Cluster Robust Testing using R and SPSS Add-on

Francis L. Huang, PhD & Xintong Li, PhD

2020.09.22 / updated 2021.05.17

The following is an example testing the comparability of the clustSE function in R and the SPSS add-on macro (available from http://faculty.missouri.edu/huangf/CR2/) with results from other R packages.

## 1. Load in dataset

The auto dataset can be downloaded from the Stata website. For the current example,

dat <- rio::import("http://www.stata-press.com/data/r13/auto7.dta")  
dat2 <- dat[, c('mpg', 'price', 'manufacturer', 'rep78', 'headroom', 'foreign')] #show only vars of interest  
dim(dat2)

## [1] 74 6

head(dat2)

## mpg price manufacturer rep78 headroom foreign  
## 1 17 4749 AMC 3 3.0 0  
## 2 22 4099 AMC 3 2.5 0  
## 3 22 3799 AMC NA 3.0 0  
## 4 17 9690 Audi 5 3.0 1  
## 5 23 6295 Audi 3 2.5 1  
## 6 25 9735 BMW 4 2.5 1

This is a small dataset (n = 74) with manufacturer as the cluster variable. We will only focus on a few variables predicting price using mpg, rep78 (repair record 1978) which has 5 missing values, headroom, and foreign. foreign is a cluster level dummy coded variable.

The dataset is useful for testing as it has:

1. a few clusters (23)
2. has single cases in a cluster (see below): you shouldn’t have these in CRTs anyway but may have these when analyzing secondary datasets
3. has missing values on rep78
4. a large ICC (ICC = .78) (not the usual case though)

The functions below will work even if the cluster variable is not sorted.

length(table(dat2$manufacturer))

## [1] 23

table(dat2$manufacturer)

##   
## AMC Audi BMW Buick Cad. Chev. Datsun Dodge Fiat Ford   
## 3 2 1 7 3 6 4 4 1 2   
## Honda Linc. Mazda Merc. Olds Peugeot Plym. Pont. Renault Subaru   
## 2 3 1 6 7 1 5 6 1 1   
## Toyota Volvo VW   
## 3 1 4

## 2. Run the regression model

For the current example, we are not so interested in the coefficents themselves but the comparability of the standard errors using different functions.

ols1 <- lm(price ~ mpg + rep78 + headroom + foreign, data = dat)  
summary(ols1)

##   
## Call:  
## lm(formula = price ~ mpg + rep78 + headroom + foreign, data = dat)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2999.5 -1546.5 -681.5 707.4 10139.0   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 11357.44 2193.15 5.179 2.42e-06 \*\*\*  
## mpg -302.47 63.79 -4.741 1.23e-05 \*\*\*  
## rep78 458.69 400.34 1.146 0.256   
## headroom -204.91 408.42 -0.502 0.618   
## foreign 921.85 894.24 1.031 0.306   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2566 on 64 degrees of freedom  
## (5 observations deleted due to missingness)  
## Multiple R-squared: 0.2696, Adjusted R-squared: 0.2239   
## F-statistic: 5.905 on 4 and 64 DF, p-value: 0.0004147

nobs(ols1) #with up with 74 - 5 = 69 observations

## [1] 69

## 

## 3. Use clustSE function

Load in the clustSE function from http://faculty.missouri.edu/huangf/CR2/:

source("http://faculty.missouri.edu/huangf/CR2/02\_CR2v1.R")

To use the function, specify:

res <- clustSE(ols1, 'manufacturer') #manufacturer is the cluster  
res #shows all results

## estimate se.unadj CR0 CR1 CR2 tCR2 dfn  
## (Intercept) 11357.4420 2193.1540 2380.1625 2511.1506 2601.8606 4.3651 64  
## mpg -302.4653 63.7914 82.6193 87.1661 91.3682 -3.3104 64  
## rep78 458.6877 400.3360 342.6625 361.5204 380.7868 1.2046 64  
## headroom -204.9053 408.4202 253.2491 267.1863 284.1111 -0.7212 64  
## foreign 921.8461 894.2364 694.6754 732.9056 764.4529 1.2059 20  
## dfBM pv.unadj CR0pv CR0pv.n CR1pv CR1pv.n CR2pv.n CR2pv  
## (Intercept) 9.76 0.0000 0.0008 0.0000 0.0012 0.0000 0.0000 0.0015  
## mpg 10.59 0.0000 0.0040 0.0005 0.0055 0.0009 0.0015 0.0073  
## rep78 9.43 0.2562 0.2121 0.1854 0.2350 0.2091 0.2328 0.2577  
## headroom 9.12 0.6176 0.4391 0.4215 0.4625 0.4460 0.4734 0.4889  
## foreign 12.29 0.3065 0.2086 0.1995 0.2318 0.2230 0.2419 0.2506

