

FINM3123 Introduction to Econometrics

Chapter 8

Heteroskedasticity

Multiple Regression Analysis: Heteroskedasticity

Consequences of heteroskedasticity for OLS

- OLS still unbiased and consistent under heteroskedasticity!
- Also, interpretation of R-squared is not changed

$$R^2 \approx 1 - \frac{\sigma_u^2}{\sigma_y^2}$$

Unconditional error variance is unaffected by heteroskedasticity
(which refers to the conditional error variance)

- heteroskedasticity invalidates variance formulas for OLS estimators
- The usual F -tests and t -tests are not valid under heteroskedasticity
- Under heteroskedasticity, OLS is no longer the best linear unbiased estimator (BLUE); there may be more efficient linear estimators

Multiple Regression Analysis: Heteroskedasticity

Heteroskedasticity-robust inference after OLS

- Formulas for OLS standard errors and related statistics have been developed that are robust to heteroskedasticity of unknown form
- These robust formulas are only valid in large samples
- 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/Eicker/Huber standard errors.
They involve the squared residuals from the regression of x_j on all other explanatory variables.

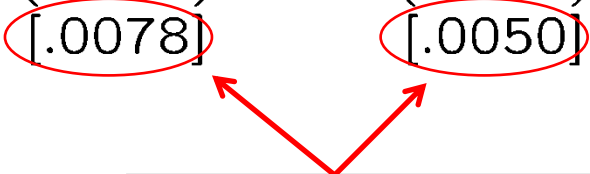
- Using these formulas, the usual t -test is valid asymptotically
- The usual F -statistic does not work under heteroskedasticity, but heteroskedasticity-robust versions are available in most software

Multiple Regression Analysis: Heteroskedasticity

Example: Hourly wage equation

$$\widehat{\log(wage)} = - .128 + .0904 \text{ educ} + .0410 \text{ exper} - .0007 \text{ exper}^2$$

(.105)	(.0075)	(.0052)	(.0001)
[.107]	[.0078]	[.0050]	[.0001]



$$H_0 : \beta_{\text{exper}} = \beta_{\text{exper}^2} = 0$$

heteroskedasticity-robust standard errors may be larger or smaller than their nonrobust counterparts. The differences are often small in practice.

$$F = 17.95$$

F-statistics are also often not too different.



$$F_{\text{robust}} = 17.99$$

If there is strong heteroskedasticity, differences may be larger. To be on the safe side, it is advisable to always compute robust standard errors.

Multiple Regression Analysis: Heteroskedasticity

Testing for heteroskedasticity

- It may still be interesting to test whether there is heteroskedasticity because then OLS may not be the most efficient linear estimator anymore

Breusch-Pagan test for heteroskedasticity

$$H_0 : Var(u|x_1, x_2, \dots, x_k) = Var(u|\mathbf{x}) = \sigma^2$$

$$Var(u|\mathbf{x}) = E(u^2|\mathbf{x}) - [E(u|\mathbf{x})]^2 = E(u^2|\mathbf{x})$$

Under MLR.4

$$\Rightarrow E(u^2|x_1, \dots, x_k) = E(u^2) = \sigma^2$$

The mean of u^2 must not vary with x_1, x_2, \dots, x_k

Multiple Regression Analysis: Heteroskedasticity

Breusch-Pagan test for heteroskedasticity (cont.)

$$\hat{u}^2 = \delta_0 + \delta_1 x_1 + \dots + \delta_k x_k + \text{error}$$

$$H_0 : \delta_1 = \delta_2 = \dots = \delta_k = 0$$

Regress squared residuals on all explanatory variables and test whether this regression has explanatory power.

$$F = \frac{R_{\hat{u}^2}^2/k}{(1 - R_{\hat{u}^2}^2)/(n - k - 1)} \sim F_{k, n-k-1}$$

A large test statistic (= a high R-squared) is evidence against the null hypothesis.

$$LM = n \cdot R_{\hat{u}^2}^2 \sim \chi_k^2$$

Alternative test statistic (= Lagrange multiplier statistic, LM). Again, high values of the test statistic (= high R-squared) lead to rejection of the null hypothesis that the expected value of u^2 is unrelated to the explanatory variables.

