LECTURE 14: STATISTICAL PROPERTIES OF ORDINARY LEAST SQUARES

MECO 7312.
INSTRUCTOR: DR. KHAI CHIONG
DECEMBER 6, 2023

1. Multivariate linear regression

Let $(y_i, x_{i1}, x_{i2}, \ldots, x_{iK}, \epsilon_i)_{i=1}^n$ be realization from some joint distribution such that the following holds:

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{iK} + \epsilon_i, \quad \text{for } i = 1, \ldots, n \tag{1}$$

Using matrix algebra:

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

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

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

$$\boldsymbol{\beta}_0 = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_K \end{bmatrix}$$

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}_0 + \boldsymbol{\epsilon}$$

$(y_i, x_{i1}, x_{i2}, \ldots, x_{iK})_{i=1}^n$ is observed, but $(\epsilon_i)_{i=1}^n$ is unobserved. The goal is to estimate $\boldsymbol{\beta}_0$, which is unknown. The OLS estimator for $\boldsymbol{\beta}_0$ is $\hat{\boldsymbol{\beta}}_{OLS} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$, which is derived by assuming $\boldsymbol{X}^T\epsilon = \boldsymbol{0}$.

What are the statistical properties of OLS estimator? We know from the last lecture that OLS is unbiased, as summarized in the theorem below.

Theorem 1. *Unbiasedness of OLS*

1.) *Data-generating process is:* $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}_0 + \boldsymbol{\epsilon}$, *where* $\boldsymbol{X}$ *and* $\boldsymbol{\epsilon}$ *are stochastic and follow some probability distributions.*

2.) *The columns of* $\boldsymbol{X}$ *is linearly independent. In another words, the rank of* $\boldsymbol{X}$ *is* K *in all realizations of* $\boldsymbol{X}$.

3.) *Exogeneity (zero conditional mean):* $\mathbb{E}[\boldsymbol{\epsilon}|\boldsymbol{X}] = \boldsymbol{0}$.

(1), (2) and (3) together imply that $\mathbb{E}[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}_0$

Besides unbiasedness, we also want to know the sampling distribution of OLS. For example, what is the variance of OLS? Is OLS asymptotically Normal? The goal is to perform statistical inference (hypothesis testing and confidence interval).

2. OLS covariance matrix

What is the sampling distribution of $\hat{\boldsymbol{\beta}}$? We must start from the data-generating process, which induces the sampling distribution. The data generating process is $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}_0 + \boldsymbol{\epsilon}$, where $(\boldsymbol{X}, \boldsymbol{\epsilon})$ follow some probability distributions. Sometimes it is easier to think of $\boldsymbol{X}$ as fixed and nonstochastic, and only $\boldsymbol{\epsilon}$ is random. For example, in scientific experiments, we sometimes treat $\boldsymbol{X}$ as being fixed. In either case, all OLS-related derivations are similar.

Now we define the variance-covariance matrix of $\hat{\boldsymbol{\beta}}$ as a $K \times K$ matrix Σ such that $\Sigma_{ii} = \text{Var}(\hat{\beta}_i) = \text{Cov}(\hat{\beta}_i, \hat{\beta}_i)$, and $\Sigma_{ij} = \text{Cov}(\hat{\beta}_i, \hat{\beta}_j)$. The variance-covariance matrix can be written as $\mathbb{E}[(\hat{\boldsymbol{\beta}} - \mathbb{E}[\hat{\boldsymbol{\beta}}])(\hat{\boldsymbol{\beta}} - \mathbb{E}[\hat{\boldsymbol{\beta}}])^T] = \mathbb{E}[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^T]$.

$(\hat{\boldsymbol{\beta}} - \mathbb{E}[\hat{\boldsymbol{\beta}}])$ is a $K \times 1$ vector, while $(\hat{\boldsymbol{\beta}} - \mathbb{E}[\hat{\boldsymbol{\beta}}])^T$ is a $1 \times K$ vector, the product of which is a $K \times K$ matrix.

