Universidade de São Paulo / Faculdade de Filosofia, Letras e Ciências Humanas

Departamento de Ciência Política

FLP-468 & FLS-6183

2º semestre / 2016

Laboratório 8 – Missing Data

O objetivo desse lab é mostrar como usar "Amelia II: A Program for Missing Data". Como o próprio nome diz, o Amelia visa amenizar o problema dos dados *missing* (variáveis faltantes). O procedimento padrão para dados *missing* é a exclusão de toda a linha do banco (*listwise deletion*). Imagine um banco com 20 variáveis (ou seja, 20 colunas). Se uma das 20 variáveis está faltando, então a linha inteira é ignorada pelos softwares de regressão. O que acontece, então, com as 19 colunas de dados presentes dessa linha? Elas são ignoradas, descartadas. Imagine se todas as informações incompletas fossem descartadas pelos paleontólogos e arqueólogos. Ainda saberíamos muito pouco sobre a evolução natural ou sobre a cultura dos povos antigos.

O Amelia utiliza os dados existentes para estimar os dados que não estão no banco. Isso é feito através de uma técnica chamada de "EMB" (*Expectation Maximization bootstraping*). A ideia básica é que o dado ausente é correlacionado com outros dados presentes. Então, ao invés de ignorar todos os casos que possuem alguma variável faltante, usaremos as variáveis existentes para preencher os "furos" do banco.

O Amélia pode ser usado como um pacote do R ou através de um GUI (Graphic User Interface), o Amelia View, para aqueles que não possuem familiaridade com o R. Apenas para não termos que entrar em detalhes da linguagem do R, vamos utilizar o GUI Amelia View.

Para instalar o Amelia View, primeiro temos que instalar o R. Caso o R não esteja instalado em seu computador, vá até o endereço abaixo, baixe e instale o programa.

https://vps.fmvz.usp.br/CRAN/

Depois vá até o link abaixo e instale o Amelia View.

https://dl.dropbox.com/u/5379343/amelia-setup.exe

Caso não for possível utilizar o programa acima, é possível abrir o Amelia View através dessas três linhas de código no R:

```
install.packages("Amelia")
library(Amelia)
AmeliaView()
```

Agora já podemos utilizar o programa. Antes de iniciarmos nossos exercícios, tente abrir e explorar o software. Se familiarize com os menus e botões disponíveis.

Neste lab, iremos usar novamente o banco de dados de Cox e Amorin Neto sobre a teoria de Duverger, mas em uma versão sem as variáveis endógenas.

Exercício 1. Abra o banco de dados "Coxappend_clean.dta".

a) rode uma regressão com o seguinte modelo:

$$Legislative\ Parties = \beta_0 + \beta_1 Log(District\ Magnitude) + \beta_2 Number\ of\ Etnic\ Groups + \beta_3 Log(District\ Magnitude) * Number\ of\ Etnic\ Groups$$

```
. reg enps c.lnml##c.eneth
```

Source	SS	df	MS		
Model	39.7248839	3	13.241628	Number of obs	= 54
Residual	69.7444016	50	1.39488803	F(3, 50)	= 9.49
				Prob > F	= 0.0000
Total	109.469285	53	2.06545822	R-squared	= 0.3629
				Adj R-squared	= 0.3247
				Root MSE	= 1.1811

enps	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
lnml	-.1911175	.2967357	-0.64	0.522	-.7871287	.4048938
eneth	-.3619712	.3486305	-1.04	0.304	-1.062216	.3382737
c.lnml#c.eneth	.4833255	.1805094	2.68	0.010	.1207617	.8458893
_cons	2.671367	.6072149	4.40	0.000	1.45174	3.890994

b) Usando o commando *margins*, encontre os efeitos marginais de $Log(District\ Magnitude)$ e de $Number\ of\ Etnic\ Groups$.

