Variable Selections through Regularizations: Study of Baseball Attendance

University of Minnesota

Tae Uk You

Introduction

Major League Baseball (MLB) is facing a problem. MLB ended 2018 regular season with total attendance of 69.97 million. This total is down 4.1 percent from previous year and marks the league's fifth decline in the last six seasons. This figure also is the lowest since 2013 and the first to fall below 70 million. Tayler (2018) indicates that average ticket prices have jumped every year for the last decade, while baseball games have been much more accessible away from the ballpark via phone, tablet or laptop. The goal of this project is to indicate which factors attribute the most for the number of attendance to predict attendance. Minnesota Twins and its home stadium, Target Field, is selected for the study. In this project, three statistical models are used, which are Ridge regression, Lasso regression, and Elastic net regression.

This paper proceeds as follows. In Literature Review, substantive findings as well as theoretical and methodological contributions to this topic in prior studies are briefly introduced. Studies about the ridge, lasso, and elastic net regression are briefly discussed. In Data Collection and Methodology section, process of data collection, occurrence of multicollinearity in the multiple regression model and its regularization through Ridge, Lasso, and Elastic net regression are described. Detailed interpretation about the regression results appear in the Data Analysis and Results section, and Discussion section provides a summary of this study, limitations, and discussions about future study.

Literature Review

Numerous studies have examined the level of demand for a game of baseball inferred from attendance numbers. Humphreys (2002) and Lee and Fort (2008) examined the variability in total league attendance by using long time-series dataset based on assumption that time-trend variables regulate other determinant's variability. Davis (2008) and Mills and Fort (2018) evaluated team-level attendance time series data. However, the method of time-series attendance analysis has limitation that is not possible to include the team-specific characteristics that could explain variations in attendance.

Anh and Lee (2014) estimated attendance function by using a panel factor model and found that the determinants of attendance have changed over time in league wise. In this study, we estimate the determinants of attendance patterns at Target Field, the home of Minnesota Twins, using regression shrinkage methods.

Duzan (2015) presented a review of ridge regression (RR) for solving the multicollinearity and introduction of its application on a predictive analysis. Marquardt and Snee (2012) showed that coefficients produced by the use of ridge regression in practice is relevant procedure for variable selection, especially when high collinearity exists between variables. Tibshirani (1996) presented lasso (Least Absolute Shrinkage Selector Operator) regression that minimizes the residual sum of squares due to the sum of the absolute value of the coefficients and proved through the simulation study that it can be applied to generalized regression models. Zou and Hastie (2003) proposed a new regression shrinkage method called, elastic net regression, and demonstrated its practicality in the analysis of data where predictors outnumber sample observations. The interpretation of these regularization methods will be discussed with details later in this paper.

Data Collection and Methodology

In this section, we define the variables used to estimate the regression model. Our dataset dates back from 2018 to 2010, which is the year of the new stadium, Target Field being opened. Most of our data were obtained from the following website: http://www.baseball-reference.com/leagues/twins, while some categorical variables were collected based on occurrence of the event. The regressor vectors x_i includes game-specific variables, opponent-specific variables, win-related variable, and event variables. Specifically,

x = *{Year, Month, Day, Night, GameNumber, Opponent,* OPP_{freq}*,* $OPP_{prvwpct}$*,* OPP_{salary}*,* OPP_{ws}*, Rank, GameBehind, OpeningDay, Holiday, Honeymoon, Statefair, Bobblehead}*

Game-specific variables are *Year(2010-2018), Month(Mar–Oct), Day(Monday-Sunday), Night(1=night game, 0=day game), and GameNumber(the sequence of 81 home games)*. Opponent-specific variables include *Opponent* that has 27 different levels, teams in this case. With respect to the opponent each game, the variable OPP_{freq} is the frequency level (High, Med, and Low) of the number of games that Twins played against over the observed period, the variable $OPP_{prvwpct}$ is the winning percentage of the opponent team in previous season, the variable OPP_{salary} is the reported total salary of the opponent team, and the variable OPP_{ws} is the number of World Series, championship game, appearances since the beginning of baseball league. The win-related variables are *Rank(division rank at the game day)* and *GameBehind(the number of winning games against the division leader)*. The event variables in *x* are *OpeningDay, Holiday, Bobblehead*, *Honeymoon,* and *Statefair*. The first three variables are the value of 1=yes and 0=no at each game when the game was opening day and holiday and when the game included a promotion of bobblehead giveaway, respectively. Given the findings in Rottenberg's (1956) attendance study, the variable *Honeymoon* received the dummy value of 1=yes and 0=no as the game took place within four years since opening of the new stadium, Target Field. The variable *Statefair* is indicated by the value of 1 and 0 when the Minnesota State Fairs occurred during the game day.

