EDUC 231D: Homework 1

Shuhan (Alice) Ai

Scenario

The principal at School 5184 heard about this super cool class you're taking where you examined the relationship between a student's family SES and their math achievement. The principal would like you to take a look at the data they have for 35 eighth grade students at their school to help them better understand the relationship at their school.

The principal set you some descriptive information about the students (included in this document) and uploaded the student data file to our BruinLearn site. The file is named hw1_sch5184_data.Rmd. Somehow there was R code included in the document that might be useful for conducting the analysis.

Use the descriptive information and data file to answer the following questions.

Submit your responses to the questions as a PDF file by **12PM on January 14**. The file name for the PDF you submit should use the following naming convention:

HW1_[LastName]_[FirstName].pdf

Set Up

To get started, you need to load some R packages and the data file.

> **Note**
>
> If you have not already installed these packages, you will first have to install them before loading the libraries.

```
# clear the R environment just in case there are things loaded that we don't want
# (start with a clean slate)
rm(list=ls())

# load packages
```

```r
library("tidyverse")
library("skimr")
library("flextable")

# load data file: make sure the file path matches where you have the file saved
hw1 <- readRDS("/Users/aishuhan/Desktop/HW1/hw1_sch5184_data.RDS")

# set a working directory for where you can save files
setwd("/Users/aishuhan/Desktop/HW1/")
```

Descriptives

Here's the descriptive information the principal sent you.

> **Note**
>
> In the data file, `bsmmatxx` is the student's math test score and `homeses` is a summary score for the student's family socioeconomic status (SES), where higher values imply higher family SES.

```r
skim(hw1)
```

Table 1: Data summary

Name	hw1
Number of rows	35
Number of columns	4
_____	
Column type frequency:	
character	2
numeric	2
_____	
Group variables	None

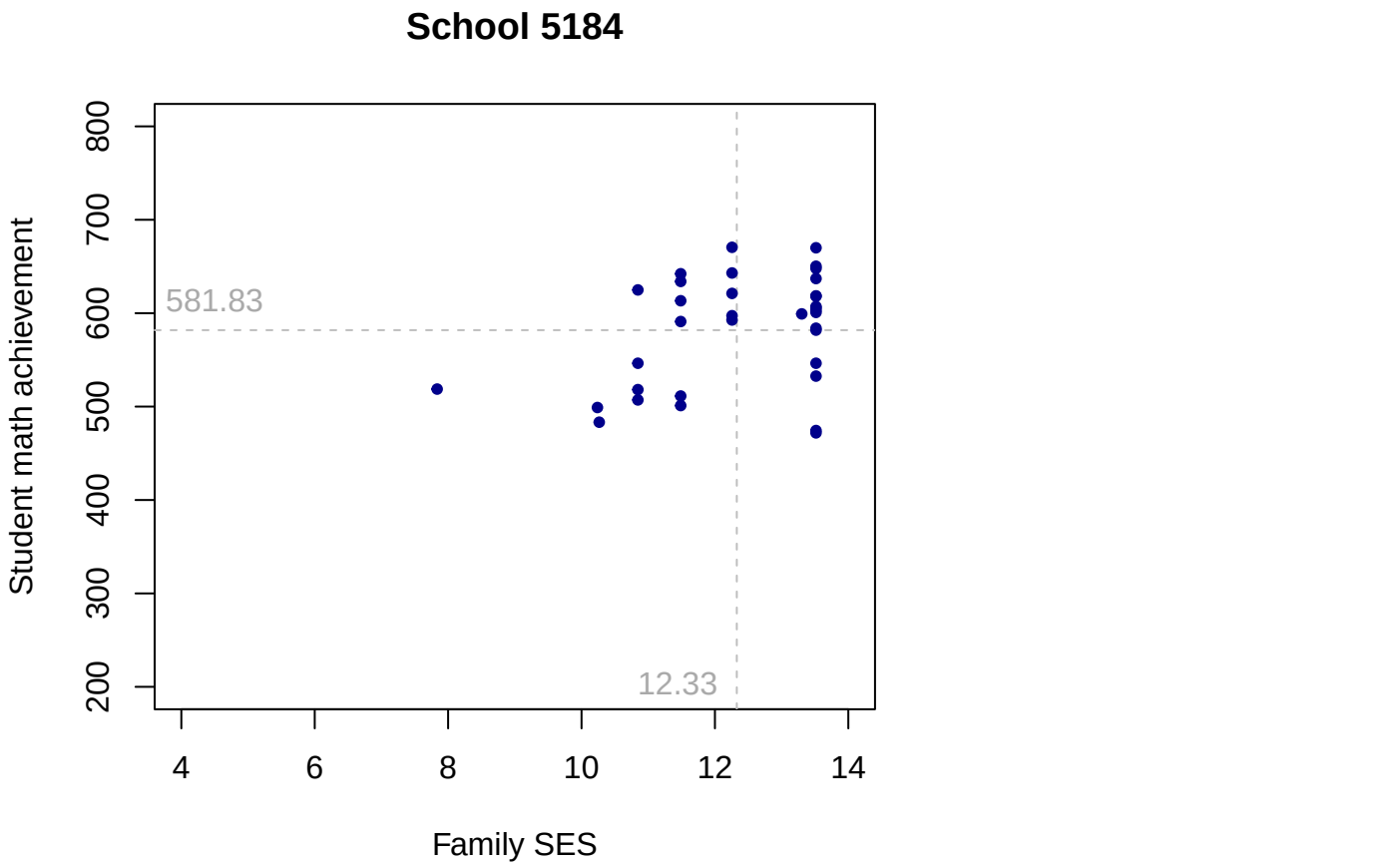
Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
idschool	0	1	4	4	0	1	0
idstud	0	1	8	8	0	35	0

