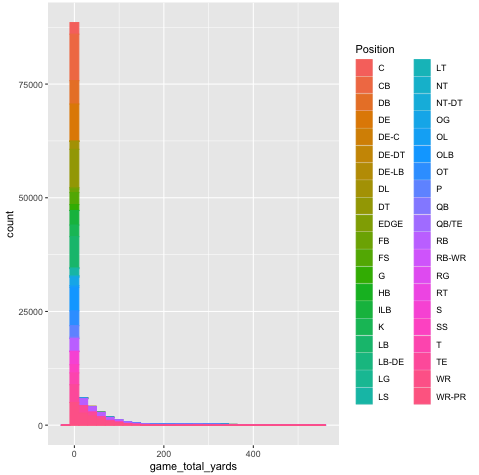
Discussion Board Post #4

Steven McWilliams

I chose to use a different dataset from my last post. My dataset consists of NFL player statistics for each game played from the 2016 to 2020 NFL seasons. My goal is to use NFL game data to build a model that calculates a player’s total yardage gained in a given game. This model is intended to be used from the coach’s point of view, meaning that the coach could input their game plan (offensive snaps, pass attempts, etc. for each player) into the model to calculate expected yardage gained in a given game for any player.

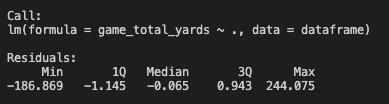
I only used a few variables of the many available in my analysis; Team, Opponent, Age, Team Season Game #, Home, Game Started, Game Offensive Snaps, Career Offensive Snaps, Season Offensive Snaps, Game Receiving Targets, Game Rushing Attempts, Game Passing Attempts, Career Passing TD, Career Total Yards, and Game Total Yards. Game Total Yards is the LHS variable and all others are RHS.

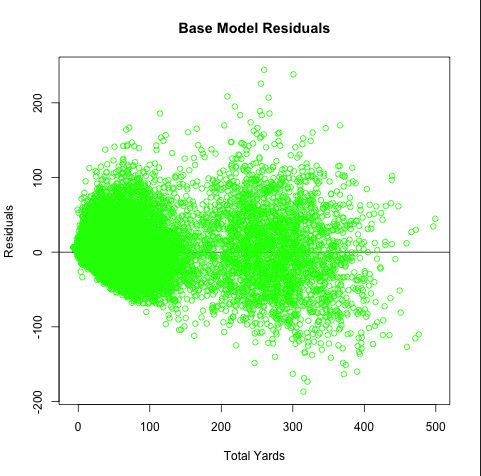
Below is a histogram of the LHS variable that I’ll be solving for. The right skew indicates that a linear model is not the best fit for this data.



**Step One**

I first built a standard linear regression model with no applied techniques to the RHS or LHS variables – this model will serve as the baseline for me to compare further adjusted models to. This model performs fairly well, but I am concerned about the heteroskedasticity in the residuals, that is what I’ll be attempting to correct for. I should note that I cannot include pictures of the model coefficients, as there are too many variables for me to screenshot. I have attached the residual outputs and model summary instead.



