Project Part 3

Chowdhury Abdul Mumin Ishmam

2021-12-18

Question

The response variable modeled is Number of wins for a premier league team in season 2017/2018

The proposed explanatory variables are Goals For(GF) and Goals Against(GA). Number of Goals For(GF), in other words, number of goals scored by a team would help predict wins in a season as the more number of goals a team scores, the higher chances are there to win game.

Number of Goals Against(GA), in other words, number of goals conceded by a team would help predict wins in a season as the more number of goals a team concedes, the higher chances are there to not win a game.

```
library(car)
```

```
## Loading required package: carData
```

```
pl <- read.csv("pl17:18.csv")
knitr::kable(pl, "pipe", col.names =c("Team", "Wins", "Goals For", "Goals Against"), align =c("l","c","c
```

Team	Wins	Goals For	Goals Against
Manchester City	32	106	27
Manchester United	25	68	28
Tottenham Hotspur	23	74	36
Liverpool	21	84	38
Chelsea	21	62	38
Arsenal	19	74	51
Burnley	14	36	39
Everton	13	44	58
Leicester City	12	56	60
Newcastle United	12	39	47
Crystal Palace	11	45	55
Bournemouth	11	45	61
West Ham United	10	48	68
Watford	11	44	64
Brighton and Hove Albion	9	34	54
Huddersfield Town	9	28	58
Southampton	7	37	56
Swansea City	8	28	56
Stoke City	7	35	68
West Bromwich Albion	6	31	56

Dataset

Reference

"Premier League Table, Form Guide & Season Archives." Premier League Table, Form Guide & Season Archives, https://www.premierleague.com/tables?co=1&se=79&ha=-1. [Accessed Dec.15, 2021]

Explanation of variables: The variable Wins is the number of wins by a premier league club in a span of 38 games for the 2017/2018 season.

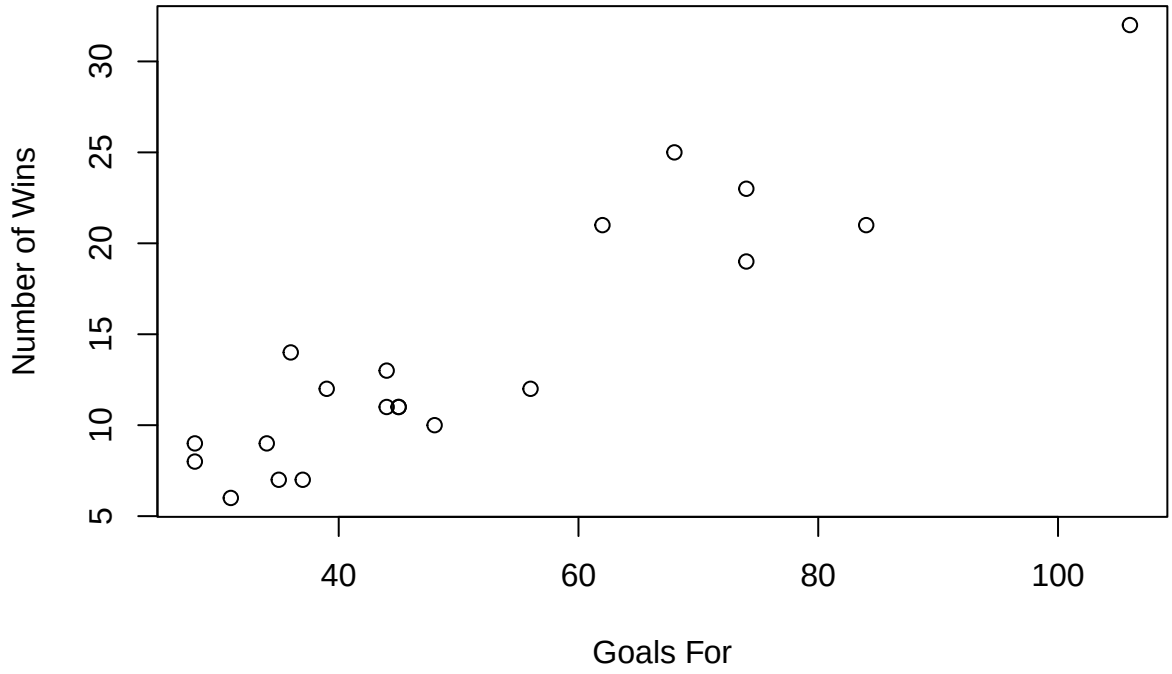
The variable Goals For(GF) is the number of goals scored by a premier league club in a span of 38 games for the 2017/2018 season.

