PROBLEM SET 2 SOLUTIONS

Problem 1 (Omitted variable bias and IV)

(a) z_i is a relevant instrument for x_i if $\text{Cov}(x_i, z_i) \neq 0$, i.e. if x_i and z_i are correlated. It is plausible to assume that z_i is relevant in this example, because people that were provided with free cigarettes ($z_i = 1$) are probably more likely to have started smoking ($x_i = 1$) than people that were not be given free cigarettes ($z_i = 0$).

z_i is exogenous if $\text{Cov}(u_i, z_i) = \mathbb{E}(u_i z_i) = 0$. This is also called an exclusion restriction, because it means that z_i has no direct effect on y_i, only an indirect effect through x_i. One can argue that z_i is exogenous here, because having 100 packs of free cigarettes should not directly affect your health, only indirectly through your smoking behaviour. However, the main concern here would probably be that there might be other indirect channels that we don't control for, in particular one might sell the 100 packs of free cigarettes and use the money to buy health services, which impacts y_i. This is clearly a concern, but unless we are given additional data (e.g. health service expenses for each individual, which we could include as an additional controls — i.e. regressors — in our model) there is not much we can do about this concern. In the following we assume that x_i is exogenous. Our results are invalid, if this assumption is violated.

(b) Let y, x, and z be the $n \times 1$ vectors with entries y_i, x_i, and z_i. Let 1_n be the $n \times 1$ vector whose entries are all equal to 1. Let $X = (1_n, x)$ and $Z = (1_n, z)$. We have

$$X'X = \begin{pmatrix} n & n_{10} + n_{11} \\ n_{10} + n_{11} & n_{10} + n_{11} \end{pmatrix} = \begin{pmatrix} 70 & 30 \\ 30 & 30 \end{pmatrix},$$

$$X'y = \begin{pmatrix} n_{00}\overline{y}_{00} + n_{01}\overline{y}_{01} + n_{10}\overline{y}_{10} + n_{11}\overline{y}_{11} \\ n_{10}\overline{y}_{10} + n_{11}\overline{y}_{11} \end{pmatrix} = \begin{pmatrix} 80 \\ 42 \end{pmatrix}. \tag{7}$$

We compute

$$(X'X)^{-1} = \frac{1}{1200} \begin{pmatrix} 30 & -30 \\ -30 & 70 \end{pmatrix}. \tag{8}$$

For the OLS estimator we thus find

$$\begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \end{pmatrix} = (X'X)^{-1}X'y = \begin{pmatrix} 0.95 \\ 0.45 \end{pmatrix}. \tag{9}$$

The fact that $\hat{\beta} > 0$ means that y_i (health) and x_i (smoking) are possitively correlated in our sample. However, since we suspect that x_i is endogenous we should not interpret $\hat{\beta} > 0$ as measuring a positive causal effect (i.e. that smoking causes better health).

(c) We have

$$\begin{aligned} Z'X &= \begin{pmatrix} n & n_{10} + n_{11} \\ n_{01} + n_{11} & n_{11} \end{pmatrix} = \begin{pmatrix} 70 & 30 \\ 20 & 10 \end{pmatrix}, \\ Z'y &= \begin{pmatrix} n_{00}\overline{y}_{00} + n_{01}\overline{y}_{01} + n_{10}\overline{y}_{10} + n_{11}\overline{y}_{11} \\ n_{01}\overline{y}_{01} + n_{11}\overline{y}_{11} \end{pmatrix} = \begin{pmatrix} 80 \\ 20 \end{pmatrix}. \end{aligned} \tag{10}$$

We compute

$$(Z'X)^{-1} = \frac{1}{100}\begin{pmatrix} 10 & -30 \\ -20 & 70 \end{pmatrix}. \tag{11}$$

For the 2SLS estimator we thus find

$$\begin{pmatrix} \hat{\alpha}_{\text{2SLS}} \\ \hat{\beta}_{\text{2SLS}} \end{pmatrix} = (Z'X)^{-1}Z'y = \begin{pmatrix} 2 \\ -2 \end{pmatrix}. \tag{12}$$

If we believe that our instrument is relevant and exogenous, and that our sample size is sufficiently large (we haven't calculated any standard error, yet), then $\hat{\beta}_{\text{2SLS}}$ is a good estimator for the true causal effect of x_i on y_i; since $\hat{\beta}_{\text{2SLS}} < 0$ this means that smoking decreases health. If we believe in this conclusion, then the fact that the OLS estimator is positive is indeed a result of the endogeneity of x_i.

