AULA 5
Análise de Dados Legislativos e Eleitorais Utilizando o Programa Stata

Professor: Ernesto Friedrich de Lima Amaral
Email: eflamaral@gmail.com
Site do curso: www.ernestoamaral.com/stata20091.html

Data: 26/05/2009
Horário: 18:00–21:00

Set mem, set matsize, set maxvar

Aprendemos a usar o "set mem" para informar o quanto de memória RAM deve ser disponibilizada pelo computador para que o Stata possa trabalhar:

```
set mem 100m, perm
```

Há ainda o comando "set matsize" que indica ao Stata o número máximo de variáveis que podem ser incluídas nos comandos do Stata. Por exemplo, esse número indica a quantidade máxima de variáveis que podem ser usadas em uma regressão.

O tamanho padrão no Stata/MP e Stata/SE é de 400 variáveis, podendo ser mudado para baixo ou para cima, com limite máximo de 11.000 variáveis. No Stata/IC, o valor inicial é de 200, com limite máximo de 800.

Por exemplo, podemos mudar o número máximo de variáveis nos comandos de estimação para 500:

```
set matsize 500
```

ou

```
set matsize 500, permanently
```

Além disso, o Stata/MP e Stata/SE permitem mudar o número máximo de variáveis no banco de dados com o comando "set maxvar". Isso não é permitido no Stata/IC.

```
set maxvar 5000
```

Vamos rodar os comandos das aulas 3 e 4

Com os comandos abaixo, criaremos novamente as variáveis que utilizamos e utilizaremos nos modelos de regressão:

```
clear
set mem 100m, perm
set matsize 500, perm

log using "C:\cursodcp\log\aula5.log", text replace

*Criando variável para indicar deputado mais velho em cada partido
use "C:\cursodcp\dados\Argentina51.dta", clear
replace p63=. if p63==99
gsort partido -p63
egen depvel=tag(partido)
sort partido
save "C:\cursodcp\dados\argoriginal.dta", replace
```

```
*Criando banco de dados que possui somente dados dos deputados mais velhos
keep if depvel==1
keep partido p34 p35 p36
rename p34 p34vel
rename p35 p35vel
rename p36 p36vel
sort partido
save "C:\cursodcp\dados\argvelho.dta", replace

*Alocando os dados dos deputados mais velhos para os demais deputados
use "C:\cursodcp\dados\argoriginal.dta", clear
merge partido using "C:\cursodcp\dados\argvelho.dta"
tab _merge
drop _merge
replace p34vel=. if depvel==1
replace p35vel=. if depvel==1
replace p36vel=. if depvel==1
save "C:\cursodcp\dados\argalocado.dta", replace

*Recodificando opinião dos deputados sobre imposto
tab p34, missing nolabel
gen imposto=p34
replace imposto=. if p34==8 | p34==9
tab imposto, missing nolabel

*Recodificando opinião dos deputados mais velhos sobre imposto
tab p34vel, missing nolabel
gen impostovel=p34vel
replace impostovel=. if p34vel==8 | p34vel==9
tab impostovel, missing nolabel

*Recodificando opinião dos deputados sobre privatização
tab p35, missing nolabel
gen privatizar=p35
replace privatizar=. if p35==8 | p35==9
tab privatizar, missing nolabel

*Recodificando opinião dos deputados mais velhos sobre privatização
tab p35vel, missing nolabel
gen privatizarvel=p35vel
replace privatizarvel=. if p35vel==8 | p35vel==9
tab privatizarvel, missing nolabel

*Recodificando opinião dos deputados sobre Mercosul
tab p36, missing nolabel
gen mercosul=p36
replace mercosul=. if p36==8 | p36==9
tab mercosul, missing nolabel

*Recodificando opinião dos deputados mais velhos sobre Mercosul
tab p36vel, missing nolabel
gen mercosulvel=p36vel
replace mercosulvel=. if p36vel==8 | p36vel==9
tab mercosulvel, missing nolabel

*Recodificando escala esquerda/direita do deputado
tab p58, missing nolabel
gen esqdir=p58
replace esqdir=. if p58==99
tab esqdir, missing nolabel
```

```
*Criando variáveis dicotômicas de privatizar (processo manual)
gen priv1=.
  replace priv1=0 if privatizar!=1 & privatizar!=.
  replace priv1=1 if privatizar==1
gen priv2=.
  replace priv2=0 if privatizar!=2 & privatizar!=.
  replace priv2=1 if privatizar==2
gen priv3=.
  replace priv3=0 if privatizar!=3 & privatizar!=.
  replace priv3=1 if privatizar==3
gen priv4=.
  replace priv4=0 if privatizar!=4 & privatizar!=.
  replace priv4=1 if privatizar==4
gen priv5=.
  replace priv5=0 if privatizar!=5 & privatizar!=.
  replace priv5=1 if privatizar==5

*Criando variáveis dicotômicas de privatizar (uso do tabulate)
tab privatizar, gen(privatizar)

*Recodificando variável imposto para 0/1
tab imposto, missing nolabel
gen impostod=.
replace impostod=0 if imposto==1
replace impostod=1 if imposto==2
tab imposto impostod

*Criando interação entre imposto e privatizar1
gen impprev1=impostod*privatizar1

*Recodificando escala esquerda/direita do partido
tab p59, missing nolabel
gen esqdirp=p59
replace esqdirp=. if p59==99
tab esqdirp, missing nolabel
```

