International School of Economics at TSU
Econometrics 2
Lasha Chochua

Problem Set 7

Instructions: You are encouraged to solve the problems before the recitation. Additionally, you are encouraged to work in groups. It is **not mandatory** to submit solutions unless stated otherwise. However, if you would like to share your solution, I would be happy to review it.

Problem 1 Consider the simple regression $Y_i = \beta_0 + \beta_1 X_i + u_i$ satisfying the least squares assumptions, but suppose X_i is unobserved. Instead, the econometrician observes $\widetilde{X}_i = X_i + w_i$, where w_i has mean zero, variance σ_w^2, $\text{Cov}(w_i, X_i) = 0$, and $\text{Cov}(w_i, u_i) = 0$.

(a) Show that the regression rewritten in terms of $\widetilde{X}_i$ takes the form $Y_i = \beta_0 + \beta_1 \widetilde{X}_i + v_i$ and derive an explicit expression for v_i.

(b) Derive $\text{Cov}(\widetilde{X}_i, v_i)$.

(c) Show that $\hat{\beta}_1 \xrightarrow{p} \dfrac{\sigma_X^2}{\sigma_X^2 + \sigma_w^2}\,\beta_1$ and explain why this is called *attenuation bias*.

(d) Suppose $\sigma_X^2 = 4$ and $\sigma_w^2 = 1$. By what percentage is $\hat{\beta}_1$ biased toward zero in large samples?

Problem 2 Suppose the true model is $Y_i = \beta_0 + \beta_1 X_i + u_i$ with all LSAs satisfied. The econometrician observes $\widetilde{Y}_i = Y_i + w_i$ instead of Y_i, where w_i is independent of X_i and u_i, has mean zero, and variance σ_w^2.

(a) Write the regression equation in terms of $\widetilde{Y}_i$ and identify the new error term.

(b) Show that $\hat{\beta}_1$ from regressing $\widetilde{Y}_i$ on X_i remains unbiased and consistent.

(c) Compare $\text{Var}(\hat{\beta}_1)$ in this case to the case with no measurement error. Conclude with a one-sentence comparison of measurement error in X versus in Y.

Problem 3 Consider the two-equation system

$$Y_i = \beta_0 + \beta_1 X_i + u_i, \qquad X_i = \gamma_0 + \gamma_1 Y_i + v_i,$$

where $\text{Cov}(u_i, v_i) = 0$, $E[u_i] = E[v_i] = 0$, $\text{Var}(u_i) = \sigma_u^2$, and $|\gamma_1 \beta_1| < 1$.

(a) Derive a closed-form expression for $\text{Cov}(X_i, u_i)$.

(b) Compute the probability limit of the OLS estimator of β_1.

(c) Show that if γ_1 and β_1 have the same sign, OLS overstates the magnitude of β_1.

Problem 4 Suppose $Y_i = \beta_0 + \beta_1 X_i + u_i$ with the LSAs holding in the population. For each of the following sample-selection mechanisms, state whether OLS on the *observed* sample is consistent for β_1 and justify briefly.

(a) Each observation is dropped independently with probability 1/2.

(b) Only observations with $X_i > 0$ are kept.

(c) Only observations with $Y_i > 0$ are kept.

(d) Only observations with $u_i > 0$ are kept.

Problem 5 A researcher has $n = 50$ i.i.d. observations and obtains

$$\widehat{Y} = 49.2 + 73.9\,X, \quad SE(\hat{\beta}_0) = 23.5, \quad SE(\hat{\beta}_1) = 16.4, \quad R^2 = 0.78.$$

A second researcher *enters every observation twice*, working with $n' = 100$.

(a) Show that the duplicated-data sample mean, variance, and covariance of X and Y equal those of the original data (use the unbiased $1/(n-1)$ formula).

(b) Conclude that $\hat{\beta}_0$, $\hat{\beta}_1$, and R^2 are *unchanged*.

(c) Show that the residual variance estimate $s^2 = SSR/(n-k)$ is approximately *halved*, and hence the reported standard errors *shrink by a factor* $\approx \sqrt{2}$.

(d) Which internal-validity condition has been violated?

Problem 6 A development economist runs an RCT in 50 villages in rural Kenya in 2018, finding that providing free deworming pills to schoolchildren raises school attendance by 6 percentage points, statistically significant at the 1% level. Treatment was randomized at the village level, take-up was 95%, and outcomes were measured by independent enumerators.

(a) Discuss the *internal validity* of the study, going through each of the five threats from the lecture. For each threat, state whether it is plausibly addressed and why.

(b) Discuss the *external validity* of generalizing the result to (i) urban Nairobi in 2018, (ii) rural Kenya in 2030, (iii) rural Bolivia in 2018.

(c) A policymaker reads the study and asks: "Should I conclude deworming raises *future earnings*?" What does the study tell us about this question, and what are the limits?

Problem 7 Consider the following hedonic price regression estimated on a sample of $n = 1{,}200$ home sales:

$$\log(price) = \beta_1\, sqft + \beta_2\, sqft^2 + \beta_3\, age + \beta_4 + e,$$

where *price* is in thousands of USD, $sqft$ is house size in hundreds of square feet, and *age* is the house's age in years.

The OLS point estimates are:

$$\hat{\beta}_1 = 0.060, \qquad \hat{\beta}_2 = -0.0008, \qquad \hat{\beta}_3 = -0.004, \qquad \hat{\beta}_4 = 4.50.$$

The heteroskedasticity-robust covariance matrix of $\hat{\beta}$ is:

$$\hat{V}_{\hat{\beta}} = \begin{pmatrix} 4.00 & -0.80 & 0 & 0 \\ -0.80 & 0.25 & 0 & 0 \\ 0 & 0 & 1.00 & 0 \\ 0 & 0 & 0 & 100 \end{pmatrix} \times 10^{-5}.$$

(The zero entries reflect negligible empirical covariances and are set to zero to keep the arithmetic clean – do not worry about their realism.)

(a) Interpret each coefficient. Why does the presence of $sqft^2$ with $\hat{\beta}_2 < 0$ make economic sense?

(b) Define $\theta_1 = 100(\hat{\beta}_1 + 2\hat{\beta}_2\, sqft)$ as the percentage change in price from adding 100 square feet, evaluated at $sqft = 20$ (a 2000 sqft house). Compute $\hat{\theta}_1$ and its standard error using the Delta Method. Construct a 95% confidence interval.

(c) Define $\theta_2 = -\hat{\beta}_1/(2\hat{\beta}_2)$ as the house size (in hundreds of sqft) that **maximizes** $\log(price)$. Compute $\hat{\theta}_2$ and its standard error. Construct a 95% confidence interval. Comment on its economic plausibility.

(d) For part (c), explain in one sentence why the standard error is considerably less informative than the one in part (b), even though both use the same (β_1, β_2) block of the covariance matrix.