1. Do Exercise 0.6 on page 15 of the textbook. (Shown below for those who haven’t been able to purchase a book yet.)

0.6 Wine model. In his book *Super Crunchers: Why Thinking by Numbers Is the New Way to Be Smart*, Ian Ayres writes about Orley Ashenfelter, who has gained fame and generated considerable controversy by using statistical models to predict the quality of wine. Ashenfelter developed a model based on decades of data from France’s Bordeaux region, which Ayres reports as

\[
\text{WineQuality} = 12.145 + 0.00117 \text{WinterRain} + 0.0614 \text{AverageTemp} - 0.00386 \text{HarvestRain} + \varepsilon
\]

where *WineQuality* is a function of the price, rainfall is measured in millimeters, and temperature is measured in degrees Celsius.

a. Identify the response variable in this model. Is it quantitative or categorical?

*The response variable is WineQuality, quantitative.*

b. Identify the explanatory variables in this model. Are they quantitative or categorical?

*The explanatory variables are WinterRain, AverageTemp and HarvestRain. They are all quantitative.*

c. According to this model, is higher wine quality associated with more or with less winter rainfall?

*Higher wine quality is associated with more winter rainfall, because the coefficient is positive.*

d. According to this model, is higher wine quality associated with more or with less harvest rainfall?

*Higher wine quality is associated with less harvest rainfall, because the coefficient is negative.*

e. According to this model, is higher wine quality associated with more or with less average growing season temperature?

*Higher wine quality is associated with more (higher) average growing season temperature, because the coefficient is positive.*

f. Are the data that Ashenfelter analyzed observational or experimental? Explain.

*This is an observational study. None of the explanatory variables could have been randomly assigned, because rainfall and temperature are controlled by nature, not by an experimenter.*

2. Explain why cause and effect conclusions generally cannot be made based on observational studies, but can be made based on randomized experiments.

*With observational studies there can be confounding variables that could explain any observed relationships between explanatory and response variables. In other words, there could be differences among individuals in the different levels or groups for the explanatory variable that also have an effect on the response variable. But with a randomized experiment, the values of potential...*
confounding variables should be at least approximately evened out among the different levels of the explanatory variable(s). As an example, suppose it is found that people who eat a vegetarian diet get fewer colds. It could be that those people are just more health conscious in general than non-vegetarians, and that it is the health consciousness and not the vegetarian diet that is responsible for getting fewer colds.

3. Suppose researchers want to know whether drinking at least two cups of coffee a day increases the risk of kidney cancer.
   a. Could they conduct a randomized experiment to test this? Explain.

   No. You cannot randomly assign people to drink certain amounts of coffee, at least not for very long! People would not want to cooperate.

   b. If they conducted an observational study and found that those who drink at least two cups of coffee a day had a higher rate of kidney cancer than those who did not drink at least two cups of coffee a day, could they conclude that drinking coffee increases the risk of kidney cancer? Explain why or why not.

   No, because you cannot make a causal conclusion based on an observational study. For instance, in this example it could be that people who drink lots of coffee are also more likely to drink alcohol, and that alcohol (and not coffee) affects the chance of getting kidney cancer.

   c. Identify a possible confounding variable in this situation, and explain how it meets the two criteria for confounding variables (given in Lecture 1).

   An example is given in part (b) – drinking alcohol. People who don’t drink coffee may also drink less alcohol than those who do drink at least two cups of coffee a day, so that meets the condition that the confounding variable is related to the explanatory variable. And alcohol consumption is likely to affect the kidneys and thus perhaps the likelihood of getting kidney cancer, so the confounding variable (alcohol consumption) may have an effect on the response (getting kidney cancer or not).

