Project 1

Contents

Data Exploration

Data Preparation

```
##            variable q_zeros p_zeros q_na  p_na q_inf p_inf    type unique
## 1             INDEX       0    0.00    0  0.00     0     0 integer    259
## 2    TEAM_BATTING_H       0    0.00    0  0.00     0     0 integer    199
## 3   TEAM_BATTING_2B       0    0.00    0  0.00     0     0 integer    137
## 4   TEAM_BATTING_3B       0    0.00    0  0.00     0     0 integer     91
## 5   TEAM_BATTING_HR       1    0.39    0  0.00     0     0 integer    141
## 6   TEAM_BATTING_BB       0    0.00    0  0.00     0     0 integer    185
## 7   TEAM_BATTING_SO       2    0.77   18  6.95     0     0 integer    206
## 8   TEAM_BASERUN_SB       1    0.39   13  5.02     0     0 integer    153
## 9   TEAM_BASERUN_CS       1    0.39   87 33.59     0     0 integer     68
## 10 TEAM_BATTING_HBP       0    0.00  240 92.66     0     0 integer     17
## 11  TEAM_PITCHING_H       0    0.00    0  0.00     0     0 integer    210
## 12 TEAM_PITCHING_HR       1    0.39    0  0.00     0     0 integer    144
## 13 TEAM_PITCHING_BB       0    0.00    0  0.00     0     0 integer    199
## 14 TEAM_PITCHING_SO       2    0.77   18  6.95     0     0 integer    205
## 15  TEAM_FIELDING_E       0    0.00    0  0.00     0     0 integer    168
## 16 TEAM_FIELDING_DP       0    0.00   31 11.97     0     0 integer     90
```

```
##            variable q_zeros p_zeros q_na  p_na q_inf p_inf    type unique
## 7   TEAM_BATTING_SO       2    0.77   18  6.95     0     0 integer    206
## 14 TEAM_PITCHING_SO       2    0.77   18  6.95     0     0 integer    205
## 5   TEAM_BATTING_HR       1    0.39    0  0.00     0     0 integer    141
## 8   TEAM_BASERUN_SB       1    0.39   13  5.02     0     0 integer    153
## 9   TEAM_BASERUN_CS       1    0.39   87 33.59     0     0 integer     68
## 12 TEAM_PITCHING_HR       1    0.39    0  0.00     0     0 integer    144
## 1             INDEX       0    0.00    0  0.00     0     0 integer    259
## 2    TEAM_BATTING_H       0    0.00    0  0.00     0     0 integer    199
## 3   TEAM_BATTING_2B       0    0.00    0  0.00     0     0 integer    137
## 4   TEAM_BATTING_3B       0    0.00    0  0.00     0     0 integer     91
## 6   TEAM_BATTING_BB       0    0.00    0  0.00     0     0 integer    185
## 10 TEAM_BATTING_HBP       0    0.00  240 92.66     0     0 integer     17
## 11  TEAM_PITCHING_H       0    0.00    0  0.00     0     0 integer    210
## 13 TEAM_PITCHING_BB       0    0.00    0  0.00     0     0 integer    199
## 15  TEAM_FIELDING_E       0    0.00    0  0.00     0     0 integer    168
## 16 TEAM_FIELDING_DP       0    0.00   31 11.97     0     0 integer     90
```

```
##            variable q_zeros p_zeros q_na  p_na q_inf p_inf    type unique
## 1             INDEX       0    0.00    0  0.00     0     0 integer   2276
## 2       TARGET_WINS       1    0.04    0  0.00     0     0 integer    108
## 3    TEAM_BATTING_H       0    0.00    0  0.00     0     0 integer    569
## 4   TEAM_BATTING_2B       0    0.00    0  0.00     0     0 integer    240
## 5   TEAM_BATTING_3B       2    0.09    0  0.00     0     0 integer    144
## 6   TEAM_BATTING_HR      15    0.66    0  0.00     0     0 integer    243
## 7   TEAM_BATTING_BB       1    0.04    0  0.00     0     0 integer    533
## 8   TEAM_BATTING_SO      20    0.88  102  4.48     0     0 integer    822
## 9   TEAM_BASERUN_SB       2    0.09  131  5.76     0     0 integer    348
## 10  TEAM_BASERUN_CS       1    0.04  772 33.92     0     0 integer    128
## 