Applications of Data Analysis (EC969)

Simonetta Longhi and Alita Nandi

ISER, University of Essex

Week 1 Lecture 1

Structure of Each Class

- Part I: Discussion of specific economic/survey problems, and how to analyse them
- Part II: Hands-on exercise (estimations etc. using Stata; some parts of the exercises are optional)
- Part III: Discussion of the empirical results of the analysis, interpretation, etc.

Structure of the Course

- Week 1: Stata, datasets, and basic data management. Cross-sectional regression. Prepare the dataset which we will use for the rest of the course.
- Week 2: survey sample design, non response issues, selection issues
- Week 3: wage regressions for panel data; applications to marriage and to unemployment scarring
- Week 4: models for limited dependent variables in panel data; applications to unemployment scarring and unemployment persistence
- Week 5: event history analysis

We will assume basic knowledge of econometric techniques
(consult books in background reading list)

Structure of the Exam

Assessment

Whichever is the greater:
EITHER: 50% course work mark, 50% exam mark,
OR: 100% exam mark

Coursework: One term paper

Exam duration and period

2 hour exam during summer examination period

Term Paper

- Spring term paper: now available on the course website
- Deadline: 6th May, 12 noon
- Hand one paper copy with cover sheet to Claire Cox **and** submit one copy online (via MyEssex page)
- Cover sheet available online (http://www.essex.ac.uk/economics/documents/cover_yr2,3,pgt.pdf) or in the holder outside Claire's office

Data Used in this Course

In this course we use the
British Household Panel Survey (BHPS)

→ Repeated observations on individuals over time

Available from data archive
(need to sign the UKDA form)

Panel data: allows
cross section, panel and event history analysis

Other Micro Panel Datasets

- US Panel Survey of Income Dynamics (PSID); US National Longitudinal Survey (NLS) of youth (NLSY); US Current Population Survey (CPS)
- British Cohort Survey 1970; Millennium Cohort Survey (2000); English Longitudinal Study of Ageing (ELSA)
- German Socio-Economic Panel (GSOEP);
  Swiss Household Panel (SHP);
  Household, Income and Labour Dynamics in Australia (HILDA);
  Survey of Labour & Income Dynamics (SLID) Canada

Other Micro Panel Datasets

- Multi-countries
  - European Community Household Panel (ECHP) 1994-2001
  - European Survey of Income and Living Conditions (EU-SILC) – fixed term panel
  - Cross-National Equivalent File (CNEF): PSID + GSOEP + BHPS + HILDA + SLID
- All different settings (i.e. households vs. individuals, periodicity, …)
- Keeping Track: http://www.iser.essex.ac.uk/keeptrack/index.php

Macro (Pseudo) Panel Datasets

- Repeated observations of macro/aggregated variables (e.g. unemployment, inflation rate, PPP, …) across countries or regions
- The cross-sectional component is not individuals or households, but countries or regions
- Number of cross-sections is much lower than on individual datasets (where N>>T)
  → different types of econometric problems (e.g. spatial econometrics)
- Examples: World Development Indicators (WDI); World Economic Outlook Database; Regio (EU) Database, …

British Household Panel Survey

- BHPS
- Started 1991 with a sample of 5,000 households (10,000 adults) in Great Britain;
  1999: Additional samples of 1,500x2 households from Scotland and Wales;
  2001: Additional Sample of 2,000 households from Northern Ireland
- Interviews with all adult members (aged 16+) of households + adult and children self-completion questionnaires
- Individuals are interviewed annually; most of the questions are repeated annually

Cross-section Data

Single cross section

person id	year	wage	age
1001	2005	5.2	45
1002	2005	6.9	53
1003	2005	na	21
1004	2005	4.1	21

Pooled cross sections

person id	year	wage	age
1001	2005	5.2	45
1002	2005	6.9	53
1003	2005	na	21
1004	2005	4.1	21
1005	2006	na	71
1006	2006	5.0	34
1007	2006	4.9	31

Panel Data

Balanced panel

person id	year	wage	age
1001	2005	5.2	45
1002	2005	6.9	53
1003	2005	na	21
1004	2005	4.1	21
1001	2006	5.2	46
1002	2006	7.0	54
1003	2006	4.9	22
1004	2006	na	22
1001	2007	5.3	47
1002	2007	7.0	55
1003	2007	4.9	23
1004	2007	3.9	23

