Week 10: Heteroskedasticity

Marcelo Coca Perraillon

University of Colorado
Anschutz Medical Campus

Health Services Research Methods I
HSMP 7607
2019

Updated notes are here: https://clas.ucdenver.edu/marcelo-perraillon/teaching/health-services-research-methods-i-hsmp-7607

Outline

- The problem of (conditional) unequal variance: heteroskedasticity
- Correcting and testing for heteroskedasticity
- The sandwich estimator
- Examples

Big picture

- Heteroskedasticity is so **common that we should just assume it exists**
- We can perform some tests to detected it
- The solutions depend on the **source** of heteroskedasticity
- The problem is not about the bias or consistency of the OLS estimates; the issue is that **SEs are not correct** in the presence of heteroskedasticity
- We will follow Chapter 8 of Wooldridge

Graphically

- We can see the problem graphically checking the residuals. As I said, heteroskedasticiy is **everywhere** in the linear model

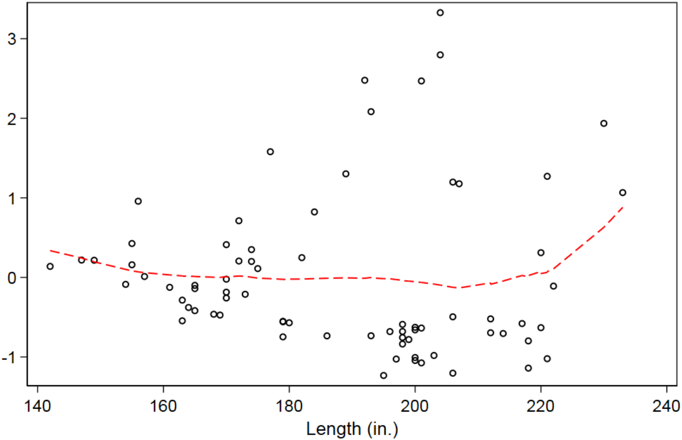
```
sysuse auto

* Do everything quietly

quietly {
   reg price length
  predict resi, rstandard
}

* See different options
scatter resi length, msymbol(Oh) msize(small) legend(off) || ///
   lowess resi length, color(red) xline() saving(g1.gph, replace)
graph export g1.png, replace
```

Graphically



- That funnel shape is quite common. Why? Well, in part it's because of fewer observations at the tails

Digression: talking Stata

- Saving graphs in a file. Stata 15 has a bunch of new commands to create “reports” using PDF, Word, or HTML (putpdf, putdocx, and dyndoc)
- Syntax not too pretty but at least you can save all graphs in one document

```
* Example

* At the start of do file:
putpdf begin

* Create graphs
scatter y x, saving(g1.gph)
graph export g1.gph

* "Put" it on PDF file
putpdf paragraph, halign(center)
putpdf image g1.png

scatter y1 x1, saving(g2.gph)
graph export g2.gph

putpdf paragraph, halign(center)
putpdf image g2.png

* Write PDF file
putpdf save filewithgraphs.pdf, replace
```

Homoskedasticity

- In the linear model $y_i = \beta_0 + \beta_1 x1_i + \cdots + xp_p + \epsilon_i$ we assumed that $\epsilon_i \sim N(0, \sigma^2)$
- That is, the error terms have all the same variance conditional on all explanatory variables: $var(\epsilon_i | x1, ..., xp) = \sigma^2$
- To simplify, we will focus on the simple linear model (only one covariate). In the presence of heteroskedasticity: $var(\epsilon_i | x_i) = \sigma_i^2$

Homoskedasticity

- In the SLR model, we can write the variance of $\hat{\beta}_1$ as
  $var(\hat{\beta}_1) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2\sigma_i^2}{\sum_{i=1}^{n}(x_i - \bar{x})^4}$
- If we have homoskedasticity the formula reduces to the one we saw in Chapter 2 (2.22):
  $var(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$
- But in the presence of heteroskedasticity we can't no longer simplify that formula
- Remember that the variances are also estimated when we estimate the coefficients

A simple solution

- The problem now is that we know that the variance depends on the value of the covariate X
- One solution is rather simple: we just **estimate the variance conditional on the values of** X
- White (1980) introduced an estimator for the variance in the presence of **unknown heteroskedasticity**. The idea is to estimate σ_i^2:

$var(\hat{\beta}_1) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2 \hat{\epsilon_i}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^4}$

