STA 207 HW-7

Due Date: 11/21 by 10:20AM

Derin Gezgin | Camel ID: 00468038

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

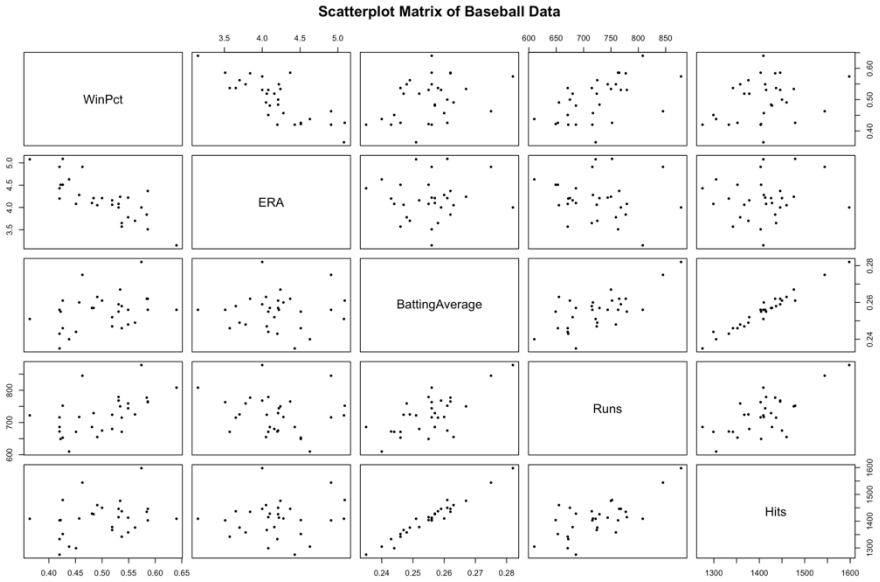
To get the data in R, do the following:

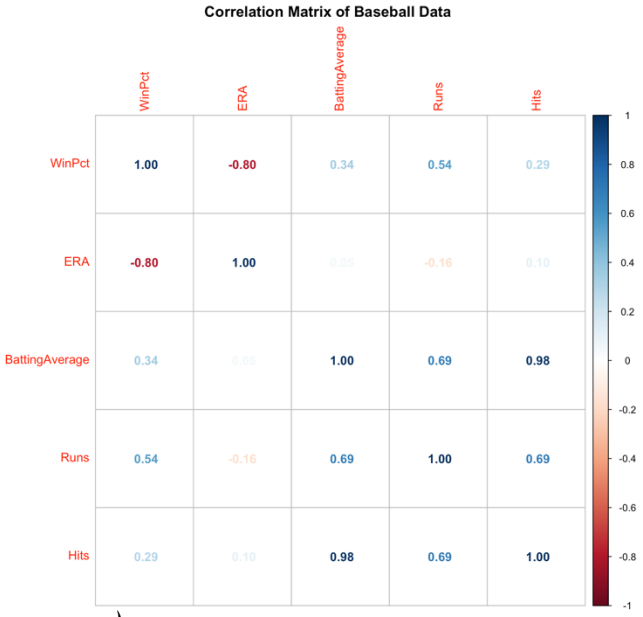
library(Stat2Data) | data(MLBStandings2016) | attach(MLBStandings2016) | names(MLBStandings2016)

```
##  [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.





Relationship between WinPct and ERA (Earned run average)

We can see the correlation coefficient value between the *WinPct* and *ERA (Earned run average)* is *-0.80*. This shows a moderately strong relationship between WinPct and ERA (Earned run average) in the negative direction. When we check the scatterplot matrix, we can also see the moderately strong linear relationship which strongly shows a downward trend.

We can conclude that there is a moderately strong relationship in the negative direction between WinPct and ERA (Earned run average).

Relationship between WinPct and BattingAverage

We can see the correlation coefficient value between the *WinPct* and *BattingAverage is 0.34*. This is a very low correlation coefficient value and is not enough to conclude that there is a relationship between these variables. Similarly, when we checked the scatterplot matrix, we cannot see any clear relationship between the variables.