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
bsmmatxx	0	1	581.83	58.89	471.87	525.68	599.23	623.01	670.52	
homeses	0	1	12.33	1.39	7.84	11.49	12.26	13.52	13.52	

```
# scatter plot
plot(NULL, main = "School 5184", xlab = "Family SES", ylab = "Student math achievement",
     ylim = c(200, 800), xlim = c(4, 14))
abline(h = mean(hw1$bsmmatxx), lty = 2, col = "grey")
abline(v = mean(hw1$homeses), lty=2, col = "grey")
points(hw1$homeses, hw1$bsmmatxx, pch=20, col = "darkblue")
text(4.5, mean(hw1$bsmmatxx), round(mean(hw1$bsmmatxx), 2), cex=1, col = "darkgrey", pos=3)
text(mean(hw1$homeses), 200, round(mean(hw1$homeses), 2), cex=1, col = "darkgrey", pos=2)
```



Question 1

How do the students in School 5184 compare to the students in School 5006 (from the Lecture 2 slides)? Based on the provided descriptive statistics, briefly describe how the distribution of students at each school compare in terms of family SES and math achievement.

- According to the descriptive statistics, students in School5184 have higher mean math scores (581.83) compared to School5006 (489.65), indicating better overall math scores. And standard deviation in School5006 (69.12) is higher than School5184 (58.89), suggesting a greater variability in math scores in School5006. The range of math scores in School5006 is wider than in School5184.
- For the family SES, students in School5184 have a relatively higher mean family SES (12.33) compared to School5006 (9.79). The variability in the family SES is slightly larger in School5184 (sd=1.39) compared to School5006 (sd= 1.16). The range of SES in School5184 is slightly broader than in School5006.

Question 2

Looking at the provided plot, how would you describe the relationship between family SES and math score? Do you see a systematic trend? If so, would you say it's steep or flat or something else? Why? How does the plotted pattern for School 5184 compare with the plotted pattern for School 5006 (from the Lecture 2 slides)?

- In general, there seems to be a positive relationship between family SES and math scores, meaning that as family SES increases, student's math score tends to increase. Compared with School5006, the slope seems to be relatively flat for School5184, meaning that family SES has a relatively lower impact on math score in School5184.

Question 3

Estimate a linear regression that predicts student math score (bsmmatxx) based on family SES (homeses) that is centered on the school mean family SES value. Before estimating the regression, create the group-mean centered version of homeses.

```
hw1 <- hw1 %>%
  mutate(homesesc = homeses - mean(homeses))

fit <- lm(bsmmatxx ~ homesesc, data = hw1)
summary(fit)
```

```
Call:
lm(formula = bsmmatxx ~ homesesc, data = hw1)

Residuals:
     Min       1Q   Median       3Q      Max
-128.494  -45.368    6.719   42.510   89.799

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  581.829      9.397   61.91   <2e-16 ***
homesesc      15.607      6.876    2.27   0.0299 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 55.6 on 33 degrees of freedom
Multiple R-squared:  0.135, Adjusted R-squared:  0.1088
F-statistic: 5.152 on 1 and 33 DF,  p-value: 0.02988
```

What is the estimate of the intercept and slope for School 5184? Briefly describe what each estimate means in language the principal and teachers at the school can understand.

