LECTURE 15: STATISTICAL PROPERTIES OF ORDINARY LEAST SQUARES

MECO 7312.
INSTRUCTOR: DR. KHAI CHIONG

1. OLS covariance matrix

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

What is the sampling distribution of $\hat{\boldsymbol{\beta}}$? We must start from the data-generating process. The sampling distribution is induced by the randomness in $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}_0 + \boldsymbol{\epsilon}$, where either $\boldsymbol{X}$ is nonstochastic (fixed) and $\boldsymbol{\epsilon}$ is random, or $(\boldsymbol{X}, \boldsymbol{\epsilon})$ are both random. In scientific experiments, we can treat $\boldsymbol{X}$ as being fixed. In either case, all OLS-related derivations are similar.

Previously, we have seen that under the *exogeneity* condition $\mathbb{E}[\boldsymbol{\epsilon}|\boldsymbol{X}] = \boldsymbol{0}$, the OLS estimator is unbiased, i.e. $\mathbb{E}[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}_0$.

Now we want to examine other features of the sampling distribution, such as the precision of the estimator – 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

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

$$= \boldsymbol{\beta}_0 + (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\epsilon} \tag{2}$$

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\epsilon} \tag{3}$$

The variance-covariance matrix is given by:

$$\mathbb{E}[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^T] = \mathbb{E}[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\epsilon}((\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\epsilon})^T] \tag{4}$$

$$= \mathbb{E}[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}] \tag{5}$$

$$= \mathbb{E}[\mathbb{E}[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}|\boldsymbol{X}]] \tag{6}$$

$$= \mathbb{E}[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\,\mathbb{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T|\boldsymbol{X}]\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}] \tag{7}$$

Now we need to assume something about $\mathbb{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T|\boldsymbol{X}]$, in particular, we assume that

$$\mathbb{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T|\boldsymbol{X}] = \sigma_0^2\boldsymbol{I} \tag{8}$$

This assumption means: $\text{Var}[\epsilon_i] = \sigma_0^2$ for all $i = 1, \ldots, n$, and that $\mathbb{E}[\epsilon_i\epsilon_j] = 0$ for all $i \neq j$. In words, the error term across all observations have the same variance σ_0^2, and the covariance of the error term across different observations is zero. That the observations are i.i.d. would imply $\mathbb{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T|\boldsymbol{X}] = \sigma_0^2\boldsymbol{I}$, but i.i.d is a stronger requirement.

When the error terms have identical variance across observations, we are said to be imposing the *homoskedasticity* assumption (as opposed to heteroskedasticity).

$$\mathbb{E}[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^T] = \mathbb{E}[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\sigma_0^2\boldsymbol{I}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}] \tag{9}$$

$$= \sigma_0^2\,\mathbb{E}[(\boldsymbol{X}^T\boldsymbol{X})^{-1}] \tag{10}$$

In the case where $\boldsymbol{X}$ is non-stochastic, then $\text{Var}(\hat{\boldsymbol{\beta}}) = \sigma_0^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$. When $\boldsymbol{X}$ is stochastic, we simply estimate $\mathbb{E}[(\boldsymbol{X}^T\boldsymbol{X})^{-1}]$ as $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$. Alternatively, we are not interested in the stochastic process governing $\boldsymbol{X}$, and so we calculate the variance-covariance matrix conditioning on $\boldsymbol{X}$. Therefore,

$$\text{Var}(\hat{\boldsymbol{\beta}}|\boldsymbol{X}) = \sigma_0^2(\boldsymbol{X}^T\boldsymbol{X})^{-1} \tag{11}$$

The precision of the OLS estimator is defined as the inverse of $\text{Var}(\hat{\boldsymbol{\beta}})$.

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

We can demonstrate this more rigorously using the FWL theorem.

Consider the regression model $\boldsymbol{y} = \boldsymbol{x}_1\beta_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{u}$. By the FWL theorem, the OLS estimator of β_1 in this regression model is equivalent to the corresponding OLS estimator in $\boldsymbol{M}_2\boldsymbol{y} = \boldsymbol{M}_2\boldsymbol{x}_1\beta_1 + \boldsymbol{\epsilon}$, where $\boldsymbol{M}_2 = \boldsymbol{I} - \boldsymbol{P}_2$, and $\boldsymbol{P}_2 = \boldsymbol{X}_2(\boldsymbol{X}_2^T\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2^T$.

