

Lab4_2

2023-02-24

Question 1 It's important to randomly assign both students and teachers to determine any causal effect of class size to prevent selection bias influencing potential given coefficients. It could be that really good teachers could improve test scores, just as really motivated students could improve their test scores as well, confounding the causal effect of class size on test scores.

*#Question 2: Average class size comparison between small vs. regular kindergarten classes
#Group data set by whether class size was categorized as small or not then summarize mean class size for*

```
class_size <- star |>
  group_by(small) |>
  summarise(mean = mean(class_size))
class_size
```

```
## # A tibble: 2 x 2
##   small mean
##   <dbl> <dbl>
## 1     0  22.5
## 2     1  15.1
```

The average class size for small classes was roughly 15 students per class; for regular classes, about 22-23 students per class.

#Question 3: Converting different test scores into a standardized unit

#3A: Creating standardized test scores

#creating control variables

```
control <- star |>
  filter(small == 0) |>
  select(read, math, listen, wordskill)
```

#Math

```
math_ctrl_mean <- mean(control$math, na.rm = TRUE)
math_ctrl_sd <- sd(control$math, na.rm = TRUE)
```

```
math_stdrd <- (star$math - math_ctrl_mean)/math_ctrl_sd
```

#Reading

```
read_ctrl_mean <- mean(control$read, na.rm = TRUE)
read_ctrl_sd <- sd(control$read, na.rm = TRUE)
```

```
read_stdrd <- (star$read - read_ctrl_mean)/read_ctrl_sd
```

#Wordskill

```
wordskill_ctrl_mean <- mean(control$wordskill, na.rm = TRUE)
wordskill_ctrl_sd <- sd(control$wordskill, na.rm = TRUE)
```

```
wordskill_stdrd <- (star$wordskill - wordskill_ctrl_mean)/wordskill_ctrl_sd
```

```

#Listening
listen_ctrl_mean <- mean(control$listen, na.rm = TRUE)
listen_ctrl_sd <- sd(control$listen, na.rm = TRUE)

listen_stdrd <- (star$listen - listen_ctrl_mean)/listen_ctrl_sd

#Reporting summary statistics for four standardized test score variables

summary(math_stdrd)

##      Min.   1st Qu.   Median     Mean   3rd Qu.     Max.
## -3.49539 -0.62800  0.01395  0.05194  0.63451  3.05253

sd(math_stdrd)

## [1] 1.021684

summary(read_stdrd)

##      Min.   1st Qu.   Median     Mean   3rd Qu.     Max.
## -2.47023 -0.68020 -0.07287  0.05089  0.56643  6.12831

sd(read_stdrd)

## [1] 1.0152

summary(wordskill_stdrd)

##      Min.   1st Qu.   Median     Mean   3rd Qu.     Max.
## -2.79641 -0.76367 -0.15934  0.04772  0.69222  4.40060

sd(wordskill_stdrd)

## [1] 1.012056

summary(listen_stdrd)

##      Min.   1st Qu.   Median     Mean   3rd Qu.     Max.
## -4.22024 -0.61906 -0.01382  0.03368  0.71247  4.07155

sd(listen_stdrd)

## [1] 1.00281

#Creating standardized test score across all scores

star$sat_index <- (math_stdrd + read_stdrd + wordskill_stdrd + listen_stdrd)/4
summary(star$sat_index)

##      Min.   1st Qu.   Median     Mean   3rd Qu.     Max.
## -2.43578 -0.59900 -0.04005  0.04606  0.56693  3.92905

#3C: Histogram of sat_index for small and regular kindergarten classes

histosat <- star |>
ggplot(aes(x = sat_index,
  fill = factor(small, labels=c("Small", "Regular")),
  y = ..density..)) +
  geom_histogram(alpha = 0.5, position = "identity") +
  labs(x = "Standardized Test Score",

```

```

y = "Frequency",
fill = "Class Size",
title = "Standardized SAT Scores for Small vs. Regular Classrooms")

```

```
histosat
```

```
## Warning: The dot-dot notation (`..density..`) was deprecated in ggplot2 3.4.0.
```

```
## i Please use `after_stat(density)` instead.
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

Standardized SAT Scores for Small vs. Regular Classrooms



```
ggsave("histostat.png")
```

```
## Saving 6.5 x 4.5 in image
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

There is some overlap between the standardized test scores for small and regular classrooms.

#4A: Aggregating the dataset by teacher information

```

by_class <- star |>
  group_by(teacher_id, school_id, small, teacher_masters, teacher_white, teacher_black, te

classes <- by_class |>
  summarise(sat_index = mean(sat_index),
            class_size = mean(class_size))

```

```

## `summarise()` has grouped output by 'teacher_id', 'school_id', 'small',
## 'teacher_masters', 'teacher_white', 'teacher_black'. You can override using the
## `.groups` argument.