11 TEAM_BATTING_HBP       0    0.00 2085 91.61     0     0 integer     55
## 12  TEAM_PITCHING_H       0    0.00    0  0.00     0     0 integer    843
## 13 TEAM_PITCHING_HR      15    0.66    0  0.00     0     0 integer    256
## 14 TEAM_PITCHING_BB       1    0.04    0  0.00     0     0 integer    535
## 15 TEAM_PITCHING_SO      20    0.88  102  4.48     0     0 integer    823
## 16  TEAM_FIELDING_E       0    0.00    0  0.00     0     0 integer    549
## 17 TEAM_FIELDING_DP       0    0.00  286 12.57     0     0 integer    144
```

```
##            variable q_zeros p_zeros q_na  p_na q_inf p_inf    type unique
## 11 TEAM_BATTING_HBP       0    0.00 2085 91.61     0     0 integer     55
## 10  TEAM_BASERUN_CS       1    0.04  772 33.92     0     0 integer    128
## 17 TEAM_FIELDING_DP       0    0.00  286 12.57     0     0 integer    144
## 9   TEAM_BASERUN_SB       2    0.09  131  5.76     0     0 integer    348
## 8   TEAM_BATTING_SO      20    0.88  102  4.48     0     0 integer    822
## 15 TEAM_PITCHING_SO      20    0.88  102  4.48     0     0 integer    823
## 1             INDEX       0    0.00    0  0.00     0     0 integer   2276
## 2       TARGET_WINS       1    0.04    0  0.00     0     0 integer    108
## 3    TEAM_BATTING_H       0    0.00    0  0.00     0     0 integer    569
## 4   TEAM_BATTING_2B       0    0.00    0  0.00     0     0 integer    240
## 5   TEAM_BATTING_3B       2    0.09    0  0.00     0     0 integer    144
## 6   TEAM_BATTING_HR      15    0.66    0  0.00     0     0 integer    243
## 7   TEAM_BATTING_BB       1    0.04    0  0.00     0     0 integer    533
## 12  TEAM_PITCHING_H       0    0.00    0  0.00     0     0 integer    843
## 13 TEAM_PITCHING_HR      15    0.66    0  0.00     0     0 integer    256
## 14 TEAM_PITCHING_BB       1    0.04    0  0.00     0     0 integer    535
## 16  TEAM_FIELDING_E       0    0.00    0  0.00     0     0 integer    549
```

```
##            variable q_zeros p_zeros q_na p_na q_inf p_inf    type unique
## 1             INDEX       0    0.00    0    0     0     0 integer   2276
## 2       TARGET_WINS       1    0.04    0    0     0     0 integer    108
## 3    TEAM_BATTING_H       0    0.00    0    0     0     0 integer    569
## 4   TEAM_BATTING_2B       0    0.00    0    0     0     0 integer    240
## 5   TEAM_BATTING_3B       2    0.09    0    0     0     0 integer    144
## 6   TEAM_BATTING_HR      15    0.66    0    0     0     0 integer    243
## 7   TEAM_BATTING_BB       1    0.04    0    0     0     0 integer    533
## 8   TEAM_BATTING_SO      20    0.88    0    0     0     0 numeric    823
## 9   TEAM_BASERUN_SB       2    0.09    0    0     0     0 numeric    349
## 10  TEAM_BASERUN_CS       1    0.04    0    0     0     0 numeric    129
## 11 TEAM_BATTING_HBP       0    0.00    0    0     0     0 numeric     56
## 12  TEAM_PITCHING_H       0    0.00    0    0     0     0 integer    843
## 13 TEAM_PITCHING_HR      15    0.66    0    0     0     0 integer    256
## 14 TEAM_PITCHING_BB       1    0.04    0    0     0     0 integer    535
## 15 TEAM_PITCHING_SO      20    0.88    0    0     0     0 numeric    824
## 16  TEAM_FIELDING_E       0    0.00    0    0     0     0 integer    549
```

```
## 17 TEAM_FIELDING_DP       0    0.00    0    0     0     0 numeric    145
## 18      HBP_missing     191    8.39    0    0     0     0 numeric      2
## 19       CS_missing    1504   66.08    0    0     0     0 numeric      2
## 20       DP_missing    1990   87.43    0    0     0     0 numeric      2
```