Multiple Regression Analysis: Heteroskedasticity

Example: heteroskedasticity in housing price equations

$$\widehat{price} = -21.77 + .0021 \text{ lotsize} + .123 \text{ sqrft} + 13.85 \text{ bdrms}$$

(29.48) (.0006) (.013) (9.01)

heteroskedasticity

$$\Rightarrow R_{\hat{u}^2}^2 = .1601, p\text{-value}_F = .002, p\text{-value}_{LM} = .0028$$

$$\widehat{\log(price)} = -1.30 + .168 \log(lotsize) + .700 \log(sqrft) + .037 \text{ bdrms}$$

(.65) (.038) (.093) (.028)

$$\Rightarrow R_{\hat{u}^2}^2 = .0480, p\text{-value}_F = .245, p\text{-value}_{LM} = .2390$$

In the logarithmic specification, homoskedasticity cannot be rejected

Multiple Regression Analysis: Heteroskedasticity

White test for heteroskedasticity

Regress squared residuals on all explanatory variables, their squares, and interactions (here: example for k=3)

$$\hat{u}^2 = \delta_0 + \delta_1 x_1 + \delta_2 x_2 + \delta_3 x_3 + \delta_4 x_1^2 + \delta_5 x_2^2 + \delta_6 x_3^2 \\ + \delta_7 x_1 x_2 + \delta_8 x_1 x_3 + \delta_9 x_2 x_3 + error$$

$$H_0 : \delta_1 = \delta_2 = \dots = \delta_9 = 0$$

$$LM = n \cdot R_{\hat{u}^2}^2 \sim \chi_9^2$$

The White test detects more general deviations from heteroskedasticity than the Breusch-Pagan test


Disadvantage of this form of the White test:

- Including all squares and interactions leads to a large number of estimated parameters (e.g. k=6 leads to 27 parameters to be estimated)

Multiple Regression Analysis: Heteroskedasticity

Alternative form of the White test

$$\hat{u}^2 = \delta_0 + \delta_1 \hat{y} + \delta_2 \hat{y}^2 + error$$



This regression indirectly tests the dependence of the squared residuals on the explanatory variables, their squares, and interactions, because the predicted value of y and its square implicitly contain all of these terms.

$$H_0: \delta_1 = \delta_2 = 0, \quad LM = n \cdot R_{\hat{u}^2}^2 \sim \chi_2^2$$

Example: heteroskedasticity in (log) housing price equations

$$R_{\hat{u}^2}^2 = .0392, LM = 88(.0392) \approx 3.45, p\text{-value}_{LM} = .178$$

Multiple Regression Analysis: Heteroskedasticity

Weighted least squares estimation

Assume heteroskedasticity is known up to a multiplicative constant

$$Var(u_i|\mathbf{x}_i) = \sigma^2 h(\mathbf{x}_i), \quad h(\mathbf{x}_i) = h_i > 0$$

← The functional form of the heteroskedasticity is known

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

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

$$\Leftrightarrow y_i^* = \beta_0 x_{i0}^* + \beta_1 x_{i1}^* + \cdots + \beta_k x_{ik}^* + u_i^* \quad \leftarrow \text{Transformed model}$$

Multiple Regression Analysis: Heteroskedasticity

Example: Savings and income

$$sav_i = \beta_0 + \beta_1 inc_i + u_i, \quad Var(u_i | inc_i) = \sigma^2 inc_i$$

$$\left[\frac{sav_i}{\sqrt{inc_i}} \right] = \beta_0 \left[\frac{1}{\sqrt{inc_i}} \right] + \beta_1 \left[\frac{inc_i}{\sqrt{inc_i}} \right] + u_i^*$$

Note that this regression model has no intercept

The transformed model is homoskedastic

$$E(u_i^{*2} | \mathbf{x}_i) = E \left[\left(\frac{u_i}{\sqrt{h_i}} \right)^2 | \mathbf{x}_i \right] = \frac{E(u_i^2 | \mathbf{x})}{h_i} = \frac{\sigma^2 h_i}{h_i} = \sigma^2$$

If the other Gauss-Markov assumptions hold as well, OLS applied to the transformed model is the best linear unbiased estimator!

