

## the Tarzan

[R] + applied economics.

[About](#)
[ECNS 561](#)
[Nuts'n Bolts](#)
[Resources](#)

« Heteroskedasticity-Robust and Clustered Standard Errors in R | TikZ diagrams for economists: A price ceiling »

## Clustered Standard Errors in R

So, you want to calculate clustered standard errors in R (a.k.a. cluster-robust, huber-white, White's) for the estimated coefficients of your OLS regression? This post shows how to do this in both Stata and R:

## Overview

Let's say that you want to relax the Gauss-Markov homoskedasticity assumption, and account for the fact that there may be several different covariance structures within your data sample that vary by a certain characteristic – a “cluster” – but are homoskedastic within each cluster.

*For example, say you have a panel data set with a bunch of different test scores from different schools around the country. You may want to cluster your sample by state, by school district, or even by town. Economic theory and intuition will guide you in this decision.*

So, similar to heteroskedasticity-robust standard errors, you want to allow more flexibility in your variance-covariance (VCV) matrix (Recall that the diagonal elements of the VCV matrix are the squared standard errors of your estimated coefficients). The way to accomplish this is by using clustered standard errors. The formulation is as follows:

$$\text{Variance}_{cluster} = \left( \frac{M}{M-1} \frac{N-1}{N-K} \right) (X'X)^{-1} \sum_{j=1}^M \{X_M \hat{\epsilon}_M \hat{\epsilon}_M' X_M\} (X'X)^{-1}$$

where  $M$  = number of unique clusters (e.g. number of school districts)  $N$  = number of observations, and  $K$  = the number of regressors (including the intercept). (See pages 312-313 of Angrist and Pischke's **Mostly Harmless Econometrics** (Princeton University Press, 2009) for a better explanation of the summation notation.) This returns a Variance-covariance (VCV) matrix where the diagonal elements are the estimated cluster-robust coefficient variances — the ones of interest. Estimated coefficient standard errors are the square root of these diagonal elements.

Stata makes it very easy to calculate, by simply adding `, cluster(state)` to the end of your regression command.

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](#) and [here](#).

```
cl <- function(dat, fm, cluster){
  require(sandwich, quietly = TRUE)
  require(lmtest, quietly = TRUE)
  M <- length(unique(cluster))
  N <- length(cluster)
  K <- fm$rank
  dfc <- (M/(M-1))*((N-1)/(N-K))
  ujl <- apply(estfun(fm), 2, function(u)
    vcovCL <- dfc*sandwich(fm, meat=cru
    coeftest(fm, vcovCL) }
```