Huber-White robust standard errors

- In the previous equation, $\hat{\epsilon_i}^2$ is the **estimated residual** of the regression
- The estimation proceeds in two steps: **1)** Estimate the original regression of Y on X and **2)** Obtain the residuals to estimate the **robust** variance
- (Remember what I told you about the term "robust" in statistics. Always ask yourself robust to what? In this case, robust to heteroskedasticity problems)
- In matrix notation, the variance-covariance matrix is $var(\hat{\beta}) = \sigma^2(X'X)^{-1}$
- The **Huber-White robust variance-covariance** matrix is
- $var(\hat{\beta}_{rob}) = (X'X)^{-1}X'\hat{\Sigma}(X'X)^{-1}$

Huber-White robust standard errors

- $var(\hat{\beta}_{rob}) = (X'X)^{-1}X'\hat{\Sigma}(X'X)^{-1}$
- $\hat{\Sigma}$ is the variance-covariance matrix from the original model
- The way the formula looks is the reason why Huber-White robust standard errors are (affectionately?) referred to as the **sandwhich estimator**
- The intuition is that we will correct for the heteroskedasticity problem in (sort of) the same way we diagnose the problem: we will, empirically, estimate a variance using the residuals
- Importantly, we do not need to know the **source** of heteroskedasticity

Example

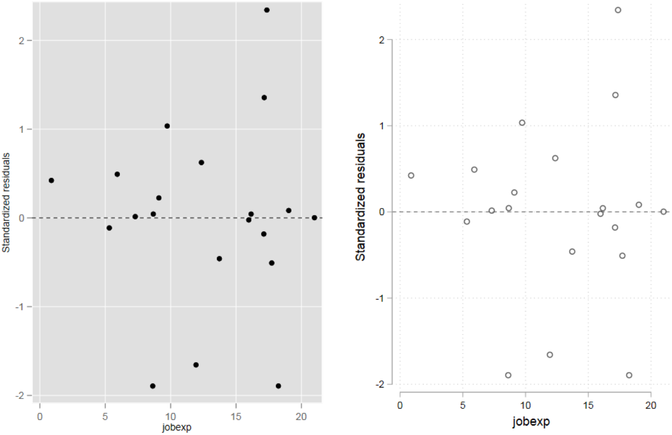
- Another example just to show you some graphs...

```
* Load data
use https://www3.nd.edu/~rwilliam/statafiles/reg01.dta, clear
reg  income jobexp
   Source |       SS           df       MS      Number of obs   =        20
-------------+----------------------------------   F(1, 18)        =      1.39
       Model |  130.495675         1  130.495675   Prob > F        =    0.2538
    Residual |   1689.9298        18  93.8849889   R-squared       =    0.0717
-------------+----------------------------------   Adj R-squared   =    0.0201
       Total |  1820.42548        19  95.8118671   Root MSE        =    9.6894

------------------------------------------------------------------------------
      income |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      jobexp |   .4799311   .4070792     1.18   0.254    -.3753106    1.335173
       _cons |   18.34387   5.586783     3.28   0.004     6.606476    30.08127
------------------------------------------------------------------------------
predict res if e(sample), rstandard

search plotplain
set scheme plottig
scatter res jobexp, jitter(2) yline(0) saving(res1.gph, replace)
set scheme plotplain
scatter res jobexp, jitter(2) yline(0) saving(res2.gph, replace)
graph combine res1.gph res2.gph
graph export het1.png, replace
```

Example



- The residual at each point of job experience is different

Example II

- Using the mksp1 dataset we saw that it's likely there is a hetoskedasticity problem

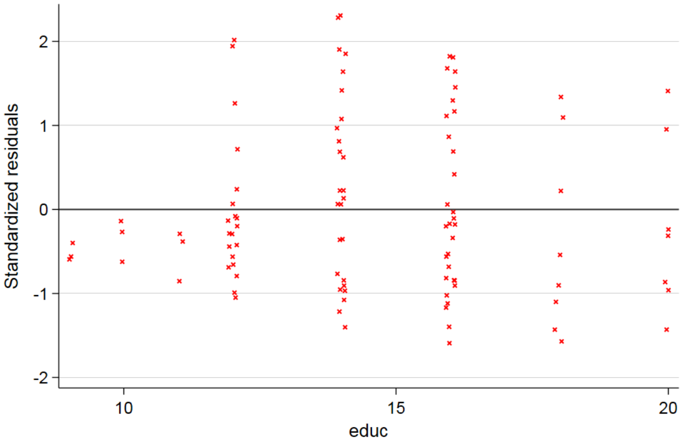
```
* Load data
webuse mksp1

* Regress educ on income
reg income educ
    Source |       SS           df       MS      Number of obs   =       100
-------------+----------------------------------   F(1, 98)        =     10.34
       Model |  2.7896e+09         1  2.7896e+09   Prob > F        =    0.0018
    Residual |  2.6433e+10        98   269719984   R-squared       =    0.0955
-------------+----------------------------------   Adj R-squared   =    0.0862
       Total |  2.9222e+10        99   295173333   Root MSE        =     16423

------------------------------------------------------------------------------
      income |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   2001.493   622.3571     3.22   0.002     766.4461    3236.541
       _cons |   14098.23   9221.392     1.53   0.130    -4201.327    32397.78
------------------------------------------------------------------------------
predict incres, rstandard

set scheme lean2
scatter incres educ, yline(0) jitter(2) msymbol(x) mcolor(red)
```

