ECON 468 Final Review

Tiffany Yong

November 27, 2022

This is the notes packet for the final exam review session on December 1st, 2022 for the Econometrics I - Honours class (ECON 468, McGill University) given by Saraswata Chaudhuri. The topic covered in this class is linear regression and topics in regression analysis.

Contents

1 Review on Statistical Theory

1.1 Useful Theorems

Theorem 1 (Law of Iterated Expectation)**.** *Let X and Y be two random variables. Then,*

$$E[Y] = E[E[Y|X]].$$

Theorem 2 (Law of Total Variance)**.** *Let X and Y be two random variables. Then,*

$$Var(Y) = E[Var(Y|X)] + Var(E[Y|X]).$$

Definition 3 (Convergence in Probability)**.** We say that X_n converges to μ in probability, denoted $X_n \xrightarrow[n\to\infty]{P} \mu$ if for any $\epsilon > 0$ and $\delta > 0$, there exists N such that

$$P(|X_n - \mu| > \epsilon) < \delta \quad \forall n > N.$$

Theorem 4 (Weak Law of Large Numbers)**.** *Let $X_1, X_2, \ldots, X_n$ be i.i.d copies of a random variable X, with $\mu_X = E[X]$. Let $\overline{X}$ be the sample mean of these n observations. Then,*

$$\overline{X} \xrightarrow[n\to\infty]{P} \mu_X.$$

Definition 5 (Convergence in Distribution)**.** We say that a sequence of random variables $X_1, X_2, \ldots$ converges in distribution to Z, denoted $X_n \xrightarrow{d} Z$ if

$$\lim_{n\to\infty} F_n(x) = F(x) \quad \forall x \in \mathbb{R} \text{ s.t. } F \text{ continuous},$$

where $F_n(x)$ is the CDF of X_n and $F(x)$ is the CDF of Z.

Theorem 6 (Central Limit Theorem)**.** *Let $X_1, X_2, \ldots, X_n$ be i.i.d copies of a random variable X, with $\mu_X = E[X]$ and $\sigma_X^2 = Var(X)$. Let $\overline{X}$ be the sample mean of these n observations. Then,*

$$\frac{\overline{X} - \mu_X}{\sqrt{\sigma_X^2/n}} \xrightarrow{d} N(0,1).$$

2 OLS Regression

2.1 Derivation of OLS estimators

The OLS estimators

$$\hat{\beta} = \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_k \end{pmatrix}$$

are the estimators that minimise the sum of squared errors:

$$\hat{\beta} \coloneqq \arg\min_b \sum_{i=1}^{n} (y_i - X_i^T b)^2 = \arg\min_b f(b).$$

Minimising this sum $f(b)$ requires a little bit of calculus for optimisation. We take the partial derivative of $f(b)$ with respect to each b_l, where $1 \leq l \leq k$. Then, we get k equations of the form:

$$\frac{\partial}{\partial b_l} = -2 \sum_{i=1}^{n} (y_i - X_i^T b) x_{li}$$

In order to minimise $f(b)$, we use the first order conditions and set each partial derivative to 0, thus getting:

$$-2 \sum_{i=1}^{n} x_{li}(y_i - X_i^T \hat{\beta}) = 0 \quad \forall 1 \leq l \leq k$$

Since this is true for all $1 \leq l \leq k$, we can turn these k equations into a single matrix condition:

$$\begin{pmatrix} \sum_{i=i}^{n} x_{1i}(y_i - X_i^T \hat{\beta}) \\ \sum_{i=i}^{n} x_{2i}(y_i - X_i^T \hat{\beta}) \\ \vdots \\ \sum_{i=i}^{n} x_{ki}(y_i - X_i^T \hat{\beta}) \end{pmatrix} = 0$$

This is algebraically equivalent to the single condition:

$$\sum_{i=1}^{n} X_i(y_i - X_i^T \hat{\beta}) = 0$$

Then, solving for $\hat{\beta}$, with $A_i \coloneqq \left(\sum_{i=1}^{n} X_i X_i^T\right)^{-1} X_i$, we get:

$$\begin{aligned} &\sum_{i=1}^{n} X_i(y_i - X_i^T \hat{\beta}) = 0 \\ &\Rightarrow \sum_{i=1}^{n} X_i y_i - \sum_{i=1}^{n} X_i X_i^T \hat{\beta} = 0 \\ &\Rightarrow \hat{\beta} = \left(\sum_{i=1}^{n} X_i X_i^T\right)^{-1} \sum_{i=1}^{n} X_i y_i \\ &\qquad = \sum_{i=1}^{n} A_i y_i \end{aligned}$$

The setup here is n observations, with k regressors, so:

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \quad X = \begin{pmatrix} 1 & 1 & \dots & 1 \\ x_{11} & x_{12} & \dots & x_{1n} \\ x_{21} & x_{22} & \dots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{k1} & x_{k2} & \dots & x_{kn} \end{pmatrix}$$

Now that we have derived the general form of $\hat{\beta}$, we can also derive the special case where $k = 1$, and our regression has the form:

$$y_i = \beta_0 + \beta_1 x_{1i} + u_i.$$

Since one of the first order conditions give us $\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - x_{1i}\hat{\beta}_1) = 0$, we have that

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{X}.$$

Then, to get $\hat{\beta}_1$

$$\begin{aligned}
\hat{\beta} &= \left(\sum_{i=1}^{n} X_i X_i^T\right)^{-1} \sum_{i=1}^{n} X_i y_i \\
&= \left(\sum_{i=1}^{n} \begin{pmatrix} 1 \\ x_{1i} \end{pmatrix} \begin{pmatrix} 1 & x_{1i} \end{pmatrix}\right)^{-1} \sum_{i=1}^{n} \begin{pmatrix} 1 \\ x_{1i} \end{pmatrix} y_i \\
&= \begin{pmatrix} n & \sum_{i=1}^{n} x_{1i} \\ \sum_{i=1}^{n} x_{1i} & \sum_{i=1}^{n} x_{1i}^2 \end{pmatrix}^{-1} \begin{pmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_{1i} \end{pmatrix} y_i \\
&= \frac{1}{n\sum_{i=1}^{n} x_{1i}^2 - (\sum_{i=1}^{n} x_{1i})^2} \begin{pmatrix} \sum_{i=1}^{n} x_{1i}^2 & -\sum_{i=1}^{n} x_{1i} \\ -\sum_{i=1}^{n} x_{1i} & n \end{pmatrix} \begin{pmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_{1i} y_i \end{pmatrix} \\
&= \frac{1}{n(\sum_{i=1}^{n} x_{1i}^2 - \overline{x}_1^2)} \begin{pmatrix} \hat{\beta}_0 \cdot n(\sum_{i=1}^{n} x_{1i}^2 - \overline{x}_1^2) \\ -\sum_{i=1}^{n} x_{1i} \sum_{i=1}^{n} y_i + n\sum_{i=1}^{n} x_{1i} y_i \end{pmatrix} \\
&= \frac{1}{n\sum_{i=1}^{n}(x_{1i} - \overline{x}_1)^2} \begin{pmatrix} \hat{\beta}_0 \cdot n\sum_{i=1}^{n}(x_{1i} - \overline{x}_1)^2 \\ n(\sum_{i=1}^{n} x_{1i} y_i - n\overline{x}_1 \overline{y}) \end{pmatrix} \\
&= \frac{1}{n\sum_{i=1}^{n}(x_{1i} - \overline{x}_1)^2} \begin{pmatrix} \hat{\beta}_0 \cdot n\sum_{i=1}^{n}(x_{1i} - \overline{x}_1)^2 \\ n\sum_{i=1}^{n}(x_{1i} - \overline{x}_1)(y_i - \overline{y}) \end{pmatrix} \\
&= \begin{pmatrix} \hat{\beta}_0 \\ \frac{\sum_{i=1}^{n}(x_{1i} - \overline{x}_1)(y_i - \overline{y})}{\sum_{i=1}^{n}(x_{1i} - \overline{x}_1)^2} \end{pmatrix}
\end{aligned}$$

The inverse of a 2×2 matrix $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$ is

$$\frac{1}{ad - bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$$

The following identity $(*)$ simplifies the equations:

$$\begin{aligned}
&\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) \\
&= \sum_{i=1}^{n} x_i y_i - \overline{x}\sum_{i=1}^{n} y_i - \overline{y}\sum_{i=1}^{n} x_i + n\overline{x}\overline{y} \\
&= \sum_{i=1}^{n} x_i y_i - n\overline{x}\overline{y}
\end{aligned}$$

This is very messy, but the key part of this derivation is that you ultimately get

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_{1i} - \overline{x}_1)(y_i - \overline{y})}{\sum_{i=1}^{n}(x_{1i} - \overline{x}_1)^2} = \frac{s_{XY}}{s_X^2}.$$

Getting an exact value of $\hat{\beta}_0$ and $\hat{\beta}_1$ allows us to do other analysis. Specifically, it allows us to test whether this estimator is even a useful measure of causal effect of X on Y. But to do that, we also need a few more additional assumptions briefly covered in the next section.

2.2 *Assumptions of OLS*

There are 4 key assumptions in order for the OLS estimator to have useful causal properties with a small sample:

1. $E(u_i|X_i) = 0$, i.e. X is exogenous[1]
2. $(X_i, Y_i), \forall 1 \leq i \leq n$ are independent and identically distributed[2]
3. Large outliers are unlikely: X_i and Y_i have nonzero finite fourth moments[3]
$$E[||X||^4] < \infty \qquad E[u^4] < \infty$$
4. With multiple regressors, there is no perfect multicollinearity, so there is no linear dependence between the regressors[4]

When these assumptions are satisfied, we can derive the following theorems.