Unbalanced panel

person id	year	wage	age
1001	2005	5.2	45
1002	2005	6.9	53
1004	2005	4.1	21
1001	2006	5.2	46
1002	2006	7.0	54
1003	2006	4.9	22
1004	2006	na	22
1001	2007	5.3	47
1003	2007	4.9	23
1004	2007	3.9	23

Types of Longitudinal Data

- Administrative data: collected for administrative purposes, limited information, but accurate (not a sample) e.g. benefit data
- Surveys: collected for research purposes, a sample of persons ('panel') is followed over time more information, but with error
  - Retrospective: cheaper, faster, recollection errors
  - Prospective: more expensive, data collected from a sequence of interviews ('waves')

Prospective Surveys

- Fixed life (‘rotating panel’): people are interviewed only a certain number of times (Labour Force Survey)
- Indefinite life: no scheduled end (BHPS)

BHPS

- The BHPS is an indefinite life panel survey without sample replacement by drawing of new samples
- People are ‘followed’ and re-interviewed annually (face-to-face + self-completion questionnaires)
- BHPS is mostly a prospective survey, with some retrospective elements
  (between waves job and employment histories; lifetime employment and job histories; fertility and marital histories; ...)
- Will be matched to administrative data

BHPS Following Rules

- ‘Following rules’ specify who should be eligible to be interviewed at each wave
- Required to maintain representativeness of original population and their descendants
- The BHPS sample consists of:
  - Original Sample Members (OSM): members of original (1991 or 1999-2001 for the booster samples) households, and their natural descendants born since the start of the panel
  - Temporary Sample Members (TSM): at each wave, the current co-residents of OSMs are also eligible for interview
  - Permanent Sample Members (PSM): TSMs who had a child with an OSM
- OSMs and PSMs are eligible for interview each wave so long as they remain in scope, while TSMs are not followed if they no longer live with a sample member

Development of the Sample over Time

- Sample is reduced by:
  - Attrition: refusal and non-contact
  - Becoming ineligible: deaths and moves out of scope (e.g. abroad)
- Sample is increased by:
  - New births
  - Other new (temporary) entrants
  - Additional samples (Scotland and Wales wave 9; Northern Ireland wave 11)
- Balanced panel continuously reduces over time

We will discuss these issues next week

Structure of the BHPS Dataset

- Dataset is divided into a set of separate files: by wave and unity of analysis (individual adults, young people, households, etc.)
  → often when doing analysis we need to combine different data files
  (we will learn it this week)

During this course we will use only a sub-set of all files available; look at the documentation for more info on the dataset:
http://www.iser.essex.ac.uk/survey/bhps
BHPS courses: http://www.iser.essex.ac.uk/survey/bhps/courses

Naming Conventions

- Wave-specific variables and files have a wave specific prefix ‘w’ where w = a (wave 1), b (wave 2), etc.
- Roots of file and variable names are constant across waves (e.g. aindresp.dta, bindresp.dta, …; aage, bage, …)
- Not always the case (e.g. GSOEP)

More About BHPS

- BHPS online documentation:
  http://www.iser.essex.ac.uk/survey/bhps

- Two-days workshop ‘Introduction to BHPS using STATA’:
  http://www.iser.essex.ac.uk/survey/bhps/courses
  go to the bottom of the page to download the material

- Working paper describing the BHPS:
  http://www.longitudinal.stir.ac.uk/wp/lda_2006_2.pdf

