Applications of Econometric & Data Science Methods

Homework 2

Due Date: Monday, April 11, 18:30 CST

Note: all vectors are column vectors.

## 1. [35 pts] Frisch-Waugh Theorem with a Realized Sample

Suppose we have $n$ observations of an outcome with $p$ covariates: $\{(y_i, x_i)\}_{i=1}^n$. In matrix form, we observe $y$ and $X$. We divide our covariates in two groups, so that $x_i = (x_{i1}, x_{i2})$ for all $i$, or $X = (X_1, X_2)$. Suppose the first $p_1$ covariates are in group 1. Let $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2)$ be the best linear predictor of $y$ using $X$ and consider the following matrices:

$$P_1 = X_1(X_1'X_1)^{-1}X_1' \; ; \quad M_1 = I - P_1.$$

Let $\tilde{X}_2 = M_1 X_2$ and $\tilde{y} = M_1 y$.

### a. [3 pts]

What are the dimensions of $y$? Of $X_2$? Of $\hat{\beta}_1$?

### b. [5 pts]

Show that the *normal equations* – the ones that give rise to $\hat{\beta}$ – are:

$$\begin{aligned} X_1'X_1\hat{\beta}_1 + X_1'X_2\hat{\beta}_2 &= X_1'y \\ X_2'X_1\hat{\beta}_1 + X_2'X_2\hat{\beta}_2 &= X_2'y \end{aligned}$$

### c. [5 pts]

Show that

$$X_1\hat{\beta}_1 + P_1X_2\hat{\beta}_2 = P_1y.$$

## d. [5 pts]

Show that

$$\tilde{y} = \tilde{X}_2 \hat{\beta}_2 + \hat{\epsilon},$$

where $\hat{\epsilon}$ is the error from the projection of $y$ on $X$.

## e. [5 pts]

Show that

$$\hat{\beta}_2 = \left( \tilde{X}_2' \tilde{X}_2 \right)^{-1} \tilde{X}_2' \tilde{y}.$$

## f. [5 pts]

Explain your result in words.

## g. [7 pts]

Observation $i$, $(y_i, x_i)$, is a realization of random vector $(Y_i, \mathbf{X}_i)$ where $\mathbf{X}_i = (\mathbf{X}_i^1, \mathbf{X}_i^p)$, so that our data as a whole can be seen as a realization of random vector $\big((Y_1, \mathbf{X}_1), (Y_2, \mathbf{X}_2), \ldots, (Y_n, \mathbf{X}_n)\big)$.

Answer Yes or No.

(i) [2 pts] Does your result require that $(Y_i, \mathbf{X}_i)$ and $(Y_j, \mathbf{X}_j)$ be identically distributed for all $i, j \in \{1, \ldots, n\}$?

(ii) [2 pts] Does your result require that $\big\{(Y_1, \mathbf{X}_1), (Y_2, \mathbf{X}_2), \ldots, (Y_n, \mathbf{X}_n)\big\}$ be mutually independent?

(iii) [1 pts] Does your result require that $\big((Y_1, \mathbf{X}_1), (Y_2, \mathbf{X}_2), \ldots, (Y_n, \mathbf{X}_n)\big)$ be i.i.d.?

(iv) [2 pts] Suppose there are random variables $(W_1, \ldots, W_n)$ whose realizations you do not observe and that, for some $i \in \{1, \ldots, n\}$,

$$\mathrm{Cov}(\mathbf{X}_i, W_i) \equiv \begin{pmatrix} \mathrm{Cov}(\mathbf{X}_i^1, W_i) \\ \vdots \\ \mathrm{Cov}(\mathbf{X}_i^p, W_i) \end{pmatrix} > \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}.$$

Does your result require that $\mathrm{Cov}(Y_i, W_i) = 0$?

## 2. [35 pts] Measurement Error in OLS

Consider an outcome $y$ and covariates $x_1, \dots, x_K$. These are all random variables. Now fix $\beta = (\beta_0, \dots, \beta_K) \in \mathbb{R}^{K+1}$ such that

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_K x_K + \epsilon$$

where $\mathbb{E}[\epsilon] = 0$ and $\mathbb{E}[\epsilon x_k] = 0$ for each $k \in \{1, \dots, K\}$.

Suppose we have an i.i.d. sample $\{(y_i, 1, x_{1i}, \dots, x_{Ki})\}_{i=1}^N$. Let $x_i = (1, x_{1i}, \dots, x_{Ki})$, $X = (x_1, \dots, x_N)$ and $\mathbf{y} = (y_1, \dots, y_N)$. $X$ has dimensions $N \times (K+1)$. Assume away multicollinearity in population and in sample. Under these conditions, recall that

$$\sqrt{N}(\hat{\beta} - \beta) \xrightarrow{d} N\left(0,\ \mathbb{E}[x_i x_i']^{-1}\mathbb{E}[\epsilon_i^2 x_i x_i']\mathbb{E}[x_i x_i']^{-1}\right),$$

where $\xrightarrow{d}$ denotes convergence in distribution as $N$ grows to infinity, and $\hat{\beta} = (X'X)^{-1}X'\mathbf{y}$.

### a. [5 pts] Homoskedasticity

Suppose in addition that

(i) $\mathrm{Var}(y_i \mid x_i) = \sigma^2$

(ii) $\mathbb{E}[y_i \mid x_{1i}, \dots, x_{Ki}] = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_K x_{Ki}$.