The variable Goals Against(GA) is the number of goals conceded by a premier league club in a span of 38 games for the 2017/2018 season.

Scatterplots:

Following are the scatterplots and summary for predicting Goals for and Number of wins

```
#Scatterplot of GF and W
plot(pl$GF, pl$W, xlab = "Goals For", ylab= "Number of Wins")
```




```
pl2.lm <- lm(W ~ GF , data = pl)
summary(pl2.lm)
```

```
##
## Call:
## lm(formula = W ~ GF, data = pl)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.640 -2.132 -0.339  1.682  5.617
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.82297    1.65294  -1.103    0.285
## GF           0.31185    0.03014  10.348 5.26e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 2.754 on 18 degrees of freedom
## Multiple R-squared:  0.8561, Adjusted R-squared:  0.8481
## F-statistic: 107.1 on 1 and 18 DF,  p-value: 5.262e-09
```

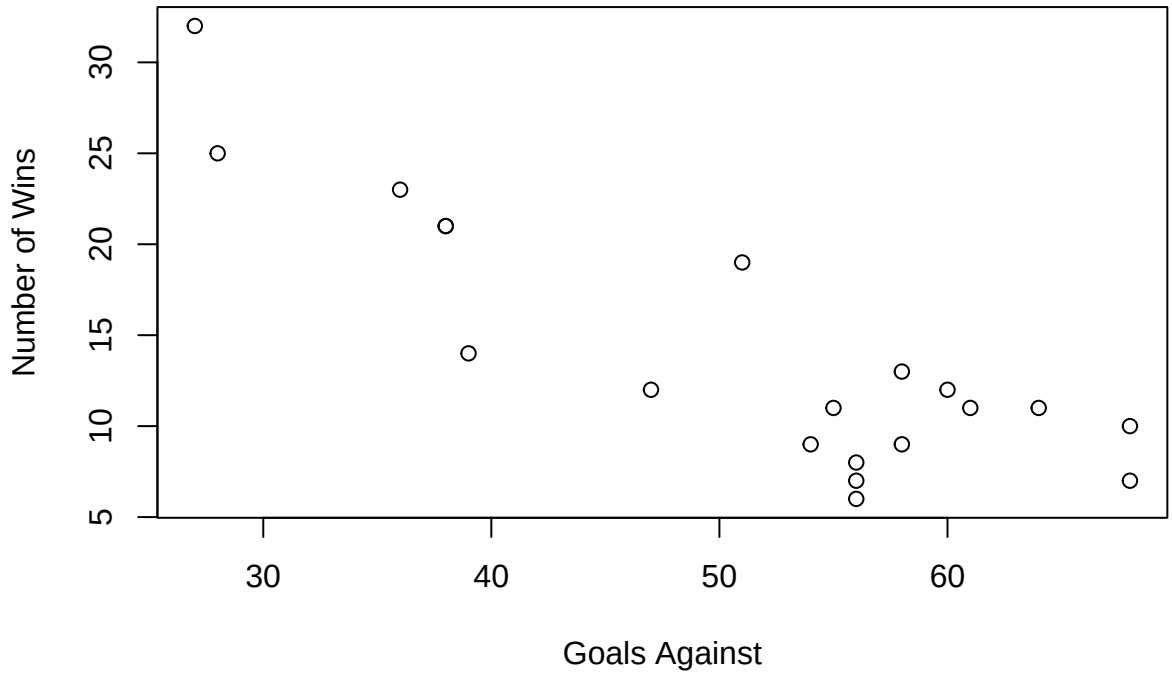
```
# R^2 = 0.8561, Adj_R^2 = 0.8481
cor(pl$GF, pl$W)
```

```
## [1] 0.9252534
```

Since $r^2 = 0.8561$, and correlation 0.9252 therefore there is a positive relationship of Goals For and Number of Wins.

