**Regression Analysis Project Exam**

Brennan Poole

**Section 1: Introduction**

For my project, I’ve decided to take a statistical deep dive into one of my favorite hobbies, watching football. Football is something that has always been very important to me and my family throughout my life, and something that has made me feel great emotions as well, so it only felt natural that it be something that I research and perform analysis on. One aspect of the game that’s very important to not only the fans, but the teams themselves is a team’s ability to pass the ball, and the receiver’s ability to catch the ball. These statistics can be very important to teams when building their offense, teams need to balance spending with their needs to try and put together the best roster to win. When doing so teams perform statistical analyses to determine how much a player should be paid and which players they can bring in to fulfil a needed role on the team. I decided to analyze data from the top 100 pass catchers in the NFL the last two seasons to see if I could determine how many receptions a player will have in a season. My hypothesis for this data is that the number of targets as well as pro bowl and all pro team selections will be the most important things when determining the number of receptions a player gets in a season, because logically, the more times the ball is thrown to a player the more catches they should have, and players selected to all pro and pro bowl teams are considered the best in the league and therefore should have the most catches

The data I used was from pro-football-reference.com, a company that collects tons of data from various professional and collegiate sports leagues. Initially there were 19 attributes to my data set, and I chose the ones I believed to be the most important when determining the number of receptions and trimmed it down to 10 variables with 9 predictors and my target variable. I then removed a few points that had N/A values to make the data analysis easier and turned 3 of my variables into factor variables. These are the variables I chose:

|  |  |
| --- | --- |
| **Variable Name** | **Variable Description** |
| Age | A player’s age |
| Position (Pos) (Factor with 3 categories) | A player’s position |
| Receptions (Rec) (Target variable) | The number of receptions a player had in the season |
| Yards (Yds) | The number of yards a player had in the season |
| First Down Receptions (X1D) | The number of catches a player had in the season that resulted in a first down |
| Catch Percentage (Ctch.) | The percentage of passes thrown at a player that were caught |
| Touchdowns (TD) | The number of catches a player had in the season that resulted in a touchdown |
| Targets (Tgt) | The number of times the ball was thrown to a player |
| Pro Bowl (PB) (Factor with 2 categories) | Was the player selected for a pro bowl that season |
| All Pro (AP) (Factor with 2 categories) | Was the player selected for an all pro team that season |

A group of small colored text

Description automatically generated with medium confidenceHere is the structure and summary of the data. All variables are numeric except for position, pro bowl, and all pro. Position is split into 3 categories (Running Backs, Tight Ends, and Wide Receivers) with a large majority of the players with the most receptions being wide receivers with 116 compared to 30 running backs and 43 tight ends. The Pro Bowl variable is split into 2 categories, yes and no, with only 36 players having been selected to a pro bowl team and the rest not. The same is true for All Pro with only 19 of these players being selected to an all pro team and the rest not.

A graph of a distribution of targets

Description automatically generatedThe following are histograms of each variable to determine whether any transformations are necessary as well as box plots of the factor variables and their relationship to the target.

A graph of a number of different sizes

Description automatically generated with medium confidenceA graph of age distribution

Description automatically generatedA graph of reception and reception

Description automatically generated

A graph of reception and a table of tables

Description automatically generated with medium confidenceA graph of a distribution of touchdowns

Description automatically generatedA graph of a distribution of yards

Description automatically generatedA graph of reception and position

Description automatically generatedA graph of a distribution of a number of percent

Description automatically generated

A graph of reception and pro bowl

Description automatically generated

It can be seen from these box plots that players that play wide receiver generally have more receptions than the two other major pass catching positions, which makes sense as the wide receiver’s entire role is centered around catching the ball. Along with that players that are selected to pro bowl and all pro teams generally have more receptions than those who are not, which also makes sense as players seleced to these teams are seen as the best in the league so they should generally catch the ball more.

