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 – Classical Errors-in-Variables

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?

Solution

(a) Substituting $X_i = \widetilde{X}_i - w_i$ into the population regression:

$$Y_i = \beta_0 + \beta_1(\widetilde{X}_i - w_i) + u_i = \beta_0 + \beta_1 \widetilde{X}_i + \underbrace{(u_i - \beta_1 w_i)}_{v_i}.$$

(b) Using $\widetilde{X}_i = X_i + w_i$ and the orthogonality assumptions:

$$\text{Cov}(\widetilde{X}_i, v_i) = \text{Cov}(X_i + w_i,\, u_i - \beta_1 w_i) = -\beta_1 \,\text{Cov}(w_i, w_i) = -\beta_1 \sigma_w^2.$$

(c) The variance of the observed regressor is $\sigma^2_{\widetilde{X}} = \sigma^2_X + \sigma^2_w$. By the OLS probability limit:

$$\hat{\beta}_1 \xrightarrow{p} \beta_1 + \frac{\mathrm{Cov}(\widetilde{X}_i, v_i)}{\sigma^2_{\widetilde{X}}} = \beta_1 - \frac{\beta_1 \sigma^2_w}{\sigma^2_X + \sigma^2_w} = \frac{\sigma^2_X}{\sigma^2_X + \sigma^2_w}\,\beta_1.$$

The multiplicative factor lies in $(0, 1)$, so $|\hat{\beta}_1| < |\beta_1|$ in probability – the estimate is shrunk toward zero, hence *attenuation*.

(d) The factor equals $4/(4+1) = 0.8$, so $\hat{\beta}_1$ converges to $0.8\beta_1$, a **20% downward bias**.

Problem 2 – Measurement Error in Y vs in X

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^2_w.

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

Solution

(a) Substituting $Y_i = \widetilde{Y}_i - w_i$:

$$\widetilde{Y}_i = \beta_0 + \beta_1 X_i + (u_i + w_i) \equiv \beta_0 + \beta_1 X_i + \tilde{u}_i.$$

(b) Since w_i is independent of X_i and $E[w_i] = 0$, we have $E[\tilde{u}_i \mid X_i] = E[u_i \mid X_i] + E[w_i] = 0$. The first LSA holds for the rewritten model, so OLS is unbiased and consistent.

(c) The composite error has variance $\mathrm{Var}(\tilde{u}_i) = \sigma^2_u + \sigma^2_w > \sigma^2_u$, so $\mathrm{Var}(\hat{\beta}_1) = \dfrac{\sigma^2_u + \sigma^2_w}{n\sigma^2_X}$ – larger than the no-error variance $\sigma^2_u/(n\sigma^2_X)$.

Comparison. Measurement error in X is *dangerous* (biases $\hat{\beta}_1$ toward zero); measurement error in Y is merely *costly* (inflates variance, no bias).

Problem 3 – Simultaneous Causality Bias

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.

Solution

(a) Take $\text{Cov}(\cdot, u_i)$ of the second equation:

$$\text{Cov}(X_i, u_i) = \gamma_1 \text{Cov}(Y_i, u_i) + \text{Cov}(v_i, u_i) = \gamma_1 \text{Cov}(Y_i, u_i).$$

From the first equation, $\text{Cov}(Y_i, u_i) = \beta_1 \text{Cov}(X_i, u_i) + \sigma_u^2$. Substituting and solving:

$$\text{Cov}(X_i, u_i) = \frac{\gamma_1 \sigma_u^2}{1 - \gamma_1 \beta_1}.$$

(b) $\hat{\beta}_1 \xrightarrow{p} \beta_1 + \dfrac{\text{Cov}(X_i, u_i)}{\sigma_X^2} = \beta_1 + \dfrac{\gamma_1 \sigma_u^2}{(1 - \gamma_1 \beta_1)\, \sigma_X^2}.$

(c) If $\gamma_1, \beta_1 > 0$, the assumption $\gamma_1 \beta_1 < 1$ keeps the denominator positive, so the bias term is positive and $\text{plim}\, \hat{\beta}_1 > \beta_1 > 0$ – magnitude is overstated. The same logic applies (with both negative) to the case $\gamma_1, \beta_1 < 0$.