[1] This is commonly violated if there is some important omitted variable, measurement error, or simultaneous causality, covered in 3.

[2] This is commonly violated in time series data, which has the structure
$$y_t = \beta_0 + \beta_1 y_{t-1} + u_t.$$

[3] The fourth moment of a random variable is a measure of kurtosis, which is how heavy the tails are.

[4] This is commonly violated if you fall into the "dummy variable trap".

2.3 OLS is Unbiased

Theorem 7. *The OLS estimator $\hat{\beta}_1$ is unbiased.*

Proof.

$$\begin{aligned}
y_i - \overline{y} &= (\beta_0 + \beta_1 x_i + u_i) - (\beta_0 + \beta_1 \overline{x} + \overline{u}) \\
&= \beta_1(x_i - \overline{x}) + (u_i - \overline{u}) \\
\hat{\beta}_1 &= \frac{\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^n (x_i - \overline{x})^2} \\
&= \frac{\sum_{i=1}^n (x_i - \overline{x}_1)(\beta_1(x_i - \overline{x}) + (u_i - \overline{u}))}{\sum_{i=1}^n (x_i - \overline{x})^2} \\
&= \frac{\beta_1 \sum_{i=1}^n (x_i - \overline{x})^2 + \sum_{i=1}^n (x_i - \overline{x})(u_i - \overline{u})}{\sum_{i=1}^n (x_i - \overline{x})^2} \\
&= \beta_1 + \frac{\sum_{i=1}^n (x_i - \overline{x})(u_i - \overline{u})}{\sum_{i=1}^n (x_i - \overline{x})^2} \\
&= \beta_1 + \frac{\sum_{i=1}^n (x_i - \overline{x})u_i}{\sum_{i=1}^n (x_i - \overline{x})^2} \\
\Rightarrow E(\hat{\beta}_1|X) &= \beta_1 + E\left[\frac{\sum_{i=1}^n (x_i - \overline{x})u_i}{\sum_{i=1}^n (x_i - \overline{x})^2} \,\middle|\, X\right] \\
&= \beta_1 + \frac{\sum_{i=1}^n (x_i - \overline{x})}{\sum_{i=1}^n (x_i - \overline{x})^2} E(u_i|X) \\
&= \beta_1 + \frac{\sum_{i=1}^n (x_i - \overline{x})}{\sum_{i=1}^n (x_i - \overline{x})^2} E(u_i|X_i) \\
&= \beta_1
\end{aligned}$$

$$\begin{aligned}
&\sum_{i=1}^n (x_i - \overline{x})(u_i - \overline{u}) \\
&= \sum_{i=1}^n (x_i - \overline{x})u_i - \sum_{i=1}^n (x_i - \overline{x})\overline{u} \\
&= \sum_{i=1}^n (x_i - \overline{x})u_i - \left(\sum_{i=1}^n x_i - n\overline{x}\right)\overline{u} \\
&= \sum_{i=1}^n (x_i - \overline{x})u_i - (0)\overline{u} \\
&= \sum_{i=1}^n (x_i - \overline{x})u_i
\end{aligned}$$

Since we have that $E(\hat{\beta}_1|X) = \beta_1$, we have proven that $\hat{\beta}_1$ is unbiased. □

This holds in more generality when $k \geq 2$ as well, but the above derivation is still useful for when we are dealing with simple regression.

Theorem 8. *The OLS estimator $\hat{\beta}$ is unbiased.*

Proof.

$$\begin{aligned}
E(\hat{\beta}) &= E(E(\hat{\beta}|X)) \\
&= E\left[E\left[\left(\sum_{i=1}^{n} X_i X_i^T\right)^{-1} \sum_{i=1}^{n} X_i y_i \;\middle|\; X\right]\right] \\
&= E\left[E\left[\left(\sum_{i=1}^{n} X_i X_i^T\right)^{-1} \left(\sum_{i=1}^{n} X_i X_i^T\right)\beta + \left(\sum_{i=1}^{n} X_i X_i^T\right)^{-1} \sum_{i=1}^{n} X_i u_i \;\middle|\; X\right]\right] \\
&= \beta + E\left[E\left[\left(\sum_{i=1}^{n} X_i X_i^T\right)^{-1} \sum_{i=1}^{n} X_i u_i \;\middle|\; X\right]\right] \\
&= \beta + E\left[\left(\sum_{i=1}^{n} X_i X_i^T\right)^{-1} \sum_{i=1}^{n} X_i E\left[u_i \mid X\right]\right] \\
&= \beta + 0 \\
&= \beta
\end{aligned}$$

□

2.4 OLS is Consistent

Theorem 9. *The OLS estimator $\hat{\beta}$ is consistent.*

Proof. We first note that $S_{XX^T}^{-1} \coloneqq \operatorname{plim}_{n\to\infty}\left(\frac{1}{n}\sum_{i=1}^{n} X_i X_i^T\right)^{-1} = E[XX^T]^{-1}$ is finite and nonrandom, since every element in the matrix is:

$$\left(\frac{1}{n}\sum_{i=1}^{n} X_i X_i^T\right)_{ab} = \frac{1}{n}\sum_{i=1}^{n} x_{ai} x_{bi}$$

and by our assumption, variance must be finite.[5]

[5] If kurtosis is finite, variance must be finite as well.

$$\begin{aligned}
\operatorname*{plim}_{n\to\infty} \hat{\beta} - \beta &= \operatorname*{plim}_{n\to\infty}\left(\sum_{i=1}^{n} X_i X_i^T\right)^{-1} \sum_{i=1}^{n} X_i y_i - \beta \\
&= \operatorname*{plim}_{n\to\infty}\left(\sum_{i=1}^{n} X_i X_i^T\right)^{-1} \sum_{i=1}^{n} X_i u_i \\
&= \operatorname*{plim}_{n\to\infty}\left(\frac{1}{n}\sum_{i=1}^{n} X_i X_i^T\right)^{-1} \frac{1}{n}\sum_{i=1}^{n} X_i u_i \\
&= S_{XX^T}^{-1} \operatorname*{plim}_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} X_i u_i
\end{aligned}$$

If $E(u_i|X_i) = 0$, then $E(X_i u_i) = E(X_i E(u_i|X_i)) = 0$. Since variance is finite, by the Law of Large Numbers,

$$\operatorname*{plim}_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} X_i u_i = 0.$$

Then, since S_{XX^T} is finite, we must have that

$$\operatorname*{plim}_{n\to\infty} \hat{\beta} - \beta = 0$$

□

2.5 OLS is BLUE

Definition 10 (Homoskedasticity)**.** We say that the errors u_i from a regression are homoskedastic if

$$Var(u_i|X_i) = \sigma_0^2$$

where σ_0^2 is some constant.

Definition 11 (Best Linear Unbiased Estimator)**.** We say that an estimator $\hat{\beta}$ is the Best Linear Unbiased Estimator (BLUE) if $\hat{\beta}$ is unbiased, and more efficient than any other linear unbiased estimator $\tilde{\beta}$, such that:

$$Var(\tilde{\beta}) > Var(\hat{\beta})$$

Theorem 12. *$Var(\hat{\beta})$ is $\sum_{i=1}^{n} A_i Var(u_i^2|X)A_i^T$.*

Proof.

$$\begin{aligned}
Var(\hat{\beta}) &= E(Var(\hat{\beta}|X)) + Var(E(\hat{\beta}|X)) \\
&= Var(\hat{\beta} \mid X) \\
&= Var\left(\left(\sum_{i=1}^{n} X_i X_i^T\right)^{-1} \sum_{i=1}^{n} X_i y_i \;\middle|\; X\right) \\
&= Var\left(\beta + \left(\sum_{i=1}^{n} X_i X_i^T\right)^{-1} \sum_{i=1}^{n} X_i u_i \;\middle|\; X\right) \\
&= Var\left(\sum_{i=1}^{n} A_i u_i \;\middle|\; X\right) \\
&= \sum_{i=1}^{n} Var(A_i u_i|X) \\
&= \sum_{i=1}^{n} \left(E\left((A_i u_i)^2 \mid X\right) + \left(E\left(A_i u_i \mid X\right)\right)^2\right) \\
&= \sum_{i=1}^{n} \left(E(A_i u_i u_i^T A_i^T \mid X) + (A_i E(u_i|X))\right)^2 \\
&= \sum_{i=1}^{n} E(A_i u_i^2 A_i^T \mid X) \\
&= \sum_{i=1}^{n} A_i E(u_i^2|X) A_i^T \\
&= \sum_{i=1}^{n} A_i Var(u_i^2|X) A_i^T
\end{aligned}$$

□

Since we will assume homoskedasticity for Gauss-Markov, we want to derive $Var(\hat{\beta})$ for homoskedastic errors as well.

Theorem 13. *With homoskedastic errors such that $Var(u_i^2|X) = \sigma_0^2$, we have that $Var(\hat{\beta})$ is $\sigma_0^2 \left(\sum_{i=1}^n X_i X_i^T\right)^{-1}$.*

Proof.

$$\begin{aligned}
Var(\hat{\beta}) &= \sum_{i=1}^n A_i Var(u_i^2|X) A_i^T \\
&= \sum_{i=1}^n A_i \sigma_0^2 A_i^T \\
&= \sigma_0^2 \left(\sum_{i=1}^n X_i X_i^T\right)^{-1} \sum_{j=1} X_j X_j^T \left(\sum_{i=1}^n X_i X_i^T\right)^{-1} \\
&= \sigma_0^2 \left(\sum_{i=1}^n X_i X_i^T\right)^{-1}
\end{aligned}$$

□

For the special case where $k = 1$, we can get an even more explicit form of $Var(\hat{\beta})$. I will be skipping steps in the derivation that are covered before.

