Lecture 16. Heteroskedasticity

In the CLR model

\[ Y_i = \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_K X_{iK} + u_i, \quad i = 1, \ldots, n \]

one of the assumptions was

Assumption 3 (Homoskedasticity)
All \( u_i \)'s have the same variance, i.e. for \( i = 1, \ldots, n \)

\[ Var(u_i) = E(u_i^2) = \sigma^2 \]
When is this a bad assumption?

If omitted variables are not correlated with the included variables (assumption 1), but have a different order of magnitude for (groups of) observations.

- 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 behavior. For instance, high income individuals have more discretion in their spending.
Example of second case: Relation between income and experience.

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

See graphs

Note

• Variation in income (in 1000$) increases with work experience
• Variation in relative income first increases and then decreases

Is consistent because income is higher if more work experience
Salary (1000$) and work experience (years since Ph.D.)
Log(Salary) and work experience
(years since Ph.D.)
Note that log transformation reduces variation in income with experience. Why?

If variation in income increases proportionally with income level, then variation in relative income does not change with income level.

Example

Income with work experience 4 years: 30, 40, 60 with absolute difference 10, 30, relative difference 33\%, 100\% and log difference 0.29, 0.69 (all relative to lowest)

Income at work experience 8 years: 90, 120, 180 with absolute difference 30, 90, relative difference 33\%, 100\% and log difference 0.29, 0.69 (all relative to lowest)

Often after log transformation variation is constant (but not in example).
Is there heteroskedasticity if we estimate the model

\[ Y = \beta_1 + \beta_2 X + \beta_3 X^2 + u \]

with

\[ Y = \text{log salary} \]

\[ X = \text{work experience} \]

See output and graph

- Interpretation of regression coefficients
- Nature of relation: Maximum at 35 years work experience
- Heteroskedasticity: Plot squared OLS residuals against \( X \)

All examples for cross-sections, but heteroskedasticity also important in time-series data, e.g. volatility in the stock market. This is like case 2 but omitted variable is news/information.
<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>Std. Error</th>
<th>t-Statistic</th>
<th>Prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>3.809365</td>
<td>0.041338</td>
<td>92.15104</td>
<td>0.0000</td>
</tr>
<tr>
<td>YEARS</td>
<td>0.043853</td>
<td>0.004829</td>
<td>9.081645</td>
<td>0.0000</td>
</tr>
<tr>
<td>YEARS2</td>
<td>-0.000627</td>
<td>0.000121</td>
<td>-5.190657</td>
<td>0.0000</td>
</tr>
</tbody>
</table>

- R-squared: 0.536179
- Adjusted R-squared: 0.531943
- S.E. of regression: 0.206962
- Sum squared resid: 9.380504
- Log likelihood: 36.20452
- Durbin-Watson stat: 1.434005

- Mean dependent var: 4.325410
- S.D. dependent var: 0.302511
- Akaike info criterion: -0.299140
- Schwarz criterion: -0.253158
- F-statistic: 126.5823
- Prob(F-statistic): 0.000000
Squared OLS residuals and work experience
Effect of heteroskedasticity on OLS estimators and tests

• OLS estimators are unbiased (only assumptions 1 and 2 are needed)
• The usual formula for the sampling variance is wrong (assumption 3 was used in derivation)
• The OLS estimators not Best Linear Unbiased (BLU), i.e. better estimators may exist
• The t- and F-tests cannot be used

Often standard errors reported by regression program are too small, e.g. estimates of regressions coefficients seem more significant than they really are. This is case in simple regression model and if error variance increases with $X$. 
How do we detect heteroskedasticity?

- Plot of squared OLS residuals against regressors
- Tests

For test we must specify a model for the heteroskedasticity

\[ \text{Var}(u_i) = \sigma_i^2 \]

We cannot estimate these variances as parameters. Why not?

Models

\[ \sigma_i^2 = \alpha_1 + \alpha_2 Z_{i2} + \cdots + \alpha_L Z_{iL} \] (Breusch-Pagan)

\[ \sigma_i = \alpha_1 + \alpha_2 Z_{i2} + \cdots + \alpha_L Z_{iL} \] (Glesjer)

\[ \ln \sigma_i^2 = \alpha_1 + \alpha_2 Z_{i2} + \cdots + \alpha_L Z_{iL} \] (Harvey-Godfrey)

The \( Z \)'s may be regressors or squares or products of regressors
Choice of $Z$’s

- If heteroskedasticity because of size differences, choose size, e.g. population
- If no clear choice, choose $X_2, \ldots X_K, X_2^2, \ldots, X_K^2, X_1X_2, \ldots$
  i.e. regressors, their squares and cross-products. BP test with this choice is White test
Test

1. Estimate by OLS and obtain OLS residuals $e_i, i = 1, \ldots, n$

2. Estimate linear regression of $e_i^2 (BP)$, $|e_i| (G)$, or $\ln e_i^2 (HG)$ on constant and $Z_{i2}, \ldots, Z_{iL}$ and compute the $R^2$ of this regression,

3. Compute the test statistic $LM = n R^2$ for the hypothesis $H_0 : \alpha_2 = 0, \ldots, \alpha_L = 0$. If $H_0$ is true (homoskedastic errors) then $LM$ has a $\chi^2$ distribution with $L - 1$ degrees of freedom. Use this to obtain critical value.

