The goal of the project is to predict number of wins of each basketball team using their performance statistics. The performance statistics are thing like hits by betters, walks by betters and errors, Each of these measure has positive or negative impact to wins. The data set contain 15 measures and each of measures has about 2000 records. Because of the variation of the team ability, most of measures range from hundreds to thousands and most of measures also has mean and median around hundreds to thousands. In addition, some measures have missing values like strikeout by batters and strikeout by pitchers which may affect the analysis later. Among all measures, batters hits by pitch has most missing values; 2058 missing values. Thus this measure could mislead the analysis. On the other hand, some measures are correlated. For example. Homeruns by batters and homeruns allowed are highly correlated whose correlation is 0.97165. Other correlated variable also include walks by batters and walks allowed.

Before start modeling, we need to clean data. Fist we need to remove batters hits by pitch because it has to many missing values and for other measures with missing values like stolen bases, we will impute the variable using predictive mean matching method. For highly correlated variables, we will remove one variable for each pair of correlated variables. Specifically, homeruns allowed and walks allowed will be removed. We will also do data transformation when modeling.

We build threes models. First we build a model according to all variables. We see in the first model, most variable has positive coefficient but some of them doesn't make sense. For example, caught stealing has positive coefficient but it doesn't make sense that a team has a lot of caught stealing will expect to have more wins. In addition, double play has negative coefficient which doesn't make sense since it's a very good move in the game. Although, the p-value of the model suggest the model is statistically significant, some variable doses not make sense intuitively and statistically, we build a second model using only those most significant variables. In the second model, we see most variable's coefficient are increased compare to the fist model. Finally, we build the third using box-cox transformation.

Among all three models, we see p-value are all statistical significant, all R square are around 0.35 with the first model has highest value since it has more variables. The R square indicate it that it doesn't have a very high explanation of the model variation. However the Q-Q plot shows that the residual are normally distribution and the residual plot also not show a clear trend. So the model meets the linear model assumption very well. Since model 2 and model 3 has similar statistics, we decide to use model 2 since the Q-Q plot for model 2 is better than model 3. We print the result using model 2.

```
        fit      lwr      upr
1  61.48862 36.514975  86.46227
2  63.67474 38.717636  88.63185
3  73.83651 48.898054  98.77497
4  89.51893 64.578281 114.45957
5  62.29474 37.317316  87.27216
6  68.75403 43.787064  93.72100
7  84.02502 59.058804 108.99124
8  77.39563 52.451508 102.33975
9  69.05880 44.105203  94.01239
10 74.35175 49.407610  99.29590
11 69.67720 44.714221  94.64018
12 82.91626 57.960845 107.87167
13 81.05194 56.087042 106.01683
14 84.17889 59.215180 109.14260
15 86.91114 61.937809 111.88447
16 78.53584 53.586921 103.48476
17 74.09092 49.143836  99.03801
18 78.66637 53.722303 103.61044
19 72.38247 47.409252  97.35569
20 91.98883 67.024054 116.95361
21 81.43269 56.483372 106.38201
```

