Moneyball Data Assignment

Cesar Espitia

CUNY SPS Data 621

Table of Contents

Abstract

This assignment focused on the most American sport – baseball. The contains approximately 2200 records spanning nearly 150 years with each record indicative of a baseball team in a particular year. The variables for the data are performance metrics for each team from both the team at bat or the team on the field. The purpose is to analyze the data, perform any data manipulation / clean-up and build three (3) linear regression models using only the data (or derivatives) to predict if the team won. The chosen model provided a $R^2 = 0.3136$.

Keywords: moneyball, data621

Moneyball Data Assignment

The following is the analysis and write-up based upon my interpretation of the data and final model used to predict if a baseball team won based upon the present variables.

[The body of your paper uses a half-inch first line indent and is double-spaced. APA style provides for up to five heading levels, shown in the paragraphs that follow. Note that the word *Introduction* should not be used as an initial heading, as it's assumed that your paper begins with an introduction.]

Data Exploration

The purpose of this step is to get a sense of the data or 'feel.' The following information describes the data from different angles including completeness, statistical summaries, visuals to determine the shape and effect of each variable and other items deemed pertinent.

Summary Statistics

The first step is to look at the data to determine some items including completeness and the shape of each variable.   The following are the results of summarizing the data in a table and the visualization of each variables density function (PDF).

Table 1

Summary Statistics for Moneyball Training Data

Variable	Min	1Q	Med	Mean	3Q	Max	NA
TARGET_WINS	12.00	71.00	82.00	80.83	92.00	146.00	0
TEAM_BATTING_H	992	1383	1454	1470	1538	2554	0
TEAM_BATTING_2B	69.0	208.0	238.0	241.3	273.0	458.0	0
TEAM_BATTING_3B	0.00	34.00	47.00	55.27	72.00	223.00	0
TEAM_BATTING_HR	0.00	42.00	102.00	99.66	147.00	264.00	0
TEAM_BATTING_BB	12.0	451.0	512.0	501.8	580.0	878.0	0
TEAM_BATTING_SO	0.0	557.5	735.6	735.9	925.0	1399.0	102
TEAM_BASERUN_SB	0.0	67.0	106.0	124.8	151.0	697.0	131
TEAM_BASERUN_CS	7.00	44.00	52.80	52.83	54.50	201.00	772
TEAM_BATTING_HBP	29.00	59.36	59.36	59.36	59.36	95.00	2085
TEAM_PITCHING_H	1137	1419	1518	1769	1682	30132	0

TEAM_PITCHING_HR	0.0	50.0	107.0	105.7	150.0	343.0	0
TEAM_PITCHING_BB	119.0	476.0	537.0	553.3	611.0	3645.0	0
TEAM_PITCHING_SO	0.0	626.0	817.7	818.1	957.0	19278.0	102
TEAM_FIELDING_E	65.0	127.0	159.0	245.8	249.0	1898.0	0
TEAM_FIELDING_DP	52.0	134.0	146.4	146.4	161.5	228.0	286

Note: Source: moneyball-training-data.csv

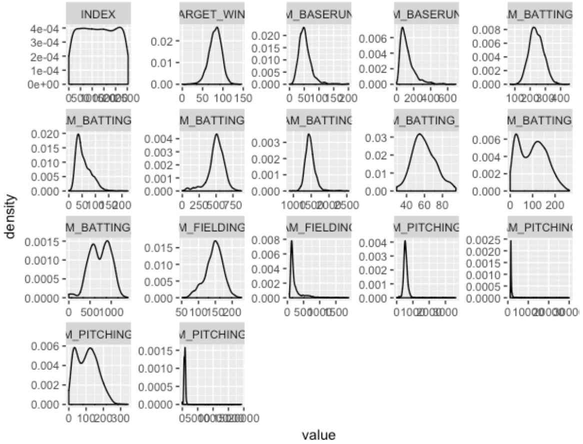

Figure 1. PDF for Each Dataframe Variable.

In looking at both, Table 1, Figure 1 and Appendix B (correlation matrix) together, we can note specific items that may skew our model building results.

