Seminário de Investigação 2014-2015 (MEPP)

Utilização de software estatístico como suporte ao trabalho empírico

Carlos Farinha Rodrigues

100 1911-2011 ANOS
100 ANOS A PENSAR NO FUTURO

Objectivo:

- Discutir como lidar com a parte empírica da tese de mestrado e as potencialidades do software estatístico na elaboração da tese.

Introdução:

- O trabalho empírico é subsidiário de uma 'ideia', da resposta a uma questão que é o centro da tese.
- Não fazemos trabalho empírico para fazer trabalho empírico.

Introdução:

- O trabalho empírico serve para comprovar uma ideia, para responder a uma questão concreta.
  - O aumento das pensões mínimas ajuda a reduzir a pobreza.
  - Uma certa campanha publicitária permite ganhar novos clientes para um produto.

A questão dos dados:

- Qual o tipo de dados ideal para o problema que eu quero resolver.
  - Para analisar o problema das pensões mínimas basta-me os dados administrativos do sistema de pensões ou preciso de algo mais complexo que relacione estas com o conjunto dos rendimentos das famílias.

A questão dos dados:

- Os dados de suporte existem ou tenho que os construir ?
  - Obtenção de dados publicados.
  - Obtenção de bases de dados.
  - Realização de um inquérito.

A questão dos dados:

- Dados Publicados *versus* bases de dados.
- Três etapas no tratamento dos dados:

1. Obtenção dos dados.
2. Organização dos dados.
3. Modelização dos dados.

Slide 7

A questão dos dados:

- Qualquer uma destas etapas exige tempo e trabalho.
- Necessidade de resolver a questão numa fase preliminar da dissertação.
- Obter dados já existentes pode levar tempo: contratos / protocolos / tempo de espera...

Slide 8

A questão dos dados:

- Geralmente o acesso a bases de dados estatísticos envolve a apresentação de um projecto validade pela Universidade.

Slide 9

Regras de conduta quanto à utilização dos dados:

- Existe um conjunto de questões éticas associadas à utilização de dados.
- Respeito dos objectivos estabelecidos no contrato de acesso.

Slide 10

Regras de conduta quanto à utilização dos dados:

- Respeito dos objectivos estabelecidos no contrato de acesso.

1. Limitar o uso de dados ao objectivo solicitado .
2. Mencionar sempre quem cedeu os dados (e a versão com que se está a trabalhar) .
3. Não utilizar os dados para fins comerciais ou outros não estabelecidos.

Slide 11

Regras de conduta quanto à utilização dos dados:

- Respeito dos objectivos estabelecidos no contrato de acesso.

4. Respeitar as regras de confidencialidade e de anonimização.
5. Destruição dos dados no fim do período estabelecido.

Slide 12

Regras de conduta quanto à realização do trabalho empírico:

- Os resultados obtidos têm de poder ser verificados pela comunidade científica.
- Os resultados apresentados têm de poder ser duplicados por outros investigadores.

Regras de conduta quanto à realização do trabalho empírico:

- As diferentes hipóteses assumidas quanto ao tratamento dos dados tem de ser apresentadas e justificadas.
  - ✓ Registo diário e pormenorizado do trabalho empírico.
  - ✓ Hipóteses assumidas e hipóteses abandonadas.

Regras de conduta quanto à realização do trabalho empírico:

- Por exemplo a base de dados das famílias utilizada para o estudo das pensões evidenciava vários pensionistas com pensões inferiores aos mínimos legais e foram corrigidos.
- Tratamento de missings/quebras de série, etc.
- Porque restringir a análise ao Continente.

Escolha de software:

- A escolha de software depende obviamente do trabalho a desenvolver e da estrutura da informação estatística que vão utilizar.

Escolha de software:

- Três níveis:
  1. Excel.
  2. SPSS, Stata, SAS, TSP,...
  3. Gauss.

➢ Para uma discussão dos níveis de popularidade dos vários programas consulte:

The Popularity of Data Analysis Software by Robert A. Muenchen
(http://r4stats.com/popularity)

Trabalhar em spss (ou stata, ou sas ...)

- Todos estes programas tem um interface com o utilizador geralmente assente num sistema de menus, numa worksheet onde são apresentados os dados e uma janela de output.

