STAT 420 Final Project: Does College Quarterback Experience Translate to Professional Success?

#### Ajay Dugar (dugar3), Sam Swislow (swislow2), Sairaj Vorugant (sairajv2)

#### August 3, 2019

Introduction

In the National Football League(NFL), the quarterback position reigns supreme. For teams without a great quarterback, their main concern is getting one. For teams with a great quarterback, their main concern is maximizing their window to win games. Finding a franchise quarterback isn’t an exact science, but as a group of NFL fans, we wanted to see if a quarterback’s NFL success is based on their college experience. In order to model this fact, we used a dummy variable to divide the quarterbacks into two groups, quarterbacks with less than 26 college starts (0) and those with 26 or more quarterback starts (1) (26 games = 2 full regular college seasons).

The college dataset was found from [Sports-Reference College Football](https://www.sports-reference.com/cfb/) website and the NFL dataset was found from [Pro Football Reference’s website](https://www.pro-football-reference.com/). We examined the set of quarterbacks in the NFL (200 pass attempts per year minimum) dating back from 2005 and recorded their career average QBR. We then fit a multiple linear regression selected using backwards BIC. For the starting full model, we regressed the independent variables of wheter they were at least a two season starter (dummy variable), passing yards per attempt, touchdown to interception ratio, completion percentage, rushing yards per attempt, and rushing yards per game, as well as all three-way interactions, on the dependent variable of their professional career average QBR. To be clear, all of the independent variables are from the player’s college career.

We will look at whether the overall model is a good predictor of NFL QBR, as well as whether the coefficient of being a two season starter is statistically significant.

Methods

Data Preprocessing

In order to generate the dataset, we pulled the data from the websites mentioned above and combined them. We took the average QBR of every quarterback for every year of their NFL career. For players who had incomplete data, this data was found from a variety of different sources (old college media guides, calculated by hand, etc.).

Data Analysis and Model Building

QBR is a proprietary statistic developed by ESPN to measure quarterback performance that takes into account their overall contributions to winning including passing, rushing, turnovers, penalties, with respect to the conditions on the field, such as time left in the game, down, and distance. We chose to use QBR because it is the most precise statistic when it comes to quarterback play, by analyzing each player on a play-to-play level.

To build the model, we first had to ensure that there was no serious multicollinearity within the data. Below is the correlation matrix:

## Avg.NFL.QBR Two.Season.Starter Pass.Yds.Per.Att  
## Avg.NFL.QBR 1.00000000 -0.028614757 0.024067364  
## Two.Season.Starter -0.02861476 1.000000000 0.009823692  
## Pass.Yds.Per.Att 0.02406736 0.009823692 1.000000000  
## TD.Int.Ratio 0.06132287 0.011189963 0.580303832  
## Comp.Pct -0.10601655 0.036837952 0.570434923  
## Rush.Yds.Per.Att 0.21096977 -0.090587554 0.160206618  
## Rush.Tds.Per.G 0.24767753 -0.065180742 0.274635362  
## TD.Int.Ratio Comp.Pct Rush.Yds.Per.Att  
## Avg.NFL.QBR 0.06132287 -0.10601655 0.21096977  
## Two.Season.Starter 0.01118996 0.03683795 -0.09058755  
## Pass.Yds.Per.Att 0.58030383 0.57043492 0.16020662  
## TD.Int.Ratio 1.00000000 0.44239149 0.17282138  
## Comp.Pct 0.44239149 1.00000000 0.03853563  
## Rush.Yds.Per.Att 0.17282138 0.03853563 1.00000000  
## Rush.Tds.Per.G 0.18045890 0.12295700 0.55644824  
## Rush.Tds.Per.G  
## Avg.NFL.QBR 0.24767753  
## Two.Season.Starter -0.06518074  
## Pass.Yds.Per.Att 0.27463536  
## TD.Int.Ratio 0.18045890  
## Comp.Pct 0.12295700  
## Rush.Yds.Per.Att 0.55644824  
## Rush.Tds.Per.G 1.00000000

With no correlations between the indepedent variables greater than 0.8, we can safely say that there is not any serious multicollineary in the data.

