Applied Macro and Financial Econometrics
Week 1 Review

Thomas T. Yang
Research School of Economics
Australian National University

S2 2023

Outline

- Review
  1. OLS estimator with a single X
  2. OLS estimator with multiple X
  3. Maximum Likelihood Estimation

Review: OLS Estimator with a Single X

The linear model:

$$Y = \beta_0 + X\beta_1 + U.$$

The parameter β_1 is the population regression coefficient
Because we do not observe the population, we do not know β_1
How can we learn about it?

Learning about an unknown population coefficient is the goal of statistical inference

When the unknown population coefficient is part of a linear model, then there is one dominant method to learn about the unknown population coefficient: estimation via OLS

The whole point of EMET2007 or EMET8005 was to expose you to OLS estimation

OLS is one of many ways to estimate β_1

Let's do a brief review of OLS and its properties . . .

(EMET2007 lecture 4 or somewhere in EMET8005)

Definition

The **Ordinary Least Squares (OLS) estimators** are defined by

$$\hat{\beta}_0, \hat{\beta}_1 := \underset{b_0, b_1}{\operatorname{argmin}} \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i)^2$$

In words

- we look at the rhs as a function in b_0 and b_1
- that function happens to be quadratic
- we find the values of b_0 and b_1 that minimize that function
- the values that minimize that function are called solution
- we give the solution a specific name: $\hat{\beta}_0$ and $\hat{\beta}_1$

The mathematics of finding the solution

The basic approach is *multivariate calculus* which you know from high school or EMET1001 or both

First step: differentiate $L = \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i)^2$ wrt b_0, b_1

$$\begin{aligned} \frac{\partial L}{\partial b_0} &= -\sum_{i=1}^{n} 2 \left(Y_i - b_0 - b_1 X_i \right), \\ \frac{\partial L}{\partial b_1} &= -\sum_{i=1}^{n} 2 X_i \left(Y_i - b_0 - b_1 X_i \right). \end{aligned}$$

Second step: set derivatives equal zero (obtain the foc)

$$\begin{aligned} -\sum_{i=1}^{n} 2 \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i \right) &= 0, \\ -\sum_{i=1}^{n} 2 X_i \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i \right) &= 0. \end{aligned}$$

Third step: solve

$$\begin{aligned} n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} X_i &= \sum_{i=1}^{n} Y_i, \\ \hat{\beta}_0 \sum_{i=1}^{n} X_i + \hat{\beta}_1 \sum_{i=1}^{n} X_i^2 &= \sum_{i=1}^{n} X_i Y_i. \end{aligned}$$

Fourth step:

$$\begin{aligned} \hat{\beta}_0 &= \bar{Y} - \hat{\beta}_1 \bar{X}, \\ (\bar{Y} - \hat{\beta}_1 \bar{X}) \sum_{i=1}^{n} X_i + \hat{\beta}_1 \sum_{i=1}^{n} X_i^2 &= \sum_{i=1}^{n} X_i Y_i. \end{aligned}$$

End result:

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

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

The OLS estimator of the slope is equal to the ratio of sample covariance and sample variance!
The OLS estimators are functions of the sample data only
Given the sample data (X_i, Y_i) we can first compute the rhs for $\hat{\beta}_1$ and then we can compute the rhs for $\hat{\beta}_0$
Computer programs such as Stata easily calculate the rhs for you

Given an infinitely large set of possible estimators for β_1, why would we use this complicated looking OLS procedure?
As you know already, it turns out that the OLS estimator has some desirable properties
We assess 'goodness' of an estimator by three properties:

1. bias
2. variance
3. consistency

Let's look at these in turn

Definition

An estimator $\hat{\theta}$ for an unobserved population parameter θ is **unbiased** if its expected value is equal to θ, that is

$$E[\hat{\theta}] = \theta$$

If we draw lots of random samples of size n we obtain lots of estimates $\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3, \ldots$

If the estimator $\hat{\theta}$ is unbiased, then the mean of these estimates will be equal to θ

Note that this is only a thought exercise, in reality we will not draw lots of random samples (we only have one available)

Definition

An estimator $\hat{\theta}$ for an unobserved population parameter θ has **minimum variance** if its variance is (weakly) smaller than the variance of any other estimator of θ. Sometimes we will also say that the estimator is **efficient**.

