Heteroskedasticity-Robust and Clustered Standard Errors in R

Recall that if heteroskedasticity is present in our data sample, the OLS estimator will still be unbiased and consistent, but it will **not** be efficient. Specifically, estimated standard errors will be biased, a problem we cannot solve with a larger sample size.

To correct for this bias, it may make sense to adjust your estimated standard errors. Two popular ways to tackle this are to use:

1. "Robust" standard errors (a.k.a. White's Standard Errors, Huber–White standard errors, Eicker–White or Eicker–Huber–White)
2. Clustered Standard Errors

In practice, heteroskedasticity-robust and clustered standard errors are usually larger than standard errors from regular OLS — however, this is not always the case. For further detail on when robust standard errors are smaller than OLS standard errors, see Jorn-Steffen Pische's response on Mostly Harmless Econometrics' Q&A blog.

"Robust" standard errors

The following example will use the `CRIME3.dta`

Because one of this blog's main goals is to translate STATA results in R, first we will look at the `robust` command in STATA. For backup on the calculation of heteroskedasticity-robust standard errors, see the following link: http://www.stata.com/support/faqs/stat/cluster.html. The formulation is as follows:

$$Variance_{robust} = \left(\frac{N}{N-K}\right)(X'X)^{-1}\sum_{i=1}^{N}\{X_i X_i' \hat{\varepsilon}_i^2\}(X'X)^{-1}$$

where $N =$ number of observations, and $K =$ the number of regressors (including the intercept). This returns a Variance-covariance (VCV) matrix where the diagonal elements are the estimated heteroskedasticity-robust coefficient variances — the ones of interest. Estimated coefficient standard errors are the square root of these diagonal elements.

STATA:

```
reg cmrdrte cexec cunem if year==93, robust
```

R:

The following bit of code was written by Dr. Ott Toomet (mentioned in the Dataninja blog). I added a degrees of freedom adjustment so that the results mirror STATA's `robust` command results.

```
## Heteroskedasticity-robust standard error calculation.
summaryw <- function(model) {
s <- summary(model)
X <- model.matrix(model)
u2 <- residuals(model)^2
XDX <- 0

## Here one needs to calculate X'DX. But due to the fa
## D is huge (NxN), it is better to do it with a cycle
for(i in 1:nrow(X)) {
XDX <- XDX + u2[i]*X[i,]%*%t(X[i
}

# inverse(X'X)
XX1 <- solve(t(X)%*%X)

# Variance calculation (Bread x
varcovar <- XX1 %*% XDX %*% XX1

# degrees of freedom adjustment
dfc <- sqrt(nrow(X))/sqrt(nrow()

# Standard errors of the coeffic
# square roots of the diagonal e
stdh <- dfc*sqrt(diag(varcovar))

t <- model$coefficients/stdh
p <- 2*pnorm(-abs(t))
results <- cbind(model$coefficients, stdh, t, p)
dimnames(results) <- dimnames(s$coefficients)
results
}
```


To use the function written above, simply replace `summary()` with `summaryw()` to look at your regression results — like this:

```
require(foreign)
mrdr = read.dta(file="/Users/kevingoulding/data/MURDER.dta")

# run regression
reg4 = lm(cmrdrte ~ cexec + cunem, data = subset(mrdr,year == 93))

# see heteroskedasticity-robust standard errors
summaryw(reg4)
```

Search this blog
Search...

Contributors

Kevin Goulding

Categories
Econometrics
Econometrics with R
Numpy
Python
R tips & tricks
Surviving Graduate Econometrics with R
TikZ for Economists
Visualizing Data with R
White Papers

Twitter feed
RT @gappy3000: This post, apparently about #julialang and #pydata, explains why #rstats has become the standard of data analysis http:// ... 3 years ago
RT @justinwolfers: "if prediction markets are really as valuable as economists think, then..more experimentation could prove worthwhile. ... 3 years ago
RT @vsbuffalo: For me the biggest victory is for statistics and empiricism. Go Nate Silver and @fivethirtyeight for a brilliant forecast ... 3 years ago
Follow @baha_kev

Tag Cloud
cluster-robust Econometrics heteroskedasticity LaTeX Numpy Parallel Computing plots Python R STATA tex TikZ

These results should match the STATA output exactly.

