ECON 5360 Class Notes
Heteroscedasticity

1 Introduction

In this chapter, we focus on the problem of heteroscedasticity within the multiple linear regression model. Throughout, we assume that all other classical assumptions are satisfied. Assume the model is

$$Y = X\beta + \epsilon \tag{1}$$

where

$$E(\epsilon\epsilon') = \sigma^2\Omega = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 & 0 \\ 0 & \sigma_2^2 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & \sigma_{n-1}^2 & 0 \\ 0 & 0 & \cdots & 0 & \sigma_n^2 \end{bmatrix}. \tag{2}$$

Heteroscedasticity is a common occurrence in cross-sectional data. It can also occur in time series data (e.g., AutoRegressive Conditional Heteroscedasticity, ARCH).

2 Ordinary Least Squares

We now examine several results related to OLS when heteroscedasticity is present in the model.

2.1 Summary of Findings

1. $b = (X'X)^{-1}X'Y$ is unbiased and consistent.
2. $var(b) = \sigma^2(X'X)^{-1}X'\Omega X(X'X)^{-1}$ is the correct formula.
3. $var(b) = \sigma^2(X'X)^{-1}$ is the incorrect formula.
4. $b \overset{asy}{\sim} N(\beta, \frac{\sigma^2}{n}Q^{-1}\tilde{Q}Q^{-1})$ where $\text{plim}\frac{1}{n}(X'X) = Q$ and $\text{plim}\frac{1}{n}(X'\Omega X) = \tilde{Q}$.

2.2 White's Estimator of Var(b)

If we continue to use OLS, we need a good estimate of $var(b) = \sigma^2(X'X)^{-1}X'\Omega X(X'X)^{-1}$. White (1980) suggests that if we don't know the form of Ω, we can still find a consistent estimate of $X'\Omega X$, that is,

$$S_0 = \frac{1}{n}\sum\nolimits_{i=1}^{n} e_i^2 x_i x_i'$$

will converge in probability to $\frac{\sigma^2}{n}X'\Omega X$, where the e_i are the OLS residuals. Therefore, White's asymptotic estimate of $var(b)$ is

$$est.asy.var(b) = (X'X)^{-1}nS_0(X'X)^{-1}.$$

Davidson and McKinnon have shown that White's estimator can be unreliable in small samples and have suggested appropriate modifications.

2.3 Gauss Example

In this application, we are interested in measuring the degree of technical inefficiency of rice farmers in the Ivory Coast. The data are both cross-sectional ($N = 154$ farmers) and time series ($T = 3$ years). The model is

$$\ln(1/TE) = \alpha + X\beta + Z\gamma + \epsilon$$

where TE represents technical efficiency (i.e., ratio of actual production to the efficient level from a production frontier), X is a set of managerial variables (e.g., years of experience, gender, age, education, etc.) and Z is a set of exogenous variables (i.e., erosion, slope, weed density, pests, region dummies, year dummies, etc.). The main point of the exercise is to see whether technical inefficiency is related to the managerial characteristics of the rice farmers, once we have accounted for aspects of the production process outside their control. See Gauss example 1 for further details.

3 Testing for Heteroscedasticity

All the tests below are based on the OLS residuals. This makes sense, at least asymptotically, because $b \xrightarrow{p} \beta$.

3.1 Graphical Test

As a first step, it may be useful to graph e_i^2 or e_i against any variable suspected of being related to the heteroscedasticity. If you are unsure which variable is responsible, you can plot against $\hat{Y}_i = X_i b$, which is simply a weghted sum of all X.

3.2 White's Test

The advantage of White's test for heteroscedasticity (and similarly White's estimator of $var(b)$) is that you do not need to know the specific form of Ω. The null hypothesis is H_0: $\sigma_i^2 = \sigma^2, \forall i$ and the alternative is that the null is false. The motivation for the test is that if the null is true $s^2(X'X)^{-1}$ and $s^2(X'X)^{-1}X'\Omega X(X'X)^{-1}$

are both consistent estimators of $var(b)$, while if the null is false, the two estimates will diverge. The test procedure is

- Regress e_i^2 on all the crosses and squares of X. The test statistic is $W = nR^2 \overset{asy}{\sim} \chi^2(P-1)$, where P is the total number of regressors, including the constant.