res[,c('estimate', 'CR2', 'tCR2', 'dfBM', 'CR2pv')] #only selected columns

## estimate CR2 tCR2 dfBM CR2pv  
## (Intercept) 11357.4420 2601.8606 4.3651 9.76 0.0015  
## mpg -302.4653 91.3682 -3.3104 10.59 0.0073  
## rep78 458.6877 380.7868 1.2046 9.43 0.2577  
## headroom -204.9053 284.1111 -0.7212 9.12 0.4889  
## foreign 921.8461 764.4529 1.2059 12.29 0.2506

## 4. Compare results with other R functions

For comparability, we can use the vcovCL function in the sandwich (Zeileis, 2006) package.

library(sandwich) #for CR  
library(lmtest) #to get the inferential statistics  
vc1 <- vcovCL(ols1, type = 'HC2', cluster = dat$manufacturer)  
#sqrt(diag(vc1)) #CRSEs  
coeftest(ols1, vc1)

##   
## t test of coefficients:  
##   
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 11357.442 2601.861 4.3651 4.733e-05 \*\*\*  
## mpg -302.465 91.368 -3.3104 0.001534 \*\*   
## rep78 458.688 380.787 1.2046 0.232802   
## headroom -204.905 284.111 -0.7212 0.473403   
## foreign 921.846 764.453 1.2059 0.232299   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

The CR2 standard errors are the same (in the vcovCL function, specify both the cluster option and type = 'HC2'). The *p* values though are not the same as a result of different degrees of freedom used.

We now use the newer clubSandwich (Pustejovsky, 2018) package:

library(clubSandwich)  
vc2 <- vcovCR(ols1, cluster = dat$manufacturer, type = 'CR2')  
coef\_test(ols1, vc2)

## Coef. Estimate SE t-stat d.f. p-val (Satt) Sig.  
## 1 (Intercept) 11357 2601.9 4.365 9.76 0.00149 \*\*  
## 2 mpg -302 91.4 -3.310 10.59 0.00730 \*\*  
## 3 rep78 459 380.8 1.205 9.43 0.25774   
## 4 headroom -205 284.1 -0.721 9.12 0.48886   
## 5 foreign 922 764.5 1.206 12.29 0.25057

The results are the same for both the standard errors and the degrees of freedom (differences due to rounding).

## 5. Using the SPSS Add-on

Using the graphical, menu-driven SPSS add-on with the CR2 estimator and the Bell and McCaffrey (2002) dof adjustment results in:

Model Specification  
 CR Type CR2  
 DF Type BM  
  
 Cluster-Robust Regression Coefficient Tests  
 Estimate CR\_S.E. t\_value d.f. Sig.  
 Constant 11357.4420 2601.8606 4.3651 9.7645 .0015  
 mpg -302.4653 91.3682 -3.3104 10.5896 .0073  
 rep78 458.6877 380.7868 1.2046 9.4253 .2577  
 headroom -204.9053 284.1111 -.7212 9.1170 .4889  
 foreign 921.8461 764.4529 1.2059 12.2867 .2506

SPSS results are the same as previously shown using R.

**References**

Bell, R., & McCaffrey, D. (2002). Bias reduction in standard errors for linear regression with multi-stage samples. *Survey Methodology, 28,* 169–182.

Pustejovsky, J. (2018). clubSandwich: Cluster-robust (sandwich) variance estimators with small-sample corrections. https://CRAN.R-project.org/package=clubSandwich

Zeileis, A. (2006). Object-oriented computation of sandwich estimators. *Journal of Statistical Software, 16*(9), 1–16.

— END