Pouria Khoushehchin

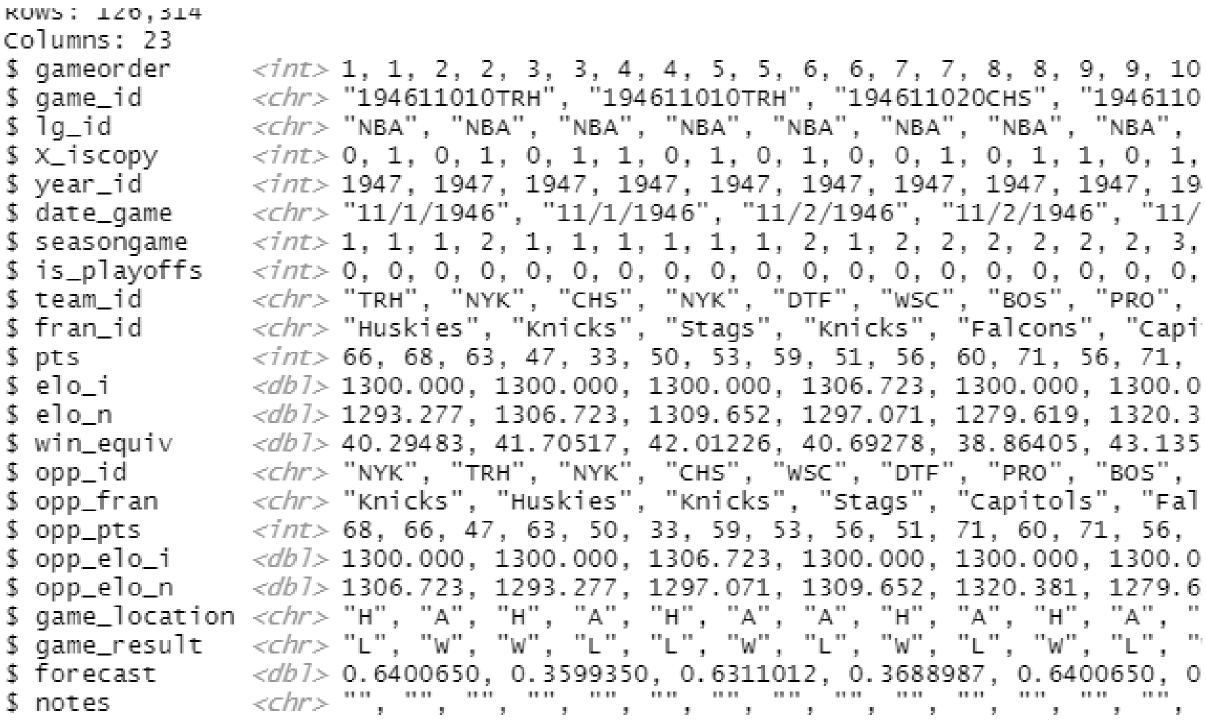
Final Project

CMSE 381

NBA Playoff Prediction

The purpose of this research is to understand the data of games in NBA and be able to predict the outcome of it. The forecast column of this Data is Elo-based chances of winning for the tea, in the team_id column, based on elo ratings and game location.

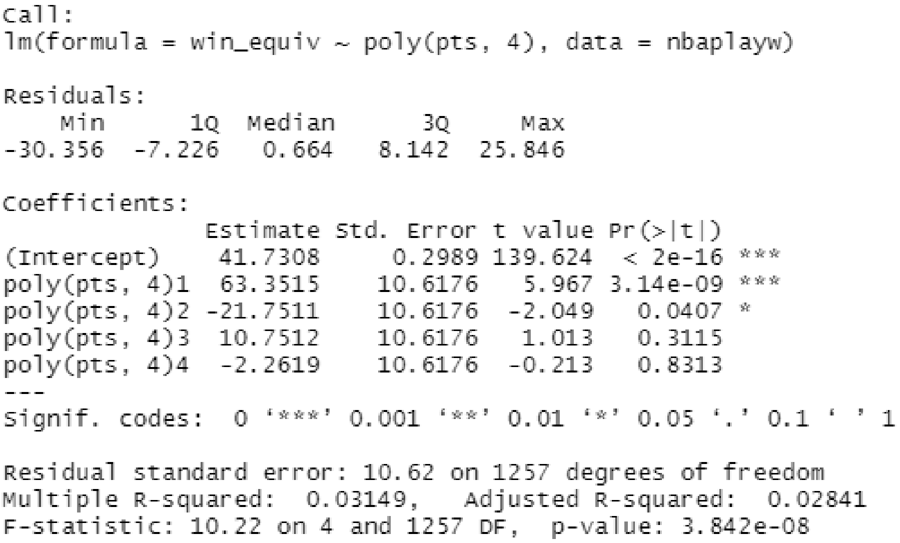
Let's take a glimpse at the data:



