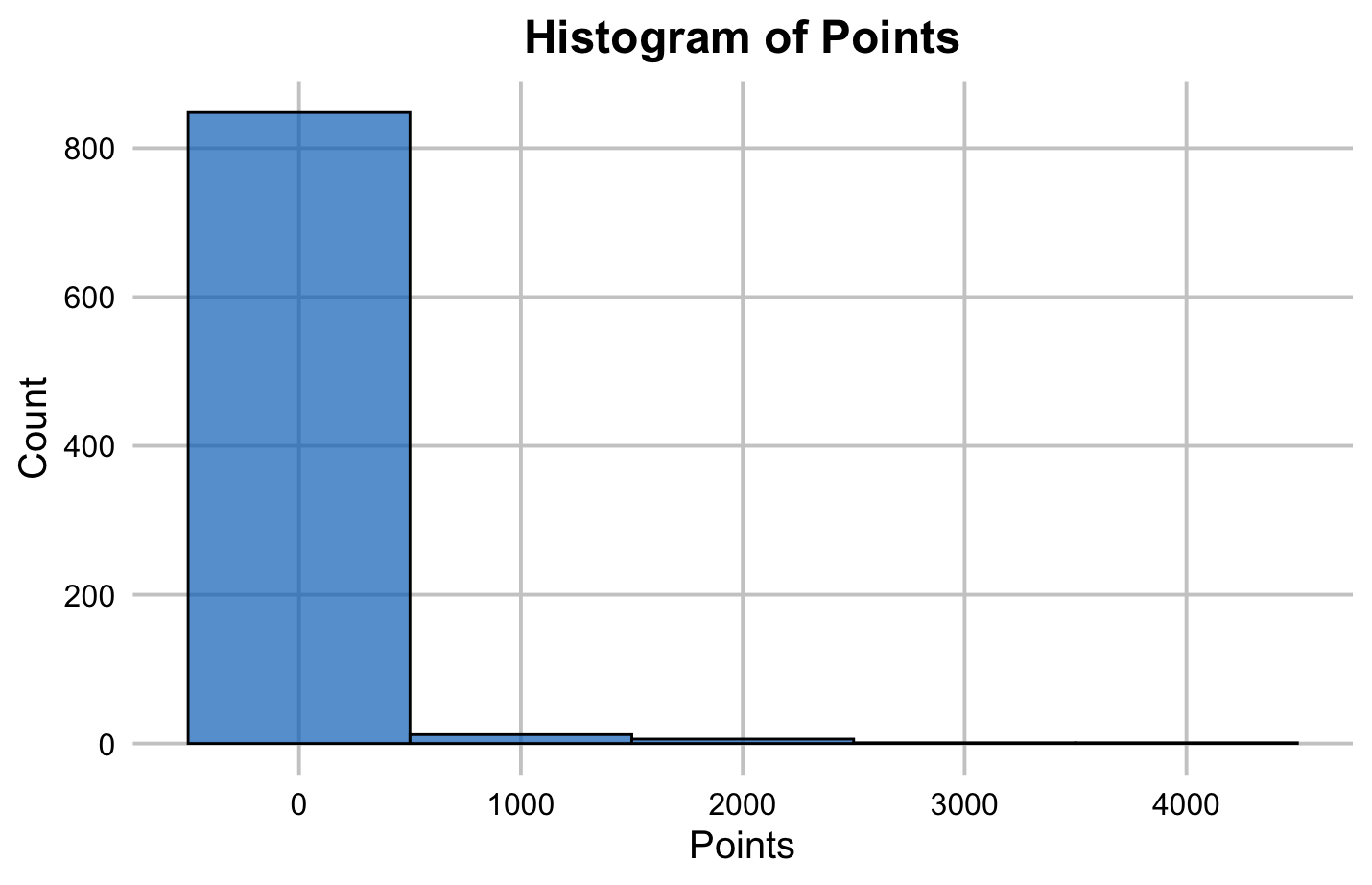
**Evaluating F1 Drivers Statistics**

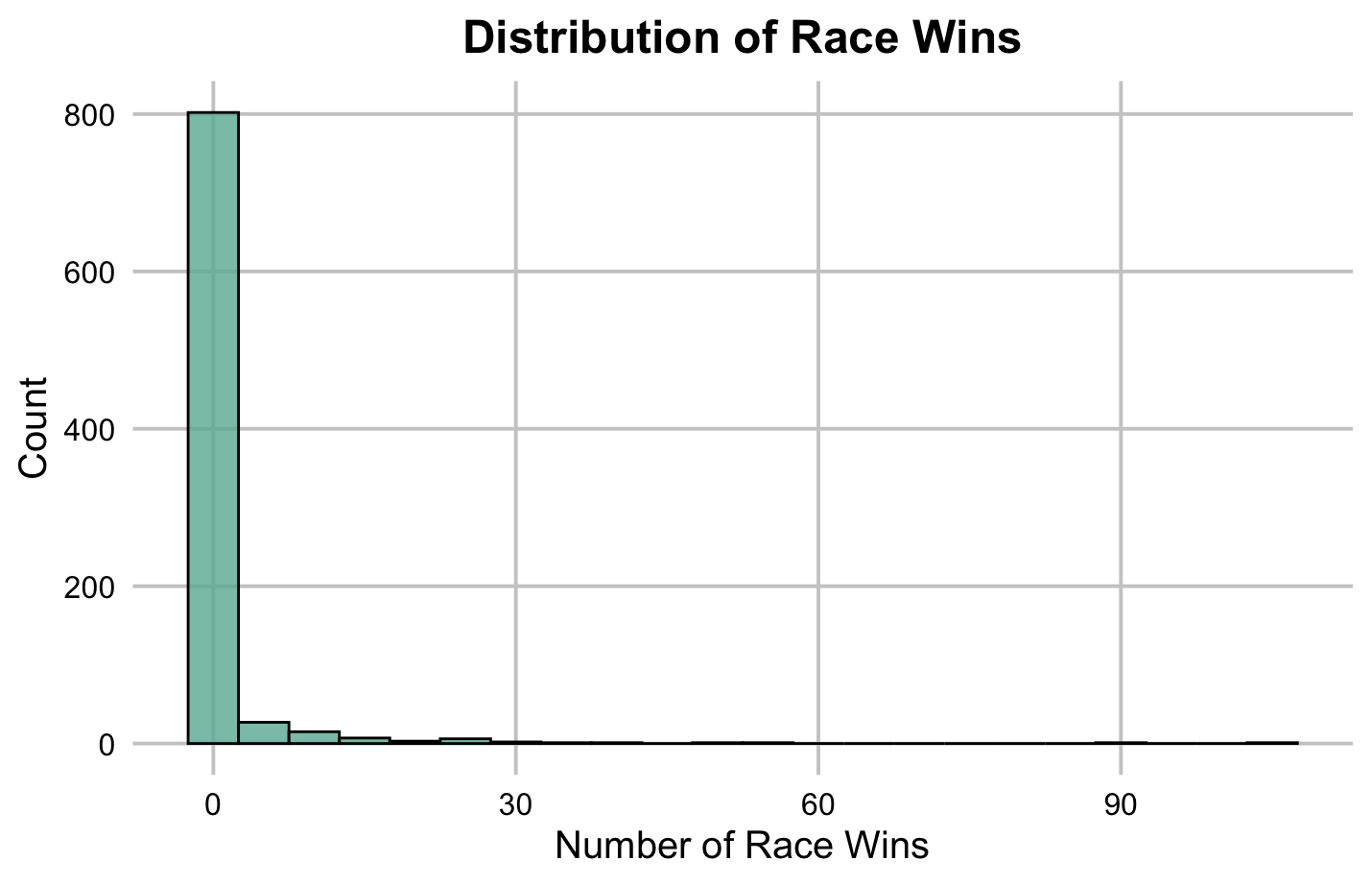
Nirav Naidu

**Introduction**

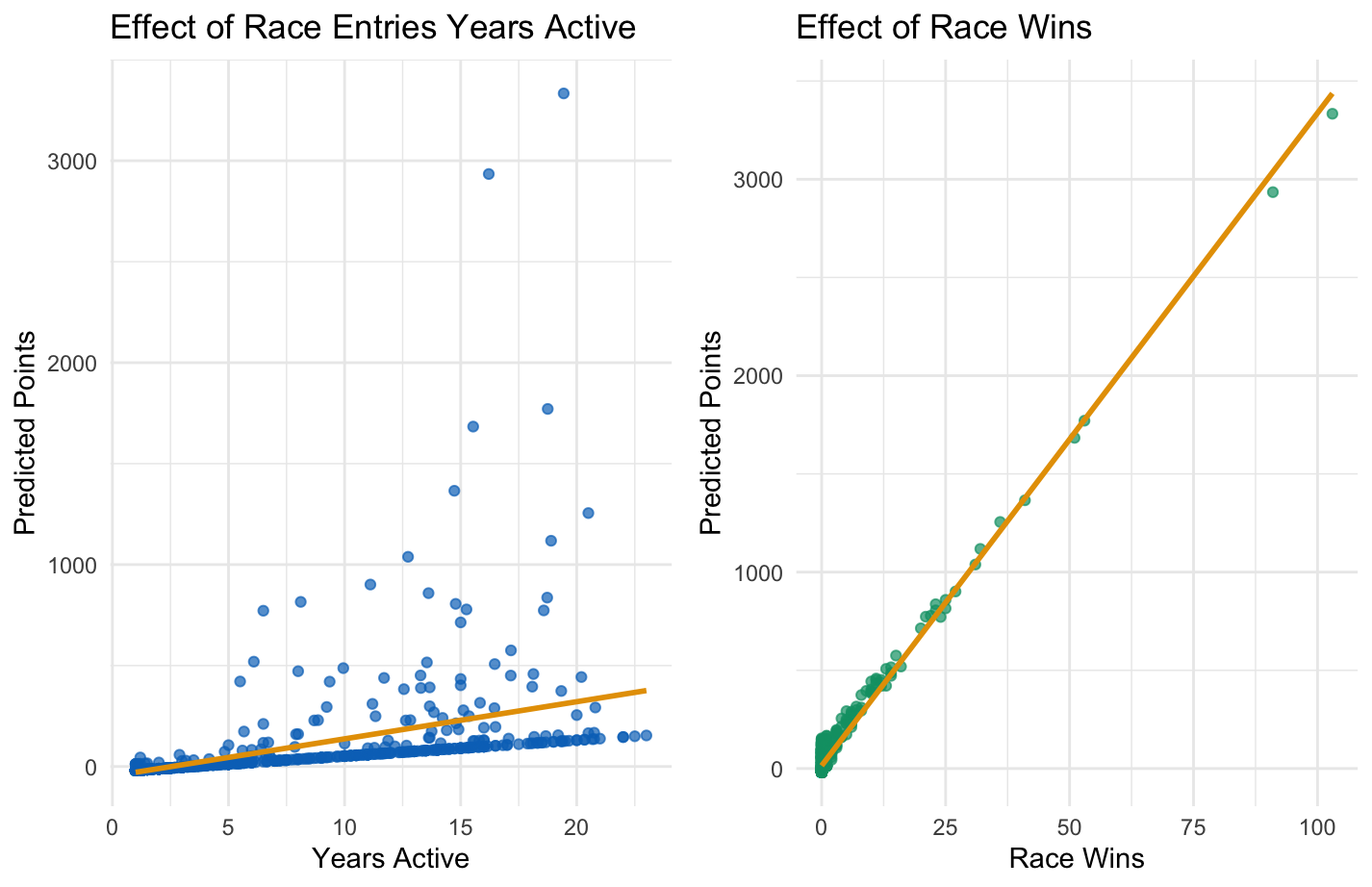
The data set that was utilized is “F1Drivers\_Dataset”, a data set from Kaggle. This data set allows you to see various statistics of Formula 1 (F1) drivers like championships, race wins, race entries, podiums, win rate, race starts, pole positions, etc. There are multiple variables that are being assessed. First, we assessed a relationship between points and race wins. Next, we assessed a multiple regression model using a new variable that we created, a ratio between race entries and years active, along with points and race wins. Finally, we assessed a relationship between race starts, nationality, and race entries.

**Histograms**

Using the histogram to the left, we can see the distribution for driver points in their career. We can see that most drivers fall within the 0-1000 range. There are some drivers that have points that are greater than 2000 points.

Using the histogram to the left, we can see the distribution for race wins in their career. We can see that most drivers fall under 10 race wins, and almost all drivers fall under 30 race wins.

