Summer Poissonnier

February 18, 2021

CAP 4773

Assignment 2

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

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

1) First, we will use simple linear regression to fit a model and make predictions about a college's out-of-state tuition from the percentage of students from the top 10% of their high school class.

   a. (5 points) Provide the names of the feature and response variables in this simple linear regression model.

      Feature: **Percentage of students from the top 10% of their high school class (Top10perc)**

      Response: **Out of state tuition (Outstate)**

   b. (5 points) Provide the y-intercept and slope of the least squares regression line.

      y-intercept: **6906.5**

      Slope: **128.2**

   c. (5 points) Use the names of the feature and response variable from 1a and the estimated y -intercept and slope from 1b to write the formula for the least squares regression line.

      Y = 6906.5 + 128.2X
      Outstate = 6906.5 + 128.2Top10perc

   d. (5 points) Obtain 95% confidence intervals for β_0 and β_1. Ensure that you use correct notation (as in class) when writing out your confidence intervals.

      95% confidence interval for β_0: **[6471.424, 7341.493]**

      95% confidence interval for β_1: **[114.946, 141.541]**

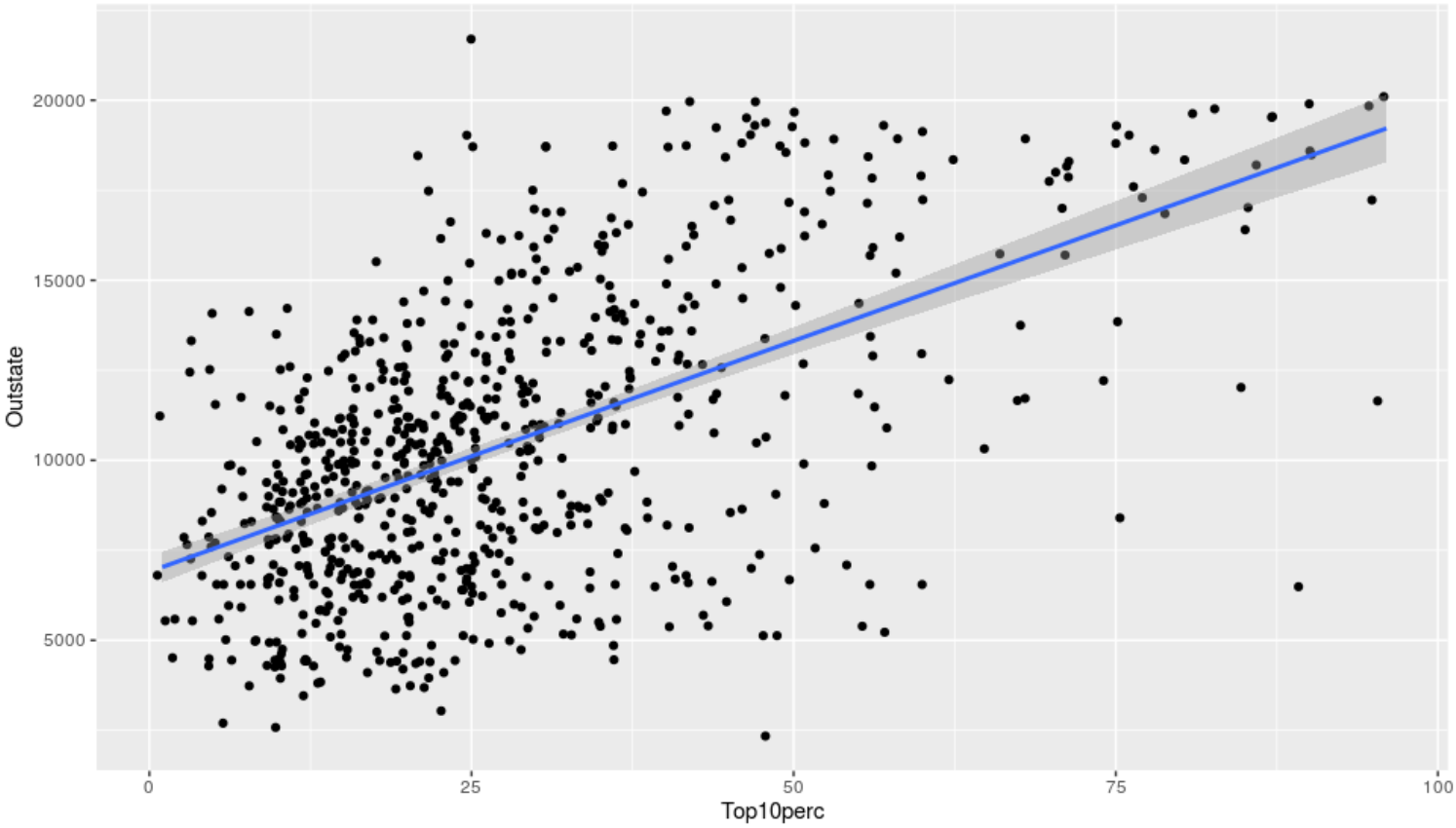
e. (5 points) Obtain a point estimate and a 95% confidence interval for the predicted out-of-state tuition of a college when 33% of its students are from the top 10% of their high school class.