Theorem 14. *With homoskedastic errors such that $Var(u_i^2|X) = \sigma_0^2$, when $k = 1$, $Var(\hat{\beta}_1) = \frac{\sigma_0^2}{\sum_{i=1}^n (x_{1j} - \bar{x}_1)^2}$*

Proof.

$$\begin{aligned}
Var(\hat{\beta}) &= \sigma_0^2 \left(\sum_{i=1}^n X_i X_i^T\right)^{-1} \\
&= \sigma_0^2 \left(\sum_{i=1}^n \begin{pmatrix} 1 \\ x_{1i} \end{pmatrix} \begin{pmatrix} 1 & x_{1i} \end{pmatrix}\right)^{-1} \\
&= \sigma_0^2 \begin{pmatrix} n & \sum_{i=1}^n x_{1i} \\ \sum_{i=1}^n x_{1i} & \sum_{i=1}^n x_{1i}^2 \end{pmatrix}^{-1} \\
&= \sigma_0^2 \frac{1}{n \sum_{i=1}^n x_{1i}^2 - (\sum_{i=1}^n x_{1i})^2} \begin{pmatrix} \sum_{i=1}^n x_{1i}^2 & -\sum_{i=1}^n x_{1i} \\ -\sum_{i=1}^n x_{1i} & n \end{pmatrix}
\end{aligned}$$

Here, $Var(\hat{\beta}_1)$ is the bottom right term in the variance-covariance

matrix, so:

$$\begin{aligned} Var(\hat{\beta}_1) &= \frac{n\sigma_0^2}{n\sum_{i=1}^n x_{1i}^2 - (\sum_{i=1}^n x_{1i})^2} \\ &= \frac{n\sigma_0^2}{n\sum_{i=1}^n x_{1i}^2 - n^2\overline{x}_1} \\ &= \frac{\sigma_0^2}{\sum_{i=1}^n x_{1i}^2 - n\overline{x}_1} \\ &= \frac{\sigma_0^2}{\sum_{i=1}^n (x_{1i} - \overline{x}_1)^2} \end{aligned}$$

□

Theorem 15 (Gauss-Markov Theorem)**.** *If $E(u_i|X_i) = 0$ and the errors of the regression are homoskedastic, then the OLS estimator $\hat{\beta}$ is the Best Linear Unbiased Estimator (BLUE).*

Proof. [6] Let $\tilde{\beta}$ be any other linear unbiased estimator of β. Since $\tilde{\beta}$ is a linear estimator of β, we have that

$$\begin{aligned} \tilde{\beta} &= \sum_{i=1}^n c_i y_i \\ &= \sum_{i=1}^n c_i X_i^T \beta + \sum_{i=1}^n c_i u_i \end{aligned}$$

We define $a_i := c_i - \left(\sum_{i=1}^n X_i X_i^T\right)^{-1} X_i$, so:

$$\begin{aligned} \sum_{i=1}^n a_i y_i &= \sum_{i=1}^n c_i y_i - \left(\sum_{i=1}^n X_i X_i^T\right) \sum_{i=1}^n X_i y_i \\ &= \tilde{\beta} - \hat{\beta} \end{aligned}$$

Since $\tilde{\beta}$ is an unbiased estimator of β, we must also have that $E(\tilde{\beta}|X) = 0$, implying that

$$E\left(\sum_{i=1}^n c_i X_i^T \beta + \sum_{i=1}^n c_i u_i \,\middle|\, X\right) = \beta$$

This is only possible if $\sum_{i=1}^n c_i X_i^T = I$ for all $1 \leq i \leq n$. With this condition,

$$\begin{aligned} \sum_{i=1}^n a_i X_i^T &= \sum_{i=1}^n c_i X_i^T - \left(\sum_{i=1}^n X_i X_i^T\right)^{-1} \sum_{i=1}^n X_i X_i^T \\ &= I - I = 0 \\ \Rightarrow \sum_{i=1}^n a_i y_i &= \sum_{i=1}^n a_i X_i^T \beta + \sum_{i=1}^n a_i u_i \\ &= \sum_{i=1}^n a_i u_i \end{aligned}$$

[6] This is an alternative proof, because I really did not want to use first order conditions.

Notice that:

$$\begin{aligned}
Var(\tilde{\beta}) &= Var(\hat{\beta} + (\tilde{\beta} - \hat{\beta})) \\
&= Var\left(\hat{\beta} + \sum_{i=1}^{n} a_i u_i\right) \\
&= Var(\hat{\beta}) + Var\left(\sum_{i=1}^{n} a_i u_i\right) + Cov\left(\hat{\beta}, \tilde{\beta} - \hat{\beta}\right) \\
Cov\left(\hat{\beta}, \tilde{\beta} - \hat{\beta}\right) &= Cov\left(\hat{\beta} - \beta, \tilde{\beta} - \hat{\beta}\right) \\
&= E\left(\left(\sum_{i=1}^{n} X_i X_i^T\right)^{-1} \sum_{i=1}^{n} X_i u_i \left(\sum_{i=1}^{n} a_i u_i\right)^T\right) \\
&= E\left(\left(\sum_{i=1}^{n} X_i X_i^T\right)^{-1} \sum_{i=1}^{n} X_i u_i u_i^T a_i^T\right) \\
&= E\left(\sigma_0^2 \left(\sum_{i=1}^{n} X_i X_i^T\right)^{-1} \sum_{i=1}^{n} X_i a_i^T\right) \qquad \text{If } \textstyle\sum_{i=1}^{n} a_i X_i^T = 0, \text{ then } \sum_{i=1}^{n} X_i a_i^T = 0. \\
&= 0 \\
\Rightarrow Var(\tilde{\beta}) &= Var(\hat{\beta}) + Var\left(\sum_{i=1}^{n} a_i u_i\right) \\
&\geq Var(\hat{\beta})
\end{aligned}$$

□

With the mean and variance of $\hat{\beta}$ derived from before, we can find the distribution of $\hat{\beta}$ to conduct hypothesis tests.

Theorem 16. $\hat{\beta} \sim N(\beta, \sigma_0^2 \left(\sum_{i=1}^{n} X_i X_i^T\right)^{-1})$

Proof. This follows from the Central Limit Theorem. □

2.6 Test of Linear Restrictions

A vector of r linear restrictions on the regressors can always be written in the form

$$R\beta = r.$$

The way to test if these conditions are fulfilled is to calculate the Wald statistic:

$$W(\hat{\beta}) = (R\hat{\beta} - r)^T (RVar(\hat{\beta})R^T)^{-1}(R\hat{\beta} - r)$$

Then, asymptotically,

$$W(\hat{\beta}) \sim \chi^2(r).$$

2.7 Large Sample Properties of OLS

When we deal with large samples of observations, we can weaken[7] the assumption that $E(u|X) = 0$, and replace it with the assumption that

$$E(Xu) = 0.$$

We can still obtain consistency and asymptotic normality with this weaker assumption, when we have a sufficiently large sample.

[7] This is a weaker assumption because:

$$\begin{aligned} E(Xu) &= E[E(Xu|X)] \\ &= E[XE(u|X)] \\ &= 0 \end{aligned}$$

$$E(u|X) = 0 \Rightarrow E(Xu) = 0$$

Theorem 17 (Large Sample Consistency)**.** *When n large, the OLS estimator $\hat{\beta}$ is consistent with the weakened assumption that $E[Xu] = 0$.*

Proof. From the proof of consistency of $\hat{\beta}$ with $E(u|X) = 0$ in 2.4, we again note that $S_{XX^T}^{-1} := \text{plim}_{n\to\infty}\left(\frac{1}{n}\sum_{i=1}^{n} X_i X_i^T\right)^{-1} = E[XX^T]^{-1}$ is finite and nonrandom. Then,

$$\begin{aligned} \operatorname*{plim}_{n\to\infty}(\hat{\beta} - \beta) &= \operatorname*{plim}_{n\to\infty}\left(\frac{1}{n}\sum_{i=1}^{n} X_j X_j^T\right)^{-1} \operatorname*{plim}_{n\to\infty}\left(\frac{1}{n}\sum_{i=1}^{n} X_i u_i\right) \\ &= S_{XX^T}^{-1} \operatorname*{plim}_{n\to\infty}\left(\frac{1}{n}\sum_{i=1}^{n} X_i u_i\right) \end{aligned}$$

By the Weak Law of Large Numbers,

$$\operatorname*{plim}_{n\to\infty}\left(\frac{1}{n}\sum_{i=1}^{n} X_i u_i\right) = E[Xu] = 0.$$

Thus, we have that

$$\operatorname*{plim}_{n\to\infty}(\hat{\beta} - \beta) = 0$$

□

Theorem 18 (Large Sample Asymptotic Distribution)**.** *When n large, with $E[Xu] = 0$, we have that $\frac{1}{\sqrt{n}}(\hat{\beta} - \beta) \sim N\left(0, S_{XX^T}^{-1} Var(Xu) S_{XX^T}^{-1}\right)$*

Proof. By the Central Limit Theorem[8],

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i u_i \xrightarrow{d} N(0, Var(Xu)).$$

Then, with the definition of $S_{XX^T}^{-1}$ above and by the Law of Large Numbers[9],

$$\begin{aligned} \sqrt{n}(\hat{\beta} - \beta) &= \left(\frac{1}{n}\sum_{i=1}^{n} X_i X_i^T\right)^{-1} \left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i u_i\right) \\ &\xrightarrow{d} S_{XX^T}^{-1} N(0, Var(Xu)) \\ &\sim N(0, S_{XX^T}^{-1} Var(Xu) S_{XX^T}^{-1}) \end{aligned}$$

□

[8] With the CLT:

$$\begin{aligned} &\frac{\frac{1}{n}\sum_{i=1}^{n} X_i u_i - E[Xu]}{\sqrt{Var(Xu)/n}} \sim N(0,1) \\ &\Rightarrow \frac{1}{\sqrt{n}} \frac{\sum_{i=1}^{n} X_i u_i}{\sqrt{Var(Xu)}} \sim N(0,1) \\ &\Rightarrow \frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i u_i \sim N(0, Var(Xu)) \end{aligned}$$

[9] Convergence in probability implies convergence in distribution.