Utilizando peso nas regressões

Primeiramente, vamos excluir as observações que possuem "missing" em nossas variáveis de interesse ("esqdir", "impostod" e "privatizar"):

```
keep if esqdir!=. & impostod!=. & privatizar!=.
display _N
```

É bom entender a variável "peso" de nosso banco de dados. Nesse caso, o número de observações (*Obs*) é o mesmo que a soma dos pesos (*Sum of Wgt.*):

```
sum peso, detail
```

O "Obs" e "Sum of Wgt" são iguais porque esse peso foi construído de forma que alguns pesos possuem valor acima de uma unidade, e outros pesos têm valor abaixo de uma unidade:

```
tab peso, missing
```

A ajuda no Stata mostra as diferentes opções de uso do peso:

```
help weight
```

Há o peso "iweight" (peso importância) que não tem uma explicação estatística formal. Esse peso é utilizado por programadores que precisam implementar técnicas analíticas próprias:

```
char privatizar[omit] 3
xi: regress esqdir impostod i.privatizar [iweight=peso]
```

A regressão pode utilizar o peso "aweight" (peso analítico) que é inversamente proporcional à variância da observação. Com esse peso, o número de observações na regressão é automaticamente escalonado para permanecer o mesmo que o número de observações no banco. Esse peso é utilizado para estimar uma regressão linear quando os dados são médias observadas:

– Temos um banco da seguinte forma:

```
group    x      y    n
  1     3.5   26.0   2
  2     5.0   20.0   3
```

– Ao invés de:

```
group   x      y
  1     3     22
  1     4     30
  2     8     25
  2     2     19
  2     5     16
```

– A regressão seria estimada assim:

```
regress y x [aweight=n]
```

– Em nosso exemplo, o uso de "aweight" seria feito da seguinte forma:

```
xi: regress esqdir impostod i.privatizar [aweight=peso]
```

De uma forma geral, não é correto utilizar o "aweight" como um peso amostral, porque as fórmulas utilizadas por esse comando assumem que pesos maiores se referem a observações medidas de forma mais acurada. Uma observação em uma amostra não é medida de forma mais cuidadosa que nenhuma outra observação, já que todas fazem parte do mesmo plano amostral.

Usar o "aweight" para especificar pesos amostrais fará com que o Stata estime valores incorretos de variância e de erros padrões para os coeficientes, assim como valores incorretos de "p" para os testes de hipótese.

O ideal em nosso caso é utilizar o peso "pweight", o qual usa o peso amostral como o número de observações na população que cada observação representa. Com isso, são estimadas proporções, médias e parâmetros da regressão corretamente. Há o uso de uma técnica de estimação robusta da variância que automaticamente ajusta para as características do plano amostral, de tal forma que variâncias, erros padrões e intervalos de confiança são calculados de forma mais precisa.

