Lecture 12 Multiple Regression Analysis - Heteroskedasticity

Homoskedasticity and heteroskedasticity

Let's recall our population model

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

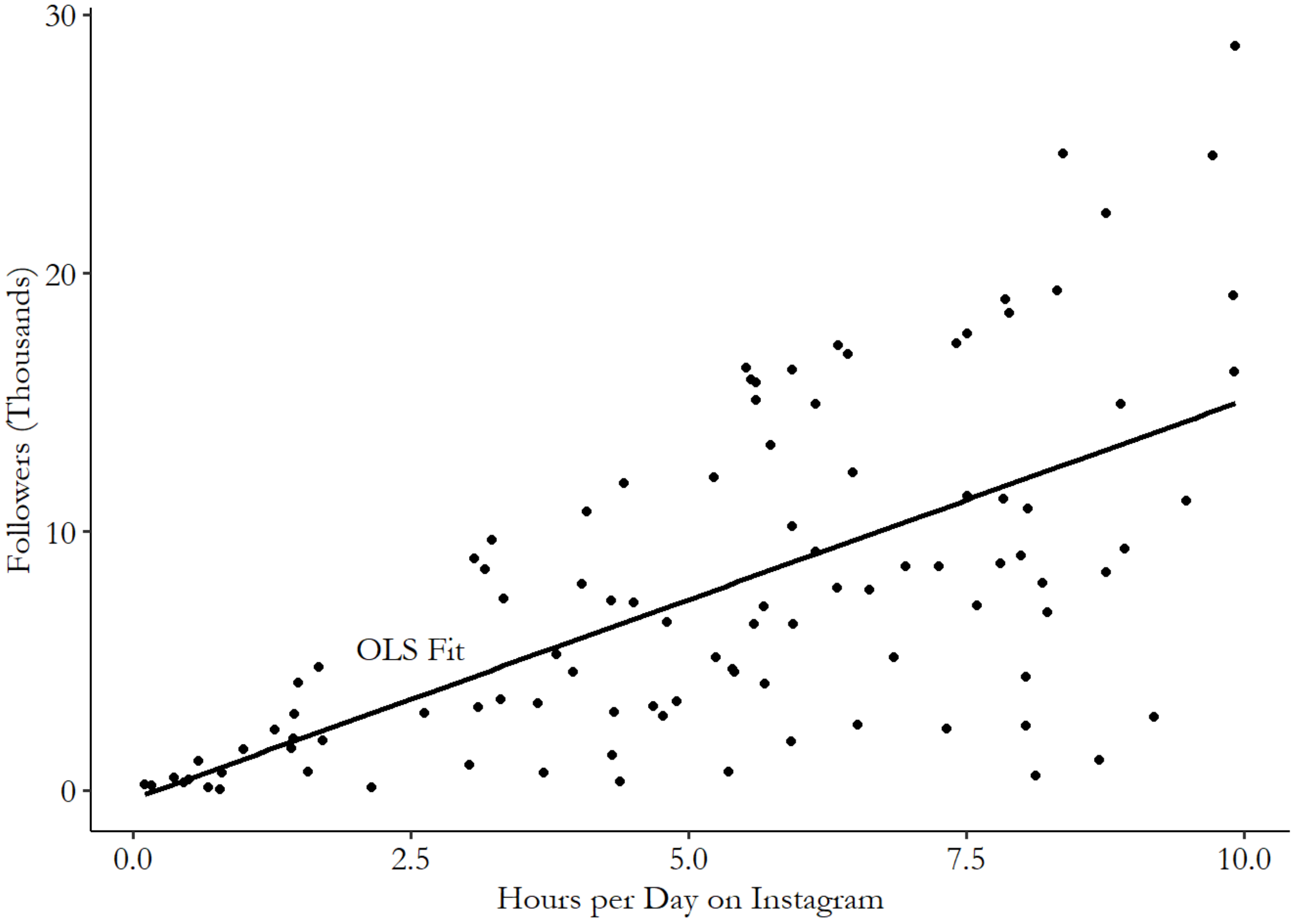
and the assumptions we made about our it.

