University of Bamberg

Migration and Labor Market

STATA I:

Introduction to the Basics

Bamberg, May, 25, 2020

Dr. Ehsan Vallizadeh

Department of Economics – Integration of European Labor Markets

Content

1. What do we have to do for the paper
2. Brief introduction into the datasets
3. Introduction into STATA
   - STATA Software Package
   - STATA Basics: three files
   - Getting started
   - The STATA Menus
   - The Grammar of STATA
   - Working with Do-Files
   - Organizing your session
   - Processing your data
   - Describe your data
   - Making graphs

What do we have to do for the paper?

University of Bamberg

1) State of research (brief literature review)
2) Defining the question
   - Wage effects of immigration
   - Employment effects
3) Preparing the dataset
   - Generating Dummy Variables
   - Generating/Transforming other Variables
4) Describing the data
   - Data sources
   - Descriptive statistics
   - Graphs
5) Running simple regression models
6) Presentation in class and drafting the paper

Recall the Borjas (2003)-Model

$$y_{ijt} = \theta p_{ijt} + s_i + x_j + \pi_t + (s_i \times x_j) + (s_i \times \pi_t) + (x_j \times \pi_t) + \varphi_{ijt},$$

where

- y_{ijt} is the dependent variable (e.g. log wage, unemployment rate) in education category *i,* work-experience category *j,* at time *t*
- s_i is an education dummy
- x_j is an work experience dummy
- π_t is a time dummy
- plus many interaction dummies

This is what we want to estimate. For this we need certain tools, inter alia STATA knowledge.

Data preparation with STATA

What is important to know about the data

- The data are derived from micro administrative data (cumbersome)
- **Denmark:** IDA register data (social security data)
  - whole universe of the population
- **Germany:** IEB register data (social security records)
  - 5% sample
- **UK:** Labour Force Survey data
  - 0,5-1% sample of population
- Differences
  - Country of birth (DK, UK), nationality and past nationality (DE)
  - Education: degrees (DE and DK), years of schooling (UK) due to educational systems
- Many, many details

What is important to know about the data

Literature reference:

Brücker, Herbert; Hauptmann, Andreas, Jahn, Elke, Upward, Richard (2014), “Migration and imperfect labour markets. Theory and cross-country evidence from Denmark, Germany and the UK”, in: *European Economic Review*, Vol. 66, pp. 205-225.

How to organize the datafile?

Variable **Definition**

- **Year:** 19XX to 20XX
- **Education:** 3 groups by degree (DE, DK) or years of schooling (UK)
- **Experience:** 4 groups by work experience (0-5, 6-10, 11-20, over 20)
- **LHqjt:** Number of workers by nationality, education, workexperience and year
- **UHqjt:** Number of unemployed workers by nationality, education and workexperience and year
- **NHqjt:** Labour force (L+U) by nationality, education, work experience and year
- **Hsumwage:** Wage-sum of native workers (LHt*wht)
- **Fsumwage:** Wage-sum of immigrant workers (LFt*wft)

How to organize the datafile?

University of Bamberg

Variable	Definition
• **wHqjt**:	wage of workers by nationality *H*, education *q*, work experience *j* and year *t*
• **wHqt**:	mean wage aggregate by nationality and education and year
• **wHt**:	mean wage aggregate by nationality and year
• **wt**	mean wage aggregate by year Same for **L**, **U**, and **N**
Indices:	
• **H, F**:	Natives, Immigrants
• **q**:	education
• **j**:	work experience
• **t**:	time (year)

Working with STATA: Introduction into the Basics

Helpful reading

Good STATA textbooks

- Ulrich Kohler / Frauke Kreuter: *Datananalyse mit STATA: allgemeine Konzepte der Datenanalyse und ihre praktische Anwendung*, 4th Edition. (German)

- Ulrich Kohler and Frauke Kreuter: *Data Analysis Using Stata*, 2012, 3rd Edition. (English)

STATA Software Package

University of Bamberg

1. Connect to Bamberg University data-net via cable, WLAN or VPN

2. Switch to

   \\lizenz01.rz.uni-bamberg.de\Stata_SoWi (for Windows)

   smb://lizenz01.rz.uni-bamberg.de/Stata_SoWi (for Mac)

3. Insert your username in format uni-Bamberg.de\BA-Kennung

4. Use START_Stata_SoWi_Win (_Mac) button to start Software (doubleclick)
   Note that this may need several minutes

5. Support: softwarebeschaffung.rz@uni-bamberg.de;
   +49 951 863-1327 (-1333)

Note: there are only limited licenses. Close Stata if you don't use it!

Structure of STATA: Three files

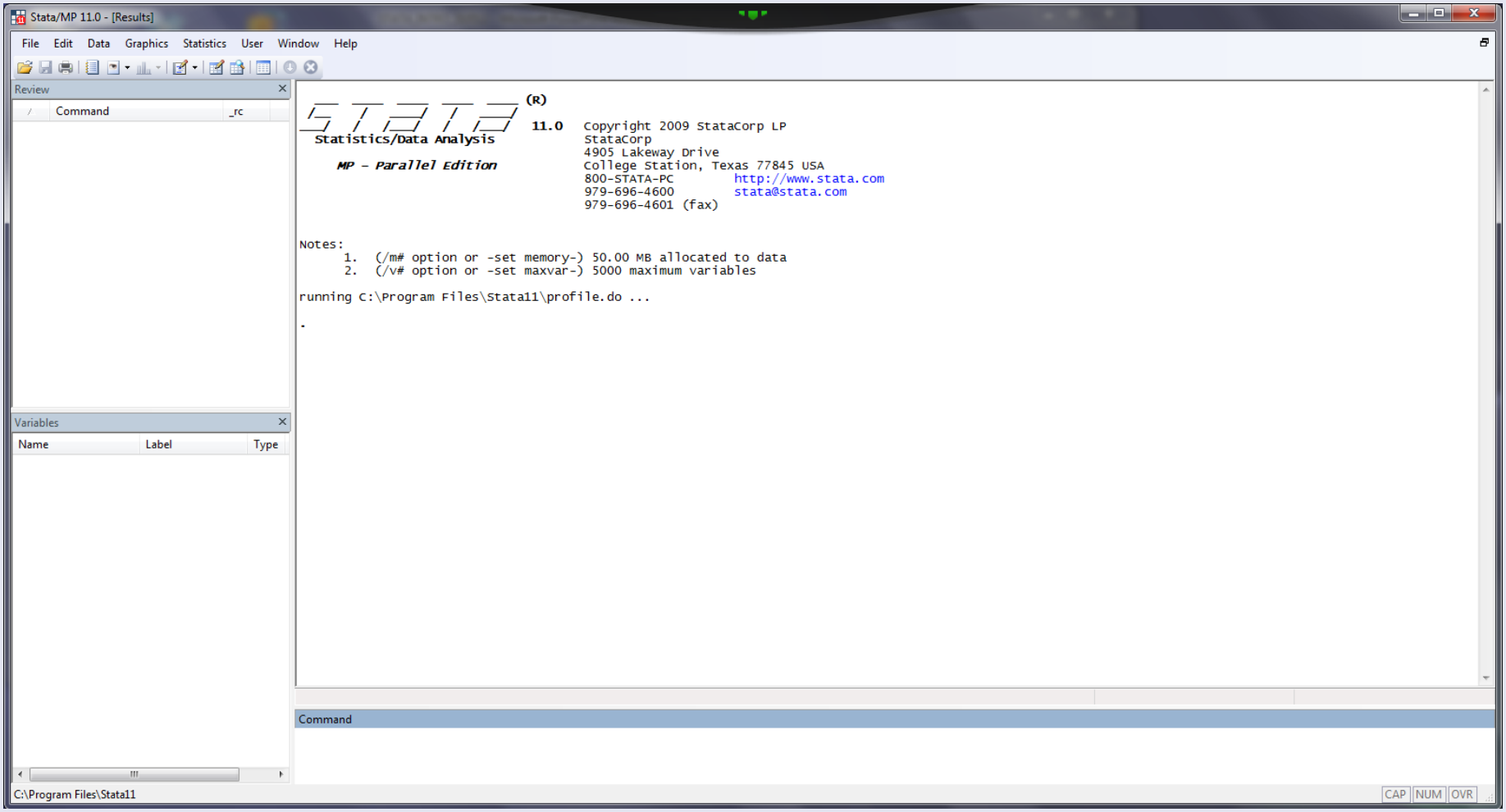
1. The **DATA file (.dta)** where you have your data.
   - **DATA BROWSER:** watching your data
   - **DATA EDITOR**: editing your data

2. The **DO file (.do)** where you run and save your commands of any session.
   Very useful (i) to organize your data set, (ii) to see what you have done in the last session, (iii) to replicate what you have done in last session, (iv) to exchange work with your collaborators.
   - You write and run your commands with the **DO FILE EDITOR**

3. The **LOG file (.scmf)** which automatically reports all things which you have done during your session. Is automatically saved after your session. Useful if something goes wrong.

