CAPÍTULO 8 – ANÁLISE DE REGRESSÃO MÚLTIPLA – O PROBLEMA DA INFERÊNCIA

1-A HIPÓTESE DE NORMALIDADE

Considere novamente o modelo de regressão múltipla com duas variáveis explicativas dado por:

$$Y_i = \beta_1 + \beta_2 . X_{2i} + \beta_3 . X_{3i} + \varepsilon_i \quad \text{para i = 1, 2, ..., n} \qquad (1)$$

A hipótese de normalidade dos erros e variância constante garante que os estimadores por mínimos quadrados dos β's são:

- Iguais aos estimadores de máxima verossimilhança;
- São os melhores estimadores LINEARES não tendenciosos dos β's, ou seja, são os BLUEs (Best Linear Unbiased Estimators).
- Um múltiplo da variância amostral, mais precisamente: $(n-3)\frac{\hat{\sigma}^2}{\sigma^2}$ tem distribuição Qui-Quadrado com n – 3 graus de liberdade e esta distribuição é independente da dos estimadores MQO dos β's.
- Acima $\hat{\sigma}^2 = \frac{RSS}{n-3} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-3} = \frac{\sum_{i=1}^{n}(\hat{\varepsilon}_i)^2}{n-3}$ é a variância amostral

Pelos comentários acima é fácil perceber que testes de hipóteses e intervalos de confiança para os β's podem ser baseados em estatísticas t calculadas a partir de seus estimadores MQO. Especificamente:

$$T = \frac{\hat{\beta}_j - \beta_j}{ep(\hat{\beta}_j)} \quad \text{para j} = 1,2,3 \qquad (2)$$

As estatísticas t dadas na equação (2) têm distribuição t de Student com n-3 graus de liberdade. Também, na definição (2), os denominadores são os erros padrão de cada estimador MQO, que são dados (vide apêndice C do Gujarati) pelos elementos da diagonal da matriz $V = \sigma^2 (X^t X)^{-1}$ onde $\sigma^2 = \text{VAR}(\varepsilon_i)$ para $\text{i} = 1,2,...,\text{n}$. Na prática, σ^2 é estimado por RSS/(n-3) neste caso particular do modelo com duas variáveis explicativas e termo constante.

Nota – caso geral

Considere o modelo de regressão com termo constante e k-1 variáveis explicativas.

$$Y_i = \beta_1 + \beta_2.X_{2i} + ... + \beta_k.X_{ki} + \varepsilon_i \quad \text{para i} = 1, 2, ...\text{n} \qquad (3)$$

Note que existem k+1 parâmetros a estimar – os k coeficientes β_ι e a variância do erro σ^2.

O estimador da variância é:

$$\hat{\sigma}^2 = \frac{RSS}{n-k} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-k} = \frac{\sum_{i=1}^{n}(\hat{\varepsilon}_i)^2}{n-k}, \text{ a variância amostral} \qquad (4)$$

Também:

$(n-k)\frac{\hat{\sigma}^2}{\sigma^2}$ tem distribuição Qui-Quadrado com (n-k) graus de liberdade, e é independente das estatísticas t baseadas nos estimadores dadas por:

$$T = \frac{\hat{\beta}_j - \beta_j}{ep(\hat{\beta}_j)} \quad \text{para j} = 1,2,3...,\text{k} \quad \text{têm distribuição t de Student com (n - k) graus de liberdade}$$

A estatística T serve para testar premissas sobre os coeficientes parciais individuais da regressão, e para encontrar intervalos de confiança para estes coeficientes. No entanto, veremos que estes não são únicos testes interessantes que podemos (ou devemos) fazer no contexto de regressão múltipla.

Exemplo 8.1.

Considere a regressão múltipla do exemplo 7.1. (carros). Os resultados são novamente mostrados aqui.

```
The regression equation is
milhas_por_galao = 52,9 - 0,0990 hp - 0,00698 peso


282 cases used, 9 cases contain missing values


Predictor         Coef     SE Coef        T      P
Constant        52,904       1,365    38,76  0,000
hp            -0,09897     0,02062    -4,80  0,000
peso       -0,0069813   0,0006945   -10,05  0,000


S = 4,33857   R-Sq = 58,5%   R-Sq(adj) = 58,2%
```

Vamos examinar as estatísticas t para cada um dos coeficientes estimados e testar a (importantíssima) hipótese de que o coeficiente é nulo, o que equivale a dizer que a variável correspondente não tem, ignorando o efeito das outras variáveis, efeito sobre a variável dependente.

Assim, as hipóteses que desejamos testar têm todas o aspecto:

H_0: $\beta_i = 0$ versus H_1: $\beta_i \neq 0$ para i = 1, 2, 3 neste caso. Ou seja, estamos interessados no teste de hipóteses bilateral.