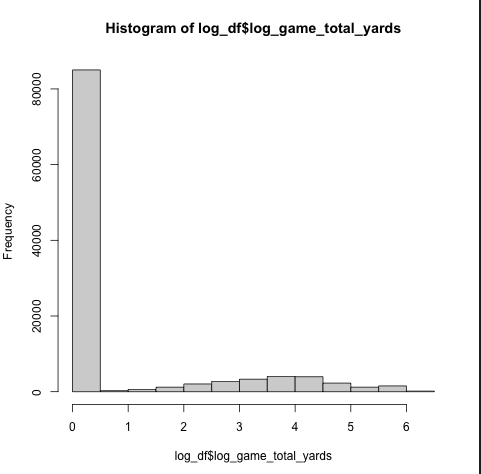




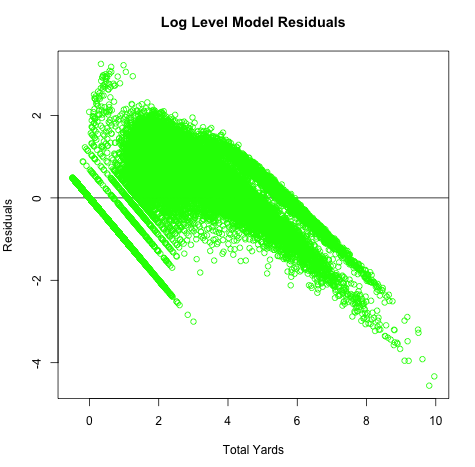
The histogram and heteroskedastic residuals indicate that the relationship between the RHS and LHS variables is not completely linear, meaning a log LHS variable or quadratic RHS variables may result in a more accurate model.

**Step Two**

First, I constructed a Log-Level model. All instances (313) of negative yardage gained in a game had to be removed in order for the log function to work and not return. NaNs. I then applied the log1p function calculate the log of the LHS variable. This transformation of the data resulted in a more normal distribution, but the amount of player-game with zero yards gained is substantially impacting the model.



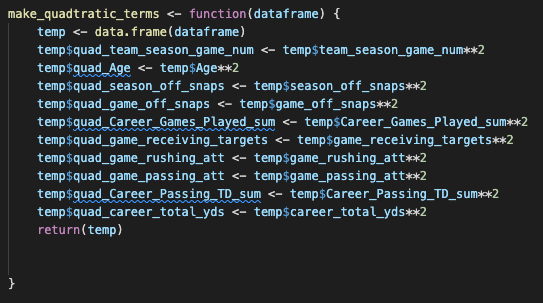
While the Log-Level model results in less of a fan shape, there is still evidence of a non-linear relationship due to the n-shaped pattern in the residuals.



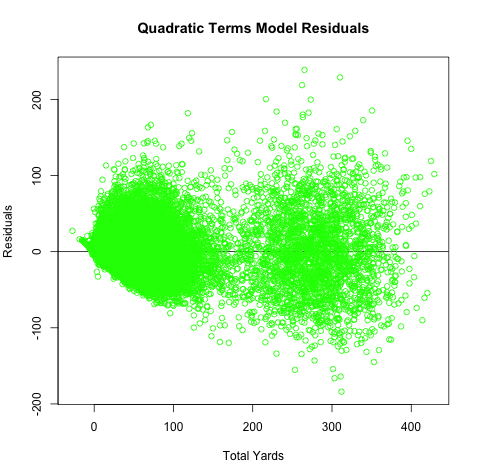
Further, the Log-Level model results in a lower r-squared and lower f-statistic than the Base Model. Of note, the residual standard error is much lower than the base model at 3.25 yards compared to the Base Model’s RSE of 13.81.



This data presents many opportunities to capitalize on diminishing returns with variables like Age, Career/Game/Season Snaps, Passing/Rushing/Receiving Attempts, and Season Game #. In order to do so, I created a quadratic model and applied the quadratic function to many of the numerical variables, shown below.



The residuals of the Quadratic Terms model are still heteroskedastic, albeit in a tighter formation than the Base model.



The Quadratic Terms model performed better than base, with a RSE of 13.56 and an r-squared of 0.92, higher than the Log-Level and Base model.



**Step Three**

While the Quadratic model is slightly better than Base, it doesn’t provide any material value over the Base model and neither does the Log-Level model. Despite high r-squared values in the model summaries, all residual plots show that there is a non-linear relationship within the data. Despite using dummy variables for each categorical variable, the non-linear relationship could not be identified.

**Step Four**

While the categorical variables include Team, Opponent, Position, Game Started and Home. I would have liked to have a categorical for Playoff Team, indicated that the team was in the playoffs the prior year. This could be used for both players and the Team/Opponent. This would be intriguing to including in the model because it is a categorical value that captures the performance of a team, which is unavailable otherwise. This unique characteristic may lead to a stronger connection between teams and yardage gained by a player in a given game.