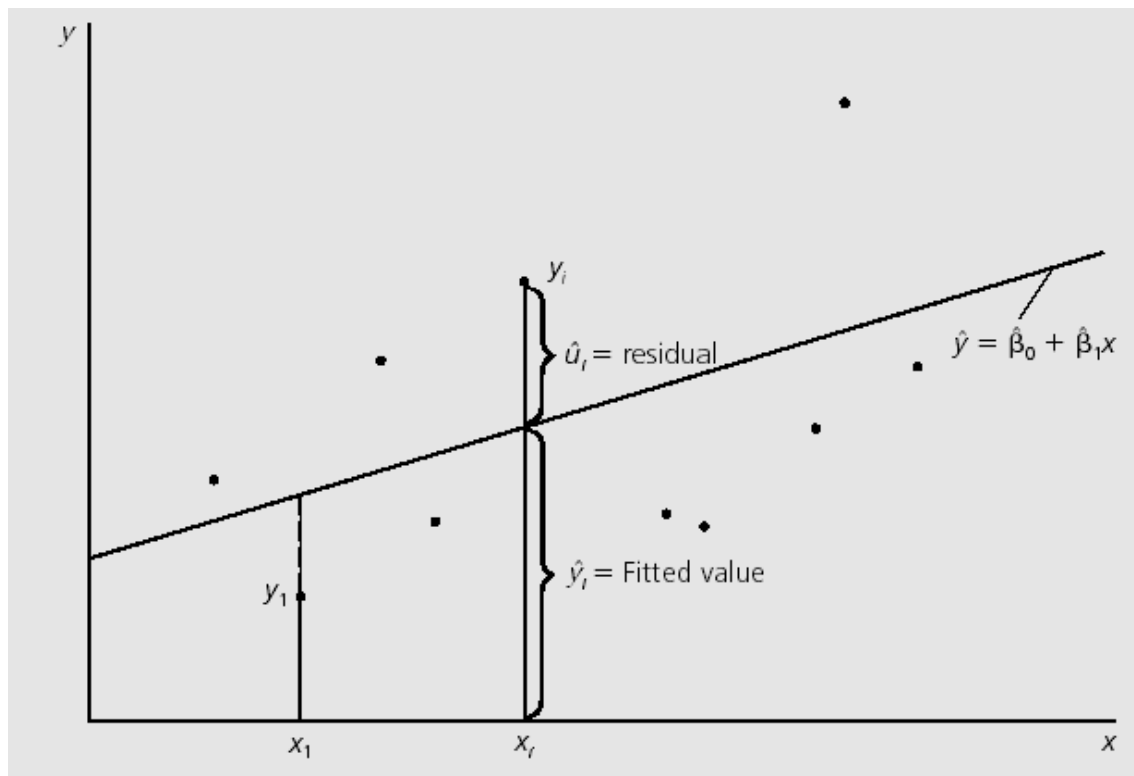


# ***ECONOMETRIA USANDO O STATA***



HENRIQUE D. NEDER – PROFESSOR ASSOCIADO INSTITUTO DE ECONOMIA –  
UNIVERSIDADE FEDERAL DE UBERLÂNDIA

## Sumário

Prefácio .....	- 3 -
1. Operações básicas no STATA.....	- 4 -
2. Regressão Linear Simples .....	- 20 -
3. Relações não lineares.....	- 31 -
4. Uma rotina de simulação de regressões utilizando re-amostragem .....	- 36 -
5. Introdução a regressão múltipla .....	- 39 -
6. O significado dos coeficientes beta (padronizados).....	- 44 -
7. Multicolinearidade .....	- 45 -
8. Apresentação de resultados de regressões no STATA .....	- 49 -
9. O teste Wald.....	- 50 -
10. Testes de hipóteses conjuntas .....	- 54 -
11. Calculando resíduos, valores preditos e predições por intervalos .....	- 54 -
12. Especificando a forma funcional .....	- 57 -
13. Erros não i.i.d.....	- 62 -
14. Regressão com variáveis dummies .....	- 67 -
15. Regressão com Variáveis Instrumentais .....	- 79 -
16. Método dos Momentos Generalizados (GMM) .....	- 98 -
17. Simulação .....	- 113 -
Referencias bibliográficas .....	- 119 -

## Índice de Figuras

Figura 1 – Ativação do editor de dados .....	- 5 -
Figura 2 – Janela do editor de dados.....	- 5 -
Figura 3 – Ativação de um comando de acordo com a ultima ação executada via menu .....	- 7 -
Figura 4 – Histograma de frequências para a variável price .....	- 9 -
Figura 5 – Construção de um gráfico de barras através do menu do Stata .....	- 12 -

Figura 6 – Gráfico de barras para médias de duas variáveis e grupos - 13 -	
Figura 7 – Gráfico de barras por seqüência de comandos no do-file editor ..... - 14 -	
Figura 8 – Histograma dos resíduos da regressão ..... - 22 -	
Figura 9 – Obtenção do numero total de amostras de mesmo tamanho n através do Excel ..... - 35 -	
Figura 10 – Reta de regressão da população e retas de regressão de amostras..... - 39 -	
Figura 11 – Omissão de variáveis relevantes no modelo – verificação do calculo do viés da estimativa ..... - 60 -	
Figura 12 – Teste de diferença de médias para subgrupos populacionais..... - 69 -	
Figura 13 – Regressão OLS da variável endógena (educ) com o instrumento (fatheduc)..... - 89 -	
Figura 14 - Diagrama de Decisão para escolha do método de estimação..... - 99 -	

## **Prefácio**

Este livro pretende explicar econometria através de um software (Stata) que vem sendo utilizado intensivamente por pesquisadores no país e em todo o mundo. São tratados diversos tópicos básicos e alguns mais avançados permitindo ao leitor um aprendizado essencialmente aplicado da disciplina. Na verdade, baseia-se bastante em um conjunto de obras teóricas e aplicadas relacionadas ao uso deste software. Apesar de adotar um enfoque bastante aplicado, em muitas de suas passagens são discutidos aspectos da teoria econométrica, principalmente aqueles que são essenciais para cada modelo e metodologia.

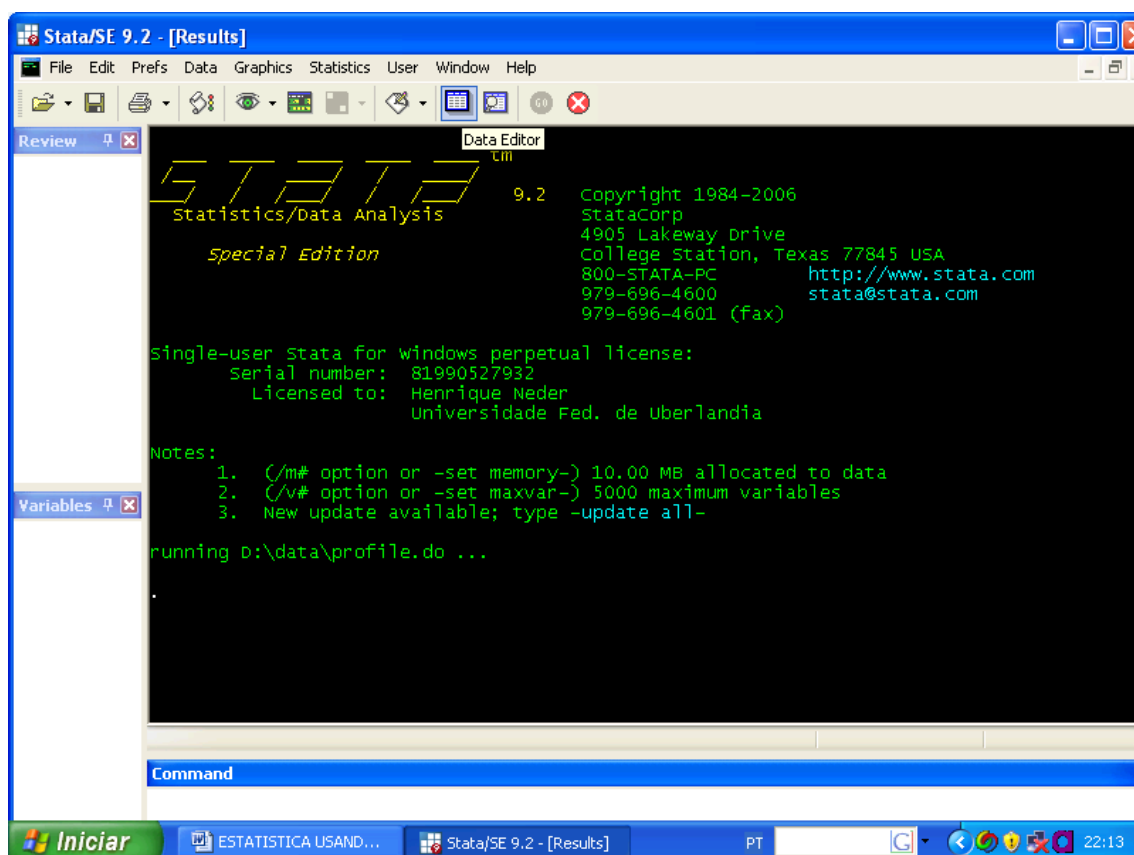
Baseei-me também em minha experiência de alguns anos ensinando econometria em laboratório, para cursos de graduação e pós-graduação na Universidade Federal de Uberlândia. Pude constatar que muitos alunos interessaram-se pelo estudo nesta área, apesar de anteriormente terem passado por inúmeras dificuldades de aprendizado. Muitas vezes a aridez dos métodos e conceitos é apresentada de uma forma mais intuitiva, colaborando para uma motivação maior e ao mesmo tempo garantindo uma compreensão mais adequada da complexidade do tema. Como estratégia didática, este é um

bom ponto de partida para uma motivação no aprendizado, para em seguida partir para abordagens mais aprofundadas.

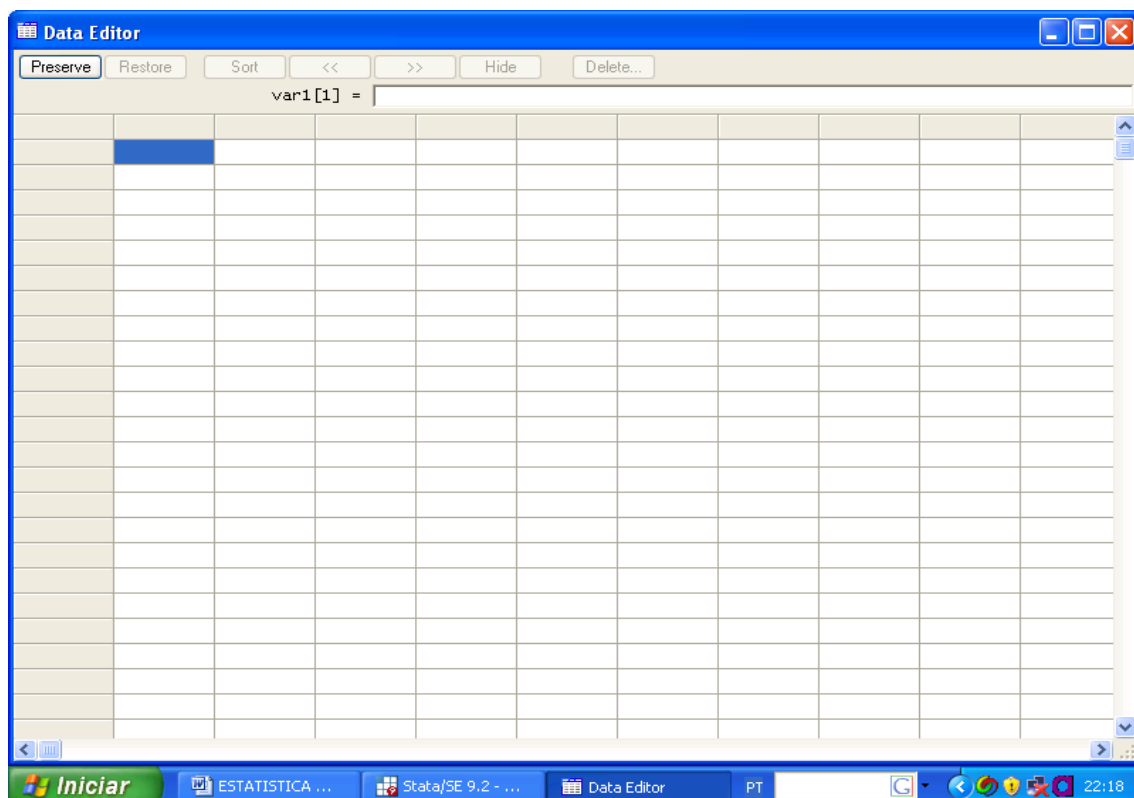
Vimos observando uma elevação da demanda e interesse por aplicações econométricas, tanto porque as possibilidades de uso tornam-se mais freqüentes, em função da melhoria dos métodos de coleta de informações, como também pela utilização generalizada de recursos computacionais, que facilitam enormemente estas aplicações e sua interpretação. Além disto, a maior disponibilidade de recursos e materiais de aprendizado através da internet facilita bastante a replicação de resultados de trabalhos acadêmicos. Tentaremos explorar estas facilidades, possibilitando ao leitor ampliar o uso das diversas técnicas, muitas delas que vão além dos limites deste trabalho.

## **1. Operações básicas no STATA**

Existem diversas formas de introduzir dados no STATA. A primeira delas é através da digitação direta dos mesmos no editor de dados do STATA. Este editor é ativado a partir de um botão, conforme mostra a Figura 1. Com a ativação do editor de dados surge uma nova janela que é uma matriz, cujas linhas representam as observações (casos) e as colunas referem-se às variáveis. Normalmente dados estatísticos são apresentados na forma bruta de um conjunto de indivíduos (que são as observações-linhas) com informações para diversas características (que são as variáveis-colunas). Por exemplo, podemos ter um arquivo de automóveis sendo que as diversas linhas deste arquivo referem-se às marcas de cada automóvel e as colunas são as variáveis (como preço, comprimento, consumo de combustível, comprimento, etc.). Na Figura 2 podemos ver que a janela do editor de dados é uma matriz, com cada célula designada por uma linha e uma coluna. A partir da ativação da janela do editor de dados os mesmos podem ser introduzidos nas células da matriz.



**Figura 1 – Ativação do editor de dados do Stata do Stata**



**Figura 2 – Janela do editor de dados**

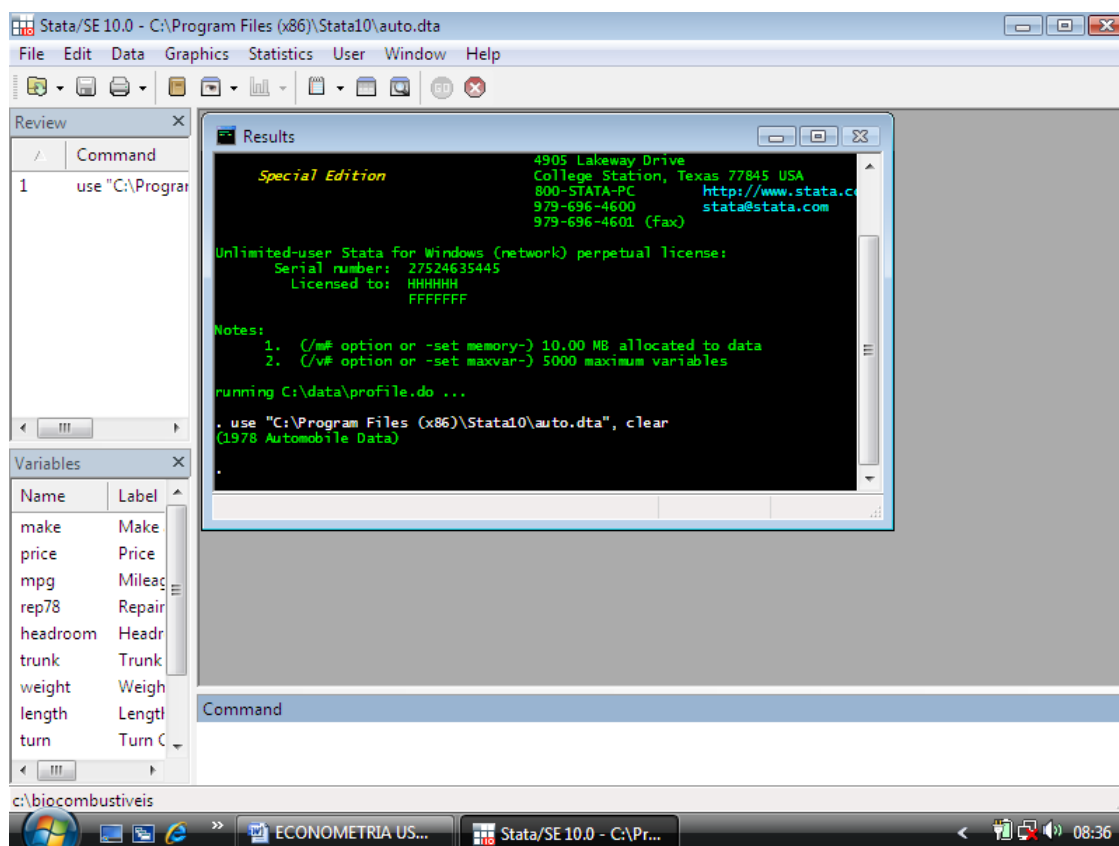
Uma segunda forma de se introduzir dados no STATA é a abertura de arquivos já preparados no formato do software. Estes arquivos de dados têm uma extensão .dta e utilizaremos um arquivo de exemplo que poderá ser encontrado no diretório c:\arquivos de programas\Stata10 denominado auto.dta. Para carregar este arquivo vá até o menu File – Open e busque o arquivo auto neste caminho. Quando o arquivo for carregado aparecerá na janela de resultados a mensagem:

```
. use "C:\Program Files (x86)\Stata10\auto.dta", clear  
(1978 Automobile Data)
```

Toda vez que executarmos uma ação via menus, automaticamente o STATA mostrará o comando correspondente na janela de resultados. O comando **use** é o comando de abertura (carregamento) de arquivos. Poderemos executar a mesma operação de abertura destes mesmos dados através da execução deste comando na janela de comandos. Através da tecla PgUp quando o cursor estiver sobre a janela de comandos irá surgir toda a seqüência de comandos anteriormente executada. Um deles é o comando:

```
. use "C:\Program Files (x86)\Stata10\auto.dta", clear
```

Quando este surgir na janela de revisão (review) localizada na parte superior esquerda, conforme mostra a Figura 3, basta clicar em cima da linha nesta janela que este comando será transferido para a janela de comandos (localizada na parte inferior da tela do Stata) e através da tecla enter os dados poderão ser novamente carregados para a memória.



**Figura 3 – Ativação de um comando de acordo com a ultima ação executada via menu**

Formas alternativas de carregar este arquivo de dados (auto.dta) são através do comando `syuse auto` ou do menu `file – Example Datasets – Example datasets installed with Stata` e escolher o arquivo `auto.dta` clicando no link `use`. Nesta área encontram-se todos os arquivos de exemplo do Stata e outros arquivos que são utilizados nos manuais (Stata documentation).

Observe na Figura 3 que a esquerda fica a janela de variáveis, mostrando os nomes das mesmas que estão no arquivo `auto`. Para iniciar uma operação estatística no Stata ativemos o menu `Statistics` e em seguida o menu `Summaries, Tables & tests` e `Summary statistics` e finalmente `Summary statistics` novamente. Introduza a variável `price` na janela de variáveis e ative o botão `ok`. O comando ativado na janela de resultados do STATA será:

**. summarize price**

Variable	Obs	Mean	Std. Dev.	Min	Max
price	74	6165.257	2949.496	3291	15906

Sem precisar recorrer a uma operação via menus, podemos invocar diretamente o comando summarize através da janela de comandos. Para isto basta digitar nesta janela o comando **summarize price** e digitar enter.

Serão apresentados na tabela o numero de observações (74), a media da variável preço (6165.257), o desvio padrão (2949.496), o valor mínimo (3291) e o valor máximo (15906). Outra opção é executarmos o comando summarize com a opção detail (após a vírgula). Devemos então digitar o seguinte na janela de comandos e apertar a tecla <enter>:

### . summarize price, detail

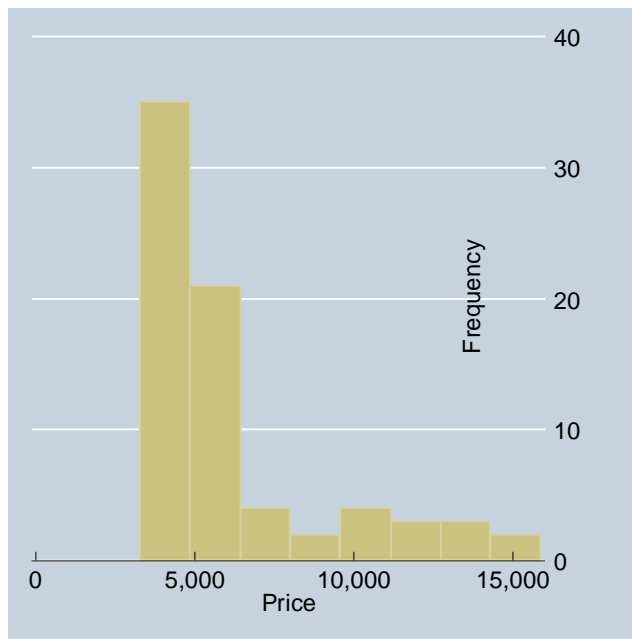
Aparecerá a seguinte tabela na janela de resultados

Price				
Percentiles		Smallest		
1%	3291	3291		
5%	3748	3299		
10%	3895	3667	Obs	74
25%	4195	3748	Sum of Wgt.	74
50%	5006.5		Mean	6165.257
		Largest	Std. Dev.	2949.496
75%	6342	13466		
90%	11385	13594	Variance	8699526
95%	13466	14500	Skewness	1.653434
99%	15906	15906	Kurtosis	4.819188

Na janela de resultados serão apresentados diversos percentis e outras estatísticas tais como a variância, a assimetria (skewness) e a curtose. O Percentil 75 % (ou terceiro quartil) da variável preço é igual a 6342, o que significa que 75 % das observações têm preço menor ou igual a este valor. Já o percentil 10 (ou primeiro decil) é igual a 3895 indicando que 10 % das observações têm preço menor ou igual este valor. Neste exemplo, a assimetria da variável preço é igual a 1,6534 indicando que a distribuição do preço tem uma assimetria positiva, o que quer dizer que temos uma concentração de frequência de marcas de automóveis com preços mais reduzidos.



Esta última informação pode ser confirmada graficamente através do menu graphics, histogram, selecionar a variável **price** na janela variables, indicar a opção frequencies na janela y-axis e ativando o botão ok. O STATA criará o gráfico do histograma para a variável **price**. Podemos observar, confirmando a informação de um coeficiente de assimetria positivo, que o histograma apresenta as maiores freqüências absolutas em valores mais reduzidos para a variável.



**Figura 4 – Histograma de freqüências para a variável price**

Observa-se no histograma acima uma elevada concentração de freqüência para preços mais reduzidos (em torno de 5000) e uma baixa concentração de freqüências para valores elevados da variável price. Isto confirma o valor positivo para o coeficiente de assimetria.

Outra forma de obterem-se valores de estatísticas descritivas no STATA é através do comando tabstat. Antes de mostrarmos como funciona este comando é importante notar que para qualquer comando podemos pedir uma ajuda para o STATA com o comando help. Para isto digitaremos na linha de comando:

**. help tabstat**

Aparecerá uma janela com a sintaxe do comando:

```
tabstat varlist [if] [in] [weight] [, options]
```

Esta mensagem está indicando a sintaxe do comando tabstat. Nesta sintaxe, o que tiver com cor preta é palavra obrigatória no comando. O termo varlist indica que após a palavra obrigatória temos que colocar uma lista de variáveis. Em seguida a sintaxe indica uma serie de opções (não obrigatórias) que aparecem entre colchetes. Devemos digitar a palavra obrigatória tabstat seguida do nome da(s) variável(eis) com a possibilidade de introduzirmos um filtro condicional por if ou in, designação de pesos e outras opções. Veremos adiante detalhadamente como utilizar os filtros do arquivo para cálculos e sobre a utilização de pesos. Neste visor do **help** aparecem inúmeras opções que o usuário do STATA irá com o passar do tempo utilizando com maiores detalhes e aperfeiçoando o comando de acordo com as suas necessidades. Por exemplo, poderemos digitar na linha de comandos:

```
. tabstat price if foreign == 1, s(count min max mean sd cv sk p0 p50 p75 p90)
```

Aparece na janela de resultados do Stata:

variable	N	min	max	mean	sd	cv	skewness
price	22	3748	12990	6384.682	2621.915	.4106571	1.215236

variable	p10	p50	p75	p90
price	3895	5759	7140	9735

Podemos também solicitar as estatísticas para diversas variáveis com:

```
. tabstat price length weight if foreign == 1, s(count min max mean sd cv sk p0 p50 p75 p90)
```

stats	price	length	weight
N	22	22	22
min	3748	142	1760
max	12990	193	3420
mean	6384.682	168.5455	2315.909
sd	2621.915	13.68255	433.0035
cv	.4106571	.0811802	.1869691
skewness	1.215236	.0809646	1.056582

p10		3895	154	1930
p50		5759	170	2180
p75		7140	175	2650
p90		9735	189	2830

---

Verificamos que o coeficiente de assimetria (de Pearson) para a variável price é igual a 1.2152, indicando uma assimetria levemente positiva para a distribuição desta variável. A variável length (comprimento do carro) já tem uma assimetria um pouco menos acentuada que a variável price. O coeficiente de variação da variável price (0.4106) é maior do que o coeficiente de variação da variável length (0.0809) indicando que a primeira tem uma distribuição com dispersão relativa mais elevada que a segunda. Geralmente variáveis com coeficiente de variação mais elevados tem assimetrias mais acentuadas, mas isto não ocorre sempre. Podemos encontrar casos de distribuições com elevada dispersão relativa (coeficiente de variação), mas com baixa assimetria.

O valor do percentil 90 para a variável price é igual 9735 significando que temos 90 % dos automóveis com preços inferiores a este valor.

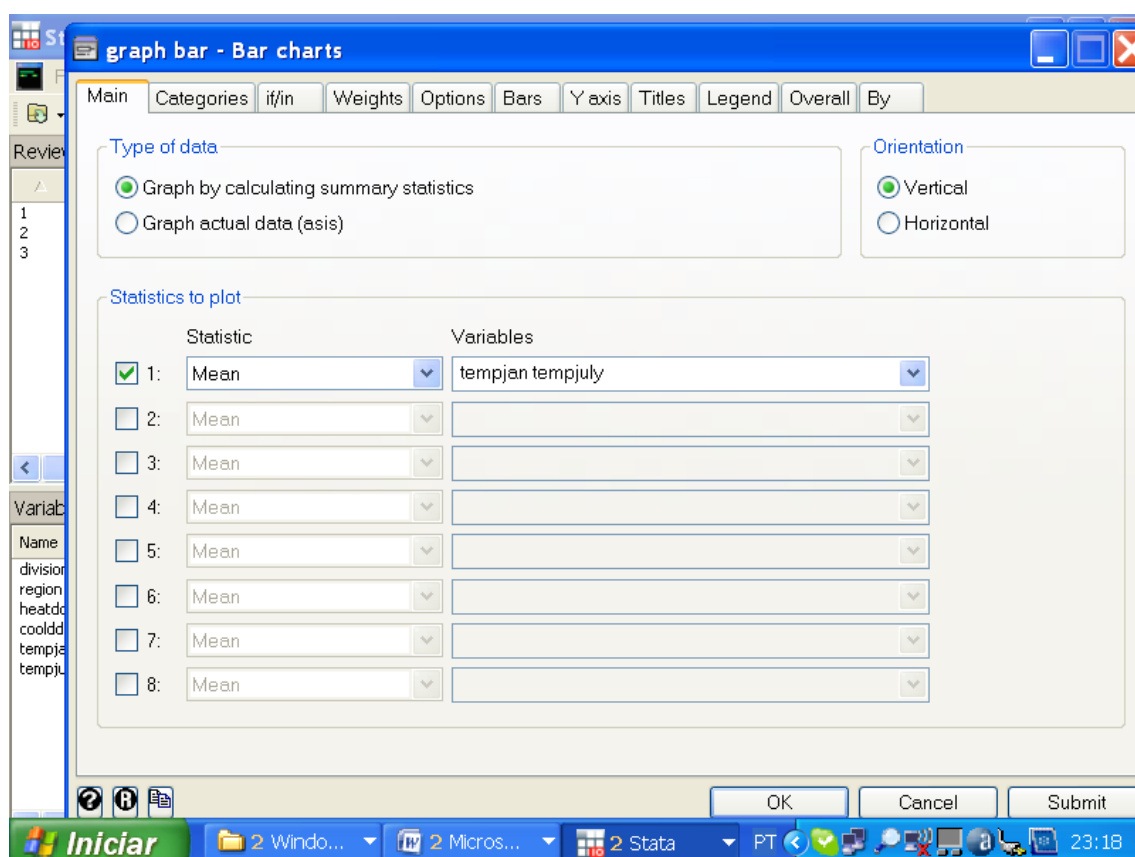
Vejamos introdutoriamente alguns dos recursos gráficos do STATA. Escreva na janela de comandos:

**. sysuse citytemp**

Vamos construir um gráfico de barras para a temperatura mensal por região.

Vá ao menu Graphics => Bar chart => summary statistics

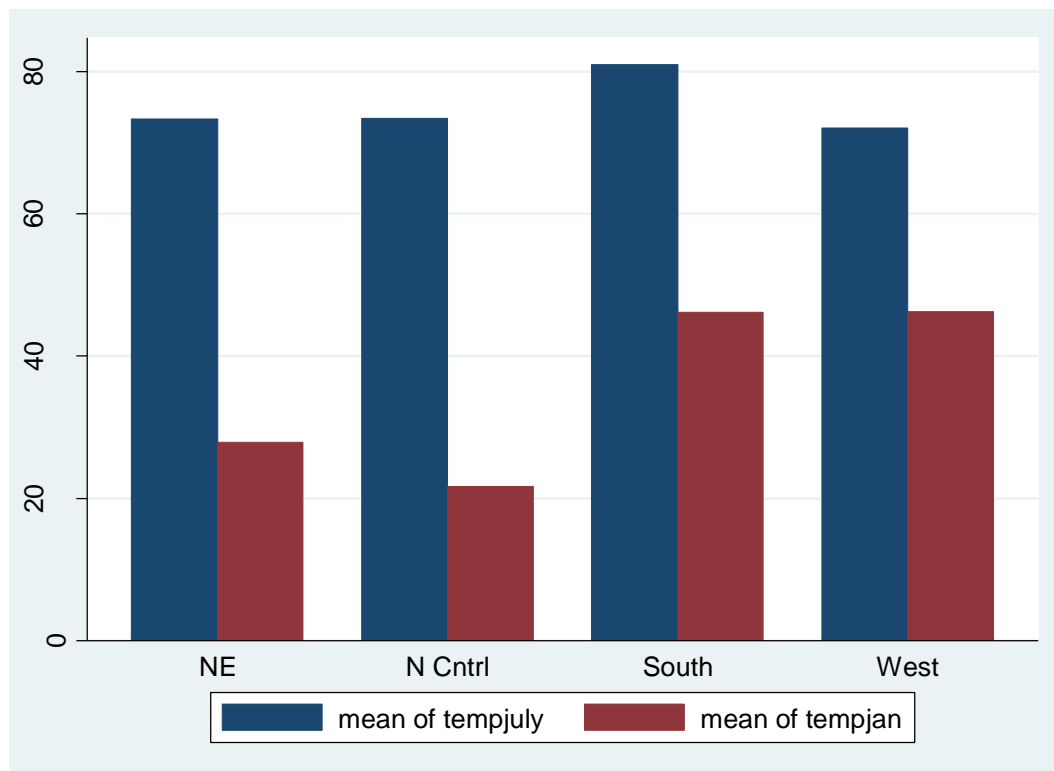
Coloque na janela de variáveis tempjuly tempjan



**Figura 5 – Construção de um gráfico de barras através do menu do Stata**

Na aba da janela Categories marque a opção Group 1 e selecione a variável de grupo **region**

Aparecerá o seguinte gráfico:



**Figura 6 – Gráfico de barras para médias de duas variáveis e grupos**

A propósito, para transferir um gráfico gerado pelo STATA no Word, proceda da seguinte forma:

- 1) Após gerar o gráfico no STATA salve o mesmo utilizando o menu File => Save graph e salve como tipo:

EPS with TIFF preview (\*.eps)

que é uma forma de transporte de gráfico (encapsulated). Esta forma é facilmente importada no Word através do menu Inserir => Imagem.

