Heteroskedasticity

8 Chapter

Ketevani Kapanadze

Brno, 2020

Consequences of Heteroskedasticity for OLS

- **<u>Consequences of heteroscedasticity for OLS</u>**
  - OLS still unbiased and consistent under heteroscedastictiy!
  - Also, interpretation of R-squared is not changed
  - 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

Heteroskedasticity-Robust Inference after OLS Estimation

- **<u>Heteroscedasticity-robust inference after OLS</u>**
  - Formulas for OLS standard errors and related statistics have been developed that are robust to heteroscedasticity of unknown form
  - <u>All formulas are only valid in large samples</u>
  - 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 <u>White/Eicker standard errors</u>. They involve the squared residuals from the regression and from a regression of x_j on all other explanatory variables.

  - Using thes formula, 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

Heteroskedasticity-Robust Inference after OLS Estimation

- **Example: Hourly wage equation**

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

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-statistics are also often not too different.

$$F_{robust} = 17.99$$

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

Testing for Heteroskedasticity

- <u>**Testing for heteroscedasticity**</u>
  - 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**

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

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

Under MLR.4

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

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

Testing for Heteroskedasticity

- **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 expla-natory 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 obtained by regressing residuals from unrestricted model to all explanatory variables). 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.

Testing for Heteroskedasticity

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

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

Heteroscedasticity

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

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

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

In the logarithmic specification, homoscedasticity cannot be rejected – benefit of using the logarithmic functional form

Testing for Heteroskedasticity

- **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)

Testing for Heteroskedasticity

- **Alternative form of the White test**

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

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.

$$H_0 : \delta_1 = \delta_2 = 0,\ \ LM = n \cdot R_{\widehat{u}^2} \sim \chi^2_2$$

- **Example: Heteroscedasticity in (log) housing price equations**

$$R^2_{\widehat{u}^2} = .0392, LM = 88(.0392) \approx 3.45, p\text{--}value_{LM} = .178$$

Weighted Least Squares Estimation

- **Heteroscedasticity is known up to a multiplicative constant**

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

The functional form of the heteroscedasticity is known

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

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

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

Transformed model

Weighted Least Squares Estimation

- **Example: Savings and income**

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

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

Note that this regression model has no intercept

- **The transformed model is homoscedastic**

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

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

Weighted Least Squares Estimation

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

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

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

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)**

Weighted Least Squares Estimation

- **Unknown heteroscedasticity function (feasible GLS)**

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

Assumed general form of heteroscedasticity; exp-function is used to ensure positivity

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

Multiplicative error (assumption: independent of the explanatory variables)

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

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

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

Use inverse values of the estimated heteroscedasticity funtion as weights in WLS

Feasible GLS is consistent and asymptotically more efficient than OLS.

Weighted Least Squares Estimation

- **Example: Demand for cigarettes**
- **Estimation by OLS**

Smoking restrictions in restaurants

Cigarettes smoked per day

Logged income and cigarette price

$$\widehat{cigs} = -\underset{(24.08)}{3.64} + \underset{(.728)}{.880}\ \log(income) - \underset{(5.773)}{.751}\ \log(cigpric)$$

$$- \underset{(.167)}{.501}\ educ - \underset{(.160)}{.771}\ age - \underset{(.0017)}{.0090}\ age^2 - \underset{(1.11)}{2.83}\ restaurn$$

$$n = 807, R^2 = .0526, p{-}value_{Breusch-Pagan} = .000$$

Reject homo-scedasticity

Weighted Least Squares Estimation

- **Estimation by FGLS**

Now statistically significant

$$\widehat{cigs} = -\underset{(17.80)}{5.64} + \underset{(.44)}{1.30}\,\log(income) - \underset{(4.46)}{2.94}\,\log(cigpric)$$

$$- \underset{(.120)}{.463}\,educ + \underset{(.097)}{.482}\,age - \underset{(.0009)}{.0056}\,age^2 - \underset{(.80)}{3.46}\,restaurn$$

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

- **Discussion**
  - The income elasticity is now statistically significant; other coefficients are also more precisely estimated (without changing qualit. results)

Weighted Least Squares Estimation

- **<u>What if the assumed heteroscedasticity function is wrong?</u>**
  - If the heteroscedasticity 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**

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

  - 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 heteroscedasticity, it is still often better to use a wrong form of heteroscedasticity in order to **increase efficiency**

Next Class

- **Endogenous regressors and instrumental variables**
- **Multiple Choice Quiz ☺**

13.03.2020