Para o termo constante:

$T = \frac{52{,}904 - 0}{1{,}365} = 38{,}76$ que deve ser comparado com o valor bilateral de uma distribuição t com 282-3 = 279 graus de liberdade, o que é, na prática, uma distribuição N(0,1). Para um nível de significância 95%, o percentil 97,5% da N(0,1) é 1,96, e claramente 38,76 está MUITO fora do intervalo (-1,96, +1,96). Para a distribuição t_{279} o percentil 97,5% é 1,968 e as conclusões são as mesmas. Se olharmos para o valor-p apresentado chegamos à mesma conclusão, pois o valor p é menor que 0,000, significando que a probabilidade de obter uma estatística t maior ou igual a 38,76 (em módulo) é virtualmente zero.

Para o coeficiente β_1 (ou, alternativamente, para a variável HP):

$$T = \frac{-0{,}09897 - 0}{0{,}02062} = -4{,}80$$

Por argumentos análogos aos anteriores (atmbém comparamos a estatística T com os percentis da distribuição t com 279 graus de liberdade, nota-se que o valor da estatística T é muito grande (em módulo) e assim rejeita-se fortemente a hipótese de que a variável HP não tenha (individualmente) efeito sobre a variável MPH (consumo de combustível). Além disso, este efeito é negativo – quanto mais potente o carro (quanto maior o valor de HP), menos milhas por galão o carro faz, ou seja, MAIOR o consumo de combustível.

Para o coeficiente β_2 (ou seja, para a variável MOTOR, que mede o tamanho do motor):

$$T = \frac{-0{,}0069813 - 0}{0{,}0006945} = -10{,}05$$

Novamente, conclui-se que o efeito da variável é altamente significante e também negativo – quanto maior o tamanho do motor, menor o valor de MPH (ou seja, maior o consumo).

Testes de Hipóteses e Intervalos de Confiança

Certamente você já viu a relação estreita que existe entre testes de hipóteses e intervalos de confiança. No contexto mostrado anteriormente, um intervalo de confiança (1-α)% para β_i é dado por:

$$\left(\hat{\beta}_i - t_{1-\frac{\alpha}{2},n-k}.ep\left(\hat{\beta}_i\right), \hat{\beta}_i + t_{1-\frac{\alpha}{2},n-k}.ep\left(\hat{\beta}_i\right)\right) \quad (5)$$

Onde $t_{1-\frac{\alpha}{2},n-k}$ é o percentil 1-α/2 da distribuição t de Student com n-k graus de liberdade, e ep(.) indica o erro padrão do respectivo estimador por mínimos quadrados. No caso particular da regressão com termo constante e duas variáveis explicativas, k =3.

Por exemplo, um IC 95% para o coeficiente β_1 (correspondente à variável HP) é:

$$\left(\hat{\beta}_i - t_{1-\frac{\alpha}{2},n-k}.ep\left(\hat{\beta}_i\right), \hat{\beta}_i + t_{1-\frac{\alpha}{2},n-k}.ep\left(\hat{\beta}_i\right)\right) = (-0{,}09897 - 1{,}968 * 0{,}02062, -0{,}09897 + 1{,}968 * 0{,}02062) =$$
$$= (-0{,}13955, -0{,}05839)$$

Note que este intervalo NÃO INCLUI ZERO.

Então em geral podemos dizer:

Teste de Hipótese	Intervalo de Confiança
$H_0: \beta_i = 0$ versus $H_1: \beta_i \neq 0$ O parâmetro é significante (isto é, diferente de zero se o IC NÃO INCLUI ZERO)	$\left(\hat{\beta}_i - t_{1-\frac{\alpha}{2},n-k}.ep\left(\hat{\beta}_i\right), \hat{\beta}_i + t_{1-\frac{\alpha}{2},n-k}.ep\left(\hat{\beta}_i\right)\right)$

2 – Testando a Normalidade dos Resíduos – o teste de Jarque e Bera

Para aplicar corretamente o teste t é necessário que os erros do modelo sejam Normais. No entanto, os erros não são diretamente observáveis – o melhor que podemos fazer é basear testes nos seus estimadores, os resíduos.

O teste para Normalidade é o teste de Jarque-Bera, que só é **válido para grandes amostras** (vide Gujarati – capítulo 5). Este teste é baseado na assimetria e na curtose de uma variável aleatória. Sabemos que, se uma variável é Normal, sua assimetria é 0 e sua curtose é 3.

A estatística de Jarque e Bera é definida como:

$$JB = n\left\{\frac{SKEW^2}{6} + \frac{(K-3)^2}{24}\right\} \quad (6)$$

Onde n é o tamanho da amostra, SKEW é o coeficiente de assimetria e K é o coeficiente de curtose. A estatística JB será "grande" (muito maior que zero) se houver evidências de não

normalidade, tanto pelo lado da assimetria quanto pelo lado da curtose. **Sob a hipótese nula de que os resíduos têm distribuição Normal, a estatística JB é assintoticamente Qui-Quadrado com 2 graus de liberdade.**

Comenta-se que a aproximação Qui-quadrado para a estatística JB é ruim e muitas vezes leva, indevidamente, à rejeição da hipótese nula de normalidade dos dados. Os p-valores para um conjunto de dados e nível de significância especificado podem ser obtidos em MATLAB (através de simulação de Monte Carlo) usando a função jbtest. Para mais informações veja:

http://www.mathworks.com/access/helpdesk/help/toolbox/stats/jbtest.html

Exemplo 8.2.

