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) (20 points) Make scatterplot matrix and correlation matrix for WinPct, ERA, BattingAverage, Runs, and Hits. Discuss the relationship between each pair of variables.

The scatterplot matrix in Fig. 1 and correlation matrix in Fig. 2 below show that the proportion of games won has a

- negative, linear, somewhat strong relation with the earned runs allowed per 9 innings, there are no obvious outliers but there is a slight curve (non-linear) pattern
- positive, moderate, linear relationship between with team's batting average, there are some potential outliers and there is a curve too
- positive, moderate, linear relation with the number of runs, there are some potential outliers but no curve
- positive, weak, linear relation with the number of hits, there are some potential outliers and there might be a slight curve too

Looking at other pair of variables, we see that

- earned runs allowed per 9 innings has very low, positive, linear relation with team's batting average and number of hits, but it has a very low, negative, linear relation with the number of runs. There are some potential outliers, but no curve
- team's batting average has a moderate, positive linear relation with the number of runs, and strong, positive relation with the number of hits. There are no obvious outliers and no curve.
- Number of runs has a positive, moderate, linear relation with the number of hits. There are no obvious outliers and no curve.

Fig.1 Matrix Scatterplot

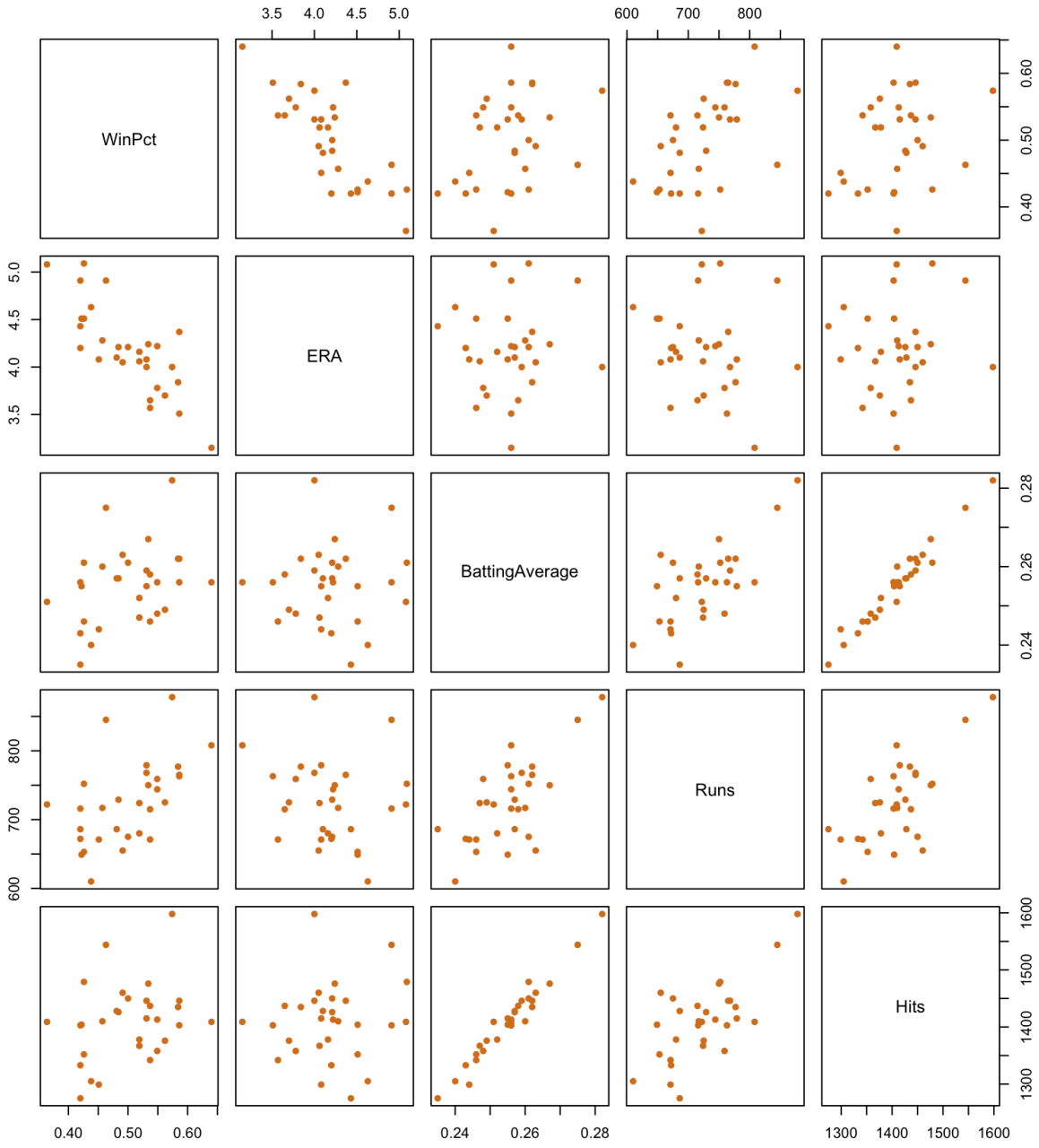
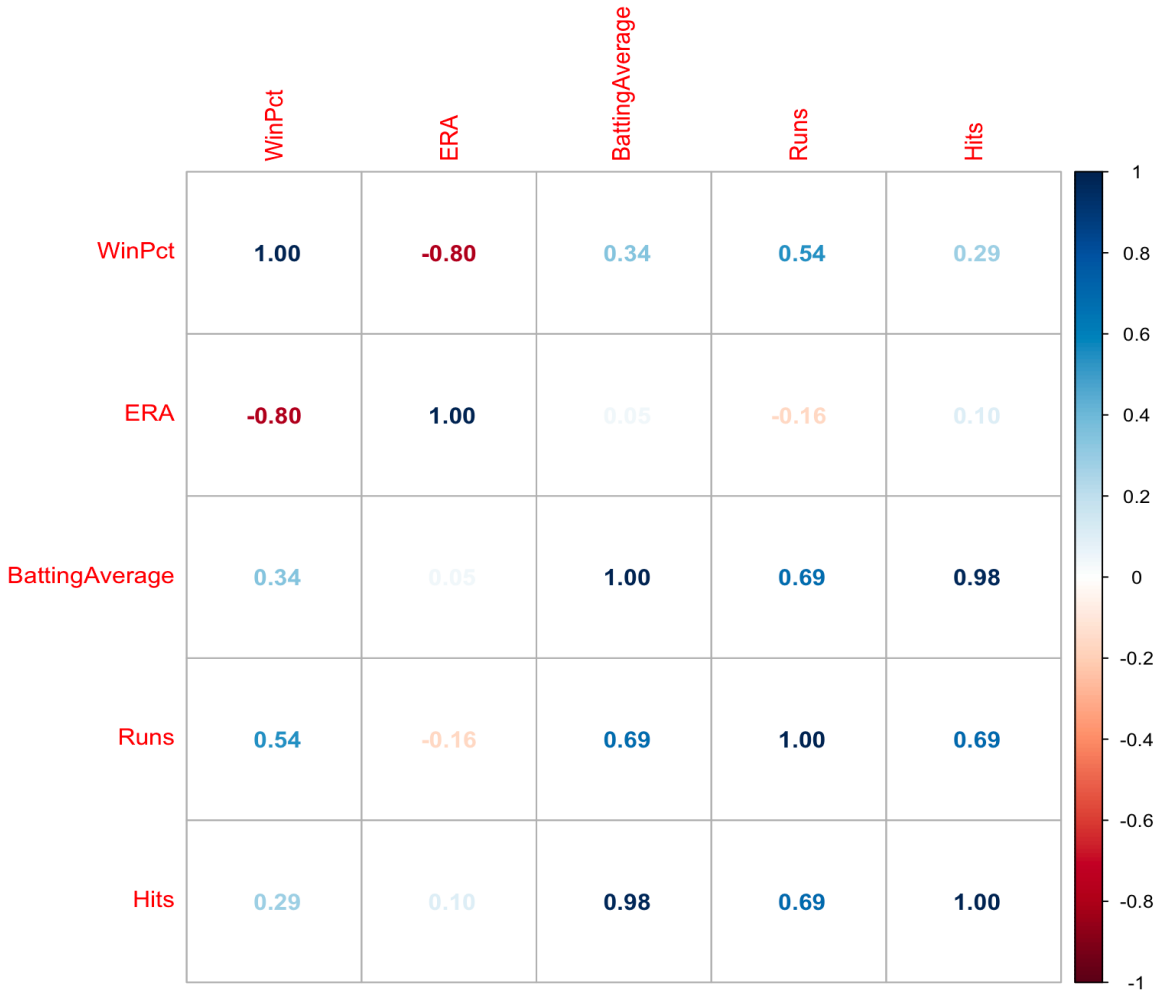


Fig. 2 Correlation Matrix



b) (5 points) Regressing WinPct on the four predictors ERA, BattingAverage, Runs, and Hits and report the fitted model.

The fitted model is

$\hat{Y} = 0.362 - 0.108X1 + 2.381X2 + 0.0003X3 - 0.0002X4$, where

Y WinPct is the proportion of games won and predictor variables,

X1: ERA: Earned run average (earned runs allowed per 9 innings),

X2: Batting Average: Team batting average,

X3: Runs: Number of runs scored,

X4: Hits: Number of hits.

c) (5 points) Is there multi-collinearity present in the above model.