For Exercises 4 to 7: Refer to the data set MedGPA accompanying the book and used as an example in class. You can download the data sets from the website accompanying the book, linked to the course webpage. The full list of 11 variables provided in the MedGPA dataset is as follows:

- AcceptStatus: A = accepted to medical school or D = denied admission
- Acceptance Indicator for Accept: 1 = accepted or 0 = denied
- Sex: F = female or M = male
- BCPM: Bio/Chem/Physics/Math grade point average
- GPA: College grade point average
- VR: Verbal reasoning (subscore)
- PS: Physical sciences (subscore)
- WS: Writing sample (subcore)
- BS: Biological sciences (subscore)
- MCAT: Score on the MCAT exam (sum of CR + PS + WS + BS)
- Apps: Number of medical schools applied to

(In case you aren’t familiar with it, MCAT is the Medical School Admissions Test)
4. Specify whether each of the 11 variables is Categorical or Quantitative. If the variable is Categorical, state whether either of the terms “binary” or “ordinal” are appropriate for that variable.

AcceptStatus, Acceptance and Sex are all Categorical and binary; none are ordinal
BCPM, GPA, VR, PS, WS, BS, MCAT, Apps are all quantitative.

5. Suppose the following research questions are of interest. In each case, specify which of the variables in this data set would be the response variable, and which would be the explanatory variable(s). Note that each question has only one response variable, but could have multiple explanatory variables.

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

Response: VR; Explanatory: Sex

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

Response: AcceptStatus or Acceptance; Explanatory: Sex

c. Is GPA a good predictor of MCAT score?

Response: MCAT, Explanatory: GPA

d. Which of the subscores on the MCAT is the best predictor of whether or not someone is admitted to medical school?

Response: AcceptStatus or Acceptance; Explanatory: all of the subscores of the MCAT (VR, PS, WS, BS)

e. Which of the variables BCPM GPA, full College GPA or MCAT score is the best predictor of whether someone gets into medical school?

Response: AcceptStatus or Acceptance; Explanatory: BCPM GPA, full College GPA, MCAT score

6. Read the data set into R.

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

```r
> denyGPA = MedGPA$GPA[ MedGPA$Accept == "D" ]
> acceptGPA = MedGPA$GPA[ MedGPA$Accept == "A" ]
> summary(denyGPA)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.     
> summary(acceptGPA)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.     
3.140   3.545   3.715   3.693   3.888   3.970     
```

b. Write a few sentences comparing the GPAs for the two groups (admitted and denied).

[There are various features you could mention; here are some of them.] The mean and all numbers in the five-number summary are higher for those who were admitted than for those who
were denied. The mean GPAs are 3.693 (admitted) and 3.385 (denied), resulting in a mean that is about 0.31 points higher in GPA for those admitted. The respective medians are 3.715 and 3.380, for a difference of about 0.335 GPA points. All of the students who were admitted had GPAs of 3.14 or higher, whereas some of those who were denied had GPAs below 3.0. Based on the lower quartile, 75% of those admitted had GPAs of at least 3.545, whereas (based on the upper quartile) 75% of those denied had GPAs of 3.61 or lower.

7. Continue using the MedGPA data set in R.
   a. Find the regression equation for predicting MCAT score based on GPA. Write the equation.

   ```r
   > lm(MedGPA$MCAT ~ MedGPA$GPA)
   Call:
   lm(formula = MedGPA$MCAT ~ MedGPA$GPA)
   Coefficients:
    (Intercept)    MedGPA$GPA
   3.923          9.104
   ```

   The equation is \( \hat{Y} = 3.923 + 9.104(GPA) \).

   b. Interpret the intercept and slope values in the context of this example.

   The slope of 9.104 means that for each additional point in GPA, predicted MCAT score increases by 9.104. (This is a very large increase; scores can only range from 3 to 45. It might make more sense to think about this in terms of GPA going up by 0.1 and predicted MCAT going up by 0.91 because GPAs for students where this is relevant generally only range from about 2.7 to 4.0.) The intercept of 3.923 does not have a useful meaning in this example. It would be the predicted MCAT score for an applicant with a GPA of 0, but no one would have a GPA of 0.

   c. Use the equation to predict the MCAT score for someone who has a GPA = 3.0 and for someone who has a GPA of 4.0.

   For a GPA of 3.0, \( \hat{Y} = 3.923 + 9.104(3.0) = 31.235 \).

   For a GPA of 4.0, \( \hat{Y} = 3.923 + 9.104(4.0) = 40.339 \).