NBA Wins Analysis

Team № 09

Dennis Felson
Krishna Agarwal
Harman Kaur
Vishnu Regimon Nair
Anna Iuferova
Nikitha Patel

Course 603
Statistical Modelling with Data

Calgary 2021

UNIVERSITY OF CALGARY

Introduction

The purpose of the project :

investigate the factors that affect total wins for teams in the National Basketball Association (NBA).

Data Source

The Dataset is cross-sectional from the 82 games , containing observations from 30 teams during the 2018-2019 season

A linear regression model contains 14 independent variables:

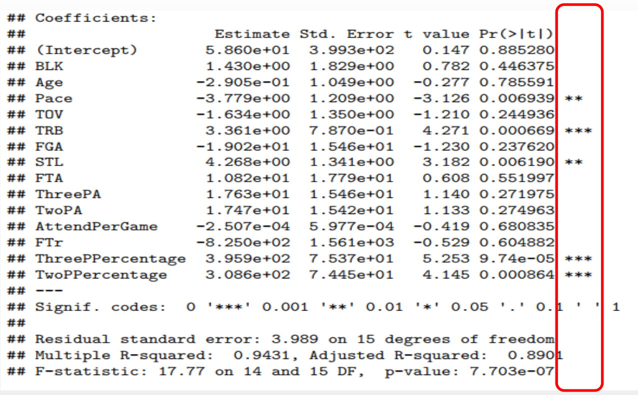
Dependant variable - Total Number of Wins (W - Discrete variable)

1. Average number of Blocks per Game (BLK - Continuous variable)
2. Average Age of the Roster (Age - Continuous variable)
3. Average number of Possessions per game (Pace - Continuous variable)
4. Average number of Turnovers per game (TOV - Continuous variable)
5. Average number of Total Rebounds per game (TRB - Continuous variable)
6. Average number of Field Goal Attempts per game (FGA - Continuous variable)
7. Average number of Steals per game (STL - Continuous variable)
8. Average number of Free Throw Attempts per game (FTA - Continuous variable)
9. Average number of Three Point Attempts per game (ThreePA - Continuous variable)
10. Average number of Two Point Attempts per game (TwoPA - Continuous variable)
11. Average Attendance per game (AttendPerGame - Continuous variable)
12. Average Free Throw Rate per game (FTr - Continuous variable)
13. Average Three Point Percentage per game (ThreePPercentage - Continuous variable)
14. Average Two Point Percentage per game (TwoPPercentage - Continuous variable)

Modeling

- **Data Wrangling**

- **Build a full model with all the independent variables**

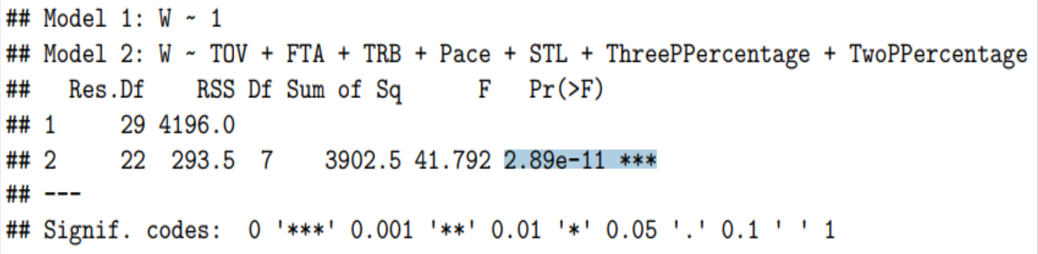


```
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       5.860e+01  3.993e+02   0.147 0.885280
## BLK               1.430e+00  1.829e+00   0.782 0.446375
## Age              -2.905e-01  1.049e+00  -0.277 0.785591
## Pace             -3.779e+00  1.209e+00  -3.126 0.006939 **
## TOV              -1.634e+00  1.350e+00  -1.210 0.244936
## TRB               3.361e+00  7.870e-01   4.271 0.000669 ***
## FGA              -1.902e+01  1.546e+01  -1.230 0.237620
## STL               4.268e+00  1.341e+00   3.182 0.006190 **
## FTA               1.082e+01  1.779e+01   0.608 0.551997
## ThreePA           1.763e+01  1.546e+01   1.140 0.271975
## TwoPA             1.747e+01  1.542e+01   1.133 0.274963
## AttendPerGame    -2.507e-04  5.977e-04  -0.419 0.680835
## FTr              -8.250e+02  1.561e+03  -0.529 0.604882
## ThreePPercentage  3.959e+02  7.537e+01   5.253 9.74e-05 ***
## TwoPPercentage    3.086e+02  7.445e+01   4.145 0.000864 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.989 on 15 degrees of freedom
## Multiple R-squared:  0.9431, Adjusted R-squared:  0.8901
## F-statistic: 17.77 on 14 and 15 DF,  p-value: 7.703e-07
```