O uso do peso "pweight" (peso amostral) que é o inverso da probabilidade da observação ser incluída no banco, devido ao desenho amostral, é realizado da seguinte forma:

```
xi: regress esqdir impostod i.privatizar [pweight=peso]
```

Utilizando peso nas tabelas

Os pesos também podem ser utilizados na elaboração de tabelas. Com o "aweight" e "iweight", o total na tabela é o mesmo número de observações no banco:

```
tab esqdir [aweight=peso]
tab esqdir [iweight=peso]
```

O "fweight" (peso freqüência) expande os resultados da amostra para o tamanho populacional. O uso desse peso é importante na amostra do Censo Demográfico e na Pesquisa Nacional por Amostra de Domicílios (PNAD) do Instituto Brasileiro de Geografia e Estatística (IBGE) para expandir a amostra para o tamanho da população do país, por exemplo.

No caso do banco sobre os legisladores, não há a expansão da amostra para um universo de legisladores. Além disso, o "fweight" só funciona quando não há decimais no peso.

Pra utilizar o "fweight" nesse banco, poderíamos arredondar o peso (opção "round"), mas isso tornaria todos os pesos iguais a uma unidade:

```
gen peso2=round(peso)
tab peso2
```

Em decorrência disto, é feito o exercício abaixo, em que a tabela mostra o total populacional 10 milhões de vezes maior que o número de observações no banco:

```
gen peso3=peso*10000000
tab esqdir [fweight=peso3]
```

De qualquer forma, o correto aqui é estimar as tabelas de freqüência com o "aweight", ao invés do "fweight".

Regressão logística

Primeiramente, vamos transformar nossa escala de esquerda/direita (esqdir) em variáveis dicotômicas (esquerda e direita):

```
*Criando variável dicotômica de deputado de esquerda
gen esquerda=.
replace esquerda=1 if esqdir>=1 & esqdir<=4
replace esquerda=0 if esqdir>=5 & esqdir<=9

*Criando variável dicotômica de deputado de direita
gen direita=.
replace direita=1 if esqdir>=5 & esqdir<=9
replace direita=0 if esqdir>=1 & esqdir<=4

*Tabelas das novas variáveis
tab esqdir esquerda, missing
tab esqdir direita, missing
```

Vamos retomar a regressão de mínimos quadrados ordinários:

```
*Regressão OLS com esqdir como variável dependente
xi: regress esqdir impostod i.privatizar
xi: regress esqdir impostod i.privatizar [pweight=peso]
```

Agora vamos rodar regressão logística, utilizando "esquerda" como variável dependente. Podemos usar o comando "logit" para mostrar os coeficientes, ou o comando "logistic" para mostrar as razões de chances (*odds ratios*):

```
*Regressão logística com esquerda como variável dependente
*"Privatizaria todos os serviços públicos" (privatizar=1) como referência
char privatizar[omit]1
xi: logit esquerda impostod i.privatizar [pweight=peso]
xi: logistic esquerda impostod i.privatizar [pweight=peso]
```

Veja que a razão de chances de "impostod" estimado por "logistic" (0,0695398) é igual ao exponencial do coeficiente estimado por "logit" (–2,665855):

```
di exp(-2.665855)
```

Vamos rodar a mesma regressão logística, utilizando "direita" como variável dependente:

```
*Regressão logística com direita como variável dependente
*"Não privatizaria nenhum serviço público" (privatizar=5) como referência
char privatizar[omit]5
xi: logit direita impostod i.privatizar [pweight=peso]
xi: logistic direita impostod i.privatizar [pweight=peso]
```

Agora vamos usar o comando "glm" para gerar os mesmos coeficientes do comando "logit":

```
*Usando comando GLM para estimar regressão logística (coeficientes)
char privatizar[omit]1
xi: glm esquerda impostod i.privatizar [pweight=peso], family(binomial)
link(logit)
xi: logit esquerda impostod i.privatizar [pweight=peso]
```