1. Linear CEF
    1. Linear model
    2. Error term is mean-independent of predictors
2. Random Sampling
3. No Perfect Collinearity
4. Homoskedasticity
5. Normality

In the previous lecture, we showed that we can relax the normality assumption. We needed this assumption for deriving the exact sampling distribution of the OLS estimator. However, even if the error term is not normally distributed, the large sample properties of the OLS guarantee that its sampling distribution is approximately normal, at least in large samples. As a consequence of that, we can still use the usual t and F statistics, with the understanding that they will be distributed *approximately* as t and F random variables, respectively.

In this lecture, we will relax another assumption about the error term: homoskedasticity. The homoskedasticity (or constant variance) assumption states that the variance of the error term is constant, conditional on the predictors, $\mathrm{Var}(U \mid \mathbf{x}) = \sigma^2$. The failure of homoskedasticity is called *heteroskedasticity*. Homoskedasticity fails whenever the variance of the error term changes across different segments of the population (observations), which are determined by the different values of the predictors.

For example, suppose we are interested in the effect of hours per day spent on Instagram by a given person on the number of followers that person has. We might find that people who spend very little time on Instagram have very few followers and there is very little variation in the number of followers for these people. On the other hand, heavy Instagram users might on average have more followers, however, there will also be a lot of variation in the number of followers. In this case, the homoskedasticity assumption is violated. The picture below illustrates the idea.



Consequences of heteroskedasticity for OLS

Why do we need homoskedasticity, anyway? We do not need it for the unbiasedness or consistency. Just like with the normality assumption, the homoskedasticity assumption was needed for statistical inference. Even if we can drop the normality assumption, we still need to know that variance of the OLS estimator. We need the variance, in turn, to compute the standard errors and conduct hypotheses testing. It turns out that without the homoskedasticity assumption, the usual OLS standard errors will be biased and the t and F statistics computed using those standard errors will not have the t and F distributions, even asymptotically.

Likewise, the failure of homoskedasticity will invalidate the results of the Gauss-Markov theorem. This should not be surprising since homoskedasticity is a part of the Gauss-Markov assumptions. Under heteroskedasticity, OLS is no longer the best linear unbiased estimator.

Let's recall the derivation of the variance of the OLS estimator to see where exactly the homoskedasticity assumption plays a role. Our starting point is the regression anatomy formula:

$$\hat{\beta}_j = \frac{\widehat{\text{Cov}}(Y, U_j)}{\widehat{\text{Var}}(U_j)},$$

where U_j is the error term from the regression of X_j on all other predictors:

$$X_j = \gamma_0 + \gamma_1 X_1 + \ldots + \gamma_{j-1} X_{j-1} + \gamma_{j+1} X_{j+1} + \ldots + \gamma_k X_k + U_j.$$

During the proof of the formula, we showed that

$$\widehat{\mathrm{Cov}}(Y, U_j) = \frac{1}{n}\sum_{i=1}^{n} y_i \hat{u}_{ij}.$$

Substituting for y_i, we get

$$\begin{aligned}\widehat{\mathrm{Cov}}(Y, U_j) &= \frac{1}{n}\sum_{i=1}^{n}(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + u_i)\hat{u}_{ij} \\ &= \frac{1}{n}\sum_{i=1}^{n}(\beta_j x_{ij}\hat{u}_{ij} + u_i\hat{u}_{ij}) \\ &= \frac{1}{n}\beta_j \sum_{i=1}^{n} x_{ij}\hat{u}_{ij} + \frac{1}{n}\sum_{i=1}^{n} u_i\hat{u}_{ij}.\end{aligned}$$

On the other hand, we have shown that this covariance can be written as

$$\begin{aligned}\widehat{\mathrm{Cov}}(Y, U_j) &= \frac{1}{n}\hat{\beta}_j \sum_{i=1}^{n} x_{ij}\hat{u}_{ij} \\ &= \frac{1}{n}\hat{\beta}_j \sum_{i=1}^{n} \hat{u}_{ij}^2.\end{aligned}$$

We denote $SSR_j \equiv \sum_{i=1}^{n} \hat{u}_{ij}^2$.

