Bowling Green State University

NFL Quarterback Salary Analysis

Geri Dimas

Gaurav Sitaula

Zheng Zheng

Stat 5020

Dr. Arthur Yeh

December 6, 2016

Section 1:Executive Overview

In this study the objective will be to look at the salaries of quarterbacks in the National Football League (NFL) and possible variables that can help to predict what the salary could be in future years. The sample collected is the top 50 quarterbacks from the past three years, 2013 to 2015. The response variable, or the variable of which we hope to predict is salary. The variable salary was collected for 150 quarterbacks with the salary adjusted for inflation, to give the same value in dollars if these quarterbacks were all being paid in the current year of 2016[1]. Initially 13 predictor variables were collected, meaning these 13 variables were chosen due to their potential ability to help predict salary. The final and best model used to predict salary (transformed by raising to$\propto = 0.3$) for a NFL quarterback in a given season was experience (EXP) (transformed by LN), number of touchdowns (TD), number of Interceptions (INT) and the total number of teams played for (Num_Teams). In a more formal equation this model can be represented as:

$$Salary^{0.3} = 0.6932 + 0.5548LN(Exp) - 0.2344(NumTeams) + 0.00889(TD) + 0.02498(INT)$$

However, the slope parameters of this model do not have a direct interpretation because it is the transformed version of the reduced model, and does not contain the "influential" observations. Although we cannot interpret this model, we are able to interpret the reduced (Figure 6A) model in a practical and meaningful way.

Salary = 0.716 + 1.1092 Experience - 2.570 Num_Teams + 0.0997 TD + 0.1877 Int

In this example the intercept is 0.716, which means that this is the predicted average salary for a player when all the other parameters are 0. This could be interpreted as the starting salary of a first time "rookie" quarterback, which would be on average 0.716 X 1,000,000=$716,000. The slope associated with EXP is 1.1092, which indicates that for every one season increased in the experience of the player while holding Num_Teams, TD and INT constant, the salary increases on average by $1.1092 \times 1{,}000{,}000 = \$1{,}109{,}200$. The slope associated with Num_Teams is -2.570 which indicates that for every one number of teams changed by the player while holding EXP, TD and INT constant, the salary decreases on average by $2.570 \times 1{,}000{,}000 = \$2{,}570{,}000$. The slope associated with TD is 0.0997, which indicates that for every one increased in the passing touchdown of the player while holding EXP, Num_Teams and INT constant, the salary increases on average by $0.0997 \times 1{,}000{,}000 = \$99{,}700$. The slope associated with INT, is 0.1877, which indicates that for every one increased in the interception thrown by the player while holding EXP, Num_Teams and TD constant, the salary increases on average by $0.1877 \times 1{,}000{,}000 = \$187{,}700$.

After looking at many different models, this model appeared to be the best fit, and therefore the model used. One of the main reasons this model was used was due to the fact that it had the least amount of transformations needed in comparison to otherwise equally acceptable models, by trial and error. There existed models with more predictor variables and less predictor variables. The models with more predictor variables (greater than 4) did not offer any significant advancement for the model, so the simpler model was chosen. The models with less predictor variables (less than 4) did not come close in comparison to having a significant model and thus the larger model was chosen (with 4 predictor variables). Overall this model meets not only the requirements for a good model, but is also the simplest model possible based on our analysis.

---

1 http://data.bls.gov/cgi-bin/cpicalc.pl?cost1=1.80&year1=2013&year2=2016

Section 2:Data Collection

The data selected for this regression analysis was all obtained from the official National Football League website under the statistics page.[2] Each season (year) considered was looked up and then each individual quarterback (the top 50) for that year was recorded with the associated statistics. The original 13 variables gathered were; salary, age, experience (EXP), number of teams played for (Num_Teams), total games played (G), passing attempts (ATT), completed passes (COMP), percent of attempts completed (PCt), passing yards(PYards), rushing yards(Ryards), touchdowns(TD), interceptions(INT), sacks(SCK), and fumbles(FUM). This data was chosen because it is well known that the measured quality of a quarterback is often associated with the variables described above. Generally we assume that the better the overall statistics for a quarterback are the more they are likely to be paid. Since we were interested in being able to predict possible future salaries of the NFL quarterbacks, these variables seem to be appropriate and could make good predictors of salary.

The topic of quarterback salaries in the NFL is of interest to our group because when the project first started it was the beginning of football season, and by the time this semester and project are over it will be the end of football season, which means it is almost time for contract negotiations. Given the timing of these salary negotiations we thought it would be interesting to see if we would be able to get a good idea based on the end of 2016 football season what some quarterbacks will be paid. This intern would be a good indication as to how much money each team would have remaining for the rest of the players. This study's draw of interest does not end with our group, but could really be of interest to a large number of groups. The first would be quarterbacks themselves, being able to predict possible future salaries based off performance would help these quarterbacks know what they are "worth" as a player. In addition this would draw interest from the NFL teams coaches and those in charge of the players salaries, for similar reasons. If there were a model that could get a fairly decent prediction of future salaries based off of performance, the model could be used to negotiate contracts. The last group that might find this study interesting would be other players on the teams. The NFL has a current salary cap of 155.27[3] million, meaning if other players on the team know about how much the quarterback(s) will be making, it could give an idea to other players how much money is "left" for the remaining players on the team, and how much they might be able to themselves negotiate.

---

2 http://www.nfl.com/stats/categorystats?archive=true&conference=null&statisticPositionCategory=QUARTERBACK&season=2015&seasonType=REG&experience=&tabSeq=1&qualified=false&Submit=Go

3 https://www.nflpa.com/news/all-news/2016-adjusted-team-salary-caps

Section 3:Regression Modeling

After running a multiple linear regression analysis on this data, the final model decided upon was a reduced model made up of; salary (re-expressed to the power of ∝=0.3), verses EXP (to the LN), TD, INT and Num_Teams . Although this is the final model we decided upon, there were many steps taken to arrive at this final model.

The first thing we had to do was look at all of the variables for the full model, which summed to 14 total variables (1 response and 13 predictor variables) and then from there we could start with the variable selection process. The reason we started with variable selection was because we wanted to get rid of any multicollinearity we might have had between our predictor variables. This helped later when we started the rest of our analysis, because we no longer had to worry about the possibility of high correlation between two predictor variables. In order to do this, we looked at the correlation matrix between all the predictor variables (full model) to see which, if any, had strong correlation with one another. As seen in Figure 1A, a few variables seemed to have high correlation and thus indicated that we might have some issues with multicollinearity. Given the strong correlation, looking at the VIF (variance inflation factor) value allowed us to confirm this multicollinearity, and gave a better idea of how serious a problem it may be. Looking at the VIF's of all our predictor variables (Figure 2A) it become apparent which of our predictor variables were strongly correlated, and why. Looking first at the variables of age and experience, we had a strong correlation in our matrix, where both predictor variables had a high VIF. This would make sense because the older you are, the more experience you are likely to have, and so we decided to use experience, and drop the variable age. Next looking at the variables games played, passing attempts, passing completed, total passing yards and total number of touchdowns, it made sense that these had a strong correlation amongst themselves and against the total number of games played since the more games you play, the more opportunity you have for each of these. Thus it would make sense to drop the variable total games played. Also, since the number of passes attempted and passes completed would be high correlated with the ratio of the percentage of passing yards competed, it would make sense to include this simpler variable, that captures both of these in one variable (CPt). The last variable that appeared to be strongly correlated with the remaining eight variables (with a high VIF value) was the variable passing yards, so we decided to drop this variable as well. After dropping the predictor variables mentioned above, we re-run the VIF's (Figure 3A) and correlation matrix (Figure 4A), and all the remaining variables values were fairly low (close to 1). After checking the correlations and VIF's of the 13 original variables we were then left with 8 predictor variables that could be used in the rest of our variable selection process.

After we had successfully taken care of any possible multicollinearity between our variables, we could finish the variable selection process by taking the "all possible regressions model" approach, which included looking at four specific values R^2, ${R^2}_{adj}$, the estimate of error standard deviation and Mallow's $C_p$criterion (Figure 5A). After running the best subsets for the full model of salary verses EXP, Num_Teams, Pyards,CPt, Ryards, TD, INT, SCK and FUM, the best model was found (Figure 5A-1). Looking at all the criterion, and comparing all models, we could see that the best model included 4 predictor variables: EXP, Num_Teams, TD and INT. Although in this method the model appeared to be the best, there was one area of concern; the Mallows $C_p$value was below the desired amount of 5, which could indicate some instability in our model. Given this concern, it was decided to run another method of variable selection, the stepwise (Figure 6A), to compare and see if indeed this appeared to be the best model. After running the stepwise with an ∝= 0.15 to enter and ∝=0.15 to remove, it was confirmed that this was indeed considered to be the best model. The coefficient of determination (R^2 value) is 72.5%. It provides information about the strength of the linear regression and measures how close the data are to the fitted regression line. In our case, it indicates that 72.5% of the variability in the response variable (salary) that can be explained by the variabilities in predictor variables. Looking at the individual VIF and p-values of the variables and model, everything seemed to pass the hypothesis testing of p-value <0.05 where:

$$\mathrm{H}_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$$

$$\text{vs. } \mathrm{H}_a: \text{Not all } \beta_j\text{'s are 0.}$$

As we can see in Figure 6A the p-value $(0.0) < \alpha(0.05)$, so we rejected H_0 and accepted H_a. Hence, we had sufficient evidence at 5% level of significance to suggest that not all the slope parameters were equal to 0. After we had selected our best model, we wanted to check our assumptions to see which, if any of them were violated. If none had been violated, then we would be able to move on to comparing the full model verses the reduced model however, if any of the assumptions were violated, re-expressions would be needed for the model.

After we had selected our variables, we wanted to check the assumption and make sure that none of the four were violated. Starting with the model assumption we looked for a few things, the first of which (in this case) was the lack-of-fit. Although in our data we only had a few cases (two specifically) where we had replicates of x variables, we still needed to check and therefore, consider the lack-of-fit test. First looking at the residuals verses fitted values plot (Figure 1B) we saw a few things. The first was that there seemed to be an overall good distribution both below and above the line, where the points above the line seemed to be under fitting (negative residuals), and the points below were over fitted (positive residuals). The second thing we could take away from this plot was that there didn't seem to be an overall pattern (up and down) to the observations, however we did notice a slight right shaped megaphone, which we discuss in greater detail later (for the constant variance assumption). From this alone, it was not clear that the model assumption was violated and therefore it was not clear if this model could be modeled by a linear relationship. The second thing we had to consider in this case was the lack-of-fit test (Figure 2B), which was tested by the hypothesis:

$$Ho: y = \beta_0 + \beta 1_1 x_1 + \beta_2 x_2 + \cdots + \beta_K x_K + \varepsilon$$

$$\text{Vs. } Ha: y \neq \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_K x_K + \varepsilon$$