- **Multi-Collinearity test:**
  Age, BLK, TOV, RB, Pace, STL, FTA, AttendPerGame, ThreePPercentage, TwoPPercentage

Do not show multicollinearity,
Will use STEPWISE/ FORWARD/ BACKWARD regressions

- **Global F-test**
  Level of significance $\alpha = 0.05$
  $H0 : \beta1 = \beta2 = \beta3 = \beta4 = \beta5 = \beta6 = \beta7 = 0$
  $Ha : at\ least\ one\ \beta i \neq 0$



```
## Model 1: W ~ 1
## Model 2: W ~ TOV + FTA + TRB + Pace + STL + ThreePPercentage + TwoPPercentage
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     29 4196.0
## 2     22  293.5  7    3902.5 41.792 2.89e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

P-value<0.05, so We reject the Null Hypothesis.

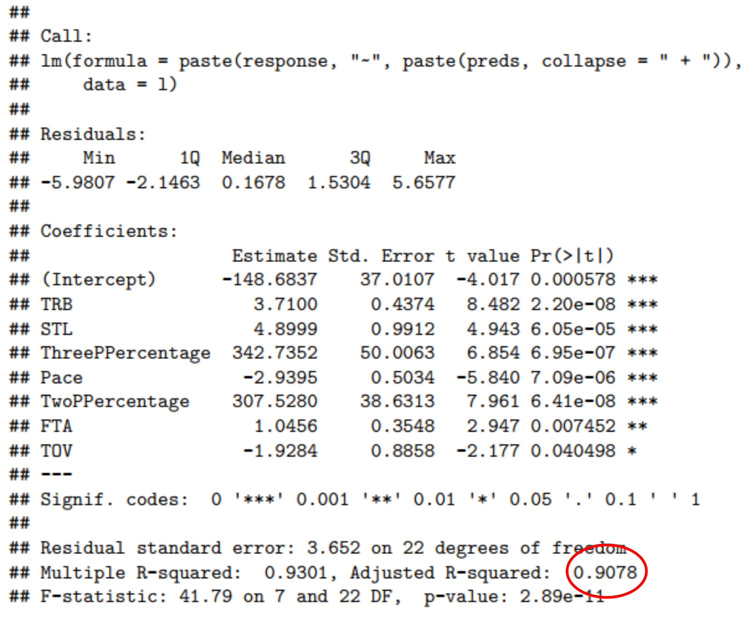
Number of wins depends on at least one of the independent variable

Modeling

- **STEPWISE, FORWARD, BACKWARD REGRESSION MODELS**

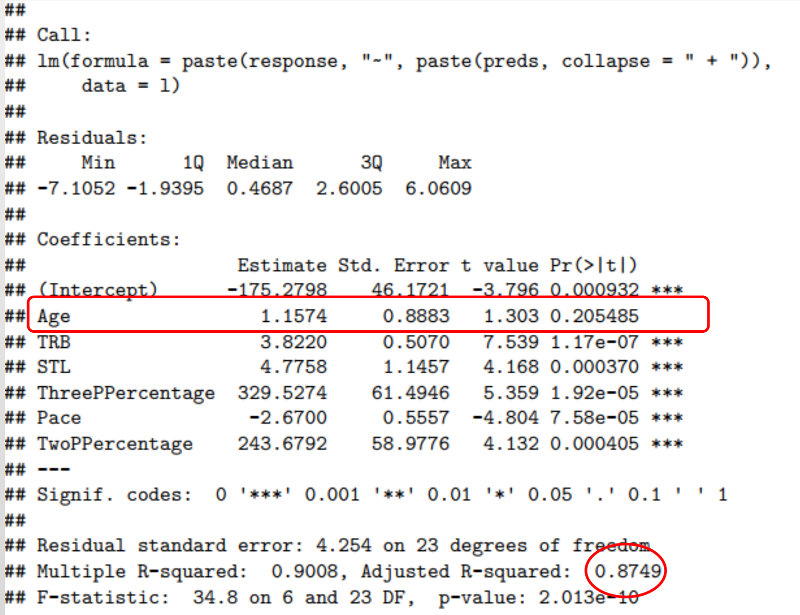
  (pent = 0.05, prem = 0.1)

STEPWISE



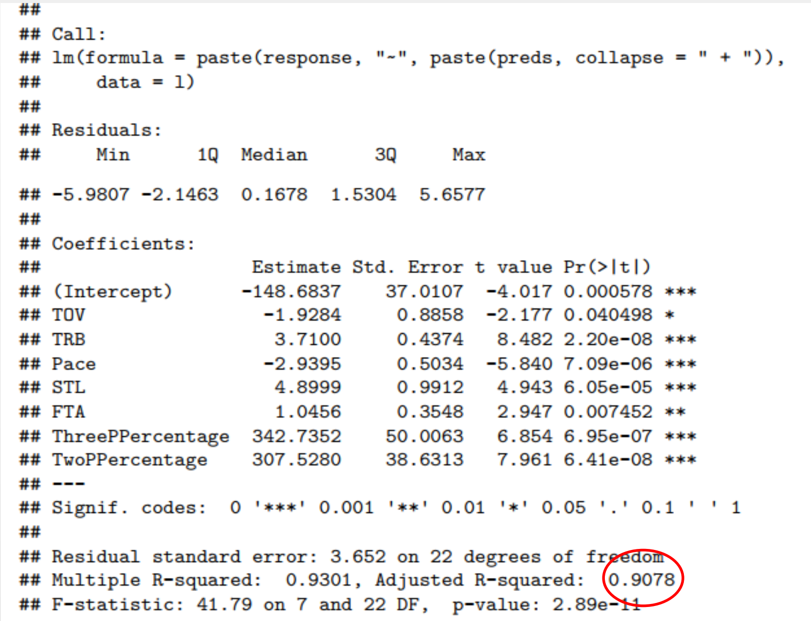
```
##
## Call:
## lm(formula = paste(response, "~", paste(preds, collapse = " + ")),
##     data = l)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.9807 -2.1463  0.1678  1.5304  5.6577
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -148.6837    37.0107  -4.017 0.000578 ***
## TRB                 3.7100     0.4374   8.482 2.20e-08 ***
## STL                 4.8999     0.9912   4.943 6.05e-05 ***
## ThreePPercentage  342.7352    50.0063   6.854 6.95e-07 ***
## Pace               -2.9395     0.5034  -5.840 7.09e-06 ***
## TwoPPercentage    307.5280    38.6313   7.961 6.41e-08 ***
## FTA                 1.0456     0.3548   2.947 0.007452 **
## TOV                -1.9284     0.8858  -2.177 0.040498 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.652 on 22 degrees of freedom
## Multiple R-squared:  0.9301, Adjusted R-squared:  0.9078
## F-statistic: 41.79 on 7 and 22 DF,  p-value: 2.89e-11
```

FORWARD



