Tennessee STAR Demo

Experiments and Causality

December 21, 2018

The Tennessee STAR experiment is a shining example of a well run, well executed randomized controlled experiment. Here is code, and data that reproduces the tables that are published in *Mostly Harmless Econometrics* written by Angrist and Pischke. These tables are, themselves, reproductions of tables published in Kruger (1999).

Download the data

```
# browseURL("https://drive.google.com/open?id=0BxwM1dZBYvxBc0ZMbkNXYno1OE0")
```

Load Packages and Set Options and Load Data

```
library(data.table)                     # for data manipulation
library(foreign)                        # to read spss file
library(sandwich)                       # for creating robust SEs
library(lmtest)                         # for easy testing with robuse SEs
library(multiwayvcov)                   # for clustered SEs
library(stargazer)                      # for printing tables

options(digits=3)                       # limit the printed digits

rm(list = ls())

# set the path for your machine
##path = "/Users/douglashughes/Google Drive/1 - Current Work/David Reiley-Broockman Experiments Class/Da
path = "./"
file = "STAR_Students.sav"
# set a copy aside that won't be manipulated
d_raw <- read.spss(paste0(path, file), to.data.frame = TRUE)
d_raw <- data.table(d_raw)

head(d_raw)
```

Data cleaning

We begin with some basic cleaning of the data set. We could have done this before sharing the dataset with you, but thought instead that this might be instructive for you to see!

What we have is a worse case scenario where one receives a package of data with little information about the data, how its values are stored, etc. So, we will define these things for ourselves. In fact, there are for (or more) tests in the data, and Angrist and Pischke report that the `percentile` variable they use is built from a student's performance on three of these tests. Specifically which three is not detailed in the book. As a result, our answers deviate slightly from those reported in *Mostly Harmless Econometrics* but they're close.

```
# this is the worst case scenario, when you get a dataset that
# has no information and no meta data. let us clean it a bit for you
# and then we'll describe what we've got.

##
## clean up data
##

d <- d_raw[ ,
            .(race_wa    = race %in% c("WHITE", "ASIAN"),
                 age_1985   = 1985-birthyear,
                 gender     = gender,
                 free_lunch = relevel(gkfreelunch, ref = "NON-FREE LUNCH"),
                 class_size = gkclasssize,
                 read_score = gktreadss,
                 math_score = gktmathss,
                 list_score = gktlistss,
                 school_id  = factor(gkschid),
                 treatment  = relevel(gkclasstype, ref = "REGULAR CLASS")
                )
        ]


## drop rows with any missingness.
# nrow(d)                                 # how many observations? 11601
d <- na.omit(d)                           # drop row with ANY missing
# nrow(d)                                 # now, how many observations? 5708

# cleaned up.
```

Variables to Consider

We aren't going to use everything, but we'll use some of them.

- race_wa: student race is white or asian
- age_1985: student's age in 1985
- gender: students gender
- free_lunch: did the student receive free lunch based on income
- class_size: student's class size
- read_score: student score on stanford reading test
- math_score: student score on stanford math test
- list_score: student score on stanford listening test
- school_id: school ID
- treatment: what treatment did the student receive?
  - `:: SMALL SIZE CLASS`
  - `:: REGULAR SIZE CLASS`
  - `:: REGULAR SIZE CLASS + AIDE`

Create Outomce Variable

With the data mostly cleaned up and renamed, we can now take create the outcome variable. We are creating this variable in the following way:

1. We are summing student's performance on each of the reading, math, and listening tests to create a single, composite score.
2. With this score, we are calculating the empirical cumulative distribution function (CDF) for each individual student compared to the entire set of students. An empirical CDF is really just a student's percentile.
3. We should note too, we aren't education experts, and aren't familiar with the Stanford Tests. (*Go Berkeley*). It might be more approprite to find a student's empirical CDF within each test and then, average these. Or it might not.

Create a student test percentile

```r
d[ , test_score := read_score + math_score + list_score]

# create a function that will evaluate the students' percentiles
pct_fun <- ecdf(d$test_score)

d[ , percentile := pct_fun(test_score) ]
```

Produce Summary Statistics

```r
# you'll see that for category and numeric variables we are doing
# slightly different things. tables for categories, and means for
# numeric.
#
# d[ , table(free_lunch, treatment)]
# d[ , table(race_wa, treatment)]
# d[ , mean(age_1985), by = treatment]
# d[ , mean(class_size), by = treatment]
# d[ , mean(percentile), by = treatment]

# we could make this slightly more consistent if we do the following:
summary_table <- d[ , .(
       free_lunch  = mean(free_lunch == "FREE LUNCH"),
       white_asian = mean(race_wa),
       age_1985    = mean(age_1985),
       class_size  = mean(class_size),
       percentile  = mean(percentile)
       ),
    by = treatment
  ]
```

In this summary statistics table, we have three difference groups, each corresponding to one of the treatments that students in classrooms were assigned to receive. The "REGULAR + AIDE CLASS" group is the group of classrooms that were the normal size, but were assigned to receive a teachers' aide; the "SMALL CLASS" are the classrooms that were capped at size 15; and, the "REGULAR CLASS" are the regular sized classrooms.