Looking at the p-value associated with this "Lack of fit" we found that the p-value was less than 0.05 (p-value=0.011 <∝(0.05)), which meant we had sufficient evidence at a 5% level of significant to suggest that the linear model assumption was violated. Next we wanted to see what other assumptions may have been violated (if any); looking next at the constant variance assumption we could see that this too might have been violated. Checking the constant variance assumption was done by looking at the residuals verses fitted plot (Figure3B), which we also used in looking at the model assumption. As mentioned above we saw a slight right megaphone shape, which was an indication that this assumption might have been violated, meaning we didn't see the residuals bounded by a "uniform band" as we would have liked. This right shaped megaphone indicated that the constant variance assumption was possibly violated and that the error variance increased as the number of observations increased. Since we had a possible violation of the constant variance assumption there were two things we wanted to do; the first was plot the residuals verses each of our predictor variables (Figure 4B) to see where perhaps this violation was coming from, and the second was using a simple hypothesis test to confirm this violation. Plotting the residuals verses each individual predictor variables we could see that the predictor variables of EXP, Num_Team and INT appeared to have a non-linear relationship, which meant that this megaphone shape we saw in the residuals might have been due to one of these. Next we tested the following hypothesis to see if the constant variance was violated (Figure5B):

$$H_0: Variance\ is\ constant$$

$$\text{vs.}H_a : Variance\ is\ not\ constant$$

After calculating the Q statistic of this model, we saw that the p-value= 0.9993 > $\alpha(0.05)$ indicating that we failed to reject H_0, hence, we had sufficient evidence at 5% level of significance to suggest that the constant variance assumption was not violated. The next assumption to check was the normality assumption, which was done by looking at the Anderson-Darling (used in this analysis) test and associated p-value for the normality plot of the residuals (Figure6B). Visually we could see that there were quit a few observations that deviated from our linear model line, and thus the normality assumption was most likely violated. To confirm this we tested the hypothesis:

$$H_0: F(x) is\ normally\ distributed$$

$$\text{vs.}H_a: F(x) is\ not\ normally\ distributed$$

Using the Anderson-darling testing statistic, we saw that our normality assumption was indeed violated since p-value= (0.01) $<\propto$ (0.05) so we rejected H_0 , hence we had sufficient evidence at 5% level of significant to conclude that the residuals were not normally distributed which meant that the normality assumption was violated. The last assumption to be tested was the independence assumption, which looked at the time order of the observations and checked if these observations were independent from each other. Looking at the residuals verses time order plot (Figure7B) we saw that there did not seem to be any particular order, therefore it appeared (by visual inspection) that the independence assumption was not violated, however confirmed this by using the Durbin-Watson test. The Durbin-Watson procedure tests that

$$H_0: \rho = 0$$

$$\text{Vs. } H_a: \rho > 0.$$

The d value in this case was 1.84910(Figure 7B), where the$[d_u, d_l]$ is [1.61,1.74], and therefore the d-value fell above the upper limit, which meant we did not reject H_0, since we did not have sufficient evidence to conclude that the observations were positively correlated , so the independence assumption was not violated. Although the constant variance and independent assumption was not violated, the linear model assumption and the normality assumption were, and therefore we wanted to consider some possible re-expressions, that would help to validate our violated assumptions.

Section 3:Regression Modeling

After some trial and error, we were able to find a (reasonable) re-expression, which did not violate any of the model assumptions. In the final model we were able to re-express our response variable of Y, Salary, to the power of 0.3, and take the natural log of the predictor variable EXP, which helped to remedy the previously violated assumptions. The first assumption we checked with this new model was the linear model assumption as previously done by looking at the residuals verses fitted plot and lack-of-fit test. We were able to determine that this assumption was indeed valid. First looking at the residuals verses fitted values plot (Figure 1C) we saw a few things. The first was that there seemed to be an overall good distribution both below and above the line, where the points above the line seemed to be under fitting (negative residuals), and the points below were over fitted (positive residuals). The second thing we could take away from this plot was that there didn't seem to be an overall pattern (up and down) to the observations. From this alone, it was not clear if the model assumption was valid and therefore it was not clear if this model followed a linear relationship, which is why we also had to check the lack-of-fit test (Figure 2C), which was tested by the hypothesis (Figure 1G). Looking at the p-value associated with this "Lack of fit" test we found that the p-value was greater than 0.05 (p-value=0.054 >0.5), which meant that we did not have sufficient evidence to suggest that the linear model assumption was violated. Although this value (0.054) barely meet the requirements (0.05), we still no longer had sufficient evidence to reject the null hypothesis, meaning the model assumption was adequate.

Next looking at the constant variance assumption we saw that this was too, still adequate. Checking the constant variance assumption was done by looking at the residuals verses fitted plot (Figure3C), which we also used in looking at the model assumption. Now we saw that the residuals were bounded by a "uniform band" as we wanted, indicating that this assumption was valid, however there did appear to be a slight left megaphone shape, which might have indicated a violation. Since we had a possible violation of the constant variance assumption we used a simple hypothesis test (Figure 2G) to either confirm or deny this possible violation (Figure4C). After calculating the Q statistic of this model we saw that p-value= 0.8365 $> \propto$ (0.05), indicating that the we did not have sufficient evidence to reject the null hypothesis and thus the constant variance assumption was not violated.

The next assumption to check was the normality assumption, which was done by looking at the Anderson-Darling (used in this analysis) test and associated p-value for the normality plot of the residuals (Figure5C). Visually we saw that there were still quit a few observations that deviated from our linear model line, however not as many observations as before, and thus we wanted to test the hypothesis(Figure 3G).Using the Anderson-darling testing statistic with a p-value >∝(0.05), we saw that our normality assumption was no longer violated since 0.734 $> \propto$ (0.05), and therefore there was not significant evidence to support that the null hypothesis was violated, which meant that the normality assumption was valid. The last assumption to be tested was the independence assumption, looking at the residuals verses time order plot (Figure6C) we saw that there did not appear to be any particular order, therefore (by visual inspection) the independence assumption was not violated. This can was confirmed by using the Durbin-Watson test, where the hypothesis tested was the same as previously mentioned(Figure 4G).The d value in this case was 2.03230(Figure 6C), where the$[d_u, d_l]$ is [1.61,1.74], and therefore the d-value fell above the upper limit, which meant we did not reject H_0. Hence, we did not have sufficient evidence to conclude that the observations were positively correlated and that meant that the independence assumption was not violated. After checking all four assumptions we saw that this re-expression indeed was successful in remedying the assumption violations we original saw in this model. Now that we had a reduced model that passed all four assumptions, we were able to compare the reduced verses the full model to see if there was sufficient evidence to support the claim that none of the remaining four variables were significantly different from 0.

Now that we had confirmed that our re-expressed model passed the assumptions we were able to move on to testing the reduced four (predictor) variable model verses the full eight (predictor) variable model. This was done by testing the F-statistic hypotheses:

$$\mathrm{H}_0: E(y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_L X_L$$

$$\mathrm{Vs.H}_a: E(y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_K X_K.$$

Even before calculating the F-statistic and testing this hypothesis we first looked at a few values from the full model regression analysis in comparison to that of the reduced model. Looking at the ${R^2}_{adj}, S, R^2$, and coefficient estimators between these two models, it became apparent that the full model didn't improve our model a whole lot, in some cases it actually made the model worse. First looking at the reduced model (Figure 1D), we saw that the ${R^2}_{adj} = 70.6\%, S = 0.316540, R^2 = 71.46\%$, where as in the full model (Figure 2D) we saw that the ${R^2}_{adj} = 70.23\%, S = 0.318893, R^2 = 71.83\%$. Not only did the full model have a lower ${R^2}_{adj}$, but the increase seen in the R^2 was not significant enough, the S increased from the reduced model to the full model and thus this suggested that the full model's additional four variables did not add any significant improvement to the model. This was confirmed when we calculated the F-statistic(Figure 3D) with a p-value $0.75998 > \propto(0.05)$, we did not reject the null hypothesis since there was not sufficient evidence at a 5% level of significant to suggest that at least one of the four slope parameters $\beta_5, \beta_6, \beta_7$ and β_8 was significantly different from 0. What this conclusion suggested in practical meaning was that with the model already containing LN(experience), Num_teams, TD and INT, it did not really add any significance to the linear model. So between the reduced model with LN(experience), Num_teams, TD and INT as the predictor variables, and the full model with additional PCt, FUM, RYards and SCK as predictor variables, the reduced model was just as good as the full model and was the one used since it is simpler (with 4 fewer predictor variables). After we had confirmed that our reduced model was sufficient, we looked at some possible ways to improve the model further, by considering outliers, high leverage and influential points.

Starting with the outliers for this model (Figure 1E), we could see 6 observations that were unusual with respect to the salary (y-variable), these were found using a threshold value t=2. In addition to looking at the studentized residuals we could also look at the deleted residuals to see if we missed any additional outliers. These "missed" outliers could be due to an outlier failing to produce a large studentized. These outliers and their associated sized residual can be seen in Figure 2E for the studentized residuals, most of these outliers were only slightly outside the threshold bound of -2 to 2. Looking next at the deleted residuals (Figure 3E), using a threshold value t= 1.976, we saw that we then had 7 outliers, which meant we had an additional outlier to consider. Although these were important to note, because outliers such as these could still affect the regression analysis, they do not usually affect the analysis to the extent that they actually "distort" the models' overall "picture". The next unusual observations we wanted to look at, and make note of were those that are considered to be high leverage, where these observations could potentially drastically distort the model's regression analysis and thus the overall "picture". In this model's case 12 observations (about 8% of all observations) were found that were considered to be high leverage points (Figure 4E), using a threshold value hii= 0.0667. Although both outliers and high leverage points are important to make note of, the unusual observations that we were most interested in were the ones that are considered to be influential points. Influential points are important because these observations are the ones that can totally "distort" the overall regression analysis because either with or without these values the regression can be drastically different.

First we looked at the DFITS, where observations were considered to be influential on a single fitted value, with a threshold t =0.36514837 we saw that we had 10 influential observations (Figure 5E). The last type of unusual observations we wanted to look for were influential points that are influential on all the fitted values. This was done by using Cook's distance where $F\ (.50, 5, 145)$=0.87 as seen in Figure 6E, none of our observations appeared to have influence on all the fitted values. After looking at all the diagnostic measures it appeared that there were 10 observations that had the potential to greatly influence our overall model. The way we measured if these influential points and exactly how much they impacted our model was done by comparing the model with these 10 observations (Figure 7E) and then without (Figure 8E) and see if there had been a great influence on the model as a whole in regards to parameter estimations and the overall prediction. After looking at the ${R^2}_{adj}, S, R^2$, and coefficient estimates, it appeared that we saw an overall improvement in our model when we removed these 10 observations where ${R^2}_{adj} = 79.57\%, S = 0.270559, R^2 = 80.15\%$, these are all improvements from the model with the influential points. The coefficient estimates also changed, which made sense, because without these influential points the final regression too would change to reflect the new data. After removing these observations we had to check all of our model assumption to make sure that this new model did not violate any of them.

First Looking at the new lack-of-fit value (Figure 8E) we saw that this value for p was 0.063>∝(0.05) with the hypothesis tested(Figure 1G), which meant that the model assumption was not violated. Next we looked at the residual verses order plot and we saw that overall there was a slight left megaphone shape (Figure 9E), however when we tested the Q statistics (Figure 10E), we had a p-value= 0.6856>∝(0.05) with the hypothesis(Figure 2G), so we failed to reject Ho, and the constant variance assumption was not violated. Next we tested the normality assumption by using the Anderson-Darling test and normality plot (Figure 11E) and we saw that the p-value 0.092>∝(0.05) with the hypothesis(Figure 3G), which meant that we failed to reject the null hypothesis and thus the normality assumption was not violated. Lastly, checking the independence assumption, plotting the residuals verse time order (Figure12E), there didn't appear to be any distinctive pattern, and using the Durbin- Watson statistic we calculated a value of 2.23340 which fell above the reference interval [1.61,1.74], meaning that we failed to reject Ho(Figure 4G), and thus the independence assumption was not violated.

Section 4:Summary and Conclusion

This model has many possible uses for real-life as mentioned earlier, quarterbacks, coaches, and other players on the team may all be interested in the real-life application of this model. The National Football League (NFL) has a league-wide salary cap, most recently in 2016 announced to be at $155.27 million. However, each individual team has their own individual adjusted team salaries caps based off of this league cap plus any additional carryover the team would have had from the previous year[4]. One of the biggest questions people have when looking at the salaries for the NFL is how each team decides to allocate their adjusted salary caps. There are many factors that can play into this, but one of the most prominent universally agreed upon factors is the player's position. In particular the position of quarterback receives the highest gross income for most teams. [5] Since most teams spend a large quantity first and foremost on their quarterbacks, being able to predict a quarterback's possible future salary would allow people to see how much of the remaining cap is readily available for the remaining players on a team. We believe this model has been able to do a decent job at predicting quarterbacks annual salaries in the NFL based on the four predictor variables; experience, touchdowns, interceptions and number of teams played for.

