Introductory Econometrics

Heteroskedasticity

Monash Econometrics and Business Statistics

2022

Recap

The multiple regression model

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik} + u_i, \ \ i = 1, 2, ...n.$$

A1 model is linear in parameters: $y = X\beta + u$.

A2 columns of X are linearly independent.

A3 conditional mean of errors is zero: $E(u|X) = 0$.

A4 homoskedasticity and no serial correlation: $Var(u|X) = \sigma^2 I_n$.

A5 errors are normally distributed: $u|X \sim N(0, \sigma^2 I_n)$.

What's next?

If all assumptions hold, linear regression can do amazing things.

But what can we do if one of the assumptions does not hold?

Homoskedasticity

The multiple regression model

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik} + u_i, \;\; i = 1, 2, ...n.$$

A4 homoskedasticity and no serial correlation: $Var(u|X) = \sigma^2 I_n$.

$$Var(u|X) = \begin{pmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{pmatrix}$$

Lecture Outline

1. Definition of heteroskedasticity

A4 homoskedasticity and no serial correlation: $Var(u|X) = \sigma^2 I_n$.

A4(a) homoskedasticity: $Var(u_1|X) = ... = Var(u_n|X) = \sigma^2$.

A4(b) no serial correlation: $Cov(u_{i,} u_j|X) = 0$ for all $i \neq j$.

When A4(a) does not hold, the error terms in u are heteroskedastic:

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

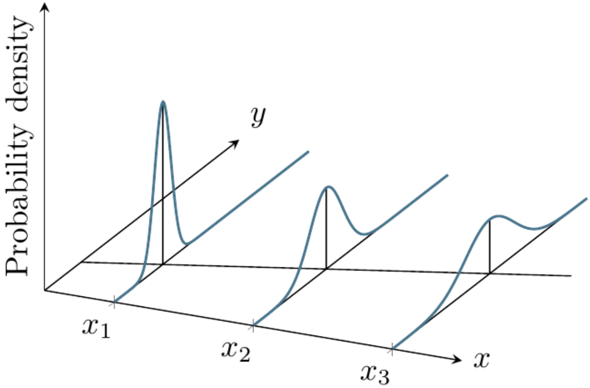
2. Causes of heteroskedasticity

Example:

- Study of household expenditure on food:
  The variation in expenditure on food for the population of high income households may be greater than the variation for the population of low income households.

$$Var(y|income = low) < Var(y|income = middle) < Var(y|income = high).$$

- A 3D graphical representation of heteroskedasticity:



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

2. Causes of heteroskedasticity

Other examples:

- Market volatility:
  The variance of a company's share price may be greater during periods of economic instability than during periods of economic stability.
- Information on group level instead of individual data:
  The variance of crime levels in different districts depends inversely on the population of each district.

3. Consequences of heteroskedasticity

- Heteroskedasticity does not affect A1-A3:
  - the OLS estimator remains unbiased.
- Heteroskedasticity violates A4:
  - the OLS estimator is no longer BLUE.
  - $Var(\hat{\beta}) \neq \sigma^2(X'X)^{-1}$.
    - default standard errors are incorrect.
      - default t and F tests are incorrect.

4. Detection heteroskedasticity

4.1 Informal analysis

4.2 The Breusch-Pagan test for heteroskedasticity

4.3 The White test for heteroskedasticity

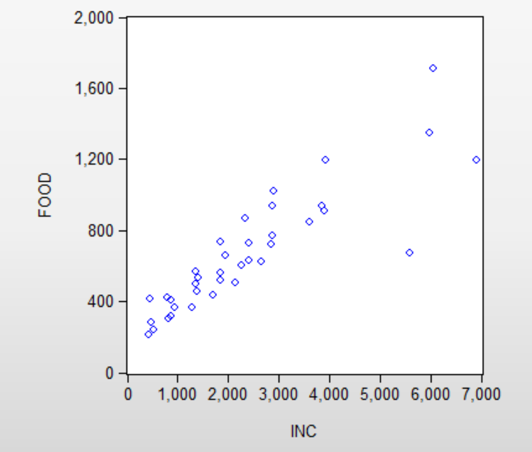
4.1 Informal analysis

Example:

- Study of household expenditure on food:
  The variation in expenditure on food for the population of high income households may be greater than the variation for the population of low income households.

$$Var(y|income = low) < Var(y|income = middle) < Var(y|income = high).$$

- Scatter plot of household food expenditure against income:



- The variation in expenditure on food does increase with income.