Key methodology that is used in this paper is solving multicollinearity problem of data through regularization. With the simple regression analysis, we use ordinal least square (OLS) to estimate $\hat{\beta}$ in the linear model by minimizing the mean square error (MSE) that is difference between *y* and $X\hat{\beta}$. To obtain $\hat{\beta}$ parameter estimators, we minimize the following function, which is Residual Sum of Square (RSS):

$$L_{OLS}(\hat{\beta}) = \sum_{i=1}^{n} (y_i - x_i'\hat{\beta})^2 = ||y - X\hat{\beta}||^2 = RSS$$

The OLS model can achieve unbiased estimators however results in high variance of model. Duzan (2015) explained that complex model with high variance can cause following problems: 1) multicollinearity may exist due to strong correlations between independent variables, and 2) the variance of the regression coefficient becomes large, and the prediction accuracy of the estimated regression equation becomes low. It becomes difficult to judge which of the many independent variables predicts the response the most. Schroeder (1990) found that the problem of multicollinearity appears most researches in social sciences for complex interrelationships between social determinants.

When this multicollinearity occurs, one of available solutions is regularization. The regularization method reduces variance through a trade-off losing unbiasedness of the model. According to the model complexity trade-off plot that Yu (2006) introduced in figure 1, the unbiased OLS model is somewhere in the right side of the graph having unbiasedness but high variance estimator with high complexity. The regularization method will move the complexity to the optimal point in the middle of the graph to reduce variance by costing unbiasedness. Through this regularization method, the fitted line becomes to be less complex and be regularized to explain data points more generally despite biasedness. Unlike subset selection method where prediction accuracy may decrease by fitting only selected variables and disregarding unselected ones, regularization methods can reduce variability and still hold predictability of model.

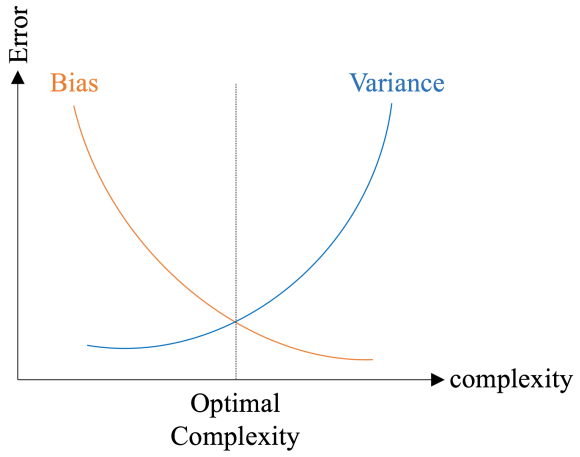


Figure 1. Yu's model complexity trade-off plot (2006)

In this paper, three regularization methods are used, including ridge, lasso, and elastic net regression. First, the ridge regression decreases model complexity by penalizing the predictors' coefficients that are far from zero to enforce them to be close to zero, while all variables remain in model. Like following equation, the ridge regression estimates $\hat{\beta}$ by minimizing the residual sum of squares with added penalty that is the size of coefficient estimates, which is called L2 penalty.

$$L_{ridge}(\hat{\beta}) = ||y - X\hat{\beta}||^2 + \lambda||\hat{\beta}||^2 = RSS + \lambda||\hat{\beta}||^2$$

The parameter λ is a tuning parameter that controls the strength of penalty. When λ equals to zero, the equation is basically OLS. As λ is larger, the penalized coefficients start to shrink toward zero with lower variance. (Ogutu, 2012)

Similarly, lasso regression also adds penalty to the non-zero coefficients but in different form, which is called L1 penalty, to the OLS equation. It penalizes sum of absolute values of coefficients.

$$L_{lasso}(\hat{\beta}) = ||y - X\hat{\beta}||^2 + \lambda||\hat{\beta}||_1 = RSS + \lambda||\hat{\beta}||_1$$

Unlike ridge regression, lasso regression enables to shrink some coefficients to completely zero as λ gets larger. Thus, Ogutu(2012) indicates that lasso regression can be robustly used where there are many predictors but only a few of them are significant to the response.

The elastic net, derived by Zou and Hastie (2003), combines both ridge and lasso constraints together. Notice α is another tuning parameter that makes it ridge when α is 0 and lasso when α is 1.

$$L_{elasticnet}(\hat{\beta}) = \frac{RSS}{2n} + \lambda(\frac{1-\alpha}{2}||\hat{\beta}||^2 + \alpha||\hat{\beta}||_1)$$

Therefore identifying the parameters α value between 0 and 1 with tuning the parameter λ value can optimize the elastic net model.

To choose the tuning parameter λ that produces the lowest variance in all three methods, repeated cross validation is used. In each repetition, folds are split in different way. The reason behind selecting the repeated cross validation approach over only one cross validation set is to lower the bias

since every fold of data is used to fit the models. By using repeated cross validation, our data is divided into train and validation set, and the three regularization methods are performed with the train set. Afterwards, model assessment metrics, RMSE and R^2, are computed to check performance of the three methods.

Analysis of Data and Results

Starting with the basics, we begins with testing the simple linear model of attendance:

$$\widehat{y_i} = \hat{\beta}_0 + x_1\hat{\beta}_1 + x_2\hat{\beta}_2 + \ldots + x_p\hat{\beta}_p \quad (1)$$

where y_i is the attendance per game *i* (=1,…,729); x_{ij} is p predictors = $(x_{11}, x_{22}, \ldots x_{ip})$; and β is a vector of regression coefficients $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_p)$. To achieve a parsimonious model, the linear model excludes *Opponent* (27 different levels) variable for testing.

