta = -2d(\beta')X'Y + 2d(\beta')X'X\beta \cong 0 \text{ ou ainda, } DZ/d\beta = d(\beta')(X'X\beta - X'Y) \cong 0$$

onde

$$\beta = \begin{vmatrix} \beta_0 \\ \\ \beta_1 \end{vmatrix} = \begin{vmatrix} b_0 \\ \\ b_1 \end{vmatrix}$$

Portanto, a diferencial de Z será identicamente nula para $X'Y = X'X\beta$.

Assim: $\beta = (X'X)^{-1} X'Y$

onde:

$$X'X = \begin{vmatrix} n & \sum X_i \\ \\ \sum X_i & \sum X_i^2 \end{vmatrix}_{(2x2)} \qquad X'Y = \begin{vmatrix} \sum Y_i \\ \\ \sum X_i Y_i \end{vmatrix}_{(2x1)}$$

Assim, podemos verificar que:

$$\begin{vmatrix} 1 & 1 & 1 & . & . & . & 1 \\ \\ X_1 & X_2 & X_3 & . & . & . & X_n \end{vmatrix} \begin{vmatrix} 1 & X_1 \\ 1 & X_2 \\ 1 & X_3 \\ . & . \\ . & . \\ 1 & X_n \end{vmatrix} = \begin{vmatrix} n & \Sigma X_i \\ \\ \Sigma X_i & \Sigma X_i^2 \end{vmatrix}$$

Para encontrar a inversa de X'X, primeiro calcula-se o Det (X'X) = ad – bc. Depois, calcula-se:

$$A = \frac{d}{det} \qquad B = \frac{-b}{det} \qquad C = \frac{-c}{det} \qquad D = \frac{a}{det}$$

Assim, a inversa de X'X é dada por:

$$(X'X)^{-1} = \begin{vmatrix} A & B \\ C & D \end{vmatrix}$$

Para encontrar X'Y, faz-se:

$$\begin{vmatrix} 1 & 1 & 1 & . & . & . & 1 \\ X_1 & X_2 & X_3 & . & . & . & X_n \end{vmatrix}_{(2xn)} \begin{vmatrix} Y_1 \\ Y_2 \\ Y_3 \\ . \\ . \\ Y_n \end{vmatrix}_{(nx1)} = \begin{vmatrix} \Sigma Y_i \\ \Sigma X_i Y_i \end{vmatrix}_{(2x1)}$$

Para calcular o vetor de coeficientes, procede-se da seguinte maneira:

$$\underline{\beta} = (X'X)^{-1} X'Y$$

$$\beta = \begin{vmatrix} b_0 \\ b_1 \end{vmatrix} = \frac{1}{n\sum X_i^2 - (\sum X_i)^2} \begin{vmatrix} \sum X_i^2 & -\sum X_i \\ -\sum X_i & n \end{vmatrix} \times \begin{vmatrix} \sum Y_i \\ \sum X_i Y_i \end{vmatrix}$$

$$\beta = \begin{vmatrix} \overline{Y} - b_1 \overline{X} \\ \dfrac{\sum X_i Y_i - (\sum X_i)(\sum Y_i)/n}{\sum X_i^2 - (\sum X_i)^2/n} \end{vmatrix} = \begin{vmatrix} b_0 = \overline{Y} - b_1 \overline{X} \\ b_1 = S_{XY}/S_{XX} \end{vmatrix}$$

3.9 - PRESSUPOSTOS BÁSICOS DA ANÁLISE DE REGRESSÃO

Quando um modelo de regressão é selecionado, deve-se verificar se ele é adequado para os propósitos a que se destina. Uma ou mais características do modelo podem não se ajustar aos dados da amostra. Então, é importante investigar a aptidão do modelo, antes de qualquer análise mais aprofundada dos resultados. O exame inicial é realizado quanto ao ajustamento do modelo, através dos testes de variância e de inferência estatística. O modelo numérico gerado na análise não pode ser generalizado e aceito em qualquer situação. Para que possa ser empregado na estimação de valores, deve obedecer a algumas exigências. As condições que devem ser satisfeitas são chamadas de pressupostos, hipóteses ou condições básicas. São elas:

1. Relação linear entre a variável dependente e as independentes;
2. Normalidade dos resíduos;
3. Média zero dos resíduos [$(\Sigma e_i)/n = 0$];
4. Homocedasticidade dos resíduos (variância constante);
5. Independência serial dos resíduos (não existe autocorrelação dos erros);
6. As variáveis importantes foram incluídas;
7. Não existem observações espúrias (outliers);
8. Não-aleatoriedade das variáveis independentes (são fixas, só a dependente é aleatória);
9. Não existência de relação exata (colinearidade perfeita) entre quaisquer das variáveis independentes;
10. Número de observações maior que o número de coeficientes a ser estimado.

