Introductory Econometrics

Lecture 13: Multiple Regression Analysis - Heteroskedasticity Pt. 2

Alex Alekseev

January 23, 2023

University of Regensburg, Department of Economics

Testing for heteroskedasticity

Recap

- The **heteroskedasticity-robust standard errors** provide a simple method for computing t statistics that are asymptotically t distributed whether or not heteroskedasticity is present
- The heteroskedasticity-robust F statistic is also available
- Implementing these tests does not require knowing whether or not heteroskedasticity is present
- There are good reasons for having simple **tests** that can detect its presence
- If heteroskedasticity is present, the OLS estimator is no longer the best linear unbiased estimator
- It is possible to obtain a **better estimator** than OLS when the form of heteroskedasticity is known

Population model

- As usual, we start with our population model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + U$$

- And make the following assumptions about it
  1. Linear CEF
     1.1 Linear model
     1.2 Error term is mean-independent of predictors
  2. Random Sampling
  3. No Perfect Collinearity
- We will not assume homoskedasticity
- We will not be assuming normality either

Population model

- As usual, we start with our population model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + U$$

- And make the following assumptions about it
  1. Linear CEF
     1.1 Linear model
     1.2 Error term is mean-independent of predictors
  2. Random Sampling
  3. No Perfect Collinearity
- We will not assume homoskedasticity
- We will not be assuming normality either
- The assumptions we made guarantee that OLS is unbiased and consistent

Hypothesis

- The null hypothesis in our test will be **homoskedasticity**:

$$H_0 : \mathrm{Var}(U \mid \mathbf{x}) = \sigma^2.$$

- If we cannot reject the null, we would conclude that heteroskedasticity is not a problem
- If we do reject the null, we would have to correct for heteroskedasticity

Re-write the hypothesis

- Recall that since the error term has zero mean, we have that

$$\mathrm{Var}(U \mid \mathbf{x}) = \mathbb{E}[U^2 \mid \mathbf{x}].$$

- Hence our null hypothesis can be re-written as

$$H_0 : \mathbb{E}[U^2 \mid \mathbf{x}] = \mathbb{E}[U^2] = \sigma^2.$$

Main idea

- Then the idea of the test is to see whether the expectation of U^2 is **associated with any of the predictors**
- If the null hypothesis is false, the conditional expectation of U^2 can be any function of the predictors
- A simple approach is to assume a **linear relationship**

$$U^2 = \delta_0 + \delta_1 X_1 + \ldots + \delta_k X_k + V,$$

where V is an error term that is mean-independent of all the predictors

Test

- In this specification, the null hypothesis is equivalent to saying that

$$H_0 : \delta_1 = \delta_2 = \ldots = \delta_k = 0.$$

- And we know how to test this hypothesis using an F-test for the **overall significance**
- The F statistic for this regression can be shown to have an asymptotic F distribution

Residuals

- We cannot just run the regression above because we **do not observe** the error term U
- As before, we do the next best thing and replace it with **residuals** $\hat{U}$:

$$\hat{U}^2 = \delta_0 + \delta_1 X_1 + \ldots + \delta_k X_k + W.$$

- We can then estimate this model and do an F-test for the overall significance
- It turns out that using the OLS residuals in place of the errors does not affect the large sample distribution of the F statistic

Breusch-Pagan test

- The described procedure for testing for heteroskedasticity is often called the **Breusch-Pagan test for heteroskedasticity (BP test)**
- If an F-test shows that the coefficients are jointly not statistically significant, then heteroskedasticity is likely not a problem
- If an F-test rejects the null, though, we would need to correct for heteroskedasticity

Breusch-Pagan test

- The described procedure for testing for heteroskedasticity is often called the **Breusch-Pagan test for heteroskedasticity (BP test)**
- If an F-test shows that the coefficients are jointly not statistically significant, then heteroskedasticity is likely not a problem
- If an F-test rejects the null, though, we would need to correct for heteroskedasticity

Note

The original Breusch-Pagan test actually uses a different test statistic called an LM statistic instead of an F statistic. The LM statistic is computed as nR^2 and has an asymptotic χ^2 distribution. In practice, using either an LM or F tests tends to give similar results.