4.2 Testing for heteroskedasticity

Consider the linear regression equation

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik} + u_i, \;\; i = 1, 2, ...n.$$

- By definition,

$$Var(u_i | x_{i1}, \ldots, x_{ik}) = E(u_i^2 | x_{i1}, \ldots, x_{ik}) - [E(u_i | x_{i1}, \ldots, x_{ik})]^2 .$$

- Under A3 it holds that $E(u_i | x_{i1}, \ldots, x_{ik}) = 0$, so

$$Var(u_i | x_{i1}, \ldots, x_{ik}) = E(u_i^2 | x_{i1}, \ldots, x_{ik}).$$

- Under the homoskedasticity assumption,

$$Var(u_i | x_{i1}, \ldots, x_{ik}) = E(u_i^2 | x_{i1}, \ldots, x_{ik}) = \sigma^2, \;\; i = 1, 2, ..., n.$$

4.2 Testing for heteroskedasticity

Suppose $Var(u_i|x_{i1},\ldots,x_{ik})$ depends on a set of variables $(z_{i1},\ldots,z_{iq})$:

$$E(u_i^2|x_{i1},x_{i2},....,x_{ik}) \quad = \quad \delta_0+\delta_1 z_{i1}+\delta_2 z_{i2}+.....+\delta_q z_{iq}.$$

- The H_0 that $Var(u_i|x_{i1},\ldots,x_{ik})$ is constant, is equivalent to

$$H_0 : \delta_1=\delta_2=......=\delta_q=0.$$

- Because we don't observe u_i^2, we can't estimate the equation

$$\begin{aligned} u_i^2 &= E(u_i^2|x_{i1},x_{i2},....,x_{ik})+\epsilon_i \\ &= \delta_0+\delta_1 z_{i1}+\delta_2 z_{i2}+.....+\delta_q z_{iq}+\epsilon_i. \end{aligned}$$

4.2 The Breusch-Pagan test for heteroskedasticity

1. Obtain the OLS residuals $\widehat{u}_i$ for $i = 1, \ldots, n$ from the model:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + u_i, \; i = 1, \ldots, n.$$

2. Obtain the R-squared $R^2_{\widehat{u}^2}$ from the auxiliary regression:

$$\widehat{u}_i^2 = \gamma_0 + \gamma_1 z_{i1} + \gamma_2 z_{i2} + \cdots + \gamma_q z_{iq} + v_i, \; i = 1, \ldots, n.$$

3. Under $H_0 : \delta_1 = \cdots = \delta_q = 0$, we have the test statistic:

$$BP = nR^2_{\widehat{u}^2} \overset{asy}{\sim} \chi^2(q).$$

4. Reject H_0 in favor of $H_1 : \delta_j \neq 0$ for some j $= 1, \ldots, q$, if

$$BP_{calc} > \chi^2_{crit}(q).$$

4.2 The Breusch-Pagan test for heteroskedasticity

- An alternative way to conduct the BP test is to estimate

$$\widehat{u}_i^2 = \gamma_0 + \gamma_1 z_{i1} + \gamma_2 z_{i2} + ... + \gamma_q z_{iq} + v_i, \;\; i = 1, \ldots, n,$$

  and perform a standard F test of $H_0 : \gamma_1 = \gamma_2 = ... = \gamma_q = 0$.

- In both versions of the BP test, the variables $(z_{i1}, z_{i2}, ..., z_{iq})$:
  - can be a subset of the regressors $(x_{i1}, x_{i2}, ..., x_{ik})$.
  - can include variables that do not predict y.
  - as long as they may affect the variance.