Using 10-fold repeated cross validations, training data is broken into 10 parts and model is made from 9 parts, while 1 parts is used for the error estimation. And this is repeated 5 times with a different part used for error estimation, root mean square error (RMSE).

After modeling based on the training sets, the fitted linear model marks RMSE of 3531.633 with the coefficient of determination, R^2, of 0.7579. In other words, 75% of variability to the response is explained by this model. The linear model assumptions were also checked in Appendix 1.

The table in Appendix 2 shows the summary of the linear model with estimated coefficients each variable at 0.05 significance level. From this summary, we found that the variables *GameNumber, MonthMay, MonthJun, MonthJul, DaySaturday, OpeningDay, Holiday, Honeymoon, Bobblehead, OppfreqLow, OPPws*, and *GameBehind* have positive impact on the response. Attendance increased as season progressed especially on month of May, June, and July; more people visited Target Field on Saturdays, holidays, and Opening Days; more fans were motivated to visit ballpark when bobblehead promotion existed and when opponent was historically a winning franchise team or not playing in Minneapolis relatively often; and obviously attendance increased when the Twins were having a

winning record. On the other hand, attendance increased as year progressed since 2010; less fans visited the ballpark on Mondays, Tuesdays, and Wednesdays; and interestingly less people came to ballpark when the Minnesota State Fair was taking place.

Given the method used in Schroeder's study, we adopt variance inflation factors (VIF) as the diagnosing tool to assess the multicollinearity in this linear model (1). Fox (1984) explained, "The VIF indicates the degree to which the precision of the model is degraded by multicollinearity." Table 2 provides the values of VIF that are greater than 3 in our model. The result shows that the model has a problem of multicollinearity and is needed to reduce unnecessary variables. To solve overfitting of the model, we select regularization methods, including ridge regression, lasso regression, and elastic net regression.

Table 2. Variance Inflation Factors of independent variables in the model

Year	***GameNumber***	***MonthAug***	***MonthJul***	***MonthJun***	***MonthMay***
4.732617	34.52727	24.797333	15.562222	7.726418	3.459061
MonthOct	***MonthSep***	***Honeymoon1***	***Rank***	***GameBehind***	
4.207924	41.040993	4.157073	3.178986	4.763372	

First, ridge regression was used to try shrinking coefficients while all variables remain in the model. In R, we use the function `train()` to compute the ridge regression model ($\alpha = 0$) because this function is advantageous to allow tuning the parameter α for lasso ($\alpha = 1$) and elastic net ($0 < \alpha < 1$). The λ value is estimated by using 10-fold repeated cross validation approach that is a penalty to the coefficients. After running the model, we found the best value of λ is 500 as demonstrated in the Figure 2. Notice that the error increased as λ value increased after value of 500.

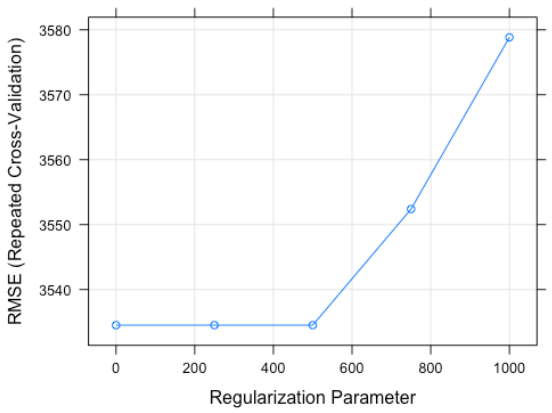


Figure 2. Trend of RMSE depending on λ value through ridge regression

The process of coefficients shrinkage through the ridge regression method is presented via Figure 3. Each line shows coefficients for one variable for different value of λ. Notice variables such as *GameNumber* has very large coefficients at the beginning. By penalizing the coefficients with λ value, the coefficients are shrunken and become toward zero as λ increases. This clearly shows that increasing λ value helps shrink coefficient values while all 27 independent variables are included in the model.

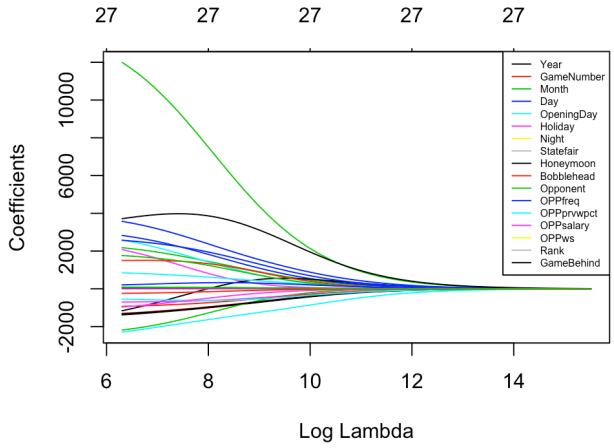


Figure 3. Coefficient shrinkage depending λ value through ridge regression

We also report the plot that explains how value of all coefficients grow as variability of the model to the response increases in Figure 4.