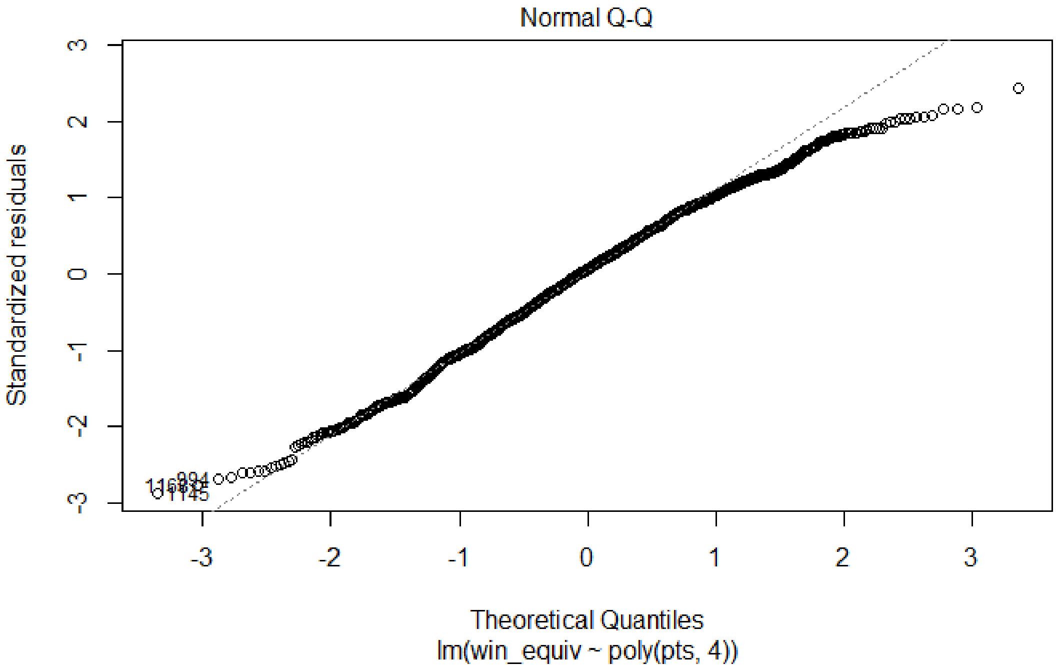
```
Rows: 126,314
Columns: 23
$ gameorder     <int> 1, 1, 2, 2, 3, 3, 4, 4, 5, 5, 6, 6, 7, 7, 8, 8, 9, 9, 10
$ game_id       <chr> "194611010TRH", "194611010TRH", "194611020CHS", "1946110
$ lg_id         <chr> "NBA", "NBA", "NBA", "NBA", "NBA", "NBA", "NBA", "NBA",
$ X_iscopy      <int> 0, 1, 0, 1, 0, 1, 1, 0, 1, 0, 1, 0, 0, 1, 0, 1, 1, 0, 1,
$ year_id       <int> 1947, 1947, 1947, 1947, 1947, 1947, 1947, 1947, 1947, 19
$ date_game     <chr> "11/1/1946", "11/1/1946", "11/2/1946", "11/2/1946", "11/
$ seasongame    <int> 1, 1, 1, 2, 1, 1, 1, 1, 1, 1, 2, 1, 2, 2, 2, 2, 2, 2, 3,
$ is_playoffs   <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
$ team_id       <chr> "TRH", "NYK", "CHS", "NYK", "DTF", "WSC", "BOS", "PRO",
$ fran_id       <chr> "Huskies", "Knicks", "Stags", "Knicks", "Falcons", "Capi
$ pts           <int> 66, 68, 63, 47, 33, 50, 53, 59, 51, 56, 60, 71, 56, 71,
$ elo_i         <dbl> 1300.000, 1300.000, 1300.000, 1306.723, 1300.000, 1300.0
$ elo_n         <dbl> 1293.277, 1306.723, 1309.652, 1297.071, 1279.619, 1320.3
$ win_equiv     <dbl> 40.29483, 41.70517, 42.01226, 40.69278, 38.86405, 43.135
$ opp_id        <chr> "NYK", "TRH", "NYK", "CHS", "WSC", "DTF", "PRO", "BOS",
$ opp_fran      <chr> "Knicks", "Huskies", "Knicks", "Stags", "Capitols", "Fal
$ opp_pts       <int> 68, 66, 47, 63, 50, 33, 59, 53, 56, 51, 71, 60, 71, 56,
$ opp_elo_i     <dbl> 1300.000, 1300.000, 1306.723, 1300.000, 1300.000, 1300.0
$ opp_elo_n     <dbl> 1306.723, 1293.277, 1297.071, 1309.652, 1320.381, 1279.6
$ game_location <chr> "H", "A", "H", "A", "H", "A", "A", "H", "A", "H", "A", "
$ game_result   <chr> "L", "W", "W", "L", "L", "W", "L", "W", "L", "W", "L", "
$ forecast      <dbl> 0.6400650, 0.3599350, 0.6311012, 0.3688987, 0.6400650, 0
$ notes         <chr> "", "", "", "", "", "", "", "", "", "", "", "", "", "",
```