O comando "glm" também pode gerar as mesmas razões de chance do comando "logistic":

```
*Usando comando GLM para estimar regressão logística (razões de chances)
char privatizar[omit]1
xi: glm esquerda impostod i.privatizar [pweight=peso], family(binomial)
link(logit) eform
xi: logistic esquerda impostod i.privatizar [pweight=peso]
```

Testes de diferença entre modelos

Para fazer testes de diferença entre modelos, eu copiei o arquivo que está gravado nesse site (http://www.stata.com/quest/quest/t/tablesq.ado) para o diretório C:\cursodcp\pacotes. Depois de realizado o download, utilize o seguinte comando para instalar esse programa no Stata:

```
do "C:\cursodcp\pacotes\tablesq.ado"
```

Agora podemos realizar testes com o comando "tablesq" para avaliar se há diferença estatisticamente significativa entre os modelos. Podemos estimar um modelo que tem somente "impostod" e depois outro modelo que insere as variáveis dicotômicas de "privatizar":

```
*Modelo 1
xi: logistic esquerda impostod [pweight=peso]

*Modelo 2
xi: logistic esquerda impostod i.privatizar [pweight=peso]
```

Com base nos graus de liberdade (*degrees of freedom*) e nos logaritmos de verossimilhança (*log-likelihood*), reportados pelos modelos, podemos fazer o teste de significância de qui-quadrado. Esse teste indicará se a diferença entre os modelos é estatisticamente significante.

No nosso caso, se o teste for significante, podemos argumentar que nosso modelo é mais robusto com a inclusão das variáveis dicotômicas de "privatizar":

```
*Modelo 1
*Degrees of freedom=1
*Log-likelihood=-41.346034

*Modelo 2
*Degrees of freedom=5
*Log-likelihood=-37.912895

*Diferença entre DF (modelo 2 - modelo 1):
di 5-1

*Diferença entre logs (modelo 2 - modelo 1):
di (-37.912895)-(-41.346034)

*Teste de significância de qui-quadrado:
tablesq X 4 3.433139
```

Nesse caso, o teste "tablesq" não foi significativo. Ou seja, o teste está dizendo que não há melhora estatisticamente significante em nosso modelo ao acrescentar as variáveis dicotômicas de "privatizar".

Também podemos fazer teste de diferença entre modelos usando o comando "fitstat". Podemos fazer o download no seguinte site (http://ideas.repec.org/c/boc/bocode/s407201.html). Eu gravei o arquivo ".ado" no diretório C:\cursodcp\pacotes. Utilize o seguinte comando para instalar esse programa no Stata:

```
do "C:\cursodcp\pacotes\fitstat.ado"
```

Para utilizar o comando "fitstat", primeiro rodamos o modelo 1 e salvamos os resultados da regressão:

```
*Modelo 1
xi: logistic esquerda impostod [pweight=peso]

*Salvando resultados do modelo 1
fitstat, saving(1)

*Também podemos salvar os resultados do modelo 1 da seguinte forma
quietly fitstat, saving(1)
```

Em seguida, rodamos o modelo 2 e o comparamos com os resultados do modelo 1:

```
*Modelo 2
xi: logistic esquerda impostod i.privatizar [pweight=peso]

*Comparando modelo 1 com modelo 2
fitstat, using(1)
```

A informação final do resultado do comando acima coloca que a diferença de 10,07 no teste BIC fornece forte amparo para o modelo salvo (*Difference of 10.070 in BIC' provides very strong support for saved model*). Nosso modelo salvo é justamente o modelo 1. Mais uma vez, o teste indica que a inclusão das variáveis dicotômicas de "privatizar" NÃO OCASIONOU um acréscimo explicativo para nossa variável dependente (esquerda).

Se o mesmo teste for feito para as regressões de mínimos quadrados ordinários, em que nossa variável dependente é a escala esquerda/direita (esqdir), e não a variável dependente dicotômica (esquerda) temos:

```
xi: regress esqdir impostod [pweight=peso]
quietly fitstat, saving(1)
xi: regress esqdir impostod i.privatizar [pweight=peso]
fitstat, using(1)
```

A informação final do resultado do comando acima coloca que a diferença de 0,743 no teste BIC fornece fraco amparo para o modelo salvo (*Difference of 0.743 in BIC' provides weak support for saved model*). Nosso modelo salvo é o modelo 1. Nesse caso, o teste indica que a inclusão das variáveis dicotômicas de "privatizar" OCASIONOU um acréscimo explicativo para nossa variável dependente (esqdir).