Point estimate: **11138.5**

95% confidence interval: **[10893.16, 11383.84]**

I changed my code to this:
predict(lm.fit, data.frame(Top10perc = c(33)), interval = "confidence")

f. (10 points) Create a scatterplot with the feature on the x axis and the response on the y axis. Remember to avoid overplotting. Overlay the scatterplot with the least squares regression line and 95% confidence bands. *Hint:* Look at the arguments for `geom_smooth()`.



2) Next, we will perform a hypothesis test to evaluate the linear relationship between a college's out-of-state tuition and the percentage of students from the top 10% of their high school class.

a. (5 points) State the null and alternative hypotheses using mathematical notation (as in class).

H_0: $\boldsymbol{\beta_1 = 0}$

H_A: $\boldsymbol{\beta_1 \neq 0}$

b. (5 points) Provide the test statistic and associated p value.

Test statistic: **18.93**

p value: **< 2.2e-16**

c. (5 points) Draw an appropriate conclusion based on the results of your hypothesis test.

The p value of the data is very small-thus indicating that we will reject $\mathbf{H_0}$ and we can conclude that $\boldsymbol{\beta_1 \neq 0}$.

d. (5 points) Briefly explain what this conclusion means about the linear relationship between a college's out-of-state tuition and the percentage of students from the top 10% of their high school class.

This conclusion means that the linear relationship between a college's out-of-state tuition and the percentage of students from the top 10% of their high school class is directly related. As X increases, Y increases and vise versa.

3) Last, we will examine the fit of the least squares linear model for a college's out-of-state tuition from the percentage of students from the top 10% of their high school class.

a. (20 points) Create a plot that shows the data points, the least squares regression line, and vertical line segments connecting the data points to the least squares regression line. Here are step-by step instructions for generating this plot in **`ggplot2`**:

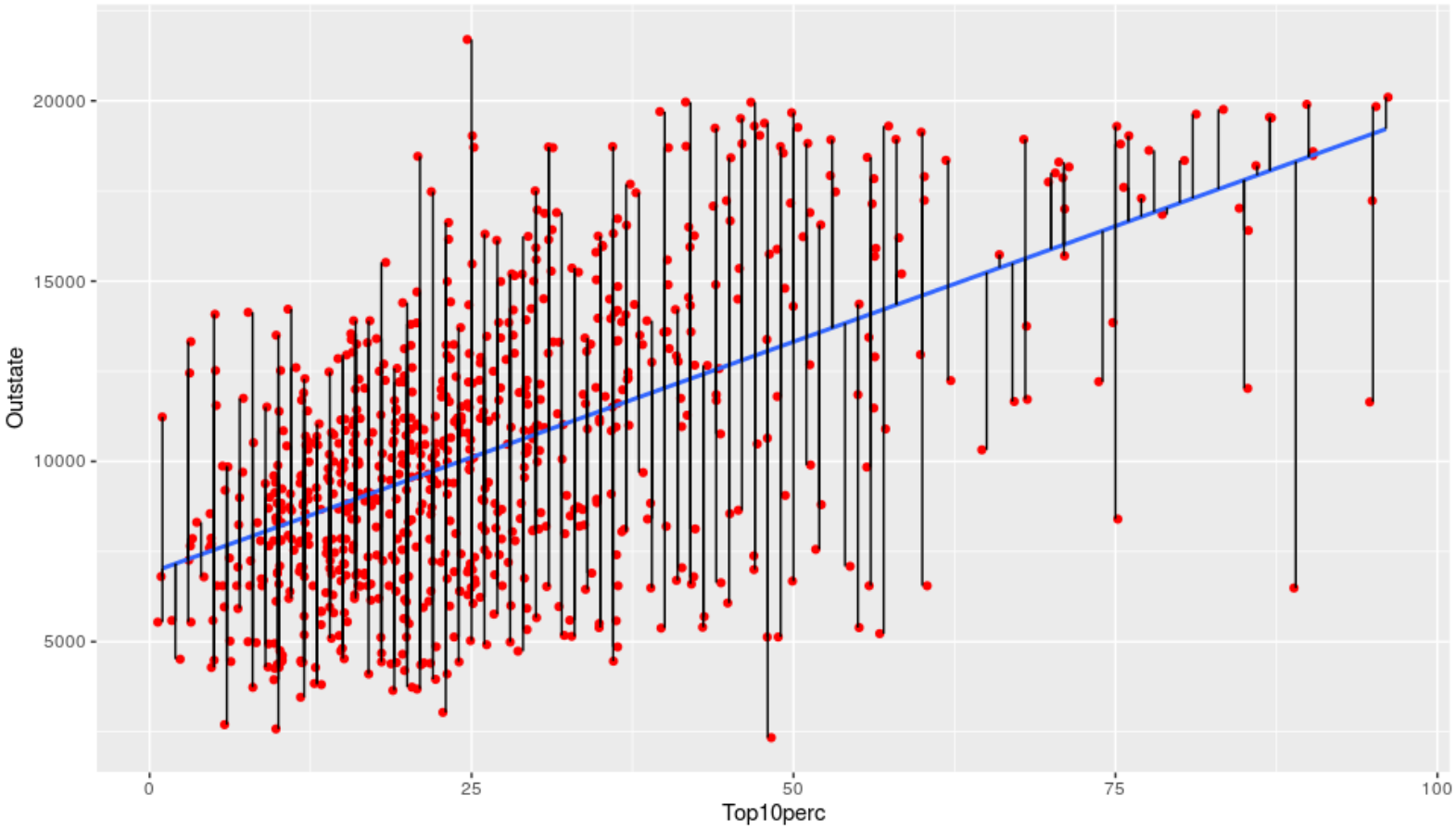
1. Call **`ggplot()`** with the **`data`** argument set to the fitted linear model. For ease, set the aesthetic mapping arguments **`x`** to the feature and **`y`** to the response within **`ggplot()`**.