From Fig. 1 and 2, we see that hits and batting average are highly correlated so these two might have multicollinearity. The number of hits and runs also show a moderate, linear relation, so this might also lead to multi-collinearity. Using the variance inflation factor,

ERA	BattingAverage	Runs	Hits
1.218878	25.270853	2.164131	26.903838

We see that batting average and number of hits show multi-collinearity.

d) (5 points) Report ANOVA of the above model.

Table-1: ANOVA for the full model.

Source of Variation	**Df**	**SS**	**MS**	**F**	**P-value**
Regression	4	0.1049829	0.0262455	29.79058	0.000000003479
Error	25	0.022018	0.000881		
Total	29	0.127001			

e) (5 points) Show all steps for testing the overall fit of the regression model.

H_0: $\beta_1 = 0, \beta_2 = 0, \beta_3 = 0, \beta_4 = 0$ (all four predictors are not statistically significant)

H_1: $\beta_j \neq 0, for\ atleast\ one\ j, where\ j = 1,2,3,4$ (at least one of the four predictors are statistically significant)

The default level of significant $\alpha = 0.05$

$F = 29.8$

P-value=3.479e-09

P-value < α → Reject H_0

Conclusion: At least one of the earned runs allowed per 9 innings, team batting average, number of runs scored, and number of hits is statistically significant in predicting or explaining variance in the proportion of games won and predictor variables.

f) (5 points) Let a reduced model-1 be the one with Hits predictor removed from the original model. Report the fitted model.

The fitted reduced model-1 is

$$\hat{Y} = 0.426 - 0.110X1 + 1.144X2 + 0.0003X3$$

Where Y, X2, and X3 are as defined above.

g) (5 points) Is there multi-collinearity present in the reduced model-1.

Reduced Model-1

The variance inflation factors are:

ERA	BattingAverage	Runs
1.077534	2.000594	2.049241

No multicollinearity expected from Fig. 1 and 2 and above variance inflation factors confirm that.

h) (5 points) Report ANOVA for the reduced model-1.

Table-2: ANOVA for the Reduced Model-1.

Source of Variation	**Df**	**SS**	**MS**	**F**	**P-value**
Regression	3	0.1047991	0.03493304	40.91	0.000000000545
Error	26	0.022202	0.000853917		
Total	29	0.127001			

h) (5 points) Show all steps for testing the overall fit for reduced model-1.

$H_0: \beta_1 = 0, \beta_2 = 0, \beta_3 = 0$ (all threepredictors are not statistically significant)

$H_1: \beta_j \neq 0, for\ atleast\ one\ j, where\ j = 1,2,3$ (at least one of the three predictors are statistically significant)

The default level of significant $\alpha = 0.05$

$F = 40.91$

P-value=0.000000000545

P-value < α → Reject H_0

Conclusion: At least one of the earned runs allowed per 9 innings, team batting average, and number of runs scored is statistically significant in predicting or explaining variance in the proportion of games won and predictor variables.

i) (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.
The fitted reduced model-2 is
$\hat{Y} = 0.604 - 0.106X1 + 0.0005X3,$
Where Y, X1, and X3 are defined above.

j) (10 points) Is there multi-collinearity present in the reduced model-2. (5 points) Report ANOVA for reduced model-2.

For Reduced Model-2, the variance inflation factors are:

ERA	Runs
1.02641	1.02641

No multicollinearity expected from Fig. 1 and 2, and the VIFs.

Table-3: ANOVA for the Reduced Model-2.

Source of Variation	**Df**	**SS**	**MS**	**F**	**P-value**
Regression	2	0.1029736	0.0514868	57.86	0.000000000173
Error	27	0.024027	0.000890		
Total	29	0.127001			

k) (5 points) Show all steps for testing the overall fit for reduced model-2.

H_0: $\beta_1 = 0, \beta_3 = 0$ (both the predictors are not statistically significant)

H_1: $\beta_j \neq 0, for\ atleast\ one\ j, where\ j = 1,3$ (at least one of the two predictors is statistically significant)

The default level of significant $\alpha = 0.05$

$F = 57.86$

P-value=0.000000000173

P-value > α →Reject H_0

Conclusion: At 5% level of significance, at least of og Hits and Batting Average is statistically significant predictor for WinPct.

l) (20 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

Table-3 Comparing Full Model with Two Reduced Models. The red values shows the best model picked according to the specific criterion.

Model	R^2_{adj}	$\hat{\sigma}$	AIC	AICc	BIC
Full model	0.7989	0.02968	-3.979213	-5.695351	-3.698973
Reduced Model-1	0.805	0.02922	-4.037567	-5.792111	-3.804034
Reduced Model-2	0.7968	0.02983	-4.025216	-5.80976	-3.83839
Criteria	Highest is better	Lowest is better	Lowest is better	Lowest is better	Lowest is better
Best Model	Reduced Model-1	Reduced Model-1	Reduced Model-1	Reduced Model-2	Reduced Model-2