**Regression “Lab”**

NASCAR driver Matt Kenseth won the 2012 Daytona 500. His win was no surprise because for the 2011 season he finished fourth in the point standings with 2330 points. In 2011, he earned $6.184 million dollars by winning three Poles (fastest driver in qualifying), winning three races, finishing in the top five 12 times, and finishing in the top ten 20 times. NASCAR’s point system in 2011 allocated 43 points to the driver who finished first, 42 points to the driver who finished second, on down to 1 point for the driver who finished last. In addition, any driver who led a lap received 1 bonus point, the driver who led the most laps received an additional bonus point, and the race winner was awarded 3 bonus points. But, the maximum number of points a driver could earn in any race was 48. The accompanying spreadsheet contains the 2011 data for the top 35 NASCAR drivers. The variables include the drivers points for the season (POINTS), the number of poles won (POLES), the number of races won (WINS), the number of top 5 finishes (TOP5), and the number of top 10 finishes (TOP10), along with each drivers winnings (in millions of $).

Suppose that you want to predict the winnings for a NASCAR driver based on his performance at the races. Use Excel or R to complete the following tasks.

1. Create scatterplots for WINNINGS vs each of the explanatory variables in the data set. (Copy and paste the graphs into a document for submission)

A graph with blue dots

Description automatically generated

1. Create a correlation matrix or correlation plot for all of the variables in the data set. (Copy and paste the correlation matrix or plot into a document for submission)

A screenshot of a graph

Description automatically generated

1. Based on the information generated above, do you believe that it is reasonable to use a linear regression model to predict winnings? Which of the explanatory variables appears to have the strongest relationship with winnings?

It is reasonable to use a linear regression model to predict winnings because there are many variables that have a correlation that is greater than 0.8. The variable with the strongest relationship is Top 10 with a correlation of 0.9.

1. Fit a linear regression model that includes all of the explanatory variables. (Copy and paste the Coefficients table, the Residual by Predicted plot, and the Q-Q Plot into a document for submission)

A screenshot of a computer screen

Description automatically generated

A graph with blue dots

Description automatically generated

A graph with blue dots

Description automatically generated

1. Do you believe that this model could be useful for predicting winnings? Why or why not? Do you have any concerns about the model? What are they?

This model could be useful for predicting winnings. However, it could be made better since the p-values for points, poles, and wins are large. I am going to remove these variables in the next model I make.

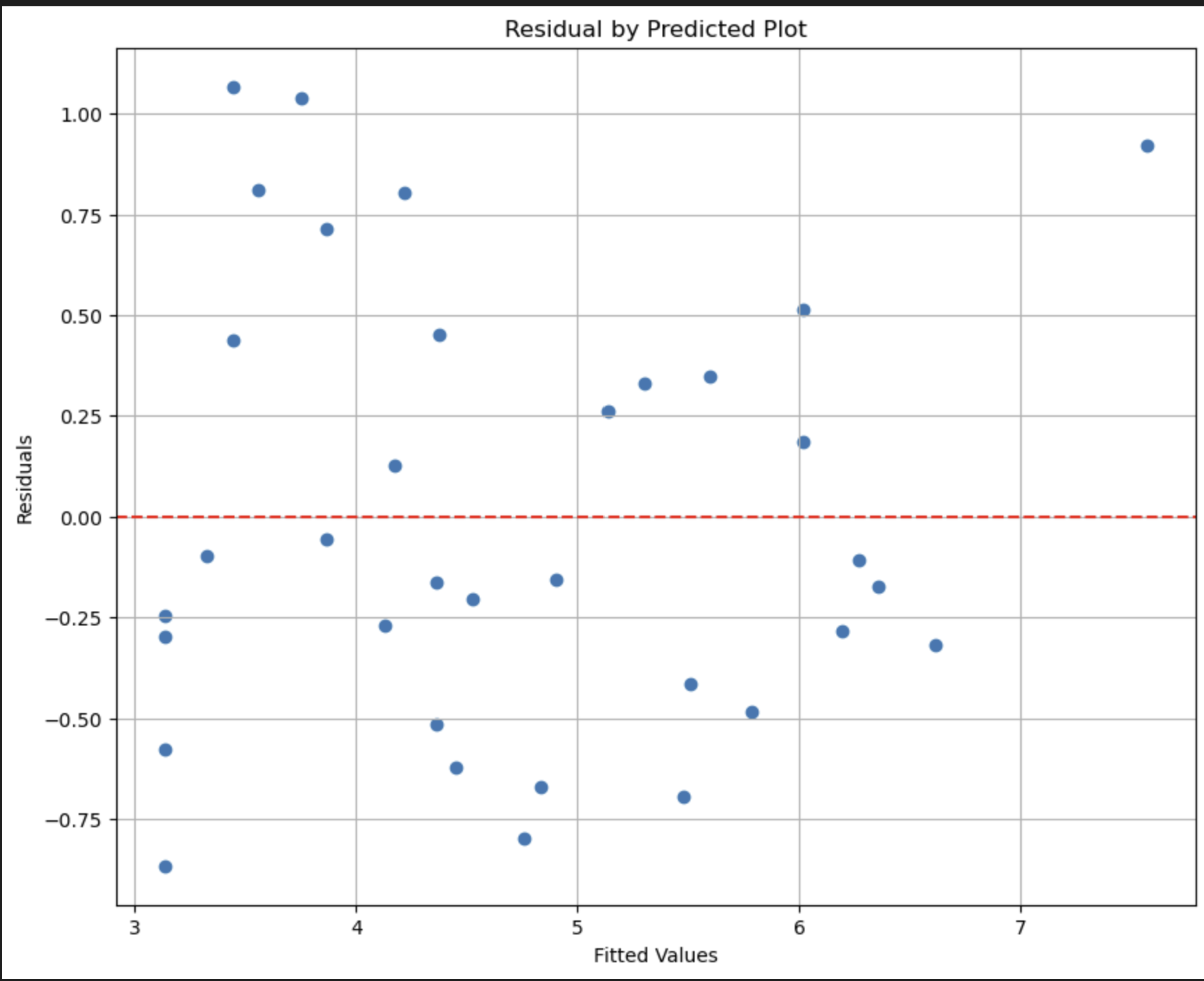
1. Now systematically remove non-significant variables from the model either “by hand” or using automated variable selection techniques. Once you are satisfied with your reduced model, report the Coefficients table, Residual by Predicted plot, and Q-Q plot. How does this model compare to the full model? Which model would you consider the “best”? Why? Do you have any concerns about this model that may still need to be addressed?

A screenshot of a computer

Description automatically generated

A graph with blue dots

Description automatically generated



I would consider this model the better one because it has a higher adjusted R squared value. Compared to the original model, the R squared value decreased a few points because I removed variables. The only concern that I have with this new model is that Top 5 and Top 10 are highly correlated to one another which could throw off the new model's accuracy.

1. Ignoring any additional concerns, based on this model, what advice would you give to a NASCAR driver who is interested in increasing his winnings? Include a statement about how much additional money he could expect to win if he followed your advice.

I suggest that a NASCAR driver should focus on placing at least in the top 10 in each race. If the driver places in the top 10, their earnings will increase by 11.68%, and if the driver places in the top 5, their earnings will increase by an additional 7.38%.

1. Suppose that a driver has 22 top 10 finishes, 13 top 5 finishes, 3 wins, and no poles with a total of 2345 points for the season. How much should he expect to win?

Earnings = 3.1134 + 7.022e-05(2345) + 0.0034(3) + 0.0703(13) + 0.1127(22) = $6.682 million