This is test is called the Lagrange Multiplier (LM) test for heteroskedasticity of a particular form.
Example: BP test with $Z_2 = X_2$ is years and $Z_3 = X_3$ is years squared.

\[ R^2 = 0.0747 \quad LM = nR^2 = 222 \times 0.0747 = 16.59 \]

Critical value for 5% and chi-squared distribution with 2 df is 5.99 (see book)

White test: Add $Z_4 = X_2 X_3$ is years cubed

\[ R^2 = 0.0810 \quad LM = nR^2 = 222 \times 0.0810 = 17.98 \]

Critical value for 5% and chi-squared distribution with 3 df is 7.81
Dependent Variable: RESID2  
Method: Least Squares  
Date: 11/07/01   Time: 14:13  
Sample: 1 222  
Included observations: 222

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>Std. Error</th>
<th>t-Statistic</th>
<th>Prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>-0.011086</td>
<td>0.013473</td>
<td>-0.822858</td>
<td>0.4115</td>
</tr>
<tr>
<td>YEARS</td>
<td>0.006084</td>
<td>0.001574</td>
<td>3.865764</td>
<td>0.0001</td>
</tr>
<tr>
<td>YEARS2</td>
<td>-0.000129</td>
<td>3.94E-05</td>
<td>-3.270700</td>
<td>0.0012</td>
</tr>
</tbody>
</table>

R-squared 0.074714  
Adjusted R-squared 0.066264  
S.E. of regression 0.067451  
Sum squared resid 0.996378  
Log likelihood 285.0957  
Durbin-Watson stat 1.707896

Mean dependent var 0.042255  
S.D. dependent var 0.069804  
Akaike info criterion -2.541402  
Schwarz criterion -2.495420  
F-statistic 8.841813  
Prob(F-statistic) 0.000203
# Regression Results

**Dependent Variable:** RESID2  
**Method:** Least Squares  
**Date:** 11/07/01  
**Time:** 14:25  
**Sample:** 1 222  
**Included observations:** 222

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>Std. Error</th>
<th>t-Statistic</th>
<th>Prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>-0.027664</td>
<td>0.019079</td>
<td>-1.449980</td>
<td>0.1485</td>
</tr>
<tr>
<td>YEARS</td>
<td>0.010395</td>
<td>0.003853</td>
<td>2.698166</td>
<td>0.0075</td>
</tr>
<tr>
<td>YEARS2</td>
<td>-0.000380</td>
<td>0.000209</td>
<td>-1.821374</td>
<td>0.0699</td>
</tr>
<tr>
<td>YEARS3</td>
<td>3.99E-06</td>
<td>3.25E-06</td>
<td>1.225781</td>
<td>0.2216</td>
</tr>
</tbody>
</table>

**R-squared:** 0.081048  
**Mean dependent var:** 0.042255  
**Adjusted R-squared:** 0.068402  
**S.D. dependent var:** 0.069804  
**Akaike info criterion:** -2.539262  
**Schwarz criterion:** -2.477953  
**F-statistic:** 6.408914  
**Prob(F-statistic):** 0.000352

---

**Included observations:** 222

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>Std. Error</th>
<th>t-Statistic</th>
<th>Prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>-0.027664</td>
<td>0.019079</td>
<td>-1.449980</td>
<td>0.1485</td>
</tr>
<tr>
<td>YEARS</td>
<td>0.010395</td>
<td>0.003853</td>
<td>2.698166</td>
<td>0.0075</td>
</tr>
<tr>
<td>YEARS2</td>
<td>-0.000380</td>
<td>0.000209</td>
<td>-1.821374</td>
<td>0.0699</td>
</tr>
<tr>
<td>YEARS3</td>
<td>3.99E-06</td>
<td>3.25E-06</td>
<td>1.225781</td>
<td>0.2216</td>
</tr>
</tbody>
</table>

**R-squared:** 0.081048  
**Mean dependent var:** 0.042255  
**Adjusted R-squared:** 0.068402  
**S.D. dependent var:** 0.069804  
**Akaike info criterion:** -2.539262  
**Schwarz criterion:** -2.477953  
**F-statistic:** 6.408914  
**Prob(F-statistic):** 0.000352
Estimation with heteroskedasticity

• Use OLS but get correct standard errors
• Find a better estimation procedure

It is possible to derive the correct standard error of OLS estimator. Formula does not depend on the model for heteroskedasticity.

These standard errors are called heteroskedasticity-consistent standard errors. Many regression programs have this option.

Example: See output.

Note differences are small (wrong standard errors are here too large).
### Regression Output

**Dependent Variable:** LNSALARY  
**Method:** Least Squares  
**Date:** 11/07/01  
**Time:** 13:49  
**Sample:** 1 222  
**Included observations:** 222  

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>Std. Error</th>
<th>t-Statistic</th>
<th>Prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>3.809365</td>
<td>0.026119</td>
<td>145.8466</td>
<td>0.0000</td>
</tr>
<tr>
<td>YEARS</td>
<td>0.043853</td>
<td>0.004361</td>
<td>10.0599</td>
<td>0.0000</td>
</tr>
<tr>
<td>YEARS2</td>
<td>-0.000627</td>
<td>0.000118</td>
<td>-5.322369</td>
<td>0.0000</td>
</tr>
</tbody>
</table>

- **R-squared:** 0.536179  
- **Mean dependent var:** 4.325410  
- **Adjusted R-squared:** 0.531943  
- **S.D. dependent var:** 0.302511  
- **S.E. of regression:** 0.206962  
- **Akaike info criterion:** -0.299140  
- **Schwarz criterion:** -0.253158  
- **F-statistic:** 126.5823  
- **Prob(F-statistic):** 0.000000  
- **Durbin-Watson stat:** 1.434005