Econometrics

Homework 7

Problem 1. that (Y_i, X_i, Z_i), $i = 1, ..., n$ is a sequence of i.i.d. discrete random vectors and $Y_i \in \{0, 1, 2\}$, $Z_i \in \{0, 1\}$ and $X_i \in \{0, 1\}$.

(i) Show that for any $a \in \{0, 1\}$, we have

$$\begin{aligned} \mathrm{E}[Y_i|X_i = a] =& \mathrm{E}[Y_i|X_i = a, Z_i = 0] \mathrm{P}[Z_i = 0|X_i = a] \\ &+ \mathrm{E}[Y_i|X_i = a, Z_i = 1] \mathrm{P}[Z_i = 1|X_i = a]. \end{aligned}$$

(ii) Show $\mathrm{E}[Z_i X_i] = \mathrm{P}[Z_i = 1, X_i = 1]$.

(iii) Show $\mathrm{E}[\mathrm{E}[Z_i|X_i = 1] X_i] = \mathrm{E}[Z_i X_i]$.

(iv) Show that $\hat{\theta} = \frac{\sum_{i=1}^n Z_i X_i}{\sum_{i=1}^n X_i}$ is a consistent estimator of $\theta = \mathrm{P}[Z_i = 1|X_i = 1]$.

(v) Find a formula for σ^2 such that

$$\sqrt{n}\left(\hat{\theta} - \theta\right) \to_d \mathrm{N}\left(0, \sigma^2\right).$$

Solution.

(i) By LIE, we have $\mathrm{E}[Y|X] = \mathrm{E}[\mathrm{E}[Y|X, Z]|X]$. Notice that $\mathrm{E}[Y|X, Z]$ is a function of (X, Z). Once we know $X = a$, the randomness of $\mathrm{E}[Y|X = a, Z]$ is due to the randomness of Z solely. We now have

$$\mathrm{E}[Y|X = a] = \mathrm{P}[Z = 1|X = a] \mathrm{E}[Y|X = a, Z = 1] + \mathrm{P}[Z = 0|X = a] \mathrm{E}[Y|X = a, Z = 0].$$

(ii)

$$\begin{aligned} \mathrm{E}[ZX] =& \mathrm{P}[X = 1, Z = 1] \cdot 1 + \mathrm{P}[X = 1, Z = 0] \cdot 0 \\ &+ \mathrm{P}[X = 0, Z = 1] \cdot 0 + \mathrm{P}[X = 0, Z = 0] \cdot 0 \\ =& \mathrm{P}[X = 1, Z = 1]. \end{aligned}$$

(iii) Notice that $\mathrm{E}[Z|X = 1]$ is a constant.

$$\begin{aligned} \mathrm{E}[\mathrm{E}[Z|X = 1] X] =& \mathrm{E}[Z|X = 1] \mathrm{E}[X] \\ =& \mathrm{P}[Z = 1|X = 1] \mathrm{P}[X = 1] \\ =& \mathrm{P}[Z = 1, X = 1] \\ =& \mathrm{E}[ZX], \end{aligned}$$

where the last equality follows from Part (ii).

(iv) By Slutsky's lemma and Part (iii), we have

$$\hat{\theta} = \frac{\sum_{i=1}^n Z_i X_i}{\sum_{i=1}^n X_i} = \frac{\frac{1}{n}\sum_{i=1}^n Z_i X_i}{\frac{1}{n}\sum_{i=1}^n X_i} \to_p \frac{\mathrm{E}[ZX]}{\mathrm{E}[X]} = \mathrm{P}[Z = 1|X = 1].$$

(v) Denote $\epsilon_i = Z_i - \mathrm{E}[Z|X = 1]$. Now we have

$$\hat{\theta} = \frac{\sum_{i=1}^n Z_i X_i}{\sum_{i=1}^n X_i} = \frac{\sum_{i=1}^n (\mathrm{E}[Z|X = 1] + \epsilon_i) X_i}{\sum_{i=1}^n X_i} = \mathrm{E}[Z|X = 1] + \frac{\sum_{i=1}^n \epsilon_i X_i}{\sum_{i=1}^n X_i}$$

which gives

$$\sqrt{n}\left(\hat{\theta} - \theta\right) = \frac{\frac{1}{\sqrt{n}}\sum_{i=1}^n \epsilon_i X_i}{\frac{1}{n}\sum_{i=1}^n X_i}.$$

By LLN, $\frac{1}{n}\sum_{i=1}^n X_i \to_p \mathrm{E}[X]$. By Part (iii),

$$\mathrm{E}[\epsilon_i X_i] = \mathrm{E}[(Z_i - \mathrm{E}[Z_i|X_i = 1]) X_i] = 0.$$

