Heteroscedasticity

Chapter 7 (Ch. 8 of Textbook)

Wooldridge: Introductory Econometrics:
A Modern Approach, 5e

Multiple Regression Analysis: Heteroscedasticity

- **Motivation**
  - Assumption MLR.5
    $$Var(u_i \mid x_{i1}, x_{i2}, \ldots, x_{ik}) = \sigma^2$$
  - With cross-section data this assumption is not verified often
  - The conditional variance of the error term depends on the explanatory variables $\Rightarrow$ **heteroscedasticity**
    $$Var(u_i \mid x_{i1}, x_{i2}, \ldots, x_{ik}) = h(x_{i1}, x_{i2}, \ldots, x_{ik}) = \sigma_i^2 \neq \sigma^2$$
  - It is a issue of the **conditional variance**

Multiple Regression Analysis: Heteroscedasticity

- **Consequences of heteroscedasticity for OLS**
  - OLS still **unbiased and consistent** under heteroscedastictiy!
  - 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 heteroscedasticity (which refers to the conditional error variance)

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

Multiple Regression Analysis: Heteroscedasticity

- **Heteroscedasticity-robust inference after OLS**
  - Formulas for OLS standard errors and related statistics have been developed that are **robust to heteroscedasticity** of unknown form
  - All formulas are only valid in large samples
  - Formula for heteroscedasticity-robust OLS standard error

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

Also called White/Eicker standard errors. They involve the squared residuals from the regression and from a 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 heteroscedasticity**, but heteroscedasticity robust versions are available in most software

Multiple Regression Analysis: Heteroscedasticity

- **Example: Hourly wage equation**

$$\widehat{\log}(wage) = \underset{\substack{(.105)\\ [.107]}}{-\ .128} + \underset{\substack{(.0075)\\ [.0078]}}{.0904}\ educ + \underset{\substack{(.0052)\\ [.0050]}}{.0410}\ exper - \underset{\substack{(.0001)\\ [.0001]}}{.0007}\ exper^2$$

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

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

$$F = 17.95$$

$$F_{robust} = 17.99$$

F-statistics are also often not too different.

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

Multiple Regression Analysis: Heteroscedasticity

- **Example: Hourly wage equation in EVIEWS**

Dependent Variable: LOG(WAGE)
Method: Least Squares
Included observations: 526

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.390483	0.102210	3.820413	0.0001
FEMALE	-0.337187	0.036321	-9.283424	0.0000
EDUC	0.084136	0.006957	12.09407	0.0000
EXPER	0.038910	0.004824	8.066683	0.0000
EXPER^2	-0.000686	0.000107	-6.388842	0.0000

R-squared	0.399590	Mean dependent var	1.623268
Adjusted R-squared	0.394981	S.D. dependent var	0.531538
S.E. of regression	0.413446	Akaike info criterion	1.080882
Sum squared resid	89.05862	Schwarz criterion	1.121427
Log likelihood	-279.2720	Hannan-Quinn criter.	1.096757
F-statistic	86.68521	Durbin-Watson stat	1.775544
Prob(F-statistic)	0.000000		

Dependent Variable: LOG(WAGE)
Method: Least Squares
Included observations: 526
White heteroskedasticity-consistent standard errors & covariance

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.390483	0.108598	3.595658	0.0004
FEMALE	-0.337187	0.036184	-9.318715	0.0000
EDUC	0.084136	0.007690	10.94104	0.0000
EXPER	0.038910	0.004675	8.322568	0.0000
EXPER^2	-0.000686	0.000100	-6.828754	0.0000

R-squared	0.399590	Mean dependent var	1.623268
Adjusted R-squared	0.394981	S.D. dependent var	0.531538
S.E. of regression	0.413446	Akaike info criterion	1.080882
Sum squared resid	89.05862	Schwarz criterion	1.121427
Log likelihood	-279.2720	Hannan-Quinn criter.	1.096757
F-statistic	86.68521	Durbin-Watson stat	1.775544
Prob(F-statistic)	0.000000	Wald F-statistic	81.96798
Prob(Wald F-statistic)	0.000000		

Multiple Regression Analysis: Heteroscedasticity

- **Testing for heteroscedasticity**
  - It may still be interesting whether there is heteroscedasticity because then OLS may not be the most efficient linear estimator anymore
- **Breusch-Pagan test for heteroscedasticity**

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

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

Under MLR.4

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

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

Multiple Regression Analysis: Heteroscedasticity

- **Breusch-Pagan test for heteroscedasticity (cont.)**

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

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

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

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

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

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

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: Heteroscedasticity

- **Example:**

$\hat{u}^2$ With $\hat{u}$ the residual for the regression of log(wage)

Test Equation:
Dependent Variable: RESID^2
Method: Least Squares
Included observations: 526