Ainda não sabemos qual o melhor modelo. Vamos explorar um pouco mais nossas possibilidades...

Vamos agora estimar um modelo 1 que tem somente as variáveis dicotômicas de "privatizar", um modelo 2 que tem somente a variável "impostod", e um modelo 3 com as variáveis "privatizar" e "impostod". Vamos realizar testes entre esses modelos:

```
*Modelo 1
xi: logistic esquerda i.privatizar [pweight=peso]
quietly fitstat, saving(1)

*Modelo 2
xi: logistic esquerda impostod [pweight=peso]
quietly fitstat, saving(2)

*Modelo 3
xi: logistic esquerda i.privatizar impostod [pweight=peso]

*Teste entre modelo 1 e modelo 3
fitstat, using(1)

*Teste entre modelo 2 e modelo 3
fitstat, using(2)
```

O teste entre o modelo 1 (privatizar) e o modelo 3 (privatizar e imposto) informa que a diferença de 5,506 no teste BIC indica apoio positivo ao modelo corrente (nesse caso, o modelo 3). Ou seja, o modelo fica melhor com a inclusão da variável "impostod".

O teste entre o modelo 2 (imposto) e o modelo 3 (privatizar e imposto) informa que a diferença de 10,07 no teste BIC evidencia apoio forte para o modelo salvo (nesse caso, o modelo 2). Ou seja, o modelo fica melhor somente com a variável explicativa "impostod".

Nos modelos com "privatizar" (modelos 1 e 3), somente os coeficientes que mostram a relação entre privatizar=1 e privatizar=5 são significantes. Se uma tabela de freqüência de "privatizar" for estimada, veremos que há somente 7 observações em "privatizar" iguais a 1, e também 7 observações iguais a 5:

```
tab privatizar
```

O que o teste está dizendo é que o pequeno número de dados não nos possibilita concluir que essa variável realmente tem uma influência na ideologia do deputado (esquerda/direita).

Há ainda um teste chamado de "lrtest" para fazer comparações entre modelos de regressão logísticos. O teste "lrtest" não funciona com o "pweight", mas somente com o "aweight" e "iweight". Por seu lado, o comando "logistic" não funciona com "aweight". Por isso, será usado o "iweight":

```
*Modelo A
xi: logistic esquerda impostod [iweight=peso]
estimates store A

*Modelo B
xi: logistic esquerda impostod i.privatizar [iweight=peso]
estimates store B

*Teste entre modelo A e modelo B
lrtest A B

Likelihood-ratio test                                 LR chi2(4)  =      6.82
(Assumption: A nested in B)                           Prob > chi2 =    0.1460
```

O teste "lrtest" indica que a probabilidade acima do valor calculado de qui-quadrado (6,82), levando em consideração a diferença entre os graus de liberdade dos modelos (4), é grande (0,1460). Isso indica mais uma vez que as variáveis dicotômicas de "privatizar" não trazem melhoras estatisticamente significantes para o nosso modelo.

O teste "lrtest" pode ser usado para comparar os modelos de regressão de mínimos quadrados ordinários:

```
*Modelo A
xi: regress esqdir impostod [aweight=peso]
estimates store A

*Modelo B
xi: regress esqdir impostod i.privatizar [aweight=peso]
estimates store B

*Teste entre modelo A e modelo B
lrtest A B

Likelihood-ratio test                                 LR chi2(4)  =     16.19
(Assumption: A nested in B)                           Prob > chi2 =    0.0028
```

No caso dos modelos de regressão de mínimos quadrados ordinários, da mesma forma como ocorreu no caso do teste "fitstat", o teste "lrtest" indica que as variáveis dicotômicas de "privatizar" ocasionaram um acréscimo explicativo para nossa variável dependente (esqdir).

