In NASCAR, driver rating is a metric used to evaluate the performance of drivers in races. It provides a comprehensive assessment of a driver's overall competitiveness, efficiency, and consistency during a race. The driver rating is based on several key performance factors and is designed to offer a more objective view of a driver's abilities.

The driver rating is calculated using a formula that combines these aspects such as wins, finish position, average running position, and multiple other components into a single numerical value. Although the exact formula for calculating driver rating has been tweaked over the years, and explanation and example using the 2006 structure is available at <https://www.driveraverages.com/drvavg/nascar-driverratings.php>.

Suppose that, given the complexity of the driver rating formula, you wish to build an approximation to this metric using a statistical model. To do so, you have access to data from the NASCAR website (<https://www.nascar.com/stats/>). This data shows the season statistics from 2007-2022 with variables such as wins, average start, average mid race, average finish, average position, pass difference, green flag passes, green flag passed, quality passes, percent of quality passes, number of fastest laps, laps in top 15, percent of laps in top 15, laps led, percent of laps led, total laps, driver rating, and points.

For this exercise, you will be examining the relationship between a driver’s average running position across all races they competed in (AvgPos), and their average driver rating (DriverRating) for a season. (Note that a separate driver’s rating is calculated for each race. The DriverRating variable represents the average of these values.)

1. Create a scatterplot with a smoother displaying the relationship between driver rating and AvgPos. Discuss the trends you see in the relationship between the two variables.

A graph showing a line of dots

Description automatically generated

There appears to be a strong negative correlation that displays curvature. As AvgPos increases, DriverRating decreases.

1. Regardless of what you observed in the scatterplot, fit a simple linear model, using AvgPos to predict DriverRating. Report the least squares regression equation of the linear model.

A close-up of a number

Description automatically generated

DriverRating = 121.86906 – 2.50019(AvgPos)

1. Plot the residuals vs. fitted values of the model. Use the plot to check the linearity assumption for the linear model. Does it seem reasonably met? If not, explain why.

A graph of a graph showing a number of values

Description automatically generated

The linearity assumption does not seem reasonably met due to the concave up curvature in the residuals vs. fitted values plot.

1. One way to fix the lack of linearity is by applying a transformation to at least one of the variables used in the model.
   1. Use natural logs of driver rating and/or average position to find a combination of variables that display a relationship that should satisfy the linearity assumption.

DriverRating ~ log(AvgPos)

* 1. Using proper notation, record the estimated equation for the model that you think best models the relationship between the variables (from those you investigated in the previous question).

DriverRating = 235.3411 – 56.1008\*log(AvgPos)

* 1. Create a new residual vs. fitted values plot for the model you chose. Does the linearity assumption seem to be met now? How has the transformation altered the model? Are there any new issues?

A graph of black circles

Description automatically generated

Linearity assumption seems to be met now by taking the natural log of AvgPos. The intercept and slope are significantly higher than in the original model.

There is still slight curvature and slight inconsistency in variance.

1. Another method of transforming the data is by taking the square root of our variable(s).
   1. Use square roots of driver rating and/or average position to find a combination of variables that display a relationship that should satisfy the linearity assumption.

DriverRating ~ sqrt(AvgPos)

* 1. Using proper notation, record the estimated equation for the model (using square roots) that you think best models the relationship between the variables.

DriverRating = 178.2119 – 24.2353\*sqrt(AvgPos)

* 1. Create a new residual vs. fitted values plot for the model you chose. Does the linearity assumption seem to be met now? How has the transformation altered the model? Are there any new issues?

A graph of black circles

Description automatically generated

The linearity assumption does seem to be met now. The intercept and slope are higher than in the original model but lower than in the model using the natural log. There does not appear to be any new issues aside from some unusual observations seen in the Q-Q plot.

1. Another way to deal with curvature in a relationship is to use polynomial regression.
   1. Investigate models such as quadratic and cubic regression to see if a polynomial can adequately model the relationship between driver rating and average position.

DriverRating ~ AvgPos + AvgPos^2

* 1. Using proper notation, record the estimated equation for the polynomial regression model that you think best models the relationship between the variables.

DriverRating = 145.724 – 4.718398(AvgPos) + 0.044362(AvgPos^2)

* 1. Create a new residual vs. fitted values plot for the polynomial model you chose. Does the linearity assumption seem to be met now? How has the transformation altered the model? Are there any new issues?

A graph of black dots

Description automatically generated

Linearity assumption is now met. The intercept and slope are more similar to the original model. There are no new issues besides some inconsistency with variance.

1. In 2022, Ross Chastain’s AvgPos was 12.4. Using all three model equations, provide a prediction for Chastain’s driver rating for the season.

DriverRating = 235.3411 – 56.1008\*log(12.4) = 94.096

DriverRating = 178.2119 – 24.2353\*sqrt(12.4) = 92.871

DriverRating = 145.724 – 4.718398(12.4) + 0.044362(12.4^2) = 94.037

1. In reality, Ross Chastain ended the 2022 season with a driver rating of 92.8. Calculate the residual for each of your estimates.
2. Which of the three models, log-transformation, square root-transformation, or polynomial regression, do you think is the best model? You should justify your answer using statistically sound methods. (e.g., by assessing model assumptions, strength of fit, ease of use)

All three models are fairly comparable. The polynomial model has the highest Adjusted R-squared value with a very flat residual vs fitted plot line. The square root model also has little curvature in the residual vs. fitted plot and the second highest Adjusted r-squared

Log Model: Adjusted R-squared: 0.9611

Sqrt Model: Adjusted R-squared: 0.9618

Polynomial model: Adjusted R-squared: 0.9653