We want to check the probability of correlation between how many points a team gets in NBA playoffs versus the equivalent number of wins in a 82-game season for elo_n quality. There must be a difference between a Home game versus Away game, so we are going to separate those two and go through it separately.

The data is too big to use all of it, so we limit it to year of 2000 only. For the first part we are going to check the Home playoff games. If we try to fit the data with 4th degree polynomial we get:

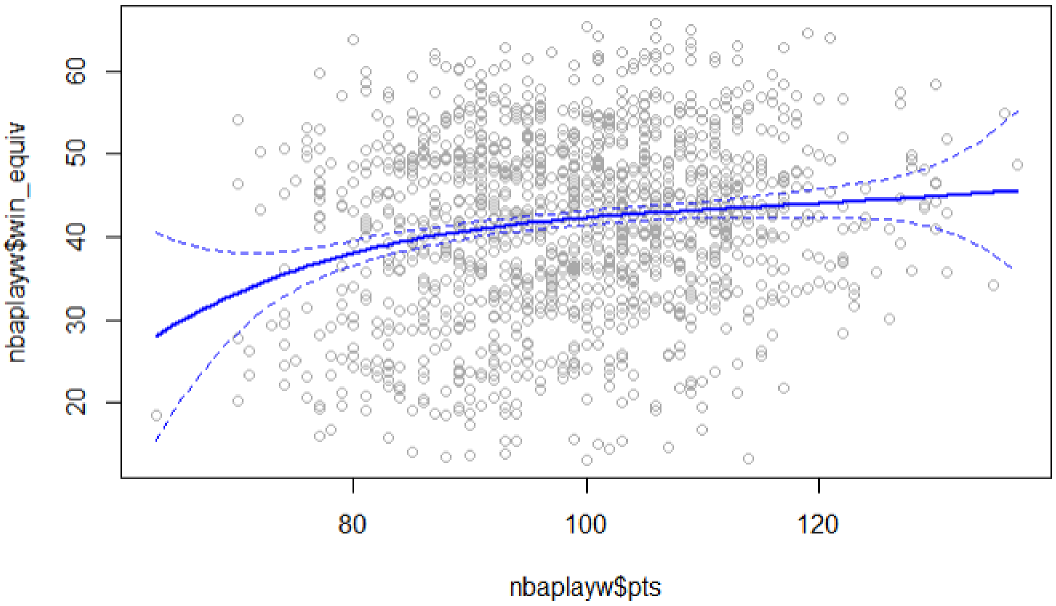


```
Call:
lm(formula = win_equiv ~ poly(pts, 4), data = nbaplayw)

Residuals:
    Min      1Q  Median      3Q     Max 
-30.356  -7.226   0.664   8.142  25.846 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)      41.7308     0.2989 139.624  < 2e-16 ***
poly(pts, 4)1    63.3515    10.6176   5.967 3.14e-09 ***
poly(pts, 4)2   -21.7511    10.6176  -2.049   0.0407 *  
poly(pts, 4)3    10.7512    10.6176   1.013   0.3115    
poly(pts, 4)4    -2.2619    10.6176  -0.213   0.8313    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.62 on 1257 degrees of freedom
Multiple R-squared:  0.03149,	Adjusted R-squared:  0.02841 
F-statistic: 10.22 on 4 and 1257 DF,  p-value: 3.842e-08
```