4.2 The Breusch-Pagan test for heteroskedasticity

Example:

$$nettfa_i = \beta_0 + \beta_1 inc_i + \beta_2 age_i + \beta_3 age_i^2 + u_i, \;\; i = 1, \ldots, n,$$

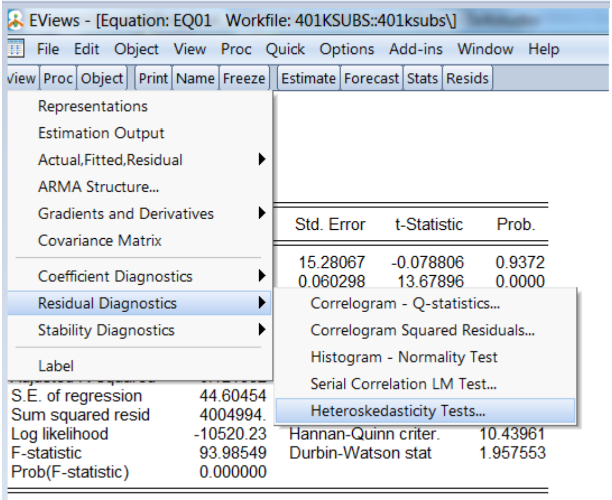
where *nettfa* is individual net financial assets (wealth) in \$1,000s, age is in years and inc is current income in \$1,000s.

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 steps yourself rather than to use Eviews options in order to make sure you understand the test.

4.3 The White test for heteroskedasticity

- The null hypothesis for the White test is the same as in BP:

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

- However, its alternative hypothesis is different:

$$H_1 : \text{ the variance is a smooth unknown function of } x_{i1}, \ldots, x_{ik}.$$

- 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.

4.3 The White test for heteroskedasticity

Example with $k = 3$:

1. Obtain the OLS residuals $\widehat{u}_i$ for $i = 1, \ldots, n$ from the model:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + u_i, \;\; i = 1, \ldots, n.$$

2. Obtain the R-squared $R^2_{\widehat{u}^2}$ from the auxiliary regression:

$$\begin{aligned} \widehat{u}_i^2 = \gamma_0 + \gamma_1 x_{i1} + \gamma_2 x_{i2} + \gamma_3 x_{i3} + \alpha_1 x_{i1}^2 + \alpha_2 x_{i2}^2 + \alpha_3 x_{i3}^2 \\ + \lambda_1 x_{i1} x_{i2} + \lambda_2 x_{i1} x_{i3} + \lambda_3 x_{i2} x_{i3} + v_i, \;\; i = 1, \ldots, n. \end{aligned}$$

3. Under $H_0 : \gamma_1 = \gamma_2 = \gamma_3 = \alpha_1 = \alpha_2 = \alpha_3 = \lambda_1 = \lambda_2 = \lambda_3 = 0$:

$$W = nR^2_{\widehat{u}^2} \overset{asy}{\sim} \chi^2(9).$$

4. Reject H_0 in favor of H_1 : conditional heteroskedasticity, if

$$W_{calc} > \chi^2_{crit}(9).$$

4.3 The White test for heteroskedasticity

The auxiliary regression may have $k + k(k+1)/2$ regressors.

- Omit the cross-terms from the auxiliary regression.
- Or use a special case of the White test with:

2. Estimate the following auxiliary regression in step 2 instead:

$$\widehat{u}_i^2 = \gamma_0 + \gamma_1 \widehat{y}_i + \gamma_2 \widehat{y}_i^2 + v_i, \; i = 1, \ldots, n,$$

   where $\widehat{y}_i$ is the predicted value of y_i from the model in step 1.

3. Step 3 and 4 test $H_0 : \gamma_1 = \gamma_2 = 0$ versus $H_1 : \gamma_1$ and/or $\gamma_2 \neq 0$.

The logic behind using $\widehat{y}_i$ at step 2 is that

$$\begin{aligned}
\widehat{y}_i &= \widehat{\beta}_1 x_{i1} + \widehat{\beta}_2 x_{i2} + \widehat{\beta}_3 x_{i3,} \\
\widehat{y}_i^2 &= (\widehat{\beta}_1 x_{i1} + \widehat{\beta}_2 x_{i2} + \widehat{\beta}_3 x_{i3,})^2 \\
&= \widehat{\beta}_1^2 x_{i1}^2 + \widehat{\beta}_2^2 x_{i2}^2 + \widehat{\beta}_3^2 x_{i3}^2 + 2\widehat{\beta}_1 \widehat{\beta}_2 x_{i2} x_{i1} + 2\widehat{\beta}_1 \widehat{\beta}_3 x_{i1} x_{i3} + 2\widehat{\beta}_2 \widehat{\beta}_3 x_{i2} x_{i3}.
\end{aligned}$$

