

LECTURE 14: STATISTICAL PROPERTIES OF ORDINARY LEAST SQUARES

MECO 7312.
INSTRUCTOR: DR. KHAI CHIONG
DECEMBER 1, 2021

1. Multivariate linear regression

Recall the linear regression model:

$$(1) \quad y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{iK} + \epsilon_i$$

Using Matrix Algebra:

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

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

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

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

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

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

What are the statistical properties of OLS estimator? We want OLS to be unbiased, consistent, efficient (achieves lowest variance). We also want to know the sampling

distribution of OLS. For example, what is the variance of OLS? Is OLS asymptotically Normal? Such information will allow us to perform statistical inference involving OLS (hypothesis testing and confidence interval).

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

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

3.) *Exogeneity (zero conditional mean): $\mathbb{E}[\boldsymbol{\epsilon}|\mathbf{X}] = \mathbf{0}$. A sufficient condition is the independence of $\boldsymbol{\epsilon}$ and \mathbf{X} .*

\implies *OLS is unbiased $\mathbb{E}[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}_0$*

2. OLS covariance matrix

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

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 $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}_0 + \boldsymbol{\epsilon}$, where either \mathbf{X} is nonstochastic (fixed) and $\boldsymbol{\epsilon}$ is random, or $(\mathbf{X}, \boldsymbol{\epsilon})$ are both random. In scientific experiments, we can treat \mathbf{X} as being fixed. In either case, all OLS-related derivations are similar.

Previously, we have seen that under the *exogeneity* condition $\mathbb{E}[\boldsymbol{\epsilon}|\mathbf{X}] = \mathbf{0}$, 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.

Specifically, 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

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

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

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

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

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

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

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

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

$$(9) \quad \mathbb{E}[\boldsymbol{\epsilon} \boldsymbol{\epsilon}^T | \mathbf{X}] = \sigma_0^2 \mathbf{I}$$

This assumption means: $\text{Var}[\epsilon_i] = \sigma_0^2$ for all $i = 1, \dots, n$, 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 $\mathbb{E}[\boldsymbol{\epsilon} \boldsymbol{\epsilon}^T | \mathbf{X}] = \sigma_0^2 \mathbf{I}$, 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).

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

$$(11) \quad = \sigma_0^2 \mathbb{E}[(\mathbf{X}^T \mathbf{X})^{-1}]$$

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

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

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) $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}_0 + \boldsymbol{\epsilon}$,
- (ii) $\mathbb{E}[\mathbf{X}|\boldsymbol{\epsilon}] = \mathbf{0}$
- (iii) $\mathbb{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T|\mathbf{X}] = \sigma_0^2\mathbf{I}$

These assumptions implied that $\text{Var}(\epsilon_i) = \mathbb{E}[\epsilon_i^2] = \sigma_0^2$ for $i = 1, \dots, n$. As such, a reasonable estimator for σ_0^2 would be to estimate each $\mathbb{E}[\epsilon_i^2]$ with $\hat{\epsilon}_i^2$, for each $i = 1, \dots, n$ (without imposing i.i.d). Therefore, consider the estimator $s^2 = \frac{1}{n} \sum_{i=1}^n \hat{\epsilon}_i^2$, where $\hat{\boldsymbol{\epsilon}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}$ is the residual.

It turns out that s^2 is a biased estimator of σ_0^2 , but we are not too far off the mark. In particular, $\mathbb{E}[s^2] = \frac{n-K}{n}\sigma_0^2$, and therefore, an unbiased estimator of σ_0 would be:

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

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, 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 to check that R's built-in OLS estimator (`lm`) uses exactly the formula $(\frac{1}{n-K} \sum_{i=1}^n \hat{\epsilon}_i^2)(\mathbf{X}^T\mathbf{X})^{-1}$ 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 $\hat{\boldsymbol{\beta}}$ is:

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

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

To obtain a simplified expression of $\mathbb{E}[(\hat{\beta} - \beta_0)(\hat{\beta} - \beta_0)^T]$, we made the homoskedastic assumption: $\mathbb{E}[\epsilon\epsilon^T | \mathbf{X}] = \sigma_0^2 \mathbf{I}$. The assumption of constant variance across observations is not plausible. It is usually the case 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:

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

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) proposed a heteroskedastic-consistent estimator of the variance-covariance matrix. That is, an estimator S^2 that converges in probability to $\mathbb{E}[(\hat{\beta} - \beta_0)(\hat{\beta} - \beta_0)^T] = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbb{E}[\epsilon\epsilon^T] \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1}$, without making the assumption $\mathbb{E}[\epsilon\epsilon^T | \mathbf{X}] = \sigma_0^2 \mathbf{I}$.

The idea is simple, instead of assuming $\mathbb{E}[\epsilon\epsilon^T | \mathbf{X}] = \sigma_0^2 \mathbf{I}$, why not just estimate $\mathbb{E}[\epsilon\epsilon^T]$?

Now in the presence of heteroskedasticity, $\mathbb{E}[\epsilon_i^2] = \sigma_i^2$ for $i = 1, \dots, n$, where σ_i^2 is unknown. How about we estimate $\mathbb{E}[\epsilon_i^2]$ by $\hat{\epsilon}_i^2$? The White's heteroskedastic-consistent estimator of the variance-covariance matrix is:

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

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. This is the estimator that Stata

implements by default with the command `reg y x, robust`. 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.

There are other heteroskedastic-consistent variance-covariance estimator. The one implemented in 15 is called HC0. Stata uses HC1, where:

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

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

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 $\hat{\epsilon} = \mathbf{y} - \mathbf{X}\hat{\beta}$. Then we run the auxiliary regression $\hat{\epsilon}^2 = \mathbf{X}\boldsymbol{\gamma} + \boldsymbol{\eta}$. Under the null hypothesis of homoskedasticity, the test statistic nR^2 is asymptotically distributed as χ_{K-1}^2 , 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 could have correlated error terms, whereas observations from different groups have uncorrelated error terms.

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

For example, with two groups:

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

We can then estimate $\mathbb{E}[\epsilon_i \epsilon_j]$ using $\hat{\epsilon}_i \hat{\epsilon}_j$ obtained from OLS.

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

These assumptions are collectively called the Classical Linear Regression Model.

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

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

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 $\mathbf{u} \sim \mathcal{N}(\mu, \Sigma)$, 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,

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

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

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

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 $\hat{\beta}_j \sim \mathcal{N}(0, \sigma_0^2(\mathbf{X}^T \mathbf{X})_{jj}^{-1})$. If we know σ_0^2 , then a t -test statistic $\frac{\hat{\beta}_j}{\sqrt{\sigma_0^2(\mathbf{X}^T \mathbf{X})_{jj}^{-1}}}$ 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, $\frac{\hat{\beta}_j}{\sqrt{\hat{\sigma}^2(\mathbf{X}^T \mathbf{X})_{jj}^{-1}}}$ has a Student t 's distribution with $n - K$ degrees of freedom. Recall that if $Z \sim \mathcal{N}(0, 1)$ and if $S^2 \sim \chi_d^2$, then $\frac{Z}{\sqrt{S^2/d}}$ has a Student's t distribution with d degrees of freedom.

6. Summary

- (i) Exogeneity alone guarantees unbiasedness. Exogeneity can be violated under many circumstances – whenever the regressor is correlated with the error term.
- (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 $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$, but we estimate $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\alpha} + \boldsymbol{\epsilon}$. It is straightforward to see why OLS is still unbiased.