3 Validity of OLS

3.1 Omitted variable bias

Suppose that the true form of the dependent variable is:

$$Y = \beta_0 + \beta_1 X + \beta_2 Z + u$$

but you have conducted a multiple regression of the form

$$Y = \gamma_0 + \gamma_1 X + v.$$

What is the effect of underspecifying your regression?

Theorem 19. *When a regression is underspecified, the OLS estimator we obtain is no longer consistent.*

Proof. Based on the true functional form, we actually have that $v = \beta_2 Z + u$. Thus,

$$\begin{aligned}
\hat{\gamma}_1 &= \frac{\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^n (x_i - \overline{x})^2} \\
&= \frac{\sum_{i=1}^n (x_i - \overline{x}) y_i}{\sum_{i=1}^n (x_i - \overline{x})^2} \\
&= \frac{\sum_{i=1}^n (x_i - \overline{x})(\beta_0 + \beta_1 x_i + \beta_2 z_i + u_i)}{\sum_{i=1}^n (x_i - \overline{x})^2} \\
&= \frac{\beta_1 \sum_{i=1}^n (x_i - \overline{x}) x_i + \beta_2 \sum_{i=1}^n (x_i - \overline{x}) z_i + \sum_{i=1}^n (x_i - \overline{x}) u_i}{\sum_{i=1}^n (x_i - \overline{x})^2} \\
&= \beta_1 + \beta_2 \frac{\sum_{i=1}^n (x_i - \overline{x}) z_i}{\sum_{i=1}^n (x_i - \overline{x})^2} + \frac{\sum_{i=1}^n (x_i - \overline{x}) u_i}{\sum_{i=1}^n (x_i - \overline{x})^2} \\
E(\hat{\gamma}_1) &= E(E(\hat{\gamma}_1 \mid X)) \\
&= E\left(E\left(\beta_1 + \beta_2 \cdot \frac{\sum_{i=1}^n (x_i - \overline{x}) z_i}{\sum_{i=1}^n (x_i - \overline{x})^2} + \frac{\sum_{i=1}^n (x_i - \overline{x}) u_i}{\sum_{i=1}^n (x_i - \overline{x})^2} \,\middle|\, X \right) \right) \\
&= \beta_1 + \beta_2 \cdot \frac{Cov(X, Z)}{Var(X)}
\end{aligned}$$

Thus, by the Weak Law of Large Numbers, we have that:

$$(\hat{\gamma}_1 - \beta_1) \xrightarrow{P} \beta_2 \cdot \frac{Cov(X, Z)}{Var(X)}.$$

This implies that unless $Cov(X, Z)$ happens to be 0, our estimator $\hat{\gamma}_1$ is no longer consistent. □

Notice that

$$\begin{cases} \hat{\gamma}_1 - \beta_1 > 0 & \text{if } \operatorname{sign} Cov(X, Z) = \operatorname{sign} \beta_2 \\ \hat{\gamma}_1 - \beta_1 < 0 & \text{if } \operatorname{sign} Cov(X, Z) = \operatorname{sign} \beta_2 \end{cases}$$

which allows us to figure out if our bias is upwards or downwards.

3.2 *Measurement Error*

There are two models of measurement error: the classical measurement error and the "best guess" measurement error.

For the classical measurement error, we assume that the data we obtain $\tilde{X}$ is the combination of the regressors X and some random noise term v:

$$\tilde{X}_i = X_i + v_i$$

where $Cov(X_i, v_i) = 0$ and $Cov(u_i, v_i) = 0$. Let $\sigma_v^2 = Var(v_i)$ and $\sigma_X^2 = Var(X_i)$. Then, although the truth is

$$Y_i = \beta_0 + \beta_1 X_1 + u_i$$

our actual regression is

$$\begin{aligned} Y_i &= \beta_0 + \beta_1(X_i - v_i) + u_i \\ \Rightarrow Y_i &= \beta_0 + \beta_1 X_i + (u_i - \beta_1 v_i) \\ Y &= \beta_0 + \beta_1 \tilde{X}_i + \tilde{u}_i \end{aligned}$$

where $\tilde{u}_i = u_i - \beta_1 v_i$. Thus, the least squares assumption 1 is violated, where:

$$E(\tilde{u}_i|\tilde{X}_i) = E(\tilde{u}_i|v_i) = E(u_i + \beta_1 v_i|v_i) = -\beta_1 v_i \neq 0.$$

This results in biased estimators, since from 2.3 and by the Law of Large Numbers,

$$\begin{aligned} \operatorname*{plim}_{n\to\infty} \hat{\beta}_1 &= \beta_1 + \operatorname*{plim}_{n\to\infty} \frac{\sum_{i=1}^n (\tilde{x}_i - \tilde{\bar{x}})\tilde{u}_i}{\sum_{i=1}^n (\tilde{x}_i - \tilde{\bar{x}})^2} \\ &= \beta_1 + \frac{Cov(\tilde{X}_i, \tilde{u}_i)}{Var(\tilde{X})} \\ &= \beta_1 + \frac{Cov(X_i + v_i, u_i) - \beta_1 Cov(X_i + v_i, -v_i)}{Var(\tilde{X})} \\ &= \beta_1 \left(1 - \frac{\sigma_v^2}{\sigma_X^2 + \sigma_v^2}\right) \\ &= \beta_1 \frac{\sigma_X^2}{\sigma_X^2 + \sigma_v^2} \end{aligned}$$

This gives us that $\hat{\beta}_1$ is biased towards 0.

The other model of measurement error is the "best guess" model, where X_i is not known, but we use another variable W_i where $E(u_i \mid W_i) = 0$ to "guess" X_i with:

$$\tilde{X} = E(X_i \mid W_i).$$

Then, our regression is of the form:

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 \tilde{X}_i + \beta_1(X_i - \tilde{X}_i) + u_i \\ Y_i &= \beta_0 + \beta_1 \tilde{X}_i + \tilde{u}_i. \end{aligned}$$

Here, there will not be biased estimators, since:

$$\begin{aligned}\operatorname*{plim}_{n\to\infty}\hat{\beta}_1 &= \beta_1 + \operatorname*{plim}_{n\to\infty}\frac{\sum_{i=1}^{n}(\tilde{x}_i-\tilde{\bar{x}})\tilde{u}_i}{\sum_{i=1}^{n}(\tilde{x}_i-\tilde{\bar{x}})^2}\\ &= \beta_1 + \frac{Cov(\tilde{X}_i,\tilde{u}_i)}{Var(\tilde{X})}\\ &= \beta_1 + \frac{Cov(\tilde{X}_i,u_i)+\beta_1 Cov(\tilde{X}_i,X_i-\tilde{X}_i)}{Var(\tilde{X})}\end{aligned}$$

- $Cov(\tilde{X}_i, u_i) = 0$, since $\tilde{X}_i$ is a function of W, and $E(u_i|W_i) = 0$.
- $Cov(\tilde{X}_i, X_i - \tilde{X}_i) = 0$, since $\tilde{X}_i = E(X_i|W_i)$

So,

$$\operatorname*{plim}_{n\to\infty}\hat{\beta}_1 = \beta_1.$$

3.3 Missing Data

When collecting data, it is both impossible and inefficient to collect data for the entire population — after all, the purpose of statistics is to draw inference from the large population with a smaller sample. Thus, there will be some missing data from the population.

Depending on the cause of the missing data, there will be different effects on the validity of the regression. There are 3 types of missing data:

1. Missing Completely At Random (MCAR)

   Data is missing completely at random when they are missing for random reasons unrelated to values of X or Y.

   For example, when you randomly draw a sample from the entire population to take a survey on overwork, this is an example of data missing completely at random.

2. Missing At Random (MAR)

   Data is missing at random when they are missing based on X.

   For the survey on overwork, even if you send the survey to the random sample of the population, it is possible that people's response rates are dependent on their profession. Here, we can include that as one of our regressors, and the data is missing at random.

3. Missing Not At Random

   Data is missing not at random when they are missing for reasons related to Y outside of X.

   In a similar train of thought from above, it is possible that overworked people are less likely to have the time to fill in surveys, so here the data will be missing not at random.

3.4 Simultaneous Causality

Another possible problem that arises affecting the validity of the OLS procedure is that not only does X cause Y, but Y causes X. Here, there are two data-generating processes:

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

$$X_i = \gamma_0 + \gamma_1 Y_i + v_i$$

This leads to correlation between the regressor and the error term, since:

$$\begin{aligned} Cov(X_i, u_i) &= Cov(\gamma_0 + \gamma_1 Y_i + v_i, u_i) \\ &= \gamma_1 Cov(Y_i, u_i) + Cov(v_i, u_i) \\ &= \gamma_1 Cov(\beta_0 + \beta_1 X_i + u_i, u_i) \\ &= \gamma_1 \beta_1 Cov(X_i, u_i) + \gamma_1 \sigma_u^2 \\ \Rightarrow Cov(X_i, u_i) &= \gamma_1 \frac{\sigma_u^2}{1 - \gamma_1 \beta_1} \end{aligned}$$

This covariance term is not 0 unless $\gamma_1 = 0$.

