Multiple Linear Regression
Part B: Ordinary Least Squares

Thomas Wiemann
University of Chicago

Econometrics
Econ 21020

Updated: May 18, 2022

Summary

In Part A, we introduced BLP($Y|X$) as approximation to $E[Y|X]$.

- ▷ BLP-coefficients are well-defined when $E\left[XX^\top\right]^{-1}$ exists;
- ▷ Used the Frisch-Waugh Theorem for subvector analysis;
- ▷ Discussed interpretation using a generalized Yitzhaki's Theorem;

The BLP and its coefficients β are theoretical concepts.

In Part B, we bridge the gap between BLP and real data using statistics.

- ▷ Develop the *ordinary least squares* estimator;
- ▷ Analyze its statistical properties under an iid sample;
- ▷ Use matrix calculus for implementation.

Outline

Outline

1. **Ordinary Least Squares**
2. Estimator Properties
   - ▷ Bias
   - ▷ Consistency
   - ▷ Asymptotic Distribution
3. Implementation

Ordinary Least Squares

Let Y be a random variable and $X = (1, X_1, \dots, X_k)^\top$ be a random vector. Consider a random sample $(Y^1, X^1), \dots, (Y^n, X^n) \overset{iid}{\sim} (Y, X)$.

From Lecture 8A, we know that the BLP-coefficients are given by

$$\beta = E[XX^\top]^{-1}E[XY], \tag{1}$$

whenever $E[XX^\top]^{-1}$ exists.

This suggests the sample analogue estimator

$$\hat{\beta}_n = \left(\frac{1}{n}\sum X^i X^{i\top}\right)^{-1}\left(\frac{1}{n}\sum X^i Y^i\right) \tag{2}$$

Notation: *Superscripts – i.e., $X^1, \dots X^n$ – are used as sample indices throughout.*

Ordinary Least Squares (Contd.)

The estimator $\hat{\beta}_n$ is known as *ordinary least squares* (OLS). This is because it can also be motivated as solutions to the least-squares sample criterion:

$$\hat{\beta}_n = \underset{\beta \in \mathbb{R}^{k+1}}{\arg\min} \ \frac{1}{n} \sum_{i=1}^{n} \left(Y^i - X^{i\top} \beta \right)^2 , \tag{3}$$

whenever $E[\sum_{i=1}^{n} X^i X^{i\top}]^{-1}$ exists. In particular, we have:

$$R_n(\beta) = \frac{1}{n} \sum (Y^i - X^{i\top}\beta)^2 = \frac{1}{n} \sum (Y^{i2} - 2Y^i X^{i\top}\beta + \beta^\top X^i X^{i\top} \beta)$$

$$= \frac{1}{n} \sum Y^{i2} - 2\frac{1}{n} \sum Y^i X^{i\top}\beta + \beta^\top \left(\frac{1}{n} \sum X^i X^{i\top}\right)\beta$$

FOC: $$\frac{\partial R_n(\beta)}{\partial \beta} = -2 \frac{1}{n} \sum Y^i X^{i\top} + 2\beta^\top \left(\frac{1}{n} \sum X^i X^{i\top}\right) = 0^\top$$

$$\Rightarrow \quad \left(\frac{1}{n} \sum X_i X_i^\top\right) \beta = \frac{1}{n} \sum X^i Y^i$$

$$\Rightarrow \quad \beta = \left(\frac{1}{n} \sum X^i X^{i\top}\right)^{-1} \left(\frac{1}{n} \sum X^i Y^i\right)$$

Ordinary Least Squares (Contd.)