We can conclude that there is not a significant relationship between WinPct and BattingAverage.

Relationship between WinPct and Runs

We can see the correlation coefficient value between the *WinPct* and *Runs* is 0.54. This is a very borderline value to be classified as a weak-moderate relationship in the positive direction. We can check the scatterplot to have a better assessment. In the scatterplot matrix, we can see that the relationship is not really strong but we can see a very slight relationship. It is more than what we had in WinPct v. Runs

We can conclude that there is not a significant relationship between WinPct v. Runs. But we can say that there might be a weak-moderate relationship in the positive direction.

Relationship between WinPct and Hits

We can see the correlation coefficient value between the *WinPct* and *Hits* is 0.29. As this is a very insignificant and small number, we can say that there is no relationship between WinPct and Hits. In the scatterplot matrix, we can also see that there is no relationship at all between WinPct and Hits.

We can conclude that there is no relationship between WinPct and Hits.

Relationship between ERA (Earned run average) and BattingAverage

We can see the correlation coefficient value between the *ERA (Earned run average)* and *BattingAverage* is 0.05. This is a near-0 value which shows that there is no relationship between these variables. In the scatterplot, we can see that there is no visible pattern at all and it seems like a random distribution.

We can conclude that there is no significant relationship between ERA (Earned run average) and BattingAverage.

Relationship between ERA (Earned run average) and Runs

We can see the correlation coefficient value between the *ERA (Earned run average)* and *Runs* is -0.16. This is a near-0 value which shows that there is no relationship between these variables. In the scatterplot, there is no visible pattern and the points are mostly randomly distributed.

We can conclude that there is no significant relationship between ERA (Earned run average) and Runs

Relationship between ERA (Earned run average) and Hits

We can see the correlation coefficient value between the *ERA (Earned run average)* and *Hits* is 0.10. This is a near-0 value which shows that there is no relationship between these variables. In the scatterplot, there is no visible pattern and the points are mostly randomly distributed.

We can conclude that there is no significant relationship between ERA (Earned run average) and Hits.

Relationship between BattingAverage and Runs

We can see the correlation coefficient value between the *BattingAverage* and *Runs* is 0.69. This is a value over 0.5 and the value is in an area that can be considered as a moderate relationship in the positive direction. In the scatterplot, we can see that there is a moderate relationship between these variables.

We can conclude that there is a moderately strong relationship -in the positive direction- between BattingAverage and Runs

Relationship between BattingAverage and Hits

We can see the correlation coefficient value between the *BattingAverage* and *Runs* is 0.98. This a value extremely close to 1 which shows that there is an extremely strong relationship in the positive direction between BattingAverage and Hits. When we check the scatterplot matrix, we can see that the scatterplot of BattingAverage v. Hits also support this correlation coefficient value.

We can conclude that there is an extremely strong relationship in the positive direction, between BattingAverage and Hits.

Relationship between Runs and Hits

We can see the correlation coefficient value between the *Runs* and *Hits* is 0.69. This is a value over 0.5 and the value is in an area that can be considered as a moderate relationship in the positive direction. In the scatterplot, we can see that there is a moderate relationship between these variables.

We can conclude that there is a moderately strong relationship -in the positive direction- between Runs and Hits

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

The fitted model is: $\widehat{\boldsymbol{Y}} = \mathbf{0.362} - \mathbf{0.108}\boldsymbol{x_1} + \mathbf{2.381}\boldsymbol{x_2} + \mathbf{0.0003}\boldsymbol{x_3} - \mathbf{0.0002}\boldsymbol{x_4}$

Where;

Y: Proportion of games won

x_1: Earned run average (earned runs allowed per 9 innings) | x_2: Team batting average

x_3: Number of runs scored | x_4: Number of hits

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

"I am not going to add the scatterplot and correlation matrices to save from space. I am directly referring to those matrices on part a."

From the scatterplot and correlation matrices, and the Variance Inflation Factor (VIF) values we can check the predictor variables individually;

Earned Run Average (ERA)

We can see that earned run average (ERA) and team batting average has a correlation coefficient of 0.05 which shows nearly no correlation at all. We can see a similar result in the Earned Run Average (ERA) v. team batting average scatterplot.