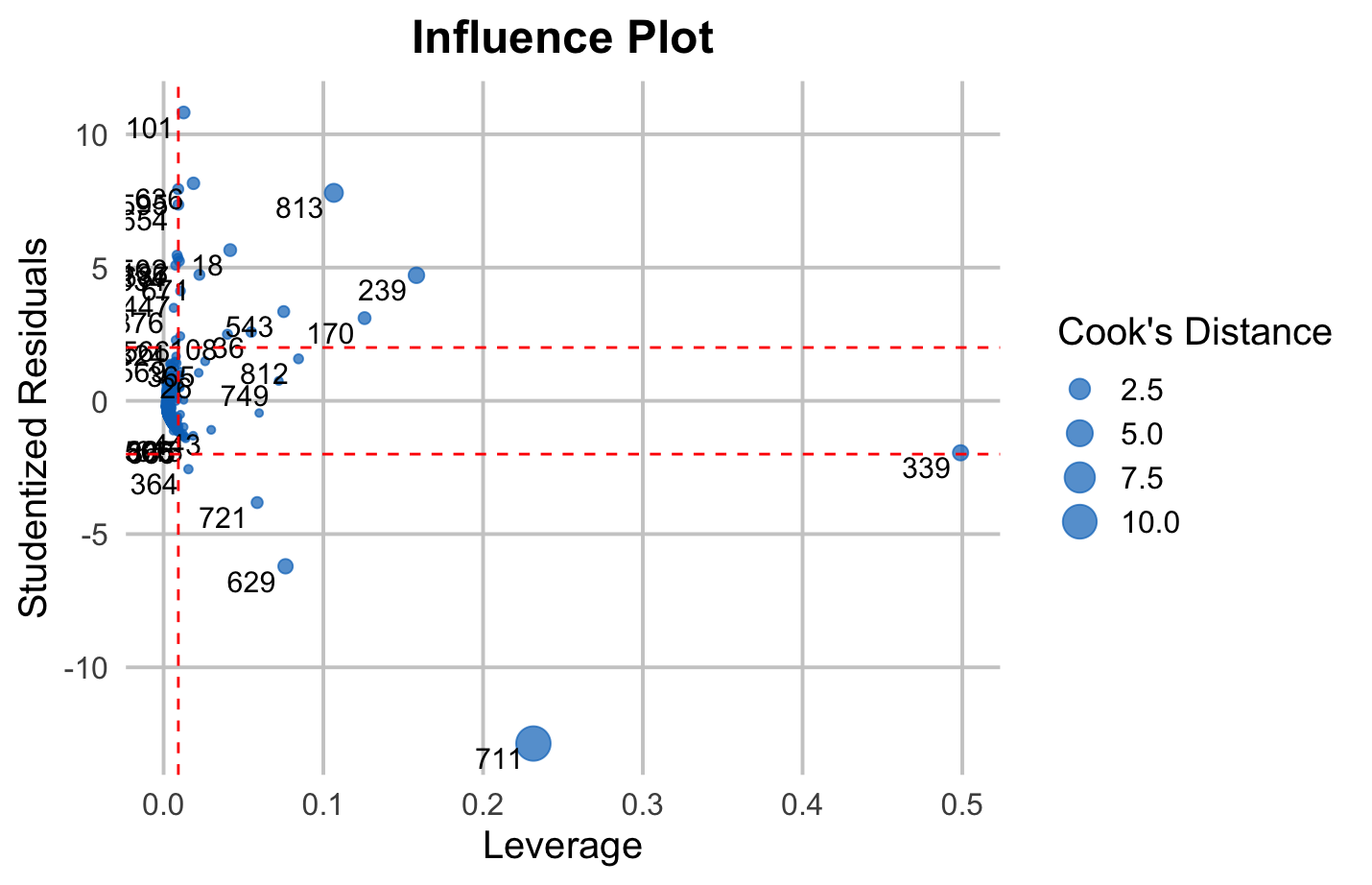
**Multiple Regression Model**



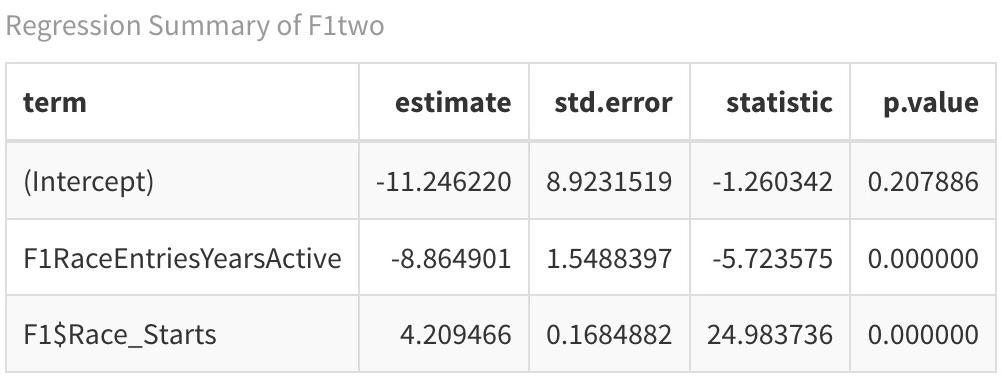
Using the model to the right, we can see the multiple regression model. The new variable that was created is the ratio between race entries and years active. Generally, if a driver has been in F1 for a longer period, they tend to have more race entries.

