Class Oct 10

Kevin R Foster, CCNY, ECO B2000
Fall 2013

Exam 1 is Oct 17 in NAC 6-150

Heteroskedasticity-Consistent Errors in SPSS

The Stock and Watson textbook uses heteroskedasticity-consistent errors (sometimes called Eicker-Huber-White errors, after the authors who figured out how to calculate them). However SPSS does not have an internal option on a drop-down list to compute heteroskedasticity-consistent standard errors. However with just a bit more work we can still produce the desired output.

How can we get heteroskedasticity consistent standard errors? Google (our goddess). I found an SPSS macro, written by Andrew F. Hayes at Ohio State University, who wrote the code and provided documentation. Download the macro, `hcreg.sps`, (from InYourClass, in the "Kevin Foster SPSS" Group) and start up SPSS. Before you do the regressions, click "`File`" then "`Open`" then "`Syntax…`". Find the file that you downloaded (`hcreg.sps`) and open it. This will open the SPSS Syntax Editor. All you need to do is choose "`Run`" from the top menu then "`All`". There should not be any errors. You need to run this macro each time you start up SPSS but it will stay in memory for the entire session until you close SPSS.

The macro does not add extra options to the menus, however. To use the new functionality we need to write a bit of SPSS syntax ourselves. For example, suppose we are using the PUMS dataset and want to regress commute time (JWMNP) on other important variables, such as Age, gender, race/ethnicity, education, and borough.

We will have to use the "`Name`" of the variable rather than the label. This is inconvenient but not a terrible challenge. Age conveniently has name "`Age`" but the gender dummy has name "`female`"; the race/ethnicity variables are "`africanamerican`" "`nativeamerican`" "`asianamerican`" "`raceother`" and "`Hispanic`"; education is "`educ_hs`" "`educ_somecoll`" "`educ_collassoc`" "`educ_coll`" and "`educ_adv`"; boroughs are "`boro_bx`" "`boro_si`" "`boro_bk`" and "`boro_qns`". (Note that we leave one out for education and borough.)

Go back to the SPSS Syntax Editor: from the Data View choose "File" "New" "Syntax". This will re-open the editor on a blank page. Type:

```
HCREG dv = JWMNP/iv = Age female africanamerican nativeamerican
asianamerican raceother Hispanic educ_hs educ_somecoll educ_collassoc
educ_coll educ_adv boro_bx boro_si boro_bk boro_qns.
```

Then go to "Run" on the top menu and choose "All" and watch it spit out the output.

Your output should look like this,

```
Run MATRIX procedure:


HC Method
 3


Criterion Variable
 JWMNP


Model Fit:
       R-sq          F        df1        df2          p
      .0475   491.2978    16.0000 132326.000      .0000


Heteroscedasticity-Consistent Regression Results
              Coeff     SE(HC)          t      P>|t|
Constant    26.7397      .3700    72.2637      .0000
Age           .0450      .0054     8.3550      .0000
female       -.2820      .1404    -2.0085      .0446
africana     7.9424      .1999    39.7312      .0000
nativeam     4.2621     1.3060     3.2635      .0011
```

```
asianame     5.2494      .2270    23.1237      .0000
raceothe     3.5011      .2720    12.8696      .0000
Hispanic     1.9585      .2269     8.6317      .0000
educ_hs     -1.1125      .2701    -4.1192      .0000
educ_som     -.7601      .2856    -2.6611      .0078
educ_col      .2148      .3495      .6145      .5389
educ_c_1     1.1293      .2720     4.1517      .0000
educ_adv    -1.3747      .2847    -4.8281      .0000
boro_bx      8.3718      .2564    32.6485      .0000
boro_si     12.7391      .3643    34.9712      .0000
boro_bk      9.6316      .1882    51.1675      .0000
boro_qns    10.2350      .1932    52.9754      .0000

------ END MATRIX -----
```

Did that seem like a pain? OK, here's an easier way that also adds some more error-checking so is more robust.

