Week 7 homework problems

1. As an extreme example of what can happen when an important predictor is excluded from a model, consider the data produced by the following code:

```
set.seed(100)
n <- 200
x1 <- runif(n,0,10)
x2 <- -x1+rnorm(n,0,1)
Y <- 0.1*x1+10*x2+rnorm(n,0,3)
```

Fit the linear models with x_1 and x_2 first and then with x_1 only. Comment on what happens by filling in each blank with one of the choices that follow it in the following paragraph:

*The two models differ in that the estimated effect of X_1 changes from a slight _____ (positive/negative) slope in the full model to a _____ (steep/gradual) negative slope in the reduced model. It is clear from the way the data are generated that as X_1 increases, the mean of _____ ($Y/X_1/X_2$) increases and the mean of X_2 decreases. Hence, when X_2 is left out of the model, an increase in X_1 corresponds to an **uncontrolled** _____ (increase/decrease) in X_2 so that Y responds to **both** movements. Since the association between Y and X_2 is very strong, the effect of the increase in X_1 is swamped by the effect of the decrease in X_2 and the mean of _____ ($Y/X_1/X_2$) decreases. This effect is (incorrectly) imputed, by the reduced model, to _____ ($Y/X_1/X_2$) since it is the only term in the model.*

2. Do problem **4.2**.

3. The `cruise.csv` file on Canvas contains data on 158 cruise ships in operation worldwide as of 2013. We will use `Capacity` (passenger capacity in 100s) as the response and `Length` (in 100s of feet) and `Crew` (in 100s) as predictors. Download the data and do the following.

   a. Fit the model with both predictors and their interaction. Perform a test on the significance of the interaction's coefficient, including a test statistic and p-value.

   b. Interpret the interaction's estimated effect by finishing the following sentence:

      For every additional hundred feet of length of a ship, the mean passenger capacity increases by ____ when there are 4 (hundred) crew, by ____ when there are 8 (hundred) crew, and by ____ when there are 12 (hundred) crew.

   c. Perhaps the interaction is significant because increasing the lengths of ships that serve high-end customers (as represented by ships with high numbers of crew) does not increase capacity much, while increasing lengths of ships that serve low-end customers (that put comparatively fewer crew aboard) makes a bigger difference for capacity. But the inter-relationships between all the variables makes it hard to know. To reduce these inter-relationships, calculate a new variable `CPP` (crew per passenger), by dividing `Crew` by `Capacity`. `CPP` is now maybe a good proxy variable for the "fancy-ness" of the ship. Fit the model that contains `Capacity`, `Length`, and `CPP`, and the interaction between `Length` and `CPP`. Repeat part **b** by completing the following sentence:

      For every additional hundred feet of length of a ship, the mean passenger capacity increases by ____ when there are 0.3 crew per passenger, by ____ when there are 0.5 crew per passenger, and by ____ when there are 0.7 crew per passenger.

   Does this result lend support to our theory above?

d. In a scatter plot matrix of `Capacity`, `Length`, and `CPP`, there appear to be trends between `Length` and `Capacity` and also between `Length` and `CPP`. Find the variance inflation factors for these in the interaction model you just fit. What do they tell you?