NA: These incomplete cases will cause any correlation exercise to be incorrect or not possible. There are a few ways to deal with NAs including imputing the missing data or ignoring the variable altogether. For the purposes of this analysis, the variable TEAM_BATTING_HBP will be ignored as 94% of the data is missing, for the other 5 variables, the mean will be used in imputing the missing data.

PDF: Figure 1 shows the PDF of each variable, this allows us to see if the data is normal or night. 6 variables show typical normal density functions but others like TEAM_PICTHING_BB show left skewness and other show bimodality. For the purposes of this analysis, the variables will be normalized around each variable's mean to remove the effects of order of magnitude and skewness.

Outliers: These data points can cause unintended consequences when running our models as they can skew the coefficients and produce significance in the variable when it may actually not be. In this data, the variable TEAM_PITCHING_ SO has a max value of 19,278 which is one or two orders of magnitude bigger than the other variables. There are 2 teams that have a value over 10k and 4 teams that are over 3,645 which is the max of variable TEAM_PITCHING_ BB. For the purposes of this analysis, these four cases will be ignored.

Correlation: Lastly, we look for correlated variables that we can make decisisons on to determine which one is more important than the other. Correlated variables bloat the model and don't produce any more insight than ignoring one of the two. In our data, TEAM_BATTING_SO is correlated to TEAM_BATTING_3B and TEAM_PITCHING_HR is correlated to TEAM_BATTING_HR. In both cases, a Pearson correlation test was done to ensure that the correlation is significant, in both cases they were. For the purposes of this analysis, TEAM_BATTING_SO will be kept as the number of negative impact variables is small and TEAM_BATTING_HR as a toss-up, in this set, they both have equal validity.

Data Preparation

The purpose of this step is to take the findings from the exploration and transform the data as needed. The following information describes the transformations done in order to prepare the data for model building and model selection.

NA: All missing values were imputed using the mean within each column even though it is not the most adequate for this data. The nearest neighbor method would have been more valid if the variable year were present as this variable would help sort the data ascending and take into account seasonality of the data. The seasonality in this case would refer to how the game has evolved from the late 1800’s to this century. As this information is not known using the mean is the most efficient and effective way for this data.

Variable Creation: For this dataset, three variables were created that were deemed unique to the data. TEAM_BP_1B was created as single base hits are not called out explicitly in the data but may be important in the model to predict wins. This was taking by using the following equation:

$$TEAM\_BP\_1B = TEAM\_BATTING\_H - \Sigma(2B, 3B, HR)_{BATTING}$$

Which takes all base hits including homeruns and removes each singled out base hit column. The two remaining variables are ratios of existing data.

TEAM_BP_SO and TEAM_BP_BB each take the batting data for SO and BB and divide it by the pitching data for SO and BB. The thought behind these two variables is that the numerical values show the magnitude of each side, but a ratio shows if the team does better at either being struck out or causing the other team to strike out, or if they achieve walks versus giving them out. If the ratio SO < 1 and ratio BB > 1, the expectation is that these two combine are a positive contribution to the number of wins, the reverse would indicate that the team should have a lower value of wins.

Variable Deletion / Data Deletion: For this dataset, two variables were removed that were deemed unnecessary for model building. As noted in the exploration step on Page 6, TEAM_BATTING_3B and TEAM_BATTING_HR were removed from the data as they are the most highly correlated variables. The column INDEX is also removed, as it is a key variable for the data and does not add any value. TEAM_PITCHING_ SO will have 4 outlier rows of data removed (as noted in Page 6) which reduces the overall dimension of the dataframe. Lastly, there is one team that didn't win any games in the dataset, this outlier will be removed from TARGET_WINS.

Correlation Check: Once these manipulations are done, a side-by-side comparison of the correlations matrix is done to ensure no inadvertent effects to the data.

