SOSC 5340 Tutorial One

Standard Errors and Bootstrap

Yabin YIN
HKUST

Feb, 2021

Set working directory to the current directory

Remark: Need to save current R file before using *getActiveDocumentContext*

R Packages

R packages for robust and cluster standard errors:

- *sandwich*: https://cran.r-project.org/web/packages/sandwich
- *estimatr*: https://cran.r-project.org/web/packages/estimatr
- *clubSandwich*: https://cran.r-project.org/web/packages/clubSandwich
- Read the *reference manual* and *vignettes*.
- We will focus on the *sandwich* package. Please try other packages yourself.

R packages for output tables:

- *texreg*: https://cran.r-project.org/web/packages/texreg/texreg.pdf
- *stargazer*: https://cran.r-project.org/web/packages/stargazer/stargazer.pdf
- *starpolishr*: https://github.com/ChandlerLutz/starpolishr

Latex - *overleaf*: https://www.overleaf.com/ online platform - *Latex*: https://www.latex-project.org/get/

Empirical Example: *Aghion, Van Reenen, and Zingales (2013 AER)*

Aghion, Van Reenen, and Zingales (2013) studied the relationship between institutional ownership and innovation. We replicate column 1 of Table 1 of this paper (see page 283).

Robust Standard Errors

```
# require the packages
library(sandwich)
library(lmtest)
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
library(texreg)
```

```
## Version:  1.37.5
## Date:     2020-06-17
## Author:   Philip Leifeld (University of Essex)
##
## Consider submitting praise using the praise or praise_interactive functions.
## Please cite the JSS article in your publications -- see citation("texreg").
```

```r
library(stargazer)
```

```
##
## Please cite as:
```

```
##  Hlavac, Marek (2018). stargazer: Well-Formatted Regression and Summary Statistics Tables.
```

```
##  R package version 5.2.2. https://CRAN.R-project.org/package=stargazer
```

```r
library(starpolishr)
library(tinytex)

# load the data: from the "sandwich" package
data("InstInnovation")

# classic s.e.
lm_classic <- lm(log(cites+1)~institutions+log(I(capital/employment)+1)+log(sales+1)
                 +factor(industry)+factor(year), # industry dummies and time dummies
                 data = InstInnovation)
# robust s.e.
lm_r_sandwich <- coeftest(lm_classic, vcov. = vcovHC(lm_classic, type = "HC0"))

# show the results
stargazer(lm_classic, lm_classic,
          se = list(summary(lm_classic)$coefficients[,2],
                             lm_r_sandwich[,2]),
          column.labels = c("ln(Cites) classic", "ln(Cites) robust"),
          dep.var.labels.include = F,
          keep = c("institutions", 'capital/employment', 'sales'),
          covariate.labels = c("Share of institutions", "ln(K/L)", "ln(Sales)"),
          digits = 4, no.space=TRUE, column.sep.width = "1pt",
          add.lines = list(c('Industry FE', 'Y', 'Y'),
                           c('Year FE', 'Y', 'Y')),
          omit.stat = c("ser","f"),
          type = 'text')
```

```
##
## ========================================================
##                              Dependent variable:
##                       ----------------------------------
##                       ln(Cites) classic ln(Cites) robust
##                              (1)              (2)
## --------------------------------------------------------
## Share of institutions    0.0060***        0.0060***
##                          (0.0010)          (0.0011)
## ln(K/L)                  0.4304***        0.4304***
##                          (0.0391)          (0.0408)
## ln(Sales)                0.6123***        0.6123***
##                          (0.0138)          (0.0155)
## --------------------------------------------------------
## Industry FE                   Y                Y
## Year FE                       Y                Y
## Observations                6,208            6,208
## R2                         0.5753           0.5753
## Adjusted R2                0.5650           0.5650
## ========================================================
## Note:                        *p<0.1; **p<0.05; ***p<0.01
```

```
# alternative way: screenreg from texreg
screenreg(list(lm_classic, lm_classic),
          custom.model.names = c("ln(Cites) classic", "ln(Cites) robust"),
          custom.coef.names = c("Share of institutions", "ln(K/L)", "ln(Sales)"),
          override.se = list(summary(lm_classic)$coefficients[,2],
                             lm_r_sandwich[,2]),
          override.pvalues = list(summary(lm_classic)$coefficients[,4],
                                  lm_r_sandwich[,4]),
          omit.coef = c("(Intercept)|(industry)|(company)|(year)"),
          custom.gof.rows = list('Industry FE' = c('Y', 'Y'),
                           'Year FE' = c('Y', 'Y')),
          stars = c(0.01, 0.05, 0.1),
          digits = 4)
```

```
##
## ==========================================================
##                        ln(Cites) classic  ln(Cites) robust
## ----------------------------------------------------------
## Share of institutions     0.0060 ***         0.0060 ***
##                          (0.0010)           (0.0011)
## ln(K/L)                   0.4304 ***         0.4304 ***
##                          (0.0391)           (0.0408)
## ln(Sales)                 0.6123 ***         0.6123 ***
##                          (0.0138)           (0.0155)
## ----------------------------------------------------------
## Industry FE               Y                  Y
## Year FE                   Y                  Y
## R^2                       0.5753             0.5753
## Adj. R^2                  0.5650             0.5650
## Num. obs.              6208               6208
## ==========================================================
## *** p < 0.01; ** p < 0.05; * p < 0.1
```