We can see that earned run average (ERA) and number of runs scored has a correlation coefficient of -0.16 which also shows nearly no correlation between these variables. We can see a similar result in the Earned Run Average (ERA) v. Runs scatterplot where there is no significant relationship at all.

Lastly, we can examine the relationship between earned run average (ERA) and number of hits. When we check the correlation matrix, we can see the correlation coefficient between these variables is 0.10 which again shows nearly no correlation at all. Similarly, in the scatterplot, we can also see a similar result.

The VIF value for earned run average (ERA) is 1.219. As the VIF value is between 1 and 4, this shows a slightly moderate multicollinearity between Earned Run Average and other predictor variables (Team batting average, number of runs scored, number of hits).

Considering the cross-check of scatterplots and correlation coefficient values between the Earned Run Average and other predictor variables, and the VIF value, we can conclude that there is a moderate correlation. But this is not enough to warrant corrective measures.

Team batting average

We can see that team batting average and earned run average (ERA) has a correlation coefficient of 0.05 which shows nearly no correlation at all. We can see a similar result in the team batting average v. Earned Run Average (ERA) scatterplot.

We can see that team batting average and number of runs scored has a correlation coefficient of 0.69. This is a value over 0.5 which can be considered as a moderate relationship in the positive direction. The scatterplot of team batting average v. number of runs scored also shows us a similar result.

We can see that team batting average and number of hits, has a correlation coefficient value of 0.98 which is a sign of an extremely strong relationship in the positive direction. We can also see this relationship in the scatterplot of Team Batting Average v. number of hits scatterplot.

The VIF value for team batting average is 25.271. As the VIF value is between significantly larger 4, this shows a strong multicollinearity between team batting average and other predictor variables (earned run average (ERA), number of runs scored, number of hits).

Considering the cross-check of scatterplots and correlation coefficient values between the Team Batting Average and other predictor variables, and the VIF value, we can conclude that there is a strong correlation. This shows a problem of multi-correlation. We can say that coefficients are poorly estimated, and the p-values are questionable.

Number of runs scored

We can see that number of runs scored and earned run average (ERA) has a correlation coefficient of -0.16 which also shows nearly no correlation between these variables. We can see a similar result in the Runs v. Earned Run Average (ERA) scatterplot where there is no significant relationship at all.

We can see that number of runs scored and team batting average has a correlation coefficient of 0.69. This is a value over 0.5 which can be considered as a moderate relationship in the positive direction. The scatterplot of number of runs scored v. team batting average also shows us a similar result.

Lastly, we can see the number of runs scored and number of hits has a correlation coefficient of 0.69. This is a value over 0.5 which can be considered as a moderate relationship in the positive direction. The scatterplot of number of runs scored v number of hits also shows us a similar result.

The VIF value for number of runs scored is 2.164. As the VIF value is between 1 and 4, this shows a slightly moderate multicollinearity between Number of runs scored and other predictor variables (Earned Run Average (ERA), Team batting average, number of hits).

Considering the cross-check of scatterplots and correlation coefficient values between the Number of Runs Scored and other predictor variables, and the VIF value, we can conclude that there is a moderate correlation. But this is not enough to warrant corrective measures.

Number of hits

When we check the correlation matrix, we can see the correlation coefficient between number of hits and earned run average (ERA) variables is 0.10 which again shows nearly no correlation at all. Similarly, in the scatterplot of number of hits v. earned run average (ERA), we can also see a similar pattern.

When we check the correlation matrix, we can see the correlation coefficient between number of hits and team batting average is 0.98 which is a sign of an extremely strong relationship in the positive direction. We can also see this relationship in the scatterplot of Number of Hits v. Team Batting Average.

Lastly, we can see the number of hits and number of runs scored has a correlation coefficient of 0.69. This is a value over 0.5 which can be considered as a moderate relationship in the positive direction. The scatterplot of number of hits v. number of runs scored also shows us a similar result.