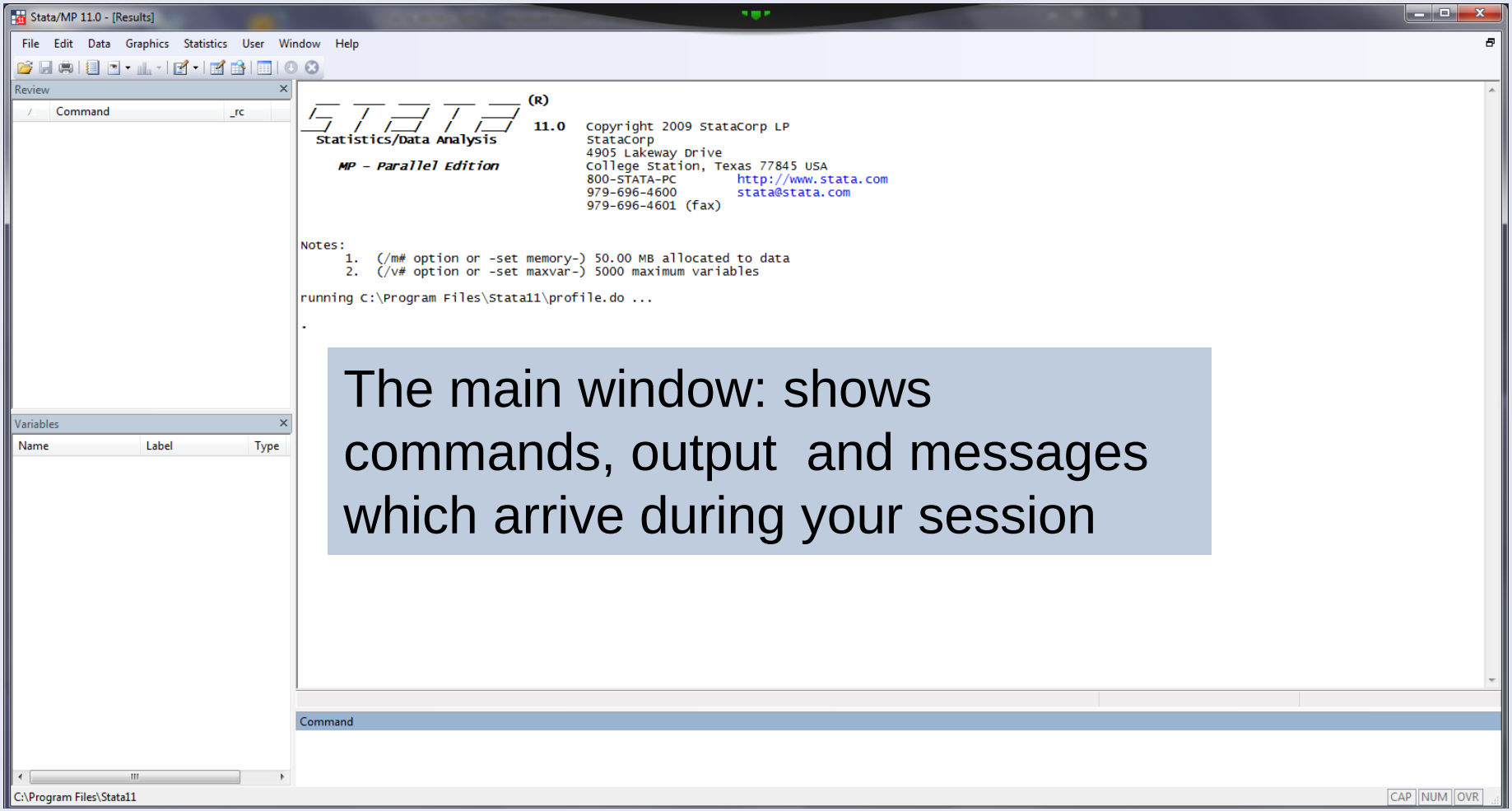
Getting started: the STATA empty window

University of Bamberg



Getting started: the STATA empty window

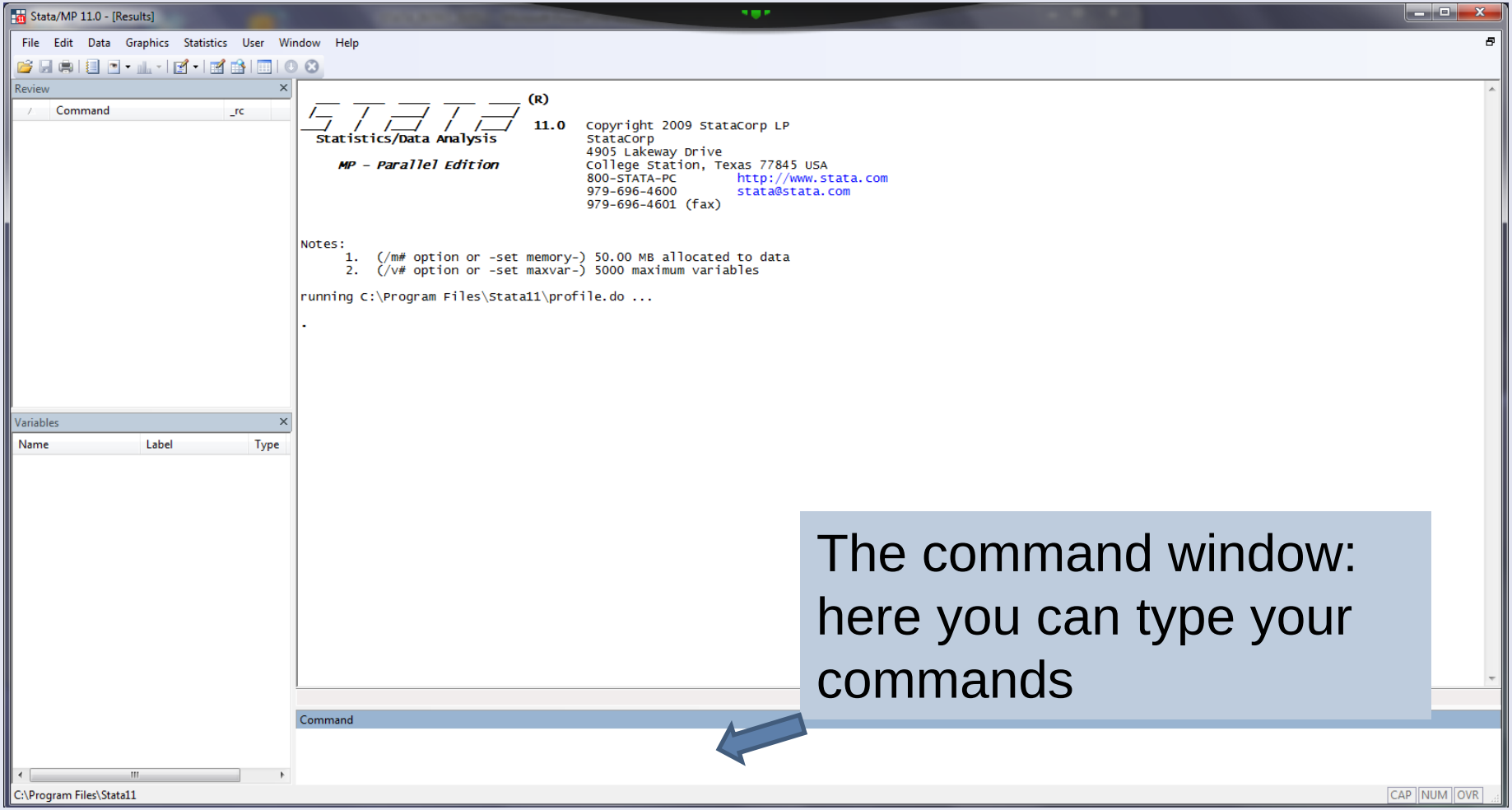
University of Bamberg



The main window: shows commands, output and messages which arrive during your session