Files Used in this Course

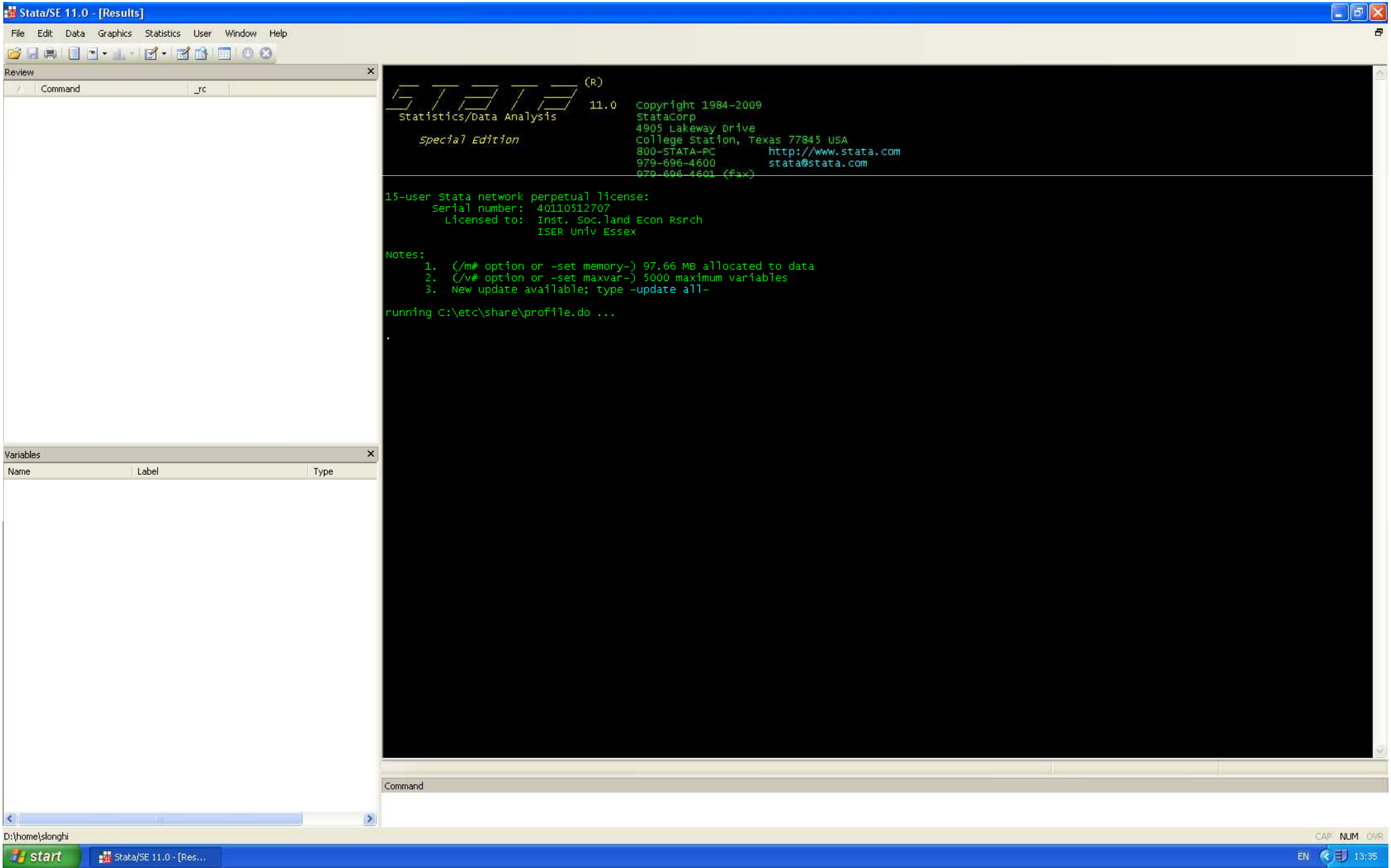
- wINDRESP: contains data for respondent adults
- wJOBHIST: contains data on wave-to-wave job histories
- XWAVEID: contains information on results of interview, useful for matching individuals between waves