Heteroskedasticity-robust LM Test

It may also be important to calculate heteroskedasticity-robust restrictions on your model (e.g. an F-test).

Clustered Standard Errors

Let's say that you want to relax your homoskedasticity assumption, and account for the fact that there might be a bunch of covariance structures that vary by a certain characteristic – a "cluster" – but are homoskedastic within each cluster. Similar to heteroskedasticity-robust standard errors, you want to allow more flexibility in your variance-covariance (VCV) matrix. The result is clustered standard errors, a.k.a. cluster-robust.

STATA:

```
use wr-nevermar.dta
reg nevermar impdum, cluster(state)
```

R:

In R, you first must run a function here called `cl()` written by Mahmood Ara in Stockholm University – the backup can be found here.

```
cl   <- function(dat,fm, cluster){
            attach(dat, warn.conflicts = F)
            library(sandwich)
            M <- length(unique(cluster))
            N <- length(cluster)
            K <- fm$rank
            dfc <- (M/(M-1))*((N-1)/(N-K))
            uj  <- apply(estfun(fm),2, function(x) tapply(x, cluster, sum));
            vcovCL <- dfc*sandwich(fm, meat=crossprod(uj)/N)
            coeftest(fm, vcovCL) }
```

After running the code above, you can run your regression with clustered standard errors as follows:

```
nmar = read.dta("http://www.montana.edu/econ/cstoddard/562/wr-nevermar.dta")

# Run a plain linear regression
regt = lm(nevermar ~ impdum, data = nmar)

# apply the 'cl' function by choosing a variable to cluster on.
# here, we are clustering on state.
cl(nmar, regt, nmar$state)
```


Share this:

Share

Like

Be the first to like this.

Related

Clustered Standard Errors in R
In "Econometrics with R"

Calculate an OLS regression using matrices in Python using Numpy
In "Econometrics"

Calculate OLS regression manually using matrix algebra in R
In "Econometrics with R"

Posted on May 28, 2011 at 7:43 am in Econometrics with R | RSS feed | Reply | Trackback URL

Tags: R, STATA

18 Comments to "Heteroskedasticity-Robust and Clustered Standard Errors in R"

mr science
June 27, 2011 at 3:41 pm

can I use 3 for pi ???

Reply

Kevin Goulding
June 29, 2011 at 4:05 pm

@mr science

You may use 3 for pi, but why would you when R has the value of pi stored inside it already – thru 14 decimal places. Just type the word `pi` in R, hit [enter] — and you're off and running!

Reply

econ
August 19, 2011 at 7:05 pm

Kevin, what would be the reason why heteroskadisticy-robust and clustered errors could be smaller than regular OLS errors?

Reply

Kevin Goulding
August 22, 2011 at 10:24 am

Hi econ – Robust standard errors have the potential to be smaller than OLS standard errors if outlier observations (far from the sample mean) have a low variance; generating an upward bias in OLS standard errors. For a more detailed discussion of this phenomenon, see Jorn-Steffen Pische's response on Mostly Harmless Econometrics' Q&A blog. I've added a similar link to the post above. Hope this helps. -Kevin

Reply

econ
August 22, 2011 at 9:22 pm

Thanks, Kevin. Help much appreciated.

Sohail Farooq
October 10, 2011 at 5:25 am

Dear Kevin, I have a problem of similar nature.
let suppose I run the same model in the following way.
1) xtreg Y X1 X2 X3, fe robust cluster(country)
2) xtreg Y X1 X2 X3, fe robust
3) xtreg Y X1 X2 X3, fe cluster(country)
4) xtreg Y X1 X2 X3, fe

In first 3 situations the results are same. but in the last situation (4th, i.e. without robust and cluster at country level) for X3 the results become significant and the Standard errors for all of the variables got lower by almost 60%. so can you please guide me that what's the reason for such strange behaviour in my results? your help is highly appreciable.