Getting started: the STATA empty window

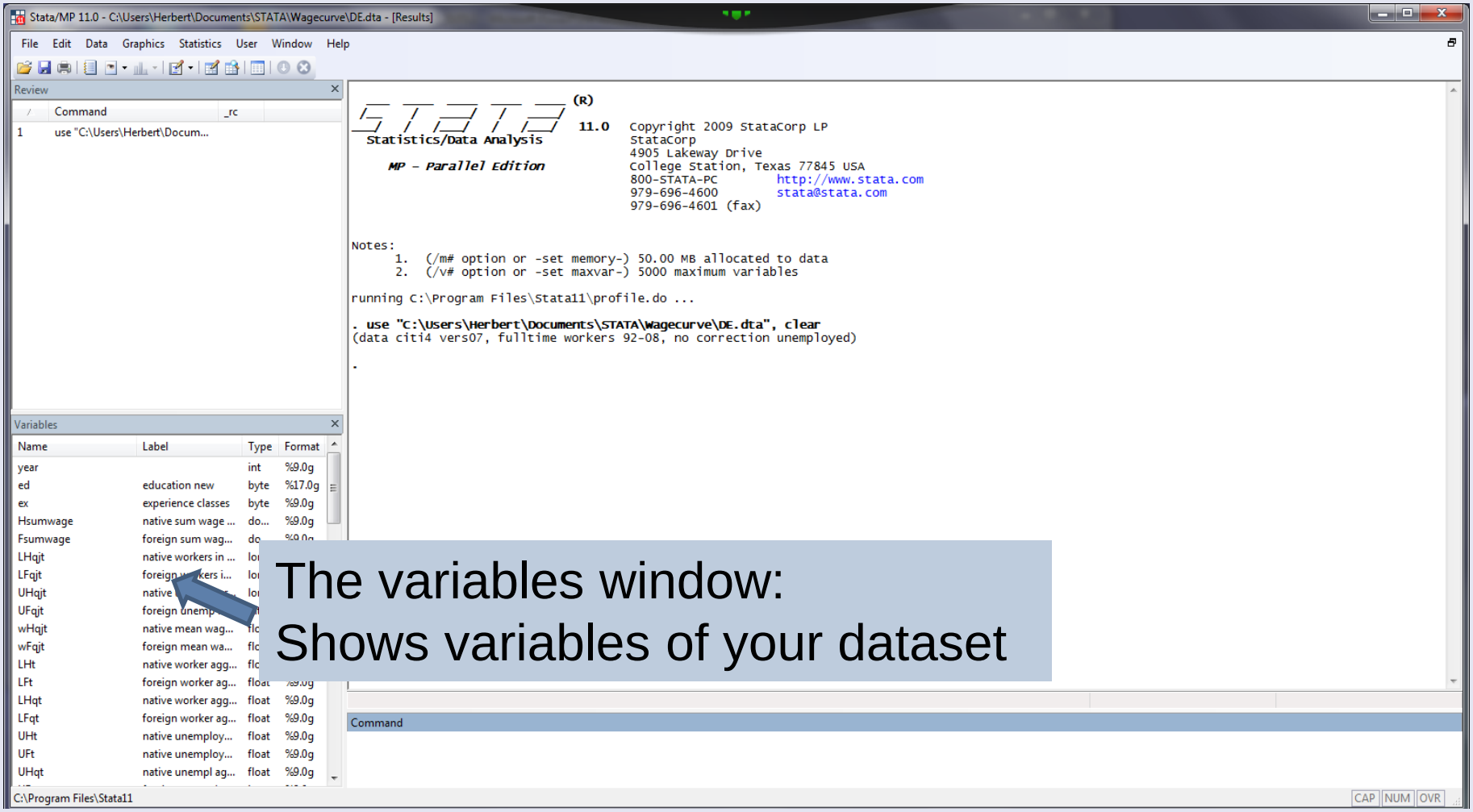
University of Bamberg



The command window: here you can type your commands

Getting started: the STATA empty window

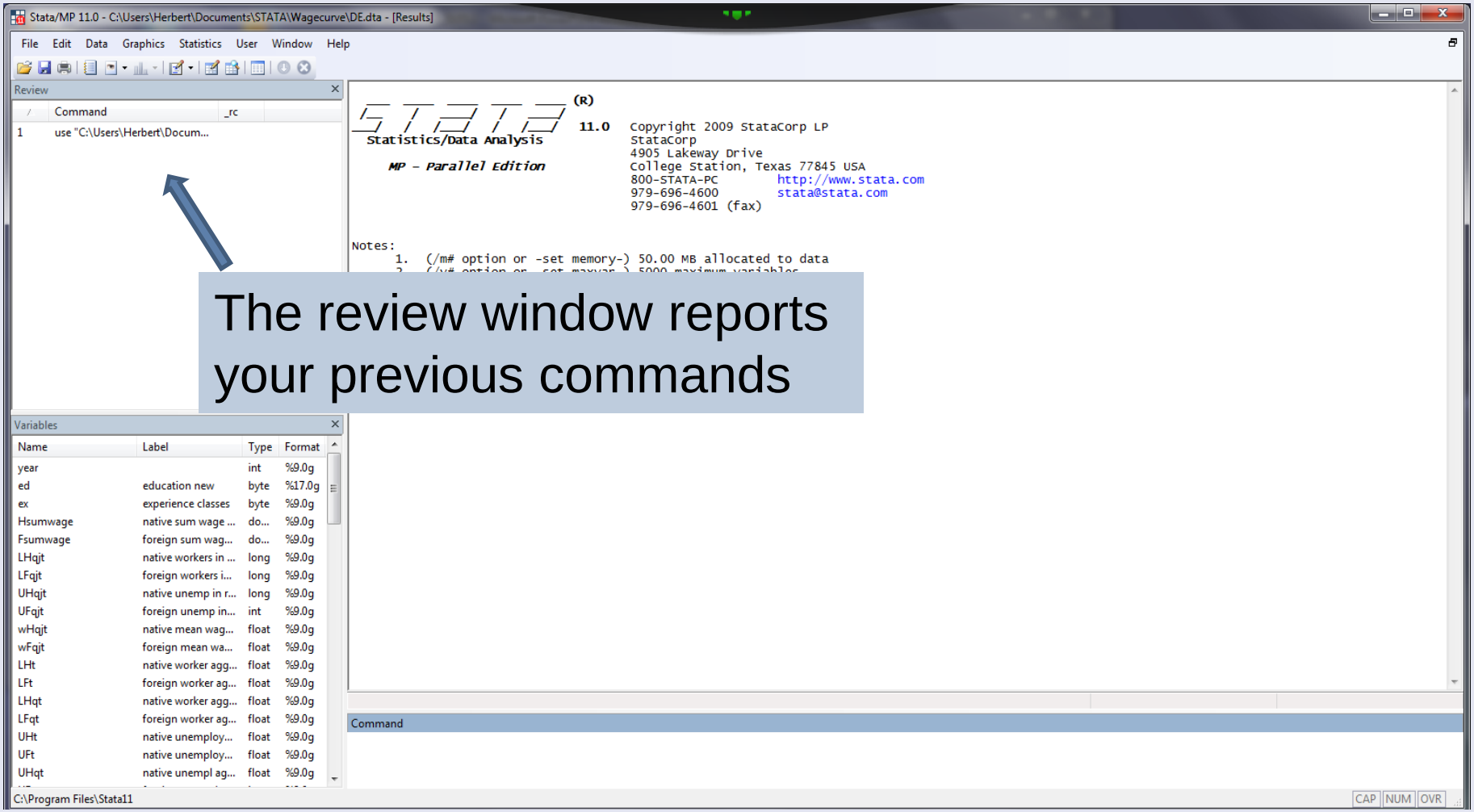
University of Bamberg



The variables window:
Shows variables of your dataset

Getting started: the STATA empty window

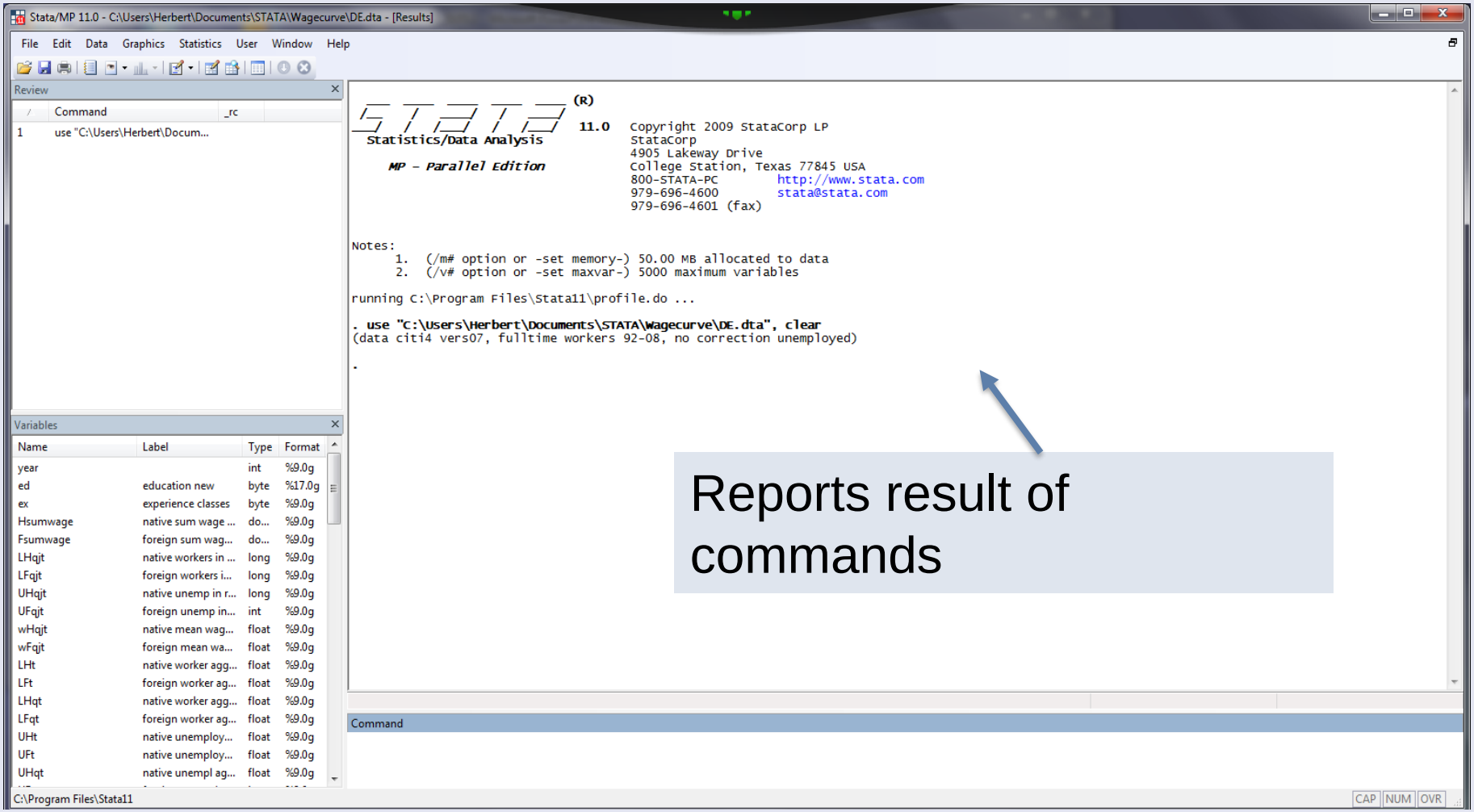
University of Bamberg



The review window reports your previous commands

Getting started: the STATA empty window

University of Bamberg

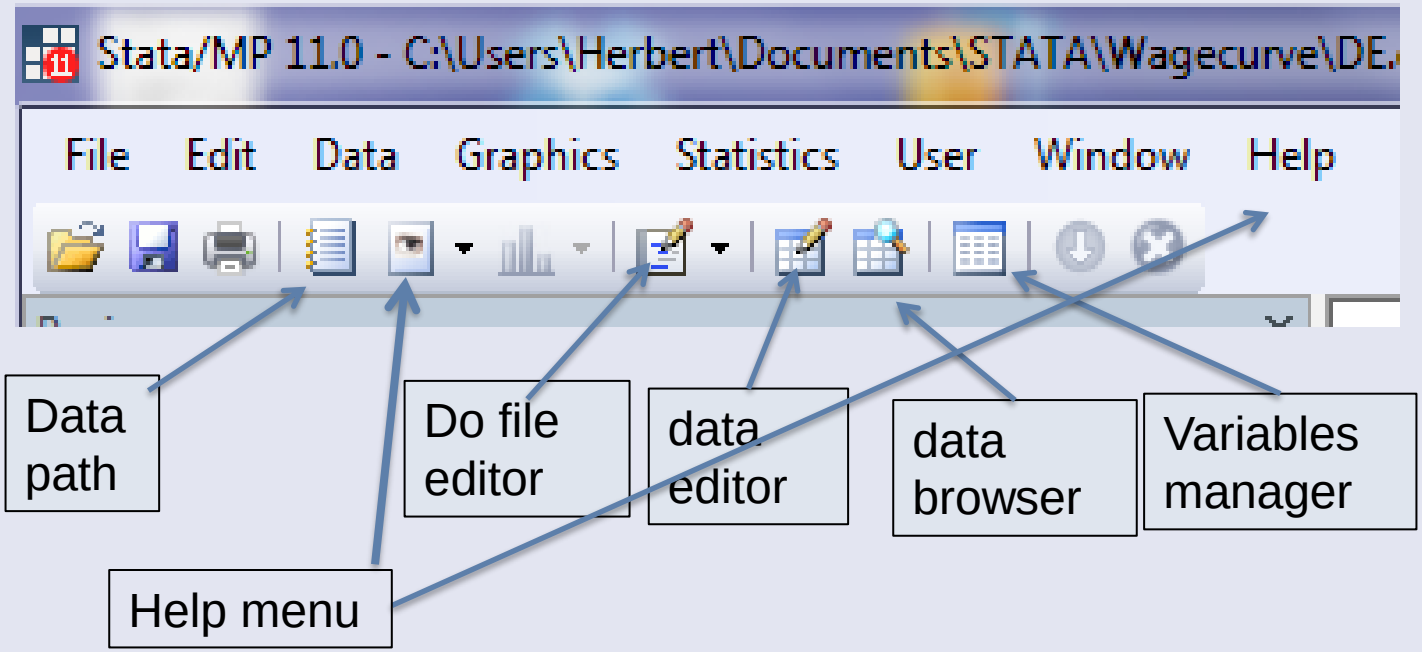


Reports result of commands

Getting started

- In principle, you can start your STATA session by (i) loading your data set and (ii) typing your commands in the command window.
- It is however recommended to use the **DO FILE EDITOR** right from the beginning.
- But let's look at the **STATA menus** first.

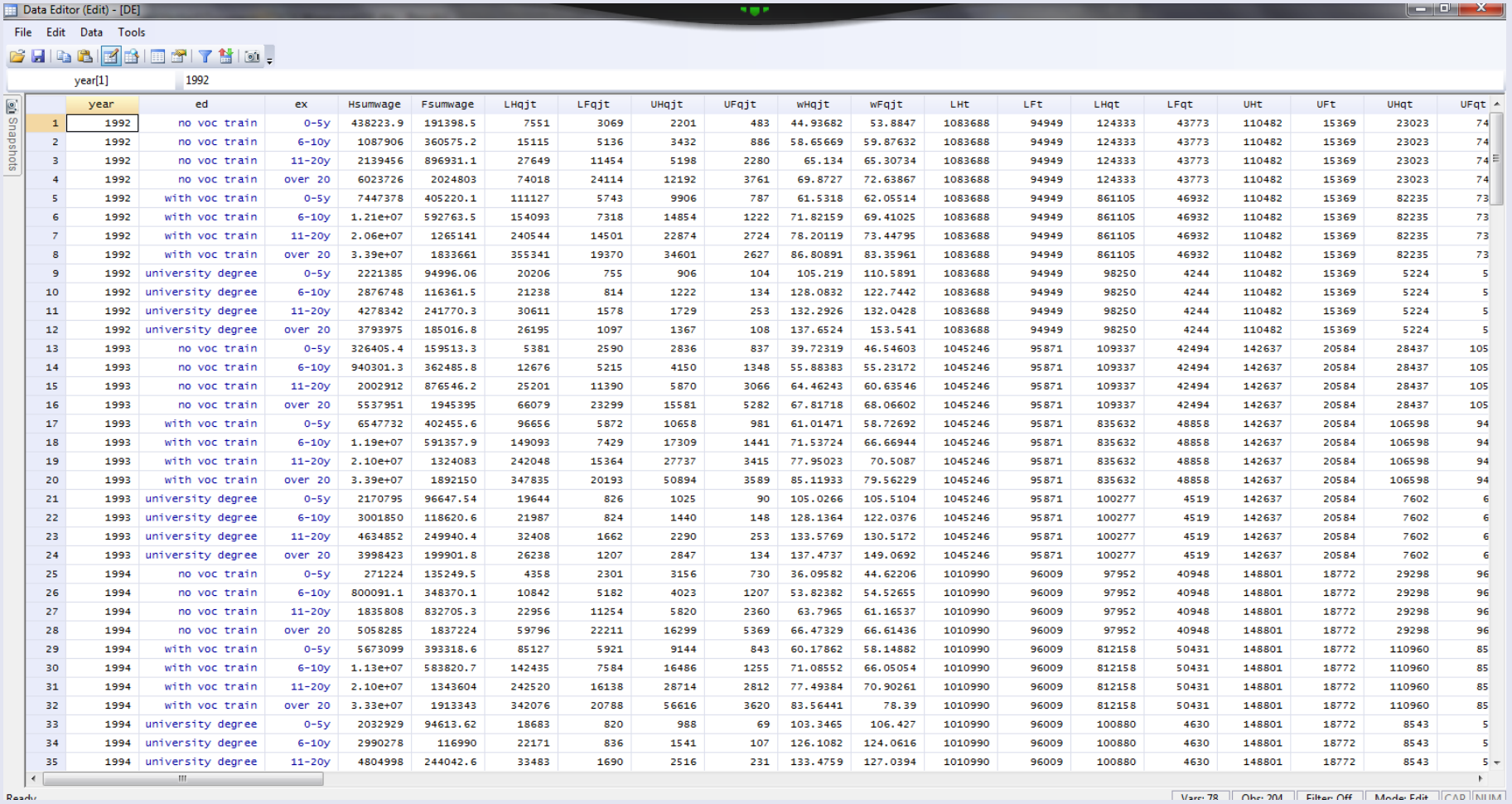
The STATA Menus



- For watching your data and changing your data by hand you need the DATA BROWSER and the DATA EDITOR.
- For starting and running your DO files you need the DO FILE EDITOR.
- The other menus are not relevant for the beginning.