I determined that both the Targets and Touchdown variables were both too skewed to the right and don’t follow normal distributions, so I decided to perform a logrithmic transformation on them and it resulted in the following.

A graph of a log of targets

Description automatically generatedA graph of a log of touch downs

Description automatically generated with medium confidence

A graph of data analysis

Description automatically generated with medium confidence

A number and numbers on a white background

Description automatically generated

I then created an added variable plot and a correlation table to determine which of the variables had the highest correlation with the target variable. It can be seen from these figures that Yards (Yds), first down catches (X1D), ln(touchdowns), and ln(targets) are all show high correlation with receptions. With Yards, first down catches and ln(targets) all being very highly correlated with correlation coefficients of .8417, .8589, and .9090 respectively.

2. Regression Analysis

After cleaning and preparing all the data, the next step was to create a full model and see how it performed as well as check the assumptions for the model.

A screenshot of a computer

Description automatically generated

The full model appears to be a very well performing model with very high R^2 and adjusted R^2 values of .9751 and .9737 as well as residual standard error of 3.323. The biggest problem with this model is that most of the variables don’t appear to be significant when looking at the t-values and p-values. Age, position, yards, and all pro selection all have high p-values meaning that they do not pass a t-test for significance when being used to predict the number of receptions this indicates that the model may suffer from overfitting will be addressed later in my write up.

The next important step is to check the assumptions, then to find and potentially remove any outliers. The first assumption is Linearity, looking at the added variable plot again, it appears that all predictor variables are independed as there is no curve forming in any of the plots against the target variable.

A graph of data analysis

Description automatically generated with medium confidence

The next assumption is Independence, to check this I used the VIF function I found while researching for this project that helps show which variables may have multicollinearity.

A number and numbers on a white background

Description automatically generated

It can be seen here by looking at the adjusted VIF values, which is the column on the far right, that there appears to be no multicollinearity in this model. Generally, this adjusted VIF value is compared to 5 or 10 to determine if a variable is causing problems with independence, but all these values being below 5 is a good sign for the model’s independence.

The third assumption of homeoscedasticity was where I ran into some trouble, when checking for this assumption you must look at the graph of residuals vs fitted values for the model. For the model to pass the assumption, this graph should show to have basically no pattern, however the graph for the full model showed a discernible pattern.

A graph with black dots and red line

Description automatically generated

The solution to this problem is to perform a transformation on the predictor variable, in this case I determined that a logrithmic transformation was best to remove the heteroscadasticity. The new model yielded the Residuals vs Fitted graph in the figure below.

A graph of a number of values

Description automatically generated with medium confidence

A screenshot of a computer

Description automatically generatedA graph of data analysis

Description automatically generated with medium confidenceThis new model passes both the Linearity and Independence assumptions as can be seen in the figures below.

A graph of a graph

Description automatically generated with medium confidenceThe final assumption to be checked is Normality, this can be seen in the normal Q-Q plot. The points on this plot should be relatively linear which is the case here besides a few points that can be seen at the bottom, observations 49,76, and 97 appear to be high leverage points and will be investigated as to whether they should be removed.

A screenshot of a computer

Description automatically generated

A number and numbers on a white background

Description automatically generated

The new model with ln(rec) as the target yields the above summary and correlation table. This model performs even better than the first with incredibly high R^2 and adjusted R^2 values of .9993 and .9992 and standard error of .0081 but shows even fewer variables that are significant in determining the target, meaning that there is even more of an overfitting problem for this model.