Mas existe uma outra forma de transportar gráficos do Stata para o Word. Após gerar o gráfico no Stata, vá ao menu Edit => Copy Graph, passe para o Word, localize o cursor onde deseja copiar o gráfico e cole o gráfico.

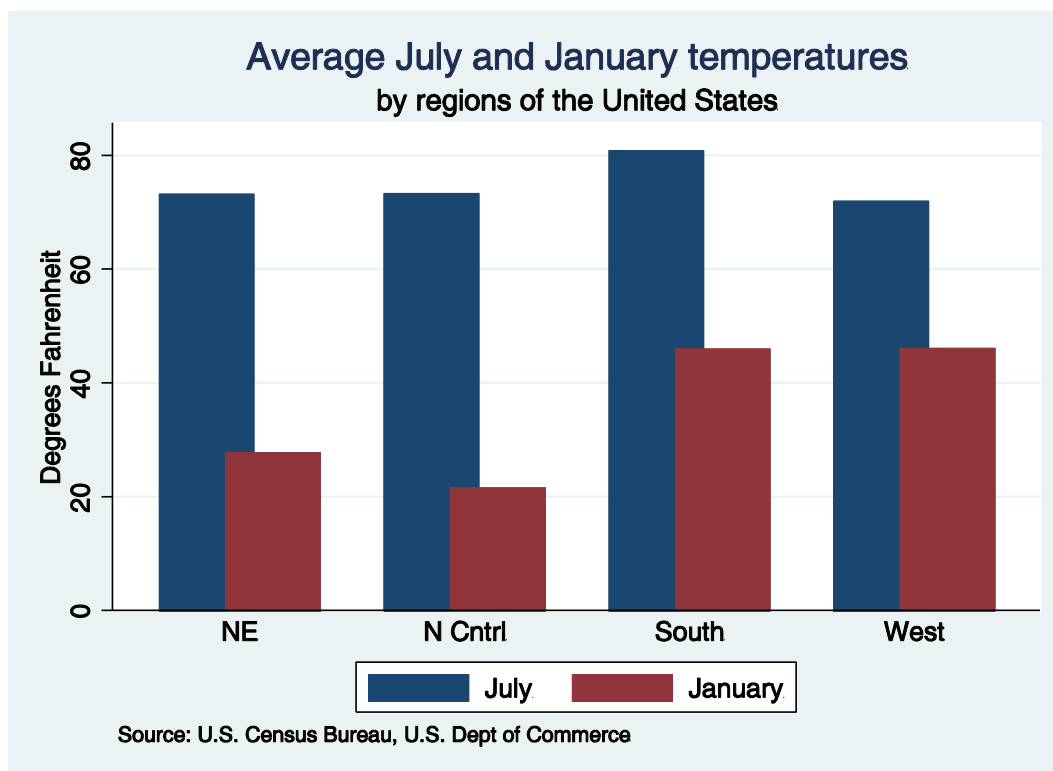
No manual de gráficos do STATA 9.2 temos um exemplo de gráfico mais detalhado. Abra o editor de do file (Window => Do-file editor) e escreva o seguinte:

```
graph bar (mean) tempjuly tempjan, over(region)           ///  
    bargap(-30)                                           ///  
    legend( label(1 "July") label(2 "January"))           ///  
    ytitle("Degrees Fahrenheit")                          ///  
    title("Average July and January temperatures")        ///  
    subtitle("by regions of the United States")            ///  
    note("Source: U.S. Census Bureau, U.S. Dept of Commerce")
```

Os caracteres ///  
 servem para indicar que o comando continua na próxima linha. Isto é útil quando trabalhamos com comandos longos (no do-file editor) garantindo que ele seja executado como um único comando. Neste comando estamos definindo varias opções para o gráfico, tais como a legenda, o título do eixo y (eixo vertical), o título do gráfico, um sub-título e uma nota do gráfico.

Execute esta seqüência de linhas utilizando o menu do editor de do-file tools => do to botton ou selecione a seqüência de linhas e tools => do Selection

Surgirá na tela o seguinte gráfico:



**Figura 7 – Gráfico de barras por seqüência de comandos no do-file editor**

Outros gráficos importantes para análise estatística também podem ser feitos. Um exemplo é o chamado diagrama de dispersão (scatterplot).

```
. sysuse auto
. twoway (scatter weight price)
```

Podemos colocar labels (etiquetas) considerando que a variável que dá nome aos carros é make:

```
. twoway (scatter weight price, mlabel(make))
```

Este diagrama de dispersão mostra como é a relação bivariada entre as duas variáveis. De uma forma analítica esta relação bivariada pode ser avaliada pelo coeficiente de correlação.

```
. correlate price weight
```

Veja que o coeficiente de correlação (neste caso, o de Pearson) é 0,5386 indicando uma relação linear relativamente forte e positiva entre as duas variáveis.

Podemos também construir uma matriz de correlação, utilizando o mesmo comando correlate:

```
. correlate price mpg rep78 weight length
```

	price	mpg	rep78	weight	length
price	1.0000				
mpg	-0.4559	1.0000			
rep78	0.0066	0.4023	1.0000		
weight	0.5478	-0.8055	-0.4003	1.0000	
length	0.4425	-0.8037	-0.3606	0.9478	1.0000

Observe que o preço e o numero de milhas por galão (mpg) tem correlação negativa (-0.4559) como esperada. Já peso (weight) e comprimento (length) têm correlação altamente positiva. Observe que os resultados os parecem abaixo da diagonal principal porque a matriz de correlação é simétrica.

Vamos aprender um pouco da capacidade de calculo matricial do STATA. Para ter uma idéia geral destes recursos vá para o menu Help => Contents => Programming and matrices => Matrices => Summary of matrix

comands. Você verá que temos um grande numero de comandos para matrizes. Vamos entrar no “Inputting matrices by hand” (introduzindo matrizes manualmente) que corresponde ao comando matrix define. Clicando no termo em azul **matriz define** aparecerá o help para este comando com a sua estrutura básica (sintaxe) e inúmeros exemplos de sua utilização. Por exemplo, digite na janela de comandos:

```
. matrix input mymat = (1,2\3,4) e aperte a tecla enter  
. matrix list mymat
```

Aparecerá a matriz digitada na janela de resultados.

Agora vamos calcular a inversa desta matriz:

```
. matrix B = inv(mymat)  
. matrix list B
```

O comando matrix list simplesmente lista a matriz B na janela de resultados.

```
. matrix C = mymat*B  
. matrix list C
```

Como não poderia ser de outra forma, a matriz C é a matriz identidade.

Agora vamos resolver um sistema de equações lineares no Stata:

$$\begin{aligned} 3x + 7y - 2z &= 3 \\ x - 2y + z &= 1 \\ 2x + 3y - 4z &= -4 \end{aligned}$$

Resolvendo por Laplace:

```
. matrix A = (3,7,-2\1,-2,1\2,3,-4)  
. matrix A1 = (3,7,-2\1,-2,1\ -4,3,-4)  
. matrix A2 = (3,3,-2\1,1,1\2,-4,-4)  
. matrix A3 = (3,7,3\1,-2,1\2,3,-4)  
. scalar X = det(A1)/det(A)  
. scalar Y = det(A2)/det(A)  
. scalar Z = det(A3)/det(A)  
. disp X, Y, Z
```



O comando scalar X = expressão calcula um valor escalar de acordo com a expressão e armazena em uma localização de memória chamada X. E se o sistema for indeterminado, o que irá acontecer? É o caso em que o determinante da matriz A é igual a zero quando as colunas ou linhas da matriz A são linearmente dependentes. Por exemplo, se tivermos o seguinte sistema:

$$\begin{aligned} -x + 0,5y - 1z &= 3 \\ 2x - 0,5y + 1,5z &= 1 \\ x + 1,5y - 1z &= -4 \end{aligned}$$

Repare que a coluna 1 é igual a soma do dobro da coluna 2 e do dobro da coluna 3, sendo portanto a coluna 1 uma combinação linear das colunas 2 e 3 e desta forma as colunas 1, 2 e 3 são vetores linearmente dependentes.

```
. matrix A = (-1,0.5,-1\2,-0.5,1.5\1,1.5,-1)  
. disp det(A)
```

Como o determinante de A é nulo o sistema fica indeterminado. Já que iremos tratar do tema econometria suponhamos que desejamos fazer uma regressão a partir dos dados do arquivo de automóveis:

```
. sysuse auto  
. regress price mpg weight length foreign
```

Source	SS	df	MS	Number of obs = 74		
Model	348708940	4	87177235	F( 4, 69)	=	21.01
Residual	286356456	69	4150093.57	Prob > F	=	0.0000
Total	635065396	73	8699525.97	R-squared	=	0.5491
				Adj R-squared	=	0.5230
				Root MSE	=	2037.2

	price	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
mpg		-13.40719	72.10761	-0.19	0.853	-157.2579	130.4436
weight		5.716181	1.016095	5.63	0.000	3.689127	7.743235
length		-92.48018	33.5912	-2.75	0.008	-159.4928	-25.46758
foreign		3550.194	655.4564	5.42	0.000	2242.594	4857.793
_cons		5515.58	5241.941	1.05	0.296	-4941.807	15972.97

Vamos tentar chegar aos resultados anteriores utilizando álgebra linear:

```
. gene const = 1
. mkmat const mpg weight length foreign, matrix(X)
. matrix list X
. mkmat price, matrix(Y)
. matrix B = inv(X'*X)*X'*Y
. matrix list B
```

Na lista de comandos anteriores o comando `gen constant = 1` gera uma variável com valor constante igual a 1, o segundo comando gera uma matriz com as colunas sendo as variáveis `mpg`, `weight`, `length` e `foreign`, ou seja, as variáveis explicativas do modelo de regressão. Este comando monta uma matriz `X` a partir dos dados do STATA. O terceiro comando (`matrix list X`) lista a matriz `X`. O quarto comando monta uma matriz `Y` a partir dos dados da variável `price`. O quarto comando é a equação matricial para o cálculo das estimativas de mínimos quadrados ordinários resultando em um vetor `B`. Compare os resultados deste cálculo matricial com os resultados do comando `regress`. Vamos agora confirmar um desenvolvimento que está no apêndice E do Wooldridge. Neste apêndice o autor desenvolve a formulação de álgebra linear para o modelo de regressão múltipla. A obtenção dos resíduos da equação é feita através da relação:

$$\hat{u} = Y - X\hat{\beta}$$

```
. matrix u = Y-X*B
. matrix list u
```

Observe que os valores desta matriz são praticamente nulos, o que é esperado pela teoria da regressão. De acordo com Wooldridge isto significa que os resíduos de mínimos quadrados ordinários sempre somam zero quando um intercepto é incluído na equação. Verifiquemos esta afirmação com o seguinte comando:

```
. svmat u, name(u)
```

Este último comando transforma o vetor de resíduos `u` em uma variável `u1`.

```
. tabstat u1, s(sum)
```

Verifique que a soma dos resíduos é igual a -.0002327, diferente de zero devido a um erro de aproximação de cálculos internos do STATA. Vamos verificar uma outra afirmação do mesmo Wooldridge neste apêndice, pagina 678: a covariância amostral entre cada variável independente e os resíduos é igual a zero.

**. correlate mpg weight length foreign u1, covariance**

	mpg	weight	length	foreign	u1
mpg	33.472				
weight	-3629.43	604030			
length	-102.514	16370.9	495.79		
foreign	1.04739	-212.029	-5.84265	.211773	
u1	-.000021	-.000727	-.00002	-9.3e-07	3.9e+06

Observe os valores da ultima linha da matriz acima: todos os seus valores são praticamente nulos confirmando a informação anterior. Como o STATA realiza cálculos estatísticos? Existem diversos tipos de funções, entre elas as funções estatísticas. Vá ao menu help => Contents => Basics => Syntax => Expressions and Functions => Functions => Probability distributions and density functions. Você encontrará uma lista de funções entre elas: normal(z) e invnormal(p). Suponhamos que queiramos determinar qual é o valor acumulado de probabilidade para a função densidade de probabilidade normal padrão na abscissa  $z = 1.34$ .

**. disp normal(1.34)**  
**.90987733**

Agora, para conferir vamos determinar qual é o valor da variável normal padrão  $z$  que deixa uma probabilidade acumulada até – infinito de 0,90987733.

**. disp invnormal(.90987733)**  
**. 1.34**

Vamos agora calcular o valor da seguinte expressão:

$$Y = \ln(525^2 - 21^{2/3}) - \arctg(-2,33)$$

```
. disp "Y =", ln(525^2- 21^(2/3)) - atan(-2.33)  
. Y = 13.692156
```

Desejamos calcular o valor de probabilidade de 100 ou mais sucessos em uma distribuição binomial com 300 tentativas e com probabilidade de sucesso = 0,34.

```
. disp "prob = ", Binomial(300,100,.34)  
. prob = .61741607
```

## 2. Regressão Linear Simples

A partir deste ponto iremos utilizar dados do livro do Wooldridge e do Baum<sup>1</sup>. Primeiramente iremos abrir os dados denominados CEOSAL1.DTA do Wooldridge.

```
. use http://fmwww.bc.edu/ec-p/data/wooldridge/CEOSAL1, clear  
. clear
```

O STATA é um pouco “temperamental” quando se trata de abrir um arquivo a partir da Internet. O comando use http: é utilizado para isto. Se houver problemas da execução do comando acima, vá para o menu Edit => Preferences => General Preferences => Internet e preencha os dados do seu servidor de Proxy.

---

<sup>1</sup> Os dados do Baum podem ser baixados através do site <http://www.stata-press.com/data/mus.html> ou através da execução dos 3 comandos, em sequência, na janela de comandos do Stata:

```
net from http://www.stata-press.com/data/mus  
net install mus  
net get mus
```

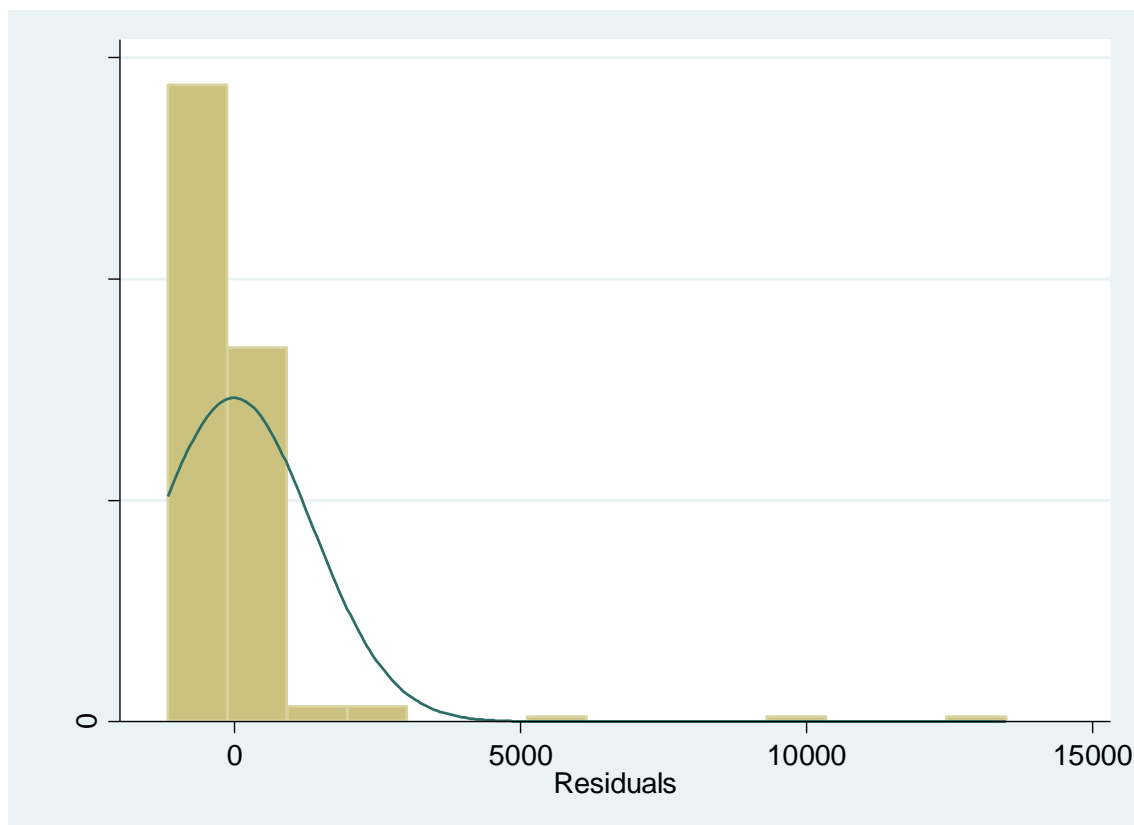
Os dados do Wooldridge podem ser baixados de: [http://websites.swlearning.com/cgi-wadsworth/course\\_products\\_wp.pl?fid=M20bl&flag=instructor&product\\_isbn\\_issn=9780324581621&discipline\\_number=413](http://websites.swlearning.com/cgi-wadsworth/course_products_wp.pl?fid=M20bl&flag=instructor&product_isbn_issn=9780324581621&discipline_number=413)

Neste arquivo a variável salary representa o salário anual em milhares de dólares e a variável roe representa um retorno em percentual (ver pg. 32 Wooldridge).

```
. regress salary roe
. ereturn list
. matrix b = e(b)
. disp b[1,2] + b[1,1]*30
```

Este o valor predito da variável dependente do modelo para um retorno igual 30 %. Isto não quer dizer que uma particular observação que tenha roe = 30 tenha também o valor de salary = 1518,221. Existem muitos outros fatores que afetam o salário além da variável roe. Como calcular a reta de predição da regressão e os valores dos resíduos no STATA?

```
. predict yhat
. predict residuos, residuals
. edit salary roe yhat residuos
. twoway (scatter salary roe) (scatter yhat roe)
. twoway (scatter roe residuos)
. histogram residuos, normal
```



## Figura 8 – Histograma dos resíduos da regressão

Observe que os resíduos desta regressão têm uma distribuição que difere bastante da distribuição normal apresentando valores discrepantes elevados como mostra o histograma em sua parte direita próximo do valor 10000 e 13000.

Vamos agora discutir alguns aspectos teóricos do modelo de regressão linear simples contidos no capítulo 2 do Wooldridge e no capítulo 4 do Baum. Por exemplo, na página 24 do Wooldridge é apresentada a seguinte equação:

$$wage = \beta_0 + \beta_1 educ + u \quad (2.1)$$

Aconselho ao leitor a ir se familiarizando com o uso de um editor de equações (proponho o uso do Mathtype 6.0). Em primeiro lugar vamos fazer a observação de que a linearidade de (2.1) implica que uma mudança unitária em  $x$  tem o mesmo efeito em  $y$ , não importando o valor inicial de  $x$ . Isto não é realístico para muitas aplicações econômicas, que tem relações não lineares. Outro ponto a ser destacado é que  $\beta_1$  mede o efeito de  $x$  sobre  $y$  mantendo os outros fatores (em  $u$ ) fixos. A pergunta que se faz é: como podemos esperar saber em geral sobre o efeito *ceteris paribus* de  $x$  sobre  $y$ , considerando os outros fatores fixos, quando estamos ignorando todos estes outros fatores? Na verdade, só estamos ignorando a existência destes fatores na parte determinística da equação. Eles são representados pela parte aleatória, representada pelo termo de erro  $u$ .

Como veremos adiante, só poderemos ter estimadores confiáveis de  $\beta_0$  e  $\beta_1$  a partir de uma amostra aleatória dos dados quando estabelecemos uma hipótese restritiva a respeito de como  $u$  está relacionada à variável explicativa  $x$ . Sem que tal restrição seja satisfeita, não somos capazes de estimar o efeito *ceteris paribus*  $\beta_1$ . Outra hipótese a ser utilizada é a de que a média do termo de erro é zero. Se o termo de intercepto  $\beta_0$  é incluído na

equação não perdemos nada se assumimos esta hipótese.<sup>2</sup> Matematicamente, ela é definida como:

$$E(u) = 0 \quad (2.2)$$

Podemos sempre redefinir o intercepto na equação (2.1) para fazer com que (2.2) seja verdadeiro. Se  $u$  e  $x$  são não correlacionados, então, como variáveis aleatórias, são não linearmente relacionadas. Mas correlação (coeficiente de correlação de Pearson product-moment) mede apenas a dependência linear entre  $u$  e  $x$ . Por exemplo,  $u$  e  $x$  podem ser não correlacionados, mas  $u$  pode ser correlacionado com funções de  $x$ , como  $x^2$ . De acordo com Wooldridge, esta possibilidade pode causar problemas para alguns casos de regressão, para interpretar o modelo e para derivar suas propriedades estatísticas. Uma melhor hipótese (ao invés de somente  $u$  e  $x$  serem não correlacionados) envolve o valor esperado de  $u$  dado  $x$ , ou seja, um valor esperado condicional. A hipótese agora é a de que o valor médio de  $u$  não depende do valor de  $x$  (ou de um dado valor de  $x$ ). Isto é expresso matematicamente por:

$$E(u | x) = E(u) = 0 \quad (2.3)$$

Esta expressão é denominada hipótese da media condicional nula. Para qualquer valor de  $x$  dado (dado no sentido de condicionado), a média das variáveis não observáveis é a mesma e, portanto precisa ser igual ao valor médio de  $u$  na população inteira. Ela indica que o nível médio das variáveis não observáveis é o mesmo não importando o valor que tomamos para a variável independente  $x$ . No exemplo da equação (2.1) para a relação entre salário e educação, podemos considerar que  $u$  representa o efeito de variáveis não observáveis como habilidades inatas do trabalhador e que independem do seu nível de instrução.

---

<sup>2</sup> Suponhamos que na equação  $y = \beta_0 + \beta_1 x + u$  para a população, o valor médio de  $u$  seja igual a  $\mu \neq 0$ . Então o modelo a ser estimado passa a ser  $y = \beta'_0 + \beta_1 x + u'$ , com  $\beta'_0 = \beta_0 + \mu$  e  $u' = u - \mu$  com  $E[u'] = 0$ . Do que se conclui que sempre haverá um valor do termo constante, de tal forma que  $E[u] = 0$ .

Então se pegarmos os trabalhadores com 10 anos de estudo, estes vão ter em media o mesmo valor de habilidades inatas que os trabalhadores com 15 anos de estudo. Podemos interpretar as médias condicionais como médias para a variável habilidade inata (supondo que esta poderia ser mensurável) para grupos com valores fixos da variável anos de estudo. Esta é então, a hipótese restritiva que se faz nos modelos de regressão. Hipótese que, convenhamos, não traz grandes problemas em termos do realismo deste modelo e de sua adequação. Baum mostra na pagina 72 que esta hipótese pode ser apresentada na forma matricial:

$$\begin{aligned} E[\mathbf{x}'\mathbf{u}] &= \mathbf{0} \\ E[\mathbf{x}'(\mathbf{y} - \mathbf{x}\boldsymbol{\beta})] &= \mathbf{0} \end{aligned} \tag{2.4}$$

As expressões em negrito referem-se a vetores (ou matrizes). A primeira expressão de (2.4) indica que o produto matricial da matriz transposta de  $\mathbf{x}$  pelo vetor dos resíduos é igual ao vetor nulo. Vamos pegar o nosso exemplo que já está na memória do STATA:

```
. gen const = 1
. mkmat const roe, matrix(x)
. mkmat resíduos, matrix(u)
. matrix r = x'*u
. matrix list r
```

A expressão (2.4) está relacionada com um método de estimação (distinto do método de mínimos quadrados, mas que, como veremos, conduz aos mesmos resultados) denominado método dos momentos. Porque se chama método dos momentos. Momento é um conceito originário da Física que representa o produto de uma força por um braço de alavanca perpendicular a esta força. Então, por exemplo, se aplicarmos uma força de 10 kgf sobre um braço de alavanca de comprimento igual a 2 metros, teremos um momento de 20 kgf x m. Em analogia a este conceito físico de momento, os estatísticos definem momento como o produto de dois vetores. Se estes dois vetores forem ortogonais (formarem entre si um ângulo de 90º) então se diz que o momento é igual a 0. Isto ocorre porque o valor escalar de um vetor que é o produto vetorial de dois vetores é dado pela expressão  $|z| = |x||y|\cos\hat{\theta}$  onde  $|z|$  é o



módulo do vetor  $z$  e  $\hat{\theta}$  é o ângulo formado pelos dois vetores  $x$  e  $y$ . Se os vetores  $x$  e  $y$  forem ortogonais  $\cos \hat{\theta} = 0$  e o momento entre eles é nulo<sup>3</sup>.

O método de estimação conhecido como método dos momentos estabelece que os estimadores dos parâmetros da regressão são obtidos de forma que todos os momentos de cada variável independente com o vetor de erros é nulo. Dizer que estes momentos são nulos é o mesmo que dizer que as correlações entre as variáveis independentes e os resíduos são todas nulas, porque quando duas variáveis são representadas na forma vetorial e estes vetores são ortogonais, esta é a situação de correlação nula. Resumindo, a condição de média condicional nula dada pela expressão (2.3) é equivalente a correlação nula entre variáveis independentes e termo de erro da regressão, que por sua vez implica em ortogonalidade entre as variáveis independentes e o termo de erro e isto resulta finalmente em momentos nulos.

É bom advertir que toda esta discussão teórica a respeito de uma importante restrição do modelo de regressão linear refere-se em princípio à população e não a amostra. Os erros devem ser não correlacionados às variáveis independentes no conjunto da população (e não apenas em uma dada amostra). Estamos aqui derivando um método de estimação que será aplicado a sua contraparte amostral. Estabelecemos para a amostra que dispomos uma condição de restrição que é a de que os momentos são todos nulos e fazemos a seguinte pergunta: quais seriam as fórmulas matemáticas para nossos estimadores de parâmetros se impusermos a condição de que os momentos sejam todos nulos?<sup>4</sup>

---

<sup>3</sup> Os estatísticos também utilizam este conceito físico de momento para definir os momentos de uma distribuição. Assim existe o primeiro momento que é a média, o segundo momento que é a variância  $\sigma^2$ , o terceiro momento  $\sigma^3$  e assim indefinidamente. Diz-se que duas distribuições (dadas por duas funções densidade) são equivalentes se todos os momentos de uma distribuição forem iguais aos momentos correspondentes da outra distribuição. Esta é uma idéia mais ampla do que uma simples comparação de médias e variâncias.

<sup>4</sup> Desta forma fica mais inteligível a afirmação anterior (e confirmação através do exemplo com o Stata) de que o método dos mínimos quadrados ordinários (que produz os mesmos resultados de estimativas

Assim para  $k$  variáveis independentes, teríamos  $k$  equações de momento e se estabelecermos a restrição simultânea de que todas estas equações são iguais a zero temos  $k$  condições de momento que resultam em  $k$  estimadores de parâmetros. Estas  $k$  condições de momento são dadas pela expressão

$$E[\mathbf{x}'\mathbf{u}] = \mathbf{0}$$

em que os termos em negrito representam vetores. Em termos escalares, este produto vetorial seria dado por:

$$x_{11}u_1 + x_{12}u_2 + \dots + x_{1n}u_n = 0$$

$$x_{21}u_1 + x_{22}u_2 + \dots + x_{2n}u_n = 0$$

.....

$$x_{k1}u_1 + x_{k2}u_2 + \dots + x_{kn}u_n = 0$$

Estas são as condições de  $k$  momentos. Voltando a formulação vetorial  $\mathbf{x}'\mathbf{u} = 0$  implica em:

$$\mathbf{X}'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \mathbf{0}$$

$$\mathbf{X}'\mathbf{y} - \mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \mathbf{0}$$

$$\boldsymbol{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

que é igual a fórmula para os estimadores de MQO.

Observe que não derivamos esta expressão a partir da esperança matemática do produto vetorial ( $E(\mathbf{x}'\mathbf{u})$ ) mas do argumento do operador esperança  $\mathbf{x}'\mathbf{u}$ . Da mesma forma observe que os valores da matriz  $r$  (resultado da última seqüência de comandos – pg 24) não são os valores de  $E[\mathbf{x}'\mathbf{u}]$ . São os valores de  $\mathbf{x}'\mathbf{u}$ . Acredito que os valores de  $E[\mathbf{x}'\mathbf{u}]$  somente poderiam ser calculados se dispuséssemos de todos os pontos-população e

---

que o método dos momentos) sempre produz resíduos com média zero e não correlacionados com as variáveis independentes. Este é um resultado que fica imposto pelo próprio método de estimação.

seleccionássemos todas as amostras possíveis de tamanho  $n$  a partir desta população. E para cada uma destas amostras calculássemos uma matriz  $\mathbf{x}'\mathbf{u}$ , finalizando com o calculo de uma media destas matrizes. De acordo com a teoria, este resultado teria que ser um vetor coluna nulo.

Vamos discutir algumas propriedades algébricas dos estimadores OLS:

$$1) \sum_{i=1}^n \hat{u}_i = 0 \quad (2.5)$$

Esta propriedade advém do próprio método de obtenção das estimativas OLS. Podemos confirmar esta propriedade através do comando STATA:

```
.summ residuos
Variable |      Obs      Mean    Std. Dev.      Min      Max
-----+-----
residuos |      209    2.85e-06    1363.266   -1160.168   13499.89
```

Observe que o valor da soma dos resíduos estimados é praticamente nula (sempre existe algum erro de aproximação em qualquer software).

2) A covariância amostral entre os regressores e os resíduos OLS é zero. Esta propriedade também advém do próprio método de obtenção das estimativas OLS. Para confirmar isto no STATA:

```
. correlate residuos roe, covariance
```

```
          |  residuos      roe
-----+-----
residuos |  1.9e+06
roe      | -7.9e-06    72.565
```

Novamente temos a covariância entre o regressor (roe) e os resíduos igual à praticamente zero. Esta propriedade dos estimadores OLS pode ser expressa, a partir da propriedade 1, da seguinte forma:

$$\begin{aligned} \text{cov}(xu) &= E[(x - E[x])(u - E(u))] = E[(x - E(x))u] = E[xu] - E[E(x)u] = \\ &= E[xu] - E[x]E[u] = E[xu] \end{aligned} \quad (2.6)$$

De acordo com Wooldridge, as propriedades 1 e 2 podem ser formuladas da seguinte forma:

$$\begin{aligned} E(y - \beta_0 - \beta_1 x) &= 0 \\ E[x(y - \beta_0 - \beta_1 x)] &= 0 \end{aligned} \quad (2.7)$$

Estas duas equações implicam em duas restrições para a distribuição conjunta de probabilidade de (x,y) da população. Para uma amostra de **n** observações, temos:

$$\begin{aligned} n^{-1} \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) &= 0 \\ n^{-1} \sum_{i=1}^n x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) &= 0 \end{aligned} \quad (2.8)$$

Lembremo-nos que esperança matemática é uma média. A dupla de expressões anteriores refere-se à população de pontos. Esta ultima dupla de expressões refere-se à média calculada na amostra. Repare que covariância é também uma média (e também uma esperança). Ver na pg 29 do Wooldridge como é deduzida, a partir destas expressões, a fórmula para a estimativa do parâmetro do termo de inclinação da regressão:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \quad (2.9)$$

Uma importante observação do mesmo autor é que a expressão acima indica que a estimativa OLS do parâmetro de inclinação é igual à covariância entre x e y dividida pela variância de x. Como a variância de x é sempre positiva, então se x e y são positivamente correlacionados, então  $\hat{\beta}_1$  é positivo; se x e y são negativamente correlacionados, então  $\hat{\beta}_1$  é negativo. Observe também que pela expressão anterior, não podemos calcular  $\hat{\beta}_1$  se

$$\sum_{i=1}^n (x_i - \bar{x})^2 = 0, \text{ ou seja, quando } x \text{ é constante na amostra.}$$

3) O ponto  $(\bar{x}, \bar{y})$  está sempre na linha de regressão OLS. Confirmando isto no STATA:

**. regress salary roe**

```

. summa salary
. return list
. scalar ytraco = r(mean)
. summa roe
. scalar xtraco = r(mean)
. matrix b = e(b)
. matrix list b
. scalar ypredic = b[1,2] + b[1,1]*xtraco
. disp ypredic, ytraco
1281.1196 1281.1196

```

Também pode ser observado que dizer que a covariância entre  $x$  e  $u$  é zero é o mesmo que dizer que:

$$\sum_{i=1}^n x_i \hat{u}_i = 0 \quad (2.10)$$

pois considerando-se esta propriedade e a propriedade (1) acima é fácil demonstrar a propriedade (2), segundo um caminho inverso ao seguido na expressão (2.7).

