Dr. Markus Haas
Department of Empirical Research and Econometrics

TSP Exercise Session - Problem Set 9

Multiple Regression II

Preparations

Please create a new folder for this exercise session with your name in directory T:. Then go to L:\Intermediate Econometrics\PC3 and copy the files into your folder.

1) The Munich Rent Index

The data set in "hpm.dat" contains information on housing prices in Munich from the Munich Rent Index. We will analyze how the net rent depends on the following variables:

nm	net rent in EUROs
wfl	living space in m^2
$badextra$	specially equipped bathroom (1=yes, 0=no)
zh	central heating available (1=yes, 0=no)

(a) Estimate the following model:

$$nm = \beta_0 + \beta_1 \cdot zh + \beta_2 \cdot wfl + \beta_3 \cdot wfl^2 + \beta_4 \cdot badextra + \beta_5 \cdot wfl \cdot bad + u$$

How does the net rent depend on the living space and on the existence of a specially equipped bathroom? Are those effects statistically significant?

(b) Create a plot placing living space in m^2 on the x-axis and the residuals from the estimation in subquestion (a) on the y-axis. How do you interpret the plot?

(c) Can you detect further indication of the above-mentioned problem in the TSP output of your estimation? If there existed such a problem, would your results from subquestion (a) still be valid? Which solutions could you think of?

(d) Repeat the estimation from subquestion (a) for the **logarithmic** net rent. How does the interpretation of the estimated coefficients change? What about the above-mentioned problem?

2) College grade point average (Wooldridge, Chapter 7)

Use the data in “gpa2.raw” for this exercise and consider the following equation:

$$\begin{aligned} colgpa = \beta_0 + \beta_1 \cdot hsize + \beta_2 \cdot hsize^2 + \beta_3 \cdot hsperc + \beta_4 \cdot sat \\ + \beta_5 \cdot female + \beta_6 \cdot athlete + u, \end{aligned}$$

where $colgpa$ is cumulative college grade point average, $hsize$ is size of high school graduating class, $hsperc$ is academic percentile in graduating class, sat is combined SAT score, $female$ is a gender dummy (1=female, 0=male), and $athlete$ is a binary variable with 1=student-athletes and 0=student-nonathlets.

(i) What are your expectations for the coefficients in this equation? Which ones are you unsure about?

(ii) Estimate the equation from above and report the results in the usual form. What is the estimated GPA differential between athletes and nonathletes? It is statistically significant?

(iii) Drop sat from the model and reestimate the equation. Now, what is the effect of being an athlete? Discuss why the estimate is different than that obtained in part (ii).

(iv) In the model for $colgpa$, allow the effect of being an athlete to differ by gender and test the null hypothesis that there is a no ceteris paribus difference between women athletes and women nonathletes.

(v) Does the effect of sat on $colgpa$ differ by gender? Justify your answer.

3) Multicollinearity and Partitioned Regression

Consider the following regression model:

$$y_i = 1 + 0.15 \cdot x_i + 0.15 \cdot z_i + u_i$$

with:

- u_i and $x_i \sim N(0, 1)$
- $z_i = \lambda \cdot x_i + v_i$, $\lambda = const.$, $v_i \sim N(0, 1)$

(a) Generate a data set (N=100) with $\lambda = 3.042$. How large is the expected correlation between x and z?

(b) Estimate the model using OLS. Which problems might occur during the estimation? Which problems actually do occur in the case of your estimation?

(c) Estimate β_x by the use of partitioned regression.

Appendix: TSP-commands

do	**do** indexname **=** startvalue **to** endvalue **by** increment **;** program code **enddo ;** → specifies a conventional **loop**. The statements between the do and enddo statement are executed repetitively as many times as specified by the information given on the do statement.
dot	→ creates a “dot” loop, which is like a regular “do” loop, except that the values of the index are a series of character strings (names), e.g.: **dot** variable1 variable2 variable3 **;** genr new. = .*100 ; **enddot ;** → generates three new variables called newvariable1, newvariable2, etc.
random	**random (**mean=0,stdev=1**)** variable1**;** **random (**mean=10,poisson**)** variable2**;** etc. → draws a **random variable** which follows a certain distribution; if no options are specified the default setting is normal distribution