Chapter 8: Heteroskedasticity

1. Heteroskedasticity is present when the variance of error term conditional on regressor is not constant:

\[ E(u^2|X) = h(X) \neq \text{constant} \quad \text{(heteroskedasticity)} \]  

2. Heteroskedasticity implies that the variance of \( Y \) depends on \( X \) as well

\[ \text{var}(Y|X) \neq \text{constant} \]  

For instance, heteroskedasticity is present if the spread of consumption \( Y \) changes as income \( X \) changes.

3. The OLS estimator is consistent as long as the regressor is exogenous. So heteroskedasticity does NOT cause the OLS estimator to be inconsistent. Omitted variable can cause the OLS estimator to be inconsistent.

4. When heteroskedasticity is present, the conventional formula for variance of the OLS estimator is wrong. Therefore, the conventional standard error, \( t \) value and \( p \)-value are all wrong in presence of heteroskedasticity. To see this, recall that in a simple regression

\[ \hat{\beta}_1 = \beta_1 + \frac{\sum_i d_i u_i}{\sum_i d_i^2} \]  

where \( d_i \equiv x_i - \bar{x} \). It follows that

\[ \text{var}(\hat{\beta}_1)_{\text{robust}} = \text{var} \left( \frac{\sum_i d_i u_i}{\sum_i d_i^2} \right) \]  

\[ = \frac{\sum_i d_i^2 E(u_i^2|X)}{(\sum_i d_i^2)^2} \]  

\[ = \frac{\sum_i d_i^2 h_i}{(\sum_i d_i^2)^2} \]  

Equation (6) is the formula for heteroskedasticity-robust variance. The Stata command \texttt{reg y x, r} uses (6) to compute the heteroskedasticity robust standard error, \( t \) value and \( p \)-value.

5. Without the option \( r \), Stata command \texttt{reg y x} uses below formula to compute the
conventional standard error, t value and p-value.

\[
\text{var}(\hat{\beta}_1)_{\text{conventional}} = \frac{\sigma^2}{\sum_i d_i^2} \tag{7}
\]

Formula (7) is valid only if homoskedasticity holds, i.e.,

\[E(u^2|X) = \sigma^2 \text{ (homoskedasticity)} \tag{8}\]

6. Exercise: show that (6) reduces to (7) when \(h_i = \sigma^2\).

7. The common practice in modern econometrics is to report heteroskedasticity-robust standard error, t value and p-value as much as possible, since heteroskedasticity is norm.

8. There are several formal tests for heteroskedasticity. We focus on one of them, the Breusch-Pagan test, which involves two steps

(a) First regress \(y\) onto \(x_1, \ldots, x_k\) and save the residual \(\hat{u}\).
(b) Then regress squared residual \(\hat{u}^2\) on \(x_1, \ldots, x_k\) and compute

\[LM = nR^2 \tag{9}\]

This is the LM (likelihood multiplier) form of Breusch-Pagan test. Under the null hypothesis of homoskedasticity, the LM statistic follows chi-squared distribution with degree of freedom of the number of regressors.

\[LM \sim \chi^2_k \text{ (under } H_0 \text{ homoskedasticity)} \tag{10}\]

Homoskedasticity is rejected when the p-value is less than 0.05.

(c) Exercise: Find \(LM\) if regressing residual (not squared) \(\hat{u}\) on \(x_1, \ldots, x_k\)

9. OLS is no longer the best (most efficient) estimator when heteroskedasticity is present. The best estimator is weighted least squares (WLS). WLS is better than OLS since its variance is smaller, and confidence interval is narrower.