Next, let us reduce the full model using backwards BIC. The summary of the reduced model is below:

full\_model = lm(Avg.NFL.QBR ~ (Two.Season.Starter + Pass.Yds.Per.Att + TD.Int.Ratio + Comp.Pct + Rush.Yds.Per.Att + Rush.Tds.Per.G)^3, data = qb\_data)  
reduced\_model = step(full\_model, direction = "backward", k = log(nrow(qb\_data)), trace = 0)  
summary(reduced\_model)

##   
## Call:  
## lm(formula = Avg.NFL.QBR ~ Two.Season.Starter + Pass.Yds.Per.Att +   
## TD.Int.Ratio + Comp.Pct + Rush.Yds.Per.Att + Rush.Tds.Per.G +   
## Two.Season.Starter:Pass.Yds.Per.Att + Two.Season.Starter:TD.Int.Ratio +   
## Two.Season.Starter:Comp.Pct + Two.Season.Starter:Rush.Yds.Per.Att +   
## Two.Season.Starter:Rush.Tds.Per.G + Pass.Yds.Per.Att:Comp.Pct +   
## TD.Int.Ratio:Rush.Yds.Per.Att + TD.Int.Ratio:Rush.Tds.Per.G +   
## Comp.Pct:Rush.Yds.Per.Att + Comp.Pct:Rush.Tds.Per.G + Two.Season.Starter:Pass.Yds.Per.Att:Comp.Pct +   
## Two.Season.Starter:TD.Int.Ratio:Rush.Yds.Per.Att + Two.Season.Starter:TD.Int.Ratio:Rush.Tds.Per.G +   
## Two.Season.Starter:Comp.Pct:Rush.Yds.Per.Att + Two.Season.Starter:Comp.Pct:Rush.Tds.Per.G,   
## data = qb\_data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -40.307 -10.388 0.547 10.694 37.965   
##   
## Coefficients:  
## Estimate Std. Error  
## (Intercept) -671.8653 245.3013  
## Two.Season.Starter 776.1106 249.5576  
## Pass.Yds.Per.Att 146.5818 46.7097  
## TD.Int.Ratio -27.1281 12.9018  
## Comp.Pct 14.8998 5.2813  
## Rush.Yds.Per.Att 479.6425 151.4194  
## Rush.Tds.Per.G -3956.6905 1213.0988  
## Two.Season.Starter:Pass.Yds.Per.Att -156.0554 47.3429  
## Two.Season.Starter:TD.Int.Ratio 28.3944 13.0927  
## Two.Season.Starter:Comp.Pct -15.9518 5.3433  
## Two.Season.Starter:Rush.Yds.Per.Att -475.5355 151.5060  
## Two.Season.Starter:Rush.Tds.Per.G 4004.5137 1215.4625  
## Pass.Yds.Per.Att:Comp.Pct -2.6626 0.8542  
## TD.Int.Ratio:Rush.Yds.Per.Att 28.4210 9.6843  
## TD.Int.Ratio:Rush.Tds.Per.G -124.7030 51.3785  
## Comp.Pct:Rush.Yds.Per.Att -8.7842 2.7883  
## Comp.Pct:Rush.Tds.Per.G 68.6912 21.3272  
## Two.Season.Starter:Pass.Yds.Per.Att:Comp.Pct 2.8053 0.8640  
## Two.Season.Starter:TD.Int.Ratio:Rush.Yds.Per.Att -28.1807 9.7017  
## Two.Season.Starter:TD.Int.Ratio:Rush.Tds.Per.G 123.0435 51.7630  
## Two.Season.Starter:Comp.Pct:Rush.Yds.Per.Att 8.7141 2.7897  
## Two.Season.Starter:Comp.Pct:Rush.Tds.Per.G -69.1838 21.3651  
## t value Pr(>|t|)   
## (Intercept) -2.739 0.00679 \*\*  
## Two.Season.Starter 3.110 0.00218 \*\*  
## Pass.Yds.Per.Att 3.138 0.00199 \*\*  
## TD.Int.Ratio -2.103 0.03691 \*   
## Comp.Pct 2.821 0.00533 \*\*  
## Rush.Yds.Per.Att 3.168 0.00181 \*\*  
## Rush.Tds.Per.G -3.262 0.00133 \*\*  
## Two.Season.Starter:Pass.Yds.Per.Att -3.296 0.00118 \*\*  
## Two.Season.Starter:TD.Int.Ratio 2.169 0.03144 \*   
## Two.Season.Starter:Comp.Pct -2.985 0.00323 \*\*  
## Two.Season.Starter:Rush.Yds.Per.Att -3.139 0.00199 \*\*  
## Two.Season.Starter:Rush.Tds.Per.G 3.295 0.00119 \*\*  
## Pass.Yds.Per.Att:Comp.Pct -3.117 0.00213 \*\*  
## TD.Int.Ratio:Rush.Yds.Per.Att 2.935 0.00378 \*\*  
## TD.Int.Ratio:Rush.Tds.Per.G -2.427 0.01622 \*   
## Comp.Pct:Rush.Yds.Per.Att -3.150 0.00191 \*\*  
## Comp.Pct:Rush.Tds.Per.G 3.221 0.00152 \*\*  
## Two.Season.Starter:Pass.Yds.Per.Att:Comp.Pct 3.247 0.00140 \*\*  
## Two.Season.Starter:TD.Int.Ratio:Rush.Yds.Per.Att -2.905 0.00414 \*\*  
## Two.Season.Starter:TD.Int.Ratio:Rush.Tds.Per.G 2.377 0.01852 \*   
## Two.Season.Starter:Comp.Pct:Rush.Yds.Per.Att 3.124 0.00209 \*\*  
## Two.Season.Starter:Comp.Pct:Rush.Tds.Per.G -3.238 0.00144 \*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 17.03 on 177 degrees of freedom  
## Multiple R-squared: 0.16, Adjusted R-squared: 0.06037   
## F-statistic: 1.606 on 21 and 177 DF, p-value: 0.05228

