Rental Bikes Prediction Final Project

by Maisy Song

Abstract

As the world becomes harder to own a car both physically and financially in a big metropolitan area, people start to prefer public transportation more. This project was about data analysis of the number of bike rentals in Seoul, South Korea. R Studio was used to analyze the dataset and the goal was to use linear regression in R to extract an accurate model to predict the number of bike rentals. Total of three models were set up initially and performed model selection to select the most accurate model.

Introduction

The objective of this data analysis project was to predict bike rentals per hour in Seoul, South Korea with a given bike rental dataset. This dataset of rental bikes has a total of 15 variables and 544 observations.

The dataset used was a subset of the original data that was considered more important and relevant. The subset of the given data included the following variables: Month, the month of the data recorded(1-12), TemperatureC, the temperature of the hour in Celsius, Rented.Bike.Count, which is the number of bikes rented each hour, Hour, which is the recorded hour of the day(0-23), Humidity, the percentage of humidity, Wind.speed.ms, measured in m/s, Rainfallmm, measured in mm, Solar.Radiation..MJ.m2., solar radiation of the hour measured in MJ/m^2, and the Seasons the four seasons: Autumn, Spring, Summer, Winter.

Results

Section 1. Models and Model Selection

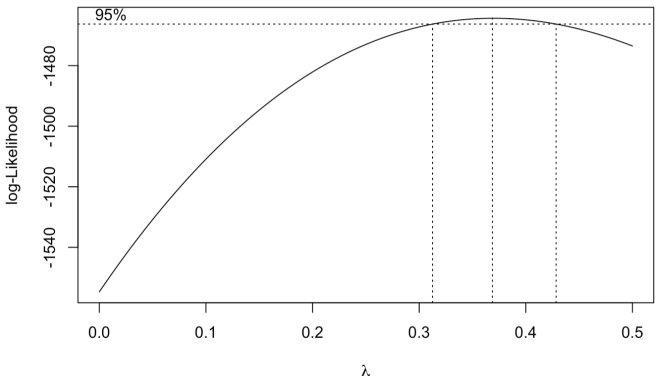
The first saturated model was a fitted linear model with additive first-order terms of all the variables in the dataset. The first model was *Rented.Bike.Count = Month +TemperatureC + Hour + Humidity + Wind.speed.ms + Rainfallmm + Solar.Radiation..MJ.m2. + Seasons*. From the summary of the first model, the estimated slope for the variable 'TemperatureC' was 22.6314. Every 1 degree increase in temperature, the number of rented bikes per hour is estimated to increase 22.6314 with the rest of the variables held constant. The 'Seasons' variable had four categories of 'Autumn', 'Spring', 'Summer', and 'Winter'. The baseline of the variable was 'Autumn'. When the 'Seasons' variable was 'Spring', the number of rented bikes per hour was estimated to decrease 189.7555, with all other variables held constant. Same goes for 'Summer' with the estimated decrease value of 125.8787 and 'Winter'

with the estimated decrease value of 464.4512. The adjusted R-squared value of the first model was 0.5665. The additive second model was *Rented.Bike.Count = TemperatureC + Hour + Humidity + Rainfallmm + Solar.Radiation..MJ.m2 + Seasons* which was the result of eliminating couple variables that have higher p-values than $\alpha = 0.1$. The adjusted R-squared model of the second model was 0.5681 which shows that the elimination of variables made the linear model more accurate. Another model taken in for consideration was a linear model with added interaction terms of 'Hour' with 'TemperatureC' and 'TemperatureC' with 'Humidity': *Rented.Bike.Count = Hour + TemperatureC + Humidity + Solar.Radiation..MJ.m2.. + Rainfallmm + Seasons + (TemperatureC)(Humidity) + (Hour)(TemperatureC)*. The adjusted-R-squared value for the third model was 0.6183, which was higher than that of the second model.

Method of model selection that was used was the ANOVA test. ANOVA test with the second model and third model was performed to see if addition of interaction terms is necessary. With a p-value of 1.862e-15, the third model with the addition of two interaction terms was selected over the second model with only first-order terms.

