

Seminário de Investigação

Tema C:

Bases de Dados Quantitativas, utilização de Software Estatístico e questões éticas da utilização da informação na elaboração do TFM

Vítor Escária
(+ Paulo Madruga e Carlos Farinha)

OUT 2017

Objectivo da Sessão

O acesso e utilização de informação quantitativa, as potencialidades dos pacotes de software e questões éticas na utilização de informação quantitativa quando se faz trabalho empírico em economia ...

... a partir das nossas experiências práticas

Tópicos

- Trabalho empírico em economia
- Obtenção/ preparação dos dados
- Fontes de informação
- Pacotes de software
- Questões éticas

Trabalho Empírico

- Definição de ***objecto de interesse*** – do modelo teórico, da literatura
- Definição de hipóteses a testar – a ***tese***
- ***Obtenção e preparação*** de dados
 - Recolha e sistematização de dados
 - Instrumentos e métodos de medida
- ***Teste das hipóteses*** – confirmação - Modelação e análise de dados
 - análises univariadas, bivariadas, multivariadas
- **Apresentação dos resultados**

Trabalho Empírico

- **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**
- **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 ação do banco central reduz os custos de financiamento das empresas?**

Obtenção/preparação dos dados

- **Sem dados, o trabalho empírico não existe.**
- **A natureza dos dados existentes determina os instrumentos e métodos de medida, bem como as técnicas de modelação e análise a usar...**
- **... para testar a hipótese teórica (falsificação, apoio).**

Obtenção/preparação dos dados

- **Informação secundária - produzida por entidades do Sistema Estatístico/ outras entidades**
- **Informação primária - recolha directa**
 - Inquéritos
 - Estudos de caso

Obtenção dos dados : Fontes de informação estatística secundária mais comuns

- **Entidades sistema Estatístico**
 - **Nacionais:**
 - INE, Banco de Portugal, DGO, etc...
 - **Internacionais**
 - Eurostat, OCDE, Comissão Europeia, Nações Unidas, Banco Mundial, FMI, etc...
- **Outras fontes**
 - **Datastream, Bloomberg, Reuters, Dun & Bradstreet**

Dados de fontes estatísticas

- **Alguns cuidados a ter:**
 - **Metainformação:**
 - informação acerca da construção e especificações das estruturas;
 - descrição dos pormenores concretos
 - **Nomenclaturas**
 - convenções para denominações e regras
 - **Problemas/exemplos**
 - Fontes diferentes para a mesma variável
 - Valores originais vs valores normalizados
 - Comparações internacionais
 - Quebras de séries

Obtenção dos dados

Recolha directa/Inquéritos

- **Alguns cuidados a ter:**
 - seleção da população a inquirir e da amostra
 - metodologia de inquirição
 - realização de um teste ao inquérito e de uma simulação de apuramentos
 - **Alguns problemas:**
 - tipos de questões: abertas vs fechadas
 - codificação de respostas
 - questões fechadas: escalas de medida
 - custos monetários e em termos de tempo da realização

Tipos de dados

- **Perfil de observação das unidades e dimensão temporal**
 - seccionais (*cross section*) – várias unidades de observação (e.g. indivíduos, empresas, países) “fotografados” num determinado momento (ou período);
 - temporais (*time series*) – a mesma unidade “filmada” ao longo de vários períodos ou momentos (e.g. anos, trimestres, meses ou até segundos nas finanças);
 - longitudinais ou em painel (*panel*) – cruzam os dois tipos anteriores:
 - “painéis largos” – muitas unidades de observação e poucas observações temporais;
 - “painéis profundos” - poucas unidades de observação e muitas observações temporais.

Tipos de dados

- **Nível de agregação:**
 - **agregados** – combina valores de diversas unidades estatísticas (e.g. dados macro);
 - **microdados** – os valores das variáveis respeitam à observação da unidade estatística individual.

Modelação e análise de dados

Tipos de análises

- **Estática – num mesmo período; podem ser comparadas várias unidades estatísticas (usa dados seccionais).**
- **Estática comparada – compara situações para a mesma unidade estatística em dois (ou três) “momentos” diferentes (usa dados seccionais e/ou painéis pouco profundos).**
- **Dinâmica – ao longo do tempo:**
 - para dados agregados (usa séries temporais);
 - para microdados (usa painéis profundos).