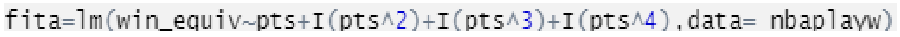
We can see the 4th degree polynomial has the highest p-value of 0.8313 that if our alpha number be 0.05, we say can that there is a probability of correlation between how many points a team gets in NBA playoffs versus the equivalent number of wins in a 82-game season for elo_n quality.

So, if we predict the outcome of this nonlinear equation we get the following outcome:



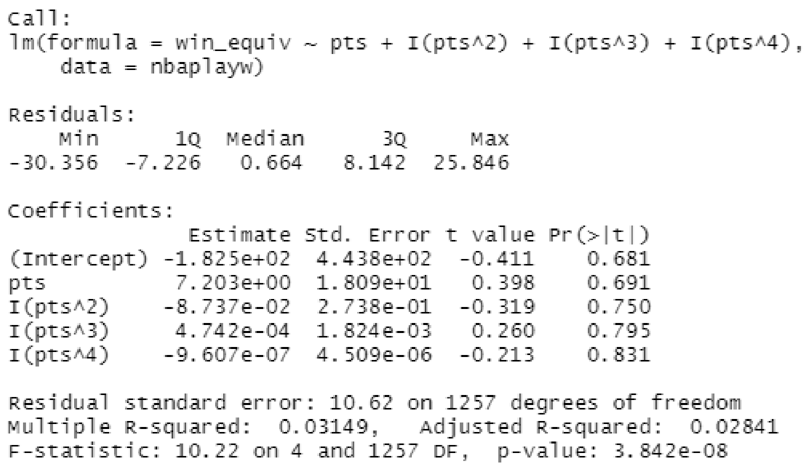
The dashed lines are the errors that can be happening. We can see the fitted line is good and has the p-value of 0.83.

Now we need to figure out the equation that works with this current fitted line.



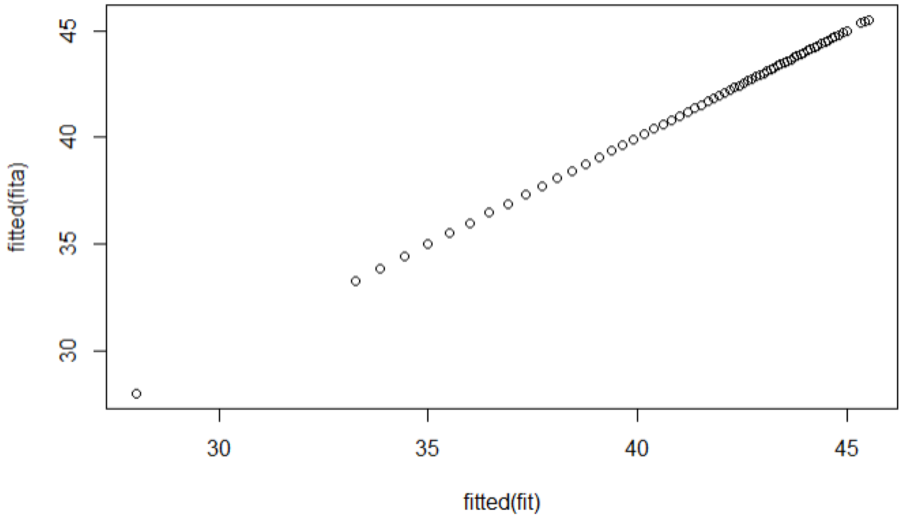
```
fita=lm(win_equiv~pts+I(pts^2)+I(pts^3)+I(pts^4),data= nbaplayw)
```

