Lectures 9 & 10

Ordinary Least Squares Estimation

Oscar Volpe

10/25/2021 & 10/27/2021

1. Constructing the OLS Estimator
   - Derivation
   - Measures of Fit
2. Properties of the OLS Estimator
   - Gauss-Markov Theorem
   - Unbiasedness
   - Consistency
   - Asymptotic Normality
3. *Example:* Control Variables & Hypothesis Testing

Solving for the BLP

Consider an i.i.d. sample $\{X_i, Y_i\}_{i=1}^n$, where $Y_i \in \mathbb{R}$ and $X \in \mathbb{R}^{k+1}$. To estimate β, we solve a sample analogue of the least-squares problem:

$$\hat{\beta}_n \in \underset{b}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} (Y_i - X_i' b)^2 \tag{1}$$

Solving this minimization problem gives us an estimator for β_1:

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

This estimator is called the *ordinary least squares* (OLS) estimator.

- We require that $\frac{1}{n} \sum_{i=1}^{n} X_i X_i'$ is invertible, which means there can be no perfect collinearity *within the sample*. This assumption can fail!
- One solution when there multicollinearity is to run *Ridge regression*.

Residuals

The optimality conditions from the ordinary least squares problem are:

$$\frac{1}{n}\sum_{i=1}^{n} X_i(Y_i - X_i'\hat{\beta}_n) = 0$$

We define $\hat{U}_i = Y_i - X_i'\hat{\beta}_n$ to be the ith *residual*. It follows that:

$$\sum_{i=1}^{n} X_i\hat{U}_i = \mathbf{0}_{k+1}$$

Define the *predicted value* (or *fitted value*) of Y_i to be $\hat{Y}_i = X_i'\hat{\beta}_n$.

- Therefore, the residuals $\{\hat{U}_i\}_{i=1}^n$ are given by $\hat{U}_i = Y_i - \hat{Y}_i$.
- *Note:* as long as there is a constant in the model, we have $X_{i,1} = 1$. It follows that $\sum_{i=1}^n \hat{U}_i = 0$, i.e. the sum of residuals equals zero.

Decomposing the *TSS*

Suppose we want to measure how well $\{\hat{Y}_i\}_{i=1}^n$ approximates $\{Y_i\}_{i=1}^n$. Just as with simple linear regression, we define the following terms:

- $\text{TSS} = \sum_{i=1}^n (Y_i - \bar{Y}_n)^2$ is the *total sum of squares*
- $\text{ESS} = \sum_{i=1}^n (\hat{Y}_i - \bar{Y}_n)^2$ is the *explained sum of squares*
- $\text{SSR} = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^n \hat{U}_i^2$ is the *sum of squared residuals*

Note that we can decompose the total sum of squares (TSS) as:

$$\sum_{i=1}^n (Y_i - \bar{Y}_n)^2 = \sum_{i=1}^n (\hat{Y}_i - \bar{Y}_n + \hat{U}_i)^2$$

$$= \underbrace{\sum_{i=1}^n (\hat{Y}_i - \bar{Y}_n)^2}_{\text{ESS}} + \underbrace{2\sum_{i=1}^n \hat{U}_i(\hat{Y}_i - \bar{Y}_n)}_{\text{equal to 0}} + \underbrace{\sum_{i=1}^n \hat{U}_i^2}_{\text{SSR}}$$

Coefficient of Determination

The *coefficient of determination* (or R-squared) is defined to be:

$$R^2 = \frac{\text{ESS}}{\text{TSS}} = 1 - \frac{\text{SSR}}{\text{TSS}}$$

Intuitively, if the model fits the data well, then much of the variation in Y_i is captured by the variation in $\hat{Y}_i$. In this case, the R-squared is large.

- Since $\text{TSS} = \text{ESS} + \text{SSR}$, we know that $0 \leq R^2 \leq 1$.
- $R^2 = 1$ if $\text{SSR} = 0$, i.e. all data points lie on a line.
- $R^2 = 0$ if $\text{ESS} = 0$, i.e. X_i does not help us to predict Y_i.