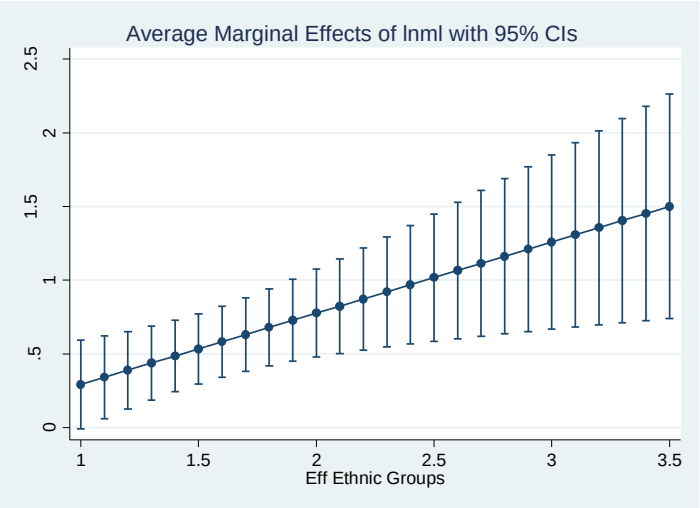
. quietly margins, dydx(lnml) at(eneth=(1(.1)3.5))

. marginsplot

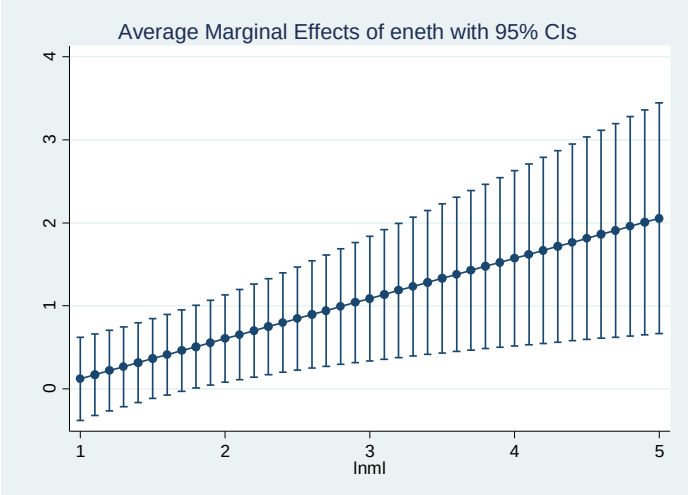
. quietly margins, dydx(eneth) at(lnml=(1(.1)5))

. marginsplot

Efeitos marginais da magnitude do distrito em diferentes valores de heterogeneidade social



Efeitos marginais da heterogeneidade social em diferentes valores de magnitude do distrito



Salve os resultados desse exercício em um documento separado, eles serão usados posteriormente.

Exercício 2. No mesmo banco, delete aleatoriamente um quarto (25%) das observações da variável "enps". O do-file "Deleta aleatoriamente.do" contém um método rápido para fazer isso. Salve o novo banco com o nome "Coxappend_new.dta". Rode novamente os dados do exercício anterior. Quais diferenças houveram em relação ao exercício 1?

```
. reg enps c.lnml##c.eneth
```

Source	SS	df	MS
Model	34.9369445	3	11.6456482
Residual	52.2502162	36	1.4513949
Total	87.1871607	39	2.23556822

Number of obs = 40
F(3, 36) = 8.02
Prob > F = 0.0003
R-squared = 0.4007
Adj R-squared = 0.3508
Root MSE = 1.2047

enps	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
lnml	-.2785821	.3195512	-0.87	0.389	-.926662	.3694978
eneth	-.5547909	.3718237	-1.49	0.144	-1.308884	.1993025
c.lnml#c.eneth	.5341922	.1918992	2.78	0.009	.1450025	.9233819
_cons	3.038288	.6662292	4.56	0.000	1.687113	4.389464

As principais diferenças são: o n passou a ser 40, sendo que antes era 54. Os coeficientes dos modelos praticamente não se alteraram. Para uma comparação mais sistemática, podemos avaliar se a diferencia dos coeficientes obtidos são estatisticamente significativas. Podemos fazer isso através de um teste t. Abaixo está o valor para a variável lnml:

$$t = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{(n_1-1){s_1}^2+(n_2-1)s_2^2}{n_1+n_2-2}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$$

Onde:

$\overline{x_1}$ = -.1911175

$\overline{x_2}$= -.2785821

${s_1}^2$ = (.2967357)^2 = .08805208

${s_2}^2$ = (.3195512)^2 = .10211297

n_1= 54

n_2= 40

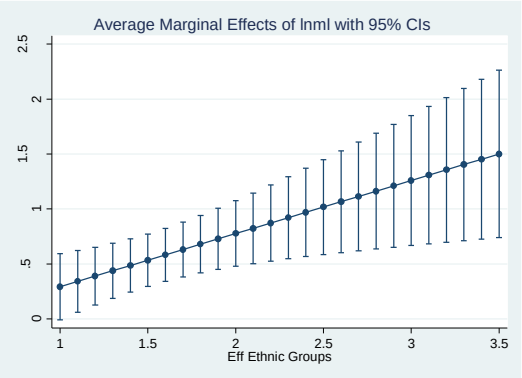
$$t = \frac{-.1911175 - (-.2785821)}{\sqrt{\frac{(54-1)08805208 +(40-1).10211297}{54+40-2}\left(\frac{1}{54}+\frac{1}{40}\right)}}$$

