STA 207 HW-7

Due Date: 11/21 by 10:20AM

Problem: Major League Baseball (MLB) data

MLBStandings2016 is the Major League Baseball (MLB) standings and team statistics for the 2016 season. Data for all 30 Major League Baseball (MLB) teams for the 2016 regular season. This data includes team batting statistics (BattingAvg through SLG) and team pitching statistics (ERA through WHIP).

The variables in the data are:

- Team: Team name
- League:AL=American or NL=National
- Wins:Number of wins for the season (out of 162 games)
- Losses:Number of losses for the season
- WinPct:Proportion of games won
- BattingAverage:Team batting average
- Runs:Number of runs scored
- Hits:Number of hits
- HR:Number of home runs hit
- Doubles:Number of doubles hit
- Triples:Number of triples hit
- RBI:Number of runs batted in
- SB:Number of stolen bases
- OBP:On base percentage
- SLG:Slugging percentage
- ERA:Earned run average (earned runs allowed per 9 innings)
- HitsAllowed:Number of hits against the team
- Walks:Number of walks allowed
- StrikeOuts:Number of strikeouts (by the team's pitchers)
- Saves:Number of games saved (by the team's pitchers)

- WHIP:Number of walks and hits per inning pitched

To get the data in R, do the following:

```
library(Stat2Data)

data(MLBStandings2016)

attach(MLBStandings2016)

names(MLBStandings2016)
```

```
##  [1] "Team"          "League"        "Wins"          "Losses"
##  [5] "WinPct"        "BattingAverage" "Runs"         "Hits"
##  [9] "HR"            "Doubles"       "Triples"       "RBI"
## [13] "SB"            "OBP"           "SLG"           "ERA"
## [17] "HitsAllowed"   "Walks"         "StrikeOuts"    "Saves"
## [21] "WHIP"
```

Answer the following questions:

a) (10 points) Make scatterplot matrix and correlation matrix for WinPct , ERA, BattingAverage, Runs, and Hits. Discuss the relationship between each pair of variables.
b) (5 points) Regressing WinPct on the four predictors ERA, BattingAverage, Runs, and Hits and report the fitted model.
c) (5 points) Is there multi-collinearity present in the above model.
d) (5 points) Report ANOVA of the above model.
e) (5 points) Show all steps for testing the overall fit of the regression model.
f) (5 points) Let a reduced model-1 be the one with Hits predictor removed from the original model. Report the fitted model.
g) (5 points) Is there multi-collinearity present in the reduced model-1.
h) (5 points) Report ANOVA for the reduced model-1.
i) (5 points) Show all steps for testing the overall fit for reduced model-1.
j) (5 points) Let a reduced model-2 be the one with Hits and Batting Average predictors removed from the original model. Report the fitted model.
k) (5 points) Is there multi-collinearity present in the reduced model-2. (5 points) Report ANOVA for reduced model-2.
l) (5 points) Show all steps for testing the overall fit for reduced model-2.
m) (10 points) Show hypothesis test comparing reduced model-1 with the original (full) model.
n) (10 points) Show hypothesis test comparing reduced model-1 with the original (full) model.
o) (10 points) Compare original model, reduced model-1 and reduced model-2 using the following:

- Adjusted R^2 ,
- Residual Standard Errors $\hat{\sigma}$
- AIC, AICc, and BIC

Data Analysis Project (One per group) For the data analysis project, do the following:

a.) (10 points) Report the dataset you are using with your research goals. Explain response and predictor variables.

b.) (10 points) Show the scatterplot and correlation matrices. Discuss the linear relation between all pair of variables.

c.) (10 points) Report the fitted MLR model. Which predictors are statistically significant and why?

d.) (5 points) Report the goodness of fit of this model.

e.) (15 points) Test and report the LINE conditions.