AULA 2
CURSO DE INTRODUÇÃO AO STATA

Professor: Ernesto Friedrich de Lima Amaral (DCP/UFMG)
Email: eflamaral@gmail.com
Site do curso: www.ernestoamaral.com/stata20092b.html
Data: 05/12/2009
Horário: 8:00 às 12:00 e 14:00 às 18:00

Inclusão de observações de outros bancos (*append*)

Vamos utilizar o comando "append" (inclusão de observações) para juntar os bancos de Minas Gerais e Goiás:

```
use "C:\curso\dados\pes2007MG.dta", clear
append using "C:\curso\dados\pes2007GO.dta", nolabel
save "C:\curso\dados\pes2007MGGO.dta", replace
```

Acima foi utilizada a opção "nolabel" para excluir os rótulos das categorias do segundo banco. Isso é importante de ser realizado quando determinadas variáveis possuem rótulos diferentes entre os bancos que estão sendo agrupados.

Inclusão de variáveis de outros bancos (*merge*)

Suponha que queremos incluir as variáveis de domicílio de Minas Gerais no banco de indivíduos. Para isto, utilizamos o comando "merge" (inclusão de variáveis).

Antes de tudo, é preciso ordenar os bancos pelas variáveis de identificação. No caso da PNAD, as variáveis de identificação são número de controle (v0102) e número de série (v0103).

No entanto, incluímos também ano de referência (v0101) e Unidade da Federação (uf), simplesmente porque elas aparecem nos dois bancos. Se estas duas variáveis não fossem incluídas no comando, o resultado acabaria sendo o mesmo, já que os dados do domicílio seriam gravados sobre os dados de indivíduos.

Primeiramente, ordenamos o banco de domicílios:

```
use "C:\curso\dados\dom2007MG.dta", clear
sort v0101 uf v0102 v0103
save "C:\curso\dados\dom2007MG.dta", replace
```

Em seguida, ordenamos o banco de indivíduos e realizamos a junção dos dados:

```
use "C:\curso\dados\pes2007MG.dta", clear
sort v0101 uf v0102 v0103
merge v0101 uf v0102 v0103 using "C:\curso\dados\dom2007MG.dta"
```

Note que a variável "_merge" é criada automaticamente pelo Stata e pode assumir os seguintes resultados:

_merge=1 observações somente do banco de dados mestre (*master data*)
_merge=2 observações somente do banco de dados secundário (*using data*)
_merge=3 observações dos dois bancos (*master* e *using*)

O ideal é que todas as observações de "_merge" tenham valor igual 3, já que isso seria um indício de que houve a junção de dados de domicílios e pessoas em todos os casos.

No entanto, observamos que há 2.906 casos que são oriundos somente do banco de domicílios:

```
tab _merge
```

_merge	Freq.	Percent	Cum.
2	2,906	7.41	7.41
3	36,320	92.59	100.00
Total	39,226	100.00	

É preciso investigar porque isso aconteceu. Vamos olhar todas variáveis para os casos em que "_merge" é igual a 2:

```
browse if _merge==2
```

Observando o banco, vemos que a primeira variável de domicílio (tipo de entrevista – v0104) possui valores diferentes de "missing", mas as variáveis seguintes apresentam valores em branco. Vamos ver os valores dessa variável quando "_merge" é igual a 3:

```
tab v0104 if _merge==3, missing
```

Os 36.320 casos indicam que a entrevista foi realizada (v0104=1).

Agora vamos ver a tabela desta variável "v0104" somente no caso de "_merge" ser igual a 2:

```
tab v0104 if _merge==2, missing
```

Perceba que neste segundo caso, as observações são de entrevistas não realizadas (v0104 diferente de 1):
TIPO A UNIDADE OCUPADA
v0104=02 (Fechada)
v0104=03 (Recusa)
v0104=04 (Outra)

TIPO B UNIDADE VAGA
v0104=05 (Em condições de ser habitada)
v0104=06 (Uso ocasional)
v0104=07 (Construção ou reforma)
v0104=08 (Em ruínas)

TIPO C UNIDADE INEXISTENTE
v0104=09 (Demolida)
v0104=10 (Não foi encontrada)
v0104=11 (Não residencial)
v0104=12 (Fora do setor)

Ou seja, o total de observações em que as entrevistas não foram realizadas é o mesmo que o número de vezes em que "_merge" é igual a dois.

