Niclas Knecht (niclas.knecht@u-bordeaux.fr)
Econometrics
Spring 2024

Problem Set 2

Problem 1

Which of the following can cause OLS estimators to be biased?

a) Heteroskedasticity.

b) Omitting an important variable.

c) A sample correlation coefficient of 0.95 between two independent variables both included in the model.

Problem 2 (Stata)

Use the data set **td2_wages.dta** from Blackburn and Neumark (1992) for this exercise. It contains information on monthly earnings, education, several demographic variables, and IQ scores for 935 men in 1980. To account for omitted variable bias, we add IQ and KWW ("Knowledge of the world of work", a test score) to a standard log-wage (in the data as *lwage*) equation.
Our primary interest is in what happens to the estimated return to education:

$$log(wage) = \beta_0 + \beta_1\, educ + \beta_2\, exper + \beta_3\, tenure + \beta_4\, ability + u.$$

a) Estimate by OLS the wage equation without controlling for ability. What is the estimated return to education in this case?

b) Use the variable IQ as a proxy for ability. What is the estimated return to education in this case? Is it necessary to control for it?

c) Use the variable KWW as a proxy for ability instead of IQ. What is the estimated return to education in this case?

d) Now use IQ and KWW together as proxy variables. What happens to the estimated return to education? What do you conclude?

Problem 3

In a study relating college grade point average (GPA) to time spent in various activities, you distribute a survey to several students. The students are asked how many hours they spend each week in four activities: studying, sleeping, working, and leisure. Any activity is put into one of the four categories, so that for each student the sum of hours for the four activities must be equal to 168. Consider the model

$$GPA = \beta_0 + \beta_1\, study + \beta_2\, sleep + \beta_3\, work + \beta_4\, leisure + u.$$

a) Does it make sense in this model to hold *sleep*, *work*, and *leisure* fixed, while changing *study*?

b) Explain why this model violates the assumption of no perfect collinearity.

c) How could you reformulate the model so that its parameters have a useful interpretation and the model satisfies the assumption of no perfect collinearity?

Problem 4

Which (and if so, why) of the following are consequences of heteroskedasticity?

a) The OLS estimators, β_j, are inconsistent.

b) The usual F-statistic no longer has an F-distribution.

c) The OLS estimators are no longer BLUE.

Problem 5 (Stata)

Use the data in **td2_price.dta** and the following model of house prices:

$$price = \beta_0 + \beta_1 \, lotsize + \beta_1 \, sqrft + \beta_3 \, bdrms + u,$$

where *lotsize* is the size of the lot (land) in square feet, *sqrft* is the size of the house in square feet, and *bdrm* is the number of bedrooms.

a) Estimate the equation with both normal and heteroskedasticity-robust standard errors and discuss any important differences with the usual standard errors.

b) Create a scatterplot between the residuals and the fitted values for both cases. What can we infer from the graph?

c) Repeat a) but transforming the continuous variables into logarithms, such that the elasticities of *price* with respect to *lotsize* and *sqrft* are constant. Report your results.

d) What does this example suggest about heteroskedasticity and the transformation used for the dependent variable?

e) Apply the full White's test for heteroskedasticity (Note: you cannot use robust standard errors). Use the chi-square form of the statistic and obtain the p-value. What do you conclude?

f) Apply the Breusch-Pagan test for heteroskedasticity to the same equation. What do you infer?

Problem 6

a) True or False? If $y = x\beta + e$, $e \in \mathbb{R}$, and $\mathbb{E}[e|x] = 0$, then $\mathbb{E}[x^2 e] = 0$.

b) True or False? If $y = x\beta + e$, $e \in \mathbb{R}$, and $\mathbb{E}[ex] = 0$, then $\mathbb{E}[x^2 e] = 0$.

c) True or False? If $y = x\beta + e$, $e \in \mathbb{R}$, and $\mathbb{E}[e|x] = 0$, then e is independent of x.

Problem 7

Let $\hat{y} = (\hat{y}_1, ..., \hat{y}_n)'$ be a vector of fitted values from an OLS regression of y on some regressors including a constant. Coefficient of determination, or R-squared, is defined as follows:

$$R^2 = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2}.$$

a) Show that $\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n \hat{u}_i^2$. Interpret R^2.

b) Prove that $\hat{y}'\hat{y} = \hat{y}'y$.

c) Use the result from b) and that $\sum_{i=1}^n \hat{y}_i = \sum_{i=1}^n y_i$, to prove that R^2 (in the model with the constant term) is the square of the sample correlation between y and $\hat{y}$.

d) Consider two OLS regressions

$$y = \tilde{\beta}_0 + x_1 \tilde{\beta}_1 + \tilde{u},$$
$$y = \hat{\beta}_0 + x_1 \hat{\beta}_1 + x_2 \hat{\beta}_2 + \hat{u},$$

with R-squared equal to R_1^2 and R_2^2, respectively. Show that $R_2^2 \geq R_1^2$. Is there a case when there is equality, i.e., $R_2^2 = R_1^2$?

Problem 8 (Stata)

Using the dataset **td2_sales.dta** estimate the following model:

$$rdintens = \beta_0 + \beta_1 \, sales + \beta_2 sales^2 + u,$$

where the variable *rdintens* is expenditures on research and development (R&D) as a percentage of sales. Sales are measured in millions of dollars.

a) At what point does the marginal effect of *sales* on *rdintens* become negative?

b) Would you keep the quadratic term in the model? Explain.

c) Define *salesbil* as sales measured in billions of dollars: $salesbil = sales/1000$. Rewrite the estimated equation with *salesbil* and $salesbil^2$ as independent variables instead of *sales* and $sales^2$. Report your results, including standard errors and R-squared. (*Hint*: Note that $salesbil^2 = sales^2/1000^2$.)

d) For the purpose of reporting the results, which equation do you prefer?