

BSM946: Topic 2: Dummy Variables and Heteroskedasticity

David Morris

Economics, Finance and Entrepreneurship Group

d.morris5@aston.ac.uk

House Keeping

① Workshops

- Learning by doing...
- and consolidation!!!

② Logs

- Notes on BB

③ Dissertations

- Happy to discuss models and work etc

Today's Plan

Basics

- ① Introduction to dummy variables
- ② Look at p-values

Heteroskedasticity

- ③ What is heteroskedasticity
- ④ How do you spell heteroskedasticity?!
- ⑤ What is heteroskedasticity?
- ⑥ How to deal with heteroskedasticity

Dummy Variables

To start with:

- What are dummy variables?
- How do we know if a variable is a dummy variable?
- Why can we not include all dummy variables?
- Interpreting a dummy variable?

Dummy variables

Where the variable takes only one of two values which are mutually exclusive

In practice this means interested in variables that split the sample into two distinct groups in the following way

- $D = 1$ if the criterion is satisfied
- $D = 0$ if not

Eg. Male/Female

- so that the dummy variable “Male” would be coded 1 if male and 0 if female

(though could equally create another variable “Female” coded 1 if female and 0 if male)

VERY FREQUENTLY USED!

Example

Suppose we are interested in the gender pay gap

Model is $\text{Log Wage} = \beta_0 + \beta_1 \text{Age} + \beta_2 \text{Male}$

- where $\text{Male} = 1$ if male or 0 if female

For men therefore the predicted wage is:

- $\text{Log Wage}_{\text{men}} = \beta_0 + \beta_1 \text{Age} + \beta_2 * 1$
- $\text{Log Wage}_{\text{men}} = \beta_0 + \beta_1 \text{Age} + \beta_2$

For women therefore the predicted wage is:

- $\text{Log Wage}_{\text{women}} = \beta_0 + \beta_1 \text{Age} + \beta_2 * 0$
- $\text{Log Wage}_{\text{women}} = \beta_0 + \beta_1 \text{Age}$

Interpretation

The coefficients on dummy variables measure the average difference between the group coded with the value “1” and the group coded with the value “0”

β_2 showed the difference in the wage received by men, compared to the base group of women

- Group coded zero is the “default” or “base group”

Categorical Variable

Similar to a dummy variable but with more than two options

Categorical Variable

Is an artificial variable created to represent an attribute with two or more distinct categories/levels

Why is it used?

- Regression analysis treats all independent (X) variables in the analysis as numerical
- If we want to measure someones education, we may code that as 1=degree, 2= A levels, 3=GCSEs 4=No qualifications....
 - 1 is not half as big as 2, 2 is not double 4 etc....

Therefore if we were interested in the impact of education on wages, we could turn this categorical variable into numerous dummy variables and add them to the analysis

$$\text{Wage} = \beta_0 + \beta_1 \text{Degree} + \beta_2 \text{Alevels} + \beta_3 \text{GCSEs} + \beta_4 \text{Age} + \varepsilon$$

What happened to the fourth education level, no qualifications?

- Well, it turns out we don't need a fourth dummy variable to represent it
- Setting Degree, A-levels and GCSEs to '0' indicates the individual has no degree

Notice that this coding only works if the four levels are mutually exclusive (so not overlap) and exhaustive (no other levels exist for this variable)

Recap

- 1 Dummy variables assign the numbers '0' and '1' to indicate membership
- 2 Membership groups must be **mutually exclusive** and **exhaustive**
- 3 The number of dummy variables necessary to represent a single attribute variable is equal to the number of levels (categories) in that variable minus one (**n-1**)

STATA Example

Throughout the first part of today we will be using the following STATA example...

Source	SS	df	MS	Number of obs = 3680		
Model	1178662.21	7	168380.315	F(7, 3672) = 280.43		
Residual	2204768.3	3672	600.427098	Prob > F = 0.0000		
Total	3383430.51	3679	919.660372	R-squared = 0.3484		
				Adj R-squared = 0.3471		
				Root MSE = 24.504		

	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
edyears	3.238817	.2102537	15.40	0.000	2.826591	3.651042
age	.4713265	.0331272	14.23	0.000	.4063769	.5362761
female	-17.10797	.9413247	-18.17	0.000	-18.95354	-15.2624
egp1	16.6732	2.077593	8.03	0.000	12.59985	20.74655
egp2	12.5756	1.441205	8.73	0.000	9.74996	15.40124
egp3	-.5958906	1.306433	-0.46	0.648	-3.157296	1.965515
egp4	-.0258211	1.450667	-0.02	0.986	-2.870013	2.818371
_cons	66.23801	1.743013	38.00	0.000	62.82064	69.65538

STATA Example

What is this regression investigating?

