STAT 608, Spring 2022 - Assignment 5
SOLUTIONS

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

   (a) USING TABLE 7.4, THE BEST MODEL BASED ON ALL THREE CRITERIA IS THE SECOND MODEL, WITH PREDICTORS X_1 AND X_2. WE NOTICE THAT R^2_{ADJ} IS HIGHEST FOR THIS MODEL (IT IS THE SAME VALUE FOR THE THIRD MODEL, BUT BY DEFAULT WE'LL CHOOSE THE SIMPLER MODEL), AND AIC AND BIC ARE SMALLEST FOR THE SECOND MODEL.

   (b) BOTH AIC AND BIC TAKE A STEP FORWARD TO ADD X_3, AND THEN STOP.

   (c) IT APPEARS THAT X_3 IS A PRETTY GOOD VARIABLE FOR EXPLAINING Y AND DOES A BETTER JOB BY ITSELF THAN EITHER X_1 OR X_2 ALONE, WHILE THE LINEAR COMBINATION OF X_1 AND X_2 TOGETHER EXPLAIN Y PERFECTLY. BUT SINCE X_3 WORKS PRETTY WELL, NEITHER X_1 NOR X_2 ALONE HELPS X_3 ENOUGH TO BE INCLUDED IN THE MODEL IN THE NEXT STEP FORWARD.

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:

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

   (a) Interpret each of the four regression parameters.

   β_0 IS THE MEAN RESPONSE FOR GROUP A. (NOTICE THAT THIS IS THE PARAMETER, NOT THE STATISTIC, SO I DON'T HAVE TO SAY IT'S THE "APPROXIMATE" MEAN OF GROUP A.) β_1 IS THE MEAN DIFFERENCE IN RESPONSE BETWEEN GROUPS B AND A. β_2 IS THE MEAN DIFFERENCE IN RESPONSE BETWEEN GROUPS C AND A. β_3 IS THE MEAN DIFFERENCE IN RESPONSE BETWEEN GROUPS D AND A.

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

First, note that $\beta_1 = \mu_B - \mu_A$. The confidence interval for β_1 is

$$\begin{aligned}\hat{\beta}_1 \pm t^*_{n-p-1}\text{SE}\left(\hat{\beta}_1\right) &= -11.5 \pm t_{196,0.025}\sqrt{19.45^2 \times 0.04} \\ &= -11.5 \pm 1.97 \times 3.89 \\ &= (-19.1633, -3.8367)\end{aligned}$$

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

The parameter of interest here is $\beta_0 + \beta_1$. Note that we can write $\mathbf{A}\hat{\boldsymbol{\beta}} = \hat{\beta}_0 + \hat{\beta}_1$ with $\mathbf{A} = [1\ 1\ 0\ 0]$. We know that for a matrix of constants $\mathbf{A}$, $\text{Var}\left(\mathbf{A}\hat{\boldsymbol{\beta}}\right) = \mathbf{A}\text{Var}\left(\hat{\boldsymbol{\beta}}\right)\mathbf{A}' = \sigma^2\mathbf{A}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{A}' = 0.02\hat{\sigma}^2$.

$$\begin{aligned}\hat{\beta}_0 + \hat{\beta}_1 \pm t^*_{n-p-1}\text{SE}\left(\hat{\beta}_0 + \hat{\beta}_1\right) &= 37.5 - 11.5 \pm t_{196,0.025}\sqrt{19.45^2 \times 0.02} \\ &= 26 \pm 1.97 \times 2.75 \\ &= (20.58, 31.42)\end{aligned}$$

3. Chapter 6, Question 3

(a) No, this is not a valid model, for the following reasons:

- The plot of the residuals versus the fitted values does not have a nice random scatter.
- The smooth line through the Scale-Location plot has increasing slope.
- There are several high leverage points which are also high residual points, meaning they strongly influence the model.
- The scatterplot matrix suggests strong correlations among predictors, so I anticipate high VIF values.

(b) Because the relationships between the predictors are not all linear, we aren't able to say what is wrong with the model. We need to use transformations before searching for model improvements.

(c) Point 223 is identified by the cutoff in R as having a Cook's Distance too large; I am willing to guess that points 222 and 67 would also have Cook's Distance values that our cutoff would flag. But it appears that point 229 should also be investigated. It has a very large residual and large fitted value, so it stands apart from many other points, like points 222 and 223, so all these points probably need to be pulled in by transformations.

(d) This model is certainly an improvement, in that the Scale-Location plot line looks much more horizontal, and the residuals vs. fitted values plot looks much more random. However, the model is still not valid, because

- The VIF values still exceed 5, so the predictors are explaining much of the same information.

- There are still several influential points: 66, 67, and 88 are bad outliers for this model.
- I'd prefer the residuals vs. fitted values plot to have a smooth line with a more horizontal slope.

(e) Our hypotheses are $H_0 : \beta_4 = \beta_6 = 0$ vs. H_a: At least one of the two parameters is not zero. If the full model were valid, the following p-value would be valid:

$$F = \frac{\left(0.1781^2 \times 228 - 0.1724^2 \times 226\right)/2}{0.1724^2} \quad \text{(p-value} = 0.0002)$$

So, we would have very strong evidence that the two models were significantly different, even though each individual predictors is not significant.