| | | | |
|---|---|---|---|
| 22 | 84.27248 | 59.326310 | 109.21865 |
| 23 | 78.00715 | 53.068815 | 102.94548 |
| 24 | 72.64091 | 47.691671 | 97.59015 |
| 25 | 84.69136 | 59.754457 | 109.62826 |
| 26 | 90.45868 | 65.518348 | 115.39901 |
| 27 | 55.83449 | 30.727730 | 80.94126 |
| 28 | 76.41529 | 51.468891 | 101.36169 |
| 29 | 83.32339 | 58.368428 | 108.27836 |
| 30 | 77.12495 | 52.165679 | 102.08423 |
| 31 | 89.29254 | 64.347984 | 114.23710 |
| 32 | 84.79803 | 59.854306 | 109.74176 |
| 33 | 80.06604 | 55.126563 | 105.00551 |
| 34 | 80.83789 | 55.893229 | 105.78255 |
| 35 | 78.38477 | 53.445595 | 103.32394 |
| 36 | 86.23676 | 61.262732 | 111.21078 |
| 37 | 75.57683 | 50.640237 | 100.51342 |
| 38 | 89.04392 | 64.087337 | 114.00051 |
| 39 | 85.54446 | 60.602615 | 110.48631 |
| 40 | 91.18025 | 66.214563 | 116.14594 |
| 41 | 84.73920 | 59.796462 | 109.68194 |
| 42 | 91.41862 | 66.464427 | 116.37281 |
| 43 | 29.95729 | 4.504833 | 55.40974 |
| 44 | 106.15781 | 81.108336 | 131.20729 |
| 45 | 95.28580 | 70.302997 | 120.26859 |
| 46 | 91.66196 | 66.699343 | 116.62459 |
| 47 | 100.88340 | 75.902086 | 125.86471 |
| 48 | 76.98135 | 52.032759 | 101.92993 |
| 49 | 68.41863 | 43.467491 | 93.36977 |
| 50 | 80.37283 | 55.421894 | 105.32376 |
| 51 | 78.73984 | 53.791877 | 103.68780 |
| 52 | 87.38978 | 62.440488 | 112.33908 |
| 53 | 75.18487 | 50.240889 | 100.12884 |
| 54 | 74.00575 | 49.064638 | 98.94687 |
| 55 | 75.59246 | 50.654311 | 100.53062 |
| 56 | 79.74824 | 54.812973 | 104.68350 |
| 57 | 90.96616 | 66.001755 | 115.93057 |
| 58 | 77.23403 | 52.272949 | 102.19511 |
| 59 | 62.99083 | 38.014219 | 87.96744 |
| 60 | 78.21019 | 53.262693 | 103.15769 |
| 61 | 88.68404 | 63.729915 | 113.63816 |
| 62 | 70.58663 | 45.616201 | 95.55705 |
| 63 | 88.53905 | 63.596556 | 113.48155 |
| 64 | 87.54466 | 62.574560 | 112.51475 |
| 65 | 86.90231 | 61.947953 | 111.85666 |
| 66 | 108.12364 | 83.076620 | 133.17067 |
| 67 | 72.56794 | 47.622527 | 97.51336 |
| 68 | 79.19379 | 54.236334 | 104.15125 |
| 69 | 78.13997 | 53.180052 | 103.09989 |
| 70 | 85.09249 | 60.135897 | 110.04908 |
| 71 | 82.94575 | 57.979320 | 107.91219 |
| 72 | 71.90785 | 46.921733 | 96.89397 |
| 73 | 77.03308 | 52.059140 | 102.00702 |
| 74 | 89.38295 | 64.389179 | 114.37672 |
| 75 | 81.92393 | 56.955324 | 106.89253 |
| 76 | 83.20713 | 58.254211 | 108.16006 |
| 77 | 80.83363 | 55.895126 | 105.77213 |
| 78 | 84.70283 | 59.759201 | 109.64646 |
| 79 | 74.68219 | 49.725960 | 99.63842 |
| 80 | 78.01550 | 53.064886 | 102.96611 |
| 81 | 87.11763 | 62.163989 | 112.07128 |
| 82 | 88.10592 | 63.157480 | 113.05435 |
| 83 | 97.17046 | 72.202196 | 122.13871 |
| 84 | 74.54375 | 49.570860 | 99.51663 |
| 85 | 82.27063 | 57.329101 | 107.21215 |