The disadvantage of the test is that since it is so general, it can easily detect other sorts of misspecifications other than heteroscedasticity. Also the test in nonconstructive, in the sense that once heteroscedasticity is found, the test does not provide guidance in how to find an optimal estimator.

3.3 Goldfeld-Quandt Test

The Goldfeld-Quandt test addresses the disadvantage of White's test. It is a more powerful test that assumes the sample can be divided into two groups – one with a low error variance and the other with a high error variance. The trick is to find the variable on which to sort the data. The hypotheses are

$$\begin{aligned} H_0 &: \quad \sigma_i^2 = \sigma^2, \forall i \\ H_A &: \quad \sigma_n^2 \leq \sigma_{n-1}^2 \leq ... \leq \sigma_1^2 \end{aligned}$$

The test procedure is

1. Order the observations in ascending order according to the size of the error variances.
2. Omit r central observations (often $r = n/3$).
3. Run two separate regressions – first $(n-r)/2$ observations and last $(n-r)/2$ observations.
4. Form the statistic $F = (e_1'e_1/(n_1-k))/(e_2'e_2/(n_2-k)) \sim F(n_1-k, n_2-k)$, which requires that $\epsilon \sim N(0, \sigma^2\Omega)$.
5. Reject or fail to reject the null hypothesis.

3.4 Breusch-Pagan Test

One drawback of the Goldfeld-Quandt test is that you need to choose only one variable related to the heteroscedasticity. Often there are many candidates. The Breusch-Pagan test allows you to choose a vector, z_i, of variables causing the heteroscedasticity. The hypotheses are

$$\begin{aligned} H_0 &: \quad \sigma_i^2 = \sigma^2, \forall i \\ H_A &: \quad \sigma_i^2 = \sigma^2 f(\alpha_0 + \alpha' z_i). \end{aligned}$$

The test statistic is

$$LM = \frac{g'Z(Z'Z)^{-1}Z'g}{2} \overset{asy}{\sim} \chi^2(P-1)$$

where $g_i = (e_i^2/\hat{\sigma}^2) - 1$ and $Z_i = (1, z_i)$. If Z the regressors from White's test, then the two tests are algebraically equivalent.

3.5 Gauss Example (cont.)

We now perform the three tests for heteroscedasticity using the Ivory Coast rice-farming data. The Goldfeld-Quandt test will not work because after sorting, the smaller X matrix is not of full rank. White's test will not work either because there are too many variables. See Gauss example 2 for the results from the Breusch-Pagan test.

4 Generalized Least Squares

4.1 Ω is Known

Assume that the variance-covariance matrix of the errors is known (apart from the scalar σ^2) and is given by (2). We learned that the efficient estimator is

$$\begin{aligned} \hat{\beta} &= (X'\Omega^{-1}X)^{-1}(X'\Omega^{-1}Y) \\ &= (X'P'PX)^{-1}(X'P'PY) \end{aligned}$$

where $P\Omega P' = I$ and

$$P = \begin{bmatrix} 1/\sigma_1 & 0 & \cdots & 0 \\ 0 & 1/\sigma_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1/\sigma_n \end{bmatrix}.$$

GLS can be interpreted as "weighted least squares" because the transformation matrix P weights every observation by the inverse of its error standard deviation. Therefore, observations with the most inherent uncertainty get the smallest weight.

- Example. Let the model be

  $$Y_i = \beta X_i + \epsilon_i$$

  where

  $$\sigma_i^2 = \sigma^2 X_i^2.$$

The GLS estimator is therefore

$$\hat{\beta} = \frac{\sum_i \frac{1}{X_i^2} X_i Y_i}{\sum_i \frac{1}{X_i^2} X_i^2} = \frac{1}{n} \sum_i Y_i / X_i$$

or the average y-x ratio.

4.2 Ω is Unknown

There are too many σ_i^2 elements to estimate with a sample size equal to n. Therefore, we need to restrict σ_i^2 so that it is a function of a smaller number of parameters (e.g., $\sigma_i^2 = \sigma^2 X_i^2$ or $\sigma_i^2 = f(\alpha' z_i)$).

4.2.1 Two-Step Estimation

Since Ω is unknown, we need to estimate it. Let's refer to

$$\hat{\beta}_{FGLS} = (X'\hat{\Omega}^{-1}X)^{-1}(X'\hat{\Omega}^{-1}Y)$$

as the feasible GLS estimator. Consider the following two-step procedure for calculating $\hat{\beta}_{FGLS}$:

1. Estimate the regression model $e_i^2 = f(\alpha' z_i) + \upsilon_i$. Use $\hat{\alpha}$ to obtain the estimates $\hat{\sigma}_i^2 = f(\hat{\alpha}' z_i)$.
2. Calculate $\hat{\beta}_{FGLS}$.