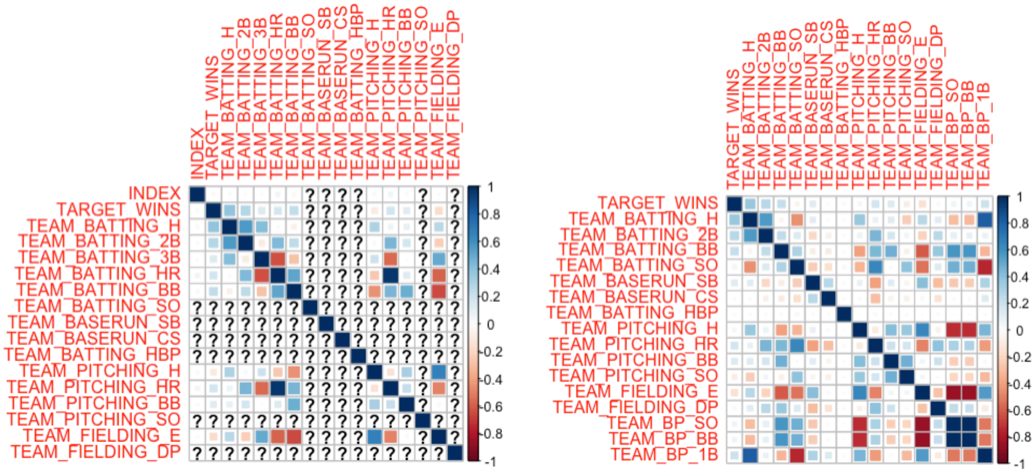

Figure 2. Correlation Comparison Before and After.

As is noted, before due to the amount of NA data in the original dataset, not all correlations were visible, however, those that were, can be noticed to have lesser correlations now that the information was manipulated and updated. There seems to be some correlation in the data from the variables created, but it will be ignored for now.

Normalization: Each variable was then normalized using the mean. The built-in function scale was used, in reviewing the summary statistics of this new data frame the Mean is 0 for each variable with SD 1.

Model Building

The purpose of this step is to take the modified dataset and beging exploring potential models that will be used on the final dataset provided. The following information describes the 3 models built for this step and the relevant analysis to provide reasons for model selection in the next step.

MODEL 1

The first model takes in the data as manipulated in step two (with variables imputed and removed). In this first model, we have an $R^2 = 0.3141$ and p-value < 0.05. The data in Figure 3, shows that there is not heteroscedastic and has a positive trend on the predicted vs fitted values.

```
Residuals:
    Min      1Q  Median      3Q     Max 
-3.0246 -0.5325  0.0077  0.5362  4.2954 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)      -4.514e-16  1.744e-02   0.000 1.000000    
TEAM_BATTING_H    1.485e+00  1.438e-01  10.326  < 2e-16 *** (EXPECTED POS)
TEAM_BATTING_2B  -4.170e-01  5.600e-02  -7.447 1.36e-13 *** (EXPECTED POS)*
TEAM_BATTING_BB   2.386e-01  4.623e-02   5.161 2.67e-07 *** (EXPECTED POS)
TEAM_BATTING_SO  -7.959e-02  3.922e-02  -2.029 0.042573 *   (EXPECTED NEG)
TEAM_BASERUN_SB   1.867e-01  2.457e-02   7.600 4.34e-14 *** (EXPECTED POS)
TEAM_BASERUN_CS  -1.037e-02  1.930e-02  -0.537 0.591081     (EXPECTED NEG)
TEAM_BATTING_HBP  1.725e-02  1.746e-02   0.988 0.323400     (EXPECTED POS)
TEAM_PITCHING_H  -1.530e-02  3.628e-02  -0.422 0.673320     (EXPECTED NEG)
TEAM_PITCHING_HR -2.192e-01  6.144e-02  -3.567 0.000368 *** (EXPECTED NEG)
TEAM_PITCHING_BB -1.297e-01  4.642e-02  -2.794 0.005247 **  (EXPECTED NEG)
TEAM_PITCHING_SO  8.645e-02  3.145e-02   2.748 0.006038 **  (EXPECTED POS)
TEAM_FIELDING_E  -5.843e-01  4.738e-02 -12.333  < 2e-16 *** (EXPECTED NEG)
TEAM_FIELDING_DP -1.600e-01  2.120e-02  -7.551 6.27e-14 *** (EXPECTED POS)*
TEAM_BP_SO       -1.025e+00  3.072e-01  -3.337 0.000859 *** (EXPECTED POS)
TEAM_BP_BB        5.999e-01  3.065e-01   1.957 0.050447 .  
TEAM_BP_1B       -9.889e-01  1.345e-01  -7.352 2.73e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8282 on 2239 degrees of freedom
Multiple R-squared:  0.3189,	Adjusted R-squared:  0.3141 
F-statistic: 65.53 on 16 and 2239 DF,  p-value: < 2.2e-16
```