Importantly, R-squared does not tell us anything about the causal relationship between X and Y. It simply measures goodness of fit.

Gauss-Markov Assumptions

For multiple linear regression, the *Gauss-Markov assumptions* are:

(1) The model is $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + U = X'\beta + U$.

(2) We observe an *iid* sample $\{X_i, Y_i\}_{i=1}^n$ of X and Y.

(3) There is no perfect collinearity in the sample (i.e. a unique $\hat{\beta}_n$ exists).

(4) Suppose $E(U|X) = 0$ (i.e. conditional expectation is linear).

(5) The conditional variance is constant: $\text{Var}(U|X) = \sigma^2$.

With these assumptions, we can prove the Gauss-Markov theorem, i.e. that the OLS estimator $\hat{\beta}_n$ is the *best linear unbiased estimator*. Also:

- Unbiasedness of $\hat{\beta}_n$ follows from assumptions $(1) - (4)$.
- Consistency of $\hat{\beta}_n$ comes directly from BLP assumptions.

Statement of the Theorem

Theorem (Gauss-Markov Theorem)

Suppose that all of the Gauss-Markov assumptions are satisfied. Then the OLS estimator $\hat{\beta}_n$ will be the best linear unbiased estimator for β.

Interpretation: $\hat{\beta}_n$ has the "smallest" variance among the class of estimators that are both linear and unbiased (conditional on $\{X_i\}_{i=1}^n$).

We must show $\text{Var}(\hat{\beta}_n | \{X_i\}_{i=1}^n)$ is "smaller" than $\text{Var}(\tilde{\beta} | \{X_i\}_{i=1}^n)$, where:

- $\tilde{\beta}$ is linear, i.e. $\tilde{\beta}$ has the form $\mathbf{A}(\{X_i\}_{i=1}^n) Y$
- $\tilde{\beta}$ is unbiased, i.e. $\mathbb{E}[\tilde{\beta} | \{X_i\}_{i=1}^n] = \beta$

To prove this, first note that $\hat{\beta}_n$ can be written as $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$, where:

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} \quad \text{and} \quad \mathbf{X} = \begin{bmatrix} 1 & X_{1,1} & X_{1,2} & \cdots & X_{1,k} \\ 1 & X_{2,1} & X_{2,2} & \cdots & X_{2,k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n,1} & X_{n,2} & \cdots & X_{n,k} \end{bmatrix} \quad \text{and} \quad \mathbf{U} = \begin{bmatrix} U_1 \\ U_2 \\ \vdots \\ U_n \end{bmatrix}$$

Proving the Theorem

Step 1

The goal is show that $\text{Var}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}|\{X_i\}_{i=1}^n] \leq \text{Var}[\mathbf{AY}|\{X_i\}_{i=1}^n]$ for any linear estimator $\tilde{\beta} = \mathbf{AY}$ that is unbiased for β. As a first step, note:

- We require that $\tilde{\beta}$ satisfies: $\beta = \mathbb{E}[\tilde{\beta}|\{X_i\}_{i=1}^n] = \mathbb{E}[\mathbf{AY}|\{X_i\}_{i=1}^n] = \mathbf{AX}\beta$.
- As this equality holds for all β, it follows that $\mathbf{AX} = I_{k+1}$.

Step 2

As a next step, note that we can write $\text{Var}(\mathbf{AY}|\{X_i\}_{i=1}^n)$ to be:

$$\text{Var}(\mathbf{AY}|\{X_i\}_{i=1}^n) = \mathbf{A}\text{Var}(\mathbf{Y}|\{X_i\}_{i=1}^n)\mathbf{A}' = \mathbf{A}\text{Var}(U|\{X_i\}_{i=1}^n)\mathbf{A}' = \sigma^2\mathbf{AA}'$$

In the case of OLS, $\mathbf{A} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, so the variance is $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$.