The STATA Menus: The Data Editor/ Browser

University of Bamberg

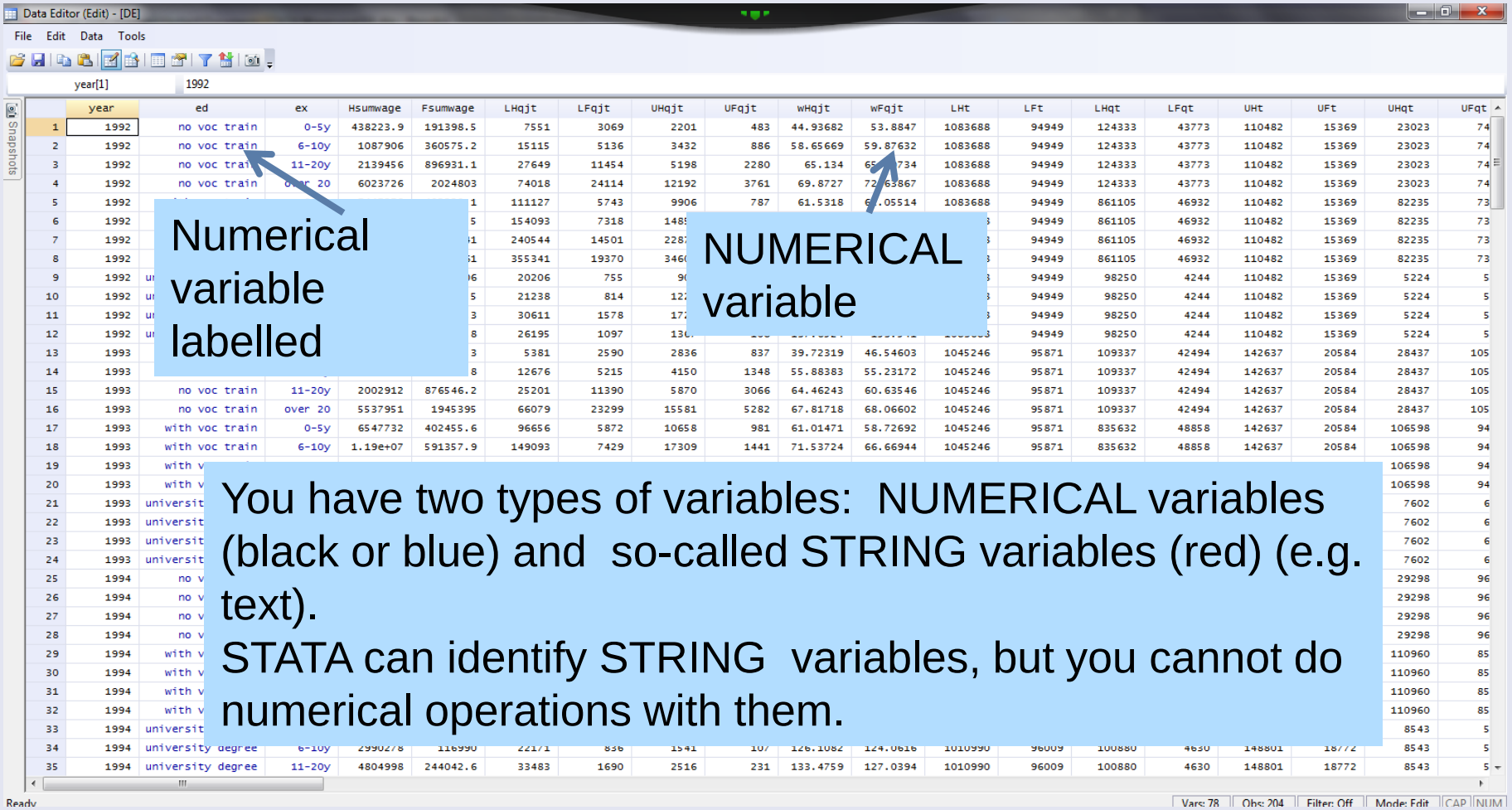


	year	ed	ex	Hsumwage	Fsumwage	LHqjt	LFqjt	UHqjt	UFqjt	wHqjt	wFqjt	LHt	LFt	LHqt	LFqt	UHt	UFt	UHqt	UFqt
1	1992	no voc train	0-5y	438223.9	191398.5	7551	3069	2201	483	44.93682	53.8847	1083688	94949	124333	43773	110482	15369	23023	74
2	1992	no voc train	6-10y	1087906	360575.2	15115	5136	3432	886	58.65669	59.87632	1083688	94949	124333	43773	110482	15369	23023	74
3	1992	no voc train	11-20y	2139456	896931.1	27649	11454	5198	2280	65.134	65.30734	1083688	94949	124333	43773	110482	15369	23023	74
4	1992	no voc train	over 20	6023726	2024803	74018	24114	12192	3761	69.8727	72.63867	1083688	94949	124333	43773	110482	15369	23023	74
5	1992	with voc train	0-5y	7447378	405220.1	111127	5743	9906	787	61.5318	62.05514	1083688	94949	861105	46932	110482	15369	82235	73
6	1992	with voc train	6-10y	1.21e+07	592763.5	154093	7318	14854	1222	71.82159	69.41025	1083688	94949	861105	46932	110482	15369	82235	73
7	1992	with voc train	11-20y	2.06e+07	1265141	240544	14501	22874	2724	78.20119	73.44795	1083688	94949	861105	46932	110482	15369	82235	73
8	1992	with voc train	over 20	3.39e+07	1833661	355341	19370	34601	2627	86.80891	83.35961	1083688	94949	861105	46932	110482	15369	82235	73
9	1992	university degree	0-5y	2221385	94996.06	20206	755	906	104	105.219	110.5891	1083688	94949	98250	4244	110482	15369	5224	5
10	1992	university degree	6-10y	2876748	116361.5	21238	814	1222	134	128.0832	122.7442	1083688	94949	98250	4244	110482	15369	5224	5
11	1992	university degree	11-20y	4278342	241770.3	30611	1578	1729	253	132.2926	132.0428	1083688	94949	98250	4244	110482	15369	5224	5
12	1992	university degree	over 20	3793975	185016.8	26195	1097	1367	108	137.6524	153.541	1083688	94949	98250	4244	110482	15369	5224	5
13	1993	no voc train	0-5y	326405.4	159513.3	5381	2590	2836	837	39.72319	46.54603	1045246	95871	109337	42494	142637	20584	28437	105
14	1993	no voc train	6-10y	940301.3	362485.8	12676	5215	4150	1348	55.88383	55.23172	1045246	95871	109337	42494	142637	20584	28437	105
15	1993	no voc train	11-20y	2002912	876546.2	25201	11390	5870	3066	64.46243	60.63546	1045246	95871	109337	42494	142637	20584	28437	105
16	1993	no voc train	over 20	5537951	1945395	66079	23299	15581	5282	67.81718	68.06602	1045246	95871	109337	42494	142637	20584	28437	105
17	1993	with voc train	0-5y	6547732	402455.6	96656	5872	10658	981	61.01471	58.72692	1045246	95871	835632	48858	142637	20584	106598	94
18	1993	with voc train	6-10y	1.19e+07	591357.9	149093	7429	17309	1441	71.53724	66.66944	1045246	95871	835632	48858	142637	20584	106598	94
19	1993	with voc train	11-20y	2.10e+07	1324083	242048	15364	27737	3415	77.95023	70.5087	1045246	95871	835632	48858	142637	20584	106598	94
20	1993	with voc train	over 20	3.39e+07	1892150	347835	20193	50894	3589	85.11933	79.56229	1045246	95871	835632	48858	142637	20584	106598	94
21	1993	university degree	0-5y	2170795	96647.54	19644	826	1025	90	105.0266	105.5104	1045246	95871	100277	4519	142637	20584	7602	6
22	1993	university degree	6-10y	3001850	118620.6	21987	824	1440	148	128.1364	122.0376	1045246	95871	100277	4519	142637	20584	7602	6
23	1993	university degree	11-20y	4634852	249940.4	32408	1662	2290	253	133.5769	130.5172	1045246	95871	100277	4519	142637	20584	7602	6
24	1993	university degree	over 20	3998423	199901.8	26238	1207	2847	134	137.4737	149.0692	1045246	95871	100277	4519	142637	20584	7602	6
25	1994	no voc train	0-5y	271224	135249.5	4358	2301	3156	730	36.09582	44.62206	1010990	96009	97952	40948	148801	18772	29298	96
26	1994	no voc train	6-10y	800091.1	348370.1	10842	5182	4023	1207	53.82382	54.52655	1010990	96009	97952	40948	148801	18772	29298	96
27	1994	no voc train	11-20y	1835808	832705.3	22956	11254	5820	2360	63.7965	61.16537	1010990	96009	97952	40948	148801	18772	29298	96
28	1994	no voc train	over 20	5058285	1837224	59796	22211	16299	5369	66.47329	66.61436	1010990	96009	97952	40948	148801	18772	29298	96
29	1994	with voc train	0-5y	5673099	393318.6	85127	5921	9144	843	60.17862	58.14882	1010990	96009	812158	50431	148801	18772	110960	85
30	1994	with voc train	6-10y	1.13e+07	583820.7	142435	7584	16486	1255	71.08552	66.05054	1010990	96009	812158	50431	148801	18772	110960	85
31	1994	with voc train	11-20y	2.10e+07	1343604	242520	16138	28714	2812	77.49384	70.90261	1010990	96009	812158	50431	148801	18772	110960	85
32	1994	with voc train	over 20	3.33e+07	1913343	342076	20788	56616	3620	83.56441	78.39	1010990	96009	812158	50431	148801	18772	110960	85
33	1994	university degree	0-5y	2032929	94613.62	18683	820	988	69	103.3465	106.427	1010990	96009	100880	4630	148801	18772	8543	5
34	1994	university degree	6-10y	2990278	116990	22171	836	1541	107	126.1082	124.0616	1010990	96009	100880	4630	148801	18772	8543	5
35	1994	university degree	11-20y	4804998	244042.6	33483	1690	2516	231	133.4759	127.0394	1010990	96009	100880	4630	148801	18772	8543	5

The difference between the data browser and the data editor is that you can manipulate data in the editor and only watch them in the browser.

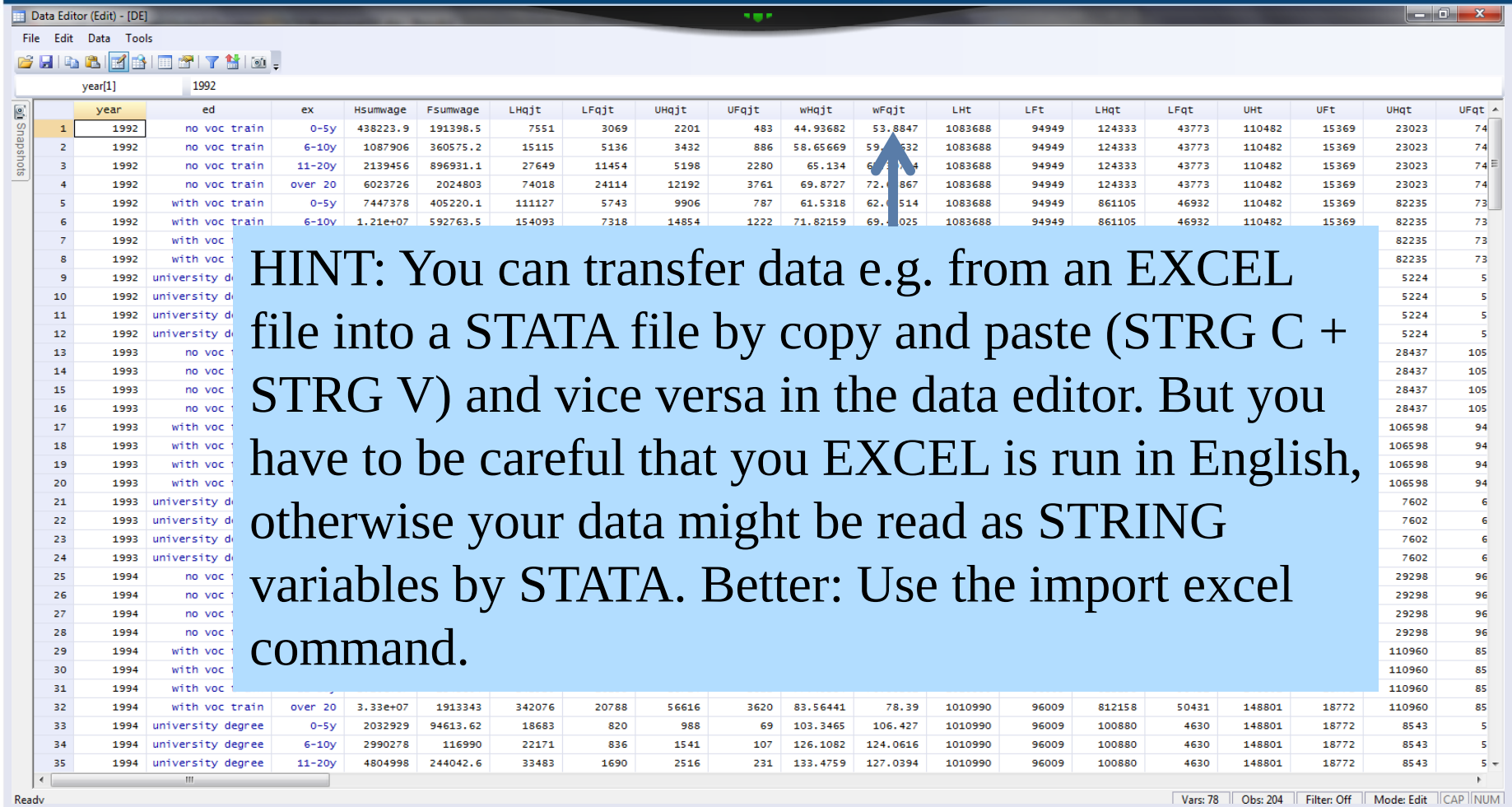
The STATA Menus: The Data Editor/ Browser

University of Bamberg



	year	ed	ex	Hsumwage	Fsumwage	LHqjt	LFqjt	UHqjt	UFqjt	wHqjt	wFqjt	LHt	LFt	LHqt	LFqt	UHt	UFt	UHqt	UFqt
1	1992	no voc train	0-5y	438223.9	191398.5	7551	3069	2201	483	44.93682	53.8847	1083688	94949	124333	43773	110482	15369	23023	74
2	1992	no voc train	6-10y	1087906	360575.2	15115	5136	3432	886	58.65669	59.87632	1083688	94949	124333	43773	110482	15369	23023	74
3	1992	no voc train	11-20y	2139456	896931.1	27649	11454	5198	2280	65.134	65.[illegible]734	1083688	94949	124333	43773	110482	15369	23023	74
4	1992	no voc train	over 20	6023726	2024803	74018	24114	12192	3761	69.8727	72.[illegible]3867	1083688	94949	124333	43773	110482	15369	23023	74
5	1992					[illegible]1	111127	5743	9906	787	61.5318	6[illegible].05514	1083688	94949	861105	46932	110482	15369	82235	73
6	1992					[illegible]5	154093	7318	1485					94949	861105	46932	110482	15369	82235	73
7	1992					[illegible]41	240544	14501	2287					94949	861105	46932	110482	15369	82235	73
8	1992					[illegible]61	355341	19370	3460					94949	861105	46932	110482	15369	82235	73
9	1992	u				[illegible]06	20206	755	90					94949	98250	4244	110482	15369	5224	5
10	1992	u				[illegible]5	21238	814	122					94949	98250	4244	110482	15369	5224	5
11	1992	u				[illegible]3	30611	1578	172					94949	98250	4244	110482	15369	5224	5
12	1992	u				[illegible]8	26195	1097	130					94949	98250	4244	110482	15369	5224	5
13	1993					[illegible]3	5381	2590	2836	837	39.72319	46.54603	1045246	95871	109337	42494	142637	20584	28437	105
14	1993					[illegible]8	12676	5215	4150	1348	55.88383	55.23172	1045246	95871	109337	42494	142637	20584	28437	105
15	1993	no voc train	11-20y	2002912	876546.2	25201	11390	5870	3066	64.46243	60.63546	1045246	95871	109337	42494	142637	20584	28437	105
16	1993	no voc train	over 20	5537951	1945395	66079	23299	15581	5282	67.81718	68.06602	1045246	95871	109337	42494	142637	20584	28437	105
17	1993	with voc train	0-5y	6547732	402455.6	96656	5872	10658	981	61.01471	58.72692	1045246	95871	835632	48858	142637	20584	106598	94
18	1993	with voc train	6-10y	1.19e+07	591357.9	149093	7429	17309	1441	71.53724	66.66944	1045246	95871	835632	48858	142637	20584	106598	94
19	1993	with v																106598	94
20	1993	with v																106598	94
21	1993	universit																7602	6
22	1993	universit																7602	6
23	1993	universit																7602	6
24	1993	universit																7602	6
25	1994	no v																29298	96
26	1994	no v																29298	96
27	1994	no v																29298	96
28	1994	no v																29298	96
29	1994	with v																110960	85
30	1994	with v																110960	85
31	1994	with v																110960	85
32	1994	with v																110960	85
33	1994	universit																8543	5
34	1994	university degree	6-10y	2990278	116990	22171	836	1541	107	126.1082	124.0616	1010990	96009	100880	4630	148801	18772	8543	5
35	1994	university degree	11-20y	4804998	244042.6	33483	1690	2516	231	133.4759	127.0394	1010990	96009	100880	4630	148801	18772	8543	5