Tipo de entrevista	Freq.	Percent	Cum.
2	286	9.84	9.84
3	112	3.85	13.70
4	16	0.55	14.25
5	1,182	40.67	54.92
6	862	29.66	84.58
7	202	6.95	91.53
8	67	2.31	93.84
9	73	2.51	96.35
10	30	1.03	97.38
11	76	2.62	100.00
Total	2,906	100.00	

No decorrer do banco, algumas variáveis de domicílio apresentam valores diferentes de "missing", a partir da variável "v4105". Estas são variáveis derivadas que são possíveis de ser captadas mesmo não havendo realização da entrevista.

Vimos que não houve problema na realização do "merge" entre bancos de domicílios e pessoas. Por isso, podemos excluir os casos em que "_merge" é igual a dois e salvar o banco:

```
drop if _merge==2
drop _merge
save "C:\curso\dados\pesdom2007MG.dta", replace
```

Gráficos

De uma forma geral, é bom olhar o menu *Graphics* para explorar os diversos tipos de gráficos elaborados pelo Stata. Aqui vou colocar alguns exemplos:

– Gráfico de barras:
Média do rendimento mensal no trabalho principal (v4718), de todos trabalhos (v4719) e de todas as fontes (v4720) por raça (branca X não branca).

Recodificando variável cor/raça:

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

Transformando não declaração de rendimento (999.999.999.999) em "missing":

```
replace v4718=. if v4718==999999999999
replace v4719=. if v4719==999999999999
replace v4720=. if v4720==999999999999
```

Definindo o rótulo da variável cor/raça:

```
label define branca 0 "Preta/Parda" 1 "Branca"
label values branca branca
```

Gerando o gráfico de barras:

```
graph bar (mean) v4718 v4719 v4720, over(branca) ///
///
title("Média de rendimento de trabalho principal, de todos trabalhos," ///
"de todas as fontes por cor/raça em Minas Gerais, PNAD 2007" ///
, size(medlarge)) ///
///
ytitle("Média de rendimento") ///
///
blabel(bar, format(%9,2fc)) ///
///
bar(1, fcolor(dknavy) lcolor(dknavy)) ///
bar(2, fcolor(gray) lcolor(gray)) ///
bar(3, fcolor(dkgreen) lcolor(dkgreen)) ///
///
legend(title(Tipos de rendimento, size(medsmall)) ///
label(1 "Trabalho principal") ///
label(2 "Todos trabalhos") ///
label(3 "Todas as fontes"))
```

Salvando o gráfico como figura:

```
graph export "C:\curso\grafs\medrenda.wmf", replace
```

– Histogramas:

Transformando anos de estudo não-determinados (v4803=17) em "missing":

```
gen estudo=.
replace estudo=v4803-1 if (v4803>=1 & v4803<=16)
```

Histograma de anos de estudo:

```
histogram estudo, discrete frequency ///
color(black) lcolor(gray) ///
title("Histograma de anos de estudo em Minas Gerais, PNAD 2007", ///
color(black)) ///
ytitle(Frequência) ///
ylabel(0(2000)8000, grid glcolor(black) glpattern(dash)) ///
xtitle(Anos de estudo) ///
xlabel(1(1)17 17 " ") ///
mcolor(black) fcolor(black) lcolor(white) ///
plotregion(lcolor(black)) ///
graphregion(color(white) icolor(white))
```

Salvando histograma como figura:

```
graph export "C:\curso\grafs\histestudo.wmf", replace
```

Histograma de rendimento do trabalho principal (v4718):

```
histogram v4718, frequency color(black) ///
ytitle(Frequência) ///
xtitle(Rendimento no trabalho principal) ///
title("Histograma de rendimento no trabalho principal" ///
"em Minas Gerais, PNAD 2007")
```

Salvando histograma como figura:

```
graph export "C:\curso\grafs\histv4718.wmf", replace
```

Histograma do logaritmo de rendimento no trabalho principal:

```
gen lnv4718=ln(v4718)

histogram lnv4718, frequency color(black) ///
ytitle(Frequência) ///
ylabel(0(1000)4000, format(%9,0fc)) ///
xlabel(2(1)10) ///
xtitle(Log do rendimento no trabalho principal) ///
title("Histograma do logaritmo de rendimento no trabalho principal" ///
"em Minas Gerais, PNAD 2007")
```

