Summer Poissonnier

January 28, 2021

CAP 4773

Assignment 1

Homework 1: Due Thursday, February 11 at 11:59pm

There are three parts to this homework assignment, each with multiple questions. Please insert answers and plots under their corresponding questions, then save the document as a pdf and upload it to CANVAS. *Providing your R code is not required, but may be helpful when assigning partial credit.*

We will be using the `College` dataset in the `ISLR` package for this assignment.

1) First, we will explore the dataset.

   a. [5 points] How many colleges are in the dataset?

777

   b. [5 points] How many features are there for each college?

18

   c. [5 points] Which feature(s) is(are) categorical?

Private

   d. [5 points] Which feature(s) is(are) numerical?

Apps, Accept, Enroll, F.Undergrad, P.Undergrad, Top10perc, Top25perc, Outstate, Room.Board, Books, Personal, PhD, Terminal, S.F.Ratio, perc.alumni, Expend, Grad.Rate

   e. [5 points] How many colleges are private?

565

   f. [5 points] What is the mean graduation rate across colleges?

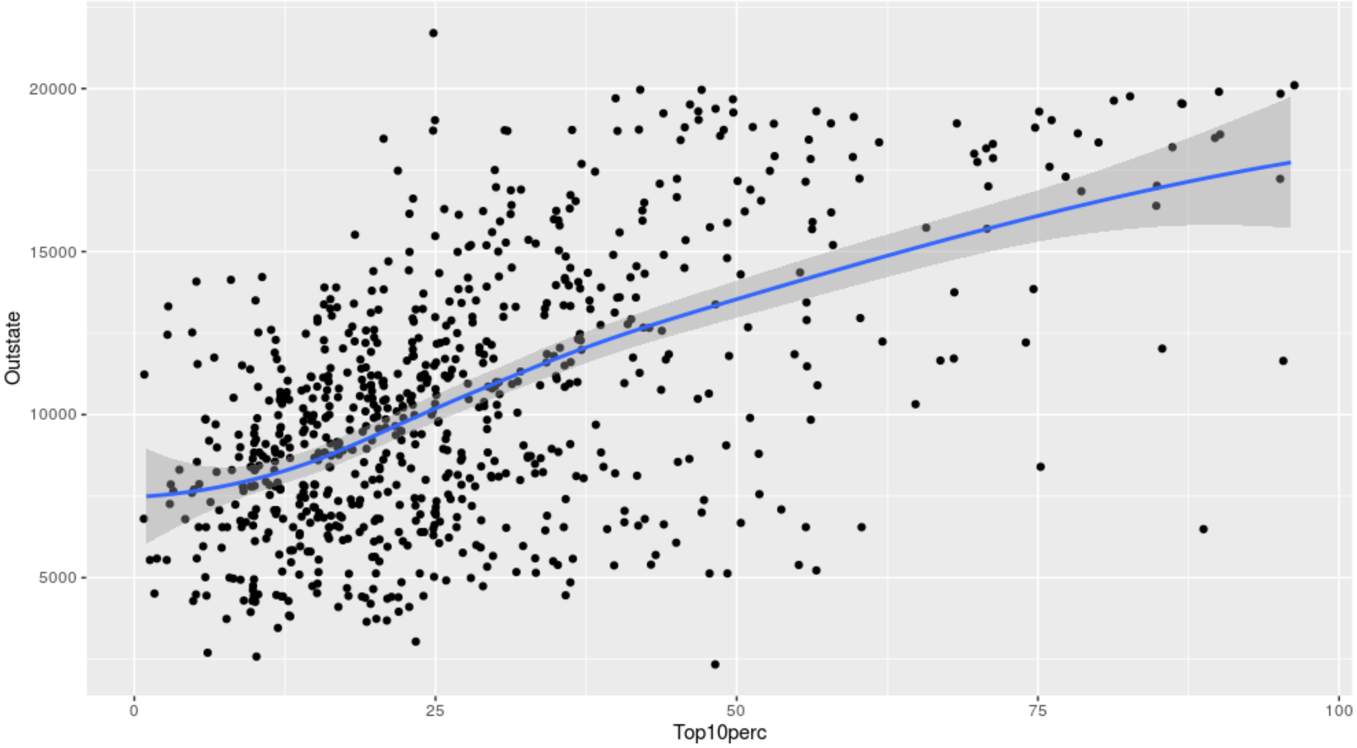
65.46%

g. [5 points] What is the maximum number of undergraduate students at a college? *Hint:* Make sure that you include both full-time and part-time students in your calculation.