The accuracy, and overall effectiveness of this model for prediction can be seen in practice by looking at two different examples. The first example is of one observation from the current data (Figure 1F), where we can see that the observation 1 overall seems to have been fairly accurate in predicting the salary of Drew Brees. Starting with this observation 1 (Drew Brees) we can see that in this model the confidence interval says that we are 95% confident that the average salary for quarterbacks with 15 years Experience ($e^{2.70805}$), having played for 2 teams, 32 touchdowns and 11 interceptions is between $ 13,764,108 ($2.19596^{(\frac{1}{0.3})}$ x $1,000,000) and $17,909,247 ($2.37642^{(\frac{1}{0.3})}$x $1,000,000)The prediction interval says that we are 95% confident that the salary of a quarterback with 15 years Experience ($e^{2.70805}$), having played for 2 teams, 32 touchdowns and 11 interceptions is between $6,379,422 ($1.74355^{(\frac{1}{0.3})}x1,000,000) and $32,015,195 ($2.82883^{(\frac{1}{0.3})}$x $1,000,000). Given that the actual salary in this case (Drew Brees) was $24,400,000, this means that the prediction interval was successful in capturing the true salary, while the confidence interval was not successful in capturing it.

Next looking at an observation from outside the (collected) data we gathered data for Drew Brees' 2012 season. His salary for that year was $10,949,960 where he had 13 years experience, had played for 2 teams with 43 touchdowns and 19 interceptions. Looking at the prediction for his salary (Figure 2F), we can see that this time both the confidence interval (at a 95% level) and the prediction interval (at a 95% level) were both successful in capturing the true salary. This time the confidence interval says that we are 95% confident that the average salary for quarterbacks with 13 years Experience ($e^{2.56495}$), having played for 2 teams, 43 touchdowns and 19 interceptions is between $ 17,929,351 ($2.37722^{(\frac{1}{0.3})}$ x $1,000,000)and $25,162,247 ($2.63163^{(\frac{1}{0.3})}$x $1,000,000)The prediction interval says that we are 95% confident that the salary of a quarterback with 15 years Experience ($e^{2.70805}$), having played for 2 teams, 32 touchdowns and 11 interceptions is

[4] https://www.nflpa.com/news/all-news/2016-adjusted-team-salary-caps
[5] http://www.businessinsider.com/nfl-highest-paid-positions-2014-9

between \$9,333,984 ($1.95443^{(\frac{1}{0.3})}$x\$1,000,000) and \$41,345,588 ($3.05442^{(\frac{1}{0.3})}$x \$1,000,000).

References

CPI Inflation Calculator. (n.d.). Retrieved from http://data.bls.gov/cgi-bin/cpicalc.pl?cost1=1.80&year1=2013&year2=2016

News & Events. (n.d.). Retrieved from NFLPA: https://www.nflpa.com/news/all-news/2016-adjusted-team-salary-caps

NFL. (n.d.). Retrieved from ESPN: http://www.espn.com/nfl/

NFL Team Contracts and Payroll. (n.d.). Retrieved from Spotrac: http://www.spotrac.com/nfl/

Quarterback Contracts. (n.d.). Retrieved from Over The Cap: http://overthecap.com/position/quarterback/

Sports. (n.d.). Retrieved from Business Insider: http://www.businessinsider.com/nfl-highest-paid-positions-2014-9

statistics. (n.d.). Retrieved from NFL: http://www.nfl.com/stats/categorystats?archive=true&conference=null&statisticPositionCategory=QUARTERBACK&season=2015&seasonType=REG&experience=&tabSeq=1&qualified=false&Submit=Go

Appendix A : Multicolinearity and Variable Selection

	Salary	Age	Experience	Num_Teams	G
Age	0.492				
Experience	0.582	**0.961**			
Num_Teams	-0.246	0.500	0.451		
G	0.515	0.087	0.176	-0.287	
ATT	0.614	0.152	0.238	-0.282	**0.951**
Comp	0.646	0.189	0.278	-0.278	**0.941**
PCt	0.424	0.326	0.380	0.003	0.376
P_yards	0.637	0.185	0.273	-0.277	**0.945**
R_yards	-0.039	-0.333	-0.259	-0.208	0.435
TD	0.614	0.234	0.311	-0.249	**0.860**
Int	0.440	0.014	0.073	-0.274	0.729
Sck	0.325	-0.094	-0.006	-0.303	0.836
FUM	0.269	-0.068	-0.013	-0.194	0.697

	ATT	Comp	PCt	P_yards	R_yards
Comp	**0.993**				
PCt	0.384	0.464			
P_yards	**0.986**	**0.990**	0.451		
R_yards	0.281	0.252	0.000	0.278	
TD	**0.900**	**0.916**	0.472	**0.937**	0.250
Int	0.781	0.751	0.147	0.740	0.162
Sck	0.787	0.767	0.264	0.757	0.569
FUM	0.676	0.654	0.213	0.664	0.410

	TD	Int	Sck
Int	0.584		
Sck	0.625	0.631	
FUM	0.598	0.534	0.691

Figure 1A

Predictor	Coef	SE Coef	T	P	VIF
Constant	7.268	8.595	0.85	0.399	
Age	-0.2436	0.2520	-0.97	0.335	**16.073**
Num_Teams	-2.3997	0.2884	-8.32	0.000	1.816
Experience	1.2475	0.2474	5.04	0.000	15.995
G	-0.6584	0.2395	-2.75	0.007	**17.503**
ATT	-0.00126	0.02334	-0.05	0.957	**273.591**
Comp	0.03633	0.03763	0.97	0.336	**301.480**
PCt	0.0082	0.1086	0.08	0.940	3.298
P_yards	0.001718	0.001765	0.97	0.332	**89.198**
R_yards	0.004399	0.002618	1.68	0.095	2.455
TD	-0.13226	0.08532	-1.55	0.123	12.696
Int	0.05948	0.09720	0.61	0.542	3.500
Sck	-0.08756	0.04531	-1.93	0.055	5.360
FUM	0.0472	0.1282	0.37	0.713	2.312

Figure 2A

Appendix A : Multicolinearity and Variable Selection

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	-4.37	4.53	-0.96	0.337	
Experience	1.0792	0.0935	11.54	0.000	2.04
Num_Teams	-2.580	0.293	-8.81	0.000	1.67
PCt	0.0910	0.0767	1.19	0.237	1.47
R_yards	0.00018	0.00240	0.08	0.940	1.84
TD	0.0919	0.0435	2.11	0.036	2.94
Int	0.2253	0.0806	2.80	0.006	2.14
Sck	-0.0332	0.0390	-0.85	0.396	3.53
FUM	0.050	0.133	0.38	0.707	2.23

Figure 3A

	Salary	Experience	Num_Teams	PCt	R_yards	TD
Experience	0.582					
Num_Teams	-0.246	0.451				
PCt	0.424	0.380	0.003			
R_yards	-0.039	-0.259	-0.208	0.000		
TD	0.614	0.311	-0.249	0.472	0.250	
Int	0.440	0.073	-0.274	0.147	0.162	0.584
Sck	0.325	-0.006	-0.303	0.264	0.569	0.625
FUM	0.269	-0.013	-0.194	0.213	0.410	0.598

	Int	Sck
Sck	0.631	
FUM	0.534	0.691

Figure 4A

Appendix A : Multicolinearity and Variable Selection

Response is Salary

Vars	R-Sq	R-Sq (adj)	R-Sq (pred)	Mallows Cp	S	Experience	Num_Teams	PCts	R_yards	TD	Int	Sck	FUM
1	37.7	37.3	36.1	178.6	5.2725					X			
1	33.9	33.4	31.9	198.6	5.4329	X							
2	66.4	65.9	64.7	31.3	3.8882	X	X						
2	54.6	54.0	52.6	92.4	4.5154	X				X			
3	71.1	70.5	69.4	8.5	3.6153	X	X			X			
3	71.0	70.4	69.0	9.1	3.6224	X	X				X		
4	72.5	71.8	70.3	3.1	3.5371	X	X			X	X		
4	71.8	71.1	69.7	6.8	3.5824	X	X	X			X		
5	72.8	71.8	70.3	3.9	3.5345	X	X	X		X	X		
5	72.6	71.7	70.0	4.5	3.5423	X	X			X	X	X	
6	72.9	71.8	70.0	5.2	3.5373	X	X	X		X	X	X	
6	72.8	71.7	70.0	5.7	3.5447	X	X	X	X	X	X		
7	72.9	71.6	69.5	7.0	3.5479	X	X	X		X	X	X	X
7	72.9	71.6	69.6	7.1	3.5497	X	X	X	X	X	X	X	
8	72.9	71.4	69.1	9.0	3.5604	X	X	X	X	X	X	X	X

Figure 5A

Response is Salary

Vars	R-Sq	R-Sq (adj)	R-Sq (pred)	Mallows Cp	S	Experience	Num_Teams	PCts	R_yards	TD	Int	Sck	FUM
1	37.7	37.3	36.1	178.6	5.2725					X			
1	33.9	33.4	31.9	198.6	5.4329	X							
2	66.4	65.9	64.7	31.3	3.8882	X	X						
2	54.6	54.0	52.6	92.4	4.5154	X				X			
3	71.1	70.5	69.4	8.5	3.6153	X	X			X			
3	71.0	70.4	69.0	9.1	3.6224	X	X				X		
4	72.5	71.8	70.3	3.1	3.5371	X	X			X	X		
4	71.8	71.1	69.7	6.8	3.5824	X	X	X			X		
5	72.8	71.8	70.3	3.9	3.5345	X	X	X		X	X		
5	72.6	71.7	70.0	4.5	3.5423	X	X			X	X	X	
6	72.9	71.8	70.0	5.2	3.5373	X	X	X		X	X	X	
6	72.8	71.7	70.0	5.7	3.5447	X	X	X	X	X	X		
7	72.9	71.6	69.5	7.0	3.5479	X	X	X		X	X	X	X
7	72.9	71.6	69.6	7.1	3.5497	X	X	X	X	X	X	X	
8	72.9	71.4	69.1	9.0	3.5604	X	X	X	X	X	X	X	X

Figure 5A-1

Appendix A : Multicolinearity and Variable Selection

Stepwise Selection of Terms

α to enter = 0.15, α to remove = 0.15

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	4	4791.84	1197.96	95.75	0.000
Experience	1	2190.38	2190.38	175.07	0.000
Num_Teams	1	1013.31	1013.31	80.99	0.000
TD	1	101.72	101.72	8.13	0.005
Int	1	94.19	94.19	7.53	0.007
Error	145	1814.11	12.51		
Total	149	6605.95			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
3.53710	72.54%	71.78%	70.34%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	0.716	0.836	0.86	0.393	
Experience	1.1092	0.0838	13.23	0.000	1.66
Num_Teams	-2.570	0.286	-9.00	0.000	1.61
TD	0.0997	0.0350	2.85	0.005	1.92
Int	0.1877	0.0684	2.74	0.007	1.57