Remark: Using *stragazer* function, we can make publishable tables showing coefficients, robust standard errors and other information, we can change labels of dependent variables and independent variables, omit variables and statistics, and add customized information.

When using *screenreg* (also, *texreg* or *htmlreg*) function to show regression results with robust standard errors, we have to put the fitted model with classic s.e. in the *list()* then use the *override.se* and *override.pvalues* arguments to override the classic s.e. and p-value, otherwise we cannot get R^2 and *Num. obs.* and other statistics.

Cluster Standard Errors

Now suppose that we are concerned that

- *(i)* firms within the same four-digit industry might be correlated, so we have to adjust the standard errors by clustering at the four-digit industry level;

- *(ii)* there might be persistence over time for each firm, so we have to cluster at the firm level;
- *(iii)* there are macro common shocks to all firms in a given year, so we have to cluster at the year level;
- *(iv)* how about clustering at both firm and year levels?

```
# cluster: industry
lm_clu_industry <- coeftest(lm_classic,
                            vcov. = vcovCL(lm_classic,
                                           cluster = InstInnovation$industry,
                                           type = "HC0"))
# cluster: firm
lm_clu_firm <- coeftest(lm_classic,
                        vcov. = vcovCL(lm_classic,
                                       cluster = InstInnovation$company,
                                       type = "HC0"))
# cluster: year
lm_clu_year <- coeftest(lm_classic,
                        vcov. = vcovCL(lm_classic,
                                       cluster = InstInnovation$year,
                                       type = "HC0"))
# cluster: firm + year
lm_clu_twoway <- coeftest(lm_classic,
                          vcov. = vcovCL(lm_classic,
                                         cluster = InstInnovation[,c("company",
                                                                     "year")],
                                         type = "HC0"))
```

```
## Warning in sqrt(diag(se)): NaNs produced
```

```
# show the results
stargazer(lm_classic, lm_classic, lm_classic, lm_classic,
          se = list(lm_clu_industry[,2], lm_clu_firm[,2],
                    lm_clu_year[,2], lm_clu_twoway[,2]),
          column.labels = c("ln(Cites) ind", "ln(Cites) firm",
                                 "ln(Cites) year", "ln(Cites) fi+ye"),
          dep.var.labels.include = F,
          keep = c("institutions", 'capital/employment', 'sales'),
          covariate.labels = c("Share of institutions", "ln(K/L)", "ln(Sales)"),
          digits = 4, no.space=TRUE, column.sep.width = "1pt",
          add.lines = list(c('Industry FE', 'Y', 'Y', "Y", 'Y'),
                           c('Year FE', 'Y', 'Y', "Y", 'Y')),
          omit.stat = c("ser","f"),
          type = 'text')
```

```
##
## =================================================================================
##                                           Dependent variable:
##                       -----------------------------------------------------------
##                       ln(Cites) ind ln(Cites) firm ln(Cites) year ln(Cites) fi+ye
##                            (1)           (2)            (3)             (4)
## ---------------------------------------------------------------------------------
## Share of institutions   0.0060***     0.0060***      0.0060***       0.0060***
##                         (0.0020)       (0.0020)       (0.0015)       (0.0023)
## ln(K/L)                 0.4304***     0.4304***      0.4304***       0.4304***
##                         (0.1603)       (0.0854)       (0.0458)       (0.0879)
## ln(Sales)               0.6123***     0.6123***      0.6123***       0.6123***
##                         (0.0642)       (0.0326)       (0.0650)       (0.0711)
## ---------------------------------------------------------------------------------
## Industry FE                 Y             Y              Y               Y
## Year FE                     Y             Y              Y               Y
## Observations              6,208         6,208          6,208           6,208
## R2                       0.5753         0.5753         0.5753         0.5753
## Adjusted R2              0.5650         0.5650         0.5650         0.5650
## =================================================================================
## Note:                                                 *p<0.1; **p<0.05; ***p<0.01
```