Multiple Regression Analysis: Heteroskedasticity

OLS in the transformed model is called **weighted least squares (WLS)**

$$\min \sum_{i=1}^n \left(\left[\frac{y_i}{\sqrt{h_i}} \right] - b_0 \left[\frac{1}{\sqrt{h_i}} \right] - b_1 \left[\frac{x_{i1}}{\sqrt{h_i}} \right] - \dots - b_k \left[\frac{x_{ik}}{\sqrt{h_i}} \right] \right)^2$$

$$\Leftrightarrow \min \sum_{i=1}^n (y_i - b_0 - b_1 x_{i1} - \dots - b_k x_{ik})^2 / h_i$$

Observations with a large variance get a smaller weight in the optimization problem

Why is WLS more efficient than OLS in the original model?

- Observations with a large variance are less informative than observations with small variance and therefore should get less weight

WLS is a special case of **generalized least squares (GLS)**

Multiple Regression Analysis: Heteroskedasticity

Example: Financial wealth equation

Net financial wealth

Assumed form of heteroskedasticity:

$$\text{Var}(u|inc, age, male, e401k) = \sigma^2 inc$$

WLS estimates have considerably smaller standard errors (which is in line with the expectation that they are more efficient).

Participation in 401K pension plan

TABLE 8.1 Dependent Variable: *netffa*

Independent Variables	(1) OLS	(2) WLS	(3) OLS	(4) WLS
<i>inc</i>	.821 (.104)	.787 (.063)	.771 (.100)	.740 (.064)
$(age - 25)^2$	—	—	.0251 (.0043)	.0175 (.0019)
<i>male</i>	—	—	2.48 (2.06)	1.84 (1.56)
<i>e401k</i>	—	—	6.89 (2.29)	5.19 (1.70)
<i>intercept</i>	-10.57 (2.53)	-9.58 (1.65)	-20.98 (3.50)	-16.70 (1.96)
Observations	2,017	2,017	2,017	2,017
R-squared	.0827	.0709	.1279	.1115

© Cengage Learning, 2013

Multiple Regression Analysis: Heteroskedasticity

Important special case of heteroskedasticity

- If the observations are reported as averages at the city/county/state/country/firm level, they should be weighted by the size of the unit

Average contribution to pension plan in firm i Average earnings and age in firm i Percentage firm contributes to plan heteroskedastic error term

$$\overline{contrib}_i = \beta_0 + \beta_1 \overline{earns}_i + \beta_2 \overline{age}_i + \beta_3 mrate_i + \overline{u}_i$$

$\Rightarrow Var(\overline{u}_i) = Var\left(\frac{1}{m_i} \sum_{e=1}^{m_i} u_{i,e}\right) = \sigma^2 / m_i$ Error variance if errors are homoskedastic at the employee level

If errors are homoskedastic at the employee level, WLS with weights equal to firm size m_i should be used. If the assumption of homoskedasticity at the employee level is not exactly right, one can calculate robust standard errors after WLS (i.e. for the transformed model).