$$t = -0{,}13$$

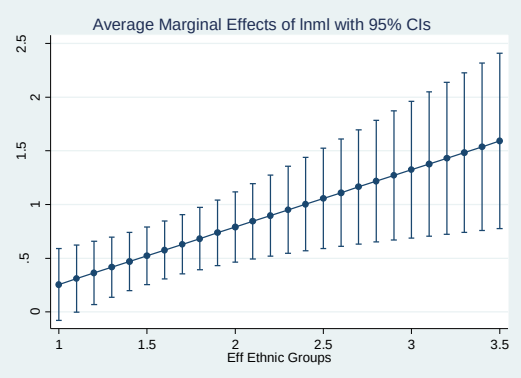
O valor do teste t é muito baixo para rejeitar a hipótese nula de diferença igual a zero. O mesmo ocorre com a variável eneth (t = -0,21) e com o termo interativo (t = 0,349). Por outro lado, a variância dos coeficientes aumentaram no banco com missing, tornando as conclusões menos precisas. Podemos verificar isso olhando para os efeitos marginais:

Efeitos marginais da magnitude do distrito em diferentes valores de heterogeneidade social

Original

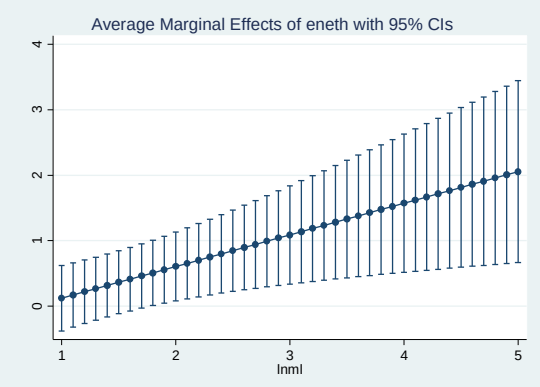


Missing

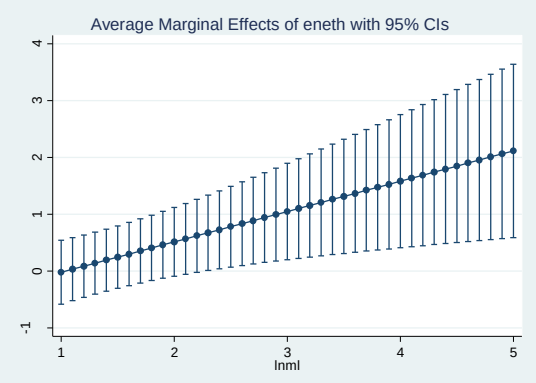


Efeitos marginais da heterogeneidade social em diferentes valores de magnitude do distrito

Original



Missing



Podemos ver que a inclinação das curvas de efeito marginal pouco mudou em relação ao modelo original, contudo, como os intervalos de confiança aumentaram a variável passou a ser estatisticamente significativa em momentos diferentes. No caso de lnml (log da magnitude do distrito), no modelo original, a variável possui influência sobre o número de partidos quando eneth é maior que cerca de 1,1 (justamente o valor mínimo da variável eneth), mostrando que a magnitude dos distritos possui influência sobre o número de partidos em qualquer sociedade. A mesma conclusão não pode ser tirada no banco com missing, pois lnml passa a ser significativo somente quando eneth é maior que cerca de 1,2. Um analista que dispõe apenas da amostra com