Os valores preditos e os resíduos são não correlacionados na amostra. Podemos verificar isto:

**. correlate yhat resíduos, covariance**

Podemos ver a estimação OLS como um método de decompor cada  $y_i$  em duas partes: um valor predito e um resíduo.

Podemos calcular 3 medidas:

$$\begin{aligned}
SST &= \sum_{i=1}^n (y_i - \bar{y})^2 \\
SSE &= \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 \\
SSR &= \sum_{i=1}^n \hat{u}_i^2 \\
SST &= SSE + SSR
\end{aligned} \quad (2.11)$$

Estes valores são automaticamente calculados pelo STATA e dispostos na saída do comando regress. Por exemplo, SSE no nosso exemplo é igual a

5166419.04 e SSR é igual 386566563. Este é um caso de estimação de regressão bastante atípico, pois a soma dos quadrados dos resíduos (diferença entre os valores ajustados – estimados na reta - e o valor de y observados) é bastante superior a soma dos quadrados das diferenças dos valores ajustados e o valor de y médio. Podemos verificar estes valores indiretamente através dos seguintes comandos STATA:

```
use http://fmwww.bc.edu/ec-p/data/wooldridge/CEOSAL1, clear
regress salary roe
predict yhat
predict residuos, residuals
quietly summa salary
gen totais = (salary - r(mean))^2
gen regressao = (yhat - r(mean))^2
gen erro = residuos^2
quietly summ totais
scalar sst = r(sum)
quietly summ regressao
scalar sse = r(sum)
quietly summ erro
scalar SSR = r(sum)
scalar soma = sse + SSR
disp sst, soma
```

Seria conveniente ler os comentários do Wooldridge sobre o  $R^2$  (pagina 40). Vamos agora verificar o que ocorre quando mudamos a unidade de medida de uma das variáveis da regressão. Suponhamos que agora vamos regredir os salários (que eram anteriormente medidos em milhares de dólares) em dólares. Para isto criemos uma nova variável – salardol:

```
. gen salardol = salary*1000
. regress salardol roe
```

Como pode ser observado as estimativas dos parâmetros também ficam multiplicadas por mil, em comparação com as estimativas anteriores. O que acontece quando mudamos a unidade de medida de uma variável independente? Vamos definir uma nova variável roedec (agora os retornos estão em valores decimais e não em porcentagem, como anteriormente):

```
. gen roedec = roe / 100
```

## **. regress salary roedec**

Neste caso, o valor da estimativa do parâmetro de intercepto permanece o mesmo, alterando-se apenas (fica multiplicado por 100) a estimativa do parâmetro  $\beta_1$ . Se a variável independente é multiplicada ou dividida por uma constante  $c$ , então o coeficiente de inclinação OLS é também dividido ou multiplicado por  $c$  respectivamente. Observe que o valor da estimativa do intercepto não se altera quando mudamos apenas a unidade de medida da variável independente (sem alterar a unidade de medida da variável dependente). Isto se explica porque a interpretação do termo de intercepto (é como o próprio nome diz) o valor da ordenada do ponto em que a reta de regressão da amostra corta o eixo das ordenadas (ou seja, o valor predito de salary quando roedec = 0). Mas quando roedec = 0 também roe = 0 e qualquer que seja o valor da inclinação da reta, ela terá que passar pelo mesmo ponto de intercepto com o eixo das ordenadas.

Outro ponto importante a ser considerado é que o valor de  $R^2$  não se altera com a alteração das unidades de medida das variáveis (seja a dependente como a independente). Alguém pode explicar porque?

### **3. Relações não lineares**

Vimos que o modelo de regressão linear supõe que o coeficiente de inclinação estimado pode ser interpretado como sendo igual à variação na variável dependente devido a uma variação unitária de  $x$ . No entanto, uma deficiência do modelo linear é que esta variação unitária em  $x$  pode ser considerada a partir de qualquer nível de  $x$ . Isto não é uma situação razoável nem realística em muitas situações práticas do mundo econômico. Podemos considerar um modelo que dá (aproximadamente) um efeito de percentagem constante sobre a variável  $y$  devido a uma variável unitária em  $x$ :

$$\log(y) = \beta_0 + \beta_1 x + u \quad (3.1)$$

Se  $\Delta u = 0$  então:

$$\% \Delta y \approx (100\beta_1)\Delta x \quad (3.2)$$

Vamos verificar isto, considerando-se os nossos dados:

**. use <http://fmwww.bc.edu/ec-p/data/wooldridge/wage1>, clear**  
**. regress lwage educ**

Source	SS	df	MS	Number of obs = 526		
Model	27.5606288	1	27.5606288	F( 1, 524)	=	119.58
Residual	120.769123	524	.230475425	Prob > F	=	0.0000
Total	148.329751	525	.28253286	R-squared	=	0.1858
				Adj R-squared	=	0.1843
				Root MSE	=	.48008

lwage	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
educ	.0827444	.0075667	10.94	0.000	.0678796	.0976091
_cons	.5837727	.0973358	6.00	0.000	.3925563	.7749891

O coeficiente da variável educ pode ter uma interpretação percentual quando é multiplicado por 100: o salário aumenta 8,3 % para cada aumento adicional de anos de educação. Este impacto é sobre o valor dos salários e não sobre o valor do logaritmo dos salários. No entanto, a equação (8) acima não captura todas as não linearidades que existem na relação existente entre salários e grau de instrução. Wooldridge chama a atenção que podem existir “efeitos de diploma” de tal forma que o décimo segundo ano (que corresponde ao termino do curso superior) pode ter mais impacto do que o décimo primeiro ano.

Outro uso de logarítmicos é a obtenção de modelos de elasticidade constante.

**. use <http://fmwww.bc.edu/ec-p/data/wooldridge/ceosal1>, clear**  
**. regress lsalary lsales**

Aqui temos um modelo log-log do tipo:

$$\log(\text{salary}) = \beta_0 + \beta_1 \log(\text{sales}) + u \quad (3.3)$$

Podemos verificar pela saída deste modelo que o valor da elasticidade é igual 0,257, o que significa que uma variação de 1 % nas vendas acarreta uma variação de 0,257 % no salário. Um comentário final do Wooldridge sobre modelos com logaritmos é que a mudança de unidade de medida da variável



dependente não afeta o valor do coeficiente de inclinação ( $\beta_1$ ) da variável  $x$ . Isto se explica da seguinte forma: se tivermos inicialmente uma equação  $\log(y_i) = \beta_0 + \beta_1 x_i + u_i$  e adicionarmos  $\log(c_1)$  a ambos os lados da equação teremos  $\log(c_1) + \log(y_i) = [\log(c_1) + \beta_0] + \beta_1 x_i + u_i$  ou  $\log(c_1 \cdot y_i) = [\log(c_1) + \beta_0] + \beta_1 x_i + u_i$ . Portanto, o coeficiente de inclinação da reta continua sendo o mesmo, alterando-se somente o coeficiente de intercepto.

Considerando a equação

$$y = \beta_0 + \beta_1 x + u \quad (3.4)$$

podemos dizer que ela é linear nos seus parâmetros. Desta forma, as variáveis  $y$  e  $x$  podem representar qualquer função matemática. Por exemplo, podemos supor que  $x = \sqrt{cons}$

e a nossa equação fica sendo:

$$y = \beta_0 + \beta_1 \sqrt{cons} + u \quad (3.5)$$

podemos concluir que o modelo de regressão linear simples independe da forma em que as variáveis  $x$  e  $y$  são definidas. No entanto, existem diversos modelos que não são lineares nos seus parâmetros. Um exemplo desses modelos é:

$$cons = 1/(\beta_0 + \beta_1 inc) + u \quad (3.6)$$

observe que não podemos transformar esse modelo em uma equação linear, mesmo que seja aplicada alguma transformação matemática sobre as suas variáveis originais. Para esse tipo de modelo não é possível aplicar os métodos convencionais utilizados para regressões lineares simples (ou múltipla).

Vamos agora interpretar o modelo de regressão linear considerando que as estimativas de mínimos quadrados ordinários são obtidas a partir de uma amostra aleatória simples selecionada aleatoriamente de uma população (universo). Para isto vamos supor uma determinada população de observações. A partir desta população de observações vamos estimar a reta populacional através do método dos mínimos quadrados ordinários. Esta será a

reta da população e a mesma será definida através dos parâmetros da população.

```
. use http://fmwww.bc.edu/ec-p/data/wooldridge/ceosal1, clear
. cd "d:\TEXTOS DOWNLOAD\WOOLDRIDGE DATA FILES\"
. save ceosal1, replace
. use ceosal1, clear
. regress salary roe
. sample 50
. regress salary roe
. use ceosal1, clear
. sample 50
. regress salary roe
. use ceosal1, clear
. sample 50
. regress salary roe
```

Observe que estamos selecionando uma amostra aleatória simples de 50 observações a partir de uma população de 209 observações. A primeira regressão refere-se a população com um todo, ou seja, a regressão refere-se a reta populacional. Em outras palavras, os coeficientes desta primeira regressão são os parâmetros verdadeiros da reta populacional. Na segunda regressão já estimamos a partir de uma amostra de 50 observações. Os resultados desta segunda regressão se referem às estimativas de parâmetros baseados em uma única amostra. A terceira regressão é uma repetição da mesma operação de amostragem de 50 observações selecionadas aleatoriamente a partir da população de 209 observações.

Verifique que quando selecionamos amostras distintas da mesma população as estimativas dos parâmetros variam. Imaginemos que pudéssemos selecionar todas as amostras possíveis de 50 observações a partir de uma mesma população de 209 observações e suponhamos que para cada uma destas amostras obtenhamos uma regressão amostral pelo método dos mínimos quadrados ordinários. Teremos então um número que corresponde ao total de amostras. Este número pode ser calculado pela análise combinatória como:

56227942197269400000000000000000000000000000000000000



#### 4. Uma rotina de simulação de regressões utilizando re-amostragem

Construímos uma rotina (do-file STATA) que irá simular a seleção de diversas amostras a partir de uma mesma população. Para cada uma destas amostras será realizada uma regressão e os coeficientes desta regressão serão armazenados em uma matriz. Desta forma, são selecionadas 5000 amostras de mesmo tamanho  $n = 50$  selecionadas a partir de uma população fixa de tamanho  $N = 209$ . Iniciamos a rotina com a realização de uma regressão para toda a população. São então calculados, por mínimos quadrados ordinários, os valores dos verdadeiros parâmetros da população. A seguir definimos os comandos da própria rotina que é a seqüência que vai do comando **program define** até o comando **end**. Dentro desta rotina existem instruções (comandos) para ler o arquivo original, selecionar uma amostra aleatória de tamanho  $n = 50$  sem reposição (comando **sample**) e realizar a regressão para esta amostra. A seguir inicializamos os valores de quatro matrizes com valores nulos. Finalmente replicamos a rotina 5000 vezes através do comando **forvalues**.

---

```
*  UMA  SIMULAÇÃO  SOBRE  INFERENCIA  EM  REGRESSAO  LINEAR
SIMPLES
*  NESTE  PRIMEIRO  CASO  ESTAMOS  ASSUMINDO  QUE  ESTAMOS
REALIZANDO
*  1000  SELEÇÕES  DE  AMOSTRAS  ALEATORIAS  SIMPLES  SEM
REPOSIÇÃO,
*  COM  TAMANHO  n = 50 e de uma população de tamanho N = 209

set matsize 5000
set more off
cd "d:\TEXTOS DOWNLOAD\WOOLDRIDGE DATA FILES\"

*/ estes sao os valores dos parametros da regressao
*/ estamos calculando por OLS para a população de 209
observações

use CEOSAL1, clear
regress salary roe
matrix e = e(b)
disp "o valor do parametro de inclinacao - beta 1 - é: ",
e[1,1]
```

```

scalar b1 = e[1,1]

capture program drop simulacao
program define simulacao
use CEOSAL1, clear
qui bsample 50
qui regress salary roe
end

matrix B0 = J(5000,1,0)
matrix B1 = J(5000,1,0)
matrix LI = J(5000,1,0)
matrix LS = J(5000,1,0)

forvalues i=1(1)5000 {
simulacao
matrix b = e(b)
matrix v = e(V)
matrix B0[`i',1] = b[1,2]
matrix B1[`i',1] = b[1,1]
matrix LI[`i',1] = b[1,1]-
invttail(48,.025)*sqrt(v[1,1])*sqrt((209-50)/(209-1))
matrix LS[`i',1] =
b[1,1]+invttail(48,.025)*sqrt(v[1,1])*sqrt((209-50)/(209-
1))
}

clear
svmat B0, names(B0)
svmat B1, names(B1)
svmat LI, names(LI)
svmat LS, names(LS)

gen controle = .
replace controle = 1 if b1 >= LI1 & b1 <= LS1
replace controle = 0 if b1 < LI1 | b1 > LS1
gen parametro = b1

summ B0
summ B1
summ controle

```

---

O objetivo dessa rotina é o de mostrar que a média das estimativas de mínimos quadrados ordinários para diversas amostras tende a ser igual aos valores verdadeiros dos parâmetros da população. Além disso, tentamos

demonstrar empiricamente que aproximadamente 95 % dos intervalos de confiança obtidos a partir das regressões amostrais contem os parâmetros da população.

Note que no interior da rotina (que se inicia com o comando `program` define e finaliza com o comando `end`) utilizamos o comando `bsample`. Este comando seleciona aleatoriamente amostras com reposição para um dado tamanho  $n$ , ao contrário do comando `sample` que seleciona amostras aleatórias sem reposição. Havíamos simulado com o uso do comando `sample` e verificamos posteriormente que quando utilizamos o comando `bsample` (amostras com reposição de tamanho 50 selecionadas da população de tamanho 209) que a média das estimativas dos parâmetros  $\beta_0$  e  $\beta_1$  se aproximam mais destes parâmetros. Isto seria uma “mostração” da propriedade de não viés dos estimadores OLS.

Vamos agora construir um gráfico representando diversas retas amostrais juntamente com a reta da população. Para isto, devemos executar a seguinte rotina.

```
*GRAFICOS DA RETA DA POPULACAO E DAS RETAS DAS AMOSTRAS
set more off
cd "d:\TEXTOS DOWNLOAD\WOOLDRIDGE DATA FILES\"

qui use CEOSAL1, clear
qui regress salary roe
matrix bpop = e(b)
matrix vpop = e(V)

qui use CEOSAL1, clear
sample 50, count
qui regress salary roe
matrix bamo1 = e(b)
matrix vamo1 = e(V)

qui use CEOSAL1, clear
sample 50, count
qui regress salary roe
matrix bamo2 = e(b)
matrix vamo2 = e(V)

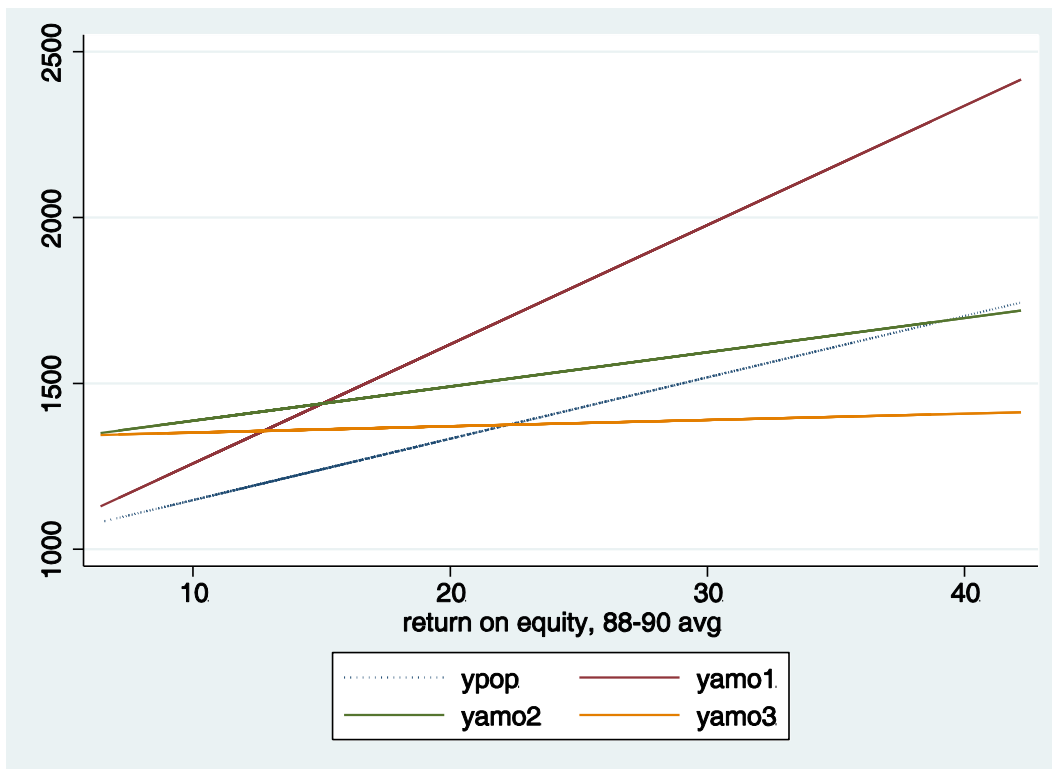
qui use CEOSAL1, clear
sample 50, count
```

```

qui regress salary roe
matrix bamo3 = e(b)
matrix vamo3 = e(V)

gen ypop = bpop[1,2]+bpop[1,1]*roe
gen yamo1 = bamo1[1,2]+bamo1[1,1]*roe
gen yamo2 = bamo2[1,2]+bamo2[1,1]*roe
gen yamo3 = bamo3[1,2]+bamo3[1,1]*roe
twoway (line ypop roe, lpattern(dot)) (line yamo1 roe)
(line yamo2 roe) (line yamo3 roe)

```



**Figura 10 – Retas de regressão da população e retas de regressão de amostras.**

## 5. Introdução a regressão múltipla

O mesmo comando **regress** utilizado para regressão linear simples também será utilizado para regressão linear múltipla. A única diferença é que agora utilizaremos um número maior de variáveis independentes. Vamos considerar a seguinte sequência de comandos:

```

cd "D:\DADOS_BAUM\"
use hprice2a.dta, clear
summa price lprice lnox ldist rooms stratio

```

```
regress lprice lnox ldist rooms stratio
```

Source	SS	df	MS	Number of obs = 506		
Model	49.3987735	4	12.3496934	F( 4, 501) = 175.86		
Residual	35.1834974	501	.070226542	Prob > F = 0.0000		
Total	84.5822709	505	.167489645	R-squared = 0.5840		
				Adj R-squared = 0.5807		
				Root MSE = .265		

lprice	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
lnox	-.95354	.1167418	-8.17	0.000	-1.182904	-.7241762
ldist	-.1343401	.0431032	-3.12	0.002	-.2190255	-.0496548
rooms	.2545271	.0185303	13.74	0.000	.2181203	.2909338
stratio	-.0524512	.0058971	-8.89	0.000	-.0640373	-.0408651
_cons	11.08387	.3181115	34.84	0.000	10.45887	11.70886

Os dados utilizados provêm de um arquivo do livro do Baum. Referem-se a preços de casas de 506 áreas comunitárias de Boston, para as quais a variável de resposta (dependente) é o logaritmo do preço mediano dos domicílios de família única em cada comunidade (**lprice**). A variável **rooms** é o número médio de salas por domicílio. Considera-se também uma medida de poluição do ar (**lnox**) como fator que influencia os preços. Além disso, outra variável independente considerada é a distância da comunidade ao centro de emprego (**ldist**). Finalmente a última variável independente é a relação média aluno-professor nas escolas locais (**stratio**). De acordo com a definição dessas variáveis podemos esperar qual deverá ser o sinal dos coeficientes da regressão linear múltipla.

Podemos considerar nesta última saída do STATA, os seguintes pontos. Do lado esquerdo dos resultados é apresentada a tabela de análise de variância referente a regressão. Na coluna SS estão os valores das somas dos quadrados do modelo, dos resíduos e a soma dos quadrados total. Na segunda coluna temos o número de graus de liberdade referentes ao cálculo de cada soma de quadrado. Na terceira coluna da tabela de análise de variância temos o valor do “mean square”, que é a divisão da soma dos quadrados pelo número de graus de liberdade correspondente. O resultado da divisão entre o SS do modelo e o SS total corresponde ao valor do coeficiente de determinação da regressão ( $R^2$ ).



Na parte direita da listagem da regressão temos as seguintes informações: o número de observações utilizadas na estimativa (506), o valor da estatística F que é utilizada no teste de significância simultânea dos parâmetros da regressão (exceto o termo de intercepto), o p-value correspondente a este valor, o valor do coeficiente de determinação, o valor do coeficiente de determinação ajustado e a raiz quadrada do MSE (mean square error). Este último é o erro médio quadrático e o valor 0,265 corresponde ao valor de uma estimativa para  $\sigma$ , o desvio padrão dos erros, que é um dos parâmetros da regressão. Este valor é baixo, quando comparado a média da variável dependente y (lprice) que é 9,94.

Podemos observar no modelo acima que todas as variáveis independentes são significativas. De fato, para todas elas os valores das estatísticas t são elevados e os correspondentes p-values são baixíssimos. Podemos interpretar estes últimos como o menor nível de significância para o qual podemos rejeitar a hipótese nula (o valor do parâmetro – coeficiente da regressão – é igual a zero).

O valor do  $R^2$  ajustado é 0,5807. A utilidade desta informação é para a comparação entre modelos com distintos números de variáveis. Quando elevamos o número de regressores na equação, o valor do  $R^2$  não ajustado se eleva, o que pode levar a uma conclusão espúria que um modelo com maior número de variáveis independentes é melhor do que um modelo com menor número de variáveis independentes. Esta conclusão pode ser espúria porque podemos estar adicionando variáveis sem sentido (*non sense*) ao modelo restrito. E qualquer variável acrescentada (mesmo que *non sense*) estará elevando o valor do  $R^2$  não ajustado, por razões matemáticas do OLS.

Assim, para dar um exemplo extremo, no exemplo dos preços de domicílios, se adicionarmos ao modelo a variável valor médio do dia de nascimento (de 1 a 365) das pessoas de cada comunidade, o valor do  $R^2$  não ajustado irá se elevar, mesmo que a introdução desta variável como fator explicativo dos preços das residências não tenha o menor sentido. Para resolver este problema de comparação de ajuste de modelos com distinto

numero de variáveis foi proposto o  $R^2$  ajustado, que não é afetado pelo numero de variáveis (veja detalhes sobre este indicador no Wooldridge e Baum). Uma questão importante que Baum levanta é que a adição do regressor eleva o  $R^2$  não ajustado apenas quando este é linearmente independente em relação às colunas previas da matriz X. Também é importante notar que Baum considera que o  $R^2$  ajustado é uma abordagem não estatística de escolha entre modelos *non nested*. O que são modelos *nested*?

Se temos dois modelos da seguinte forma:

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon_1$$

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \varepsilon_2$$

Estes são modelos *nested* porque as variáveis independentes do primeiro modelo formam um subconjunto das variáveis independentes do primeiro modelo. Agora estes dois modelos:

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_6 X_6 + \varepsilon_1$$

$$y = \beta_0 + \beta_1 X_1 + \beta_4 X_4 + \beta_5 X_5 + \varepsilon_2$$

são não *nested* pois os parâmetros (e variáveis) do primeiro modelo não estão contidos no conjunto de parâmetros (e variáveis) do segundo modelo (e vice-versa). Em suma, modelos *nested* são aqueles que podem ser obtidos a partir da simples inclusão de variáveis no primeiro modelo para obter o seguinte modelo.

Veremos adiante que para a seleção do melhor modelo entre pares de modelos *nested* temos uma abordagem bem estatística que são os testes Wald. Quando os modelos não são *nested* resta-nos a abordagem não estatística (e, portanto de menor poder) que é a do  $R^2$  ajustado. Veremos também adiante que existem outras abordagens não estatísticas para a comparação entre modelos *nested* e *non nested* que são os diversos critérios de ajuste: AIC, BIC, etc. Para isto execute o seguinte comando:

```
. estat ic
```

```
-----
```

Model	Obs	ll (null)	ll (model)	df	AIC	BIC
-------	-----	-----------	------------	----	-----	-----

-----+-----						
.	506	-265.4135	-43.49514	5	96.99028	118.123
-----						

Baum chama a atenção (pg 79) para o fato que os resultados de estimação por mínimos quadrados ordinários podem ser considerados como estimadores de máxima verossimilhança (método que iremos tratar mais tarde) do vetor  $\beta$  de parâmetros e o parâmetro adicional  $\sigma_u^2$ . O grau para o qual o nosso modelo ajustado melhora em relação ao modelo nulo (o modelo  $y = \mu + u_i$ , apenas com o termo de intercepto-constante e sem regressores) na explicação da variação da variável dependente é medido pelo maior valor (em termos absolutos) do  $ll(\text{model})$  em relação ao  $ll(\text{null})$ .

Será discutido a frente um teste (portanto, um procedimento estatístico) baseado na função de verossimilhança aplicado aos modelos. As medidas AIC (Akaike information Criteria) e BIC (Bayesian Information Criteria ou Schwarz Criteria) levam em conta (como o  $R^2$  ajustado) o grau de ajuste do modelo como sua parcimônia.<sup>5</sup>

Atenção para a análise desenvolvida por Wooldridge (pág. 87), sobre inclusão de variáveis irrelevantes no modelo de regressão. Sua conclusão é que esta inclusão não altera o fato de que os estimadores dos parâmetros das variáveis relevantes continuam sendo não viesados, mas isto pode causar efeitos indesejáveis nas variâncias dos estimadores OLS. Este é o chamado problema de **sobre especificação** do modelo. O problema oposto é quando omitimos uma variável relevante (ou seja, esta variável afeta o valor de  $y$  na população). Este é o chamado problema de subespecificação do modelo ou problema do viés de variável omitida. Este problema causa viés das estimativas de mínimos quadrados ordinários.

---

<sup>5</sup> Sobre o critério da parcimônia, bastante caro na análise econométrica, parece ter surgido ancestralmente com a proposição da navalha de Okham, nome proveniente de filósofo medieval que considerou que se duas teorias são concorrentes para explicar a realidade, deve-se escolher a menos complexa.

## 6. O significado dos coeficientes beta (padronizados)

Vamos executar a seguinte sequência de comandos no do-file editor:

```
regress lprice lnox ldist rooms stratio  
  
regress, beta  
  
qui summ lprice  
qui gene zlprice = (lprice - r(mean))/ r(sd)  
qui summ lnox  
qui gene zlnox = (lnox - r(mean))/ r(sd)  
qui summ ldist  
qui gene zldist = (ldist - r(mean))/ r(sd)  
qui summ rooms  
qui gene zrooms = (rooms - r(mean))/ r(sd)  
qui summ stratio  
qui gene zstratio = (stratio - r(mean))/ r(sd)  
  
regress zlprice zlnox zldist zrooms zstratio
```

O segundo comando (**regress, beta**) calculará os coeficientes betas da regressão, que são os coeficientes padronizados. Eles são calculados a partir dos valores das variáveis (tanto a dependente como as independentes) padronizados. Padronizar os valores de uma variável significa subtrair de cada valor original a sua média e dividir o resultado pelo seu desvio padrão. Fazemos estes cálculos nos comandos seguintes para “verificar” o cálculo através do comando **regress, beta**. Como interpretar os coeficientes betas?

Os valores destes coeficientes servem para verificar quais são as variáveis que tem maior “força” explicativa em y, independentemente de suas distintas escalas ou unidades de medida utilizadas. Sabemos (vimos em exercício anterior) quais são os efeitos da multiplicação de variáveis (dependente e independentes) por um valor constante. Isto conduziu a conclusão que os coeficientes (não padronizados) dependem da escala de mensuração das variáveis. Os coeficientes betas são invariantes ao fator escala de mensuração das variáveis e podem ser utilizados para avaliar a sua maior ou menor importância explicativa na variável dependente.

## 7. Multicolinearidade

O problema da multicolinearidade surge quando um dos regressores é combinação linear de outros regressores. Isto fará com que a matrix  $X'X$  seja singular (com determinante nulo) o que impede completamente a solução matricial do método dos mínimos quadrados ordinários. Podemos também dizer que nem todas as estimativas dos parâmetros da regressão são numericamente identificáveis. Este é o caso da multicolinearidade perfeita, quando não podemos inverter a matriz  $X'X$ . O programa STATA detecta automaticamente uma situação de multicolinearidade perfeita.

No entanto, para a quase-multicolinearidade devemos ter um tratamento mais cuidadoso. Tanto a multicolinearidade perfeita como a quase multicolinearidade afetam (de forma negativa) as estimativas da regressão e podem nos conduzir a falsas conclusões a partir das mesmas. Este é um problema para o qual não é dada muita atenção em muitos trabalhos econométricos e que, portanto perdem a sua validade analítica. O STATA elimina automaticamente regressores que formam relações de dependência linear perfeita com outros regressores e dá uma mensagem no início do relatório da regressão.

Já em uma situação de quase-multicolinearidade este procedimento automático não ocorre e neste caso temos que ter mais cautela. De acordo, com Baum (pg 85) pequenas mudanças na matriz de dados podem causar grandes mudanças nas estimativas dos parâmetros, desde que elas são proximamente não identificadas. Pode ocorrer que o ajuste geral da regressão é muito bom (elevado  $R^2$  e  $R^2$  ajustado) e os coeficientes podem ter erros padrões muito elevados e talvez sinais incorretos ou grandezas implausivelmente grandes.

Os econométristas desenvolveram um teste muito útil para detectar uma situação de quase multicolinearidade: o teste vif (*variance inflation factor*). A lógica deste teste baseia-se na idéia de que quando um regressor não é

ortogonal aos outros regressores<sup>6</sup> a variância do respectivo parâmetro fica inflacionada. Vamos executar o comando STATA para verificar esta situação:

**. estat vif**

Variable	VIF	1/VIF
-----+-----		
lnox	3.98	0.251533
ldist	3.89	0.257162
rooms	1.22	0.820417
stratio	1.17	0.852488
-----+-----		
Mean VIF	2.56	

Uma regra de bolso para verificar se existe multicolinearidade entre o conjunto de regressores é que a media do vif não deve ser maior do que 1 ou que o maior vif não é maior do que 10. No caso anterior podemos concluir que não existe evidencia de multicolinearidade, pois o maior vif não é maior do que 4.

