STAT 3008: Applied Regression Analysis
2019-20 Term 2
Assignment #3

**Revised (Prob 1(b), 1(d), Prob 3 negative AIC and BIC values)**

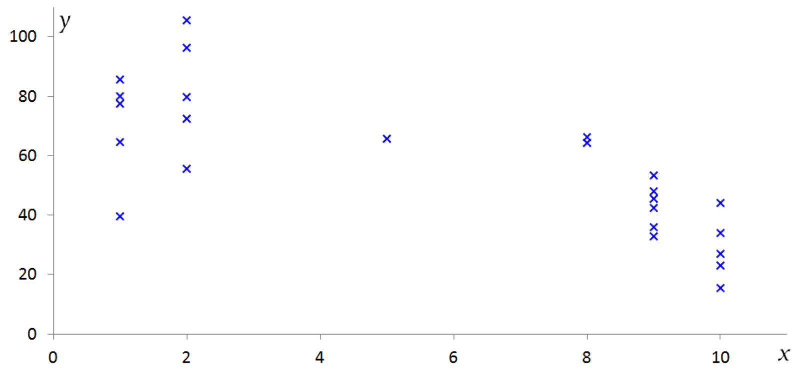
**Due: April 27th, 2020 (Monday) at 5:30pm [Extended from Apr 24th (Fri)]**

This assignment covers material from Chapter 5 to 6 of the lecture notes.

** Please submit the hardcopy of the R-code and R-outputs for Problem 2 and 4 (Quick and dirty is good enough, *R* markdown NOT recommended)

You need to show your calculation in details order to obtain full scores.

**Problem 1 [25 points]**: Consider the scatterplot below with data $\{(x_i, y_i), i = 1, 2, \ldots, 24\}$:




Suppose the data is fitted to a quadratic regression with mean function

$$E(Y \mid X = x) = \beta_0 + \beta_1 x + \beta_2 x^2$$

In matrix form, $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$ => $\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_{24} \end{pmatrix} = \begin{pmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ \vdots & \vdots & \vdots \\ 1 & x_{24} & x_{24}^2 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_{24} \end{pmatrix}$

Given that $\bar{x} = 5.833333$, $\bar{y} = 56.6275$, $\sum_{i=1}^{24} y_i^2 = \mathbf{Y'Y}$ = 89882.2642.

$$\mathbf{X'X} = \begin{pmatrix} 24 & 140 & 1164 \\ 140 & 1164 & 10568 \\ 1164 & 10568 & 98268 \end{pmatrix}, \quad (\mathbf{X'X})^{-1} = \begin{pmatrix} 0.450020 & -0.242619 & 0.020761 \\ -0.242619 & 0.167181 & -0.015105 \\ 0.020761 & -0.015105 & 0.001389 \end{pmatrix}, \quad \mathbf{X'Y} = \begin{pmatrix} 1359.06 \\ 6322.83 \\ 47452.65 \end{pmatrix}$$

(a) [6 points] Compute the OLS estimates for $\beta_o$, $\beta_1$ and $\beta_2$.

(b) [5 points] Compute the RSS and show that $\hat{\sigma} = 13.99$ ***(12.65 also acceptable)***.

(c) [5 points] Let $x^*$ be the optimal value of $x$ which maximizes the response y. What is the point estimate of $x^*$?

(d) [6 points] Construct an ANOVA table to test if $\beta_2 = \beta_1 = 0$ **[not $\beta_o = \beta_1 = 0$]**
*(No testing procedure is required, only the ANOVA table is sufficed).*

(e) [3 points] Suppose $x$ is an experimental EV. Before performing the experiment to obtain the response $y$, it's known that the optimal value of $x$ is somewhere in the middle of the interval [1.0, 10.0]. Briefly comment on whether the current choice of EV values $\{x_i$, $i = 1,2, \ldots, 24\}$ is reasonable.

**Problem 2 [33 points]**: The data '*salary*' from the alr3 library contains salary and other characteristics of all faculty members in a small Midwestern college in early 1980s. Below are the description of selected variables in the data file:

| Variable | Notation | Description |
|---|---|---|
| Sex | $\boldsymbol{S}$ | 1 = Female, 0 = Male |
| Rank | $\boldsymbol{R}$ | 1 = Assistant Professor, 2 = Associate Professor, 3 = Full Professor |
| Year | $\boldsymbol{X}$ | Number of years in current rank |
| Salary | $\boldsymbol{Y}$ | Annual salary (in US$) |

```
> library(car); library(alr3); S<-salary$Sex; R<-salary$Rank; X<-salary$Year; Y<-salary$Salary
```

Let $U_2$ and $U_3$ be the dummy variables for Associate Professor and Full Professor respectively.