If you look carefully at the title row of the table we can see the dependent variable

- In this case this is **wages**

Thus this is a model trying to *calculate how different variables impact the wage an individual is paid*

There is a model trying to *calculate how different variables impact the wage an individual is paid*

But what variables are included?

The independent variables can be seen in the furthest left column, **these are the factors that we are testing to see if they are related to an individual's wage**

- You can see we have included the age of the individual ([age](#))
- The years they have been working ([edyears](#))
- Their gender ([female](#))
- A series of dummy variables to capture training conducted while in work ([egp1](#), [egp2](#), [egp3](#), [egp4](#))

Coefficients: Independent Variables

Continuous Independent Variables

When there is a continuous independent variable, β represents the difference in the predicted value of Y for each one-unit difference in the given independent variable, if all other independent variables remain constant

This means that if β_1 differed by one unit, and all other β values did not differ, Y will differ by β_1 units, on average

STATA Example: Independent Variables

Source	SS	df	MS	Number of obs = 3680		
Model	1178662.21	7	168380.315	F(7, 3672) = 280.43		
Residual	2204768.3	3672	600.427098	Prob > F = 0.0000		
Total	3383430.51	3679	919.660372	R-squared = 0.3484		
				Adj R-squared = 0.3471		
				Root MSE = 24.504		

wage	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
edyears	3.238817	.2102537	15.40	0.000	2.826591	3.651042
age	.4713285	.0331272	14.23	0.000	.4063769	.5362761
female	-17.10797	.9413247	-18.17	0.000	-18.95354	-15.2624
egp1	16.6732	2.077593	8.03	0.000	12.59985	20.74655
egp2	12.5756	1.441205	8.73	0.000	9.74996	15.40124
egp3	-.5958906	1.306433	-0.46	0.648	-3.157296	1.965515
egp4	-.0258211	1.450667	-0.02	0.986	-2.870013	2.818371
_cons	66.23801	1.743013	38.00	0.000	62.82064	69.65538

In the STATA output the β coefficient is reported in the coefficients column

- The continuous independent variable years of education has been highlighted here

The coefficient here suggests that for an extra year of education, the weekly wage of an individual increases by £3.23

Coefficients: Independent Variables

Dummy Variables

What if the independent variable is not continuous?

If the variable is a categorical variable, and thus coded as 0 or 1, a one unit difference represents switching from one category to the other

In the case of gender, the individual in this dataset is either identified as male (0) or female (1), there is not a continuous scale between them

Coefficients: Independent Variables

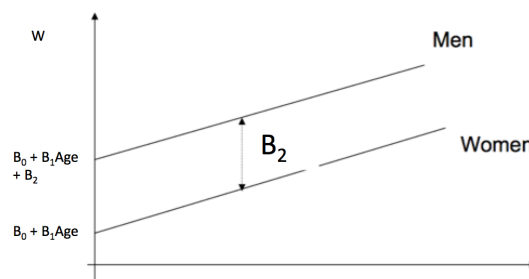
β is then the average difference in the dependent variable between the category for which the independent variable = 0 (the reference group) and the category for which the variable = 1 (the comparison group)

In this case females are 1 and males are 0

- So the β value shows the average difference in weekly wage, from being female rather than male

So compared to males in the sample, we would expect female earn £17.10 less per week, holding everything else constant

STATA Example: Dummy Variables



The coefficient on the female dummy variable suggest that compared to males in the sample, we would expect female earn £17.10 less per week, holding everything else constant

Coefficients: Further Notes

Further Notes

It is important to keep in mind that **each coefficient is influenced by the other variables in a regression model**

Therefore, **each coefficient will change when other variables are added to or deleted from the model**, and so care must be taken when comparing coefficients across different specifications

REMINDER

This interpretation is only correct for simple OLS regressions

More complex regressions will not have as simple an interpretation with regards to the coefficients

So we have talked about different types of independent variables and how to interpret their coefficients

We can now comment on the **size** and **sign** of these coefficients

The next important thing to comment on is the **significance** of these coefficients

- To do this the easiest way is to look at the **p-values**

Interpreting P-Values

The p-value for each term tests the null hypothesis that the coefficient is equal to zero (**no effect**)

A low p-value (< 0.05) indicates that you can **reject the null hypothesis**

In other words, a predictor that has a **low p-value is significantly related** to changes in the dependent variable in this sample

Conversely, a **larger (insignificant) p-value** suggests that changes in the predictor are **not associated** with changes in the dependent variable

Let's look at an example from STATA...