Example



- Some evidence of unequal variances conditional on education (but nothing terrible)

Huber-White robust SEs in Stata

- The option vce(robust) or simply robust uses the sandwich estimator

```
reg income educ, vce(robust)
* same as reg income educ, robust

Linear regression                               Number of obs     =        100
                                                F(1, 98)          =      13.84
                                                Prob > F          =     0.0003
                                                R-squared         =     0.0955
                                                Root MSE          =      16423

------------------------------------------------------------------------------
             |               Robust
      income |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   2001.493   538.0771     3.72   0.000     933.6971     3069.29
       _cons |   14098.23   7680.933     1.84   0.069    -1144.337    29340.79
------------------------------------------------------------------------------
```

- Conclusions won't change but notice that CIs are narrower. SEs went down
- **Chose this example on purpose**. You always hear that SEs go up, not down, but not always the case (!)

Huber-White robust SEs in Stata

- Compare models; some tests will of course change now that we have different SEs

```
qui reg income educ
est sto m1

qui reg income educ, robust
est sto m2

est table m1 m2, se stats(N F)
----------------------------------------
    Variable |     m1           m2
-------------+--------------------------
        educ |  2001.4935    2001.4935
             |  622.35711    538.07705
       _cons |  14098.225    14098.225
             |   9221.392    7680.9332
-------------+--------------------------
           N |        100          100
           F |  10.342584    13.836282
----------------------------------------
```

- Note that Stata calculates a different F statistics

Huber-White robust SEs in Stata

- Compare models; some tests will of course change now that we have different SEs

```
qui reg income educ
test educ= 900
 ( 1)  educ = 900
       F(  1,    98) =    3.13
            Prob > F =    0.0799

qui reg income educ, robust
test educ= 900
 ( 1)  educ = 900
       F(  1,    98) =    4.19
            Prob > F =    0.0433
```

- Since SEs haven changed, tests can change

The good and the bad of the sandwich

- **Good**: We do **not need to know the source** of unequal variance
- **Great**: The sandwhich estimator is **asymptotically unbiased**
- **Fantastic**: The sandwhich estimator is asymptotically unbiased **even in the presence of homoskedasticity**
- If we often suspect heteroskedasticity and the sandwich estimator is asymptotically valid *even in the presence of homoskedasticity*, **why not always use the robust SEs?**
- Well... many researchers add the option robust to every single model for "insurance"
- **The bad**: The only drawback is that if the homoskedasticity assumption is valid, in **smaller samples the robust SEs may be biased**
- But... We seldom work with "small" samples anymore so you could just add the robust option by default

Testing for heteroskedasticity

- If **small samples and unequal variance in doubt**, useful to have a test for heteroskedasticity rather than just assume it
- The null hypothesis is $H_0 : var(\epsilon | x_1, x_2, ..., x_p) = \sigma^2$ (that is, homoskedasticity)
- As usual with hypothesis testing, we will look at the data to provide evidence that the variance is not equal conditional on $x_1, x_2, ..., x_p$
- Recall the basic formula of the variance: $var(X) = E[(X - \bar{X})^2] = E[X^2] - (E[X])^2$
- Since $E[\epsilon] = 0$ we can rewrite the null as: $H_0 : E(\epsilon^2 | x_1, x_2, ..., x_p) = E[\epsilon^2] = \sigma^2$ (think of σ^2 here as a constant)
- If you see the problem this way, it looks a lot easier. We need to figure out if the $E[\epsilon^2]$ is **related** to one or more of the explanatory variables (we will use $E[\hat{\epsilon}^2]$). If not, we can assume homoskedasticity

Testing for heteroskedasticity

- By related, it could be in **any functional form**, but start with a linear relationship
- The model becomes:
- $\epsilon^2 = \gamma_0 + \gamma_1 x_1 + \cdots + \gamma_p x_p + u$
- If we reject $H_0 : \gamma_0 = \gamma_1 = ... = \gamma_p = 0$ then there is **evidence of unequal variance**
- Of course, we do not observe ϵ^2 so we need to work with $\hat{\epsilon}^2$
- The test is an F-test of the **overall significance of the model**
- As you probably suspect, Stata has a command for that

