Problem Set 4: OLS

Review the Concepts and Proofs

1. Show that $E(u|x) = 0$ implies $E(x'u) = 0$.
2. Explain the omitted variable bias.
3. State the consequences of the different types of measurement errors for estimation.
4. What does redundancy mean?
5. Derive an estimator for the variance of $\hat{\boldsymbol{\beta}}$ for the cases of homoscedastic and heteroscedastic errors. Which one would you prefer?
6. Discuss the difference between exogenous and endogenous variables. Give economic examples for both cases.
7. What does causality mean? Can you interpret OLS estimates as causal?

Exercises

1. Consider the linear model $y_i = \boldsymbol{x}_i\boldsymbol{\beta} + u_i$. Assume that the OLS assumptions hold and that the sample is random.

   (a) Explain the OLS assumptions briefly and state which ones you need to obtain (i) identification of $\boldsymbol{\beta}$, (ii) unbiasedness of $\hat{\boldsymbol{\beta}}_{OLS}$, (iii) consistency of $\hat{\boldsymbol{\beta}}_{OLS}$, and (iv) asymptotic normality of $\hat{\boldsymbol{\beta}}_{OLS}$.

   (b) Derive the OLS estimator in matrix notation. Show that $\hat{\boldsymbol{\beta}}_{OLS}$ is consistent and asymptotically normally distributed.

2. Consider the following model for the average grade of participants of one of the economic-master programs at the university of Kiel (higher average grade means student performs worse):

   $$avgrade_i = \beta_0 + \beta_1 numsem_i + \beta_2 worktime_i + \sum_{j=1}^{J} \gamma_j famground_{j,i} + \sum_{k=1}^{K} \delta_k country_{k,i} + u_i$$

   where $numsem_i$ is the number of semesters needed for completion of the program, $worktime_i$ denotes the average hours worked per week, $famground_{j,i}$ are J family background variables (e.g. financial situation, education of the parents), $country_{k,i}$ are K country dummies.

(a) The ability of a participant is likely to be correlated with *numsem*. Is β_1 likely to be upward or downward biased?

(b) How would you estimate this model to reduce the bias?

3. Consider a standard log(*wage*) equation for men under the assumption that all explanatory variables are exogenous:

$$\begin{aligned} &\log(wage) = \beta_0 + \beta_1 married + \beta_2 educ + \boldsymbol{z}\gamma + u, \\ &\mathrm{E}(u|married, educ, \boldsymbol{z}) = 0, \end{aligned} \tag{1}$$

where *married* is a dummy variable that takes value 1 if a man is married and 0 otherwise, *educ* denotes the number of years of schooling, and $\boldsymbol{z}$ contains factors other than marital status and education that can affect wage. Where β_1 is small, $100 \cdot \beta_1$ is approximately the ceteris paribus percentage difference in wages between married and unmarried men. When β_1 is large, it might be preferable to use the exact percentage difference in $\mathrm{E}(wage|married, educ, \boldsymbol{z})$. Call this θ_1.

(a) Show that, if u is independent of all explanatory variables in equation (1), then $\theta_1 = 100 \cdot [\exp(\beta_1) - 1]$.
Hint: Find $\mathrm{E}(wage|married, educ, \boldsymbol{z})$ for both $married = 1$ and $married = 0$ and find the percentage difference.

(b) A natural, consistent, estimator of θ_1 is $\hat{\theta}_1 = 100 \cdot [\exp(\hat{\beta}_1) - 1]$, where $\hat{\beta}_1$ is the OLS estimator from equation (1).Use the delta method to show that asymptotic standard error of $\hat{\theta}_1$ is $[100 \cdot \exp(\hat{\beta}_1)] \cdot \mathrm{se}(\hat{\beta}_1)$.

(c) Repeat parts a and b by finding the exact percentage change in $\mathrm{E}(wage|married, educ, \boldsymbol{z})$ for any given change in $educ$, $\Delta educ$. Call this θ_2. Explain how to estimate θ_2 and obtain its asymptotic standard error.