Problem 4 – Sample Selection: Three Cases

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.

Solution

(a) Consistent. Selection is independent of (X_i, u_i), so the kept sub-sample is i.i.d. from the same population. Only n shrinks.

(b) Consistent. Selection depends only on X_i, so $E[u_i \mid X_i, \text{selected}] = E[u_i \mid X_i] = 0$. The conditional mean independence assumption survives.

(c) Inconsistent. Selection on Y_i is selection on the dependent variable. Since Y_i depends on u_i, conditioning on $Y_i > 0$ induces $E[u_i \mid X_i, \text{selected}] \neq 0$, generating sample selection bias.

(d) Inconsistent. Selection directly on the error gives $E[u_i \mid \text{selected}] = E[u_i \mid u_i > 0] > 0$, violating the first LSA.

Problem 5 – The Duplicated-Data Mistake

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?

Solution

(a) Let the original sample have sums $S_X = \sum X_i$ and $S_{XX} = \sum X_i^2$. With each observation duplicated: $S'_X = 2S_X$, $S'_{XX} = 2S_{XX}$, and $n' = 2n$. Then $\bar{X}' = S'_X/n' = \bar{X}$, and similarly $\bar{Y}' = \bar{Y}$. For the variance,

$$s'^2_X = \frac{1}{n'-1}\sum (X_i - \bar{X}')^2 = \frac{2}{2n-1}\sum (X_i - \bar{X})^2 \approx s^2_X$$

(exact in the limit, off by a factor $(2n-2)/(2n-1)$ for finite n). The same holds for s^2_Y and s_{XY}.

(b) OLS slope and intercept depend only on means and (co)variances, which are unchanged: $\hat{\beta}_1' = s_{XY}/s_X^2 = \hat{\beta}_1$ and $\hat{\beta}_0' = \bar{Y} - \hat{\beta}_1 \bar{X}$. The $R^2 = s_{XY}^2/(s_X^2 s_Y^2)$ also unchanged.

(c) The sum of squared residuals doubles ($SSR' = 2\,SSR$) but the degrees of freedom roughly double too: $n' - 2 = 2n - 2$. So

$$s'^2 = \frac{2\,SSR}{2n-2} = \frac{SSR}{n-1} \approx \frac{n-2}{n-1}\, s^2 \approx s^2.$$

However, the variance of $\hat{\beta}_1$ is $s'^2/\sum(X_i' - \bar{X}')^2 = s^2/(2\sum(X_i - \bar{X})^2)$ – *halved*. Standard errors therefore shrink by $\sqrt{2}$:

$$SE'(\hat{\beta}_1) \approx 16.4/\sqrt{2} \approx 11.6, \qquad SE'(\hat{\beta}_0) \approx 23.5/\sqrt{2} \approx 16.6.$$

(d) The i.i.d. (random sampling) assumption is violated – duplicated observations are perfectly correlated, not independent. The reported standard errors are inconsistent (too small), so confidence intervals and t-tests are invalid even though point estimates are unchanged.

Problem 6 – Internal vs External Validity in Practice

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?

Solution

(a) Internal validity.

- *Omitted variables:* Random assignment ensures, in expectation, that treated and control villages are balanced on all observed and unobserved characteristics. With 50 villages, residual imbalance is possible but unsystematic. **Plausibly addressed.**
- *Functional form:* The treatment effect is estimated as a simple difference in means – no parametric form is imposed. **Not a concern.**

- *Errors in variables:* Outcomes were measured by independent enumerators (reduces bias from teacher misreporting). Treatment assignment is a clean binary variable. **Plausibly addressed.**
- *Sample selection:* If the 50 villages were sampled representatively from the rural Kenyan target population, no concern within that population. If villages were chosen for convenience, the external-validity question gets harder but internal validity for the studied villages still holds.
- *Simultaneous causality:* Reverse causality is impossible – random treatment assignment severs any potential reverse channel from attendance to deworming. **Eliminated by design.**

Internal validity is *strong*: this is roughly the gold standard for causal inference.

(b) External validity.