The VIF value for number of hits is 26.904. As the VIF value is between significantly larger 4, this shows a strong multicollinearity between number of hits and other predictor variables (earned run average (ERA), team batting average, number of runs scored).

Considering the cross-check of scatterplots and correlation coefficient values between the Number of Hits and other predictor variables, and the VIF value, we can conclude that there is a strong correlation. This shows a problem of multi-correlation. We can say that coefficients are poorly estimated, and the p-values are questionable.

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

Source of Variation	Degrees of Freedom (DF)	Sum of Squares (SS)	Mean Squares (MS)	F-Statistic	P-Value
Regression	4	0.105	0.026	29.80005	~0
Error	25	0.022	0.000881		
Total	29	0.127	0.004		

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

H_0: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$

H_a: At least one of $B_j \neq 0$ for $j = 1,2,3,4$

$\alpha = 0.05$

F = 29.8 ~ $F_{4,25}$

P-Value = 0.00000000347 ~ 0

P-Value ≤ 0.05 → Reject H_0

Conclusion

At a 5% significance level, at least one of the 4 predictors above explain significant amount of variability in Y.

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 model is: $\widehat{Y} = 0.426 - 0.1099x_1 + 1.144x_2 + 0.0003x_3$

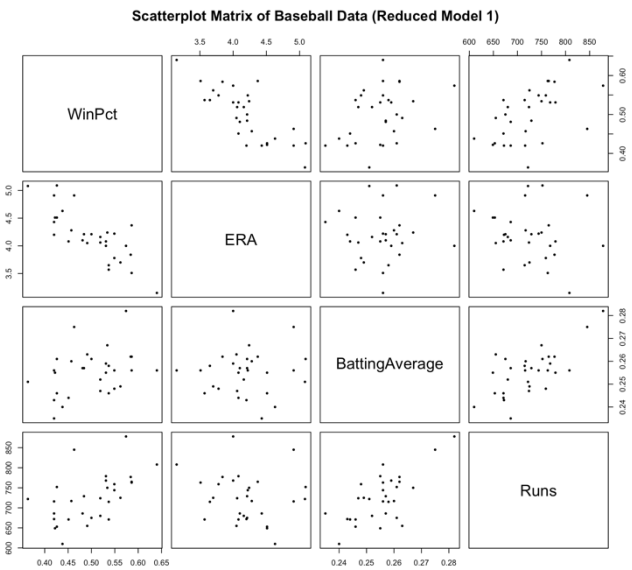
Where;

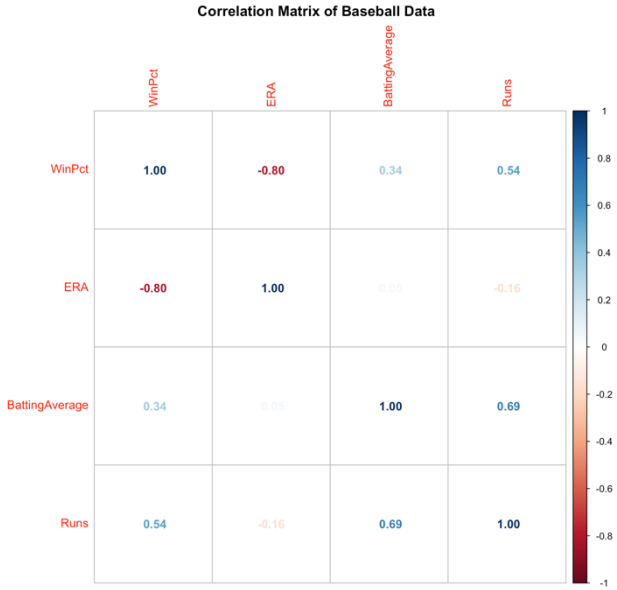
Y: Proportion of games won

x_1: Earned run average (earned runs allowed per 9 innings) | x_2: Team batting average

x_3: Number of runs scored

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





From the scatterplot and correlation matrices, and the Variance Inflation Factor (VIF) values we can check the predictor variables individually;

Earned Run Average (ERA)

We can see that earned run average (ERA) and team batting average has a correlation coefficient of 0.05 which shows nearly no correlation at all. We can see a similar result in the Earned Run Average (ERA) v. team batting average scatterplot.