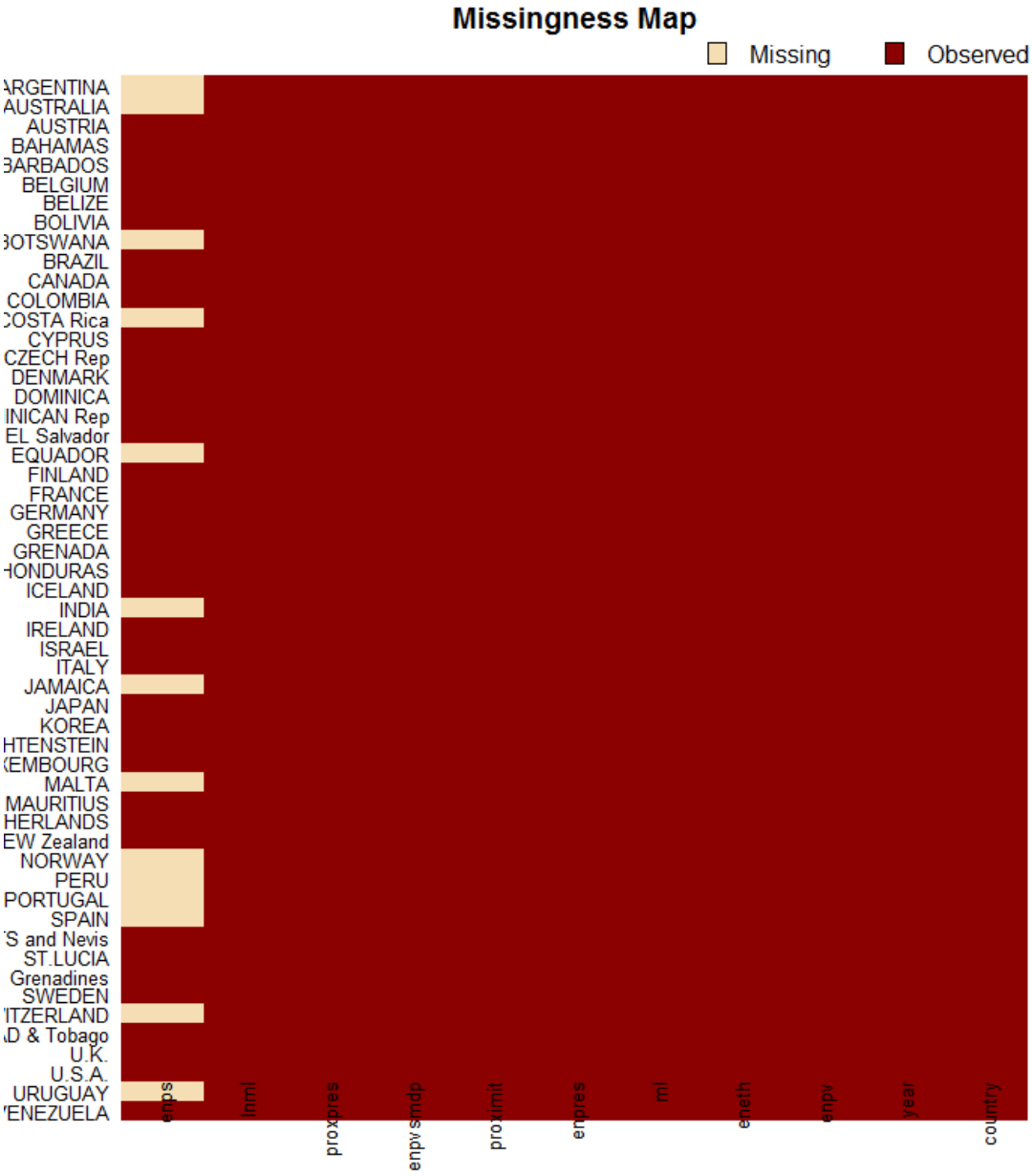
os missings pode concluir (equivocadamente) que em sociedades muito homogêneas, a magnitude do distrito não influencia o número de partidos legislativos.

No caso de eneth, no banco original ela influencia o número de partidos quando a lnml é igual a 1,7 (o equivalente a 6,05 cadeiras). Já no banco com missing, esse valor sobe para 2,2; o equivalente a nove cadeiras. Uma mudança marginal.

Exercício 3. Agora iremos explorar os recursos do Amelia. Entre no Amélia e importe o banco "Coxappend_new".

Agora já estamos prontos para imputar observações nos campos *missing* no banco. O Amelia irá usar todos os campos do banco de dados para fazer a estimação. Contudo alguns campos são especiais. Dois campos são particularmente importantes: o campo que mostra o país analisado (*unit variable*, *id variable* ou *cross-section variable*) e o ano de análise (*time series variable*). Atribua o campo "*country*" como *cross-section variable* e o campo "year" como *time series variable*.

Mostre um *missingness map* dos dados.



Exercício 4. Peça para o Amelia imputar valores nos campos *missing*.