Considere a regressão múltipla de MPH e HP e Motor mostrada no exemplo 8.1. Os resíduos desta regressão têm as seguintes estatísticas descritivas:

Descriptive Statistics: residuos

Variable	N	N*	Mean	SE Mean	StDev	Variance	Minimum	Q1
residuos	282	9	2,16702E-14	0,257	4,323	18,689	-9,540	-2,672

Variable	Median	Q3	Maximum	Skewness	Kurtosis
residuos	-0,511	2,342	14,860	0,60	0,77

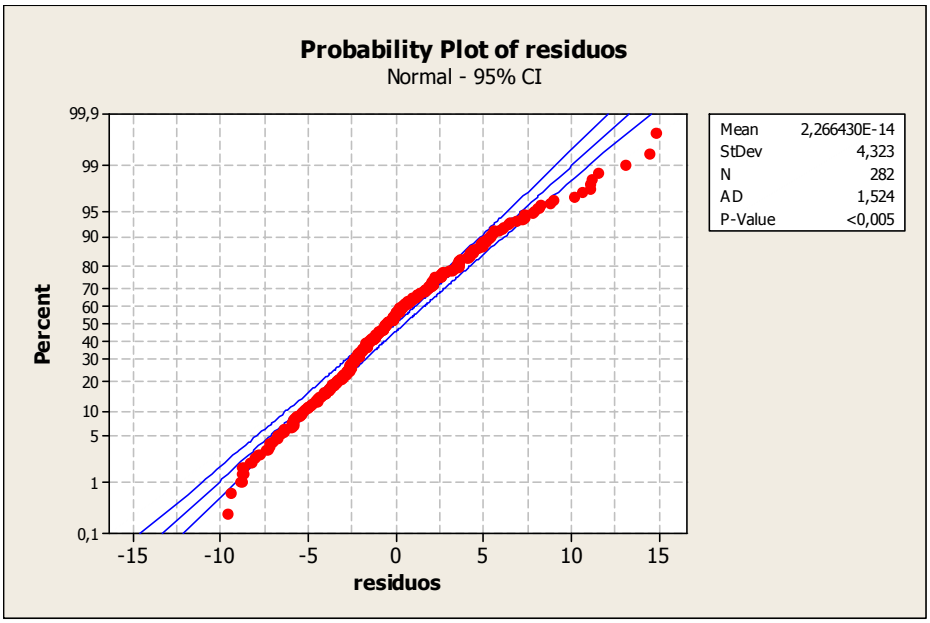
A estatística de Jarque e Bera é:

$$JB = n\left\{\frac{SKEW^2}{6} + \frac{(K-3)^2}{24}\right\} = (282-9)\left\{\frac{(0{,}60)^2}{6} + \frac{(0{,}77-3)^2}{24}\right\} = 273 * \{0{,}2672\} = 72{,}946$$

Neste caso a estatística JB é enorme (se comparada a uma variável Qui-Quadrado com 2 graus de liberdade), e rejeita-se fortemente a hipótese de Normalidade dos resíduos.

Na verdade, este resultado é corroborado pelo próximo gráfico, um QQ-Plot dos resíduos da regressão versus a distribuição Normal. Note como os resíduos se afastam da linha reta e como estão fora das linhas que indicam o intervalo de confiança.

Figura 1 – QQ-Plot dos Resíduos versus Distribuição Normal



3 – Teste da Significância Geral do Modelo de Regressão

Suponha que, no modelo com o termo constante e duas variáveis explicativas queremos testar a hipótese conjunta de que ambos os coeficientes das variáveis explicativas são nulos, isto é:

$$H_0: \beta_2 = \beta_3 = 0$$

Você poderia pensar em resolver através de dois testes t, isto é, testar primeiro se $\beta_2 = 0$ e depois se $\beta_3 = 0$. No entanto, isso não funciona, pois os testes não são "independentes", e o problema por trás disso é o fato da $COV(\hat{\beta}_2, \hat{\beta}_3) \neq 0$. Embora seja verdade que:

$$\beta_i \in \left(\hat{\beta}_i - t_{1-\frac{\alpha}{2},n-3}.ep(\hat{\beta}_i), \hat{\beta}_i + t_{1-\frac{\alpha}{2},n-3}.ep(\hat{\beta}_i) \right) \text{ com probabilidade } (1-\alpha) \quad \text{para i} = 2,3$$

Não é verdade que (simultaneamente):

$$(\beta_2, \beta_3) \in \left(\hat{\beta}_2 \pm t_{1-\frac{\alpha}{2},n-3}.ep(\hat{\beta}_2), \hat{\beta}_3 \pm t_{1-\frac{\alpha}{2},n-3}.ep(\hat{\beta}_3) \right) \quad \text{com probabilidade } (1-\alpha)*(1-\alpha)$$

Então, como testar a hipótese nula simultânea $H_0: \beta_2 = \beta_3 = 0$? A resposta é o teste F, que "sai" da tabela de Análise de Variância mostrada nos diagnósticos de um modelo de regressão múltipla.