Following are the scatterplots and summary for predicting Goals Against and Number of wins

```
#Scatterplot of GA and W
plot(pl$GA, pl$W, xlab = "Goals Against", ylab= "Number of Wins")
```




```
pl3.lm <- lm(W ~ GA , data = pl)
summary(pl3.lm)
```

```
##
## Call:
## lm(formula = W ~ GA, data = pl)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.8590 -3.5427  0.6529  2.3981  6.2832
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 38.89688    3.55294   10.95 2.18e-09 ***
## GA          -0.48815    0.06789   -7.19 1.08e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 3.69 on 18 degrees of freedom
## Multiple R-squared:  0.7417, Adjusted R-squared:  0.7274
## F-statistic: 51.69 on 1 and 18 DF,  p-value: 1.083e-06
```

```
# R^2 = 0.7417, Adj_R^2 = 0.7274
cor(pl$GA, pl$W)
```

```
## [1] -0.8612375
```

Since $r^2 = 0.7471$, and correlation of -0.8612 therefore there is a negative relationship of Goals Against and Number of Wins.

Preliminary Model

The following is the the model predicting Number of wins from Goals For(X1)

```
lm.x1<- lm(W~GF, data = pl)
lm.x1
```

```
##
## Call:
## lm(formula = W ~ GF, data = pl)
##
## Coefficients:
## (Intercept)           GF
##     -1.8230       0.3118
```

Therefore, the regression line for predicting wins from goals for is:

$\hat{y} = -1.8230 + 0.3118X_1$

The following is the the model predicting Number of wins from Goals Against(X2)

```
lm.x2<- lm(W~GA, data = pl)
lm.x2
```

```
##
## Call:
## lm(formula = W ~ GA, data = pl)
##
## Coefficients:
## (Intercept)           GA
##     38.8969      -0.4882
```

Therefore, the regression line for predicting wins from goals for is:

$\hat{y} = 38.8969 - 0.4882X_2$

The following is the the model predicting Number of wins from Goals For(X1) and Goals Against(X2)

```
lm.x1x2<- lm(W~GF+GA, data = pl)
lm.x1x2
```

```
##
## Call:
## lm(formula = W ~ GF + GA, data = pl)
##
## Coefficients:
## (Intercept)           GF           GA
##     15.5880       0.2129      -0.2431
```

Therefore, the regression line for predicting wins from goals for and goals against is:

$\hat{y} = 15.5880 + 0.2129X_1 - 0.2431X_2$

```
summary(lm.x1x2)
```

```
##
## Call:
## lm(formula = W ~ GF + GA, data = pl)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2339 -0.7228  0.3184  0.9063  2.1453
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 15.58795    3.05319   5.105 8.79e-05 ***
## GF           0.21292    0.02406   8.849 9.00e-08 ***
## GA          -0.24313    0.04046  -6.009 1.41e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.604 on 17 degrees of freedom
## Multiple R-squared:  0.9539, Adjusted R-squared:  0.9485
## F-statistic:   176 on 2 and 17 DF,  p-value: 4.353e-12
```

As mentioned above in dataset part, the ${R_{adj}}^2$ for predicting wins from goals for is 0.8481 and predicting wins from goals against is 0.7274. Above we see that the ${R_{adj}}^2$ for predicting Wins from Goals for and Goals against combined increased from both the models to 0.9485.

The following is the second-order model:

```
pl.full <- lm(W~pl$GF+pl$GA+I(pl$GF^2)+I(pl$GA^2)+pl$GF*pl$GA, data= pl)
pl.full
```

```
##
## Call:
## lm(formula = W ~ pl$GF + pl$GA + I(pl$GF^2) + I(pl$GA^2) + pl$GF *
##     pl$GA, data = pl)
##
## Coefficients:
## (Intercept)        pl$GF        pl$GA   I(pl$GF^2)   I(pl$GA^2)  pl$GF:pl$GA
##  29.0398430    0.2684704   -0.8450661   -0.0007179    0.0058767    0.0001757
```

Therefore, the regression line for the full second-order model is:

$\hat{y} = 29.0398 + 0.2685X_1 - 0.8451X_2 - 0.0007179X_1^2 + 0.0059X_2^2 + 0.0001757X_1X_2$

