AULA 3
CURSO DE INTRODUÇÃO AO STATA

Professor: Ernesto Friedrich de Lima Amaral (DCP/UFMG)
Email: eflamaral@gmail.com
Site do curso: www.ernestoamaral.com/stata20092c.html
Data: 18/12/2009
Horário: 8:00 às 12:00

Passos para realização de inferência estatística

A intenção até agora foi de mostrar como os diferentes passos para realização de inferência estatística podem ser realizados no Stata. A continuação desse exercício é justamente a exploração das possibilidades de realização de regressão estatística neste pacote. Estes foram os passos abordados:

– Organização do banco de dados.
– Recodificação de variáveis.
– Histograma.
– Gráfico de dispersão (variáveis independentes e variável dependente).
– Exploração de outros gráficos.
– Testes de significância: testes de correlação e significância entre variáveis.
– Agora vamos para as regressões estatísticas.

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.

Abrindo o banco e recodificando variáveis de interesse:
– Anos de estudo (v4803)
– Rendimento no trabalho principal (v4718)

```
use "C:\curso\dados\pes2007MG.dta", clear
gen estudo=.
replace estudo=v4803-1 if v4803>=1 & v4803<=16
replace v4718=. if v4718==999999999999
gen lnv4718=ln(v4718)
```

Usaremos o logaritmo do rendimento no trabalho principal (lnv4718) como variável dependente, e anos de estudo como variável independente:

```
twoway (lfit lnv4718 estudo) (scatter lnv4718 estudo)

anova lnv4718 estudo
```

```
                  Number of obs =   16302     R-squared     =  0.2315
                  Root MSE      = .769616     Adj R-squared =  0.2308

    Source |  Partial SS    df       MS           F     Prob > F
-----------+------------------------------------------------------
     Model |  2906.08471    15  193.738981     327.09     0.0000
           |
    estudo |  2906.08471    15  193.738981     327.09     0.0000
           |
  Residual |  9646.33992 16286  .592308726
-----------+------------------------------------------------------
     Total |  12552.4246 16301  .770040159
```

– *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".

O logaritmo do rendimento no trabalho principal é significantemente diferente entre as categorias de anos de estudo.

O mesmo exercício pode ser usado para a variável independente que indica a cor/raça dos indivíduos (v0404), e sua relação com a variável dependente "lnv4718". Vamos utilizar a dicotomização entre brancos e não-brancos:

```
gen branca=.
replace branca=0 if v0404==4 | v0404==8
replace branca=1 if v0404==2

twoway (lfit lnv4718 branca) (scatter lnv4718 branca)

anova lnv4718 branca
```

```
                  Number of obs =   16297     R-squared     =  0.0387
                  Root MSE      = .859632     Adj R-squared =  0.0386

    Source |  Partial SS    df       MS           F     Prob > F
-----------+------------------------------------------------------
     Model |  484.121208     1  484.121208     655.13     0.0000
           |
    branca |  484.121208     1  484.121208     655.13     0.0000
           |
  Residual |  12041.4794 16295  .738967746
-----------+------------------------------------------------------
     Total |  12525.6006 16296  .768630377
```

Apesar do teste de F ser significativo, a capacidade de explicação de cor/raça na variação do rendimento no trabalho principal é bem inferior que aquela encontra pelos anos de estudo.

Podemos ainda incluir as duas variáveis independentes na explicação do rendimento:

```
anova lnv4718 branca estudo
```

Regressão linear simples

Agora vamos utilizar a variável recodificada de anos de estudo (estudo) como variável independente para explicar o rendimento no trabalho principal (lnv4718):

```
regress lnv4718 estudo
```

Também podemos utilizar a variável "branca" como variável independente na explicação de "lnv4718":

```
regress lnv4718 branca
```