The White test for heteroskedasticity

Motivation for the test

- It turns out that the homoskedasticity assumption can be replaced with the weaker assumption
- The assumption is that the squared error U^2 is uncorrelated with all the predictors, the squares of the predictors, and all the interactions between the predictors
- This observation motivates **the White test**

Example

- For example, with three predictors, the estimated equation will look like

$$\begin{aligned}\hat{U}^2 &= \delta_0 + \delta_1 X_1 + \delta_2 X_2 + \delta_3 X_3 \\ &+ \delta_4 X_1^2 + \delta_5 X_2^2 + \delta_6 X_3^2 \\ &+ \delta_7 X_1 X_2 + \delta_8 X_1 X_3 + \delta_9 X_2 X_3 \\ &+ W.\end{aligned}$$

- Compare this with the Breusch-Pagan test, which used only the main terms and no squares or interactions
- The White test for heteroskedasticity is then the test of the overall significance of the coefficients in the above regression
- One issue with the the White test, however, is that the number of terms grows very quickly with the number of predictors

Note

An important caveat when using a test for heteroskedasticity (either the Breusch-Pagan or White) is that we interpret the rejection of the null hypothesis as evidence **in favor of heteroskedasticity**. This interpretation relies on our **assumptions to be true**. If, however, one of those assumptions is violated, for example, we omit a relevant predictor from the model, then a test for heteroskedasticity can reject the null **even if the variance is constant**.

Weighted least squares

Assumptions

- Suppose we can write the variance on the error term as

$$\mathrm{Var}(U \mid \mathbf{x}) = \sigma^2 h(\mathbf{x}),$$

  where $h(\mathbf{x})$ is some **known** function of the predictors, which determines the form of heteroskedasticity, and σ^2 is a parameter we estimate
- In our random sample, we can write

$$\sigma_i^2 = \mathrm{Var}(U_i \mid \mathbf{x}_i) = \sigma^2 h(\mathbf{x}_i) = \sigma^2 h_i$$

Transformation

- Let's consider our sample regression

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_k X_{ik} + U_i.$$

- We know that U_i error terms are heteroskedastic
- But what about $U_i^* \equiv U_i/\sqrt{h_i}$?
- Its variance is

$$\begin{aligned}
\text{Var}(U_i^* \mid \mathbf{x}_i) &= \mathbb{E}[(U_i^*)^2 \mid \mathbf{x}_i] \\
&= \mathbb{E}[(U_i/\sqrt{h_i})^2 \mid \mathbf{x}_i] \\
&= \mathbb{E}[U_i^2/h_i \mid \mathbf{x}_i] \\
&= 1/h_i \mathbb{E}[U_i^2 \mid \mathbf{x}_i] \\
&= 1/h_i \sigma^2 h_i \\
&= \sigma^2.
\end{aligned}$$

Transformed model

- But how can we get to that transformed error term?
- We just take our model and scale all the variables by $1/\sqrt{h_i}$:

$$\begin{aligned} Y_i/\sqrt{h_i} &= \beta_0/\sqrt{h_i} + \beta_1 X_{i1}/\sqrt{h_i} + \beta_2 X_{i2}/\sqrt{h_i} + \dots \\ &\quad + \beta_k X_{ik}/\sqrt{h_i} + U_i\sqrt{h_i} \\ Y_i^* &= \beta_0 X_{i0}^* + \beta_1 X_{i1}^* + \beta_2 X_{i2}^* + \dots \\ &\quad + \beta_k X_{ik}^* + U_i^*, \end{aligned}$$

where $X_{i0}^* \equiv 1/\sqrt{h_i}$

Properties of the transformed model

$$Y_i^* = \beta_0 X_{i0}^* + \beta_1 X_{i1}^* + \beta_2 X_{i2}^* + \ldots + \beta_k X_{ik}^* + U_i^*,$$