Therefore,

$$\text{Var}(\hat{\beta}_1) = \sigma_0^2((\boldsymbol{M}_2\boldsymbol{x}_1)^T(\boldsymbol{M}_2\boldsymbol{x}_1))^{-1} \tag{12}$$

$$= \frac{\sigma_0^2}{(\boldsymbol{M}_2\boldsymbol{x}_1)^T(\boldsymbol{M}_2\boldsymbol{x}_1)} \tag{13}$$

Therefore, $\hat{\beta}_1$ becomes more precise when the squared Euclidean length of the vector $\boldsymbol{M}_2\boldsymbol{x}_1$ is large.

The squared Euclidean length of the vector $\boldsymbol{M}_2\boldsymbol{x}_1$ is just the sum of squared residuals from the regression:

$$\boldsymbol{x}_1 = \boldsymbol{X}_2\boldsymbol{\alpha} + \text{residuals} \tag{14}$$

Thus, when $\boldsymbol{x}_1$ can be explained by $\boldsymbol{X}_2$, $(\boldsymbol{M}_2\boldsymbol{x}_1)^T(\boldsymbol{M}_2\boldsymbol{x}_1)$ becomes small, and consequently, $\hat{\beta}_1$ becomes less precise. The intuition is that, we cannot estimate the effect of a regressor on the dependent variable well if that regressor can be explained by the other regressors – this regressor does not add new orthogonal information. For instance, if the dependent variable is sales across customers, and the explanatory variables are demographics regressors, we may find that it is very hard to disentangle the effects between two highly correlated regressors, say income and education level.

2. Estimating the variance of the error terms

How do we estimate σ_0^2? Using the method of moments, and the moment condition that $\mathbb{E}[\epsilon_i^2] = \sigma_0^2$, one estimator is $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n \hat{\epsilon}_i^2$, where $\hat{\boldsymbol{\epsilon}} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}$ is the fitted residual.

It can be shown that $\hat{\sigma}^2$ is a consistent estimator of σ_0, however $\hat{\sigma}^2$ is a biased estimator. In particular, $\mathbb{E}[\hat{\sigma}^2] = \frac{n-k}{n}\sigma_0^2$, and therefore, an unbiased estimator of σ_0 would be:

$$s^2 = \frac{1}{n-k}\sum_{i=1}^n \hat{\epsilon}_i \tag{15}$$

To see this, note that:

$$\hat{\boldsymbol{\epsilon}} = \boldsymbol{M}\boldsymbol{y} \tag{16}$$

$$= \boldsymbol{M}(\boldsymbol{X}\boldsymbol{\beta}_0 + \boldsymbol{\epsilon}) \tag{17}$$

$$= \boldsymbol{M}\boldsymbol{\epsilon} \tag{18}$$

Now consider the estimator $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n \hat{\epsilon}_i^2 = \frac{1}{n}\text{Trace}(\hat{\boldsymbol{\epsilon}}\hat{\boldsymbol{\epsilon}}^T)$, where the Trace of a matrix is just the sum of the diagonal elements. Now we want to compute $\mathbb{E}[\hat{\sigma}^2]$. Assume that $\boldsymbol{X}$ is non-stochastic (or we implicitly condition on $\boldsymbol{X}$, i.e. $\mathbb{E}[\hat{\sigma}^2|\boldsymbol{X}]$).

$$\mathbb{E}[\text{Trace}(\hat{\boldsymbol{\epsilon}}\hat{\boldsymbol{\epsilon}}^T)] = \mathbb{E}[\text{Trace}(\boldsymbol{M}\boldsymbol{\epsilon}(\boldsymbol{M}\boldsymbol{\epsilon})^T)] \tag{19}$$

$$= \text{Trace}(\mathbb{E}[\boldsymbol{M}\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T\boldsymbol{M}]) \tag{20}$$

$$= \text{Trace}(\boldsymbol{M}\,\mathbb{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T]\boldsymbol{M}) \tag{21}$$

$$= \sigma_0^2\text{Trace}(\boldsymbol{M}\boldsymbol{M}) \tag{22}$$

$$= \sigma_0^2\text{Trace}(\boldsymbol{M}) \tag{23}$$

Now:

$$\text{Trace}(\boldsymbol{M}) = \text{Trace}(\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T) \tag{24}$$

$$= \text{Trace}(\boldsymbol{I}) - \text{Trace}(\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T) \tag{25}$$

$$= n - \text{Trace}(\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}) \tag{26}$$

$$= n - k \tag{27}$$

Therefore $\mathbb{E}[\hat{\sigma}^2] = \mathbb{E}[\frac{1}{n}\sum_{i=1}^n \hat{\epsilon}_i^2] = \frac{n-k}{n}\sigma_0^2$. An unbiased estimator of σ_0^2 is:

$$s^2 = \frac{1}{n-k}\sum_{i=1}^n \hat{\epsilon}_i \tag{28}$$

To see the intuition behind the factor $\frac{1}{n-k}$. Suppose there are as many regressors as there are observations, then the residuals are always zero, and we would not have obtained any information on σ_0^2. Of course this is an extreme case. In practice we confine ourselves to the case $k < n$. The very fact that we choose $\hat{\boldsymbol{\beta}}$ in such a way that the sum of squared residuals is minimized is the cause of the fact that the squared residuals $\hat{\boldsymbol{\epsilon}}$ are smaller (on average) than the squared error terms $\boldsymbol{\epsilon}$.

2.1. Leverage

If we denote the i-th diagonal of $\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$ as h_i, then h_i is also known as the *leverage* of the observation i.

$$\text{Var}(\hat{\epsilon}_i) = \mathbb{E}[\hat{\epsilon}_i{}^2] \tag{29}$$

$$= \mathbb{E}[(\hat{\boldsymbol{\epsilon}}\hat{\boldsymbol{\epsilon}}^T)_{ii}] \tag{30}$$

$$= (\boldsymbol{M}\,\mathbb{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T]\boldsymbol{M})_{ii} \tag{31}$$

$$= \sigma_0^2\boldsymbol{M}_{ii} \tag{32}$$

$$= \sigma_0^2(1 - h_i) \tag{33}$$

Therefore, when an observation has a high leverage, $1 - h_i$ is small, and the residual $\hat{\epsilon}_i$ is close to zero. Even though it must be the case that $\sum_i \hat{\epsilon}_i = 0$, and $\mathbb{E}[\hat{\epsilon}_i] = 0$, not all residuals are equally small – some observations have smaller residuals – OLS prioritize these observations (those with high leverage). By the way, can you show that $\sum_i \hat{\epsilon}_i = 0$, and $\mathbb{E}[\hat{\epsilon}_i] = 0$?

Leverage is a measure of how influential an observation is – how much the OLS estimate changes when the point is removed. Data points with high leverage or influence force the regression line to be close to the point. Leverage is not to be confused with the notion of outliers.

3. Heteroskedasticity-consistent covariance matrix estimator

Recall that the variance-covariance matrix of the OLS estimator $\boldsymbol{\beta}$ is:

$$\mathbb{E}[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^T] = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\,\mathbb{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T]\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1} \tag{34}$$

(Either treat $\boldsymbol{X}$ to be fixed, or implicitly condition on $\boldsymbol{X}$).

Heteroskedasticity means:

$$\mathbb{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T] = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{bmatrix}$$

We have reasons to believe that some observations have more noise than others – $\text{Var}(\epsilon_i)$ differs across i.

Heteroskedasticity does **not** cause OLS to be biased, but the estimator for the variance-covariance matrix of OLS would be biased and wrong. Therefore, we still get the same estimate regardless of whether we assume heteroskedasticity or not, but our inference (hypothesis test, confidence interval, etc) would be wrong.

In one of the most cited paper of all time in Economics, Halber White (1980) introduced the following estimator:

$$(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\hat{\Sigma}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1} \tag{35}$$

Where:

$$\hat{\Sigma} = \begin{bmatrix} \hat{\epsilon}_1^2 & 0 & \cdots & 0 \\ 0 & \hat{\epsilon}_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \hat{\epsilon}_n^2 \end{bmatrix}$$

White shows that this is a consistent estimator for the variance-covariance matrix of OLS estimators, i.e. $\mathbb{E}[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^T]$. This is the estimator that Stata implements by default with the command `reg y x, robust`.

The intuition behind this estimator can be seen by recalling Equation 33: $\text{Var}(\epsilon_i) = \mathbb{E}[\hat{\epsilon_i}^2] = \sigma_0^2(1 - h_i)$. Asymptotically when n is large, we can ignore the leverage factor. There are other heteroskedastic-consistent variance-covariance estimator, such as using:

$$\hat{\Sigma} = \begin{bmatrix} \hat{\epsilon}_1^2(1-h_1) & 0 & \cdots & 0 \\ 0 & \hat{\epsilon}_2^2(1-h_2) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \hat{\epsilon}_n^2(1-h_n) \end{bmatrix}$$

White's estimator is the standard estimator. None of these estimators have any finite-sample guarantee (unbiasedness), but asymptotically they are all equivalent and consistent. Therefore, there is no formal mathematical reason to think that the latter estimator is better.

