8.3 Testing for Heteroskedasticity

1. Do we really need to know whether there is heteroskedasticity to implement the methods? *No*, but it helps to have a method.

2. Idea: We want to test whether the regression error, conditional on the x's is homoskedastic or heteroskedastic. Let's start with a model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + u$$

Under MLR.1 to MLR.4 we have consistent and unbiased estimates. We add MLR.5

$$H_0 : Var(u|x_1, x_2, x_3, \ldots) = \sigma^2$$

3. If H_0 is false, then the conditional expectation function of u can be any function of any of the x_k. Let us assume it is a linear function of the following form:

$$u^2 = \delta_0 + \delta_1 x_1 + \delta_2 x_2 + \cdots + \delta_k x_k + v$$

4. Null hypothesis under homoskedasticity implies:

$$H_0 : \delta_0 = \delta_1 = \ldots = \delta_k = 0$$

5. But we dont know the true u^2 so we use the sample proxy to obtain:

$$\hat{u}^2 = \delta_0 + \delta_1 x_1 + \delta_2 x_2 + \cdots + \delta_k x_k + v$$

6. The F statistic is:

$$F = \frac{R^2_{\hat{u}^2}/k}{(1 - R^2_{\hat{u}^2})/(n - k - 1)}$$

where k are the number of regressors in step 5. F stat has approximately $F_{k,n-k-1}$ distribution under the null hypothesis.

7. LM statistic is called the **Breusch-Pagan test** for heteroskedasticity:

$$LM = n \cdot R^2_{\hat{u}^2}$$

8. **Form of heteroskedasticity**: If we think that variance of the error term depends upon only some of the x^s , we repeat step 5 but only on selected x and construct the statistic.

8.3.1 White Test for Heteroskedasticity

1. Idea: White suggested to test different form of heteroskedasticity by including squared terms and cross products in the following equation:

$$\begin{aligned}\hat{u}^2 =& \delta_0 + \delta_1 x_1 + \delta_2 x_2 + \delta_3 x_3 + \delta_4 x_1^2 + \delta_5 x_2^2 + \delta_6 x_3^2 \\ &+ \delta_7 x_1 x_2 + \delta_8 x_1 x_3 + \delta_9 x_2 x_3 + v\end{aligned}$$

2. The null hypothesis for heteroskedasticity implies that all δ's are zero except δ_0, thus 9 exclusion restrictions.

   Naturally, it uses a lot of degrees of freedom even with a small size of the model.

We could alternatively use OLS fitted values to get around the degree of freedom issue by running the following model:

$$\begin{aligned}\hat{u}^2 =& \delta_0 + \delta_1 \hat{y} + \delta_2 \hat{y}^2 + v \\ \hat{y} =& \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3\end{aligned}$$

In order to test heteroskedasticity, we merely test whether δ_0 and δ_1 are zeros or not and conserve degrees of freedom.

3. **Caution**: If MLR.4 is violated - in particular if the functional form of $E(y|\mathbf{x})$ is misspecified - then a test for heteroskedasticity can reject the null H_0 even if $Var(y|\mathbf{x})$ is homoskedastic.

8.4 Weighted Least Square Estimation

1. **Idea**: If you could not certainly tell whether there is heteroskedasticity, then you could alternatively weight your OLS estimates according to the form of heteroskedasticity.

8.4.1 Heteroskedasticity is Known to a Multiplicative Constant

Consider the scenario when you know the functional form $h(\mathbf{x})$ of conditional variance with an unknown population variance σ^2

$$Var(u|\mathbf{x}) = \sigma^2 h(\mathbf{x})$$

For instance, you want to know the effect of individual savings on income, but you suspect that conditional error variance is linearly related to income of the individual.

$$sav_i = \beta_0 + \beta_i inc_i + u_i \qquad Var(u_i|inc_i) = \sigma^2 inc_i$$

Idea: We use our knowledge of the conditional variance and transform the original regression

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + u_i$$

into one with homoskedastic errors.

