AULA 4
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: 25/05/2009
Horário: 18:00–21:00

Recursos disponíveis online

– Stata:
http://www.stata.com/links

– Centro de População da Carolina (CPC) da Universidade da Carolina do Norte de Chapel Hill (UNC):
http://www.cpc.unc.edu/services/computer/presentations/statatutorial

– Serviços de Tecnologia Acadêmica (ATS) da Universidade da Califórnia de Los Angeles (UCLA):
http://www.ats.ucla.edu/stat/stata/sk

– Site com explicações e exemplificações de comandos diversos de inferência estatística:
http://www.ats.ucla.edu/stat/stata/whatstat/whatstat.htm

Algumas observações

Estamos lembrando de colocar o "log using" para começar a salvar os resultados:

```
log using "C:\cursodcp\log\aula4.log", text replace
```

Além disso, é sempre bom lembrar-se de escrever "log close" no final do arquivo, para que esses resultados realmente sejam salvos. Ou seja, no final da aula rode todo o programa novamente para salvar o "log" completo:

```
log close
```

Encerrar o Stata:

```
exit
```

Se houver um banco de dados aberto no Stata, o ideal é digitar o seguinte comando para encerrar o programa sem salvar os dados:

```
exit, clear
```

Passos para realização de inferência estatística

– Recodificação de variáveis.
– Histograma.
– Gráfico de dispersão (variáveis independentes e variável dependente).
– Testes de significância: testes de correlação e significância entre variáveis.
– Regressão estatística: modelos de mínimos quadrados ordinários (ordinary least squares).

Vamos fazer a mesma alocação de informações entre registros da Aula 3

Vamos supor que queremos colocar algumas respostas do deputado mais velho de cada partido no registro dos outros deputados daquele mesmo partido. O pressuposto é de que o deputado mais velho teria influência sobre os demais deputados de seu partido em temas polêmicos ("p34": opinião sobre imposto; "p35": opinião sobre privatizações; e "p36": opinião sobre Mercosul):

```
use "C:\cursodcp\dados\Argentina51.dta", clear
replace p63=. if p63==99
gsort partido -p63
browse partido p63
egen depvel=tag(partido)
browse partido p63 depvel
sort partido
save "C:\cursodcp\dados\argoriginal.dta", replace

keep if depvel==1
keep partido p34 p35 p36
rename p34 p34vel
rename p35 p35vel
rename p36 p36vel
browse
sort partido
save "C:\cursodcp\dados\argvelho.dta", replace

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
```

Teste de Qui-quadrado (*Chi-square test*)

Antes de realizar o teste de qui-quadrado, é necessário recodificar as variáveis de interesse para colocar os valores "não sabe" e "não respondeu" como "missing":

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

*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

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

```
*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

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

*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
```

O teste de qui-quadrado é usado quando se deseja saber se há relação entre duas variáveis categóricas. A opção "chi2" é usada com o comando "tabulate" para obter o teste estatístico e seu valor "p".

```
tabulate imposto impostovel, chi2
tabulate privatizar privatizarvel, chi2
tabulate mercosul mercosulvel, chi2
```

Observamos que não há correlação significativa entre as respostas dos legisladores e as respostas do legislador mais velho no mesmo partido, em relação a opiniões sobre imposto (p34), privatização (p35) e Mercosul (p36).

Teste de "t" (*t-test*)

O teste de "t" (*t-test*) permite verificar se a média de uma variável com distribuição normal difere significativamente de um valor hipotético. Por exemplo, podemos testar se a média da auto-classificação dos deputados na escala esquerda/direita (esqdir) difere significantemente de 5,5:

```
tab p58, missing nolabel
gen esqdir=p58
replace esqdir=. if p58==99
tab esqdir, missing nolabel
histogram esqdir

ttest esqdir=5.5
```

A média da variável "esqdir" é 4,525, o que é estatisticamente significante diferente do valor de 5,5. Poderíamos concluir que nesse grupo de legisladores, eles se auto-classificam mais próximos à esquerda (ou centro-esquerda) do que a média geral da escala do questionário.

Análise de Variância (ANOVA)

Podemos utilizar o comando de análise de variância (anova) para estimar a relação entre uma variável independente categórica e uma variável dependente com distribuição normal. Esse comando testa se há diferenças entre as médias da variável dependente nos diversos níveis da variável independente.