Na verdade estes são os casos de evidencia de perfeita colinearidade. De acordo com Baum, os casos de quase-colinearidade somente podem ser checados por um comando que não é oficial ao STATA: o comando **coldiag2**. Para a instalação deste comando execute:

**. findit coldiag2**

Surgirá uma tela de display e siga para a instalação. Para a utilização adequada deste procedimento execute o comando:

**. help coldiag2**

Wooldridge desenvolve uma importante discussão a partir da expressão (pg. 94):

---

<sup>6</sup> O ideal é que os regressores sejam perfeitamente ortogonais. A ortogonalidade é uma representação geométrica das variáveis em um espaço n-euclidiano. Se duas variáveis tem correlação nula elas serão representadas neste espaço como dois vetores ortogonais. Então um sistema de variáveis que não tem dependência linear entre si devem formar um conjunto de vetores ortogonais entre si.

$$Var(\hat{\beta}_j) = \frac{\sigma^2}{SST_j(1 - R_j^2)}$$

Nesta expressão observa-se que a variância do estimador OLS do parâmetro  $\hat{\beta}_j$  depende de três fatores:

1) A variância dos erros da regressão. Este é um componente da expressão que é desconhecido, mas para o qual pode ser obtida uma estimativa. Para uma dada variável dependente  $y$ , só há uma maneira de reduzir a variância do erro que é incluir na equação mais variáveis independentes (ou seja, extrair alguns fatores do termo de erro). Mas isso nem sempre é possível e nem desejável.

2) a variação total amostral em  $x_j$  ( $SST_j$ ). De acordo com expressão acima quanto maior a variação total amostral em  $x_j$  menor é a variância do estimador de  $\hat{\beta}_j$ . Uma maneira de obter isto é aumentar o tamanho da amostra.

3) As relações lineares entre as variáveis independentes,  $R_j$ . Esse é o fator que se refere ao problema da multicolinearidade. O valor de  $R_j^2$  é o coeficiente de determinação da regressão da variável independente  $x_j$  com todas as outras variáveis independentes. Para um dado valor de  $\sigma^2$  e  $SST_j$ , o menor valor de  $Var(\hat{\beta}_j)$  é obtido quando  $R_j^2 = 0$ . Isto ocorre quando  $x_j$  tem correlação nula com todas as outras variáveis independentes, o que é um caso muito difícil de acontecer. No outro extremo quando  $R_j^2 = 1$  temos uma situação de multicolinearidade perfeita. Mas o problema das variâncias elevadas também pode ocorrer devido à baixa variabilidade das variáveis independentes. Este problema é designado na literatura econométrica como micronumerosidade. Ambos os problemas podem ser atenuados com a utilização de maiores amostras. Outra alternativa é eliminar a variáveis independentes do modelo. Mas nesse caso, podemos estar eliminando variáveis relevantes.

Para simular esses resultados vamos considerar os dados do exemplo anterior. Considere a seguinte sequência de comandos:

```

use http://www.stata-press.com/data/imeus/hprice2a, clear
gen varteste1 = 50
regress lprice lnox ldist rooms stratio varteste1
replace varteste1 = 49 in 150/152
regress lprice lnox ldist rooms stratio varteste1
estat vif
gen varteste2 = 2*lnox + 3*ldist
regress lprice lnox ldist rooms stratio varteste2
estat vif
replace varteste2 = varteste2 + 1 in 150/152
regress lprice lnox ldist rooms stratio varteste2
estat vif

```

Observe que primeiramente criamos uma variável com valor constante igual a 50 e utilizamos esta variável como regressor. O STATA automaticamente elimina esta variável da regressão. Após isso fazemos uma pequena alteração da variável constante e rodamos a regressão. Desta vez a regressão é estimada. Em seguida criamos uma variável que é combinação linear de 2 variáveis preexistentes no modelo. O STATA elimina automaticamente uma das variáveis que forma combinação linear com as outras duas.

Fazemos o teste do fator de inflação de variância (vif) e os resultados mostram não haver multicolinearidade. O STATA eliminou a variável porque ele detecta automaticamente multicolinearidade perfeita. A seguir alteramos dois valores da variável criada anteriormente. Criamos assim uma situação de quase multicolinearidade perfeita. Ao rodar a modelo, o STATA não elimina nenhuma variável. Mas o teste de fator de inflação de variância acusa problema de multicolinearidade. É importante notar que os três fatores que podem elevar a variância dos estimadores dos betas podem influenciar conjuntamente.

Podemos imaginar uma situação em que temos simultaneamente baixa variabilidade dos regressores, quase multicolinearidade e elevado valor da variância dos erros. É importante notar que os três fatores que podem elevar a variância dos estimadores dos betas podem influenciar conjuntamente.



## 8. Apresentação de resultados de regressões no STATA

A sintaxe a seguir, executada através de um arquivo do-file permite especificar diversos modelos e apresentá-los de uma forma conjunta em uma única tabela, permitindo uma melhor comparabilidade de seus resultados.

```
gene rooms2 = rooms^2
qui regress lprice rooms
estimates store model1
qui regress lprice rooms rooms2 ldist
estimates store model2
qui regress lprice ldist stratio lnox
estimates store model3
qui regress lprice lnox ldist rooms stratio
estimates store model4

estimates table model1 model2 model3 model4, stat(r2_a rmse) ///
b(%7.3g) se(%6.3g) p(%4.3f)
estimates table model1 model2 model3 model4, stat(r2_a rmse ll) /// b(%7.3g)
star title("Models of median housing pricing")
```

Variable	model1	model2	model3	model4
rooms	.369 .0201 0.000	-.821 .183 0.000		.255 .0185 0.000
rooms2		.0889 .014 0.000		
ldist		.237 .0255 0.000	-.157 .0505 0.002	-.134 .0431 0.002
stratio			-.0775 .0066 0.000	-.0525 .0059 0.000
lnox			-1.22 .135 0.000	-.954 .117 0.000
_cons	7.62 .127 0.000	11.3 .584 0.000	13.6 .304 0.000	11.1 .318 0.000
r2_a	.399	.5	.424	.581
rmse	.317	.289	.311	.265

legend: b/se/p

Para acessar todos os recursos deste procedimento execute o comando:

```
. help estimates table
```

## 9. O teste Wald

Nesta parte do texto seguiremos basicamente o livro do Baum, seção 4.5. Consideremos uma equação de regressão dada na sua forma matricial:

$$\mathbf{y} = \mathbf{x}\boldsymbol{\beta} + \mathbf{u} \quad (9.1)$$

Vamos também considerar um conjunto de restrições lineares impostas ao vetor de parâmetros  $\boldsymbol{\beta}$ :

$$\mathbf{R}\boldsymbol{\beta} = \mathbf{r} \quad (9.2)$$

Por exemplo, se tivermos a seguinte equação de regressão:

$$y = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + u$$

Se testarmos a hipótese  $H_0\{\beta_3=0\}$  então:

$$\mathbf{R} = (0010) \text{ e } \mathbf{r} = (0)$$

Se testarmos a hipótese  $H_0\{\beta_3 = \beta_4=0\}$  então:

$$\mathbf{R} = (001-1) \text{ e } \mathbf{r} = (0)$$

O teste F (ANOVA) de que os coeficientes (exceto o de intercepto) são todos nulos, ou seja,  $H_0\{\beta_2 = \beta_3 = \beta_4=0\}$  então:

$$\mathbf{R} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

$$\mathbf{r} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

De forma que  $\mathbf{R}\boldsymbol{\beta} = \mathbf{r}$  fica sendo como:

$$\mathbf{R}\boldsymbol{\beta} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

$$\begin{pmatrix} \beta_2 \\ \beta_3 \\ \beta_4 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

O próximo comando importa diretamente pela net um arquivo de dados armazenado em um servidor:

```
. use http://www.stata-press.com/data/imeus/hprice2a, clear
. regress lprice lnox ldist rooms stratio
. rooms test
```

```
( 1)   rooms = 0

      F(   1,   501) =   188.67
      Prob > F =       0.0000
```

Em termos de comandos STATA este comando é equivalente a:

```
. test _b[rooms] = 0
```

O resultado do comando anterior apresenta a estatística F ao invés da t de Student da tabela da regressão. É uma F(1,N-k) graus de liberdade. O STATA reporta a estatística F porque muitas hipóteses podem envolver mais do que uma restrição no vetor de coeficientes (e portanto mais do que um grau de liberdade). Assim, para não haver perda de generalidade a estatística F é apresentada.

Vamos supor agora que queremos testar se o parâmetro é igual a um determinado valor distinto de zero (uma constante):

```
. test rooms = 0.50
```

```
( 1)   rooms = .5
```

```
F( 1, 501) = 175.49
Prob > F = 0.0000
```

Neste caso podemos rejeitar a hipótese nula  $H_0 \{ \beta_{rooms} = 0.50$

Podemos também considerar um teste Wald para uma combinação linear de parâmetros:

**. lincom rooms + ldist + stratio**

```
( 1)  ldist + rooms + stratio = 0
```

lprice	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
(1)	.0677357	.0490714	1.38	0.168	-.0286753	.1641468

A soma estimada dos três coeficientes é 0,068. Observe também que o intervalo de confiança de 95 % de probabilidade para a estimativa da soma inclui o valor zero. Isto está em concordância com o valor do p-value (0,168) o que faz com que aceitemos a hipótese nula de que a soma dos coeficientes é igual a zero.

Verifique a igualdade de resultados dos seguintes dois comandos:

**. test ldist = stratio**

**. lincom ldist – stratio**

Você vai perceber que o p-value para a estatística t (do comando test) e da estatística F (do comando lincom) são os mesmos. Compare também os resultados de:

**. test lnox = 10\*stratio**

**. lincom lnox - 10\*stratio**

Não podemos utilizar os mesmos procedimentos para testar a hipótese de que o produto de dois coeficientes é igual a uma dada constante. Podemos considerar a estimação de um determinado modelo sujeito a uma restrição quanto a seus parâmetros. Por exemplo, podemos forçar que a soma de três parâmetros seja nula. Isto é distinto de estimar o mesmo modelo livremente, sem restrições. De acordo com Baum, podemos proceder de duas formas:

1) Introduzir a restrição na própria equação do modelo e estimar o modelo restrito. Por exemplo, suponhamos que temos o seguinte modelo:

$$y = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + u$$

E queremos restringi-lo com a seguinte expressão:

$$\beta_2 = 1,5\beta_3$$

Então, o modelo restrito fica sendo:

$$y = \beta_1 X_1 + \beta_3 (1,5 X_2 + X_3) + \beta_4 X_4 + u$$

Desta forma, vamos estimar o modelo restrito, por OLS com os regressores  $X_1$ ,  $1,5X_2+X_3$  e  $X_4$ .

Esta estratégia pode se tornar difícil para modelos mais complicados sujeitos a restrições.

2) Outra forma é aplicar o comando STATA **constraint** para impor cada restrição

ao modelo e posteriormente executar o comando **cnsreg** para estimar a equação do modelo com as restrições impostas.

**. constraint def 1 ldist + rooms +stratio = 0**

**. cnsreg lprice lnox ldist rooms stratio, constraint(1)**

Constrained linear regression					Number of obs = 506	
					F( 3, 502) = 233.42	
					Prob > F = 0.0000	
					Root MSE = .26524	
( 1) ldist + rooms + stratio = 0						
lprice	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
lnox	-1.083392	.0691935	-15.66	0.000	-1.219337	-.9474478
ldist	-.1880712	.0185284	-10.15	0.000	-.2244739	-.1516684
rooms	.2430633	.01658	14.66	0.000	.2104886	.2756381
stratio	-.0549922	.0056075	-9.81	0.000	-.0660092	-.0439752
_cons	11.48651	.1270377	90.42	0.000	11.23691	11.7361

Para checar se efetivamente o modelo foi estimado com a restrição imposta execute:

**. disp \_b[ldist] + \_b[rooms] + \_b[stratio]**

É importante destacar que restrições aos modelos não melhoram o grau de ajuste aos dados. Observe que no nosso caso o valor do Root MSE se elevou ligeiramente em relação ao modelo original (pagina 34).

## 10. Testes de hipóteses conjuntas

Muitas vezes queremos testar uma hipótese envolvendo restrições múltiplas no vetor de coeficientes. Na presença de um elevado grau de multicolinearidade encontramos um resultado de regressão em que todos os coeficientes individualmente são não significativos ao passo que o teste conjunto (teste F) revela-se significativo. A estatística de teste F tem tantos graus de liberdade no numerador quantas forem as restrições no vetor de coeficientes.

```
. qui regress lprice lnox ldist rooms stratio  
. test lnox ldist
```

```
( 1)  lnox = 0  
( 2)  ldist = 0  
  
      F( 2, 501) = 58.95  
      Prob > F = 0.0000
```

Rejeitamos a hipótese nula de que os coeficientes das variáveis **lnox** e **ldist** são simultaneamente (conjuntamente) iguais a zero. Mais rigorosamente, rejeitamos a hipótese conjunta de que o modelo excluindo as variáveis **lnox** e **ldist** é mais corretamente especificado do que o modelo completo.

## 11. Calculando resíduos, valores preditos e predições por intervalos

Vimos na introdução deste trabalho que o STATA calcula facilmente os resíduos de uma regressão com um comando de pós-estimação (executado logo após o comando regress) chamado predict, que já vimos anteriormente. Este comando permite calcular tanto os resíduos como os valores preditos para

cada observação da amostra. No caso das predições podemos calcular predições dentro da amostra, mas também podemos calcular predições fora da amostra.

Por exemplo, podemos considerar o cálculo da predição para a variável de resposta (variável dependente) considerando que determinados valores das variáveis independentes não são observados na amostra. No exemplo da determinação dos salários pode ser que não encontremos na amostra nenhum indivíduo com o valor de anos de estudo igual a 12. Mesmo assim podemos calcular o valor predito, bastando para isto introduzir na equação da regressão estimada este valor para a variável número de anos de estudo.

De acordo com Baum um modelo de regressão bem especificado deveria gerar predições razoáveis para qualquer amostra da população, mesmo para amostras aleatórias bem distintas daquela a partir da qual geramos as estimativas dos parâmetros. Baum recomenda a utilização da opção `double` no comando `predict` para gerar os valores preditos com toda a precisão numérica do software. Por exemplo, após executar uma regressão:

```
sysuse auto
regress price weight length foreign
predict double predprice
```

Se realizarmos a regressão em apenas uma parte da amostra, por exemplo:

```
drop predprice
regress price weight length if foreign == 1
predict double predprice
```

Isto irá gerar predições para toda a amostra de automóveis inclusive para os automóveis nacionais (`foreign = 0`) que não entraram na estimação da regressão. Se quisermos realizar predições apenas para a amostra utilizada na regressão temos que utilizar a opção `if e(sample)`:

```
drop predprice
regress price weight length if foreign == 1
predict double predprice if e(sample)
```

Vamos agora tratar da estimação de predições do modelo considerando a construção de intervalos de confiança para estas predições. Seguindo o desenvolvimento em Wooldridge (2006) podemos considerar a equação de estimação:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \dots + \hat{\beta}_k x_k$$

Consideremos que  $c_1, c_2, \dots, c_k$  sejam valores particulares para as variáveis independentes (estes podem ser valores da amostra ou valores fora da amostra). O parâmetro de predição que queremos estimar é:

$$\theta_0 = \beta_0 + \beta_1 c_1 + \beta_2 c_2 + \dots + \beta_k c_k = E(y | x_1 = c_1, x_2 = c_2, \dots, x_k = c_k)$$

Cujo estimador é:

$$\hat{\theta}_0 = \hat{\beta}_0 + \hat{\beta}_1 c_1 + \hat{\beta}_2 c_2 + \dots + \hat{\beta}_k c_k$$

Consideremos agora a equação da reta de regressão:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + u$$

Podemos isolar o valor:

$$\beta_0 = \theta_0 - \beta_1 c_1 - \beta_2 c_2 - \dots - \beta_k c_k$$

E substituir na equação anterior, resultando em:

$$y = \beta_0 + \beta_1 (x_1 - c_1) + \beta_2 (x_2 - c_2) + \dots + \beta_k (x_k - c_k) + u$$

Então o valor predito  $\beta_0$  é obtido da equação de  $y$  sobre  $(x_1 - c_1), (x_2 - c_2), \dots, (x_k - c_k)$  e também podemos obter seu erro padrão e conseqüentemente o intervalo de confiança da predição. Vamos seguir o exemplo 6.5, pagina 199 do Wooldridge(2006).

use <http://fmwww.bc.edu/ec-p/data/wooldridge/gpa2>  
regress colgpa sat hsperc hsize hsize\_sq

Suponhamos que desejamos estimar a predição para a variável dependente quando  $\text{sat} = 1200$ ,  $\text{hsperc} = 30$ ,  $\text{hsize} = 5$



Uma forma de obter indiretamente o valor da predição é substituir os valores das variáveis nos resultados da regressão anterior:

```
regress colgpa sat hsperc hsize hsizeSq
```

```
matrix b = e(b)
```

```
matrix list b
```

```
disp b[1,1]*1200 + b[1,2]*30 + b[1,3]*5 + b[1,4]*25 + b[1,5]
```

Mas desta forma estimamos apenas o valor predito mas não estimamos o seu erro padrão que pela regressão anterior é .0198778.

Mas outra forma é empregar o desenvolvimento teórico do Woodridge:

```
gen sat0 = Sat - 1200
```

```
gen hsperc0 = hsperc - 30
```

```
gen hsize0 = hsize - 5
```

```
gen hsizeSq0 = hsizeSq - 25
```

```
regress colgpa sat0 hsperc0 hsize0 hsizeSq0
```

```
matrix b = e(b)
```

```
matrix v = e(V)
```

```
matrix list v
```

```
disp "lim inf = ", b[1,5] - sqrt(v[5,5])*1.96
```

```
disp "lim sup = ", b[1,5] + sqrt(v[5,5])*1.96
```

Temos assim os resultados de um intervalo de confiança de 95 % de probabilidade para o valor predito. Wooldridge(2006) chama a atenção que a variância desta predição é mínima quando estamos predizendo para os valores médios das variáveis independentes.

## 12. Especificando a forma funcional

A consistência do estimador da regressão linear requer que a função de regressão da amostra corresponda à função de regressão subjacente ou o verdadeiro modelo de regressão para a variável de resposta (dependente)  $y$ :

$$y_i = x_i\beta + u_i \quad (12.1)$$

A teoria econômica freqüentemente fornece um guia na especificação do modelo, mas pode ser que ela não indique explicitamente como uma variável específica entre no modelo ou identifique a forma funcional. O modelo deve ser estimado em níveis para as variáveis; ou em uma estrutura logarítmica; como um polinômio em um ou mais dos regressores? Em geral a teoria se cala frente a estes pontos específicos e temos que utilizar estratégias empíricas.

### 12.1 Omissão de variáveis relevantes do modelo (subespecificação)

Suponha que o verdadeiro modelo (população) é:

$$y = x_1\beta_1 + x_2\beta_2 + u \quad (12.1)$$

com  $k_1$  e  $k_2$  regressores em dois subconjuntos, mas regredimos  $y$  somente sobre as variáveis  $x_1$  :

$$y = x_1\beta_1 + u \quad (12.2)$$

A solução de mínimos quadrados ordinários é:

$$\begin{aligned} \hat{\beta}_1 &= (X_1'X_1)^{-1}X_1'y = (X_1'X_1)^{-1}X_1'(\beta_1X_1 + \beta_2X_2 + u) \\ &= \beta_1 + (X_1'X_1)^{-1}X_1'X_2\beta_2 + (X_1'X_1)^{-1}X_1'u \end{aligned} \quad (12.3)$$

A menos que  $X_1'X_2 = 0$  ou  $\beta_2$  a estimativa de  $\hat{\beta}_1$  é viesada, desde que:

$$E[\hat{\beta}_1 | X] = \beta_1 + P_{1.2}\beta_2 \quad (12.4)$$

onde:

$$P_{1.2} = (X_1'X_1)^{-1}X_1'X_2 \quad (12.5)$$

$P_{12}$  é uma matriz  $k_1 \times k_2$  refletindo a regressão de cada coluna de  $X_2$  nas colunas de  $X_1$ .

Se  $k_1=k_2$  e a variável única em  $\mathbf{X}_2$  é correlacionada com a variável única em  $\mathbf{X}_1$ , podemos prever a direção do viés. Mas se tivermos múltiplas variáveis em cada conjunto não podemos prever a natureza do viés dos coeficientes.

Consideremos a seguinte sequência de comandos para simular uma situação de omissão de variáveis. Nesta sequência de comandos iremos inicialmente fazer uma regressão na população e calcular os verdadeiros parâmetros do modelo de regressão verdadeiro. Em seguida, iremos selecionar uma amostra aleatória simples desta população e estimaremos os parâmetros com um modelo completo e com um modelo com omissão de variáveis. Finalmente, vamos empregar a teoria exposta anteriormente, para verificar que a estimativa de mínimos quadrados ordinários para o parâmetro  $\beta_1$  é viesada e que o tamanho do viés para cada amostra é igual ao valor do produto matricial  $\mathbf{P}_{1.2}\beta_2 + (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{u}$  sendo que a esperança matemática condicional aos valores da matriz de variáveis independente  $\mathbf{X}$  do ultimo termo desta soma é zero.

#### \* OMISSAO DE VARIÁVEIS RELEVANTES NO MODELO

```
matrix drop _all
```

\* Vamos considerar o arquivo gpa2 do Wooldridge como dados de uma população

use <http://fmwww.bc.edu/ec-p/data/wooldridge/gpa2>

\*Vamos verificar o tamanho N da população e calcular os valores dos parâmetros

```
count
```

```
regress colgpa hsperc sat hsize
```

```
matrix bpop = e(b)
```

```
matrix list bpop
```

```
matrix betapop = e(b)
```

```
matrix betapop = betapop'
```

```
matrix list betapop
```

```
matrix beta1pop = J(2,1,0)
```

```
matrix beta1pop[2,1] = betapop[1,1]
```

```
matrix beta1pop[1,1] = betapop[4,1]
```

```
matrix beta2pop = J(2,1,0)
```

```
matrix beta2pop[1,1] = betapop[2,1]
```

```
matrix beta2pop[2,1] = betapop[3,1]
```

```
predict residuo, residuals
```

\* vamos selecionar uma amostra aleatória de tamanho n = 50

```
sample 50, count
```

```
regress colgpa hsperc sat hsize
```

regress colgpa hsperc

\* vamos gerar o valor da estimativa viesada do parâmetro beta1

matrix b = e(b)

matrix list b

gen const = 1

mkmat residuo, matrix(u)

mkmat const hsperc, matrix(X1)

mkmat sat hsize, matrix(X2)

mkmat colgpa, matrix(Y)

\* Vamos calcular a estimativa do parâmetro beta1 nesta ultima regressão

\* (com omissão da variável sat) utilizando álgebra linear e empregando

\* a expressão da pagina 116 do Baum

matrix betahat1 = inv(X1'\*X1)\*X1'\*Y

matrix list betahat1

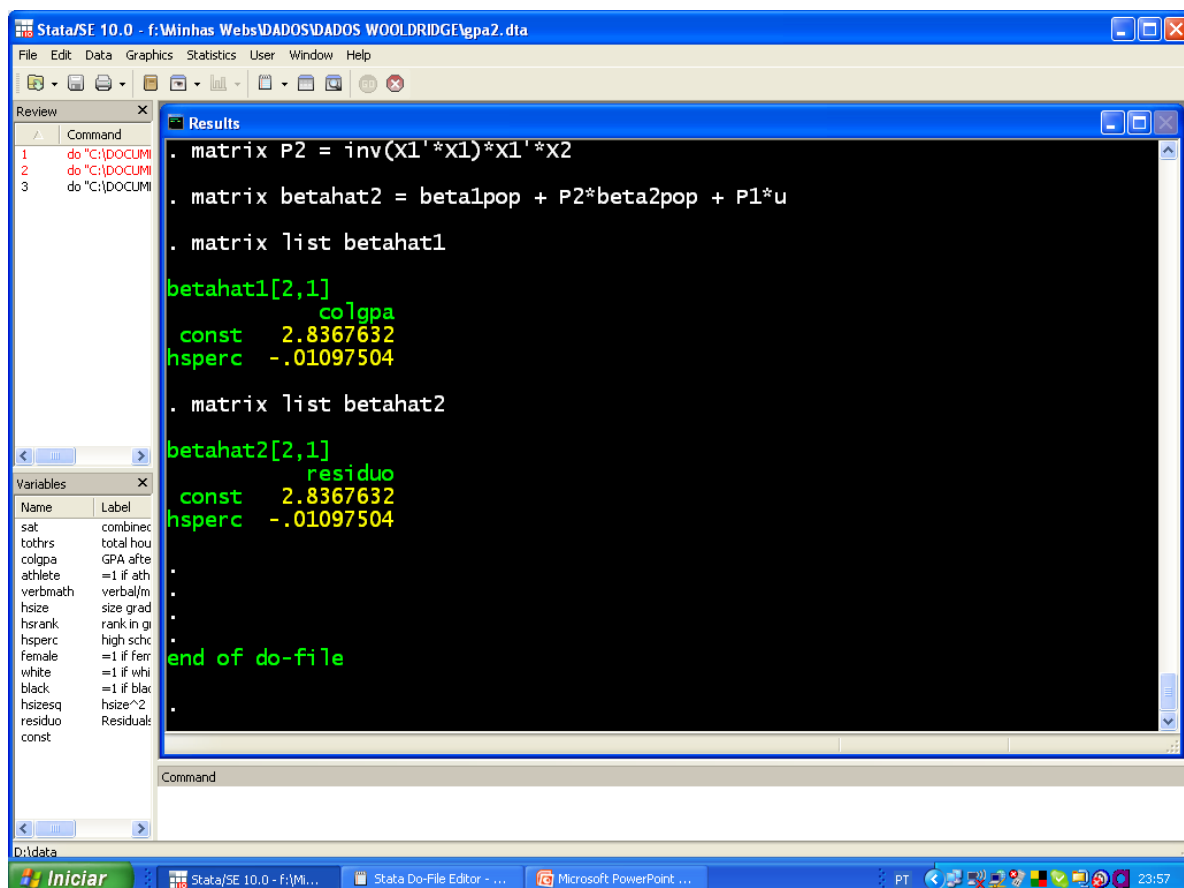
matrix P1 = inv(X1'\*X1)\*X1'

matrix P2 = inv(X1'\*X1)\*X1'\*X2

matrix betahat2 = beta1pop + P2\*beta2pop + P1\*u

matrix list betahat1

matrix list betahat2



The screenshot shows the Stata/SE 10.0 interface. The main window displays a do-file with the following commands:

```
1 do "C:\DOCUMENTOS\WOOLDRIDGE\gpa2.dta"
2 do "C:\DOCUMENTOS\WOOLDRIDGE\gpa2.dta"
3 do "C:\DOCUMENTOS\WOOLDRIDGE\gpa2.dta"
```

The Results window shows the output of the commands:

```
. matrix P2 = inv(x1'*x1)*x1'*x2
. matrix betahat2 = beta1pop + P2*beta2pop + P1*u
. matrix list betahat1
      colgpa
const      2.8367632
hsperc     -.01097504
. matrix list betahat2
      residuo
const      2.8367632
hsperc     -.01097504
.
.
.
end of do-file
```

The Variables window shows a list of variables:

Name	Label
sat	combined
tothrs	total hou
colgpa	GPA afte
athlete	=1 if ath
verbmth	verbal/m
hsize	size grad
hsrank	rank in g
hsperc	high sch
female	=1 if fem
white	=1 if whi
black	=1 if bla
hsizesq	hsize^2
residuo	Residual
const	

Figura 11 – Omissão de variáveis relevantes no modelo – verificação do cálculo do viés da estimativa

Wooldridge (2006) apresenta na pg 90 um quadro resumo para modelos de 2 variáveis:

	<b>Corr(<math>x_1, x_2 &gt; 0</math>)</b>	<b>Corr(<math>x_1, x_2 &lt; 0</math>)</b>
$\beta_2 > 0$	Viés positivo	Viés negativo
$\beta_2 < 0$	Viés negativo	Viés positivo

Se a correlação entre  $X_1$  e  $X_2$  é nula na população, as estimativas de regressão são consistentes, mas provavelmente serão viesadas em amostras finitas. Quando existe correlação entre as variáveis  $X_2$  e  $X_1$  então automaticamente existirá correlação entre  $X_1$  (a variável não omitida) e os erros do modelo (pois estes conterão a variável omitida). Se  $\text{corr}(X_1, X_2)$  é diferente de zero então  $\text{corr}(X_1, X_2 + v_i)$  é diferente de zero sendo que erro do modelo na população =  $X_2 + v_i$ . Então, o modelo  $Y = \beta_1 X_1 + \beta_2 X_2 + u$  produzirá estimativas inconsistentes (ou seja, mesmo para amostras grandes ocorrerá viés). No entanto se  $\text{corr}(X_1, X_2) = 0$  as estimativas de  $\beta_1$  e  $\beta_2$  serão viesadas para pequenas amostras e tenderão a ser não viesadas para grandes amostras (serão consistentes, apesar de viesadas para pequenas amostras).

Isto porque neste caso, apesar de estarmos excluindo uma variável relevante no modelo populacional ( $Y = \beta_1 X_1 + \beta_2 X_2 + u$ ) e executando um modelo na amostra correspondente a  $Y = \beta_1 X_1 + \varepsilon$ , onde  $\varepsilon$  é o valor do resíduo, como  $\text{corr}(X_1, X_2)$  é igual a zero, então também  $\text{corr}(X_2, \text{erro}) = 0$ , não violando este pressuposto crucial do modelo de regressão múltipla ao qual temos sempre nos referido. Quando este pressuposto não é violado, os estimadores OLS são consistentes (mesmo que viesados) para pequenas amostras.

Temos uma interessante expressão na pagina 160 do Wooldridge:

$$p \lim \hat{\beta}_1 = \beta_1 + \text{cov}(X_1, u) / \text{var}(X_1) \quad (12.6)$$

Ou seja, o limite em probabilidade do estimador de  $\beta_1$  ( $\hat{\beta}_1$ ) converge em probabilidade para  $\beta_1$  quando a covariância entre  $X_1$  e  $u$  é zero. O que significa convergir em probabilidade?

Na página 10 do apêndice C do Wooldridge (as versões em Português deste apêndice não estão no livro, mas podem ser obtidas do site Editora Thomson) temos a definição de um estimador consistente:

$W_n$  um estimador de  $\theta$  com base em uma amostra  $Y_1, Y_2, \dots, Y_n$  de tamanho  $n$ . Então,  $W_n$  será um estimador consistente de  $\theta$  se, para cada  $\varepsilon > 0$

$$P(|W_n - \theta| > \varepsilon) \rightarrow 0 \text{ conforme } n \rightarrow \infty \quad (12.7)$$

Esta expressão acima indica que o limite em probabilidade do estimador do parâmetro é igual ao parâmetro. Ela significa que a medida que aumentamos o tamanho  $n$  da amostra a probabilidade de que a diferença em valor absoluto entre o valor do estimador (estimativa) do parâmetro e o próprio parâmetro seja maior do que um pequeno valor  $\varepsilon > 0$ , arbitrário, esta probabilidade tende a zero. Esta expressão é resumida pelos estatísticos como plim (ou limite em probabilidade). Então dizer que um estimador é consistente é dizer que o plimite deste estimador é o próprio parâmetro quando  $n$  tende a infinito.

### 13. Erros não i.i.d.

Já é chegado o momento de fazermos uma síntese esquemática do conjunto de procedimentos que tratamos agora. O modelo de regressão linear múltipla considera a não violação das seguintes hipóteses:

HIPOTESE 1: LINEARIDADE NOS PARAMETROS

HIPOTESE 2: AMOSTRAGEM ALEATORIA

HIPOTESE 3: MEDIA CONDICIONAL ZERO

HIPOTESE 4: COLINEARIDADE NÃO PERFEITA

Sob estas 4 hipóteses os estimadores OLS são estimadores não viesados dos parâmetros da população. Vimos que em algumas situações uma ou mais

destas quatro hipóteses são violadas e os estimadores OLS tornam-se viesados. No caso de ser violada a hipótese 2 e os dados não serem provenientes de uma amostra aleatória simples existem técnicas econométricas que tentam incorporar distintos delineamentos de amostragem para realizar estimativas não viesadas. Veremos adiante a utilização dos comandos `svyreg`, `svylogit` e outros que incorporam as informações do delineamento das amostras para garantir um não viés das estimativas dos parâmetros. A principal hipótese que é violada na prática econométrica é a hipótese 3. Vimos que esta hipótese é violada quando omitimos variáveis relevantes no modelo.

