LECTURE 13: STATISTICAL PROPERTIES OF ORDINARY LEAST SQUARES

MECO 7312.
INSTRUCTOR: DR. KHAI CHIONG
NOVEMBER 20, 2024

1. Linear regression models

Let (Y, X, ϵ) be random variables such that:

$$Y = a + bX + \epsilon \tag{1}$$

a and b are unknown parameters, where $\mathbb{E}[\epsilon|X] = 0$ and $\mathbb{E}[\epsilon] = 0$. Show that $\mathbb{E}[\epsilon|X] = 0$ implies the following: (i) $\mathbb{E}[X\epsilon] = 0$, (ii) $\text{Cov}(X, \epsilon) = 0$, and (iii) $\mathbb{E}[\epsilon] = 0$. However, $\text{Cov}(X, \epsilon) = 0$ and $\mathbb{E}[\epsilon] = 0$ does not imply that $\mathbb{E}[\epsilon|X] = 0$. On the other hand, the independence of X and ϵ, along with $\mathbb{E}[\epsilon] = 0$, implies that $\mathbb{E}[\epsilon|X] = 0$, which is also known as the conditional mean independence.

Suppose n i.i.d random samples: (y_i, x_i, ϵ_i) for $i = 1, \ldots, n$ are drawn from the data-generating model, but we only observe $(y_i, x_i)_{i=1}^n$ as our dataset.

Three ways of estimating a and b, all leading to the same estimators!

Method of moments.

$$\mathbb{E}[\epsilon X] = 0 \tag{2}$$

$$\mathbb{E}[XY - aX - bX^2] = 0 \tag{3}$$

$$\mathbb{E}[\epsilon] = 0 \tag{4}$$

$$\mathbb{E}[Y - a - bX] = 0 \tag{5}$$

The Method of Moments estimators are:

$$\hat{b} = \frac{\frac{1}{n}\sum_{i=1}^n x_i y_i - \bar{x}\bar{y}}{\frac{1}{n}\sum_{i=1}^n x_i^2 - \bar{x}^2}$$

$$\hat{a} = \bar{y} - \hat{b}\bar{x}$$

Or we can use Maximum Likelihood Estimator, but we have to additionally assume that $\epsilon \sim \mathcal{N}(0, \sigma^2)$, or equivalently, $Y - a - bX \sim \mathcal{N}(0, \sigma^2)$. Therefore,

$$L(a, b, \sigma | x_1, y_1, \ldots, x_n, y_n) = \prod_{i=1}^{n} \phi\left(\frac{y_i - a - bx_i}{\sigma}\right) \tag{6}$$

$$\underset{a,b,\sigma}{\operatorname{argmax}} \sum_{i=1}^{n} \log \phi\left(\frac{y_i - a - bx_i}{\sigma}\right) \tag{7}$$

Where ϕ is the pdf of the standard Normal.

The third method is to minimize the sum of squared errors using calculus: $\min_{a,b} \sum_{i=1}^{n} (y_i - a - bx_i)^2$.

1.1. Multivariate linear regression

Now consider:

$$Y = \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_K X_K + \epsilon \tag{8}$$

Suppose we draw n i.i.d random samples: $(y_i, x_{i1}, x_{i2}, \ldots, x_{iK}, \epsilon_i)$ for $i = 1, \ldots, n$ from the data-generating process above. However, ϵ_i is unobserved, we only observe $(y_i, x_{i1}, x_{i2}, \ldots, x_{iK})$, which we refer to as the "dataset." The dataset is related as follows.

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{iK} + \epsilon_i \tag{9}$$

We can manipulate this equation using Matrix Algebra.

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$$

$$\boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}_1, & \boldsymbol{X}_2, \cdots, & \boldsymbol{X}_K \end{bmatrix}$$

Where $\boldsymbol{X}_k$ is a $n \times 1$ column vector containing the k-th explanatory variable. Other names for explanatory variable: features (used by computer scientists), covariates, regressors (used by economists).

$$\boldsymbol{X}_k = \begin{bmatrix} x_{1k} \\ \vdots \\ x_{nk} \end{bmatrix}$$

$(x_{1k}, x_{2k}, \ldots, x_{ik}, \ldots, x_{nk})$ are called *observations* for the k-th covariate.

$$\boldsymbol{X} = \begin{bmatrix} x_{11} & \cdots & x_{1K} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nK} \end{bmatrix}$$

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_K \end{bmatrix}$$

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

$\boldsymbol{X}\boldsymbol{\beta}$ is the matrix product of a $n \times k$ matrix with a $k \times 1$ matrix, resulting in a $n \times 1$ matrix.