After running the code above, you can run

```
require(foreign)
nmar = read.dta("http://www.montana.edu/econ/cst

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

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

*Thoughts, comments, ideas? Let me know; I'm always appreciative of feedback. You can contact me via e-mail at [kevingoulding@gmail.com](mailto:kevingoulding@gmail.com).*

## Search this blog

## 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

## Twitterfeed

RT @gappy3000: This post, apparently about #julialang and #pydata, explains why #rstats has become the standard of data analysis [http:// ... 3 years ago](#)

RT @justinwolffers: "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

[Follow](#)

## Follow “the Tarzan”

Get every new post delivered to your Inbox.

Join 78 other followers

Enter your email address

Sign me up

Build a website with WordPress.com

Clustered standard errors as follows:

About these ads

Share this:



Be the first to like this.

Related

[Heteroskedasticity-Robust and 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 June 11, 2011 at 5:17 pm in [Econometrics with R](#) | [RSS feed](#) | [Reply](#) | [Trackback URL](#)

Tags: [cluster-robust](#), [heteroskedasticity](#), [R](#), [STATA](#)


29 Responses to “Clustered Standard Errors in R”



*Riccardo Klinger*  
September 12, 2011 at 2:59 pm

Dear Kevin, thank you for your blog (especially this entry). one question: do you know a procedure to use multiple input variables and not only one like in the example given by Mahmood Ara?  
my goal is to use a formula like  $y \sim x_1 + x_2 + x_1 * x_2 \dots$  any ideas will be highly motivating....


Reply



*Kevin Goulding*  
September 12, 2011 at 3:48 pm

Hi Riccardo – the procedure outlined above should work fine for a regression with multiple independent variables. For example, if instead of “`regt = lm(nevermar ~ impdum, data = nmar)`” it was “`regt = lm(nevermar ~ impdum + x2 + x3, data = nmar)`”, the procedure to output clustered standard errors would be exactly the same: “`cl(nmar, regt, nmar$state)`” . If your question is how to cluster around more than one variable (e.g. multiple clusters), then I suggest you look at the function “`mclx()`” that is also in the Mahmoon Ara documentation. Hope this helps. -Kevin


Reply



*Riccardo Klinger*  
September 13, 2011 at 12:00 am

Hi Kevin, thanks for your very quick reply. I thought so as well but the sandwich estimator function throwed me following error: Fehler in bread. %\*% meat. : nicht passende Argumente (Failure in bread. %\*% meat, arguments not matching). this was given me the idea about to use this function only in a one-variable-regression way. examining the results of the used bread and meat functions just showed me a NA problem for the t-values, P values and so on in the linematching (lm) results for the multiple variable model...  
i hope it will work when the problem is solved.

Reply




*Riccardo Klinger*  
September 13, 2011 at 1:56 am

dear kevin,

due to some singularity in my data i was receiving this error. now i have solved this and everything is working fine. Yet: is there any possibility to get an R-squared value?

all the best, riccardo

Reply



*STS*  
December 14, 2011 at 9:49 am

Hi,

I have the same problem (singularities within the data) and wonder how you solved it. Is there any other

way than dropping some/those variables?

best, sjard.

Reply



**Kevin Goulding**

December 14, 2011 at 4:43 pm

Hi STS — What is causing the singularities in your data? This could be due to overspecification (e.g., too many regressors for too little data), or a “dummy variable trap”, or all the variation in one regressor is captured in another. You will have to look closely at your data to identify what is causing the singularity. Another solution could be to add more data (increase sample size). Without understanding the nature of the singularity, it is difficult for me to help you. HTH, Kevin



**ricckli**

September 13, 2011 at 4:44 am

in the original script by Arai he is dividing the results of sandwich by dfcw. have you missed that? best regards, riccardo

Reply



**ricckli**

September 13, 2011 at 7:14 am

one more thing: you`re not using “dat” in your function.  
the original function by Arai uses a correction for the degrees of freedom, which you are not using. so the function only depends on clusters and the model....

Reply



**riccardo**

December 14, 2011 at 4:26 pm

Just use some stepwise function to create a good linear model. Than there will be no singularities. And one more thing : excuse my question about the r-squared. Of course it is given by lm() itself, isn't it?

Reply



**Sach**

August 6, 2012 at 11:18 am

Hi Kevin,

I was just wondering about a little thing on clustered standard errors. i was hoping to hear an opinion from you on this.

IF we cluster the standard erros, do we have to cluster them at the unit of observations (say, villages). Or, if 5 to 6 villages are comprised in a county, would it also be possible to cluster the according to the number of counties? The latter makes more sense to me since we would allow for clustering within the cluster group (region) which seems reasonable to me (imagine the county decides to build a new railway station, all villages would be affected). So, provides that we think that the no of clusters is high enough, could we also cluster at a unit that is NOT the unit of the observations in one's dataset?

I am really curious to read your views on that!

Reply



**Kevin Goulding**

August 7, 2012 at 11:57 am

Hi Sach, great question. The short answer is yes. We can cluster at any “grouping level” we want, and can cluster on multiple groupings. Nearly always it makes the most sense to group at a level that is not at the unit-of-observation level. For example, if you have individual test scores for students across the country, you might want to cluster the standard errors by state, school district, region, or some combination thereof. To take your question a bit further, the only time you might want to cluster standard errors at the observation level is if you have a time series component in your data (i.e. repeated measures). Although if this were the case, you might want to investigate exactly how observations for the same individual vary over time — are they autocorrelated, for example. This might lead you to an approach that explicitly models autocorrelation at the individual level, and while it is related to clustering standard errors, is not the same. To conclude, it is highly unlikely that you will want to cluster standard errors at the unit-of-observation level. Hope this helps.

Reply



**Max**

October 14, 2012 at 12:35 pm

Hey, Kevin!

This is great post. I do have a little problem applying the procedure, though. Eventhough I included the packages zoo and sandwich, when I run c1 I am told that the function “coefest” couldn't be found. Perhaps this is a stupid questions which has nothing to do with the thread but it would be very nice, if you answered anyway. Thx – me

Reply



**Kevin Goulding**

October 14, 2012 at 2:37 pm

Hey; enter ``require(lmtest)`` and see if that does it. If you haven't already installed that package you will need to enter ``install.packages("lmtest", dep=TRUE)``. HTH, Kevin

Reply



**exl022**

October 15, 2012 at 7:50 pm

Hey, Kevin, great post. Thanks for doing this. I have seen similar posts on robust and clustered SEs, and there are often annoying small differences between results from that code that stata. This was pretty much spot on. I had a quick question about clustered SEs and other models, say negative binomial and

poisson. Do you know if this code works for non `lm()` models as well?

Reply



**Kevin Goulding**

October 16, 2012 at 10:29 am

I do not think this code will work with other non-`lm()` models; however, you are welcome to give it a shot.

Reply



**childofforest**

June 14, 2013 at 5:12 pm

Hi, I get the error message `Error in tapply(x, cluster, sum) : arguments must have same length` is there anyway to get around this?

Reply



**cbgibson**

November 5, 2013 at 4:58 pm

Yes — I'm getting the same error. Any luck finding a solution?

Reply



**Cari Bogulski (@caribogulski)**

June 13, 2014 at 1:50 pm

I was getting the same error, and realized this is due to NA values in one of the variables in the original `lm()` I ran, which made the length of the vector I was trying to cluster on (which had no NAs) a different length. After I limited the data in the `lm()` to only those rows that had no NAs in any of the variables I was inputting into the `lm`, the `cl()` function worked...although my Stata output is still not quite matching the R output yet...



**Dustin**

October 17, 2014 at 4:39 pm

I'm getting the "Error in `tapply(x, cluster, sum) : arguments must have same length`" I tried the code Aks suggests to deal with NA's and got "Error in `data$flpstate : object of type 'closure' is not subsettable`" Any suggestions would be greatly appreciated.



**cjohnson**

October 30, 2013 at 6:36 am

Kevin, I am having trouble loading the sandwich package. I receive the following error message: `Error in loadNamespace(i, c(lib.loc, .libPaths()), versionCheck = vI[[i]]) : there is no package called 'zoo'` Any suggestions?

Reply



**cjohnson**

October 30, 2013 at 6:43 am

It looks like the "zoo" package is not available for R 3.0.2. I've had trouble installing other packages as well, so I'm now thinking I should install an older version of R. Which version are you using?

Reply



**Kevin Goulding**

October 30, 2013 at 7:51 am

I believe the version of R I used for this was v2.14 or thereabouts.



**Aks**

April 12, 2014 at 11:51 pm

Hi Kevin,

The function does not estimate when the residuals have NA in them.

so when it computes `estfun(fm)`, it has missing rows and hence the number of rows in `estfun(fm)` is less than the number of rows in data.

this causes error in the next step of computing `uj` because rows of data are not the same as rows of `estfun`.

I tried amending the code as follows

