Heteroscedasticity Problem

1 Nature and Consequences of heteroscedasticity for OLS

- Heteroskedasticity (broadly) -

- Heteroskedasticity (in econometrics) -

1.1 Nature of Heteroskedasticity

1.2 Consequences of Heteroskedasticity

1. Does not affect the biasedness of the OLS estimators

2. Does not affect the value of R^2 and $adj.R^2$

3. Make the estimated value of $Var(\widehat{\beta}_{OLS})$ wrong

4. Affect the correctness of our inference

1.3 How can the estimated value of $Var(\widehat{\beta}_{OLS})$ be wrong?

Suppose

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

Given that assumption 1 to 4 are true, but assumption 5 (homoskedasticity) is violated. Thus,

$$Var(u_i|x_i) =$$

And from the OLS estimation steps, we can write

$$\begin{aligned}
\hat{\beta}_1 &= \beta_1 + \frac{\sum_i (x_i - \overline{x}) u_i}{\sum_i (x_i - \overline{x})^2} \\
Var(\hat{\beta}_1) &= \frac{1}{\left(\sum_i (x_i - \overline{x})^2\right)^2} Var(\sum_i (x_i - \overline{x}) u_i) \\
&= \frac{1}{SST^2} Var[(x_1 - \overline{x})u_1 + (x_2 - \overline{x})u_2 + ... + (x_n - \overline{x})u_n] \\
&= \frac{1}{SST^2} [Var(x_1 - \overline{x})u_1 + Var(x_2 - \overline{x})u_2 + ... + Var(x_n - \overline{x})u_n] \\
&= \frac{1}{SST^2} [(x_1 - \overline{x})^2 \sigma_1^2 + (x_2 - \overline{x})^2 \sigma_2^2 + ... + (x_n - \overline{x})^2 \sigma_n^2] \\
&= \frac{\sum_i (x_i - \overline{x}) \sigma_i^2}{SST^2} \text{ if we have heteroskedasticity!}
\end{aligned}$$

Compared with

$Var(\hat{\beta}_1) =$ under homoskedasticity.

1.4 Two types of remedies

1. Passive

2. Active

2 Testing for heteroskedasticity

- The main point -

Suppose

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + u$$

Assume that assumption 1 to 4 are true. Our hypotheses to test for heteroskeroskedasticity would be

We know that $Var(u|\mathbf{x}) = E(u^2|\mathbf{x}) - [E(u|\mathbf{x})]^2$. But _____ according to assumption 4. Thus, H_0 and H_a can be written as

2.1 Breusch-Pagan test (BP test)

To perform the Breusch-Pagan Test in STATA

STATA commands (in case $k = 4$):

regress y x_1 x_2 x_3 x_4

predict u_hat, residual

generate u_hat_sq = u_hat^2

regress u_hat_sq x_1 x_2 x_3 x_4

** Then, check the F-statistic on the top right-hand corner of the result table.

Example: Finding the determinants of GPA.

```
. regress  termgpa attend priGPA final frosh soph
```

Source	SS	df	MS		
Model	226.077541	5	45.2155081	Number of obs	= 680
Residual	142.319996	674	.211157264	F(5, 674)	= 214.13
				Prob > F	= 0.0000
Total	368.397537	679	.542558964	R-squared	= 0.6137
				Adj R-squared	= 0.6108
				Root MSE	= .45952

termgpa	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
attend	.046594	.0036082	12.91	0.000	.0395093	.0536787
priGPA	.5329307	.0403281	13.21	0.000	.4537468	.6121146
final	.0503197	.0040339	12.47	0.000	.0423992	.0582403
frosh	.0974307	.0560211	1.74	0.082	-.0125662	.2074276
soph	.0689273	.0467006	1.48	0.140	-.0227689	.1606236
_cons	-1.361077	.1316861	-10.34	0.000	-1.619642	-1.102513

```
. predict u_hat, residual

. generate u_hat_sq = u_hat^2

. regress u_hat_sq  attend priGPA final frosh soph

regress u_hat_sq attend priGPA final frosh soph
```