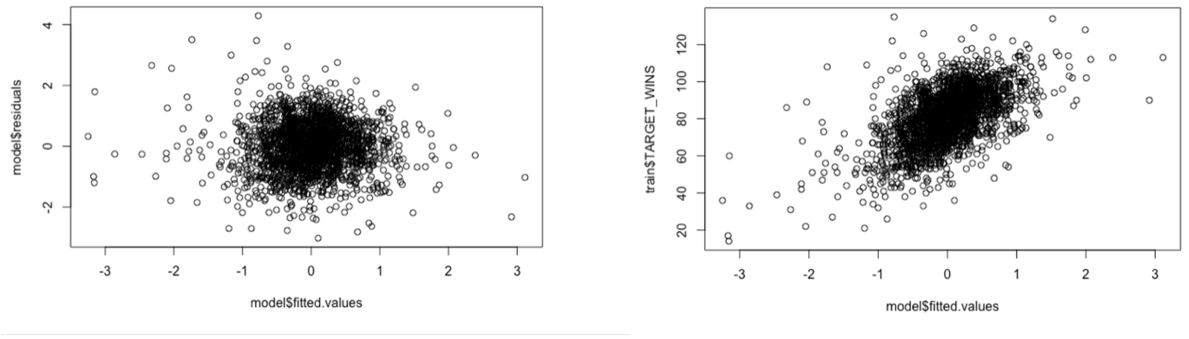

Figure 3. Model Check for Residual Shape and Model vs. Actuals

What is peculiar in the results however, are that some variables have factor that is counterintuitive to the expected impact on TARGET_WINS. TEAM_BATTING_2B and TEAM_FIELDING_DP are both negative but should be positive contributors. For now, they will be left in.

MODEL 2

The second model only takes into account the variables noted of significance from Model 1 (p-value < 0.05). This means that TEAM_BASERUN_CS, TEAM_BATTING_HBP, TEAM_BP_BB and TEAM_PITCHING_H will be removed from the dataset for Model 2. In this second model, we have an $R^2 = 0.3136$ and p-value < 0.05 which is only a marginal improvement in the model capability.

```
Residuals:
    Min      1Q  Median      3Q     Max 
-3.0245 -0.5350  0.0061  0.5278  4.3289 

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)    
(Intercept)      -6.386e-01  4.830e-01  -1.322 0.186243    
TEAM_BATTING_H    1.077e-02  1.019e-03  10.568  < 2e-16 ***
TEAM_BATTING_2B  -9.259e-03  1.188e-03  -7.792 9.99e-15 ***
TEAM_BATTING_BB   2.152e-03  3.631e-04   5.926 3.58e-09 ***
TEAM_BATTING_SO  -3.058e-04  1.671e-04  -1.831 0.067285 .  
TEAM_BASERUN_SB   2.189e-03  2.681e-04   8.164 5.35e-16 ***
```

```
TEAM_PITCHING_HR -3.835e-03  9.827e-04  -3.903 9.79e-05 ***
TEAM_PITCHING_BB -8.683e-04  2.490e-04  -3.487 0.000498 ***
TEAM_PITCHING_SO  1.588e-04  5.836e-05   2.721 0.006566 **
TEAM_FIELDING_E  -2.887e-03  2.122e-04 -13.605  < 2e-16 ***
TEAM_FIELDING_DP -6.884e-03  8.385e-04  -8.210 3.68e-16 ***
TEAM_BP_SO       -3.128e+00  3.637e-01  -8.601  < 2e-16 ***
TEAM_BP_1B       -8.166e-03  1.065e-03  -7.669 2.57e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8285 on 2243 degrees of freedom
Multiple R-squared:  0.3173,    Adjusted R-squared:  0.3136
F-statistic: 86.85 on 12 and 2243 DF,  p-value: < 2.2e-16
```