```
##
## Call:
## lm(formula = paste(response, "~", paste(preds, collapse = " + ")),
##     data = l)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.1052 -1.9395  0.4687  2.6005  6.0609
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -175.2798    46.1721  -3.796 0.000932 ***
## Age                 1.1574     0.8883   1.303 0.205485
## TRB                 3.8220     0.5070   7.539 1.17e-07 ***
## STL                 4.7758     1.1457   4.168 0.000370 ***
## ThreePPercentage  329.5274    61.4946   5.359 1.92e-05 ***
## Pace               -2.6700     0.5557  -4.804 7.58e-05 ***
## TwoPPercentage    243.6792    58.9776   4.132 0.000405 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.254 on 23 degrees of freedom
## Multiple R-squared:  0.9008, Adjusted R-squared:  0.8749
## F-statistic:  34.8 on 6 and 23 DF,  p-value: 2.013e-10
```

BACKWARD



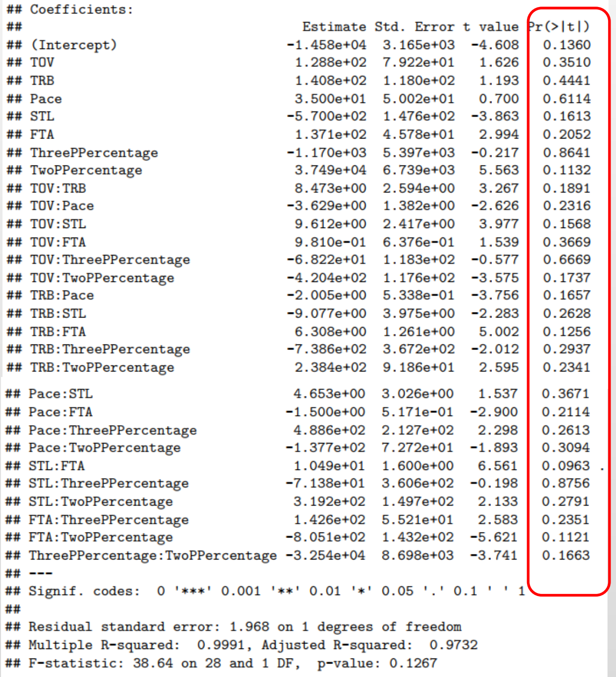
```
##
## Call:
## lm(formula = paste(response, "~", paste(preds, collapse = " + ")),
##     data = l)
##
## Residuals:
##     Min      1Q  Median      3Q     Max