Equating these two expressions, we get

$$\begin{aligned}\frac{1}{n}\hat{\beta}_j \sum_{i=1}^{n} x_{ij}\hat{u}_{ij} &= \frac{1}{n}\beta_j \sum_{i=1}^{n} x_{ij}\hat{u}_{ij} + \frac{1}{n}\sum_{i=1}^{n} u_i\hat{u}_{ij} \\ \hat{\beta}_j &= \beta_j + \frac{\frac{1}{n}\sum_{i=1}^{n} u_i\hat{u}_{ij}}{\frac{1}{n}\sum_{i=1}^{n}\hat{u}_{ij}^2} \\ \hat{\beta}_j &= \beta_j + \frac{\sum_{i=1}^{n} u_i\hat{u}_{ij}}{SSR_j}.\end{aligned}$$

If we denote

$$w_{ij} \equiv \frac{\hat{u}_{ij}}{SSR_j},$$

the formula becomes

$$\hat{\beta}_j = \beta_j + \sum_{i=1}^{n} w_{ij} u_i.$$

Now we can compute the variance of $\hat{\beta}_j$ conditional on the predictors:

$$
\begin{aligned}
\operatorname{Var}(\hat{\beta}_j \mid \mathbf{x}) &= \operatorname{Var}(\sum_{i=1}^{n} w_{ij} u_i \mid \mathbf{x}) \\
&= \sum_{i=1}^{n} \operatorname{Var}(w_{ij} u_i \mid \mathbf{x}) \\
&= \sum_{i=1}^{n} w_{ij}^2 \operatorname{Var}(u_i \mid \mathbf{x}) \\
&= \frac{\sum_{i=1}^{n} \hat{u}_{ij}^2 \operatorname{Var}(u_i \mid \mathbf{x})}{SSR_j^2}.
\end{aligned}
$$

If we assume homoskedasticity, $\operatorname{Var}(u_i \mid \mathbf{x}) = \sigma^2$, the formula becomes

$$
\operatorname{Var}(\hat{\beta}_j \mid \mathbf{x}) = \frac{\sum_{i=1}^{n} \hat{u}_{ij}^2 \sigma^2}{SSR_j^2} = \frac{\sigma^2}{SSR_j} = \frac{\sigma^2}{SST_j(1 - R_j^2)}.
$$

We can then estimate this variance by estimating σ^2 as

$$
\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} \hat{u}_i^2}{n - (k+1)},
$$

which leads to

$$
\widehat{\operatorname{Var}}(\hat{\beta}_j \mid \mathbf{x}) = \frac{\hat{\sigma}^2}{SST_j(1 - R_j^2)} = \frac{1}{n - (k+1)} \frac{\sum_{i=1}^{n} \hat{u}_i^2}{SST_j(1 - R_j^2)}.
$$

Thus, our derivation of the variance of the OLS estimator, as well as other quantities that depend on that variance, such as test statistics, does rely on homoskedasticity. What should we do, then, if homoskedasticity fails?

Heteroskedasticity-robust inference after OLS estimation

Hypotheses testing after estimation is an important part of any econometric analysis. However, the usual OLS inference will in general be invalid in the presence of heteroskedasticity. Does this mean that we should abandon OLS? Fortunately, OLS is still useful. There are methods to adjust standard errors, as well as test statistics, so that they are valid in the presence of heteroskedasticity of *unknown form*. These methods are called as heteroskedasticity-robust (heteroskedasticity-consistent) procedures because they are valid, at least in large samples, whether or not the errors have constant variance, and we do not need to know which is the case.

Let's return to the formula for the variance of $\hat{\beta}_j$

$$
\operatorname{Var}(\hat{\beta}_j \mid \mathbf{x}) = \frac{\sum_{i=1}^{n} \hat{u}_{ij}^2 \operatorname{Var}(u_i \mid \mathbf{x})}{SSR_j^2}.
$$

The key insight of the heteroskedasticity-robust methods is to estimate each individual variance $\operatorname{Var}(u_i \mid \mathbf{x})$ using $\hat{u}_i^2$. A careful proof of why this method works is fairly technical and is not given here. However, an intuition is that this method can be thought of as estimating the variance of each individual error term u_i using a single data point i.