First, let's consider the `class_size` varible. If the treatment worked, then the classrooms that are in the treatment group should have at most 15 students in expectation. Indeed, it appears this is the case, with 15.1 students. Also, if the randomization worked, then the classroom size of the other classrooms should be roughly the same, since classes in these two groups were just randomly assiged either to receive an aide or not receive an aide. Indeed, they are quite close with `class_size`s of 22.7 and 22.3.

The other variables in this summary table, `free_lunch`, `white_asian`, and `percentile` are further checks to assess whether randomization was producing reasonably balanced groups. Although we cannot measure potential outcoems for the students in all of these groups (since they for at some of the groups these are unmeasureable because they will not be realized), what we *can* do is to check that our randomization proudced groups that look roughly the same *on observables*. Indeed, in this set, it appears that on observables the classroom groups are quite similar; and so, we don't have any indication that our randomization went awry, or that we have difference in the potential outcomes *before* treatment (which would be a form of experimenter induced selection bias).

```
stargazer(summary_table, summary = FALSE, header = FALSE,
          title = "Summary Table of Key Covariates",
          digits = 2)
```

Table 1: Summary Table of Key Covariates

	treatment	free_lunch	white_asian	age_1985	class_size	percentile
1	REGULAR + AIDE CLASS	0.50	0.67	5.25	22.70	0.48
2	SMALL CLASS	0.47	0.68	5.25	15.10	0.54
3	REGULAR CLASS	0.47	0.68	5.24	22.30	0.49

Estimate Effects

As we have made a point of saying in the course, there are number of mechanical ways to estimate an average treatment effect. One simple way is using the difference in means between the two groups.

```
diff_mean <- d[ , .(group_mean   = mean(percentile),
                    group_var     = var(percentile),
                    observations = .N
                    )
               , by = .(treatment)]
diff_mean
```

```
##               treatment group_mean group_var observations
## 1: REGULAR + AIDE CLASS      0.482    0.0789         2009
## 2:          SMALL CLASS      0.536    0.0859         1716
## 3:        REGULAR CLASS      0.491    0.0840         1983
```

With this information, we can easily calculate the difference beween groups as well as the standard error of the treament effect (e.g. Field Experiments equation 3.6). Neat!

Here, we compare the difference between the `REGULAR + AIDE` and the `SMALL` against the `REGULAR` classes.

```
diff_mean[ , difference := group_mean - group_mean[3]]
diff_mean[ , diff_se    := sqrt((group_var/observations) + (group_var[3]/observations[3]))]
diff_mean[ , t          := difference / diff_se]
diff_mean[ , p_val      := 2*(1-pnorm(abs(t)))]
```

```
diff_mean[ , .(treatment, difference, diff_se, t, p_val)]
```

```
##               treatment difference diff_se     t    p_val
## 1: REGULAR + AIDE CLASS   -0.00944 0.00903 -1.04 2.96e-01
## 2:          SMALL CLASS    0.04481 0.00961  4.66 3.14e-06
## 3:        REGULAR CLASS    0.00000 0.00920  0.00 1.00e+00
```

But, somewhat more easily, and more easily extensible to address a number of problems that might arise, we can also use a linear model which will provide us with identical p-values, and can still be interpreted as a causal estimand.

```
?lm
m1 <- lm(percentile ~ treatment, data = d)
m2 <- lm(percentile ~ treatment + school_id, data = d)
m3 <- lm(percentile ~ treatment + race_wa + gender + free_lunch + school_id,
         data = d)
```