## -5.9807 -2.1463  0.1678  1.5304  5.6577
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -148.6837    37.0107  -4.017 0.000578 ***
## TOV                -1.9284     0.8858  -2.177 0.040498 *
## TRB                 3.7100     0.4374   8.482 2.20e-08 ***
## Pace               -2.9395     0.5034  -5.840 7.09e-06 ***
## STL                 4.8999     0.9912   4.943 6.05e-05 ***
## FTA                 1.0456     0.3548   2.947 0.007452 **
## ThreePPercentage  342.7352    50.0063   6.854 6.95e-07 ***
## TwoPPercentage    307.5280    38.6313   7.961 6.41e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.652 on 22 degrees of freedom
## Multiple R-squared:  0.9301, Adjusted R-squared:  0.9078
## F-statistic: 41.79 on 7 and 22 DF,  p-value: 2.89e-11
```

$$\hat{Y}_{wins} = \hat{\beta}_0 + \hat{\beta}_1 TOV + \hat{\beta}_2 PACE + \hat{\beta}_3 TRB + \hat{\beta}_4 STL + \hat{\beta}_5 FTA + \hat{\beta}_6 ThreePPercentage + \hat{\beta}_7 TwoPPercentage$$

first order model

Modeling

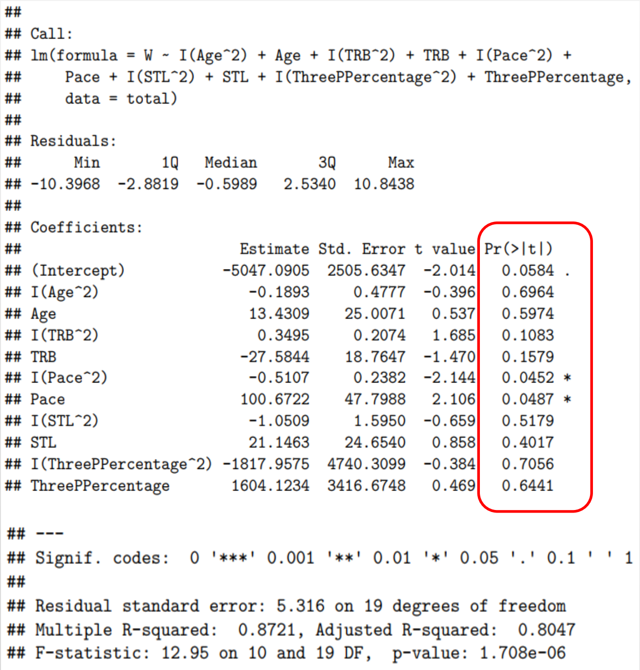
- **Interaction model**



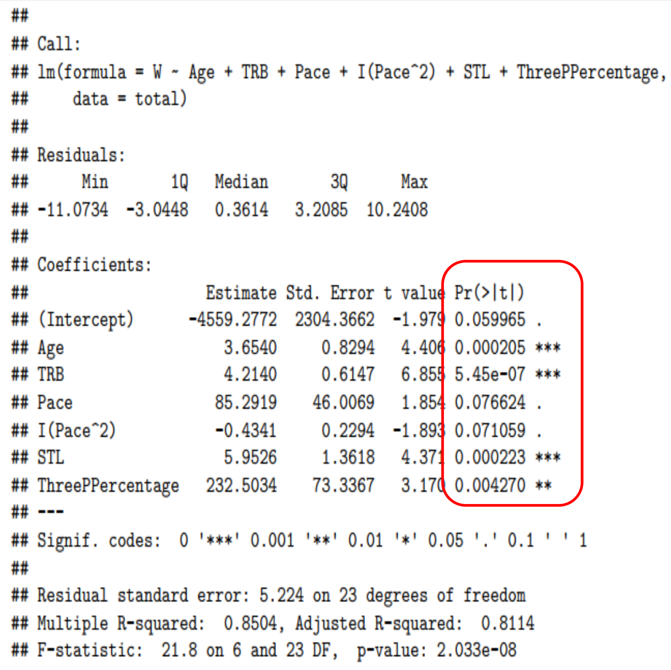
```
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     -1.458e+04  3.165e+03  -4.608   0.1360
## TOV                              1.288e+02  7.922e+01   1.626   0.3510
## TRB                              1.408e+02  1.180e+02   1.193   0.4441
## Pace                             3.500e+01  5.002e+01   0.700   0.6114
## STL                             -5.700e+02  1.476e+02  -3.863   0.1613
## FTA                              1.371e+02  4.578e+01   2.994   0.2052
## ThreePPercentage                -1.170e+03  5.397e+03  -0.217   0.8641
## TwoPPercentage                   3.749e+04  6.739e+03   5.563   0.1132
## TOV:TRB                          8.473e+00  2.594e+00   3.267   0.1891
## TOV:Pace                        -3.629e+00  1.382e+00  -2.626   0.2316
## TOV:STL                          9.612e+00  2.417e+00   3.977   0.1568
## TOV:FTA                          9.810e-01  6.376e-01   1.539   0.3669
## TOV:ThreePPercentage            -6.822e+01  1.183e+02  -0.577   0.6669
## TOV:TwoPPercentage              -4.204e+02  1.176e+02  -3.575   0.1737
## TRB:Pace                        -2.005e+00  5.338e-01  -3.756   0.1657
## TRB:STL                         -9.077e+00  3.975e+00  -2.283   0.2628
## TRB:FTA                          6.308e+00  1.261e+00   5.002   0.1256
## TRB:ThreePPercentage            -7.386e+02  3.672e+02  -2.012   0.2937
## TRB:TwoPPercentage               2.384e+02  9.186e+01   2.595   0.2341
## Pace:STL                         4.653e+00  3.026e+00   1.537   0.3671
## Pace:FTA                        -1.500e+00  5.171e-01  -2.900   0.2114
## Pace:ThreePPercentage            4.886e+02  2.127e+02   2.298   0.2613
## Pace:TwoPPercentage             -1.377e+02  7.272e+01  -1.893   0.3094
## STL:FTA                          1.049e+01  1.600e+00   6.561   0.0963 .
## STL:ThreePPercentage            -7.138e+01  3.606e+02  -0.198   0.8756
## STL:TwoPPercentage               3.192e+02  1.497e+02   2.133   0.2791
## FTA:ThreePPercentage             1.426e+02  5.521e+01   2.583   0.2351
## FTA:TwoPPercentage              -8.051e+02  1.432e+02  -5.621   0.1121
## ThreePPercentage:TwoPPercentage -3.254e+04  8.698e+03  -3.741   0.1663
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.968 on 1 degrees of freedom
## Multiple R-squared:  0.9991, Adjusted R-squared:  0.9732
## F-statistic: 38.64 on 28 and 1 DF,  p-value: 0.1267
```

No interaction terms at level of significance alpha = 0.05

- **High order regression model**



