Computer Homework 2

Due April 13 at 2pm

Professor Malthouse

Work the problems and then enter the answers into Canvas. You may discuss the problems with your group members.

1. Use the auto data set from JWHT problem 3.9 on page 122. Type the following:

```
auto = read.table("http://www-bcf.usc.edu/~gareth/ISL/Auto.data",
  header=T, na.strings="?")
auto$origin = factor(auto$origin, 1:3, c("US", "Europe", "Japan"))
```

   (a) Regress `mpg` on `cylinders`, `displacement`, `weight`, and `year`. Comment on the signs of the estimated coefficients and note which are significantly different from 0. What is value of R^2?

   (b) Compute the variance inflation factors. What do they tell you?

   (c) Drop `weight` from the model. What happens to the parameter estimates and R^2?

   (d) Drop `weight` and `displacement` from the model. What happens to the parameter estimates and R^2?

2. JWHT problem 3.14a–f on page 125. For part (c)–(e), are the parameters "covered" by the 95% confidence intervals?

3. Consider the quality control data set discussed in class.

   (a) How much variation is left unexplained by the intercept model? (this will be called the *null deviance*)

   (b) How much variation is explained by adding `rate` to the intercept model?

   (c) How much additional variation is explained by adding `am` to a model that already has `rate` in it?

   (d) How much variation is unexplained by a model having both predictors?

   (e) How much less variation is explained if we drop `rate` from a model with both predictors in it?

   (f) Compute R^2 for the two-predictor model "by hand" using only the numbers you have found above. Confirm your answer by having R compute it.

   (g) Compute the F statistic for the overall test of significance by hand.

   (h) Using the two-variable model, compute the F statistic to test $H_0 : \beta_1 = 0$ by hand (where β_1 is for rate) (hint: it is in the `drop1` output).

4. For this problem you may work in groups of 6. There is nothing to turn in for now, but I expect you to come to class ready to discuss this. How does the type and amount of crime around a Divvy bike station affect the **demand** for bikes, as measured by the number of rentals per time period? We also want to assess how other independent variables are related to demand. I will be providing you with a data set giving the demand at $n = 300$ bike stations, but I want you to think about what results you would expect before looking at the data. I do **not** expect you to read other research articles. Instead, I want you to think about what could happen in the model and explanations for why. Your data will have data on how often 31 different crimes occurred in the area around the bike share station during the previous year—I have lagged the crime data to avoid problems with reverse causality. Many of the crimes are rare, so we will focus on the following eight (you could use the other if you want): theft, battery, deceptive practice, assault, burglary, robbery, criminal trespassing, narcotics, and homicide. You might want to google these terms for more precise definitions. For example, my understanding is that assault involves a threat, but not bodily harm, while battery implies harm. Deceptive practice is sometimes called fraud, and an example is passing bad checks or trying to withdraw model from the bank as someone else. You will also have data on: number of bus stops in the area, number of train stops in the area, station capacity (number of bikes), number of marked bike routes, number of businesses in the area, population density, park area, percent minority, average education level, and average per capita income. For this problem, I want you to develop a theory to explain how different types of crime will affect demand. It may be that some types of crime have no relationship with demand, and your theory should allow for this. Why might some types have an association and others not? Another consideration is that you have actual crime statistics from the Chicago Police Department instead of *perceptions* about crime. The two could be different (why?). Your next assignment will be to test your theory against the data.