621 HW 1 v 4

Jeff Nieman, Scott Karr, James Topor, Armenoush

June 13, 2016

Eliminate HBP, CS, and pitching HR's.

Build model for batting SO using Gelman approach

Build model for Pitching SO

Build model for SB

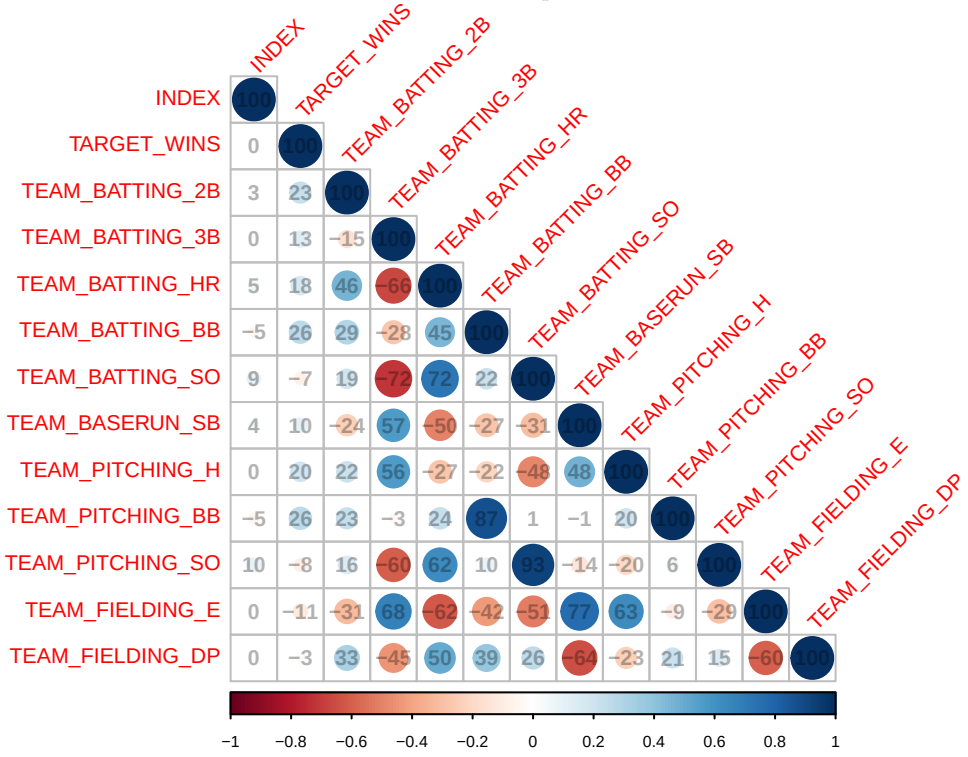
Build model to replace DP

Eliminate unhistorical outliers - DO THIS FOR THE EVAL DATA AS WELL

SINGLE PREDICTOR ANALYSIS & TRANSFORMATIONS

Model SMK Generlized Equation Review descriptive statistics to confirm each variable is within acceptable bounds and contains no missing data. Review Density plots of 13 variables for skewness to identify which may require transformation.

Evaluate Correlations Evaluate Correlation between predictors so as to not introduce collinearity into the model.



```
## [1] 2157
```

Model Selection Strategy Two common strategies for adding or removing variables in a multiple regression model are called backward elimination and forward selection. These techniques are often referred to as stepwise model selection strategies, because they add or delete one variable at a time as they "step" through the candidate predictors. Model 1 uses the forward selection strategy which adds variables two-at-a-time until variables cannot be found that improve the model as measured by adjusted R^2. Diez, D.M., Barr, C.D., & Çetinkaya-Rundel, M. (2015). OpenIntro Statistics (3rd Ed). pg. 378