Below is the ANOVA test for the full second-order model stated above:

```
summary(pl.full)
```

```
##
## Call:
## lm(formula = W ~ pl$GF + pl$GA + I(pl$GF^2) + I(pl$GA^2) + pl$GF *
##     pl$GA, data = pl)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

```
## -2.46035 -0.58739 -0.00287  0.73259  2.33362
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 29.0398430 22.8783163   1.269    0.225
## pl$GF        0.2684704  0.3720236   0.722    0.482
## pl$GA       -0.8450661  0.5840846  -1.447    0.170
## I(pl$GF^2)  -0.0007179  0.0016073  -0.447    0.662
## I(pl$GA^2)   0.0058767  0.0040578   1.448    0.170
## pl$GF:pl$GA  0.0001757  0.0043148   0.041    0.968
##
## Residual standard error: 1.532 on 14 degrees of freedom
## Multiple R-squared:  0.9654, Adjusted R-squared:  0.953
## F-statistic:  78.1 on 5 and 14 DF,  p-value: 1.012e-09
```

LEVEL OF SIGNIFICANCE: $\alpha = 0.05$

HYPOTHESES: $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$ vs H_A : At least one $\beta_i \neq 0$ for i = 1,. . . ,5

DECISION RULE: Reject H_0 if p-value is $\leq \alpha$

TEST STATISTIC: F = 78.1

P-VALUE: ≈ 0

CONCLUSION: As p-value is $\leq \alpha = 0.05$, reject H_0. Conclude there is sufficient evidence atleast one of the model terms is significant or does a sufficient job at explaining the variation on the number of wins.

Model Refinement

Below is the summary for the full model:

```
summary(pl.full)
```

```
##
## Call:
## lm(formula = W ~ pl$GF + pl$GA + I(pl$GF^2) + I(pl$GA^2) + pl$GF *
##     pl$GA, data = pl)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.46035 -0.58739 -0.00287  0.73259  2.33362
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 29.0398430 22.8783163   1.269    0.225
## pl$GF        0.2684704  0.3720236   0.722    0.482
## pl$GA       -0.8450661  0.5840846  -1.447    0.170
## I(pl$GF^2)  -0.0007179  0.0016073  -0.447    0.662
## I(pl$GA^2)   0.0058767  0.0040578   1.448    0.170
## pl$GF:pl$GA  0.0001757  0.0043148   0.041    0.968
##
## Residual standard error: 1.532 on 14 degrees of freedom
## Multiple R-squared:  0.9654, Adjusted R-squared:  0.953
## F-statistic:  78.1 on 5 and 14 DF,  p-value: 1.012e-09
```

None of our model terms seems significant on the summary, therefore we will start eliminating the higher order model terms.

```
vif(pl.full)
```

```
##       pl$GF       pl$GA  I(pl$GF^2)  I(pl$GA^2) pl$GF:pl$GA
##   492.83227   429.54307   146.27837   190.22887    67.46585
```

```
pl.full1 <- lm(W~GF + GA + I(GF^2) + GF*GA, data= pl)
pl.full1
```

```
##
## Call:
## lm(formula = W ~ GF + GA + I(GF^2) + GF * GA, data = pl)
##
## Coefficients:
## (Intercept)           GF           GA      I(GF^2)        GF:GA
##   -0.301628     0.639637    -0.029274    -0.002058    -0.004348
```

```
summary(pl.full1)
```

```
##
## Call:
## lm(formula = W ~ GF + GA + I(GF^2) + GF * GA, data = pl)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.9181 -0.5532  0.4120  0.9667  1.9346
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.301628  11.008928  -0.027   0.9785
## GF           0.639637   0.279349   2.290   0.0369 *
## GA          -0.029274   0.159952  -0.183   0.8572
## I(GF^2)     -0.002058   0.001361  -1.511   0.1515
## GF:GA       -0.004348   0.003084  -1.410   0.1790
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.587 on 15 degrees of freedom
## Multiple R-squared:  0.9602, Adjusted R-squared:  0.9496
## F-statistic: 90.48 on 4 and 15 DF,  p-value: 2.586e-10
```

Therefore, we can see that still only GF is significant but none else is. We will now eliminate the GF^2.