- The intercept represents the predicted math score (581.83) for a student whose family SES is equal to the average SES in School5184. The slope indicates that for every one-unit increase in family SES (compared to the school average SES), the expected math score is predicted to increase by about 15.61 points.

How do the estimated intercept and slope for School 5184 compare to the estimated intercept and slope for School 5006 (from the Lecture 2 slides)?

- The intercept for School5184 (581.83) is higher than School5006 (489.65), meaning that students in School5184 perform better in math compared to students in School5006, even when both schools have students with average SES.
- The slope for School5184 (15.61) is lower than that of School5006 (25.29), suggesting that family SES has a weaker influence on math score in School5184 compared to School5006.

Feedback from Jordan for Q3:

"even when both schools have students with average SES" is not technically correct, or it's at least misleading, because SES was centered on each school's mean SES value instead of one common mean across both schools. So what constitutes "average" differs from one school to the next."

Question 4

Construct an approximate 95% confidence interval for the slope parameter for School 5184. Is it plausible that there may be no systematic linear relationship between family SES and math achievement in this school? Put differently, is a value of 0 for the slope parameter consistent with the data from School 5184?

> **Note**
>
> Show your work for constructing the 95% confidence interval.

```
slope_estimate <- coef(summary(fit))["homesesc", "Estimate"]
slope_se <- coef(summary(fit))["homesesc", "Std. Error"]
lower_bound <- slope_estimate - 2*slope_se
upper_bound <- slope_estimate + 2*slope_se

print(lower_bound)
```

```
[1] 1.855058
```

```
print(upper_bound)
```

```
[1] 29.35826
```

- The lower bound of the approximate confidence interval for the slope parameter of School5184 is 1.86, and the upper bound is 29.36. The slope value 0 would imply no systematic linear relationship between family SES and math achievement. Since 0 is not included in the confidence interval, it is not plausible that there is no systematic linear relationship between family SES and math achievement.

Comparing the 95% confidence interval for the slope parameter for School 5184 with the 95% interval for the slope parameter for School 5006, do you think there is evidence that the slope parameters for these two schools differ? Why or why not?

```
slope_estimate_5006 <- 25.294
slope_se_5006 <- 10.181
lower_bound_5006 <- slope_estimate_5006 - 2 * slope_se_5006
upper_bound_5006 <- slope_estimate_5006 + 2 * slope_se_5006
print(lower_bound_5006)
```

```
[1] 4.932
```

```
print(upper_bound_5006)
```

```
[1] 45.656
```

- The approximate 95% confidence interval for slope of the School5184 is [1.86, 29.36], and the approximate 95% confidence interval for slope of the School5006 is [4.93, 45.66], the overlap of intervals is [4.93, 29.36]. The overlap indicates that it is plausible the true slope parameters for the two schools are similar. So, there is no strong evidence to suggest that the slope parameters for the School5184 and School5006 are different. The overlapping intervals indicate that the relationship between family SES and math achievement might be similar for both schools. And the non-overlapping parts indicate that there may still be differences in the magnitude of the slopes.

Question 5

Construct an approximate 95% confidence interval for the intercept parameter for School 5184. Comparing the 95% confidence interval for the intercept parameter for School 5184 with the 95% interval for the intercept parameter for School 5006, do you think there is evidence that the intercept parameters for these two schools differ? Why or why not?

> **Note**
>
> Show your work for constructing the 95% confidence interval.

```
intercept_estimate_5184 <- coef(summary(fit))["(Intercept)", "Estimate"]
intercept_se_5184 <- coef(summary(fit))["(Intercept)", "Std. Error"]
lower_bound_intercept5184 <- intercept_estimate_5184 - 2 * intercept_se_5184
upper_bound_intercept5184 <- intercept_estimate_5184 + 2 * intercept_se_5184
print(lower_bound_intercept5184)
```

```
[1] 563.0341
```

```
print(upper_bound_intercept5184)
```

```
[1] 600.6235
```

```
intercept_estimate_5006 <- 489.648
intercept_se_5006 <- 11.626
lower_bound_intercept5006 <- intercept_estimate_5006 - 2 * intercept_se_5006
upper_bound_intercept5006 <- intercept_estimate_5006 + 2 * intercept_se_5006
print(lower_bound_intercept5006)
```

```
[1] 466.396
```

```
print(upper_bound_intercept5006)
```

```
[1] 512.9
```

- The approximately 95% confidence interval for the intercept of the School5184 is [563.034, 600.624], and the approximately 95% confidence interval for the intercept of the School5006 is [466.396, 512.9]. The confidence intervals for the intercepts of the two schools do not overlap. So there is evidence that the intercepts for the two schools are different.