Lembre-se que podemos particionar a soma de quadrados total (vide capítulo 7) como:

$$SST = SYY = \sum (y_i - \bar{y})^2 = \text{SSReg} + \text{RSS} \qquad (7)$$

Onde: SST é a soma dos quadrados total, SSReg é a soma dos quadrados devido à regressão e RSS é a soma do quadrado dos resíduos. Os graus de liberdade associados a cada um destes termos são, respectivamente, (n -1), 2 e (n-3).

A tabela ANOVA pode ser montada da seguinte maneira (no modelo com termo constante e 2 variáveis explicativas):

Fonte de Variação	Graus de Liberdade (gl)	Mean Squared (MSQ)
Regressão (SSREg)	2	MSReg= SSReg/2
Resíduos (RSS)	n-3	$\hat{\sigma}^2 = \frac{RSS}{n-3}$
Total (SST = SYY)	n-1	$S^2 = \frac{SYY}{n-1} = \frac{1}{n-1}\sum (y_i - \bar{y})^2$

Sob a premissa de normalidade dos erros e sob a hipótese nula $\beta_2 = \beta_3 = 0$ segue que a estatística:

$$F = \frac{SS\,\text{Reg}/(2)}{RSS/(n-3)} \quad \text{tem distribuição F com 2 graus no numerador e (n - 3) no denominador}$$

Se F é "grande" rejeita-se a hipótese nula de igualdade $\beta_2 = \beta_3 = 0$, ou seja, rejeita-se EM CONJUNTO a hipótese de que as variáveis correspondentes não afetam a variável dependente.

Exemplo 8.3.

Para os dados de carros (Exemplo 8.1), a tabela ANOVA é dada a seguir. Investigue a significância conjunta dos coeficientes β_2 e β_3.

```
Analysis of Variance

Source           DF       SS      MS       F      P
Regression        2   7394,3  3697,2  196,41  0,000
Residual Error  279   5251,7    18,8
Total           281  12646,0
```

Solução

Verifique que os valores dados na coluna MS realmente correspondem à coluna SS (soma de quadrados) dividida pelos respectivos graus de liberdade (coluna gl).

A estatística F é dada por:
$F = \frac{SS\,\mathrm{Re}\,g/(2)}{RSS/(n-3)} = \frac{7394{,}7/(2)}{5251{,}7/(279)} = \frac{3697{,}35}{18{,}8233} = 196{,}4241$ como indicado na saída do "software" (a menos de erros de arredondamento). O percentil 99% da distribuição F(2,279) é 4,68 e então a estatística F é muito maior que este percentil, indicando que β_2 e β_3 são conjuntamente diferentes de zero.

Teste F – extensão para o modelo com k variáveis

Considere o modelo geral (com k-1 variáveis explicativas) dado pela equação (3):

$$Y_i = \beta_1 + \beta_2.X_{2i} + ... + \beta_k.X_{ki} + \varepsilon_i \quad \text{para i} = 1, 2, ...\text{n} \qquad (3)$$

Suponha que desejamos testar a hipótese de que todas as variáveis são conjuntamente não significantes, ou seja, a hipótese nula é:

$$H_0: \beta_2 = \beta_3 = ... = \beta_k = 0$$

A hipótese alternativa é: H_1: nem todos os coeficientes são zero.

A estatística de teste é:

$$F = \frac{SS\,\mathrm{Re}\,g/(k-1)}{RSS/(n-k)} \quad \text{tem distribuição F com (k - 1) graus no numerador e (n - k) no denominador}$$

Se F > $F_{1-\alpha}$(k-1, n-k) onde $F_{1-\alpha}$(k-1, n-k) é o percentil (1-α)% da distribuição F(k-1, n-k) REJEITA-SE a hipótese nula. Do contrário, a hipótese nula de que os parâmetros são conjuntamente iguais a zero não é rejeitada.

A relação entre a estatística F e o R^2

Lembre-se que o coeficiente de determinação R^2 foi definido como:

$$R^2 = \frac{SS\,\mathrm{Re}\,g}{SST} = \frac{SS\,\mathrm{Re}\,g}{SYY} = \frac{SYY - RSS}{SYY} = 1 - \frac{RSS}{SYY}$$

Pela definição da estatística F:

$$F = \frac{SS\,\mathrm{Re}\,g/(k-1)}{RSS/(n-k)} = \frac{\mathrm{n-k}}{\mathrm{k-1}}\left(\frac{SS\,\mathrm{Re}\,g}{RSS}\right) = \frac{\mathrm{n-k}}{\mathrm{k-1}}\left(\frac{SS\,\mathrm{Re}\,g}{SYY - SS\,\mathrm{Re}\,g}\right) = \frac{\mathrm{n-k}}{\mathrm{k-1}}\left(\frac{SS\,\mathrm{Re}\,g/SYY}{(SYY - SS\,\mathrm{Re}\,g)/SYY}\right) =$$

$$= \frac{\mathrm{n-k}}{\mathrm{k-1}}\left(\frac{R^2}{1-R^2}\right) \qquad (8)$$

A relação entre R^2 e F é direta. Se $R^2 = 0$, F também é. Se R^2 tende a 1, F tende a infinito. Quanto maior o R^2, maior o valor da estatística F.

