# Chapter 8 Heteroskedasticity

RecallMLR 5 Homskedasticity –error u has the same variance given any values of the explanatory variables

$Var(u|x1, . . .,xk) = \sigma^2$
or $E(UU') = \sigma^2 I$

Suppose other GM assumptions hold but have heteroskedasticity.

$Var(u_i|x_i) = \sigma_i^2$

Why might economic data have heteroskedasticity?

1. Skewness of one of the x's
2. Outliers
3. Binary y

4. Error learning models—suppose have data on hours of typing practice and typing errors for a bunch of different people. Higher variance at lower hours, but as people learn, variance in errors falls.
5. X is income, Y is something like savings—as income grows, have more scope for choice about what to do with income

6. Specification errors—functional form
7. Clustered data—data on individuals, some variables are state level averages

What problems does this violation cause?

Do this in a two variable case:

$$\hat{\beta}_1 = \frac{\sum_i (x_{i1} - \bar{x}_1) y_i}{\sum_i (x_{i1} - \bar{x}_1)^2}$$

OLS Estimator

$$= \frac{\sum_i (x_{i1} - \bar{x}_1)(\beta_1 (X_{i1} - \bar{x}_1) + u_i)}{\sum_i (x_{i1} - \bar{x}_1)^2}$$

$$= \beta_1 + \frac{\sum_i (x_{i1} - \bar{x}_1) u_i}{\sum_i (x_{i1} - \bar{x}_1)^2}$$

- If we take the expected value of this, as long as we have exogeneity, **OLS estimator is unbaised**
- **R$^2$ is fine** even with heteroskedasticity

However, if have heteroskedasticity, usual estimator for **var($\hat{\beta}$) is biased**.

Recall **var($\hat{\beta}$)** Derivation

$$\hat{\beta} = \beta_1 + \frac{\sum_i (x_{i1} - \bar{x}_1)u_i}{\sum_i (x_{i1} - \bar{x}_1)^2}$$

Take the variance of this:

var($\hat{\beta}_2$) =

$$0 + \frac{1}{SST_x^2} var(\sum(x_i - \bar{x})u_i) = \frac{1}{SST_x^2}(\sum(x_i - \bar{x})^2 var(u_i)$$

$$= \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sigma_i^2}{SST_x^2} \text{ where } SST_x = \sum_{i=1}^{n}(x_i - \bar{x})^2$$

Or $\boxed{\text{Var}(\hat{\beta}_2) = \frac{\sum x_i^2 \sigma_i^2}{\left(\sum x_i^2\right)^2}}$

Under homoskedasticity, we get

Var($\hat{\beta}_2$) = $\frac{\sigma^2}{\left(\sum x_i^2\right)}$ = (or $\sigma^2$(X'X)$^{-1}$ in multivariate case)

- Recall that $\sigma^2 = 1/n \sum_{i=1}^{n} u_i$
- However, we don't know the true $u_i$
- Instead we have $\hat{u}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$
- Recall that simply replacing $u_i$ with $\hat{u}_i$ leads to a biased estimator of $\sigma^2$ (see Mark's notes)
- Instead we use $\hat{\sigma}^2 = \frac{1}{n-k-1\sum_{i=1}^{n}\hat{u}_i}$

So the usual estimator for the variance of the estimated coefficients is

$$\boxed{\widehat{var(\hat{\beta}_2)} = \frac{\hat{\sigma}^2}{\left(\sum x_i^2\right)}}$$

In multivariate case $\widehat{var(\hat{\beta}_J)} = \hat{\sigma}^2$(X'X)$^{-1}$

But under heteroskedasticity, this is biased.  Still the case that

$$\text{Var}(\hat{\beta}_2) = \frac{\sum x_i^2 \sigma_i^2}{\left(\sum x_i^2\right)^2}$$

But assuming equal variance $\sigma_i^2 = \sigma^2$ will result in biased standard errors

Why do we care if the estimator for this is biased?

- If the estimator for this is biased inferences will be wrong—LM, F, t stats will be wrong.

## 8.3 Testing for Heteroskedasticity:

$y = \beta_o + \beta_1 x_1 + \ldots. + \beta_k x_k + u$

**A. BREUSCH-PAGAN TEST** (this is the Koenker version—there are several BP tests, but Wooldridge calls this version the BP test)

What is Heteroskedasticity?
Ho = $\mathrm{var}(u|x_1, x_2, \ldots .x_k) = \sigma^2$
Alternative—not identical

Matrix form $E(UU') = \sigma^2 I$—draw this out

So if Ho is true and u has zero conditional mean $\mathrm{var}(u|X) = E(u^2|X)$
so Ho = $E(u^2| x_1, x_2, \ldots .x_k) = E(u^2)= \sigma^2$

**Use an LM test procedure for this**

Steps:

1. Run restricted regression (regular OLS model with the restriction of homoskedasticity)→ $\tilde{u}$ Number of restrictions is k
2. Run auxillary regression $\tilde{u}^2 = \delta 0 + \delta_1 x_1 + \ldots . + \delta_k x_k + error$
3. LM = n* $R^2$ from above regression
4. Compare to $\chi^2_k$