We do this by dividing $\frac{u_i}{\sqrt{h_i}}$ such that the conditional variance of this transformed error is

$$E\left[\frac{u_i}{\sqrt{h_i}}\right] = \frac{\sigma^2 h_i}{h_i} = \sigma^2$$

We obtain this transformed error from the transformed regression of the following form (remember chapter 6)

$$\frac{y_i}{\sqrt{h_i}} = \frac{\beta_0}{\sqrt{h_i}} + \frac{\beta_1}{\sqrt{h_i}} x_{i1} + \frac{\beta_2}{\sqrt{h_i}} x_{i2} + \ldots + \frac{\beta_k}{\sqrt{h_i}} x_{ik} + u_i$$

In our savings example:

$$\frac{sav_i}{\sqrt{inc_i}} = \frac{\beta_0}{\sqrt{inc_i}} + \frac{\beta_1 inc_i}{\sqrt{inc_i}} + \frac{u_i}{\sqrt{inc_i}}$$

This is an example of the **Generalized Least Square estimation of a model** whose original version had heteroskedasticity. The transformed equation satisfies all the Gauss-Markov assumptions

1. **MLR.1** : linear in parameters. The estimated parameters in this case are $\beta_0^* = \frac{\beta_0}{\sqrt{inc_i}}$. The model is linear in the transformed parameters.

2. **MLR.2**: Random sampling merely implies that when we are recording income and savings of people, they are representative of the population.

3. **MLR.3**: No perfect collinearity which in this case comes automatically cause there is no other variable.

4. **MLR.4**: Zero conditional mean implies that the transformed error from the regression has zero mean. Since its from the regression, its already conditioned on income.

5. **MLR.5**: Homoskedasticity i.e. $Var(u_i|\mathbf{x}_i) = \sigma^2$

6. **MLR.6**: Normality i.e. $u_i \ N(0, \sigma^2)$

Does it help inference? Yes since the transformed estimates follow all the MLR assumptions. But interpret the original estimates. This particular GLS is the *weighted least square* since the weights in this case are $\frac{1}{h_i}$.

Idea: WLS gives less weight to observations with higher variance. A weighted least squares estimator can be defined for any set of positive weights. OLS is the special case that gives equal weight to all observations.

Best procedure: is when you weight each squared residual by the inverse of the conditional variance of u_i given $\mathbf{x}_i$

Example 8.6: Financial Wealth Equation

Consider an ordinary least square estimation for the following equation:

$$nettfa = \beta_0 + \beta_1 inc + u$$

```
> summary(lm(nettfa ~ inc, k401ksubs))

Call:
lm(formula = nettfa ~ inc, data = k401ksubs)

Residuals:
    Min      1Q  Median      3Q     Max
-504.39  -18.10   -4.29    6.73 1475.04

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -20.17948    1.17643  -17.15   <2e-16 ***
inc           0.99991    0.02554   39.15   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 59.26 on 9273 degrees of freedom
Multiple R-squared:  0.1418,    Adjusted R-squared:  0.1417
F-statistic:  1532 on 1 and 9273 DF,  p-value: < 2.2e-16
```

If we use data on only single people to know their motivations for savings, we do the following:

```
> mod = lm(nettfa ~ inc, data = k401ksubs, subset = (fsize==1))
> summary(mod)

Call:
lm(formula = nettfa ~ inc, data = k401ksubs, subset = (fsize ==
    1))

Residuals:
    Min      1Q  Median      3Q     Max
-185.12  -12.85   -4.85    1.78 1112.66

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -10.5709     2.0607   -5.13 3.18e-07 ***
inc           0.8207     0.0609   13.48  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 45.59 on 2015 degrees of freedom
Multiple R-squared:  0.08267,   Adjusted R-squared:  0.08222
F-statistic: 181.6 on 1 and 2015 DF,  p-value: < 2.2e-16
```

Now we can report heteroskedasticity consistent standard errors instead of the usual standard errors if we suspect the possibility of heteroskedasticity.

```
> coeftest(mod, vcov = vcovHC(mod, type="HC1"))

t test of coefficients:

             Estimate Std. Error t value  Pr(>|t|)
(Intercept) -10.57095    2.53027 -4.1778 3.069e-05 ***
inc           0.82068    0.10359  7.9221 3.826e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Alternatively we can run a weighted least squares regression if we