Start with p.Hits & p.Walks

Add b.Singles & b.Doubles

Removed p.Hits

Added Stolen Bases and Double Plays

Added b.Walks and p.Strikeouts

Remove b.Walks

Add b.StrikeOuts, & b.Slugging

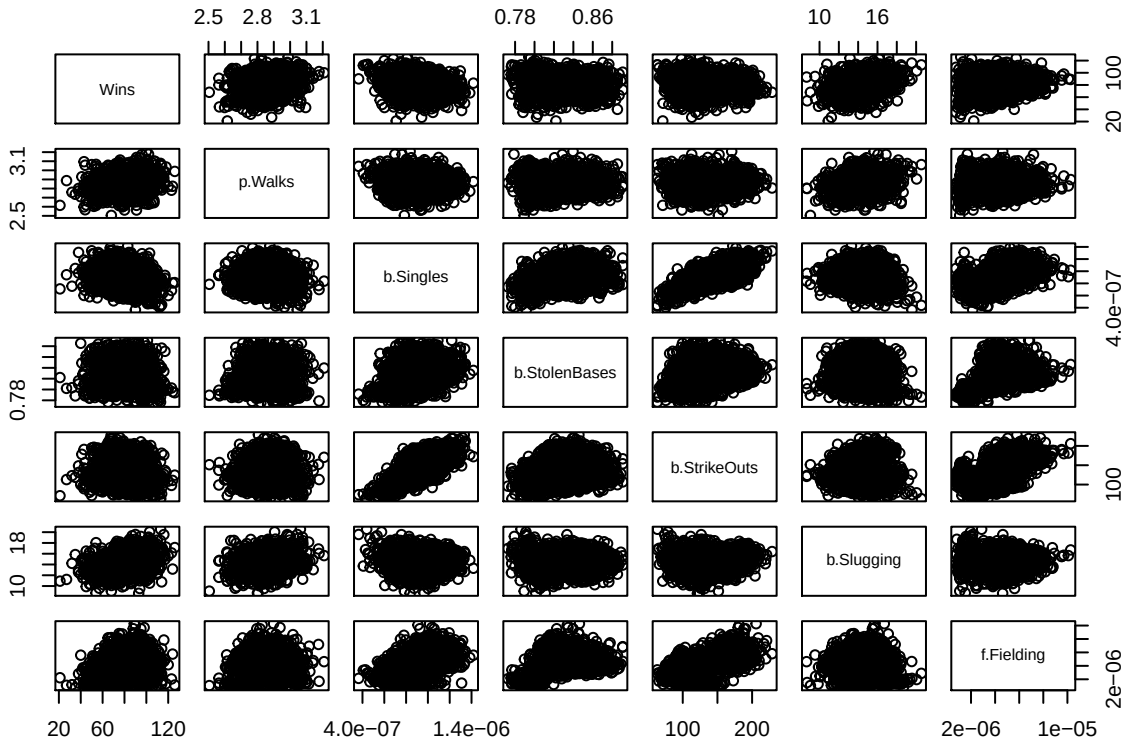
Add b.Fielding

$$\widehat{wins} = \hat{\beta}_0 + \hat{\beta}_1 \times p.Walks\hat{\beta}_2 \times b.Singles + \hat{\beta}_3 \times b.Doubles + \hat{\beta}_4 \times b.StolenBases +$$

$$\hat{\beta}_5 \times f.DoublePlays + \hat{\beta}_6 \times b.StrikeOuts + \hat{\beta}_7 \times b.Slugging + \hat{\beta}_8 \times b.FieldingYield +$$

nrow(lm.smk)

```
#VARIABLES
#variables have been transformed first as individual predictors
Wins  <- mW
Index <- lm.smk$INDEX
p.Walks <- tmp.BB
b.Singles <- tmb.1B
b.Doubles <- tmb.2B
b.StolenBases <- tmb.SB
b.StrikeOuts <- tmb.SO
b.Slugging <- tmb.SL
f.Fielding <- tmf.FY

m1 <- lm(Wins ~ -Index+p.Walks+b.Singles+b.Doubles+b.StolenBases+b.StrikeOuts+b.Slugging+f.Fielding)

#PAIRWISE PLOT
par(mfrow=c(1,1))
pairs(Wins ~ p.Walks+b.Singles+b.StolenBases+b.StrikeOuts+b.Slugging+f.Fielding)
```



