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:

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

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

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?

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.

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**?

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*.

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*.

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.

$SE\left(q_j\right)^2$ 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

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

Calcule a variância, $SE(q)^2$, e o erro padrão, $SE(q)$, de todos os parâmetros rodados no exercício 9.

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

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