(d) We assume $\text{Var}(u_i|z_i) = 1/10$, i.e. homoscedasticity with $\sigma^2 = 1/10$. Under homoscedasticty and for known σ^2 the variance of the 2SLS estimator can be estimated via

$$\widehat{\text{Var}}\begin{pmatrix} \hat{\alpha}_{\text{2SLS}} \\ \hat{\beta}_{\text{2SLS}} \end{pmatrix} = \sigma^2(X'P_ZX)^{-1} = \sigma^2\left[(X'Z)(Z'Z)^{-1}(Z'X)\right]^{-1}. \tag{13}$$

We already calculated $Z'X$ above (and we have $X'Z = (Z'X)'$). We find

$$\begin{aligned} Z'Z &= \begin{pmatrix} n & n_{01}+n_{11} \\ n_{01}+n_{11} & n_{01}+n_{11} \end{pmatrix} = \begin{pmatrix} 70 & 20 \\ 20 & 20 \end{pmatrix}, \\ (Z'Z)^{-1} &= \frac{1}{1000}\begin{pmatrix} 20 & -20 \\ -20 & 70 \end{pmatrix}. \end{aligned} \tag{14}$$

Combing these results we obtain

$$\begin{aligned} \widehat{\text{Var}}\begin{pmatrix} \hat{\alpha}_{\text{2SLS}} \\ \hat{\beta}_{\text{2SLS}} \end{pmatrix} &= \frac{1}{10}\begin{pmatrix} 70 & 30 \\ 30 & 13 \end{pmatrix}^{-1} \\ &= \begin{pmatrix} 0.13 & -0.3 \\ -0.3 & 0.7 \end{pmatrix}. \end{aligned} \tag{15}$$

Thus, $\widehat{\text{std}}(\hat{\beta}) = \sqrt{0.7} = 0.8367$.

(e) The t-test statistics for testing $H_0 : \beta = 0$ reads $t = \frac{\hat{\beta}}{\widehat{\text{std}}(\hat{\beta})} = -2.39$. Since $|t| > 1.96$ we reject H_0 at 95% confidence level (= 5% significance level).

Problem 2 (Measurement error and IV)

(a) We have

$$y_i = p_i\beta + u_i,$$

where $u_i = \varepsilon_i - v_i\beta$. We find

$$\hat{\beta}_{\text{OLS}} - \beta = \frac{\frac{1}{n}\sum_{i=1}^n p_i u_i}{\frac{1}{n}\sum_{i=1}^n p_i^2} \to_p \frac{\mathbb{E}p_i u_i}{\mathbb{E}p_i^2} = \frac{-\beta\mathbb{E}v_i^2}{\mathbb{E}(p_i^*)^2 + \mathbb{E}v_i^2} = \frac{-\beta\sigma_v^2}{1+\sigma_v^2} \neq 0,$$

as $n \to \infty$. Here, we used the WLLN and the continuous mapping theorem. Thus, $\hat{\beta}_{\text{OLS}}$ is not consistent.

(b) As $n \to \infty$ we have

$$\hat{\gamma} = \frac{\sum_{i=1}^n z_i p_i}{\sum_{i=1}^n z_i^2} \to_p \frac{\mathbb{E}z_i p_i}{\mathbb{E}z_i^2} = \frac{\mathbb{E}z_i p_i^*}{\mathbb{E}z_i^2} = \frac{\rho}{1} = \rho,$$
$$\hat{\pi} = \frac{\sum_{i=1}^n z_i y_i}{\sum_{i=1}^n z_i^2} \to_p \frac{\mathbb{E}z_i y_i}{\mathbb{E}z_i^2} = \frac{\beta\mathbb{E}z_i p_i^*}{\mathbb{E}z_i^2} = \frac{\beta\rho}{1} = \beta\rho,$$

where we again used the WLLN and the continuous mapping theorem.

(c) A consistent estimator for β is given by

$$\hat{\beta}_{\text{IV}} = \frac{\hat{\pi}}{\hat{\gamma}}.$$

This is the standard IV (or 2SLS) estimator expressed in terms of the reduced form estimators $\hat{\pi}$ and $\hat{\gamma}$. In the standard 2SLS setting the first stage consists of obtaining $\hat{\gamma}$ and thus $\hat{p}_i = \hat{\gamma} z_i$, followed by the second stage, where one regresses y_i on $\hat{p}_i$, which gives the same $\hat{\beta}_{\text{IV}}$ as above.

(d) z_i is a relevant instrument if $\rho \neq 0$.
The effective error term is $u_i = \varepsilon_i - v_i\beta$, which contains both ε_i and v_i, i.e. exogeneity of z_i requires $\mathbb{E}(z_i\varepsilon_i) = 0$ and $\mathbb{E}(z_i v_i) = 0$.

(e) 2SLS still works, i.e. one regresses y_i on a constant, $\hat{p}_i = \hat{\gamma} z_i$ and w_i. The resulting second stage estimator for β_1, β_2, β_3 is consistent.