Se você for no diretório onde estão os dados, você irá ver que o Amelia criou cinco novos bancos de dados. Isso ocorre porque não podemos simplesmente tratar os dados imputados da mesma forma que os dados observados. A estimação produz um grau de incerteza maior do que os dados observados (que também possuem um grau de incerteza, deve-se lembrar). Para demonstrar a incerteza adicional que a imputação gera, o Amelia utiliza uma estratégia chamada de "*multiple inputation*", ou MI. Basicamente, são gerados vários bancos, com vários valores possíveis das variáveis *missing*. Voltaremos a isso abaixo.

Por enquanto, importe um único banco de dados gerado pelo Amelia no Stata. Compare esses dados com o banco de dados original, "Coxappend_clean.dta". Você acha que o Amelia fez um bom trabalho na estimação? Justifique sua resposta.

	enps	enpsIm~d
1.	2.3714	3.270904
2.	2.3838	2.342601
9.	1.347	1.257595
13.	2.2086	2.166253
20.	5.7849	8.361856
28.	1.69	3.272684
32.	1.6	1.816764
37.	1.9996	2.221955
41.	3.0939	3.241733
42.	2.3182	2.293425
43.	3.409	3.597975
44.	2.8083	3.09099
49.	5.2625	5.166435
53.	3.3484	2.633486

Na lista acima, estamos comparando as observações que foram transformadas em missing para permitir imputação com os valores verdadeiros no banco original (enps). Novamente, fica difícil avaliar os resultados da imputação e precisamos fazer uma análise mais detalhada.

Exercício 5. Rode a regressão do **exercício 1** novamente. Quais diferenças você pode observar em relação aos resultados do **exercício 1**? E em relação ao **exercício 2**?

```
. do "C:\Users\dcplab\AppData\Local\Temp\STD01000000.tmp"

. reg enpsImputed c.lnml##c.eneth
```

Source	SS	df	MS
Model	40.8079473	3	13.6026491
Residual	101.385727	50	2.02771454
Total	142.193674	53	2.68289951

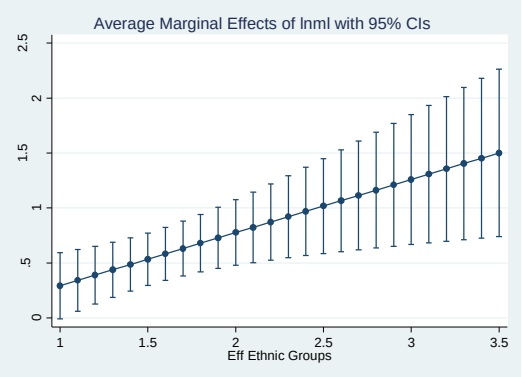
```
Number of obs =      54
F(  3,    50) =    6.71
Prob > F      =  0.0007
R-squared     =  0.2870
Adj R-squared =  0.2442
Root MSE      =   1.424
```

enpsImputed	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
lnml	-.19585	.3577697	-0.55	0.587	-.9144516	.5227517
eneth	-.1665786	.4203384	-0.40	0.694	-1.010853	.6776959
c.lnml#c.eneth	.4764755	.2176374	2.19	0.033	.0393379	.9136131
_cons	2.514064	.7321097	3.43	0.001	1.043578	3.984549

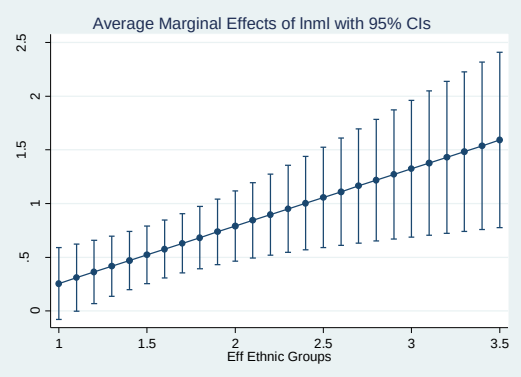
Efeitos marginais da magnitude do distrito em diferentes valores de heterogeneidade social

Efeitos marginais da magnitude do distrito em diferentes valores de heterogeneidade social