Step 3

We must show $\text{Var}(\hat{\beta}_n|\{X_i\}_{i=1}^n)$ is smaller than $\text{Var}(\tilde{\beta}|\{X_i\}_{i=1}^n)$. To do so, we can show that $\sigma^2\mathbf{AA}' - \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$ is a positive semidefinite matrix.

Key Points of the Theorem

The Gauss-Markov Theorem says that, under homoskedasticity, the OLS estimator is the *best* among those that are *linear* and *unbiased*.

- *best* means having the smallest conditional variance $\text{Var}(\mathbf{AY}|\{X_i\}_{i=1}^{n})$
- the result only compares *linear* and *unbiased* estimators
- key assumption: *homoskedasticity*

Nonetheless, this theorem validates the use of OLS among a large class of estimators, and it also suggests instances where we may deviate from OLS.

1. Constructing the OLS Estimator
   - Derivation
   - Measures of Fit
2. Properties of the OLS Estimator
   - Gauss-Markov Theorem
   - Unbiasedness
   - Consistency
   - Asymptotic Normality
3. *Example:* Control Variables & Hypothesis Testing

Bias of $\hat{\beta}_n$

Theorem (Unbiasedness of the OLS Estimator)

Let $\{X_i, Y_i\}_{i=1}^n$ be an i.i.d. sample, and let $Y_i = X_i'\beta + U_i$ be the model under consideration. If there is no collinearity in X_i within the sample and if $E(U_i|X_i) = 0$, then the OLS estimator $\hat{\beta}_n$ is unbiased for β.

Sketch of the Proof

Recall that $\hat{\beta}_n = \left(\frac{1}{n}\sum_{i=1}^n X_i X_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^n X_i Y_i\right)$, where $Y_i = X_i'\beta + U_i$.

$$E(\hat{\beta}_n|\{X_i\}_{i=1}^n) = \beta + \left(\frac{1}{n}\sum_{i=1}^n X_i X_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^n X_i E(U_i|\{X_i\}_{i=1}^n)\right) = \beta$$

where the last equality holds because our sample is *i.i.d.* and $E(U_i|X_i) = 0$. Finally, by the Law of Iterated Expectations, we write:

$$E(\hat{\beta}_n) = E(E(\hat{\beta}_n|\{X_i\}_{i=1}^n)) = \beta$$

Consistency of $\hat{\beta}_n$

Theorem (Consistency of the OLS Estimator)

Let $\{X_i, Y_i\}_{i=1}^n$ be an i.i.d. sample, and let $Y_i = X_i'\beta + U_i$ be the model under consideration. If there is no collinearity in X_i within the sample and if $E(X_i U_i) = 0$, then the OLS estimator $\hat{\beta}_n$ is consistent for β.

Sketch of the Proof

How do we show that $\hat{\beta}_n \xrightarrow{p} \beta$?

- For simplicity, let $\hat{A} = \frac{1}{n}\sum_{i=1}^n X_i X_i'$ and $\hat{B} = \frac{1}{n}\sum_{i=1}^n X_i Y_i$.
- By the WLLN, $\hat{A} \xrightarrow{p} E(X_i X_i')$ and $\hat{B} \xrightarrow{p} E(X_i Y_i)$.
- By the CMT, $\hat{A}^{-1}\hat{B} \xrightarrow{p} E(X_i X_i')^{-1} E(X_i Y_i)$, which equals β.

1. Constructing the OLS Estimator
   - Derivation
   - Measures of Fit

2. Properties of the OLS Estimator
   - Gauss-Markov Theorem
   - Unbiasedness
   - Consistency
   - Asymptotic Normality

3. *Example:* Control Variables & Hypothesis Testing

Limiting Distribution of $\hat{\beta}_n$ *(Part 1)*

Theorem (Limiting Distribution of the OLS Estimator)

Let $\{X_i, Y_i\}_{i=1}^n$ be an i.i.d. sample, and let $Y_i = X_i'\beta + U_i$ be the model under consideration. If there is no collinearity in X_i within the sample and if both $E(X_i U_i) = 0$ and $Var(X_i U_i)$ exists, then:

$$\sqrt{n}(\hat{\beta}_n - \beta) \xrightarrow{d} N(0, \Omega), \quad \text{where: } \Omega = E(X_i X_i')^{-1} Var(X_i U_i) E(X_i X_i')^{-1}$$