Regression Equation

Salary = 0.716 + 1.1092 Experience - 2.570 Num_Teams + 0.0997 TD + 0.1877 Int

Figure 6A

Appendix B : Testing Assumption for selected "best" Reduced Model

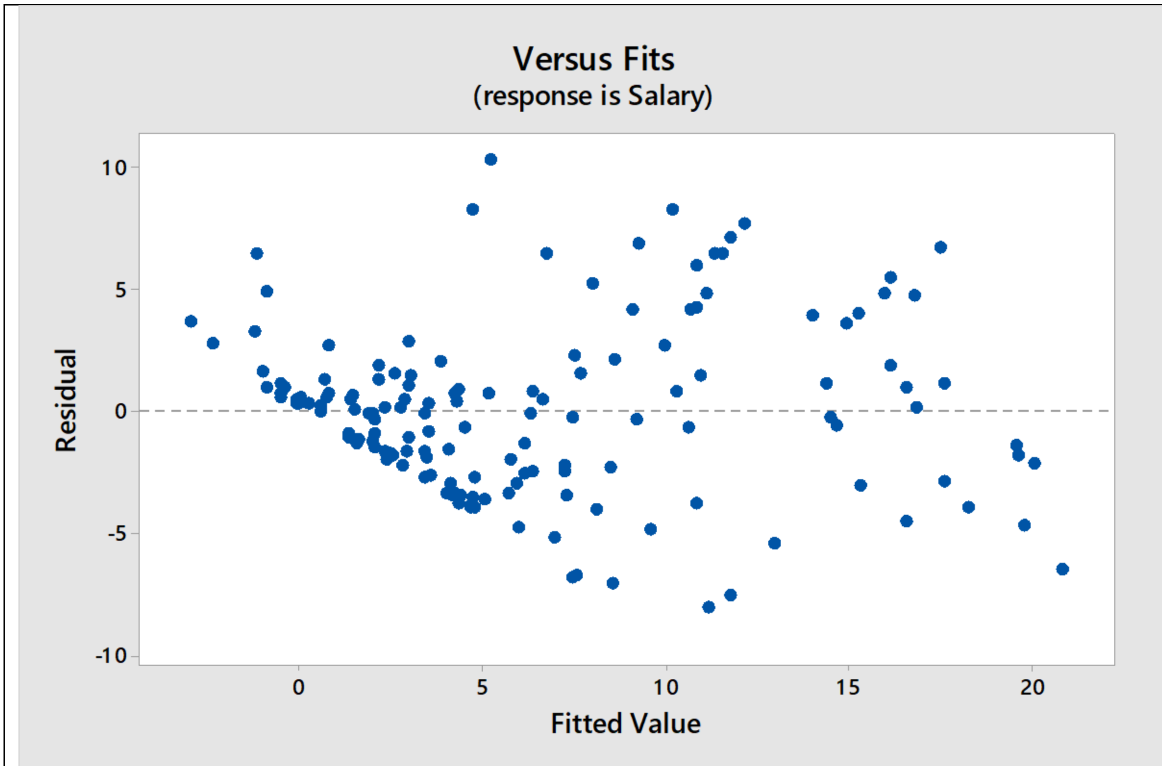

Figure 1B

Appendix B : Testing Assumption for selected “best” Reduced Model

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	4	4791.84	1197.96	95.75	0.000
Experience	1	2190.38	2190.38	175.07	0.000
Num_Teams	1	1013.31	1013.31	80.99	0.000
TD	1	101.72	101.72	8.13	0.005
Int	1	94.19	94.19	7.53	0.007
Error	145	1814.11	12.51		
Lack-of-Fit	144	1814.11	12.60	5142.02	0.011
Pure Error	1	0.00	0.00		
Total	149	6605.95			

Figure 2B

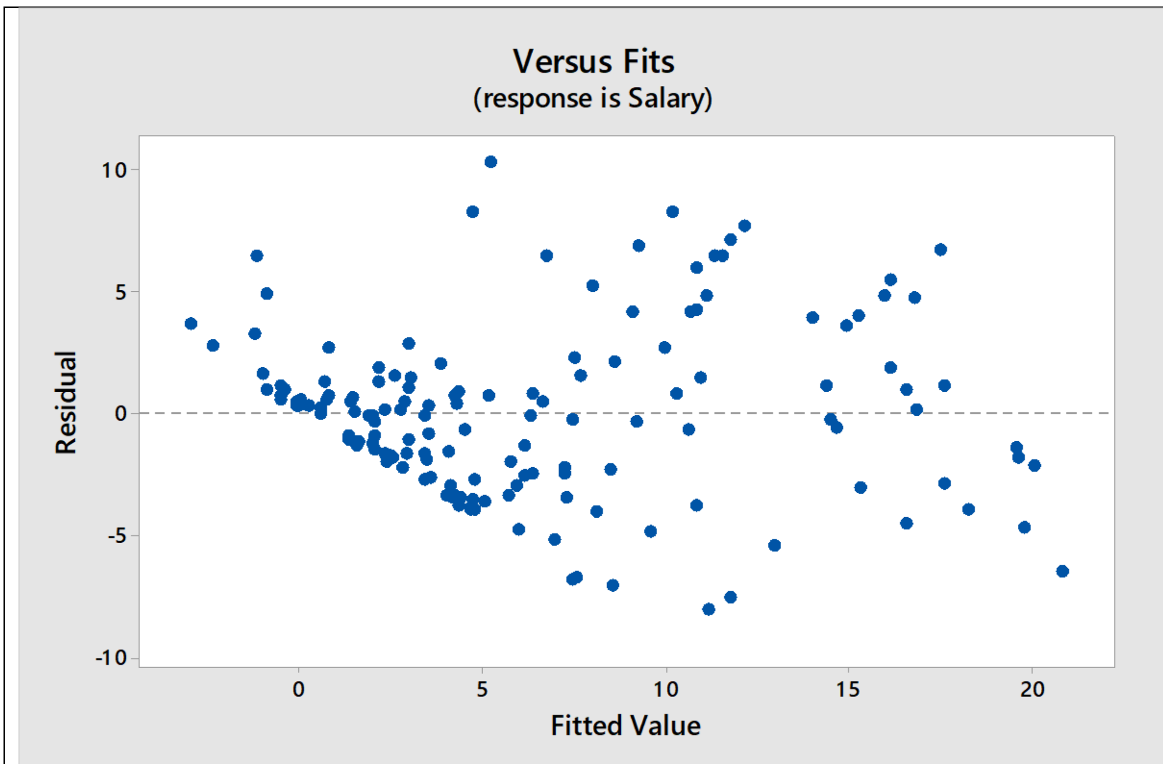

Figure 3B

Appendix B : Testing Assumption for selected “best” Reduced Model

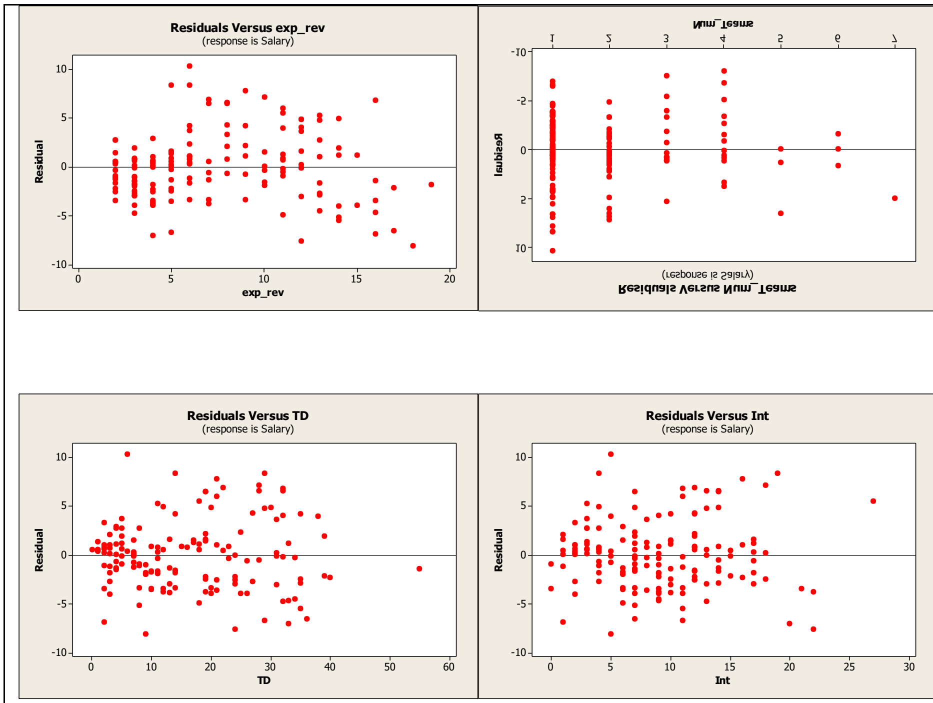

Figure 4B

Appendix B : Testing Assumption for selected “best” Reduced Model

$$h = \frac{\sum_{i=1}^{n} i \times \hat{e}_i^2}{\sum_{i=1}^{n} \hat{e}_i^2} = 59.4476$$

$$Q = \left(\frac{6n}{n^2 - 1}\right)^{\frac{1}{2}} * \left(h - \frac{n+1}{2}\right) = 0.200004445 * -16.0524 = -3.210551353$$

p-value= P(Z>Q) = 0.9993 > 0.05

Figure 5B

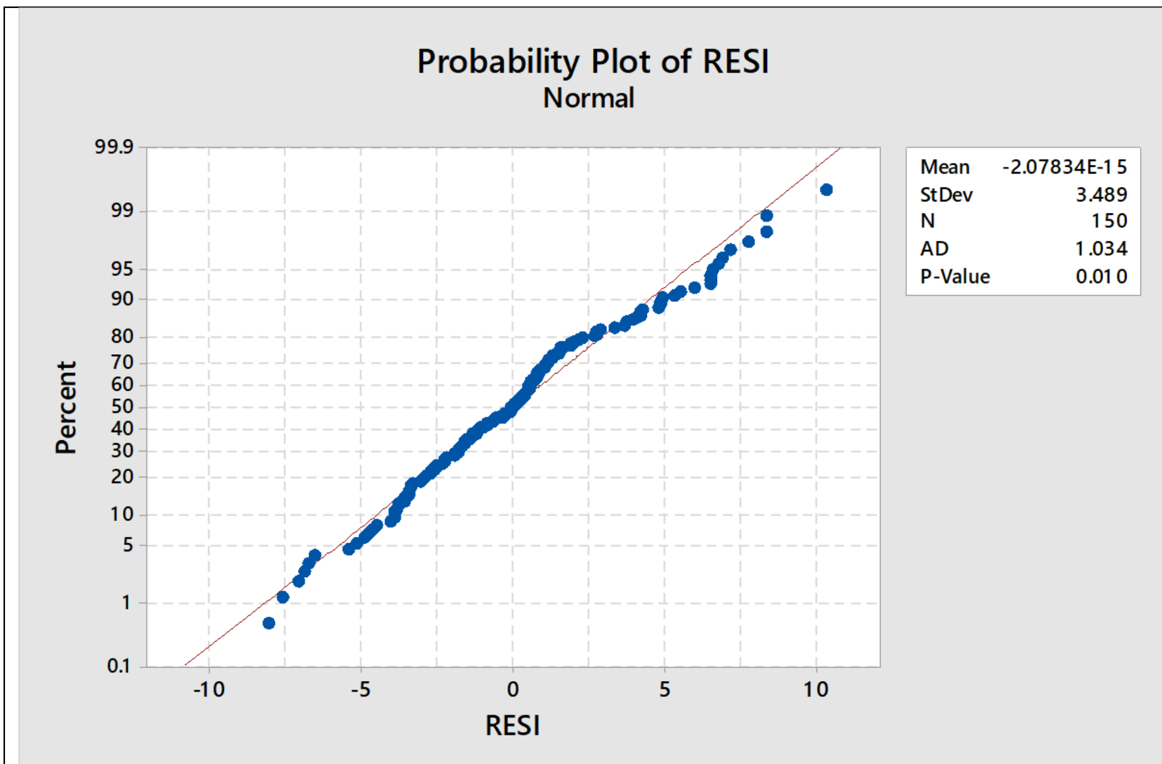

Figure 6B

Appendix B : Testing Assumption for selected "best" Reduced Model

```
Durbin-Watson Statistic

Durbin-Watson Statistic =  1.84910
```