Then a valid estimator of $\mathrm{Var}(\hat{\beta}_j \mid \mathbf{x})$, for heteroskedasticity of any form (including homoskedasticity), is

$$\widehat{\mathrm{Var}}(\hat{\beta}_j \mid \mathbf{x}) = \frac{\sum_{i=1}^{n} \hat{u}_{ij}^2 \hat{u}_i^2}{SSR_j^2}.$$

The square root of this quantity is the *heteroskedasticity-robust standard error* for $\hat{\beta}_j$. This approach was developed in the works of Eicker, Huber, and White, and often these standard errors are referred by the names of these authors. But most often, they are simply referred to as *robust* standard errors. We will call the variance above *HC0*. We will also call the usual OLS variance *NHC*.

Note

Strictly speaking, the terms such as HC0 refer to the full variance-covariance matrix of the OLS estimator. Here we use them only to refer to the variances of each coefficient estimate, for simplicity.

Once we have heteroskedasticity-robust standard errors, it is simple to construct a heteroskedasticity-robust t statistic. Recall that the general form of the t statistic is

$$t = \frac{\text{estimate} - \text{hypothesized value}}{\text{standard error}}$$

Since we are still using the OLS estimates and we have chosen the hypothesized value ahead of time, the only difference between the usual OLS t statistic and the heteroskedasticity-robust t statistic is in how the standard error is computed. It is also possible to obtain an F statistic that is robust to heteroskedasticity of an unknown, arbitrary form, but we will not do it here.

Modifications of robust standard errors

There are several modifications of the HC0 standard errors that improve upon their small sample properties. The first small-sample modification is called *HC1* and is given by

$$HC1 = \frac{n}{n-(k+1)} \frac{\sum_{i=1}^{n} \hat{u}_{ij}^2 \hat{u}_i^2}{SSR_j^2} = \frac{n}{n-(k+1)} HC0$$

The motivation for this is that if the squared OLS residuals were the same for all observations (this would be the strongest possible form of homoskedasticity in a sample), $\hat{u}_i^2 = \hat{u}^2$, we would get the usual OLS standard errors:

$$\begin{aligned} HC1 &= \frac{n}{n-(k+1)} \frac{\sum_{i=1}^{n} \hat{u}_{ij}^2 \hat{u}^2}{SSR_j^2} = \frac{n}{n-(k+1)} \frac{\hat{u}^2}{SSR_j} \\ NHC &= \frac{1}{n-(k+1)} \frac{\sum_{i=1}^{n} \hat{u}^2}{SSR_j} = \frac{n}{n-(k+1)} \frac{\hat{u}^2}{SSR_j}. \end{aligned}$$

Since $n/(n-(k+1)) > 1$, we have that $HC1 > HC0$. In words, the HC1 standard errors will be always be larger than HC0 standard errors.

Two other modifications of the HC0 standard errors are motivated by the role of outliers and influential observations. A key insight here is that even if the errors are homoscedastic, the residuals will not be. If error terms have a constant variance σ^2, it turns out that

$$\texttt{Var}(\hat{u}_i) = \sigma^2(1-h_{ii}),$$

where h_{ii} are the diagonal elements of the so-called *hat matrix* $\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. It can be shown that $0 \leqslant 1-h_{ii} \leqslant 1-1/n$, so that $\texttt{Var}(\hat{u}_i)$ will always be lower than $\texttt{Var}(u_i)$.

This suggests modifying the HC0 standard errors in the following way:

$$HC2 = \frac{1}{1-h_{ii}} HC0.$$

Even if $\hat{u}_i^2$ is a biased estimator of $\texttt{Var}(u_i)$, then $\hat{u}_i^2/(1-h_{ii})$ will be less biased. It easy easy to see that $HC2 > HC0$.

Hat matrix

Why is $\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ called the hat matrix? Recall the OLS estimator formula in the matrix notation:

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}.$$

Also recall that the fitted values can be found as

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}.$$

The matrix $\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, when applied to a vector of outcomes $\mathbf{y}$ produces the fitted (or *hat*) values, hence the name *hat matrix*.

Yet another modification of the HC0 standard errors is given by

$$HC3 = \frac{1}{(1-h_{ii})^2} HC0.$$

Dividing by the square of $1-h_{ii}$ increases the variance even further, which is thought to adjust for the excessive influence of observations with large variances. The derivation of this formula relies on a so-called *jackknife* estimator, which we will not cover here. Since $1-h_{ii} < 1$, we have that $HC3 > HC2$.