```
##
## Call:
## lm(formula = W ~ I(Age^2) + Age + I(TRB^2) + TRB + I(Pace^2) +
##     Pace + I(STL^2) + STL + I(ThreePPercentage^2) + ThreePPercentage,
##     data = total)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.3968  -2.8819  -0.5989   2.5340  10.8438
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -5047.0905  2505.6347  -2.014   0.0584 .
## I(Age^2)                -0.1893     0.4777  -0.396   0.6964
## Age                     13.4309    25.0071   0.537   0.5974
## I(TRB^2)                 0.3495     0.2074   1.685   0.1083
## TRB                    -27.5844    18.7647  -1.470   0.1579
## I(Pace^2)               -0.5107     0.2382  -2.144   0.0452 *
## Pace                   100.6722    47.7988   2.106   0.0487 *
## I(STL^2)                -1.0509     1.5950  -0.659   0.5179
## STL                     21.1463    24.6540   0.858   0.4017
## I(ThreePPercentage^2) -1817.9575  4740.3099  -0.384   0.7056
## ThreePPercentage       1604.1234  3416.6748   0.469   0.6441

## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.316 on 19 degrees of freedom
## Multiple R-squared:  0.8721, Adjusted R-squared:  0.8047
## F-statistic: 12.95 on 10 and 19 DF,  p-value: 1.708e-06
```