4- A Função de Produção Cobb-Douglas

Exemplos com a função de produção Cobb-Douglas aparecem com frequência no livro, e neste capítulo ela surge para motivar a idéia de mínimos quadrados restritos. Os comentários a seguir sobre Cobb-Douglas foram retirados principalmente da Wikipedia.

Uma Função de produção é usada em Economia para representar o relacionamento de uma determinada saída e às diversas entradas. Na prática, a produção Y é geralmente representada como função dos "inputs" capital K e trabalho L, ou seja, pode-se escrever uma função de produção genérica como Y = F(K, L).

Uma **função de produção Cobb-Douglas** tem a forma:

$$Y = F(K, L) = A.L^{\alpha}.K^{\beta} \qquad (9)$$

Onde: Y = saída = produção

L = entrada = trabalho

K = entrada de capital

A, α e β são constantes determinadas pela tecnologia. A é a produtividade total dos fatores.

Se **α + β = 1**, a função de produção tem **retornos constantes de escala** (se L e K forem aumentados 20%, Y aumenta 20%).

Se **α + β é menor que 1**, os retornos de escala são decrescentes, e se forem maiores que 1, os retornos à escala estão aumentando, se dobramos o uso de insumos, a produção cresce mais que duas vezes, por exemplo.

Considere agora uma função de produção Cobb-Douglas na forma estocástica (ou seja, inserindo-se um termo de erro):

$$Y_i = \beta_1 . X_{2i}^{\beta_2} . X_{3i}^{\beta_3} . e^{\varepsilon_i} \quad (10)$$

Onde Y é a produção, X_2 e X_3 os insumos mão-de-obra e capital, e ε um erro estocástico.

Pode-se aplicar o logaritmo aos dois lados da equação (10) e encontramos uma forma funcional bem mais "simpática":

$$\ln(Y_i) = \ln(\beta_1) + \beta_2 . \ln(X_{2i}) + \beta_3 . \ln(X_{3i}) + \varepsilon_i = \alpha + \beta_2 . \ln(X_{2i}) + \beta_3 . \ln(X_{3i}) + \varepsilon_i \quad (11)$$

Onde $\alpha = \ln(\beta_1)$.

A equação (11) é um modelo de regressão linear múltipla onde os regressores são $\ln(X_2)$ e $\ln(X_3)$ e a variável dependente é ln(Y). O que significam os coeficientes β_2 e β_3 nesta equação? β_2 é a elasticidade da produção em relação ao insumo X_2 (mão de obra), ou seja, é a variação PERCENTUAL na produção quando há uma variação de 1% em X_2 e X_3 é mantido constante. Analogamente, β_3 é a elasticidade da produção em relação ao insumo capital mantida constante a mão de obra.

Uma hipótese que faz sentido testar do ponto de vista econômico é a de retornos constantes de escala. Esta hipótese se traduz estatisticamente em H_0: $\beta_2 + \beta_3 = 1$.

Podemos atacar este problema de duas maneiras: a primeira, e mais simples, é através de um teste t. A segunda é mais geral, e usa um teste F.

Abordagem do teste t para a hipótese H_0: $\beta_2 + \beta_3 = 1$

A estatística de teste é apenas:

$$t = \frac{\hat{\beta}_2 + \hat{\beta}_3 - (\beta_2 + \beta_3)}{ep(\hat{\beta}_2 + \hat{\beta}_3)} = \frac{\hat{\beta}_2 + \hat{\beta}_3 - (\beta_2 + \beta_3)}{\sqrt{VAR(\hat{\beta}_2) + VAR(\hat{\beta}_3) + 2COV(\hat{\beta}_2, \hat{\beta}_3)}}$$

E sob a hipótese nula H_0: $\beta_2 + \beta_3 = 1$ esta estatística se reduz a:

$$t = \frac{\hat{\beta}_2 + \hat{\beta}_3 - 1}{\sqrt{VAR(\hat{\beta}_2) + VAR(\hat{\beta}_3) + 2COV(\hat{\beta}_2, \hat{\beta}_3)}}$$

Esta estatística deve ser comparada com o percentil (1-α)% da distribuição t com n-k graus de liberdade (k =3 aqui). Se o valor da estatística t for superior ao percentil crítico, rejeitamos a hipótese nula, do contrário a hipótese nula não é rejeitada.