Relationships between different modifications

In general, we cannot say whether the robust standard errors HC0 will be higher or lower than the usual OLS standard errors. In practice, HC0 tends to be higher, but this is not guaranteed. Also, we cannot in general say whether HC2 is higher or lower than HC1.

At this point, we might be wondering which heteroskedasticity-robust standard errors we should be using: HC0, HC1, HC2, or HC3. Since all forms have only asymptotic justification, they are asymptotically equivalent, and there is no strong theoretical reason to prefer one over the other. However, practice suggests that HC3 standard errors tend to perform better than others.

Stata vs. R

Stata and R have different defaults for heteroskedasticity-robust standard errors. Stata uses HC1 by default (it refers to them as simply *robust*), while the R package `sandwich` uses HC3 by default.

Sandwiches

The heteroskedasticity-robust estimators of the variances of OLS are often called the *sandwich* estimators. The motivation for this name comes from the matrix notation for the OLS and how we derive the full variance-covariance matrix of the vector of coefficients $\hat{\beta}$. Previously, we have shown that we can write the variance matrix as

$$\mathrm{Var}(\hat{\beta} \mid \mathbf{X}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\Omega\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1},$$

where $\Omega \equiv \mathrm{Var}(\mathbf{u} \mid \mathbf{X}) = \mathbb{E}[\mathbf{u}\mathbf{u}' \mid \mathbf{X}]$ is the variance matrix of the vector of error terms.

Under homoskedasticity, we assume that $\Omega = \sigma^2\mathbf{I}$, where $\mathbf{I}$ is the identity matrix. The homoskedasticity assumption allows us to considerably simplify the formula for the variance matrix of $\hat{\beta}$

$$\begin{aligned}\mathrm{Var}(\hat{\beta} \mid \mathbf{X}) &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\sigma^2\mathbf{I}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}.\end{aligned}$$

This is the usual variance matrix of the OLS estimator.

However, in general this simplification is not possible if homoskedasticity fails. In the case of heteroskedasticity, we have to work with the general formula

$$\mathrm{Var}(\hat{\beta} \mid \mathbf{X}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\Omega\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$$

in which matrix $\mathbf{X}'\Omega\mathbf{X}$ is sandwiched between the two $(\mathbf{X}'\mathbf{X})^{-1}$ matrices. Different heteroskedasticity-robust variance estimators correspond to different assumptions about how to estimate Ω.

For the HC0 standard errors, the estimate of the Ω matrix is proposed to be

$$\hat{\Omega} = \mathrm{diag}(\hat{u}_i^2),$$

where $\mathrm{diag}(\hat{u}_i^2)$ denotes a diagonal matrix with elements $\hat{u}_i^2$ in the diagonal and zeros everywhere else.

Likewise, the other heteroskedasticity robust estimators make the following assumptions

$$HC1 : \hat{\Omega} = \frac{n}{n-(k+1)}\mathrm{diag}(\hat{u}_i^2)$$
$$HC2 : \hat{\Omega} = \mathrm{diag}\left(\frac{\hat{u}_i^2}{1-h_{ii}}\right)$$
$$HC3 : \hat{\Omega} = \mathrm{diag}\left(\frac{\hat{u}_i^2}{(1-h_{ii})^2}\right)$$

Example

Let's see how different heteroskedasticity-robust standard errors work in the trade example. We will estimate the gravity model

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

but vary how we estimate the standard errors. We will look at the usual OLS standard errors, as well as the four HC kinds we considered before. The table below shows the results.

	Model 1	Model 2	Model 3	Model 4	Model 5
(Intercept)	4.670	4.670	4.670	4.670	4.670
	(2.181)	(1.821)	(1.880)	(1.929)	(2.051)
log(gdp)	0.976	0.976	0.976	0.976	0.976
	(0.064)	(0.057)	(0.058)	(0.059)	(0.062)
log(distance)	-1.075	-1.075	-1.075	-1.075	-1.075
	(0.157)	(0.153)	(0.158)	(0.163)	(0.173)
Num.Obs.	48	48	48	48	48
Std.Errors	Constant	HC0	HC1	HC2	HC3

As you can see, the coefficient estimates stay the same, only the standard errors change. In this example, the HC0 standard errors tend to be lower than the usual ("constant") standard errors, although this is not always the case.