**B. WHITE TEST**—test with weaker assumptions that BP test:

Instead of Ho = $\mathrm{var}(u| x_1, x_2, \ldots .x_k) = \sigma^2$
Tests that $\mathrm{Corr}(u^2, (x_i, x_i^2, x_i x_j) = 0$—turns out this tests for all forms of heteroskedasticity that could invalidate OLS standard errors

**Again, another LM test procedure**

Steps:

1. Run restricted regression (regular OLS model)→ $\tilde{u}$ Number of restrictions is k

2. Run auxillary regression

   $\tilde{u}^2 = \delta 0 + \delta_1 x_1 + .... + \delta_k x_k + \delta_{k+1} x_1^2 + ... + \delta_{k+k} x_k^2 + \delta_{2k+1} x_1 x_2 + ... + error$

3. LM = n* $R^2$ from above regression

Chews up lots of degrees of freedom. . .

## C. One more LM test procedure—Special case of White test:

1. Construct fitted values of y

   $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + ... + \hat{\beta}_k x_{ik}$

   If square these fitted values, get a particular function of all squares and cross-products of x's

2. Estimate

   $\hat{u}^2 = \delta_0 + \delta_1 \hat{y}_1 + \delta_2 \hat{y}^2 + + error$

3. LM = n* $R^2$ from above regression $\sim\chi^2_2$

Solutions to Heteroskedasticity

1. Generate Robust standard errors t, F, LM statistics that are valid in presence of heteroskedasticity of unknown form.

2. Use Weighted Least Squares (more generally, GLS) with an aribitrary covariance matrix—less commonly used in general (FGLS)

3. Use Weighted Least Squares (more generally, GLS) with a specific covariance matrix—less commonly used

Solution #1. Heteroskedastic Robust Inference after OLS estimation (Section 8.2)

Recall with only one X:

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

Other GM assumptions hold but have heteroskedasticity. $\mathrm{Var}(u_i|x_i) = \sigma_i^2$

$$\mathrm{var}(\hat{\beta}_1) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2\sigma_i^2}{SST_x^2} \text{ where } SST_x = \sum_{i=1}^{n}(x_i - \bar{x})^2$$

Note that if $\sigma_i^2 = \sigma^2$, then we have usual form $\frac{\sigma^2}{SST_x}$

Use an estimator for $\sigma^2$

$$\hat{\sigma}^2 = \frac{1}{n-k-1\sum_{i=1}^{n}\hat{u}_i}$$

What if have heteroskedasticity? Then need a consistent estimate of var($\hat{\beta}_1$)

White shows can use the following:

1. Regress y on x → Predicted errors $\hat{u}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1}$
2. var($\hat{\beta}$) $\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2 \hat{u}_i^2}{SST_x^2}$ (put hat over var)
3. This estimator is consistent—see notes in Wooldridge on this proof.

   Converges in probability to $\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sigma_i^2}{SST_x^2}$

Extension to multiple regression framework:

Huber-White or White or Huber or Robust standard errors:

var( $\hat{\beta}_j$ ) = $\frac{\sum_{i=1}^{n}\hat{r}_{ij}^{2}\hat{u}_i^2}{SSR_j^2}$ where $r_{ij}$ is ith residual from regressing $x_j$ on all other x's
(hat over var)

and SSR is sum of squared residuals from xj on all other x's

```
STATA:   reg yvar xvar1 xvar2, robust
```

These standard errors and the associated t-stats are only valid as sample size gets large. IF homoskedasticity holds AND sample size is small, may not want to add robust option. But most of time will estimate everything with robust standard errors.

What about testing restrictions?

- WALD statistic—Recall that we talked about F tests. There is a heteroskedasticity robust version of F that is a Wald statistic.
- LM tests—see book

**Solution #2 Weighted Least Squares**

If know the form of the variance, can use a different estimator: WLS.
WLS is more efficient than OLS IF know the form of the variance.
If don't know the form of the variance, is not necessarily more efficient.
But if have STRONG heteroskedasticity, WLS can be more efficient.

In practice, mostly use OLS with robust standard errors. But good to see how WLS works—a special case of GLS.

**Suppose have heteroskedasticity is known up to a multiplicative constant**

GENERALIZED LEAST SQUARES—GLS

- Here also called WEIGHTED least squares—
- essentially, minimize sum of squared errors, where weight each observation by 1/hi—
- give more weight to obs with lower variance.
- Figuring out the weights is somewhat arbitrary unless for some reason KNOW the form of heterosked.

One case where do know form of Heterosked—when have **averages of individual level data**—instead of estimating individuals, have firm or state or country averages.

Suppose that individual's errors are uncorrelated. Variance decreases with size of group.

In this case, **hi = 1/mi** where mi is number of members in the group (firm or state or country). Homework has you work through this in more detail.