Multiple Regression Analysis: Heteroskedasticity

Unknown heteroskedasticity function (feasible GLS)

$$Var(u|\mathbf{x}) = \sigma^2 \exp(\delta_0 + \delta_1 x_1 + \dots + \delta_k x_k) = \sigma^2 h(\mathbf{x})$$

Assumed general form of heteroskedasticity; the exponential function is used to ensure positivity

$$u^2 = \sigma^2 \exp(\delta_0 + \delta_1 x_1 + \dots + \delta_k x_k) \cdot v$$

Multiplicative error (assumption: independent of the explanatory variables)

$$\Rightarrow \log(u^2) = \alpha_0 + \delta_1 x_1 + \dots + \delta_k x_k + e$$

$$\log(\hat{u}^2) = \hat{\alpha}_0 + \hat{\delta}_1 x_1 + \dots + \hat{\delta}_k x_k + error$$

Use inverse values of the estimated heteroskedasticity function as weights in WLS

$$\Rightarrow \hat{h}_i = \exp(\hat{\alpha}_0 + \hat{\delta}_1 x_1 + \dots + \hat{\delta}_k x_k)$$

Feasible GLS is consistent and asymptotically more efficient than OLS.

Multiple Regression Analysis: Heteroskedasticity

Example: demand for cigarettes

■ Estimation by OLS

Cigarettes smoked per day

Logged income and cigarette price

$$\widehat{cigs} = - \frac{3.64}{(24.08)} + \frac{.880}{(.728)} \log(income) - \frac{.751}{(5.773)} \log(cigpric)$$
$$- \frac{.501}{(.167)} educ - \frac{.771}{(.160)} age - \frac{.0090}{(.0017)} age^2 - \frac{2.83}{(1.11)} restaurn$$

Smoking restrictions in restaurants

Reject homoskedasticity

$n = 807, R^2 = .0526, p\text{-value}_{Breusch-Pagan} = .000$

Multiple Regression Analysis: Heteroskedasticity

Estimation by FGLS

Now statistically significant

$$\begin{aligned} \widehat{cigs} = & - \frac{5.64}{(17.80)} + \frac{1.30}{(.44)} \log(income) - \frac{2.94}{(4.46)} \log(cigpric) \\ & - .463 \frac{educ}{(.120)} + .482 \frac{age}{(.097)} - .0056 \frac{age^2}{(.0009)} - 3.46 \frac{restaurn}{(.80)} \end{aligned}$$

$$n = 807, R^2 = .1134$$

Discussion

- The income elasticity is now statistically significant; other coefficients are also more precisely estimated (without changing the quality of the results)

Multiple Regression Analysis: Heteroskedasticity

What if the assumed heteroskedasticity function is wrong?

- If the heteroskedasticity function is misspecified, WLS is still consistent under MLR.1 – MLR.4, but robust standard errors should be computed
- WLS is consistent under MLR.4 but not necessarily under MLR.4'

$$E(u_i | \mathbf{x}_i) = 0 \quad \Rightarrow \quad E\left(u_i / \sqrt{h(\mathbf{x}_i)} \mid \mathbf{x}_i\right) = 0$$

- If OLS and WLS produce very different estimates, this typically indicates that some other assumptions (e.g. MLR.4) are wrong
- If there is strong heteroskedasticity, it is still often better to use a wrong form of heteroskedasticity in order to increase efficiency

Multiple Regression Analysis: Heteroskedasticity

WLS in the linear probability model

$$P(y = 1|\mathbf{x}) = p(\mathbf{x}) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

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

In the LPM, the exact form of heteroskedasticity is known

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

Use inverse values as weights in WLS

Discussion

- Infeasible if LPM predictions are below zero or greater than one
- If such cases are rare, they may be adjusted to values such as .01/.99
- Otherwise, it is probably better to use OLS with robust standard errors

Summary

- Testing for heteroskedasticity
 - Breusch-Pagan test
 - White test
 - Alternative form of the White test
- If heteroskedasticity exists, corrective measures
 - heteroskedasticity-robust standard error
 - WLS if heteroskedasticity is known to a multiplicative constant
 - If observations are reported as averages at the group level
 - Linear probability model
 - FGLS for unknown heteroskedasticity function