Chapter 8
Heteroscedasticity

The fourth assumption in the estimation of the coefficients via ordinary least squares is the one of homoscedasticity. This means that the error terms u_i in the linear regression model have a constant variance across all observations i,

$$\sigma_{u_i}^2 = \sigma_u^2 \quad \text{for all } i. \tag{8.1}$$

When this assumption does not hold, and $\sigma_{u_i}^2$ changes across i we say we have an heteroscedasticity problem. This chapter discusses the problems associated with heteroscedastic errors, presents some tests for heteroscedasticity and points out some possible solutions.

8.1 Heteroscedasticity and its implications

What happens if the errors are heteroscedasticity? The good news is that under heteroscedastic errors, OLS is still unbiased. The bad news is that we will obtain the incorrect standard errors of the coefficients. This means that the t and the F tests that we discussed in earlier chapters are no longer valid. Figure 8.1 shows the regression equation `wage` = β_0 + β_1`educ` + u with heteroscedastic errors. The variance of u_i increases with higher values of `educ`.

8.2 Testing for heteroscedasticity

8.2.1 Breusch-Pagan test

Given the linear regression model

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots + \beta_K + u \tag{8.2}$$

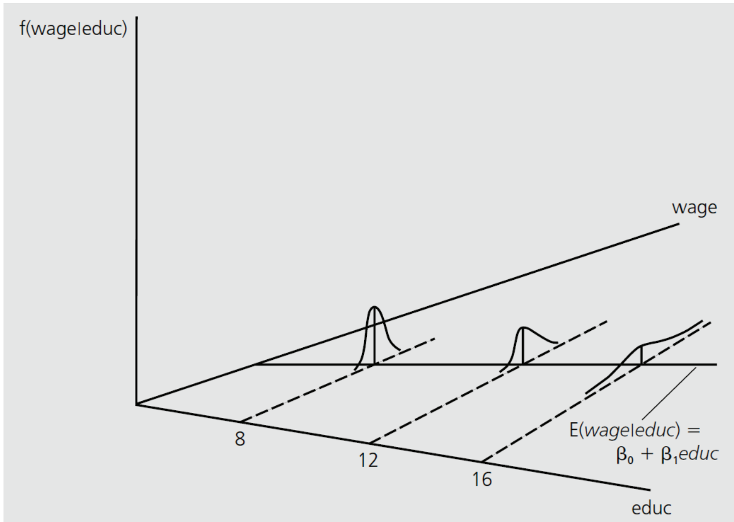

Fig. 8.1 `wage` = β_0 + β_1`educ` + u with heteroscedastic errors.

we know that OLS is unbiased and consistent if we assume $E[u|X_2, X_3, \ldots, X_K] = 0$. Let the null hypothesis that we have homoscedastic errors be

$$H_0 : Var[u|X_2, X_3, \ldots, X_K] = \sigma^2. \tag{8.3}$$

Because we are assuming that u has zero conditional expectation, $\mathrm{Var}[u|X_2, X_3, \ldots, X_K] = E[u^2|X_2, X_3, \ldots, X_K]$, and so the null hypothesis of homoscedasticity is equivalent to

$$H_0 : E[u^2|X_2, X_3, \ldots, X_K] = \sigma^2. \tag{8.4}$$

This shows that if we want to test for violation of the homoscedasticity assumption, we want to test whether $E[u^2|X_2, X_3, \ldots, X_K]$ is related to one or more of the independent variables. If H_0 is false, $E[u|X_2, X_3, \ldots, X_K]$ can be any function of the independent variables. A simple approach is to assume a linear function

$$u^2 = \delta_1 + \delta_2 X_1 + \delta_3 X_3 + \cdots + \delta_K X_K + \varepsilon, \tag{8.5}$$

where ε is an error term with mean zero given X_2, X_3, $\ldots$, X_K. The null hypothesis for homoscedasticity is:

$$H_0 : \delta_1 = \delta_2 = \delta_3 = \cdots = \delta_K = 0. \tag{8.6}$$

Under the null, it is reasonable to assume that ε is independent of $X_2, X_3, \ldots, X_K$. To be able to implement this test, we follow a two step procedure. In the first step we estimate Equation 5.14 via OLS. We estimate the residuals, square them and then estimate the following equation:

$$\hat{u}^2 = \delta_1 + \delta_2 X_1 + \delta_3 X_3 + \cdots + \delta_K X_K + error. \tag{8.7}$$