Testing for heteroskedasticity, example

- Let's go back to the income, education, and age dataset and estimate the model

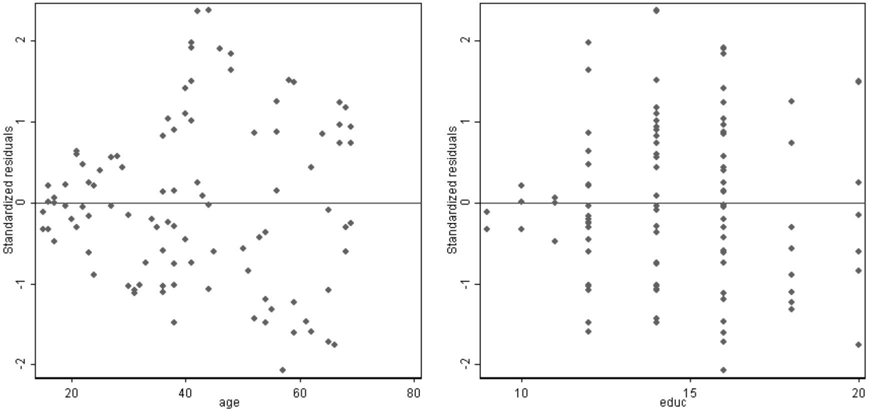
  $income = \beta_0 + \beta_1 educ + \beta_2 age + \epsilon$

```
* Get residuals
qui reg income age edu
predict incress, rstandard

* Combine the plots
scatter incress age, yline(0) legend(off) saving(r1.gph, replace)
scatter incress educ, yline(0) legend(off) saving(r2.gph, replace)

* Export plot
graph combine r1.gph r2.gph, row(1) ysize(10) xsize(20)
graph export rall.png, replace
```

Testing for heteroskedasticity, example



- Clearly, we suspect unequal variance conditional on both age and education

Testing for heteroskedasticity, example

- We use the post-estimation command hettest and confirm that we do reject the null:

```
 reg income age edu

      Source |       SS           df       MS      Number of obs   =       100
-------------+----------------------------------   F(2, 97)        =     14.71
       Model |  6.8005e+09         2  3.4002e+09   Prob > F        =    0.0000
    Residual |  2.2422e+10        97   231151328   R-squared       =    0.2327
-------------+----------------------------------   Adj R-squared   =    0.2169
       Total |  2.9222e+10        99   295173333   Root MSE        =     15204

------------------------------------------------------------------------------
      income |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   440.2441   105.6871     4.17   0.000     230.4845    650.0037
        educ |   706.8841   654.6241     1.08   0.283    -592.3636    2006.132
       _cons |   14800.35   8538.327     1.73   0.086     -2145.86    31746.57
------------------------------------------------------------------------------

estat hettest, rhs
Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: age educ

         chi2(2)      =     9.86
         Prob > chi2  =   0.0072
```

By hand

- Not exactly the same as the Breusch-Pagan but relatively close (p-value of F test: 0.0012)
- The BP regress agains all regressors, squares, and cross-products (interactions)

```
qui reg income age edu
* Get square of residuals
predict r1, rstandard
gen r12 = r1^2
*  Regress
   Source |       SS           df       MS      Number of obs   =       100
-------------+----------------------------------   F(2, 97)        =      7.51
       Model |  20.0399037         2  10.0199519   Prob > F        =    0.0009
    Residual |  129.416801        97  1.33419383   R-squared       =    0.1341
-------------+----------------------------------   Adj R-squared   =    0.1162
       Total |  149.456705        99  1.50966369   Root MSE        =    1.1551

------------------------------------------------------------------------------
         r12 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0270132   .0080294     3.36   0.001     .0110771    .0429494
        educ |   .0047798    .049734     0.10   0.924    -.0939284     .103488
       _cons |   -.188754   .6486853    -0.29   0.772    -1.476215    1.098707
------------------------------------------------------------------------------
```

- As suspected, the problem is age and not so much education

Using Breusch-Pagan

- We can also test for age or education separately

```
qui reg income age edu

estat hettest age

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: age

         chi2(1)      =     9.86
         Prob > chi2  =   0.0017

estat hettest edu

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: educ

         chi2(1)      =     2.39
         Prob > chi2  =   0.1219
```

