**Week 6 Lab Handout-Heteroskedasticity**

**PA 5032 – Applied Regression**

February 26th, 2021

Wenchen Wang

**PART A: Detecting Heteroskedasticity ~20min**

**PART B: Correcting Heteroskedasticity~ 15min**

**PART C: Report 2 Questions?**

**Data:** EITC2.dta (On Class Canvas Site)

**Contents:** 13,741 observations

**Variables:** children = number of children

earn = annual earnings

ed = years of education

tenure = years with current company

nonwhite = dummy variable where 1 = Hispanic/Black

age = age of women

age2: squared of age

lnearn: log of earn

**PART A: Detecting Heteroscedasticity**

The A5 assumption from Gauss-Markov assumptions holds**: that there is constant variance in the error term of the model.** This is a dubious assumption.

Consequences of Heteroskedasticity:

* Estimates are unbiased;
* SEs on the estimates will be wrong;
* unreliable hypothesis testing and confidence: could be too big & harder to find significant results; or could be too small & too easy to find significant results.

There are three main ways to check for heteroscedasticity:

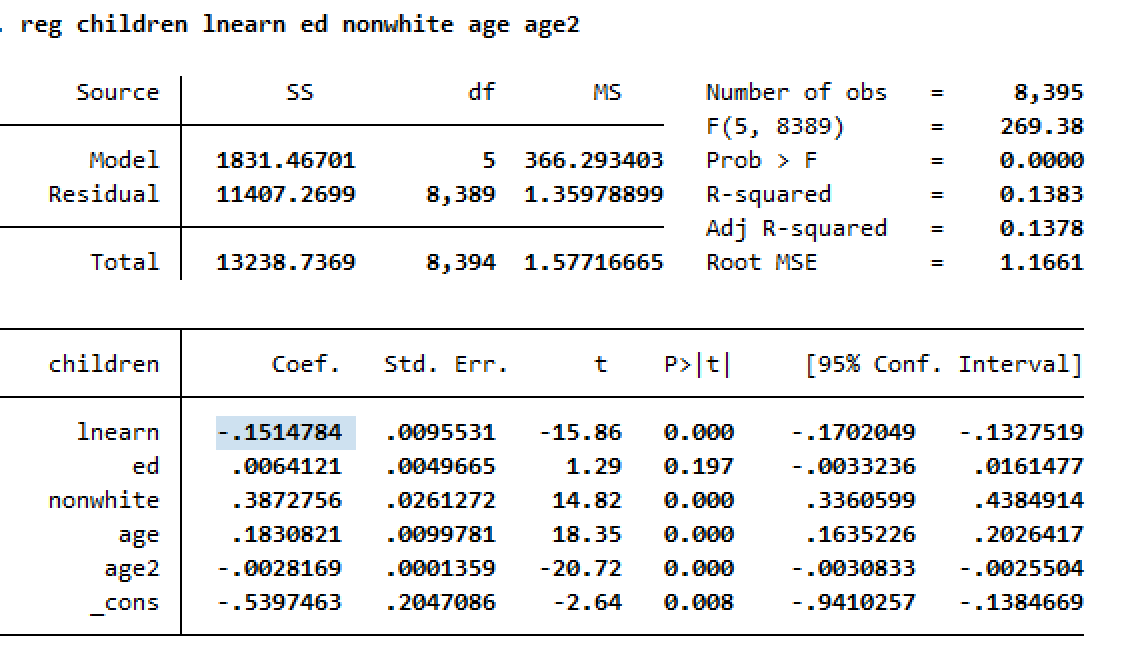
1. Eye-balling residuals

2. Breusch-Pagan test

3. White Test

1. Eye-Ball the Residuals

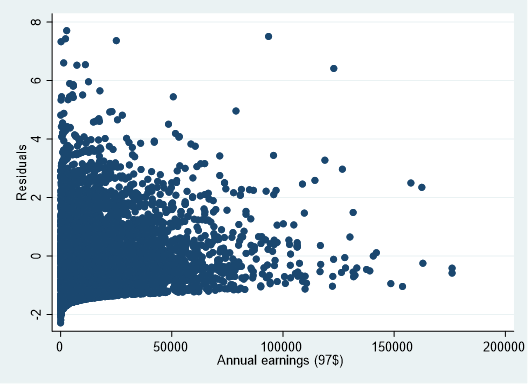
*reg children lnearn ed nonwhite age age2*