For our analysis, it's useful to rewrite $\hat{\beta}_n$ using $\varepsilon^i \equiv Y^i - \text{BLP}(Y^i|X^i)$:

$$\hat{\beta}_n = \left(\cancel{\tfrac{1}{n}} \sum X^i X^i\right)^{-1} \left(\cancel{\tfrac{1}{n}} \sum X^i Y^i\right) \qquad \Leftrightarrow Y^i = X^{iT}\beta + \varepsilon^i$$

$$= \left(\sum X^i X^i\right)^{-1} \left(\sum X^i Y^i\right) \tag{4}$$

$$= \left(\sum X^i X^i\right)^{-1} \left(\sum X^i \left[X^{iT}\beta + \varepsilon^i\right]\right)$$

$$= \underbrace{\left(\sum X^i X^i\right)^{-1} \left(\sum X^i X^{iT}\right)}_{I_{k+1}} \beta + \left(\sum X^i X^i\right)^{-1} \left(\sum X^i \varepsilon^i\right)$$

$$= \beta + \left(\sum X^i X^i\right)^{-1} \left(\sum X^i \varepsilon^i\right)$$

Outline

1. Ordinary Least Squares
2. **Estimator Properties**
   - ▷ **Bias**
   - ▷ Consistency
   - ▷ Asymptotic Distribution
3. Implementation

Bias

Our analysis of the OLS estimator begins with its bias.

We assume here that X is continuous to ensure existence of $E[\sum_{i=1}^{n} X^i X^{i\top}]^{-1}$ (for $n > k+1$) when $E[XX^\top]^{-1}$ exists.

The bias of $\hat{\beta}_n$ when X is continuous and $E[XX^\top]^{-1}$ exists is given by

$$\begin{aligned}
\mathrm{Bias}(\hat{\beta}_n) = E[\hat{\beta}_n] - \beta &= E\left[\cancel{\beta} + \left(\sum X^i X^{i\top}\right)^{-1}\left(\sum X^i \varepsilon^i\right)\right] - \cancel{\beta} \\
&= E\left[\left(\sum X^i X^{i}\right)^{-1}\left(\sum X^i \varepsilon^i\right)\right] \\
&= E\left[E\left[\left(\sum X^i X^{i}\right)^{-1}\left(\sum X^i \varepsilon^i\right) \mid (X^i)_{i=1}^{n}\right]\right] \\
&= E\left[\left(\sum X^i X^{i\top}\right)^{-1}\left(\sum X^i E[\varepsilon^i \mid (X^i)_{i=1}^{n}]\right)\right] \\
&\overset{iid}{=} E\left[\left(\sum X^i X^{i\top}\right)^{-1}\left(\sum X^i \underbrace{E[\varepsilon^i \mid X^i]}_{\neq 0 \text{ in general!}}\right)\right] \neq 0 \text{ in general!}
\end{aligned} \tag{5}$$

Bias

Hence, if $E[\varepsilon^i|X^i] = 0$, then $\text{Bias}(\hat{\beta}_n) = 0$.

- ▷ Does $E[\varepsilon^i|X^i] = 0$ hold generally? No: $E[\varepsilon^i X^i] = 0 \not\Rightarrow E[\varepsilon^i|X^i] = 0$.
- ▷ When do we know that $E[\varepsilon^i|X^i] = 0$? Special case: Linear $E[Y|X]$.

Many textbooks state that the OLS estimator $\hat{\beta}_n$ is unbiased for β.

- ▷ Importantly: Strong assumption are made along the way!
- ▷ We *only* showed $\text{Bias}(\hat{\beta}_n) = 0$ if $E[Y|X]$ linear *and* X is continuous.

Generally, little reason to believe $\text{Bias}(\hat{\beta}_n) = 0$ in economic applications:

- ▷ Economic theory rarely implies linear $E[Y|X]$ with continuous X.
- ▷ Horrible news? No: Most estimators are biased in practice...

Outline

1. Ordinary Least Squares
2. **Estimator Properties**
   - ▷ Bias
   - ▷ **Consistency**
   - ▷ Asymptotic Distribution
3. Implementation

Consistency

Theorem 1 ensures OLS satisfies the minimum requirement: Consistency.

Theorem 1

Let Y be a random variable and $X = (1, X_1, \ldots, X_k)^\top$ be a random vector such that $E[XX^\top]^{-1}$ exists, and let β denote the BLP$(Y|X)$-coefficient. If $\hat{\beta}_n$ are the OLS estimators constructed using $(Y^1, X^1), \ldots, (Y^n, X^n) \overset{iid}{\sim} (Y, X)$, then

$$\hat{\beta}_n \xrightarrow{p} \beta. \tag{6}$$

Since the OLS estimators are continuous functions of moments of (Y, X), we can prove this straightforwardly using the WLLN and CMT.