We can see that earned run average (ERA) and number of runs scored has a correlation coefficient of -0.16 which also shows nearly no correlation between these variables. We can see a similar result in the Earned Run Average (ERA) v. Runs scatterplot where there is no significant relationship at all.

The VIF value for earned run average (ERA) is 1.078. As the VIF value is between 1 and 4, this shows a slightly moderate multicollinearity between Earned Run Average and other predictor variables (Team batting average, number of runs scored).

Considering the cross-check of scatterplots and correlation coefficient values between the Earned Run Average and other predictor variables, and the VIF value, we can conclude that there is a moderate correlation. But this is not enough to warrant corrective measures.

Team batting average

We can see that team batting average and earned run average (ERA) has a correlation coefficient of 0.05 which shows nearly no correlation at all. We can see a similar result in the team batting average v. Earned Run Average (ERA) scatterplot.

We can see that team batting average and number of runs scored has a correlation coefficient of 0.69. This is a value over 0.5 which can be considered as a moderate relationship in the positive direction. The scatterplot of team batting average v. number of runs scored also shows us a similar result.

The VIF value for team batting average is 2.001. As the VIF value is between 1 and 4, this shows a slightly moderate multicollinearity between Team batting average and other predictor variables (Earned Run Average (ERA), number of runs scored).

Considering the cross-check of scatterplots and correlation coefficient values between the Earned Run Average and other predictor variables, and the VIF value, we can conclude that there is a moderate correlation. But this is not enough to warrant corrective measures.

Number of runs scored

We can see that number of runs scored and earned run average (ERA) has a correlation coefficient of -0.16 which also shows nearly no correlation between these variables. We can see a similar result in the Runs v. Earned Run Average (ERA) scatterplot where there is no significant relationship at all.

We can see that number of runs scored and team batting average has a correlation coefficient of 0.69. This is a value over 0.5 which can be considered as a moderate relationship in the positive direction. The scatterplot of number of runs scored v. team batting average also shows us a similar result.

The VIF value for number of runs scored is 2.049. As the VIF value is between 1 and 4, this shows a slightly moderate multicollinearity between Team batting average and other predictor variables (Earned Run Average (ERA), number of runs scored).

Considering the cross-check of scatterplots and correlation coefficient values between the Earned Run Average and other predictor variables, and the VIF value, we can conclude that there is a moderate correlation. But this is not enough to warrant corrective measures.

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

Source of Variation	Degrees of Freedom (DF)	Sum of Squares (SS)	Mean Squares (MS)	F-Statistic	P-Value
Regression	3	0.105	0.0349	40.908	~0
Error	26	0.0222	0.000847		
Total	29	0.127	0.00437		

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

H_0: $\beta_1 = \beta_2 = \beta_3 = 0$

H_a: At least one of $B_j \neq 0$ for $j = 1,2,3$

$\alpha = 0.05$

F = 40.908 ~ $F_{3,26}$

P-Value = 0.000000000545 ~ 0

P-Value ≤ 0.05 → Reject H_0

Conclusion

At a 5% significance level, at least one of the 3 predictors above explain significant amount of variability in Y.

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.

