ECO375 Tutorial 7
Heteroscedasticity

Matt Tudball

University of Toronto Mississauga

November 9, 2017

Review: Heteroscedasticity

- Consider again the standard multiple regression model in which an outcome y_i is linearly related to some explanatory variables $x_{i1}, x_{i2}, ..., x_{ik}$ such that

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

- Recall that an assumption needed for efficiency of the OLS estimator and construction of t-tests, F-tests and valid confidence intervals was MLR.5:

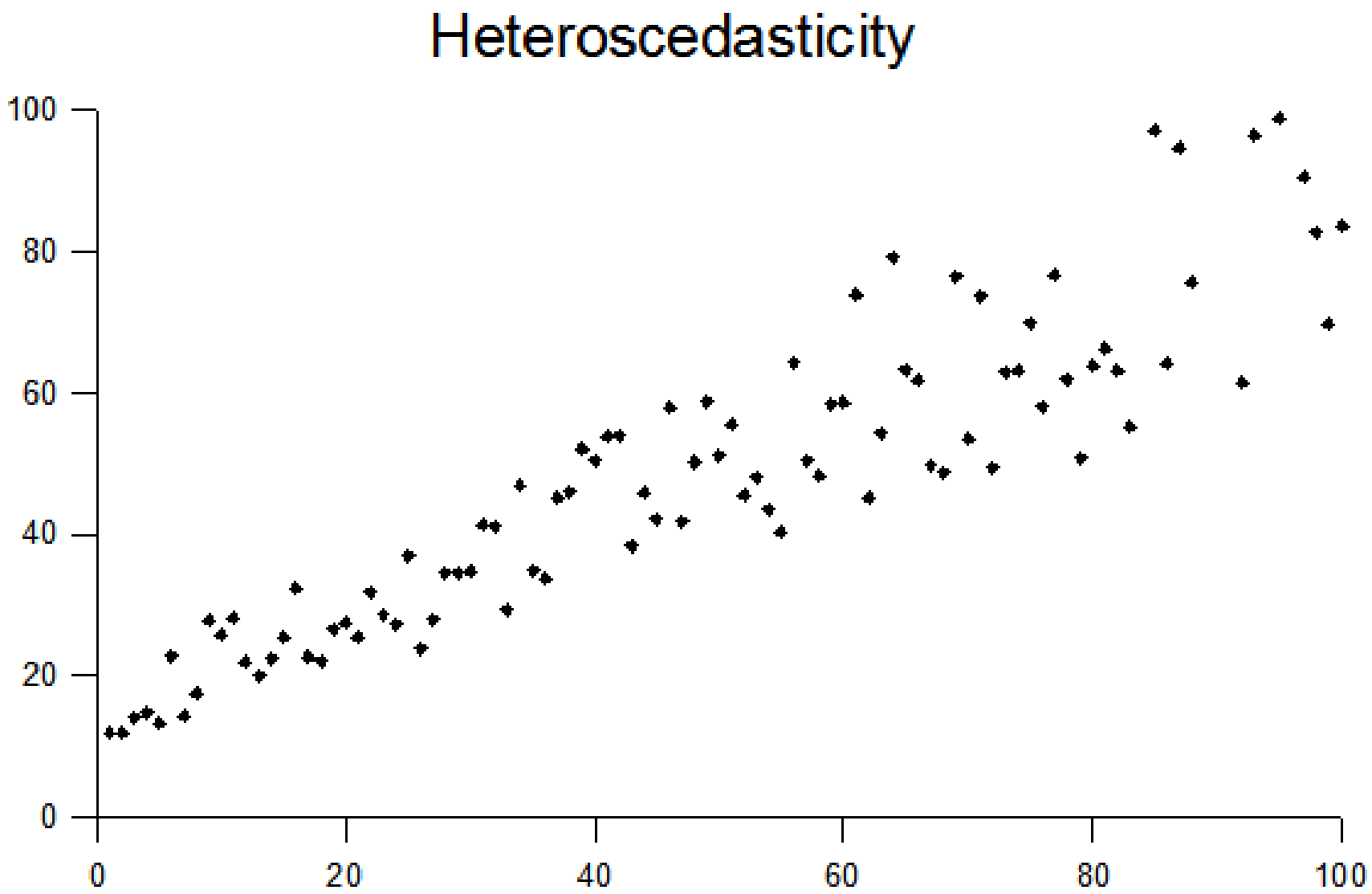
MLR.5 The error u_i homoscedastic, i.e. it has the same variance for all x_i $\text{Var}(u_i|x_{i1}, x_{i2}, ..., x_{ik}) = \sigma^2$

- This says that the unobservable error term u_i has the same variance for all observations i.

