STAT 608, Spring 2022 - Assignment 5

1. Chapter 7, Question 1 parts (a), (b), and (c).

2. (Old Qualifying Exam Question) A randomized trial was conducted to investigate the relationship between a continuous response y and four treatments A, B, C, and D. The sample size was $n = 200$, with 50 observations in each of the four treatment groups. Let y be the 200×1 vector of response values, ordered so that the first 50 entries are for treatment group A, the next 50 for B, then C, and finally D. The regression model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$ was fit, where $\mathbf{X}$ is the 200×4 design matrix given by

$$\mathbf{X} = \begin{bmatrix} \mathbf{1} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{1} & \mathbf{1} & \mathbf{0} & \mathbf{0} \\ \mathbf{1} & \mathbf{0} & \mathbf{1} & \mathbf{0} \\ \mathbf{1} & \mathbf{0} & \mathbf{0} & \mathbf{1} \end{bmatrix}$$

and where each entry is a column vector of length 50. The estimated regression coefficients were $\hat{\boldsymbol{\beta}}' = [37.5, -11.5, 1.0, -27.7]$, with standard errors $2.75, 3.89, 3.89, 3.89$, and residual standard deviation $\hat{\sigma} = 19.45$. Also:

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} 0.02 & -0.02 & -0.02 & -0.02 \\ -0.02 & 0.04 & 0.02 & 0.02 \\ -0.02 & 0.02 & 0.04 & 0.02 \\ -0.02 & 0.02 & 0.02 & 0.04 \end{bmatrix}$$

   (a) Interpret each of the four regression parameters.

   (b) What is an approximate 95% confidence interval for the mean difference in response between treatment groups B and A (so, the difference $\mu_B - \mu_A$)?

   (c) What is an approximate 95% confidence interval for the mean response in treatment group B?

3. Chapter 6, Question 3

4. (From Weisberg, 2005) We are interested in the linear model $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + e$.

   (a) Suppose we fit the model above to data for which $x_1 = 2.2x_2$ with no error (that is, all residuals $= 0$). For example, X_1 could be a weight in pounds, and X_2 the weight of the same object in kg. Describe the appearance of the added-variable plot for x_2 after x_1 had been added to the above model. Explain why. Assume that Y has a correlation with the predictors that is neither 0 nor 1. Hint: Think about what goes on the x-axis and the y-axis of the added variable plot. You should notice something interesting about one of those residual vectors.

   (b) Again referring to the model above, this time suppose that Y and X_1 are perfectly correlated, so $Y = 3X_1$, without any error. Describe the appearance of the added-variable plot for x_2 after x_1 had been added to the model. Explain. Assume this time that the correlation between the predictors is between 0 and 1.

5. Suppose we are interested in the linear model $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + e_i$. Also suppose the columns $\mathbf{x}_1$ and $\mathbf{x}_2$ of the design matrix for this model have mean 0 and *length* 1. (That is, $\mathbf{x}_1'\mathbf{x}_1 = 1$ and $\mathbf{x}_2'\mathbf{x}_2 = 1$. This is a very particular situation that is unlikely to happen in practice; it just makes our arithmetic easier for a moment.). Then if r is the correlation between $\mathbf{x}_1$ and $\mathbf{x}_2$, we have the following:

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} n & 0 & 0 \\ 0 & 1 & r \\ 0 & r & 1 \end{bmatrix} \text{ and } \left(\mathbf{X}'\mathbf{X}\right)^{-1} = \begin{bmatrix} 1/n & 0 & 0 \\ 0 & 1/(1-r^2) & -r/(1-r^2) \\ 0 & -r/(1-r^2) & 1/(1-r^2) \end{bmatrix}$$

   (a) In our setup where the predictors have mean 0 and length 1, explain why SXX $= 1$. Use that to show that the VIF formula on page 203 matches $\sigma^2 \left(\mathbf{X}'\mathbf{X}\right)^{-1}$ (above).

   (b) Determine what values of r will make the variance of $\hat{\beta}_1$ and $\hat{\beta}_2$ large. Explain why, using what you know about the variance of the vector $\hat{\boldsymbol{\beta}}$.

6. In a study on weight gain in rabbits, researchers randomly assigned rabbits to 1, 2, or 3 mg of one of dietary supplements A or B (one rabbit to each level of each supplement, which is not enough, of course). Consider the linear model $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + e$, where x_1 is the dosage level of the supplement, and x_2 is a dummy variable indicating the type of supplement used.

   (a) Compute the variance inflation factor for variable x_1. You should be able to do this completely without the use of statistical software. Explain, using the word "orthogonal," why the variance inflation factor is the value computed.

   (Hint: To get started, you might go ahead and use R and the `vif()` function. You'll have to invent a response vector y; try

```
y <- c(1, 2, 3, 4, 5, 6)
```

   to get you started. Notice that the VIF is the same no matter what values you use for y. Why? Then you might look at the formulas for VIF and notice that correlation is part of that formula. Calculate correlations between vectors to see what happens. Then you'll see what is orthogonal to what.)

   (b) Now suppose that the researcher used levels 1, 2, and 3 for supplement A, and levels 2, 3, and 4 for supplement B. Use software if desired. What is the variance inflation factor for variance x_1 in this case? Is it larger or smaller than in part (a) above? Why?