- ExtraData: a file we have create specifically for this course

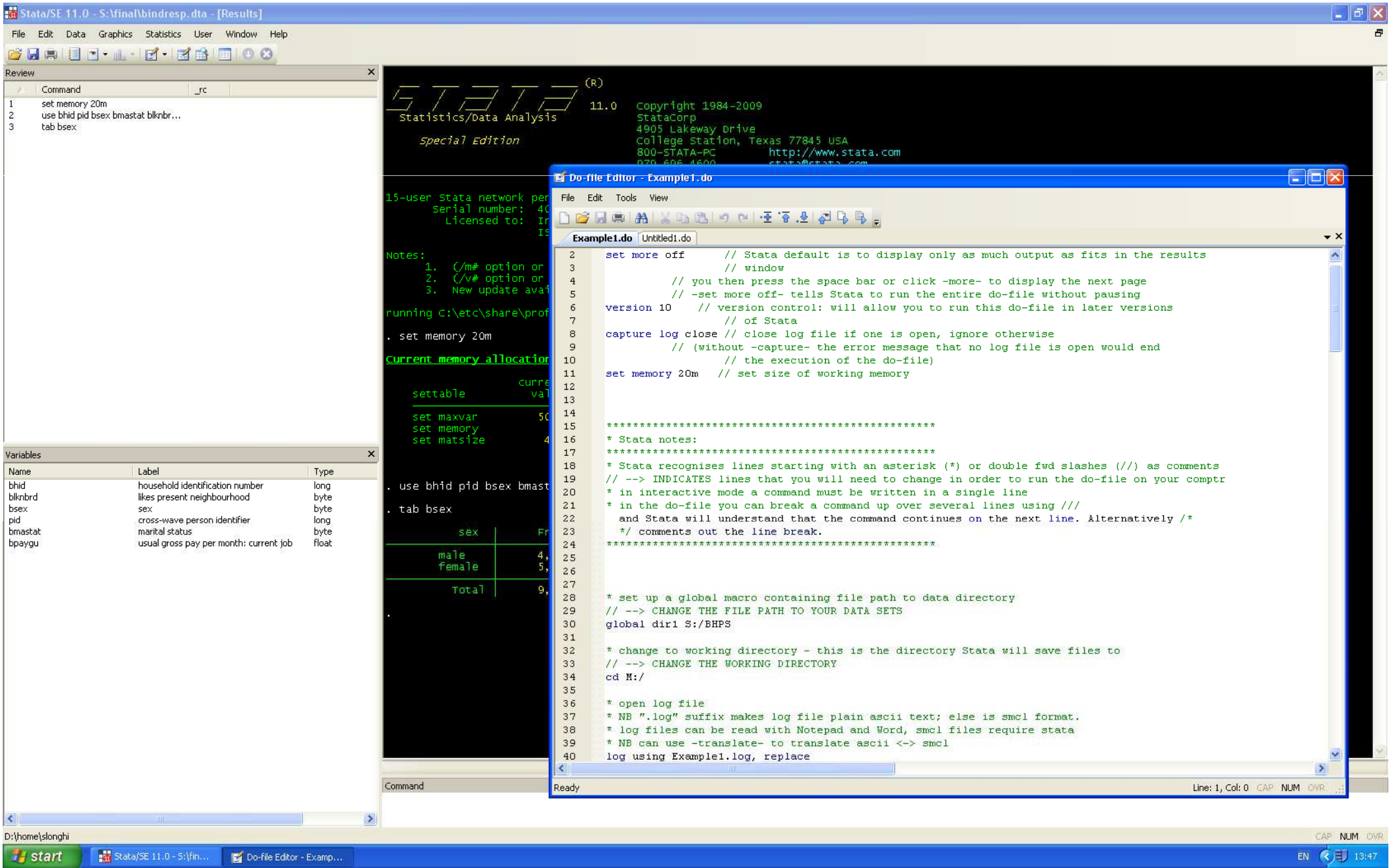
Hands-on Examples: Worksheets

- Worksheet 1: how does Stata look like
- Worksheet 2: how does the BHPS look like
  - Upload and inspect the data
  - Run cross-section regressions, test coefficient restrictions and save the results in a table