No entanto, vimos que os estimadores não necessitam ser apenas não viesados. Além disto, estes estimadores devem ser eficientes (com mínima variância possível) e devem ser consistentes (para grandes amostras os valores das estimativas tendem a convergir em probabilidade com os valores dos parâmetros desconhecidos).

Até o presente momento não discutimos o comportamento dos erros dos modelos, ou seja, não estabelecemos nenhuma restrição a respeito da distribuição dos erros assim como nada dissemos a respeito da relação entre os erros para todas as observações. No capítulo 6 do livro do Baum (Regression with non – i.i.d. errors) são discutidas hipóteses restritivas a respeito dos erros dos modelos. Anteriormente, na seção 4.2.2 pg 73) o autor analisa a distribuição amostral das estimativas da regressão.

Os estimadores OLS são, de acordo com o Teorema de Gauss-Markov, os melhores estimadores lineares não viesados. Isto quer dizer que dentro da classe de estimadores lineares e não viesados, os estimadores OLS são aqueles que possuem a menor variância possível. Além disso, se os erros são i.i.d. (independentes e identicamente distribuídos) os estimadores OLS são consistentes e a sua distribuição amostral para grandes amostras é normal com média igual ao valor dos parâmetros e a matriz de variância-covariância dos estimadores  $\hat{\beta}$  pode ser estimada de forma consistente de acordo com a expressão:

$$VCE(\hat{\beta}) = s^2 (X'X)^{-1} = \frac{\hat{u}'\hat{u}}{N-k} (X'X)^{-1} \quad (13.1)$$

E o que acontece quando os erros do modelo de regressão não são i.i.d.? Se os erros não são i.i.d. e a hipótese 3 não é violada (média condicional dos erros é nula) os estimadores OLS continuam sendo não viesados e a distribuição amostral destes estimadores para grandes amostras continua sendo normal, sendo estes estimadores além de não viesados também consistentes, ou seja, continuam convergindo em probabilidade para os verdadeiros valores dos parâmetros. Mas, neste caso, a matriz de variância-covariância dos estimadores não pode ser consistentemente estimada pela expressão anterior.

Quando a hipótese i.i.d. falha? Erros não i.i.d. ocorrem em três situações básicas possíveis:

- 1) Erros com distribuição variante de acordo com as observações. O caso mais comum é quando os erros têm variância não constante (conhecido como heterocedasticidade).<sup>7</sup>
- 2) Erros não independentes. Sabemos da teoria estatística que quando duas variáveis aleatórias são independentes sua correlação é nula. No entanto, a recíproca não é verdadeira: podem existir casos de variáveis aleatórias com correlação nula, mas que são dependentes. No entanto, na teoria econometria estamos apenas interessados em verificar se os erros são ou não independentes e a maneira mais simples e direta de fazer isto é ver se o coeficiente de correlação entre os erros é igual a zero ou não. Caso não seja igual a zero estamos certos de que existe não independência (apesar de que se o coeficiente de correlação for nulo não nos dá certeza de que os erros são independentes). É importante destacar que a hipótese de independência dos erros é bem distinta da hipótese de média condicional dos erros nula (Hipótese 3).

---

<sup>7</sup> Na verdade estamos considerando que a distribuição não se altera até o segundo momento representado pela variância. O primeiro momento que é a média dos erros também é constante pois sabemos que OLS geram erros com média nula. Então, restringindo para variância dos erros constantes estamos fixando a distribuição dos erros para os seus dois primeiros momentos.



No primeiro caso os erros são independentes entre si enquanto que no segundo caso os regressores são não correlacionados com os erros.<sup>8</sup>

3) Podemos ter um terceiro caso em que ocorre uma situação híbrida: autocorrelação dos erros não nula e variância não constante.

Estas três situações conduzem a estimativas não consistentes para a matriz de variância-covariância. Ou seja, se tivermos uma situação em que não é violada a hipótese de média condicional dos erros nula, os estimadores dos parâmetros da regressão não serão viesados, no entanto, não teremos mais confiança das estimativas dos erros padrões destes estimadores, mesmo para grandes amostras. Veremos em um capítulo à parte como “corrigir” estes problemas decorrentes do aparecimento de erros não i.i.d. Para o momento queremos destacar um importante aspecto mostrado por Baum: podemos usar estimativas de ponto OLS consistentes com um estimador diferente da matriz de variância-covariância dos estimadores dos parâmetros (VCE) que leve em conta os erros não i.i.d. Esta é a chamada abordagem robusta. O significado do termo robusto advém do fato de que nesta abordagem não impomos nenhuma restrição ao comportamento dos erros do modelo. Em outra abordagem podemos especificar como os erros se desviam da hipótese i.i.d.. Esta é a chamada abordagem eficiente. O STATA possui uma opção (robust) que implementa a primeira abordagem. Através dessa opção, o STATA obtém as estimativas das variâncias pelo método de Huber-White sandwich.

Utilizamos a abordagem robusta quando desconhecemos a forma em que o processo de erros do modelo se desvia na hipótese i.i.d. (tanto que se refere à mudança na sua distribuição como no que se refere a não independência dos erros ou em ambos os casos). Quando conhecemos a maneira em que o processo de erro se desvia da hipótese i.i.d., utilizamos a abordagem eficiente que é mais complexa e exige mais perícia por parte do econometrista. Como verificar se devemos optar por uma ou outra abordagem ou continuarmos com as nossas estimativas da VCE obtidas por OLS? Um primeiro passo é testar se

---

<sup>8</sup> Não esquecer que a hipótese 3 também pode ser interpretada desta forma (independência entre regressores e erros).

os erros são de fato homocedásticos. Para isso, utilizamos o comando **estat hettest, iid**. Vejamos um exemplo:

```
use http://www.stata-press.com/data/imeus/fertil2, clear qui regress ceb
age agefbrth usemeth
estimates store nonRobust
hettest, rhs
summa ceb age agefbrth usemeth children if e(sample)
regress ceb age agefbrth usemeth, robust
estimates store Robust
estimates table nonRobust Robust, b(%9.4f) t(%5.2f) ///
title(Estimativas de CEB com erros padroes OLS e robustos)
```

```
sysuse auto, clear
count
gen lprice = ln(price)
regress lprice mpg weight length foreign
estimates store nonRobust
hettest, rhs
regress lprice mpg weight length foreign, robust
estimates store Robust
estimates table nonRobust Robust, b(%9.4f) t(%5.2f) ///
title(Estimativas de CEB com erros padroes OLS e robustos)
```

No do file acima estamos considerando dois exemplos distintos. No primeiro exemplo, estamos estimando uma regressão com resíduos heterocedásticos. Isto pode ser observado através do resultado do comando **estat hettest**, cujo p-value é igual a zero, rejeitando portanto a hipótese nula de homocedasticidade. Temos então uma situação em que os resíduos (e provavelmente os erros) não são i.i.d. Nesse caso, rodamos o mesmo modelo com opção robusta e pedimos ao STATA para apresentar uma tabela com os resultados do modelo não robusto e do modelo robusto. Verifica-se que os erros padrões das estimativas dos parâmetros diferem bastante de um modelo para o outro.

Em seguida, com a base de dados de automóveis carregada como arquivo de sistema, estimamos também dois modelos. Após a estimativa do primeiro modelo realizamos o teste referente a heterocedasticidade dos erros. Observa-se que o valor do p-value para esse teste é igual a 0,1883, aceitando-se portanto a hipótese nula homocedasticidade. Nesse caso, não temos um problema de erros não i.i.d., pelo menos no que se refere a não constância da variância dos erros. Mesmo assim, estimamos a equação de regressão pelas

duas formas: não robusta e robusta. Observa-se que os erros padrões e as estatísticas  $t$  não se diferenciam substancialmente para os dois modelos, o que comprova que não estamos violando a hipótese i.i.d. com estes últimos dados.

Quanto a este último resultado, Baum tece um comentário importante: se a hipótese de homocedasticidade é válida, o estimador simples da VCE (matriz de variância-covariância dos estimadores dos coeficientes da regressão) é mais eficiente do que a versão robusta. Ou seja, nesse caso não é conveniente utilizar os estimadores robustos de Huber-White-sandwich. Em amostras pequenas, quando confiamos na hipótese de variância constante dos erros, temos que também confiar nos estimadores simples da VCE. No entanto, como é fácil obter os resultados robustos pelo STATA, em grandes amostras, está se tornando comum reportar sempre os resultados com esta opção.

## **14. Regressão com variáveis dummies**

As variáveis dummies são utilizadas basicamente em três tipos de aplicação: 1) representação de efeitos de fatores qualitativos na variável dependente; 2) representação de efeitos sazonais em análise de séries temporais; 3) avaliação e teste de mudanças estruturais em séries temporais. Visaremos nesta seção discutir apenas o primeiro caso. Quando dizemos fatores qualitativos estamos nos referindo a todos os tipos de variáveis que representam categorias: variáveis nominais e variáveis ordinais. Em estatística, estas variáveis são denominadas variáveis categóricas. Um exemplo pode ser a variável sexo que será utilizada como variável independente em um modelo de regressão múltipla para explicar o valor da renda em uma amostra de trabalhadores. A variável sexo terá valor 1 para sexo = masculino e valor zero para sexo = feminino. Desta forma, uma dimensão qualitativa (sexo) é representada numericamente como uma variável binária (ou booleana). Fazendo esta transformação podemos empregar o método OLS para os dados e estimar os parâmetros do modelo de regressão. Vamos considerar inicialmente um exemplo do capítulo 7 do Wooldridge.

use <http://fmwww.bc.edu/ec-p/data/wooldridge/CEOSAL1>, clear

regress wage female educ exper tenure

Source	SS	df	MS	Number of obs = 526		
Model	2603.10658	4	650.776644	F( 4, 521)	=	74.40
Residual	4557.30771	521	8.7472317	Prob > F	=	0.0000
				R-squared	=	0.3635
				Adj R-squared	=	0.3587
				Root MSE	=	2.9576
Total	7160.41429	525	13.6388844			

wage	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
female	-1.810852	.2648252	-6.84	0.000	-2.331109	-1.290596
educ	.5715048	.0493373	11.58	0.000	.4745802	.6684293
exper	.0253959	.0115694	2.20	0.029	.0026674	.0481243
tenure	.1410051	.0211617	6.66	0.000	.0994323	.1825778
_cons	-1.567939	.7245511	-2.16	0.031	-2.991339	-.144538

Neste modelo estamos querendo verificar o efeito da variável sexo sobre o nível de renda controlando com as variáveis educação, tempo de experiência e tempo de permanência no emprego. A variável female tem valor igual a um quando o sexo é feminino e igual a zero em caso contrário. Observa-se que o valor do coeficiente estimado para a variável female é igual a -1,81, o que significa que as mulheres ganham em média 1,81 dólares a menos do que os homens com os mesmos níveis de educação, experiência e permanência no emprego. Controlando-se o efeito destas três últimas, o valor de -1,81 refere-se ao efeito da variável sexo exclusivamente. Podemos também estimar uma regressão sem essas variáveis de controle, tendo apenas sexo como variável independente:

regress wage female

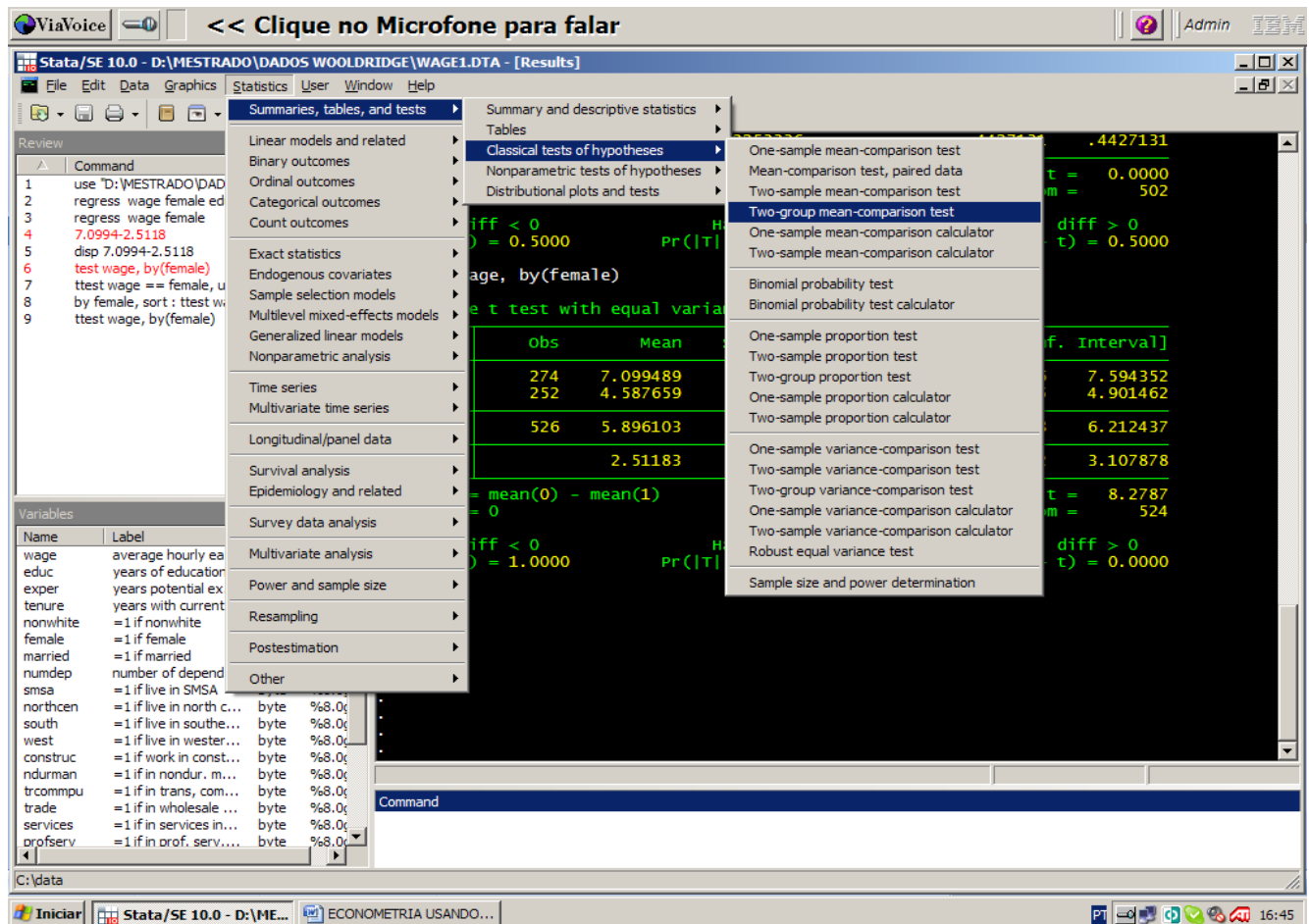
Source	SS	df	MS	Number of obs = 526		
Model	828.220467	1	828.220467	F( 1, 524)	=	68.54
Residual	6332.19382	524	12.0843394	Prob > F	=	0.0000
				R-squared	=	0.1157
				Adj R-squared	=	0.1140
				Root MSE	=	3.4763
Total	7160.41429	525	13.6388844			

wage	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
female	-2.51183	.3034092	-8.28	0.000	-3.107878	-1.915782
_cons	7.099489	.2100082	33.81	0.000	6.686928	7.51205

Neste caso, o valor do termo de intercepto (7.0994) deve ser interpretado como o salário médio dos homens na amostra (female=0). O valor do coeficiente

estimado para a variável dummy (female) corresponde à diferença entre o salário médio das mulheres e dos homens. Desta forma, o salário médio das mulheres é igual a  $7.0994 + (-2.5118) = 4.5876$ . O resultado dessa regressão é comparável a um teste t para comparação de médias populacionais para dois grupos. Podemos fazer esse teste no STATA utilizando o comando `ttest`, de acordo com o seguinte menu:



**Figura 12 – Teste de diferença de médias para subgrupos populacionais**

```
. ttest wage, by(female)
```

Two-sample t test with equal variances

Group	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf. Interval]
0	274	7.099489	.2513666	4.160858	6.604626 7.594352
1	252	4.587659	.1593349	2.529363	4.273855 4.901462
combined	526	5.896103	.1610262	3.693086	5.579768 6.212437
diff		2.51183	.3034092		1.915782 3.107878

diff = mean(0) - mean(1) t = 8.2787

Ho: diff = 0	degrees of freedom =	524
Ha: diff < 0	Ha: diff != 0	Ha: diff > 0
Pr(T < t) = 1.0000	Pr( T  >  t ) = 0.0000	Pr(T > t) = 0.0000

O resultado para esse teste bilateral pode ser visto na tabela acima com o valor da estatística  $t = 8,278$  que corresponde em valor absoluto ao mesmo valor na mesma estatística para o coeficiente da variável binária da regressão anterior. De acordo com esse teste, rejeita-se a hipótese nula de igualdade de médias dos salários para homens e mulheres na população. Conclui-se, portanto, que estimar uma regressão com uma variável independente que é uma dummy corresponde em termos estatísticos a realizar um teste de diferença de médias para os dois grupos populacionais.

Na tabela acima (saída para o comando `ttest wage, by(female)`) verificamos que temos 274 observações na amostra com homens e 252 observações com mulheres. O valor médio para os homens é 7,099 e para as mulheres é 4,567. Temos também os valores dos erros padrões e desvios padrões assim como os intervalos de confiança para estas estimativas. Além disso, a tabela mostra o valor da média dos dados combinados (homens e mulheres juntos). Na linha seguinte temos a diferença entre as médias amostrais (2,518) que não por coincidência é igual a estimativa do coeficiente da variável dummy female. Verificamos que o seu intervalo de confiança de 95 % de probabilidade não contém o zero e, portanto, este valor é significativamente distinto de zero ao nível de 5 %. Isto é comprovado pelos resultados dos testes de hipótese na parte de baixo da tabela. A hipótese nula é que o valor da diferença entre as médias na população é igual a zero. Temos 524 graus de liberdade que é igual ao número de observações (626) menos dois graus de liberdade perdidos, pois estamos estimando duas médias.

Baum chama a atenção para o fato de que regressão com variáveis qualitativas corresponde a uma análise de variância (ANOVA). O mesmo autor sugere que devemos criar as variáveis binárias utilizando o comando `tabulate` com a opção `generate`. Por exemplo:

```
use http://www.stata-press.com/data/imeus/NEdata, clear
mean dpipc, over(state)
```

Mean estimation	Number of obs	=	120
CT: state = CT			

```

MA: state = MA
ME: state = ME
NH: state = NH
RI: state = RI
VT: state = VT

```

	Over	Mean	Std. Err.	[95% Conf. Interval]	
dpipc					
	CT	22.32587	1.413766	19.52647	25.12527
	MA	19.77681	1.298507	17.20564	22.34798
	ME	15.17391	.9571251	13.27871	17.06911
	NH	18.66835	1.193137	16.30582	21.03088
	RI	17.26529	1.045117	15.19586	19.33473
	VT	15.73786	1.020159	13.71784	17.75788

```
. tabulate state, generate(NE)
```

state	Freq.	Percent	Cum.
CT	20	16.67	16.67
MA	20	16.67	33.33
ME	20	16.67	50.00
NH	20	16.67	66.67
RI	20	16.67	83.33
VT	20	16.67	100.00
Total	120	100.00	

Este último comando cria variáveis binárias correspondentes a cada estado. Se a observação pertencer a um determinado estado, a variável binária correspondente a esse estado terá valor um. Em caso contrário terá valor zero. No entanto, em uma regressão teremos que eliminar uma destas seis variáveis binárias para não haver problema de multicolinearidade perfeita. Esse problema existe porque se utilizarmos 6 variáveis binárias para representar seis categorias, uma dessas variáveis será combinação linear perfeita das demais. Vamos então estimar o modelo de regressão para a variável renda disponível per capita (dpipc) desconsiderado a primeira variável binária correspondente ao primeiro estado (CT).

```
. regress dpipc NE2-NE6
```

Source	SS	df	MS			
Model	716.218512	5	143.243702	Number of obs =	120	
Residual	3099.85511	114	27.1917115	F( 5, 114) =	5.27	
Total	3816.07362	119	32.0678456	Prob > F =	0.0002	
				R-squared =	0.1877	
				Adj R-squared =	0.1521	
				Root MSE =	5.2146	

dpipc	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
NE2	-2.549057	1.648991	-1.55	0.125	-5.815695	.7175814
NE3	-7.151959	1.648991	-4.34	0.000	-10.4186	-3.88532
NE4	-3.65752	1.648991	-2.22	0.029	-6.924158	-.3908815
NE5	-5.060575	1.648991	-3.07	0.003	-8.327214	-1.793937

NE6		-6.588007	1.648991	-4.00	0.000	-9.854646	-3.321369
_cons		22.32587	1.166013	19.15	0.000	20.01601	24.63573

Nesse caso, a categoria de referência é o estado CT. Por exemplo, o segundo estado (MA) tem renda disponível per capita -2.5490 inferior a categoria de referência. Podemos verificar esse resultado pela tabela de médias anterior:  $19.7768 - 22.3258 = -2.5490$ . De acordo com Baum podemos verificar a relevância do fator e estados através da estatística F da análise de variância. O valor desta estatística (5.27) e o seu respectivo p-value (0.0002) descartam a hipótese nula de não relevância para esse fator. Nesse caso, estamos fazendo um teste de significância conjunta para todas as 5 variáveis binárias correspondentes a este único fator. Então, para testarmos a significância de uma variável qualitativa em um modelo de regressão, temos que testar a significância conjunta de todas as variáveis binárias correspondentes a esta variável qualitativa.

Vamos agora ver uma regressão com duas variáveis qualitativas.

```
use http://www.stata-press.com/data/imeus/nlsw88, clear
keep if !missing(wage + race + union)
summarize wage race union tenure, sep(0)
tabulate race, generate(R)
gen lwage = ln(wage)
regress lwage R1 R2 union
```

Source	SS	df	MS	Number of obs =	1878
Model	29.3349228	3	9.77830761	F( 3, 1874) =	38.73
Residual	473.119209	1874	.252464893	Prob > F =	0.0000
Total	502.454132	1877	.267690001	R-squared =	0.0584
				Adj R-squared =	0.0569
				Root MSE =	.50246

lwage	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
R1	-.0349326	.1035125	-0.34	0.736	-.2379444 .1680793
R2	-.2133924	.1049954	-2.03	0.042	-.4193126 -.0074721
union	.239083	.0270353	8.84	0.000	.1860606 .2921054
_cons	1.913178	.1029591	18.58	0.000	1.711252 2.115105

```
test R1 R2
```

```
( 1) R1 = 0
( 2) R2 = 0
```

```
F( 2, 1874) = 23.25
Prob > F = 0.0000
```



O primeiro comando importa automaticamente um arquivo da internet. Este é um arquivo de dados de uma pesquisa de dados longitudinais dos EUA. O segundo comando restringe os dados a uma sub-amostra eliminando os missing values para as variáveis wage race e union simultaneamente. Este procedimento é absolutamente dispensável já que o comando regress elimina automaticamente estes missing values para aplicar OLS. O terceiro comando mostra as estatísticas descritivas para as variáveis do modelo. Este é mais um procedimento de verificação que deve ser convenientemente antecipado ao modelo. Devemos ficar sempre atentos a eventuais erros nos dados que perturbarão as estimativas. No final fazemos um teste conjunto de significância para as dummies de raça através do comando test. Rejeitamos a hipótese nula de não significância dos dois parâmetros (coeficientes) correspondentes a raça. Portanto, elas devem permanecer no modelo.

Neste modelo estamos assumindo que o efeito dos dois fatores qualitativos sobre o logaritmo dos salários (lwage) é aditivo, ou seja, eles atuam cada um sobre a variável dependente de forma separada e independente. Temos 3 raças (White, Black e others) e 2 categorias de sindicalização (union = 1 e union = 0). Assim temos uma matriz de 3x2 cruzamentos de categorias. Nos resultados do modelo podemos ver que o efeito médio da passagem de raça = others para raça = White é negativo. No entanto esta estimativa não é significativa ( $t = -0,34$ ). Já o efeito médio da passagem de raça = Others para raça = Black é negativo e significativo. De acordo com Baum (podemos experimentar isto) a magnitude dos coeficientes (e das estatísticas  $t$ ) irá depender da escolha (arbitrária) da categoria de referencia. O que interessa no caso é se devemos ou não manter o conjunto de dummies para a dimensão qualitativa e isto pode ser verificado através do teste conjunto feito anteriormente. Ou seja, as estatísticas  $t$  para cada dummy individual não tem muita importância analítica, a não ser que elas revelam diferenças (significativas ou não) entre contrastes (baseados na comparação entre cada categoria e a categoria de referencia). Em outros termos, ou mantemos ou removemos conjuntamente as dummies referentes a cada dimensão qualitativa – não podemos manter ou remove-las individualmente. Não existe sentido

nenhum econométrico em remover-se uma dummy e deixar o sistema de dummies truncado.

Se a pessoa é Black espera-se que lwage seja 0,213 menor que uma pessoa da categoria de referencia independente do seu status de sindicalização. Se a pessoa é sindicalizada espera-se que tenha um valor para lwage 0,239 superior a uma pessoa não sindicalizada independente da sua raça. Vamos transcrever aqui uma importante passagem do Baum:

“Como este modelo de regressão prediria a renda de um sindicalizado negro em relação à classe excluída (raça = outras e union = não sindicalizado)? Ele predirá meramente como a soma desses dois efeitos, ou  $0,239 - 0,213 = 0,026$ , desde que o efeito de sindicalização é ligeiramente superior ao efeito de raça. Temos uma tabela em duas dimensões 3 x 2 de categorias de raça e sindicalização. Podemos preencher as seis células desta tabela a partir dos 4 coeficientes estimados da regressão. Para que essa abordagem seja viável, precisamos assumir independência dos efeitos qualitativos de forma que o efeito conjunto (refletido por uma célula dentro da tabela) é a soma dos efeitos marginais. O efeito de ser Black e um membro de sindicato é tomado como a soma dos efeitos de ser Black, independente do status de sindicalização, e o de ser sindicalizado, independente da raça.”

Efeitos aditivos independentes dos fatores qualitativos (situação de referencia: union = não sindicalizado e raça = others)

	Branco	Negro	Outra
Sindicalizado	.239083-.0349326 =.2041504	.239083-.2133924 = .0256906	.239083
Não sindicalizado	-.0349326	-.2133924	0

Se relaxarmos esta hipótese de independência aditiva dos efeitos dos fatores qualitativos sobre a variável dependente temos que supor interação entre as variáveis qualitativas. Nesse caso, podemos considerar uma situação no mercado de trabalho: um trabalhador jovem tem maiores dificuldades de obter emprego do que o trabalhador com mais experiência. Ao mesmo tempo, um trabalhador pertencente a um grupo social discriminado também terá barreiras quanto ao acesso ao mercado de trabalho. Um trabalhador jovem que também

pertence a um grupo discriminado poderá ter uma dificuldade potencializada. Talvez os efeitos de ser jovem e ao mesmo tempo discriminado podem ser maiores do que a simples soma dos efeitos individuais de ser jovem e de ser discriminado. Então aqui devemos supor que as duas categorias interagem. Em termos da tabela anterior, temos que definir um modelo em que seja possível obter os valores de suas células diretamente a partir dos coeficientes no modelo. Como devemos considerar esse efeito de interação de variáveis qualitativas em modelo de regressão? No modelo anterior, se temos uma variável (raça) com duas variáveis dummies e outra variável (union) com uma variável dummy, temos que considerar dois termos de interação: a interação de cada dummy de raça com a dummy de union.

```
. generate Rlu = R1*union
(368 missing values generated)
```

```
. generate R2u = R2*union
(368 missing values generated)
```

```
. regress lwage R1 R2 union Rlu R2u
```

Source	SS	df	MS	Number of obs = 1878		
Model	33.3636017	5	6.67272035	F( 5, 1872) = 26.63		
Residual	469.09053	1872	.250582548	Prob > F = 0.0000		
				R-squared = 0.0664		
				Adj R-squared = 0.0639		
				Root MSE = .50058		
Total	502.454132	1877	.267690001			

lwage	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
R1	-.1818955	.1260945	-1.44	0.149	-.4291962	.0654051
R2	-.4152863	.1279741	-3.25	0.001	-.6662731	-.1642995
union	-.2375316	.2167585	-1.10	0.273	-.6626452	.187582
Rlu	.4232627	.2192086	1.93	0.054	-.0066561	.8531816
R2u	.6193578	.2221704	2.79	0.005	.1836302	1.055085
_cons	2.07205	.1251456	16.56	0.000	1.82661	2.317489

```
. test Rlu R2u
```

```
( 1) Rlu = 0
( 2) R2u = 0
```

```
F( 2, 1872) = 8.04
Prob > F = 0.0003
```

Repare que o teste conjunto de significância para os dois termos de interação rejeita a hipótese nula e desta forma verificamos a relevância da interação para o modelo. É conveniente sempre partimos de uma especificação mais geral com todos os termos e realizarmos em seguida os testes de significância conjunta para cada conjunto de dummies.

## Regressão com fatores qualitativos e quantitativos

Vamos considerar agora um modelo de regressão que contenha como regressores, tanto variáveis dummies como variáveis quantitativas. Para isso iremos utilizar o mesmo arquivo de dados empregado anteriormente.

regress lwage R1 R2 union tenure

Source	SS	df	MS	Number of obs = 1868		
Model	77.1526731	4	19.2881683	F( 4, 1863) = 85.88		
Residual	418.434693	1863	.224602626	Prob > F = 0.0000		
Total	495.587366	1867	.265445831	R-squared = 0.1557		
				Adj R-squared = 0.1539		
				Root MSE = .47392		

lwage	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
R1	-.070349	.0976711	-0.72	0.471	-.2619053	.1212073
R2	-.2612185	.0991154	-2.64	0.008	-.4556074	-.0668297
union	.1871116	.0257654	7.26	0.000	.1365794	.2376438
tenure	.0289352	.0019646	14.73	0.000	.0250823	.0327882
_cons	1.777386	.0975549	18.22	0.000	1.586058	1.968715

( 1) R1 = 0  
( 2) R2 = 0

F( 2, 1863) = 29.98  
Prob > F = 0.0000

Podemos verificar claramente que os resultados deste último modelo mostram que o mesmo explica um percentual mais elevado de variação da variável dependente do que o modelo anterior. Como podemos interpretar o coeficiente correspondente a variável tenure? Ele significa que aproximadamente uma variação de mais um ano no emprego atual irá acarretar um acréscimo de 2,89% no salário. E como podemos interpretar o valor do termo constante? Podemos dizer que o valor médio do logaritmo do salário para um trabalhador não sindicalizado de outras raças e com zero anos no emprego atual é de US\$1,78. O que aconteceria se o teste F acima (de exclusão simultânea das variáveis R1 e R2) não rejeitar essa hipótese nula? Em caso de rejeição da hipótese nula, como ocorreu acima, temos seis perfis distintos (linhas retas paralelas) no plano [log(wage), tenure]. Cada um desses perfis corresponde a uma combinação de um valor para a variável raça e um valor para a variável union. Como o coeficiente da variável union é significativo, em caso de não rejeição da hipótese nula, os seis perfis se tornarão apenas dois perfis (2 retas

paralelas no mesmo plano) correspondentes a cada um dos estados de sindicalização.

