Week 10: Heteroskedasticity II

Marcelo Coca Perraillon

University of Colorado
Anschutz Medical Campus

Health Services Research Methods I
HSMP 7607
2017

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

Outline

- Dealing with heteroskedasticy of known form (old fashioned but worth going over it)
- Weighted least squares
- Lowess once again
- Examples

Heteroskedasticity source is know: multiplicative constant

- Suppose that we know or suspect that the variance is a function of some or all the explanatory variables
- For example: $var(\epsilon|x_1, ...x_p) = \sigma^2 f(x_1, ..., x_p)$
- $f(x_1, ..., x_p) > 0$ because the variance has to be positive. For the moment, we will assume that we know the functional form for $f(x_1, ..., x_p)$
- Another way of writing this for an observation i: $\sigma_i^2 = var(\epsilon_i|x_{1i}, ..., x_{1i}) = \sigma^2 f(x_{1i}, ..., x_{pi})$
- Note that σ^2 is constant on the right side (no subscript i) but it varies according to the values of $x_{1i}, ..., x_{pi}$

Example

- Let's go back to the income and age dataset and estimate the model $income = \beta_0 + \beta_2 age + \epsilon$

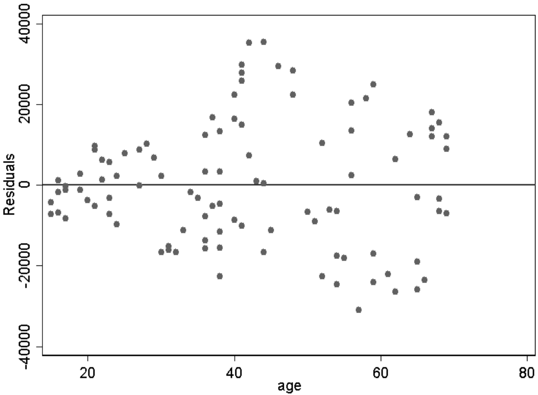
```
 webuse mksp1, clear
reg income age

      Source |       SS           df       MS      Number of obs   =       100
-------------+----------------------------------   F(1, 98)        =     28.21
       Model |  6.5310e+09         1  6.5310e+09   Prob > F        =    0.0000
    Residual |  2.2691e+10        98   231542958   R-squared       =    0.2235
-------------+----------------------------------   Adj R-squared   =    0.2156
       Total |  2.9222e+10        99   295173333   Root MSE        =     15217

------------------------------------------------------------------------------
      income |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   494.4258   93.09552     5.31   0.000     309.6808    679.1709
       _cons |    22870.1   4133.273     5.53   0.000     14667.75    31072.45
------------------------------------------------------------------------------

predict res, res
 scatter res age, yline(0)
```

Example

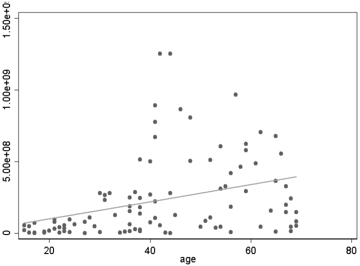


- Assuming that the residual variance is a function of age is a reasonable assumption
- We saw last class that the graphs and the heteroskedastic tests pointed towards age as the source of the problem

Remember the Breusch-Pagan test?

- The Breusch-Pagan test models ${\epsilon_i}^2 = \gamma_0 + \gamma age_i + u_i$

```
qui reg income age
predict ires, rstandard
gen ires2 = ires^2
scatter ires2 age || lfit ires2 age, legend(off)
```



- The square of the residual could be assumed to be a linear function of age

Multiplicative constant

- We will assume that $f(age) = age$, so $var(\epsilon_i|age_i) = \sigma^2 age_i$
- Age is always positive so no risk of getting a negative variance (otherwise, we could take the square).
- The standard error is, of course, $\sigma\sqrt{age_i}$
- Once we assume a functional form for $f(age)$ the rest is not too complicated
- The idea is very simple: we will **transform the variables** in the original model in such a way that the **variance of the new model will be constant** given values of age