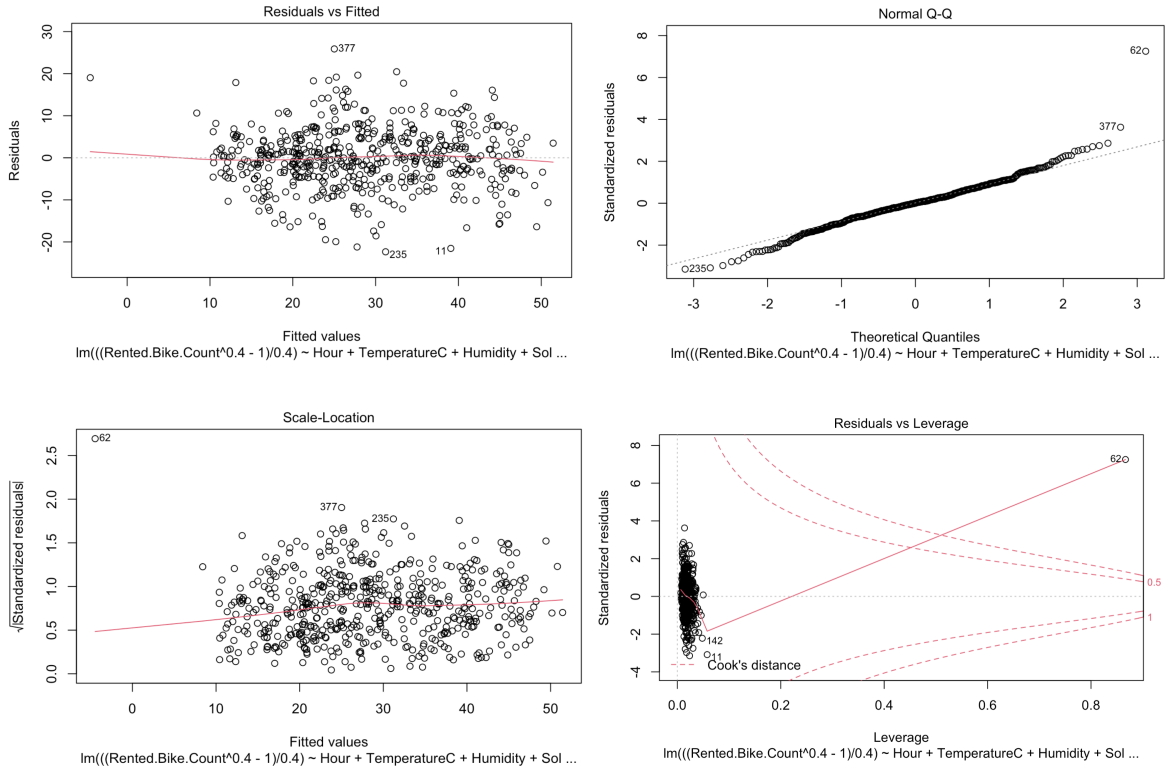
Section 2. Model Transformation

The Box-Cox transform assessed whether the model/variables need any transformation. It evaluated a lambda value where the response variable was transformed according to it. The following graph is a Box-Cox transform graph for the third model:



This Box-Cox graph shows that the lambda is near $\lambda = 0.4$. According to this lambda value, the transformed model is the following: *((Rented.Bike.Count$^{0.4}$ - 1)/0.4) = Hour + TemperatureC + Humidity + Solar.Radiation..MJ.m2.. + Rainfallmm + Seasons + (TemperatureC)(Humidity) + (Hour)(TemperatureC)*

To analyze the transformed model, four graphs were plotted with the 'plot()' function in R.



From the first graph, Residuals vs Fitted, the fitted values seem reasonable to assume that there is a true linear relationship between the predictor variables and the transformed response variable due to randomly scattered points and average of the points being flat and near 0. It was also reasonable to say that there was an equal variance. The second graph, Normal Q-Q, showed that the assumption that the true error terms are normally distributed seems reasonable. From the Scale-Location graph, the red line was flat so it was reasonable to assume the transformed response variable has an equal variance. The last graph assessed the existence of any unusual points. It seemed like there were some points that had high residuals. There is one influential point, point 62, where it had both a high standardized residual and high leverage.

Section 3: Comparing First Model and Final Model

Variance Inflation Factors(vif) function in R is a function used to calculate the collinearity of the variables in the linear model. The Variance Inflation Factors for the first model were calculated to be 1.084704 for ‘Month’, 1.358481 for ‘TemperatureC’, 1.120060 for ‘Hour’, 1.695711 for ‘Humidity’, 1.328152 for ‘Wind.speed.ms’, 1.057349 for ‘Rainfallmm’, and 1.703043 for ‘Solar.Radiation..MJ.m2’. These values do not show any concerns since all factors were all slightly above 1. The Variance Inflation Factors for the final model were 2.113449 for ‘Hour’, 16.797243 for ‘TemperatureC’, 2.676192 for ‘Humidity’, 1.692868 for ‘Solar.Radiation..MJ.m2’, 1.080575

for ‘Rainfallmm’, 13.574262 for ‘TemperatureC*Humidity’, and 5.347797 for ‘Temperature*Hour’. There were some concerns for collinearity with variables, ‘TemperatureC’ and ‘TemperatureC*Humidity’ since both values were higher than 10. It was reasonable that these values are high enough to raise a concern for collinearity because of the interaction term ‘TemperatureC*Humidity’.

Finally, F-test was performed with the final(transformed) model with the following hypotheses:

H_o: $\beta_i = 0$ where $i = \{0, 1, \ldots, 10\}$

H_a: at least one of the coefficients are not 0

There was sufficient evidence to reject the null hypothesis at $\alpha = 0.01$ with the p-value of 2.2e-16 and conclude that there was at least one of the coefficients that is not 0. The F-statistic is also quite big, 102.9, which as also greater than the F-statistic from the first model, 71.97. The adjusted R-squared for the final model shown above was 0.6524, whereas the reported adjusted R-squared value for the first model was 0.5665.

Discussion

The adjusted R-squared value shows that the final model predicts the number of rented bikes more accurately than the first model. The performed tests in this project was to find a model that accurately predicts the number of rented bikes on a given day. I started from the first saturated model with all the first-order of the variables. For this first model, the number of coefficients was not necessarily a concern for the size of the dataset because there were more than 500 observations in the dataset. With trial and error, I added and eliminated many variables, interaction terms, and/or polynomial terms to see if the model would predict response variable better. I came to a conclusion that the model with the two interaction terms ‘TemperatureC*Humidiy’ and ‘TemperatureC*Hour’ was a more accurately fitting model than any other models I fitted.

I think the bikes data was a good sample of the population because there were more than 500 observations. However, I think the data itself was really scattered, since riding bikes is usually an outdoor activity and the data was taken all year round. For example, the Residuals vs. Leverage graph showed a good amount of points that are considered high residuals. This was because since riding bikes is an outdoor activity, it heavily depends on the weather, especially Next time, I would try to focus more on a specific time of the year or specific time of the day.