Economics 326
Methods of Empirical Research in Economics

Lecture 18: The asymptotic variance of OLS and heteroskedasticity

Vadim Marmer
University of British Columbia

March 24, 2009

Asymptotic normality

- In the previous lecture, we showed that when the data are iid and the regressors are exogenous:

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_i + U_i, \\ EU_i &= E\left(X_i U_i\right) = 0, \end{aligned}$$

the OLS estimator of β_1 is asymptotically normal:

$$\sqrt{n}\left(\hat{\beta}_{1,n} - \beta_1\right) \to_d N\left(0, V\right),$$

$$V = \frac{E\left(\left(X_i - EX_i\right)^2 U_i^2\right)}{\left(Var\left(X_i\right)\right)^2}.$$

- For the purpose of hypothesis testing, we need to obtain a consistent estimator of the asymptotic variance V:

$$\hat{V}_n \to_p V.$$

Homoskedastic errors

- Let's assume that the errors are homoskedastic:

$$E\left(U_i^2|X_i\right) = \sigma^2 \text{ for all } X_i\text{'s.}$$

- In this case, the asymptotic variance can be simplified using the Law of Iterated Expectation:

$$\begin{aligned}
E\left(\left(X_i - EX_i\right)^2 U_i^2\right) &= EE\left[\left(X_i - EX_i\right)^2 U_i^2|X_i\right] \\
&= E\left(\left(X_i - EX_i\right)^2 E\left[U_i^2|X_i\right]\right) \\
&= E\left(\left(X_i - EX_i\right)^2 \sigma^2\right) \\
&= \sigma^2 E\left(X_i - EX_i\right)^2 = \sigma^2 Var\left(X_i\right).
\end{aligned}$$

Homoskedastic errors

- Thus, when the errors are homoskedastic with $EU_i^2 = \sigma^2$,

$$V = \frac{E\left(\left(X_i - EX_i\right)^2 U_i^2\right)}{\left(Var\left(X_i\right)\right)^2} = \frac{\sigma^2 Var\left(X_i\right)}{\left(Var\left(X_i\right)\right)^2} = \frac{\sigma^2}{Var\left(X_i\right)}.$$

- Let $\hat{U}_i = Y_i - \hat{\beta}_{0,n} - \hat{\beta}_{1,n} X_i$, where $\hat{\beta}_{0,n}$ and $\hat{\beta}_{1,n}$ are the OLS estimators of β_0 and β_1.
- A consistent estimator for the asymptotic variance can be constructed by using the **Method of Moments**.

$$\begin{aligned}
\hat{\sigma}_n^2 &= \frac{1}{n}\sum_{i=1}^{n} \hat{U}_i^2, \\
\widehat{Var}\left(X_i\right) &= \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}_n\right)^2, \text{ and} \\
\hat{V}_n &= \frac{\hat{\sigma}_n^2}{\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}_n\right)^2}.
\end{aligned}$$

Homoskedastic errors

$$\hat{V}_n = \frac{\hat{\sigma}_n^2}{\frac{1}{n}\sum_{i=1}^n \left(X_i - \bar{X}_n\right)^2}, \qquad \hat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^n \hat{U}_i^2, \qquad \hat{U}_i = Y_i - \hat{\beta}_{0,n} - \hat{\beta}_{1,n}X_i.$$

- When proving the consistency of OLS in Lecture 16, we showed that

$$\frac{1}{n}\sum_{i=1}^n \left(X_i - \bar{X}_n\right)^2 \to_p Var\left(X_i\right),$$

and to establish $\hat{V}_n \to_p V$, we need to show that $\hat{\sigma}_n^2 \to_p \sigma^2$.

- Note that the LLN cannot be applied directly to

$$\frac{1}{n}\sum_{i=1}^n \hat{U}_i^2$$

because $\hat{U}_i$'s are not iid: they are dependent through $\hat{\beta}_{0,n}$ and $\hat{\beta}_{1,n}$.

