Introductory Econometrics

Heteroskedasticity

Monash Econometrics and Business Statistics

Semester 2, 2018

Recap

- We have studied the multiple regression model and learnt that when:

1. model is linear in parameters: $y = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$
2. conditional mean of errors is zero: $E(\mathbf{u} \mid \mathbf{X}) = \mathbf{0}$
3. columns of $\mathbf{X}$ are linearly independent

$\Rightarrow$ then the OLS estimator $\widehat{\boldsymbol{\beta}}$ is an unbiased estimator of $\boldsymbol{\beta}$

Recap

- We have studied the multiple regression model and learnt that when:

1. model is linear in parameters: $y = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$
2. conditional mean of errors is zero: $E(\mathbf{u} \mid \mathbf{X}) = \mathbf{0}$
3. columns of $\mathbf{X}$ are linearly independent

$\Rightarrow$ then the OLS estimator $\widehat{\boldsymbol{\beta}}$ is an unbiased estimator of $\boldsymbol{\beta}$

- if in addition,

4. sample is random and errors are homoskedastic: $Var(\mathbf{u} \mid \mathbf{X}) = \sigma^2 \mathbf{I}_n$,

$\Rightarrow$ then $\widehat{\boldsymbol{\beta}}$ is the BLUE and $Var(\widehat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$

Recap

- We have studied the multiple regression model and learnt that when:

1. model is linear in parameters: $y = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$
2. conditional mean of errors is zero: $E(\mathbf{u} \mid \mathbf{X}) = \mathbf{0}$
3. columns of $\mathbf{X}$ are linearly independent

$\Rightarrow$ then the OLS estimator $\widehat{\boldsymbol{\beta}}$ is an unbiased estimator of $\boldsymbol{\beta}$

- if in addition,

4. sample is random and errors are homoskedastic: $Var(\mathbf{u} \mid \mathbf{X}) = \sigma^2 \mathbf{I}_n$,

$\Rightarrow$ then $\widehat{\boldsymbol{\beta}}$ is the BLUE and $Var(\widehat{\boldsymbol{\beta}}) = \sigma^2 (\mathbf{X}'\mathbf{X})^{-1}$

- If, in addition to the above,

5. errors are normally distributed,

$\Rightarrow$ then conditional on $\mathbf{X}$, $\widehat{\boldsymbol{\beta}}$ is normally distributed, and we can use the usual t and F tests to make inferences based on the OLS estimator

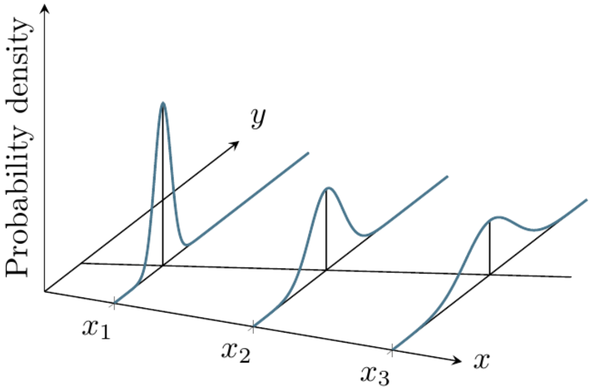
Lecture Outline

- Heteroskedasticity:
  1. Definition of heteroskedasticity and its consequences for OLS (textbook reference 8.1)
  2. Testing for heteroskedasticity (textbook reference 8.3)
     - 2.1 Breusch-Pagan test
     - 2.2 White test
  3. Heteroskedasticity robust standard errors (a simplified version of 8.2)
  4. Weighted least squares when heteroskedasticity is known up to a multiplicative constant (textbook reference 8.4, only the first part)
- We will not cover heteroskedasticity robust LM tests (the last part of section 8.2), Feasible GLS and the consequences of wrong specification of the variance function (the last two parts of section 8.4) and the linear probability model (section 8-5).