Essas condições devem ser garantidas para que o modelo encontrado e a própria análise de regressão tenha validade. Para qualquer processo de estimação empregado (Mínimos Quadrados, Máxima Verossimilhança), essas condições básicas devem ser respeitadas. São apresentados a seguir os problemas que advém da violação destas condições, testes para detecção e soluções que podem ser empregadas para resolver ou amenizar o problema. Além disso, são usados testes para a verificação da validade ou não do modelo de regressão proposto e das variáveis que o compõe, individualmente.

Os testes estatísticos geralmente não indicam respostas do tipo “sim/não” (determinístico), mas do tipo “melhor/pior” (probabilístico), e a análise da gravidade da situação cabe ao avaliador. Neste sentido, inicialmente devem ser testadas com prioridade a normalidade dos resíduos e a existência de outliers, que são razoavelmente fáceis de serem examinadas e estão ligadas a outros problemas. Caso não sejam satisfeitas as condições, o modelo deve ser descartado, sem que se perca tempo com os outros testes.

1. Linearidade da relação entre as variáveis

A análise de regressão baseia-se no "modelo linear clássico", como são as seguintes equações:

$$\hat{Y} = a + bX; \quad \hat{Y} = a_0 + a_1X_1 + a_2X_2 + ... + a_kX_k$$

Se alguma variável X_i mantém relacionamento não linear com Y, então surgem problemas na estimativa da equação. A não-linearidade pode se verificar através de gráficos de resíduos contra as variáveis presentes no modelo (Y, X_1, X_2, ..., X_k). Espera-se que não haja forma definida para os pontos. Se forem detectadas tendências, deve-se linearizar a relação, usando-se transformações nas variáveis, tais como: logarítmos, inversas ou potências.

2. Normalidade dos resíduos

A análise de regressão baseia-se na hipótese de que os erros seguem uma distribuição normal (distribuição de Gauss). A condição de normalidade dos resíduos não é necessária para a obtenção dos estimadores de mínimos quadrados, mas é fundamental para a definição de intervalos de confiança e testes de significância. Ou seja, em presença de falta de normalidade, os estimadores são não-tendenciosos, mas os testes não têm validade, principalmente em amostras pequenas. Entretanto, pequenas fugas da normalidade não causam grandes problemas.

A não-normalidade dos resíduos pode ser causada por violações de outras condições básicas, tais como a heterocedasticidade (variância não-constante dos erros) ou a escolha de um modelo incorreto para a equação.

O teste mais simples e rápido é o teste gráfico, comparando a freqüência acumulada dos resíduos padronizados (ei/s, onde s é o desvio-padrão dos resíduos) com a curva Normal. O aspecto do gráfico varia de acordo com o software estatístico, mas sempre existe uma linha-base, representando a curva Normal, e os resíduos acumulados do modelo ajustado devem aproximar-se desta linha. Fugas significativas são indicadores de não-normalidade dos resíduos, e o modelo deve ser descartado, buscando-se outra configuração para a equação.

Também é bastante comum o emprego de um teste numérico. Pelas propriedades da Normal, 68% dos resíduos devem estar no intervalo (-1, 1), 90% no intervalo (-1,64; 1,64)

e 95% no intervalo (-1,96; 1,96). Então, se os resíduos corresponderem a estes limites, aproximadamente, pode-se dizer que a normalidade está garantida.

Gráfico: **gráfico de probabilidade normal** (*Normal Probability Plot*). Existem dois tipos de gráficos de probabilidade normal:

1º tipo: representa a probabilidade acumulada que seria de esperar se a distribuição fosse normal, em função da probabilidade observada acumulada dos erros (**Normal P-P Plot**);

2º tipo: representa o quantil de probabilidade esperado se a distribuição fosse normal em função dos resíduos (**Normal Q-Q Plot**).

Se os erros possuírem distribuição Normal, todos os pontos dos gráficos devem posicionarem-se mais ou menos sobre uma reta.

Testes: Kolmogorov-Smirnov, Shapiro-Wilk, Jarque-Bera, Lilliefors.