By CLT, $\frac{1}{\sqrt{n}}\sum_{i=1}^n \epsilon_i X_i \to_d \mathrm{N}(0, \mathrm{E}[\epsilon_i^2 X_i^2])$. By Slutsky's lemma and the lemma on Page 7 of Lecture 17, we have

$$\sqrt{n}\left(\hat{\theta} - \theta\right) \to_d \mathrm{N}\left(0, \frac{\mathrm{E}[\epsilon_i^2 X_i^2]}{\mathrm{E}[X_i]^2}\right).$$

Problem 2. Let $\{(Y_i, X_i, D_i)\}_{i=1}^n$ be a sequence of i.i.d. observations. D_i is a dummy variable. Consider the following binary choice model:

$$Y_i = 1(\beta_0 + \beta_1 X_i + \beta_2 X_i D_i \geqslant U_i),$$

where the conditional CDF of U_i is given by

$$\mathrm{P}[U_i \leqslant t|X_i, D_i] = \frac{\exp(t)}{1 + \exp(t)}.$$

(i) Define and derive the expression of the log-likelihood function for the i.i.d. observations $\{(Y_i, X_i, D_i)\}_{i=1}^n$.

(ii) Derive the average derivative (or average partial effect) with respect to X_i in terms of the observations and the parameters.

(iii) Let the MLE's for β_0, β_1 and β_2 be denoted by $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\beta}_2$. Provide an estimator of the average derivative in (ii).

Solution.

(i) Define

$$G(t) = \frac{\exp(t)}{1 + \exp(t)}.$$

Then by the chain rule for differentiation, we have

$$g(t) = \frac{\mathrm{d}G(t)}{\mathrm{d}t} = \frac{\exp(t)}{(1+\exp(t))^2}.$$

By construction of the model, we have

$$\begin{aligned}\mathrm{P}[Y_i = 1|X_i, D_i] =& \mathrm{P}[\beta_0 + \beta_1 X_i + \beta_2 X_i D_i \geqslant U_i | X_i, D_i] \\ =& \frac{\exp(\beta_0 + \beta_1 X_i + \beta_2 X_i D_i)}{1 + \exp(\beta_0 + \beta_1 X_i + \beta_2 X_i D_i)} \\ =& G(\beta_0 + \beta_1 X_i + \beta_2 X_i D_i)\end{aligned}$$

and

$$\mathrm{P}[Y_i = 0|X_i, D_i] = 1 - G(\beta_0 + \beta_1 X_i + \beta_2 X_i D_i).$$

Denote $Z = \{(Y_i, X_i, D_i)\}_{i=1}^n$ for simplicity. The likelihood function is

$$L(b_0, b_1, b_2; Z) = \prod_{i=1}^n G(b_0 + b_1 X_i + b_2 X_i D_i)^{Y_i} (1 - G(\beta_0 + \beta_1 X_i + \beta_2 X_i D_i))^{1-Y_i}$$

and the corresponding log-likelihood function is

$$\ell(b_0, b_1, b_2; Z) = \sum_{i=1}^n \{Y_i \log(G(b_0 + b_1 X_i + b_2 X_i D_i)) + (1 - Y_i)\log(1 - G(b_0 + b_1 X_i + b_2 X_i D_i))\}$$

(ii)

$$\begin{aligned}\frac{\partial \mathrm{E}[Y_i|X_i = x, D_i = d]}{\partial x} =& \frac{\partial \mathrm{P}[Y_i = 1|X_i = x, D_i = d]}{\partial x} \\ =& g(\beta_0 + \beta_1 x + \beta_2 x d)(\beta_1 + \beta_2 d).\end{aligned}$$

The average derivative is

$$\mathrm{E}[g(\beta_0 + \beta_1 X_i + \beta_2 X_i D_i)(\beta_1 + \beta_2 D_i)]. \tag{1}$$

(iii) The "sample analogue" of (1) estimator is

$$\frac{1}{n}\sum_{i=1}^n g\left(\hat\beta_0 + \hat\beta_1 X_i + \hat\beta_2 X_i D_i\right)\left(\hat\beta_1 + \hat\beta_2 D_i\right).$$

Problem 3. In this question, you will derive the asymptotic distribution of the OLS estimator under endogeneity. Consider the usual linear regression model (without intercept) $Y_i = \beta X_i + U_i$. Assume, however, that X_i is endogenous:

$$E(X_i U_i) = \mu \neq 0,$$

where μ is unknown. Let $\hat\beta_n$ denote the OLS estimator of β. Make the following additional assumptions:

A1. Data are iid.

A2. $0 < Q = E(X_i^2) < \infty$.

A3. $0 < E(U_i - \delta X_i) X_i^2 < \infty$, where $\delta = Q^{-1}\mu$.

(i) Find the probability limit of $\hat{\beta}_n$.