Additionally, if $E(X_i|U_i) = 0$ and $Var(U_i|X_i) = \sigma^2$, then:

$$\sqrt{n}(\hat{\beta}_n - \beta) \xrightarrow{d} N(0, \sigma^2 E(X_i X_i')^{-1})$$

Sketch of the Proof

Recall that $\hat{\beta}_n = \left(\frac{1}{n}\sum_{i=1}^n X_i X_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^n X_i Y_i\right)$, where $Y_i = X_i'\beta + U_i$.

$$\sqrt{n}(\hat{\beta}_n - \beta) = \left(\frac{1}{n}\sum_{i=1}^n X_i X_i'\right)^{-1}\left(\sqrt{n} \times \frac{1}{n}\sum_{i=1}^n X_i U_i\right)$$

Limiting Distribution of $\hat{\beta}_n$ *(Part 2)*

Sketch of the Proof (Continued)

Since $E(X_i U_i) = 0$, the Central Limit Theorem guarantees that:

$$\sqrt{n}\Big(\frac{1}{n}\sum_{i=1}^{n} X_i U_i\Big) \xrightarrow{d} N\big(0, \text{Var}(X_i U_i)\big)$$

Note that $\left(\frac{1}{n}\sum_{i=1}^{n} X_i X_i'\right)^{-1} \xrightarrow{p} E(X_i X_i')^{-1}$. By Slutsky's theorem, write:

$$\sqrt{n}(\hat{\beta}_n - \beta) \xrightarrow{d} N(0, \Omega), \quad \text{where: } \Omega = E(X_i X_i')^{-1}\text{Var}(X_i U_i)E(X_i X_i')^{-1}$$

Notice that, if $E(X_i | U_i) = 0$ and $\text{Var}(U_i | X_i) = \sigma^2$, then:

$$\text{Var}(X_i U_i) = E(U_i^2 X_i X_i') = E(E(U_i^2 | X_i) X_i X_i') = \sigma^2 E(X_i X_i')$$

In this case, we conclude that $\Omega = \sigma^2 E(X_i X_i')^{-1}$, as desired.

Estimating Ω under Homoskedasticity

Under the Gauss-Markov assumptions, we know that $\Omega = \sigma^2 E(X_i X_i')^{-1}$.

- A natural estimator for $E(X_i X_i')^{-1}$ is $\left(\frac{1}{n}\sum_{i=1}^n X_i X_i'\right)^{-1}$.
  - **Is it consistent?** Yes! (Apply WLLN and CMT)
  - **Is it unbiased?** No! (Upward bias by Jensen's inequality)
- We select $\hat{\sigma}^2 = \frac{\text{SSR}}{n-k-1}$ as an estimator for σ^2.
  - It turns out that $\hat{\sigma}^2$ is both *consistent* and *unbiased* for σ^2.
  - Dividing by $n-k-1$ rather than n is called a *degrees of freedom* correction. Subtracting $k+1$ from n guarantees unbiasedness.

Given these observations, we can estimate Ω with $\hat{\Omega}$, where:

$$\hat{\Omega}_n = \hat{\sigma}^2 \left(\frac{1}{n}\sum_{i=1}^n X_i X_i'\right)^{-1}$$

Note that $\hat{\Omega}_n$ is a consistent estimator for Ω, i.e. $\hat{\Omega}_n \xrightarrow{p} \Omega$, by the CMT.

Context: Free Preschool

Suppose that an organization implements a high-quality preschool program for children in under-resourced households. You collect data about:

- X_1: a dummy variable for participation in the program
- X_2: parental income
- X_3: parents' years of educational attainment
- W: weekly earnings of children in adulthood
- H: hours/week worked by children in adulthood

You want to understand the long-term impact of access to free preschool. Choosing log hourly wages as your outcome, you write down the model:

$$\log(W/H) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + U$$

Can we (and *should we*) give a causal interpretation to β_1? How might our interpretation of β_1 change if we don't control for X_2 and/or X_3?