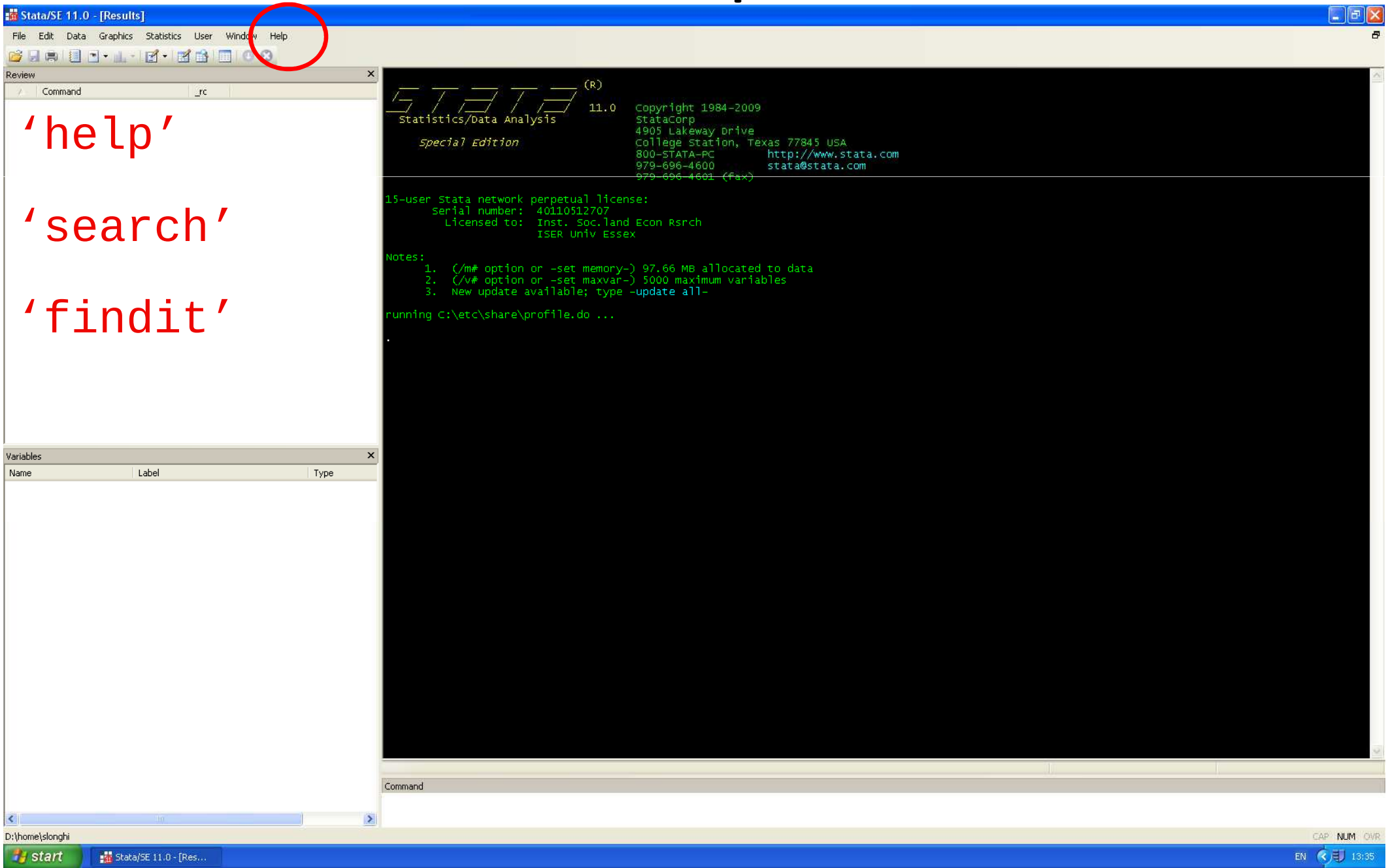
Stata Windows



‘.do’ Files


Stata/SE 11.0 - S:\final\bindresp.dta - [Results]
File Edit Data Graphics Statistics User Window Help
Review
Command _rc
1 set memory 20m
2 use bhid pid bsex bmastat blknbr…
3 tab bsex
Variables
Name Label Type
bhid household identification number long
blknbrd likes present neighbourhood byte
bsex sex byte
pid cross-wave person identifier long
bmastat marital status byte
bpaygu usual gross pay per month: current job float
Statistics/Data Analysis
Special Edition
11.0
Copyright 1984-2009
StataCorp
4905 Lakeway Drive
College Station, Texas 77845 USA
800-STATA-PC http://www.stata.com
Notes:
. set memory 20m
. use bhid pid bsex bmast
. tab bsex
sex
male 4,
female 5,
Total 9,
Do-file Editor - Example1.do
File Edit Tools View
Example1.do Untitled1.do
2 set more off // Stata default is to display only as much output as fits in the results
3 // window
4 // you then press the space bar or click -more- to display the next page
5 // -set more off- tells Stata to run the entire do-file without pausing
6 version 10 // version control: will allow you to run this do-file in later versions
7 // of Stata
8 capture log close // close log file if one is open, ignore otherwise
9 // (without -capture- the error message that no log file is open would end
10 // the execution of the do-file)
11 set memory 20m // set size of working memory
15 **
16 * Stata notes:
17 **
18 * Stata recognises lines starting with an asterisk (*) or double fwd slashes (//) as comments
19 // --> INDICATES lines that you will need to change in order to run the do-file on your comptr
20 * in interactive mode a command must be written in a single line
21 * in the do-file you can break a command up over several lines using ///
22 and Stata will understand that the command continues on the next line. Alternatively /*
23 */ comments out the line break.
24 **
28 * set up a global macro containing file path to data directory
29 // --> CHANGE THE FILE PATH TO YOUR DATA SETS
30 global dir1 S:/BHPS
32 * change to working directory - this is the directory Stata will save files to
33 // --> CHANGE THE WORKING DIRECTORY
34 cd M:/
36 * open log file
37 * NB ".log" suffix makes log file plain ascii text; else is smcl format.
38 * log files can be read with Notepad and Word, smcl files require stata
39 * NB can use -translate- to translate ascii <-> smcl
40 log using Example1.log, replace
Ready
Line: 1, Col: 0 CAP NUM OVR
Command
D:\home\slonghi
start
Stata/SE 11.0 - S:\fin…
Do-file Editor - Examp…
EN 13:47

Help


Stata/SE 11.0 - [Results]
File Edit Data Graphics Statistics User Window Help
Review
Command _rc
'help'
'search'
'findit'
Variables
Name Label Type