By fitting this line, we get an outcome that is a little bit different:

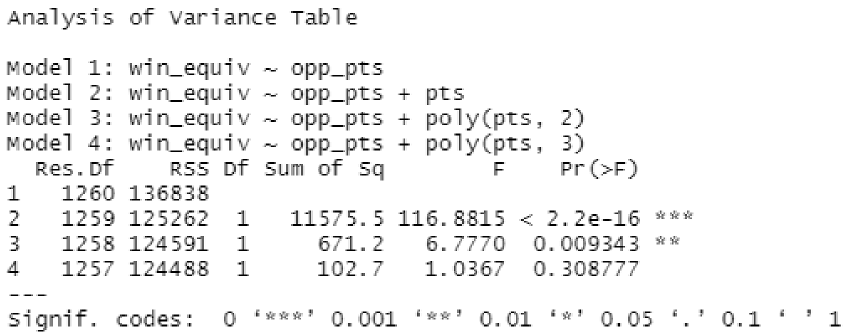


```
Call:
lm(formula = win_equiv ~ pts + I(pts^2) + I(pts^3) + I(pts^4),
    data = nbaplayw)

Residuals:
    Min      1Q  Median      3Q     Max 
-30.356  -7.226   0.664   8.142  25.846 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.825e+02  4.438e+02  -0.411    0.681
pts          7.203e+00  1.809e+01   0.398    0.691
I(pts^2)    -8.737e-02  2.738e-01  -0.319    0.750
I(pts^3)     4.742e-04  1.824e-03   0.260    0.795
I(pts^4)    -9.607e-07  4.509e-06  -0.213    0.831

Residual standard error: 10.62 on 1257 degrees of freedom
Multiple R-squared:  0.03149,   Adjusted R-squared:  0.02841 
F-statistic: 10.22 on 4 and 1257 DF,  p-value: 3.842e-08
```

However, when we graph the two fitted equation against each other we get a straight line.



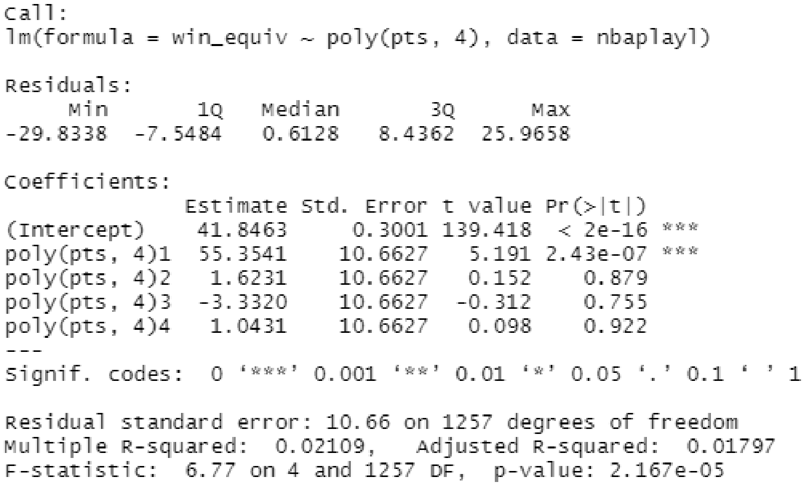
There is a chance that bringing in opponent points can improve our prediction.



