621 HW 1 v 4

Jeff Nieman, Scott Karr, James Topor, Armenoush

June 13, 2016

CONSOLIDATE VARIABLES, REMOVE NA'S & OUTLIERS

Create new column for batting singles and eliminating hits for batting

Eliminate HBP, CS, and pitching HR's.

Build model for batting SO using Gelman approach

Build model for Pitching SO

Build model for SB

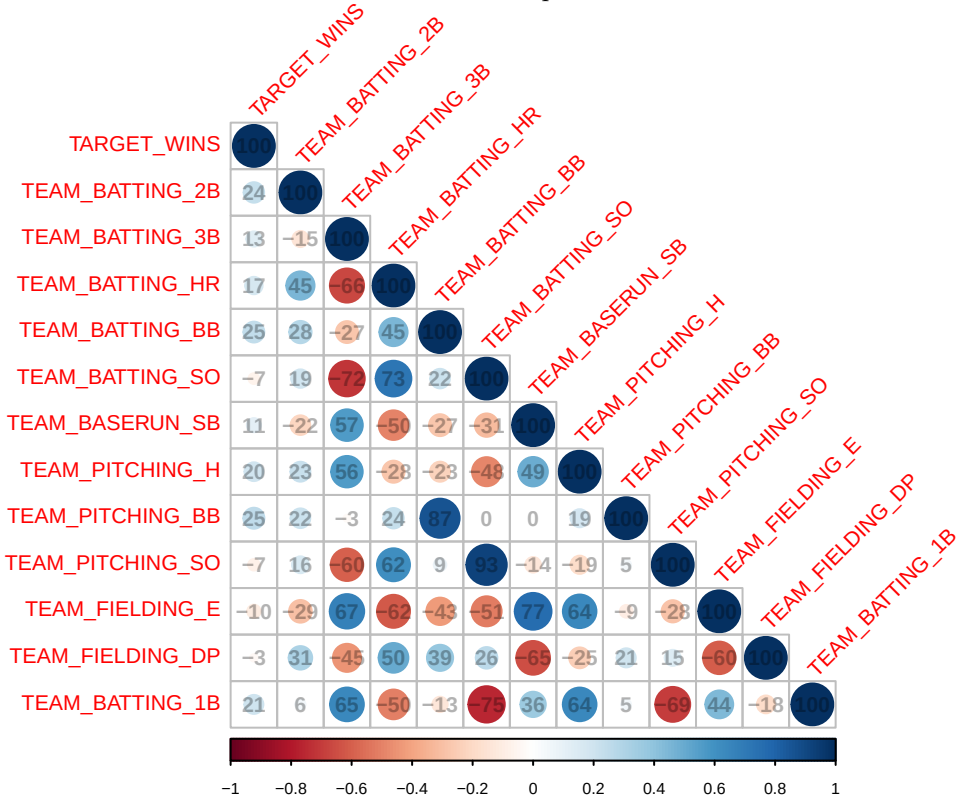
Build model to replace DP

Eliminate unhistorical outliers

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.



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 one-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

$$\widehat{wins} = \hat{\beta}_0 + \hat{\beta}_1 \times p.Hits + \hat{\beta}_2 \times p.Walks$$

Add b.Singles & b.Doubles

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

Removed p.Hits

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

Added Stolen Bases and Double Plays

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

$$\hat{\beta}_5 \times f.DoublePlays +$$

Added b.Walks and p.Strikeouts

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

$$\hat{\beta}_5 \times f.DoublePlays + \hat{\beta}_6 \times p.Walks + \hat{\beta}_7 \times p.StrikeOuts +$$