```

```
#reporting summary variables
```

```
summarysmall <- classes |>
  filter(small == 1) |>
  summary()
```

```
summarysmall
```

```
##      teacher_id      school_id      small  teacher_masters
## Min.   :11203801  Min.   :112038  Min.    :1  Min.    :0.0000
## 1st Qu.:17632902  1st Qu.:176329  1st Qu.:1  1st Qu.:0.0000
## Median :21252204  Median :212522  Median :1  Median :0.0000
## Mean   :21003518  Mean    :210035  Mean    :1  Mean    :0.3228
## 3rd Qu.:24476902  3rd Qu.:244769  3rd Qu.:1  3rd Qu.:1.0000
## Max.   :26494505  Max.    :264945  Max.    :1  Max.    :1.0000
## teacher_white  teacher_black  teacher_experience  sat_index
## Min.    :0.0000  Min.    :0.0000  Min.    : 0.000  Min.    :-0.93613
## 1st Qu.:1.0000  1st Qu.:0.0000  1st Qu.: 4.500  1st Qu.: -0.21783
## Median :1.0000  Median :0.0000  Median : 8.000  Median : 0.05043
## Mean    :0.8661  Mean    :0.1339  Mean    : 9.024  Mean    : 0.16308
## 3rd Qu.:1.0000  3rd Qu.:0.0000  3rd Qu.:13.000  3rd Qu.: 0.54551
## Max.    :1.0000  Max.    :1.0000  Max.    :27.000  Max.    : 1.51173
##      class_size
## Min.    :12.00
## 1st Qu.:14.00
## Median :15.00
## Mean    :14.96
## 3rd Qu.:16.00
## Max.    :17.00
```

```
summaryreg <- classes |>
  filter(small == 0) |>
  summary()
```

```
summaryreg
```

```
##      teacher_id      school_id      small  teacher_masters
## Min.   :11203802  Min.   :112038  Min.    :0  Min.    :0.0000
## 1st Qu.:17733278  1st Qu.:177333  1st Qu.:0  1st Qu.:0.0000
## Median :21653604  Median :216536  Median :0  Median :0.0000
## Mean   :21161869  Mean    :211619  Mean    :0  Mean    :0.3636
## 3rd Qu.:24477402  3rd Qu.:244774  3rd Qu.:0  3rd Qu.:1.0000
## Max.   :26494503  Max.    :264945  Max.    :0  Max.    :1.0000
## teacher_white  teacher_black  teacher_experience  sat_index
## Min.    :0.0000  Min.    :0.0000  Min.    : 0.000  Min.    :-1.32666
## 1st Qu.:1.0000  1st Qu.:0.0000  1st Qu.: 5.000  1st Qu.: -0.33247
## Median :1.0000  Median :0.0000  Median :10.000  Median : 0.02808
## Mean    :0.8232  Mean    :0.1717  Mean    : 9.434  Mean    : 0.00967
## 3rd Qu.:1.0000  3rd Qu.:0.0000  3rd Qu.:13.000  3rd Qu.: 0.30099
## Max.    :1.0000  Max.    :1.0000  Max.    :27.000  Max.    : 1.70300
##      class_size
## Min.    :15.00
## 1st Qu.:21.00
## Median :22.00
## Mean    :22.37
## 3rd Qu.:24.00
## Max.    :28.00
```

```
#4B: Creating linear model for teacher experience and class size, constructing 95% confidence interval
mod1 <- lm(teacher_experience ~ small, data = classes)
summary(mod1)
```

```
##
## Call:
## lm(formula = teacher_experience ~ small, data = classes)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -9.4343 -4.4343 -0.4343  3.5657 17.9764
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9.4343     0.4157  22.697  <2e-16 ***
## small        -0.4107     0.6649  -0.618   0.537
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.849 on 323 degrees of freedom
## Multiple R-squared:  0.00118,    Adjusted R-squared:  -0.001912
## F-statistic: 0.3815 on 1 and 323 DF,  p-value: 0.5372
```

```
#95% Confidence interval
modlcoef <- coeftest(mod1, vcov = vcovHC(mod1, type="HC1"))
modlse <- modlcoef[4]
estdiff <- modlcoef[2]
lb_cf1 <- estdiff - 1.96*modlse
ub_cf1 <- estdiff + 1.96*modlse

lb_cf1 #and
```

```
## [1] -1.713294
```

```
ub_cf1
```

```
## [1] 0.8918514
```