STATA Example

Source	SS	df	MS			
Model	1178662.21	7	168380.315			
Residual	2204768.3	3672	600.427098			
Total	3383430.51	3679	919.660372			

					Number of obs =	3680
					F(7, 3672) =	280.43
					Prob > F =	0.0000
					R-squared =	0.3484
					Adj R-squared =	0.3471
					Root MSE =	24.504

	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
wage						
edyears	3.238817	.2102537	15.40	0.000	2.826591	3.651042
age	.4713265	.0331272	14.23	0.000	.4063769	.5362761
female	-17.10797	.9413247	-18.17	0.000	-18.95354	-15.2624
egp1	16.6732	2.077593	8.03	0.000	12.59985	20.74655
egp2	12.5756	1.441205	8.73	0.000	9.74996	15.40124
egp3	-.5958906	1.306433	-0.46	0.648	-3.157296	1.965515
egp4	-.0258211	1.450667	-0.02	0.986	-2.870013	2.818371
_cons	66.23801	1.743013	38.00	0.000	62.82064	69.65538

You can see the p-values for this STATA output in the red box

What do the p-values show?

STATA Example

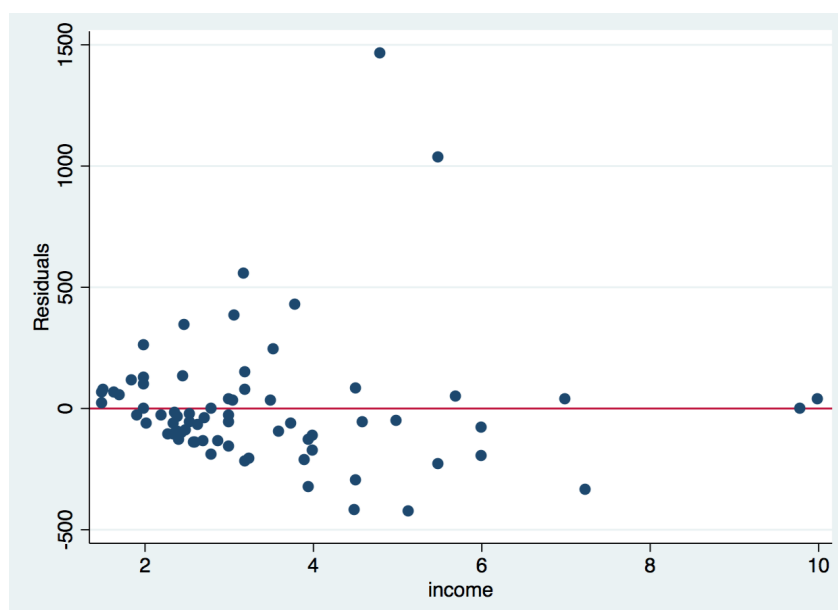
The first 5 independent variables have coefficients that are **highly significant**, they have p-values shown here as 0.000

- This suggests that these variables are significantly related to changes in the dependent variable in this sample

The next 2 independent variables (**egp3 and egp4**) have coefficients that are **not significant at the 5% level**, they have p-values shown here as > 0.05

- This suggests that these variables are not significantly related to changes in the dependent variable in this sample

Heteroskedasticity



“Every serious subject has its jargon. Economists need to know about heteroskedasticity. I take this example because it is virtually impossible to pronounce, and impossible to use the word in front of a class without everyone bursting out into laughter. Indeed, most spell-check programmes reject it, and offer improbable or embarrassing alternatives”

Quote from John Kay

- We will spell it with a k!

What is Heteroskedasticity?