In EMET2007 Juergen or EMET8005 Tue gave this definition of consistency:

Definition

An estimator $\hat{\theta}$ for an unobserved population parameter θ is **consistent** if it converges in probability to θ.

Consistency is difficult to understand
Here is a useful way to look at it:
If, in a thought experiment, you observe the entire population and apply your estimator to it, you want the resulting estimate to be equal to θ

Let's be slightly more technical (and therefore more precise)

Definition (Convergence in Probability)

Let $\hat{\theta}$ be an estimator of θ. We say that $\hat{\theta}$ **converges in probability** to θ if

$$\Pr(|\hat{\theta} - \theta| > \varepsilon) \to 0 \text{ for all } \varepsilon > 0.$$

We write $\hat{\theta} \xrightarrow{p} \theta$ and say that $\hat{\theta}$ is a **consistent** estimator of the population parameter θ.

Important result about the 'goodness' of the OLS estimator: (EMET2007 lecture 6)

Theorem

Under OLS Assumptions 1 through 4a, the OLS estimator

$$\hat{\beta}_0, \hat{\beta}_1 := \underset{b_0, b_1}{argmin} \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i)^2$$

is BLUE.

The Gauss-Markov theorem provides a theoretical justification for using OLS

Digression: do you remember the 4 Assumptions?

1. conditional mean independence (CMI):

$$E[u_i|X_i] = E[u_i]$$

2. sample data are i.i.d. draw from population distribution
3. finite fourth moments (large outliers are unlikely)
4. homoskedasticity

Putting together the pieces of the puzzle:

- our research objective is the causal effect of X_i on Y_i
- generically there exists a functional relationship between the two: $Y_i = f(X_i, u_i)$
- to make our lives easier, we assume that $f(\cdot)$ is linear
- then the causal effect boils down to the parameter β_1 and can be interpreted as the *average* causal effect
- to estimate that parameter we use OLS
- we obtain the estimate $\hat{\beta}_1$ which is our estimate of the causal effect β_1
- by the Gauss-Markov theorem, the estimate $\hat{\beta}_1$ is 'good' as long as the four OLS Assumptions are satisfied

Review: OLS Estimator with Multiple X

The True Model

- Let X be an $n \times k$ matrix where we have observations on k independent variables for n observations. Since our model will usually contain a constant term, one of the columns in the X matrix will contain only ones. This column should be treated exactly the same as any other column in the X matrix
- Let Y be an $n \times 1$ vector of observations on the dependent variable
- Let ϵ be an $n \times 1$ vector of disturbances or errors.
- Let β be an $k \times 1$ vector of unknown population parameters that we want to estimate.

The True Model continued

- The model will look something like the following:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ \vdots \\ Y_n \end{bmatrix}_{n\times 1} = \begin{bmatrix} 1 & X_{21} & X_{31} & \cdots & X_{k1} \\ 1 & X_{22} & X_{32} & \cdots & X_{k2} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & X_{2n} & X_{3n} & \cdots & X_{kn} \end{bmatrix}_{n\times k} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \vdots \\ \beta_k \end{bmatrix}_{k\times 1} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \vdots \\ \epsilon_n \end{bmatrix}_{n\times}$$

- Or simply

$$Y = X\beta + \epsilon.$$

- or

$$Y_i = \beta_1 + X_{2i}\beta_2 + X_{3i}\beta_3 + \cdots + X_{ki}\beta_k + \epsilon_i,$$

for $i = 1, 2, ..., n$.

The True Model continued

- The model will look something like the following:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ \vdots \\ Y_n \end{bmatrix}_{n\times 1} = \begin{bmatrix} X_1'\beta \\ X_2'\beta \\ \vdots \\ \vdots \\ X_n'\beta \end{bmatrix}_{n\times 1} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \vdots \\ \epsilon_n \end{bmatrix}_{n\times 1}, \text{ with } X_i = \begin{bmatrix} 1 \\ X_{i2} \\ \vdots \\ \vdots \\ X_{ik} \end{bmatrix}.$$

- Since

$$Y_i = X_i'\beta + \epsilon_i.$$

  for $i = 1, 2, ..., n$.