3. Média zero dos resíduos

A média dos erros é sempre nula (bem como a soma dos erros), se existir a constante na equação, porque o processo de estimação ajusta o valor desta constante para que isto ocorre.

4. Homocedasticidade

Homocedasticidade é a variância constante dos resíduos. Esta é uma prioridade fundamental que deve ser garantida, sob pena de invalidar toda a análise estatística. Deseja-se que os erros sejam aleatórios, ou seja, não devem ser relacionados com as características das variáveis em estudo.

As conseqüências da heterocedasticidade são que as estimativas dos parâmetros da regressão (a_0, a_1, a_2, ..., a_k) não são tendenciosas mas são ineficientes e as estimativas das variâncias são tendenciosas. Os testes t e F tendem a dar resultados incorretos. Neste caso, os resultados não são confiáveis, ou seja, o modelo pode parecer bom, mas ele não é adequado aos dados, na verdade.

A heterocedasticidade pode ser verificada através de gráficos de resíduos (erros) contra os valores reais e contra os valores calculados pela equação. Se os pontos estão distribuídos aleatoriamente, sem demonstrar um comportamento definido, há homocedasticidade. Mas se existe alguma tendência (crescimento/decrescimento/ oscilação), então há heterocedasticidade. Havendo heterocedasticidade, podem ser tentadas transformações nas variáveis (geralmente logarítmicas) ou outras soluções mais

complexas. O modelo deve ser modificado. A análise gráfica da nuvem de pontos é simples e pode ser muito útil.

Gráficos e testes: Golfeld-Quandt, Breush-Pagan, White, Bartlett, Levene, Brown-Forsythe, Welch.

5. Independência serial dos resíduos (não-autocorrelação)

Existe autocorrelação quando os erros são correlacionados com os valores anteriores ou posteriores na série. Este problema também é chamado de correlação serial.

Pode surgir por especificação incorreta do modelo de regressão, devido a erros na forma do modelo ou por exclusão de variáveis independentes importantes. Isto ocorre principalmente em aplicações envolvendo séries temporais. Se a autocorrelação ocorrer por omissão de variáveis importantes, elas mesmas correlacionadas linearmente, pode ocorrer autocorrelação também em amostragens tipo *cross-section* (curto intervalo de tempo).

Existindo autocorrelação, os estimadores ordinários de mínimos quadrados não são mais os melhores estimadores lineares não-tendenciosos (as variâncias amostrais dos coeficientes estimados para a equação serão excessivamente grandes, essas variâncias serão subestimadas, as fórmulas perderão a validade e serão obtidas predições ineficientes). Neste caso existirão outros métodos que produzem menor variância amostral nos estimadores.

Além disso, em presença de correlação serial, os testes de significância (t e F) e os intervalos de confiança dos coeficientes da regressão também oferecem conclusões incorretas, isto é, as regiões de aceitação e os intervalos de confiança podem ser mais largos ou mais estreitos do que os calculados, dependendo da tendência ser positiva ou negativa.

Pode-se detectar a autocorrelação através de gráficos dos resíduos contra os valores da variável dependente. Se os resíduos parecem seguir um padrão, há autocorrelação.

Se os dados são obtidos em uma seqüência de tempo, é interessante construir um gráfico dos resíduos contra o tempo, mesmo que este não tenha sido incluído no modelo. Desta forma, verifica-se se há relação entre os erros no tempo. Havendo esta relação, pode-se interpretar como sendo o caso de uma variável omitida (tempo). Quando os termos de erro são independentes, pode-se esperar resíduos flutuando em torno de zero. Se aparecer uma alternância excessiva de resíduos positivos ou negativos, ou poucas alterações, podem existir problemas de falta de aleatoriedade.

Teste: Durbin-Watson.

6. Variáveis importantes ficaram fora da modelagem

As variáveis importantes devem participar da análise, devendo o pesquisador selecionar quais os aspectos prioritários para o estudo, coletando as informações que permitem quantificar estes aspectos.

Se uma variável não é utilizada, por impossibilidade de medição, devem ser usadas outras variáveis que possam simular os efeitos da verdadeira. Essas variáveis substitutas são conhecidas como variáveis *proxy*.

O gráfico comparando os erros com as variáveis não incluídas pode revelar sua importância. Se existe um padrão de variação sistemática dos resíduos, possivelmente é causado pela variável não incluída no modelo devendo esta ser incluída.