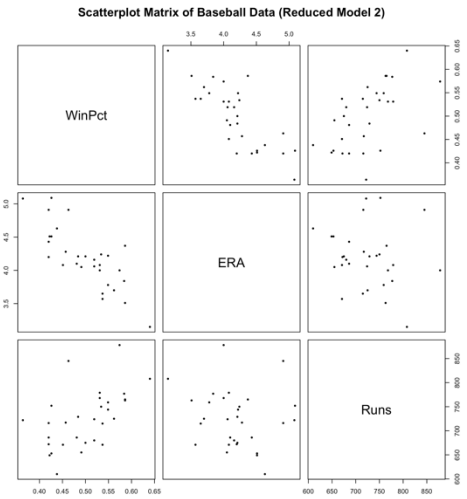
The fitted model is: $\hat{Y} = 0.604 - 0.106x_1 + 0.000468x_2$

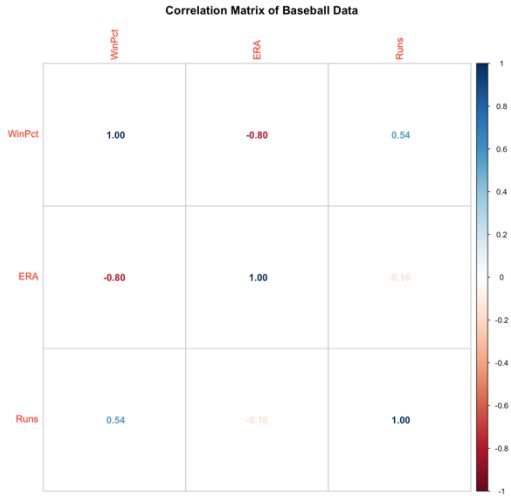
Where;

Y: Proportion of games won

x_1: Earned run average (earned runs allowed per 9 innings) | x_2: Number of runs scored

k) (5 points) Is there multi-collinearity present in the reduced model-2.





From the scatterplot and correlation matrices, and the Variance Inflation Factor (VIF) values we can check the predictor variables individually;

Earned Run Average (ERA)

We can see that earned run average (ERA) and number of runs scored has a correlation coefficient of -0.16 which also shows nearly no correlation between these variables. We can see a similar result in the Earned Run Average (ERA) v. Runs scatterplot where there is no significant relationship at all.

The VIF value for earned run average (ERA) is 1.026. As the VIF value is between 1 and 4, this shows a very slight moderate multicollinearity between Earned Run Average and other predictor variables (number of runs scored).

Considering the cross-check of scatterplots and correlation coefficient values between the Earned Run Average and other predictor variables, and the VIF value, we can conclude that there is a (very slight) moderate correlation. But this is not enough to warrant corrective measures. Also, the value is really close to 1 which is our threshold.

Number of runs scored

We can see that number of runs scored and earned run average (ERA) has a correlation coefficient of -0.16 which also shows nearly no correlation between these variables. We can see a similar result in the Runs v. Earned Run Average (ERA) scatterplot where there is no significant relationship at all.

The VIF value for number of runs scored is 1.026. As the VIF value is between 1 and 4, this shows a very slight moderate multicollinearity between number of runs scored and other predictor variables (number of runs scored).

Considering the cross-check of scatterplots and correlation coefficient values between the Number of runs scored and other predictor variables, and the VIF value, we can conclude that there is a (very slight) moderate correlation. But this is not enough to warrant corrective measures. Also, the value is really close to 1 which is our threshold.

l) (5 points) Report ANOVA for reduced model-2.

Source of Variation	Degrees of Freedom (DF)	Sum of Squares (SS)	Mean Squares (MS)	F-Statistic	P-Value
Regression	2	0.103	0.0515	57.857	~0
Error	27	0.024	0.00089		
Total	29	0.127	0.00438		

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

H_0: $\beta_1 = \beta_2 = 0$

H_a: At least one of $B_j \neq 0$ for $j = 1,2$

$\alpha = 0.05$

F = 57.857 ~ $F_{2,27}$

P-Value = 0.000000000173 ~ 0

P-Value ≤ 0.05 → Reject H_0

Conclusion

At a 5% significance level, at least one of the 2 predictors above explain significant amount of variability in Y.

~~**n) (10 points) Show hypothesis test comparing reduced model-1 with the original (full) model.**~~

~~**o) (10 points) Show hypothesis test comparing reduced model-2 with the original (full) model.**~~

p) (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**

Model	AIC	AICc	BIC	R^2_{adj}	RSE ($\hat{\sigma}$)
Model-0	-3.979	-5.695	-3.699	0.799	0.0297
Model-1	-4.038	-5.792	-3.804	0.805	0.0292
Model-2	-4.025	-5.81	-3.838	0.797	0.0298

In this case, we would look for smallest AIC, AICc, RSE and maximum BIC, and RSE. I highlighted the best values for each class.