**Extrapolation Check**

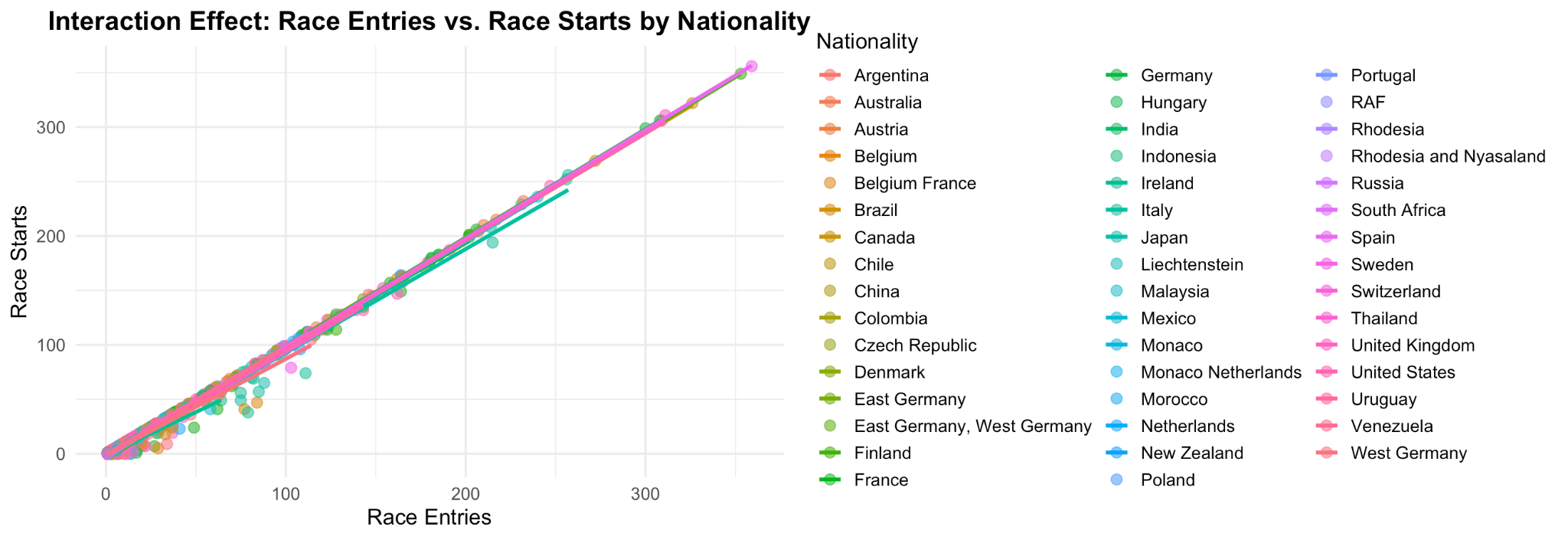
Using the graphic to the left, we can see the first 13 values of the extrapolation check. The whole extrapolation check has 868 rows, which corresponds to all the data in the dataset. The reason why extrapolation is risky is because our predictions can be flat out wrong.