Now we can derive the variance-covariance matrix of the OLS estimator. Recall that

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

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

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

$$\mathbb{E}[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^T] = \mathbb{E}[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\epsilon}((\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\epsilon})^T] \tag{5}$$

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

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

$$= \mathbb{E}[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\,\mathbb{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T|\boldsymbol{X}]\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}] \tag{8}$$

We have used the fact that $(A^{-1})^T = (A^T)^{-1}$ for a square invertible matrix A, and $(AB)^T = B^TA^T$. Now we need to assume something about $\mathbb{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T|\boldsymbol{X}]$, in particular, we assume that

$$\mathbb{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T|\boldsymbol{X}] = \sigma_0^2\boldsymbol{I} \tag{9}$$

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).

$$\mathbb{E}[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^T] = \mathbb{E}[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\sigma_0^2\boldsymbol{I}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}] \tag{10}$$

$$= \sigma_0^2\,\mathbb{E}[(\boldsymbol{X}^T\boldsymbol{X})^{-1}] \tag{11}$$

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,

$$\text{Var}(\hat{\boldsymbol{\beta}}|\boldsymbol{X}) = \sigma_0^2(\boldsymbol{X}^T\boldsymbol{X})^{-1} \tag{12}$$

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.

3. Estimating the variance of the error terms

How do we estimate σ_0^2? Let's recall all the assumptions we have imposed so far:

(i) $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}_0 + \boldsymbol{\epsilon}$,

(ii) $\mathbb{E}[\boldsymbol{\epsilon}|\boldsymbol{X}] = \boldsymbol{0}$

(iii) $\mathbb{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T|\boldsymbol{X}] = \sigma_0^2\boldsymbol{I}$

These assumptions implied that $\text{Var}(\epsilon_i) = \mathbb{E}[\epsilon_i^2] = \sigma_0^2$ for $i = 1, \ldots, n$. If ϵ_i is i.i.d, then we can estimate σ_0^2 as the sample variance estimator $s^2 = \frac{1}{n}\sum_{i=1}^n \epsilon_i^2$. However, we don't observe ϵ_i. Consider replacing ϵ_i with $\hat{\epsilon}_i$. That is, consider the estimator $s^2 = \frac{1}{n}\sum_{i=1}^n \hat{\epsilon}_i^2$, where $\hat{\boldsymbol{\epsilon}} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}$ is the residual.

It turns out that s^2 is a biased estimator of σ_0^2, but we are not too far off. In particular, we can show that $\mathbb{E}[s^2] = \frac{n-K}{n}\sigma_0^2$. This is true without the i.i.d assumption on ϵ_i. Therefore, an unbiased estimator of σ_0 would be:

$$\hat{\sigma}^2 = \frac{1}{n-K}\sum_{i=1}^n \hat{\epsilon}_i^2 \tag{13}$$

The factor $\frac{1}{n-K}$ appears because we cannot simply estimate the true error term ϵ_i with the residual $\hat{\epsilon}_i$. The residual is an underestimate of ϵ_i because OLS tries to minimize the the sum of squared residuals. Therefore, we have to inflate $\frac{1}{n}\sum_{i=1}^n \hat{\epsilon}_i^2$ with the factor $\frac{n}{n-K} > 1$ to achieve an unbiased estimate. When K is large relative to n, this inflation factor is large: when we have many regressors, we can fit the dependent variable very well, leaving very little for the residuals.

Now we use R or Python to check that built-in OLS estimators use exactly the formula $(\frac{1}{n-K}\sum_{i=1}^n \hat{\epsilon}_i^2)(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ to calculate the standard errors of the OLS estimates.

4. Heteroskedasticity-consistent covariance matrix estimator

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

$$\mathbb{E}[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^T] = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\,\mathbb{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T]\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1} \tag{14}$$

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

To obtain a simplified expression of $\mathbb{E}[(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)^T]$, we made the homoskedastic assumption: $\mathbb{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T|\boldsymbol{X}] = \sigma_0^2\boldsymbol{I}$. Now we want to relax this assumption of constant variance across observations to allow for the fact that some observations have more noise than others – $\text{Var}(\epsilon_i)$ differs across i.