Salvando histograma como figura:

```
graph export "C:\curso\grafs\histlnv4718.wmf", replace
```

– Gráfico de caixa (box-plot):

Distribuição de anos de estudo:

```
label variable estudo "Anos de estudo"

graph hbox estudo, ylabel(0(1)16) ///
title("Distribuição de anos de estudo em Minas Gerais, 2007" ///
, size(large)) ///
note(Fonte: PNAD 2007)
```

Salvando gráfico de caixa de anos de estudo:

```
graph export "C:\curso\grafs\boxestudo.wmf", replace
```

– Gráfico de dispersão (scatterplot):

Idade do indivíduo (v8005) pelo logaritmo do rendimento no trabalho principal (lnv4718):

```
twoway (scatter lnv4718 v8005) (lfit lnv4718 v8005, lwidth(vthick)), ///
legend(label(1 "Logaritmo do rendimento") label(2 "Tendência Linear")) ///
title("Gráfico de dispersão da idade do indivíduo" ///
"pelo logaritmo do rendimento no trabalho principal", ///
color(black) size(large)) ///
xtitle("Idade") ///
xlabel(0(20)100) ///
ytitle("Logaritmo do rendimento") ///
ylabel(1(1)11, grid glcolor(gray) glpattern(dash)) ///
plotregion(lcolor(black)) ///
graphregion(color(white) icolor(white))
```

Salvando gráfico de dispersão:

```
graph export "C:\curso\grafs\idadeXtrab.wmf", replace
```

Utilização do comando foreach para recodificação e criação de grupos de variáveis

As seguintes variáveis possuem o código 999.999.999.999 quando não houve declaração de renda:
– Rendimento mensal no trabalho principal (v4718)
– Rendimento mensal de todos trabalhos (v4719)
– Rendimento mensal de todas as fontes (v4720)
– Rendimento mensal domiciliar (v4721)
– Rendimento mensal familiar (v4722)

Abra o banco de dados original, antes de rodar os comandos abaixo:

```
use "C:\curso\dados\pes2007MG.dta", clear
```

Com o intuito de reclassificar as respostas "sem declaração" para "missing", poderíamos usar todas essas linhas de comando:

```
replace v4718=. if v4718==999999999999
replace v4719=. if v4719==999999999999
replace v4720=. if v4720==999999999999
replace v4721=. if v4721==999999999999
replace v4722=. if v4722==999999999999
```

No entanto, podemos simplesmente escrever:

```
foreach x of varlist v4718-v4722 {
  replace `x'=. if `x'==999999999999
}
```

Depois de recodificar os valores "sem declaração" das variáveis v4718 a v4722, podemos gerar novas variáveis para informar as médias de renda por sexo (v0302):

```
sort v0302
foreach x of varlist v4718-v4722 {
  by v0302: egen med`x' = mean(`x')
}

browse v0302 v4718-v4722 med*
browse v0302 v4718-v4722 med*, nolabel
```

Criação de bancos de dados agrupados por categorias de variáveis (*collapse*)

Vamos criar e salvar um banco de dados que vai indicar médias de rendimentos (v4718-v4722) por sexo (v0302) e cor/raça (branca). Abra o banco de dados original, antes de rodar os comandos abaixo:

```
use "C:\curso\dados\pes2007MG.dta", clear
```

Criando variável cor ou raça: branca X preta/parda:

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

Recodificando rendimento:

```
foreach x of varlist v4718-v4722 {
  replace `x'=. if `x'==999999999999
}
```

Calculando médias de rendimentos (v4718-v4722), além do número de observações não-missings e da mediana do rendimento do trabalho principal (v4718), pelas categorias de sexo (v0302) e de cor/raça (branca):

```
sort v0302 branca
collapse (mean) v4718-v4722 (count) c4718=v4718 (median) med4718=v4718, ///
by(v0302 branca)
browse
```

Salvando os dados agrupados:

```
save "C:\curso\dados\collapse.dta", replace
```