All of the individual t-tests for the coefficients are all significant at the \(\alpha = 0.05\) level. The adjusted \(R^2\) value is 0.06037. Let’s look at whether this model is a true improvement, based on an F-test with the full model.

anova(full\_model, reduced\_model)

## Analysis of Variance Table  
##   
## Model 1: Avg.NFL.QBR ~ (Two.Season.Starter + Pass.Yds.Per.Att + TD.Int.Ratio +   
## Comp.Pct + Rush.Yds.Per.Att + Rush.Tds.Per.G)^3  
## Model 2: Avg.NFL.QBR ~ Two.Season.Starter + Pass.Yds.Per.Att + TD.Int.Ratio +   
## Comp.Pct + Rush.Yds.Per.Att + Rush.Tds.Per.G + Two.Season.Starter:Pass.Yds.Per.Att +   
## Two.Season.Starter:TD.Int.Ratio + Two.Season.Starter:Comp.Pct +   
## Two.Season.Starter:Rush.Yds.Per.Att + Two.Season.Starter:Rush.Tds.Per.G +   
## Pass.Yds.Per.Att:Comp.Pct + TD.Int.Ratio:Rush.Yds.Per.Att +   
## TD.Int.Ratio:Rush.Tds.Per.G + Comp.Pct:Rush.Yds.Per.Att +   
## Comp.Pct:Rush.Tds.Per.G + Two.Season.Starter:Pass.Yds.Per.Att:Comp.Pct +   
## Two.Season.Starter:TD.Int.Ratio:Rush.Yds.Per.Att + Two.Season.Starter:TD.Int.Ratio:Rush.Tds.Per.G +   
## Two.Season.Starter:Comp.Pct:Rush.Yds.Per.Att + Two.Season.Starter:Comp.Pct:Rush.Tds.Per.G  
## Res.Df RSS Df Sum of Sq F Pr(>F)  
## 1 157 45321   
## 2 177 51352 -20 -6031 1.0446 0.4141

We have an F-statistic of 1.0446, witha corresponding p-value of 0.4141. So this means that the additional coefficients from the full model don’t add any significant explanatory power to the reduced model. So we can continue with the reduced model.