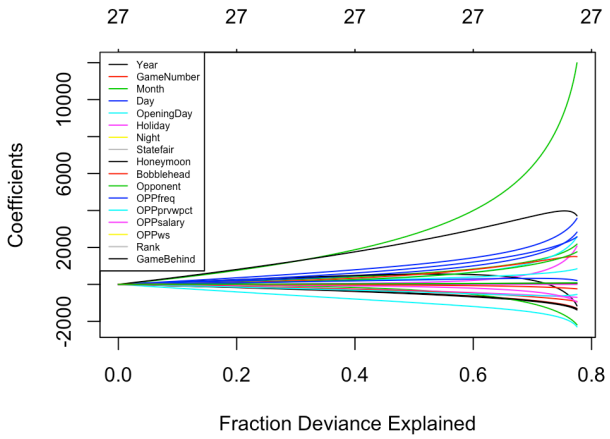


Figure 4. Trend of coefficients on variability of model through ridge regression

Noticeably, the coefficients of all variables gradually grow and explain about 70% of deviance with balance. By reaching at 75% of variability of model, the variable *Month*'s coefficient becomes to be highly inflated, that represents a problem of overfitting.

Similarly, we select the best cross validated lambda with parameter α set to 1 for lasso and parameter α to be mixed between 0 and 1 for elastic net, portrayed in Figure 5.

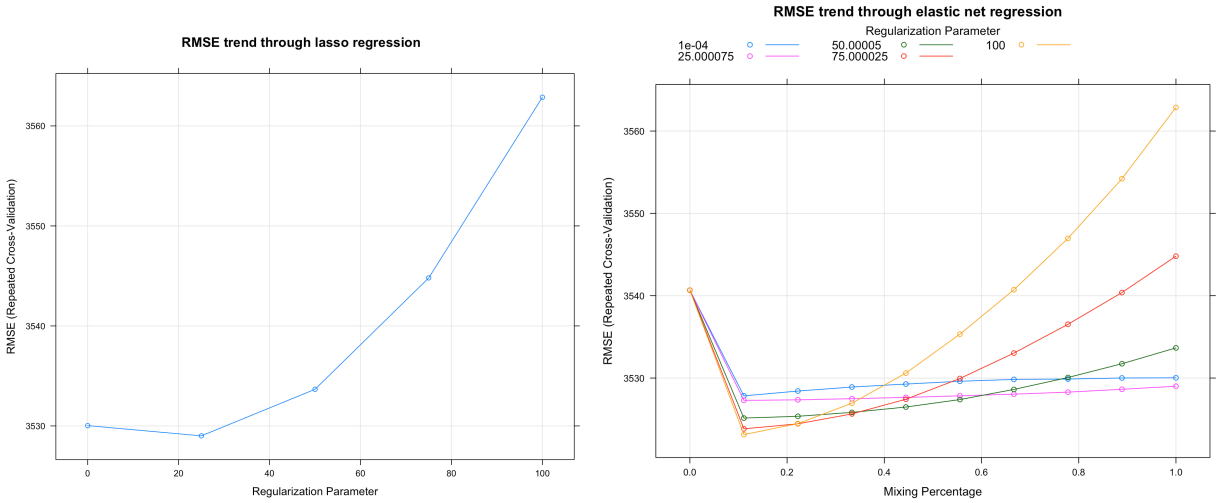


Figure 5. Trend of RMSE depending on λ value through lasso and elastic regression

Notice that the best value of lambda for lasso is 25 that minimizes the RMSE the most and the RMSE starts to increase upright as more L1 penalty is added. For elastic net, each line represents different value of lambda and is evaluated at each mixing percentage of α. It turns out that when lambda equals to 100 with the mixing percentage of ridge and lasso at 0.111 RMSE is minimized.

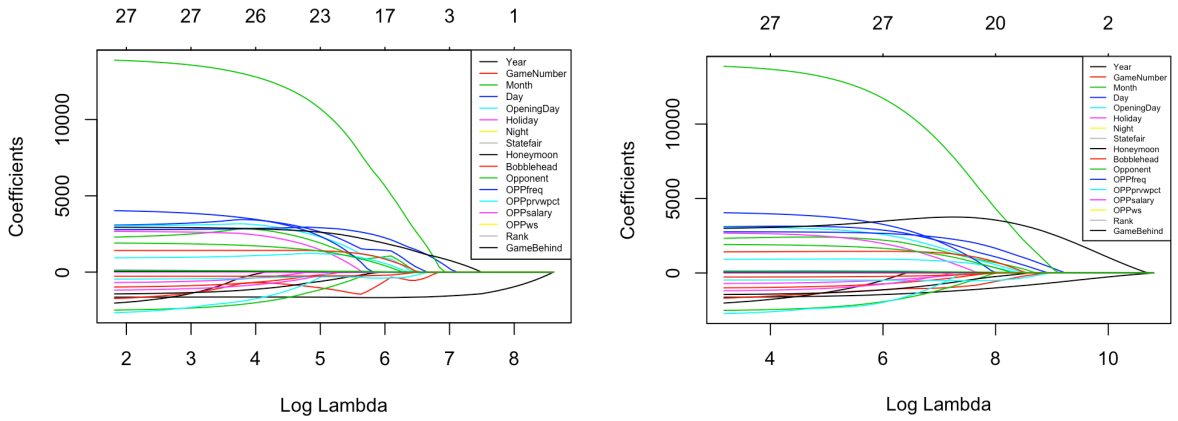


Figure 6. Coefficient shrinkage depending λ value through lasso and elastic net regression