Regressão múltipla

Podemos utilizar as variáveis dicotômicas de grupos de anos de estudo (v4838) melhor entender o seu efeito no rendimento (lnv4718). Primeiramente, vamos transformar em "missing" os casos "não-determinados" (v4838=7):

```
gen grestudo=.
replace grestudo=v4838 if (v4838>=1 & v4838<=6)
tab v4838 grestudo, missing
```

Qual deveria ser a categoria de referência? Geralmente deve ser a que possui mais observações:

```
tab grestudo
```

Nesse caso, deveríamos utilizar as pessoas de 4 a 7 anos de estudo (v4838 igual a 3). Por isso, omitimos ("omit") essa terceira categoria antes de rodar a regressão:

```
char grestudo[omit] 3
xi: regress lnv4718 i.grestudo
```

Uma forma de facilitar a leitura dos resultados, seria colocar a categoria 1 ou 6 de grupos de estudo (grestudo) como referência. Como pessoas com 15 anos ou mais de escolaridade são poucos casos (grestudo=6), vamos utilizar a categoria das pessoas sem instrução e com menos de um ano de estudo (grestudo=1) como referência.

```
char grestudo[omit] 1
xi: regress lnv4718 i.grestudo
```

Podemos também incluir o indicador de raça branca ("branca") e as variáveis dicotômicas de grupos de estudo ("grestudo") conjuntamente na explicação de rendimento ("lnv4718"):

```
char grestudo[omit] 1
xi: regress lnv4718 branca i.grestudo
```

Podemos testar interações entre "branca" e as variáveis dicotômicas de "grestudo", o que acaba não gerando resultados estatisticamente significantes:

```
char grestudo[omit] 1
xi: regress lnv4718 i.grestudo*branca
```

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

```
char grestudo[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 do logaritmo do rendimento mensal no trabalho principal (lnv4718), vamos utilizar os logaritmos do rendimento mensal de todos trabalhos (lnv4719) e do rendimento mensal de todas as fontes (lnv4720):

```
replace v4719=. if v4719==999999999999
gen lnv4719=ln(v4719)

replace v4720=. if v4720==999999999999
gen lnv4720=ln(v4720)
```

Cada modelo terá como variáveis independentes raça e grupos de estudo:

```
char grestudo[omit] 1
xi: mvreg lnv4718 lnv4719 lnv4720 = branca i.grestudo
```

Utilizando peso nas regressões

Vamos ver as estatísticas descritivas do peso da pessoa (v4729) de nosso banco de dados:

```
sum v4729, detail
```

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 grestudo[omit] 1
xi: regress lnv4718 branca i.grestudo [iweight=v4729]
```

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 lnv4718 branca i.grestudo [aweight=v4729]
```

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 peso "pweight" (peso amostral) é o inverso da probabilidade da observação ser incluída no banco, devido ao desenho amostral. A sua utilização ocorre da seguinte forma:

```
xi: regress lnv4718 branca i.grestudo [pweight=v4729]
```

Utilizando peso nas tabelas

Os pesos também podem ser utilizados na elaboração de tabelas. Com o "aweight" (peso analítico), o total na tabela é o mesmo número de observações no banco. Vamos tabular a variável sexo (v0302):

```
tab v0302 [aweight=v4729]
count
```

O "iweight" (peso importância) e o "fweight" (peso freqüência) expandem 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.

```
tab v0302 [iweight=v4729]
tab v0302 [fweight=v4729]
```

É bom lembrar que o "fweight" só funciona quando não há decimais no peso.

Regressão logística

Vamos elaborar uma regressão logística em que nossa variável dependente binária indicará se a mulher teve filho no último ano.

As variáveis utilizadas para criar nossa variável dependente são:
– Mês de nascimento do último filho tido nascido vivo (v1181)
– Ano de nascimento do último filho tido nascido vivo (v1182)

```
sum v1181
sum v1182
```

Como a data de referência da PNAD que estamos trabalhando é de 29 de setembro de 2007, os filhos nascidos entre outubro de 2006 e setembro de 2007 devem ser considerados na variável dependente:

```
gen filho=.