One of the assumptions of OLS is Homoskedasticity

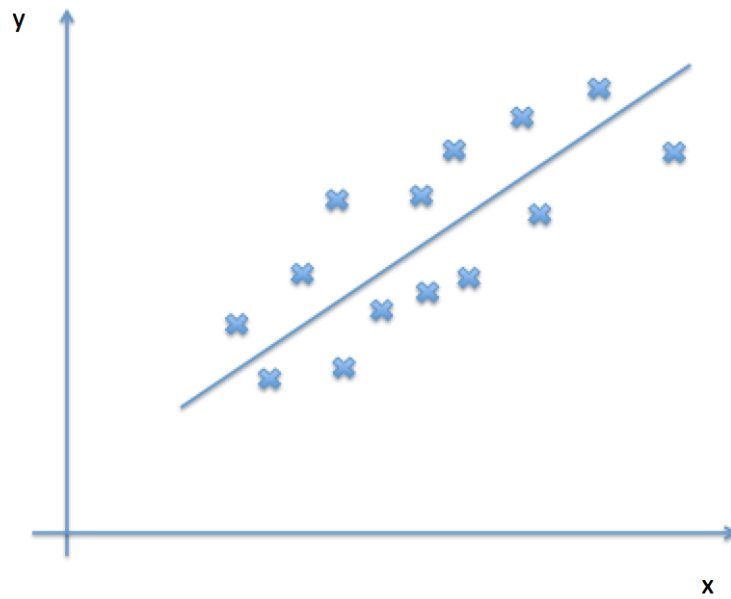
All u_i 's have the same variance, i.e. for $i = 1, \dots, n$

$$\text{Var}(u_i) = E(u_i^2) = \sigma^2$$

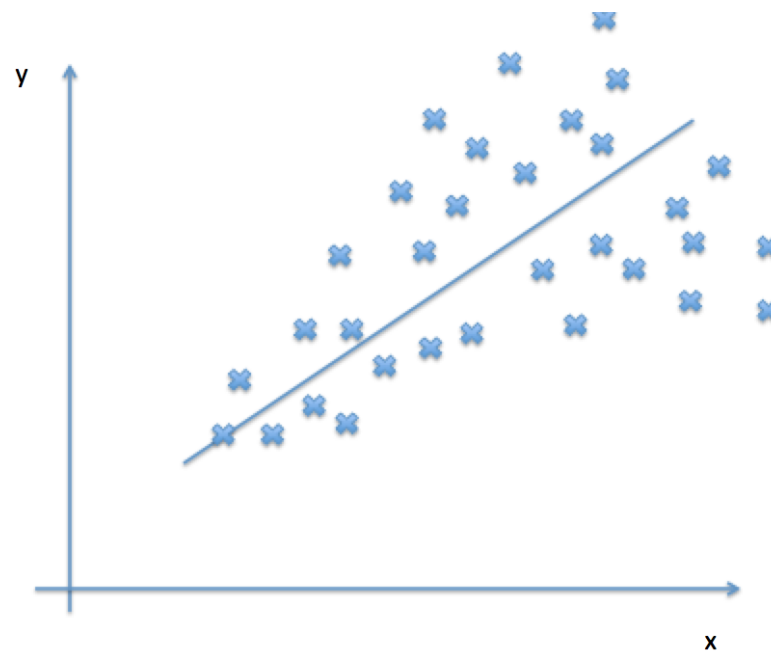
Variance of the errors given the independent variables, X , are constant

$$\text{Var}(u_i | X_i) = \sigma^2$$

Homoskedastic



Heteroskedastic



Why is it important?

Reason 1: BLUE doesn't hold

If there are heteroskedastic errors BLUE does not hold for OLS

OLS will no longer be the BEST

- B will not hold

Sampling variance will be lower in other estimators

- GLS
- WLS

Why is it important?

Reason 2: Standard errors will be biased

Software assumes homoskedasticity

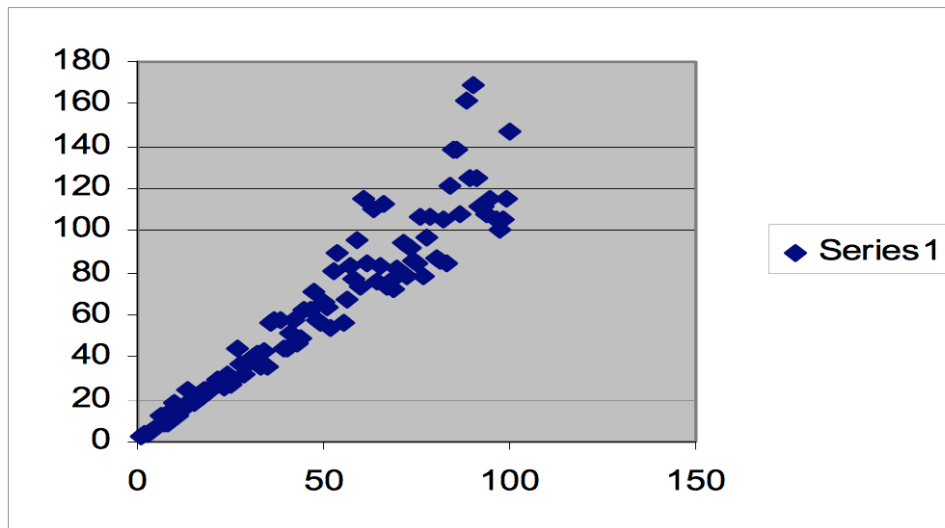
If there are heteroskedastic errors software does not take this into account