Abordagem do teste F para a hipótese H_0: $\beta_2 + \beta_3 = 1$

Uma abordagem direta para o teste desta hipótese seria incorporá-la desde o início no modelo, calculando um modelo "restrito".

Se H_0 é verdadeiro, então obviamente: $\beta_2 = 1- \beta_3$ e $\beta_3 = 1-\beta_2$. Podemos usar qualquer uma destas igualdades diretamente no modelo e reescrever a função de produção Cobb-Douglas dada pela equação (11):

$$\begin{aligned}
&\ln(Y_i) = \alpha + \beta_2.\ln(X_{2i}) + \beta_3.\ln(X_{3i}) + \varepsilon_i = \alpha + (1-\beta_3).\ln(X_{2i}) + \beta_3.\ln(X_{3i}) + \varepsilon_i \\
&\ln(Y_i) = \alpha + \beta_3.\{\ln(X_{3i}) - \ln(X_{2i})\} + \ln(X_{2i}) + \varepsilon_i \\
&\ln(Y_i) - \ln(X_{2i}) = \alpha + \beta_3.\{\ln(X_{3i}) - \ln(X_{2i})\} + \varepsilon_i \qquad (12) \\
&\ln\left(\frac{Y_i}{X_{2i}}\right) = \alpha + \beta_3.\ln\left(\frac{X_{3i}}{X_{2i}}\right) + \varepsilon_i
\end{aligned}$$

Ou seja, após aplicarmos a restrição $\beta_2 = 1- \beta_3$, a equação (11) torna-se uma regressão linear simples que relaciona as variáveis Y/X_2 e X_3/X_2. A equação (12) nos permite estimar β_3 e β_2 é encontrado aplicando a restrição da soma ser igual a um.

O procedimento descrito em (12) é conhecido como *mínimos quadrados restritos (MQR)* e pode ser generalizado para modelos com um número qualquer de variáveis explicativas sujeitas a uma restrição de igualdade.

A questão que resta é: como comparar os modelos restrito e irrestrito? A resposta está no teste F.

Sejam:

- RSS_{SEM} a soma do quadrado dos resíduos do modelo sem restrições;
- RSS_{COM} a soma do quadrado dos resíduos do modelo com restrições;
- m = número de restrições lineares (= 1 na situação mostrada aqui);
- k = número de parâmetros da regressão SEM restrições;
- n = número de observações

A estatística: $F = \dfrac{(RSS_{COM} - RSS_{SEM})/m}{RSS_{SEM}/(n-k)}$ tem distribuição F(m, n-k).

Exemplo 8.4.

Este exemplo tenta reproduzir o Exemplo 8.3. do Gujarati para a economia brasileira, obtendo uma função de produção para a nossa economia. Os dados foram obtidos no site do ipeadata.

Considere as seguintes variáveis:

PIB = PIB anual em reais de 1980

Capital = Capital fixo - formação bruta – em R$ de 1980

População = População residente em 1º de julho em milhares de habitantes

PEA = População Economicamente Ativa em milhares de pessoas - média dos valores mensais - séries antiga e nova mescladas – disponível apenas a partir de 1982

A idéia é ajustar uma função de produção Cobb-Douglas ao PIB usando Capital e População (ou a PEA) como os insumos. Também ajustaremos o modelo restrito, que corresponde aos retornos constantes de escala na função de produção.

Modelo 1 (Irrestrito)

Regression Analysis: log_PIB versus log_K; log_pop

```
The regression equation is
log_PIB = - 13,9 + 0,447 log_K + 1,31 log_pop


Predictor       Coef  SE Coef       T      P
Constant   -13,8654   0,4282  -32,38  0,000
log_K       0,44685  0,03156   14,16  0,000
log_pop     1,31167  0,03588   36,56  0,000


S = 0,0325001   R-Sq = 99,3%   R-Sq(adj) = 99,3%


Analysis of Variance

Source          DF      SS      MS        F      P
Regression       2  5,5498  2,7749  2627,13  0,000
Residual Error  37  0,0391  0,0011
Total           39  5,5889


Source   DF  Seq SS
log_K     1  4,1382
log_pop   1  1,4116
```

A equação do modelo irrestrito mostrada acima nos diz que o modelo ajustado é:
PIB = $(9{,}51 * 10^{-7}) * (\text{CAPITAL})^{0{,}44685} (\text{POPULACAO})^{1{,}31167}$. Nesta análise os coeficientes de CAPITAL e POPULAÇÃO são, respectivamente, 0,4469 e 1,3117.

Pelo modelo acima, as elasticidade produção/trabalho e produção capital são, aproximadamente, 1,31 e 0,45. Sua soma (1,76) sugere que a economia brasileira, no período 1970-2009, exiba retornos de escala crescentes. No entanto, deve-se testar de 1,76 é estatisticamente diferente de 1. Para verificar isso, considere o modelo 2 (modelo restrito) a seguir.

Modelo 2

É o modelo restrito, no qual PIB e População estão divididos por Capital e um modelo de regressão simples é ajustado. Neste modelo está implícita a restrição $\beta_2 = 1- \beta_3$ e o modelo reproduz a equação (12).

