PROBLEM SET 2

Problem 1 (Omitted variable bias and IV)

Let y_i be a measure of good health, and x_i be an indicator for smoking ($x_i = 1$ if person i smokes, $x_i = 0$ otherwise). We want to estimate the structural model

$$y_i = \alpha + \beta x_i + u_i.$$

You collect data on y_i and x_i from randomly selected individuals in Strangetown. To your astonishment you find that y_i and x_i are positively correlated in your dataset. However, you also notice that both cigarettes and health services are very expensive in Strangetown, because of high taxes on cigarettes and a shortage of medical personal. You therefore suspect that x_i is endogenous, because only people with high income can afford medical services and smoking (usually low income individuals have a higher propensity to smoke, but Strangetown is a little strange). You don't observe income.

A couple of years ago a big tobacco company randomly selected individuals as part of a publicity campain in Strangetown and provided the selected individuals with 100 packs of cigarettes for free. You observe an indicator z_i of whether individual i was selected ($z_i = 1$) or not ($z_i = 0$).

The total sample size is $n = 70$. We observe $n_{00} = 30$ individuals with $x_i = z_i = 0$; the average of the health outcome y_i in that group is observed to be $\overline{y}_{00} = 1.0$. We observe $n_{01} = 10$ individuals with $x_i = 0$ and $z_i = 1$; the average outcome in that group is $\overline{y}_{01} = 0.8$. We observe $n_{10} = 20$ individuals with $x_i = 1$ and $z_i = 0$; the average outcome in that group is $\overline{y}_{10} = 1.5$. Finally, we observe $n_{11} = 10$ individuals with $x_i = z_i = 1$; the average outcome in that group is $\overline{y}_{11} = 1.2$.

(a) Argue why z_i is probably a good instrument for x_i. What does it mean that z_i is a relevant instrument? Is it reasonable to assume that z_i is relevant? What does it mean that z_i is exogenous? Is it reasonable to assume that z_i is exogenous?

(b) Calculate the OLS estimator for α and β. Interpret the OLS estimator for β.

(c) Calculate the 2SLS estimator for α and β. Compare to the OLS estimator for β and interpret the difference. (Remember: we can calculate the 2SLS estimator in the exactly identified case as $(Z'X)^{-1}Z'y$)

(d) Assume that $\text{Var}(u_i|z_i) = 1/10$. Calculate the estimated standard error of the 2SLS estimator for β.

(e) Test the hypothesis $H_0 : \beta = 0$. Do you reject the hypothesis when employing a large sample t-test with size 5%?

Problem 2 (Measurement error and IV)

Consider a linear regression model

$$y_i = p_i^* \beta + \varepsilon_i,$$

where p_i^* is the only regressor and $\beta \neq 0$. However, p_i^* is not observed, only p_i is observed, which is a noisy version of p_i^* that is contaminated with measurement error v_i, namely

$$p_i = p_i^* + v_i.$$

In addition, we observe a single instrumental variable z_i, i.e. the three observed variables are y_i, p_i, and z_i. We assume that $(z_i, p_i^*, \varepsilon_i, v_i)$ are independent and identically distributed across observations $i = 1, \ldots, n$, and that

$$\begin{pmatrix} z_i \\ p_i^* \\ \varepsilon_i \\ v_i \end{pmatrix} \sim \mathcal{N}\left[\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho & 0 & 0 \\ \rho & 1 & 0 & 0 \\ 0 & 0 & \sigma_\epsilon^2 & 0 \\ 0 & 0 & 0 & \sigma_v^2 \end{pmatrix} \right],$$

where $0 < \rho < 1$, $\sigma_\epsilon > 0$, and $\sigma_v > 0$ are unknown parameters. Consider the following estimators

$$\hat{\beta}_{\text{OLS}} = \frac{\sum_{i=1}^n p_i y_i}{\sum_{i=1}^n p_i^2}, \qquad \hat{\gamma} = \frac{\sum_{i=1}^n z_i p_i}{\sum_{i=1}^n z_i^2}, \qquad \hat{\pi} = \frac{\sum_{i=1}^n z_i y_i}{\sum_{i=1}^n z_i^2}.$$

These are three OLS estimators obtained from regressing y_i on p_i, p_i on z_i, and y_i on z_i, respectively.

(a) Derive the probability limit of $\hat{\beta}_{\text{OLS}}$ as $n \to \infty$. Is $\hat{\beta}_{\text{OLS}}$ a consistent estimator for β?

(b) Derive the probability limits of $\hat{\gamma}$ and $\hat{\pi}$ as $n \to \infty$.

(c) How can $\hat{\gamma}$ and $\hat{\pi}$ be combined to obtain a consistent estimator for β? Define $\hat{p}_i = \hat{\gamma} z_i$ (the predicted values of the regression of p_i on z_i). How can your proposed estimator for β be obtained by a regression that involves $\hat{p}_i$?

(d) What is the condition on ρ that guarantees that z_i is a relevant instrument for p_i here? Would z_i be an exogenous instrument if $\mathbb{E}(z_i \varepsilon_i) \neq 0$? Would z_i be an exogenous instrument if $\mathbb{E}(z_i v_i) \neq 0$?

(e) Assume that the original model reads

$$y_i = \beta_1 + p_i^* \beta_2 + w_i \beta_3 + \varepsilon_i,$$

where we now also include a constant and one additional regressor w_i, which is assumed to be exogenous (i.e. uncorrelated with ε_i) and uncorrelated with v_i. Otherwise, all distributional assumptions on $(z_i, p_i^*, \varepsilon_i, v_i)$ are unchanged. Describe how one can consistently estimate β_1, β_2 and β_3 in that case (no proof required)?

Problem 3 (Applied 2SLS)

In Problem Set 1, we found that the wages of working married women were significantly affected by years of education. However, from our understanding of econometrics, we may fear that we've reached a premature conclusion.

$$\log(wage) = \beta_0 + \beta_1 educ + u \tag{1}$$

1) If we estimate the simple wage equation above on a basis of a random sample of 5000 adults, will we obtain a consistent estimate of β_1? [Hint: Can you think of any omitted variables, correlated with *educ*?]

2) If we had unlimited funds, and wanted to figure out β_1 by estimating the equation above, what kind of data should we collect?

3) Can you think of any candidates for an instrumental variable (IV) of *educ*?

4) Angrist & Krueger (1991) utilize information from the 1970 and 1980 national census of males to estimate the returns to education. Use the "NEW7080.dta" dataset to replicate column 1 of table IV on page 999.

5) Angrist & Krueger (1991) uses quarter of birth - i.e. four dummies for quarter of birth *QTR1*, *QTR2*, *QTR3* and *QTR4* - as instruments for educational level. Write up and discuss the assumptions which are required for these instruments to be valid?

6) Write up and estimate the reduced form (stage 1) for *educ*. Interpret your results.

7) How could you assess the strength of *QTR1 – QTR4* as instruments for *educ*? What do you conclude about identification of model (1) using *QTR1 – QTR4* as instruments for *educ*?

8) Estimate model (1) by two stage least squares (2SLS) using *QTR1 – QTR4* as instruments for *educ*.

9) Are there any important differences in the OLS and 2SLS estimates?

10) Replicate column 2 of table IV using quarter of birth interacted with birth year as instruments for education.

11) EXTRA: If time permits, try replicating the remaining columns in table IV.