Our entire dataset are contained in the data matrix $[\boldsymbol{y}, \boldsymbol{X}]$.

2. Ordinary Least Squares (OLS) estimator

How do we estimate the $\boldsymbol{\beta}$? If we were to use Method of Moments, we need at least K number of moments conditions.

The assumption we need is that $\boldsymbol{X}$ is *exogenous*, also known as the *conditional mean independence assumption*: $\mathbb{E}[\epsilon|X_1] = 0$, $\mathbb{E}[\epsilon|X_2] = 0$, $\ldots$, $\mathbb{E}[\epsilon|X_K] = 0$. The error term is (conditionally mean) independent of each of the K explanatory variable.

The sample moment conditions can be written as: $\sum_{i=1}^{n} x_{i1}\epsilon_i = 0$, $\sum_{i=1}^{n} x_{i2}\epsilon_i = 0$, $\ldots$, $\sum_{i=1}^{n} x_{ik}\epsilon_i = 0$. In matrix notation:

$$\boldsymbol{X}_1^T \boldsymbol{\epsilon} = 0 \tag{10}$$

$$\vdots$$

$$\boldsymbol{X}_k^T \boldsymbol{\epsilon} = 0 \tag{11}$$

Now $\boldsymbol{X}_k^T$ is the matrix transpose of $\boldsymbol{X}_k$, therefore, $\boldsymbol{X}_k^T$ is a $1 \times n$ row vector. $\boldsymbol{X}_k^T = \begin{bmatrix} x_{1k}, & x_{2k}, \cdots, & x_{nk} \end{bmatrix}$.

Finally, all the sample moment conditions can be summarized as just:

$$\boldsymbol{X}^T \boldsymbol{\epsilon} = \mathbf{0}_K \tag{12}$$

Where $\mathbf{0}_K$ is a $K \times 1$ vector of zeros. $\boldsymbol{X}^T$ is a $K \times n$ matrix, while $\boldsymbol{\epsilon}$ is a $n \times 1$ matrix, therefore their matrix product has dimension $K \times 1$.

Now, we can derive the OLS (Ordinary Least Square) estimators:

$$\boldsymbol{X}^T \boldsymbol{\epsilon} = \mathbf{0} \tag{13}$$

$$\boldsymbol{X}^T (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) = \mathbf{0} \tag{14}$$

$$\boldsymbol{X}^T \boldsymbol{y} - \boldsymbol{X}^T \boldsymbol{X}\boldsymbol{\beta} = \mathbf{0} \tag{15}$$

$$(\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{y} - (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{X}\boldsymbol{\beta} = \mathbf{0} \tag{16}$$

$$(\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{y} - \boldsymbol{\beta} = \mathbf{0} \tag{17}$$

$$\boldsymbol{\beta} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{y} \tag{18}$$

Note that the right-hand side of $\boldsymbol{\beta} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{y}$ consists entirely of the components of the data matrix. Therefore this is a valid estimator.

Although the OLS estimator can be derived under a weaker assumption of zero covariance between ϵ and the covariates, i.e. $\mathbb{E}[X_k \epsilon] = 0$ for $k = 1, \ldots, K$, we use the stronger assumption of conditional mean independence because zero covariance does not guarantee the unbiasedness of OLS estimator – see Section 4.

2.1. OLS simulation

It is instructive to implement OLS estimators in a programming language of your choice. For this section, we refer to the Python Notebook or R Markdown.

Let the true model and data-generating process be $Y = 2 - 3X_1 + 0.5X_2 + \epsilon$, where $\epsilon \sim \mathcal{N}(0, 2)$, $X_1 \sim$ Exponential(0.5), and $X_2 \sim \mathcal{N}(-1, 1)$. We now generate dataset from this DGP assuming random i.i.d sampling with a sample size of $n = 1000$.