Provided $\hat{\alpha}$ is a consistent estimate of α in step #1, then $\hat{\beta}_{FGLS}$ will be asymptotically efficient at step #2. It may be possible to iterate steps #1 and #2 further, but nothing is gained asymptotically. Sometimes it may be necessary to transform the regression model in step #1 (e.g., take natural logs of $\sigma_i^2 = \exp(\alpha' z_i)$).

4.2.2 Maximum Likelihood Estimation

Write the heteroscedasticity generally as $\sigma_i^2 = \sigma^2 f_i(\alpha)$. The (normal) log likelihood function is

$$\ln L(\beta, \sigma^2, \alpha) = -\frac{n}{2}(\ln(2\pi) + \ln(\sigma^2)) - 0.5 \sum_{i=1}^{n} [\ln f_i(\alpha) + \frac{1}{\sigma^2} \frac{(y_i - x_i'\beta)^2}{f_i(\alpha)}].$$

The first-order conditions are

$$\frac{\partial \ln L}{\partial \beta} = \frac{1}{2\sigma^2} \sum_{i=1}^{n} \frac{2x_i y_i - 2x_i x_i'\beta}{f_i(\alpha)} = 0 \Longrightarrow \sum_{i=1}^{n} \frac{x_i \epsilon_i}{f_i(\alpha)} = 0 \quad (3)$$

$$\frac{\partial \ln L}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^{n} \frac{\epsilon_i^2}{f_i(\alpha)} = 0 \Longrightarrow \sigma^2 = \frac{1}{n} \sum_{i=1}^{n} \frac{\epsilon_i^2}{f_i(\alpha)} \quad (4)$$

$$\frac{\partial \ln L}{\partial \alpha} = -\frac{1}{2} \sum_{i=1}^{n} \frac{g_i(\alpha)}{f_i(\alpha)} + \frac{1}{2\sigma^2} \sum_{i=1}^{n} \frac{\epsilon_i^2 g_i(\alpha)}{f_i(\alpha)^2} = 0 \quad (5)$$

where $g_i(\alpha) = \partial f_i(\alpha)/\partial\alpha$. Notice that equation (3) gives the normal equation for GLS. Solving equations (3) through (5) jointly for $\theta = \{\beta, \sigma^2, \alpha\}$ will produce the maximum likelihood estimates of the model. This can be accomplished in a couple of different ways.

1. Brute force. Use one of the nonlinear optimization algorithms (e.g., Newton-Raphson) to maximize the likelihood function.

2. Oberhofer and Kmenta two-step estimator. Start with a consistent estimator of β. Use that estimate to obtain estimates of σ^2 and α. Iterate back and forth until convergence.

The (efficient) asymptotic ML variance is given by the negative inverse of the information matrix

$$asy.var.(\hat{\theta}_{ML}) = -E[\frac{\partial^2 \ln L}{\partial\theta\partial\theta'}]^{-1}$$

and is given as equation (11-21) in Greene. If this matrix is not working well in the nonlinear optimization algorithm or is not invertible, one could simply use the negative inverse Hessian (without expectations) or the outer product of the gradients (OPG).

4.3 Model Based Test for Heteroscedasticity

As a final note, rather than use the OLS residuals to test for heteroscedasticity, one could test the null hypothesis H_0: $\alpha = 0$ using one of the classical asymptotic tests. For example, the likelihood ratio test would use

$$LR = -2[\ln(L_R) - \ln(L_U)] \overset{asy}{\sim} \chi^2(J)$$

where L_R is the likelihood value with homoscedasticity imposed (i.e., $\alpha = 0$) and L_U is the likelihood value allowing for heteroscedasticity (i.e., $\alpha \neq 0$).

4.4 Gauss Application (cont.)

Using the Ivory Coast rice-farming example, we now calculate feasible GLS and ML estimates of β and γ. The heteroscedasticity is assumed to follow $\sigma_i^2 = \sigma^2 \exp(\alpha' z_i)$, where $z_i = (1, region1_i, region2_i)$. See Gauss example 3 for further details.