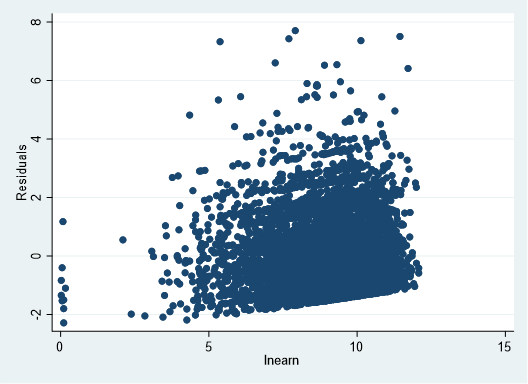
How can we interpret the coefficient on lnearn (Interpreting logged independent variable)?

*predict res, resid*

*scatter res earn*

**

*rvpplot lnearn*



How to eyeball the residual plot for heteroskedasticity?

1. The Breusch-Pagan Test

This test present evidence against the null hypothesis that the variance of the error term or residuals should be the same for all values:

H0: Var(ε|x1,x2,…xk)=σ2

In the default version of the test (Breusch-Pagan), the null hypothesis also includes the assumption that the residuals are normally distributed:

E(ε2∣∣x1,x2,…xk)=σ2

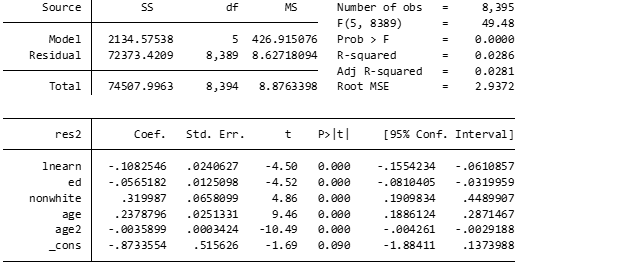
Breusch-Pagan by Hand:

Step 1: Generate a squared residual term

*gen res2=res\*res*

Step 2: Regress this squared residual on the control variables

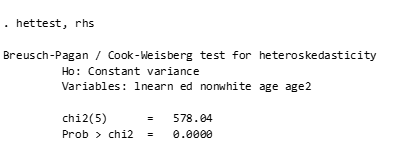
*reg res2 lnearn ed nonwhite age age2*



How do we interpret the results here? Do we reject or fail to reject the Null Hypothesis?

Breusch-Pagan via Stata (hettest):

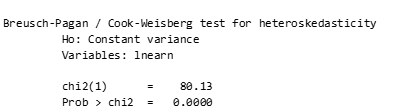
*reg children lnearn ed nonwhite age age2*



rhs: specifies that tests for heteroskedasticity be performed for the right-hand-side (explanatory) variables of the fitted regression model.

OR to look at just one specific variable that you are concerned is causing the heteroskedasticity

*hettest lnearn*



Do we reject the null hypothesis for either of these tests?

**\*NOTE: If you Reject the null, there is heteroscedasticity**

1. White Test

Now let’s say we really think the true relationship between number of children and earnings is flipped, that the number of children you have more so dictates your earnings. Let’s now try running our full model, but with earnings as the key dependent to see if that helps eliminate the heteroskedasticity. We will need to recapture the residuals.