Nesse caso, usaremos a auto-classificação do deputado (esqdir) como variável dependente, e sua opinião sobre imposto (p34) como variável independente:

```
twoway scatter esqdir imposto
twoway (lfit esqdir imposto) (scatter esqdir imposto)

anova esqdir imposto

             Number of obs =      82     R-squared     =  0.1475
             Root MSE      = 1.27353     Adj R-squared =  0.1368

    Source |  Partial SS    df       MS           F     Prob > F
-----------+----------------------------------------------------
     Model |  22.4438782     1  22.4438782      13.84     0.0004
           |
   imposto |  22.4438782     1  22.4438782      13.84     0.0004
           |
  Residual |  129.751244    80  1.62189055
-----------+----------------------------------------------------
     Total |  152.195122    81  1.87895212
```

– *Partial SS* é o mesmo que "partial sum of squares", ou soma parcial dos quadrados.
– *Df* é o mesmo que "degrees of freedom", ou graus de liberdade.
– *MS* é o mesmo que "mean square", ou média dos quadrados.
– *F* é o mesmo que "F Value", ou "F-test", ou teste de "F".

A média de auto-classificação na escala esquerda/direita é significantemente diferente entre as duas opiniões sobre imposto.

O mesmo exercício pode ser usado para a variável independente que indica a opinião dos deputados sobre privatização (p35), e sua relação com a variável dependente "esqdir":

```
twoway scatter esqdir privatizar
twoway (lfit esqdir privatizar) (scatter esqdir privatizar)

anova esqdir privatizar

             Number of obs =      79     R-squared     =  0.2331
             Root MSE      = 1.30119     Adj R-squared =  0.1916

    Source |  Partial SS    df       MS           F     Prob > F
-----------+----------------------------------------------------
     Model |  38.0773827     4  9.51934568       5.62     0.0005
           |
privatizar |  38.0773827     4  9.51934568       5.62     0.0005
           |
  Residual |  125.289706    74  1.69310413
-----------+----------------------------------------------------
     Total |  163.367089    78  2.09444985
```

A média da escala esquerda/direita também difere significantemente entre as cinco opiniões sobre privatização. Porém, não sabemos em quais níveis da variável de privatização essas médias variam.

Para saber a média da escala esquerda/direita para cada opinião de privatização, é possível combinar o comando "tabulate" com "summarize":

```
tabulate privatizar, summarize(esqdir)
```

ou

```
tab privatizar, sum(esqdir)
```

```
            |          Summary of esqdir
 privatizar |        Mean   Std. Dev.       Freq.
------------+------------------------------------
          1 |         5.9   1.3703203          10
          2 |   4.4583333   1.2846643          24
          3 |   4.6470588   1.2764143          34
          4 |   4.3333333   .57735027           3
          5 |           3   1.5118579           8
------------+------------------------------------
      Total |   4.5696203   1.4472214          79
```

Vemos que os deputados favoráveis a privatizar todos os serviços públicos (privatizar=1) são os que mais se classificam à direita (média=5,9).

Nos resultados gerados pelo comando "anova", os valores do teste "F" (*F-test*) dos modelos (*Model*) são os mesmos das variáveis independentes ("imposto" e "privatizar"). Isso aconteceu porque realizamos a análise com somente uma variável independente por comando. Se outras variáveis forem inseridas, o teste "F" será diferente no *Model* e nas variáveis independentes. Esse é o caso de:

```
anova esqdir imposto
anova esqdir privatizar
anova esqdir imposto privatizar
```

Correlação

O comando de correlação (corr) é útil para testar se há relação linear entre duas ou mais variáveis com intervalos normalmente distribuídos. Podemos testar nossas duas variáveis independentes ("imposto" e "privatizar"), mesmo sabendo que elas não têm distribuição normal:

```
hist imposto
hist privatizar

corr imposto privatizar
```

Também é possível estimar a significância dessa correlação:

```
pwcorr imposto privatizar, sig
```

Na realidade esse teste de correlação também pode ser realizado para variáveis dicotômicas (*dummy variables*). Isso seria mais aplicável ao nosso caso, já que "privatizar" é uma variável categórica que tem cinco valores, e "imposto" tem dois valores. Para realizar a correlação entre "imposto" separadamente com os cinco valores de "privatizar" podemos escolher diferentes estratégias.