When this assumption is violated, we say that the error terms are heteroskedastic, or there is heteroskedasticity. Heteroskedasticity means:

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

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.

Halbert White (1980) proposed a heteroskedastic-consistent estimator of the variance-covariance matrix. That is, an estimator S^2 that converges in probability to $\mathbb{E}[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^T|\boldsymbol{X}] = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\,\mathbb{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T|\boldsymbol{X}]\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}$, without making the assumption $\mathbb{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T|\boldsymbol{X}] = \sigma_0^2\boldsymbol{I}$.

The idea is simple, instead of assuming $\mathbb{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T|\boldsymbol{X}] = \sigma_0^2\boldsymbol{I}$, why not just estimate $\mathbb{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T|\boldsymbol{X}]$? Now in the presence of heteroskedasticity, $\mathbb{E}[\epsilon_i^2] = \sigma_i^2$ for $i = 1, \ldots, n$, where σ_i^2 is unknown, how about we estimate $\mathbb{E}[\epsilon_i^2]$ with $\hat{\epsilon}_i^2$? The White's heteroskedastic-consistent estimator of the variance-covariance matrix is:

$$S^2 = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\hat{\Sigma}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1} \tag{15}$$

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 under some conditions, this is a consistent estimator of the variance-covariance matrix of OLS estimators. There are other heteroskedastic-consistent variance-covariance estimator. The one implemented in Equation (15) is called HC0.[1] Stata uses HC1 below, where:

$$\hat{\Sigma} = \begin{bmatrix} (\frac{n}{n-K})\hat{\epsilon}_1^2 & 0 & \cdots & 0 \\ 0 & (\frac{n}{n-K})\hat{\epsilon}_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & (\frac{n}{n-K})\hat{\epsilon}_n^2 \end{bmatrix}$$

The intuition behind this estimator can be seen by recalling the previous section that $\mathbb{E}[\frac{1}{n}\sum_{i=1}^n \hat{\epsilon}_i^2] = \frac{n-K}{n}\sigma_0^2$, and therefore $\sigma_0^2 = \mathbb{E}[\frac{1}{n}\sum_{i=1}^n (\frac{n}{n-K})\hat{\epsilon}_i^2]$.

These estimators are consistent under heteroskedasticity, therefore for large n, they are all asymptotically equivalent. However none of these estimators have any finite-sample guarantee (unbiasedness).

The Breusch-Pagan test can be used to test for heteroskedasticity. First, we obtain the residuals from $\hat{\boldsymbol{e}} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}$. Then we run the auxiliary regression $\hat{\boldsymbol{e}}^2 = \boldsymbol{X}\boldsymbol{\gamma} + \boldsymbol{\eta}$. Under the null hypothesis of homoskedasticity, the test statistic nR^2 is asymptotically distributed as χ^2_{K-1}, where R^2 is the R-squared from the auxiliary regression.

4.1. Clustered standard errors

In empirical research, we often hear that "we clustered the standard errors at the level of counties, states, industries, etc." This means that the authors are assuming a block-diagonal structure for $\mathbb{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T]$.

Observations within the same group or cluster could have correlated error terms, whereas observations from different groups have uncorrelated error terms.

$$\mathbb{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T]_{ij} = \mathbb{E}[\epsilon_i\epsilon_j] = \begin{cases} 0 & \text{if } i \text{ and } j \text{ does not belong to the same cluster} \\ \sigma_{i,j}^2 & \text{if } i \text{ and } j \text{ belongs to the same cluster} \end{cases}$$

For example, with two groups:

[1] In the R Markdown that accompanies this lecture, we show how to implement heteroskedastic-robust standard errors. We verify that the estimator constructed here yields the same results as those implemented by existing packages.

