Assignment 6

1. [3+3+3+3 = 12 pts] Suppose the covariances between X_1 and X_2 is 5, between X_1 and X_3 is 4, and between X_2 and X_3 is 0. Moreover, the variances of X_1, X_2, X_3 are 49, 25,9, respectively.

   (a) Write the 3×3 covariance (Σ) and correlation (R) matrices of the random vector $X = (X_1, X_2, X_3)$.

   For the next parts use the fact that for any scalars a_1, a_2, a_3:

$$\text{Var}(a_1X_1 + a_2X_2 + a_3X_3) =$$
$$a_1^2\text{Var}(X_1) + a_2^2\text{Var}(X_2) + a_3^2\text{Var}(X_3) + 2a_1a_2\text{Cov}(X_1, X_2) + 2a_1a_3\text{Cov}(X_1, X_3) + 2a_2a_3\text{Cov}(X_2, X_3).$$

   (b) Use the above formula and compute the (numerical) value of $\text{Var}(a_1X_1 + a_2X_2 + a_3X_3)$ for the following choices of $\mathbf{a} = (a_1, a_2, a_3)$:

   1. $\mathbf{a} = (2, -2, 0)$,
   2. $\mathbf{a} = (2, -1, 5)$,
   3. $\mathbf{a} = (-1, 1, 2)$,
   4. $\mathbf{a} = (-\beta_1, -\beta_2, 1)$.

   In the last case, β_1 and β_2 are unknown. Explain why you or someone might be interested in choosing them so that the variance of the relevant linear combination of the three random variables is *minimized*. (Hint: Think in terms of prediction, least squares method,.).

   (c) Derive the *normal equations* for minimizing

$$Q(b_1, b_2) = \text{Var}(X_3 - b_1X_1 - b_2X_2),$$

   assuming that the random variables have means equal to zero.

   (d) Write the equations in (c) in matrix form first, then solve it for b_1 and b_2.

2. (3+3+3+3+3 = 15 pts]

Consider the linear regression model $y = \beta_0 + \beta_1x_1 + \beta_2x_2 + e$. Suppose that the second and third columns of the design matrix X have mean zero and length one, that is $x_1^Tx_1 = 1$ and the same for x_2, and the sample correlations between these two columns is ρ.

(a). Compute the X^TX matrix and show that

$$X'X = \begin{pmatrix} n & 0 & 0 \\ 0 & 1 & \rho \\ 0 & \rho & 1 \end{pmatrix}$$

(b) Show that

$$(X'X)^{-1} = \begin{pmatrix} 1/n & 0 & 0 \\ 0 & c & -c\rho \\ 0 & -c\rho & c \end{pmatrix}$$

Where $c = \frac{1}{1-\rho^2}$

(c) Compute and express $Var(\widehat{\beta_1} - \widehat{\beta_2})$ in terms of ρ, σ^2.

(d) Determine the range of value(s) of ρ that will make the variance of the least squares estimators of β_1 and β_2 larger than $5\sigma^2$.

(e) Discuss the possible connection between the result in (d) and multicollinearity (Hint: use variance inflation factor).

Question 3 [3+3+3+3=12 points]: Solve question 1, Chapter 7 (page 252)

Question 4 [3+3 = 6 points]: 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:

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

(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 $\sigma^2 (\mathbf{X}'\mathbf{X})^{-1}$ (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 $\hat{\boldsymbol{\beta}}$.

Question 5 [3+2=5 points]: In a study on weight gain in rabbits, researchers randomly assigned rabbits to 1,2, or 3mg 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.

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

$$\text{y} < -\text{c}(1, 2, 3, 4, 5, 6)$$

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.)