First do a regular OLS regression with drop-down menus in SPSS. Do the same regression as above, with travel time as dependent and the other variables as independent, and note that just before the output you'll see something like this,

```
REGRESSION
/MISSING LISTWISE
/STATISTICS COEFF OUTS R ANOVA
/CRITERIA=PIN(.05) POUT(.10)
/NOORIGIN
/DEPENDENT JWMNP
/METHOD=ENTER Age female africanamerican nativeamerican asianamerican
raceother Hispanic educ_hs educ_somecoll educ_collassoc educ_coll educ_adv
boro_bx boro_si boro_bk boro_qns.
```

This is the SPSS code that your drop-down menus created. You can ignore most of it but realize that it gives a list of all of the variable names (after "`/METHOD=ENTER`") so you can do this regression and just copy-and-paste that generated list into the hcreg syntax.

The other advantage of doing it this way first is that this will point out any errors you make. If you put in too many dummy variables then SPSS will take one out (and note that in "`Variables Removed`" at the beginning of the output). If that happens then take that out of the list from `hcreg` or else that will cause errors. If the SPSS regression finds other errors then those must be fixed first before using the `hcreg` syntax.

The general template for this command is "`HCREG`", the name of the macro, then "`DV = `" with the name of the **D**ependent **V**ariable, "`IV = `" with the names of the **I**ndependent **V**ariables, and then a period to mark the end of a command line.

The macro actually allows some more fanciness. It contains 4 different methods of computing the heteroskedasticity-consistent errors. If you follow the "`IV = `" list with "`/method = `" and a number from 1 to 5 then you will get slightly different errors. The default is method 3. If you type "`/method = 5`" then it will give the homoskedastic errors (the same results as if you did the ordinary regression with the SPSS menus).

The macro additionally allows you to set the constant term equal to zero by adding "`/constant = 0`"; "`/covmat = 1`" to print the entire covariance matrix; or "`/test = q`" to test if the last q variables all have coefficients equal to zero. Prof. Hayes did a very nice job, didn't he? Go to his web page for complete documentation.

The Syntax Editor can be useful for particular tasks, especially those that are repetitive. Many of the drop-down commands offer a choice of "Paste Syntax" which will show the syntax for the command that you just implicitly created with the menus, which allows you to begin to learn some of the commands. The Syntax Editor also allows you to save the list of commands if you're doing them repeatedly.

This syntax, to perform the regressions, is

```
HCREG dv = JWMNP/iv = Age female africanamerican nativeamerican
asianamerican raceother Hispanic educ_hs educ_somecoll educ_collassoc
educ_coll educ_adv boro_bx boro_si boro_bk boro_qns.
```

```
HCREG dv = JWMNP/iv = Age female africanamerican nativeamerican
asianamerican raceother Hispanic educ_hs educ_somecoll educ_collassoc
educ_coll educ_adv boro_bx boro_si boro_bk boro_qns
/method = 5 .
```

Do those in SPSS and with the regression with the drop menus for comparison. You will see that the results, between the homoskedastic method=5 and the choosen-from-drop-lists, are identical. More precisely, all of the coefficient estimates are the same in every version but the standard errors (and therefore t statistics and thus p-values or Sig) are different between the two hcreg versions (but hcreg method 5 delivers the same results as SPSS's drop down menus).

Heteroskedasticity-Consistent Errors in R

These are HCerrors, in the "sandwich" package, which depends on "zoo" package; probably the easiest implementation is via the "lmtest" package. So install those 3.

On my Win7 machine, I find it easiest to download the packages, "Install from local zip file", [that way I don't need a wifi connection every time] then drop in these commands,

```
library("zoo")
library("sandwich")
library("lmtest")
```

Then load in the data; for example if it's the PUMS data that I gave,

```
dat1 = read.csv("ACS_2008_2011_NYC_med.csv")
```

I'll show a very simple (simplistic, even) regression of $Wage_i = \beta_0 + \beta_1 Age_i + \varepsilon_i$.

