Problem Set: Causality

Problem 1. *You wish to estimate the ceteris paribus effect of x_1 on y using controls x_2 and x_3 (e.g., y = exam score, x_1 = attendance, x_2 = prior GPA, x_3 = SAT/ACT). Let $\tilde{\beta}_1$ be the slope from the simple regression of y on x_1, and let $\hat{\beta}_1$ be the slope from the multiple regression of y on (x_1, x_2, x_3).*

1. *If x_1 is highly correlated with (x_2, x_3) and x_2, x_3 have large partial effects on y, do you expect $\tilde{\beta}_1$ and $\hat{\beta}_1$ to be close or far apart? Explain briefly.*

2. *If x_1 is uncorrelated with (x_2, x_3) but x_2 and x_3 are highly correlated with each other, will $\tilde{\beta}_1$ and $\hat{\beta}_1$ tend to be similar or quite different? Explain.*

3. *If x_1 is highly correlated with (x_2, x_3) and x_2, x_3 have small partial effects on y, which is likely smaller: $\mathrm{se}(\tilde{\beta}_1)$ or $\mathrm{se}(\hat{\beta}_1)$? Explain.*

4. *If x_1 is uncorrelated with (x_2, x_3), x_2, x_3 have large partial effects on y, and x_2, x_3 are highly correlated, which is likely smaller: $\mathrm{se}(\tilde{\beta}_1)$ or $\mathrm{se}(\hat{\beta}_1)$? Explain.*

Problem 2. *Consider the population model*

$$y \;=\; \beta_0 + \beta_1 x + \beta_2 x^2 + u, \qquad E[u \mid x] = 0,$$

where the regressor x is standard normal in the population:

$$E[x] = 0, \quad \mathrm{Var}(x) = E[x^2] = 1, \quad E[x^3] = 0$$

We study what can be said about the OLS estimator of β_1 when x^2 is omitted and we run the simple regression of y on x.

1. *Show that we can write*
$$y \;=\; \alpha_0 + \beta_1 x + v$$
with $E[v] = 0$. Identify v and the new intercept α_0.

2. *Show that $E[v \mid x]$ depends on x unless $\beta_2 = 0$.*

3. *Show that $\mathrm{Cov}(x, v) = 0$.*

4. *Let $\hat{\beta}_1$ be the slope from the simple regression of y_i on x_i. Is $\hat{\beta}_1$ consistent for β_1? Is it unbiased? Explain.*

5. *Argue that estimating β_1 has value in the following sense: β_1 is the partial effect of x on y evaluated at $x = 0$, the population mean of x.*

6. *Explain why being able to consistently estimate both β_1 and β_2 is more informative than estimating only β_1.*

Problem 3. *You wish to estimate the effect of alcohol consumption on college GPA. For each student you observe colGPA (college GPA), alcohol (alcohol consumption), attend (percent of lectures attended), SAT (standardized test score), and hsGPA (high school GPA).*

1. *Should attend be included alongside alcohol in a multiple regression (e.g., $colGPA = \beta_0 + \beta_1 alcohol + \beta_2 attend + u$)? Explain how including attend affects the interpretation of β_1 (the coefficient on alcohol).*

2. *Should SAT and hsGPA be included as additional controls? Briefly justify your answer (e.g., relevance, omitted-variable bias, and interpretation of partial effects).*

Problem 4. *Two regressions of fourth-grade math performance (math4) on school characteristics were estimated. The key regressor is lexppp, the log of per-pupil expenditures.*

Model A (without reading):

$$\widehat{math4} = \underset{(59.24)}{24.49} + \underset{(4.04)}{9.01}\, lexppp - \underset{(0.071)}{0.422}\, free - \underset{(5.358)}{0.752}\, lmedinc - \underset{(0.161)}{0.274}\, pctsgle,$$
$$n = 229,\ R^2 = 0.472,\ adjR^2 = 0.462.$$

Model B (adds reading):

$$\widehat{math4} = \underset{41.70}{149.38} + \underset{(2.82)}{1.93}\, lexppp - \underset{(0.054)}{0.060}\, free - \underset{(3.76)}{10.78}\, lmedinc - \underset{(0.111)}{0.397}\, pctsgle + \underset{(0.042)}{0.667}\, read4,$$
$$n = 229,\ R^2 = 0.749,\ adjR^2 = 0.743,$$

where free is % children eligible for free lunch, medinc is median family income in the same zipcode, and pctsgle % of children not in married-couple families.

