TSP Exercise Session - Problem Set 7

Multiple Regression I

Preparations

Please create a new folder for this exercise session with your name in directory T:. Then go to L:\Intermediate Econometrics\PC2 and copy the files into your folder.

1) Estimating the effects of smoking during pregnancy on infant health

(Based on Wooldridge, Computer Exercise C 3.1, p. 110)

A problem of interest to health officials (and others) is to determine the effects of smoking during pregnancy on infant health. One measure of infant health is birth weight; a birth rate that is too low can put an infant at risk for contracting various illnesses. Since factors other than cigarette smoking that affect birth weight are likely to be correlated with smoking, we should take those factors into account. For example, higher income generally results in access to better prenatal care, as well as better nutrition for the mother. An equation that recognizes this is:

$$bwght = \beta_0 + \beta_1 \cdot cigs + \beta_2 \cdot faminc + u$$

with **bwght** birth weight, ounces; **cigs** cigs smoked per day while pregnant; **faminc** 1988 family income, $ 1000s

(a) What are the most likely signs for β_1 and β_2?

(b) Do you think **cigs** and **faminc** are likely to be correlated? Explain why the correlation might be positive or negative.

(c) Now estimate the equation with and without **faminc**, using the data in "pc2_01.xls". Report the results in equation form, including the sample size and R^2. Discuss your results, focusing on whether adding **faminc** substantially changes the estimated effect of **cigs** on **bwght**.

(d) Generate a dummy variable indicating whether the mother has been smoking while pregnant ($dummy = 1$) or not ($dummy = 0$). Then estimate the following equation:

$$bwght = \beta_0 + \beta_1 \cdot dummy + \beta_2 \cdot faminc + u$$

(e) Compare the estimated β_1 with your results from part (c). Why do the estimated coefficients differ from each other? How can you interprete them?

2) Mincerian Earnings Equation

The data set "pc2_02.xls" contains labor-market-related information on 974 full-time employees from the year 1996. The data is taken from the German Socio-Economic Panel Study (SOEP). It contains an identification number (ID), gender (GESCHL: 0=male, 1=female), age (ALTER), number of years of education (BILDUNG), yearly gross wage (EINK), hours worked (STUNDEN), sector (SEKTOR: 1=primary, 2=secondary, 3=tertiary; OEFFD: 0=private, 1=public) and size of the firm (BGROESSE) as well as the employee's origin (SAMPLE: 1=West German, 2= East German, 3=guest worker).

We will analyze the factors influencing wages in Germany by estimating a Mincerian earnings function of the following form:

$$ln(w_i) = \beta_0 + \beta_1 \cdot X_i + \beta_2 \cdot X_i^2 + \beta_3 \cdot E_i + \gamma \cdot Z_i + \epsilon$$

with w_i wage rate; X_i work experience; E_i number of years of education; Z_i further variables; β_3 rate of return to education

(a) Generate the logarithmic wage rate per hour. Generate the potential work experience of the employees according to the following rule: potential work experience = age - number of years of education - 6. Run a regression of the logarithmic hourly wage rate on work experience and squared work experience (no other variables). Display the predicted wage profile graphically. After how many years of work experience do you earn the highest wage rate?

(b) Extend the model from part (a) by adding a gender dummy and years of education. Estimate the new model using OLS. Do women, on average, earn less than men?

(c) Now, calculate the average wage rates of men and women separately. Do they differ? Furthermore, do women with similar educational background and similar potential work experience earn less than men? To answer this question, run estimations for women and men separately. Why need the results not necessarily be caused by gender discrimination? Can you think of other possible explanations?

(d) Calculate the rate of return to education for West Germans and guest workers. Is the discrepancy statistically different from zero? (**Hint:** Look at the interaction effect.) Based on your results, would you say there exists discrimination against guest workers on the German labor market?

(e) Generate a dummy variable for the sector of the company. In which sector are the highest wage rates being paid? In which sector do full-time employees earn the lowest wage rates? Test the hypothesis that the sector dummies jointly do not have any influence on the wage rate.

3) Discrimination against certain customer groups by fast food restaurants

(Based on Wooldridge, Computer Exercise C 3.8, p. 112)

Use the data in "pc2_03.xls" to answer this question. These are ZIP code-level data on prices for various items at fast-food restaurants, along with characteristics of the zip code population, in New Jersey and Pennsylvania. The idea is to see whether fast-food restaurants charge higher prices in areas with larger concentration of blacks. The data set contains the following variables:

$psoda$	price of medium soda
$prpblck$	proportion black, zipcode
$prppov$	proportion in poverty, zipcode
$income$	median family income, zipcode

(a) Find the average values of **prpblck** and **income** in the sample, along with their standard deviations. What are the units of measurement of **prpblck** and **income**?

(b) Consider a model to explain the price of soda, **psoda**, in terms of the proportion of the population that is black and median income:

$$psoda = \beta_0 + \beta 1 \cdot prpblck + \beta 2 \cdot income + u$$

Estimate this model by OLS and report the results in equation form, including the sample size and R^2. Interpret the coefficient on **prpblck**. Do you think it is economically large?

(c) Compare the estimate from part (b) with the simple regression from **psoda** on **prpblck**. Is the discrimination effect larger or smaller when you control for income?

(d) A model with a constant price elasticity with respect to income may be more appropriate. Report the estimates of the model:

$$log(psoda) = \beta_0 + \beta 1 \cdot prpblck + \beta 2 \cdot log(income) + u$$

If **prpblck** increases by 0.20 (20 percentage points), what is the estimated change in **psoda**?

(e) Now add the variable **prppov** to the regression in part (d). What happens to $\hat{\beta}_{prpblck}$?

(f) Find the correlation between **log(income)** and **prppov**. Is it roughly what you expected?

(g) Evaluate the following statement: "Because **log(income)** and **prppov** are so highly correlated, the have no business being in the same regression."

4) Appendix: TSP-commands

cdf	CDF calculates and prints tail probabilities (P-values or significance levels) or critical values for several cumulative distribution functions, e.g.: **cdf (f, df1=**numerator degrees of freedom for F**, df2=**denominator degrees of freedom for F**)** test statistic **;** **cdf(t,df=**degrees of freedom for T**)** test statistic **;** etc.
frml analyz	**frml** testname1 variable1 **;** **frml** testname2 variable2-0.5 **;** **analyz(noconstr)** testname1 **;** → tests whether variable1 is statistically different from zero **analyz(noconstr)** testname2 **;** → tests whether variable2 is statistically different from 0.5 **analyz(noconstr)** testname1 testname2 **;** → tests for the two variables **jointly** → analyz **always** tests for the hypothesis, that the variables or equations defined by the frml command are equal to **zero**
olsq	TSP stores most of the results from an OLS regression in data storage for your later use. → e.g. the number of observations is stored in the variable **@nob**, the sum of squared residuals is stored in the variable **@ssr**, etc.
dummy	**dummy** variable variable1 variable2 variable3 ; **dummy** creates a set of dummy variables (variable1, variable2, and variable3) corresponding to the different values in the variable taken into account