```
> mod2 = lm(nettfa ~ inc, data = k401ksubs, subset = (fsize==1), weights = 1/inc)
> summary(mod2)

Call:
lm(formula = nettfa ~ inc, data = k401ksubs, subset = (fsize ==
    1), weights = 1/inc)

Weighted Residuals:
    Min      1Q  Median      3Q     Max
-23.469  -2.339  -1.086   0.352 178.220

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -9.58070    1.65328  -5.795 7.91e-09 ***
inc          0.78705    0.06348  12.398  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.219 on 2015 degrees of freedom
Multiple R-squared:  0.07088,   Adjusted R-squared:  0.07042
F-statistic: 153.7 on 1 and 2015 DF,  p-value: < 2.2e-16
```

We suspect that savings to 401K start only after individuals reach a certain age after which such contributions increase. We suspect a quartratic relation for all age range which increases with an upward slope after age 25.

```
> mod3 = lm(nettfa ~ inc + I((age - 25) * (age-25)) + male + e401k, data = k4
01ksubs, subset = (fsize==1))
> summary(mod3)

Call:
lm(formula = nettfa ~ inc + I((age - 25) * (age - 25)) + male +
    e401k, data = k401ksubs, subset = (fsize == 1))

Residuals:
    Min      1Q  Median      3Q     Max
-176.04  -14.17   -3.15    6.01 1111.42

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)               -20.984990   2.472022  -8.489   <2e-16 ***
inc                         0.770583   0.061452  12.540   <2e-16 ***
I((age - 25) * (age - 25))  0.025127   0.002593   9.689   <2e-16 ***
male                        2.477927   2.047776   1.210   0.2264
e401k                       6.886223   2.123275   3.243   0.0012 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 44.49 on 2012 degrees of freedom
Multiple R-squared:  0.1279,    Adjusted R-squared:  0.1261
F-statistic: 73.75 on 4 and 2012 DF,  p-value: < 2.2e-16
```

A WLS version of the expanded model is

```
> mod4 = lm(nettfa ~ inc + I((age - 25) * (age-25)) + male + e401k, data = k4
01ksubs, subset = (fsize==1), weights = 1/inc)
> summary(mod4)

Call:
lm(formula = nettfa ~ inc + I((age - 25) * (age - 25)) + male +
    e401k, data = k401ksubs, subset = (fsize == 1), weights = 1/inc)

Weighted Residuals:
    Min      1Q  Median      3Q     Max
-26.613  -2.491  -0.803   0.934 178.052

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                -16.702521   1.957995  -8.530  < 2e-16 ***
inc                          0.740384   0.064303  11.514  < 2e-16 ***
I((age - 25) * (age - 25))   0.017537   0.001931   9.080  < 2e-16 ***
male                         1.840529   1.563587   1.177  0.23929
e401k                        5.188281   1.703426   3.046  0.00235 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.065 on 2012 degrees of freedom
Multiple R-squared:  0.1115,    Adjusted R-squared:  0.1097
F-statistic: 63.13 on 4 and 2012 DF,  p-value: < 2.2e-16
```

Two cases when WLS works better than OLS:

1. Consider a model of employee (e) contribution to a savings fund in a firm (i). In this case, even with homoskedastic errors if you dont have employee level data for all firms, then you could work with average contribution of employees in one firm to examine cross firm behavior. But running an OLS would be wrong cause the error in the second model are not homoskedastic:

$$contrib_{ie} = \beta_0 + \beta_1 earns_{ie} + \beta_2 age_{ie} + \beta_3 mrate_{ie} + u$$
$$\bar{contrib}_{ie} = \beta_0 + \beta_1 \bar{earns}_{ie} + \beta_2 \bar{age}_{ie} + \beta_3 \bar{mrate}_{ie} + u$$

2. Similar issue arises when 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, weighted least squares with weights equal to the population is appropriate.

If heteroskedasticity exists at the individual level, then the proper weighting depends on the form of heteroskedasticity