STAT 151A
Homework 1
Xuanpei Ouyang
3032360371

Chapter 1:

(a) Can this result be taken as evidence that completing homework assignments *causes* higher grades on the final exam? Why or why not?

- No.
- This result cannot be taken as evidence that completing homework assignment causes higher grades on the final exam. Since the instructor just passively collected the data and documented the response variable - students' final exam scores without manipulation of explanatory variable - the amount of homework completed, the study is an observational study. An observational study does not provide a causation proof and it only provides an evidence of causation.

(b) Is it possible to design an experimental study that could provide more convincing evidence that completing homework assignments causes higher exam grades? If not, why not? If so, how might such an experiment be designed?

- Yes. (It will depends on different assumptions)
- Suppose that the instructor teaches can randomly divide the classes into two sections, we can design an experimental study that provide more convincing evidence. For example, the instructor can use computer software to randomly divide students into two sections. Take one section as control group and assign normal assignments and other section as experimental group and assign more assignments. Then, the instructor can compare the final exam scores between the controlled group and experimental group to make stronger conclusion.

(c) Is it possible to marshal stronger observational evidence that completing homework assignments causes higher exam grades? If not, why not? If so, how?

- Yes. (But it is difficult to do so)
- We can find more existing data about homework and exam grades. For example, we can find two data about student's exam scores for the same class with the same professor in different year. The first class is in earlier year when instructor didn't assign homework and for the second class when instructor changed to assign homework every week. As a result, we compared the scores in these two dataset and get stronger observational evidence.
- Or we can also collect more information and control the confounding variables such as class levels (freshman, sophomore, junior, senior). For example, we can separate data based on students' different class levels and analyze the data, calculate the fitted model and make conclusion separately.

Chapter 2:

1) Explain the assumptions of linear regression.

- There are three assumptions for linear regression. First, the conditional distribution of response variable on explanatory variable(s) need to be normally distributed. Second, the variances of the conditional distribution of response variable given explanatory variable(s) need to be the same even for different values of explanatory variables. Third, the mean(s) of the conditional distribution of response variable given explanatory variable(s) need to be linear.

2) Explain the tradeoff between variance and bias.

- Variance and bias are two sources of error. Variance is the change of the conditional sample mean when we apply the regression on different intervals of sample and can be minimized by using a same number of relatively wide bins. Bias is the difference between the average value of intervals and each sample point in that interval and Bias can be minimized by using a large number of narrow bins. There is a tradeoff between variance and bias because we would like to minimize both but the ways to minimize them contradict each other. Larger bins can reduce variance but increase the bias, small bins can reduce bias but increase the variance.

3) Explain lowess method for regression. What are the advantages of lowess over local averaging.

- Lowess (Locally weighted regression) method uses local averaging smoothing techniques and fits a locally weighted least square regression line. Lowess method also give different weights to observations: more weights to nearby values to the focal x and less weights to values far away from the focal x.

- Compared to local averaging, Lowess can reduces rough jumps between different intervals and thus produce smoother regression lines, reduces boundary bias, i.e., reduce the flatten part of first and last few local averages and also make the regression more resistant to outlier.

4) When do we use non parametric naïve regression. Name one the method in this category. Explain how it works.

- We use non-parameter naive regression when it is difficult to satisfy the assumptions of linear regression or there is few satisfied assumptions for the distribution we want to model.

- One example in non-parametric naive regression methods is local averaging.
- Local averaging partition the data values based on a fixed width window into overlapping or non overlapping intervals. And then use the average conditional response means Y values on the explanatory X value(s) within the same interval and calculate linear regression lines by connecting the average Y values for all the intervals to get the non parametric regression line.