Reorganização do arranjo (formato) dos bancos de dados

O comando "reshape" muda o formato dos bancos de dados, segundo a necessidade do estudo em questão. Este é o exemplo da ajuda do Stata:

```
(wide form)
i        ....... x_ij ........
id  sex   inc80   inc81   inc82
------------------------------
1    0    5000    5500    6000
2    1    2000    2200    3300

(long form)
i     j           x_ij
id   year   sex    inc
-----------------------
1     80     0   5000
1     81     0   5500
1     82     0   6000
2     80     1   2000
2     81     1   2200
2     82     1   3300
```

O comando para mudar o formato dos bancos de "wide" (amplo) para "long" (dados em painel) é:

```
reshape long inc, i(id) j(year)
```

O comando para mudar o formato dos bancos de "long" (dados em painel) para "wide" (amplo) é:

```
reshape wide inc, i(id) j(year)
```

Primeiramente, vamos limpar os dados de Minas Gerais, provenientes da PNAD de 2006:

```
use "C:\curso\dados\pes2006MG.dta", clear
keep v0101 uf v0302 v0404 v4718-v4720
save "C:\curso\dados\pes2006MGpeq.dta", replace
```

Agora vamos limpar os dados de Minas Gerais, provenientes da PNAD de 2007 e agrupar com os dados de 2006:

```
use "C:\curso\dados\pes2007MG.dta", clear
keep v0101 uf v0302 v0404 v4718-v4720
append using "C:\curso\dados\pes2006MGpeq.dta"
save "C:\curso\dados\peq2006-2007MG.dta", replace
```

Organizando variável sexo (v0302):

```
gen mulher=.
replace mulher=1 if v0302==4
replace mulher=0 if v0302==2
label define mulher 0 "Homem" 1 "Mulher"
label values mulher mulher
```

Organizando variável cor/raça (v0404):

```
gen branca=.
replace branca=0 if v0404==4 | v0404==8
replace branca=1 if v0404==2
label variable branca "Cor ou raça: branca X preta/parda"
label define branca 0 "Preta/Parda" 1 "Branca"
label values branca branca
```

Organizando variáveis de rendimento (v4718, v4719 e v4720):

```
*Rendimento mensal no trabalho principal (v4718)
foreach x of varlist v4718-v4720 {
  replace `x'=. if `x'==999999999999
  gen ln`x'=ln(`x')
}
```

Salvando dados recodificados:

```
save "C:\curso\dados\peq2006-2007MGrec.dta", replace
```

Agora vamos utilizar o comando "collapse" para calcular as médias de rendimentos (v4718-v4720) por ano, sexo e cor/raça:

```
sort v0101 mulher branca
collapse (mean) v4718-v4720 lnv4718-lnv4720, by(v0101 mulher branca)
```

Salvando os dados agrupados no formato "long":

```
save "C:\curso\dados\collapselong.dta", replace
```

Vamos organizar o ano de referência (v0101):

```
gen ano=.
replace ano=0 if v0101==2006
replace ano=1 if v0101==2007
```

Neste momento, temos os dados em painel (long). Desta forma, podemos estimar modelos de regressão que calculam o impacto do sexo (mulher), cor/raça (branca) em cada um dos rendimentos (v4718, v4719 e v4720), ao longo do tempo (ano). Podemos estimar um modelo da seguinte forma:

```
reg lnv4718 mulher branca ano
```

Podemos passar este banco de dados para o formato amplo (wide):

```
(wide form)
i        ....... x_ij ........
id  sex   inc80   inc81   inc82
------------------------------
1    0    5000    5500    6000
2    1    2000    2200    3300

(long form)
i     j           x_ij
id   year   sex    inc
----------------------
1     80     0   5000
1     81     0   5500
1     82     0   6000
2     80     1   2000
2     81     1   2200
2     82     1   3300
```

Antes disso, vamos apagar a variável "v0101", já que ela foi recodificada para "ano":

```
drop v0101
```

O comando para mudar o formato dos bancos de "long" (dados em painel) para "wide" (amplo) é:

```
reshape wide v4718-v4720 lnv4718-lnv4720, ///
i(mulher branca) j(ano)
```

Neste novo formato, podemos estimar correlações entre as variáveis de rendimento nos dois anos em análise.

Alocação de informações entre registros

Vamos supor que queremos colocar algumas respostas da pessoa de referência do domicílio (v0401=1) no registro dos outros residentes no mesmo domicílio (2<=v0401<=8). A idéia é de que essa pessoa de referência teria relação de influência com os demais habitantes no domicílio, principalmente com os filhos (v0401=3).