Results

Based on the reduced model, it appears that we don’t have a good regression due to the adjusted \(R^2\) value of 0.06037. So we should see if we have any influential points that are throwing off the regression. Well rerun the regression after removing influential points (based of the Cook’s Distance metric)

cd = cooks.distance(reduced\_model)  
cutoff = 4 / nrow(qb\_data)  
influential <- as.numeric(names(cd)[(cd > cutoff)])  
qb\_data2 <- qb\_data[-influential, ]

We removed 19 points. So let’s rerun the reduced model on the reduced data.

reduced\_model2 = lm(Avg.NFL.QBR ~ Two.Season.Starter + Pass.Yds.Per.Att + TD.Int.Ratio +   
 Comp.Pct + Rush.Yds.Per.Att + Rush.Tds.Per.G + Two.Season.Starter:Pass.Yds.Per.Att +   
 Two.Season.Starter:TD.Int.Ratio + Two.Season.Starter:Comp.Pct +   
 Two.Season.Starter:Rush.Yds.Per.Att + Two.Season.Starter:Rush.Tds.Per.G +   
 Pass.Yds.Per.Att:Comp.Pct + TD.Int.Ratio:Rush.Yds.Per.Att +   
 TD.Int.Ratio:Rush.Tds.Per.G + Comp.Pct:Rush.Yds.Per.Att +   
 Comp.Pct:Rush.Tds.Per.G + Two.Season.Starter:Pass.Yds.Per.Att:Comp.Pct +   
 Two.Season.Starter:TD.Int.Ratio:Rush.Yds.Per.Att + Two.Season.Starter:TD.Int.Ratio:Rush.Tds.Per.G +   
 Two.Season.Starter:Comp.Pct:Rush.Yds.Per.Att + Two.Season.Starter:Comp.Pct:Rush.Tds.Per.G, data = qb\_data2)  
summary(reduced\_model2)

##   
## Call:  
## lm(formula = Avg.NFL.QBR ~ Two.Season.Starter + Pass.Yds.Per.Att +   
## TD.Int.Ratio + Comp.Pct + Rush.Yds.Per.Att + Rush.Tds.Per.G +   
## Two.Season.Starter:Pass.Yds.Per.Att + Two.Season.Starter:TD.Int.Ratio +   
## Two.Season.Starter:Comp.Pct + Two.Season.Starter:Rush.Yds.Per.Att +   
## Two.Season.Starter:Rush.Tds.Per.G + Pass.Yds.Per.Att:Comp.Pct +   
## TD.Int.Ratio:Rush.Yds.Per.Att + TD.Int.Ratio:Rush.Tds.Per.G +   
## Comp.Pct:Rush.Yds.Per.Att + Comp.Pct:Rush.Tds.Per.G + Two.Season.Starter:Pass.Yds.Per.Att:Comp.Pct +   
## Two.Season.Starter:TD.Int.Ratio:Rush.Yds.Per.Att + Two.Season.Starter:TD.Int.Ratio:Rush.Tds.Per.G +   
## Two.Season.Starter:Comp.Pct:Rush.Yds.Per.Att + Two.Season.Starter:Comp.Pct:Rush.Tds.Per.G,   
## data = qb\_data2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -39.857 -9.419 0.316 10.023 36.448   
##   
## Coefficients: (6 not defined because of singularities)  
## Estimate Std. Error  
## (Intercept) 2.455e+03 2.581e+03  
## Two.Season.Starter -2.272e+03 2.576e+03  
## Pass.Yds.Per.Att -2.489e+01 2.499e+01  
## TD.Int.Ratio 6.332e+01 7.015e+01  
## Comp.Pct -3.818e+01 4.048e+01  
## Rush.Yds.Per.Att -6.374e+01 7.665e+01  
## Rush.Tds.Per.G 5.213e+01 9.896e+01  
## Two.Season.Starter:Pass.Yds.Per.Att 4.629e+00 1.808e+01  
## Two.Season.Starter:TD.Int.Ratio -6.429e+01 7.008e+01  
## Two.Season.Starter:Comp.Pct 3.584e+01 4.038e+01  
## Two.Season.Starter:Rush.Yds.Per.Att 6.255e+01 7.617e+01  
## Two.Season.Starter:Rush.Tds.Per.G NA NA  
## Pass.Yds.Per.Att:Comp.Pct 3.292e-01 3.198e-01  
## TD.Int.Ratio:Rush.Yds.Per.Att 4.133e-01 6.403e-01  
## TD.Int.Ratio:Rush.Tds.Per.G -7.177e-01 6.768e+00  
## Comp.Pct:Rush.Yds.Per.Att 1.391e-02 1.408e-01  
## Comp.Pct:Rush.Tds.Per.G -5.841e-01 1.701e+00  
## Two.Season.Starter:Pass.Yds.Per.Att:Comp.Pct NA NA  
## Two.Season.Starter:TD.Int.Ratio:Rush.Yds.Per.Att NA NA  
## Two.Season.Starter:TD.Int.Ratio:Rush.Tds.Per.G NA NA  
## Two.Season.Starter:Comp.Pct:Rush.Yds.Per.Att NA NA  
## Two.Season.Starter:Comp.Pct:Rush.Tds.Per.G NA NA  
## t value Pr(>|t|)  
## (Intercept) 0.951 0.343  
## Two.Season.Starter -0.882 0.379  
## Pass.Yds.Per.Att -0.996 0.321  
## TD.Int.Ratio 0.903 0.368  
## Comp.Pct -0.943 0.347  
## Rush.Yds.Per.Att -0.832 0.407  
## Rush.Tds.Per.G 0.527 0.599  
## Two.Season.Starter:Pass.Yds.Per.Att 0.256 0.798  
## Two.Season.Starter:TD.Int.Ratio -0.917 0.360  
## Two.Season.Starter:Comp.Pct 0.888 0.376  
## Two.Season.Starter:Rush.Yds.Per.Att 0.821 0.413  
## Two.Season.Starter:Rush.Tds.Per.G NA NA  
## Pass.Yds.Per.Att:Comp.Pct 1.030 0.305  
## TD.Int.Ratio:Rush.Yds.Per.Att 0.645 0.520  
## TD.Int.Ratio:Rush.Tds.Per.G -0.106 0.916  
## Comp.Pct:Rush.Yds.Per.Att 0.099 0.921  
## Comp.Pct:Rush.Tds.Per.G -0.343 0.732  
## Two.Season.Starter:Pass.Yds.Per.Att:Comp.Pct NA NA  
## Two.Season.Starter:TD.Int.Ratio:Rush.Yds.Per.Att NA NA  
## Two.Season.Starter:TD.Int.Ratio:Rush.Tds.Per.G NA NA  
## Two.Season.Starter:Comp.Pct:Rush.Yds.Per.Att NA NA  
## Two.Season.Starter:Comp.Pct:Rush.Tds.Per.G NA NA  
##   
## Residual standard error: 16.23 on 164 degrees of freedom  
## Multiple R-squared: 0.1217, Adjusted R-squared: 0.04137   
## F-statistic: 1.515 on 15 and 164 DF, p-value: 0.105