Show that

$$\sqrt{N}(\hat{\beta} - \beta) \xrightarrow{d} N\left(0,\ \sigma^2 \mathbb{E}[x_i x_i']^{-1}\right).$$

### b. [15 pts] Measurement Error in the Outcome Variable

Suppose that we can only work with an imperfect measure of $y$: $\tilde{y}$, where

$$\tilde{y} = y + u_y.$$

(i) [6 pts] What assumption(s) is (are) required to consistently estimate $\beta$ by simply working with $(\tilde{y}_1, \dots, \tilde{y}_N)$ instead of the (unobservable) $y_i$'s?

(ii) [6 pts] Suppose your assumption(s) in (i) hold. For simplicity, assume homoskedasticity and suppose that $\mathrm{Var}(u_y \mid x_1, \dots, x_K) = \sigma_u^2$. Moreover, suppose that $u_y$ and $\epsilon$ are uncorrelated. How does the asymptotic variance of $\tilde{\beta} \equiv (X'X)^{-1}X'\tilde{\mathbf{y}}$ compare with that of $\hat{\beta}$?

(iii) [3 pts] Under the assumptions in (ii), if you had access to both $y$ and $\tilde{y}$, would it be a good idea to work with $\tilde{y}$? Justify you answer.

## c. [15 pts] Measurement Error in the Covariates

Suppose instead that we observe $y, x_1, \ldots, x_{K-1}$ perfectly, but we can only work with an imperfect measure of $x_K$: $\tilde{x}_K$, where

$$\tilde{x}_K = x_K + u_K.$$

(i) [5 pts] What assumption(s) is (are) required to consistently estimate $\beta$ by simply working with $(\tilde{x}_{K1}, \ldots, \tilde{x}_{KN})$ instead of the (unobservable) $x_{Ki}$'s?

(ii) [1 pts] If your assumption(s) in (i) hold and $u_K$ and $\epsilon$ are uncorrelated. Is it a good idea to work with $\tilde{x}_K$ if you had access to $\tilde{x}_K$ and $x_K$? Why, or why not?

(iii) [4 pts] Clear your assumptions and suppose that $\text{Cov}(x_K, u_K) = 0$. Can $\beta$ be consistently estimated by working with $\tilde{x}_K$ instead of $x_K$? Why, or why not?

(iv) [5 pts] Clear all assumptions. Consider now the simpler case where $K = 1$. Show that the probability limit of $\tilde{\beta}_1$ when using $\tilde{x}_1$ instead of $x_1$ is

$$\operatorname*{plim}_{N \to \infty} \tilde{\beta}_1 = \frac{\beta_1[\text{Var}(x_{1i}) + \text{Cov}(x_{1i}, u_{1i})] + \text{Cov}(\epsilon_i, u_{1i})}{\text{Var}(x_{1i}) + \text{Var}(u_{1i}) + 2\text{Cov}(x_{1i}, u_{1i})}.$$

What happens when the measurement error is "random", i.e. independent of $x_1$ and $\epsilon$? In metrics slang, this is called the *attenuation bias*.

## 3. [30 pts] Checking the Exclusion Restriction

Consider an outcome $y$ and $K$ covariates, $x_1, \dots, x_K$. These are all random variables. You know that, holding everything else constant (*caeteris paribus*), an increase in the realization of $x_k$ of one unit increases the realized outcome by $\beta_k$ units, for all $k \in \{1, \dots, K\}$.

Let $\epsilon = y - \big(\beta_0 + \beta_1 x_1 + \cdots + \beta_K x_K\big)$, where $\mathbb{E}[\epsilon] = 0$ (or $\beta_0 \equiv \mathbb{E}[y - (\beta_1 x_1 + \cdots + \beta_K x_K)]$). While $x_1, \dots, x_{K-1}$ are all exogenous, you suspect that $x_K$ is endogenous, i.e. correlated with $\epsilon$. To measure $\beta_0, \dots, \beta_K$ you come up with a plausible instrument $z_K$. It is a strong instrument, and you would like to test whether it is also excluded, i.e. uncorrelated with $\epsilon$.

### a. [10 pts] An ideal case

Suppose someone gave you $\beta_0, \dots, \beta_K$.

(i) [3 pts] Do you then observe the realizations of $\epsilon$ in any given dataset containing realizations of $y$ and the $K$ covariates? Answer Yes or No (one word).

(ii) [7 pts] For a given sample with $n$ observations, you can form the statistic $\widehat{\text{Cov}}(z_K, y - [\beta_0 + \beta_1 x_1 + \cdots + \beta_K x_K])$, a sample covariance. Derive the asymptotic distribution of this statistic. How would you test for the exclusion restriction of $z_K$?

*Hint: Use the central limit theorem and the delta method. You DO NOT need to give an explicit form for the variance-covariance matrix.*

### b. [20 pts] Real life

Now you do not know what $\beta_0, \dots, \beta_K$ are. Your goal is the same: to check whether $z_K$ satisfies the exclusion restriction.

Let $x = (1, x_1, \dots, x_K)$ and $z \equiv (1, x_1, \dots, x_{K-1}, z_K)$.

(i) [7 pts] Show that the residual from the IV regression that uses $z$ as an instrument for $x$ is:

$$\epsilon^{IV} = y - x'\mathbb{E}[zx']^{-1}\mathbb{E}[zy]$$

(ii) [3 pts] What is $\mathbb{E}[z\epsilon^{IV}]$?

(iii) [10 pts] Reflect and comment: Is your goal met, i.e. have you found a way to test the exclusion restriction of $z$?