Homework 1

1. An article under the headline “26 Million Homeless, Survey Says” in the December 28, 1994, issue of the *Columbus Dispatch* reported that 13.5 million Americans have been homeless for at least a few days sometime during their lives, based on a study by Professor Bruce Link at Columbia University. Another 12.5 million have stayed off the street only by moving in with friends or family for periods that ranged from a few days to a year. The estimates were based on a telephone survey of 1,507 adults that asked whether they recalled ever being homeless, for how long and where they slept — in the street, a shelter, abandoned buildings or someone else’s home. Comment on this article and the results of this survey, paying particular attention to possible weaknesses in the sampling strategy. Be as specific as possible.

2. The file `railtrailproperty.mpj` gives data for 104 houses in Northampton, MA that were sold in 2007. The file gives the actual sale price in 2007, with Zillow ([www.zillow.com](http://www.zillow.com)) estimates for 1998, 2011, and 2014. All prices are in thousands of 2014 dollars. Also given are lot size and living space (in thousands of square feet), the distance in miles from the Northampton Rail Trail, and measures of walkability and bikeability (ranging from 0-100 and coming from [www.walkscore.com](http://www.walkscore.com)). You should view these data as a snapshot of an ongoing stable random process related to the value of homes in the northeastern U.S., and hence inference is meaningful considering that as the population.

   (a) Calculate the mean and standard deviation of the 2014 price, and of lot size. Treating the given data as being a random sample from the general population of northeastern U.S. homes over time, for each variable, give an interval within which we would expect roughly 95% of all homes to fall. For which of these variables do you think the interval is likely to be more useful? Why?

   (b) Compute the correlation coefficients between (i) 2014 price and living space, and (ii) 2014 price and walkability score. Which variable, living space or walkability, seems to have the stronger relationship with 2014 price? Explain how you got your answer, and (from a real estate point of view) why you think the answer turned out the way it did.

   (c) Now, compute covariances, rather than correlations, of the two pairs of variables from part (b). Do the results surprise you? Should they? How do you account for the apparent differences in the results between parts (b) and (c)?

   (d) Construct and report the regression line that allows you to predict 2014 price from living space. What is the interpretation of the slope of the line?

   (e) Construct and report the regression line that allows you to predict 2014 price from rail trail distance. What is the interpretation of the slope of the line?

   (f) Construct side-by-side boxplots of 2014 price, lot size, living space, and rail trail distance separated by number of bedrooms of the home. How would you explain the patterns you see from a real estate point of view?

3. The file `hdtv.mpj` gives data from pages 42-43 of the September, 2016 issue of *Consumer Reports* about HD television sets. The file gives the suggested retail price, *Consumer Reports* rating score, test results for HD picture quality and viewing angle.
(5 = Excellent, 4 = Very good, 3 = Good, 2 = Fair), whether the reception type of the set is Ultra HD and whether the set is High Dynamic Range-enabled, respectively (1 = Yes, 0 = No), and screen size (diagonal length in inches). Using any statistics or graphics that you wish, explore the relationships in the data between price and the characteristics of the television. Do you “get what you pay for” in any reasonable sense? Are there particular sets that seem particularly under- or overpriced? How might you use your results in deciding what television to buy?

Reminder: The files railtrailproperty.mpj and hdtv.mpj, and other data files that you will need (or I will use) during the semester, can be obtained from the class web site. They also can be obtained locally on the x drive (x:\sor\jsimonof\1305). I recommend that you word process this and future homeworks, but this is not required; you should make sure that the homework is neat, however (you will lose credit for sloppiness).

Note: You might be uncomfortable doing this homework for a few reasons. First, it will probably take you longer to complete than (most of) the other homeworks, so you should leave yourself enough time to complete it. You might also find it open–ended, leaving you with a sense of not knowing exactly how to proceed. This is perfectly normal, and, in fact, desirable. Data analysis is by its nature amorphous and open–ended, and one of the purposes of this homework is to give you some experience with that. In any data analysis problem you should try to focus on the questions that might be of interest to you, and how you might use statistical methods to help you to answer those questions. You’ll also notice when doing this homework that we have not discussed much of this material in class; one of the purposes of this homework is to make sure that everyone is up to speed on basic data analysis, and the use of Minitab, before we get further into more complex data analysis later on in the semester. You should show your work and justify your statements in this and future homeworks, or you could lose credit.

Homework due: September 25