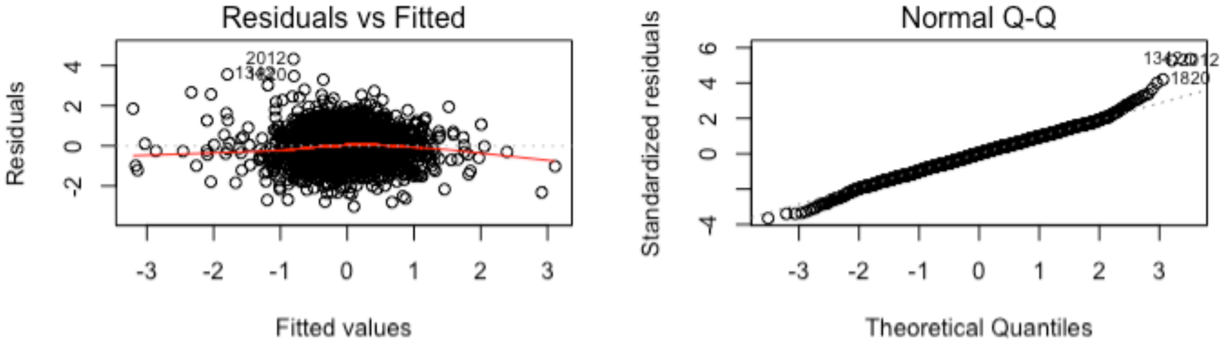

Figure 4. Model 2 Plots (Residuals vs Fitted and QQ)

MODEL 3

The third model takes the second model and removes the variables noted in Model 1 as contrary to their expected impact (p-value < 0.05). In this third model, we have an $R^2 = 0.2721$ and p-value < 0.05 which is not an improvement in the model capability.

```
Residuals:
    Min      1Q  Median      3Q     Max 
-3.0276 -0.5828  0.0284  0.5658  3.8003 

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)    
(Intercept)      -1.444e+00  4.919e-01  -2.935  0.00337 ** 
TEAM_BATTING_H    3.399e-03  4.051e-04   8.390  < 2e-16 ***
TEAM_BATTING_BB   2.139e-03  3.725e-04   5.743 1.06e-08 ***
TEAM_BATTING_SO  -4.634e-04  1.657e-04  -2.796  0.00521 ** 
TEAM_BASERUN_SB   3.153e-03  2.626e-04  12.005  < 2e-16 ***
TEAM_PITCHING_HR  6.395e-04  6.843e-04   0.935  0.35011    
```

```
TEAM_PITCHING_BB -1.194e-03  2.547e-04  -4.688 2.92e-06 ***
TEAM_PITCHING_SO  1.715e-04  5.996e-05   2.860  0.00427 **
TEAM_FIELDING_E  -2.574e-03  2.065e-04 -12.469  < 2e-16 ***
TEAM_BP_SO       -2.737e+00  3.523e-01  -7.770 1.19e-14 ***
TEAM_BP_1B       -1.025e-03  5.446e-04  -1.882  0.06001 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8532 on 2245 degrees of freedom
Multiple R-squared:  0.2753,   Adjusted R-squared:  0.2721
F-statistic: 85.29 on 10 and 2245 DF,  p-value: < 2.2e-16
```