(ii) Re-write the model as $Y_i = (\beta + \delta)X_i + (U_i - \delta X_i)$ and find $E(X_i(U_i - \delta X_i))$.

(iii) Using the result in (ii), derive the asymptotic distribution of $\hat{\beta}_n$ and find its asymptotic variance. Explain how this result differs from the asymptotic normality of OLS with exogenous regressors.

(iv) Can $\hat{\beta}_n$ and its asymptotic distribution be used for constructing a confidence interval about β? Explain why or why not.

(v) Suppose that the errors U_i's are homoskedastic:

$$E\left(U_i^2 | X_i\right) = \sigma^2 = constant.$$

Consider the usual estimator of the asymptotic variance of OLS designed for a model with homoskedastic errors and exogenous regressors:

$$\left(n^{-1} \sum_{i=1}^{n} \left(Y_i - \hat{\beta}_n X_i\right)^2\right) \left(n^{-1} \sum_{i=1}^{n} X_i^2\right)^{-1}.$$

Is it consistent for the asymptotic variance of the OLS estimator if X_i's are in fact endogenous? Explain why or why not.

Solution.

(i) Write

$$\begin{aligned} \hat{\beta}_n &= \beta + \frac{\frac{1}{n}\sum_{i=1}^n X_i U_i}{\frac{1}{n}\sum_{i=1}^n X_i^2} \\ &\to_p \beta + Q^{-1}\mu \\ &= \beta + \delta, \end{aligned}$$

where convergence of $n^{-1}\sum_{i=1}^n X_i^2 \to_p Q$ and $n^{-1}\sum_{i=1}^n X_i U_i \to_p E(X_i U_i) = \mu$ hold by the WLLN.

(ii)

$$\begin{aligned} E\left(X_i(U_i - \delta X_i)\right) &= E\left(X_i U_i\right) - E\left(X_i^2\right) Q^{-1}\mu \\ &= \mu - QQ^{-1}\mu \\ &= 0. \end{aligned}$$

(iii) Write

$$\hat{\beta}_n - (\beta + \delta) = \frac{\frac{1}{n}\sum_{i=1}^n X_i \epsilon_i}{\frac{1}{n}\sum_{i=1}^n X_i^2},$$

where

$$\epsilon_i = U_i - \delta X_i$$

and uncorrelated with X_i by the result in (ii). Furthermore, $X_i\epsilon_i$ satisfies the assumptions of the CLT. Hence, this is a regression with all the usual assumptions, however, it has a new regression coefficient $\beta+\delta$ and new errors ϵ_i's. We have:

$$\sqrt{n}\left(\hat{\beta}_n-(\beta+\delta)\right)\to_d N\left(0,Q^{-2}E\left(U_i-\delta X_i\right)^2X_i^2\right).$$

Comparing to the case with exogenous regressors, the center of the asymptotic distribution is shifted by δ. Also, the asymptotic variance depends on δX_i through $E\left(U_i-\delta X_i\right)^2X_i^2$.

(iv) Asymptotic inference about β based on the OLS estimator will be invalid since the asymptotic distribution of the OLS estimator is centered at $\beta+\delta$. The OLS estimator can be only used for testing hypotheses about $\beta+\delta$.

(v) First, we need to describe the probability limit of the estimator proposed. Write:

$$\begin{aligned}n^{-1}\sum_{i=1}^n\left(Y_i-\hat{\beta}_nX_i\right)^2 &= n^{-1}\sum_{i=1}^n\left((U_i-\delta X_i)+\left(\beta+\delta-\hat{\beta}_n\right)X_i\right)^2\\ &= n^{-1}\sum_{i=1}^n\left(\epsilon_i+\left(\beta+\delta-\hat{\beta}_n\right)X_i\right)^2,\end{aligned}$$

where

$$\epsilon_i=U_i-\delta X_i.$$

In view of the result in (i), $\beta+\delta-\hat{\beta}_n\to_p 0$, and therefore

$$n^{-1}\sum_{i=1}^n\left(Y_i-\hat{\beta}_nX_i\right)^2\to_p E\left(\epsilon_i^2\right).$$

Hence, the proposed estimator converges in probability to $E\left(U_i-\delta X_i\right)^2Q^{-1}$. This would be the same as the asymptotic variance in (iii) if the errors $\epsilon_i=U_i-X_i'\delta$ were homoskedastic. It is given that U_i's are homoskedastic. However, even if U_i's are homoskedastic, $\epsilon_i=U_i-\delta X_i$ would be heteroskedastic:

$$E(\epsilon_i^2|X_i)=\sigma^2+(\delta X_i)^2-2E\left(U_i|X_i\right)\delta X_i\neq constant,$$

unless $E\left(U_i|X_i\right)=0.5\delta X_i$. Since $\delta=Q^{-1}\mu$, and $\mu=E\left(X_iU_i\right)$, the law of iterated expectation implies that if $E\left(U_i|X_i\right)=0.5\delta X_i$, then

$$\begin{aligned}\mu &= E\left(X_iU_i\right)\\ &= E\left(X_iE(U_i|X_i)\right)\\ &= E\left(X_i\times 0.5\delta X_i\right)\\ &= 0.5Q\delta\\ &= 0.5Q\times Q^{-1}\mu\\ &= 0.5\mu.\end{aligned}$$