Trabalhar em spss (ou stata, ou sas …)

- Todos estes programas tem igualmente um outro interface que permite ao utilizador escrever e introduzir rotinas, submete-las e obter os resultados.
- Muitos dos comandos mais potentes e/ou menos usuais somente podem ser passados para o programa por esta via.

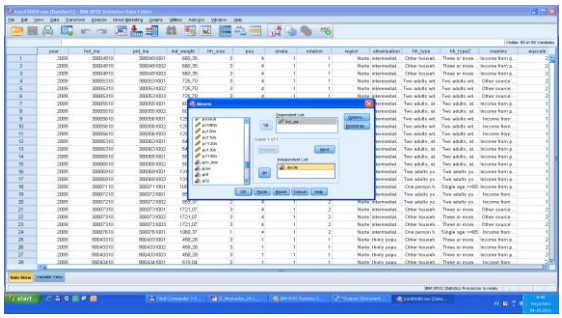
Slide 19

Trabalhar em spss (ou stata, ou sas …)

- Geralmente estes programas têm um sistema de interacção entre o sistema de menus e a lógica de programação de rotinas.
- É possível utilizar o sistema de menus para escrever parte das rotinas a utilizar.

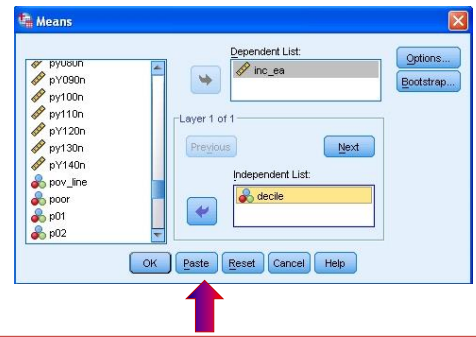
Slide 20

Trabalhar em spss (ou stata, ou sas …)



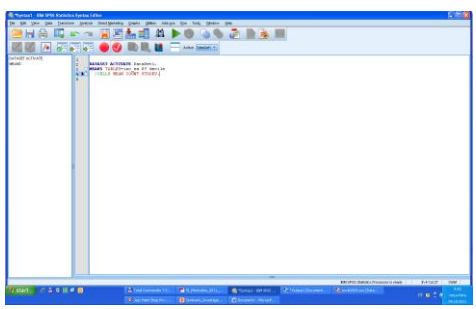
Slide 21

Trabalhar em spss (ou stata, ou sas …)



Slide 22

Trabalhar em spss (ou stata, ou sas …)



Slide 23

Trabalhar em spss (ou stata, ou sas …)

```
SET printback=listing messages=listing.
Title  *****          rsi_model_09#001          ****.

* ********************************************************************************.
* rsi_model_09#001:                                                              *.
*   Simulation of RSI based on SILC 2009                                         *.
*   Builds the individual datafile from silc files                               *.
* ********************************************************************************.
* @cfr2011 - version 24-09-2011                                                  *.
* ********************************************************************************.


DATASET CLOSE ALL.
GET FILE='c:\temp\icor2009r.sav'/KEEP hid_ine pid_ine rb010 rb030 rb050 rb080 rb090 rb220 rb230 rb240.
DATASET NAME DataSet1 WINDOW=FRONT.

IF (rb080 ne rb010)age=(rb010-1)-rb080.
IF (rb080 eq rb010)age=0.
FORMATS age (f3.0).
VARIABLE LABELS age 'Age at the end of the income reference period'.
EXECUTE.

Rename vars (rb010 rb030 rb050 rb090 rb220 rb230 rb240 = year pid ind_weight sex pid_father pid_mother pid_partner).
execute.

compute hid = trunc(pid/100).
variable label hid 'Household ID'.
execute.
```

Slide 24

Trabalhar em spss (ou stata, ou sas …)

```
* ********************************************************************************************.
* excludes persons born in 2009 and recompute the weights                                     *.
* ********************************************************************************************.

AGGREGATE  /OUTFILE=* MODE=ADDVARIABLES /BREAK=year /ind_weight_sum=SUM(ind_weight).

SELECT IF (age > 0).
EXECUTE.

AGGREGATE
  /OUTFILE=* MODE=ADDVARIABLES
  /BREAK=year
  /ind_weight_sum2=SUM(ind_weight).

COMPUTE  ind_weight = ind_weight  * (ind_weight_sum / ind_weight_sum2).
EXECUTE.
```