Heteroskedasticity

- Sometimes there is a good reason to doubt the assumption of equal variance for all errors. Here are some examples:
- In the study of food consumption, income is an important explanatory variable. It is unreasonable to assume that the variance of food consumption is the same for poor and rich people
- In many cases we do not have individual data (for confidentiality reasons), but we get information on averages over groups of individuals.
- For example, we can get incidences of crime per 1000 people, employment rate and income per capita in each district. These are averages, but different districts have different populations, so there is a good reason to believe that variances of these averages depend inversely on the population of each district
- In finance, some unpredicted news increase the volatility of the market (i.e. the variance of the market return) and this can last for several days (a large part of financial econometrics ETC3460 is about modelling this phenomenon)

- A 3D graphical representation of heteroskedasticity:



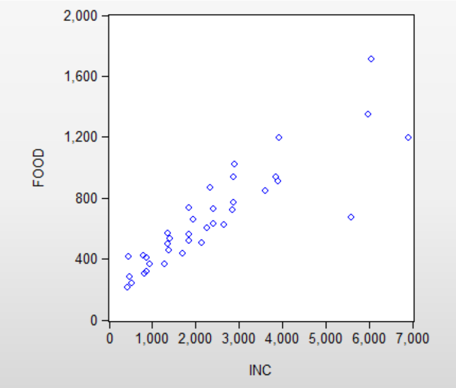
- In this example, the variance of u is getting larger as x increases

Consequences of heteroskedasticity (HTSK) for OLS

- HTSK does not affect the first 3 assumptions on the Recap slide, therefore **the OLS estimator will remain unbiased**
- HTSK violates assumption 4, therefore **the OLS estimator will no longer be BLUE and** $Var(\widehat{\beta}) \neq \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$. This means that the default standard errors reported by the statistical package for the OLS estimator will be incorrect
- Even if errors are Normally distributed, **the t and F tests based on the default OLS standard errors will be unreliable**
- Fortunately, if we detect HTSK, we have ways to conduct reliable inference based on the OLS estimator, or even obtain a more efficient estimator than the OLS estimator

Detecting HTSK

- As always, step 1: think about the problem!
- If we only have one x, the scatter plot can give us a clue:



- However, we hardly ever have only one x :-(

Testing for HTSK

- Since $E(u_i \mid x_{i1}, \ldots, x_{ik}) = 0$,
  $Var(u_i \mid x_{i1}, \ldots, x_{ik}) = E(u_i^2 \mid x_{i1}, \ldots, x_{ik})$
- If we suspect that variance can change with some subset of the independent variables, or even some exogenous variables that do not affect the mean, but can affect the variance, then, if we had u_i, we could square it and estimate the conditional expectation function of u_i^2. But u_i is unknown :-(
- Australian econometricians Trevor Breusch and Adrian Pagan, and the American econometrician Hal White showed that we can use the OLS residuals instead, and in large samples this will give us reliable results :-)



Breusch-Pagan test

$$\begin{aligned}
y_i &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + u_i \text{ for } i = 1, \ldots, n \\
H_0 &: E(u_i^2 \mid x_{i1}, x_{i2}, \ldots, x_{ik}) = \sigma^2 \text{ for } i = 1, \ldots, n \\
H_1 &: E(u_i^2 \mid x_{i1}, x_{i2}, \ldots, x_{ik}) = \delta_0 + \delta_1 z_{i1} + \delta_2 z_{i2} + \cdots + \delta_q z_{iq}
\end{aligned}$$

where $z_{i1}, \ldots, z_{iq}$ are a subset of $x_{i1}, x_{i2}, \ldots, x_{ik}$. In fact the z variables can include some variables that do not appear in the conditional mean, but may affect the variance (note the difference with the book)

1. Estimate the model by OLS as usual. Obtain $\hat{u}_i$ for $i = 1, \ldots, n$ and square them.
2. Regress $\hat{u}_i^2$ on a constant, $z_{i1}, \ldots, z_{iq}$. Denote the R^2 of this *auxiliary regression* by $R^2_{\hat{u}^2}$.
3. Under H_0, the statistic $n \times R^2_{\hat{u}^2}$ has a χ^2 distribution with q degrees of freedom in large samples. This statistic is called the Lagrange multiplier (LM) statistic for HTSK.
4. Given the desired level of significance, we obtain the cv of the test from the χ^2 table, and reject H_0 if the value of the test statistic is larger than the cv.