The 95 % confidence interval for the estimated difference in teacher experience based on class size is between a -1.71 decrease in years of teacher experience and an .89 increase in years of experience. The regression coefficient is not statistically significant.

```
#4C: Teacher Masters
mod2 <- lm(teacher_masters ~ small, data = classes)
summary(mod2)
```

```
##
## Call:
## lm(formula = teacher_masters ~ small, data = classes)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.3636 -0.3636 -0.3228  0.6364  0.6772
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.36364     0.03392  10.721  <2e-16 ***
```

```
## small      -0.04080    0.05426  -0.752    0.453
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4773 on 323 degrees of freedom
## Multiple R-squared:  0.001747, Adjusted R-squared:  -0.001343
## F-statistic: 0.5654 on 1 and 323 DF, p-value: 0.4526
```

#95% Confidence interval

```
mod2coef <- coeftest(mod2, vcov = vcovHC(mod2, type="HC1"))
mod2se <- mod2coef[4]
estdiff <- mod2coef[2]
lb_cf2 <- estdiff - 1.96*mod2se
ub_cf2 <- estdiff + 1.96*mod2se
```

```
lb_cf2 #and
```

```
## [1] -0.1464958
```

```
ub_cf2
```

```
## [1] 0.06489236
```

The 95 % confidence interval for the estimated difference in a teacher having a master's degree based on class size is between a -0.14 decrease in the likelihood of a teacher having a master's and a 0.06 increase in the likelihood of a given teacher having a master's. The regression coefficient is not statistically significant.

#4c: Teacher white

```
mod3 <- lm(teacher_white ~ small, data = classes)
summary(mod3)
```

```
##
## Call:
## lm(formula = teacher_white ~ small, data = classes)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.8661  0.1339  0.1768  0.1768  0.1768
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.82323    0.02609  31.552  <2e-16 ***
## small        0.04291    0.04174   1.028   0.305
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3671 on 323 degrees of freedom
## Multiple R-squared:  0.003261, Adjusted R-squared:  0.0001756
## F-statistic: 1.057 on 1 and 323 DF, p-value: 0.3047
```

#95% Confidence interval

```
mod3coef <- coeftest(mod3, vcov = vcovHC(mod3, type="HC1"))
mod3se <- mod3coef[4]
estdiff <- mod3coef[2]
lb_cf3 <- estdiff - 1.96*mod3se
ub_cf3 <- estdiff + 1.96*mod3se
```

```
lb_cf3 #and
```

```
## [1] -0.03690069
```

```
ub_cf3
```

```
## [1] 0.1227195
```

The 95 % confidence interval for the estimated difference in a teacher being white based on class size is between a -0.03 decrease in the likelihood of a teacher being white and a 0.12 increase. The regression coefficient is not statistically significant.

```
#4C: Teacher Black
```

```
mod4 <- lm(teacher_black ~ small, data = classes)
```

```
summary(mod4)
```

```
##
```

```
## Call:
```

```
## lm(formula = teacher_black ~ small, data = classes)
```

```
##
```

```
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max  
## -0.1717 -0.1717 -0.1717 -0.1339  0.8661
```

```
##
```

```
## Coefficients:
```

```
##              Estimate Std. Error t value Pr(>|t|)  
## (Intercept)  0.17172    0.02590   6.631 1.4e-10 ***  
## small       -0.03786    0.04143  -0.914   0.361
```

```
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
```

```
## Residual standard error: 0.3644 on 323 degrees of freedom
```

```
## Multiple R-squared:  0.002579, Adjusted R-squared:  -0.0005088
```

```
## F-statistic: 0.8352 on 1 and 323 DF, p-value: 0.3614
```

```
#95% Confidence interval
```

```
mod4coef <- coeftest(mod4, vcov = vcovHC(mod4, type="HC1"))
```

```
mod4se <- mod4coef[4]
```

```
estdiff <- mod4coef[2]
```

```
lb_cf4 <- estdiff - 1.96*mod4se
```

```
ub_cf4 <- estdiff + 1.96*mod4se
```

```
lb_cf4
```

```
## [1] -0.1172655
```

```
ub_cf4
```

```
## [1] 0.04154774
```

The 95 % confidence interval for the estimated difference in a teacher being Black based on class size is between a -0.11 decrease in the likelihood of a teacher being white and a 0.04 increase. The regression coefficient is not statistically significant.

Question 4D As reported above, none of the regression coefficients in the linear models exploring the relationship between each of the teacher characteristics and class size were statistically significant. The coefficient magnitudes are quite small compared to the intercept, suggesting that they are not practically significant. This suggests that the randomization was successful in balancing teacher characteristics.