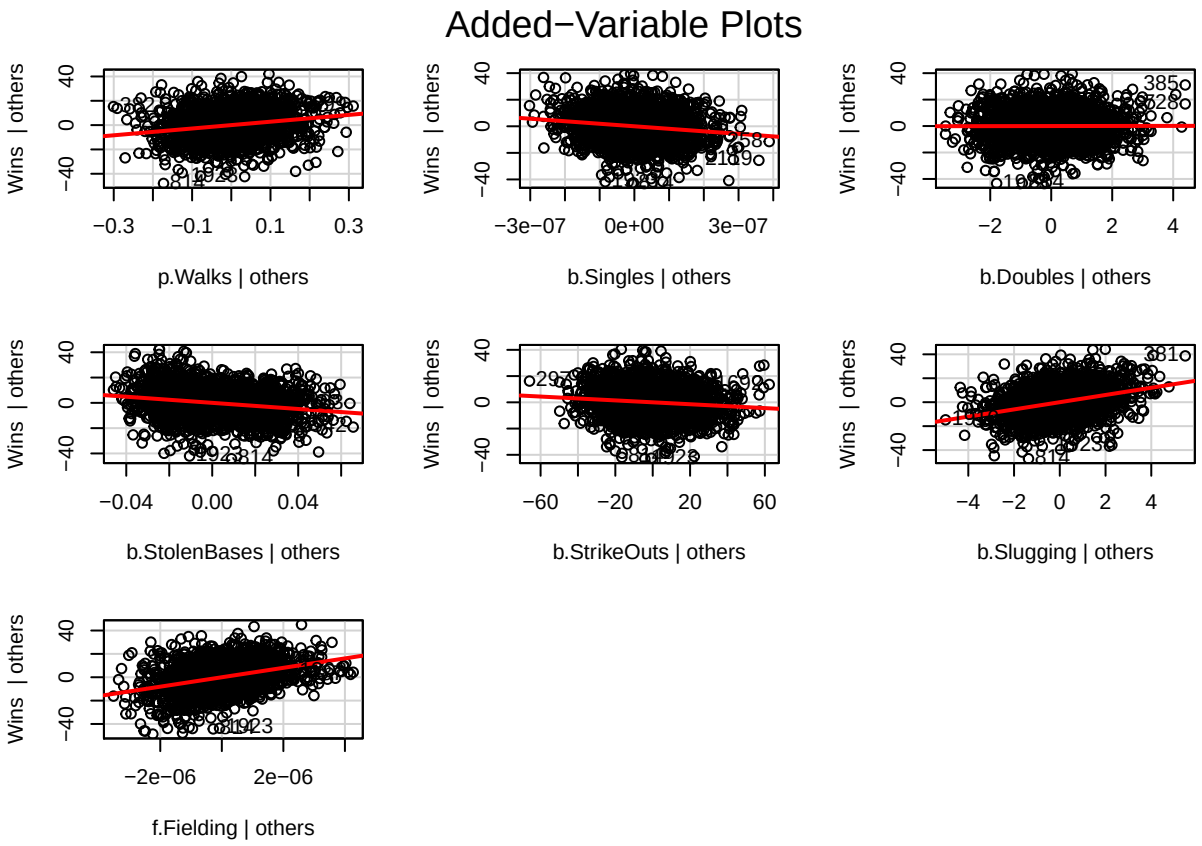

```
#MODEL DIAGNOSTICS
summary(m1)
```

```
##
## Call:
## lm(formula = Wins ~ -Index + p.Walks + b.Singles + b.Doubles +
##     b.StolenBases + b.StrikeOuts + b.Slugging + f.Fielding)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -43.318  -7.845  -0.053   7.816  39.368
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)    6.703e+01  1.267e+01   5.289 1.36e-07 ***
## p.Walks        2.793e+01  3.000e+00   9.311  < 2e-16 ***
## b.Singles     -1.900e+07  2.670e+06  -7.117 1.50e-12 ***
## b.Doubles      3.065e-02  2.130e-01   0.144    0.886
## b.StolenBases -1.211e+02  1.299e+01  -9.318  < 2e-16 ***
## b.StrikeOuts  -7.317e-02  1.409e-02  -5.195 2.24e-07 ***
## b.Slugging     3.015e+00  1.829e-01  16.485  < 2e-16 ***
## f.Fielding     4.033e+06  2.259e+05  17.853  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.95 on 2149 degrees of freedom
## Multiple R-squared:  0.3168, Adjusted R-squared:  0.3146
## F-statistic: 142.4 on 7 and 2149 DF,  p-value: < 2.2e-16
```