This table provides each variable in the new model and how its coefficient can be interpreted in terms of the target variable.

|  |  |  |
| --- | --- | --- |
| **Variable** | **Interpretation in terms of lnRec** | **Interpretation in terms of Receptions** |
| Age | For every 1 unit increase in Age, lnRec decreases by .000121 | For every 1 unit increase in Age, Receptions decreases by 1.00 |
| PosTE | If a player is a tight end, lnRec increases by .01135 compared to running backs | If a player is a tight end, their number of receptions increases by 1.0114 compared to running backs |
| PosWR | If a player is a wide receiver lnRec, increases by .00949 compared to running backs | If a player is a wide receiver, their number of receptions increases by 1.010 compared to running backs |
| lnTgt | For every 1 unit increase in lnTgt, lnRec increases by .9965 | For every 1 unit increase in lnTgt, Receptions increases by 2.7088 |
| Yds | For every 1 unit increase in Yards, lnRec decreases by .000001568 | For every 1 unit increase in Yards, Receptions decreases by 1.00 |
| lnTD | For every 1 unit increase in lnTD, ln Rec decreases by .001164 | For every 1 unit increase in Touchdowns, Receptions increases by 1.0012 |
| X1D | For every 1 unit increase in X1D, lnRec increases by .0001469 | For every 1 unit increase in X1D, Receptions increases by 1.0001 |
| Ctch. | For every 1 unit increase in Ctch., lnRec increases by .01472 | For every 1 unit increase in Ctch., Receptions increases by 1.0148 |
| PBy | If a player is selected to the Pro Bowl, lnRec increases by .0007910 compared to those that aren’t | If a player is selected to the Pro Bowl, Receptions increases by 1.0008 compared to those that aren’t |
| APy | If a player is selected to be an All Pro, lnRec increases by .0007820 compared to those that aren’t | If a player is selected to be an All Pro, Receptions Increases by 1.0008 compared to those that aren’t |

Next, I investigated the 3 high leverage points in the new model (49,76, and 97) to determine whether they should be removed. To do so I created 3 separate models, each without one of the points, then calculated the standardized residual value, the leverage value, and the cook’s distance.

|  |  |  |  |
| --- | --- | --- | --- |
| Outlier | Standardized Residual | Leverage Value (hii) | Cook’s Distance |
| 49 | -4.19518 | .06314 | .1078 |
| 76 | -.5210 | .0246 | .0006 |
| 97 | -1.1310 | .09765 | .01278 |

The values in this table must be compared to certain thresholds to determine their leverage.

* If the absolute value of each outlier’s standardized residual is greater than 2 then it is considered an outlier. This is only the case for observation 49 making it an outlier.
* If the leverage value is greater than 3(k+1)/n then the point is considered to be high leverage. In this case, 3(k+1)/n = .16042. This again is not the case for any of these 3 points, so they are not high leverage.
* If the cook’s distance for any of the points is greater than .05 and greater than the F value for the model, which in this case is 1.8816, then it could be considered to have high influence, however again this is not the case for any of the points as observation 49 nay have a cook’s distance greater than .05 it is not greater than the F statistic of 1.8896 which means it has some influence but is not highly influential meaning it does not need to be removed.

I created a 95% confidence interval for receptions to estimate the population parameter for the top 100 pass catchers in the NFL. The interval is (64.943, 70.825) meaning that there is a 95% certainty that the population mean falls within this interval, or that 95% of pass catchers will catch an amount that’s within this interval. This can be useful information for teams to see which players are outliers either way or if a player is under or over performing, this is especially important when it comes time for contract negotiations.

The final step is model selection and validation. I came up with 4 different models to determine which was best. The first of these came from forwards and backwards selection, one was my own that I chose based on which variables had the highest correlation coefficient compared to the target, the third was the same hand picked model with an added interaction term between yards and lnTDs because I believe there is probably a strong relationship between a players yardage count and the number of touchdowns they score, and the full model.

The first model comes from forwards stepwise selection using AIC as the selection criteria. Interestingly enough, forwards stepwise selection using AIC and BIC as the selection criteria and backwards selection using both AIC and BIC as the selection criteria both yielded the exact same model.

A screenshot of a computer

Description automatically generated

The next model is my hand-picked model that I choose from the variables that were highly correlated with the target.