Source	SS	df	MS	Number of obs =	680
				F(5, 674) =	14.54
Model	8.22606613	5	1.64521323	Prob > F =	0.0000
Residual	76.2624962	674	.113149104	R-squared =	0.0974
				Adj R-squared =	0.0907
Total	84.4885623	679	.124430872	Root MSE =	.33638

u_hat_sq	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
attend	-.0088079	.0026413	-3.33	0.001	-.0139941	-.0036218
priGPA	-.1454432	.029521	-4.93	0.000	-.2034074	-.0874791
final	.0061879	.0029529	2.10	0.036	.0003899	.0119859
frosh	-.1077493	.0410085	-2.63	0.009	-.1882692	-.0272294
soph	-.0975658	.0341858	-2.85	0.004	-.1646892	-.0304423
_cons	.7368933	.0963968	7.64	0.000	.5476191	.9261674

Alternatively, you can use the following set of STATA commands:
regress y x_1 x_2 x_3 x_4
estat hettest x_1 x_2 x_3 x_4

```
. regress termgpa attend priGPA final frosh soph

      Source |       SS       df       MS              Number of obs =     680
-------------+------------------------------           F(  5,   674) =  214.13
       Model |  226.077541     5  45.2155081            Prob > F      =  0.0000
    Residual |  142.319996   674  .211157264           R-squared     =  0.6137
-------------+------------------------------           Adj R-squared =  0.6108
       Total |  368.397537   679  .542558964           Root MSE      =  .45952

------------------------------------------------------------------------------
     termgpa |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      attend |    .046594   .0036082    12.91   0.000     .0395093    .0536787
      priGPA |   .5329307   .0403281    13.21   0.000     .4537468    .6121146
       final |   .0503197   .0040339    12.47   0.000     .0423992    .0582403
       frosh |   .0974307   .0560211     1.74   0.082    -.0125662    .2074276
        soph |   .0689273   .0467006     1.48   0.140    -.0227689    .1606236
       _cons |  -1.361077   .1316861   -10.34   0.000    -1.619642   -1.102513
------------------------------------------------------------------------------

. estat hettest attend priGPA final frosh soph

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: attend priGPA final frosh soph

         chi2(5)      =    93.90
         Prob > chi2  =   0.0000
```

If the null hypothesis is rejected (we have the heteroskedasticity problem), we can use the ", robust" option in STATA. This option gives us the correct standard error, or "heteroskedasticity-robust standard error". We can now use the t-statistics in this case.

```
. regress termgpa attend priGPA final frosh soph, robust

Linear regression                               Number of obs =     680
                                                F(  5,   674) =  234.28
                                                Prob > F      =  0.0000
                                                R-squared     =  0.6137
                                                Root MSE      =  .45952

------------------------------------------------------------------------------
             |               Robust
     termgpa |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      attend |    .046594   .0044101    10.57   0.000     .0379348    .0552532
      priGPA |   .5329307   .0426426    12.50   0.000     .4492023     .616659
       final |   .0503197   .0041066    12.25   0.000     .0422564     .058383
       frosh |   .0974307   .0633543     1.54   0.125    -.0269648    .2218262
        soph |   .0689273   .0520495     1.32   0.186    -.0332713    .1711259
       _cons |  -1.361077   .1448208    -9.40   0.000    -1.645431   -1.076723
------------------------------------------------------------------------------
```

2.2 The White Test

Similar to the Breush-Pagan test, but is stricter because it does not allow $\widehat{u}^2$ to be correlated with x^2 or interactions among different x_s.

The White Test (special case) (save degree of freedom)

1. Get $\hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta}_k$ by OLS.
2. Calculate $\hat{u}_i^2 = [y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + ... + \hat{\beta}_k x_k)]^2$
3. Calculate $\hat{y}_i = (\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + ... + \hat{\beta}_k x_k)$
4. Calculate $\hat{y}_i^2$
5. Estimate $\hat{u}_i^2 = \gamma_0 + \gamma_1 \hat{y}_i + \gamma_2 \hat{y}^2 + error$ (keep R^2 of this regression)
6. $LM = nR^2$
7. If $p - value > significance\ level$, cannot reject H_o.

3 Remedial measures

As mentioned before, there are 2 types of remedies – passive and active.

- The passive remedies just re-calculate the *std.err.* or $\hat{\beta}$ using the heteroskedasticity-robust standard error formula(s).
- The active remedies include the "weighted least squares (WLS) estimators", "generalized least squares (GLS) estimators", or "feasible GLS estimator".

3.1 Weighted Least Squares (WLS)

We assume that the heteroskedasticity may take the pattern

From

$$\begin{aligned} Var(u_i|\mathbf{x}) &= E(u_i^2|\mathbf{x}) - [E(u_i|\mathbf{x})]^2 \\ &= \end{aligned}$$

We get

To make the error term become homoskedastic, we weight every term by $\sqrt{h_i}$.