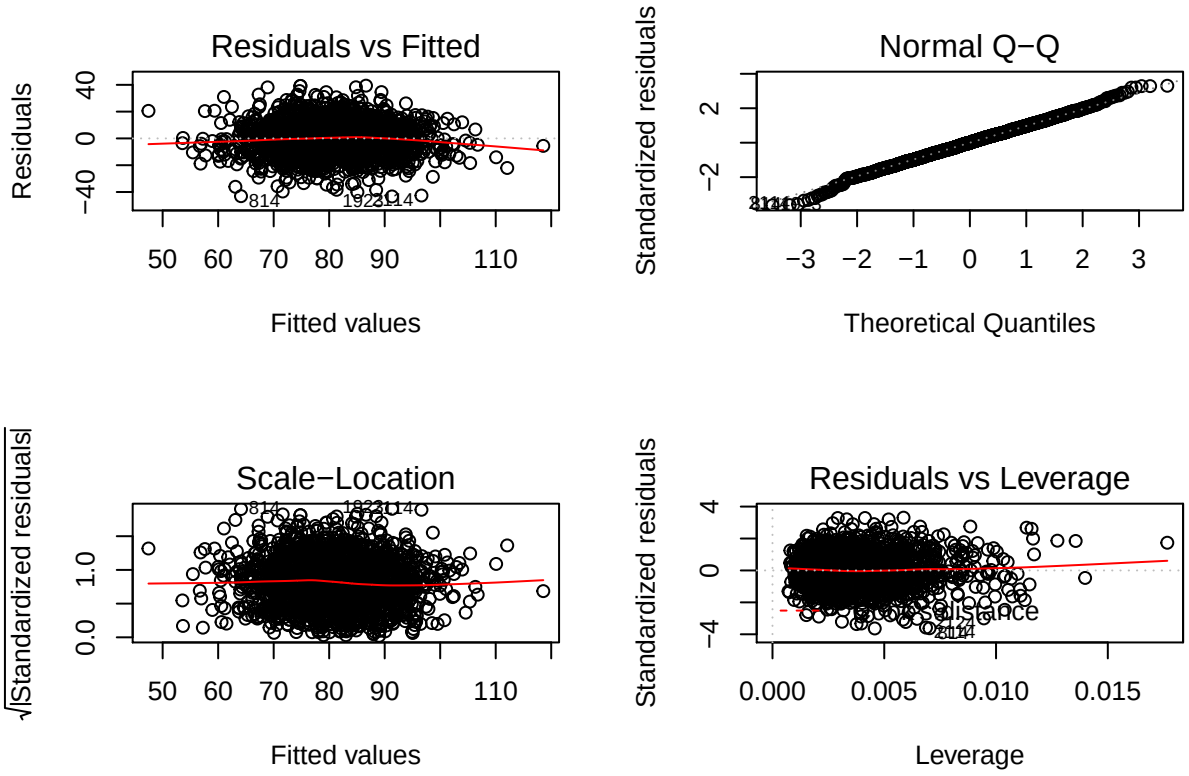
```
#DIAGNOSTIC1. show collinearity of variables after checking p-values to < 0.05.
vif(m1)
```

```
##       p.Walks     b.Singles     b.Doubles b.StolenBases  b.StrikeOuts
##      1.106592      2.857435      1.469729      1.205956      3.210354
##    b.Slugging    f.Fielding
##      1.301083      1.801535
```

