Programming, Data Management and Visualization

Alexander Ahammer*

β version
Last updated: Monday 18th January, 2021
Please check this document regularly

1. Goals

In this class you will learn advanced concepts in programming and data management using the statistical software package Stata. We focus on Stata because almost all subsequent courses in the econometrics curriculum use this software. However, note that many topics we cover are highly relevant for any statistical programming suite, even though the commands and concepts may differ slightly. However, Stata is not object-oriented as most other common languages, such as `R` or Python, but rather procedural or function-oriented (which makes it also much easier to learn). Upon successful completion, you are capable to handle Stata and understand data management at a level required for the subsequent courses in the JKU econometrics curriculum.

We start by discussing how to set up projects, arrange codes, and how to work with functions, macros, scalars, and matrices. Based on these preliminaries we will cover topics in data management (specifically how to combine, reorganize, and clean data), programming (e.g., how to use loops), as well as data analysis and visualization. Interested students will be provided with material teaching them to program an OLS estimator with Mata.

A specific emphasis is placed on econometrics and big data. Most examples and exercises will be based on exemplary data from the *Austrian Social Security Database.* We will discuss several particularities in handling big data sets throughout the course of the lecture.

2. Prerequisites

This class is held as a standard KS worth 4 ECTS. Ideally it should be taken in the first year by all *Economics* or *Management and Applied Economics* students, indeed students from other MSc programs and motivated BSc students are highly welcome as well. Economics PhD students may be able to use this class for their generic competences, make sure to confirm this with your advisor before registering.

Importantly, basic knowledge in econometrics and statistics at the level of *Intermediate Econometrics* (formerly *Econometrics I*) is a prerequisite. If you don't have prior knowledge in statistical programming, I advise you to invest time in learning Stata before the class starts. It is a

*Department of Economics, Johannes Kepler University Linz, ✉ alexander.ahammer@jku.at. I do not offer regular office hours, if you require an appointment please contact me via email.

very intuitive language with a steep learning curve (especially if you know other programming languages).

3. Curriculum change

The introduction of this course came along with a change in the Master curriculum. For all students who took at least one course in the module *'Methods in Economics'* before October 1, 2018 this course is **not compulsory.** For all others **it is.**

4. Topics

This is a coarse list of all topics we will cover this semester. Topics may change as we go along. If you are interested in additional contents (e.g., for your thesis), please let me know as soon as possible — there may be time to cover them.

(A) **Elementary concepts and data organization** [slides]

How to set up and organize a project, replicability, data types, memory, importing and exporting data.

(B) **Programming preliminaries** [slides]

Writing do-files efficiently using lists, logical qualifiers, strings, observation numbering; functions, macros, scalars, and matrices; loops.

(C) **Data management** [slides]

Data validation; reorganizing and combining datasets, useful data management commands.

(D) **Reporting results** [slides]

Store, save, and reuse computed results; automate reporting of estimation output and graphs, produce publication-ready tables with Stata.

(E) **Data analysis and visualization** [slides]

Summary statistics, cross-tabulations, graphs, geographical maps.

Don't print the slides until right before the corresponding lecture, because they are updated continuously. Each slide set has a time stamp for reference.

5. Organization and Grading

Lectures are delivered in weekly meetings. Due to COVID, classes will be offered exclusively online for now. Every two weeks, students have to solve a short problem set and send it in via email to me. The class concludes with a take-home exam at the end of the semester. The schedule can be found in Table 1. Please check this syllabus regularly, as links to slides and other materials will be published here.

Grading follows the Austrian scholastic system, with 1 or *Sehr Gut* being the best and 5 or *Nicht Genügend* being the worst and only failing grade. There are 5 weekly problem sets à 6 points, amounting to a total of 30 points. The take-home final exam is worth 30 points too. In total you can therefore reach 60 points, with at least 31 points required to pass the course (with 38 points or more you get a *'Befriedigend,'* 46 or more a *'Gut,'* and 54 or more a *'Sehr Gut'*).

Table 1: Course schedule (will be updated regularly, check back here).

W	Date	Time	Venue	Topic	Slides	Problem Sets
0	Mo, 05.10.2020	13:45–14:15	Zoom	Introduction	Intro	
1	Mo, 12.10.2020	13:45–15:15	Zoom	Module A	[A]	
2	Mo, 19.10.2020	13:45–15:15	Zoom	Module A	[A]	PS1
3	Mo, 09.11.2020	13:45–15:15	Zoom	Module B	[B]	
	Mo, 16.11.2020					PS2
4	Mo, 23.11.2020	13:45–15:15	Zoom	Module B	[B]	
	Mo, 30.11.2020					PS3
5	Mo, 07.12.2020	13:45–15:15	Zoom	Module C	[C]	
8	Mo, 14.12.2020	13:45–15:15	Zoom	Module C	[C]	PS4
8	Mo, 11.01.2021	13:45–15:15	Zoom	Module D	[D]	
9	Mo, 18.01.2021	13:45–15:15	Zoom	Module E	[E]	PS5
10	Mo, 25.01.2021	13:45–15:15	Zoom	Module E	[E]	

6. TAKE-HOME EXAM

The take-home exam accounts for the other half of your grade. It will essentially be a longer problem set; based on transforming, organizing, and analyzing real-world data. The exam requires you to apply methods we have learned during the semester, but there will also be new challenges you have not yet encountered, allowing you to demonstrate your acquired problem solving skills.

The take-home exam should be submitted in the form of a seminar paper, where you verbally explain every piece of your code and interpret each table and graph you generate. There will be more than one exam sheet, and you will be randomly assigned one of them. You will work alone on your exam, cheating (especially copying from others) will be *rigorously punished.* I reserve the right to invite students for an oral exam if there is reasonable suspicion of academic fraud.

The take-home exam will be uploaded on **January 26, 2021** and is due **March 1, 2021, midnight sharp.**

7. LITERATURE

The main reference for this class is my slide set. However, the structure is loosely based on

> Christopher F. Baum (2009), 'An Introduction to Stata Programming,' first or second edition, Stata Press, College Station, Texas.

In the future I may provide additional references for modules E (data analysis and visualization) and F (programming an estimation command). A very interesting read is *Code and Data for the Social Sciences: A Practitioner's Guide* by Matthew Gentzkow and Jesse Shapiro (both Chicago Booth), I recommend everybody who writes an empirical econ thesis or aspires a career in academia to study this: [Link]

Check also these **Stata cheat sheets,** they are extremely helpful: [Link]

8. DATA DECLARATION

As a reminder, here is the data declaration you signed:

> I, *[your name]*, will receive different random samples drawn from the *Austrian Social Security Database* and the *Upper Austrian Health Insurance Fund Database* to solve exercises for the 2020/'21 Johannes Kepler University Linz class *Programming, Data Management and Visualization* taught by Alexander Ahammer. I hereby declare, that I will
>
> – not search for individual-level information in the data,
>
> – delete the datasets immediately and completely after the end of the course, and
>
> – make sure, that none of these data will be passed on to third parties.
>
> I am aware that I am liable to prosecution if I violate any point of this declaration.

You can find all datasets in this password-protected Dropbox folder: `[Link]`