The two plots in Figure 6 illustrate the process of coefficient shrinkage for lasso and elastic net method, respectively. The coefficients are shrunk towards zero as λ value increases for both methods. Especially, coefficient of the *Month* variable is reduced more rapidly than other variables for both methods. Besides performing coefficient shrinkage, the lasso and elastic net method also perform a feature selection as the number of variables that are explained at each different value of lambda decreases, unlike the ridge regression.

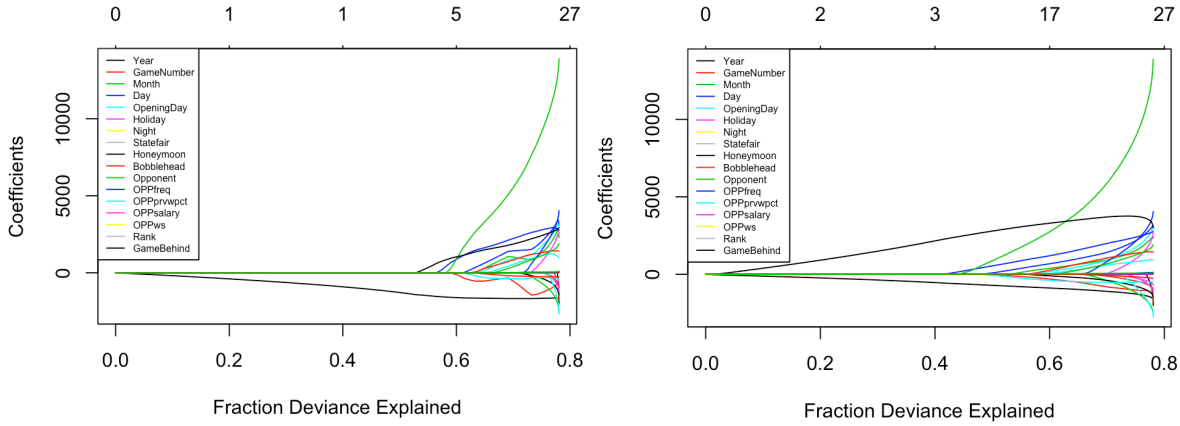


Figure 7. Trend of coefficients on variability of model through lasso and elastic net regression

The two plots that explain the variability of model depending on the trend of coefficients of selected variables through lasso and elastic net method are also presented in Figure 7. The left plot of lasso regression shows that 60% of variability is explained by only 5 variables. It is notable to observe that the coefficient of *Month* variable grows very aggressively while only 10% of variability is being

added in both methods. Obviously, we can see that other variables whose coefficients grow more gradually, such as *GameBehind* and *Year*, performed better than *Month* variable.

After using different shrinkage methods, all three models and the simple linear are compared in terms of RMSE and R^2.

Table 3. Comparison of linear and different regularization models

	Linear	Ridge	Lasso	Elastic Net
RMSE	3531.633	3534.511	3529.008	3517.88
R-Squared	0.7579738	0.758915	0.7578321	0.7603406

From the Table 3 above, we can clearly see that RMSEs are reduced by the models using lasso and elastic net compared to the linear model, and the variabilities of two models to our response, attendance are improved. The elastic net that combines the methods of ridge and lasso penalties is shown to be the best model. Noticeably, RMSE and R^2 are not improved through the ridge regression. This result explains characteristic of lasso and elastic net shrinkage methods that they work robustly where only a few variables dominate variability of the model with high coefficients while others hardly affect the response, which is exactly the case of our dataset.

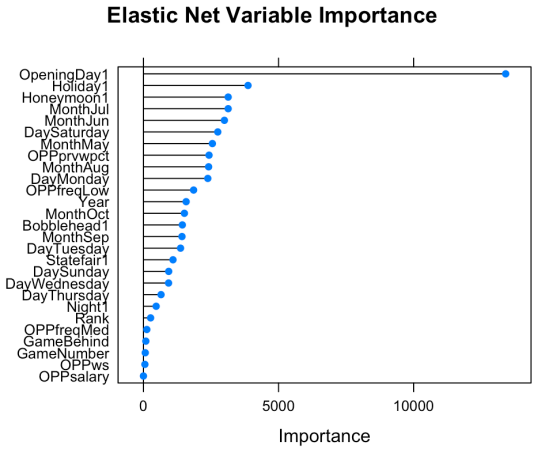


Figure 8. Variable importance through elastic net method

With choosing the best shrinkage method of elastic net regression, we explores to see which variables are important to explain the variability of the model. The Figure 8 illustrates importance of

each variable with the scale of different coefficients. Notice that *OpeningDay* is dominantly significant variable followed by the second running group of *Holiday*, *Honeymoon*, *Month* from *May* to *August*, and *DayMonday*. And coefficients of some variables, such as *GameNumer*, *GameBehind*, OPP_{ws}, OPP_{salary} value zero in this model. The variable importance plots using ridge and lasso regression are also reported in Appendix 3.