4 Panel data

4.1 Setup for Panel Data

- Panel data consists of observations for n different entities across T different time periods, where $T \geq 2$.

$$\begin{pmatrix} (X_{11}, Y_{11}) & (X_{12}, Y_{12}) & \dots & (X_{1T}, Y_{1T}) \\ (X_{21}, Y_{21}) & (X_{22}, Y_{22}) & \dots & (X_{2T}, Y_{2T}) \\ \vdots & \vdots & \ddots & \vdots \\ (X_{n1}, Y_{n1}) & (X_{n2}, Y_{n2}) & \dots & (X_{nT}, Y_{nT}) \end{pmatrix}$$

- The panel is balanced if all variables are observed for each entity and time period, and a panel is unbalanced if not.

Often we want to determine the effects of a policy X, but we only have panel data. Directly performing OLS regression with this data will almost certainly result in omitted variable bias, since different entities and time periods will have different characteristics, which might affect our dependent variable Y.

Then, our challenge is to extricate the effects of a change in X from the effects of other changes across time or across different entities.

4.2 *Fixed Effects Regression*

If the effects on Y across different entities is constant across time, we can use the Fixed Effects Regression.

A special case of the fixed effects regression is the "before and after" comparison, when there are only 2 time periods.

This method makes a lot of intuitive sense — if you want to see the effect on Y of a change in policy X across different entities, you should regress the change in X against their change in Y, to account for baseline differences in Y. Mathematically, suppose each entity had a different fixed effect Z_i, so we have data of the form:

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \beta_2 Z_i + u_{it}$$

where $t \in \{1, 2\}$. We saw in 3.1, if we were unable to observe Z_i, it would lead to the the omitted variable bias. However, with the method of first differences:

$$\begin{aligned} Y_{i2} - Y_{i1} &= (\beta_0 - \beta_0) + \beta_1(X_{i2} - X_{i1}) + \beta_2(Z_i - Z_i) + (u_{i2} - u_{i1}) \\ Y_{i2} - Y_{i1} &= \beta_1(X_{i2} - X_{i1}) + (u_{i2} - u_{i1}) \\ \Delta Y_i &= \beta_1 \Delta X_i + \Delta u_i. \end{aligned}$$

Then, regressing ΔY_i against ΔX_i will sidestep this problem. This method also extends to if there are k regressors.

But what if there are more than 2 time periods? Our data is still of the form

$$Y_{it} = \beta_0 + \beta_1 X_{1,it} + \cdots + \beta_k X_{k,it} + \beta_{k+1} Z_i + u_{it}$$

but $1 \leq t \leq T$. We notice above that the intercept value $\beta_0 + \beta_{k+1} Z_i$ was the value that disappeared in comparisons across time, so we define $\alpha_i = \beta_0 + \beta_{k+1} Z_i$ as the entity fixed effect. Then our data is of the form:

$$Y_{it} = \beta_1 X_{1,it} + \cdots + \beta_k X_{k,it} + \alpha_i + u_{it}.$$

From here, we have two options. We could rewrite this equation into binary variables for each entity, so:

$$Y_{it} = \beta_0 + \beta_1 X_{1,it} + \cdots + \beta_k X_{k,it} + \gamma_2 \mathbb{1}\{i = 2\} + \cdots + \gamma_n \mathbb{1}\{i = n\} + u_{it}$$

and conduct an OLS using the techniques we have learnt before. We drop $\gamma_1 \mathbb{1}\{i = 1\}$ to avoid multicollinearity, so this should be a valid OLS. However, this OLS has $k + n$ regressors, so this can get quite cumbersome.

The better option is to use the "entity-demeaned" OLS algorithm. We first calculate the following variables:

$$\overline{Y}_i = \frac{1}{T} \sum_{t=1}^{T} Y_{it} \qquad \overline{X}_{j,i} = \frac{1}{T} \sum_{t=1}^{T} X_{j,it} \qquad \overline{u}_i = \frac{1}{T} \sum_{t=1}^{T} u_{it}$$

for each $1 \leq j \leq k$ and notice that $\overline{\alpha}_i = \frac{1}{T}\sum_{t=1}^{T} \alpha_i = \alpha_i$. Then, we take the entity-demeaned variables and defining $\tilde{Y}_{it} = Y_{it} - \overline{Y}_i$, $\tilde{X}_{j,it} = X_{j,it} - \overline{X}_{j,i}$, $\tilde{u}_{it} = u_{it} - \overline{u}_i$:

$$Y_{it} - \overline{Y}_i = \beta_1(X_{1,it} - \overline{X}_{1,i}) + \cdots + \beta_k(X_{k,it} - \overline{X}_{k,i}) + (\alpha_i - \overline{\alpha}_i) + (u_{it} - \overline{u}_i)$$
$$Y_{it} - \overline{Y}_i = \beta_1(X_{1,it} - \overline{X}_{1,i}) + \cdots + \beta_k(X_{k,it} - \overline{X}_{k,i}) + (u_{it} - \overline{u}_i)$$
$$\tilde{Y}_{it} = \beta_1\tilde{X}_{1,it} + \cdots + \beta_k\tilde{X}_{k,it} + \tilde{u}_{it}$$

We can estimate β_i with OLS, and this is identical to the OLS estimator we get from using the binary variables method above.

4.3 *Time Effects Only Regression*

If the effects on Y across different times is constant across different entities, we can use the time effects only regression. This is directly analogous to the fixed effects only regression we analysed above, where each time period has a different time effect S_t, and our data is of the form:

$$Y_{it} = \beta_0 + \beta_1 X_{1,it} + \cdots + \beta_k X_{k,it} + \beta_{k+1} S_t + u_{it}.$$

Then, defining $\lambda_t = \beta_0 + \beta_{k+1} S_t$ as the time fixed effect, we get the equation:

$$Y_{it} = \beta_1 X_{1,it} + \cdots + \beta_k X_{k,it} + \lambda_t + u_{it}.$$

From here, we could solve this with the binary variables OLS

$$Y_{it} = \beta_1 X_{1,it} + \cdots + \beta_k X_{k,it} + \delta_2 \mathbb{1}\{t = 2\} + \cdots + \delta_T \mathbb{1}\{t = T\} + u_{it}.$$

4.4 *Difference in Differences*

When there are both fixed and time effects, then we use the method of difference in differences (DiD) to handle both kinds of variation.

It is crucial to notice here that the fixed effects must be constant across time and the time effects must be constant across entities for this to be valid, since we will not account for any interaction term.

We will analyse the algebra for $k = 1$, but the $k \geq 2$ case extends easily. The combined regression model will be:

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \alpha_i + \gamma_t + u_{it}$$

where $\alpha_i = \beta_2 Z_i$ and $\gamma_t = \beta_3 S_t$.

We take the difference between the data and its within-entity mean over time, as well as the difference between the across-time mean over entities and the total mean:

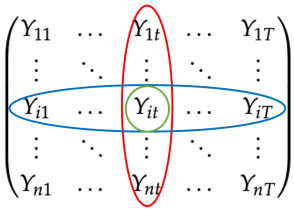

Figure 1: The observation circled in green is the data, the mean over the observations circled in blue is the within-entity mean over time, the mean over the observations circled in red is the across-time mean over entities and the total mean is the mean of all observations of Y.

$$\begin{aligned} Y_{it} - \overline{Y}_{i\cdot} &= (\beta_0 + \beta_1 X_{1,it} + \alpha_i + \gamma_t + u_{it}) - (\beta_0 + \beta_1 \overline{X}_{1,i\cdot} + \alpha_i + \overline{\gamma} + \overline{u}_{i\cdot}) \\ &= \beta_1(X_{it} - \overline{X}_{i\cdot}) + (\gamma_t - \overline{\gamma}) + (u_{it} - \overline{u}_i) \\ \overline{Y}_{\cdot t} - \overline{Y} &= (\beta_0 + \beta_1 \overline{X}_{1,\cdot t} + \overline{\alpha} + \gamma_t + \overline{u}_{\cdot t}) - (\beta_0 + \beta_1 \overline{X} + \overline{\alpha} + \overline{\gamma} + \overline{u}) \\ &= \beta_1(\overline{X}_{1,\cdot t} - \overline{X}) + (\gamma_t - \overline{\gamma}) + (\overline{u}_{\cdot t} - \overline{u}) \end{aligned}$$

Then, taking the difference of differences,

$$\begin{aligned} Y_{it} - \overline{Y}_{i\cdot} - \overline{Y}_{\cdot t} - \overline{Y} &= \beta_1(X_{it} - \overline{X}_{i\cdot}) + (u_{it} - \overline{u}_i) - \beta_1(\overline{X}_{1,\cdot t} - \overline{X}) + (\overline{u}_{\cdot t} - \overline{u}) \\ &= \beta_1(X_{it} - \overline{X}_{i\cdot} - \overline{X}_{1,\cdot t} + \overline{X}) + (u_{it} - \overline{u}_i - \overline{u}_{\cdot t} + \overline{u}) \end{aligned}$$

This allows us to estimate β_1 without knowing anything about α_i or γ_t. This procedure is much more sensible and easy to parse with only $n = 2$, $T = 2$, and an example.

Example 20. Minimum Wage and Unemployment[10,11]

[10] Card and Krueger [1994]

[11] 1/3rds of the 2021 Nobel Memorial Prize in Economic Sciences.

- **Goal:** You want to find out if increasing the minimum wage in New Jersey has an effect on unemployment rates.
- **Problem:** If we naively compare the unemployment rates in New Jersey before and after the change in minimum wage, we are conducting a fixed effects regression.