(R)
Statistics/Data Analysis
11.0
Special Edition
Copyright 1984-2009
StataCorp
4905 Lakeway Drive
College Station, Texas 77845 USA
800-STATA-PC http://www.stata.com
979-696-4600 stata@stata.com
979-696-4601 (fax)
15-user Stata network perpetual license:
Serial number: 40110512707
Licensed to: Inst. Soc.land Econ Rsrch
ISER Univ Essex
Notes:
1. (/m# option or -set memory-) 97.66 MB allocated to data
2. (/v# option or -set maxvar-) 5000 maximum variables
3. New update available; type -update all-
running C:\etc\share\profile.do ...
Command
D:\home\slonghi
CAP NUM OVR
start
Stata/SE 11.0 - [Res...
EN
13:35

Some Stata tips/good practice

Shall we use the drop-down menu, type the commands iteratively, or use do files?

The menus are good to start and have a feeling of how the command looks like (and for those commands with a lot of options, such as graphs),

but it is better to soon learn the structure of each command

Stata Commands

Most Stata commands have the following form:
command [optional qualifiers], [options]

Commands and options can usually be abbreviated, sometimes to one letter!
command, options
co, opt

Some commands can be used in different 'versions'
(e.g. the command **use**)

Housekeeping

- **`clear`** → clears the working directory from any data that might already be open (Note: in Stata11 use 'clear all')
- **`set memory 20m`** → set the size of the working memory to be able to open large files
- **`set more off`** → to run the entire do-file, without pausing after each page and displaying the **`more`** message
- **`version 11`** → specifies the version of Stata in which the do-file was written, so that you can later run this file with a different version of Stata

Reproducibility

```
log using Example1.log, replace
```

- Use log files to document what you are doing and to keep a record of all the models you have estimated. A log file saves all the results that appear on the screen.
- The extension **.log** means that the file should be written in ascii, which can be read in Word and other programmes
- At the beginning of the programme close the log file if one is already open **capture log close**
- You can **rewrite** (over the old file) or **append** the new results to the old ones

Portability

```
global dir1 "S:/BHPS"
```

- To make our do-files more portable (between computers or collaborators), it is a good idea to store the file path to the data directory in a global macro
- Throughout the do-file we then refer to the contents of this global macro. This means that if we change to a different computer, we will simply have to change the file path at the beginning of the do-file and not have to worry about the rest of the code.