This means that standard errors will be larger than software calculates

- Inference in t-stats, f-stats, p-values etc. will thus be wrong

Heteroskedasticity

With real data it may look something like this....



When might this happen?

- Cross-sectional data on units of different size
 - e.g. states, cities. Omitted variables may be larger for more populous states or cities
- Cross-sectional data on units at different points in time
 - Omitted variables may be more important at some points in time
- Cross-sectional data on units that face different restrictions on their behaviour
 - For instance, high income individuals have more discretion in their spending

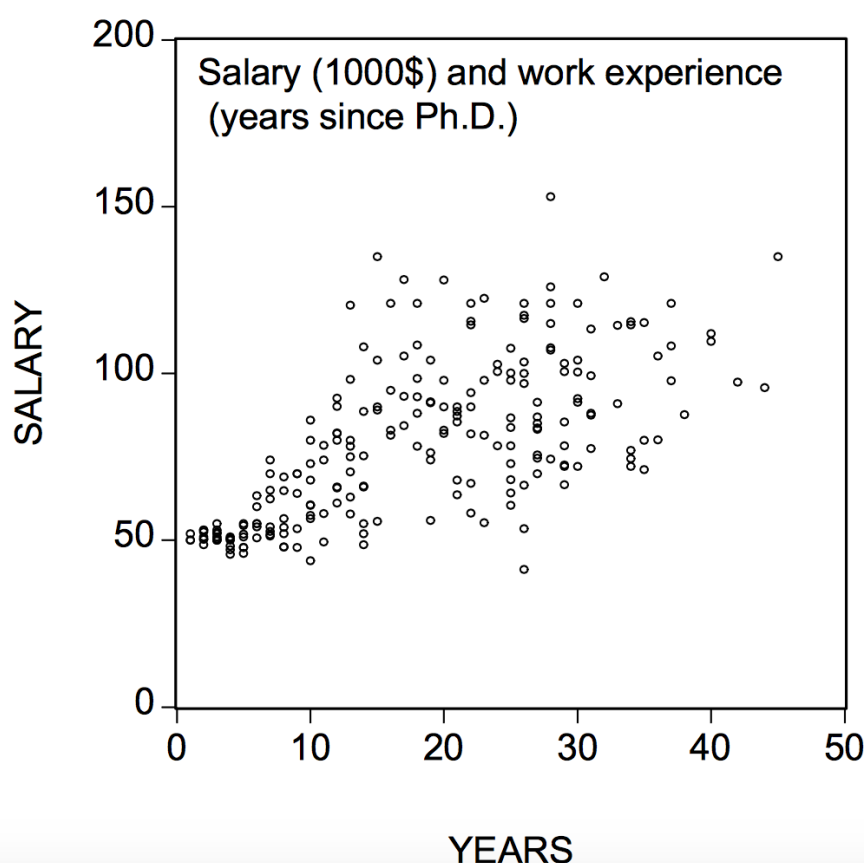
Classic example is the height of mice and basketball players....

Example

Relation between income and experience

Data on 222 university professors for 7 US schools (UC Berkeley, UCLA, UCSD, Illinois, Stanford, Michigan, Virginia)

- Variation in income (in 1000\$) increases with work experience
- Variation in relative income first increases and then decreases
 - Any form of variation with an X characteristic is heteroskedastic
 - Does not just have to be ever increasing



Heteroskedasticity

Common issue

Can be very problematic

We will:

- Look at how to test for it
- Look at how to treat it

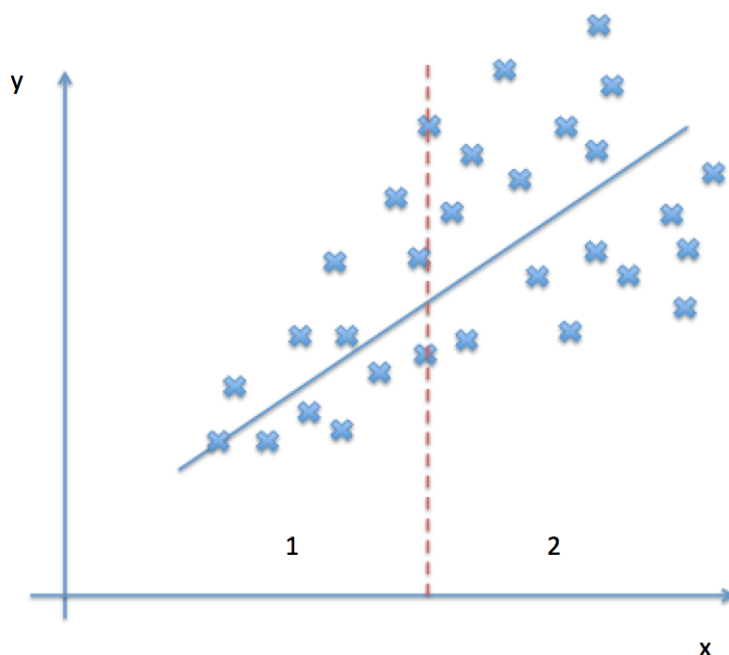
Testing: Plan A

Graphs:

- 1 Plot residuals against fitted values
- 2 Plot of squared OLS residuals against regressors
 - Try to find the problematic variable
- Residuals against Y: **rvfplot**
 - Plots the residuals against the fitted values
- Residuals against X: **rvpplot x**
 - Plots the residuals against the x variable selected

Testing: Plan B

Goldfeld-Quandt Test



Testing: Plan B

Can simply test if the sum of the squared errors in the first half of the sample is equal to the second half of the sample

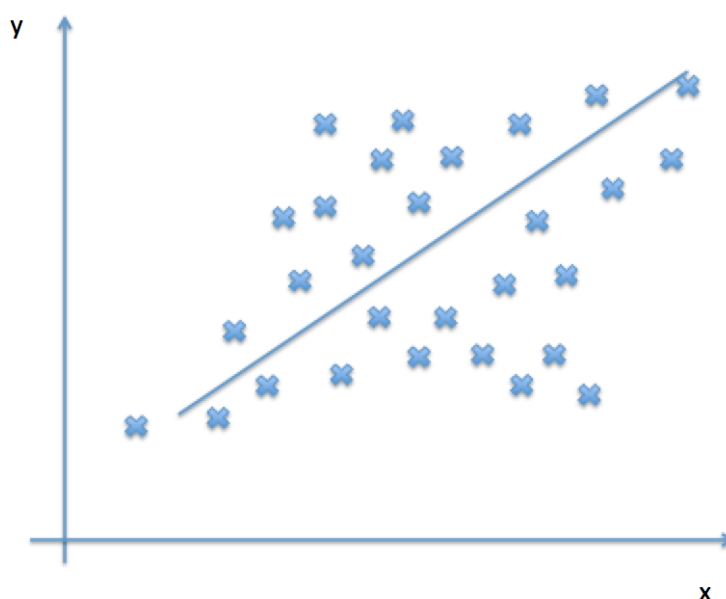
Formally:

$$\frac{\sum \hat{u}_2^2 / N_2}{\sum \hat{u}_1^2 / N_1}$$

Issues:

- ① Linear
- ② Tests one X variable

Non-linear Heteroskedasticity



Testing: Plan B

In STATA:

Run regression....

```
predict uhat, residual
```

```
estat sum X
```

- To find the mean of the X variable of interest

```
gen groupA=X< Mean from above
```

```
gen uhatsq = uhat^2
```

```
sdtest uhatsq, by(groupA)
```

Testing: Plan C

Breusch-Pagan Test

$$y = \alpha + \beta_1 X_1 + \dots + \beta_p X_p + u_i$$

If we have $\text{Var}(u_i | X_i) \neq \sigma^2$ then there is some relationship between the X terms and the u terms

We can test this.....

$$\hat{u}^2 = \delta_0 + \delta_1 X_1 + \dots + \delta_p X_p + v$$

If the δ term on any of the independent variables is significant..... there is a relationship between the error term and that variable

- i.e. there is heteroskedasticity

Testing: Plan C

Are we just interested in the relationship of one independent variable and the errors?

No!