N=150, k=4 [1.61,1.74]

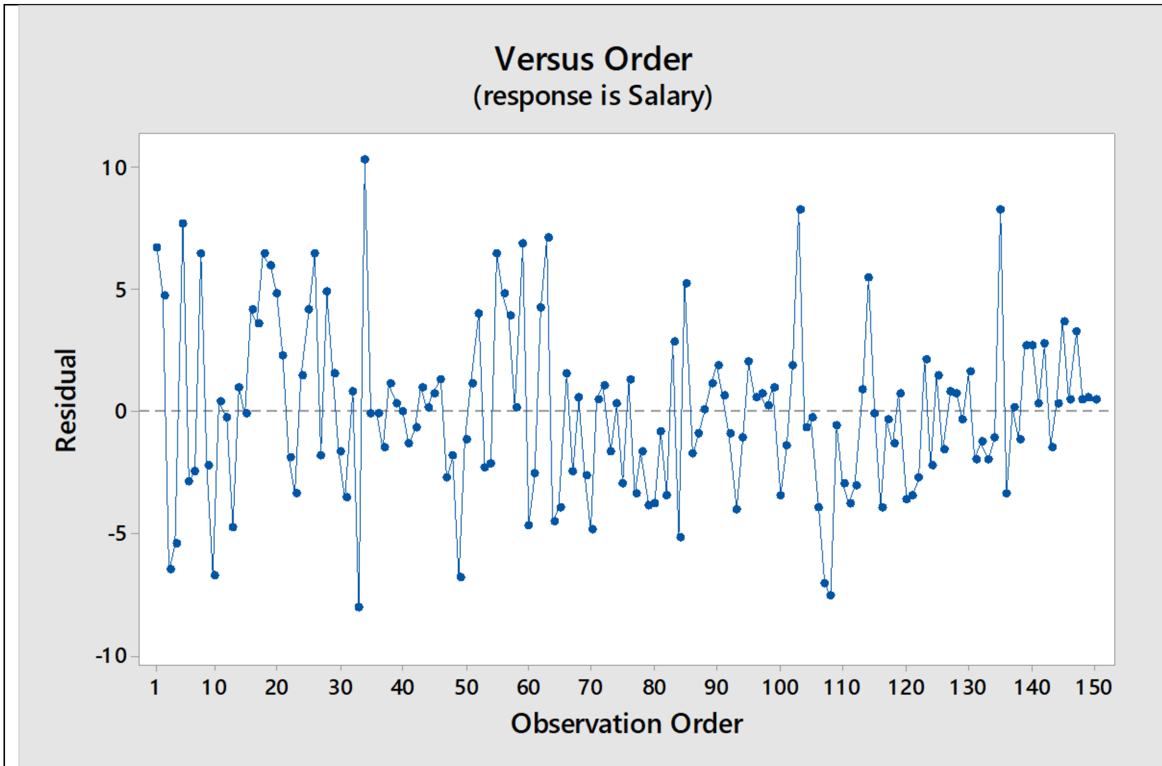

Figure 7B

Appendix C : Testing Assumption for Selected "Best" Re-expressed Model

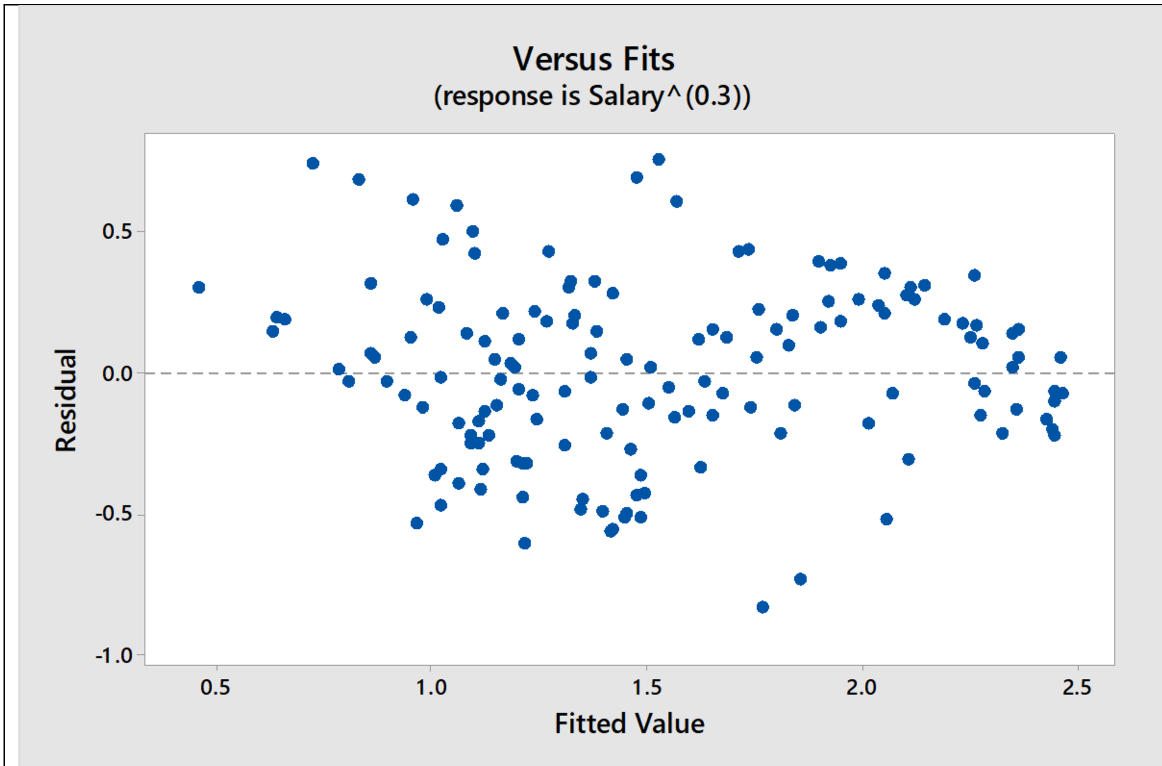

Figure 1C

Appendix C : Testing Assumption for Selected “Best” Re-expressed Model

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	4	36.3723	9.0931	90.75	0.000
Ln(Experience)	1	15.5557	15.5557	155.25	0.000
Num_Teams	1	6.1806	6.1806	61.68	0.000
TD	1	1.0151	1.0151	10.13	0.002
Int	1	1.2043	1.2043	12.02	0.001
Error	145	14.5286	0.1002		
Lack-of-Fit	144	14.5282	0.1009	215.70	0.054
Pure Error	1	0.0005	0.0005		
Total	149	50.9009			

Figure 2C

Appendix C : Testing Assumption for Selected "Best" Re-expressed Model

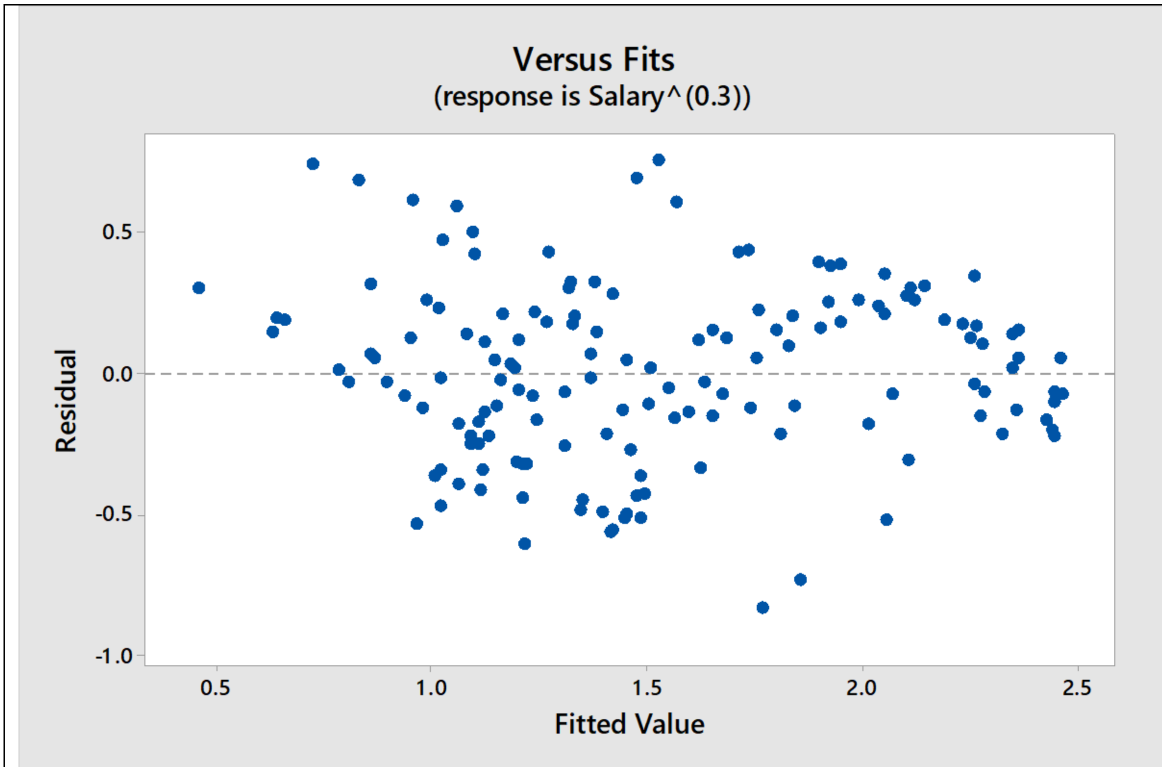

Figure 3C

$$h = \frac{\sum_{i=1}^{n}(n-i+1)\times \hat{e}_i^2}{\sum_{i=1}^{n}\hat{e}_i^2} = 70.5989$$

$$Q = \left(\frac{6n}{n^2-1}\right)^{\frac{1}{2}} * \left(h - \frac{n+1}{2}\right) = 0.200004445 * -5.4011 = -1.080244008$$

p-value= P(Z>Q) = 0.8599 > 0.05

Figure 4C

Appendix D : Testing Reduced Vs. Full Model

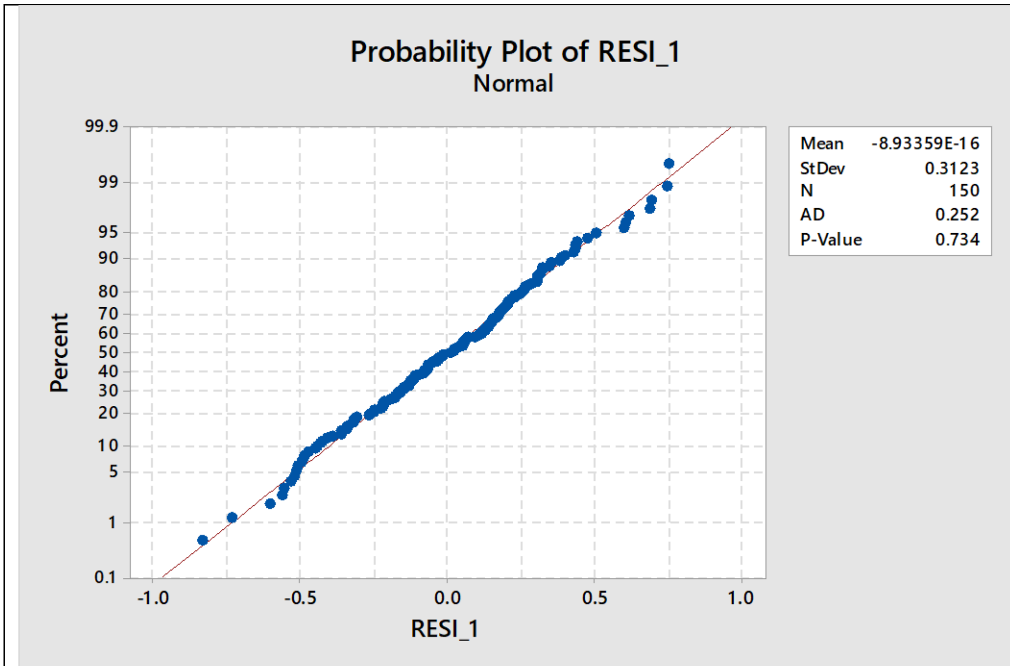

Figure 5C

```
Durbin-Watson Statistic =   2.03230
```

N=143, k=4 [1.61,1.74]

