Assignment 1

Math 158, Linear Models
Spring 2018

Due: Wednesday, January 24, 2018, noon, to Sakai

Important Note:

You should work to turn in assignments that are clear, communicative, and concise. Part of what you need to do is not print pages and pages of output. Additionally, you should remove these exact sentences and the information about HW scoring below.

HW advice

General notes on homework assignments (also see syllabus for policies and suggestions):

- all homework assignments should be done in R and printed from a pdf file. If you are not familiar with LaTeX, you may use R Markdown to compile to a word document. please be neat and organized which will help me, the grader, and you (in the future) to follow your work.
- be sure to include your name on all pages, and staple them together *prior* to turning in the assignment
- please include at least the number of the problem, or a summary of the question (which will also be helpful to you in the future to prepare for exams).
- it is strongly recommended that you look at the questions as soon as you get the assignment. This will help you to start thinking how to solve them!
- for R problems, it is required to use R Markdown (or R Sweave)
- in case of questions, or if you get stuck please don't hesitate to email me (though I'm much less sympathetic to such questions if I receive emails within 24 hours of the due date for the assignment).

Homework assignments

will be graded out of 5 points, which are based on a combination of accuracy and effort. Below are rough guidelines for grading.

Score & Description

5 All problems completed with detailed solutions provided and 75% or more of the problems are fully correct.

4 All problems completed with detailed solutions and 50-75% correct; OR close to all problems completed and 75%-100% correct. **One point will be deducted if superfluous information is printed or if assignment is excessively long – ask me or your mentor Bradley if you want clarification on "excessively long."**

3 Close to all problems completed with less than 75% correct

2 More than half but fewer than all problems completed and ¿ 75% correct

1 More than half but fewer than all problems completed and ¡ 75% correct; OR less than half of problems completed

0 No work submitted, OR half or less than half of the problems submitted and without any detail/work shown to explain the solutions.

Summary

The tasks in this homework assignment include reviewing ideas of simple linear regression and building linear models in R. The `broom` package in R will allow you to work directly with the output of the linear model.

Assignment

1. In a simulation exercise, a regression model with $\beta_0 = 100$, $\beta_1 = 20$, and $\sigma^2 = 25$ is applied.

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

   where β_0 and β_1 are parameters, X_i is a known constant, and ϵ_i is a random error terms with mean 0 and variance σ^2.

   An observation on Y will be made for $X = 5$.

   (a) Can you state the exact probability that Y will fall between 195 and 205? Explain.

   (b) If the normal error regression model is applicable, can you now state the exact probability that Y will fall between 195 and 205? If so, state it. The function `xpnorm` in the `mosaic` package might be helpful.

2. An analyst in a large corporation studied the relation between current annual salary (Y) and age (X) for the 46 computer programmers presently employed in the company. The analyst concluded that the relation is curvilinear, reaching a maximum at 47 years. Does this imply that the salary for a programmer increases until age 47 and then decreases? Explain.

3. In a study of the relationship for senior citizens between physical activity and frequency of colds, participants were asked to monitor their weekly time spent in exercise over a five-year period and the frequency of colds. The study demonstrated that a negative statistical relation exists between time spent in exercise and frequency of colds. The investigator concluded that increasing the time spent in exercise is an effective strategy for reducing the frequency of colds for senior citizens.

   (a) Were the data obtained in the study observational or experimental data?

   (b) Comment on the validity of the conclusions reached by the investigator.

   (c) Identify two or three other explanatory variables that might affect both the time spent in exercise and the frequency of colds for senior citizens simultaneously.

   (d) How might the study be changed so that a valid conclusion about causal relationship between amount of exercise and frequency of colds can be reached?

4. Refer to regression model:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

   where β_0 and β_1 are parameters, X_i is a known constant, and ϵ_i is a random error terms with mean 0 and variance σ^2.

(a) What is the implication for the regression function if $\beta_1 = 0$, so that the model is $Y_i = \beta_0 + \epsilon_i$? How would the regression function plot on a graph?

(b) Derive the least squares estimator of β_0 for this model. (Calculus needed)

5. Consider the Wii Mario Kart dataset in the openintro package in R. We'll regress the auction price on the number of bids.

(a) Regress auction price (Y) on number of bids (X). Use the summary command to produce the estimated regression function.

(b) Give a point estimate for the average auction price of a Mario Kart Wii game which gets 15 bids.

(c) What is the point estimate for the change in average auction price when a game has one additional bid? Explain.

(d) Do the residuals sum to zero? Show your work (in R). Estimate σ and σ^2.

(e) Provide a histogram of the residuals (use the lattice function histogram with at least 20 breaks (i.e., bins)). Comment on the residuals.

(f) Do you think your answers to parts (b) and (c) were valid for the population from which these data came?

The scaffolding for the R code is below. Do two things, (1) toggle to `eval=TRUE`, (2) fill in the blanks. For example, the difference between which two variables in the marioKart data is the auction price?

```
# only do the install command once ever
# install.packages("openintro")
require(openintro)
require(dplyr)
require(broom)
require(ggplot2)
data("marioKart")
marioKart <- marioKart %>% mutate(aucPr = ____ - ____)

ggplot(marioKart, aes(x=____, y=____)) +
  geom_point() +
  geom_smooth(method="lm", se=FALSE)

eBay_lm <- lm(____ ~ ____, data=marioKart)
tidy(____)

augment(____) %>%
  ggplot(aes(x=.resid)) +
  geom_histogram()
```

```
augment(____) %>%
  select(.resid) %>%
  skim()
```