# STATS 191: Homework Assignment 5

*Dr. Pratheepa Jeganathan*

*11/01/2019*

**You may discuss homework problems with other students, but you have to prepare the written assignments yourself.**

**Please combine all your answers, the computer code and the figures into one PDF file, and submit a copy to gradescope.**

Please use **newpage** to write solution for each part of a question.

Please specify the page number for each part of a question in gradescope.

**Grading scheme:** $\{0, 1, 2\}$ **points per question, total of 36.**

Due date: 11:59 PM November 08, 2019 (Friday evening).

## Question 1 (This question is based on lectures 17 and 18 and Chapter 4 from the textbook CH)

The data set `state.x77` in R contains the following statistics (among others) related to the 50 states of the United States of America:

- `Population`: population estimate (1975)
- `Income`: per capita income (1974)
- `Illiteracy`: illiteracy (1970, percent of population)
- `HS.Grad`: percent high school graduates (1970)

```
state.data = data.frame(state.x77)
```

We are interested in the relation between `Income` and other 3 variables (`Population`, `Illiteracy`, `HS.Grad`).

(1) Produce a 4 by 4 scatter plot of the variables above.

(2) Fit a multiple linear regression model to the data with `Income` as the response variable, and `Population`, `Illiteracy`, `HS.Grad` as the predictor variables. Comment on the significance of the variables in the model using the result of `summary()`.

(3) Produce standard diagnostic plots of the multiple regression fit in part 2.

(4) Plot DFFITS of the observations and find observations which have high influence, using critical value 0.5.

(5) Plot Cook's distance of the observations and find observations which have high influence, using critical value 0.1. Compare this results with the result of part (4).

(6) Find **states** with outlying predictors by looking at the leverage values. Use critical value 0.3.

(7) Find outliers, if any, in the response. Remove them from the data and refit a multiple linear regression model and compare the result with the previous fit in part (2).

(8) As a summary, find all the influential **states** using `influence.measures` function.

## Question 2 (This question is based on lectures 17 and 18 and Chapter 4 from the textbook CH) CH Exercise 4.11 in page 126.

Consider the Scottish hills races data.

```
url = 'http://www.statsci.org/data/general/hills.txt'
races.table = read.table(url,
  header=TRUE, sep='\t')
```

Choose an observation index $i$ (e.g. $i = 33$, which corresponds to the outlying observation number 33) and create an dummy variable $U$, where all the values of $U$ are zero except for its i-th (33rd value) value which is one.

```
U = rep(0, nrow(races.table))
U[33] = 1
```

Now consider comparing the following models:

$$\mathrm{H}_0 : \mathrm{Time} = \beta_0 + \beta_1 \mathrm{Distance} + \beta_2 \mathrm{Climb} + \epsilon, \quad \text{say,} \quad \text{Model 1}$$

$$\mathrm{H}_0 : \mathrm{Time} = \beta_0 + \beta_1 \mathrm{Distance} + \beta_2 \mathrm{Climb} + \beta_3 U + \epsilon. \quad \text{say,} \quad \text{Model 2}$$

Let $t_i^* = \dfrac{\hat{Y}_i - \hat{Y}_{(i),i}}{\hat{\sigma}_{(i)}\sqrt{1 - H_{ii}}}$ be the i-th externally standardized residual (studentized residuals), where $\hat{\sigma}^2_{(i)}$ is the mean squared error of fit without the i-th observation, $H_{ii}$ is the i-th leverage value (That is the i-th diagonal element of $\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1} X^T$). Verify using the 33rd observation that

(1) The t-test statistic value for testing $\beta_3 = 0$ in Model 2 is the same as the i-th (33rd) externally standardized residual (studentized residuals) obtained from Model 1.

(2) The F-test statistics value for testing Model 1 versus Model 2 reduces to the square of the i-th (33rd) externally standardized residual (studentized residuals).

(3) Fit Model 1 to the Scottish hills races data without the i-th (33rd) observation.

(4) Show that the estimates of $\beta_0, \beta_1, \beta_2$ in Model 2 are the same as those obtained in part (3). [Hence adding an indicator variable for the i-th observation is equivalent to deleting the corresponding observation!]

## Question 3 (This question is based on lectures 10-18 and Chapters 3 and 4 from the textbook CH)

This question will review some of the fundamental concepts of the multiple linear regression model.

(1) Define the multiple linear regression model. [Use $Y$ as response and $X_1, \cdots, X_p$ as predictors and $n$ observations.]

(2) Write down the assumptions of multiple linear regression model.

(3) What is the **regression function** in the multiple linear regression model?

(4) What about the regression function makes this model a **linear** model?

(5) What function might you minimize to estimate parameters in your multiple linear regression model. [No need to give too detailed an algorithm.]

(6) Give an example of a regression function you might call **nonlinear**.