Advanced Econometrics, HT, Problem Set 5

Question 1

We observe a random sample (Y_i, X_i, Z_i), $i = 1, \ldots, n$, where n is the sample size, and Y_i, X_i and Z_i are three scalar variables with finite second moments. For simplicity we assume that $\mathbb{E}Y_i = \mathbb{E}X_i = \mathbb{E}Z_i = 0$. We assume that

$$Y_i = X_i\beta + U_i$$

holds, where U_i is a mean zero error term and β is the parameter of interest. We also define $\gamma_{xx} = \mathrm{Var}(X_i)$, $\gamma_{xy} = \mathrm{Cov}(X_i, Y_i)$, $\gamma_{xz} = \mathrm{Cov}(X_i, Z_i)$ and $\gamma_{yz} = \mathrm{Cov}(Y_i, Z_i)$. We assume that $\gamma_{xx} > 0$.

(a) Write down consistent estimators for γ_{xx}, γ_{xy}, γ_{xz}, and γ_{yz}. Explain why they are consistent.

(b) Assume that X_i is exogenous, i.e. that $\mathbb{E}(X_iU_i) = 0$. Use this assumption to express β as a function of γ_{xx} and γ_{xy}. Use this expression for β and your result in (a) to provide a consistent estimator for β. Prove that your estimator for β is indeed consistent.

(c) Assume that X_i is endogenous, i.e. $\mathbb{E}(X_iU_i) \neq 0$, but that the instrument Z_i satisfies the exclusion restriction $\mathbb{E}(Z_iU_i) = 0$ and relevance assumption $\mathbb{E}(X_iZ_i) \neq 0$. Use these two assumptions on Z_i to express β as a function of γ_{xz} and γ_{yz}. Use this expression for β and to write down a consistent estimator for β, which we denote by $\widehat{\beta}_{\mathrm{IV}}$.

(d) Under the assumptions in (c), show that $\widehat{\beta}_{\mathrm{IV}}$ is asymptotically normal, i.e. show that $\sqrt{n}(\widehat{\beta}_{\mathrm{IV}} - \beta) \Rightarrow \mathcal{N}(0, \Sigma_{\mathrm{IV}})$ as $n \to \infty$. Provide a formula for the asymptotic variance Σ_{IV}.

Question 2

Consider the model with outcome $Y \in \mathbb{R}$, binary treatment $D \in \{1, 0\}$, and binary instrument $Z \in \{1, 0\}$. We observe the sample (Y_i, D_i, Z_i), $i = 1, \ldots, n$. For $z, d \in \{1, 0\}$

let

$$N_{zd} = \sum_{i=1}^{n} \mathbb{1}\left\{Z_i = z \,\&\, D_i = d\right\}$$

be the number of observations with instrument value z and treatment status d. We have $n = N_{11} + N_{10} + N_{01} + N_{00}$. Similarly, let

$$\overline{Y}_{zd} = \frac{1}{N_{zd}} \sum_{i=1}^{n} \mathbb{1}\left\{Z_i = z \,\&\, D_i = d\right\} Y_i$$

be the average outcome of units with instrument value z and treatment status d. Let $\widehat{\gamma}_1$ and $\widehat{\delta}_1$ be the OLS estimates for γ_1 and δ_1 in the first stage and reduced form regressions

$$D_i = \gamma_0 + \gamma_1 Z_i + \nu_i,$$
$$Y_i = \delta_0 + \delta_1 Z_i + v_i.$$

Let $\widehat{\beta} = \widehat{\delta}_1 / \widehat{\gamma}_1$ be the 2SLS estimator for the parameter β in the structural equation

$$Y_i = \alpha + \beta D_i + \epsilon_i. \tag{1}$$

(a) Provide expressions for $\widehat{\gamma}_1$, $\widehat{\delta}_1$, $\widehat{\beta}$ as functions of $(N_{zd}, \overline{Y}_{zd} : z, d \in \{1, 0\})$.

(b) Show how the formula for $\widehat{\beta}$ simplifies for the case that $N_{01} = 0$ and $N_{10} = 0$. This is the case where we only have compliers (i.e. we have $D = Z$, or $D(z) = z$ in potential outcome notation). Interpret the formula for $\widehat{\beta}$ in that case.

(c) Show how the formula for $\widehat{\beta}$ also simplifies for the case that $N_{00} = 0$ and $N_{11} = 0$. This is the case where we only have defiers (i.e. we have $D = 1 - Z$, or $D(z) = 1 - z$ in potential outcome notation). Interpret the formula for $\widehat{\beta}$ in that case.