31643 + 21836 = 53,479

2) Second, we will examine the relationship between the percentage of students from the top 10% of their high school class and the out-of-state tuition at a college.

a. [10 points] Create a scatterplot with the percentage of students from the top 10% of their high school class on the x axis and the out-of-state tuition on the y axis. Overlay the points with a smoothed line and 95% confidence bands. Remember to avoid overplotting.



b. [5 points] Is the correlation between the percentage of students from the top 10% of their high school class and the out-of-state tuition positive or negative?

Positive Correlation

c. [5 points] Explain what this correlation means about the relationship between the percentage of students from the top 10% of their high school class and the out-of-state tuition at a college.

This positive correlation means that the students that are in the top 10 percent of their class attend the colleges with the higher out of state tuition; likely because student who are in the top 10 percent got into more expensive colleges like Yale or Harvard and are more likely to go out of state.

d. [5 points] If you were to perform a hypothesis test to evaluate this relationship, what would be the null hypothesis?

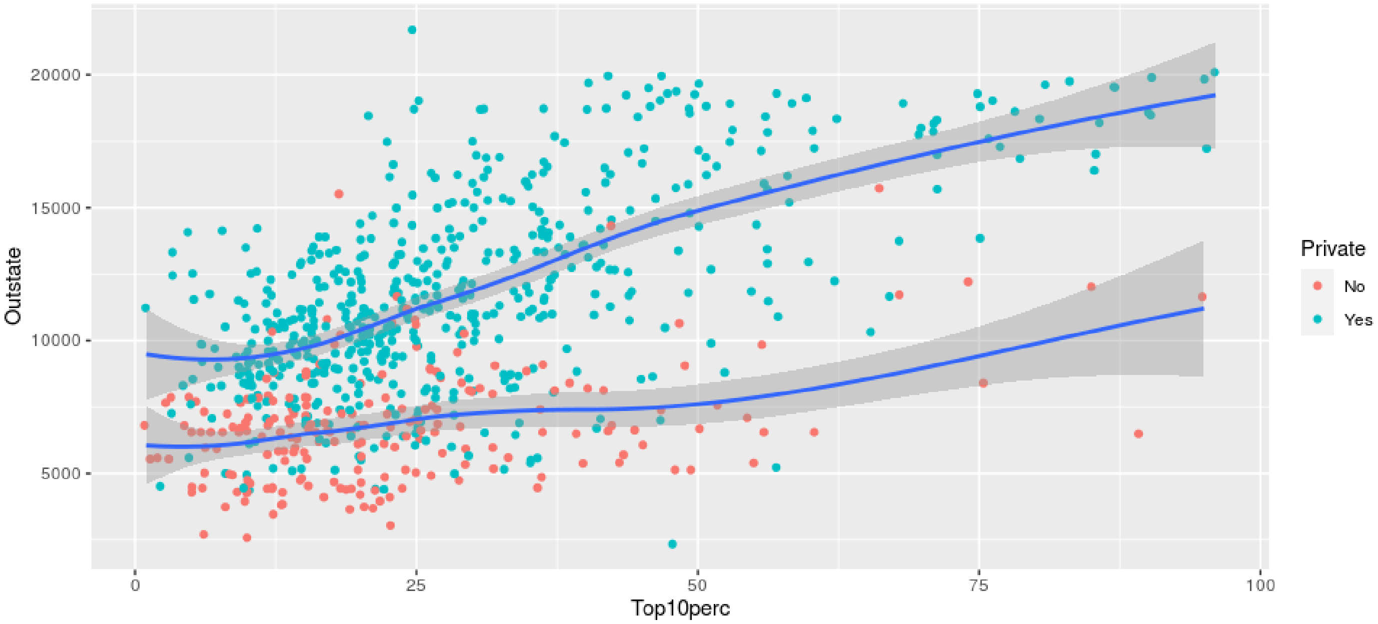
If the percentage of students from the top ten percent of their class is highest, then they are less likely to pay higher out-of-state tuition fees. Or percentage of students in the top ten percent has no effect on out of state tuition.

e. [5 points] Why is the 95% confidence interval widest when the percentage of students from the top 10% of their high school class is largest?