Let’s look quickly at the first model that we estimated, the two group difference between the treamtent and the control group.

```
stargazer(m1, m2,
          type = "text", omit = "school_id",
          apply.coef = function(x) x * 100,
          apply.se = function(x) x * 100,
          add.lines = list(c("School Fixed Effects", "No", "Yes")))
```

```
##
## ================================================================================
##                                               Dependent variable:
##                               --------------------------------------------------
##                                                    percentile
##                                         (1)                       (2)
## --------------------------------------------------------------------------------
## treatmentSMALL CLASS                  4.480***                 5.340***
##                                       (0.949)                   (0.847)
##
## treatmentREGULAR + AIDE CLASS          -0.944                   -0.257
##                                       (0.911)                   (0.814)
##
## Constant                             49.100***                 20.300***
##                                       (0.646)                   (3.540)
##
## --------------------------------------------------------------------------------
## School Fixed Effects                     No                      Yes
## Observations                           5,708                     5,708
## R2                                     0.006                     0.243
## Adjusted R2                            0.006                     0.232
## Residual Std. Error              0.288 (df = 5705)         0.253 (df = 5627)
## F Statistic                   18.300*** (df = 2; 5705) 22.600*** (df = 80; 5627)
## ================================================================================
## Note:                                                *p<0.1; **p<0.05; ***p<0.01
```

In `m1` rather than estimateing the means in each group, instead we’re estimating the difference (the *ATE*) beween the category listed in the model report and the reference category (which in this case we have set to be the “regular” classroom group).

What we haven’t done, is to correct the standard errors. This is in the next section.

Correct the Standard Errors

There are a lot of reasons to suspect that students' performances might not satisfy the assumptions that go in to building normal, Gaussian distributions of the residuals.

- One that comes to mind immediatly is that perhaps students who are assigned to treatment have greater variance in their outcomes than students who are assigned to control. This could happen if there are some students who really excel in the small classroom.
- Another, that we aren't going to correct for in this document, but that we **certainly** should, is that the outcomes within a school might also be correlated in a way that our model is not accounting for. Including a fixed effect for each school effectively *de-means* the school effects so that we have a more precise estimate of the treatment effect within each school, but it doesn't address any of the *within* school correlation that might exist.
- As we talk about in the course, if we have relatively high Inter-Cluster Correlation, then our standard errors are inappropriately enthuiastic about rejecting the null hypothesis (they're too small). This is because our standard errors are behaving as though we have `N` observations, when in fact functionally we might have (many) fewer useful observations.

Correct standard errors to be robust SEs

```
# here's what is happening:
# 1. we are using the `vcovHC` function from the library `sandwich`
#    to estimate the white heteroskedastic-consistent standard errors

m1.vcovHC <- vcovHC(m1)  # from library(sandwich)
m2.vcovHC <- vcovHC(m2)
m3.vcovHC <- vcovHC(m3)


# 2. with these, we can use the `coeftest` function from the `lmtest`
#    package to perform hypothesis tests.
#    these are the `robust` standard errors.

# wald test is: coef / se

# library(lmtest)
?coeftest
r1 <- coeftest(m1, vcov = vcovHC(m1, type = "const"))
r2 <- coeftest(m1, vcov = vcovHC(m1, type = "HC3"))

stargazer(r1, r2, type = "text")
```

```
##
## ==========================================================
##                                   Dependent variable:
##                               ----------------------------
##
##                                    (1)            (2)
## ----------------------------------------------------------
## treatmentSMALL CLASS             0.045***      0.045***
##                                  (0.009)        (0.010)
##
## treatmentREGULAR + AIDE CLASS     -0.009        -0.009
##                                  (0.009)        (0.009)
##
## Constant                         0.491***      0.491***
##                                  (0.006)        (0.007)
##
## ==========================================================
## ==========================================================
## Note:                          *p<0.1; **p<0.05; ***p<0.01
```

```r
coeftest(m1, vcov = m1.vcovHC)
```

```
##
## t test of coefficients:
##
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     0.49142    0.00651   75.49  < 2e-16 ***
## treatmentSMALL CLASS            0.04481    0.00962    4.66  3.2e-06 ***
## treatmentREGULAR + AIDE CLASS -0.00944    0.00904   -1.04      0.3
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
coef(m1) / sqrt(diag(m1.vcovHC))
```

```
##                   (Intercept)          treatmentSMALL CLASS
##                         75.49                          4.66
## treatmentREGULAR + AIDE CLASS
##                         -1.04
```

```r
coeftest(m1, vcov = m1.vcovHC)
```

```
##
## t test of coefficients:
##
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     0.49142    0.00651   75.49  < 2e-16 ***
## treatmentSMALL CLASS            0.04481    0.00962    4.66  3.2e-06 ***
## treatmentREGULAR + AIDE CLASS -0.00944    0.00904   -1.04      0.3
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
coeftest(m1, vcov = vcovHC(m1, type = "const"))
```

```
##
## t test of coefficients:
##
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     0.49142    0.00646   76.06  < 2e-16 ***
## treatmentSMALL CLASS            0.04481    0.00949    4.72  2.4e-06 ***
## treatmentREGULAR + AIDE CLASS -0.00944    0.00911   -1.04      0.3
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# 3. to print more nicely,  we are taking the square-root of the diagonals
# of this heteroskedastic consistent variance covariance matrix, which
# provides the standard errors for each of the coefficients.

rse1 <- sqrt(diag(m1.vcovHC))
```

```r
rse2 <- sqrt(diag(m2.vcovHC))
rse3 <- sqrt(diag(m3.vcovHC))
```

General Loop for Inference

There is a general, three-step process for drawing inference.