2. Add a layer with a scatterplot, remembering to avoid overplotting. Color all of the points in red so that they will be easy to see.

3. As you did for question 1f, add a layer with the least squares regression line. However, this time do not display the confidence interval.

4. Add a final layer with vertical line segments connecting data points to the least squares regression line. To do this, use the **`geom_segment()`** function, and set the aesthetic mapping arguments **`xend`** to the feature variable and **`yend`** to **`.fitted`** (these are your fitted values from the least squares regression line).

5. If you would like, you can use themes to customize the background color, remove gridlines, etc. of your plot. This part is optional, and you will receive full credit as long as you complete parts 1-4.



b. (5 points) Briefly explain what the vertical line segments in your plot from 3b represent.

The vertical line segments in the plot represent the standard deviation of residuals.

c. (5 points) Provide the RSE and R^2 statistic for the linear model.

RSE = **3329**

R^2 = **0.3162**

d. (5 points) What is the average deviation between observed and predicted out-of-state tuition in dollars?

The average deviation between observed and predicted out of state tuition is: $3329.

e. (5 points) What proportion of variation in a college's out-of-state tuition can be explained by the percentage of students from the top 10% of their high school class?

31.62% of variation in out of state tuition can be explained by the percentage of students from the top 10% of their class.

f. (5 points) What strategy could you use to improve the fit of the linear model to the data? *Hint:* Think about what we observed in Homework 1 question 3.

To improve the fit of the linear model to the data, we could include more features into the model. In homework 1 question 3, we included the public and private colleges feature which gave us more data-this could be used to improve the fit of the linear model.

Code:

```
# Question 1 a-c

lm.fit <- lm(Outstate ~ Top10perc)
lm.fit

# Question 1 d

confint(lm.fit)

# Question 1 e

predict(lm.fit, data.frame(Outstate = c(33)), interval = "confidence")

# Question 1 f

ggplot(data=College)+geom_point(mapping=aes(x=Top10perc,y=Outstate),
position="jitter")+geom_smooth(mapping=aes(x=Top10perc,y=Outstate),
method='lm', formula= y~x)


#Question 2 b

summary(lm.fit)
```

 # Question 3 a

```
ggplot(data = College, mapping=aes(x=Top10perc,y=Outstate))+
  geom_point(position = "jitter", color = "red")+
  geom_smooth(method = 'lm', se = FALSE)+
  geom_segment(mapping = aes(xend = Top10perc,
  yend = fitted.values(lm.fit)))
```

 # Question 3 c-e

```
summary(lm.fit)
```