The dataset is $(y_1, x_{i1}, x_{i2}, \epsilon_i)_{i=1}^n$. Since y_i is related to the other variables, we generate y_i through $y_i = 2 - 3x_{i1} + 0.5x_{i2} + \epsilon_i$. The true coefficient/parameters are therefore $\boldsymbol{\beta} = [2, -3, 0.5]^T$. After we generated the data matrix, we stack them according to the data matrix $[\boldsymbol{y}, \boldsymbol{X}]$. We compute the OLS estimator $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$, and compare it to the true value.

3. Multicollinearity

The OLS estimator is $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$.

The matrix $(\boldsymbol{X}^T\boldsymbol{X})$ needs to be invertible. $(\boldsymbol{X}^T\boldsymbol{X})$ is invertible if and only if $\text{rank}(X) = K \leq n$, i.e. the columns of X are linearly independent and the number of rows is greater than the number of columns.

Suppose a particular column of $\boldsymbol{X}$ can be written as a linear function of some other columns of $\boldsymbol{X}$ (for example, $\boldsymbol{X}_k = \lambda_1\boldsymbol{X}_1 + \lambda_2\boldsymbol{X}_2$), then we say that there is a **perfect multicollinearity**. The regressors are linearly dependent. $(\boldsymbol{X}^T\boldsymbol{X})$ does not have an inverse – OLS estimator is ill-defined.

In general, even when there is no exact linear relationship between the regressors, OLS estimator will run into problem when one of the regressors are highly correlated with another regressor. This is the **multicollinearity** problem. The inverse $(\boldsymbol{X}^T\boldsymbol{X})$ is *almost singular*. Computation of the inverse of an almost singular matrix is highly unstable and numerically imprecise. When one of the regressors are too similar to another regressor, we cannot separately identify their respective coefficients.

Consider the simulation exercise before. Let the (true) data-generating process be $Y_i = 2 - 4X_{i1} + 0.5X_{i2} + \epsilon_i$ for $i = 1, \ldots, 1000$, where $\epsilon_i \sim$ i.i.d $\mathcal{N}(0, 2)$, $X_{i1} \sim$ i.i.d Exponential(0.5), and $X_{i2} = 5 - 2X_{i1}$.

Now let $X_{i2} = 5 - 2X_{i1} + v_i$, where $v_i \sim \mathcal{N}(0, 0.1)$.

Multicollinearity can be detected by calculating the condition number of the matrix $(\boldsymbol{X}^T\boldsymbol{X})$. When the condition number is high, the matrix is ill-conditioned and almost singular.[1]

4. Unbiasedness of OLS estimators

What does unbiasedness mean here? Recall the simulation exercise before – we get different OLS estimates in different simulation when we draw a different random

[1] The condition number is computed by finding the square root of the maximum eigenvalue divided by the minimum eigenvalue of the matrix. If the condition number is above 30, the regression may have significant multicollinearity. The condition number of a matrix indicates the potential sensitivity of the computed inverse to small changes in the original matrix.

sample from the DGP. What is the average of those OLS estimates over infinitely many simulations?

To examine the unbiasedness of the OLS estimator, we need a ground truth, and say that it is unbiased with respect to a data-generating process.

DGP: $(y_i, x_{i1}, x_{i2}, \ldots, x_{iK}, \epsilon_i)_{i=1}^n$ are generated from some joint distribution that obeys the equation $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}_0 + \boldsymbol{\epsilon}$. We can be agnostic about this joint distribution, in particular, ϵ_i may not even be i.i.d across i. For instance, $(x_{i1}, x_{i2}, \ldots, x_{iK}, \epsilon_i)_{i=1}^n$ could be generated i.i.d from the distribution $f(x_1, x_2, \ldots, x_K, \epsilon)$, or from some non-i.i.d distribution $f((x_{i1}, x_{i2}, \ldots, x_{iK}, \epsilon_i)_{i=1}^n)$.