#Building multilinear regression model using school fixed effects

```
mod5 <- lm(sat_index ~ small + factor(school_id), data = classes)
summary(mod5)
```

```
##
## Call:
## lm(formula = sat_index ~ small + factor(school_id), data = classes)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.03037 -0.18519 -0.01429  0.17938  1.16597
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -0.85102    0.20512   -4.149 4.61e-05 ***
## small           0.16180    0.04089    3.957 9.95e-05 ***
## factor(school_id)123056  0.74298    0.28944    2.567 0.010855 *
## factor(school_id)128068  0.59125    0.28944    2.043 0.042150 *
## factor(school_id)128076  0.53023    0.28944    1.832 0.068183 .
## factor(school_id)128079  0.56178    0.28944    1.941 0.053418 .
## factor(school_id)130085  0.80290    0.27083    2.965 0.003331 **
## factor(school_id)159171  1.38135    0.25890    5.335 2.17e-07 ***
## factor(school_id)161176  0.54775    0.27083    2.022 0.044218 *
## factor(school_id)161183  1.24442    0.25066    4.964 1.29e-06 ***
## factor(school_id)162184  0.65436    0.27083    2.416 0.016420 *
## factor(school_id)164198  1.11146    0.28944    3.840 0.000157 ***
## factor(school_id)165199  1.67132    0.28944    5.774 2.33e-08 ***
## factor(school_id)166203  0.55576    0.27083    2.052 0.041230 *
## factor(school_id)168211  0.82023    0.25890    3.168 0.001729 **
## factor(school_id)168214  1.30453    0.28944    4.507 1.02e-05 ***
## factor(school_id)169219  1.44028    0.28944    4.976 1.22e-06 ***
## factor(school_id)169229  1.27111    0.24465    5.196 4.30e-07 ***
## factor(school_id)169231  0.23688    0.28944    0.818 0.413931
## factor(school_id)169280  1.06625    0.28944    3.684 0.000283 ***
## factor(school_id)170295  1.02704    0.27083    3.792 0.000188 ***
## factor(school_id)173312  1.70604    0.28944    5.894 1.24e-08 ***
## factor(school_id)176329  1.26890    0.27083    4.685 4.64e-06 ***
## factor(school_id)180344  0.83315    0.25890    3.218 0.001465 **
## factor(school_id)189378  0.65366    0.27083    2.413 0.016535 *
## factor(school_id)189382  0.93602    0.27083    3.456 0.000646 ***
## factor(school_id)189396  0.48924    0.28944    1.690 0.092245 .
## factor(school_id)191411  0.86528    0.28944    2.989 0.003079 **
## factor(school_id)193422  1.18801    0.28944    4.104 5.52e-05 ***
## factor(school_id)193423  0.98832    0.27083    3.649 0.000322 ***
## factor(school_id)201449  1.34185    0.25066    5.353 1.99e-07 ***
## factor(school_id)203452  0.96821    0.25890    3.740 0.000229 ***
## factor(school_id)203457  1.27307    0.28944    4.398 1.63e-05 ***
## factor(school_id)205488  1.02587    0.28944    3.544 0.000471 ***
## factor(school_id)205489  1.32846    0.27083    4.905 1.70e-06 ***
## factor(school_id)205490 -0.10017    0.28944   -0.346 0.729589
## factor(school_id)205491  0.94005    0.27083    3.471 0.000613 ***
## factor(school_id)205492  0.91019    0.27083    3.361 0.000902 ***
## factor(school_id)208501  0.80317    0.27083    2.966 0.003320 **
```



```
## factor(school_id)208503 0.44805 0.28944 1.548 0.122916
## factor(school_id)209510 0.67458 0.25890 2.606 0.009734 **
## factor(school_id)212522 0.94284 0.27083 3.481 0.000591 ***
## factor(school_id)215533 1.22527 0.25066 4.888 1.84e-06 ***
## factor(school_id)216536 0.84208 0.25890 3.253 0.001305 **
## factor(school_id)218562 1.14545 0.27083 4.229 3.31e-05 ***
## factor(school_id)221571 0.14005 0.25890 0.541 0.589027
## factor(school_id)221574 0.44891 0.27083 1.658 0.098696 .