Primeiro, poderíamos criar manualmente variáveis dicotômicas para cada valor de "privatizar":

```
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

corr imposto priv1 priv2 priv3 priv4 priv5
pwcorr imposto priv1 priv2 priv3 priv4 priv5, sig
```

Também podemos usar as opções "xi" e "i." para calcular a correlação entre "imposto" e "privatizar" rapidamente. Nesse caso, uma categoria da variável "privatizar" é omitida:

```
xi: corr imposto i.privatizar
xi: pwcorr imposto i.privatizar, sig
```

O Stata também gera automaticamente variáveis dicotômicas com o uso do comando "tabulate" e levando em consideração as categorias de uma determinada variável. No caso abaixo, serão criadas cinco variáveis ("privatizar1", "privatizar2", "privatizar3", "privatizar4" e "privatizar5"):

```
tab privatizar, gen(privatizar)

corr imposto privatizar1 privatizar2 privatizar3 privatizar4 privatizar5
pwcorr imposto privatizar1 privatizar2 privatizar3 privatizar4 privatizar5, sig
```

```
             |  imposto privat~1 privat~2 privat~3 privat~4 privat~5
-------------+------------------------------------------------------
     imposto |   1.0000
             |
             |
 privatizar1 |   0.2730   1.0000
             |   0.0203
             |
 privatizar2 |  -0.0973  -0.2607   1.0000
             |   0.4159   0.0180
             |
 privatizar3 |  -0.0157  -0.3397  -0.5715   1.0000
             |   0.8959   0.0018   0.0000
             |
 privatizar4 |   0.0732  -0.0767  -0.1291  -0.1682   1.0000
             |   0.5413   0.4934   0.2479   0.1310
             |
 privatizar5 |  -0.1612  -0.1294  -0.2178  -0.2837  -0.0641   1.0000
             |   0.1761   0.2465   0.0494   0.0098   0.5674
```

Ao elevar ao quadrado a correlação e multiplicar por 100, podemos determinar a porcentagem de variabilidade conjugada entre as variáveis.

Podemos realizar esse cálculo para os deputados que pensam que "todos serviços públicos deveriam ser privatizados" ("privatizar" igual a 1):

```
display (0.2730*0.2730)*100
```

O resultado indica que a variável "imposto" tem uma variação de 7,45% com a opção "1" da variável "privatizar".

Vamos fazer esse cálculo para os deputados que pensam que o correto é "privatizar todos os serviços públicos, com exceção dos que tiveram uma incidência para a maioria da população" ("privatizar" igual a 3), em que a correlação com "imposto" foi de –0,0157:

```
display (-0.0157*-0.0157)*100
```

Isso significa que a variável "imposto" tem uma variação de apenas 0,025% com a opção "3" da variável "privatizar".

De uma forma geral, as correlações entre "imposto" e "privatizar" são pequenas, algo corroborado pela baixa significância dos testes estatísticos realizados com comando "pwcorr". Em outras palavras, por não estarem correlacionadas, essas variáveis poderiam ser inseridas conjuntamente em um modelo de regressão como variáveis independentes.

Voltando ao exemplo do deputado mais velho

Voltando à nossa variável de influência de opinião do deputado mais velho (impostovel, privatizarvel, mercosulvel) nos demais deputados do partido (imposto, privatizar, mercosul), utilizamos anteriormente a opção de qui-quadrado dentro de "tabulate". Também podemos usar os comandos "corr" e "pwcorr":

```
*Opinião sobre imposto
tabulate imposto impostovel, chi2
corr imposto impostovel
pwcorr imposto impostovel, sig

*Opinião sobre privatização
tabulate privatizar privatizarvel, chi2
corr privatizar privatizarvel
pwcorr privatizar privatizarvel, sig

*Opinião sobre Mercosul
tabulate mercosul mercosulvel, chi2
corr mercosul mercosulvel
pwcorr mercosul mercosulvel, sig
```

No caso da opinião sobre imposto e privatização, os testes de qui-quadrado e correlação não foram significativos entre as opiniões de todos deputados e as opiniões dos deputados mais velhos de cada partido.