Second Worksheet

- Uses data on respondents from the first wave of the BHPS (aindresp)
- Shows:
  - Different ways to upload the data
  - How to inspect the data
  - Recode values, create variables and labels
  - Delete variables and cases
  - Cross-section wage regression
  - Tests on regression coefficients
  - Saving results into output tables

Results

	Model1	Model2	Model3	Model4 Men	Model4 Women
Age	0.131***	0.131***	0.128***	0.166***	0.098***
	(0.005)	(0.006)	(0.006)	(0.008)	(0.009)
Women	-0.720***	-0.720***	-0.348***		
	(0.019)	(0.019)	(0.038)		
Married/cohab	0.077***	0.077***	0.359***	0.209***	-0.084**
	(0.024)	(0.024)	(0.029)	(0.027)	(0.036)
1st degree	0.805***	0.805***	0.787***	0.562***	1.016***
	(0.036)	(0.035)	(0.035)	(0.038)	(0.059)
hnd,hnc,teaching	0.690***	0.690***	0.672***	0.480***	0.880***
	(0.044)	(0.038)	(0.038)	(0.044)	(0.062)
a level	0.452***	0.452***	0.435***	0.330***	0.506***
	(0.031)	(0.030)	(0.030)	(0.033)	(0.054)
o level	0.271***	0.271***	0.259***	0.208***	0.296***
	(0.027)	(0.028)	(0.027)	(0.032)	(0.042)
cse	0.173***	0.173***	0.161***	0.225***	0.108
	(0.045)	(0.047)	(0.046)	(0.050)	(0.074)
Married woman			-0.534***		
			(0.044)		
…					
Observations	5145	5145	5145	2557	2588
Adj_R2	.3848995	.3848995	.4038982	.4530005	.1922694

Robust standard errors in parenthesis, (except Model1); * Significant at 10%, ** Significant at 5%, *** Significant at 1%

	Model1	Model2	Model3	Model4 Men	Model4 Women
Age	0.131***	0.131***	[illegible]28***	0.166***	0.098***
	(0.005)	(0.006)	[illegible]	(0.008)	(0.009)
Women	-0.720***	-0.720***	-0.348[illegible]		
	(0.019)	(0.019)	(0.038)		
Married/cohab	0.077***	0.077***	0.359***	0.2[illegible]	-0.084**
	(0.024)	(0.024)	(0.029)	(0.027)	(0.036)
1st degree	0.805***	0.805***	0.787***	0.562***	1.016***
	(0.036)	(0.035)	(0.035)	(0.038)	(0.059)
hnd,hnc,teaching	0.690***	0.690***	0.672***	0.480***	0.880***
	(0.044)	(0.038)	(0.038)	(0.044)	(0.062)
a level	0.452***	0.452***	0.435***	0.330***	0.506***
	(0.031)	(0.030)	(0.030)	(0.033)	(0.054)
o level	0.271***	0.271***	0.259***	0.208***	0.296***
	(0.027)	(0.028)	(0.027)	(0.032)	(0.042)
cse	0.173***	0.173***	0.161***	0.225***	0.108
	(0.045)	(0.047)	(0.046)	(0.050)	(0.074)
Married woman			-0.534***		
			(0.044)		
…					
Observations	5145	5145	5145	2557	2588
Adj_R2	.3848995	.3848995	.4038982	.4530005	.1922694

Using Robust Standard Errors

Robust standard errors in parenthesis, (except Model1); * Significant at 10%, ** Significant at 5%, *** Significant at 1%