Lab 2 – Dummy, Heteroskedasticity, Specification Issues

1 The scaling issue.

1. Download the datafile "BWGHT.dta" from your EE325 Moodle page.
2. Open the STATA software program
3. type: regress bwghtlbs cigs faminc
4. type: gen faminc_thb = faminc*33200
5. type: regress bwghtlbs cigs faminc_thb

2 Does "beauty" help increase wage?

1. Download the datafile "beauty.xlsx" from your EE325 Moodle page.
2. Open the STATA software program. Click on the "Data Editor" icon.
3. Copy the entire dataset from the excel file and paste it onto the STATA's Data Editor page.

   Choose "Treat first row as variable names".
4. Save the new STATA dataset.

   Choose File -> Save As -> (then name the new dataset "beauty_stata")
5. Open a new do-file and save it.

   Choose "New Do-file Editor" icon

   On the Do–file's top panel, choose File -> Save As -> (then name the new do-file "second_stata_lab")

   - **To explore and understand the data**
6. type: browse
7. type: sum
8. type: tab look
9. type: tab look female

   - **We want to test whether "good look" has a positive impact on wage**
10. type: gen belavg = 0
11. type: replace belavg = 1 if look < 3

12. type: gen log_wage = log(wage)
13. regress log_wage belavg abvavg

    - **Seems like we may have the omitted variable bias. Let's take into account other variables.**

14. type: regress log_wage abvavg belavg educ
15. type: regress log_wage abvavg belavg educ exper expersq
16. type: regress log_wage abvavg belavg educ exper expersq bigcity
17. type: regress log_wage abvavg belavg educ exper expersq bigcity black
18. type: regress log_wage abvavg belavg educ exper expersq bigcity black union
19. type: regress log_wage abvavg belavg educ exper expersq bigcity black union female

    - **Now, let's export the data into a formal format**

20. type: ssc install outreg2

    (This command is to install a command called "outreg2". Once you install this command, your computer will recognize it. So, no need to reinstall in the future.)

21. type: regress log_wage abvavg belavg educ
22. type: outreg2 using statalab2.docx
23. type: regress log_wage abvavg belavg educ exper expersq bigcity
24. type: outreg2 using statalab2.docx, append

    - **Check if we have the heteroskedasticity problem. (Let's use the White Test (special case))**

25. type: regress log_wage abvavg belavg educ exper expersq bigcity black union female
26. type: predict u_hat, resid
27. type: predict y_hat, xb
28. type: gen u_hat_sq = u_hat^2
29. type: gen y_hat_sq = y_hat^2
30. type: regress u_hat_sq y_hat y_hat_sq
31. Calculate $LM = nR^2$
32. Do we reject H_0 : homoskedasticity at 5% level of confidence?

33. Now, try using the ready-made test by STATA
34. type: regress log_wage abvavg belavg educ exper expersq bigcity black union female
35. type: estat hettest
36. Do we reject H_0 : homoskedasticity at 5% level of confidence?

- **Should we believe that the value of β are the same for female and male?** (Chow Test)
- Chow statistic is a type of F-statistic $F = \frac{SSR_p-(SSR_1+SSR_2)}{SSR_1+SSR_2} \cdot \frac{[n-2(k+1)]}{k+1}$

37. To get SSR_p: regress log_wage abvavg belavg educ exper expersq bigcity black union
38. To get SSR_1: regress log_wage abvavg belavg educ exper expersq bigcity black union if female == 0
39. To get SSR_2: regress log_wage abvavg belavg educ exper expersq bigcity black union if female == 1
40. What is the value of the F-statistic? Can we reject H_0 (can use the same model)?

3 Labor Force Participation of Female

1. Download the "MORA.DTA" dataset from your EE325 Moodle page and open it in STATA.

   Choose File -> Open -> (then direct to the location of the file)
2. type: des
3. type: regress inlf nwifeinc educ exper expersq age kidslt6 kidsge6
4. type: estat hettest
5. type: regress inlf nwifeinc educ exper expersq age kidslt6 kidsge6, robust
6. type: predict y_hat, xb
7. type: twoway scatter inlf educ || line y_hat educ
8. type: sort educ

9. type: twoway scatter inlf educ || line y_hat educ

   - **We need to "hold other things constant". Suppose nwifeinc = 30, exper = 10, age = 35, kidslt6 = 0, kidsage6 = 0.**