## factor(school_id)225585 0.65390 0.27077 2.415 0.016470 *
## factor(school_id)228606 1.09829 0.27083 4.055 6.73e-05 ***
## factor(school_id)230612 1.36672 0.28944 4.722 3.94e-06 ***
## factor(school_id)231616 1.01846 0.28944 3.519 0.000517 ***
## factor(school_id)234628 0.97852 0.25890 3.780 0.000197 ***
## factor(school_id)244697 0.10599 0.25890 0.409 0.682619
## factor(school_id)244708 0.05013 0.25066 0.200 0.841660
## factor(school_id)244723 0.07303 0.25890 0.282 0.778118
## factor(school_id)244727 0.62049 0.27077 2.292 0.022779 *
## factor(school_id)244728 0.18125 0.28944 0.626 0.531762
## factor(school_id)244736 0.86994 0.28944 3.006 0.002926 **
## factor(school_id)244745 1.07229 0.27077 3.960 9.82e-05 ***
## factor(school_id)244746 0.76322 0.28944 2.637 0.008903 **
## factor(school_id)244755 0.81234 0.24463 3.321 0.001035 **
## factor(school_id)244764 0.63017 0.32368 1.947 0.052689 .
## factor(school_id)244774 0.42030 0.25890 1.623 0.105784
## factor(school_id)244776 0.39342 0.24009 1.639 0.102575
## factor(school_id)244780 1.69034 0.28944 5.840 1.65e-08 ***
## factor(school_id)244796 0.32779 0.27083 1.210 0.227328
## factor(school_id)244799 0.47331 0.27083 1.748 0.081787 .
## factor(school_id)244801 0.64923 0.27083 2.397 0.017273 *
## factor(school_id)244806 1.24893 0.24002 5.203 4.14e-07 ***
## factor(school_id)244818 0.40085 0.27077 1.480 0.140048
## factor(school_id)244831 0.54744 0.28944 1.891 0.059757 .
## factor(school_id)244839 1.30757 0.25890 5.050 8.61e-07 ***
## factor(school_id)252885 1.01454 0.27077 3.747 0.000223 ***
## factor(school_id)253888 0.86472 0.28944 2.988 0.003098 **
## factor(school_id)257899 0.71585 0.25890 2.765 0.006126 **
## factor(school_id)257905 1.22951 0.24465 5.025 9.69e-07 ***
## factor(school_id)259915 0.71096 0.28944 2.456 0.014732 *
## factor(school_id)261927 1.12293 0.27077 4.147 4.64e-05 ***
## factor(school_id)262937 1.73670 0.27083 6.412 7.31e-10 ***
## factor(school_id)264945 1.32069 0.25890 5.101 6.77e-07 ***
```

```
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
```

```
## Residual standard error: 0.3545 on 245 degrees of freedom
```

```
## Multiple R-squared:  0.6528, Adjusted R-squared:  0.5408
```

```
## F-statistic: 5.831 on 79 and 245 DF,  p-value: < 2.2e-16
```

```
#95% confidence interval
```

```
mod5coef <- coeftest(mod5, vcov = vcovHC(mod5, type="HC1"))
```

```
mod5se <- mod5coef[2,2]
```

```
estdiff <- mod5coef[2]
```

```
lb_cf5 <- estdiff - 1.96*mod5se
```

```
ub_cf5 <- estdiff + 1.96*mod5se
```

```
#Bar chart showing the difference in standardized test score between small and regular classes
control_bar <- mean(classes$sat_index[classes$small == 0])
treat_bar <- control_bar + estdiff
```

```
#create a dataframe
data_for_graph <- data_frame(c(control_bar, treat_bar),
                              c(NA, mod5se),
                              c("Control", "Treatment"),
                              c(NA, lb_cf5),
                              c(NA, ub_cf5))
```

```
## Warning: `data_frame()` was deprecated in tibble 1.1.0.
## i Please use `tibble()` instead.
```

```
names(data_for_graph)[1] <- "SAT_Index"
names(data_for_graph)[2] <- "SE"
names(data_for_graph)[3] <- "Groups"
names(data_for_graph)[4] <- "Lower_bound"
names(data_for_graph)[5] <- "Upper_bound"
```

```
#create graph
```

```
bar <- data_for_graph |>
  ggplot(aes(x = Groups, y = SAT_Index, fill = Groups, label = round(SAT_Index, 2))) +
  geom_bar(stat = "identity", show.legend = FALSE, width = .6) +
  geom_errorbar(aes(ymin = Lower_bound, ymax = Upper_bound), width=.1, size = 0.7, color="black",
  labs(y = "SAT Index Score",
        title = "Estimated Effect of STAR Treatment on Test Scores") +
  scale_fill_manual(values = c("red", "blue")) +
  geom_text(position = position_dodge(width = .9), # move to center of bars
            vjust = -0.5, # nudge above top of bar
            size = 3)
```

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use `linewidth` instead.
```

```
ggsave("bar.png")
```

```
## Saving 6.5 x 4.5 in image
```

```
bar
```