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.037480	0.068360	0.548268	0.5837
FEMALE	-0.013118	0.024292	-0.540000	0.5894
EDUC	0.005509	0.004653	1.183964	0.2370
EXPER	0.008761	0.003226	2.715742	0.0068
EXPER^2	-0.000169	7.18E-05	-2.358137	0.0187

R-squared	0.018981	Mean dependent var	0.169313
Adjusted R-squared	0.011449	S.D. dependent var	0.278118
S.E. of regression	0.276521	Akaike info criterion	0.276402
Sum squared resid	39.83772	Schwarz criterion	0.316946
Log likelihood	-67.69362	Hannan-Quinn criter.	0.292277
F-statistic	2.520076	Durbin-Watson stat	1.967219
Prob(F-statistic)	0.040375		

Test statistic

Multiple Regression Analysis: Heteroscedasticity

- **Example: Heteroscedasticity in housing price equations**

$$\widehat{price} = -\underset{(29.48)}{21.77} + \underset{(.0006)}{.0021}\ lotsize + \underset{(.013)}{.123}\ sqrft + \underset{(9.01)}{13.85}\ bdrms$$

n=88

Heteroscedasticity

$$\Rightarrow R_{\hat{u}^2} = .1601,\ \boxed{p\text{−}value_F = .002,\ p\text{−}value_{LM} = .0028}$$

$$\widehat{\log}(price) = -\underset{(.65)}{1.30} + \underset{(.038)}{.168}\ \log(lotsize) + \underset{(.093)}{.700}\ \log(sqrft) + \underset{(.028)}{.037}\ bdrms$$

n=88

$$\Rightarrow R_{\hat{u}^2} = .0480,\ \boxed{p\text{−}value_F = .245,\ p\text{−}value_{LM} = .2390}$$

In the logarithmic specification, homoscedasticity cannot be rejected

Multiple Regression Analysis: Heteroscedasticity

- **White test for heteroscedasticity**

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

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

$$+ \delta_7 x_1 x_2 + \delta_8 x_1 x_3 + \delta_9 x_2 x_3 + error$$

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

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

The White test detects more general deviations from heteroscedasticity 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)

$\hat{u}^2$ With $\hat{u}$ the residual for the regression of log(wage)

Test Equation:
Dependent Variable: RESID^2
Method: Least Squares
Included observations: 526
Collinear test regressors dropped from specification

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.558137	0.268879	2.075795	0.0384
FEMALE	-0.125279	0.140393	-0.892343	0.3726
FEMALE*EDUC	0.012647	0.010043	1.259261	0.2085
FEMALE*EXPER	-0.003751	0.006530	-0.574430	0.5659
FEMALE*EXPER^2	4.52E-05	0.000146	0.310811	0.7561
EDUC^2	0.002544	0.001125	2.260940	0.0242
EDUC*EXPER	-2.69E-05	0.001256	-0.021411	0.9829
EDUC*EXPER^2	1.40E-05	2.73E-05	0.511937	0.6089
EDUC	-0.066880	0.033663	-1.986760	0.0475
EXPER^2	0.000813	0.001351	0.601774	0.5476
EXPER*EXPER^2	-2.99E-05	4.44E-05	-0.673824	0.5007
EXPER	-0.005241	0.021205	-0.247149	0.8049
EXPER^2^2	2.50E-07	4.70E-07	0.531774	0.5951

R-squared	0.037890	Mean dependent var	0.169313
Adjusted R-squared	0.015385	S.D. dependent var	0.278118
S.E. of regression	0.275970	Akaike info criterion	0.287356
Sum squared resid	39.06984	Schwarz criterion	0.392772
Log likelihood	-62.57470	Hannan-Quinn criter.	0.328631
F-statistic	1.683600	Durbin-Watson stat	1.970279
Prob(F-statistic)	0.066942		

Test statistic

Multiple Regression Analysis: Heteroscedasticity

- **Alternative form of the White test – Simplified White**

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

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.

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

- **Example: Heteroscedasticity 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: Heteroscedasticity

Example of Simplified White

$\hat{u}^2$ With $\hat{u}$ the residual for the regression of log(wage)

Dependent Variable: RESID^2
Method: Least Squares
Included observations: 526

	Variable	Coefficient	Std. Error	t-Statistic	Prob.
	C	0.233483	0.209660	1.113627	0.2660
$\hat{y}$	FIT	-0.187119	0.263787	-0.709356	0.4784
$\hat{y}^2$	FIT^2	0.087191	0.081226	1.073446	0.2836

R-squared	0.014904	Mean dependent var	0.169313
Adjusted R-squared	0.011137	S.D. dependent var	0.278118
S.E. of regression	0.276565	Akaike info criterion	0.272944
Sum squared resid	40.00326	Schwarz criterion	0.297271
Log likelihood	-68.78420	Hannan-Quinn criter.	0.282469
F-statistic	3.956441	Durbin-Watson stat	1.943744
Prob(F-statistic)	0.019706		

Test statistic