Proof. $\beta_n = \left(\frac{1}{n}\sum X^i X^{iT}\right)^{-1}\left(\frac{1}{n}\sum X^i Y^i\right)$

1. $A_n \equiv \frac{1}{n}\sum X^i X^i$, $B_n \equiv \frac{1}{n}\sum X^i Y^i$

2. $g(a,b) = a^{-1}b$

3. By WLLN, $A_n \xrightarrow{p} E[XX^T]$

   By WLLN, $B_n \xrightarrow{p} E[XY]$

4. By CMT,

$$g(A_n, B_n) \xrightarrow{p} E[XX^T]^{-1}E[XY] = \beta$$

whenever $E[XX^T]^{-1}$ exists.

$\square$

Outline

Theorem 2 shows that OLS is asymptotically normal.

Theorem 2

Let Y be a random variable and $X = (1, X_1, \ldots, X_k)^\top$ be a random vector such that $E[XX^\top]^{-1}$ exists, and let β denote the $BLP(Y|X)$-coefficient. If $\hat{\beta}_n$ are the OLS estimators constructed using $(Y^1, X^1), \ldots, (Y^n, X^n) \overset{iid}{\sim} (Y, X)$, then

$$\sqrt{n}\left(\hat{\beta}_n - \beta\right) \xrightarrow{d} N\left(0,\ \Sigma\right), \tag{7}$$

where

$$\Sigma = E\left[XX^\top\right]^{-1} E\left[XX^\top \varepsilon^2\right] E\left[XX^\top\right]^{-1}, \tag{8}$$

with $\varepsilon \equiv Y - BLP(Y|X)$.

Asymptotic Distribution (Contd.)

Proof.

$$\sqrt{n}(\hat{\beta}_n - \beta) = \sqrt{n}\left(\frac{1}{n}\sum X^i X^{iT}\right)^{-1}\left(\frac{1}{n}\sum X^i \varepsilon^i\right)$$
$$= \left(\frac{1}{n}\sum X^i X^{iT}\right)^{-1}\sqrt{n}\left(\frac{1}{n}\sum X^i \varepsilon^i\right)$$

By WLLN + CMT, $\left(\frac{1}{n}\sum X^i X^{iT}\right)^{-1} \xrightarrow{p} E[XX^T]^{-1}$ whenever $E[XX^T]^{-1}$ exists.

By CLT, $\sqrt{n}\left(\frac{1}{n}\sum X^i\varepsilon^i - \underbrace{E[X\varepsilon]}_{=0}\right) \xrightarrow{d} N(0, \underbrace{Var(X\varepsilon)}_{})$

$$= E[\varepsilon X X^T \varepsilon] - \underbrace{E[X\varepsilon]}_{=0} E[\varepsilon X^T]$$
$$= E[XX^T\varepsilon]$$

By Slutsky's Theorem,

$$\sqrt{n}(\hat{\beta}_n - \beta) \xrightarrow{d} E[XX^T]^{-1} N(0, E[XX^T\varepsilon^2])$$
$$\overset{d}{=} N(0, E[XX^T]E[XX^T\varepsilon^2]E[XX^1])$$

OLS Covariance Estimation

Theorem 2 is of no practical use unless we can replace the expression for the asymptotic variance by a consistent estimator. Fortunately, we can.

Theorem 3

Let Y be a random variable and $X = (1, X_1, \ldots, X_k)^\top$ be a random vector such that $E[XX^\top]^{-1}$ exists, and let β denote the BLP$(Y|X)$-coefficient. If $\hat{\beta}_n$ is the OLS estimator constructed using $(Y^1, X^1), \ldots, (Y^n, X^n) \overset{iid}{\sim} (Y, X)$, then

$$\sqrt{n}\widehat{\Sigma}_n^{-\frac{1}{2}}\left(\hat{\beta}_n - \beta\right) \xrightarrow{d} N\left(0, \mathbf{I}_{k+1}\right), \tag{9}$$

where

$$\widehat{\Sigma}_n = \left(\frac{1}{n}\sum_{i=1}^n X^i X^{i\top}\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^n X^i X^{i\top}\hat{\varepsilon}^{i2}\right)\left(\frac{1}{n}\sum_{i=1}^n X^i X^{i\top}\right)^{-1} \tag{10}$$

and $\hat{\varepsilon}^i = Y^i - X^{i\top}\hat{\beta}_n$.