```
fm <- lm(y~x, wt=wt, na.action=na.exclude)
```

```
uj <- apply(estfun(fm), 2, function(x) tapply(x, data$flpstate, sum, na.rm=TRUE))
```

In this modified code, at least the code runs and computes standard errors, but the clustered se is not the same as obtained in stata.

Reply



**Sam**

April 27, 2014 at 6:37 am

Hi Kevin,

I'm having an issue using your function: <http://stackoverflow.com/questions/23313907/clustering-standard-errors>

If you have any suggestions, would be very grateful!

Best,

Sam

Reply



**Alex**

May 2, 2014 at 10:14 am

Thanks for this! Does the F-statistic change? If so, how do I get that? I am assuming the R-squared stays the same?

Reply



Can Celebi  
February 22, 2015 at 1:35 pm

Hey Kevin,  
Why don't we use the bread part of the sandwich function i.e  $(X'X)^{-1}$  parts around the meat in this code?  
The original paper states that under some assumed conditions  $S(\text{teta})$  i.e the sandwich function reduces to the bread function which mean that we don't have the  $(X'X)^{-1}$  part you have written on the equation. Is this correct?

Reply



MichaelChirico  
October 4, 2015 at 4:54 pm

Both backup links appear dead. I believe this is the referred overview:

[http://www.ne.su.se/polopoly\\_fs/1.216115.1426234213!/menu/standard/file/clustering1.pdf](http://www.ne.su.se/polopoly_fs/1.216115.1426234213!/menu/standard/file/clustering1.pdf)

Reply

Trackbacks

*cluster robust standard errors in R « R in finance*  
September 22, 2011 at 1:48 pm

*Fama-MacBeth and Cluster-Robust (by Firm and Time) Standard Errors in R « landroni*  
June 2, 2012 at 2:20 pm

Leave a Reply

Enter your comment here...

Tags

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

Calendar

June 2011						
M	T	W	T	F	S	S
		1	2	3	4	5
6	7	8	9	10	11	12
13	14	15	16	17	18	19
20	21	22	23	24	25	26
27	28	29	30			
« May						Jul »

Archives

October 2012  
February 2012  
July 2011  
June 2011  
May 2011

Blogroll

Documentation  
Plugins  
Suggest Ideas  
Support Forum  
Themes  
WordPress Blog  
WordPress Planet