We can then easily compute the F statistic for the joint significance of all variables X_2, X_3, $\ldots$, X_K. Using OLS residuals in place of the errors does not affect the large sample distribution of the F statistic. An additional LM statistic to test for heteroscedasticity can be constructed based on the $R^2_{\hat{u}^2}$ obtained from Equation 8.7:

$$LM = n \cdot R^2_{\mathrm{u}^2}. \tag{8.8}$$

Under the null hypothesis, LM is distributed asymptotically as χ^2_{K-1}. This LM version of the test is called the Breusch-Pagan test for heteroscedasticity.

8.2.2 Breusch-Pagan test in Gretl

As an example, consider once again our wage equation

$$\texttt{wage} = \beta_1 + \beta_2 \texttt{educ} + u \tag{8.9}$$

Once we estimated the model in Gretl

$$\widehat{\text{wage}} = \underset{(77.715)}{146.952} + \underset{(5.6950)}{60.2143}\,\texttt{educ}$$

$$N = 935 \quad \bar{R}^2 = 0.1060 \quad F(1,933) = 111.79 \quad \hat{\sigma} = 382.32$$

(standard errors in parentheses)

In the regression output window, go to `Tests` → `Heteroskedasticity` → `Breusch-Pagan` to obtain

```
Breusch-Pagan test for heteroskedasticity
OLS, using observations 1-935
Dependent variable: scaled uhat^2

             coefficient   std. error   t-ratio   p-value
  --------------------------------------------------------
  const       -0.885844    0.450097     -1.968    0.0494   **
  educ         0.140019    0.0329833     4.245    2.40e-05 ***

  Explained sum of squares = 88.3581

Test statistic: LM = 44.179066,
with p-value = P(Chi-square(1) > 44.179066) = 0.000000
```

Notice how Gretl reports the auxiliary regression presented in Equation 8.7 and the LM statistic from Equation 8.8. The large LM statistic associated with a small p-value (below 0.05 or 5%) indicates that we reject the null hypothesis of homoscedasticity. Hence, we have heteroscedaticity in the model of Equation 8.9.

8.2.3 White test

White (1980) proposed a test for heteroscedasticity that that adds the squares and cross products of all the independent variables to Equation 8.2. In a model with only three independent variables, the White test is based on the estimation of:

$$\begin{aligned}\hat{u}^2 = {} & \delta_1 + \delta_2 X_2 + \delta_3 X_3 + \delta_4 X_4 + \delta_5 X_2^2 + \delta_6 X_3^2 + \delta_7 X_4^2 \\ & \delta_8 X_2 \cdot X_3 + \delta_9 X_2 \cdot X_4 + \delta_{10} X_3 \cdot X_4 + error.\end{aligned} \tag{8.10}$$

Compared with the Breusch-Pagan test (see Equation 8.7), Equation 8.10 has more regressors. The White test for heteroscedasticity is based on the *LM* statistic for testing that all the δ_j in Equation 8.10 are zero, except for the intercept.

8.2.4 White test in Gretl

We not use Gretl to test for heteroscedasticity in Equation 8.9 using the White test. In the regression output window, go to `Tests` $\rightarrow$ `Heteroskedasticity` $\rightarrow$ `White's test` to obtain

```
White's test for heteroskedasticity
OLS, using observations 1-935
Dependent variable: uhat^2

             coefficient    std. error    t-ratio    p-value
  ----------------------------------------------------------
  const      -126650        435765       -0.2906     0.7714
  educ         20049.7       63065.6      0.3179     0.7506
  sq_educ         13.2563     2234.63     0.005932   0.9953

  Unadjusted R-squared = 0.018950

Test statistic: TR^2 = 17.717812,
with p-value = P(Chi-square(2) > 17.717812) = 0.000142
```

Consistent with the Breusch-Pagan test, here the White test has a large *LM* statistic (labeled `TR^2` following $LM = n \cdot R^2_{\mathrm{u}^2}$ as in Equation 8.8) associated with a small p-value (smaller than 5%). Hence, we reject the null of homoscedasticity and conclude that our model is heteroscedastic.