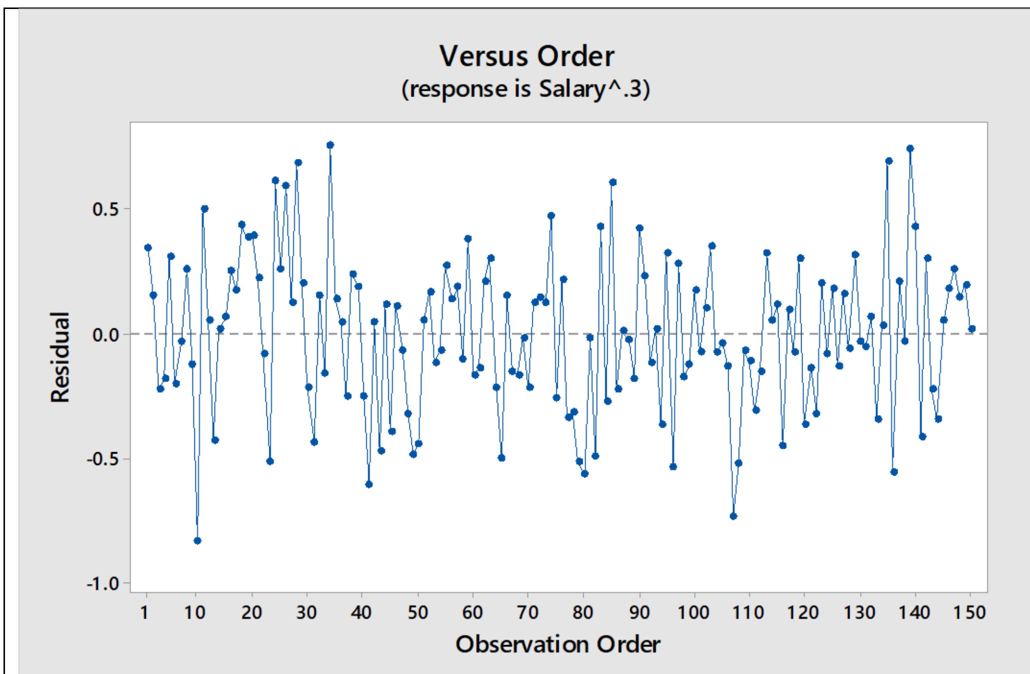

Figure 6C

Appendix D : Testing Reduced Vs. Full Model

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	8	36.5622	4.5703	44.94	0.000
Ln(Experience)	1	13.0625	13.0625	128.45	0.000
Num_Teams	1	5.9817	5.9817	58.82	0.000
PCt	1	0.0764	0.0764	0.75	0.388
R_yards	1	0.0283	0.0283	0.28	0.599
TD	1	0.3501	0.3501	3.44	0.066
Int	1	0.8360	0.8360	8.22	0.005
Sck	1	0.0087	0.0087	0.09	0.770
FUM	1	0.0052	0.0052	0.05	0.821
Error	141	14.3387	0.1017		
Total	149	50.9009			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.318893	71.83%	70.23%	68.10%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	0.377	0.405	0.93	0.354	
Ln(Experience)	0.5048	0.0445	11.33	0.000	2.02
Num_Teams	-0.2026	0.0264	-7.67	0.000	1.69
PCt	0.00599	0.00692	0.87	0.388	1.49
R_yards	0.000112	0.000212	0.53	0.599	1.79
TD	0.00716	0.00386	1.86	0.066	2.89
Int	0.02069	0.00722	2.87	0.005	2.14
Sck	0.00102	0.00349	0.29	0.770	3.53
FUM	0.0027	0.0120	0.23	0.821	2.25

Regression Equation

Salary^(0.3) = 0.377 + 0.5048 Ln(Experience) - 0.2026 Num_Teams + 0.00599 PCt
+ 0.000112 R_yards + 0.00716 TD + 0.02069 Int + 0.00102 Sck + 0.0027 FUM

Figure 1D

$$F = \frac{(SSR_F - SSR_R)/(K-L)}{SSE_F/(n-K-1)}$$

n = 150, K=8, L=4

$$F = \frac{(36.5622 - 36.3723)/(8-4)}{14.3387/(150-8-1)}$$

F=.466846715

Cumulative Distribution Function

```
F distribution with 4 DF in numerator and 141 DF in denominator

      x  P( X ≤ x )
0.46685    0.240023
```

$$P - value = P\left(F_{4,141} \geq .240023\right) = 1 - .240023 = .759977$$

$$.759977 > 0.05$$

Figure 3D

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	0.7557	0.0768	9.84	0.000	
Ln(Experience)	0.5003	0.0402	12.46	0.000	1.66
Num_Teams	-0.2022	0.0257	-7.85	0.000	1.63
TD	0.00990	0.00311	3.18	0.002	1.90
Int	0.02122	0.00612	3.47	0.001	1.57

Regression Equation

```
Salary^(0.3) = 0.7557 + 0.5003 Ln(Experience) - 0.2022 Num_Teams + 0.00990 TD + 0.02122 Int
```

Figure 2D

Appendix E : Outlier, High Leverage and Influential Points

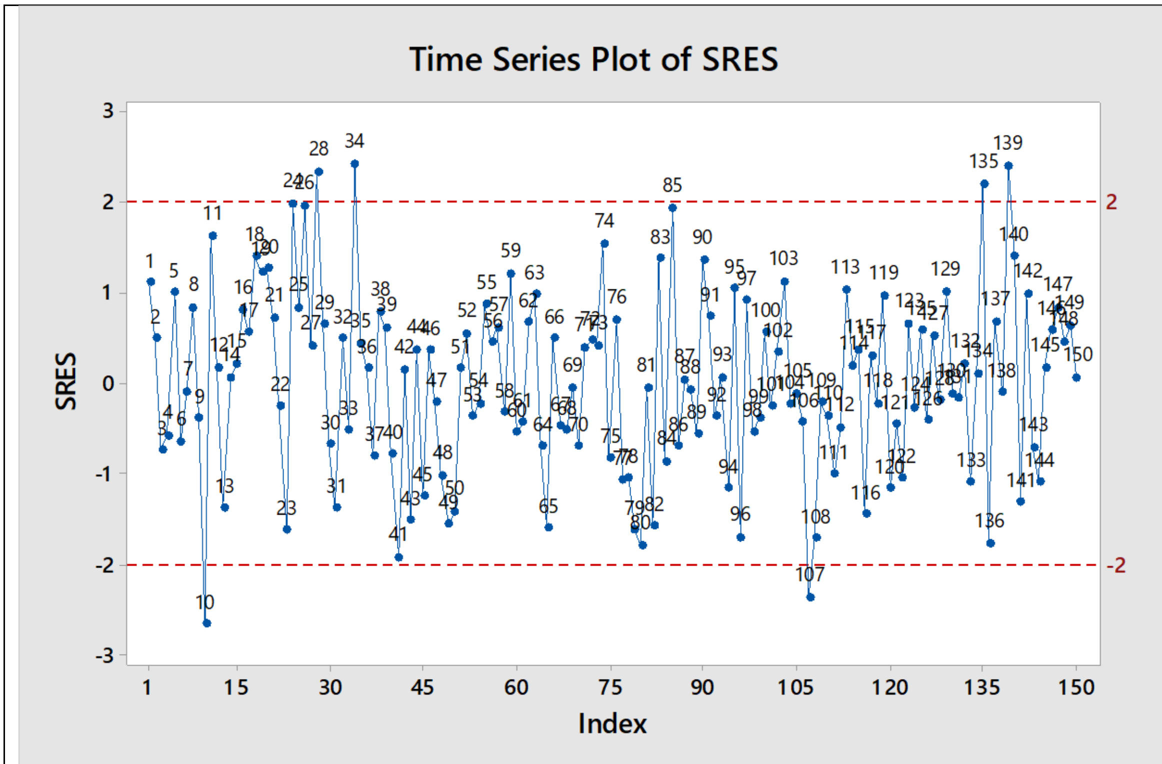

Figure 1E- 6 outliers

Appendix E : Outlier, High Leverage and Influential Points

Fits and Diagnostics for Unusual Observations

Obs	Salary^(0.3)	Fit	Resid	Std Resid		
10	0.9317	1.7676	-0.8359	-2.67	R	
15	1.4320	1.3673	0.0647	0.22		X
28	1.5146	0.8272	0.6874	2.33	R	X
34	2.2778	1.5243	0.7536	2.41	R	
78	0.8788	1.1917	-0.3129	-1.05		X
101	2.3856	2.4628	-0.0772	-0.26		X
107	1.1225	1.8542	-0.7317	-2.36	R	
114	2.5142	2.4568	0.0574	0.20		X
135	2.1616	1.4706	0.6911	2.21	R	
139	1.4624	0.7176	0.7449	2.40	R	

