STAT 113 – NASCAR Transformations Worksheet

Description: The dataset nascar\_df.csv includes statistics of NASCAR drivers from races since 2007 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.

Goal: To learn what regressions are appropriate for certain models.

Here is a scatterplot of the relationship between Average Start and Driver Rating. Average Start is the sum of the driver’s starting positions divided by the number of starts. Driver Rating is a formula combining several categories to rate the nascar drivers.

Discuss with your neighbors if there are any trends you see.

Chart, scatter chart

Description automatically generated

The correlation for AvgStart and DriverRating is -0.8948724. What does this correlation mean?

Let’s look at the linear regression model formula. Write down the linear model formula, as well as examine the residual plots to determine if this model is the appropriate fit. Also, write down the adjusted R2 for the model.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Coefficients: | | | | |
|  | Estimate | Std. Error | t value | Pr(>|t|) |
| (Intercept) | 112.12259 | 0.86962 | 128.93 | <2e-16 |
| AvgStart | -2.09943 | 0.03144 | -66.77 | <2e-16 |

|  |  |
| --- | --- |
| Residual standard error: | 10.72 on 1109 degrees of freedom |
| Multiple R-squared: | 0.8008 |
| Adjusted R-squared: | 0.8006 |
| F-statistic: | 4458 on 1 and 1109 DF |
| p-value: | <2.2e-16 |

Chart, scatter chart

Description automatically generated

Look at the original scatterplot again. Is a linear regression model the most appropriate model to use for this association? Why or why not?

STAT 213 – NASCAR Transformations Worksheet

Description: The dataset nascar\_df.csv includes statistics of NASCAR drivers from races since 2007 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.

Goal: To learn what regression are appropriate for certain models.

Driver Rating vs Average Finish

Here is the scatterplot of Driver Rating vs Average Finish. Driver Rating is a formula combining several categories to rate the nascar drivers. Average Finish is the sum of the driver’s finishing positions divided by the number of finishes.

Example 1: Logarithmic Transformation for AvgFinish vs DriverRating

Discuss with your neighbors if there are any trends you see.

Chart, scatter chart

Description automatically generated

Try producing a logarithmic regression model for AvgFinish ~ Driver Rating by applying log to y and/or x to improve the fit. Then create the residual plot to determine if this model is the appropriate fit. Also, write down the adjusted R2 for the model.

Given your new logarithmic transformation made, predict the Average Finish when Driver Rating = 80.

STAT 113 – NASCAR Predicting Driver Rating Worksheet

Description: The dataset nascar\_df.csv includes statistics of NASCAR drivers from races since 2007 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.

Goal: Learn how to find the best variable to predict the response variable.

Let’s pick three correlated variables with Driver Rating. In this case, let’s use Average Start, Total Laps, and Percent of Laps In Top 15.

Look at each of the scatterplots.

Chart, scatter chart

Description automatically generatedChart, scatter chart

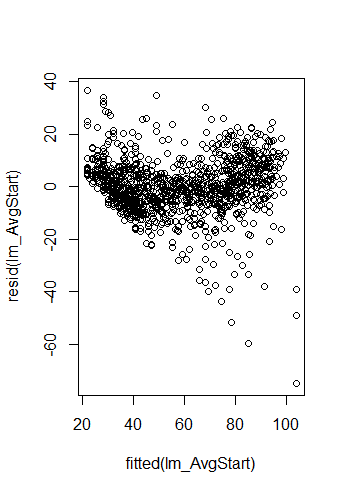
Description automatically generatedChart, scatter chart

Description automatically generated

Discuss with your neighbors and write which variable is the best to predict Driver Rating based on these scatterplots. Why?

Let’s use AvgStart to predict DriverRating. Look at the linear regression model formula and the residual vs fitted model of the best variable to confirm it is the best variable to predict Driver Rating.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Coefficients | | | | |
|  | Estimate | Std. Error | t value | Pr(>|t| |
| (Intercept) | 112.12259 | 0.86962 | 128.93 | <2e-16 |
| AvgStart | -2.09943 | 0.03144 | -66.77 | <2e-16 |



Use the AvgStart vs DriverRating scatterplot to predict Driver Rating when the best variable = 20.

STAT 213 – NASCAR Predicting Driver Rating Worksheet

Description: The dataset nascar\_df.csv includes statistics of NASCAR drivers from races since 2007 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.

Goal: To see if more than one variable could be used to better predict the response variable.

We will be using the variables Average Start, Percentage of Laps in Top 15, and Total Laps to see if we can better predict Driver Rating.

Chart

Description automatically generated

Create a linear regression model formula for just Average Start and Driver Rating and report the adjusted r-squared value.

Now let’s pick another variable with Average Start for a multi-linear regression model formula for predicting Driver Rating. Here, we will use Percent of Laps in Top 15. Create the multi-linear regression model formula for Average Start and Percent of Laps in Top 15 predicting Driver Rating, create the residual vs fitted model, and report the new adjusted r-squared value.

Discuss with your neighbors and write down if this multi-linear regression model better predicts Driver Rating. Why or why not?

Now let’s use all three variables for another multi-linear regression model formula for Driver Rating. Create the multi-linear regression model formula, create a residual vs fitted model, and report the new adjusted r-squared value.

Discuss with your neighbors and write down if this multi-linear regression model better predicts Driver Rating. Why or why not? And which multi-linear regression model would be the best model for predicting Driver Rating?

Challenge:

Create a model given all of the variables from nascar\_df.csv that can predict Driver Rating better than what we have found. Here is the correlation of each variable to get you started. Try to get an adjusted r-squared value higher than 0.9247! Share with your neighbors the best model you made!!

Chart

Description automatically generated