- *(i) Urban Nairobi 2018.* Worm prevalence is much lower in urban areas (better sanitation, water). The mechanism – reduced parasitic illness → better attendance – is unlikely to operate at the same magnitude. **Weak generalizability.**
- *(ii) Rural Kenya 2030.* If sanitation and parasite prevalence remain similar, the effect should largely persist. If government deworming programs have already scaled up, the marginal effect of additional pills will be smaller. **Moderate generalizability** (depends on what changed).
- *(iii) Rural Bolivia 2018.* Different worm species, different schooling system, different climate. The *qualitative* mechanism (treat illness → improve attendance) may transfer; the *quantitative* magnitude almost certainly differs. **Mechanism may transfer; magnitude likely will not.**

(c) The study estimates the causal effect on a *short-run, intermediate outcome* (attendance), not on earnings. Concluding anything about future earnings requires (i) a credible mapping from additional schooling to wages and (ii) the assumption that deworming-induced attendance has the same returns as other forms of attendance. Both are *additional assumptions outside the experiment*. The honest answer: the study tells us deworming raises attendance; long-run earnings effects require either follow-up data or extrapolation through outside evidence (which is what Baird et al. and the deworming literature have attempted).

Problem 7 – Delta Method

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:

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

(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 $\widehat{\theta}_1$ 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 $\widehat{\theta}_2$ 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.

Solution

Part (a): Interpretation

- $\hat{\beta}_1 = 0.060$ is the marginal log-price effect of adding 100 sqft at $sqft = 0$; by itself it would suggest a 6% price increase per 100 sqft, but this is only the **linear** piece.
- $\hat{\beta}_2 = -0.0008$ captures **diminishing returns to size**: each additional 100 sqft adds less value than the previous one. The negative sign makes economic sense because buyers value the first few rooms much more than the twentieth room – the wage-experience analogue is exactly the same concavity argument.
- $\hat{\beta}_3 = -0.004$: each additional year of age reduces log price by about 0.4%.
- $\hat{\beta}_4 = 4.50$ is the intercept (baseline log price).

Because $\hat{\beta}_2 < 0$, the quadratic $\hat{\beta}_1\, sqft + \hat{\beta}_2\, sqft^2$ is **concave**, so it has an interior maximum – the house size beyond which additional square footage actually reduces log price (perhaps because maintenance costs dominate, or the market has few buyers for very large homes).

Part (b): Percentage Price Effect at $sqft = 20$

Step 1 – Function and point estimate. Define $r(\beta) = 100(\beta_1 + 40\beta_2)$, since $2 \cdot sqft = 40$ at $sqft = 20$. Then:

$$\hat{\theta}_1 = 100(0.060 + 40(-0.0008)) = 100(0.060 - 0.032) = 100(0.028) = 2.8.$$

So adding 100 sqft to a 2000-sqft home raises log price by about 2.8%.

Step 2 – Jacobian. Differentiate $r(\beta)$ with respect to each of the four coefficients:

$$R = \begin{pmatrix} \partial r/\partial \beta_1 \\ \partial r/\partial \beta_2 \\ \partial r/\partial \beta_3 \\ \partial r/\partial \beta_4 \end{pmatrix} = \begin{pmatrix} 100 \\ 4000 \\ 0 \\ 0 \end{pmatrix}.$$

This is a 4×1 column. The nonzero slots are positions 1 and 2, so the active block of $\hat{V}_{\hat{\beta}}$ is the top-left 2×2 submatrix.

Step 3 – Standard error. Compute the sandwich $R'\hat{V}_{\hat{\beta}}R$, using only the relevant block:

$$s(\hat{\theta}_1)^2 = \begin{pmatrix} 100 & 4000 \end{pmatrix} \begin{pmatrix} 4.00 & -0.80 \\ -0.80 & 0.25 \end{pmatrix} \begin{pmatrix} 100 \\ 4000 \end{pmatrix} \times 10^{-5}.$$

Inner product first:

$$\begin{pmatrix} 4.00(100) + (-0.80)(4000) \\ -0.80(100) + 0.25(4000) \end{pmatrix} = \begin{pmatrix} 400 - 3200 \\ -80 + 1000 \end{pmatrix} = \begin{pmatrix} -2800 \\ 920 \end{pmatrix}.$$

Then:

$$100(-2800) + 4000(920) = -280{,}000 + 3{,}680{,}000 = 3{,}400{,}000.$$

So $s(\hat{\theta}_1)^2 = 3{,}400{,}000 \times 10^{-5} = 34$, and $s(\hat{\theta}_1) = \sqrt{34} \approx 5.83$.