Looking at this model, we see that we have a large amount of coefficients that are now not defined. Let’s cut down the model again.

reduced\_model2 = lm(Avg.NFL.QBR ~ Two.Season.Starter + Pass.Yds.Per.Att + TD.Int.Ratio +   
 Comp.Pct + Rush.Yds.Per.Att + Rush.Tds.Per.G + Two.Season.Starter:Pass.Yds.Per.Att +   
 Two.Season.Starter:TD.Int.Ratio + Two.Season.Starter:Comp.Pct +   
 Two.Season.Starter:Rush.Yds.Per.Att +   
 Pass.Yds.Per.Att:Comp.Pct + TD.Int.Ratio:Rush.Yds.Per.Att +   
 TD.Int.Ratio:Rush.Tds.Per.G + Comp.Pct:Rush.Yds.Per.Att +   
 Comp.Pct:Rush.Tds.Per.G, data = qb\_data2)  
summary(reduced\_model2)

##   
## Call:  
## lm(formula = Avg.NFL.QBR ~ Two.Season.Starter + Pass.Yds.Per.Att +   
## TD.Int.Ratio + Comp.Pct + Rush.Yds.Per.Att + Rush.Tds.Per.G +   
## Two.Season.Starter:Pass.Yds.Per.Att + Two.Season.Starter:TD.Int.Ratio +   
## Two.Season.Starter:Comp.Pct + Two.Season.Starter:Rush.Yds.Per.Att +   
## Pass.Yds.Per.Att:Comp.Pct + TD.Int.Ratio:Rush.Yds.Per.Att +   
## TD.Int.Ratio:Rush.Tds.Per.G + Comp.Pct:Rush.Yds.Per.Att +   
## Comp.Pct:Rush.Tds.Per.G, data = qb\_data2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -39.857 -9.419 0.316 10.023 36.448   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 2.455e+03 2.581e+03 0.951 0.343  
## Two.Season.Starter -2.272e+03 2.576e+03 -0.882 0.379  
## Pass.Yds.Per.Att -2.489e+01 2.499e+01 -0.996 0.321  
## TD.Int.Ratio 6.332e+01 7.015e+01 0.903 0.368  
## Comp.Pct -3.818e+01 4.048e+01 -0.943 0.347  
## Rush.Yds.Per.Att -6.374e+01 7.665e+01 -0.832 0.407  
## Rush.Tds.Per.G 5.213e+01 9.896e+01 0.527 0.599  
## Two.Season.Starter:Pass.Yds.Per.Att 4.629e+00 1.808e+01 0.256 0.798  
## Two.Season.Starter:TD.Int.Ratio -6.429e+01 7.008e+01 -0.917 0.360  
## Two.Season.Starter:Comp.Pct 3.584e+01 4.038e+01 0.888 0.376  
## Two.Season.Starter:Rush.Yds.Per.Att 6.255e+01 7.617e+01 0.821 0.413  
## Pass.Yds.Per.Att:Comp.Pct 3.292e-01 3.198e-01 1.030 0.305  
## TD.Int.Ratio:Rush.Yds.Per.Att 4.133e-01 6.403e-01 0.645 0.520  
## TD.Int.Ratio:Rush.Tds.Per.G -7.177e-01 6.768e+00 -0.106 0.916  
## Comp.Pct:Rush.Yds.Per.Att 1.391e-02 1.408e-01 0.099 0.921  
## Comp.Pct:Rush.Tds.Per.G -5.841e-01 1.701e+00 -0.343 0.732  
##   
## Residual standard error: 16.23 on 164 degrees of freedom  
## Multiple R-squared: 0.1217, Adjusted R-squared: 0.04137   
## F-statistic: 1.515 on 15 and 164 DF, p-value: 0.105

Interestingly enough, both the model’s significance and the adjusted \(R^2\) value decreased after removing the influential points. Additionally, all the three-way interections were not significant.

Next, we can take a look at the assumptions of the MLR model:

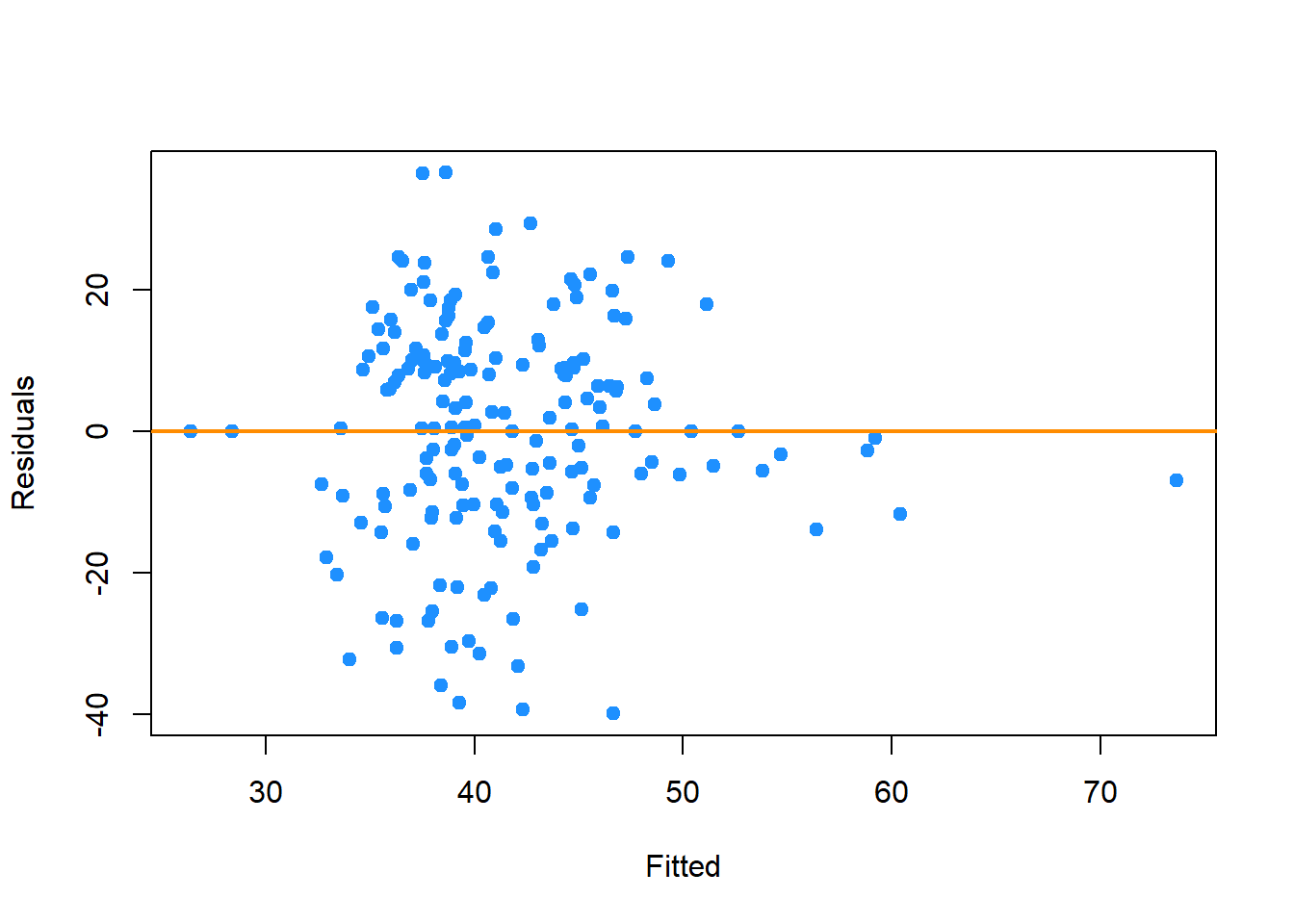
library(lmtest)

## Loading required package: zoo

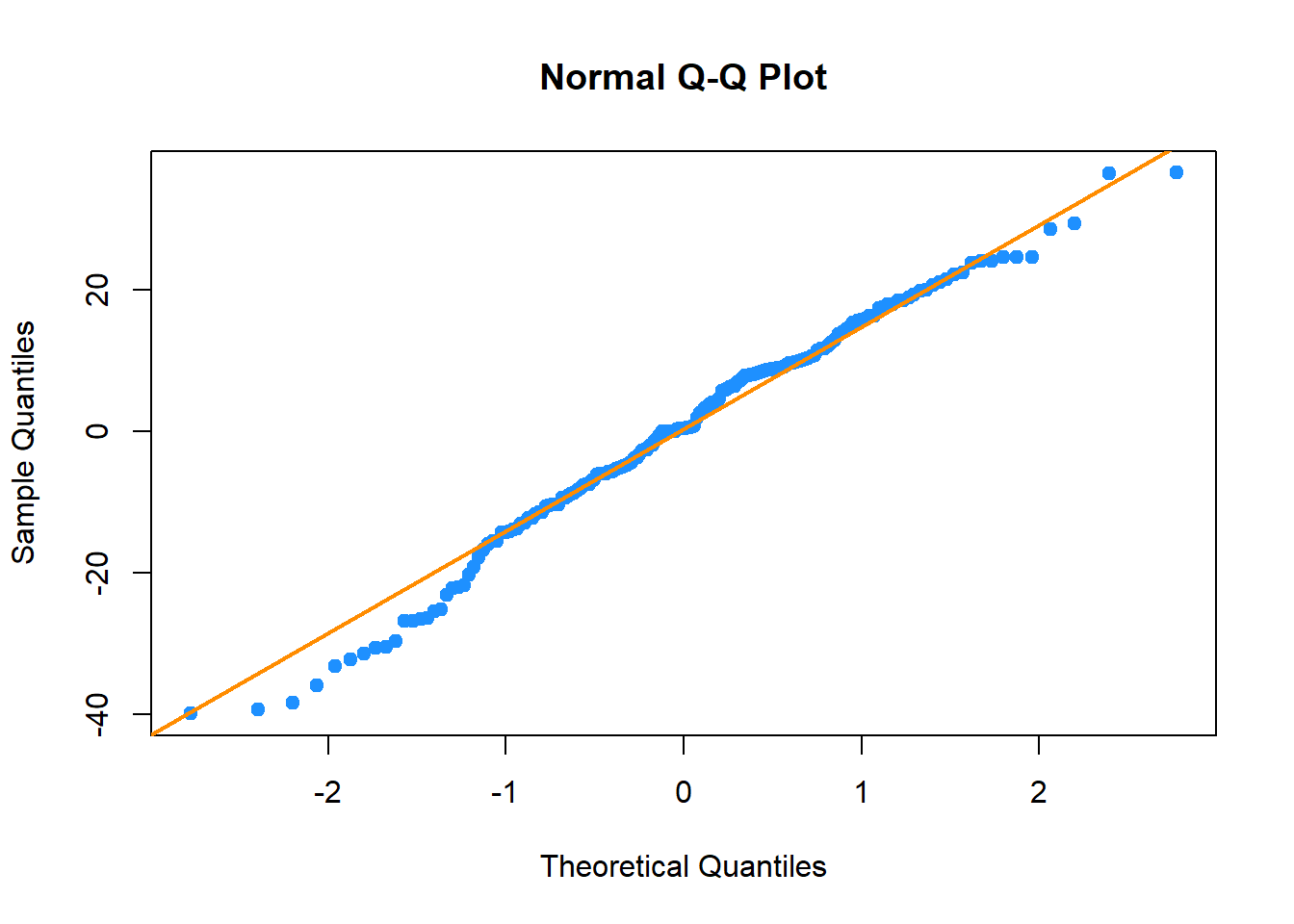
##   
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':  
##   
## as.Date, as.Date.numeric