Multiplicative constant

- The original model is $income_i = \beta_0 + \beta_1 age_i + \epsilon_i$
- What about if we divide the model by $\frac{1}{\sqrt{age}}$ to obtain:
- $\frac{income_i}{\sqrt{age_i}} = \frac{\beta_0}{\sqrt{age_i}} + \beta_1 \frac{age_i}{\sqrt{age_i}} + \frac{\epsilon_i}{\sqrt{age_i}}$?
- It looks a bit odd and arbitrary but it turns out that this transformation makes the model have constant variance (homoskedastic)
- Remember that we **assumed** that the **true** variance conditional on age is $var(\epsilon_i|age_i) = E(\epsilon_i{}^2|age_i) = \sigma^2 age_i$. So what is the expected value of the transformed variance?
- $E[(\frac{\epsilon_i}{\sqrt{age_i}}|age_i)^2] = \frac{E[\epsilon_i|age_i{}^2]}{age_i} = \frac{\sigma^2 age_i}{age_i} = \sigma^2$
- If confused, it's easier if you remove the conditioning on age:
- $E[(\frac{\epsilon_i}{\sqrt{age_i}})^2] = \frac{E[\epsilon_i{}^2]}{age_i} = \frac{\sigma^2 age_i}{age_i} = \sigma^2$

Big picture

- Remember: we assumed the variance depends on one or more covariates: $\sigma^2 f(x_{1i}, ..., x_{pi})$
- In the example with only one explanatory variable, we assumed the simplest functional form: $\sigma^2 age_i$
- We transformed the data to come up with a new model that has constant variance
- Of course, we do make an assumption: **we assume that we have a good model of the source of heteroskedasticity**
- If the assumption is wrong, then the expected value of the variance in the transformed model no longer is constant. This is a **strong assumption** that can't be verified with the data
- We do this to have better estimates of the variance-covariance matrix; the **new parameters do not have a useful interpretation**

Big picture: weighted least squares

- We will get back to this shortly but the way we will estimate this model in Stata is by **weighting** the regression by $\frac{1}{age}$
- The weight is proportional to the inverse of the variance $var(\epsilon_i|age_i) = \sigma^2 age_i$
- The intuition is actually very simple: we are giving **less importance to observations that have a higher variance**. For older people, $\frac{1}{age}$ is lower than for younger people
- This is what we want since we assumed (based on some evidence) that the variance is a linear function of age
- If we were to transform the variables, we would have to divide all the variables by $\frac{1}{\sqrt{age}}$

Example

- Stata implementation is fairly easy; we use the option [aw] to incorporate the weights

```
gen w = 1/age
qui reg income age educ
est sto orig
qui reg income age educ [aw=w]
est sto weig
est table orig weig, se p stats(N)
----------------------------------------
    Variable |    orig         weig
-------------+--------------------------
         age |  440.24407    460.26434
             |  105.68708    102.59664
             |     0.0001       0.0000
        educ |  706.88408    780.33877
             |  654.62413    575.55667
             |     0.2829       0.1783
       _cons |  14800.355    12902.949
             |  8538.3265    6716.1242
             |     0.0862       0.0576
-------------+--------------------------
           N |        100          100
----------------------------------------
```

- Focus on SEs; remember, we care about the **new variance-covariance matrix**

Example

- So how doe this compare to the sandwich?

```
qui reg income age educ, robust
est sto sand
est table orig weig sand, se p stats(N)
-----------------------------------------------------
    Variable |    orig         weig         sand
-------------+---------------------------------------
         age |  440.24407    460.26434    440.24407
             |  105.68708    102.59664    94.815869
             |     0.0001       0.0000       0.0000
        educ |  706.88408    780.33877    706.88408
             |  654.62413    575.55667    612.81005
             |     0.2829       0.1783       0.2515
       _cons |  14800.355    12902.949    14800.355
             |  8538.3265    6716.1242    7245.2375
             |     0.0862       0.0576       0.0438
-------------+---------------------------------------
           N |        100          100          100
-----------------------------------------------------
```

