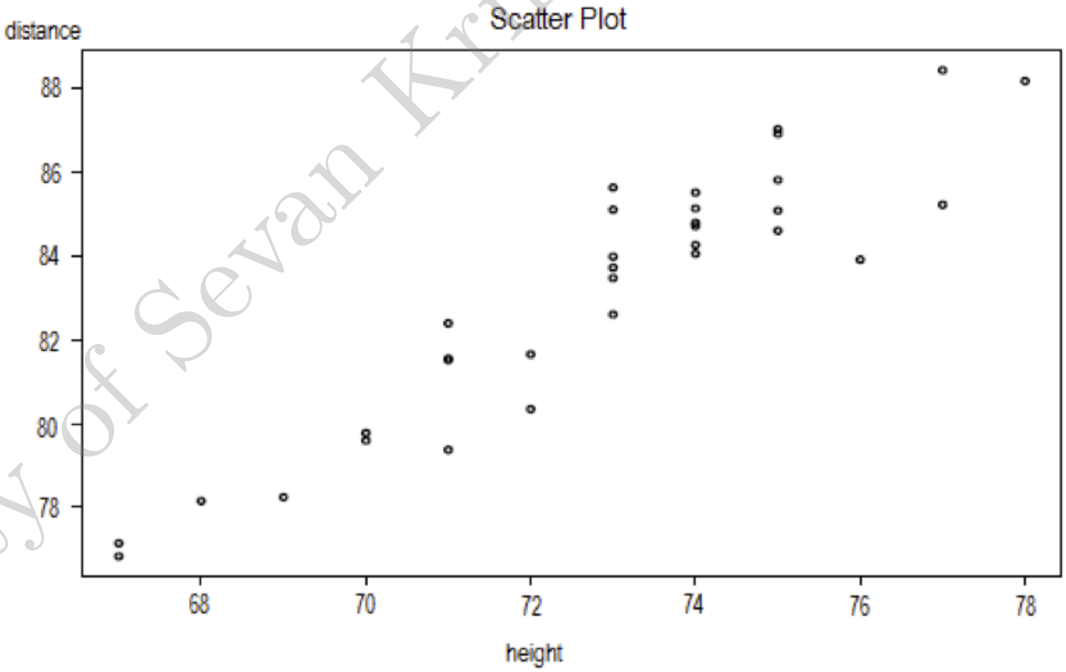
STATS 210P

Homework 1

This file is private intellectual property and cannot be distributed, sold, or reproduced without written permission of the owner.

1. A high school track & field coach wanted to asses the relationship between an athletes height and how far they can jump in the long jump event (both in inches). They collect data on each athletes height and how far they can jump. Let the height of the athlete be the explanatory variable (X) and the length of the jump be the response (Y).
The scatterplot of the data is as follows:



a. Based on the scatterplot, is there a positive or negative association between height of athlete and length of jump?

b. Based on the scatterplot, state why using a linear regression equation is justified to asses the relationship between height and length of jump.

c. Which of the following (pick one only) could be a possible value of the correlation coefficient between distance and height? Explain in a sentence or two.
Choices are: -1, -0.7, 0, 0.1, 0.7, 1.

Now say a simple linear regression model is fit to the data. Fitting the simple linear regression model, the estimated regression equation is:

$$\hat{Y} = 6.4285 + 1.0534X$$

d. What type of variable is the response (categorical or quantitative)? How about the explanatory variable?

e. What is the predicted length of the jump for an athlete who is 72 inches tall?

f. Interpret what the 1.0534 represents.

g. Does the intercept of 6.4285 inches have any useful interpretation to the coach?

h. Can the coach conclude the taller the athlete is will cause them to jump farther?

i. The original units of measurement were Y=distance length in inches and X=height in inches. Now say the *response* variable is recorded in feet NOT inches (there are 12 inches in one feet). What will happen to the intercept estimate of 6.4285? Will it stay the same, increase, or decrease? Explain in a sentence or two.

j. The original units of measurement were Y=distance length in inches and X=height in inches. Now say the *explanatory* variable is recorded in feet NOT inches (there are 12 inches in one feet). What will happen to the intercept estimate of 6.4285? Will it stay the same, increase, or decrease? Explain in a sentence or two.

k. Continue with the situation in part j. Let ρ_1 be the correlation coefficient between distance in inches and height in inches. Let ρ_2 be the correlation coefficient between distance in inches and height in feet. Is ρ_1 equal to, less than, or greater than ρ_2. Explain in a sentence.

2. Suppose researchers want to investigate if experiencing mother nature more often reduces a persons blood pressure. They define the response variable to be blood pressure (systolic), and the explanatory variable to be how many hours, on average, someone spends outdoors each day.

a. Could the researchers conduct a randomized experiment to test this? Explain briefly.

b. Say an observational study was conducted, and it was found that the people who spend more time outdoors tended to have lower blood pressure.

Could this be used to conclude that spending time outdoors causes blood pressure to decrease? Explain briefly.

c. In a few sentences, discuss why the number of miles (on average) someone jogs each day could be a possible confounder.

For questions 3-6 refer to the MedGPA.txt dataset on the class website. The MCAT (medical college admission test) is a test that is taken by students who want to attend medical school in the USA.

The description of each variable in this dataset is as follows:

- AcceptStatus: A= accepted to med school and D=denied
- Acceptance: 1=accepted and 0=denied
- Sex: F=female and M=male
- BCPM: Bio/Chem/Physics/Math grade point average
- GPA: College grade point average
- VR: Verbal reasoning subscore on the MCAT
- PS: Physical science subscore
- WS: Writing sample subscore
- BS: Biological science subscore
- MCAT: Score on the MCAT (sum of VR, PS, WS, and BS)
- Apps: Number of medical schools applied to

3. Specify which of the variables listed are quantitative and which are categorical.

4. In each part below, it will be stated what the goal of the study is. In each study, state what is the response variable and what is the explanatory variable (from the list above).

a. Do verbal reasoning scores differ on average for males and females?

b. Are equal proportions of males and females accepted into medical school?

c. Is GPA a good predictor of MCAT scores?

d. Is BCPM a good predictor of whether or not someone gets accepted into med school?

5. Read the MedGPA data set into R.

a. Compute the mean and five number summary for GPA separately for those who were admitted and for those who were denied admission to med school.

b. In a couple of sentences compare the GPAs for the two groups.

6. Continuing with the MedGPA dataset

a. Find the estimated regression equation for predicting MCAT scores based on GPA. Write the equation using proper notation and also show the output from R.

b. Write out the sum of squared errors (SSE, the objective function) that is going to be minimized with respect to the β coefficients to obtain our estimated regression equation. That is to say, write out the theoretical form of the SSE (sum of the residuals squared) using the β coefficients that is to be minimized to find the estimated regression line.

c. Interpret the slope value in context of the problem.

d. Interpret the intercept value in context of the problem. Is this a useful interpretation?

e. Use the equation from part (a) to predict the MCAT score for someone that has a GPA of 3.0. Now predict the MCAT score form someone with a GPA of 4.0.

f. What is the predicted difference in MCAT scores for two people who differ in GPA by 2.0?

g. Can we conclude that increasing GPA scores will increase MCAT scores? Explain in a sentence or two.

7. In a sentence or two, discuss why a casual relationship cannot be concluded in observational studies but can be in randomized experiments.