Proof of $\hat{\sigma}_n^2 \to_p \sigma_n^2$

- First, write

$$\begin{aligned}\hat{U}_i &= Y_i - \hat{\beta}_{0,n} - \hat{\beta}_{1,n} X_i \\ &= (\beta_0 + \beta_1 X_i + U_i) - \hat{\beta}_{0,n} - \hat{\beta}_{1,n} X_i \\ &= U_i - \left(\hat{\beta}_{0,n} - \beta_0\right) - \left(\hat{\beta}_{1,n} - \beta_1\right) X_i .\end{aligned}$$

- Now,

$$\hat{\sigma}_n^2 = \frac{1}{n} \sum_{i=1}^{n} \hat{U}_i^2 = \frac{1}{n} \sum_{i=1}^{n} \left(U_i - \left(\hat{\beta}_{0,n} - \beta_0\right) - \left(\hat{\beta}_{1,n} - \beta_1\right) X_i\right)^2 .$$

Proof of $\hat{\sigma}_n^2 \to_p \sigma^2$

- We have

$$\begin{aligned}
\hat{\sigma}_n^2 &= \frac{1}{n}\sum_{i=1}^{n}\left(U_i - \left(\hat{\beta}_{0,n} - \beta_0\right) - \left(\hat{\beta}_{1,n} - \beta_1\right)X_i\right)^2 \\
&= \frac{1}{n}\sum_{i=1}^{n} U_i^2 + \left(\hat{\beta}_{0,n} - \beta_0\right)^2 + \left(\hat{\beta}_{1,n} - \beta_1\right)^2 \frac{1}{n}\sum_{i=1}^{n} X_i^2 \\
&\quad - 2\left(\hat{\beta}_{0,n} - \beta_0\right)\frac{1}{n}\sum_{i=1}^{n} U_i - 2\left(\hat{\beta}_{1,n} - \beta_1\right)\frac{1}{n}\sum_{i=1}^{n} U_i X_i \\
&\quad + 2\left(\hat{\beta}_{0,n} - \beta_0\right)\left(\hat{\beta}_{1,n} - \beta_1\right)\frac{1}{n}\sum_{i=1}^{n} X_i.
\end{aligned}$$

- By the LLN,

$$\frac{1}{n}\sum_{i=1}^{n} U_i^2 \to_p EU_i^2 = \sigma^2.$$

- Because $\hat{\beta}_{0,n}$ and $\hat{\beta}_{1,n}$ are consistent,

$$\hat{\beta}_{0,n} - \beta_0 \to_p 0 \text{ and } \hat{\beta}_{1,n} - \beta_1 \to_p 0.$$

Homoskedastic errors

- Thus, when the errors are homoskedastic,

$$\hat{V}_n = \frac{\hat{\sigma}_n^2}{\frac{1}{n}\sum_{i=1}^n \left(X_i - \bar{X}_n\right)^2}, \text{ with } \hat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^n \hat{U}_i^2,$$

is a consistent estimator of $V = \frac{\sigma^2}{Var(X_i)}$.
- Note that

$$s^2 = \frac{1}{n-2}\sum_{i=1}^n \hat{U}_i^2 \to_p \sigma^2,$$

and therefore

$$\hat{V}_n = \frac{s^2}{\frac{1}{n}\sum_{i=1}^n \left(X_i - \bar{X}_n\right)^2}$$

is also a consistent estimator of $V = \frac{\sigma^2}{Var(X_i)}$.
- This version has an advantage over the one with $\hat{\sigma}_n^2$: in addition to being consistent, s^2 is also an unbiased estimator of σ^2 if the regressors are strongly exogenous.

Homoskedastic errors: Asymptotic approximation

- Recall that $\sqrt{n}\left(\hat{\beta}_{1,n} - \beta_1\right) \rightarrow_d N\left(0, V\right)$ is used as the following approximation:

$$\hat{\beta}_{1,n} \overset{a}{\sim} N\left(\beta_1, \frac{V}{n}\right),$$

where $\overset{a}{\sim}$ denotes approximately in large samples. Thus, the variance of $\hat{\beta}_{1,n}$ can be taken as approximately V/n.

- Note that, with $\hat{V}_n = \frac{s^2}{\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}_n\right)^2}$ we have

$$\frac{\hat{V}_n}{n} = \frac{s^2}{\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}_n\right)^2}\frac{1}{n} = \frac{s^2}{\sum_{i=1}^{n}\left(X_i - \bar{X}_n\right)^2}.$$

Homoskedastic errors: Asymptotic approximation

$$\frac{\hat{V}_n}{n} = \frac{s^2}{\sum_{i=1}^{n} \left(X_i - \bar{X}_n\right)^2}$$

- Thus, in the case of homoskedastic errors we have the following asymptotic approximation:

$$\hat{\beta}_{1,n} \overset{a}{\sim} N\left(\beta_1, \frac{s^2}{\sum_{i=1}^{n} \left(X_i - \bar{X}_n\right)^2}\right).$$

- In finite samples, we have the same result exactly, when the regressors are **strongly exogenous** and the errors are **normal**.