**Influence Plot**

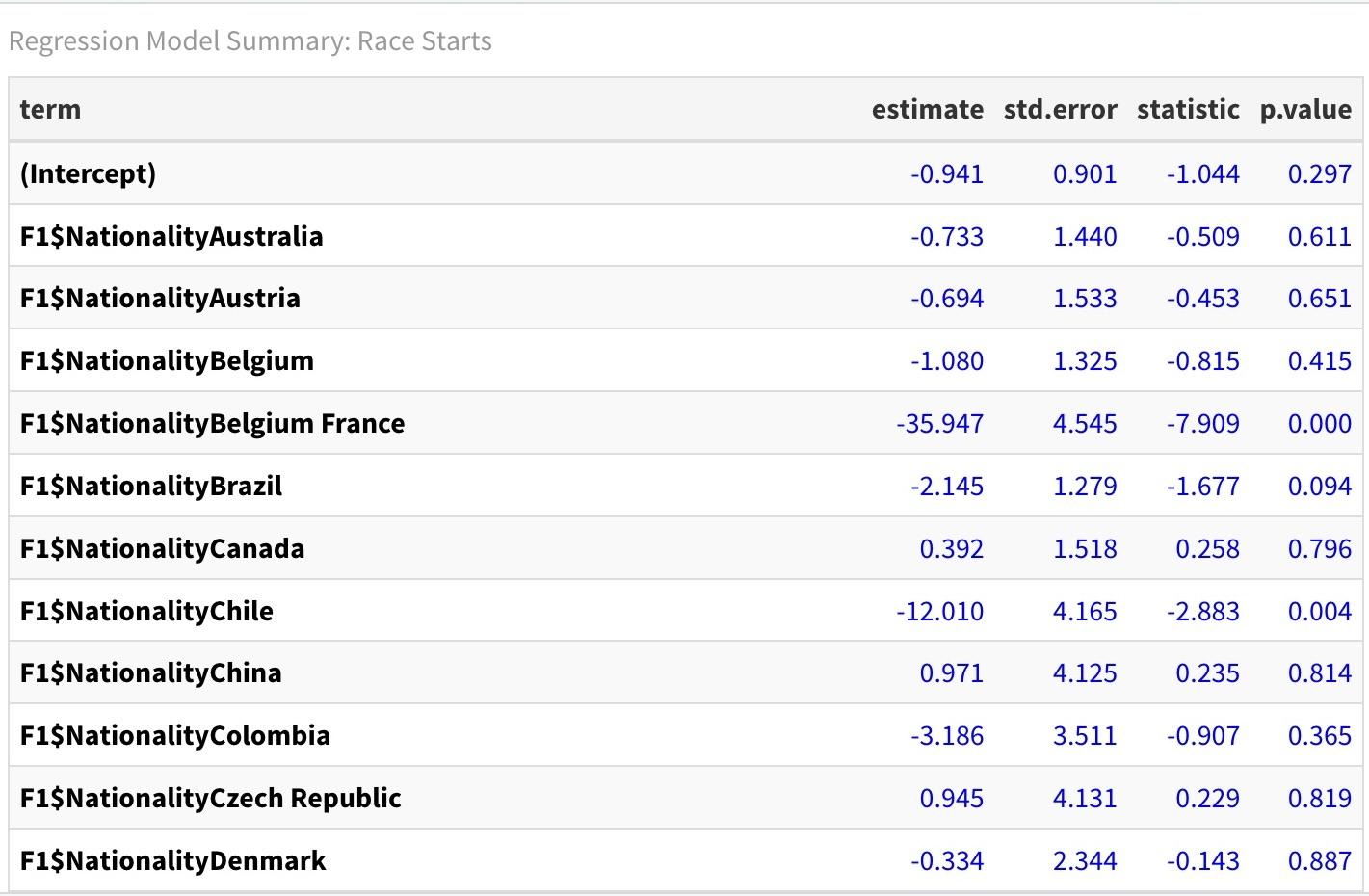
Using the plot to the left, we can see the influence plot of the model. We can see all the points, but the ones we care about are the ones that have high influence. The point that has the highest influence is 711.This means that this point is pulling this model in the negative direction.

**Regression Summary of the Model 1**

The table to the right is the regression summary for model 1. We can see that the p-value for the ratio of race entries to years active is 0, and the p-value for race starts is 0, meaning that both variables are statistically significant and unlikely to happen due to random chance. The r-squared is 0.6841 and the adjusted r-squared is 0.6834. Since this is a multiple regression model, we want to use the adjusted r-squared, as it adjusts to the new variables that were added to the model. This means that 68.34% of the variation in points is explained by the variation in the ratio between race entries and years active and race starts.

**Indicator Variable Model with Interaction**

We can see the plot of the interaction effect of race entries vs. race starts by nationality. Nationality lets us see the number of drivers that entered and started an F1 race. Since this is an interaction, this affects the slope of the line.

**Indicator Variable Model Summary**

This is the summary statistics for the indicator variable model. Since nationality is being used as the categorical variable, there are many levels to this. This summary is not representative of the visualization but is still useful. Countries like Australia, Belgium, and Canada have large p-values, meaning that they do not bring in new or useful information to the model. Chile and Belgium-France have low p-values, meaning that they bring in new and useful information.

**References**

* ChatGPT was used to make the graphics look better.
* A point was earned during class for answering a question.