The STATA Menus: The Data Editor/ Browser



	year	ed	ex	Hsumwage	Fsumwage	LHqjt	LFqjt	UHqjt	UFqjt	wHqjt	wFqjt	LHt	LFt	LHqt	LFqt	UHt	UFt	UHqt	UFqt
1	1992	no voc train	0-5y	438223.9	191398.5	7551	3069	2201	483	44.93682	53.8847	1083688	94949	124333	43773	110482	15369	23023	74
2	1992	no voc train	6-10y	1087906	360575.2	15115	5136	3432	886	58.65669	59.[illegible]32	1083688	94949	124333	43773	110482	15369	23023	74
3	1992	no voc train	11-20y	2139456	896931.1	27649	11454	5198	2280	65.134	6[illegible]4	1083688	94949	124333	43773	110482	15369	23023	74
4	1992	no voc train	over 20	6023726	2024803	74018	24114	12192	3761	69.8727	72.[illegible]867	1083688	94949	124333	43773	110482	15369	23023	74
5	1992	with voc train	0-5y	7447378	405220.1	111127	5743	9906	787	61.5318	62.[illegible]514	1083688	94949	861105	46932	110482	15369	82235	73
6	1992	with voc train	6-10y	1.21e+07	592763.5	154093	7318	14854	1222	71.82159	69.[illegible]025	1083688	94949	861105	46932	110482	15369	82235	73
7	1992	with voc																82235	73
8	1992	with voc																82235	73
9	1992	university d																5224	5
10	1992	university d																5224	5
11	1992	university d																5224	5
12	1992	university d																5224	5
13	1993	no voc																28437	105
14	1993	no voc																28437	105
15	1993	no voc																28437	105
16	1993	no voc																28437	105
17	1993	with voc																106598	94
18	1993	with voc																106598	94
19	1993	with voc																106598	94
20	1993	with voc																106598	94
21	1993	university d																7602	6
22	1993	university d																7602	6
23	1993	university d																7602	6
24	1993	university d																7602	6
25	1994	no voc																29298	96
26	1994	no voc																29298	96
27	1994	no voc																29298	96
28	1994	no voc																29298	96
29	1994	with voc																110960	85
30	1994	with voc																110960	85
31	1994	with voc																110960	85
32	1994	with voc train	over 20	3.33e+07	1913343	342076	20788	56616	3620	83.56441	78.39	1010990	96009	812158	50431	148801	18772	110960	85
33	1994	university degree	0-5y	2032929	94613.62	18683	820	988	69	103.3465	106.427	1010990	96009	100880	4630	148801	18772	8543	5
34	1994	university degree	6-10y	2990278	116990	22171	836	1541	107	126.1082	124.0616	1010990	96009	100880	4630	148801	18772	8543	5
35	1994	university degree	11-20y	4804998	244042.6	33483	1690	2516	231	133.4759	127.0394	1010990	96009	100880	4630	148801	18772	8543	5

HINT: You can transfer data e.g. from an EXCEL file into a STATA file by copy and paste (STRG C + STRG V) and vice versa in the data editor. But you have to be careful that you EXCEL is run in English, otherwise your data might be read as STRING variables by STATA. Better: Use the import excel command.

The Grammar of STATA

- General Structure of STATA commands

[prefix :] command [varlist] [if] [in] [weight] [, options]

The Grammar of STATA

University of Bamberg

- We will concentrate on:

[prefix :] **command [varlist] [if]** [in] [weight] **[, options]**

Grammar of STATA: Commands

[prefix :] **command** **[varlist] [if]** [in] [weight] **[, options]**

- What you want to do?

<u>Examples:</u>

- Run a regression: ***regress***
- Make summary statistics: ***summarize***
- Produce graphs: ***graph twoway line***

Grammar of STATA: Using abbreviations

University of Bamberg

[prefix :] **command** **[varlist] [if]** [in] [weight] **[, options]**

Examples:

- ***r*** or ***reg*** instead of ***regress***
- ***s*** or ***sum*** instead of ***summarize***
- ***g*** instead of ***graph***

Grammar of STATA: Variables

[prefix :] **command [varlist] [if]** [in] [weight] **[, options]**

- ***varlist*** is a list of variables, which you have in your dataset

Examples:

- Foreign wage: **wft**
- Foreign labor force: **Lft**

- You can also use list of variables which comprises a number of variables you have defined before (see working with ***macros*** below).

Grammar of STATA: the if-condition

[prefix :] **command [varlist]** **[if]** [in] [weight] **[, options]**

- The ***if*** condition constrains what you want to do to a certain condition you have defined.

<u>Example:</u>

- You want to replace a variable with certain values ***if*** the individuals have higher education:

```
replace ed3 = 1 if education == 3
```

Grammar of STATA: the if-condition

[prefix :] **command [varlist]** **[if]** [in] [weight] **[, options]**

<u>Another example:</u>

- You want to restrict a regression to a subsample, e.g. to **foreigners** (**Nft**) only:

```
reg  wqjt mqjt Dqjt Djt Dt if Nft > 0
```

Grammar of STATA: Options

[prefix :] **command [varlist] [if]** [in] [weight] **[, options]**

- ***Options*** are very often complementing commands, e.g. you can run a regression with fixed effects.

Example:

xtreg wqjt mqjt Dqjt Djt, **fe**

where **fe** is the option for fixed effects.

Grammar of STATA: Types of variables

University of Bamberg

- There are two types of variables (data):
- **numerical variables**, e.g.: 0, 1, 501, 0.5, -12 etc.
- **string variables**, e.g.: no voc train , male, female etc.

- How to deal with the data types:
- Numerical variables: you can do all mathematical operations,

  e.g. var1 + var2, var1/var2, var1*var2 etc.

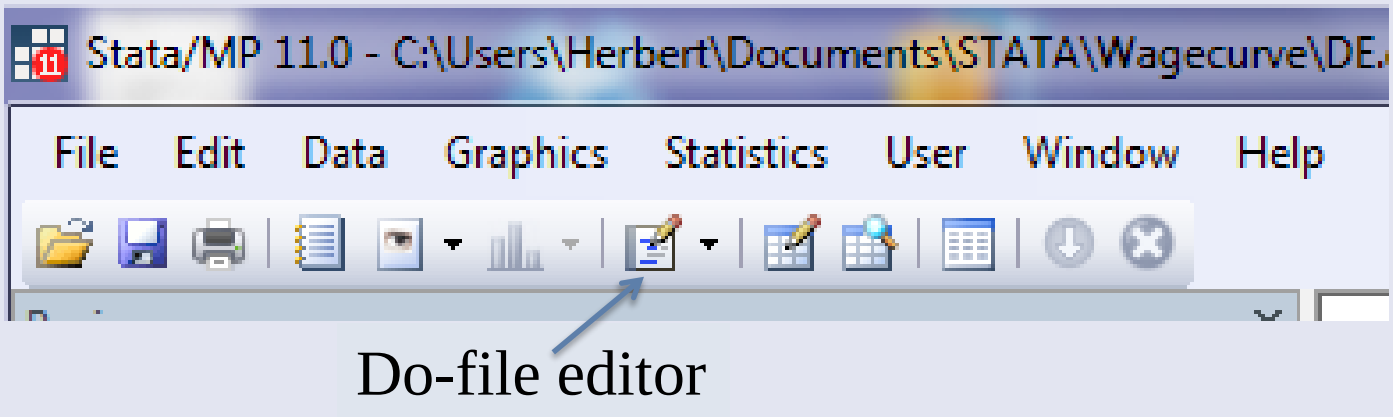
- String variables: You have to use quotation marks for identification, e.g.
- var1 = 1 if sex == “female”

Working with DO FILES: General issues

University of Bamberg

- The standard approach is to start your work with a **DO FILE**
- Click on the **DO FILE editor button** after starting STATA
- Load an existing **DO FILE** or start a new one
- Save the DO File at the end of your session

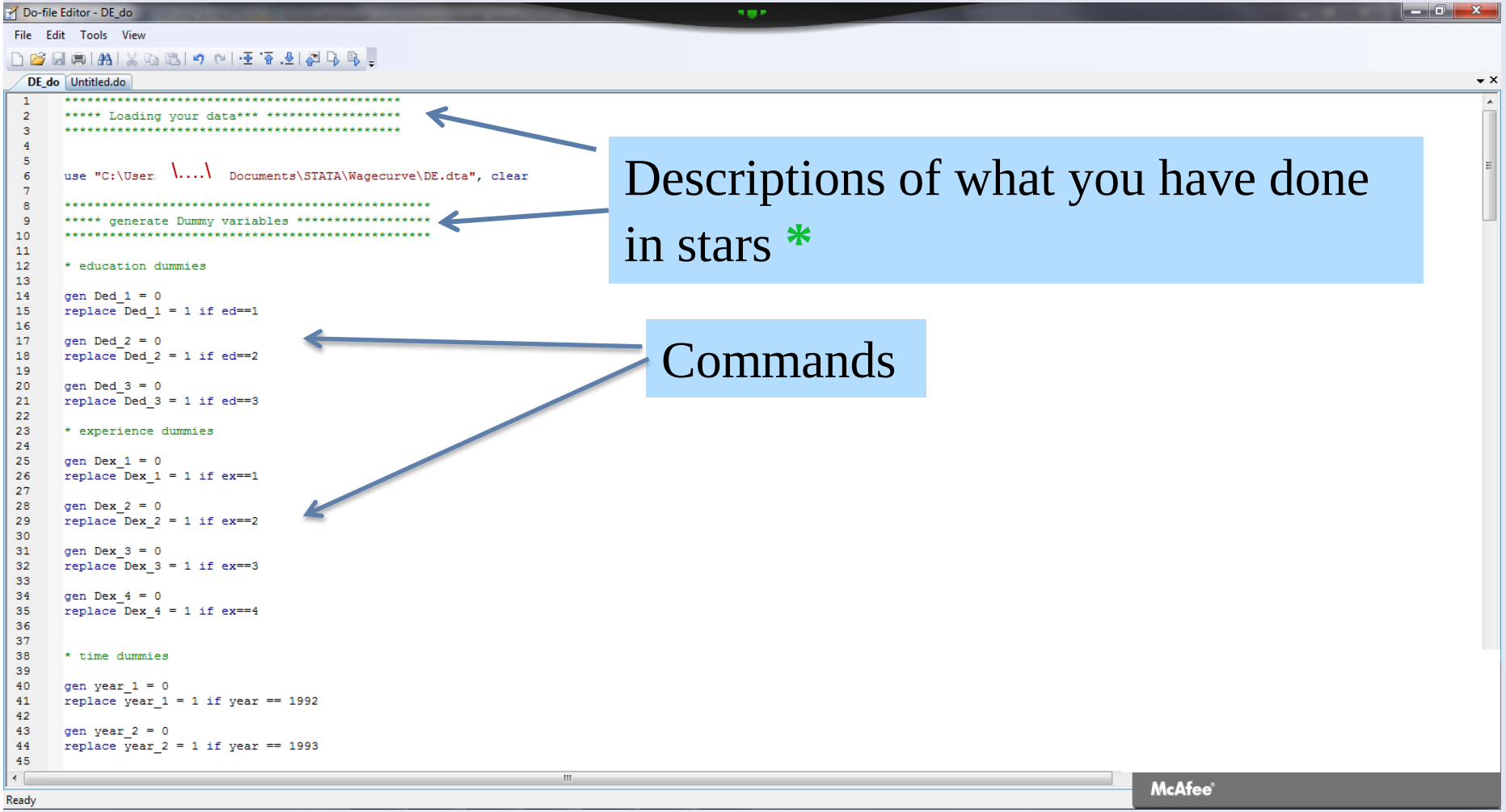
Open your DO FILE editor



- After starting STATA click on the DO FILE editor button

How does a DO FILE look like

University of Bamberg

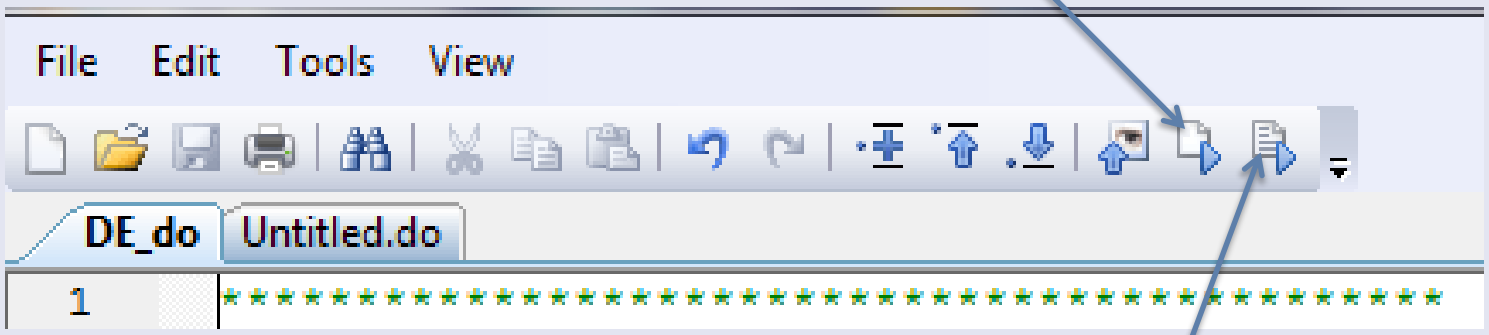


Descriptions of what you have done in stars *

Commands

DO FILE Menu

Clicking this button runs the entire DO FILE (not recommended)



Clicking this button runs a selection of marked commands (recommended)

- Note: STATA stops the DO File execution after the first mistake in your commands. That makes it advisable to proceed step by step.

Organizing your session: General issues

University of Bamberg

- The basis for all what you do is your **Do-File** which you open in all sessions first
- Work with a **small data set** and do the generation of all additional variables based on your Do-File in the beginning of each session. Save only the small dataset. That is efficient from the data management side and reduces the risk that you delete/change important variables which you cannot restore. Change somewhere a saved version of your dataset
- Use a **LOG File** that you can see in case of a problem what you have actually done. You seldomly look at it, but you might miss it in an important case
- At the end of session save the Do-file and close Log file
- Close **data-file** but **do not** save it: this can destroy your dataset!