Remark: The coefficients do not vary across columns because we only adjust the standard errors that are changing across columns.

Output latex table

You may copy paste the output latex into your .tex file or directly output that to a file.

```
##
## % Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harva
## % Date and time: Wed, Feb 17, 2021 - 23:48:31
## \begin{table}[!htbp] \centering
##   \caption{Robust Standard Error}
##   \label{}
## \begin{tabular}{@{\extracolsep{1pt}}lcc}
## \\[-1.8ex]\hline
## \hline \\[-1.8ex]
##  & \multicolumn{2}{c}{\textit{Dependent variable:}} \\
## \cline{2-3}
##  & ln(Cites) classic & ln(Cites) robust \\
## \\[-1.8ex] & (1) & (2)\\
## \hline \\[-1.8ex]
##  Share of institutions & 0.0060$^{***}$ & 0.0060$^{***}$ \\
##   & (0.0010) & (0.0011) \\
##   ln(K/L) & 0.4304$^{***}$ & 0.4304$^{***}$ \\
##   & (0.0391) & (0.0408) \\
##   ln(Sales) & 0.6123$^{***}$ & 0.6123$^{***}$ \\
##   & (0.0138) & (0.0155) \\
##  \hline \\[-1.8ex]
## Industry FE & Y & Y \\
## Year FE & Y & Y \\
## Observations & 6,208 & 6,208 \\
## R$^{2}$ & 0.5753 & 0.5753 \\
## Adjusted R$^{2}$ & 0.5650 & 0.5650 \\
## \hline
## \hline \\[-1.8ex]
## \textit{Note:}  & \multicolumn{2}{r}{$^{*}$p$<$0.1; $^{**}$p$<$0.05; $^{***}$p$<$0.01} \\
## \end{tabular}
## \end{table}
```

Remark:

- in Stargazer, using *star_tex_write* to output latex code as *.tex* file
- in texreg, add *file=XXX.tex* to output latex code as *.tex* file

Last, in Rmarkdown, you can also directly output the table as a LaTeXtable

(note the *results='asis'* option; that is how you output table as LaTeXdirectly)

Table 1: Cluster Standard Error

	ln(Cites) ind	ln(Cites) firm	ln(Cites) year	ln(Cites) fi+ye
Share of institutions	0.0060^{***}	0.0060^{***}	0.0060^{***}	0.0060^{***}
	(0.0020)	(0.0020)	(0.0015)	(0.0023)
ln(K/L)	0.4304^{***}	0.4304^{***}	0.4304^{***}	0.4304^{***}
	(0.1603)	(0.0854)	(0.0458)	(0.0879)
ln(Sales)	0.6123^{***}	0.6123^{***}	0.6123^{***}	0.6123^{***}
	(0.0642)	(0.0326)	(0.0650)	(0.0711)
Industry FE	Y	Y	Y	Y
Year FE	Y	Y	Y	Y
R^2	0.5753	0.5753	0.5753	0.5753
Adj. R^2	0.5650	0.5650	0.5650	0.5650
Num. obs.	6208	6208	6208	6208