Regression Analysis: C22 versus C23

```
The regression equation is
C22 = - 7,17 + 0,745 C23


Predictor     Coef  SE Coef      T      P
Constant   -7,168    1,916  -3,74  0,001
C23        0,7447   0,1600   4,65  0,000


S = 0,167932   R-Sq = 36,3%   R-Sq(adj) = 34,6%


Analysis of Variance

Source           DF       SS       MS      F      P
Regression        1  0,61108  0,61108  21,67  0,000
Residual Error   38  1,07165  0,02820
Total            39  1,68273
```

A equação do modelo ajustado é:
(PIB/CAPITAL) = (0,00077*(POPULACAO/CAPITAL)^{0,7447}

O coeficiente de POPULAÇÃO é $\beta_3 = 0,7447$ e por $\beta_2 = 1- \beta_3$ segue que o coeficiente de capital é $\beta_2 = 0,2553$.

A soma do quadrado dos resíduos do modelo com restrições é: $RSS_{COM} = 1,07165$ (com 38 graus de liberdade). Do modelo 1 vemos que a soma do quadrado dos resíduos do modelo irrestrito é $RSS_{SEM} = 0,0391$ com 37 graus de liberdade.

Portanto, a estatística F é:

$$F = \frac{(RSS_{COM} - RSS_{SEM})/m}{RSS_{SEM}/(n-k)} = \frac{1,07165 - 0,0391}{0,0391/(40-3)} = \frac{(37)(1,0326)}{0,0391} = \frac{38,2044}{0,0391} = 977,09$$

F deve ser comparado com um percentil apropriado da distribuição F(1, 37) e é, sem qualquer dúvida, significativo.

Qual a conclusão? A hipótese nula $\beta_2 = 1- \beta_3$ (alternativamente $\beta_2 + \beta_3 = 1$, o que significa retornos constantes de escala) é fortemente rejeitada neste período para a economia brasileira, e os dados indicam que há retornos crescentes de escala.

Modelo 3 (Irrestrito)

É semelhante ao modelo 1, mas agora usamos a PEA ao invés da população total. Note que o número de observações disponíveis caiu para 28 (período entre 1982 e 2009).

Regression Analysis: log_PIB versus log_K; log_PEA

```
The regression equation is
log_PIB = - 8,51 + 0,292 log_K + 1,06 log_PEA


28 cases used, 12 cases contain missing values


Predictor      Coef  SE Coef      T      P
Constant   -8,5146   0,9060  -9,40  0,000
log_K      0,29156  0,08189   3,56  0,002
log_PEA    1,05825  0,09231  11,46  0,000


S = 0,0438860   R-Sq = 96,0%   R-Sq(adj) = 95,7%


Analysis of Variance

Source           DF       SS       MS       F      P
Regression        2  1,16293  0,58146  301,91  0,000
Residual Error   25  0,04815  0,00193
Total            27  1,21108


Source   DF   Seq SS
log_K     1  0,90979
log_PEA   1  0,25313
```

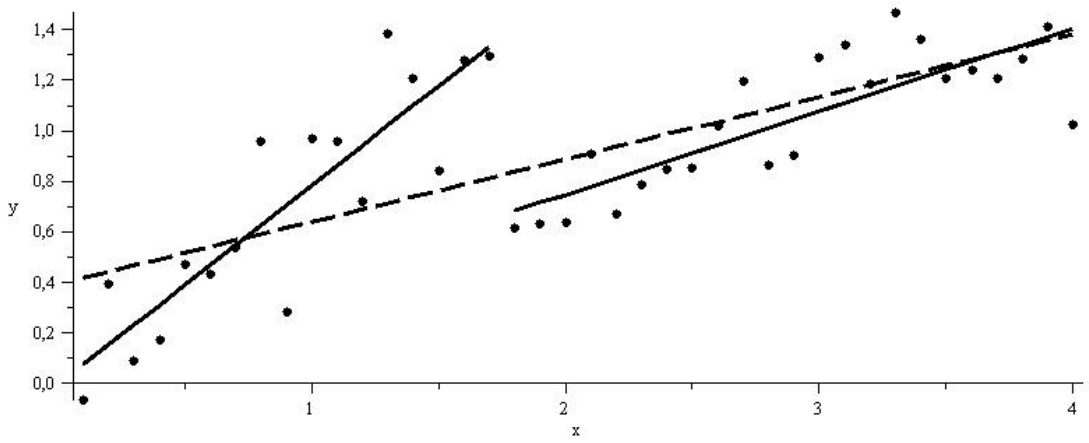
A equação do modelo irrestrito mostrada acima nos diz que o modelo ajustado é:
PIB = (0,0002)*(CAPITAL)^{0,2916}(PEA)^{1,0583} . Nesta análise os coeficientes de CAPITAL e POPULAÇÃO são, respectivamente, 0,2916 e 1,0583.