| | | | |
|---|---|---|---|
| 86 | 83.19936 | 58.249729 | 108.14898 |
| 87 | 84.39458 | 59.443537 | 109.34561 |
| 88 | 83.65631 | 58.722213 | 108.59041 |
| 89 | 89.87099 | 64.922053 | 114.81992 |
| 90 | 91.09855 | 66.145593 | 116.05150 |
| 91 | 83.18296 | 58.226241 | 108.13969 |
| 92 | 83.57297 | 58.365967 | 108.77997 |
| 93 | 73.94556 | 48.989341 | 98.90178 |
| 94 | 88.00328 | 63.036080 | 112.97049 |
| 95 | 87.62485 | 62.661852 | 112.58784 |
| 96 | 84.27068 | 59.309802 | 109.23155 |
| 97 | 84.86999 | 59.919992 | 109.81998 |
| 98 | 99.29347 | 74.295489 | 124.29146 |
| 99 | 84.28220 | 59.323876 | 109.24053 |
| 100 | 85.04485 | 60.077711 | 110.01200 |
| 101 | 79.11703 | 54.163478 | 104.07059 |
| 102 | 74.87355 | 49.919985 | 99.82712 |
| 103 | 84.13077 | 59.189649 | 109.07190 |
| 104 | 82.96378 | 58.006713 | 107.92086 |
| 105 | 78.02534 | 53.060295 | 102.99038 |
| 106 | 69.69757 | 44.731039 | 94.66410 |
| 107 | 66.03548 | 40.944943 | 91.12601 |
| 108 | 78.52656 | 53.569503 | 103.48363 |
| 109 | 86.83456 | 61.885955 | 111.78316 |
| 110 | 59.10053 | 34.063259 | 84.13781 |
| 111 | 84.98945 | 60.045854 | 109.93304 |
| 112 | 88.75162 | 63.798185 | 113.70505 |
| 113 | 95.00767 | 70.060546 | 119.95480 |
| 114 | 93.25532 | 68.313825 | 118.19682 |
| 115 | 82.53415 | 57.600057 | 107.46824 |
| 116 | 79.69012 | 54.746803 | 104.63343 |
| 117 | 85.62777 | 60.666365 | 110.58917 |
| 118 | 82.13002 | 57.193832 | 107.06620 |
| 119 | 75.30252 | 50.352647 | 100.25239 |
| 120 | 77.56087 | 52.577006 | 102.54473 |
| 121 | 96.40398 | 71.425088 | 121.38287 |
| 122 | 69.74443 | 44.778394 | 94.71046 |
| 123 | 67.88779 | 42.926013 | 92.84957 |
| 124 | 64.65830 | 39.646368 | 89.67024 |
| 125 | 67.15112 | 42.195661 | 92.10658 |
| 126 | 89.16025 | 64.202599 | 114.11790 |
| 127 | 89.79150 | 64.821180 | 114.76181 |
| 128 | 77.02210 | 52.077416 | 101.96678 |
| 129 | 93.12606 | 68.164761 | 118.08736 |
| 130 | 91.44315 | 66.484181 | 116.40211 |
| 131 | 85.38163 | 60.433830 | 110.32943 |
| 132 | 81.11785 | 56.172780 | 106.06292 |
| 133 | 81.60071 | 56.657622 | 106.54380 |
| 134 | 83.77266 | 58.815499 | 108.72981 |
| 135 | 87.11397 | 62.162460 | 112.06547 |
| 136 | 73.19837 | 48.215390 | 98.18134 |
| 137 | 73.97182 | 49.029828 | 98.91381 |
| 138 | 78.88143 | 53.945026 | 103.81784 |
| 139 | 91.17467 | 66.207938 | 116.14140 |
| 140 | 81.65655 | 56.720929 | 106.59217 |
| 141 | 65.10875 | 40.136790 | 90.08071 |
| 142 | 71.09900 | 46.135961 | 96.06205 |
| 143 | 89.64738 | 64.685671 | 114.60910 |
| 144 | 71.57347 | 46.621825 | 96.52511 |
| 145 | 71.18772 | 46.237764 | 96.13767 |
| 146 | 71.09695 | 46.152211 | 96.04169 |
| 147 | 76.43929 | 51.500954 | 101.37762 |
| 148 | 78.82154 | 53.879571 | 103.76351 |
| 149 | 79.14428 | 54.181950 | 104.10662 |