Esse tipo de exercício é muito utilizado em estudos que alocam informações dos pais (renda, educação e outros) para as crianças no mesmo domicílio.

Primeiramente, vamos criar um banco menor com variáveis que nos interessam da pessoa de referência na unidade domiciliar (v0401=1):

```
use "C:\curso\dados\pes2007MG.dta", clear
keep v0102 v0103 v0301 v0302 v8005 v0401 v0404 v4803 v4718-v4720 v4838
keep if v0401==1
```

Agora vamos renomear as variáveis do chefe do domicílio:

```
foreach x of varlist v0301-v4838 {
  rename `x' chef`x'
}
```

Ordenando e salvando o banco:

```
sort v0102 v0103
save "C:\curso\dados\pes2007MGchefe.dta", replace
```

Agora vamos abrir o banco original novamente e juntá-lo (merge) ao banco dos chefes dos domicílios:

```
use "C:\curso\dados\pes2007MG.dta", clear
sort v0102 v0103
merge v0102 v0103 using "C:\curso\dados\pes2007MGchefe.dta"
```

Vamos verificar se o "merge" ocorreu corretamente:

```
tab _merge
drop _merge
```

Salvando o banco com dados alocados:

```
save "C:\curso\dados\pes2007MGaloc.dta", replace
```

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

Lembre de ter o banco com dados alocados aberto:

```
use "C:\curso\dados\pes2007MGaloc.dta", clear
```

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":

```
*Grupos de anos de estudo (v4838)
replace v4838=. if v4838==7
replace chefv4838=. if chefv4838==7
```

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".

Vamos testar se há correlação significativa no nível de 95% entre a escolaridade da pessoa de referência do domicílio (chefv4838) e a escolaridade do filho (v4838, quando v0401 igual a 3):

```
tabulate v4838 chefv4838 if v0401==3, chi2
```

Observamos que há correlação significativa entre a escolaridade dos filhos e do chefe do domicílio.

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.

Recodificando a variável de anos de estudo:

```
gen estudo=.
replace estudo=v4803-1 if v4803>=1 & v4803<=16
```

Mesmo sabendo que essa variável não tem uma distribuição normal, vamos testar se a média de anos de estudo da população difere significantemente de 7,5, considerando uma escala que varia de 0 a 15:

```
histogram estudo

ttest estudo=7.5
```

A média da variável "estudo" é 5,895, com probabilidade estatisticamente significante de ser diferente de 7,5 e menor que 7,5.

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 se cor/raça (v0404) está correlacionada com anos de estudo (estudo). O correto seria transformar a variável cor/raça em variáveis dicotômicas:

```
use "C:\curso\dados\pes2007MG.dta", clear
gen estudo=.
replace estudo=v4803-1 if v4803>=1 & v4803<=16
tab v0404, gen(cor)
```

Em seguida é realizada o teste de correlação, em que é também possível estimar a significância:

```
pwcorr estudo cor?, sig
```

Também podemos usar as opções "xi" e "i." para calcular a correlação entre "estudo" e cor/raça rapidamente. Nesse caso, uma categoria da variável cor/raça (v0404) é omitida:

```
xi: pwcorr estudo i.v0404, sig
```

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 anos de estudo e brancos (v0404=2):

```
display (0.1437*0.1437)*100
```

O resultado indica que a variável anos de estudo tem uma variação de apenas 2,06% com o indicador de cor/raça branca.

De uma forma geral, as correlações entre anos de estudo e cor/raça são pequenas, conforme estimado pelo comando "pwcorr". Em outras palavras, por não estarem correlacionadas, essas variáveis poderiam ser inseridas conjuntamente em um modelo de regressão como variáveis independentes.

Os anos de estudo estão fortemente correlacionados com o logaritmo do rendimento no trabalho principal (lnv4718), o qual poderia ser a variável dependente do modelo:

```
replace v4718=. if v4718==999999999999
gen lnv4718=ln(v4718)

pwcorr lnv4718 estudo, sig

display (0.4282*0.4282)*100
```

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 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, 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
```

Passos para realização de inferência estatística

A intenção 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.
– Regressão estatística (continuação...).