Testing Linear Restrictions *(Overview)*

Given our data $\{W_i, H_i, X_{i,1}, X_{i,2}, X_{i,3}\}_{i=1}^n$, we run the regression:

$$\begin{aligned}\log(W_i/H_i) &= \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \beta_3 X_{i,3} + U_i \\ &= X_i'\beta + U_i\end{aligned}$$

where $X_i = (1, X_{i,1}, X_{i,2}, X_{i,3})'$ and $\beta = (\beta_0, \beta_1, \beta_2, \beta_3)'$. From running OLS, we obtain an estimate $\hat{\beta}_n$ for β, which we know must satisfy:

$$\sqrt{n}(\hat{\beta}_n - \beta) \xrightarrow{d} N(0, \Omega)$$

Suppose we want to test whether β satisfies some linear restriction:

$$H_0 : r'\beta = c \quad \text{versus} \quad H_1 : r'\beta \neq c$$

Under H_0, $T_n = \frac{n(r'\hat{\beta}_n - c)}{\sqrt{r'\hat{\Omega}_n r}} \xrightarrow{d} N(0,1)$. So, our test is: $\mathbb{I}\{|T_n| > z_{1-\alpha/2}\}$.

Testing Linear Restrictions *(Examples)*

Suppose we run OLS to estimate the estimate our full linear model:

$$\log(W_i/H_i) = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \beta_3 X_{i,3} + U_i$$

Suppose that n is "large" so that asymptotic approximations are good.

Question 1: How do we test whether β_1 is significantly different from 0?

Question 2: How do we test whether $\beta_2 = \beta_3$?

Question 3: Can we write a 95% confidence interval for $\theta = \beta_2 - \beta_3$?

Question 4: Can we write one for $E[\log(\frac{W}{H})|(X_1, X_2, X_3) = (1, 20k, 12)]$?

Testing Exclusion Restrictions

Suppose we want to know whether controlling for parental income and education really matters? In other words, is it likely that $\beta_2 = \beta_3 = 0$?

$$H_0 : \beta_2 = 0 \text{ and } \beta_3 = 0 \quad \text{versus} \quad H_1 : \text{either } \beta_2 \neq 0 \text{ or } \beta_3 \neq 0$$

Idea: check whether these exclusion restrictions hold by seeing if the restricted model fits the data almost as well as the unrestricted model.

$$\textit{Unrestricted:} \quad \log(W/H) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + U$$
$$\textit{Restricted:} \quad \log(W/H) = \tilde{\beta}_0 + \tilde{\beta}_1 X_1 + V$$

Take SSR_U for the unrestricted model and SSR_R for the restricted model.

- If SSR_U is "close" to SSR_R, then we don't get much more information about Y when we control for X_2 and X_3.
- In this case, it is more likely that $\beta_2 = 0$ and $\beta_3 = 0$.

F-Tests

We measure the decrease in fit by constructing an F-statistic:

$$F_n = \frac{(\mathrm{SSR}_R - \mathrm{SSR}_U)/q}{\mathrm{SSR}_U/(n-k-1)} \xrightarrow{d} F_{q,n-k-1},$$

Here, $q = 2$ (*number of restrictions*) and $k = 3$ (*number of variables*).

We reject H_0 in favor of H_1 when F_n is "large": $F_n > F_{2,n-4,1-\alpha}$.

- If $F_n > F_{2,n-4,0.95}$, then controlling for parental income and education significantly changes the model (at a 5% significance level).
- If $F_n \leq F_{2,n-4,0.95}$, one may infer that the controls are likely irrelevant.

Note: we can test a variety of linear restrictions using an F-test.

- *Example:* to test whether $\beta_3/\beta_2 = 1$ and $\beta_1 = 0$, you could compare the restricted and unrestricted models, then compute an F-statistic.