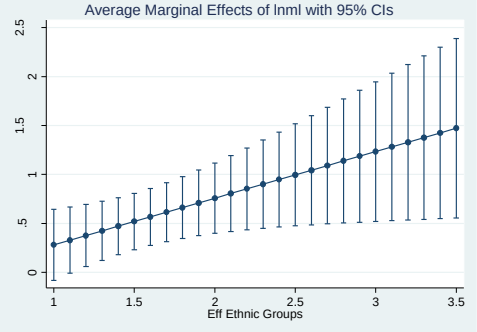
Original



Missing



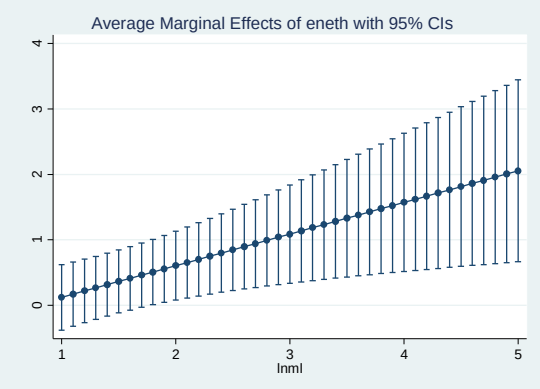
Imputed 1



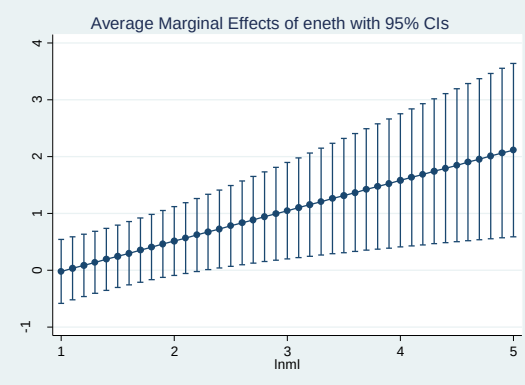
Podemos ver que nessa amostra aleatória gerada pelo Amelia, a inclinação não se alterou em relação ao padrão original e ao banco de dados missing. Quanto à variância, não houve melhoras significativas em relação ao modelo missing. Temos que considerar que esse resultado é em grande parte resultado do modo aleatório que criamos os dados missing. Caso os missing tivessem um padrão (os países pobres tivessem chance maior de missing, por exemplo), então haveria diferenças mais claras entre os modelos originais, missing e com imputação. Também temos que considerar que o Amélia utiliza modelos aleatórios (bootstrap) para gerar os dados, então, existe uma chance de a imputação não produzir boas estimativas. Mas esse fator é aliviado pelo fato de que o Amelia utiliza múltiplas imputações e se trabalha com uma média (ver mais sobre isso abaixo).

Efeitos marginais da heterogeneidade social em diferentes valores de magnitude do distrito

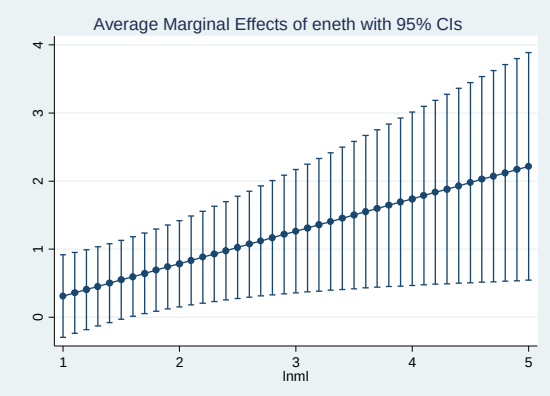
Original



Missing



Imputed 1



Exercício 6. Apenas utilizar um dos bancos de dados gerados pelo Amelia não é o recomendado pela estratégia de *multiple imputation*. O recomendado é rodar a o mesmo modelo em todos os bancos gerados e fazer uma média dos resultados.

Então, mão na massa! Rode o modelo do exercício 1 nos cinco bancos gerados pelo Amelia. **Guarde os valores dos parâmetros e dos erros padrões do modelo, eles serão usados abaixo**. O uso do Ctrl+C Ctrl+V no do file do Stata está liberado, mas aqueles que fizerem isso usando um loop ganharão *extra credit*.

Veja o script com no do-file. Repare que o STATA permite criar um vetor (array) com os nomes dos bancos a serem importados e rodar uma mesma rotina em cada um deles. Podemos salvar os valores dos coeficientes criando uma matriz de coeficientes e uma matriz de variâncias dos bancos. Isso é possível porque o comando "regress" guarda os valores principais do modelo (coeficientes, matriz de variância-covariância, etc) na memória do computador. Abaixo estão as matrizes com os valores dos coeficientes e da variância dos coeficientes nos cinco modelos dos

```
. * Mostra as matrizes criadas
. matrix list coefs // matrix dos coeficients beta

coefs[5,4]
                                c.lnml#
          lnml       eneth     c.eneth       _cons
y1  -.19584997  -.16657864   .47647549   2.5140637
y1  -.19420308  -.22134815    .4640263   2.6394058
y1  -.15286247  -.20765331    .4435338   2.5840254
y1  -.17586134  -.15827511   .45043721   2.5023335
y1  -.18907845  -.20334146   .46445653   2.5325167

. matrix list Vars // matrix da variância dos coeficients beta

Vars[5,4]
                             c.lnml#
         lnml      eneth    c.eneth      _cons
r1  .12799918  .17668438  .04736605  .53598454
r1  .11975537  .16530499  .04431543  .50146435
r1  .12894198  .17798578  .04771493  .53993243
r1  .13739114  .18964863  .05084154  .57531249
r1   .1217524   .1680616  .04505443  .50982672
```