7. Outliers ou observações discrepantes

Outliers são elementos com comportamento diferente dos demais. Os resíduos são grandes, em módulo. É extremamente importante controlar os outliers porque, em virtude da forma de estimação da equação, geralmente por mínimos quadrados, um erro grande modifica significativamente os somatórios, alterando os coeficientes da equação.

Não existem limites fixos, mas geralmente se adota o intervalo de 2 desvios-padrão em torno da média dos erros. Como a média tende a ser zero, os resíduos padronizados devem estar no intervalo [-2; +2].

Uma das formas mais simples e eficiente de análise dos *outliers* é o exame dos gráficos de resíduos. Se forem encontrados *outliers*, não se recomenda a exclusão automática, mesmo que a amostra seja grande. O primeiro passo é verificar se existem erros de leitura, digitação ou cálculo, ocorridos na transcrição para o computador. Em seguida, deve-se confirmar os dados na fonte da informação.

É possível que esteja faltando alguma variável na análise, que pudesse explicar as diferenças encontradas.

8. Não-aleatoriedade das variáveis independentes

As variáveis independentes devem ser determinísticas, ou seja, não podem conter parcelas aleatórias. É fácil perceber que a dependente é aleatória, pois ela decorre da soma de um grupo de variáveis determinísticas com um erro aleatório (o erro segue a distribuição Normal):

$$\hat{Y} = (a_0 + a_1X_1 + a_2X_2 + ... + a_kX_k) + \varepsilon$$

Entretanto, nem sempre se pode garantir que as variáveis independentes sejam determinísticas. Existem variáveis que dependem de julgamento pessoal do avaliador, tais como estado de conservação, padrão de qualidade da construção e da vizinhança, e que estão sujeitas a erros de medição aleatórios (por exemplo, vistorias em dias diferentes, com condições climáticas distintas ou a diferença constatada entre as primeiras e últimas vistorias, refletindo o desgaste físico do profissional). Alguns erros de medição também podem conduzir à aleatoriedade em variável *proxy*, tais como o uso da idade legal (do Habite-se) para representar a idade física ou a distância em linha reta para representar a acessibilidade a algum ponto importante.

A prova de não-aleatoriedade é mais difícil. Cada variável deve ser confrontada com curvas de distribuições de probabilidade (Normal, t, F, binomial, etc), verificando se há ajuste de algumas delas. Contudo, para finalidade prática, pode-se aceitar pequenas influências aleatórias, e a não-aleatoriedade pode ser presumida, se o modelo for aceito nas análises de *outliers* e normalidade dos resíduos.

9. Multicolinearidade (em caso de regressão múltipla)

Duas ou mais variáveis são colineares se possuem relação exata, ou seja, se um dos vetores é uma combinação linear dos outros (como se fossem retas paralelas). A correlação exata raramente ocorre, mas correlações fortes (acima de 0,8) já são perigosas.

Apenas a correlação entre variáveis independentes é problemática. A relação forte de cada uma das variáveis independentes (X_i) com a variável dependente (Y) é desejável. Quando existem mais de duas variáveis independentes relacionadas fortemente fala-se em multicolinearidade.

A multicolinearidade afeta significativamente os coeficientes da equação de regressão, alterando o valor e até o sinal em relação ao que ocorreria se não houvesse este problema. Na presença de correlação alta, os coeficientes de regressão estimados tendem a ser imprecisos e as estimativas dos coeficientes variam bastante de uma amostra para outra. Quando ocorre colinearidade, as estimativas dos mínimos quadrados ainda são não-tendenciosas e eficientes, porém o erro padrão dos coeficientes tende a ser grande, e o teste baseado na estatística “t” de Student calculará significância menor que a real.

Outro efeito da colinearidade é que se torna difícil obter interpretações sobre o efeito isolado de cada uma das variáveis.

Nos casos de correlação alta, uma das alternativas é a remoção da variável mais afetada. Isso pode introduzir tendências, sendo mais adequado substituir esta variável por

outra menor colinear, mas que tenha aproximadamente a mesma construção teórica, ou por uma variável que seja a combinação das colineares (A/B, A + B, A*B, etc.).

Nem sempre a remoção ou substituição da variável afetada é uma boa solução. Quando se trabalha com predição de valores e existem indicações de que a colinearidade encontrada continuará no futuro, o modelo poderá apresentar bons resultados.

A verificação da existência de colinearidade é realizada através do exame da matriz de correlação, relacionando todas as variáveis da análise, ou por meio de outros critérios,
Nota: A definição do limite aceitável de correlação é arbitrária, e pode-se adotar limites mais conservadores, como 0,5.