$$\mathbb{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T] = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1n_1} & 0 & 0 & \cdots & 0 \\ \sigma_{12} & \sigma_2^2 & \cdots & \sigma_{2n_1} & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \sigma_{n_1 1}^2 & \sigma_{n_1 2}^2 & \cdots & \sigma_{n_1 n_1}^2 & 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & \sigma_{n_1+1}^2 & \sigma_{n_1+1,n_1+2} & \cdots & \sigma_{n_1+1,n} \\ 0 & 0 & \cdots & 0 & \sigma_{n_1+2,n_1+1} & \sigma_{n_1+2}^2 & \cdots & \sigma_{n_1+2,n} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & \sigma_{n,n_1+1} & \sigma_{n,n_1+2} & \cdots & \sigma_n^2 \end{bmatrix}$$

We can then estimate $\mathbb{E}[\epsilon_i \epsilon_j]$ using $\hat{\epsilon}_i \hat{\epsilon}_j$ obtained from OLS. As with heteroskedasticity, not using clustered standard errors when the DGP has correlated errors will not lead to biased estimates of OLS, but would rather lead to incorrect inference and standard errors.

4.2. Serial correlation

In the presence of serial correlation, the error terms are correlated across observations, i.e. $\mathbb{E}[\epsilon_i \epsilon_j] = \text{Cov}(\epsilon_i, \epsilon_j) \neq 0$. The off-diagonals of $\mathbb{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T]$ are non-zero. This is quite common in time-series (but not in cross-sectional data).

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 $\hat{\epsilon}_{t-1}$.

5. 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, $\boldsymbol{u}|\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{0}, \sigma_0^2 \boldsymbol{I})$.

These assumptions are collectively called the Classical Linear Regression Model.

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

$$\hat{\boldsymbol{\beta}}|\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\beta}_0, \sigma_0^2 (\boldsymbol{X}^T\boldsymbol{X})^{-1}) \tag{16}$$

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.[2] 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. Suppose we want to test whether one of the coefficients is zero, i.e. $H_0 : \beta_{0j} = 0$ versus $H_1 : \beta_{0j} \neq 0$. Under the null, we have an estimator that is Normally distributed as $\hat{\beta}_j \sim \mathcal{N}(0, \sigma_0^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}_{jj})$. If we know σ_0^2, then a t-test statistic $\frac{\hat{\beta}_j}{\sqrt{\sigma_0^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}_{jj}}}$ has a $\mathcal{N}(0,1)$ under the null. If we had to estimate σ_0^2, then it turns out that, our estimator $\hat{\sigma}^2 = \frac{1}{n-K}\sum_{i=1}^n \hat{\epsilon}_i^2$ has a Chi-squared distribution, and therefore under the null, $\frac{\hat{\beta}_j}{\sqrt{\hat{\sigma}^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}_{jj}}}$ has a Student t's distribution with $n-K$ degrees of freedom.[3]

If we are unwilling to assume Normal error terms, then there are two alternative approaches: (1) bootstrapping, (2) asymptotics. Asymptotic sampling distribution. Under the assumption that $\frac{\boldsymbol{X}^T\boldsymbol{X}}{n} \to_p Q$ as $n \to \infty$, where Q is a positive definite matrix,

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \sigma_0^2 Q^{-1}\right) \tag{17}$$

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

$$\hat{\boldsymbol{\beta}} \sim \mathcal{N}(\boldsymbol{\beta}_0, \sigma_0^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}) \tag{18}$$

6. Summary

In the companion Python notebook, we can run Monte Carlo simulation of a linear regression DGP to see the following point:

(i) Exogeneity alone guarantees unbiasedness.

(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.

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

[3] Recall that if $Z \sim \mathcal{N}(0,1)$ and if $S^2 \sim \chi^2_d$, then $\frac{Z}{\sqrt{S^2/d}}$ has a Student's t distribution with d degrees of freedom.

(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, but over-specification increases 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.