(a) The basic idea of WLS is to run regression using transformed variables. Variables are transformed so that the new error term is homoskedastic (having constant conditional variance).
(b) Let the original regression with heteroskedastic error term be

\[ Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k + u \quad E(u^2|X_1, \ldots, X_k) = h \]  

(11)

(c) Consider the new (transformed) error term defined as

\[ \tilde{u} = \frac{u}{\sqrt{h}} \]  

(12)

We can show the new error term has constant conditional variance

\[ E(\tilde{u}^2|X_1, \ldots, X_k) = 1 \]  

(13)

(d) WLS estimator is the OLS estimator applied to the transformed regression of

\[ \frac{Y}{\sqrt{h}} = \frac{\beta_0}{\sqrt{h}} + \frac{\beta_1 X_1}{\sqrt{h}} + \ldots + \frac{\beta_k X_k}{\sqrt{h}} + \tilde{u}, \quad E(\tilde{u}^2|X_1, \ldots, X_k) = 1 \]  

(14)

So the transformation is to divide every variable by \( \sqrt{h} \). Equivalently, the transformation is to multiply every variable with a weight of \( \frac{1}{\sqrt{h}} \). Note stata specifies the weight as \( \frac{1}{h} \).

(e) Because the new error term \( \tilde{u} \) is homoskedastic, the conventional standard error, t value and p-value are valid in the transformed regression.

(f) In practice the unknown conditional variance \( h \) can be estimated in two steps

i. First regress \( \log(\hat{u}^2) \) onto \( x_1, \ldots, x_k \) and save the fitted value, denoted by \( \hat{g} \).

ii. Then compute the estimated conditional variance as

\[ \hat{h} = e^{\hat{g}} \]  

(15)

Finally, regression (14) can be fitted by OLS using weight \( \frac{1}{\sqrt{\hat{h}}} \). Stata specifies the weight as \( \frac{1}{h} \).

10. WLS is a special case of generalized least squares (GLS) estimator, which improves the OLS estimator by utilizing extra information related to the error term. OLS is no longer the best since it ignores the information of the error term.
Example, Chapter 8

1. We use data 311_smoke.dta. See Example 8.7 in textbook for detail.

2. This example illustrates how to use real data to estimate a demand function, which you learn in eco201. The dependent variable is cigarette consumption (cigs). The independent variables are log consumption (lincome), log price (lcigpric), education (educ), age and dummy variable restaurn (equals 1 if there is smoking restriction in restaurant).

3. First we want to know the percentage of people who do not smoke in our sample. The command `gen smoke = (cigs>0)` generates a dummy variable smoke, which equals 1 if cigs is greater than 0, and equals 0 otherwise. The command `tab smoke` shows that 61.59 percent people do not smoke.

4. Command

   ```
   reg cigs lincome lcigpric educ age agesq restaurn
   ```

   replicates the OLS result shown in equation 8.35 in textbook. It seems that income and price are insignificant as their $p$-values are 0.227 and 0.897, respectively. This finding is WRONG. Equation 8.35 assumes the error term has constant variance (homoskedasticity). Later we show the error term actually has nonconstant variance (heteroskedasticity).

5. We obtain heteroskedasticity robust standard error, $t$ value and $p$ value after using option r in the `reg` command. Now the $p$-value for log income becomes 0.14, smaller than before. By contrast, the $p$-value for log price is larger than before.

6. Breusch-Pagan test can be used to formally test the null hypothesis of homoskedasticity. To do so, we save the residual of regression (8.35), square it, and regress the squared residual on all regressor. It is shown that age, age squared and restaurn are significant in this regression. This is evidence that the squared error depends on those regressors, so this is evidence for heteroskedasticity.

\[
E(u^2|X) = \text{function of } X \neq \text{constant} \Rightarrow \text{Heteroskedasticity}
\]
The LM form of Breusch-Pagan test equals the sample size (saved in stata by e(N)) times R squared (saved by e(r2)).

\[ \text{LM} = nR^2. \]

In this case, Breusch-Pagan test is 32.25842, and \( p \)-value is .00001456, smaller than 0.05. So the null hypothesis of homoskedasticity is rejected. Heteroskedasticity is present in this example.

7. Next we try WLS estimates since

\[ \text{WLS is more efficient than OLS in presence of heteroskedasticity}. \]

We regress log squared residual on all regressors, compute the exponential of fitted value and use inverse of it as weight. Command

\[ \text{reg cigs lincome lcigpric educ age agesq restaurn [w=we]} \]

replicates equation 8.36 in the textbook. It shows that income is significant. The coefficient 1.29524 means that, holding other factors constant, when income rises by 1 percent, cigs is predicted to rise by 1.29524/100 \( \approx 0.013 \) or slightly more than one-tenth of a cigarette per day. In other words, a person’s income needs to fall by 10 percent in order to cut 1 cigarette consumption per day. This is a small effect. We conclude that income is statistically significant, but economically insignificant.