A screenshot of a computer program

Description automatically generated

The final model is the model with the interaction term added. A screenshot of a computer program

Description automatically generated

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **AIC** | **BIC** | **Adjusted R^2** |
| Full Model | -1272.976 | -1234.075 | .9992 |
| Stepwise Model | -1282.046 | -1262.596 | .9992 |
| Handpicked Model | -294.504 | -275.053 | .8593 |
| Interaction Term Model | -298.145 | -275.452 | .8626 |

Based on the data collected about all 4 of these models, the best model is the one chosen through the stepwise selection process. Though all 4 models seem to perform well with low AIC and BIC numbers as well as very high adjusted R^2 values, the Stepwise model is just a step above the rest with the lowest AIC and BIC values of -1282.046 and -1262.596 respectively. The only other model that comes close to this performance is the full model with AIC of -1272.976 and BIC of -1234.075. Not only is the stepwise model slightly superior in both these categories, it also avoids concerns about overfitting like the full model, and the other 2 models have, as all its predictors are shown to be significant through t-testing. Another thing worth noting, my interaction term did turn out to be significant, as seen in the summary of the 4th model, the interaction between yards and ln(touchdowns) ended up showing statistical significance in determining ln(receptions) and it performed better than the handpicked model without any interaction term.

3. Discussion and Limitations

When looking at the stepwise model I chose, it makes sense that it is the best model for predicting receptions. The only predictors in the model are position, targets, and catch percentage. The only thing that truly matters in terms of how many catches a player gets are, the position the player plays, the number of times the ball is thrown to them, and the percentage of times a player catches a ball thrown towards them all. Players playing positions that catch the most balls will catch more balls, players that get the ball thrown to them more will catch more balls, and players that have a higher catch percentage will catch more balls. My initial hypothesis was that targets, pro bowl, and all pro selections would be the most important predictors in determining a player’s number of receptions. I was partially correct as targets did end up being one of the 3 most important predictors, though pro bowl and all pro selections ended up not being very significant. This does make sense as players in my data set could have been selected to a pro bowl or all pro team from a position that catches fewer passes such as running back, or they could be selected for to a special teams role such as kick/punt returner which has nothing to do with their play on offense where they would be catching the ball. It has also happened that a player turns down their pro bowl selection and doesn’t earn one for that year despite being one of the best players at their position.

The biggest limitations of my dataset would be the fact that it doesn’t account for players being injured or sitting out for any reason. For example, if a player starts the season out playing very well and is on pace to have 100 receptions by the end of the year but gets hurt early on and ends the season with only 6 games worth of statistics compared to the usual 17, they would be penalized unfairly by my models. My solution for this would be to add more variables that consider stats per game and games played rather than just total cumulative stats and full season awards. Another major limitation is that my dataset only considers the top 100 pass catchers from the past 2 years. There may be players that missed one of those seasons due to injury that may have been on the list, also there are far more than 100 pass catching players in the NFL. My statistical analysis does not consider those players at all when determining the best model to predict receptions, the additions of hundreds more players could very well change the outcome of my analysis.

4. Conclusion and Future Work

In conclusion, the best model for predicting a player’s number from the dataset I have is the one chosen by the stepwise selection process that has ln(receptions) as the target and position, catch percentage, and ln(targets) as the predictors. As stated in the previous section, this makes sense considering the simplicity of these statistics and their importance regarding the number of receptions a player has in a year. In the future, to improve my model and analyses, I would add 2 things. First, I would add the data from several more years, and hundreds more players from each year to be able to get a full grasp on the historic data of the NFL’s pass catchers to be able to better predict the future. Second, I would add statistics that account for the number of games played by each player, things like yards per game or receptions and targets per game are something I would consider using in the future to make my analysis as accurate as possible. I would hope that anyone reading this learned something about the game that I have always loved, and that it may be useful to them or lead them into a future as a fan of the sport as well.

**References**

*2023 NFL receiving*. Pro. (n.d.). https://www.pro-football-reference.com/years/2023/receiving.htm

*2022 NFL receiving*. Pro. (n.d.-a). https://www.pro-football-reference.com/years/2022/receiving.htm