1. Estimate coefficients; then,
2. Estimate standard errors; then,
3. Test and draw inference.

When our data meets the assumptions for Ordinary Least Squares regression, we can perform both of these at the same time because the $\hat{\beta}_{OLS} = (X'X)^{-1}(X'Y)$ solution has built into it a ready made computation for the standard errors for $\hat{\beta}$ which is $\hat{SE}_{OLS} = \hat{\sigma}(X'X)^{-1}$ where $\hat{\sigma}$ is drawn from the over sum of squared residuals, $\hat{\sigma} = \epsilon'\epsilon \equiv \sum \epsilon^2$. (As a notational convention the *hat* [$\hat{\cdot}$] means that we have estimated the quantity, rather than it being a theoretic quantity.)

In general, though, we might not have these handy properties. In particular, if the errors are heteroskedastic—they have a different dispersion conditional on the X values we are examining—or if they are clustered in some way, then we'll have to make a post-estimation correction to the standard errors. I'll just note here that standard errors are *always* a function of the variance covariance matrix (in parciular, the square-root of the diagonal of this matrix).

Great… but what does this look like in practice? Three steps:

1. Estimate coefficients, e.g. `mod <- lm(Y ~ X)`; then,
2. Estimate variance covariance matrix to construct the standard errors, e.g. `v.cov <- vcovHC(mod)` which is the *Huber-White* Heteroskedastic Consistent standard error.
3. Test the ratio of these, using a *Wald* test that is just the ratio of the coefficient to the standard error, e.g. `coeftest(mod, v.cov)`.

Print Nice Tables

```r
## use stargazer to print formatted tables
stargazer(m1, m2, m3, type = "text",
          omit = "school.id",
          omit.stat = "f",
          se = list(rse1, rse2, rse3),
          header = FALSE)
```

```
##
## ===================================================================================
##                                                Dependent variable:
##                               -----------------------------------------------------
##                                                    percentile
##                                      (1)               (2)               (3)
## -----------------------------------------------------------------------------------
## treatmentSMALL CLASS              0.045***          0.053***          0.053***
##                                    (0.010)           (0.009)           (0.008)
##
## treatmentREGULAR + AIDE CLASS      -0.009            -0.003            -0.001
##                                    (0.009)           (0.008)           (0.008)
##
```

```
## race_wa                                                               0.132***
##                                                                        (0.014)
##
## genderFEMALE                                                          0.044***
##                                                                        (0.006)
##
## free_lunchFREE LUNCH                                                  -0.150***
##                                                                        (0.008)
##
## Constant                          0.491***          0.203***          0.221***
##                                    (0.007)           (0.025)           (0.028)
##
## -----------------------------------------------------------------------------------
## Observations                        5,708             5,708             5,708
## R2                                  0.006             0.243             0.316
## Adjusted R2                         0.006             0.232             0.306
## Residual Std. Error           0.288 (df = 5705) 0.253 (df = 5627) 0.240 (df = 5624)
## ===================================================================================
## Note:                                                   *p<0.1; **p<0.05; ***p<0.01
```

Make it Sparkle

```
## really spiff it up.
stargazer(m1, m2, m3, type = "latex",
          omit = "school_id",
          covariate.labels = c("Small Class", "Regular + Aide", "White/Asian",
              "Female", "Free Lunch", "Intercept"),
          add.lines = list(c("School FE", "No", "Yes", "Yes")),
          se = list(100*rse1, 100*rse2, 100*rse3),
          apply.coef = function(x) 100*x,
          dep.var.labels = "Percentile",
          notes = "All coef and se scaled by 100x",
          header = FALSE,
          column.sep.width = "2pt",
          title = "OLS Regression. Small Class Causal Effect."
          )
```

Clustered SEs

The last thing here is acknowledging that we've very likely got correlated potential outcomes within schools. Including a fixed effect term for each school removes the possiblity of inducing bias in our estimate. However, since we can't plausibly assume that the variance is the same within each of the individual school clusters, then failing to appropriately account for the empirical variance might lead us to estimate inappropriately small standard errors. Why would this be a problem? Well, if we want to falsely reject the null in only 5% of cases ($\alpha = 0.95$), then if we make the wrong assumptions we might falsely reject the null hypothesis at higher (or even lower rates). Typically, we aren't kept up at night if we're a little conservative in our estimates, but beign too *gung-ho* is a problem.