  We will fail to take into account time effects of unemployment, such as macroeconomic conditions, resulting in omitted variable bias.
- **Solution:** We collect unemployment data from a neighbouring state Pennsylvania, which did not increase their minimum wage, and perform a difference in difference procedure.

  Here we are assuming that the time effects are the same across entities, which is why we specifically focus on unemployment rates in counties of Pennsylvania and New Jersey that are on their shared border. It is unlikely that time effects will vary significantly in these neighbouring counties.

Suppose Pennsylvania is entity a and New Jersey is entity b, and time period 1 is before the rise in minimum wage, and time period 2 is after the rise in minimum wage. Then,

$$y_{a1,i} = \alpha_a + \lambda_1 + u_{a1,i} \qquad y_{a2,i} = \alpha_a + \lambda_2 + u_{a2,i}$$

$$y_{b1,i} = \alpha_b + \lambda_1 + u_{b1,i} \qquad y_{b2,i} = \beta_1 + \alpha_b + \lambda_2 + u_{b2,i}.$$

Then, we carry out the difference in differences procedure:

$$\begin{aligned} \overline{y}_{a2} - \overline{y}_{a1} &= \lambda_2 - \lambda_1 + (\overline{u}_{a2} - \overline{u}_{a1}) \\ \overline{y}_{b2} - \overline{y}_{b1} &= \beta_1 + \lambda_2 - \lambda_1 + (\overline{u}_{b2} - \overline{u}_{b1}) \\ (\overline{y}_{b2} - \overline{y}_{b1}) - (\overline{y}_{a2} - \overline{y}_{a1}) &= \beta_1 + (\overline{u}_{a2} - \overline{u}_{a1} - \overline{u}_{b2} - \overline{u}_{b1}). \end{aligned}$$

The interpretation of the difference in difference procedure is that we calculate the time effect by calculating the change in unemployment in Pennsylvania where the minimum wage did not rise. Then, if we subtract this time effect from the change in unemployment in New Jersey, we should get the change in unemployment directly caused by the rise in minimum wage, β_1.

In Card and Krueger's landmark paper, they found that increased minimum wage in New Jersey did not cause an increase in unemployment, against most traditional economic theory.

4.5 Fixed Effects Regression Assumptions

We have seen that we can write our fixed effect regression in the form:

$$Y_{it} = \beta_1 X_{it} + \alpha_i + u_{it}$$

In order for us to derive a causal effect from our fixed effect regression, we need the following 4 assumptions:

1. $E(u_{it}|X_{i1}, \ldots, X_{iT}, \alpha_i) = 0$
2. $(X_{i1}, \ldots, X_{iT}, u_{i1}, \ldots, u_{iT})$ are i.i.d. draws from their joint distribution, for all $1 \leq i \leq n$
3. Large outliers are unlikely, so X_{it} and u_{it} have finite fourth moments
4. There is no perfect multicollinearity

4.6 Cluster Standard Error

With the methods above, we can find estimates $\hat{\beta}_1$ with modified OLS procedures. But in order to verify if these estimators are statistically significant, we need to calculate the standard error of these estimates. We will cover the case of clustered standard errors for a fixed effects regression with $k = 1$.

Recall from 4.2 that our fixed effects regression is of the form:

$$\tilde{Y}_{it} = \tilde{X}_{it} + \tilde{u}_{it}.$$

Then, from 2.4,

$$\hat{\beta}_1 - \beta = \left(\sum_{i=1}^{n} \sum_{t=1}^{T} x_{it} x_{it}^{\top} \right)^{-1} \sum_{i=1}^{n} \sum_{t=1}^{T} x_{it} u_{it}$$

$$\sqrt{nT}(\hat{\beta} - \beta) = \left(\frac{1}{\sqrt{nT}} \sum_{i=1}^{n} \sum_{t=1}^{T} x_{it} x_{it}^{\top} \right)^{-1} \frac{1}{\sqrt{nT}} \sum_{i=1}^{n} \sum_{t=1}^{T} x_{it} u_{it}$$

Here, i denote the matrix inverse by $A^{\top}$, to avoid confusion with T, the number of time periods.

By the Weak Law of Large Numbers,

$$\left(\frac{1}{\sqrt{nT}}\sum_{i=1}^{n}\sum_{t=1}^{T}x_{it}x_{it}^{\top}\right)^{-1} \xrightarrow{P} S_{XX^{\top}}^{-1}$$

where $S_{XX^{\top}}^{-1} = E[XX^{\top}]^{-1}$ finite and non-random.

We assume that disturbances are uncorrelated across clusters, but possibly correlated within clusters, so we can aggregate across clusters.

$$\begin{aligned}\frac{1}{\sqrt{nT}}\sum_{i=1}^{n}\sum_{t=1}^{T}x_{it}u_{it} &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[\frac{1}{\sqrt{T}}\sum_{t=1}^{T}x_{it}u_{it}\right]\\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}Z_i\\ \Rightarrow \sqrt{nT}(\hat{\beta}-\beta) &\xrightarrow{P} S_{XX^{\top}}^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}Z_i\\ &\xrightarrow{d} N(0, S_{XX^{\top}}^{-1}Var(Z_i)S_{XX^{\top}}^{-1})\end{aligned}$$

This follows by the Central Limit Theorem, and since

$$E\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T}x_{it}u_{it}\right) = \frac{1}{\sqrt{T}}\sum_{t=1}^{T}E(x_{it}u_{it}) = 0.$$

Then, we want to calculate $\widehat{Var(Z_i)}$:

$$\begin{aligned}Var(Z_i) &= \frac{1}{T}\sum_{t=1}^{T}\sum_{k=1}^{T}(x_{it}u_{it})(x_{ik}u_{ik})\\ &= \frac{1}{T}\left[\sum_{t=1}^{T}Var(x_{it}u_{it}) + \sum_{t=1}^{T}\sum_{s\neq t}Cov(x_{it}u_{it}, x_{is}u_{is})\right]\\ \widehat{Var(Z_i)} &= \frac{1}{n-1}\sum_{i=1}^{n}(\hat{Z}_i - \overline{\hat{Z}_i})(\hat{Z}_i - \overline{\hat{Z}_i})\\ &= \frac{1}{n-1}\sum_{i=1}^{n}\hat{Z}_i^2\end{aligned}$$

where $\hat{Z}_i = \frac{1}{\sqrt{T}}\sum_{i=1}^{T}x_{it}\hat{u}_{it}$, and the final equality arises from the residuals and regressors being uncorrelated.

To see this:

$$\begin{aligned}\overline{\hat{Z}_i} &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{1}{\sqrt{T}}\sum_{t=1}^{T}x_{it}\hat{u}_{it}\\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{1}{\sqrt{T}}\sum_{t=1}^{T}x_{it}(y_{it}-\overline{y}) - \hat{\beta}_1(x_{it}-\overline{x}))\\ &= \frac{1}{\sqrt{nT}}\sum_{i=1}^{n}\sum_{t=1}^{T}(y_{it}-\overline{y})(x_{it}-\overline{x})\\ &\quad - \hat{\beta}_1\frac{1}{\sqrt{nT}}\sum_{i=1}^{n}\sum_{t=1}^{T}(x_{it}-\overline{x})^2\\ &= 0\end{aligned}$$

This gives us that:

$$\sqrt{nT}(\hat{\beta}-\beta) \xrightarrow{d} N\left(0, S_{XX^{\top}}^{-1}\left(\frac{1}{n-1}\sum_{i=1}^{n}\hat{Z}_i^2\right)S_{XX^{\top}}^{-1}\right)$$

5 *Binary Dependent Variables*

5.1 *Interpretation of Dependent Variable*

A special case that we want to learn how to deal with is when our dependent variables are binary. Suppose that we conducted a multiple regression:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki} + u_i$$

where $Y \in \{0,1\}$, often taking the form of an indicator function. Most of the time, $\hat{Y}_i \notin \{0,1\}$, so we need an interpretation for this. After all, what does it mean when $\mathbb{1}\{\text{enlisting in the military}\} = 0.8$?

- We know that

$$E(Y_i \mid X_{1i}, \ldots, X_{ki}) = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki}.$$

With our assumption that $E(u_i|X_i) = 0$.

- We also know that:

$$\begin{aligned} E(Y_i \mid X_{1i}, \ldots, X_{ki}) &= 0 \cdot P(Y_i = 0 \mid X_{1i}, \ldots, X_{ki}) + 1 \cdot P(Y_i = 1 \mid X_{1i}, \ldots, X_{ki}) \\ &= P(Y_i = 1 \mid X_{1i}, \ldots, X_{ki}) \end{aligned}$$

So we can derive a sensible interpretation that $\hat{Y}_i = P(Y_i = 1 \mid X_{1i}, \ldots, X_{ki})$.

5.2 *Linear Probability Model*

If we directly use the interpretation that

$$\hat{Y} = P(Y_i = 1 \mid X_{1i}, \ldots, X_{ki})$$

in the multiple regression model without any modifications, we get the linear probability model

$$P(Y = 1 \mid X_1, \ldots, X_k) = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + u_i.$$

Here, we can interpret the slope coefficients β_i as the change in the probability that $Y = 1$ when there is a unit change in X, and $\hat{Y}_i$ is the predicted probability that $Y = 1$. These coefficients can be estimated using all of the procedures detailed above.