4. Show that the estimator $\hat{\boldsymbol{B}} = N^{-1}\Sigma_{i=1}^{N}\hat{u}_i^2\boldsymbol{x}_i'\boldsymbol{x}_i$ is consistent for $\boldsymbol{B} = \mathrm{E}(u^2\boldsymbol{x}'\boldsymbol{x})$, if $\hat{\beta}$ is a consistent estimator for β, by showing that $N^{-1}\Sigma_{i=1}^{N}\hat{u}_i^2\boldsymbol{x}_i'\boldsymbol{x}_i = N^{-1}\Sigma_{i=1}^{N}u_i^2\boldsymbol{x}_i'\boldsymbol{x}_i + o_p(1)$. Assume that all necessary expectations exist and are finite.

5. Consider the log(*wage*) equation for individuals $i = 1, ..., N$:

$$\log(wage_i) = \beta_0 + \beta_1 age_i + \beta_2 educ_i + u,$$

where $\log(wage_i)$, age_i and $educ_i$ denote wage, age and number of years of schooling for individual i. Given that *age* is beyond an individual's control, explain how *age* can be an endogenous explanatory variable in this regression. In particular, consider the three kinds of endogeneity discussed in Lecture 3.

6. Consider estimating the effect of personal computer ownership, as represented by a binary variable, PC, on college GPA, $colGPA$. GPA means the average grade in exams. With data on SAT scores (SAT is a kind of pre-college test) and high school GPA you postulate the model

$$colGPA = \beta_0 + \beta_1 hsGPA + \beta_2 SAT + \beta_3 PC + u.$$

   (a) Why might u and PC be positively correlated?

   (b) If the given equation is estimated by OLS using a random sample of college students, is $\hat{\beta}_3$ likely to have an upward or downward asymptotic bias?

   (c) What are some variables that might be good proxies for the unobservables in u that are correlated with PC?

7. Assume that y and each x_j have finite second moments, and write the linear projection of y on $(1, x_1, \ldots, x_K)$ as

$$\begin{aligned} y &= \beta_0 + \beta_1 x_1 + \cdots + \beta_K x_K + u = \beta_0 + \boldsymbol{x}\boldsymbol{\beta} + u, \\ \mathrm{E}(u) &= 0, \qquad \mathrm{E}\left(x_j u\right) = 0, \qquad j = 1, 2, \ldots, K. \end{aligned}$$

   (a) Show that $\sigma_y^2 = \mathrm{Var}(\boldsymbol{x}\boldsymbol{\beta}) + \sigma_u^2$.

   (b) For a random draw i from the population, write $y_i = \beta_0 + \boldsymbol{x}_i\boldsymbol{\beta} + u_i$. Evaluate the following assumption, which has been known to appear in econometrics textbooks: "$\mathrm{Var}(u_i) = \sigma^2 = \mathrm{Var}(y_i)$ for all i."

   (c) Define the population R–squared by $\rho^2 \equiv 1 - \sigma_u^2/\sigma_y^2 = \mathrm{Var}(\boldsymbol{x}\boldsymbol{\beta})/\sigma_y^2$. Show that the R–squared, $R^2 = 1 -$ SSR/SST, is a consistent estimator of ρ^2, where SSR is the OLS sum of squared residuals and SST $= \sum_{i=1}^{N}(y_i - \bar{y})^2$ is the total sum of squares.

   (d) Evaluate the following statement: "In the presence of heteroskedasticity, the R–squared from an OLS regression is meaningless." (This kind of statement also tends to appear in econometrics texts.)

8. Describe what is wrong with each of the following two statements:

   (a) "The central limit theorem implies that, as the sample size grows, the error distrubtion approaches normality."

   (b) "$\hat{u}_i^2$ (the squared OLS residual) is a consistent estimator of $\mathrm{E}(u_i^2|\boldsymbol{x}_i)$ for each i."