| | | | |
|---|---|---|---|
| 150 | 84.25577 | 59.316586 | 109.19496 |
| 151 | 82.58268 | 57.635134 | 107.53023 |
| 152 | 81.82932 | 56.877527 | 106.78112 |
| 153 | 50.14708 | 23.714352 | 76.57980 |
| 154 | 70.03199 | 45.083057 | 94.98093 |
| 155 | 76.21476 | 51.269348 | 101.16017 |
| 156 | 71.01947 | 46.062305 | 95.97663 |
| 157 | 90.66366 | 65.705399 | 115.62193 |
| 158 | 74.84355 | 49.862938 | 99.82417 |
| 159 | 88.52756 | 63.569723 | 113.48539 |
| 160 | 73.85675 | 48.906619 | 98.80689 |
| 161 | 99.39923 | 74.429466 | 124.36900 |
| 162 | 104.15682 | 79.183155 | 129.13048 |
| 163 | 92.46969 | 67.509588 | 117.42980 |
| 164 | 100.87353 | 75.895652 | 125.85141 |
| 165 | 94.62287 | 69.652156 | 119.59359 |
| 166 | 87.55369 | 62.588901 | 112.51847 |
| 167 | 79.74768 | 54.795681 | 104.69968 |
| 168 | 82.85450 | 57.885612 | 107.82339 |
| 169 | 74.85876 | 49.913809 | 99.80371 |
| 170 | 82.27507 | 57.337997 | 107.21213 |
| 171 | 87.16939 | 62.210473 | 112.12831 |
| 172 | 88.33442 | 63.383783 | 113.28506 |
| 173 | 80.29577 | 55.350040 | 105.24149 |
| 174 | 94.16659 | 69.201923 | 119.13127 |
| 175 | 83.50785 | 58.566072 | 108.44963 |
| 176 | 71.90474 | 46.954311 | 96.85516 |
| 177 | 76.52101 | 51.574343 | 101.46768 |
| 178 | 69.91971 | 44.959877 | 94.87954 |
| 179 | 73.47794 | 48.537466 | 98.41842 |
| 180 | 78.15137 | 53.214247 | 103.08849 |
| 181 | 90.71027 | 65.703671 | 115.71687 |
| 182 | 88.37827 | 63.415474 | 113.34107 |
| 183 | 86.48591 | 61.548599 | 111.42321 |
| 184 | 84.51451 | 59.559818 | 109.46920 |
| 185 | 84.18151 | 58.994207 | 109.36881 |
| 186 | 100.24622 | 75.189722 | 125.30271 |
| 187 | 83.31218 | 58.350159 | 108.27419 |
| 188 | 67.30016 | 42.217671 | 92.38265 |
| 189 | 67.47508 | 42.484931 | 92.46523 |
| 190 | 115.09169 | 90.050358 | 140.13302 |
| 191 | 72.26845 | 47.313448 | 97.22345 |
| 192 | 83.05838 | 58.108052 | 108.00871 |
| 193 | 78.89088 | 53.955093 | 103.82667 |
| 194 | 79.26979 | 54.322885 | 104.21669 |
| 195 | 81.80562 | 56.851285 | 106.75997 |
| 196 | 69.25553 | 44.301879 | 94.20919 |
| 197 | 79.21506 | 54.275510 | 104.15461 |
| 198 | 82.57312 | 57.609457 | 107.53679 |
| 199 | 78.00342 | 53.059972 | 102.94686 |
| 200 | 81.29486 | 56.351308 | 106.23841 |
| 201 | 72.28043 | 47.326174 | 97.23468 |
| 202 | 77.84316 | 52.895974 | 102.79034 |
| 203 | 71.30650 | 46.347864 | 96.26514 |
| 204 | 91.59659 | 66.646872 | 116.54631 |
| 205 | 82.90431 | 57.970235 | 107.83839 |
| 206 | 81.89635 | 56.952680 | 106.84002 |
| 207 | 77.71451 | 52.759844 | 102.66918 |
| 208 | 78.06879 | 53.119940 | 103.01763 |
| 209 | 81.49022 | 56.546045 | 106.43440 |
| 210 | 73.76590 | 48.796315 | 98.73548 |
| 211 | 102.85420 | 77.866748 | 127.84165 |
| 212 | 92.71925 | 67.751227 | 117.68727 |
| 213 | 78.66968 | 53.732802 | 103.60656 |

```
214  64.73205 39.773519  89.69058
215  67.14618 42.187435  92.10493
216  82.00861 57.064893 106.95233
217  77.57519 52.618352 102.53203
218  95.56224 70.603657 120.52083
219  78.96437 54.028907 103.89983
220  78.00288 53.062363 102.94340
221  77.85396 52.903229 102.80470
222  76.13682 51.177888 101.09575
223  81.57726 56.631409 106.52311
224  73.60313 48.639426  98.56684
225  72.30206 47.064609  97.53951
226  74.24843 49.309469  99.18738
227  82.51046 57.573628 107.44730
228  78.10789 53.152122 103.06365
229  81.12438 56.184681 106.06407
230  71.67729 46.686723  96.66786
231  82.36711 57.384190 107.35003
232  92.35642 67.400407 117.31243
233  76.70101 51.741381 101.66063
234  89.56377 64.603765 114.52377
235  79.20559 54.268807 104.14238
236  74.53168 49.589026  99.47434
237  82.28975 57.339284 107.24022
238  77.18183 52.244427 102.11923
239  90.85400 65.876419 115.83157
240  71.59921 46.653149  96.54528
241  87.81774 62.876039 112.75945
242  86.55006 61.599494 111.50063
243  84.27162 59.323166 109.22007
244  83.29596 58.349715 108.24221
245  61.28326 36.308931  86.25759
246  89.81230 64.867177 114.75743
247  82.44450 57.510250 107.37875
248  86.22102 61.278074 111.16397
249  73.91320 48.967559  98.85885
250  85.50038 60.535803 110.46496
251  80.64928 55.698522 105.60003
252  77.68952 52.611922 102.76712
253  95.38387 70.415366 120.35237
254  18.73444 -6.571613  44.04049
255  69.45428 44.509829  94.39872
256  75.20836 50.268290 100.14843
257  84.65302 59.700139 109.60590
258  85.99498 61.046817 110.94314
259  79.20042 54.255460 104.14538
```