- Which one is better? With larger samples, bet is on the sandwich because it doesn't depend on knowing the form of heteroskedasticity

Example: Wooldrigde 8.1

- Model to explain net total financial wealth (nettfa) as a function of income and other covariates including age, sex, and an indicator of whether the person is eligible for 401K
- Age enters quadratic and is centered at 25
- We will replicate the models presented in Table 8.1, page 274
- Sample restricted to single people, $fsize = 1$
- We assume source of unequal variance is due to income

Example: Wooldridge 8.1

```
bcuse 401ksubs

qui reg nettfa inc
est sto m1

qui reg nettfa inc [aw=1/inc]
est sto m2

qui reg nettfa inc age252 male e401k
est sto m3

qui reg nettfa inc age252 male e401k [aw=1/inc]
est sto  m4
```

- Note that we do not need to create a weight variable; option aw takes expressions

Example: Replicate Table 8.1

```
est table m1 m2 m3 m4, se p stats(N)

--------------------------------------------------------------------
    Variable |     m1            m2            m3            m4
-------------+------------------------------------------------------
         inc |  .82068148     .78705231      .7705833     .74038434
             |      .0609     .06348144       .061452     .06430291
             |     0.0000        0.0000        0.0000        0.0000
      age252 |                              .02512668     .01753728
             |                              .00259339      .0019315
             |                                 0.0000        0.0000
        male |                              2.4779269     1.8405293
             |                              2.0477762     1.5635872
             |                                 0.2264        0.2393
       e401k |                              6.8862229     5.1882807
             |                              2.1232747     1.7034258
             |                                 0.0012        0.0024
       _cons | -10.570952    -9.5807017     -20.98499    -16.702521
             |  2.0606775     1.6532837      2.472022     1.9579947
             |     0.0000        0.0000        0.0000        0.0000
-------------+------------------------------------------------------
           N |       2017          2017          2017          2017
--------------------------------------------------------------------
                                                    legend: b/se/p
```

- In general SEs went up, not by a lot

Weighted regression

- Weighted regression is an example of **generalized least squares** or GLS
- Weighted models, not just our regular linear model, play an important role in many applied areas
- You will encounter them in survey data: each observation is given a weight because each observation represents many people in the population
- Survey weights tend to be a **black box**: they are adjusted for non-response and other factors like oversampling of certain populations (like the very old or minorities)
- The **weights add up to the population size**
- (See the article about one person influencing polls in last election because that person was given a very large weight)

Weighted regression

- Next semester, you will see that you can use the **inverse of the propensity score** to obtain a weighted treatment effect
- The weights are designed to give **more importance to observations that are similar** between treatment and control groups
- Unweighted, treatment and control are not comparable; weighted, they will become comparable (at least for the observed covariates)
- You will need to assume that *unobservables* are also balanced, which tends to be a difficult assumption to satisfy
- In other words, you'll need to assume ignorable treatment assignment or no unmeasured confounders or selection on observables or exchangeability
- Our **old friend Lowess** is also an example of a weighted model