PARA CASA:
Refaça a análise mostrada nos exemplos anteriores usando agora a PEA como indicador para a mão de obra e teste a hipótese de retornos constantes de escala.

5 – O TESTE DE CHOW – ESTABILIDADE DOS PARÂMETROS NA REGRESSÃO

O teste de Chow é um teste econométrico para verificar se os coeficientes em duas regressões lineares feitas em dois conjuntos de dados diferentes são iguais. Este teste é muito usado na análise de séries temporais para identificar a existência de uma "quebra" estrutural na série.

Na figura a seguir (fonte: Wikipedia), existe uma quebra estrutural em x = 1,7. As duas regressões (de x – 0 até x = 1,7 e de x=1,7 até x =4 se ajustam melhor aos dados que uma única reta de regressão, mostrada pela linha tracejada).



Quebras estruturais podem ser causadas por inúmeros fatores, forças externas (como as crises do petróleo em 1973 e 1979) e internas (como mudanças nas taxas de juros de um país).

A questão que nos diz respeito é: como podemos identificar estatísticamente uma mudança estrutural? A idéia é bastante simples.

Suponha que sabemos que a mudança estrutural aconteceu num certo instante T. Podemos dividir os dados da amostra em 2 blocos: do início até T-1 e de T até o final. Também podemos usar toda a amostra no nosso modelo, e isso equivale a ignorar a quebra estrutural. Então temos 3 regressões:

Regressão 1 – n_1 pontos – início da amostra até T-1

O modelo é: $Y_t = a_1 + a_2.X_t + e_{1t}$ (13)

Regressão 2 – n_2 pontos – do ponto T até o final da amostra

O modelo é: $Y_t = b_1 + b_2.X_t + e_{2t}$ (14)

Regressão 3 – $n_1 + n_2$ pontos – usa TODA a amostra

O modelo é: $Y_t = \alpha_1 + \alpha_2.X_t + e_{3t}$ (15)

A equação (15) (que usa a amostra inteira) pressupõe que não existe diferença entre os dois sub-períodos e então não haveria mudança estrutural, o que significa: $\alpha_1 = a_1 = b_1$ e $\alpha_2 = a_2 = b_2$.

As regressões (13) e (14) pressupõem que os parâmetros são diferentes nos dois períodos.

Como testar isso?

O teste de Chow pressupõe que os erros, além de independentes entre os períodos, são, dentro de cada período iid Normais com média zero e a MESMA variância σ^2. Esta hipótese de homocedasticidade é potencialmente complicada na prática, e vale a pena fazer um teste F para verificar se as variâncias nos dois sub-grupos são iguais.

Como fazer um teste de Chow

1- Calcule a regressão para a amostra inteira. Obtenha a soma do quadrado de resíduos do modelo restrito, que supõe que os coeficientes são iguais nos dois subgrupos (RSS_{COM_RESTR}), que tem n_1+n_2-k graus de liberdade.

2- Calcule a regressão para o primeiro sub-grupo. Obtenha a soma do quadrado dos resíduos (RSS_1), que tem n_1 – k graus de liberdade.

3- Calcule a regressão para o segundo sub-grupo. Obtenha a soma do quadrado dos resíduos (RSS_2), que tem n_2 – k graus de liberdade.

4- Como os dois sub-grupos são encarados como amostras independentes, as respectivas somas de quadrados podem ser somadas, e levam ao cálculo de uma soma de quadrados de resíduos "sem restrições" (pois nela não se está impondo que os coeficientes nos dois períodos sejam iguais). Assim: $RSS_{SEM\_RESTR} = RSS_1 + RSS_2$ com n_1+n_2-2.k graus de liberdade.

5- Se NÃO HÁ mudança estrutural, então RSS_{COM_RESTR} e RSS_{SEM_RESTR} devem ser essencialmente o mesmo valor. Então usa-se um teste F que leva em conta a diferença entre estas somas de quadrados:

$$F = \frac{\left(RSS_{COM\_RESTR} - RSS_{SEM\_RESTR}\right)/k}{\left(RSS_{SEM\_RESTR}\right)/\left(n_1 + n_2 - 2k\right)} \qquad (16)$$

Sob a hipótese nula (não há quebra estrutural, há estabilidade dos parâmetros), esta estatística F tem densidade F(k, n_1+n_2-2.k).

Rejeita-se a hipótese nula (isto é, indica-se que houve quebra estrutural) se a estatística F for maior que o percentil apropriado da densidade F(k, n_1+n_2-2.k).

Perigos no uso do teste de Chow

1- Teste se as variâncias dos erros das duas regressões nos sub-períodos são iguais, pois esta é uma premissa básica do teste.

2- O teste de Chow aponta que as 2 regressões são diferentes, mas não diz porque. O que causou a diferença? O coeficiente linear ou o coeficiente angular das retas, ou os dois?

3- O teste de Chow pressupõe que você saiba em que ponto ocorreu a quebra estrutural – se você não sabe quem é este ponto, o teste não vai apontá-lo para você.