As noted in both *Field Experiments* and *Mostly Harmless Econometrics* the appropriate way to assess this is to calculate the estimated variance within each of the clusters and appropritely combine these estimtates. Conceptually this is pretty simple, but getting the maths just right can be a little particular.

Table 2: OLS Regression. Small Class Causal Effect.

	Dependent variable:		
		Percentile	
	(1)	(2)	(3)
Small Class	4.480***	5.340***	5.290***
	(0.962)	(0.869)	(0.826)
Regular + Aide	−0.944	−0.257	−0.066
	(0.904)	(0.805)	(0.762)
White/Asian			13.200***
			(1.380)
Female			4.360***
			(0.647)
Free Lunch			−15.000***
			(0.797)
Intercept	49.100***	20.300***	22.100***
	(0.651)	(2.500)	(2.830)
School FE	No	Yes	Yes
Observations	5,708	5,708	5,708
R^2	0.006	0.243	0.316
Adjusted R^2	0.006	0.232	0.306
Residual Std. Error	0.288 (df = 5705)	0.253 (df = 5627)	0.240 (df = 5624)
F Statistic	18.300*** (df = 2; 5705)	22.600*** (df = 80; 5627)	31.300*** (df = 83; 5624)
Note:			*p<0.1; **p<0.05; ***p<0.01 All coef and se scaled by 100x

```
cvcov1 <- cluster.vcov(m1, ~ school_id) # from multiwayvcov
cvcov2 <- cluster.vcov(m2, ~ school_id)
cvcov3 <- cluster.vcov(m3, ~ school_id)

# coeftest(m1, vcov = vcovHC(m1))
# coeftest(m1, vcov = cluster.vcov(m1 , ~ school.id))
# coeftest(m2, cvcov2)
# coeftest(m3, cvcov3)

sec1 <- sqrt(diag(cvcov1))
sec2 <- sqrt(diag(cvcov2))
sec3 <- sqrt(diag(cvcov3))
```

And maybe print them near one another.

```
stargazer(m1, m1, m2, m2, m3, m3,
          se = list(rse1, sec2, rse2, sec2, rse3, sec3),
          type = "latex",
          omit = "school_id",
          add.lines = list(c("SE flavor", "Rob.", "Cl. Rob.", "Rob.", "Cl.Rob", "Rob.", "Cl. Rob")),
          title = "OLS Regression. Final Example.",
          omit.stat = c("ser", "f"),
          column.sep.width = "2pt",
          covariate.labels = c("Small Class", "Regular + Aide", "White or Asian Student", "Female", "Fr
          notes.align = "l",
          notes.append = TRUE,
          notes = c("", "The base category for the treatment indicators is \\textit{Normal Class Size, 
          style = "apsr"
)
```

% Table created by stargazer v.5.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Fri, Dec 21, 2018 - 15:47:54

Table 3: OLS Regression. Final Example.

	percentile					
	(1)	(2)	(3)	(4)	(5)	(6)
Small Class	0.045***	0.045***	0.053***	0.053***	0.053***	0.053***
	(0.010)	(0.015)	(0.009)	(0.015)	(0.008)	(0.014)
Regular + Aide	−0.009	−0.009	−0.003	−0.003	−0.001	−0.001
	(0.009)	(0.013)	(0.008)	(0.013)	(0.008)	(0.013)
White or Asian Student					0.132***	0.132***
					(0.014)	(0.018)
Female					0.044***	0.044***
					(0.006)	(0.006)
Free Lunch					−0.150***	−0.150***
					(0.008)	(0.010)
Intercept	0.491***	0.491***	0.203***	0.203***	0.221***	0.221***
	(0.007)	(0.008)	(0.025)	(0.008)	(0.028)	(0.015)
SE flavor	Rob.	Cl. Rob.	Rob.	Cl.Rob	Rob.	Cl. Rob
N	5,708	5,708	5,708	5,708	5,708	5,708
R^2	0.006	0.006	0.243	0.243	0.316	0.316
Adjusted R^2	0.006	0.006	0.232	0.232	0.306	0.306

*p < .1; **p < .05; ***p < .01

The base category for the treatment indicators is *Normal Class Size, No Aide.*