8. We see the WLS confidence interval for income coefficient is narrower than OLS. This is another way to show WLS is more efficient (have smaller variance) than OLS.

9. The smoking ban in restaurant is significant in lowering cigarette consumption. educ and age are significant too. This empirical study has strong policy implication. It shows that raising cigarette tax is unlikely to be effective in cutting smoking since price and income have very small effects (you learn something called income effect and substitution effect in eco315). Let (young) people get more education is likely to be effective.

10. Exercise: Do we need to report robust standard error for the WLS estimator?

11. Exercise: What does the coefficient of age and agesq imply?
. gen smoke = (cigs>0)
. tab smoke

<table>
<thead>
<tr>
<th>smoke</th>
<th>Freq.</th>
<th>Percent</th>
<th>Cum.</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>497</td>
<td>61.50</td>
<td>61.50</td>
</tr>
<tr>
<td>1</td>
<td>310</td>
<td>38.40</td>
<td>100.00</td>
</tr>
<tr>
<td>Total</td>
<td>807</td>
<td>100.00</td>
<td></td>
</tr>
</tbody>
</table>

. reg cigs lnincome lcgipric educ age ageSQ restaurn

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>8003.02506</td>
<td>6</td>
<td>1333.83751</td>
</tr>
<tr>
<td>Residual</td>
<td>143750.658</td>
<td>800</td>
<td>179.688322</td>
</tr>
<tr>
<td>Total</td>
<td>151753.683</td>
<td>806</td>
<td>188.280003</td>
</tr>
</tbody>
</table>

|             | Coef.   | Std. Err. | t     | P>|t|  | [95% Conf. Interval] |
|-------------|---------|-----------|-------|-----|----------------------|
| lnincome    | .8802682 | .7277832  | 1.19  | .227 | -.548322 to 2.308858 |
| lcgipric    | -.7500858 | 5.773343  | -1.31 | .190 | -12.08355 to 10.58183 |
| educ        | -.5014982 | .1670772  | -3.00 | .003 | -.8294397 to -.1735368 |
| age         | -.7769636 | .1601273  | -4.81 | .000 | .436384 to 1.085603  |
| ageSQ       | -.0090228 | .001743   | -5.18 | .000 | -.012443 to -.005613 |
| restaurn    | -.2825085 | 1.111794  | -2.54 | .011 | -5.007462 to -.6427678 |
| _cons       | 3.639841  | 24.07866  | -0.15 | .880 | 50.90466 to 43.62497 |

. reg cigs lnincome lcgipric educ age ageSQ restaurn, r

Linear regression

|             | Coef.   | ROBUST Std. Err. | t     | P>|t|  | [95% Conf. Interval] |
|-------------|---------|------------------|-------|-----|----------------------|
| lnincome    | .8802682 | .5980111         | 1.48  | .140 | -.2806628 to 2.050199 |
| lcgipric    | -.7508586 | .605402         | -.12  | .901 | -12.50795 to 11.09624 |
| educ        | -.5014982 | .1623935        | -3.09 | .002 | -.8202639 to -.1827305 |
| age         | -.7769636 | .1382841        | 5.57  | .000 | .499251 to 1.042136  |
| ageSQ       | -.0090228 | .0014622        | -6.17 | .000 | -.011893 to -.0061526 |
| restaurn    | -.2825085 | 1.008033        | -2.80 | .005 | -4.803786 to -8.463836 |
| _cons       | 3.639841  | 25.61564        | -0.14 | .887 | 53.92326 to 46.34357 |
. * equation (8.14) in textbook
.reg what2 lincome lcigpric educ age agesq restaurn