- The model with **transformed variables** now satisfies homoskedasticity, as well
- We can estimate this model and conduct inference as before
- We can estimate it using OLS, however, the estimators from the transformed model will be different from the ones in the original model
- The estimator from the transformed model will be BLUE, unlike the OLS estimator from the original model

Weighted least squares

- The solution we came up with is an example of the **weighted least squares (WLS) estimator**
- The estimators from the transformed model minimize the weighted sum of squared residuals, where each squared residual is weighted by $1/h_i$
- The idea is that **less weight** is given to observations with a **higher error variance**
- OLS, on the other hand, gives each observation the same weight because it is best when the error variance is constant

Equivalence

- Mathematically, the WLS estimator is the values of b_j that minimize

$$\sum_{i=1}^{n} w_i(y_i - b_0 - b_1 x_{i1} - b_2 x_{i2} - \ldots - b_k x_{ik})^2,$$

where w_i are the weights, which in our case equal to $1/h_i$

- If we take these weights inside the parentheses, we get

$$\sum_{i=1}^{n}(y_i^* - b_0 x_{i0}^* - b_1 x_{i1}^* - b_2 x_{i2}^* - \ldots - b_k x_{ik}^*)^2,$$

which is identical to the sum of squared residuals from the transformed model

Weights

- A weighted least squares estimator can be defined for **any set of positive weights** w_i
- OLS is the special case that gives equal weights to all observations
- The efficient procedure weights each squared residual by the inverse of the conditional variance of U_i given $\mathbf{x}_i$
- This is called the **inverse variance weighting**

Weights

- A weighted least squares estimator can be defined for **any set of positive weights** w_i
- OLS is the special case that gives equal weights to all observations
- The efficient procedure weights each squared residual by the inverse of the conditional variance of U_i given $\mathbf{x}_i$
- This is called the **inverse variance weighting**

Note

The use of the WLS estimator is not limited to correcting for **heteroskedasticity**. Another popular reason to use weights is to make your sample more **representative** of the population you are studying, which is common in **survey data**.

Data aggregation

When we know h

- For the WLS to solve our heteroskedasticity problem, we need to know the $h(\cdot)$ function
- While in general we do not know it and have to estimate, in some cases the form of heteroskedasticity is implied by the underlying model
- One such case is **aggregated data**

Example

- Suppose our unit of observation is a student, which we will index by j, in a given university, which we will index by i
- Let's say we are interested in the effect of the number of hours a student spends studying econometrics X_{ij} on their future income Y_{ij}
- This leads to a simple regression model

$$Y_{ij} = \beta_0 + \beta_1 X_{ij} + U_{ij}$$

- Let's assume that this model satisfies the full set of Gauss-Markov assumptions, including homoskedasticity: $\text{Var}(U_{ij} \mid X_{ij}) = \sigma^2$

Averages

- Now suppose we do not have access to **individual student data** due to privacy laws
- Instead, we have data **averaged** across all students in a given university
- Let's call the number of students in university i m_i
- We now have access only to the following averages

$$\bar{Y}_i \equiv \frac{1}{m_i}\sum_{j=1}^{m_i} Y_{ij}, \quad \bar{X}_i \equiv \frac{1}{m_i}\sum_{j=1}^{m_i} X_{ij}.$$

- Our aggregated population model then becomes

$$\bar{Y}_i = \beta_0 + \beta_1 \bar{X}_i + \bar{U}_i,$$

where $\bar{U}_i \equiv \frac{1}{m_i}\sum_{j=1}^{m_i} U_{ij}$

Variance

- Let's find the variance of this averaged error term:

$$
\begin{aligned}
\mathrm{Var}(\bar{U}_i \mid \bar{X}_i) &= \mathrm{Var}\left(\frac{1}{m_i}\sum_{j=1}^{m_i} U_{ij} \mid \bar{X}_i\right) \\
&= \frac{1}{m_i^2}\mathrm{Var}\left(\sum_{j=1}^{m_i} U_{ij} \mid \bar{X}_i\right) \\
&= \frac{1}{m_i^2}\sum_{j=1}^{m_i}\mathrm{Var}\left(U_{ij} \mid \bar{X}_i\right) \\
&= \frac{1}{m_i^2}\sum_{j=1}^{m_i}\sigma^2 \\
&= \frac{1}{m_i}\sigma^2.
\end{aligned}
$$

