Class 25 - MRM: Heteroskedasticity (part II)

Pedram Jahangiry

JON M. HUNTSMAN SCHOOL OF BUSINESS
UtahStateUniversity

Remedies for Heteroskedasticity

- If heteroskedasticity is found, the first thing to do is examine the equation carefully for specification errors.

- If there are no obvious specification errors, the heteroskedasticity is probably pure in nature and one of the following remedies should be considered.

1. Redefining the Variables
2. Heteroskedasticity-Corrected Standard Errors
3. Weighted Least Square Estimation!

1-Redefining the Variables

- Redefining variables can help avoid heteroskedasticity.
- Be careful! Redefining variables is a functional form specification change.
- In some cases, the only redefinition needed is to switch from a linear to logarithmic functional form.
- In some situations, it might be necessary to rethink project in terms of its underlying theory.

2-Heteroskedasticity-Corrected (robust) Standard Errors

- **Heteroskedasticity-corrected (HC) standard errors** are standard errors calculated specifically to avoid consequences of heteroskedasticity.
- HC standard errors are biased but are generally more accurate than uncorrected standard errors in large sample.
- HC standard errors can be used in t-tests and other hypothesis tests.

- Formula for heteroskedasticity-robust OLS standard error

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

Also called White/Huber/Eicker standard errors. They involve the squared residuals from the regression and from a regression of x_j on all other explanatory variables.

Example: Hourly wage equation

EXAMPLE 8.2 **Heteroskedasticity-Robust *F* Statistic**

Using the data for the spring semester in GPA3, we estimate the following equation:

$$\widehat{cumgpa} = 1.47 + .00114\,sat - .00857\,hsperc + .00250\,tothrs$$
$$(.23)\ (.00018)\ (.00124)\ (.00073)$$
$$[.22]\ [.00019]\ [.00140]\ [.00073]$$
$$+ .303\,female - .128\,black - .059\,white$$
$$(.059)\ (.147)\ (.141)$$
$$[.059]\ [.118]\ [.110]$$
$$n = 366,\ R^2 = .4006,\ \overline{R}^2 = .3905.$$

```
# Usual SE:
coeftest(MRM_HC)

# HC SE:
coeftest(MRM_HC, vcov=hccm(MRM_HC,type="hc0"))
```

Why do we have both **black** and **white** dummy variables here?
Any problem?

```
> H0 <- c("black","white")
> linearHypothesis(MRM_HC, H0)
Linear hypothesis test

Hypothesis:
black = 0
white = 0

Model 1: restricted model
Model 2: cumgpa ~ sat + hsperc + tothrs + female + black + white

  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    361 79.362
2    359 79.062  2   0.29934 0.6796 0.5075
```

```
> H0 <- c("black","white")
> linearHypothesis(MRM_HC, H0, vcov=hccm(MRM_HC,type="hc0"))
Linear hypothesis test

Hypothesis:
black = 0
white = 0

Model 1: restricted model
Model 2: cumgpa ~ sat + hsperc + tothrs + female + black + white

Note: Coefficient covariance matrix supplied.

  Res.Df Df      F Pr(>F)
1    361
2    359  2 0.7478 0.4741
```

3-Weighted least squares estimation

Heteroskedasticity is known up to a multiplicative constant $\mathrm{Var}(u|\mathbf{x}) = \sigma^2 h(\mathbf{x}) \quad , h(\mathbf{x}) > 0$

Idea: giving smaller weights to observations with larger variance!

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + u_i$$

$$\Rightarrow \left[\frac{y_i}{\sqrt{h_i}}\right] = \beta_0 \left[\frac{1}{\sqrt{h_i}}\right] + \beta_1 \left[\frac{x_{i1}}{\sqrt{h_i}}\right] + \cdots + \beta_k \left[\frac{x_{ik}}{\sqrt{h_i}}\right] + \left[\frac{u_i}{\sqrt{h_i}}\right]$$

$$\Leftrightarrow \; y_i^* = \beta_0 x_{i0}^* + \beta_1 x_{i1}^* + \cdots + \beta_k x_{ik}^* + u_i^*$$

← Transformed model which is homoskedastic (why?)

The β_j^* are examples of generalized least squares (GLS) estimators. If the other Gauss-Markov assumptions hold as well, WLS is the best linear unbiased estimator (**BLUE**). Observations with a large variance are less informative than observations with small variance and therefore should get less weight.