- This is just another way to write

$$Y_i = \beta_1 + X_{2i}\beta_2 + X_{3i}\beta_3 + \cdots + X_{ki}\beta_k + \epsilon_i,$$

  for $i = 1, 2, ..., n$.

Estimates

- The OLS tries to minimize the mean square errors, that is, finding $\hat{\beta}$ to

$$\min \sum_{i=1}^{n} \left(Y_i - b_1 - X_{i2} b_2 - \cdots - X_{ik} b_k \right)^2 .$$

- Let

$$\underset{n\times 1}{e} = \underset{n\times 1}{Y} - \underset{n\times k}{X}\,\underset{k\times 1}{b},$$

that is

$$e_i = Y_i - b_1 - X_{i2} b_2 - \cdots - X_{ik} b_k .$$

- The OLS is equivalent to minimizing

$$\min e'e = \min \begin{bmatrix} e_1 & e_2 & \cdots & e_n \end{bmatrix} \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix} .$$

Estimates continued

- Note

$$\min e'e = \min (Y - Xb)' (Y - Xb).$$

- A note on matrix differentiation. For

$$f(b) = \underset{1\times k}{a'} \underset{k\times 1}{b} = a_1 b_1 + a_2 b_2 + ... + a_k b_k,$$

we have

$$\underset{k\times 1}{\frac{\partial f(b)}{\partial b}} = \begin{bmatrix} \frac{\partial f(b)}{\partial b_1} \\ \frac{\partial f(b)}{\partial b_2} \\ \vdots \\ \frac{\partial f(b)}{\partial b_k} \end{bmatrix} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_k \end{bmatrix}.$$

Estimates continued

- Note

$$L = (Y - Xb)'(Y - Xb) = Y'Y - 2Y'Xb + b'X'Xb.$$

- By matrix differentiation,

$$\frac{\partial L}{\partial b} = 2X'Xb - 2X'Y.$$

- First order condition

$$\begin{aligned} \left.\frac{\partial L}{\partial b}\right|_{b=\hat{\beta}} &= 2X'X\hat{\beta} - 2X'Y = 0, \\ \Rightarrow \quad \hat{\beta} &= (X'X)^{-1} X'Y. \end{aligned}$$

Properties of OLS estimator

- X are uncorrelated with the residuals (The intution is "projection").

$$\hat{e} = Y - X\hat{\beta} = Y - X\left(X'X\right)^{-1}X'Y = \left(I_n - X\left(X'X\right)^{-1}X'\right)Y.$$

So

$$\begin{aligned} X'\hat{e} &= X'\left(I - X\left(X'X\right)^{-1}X'\right)Y = X'Y - X'X\left(X'X\right)^{-1}X'Y \\ &= X'Y - X'Y = 0. \end{aligned}$$

- The sample mean of $\hat{e}$ $\left(\frac{1}{n}\sum_{i=1}^{n}\hat{e}_i\right)$ is zero because the first column of X is a vector of ones.

Review: Maximum Likelihood Estimation

Intuition for Maximum Likelihood Estimation (MLE)

Scenario: We have a machine that produces coins. These could be fair (50% heads, 50% tails) or unfair (60% heads, 40% tails, etc.).
We are given a coin from this machine, but we don't know if it's fair or not. Our task is to estimate the "fairness" of the coin, i.e., the probability of getting a head when the coin is flipped (let's call it p).

Intuition for MLE: Flipping the Coin

We start by flipping the coin multiple times and recording the outcome of each flip.
For simplicity, let's say we flip the coin 10 times and get 7 heads and 3 tails. Now, we need to figure out what value of p makes the data we observed (7 heads and 3 tails) the most likely.

Intuition for MLE: Maximum Likelihood

We could try different values of p and calculate the probability of getting 7 heads and 3 tails for each.
If the coin is fair ($p = 0.5$), the probability of getting 7 heads and 3 tails would be relatively small. But if $p = 0.7$, the probability of getting 7 heads and 3 tails would be higher.
So, we might guess that $p = 0.7$ is a better estimate of the fairness of the coin. This is what MLE does. It finds the parameter values that make the observed data the most likely.