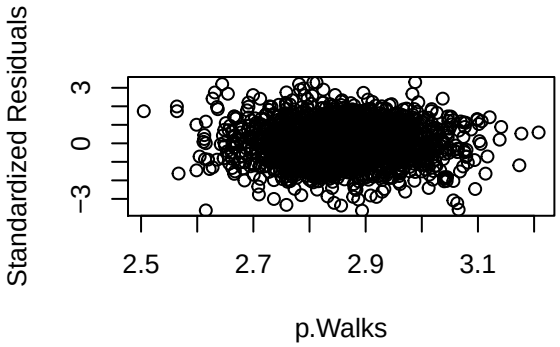
```
#p-values are all < 0.05 and no VIFs  > 5

#DIAGNOSTIC2.  generate Added Variable Plots: should show linear relationship between response & predic
par(mfrow=c(2,2))
avPlots(m1, ~.,ask=FALSE, id.n = 2)
```

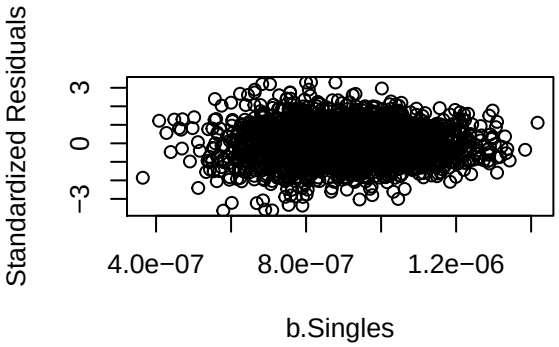


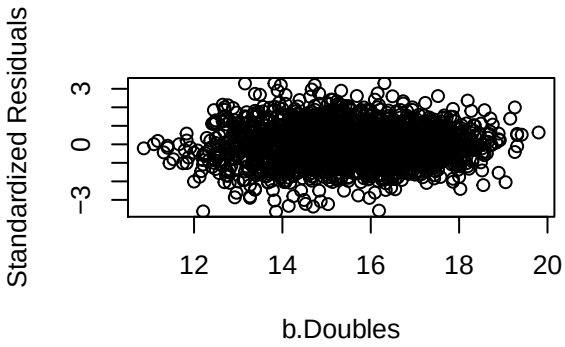

```
#relationship is linear

#DIAGNOSTIC3.  generate Summary Diagnostic Plots
par(mfrow=c(2,2))
plot(m1)
```

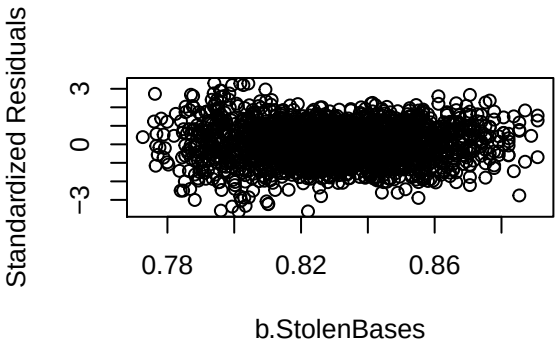



```
#Upper Left plot "Residuals vs Fitted"
# clear predictable pattern
# uniform variability for all fitted values
#Upper Right
#  normality in residuals
#Lower Right plot "Residuals vs. Leverage"
#  normal distribution, and uniform distribution of residuals
#  no significant leverage points

#DIAGNOSTIC4.  generate Standardized Residual Plots against each predictor
par(mfrow=c(2,2))
StanRest <- rstandard(m1)
plot(p.Walks,StanRest,ylab="Standardized Residuals")
plot(b.Singles,StanRest,ylab="Standardized Residuals")
plot(b.Doubles,StanRest,ylab="Standardized Residuals")
plot(b.StolenBases,StanRest,ylab="Standardized Residuals")
```

