Stata and the problem of heteroscedasticity

Every serious subject has its jargon. Economists need to know about heteroscedasticity. I take this example because it is virtually impossible to pronounce, and impossible to use the word in front of a class without everyone bursting out into laughter. Indeed, most spell-check programmes reject it, and offer improbable or embarrassing alternatives. (quote from johnkay.com)

Our Plan

- short Stata review
- When heteroscedasticity might occur
- Consequences of heteroscedasticity
- Detecting heteroscedasticity
- Dealing with heteroscedasticity

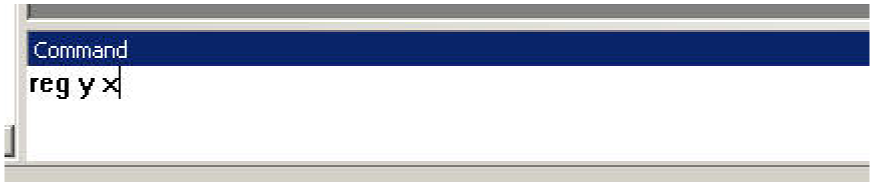
→ Call heteroscedasticity just **„HK“** for simplicity

Short Stata Review

Syntax based and GUI → syntax is faster

e.g. help hettest

Where? Command window



<u>Note:</u> set mem 100m

Before you open a data-set!

When heteroscedasticity might occur

• Errors may increase if value of explanatory variables increase

e.g. family income and family expenditures on vacations or

sales of large vs. small firms → firm size

• Errors may increase if extreme positions e.g. attitudes (hourglass shape)

• or for different subpopulations e.g. expenditures and income for white vs. black

• misspecification can cause HK e.g. instead of using Y you should use log of Y, instead of X you should use X^2..

Consequences of heteroscedasticity

First: does not result in biased estimates (this is good) but:

→But OLS estimates are no longer BLUE

<u>That:</u>

-Variance will not be the smallest anymore (bad!)

- Standard errors are biased (worse!) → affects t-test and significance

Such that significance can be too high or too low → draw wrong conclusions (really bad!)

So for OLS it might be „okay“ for some but for other regression like logistic regression HK gets really bad even affecting the parameter estimates

Detecting heteroscedasticity (1)

1. Visual Inspection

Plot the residues against the fitted values of Y <u>or</u> the suspected trouble maker!

Residues against Y → rvfplot

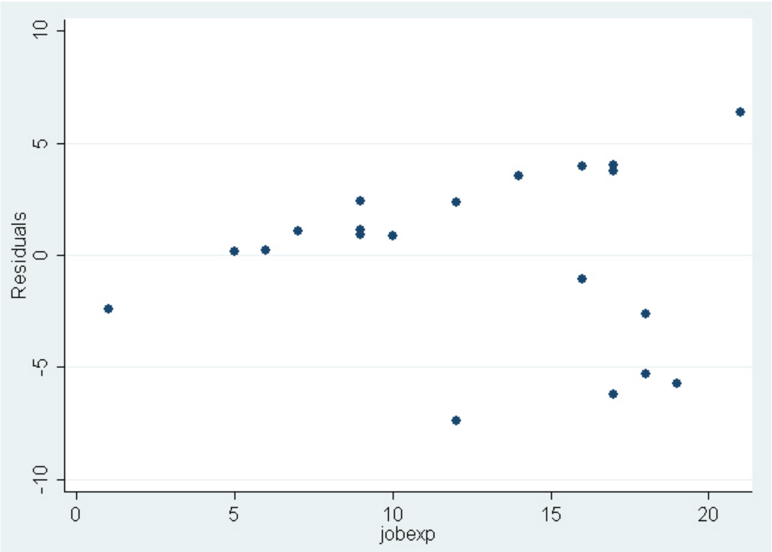
Residues against X → rvpplot x

<u>Example:</u>

Open hk.dta

reg income educ jobexp

rvpplot jobexp



Try rvfplot and rvpplot educ → what do you see?

Detecting heteroscedasticity (2)

2. Breusch-Pagan Test for HK

H_0: error variances are all equal

H_1: error variances are a multiplicative function of one or more variables

<u>Example:</u>

quietly reg income educ jobexp

estat hettest

```
Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of income

         chi2(1)      =     0.12
         Prob > chi2  =   0.7238
```

Low Chi-square value → HK not a problem (or wasn‘t a multiplicative function of the predicted values)

→ **see:** exercise to do this test manually

Detecting heteroscedasticity (3)

3. White‘s general test for HK

→ BP works well if linear forms but not for non-linear forms

But adds many terms in the test regression → sometimes a simpler test like BP is more appropiate

Example:

quietly reg income educ jobexp

estat imtest, white

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

         chi2(5)      =      8.98
         Prob > chi2  =    0.1100

Cameron & Trivedi's decomposition of IM-test
```

Source	chi2	df	p
Heteroskedasticity	8.98	5	0.1100
Skewness	2.39	2	0.3022
Kurtosis	0.98	1	0.3226
Total	12.35	8	0.1363

Detecting heteroscedasticity (4)

4. Goldfeldt-Quant test

- Useful if we can correctly identify the variable to use for sample separation but other tests are simpler and more flexible

e.g. let educ be the trouble maker

Example:

reg income educ jobexp if educ <=10

reg income educ jobexp if educ >=15

→ Use RSS and compute $F=RSS_{low}/RSS_{high}$

Here: F(3,3)=113.01/45.53=2.48< table value → so not HK!

Exercise: do the same for experience and find a cut off value!

Dealing with heteroscedasticity (1)

1. **Respecify the model / transform the variables**

- HK can be a consequence from improper model specifaction e.g use logs..

2. Use robust standard errors

- Relaxes some OLS assumptions and gives better standard errors

<u>Example:</u>

reg income educ jobexp, robust

Compare with

reg income educ jobexp

???

Dealing with heteroscedasticity (2)

3. Use Weighted Least Square (WLS)

- GLS estimation minimizes a weighted sum of squared residuals

- That error terms with large variance get a smaller weight than observations with small variance

Example:

Suspect education to be the trouble maker → use it as the weight (how to choose???)

Gen inveduc=(1/educ)^2

Reg income educ jobexp [aw = inveduc]

Where aw = analytical weight

```
(sum of wgt is   4.4265e-01)

      Source |       SS       df       MS              Number of obs =      20
-------------+------------------------------           F(  2,    17) =   86.20
       Model |  1532.21449     2  766.107244           Prob > F      =  0.0000
    Residual |  151.090319    17  8.88766581           R-squared     =  0.9102
-------------+------------------------------           Adj R-squared =  0.8997
       Total |  1683.30481    19  88.5949898           Root MSE      =  2.9812

------------------------------------------------------------------------------
      income |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   1.795724   .1555495    11.54   0.000     1.467544    2.123905
      jobexp |   .4587992   .1628655     2.82   0.012      .115183    .8024155
       _cons |  -3.159669    1.94267    -1.63   0.122    -7.258346    .9390065
------------------------------------------------------------------------------
```

Exercises (1)

BP-test by hand

White-test by hand

see: handout