Step 1: Starting your session

University of Bamberg

- Begin your Do-file with the following useful commands:

set more off

- Is not necessary but useful
- Helps data processing e.g. in long regressions, long statistics

capture log close

- closes log file if one is open
- the **capture** (**cap**) command is useful since STATA executes this command only if it is need, e.g. if a log file is open

log using "path\DE", replace

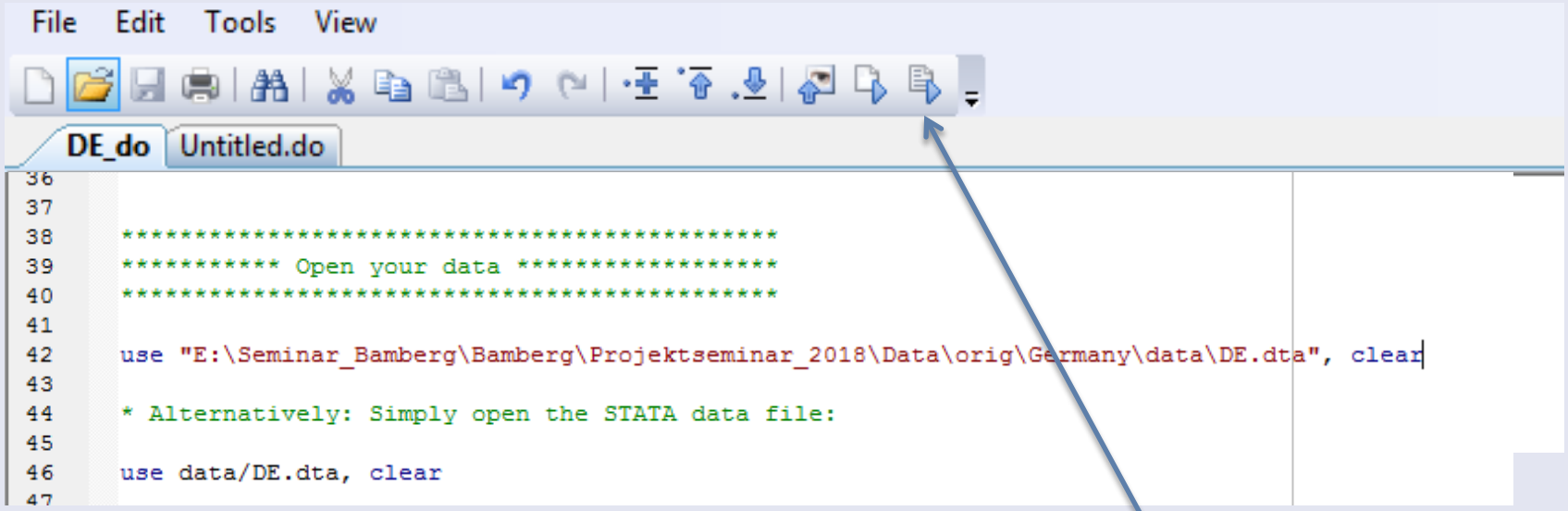
- Opens a Log-File with the name "DE"

Step 2: Loading your data

University of Bamberg

- If you have already a STATA data file:
- The **use** command loads the data
- the "**path\data\DE.dta, clear"** provides STATA the information on the path where to find the data and the name of the data file (e.g. **DE.dta**)
- the **clear** command after the **comma** clears the memory, which is needed if you have used other data sets before
- Path is the path where STATA can find your data, e.g. **C:\\Users\...\Documents\STATA\ Projectseminar_2018**

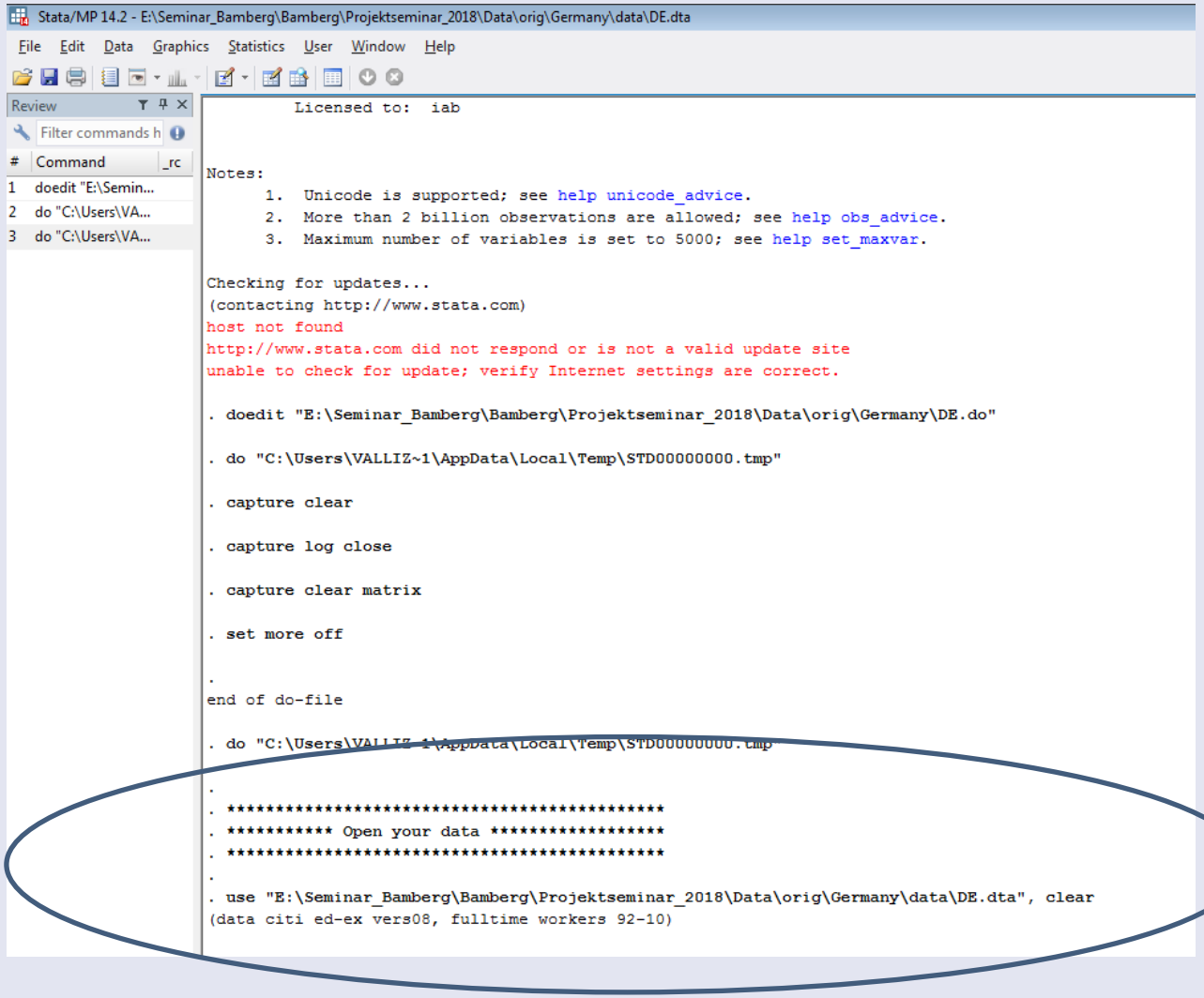
Step 2: Loading your data



1. Write the command **use „path\XXX.dta“, clear**
2. Mark the line and run the command by clicking the execution button

Step 2: Loading your data

University of Bamberg



Loading your data from EXCEL

University of Bamberg

- If you have to import data e.g. from an Excel file, use the command:

import excel using path\data\de.xlsx, firstrow

- the **import excel** tells STATA that it has to import a file with a different data format, in this case excel
- **using path\data\de.xlsx** tells data where to find the data and the file name
- The **option** after the comma **, firstrow** tells STATA that it has to treat the first row of the Excel sheet as the names (labels) of the variables. Otherwise it thinks it are data and you end up in a mess.

Step 3: Ending your session

University of Bamberg

- **Save** your **Do-file,** e.g.: ***save uk.do***
- Close your **Log-file: *log close***

- **<u>Important note:</u>** in normal cases **<u>don't save</u>** your **Data-file**. If you do this, all changes in your data set are saved and the original data set might be destroyed. You **cannot restore** the old dataset once it is saved.
- Of course, if you want to save some changes of your dataset, you can save it.

Processing your data

- Your can generate new variables and replace existing ones
- E.g. generate a numerical variable by using the information from a STRING variable
- **gen ed = .** generates the variable ed with missing values in the first place
- In the next step you can replace the values of this variables by using
- **replace ed = 1 if education ==“no voc training”**
- Which assigns the varianble a value of 1 if the person/group has no vocational training

Generating new variables

- Command:
  replace ed = 1 if education == "no voc training"
- **replace** tells STATA to replace the values of the variable, in this case of the ed variable by 1
- the **if option** tells STATA under which conditions, note that you have to use double equality sign (==) after the if option
- The **"…"** in **"no vocational training"** tells STATA that we have a STRING variable
- Then repeat this until all values of your variable are filled

Transform your data by using operators

- Useful operators in STATA:
  - \+ add
  - \- subtract
  - * multiply
  - / divide
  - ln transform into natural log
  - exp transform into exponential value

Generate Dummy Variables

Example:

- Reconsider Borjas (2003) model
- Why dummy variables?
- How to create dummy variables
- Advanced techniques to create dummy variables

Recall the Borjas (2003)-Modell

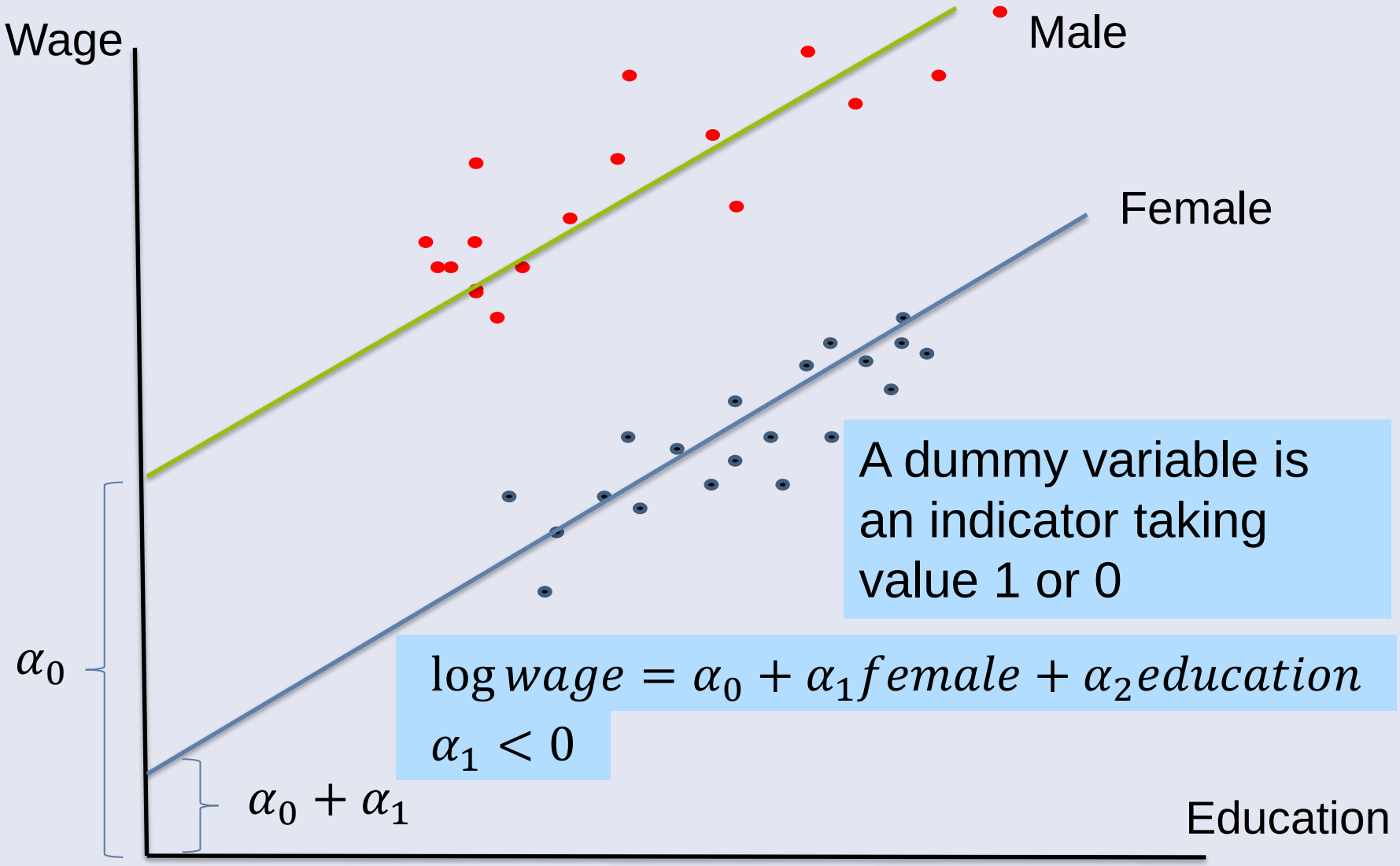
$$y_{ijt} = \theta p_{ijt} + s_i + x_j + \pi_t + (s_i \times x_j) + (s_i \times \pi_t) + (x_j \times \pi_t) + \varphi_{ijt},$$

where

- y_{ijt} is the dependent variable (e.g. log wage, unemployment rate)
- s_i is an education dummy
- x_j is an education dummy
- pi_j is a time dummy
- plus many interaction dummies

- Thus, we have to create quite a bunch of dummy variables.
- But, in the first place, what are dummy variables doing?