Build Model

Model 1

Model 1 is based on reviewing the p value from a model with all the predictors fitted. After reviewing the predictors that were statistically significant, a decission was made to fit this model with three statically significant predictors: TEAM_BATTING_H, TEAM_BASERUN_SB, and TEAM_FIELDING_E. The coefficients in this model tells us that there are possitive coefficients for TEAM_BATTING_H (Base Hits by batters) and TEAM_BASERUN_SB (Stolen bases) which makes sense since both of ther predictors will cause target wins to go up. The predictor with negative coefficient is TEAM_FIELDING_E(errors) which will indeed cause your teams wins to go down. This model has an R^2 of 0.27 and the F-statistics is 284

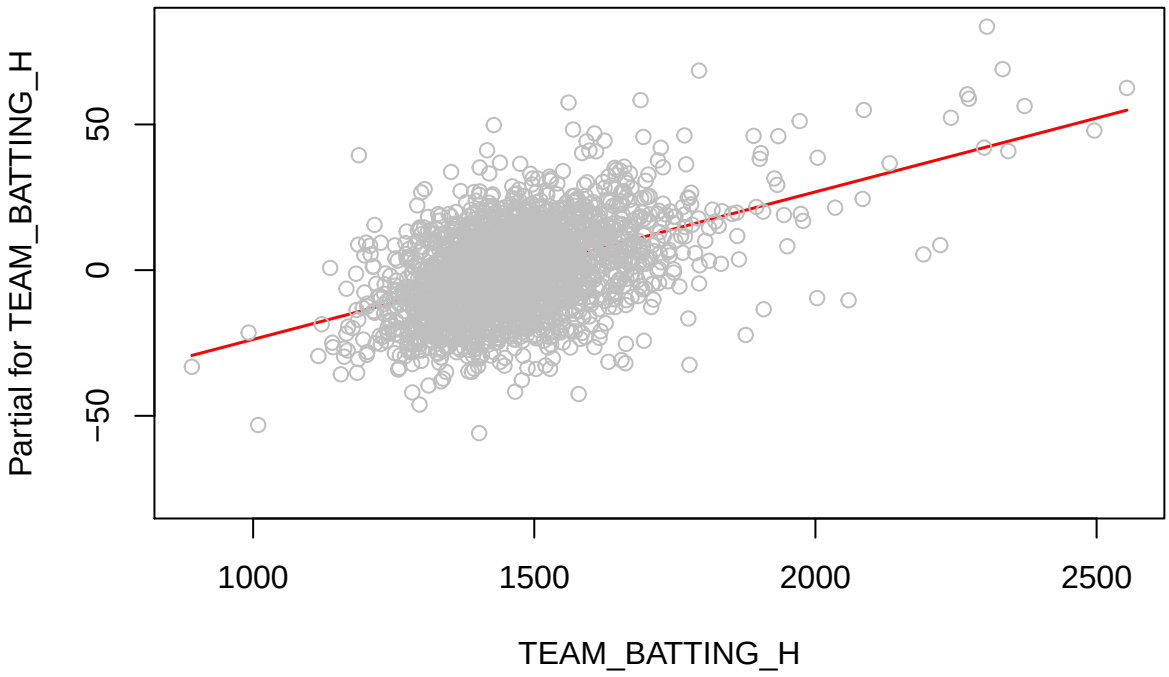
```
##
## Call:
## lm(formula = TARGET_WINS ~ (TEAM_BATTING_H + TEAM_BASERUN_SB +
##     TEAM_FIELDING_E), data = train_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -52.536  -8.953   0.136   8.776  53.676
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      8.010983   2.928195   2.736  0.00627 **
## TEAM_BATTING_H   0.050627   0.002022  25.043  < 2e-16 ***
## TEAM_BASERUN_SB  0.038590   0.003558  10.847  < 2e-16 ***
## TEAM_FIELDING_E -0.026042   0.001372 -18.987  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.44 on 2272 degrees of freedom
## Multiple R-squared:  0.2728, Adjusted R-squared:  0.2718
## F-statistic: 284.1 on 3 and 2272 DF,  p-value: < 2.2e-16
```