```
vif(pl.full1)
```

```
##        GF        GA   I(GF^2)     GF:GA
## 258.93406  30.01715  97.80697  32.11366
```

```
pl.full2 <- lm(W~GF + GA + GF*GA, data= pl)
pl.full2
```

```
##
## Call:
## lm(formula = W ~ GF + GA + GF * GA, data = pl)
##
## Coefficients:
## (Intercept)           GF           GA        GF:GA
##  14.4210905    0.2329578   -0.2171254   -0.0004862
```

```
summary(pl.full2)
```

```
##
## Call:
## lm(formula = W ~ GF + GA + GF * GA, data = pl)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.1869 -0.7330  0.2956  0.9210  2.1628
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 14.4210905  5.3311235   2.705  0.01561 *
## GF           0.2329578  0.0780286   2.986  0.00874 **
## GA          -0.2171254  0.1046453  -2.075  0.05449 .
## GF:GA       -0.0004862  0.0017951  -0.271  0.78998
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.649 on 16 degrees of freedom
## Multiple R-squared:  0.9541, Adjusted R-squared:  0.9455
## F-statistic:   111 on 3 and 16 DF,  p-value: 6.393e-11
```

Still, we can see that only GF is significant. Now, we will eliminate the final higher order term.

```
vif(pl.full2)
```

```
##       GF       GA    GF:GA
## 18.70121 11.89319 10.07215
```

```
pl.reduced <- lm(W~GF + GA, data= pl)
pl.reduced
```

```
##
## Call:
## lm(formula = W ~ GF + GA, data = pl)
##
## Coefficients:
## (Intercept)           GF           GA
##     15.5880       0.2129      -0.2431
```

```
summary(pl.reduced)
```

```
##
## Call:
## lm(formula = W ~ GF + GA, data = pl)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2339 -0.7228  0.3184  0.9063  2.1453
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 15.58795    3.05319   5.105 8.79e-05 ***
## GF           0.21292    0.02406   8.849 9.00e-08 ***
## GA          -0.24313    0.04046  -6.009 1.41e-05 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.604 on 17 degrees of freedom
## Multiple R-squared:  0.9539, Adjusted R-squared:  0.9485
## F-statistic:   176 on 2 and 17 DF,  p-value: 4.353e-12
```

Now, we can see that our model terms GF and GA are significant now. Therefore, this will be our reduced model. Reduced Model:

$\hat{y} = 15.5879 + 0.2129X_1 - 0.2431X_2$

Following are the co-efficients that are significant now and their p-values:

Goals For(GF) and their p-value: ≈ 0

Goals Against(GA) and their p-value: ≈ 0

Following will be the nested F-test,

```
anova(pl.reduced,pl.full)
```

```
## Analysis of Variance Table
##
## Model 1: W ~ GF + GA
## Model 2: W ~ pl$GF + pl$GA + I(pl$GF^2) + I(pl$GA^2) + pl$GF * pl$GA
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     17 43.715
## 2     14 32.844  3    10.871 1.5446 0.2469
```

Nested F-Test:

LEVEL OF SIGNIFICANCE: $\alpha = 0.05$

HYPOTHESES: $H_0 : \beta_3 = \beta_4 = \beta_5 = 0$ vs H_A : At least one $\beta_i \neq 0$ for i = 3,4,5

DECISION RULE: Reject H_0 if p-value is $\leq \alpha$

TEST STATISTIC: F = 1.5446

P-VALUE: 0.2469

CONCLUSION: As p-value is $\geq \alpha = 0.05$, fail to reject H_0. Conclude there is insufficient evidence atleast one of the co-efficients for $GF^2, GA^2, GF * GA$ is non-zero.

Final Model and Assesment

Below we will perform the ANOVA test for our reduced model.