1. *If your goal is the* causal *effect of spending on math achievement, explain why Model A is more relevant than Model B. Based on Model A, what is the predicted change in math4 from a 10% increase in expenditures per pupil?*

2. *When read4 is added (Model B), do any coefficients (besides that on lexppp) change in unexpected ways in sign or significance? Briefly discuss what this suggests about including read4.*

3. *How would you explain, in this context, you may* prefer *Model A even though it has a smaller adjusted R^2 than Model B?*

Problem 5. *To assess a job training program's effect on wages, consider*

$$\log(wage) \;=\; \beta_0 + \beta_1\, train + \beta_2\, educ + \beta_3\, exper + u,$$

where **train** *is a binary indicator for program participation and u captures unobserved ability. If lower-ability workers are more likely to be selected into the program (so that selection is based on unobservables contained in u), and you estimate the model by OLS, what is the likely sign of the bias in* $\hat{\beta}_1$*?*

Problem 6. *For child i in a given school district, let* $\texttt{voucher}_i \in \{0, 1\}$ *indicate selection into a school voucher program, and let* $\texttt{score}_i$ *be a subsequent standardized test score. Suppose* $\texttt{voucher}_i$ *is completely randomized—i.e., independent of all observed and unobserved determinants of* $\texttt{score}_i$*.*

1. *In the simple regression* $\texttt{score}_i = \beta_0 + \beta_1\, \texttt{voucher}_i + u_i$*, does the OLS estimator* $\hat{\beta}_1$ *provide an unbiased estimate of the program's effect?*

2. *If additional background covariates* Z_i *(e.g., family income, family structure, parents' education) are available, must you control for them to obtain an unbiased estimate of the voucher effect? Explain.*

3. *Why might you still include* Z_i *in the regression? Under what circumstances would you choose* not *to include them?*

Problem 7. *The following equation explains weekly hours of television viewing by a child in terms of the child's age, mother's education, father's education, and number of siblings:*

$$tvhours^* = \beta_0 + \beta_1\, age + \beta_2\, age^2 + \beta_3\, motheduc + \beta_4\, fatheduc + \beta_5\, sibs + u.$$

We are worried that $tvhours^*$ *is measured with error in our survey. Let tvhours denote the* reported *hours of television viewing per week.*

1. *What do the classical errors-in-variables (CEV) assumptions require in this application?*

2. *Do you think the CEV assumptions are likely to hold? Explain.*

Problem 8. *Consider the simple regression model with classical measurement error,*

$$y \;=\; \beta_0 \;+\; \beta_1 x^* \;+\; u,$$

where we have m measures on x^*. Write these as

$$z_h \;=\; x^* \;+\; e_h, \qquad h = 1, \ldots, m.$$

Assume that x^* is uncorrelated with $u, e_1, \ldots, e_m$; that the measurement errors are pairwise uncorrelated; and that they share the same variance σ_e^2. Let

$$w \;=\; \frac{z_1 + \cdots + z_m}{m}$$

be the average of the measures on x^*, so that for each observation i,

$$w_i \;=\; \frac{z_{i1} + \cdots + z_{im}}{m}$$

is the average of the m measures. Let $\hat{\beta}_1$ be the OLS estimator from the simple regression of y_i on 1 and w_i, $i = 1, \ldots, n$, using a random sample.

1. *Show that*

$$plim\big(\hat{\beta}_1\big) \;=\; \beta_1 \cdot \frac{\sigma_{x^*}^2}{\sigma_{x^*}^2 + \sigma_e^2/m}.$$

   Hint: $plim(\bar{\beta}_1) = Cov(w, y)/Var(w)$.

2. *How does the inconsistency in $\hat{\beta}_1$ compare with that when only a single measure is available (that is, $m = 1$)? What happens as m grows? Comment.*