Estimating the Linear Probability Model by Weighted Least Squares

Recall

$$Var(y|\mathbf{x}) = p(\mathbf{x})\,[1 - p(\mathbf{x})]$$

$$\Rightarrow \quad \hat{h}_i = \hat{y}_i(1 - \hat{y}_i)$$

In the LPM, the exact form of heteroskedasticity is known

Steps:

1. Estimate the model by OLS and obtain the fitted values, $\hat{y}$.
2. Determine whether all of the fitted values are inside the unit interval. If so, proceed to step (3). If not, some adjustment is needed to bring all fitted values into the unit interval.
3. Construct the estimated variances in equation $\hat{h}_i = \hat{y}_i(1 - \hat{y}_i)$
4. Estimate the equation

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + u$$

by WLS, using weights $1/\hat{h}$.

Example 8.9: Determinants of Personal Computer Ownership

The equation estimated by OLS is

$$\widehat{PC} = -.0004 + .065\,hsGPA + .0006\,ACT + .221\,parcoll$$
$$(.4905)\ (.137)\qquad (.0155)\qquad (.093)$$
$$[.4888]\ [.139]\qquad [.0158]\qquad [.087]$$
$$n = 141,\ R^2 = .0415.$$

parcoll is a binary indicator equal to unity if **at least one parent** attended college.

In this example, there are no striking differences between the **usual** and **robust** standard errors. Nevertheless, we also estimate the model by WLS. Because all of the OLS fitted values are inside the unit interval, no adjustments are needed

```
# WLS
y_hat <- predict(MRM_dummy_dep)
range(y_hat)
h_hat <- y_hat *(1-y_hat)
w     <- 1/h_hat
MRM_dummy_dep_wls <- lm(PC ~ hsGPA + ACT + parcoll, weights = w, gpa1_new)
```

How do you construct the HC standard errors?

```
=============================================================
                                Dependent variable:
                            ---------------------------
                                        PC
                                  (1)           (2)
-------------------------------------------------------------
hsGPA                           0.0654        0.0327
                               (0.1373)      (0.1299)

ACT                             0.0006        0.0043
                               (0.0155)      (0.0155)

parcoll                        0.2211**      0.2152**
                               (0.0930)      (0.0863)

Constant                        -0.0004        0.0262
                               (0.4905)      (0.4766)

-------------------------------------------------------------
Observations                      141           141
R2                               0.0415        0.0464
Adjusted R2                      0.0205        0.0256
Residual Std. Error (df = 137)   0.4860        1.0162
F Statistic (df = 3; 137)        1.9785        2.2240*
=============================================================
Note:                          *p<0.1; **p<0.05; ***p<0.01
```

Feasible GLS

Heteroskedasticity function is NOT known and must be estimated.

- in many cases we can model the function $h(x)$ and use the data to estimate the unknown parameters in this model. This results in an estimate of each h_i, denoted as $\hat{h}_i$
- This yields an estimator called the **feasible GLS (FGLS)** estimator.
- The resulting FGLS estimator is no longer unbiased but it is consistent and asymptotically efficient.

There are many ways to model heteroskedasticity, a fairly flexible approach is:

$$\mathrm{Var}(u|\mathbf{x}) = \sigma^2 \exp(\delta_0 + \delta_1 x_1 + \delta_2 x_2 + \ldots + \delta_k x_k)$$

$$\log(u^2) = \alpha_0 + \delta_1 x_1 + \delta_2 x_2 + \ldots + \delta_k x_k + e$$

Feasible GLS

A Feasible GLS Procedure to Correct for Heteroskedasticity:

1. Run the regression of y on $x_1, x_2, \ldots, x_k$ and obtain the residuals, $\hat{u}$.
2. Create $\log(\hat{u}^2)$ by first squaring the OLS residuals and then taking the natural log.
3. Run the regression in equation $\log(u^2) = \alpha_0 + \delta_1 x_1 + \delta_2 x_2 + \ldots + \delta_k x_k + e$ and obtain the fitted values, $\hat{g}$.
4. Exponentiate the fitted values $\hat{h} = \exp(\hat{g})$
5. Estimate the equation

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + u$$

by WLS, using weights $1/\hat{h}$.