8.3 What to do with heteroscedasticity?

There is a number of possible solutions when heteroscedastic errors are found. This section proposes three ways to solve the heteroscedasticity problem. First, a simple transformation of the variables; second, the use of weighted least squares; and third, the use of heteroscedasticity-robust standard errors.

8.3.1 Simple transformation of the variables

An easy way to obtain homoscedastic errors is to come up with a simple transformation of the variables. Let's revisit the estimation of Equation 8.9, but this time with a simple logarithm transformation of wages,

$$\log \texttt{wage} = \beta_1 + \beta_2 \texttt{educ} + u \tag{8.11}$$

The Gretl regression output is

$$\widehat{\texttt{logwage}} = \underset{(0.081374)}{5.97306} + \underset{(0.0059631)}{0.0598392}\,\texttt{educ}$$

$$N = 935 \quad \bar{R}^2 = 0.0964 \quad F(1,933) = 100.70 \quad \hat{\sigma} = 0.40032$$

(standard errors in parentheses)

Now, if we want to test for the existence of heteroscedasticity we go to `Tests` $\rightarrow$ `Heteroskedasticity` $\rightarrow$ `Breusch-Pagan` to obtain

```
Breusch-Pagan test for heteroskedasticity
OLS, using observations 1-935
Dependent variable: scaled uhat^2

             coefficient   std. error   t-ratio   p-value
  --------------------------------------------------------
  const       0.689778     0.329663     2.092     0.0367  **
  educ        0.0230332    0.0241578    0.9534    0.3406

  Explained sum of squares = 2.391

Test statistic: LM = 1.195499,
with p-value = P(Chi-square(1) > 1.195499) = 0.274223
```

Notice that the p-value associated with this test is above 0.05. Hence, we fail to reject the null of homoscedasticity. Compare this homoscedasticity results with the heteroscedastic errors found earlier when the variable `wage` was not in logs.

8.3.2 Weighted Least Squares

We want to estimate the following regression model

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots + \beta_K X_K + u, \tag{8.12}$$

but the errors u are heteroscedastic. When one is willing to assume that the heteroscedasticity appears as some function of X_2, X_3, $\ldots$, X_K, one can use Weighted Least Squares (WLS) to obtain homoscedastic errors. Let's say that the variance of u can be approximated using

$$u^2 = \sigma^2 \exp(\delta_1 + \delta_2 X_2 + \delta_3 X_3 + \cdots + \delta_K X_K)\eta, \tag{8.13}$$

where η is a random variable with mean equal to unity. If we assume that η is independent from $X_2, X_3, \ldots, X_K$ we have

$$\log(u^2) = \alpha_1 + \delta_2 X_2 + \delta_3 X_3 + \cdots + \delta_K X_K + \varepsilon. \tag{8.14}$$

To be able to implement this procedure, we replace the unobserved u with the OLS estimated residuals $\hat{u}$ to estimate:

$$\log(\hat{u}^2) = \alpha_1 + \delta_2 X_2 + \delta_3 X_3 + \cdots + \delta_K X_K + \varepsilon. \tag{8.15}$$

Finally, once Equation 8.15 is estimated, we obtain the fitted values and calculate the exponent to obtain

$$\hat{h}_i = \exp(\widehat{\log(\hat{u}^2)}). \tag{8.16}$$

We can use this $\hat{h}_i$ as a weight in a Weighted Least Squares regression to solve the heteroscedasticity problem. That is, we estimate the following weighted equation

$$\frac{Y}{\hat{h}} = \beta_1 \frac{1}{\hat{h}} + \beta_2 \frac{X_2}{\hat{h}} + \beta_3 \frac{X_3}{\hat{h}} + \cdots + \beta_K \frac{X_K}{\hat{h}} + \frac{u}{\hat{h}}. \tag{8.17}$$

Notice that Equation 8.17 is just Equation 8.12 divided by the weight $\hat{h}_i$. The new error term $u/\hat{h}$ should be homoscedastic.