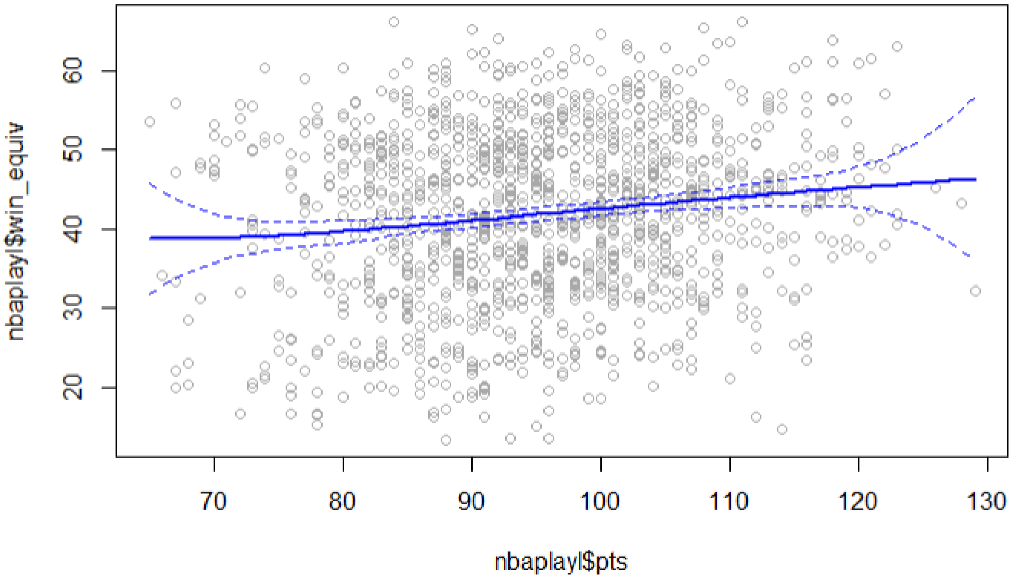
```
Analysis of Variance Table

Model 1: win_equiv ~ opp_pts
Model 2: win_equiv ~ opp_pts + pts
Model 3: win_equiv ~ opp_pts + poly(pts, 2)
Model 4: win_equiv ~ opp_pts + poly(pts, 3)
  Res.Df    RSS Df Sum of Sq        F    Pr(>F)    
1   1260 136838                                    
2   1259 125262  1   11575.5 116.8815 < 2.2e-16 ***
3   1258 124591  1     671.2   6.7770  0.009343 ** 
4   1257 124488  1     102.7   1.0367  0.308777    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The outcome is not good, and it brings the p-value of fitted equation from 0.83 to 0.3. There is a correlation there, but the new prediction lowers the previous correlation. We are not going to use opponent points in our prediction and fitted line.

For the last part of this research, we are going to try to see, the correlation in a team that is playing Away between how many games they won in a season verses points they get in playoffs.



```
Call:
lm(formula = win_equiv ~ poly(pts, 4), data = nbaplayl)

Residuals:
     Min       1Q   Median       3Q      Max 
-29.8338  -7.5484   0.6128   8.4362  25.9658 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)     41.8463     0.3001 139.418  < 2e-16 ***
poly(pts, 4)1   55.3541    10.6627   5.191 2.43e-07 ***
poly(pts, 4)2    1.6231    10.6627   0.152    0.879    
poly(pts, 4)3   -3.3320    10.6627  -0.312    0.755    
poly(pts, 4)4    1.0431    10.6627   0.098    0.922    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.66 on 1257 degrees of freedom
Multiple R-squared:  0.02109,	Adjusted R-squared:  0.01797 
F-statistic:  6.77 on 4 and 1257 DF,  p-value: 2.167e-05
```



We see that there is a correlation in a team that is playing Away between how many games they won in a season verses points they get in playoffs.

At the end we can predict how many points a team will score in playoffs when you look at their score in 82-game season in NBA. This applies for both Home and Away games.

References

Fivethirtyeight. “Fivethirtyeight/Data.” *GitHub*, 2021, github.com/fivethirtyeight/data/tree/master/nfl-elo.

Data Science Analytics. “LearningStats R Nonlinear A 112113.” *YouTube*, 15 June 2018, www.youtube.com/watch?v=u-rVXhsFyxo.