```
##
## Call:
## lm(formula = W ~ Age + TRB + Pace + I(Pace^2) + STL + ThreePPercentage,
##     data = total)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.0734  -3.0448   0.3614   3.2085  10.2408
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -4559.2772  2304.3662  -1.979 0.059965 .
## Age                  3.6540     0.8294   4.406 0.000205 ***
## TRB                  4.2140     0.6147   6.855 5.45e-07 ***
## Pace                85.2919    46.0069   1.854 0.076624 .
## I(Pace^2)           -0.4341     0.2294  -1.893 0.071059 .
## STL                  5.9526     1.3618   4.371 0.000223 ***
## ThreePPercentage   232.5034    73.3367   3.170 0.004270 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.224 on 23 degrees of freedom
## Multiple R-squared:  0.8504, Adjusted R-squared:  0.8114
## F-statistic:  21.8 on 6 and 23 DF,  p-value: 2.033e-08
```

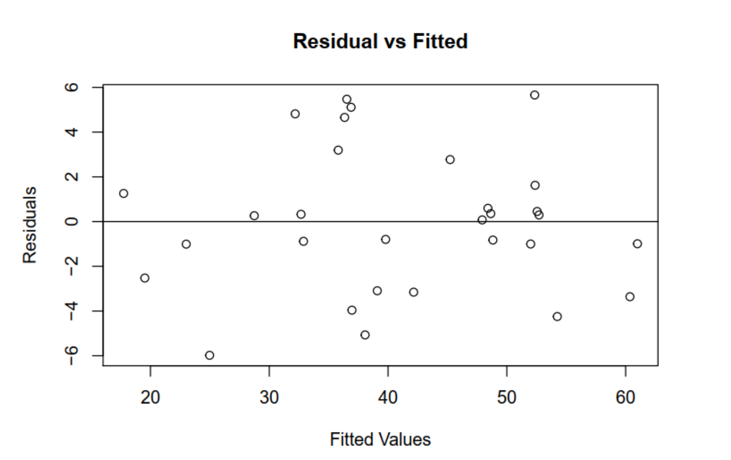
No higher order regression terms at level of significance alpha = 0.05

Modeling