Start by defining Y and X:

```
Y <- dat1$INCWAGE
X <- dat1$AGE
```

Then for the regression,

```
summary(Y)
```

```
summary(X)

regression1 <- lm(Y ~ X)

summary(regression1)

coeftest(regression1, df = Inf, vcov = vcovHC(regression1,
type = "const"))

coeftest(regression1, df = Inf, vcov = vcovHC(regression1))
```

Let me explain those lines step-by-step. The commands, `summary(Y)` and `summary(X)` just give summary statistics (quartiles, mean, stdev) – always a good habit to get into, you want to check that those seem reasonably close to what you would expect.

Then the model is defined with the command `regression1 <- lm(Y ~ X)` which sets a Linear Model (lm) of a y-variable as explained by a (list of) X variable(s).

The command `summary()` is very flexible, so if used on a `lm()` model, it knows what you want. It gives the homoscedastic errors though; to get the heteroskedastic errors takes a bit more.

The command `coeftest` will do a variety of coefficient tests; if you give it the first formulation (with `type="const"`) you get the same standard errors as in the summary. You don't often need this step, I'm just showing you each little detail. Slightly more interestingly, you can leave out the `type="const"` and use the default of `vcovHC`, to get the heteroskedasticity-consistent standard errors. (Econometricians have worked their little butts off, coming up with variations on these, so there are HC0 through HC5 just in this package, don't worry for now about which one to use or if they don't quite perfectly match the SPSS ones.)

So here is a bit of code to do a simple linear regression, 3 different times for slightly different subgroups:

```
# alt version, where labels help a bit more

summary(dat1$INCWAGE)

summary(dat1$AGE)

regression1 <- lm(dat1$INCWAGE ~ dat1$AGE)

summary(regression1)

coeftest(regression1, df = Inf, vcov = vcovHC(regression1))
```

```
# worry about zeros

subgroup2 <- (dat1$INCWAGE > 0)

summary(dat1$INCWAGE[subgroup2])

summary(dat1$AGE[subgroup2])

regression2 <- lm(dat1$INCWAGE[subgroup2] ~
dat1$AGE[subgroup2])

summary(regression2)

coeftest(regression2, df = Inf, vcov = vcovHC(regression2))


# prime-age

subgroup3 <- (dat1$INCWAGE > 0) & (dat1$AGE >= 25) &
(dat1$AGE <= 55)

summary(dat1$INCWAGE[subgroup3])

summary(dat1$AGE[subgroup3])

regression3 <- lm(dat1$INCWAGE[subgroup3] ~
dat1$AGE[subgroup3])

summary(regression3)

coeftest(regression3, df = Inf, vcov = vcovHC(regression3))
```

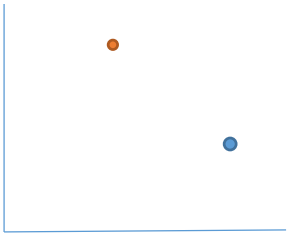
On Rankings:

We often see statistics reported that rank a number of different units based on a number of different measures of outcomes. For instance, these could be the US News ranking of colleges, or magazine rankings of city livability, or sports rankings of college teams, or any of a multitude of different things. We would hope that statistics could provide some simple formulas; we would hope in vain.

In the simplest case, if there is just a single measured variable, we can rank units based on this single measure, however even in this case there is rarely a clear way of specifying which rankings are based on differences that are large and which are small. (The statistical theory is based on "order statistics.") If the outcome measure has, for example, a normal distribution, then there will be large number of units with outcomes right around the middle, so even small measurement errors can make a big difference to ranking.

In the more complicated (and more common) case, we have a variety of measures of outcomes and want to rank units based on some amalgamation of these outcomes. A case where a large number of inputs generates a single unit output looks like a utility function from micro theory: I face a choice of hundreds (or thousands) of different goods, which I put into a single ranking: I say that the utility of some bundle of goods is higher than the utility of some other bundle and so would rank it higher (even if both were affordable).

