DSC 423: Data Analysis and Regression Assignment 08

Saransh Thakur

Id: 2020070

Problem 1

“All models are wrong, but some are useful” is an important statement in data science. It is stating that all the models will never accurately represent real-world behavior. It’s telling that if a model cannot accurately describe reality, it can still be very useful if it is close enough.

The first step, in my opinion, is to portray reality accurately and clearly in selecting the data gathered through the model's observation. Furthermore, several pointless mathematics is insufficient to articulate a good model. Furthermore, simply studying the model is not sufficient. We also need to use improved interpretation methods to interpret and apply the model.

Knowing how to interpret the results, in my opinion, is just as important as having a strong numerical procedure. When we use computers to understand the model, we can achieve great accuracy by performing basic procedures, but we also get a lot of errors as a result. As a result, this high precision model is meaningless until the numerical model does not accurately match reality. To manage this issue, we will need human interaction as well as an understanding of how to read the output of a code based on the application and objective we are working on.

Problem 2

1. We employ two ways to test for multicollinearity. We can either use cor or the VIF model to check.
2. **Using cor**

If the variables in the model are highly correlated between pairs of independent variables, the variables are multicollinear.

A screenshot of a computer

Description automatically generated with medium confidence

The putts per round and putting average are highly correlated, which means that we have multicollinearity in the data.

1. **VIF**

If the VIF is more than 10, we can assert that the variables are highly associated, and it explains 90% of the variation in the predictor.

**Text

Description automatically generated**

In this scenario, Puttsperround and putting average variables have a VIF greater than 10, indicating that they are associated with one another.

To reduce multicollinearity, delete one variable at a time that has a high VIF, therefore we will remove PUTTSPERROUND first and then check for model and VIF again.

A screenshot of a computer

Description automatically generated with medium confidence

Text

Description automatically generated

We can see that there is no multicollinearity in the data after deleting putts per round, thus we don't need to remove any more variables.

To get the final model, I delete the variable with the worst p-value, therefore the final model after eliminating the non-significant variable is

A screenshot of a computer

Description automatically generated with medium confidence

For, the final model, the F test has a p-value under a significant value of 0.05 so we fail to reject the null hypothesis and accept the alternative hypothesis that at least one Beta is not equal to zero. Then, for the t-test, all the variable lies under the significance value of 0.05 and the adjusted R-squared is 38.58 % it means that 38.58% of the variability in y is explained by our model.

1. Model from the previous assignment.

A screenshot of a computer

Description automatically generated with medium confidence

For, the final model, the F test has a p-value under a significant value of 0.05 so we fail to reject the null hypothesis and accept the alternative hypothesis that at least one Beta is not equal to zero. Then, for the t-test, all the variable lies under the significance value of 0.05 and the adjusted R-squared is 38.6 % it means that 38.6% of the variability in y is explained by our model.

After applying a log transformation to the model

Table

Description automatically generated

F-test has a very low P-value, this means that we reject the null hypothesis i.e. (all the Betas are equal to zero) and accept the alternative hypothesis that at least one is not equal to zero.

It tells us that something in the model is working.

For, Gir and birdie conversion we reject the null hypothesis and accept the alternative hypothesis, they all lie under the significance value of 0.05.

After applying a log transformation to the model, the model's quality improves, with the value of adjusted R square increasing from 38% to 54%.

To get the final model I remove the variable with the worst p-value, so after pruning the final model is

A screenshot of a computer

Description automatically generated with medium confidence

For, the final model, the F test has a p-value under a significant value of 0.05 so we fail to reject the null hypothesis and accept the alternative hypothesis that at least one Beta is not equal to zero. Then, for the t-test, all the variable lies under the significance value of 0.05 and the adjusted R-squared is 53.82 % it means that 53.82% of the variability in y is explained by our model.

Conclusion: - When we compare this model to the one, we produced in the last assignment, we can see that our final models have significantly improved. Our prior model had four variables: GIR, driving accuracy, birdie conversion, and scrambling; however, our final model has just three variables: GIR, birdie conversion, and scrambling. Our final model is parsimonious since it has fewer parameters and a higher level of goodness of fit than the prior model.

We enhance our adjusted R square from 39 to 54 percent by applying a log transformation, and we pass the f-test and p-test.

1. Residual is the difference between an observed value of the response variable and the value of the response variable predicted from the regression line.

Graphical user interface, text, application, table, Excel

Description automatically generated

Table, calendar

Description automatically generated

Text

Description automatically generated

The sum of the residuals should be Zero.

A picture containing text

Description automatically generated

The Durbin-Watson test is used to determine the independence of residuals.

In this case, we fail to reject the null hypothesis of no independence, hence we can argue that residuals are dependent on each other.

Residual plot for the model

Chart, scatter chart

Description automatically generated

The residual plot is homoscedastic, meaning it has the same variance, 95 percent of values are within two standard deviations, and any above or below two standard deviations are considered outliers. So, we have outliers in this scenario.

Chart, histogram

Description automatically generated

The residual plot is normal and most of the points lie under 2 standard deviations.

Chart, line chart

Description automatically generated

The distribution is normal because most of the observations fall on a straight line.

The residual plot for the dependent variable (Prizemoney)

Chart, scatter chart

Description automatically generated

The residual plot is heteroscedastic, which means it has a different variance, 95 percent of values are within two standard deviations, and any points above or below two standard deviations are considered outliers. So, we have outliers in this scenario.

Now, we will have a residual plot for independent variables (GIR Birdie Conversion, Scrambling).

For, GIR

Chart, scatter chart

Description automatically generated

The residual plot is homoscedastic, which means it has the same variance (by placing a point in empty space), 95 percent of values are between two standard deviations, and any points above or below two standard deviations are deemed outliers. So, we have outliers in this circumstance.

For, Birdie Conversion

Chart, scatter chart

Description automatically generated

The residual plot is homoscedastic, meaning it has the same variance, 95 percent of values are within two standard deviations, and any above or below two standard deviations are considered outliers. So, we have outliers in this scenario.

For, Scrambling

Chart, scatter chart

Description automatically generated

The residual plot is homoscedastic, which means it has the same variance (by placing a point in empty space), 95 percent of values are between two standard deviations, and any points above or below two standard deviations are deemed outliers. So, we have outliers in this circumstance.

1. Outliers are points that deviate from the overall pattern. Outliers on the y-axis have high residuals, while outliers in the x-axis are usually influential spots, implying that removing such points might modify the equation of the line.

Chart

Description automatically generated

Chart, scatter chart

Description automatically generated

Outliers in our model are points that deviate from the overall trend or fall outside the 2 standard deviations.

We can see those points 185,180, and 47 are outliers in the scatter plot above.

To address this issue, remove one outlier at a time and look at the adj-R square, f-test, and residuals plot to see if our model improves or not.

Chart, line chart

Description automatically generated

Point 47 is an influential point for the given issue, which means that deleting it will change the regression line equation. To fix this problem, we will remove the point and study the adj-r square, residuals plot, and confirm that the f-test and t-test are less than 0.05. Check to see whether our model improves; if not, retain it as an observation.