Multiple Regression Assumptions:

1. **Linearity Assumption**
   (assume that there is straight-line relationship between the predictor variables and the response variable)

No pattern at the plot
The linearity assumption holds true for our predicted model

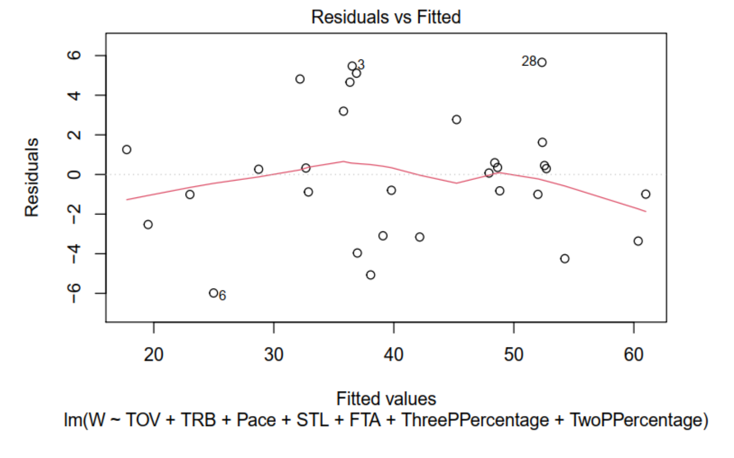


2. **Equal Variance Assumption** (Bruesch-PaganTest)

   Significance level α = 0.05
   Hypothesis:
   $H0 : Heteroscedasticity\ is\ not\ present\ (homoscedasticity)$
   $Ha : Heteroscedasticity\ is\ present$

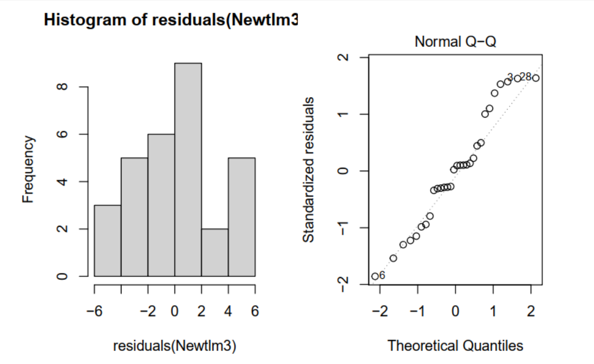
Equal Variance Assumption was met as
we fail to reject the null with p - $value$ = 0.3557.

Heteroscedasticity is not present in the model.



Modeling

3. Q-Q plot



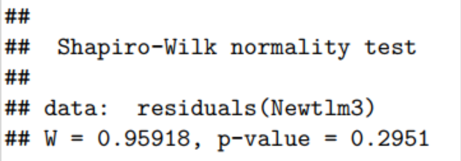
model follows a normal distribution

Shapiro-wilk test:

Significance level α = 0.05

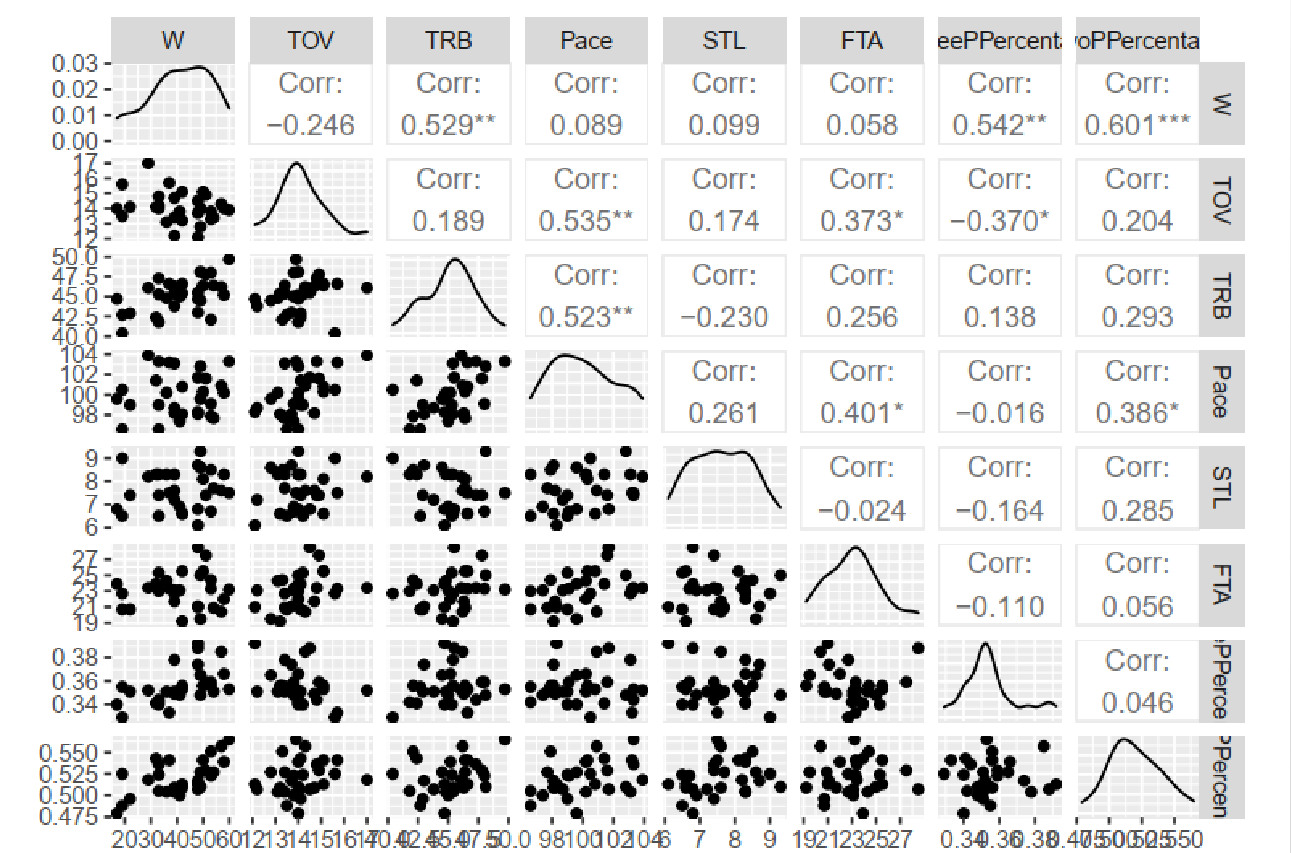
$H0 : Sample\ data\ is\ significantly\ distributed$

$Ha : Sample\ data\ is\ not\ significantly\ distributed$