Maximum Likelihood Estimation (Appendix 11.2)

- Bernoulli random variable

$$Y = \begin{cases} 1 & \text{with probability } p \\ 0 & \text{with probability } 1-p \end{cases}.$$

- n i.i.d. $y_1, y_2, ..., y_n$.
- We do not know p, and we want to estimate p using $y_1, y_2, ..., y_n$.
- The probability of it happens is:

$$\begin{aligned} & P(Y_1 = y_1, Y_2 = y_2, ..., Y_n = y_n) \\ = \; & P(Y_1 = y_1) \times P(Y_2 = y_2) \times ... \times P(Y_n = y_n) \\ = \; & p^{y_1}(1-p)^{1-y_1} \times p^{y_2}(1-p)^{1-y_2} \times ... \times p^{y_n}(1-p)^{1-y_n} \end{aligned}$$

- MLE of p is the value of p that maximizes the likelihood equation.

Maximum Likelihood Estimation (Appendix 11.2) continued

- log is a strictly increasing function.
- It is equivalent to maximize its log likelihood

$$\begin{aligned} & L = \log P\left(Y_1 = y_1, Y_2 = y_2, ..., Y_n = y_n\right) \\ = \quad & \log p^{\sum_{i=1}^{n} y_i} \left(1-p\right)^{\sum_{i=1}^{n}(1-y_i)} \\ = \quad & \log p \sum_{i=1}^{n} y_i + \log\left(1-p\right) \sum_{i=1}^{n} \left(1-y_i\right). \end{aligned}$$

- First derivative:

$$\begin{aligned} \frac{dL}{dp} &= \frac{1}{p} \sum_{i=1}^{n} y_i - \frac{1}{1-p} \sum_{i=1}^{n} \left(1-y_i\right) \\ &= \frac{\left(1-p\right) \sum_{i=1}^{n} y_i - p \sum_{i=1}^{n} \left(1-y_i\right)}{p\left(1-p\right)} \\ &= \frac{\sum_{i=1}^{n} y_i - np}{p\left(1-p\right)} \end{aligned}$$

Maximum Likelihood Estimation (Appendix 11.2) continued

- First order condition:

$$\begin{aligned} \sum_{i=1}^{n} y_i - n\hat{p} &= 0 \\ &\Longrightarrow \\ \hat{p} &= n^{-1} \sum_{i=1}^{n} y_i = \bar{y}. \end{aligned}$$

- Example, coin tossing, let

$$Y_i = \begin{cases} 1 & \text{if heads, with probability } p \\ 0 & \text{if tails, with probability } 1-p \end{cases}.$$

- Independent toss a coin n times, we estimate p as the sample average. Make sense?

Maximum Likelihood Estimation (Appendix 11.2) continued

The theory of maximum likelihood estimation says that is the most efficient estimator of p – of all possible estimators! – at least for large n. (Much stronger than the Gauss-Markov theorem). For this reason the MLE is primary estimator used for models that in which the parameters (coefficients) enter nonlinearly.

Maximum Likelihood Estimation continued

Poisson distribution

$$P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}.$$

- n i.i.d. $x_1, ..., x_n$
- The log likelihood is

$$\begin{aligned} L &= \log P(X_1 = x_1, X_2 = x_2, ..., X_n = x_n) \\ &= \sum_{i=1}^{n} \log P(X_i = x_i) \\ &= \sum_{i=1}^{n} \log\left(\frac{\lambda^{x_i} e^{-\lambda}}{x_i!}\right) \\ &= \sum_{i=1}^{n} \left[x_i \log(\lambda) - \lambda - \log(x_i!)\right] \end{aligned}$$

Maximum Likelihood Estimation continued

- First order derivative

$$\frac{dL}{d\lambda} = \sum_{i=1}^{n} \left(\frac{x_i}{\lambda} - 1 \right)$$

- Thus

$$\hat{\lambda} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Maximum Likelihood Estimation continued

Normal distribution

$$f\left(x;\mu,\sigma^2\right)=\frac{1}{\sigma\sqrt{2\pi}}\exp\left\{-\frac{\left(x-\mu\right)^2}{2\sigma^2}\right\}$$