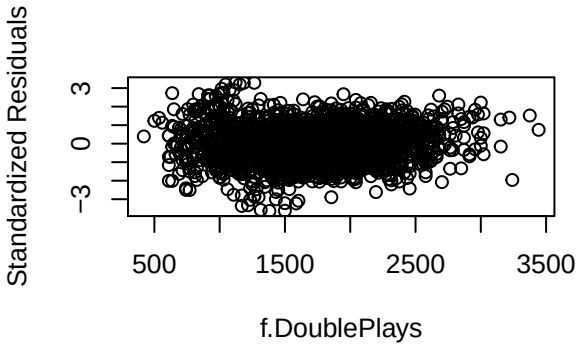


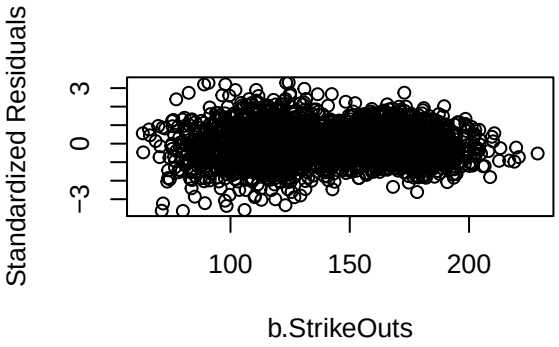


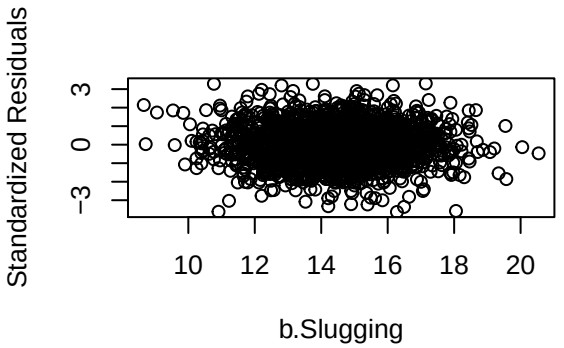


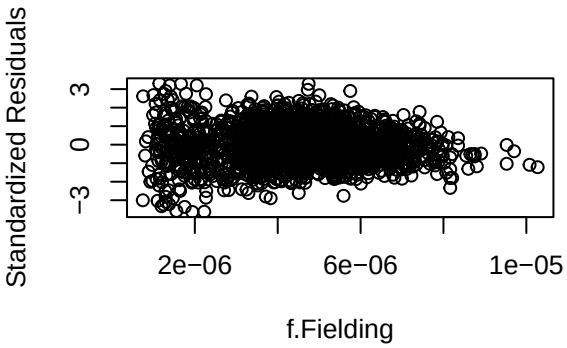



```
plot(f.DoublePlays,StanRest,ylab="Standardized Residuals")
plot(b.StrikeOuts,StanRest,ylab="Standardized Residuals")
plot(b.Slugging,StanRest,ylab="Standardized Residuals")
plot(f.Fielding,StanRest,ylab="Standardized Residuals")
```