OLS Covariance Estimation (Contd.)

Proof. Need to show: $\hat{\Sigma}_n \xrightarrow{p} \Sigma$. The rest: Just Slutsky!

$$\hat{\Sigma}_n = \left(\frac{1}{n}\sum X^i X^{iT}\right)^{-1}\left(\frac{1}{n}\sum X^i X^{iT} \hat{\varepsilon}^{i2}\right)\underbrace{\left(\frac{1}{n}\sum X^i X^{iT}\right)^{-1}}_{\xrightarrow{p} E[XX^T]^{-1} \text{ as shown before}}$$

WTS: $\frac{1}{n}\sum X^i X^{iT} \hat{\varepsilon}^{i2} \xrightarrow{p} E[XX^T\varepsilon^2]$

$$\frac{1}{n}\sum X^i X^{iT}\hat{\varepsilon}^{i2} = \frac{1}{n}\sum X^i X^{iT}\left(\hat{\varepsilon}^i - \varepsilon^i + \varepsilon^i\right)^2 = \underbrace{\frac{1}{n}\sum X^i X^{iT}\left(\hat{\varepsilon}^i - \varepsilon^i\right)^2}_{\equiv A_n} + 2\underbrace{\frac{1}{n}\sum X^i X^{iT}\left(\hat{\varepsilon}^i - \varepsilon^i\right)\varepsilon^i}_{\equiv B_n} + \underbrace{\frac{1}{n}\sum X^i X^{iT}\varepsilon^i}_{\xrightarrow{p} E[XX^T\varepsilon^2] \text{ by WLLN}}$$

Note $\hat{\varepsilon}^i - \varepsilon^i = X^{iT}(\beta - \hat{\beta}_n)$

$$B_n = \frac{1}{n}\sum X^i X^{iT}\varepsilon^i X^{iT}(\beta - \hat{\beta}_n)$$

$$= \frac{1}{n}\sum X^i X^{iT}\varepsilon^i\left[(\beta_0 - \hat{\beta}_{0n}) + X_1^i(\beta_1 - \hat{\beta}_{1n}) + \dots + X_k^i(\beta_k - \hat{\beta}_{kn})\right]$$

$$= (\beta_0 - \hat{\beta}_{0n})\left(\frac{1}{n}\sum X^i X^i \varepsilon^i\right) + (\beta_1 - \hat{\beta}_{1n})\left(\frac{1}{n}\sum X^i X^{iT}\varepsilon^i X_1^i\right) + \dots + \underbrace{(\beta_k - \hat{\beta}_{kn})}_{\xrightarrow{p} 0 \text{ by Theorem!}}\left(\frac{1}{n}\sum X^i X^{iT}\varepsilon^i X_k^i\right)$$

$\xrightarrow{p} 0$ by WLLN + CMT.

Similar for $A_n \xrightarrow{p} 0$.

□

OLS Covariance Estimation (Contd.)

Theorem 2 and 3 give inference for the *vector* $\hat{\beta}_n$.

- ▷ Often interested only in a subvector;
- ▷ E.g., the estimator $\hat{\beta}_{jn}$ of β_j.

Corollary 1 and 2 give inference for individual components of $\hat{\beta}_n$.

- ▷ Corollary 1 combines Theorem 2 + Slutsky's Theorem;
- ▷ Corollary 3 gives the standard error formula.