Review: Heteroscedasticity

- This is a fairly strong assumption and there are many circumstances in which it will not hold in practice.
- A classic example where this assumption may fail is the relationship between income and food expenditure. Poor people will generally purchase inexpensive food and so the variance of their food expenditure is low. Rich people, however, may purchase expensive or inexpensive food depending on their tastes and so the variance of their food expenditure is high.
- So for $foodexp_i = \beta_0 + \beta_1 income_i + u_i$ we will generally find that $\text{Var}(u_i|income_i)$ becomes larger as $income_i$ increases.

Review: Heteroscedasticity



Review: Heteroscedasticity

- Let's give this notion a formal definition.
- We define **heteroscedasticity** as allowing the variance of u_i to vary across x_i such that $\text{Var}(u_i|x_{i1}, x_{i2}, ..., x_{ik}) = \sigma_i^2$, where the i subscript is what allows the variance to differ across observations.
- While heteroscedasticity does not affect the consistency or unbiasedness of the OLS estimator, it does make the OLS estimator inefficient and means that our t-statistics and F-statistics may not follow their respective distributions.

Generalised Least Squares

- So what can we do about this?
- Consider the base model

$$y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik} + u_i \tag{1}$$

where $\text{Var}(u_i|x_{i1}, x_{i2}, ..., x_{ik}) = \sigma_i^2$

- We can think that there may exist a transformation p_i such that a "transformed" error term $u_i^* = p_i u_i$ has homoscedastic variance

$$\text{Var}(u_i^*|x_{i1}, x_{i2}, ..., x_{ik}) = \sigma^2$$

Generalised Least Squares

- Let's say that we know the form of p_i for each observation i.
- Then in principle we could transform our data by multiplying both sides of the base model (1) by p_i such that

$$p_i y_i = \beta_0 p_i + \beta_1 p_i x_{i1} + ... + \beta_k p_i x_{ik} + p_i u_i$$
$$y_i^* = \beta_0 x_{0i}^* + \beta_1 x_{i1}^* + ... + \beta_k x_{ik}^* + u_i^*$$

  where $\text{Var}(u_i^* | x_{i1}, x_{i2}, ..., x_{ik}) = \sigma^2$ which satisfies MLR.5.
- The estimator obtained by running OLS on this transformed model is known as the **Generalised Least Squares (GLS)** estimator.
- We will often assume that $p_i = 1/\sqrt{h_i}$ for some function h_i since this allows us to assume a form of the heteroscedasticity $\sigma_i^2 = \sigma^2 h_i$ (note that this is purely for notational convenience).

Generalised Least Squares: Example

- Consider a model which explores the relationship between house prices $price_i$, square footage $sqrft_i$, lot size $lotsize_i$ and number of bedrooms $bdrms_i$

$$price_i = \beta_0 + \beta_1 sqrtft_i + \beta_2 lotsize_i + \beta_3 bdrms_i + u_i \tag{2}$$

  and consider a possible form of heteroscedasticity $\sigma_i^2 = \sigma^2 sqrft_i$.
- This indicates that the variance of $price_i$ given $sqrft_i$, $lotsize_i$ and $bdrms_i$ is increasing in $sqrft_i$.

Generalised Least Squares: Example

- Now consider the transformation $p_i = 1/\sqrt{sqrft_i}$. Then if we divide both sides of (2) by $\sqrt{sqrft_i}$

$$\begin{aligned} price_i/\sqrt{sqrft_i} =& \beta_0/\sqrt{sqrft_i} + \beta_1\sqrt{sqrft_i} + \beta_2 lotsize_i/\sqrt{sqrft_i} \\ &+ \beta_3 bdrms_i/\sqrt{sqrft_i} + u_i/\sqrt{sqrft_i} \\ price_i^* =& \beta_0/\sqrt{sqrft_i} + \beta_1 sqrtft_i^* + \beta_2 lotsize_i^* + \beta_3 bdrms_i^* + u_i^* \end{aligned}$$

such that

$$\begin{aligned} \text{Var}(u_i^*) &= \text{Var}(u_i/\sqrt{sqrft_i}) \\ &= \text{Var}(u_i)/sqrft_i \\ &= \sigma^2 sqrft_i/sqrft_i = \sigma^2 \end{aligned}$$

which satisfies MLR.5.