Bases de dados

- **Acesso Livre**
 - Portal INE
 - Portal Banco de Portugal (BPSTAT, etc)
 - Portal Eurostat
 - Portal OCDE
 - Comissão Europeia: DG ECFIN (AMECO, KLEMS, BACH, ...)
 - OMC (Comercio internacional)
 - FMI, Banco Mundial, etc
 - CMVM /Euronext

Bases de dados

- **Disponíveis terminais ISEG ou usando *proxy* ISEG**
 - Amadeus
 - Datastream/Reuters
 - BANKSCOPE – informação sobre 23,000 bancos
 - CHELEM – dados comércio internacional , agregados macroeconómicos e balança pagamentos
 - OSIRIS – informação sobre empresas cotadas

Bases de dados

- **Disponíveis mediante protocolo**
 - **Protocolo ex-MCTES/INE – acesso microdados**

NOTA:

Obter dados já existentes pode levar tempo: contratos / protocolos / tempo de espera...=> Necessidade de resolver a questão numa fase preliminar da dissertação.

Software

- **Escolha de software**

- depende do trabalho a desenvolver e da estrutura da informação estatística que a utilizar
- Três níveis:
 - Excel
 - SPSS, Stata, SAS, TSP, Eviews, R...
 - Gauss, MATLAB

Para uma discussão dos níveis de popularidade dos vários programas consultar: The Popularity of Data Analysis Software by Robert A. Muenchen (<http://r4stats.com/popularity>)

Software de econometria e estatística

- **Trabalhar com SPSS/STATA/SAS**
 - Todos tem um interface com o utilizador geralmente assente num sistema de menus, numa folha de dados, onde são apresentados os dados, e uma janela de output
 - Todos 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
 - Geralmente 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

Software de econometria e estatística

The screenshot displays the IBM SPSS Statistics Data Editor interface. The main window shows a data table with 28 rows and 15 columns. The columns are: year, hid_in, pid_in, ind_weight, hh_size, psu, strata, rotation, region, urbanisation, hh_type, hh_type2, maininc, and eqscale. The data table is partially obscured by a 'Means' dialog box.

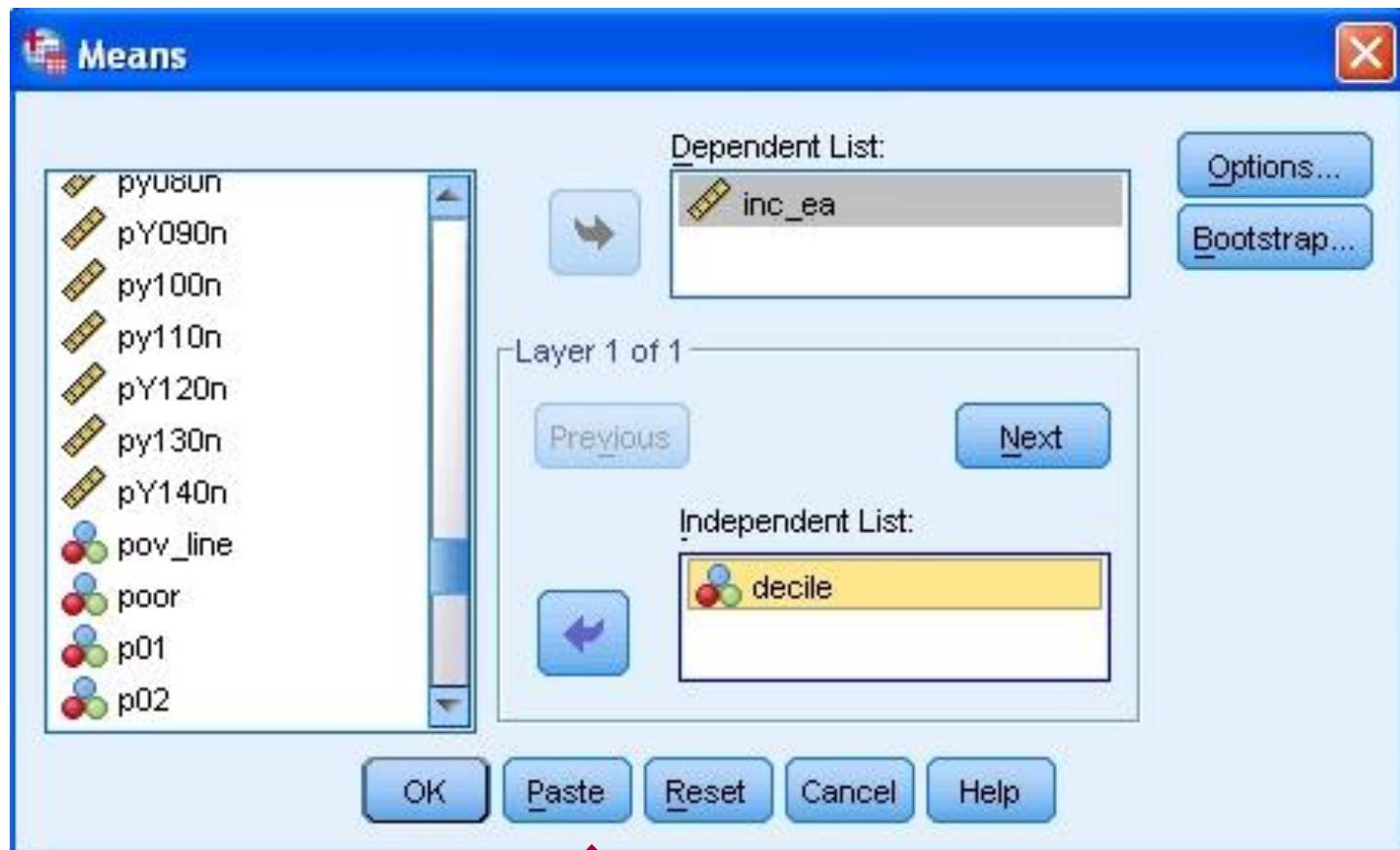
The 'Means' dialog box is open, showing the following configuration:

- Dependent List: inc_ea
- Independent List: declle
- Layer 1 of 1
- Buttons: Previous, Next, OK, Paste, Reset, Cancel, Help, Options..., Bootstrap...

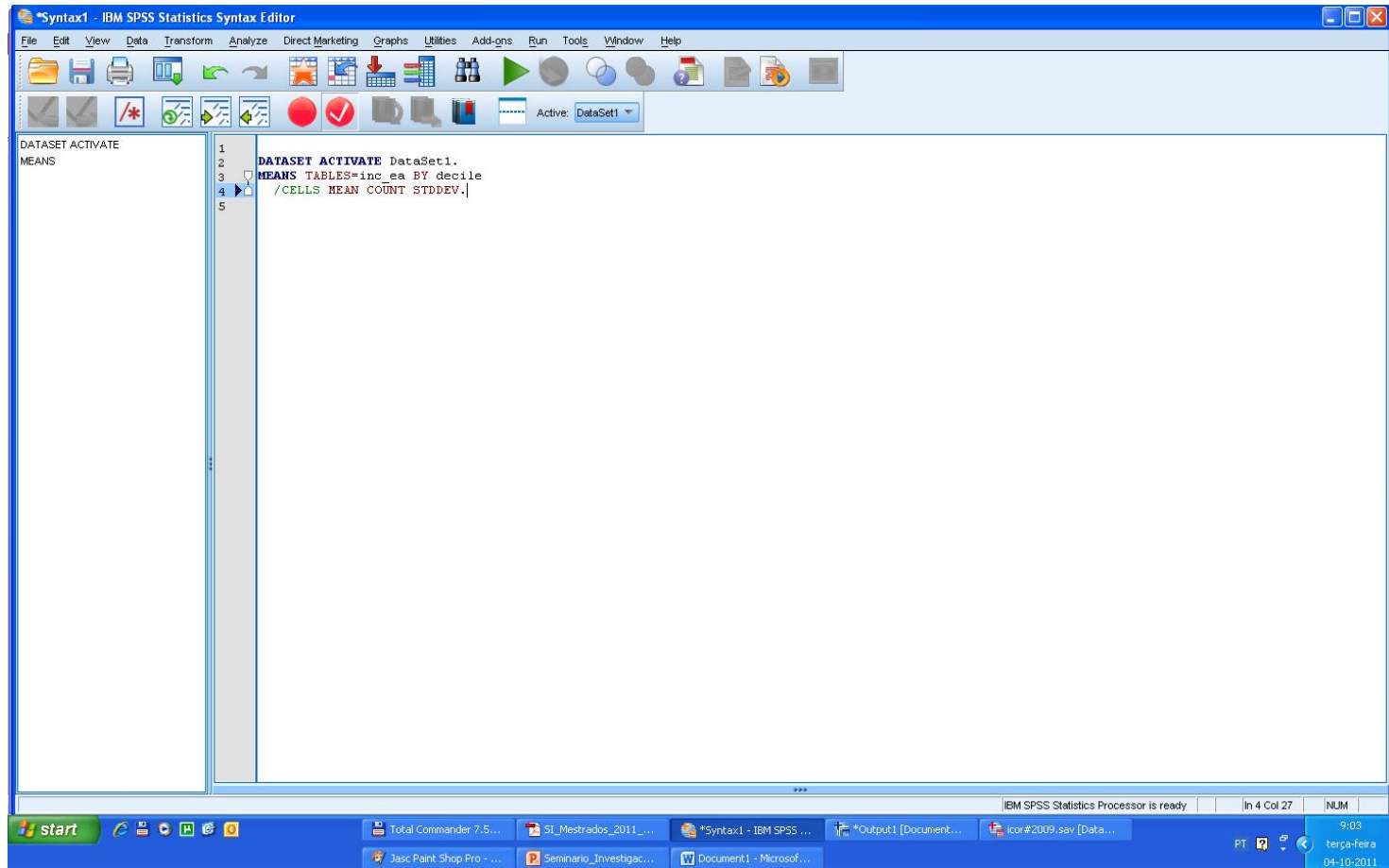
The data table shows the following data for the first 10 rows:

year	hid_in	pid_in	ind_weight	hh_size	psu	strata	rotation	region	urbanisation	hh_type	hh_type2	maininc	eqscale	
1	2009	30004910	3000491001	660,35	3	4	1	1	Norte	intermediat...	Other househ...	Three or more...	Income from p...	2
2	2009	30004910	3000491002	660,35	3	4	1	1	Norte	intermediat...	Other househ...	Three or more...	Income from p...	2
3	2009	30004910	3000491003	660,35	3	4	1	1	Norte	intermediat...	Other househ...	Three or more...	Income from p...	2
4	2009	30005310	3000531001	725,70	3	4	1	1	Norte	intermediat...	Two adults wit...	Two adults wit...	Other source ...	2
5	2009	30005310	3000531002	725,70	3	4	1	1	Norte	intermediat...	Two adults wit...	Two adults wit...	Other source ...	2
6	2009	30005310	3000531003	725,70	3	4	1	1	Norte	intermediat...	Two adults wit...	Two adults wit...	Other source ...	2
7	2009	30005510	3000551001	60	60	4	1	1	Norte	intermediat...	Two adults, at...	Two adults, at...	Income from p...	1
8	2009	30005510	3000551002	60	60	4	1	1	Norte	intermediat...	Two adults, at...	Two adults, at...	Income from p...	1
9	2009	30005610	3000561001	125	125	4	1	1	Norte	intermediat...	Two adults wit...	Two adults wit...	Income from ...	1
10	2009	30005610	3000561002	125	125	4	1	1	Norte	intermediat...	Two adults wit...	Two adults wit...	Income from ...	1

Software de econometria e estatística



Software de econometria e estatística



Software de econometria e estatística

```
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.
```

Software de econometria e estatística

- **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.
 - O mesmo programa pode ser aplicado/adaptado a outros projectos de investigação ou a outras bases de dados.

NOTA:É possível encontrar muitos procedimentos, rotinas e outros recursos passíveis de serem utilizados no nosso trabalho empírico

‘on-line’

Software de simulação e manipulação simbólica

- **Manipulação simbólica:** Serve para ajudar nos cálculos matemáticos sem ser necessário ter números (ao contrário das folhas de cálculo e dos programas de simulação).
 - Maple
 - Mathematica
- **Simulação:** Serve para simular modelos numéricos que exigem uma matemática complicada e para a qual não é, em geral, possível obter uma solução explícita.
 - Matlab.
 - Gauss.

Questões éticas

Algumas regras de conduta a observar na dissertação e no trabalho de investigação:

- Propriedade intelectual:
 - Plágio
 - Pirataria de *software* proprietário (para quê, com tanto *open source*?).
 - Não referenciação das fontes de dados ou *software* utilizados.
- Verificabilidade e replicabilidade dos resultados:
 - De preferência, as bases de dados e os códigos devem estar em acesso livre (atenção à propriedade intelectual).
 - Quando tal não é possível (e.g. confidencialidade), têm de estar para os membros do júri (*referees* para os artigos).
 - As notas metodológicas e cálculos intermédios também (e.g. Reinhart & Rogoff).

Questões éticas

- **Regras de conduta quanto à utilização de dados:**
 - **Limitar o uso de dados ao objectivo solicitado**
 - **Mencionar sempre quem cedeu os dados (e a versão com que se está a trabalhar)**
 - **Não utilizar os dados para fins comerciais ou outros não estabelecidos**
 - **Respeitar as regras de confidencialidade e de anonimização**
 - **Destruição dos dados no fim do período estabelecido**