Quanto à opinião sobre o Mercosul, o teste de qui-quadrado não foi significativo e a correlação foi significativa. Aqui é necessário explorar um pouco mais as variáveis para entender o que está acontecendo:

```
tab p36, gen(p36d)

tab p36vel, gen(p36vel)
rename p36vel3 p36vel5
rename p36vel2 p36vel4
rename p36vel1 p36vel2

pwcorr p36 p36vel2 p36vel4 p36vel5, sig
```

```
             |      p36  p36vel2  p36vel4  p36vel5
-------------+------------------------------------
         p36 |   1.0000
             |
             |
     p36vel2 |  -0.2122   1.0000
             |   0.0359
             |
     p36vel4 |  -0.0111  -0.1183   1.0000
             |   0.9134   0.2412
             |
     p36vel5 |   0.2028  -0.8698  -0.3870   1.0000
             |   0.0452   0.0000   0.0001
```

Esses resultados indicam que as correlações são mais expressivas e significativas nos extremos da escala:

– Quanto mais o deputado mais velho está pouco satisfeito com a participação da Argentina no Mercosul (p36vel2), menor é a chance dos demais deputados do mesmo partido terem a mesma opinião.

– Por outro lado, quando mais o deputado mais velho está muito satisfeito com a participação da Argentina no Mercosul (p36vel5), maior é a chance dos demais deputados do mesmo partido terem a mesma opinião.

Ou seja, é preciso avaliar de diferentes formas as relações entre as variáveis, com o intuito de entender o fenômeno social estudado.

Regressão linear simples

Primeiramente, vamos recodificar a variável "imposto" para que tenha valores 0 e 1:

```
tab imposto, missing nolabel
gen impostod=.
replace impostod=0 if imposto==1
replace impostod=1 if imposto==2
tab imposto impostod
```

Agora vamos utilizar essa nova variável imposto (impostod) como variável independente para explicar a escala esquerda/direita (esqdir):

```
regress esqdir impostod
```

Também podemos utilizar a variável "privatizar" como variável independente na explicação de "esqdir":

```
regress esqdir privatizar
```

Regressão múltipla

Podemos utilizar as variáveis dicotômicas de "privatizar" para melhor entender o seu efeito em "esqdir". Qual deveria ser a categoria de referência? Geralmente deve ser a que possui mais observações:

```
tab privatizar, nomissing nolabel
```

Nesse caso, deveríamos utilizar os deputados que pensam que o correto é "privatizar todos os serviços públicos, com exceção dos que tiveram uma incidência para a maioria da população" ("privatizar" igual a 3). Por isso, omitimos ("omit") essa terceira categoria antes de rodar a regressão:

```
char privatizar[omit] 3
xi: regress esqdir i.privatizar
```

As variáveis significativas são aquela dos deputados que concordam em "privatizar todos os serviços públicos" ("privatizar" igual a 1) e a dos deputados que pensam que "nenhum serviço público deveria ser privatizado" ("privatizar" igual a 5):

– Os deputados a favor da privatização têm maior chance de ser de direita (efeito positivo na escala esquerda/direita).

– Os deputados contra a privatização têm menor chance de ser de direita (efeito negativo na escala esquerda/direita).

Uma forma de facilitar a leitura dos resultados, seria colocar "privatizar" igual a 1 ou 5 como referência. Como "privatizar" igual a 1 e 5 tem poucos casos, a categoria de "privatizar" igual a 2 também passa a ser significante, o que pode não ser correto (já que antes somente 1, 3 e 5 eram significantes):

```
char privatizar[omit] 1
xi: regress esqdir i.privatizar

char privatizar[omit] 5
xi: regress esqdir i.privatizar
```

Podemos também incluir tanto "impostod", como as variáveis dicotômicas de "privatizar", conjuntamente na explicação de "esqdir":

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

Podemos testar uma interação entre "impostod" e "privatizar1", o que acaba não gerando resultados significantes:

```
gen impprev1=impostod*privatizar1

char privatizar[omit] 3
xi: regress esqdir impostod i.privatizar impprev1
```

Para dizer ao Stata que as próximas regressões devem utilizar a categoria padrão de "privatizar" como referência, escrevemos:

```
char privatizar[omit]
```

Regressão múltipla multivariada

Podemos utilizar o comando "mvreg" no caso de termos duas ou mais variáveis dependentes que serão preditas por uma série de outras variáveis.

Em nosso caso, além da escala esquerda/direita do próprio deputado, podemos estimar o efeito de "impostod" e "privatizar" na escala que o deputado atribui ao seu partido (p59):

```
tab p59, missing nolabel
gen esqdirp=p59
replace esqdirp=. if p59==99
tab esqdirp, missing nolabel

char privatizar[omit] 3
xi: mvreg esqdir esqdirp = impostod i.privatizar
```