Subvector Asymptotic Distribution

Corollary 1

Let Y be a random variable and $X = (1, X_1, \ldots, X_k)^\top$ be a random vector such that $E[XX^\top]^{-1}$ exists, and let $\beta = (\beta_0, \beta_1, \ldots, \beta_k)$ denote the BLP$(Y|X)$-coefficient. If $\hat{\beta}_n = (\hat{\beta}_{0n}, \hat{\beta}_{1n}, \ldots, \hat{\beta}_{kn})$ is the OLS estimator constructed using $(Y^1, X^1), \ldots, (Y^n, X^n) \overset{iid}{\sim} (Y, X)$, then

$$\sqrt{n}\left(\hat{\beta}_{jn} - \beta_j\right) \xrightarrow{d} N\left(0,\ e_j^\top \Sigma e_j\right), \quad \forall j = 0, 1, \ldots, k, \tag{11}$$

where Σ is defined by Equation (8) and e_j is the jth unit vector.

Proof. $\sqrt{n}(\hat{\beta}_{jn} - \beta_j) = \sqrt{n}(e_j^\top \hat{\beta}_n - e_j^\top \beta) = e_j^\top \sqrt{n}(\hat{\beta}_n - \beta)$

$\xrightarrow{d} e_j^\top N(0, \Sigma) \overset{d}{=} N(0, e_j^\top \Sigma e_j)$

by Theorem 2 + Slutsky's.

$\square$

Note: *$e_j^\top \Sigma e_j$ simply selects the jth diagonal entry of Σ*

Standard Error

Corollary 2

Let Y be a random variable and $X = (1, X_1, \dots, X_k)^\top$ be a random vector such that $E[XX^\top]^{-1}$ exists, and let $\beta = (\beta_0, \beta_1, \dots, \beta_k)$ denote the BLP$(Y|X)$-coefficient. If $\hat{\beta}_n = (\hat{\beta}_{0n}, \hat{\beta}_{1n}, \dots, \hat{\beta}_{kn})$ is the OLS estimator constructed using $(Y^1, X^1), \dots, (Y^n, X^n) \overset{iid}{\sim} (Y, X)$, then

$$\frac{\hat{\beta}_{jn} - \beta_j}{se\left(\hat{\beta}_{jn}\right)} \xrightarrow{d} N(0,1), \quad \forall j = 0, 1, \dots, k, \tag{12}$$

where

$$se\left(\hat{\beta}_{jn}\right) = \frac{1}{\sqrt{n}} \sqrt{e_j^\top \widehat{\Sigma}_n e_j} \tag{13}$$

with $\widehat{\Sigma}_n$ is defined by Equation (3) and e_j is the jth unit vector.

Note: *$e_j^\top \widehat{\Sigma}_n e_j$ simply selects the jth diagonal entry of $\widehat{\Sigma}_n$*

Proof.

1. $A_n = \hat{\Sigma}_n$

2. $g(a) = \dfrac{1}{\sqrt{e_j^T a e_j}}$

3. By Theorem 3, $A_n \xrightarrow{p} \Sigma$

4. By CMT,

$$g(A_n) \xrightarrow{p} \frac{1}{\sqrt{e_j^T \Sigma e_j}} \quad \text{whenever } e_j^T \Sigma e_j > 0.$$

Then, combining w/ Corollary 1, we have by Slutsky's

$$\frac{1}{\sqrt{e_j^T \hat{\Sigma}_n e_j}} \sqrt{n}(\hat{\beta}_{jn} - \beta_j) \xrightarrow{d} \frac{1}{\sqrt{e_j^T \Sigma e_j}} N(0, e_j^T \Sigma e_j) \overset{d}{=} N\left(0, \frac{e_j^T \Sigma e_j}{e_j^T \Sigma e_j}\right)$$

$$\overset{d}{=} N(0, 1)$$

□

Outline

1. Ordinary Least Squares
2. **Estimator Properties**
   - Bias
   - Consistency
   - Asymptotic Distribution
3. **Implementation**

OLS Implementation

Implementing OLS by brute force (*e.g.*, $\sum_{i=1}^n X^i X^{i\top}$) is difficult.

- ▷ Instead: Use matrix operations for straightforward computation.