3.1. Serial correlation

In the presence of serial correlation,

$$\mathbb{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T] = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{12} & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{bmatrix} \tag{36}$$

The off-diagonals are non-zero. The error terms are correlated across observations. This is quite common in time-series (but not in cross-sectional data). In Equation 36 above, we have both serial correlation and heteroskedasticity.

Serial correlation does **not** affect the unbiasedness of OLS estimators. Similar to heteroskedasticity, serial correlation results in incorrect confidence intervals and hypothesis tests.

Serial correlation is usually corrected by assuming that the serial correlation follows a specific form: $\epsilon_t = \rho\epsilon_{t-1} + u_t$. This is known as the AutoRegressive(1) errors. We can test for the presence of this kind of serial correlation using the Durbin-Watson test. Correcting for serial correlation involves differencing: we regress the difference $y_t - \rho y_{t-1}$ on the difference $x_t - \rho x_{t-1}$. The parameter ρ can be estimated consistently using OLS residuals, by regressing $\hat{\epsilon}_t$ on $\hat{\epsilon}_{t-1}$.

4. Hypothesis testing and confidence interval involving OLS estimators

Suppose that $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}_0 + \boldsymbol{u}$. Assume the following: (1) Exogeneity, $\mathbb{E}[\boldsymbol{u}|\boldsymbol{X}] = 0$, (2) No perfect multicollinearity, $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ exists, (3) Homoskedastic and no serial correlation, $\mathbb{E}[\boldsymbol{u}\boldsymbol{u}^T] = \sigma_0^2\boldsymbol{I}$, (4) Normality, $\boldsymbol{u}|\boldsymbol{X} \sim \mathcal{N}(0, \sigma_0^2\boldsymbol{I})$.

These assumptions are collectively called the Classical Linear Regression Model.

Then the OLS estimator $\hat{\boldsymbol{\beta}}$ satisfies:

$$\hat{\boldsymbol{\beta}}|\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\beta}_0, \sigma_0^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}) \tag{37}$$

This result can be derived by using the fact that a linear combination of Normal random variables is a Normal random variable, and that OLS takes linear combination of the Normal error terms. Specifically if $\boldsymbol{u} \sim \mathcal{N}(\mu, \Sigma)$, then $A + Bu \sim \mathcal{N}(A + B\mu, B\Sigma B^T)$.

Therefore, to construct hypothesis tests and confidence intervals for OLS estimates, we can simply applied what we have learned in the last few classes here.

If we are unwilling to assume Normal error terms, then there are two alternative approaches: (1) bootstrapping, (2) asymptotics.

Asymptotic sampling distribution. Under certain condition,

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \sigma_0^2\left(\frac{\boldsymbol{X}^T\boldsymbol{X}}{n}\right)^{-1}\right) \tag{38}$$

As such, the sampling distribution of $\hat{\boldsymbol{\beta}}$ for large n can be approximated as:

$$\hat{\boldsymbol{\beta}} \sim \mathcal{N}(\boldsymbol{\beta}_0, \sigma_0^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}) \tag{39}$$

5. Summary

(i) Exogeneity alone guarantees unbiasedness. Exogeneity can be violated under many circumstances – whenever the regressor is correlated with the error term.

For instance, if we regress weekly demand on weekly price, we expect high price to cause lower demand, so the coefficient is negative. But often this regression gives us positive price coefficient! This is because price is an endogenous variable. If price is set randomly (experimentation or A/B testing), then it would be exogenous. But price is set strategically by firms. When an unexpected shock drives demand higher, the firm responds by changing price, therefore, there is a correlation between price and the error term (demand shocks).

This is the foremost concern in any empirical research. The branch of statistics/econometrics dealing with this concern is called *causal inference*. Tools that fall under causal inference include (1) Instrumental Variable approach, (2) Difference-in-difference, (3) Regression discontinuity, (4) Propensity score matching, (5) experimentation and A/B testing.

(ii) Heteroskedasticity and serial correlation causes incorrect statistical inference (wrong formula for calculating the variance-covariance matrix of OLS estimator).

(iii) Multicollinearity increases the variance of OLS.

(iv) Under-specification (omission of relevant variables) causes bias since the exogeneity condition is violated. However over-specification (inclusion of irrelevant variables) does *not* cause bias. However it does increase the variance of OLS. Specifically, over-specification means the true data-generating process is $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}$, but we estimate $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\alpha} + \boldsymbol{\epsilon}$. It is straightforward to see why OLS is still unbiased.