Lowess,redux

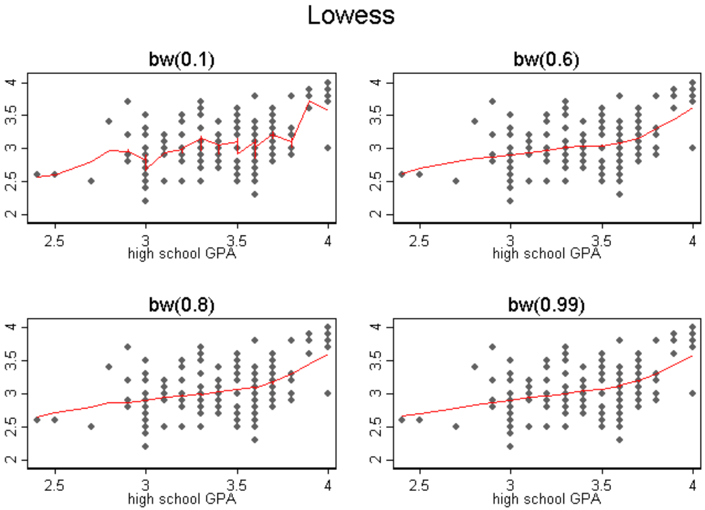
- Lowess is handy way to compute the $E[Y]$ around an area of X; less sensitive (i.e. robust) to sparse points and it's not influenced by all points (hence the **local** part). Recall that Lowess stands for **Locally Weighted Scatterplot Smoothing**
- Lowess is an example of a non-parametric method and a weighted regression
  1. For **each** point in the data, use a **window around that point** on the x-axis to calculate $E[Y]$. Use only observations *within* that window
  2. Regress y on x around window and **weigh the data** so that observations closer to the chosen point are given more weight (importance)
  3. Predict $\hat{y}$ at chosen point x
  4. **Repeat** algorithm **for all points** in the dataset
- The details change a bit but that's the essence of the method; it's a computationally intense method – needs to run a weighted regression for **each point** in dataset

Code for Lowess

- If no options, default is bw(0.8); always a good idea to try other windows

```
lowess colgpa hsgpa, bw(0.1) nograph gen(cgpa_l1)
lowess colgpa hsgpa, bw(0.6) nograph gen(cgpa_l6)
lowess colgpa hsgpa, bw(0.8) nograph gen(cgpa_l8)
lowess colgpa hsgpa, bw(0.99) nograph gen(cgpa_l9)
scatter colgpa hsgpa || line cgpa_l1 hsgpa, sort color(red) ///
        saving(l1.gph, replace) legend(off) title("bw(0.1)")
scatter colgpa hsgpa || line cgpa_l6 hsgpa, sort color(red) ///
        saving(l6.gph, replace) legend(off) title("bw(0.6)")
scatter colgpa hsgpa || line cgpa_l8 hsgpa, sort color(red) ///
        saving(l8.gph, replace) legend(off) title("bw(0.8)")
scatter colgpa hsgpa || line cgpa_l9 hsgpa, sort color(red) ///
        saving(l9.gph, replace) legend(off) title("bw(0.99)")
graph combine l1.gph l6.gph l8.gph l9.gph, title("Lowess")
graph export lowess.png, replace
```

Lowess "smoothed" college and high school grades; different bandwidths



Lowess weights

- The weights in Lowess are a bit complicated but not uncommon
- You'll encounter similar non-parametric methods in regression discontinuity (more weight to observations close to cut-off points)
- Stata has the details:

Methods and formulas

Let y_i and x_i be the two variables, and assume that the data are ordered so that $x_i \le x_{i+1}$ for $i = 1, \ldots, N-1$. For each y_i, a smoothed value y_i^s is calculated.

The subset used in calculating y_i^s is indices $i_- = \max(1, i-k)$ through $i_+ = \min(i+k, N)$, where $k = \lfloor (N \times \texttt{bwidth} - 0.5)/2 \rfloor$. The weights for each of the observations between $j = i_-, \ldots, i_+$ are either 1 (**noweight**) or the tricube (default),

$$w_j = \left\{ 1 - \left(\frac{|x_j - x_i|}{\Delta} \right)^3 \right\}^3$$

where $\Delta = 1.0001 \max(x_{i_+} - x_i, x_i - x_{i_-})$. The smoothed value y_i^s is then the (weighted) mean or the (weighted) regression prediction at x_i.

How do weights work?

- Here is an intuitive way to understand weights
- We will simulate 10 observations and estimate a model in which each observation has the same weight
- Then we will change the weight of the last observation so it's worth for 10 observations
- We will see that the new weighted model is the **same as the model in which we replicate the last observation 10 times** and run an unweighted model

How do weights work?