10. type: gen y_new = 0.585 + 30*(-0.0034) + educ*(0.0379) + 10*(0.0395) + 100*(-0.0006) + 35*(-0.0161)

11. type: twoway scatter inlf educ || line y_new educ

Serial Correlation and Heteroskedasticity in Time Series Regressions

1 The Nature of Time Series Data

TABLE 10.1

Partial Listing of Data on U.S. Inflation and Unemployment Rates, 1948–2003

Year	Inflation	Unemployment
1948	8.1	3.8
1949	−1.2	5.9
1950	1.3	5.3
1951	7.9	3.3
⋮	⋮	⋮
1998	1.6	4.5
1999	2.2	4.2
2000	3.4	4.0
2001	2.8	4.7
2002	1.6	5.8
2003	2.3	6.0

2 Examples of Time Series Regression Models

There are many time series regression models. Different models would be suitable for different types of relationship we want to estimate. Some examples of time series models are Static Model, AR (Autoregressive), ADL (Autoregressive Distributed Lag), FDL (Finite Distributed Lag), ARMA (Autoregressive Moving Average), ARCH (Autoregressive Conditional Heteroskedasticity) etc.

In this class we will talk about 2 examples 1) Static Models and 2) FDL.

2.1 Static Models

Studies a contemperaneous (occurring in the same period of time) relationship of variables.

For example:

2.2 Finite Distributed Lag Models

For example,

In general,

$$y_t = \alpha_0 + \delta_0 x_t + \delta_1 x_{t-1} + \delta_2 x_{t-2} + u_t$$

$\delta_0 = \frac{dy_t}{dx_t}$

$\delta_1 = \frac{dy_t}{dx_{t-1}}$

3 Properties of OLS under classical assumptions

Assumption TS1. Linear in Parameter – Y is linear in X.
Assumption TS2. No Perfect Collinearity
Assumption TS3. Zero Conditional Mean

***** Under Assumptions TS1 to TS3, $\hat{\beta}_{OLS}$ would be unbiased *****

Assumption TS4. Homoskedasticity
Assumption TS5. No Serial Correlation

***** Under Assumptions TS1 to TS5, $\hat{\beta}_{OLS}$ would be BLUE (best linear unbiased estimators)*****

The variance of $\hat{\beta}_{OLS}$ (under ass.TS1-TS5)

4 Properties of OLS with Serially Correlated Errors

5 Unbiasedness and Consistency

6 Efficiency and Inference

With serial correlation, $\hat{\beta}_{OLS}$ would not be BLUE ($var(\hat{\beta}_{OLS})$ would not be minimized). Consider

$$u_t = \rho u_{t-1} + e_t \ ; \ t = 1, 2, ..., n \quad \text{and} \quad |\rho| < 1$$

where u_t is from a regression model

$$y_t = \beta_0 + \beta_1 x_t + u_t.$$

7 Testing for Serial Correlation

Given the model

$$y_t = \beta_0 + \beta_1 x_{t1} + \beta_2 x_{t2} + ... + \beta_k x_{tk} + u_t$$

7.1 A "t-test" for AR(1) serial correlation with strictly exogeneous regressors

The most common type of serial correlation or autocorrelation is the AR(1) type:

To perform the test:

1. Estimate $y_t = \beta_0 + \beta_1 x_{t1} + \beta_2 x_{t2} + ... + \beta_k x_{tk} + u_t$
2. Obtain $\hat{u}_t$,$\hat{u}_{t-1}$; $\forall\ t = 1, 2, ..., n$
3. Estimate $\hat{u}_t = \rho\ \hat{u}_{t-1} + error$
4. Perform the $t - test$ for

7.2 The Durbin-Watson Test (DW test)

This implies

$$\begin{aligned} \hat{\rho} &= 0 \quad \Rightarrow \quad DW = 2 \\ \hat{\rho} &> 0 \quad \Rightarrow \quad DW < 2 \\ \hat{\rho} &< 0 \quad \Rightarrow \quad DW > 2 \end{aligned}$$

$$\begin{aligned} H_o &: \text{ no positive autocorrelation, serial-correlation} \\ H_a &: \text{ no negative serial correlation} \end{aligned}$$

To perform the test:

1. Estimate $y_t = \beta_0 + \beta_1 x_{t1} + \beta_2 x_{t2} + ... + \beta_k x_{tk} + u_t$
2. Obtain $\hat{u}_t$,$\hat{u}_{t-1}$; $\forall\ t = 1, 2, ..., n$
3. Calculate DW from eq.(2)
4. Find the critical d_L and d_u values (say, at the 5% level of significance) for the given sample size and # of regressors.
5. Follow the decision rule in the picture.