```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(Newtlm3)
## W = 0.95918, p-value = 0.2951
```

We fail to reject the Null hypothesis
Model follows the Normality assumptions



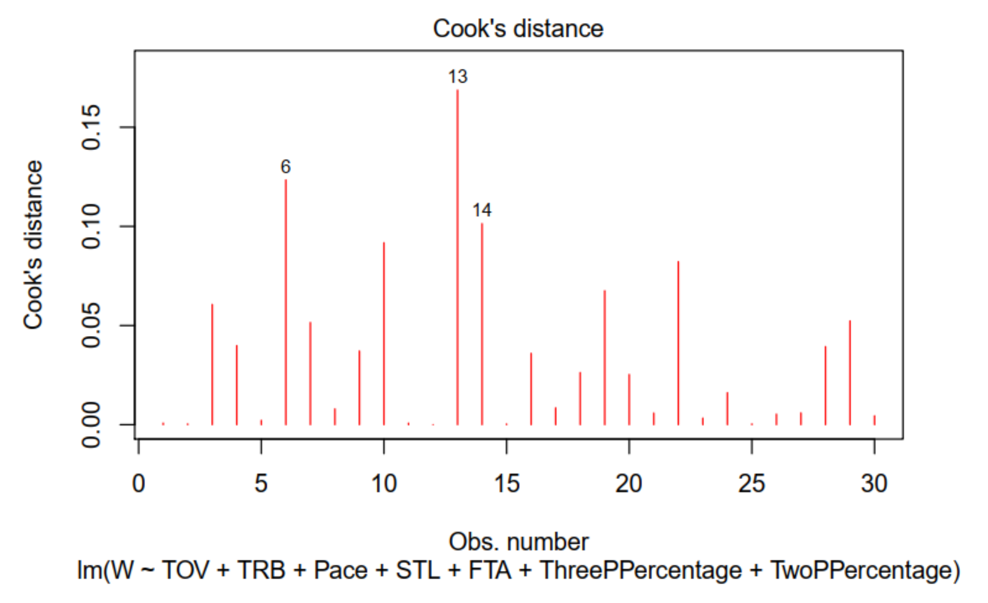
no multicollinearity in the model

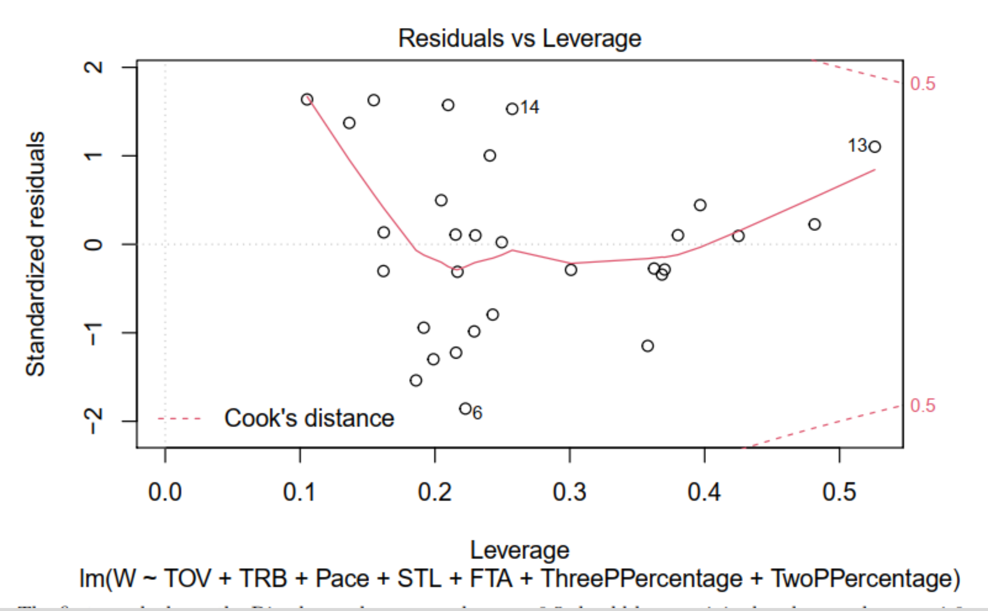
Modeling

4. Significant Outliers

- No values near 0.5, so there is no outliers according to the Cook's Distance plot.
- No values pasted the 0.5 mark at the Residual Vs. Leverage plot, no outliers

Consequently, there is no reason to change or manipulate our model further.





Conclusion

W = -148.6837 + 3.7100T*RB + 4.8999*STL + 342.7352*ThreePPercentage - 2.9395*Pace + 307.5280*TwoPPercentage + 1.0456*FTA - 1.9284*TOV’

R-Squared Adj: 0.9078
RMSE: 3.652

1. There is multi-collinearity present in Pace, FGA, FTA, ThreePA, TWOPA, and FTr. So, this variables were eliminated from the model

2. The stepwise model and backward regression procedure have high Adjusted R-squared values, showing that almost 90% of the variance in the data is explained by the model and the lowest RMSE value

3. None of the relationships in interaction model were significant.

4. Even though higher order term were significant, after combining the main effects with higher order and the interaction terms together, we found out that all the terms in our model were not significant through the partial F-test. So these terms were removed and kept only significant.

Thank you!

Modeling

1. an additive model where all the relevant variables are included
2. we will check this full-model for any signs of multicollinearity between our independent variables
3. We will utilize the ‘imcdiag’ and ‘Vif’ functions on our full-model to evaluate their variance inflation factor ratings. Redundant values will be removed.
4. Following the reduction of the model we will conduct a global F-Test to make sure that at least one our independent variables is significantly related to total wins
5. Once we have confirmed that at least one independent variable is significant, we will do three distinct stepwise procedures to create the best model These three will be regular stepwise selection using the function ‘ols_step_both_p’, backward stepwise selection using the function ‘ols_step_backward_p’ and finally forward stepwise selection utilizing the function ‘ols_step_forward_p’.
6. We will evaluate each of these models, along with their R-squared Adj. values and their Root Mean Square Error (RMSE) in order to find the combination of independent variables that create the best-fit model
7. Once we have our best combination of independent variables we will add interaction terms into the model and evaluate the significance of the interaction terms through individual t-tests. If there are no significant interaction terms then we will continue with the original best fit model. However, if any interaction terms are significant then they will be included along with the involved 2 variables unless its inclusion worsens the overall fit of the model
8. The final part in the model creation process is checking whether a higher-order-model is necessary. Any use of a higher order model will be limited because we don’t want to over-fit the model and the higher power variables are limited in terms of real-world explanation
9. Finally, once our model has passed these various checks and tests we will make sure they match five crucial assumptions:
    1. Linearity Assumption - Evaluating the Residual vs Fitted Plot
    2. Normality Assumption - Evaluating the Normal Q-Q plots; Conducting Shapiro-Wilk normality test
    3. Equal Variance Assumption - Evaluating the Residual vs Fitted and Scale-Location Plots; Conducting Breusch-Pagan test
    4. Multicollinearity - Evaluating GGpairs plot; Calculating and Evaluating the variance inflation factors (VIF)
    5. Outliers - Evaluate Cook’s distance and leverage

Notice that we do not have a test for the independence assumption since our data is not time-series. If any of these assumptions are not met we can conduct various tests and procedures to make the model better match the assumption in question.

Finally, once all the assumptions have been tested the coeffcients and the final R-Squared adjusted value will be interpreted.