(d) What happens to $\widehat{\beta}$ if $N_{10} = 0$ and $N_{00} = 0$? This is the case when we only have always takers (i.e. we have $D = 1$ for all units, or $D(z) = 1$ in potential outcome notation).

Now, assume that we initially have three separate datasets. The first dataset (labeled by a superscript "at") contains only always takers and therefore satisfies

$$N_{11}^{\text{at}} > 0, \qquad N_{10}^{\text{at}} = 0, \qquad N_{01}^{\text{at}} > 0, \qquad N_{00}^{\text{at}} = 0.$$

The second dataset (labeled by a superscript "nt") contains only never takers and therefore satisfies

$$N_{11}^{\text{nt}} = 0, \qquad N_{10}^{\text{nt}} > 0, \qquad N_{01}^{\text{nt}} = 0, \qquad N_{00}^{\text{nt}} > 0.$$

The third dataset (labeled by a superscript "c") contains only compliers and therefore satisfies

$$N_{11}^{\text{c}} > 0, \qquad N_{10}^{\text{c}} = 0, \qquad N_{01}^{\text{c}} = 0, \qquad N_{00}^{\text{c}} > 0.$$

Furthermore, let $n^{\text{at}} = N_{11}^{\text{at}} + N_{10}^{\text{at}} + N_{01}^{\text{at}} + N_{00}^{\text{at}}$ be the total number of always takers, and similarly let n^{nt} be the total number of never takes, and n^{c} be the total number of compliers. To simplify the computation we assume that

$$\frac{N_{11}^{\text{at}} + N_{10}^{\text{at}}}{n^{\text{at}}} = \frac{N_{11}^{\text{nt}} + N_{10}^{\text{nt}}}{n^{\text{nt}}} = \frac{N_{11}^{\text{c}} + N_{10}^{\text{c}}}{n^{\text{c}}}, \tag{2}$$

that is, the fraction of observations with $Z_i = 1$ is assumed to be exactly identical for the always takers, never takes, and compliers, respectively.

Finally, let $\overline{Y}_{zd}^{\text{at}}$, $\overline{Y}_{zd}^{\text{nt}}$, $\overline{Y}_{zd}^{\text{c}}$ be the average outcomes of units with instrument value z and treatment status d for the always takers, never takes, and compliers, respectively, and assume that

$$\overline{Y}_{0d}^{\text{at}} = \overline{Y}_{1d}^{\text{at}}, \qquad \overline{Y}_{0d}^{\text{nt}} = \overline{Y}_{1d}^{\text{nt}}. \tag{3}$$

The total dataset that we actually observe combines all the observations of these three datasets such that $N_{zd} = N_{zd}^{\text{at}} + N_{zd}^{\text{nt}} + N_{zd}^{\text{c}}$. However, in this combined dataset we do not actually know whether a particular observation is an always taker, a never taker or a complier. The definition of the 2SLS estimator $\widehat{\beta} = \widehat{\delta}_1/\widehat{\gamma}_1$ is unchanged. In particular, your result from part (a) is still applicable.

(e) Show that for the combined dataset we have

$$\widehat{\beta} = \overline{Y}_{11}^{\text{c}} - \overline{Y}_{00}^{\text{c}},$$

that is, $\widehat{\beta}$ obtained from the whole dataset is identical to the 2SLS estimator that would be obtained from the complier dataset only.

Remark: Let

$$\widehat{p}^{\text{at}} = \frac{N_{11}^{\text{at}} + N_{10}^{\text{at}}}{n^{\text{at}}}, \qquad \widehat{p}^{\text{nt}} = \frac{N_{11}^{\text{nt}} + N_{10}^{\text{nt}}}{n^{\text{nt}}}, \qquad \widehat{p}^{\text{c}} = \frac{N_{11}^{\text{c}} + N_{10}^{\text{c}}}{n^{\text{c}}},$$

be the fraction of observations with $Z_i = 1$ for the the always takers, never takes, and compliers, respectively. In display (2) we assumed that $\widehat{p}^{\text{at}} = \widehat{p}^{\text{nt}} = \widehat{p}^{\text{c}}$. In an actual dataset this will not be exactly the case, but under random assignment we expect that $\widehat{p}^{\text{at}} \approx \widehat{p}^{\text{nt}} \approx \widehat{p}^{\text{c}}$ in large datasets. Similarly, (3) will only hold approximately in an actual dataset under random assignment. If (2) and (3) only hold approximately, then the result in (e) will also only hold approximately in an actual dataset.