*Mulheres entre 15 e 49 anos:
replace filho=0 if v0302==4 & (v8005>=15 & v8005<=49)

*Mulheres entre 15 e 49 anos,
*com filhos entre outubro de 2006 a setembro de 2007:
replace filho=1 if ///
    ((v1181>=10 & v1181<=12) &  (v1182==2006) & (v8005>=15 & v8005<=49)) ///
                              | ((v1182==2007) & (v8005>=15 & v8005<=49))

tab filho, missing
```

Verificando se variável foi criada corretamente:

```
*Idade da pessoa se filho igual a 0 ou 1:
sum v8005 if filho==0 | filho==1

*Sexo da pessoa se filho igual a 0 ou 1:
sum v0302 if filho==0 | filho==1
```

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

```
char grestudo[omit]1
xi: logit filho branca i.grestudo [pweight=v4729]
xi: logistic filho branca i.grestudo [pweight=v4729]
```

Veja que a razão de chances de "branca" estimado por "logistic" (0,8514709) é igual ao exponencial do coeficiente estimado por "logit" (-0,1607899):

```
di exp(-0.1607899)
```

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

```
char grestudo[omit]1
xi: glm filho branca i.grestudo [pweight=v4729], family(binomial) link(logit)
xi: logit filho branca i.grestudo [pweight=v4729]
```

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

```
char grestudo[omit]1
xi: glm filho branca i.grestudo [pweight=v4729], family(binomial) link(logit)
eform
xi: logistic filho branca i.grestudo [pweight=v4729]
```

Se utilizarmos o peso inapropriado para regressão (fweight) podemos ter a falsa impressão de que temos resultados estatisticamente significantes:

```
char grestudo[omit]1
xi: glm filho branca i.grestudo [fweight=v4729], family(binomial) link(logit)
eform
xi: logistic filho branca i.grestudo [fweight=v4729]
```

Recodificando a variável idade (v8005) em grupos quinquenais:

```
gen gridade = .
replace gridade = 00 if v8005 >=  0 & v8005 <=  4
replace gridade = 05 if v8005 >=  5 & v8005 <=  9
replace gridade = 10 if v8005 >= 10 & v8005 <= 14
replace gridade = 15 if v8005 >= 15 & v8005 <= 19
replace gridade = 20 if v8005 >= 20 & v8005 <= 24
replace gridade = 25 if v8005 >= 25 & v8005 <= 29
replace gridade = 30 if v8005 >= 30 & v8005 <= 34
replace gridade = 35 if v8005 >= 35 & v8005 <= 39
replace gridade = 40 if v8005 >= 40 & v8005 <= 44
replace gridade = 45 if v8005 >= 45 & v8005 <= 49
replace gridade = 50 if v8005 >= 50 & v8005 <= 54
replace gridade = 55 if v8005 >= 55 & v8005 <= 59
replace gridade = 60 if v8005 >= 60 & v8005 <= 64
replace gridade = 65 if v8005 >= 65 & v8005 <= 69
replace gridade = 70 if v8005 >= 70 & v8005 <= 74
replace gridade = 75 if v8005 >= 75 & v8005 <= 79
replace gridade = 80 if v8005 >= 80 & v8005 <= 84
replace gridade = 85 if v8005 >= 85 & v8005 <= 120
```

Rodando a regressão de filho nascido vivo no último ano com grupos de idade como variável independente:

```
char gridade[omit]15
xi: logistic filho i.gridade [pweight=v4729]
```

Análise de regressão com gráficos

Regressão entre logaritmo do rendimento no trabalho principal (lnv4718) e raiz quadrada da idade (idraiz):

```
gen idraiz=v8005^1/2
regress lnv4718 idraiz [pweight=v4729]
```

Salvar valores preditos e resíduos:

```
predict rendap
gen rendar=lnv4718-rendap
```

Criar rótulos para variáveis:

```
label variable lnv4718 "Log do rendimento do trabalho principal"
label variable idraiz "Raiz quadrada da idade"
label variable rendap "Valores preditos de renda"
label variable rendar "Resíduo da regressão entre renda e idade"
```

Gráfico de dispersão dos valores originais de rendimento do trabalho principal (lnv4718) e valores preditos deste rendimento (rendap) pela variável original de idade (v8005):

```
twoway (scatter lnv4718 v8005) (scatter rendap v8005)
```

Gráfico de dispersão dos resíduos da regressão entre rendimento do trabalho principal e idade (rendar) pelos valores preditos de rendimento (rendap):

```
twoway (scatter rendar rendap, yline(0))
```

Ferramenta importante (Stata como interface de sistema operacional)

Alguns comandos básicos no Stata são importantes para lidar com o programa, e se assemelham aos comandos do DOS e UNIX:

– **pwd** Mostrar diretório em que se encontra.
– cd Mudar para diretório indicado.
– **sysdir** Mostrar diretório em que o Stata está instalado.
– **mkdir** Criar diretório em seu computador.
– **dir** Ver o conteúdo do diretório em que se encontra.
– **erase** Apagar arquivo no diretório especificado.
– **copy** Copiar arquivos para mesmo diretório ou diretório diferente.
– **type** Mostrar conteúdo de arquivo na tela do Stata.

Diretório em que o Stata está instalado:

```
. sysdir
   STATA:  C:\Programas\Stata10_SE\
 UPDATES:  C:\Programas\Stata10_SE\ado\updates\
    BASE:  C:\Programas\Stata10_SE\ado\base\
    SITE:  C:\Programas\Stata10_SE\ado\site\
    PLUS:  c:\ado\plus\
PERSONAL:  c:\ado\personal\
OLDPLACE:  c:\ado\
```

– UPDATES Stata dá preferência a arquivos nesse diretório.
– BASE Diretório com comandos originais.
– PLUS Armazena tudo oriundo da internet, também chamado de STBPLUS.
– PERSONAL Arquivos pessoais.
– c:\ado Comandos são armazenados em sub-pastas com a primeira letra do comando.

Nunca é bom utilizar o diretório em que o Stata está instalado para trabalhar, já que arquivos de programas podem ser apagados, ou arquivos pessoais podem ser removidos em uma atualização do programa. O recomendado é utilizar um diretório para cada projeto.

Você pode criar diretórios com o comando **mkdir** no Stata, ou no próprio Windows Explorer. Vamos supor que queremos criar um diretório para esse curso no drive C:\. Esses são os procedimentos:

```
cd C:\
mkdir curso1
cd curso1
mkdir dados
cd dados
```

A partir de agora, assim que começar o Stata, é possível mudar para o diretório recentemente criado:

```
cd C:\curso1\dados
```

Se você criar um diretório com espaço entre as palavras, é necessário utilizar aspas:

```
cd "C:\curso 1\dados"
```

Você também pode ver o conteúdo de diretórios, apagar arquivos, copiar arquivos, e examinar o conteúdo de um arquivo.

Listar o conteúdo do diretório atual:

```
dir
```

Listar os arquivos que possuem a extensão ".dta":

```
dir *.dta
```

Listar os arquivos que possuem a extensão ".dta" e nomes que começam com "prog" e possuem mais dois caracteres:

```
dir prog??.dta
```

Apagar um arquivo no diretório atual:

```
erase meuarquivo.xyz
```

Copiar arquivo no diretório atual. Isso é importante para realizar cópia de segurança, antes de realizar mudanças em um arquivo:

```
copy meuarquivo.abc meuarquivo.bak
```

Copiar um arquivo para um diretório diferente:

```
copy D:\dados\meuarquivo.dta C:\curso1\dados\meuarquivo.dta
```

Ver o conteúdo de um arquivo:

```
type meusdados.raw
```