8.3.3 Weighted Least Squares in Gretl

Consider the following model

$$\texttt{sav} = \beta_1 + \beta_2 \texttt{inc} + u \tag{8.18}$$

where `sav` is savings and `inc` is income. The Gretl output is

```
Model 1: OLS, using observations 1-100
Dependent variable: sav

             coefficient   std. error    t-ratio   p-value
  --------------------------------------------------------
  const      124.842       655.393       0.1905    0.8493
  inc          0.146628      0.0575488   2.548     0.0124  **

Mean dependent var   1582.510   S.D. dependent var   3284.902
Sum squared resid    1.00e+09   S.E. of regression   3197.415
R-squared            0.062127   Adjusted R-squared   0.052557
F(1, 98)             6.491778   P-value(F)           0.012391
Log-likelihood      -947.8935   Akaike criterion     1899.787
Schwarz criterion    1904.997   Hannan-Quinn         1901.896
```

and the Breusch-Pagan test for heteroscedasticity yields

```
Breusch-Pagan test for heteroskedasticity
OLS, using observations 1-100
Dependent variable: scaled uhat^2

             coefficient   std. error    t-ratio   p-value
  --------------------------------------------------------
  const      0.0457266     1.14381       0.03998   0.9682
  inc        9.59914e-05   0.000100436   0.9557    0.3416

  Explained sum of squares = 28.444

Test statistic: LM = 14.221987,
with p-value = P(Chi-square(1) > 14.221987) = 0.000162
```

That is, we have heteroscedastic errors.

To estimate the WLS regression

$$\frac{\texttt{sav}}{\hat{h}} = \beta_1 \frac{1}{\hat{h}} + \beta_2 \frac{\texttt{inc}}{\hat{h}} + \frac{u}{\hat{h}}, \tag{8.19}$$

in the Gretl main window we have to go to `Model` → `Other linear models` → `Heteroskedasticity corrected` to get the following computer output

```
Model 2: Heteroskedasticity-corrected, using observations 1-100
Dependent variable: sav

             coefficient   std. error    t-ratio   p-value
  --------------------------------------------------------
  const      -233.130      460.844       -0.5059   0.6141
  inc           0.185993     0.0616965    3.015    0.0033  ***

Statistics based on the weighted data:

Sum squared resid    1043.864   S.E. of regression   3.263689
R-squared            0.084866   Adjusted R-squared   0.075527
F(1, 98)             9.088089   P-value(F)           0.003276
Log-likelihood      -259.1695   Akaike criterion     522.3391
Schwarz criterion    527.5494   Hannan-Quinn         524.4478

Statistics based on the original data:

Mean dependent var   1582.510   S.D. dependent var   3284.902
Sum squared resid    1.01e+09   S.E. of regression   3205.216
```

$$\widehat{\texttt{sav}} = \underset{(460.84)}{-233.130} + \underset{(0.061697)}{0.185993}\,\texttt{inc}$$

$$N = 100 \quad \bar{R}^2 = 0.0755 \quad F(1,98) = 9.0881 \quad \hat{\sigma} = 3.2637$$

(standard errors in parentheses)

The estimates of the standard errors can now be used for inferences. The statistically significant coefficient on `inc` indicates that the marginal propensity to save out of your income is 0.18. Of every additional dollar that you make, you will save 18 cents.

8.3.4 White's heteroscedasticity-consistent standard errors

Even under the presence of heteroscedastic errors, at least in large samples a consistent estimator of the variances of the coefficients can be obtained via White's heteroscedasticity-consistent standard errors. This procedure leaves the OLS coefficients unaffected. For the estimation of Equation 8.18 you just have to make sure to select the option `Robust standard errors` in the Gretl "specify model" window when you estimate the model via OLS

```
Model 3: OLS, using observations 1-100
Dependent variable: sav
Heteroskedasticity-robust standard errors, variant HC1

             coefficient   std. error    t-ratio   p-value
  ---------------------------------------------------------
  const      124.842       528.219       0.2363    0.8137
  inc          0.146628      0.0613441   2.390     0.0187  **

Mean dependent var   1582.510   S.D. dependent var   3284.902
Sum squared resid    1.00e+09   S.E. of regression   3197.415
R-squared            0.062127   Adjusted R-squared   0.052557
F(1, 98)             5.713342   P-value(F)           0.018748
Log-likelihood      -947.8935   Akaike criterion     1899.787
Schwarz criterion    1904.997   Hannan-Quinn         1901.896
```

Notice that the constant and slope coefficients are the same as before. However, the estimated standard errors are different.