R Large residual

X Unusual X

Figure 2E

Appendix E : Outlier, High Leverage and Influential Points

$$t_{145,2.5\%} = 1.976$$

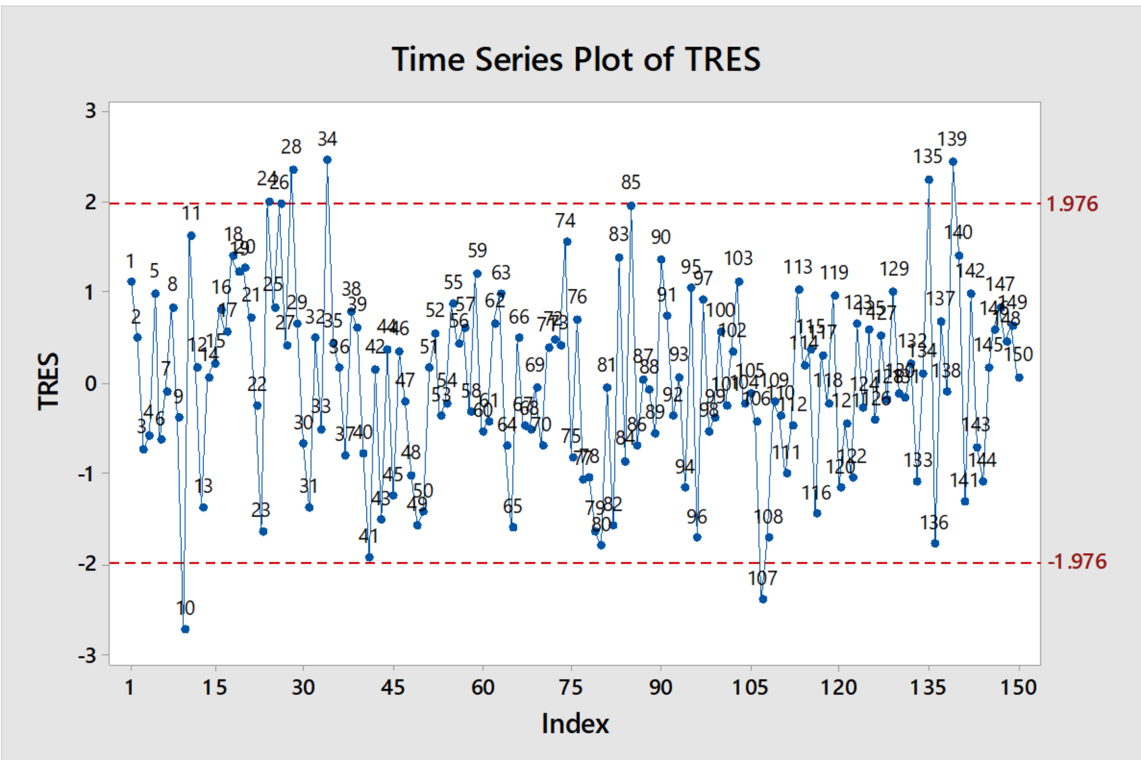

Figure 3E

Appendix E : Outlier, High Leverage and Influential Points

0.06666666667>Hii

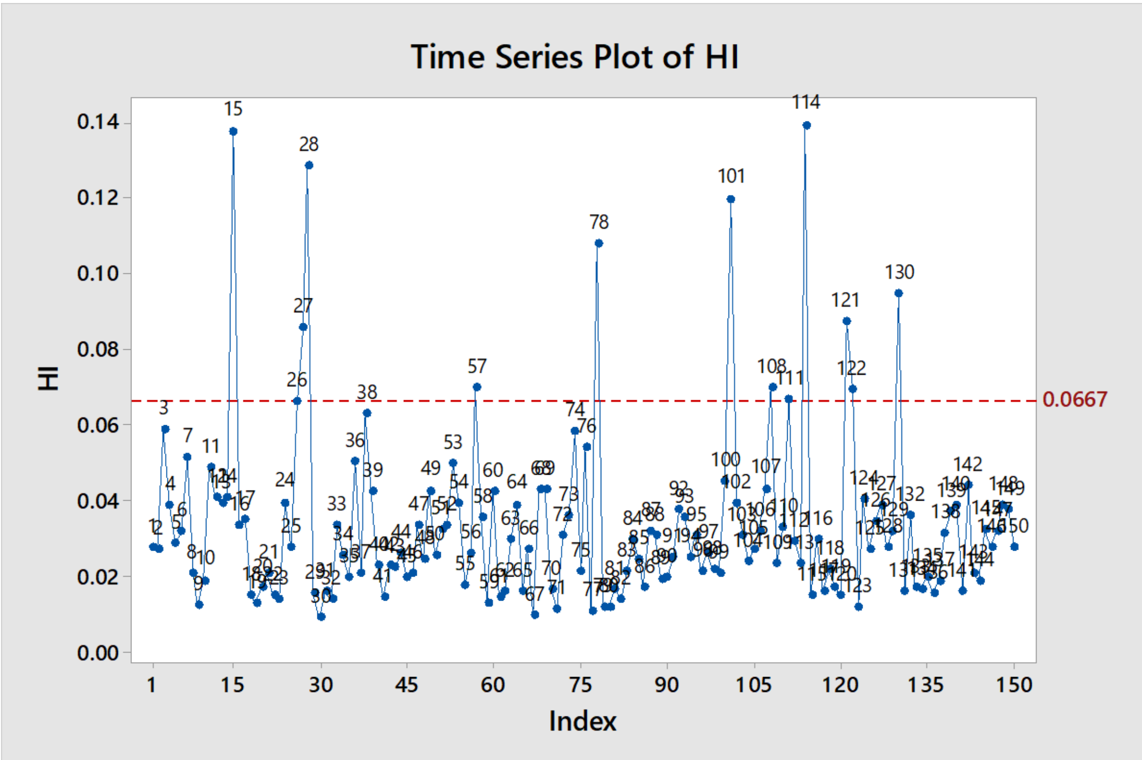

Figure 4E: 12 High leverage ,8% of all observations

Appendix E : Outlier, High Leverage and Influential Points

Threshold = .36514837

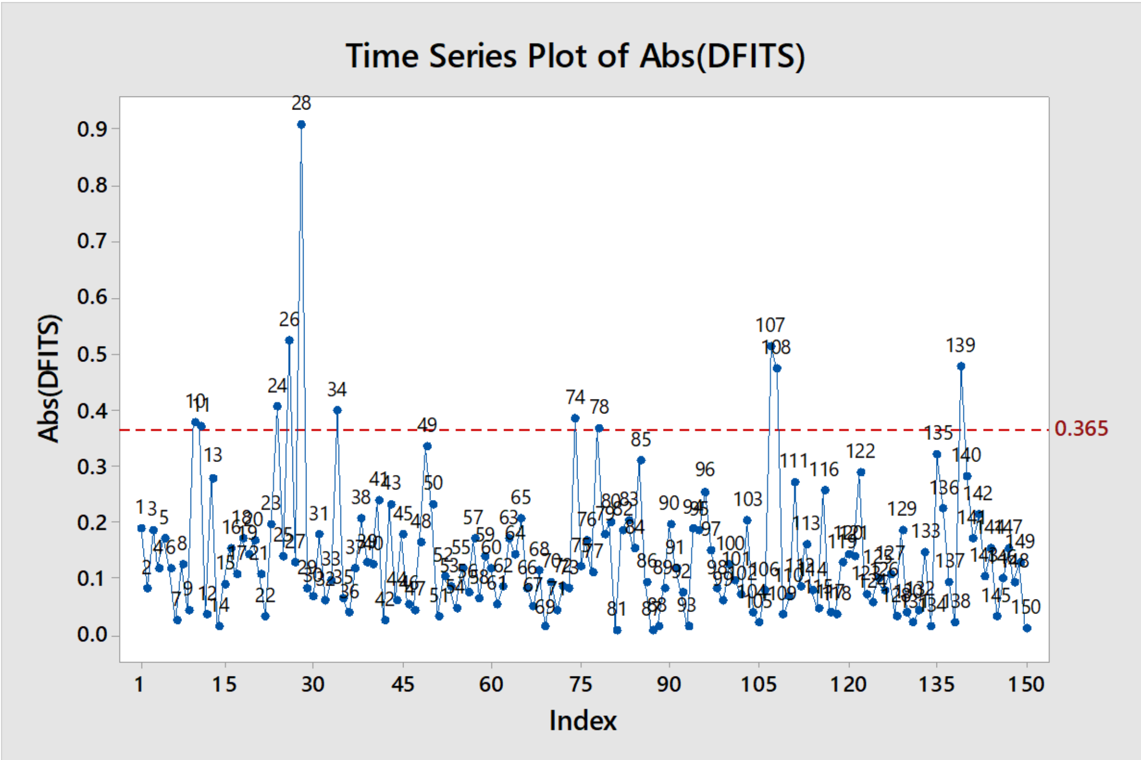

Figure 5E- DFITS 10 influential

Appendix E : Outlier, High Leverage and Influential Points

Threshold= 0.87

$F(.50,5,145)= 0.87$

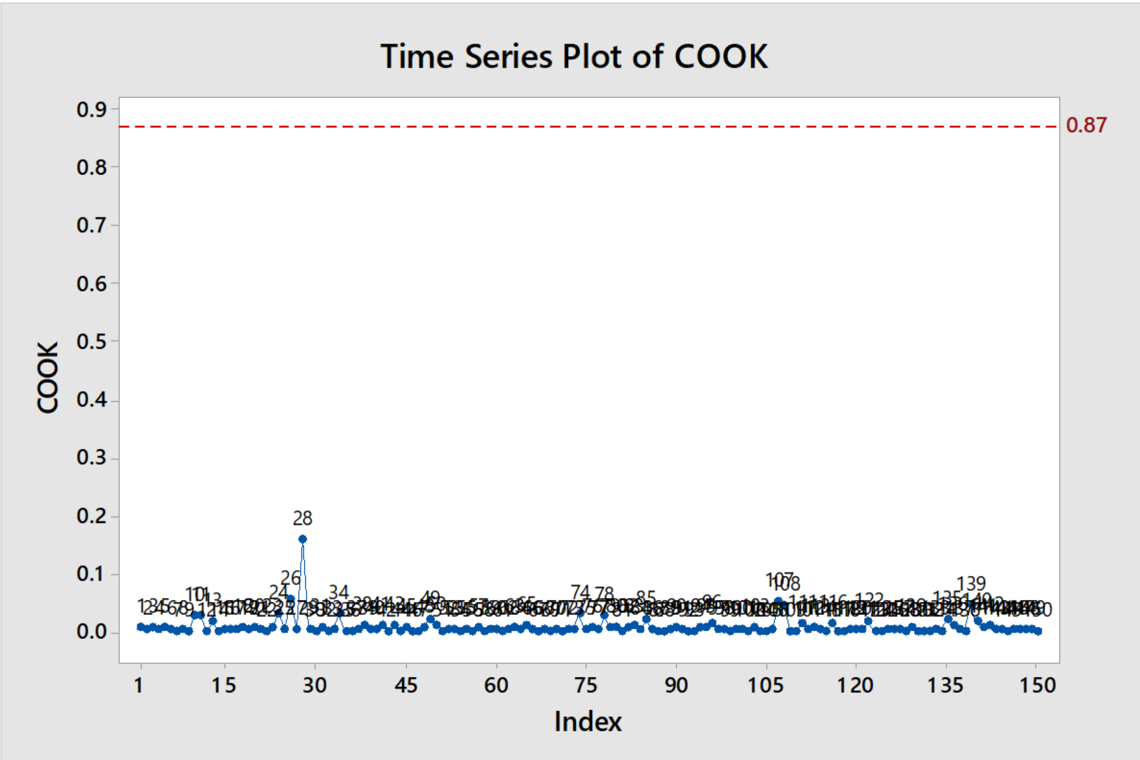

Figure 6E

Appendix E : Outlier, High Leverage and Influential Points

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	4	36.3723	9.0931	90.75	0.000
Ln(Experience)	1	15.5557	15.5557	155.25	0.000
Num_Teams	1	6.1806	6.1806	61.68	0.000
TD	1	1.0151	1.0151	10.13	0.002
Int	1	1.2043	1.2043	12.02	0.001
Error	145	14.5286	0.1002		
Lack-of-Fit	144	14.5282	0.1009	215.70	0.054
Pure Error	1	0.0005	0.0005		
Total	149	50.9009			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.316540	71.46%	70.67%	69.46%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	0.7557	0.0768	9.84	0.000	
Ln(Experience)	0.5003	0.0402	12.46	0.000	1.66
Num_Teams	-0.2022	0.0257	-7.85	0.000	1.63
TD	0.00990	0.00311	3.18	0.002	1.90
Int	0.02122	0.00612	3.47	0.001	1.57

Regression Equation

Salary^(0.3) = 0.7557 + 0.5003 Ln(Experience) - 0.2022 Num_Teams + 0.00990 TD + 0.02122 Int

Figure 7E

Appendix E : Outlier, High Leverage and Influential Points

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	4	39.9115	9.9779	136.31	0.000
ln(Exp)	1	16.3901	16.3901	223.90	0.000
Num_Teams	1	6.6584	6.6584	90.96	0.000
TD	1	0.7646	0.7646	10.45	0.002
Int	1	1.4883	1.4883	20.33	0.000
Error	135	9.8823	0.0732		
Lack-of-Fit	134	9.8819	0.0737	157.66	0.063
Pure Error	1	0.0005	0.0005		
Total	139	49.7938			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.270559	80.15%	79.57%	78.95%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	0.6932	0.0695	9.97	0.000	
ln(Exp)	0.5548	0.0371	14.96	0.000	1.71
Num_Teams	-0.2344	0.0246	-9.54	0.000	1.68
TD	0.00889	0.00275	3.23	0.002	1.96
Int	0.02498	0.00554	4.51	0.000	1.56