This seems like a sensible procedure and we appear to be done. But this approach has two problems:

1. The linear probability model implies that each unit change in X_j has a linear effect on $P(Y = 1 \mid X_1, \ldots, X_k)$ through its coefficient β_j. But this is often not a realistic assumption for most real-world applications.[12]
2. Probability is only well-defined within $[0,1]$. If the OLS estimation was conducted on data with no outliers, and we tried to predict $\hat{p} \coloneqq P(Y = 1 \mid X_1, \ldots X_k)$ with extreme data points, we could get $\hat{p} < 0$ or $\hat{p} > 1$, which is meaningless.

For this two reasons, we almost never use a linear probability model. We thus turn to nonlinear methods to solve these problems.

[12] For example, probability of having a heart attack is not very much different between ages 20 and 30, but very different between ages 50 and 60. The linear probability model cannot account for this nuance.

5.3 Probit and Logit Regression

The probit and logit regression are designed around cumulative distribution functions, since they have a nonlinear shape that forces predicted values to be between 0 and 1, resolving the issues above.

The probit regression uses the standard normal cumulative distribution function Φ. The regression conducted is such that:

$$P(Y = 1 \mid X_1, \ldots, X_k) = \Phi(\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k).$$

- To get a predicted value $\hat{p}$, we compute $z = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k$, and find the probability corresponding to this z-score in the normal distribution table.
- The interpretation of slope coefficients β_i is the change in z-value when there is a unit change in X_i, holding $\{X_j \mid i \neq j\}$ constant.
- To estimate the effect of ΔX on $\hat{p} = P(Y = 1 \mid X_1, \ldots, X_k)$, we compute $\hat{p}_1$ with the original values of X, and compute $\hat{p}_2$ with the values $X + \Delta X$. Then, $\hat{p}_2 - \hat{p}_1$ is the effect of ΔX on $\hat{p}$.

The logit regression uses the logistic regression function F. The regression conducted is such that

$$P(Y = 1 \mid X_1, \ldots, X_k) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k)}}$$

Both regressions have the exact same principles, interpretations and their estimates are very similar. The logistic c.d.f. is just easier to numerically compute than the standard normal c.d.f, since

$$P(Y = 1 \mid X_1, \ldots, X_k) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k)}}$$

is easier to solve than

$$P(Y = 1 \mid X_1, \ldots, X_k) = \int_{-\infty}^{\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx.$$

5.4 Estimation and Inference in Logit and Probit Models

Now that we can interpret probit and logit models, we want to figure out how to estimate the coefficients β_i in these models. Since these coefficients appear in a nonlinear function, they cannot be estimated directly by OLS.

One way is nonlinear least squares, where if F is the given c.d.f.,

$$\hat{\beta} = \arg\min_b \sum_{i=1}^{n} [Y_i - F(b_0 + b_1 X_{1i} + \cdots + b_k X_{ki})]^2.$$

This formulation is familiar to us, and even has two nice properties of consistency and asymptotically normal distribution. However, this estimator has high variance, so it is inefficient.

We turn to the method of maximum likelihood to determine the values of β_i.

Definition 21 (Likelihood function)**.** The likelihood function is the joint probability distribution of the data, as a function of the unknowns.

Definition 22 (Maximum Likelihood Estimator (MLE))**.** The maximum likelihood estimator of parameters β are the values $\tilde{\beta}$ which maximise the likelihood function.

Intuitively, we are given the dataset $(Y_i, X_{1i}, \ldots, X_{ki})$, and we are trying to find the values of β_i that produce this data with the highest probability.

In a probit model, the likelihood function is:

$$\begin{aligned}
&P(Y_1 = y_1, \ldots, Y_n = y_n \mid X_{1i}, \ldots, X_{X_k i}, 1 \le i \le n) \\
&= P(Y_1 = y_1 \mid X_{11}, \ldots, X_{k1}) \times \cdots \times P(Y_n = y_n \mid X_{1n}, \ldots, X_{kn}) \\
&= p_1^{y_1}(1-p_1)^{1-y_1} \times \ldots p_n^{y_n}(1-p_n)^{1-y_n}
\end{aligned}$$

Since Y_i are all independently distributed, the joint probability distribution is the product of their distributions.

Since Y_i are all binary,

$$P(Y_i = y_i \mid X_{1i}, \ldots, X_{ki}) = p_i^{y_1}(1-p_i)^{y_i}$$

where $p_i = \Phi(\beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki})$

In order to maximise the likelihood function, you can equivalently maximise the log of the likelihood function, since log is a monotonically increasing function. This gives us the following:

$$\begin{aligned}
&\ln(P(Y_1 = y_1, \ldots, Y_n = y_n \mid X_{1i}, \ldots, X_{X_k i}, 1 \le i \le n)) \\
&= \ln(p_1^{y_1}(1-p_1)^{1-y_1} \times \ldots p_n^{y_n}(1-p_n)^{1-y_n}) \\
&= \sum_{i=1}^{n} Y_i \ln(p_i) + \sum_{i=1}^{n}(1-Y_i)\ln(1-p_i) \\
&= \sum_{i=1}^{n} Y_i \ln(\Phi(\beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki})) \\
&\quad + \sum_{i=1}^{n}(1-Y_i)\ln(1-\Phi(\beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki}))
\end{aligned}$$

The log-likelihood function for the logit model is almost exactly the same, except:

$$p_i = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki})}}$$

In order to maximise both log-likelihood functions, we need to run a numerical procedure on the computer.

Under general conditions, MLEs are consistent and are asymptotically normal. More than that, the MLE is a very efficient estimator.

We also have a new measure of goodness of fit, the pseudo-R^2, which has the value:

$$\text{pseudo-}R^2 = 1 - \frac{\ln\left(f^{\max}_{probit}\right)}{\ln\left(f^{\max}_{Bernoulli}\right)}$$

where $f^{\max}_{probit}$ is the value of the maximised probit likelihood, and $f^{\max}_{Bernoulli}$ is the value of the maximised likelihood without any of the X's.

6 Instrumental Variables

6.1 Setup of Simple Instrumental Variables Regression

We know from 2.3 and 2.4 that $\hat{\beta}$ obtained from OLS is unbiased and consistent under the assumption that

$$E(u_i|X_i) = 0.$$

We have also seen in 3 that if this assumption is removed, we run into the problem that $\hat{\beta}$ can be biased and inconsistent, due to the possibility of misspecification, measurement error, missing data, and simultaneous causality.

In these cases where

$$E(u|X) \neq 0,$$

the use of an instrumental variable Z allows us to sidestep these problems. The key is finding a variable Z such that:

1. $Cor(Z_i, X_i) \neq 0$, preferably large.
2. $Cor(Z_i, u_i) = 0$.

From here, we conduct the Two-Stage Least Squares procedure, where we conduct the following two regressions:

$$X_i = \pi_0 + \pi_1 Z_i + v_i$$
$$Y_i = \beta_0 + \beta_1^{TSLS}\hat{X}_i + u_i$$

where $\hat{X}_i \overset{def}{=} \hat{\pi}_0 + \hat{\pi}_1 Z_i$ and $\hat{\pi}_0$, $\hat{\pi}_1$ are obtained from the first regression. Here, under some assumptions, the estimator $\hat{\beta}_1^{TSLS}$ is consistent, which is a significant improvement over $\hat{\beta}_1^{OLS}$. We can also derive the exact form of $\hat{\beta}_1^{TSLS}$.

Theorem 23. $\hat{\beta}_1^{TSLS} = \frac{s_{ZY}}{s_{ZX}}$

Proof. In the second stage regression, we know that:

$$\hat{\beta}_1^{TSLS} = \frac{s_{\hat{X}Y}}{s^2_{\hat{X}}}$$

by the properties of the OLS, and we have that:

$$s_{\hat{X}Y} = \hat{\pi}_1 s_{ZY} \qquad s^2_{\hat{X}} = \hat{\pi}_1^2 s^2_Z.$$

From the first stage regression, we know that:

$$\hat{\pi}_1 = \frac{s_{ZX}}{s^2_Z}.$$

Thus,

$$\hat{\beta}_1^{TSLS} = \frac{\hat{\pi}_1 s_{ZY}}{\hat{\pi}_1^2 s^2_Z} = \frac{s_{ZY}}{\hat{\pi}_1 s^2_Z} = \frac{s_{ZY}}{s_{ZX}}$$

□

Since $\hat{X} = \hat{\pi}_0 + \hat{\pi}_1 Z$,

$$Y = \beta_0 + \beta_1(\hat{\pi}_0 + \hat{\pi}_1 Z) + u_i$$

.

We will first use an example to illustrate the intuition behind the use of Z in estimation, and then derive some nice statistical properties of $\hat{\beta}_1^{TSLS}$ with some algebra.

Example 24. Veteran status and Mortality rates[13,14]

- **Goal:** You want to find out if serving in the military during the Vietnam War has an effect on mortality rates.
- **Current methods:** With OLS, we would regress

  $$X = \mathbb{1}\{\text{serving in the military}\} \text{ against } Y = \text{mortality rate}$$

- **Problem:** There are two kinds of people serving the military, volunteers and those drafted. Volunteers to serve the military tend to have lower income, and thus higher mortality rates.

  This is just one of the possibly many unobserved factors that affect both serving the military and mortality rates simultaneously, resulting in $E(u|X) \neq 0$.

[13] Angrist et al. [1996].

[14] The remaining 2/3rds of the 2021 Nobel Memorial Prize in Economic Sciences.

Using the principles above, we want to find a variable Z that is correlated with serving the military, but not correlated with mortality rates, in order to use instrumental variables estimation. Clearly, a good choice of Z is

$$Z = \mathbb{1}\{\text{was drafted into the military}\},$$

since being drafted correlates with serving the military, but draft numbers are randomly assigned, and thus have no relation with mortality rates.