Exercício 7. Agora, podemos calcular os coeficientes dos parâmetros usando a estratégia MI. Calcule a média simples dos coeficientes beta dos cinco modelos. Para aqueles que não lembram da formula da média, ela está abaixo.

$$\bar{q} = \frac{1}{m}\sum_{j=1}^{m} q_j.$$

Onde,

$\bar{q}$ é o valor médio dos parâmetros.

m é o número de bancos de dados gerados pelo Amélia.

q_j é o beta da base de dados *j*.

No do-file em anexo mostramos como calcular a média dos coeficientes utilizando a notação matricial do STATA.

```
.
. * Aqui temos a média dos coeficientes
. matrix list MeanCoefs

MeanCoefs[1,4]
                                c.lnml#
          lnml       eneth     c.eneth       _cons
c1  -.18157106  -.19143933   .45978587    2.554469
```

Exercício 8. Segundo Honaker, King & Blackwell (2011), podemos calcular a variância dos parâmetros na estratégia de MI através das médias das variâncias dos parâmetros em cada modelo mais a variância amostral dos parâmetros entre os bancos de dados, da seguinte forma:

$$SE(q)^2 = \frac{1}{m}\sum_{j=1}^{m} SE(q_j)^2 + S_q^2(1 + 1/m).$$

Onde,

$SE(q)^2$ é a variância média do parâmetro q entre todos as bases de dados.

$SE(q_j)^2$ é a variância do parâmetro q no banco de dados j.

S_q^2 é a variância amostral do parâmetro q entre todos os bancos de dados, ou seja

$$S_q^2 = \sum_{j=1}^{m}(\frac{(q_j - \bar{q})^2}{m-1})$$

Calcule a variância, $SE(q)^2$, e o erro padrão, $SE(q)$, de todos os parâmetros rodados no exercício 6.

No do-file em anexo mostramos como calcular os erros padrão dos coeficientes utilizando a notação matricial do stata.

```
. * Erro padrão dos betas de multiple imputation
. scalar list SEImputation1 SEImputation2 SEImputation3
>  SEImputation4
SEImputation1 =   .3571031
SEImputation2 =   .4199704
SEImputation3 =  .21735812
SEImputation4 =   .7321838
```

Exercício 9. Quais são as principais diferenças entre os resultados encontrados nos **exercícios 7 e 8** e os resultados dos **exercícios 1** e **2**?

Para facilitar a visualização dos dados, iremos postar as regressões com os dados completos, com os dados missing e com os dados gerados pela Amelia (na média dos cinco bancos).

Banco Completo

```
. reg enps c.lnml##c.eneth
```

Source	SS	df	MS		
Model	39.7248839	3	13.241628	Number of obs	54
Residual	69.7444016	50	1.39488803	F(3, 50)	9.49
				Prob > F	0.0000
				R-squared	0.3629
				Adj R-squared	0.3247
Total	109.469285	53	2.06545822	Root MSE	1.1811

enps	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
lnml	-.1911175	.2967357	-0.64	0.522	-.7871287	.4048938
eneth	-.3619712	.3486305	-1.04	0.304	-1.062216	.3382737
c.lnml#c.eneth	.4833255	.1805094	2.68	0.010	.1207617	.8458893
_cons	2.671367	.6072149	4.40	0.000	1.45174	3.890994

Banco Missing

```
. reg enps c.lnml##c.eneth
```

Source	SS	df	MS		
Model	34.9369445	3	11.6456482	Number of obs	40
Residual	52.2502162	36	1.4513949	F(3, 36)	8.02
				Prob > F	0.0003
				R-squared	0.4007
				Adj R-squared	0.3508
Total	87.1871607	39	2.23556822	Root MSE	1.2047

enps	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
lnml	-.2785821	.3195512	-0.87	0.389	-.926662	.3694978
eneth	-.5547909	.3718237	-1.49	0.144	-1.308884	.1993025
c.lnml#c.eneth	.5341922	.1918992	2.78	0.009	.1450025	.9233819
_cons	3.038288	.6662292	4.56	0.000	1.687113	4.389464

