LECTURE 13: STATISTICAL PROPERTIES OF ORDINARY LEAST SQUARES

MECO 7312.
INSTRUCTOR: DR. KHAI CHIONG
NOVEMBER 13, 2019

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 $\mathbb{E}[\epsilon|X] = 0$. Show that $\mathbb{E}[\epsilon|X] = 0$ implies the following: (i) $\mathbb{E}[X\epsilon] = 0$, (ii) $\text{Cov}(X, \epsilon) = 0$, and (iii) $\mathbb{E}[\epsilon] = 0$.

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.

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

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

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

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

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,

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

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

Where ϕ is the pdf of the standard Normal.

Or we can 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 observe n i.i.d random samples: $(y_i, x_{i1}, x_{i2}, \ldots, x_{iK})$ for $i = 1, \ldots, n$.

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

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*: $\mathbb{E}[\epsilon|X_1] = 0$, $\mathbb{E}[\epsilon|X_2] = 0$, ..., $\mathbb{E}[\epsilon|X_K] = 0$. The error term is (conditionally mean) independent of each of the K explanatory variable.

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

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

$$\vdots$$

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

Where $\boldsymbol{X}_k$ is the k-th column of the data matrix $\boldsymbol{X}$, it contains all n observations for the covariate X_k. Now $\boldsymbol{X}^T$ is the matrix transpose of $\boldsymbol{X}$, therefore, $\boldsymbol{X}^T$ is a $1 \times n$ row vector. $\boldsymbol{X}^T = \begin{bmatrix} x_{1k}, & x_{2k}, \cdots, & x_{nk} \end{bmatrix}$.

Finally, the 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.

2.1. OLS simulation

It is very instructive to implement OLS estimators in a programming language of your choice. For this section, please refer to the R Markdown (lecture13.rmd).

Let the (true) data-generating process be $Y_i = 2 - 3X_{i1} + 0.5X_{i2} + \epsilon_i$ 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 $x_{i2} \sim$ i.i.d $\mathcal{N}(-1, 1)$. The true coefficient/parameters are therefore $\boldsymbol{\beta} = [2, -3, 0.5]^T$. Since y_i is related to the other variables, we generate y_i through $y_i = 2 - 3x_{i1} + 0.5x_{i2} + \epsilon_i$. After we generated these numbers, we then ignore ϵ_i and 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.

Are the covariates $\boldsymbol{X}$ exogenous here? We now consider another data-generating process where the exogeneity assumption is violated.

Suppose we only regress y_i on 1 and x_{1i}, what happens to the OLS estimates? In another words, we now try to estimate the following *misspecified* model: $Y_i = \beta_1 + \beta_2 X_{i1} + v_i$, while the true model is $Y_i = \beta_1 + \beta_2 X_{i1} + \beta_3 X_{i2} + \epsilon_i$. As such, $v_i = \beta_2 X_{i2} + \epsilon_i$.

In order to correctly estimate the coefficient β_2 from $Y_i = \beta_1 + \beta_2 X_{i1} + v_i$, OLS estimator relies on the assumption that $\mathbb{E}[v_i|X_{i1}] = 0$. But because $v_i = \beta_2 X_{i2} + \epsilon_i$, we have: $\mathbb{E}[v_i|X_{i1}] = \mathbb{E}[\beta_2 X_{i2} + \epsilon_i|X_{i1}] = \beta_2\,\mathbb{E}[X_{i2}|X_{i1}]$. Therefore, we can still get consistent estimate of β_2 if $\mathbb{E}[X_{i2}|X_{i1}] = 0$, e.g. when the two covariates

are independent. We verify through the simulation that the OLS estimates are inconsistent when $(X_{1i}, X_{2i}) \sim$ i.i.d multivariate Normal with a covariance matrix of $\begin{bmatrix} 1 & 1.5 \\ 1.5 & 3 \end{bmatrix}$.

How do we estimate the variance of ϵ_i? Sum of squares residuals.

$$\mathbb{E}[\epsilon^2] \approx \frac{1}{n}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^T(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) \tag{19}$$

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. A square matrix that is not invertible is called singular. A square matrix is singular if and only if its determinant is 0.

The matrix $(\boldsymbol{X}^T\boldsymbol{X})$ has an inverse if the columns of $\boldsymbol{X}$ are linearly independent. ($(\boldsymbol{X}^T\boldsymbol{X})$ has a full column rank).

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. When one of the regressors are too similar to another regressor, we cannot separately identify their respective coefficients.

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.

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]

[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

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

Let the true data-generating process be: $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}_0 + \epsilon$. Here, we do not have to specify the the distribution of ϵ or $\boldsymbol{X}$. Let $\hat{\boldsymbol{\beta}}$ be the OLS estimator.

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

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

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

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

$$= \boldsymbol{\beta}_0 + \mathbb{E}[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\,\mathbb{E}[\boldsymbol{\epsilon}|\boldsymbol{X}]] \tag{24}$$

Expectations above are taken row-wise (for every observation $i = 1, \ldots, n$). 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 $\mathbb{E}[\boldsymbol{\epsilon}|\boldsymbol{X}] = 0$. This expression means that $\mathbb{E}[\epsilon_i|\boldsymbol{X}] = 0$ for all $i = 1, \ldots, n$.

There are two possible ways to satisfy $\mathbb{E}[\epsilon_i|\boldsymbol{X}] = 0$. EITHER, $\mathbb{E}[\epsilon_i|\boldsymbol{X}_i] = 0$ and $(\epsilon_i, \boldsymbol{X}_i)$ are i.i.d across i from some probability distributions. OR, $(\epsilon_i, \boldsymbol{X}_i)$ are *not necessarily* i.i.d across i, but $\mathbb{E}[\epsilon_i|\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n] = 0$ for each i. Therefore, in the context of time-series where i.i.d does not hold true, OLS can still be unbiased. Either of these are known as the *exogeneity* condition.

5. OLS covariance matrix

Having established that OLS is unbiased under the *exogeneity* condition, what is $\text{Var}(\hat{\boldsymbol{\beta}})$? When we take the variance of a k-dimensional vector, we mean the k-by-k variance covariance matrix.

Recall that

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.

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{\beta}_0 + \boldsymbol{\epsilon}) \tag{25}$$

$$= \boldsymbol{\beta}_0 + (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\epsilon} \tag{26}$$

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\epsilon} \tag{27}$$

The variance-covariance matrix is given by:

$$\mathbb{E}[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^T] = \mathbb{E}[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\epsilon}((\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\epsilon})^T] \tag{28}$$

$$= \mathbb{E}[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}] \tag{29}$$

$$= \mathbb{E}[\mathbb{E}[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}|\boldsymbol{X}]] \tag{30}$$

$$= \mathbb{E}[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\,\mathbb{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T|\boldsymbol{X}]\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}] \tag{31}$$

$$= \mathbb{E}[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\sigma^2\boldsymbol{I}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}] \tag{32}$$

$$= \sigma^2\,\mathbb{E}[(\boldsymbol{X}^T\boldsymbol{X})^{-1}] \tag{33}$$

An estimate of the variance-covariance matrix is then

$$\text{Var}(\hat{\boldsymbol{\beta}}) = \hat{\sigma}^2(\boldsymbol{X}^T\boldsymbol{X})^{-1} \tag{34}$$

The precision of the OLS estimator is defined as the inverse of the variance-covariance matrix. We see that the precision of the OLS estimator decreases as σ^2 increases.

Check using simulation that the variance of OLS estimators also increase when: (i) sample size n decreases, and (ii) when the collinearity between regressors increases.