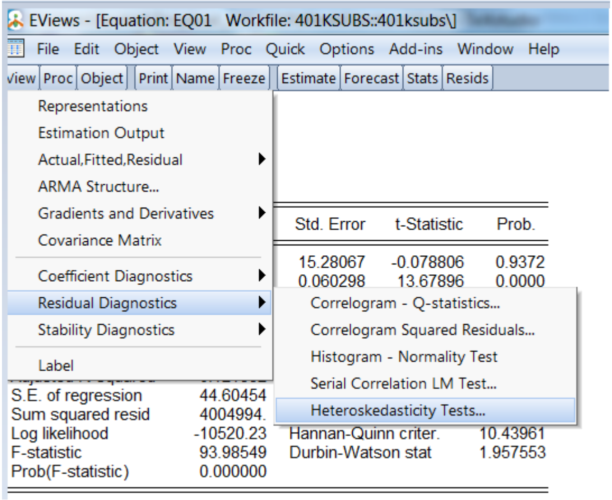
- Under H_0 the F test for the overall significance of the second regression has an $F_{q,n-q-1}$ distribution, and it can also be used to test for HTSK.
- Example: Net financial wealth in \$1,000s, predicted by age and current income in \$1,000s

Dependent Variable: NETTFA
Method: Least Squares
Included observations: 2017

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-1.2042	15.2807	-0.0788	0.9372
INC	0.8248	0.0603	13.6790	0.0000
AGE	-1.3218	0.7675	-1.7222	0.0852
AGE^2	0.0256	0.0090	2.8406	0.0045

R-squared	0.1229	Mean dependent var	13.5950
Adjusted R-squared	0.1216	S.D. dependent var	47.5906
S.E. of regression	44.6045	Akaike info criterion	10.4355
F-statistic	93.9855	Durbin-Watson stat	1.9576
Prob(F-statistic)	0.0000		

- Can save residuals, and run the OLS of $\hat{u}^2$ on a constant and all the explanatory variables, or we can use Eviews



- Choosing Breusch-Pagan test in Eviews produces:

Heteroskedasticity Test: Breusch-Pagan-Godfrey

F-statistic	4.5195	Prob. F(3,2013)	0.0036
Obs*R-squared	13.4946	Prob. Chi-Square(3)	0.0037
Scaled explained SS	1918.2328	Prob. Chi-Square(3)	0.0000

Test Equation:
Dependent Variable: RESID^2
Method: Least Squares
Included observations: 2017

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	7086.6544	11465.1527	0.6181	0.5366
INC	133.0597	45.2420	2.9411	0.0033
AGE	-591.6010	575.8554	-1.0273	0.3044
AGE^2	8.5657	6.7520	1.2686	0.2047

R-squared	0.0067	Mean dependent var	1985.6194
F-statistic	4.5195		
Prob(F-statistic)	0.0036		

- When learning, it is better to do the LM steps yourself rather than to use Eviews options in order to make sure you understand the mechanics of the test

White Test for HTSK

- White's test of course tests the same null hypothesis

$$H_0 : E(u_i^2 \mid x_{i1}, x_{i2}, \ldots, x_{ik}) = \sigma^2 \text{ for } i = 1, \ldots, n$$

  but its alternative is that the variance is a smooth unknown function of $x_{i1}, x_{i2}, \ldots, x_{ik}$.

- Hal White showed that a regression of $\hat{u}^2$ on a constant, x_1 to x_k, x_1^2 to x_k^2 and all pairwise cross products of x_1 to x_k, has the power to detect this general form of heteroskedasticity in large samples.

- Similar to the B-P test, White's test statistic is $n \times R^2_{\hat{u}^2}$, which under the null, has a χ^2 distribution with degrees of freedom as the number of explanatory variables in the auxiliary regression. We can also use the F-test of the overall significance of the auxiliary regression as well.