Banco com imputação Amelia (usando o pacote MI)

```
. miest imp regress enpsImputed lnml eneth lnml_eneth

Multiple Imputation Estimates

Model: regress
Dependent Variable: enpsImputed

Number of Observations: 54
------------------------------------------------------------------
         |      Coef.   Std. Err.       t          Df     P>|t|
------------------------------------------------------------------
    lnml |  -.18157     .3571428     -0.508     443686    0.611
   eneth |  -.19144     .4200502     -0.456     152009    0.649
lnml_eneth |  .45979    .2173924      2.115     221217    0.034
   _cons |   2.5545     .7323798      3.488      76597    0.000
------------------------------------------------------------------
```

Não é correto fazer interpretações do tipo "a variância dos parâmetros é menor no modelo com missing, portanto ele é melhor". Variância baixa pode ser resultado de vieses de seleção na amostra, tornando mais provável o erro do tipo I (rejeitar a hipótese nula quando ela é verdadeira). O correto é colocar o modelo com dados completos como "gold standard" e comparar o modelo imputado em relação à ele. Podemos ver que as estimativas pontuais se assemelham muito no modelo com dados Amelia, mais ainda que nos dados com missing. Provavelmente isso se deve à presença de uma variável muito correlacionada com a variável com dados missing (a variável "número de partidos eleitorais"). Além disso, conforme visto no curso, erros de mensuração na variável dependente tendem a provocar vieses menores que erros de mensuração nas variáveis independentes. Quanto aos erros padrão, o modelo com Amelia possui erros maiores, em parte por causa da incerteza adicional decorrente da estratégia de MI. Contudo, ao observar os dados com missing, podemos ver que nossa exclusão aleatória retirou casos onde o modelo não funcionava muito bem. Assim, é natural que a variância dos parâmetros seja menor no modelo com missing. Mas é importante lembrar que queremos modelos que funcionem em qualquer amostra e não modelos que funcionem bem em uma amostra específica. Podemos pensar, por analogia, na seguinte situação: vamos supor que um laboratório descubra que um remédio não cura a doença em 90% dos pacientes selecionados aleatóriamente, mas que esse mesmo laboratório tenha excluído esses casos, mostrando um modelo que se encaixa perfeitamente aos dados, com ótimos coeficientes entre as variáveis "dose do remédio" e "saúde do paciente". Você tomaria esse remédio?

Podemos concluir que o Amelia teve bom desempenho em prever os valores das variáveis, ainda mais porque não inserimos muitas variáveis adicionais para ajudar o programa a prever os valores missing, como os autores do programa recomendam. Podemos ver que a estratégia MI se mostrou uma estratégia prudente, aumentando os valores dos erros padrão, diminuindo a chance de cometermos o erro do tipo I devido à imputação dos dados.

Exercício 10 (extra-credit). Refaça o exercício 9 e 10 utilizando o pacote "MI" do Stata, criado por Ken Scheve. Ele pode ser baixado no link abaixo:

http://gking.harvard.edu/files/gking/files/mi.zip

Segue abaixo o output do pacote "MI":

```
. miest imp regress enpsImputed lnml eneth lnml_eneth

Multiple Imputation Estimates

Model: regress
Dependent Variable: enpsImputed

Number of Observations: 54
------------------------------------------------------------------
         |      Coef.   Std. Err.       t          Df     P>|t|
------------------------------------------------------------------
    lnml |  -.18157     .3571428     -0.508     443686    0.611
   eneth |  -.19144     .4200502     -0.456     152009    0.649
lnml_eneth |  .45979    .2173924      2.115     221217    0.034
   _cons |   2.5545     .7323798      3.488      76597    0.000
------------------------------------------------------------------
```

Em primeiro lugar, devemos observar que os valores são praticamente os mesmos dos cálculos que fizemos acima, mostrando que eles estavam corretos. O pacote MI facilita muito a implementação do Amelia no STATA (uma vez que o Amelia foi criado em ambiente R). Contudo, aprender o passo-a-passo da técnica de multiple imputation ajuda tanto a reforçar o princípios do sistema como a ajuda o pesquisador a fazer adaptações do processo, de acordo com seus interesses de pesquisa.