

DSC 423: Data Analysis and Regression

Assignment 08: Residual Analysis

Your submission must include your name and student ID. Your submission must include the honor statement: "I have completed this work independently. The solutions given are entirely my own work." Your submission must be submitted as a PDF.

1. Short Essay (10 points). Read the short PDF on George Box. Explain in your own words the significance of "all models are wrong, but some are useful" as if you were interviewing for job in data science.
2. Previously, you used the PGA tour dataset to predict Prize Money. Use a log transformation to transform Prize Money into a new response variable. Apply your knowledge of regression analysis to fit a regression model using the remaining predictors in your dataset. If necessary, remove the non-significant variables. Remember to remove one variable at a time (variable with largest p-value is removed first) and refit the model, until all variables are significant.
 - a. (10 points) Check for multicollinear. Explain your process.
 - b. (10 points) Compare this model to the one you made in the previous assignment. How did performing a log transformation impact the quality of the model? Why?
 - c. (10 points) Analyze and discuss the residual plots.
 - d. (10 points) Analyze if there are any outliers and/or influential points. If there are points in the dataset that need to be investigated, give one or more reason to support each point chosen. Discuss your answer.