Relationship with Weighted Least Squares

- GLS estimators belong to a broader class of so-called **Weighted Least Squares (WLS)** estimators.
- We can think of the transformation p_i as assigning a "weight" to each observation i.
- Observations with more variance in u_i are going to be weighted down by p_i and observations with less variance are going to be weighted up.
- This makes intuitive sense from an efficiency perspective since observations with lower variance in u_i are going to contain more precise information compared to observations with higher variance.
- The objective function of the WLS estimator with weights p_i is

$$\sum_{i=1}^{n} p_i^2 \left(y_i - \beta_0 - \beta_1 x_{i1} - \ldots - \beta_k x_{ik}\right)^2$$

Feasible Generalised Least Squares: Idea

- An obvious drawback of GLS is that the transformation $p_i = 1/\sqrt{h_i}$ which eliminates heteroscedasticity is almost always unknown.
- One approach to implementing GLS in practice is to assume a particular functional form for h_i as a function of the explanatory variables x_i, usually denoted by $h(x_i)$.
- A popular functional form is

$$h(x_i) = \exp(\delta_0 + \delta_1 x_{i1} + ... + \delta_k x_{ik})$$

- This implies a form of the variance of u such that

$$\text{Var}(u_i|x_{i1}, x_{i2}, ..., x_{ik}) = \sigma^2 \exp(\delta_0 + \delta_1 x_{i1} + ... + \delta_k x_{ik})$$

- We can therefore write a model for the variance of the form

$$\begin{aligned} u_i^2 &= \sigma^2 \exp(\delta_0 + \delta_1 x_{i1} + ... + \delta_k x_{ik}) v_i \\ \ln(u_i^2) &= \underbrace{\ln(\sigma^2) + \delta_0}_{\alpha_0} + \delta_1 x_{i1} + ... + \delta_k x_{ik} + e_i \end{aligned}$$

Feasible Generalised Least Squares: Implementation

Let's describe how to implement the **Feasible Generalised Least Squares (FGLS)** estimator described in the previous slide in practice.

1. Begin by estimating the original OLS model and recover the sample residual $\hat{u}_i$.
2. Estimate the model

$$\ln(\hat{u}_i^2) = \alpha_0 + \delta_1 x_{i1} + ... + \delta_k x_{ik} + e_i$$

3. Set

$$\hat{h}_i = \exp(\hat{\alpha}_0 + \hat{\delta}_1 x_{i1} + ... + \hat{\delta}_k x_{ik})$$

4. Transform the original model using the weights

$$\hat{p}_i = 1/\sqrt{\hat{h}_i}$$

and estimate by OLS.

In-Class Exercise 1

In this exercise we are going to implement FGLS in Stata. Download the dataset HPRICE1.dta from my website (matthewtudball.com).

```
reg price sqrft lotsize bdrms
predict uhat, residuals
gen log_uhat2 = ln(uhat^2)
reg log_uhat2 sqrft lotsize bdrms
predict reshat, xb
gen hhat = exp(reshat)
gen phat = 1/sqrt(hhat)
gen pricep = price*phat
gen sqrftp = sqrft*phat
gen lotsizep = lotsize*phat
gen bdrmsp = bdrms*phat
reg pricep phat sqrftp lotsizep bdrmsp, nocons
```

Why do I include the variable *phat* in our final regression? Why do I have *nocons* as an option in the final line, indicating that I do not want to include a constant term? (Hint: Look at Slide 9).

Robust (White) Standard Errors

- An alternative to using GLS is to use standard errors called **robust (White) standard errors**.
- The idea behind robust standard errors is to use $\hat{u}_i^2$ as an estimator for observation i's true variance σ_i^2.
- When this is used to construct an estimator for the variance of $\hat{\beta}_j$ the estimator is consistent.
- How does this compare to using GLS?
  - When the functional form of $\hat{h}_i$ is correctly specified, GLS is more efficient than robust standard errors.
  - While robust standard errors may be less efficient, they provide consistent estimators without relying on functional form assumptions.
  - **In practice always use robust standard errors.**
- To implement robust standard errors in Stata in the regression in In-Class Exercise 1, simply type

```
reg price sqrft lotsize bdrms, robust
```

Testing for Heteroscedasticity

- Since robust standard errors may be inefficient and GLS relies on assumptions that are hard to justify, we generally want to use OLS if we can rule out heteroscedasticity.
- Let's return again to the standard multiple regression model

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

- We ultimately want to test

$$H_0 : \mathbb{E}(u_i^2) = \sigma^2$$
$$H_1 : \mathbb{E}(u_i^2) = \sigma_i^2$$

- In order to test this in practice, we will need to impose some structure on the form of σ_i^2 (note that this may weaken the strength of these tests).

Testing for Heteroscedasticity

- Consider instead the alternative hypothesis

$$H_1 : \mathbb{E}(u_i^2) = \delta_0 + \delta_1 z_{i1} + ... + \delta_p z_{ip}$$

  where the variables $z_{i1}, ..., z_{ip}$ remain unspecified for now.
- Then we have a more structured form of the test in the previous slide

$$H_0 : \delta_1 = ... = \delta_p = 0$$
$$H_1 : \delta_j \neq 0 \text{ for some } j = 1, ..., p$$

- Note that since the null hypothesis has multiple restrictions this will ultimately be an F-test.

Testing for Heteroscedasticity

Let's consider how to implement a test of this form.