Appendix

## Load Data

```{r cars}
file1 <-'C:/Users/andre/Downloads/moneyball-evaluation-data.csv'
eval<-read.csv(file1)
file2 <-'C:/Users/andre/Downloads/moneyball-training-data.csv'
train<-read.csv(file2)
summary(train)
```

## Find Outlier

```{r pressure, echo=FALSE}
ggplot(stack(train), aes(x = ind, y = values)) + geom_boxplot() + coord_cartesian(ylim = c(0, 2000)) +
   theme(legend.position="none") +
   theme(axis.text.x=element_text(angle=45, hjust=1)) +
   theme(panel.background = element_rect(fill = 'grey'))
```
## Remove TEAM_BATTING_HBP since it has mostly missing values
```{r}
train <- train[,-1 ]
train <- train[, -10]
```

## Find correlations
```{r}
cor(drop_na(train))
```
TEAM_PITCHING_HR/BB and TEAM_BATTING_HR/BB are highly correlated, so we can remove one of them.

```{r}
train <- train[, -11]
train<-train[,-11]
```
```{r}
refin_data <- mice(train, m=5, maxit = 5, method = 'pmm')
refin_data <- complete(refin_data)
```
```{r}
model1 <- lm(TARGET_WINS ~., refin_data)
summary(model1)
```
```{r}
model2 <- lm(TARGET_WINS ~ TEAM_BATTING_H + TEAM_BATTING_HR +
TEAM_BATTING_SO + TEAM_BASERUN_SB + TEAM_PITCHING_H + TEAM_FIELDING_E
+ TEAM_FIELDING_DP , refin_data)
summary(model2)
plot(fitted(model2),residuals(model2),xlab = "Fitted", ylab = "Residuals")
abline(h=0)
par(mfrow = c(1,1))
qqnorm(residuals(model2),ylab = "Residuals")
qqline(residuals(model2))
```
## transfrom data
```{r}
library(caret)
library(e1071)
t = preProcess(refin_data,
                           c("BoxCox", "center", "scale"))
refin_data = data.frame(
```

```
        t = predict(t, refin_data))
```
## m3
```{r}
model3 <- lm(t.TARGET_WINS ~ t.TEAM_BATTING_H + t.TEAM_BATTING_HR +
t.TEAM_BATTING_SO + t.TEAM_BASERUN_SB + t.TEAM_PITCHING_H +
t.TEAM_FIELDING_E + t.TEAM_FIELDING_DP , refin_data)
summary(model3)
plot(fitted(model3),residuals(model3),xlab = "Fitted", ylab = "Residuals")
abline(h=0)
par(mfrow = c(1,1))
qqnorm(residuals(model3),ylab = "Residuals")
qqline(residuals(model3))
```
```{r}
eval <- eval[,-1 ]
eval <- eval[, -9]
eval<-eval[,-10]
eval<-eval[,-10]
```
```{r}
refin_val <- mice(eval, m=5, maxit = 5, method = 'pmm')
refin_val <- complete(refin_val)
```
```{r}
t = preProcess(refin_val,
                        c("BoxCox", "center", "scale"))
refin_val = data.frame(
        t = predict(t, refin_val))
```
```{r}
eval_data <- predict(model2, newdata = refin_val, interval="prediction")
summary(eval_data)
eval_data
```