- Here is the code

```
clear
set seed 1234567
set obs 10
gen x = rnormal(1, 3)
gen y = 2 + 3*x + rnormal(0,1)
gen wgt = 1

* No weights
reg y x
est sto orig
* Same weight
reg y x [aweight = wgt]
est sto samew
* Make the last observation count for 10
gen     wgt1 = wgt
replace wgt1 = 10 if _n==10
* Weighted
reg y x [aweight = wgt1]
est sto wgt1
* Expand obs
expand 10 if _n ==10
* Unweighted but expanded
reg y x
est sto expand
```

How do weights work?

- Compare models; the new weight is the same as replicating the last observation 10 times (well, 9)

```
. est table orig samew wgt1 expand

------------------------------------------------------------------
    Variable |    orig         samew          wgt1         expand
-------------+----------------------------------------------------
           x |  2.9494298    2.9494298    2.9869977    2.9869977
       _cons |  2.0051343    2.0051343    1.8564805    1.8564805
------------------------------------------------------------------
```

- Careful, **several types of weights** (inverse probability, analytical). See “help weights”
- Here, we are using **analytic weights**, their value doesn’t matter, only differences (Stata scales them)

Back to heteroskedasticity

- The weighted SEs are **more efficient** so we want to use them for statistical inference; we do not care about the new R^2 or the estimated coefficients
- The most important question is, **what if we got the functional form of the unequal variance wrong**?
- In the income, age, and education model we suspect age is the reason for unequal variance, but is $f(age_i) = age_i$ right?
- In most practical applications, we do not know of course and models are seldom so simple

Problem getting $f()$ wrong

- 1) We get the SEs wrong, of course. But we can apply robust regression to the weighted OLS estimates... (getting meta here)
- 2) If $f()$ wrong, then weighted SEs not more efficient
- So what should we do?
- In most practical applications, we do not know the exact reason why there is unequal variance
- If samples are large enough, most practitioners will use the Huber-White robust SEs. Period

Compare models

- Let's compare all options

```
* Compare models

* No correction
qui reg nettfa inc
est sto m1

* WLS
qui reg nettfa inc [aw=1/inc]
est sto m2

* Huber-White
qui reg nettfa inc, robust
est sto rob1

* No correction
qui reg nettfa inc age252 male e401k
est sto m3

* WLS
qui reg nettfa inc age252 male e401k [aw=1/inc]
est sto  m4

* Huber-White
qui reg nettfa inc age252 male e401k, robust
est sto rob2
```

Compare models

- My bet is on robust option (N= 2017)

```
est table m1 m2 rob1 m3 m4 rob2, se p stats(N F)
```

Variable	m1	m2	rob1	m3	m4	rob2
inc	.82068148	.78705231	.82068148	.7705833	.74038434	.7705833
	.0609	.06348144	.10359361	.061452	.06430291	.09957192
	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
age252				.02512668	.01753728	.02512668
				.00259339	.0019315	.00434415
				0.0000	0.0000	0.0000
male				2.4779269	1.8405293	2.4779269
				2.0477762	1.5635872	2.0583585
				0.2264	0.2393	0.2288
e401k				6.8862229	5.1882807	6.8862229
				2.1232747	1.7034258	2.2865772
				0.0012	0.0024	0.0026
_cons	-10.570952	-9.5807017	-10.570952	-20.98499	-16.702521	-20.98499
	2.0606775	1.6532837	2.5302719	2.472022	1.9579947	3.495186
	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
N	2017	2017	2017	2017	2017	2017
F	181.59949	153.71407	62.76006	73.747631	63.127351	28.960727

legend: b/se/p

Summary

- Heteroskedasticity is more common than not
- It has become the standard practice with larger sample to just add the robust option
- Careful with likelihood ratio tests, use the "test" command for testing if you use robust
- Use the tests for heteroskedasticity if in doubt
- Get the logic of weighted regression; it will come back often...