A Special Case of the White Test for HTSK

- A concern with White test is that the auxiliary regression will have $k + k(k+1)/2$ regressors, which is a very large number of restrictions
- Recall that the fitted values from OLS, $\hat{y}$ are a function of all the xs
- Thus, $\hat{y}^2$ will be a function of the squares and crossproducts of all the x's and $\hat{y}$ and $\hat{y}^2$ can proxy for all of the the x's their squares and cross products.
- A special form of the White test would be to regress the residuals squared on $\hat{y}$ and $\hat{y}^2$ and use the R^2 of this regression to form an F or LM statistic
- Note only testing for 2 restrictions now, no matter how many independent variables we have

- In the financial wealth example, choosing the White test in Eviews produces:

Heteroskedasticity Test: White

F-statistic	3.5660	Prob. F(8,2008)	0.0004
Obs*R-squared	28.2544	Prob. Chi-Square(8)	0.0004
Scaled explained SS	4016.2962	Prob. Chi-Square(8)	0.0000

Test Equation:
Dependent Variable: RESID^2
Method: Least Squares
Included observations: 2017

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	168619.	196178.	0.8595	0.3902
INC^2	0.35686	1.51943	0.2349	0.8143
INC*AGE	-62.1345	40.2406	-1.5441	0.1227
INC*AGE^2	0.88247	0.47739	1.8485	0.0647
INC	1091.73	793.845	1.3752	0.1692
AGE^2	874.156	719.881	1.2143	0.2248
AGE*AGE^2	-15.5884	11.3525	-1.3731	0.1699
AGE	-20478.0	19685.7	-1.0402	0.2984
AGE^2^2	0.09812	0.06521	1.5046	0.1326

R-squared	0.0140	Mean dependent var	1985.619
F-statistic	3.5660		2.779216
Prob(F-statistic)	0.0004		

- Eviews does not have a built in command for the alternate form of the White test, and we need to save the OLS residuals and OLS predictions and then run the auxiliary regression:

Dependent Variable: UHAT^2
Method: Least Squares
Included observations: 2017

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	277.66	1007.7	0.2755	0.7829
NETTFAHAT	29.622	93.166	0.3179	0.7506
NETTFAHAT^2	2.8195	1.7440	1.6167	0.1061
R-squared	0.0078	Mean dependent var		1985.6
F-statistic	7.8745			
Prob(F-statistic)	0.0004			

- The value of the $n \times R^2$ statistic is $2017 \times 0.0078 = 15.73$ which is larger than 5.99, the 5% critical value of the χ^2 distribution with 2 degrees of freedom. All tests lead to the same conclusion: There is significant evidence of HTSK, so, although OLS is still unbiased, it is no longer the best and OLS standard errors are unreliable

Solution 1: Robust Standard Errors

- Since OLS estimator is still unbiased, we may be happy to live with the OLS even if it is not BLUE. But the real practical problem is that t and F statistics based on OLS standard errors are unusable
- Recall the derivation of $Var(\widehat{\beta} \mid \mathbf{X})$:

$$Var(\widehat{\beta} \mid \mathbf{X}) = (\mathbf{X}'\mathbf{X})^{-1} \left[\mathbf{X}' Var(\mathbf{u} \mid \mathbf{X})\mathbf{X}\right] (\mathbf{X}'\mathbf{X})^{-1}$$

- With homoskedasticity,

$$Var(\mathbf{u} \mid \mathbf{X}) = \sigma^2 \mathbf{I}_n \Rightarrow Var(\widehat{\beta} \mid \mathbf{X}) = \sigma^2 (\mathbf{X}'\mathbf{X})^{-1}$$

- With HTSK

$$Var(\mathbf{u} \mid \mathbf{X}) = \begin{pmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{pmatrix}$$

- Therefore, with HTSK

$$Var(\widehat{\beta} \mid \mathbf{X}) = (\mathbf{X}'\mathbf{X})^{-1} \left[\mathbf{X}' \begin{pmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{pmatrix} \mathbf{X} \right] (\mathbf{X}'\mathbf{X})^{-1}$$