Define the stacked sample matrices $\mathbb{X}_n$ and $\mathbb{Y}_n$:

$$\mathbb{X}_n \equiv \begin{bmatrix} X^{1\top} \\ X^{2\top} \\ \vdots \\ X^{n\top} \end{bmatrix}, \qquad \mathbb{Y}_n \equiv \begin{bmatrix} Y^1 \\ Y^2 \\ \vdots \\ Y^n \end{bmatrix}. \tag{14}$$

Then, matrix calculus shows that we have

$$\mathbb{X}_n^\top \mathbb{X}_n = \sum_{i=1}^n X^i X^{i\top}, \qquad \mathbb{X}_n^\top \mathbb{Y}_n = \sum_{i=1}^n X^i Y^i. \tag{15}$$

The OLS estimator can then equivalently be stated as

$$\hat{\beta}_n = \left(\mathbb{X}_n^\top \mathbb{X}_n\right)^{-1} \left(\mathbb{X}_n^\top \mathbb{Y}_n\right). \tag{16}$$

OLS Implementation (Contd.)

For the OLS covariance estimator $\widehat{\Sigma}_n$, we define stacked residual vector:

$$\epsilon_n \equiv \mathbb{Y}_n - \mathbb{X}_n\hat{\beta}_n = \begin{bmatrix} Y^1 \\ Y^2 \\ \vdots \\ Y^n \end{bmatrix} - \begin{bmatrix} X^{1\top}\hat{\beta}_n \\ X^{2\top}\hat{\beta}_n \\ \vdots \\ X^{n\top}\hat{\beta}_n \end{bmatrix} = \begin{bmatrix} \hat{\varepsilon}_1 \\ \hat{\varepsilon}_2 \\ \vdots \\ \hat{\varepsilon}_n \end{bmatrix}. \tag{17}$$

By the same matrix calculus as before, we have

$$\left(\mathbb{X}_n \odot \epsilon_n\right)^\top \left(\mathbb{X}_n \odot \epsilon_n\right) = \sum_{i=1}^{n} X^i X^{i\top} \hat{\varepsilon}^{i2}, \tag{18}$$

where $\odot$ denotes element-wise multiplication (*Hadamard product*). Then

$$\widehat{\Sigma}_n = \frac{1}{n}\left(\mathbb{X}_n^\top \mathbb{X}_n\right)^{-1}\left[\left(\mathbb{X}_n \odot \epsilon_n\right)^\top \left(\mathbb{X}_n \odot \epsilon_n\right)\right]\left(\mathbb{X}_n^\top \mathbb{X}_n\right)^{-1}. \tag{19}$$

***Notation**: Strictly speaking, $\odot$ is defined only for matrices of equal dimension. We abuse the notation here to denote multiplication between each row of the matrix $\mathbb{X}_n$ with the corresponding component of the vector ϵ_n.*

OLS Estimation in R

OLS Estimation in R

```
# Compute OLS estimates
XX_inv <- solve(t(X) %*% X)
XY <- t(X) %*% Y
beta <- XX_inv %*% XY

# Compute BLP estimates
blp_yx <- X %*% beta

# Compute standard error for beta
epsilon <- c(Y - blp_yx)
XX_eps2 <- t(X * epsilon) %*% (X * epsilon)
Sigma <- XX_inv %*% XX_eps2 %*% XX_inv / n
se <- sqrt(diag(Sigma)) / sqrt(n)
```

Note: *There exists an OLS implementation in* R *– the* `lm`*-command. But importantly: Base-R does not implement the standard error of Corollary 2! So have some faith in your abilities and implement OLS yourself. See Problem 7 of Problem Set 4.*

Summary

Today, we introduced OLS as an estimator for the BLP($Y|X$).

- ▷ Showed that it is consistent and asymptotically normal;
- ▷ Derived standard errors for subvector inference.

We're now well-equipped for causal analysis under selection on observables & common support:

- ▷ Defined interesting causal parameters using the all causes model;
- ▷ Showed identification of the CATE, ATT, ATU, and ATE;
- ▷ Concluded that if (W, X) is discrete, may use the binning estimator;
- ▷ If (W, X) is continuous/mixed, we can leverage OLS to obtain approximate results.