However, the only solution to $\mu=0.5\mu$ is $\mu=0$, which contradicts the assumption that $E\left(X_iU_i\right)\neq 0$. It follows therefore that $\epsilon_i=U_i-\delta X_i$ are heteroskedastic. Hence, the estimator would be inconsistent for the asymptotic variance of the OLS estimator.

Problem 4. Consider the model

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + U_i, \tag{2}$$

where X_{1i} is an exogenous regressor and X_{2i} is an endogenous regressor. Assume that data are iid and conditions required for LLNs hold. For each of the following statements, indicate true or false, and explain your answer.

(i) Let $\hat{\beta}_1$ denote the estimated coefficient on X_1 in the OLS regression of Y against a constant, X_1, and X_2. Since X_1 is exogenous, $\hat{\beta}_1$ consistently estimates β_1.

(ii) Let $\hat{\beta}_1$ denote the estimated coefficient on X_1 in the OLS regression of Y against a constant and X_1. If $Cov(X_{1i}, X_{2i}) = 0$, then $\hat{\beta}_1$ consistently estimates β_1.

(iii) Consider the following IV estimator of β_2 that uses X_1 as an IV:

$$\hat{\beta}_2 = \frac{\sum_{i=1}^{n}(X_{1i} - \bar{X}_1)Y_i}{\sum_{i=1}^{n}(X_{1i} - \bar{X}_1)X_{2i}}.$$

If $Cov(X_{1i}, X_{2i}) \neq 0$ and $\beta_1 = 0$, then $\hat{\beta}_2$ consistently estimates β_2.

Solution.

(i) False. If X_1 and X_2 are correlated, $\hat{\beta}_1$ is inconsistent. Let $\tilde{X}_{1i}$ denote fitted residuals in the regression of X_1 against a constant and X_2:

$$\tilde{X}_{1i} = X_{1i} - \hat{\gamma}_0 - \hat{\gamma}_1 X_{2i},$$

where $\hat{\gamma}$'s denote the OLS estimators.

$$\begin{aligned} \hat{\beta}_1 &= \frac{\sum \tilde{X}_{1i} Y_i}{\sum \tilde{X}_{1i}^2} \\ &= \beta_1 + \frac{n^{-1}\sum \tilde{X}_{1i} U_i}{n^{-1}\sum \tilde{X}_{1i}^2}. \end{aligned}$$

Next,

$$n^{-1}\sum \tilde{X}_{1i} U_i = n^{-1}\sum X_{1i} U_i - \hat{\gamma}_0 n^{-1}\sum U_i - \hat{\gamma}_1 n^{-1}\sum X_{2i} U_i.$$

Since X_{1i} is exogenous,

$$n^{-1}\sum X_{1i} U_i \to_p 0.$$

We can also expect that

$$n^{-1}\sum U_i \to_p 0.$$

However, since X_{2i} is endogenous,

$$n^{-1}\sum X_{2i} U_i \to_p E X_{2i} U_i \neq 0.$$

Note also that

$$\hat{\gamma}_1 = \frac{n^{-1}\sum (X_{2i} - \bar{X}_2) X_{1i}}{n^{-1}\sum (X_{2i} - \bar{X}_2)^2} \to_p \frac{Cov(X_{2i}, X_{1i})}{Var(X_{2i})}.$$

Hence, if X_1 and X_2 are correlated, then $\hat{\beta}_1$ will be inconsistent.

(ii) True. Write

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_{1i} + V_i, \\ V_i &= \beta_2 X_{2i} + U_i. \end{aligned}$$

We have $Cov(X_{1i}, V_i) = \beta_2 Cov(X_{1i}, X_{2i}) + Cov(X_{1i}, U_i)$. Since X_1 is exogenous in the original model, $Cov(X_{1i}, U_i) = 0$. If $Cov(X_{1i}, X_{2i}) = 0$, then X_1 is uncorrelated with V in the new regression equation and, therefore, exogenous. Hence, $\hat{\beta}_1$ is a consistent estimator.

(iii) True. Since $\beta_1 = 0$, X_1 is excluded from the structural equation. By the assumption, X_1 and U are uncorrelated. Since X_1 and X_2 are correlated, X_1 is a valid IV.