In the simplest case, if there are just two outcomes that are important, I could graph these bundles as:



So how do we know whether the orange bundle is better than the blue? For some indifference curves it would be; for others it would not. Only if one bundle had more of both would there be a clear ranking possible.

I can observe individual choices and infer what the person's indifference curves look like, but what about choices by a group of people?

There is no way to generate a composite utility function that completely and successfully takes account of the information of individual choices! (This result is due to CCNY alumnus and Nobel Laureate Ken Arrow.)

Many rankings take an equal weighting of each item, but there is absolutely no good reason to do this generally: why would we believe that each measure is equally valid? Some rankings might arbitrarily choose weights, or take a separate survey to find weights (equally problematic!). You could average what fraction of measures achieve some hurdle.

One possible way around this problem is to just ask for people's rankings (let them figure out what weights to use in their own utility functions) and report some aggregation. However here again there is no single method that is guaranteed to give correct aggregations. Some surveys ask people to rank units from 1-20, then add the rankings and the unit with the lowest number wins. But what if some people rank number 1 as far ahead of all of their competitors, while others see the top 3 as tight together? This distance information is omitted from the rankings. Some surveys might, instead, give 10 points for a #1 ranking, 8 points for #2, and so on – but again this presupposes some distance between the ranks.

This is not to say that ranking is hopeless or never informative, just that there is no single path that will inerrantly give the correct result. Working through the rankings, an analyst might determine that a broad swathe of weights upon the various measures would all give similar

rankings to certain outliers. It would be useful to know that a particular unit is almost always ranked near the top while some other one is nearly always at the bottom.

Examples:

Education: College rankings try to combine student/faculty ratios, measures of selectivity, SAT scores, GPA; some add in numbers of bars near campus or the prestige of journals in which faculty publish. What is best? School teachers face efforts to rank them, by student test score improvements as well as other factors; schools and districts are ranked by a variety of measures.

Sports might seem to have it relatively easy since there is a single ranking given by pre-arranged rules, but still fans can argue: a team has a good offense because they scored a lot (even though some other team won more games); some players are better on defense but worse on offense. Sports Illustrated tried to rank the 100 all-time best sports stars, somehow comparing baseball player Babe Ruth with the race horse Secretariat! Most magazines know that rankings drive sales and give buzz. (Linkbait)

Food nutrition trades off calories, fat content, fiber, vitamin and mineral content; who is to say whether kale or blueberries are healthier? Aren't interaction effects important? Someone trying to lose weight would make a very different ranking than someone training for a marathon.

Sustainability or "green" rankings are difficult: there are so many trade-offs! If we care about global warming then we look at CO_2 emissions, but what about other pollutants? Is nuclear power better than natural gas? Ethical consumption might also consider the material conditions of workers (fair-trade coffee or no-sweatshop clothing) or other considerations.

Politics: which political party is better for the economy? Could measure stock returns or unemployment rate or GDP growth or hundreds of others. Average wage or median earnings (household or individual)? Each set of measures could give different results. You can try this yourself, get some data from FRED (http://research.stlouisfed.org/fred2/) and go wild.

Other Ignorant Beliefs

While I'm working to extirpate popular heresies, let me address another one, which is particularly common when the Olympics roll around: the extraordinary belief that outliers can give useful information about the average value. We hear these judgments all of the time: some country wins an unusual number of Olympic medals, thus the entire population of the country must be unusually skilled at this task. Or some gender/race/ethnicity is overrepresented in a certain profession thus that gender/race/ethnicity is more skilled on average. Or a school has a large number of winners of national competitions, thus the average is higher.

Statistically speaking, the extreme values of a distribution depend on many parameters such as the higher moments. If I have two distributions with the exact same mean, standard deviation, and skewness, but different values of kurtosis, then one distribution will systematically have higher extremes (by definition of kurtosis).