1. Estimate the restricted model (OLS with homoscedasticity) and obtain the sample residuals $\hat{u}_i$.
2. Use OLS to estimate

$$\hat{u}_i^2 = \delta_0 + \delta_1 z_{i1} + ... + \delta_p z_{ip} + e_i$$

3. Under the null hypothesis

$$\frac{R^2_{\hat{u}^2}/p}{(1-R^2_{\hat{u}^2})/(n-p-1)} \sim F_{n-p-1} \text{ and } nR^2_{\hat{u}^2} \sim \chi^2_p$$

Breusch-Pagan (BP) Test for Heteroscedasticity

- Which set of z_i's should we use?
- The **Breusch-Pagan (BP) Test** simply uses the set of explanatory variables x_i such that the alternative hypothesis takes the form

$$H_1 : \mathbb{E}(u_i^2) = \delta_0 + \delta_1 x_{i1} + ... + \delta_k x_{ik}$$

and the auxiliary regression is

$$\hat{u}_i^2 = \delta_0 + \delta_1 x_{i1} + ... + \delta_k x_{ik} + e_i$$

with the test statistic $nR^2_{\hat{u}^2} \sim \chi^2_k$ under the null hypothesis.

White Test for Heteroscedasticity

- Another approach allows for a more flexible relationship between x_i and σ_i^2 by specifying polynomials in the x_i's and interactions between them.
- This is known as the **White Test**.
- Consider the simple model

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

- White's version of the test uses the auxiliary regression

$$\hat{u}_i^2 = \delta_0 + \delta_1 x_{i1} + \delta_2 x_{i2} + \delta_3 x_{i1}^2 + \delta_4 x_{i2}^2 + \delta_5 x_{i1} x_{i2} + e_i$$

with the test statistic $nR^2_{\hat{u}^2} \sim \chi^2_5$.

White Test for Heteroscedasticity

- With many regressors, the number of interaction terms increases quickly. This quickly becomes a dimensionality problem.
- There is an alternative version of this same test that utilises a restricted combination of squares and interactions of the regressors.
- Let

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 \hat{x}_{i1} + ... + \hat{\beta}_k x_{ik}$$

- The auxiliary regression then becomes

$$\hat{u}_i^2 = \delta_0 + \delta_1 \hat{y}_i + \delta_2 \hat{y}_i^2 + e_i$$
$$H_0 : \delta_1 = \delta_2 = 0$$
$$nR^2_{\hat{u}^2} \sim \chi^2_2$$

- This is "restricted" in the sense that the square and interaction terms of x_i contained in $\hat{y}_i^2$ are assumed to have an identical effect on $\hat{u}_i^2$ (namely, δ_2).
- Note that rejecting the null is not proof of homoscedasticity.

In-Class Exercise 2

In this exercise we will implement both the Breusch-Pagan Test and White Test for heteroscedasticity. Load in the HPRICE1.dta dataset used in the last exercise.

- To run the Breusch-Pagan Test type the following code

```
reg price sqrft lotsize bdrms
predict uhat, residuals
gen uhat2 = uhat^2
reg uhat2 sqrft lotsize bdrms
```

  Do you reject the null hypothesis of homoscedasticity?

In-Class Exercise 2

- To run version 1 of the White Test type the following code

```
gen sqrft2 = sqrft^2
gen lotsize2 = lotsize^2
gen bdrms2 = bdrms^2
reg uhat2 c.sqrft##c.lotsize c.sqrft##c.bdrms
c.lotsize##c.bdrms ///
    sqrft2 lotsize2 bdrms2
```

  Do you reject the null hypothesis of homoscedasticity?

- To run version 2 of the White Test type the following code

```
reg price sqrft lotsize bdrms
predict yhat, xb
gen yhat2 = yhat^2
reg uhat2 yhat yhat2
```

  Do you reject the null hypothesis of homoscedasticity?

In-Class Exercise 3

This is Computer Exercise 8 from Wooldridge Chapter 8. Download the dataset GPA1.dta from my website (matthewtudball.com).

1. Use OLS to estimate a model relating *colGPA* to *hsGPA*, *ACT*, *skipped* and *PC*. Obtain (i.e. predict) the OLS residuals.
2. Compute version 2 of the White Test for heteroscedasticity. In the regression of $\hat{u}_i^2$ on $\widehat{colGPA}_i$ and $\widehat{colGPA}_i^2$, obtain the fitted values, say $\hat{h}_i$.

In-Class Exercise 3

1. Verify that the fitted values from part 2 are all strictly positive. Then, obtain the weighted least squares estimates using weights $1/\hat{h}_i$. Compare the weighted least squares estimates for the effect of skipping lectures and the effect of PC ownership with the corresponding OLS estimates. What about their statistical significance?
2. In the WLS estimation from part 3, obtain the robust standard errors. In other words, allow for the fact that the variance function estimated in part 2 might be misspecified. Do the standard errors change much from part 3?