Discussion

The purpose of this project is to understand which variables contribute the most to the response of attendance in baseball, through the process of solving multicollinearity problem between multiple variables via regularization methods of ridge, lasso, and elastic net regression. From multiple predictors in our dataset, we found that multicollinearity occurs between high correlated variables by using Variance Inflation Factors. Instead of other methods like principle components analysis or subset selection, we adopted the regularization methods that shrink coefficients of variables to reduce variance of the model with a trade-off losing unbiasedness. After running the ridge, lasso, and elastic net methods, we found that lasso and elastic net shrinkage methods were useful to our dataset as some variables' coefficients remain affective while other unnecessary variables' coefficients were shrunken to zeros. By this methodology, we found that the RMSE was reduced and R^2 was improved, compared to the simple linear model.

Some limitations of this project are discussable. First, we have insufficient dataset. This project only analyzes 9 baseball seasons, given the fact that the new stadium was opened in 2010. As more seasons will be accumulated, analysis of determinants on predicating attendance can be greatly improved with more datasets. Second, some predictors that might have great influence on attendance are not studied in this project. As the literature review section pointed out, many researchers on this subject found that ticket price has highly affected the attendance in baseball. For this project, the ticket price was not able to be collected from any source. If the ticket price is not available, defining and

identifying metrics that can measure fans' ticket purchasing behaviors can greatly improve the analysis in future research.

Given the problem of decreasing attendance in baseball, MLB league office and each team are encouraged to investigate major and minor factors that affect attendance, as they have a great advantage of possessing many resources and information. Although predicting attendance can never be accurate, their efforts, just like what this project suggested, can allow them to identify key variables and utilize them when they plan optimal business strategies in future.

Works Cited

Ahn, S. C., & Lee, Y. H. (2014). Major League Baseball Attendance: Long-Term Analysis Using Factor Models. *Journal of Sports Economics*, 1-27.

Davis, M. C. (2008). The interaction between baseball attendance and winning percentage: A VAR analysis. *International Journal of Sport Finance, 3*, 58-73.

Duzan, H. (2015). Ridge Regression for Solving the Multicollinearity Problem . *Journal of Applied Science , 15*(3), 392-404.

Fox, J. (1984). Linear Statistical Models and Related Methods: With applications to social research.

Humphreys, B. R. (2002). Alternative measures of competitive balance in sports leagues. *Journal of Sports Economics, 3*, 133–148.

Lee, Y. H., & Fort, R. (2008). Attendance and the uncertainty of outcome hypothesis in Baseball. *Review of Industrial Organization, 33*, 281–295.

Marquardt, D. W., & Snee, R. D. (2012). Ridge Regression in Practice. *The American Statistician , 12*, 3-20.

Mills, B. M., & Fort, R. (2018). Team-Level Time Series Analysis in MLB, the NBA, and the NHL: Attendance and Outcome Uncertainty. *Journal of Sports Economics, 19*(7), 911-933.

Ogutu, J. (2012). Genomic selection using regularized linear regression models: ridge regression, lasso, elastic net and their extensions. *15th European workshop on QTL mapping and marker assisted selection (QTLMAS), 2*, 19-20.

Schroeder, M. A. (1990). Diagnosing and Dealing with Multicollinearity . *Western Journal of Nursing Research , 12*(2), 175-187.

Tayler, J. (2018). *Rob Manfred and MLB's Declining Attendance Issue.* NY: Sports Illustrated Magazine.

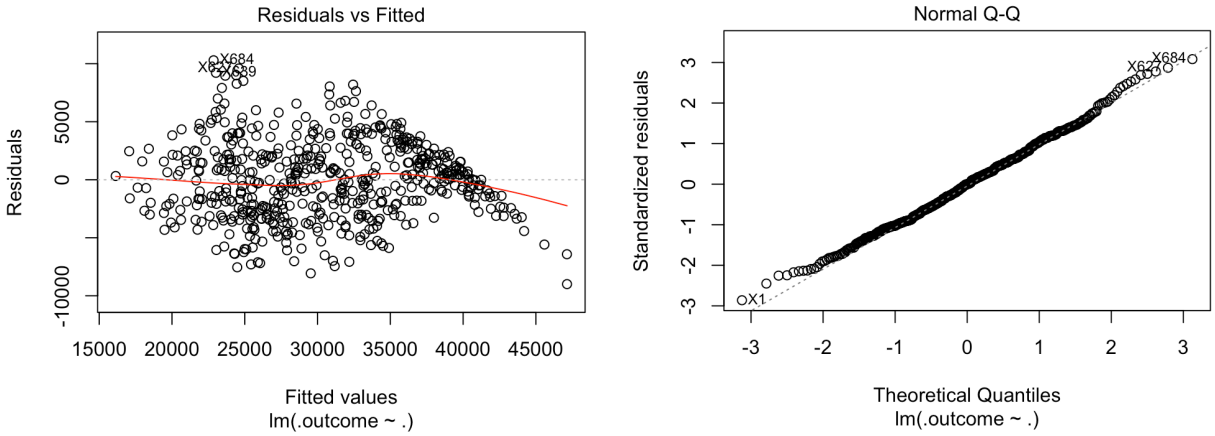
Tibshirani, R. (1996). Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society* , 267-288.