Trabalhar em stata (ou spss, ou sas …)

```
*********************************************************************************
* PROJECT:          EUROMODupdate: construct a EUROMOD database from SILC database
* DO-FILE NAME:     00_MAIN.do
* TEMPLATE VERSION: 07-100219
* DESCRIPTION:      Main do-file to set the main parameters (country, paths) and  call sub-scripts
*********************************************************************************
* COUNTRY:          Portugal
* SILC VERSION:     EU-SILC 2007
* NATIONAL MODELLERS:  Carlos Farinha
* LAST UPDATE:      18/09/2010
*********************************************************************************

version 9 // this is to use Stata 9
clear
set logtype smcl
set more off
set mem 200m

*********************************************************************************
*
global country = "PT"

display in y "Country selected: ${country}"
```


Trabalhar em stata (ou spss, ou sas …)

```
*********************************************************************************
* Define the SILC year, i.e. when was collected
* (will be used in the name of output files)
*********************************************************************************

global silcyr = 2007
global silc_UDByr = "UDB_c07"
global silc_ver = "_ver 2007-3 from 01-03-10"

*********************************************************************************
* Define EUROMOD database source, i.e. x in CC_year_x# (eg. uk_2006_a1)
* (will be used in the name of the final output file)
*********************************************************************************

global data_source = "a"                                        // [TO DO]!

*********************************************************************************
* Define EUROMOD database version, i.e. # in CC_year_x# (eg. uk_2006_a1)
*********************************************************************************

global data_ver = 1                                             // [TO DO]!

global path "C:\EuromodPT"

* folder where original EU-SILC data are stored and output files *will be* stored
global path_data "${path}\Data"
```


Vantagens da utilização de "ficheiros de comandos"

- **Porquê utilizar ficheiros de sintaxe em SPSS (*.sps) , ficheiros DO em Stata (*.do) , etc.**

Vantagens da utilização de "ficheiros de comandos"

- Uma vez aprendida a linguagem de programação poupa tempo. É mais simples alterar um elemento no programa e mandar executá-lo que repetir toda a sequência do sistema de menus.
- O programa permite perceber a lógica de tratamento dos dados que foi seguida e a modelização efectuada. O(s) programa(s) permite(m) identificar a estratégia implementada para a parte empírica do trabalho.

Vantagens da utilização de "ficheiros de comandos"

- O mesmo programa pode ser aplicado/adaptado a outros projectos de investigação ou a outras bases de dados.
- Desde que devidamente comentados os ficheiros de comandos possibilitam que outros investigadores (ou o próprio algum tempo depois…) entendam as opções seguidas, validem os resultados obtidos e eventualmente repliquem os mesmos procedimentos em outros projectos.

Não somos os primeiros a utilizar um dado software estatístico para fazer investigação…

- Raramente o programa escolhido vem com aquilo que pretendemos pré-definido.
- É possível encontrar muitos procedimentos, rotinas e outros recursos passiveis de serem utilizados no nosso trabalho empírico '*on-line*'.

Slide 31

Não somos os primeiros a utilizar um dado software estatístico para fazer investigação…
…. alguns '*sites*' a explorar.

- 'Sites' Oficiais.
  - www.spss.com
  - www.stata.com
  - www.sas.com

Slide 32

Não somos os primeiros a utilizar um dado software estatístico para fazer investigação…
…. alguns '*sites*' a explorar.

- Outros '*Sites*' (SPSS).
  - http://www.fpce.uc.pt/niips/spss_prc/
  - http://www.spsstools.net/
  - http://listserv.uga.edu/archives/spssx-l.html

Slide 33

Não somos os primeiros a utilizar um dado software estatístico para fazer investigação…
…. alguns '*sites*' a explorar.

- Outros '*Sites*' (STATA).
  - http://www.stata.com/links/resources1.html
  - http://www.stata.com/statalist/
  - http://www.stata.com/statalist/archive/

Slide 34

Seminário de Investigação 2014-2015 (MEPP)

Utilização de software estatístico como suporte ao trabalho empírico

Carlos Farinha Rodrigues

100 1911-2011 ANOS
100 ANOS A PENSAR NO FUTURO