(a) [8 points] Assume that the impact of the number of years in current rank (Year $\boldsymbol{X}$) is the same for different sex and ranks, we construct a linear model to explain the salary by the 3 other variables:

(Model 1) $E(Y \mid S=s, R=j, X=x) = \eta_0 + \eta_1 s + \beta x + \sum_{j=2}^{3}\left(\eta_{0j}U_j + \eta_{1j}U_j s\right)$

Compute the OLS estimates for the parameters $(\eta_0, \eta_1, \beta, \eta_{02}, \eta_{03}, \eta_{12}, \eta_{13}, \sigma^2)$ .

(b) [2 points] Suppose Mary received an offer as Assistant Professor from that college in early 1980s right after she received her PhD. Estimate the annual salary (in US$) offered by the college to her.

(c) [3 points] What is the RSS of the model in part (a)?

(d) [8 points] Construct an ANOVA table for the hypotheses on whether Rank is important to explain the Salary. That is,

$H_o$: $E(Y \mid S=s, R=j, X=x) = \eta_0 + \eta_1 s + \beta x$ vs

$H_1$: $E(Y \mid S=s, R=j, X=x) = \eta_0 + \eta_1 s + \beta x + \sum_{j=2}^{3}\left(\eta_{0j}U_j + \eta_{1j}U_j s\right)$

(e) [3 points] What are the (I) decision and (II) conclusion you would draw from the results in part (d)?

[Part (f) to (i)] Suppose ANOVA is used to test the hypothesis on whether salary for male and female are the same for all the 3 ranks in Model 1.

(f) [1 point] What is the mean function for $H_o$?

(g) [1 point] What is the mean function for $H_1$?

(h) [4 points] Construct the corresponding ANOVA table.

(i) [3 points] What are the (I) decision and (II) conclusion you would draw from the results in part (h)?

**Problem 3 [21 points] (Modified from Final Exam 2014-15 Term2)**: Consider a multiple linear regression with 4 terms: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + e$

The table below shows the AIC and BIC for models with different subsets of terms.

| Model | x1 | x2 | x3 | x4 | AIC | BIC |
|---|---|---|---|---|---|---|
| 1 | | | | | -68.1 | -65.5 |
| 2 | x | | | | -151.0 | -145.8 |
| 3 | | x | | | -121.8 | -116.6 |
| 4 | x | x | | | -609.1 | -601.3 |
| 5 | | | x | | -148.8 | -143.6 |
| 6 | x | | x | | -149.3 | -141.5 |
| 7 | | x | x | | -448.7 | -440.9 |
| 8 | x | x | x | | -608.2 | -597.8 |
| 9 | | | | x | -66.7 | -61.5 |
| 10 | x | | | x | -150.9 | -143.1 |
| 11 | | x | | x | -120.3 | -112.5 |
| 12 | x | x | | x | -7317.1 | -7306.7 |
| 13 | | | x | x | -148.1 | -140.3 |
| 14 | x | | x | x | -149.1 | -138.6 |
| 15 | | x | x | x | -459.5 | -449.1 |
| 16 | x | x | x | x | -7317.6 | -7304.6 |

**Revised AIC & BIC values at the table**: AIC and BIC should be negative in the table instead of positive in the original assignment, since BIC>AIC for each model.

*(For instance, **AIC = -121.8** for Model 3:* $y = \beta_0 + \beta_2 x_2 + e$

***BIC = -7304.6*** *for Model 16:* $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + e$ *)*

(a) [5 points] Implement the forward selection method using the AIC. Show your steps in details on how you come up with the parsimonious model.

(b) [10 points] Repeat part (a) if the backward selection method is implemented using the BIC.

(c) [2 points] What is the sample size $n$?

(d) [4 points] Do you think multicollinearity exist in Model 16? If so, identify the terms which are highly collinear with each other.

**Problem 4 [21 points]**: Consider the Berkeley Guidance Study data mentioned in Section 4.1. Suppose we want to model the height of girls at age 18 by 6 other variables taken at age 2 and age 9 (x1 to x6 in the R codes below):

```
library(alr3)
y<-BGSgirls$HT18      # height at age 18 (in cm)
x1<-BGSgirls$WT2      # weight at age 2 (in kg)
x2<-BGSgirls$HT2      # height at age 2 (in cm)
x3<-BGSgirls$WT9      # weight at age 9 (in kg)
x4<-BGSgirls$HT9      # height at age 9 (in cm)
x5<-BGSgirls$LG9      # leg circumference at age 9 (in cm)
x6<-BGSgirls$ST9      # strength at age 9 (in kg)
```

(a) [8 points] Based on the *stepAIC* function in *R (similar to those from Ch6 p26 and p32)*, show that the parsimonious model based on AIC is the same regardless of the (forward/backward) model selection methods

(b) [8 points] Repeat part (a) based on BIC. How do those parsimonious models differ from the one in part (a).

(c) [5 points] Note that leg circumference at age 9 (variable x5) is not included in the parsimonious model in part (a) because of multicollinearity. What is the value of variance inflation factor $VIF_5$ in the full model (i.e. model with all the 6 terms)?

- **End of the Assignment** -