So need to test that all the δ terms equal 0

Requires an F-test

Formally:

$$\frac{R^2 / \text{Number of Variables}}{(1 - R^2) / (\text{Degrees of freedom})}$$

where degrees of freedom = $N - P - 1$

Testing: Plan C

Breusch-Pagan Test

- H_0 : error variances are all equal
- H_1 : error variances are a multiplicative function of one or more variables

To test:

`regress income educ jobexp estat hettest`

Low Chi-square suggests not a problem (or wasn't a multiplicative function of the predicted values)

Testing: Plan D

White's general test for Heteroskedasticity

BP works well if linear forms of relationships between the X variables and the errors....but not for non-linear forms

Rather than...

$$\hat{u}^2 = \delta_0 + \delta_1 X_1 + \dots + \delta_p + v$$

Can include squares to give....

$$\hat{u}^2 = \delta_0 + \delta_1 X_1 + \dots + \delta_p + v + \gamma_1 X_1^2 + \dots + \gamma_k X_k^2$$

And interactions between X variables....

$$\hat{u}^2 = \delta_0 + \delta_1 X_1 + \dots + \delta_p + v + \gamma_1 X_1^2 + \dots + \gamma_k X_k^2 + \rho_1 X_1 + \dots + \rho_p$$

White simplifies this to:

$$\hat{u}^2 = \delta_0 + \delta_1 \hat{y}_1 + \delta_2 \hat{y}_1^2 + v$$

Similar to BP

- But adds many terms in the test regression
- Sometimes a simpler test like BP is more appropriate

Again:

- H_0 : error variances are all equal
- H_1 : error variances are a multiplicative function of one or more variables

To test:

`regress income educ jobexp estat imtest, white`

Dealing with heteroskedasticity: Option 1

Respecify the model / transform the variables

Heteroskedasticity can be a consequence from improper model specification

- e.g use logs

Dealing with heteroskedasticity: Option 2

Use Weighted Least Square (WLS)

The best option as you can model the heteroskedasticity

BUT....

- Need to know what to weight by
- Can take some playing around with....

Dealing with heteroskedasticity: Option 3

Use robust standard errors (Huber-White sandwich estimators)

Relaxes some OLS assumptions and gives better standard errors

`reg income educ jobexp, robust`

Robust standard errors can deal with a collection of minor concerns about failure to meet assumptions, such as minor problems about normality, heteroskedasticity, or some observations that exhibit large residuals, leverage or influence

- With the robust option, the point estimates of the coefficients are exactly the same as in ordinary OLS
- But the standard errors take into account issues concerning heterogeneity and lack of normality.

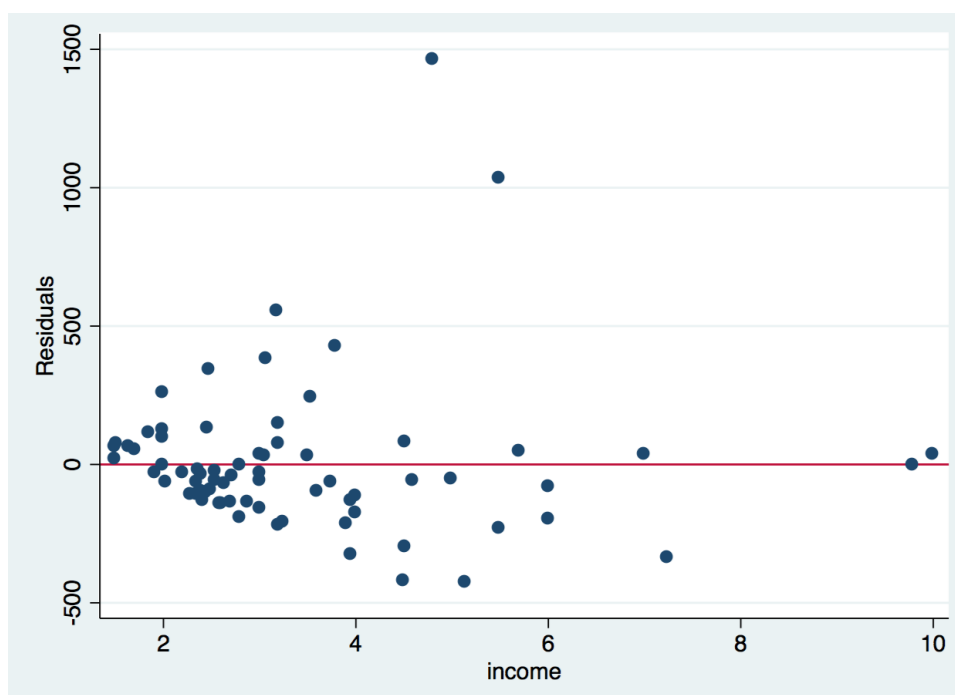
STATA Example

Going to need some data.....