Podemos nos perguntar se os perfis (retas) para trabalhadores sindicalizados e não sindicalizados no plano  $[\log(\text{wage}), \text{tenure}]$  são paralelos. Uma coisa é verificar se a variável *union* é significativa. Se for significativa, isto quer dizer que ser ou não ser sindicalizado importa na determinação do nível de salários. Mas outra coisa é verificar se para cada valor para o tempo de permanência no emprego atual (*tenure*) este efeito da variável *union* se diferencia. Por exemplo, podemos concluir que para um ano de permanência no emprego atual o diferencial de salários entre trabalhadores sindicalizados e não sindicalizados é uma grandeza  $x$ , enquanto que para dois anos de permanência esse diferencial seja uma grandeza distinta  $y$ . Isto significa que os dois perfis para a variável *union* no plano  $[\log(\text{wage}), \text{tenure}]$  seriam representados por duas retas não paralelas. Em termos econométricos esta problemática se refere a testar a significância estatística do parâmetro referente a uma variável de interação entre as variáveis *union* e *tenure*. Nesse caso necessitamos de apenas uma variável de interação por que a variável qualitativa sindicalização é representada por apenas uma dummy ( pois tem apenas duas categorias).

gen  $uTen = union * tenure$

```
regress lwage R1 R2 union tenure uTen
```

Source	SS	df	MS	Number of obs = 1868		
Model	77.726069	5	15.5452138	F( 5, 1862) = 69.27		
Residual	417.861297	1862	.224415304	Prob > F = 0.0000		
Total	495.587366	1867	.265445831	R-squared = 0.1568		
				Adj R-squared = 0.1546		
				Root MSE = .47372		

lwage	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
R1	-.0715443	.0976332	-0.73	0.464	-.2630264	.1199377
R2	-.2638742	.0990879	-2.66	0.008	-.4582093	-.0695391
union	.2380442	.0409706	5.81	0.000	.157691	.3183975
tenure	.0309616	.0023374	13.25	0.000	.0263774	.0355458
uTen	-.0068913	.0043112	-1.60	0.110	-.0153467	.001564
_cons	1.766484	.0977525	18.07	0.000	1.574768	1.9582

Podemos verificar pelos resultados da regressão acima que a variável de interação não possui um parâmetro significativo ao nível 10%. Isto quer dizer que a nossa hipótese descrita anteriormente não tem evidência empírica: com variações no tempo de permanência no emprego atual não ocorrem distintos

diferenciais de salários entre trabalhadores sindicalizados e não sindicalizados. Esses diferenciais são os mesmos independentemente do valor do tempo de permanência no emprego atual e podem ser representados por retas paralelas no plano [log(wage), tenure].

Uma outra hipótese que pode ser levantada é a de haver uma discriminação racial no processo de promoções dos trabalhadores. Isto significaria que os perfis no plano [log(wage), tenure] para diversas raças seriam distintos. Para testar esta hipótese vamos gerar 2 dummies de interação entre as duas dummies de raça e a variável tenure.

```
. gen R1ten = R1*tenure
```

```
. gen R2ten = R2*tenure
```

```
. regress lwage R1 R2 union tenure R1ten R2ten
```

Source	SS	df	MS	Number of obs =	1868
Model	77.2369283	6	12.8728214	F( 6, 1861) =	57.26
Residual	418.350438	1861	.224798731	Prob > F =	0.0000
Total	495.587366	1867	.265445831	R-squared =	0.1558
				Adj R-squared =	0.1531
				Root MSE =	.47413

lwage	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
R1	-.082753	.1395	-0.59	0.553	-.3563459 .1908398
R2	-.291495	.1422361	-2.05	0.041	-.570454 -.012536
union	.1876079	.0257915	7.27	0.000	.1370246 .2381912
tenure	.0257611	.0186309	1.38	0.167	-.0107785 .0623007
R1ten	.0024973	.0187646	0.13	0.894	-.0343045 .0392991
R2ten	.0050825	.018999	0.27	0.789	-.032179 .0423441
_cons	1.794018	.1382089	12.98	0.000	1.522957 2.065078

```
. test R1ten R2ten
```

```
( 1) R1ten = 0
```

```
( 2) R2ten = 0
```

```

F( 2, 1861) = 0.19
Prob > F = 0.8291

```

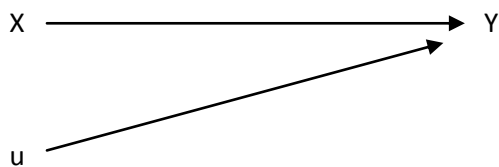
Os resultados do teste F não comprovam a nossa hipótese de discriminação.

## 15. Regressão com Variáveis Instrumentais

O que são métodos de variáveis instrumentais **(IV)**? Mais conhecidos como uma solução para regressores endógenos: variáveis explicativas correlacionadas com o termo de erro da regressão, os métodos de variáveis instrumentais são uma maneira de obter estimativas de parâmetros consistentes. Vamos primeiro considerar um diagrama de causalidade para ilustrar o problema colocado por variáveis instrumentais. Podemos usar mínimos quadrados ordinários (MQO) para estimar consistentemente o seguinte modelo:

**regressão:  $y = xb + u$**  **(15.1)**

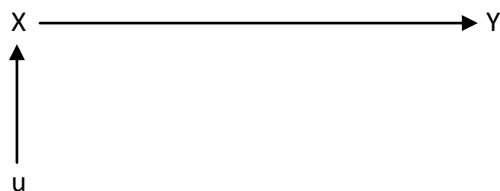
Nenhuma associação entre  $x$  e  $u$ ; MQO é consistente.



Entretanto, a regressão falha na seguinte circunstância:

**Endogeneidade:  $y = xb + u$**

Correlação entre  $x$  e  $u$ ; MQO não é consistente.



A correlação entre  $x$  e  $u$  (ou a falha na hipótese de média condicional nula  $E[u|x] = 0$ ) pode ser causada por muitos fatores.

Podemos nos referir ao problema da endogeneidade como duas ou mais variáveis determinadas conjuntamente em um modelo comportamental. Um exemplo é o modelo de equações simultâneas tal como o conhecido sistema de oferta e demanda em economia, no qual o preço e a quantidade são conjuntamente determinados no mercado.

Um choque ou perturbação tanto na oferta como na demanda afetará tanto o preço como a quantidade no mercado de forma que ambas as variáveis estão correlacionadas com uma perturbação no sistema. Regressão por MQO resultará em estimativas inconsistentes de qualquer regressão incluindo preço e quantidade.

Outra situação em que temos que utilizar variáveis instrumentais é quando temos que levar em conta fatores não observáveis relevantes e que são omitidos da equação de regressão. Tanto  $y$  como  $x$  podem ser afetados por estes fatores latentes, como, por exemplo, a habilidade.

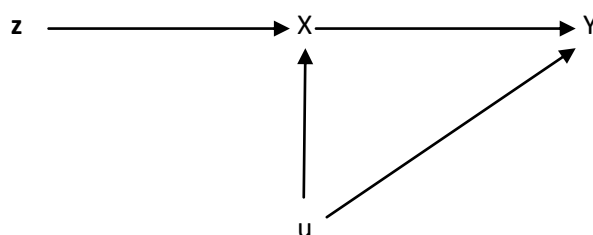
Considere a regressão de  $(\ln)$  rendimentos ( $y$ ) sobre anos de estudo ( $x$ ). O termo de erro  $u$  engloba todos os outros fatores que afetam os rendimentos tais como habilidade inata dos indivíduos ou inteligência.

Mas a habilidade é certamente correlacionada com o grau de escolaridade alcançado, causando uma correlação entre o regressor e o erro, Matematicamente, este é o mesmo problema que aquele causado pela endogeneidade ou erros de medida.

A solução deste problema por variáveis instrumentais pode ser vista como:

**Regressão de variáveis instrumentais:**  $y = \beta x + u$

$z$  não correlacionado com  $u$ , correlacionado com  $x$





A variável adicional  $\mathbf{z}$  é chamada de instrumento para  $\mathbf{x}$ . Em geral, temos muitas variáveis em  $\mathbf{x}$ , e mais de uma destas variáveis correlacionada com  $\mathbf{u}$ . Neste caso, necessitamos no mínimo tantas variáveis em  $\mathbf{z}$ , quantas forem as variáveis em  $\mathbf{x}$  correlacionadas com  $\mathbf{u}$ .

Para tratar do problema de endogeneidade em um sistema de oferta e demanda, um candidato  $\mathbf{z}$  deve afetar a quantidade ofertada, mas não deve impactar diretamente a demanda do produto. Um exemplo para um produto agrícola pode ser a temperatura ou a precipitação pluviométrica: estes fatores são claramente exógenos ao mercado, mas provavelmente importantes no processo de produção.

Consideremos o seguinte sistema de equações de “equilíbrio” de mercado:

$$\begin{aligned}q &= \beta_{1d} + \beta_{2d}p + \beta_{3d}r + u_1 \\p &= \beta_{1o} + \beta_{2o}q + u_2\end{aligned}$$

Se considerarmos a solução algébrica deste sistema de equações estruturais para as variáveis  $\mathbf{p}$  e  $\mathbf{q}$ , teremos as equações na forma reduzida, nas quais os fatores exógenos aparecerão em seus lados direitos.

No caso dos fatores latentes da equação de rendimentos, podemos escolher o instrumento  $\mathbf{z}$  como o número de anos de estudo do pai ou da mãe. Pais com maior escolaridade provavelmente têm filhos com maior escolaridade; ao mesmo tempo, fatores não observáveis que influenciam simultaneamente a renda e o nível educacional dos indivíduos não podem influenciar variáveis cujos valores são definidos no passado, como a escolaridade dos pais.

Mas porque não utilizar sempre variáveis instrumentais? Devemos considerar 3 motivos:

1. Pode ser difícil achar variáveis que servem como instrumentos válidos. Muitas variáveis que têm um efeito sobre as variáveis endógenas incluídas, também têm um efeito direto sobre a variável dependente.

2. Estimadores IV são viesados para pequenas amostras e suas propriedades para amostras finitas são freqüentemente problemáticas. Estes estimadores podem ter resultado ruim em pequenas amostras.
3. A precisão de estimadores IV é menor do que a de estimadores OLS. Na presença de instrumentos fracos (instrumentos incluídos com baixa correlação com os regressores endógenos) a perda de precisão é muito grande e as estimativas IV podem não compensar a inconsistência dos estimadores OLS. Isto sugere a necessidade de um método para determinar se um dado regressor pode ser tratado como endógeno.

Como saber se os instrumentos são fortes?

Instrumentos podem ser fracos: satisfatoriamente exógenos, mas fracamente correlacionados com os regressores endógenos. Neste caso, “a cura pode ser pior do que a doença”.

Alguns autores (ver citação em Baum, 2008), formalizaram a definição de instrumentos fracos: concluem que a estatística F da equação de primeiro estágio deve exceder 10 para que os instrumentos sejam considerados fortes. Mas este critério não é suficiente para considerar que um instrumento não seja fraco. Outros autores (Stock e Yogo, 2005) estabelecem uma regra de bolso para avaliar a fraqueza de instrumentos. Os comandos STATA `ivreg2` e `ivregress` incorporam tabulações referentes a esta regra.

Vamos carregar um arquivo de dados:

**use <http://www.stata-press.com/data/imeus/griliches>, clear**

Para abrir arquivos de dados a partir da Internet temos que primeiro configurar o STATA através do menu Edit => Preferences => General Preferences => Internet e preencher os dados de Proxy de acordo com o seu servidor.

**ivreg lw s expr tenure rns smsa \_l\* (iq= med kww age mrt), first**

Para este comando a matriz Z é formada pelas variáveis: s, expr, tenure, rns, smsa e as dummies de anos (que são os instrumentos incluídos) e as variáveis med, kww, age, mrt (que são os instrumentos excluídos). A variável endógena é iq.

Formemos primeiramente a matriz através do STATA:

```
gen const = 1  
mkmat const s expr tenure rns smsa med kww age mrt _lyear_67-  
_lyear_73, matrix(Z)
```

Podemos fazer esta operação diretamente através do menu Data => Matrices => Convert Variables to Matrix.

```
mkmat lw, matrix(y)  
mkmat const s expr tenure rns smsa iq _lyear_67- _lyear_73, matrix(X)
```

Então pela expressão (8.3) do Baum, os estimadores de IV podem ser obtidos como:

$$\hat{\beta}_{IV} = (Z'X)^{-1}Z'y \quad (15.2)$$

Esta expressão pode ser implementada pelo STATA da seguinte forma:

```
matrix betaiv = inv(Z'*X)*Z'*y
```

Este comando produz um erro de conformabilidade, pois o produto de matrizes não é possível já que o número de colunas na primeira matriz não é igual ao número de linhas da segunda matriz. Isto acontece porque o número de colunas da matriz X não é igual ao número de colunas da matriz Z, pois nesta última estamos incluindo quatro instrumentos para a variável iq.

Neste caso vamos utilizar a formulação de 2SLS (mínimos quadrados em dois estágios) apresentada na página 188 do Baum. Segundo este autor, quando temos mais de um instrumento para uma única variável endógena temos que utilizar um método que obtenha um instrumento “ótimo”. Isto é necessário, pois

aplicando o método IV através da equação  $\hat{\beta}_{IV} = (Z'X)^{-1}Z'y$  para cada instrumento, teremos estimativas distintas.

O método 2SLS baseia-se em duas operações seqüenciais: primeiramente estima-se o valor da variável dependente em uma equação em que a variável dependente é a variável endógena e as variáveis independentes são os instrumentos não incluídos. O método 2SLS permite obter este instrumento “ótimo”, utilizando a seguinte premissa: a estimativa de mínimos quadrados ordinários para a regressão do primeiro estágio fornece a combinação linear ótima para os instrumentos considerados.

A equação de primeiro estágio que teríamos para o caso deste nosso exemplo é:

$$iq = \beta_1 + \beta_2s + \beta_3expr + \beta_4tenure + \beta_5rns + \beta_6smsa + dummies + \beta_7med + \beta_8kww + \beta_9age + \beta_{10}mrt + u$$

A equação de segundo estágio será:

$$lw = \beta_1 + \beta_2s + \beta_3expr + \beta_4tenure + \beta_5rns + \beta_6smsa + dummies + \beta_7pred(iq) + u$$

Observe que na primeira equação (primeiro estágio) temos todas as variáveis exógenas da segunda equação (incluindo as dummies) e mais as variáveis instrumentais (que também são supostamente exógenas no modelo). Na segunda equação, substituímos a variável endógena  $iq$  pelo valor de sua predição na primeira equação.

De um ponto de vista de álgebra linear podemos entender este procedimento da seguinte forma: quando estimamos a primeira equação estamos calculando os valores estimados para uma matriz de regressores  $Z$  (que é a matriz  $X$  substituindo-se a variável endógena  $iq$  pelos 4 instrumentos). Neste caso as variáveis dependentes correspondem à matriz  $X$ . A equação matricial de estimação OLS  $Y_{est} = X(X'X)^{-1}X'Y$  torna-se  $X_{est} = Z(Z'Z)^{-1}Z'X$ . Para simplificar podemos chamar o produto matricial  $Z(Z'Z)^{-1}Z'$  de  $P_Z$ .

Assim, de acordo com a expressão (28), temos o seguinte desenvolvimento:

$$\begin{aligned}\hat{\beta}_{2SLS} &= (\hat{X}'X)^{-1} \hat{X}'y = \{X'Z(Z'Z)^{-1}Z'X\} \{X'Z(Z'Z)^{-1}Z'y\} \\ &= (X'P_zX)^{-1} X'P_zy\end{aligned}\tag{15.3}$$

Este cálculo pode ser realizado pelo STATA através da seguinte seqüência de comandos:

```
matrix PZ = Z*inv(Z'*Z)*Z'  
matrix beta2sls = inv(X'*PZ*X)*X'*PZ*y  
matrix list beta2sls
```

Compare estes resultados com aqueles obtidos pelo comando ivreg anterior.

Podemos também calcular as variâncias das estimativas. De acordo com Baum, se os erros forem i.i.d. um estimador consistente para grandes amostras da matriz VCE (matriz de variância-covariância dos estimadores  $\hat{\beta}$ ) é dada pela expressão:

$$Var(\hat{\beta}_{2SLS}) = \hat{\sigma}^2 \{X'Z(Z'Z)^{-1}Z'X\}^{-1} = \hat{\sigma}^2 (X'P_zX)^{-1}\tag{15.4}$$

Utilizando os recursos de álgebra linear do STATA temos a seguinte seqüência de comandos:

```
matrix u = y - X*beta2sls  
matrix sigma2 = (1/_N)*u'*u  
scalar sigma2 = sigma2[1,1]  
matrix varbeta2sls = sigma2*inv(X'*PZ*X)  
matrix list varbeta2sls  
disp sqrt(varbeta2sls[7,7])
```

Vamos a seguir percorrer alguns tópicos do capítulo 15 do Wooldridge.

Considere o problema de aptidão não observada em uma equação de salários de trabalhadores adultos. Este é um problema muito comum em modelos explicativos da renda. Consideremos que a equação da hiper-reta da população é dada pela equação:

$$\log(salario) = \beta_0 + \beta_1 educ + \beta_2 aptid + e\tag{15.5}$$

Não temos uma variável proxy para representar a variável não observável aptidão. Então colocamos esta variável no termo de erro:

$$\log(\text{salario}) = \beta_0 + \beta_1 \text{educ} + u \quad (15.6)$$

O estimador OLS de  $\beta_1$  será viesado e inconsistente se educ e aptid forem correlacionados.

Vamos descrever um modelo de regressão simples como:

$$y = \beta_0 + \beta_1 x + u \quad (15.7)$$

supomos que x e u são correlacionados.

$$\text{Cov}(x, u) \neq 0 \quad (15.8)$$

Quando x e u forem não correlacionados utiliza-se OLS.

Temos que obter uma nova variável z que não é correlacionada com u, mas que é correlacionada com a variável x. A variável z é denominada instrumento de x (ou é uma variável instrumental de x). As hipóteses necessárias para que z seja um instrumento são:

$$\begin{aligned} \text{Cov}(z, u) &= 0 \\ \text{Cov}(z, x) &\neq 0 \end{aligned} \quad (15.9)$$

Z não deve ser correlacionada com fatores não observados que afetam y.

Não podemos testar a primeira hipótese de (15.9), pois u não é observável.

Podemos testar apenas a segunda hipótese de (15.9).

Para a equação  $\log(\text{salário})$  uma variável instrumental para educ deve ser (1) não correlacionada com aptidão (e com qualquer outro fator não observável que afete o salário e (2) correlacionada com educ.

Uma variável instrumental pobre neste caso pode ser os quatro últimos dígitos da carteira de identidade do trabalhador. Esta variável certamente não será correlacionada com u, pois é aleatória. Mas também pela mesma razão não

será com a variável educ. O problema oposto seria uma Proxy da variável omitida (aptidão), por exemplo, QI.

Utilizando a equação (15.7) e calculando a covariância de  $z$  para ambos os lados desta expressão, podemos escrever:

$$\text{cov}(z, y) = \beta_1 \text{cov}(z, x) + \text{cov}(z, u) \quad (15.10)$$

Como por hipótese  $\text{cov}(z, u) = 0$  temos que:

$$\beta_1 = \frac{\text{cov}(z, y)}{\text{cov}(z, x)} \quad (15.11)$$

Para uma amostra aleatória, temos:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (z_i - \bar{z})(y_i - \bar{y})}{\sum_{i=1}^n (z_i - \bar{z})(x_i - \bar{x})} \quad (15.12)$$

Observe pela expressão (15.12) que quando  $z = x$  obtemos a estimativa OLS para  $\beta_1$ .

O estimador VI é consistente para  $\beta_1$ :  $\text{plim}(\hat{\beta}_1) = \beta_1$ , desde que as hipóteses (15.9) sejam satisfeitas.

Quando  $z$  e  $u$  forem de fato correlacionados, o estimador VI será sempre viesado (mesmo para amostras grandes). Para amostras pequenas o estimador VI pode ter um viés grande.

### ***Inferência estatística com estimador de VI***

Como as estruturas dos estimadores IV e OLS são similares, os primeiros também têm distribuição normal em amostras grandes.

Primeiramente impomos a hipótese de homocedasticidade:

$$E[u^2 | z] = \sigma^2 = \text{Var}(u) \quad (15.2)$$

Demonstra-se que sob as hipóteses (15.4), (15.5) e (15.11):

$$\lim_{n \rightarrow \infty} \text{var}(\hat{\beta}_1) = \frac{\sigma^2}{n\sigma_x^2 \rho_{x,y}^2} = \frac{\sigma^2}{SQT_x \cdot R_{x,z}^2} \quad (15.3)$$

A expressão acima se refere à variância assintótica de  $\hat{\beta}_1$ .

Como no caso de OLS, a variância assintótica de IV decresce para 0 quando  $n \rightarrow \infty$

Todas as quantidades em (15.15) podem ser consistentemente estimadas através de uma AAS.

Para estimar  $\rho_{x,y}^2$ , podemos estimar a regressão de x sobre Z e obter o valor de  $R^2$ . Para estimar  $\sigma^2$ , podemos estimar os resíduos de IV e (utilizando as estimativas de IV) empregar a expressão:

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^n \hat{u}_i^2 \quad (15.4)$$

Sob as hipóteses de Gauss-Markov a variância do estimador OLS é  $\sigma^2 / SQT_x$  enquanto que para o estimador IV é  $\sigma^2 / (SQT_x \cdot R_{x,z}^2)$

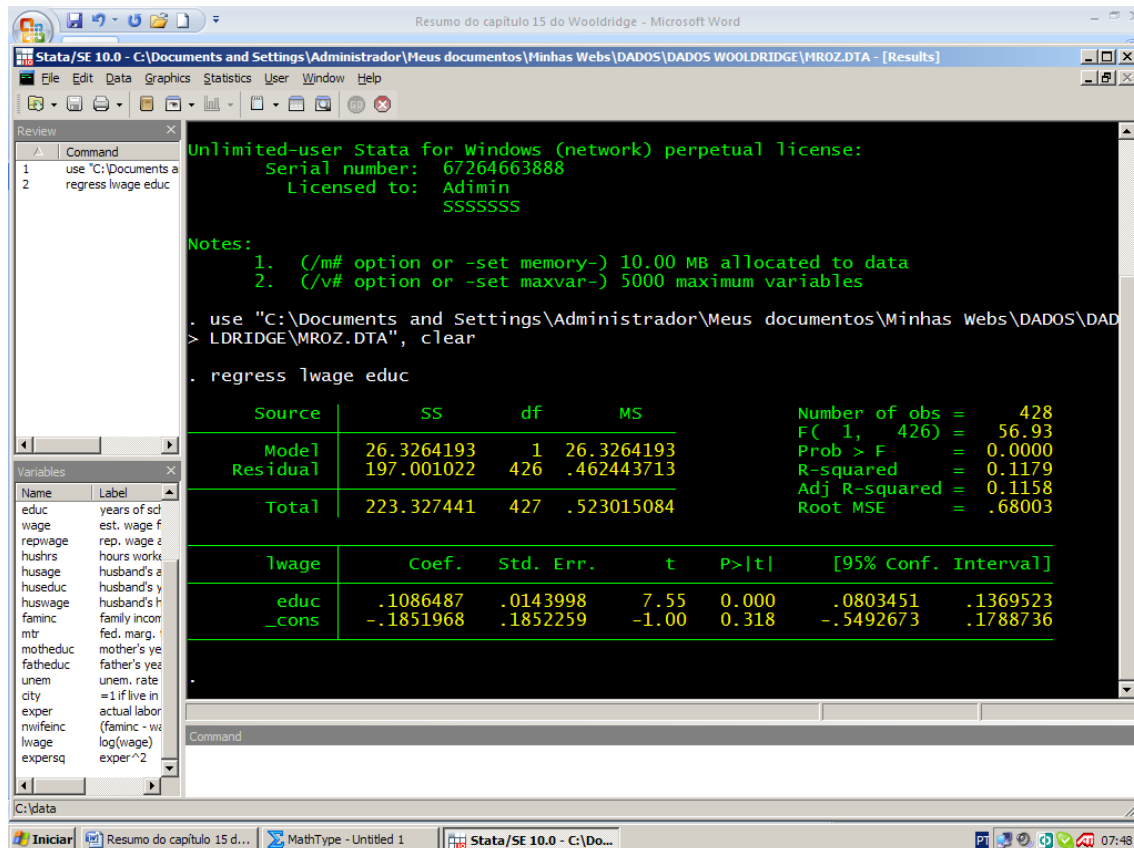
Quanto mais altamente correlacionados forem x e z, menor será a variância do estimador de IV.

No caso em que  $z = x$ , o estimador de IV se confunde com o estimador OLS.

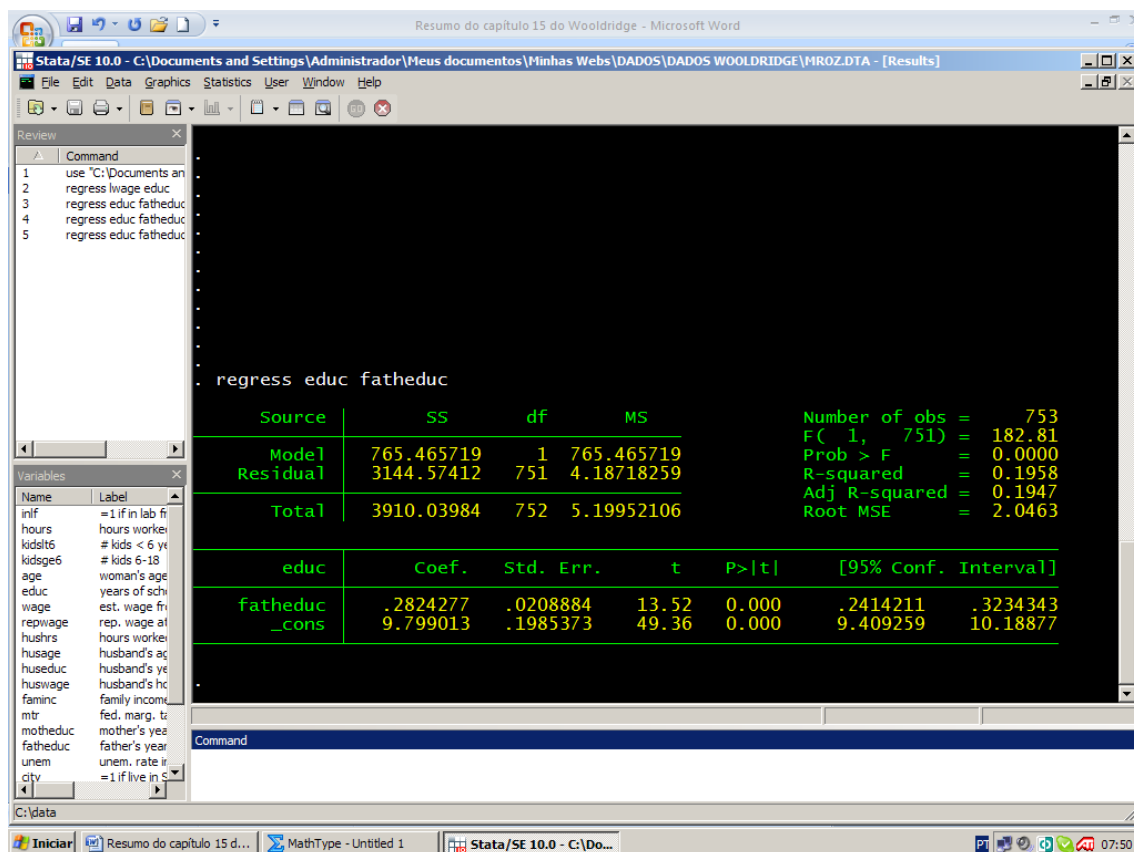


## EXEMPLO 1 – Estimador do retorno da educação para mulheres casadas.

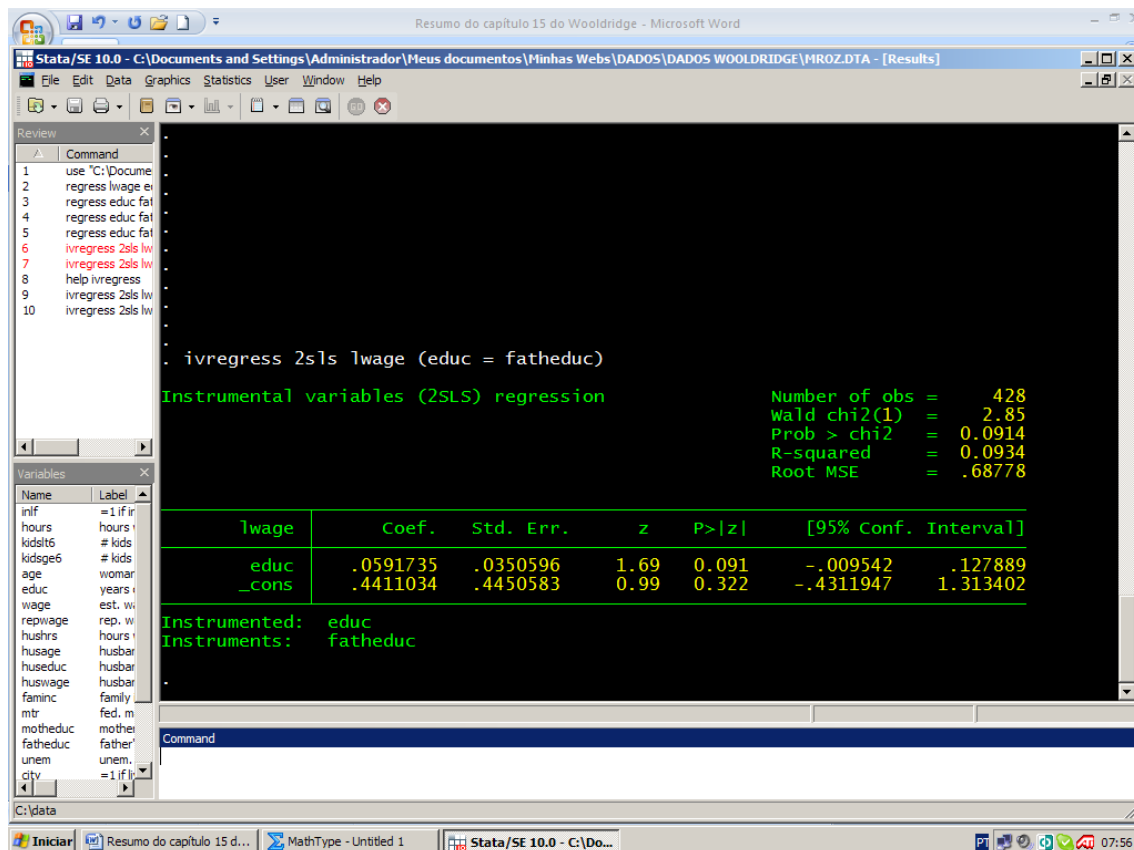
### Regressão OLS



**Figura 13** – Regressão OLS da variável endógena (educ) com o instrumento (fatheduc)



## Regressão IV:



Embora as diferenças nas estimativas de OLS e IV sejam grandes (e as estimativas dos erros padrões também o sejam) não podemos dizer que esta diferença seja estatisticamente significativa.

## EXEMPLO 15.2 Estimação do retorno da educação para homens

```

. use "C:\Documents and Settings\Administrador\Meus documentos\Minhas Webs\DADOS\WAGE2.DTA", clear
. ivregress 2sls lwage (educ = sibs)

Instrumental variables (2SLS) regression

Number of obs =      935
Wald chi2(1)    =    21.63
Prob > chi2     =    0.0000
R-squared       =      .
Root MSE       =    .42284

+-----+-----+-----+-----+-----+-----+
| lwage |   Coef.   | Std. Err. |      z   | P>|z|   | [95% Conf. Interval] |
+-----+-----+-----+-----+-----+-----+
| educ  |   .1224326 |   .0263224 |    4.65  |  0.000  |   .0708417   .1740235 |
| _cons |   5.130026 |   .3547911 |   14.46  |  0.000  |   4.434648   5.825404 |
+-----+-----+-----+-----+-----+-----+

Instrumented:  educ
Instruments:   sibs

+-----+-----+-----+-----+-----+-----+
| Name | Label |
+-----+-----+-----+-----+-----+-----+
| wage | monthly e |
| hours | average v |
| IQ | IQ score |
| KWW | knowledge |
| educ | years of e |
| exper | years of v |
| tenure | years with |
| age | age in yea |
| married | =1 if marr |
| black | =1 if black |
| south | =1 if live i |
| urban | =1 if live i |
| sibs | number of |
| brthord | birth orde |
| meduc | mother's e |
| feduc | father's e |
| lwage | natural lo |
+-----+-----+-----+-----+-----+-----+

```

**Comentário importante:** mesmo uma pequena correlação entre  $z$  e  $u$  pode causar sérios problemas para o estimador de IV.