Yu, L. (2006). A Bias-Variance-Complexity Trade-Off Framework for Complex System Modeling . *International Conference on Computational Science and Its Applications, 3980*.

Zou, H., & Hastie, T. (2004). Regression Shrinkage and Selection via the Elastic Net, with Applications to Microarrays. *J. R. Statist. Soc. B, 67*, 301-320.

Appendix

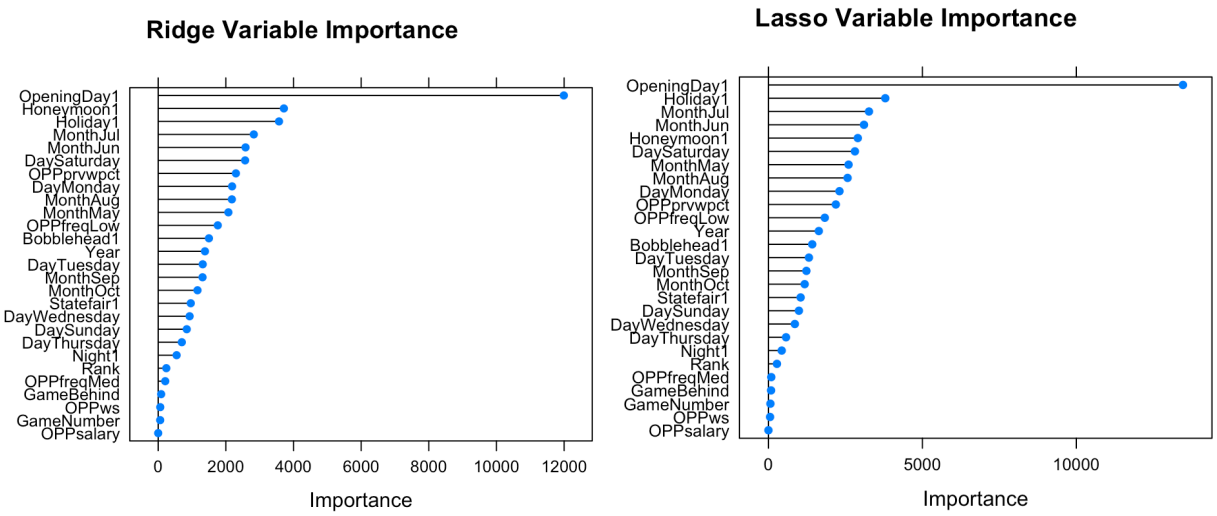
Appendix 1. Simple Linear Model Assumptions



Appendix 2. Testing statistical significance of independent variables in the linear model

	Estimate	Std.Error	t-value	p-value	Significance
(Intercept)	3.35E+06	2.51E+05	13.352	<2.0E-16	***
Year	-1.65E+03	1.24E+02	-13.264	<2.0E-16	***
GameNumber	8.27E+01	3.51E+01	2.356	0.018812	*
MonthAug	1.99E+03	1.91E+03	1.040	0.299025	
MonthJul	2.86E+03	1.49E+03	1.922	0.055105	.
MonthJun	2.90E+03	1.05E+03	2.776	0.005693	**
MonthMay	2.64E+03	7.06E+02	3.730	0.000212	***
MonthOct	-2.61E+03	2.94E+03	-0.888	0.375014	
MonthSep	-2.17E+03	2.35E+03	-0.923	0.356432	
DayMonday	-2.56E+03	6.25E+02	-4.087	5.05E-05	***
DaySaturday	2.77E+03	5.52E+02	5.012	7.33E-07	***
DaySunday	9.10E+02	6.66E+02	1.366	0.172418	
DayThursday	-7.28E+02	5.69E+02	-1.280	0.201115	
DayTuesday	-1.47E+03	5.28E+02	-2.787	0.005506	**
DayWednesday	-1.02E+03	5.79E+02	-1.765	0.078176	.
OpeningDay1	1.41E+04	1.38E+03	10.19	<2.0E-16	***
Holiday1	4.11E+03	1.36E+03	3.028	0.00258	**
Night1	-4.74E+02	4.36E+02	-1.087	0.277306	
Statefair1	-1.23E+03	6.36E+02	-1.941	0.052807	.
Honeymoon1	2.95E+03	6.05E+02	4.872	1.46E-06	***
Bobblehead1	1.42E+03	7.14E+02	1.983	0.047886	*
OPPfreqLow	1.93E+03	5.64E+02	3.426	0.000659	***
OPPfreqMed	1.26E+02	3.23E+02	0.391	0.695729	
OPPprvwpct	-2.81E+03	2.52E+03	-1.116	0.264723	
OPPsalary	1.66E-06	4.52E-06	0.367	0.713562	
OPPws	5.14E+01	2.12E+01	2.419	0.015888	*
Rank	-2.72E+02	1.77E+02	-1.539	0.124377	
GameBehind	9.26E+01	3.76E+01	2.464	0.014041	*

