Project: Regression Analysis

The “Seoul Bike Sharing Demand Data Set”, sourced from the official website of the South Korean government, pertains to bicycle sharing rentals. The original dataset, which can be found at https://archive.ics.uci.edu/ml/datasets/Seoul+Bike+Sharing+Demand comprises 13 independent variables along with one dependent variable. However, due to multicollinearity and the inclusion of a date variable, three variables have been excluded, leaving us with 10 independent variables, where two of them are dummy variables, and one dependent variable.

Given that the original dependent variable does not follow a normal distribution and includes a significant number of zeros, we’ve made modifications to enhance the data’s statistical properties. Specifically, entries containing zeros were removed, and a logarithmic transformation was performed. This results in the current dataset “Bikedata.csv”, where the dependent variable is the natural logarithm of the number of bike rentals log.Rented.Bike.Count.

The remaining variables can be summarized as follows: log.Rented.Bike.Count : logarithm of the count of bikes rented in each hour Hour : Hour of the day

Temperature : Temperature in Celsius ($^{\circ}C$)

Humidity : Humidity in percentage (%)

Windspeed : Windspeed (m/s)

Visibility : Visibility ($10m$)

Solar radiation- Megajoules per square meter MJ/m^2

Rainfall- Millimeter (mm)

Snowfall- Centimeter (cm)

Seasons- Winter, Spring, Summer, Autumn

Holiday - Holiday/No holiday

Tasks:

1. Briefly describe the relationship of the data using descriptive analysis. (i.e. scatter plot or correlation plot)

2. Determine a linear regression model of the log.Rented.Bike.Count based on all other given variables.

3. Find a suitable subset of explanatory variables for the log.Rented.Bike.Count. Summarize the regression results, including parameter estimates, statistical significance (*p*-values), confidence intervals, and a goodness-of-fit measure in a table. Interpret your results.

   Hint: For model selection, you may consider employing forward selection, backward selection, or best subset selection. These strategies can be further guided by criteria such as AIC, BIC, *adjusted-R*^{2}, or Mallows' *Cp* values. There is no standard correct model here; give a reasonable explanation of your selection process.

4. Using the selected model from task 3, create residual plots for model evaluation. Analyze these plots to check for patterns of linearity, heteroskedasticity, and normality. Further check the multicollinearity using the variance inflation factor (VIF). Briefly discuss problems that your final model may have.