- 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 the alternate form of the White test.
- Save the OLS residuals and predictions and run the 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			

- $n \times R^2 = 2017 \times 0.0078 = 15.73 > 5.99 = 5\%$ critical value $\chi^2(2)$.

5 Heteroskedasticity-robust tests

Recall that the two consequences of heteroskedasticity are:

- The OLS estimator of β is no longer BLUE.
- The standard t and F tests are no longer valid.

So we cannot conduct reliable hypothesis tests anymore!

- Hal White proposed alternative hypothesis tests which are valid in large samples, even when heteroskedasticity is present.

5.1 Heteroskedasticity-robust t tests

In the presence of heteroskedasticity

- the standard t test statistic for testing $H_0 : \beta_j = 0$ is

$$\frac{\widehat{\beta}_j}{se(\widehat{\beta}_j)} \not\sim t_{(n-k-1)}.$$

- the White t test statistic for testing $H_0 : \beta_j = 0$ is

$$\frac{\widehat{\beta}_j}{se^w(\widehat{\beta}_j)} \overset{asy}{\sim} t_{(n-k-1)}.$$

$se(\widehat{\beta}_j)$ is the OLS standard error and $se^w(\widehat{\beta}_j)$ the White standard error.

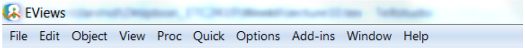
5.1 Heteroskedasticity-robust t tests

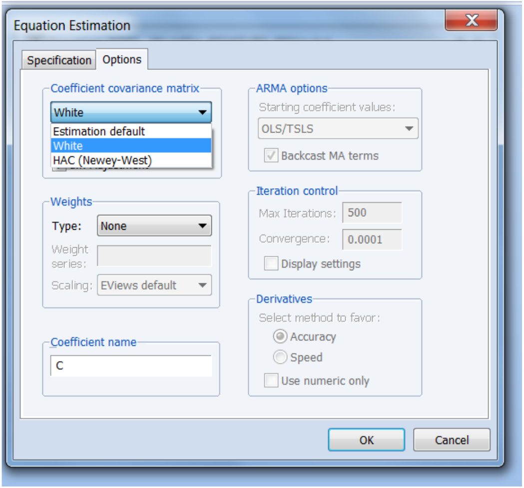
Example: the bivariate linear regression model

$$y_i = \beta_0 + \beta_1 x_i + u_i$$

- $$se(\widehat{\beta}_j) = \frac{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \widehat{\sigma}^2}}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \text{ with } \widehat{\sigma}^2 = \frac{SSR}{(n-2)}$$
- $$se^w(\widehat{\beta}_j) = \frac{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \widehat{u}_i^2}}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \text{ with } \widehat{u}_i^2 = (y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i)^2$$

- Back to the financial wealth 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		

5.2 Heteroskedasticity-robust F tests

- F test for testing multiple linear restrictions on the coefficients.
- The heteroskedasticity robust version of the F statistic is a called a heteroskedasticity-robust Wald statistic.
- Eviews reports this statistic if we choose the White option.
- Example: value of the statistic is 40.1225 with a p-value of 0.00.

5 Heteroskedasticity-robust tests

Why not always use heteroskedastic-robust test statistics?

- OLS standard errors are only valid under homoskedasticity.
- White's standard errors are valid with homo- or heteroskedasticity.
- We are never certain whether homoskedasticity holds.

However,

- The tests based on OLS standard errors are exact tests.
- Heteroskedasticity-robust tests are asymptotic.
- In small samples, heteroskedasticity-robust tests may be misleading.

Summary

- The assumption of homoskedastic errors may not be appropriate.
- 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 heteroskedasticity is found, we can still use OLS.
- But we use robust standard errors for inference.