```
summary(pl.reduced)
```

```
##
## Call:
## lm(formula = W ~ GF + GA, data = pl)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2339 -0.7228  0.3184  0.9063  2.1453
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 15.58795    3.05319   5.105 8.79e-05 ***
```

```
## GF           0.21292    0.02406   8.849 9.00e-08 ***
## GA          -0.24313    0.04046  -6.009 1.41e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.604 on 17 degrees of freedom
## Multiple R-squared:  0.9539, Adjusted R-squared:  0.9485
## F-statistic:   176 on 2 and 17 DF,  p-value: 4.353e-12
```

LEVEL OF SIGNIFICANCE: $\alpha = 0.05$

HYPOTHESES: $H_0 : \beta_3 = \beta_4 = \beta_5 = 0$ vs H_A : At least one $\beta_i \neq 0$ for i = 3,4,5

DECISION RULE: Reject H_0 if p-value is $\leq \alpha$

TEST STATISTIC: F = 176

P-VALUE: ≈ 0

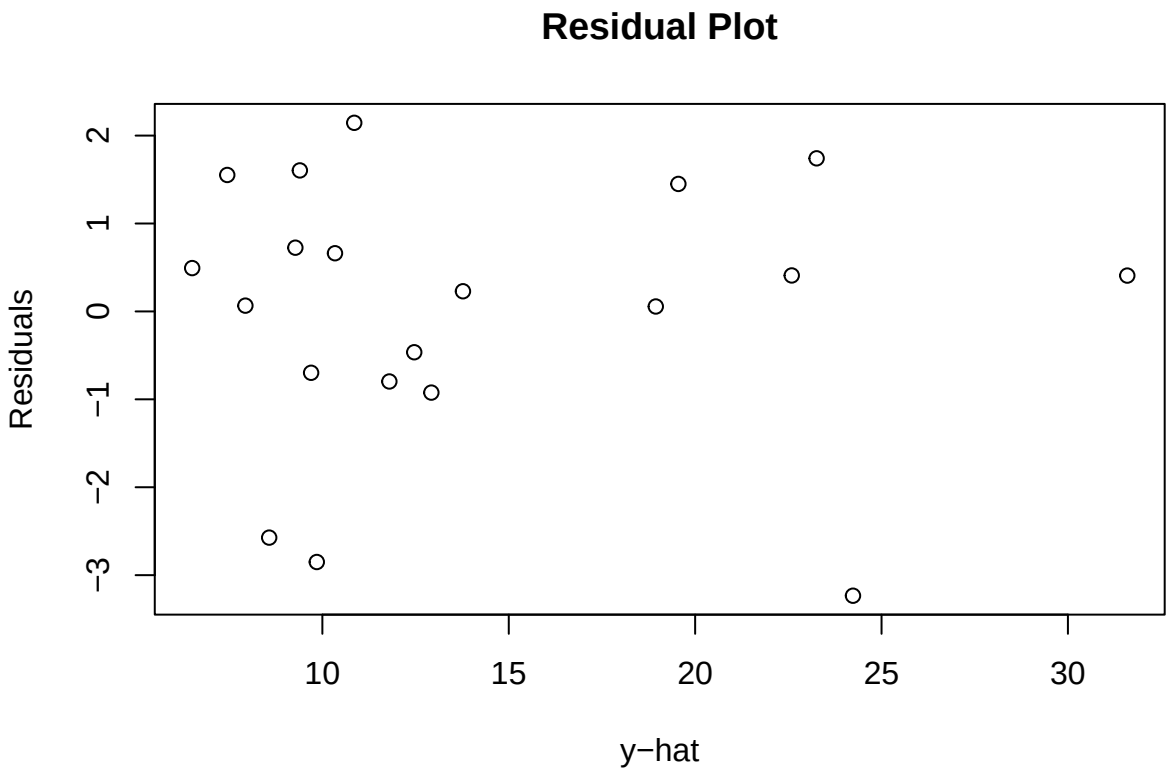
CONCLUSION: As p-value is $\leq \alpha = 0.05$, reject H_0. Conclude there is sufficient evidence atleast one of the eliminated model terms is significant or does a sufficient job at explaining the variation on the number of wins.

Final Regression equation is followed:

$\hat{y} = 15.5875 + 0.2129X_1 - 0.2431X_2$

Final model Residual plot:

```
reducedpl.res <- resid(pl.reduced)
reducedpl.fitted <- fitted.values(pl.reduced)
plot(reducedpl.fitted,reducedpl.res,xlab="y-hat",ylab="Residuals",main="Residual Plot")
```

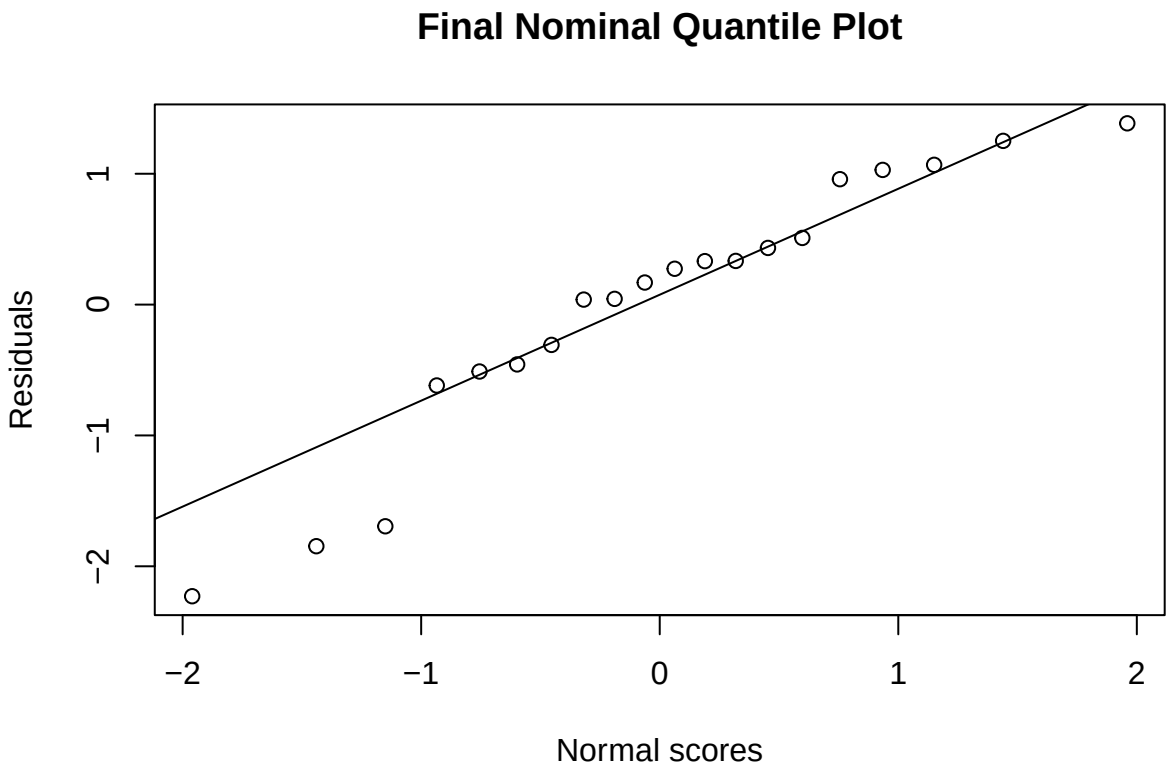


The residual plot seems to have a no discernible pattern instead of one point above 30. It is fine, as most of the points follow a patten with equal spread, we can assume that linearity and constant variance conditions

are met.

Final Model Nominal Quantile Plot:

```
reducedpl.stdres <- rstandard(pl.reduced)
qqnorm(reducedpl.stdres,ylab="Residuals",xlab="Normal scores",main="Final Nominal Quantile Plot")
qqline(reducedpl.stdres)
```



It seems that the normality is fairly reasonable assumption as most of the points fall on the straight line with just a skewness on the left tail, but since regression line is fairly robust, we assume, the normality condition is met.

Conclusion

Therefore we can conclude that Goal For and Goals Against are two explanatory variable that explains the variation in predicting the number of wins a premier league club. The tests used also suggests that individually, the predictors had a adjusted multiple co-efficient of determination of about 0.8481 and 0.7274. After predicting with the 2 predictors, the adjusted co-efficient of determination increases to 0.9485. This suggests that most of the variation in number of wins is from both predictors and therefore, combining the terms together was meaningful.

Final Regression Equation:

$\hat{y} = 15.5875 + 0.2129X_1 - 0.2431X_2$