Asymptotic T-test

- Consider testing $H_0 : \beta_1 = \beta_{1,0}$ vs $H_1 : \beta_1 \neq \beta_{1,0}$.
- Consider the behavior of T statistic under $H_0 : \beta_1 = \beta_{1,0}$. Since

$$\sqrt{n}\left(\hat{\beta}_{1,n} - \beta_1\right) \to_d N(0, V) \text{ and } \hat{V}_n \to_p V,$$

we have that

$$\begin{aligned} T = \frac{\left(\hat{\beta}_{1,n} - \beta_{1,0}\right)}{\sqrt{\hat{V}_n / n}} &= \frac{\sqrt{n}\left(\hat{\beta}_{1,n} - \beta_{1,0}\right)}{\sqrt{\hat{V}_n}} \\ &\overset{H_0}{=} \frac{\sqrt{n}\left(\hat{\beta}_{1,n} - \beta_1\right)}{\sqrt{\hat{V}_n}} \\ &\to_d \frac{N(0, V)}{\sqrt{V}} =^d N(0, 1). \end{aligned}$$

Asymptotic T-test

- We have that under $H_0 : \beta_1 = \beta_{1,0}$,

$$T = \frac{\left(\hat{\beta}_{1,n} - \beta_{1,0}\right)}{\sqrt{\hat{V}_n / n}} \to_d N\left(0, 1\right),$$

  provided that $\hat{V}_n \to_p V$ (the asymptotic variance of $\hat{\beta}_{1,n}$).
- An asymptotic size α test rejects $H_0 : \beta_1 = \beta_{1,0}$ against $H_1 : \beta_1 \neq \beta_{1,0}$ when

$$|T| > z_{1-\alpha/2},$$

  where $z_{1-\alpha/2}$ is a standard normal critical value.
- Asymptotically, the variance of the OLS estimator is known - we behave as if the variance was known.

Heteroskedastic errors

- In general, the errors are heteroskedastic: $E\left(U_i^2|X_i\right)$ is not constant and changes with X_i.
- In this case, $\hat{V}_n = \frac{s^2}{\frac{1}{n}\sum_{i=1}^n (X_i - \bar{X}_n)^2}$ is not a consistent estimator of the asymptotic variance $V = \frac{E\left((X_i - EX_i)^2 U_i^2\right)}{(Var(X_i))^2}$:

$$\begin{aligned} \frac{s^2}{\frac{1}{n}\sum_{i=1}^n \left(X_i - \bar{X}_n\right)^2} \to_p \frac{EU_i^2}{Var\left(X_i\right)} &= \frac{\left(E\left(X_i - EX_i\right)^2\right)\left(EU_i^2\right)}{\left(Var\left(X_i\right)\right)^2} \\ &\neq \frac{E\left(\left(X_i - EX_i\right)^2 U_i^2\right)}{\left(Var\left(X_i\right)\right)^2}. \end{aligned}$$

A heteroskedasticity consistent (HC) estimator of the asymptotic variance of OLS

- In the case of heteroskedastic errors, a consistent estimator of $V = \frac{E\left((X_i - EX_i)^2 U_i^2\right)}{(Var(X_i))^2}$ can be constructed as follows:

$$\hat{V}_n^{HC} = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}_n\right)^2 \hat{U}_i^2}{\left(\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}_n\right)^2\right)^2}.$$

- One can show that $\hat{V}_n^{HC} \to_p V$ when the errors are heteroskedastic or homoskedastic.
- We have the following asymptotic approximation:

$$\hat{\beta}_{1,n} \overset{a}{\sim} N\left(\beta_1, \frac{\hat{V}_n^{HC}}{n}\right),$$

and the standard errors can be computed as $SE\left(\hat{\beta}_{1,n}\right) = \sqrt{\hat{V}_n^{HC}/n}$.

HC variance estimation in Stata

- In Stata, the HC estimator of standard errors can be obtained by adding the option `robust` to the regression command:

```
. regress liver alcohol, robust

------------------------------------------------------------------------------
             |               Robust
       liver |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     alcohol |   3.586388    .550515     6.51   0.000     2.434147     4.73863
       _cons |   10.85482   2.119993     5.12   0.000     6.417625    15.29202
------------------------------------------------------------------------------
```

- Compare with the non-HC standard errors based on $\hat{V}_n$:

```
. regress liver alcohol

------------------------------------------------------------------------------
       liver |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     alcohol |   3.586388   .7541228     4.76   0.000     2.007991    5.164786
       _cons |   10.85482   2.802408     3.87   0.001     4.989313    16.72033
------------------------------------------------------------------------------
```