- **Larger schools** will have a **lower variance** since there are simply more observations to estimate the mean than in smaller schools

Weights

- This derivation shows that our (typically unknown) h function is simply $h_i = 1/m_i$
- Hence the weights we use in the WLS are $w_i = 1/h_i = m_i$
- A similar weighting arises when we are using **per capita data** at the city, county, state, or country level
- If the individual-level equation satisfies the Gauss-Markov assumptions, then the error in the per capita equation has a variance proportional to one over the size of the population
- Therefore, we should use the WLS estimator with weights equal to the population size

Note

- The procedure relies on the underlying individual equation being **homoskedastic**
- If **heteroskedasticity** exists at the individual level, then the proper weighting depends on the form of the heteroskedasticity
- To address this concern, we could weight by population but report the **heteroskedasticity-robust** statistics in the WLS estimation
- This ensures that, while the estimation is efficient if the individual-level model satisfies the Gauss-Markov assumptions, any heteroskedasticity at the individual level is accounted for through robust inference

Feasible generalized least squares

Feasible GLS

- In this section, we look at how we can model the function h and use the data to estimate the unknown parameters in this model
- This results in an estimate of each h_i, denoted as $\hat{h}_i$
- Using $\hat{h}_i$, instead of h, in the transformed model yields an estimator called the **feasible GLS (FGLS) estimator**

Modeling heteroskedasticity

- One way to model heteroskedasticity is

$$\mathrm{Var}(U \mid \mathbf{x}) = \sigma^2 \exp(\delta_0 + \delta_1 X_1 + \ldots + \delta_k X_k),$$

  where the expression $\exp(\delta_0 + \delta_1 X_1 + \ldots + \delta_k X_k)$ is our model for the h function
- Why do we use the **exponent** in this expression?

Modeling heteroskedasticity

- One way to model heteroskedasticity is

$$\text{Var}(U \mid \mathbf{x}) = \sigma^2 \exp(\delta_0 + \delta_1 X_1 + \ldots + \delta_k X_k),$$

  where the expression $\exp(\delta_0 + \delta_1 X_1 + \ldots + \delta_k X_k)$ is our model for the h function
- Why do we use the **exponent** in this expression?
- We need the variance, as well as weights, to be **positive**, and taking an exponent is one way to ensure that

Estimation

- Recall that $\text{Var}(U \mid \mathbf{x}) = \mathbb{E}[U^2 \mid \mathbf{x}]$
- Then our model is

$$\mathbb{E}[U^2 \mid \mathbf{x}] = \sigma^2 \exp(\delta_0 + \delta_1 X_1 + \ldots + \delta_k X_k)$$

- We can multiply this expression by an error term V that has a mean of one to get a regression model

$$U^2 = \sigma^2 \exp(\delta_0 + \delta_1 X_1 + \ldots + \delta_k X_k) V.$$

- Taking logs, we convert this into a linear model

$$\log(U^2) = \alpha_0 + \delta_1 X_1 + \ldots + \delta_k X_k + W,$$

where W is an error term with a zero mean and is assumed to independent of the predictors