Remove b.Walks

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

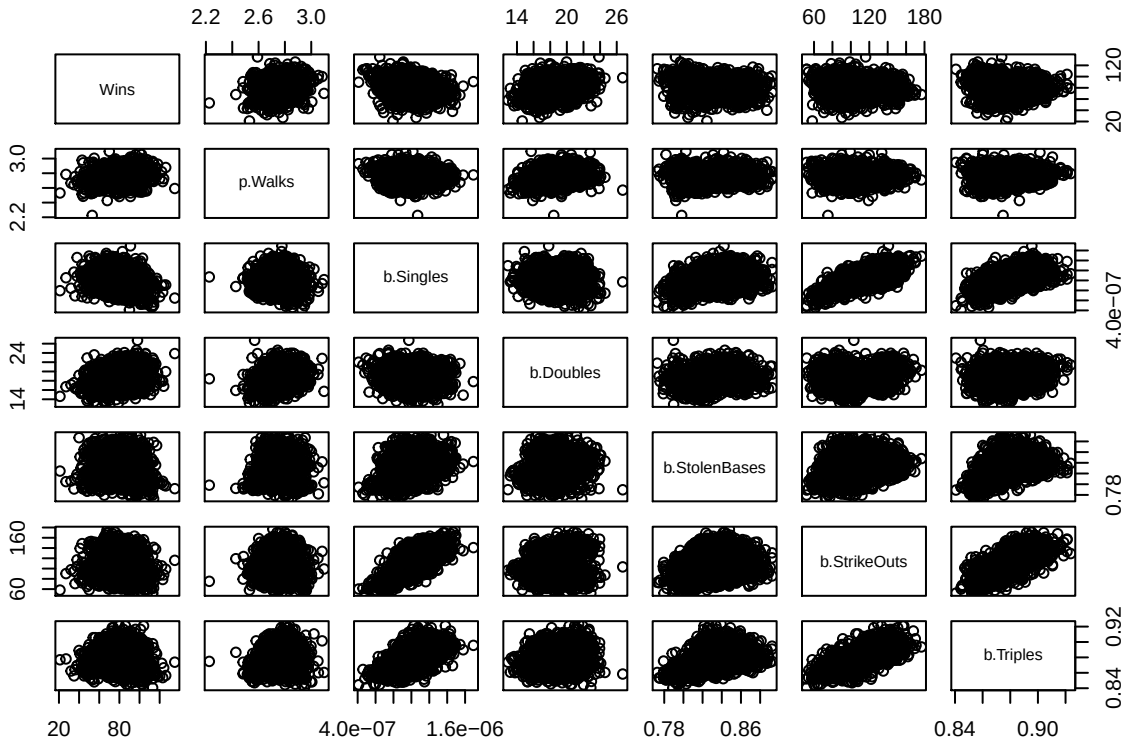
$$\hat{\beta}_5 \times f.DoublePlays + \hat{\beta}_6 \times p.StrikeOuts +$$

Add b.StrikeOuts and b.Triples

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

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

```
#VARIABLES
#variables have been transformed first as individual predictors
Wins  <- mW
p.Walks <- tmp.BB
b.Singles <- tmb.1B
b.Doubles <- tmb.2B
b.StolenBases <- tmb.SB
f.DoublePlays <- tmf.DP
b.StrikeOuts <- tmb.SO
b.Triples <- tmb.3B
m1 <- lm(Wins ~ p.Walks+b.Singles+b.Doubles+b.StolenBases+f.DoublePlays+b.StrikeOuts+b.Triples)

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



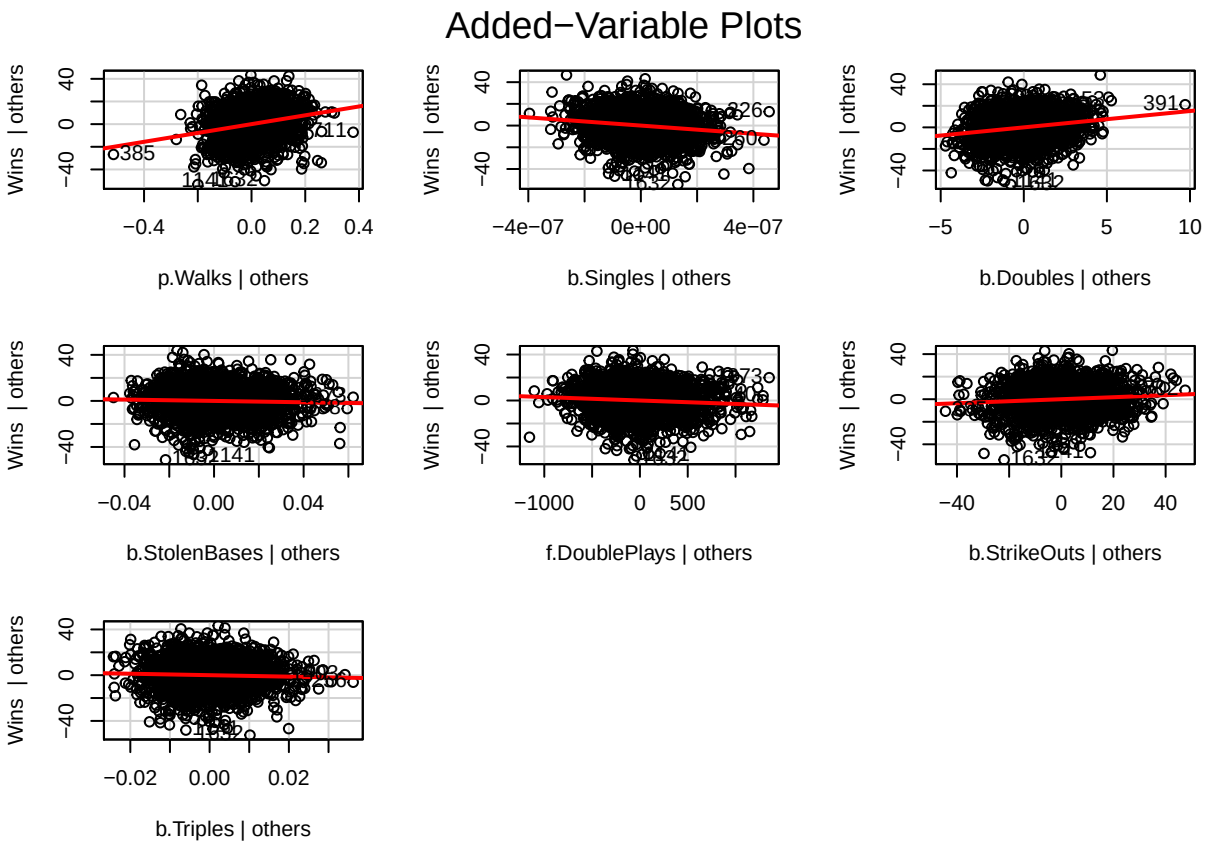