De uma forma geral, o que vimos foi que:

– Os testes "tablesq", "fitstat" e "lrtest" indicaram que as variáveis dicotômicas de "privatizar" NÃO FORAM IMPORTANTES para aumentar o poder explicativo do modelo de regressão logística sobre a variável "esquerda".

– Os testes "fitstat" e "lrtest" indicaram que as variáveis dicotômicas de "privatizar" FORAM IMPORTANTES para aumentar o poder explicativo do modelo de regressão de mínimos quadrados ordinários sobre a variável "esqdir".

Portanto a forma como a variável dependente foi incluída no modelo (esqdir ou esquerda) e o tipo de regressão escolhida (OLS ou logística), nos leva a ter opiniões diferenciadas sobre a inclusão das variáveis dicotômicas de "privatizar". Nesse caso, é necessário analisar substancialmente qual a melhor forma de categorizar nossa variável dependente, e escolher a estratégia metodológica apropriada.

Modelos de efeitos fixos

Modelos de efeitos fixos permitem a estimação de coeficientes que refletem relações dentro de uma determinada variável (ano, área, legislatura, partido), levando em consideração as demais variáveis independentes. Em outras palavras, os coeficientes estimados das variáveis independentes foram estimados após controlar por cada um dos fatores da variável definida como efeito fixo. Isso significa que para cada valor dessa variável "efeito fixo" é estimado um coeficiente que não é reportado no resultado da regressão. Essa é uma questão teórico-metodológica, já que o pesquisador assume que os efeitos das outras variáveis independentes na variável dependente só poderão ser medidos se houver o controle por essa variável "efeito fixo".

Depois de agregar os bancos de dados da Argentina, Chile e Guatemala, podemos assumir que o efeito da variável "impostod" sobre a variável "esquerda" tem que ser controlado pelo país do deputado. Nesse caso, assumimos que, dependendo do país, o impacto da opinião do deputado acerca de imposto (*impostod*) será diferenciado sobre sua colocação entre esquerda e direita (variável *esquerda*):

```
set more off

*Agrupando os bancos de dados da Argentina, Chile e Guatemala
use "C:\cursodcp\dados\Argentina51.dta", clear
append using "C:\cursodcp\dados\Chile42.dta"
append using "C:\cursodcp\dados\Guatemala52.dta"

*Mantendo somente as variáveis de interesse no banco
keep pais p34 p35 p58 peso
label variable pais "País"

*No Chile, há opção 3 na variável p34,
*que significa "não aumentar impostos".
*Essa opção será colocada como missing
*em nossa nova variável.
tab p34 pais, nolabel missing

*Recodificando opinião dos deputados sobre imposto
tab p34, missing nolabel
gen imposto=p34
replace imposto=. if p34==3 | p34==8 | p34==9
tab p34 imposto, missing nolabel

*Recodificando variável imposto para 0/1
tab imposto, missing nolabel
gen impostod=.
replace impostod=0 if imposto==1
replace impostod=1 if imposto==2
tab imposto impostod

*Recodificando escala esquerda/direita do deputado
tab p58, missing nolabel
gen esqdir=p58
replace esqdir=. if p58==98 | p58==99
tab p58 esqdir, missing nolabel
```

```
*Criando variável dicotômica de deputado de esquerda
gen esquerda=.
replace esquerda=1 if esqdir>=1 & esqdir<=4
replace esquerda=0 if esqdir>=5 & esqdir<=10
tab esqdir esquerda, missing nolabel

*Excluindo valores missing de nossas variaveis de interesse
di _N
drop if esquerda==. | impostod==. | pais==.
di _N

*Modelo de efeito fixo por país
xtreg esquerda impostod, fe i(pais)
```

Note que não utilizamos a opção de peso, porque este deveria ser constante nos diferentes valores da variável de efeito fixo escolhida (*pais*). Essa é uma exigência do comando "xtreg" do Stata.

O resultado do modelo indica que o efeito fixo por país (*F test that all u_i=0*) é significativo [*F(2,236)=4.87; Prob>F=0.0085*]. A probabilidade acima do valor de F é inferior a 0.05.

Há ainda a indicação do teste "F" para todos os coeficientes [*F(1,236)=19.13; Prob>F=0.0000*], o qual também é significativo.