10. Número de observações maior do que o número de coeficientes a ser estimado

O tamanho da amostra deve ser maior que o número de variáveis analisadas, para que a análise da regressão (estatística) possa ser empregada. Se existir quantidade igual de observações e de coeficientes a ser estimado, é caso de resolução de sistema de equações, no âmbito da matemática (determinístico). Se o número de observações for inferior, não há solução.

A diferença entre o tamanho da amostra e o número de regressores (n–k) é chamada "graus de liberdade da estimativa". Um número mínimo de observações é determinado pela Norma, mas para incrementar a precisão dos modelos, o aumento da amostra é sempre desejável, ou seja, quanto mais graus de liberdade, melhor. A exigência da Norma é de $n>3k$ e $n>2k + 5$, adotando-se como mínimo o maior destes dois parâmetros.

3.10 - ADEQUABILIDADE DO MODELO E MEDIDAS REMEDIADORAS

1) Suposições na Análise de Regressão

1. A relação entre Y e X é linear ou no mínimo bem aproximada por uma linha reta
2. Os erros tem média zero
3. Os erros tem variância constante σ^2
4. Os erros são independentes
5. Os erros são normalmente distribuídos

Pode-se sempre colocar dúvidas sobre a validade destas suposições e então conduzir análises para verificar a adequabilidade do modelo. Um modelo inadequado geralmente traz sérias conseqüências. Violações acentuadas nas suposições podem produzir um modelo instável (Diferente amostra pode conduzir a conclusões totalmente opostas).

Não temos condições de verificar desvios nas suposições do modelo simplesmente examinando as estatísticas t ou F ou o coeficiente de determinação tradicionalmente calculadas na análise de Regressão.

2) Análise de Resíduos

(a) Definição de Resíduos

O resíduo e_i é definido como a diferença entre o valor observado e ajustado pelo modelo, ou seja:

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

ERRO DO MODELO
$\varepsilon_i = Y_i - E(Y_i) = Y_i - (\alpha + \beta X_i)$

É uma medida da variabilidade não explicada pelo modelo de Regressão. Pode-se pensar os resíduos como os valores "observados" dos erros. Assim qualquer desvio nas suposições dos erros repercutirá nos resíduos. A análise de resíduos é um método efetivo para descobrir deficiências no modelo de vários tipos.

(b) Propriedades dos Resíduos

1. Média : $\bar{e} = (\sum e_i / n) = 0$
2. Variância $s_e^2 = \dfrac{\sum (e_i - \bar{e})^2}{n-2} = \dfrac{\sum e_i^{\,2}}{n-2} = \dfrac{\sum (Y_i - \hat{Y}_i)^2}{n-2} = \dfrac{SQR}{n-2} = QMR \leftarrow s^2 y/x$

(c) Resíduos Padronizados

Por conveniência analítica, os resíduos padronizados $d_i = \dfrac{e_i}{\sqrt{QMR}}$ são muitas vezes utilizados nas análise de resíduos

(d) Desvios no modelo que podem ser identificados na Análise de Resíduos

1. A função de regressão não é linear
2. Os erros não têm variância constante
3. Os erros não são independentes
4. Os erros não são normalmente distribuídos
5. O modelo ajusta bem, mas uma ou algumas observações são discrepantes("outliers")
6. Variáveis independentes importantes foram omitidas do modelo

(e) Diagnóstico pelos resíduos

1. Gráfico dos resíduos contra variável preditora (erros *versus* X - analisar as 4 figuras a seguir)
2. Gráfico dos resíduos contra valores ajustados ou preditos (erros *versus* $\hat{Y}$)

3. Gráfico dos resíduos contra o tempo ou outra sequência (para verificar independência)
4. Gráfico dos resíduos contra variáveis preditoras omitidas (para ver sua contribuição)
5. Box-Plot dos resíduos (para identificar outliers)
6. Gráfico de probabilidade normal dos resíduos