Step 4 – Confidence interval. Using the $N(0,1)$ critical value 1.96:

$$\text{CI}_{95\%}(\theta_1) = 2.8 \pm 1.96(5.83) \approx [-8.6,\ 14.2].$$

Interpretation. The point estimate says an extra 100 sqft raises log price by about 2.8% at a 2000-sqft home, but the CI is extremely wide and includes zero and negative values. We **cannot reject** the hypothesis that there is no marginal value to extra size at this point. (In a real problem set this would motivate either a larger sample or a different functional form.)

Part (c): Optimal House Size

Step 1 – Function and point estimate. Define $r(\beta) = -\beta_1/(2\beta_2)$. Then:

$$\hat{\theta}_2 = -\frac{0.060}{2(-0.0008)} = -\frac{0.060}{-0.0016} = 37.5.$$

So the log-price-maximizing size is $37.5 \times 100 = 3750$ sqft.

Step 2 – Jacobian. θ_2 is **nonlinear** in β, so R depends on β and we plug in $\hat{\beta}$:

$$R = \begin{pmatrix} -1/(2\beta_2) \\ \beta_1/(2\beta_2^2) \\ 0 \\ 0 \end{pmatrix}.$$

Evaluating at $\hat{\beta}$:

$$\hat{R}_1 = -\frac{1}{2(-0.0008)} = 625, \qquad \hat{R}_2 = \frac{0.060}{2(-0.0008)^2} = \frac{0.060}{0.00000128} \approx 46{,}875.$$

So $\hat{R} = (625,\ 46{,}875,\ 0,\ 0)'$. Same active block as part (b).

Step 3 – Standard error. Sandwich through the 2×2 top-left block:

$$s(\hat{\theta}_2)^2 = \begin{pmatrix} 625 & 46{,}875 \end{pmatrix} \begin{pmatrix} 4.00 & -0.80 \\ -0.80 & 0.25 \end{pmatrix} \begin{pmatrix} 625 \\ 46{,}875 \end{pmatrix} \times 10^{-5}.$$

Inner product:

$$\begin{pmatrix} 4.00(625) + (-0.80)(46{,}875) \\ -0.80(625) + 0.25(46{,}875) \end{pmatrix} = \begin{pmatrix} 2500 - 37{,}500 \\ -500 + 11{,}718.75 \end{pmatrix} = \begin{pmatrix} -35{,}000 \\ 11{,}218.75 \end{pmatrix}.$$

Then:

$$625(-35{,}000) + 46{,}875(11{,}218.75) \approx -21{,}875{,}000 + 525{,}878{,}906 \approx 504{,}003{,}906.$$

So $s(\hat{\theta}_2)^2 \approx 504{,}003{,}906 \times 10^{-5} \approx 5040$, and $s(\hat{\theta}_2) \approx \sqrt{5040} \approx 71.0$.

Step 4 – Confidence interval.

$$\mathrm{CI}_{95\%}(\theta_2) = 37.5 \pm 1.96(71.0) \approx [-101.6,\ 176.6].$$

Interpretation. The point estimate says the log-price-maximizing home is about 3750 sqft, but the CI is so wide that it includes negative sizes – which is economically nonsense. We have essentially **no information** about where the true optimum lies. The point estimate by itself is misleading; only the Delta Method standard error reveals how little we actually know.

Part (d): Why is (c) less informative than (b)?

Both parts use the same 2×2 block of the covariance matrix, but the Jacobian $\hat{R}$ for part (c) has much larger entries – 625 and 46,875 versus 100 and 4000 – because the gradient of a **ratio** blows up when the denominator ($\hat{\beta}_2$) is close to zero. Since $\hat{\beta}_2 = -0.0008$ is tiny, the sandwich $\hat{R}'\hat{V}_{\hat{\beta}}\hat{R}$ inherits that amplification and produces a huge variance. Linear functions are well-behaved; ratios near zero are not.