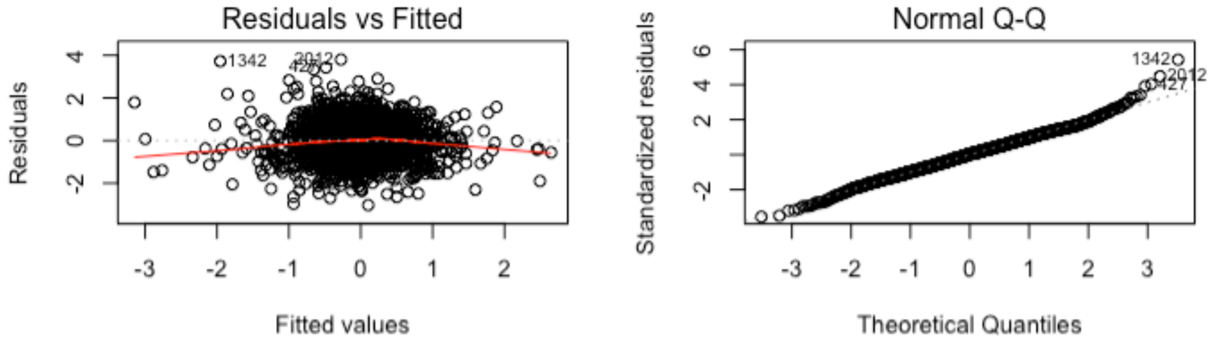

Figure 5. Model 3 Plots (Residuals vs Fitted and QQ)

The reduction in R-squared was not expected but makes sense. This means that TEAM_BATTING_2B and TEAM_FIELDING_DP were never meant to be a positive impact to TARGET_WINS but instead was meant to be negative impacts and indicates that the variables are important and that the original assumption of their impact was incorrect and needs to be updated. With this in mind, Model 2 so far seems the most appropriate so far.

Model Selection

The purpose of this step is to take the models built and build a ranking criteria to select the final model for use with the test data set. The following information describes the methodology and the results on the test data.

METHODOLOGY

Familiarity with the subject is low and therefore the methodology will be more closely related to the statistical information presented. In this case, Adjusted R^2 will be the only criteria to select the model. The reason for this is that the significance of each variable is high in models 1-3 as the adjustments for missing NA's, correlation and normalization were already taken care of in Step 2 of the process. If step 2 were not done, then it would have been hidden in the model building and taken care of between Model 1 and Model 2.

With this in mind, Model 2 is best model with an $R^2 = 0.3136$.

TEST DATA

The dataset had 2256 entries and 17 columns and was modified to fit the final variables and scaling used in Model 2 from above. This means that the same process of adjustments, imputing and scaling was done in order to be able to use the model correctly. The final predicted values are based upon a normalized value from the test data. The data is shown as follows with the corresponding upper and lower limits.

Table 2

Predicted Statstics vs Summary of TARGET_WINS in Training Data

	Predicted (Test)			Train
	fit	lwr	upr	Actual
Min.	-3.756	-4.447	-3.0667	-4.35663
1st	-3.245	-3.965	-2.5245	-0.64366
Median	-1.904	-2.791	-1.018	0.0788
Mean	-1.444	-2.48	-0.4079	0
3rd	-1.423	-2.392	-0.4531	0.72428
Max.	13.462	8962	17.9618	3.52528

This table is only meant as a comparison but it does highlight that the training data doesn’t fall in the anywhere in the upper / lower limits except in the minimum values. The spread of the data for training is also a lot tighter than the predicted values which an issue in the method of normalizing the test data. This might indicate why the training and predicted values aren’t more closely aligned.

Conclusion

Three models were presented after exploring and manipulating the data as necesarry. With only using the Adjusted R value as the criteria for selection model 2 was selected and provided an $R^2 = 0.3136$ which was adequate for the data but doesn’t necessarily indicate the best model . If more time were available, the manipulation of the predicted data would have used the same scaling values from the original training data. In rescaling the data from the test data, there is no guarantee that it will have the same means even though it might come from the same population. This might affect the predicted values of the data.

Appendix A: R Code

~~~
## Data Exploration


```{r dataexploration, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
library(e1071)
library(dplyr)
library(purrr)
library(tidyr)
library(ggplot2)
library(corrplot)
library(FactoMineR)
library(VIF)
library(knitr)
library(kableExtra)
library(Hmisc)

# read data
train = read.csv(file="data/moneyball-training-data.csv")
train2<-train

#check data
summary(train) %>% kable() %>% kable_styling()
sapply(train, function(x) sum(is.na(x)))

ntrain<-select_if(train, is.numeric)
ntrain %>%
  keep(is.numeric) %>%                     # Keep only numeric columns
  gather() %>%                             # Convert to key-value pairs
  ggplot(aes(value)) +                     # Plot the values
    facet_wrap(~ key, scales = "free") +   # In separate panels
    geom_density()

rcorr(as.matrix(train))
corrplot(cor(train), method="square")