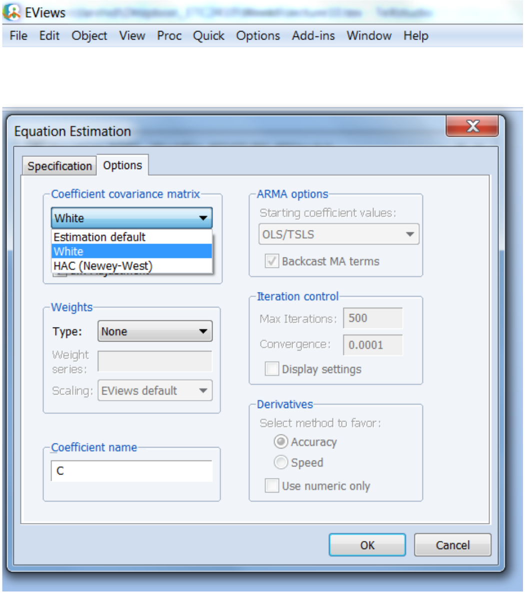
- Amazingly, White proved that:

$$\widehat{Var}(\widehat{\beta} \mid \mathbf{X}) = (\mathbf{X}'\mathbf{X})^{-1} \left[\mathbf{X}' \begin{pmatrix} \hat{u}_1^2 & 0 & \cdots & 0 \\ 0 & \hat{u}_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \hat{u}_n^2 \end{pmatrix} \mathbf{X} \right] (\mathbf{X}'\mathbf{X})^{-1}$$

  is a reliable estimator for $Var(\widehat{\beta} \mid \mathbf{X})$ in large samples

- The square root of diagonal elements of this matrix are called White standard errors or robust standard errors, which most statistical packages compute. These are reliable for inference.

- Back to the example. The option of robust standard errors is under the Options tab of the equation window:



- With this option, we get the following results:

Dependent Variable: NETTFA
Method: Least Squares
Included observations: 2017
White heteroskedasticity-consistent standard errors & covariance

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-1.2042	19.7337	-0.0610	0.9513
INC	0.8248	0.1039	7.9408	0.0000
AGE	-1.3218	1.1055	-1.1956	0.2320
AGE^2	0.0256	0.0141	1.8066	0.0710

R-squared	0.1229	Mean dependent var	13.5950
F-statistic	93.9855	Durbin-Watson stat	1.9576
Prob(F-statistic)	0.0000	Wald F-statistic	40.1225
Prob(Wald F-statistic)	0.0000		

- Compare with the original regression results:

Dependent Variable: NETTFA
Method: Least Squares
Included observations: 2017

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-1.2042	15.2807	-0.0788	0.9372
INC	0.8248	0.0603	13.6790	0.0000
AGE	-1.3218	0.7675	-1.7222	0.0852
AGE^2	0.0256	0.0090	2.8406	0.0045

R-squared	0.1229	Mean dependent var	13.5950
Adjusted R-squared	0.1216	S.D. dependent var	47.5906
S.E. of regression	44.6045	Akaike info criterion	10.4355
F-statistic	93.9855	Durbin-Watson stat	1.9576
Prob(F-statistic)	0.0000		

Solution 2: Transform the Model

a. *Logarithmic transformation of y may do the trick*: If the population model has $\log(y)$ as the dependent variable but we have used y, this kind of mis-specification can show up as heteroskedastic errors. So, if log-transformation is admissible (i.e. if y is positive), moving to a log model may solve the problem, and the OLS estimator on the log-transformed model will then be BLUE and standard errors will be useful. Of course when we consider transforming y, we should think if a log-level or a log-log model makes better sense

b. *Weighted least squares*: When there is good reason to believe that variance of each error is proportional to a known function of a single independent variable, then we can transform the model in a way to eliminate HTSK and then use OLS on the transformed model. This estimator is the *weighted least squares (WLS)* estimator, which we derive on the next slide.

Weighted Least Squares

- Suppose the model

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + u_i \text{ for } i = 1, \ldots, n \quad (1)$$

  satisfies the assumptions needed for unbiasedness of OLS, and we have

$$Var(u_i \mid x_{i1}, x_{i2}, \ldots, x_{ik}) = \sigma^2 h_i$$

  where h_i is a known function of one of x's, or a function of a variable z as long as $E(u_i \mid x_{i1}, x_{i2}, \ldots, x_{ik}, z_i) = 0$. For example $h_i = x_{i1}$, or $h_i = x_{i1}^2$, or $h_i = 1/z_i$.
- Multiplying both sides of equation (1) by $w_i = \frac{1}{\sqrt{h_i}}$ eliminates HTSK because:

- The transformed (or "weighted") model:

$$(w_i y_i) = \beta_0 w_i + \beta_1 (w_i x_{i1}) + \beta_2 (w_i x_{i2}) + \cdots + \beta_k (w_i x_{ik}) + (w_i u_i) \quad (2)$$

  satisfies all assumptions of the Gauss-Markov theorem, so the OLS estimator of its parameters is BLUE.

