# LECTURE 15: STATISTICAL PROPERTIES OF ORDINARY LEAST SQUARES

MECO 7312.
INSTRUCTOR: DR. KHAI CHIONG
NOVEMBER 26, 2019

## 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.

Specifically, we define the variance-covariance matrix of $\hat{\boldsymbol{\beta}}$ as a $K \times K$ matrix $\Sigma$ 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

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

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

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

$$\mathbb{E}[(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)^T] = \mathbb{E}[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\epsilon}((\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\epsilon})^T] \tag{4}$$
$$= \mathbb{E}[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}] \tag{5}$$
$$= \mathbb{E}[\mathbb{E}[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}|\boldsymbol{X}]] \tag{6}$$
$$= \mathbb{E}[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\,\mathbb{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T|\boldsymbol{X}]\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}] \tag{7}$$

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 $\mathbb{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T|\boldsymbol{X}]$, 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 $\sigma_0^2$, 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 $\beta_1$ 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 $\sigma_0^2$? Let's recall all the assumptions we have imposed so far:

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

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

(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 $i = 1, \ldots, n$. As such, a reasonable estimator for $\sigma_0^2$ 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 $\hat{\boldsymbol{\epsilon}} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}$ is the residual.

It turns out that $s^2$ is a biased estimator of $\sigma_0^2$, 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 $\sigma_0$ would be:

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

To show 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 $s^2 = \frac{1}{n}\sum_{i=1}^n \hat{\epsilon}_i^2 = \frac{1}{n}\text{Trace}(\hat{\boldsymbol{\epsilon}}\hat{\boldsymbol{\epsilon}}^T)$, 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:

$$\text{Trace}(\boldsymbol{M}) = \text{Trace}(\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T) \tag{24}$$
$$= \text{Trace}(\boldsymbol{I}) - \text{Trace}(\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T) \tag{25}$$
$$= n - \text{Trace}(\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}) \tag{26}$$
$$= n - K \tag{27}$$

We used the fact that the trace operator is invariant under cyclic permutations. That is, $\text{Trace}(AB) = \text{Trace}(BA)$, and $\text{Trace}(ABC) = \text{Trace}(CAB) = \text{Trace}(BCA)$.

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

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

The factor $\frac{1}{n-K}$ appears because we cannot simply estimate the true error term $\epsilon_i$ with the residual $\hat{\epsilon}_i$. The residual is an underestimate of $\epsilon_i$ 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 $(\frac{1}{n-K}\sum_{i=1}^{n}\hat{\epsilon}_i^2)(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ to calculate the standard errors of the OLS estimates.

### 2.1. Leverage

We now know that $\mathbb{E}[\frac{1}{n}\sum_{i=1}^{n}\hat{\epsilon}_i^2] = \frac{n-k}{n}\sigma_0^2 = \frac{1}{n}\sigma_0^2\text{Trace}(\boldsymbol{M})$. What about $\mathbb{E}[\hat{\epsilon}_i^2]$? It turns out that $\text{Var}(\hat{\epsilon}_i) = \mathbb{E}[\hat{\epsilon_i}^2] = \sigma_0^2(1-h_i)$, where $h_i$ is called the leverage of observation $i$. The higher the leverage of observation $i$, the smaller $\text{Var}(\hat{\epsilon_i}^2)$ it is from $\sigma_0^2$.

Hence, 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 $\sum_{i=1}^{n}\hat{\epsilon}_i = 0$, not all residuals are equally small – some observations have smaller residuals – OLS prioritize these observations (those with high leverage).

Now the leverage $h_i$ is just the $i$-th diagonal of the projection matrix $\boldsymbol{P}_X = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$.

$$\begin{aligned}
\text{Var}(\hat{\epsilon}_i) &= \mathbb{E}[\hat{\epsilon_i}^2] && (29)\\
&= \mathbb{E}[(\hat{\boldsymbol{\epsilon}}\hat{\boldsymbol{\epsilon}}^T)_{ii}] && (30)\\
&= (\boldsymbol{M}\,\mathbb{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T]\boldsymbol{M})_{ii} && (31)\\
&= \sigma_0^2\boldsymbol{M}_{ii} && (32)\\
&= \sigma_0^2(1-h_i) && (33)
\end{aligned}$$

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 $\boldsymbol{X}$).

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: $\mathbb{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T|\boldsymbol{X}] = \sigma_0^2\boldsymbol{I}$. The assumption of constant variance across observations is not plausible. It is usually the case that some observations have more noise than others – $\mathrm{Var}(\epsilon_i)$ differs across $i$.

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

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

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

Now in the presence of heteroskedasticity, $\mathbb{E}[\epsilon_i^2] = \sigma_i^2$ for $i = 1, \ldots, n$, where $\sigma_i^2$ 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:

$$S^2 = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\hat{\Sigma}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1} \tag{35}$$

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 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 35 is called HC0. Stata uses HC1, 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 $\sigma_0^2 = \frac{n}{n-K}\,\mathbb{E}[\frac{1}{n}\sum_{i=1}^n \hat{\epsilon}_i^2]$.

HC2 is[1]:

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

All 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).

[1] Using derivations from the previous section, the intuition behind this estimator is: $\mathbb{E}[\hat{\epsilon_i}^2] = \sigma_0^2(1-h_i)$, and therefore $\sigma_0^2 = \frac{1}{(1-h_i)}\,\mathbb{E}[\hat{\epsilon_i}^2]$

### 3.1. Testing for heteroskedasticity

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 $\chi^2_{K-1}$, where $R^2$ is the R-squared from the auxiliary regression.

### 3.2. 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.

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

For example, with two groups:

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

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

In Stata: `vce(cluster clustvar)`. Where `clustvar` is a variable that identifies the groups in which the error terms are allowed to correlate.

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

These assumptions are collectively called the Classical Linear Regression Model.

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

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

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:

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

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 $\hat{\beta}_j \sim \mathcal{N}(0, \sigma_0^2(\boldsymbol{X}^T\boldsymbol{X})_{jj}^{-1})$. If we know $\sigma_0^2$, then a Wald's $t$-test statistic $\frac{\hat{\beta}_j}{\sqrt{\sigma_0^2(\boldsymbol{X}^T\boldsymbol{X})_{jj}^{-1}}}$ has a $\mathcal{N}(0,1)$ under the null. If we had to estimate $\sigma_0^2$, 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, $\frac{\hat{\beta}_j}{\sqrt{\hat{\sigma}^2(\boldsymbol{X}^T\boldsymbol{X})_{jj}^{-1}}}$ 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.

## 5. 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 $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}$, 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.