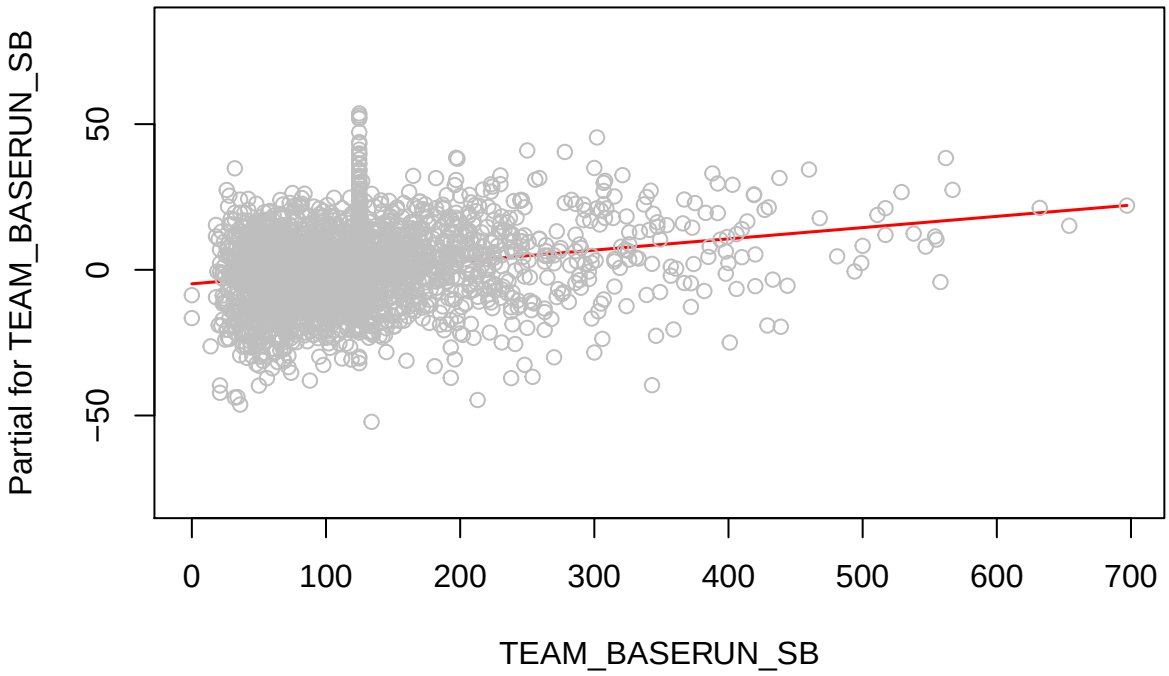
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	8.011	2.928	2.736	0.006271
TEAM_BATTING_H	0.05063	0.002022	25.04	1.979e-122
TEAM_BASERUN_SB	0.03859	0.003558	10.85	9.252e-27
TEAM_FIELDING_E	-0.02604	0.001372	-18.99	9.866e-75

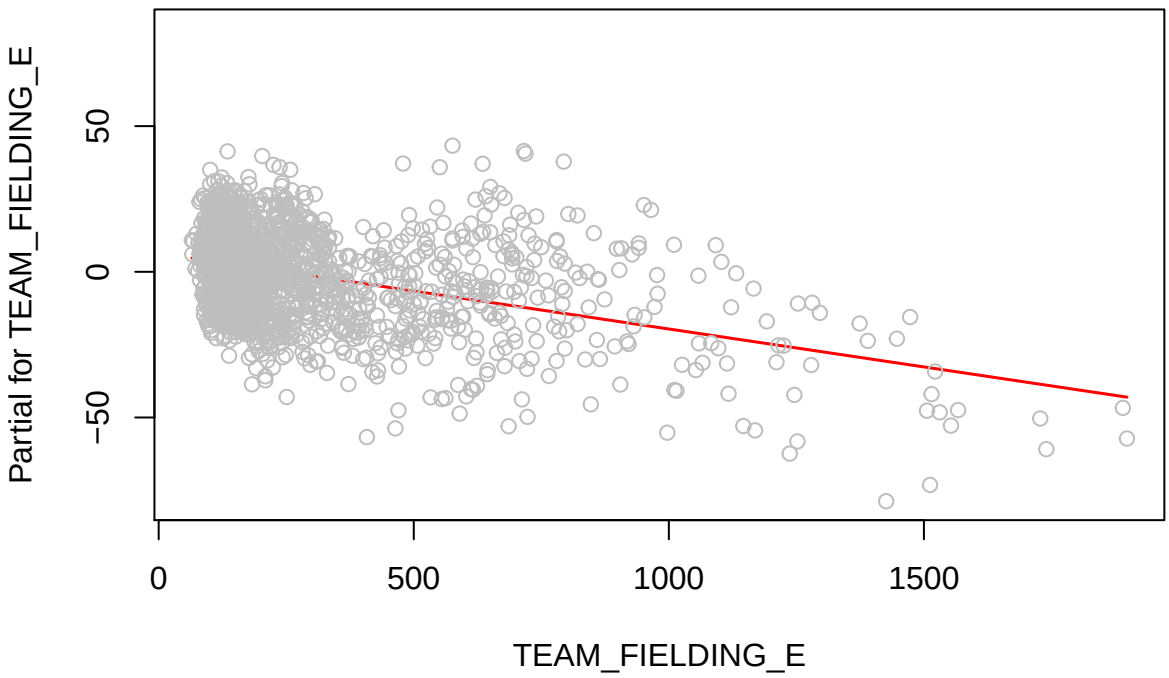
Table 2: Fitting linear model: TARGET_WINS ~ (TEAM_BATTING_H + TEAM_BASERUN_SB + TEAM_FIELDING_E)