Appendix 3. Variable importance through ridge and lasso shrinkage method



R Code

```
Att_data= read.csv("/Users/Tae/Desktop/TwinsAtt.csv",head=TRUE)
Att_data_temp = Att_data[,-12]
Att_data_temp2 = Att_data[,-13]
attach(Att_data)
# install.packages("glmnet")
# install.packages("Matrix")
library(mlbench)
library(psych)
library(caret)
library(car)
library(glmnet)
library(dplyr)   # for data cleaning


# Explore the dataset
Att_data$OpeningDay = as.factor(Att_data$OpeningDay)
Att_data$Holiday = as.factor(Att_data$Holiday)
Att_data$Night = as.factor(Att_data$Night)
Att_data$Bobblehead = as.factor(Att_data$Bobblehead)
Att_data$Statefair = as.factor(Att_data$Statefair)
Att_data$Honeymoon = as.factor(Att_data$Honeymoon)
str(Att_data)


# Explore different linear model
Att_model_0 = lm(Attendance ~.,Att_data)
summary(Att_model_0)
anova(Att_model_0)
```

```
Att_model_noopp = lm(Attendance ~.,Att_data_temp)
summary(Att_model_noopp)



# Check Multicollinearity
vif(Att_model_noopp)

# collinearlity check
pairs.panels(Att_data_temp[c(-1)])

# data partition
ind = sample(2, nrow(Att_data_temp), replace = T, prob = c(7/9,2/9))
train_data = Att_data_temp[ind==1,]
test_data = Att_data_temp[ind==2,]

# Repeated Cross Validation
custom = trainControl(method="repeatedcv",
                      number = 10,
                      repeats = 5,
                      verboseIter = T)
#verboseIter = true allows to look at the progress as the cross validation
performs.

# Linear Model
set.seed(4893)
Att_Linear_model = train(Attendance~.,
                         train_data,
                         method = 'lm',
                         trControl = custom)
Att_Linear_model$results
Att_Linear_model
summary(Att_Linear_model)
plot(Att_Linear_model$finalModel)

# Ridge Regression
set.seed(4893)
Att_Ridge = train(Attendance~.,
                  train_data,
                  method = 'glmnet',
                  tuneGrid = expand.grid(alpha=0,
                                         lambda= seq(0.0001,1000,length=5)),
                  trControl=custom)

plot(Att_Ridge)
Att_Ridge$results
plot(Att_Ridge$finalModel, xvar="lambda")
legend("topright", lwd = 1, col = 1:30, legend = colnames(Att_data[,-1]), cex
= .5)
plot(Att_Ridge$finalModel, xvar="dev")
legend("topleft", lwd = 1, col = 1:30, legend = colnames(Att_data[,-1]), cex = .5)
```

```
plot(varImp(Att_Ridge, scale=F),main="Ridge Variable Importance")

# Lasso Regression
set.seed(4893)
Att_Lasso = train(Attendance~.,
                  train_data,
                  method = 'glmnet',
                  tuneGrid = expand.grid(alpha=1,
                                         lambda= seq(0.0001,100,length=5)),
                  trControl=custom)
Att_Lasso
plot(Att_Lasso,main="RMSE trend through lasso regression")
plot(Att_Lasso$finalModel, xvar="lambda")
legend("topright", lwd = 1, col = 1:30, legend = colnames(Att_data[,-1]), cex
= .5)
plot(Att_Lasso$finalModel, xvar="dev")
legend("topleft", lwd = 1, col = 1:60, legend = colnames(Att_data[,-1]), cex = .5)
plot(varImp(Att_Lasso, scale=F),main="Lasso Variable Importance")

# Elastic Net Regression
set.seed(4893)
Att_EN = train(Attendance~.,
                  train_data,
                  method = 'glmnet',
                  tuneGrid = expand.grid(alpha=seq(0,1, length=10),
                                         lambda= seq(0.0001,100,length=5)),
                  trControl=custom)
Att_EN$results
plot(Att_EN,main="RMSE trend through elastic net regression")
plot(Att_EN$finalModel, xvar="lambda")
legend("topright", lwd = 1, col = 1:30, legend = colnames(Att_data[,-1]), cex
= .45)
plot(Att_EN$finalModel, xvar="dev")
legend("topleft", lwd = 1, col = 1:30, legend = colnames(Att_data[,-1]), cex = .5)
plot(varImp(Att_EN, scale=F),main="Elastic Net Variable Importance")
```