- Age is the source of heteroskedasticity

Testing for heteroskedasticity, example

- Correcting does change SEs but not by a lot

```
* Regular
qui reg income age edu
est sto reg

* Robust
qui reg income age edu, robust
est sto rob

* Compare
est table reg rob, se p stats(N F)
```

```
----------------------------------------
    Variable |     reg          rob
-------------+--------------------------
         age |  440.24407    440.24407
             |  105.68708    94.815869
             |     0.0001       0.0000
        educ |  706.88408    706.88408
             |  654.62413    612.81005
             |     0.2829       0.2515
       _cons |  14800.355    14800.355
             |  8538.3265    7245.2375
             |     0.0862       0.0438
-------------+--------------------------
           N |        100          100
           F |   14.71002    21.294124
----------------------------------------
                          legend: b/se/p
```

A catch 22?

- Remember the big picture. The sandwich estimator is asymptotically valid even if homokedastic variance so with large enough samples we are safe using the robust option all the time
- With small samples, we would like to test for the heteroskedastic errors
- BUT, we may not have enough power to detect heteroskedasticy with smaller sample
- We could reject the null when the null is true (Type II error)
- Not a clear solution

Back to transformations

- Remember that taking the log(y) tends to help with OLS assumptions? Could it fix the heteroskedastic problem? Yep, mostly

```
reg lincome age edu
...
------------------------------------------------------------------------------
     lincome |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0093932   .0024094     3.90   0.000     .0046113    .0141752
        educ |   .0217054   .0149237     1.45   0.149    -.007914     .0513248
       _cons |   9.895059   .1946512    50.83   0.000     9.50873     10.28139
------------------------------------------------------------------------------
estat hettest, rhs

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: age educ
         chi2(2)      =     5.00
         Prob > chi2  =   0.0821

estat hettest age
         chi2(1)      =     4.98
         Prob > chi2  =   0.0256

estat hettest educ
         chi2(1)      =     0.87
         Prob > chi2  =   0.3500
```

Back to transformations

- Since taking the log has helped with heteroskedasticity, the original and the robust model should be similar

```
* Log income, no robust
qui reg lincome age edu
est sto lm1

* Log income, robust
qui reg lincome age edu, robust
est sto lm1rob

* Compare
est table lm1 lm1rob, se p stats(N F)
----------------------------------------
    Variable |    lm1         lm1rob
-------------+--------------------------
         age |  .00939325    .00939325
             |  .00240939    .00215669
             |     0.0002       0.0000
        educ |  .02170542    .02170542
             |  .01492369    .01349306
             |     0.1491       0.1109
       _cons |  9.8950586    9.8950586
             |   .1946512    .16247044
             |     0.0000       0.0000
-------------+--------------------------
           N |        100          100
           F |  14.651729    21.599741
----------------------------------------
                           legend: b/se/p
```

Alternative: White test

- An alternative test that is popular is the **White test**
- It does use more degrees of freedom. The logic is similar to the other test
- White showed that the errors are homokedastic if ϵ^2 is uncorrelated with all the covariates, their squares, and cross products
- With three covariates, the White test will use 9 predictors rather than 3
- In my opinion, more of a Catch 22
- Easy to implement in Stata (of course)

White

- White test in Stata

```
qui reg income age edu

 estat imtest, white

White's test for Ho: homoskedasticity
         against Ha: unrestricted heteroskedasticity

         chi2(5)      =     23.77
         Prob > chi2  =    0.0002

Cameron & Trivedi's decomposition of IM-test

---------------------------------------------------
              Source |       chi2     df      p
---------------------+-----------------------------
  Heteroskedasticity |      23.77      5    0.0002
            Skewness |       3.77      2    0.1518
            Kurtosis |       2.29      1    0.1302
---------------------+-----------------------------
               Total |      29.83      8    0.0002
---------------------------------------------------
```

- Same conclusion, we reject the null

Big picture

- With large samples, robust SEs buy you insurance but with smaller samples it would be a good idea to test for heteroskedasticity
- Of course, with small samples, the power of the heteroskedasticity test is itself compromised
- No hard rules. Researchers follow different customs; some always add the robust option (I don't)
- **Careful with likelihood ratio tests in the presence of heteroskedasticity**
- Stick to robust F tests to compare nested model (use the test command in Stata)

Summary

- Robust SEs are asymptotically valid even if no heteroskedasticity
- Always suspect unequal variance; very common
- Taking the log transformation may help
- Next class, dealing with unequal variance when we know the source: weighted models
- Weighted models for dealing with heteroskedasticity is sort of old fashioned. I do want to cover weighted models because they are used a lot in survey data analysis and lately in propensity scores