The Role of Dummy Variables



A dummy variable is an indicator taking value 1 or 0

$$\log wage = \alpha_0 + \alpha_1 female + \alpha_2 education$$

$$\alpha_1 < 0$$

How to generate dummy variables?

- Generating DUMMY variables
  - Use the **gen** command, e.g.
    - **gen Ded1 = 0**
  - This creates a variable consisting only of zeros
  - Then use the **replace** command, e.g.
    - **replace Ded1 = 1 if ed == 1**
  - This replaces the zeros with 1 if the variables ed1 has a values of 1.

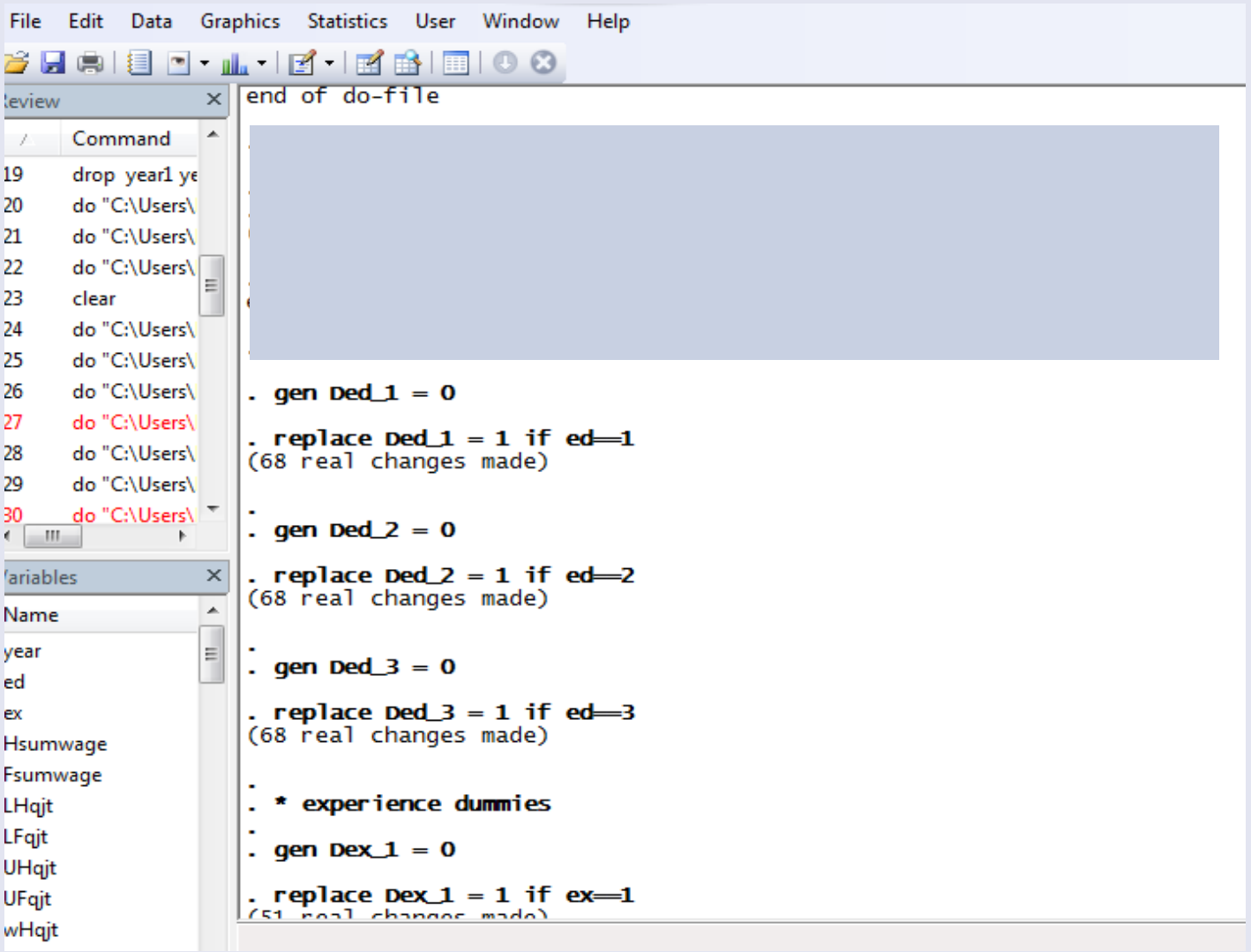
Generating Dummy Variables: DO FILE commands

University of Bamberg

```
*************************************************
***** generate Dummy variables ******************
*************************************************

* education dummies

gen Ded_1 = 0
replace Ded_1 = 1 if ed==1

gen Ded_2 = 0
replace Ded_2 = 1 if ed==2

gen Ded_3 = 0
replace Ded_3 = 1 if ed==3

* experience dummies

gen Dex_1 = 0
replace Dex_1 = 1 if ex==1

gen Dex_2 = 0
replace Dex_2 = 1 if ex==2

gen Dex_3 = 0
replace Dex_3 = 1 if ex==3

gen Dex_4 = 0
replace Dex_4 = 1 if ex==4

* time dummies

gen year_1 = 0
replace year_1 = 1 if year == 1992

gen year_2 = 0
replace year_2 = 1 if year == 1993
```

Generating Dummy Variables: STATA main window

University of Bamberg



Generating Dummy Variables

- Another example for generating dummy variables:
  - Use the **gen** command, e.g.

    gen Dt1 = 0

  - This creates a variable consisting only of zeros
  - Then use the **replace** command, e.g.

    Dt1 = 1 if year == 1991

  - This replaces the zeros with 1 if the year variable has a values of 1991
  - Note: The STATA syntax requires that you have to use after an **if** command always a double **==** for the definition of the value

Generating dummy variables: Advanced techniques

- Creating series of dummy variables if it is too cumbersome to create them individually, e.g. in case of interaction dummies. Use "**forvalues**" command
- Syntax:
  - ***forvalues i = 1/3 {***
    forvalues j = 1/4{
    gen D_ed`i'*D_ex`j'
    } }
  - i.e. for each value i = 1,2,3 and each value j = 1,2,3,4 you generate an interaction dummy by multiplying the dummy variables for education and experience. Take care of the {}!

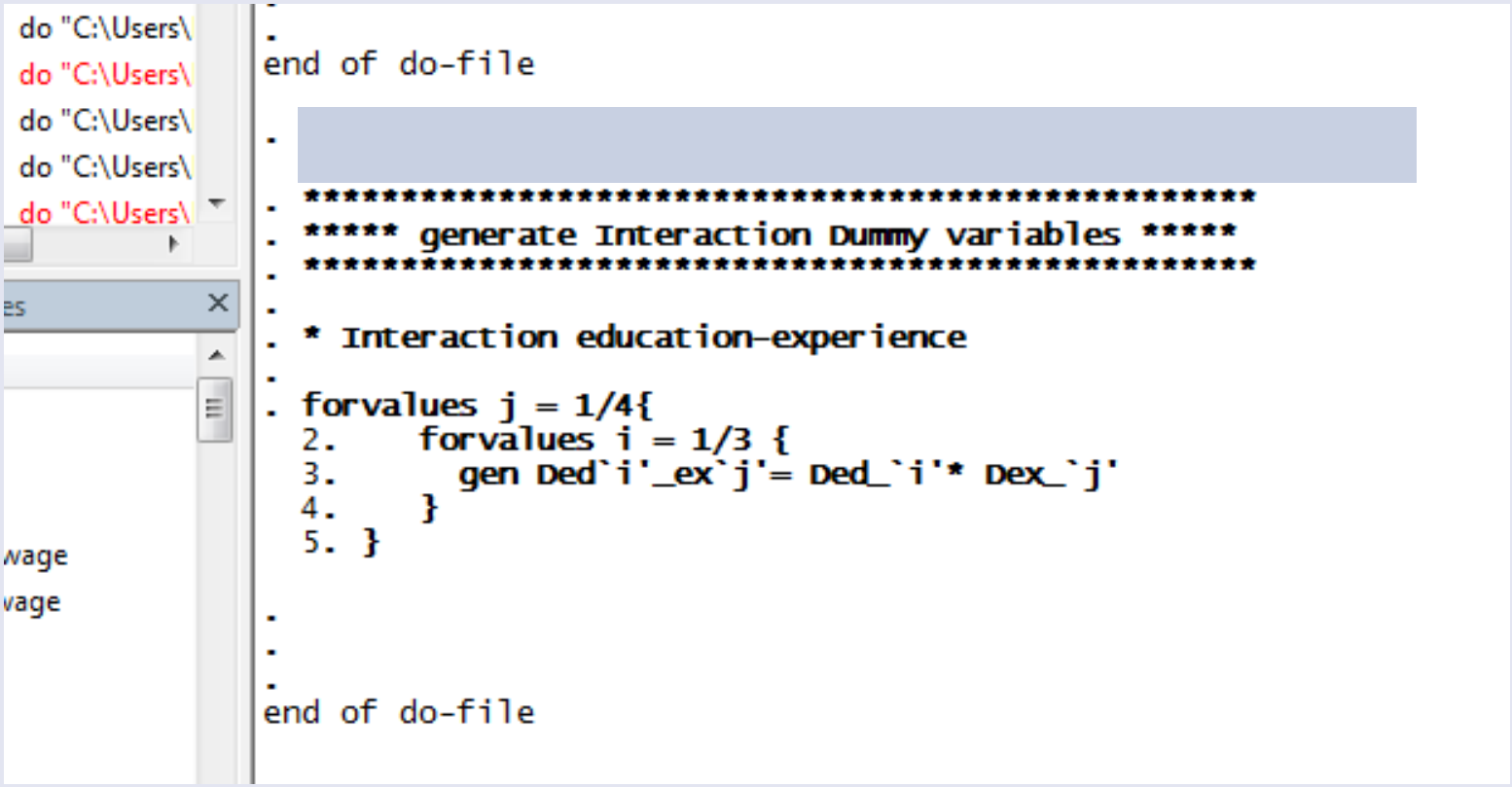
Generating Dummy Variables: Advanced techniques

University of Bamberg

```
*****************************************************
***** generate Interaction Dummy variables *****
*****************************************************

* Interaction education-experience

forvalues j = 1/4{
   forvalues i = 1/3 {
     gen Ded`i'_ex`j'= Ded_`i'* Dex_`j'
   }
}


* Interaction education-time
forvalues i = 1/3{
   forvalues t = 1/17 {
     gen Ded`i'_y`t'= Ded_`i'* year_`t'
   }
}
```

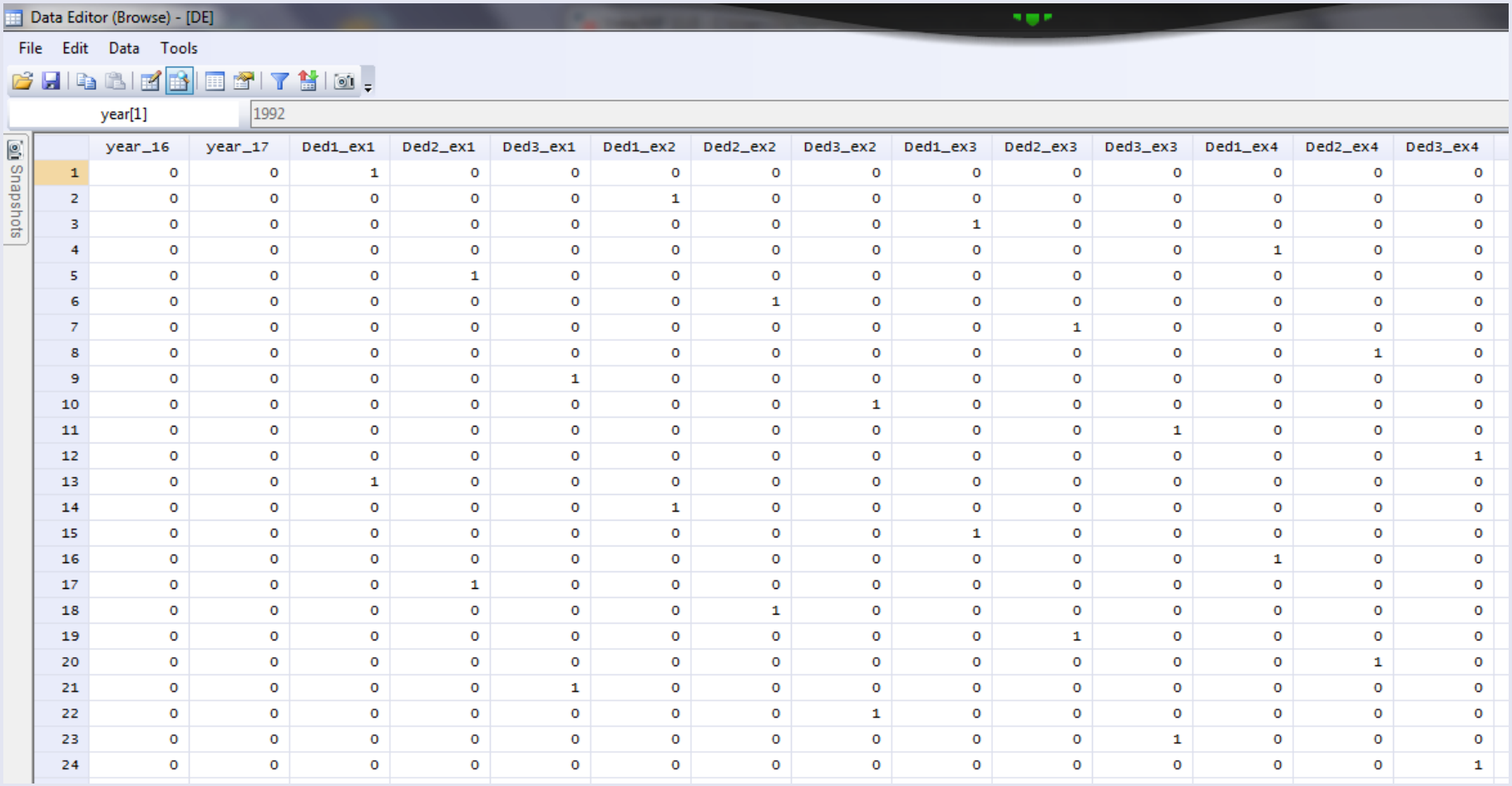
Generating Dummy Variables: Advanced techniques