We first conduct the following regression:

$$X_i = \pi_0 + \pi_1 Z_i + v_i.$$

By regressing X against Z, we can obtain an estimate $\hat{X}_i = \hat{\pi}_0 + \hat{\pi}_1 Z_i$ of all military service that is "explained" by draft number. Intuitively,

you can think of this population as the people who served the military if they were drafted, and didn't if they were not drafted. The unobserved factors like income that might make someone self-select into serving the military are no longer a confounding factor here. So, we must have that

$$E(u|\hat{X}) = 0.$$

Knowing this, we can now conduct the second stage regression:

$$Y_i = \beta_0 + \beta_1^{TSLS}\hat{X}_i + u_i.$$

Then, we can obtain the estimate $\hat{\beta}_1^{TSLS}$ to inform us whether military service directly explained by the draft has any impact on mortality rates, which is exactly what we wanted.

In Angrist, Imbens and Rubin's paper, they found that serving in the military during the Vietnam War resulted in a $\sim 25\%$ increase in mortality rates.

This sounds like a reasonable procedure, but does it give us back the desirable property of consistency?

6.2 *Statistical Properties of $\hat{\beta}_1^{TSLS}$*

Theorem 25. *$\hat{\beta}_1^{TSLS}$ is consistent.*

Proof.

$$\begin{aligned}
\hat{\beta}_1^{TSLS} &= \frac{\frac{1}{n-1}\sum_{i=1}^n (Z_i - \overline{Z})(Y_i - \overline{Y})}{\frac{1}{n-1}(Z_i - \overline{Z})(X_i - \overline{X})} \\
&= \frac{\sum_{i=1}^n (Z_i - \overline{Z})(\beta_1(X_i - \overline{X}) + (u_i - \overline{u}))}{\sum_{i=1}^n (Z_i - \overline{Z})(X_i - \overline{X})} \\
&= \frac{\beta_1 \sum_{i=1}^n (Z_i - \overline{Z})(X_i - \overline{X}) + \sum_{i=1}^n (Z_i - \overline{Z})(u_i - \overline{u})}{\sum_{i=1}^n (Z_i - \overline{Z})(X_i - \overline{X})} \\
&= \beta_1 + \frac{\frac{1}{n}\sum_{i=1}^n (Z_i - \overline{Z})(u_i - \overline{u})}{\frac{1}{n}\sum_{i=1}^n (Z_i - \overline{Z})(X_i - \overline{X})} \\
&= \beta_1 + \frac{\frac{1}{n}\sum_{i=1}^n (Z_i - \overline{Z})u_i}{s_{ZX}}
\end{aligned}$$

We know that in large samples, $\overline{Z} \simeq \mu_Z$. So, by the Law of Large Numbers,

$$\hat{\beta}_1^{TSLS} - \beta_1 \xrightarrow{P} \frac{Cov(Z_i, u_i)}{Cov(Z_i, X_i)} = 0$$

since $Cov(Z_i, u_i) = 0$, by our choice of instrument. □

Theorem 26. $\hat{\beta}_1^{TSLS} \sim N\left(\beta_1, \frac{1}{n}\frac{Var((Z_i - u_Z)u_i)}{(Cov(Z_i, X_i))^2}\right)$.

Proof. From above, we know that

$$\hat{\beta}_1^{TSLS} \cong \beta_1 + \frac{\frac{1}{n}\sum_{i=1}^{n}(Z_i - \mu_Z)u_i}{Cov(Z_i, X_i)}$$

Thus, we have that $\beta_1^{TSLS} \sim N\left(\beta_1, \frac{1}{n}\frac{Var((Z_i - u_Z)u_i)}{(Cov(Z_i,X_i))^2}\right)$. □

6.3 Setup of General IV Regression

- We may not always have 1 regressor, 1 instrument and no exogenous regressors.
- A more general setup would be have Y the dependent variable, $X_1, \ldots, X_k$ endogenous regressors, $W_1, \ldots, W_r$ included exogenous regressors or control variables uncorrelated with u_i, and $Z_1, \ldots Z_m$ the instrumental variables.
- We must have that $m \geq k$ for the IV procedure to be sensible. When $m > k$, we say that the coefficients are overidentified. When $m = k$, we say that the coefficients are exactly identified. When $m < k$, we say that the coefficients are underidentified.

Then, the TSLS regression is of the form:

$$\begin{aligned}
X_{1i} &= \pi_{1,0} + \pi_{1,1}Z_{1i} + \cdots + \pi_{1,m}Z_{mi} + \pi_{1,m+1}W_{1i} + \cdots + \pi_{1,m+r}W_{ri} + v_i \\
&\vdots \\
X_{ki} &= \pi_{k,0} + \pi_{k,1}Z_{1i} + \cdots + \pi_{k,m}Z_{mi} + \pi_{k,m+1}W_{1i} + \cdots + \pi_{k,m+r}W_{ri} + v_i \\
Y_i &= \beta_0 + \beta_1\hat{X}_{1i} + \cdots + \beta_k\hat{X}_{ki} + \beta_{k+1}W_{1i} + \beta_{k+r}W_{ri} + u_i
\end{aligned}$$

where

$$\hat{X}_{ji} = \hat{\pi}_{j,0} + \hat{\pi}_{j,1}Z_{1i} + \cdots + \hat{\pi}_{j,m}Z_{mi} + \pi_{j,m+1}W_{1i} + \cdots + \hat{\pi}_{j,m+r}W_{ri}.$$

In order for us to use the general IV regression to obtain reasonable estimates of β_i, a few assumptions must be fulfilled.

6.4 Assumptions of IV Regression

With the general IV regression, the choice of the instruments $Z_1, \ldots Z_m$ is more complicated. We still want the instruments to be relevant and exogenous, but with multiple variables, these conditions now take the form:

1. Instrument relevance
   - In general, $(\iota, \hat{X}_{1i}, \ldots, \hat{X}_{ki}, W_{1i}, W_{ri})$ cannot be perfectly multicollinear, which makes our second stage regression sensible.[15]

[15] ι is the regressor with value 1 for all observations.

- If there is only one X, at least one $\hat{\pi}_i \neq 0$ to ensure no perfect multicollinearity.

2. Instrument Exogeneity

   The instruments must be uncorrelated with the error term, so $Cor(Z_{ji}, u_i) = 0$ for all $1 \leq j \leq m$.

Other than the instruments being valid for the IV regression, we still need a few more assumptions for proper causal inference. The following assumptions are direct analogues from the assumptions we make for causal inference with OLS, as in 2.2.

3. $E(u_i|W_{1i}, \ldots, W_{ri}) = 0$
4. $(X_{1i}, \ldots, X_{ki}, W_{1i}, \ldots, W_{ri}, Z_{1i}, \ldots, Z_{mi}, Y_i)$ are i.i.d. draws from their joint distribution
5. Large outliers are unlikely, so X's, W's, Z's and Y's have nonzero finite fourth moments

With all of these assumptions, $\hat{\beta}_1^{TSLS}$ is consistent and normally distributed in large samples. We have already proven this for a special case above. The proof for large samples is not needed in your exam.[16]

[16] I think?

6.5 Instrument Validity

In order for an instrument to be valid, we want it to be both relevant and exogenous. When we come up with instruments, we want a way to test if they are valid.

1. Instrument Relevance

   Our only requirement for relevance is that $Cov(Z_i, X_i) \neq 0$. Notice that the smaller $Cov(Z_i, X_i)$ is, the larger the variance of $\hat{\beta}_1^{TSLS}$, since it is in the denominator of the variance. Thus, $\hat{\beta}_1^{TSLS}$ loses explanatory power.

   We call instruments with small $Cov(Z_i, X_i)$ weak instruments. This begs the question of how large the covariance must be to obtain a good approximation.

   When there is only one endogenous regressor X, when conducting the first regression

   $$X_i = \pi_0 + \pi_1 Z_{1i} + \cdots + \pi_m Z_{mi}$$

   we can conduct an F test on the hypothesis that $\pi_1 = \cdots = \pi_m = 0$. If the F-statistic is less than 10, this implies that the instruments are too weak for inference.

When the instruments are too weak, if the coefficients are overidentified, you can try and remove weak instruments. Otherwise, you have to find more, stronger instruments.

2. Instrument Exogeneity

   If the instruments are not exogenous, $\hat{\beta}^{TSLS}$ is inconsistent, since we relied on this assumption to prove consistency in 6.2.

   It is not possible to test if the instruments are exogenous if the coefficients are exactly identified, because there is no way to exclude any of the instruments to make a statistical comparison.

   If the coefficients are overidentified, we can test for overidentifying restrictions. We conduct the regression:

$$\hat{u}_i^{TSLS} = \delta_0 + \delta_1 Z_{1i} + \cdots + \delta_m Z_{mi} + \delta_{m+1} W_{1i} + \cdots + \delta_{m+r} W_{ri} + e_i$$

   where $\hat{u}_i^{TSLS}$ are the residuals from the second stage TSLS regression. We compute the F-statistic testing the hypothesis $\delta_1 = \cdots = \delta_m = 0$, and $J = mF$. If all instruments are exogenous, then $J \sim \chi^2_{m-k}$. This test is clearly meaningless if $m = k$.[17]

[17] And that's a wrap! Good luck on your final!

References

Joshua D. Angrist, Guido W. Imbens, and Donald B. Rubin. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91(434):444–455, 1996. ISSN 01621459. URL http://www.jstor.org/stable/2291629.

David Card and Alan Krueger. Minimum wages and employment: A case study of the fast-food industry in new jersey and pennsylvania. *American Economic Review*, 84(4):772–93, 1994. URL https://EconPapers.repec.org/RePEc:aea:aecrev:v:84:y:1994:i:4:p:772-93.