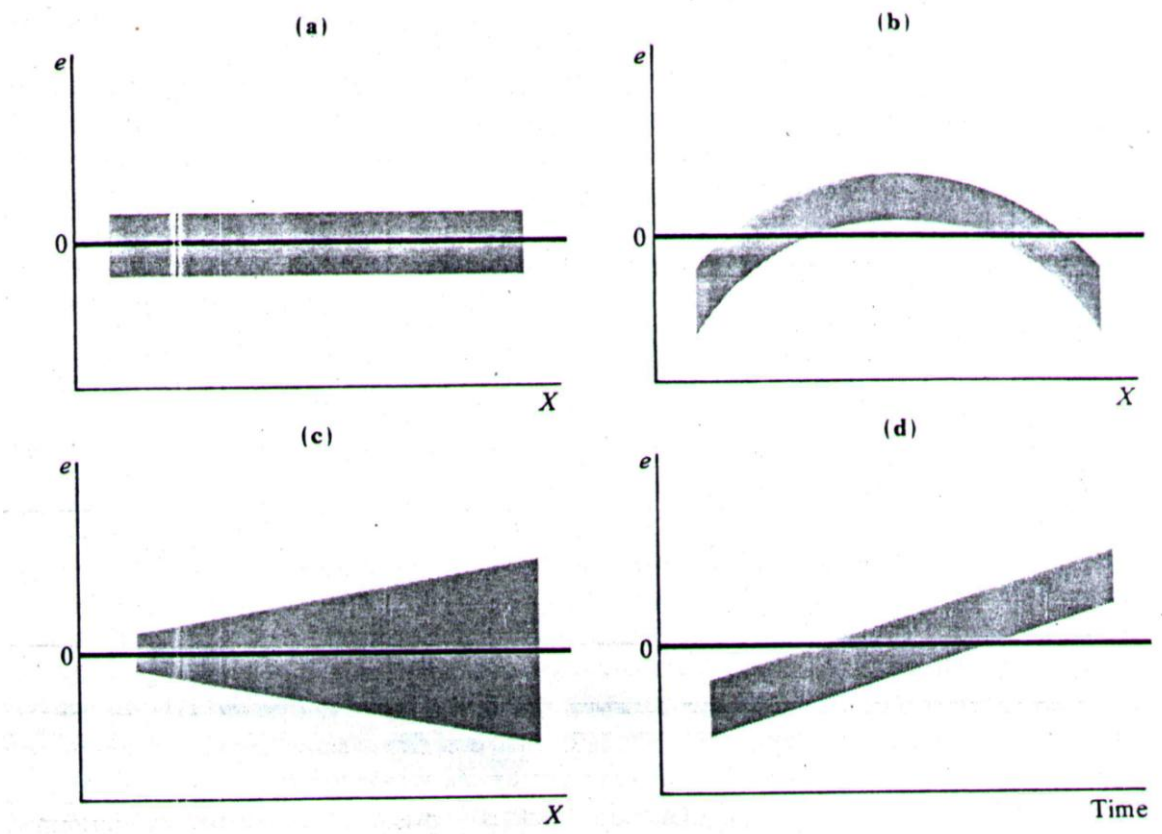
3. Medidas Remediadoras

Se o modelo de regressão linear simples (RLS) não é adequado para os dados há duas escolhas básicas:

1. Abandonar o modelo de RLS e buscar um modelo mais adequado
2. Usar alguma transformação nos dados de tal forma que o modelo de RLS seja apropriado para os dados transformados

PROBLEMAS	**SOLUÇÕES**
1) Função de regressão não linear	1) Transformação (em X) para linearização - adoção de modelo com curvatura
2) Erros com variância não constante	2) Transformação (em Y) para estabilizar a variância; mínimos quadrados ponderados
3) Não-normalidade dos erros	3) Transformação para normalizar - geralmente transformação para estabilizar a variância auxilia para normalizar
4) Não-independência dos erros correlacionados	4) Modelo que considera erros
5) Omissão de variáveis independentes	5) Ajuste de regressão múltipla

4. Gráficos de Resíduos



(a) Gráfico dos resíduos contra variável preditora quando o modelo de regressão linear é apropriado
(b) Gráfico dos resíduos contra variável preditora quando o modelo de regressão linear não é apropriado: necessidade de uma função de regressão curvilinear – falta de adequação
(c) Gráfico dos resíduos contra variável preditora identificando variância não constante dos resíduos. A variância dos resíduos cresce com X - heterocedasticidade
(d) Gráfico dos resíduos contra variável de tempo, identificando falta de independência dos resíduos. Existe uma tendência linear de acréscimo nos resíduos em função do tempo autocorrelação

Ainda, por meio do gráfico de resíduos pode-se identificar falta de aderência à normalidade (testes estatísticos, gráfico de probabilidade normal) e presença de "outliers" (Box-plot dos resíduos).