Let $\hat{\boldsymbol{\beta}}$ be the OLS estimator.

$$\mathbb{E}[\hat{\boldsymbol{\beta}}] = \mathbb{E}[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}] \tag{19}$$

$$= \mathbb{E}[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{\beta}_0 + \boldsymbol{\epsilon})] \tag{20}$$

$$= \mathbb{E}[\boldsymbol{\beta}_0 + (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\epsilon}] \tag{21}$$

$$= \boldsymbol{\beta}_0 + \mathbb{E}[\mathbb{E}[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\epsilon}|\boldsymbol{X}]] \tag{22}$$

$$= \boldsymbol{\beta}_0 + \mathbb{E}[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\,\mathbb{E}[\boldsymbol{\epsilon}|\boldsymbol{X}]] \tag{23}$$

The Law of Iterated Expectation is applied in the last two equations. It is clear that a sufficient condition for the unbiasedness of OLS estimator is that $\mathbb{E}[\boldsymbol{\epsilon}|\boldsymbol{X}] = \boldsymbol{0}$. This expression means that $\mathbb{E}[\epsilon_i|\boldsymbol{X}] = 0$ for all $i = 1, \ldots, n$. Further unpacking, it means that ϵ_i for each $i = 1, \ldots, n$ is (conditionally mean) independent from the entire matrix $\boldsymbol{X}$, i.e. $\mathbb{E}[\epsilon_i|x_{11}, \ldots, x_{ik}, \ldots, x_{nk}] = 0$.

Two possible ways to satisfy $\mathbb{E}[\epsilon_i|\boldsymbol{X}] = 0$.

(1) $(\epsilon_i, x_{i1}, x_{i2}, \ldots, x_{iK})$ are independently and identically distributed across i from some probability distributions, and $\mathbb{E}[\epsilon_i|x_{i1}, x_{i2}, \ldots, x_{iK}] = 0$. In the i.i.d case, we can drop the i subscript, and write $\mathbb{E}[\epsilon|X_1, X_2, \ldots, X_K] = 0$. Now $\mathbb{E}[\epsilon|X_1, X_2, \ldots, X_K] = 0$ implies that $\mathbb{E}[\epsilon|X_k] = 0$ for $k = 1, \ldots, K$.[2] This is what we assumed when we use the Method of Moments to derive the OLS estimator. This is a sufficient but not a necessary condition for unbiasedness.

In particular, zero covariances between the error term and the covariates do not guarantee unbiasedness, since zero covariance does not imply conditional mean independence. However, independence of the error term and the covariates would be sufficient for unbiasedness. When a covariate is correlated with the error term,

[2] By the Tower Property.

the conditional mean independence assumption does not hold, and we say that the covariate is endogenous.

(2) $(\epsilon_i, x_{i1}, x_{i2}, \ldots, x_{iK})$ are *not necessarily* i.i.d across i, but $\mathbb{E}[\epsilon_i|\boldsymbol{X}] = 0$ for each i. Therefore, in the context of time-series where i.i.d does not hold true, OLS can still be unbiased. There can be no correlation between the error term at time t and your covariates at time $1, \ldots, t, t+1, t+2, \ldots$. For example, we require that $\mathbb{E}[\epsilon_t|x_{1k}, \ldots, x_{tk}, \ldots, x_{Tk}] = 0$ for covariate k, in order for $\mathbb{E}[\epsilon_t|\boldsymbol{X}] = 0$.

4.1. Omitted variable bias

$\mathbb{E}[\boldsymbol{\epsilon}|\boldsymbol{X}] = \mathbf{0}$ is also called the exogeneity condition. Exogeneity can be violated under many circumstances – whenever the regressor is correlated with the error term. For example, a common scenario is when a variable that explains the dependent variable is omitted from the estimating equation, and this omitted variable is correlated with another explanatory variable. As such, OLS is biased (omitted variable bias). In the companion Python notebook, we worked out the direction of the biases when there is a omitted variable.

Suppose the DGP is $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$, with $\mathbb{E}[\epsilon|X_1] = 0$ and $\mathbb{E}[\epsilon|X_2] = 0$. If we omit the variable X_2, it is equivalent to the DGP $Y = \beta_0 + \beta_1 X_1 + v$, where $v = \beta_2 X_2 + \epsilon$. Assuming i.i.d, we know that a sufficient condition for unbiasedness is $\mathbb{E}[v|X_1] = 0$. Therefore, if X_1 and X_2 are independent, then OLS estimator is unbiased. If X_1 and X_2 are correlated, then $\mathbb{E}[v|X_1] = \beta_2 \mathbb{E}[X_2|X_1] \neq 0$. The direction of the bias depends on the sign of the correlation between X_1 and v, which is the sign of $\beta_2 \text{Cov}(X_1, X_2)$.