Observations	Residual Std. Error	R^2	Adjusted R^2
2276	13.44	0.2728	0.2718

The termplot below gives us a visual of the coefficients from model 1 agains the slopes of each of the predictors.




Partial for TEAM_BASERUN_SB
50
0
−50
0
100
200
300
400
500
600
700
TEAM_BASERUN_SB



R Code

```
# Data Preparation
eva_data <- read.csv(file = "moneyball-evaluation-data.csv",
    header = TRUE, sep = ",")
train_data <- read.csv(file = "moneyball-training-data.csv",
    header = TRUE, sep = ",")

eva_data_status <- df_status(eva_data)
eva_data_status[order(-eva_data_status$p_zeros), ]

train_data_status <- df_status(train_data)
train_data_status[order(-train_data_status$p_na), ]

train_data$HBP_missing <- ifelse(is.na(train_data$TEAM_BATTING_HBP),
    1, 0)
train_data$TEAM_BATTING_HBP[is.na(train_data$TEAM_BATTING_HBP)] <- mean(train_data$TEAM_BATTING_HBP,
    na.rm = TRUE)

train_data$CS_missing <- ifelse(is.na(train_data$TEAM_BASERUN_CS),
    1, 0)
train_data$TEAM_BASERUN_CS[is.na(train_data$TEAM_BASERUN_CS)] <- mean(train_data$TEAM_BASERUN_CS,
    na.rm = TRUE)

train_data$DP_missing <- ifelse(is.na(train_data$TEAM_FIELDING_DP),
```

```
    1, 0)
train_data$TEAM_FIELDING_DP[is.na(train_data$TEAM_FIELDING_DP)] <- mean(train_data$TEAM_FIELDING_DP,
    na.rm = TRUE)

train_data$TEAM_BASERUN_SB[is.na(train_data$TEAM_BASERUN_SB)] <- mean(train_data$TEAM_BASERUN_SB,
    na.rm = TRUE)
train_data$TEAM_BATTING_SO[is.na(train_data$TEAM_BATTING_SO)] <- mean(train_data$TEAM_BATTING_SO,
    na.rm = TRUE)
train_data$TEAM_PITCHING_SO[is.na(train_data$TEAM_PITCHING_SO)] <- mean(train_data$TEAM_PITCHING_SO,
    na.rm = TRUE)

df_status(train_data)
```

```
# Build Model

# This model contains all the predictor variables from our
# training dataset. All significant and non-significan
# predictors are displayed in the regression model below:
# remove the index column from the train data.
train_data <- select(train_data, -c(INDEX))
model_all_predictors <- lm(TARGET_WINS ~ ., data = train_data)
summary(model_all_predictors)

## Model 1
model1 <- lm(TARGET_WINS ~ (TEAM_BATTING_H + TEAM_BASERUN_SB +
    TEAM_FIELDING_E), data = train_data)
summary(model1)
pander(summary(model1))
termplot(model1)
```