(f) We could simply add a set of indicator (dummy) variables, with $k-1$ indicators for the k manufacturers. If, however, some manufacturers make only models with one or two different numbers of cylinders, we may find that the information carried in number of cylinders is to similar to the information carried in the set of manufacturer indicator variables, and need to drop one of those variables.

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.

In this case, X_2 is perfectly explained by X_1. This means that the residuals of X_2 regressed on X_1 are all exactly 0. So, the plotted points in the added variable plot will all have horizontal axis value 0. This means the plot will be a straight vertical line of points scattered in the vertical direction.

(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.

In this case, Y is perfectly explained by X_1. This means that the residuals of Y regressed on X_1 are all exactly 0. So, the plotted points in the added variable plot will all have vertical axis value 0. This means that plot will be a straight horizontal line of points scattered in the horizontal direction.

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).

We have that

$$\text{Var}\left(\hat{\boldsymbol{\beta}}|\mathbf{X}\right) = \sigma^2 \left(\mathbf{X}'\mathbf{X}\right)^{-1} = \begin{bmatrix} \sigma^2/n & 0 & 0 \\ 0 & \sigma^2/(1-\rho^2) & -\rho\sigma^2/(1-\rho^2) \\ 0 & -\rho\sigma^2/(1-\rho^2) & \sigma^2/(1-\rho^2) \end{bmatrix}$$

So $\text{Var}\left(\hat{\beta}_1|\mathbf{X}\right) = \text{Var}\left(\hat{\beta}_2|\mathbf{X}\right) = \sigma^2/(1-\rho^2)$. If $\rho \to 0$, then $\text{Var}\left(\hat{\beta}_1|\mathbf{X}\right) = \text{Var}\left(\hat{\beta}_1|\mathbf{X}\right) \to \sigma^2$. If $|\rho| \to 1$, then $\text{Var}\left(\hat{\beta}_1|\mathbf{X}\right) = \text{Var}\left(\hat{\beta}_2|\mathbf{X}\right) \to \infty$.

As $|\rho| \to 1$, the variables x_1 and x_2 become more correlated, which indicates both variables contain mostly the same information. This is a multicollinearity problem which will lead to estimation of $\boldsymbol{\beta}$ being less accurate.

Because the mean of X is zero,

$$\text{SXX} = \sum (x_i - \bar{x})^2 = \sum x_i^2 = 1$$

and

$$\text{Var}\left(\hat{\beta}_j\right) = \frac{1}{1-\rho^2} \times \frac{\sigma^2}{(n-1)S_{x_j}^2} = \frac{1}{1-\rho^2} \times \frac{\sigma^2}{(n-1)\frac{\text{SXX}}{n-1}} = \frac{\sigma^2}{1-\rho^2}$$

which matches the VIF show on page 203.

(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 $\hat{\boldsymbol{\beta}}$.

The variance inflation factors would all equal 1 if the predictors are all orthogonal to each other. In this case, $R^2 = 0$ which will result in VIF = 1. If r is close to -1 or 1, the VIF blows up, making the variance of the $\hat{\beta}$'s large and unstable.

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.)

The design matrix is

$$\mathbf{X} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 1 \\ 1 & 2 & 0 \\ 1 & 2 & 1 \\ 1 & 3 & 0 \\ 1 & 3 & 1 \end{bmatrix}$$

The vector $\mathbf{x}_1 - \bar{x}_1 = (-1, -1, 0, 0, 1, 1)'$ is orthogonal to vector $\mathbf{x}_2 - \bar{x}_2 = (-0.5, 0.5, -0.5, 0.5, -0.5, 0.5)'$, because their dot product is 0. Since R^2 is the square of the correlation coefficient, we have that

$$\text{VIF} = \frac{1}{1 - R_{12}^2} = \frac{1}{1 - r_{12}^2}$$

where

$$r_{12} = \text{corr}(x_1, x_2) = \frac{\text{cov}(x_1, x_2)}{\sigma_{x_1} \sigma_{x_2}}$$

where

$$\text{cov}(x_1, x_2) = \text{E}\left[(x_1 - \bar{x}_1)(x_2 - \bar{x}_2)\right] = \text{E}\left[(x_1 - 2)(x_2 - 0.5)\right] = 0$$

Therefore, VIF $= 1$.

(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?

The design matrix is

$$\mathbf{X} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 2 & 0 \\ 1 & 3 & 0 \\ 1 & 2 & 1 \\ 1 & 3 & 1 \\ 1 & 4 & 1 \end{bmatrix}$$

We have $\text{cov}(x_1, x_2) = \text{E}\left[(x_1 - \bar{x}_1)(x_2 - \bar{x}_2)\right] = 0.3$, $\sigma_{x_1} = 1.4088$, and $\sigma_{x_2} = 0.5477$, so VIF $= \frac{1}{1-0.273} = 1.375$. The variance inflation factor will be larger than in part (a) because the two variables are now correlated.