```
#MODEL DIAGNOSTICS
summary(m1)
```

```
##
## Call:
## lm(formula = Wins ~ p.Walks + b.Singles + b.Doubles + b.StolenBases +
##     f.DoublePlays + b.StrikeOuts + b.Triples)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -51.784  -8.593   0.429   9.090  43.438
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)    4.011e+01  2.898e+01   1.384 0.166518
## p.Walks        3.900e+01  3.526e+00  11.058  < 2e-16 ***
## b.Singles     -1.887e+07  2.673e+06  -7.061 2.23e-12 ***
## b.Doubles      1.514e+00  1.681e-01   9.008  < 2e-16 ***
## b.StolenBases -2.948e+01  1.757e+01  -1.678 0.093517 .
## f.DoublePlays -2.972e-03  8.008e-04  -3.712 0.000211 ***
## b.StrikeOuts   8.886e-02  2.263e-02   3.927 8.89e-05 ***
## b.Triples     -6.380e+01  3.339e+01  -1.911 0.056195 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.37 on 2164 degrees of freedom
## Multiple R-squared:  0.1574, Adjusted R-squared:  0.1547
## F-statistic: 57.77 on 7 and 2164 DF,  p-value: < 2.2e-16
```

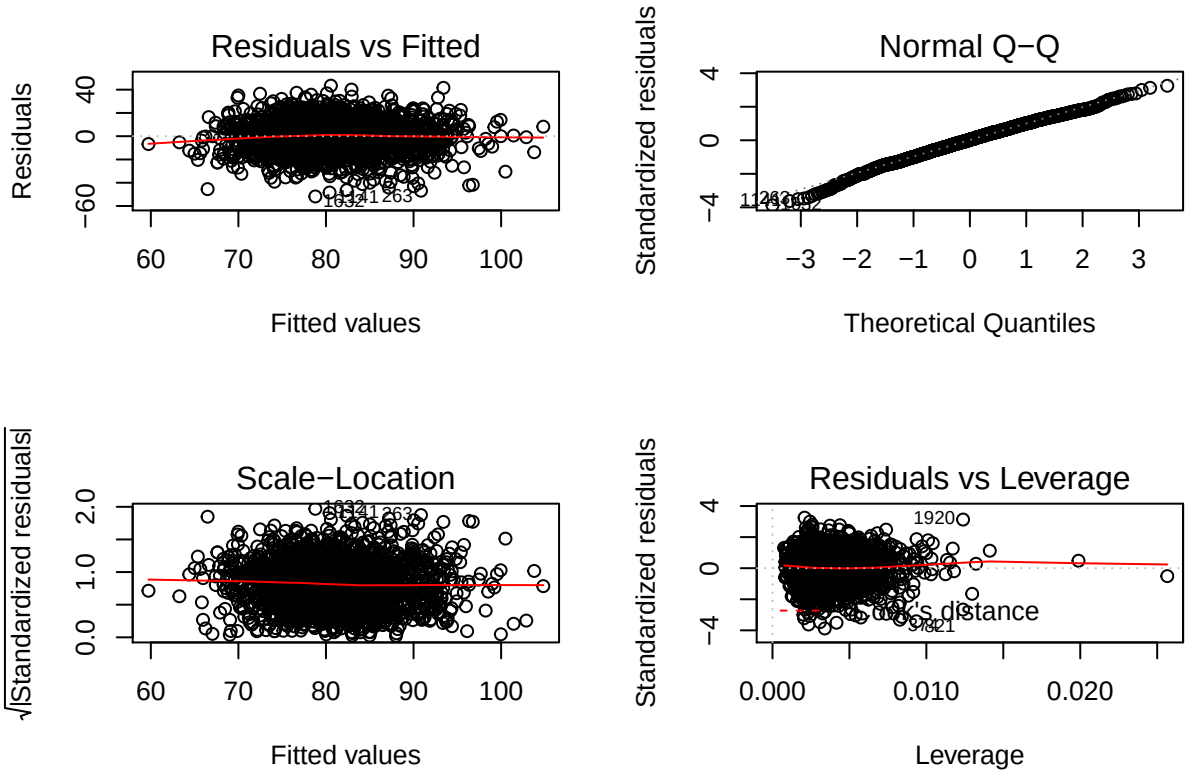
```
#diagnostic 1. show collinearity of variables after checking p-values to < 0.05.
vif(m1)
```