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>4395266.57</td>
<td>6</td>
<td>732544.429</td>
</tr>
<tr>
<td>Residual</td>
<td>105359907</td>
<td>800</td>
<td>1331949.883</td>
</tr>
<tr>
<td>Total</td>
<td>109955173</td>
<td>806</td>
<td>136420.81</td>
</tr>
</tbody>
</table>

| what2 | Coef. | Std. Err. | t     | P>|t|  | [95% Conf. Interval] |
|-------|-------|-----------|-------|------|---------------------|
| lincome | 24.63848 | 19.7216 | 1.25 | 0.212 | -14.07411 to 63.35107 |
| lcigpric | 60.97855 | 136.4487 | 0.43 | 0.667 | -246.1219 to 368.075 |
| educ | -2.384225 | 4.527535 | -0.53 | 0.599 | -11.27148 to 6.503025 |
| age | 19.41748 | 4.339068 | 4.48 | 0.000 | 10.09018 to 27.93478 |
| agesq | -21.47895 | 0.0472335 | -4.55 | 0.000 | -30.70508 to -12.20733 |
| restaurn | -71.18137 | 30.12789 | -2.36 | 0.018 | -130.3204 to -12.04232 |
| _cons | -636.303 | 652.4946 | -0.98 | 0.330 | -1917.107 to 644.5005 |

. dis "Breusch-Pagan test is \textbackslash e(N)*e(r2)"
Breusch-Pagan test is 32.25842

. dis "p-value for Breusch-Pagan test is \textbackslash ch2tail(6, e(N)*e(r2))"
p-value for Breusch-Pagan test is .000001436

* equation (8.36) in textbook
.reg cigs lincome lcigpric educ age agesq restaurn [w=we]
(analytic weights assumed)
(sum of wgt is 1.99775e+01)

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>10302.646</td>
<td>6</td>
<td>1717.10767</td>
</tr>
<tr>
<td>Residual</td>
<td>80542.1381</td>
<td>800</td>
<td>100.677698</td>
</tr>
<tr>
<td>Total</td>
<td>90844.8041</td>
<td>806</td>
<td>112.710675</td>
</tr>
</tbody>
</table>

| cigs | Coef. | Std. Err. | t     | P>|t|  | [95% Conf. Interval] |
|------|-------|-----------|-------|------|---------------------|
| lincome | 1.29524 | 1.470118 | 2.96 | 0.003 | 0.4574148 to 2.133065 |
| lcigpric | -0.900132 | 4.460144 | -0.20 | 0.842 | -11.689328 to 9.82398 |
| educ | -0.463446 | 1.201587 | -3.86 | 0.000 | -6.903099 to -2.275828 |
| age | 0.481979 | 0.096082 | 4.98 | 0.000 | 0.2919197 to 0.671976 |
| agesq | -0.005672 | 0.0009395 | -5.99 | 0.000 | -0.0074713 to -0.0037831 |
| restaurn | -3.461064 | 0.795505 | -4.35 | 0.000 | -5.027358 to -1.99541 |
| _cons | 5.63546 | 17.80314 | 0.32 | 0.752 | -29.31092 to 40.58184 |
Do File

* Do file for chapter 8
cd "I:\311"
log using 311log.txt, text replace
use 311_smoke.dta, clear
* generate smoking dummy
gen smoke = (cigs>0)
* tabulate proportion of smoking people
tab smoke
* OLS, equation (8.35) in textbook
reg cigs lincome lcigpric educ age agesq restaurn
* save residual and generate squared residual
predict uhat, re
gen uhat2 = uhat^2
* heteroskedasticity robust standard error
reg cigs lincome lcigpric educ age agesq restaurn, r
* equation (8.14) in textbook
reg uhat2 lincome lcigpric educ age agesq restaurn
* Breusch-Pagan test for heteroskedasticity
dis "Breusch-Pagan test is " e(N)*e(r2)
dis "p-value for Breusch-Pagan test is " chi2tail(6, e(N)*e(r2))
* quietly run regression (8.32) in textbook
gen luhat2 = log(uhat2)
qui reg luhat2 lincome lcigpric educ age agesq restaurn
* keep fitted value
predict ghat, xb
* equation (8.33) in textbook
gen hhat = exp(ghat)
* specify weight for stata
gen we = 1 /hhat
* WLS, equation (8.36) in textbook
reg cigs lincome lcigpric educ age agesq restaurn [w=we]