*reg earn children ed nonwhite age age2*

*predict Res, resid*

White Test (by hand):

Step 1. Square the residual

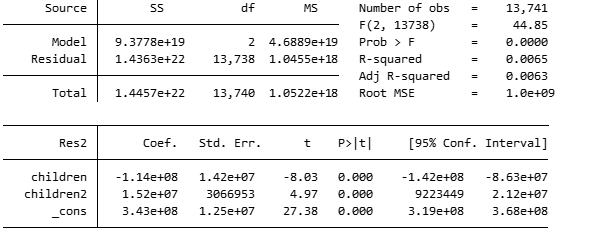
Step 2. Square the potentially problematic independent variable(s)

Step 3. Put them in a regression together

*gen Res2= Res\*Res*

*gen children2=children\*children*

*reg Res2 children children2*



Can we reject the null hypothesis of homoskedasticity?

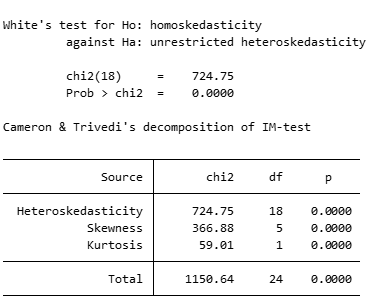
White Test (via Stata: imtest, white)

To run the white test in Stata, you just need to run a command immediately following your regression, called imtest.

\***NOTE**: When using imtest function, you need to rerun the original regression since the imtest, white command runs based upon whatever regression was run preceding it.

*reg children earn ed nonwhite age age2*

*imtest, white*



Can we reject the Null Hypothesis?

**PART B: Correcting Heteroskedasticity**

How to Correct for Heteroskedasticity:

1. Adjust SEs: Heteroscedasticity-corrected standard errors (Huber/White, Robust SEs)

2. Redefine variables: If you can change variables to remove the heteroscedasticity problem (rare case).

Let’s compare three different regression outputs to see how different correction tactics compare. A regular model, one with the adjusted SEs (robust), and one with the cluster correction.

(1) Reg y x

(2) Reg y x, vce(robust)

(3) Reg y x, vce(cluster var)

*reg earn children ed nonwhite age age2*

*outreg2 using Lab\_6.doc, replace ctitle(regular SE) bdec(3) adjr*

*reg earn children ed nonwhite age age2, vce(robust)*

*outreg2 using Lab\_6.doc, append ctitle(robust SE) bdec(3) adj*

*reg earn children ed nonwhite age age2, vce(cluster earn)*

*outreg2 using Lab\_6.doc, append ctitle(cluster SE) bdec(3) adjr*

|  |  |  |  |
| --- | --- | --- | --- |
|  | (1) | (2) | (3) |
| earnings | regular SE | robust SE | cluster SE |
|  |  |  |  |
| children | -1,801.098\*\*\* | -1,801.098\*\*\* | -1,801.098\*\*\* |
|  | (107.316) | (111.846) | (336.635) |
| ed | 3.242 | 3.242 | 3.242 |
|  | (53.113) | (57.068) | (113.817) |
| nonwhite | -1,043.619\*\*\* | -1,043.619\*\*\* | -1,043.619\*\*\* |
|  | (286.294) | (300.514) | (374.445) |
| age | -762.568\*\*\* | -762.568\*\*\* | -762.568\*\* |
|  | (110.735) | (121.287) | (379.323) |
| age2 | 11.014\*\*\* | 11.014\*\*\* | 11.014\*\* |
|  | (1.522) | (1.685) | (5.130) |
| Constant | 25,089.405\*\*\* | 25,089.405\*\*\* | 25,089.405\*\* |
|  | (1,979.628) | (2,196.376) | (11,514.908) |
|  |  |  |  |
| Observations | 13,741 | 13,741 | 13,741 |
| Adjusted R-squared | 0.036 | 0.036 | 0.036 |

Standard errors in parentheses

\*\*\* p<0.01, \*\* p<0.05, \* p<0.1

How do the three models compare? Did the coefficients change? What about the SEs and significance?

**PART C: Report 2 Questions?**