**Introduction**

For this activity, you will be examining NASCAR data and exploring relationships between key variables relating to driving performance. Using data transformations to create different linear models, you will enhance the capabilities of your models to make them more effective and accurate.

**Learning Goals**

By the end of the activity, you will have practiced:

* Assessing model effectiveness
* Checking model assumptions
* Transforming data to better fit a linear regression model

**Data**

The data can be found 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. A preview of the first six rows and ten variables of the data can be found below.

**A screenshot of a calendar

Description automatically generated**

**Methods**

For this activity, students will need to use software to create scatterplots and examine both residual and fitted values of created models. They will also need to mutate the data by applying mathematic functions to columns.

**Exercises**

For this exercise, you will be examining the relationship between a driver’s average finish for a race, and their driver rating.

1. Create a scatterplot with a smoother displaying the relationship between DriverRating and AvgFinish. Discuss the trends you see in the relationship between the two variables.

A math equation with black text

Description automatically generated

There is a strong and significant negative relationship. The scatterplot also displays significant curvature.

1. Turn this relationship into a linear model, using DriverRating to predict the AvgFinish of a driver. Report the least squares regression equation of the linear model.

A close-up of a math formula

Description automatically generated

= 45.45 – 0.35(DriverRating)

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 close up of a sign

Description automatically generated

There is obvious curvature in the residual vs. fitted values plot. Linearity assumption does not appear to hold.

1. If the linearity assumption does not hold, a linear model may not be appropriate for dealing with the relationship. One way to fix this is by taking the natural log of the explanatory variable. Create a new scatterplot below that displays the relationship between average finish and the natural log of DriverRating. What seems to have changed?

A black text with black letters

Description automatically generated

A computer screen with text

Description automatically generated

The curvature has been eliminated and the relationship is now far more linear.

1. Recreate the model using the natural log of DriverRating to predict AvgFinish of a driver and report the new least squares regression equation of the model. How has the size of a typical error changed?

A close-up of a computer code

Description automatically generated

= 104.61 – 20.00(DriverRating)

The residual standard error is reduced from 3.423 to 3.202 meaning the size of a typical error has been diminished through the transformation.

1. Create a new residual vs. fitted values plot. Does the linearity assumption seem to be met now? How has the transformation altered the model?

A black text on a white background

Description automatically generated

Looks like linearity assumption now holds as the line has been flattened in this plot as well as the scatterplot.

1. Another method of transforming the data is by taking the square root of our explanatory variable. Now using the square root of driver rating, create a scatterplot and residuals vs. fitted values plot below. How has this transformation impacted the linearity of the relationship?

A math equations on a white background

Description automatically generated

Once again, the transformation reduces the curvature of the relationship and allows for the linearity assumption to hold.

1. Report the least squares regression equation of this transformed relationship.

A close up of a text

Description automatically generated

= 65.44 – 5.40(DriverRating)

1. Does this model using the square root of driver rating seem to be better or worse than the model using the natural log of driver rating? Explain the reasoning for your conclusion.

The model using the natural log of driver rating has a slightly lower residual standard error and a slightly higher R-squared value making it more effective.

1. In 2022, Ross Chastain had a driver rating of 92.8 points. Using all three model equations, provide a prediction for Chastain’s average finish for the season.

= 45.45 – 0.35\*92.8 = 12.97

= 104.61-20.00\*ln(92.8) = 14.00

= 65.44-5.40\*sqrt(92.8) = 13.42

1. In reality, Ross Chastain ended the 2022 season with an average finish of 14.2 Calculate the residual for each of your estimates.

e1 = 14.2 - 12.97 = 1.23

e2 = 14.2 - 14 = 0.2

e2 = 14.2 – 13.42 = 0.78