# correlation test 1
cor.test(train$TEAM_BATTING_3B,train$TEAM_BATTING_SO,method="pearson")
#significant ignore

# correlation test 2
cor.test(train$TEAM_BATTING_HR,train$TEAM_PITCHING_HR,method="pearson")
#significant ignore


```

## Data Preparation

```{r datapreparation, include=FALSE}
# impute data for missing values
# use column mean for calculation

for(i in 1:ncol(train)){
  train[is.na(train[,i]), i] <- mean(train[,i], na.rm = TRUE)
}

#create THREE new variables Strikeout, Walks,
train$TEAM_BP_SO <- train$TEAM_BATTING_SO / train$TEAM_PITCHING_SO
train$TEAM_BP_BB <- train$TEAM_BATTING_BB / train$TEAM_PITCHING_BB
train$TEAM_BP_1B <- train$TEAM_BATTING_H-train$TEAM_BATTING_2B-train$TEAM_BATTING_3B-train$TEAM_BATTING_HR

#remove index, team_batting_3b and team

train <- train[, !(colnames(train) %in% c("INDEX","TEAM_BATTING_3B","TEAM_BATTING_HR"))]


train[train$TARGET_WINS==0,] <- NA
train <- train[complete.cases(train),]

# correlation
corrplot(cor(train), method="square")

trainNorm <-as.data.frame(scale(train))
summary(trainNorm)

```




## Build Models
```{r buildmodels, include=FALSE}

#MODEL 1
model <- lm(TARGET_WINS ~ ., data=trainNorm)
summary(model)
~~~

```
plot(model$residuals ~ model$fitted.values)
plot(model$fitted.values,trainNorm$TARGET_WINS)

#extract variables that are significant and rerun model
sigvars <- data.frame(summary(model)$coef[summary(model)$coef[,4] <= .05, 4])
sigvars <- add_rownames(sigvars, "vars")
colist<-dplyr::pull(sigvars, vars)

idx <- match(colist, names(train))
trainmod2 <- cbind(train[,idx], trainNorm['TARGET_WINS'])

#MODEL 2
model2<-lm(TARGET_WINS ~ ., data=trainmod2)

summary(model2)
plot(model2$residuals ~ model2$fitted.values)
plot(model2$fitted.values,trainNorm$TARGET_WINS)


par(mfrow=c(2,2))
plot(model2)

par(mfrow=c(1,2))
plot(model2$residuals ~ model2$fitted.values, main="New Reduced Var Model")
abline(h = 0)
plot(model$residuals ~ model$fitted.values, main="Orignal Model All Vars")
abline(h = 0)

#MODEL 3
#remove variables with opposite coefficients
dim(trainmod2)
trainNorm2 <- trainmod2[, !(colnames(trainmod2) %in% c("TEAM_BATTING_2B","TEAM_FIELDING_DP"))]
dim(trainNorm2)


model3<-lm(TARGET_WINS ~ ., data=trainNorm2)
summary(model3)

plot(model2$residuals ~ model3$fitted.values)
plot(model3$fitted.values,trainNorm2$TARGET_WINS)

par(mfrow=c(2,2))
plot(model3)

```

## Select Models
```{r selectmodels, include=FALSE}

cor(trainmod2)
cor(trainNorm2)
dim(trainmod2)
dim(test)

test = read.csv(file="data/moneyball-training-data.csv")
#create THREE new variables Strikeout, Walks,
test <- test[(test$TEAM_PITCHING_H <= 3456),]


for(i in 1:ncol(test)){
  test[is.na(test[,i]), i] <- mean(test[,i], na.rm = TRUE)
}
test$TEAM_BP_SO <- test$TEAM_BATTING_SO / test$TEAM_PITCHING_SO
test$TEAM_BP_BB <- test$TEAM_BATTING_BB / test$TEAM_PITCHING_BB
test$TEAM_BP_1B <- test$TEAM_BATTING_H-test$TEAM_BATTING_2B-test$TEAM_BATTING_3B-test$TEAM_BATTING_HR
test <- test[complete.cases(test),]