```
##       p.Walks     b.Singles     b.Doubles b.StolenBases f.DoublePlays
##      1.101853      2.858302      1.270195      1.745679      1.771925
##  b.StrikeOuts     b.Triples
##      3.785674      2.695211
```

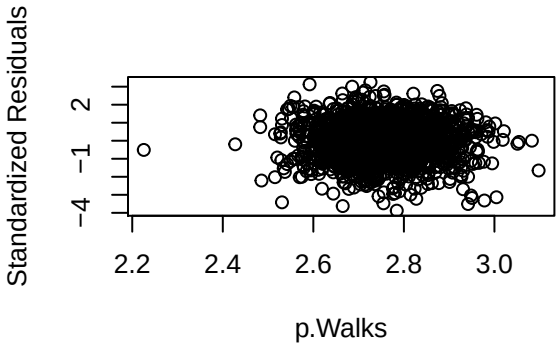
```
#If resulting coefficients are > 5, remove the variable w/ largest VIF and re-run the model.
#Subsequently remove variables either due to p-values becoming > 0.05 or VIF coefficients > 5.
#Continue until all remaining p-values are < 0.05 and no VIFs  > 5

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

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

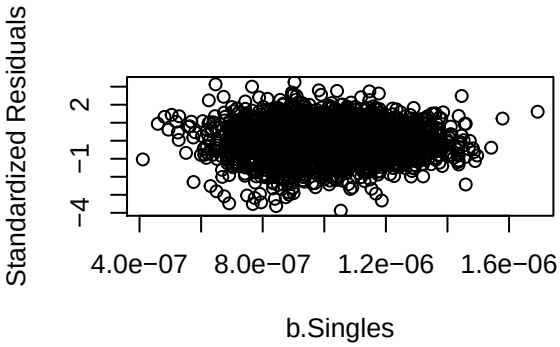


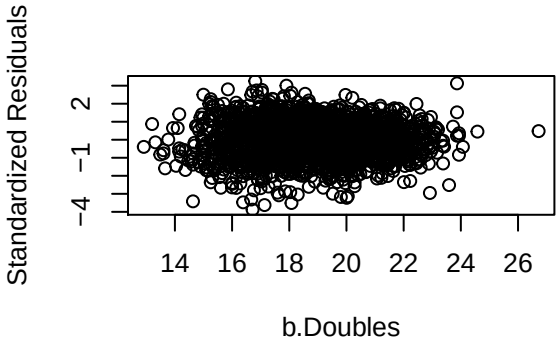

```
#relationship is linear

#diagnostic 3.  generate Summary Diagnostic Plots
par(mfrow=c(2,2))
plot(m1)
```