**Outra aplicação:** análise dos efeitos do fato de ser veterano da Guerra do Vietnã sobre os salários.

$$\log(\text{ganhos}) = \beta_0 + \beta_1 \text{veterano} + u \quad (15.5)$$

onde veterano é uma variável binária.

Estamos aqui corrigindo um problema de auto- seleção: os que ganham mais têm mais propensão a se alistar ou a decisão de se alistar esteja correlacionada com outras características que afetam os ganhos. Isto fará com que  $veterano$  e  $u$  sejam correlacionadas e as estimativas OLS sejam viesadas. Um bom IV neste caso seria o numero de sorteio dos veteranos, pois estes sendo aleatórios não estariam correlacionados com  $u$ . Entretanto como a regra adotada nos EUA era de que quanto mais baixo o numero de sorteio maior a probabilidade de ser convocado, estes números de sorteio estariam correlacionados com a probabilidade de ser veterano.

**Questão 15.1** Se alguns dos homens que receberam números baixos no sorteio militar obtivessem maior escolaridade para reduzir a probabilidade de serem selecionados, o número do sorteio seria uma boa variável instrumental de veterano em (15.14)?

Neste caso, o numero aleatório de sorteio seria correlacionado com a variável escolaridade. Como escolaridade não está na equação (15.14) ela está contida em  $u$  e portanto  $z$  é correlacionada com  $u$ , o que a torna um mau IV.

### PROPRIEDADES DA IV COM UMA VARIÁVEL INSTRUMENTAL POBRE

Embora os estimadores de VI sejam consistentes quando  $z$  e  $u$  são não correlacionados e  $z$  e  $x$  tem qualquer correlação positiva ou negativa, as estimativas de VI podem ter grandes erros padrão, especialmente se  $z$  e  $x$  forem fracamente correlacionados. O estimador de VI também pode ter um grande viés assintótico mesmo se  $z$  e  $u$  forem só moderadamente correlacionados.

$$p \lim \hat{\beta}_1 = \beta_1 + \frac{corr(z,u)}{corr(z,x)} \cdot \frac{\sigma_u}{\sigma_x} \quad (15.6)$$

Mesmo se  $corr(z,u)$  for pequena, a inconsistência no estimador IV pode ser muito grande se  $corr(z,x)$  também for pequena.

Mas  $corr(x,u) = cov(x,u) / (\sigma_x \sigma_u)$ . Então,

$$p \lim \hat{\beta}_1 = \beta_1 + \text{cov}(x, u) / \sigma_u^2 = \beta_1 + \text{corr}(x, u) \sigma_x \sigma_u / \sigma_u^2 \quad (15.7)$$

Portanto:

$$p \lim \hat{\beta}_1 = \beta_1 + \text{corr}(x, u) \cdot \frac{\sigma_u}{\sigma_x} \quad (15.8)$$

**IV** é preferível a **OLS** em termos de viés assintótico quando  $\text{corr}(z, u) / \text{corr}(z, x) < \text{corr}(x, u)$

### **EXEMPLO 18.3 A Estimação do Efeito do Hábito de Fumar sobre o Peso de Nascimento**

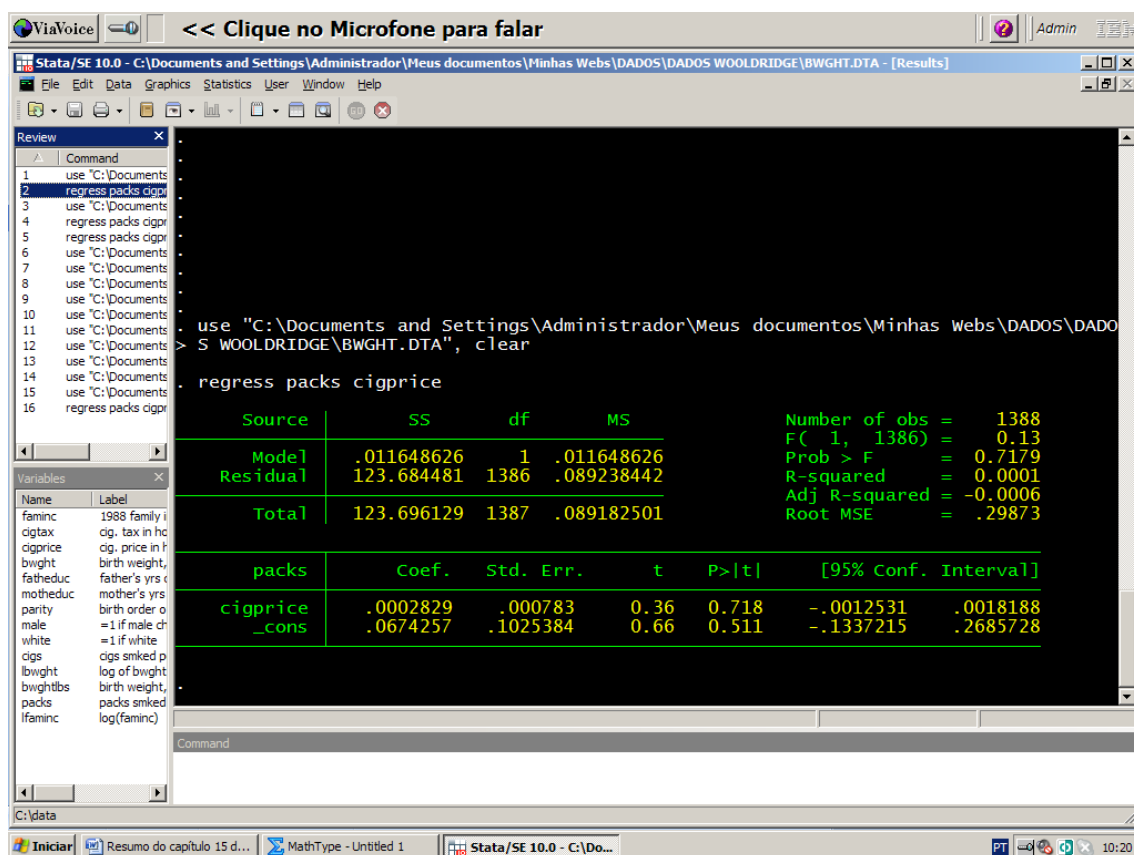
Na especificação do modelo:

$$\log(\text{pesonas}) = \beta_0 + \beta_1 \text{maços} + u \quad (15.9)$$

maços pode estar correlacionada com outros fatores relativos a saúde ou a existência de um bom procedimento pré-natal, de forma que maços e u podem estar correlacionados.

Possível IV: preço médio dos cigarros no Estado de residência, precig

Assume-se que precig e u são não correlacionados (embora o sistema de saúde estadual possa estar correlacionado com os impostos sobre cigarros).



Como  $R^2$  é muito baixo, não devemos usar cigprice como IV de maços (packs). E se o fizéssemos, o que ocorreria? Veja a seguir como ativar a opção de M2QO no STATA:

ViaVoice << Clique no Microfone para falar Admin

Stata/SE 10.0 - C:\Documents and Settings\Administrador\Meus documentos\Minhas Webs\DADOS\WORLDWIDE\BWGHT.DTA - [Results]

File Edit Data Graphics Statistics User Window Help

Review

Command

```

1 use "C:\Documents
2 regress packs cigr
3 use "C:\Documents
4 regress packs cigr
5 regress packs cigr
6 use "C:\Documents
7 use "C:\Documents
8 use "C:\Documents
9 use "C:\Documents
10 use "C:\Documents
11 use "C:\Documents
12 use "C:\Documents
13 use "C:\Documents
14 use "C:\Documents
15 use "C:\Documents
16 regress packs cigr

```

Variables

Name	Label
faminc	1988 family i
cigtax	cig. tax in ho
cigprice	cig. price in h
bwght	birth weight,
fatheduc	father's yrs o
motheduc	mother's yrs
parity	birth order o
male	=1 if male ch
white	=1 if white ch
cigs	cigs smked p
lbwght	log of bwght
bwghtlbs	birth weight,
packs	packs smked
lfaminc	log(faminc)

Statistics

- Summaries, tables, and tests
- Linear models and related
  - Binary outcomes
  - Ordinal outcomes
  - Categorical outcomes
  - Count outcomes
- Exact statistics
- Endogenous covariates
  - Single-equation instrumental-variables regression
  - Three-stage least squares
  - Probit model with endogenous covariates
  - Tobit model with endogenous covariates
- Sample selection models
- Multilevel mixed-effects models
- Generalized linear models
- Nonparametric analysis
- Time series
  - Multivariate time series
- Longitudinal/panel data
- Survival analysis
- Epidemiology and related
- Survey data analysis
- Multivariate analysis
- Power and sample size
- Resampling
- Postestimation
- Other

Results

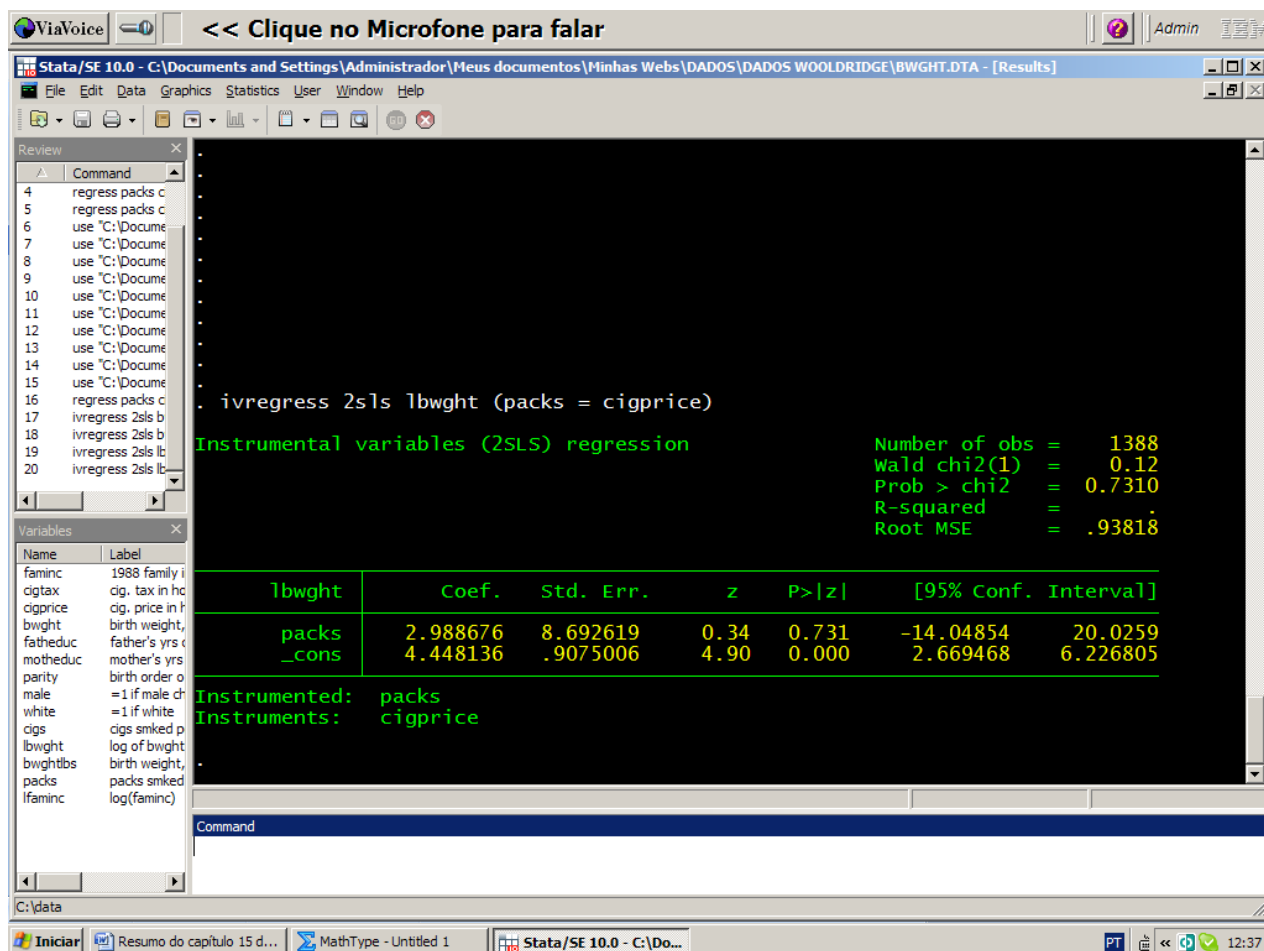
SS	df	MS	Number of obs =	1388
48626	1	.011648626	F( 1, 1386) =	0.13
34481	1386	.089238442	Prob > F =	0.7179
			R-squared =	0.0001
			Adj R-squared =	-0.0006
			Root MSE =	.29873

coef.	Std. Err.	t	P> t	[95% Conf. Interval]
.2829	.000783	0.36	0.718	-.0012531 .0018188
.4257	.1025384	0.66	0.511	-.1337215 .2685728

Command

C:\data

Iniciar Resumo do capítulo 15 d... MathType - Untitled 1 Stata/SE 10.0 - C:\Do... 10:22



As estimativas não tem significado, pois sigprice não satisfaz a única hipótese que poremos testar (15.5).

## O CALCULO DE $R^2$ APÓS A ESTIMAÇÃO DE IV

A SQR da VI pode ser maior do que SQT e isto faz com que o  $R^2$  da VI seja negativo pela formula  $R^2 = 1 - \text{SQR}/\text{SQT}$

O grau de ajuste não é um fator importante em uma estimação IV. Mesmo em uma estimação OLS um elevado  $R^2$  é de pouca importância se não pudermos estimar consistentemente  $\beta_1$ .

## ESTIMAÇÃO DE VI DO MODELO DE REGRESSAO MULTIPLA

Seja o modelo:

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + u_1 \quad (15.10)$$



Em que  $Z_1$  é exógena (a correlação entre  $Z_1$  e  $u_1$  é nula) e  $y_2$  tem uma suspeita de ser endógena (como também é  $y_1$ , pois tem correlação com  $u_1$  distinta de zero). Vamos supor que  $u_1$  contem uma variável que é correlacionada com  $y_2$ .

Um exemplo específico deste modelo é:

$$\log(\text{salario}) = \beta_0 + \beta_1 \text{educ} + \beta_2 \text{exper} + u_1 \quad (15.11)$$

Se (15.20) for estimada por OLS todos os estimadores serão viesados e inconsistentes.  $y_2 = \text{educ}$  é considerada endógena pelas razões habituais.

$Z_1$  não pode ser utilizada como variável instrumental de  $y_2$  (apesar de assumirmos que  $\text{corr}(z_1, u_1) = 0$ ) porque ela já é uma variável explicativa de (15.20).

Deve ser utilizada outra variável exógena,  $z_2$  que não está contida na equação (5.20).

Definimos as seguintes hipóteses:

$$E(u_1) = 0, \text{Cov}(z_1, u_1) = 0, \text{Cov}(z_2, u_1) = 0 \quad (15.12)$$

Em termos amostrais estas hipóteses podem ser formuladas como:

$$\begin{aligned} \sum_{i=1}^n (y_{i1} - \hat{\beta}_0 - \hat{\beta}_1 y_{i2} - \hat{\beta}_2 z_{i1}) &= 0 \\ \sum_{i=1}^n z_{i1} (y_{i1} - \hat{\beta}_0 - \hat{\beta}_1 y_{i2} - \hat{\beta}_2 z_{i1}) &= 0 \\ \sum_{i=1}^n z_{i2} (y_{i1} - \hat{\beta}_0 - \hat{\beta}_1 y_{i2} - \hat{\beta}_2 z_{i1}) &= 0 \end{aligned} \quad (15.13)$$

Estas são equações que podem ser facilmente resolvidas com os dados de  $y_1, y_2, Z_1$  e  $Z_2$ .

Se  $z_2 = y_2$  (15.22) tornam-se as condições de primeira ordem dos estimadores OLS.

A variável  $z_2$  deve ser correlacionada com  $y_2$ :

$$y_2 = \pi_0 + \pi_1 z_1 + \pi_2 z_2 + v_2 \quad (15.14)$$

onde:

$$E(v_2) = 0, Cov(z_1, v_2) = 0 \text{ e } Cov(z_2, v_2) = 0 \quad (15.15)$$

A condição de identificação fundamental (com (15.22)) é que:

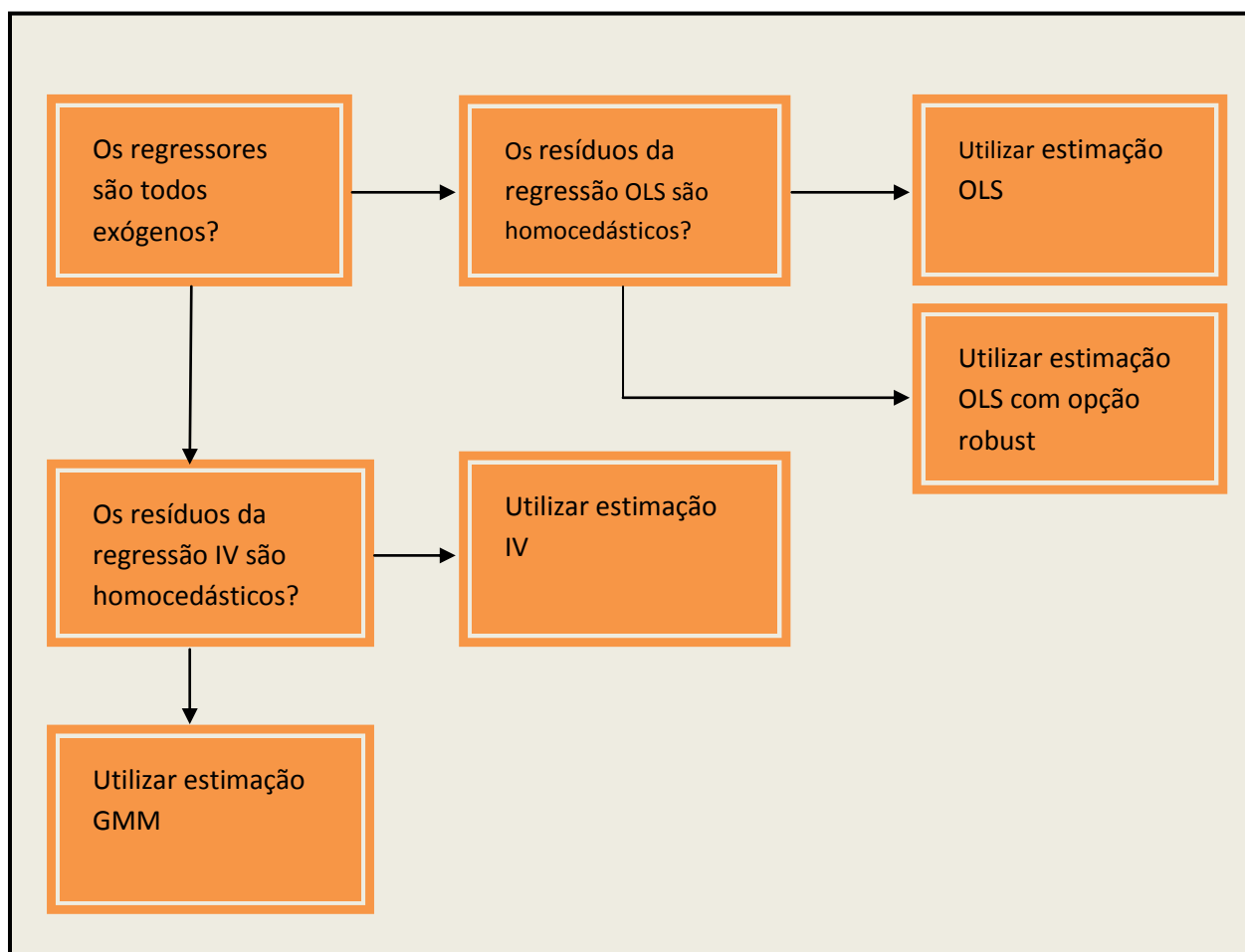
$$\pi_2 = 0 \quad (15.16)$$

## 16. Método dos Momentos Generalizados (GMM)

Vamos fazer uma pausa novamente e tentar sistematizar o que vimos até agora sobre variáveis instrumentais. Quando temos certeza de que os regressores da nossa equação não estão correlacionados com os erros podemos aplicar o método convencional de OLS. No entanto, mesmo nesse caso temos que verificar se os resíduos da regressão são homocedásticos. Então temos que realizar o teste heterocedasticidade. Caso os resíduos sejam heterocedásticos temos que realizar a regressão robusta. Isto pode ser feito utilizando a opção `robust` (após a vírgula) no comando `regress`. Caso tenhamos motivos para acreditar que um ou mais regressores sejam endógenos (tenham correlação não nula com termo de erro da equação) temos que aplicar o método das variáveis instrumentais. Então nesse caso utilizaremos o comando `ivreg` (ou através do menu `endogenous covariates`) ao invés do comando `regress`.

Mas mesmo nesse caso podemos ter uma complicação. Pode acontecer que aplicando o método das variáveis instrumentais os resíduos do modelo não sejam homocedásticos. Nesse caso temos que aplicar o método das variáveis instrumentais articulado com o método dos momentos generalizados (GMM).

Atribui-se freqüentemente o método GMM a Lars Hansen em seu paper de 1982 na revista *Econometrica*. Mas o método tem seus antecedentes nos trabalhos de Karl Pearson sobre o método dos momentos datados em 1895 e mais a frente (1928) nos trabalhos de Neyman e Egon Pearson sobre o método MCE que supera a dificuldade do método dos momentos quando temos mais condições de momentos do que parâmetros a serem estimados. O método tem, portanto, como qualquer descoberta científica, uma história bem definida.



**Figura 14** - Diagrama de Decisão para escolha do método de estimação

De acordo com Baum, o estimador padrão IV e o estimador IV em dois estágios são casos especiais do estimador GMM. A hipótese de que os instrumentos  $\mathbf{z}$  são exógenos pode ser expressa como um conjunto de condições de momento  $E[\mathbf{z}u]=0$ . Estas condições de momento expressam que a covariância entre as variáveis instrumentais e o termo de erro do modelo é igual a zero. Se tivermos  $I$  instrumentos teremos então um conjunto de  $I$  momentos expresso em equações:

$$g_i(\beta) = \mathbf{Z}_i' u_i = \mathbf{Z}_i' (y_i - x_i \beta) \quad (16.1)$$

No caso de termos o número de instrumentos (incluídos e não incluídos na equação de primeiro estágio) o resultado das estimativas GMM coincide com o resultado das estimativas padrão IV. Nesse caso temos uma variável endógena e apenas um instrumento para representá-la. Temos então  $I = K$ , o número de regressores (endógenos e exógenos) é igual ou número de instrumentos (incluídos e não incluídos). No artigo de Baum (2003) temos uma boa representação:

Regressores  $X = [X_1 \ X_2] = [X_1 \ Z_2] = [\text{Endógenos} \ \text{Exógenos}]$

Instrumentos  $Z = [Z_1 \ Z_2] = [\text{Excluídos} \ \text{Incluídos}]$

Temos  $K_1$  regressores  $X_1$  que são endógenos e  $K - K_1$  regressores remanescentes  $X_2$  que são exógenos. Os regressores exógenos que fazem parte da matriz  $X_2$  são também considerados instrumentos (incluídos). Então a matriz de regressores exógenos  $X_2$  é idêntica a matriz de instrumentos incluídos  $Z_2$ . Por outro lado, os  $K_1$  regressores  $X_1$  endógenos são representados por um conjunto de instrumentos excluídos  $Z_1$  (cujo número nem sempre é igual a  $K_1$ ). Quando temos um número de instrumentos excluídos (matriz  $Z_1$ ) igual ao número regressores endógenos (matriz  $X_1$ ) dizemos que a equação é exatamente identificada. Neste caso, temos tantas equações – as  $L$  condições de momento – quanto incógnitas – os  $K$  coeficientes na matriz (ou vetor)  $\beta$ . Para esta situação particular, o estimador GMM torna-se um estimador IV.

Quando a equação é sobre identificada, de forma que  $L > K$ , então temos mais equações do que incógnitas e não podemos achar um vetor  $\beta$  que torne nulas todas as  $K$  condições de momento. Não vamos desenvolver aqui como se obtém matematicamente (na forma matricial) o estimador GMM. Apenas apresentaremos as fórmulas para estimador do vetor de parâmetros e para variância assintótica. Sugerimos ao leitor pesquisar no Baum ou em seu artigo a respeito dos detalhes técnicos desse desenvolvimento. As expressões consideradas são:

$$\begin{aligned}\hat{\beta}_{EGMM} &= (X'Z'S^{-1}Z'X)^{-1}X'ZS^{-1}Z'y \\ V(\hat{\beta}_{EGMM}) &= \frac{1}{n}(Q'_{XZ}S^{-1}Q_{XZ})^{-1}\end{aligned}\quad (16.2)$$

onde:

S é a matriz de covariância das condições de momento.

$$S = \frac{1}{n}E[Z'uu'Z] = \frac{1}{n}E[Z'\Omega Z]$$

$$Q_{XZ} = E[X_1'Z_i]$$

$\Omega$  é a matriz de variância-covariância dos resíduos  $u$  do modelo.  $Q_{XZ}$  é a covariância entre as variáveis endógenas e os respectivos instrumentos. Para entendemos bem essa mecânica vamos desenvolver uma seqüência de comandos (do file) para o STATA. Mas antes disso vamos verificar como podemos gerar estimativas GMM em um contexto heterocedástico. Consideremos para isto a seguinte matriz de variância-covariância para os resíduos do modelo:

$$\bar{\Omega} = \begin{pmatrix} \hat{u}_1^2 & & & 0 \\ & \ddots & & \\ & & \hat{u}_i^2 & \\ & & & \ddots \\ 0 & & & & \hat{u}_n^2 \end{pmatrix}$$

Esta matriz representa um caso geral de heterocedasticidade quando não conhecemos a sua forma. As fórmulas para o estimador GMM nesse caso de heterocedasticidade geral são um pouco mais complicadas do que a expressão (15.24) para o homocedasticidade. Novamente não vamos apresentar a derivação destas fórmulas. Pedimos ao leitor para buscar nas referências anteriores citadas os detalhes técnicos para isto. As expressões matemáticas para o estimador GMM em um caso de heterocedasticidade com forma desconhecida são:

$$\hat{\beta}_{EGMM} = (X'Z'(Z'\bar{\Omega}Z)^{-1}Z'X)^{-1}X'Z(Z'\bar{\Omega}Z)^{-1}Z'y$$

$$V(\hat{\beta}_{EGMM}) = \frac{1}{n}(X'Z(Z'\bar{\Omega}Z)^{-1}Z'X)^{-1} \quad (16.3)$$

Repare que a diferença entre esta última expressão e a expressão anterior é que apenas substituímos o valor da matriz de variância-covariância das condições de momento S pelo seu valor específico para esse caso geral de heterocedasticidade. Admitimos que podemos estimar a matriz de variância-covariância dos resíduos do modelo. Finalmente agora estamos minimamente preparados, com certo embasamento teórico, para utilizarmos os recursos do STATA referentes as estimativas IV com GMM.

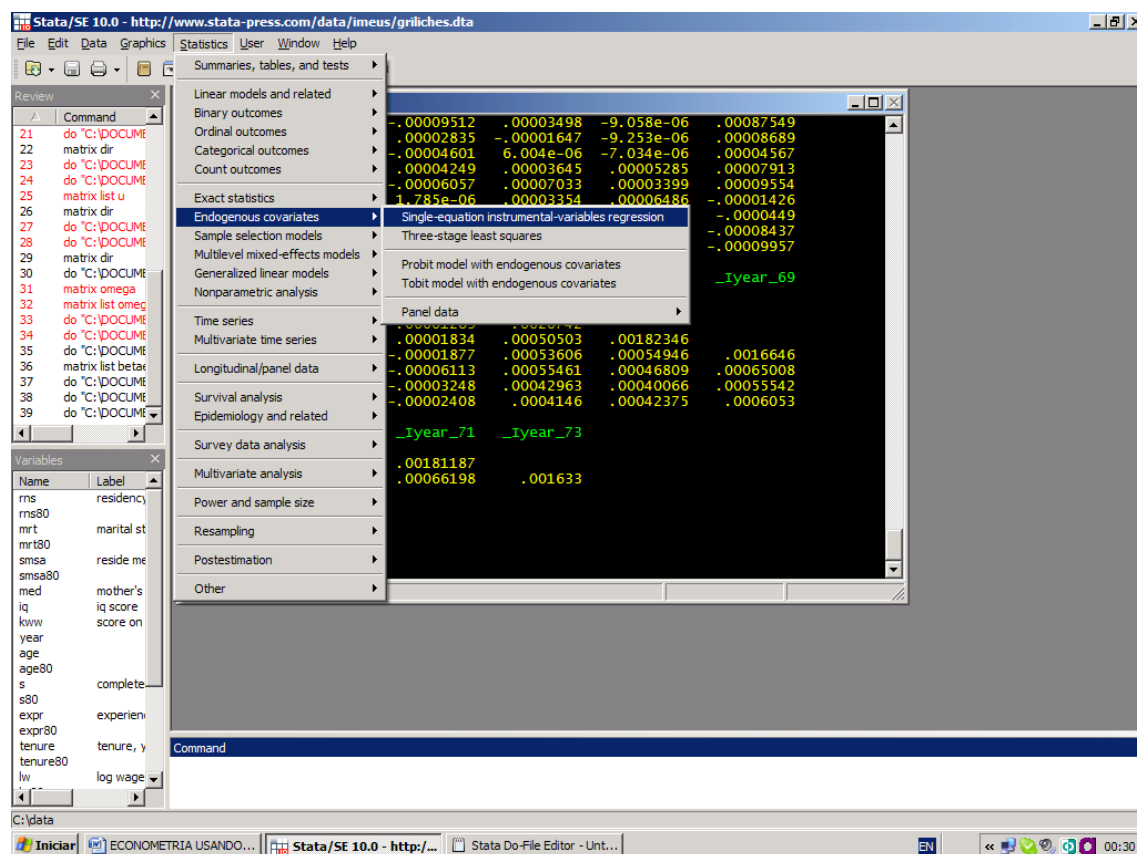
**use <http://www.stata-press.com/data/imeus/griliches>, clear**

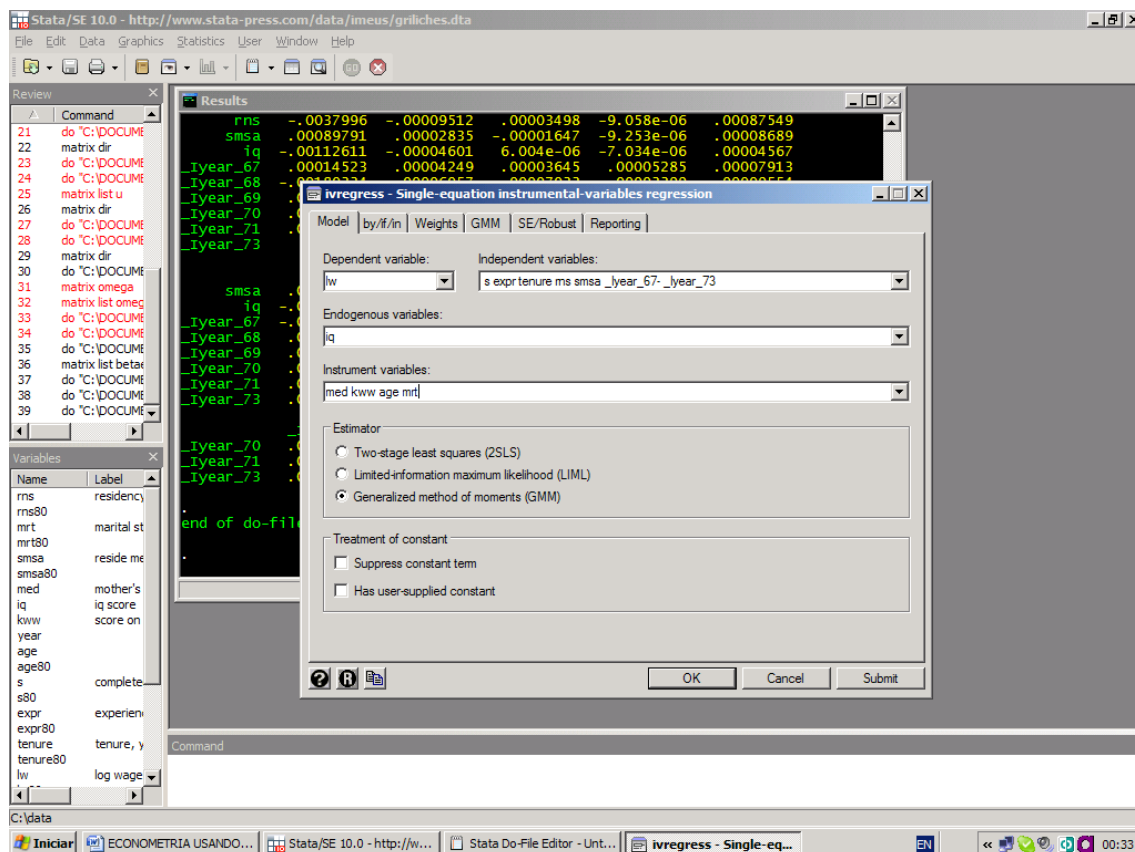
Como vimos em exemplo anterior que para esta estimação IV a matriz Z é formada pelas variáveis: s, expr, tenure, rns, smsa (que são os instrumentos incluídos) e as variáveis med, kww, age, mrt (que são os instrumentos excluídos). A variável endógena é iq e a variável dependente é lw. Vamos repetir os comandos de geração de matrizes.

```
set matsize 800
gen const = 1
mkmat const s expr tenure rns smsa med kww age mrt _lyear_67- _lyear_73,
matrix(Z)
mkmat lw, matrix(y)
mkmat const s expr tenure rns smsa iq _lyear_67- _lyear_73, matrix(X)
matrix PZ = Z*inv(Z'*Z)*Z'
matrix beta2sls = inv(X'*PZ*X)*X'*PZ*y
matrix list beta2sls
matrix u = y-X*beta2sls
matrix omega = J(_N,_N,0)
forvalues i = 1(1)758 {
    matrix omega[`i',`i'] = u[`i',1]*u[`i',1]
}
matrix betaegmm = inv(X'*Z*inv(Z'*omega*Z)*Z'X)*X'*Z*inv(Z'*omega*Z)*Z'*y
```

gmm

O comando `ivreg2` deve ser previamente instalado no STATA, pois não é um comando que vem em seu corpo principal. Para instalar execute o comando `findit ivreg2`. A versão 10 do STATA já possui um comando para realizar estimação GMM com variáveis instrumentais. Para ativar esta opção siga as figuras a seguir:





Vamos recapitular a construção do método de estimação GMM.