- n i.i.d. $x_1,...,x_n$.
- The log likelihood is

$$\begin{aligned}
L &= \log f\left(X_1=x_1,X_2=x_2,...,X_n=x_n\right)\\
&= \sum_{i=1}^{n}\log f\left(X_i=x_i\right)=\sum_{i=1}^{n}\log\frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{\left(x_i-\mu\right)^2}{2\sigma^2}\right]\\
&= \sum_{i=1}^{n}\left[-\log\sigma-\frac{1}{2}\log\left(2\pi\right)-\frac{\left(x_i-\mu\right)^2}{2\sigma^2}\right]\\
&= -n\log\sigma-\frac{n}{2}\log\left(2\pi\right)-\sum_{i=1}^{n}\frac{\left(x_i-\mu\right)^2}{2\sigma^2}
\end{aligned}$$

Maximum Likelihood Estimation continued

Normal distribution

$$f\left(x;\mu,\sigma^2\right)=\frac{1}{\sigma\sqrt{2\pi}}\exp\left\{-\frac{\left(x-\mu\right)^2}{2\sigma^2}\right\}$$

- n i.i.d. $x_1,...,x_n$.
- The log likelihood is

$$\begin{aligned}
L &= \log f\left(X_1=x_1,X_2=x_2,...,X_n=x_n\right)\\
&= \sum_{i=1}^{n}\log f\left(X_i=x_i\right)=\sum_{i=1}^{n}\log\frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{\left(x_i-\mu\right)^2}{2\sigma^2}\right]\\
&= \sum_{i=1}^{n}\left[-\log\sigma-\frac{1}{2}\log\left(2\pi\right)-\frac{\left(x_i-\mu\right)^2}{2\sigma^2}\right]\\
&= -n\log\sigma-\frac{n}{2}\log\left(2\pi\right)-\sum_{i=1}^{n}\frac{\left(x_i-\mu\right)^2}{2\sigma^2}
\end{aligned}$$

Maximum Likelihood Estimation continued

First order conditions:

$$\begin{aligned}
\left.\frac{\partial L}{\partial \sigma}\right|_{(\hat{\mu},\hat{\sigma})} &= -\frac{n}{\hat{\sigma}} + \frac{1}{\hat{\sigma}^3}\sum_{i=1}^{n}(x_i - \hat{\mu})^2 = 0 \\
\left.\frac{\partial L}{\partial \mu}\right|_{(\hat{\mu},\hat{\sigma})} &= \frac{1}{\hat{\sigma}^2}\sum_{i=1}^{n}(x_i - \hat{\mu}) = 0,
\end{aligned}$$

which yields

$$\begin{aligned}
\hat{\mu} &= \frac{1}{n}\sum_{i=1}^{n} x_i, \\
\hat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})^2 .
\end{aligned}$$

MLE for Linear Models: Normal Distribution Assumption

We often assume that the errors in a linear regression model are normally distributed:

$$y = X\beta + \epsilon$$

where X is the matrix of predictors, β is the vector of coefficients, and ϵ is the vector of errors such that $\epsilon \sim N(0, \sigma^2 I)$. Under this assumption, we can apply MLE to estimate the parameters β and σ^2.

The Likelihood Function for Linear Regression

The likelihood function for this model is given by:

$$L(\beta, \sigma^2|y) = (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}(y - X\beta)^T(y - X\beta)\right)$$

We take the natural log to simplify the expression, obtaining the log-likelihood function:

$$l(\beta, \sigma^2|y) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}(y - X\beta)^T(y - X\beta)$$

MLE for Linear Regression: Derivation of β

To find the MLE of β, we take the derivative of the log-likelihood with respect to β and set it to zero:

$$\frac{\partial l(\beta, \sigma^2|y)}{\partial \beta} = X^T(y - X\beta) = 0$$
$$\Rightarrow \hat{\beta} = (X^T X)^{-1} X^T y$$

This shows that the MLE of β is the same as the least squares estimator of β.

MLE for Linear Regression: Derivation of σ^2

To find the MLE of σ^2, we take the derivative of the log-likelihood with respect to σ^2 and set it to zero:

$$\frac{\partial l(\beta, \sigma^2|y)}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}(y - X\beta)^T(y - X\beta) = 0$$

$$\Rightarrow \hat{\sigma^2} = \frac{1}{n}(y - X\hat{\beta})^T(y - X\hat{\beta})$$

This shows that the MLE of σ^2 is the residual sum of squares divided by the number of observations.