testNorm <-as.data.frame(scale(test))

idx <- match(colist, names(testNorm))
testNorm2 <- cbind(testNorm[,idx], testNorm['TARGET_WINS'])

summary(testNorm2)

modelfinal<- predict(model2, newdata = testNorm2, interval='confidence') #data from scaling originally to get to actual wins
summary(modelfinal)

summary(trainmod2)
```

Appendix B: CORRELATION MATRIX

	TARGET_WINS	BATTING_H	BATTING_2B	BATTING_3B	BATTING_HR	BATTING_BB	BATTING_SO	BASERUN_SB	BASERUN_CS	BATTING_HBP	PITCHING_H	PITCHING_HR	PITCHING_BB	PITCHING_SO	FIELDING_E	FIELDING_DP
TARGET_WINS	1	0.39	0.29	0.14	0.18	0.23	-0.03	0.14	0.02	0.07	-0.11	0.19	0.12	-0.08	-0.18	-0.03
BATTING_H	0.39	1	0.56	0.43	-0.01	-0.07	-0.46	0.12	0.02	-0.03	0.3	0.07	0.09	-0.25	0.26	0.16
BATTING_2B	0.29	0.56	1	-0.11	0.44	0.26	0.16	-0.2	-0.1	0.05	0.02	0.45	0.18	0.06	-0.24	0.29
BATTING_3B	0.14	0.43	-0.11	1	-0.64	-0.29	-0.67	0.53	0.35	-0.17	0.19	-0.57	0	-0.26	0.51	-0.32
BATTING_HR	0.18	-0.01	0.44	-0.64	1	0.51	0.73	-0.45	-0.43	0.11	-0.25	**0.97**	0.14	0.18	-0.59	0.45
BATTING_BB	0.23	-0.07	0.26	-0.29	0.51	1	0.38	-0.11	-0.14	0.05	-0.45	0.46	0.49	-0.02	-0.66	0.43
BATTING_SO	-0.03	-0.46	0.16	**-0.67**	0.73	0.38	1	-0.25	-0.22	0.22	-0.38	0.67	0.04	0.42	-0.58	0.15
BASERUN_SB	0.14	0.12	-0.2	0.53	-0.45	-0.11	-0.25	1	0.66	-0.06	0.07	-0.42	0.15	-0.14	0.51	-0.5
BASERUN_CS	0.02	0.02	-0.1	0.35	-0.43	-0.14	-0.22	0.66	1	-0.07	-0.05	-0.42	-0.11	-0.21	0.05	-0.21
BATTING_HBP	0.07	-0.03	0.05	-0.17	0.11	0.05	0.22	-0.06	-0.07	1	-0.03	0.11	0.05	0.22	0.04	-0.07
PITCHING_H	-0.11	0.3	0.02	0.19	-0.25	-0.45	-0.38	0.07	-0.05	-0.03	1	-0.14	0.32	0.27	0.67	-0.23
PITCHING_HR	0.19	0.07	0.45	-0.57	0.97	0.46	0.67	-0.42	-0.42	0.11	-0.14	1	0.22	0.21	-0.49	0.44
PITCHING_BB	0.12	0.09	0.18	0	0.14	0.49	0.04	0.15	-0.11	0.05	0.32	0.22	1	0.49	-0.02	0.32
PITCHING_SO	-0.08	-0.25	0.06	-0.26	0.18	-0.02	0.42	-0.14	-0.21	0.22	0.27	0.21	0.49	1	-0.02	0.03
FIELDING_E	-0.18	0.26	-0.24	0.51	-0.59	-0.66	-0.58	0.51	0.05	0.04	0.67	-0.49	-0.02	-0.02	1	-0.5
FIELDING_DP	-0.03	0.16	0.29	-0.32	0.45	0.43	0.15	-0.5	-0.21	-0.07	-0.23	0.44	0.32	0.03	-0.5	1