Tutorial 1, Discussion on April 19th
Introduction to STATA[*]
OLS Recap Exercises

Empirical Banking and Finance
Konrad Adler

Summer, 2022

This tutorial has two parts: we start with a small example in order to provide a first look at the Stata environment and the way it works. In order to follow the example, you first need to download the dataset caschool.dta from eCampus. The ending .dta shows that we are dealing with a Stata dataset. In the second part, we will use the same dataset two illustrate some properties of OLS regressions discussed in class.

Data description [1] The California Standardized Testing and Reporting (STAR) dataset contains data on test performance, school characteristics and student demographic backgrounds. The data used here are from all 420 K-6 and K-8 districts in California with data available for 1998 and 1999.

Test scores are the average of the reading and math scores on the Stanford 9 standardized test administered to 5th grade students. School characteristics (averaged across the district) include enrollment, number of teachers (measured as fulltime- equivalents"), number of computers per classroom, and expenditures per student. The student-teacher ratio used here is the number of full-time equivalent teachers in the district, divided by the number of students. Demographic variables for the students also are averaged across the district. The demographic variables include the percentage of students in the public assistance program CalWorks (formerly AFDC), the percentage of students that qualify for a reduced price lunch, and the percentage of students that are English Learners (that is, students for whom English is a second language). All of these data were obtained from the California Department of Education (www.cde.ca.gov).

[*] Thanks to Ulrich Schüwer for providing the Tutorial

[1] This is the only non-finance related dataset used in this class

1 Introduction to STATA

Series in Data Set:

```
DIST_CODE: DISTRICT CODE;
READ_SCR: AVG READING SCORE;
MATH_SCR: AVG MATH SCORE;
COUNTY : COUNTY;
DISTRICT: DISTRICT;
GR_SPAN: GRADE SPAN OF DISTRICT;
ENRL_TOT : TOTAL ENROLLMENT;
TEACHERS: NUMBER OF TEACHERS;
COMPUTER: NUMBER OF COMPUTERS;
TESTSCR: AVG TEST SCORE (= (READ SCR+MATH SCR)/2 );
COMP_STU: COMPUTERS PER STUDENT ( = COMPUTER/ENRL TOT);
EXPN_STU: EXPENTITURES PER STUDENT ($'S);
STR: STUDENT TEACHER RATIO (ENRL TOT/TEACHERS);
EL_PCT: PERCENT OF ENGLISH LEARNERS;
MEAL_PCT: PERCENT QUALIFYING FOR REDUCED-PRICE LUNCH;
CALW_PCT: PERCENT QUALIFYING FOR CALWORKS;
AVGINC: DISTRICT AVERAGE INCOME (IN $1000'S);
```

If you are not familiar with Stata, have a look at the Stata Tutorial provided in the extra file.

2 OLS Recap Exercises

1. Data Preparation: Run a regression with testscr on the LHS and avginc on the RHS
   (a) Plot the regression line along with the data.
   (b) What do you notice? What would be an easy fix for this problem?
   (c) Plot the regression line with your problem fix.
2. Standard Errors: Run a regression with testscr on the LHS and avginc on the RHS
   (a) Without any adjustment to standard errors
   (b) Using the robust option
   (c) Using the vce(hc2) option
   (d) Compare the standard errors and provide a brief comment.
3. Verify the basic OLS formula $\hat{\beta}_1 = cov(testscr, avginc)/var(avginc)$ without using the reg command.
4. Someone tells you to include avginc multiplied by 2 in the regression in question 1.
   (a) What is the problem with including both avginc and avginc*2? Try running the regression.
   (b) Run the regression with avginc and avginc*2 separately with testscr as LHS variable. How does the coefficient change? What happens when you include avginc*3 or avginc*4?
   (c) What happens when you scale testscr by 2,3, or 4 and leave avginc unchanged?
5. Replication of the two-step procedure to estimate the coefficient of avginc shown in class.
   (a) Run a regression testscr on the LHS and avginc, teachers, and computer on the RHS
   (b) Obtain the same coefficient for avginc using the two-step procedure.
6. Anatomy of the OLS regression coefficient
   (a) Run a regression with testscr on the LHS and computer_discrete on the RHS.
   (b) Generate a variable called “obs” which just equals 1. Then, use the command “collapse” to compute average testscr and the sum of “obs” for each level of computer_discrete.
   (c) Run a regression with average testscr (by level of computer_discrete) on the LHS and computer_discrete on the RHS.
   (d) Run the same regression as in c) but weigh each observation by “obs” by adding [aw=obs]
   (e) Compare the coefficients of a), c), and d). What is the anatomy of an OLS regression coefficient?