Regression Equation

sal^(.3) = 0.6932 + 0.5548 ln(Exp) - 0.2344 Num_Teams + 0.00889 TD + 0.02498 Int

Figure 8E

Appendix E : Outlier, High Leverage and Influential Points

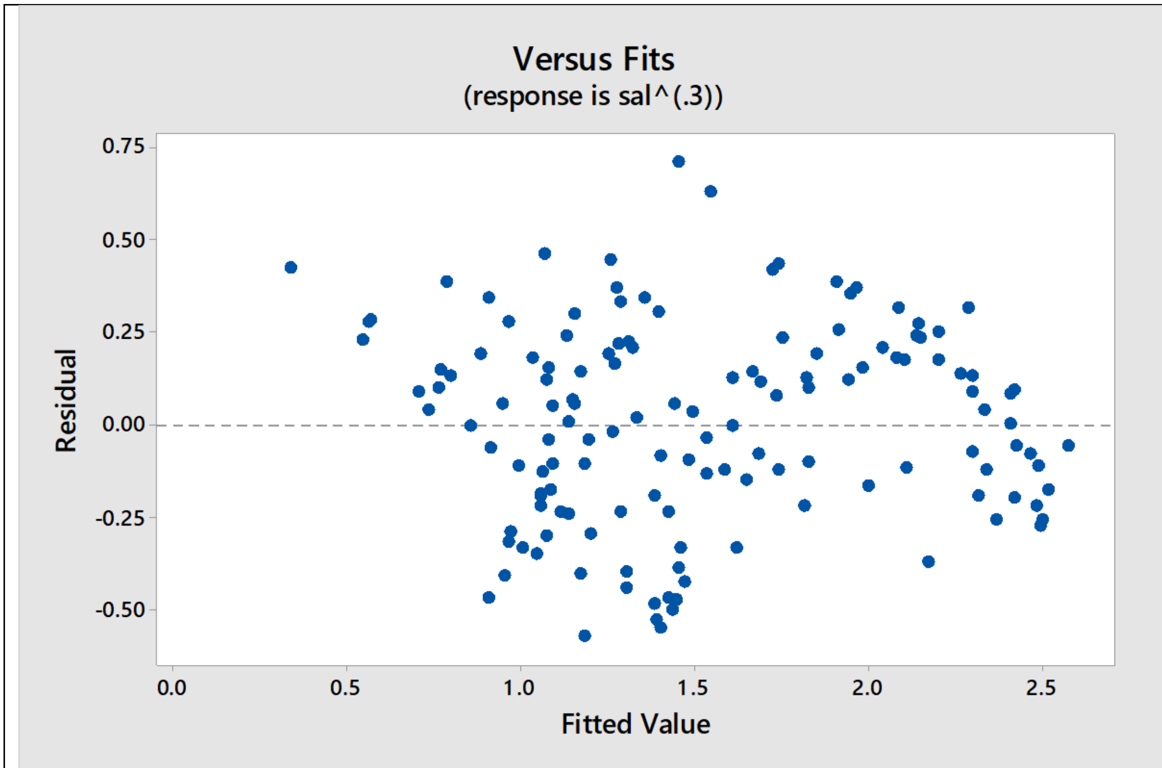

Figure 9E

N=140

$h = \frac{\sum_{i=1}^{n}(n-i+1)\times\hat{e}_i^2}{\sum_{i=1}^{n}\hat{e}_i^2} = 68.1646$

$$Q = \left(\frac{6n}{n^2-1}\right)^{\frac{1}{2}} * \left(h - \frac{n+1}{2}\right) = 0.207024949 * -2.3354 = -0.483486066$$

p-value= P(Z>Q) = 0.6856 > 0.05 so this is NOT violated

Figure 10E

Appendix E : Outlier, High Leverage and Influential Points

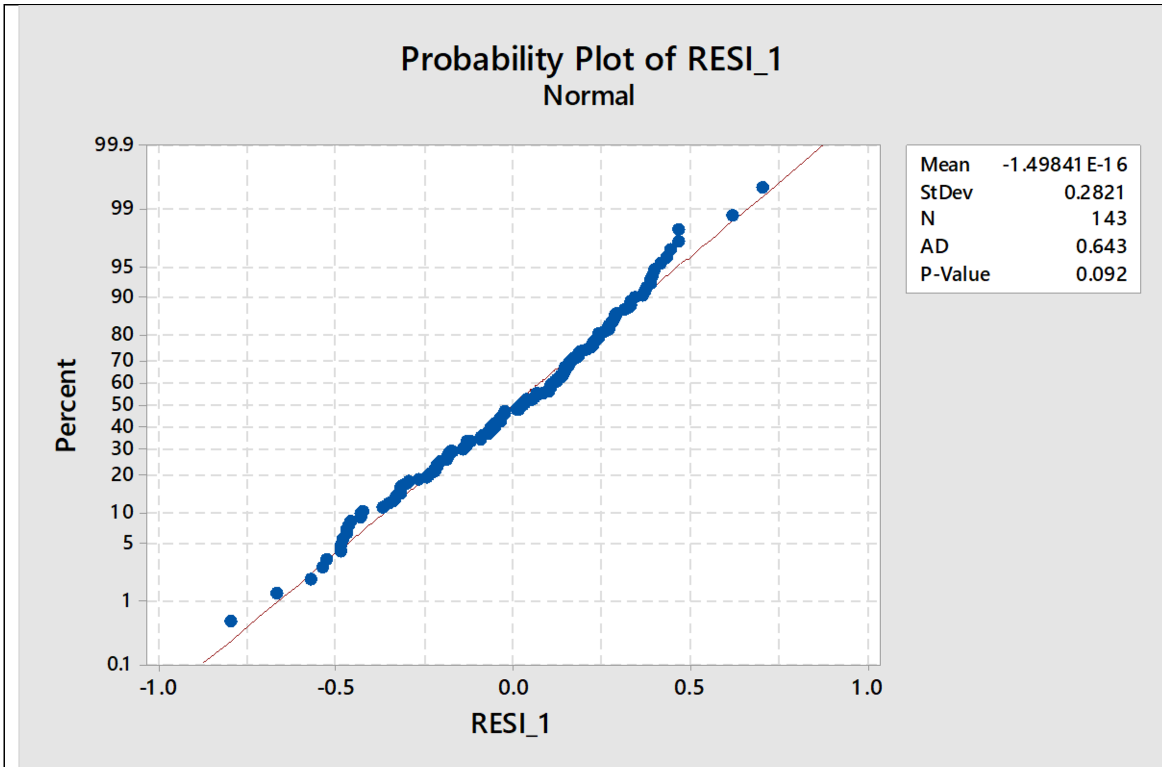

Figure 11E

Appendix E : Outlier, High Leverage and Influential Points

```
Durbin-Watson Statistic =  2.23340
```

N=140, k=4 [1.61,1.74] , Fall above to do not reject Ho

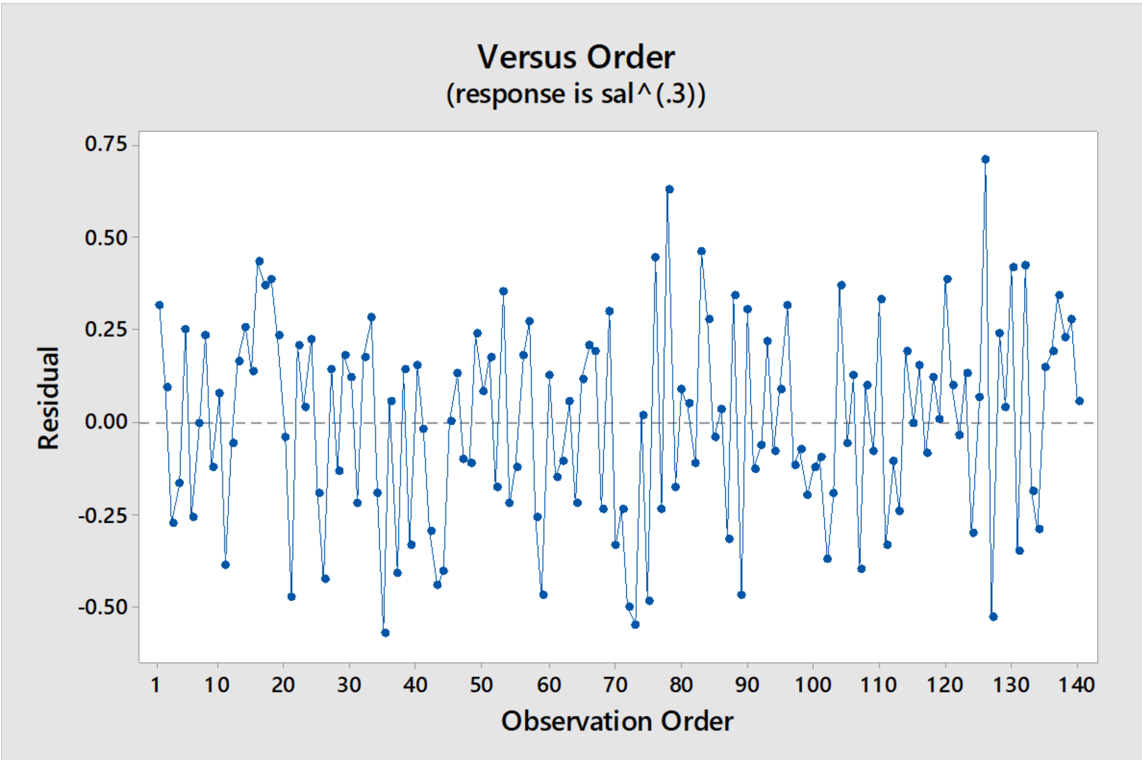

Figure 12E- Independce assumption

Appendix F: Prediction Analysis

Drew Brees 2015(Observation 1)

```
Regression Equation

sal^(.3) = 0.6932 + 0.5548 ln(Exp) - 0.2344 Num_Teams + 0.00889 TD + 0.02498 Int

Variable   Setting

ln(Exp)    2.70805

Num_Teams        2

TD              32

Int             11
```

Figure 1F

Drew Brees 2012

```
Regression Equation

Sal^(0.3) = 0.6932 + 0.5548 LN(Exp) - 0.2344 Num_Teams + 0.00889 TD + 0.02498 Int

Variable   Setting
LN(Exp)    2.56495
Num_Teams        2
TD              43
Int             19

    Fit     SE Fit        95% CI              95% PI
2.50442  0.0643201  (2.37722, 2.63163)  (1.95443, 3.05442)
```

Figure 2F

Lack of Fit Hypothesis

$$Ho: y = \beta_0 + \beta 1_1 x_1 + \beta_2 x_2 + \cdots + \beta_K x_K + \varepsilon$$

$$\text{Vs. } Ha: y \neq \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_K x_K + \varepsilon$$

Figure 1G

Constant Variance Hypothesis

$$H_0: Variance\ is\ constant$$

$$\text{vs.} H_a : Variance\ is\ not\ constant$$

Figure 2G

Normality Assumption Hypothesis

$$H_0: F(x) is\ normally\ distributed$$

$$\text{vs.} H_a: F(x) is\ not\ normally\ distributed$$

Figure 2G

Durbin-Watson Hypothesis

$$H_0: \rho = 0$$

$$\text{Vs. } H_a: \rho > 0.$$

Figure 3G

Durbin-Watson Hypothesis

$$H_0: \rho = 0$$

$$\text{Vs. } H_a: \rho > 0.$$

Figure 4G