**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 simply need to interpret the results of various models.

**Exercises**

1. The correlation between a driver’s average finish and their driver rating is -0.926. Provide an interpretation for this value.
2. Now let's turn this relationship into a linear model, using driver rating to predict the average finish of a driver. Write the equation of linear model. Also take note of the adjusted R-squared value of the model.

A screenshot of a computer error

Description automatically generated

1. Below is a scatterplot of average finish against driver rating. What trends can be identified in this plot?

A graph showing a curve

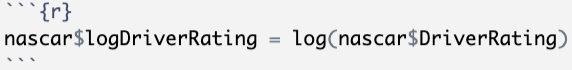
Description automatically generated

1. Here is a plot of the residuals vs. the fitted model values. Check the linearity assumption for the linear model. Does it seem reasonably met? What issues can be identified?

A graph of a number of dots

Description automatically generated

1. How can we potentially fix this issue? See the code below and provide a description for what it is doing to the data.



1. What seems to have changed in this new scatterplot that now incorporates this transformed variable of driver rating?

A graph showing a line of dots

Description automatically generated

1. Looking at both the new scatter plot and residual vs. fitted values plot, does the linearity assumption seem to be met now?

A graph of a graph showing a number of values

Description automatically generated

1. Below is a summary of the new model that takes the natural log of DriverRating to predict AvgFinish of driver. Write down new equation and compare the adjusted R-squared values. Does this transformed model appear to be more effective than before?

A screenshot of a computer program

Description automatically generated

1. Using the transformed model, what is the predicted average finish of a driver with a rating of 90?
2. Given the fact that the data is from consecutive NASCAR seasons, what issues could be present? Is there a model assumption that is may be violated? What is a possible solution to this problem?