Stat 6021: Homework Set 6

1. For this first question, you will use the dataset `swiss` which is part of the `datasets` package. 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) Create a scatterplot matrix and find the correlation between all pairs of variables for this data set. Answer the following questions based on the output:

      i. Which predictors appear to be linearly related to the fertility measure?

      ii. Do you notice if any of the predictors are highly correlated with one another? If so, which ones?

   (b) Fit a multiple linear regression with the fertiliy measure as the response variable and all the other variables as predictors. Use the `summary()` function to obtain the estimated coefficients and results from the various hypothesis tests for this model.

      i. What is being tested by the ANOVA F statistic? What is the relevant conclusion in context?

      ii. Look at the numerical values of the estimated slopes as well as their p-values. Do they seem to agree with or contradict with what you had written in your answer to part 1a? Briefly explain what do you think is going on here.

2. (You may only use R as a simple calculator or to find p-values or critical values) Data from $n = 113$ hospitals are used to evaluate factors related to the risk that patients get an infection while in the hospital. The response variable is *InfctRsk*, the percentage of patients who get an infection while hospitalized. The predictors are *Stay*, the average length of stay, *Age*, the average patient age, *Xrays*, a measure of how many Xrays are done in the hospital, and *Services*, a measure of how many different services the hospital offers. We consider the following multiple regression equation: $\mathrm{E}(InfctRsk) = \beta_0 + \beta_1 Stay + \beta_2 Age + \beta_3 Xrays + \beta_4 Services$. Some R output is shown below. You may assume the regression assumptions are met.

```
Call:
lm(formula = InfctRsk ~ Stay + Age + Xrays + Services)
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.170874   1.285745   0.133 0.894521
Stay         0.237209   0.060957   3.891 0.000173 ***
Age         -0.014071   0.022708   ----- --------
Xrays        0.020383   0.005524   3.690 0.000354 ***
Services     0.022718   0.006970   3.260 0.001493 **
---

Residual standard error: 1.04 on 108 degrees of freedom
Multiple R-squared:   ____,     Adjusted R-squared:  ____
F-statistic: 19.56 on 4 and 108 DF,  p-value: 3.96e-12
```

(a) What is the value of the estimated coefficient of the variable *Stay*? Write a sentence that interprets this value.

(b) Derive the test statistic, p-value, and critical value for the variable *Age*. What null and alternative hypotheses are being evaluated with this test statistic? What conclusion should we make about the variable *Age*?

(c) A classmate states: "The variable *Age* is not linearly related to the predicted infection risk." Do you agree with your classmate's statement? Briefly explain.

(d) Fill in the values for the ANOVA table for this regression model.

Source of Variation	df	SS	MS
Regression			
Error			
Total			*********

(e) What is the R^2 for this model? Write a sentence that interprets this value in context.

(f) What is the R^2_{adj} for this model?

3. (No R required) 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. 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
```

A classmate points out that there appears to be a contradiction in the R output, namely, while the ANOVA F statistic is significant, the t statistics for both predictors are insignificant. Is your classmate's concern warranted? Briefly explain.

4. (No R required) Recall in matrix notation, the least-squares estimators for the regression model can be written as

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_k \end{bmatrix} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}.$$

Fitted values are usually written as

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} = \boldsymbol{H}\boldsymbol{y}$$

where $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$. $\boldsymbol{H}$ is called the hat matrix. Show that $\boldsymbol{H}$ is idempotent, i.e., $\boldsymbol{H}\boldsymbol{H} = \boldsymbol{H}$.