University of Bamberg



Generating Dummy Variables: Advanced techniques

University of Bamberg



Data Editor (Browse) - [DE]

File Edit Data Tools

year[1] 1992

Snapshots

	year_16	year_17	Ded1_ex1	Ded2_ex1	Ded3_ex1	Ded1_ex2	Ded2_ex2	Ded3_ex2	Ded1_ex3	Ded2_ex3	Ded3_ex3	Ded1_ex4	Ded2_ex4	Ded3_ex4
1	0	0	1	0	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	1	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	1	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	0	1	0	0
5	0	0	0	1	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	1	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0	1	0	0	0	0
8	0	0	0	0	0	0	0	0	0	0	0	0	1	0
9	0	0	0	0	1	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	1	0	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0	0	1	0	0	0
12	0	0	0	0	0	0	0	0	0	0	0	0	0	1
13	0	0	1	0	0	0	0	0	0	0	0	0	0	0
14	0	0	0	0	0	1	0	0	0	0	0	0	0	0
15	0	0	0	0	0	0	0	0	1	0	0	0	0	0
16	0	0	0	0	0	0	0	0	0	0	0	1	0	0
17	0	0	0	1	0	0	0	0	0	0	0	0	0	0
18	0	0	0	0	0	0	1	0	0	0	0	0	0	0
19	0	0	0	0	0	0	0	0	0	1	0	0	0	0
20	0	0	0	0	0	0	0	0	0	0	0	0	1	0
21	0	0	0	0	1	0	0	0	0	0	0	0	0	0
22	0	0	0	0	0	0	0	1	0	0	0	0	0	0
23	0	0	0	0	0	0	0	0	0	0	1	0	0	0
24	0	0	0	0	0	0	0	0	0	0	0	0	0	1

Log transformation

- Econometrics is linear, while the real world is non-linear, e.g. concave, convex
- Replacing linear variable into log variable allows estimating non-linear realtionships between two variables
- Transforming variables into log variables
- Syntax:

gen ln_wijt = log(wijt)

- By using again the gen command you can transform the wage variable wijt into the natural logarithm of the wage by applying the **ln** operator

Log transformation

- Econometrics is linear, while the real world is non-linear, e.g. concave, convex
- Replacing linear variable into log variable allows estimating non-linear relationships between two variables
- Moreover, if the dependent and the independent variable are transformed into **natural logs**, we have a nice **interpretation** of the **coefficient**: it is the **elasticity**, i.e. it tells you by how many per cent the dependent variable changes if the independent variable changes by 1 per cent

Transforming data

University of Bamberg

```
************************************
*********** log wages *************
************************************

gen ln_wqjt = log(wqjt)
gen ln_wqt = log(wqt)
gen ln_wt = log(wt)

gen ln_wHqjt = log(wHqjt)
gen ln_wFqjt = log(wFqjt)
gen ln_wHqt = log(wHqt)
gen ln_wFqt = log(wFqt)
gen ln_wHt = log(wHt)
gen ln_wFt = log(wFt)
```

```
.
end of do-file
.
.
. gen ln_wqjt = log(wqjt)
. gen ln_wqt = log(wqt)
. gen ln_wt = log(wt)
.
. gen ln_wHqjt = log(wHqjt)
. gen ln_wFqjt = log(wFqjt)
. gen ln_wHqt = log(wHqt)
. gen ln_wFqt = log(wFqt)
. gen ln_wHt = log(wHt)
. gen ln_wFt = log(wFt)
.
end of do-file
.
```

Organize your data with ***macros***

University of Bamberg

- It is not convenient if you have to work with too many variables, e.g. 200 dummy variables (that is cumbersome to type some by hand)
- You can define globals, which comprise many variables
- Syntax:
  - **glo [name of global [list of variables]**
  - **glo Di Ded_1 Ded_2 D_ed3**
- i.e the global **Di** consists of the variables **Ded_1 Ded_2** and **Ded_3**
- If you want to use the global later you have to type
  - **$[globalname]**, i.e. **$Di**

Creating *globals*

University of Bamberg

```
************************************************************
***** Define globals (short-cuts) for Dummy variables *****
************************************************************

glo D_i Ded_2 Ded_3

glo D_j Dex_2 Dex_3 Dex_4

glo D_t year2 year3 year4 year5 year6 year7 year8 year9 /*
      */year10 year11 year12 year13 year14 year15 year16 year17


glo D_ij                   Ded1_ex2 Ded1_ex3 Ded1_ex4 /*
             */ Ded2_ex1 Ded2_ex2 Ded2_ex3 Ded2_ex4 /*
             */ Ded3_ex1 Ded3_ex2 Ded3_ex3 Ded3_ex4


glo D_1j      Ded1_ex2 Ded1_ex3 Ded1_ex4
glo D_2j      Ded2_ex2 Ded2_ex3 Ded2_ex4
glo D_3j      Ded3_ex2 Ded3_ex3 Ded3_ex4



glo D_it                  Ded1_y1993 Ded1_y1994 Ded1_y1995 Ded1_y1996 Ded1_y1997 Ded1_y1998 Ded1_y1999 /*
         */  Ded1_y2000 Ded1_y2001 Ded1_y2002 Ded1_y2003 Ded1_y2004 Ded1_y2005 Ded1_y2006 Ded1_y2007 Ded1_y2008 /*

         */  Ded2_y1992 Ded2_y1993 Ded2_y1994 Ded2_y1995 Ded2_y1996 Ded2_y1997 Ded2_y1998 Ded2_y1999 /*
         */  Ded2_y2000 Ded2_y2001 Ded2_y2002 Ded2_y2003 Ded2_y2004 Ded2_y2005 Ded2_y2006 Ded2_y2007 Ded2_y2008 /*

         */  Ded3_y1992 Ded3_y1993 Ded3_y1994 Ded3_y1995 Ded3_y1996 Ded3_y1997 Ded3_y1998 Ded3_y1999 /*
         */  Ded3_y2000 Ded3_y2001 Ded3_y2002 Ded3_y2003 Ded3_y2004 Ded3_y2005 Ded3_y2006 Ded3_y2007 Ded3_y2008
```

Descriptive statistics

- Any econometric analysis requires in the first step that you provide descriptive statistics to the reader. This helps to understand what's going on
- This can be easily done with the **sum** command

  sum [variable name(s)]

  sum LHijt LFijt wijt ln_wijt
- The sum command creates a table with the complete descriptive statistics, i.e. observations, mean, standard deviation, minimum, maximum

Descriptive statistics

University of Bamberg

```
************************************************
*********** Descriptive statistics *************
************************************************

sum LHqjt LFqjt mqjt wqjt ln_wqjt uqjt ln_uqjt
```

```
. sum LHqjt LFqjt mqjt wqjt ln_wqjt uqjt ln_uqjt

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
       LHqjt |        204    77113.39    101645.4        3758     382689
       LFqjt |        204    8496.632    8942.661         674      36526
        mqjt |        204    .1476157    .1279244    .0360193   .4374263
        wqjt |        204     97.5937    33.26997    47.94841   159.2809
     ln_wqjt |        204    4.525858    .3289621    3.870126   5.070669
-------------+---------------------------------------------------------
        uqjt |        204    .1632261    .1232984     .028542   .6573367
     ln_uqjt |        204   -2.060048    .7025635    -3.55638   -.4195589

.
end of do-file
```

Making graphs

- Present your data graphically
- It is usually helpful if you present the main information /variables in your data set graphically
- There are many graphical commands, use the **Graphics** menu or the respective commands

Making graphs

- The simplest way is to show the development of your variable(s) over time
- Syntax:

  graph twoway line [variable1] [variable2] if …
  graph twoway line wqjt year if ed==1 & ex == 1

- This produces a two-dimensional graph with the wage on the vertical and the year on the horizontal axis for education group 1 and experience group 1

Making graphs

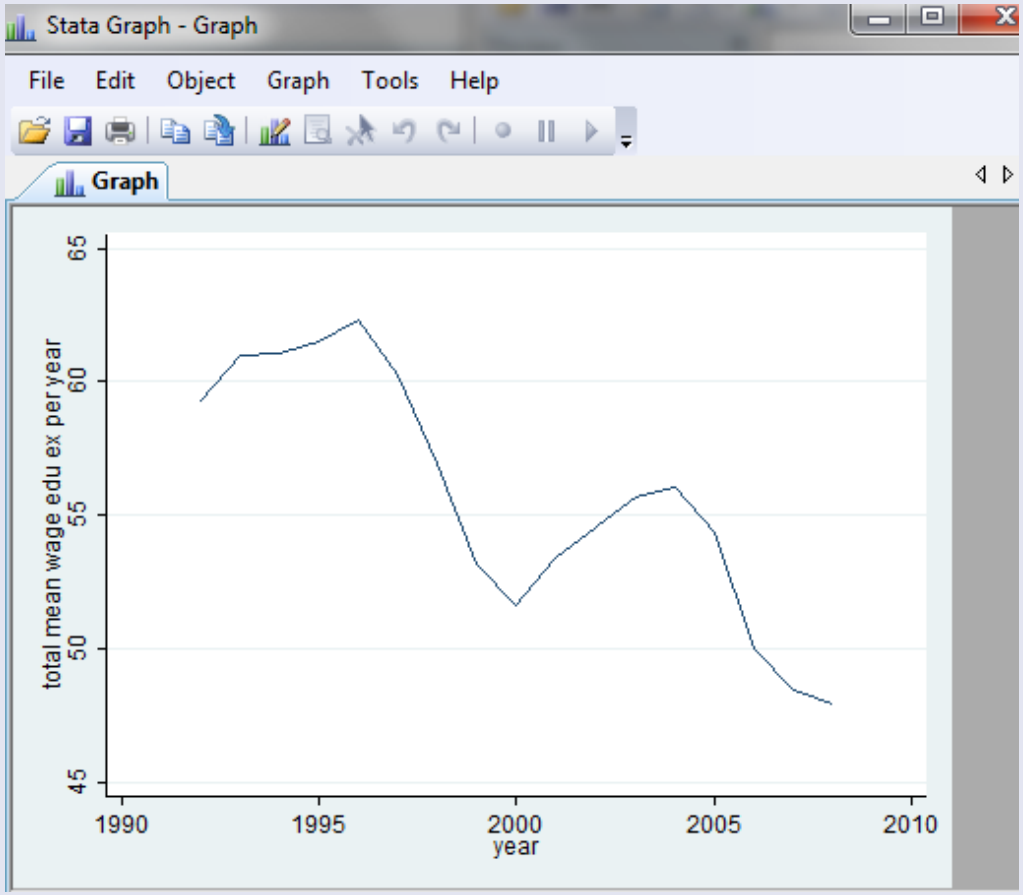
University of Bamberg

- Graph of mean wage in education 1 and experience 1

```
155
156
157  ********************************************************
158  ********** Graphs *************
159  ********************************************************
160
161
162  graph twoway line wqjt year if ed==1 & ex ==1
163
164
165  graph twoway line mqjt year if ed==1 & ex ==1
166
167
```

Making graphs

Example: line graph



Graphs with two Y-axes

- Making graphs with two variables on the y axis displayed on two y axes
- Syntax:
  - **graph twoway line ([variable1] [variable2]) ([variable 3][variable2], yaxis(2)) if …**
  - **graph twoway line (wqjt year) (mqjt year) if ed==1 & ex == 1**
- This produces a two-dimensional graph with the wage and the migration rate on the vertical axes with different scales and the year on the horizontal axis for education group 1 and experience group 1

Scatter plots

- Making graphs with scatter plots which show by point clouds the relationship between two variables
- Syntax:
  - **graph twoway scatter [variable1] [variable2] if …**
  - **graph twoway scatter wqjt mqjt)**
    if ed==1 & ex == 1
- This produces a two-dimensional scatter plot with a cloud of points which show for each observation in the sample of education group 1 and experience group 1 the values of the wage and the migration rate

Scatter plots with a regression line

- Making graphs with scatter plots which show by point clouds and a simple regression line the correlation (no causality) between two variables
- Syntax:
  - **graph twoway scatter [variable1] [variable2] || lfit [variable1] [variable2] if …**
  - **graph twoway scatter wqjt mqjt || lftit wqjt mqjt if ed==1 & ex == 1**
- This produces a two-dimensional scatter plot plus a regression line with a cloud of points which show for each observation in the sample of education group 1 and experience group 1 the correlation between the wage and the migration rate

Next Meting: June 15

- Begin: 12:00 – 14:00
- Topic: STATA II - Regression Analysis
- **All meetings are compulsory.**

THANKS FOR YOUR ATTENTION!