Sohail farooq

Reply

Kevin Goulding
February 27, 2012 at 2:35 pm

Sohail, your results indicate that much of the variation you are capturing (to identify your coefficients on X1 X2 X3) in regression (4) is "extra-cluster variation" (one cluster versus another) and likely is overstating the accuracy of your coefficient estimates due to heteroskedasticity across clusters. In short, it appears your case is a prime example of when clustering is required for efficient estimation. I would perform some analytics looking at the heteroskedasticity of your sample. HTH. -Kevin

Reply

Iva
February 27, 2012 at 2:12 pm

Hi Kevin,

This is somewhat related to the standard errors thread above. I am running an OLS regression with a dummy variable, control variable X1, interaction X1*DUMMY, and other controls. When I don't include X1 and X1*DUMMY, DUMMY is significant. When I include DUMMY, X1 and don't include the interaction term, both DUMMY and X1 are significant. When I include DUMMY, X1 and X1*DUMMY, X1 remains significant but DUMMY and X1*DUMMY become insignificant. This seems quite odd to me. Have you encountered it before? Thanks for your help and the helpful threads.

Iva

Reply

Kevin Goulding
February 27, 2012 at 2:24 pm

Iva, the interaction term X1*Dummy is highly multicollinear with both X1 & the Dummy itself. In fact, each element of X1*Dummy is equal to an element of X1 or Dummy (e.g. = 0 or = X1). I would suggest eliminating the interaction term as it is likely not relevant. Hope that helps. -Kevin

Reply

Iva
February 27, 2012 at 2:49 pm

Thanks for the quick reply, Kevin. My only concern is that if both the DUMMY and the interaction term become insignificant when included in the model, then my results may be subject to the criticism that the effect of DUMMY on the outcome variable is altogether insignificant (which however contradicts the significant coefficient of DUMMY when both only DUMMY and X1 are included and the interaction term is excluded). Do you think that such a criticism is unjustified?

Iva

Kevin Goulding

February 27, 2012 at 4:18 pm

No, I do not think it's justified. Interaction terms should only be included if there is some theoretical basis to do so. It doesn't seem like you have a reason to include the interaction term at all.

Brian Quistorff
August 8, 2012 at 8:56 am

Note, that I think this function requires "clean" data (no missing values for the variables of interest) otherwise you get an error.

Reply

mika
October 26, 2012 at 5:49 am

This code was very helpful for me as almost nobody at my school uses R and everyone uses STATA. It worked great. Thank you!

Reply

Mary
April 9, 2013 at 7:49 am

How do I get SER and R-squared values that are normally included in the summary() function?

Reply

Wim Delva
May 17, 2013 at 1:00 pm

Hi Kevin,

Thanks for sharing this code.
Unfortunately, when I try to run it, I get the following error message:
Error in tapply(x, cluster, sum) : arguments must have same length

Could it be that the code only works if there are no missing values (NA) in the variables?
If so, could you propose a modified version that makes sure the size of the variables in dat, fm and cluster have the same length?

Many thanks,

Wim

Reply

Nova Feinberg
December 4, 2013 at 11:15 pm

Oh my goodness! an incredible article dude.
Thanks Nonetheless I am experiencing issue with ur rss .

Don't know why Unable to subscribe to it. Is there anybody getting an identical rss drawback? Anyone who is aware of kindly respond.
Thnkx

Reply

Alicia
August 18, 2014 at 2:48 pm

I'm not sure where you're getting your info, but great topic. I needs to spend some time learning much more or understanding more. Thanks for wonderful info I was looking for this information for my mission.

Reply

Miguel A.
February 13, 2015 at 3:33 am

Hi, Kevin. Although this post is a bit old, I would like to ask something related to it. I have a panel-data sample which is not too large (1,973 observations). The unit of analysis is x (credit cards), which is grouped by y (say, individuals owning different credit cards). I cannot used fixed effects because I have important dummy variables. And random effects is inadequate. Therefore, I am using OLS. To control clustering in y, I have introduced a dummy variable for each y. My question is whether this is fine (instead of using (in Stata)). Thanks in advance.

Reply

Leave a Reply

Enter your comment here...

Tags

cluster-robust econometrics heteroskedasticity latex numpy parallel computing plots python r stata tex tikz

Calendar

May 2011

M	T	W	T	F	S	S
						1
2	3	4	5	6	7	8

Archives

October 2012
February 2012
July 2011

Blogroll

Documentation
Plugins
Suggest Ideas

9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29
30	31					

Jun »

June 2011
May 2011

Support Forum
Themes
WordPress Blog
WordPress Planet