Stat 6021: Homework Set 7

1. For this first question, you will continue to use the dataset `swiss` which you also used in the last homework. Load the data. For more information about the data set, type `?swiss`. The goal of the data set was to assess how fertility rates in the Swiss (French-speaking) provinces relate to a number of demographic variables.

   (a) In the previous homework, you fit a model with the fertility measure as the response variable and used all the other variables as predictors. Now, consider a simpler model, using only the last three variables as predictors: *Education*, *Catholic*, and *Infant.Mortality*. Carry out an appropriate hypothesis test to assess which of these two models should be used. State the null and alternative hypotheses, find the relevant test statistic, p-value, and state a conclusion in context. (For practice, try to calculate the test statistic by hand.)

   (b) For the model you decide to use from part 1a, assess if the regression assumptions are met.

2. (You may only use R as a simple calculator or to find p-values or critical values) The data for this question come from 113 hospitals. The key response variable is *InfctRsk*, the risk that patients get an infection while staying at the hospital. We will look at five predictors:

   - x_1: *Stay*. Average length of stay at hospital
   - x_2: *Cultures*. Average number of bacterial cultures per day at the hospital
   - x_3: *Age*. Average age of patients at hospital
   - x_4: *Census*. The average daily number of patients
   - x_5: *Beds*. The number of beds in the hospital

   Some R output is shown below. You may assume the regression assumptions are met.

```
Call:
lm(formula = InfctRsk ~ Stay + Cultures + Age + Census + Beds)


Coefficients:
```

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.2051282  1.2075929   0.170   0.8654
Stay        0.2055252  0.0660885   3.110   0.0024 **
Cultures    0.0590369  0.0103096   5.726  9.5e-08 ***
Age         0.0173637  0.0229966   0.755   0.4519
Census      0.0010306  0.0034942   0.295   0.7686
Beds        0.0004476  0.0026781   0.167   0.8676

Residual standard error: 0.9926 on 107 degrees of freedom
Multiple R-squared:  0.4765,    Adjusted R-squared:  0.4521
F-statistic: 19.48 on 5 and 107 DF,  p-value: 9.424e-14
```

```
Analysis of Variance Table

Response: InfctRsk
           Df  Sum Sq Mean Sq F value    Pr(>F)
Stay        1  57.305  57.305 58.1676 1.044e-11 ***
Cultures    1  33.397  33.397 33.8995 6.154e-08 ***
Age         1   0.136   0.136  0.1376   0.71144
Census      1   5.101   5.101  5.1781   0.02487 *
Beds        1   0.028   0.028  0.0279   0.86759
Residuals 107 105.413   0.985
```

Only use the provided R output to answer the rest of part 2.

(a) Based on the t statistics, which predictors appear to be insignificant?

(b) Based on your answer in part 2a, carry out the appropriate hypothesis test to see if those predictors can be dropped from the multiple regression model. Show all steps, including your null and alternative hypotheses, the corresponding test statistic, p-value, critical value, and your conclusion in context.

(c) Suppose we want to decide between two potential models:

- Model 1: using x_1, x_2, x_3, x_4 as the predictors for *InfctRsk*
- Model 2: using x_1, x_2 as the predictors for *InfctRsk*

Carry out the appropriate hypothesis test to decide which of models 1 or 2 should be used. Be sure to show all steps in your hypothesis test.

3. (No R required) This question is based on a data set seen in Homework Set 4.

Data from 55 college students are used to estimate a multiple regression model with response variable *LeftArm*, with predictors *LeftFoot* and *RtFoot*. All variables were measured in centimeters. You may assume the regression assumptions are met. Some R output is given below.

```
Call:
lm(formula = LeftArm ~ LeftFoot + RtFoot)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  11.7104     2.5179   4.651 2.31e-05 ***
LeftFoot      0.3519     0.2961   1.188    0.240
RtFoot        0.1850     0.2816   0.657    0.514
---

Residual standard error: 1.796 on 52 degrees of freedom
Multiple R-squared:  0.3688,    Adjusted R-squared:  0.3445
F-statistic: 15.19 on 2 and 52 DF,  p-value: 6.382e-06
```

Explain how this output indicates the presence of multicollinearity in this regression model.

4. **For practice, not required to turn in** (No R required) In Chapter 3.9, you were introduced to the unit normal and unit length scalings. In this question, we will explore another type of scaling. Let

$$\begin{aligned} y_i^* &= \frac{1}{\sqrt{n-1}}\left(\frac{y_i - \bar{y}}{s_y}\right), \\ x_{ij}^* &= \frac{1}{\sqrt{n-1}}\left(\frac{x_{ij} - \bar{x}_j}{s_j}\right), \quad j = 1, \ldots, k \end{aligned}$$

where

$$\begin{aligned} s_y &= \sqrt{\frac{\sum(y_i - \bar{y})^2}{n-1}}, \\ s_j &= \sqrt{\frac{\sum(x_{ij} - \bar{x}_j)^2}{n-1}}, \quad j = 1, \ldots, k. \end{aligned}$$

As such, the standardized regression model is as follows:

$$y_i^* = \beta_1^* x_{i1}^* + \beta_2^* x_{i2}^* + \cdots + \beta_k^* x_{ik}^* + \epsilon_i^*.$$

In this setting, we have

$$\boldsymbol{X}^* = \begin{bmatrix} x_{11}^* & \cdots & x_{1k}^* \\ x_{21}^* & \cdots & x_{2k}^* \\ \vdots & & \vdots \\ x_{n,1}^* & \cdots & x_{nk}^* \end{bmatrix}$$

and

$$\boldsymbol{Y}^* = \begin{bmatrix} y_1^* \\ \vdots \\ y_n^* \end{bmatrix}.$$

Show that the elements in the matrix $\boldsymbol{X}^{*\prime}\boldsymbol{Y}^*$ are the simple correlation between the response variable and each of the predictors, i.e., that

$$\boldsymbol{X}^{*\prime}\boldsymbol{Y}^* = \begin{bmatrix} r_{Y,1} \\ r_{Y,2} \\ \vdots \\ r_{Y,p} \end{bmatrix}.$$