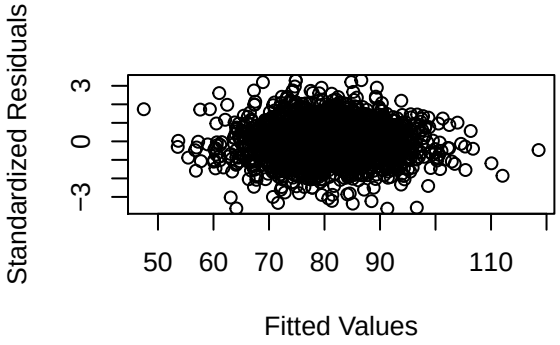




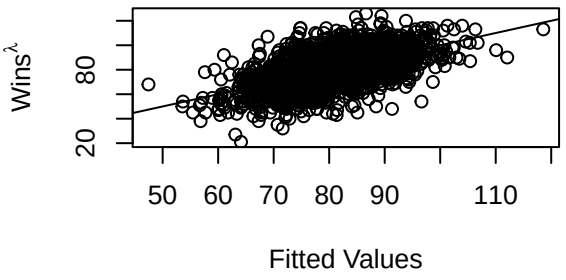


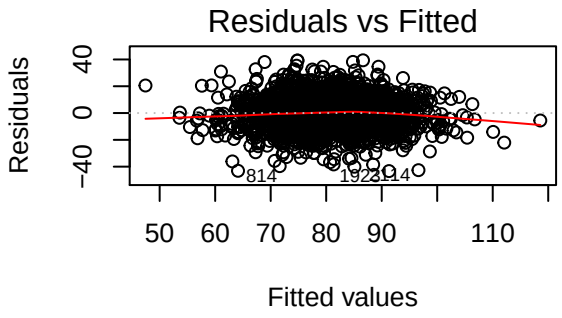


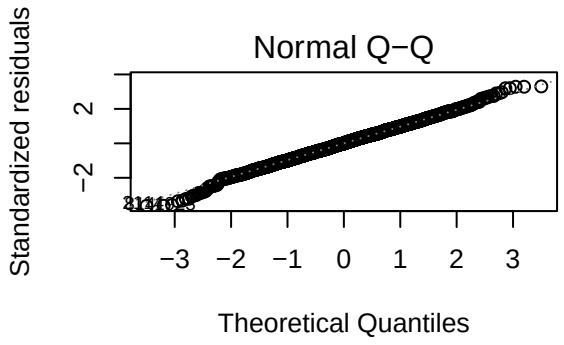
```r
plot(m1$fitted.values,StanRest,ylab="Standardized Residuals",xlab="Fitted Values")
#Examine plots for constant variability of residuals across ALL predictor.
# uniform distribution of residuals

#DIAGNOSTIC5.  generate plot of Y "response variable"" against Fitted Values "regression model"
par(mfrow = c(2,2))
```

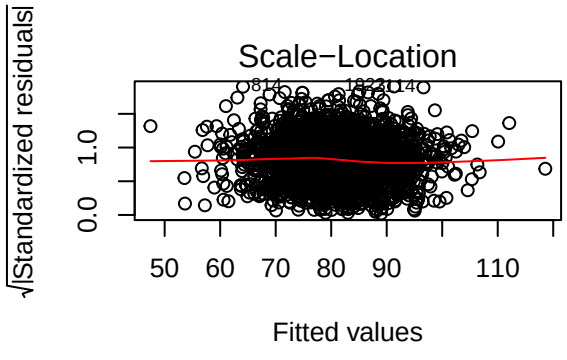



```r
plot(m1$fitted.values,Wins,xlab="Fitted Values",ylab=expression(Wins^lambda))
abline(lsfit(m1$fitted.values,Wins))
plot(m1)
```

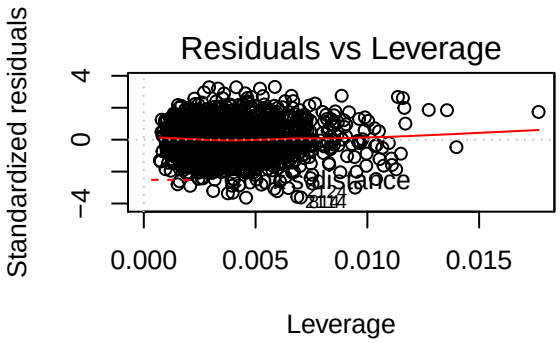








```
# normal distribution, and uniform distribution of residuals
```




```
#pred_eval.m1 <- round(predict(m1, eval_data))
#eval.BSO.imp <- impute(eval_data$TARGET_WINS, pred_eval.m1)
```