use <http://www.ats.ucla.edu/stat/stata/ado/analysis/hetdata>,
clear

```
regress exp age ownrent income incomesq
```

```
rvpplot income, yline(0)
```



The residual versus income plot shows clear evidence of heterokcedasticity

Testing heteroskedasticity 1

estat hettest

```
Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
Ho: Constant variance
Variables: fitted values of exp

chi2(1)      =    29.23
Prob > chi2  =    0.0000
```

Testing heteroskedasticity 2

estat imtest, white

```
White's test for Ho: homoskedasticity
against Ha: unrestricted heteroskedasticity
```

```
chi2(12)      =    14.33
Prob > chi2    =    0.2802
```

```
Cameron & Trivedi's decomposition of IM-test
```

Source	chi2	df	p
Heteroskedasticity	14.33	12	0.2802
Skewness	7.89	4	0.0959
Kurtosis	1.67	1	0.1964
Total	23.88	17	0.1227

Fixing Heteroskedasticity

For big issues you should:

① Transform the model

or

② Run WLS

Normally it is enough to use robust standard errors

Fixing Heteroskedasticity

Robust standard errors

regress exp age ownrent income incomesq, robust

Linear regression

Number of obs = 72
F(4, 67) = 12.51
Prob > F = 0.0000
R-squared = 0.2436
Root MSE = 284.75

exp	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval]	
age	-3.081814	3.422641	-0.90	0.371	-9.913434	3.749805
ownrent	27.94091	95.56573	0.29	0.771	-162.8091	218.6909
income	234.347	92.12261	2.54	0.013	50.46954	418.2245
incomesq	-14.99684	7.199027	-2.08	0.041	-29.36616	-.6275257
_cons	-237.1465	220.795	-1.07	0.287	-677.8551	203.5621

Summary

- Heteroskedasticity occurs when the standard deviations of the error terms are not constant and depend on the x-value
- Does not bias estimates of β
- Does cause issues with standard errors
- Can test with graphs or formal tests (GQ, BP or White)
- Can fix with transformations, WLS or robust standard errors
- Robust is the easiest!

Take away questions

Plan is to try and 'tackle' some questions as groups, as a warm up for the assignment

And by tackle I mean:

- Why
- How
- Problems

So pick one or two and post your thoughts on the discussion board!

- ① What effect do interest rates have on the exchange rate?
- ② How does income affect population growth?
- ③ What impact does having a degree have on earnings?
- ④ How does the rate of incarceration affect crime?
- ⑤ Do areas with skill shortages lead to higher wages?
- ⑥ Did this years cohort of students do better than last years on BS2247?
- ⑦ Are men's wages systematically higher than women's?
- ⑧ Why do CEO's get paid so much?

Learning Outcomes

- ① Be able to review the limitations of simple liner regression analysis and the situations where it is not the most appropriate method to use;
 - What assumption does heteroskedasticity break?
 - What are the implications of this?
- ② Have developed an awareness of which models are appropriate to use when an OLS regression may not be robust, and be able to apply these models correctly;
- ③ Command an experienced understanding of the empirical approach to economics, and be able to undertake a robust piece of econometric analysis;
 - How do we deal with heteroskedasticity?
- ④ Be able to explain the covered econometric concepts clearly in concise non-technical language

READING

Have a look at the notes on logs etc on Blackboard as well as the simulation working through heteroskedasticity

Your questions please....?

Glossary

Dummy variable - Where the variable takes only one of two values which are mutually exclusive, i.e. Job status (In work/Not in work)

Categorical variable - An artificial variable created to represent an attribute with two or more distinct and complete categories/levels

Coefficient - The element we are trying to estimate in our work, to show the relationship between the phenomenon currently being studied, normally defined as β

Glossary

P-value - the probability that the coefficient estimated is significantly different to the null hypothesis. The standard null hypothesis in regression analysis is that $\beta = 0$, thus the lower the p-value, the more likely we are to reject the null and assume that $\beta \neq 0$

Heteroskedasticity - Variance of the errors given the independent variables, X, are not constant, variance of the errors does vary in some systematic manner

Standard errors - Recall that $\hat{\beta}_i$ is a point estimate of β_i . Because of sampling variability, this estimate may be too high or too low. The standard error of β_i , gives us an indication of how much this point estimate is likely to vary from the corresponding population parameter.