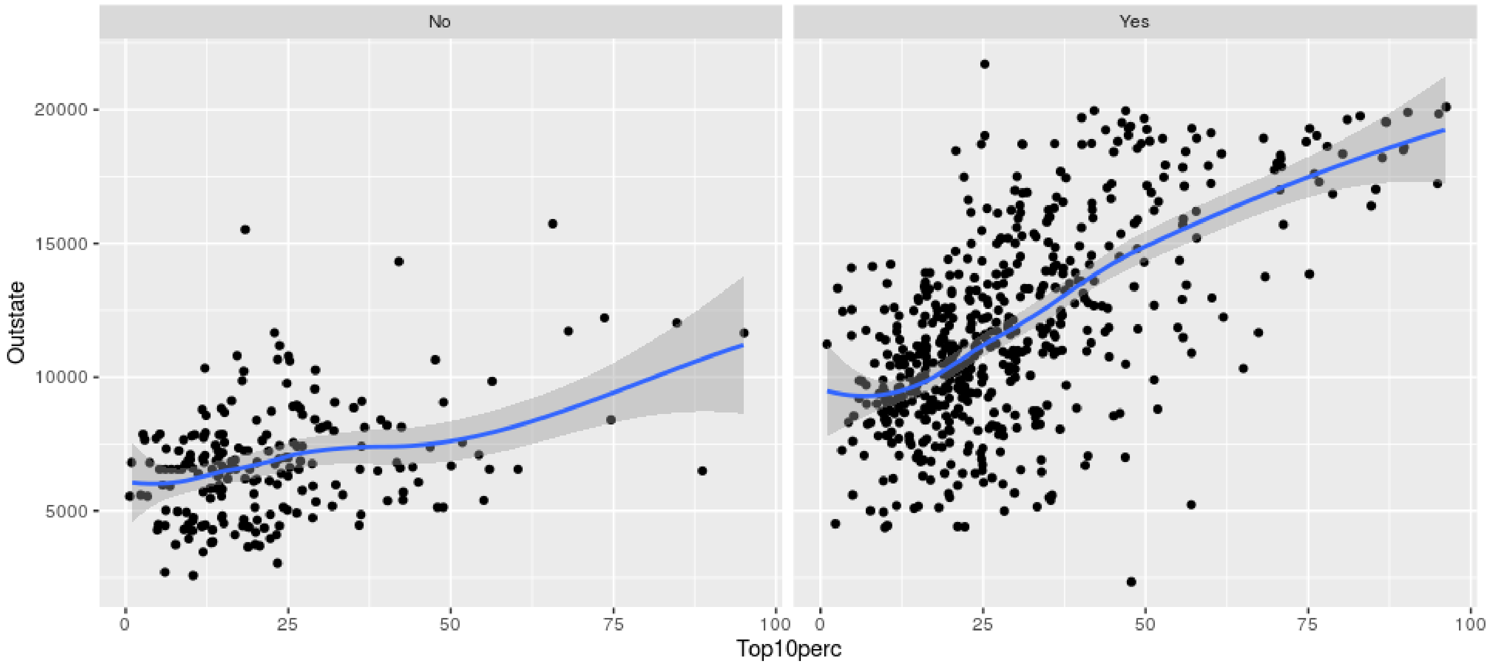
The confidence interval is widest when the percentage of students from the top ten percent of their high school class is largest because there are less students from the top ten percent of their class which causes a decrease in precision; leading to wider bands that give more availability for a mean.

3) Last, we will compare relationships between the percentage of students from the top 10% of their high school class and out-of-state tuition at public and private colleges.

a. [10 points] Using different colored points for public and private colleges, create a scatterplot with the percentage of students from the top 10% of their high school class on the x axis and the out-of-state tuition on the y axis. Overlay each set of points with a smoothed line of the same color with 95% confidence bands. Remember to avoid overplotting.



b. [10 points] Using faceting, create side-by-side scatterplots for public and private schools, with the number of students from the top 10% of their high school class on each x axis and the out-of-state tuition on each y axis. Overlay each scatterplot with a smoothed line with 95% confidence bands. Remember to avoid overplotting.



c. [5 points] Is the out-of-state tuition generally higher at public or private colleges?

Out-of-state-tuition is higher at private colleges.

d. [5 points] Is the correlation between the percentage of students in the top 10% of their high school class and the out-of-state tuition stronger for public or private colleges?

The correlation is stronger for private colleges.

e. [5 points] Why are the confidence bands generally wider for public than for private colleges?

The reason the confidence bands are generally wider for public than private colleges is because there are less students that go to public colleges that were in the top 10 percent of their class. There is less precision with the public colleges which causes greater flexibility and wider confidence bands.

```
#Question 1
summary(College)
help(College)
attach(College)

# Question 2
ggplot(data=College)+geom_point(mapping=aes(x=Top10perc,y=Outstate),
  position="jitter")+geom_smooth(mapping=aes(x=Top10perc,y=Outstate))

# Question 3 a
ggplot(data=College)+
  geom_point(mapping=aes(x=Top10perc,y=Outstate, color = Private),
  position="jitter")+
  geom_smooth(mapping=aes(x=Top10perc,y=Outstate,group=Private))

# Question 3 b
ggplot(data=College)+
  geom_point(mapping=aes(x=Top10perc,y=Outstate),
  position="jitter")+
  geom_smooth(mapping=aes(x=Top10perc,y=Outstate,group=Private))+
  facet_wrap(~ Private, ncol = 2)
```