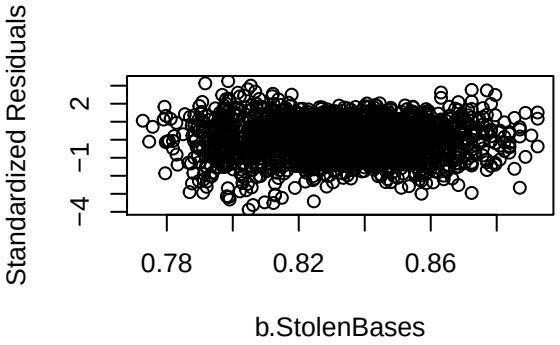



```
#Upper Left plot "Residuals vs Fitted"
#       Is there a clear predictable pattern in that plot?  If no model isn't valid
#       Do residuals have uniform variability for all fitted values?  If no model isn't valid.
#Upper Right
#       Is there normality in your residuals? If not, the model lack normality and the model isn't vali
#Lower Right plot "Residuals vs. Leverage"
#       The Y-axis in this plot represents standard deviations from the mean of the #residuals.
#       If most residuals are within 2 standard deviations of mean, the model is likely valid.
#       The plot also shows #"Cooks Distance" -- a metric for identifying high leverage outliers.
#       The worst outliers are automatically labeled by rownum in the plot. If the outliers are
#       far outside the 2 standard deviations then remove those rows from your data set, renumber
#       the #rows of the resulting dataframe so there are no gaps in the integer row numbers, and
#       then re-run your model to see if it improves.

# normal distribution, and uniform distribution of residuals
# no significant leverage points

#diagnostic 4.  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")
```