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

Bootstrap

Now, let's use a small sample of Chinese 2005 mini-census data. Our mean focus is on the relation among income, education and gender.

```
library(readstata13)
# import dataset
census <- read.dta13('2005census.dta')
# fit the model
lm_census <- lm(log(income+1)~educ*factor(female)+factor(ifwork), data = census)
summary(lm_census)
```

```
##
## Call:
## lm(formula = log(income + 1) ~ educ * factor(female) + factor(ifwork),
##     data = census)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.0858 -0.2741  0.0223  0.3156  3.6311
##
## Coefficients:
##                      Estimate Std. Error  t value Pr(>|t|)
## (Intercept)          5.981227   0.049100  121.816  < 2e-16 ***
## educ                 0.011615   0.004024    2.887  0.00391 **
## factor(female)1     -0.297371   0.059585   -4.991 6.21e-07 ***
## factor(ifwork)2     -0.541033   0.097750   -5.535 3.27e-08 ***
## factor(ifwork)3     -5.939283   0.024949 -238.058  < 2e-16 ***
## educ:factor(female)1  0.003929   0.005009    0.784  0.43281
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8257 on 5161 degrees of freedom
## Multiple R-squared:  0.9207, Adjusted R-squared:  0.9206
## F-statistic: 1.199e+04 on 5 and 5161 DF,  p-value: < 2.2e-16
```

```
# We are interested in coef. on "education+education*female"
# obtain the bootstrap estimates
bs_estimate <- c()
```

```r
set.seed(333)
for (i in 1:1000) {
  # sampling with replcement
  data <- census[sample(nrow(census), size = 2000, replace = TRUE), ]
  # run the regression with the bootstrap sample
  bootstrap <- lm(log(income+1)~educ*factor(female)+factor(ifwork), data = data)
  # save coef. on "institutions"
  bs_estimate <- c(bs_estimate, coef(bootstrap)[2]+coef(bootstrap)[6])
}

# bootstrap estimators
summary(bs_estimate)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## 0.002083 0.012585 0.015426 0.015558 0.018384 0.028811
```

```r
# standard deviation of bootstrap estimator
sd(bs_estimate) # bs s.e.
```

```
## [1] 0.004157589
```

```r
# hypothesis testing: H_0: coef. on "institution" = 0
# 0 falls into ci_bs, so fail to reject H_0.
ci_bs <- c(quantile(bs_estimate, 0.025), quantile(bs_estimate, 0.975))
ci_bs
```

```
##        2.5%       97.5%
## 0.007711255 0.023807236
```

Dignose multicolinearity

```r
library(olsrr)
```

```
##
## Attaching package: 'olsrr'

## The following object is masked from 'package:datasets':
##
##     rivers
```

```r
# fit model
lm_vif <- lm(n_child~educ+I(log(income+1))+nroom_h+nsm_h+factor(ifwork), data = census)
# show the result
screenreg(list(lm_vif),
          custom.model.names = c('number of children'),
          custom.coef.names = c("Y_schooling", "ln(income)", "n_room",
                                "area", 'ifwork_not work', 'ifwork_others'),
          override.se = list(summary(lm_vif)$coefficients[,2]),
          override.pvalues = list(summary(lm_vif)$coefficients[,4]),
          omit.coef = c("Intercept"),
          stars = c(0.01, 0.05, 0.1),
          digits = 4)
```

```
##
## ===================================
##                  number of children
## -----------------------------------
## Y_schooling        -0.0232 ***
##                    (0.0050)
## ln(income)         -0.2144 ***
##                    (0.0275)
## n_room              0.0804 ***
##                    (0.0175)
## area                0.0002
##                    (0.0005)
## ifwork_not work    -0.0528
##                    (0.1748)
## ifwork_others      -1.1719 ***
##                    (0.1679)
## -----------------------------------
## R^2                 0.0696
## Adj. R^2            0.0666
## Num. obs.        1883
## ===================================
## *** p < 0.01; ** p < 0.05; * p < 0.1
```

```
# dignose multicolinearity
ols_vif_tol(lm_vif)
```

```
##            Variables Tolerance      VIF
## 1               educ 0.9843898 1.015858
## 2 I(log(income + 1)) 0.1121323 8.918035
## 3            nroom_h 0.5898964 1.695213
## 4              nsm_h 0.5955559 1.679104
## 5    factor(ifwork)2 0.9818811 1.018453
## 6    factor(ifwork)3 0.1115065 8.968084
```

References

Aghion, Philippe, John Van Reenen, and Luigi Zingales. 2013. “Innovation and Institutional Ownership.” *American Economic Review* 103 (1): 277–304.