Example:
Suppose the calculated value of $DW = 0.80$, $n = 45$, $k = 4$.
From this, we get $d_L =$ _____ and $d_u =$ _____

8 Correcting for serial correlation

8.1 Passive way

Use the type of standard error that is robust to the serial correlation, autocorrelation problem

8.2 Active way –

Multicolinearity

1 The Nature of Multicolinearity

-

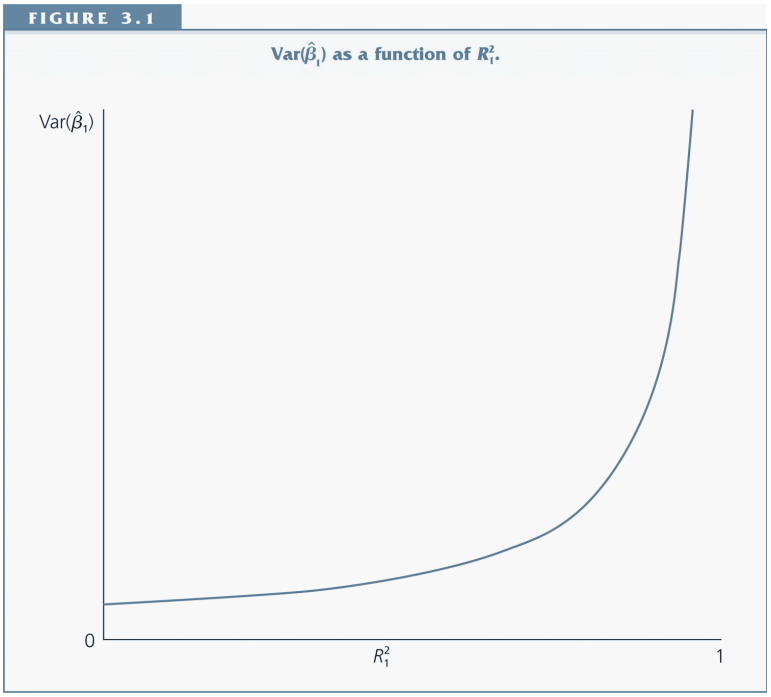
observation	x_1	x_2	$3x_1 - x_2$
1	6	18	0
2	12	36	0
3	7	21	0
4	-5	-15	0

observation	x_1	x_2	$3x_1 - x_2$
1	6	16	-2
2	12	45	9
3	7	18	-3
4	-5	-12	3

2 Consequences of Multicolinearity

2.1 The OLS estimator will still be BLUE.

2.2 The variances and covariances will be very large. This makes precise estimation difficult.



3 Detection of multicolinearity

1. There is conflicting test between t- and F-test: if we find that the conclusion derived from the two tests are inconsistent, specifically R2 is high and F-test results in statistical overall significance; whereas, at least, one null hypothesis of some t-tests cannot be rejected, it is reasonable to suspect the multicollinearity problem.

2. Correlation of regressors is greater than 0.8: the higher the correlation, the higher the variance of estimators.

3. Variance inflation factor (VIF) is greater than 10: when the regressors face the multicollinearity problem, the value of VIF might be so high that the resulting high variance of estimators adversely affects the regression analysis.

    - The VIF (variance inflation factor) to detect high multicolinearity:

4. Scatter plot of two regressors is relatively linear: when we plot the value of on regressor against another and we find that both of them tend to change in the same way, this fact might suggest the existence of multicollinearity.

.

4 Remedial Measures

1. Do nothing

2. Apply prior relationship among explanatory variables -

3. Discard some explanatory variables - the removal of the variables could mitigate the problem; but, another problem, namely specification bias problem, might occur instead. For example, suppose we want to construct the model where the production is the explained variables; and labor and capital are the explanatory ones. If there is linear relationship between labor and capital, the elimination of one variable might assuage the multicollinearity problem, but might be contrary to economic reasoning. Hence, the decision of which variables will be disposed of should be based on economic theory.

4. Collect more observations - this practice will increase _ _ _ _ _ _ _ _ _ _ _, which is the component of the variances. As a result, the variances will be lower despite high correlation among explanatory variables.

5. Transform the variables - although there is linear relationship among explanatory variables, it is not necessary that the first difference or ratio transformation of the variables will have that relationship