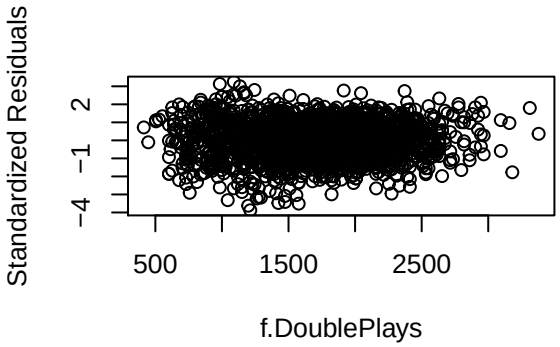


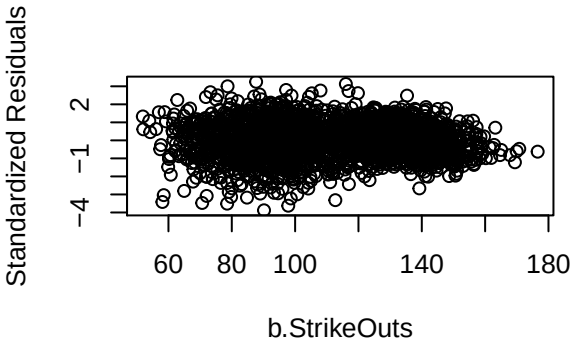


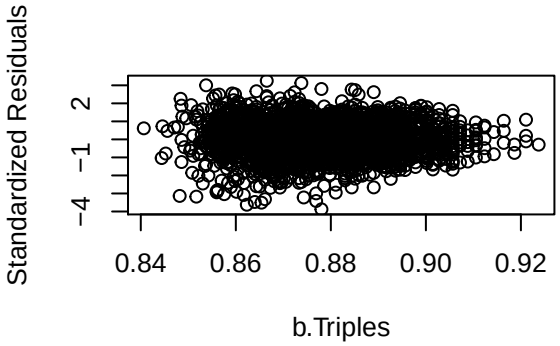


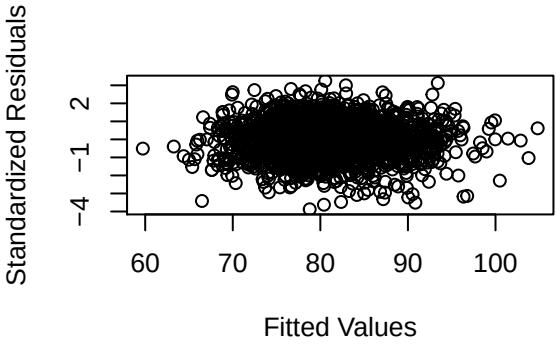



```
plot(f.DoublePlays,StanRest,ylab="Standardized Residuals")
plot(b.StrikeOuts,StanRest,ylab="Standardized Residuals")
plot(b.Triples,StanRest,ylab="Standardized Residuals")
plot(m1$fitted.values,StanRest,ylab="Standardized Residuals",xlab="Fitted Values")
```

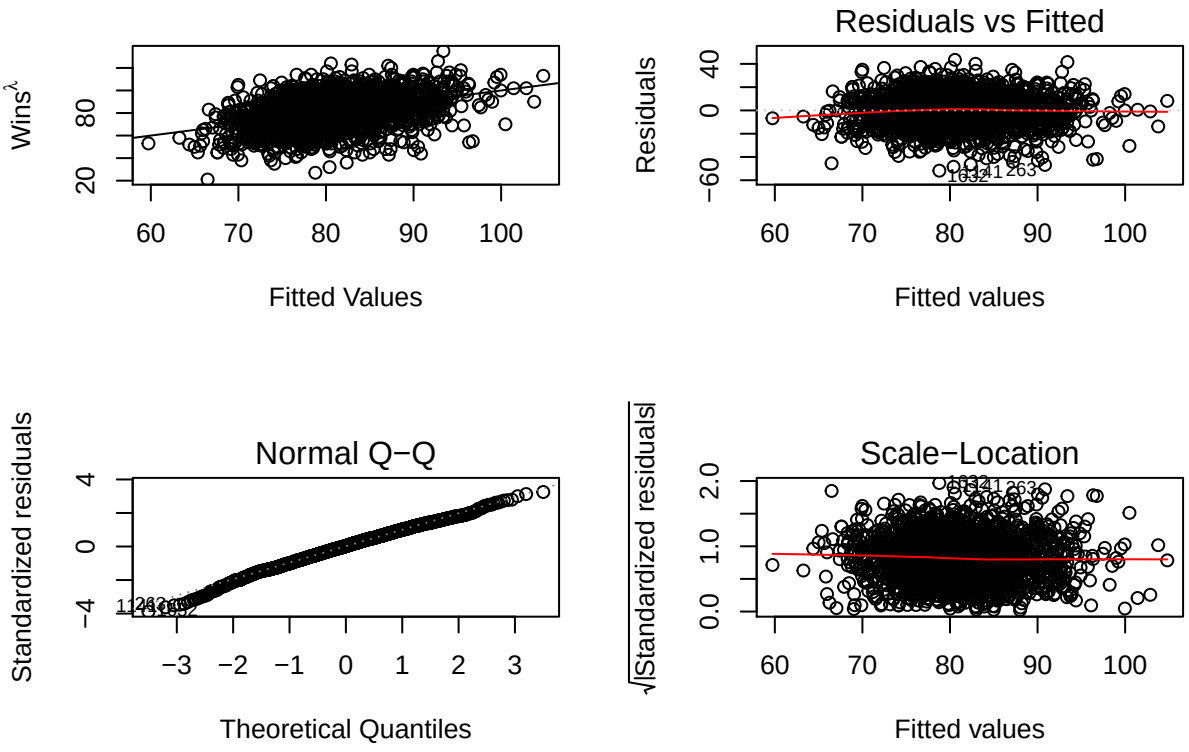








```r
#Examine plots for constant variability of residuals across ALL of the values of your predictor.
#       Constant variability means residuals are disbursed uniformly across all predictor variable's va
#       You should not see any predictable pattern or model lacks constant variability and isn't valid.

# uniform distribution of residuals

#diagnostic 5.  generate plot of Y "response variable"" against Fitted Values "regression model"
par(mfrow = c(2,2))
plot(m1$fitted.values,Wins,xlab="Fitted Values",ylab=expression(Wins^lambda))
abline(lsfit(m1$fitted.values,Wins))
plot(m1)
```




```r
#If plot doesn't shows a linear relationship with no pattern or skew the model lacks normality.

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