# Homework 3

Subrata Paul

9/7/2020

## Problem 1

Download the `simu_hw3.txt` data from canvas and read it in R. The data has four columns `x1, x2, x3` and `y`. Print the summary of the linear regression model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$$

- Is there something that surprise you? What it is?
- Why do you thing it might happend? Justify your answer. (You can use plots or some statistic for justification.)
- What model do you recommend? Run the recommended model and print the summary.

## Problem 2

Fit $y = \beta_0 + \beta_1 x_1$ model and populate the following table without using the `anova` function.

| Source | SS | df | MS |
|---|---|---|---|
| $SS_{reg}(X_1)$ | | | |
| RSS($X_1$) | | | |
| TSS | | | |

`SS`, `df`, and `MS` represent the sum of squares, degrees of freedom, and mean sum of squares, respectively. `MS` = `SS/df`.

## Problem 3

Fit $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$ model. Now, if you want, you can use the `anova` function.

| Source | SS | df | MS |
|---|---|---|---|
| $SS_{reg}(X_1, X_2)$ | | | |
| RSS($X_1, X_2$) | | | |
| TSS | | | |

## Problem 4

Define $SS_{reg}(X_2|X_1) = RSS(X_1) - RSS(X_1, X_2)$. $SS_{reg}(X_2|X_1)$ is called the extra sum of squares. Calculate $SS_{reg}(X_2|X_1)$. Can you write $SS_{reg}(X_2|X_1)$ in terms of $SS_{reg}$ of the above models?

## Problem 5

The dataset `teengamb` from `faraway` package concerns a study of teenage gambling in Britain. Fit a regression model with the expenditure on gambling as the response and the sex, status, income and verbal score as predictors. Present the output.

(a) What percentage of variation in the response is explained by these predictors?
(b) Which observation has the largest (positive) residual? Give the case number.
(c) Compute the mean and median of the residuals.
(d) Compute the correlation of the residuals with the fitted values.
(e) Compute the correlation of the residuals with the income.
(f) For all other predictors held constant, what would be the difference in predicted expenditure on gambling for a male compared to a female?

## Problem 6

In this question, we investigate the relative merits of methods for computing the coefficients. Generate some artificial data by:

```
x<-1:20
y <- x+ rnorm(20)
```

Fit a polynomial in $x$ for predicting $y$. Compute $\hat{\beta}$ in two ways — by `lm()` and by using the direct calculation described in the chapter. At what degree of polynomial does the direct calculation method fail? (Note the need for the `I()` function in fitting the polynomial, that is, `lm(y ~ x + I(x^2))`.

## Problem 7

The dataset `prostate` in the `faraway` package comes from a study on 97 men with prostate cancer who were due to receive a radical prostatectomy. Fit a model with `lpsa` as the response and `lcavol` as the predictor. Record the residual standard error and the $R^2$. Now add `lweight`, `svi`, `lbph`, `age`, `lcp`, `pgg45` and `gleason` to the model one at a time. For each model record the residual standard error and the $R^2$. Plot the trends in these two statistics.