Computer Project II

(Due on 04/07/2021 Tuesday by 11:59pm) in OneNote

We consider the 1992 baseball salary data set, , which is available from

`http://www.amstat.org/publications/jse/datasets/baseball.dat.txt`

This data set (of dimension 337×18) contains salary information (and performance measures) of 337 Major League Baseball players in 1992. More detailed information can be found at

`http://www.amstat.org/publications/jse/datasets/baseball.txt`

The data set contains the following variables.

Table 1: Variable Description for the 1992 Baseball Salary Data

Var	Columns	Description
`salary`	1 – 4	Salary (in thousands of dollars)
X_1	6 – 10	Batting average
X_2	12 – 16	On-base percentage (OBP)
X_3	18 – 20	Number of runs
X_4	22 – 24	Number of hits
X_5	26 – 27	Number of doubles
X_6	29 – 30	Number of triples
X_7	32 – 33	Number of home runs
X_8	35 – 37	Number of runs batted in (RBI)
X_9	39 – 41	Number of walks
X_{10}	43 – 45	Number of strike-outs
X_{11}	47 – 48	Number of stolen bases
X_{12}	50 – 51	Number of errors
X_{13}	53	Indicator of "free agency eligibility"
X_{14}	55	Indicator of "free agent in 1991/2"
X_{15}	57	Indicator of "arbitration eligibility"
X_{16}	59	Indicator of "arbitration in 1991/2"
`ID`	61 – 79	Player's name (in quotation marks)

The data set can be input into R by reading directly from the website, with the following R commands:

```
baseball <- read.table(file=
    "http://www.amstat.org/publications/jse/datasets/baseball.dat.txt",
    header = F,
    col.names=c("salary", "x1", "x2", "x3", "x4", "x5",
        "x6", "x7","x8", "x9", "x10", "x11", "x12", "x13",
        "x14", "x15", "x16", "ID"))
baseball$logsalary <- log(baseball$salary);
baseball <- baseball[, -c(1, 18)] # REMOVE salary AND ID
head(baseball)
```

Complete the project by following the specific instructions given below.

1. Perform EDA on the data (exclude: $x13 - x16$) and describe any interesting observation.

2. Starting with the whole model that includes all predictors (i.e., X_1, $X_2, \ldots, X_{16}$), apply one model selection procedure of your choice to select your *best* model. Provide the fitting results from your 'best' model, i.e., the table of Parameter Estimates and the ANOVA table.

3. Suppose that we are interested in the following model:

$$\textbf{Model I:} \quad \log(\texttt{salary}) = \beta_0 + \beta_1 x_{13} + \beta_2 x_{15} + \beta_3 x_3 + \beta_4 x_4 + \varepsilon. \tag{1}$$

   Fit **Model I** in (1) and output the two tables: Table of Parameter Estimates and the ANOVA table.

4. Use BIC to compare the *best* model that you found with **Model I** in (1). Which one is better according to BIC? (Hint: May use the R function `BIC`.)

5. Perform a test of $H_0 : \ \beta_3 = \beta_4 = 0$ in **Model I**.

6. Given a player who has 80 runs ($X_3 = 80$) and 120 hits ($X_4 = 120$) in total and is eligible for arbitration ($X_{15} = 1$), but not free agency yet ($X_{13} = 0$), provide a 95% prediction interval for his salary based on **Model I**.

Some helpful tips for computer projects are listed below:

- Start early and don't wait till the last day/minute;
- Create a table for all output presented.Use copy-and-paste appropriately to include necessary R output into your final report;
- Remember to interpret every result that you present; •

Place your R codes in an appendix.

- Submit your report and code in OneNote.