Assignment 5 - Multiple Linear Regression

your name goes here

Due: Wednesday, February 28, 2018, noon, to Sakai

Summary

The tasks in this homework assignment focus on understanding the decomposition of the sums of squares associated with variables in the model.

Assignment

1. State the number of degrees of freedom that are associated with each of the following extra sums of squares:
   (a) $SSR(X_1|X_2)$
   (b) $SSR(X_2|X_1, X_3)$
   (c) $SSR(X_1, X_2|X_3, X_4)$
   (d) $SSR(X_1, X_2, X_3|X_4, X_5)$
2. Define each of the following extra sums of squares:
   (a) $SSR(X_5|X_1)$
   (b) $SSR(X_3, X_4|X_1)$
   (c) $SSR(X_4|X_1, X_2, X_3)$
3. For a multiple regression model with five X variables $(X_1, X_2, X_3, X_4, X_5)$, what is the relevant extra sum of squares for testing whether or not $\beta_5 = 0$? What about whether or not $\beta_2 = \beta_4 = 0$?
4. The following regression model is being considered in a market research study:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1}^2 + \epsilon_i$$

   State the reduced models for testing whether or not
   (a) $\beta_1 = \beta_3 = 0$
   (b) $\beta_0 = 0$
   (c) $\beta_3 = 5$
   (d) $\beta_0 = 10$
   (e) $\beta_1 = \beta_2$
5. Explain in what sense the regression sum of squares $(SSR(X_1))$ is an extra sum of squares.

6. The progress report of a research analyst to the supervisor stated: "All the estimated regression coefficients in our model with three predictor variables to predict sales are statistically significant. Our new preliminary model with seven predictor variables, which includes the three variables of our smaller model, is less satisfactory because only two of the seven regression coefficients are statistically significant. Yet in some initial trials the expanded model is giving more precise sales predictions than the smaller model. The reasons for this anomaly are now being investigated." Comment.

7. Data were collected on the volume of users on the Northampton Rail Trail in Florence, Massachusetts. Variables in the data set include the number of crossings on a particular day (measured by a sensor near the intersection with Chestnut Street, volume), the average of the min and max temperature in degrees Fahrenheit for that day (avgtemp), and a dichotomous indicator of whether the day was a weekday or a weekend/holiday (weekday).

```
require(mosaicData); data(RailTrail)
RailTrail = mutate(RailTrail, daytype = ifelse(weekday==1, "Weekday", "Wkend/Holiday"))
```

Consider the following full (additive) linear model predicting the volume on the Northampton Rail Trail.

```
summary(lm(volume ~ hightemp + lowtemp + cloudcover + precip,
           data=RailTrail))$coef

##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    35.31      59.80    0.59 5.56e-01
## hightemp        6.57       1.15    5.70 1.70e-07
## lowtemp        -1.29       1.39   -0.93 3.55e-01
## cloudcover     -7.50       3.85   -1.95 5.47e-02
## precip       -100.62      42.06   -2.39 1.90e-02
```

(a) Find the extra sum of squares (the conditional SSR) for the following models:

i. precipitation
ii. cloudcover given precipitation
iii. hightemp given precipitation and cloudcover
iv. lowtemp given precipitation and cloudcover and hightemp

(b) Test whether cloudcover can be dropped from the regression model given that precipitation, hightemp, and lowtemp are retained. Use the F^* statistic and level of significance 0.01. State the hypotheses, p-value, and conclusion in terms of the problem (that is, say things about the RailTrails and an appropriate population). [Note: you should know how to do this by hand given the ANOVA table. However, R will do the test for you with the code `anova(model1, model2)`.]

(c) Test whether (given precipitation) the coefficient on hightemp is equal to the negative of the coefficient on lowtemp. Use the F^* statistic and level of significance 0.01. State the hypotheses, p-value, and conclusion in terms of the problem (that is, say things about the RailTrails and an appropriate population).

(d) Test whether both lowtemp and cloudcover can be dropped from the model given that hightemp and precipitation are retained. Use the F^* statistic and level of significance 0.01. State the hypotheses, p-value, and conclusion in terms of the problem (that is, say things about the RailTrails and an appropriate population).