plot\_fitted\_resid(reduced\_model2)



plot\_qq(reduced\_model2)



bptest(reduced\_model2)

##   
## studentized Breusch-Pagan test  
##   
## data: reduced\_model2  
## BP = 12.586, df = 15, p-value = 0.6342

shapiro.test(resid(reduced\_model2))

##   
## Shapiro-Wilk normality test  
##   
## data: resid(reduced\_model2)  
## W = 0.98557, p-value = 0.06108

Looking at the fitted vs. residual plot, there doesn’t appear to be any issues with autocorrelation between the residuals. From the QQ-Plot and the Shapiro-Wilks test, the errors appear to be normal. From the Breusch-Pagan test, there doesn’t appear to be issues with heteroskedasticity.

Is this a good model? Well, in terms of predictive power, it isn’t. We went through the correct statistical procedures and ultimately found that college statistics can’t really predict NFL success. This we can see by the fact that our adjusted \(R^2\) value is so low. But ultimately, our assumptions for model building did hold, and as good statisticians, we can’t simply make up results or come to a different conclusion that to what our data is telling us.

Discussion

So ultimately, we have a model with poor predictive power. But is it useful? Absolutely. What this model does, is verify long-held anecdotal knowledge of college quarterback evaluation. The crux of the matter is that projecting quarterback success from the college to professional level is extremely difficult. From the t-test for the signficiance of the Two.Season.Starter coefficient, it doesn’t have a statistically significant effect, even at the \(\alpha = 0.20\) level. In fact, none of our coefficients appear to do so either. Quarterback success at the NFL level is one of the hardest things to attempt to project in all of sports. Professional football is a sport with 100 years of history, and yet, we still have notable busts at the position in every draft. Successes like Tom Brady and Russell Wilson and busts like Jamarcus Russell and Ryan Leaf show that numbers aren’t everything, and that college experience isn’t necessarily a useful metric.

In the future, in order to improve on such a model, there are a couple of things that must be addressed. First, is a metric that takes into account the difficulty of opponents at the college level. Because of the wide range in challenge between good and bad teams in college football, it isn’t necessarily reflected in the raw statistics. Second, the rule changes with respect to passing yards at both the college and professional levels may have impacted these statistics with respect to time. This could be fixed by normalizing each year of each player’s data and then recalculating their overall statistics that way.

Appendix

Below is the full dataset:

View(qb\_data)