A equação a ser estimada, em notação matricial é:

$$y = X\beta + u$$

com uma linha típica:

$$y_i = X_i\beta + u_i$$

A matriz de regressores  $X$  tem dimensão  $n \times K$ , onde  $n$  é o número de observações. Alguns dos regressores são endógenos, de forma que  $E(X_i u_i) \neq 0$ . Fazemos uma partição do conjunto de regressores em  $[X_1 \ X_2]$ , com  $K_1$  regressores  $X_1$  que de acordo com a hipótese nula são endógenos e  $K_2 = (K - K_1)$  regressores  $X_2$  que são considerados exógenos.

Temos então a seguinte equação:

$$y = [X_1 \ X_2] [\beta_1' \ \beta_2']' + u$$



O conjunto de variáveis instrumentais é  $Z$  e tem dimensão  $n \times L$ . Este é o conjunto completo de variáveis que são exógenas -  $E(Z_i u_i) = 0$ . Fazemos uma partição dos instrumentos em  $[Z_1-Z_2]$ , com  $L_1$  instrumentos  $Z_1$  que são instrumentos incluídos e  $L_2=(L- L_1)$  instrumentos  $Z_2=X_2$  que são os instrumentos incluídos / regressores exógenos.

A condição de ordem para identificação da equação é:  $L \geq K$

Isto implica que precisamos ter no mínimo tantos instrumentos excluídos ( $L_1$ ) quantos forem os regressores endógenos ( $K_1$ ).

Se  $L = K$  a equação é exatamente identificada.

Se  $L > K$  a equação é sobre-identificada.

Os  $L$  instrumentos nos dão um conjunto de  $L$  momentos:

$$g_i(\beta) = Z_i' u_i = Z_i' (y_i - X_i \beta) \quad i = 1, n$$

Temos um vetor  $g_i$  que é  $L \times 1$  (resultado de uma multiplicação de uma matriz  $Z_i'$  que é  $L \times n$  por outra matriz que é  $n \times 1$ ). Dado que os  $L$  instrumentos são todos exógenos -  $E(Z_i u_i) = 0$ , temos  $L$  momentos nulos:

$$E(g_i(\beta)) = 0$$

Cada uma das  $L$  equações de momento corresponde a um momento amostral. Para um dado estimador  $\hat{\beta}$ , podemos escrever estes  $L$  momentos amostrais como:

$$\begin{aligned} \bar{g}(\beta) &= \frac{1}{n} \sum_{i=1}^n g_i(\hat{\beta}) = \frac{1}{n} \sum_{i=1}^n Z_i' (y_i - X_i \hat{\beta}) \\ &= \frac{1}{n} Z' \hat{u} \end{aligned}$$

Em notação expandida para matrizes, esta condição corresponde a:

$$\begin{pmatrix} g_1(\beta) \\ g_2(\beta) \\ \dots \\ g_l(\beta) \end{pmatrix} = \frac{1}{n} \begin{pmatrix} z_{11} & z_{21} & \dots & z_{l1} \\ z_{12} & z_{22} & \dots & z_{l2} \\ \dots & \dots & \dots & \dots \\ z_{1l} & z_{2l} & \dots & z_{ll} \end{pmatrix} \begin{pmatrix} y_1 - (\beta_1 x_{11} + \dots + \beta_k x_{1k}) \\ y_2 - (\beta_1 x_{21} + \dots + \beta_k x_{2k}) \\ \dots \\ y_n - (\beta_1 x_{n1} + \dots + \beta_k x_{nk}) \end{pmatrix}$$

O que está por trás da estimação GMM? Temos que escolher um estimador para o vetor de parâmetros  $\beta$  que torne  $\bar{g}(\beta)$  tão próximo de zero quanto possível. No caso de  $L = K$  (equação exatamente identificada) temos  $L$  condições (equações) iguais a  $K$  coeficientes (incógnitas) em  $\hat{\beta}$ . Neste caso, é possível achar uma matriz  $\hat{Q}$  que soluciona o sistema  $\bar{g}(\beta)$

Quando  $L = K$  a equação é exatamente identificada e uma solução única existe equivalente ao estimador padrão de variáveis instrumentais:

$$\hat{\beta}_{IV} = (Z'X)^{-1}Z'y$$

No caso de sobre-identificação ( $L > K$ ), podemos definir um conjunto de  $K$  instrumentos:

$$\hat{X} = Z'(Z'Z)^{-1}Z'X = P_z X$$

que é o estimador de mínimos quadrados em dois estágios (2SLS) que a despeito do seu nome é calculado por esta simples equação matricial. Se a equação é sobre-identificada ( $L \geq K$ ) temos mais equações do que incógnitas e neste caso não é possível achar uma matriz que iguale exatamente todo o conjunto de  $L$  momentos a zero. Neste caso, temos que tomar uma matriz de ponderação  $W$  ( $L \times L$ ) e utilizá-la para construir uma forma quadrática nas condições de momento.

No método 2SLS com sobre-identificação os  $L$  instrumentos disponíveis são reduzidos aos  $K$  necessários para definir a matriz  $P_z$ . De acordo com Baum(2008), na abordagem IV-GMM esta redução não é necessária e todos os  $L$  instrumentos são usados no estimador. Uma matriz de ponderação é empregada de forma que podemos determinar  $\hat{\beta}_{GMM}$  de forma que os elementos de  $\bar{g}(\hat{\beta}_{GMM})$  são tão próximos de zero quanto possível. Com  $L > K$  nem todas as  $L$  condições de momento podem ser satisfeitas e um critério de função que pondere estas condições apropriadamente é utilizada para aumentar a eficiência do estimador.

O estimador GMM minimiza o critério (função objetivo):

$$J(\hat{\beta}_{GMM}) = n\bar{g}(\hat{\beta}_{GMM})'W\bar{g}(\hat{\beta}_{GMM})$$

onde  $W$  é uma matriz de ponderação simétrica  $L \times L$ . Resolvendo através deste critério de minimização obtemos o estimador IV-GMM de uma equação sobre-identificada:

$$\hat{\beta}_{GMM} = (X'ZWZ'X)^{-1}X'ZWZ'y$$

que será idêntico para todas as matrizes  $W$  que diferem por um fator de proporcionalidade.

A consistência é garantida por qualquer matriz de ponderação  $W$  simétrica positiva e, portanto há tantos estimadores GMM como há escolhas da matriz de ponderação  $W$ . Mas a eficiência não é garantida por uma  $W$  arbitrária. Então, o último estimador será referido como estimador GMM possivelmente ineficiente. Estamos interessados em obter estimadores GMM eficientes: estimadores com mínima variância assintótica.

Qual é a escolha ótima da matriz de ponderação  $W$  que minimiza a variância do estimador GMM?

Seja  $S$  a matriz de covariância assintótica das condições de momento  $\bar{g}$  :

$$S = AVar(\bar{g}(\beta)) = \lim_{n \rightarrow \infty} \frac{1}{n} E(Z'uu'Z) =$$

$$\lim_{n \rightarrow \infty} \frac{1}{n} E(Z'\Omega Z)$$

onde  $S$  é uma matriz  $L \times L$  ,  $\bar{g}(\beta) = \frac{1}{n} Z'u$

e  $\Omega$  é a matriz de variância-covariância dos resíduos.

A fórmula geral para a distribuição do estimador GMM é:

$$V(\beta_{GMM}) = \frac{1}{n} (Q'_{XZ} W Q_{XZ})^{-1} (Q'_{XZ} W S W Q_{XZ}) (Q'_{XZ} W Q_{XZ})^{-1}$$

O estimador GMM eficiente é o estimador GMM com uma matriz de ponderação ótima que minimiza a variância assintótica do estimador. Isto é obtido pela escolha de  $W = S^{-1}$  .

Substituindo  $W$  por  $S^{-1}$  na expressão anterior do estimador GMM, temos:

$$\hat{\beta}_{GMM} = (X' Z S^{-1} Z' X)^{-1} X' Z S^{-1} Z' y$$

com variância assintótica:

$$V(\hat{\beta}_{EGMM}) = (Q'_{XZ} S^{-1} Q_{XZ})^{-1}$$

A matriz  $S$  é obtida em um primeiro estágio através da estimativa ineficiente de uma matriz diagonal  $\hat{\Omega}$  que é posteriormente introduzida na expressão:

$$\hat{S} = \frac{1}{n} \sum_{i=1}^n \hat{u}_i^2 Z_i' Z_i = \frac{1}{n} Z' \hat{\Omega} Z$$

onde  $\hat{\Omega}$  é uma matriz diagonal de resíduos ao quadrado  $u_i^2$  de  $\bar{\beta}$ , que é o estimador GMM de primeiro estágio consistente mas não necessariamente eficiente. No comando Stata `ivreg2`, este estimador de primeiro estágio é  $\hat{\beta}_{IV}$ , o estimador de variáveis instrumentais.

### ***Implicações da heterocedasticidade para o estimador IV***

Na presença de heterocedasticidade, o estimador IV é ineficiente mas consistente, enquanto que a matriz padrão estimada de covariância é inconsistente. A vantagem do GMM sobre IV é clara: se a heterocedasticidade está presente, o estimador GMM é mais eficiente que o estimador simples IV, enquanto que se não existe heterocedasticidade o estimador GMM não é pior assintoticamente que o estimador IV. No entanto, o uso do GMM tem um preço. A matriz de ponderação ótima  $\hat{S}$  é uma função dos quartos momentos e a obtenção de uma estimativa razoável para estes requer amostras muito grandes. Se o erro é homocedástico, IV é preferível ao GMM eficiente.

### ***Testes de Heterocedasticidade***

Estatísticas de Breusch-Pagan/Godfrey/Cook-Weisberg e White/Koenker são testes de heterocedasticidade em regressão OLS. Testa-se a relação entre os resíduos da regressão e  $p$  variáveis indicadores que são relacionadas a heterocedasticidade (por hipótese).

A estatística é distribuída como uma  $\chi^2$  com  $p$  graus de liberdade sob a nula de não heterocedasticidade e de que o erro da regressão é normalmente distribuído. O poder deste teste é muito sensível a hipótese de normalidade dos resíduos: Koenker propôs um teste que relaxa esta hipótese. Estes testes estão no Stata após a estimação com o comando `regress`, com `ivhetttest`, `hetttest` e `whitetst`.

Pagan e Hall mostraram que estes testes são válidos na regressão IV somente se a heterocedasticidade naquela equação e em nenhuma outra mais no sistema. As outras equações estruturais no sistema (correspondentes aos

regressores endógenos  $X_1$ ) precisam ser homocedásticas mesmo que elas não sejam explicitamente estimadas. Este teste está disponível no Stata através do comando `ivhettest` após a estimação com `ivreg`, `ivreg2` ou `ivgmm0`.

### ***Testando a relevância e validade dos instrumentos***

Como vimos, as variáveis instrumentais tem que satisfazer duas condições: precisam ser correlacionadas com os regressores endógenos e devem ser ortogonais ao processo de erro. A primeira condição pode ser testada examinando o grau de ajuste das regressões de primeiro estágio, ou o que é o mesmo, verificar o poder explicativo dos instrumentos excluídos nestas regressões. A estatística comumente usada é o  $R^2$  da regressão de primeiro estágio: a correlação parcial ao quadrado entre os instrumentos excluídos  $Z_1$  e o regressor endógeno (Bound).

Um exemplo: o pesquisador tem um modelo com dois regressores endógenos e dois instrumentos excluídos. Um dos instrumentos excluídos é altamente correlacionado com os dois regressores endógenos mas o outro instrumento excluído tem uma correlação nula (representa um processo de ruído).

O modelo está, portanto, sub-identificado: há um instrumento bom, mas dois regressores endógenos. Mas a estatística  $F$  e o  $R^2$  não revelam esta fraqueza. A solução é encontrar mais instrumentos relevantes ou eliminar o regressor endógeno da equação. A estatística de Bound só é válida quando temos apenas um regressor endógeno.

Para levar em conta diversos regressores endógenos Shea(1997) propôs “uma medida de  $R^2$  parcial que leva em conta as inter-correlações entre os instrumentos”. Para um modelo contendo um único regressor endógeno, as duas medidas de  $R^2$  são equivalentes. Se uma equação gera um grande valor do  $R^2$  parcial (Bound) e pequeno valor da medida de Shea, podemos concluir que os instrumentos tem pouca relevância para explicar os regressores endógenos e o modelo pode estar sub-especificado.

## ***Consequências de instrumentos fracos***

Vamos enumerar as principais consequências de utilização de instrumentos fracos em estimação IV:

*Aumento do viés dos coeficientes IV estimados.*

*O modelo não fica identificado com relação as variáveis endógenas.*

*Neste caso, o viés do estimador IV é o mesmo do estimador OLS – a estimação IV é inconsistente e nada se ganha com isto.*

*Para equação com um único regressor endógeno uma estatística F com valor menor do que 10 significa que os instrumentos são fracos.*

*Deve-se ser parcimonioso na escolha dos instrumentos, dado que o viés por IV é crescente com o numero de instrumentos.*

*O problema de instrumentos fracos pode aparecer mesmo quando os testes de primeiro estágio são significativos aos níveis de 5 e 1 % e se dispõe de uma amostra grande.*

## ***Testando a endogeneidade de uma variável explicativa***

Suponha a seguinte equação de regressão:

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + \beta_3 z_2 + u_1$$

onde  $y_2$  é a variável que suspeita-se que seja endógena e  $z_1$  e  $z_2$  são exógenas.

Temos a equação de  $y_2$  na forma reduzida:

$$y_2 = \pi_0 + \pi_1 z_1 + \pi_2 z_2 + \pi_3 z_3 + \pi_4 z_4 + v_2$$

Como as variáveis  $z$  são não correlacionadas com  $u_1$ ,  $y_2$  será não correlacionado com  $u_1$  se, e somente se  $v_2$  for não correlacionada com  $u_1$ .

Existem duas maneiras de testar isto:

1) Regredir  $u_1$  contra em  $v_2$  um modelo  $u_1 = \delta_1 v_2 + e_1$

onde  $e_1$  é não correlacionado com  $v_2$  e tem média 0. Então  $u_1$  e  $v_2$  serão não correlacionados se, e somente se  $\delta_1 = 0$

2) Incluir  $v_2$  como um regressor adicional na primeira equação e fazer um teste  $t$  para  $\delta_1$ :

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + \beta_3 z_2 + \delta_1 v_2 + u_1$$

Se a estimativa  $\delta_1$  for significativa (através de um teste  $t$ ) concluímos que  $y_2$  é endógena na equação (XX).

Podemos também testar a endogeneidade de múltiplas variáveis explicativas. Para cada variável suspeita de ser endógena obtemos os resíduos da equação da forma reduzida e verificamos a significância conjunta da forma estrutural usando um teste  $F$ . Se rejeitarmos a nula concluímos que pelo menos uma das variáveis explicativas é endógena (Wooldridge pg. 477).

#### \* TESTE DE ENDOGENEIDADE DE UMA UNICA VARIÁVEL EXPLICATIVA

use "c:\textos download\wooldridge data files\mroz.dta", clear

regress educ exper expersq motheduc fatheduc if hours > 0

test motheduc fatheduc

predict v2,residuals

regress lwage educ exper expersq v2

regress lwage educ exper expersq

ivregress 2sls lwage exper expersq (educ = motheduc fatheduc)

*Testes realizados através do comando ivreg2: Teste Hansen-Sargan*

- Teste de restrições de sobreidentificação.
- A hipótese nula conjunta é que os instrumentos são instrumentos válidos, isto é, não correlacionados com o termo de erro e que os instrumentos excluídos são corretamente excluídos da equação estimada.

- Sob a nula, a estatística de teste é distribuída como qui-quadrado no número de restrições de sobre-identificação.
- Uma rejeição coloca em dúvida a validade dos instrumentos.
- Para o estimador eficiente GMM, a estatística de teste é a estatística J de Hansen, que é o valor minimizado da função objetivo GMM.
- Para os estimador 2SLS, a estatística de teste é a estatística de Sargan, calculada como  $N \cdot R^2$  de uma regressão dos resíduos de IV sobre o conjunto completo de instrumentos.

#### *Testes realizados através do comando ivreg2: Estatística C*

- A estatística C, ou estatística “diferença-em-Sargan” é obtida através da opção orthog do comando ivreg2.
- Permite o teste de um subconjunto de condições de ortogonalidade, ou seja, é o teste de exogeneidade de um ou mais instrumentos.
- É definida como a diferença da estatística Hansen-Sargan da equação com o conjunto menor de instrumentos e a equação com o conjunto completo de instrumentos (incluindo os instrumentos suspeitos).
- Sob a nula de que todos os instrumentos são válidos a estatística C tem distribuição qui-quadrado no número de instrumentos testados.
- A falha em rejeitar a nula significa que o conjunto total de condições de ortogonalidade é válido.

#### *Testes realizados através do comando ivreg2: Teste de razão de verossimilhança de correlação canônica de Anderson*

- Testa se a equação é identificada, ou seja, se os instrumentos excluídos são válidos.
- A hipótese é nula é que a equação é sub-especificada.
- Sob a nula de sub-identificação, a estatística é distribuída como qui-quadrado com  $L - K + 1$  graus de liberdade ( $L$ = número de instrumentos excluídos e incluídos).
- A estatística fornece uma medida da relevância dos instrumentos e a rejeição da nula indica que o modelo é identificado.
- Importante: uma rejeição da nula deve ser interpretada com cautela, já que problemas de instrumentos fracos podem ainda estar presentes.



## 17. Simulação

Antes de começarmos a discutir a econometria propriamente dita, vamos introduzir um tema que é muito útil para as aplicações nesta área. A simulação de variáveis aleatórias facilita bastante o entendimento de processos estatísticos e econométricos. Começemos por uma variável aleatória uniforme. Na verdade a geração de números aleatórios nos computadores é realizada de uma forma determinística e por esta razão é mais adequado chamar de geração de números pseudo-aleatórios. Desta forma, instrumentos matemáticos determinísticos podem simular um processo aleatório, como por exemplo, uma distribuição uniforme que varia entre 0 e 1. O comando Stata `runiform()` é um gerador de números pseudo-aleatórios que seguem uma distribuição uniforme padrão, com parâmetros 0 e 1. De acordo com a seqüência de comandos abaixo (do-file), inicialmente definimos uma semente aleatória para que possamos reproduzir o número pseudo-aleatório em qualquer computador. O comando `scalar` gera o número pseudo-aleatório e o armazena em uma variável de memória escalar denominada `u`. Finalmente o comando `display` faz com que este número seja mostrado na janela de resultados do Stata.

```
* seleção (simulação) de uma variável aleatória uniforme
set seed 99999
scalar u = runiform()
display u
```

Vamos agora simular a seleção de 1000 números aleatórios extraídos aleatoriamente de uma distribuição uniforme com parâmetros 0 e 1.

```
* 10000 seleções aleatórias de uma distribuição uniforme
set obs 10000
set seed 99999
gen x = runiform()
edit x
summ x
```

Vemos que as 10000 seleções tem uma média .4959344 e um desvio-padrão de .2893517, bem próximos dos valores teóricos 0.5 e  $\sqrt{1/12} = .28867513$ .

A seleção aleatória de uma distribuição normal padrão pode ser feita utilizando-se a função `rnormal()`

```

* simulação de uma seleção aleatória a partir de distribuição normal
clear
set obs 10000
set seed 99999
gen uniform = runiform()
gen stnormal = rnormal()
gen normal10e4 = rnormal(10,4)
tabstat uniform stnormal normal10e4, stat(mean sd sk kurt min max) col(stat)

```

Os resultados da execução deste ultimo do-file mostram que a distribuição normal padrão tem média e desvio-padrão próximos aos valores teóricos 0 e 1, a distribuição normal tem média e desvio-padrão próximos a 10 e 4, o coeficiente de assimetria e o de curtose são próximos aos valores teóricos de 0 e 3, respectivamente, para estas duas distribuições.

### ***Aplicações de simulação para o Teorema do Limite Central***

Vamos inicialmente definir um programa com diversos comandos. Este programa é uma rotina que o Stata executará toda vez que aparecer um comando com o seu nome. Mas aqui utilizaremos o comando simulate que é um recurso que o Stata tem para poder realizar um programa um grande número de vezes. O programa que definiremos irá simular a seleção aleatória de uma amostra de tamanho  $n = 50$  de uma população com distribuição uniforme.

#### **SEQUENCIA DE COMANDOS PARA SIMULAÇÃO SOBRE O TEOREMA DO LIMITE CENTRAL PARA UMA POPULAÇÃO SIMULADA DE UMA DISTRIBUIÇÃO UNIFORME**

```

* Programa para selecionar uma amostra de tamanho 50 a partir de uma
* distribuição uniforme e calcular a média
program umaamostra, rclass
drop _all
set obs 50
gen x = runiform()
summa x
return scalar mediaparaumaamostra = r(mean)
end

* são gerados 100000 valores para a variável aleatória uniforme

set obs 100000
gen x = runiform()
histogram x, xtitle("valores selecionados aleatoriamente de uma
distribuição uniforme")

* executa o programa umaamostra 100000 vezes para obter 100000 medias
simulate xbar = r(mediaparaumaamostra), seed(99999) reps(10000)
nodots:umaamostra

```

```
summa xbar
histogram xbar, normal xtitle("média amostral para muitas amostras")
```

A primeira parte deste ultimo do-file é o programa que seleciona um número aleatório de uma distribuição uniforme. Este programa vai do comando program até o comando end. O primeiro comando (drop \_all) limpa a memória de dados do Stata. O segundo comando (set obs 50) inicializa a memória de dados com 50 observações vazias. O quarto comando (gen x = runiform() ) gera a variável aleatória uniforme para as 50 observações. O quinto comando do programa (summa x) calcula estatísticas descritivas. O sexto comando (return scalar mediaparaumaamostra = r(mean)) coloca o valor da média da amostra de 50 observações em uma variável escalar de nome r(mediaparaumaamostra) que é armazenada na memória volátil do Stata. Finalmente o ultimo comando (end) demarca o fim da rotina de nome umaamostra.

Em seguida são gerados 100000 valores para a variável aleatória uniforme. Este procedimento é interessante para ilustrar o Teorema do Limite Central cujo enunciado é:

“Suponhamos uma variável aleatória  $X$  definida em uma população (finita ou infinita). Se selecionarmos um número infinito de amostras de mesmo tamanho desta população e para cada uma destas amostras calcularmos a média da amostra  $\bar{X}$ , a distribuição destas médias amostrais terá distribuição normal mesmo que a distribuição de  $X$  na população não seja normal, quando o tamanho da amostra  $n$  tende ao infinito”.

Vamos finalizar estes exemplos de simulação para o Teorema do Limite Central com um exemplo simples retirado dos arquivos de sistema do Stata: o arquivo de automóveis (auto.dta) com uma população finita de 74 marcas de automóveis.

## SEQUENCIA DE COMANDOS PARA SIMULAÇÃO SOBRE O TEOREMA DO LIMITE CENTRAL PARA UMA POPULAÇÃO FINITA DE AUTOMOVEIS

```
clear
sysuse auto
histogram price, normal
summa price
scalar sigma2 = r(sd)^2*r(N)/(r(N)-1)
scalar mu      = r(mean)
```

```
* programa para selecionar uma amostra de tamanho 30 da população de
automoveis
```

```
program drop _all
program umaamostra, rclas
sysuse auto, clear
sample 30, count
```

```

summa price
return scalar mediaparaumaamostra = r(mean)
end

simulate xbar = r(mediaparaumaamostra), seed(99999) reps(10000)
nodots:umaamostra

summa xbar
scalar mu_xtraco = r(mean)
scalar sigma2_xtraco = r(sd)^2*(r(N)-1)/r(N)
scalar sigma2_xtraco_teor = (sigma2/30)*(74-30)/(74-1)

disp mu
disp mu_xtraco
disp sigma2_xtraco_teor
disp sigma2_xtraco

histogram xbar, normal xtitle("média amostral para muitas amostras")

```

No exemplo do do-file acima, estamos simulando a seleção aleatória de amostras sem reposição de tamanho  $n = 30$  a partir de uma população finita de tamanho  $N = 74$ . Calculamos a media da amostra para cada uma das 10000 replicações e posteriormente calculamos a media das medias amostrais, calculamos o desvio padrão das medias amostrais e comparamos com os seus respectivos valores teóricos.

É importante observar que apos simularmos as 10000 replicações calculamos: 1) a media das medias amostrais ( $\mu_{xtraco}$ ) e comparamos com a media da população ( $\mu$ ) e 2) a variância das medias amostrais calculada a partir das 10000 medias amostrais ( $\sigma^2_{xtraco}$ ) e a variância teórica ( $\sigma^2_{xtraco\_teor}$ ). Dois detalhes nestes cálculos: 1) o calculo da variancia das medias amostrais é feito através do fator  $(r(N) - 1)/r(N)$  para transformar a formula da variancia de uma amostra (obtida pelo commando summarize) na formula da variancia populacional e 2) O cálculo da variância teórica das medias amostrais é feito com correção de população finita, através do fator  $(N-n)/(N-1) = (74-30)/(74-1)$ .

### ***Uma curiosidade matemática: integração de Monte Carlo***

Vamos considerar o cálculo de uma integral definida muito difícil de ser realizado de forma analítica. Por exemplo, queremos calcular:

$$E[\exp[-\exp(Y)]] = \int_{-\infty}^{\infty} \exp[-\exp(Y)] f(Y) dy$$

E vamos supor que a variável aleatória Y tem distribuição normal padrão, ou seja.  $Y \sim N(0,1)$ . Então:

$$E[\exp[-\exp(Y)]] = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp[-\exp(Y)] \exp(-y^2 / 2) dy$$

Desta forma, estamos calculando a esperança matemática de uma função matemática da variável aleatória Y

\* Calculo de uma integral por simulação de Monte Carlo com S = 10000

```
clear all
qui set obs 10000000
set seed 99999
gen double y = invnormal(runiform())
gen double gy = exp(-exp(y))
qui summ gy, meanonly
scalar Egy = r(mean)
display "Apos 10000 simulações de Monte Carlo a estimativa de
E[exp(exp(-x))] é " Egy
```

O primeiro commando limpa a memória. O Segundo comando gera 10 milhões de observações (!! ) vazias na area de dados do Stata. O terceiro commando define a semente aleatoria (para poder reproduzir os mesmos resultados em computadores distintos), O quarto comando gera uma variável aleatória normal padrão. Ele aplica a função inversa da normal padrão a uma variável aleatória uniforme. Assim, para cada valor aleatório entre 0 e 1 que passa a ser o valor da função de distribuição cumulativa F(x) de uma normal padrão, o comando calcula o valor da função inversa da normal padrão cumulativa. O quinto comando calcula a parte da função a ser integrada ( $\exp(-\exp(y))$ ). O sexto comando calcula a media desta variável. O sétimo comando armazena o valor desta media em uma variável de memória scalar e finalmente o ultimo comando é uma instrução para mostrar na janela de resultados este ultimo valor. Conclusão: o valor esperado desta função de variavel aleatoria é aproximadamente igual a .38174164

Um outro procedimento para calcular o valor desta integral definida, de forma aproximada é através do do file abaixo que utilize o comando Stata integ de integração numérica. Mas este procedimento envolve uma elevada disponibilidade de memória e muito tempo computacional.

```
set memory 700m
clear
range x -2000000 2000000 10000000
gen y = (1/sqrt(2*acos(-1))) * exp(-exp(x)) * exp(-x^2/2)
summa y
integ y x
```

No do-file anterior o primeiro comando estabelece a memória para 700 megabytes, o segundo comando limpa a memória. O terceiro comando cria uma variável x com intervalos iguais que se inicia no primeiro argument (-2000000), termina no segundo argument (2000000) com 10 milhões de valores. O quarto comando calcula o valor do integrando. O quinto comando calcula as estatísticas sumárias para y e finalmente, o ultimo comando integra y em relação a x. O resultado da integral numérica definida é mostrado no final da execução. Observe que ele se aproxima bastante do resultado do procedimento anterior. Podemos ter maior segurança de que estamos realizando a intergração numérica em um intervalo válido utilizando a seguinte sequência de comandos:

```
clear
set memory 700m
range x -2000000 2000000 10000000
gen y = (1/sqrt(2*acos(-1))) * exp(-exp(x)) * exp(-x^2/2)
summa y
integ y x

keep if y > 0 & y != .
twoway (line y x)
clear
range x -15 5 10000000
gen y = (1/sqrt(2*acos(-1))) * exp(-exp(x)) * exp(-x^2/2)
summa y
integ y x
```

## **Referencias bibliográficas**

Baum, C. F., M. E. Schaffer, and S. Stillman. 2003. Instrumental variables and GMM: Estimation and testing. *Stata Journal* 3: 1–31.

Baum, C. F. 2006. *An Introduction to Modern Econometrics Using Stata*. College Station, TX: Stata Press.

Baum, C. F. Schaffer M.E. e Stillman, S. 2006. Enhanced routines for instrumental variables/GMM estimation and testing,

Cameron, A.C. e Trivedi, P.K. 2009. *Microeconometrics using Stata*. Stata Press Publication, StataCorp LP, College Station, Texas.

Wooldridge, J. M. 2003. *Introductory Econometrics: A Modern Approach*. 2nd ed. New York: Thomson Learning.