- More importantly, equation (2) has the same parameters as equation (1). So, OLS on the weighted model will produce BLUE of β_0 to β_k and we can test any hypotheses on these parameters based on the weighted model.

- Note that the transformed model does not have a constant term, and β_0 is the coefficient of w_i in the transformed model

- This estimator is called the weighted least squares (WLS) estimator of $\boldsymbol{\beta}$

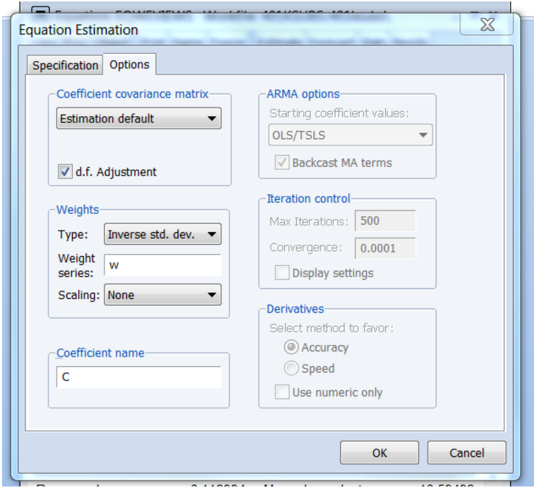
- In the financial wealth example, the auxiliary regression suggests that the variance changes with income. Since income is positive for all observations (why is this important?), we hypothesise that $Var(u_i \mid inc_i, age_i) = \sigma^2 inc_i$
- We create $w_i = \frac{1}{\sqrt{inc_i}}$ [Eviews command: `series w=1/@sqrt(inc)`] and we run the weighted regression

Dependent Variable: W*NETTFA
Method: Least Squares
Included observations: 2017

Variable	Coefficient	Std. Error	t-Statistic	Prob.
W	-3.7769	11.4808	-0.3290	0.7422
W*INC	0.7938	0.0627	12.6587	0.0000
W*AGE	-0.8890	0.5756	-1.5444	0.1227
W*(AGE^2)	0.0174	0.0067	2.5864	0.0098
R-squared	0.0810	Mean dependent var		2.1807

- The standard errors are now reliable for inference and for forming confidence intervals

- Eviews also has a built in WLS command under the option tab of the equation window. We need to enter the name of the weight series.



- The only advantage of the Eviews built in command is that it produces a set of statistics for the original model

Dependent Variable: NETTFA
Method: Least Squares
Included observations: 2017
Weighting series: W
Weight type: Inverse standard deviation (no scaling)

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-3.7769	11.4808	-0.3290	0.7422
INC	0.7938	0.0627	12.6587	0.0000
AGE	-0.8890	0.5756	-1.5444	0.1227
AGE^2	0.0174	0.0067	2.5864	0.0098

Weighted Statistics

R-squared	0.1070	Mean dependent var	2.1807

Unweighted Statistics

R-squared	0.1190	Mean dependent var	13.5950
Adjusted R-squared	0.1177	S.D. dependent var	47.5906
S.E. of regression	44.7027	Sum squared resid	4022642.0002

- Note that while heteroskedasticity is more common in cross-sectional data, it can also arise in time series data and adjustments to standard errors apply.

Summary

- The assumption of equal conditional variances for each observation may not be appropriate
- But OLS will still be unbiased even if the errors are heteroskedastic, however the usual OLS standard errors will not be correct
- We learnt how to test for heteroskedasticity
- If HTSK is found, we can still use OLS, but calculate standard errors that are robust to HTSK, and use those for inference
- If we have a reasonable idea that the HTSK is proportional to a single variable, we can use WLS, which will provide the best linear unbiased estimator for the parameters and a set of standard errors that can be used for inference