	Model1	Model2	Model3	Model4 Men	Model4 Women
Age	0.131***	0.131***	***	0.166***	0.098***
	(0.005)	(0.006)		(0.008)	(0.009)
Women	-0.720***	-0.720***	-0.3		
	(0.019)	(0.019)	(0.038)		
Married/cohab	0.077***	0.077***	0.359***		-0.084**
	(0.024)	(0.024)	(0.029)	(0.0	(0.036)
1st degree	0.805***	0.805***	0.787***	0.562***	1.016***
	(0.036)	(0.035)	(0.035)	(0.038)	(0.059)
hnd,hnc,teaching	0.690***	0.690***	0.672***	0.480***	0.880***
	(0.044)	(0.038)	(0.038)	(0.044)	(0.062)
a level	0.452***	0.452***	0.435***	0.330***	0.506***
	(0.031)	(0.030)	(0.030)	(0.033)	(0.054)
o level	0.271***	0.271***	0.259***	0.208***	0.296***
	(0.027)	(0.028)	(0.027)	(0.032)	(0.042)
cse	0.173***	0.173***	0.161***	0.225***	0.108
	(0.045)	(0.047)	(0.046)	(0.050)	(0.074)
Married woman			-0.534***		
			(0.044)		
…					
Observations	5145	5145	5145	2557	2588
Adj_R2	.3848995	.3848995	.4038982	.4530005	.1922694

Are women paid as much as men?

Robust standard errors in parenthesis, (except Model1); * Significant at 10%, ** Significant at 5%, *** Significant at 1%

	Model1	el2	Model3	Model4 Men	Model4 Women
Age	0.1	::	0.128***	0.166***	0.098***
		.006)	(0.006)	(0.008)	(0.009)
Women		-0.720***	-0.348***		
		(0.019)	(0.038)		
Married/cohab	0.077***	0.077***	0.359***	0.209***	-0.084**
	(0.024)	(0.024)	(0.029)	(0.027)	(0.036)
1st degree	0.805***	0.805***	0.787***	0.562***	1.016***
	(0.036)	(0.035)	(0.035)	(0.038)	(0.059)
hnd,hnc,teaching	0.690***	0.690***	0.672***	0.480***	0.880***
	(0.044)	(0.038)	(0.038)	(0.044)	(0.062)
a level	0.452***	0.452***	0.435***	0.330***	0.506***
	(0.031)	(0.030)	(0.030)	(0.033)	(0.054)
o level	0.271***	0.271***	0.259***	0.208***	0.296***
	(0.027)	(0.028)	(0.027)	(0.032)	(0.042)
cse	0.173***	0.173***	0.161***	0.225***	0.108
	(0.045)	(0.047)	(0.046)	(0.050)	(0.074)
Married woman			-0.534***		
			(0.044)		
…					
Observations	5145	5145	5145	2557	2588
Adj_R2	.3848995	.3848995	.4038982	.4530005	.1922694

Does the impact of being married differ between men and women?

Robust standard errors in parenthesis, (except Model1); * Significant at 10%, ** Significant at 5%, *** Significant at 1%

	Model1	Model2	Model3	Model4 Men	Model4 Women
Age	0.131**	**	0.128***	0.166***	0.098***
	(		(0.006)	(0.008)	(0.009)
Women			-0.348***		
		.019)	(0.038)		
Married/cohab		0.077***	0.359***	0.209***	-0.084**
	-4)	(0.024)	(0.029)	(0.027)	(0.036)
1st degree	0.805***	0.805***	0.787***	0.562***	1.016***
	(0.036)	(0.035)	(0.035)	(0.038)	(0.059)
hnd,hnc,teaching	0.690***	0.690***	0.672***	0.480***	0.880***
	(0.044)	(0.038)	(0.038)	(0.044)	(0.062)
a level	0.452***	0.452***	0.435***	0.330***	0.506***
	(0.031)	(0.030)	(0.030)	(0.033)	(0.054)
o level	0.271***	0.271***	0.259***	0.208***	0.296***
	(0.027)	(0.028)	(0.027)	(0.032)	(0.042)
cse	0.173***	0.173***	0.161***	0.225***	0.108
	(0.045)	(0.047)	(0.046)	(0.050)	(0.074)
Married woman			-0.534***		
			(0.044)		
…					
Observations	5145	5145	5145	2557	2588
Adj_R2	.3848995	.3848995	.4038982	.4530005	.1922694

The literature suggests that returns to education are higher for women than men. Do we find the same?

Robust standard errors in parenthesis, (except Model1); * Significant at 10%, ** Significant at 5%, *** Significant at 1%

Next Time

- Focus on complex data management
- Combine data from 2 waves in wide and long format
- Generalise long format into multiple waves

- Expected output: the file obtained at the end of this class is the (final) dataset we will be working on for the rest of the course