Using the model

- We substitute the error term with the residuals $\hat{U}$
- The model that we estimated is then

$$\log(\hat{U}^2) = \alpha_0 + \delta_1 X_1 + \ldots + \delta_k X_k + W.$$

- We need the fitted values from this model, call them $\hat{g}_i$
- Then the estimates of h_i are

$$\hat{h}_i = \exp(\hat{g}_i).$$

- Then we can use the WLS with weights given by $w_i = 1/\hat{h}_i$

Properties of FGLS

- If we could use h_i rather than its estimate, the WLS estimator would be unbiased
- In fact it would be BLUE, assuming that we have properly modeled the heteroskedasticity
- Having to **estimate** h_i using the same data means that the FGLS estimator is **no longer unbiased** (so it cannot be BLUE, either)
- Nevertheless, the FGLS estimator is **consistent** and asymptotically **more efficient** than OLS

Interpretation

- Keep in mind that the FGLS estimator is an estimator of the parameters of the original population model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + U.$$

- All of the coefficients estimates from FGLS will have **the same interpretation as from OLS**
- The magnitudes of the coefficients as well as standard errors, however, will probably differ
- If we have doubts about whether we correctly modeled the variance, we can use heteroskedasticity-robust standard errors and test statistics in the transformed equation

Different results

- If OLS and WLS produce statistically significant estimates that differ in sign, we should be **suspicious**
- Typically, this indicates that one of the other Gauss-Markov assumptions is **violated**
- In particular, the zero conditional mean assumption on the error
- Correlation between U and any predictor causes bias and inconsistency in OLS and WLS, and the biases will usually be different

Example

Trade model

- We will practice testing for heteroskedasticity and doing a FGLS using our trade gravity model

$$\ln(imports_i) = \beta_0 + \beta_1 \ln(gdp_i) + \beta_2 \ln(distance_i) + u_i$$

Trade model

- We will practice testing for heteroskedasticity and doing a FGLS using our trade gravity model

$$\ln(imports_i) = \beta_0 + \beta_1 \ln(gdp_i) + \beta_2 \ln(distance_i) + u_i$$

statistic	p.value	method
1.717	0.424	Breusch-Pagan
4.150	0.528	White's Test

Trade model

- We will practice testing for heteroskedasticity and doing a FGLS using our trade gravity model

$$\ln(imports_i) = \beta_0 + \beta_1 \ln(gdp_i) + \beta_2 \ln(distance_i) + u_i$$

statistic	p.value	method
1.717	0.424	Breusch-Pagan
4.150	0.528	White's Test

- The *p*-values from both tests show that we cannot reject the null hypothesis of homoskedasticity

FGLS

- Even though there is no substantial evidence for heteroskedasticity, we can still use the FGLS
- First, we estimate the auxiliary regression to get the predicted values $\hat{h}_i$
- Then we use these to get the weights in the WLS estimation
- We will compare the results from OLS and WLS (or FGLS)
- Just to be safe, we will also use the robust standard errors

Results

	OLS	WLS
(Intercept)	4.670	4.410
	(2.051)	(1.633)
log(gdp)	0.976	0.937
	(0.062)	(0.057)
log(distance)	-1.075	-0.904
	(0.173)	(0.143)
Num.Obs.	48	48
R2	0.886	0.880
Std.Errors	HC3	HC3

Other issues with standard errors

Independence

- Recall our assumption about **random sampling**
- It says that the observations, and in particular the error terms, are **independent** and **identically distributed**
- **Independence** means that an error term in one observation is unrelated to error terms in any other observations

Correlation

- One common way in which error terms can be correlated is **across time**
- This would be the case for **time-series** and **panel** data
- A way to correct for that is to use the heteroskedasticity- and autocorrelation-consistent (HAC) standard errors, such as **the Newey-West estimator**
- Another common way to be correlated is **across space**, which is the case if you are working with **geographic** data
- In this case, one could use **Conley spatial standard errors** to correct for correlation among geographic neighbors

Clustering

- Another common way for errors to be correlated is in a **hierarchical structure**
- Recall our earlier example with students who study econometrics and their future earnings
- It is common for students to be assigned to groups (or classrooms)
- The behavior of students in the same group is likely to be correlated because they have shared experiences
- In this case, we should use **clustered standard errors**