TU Dortmund

Introductory Case Studies

Project III: Regression Analysis

Lecturers:

Prof. Dr. Katja Ickstadt

M. Sc. Zeyu Ding

M. Sc. Yassine Talleb

Author: Bushra Tariq Kiyani

Group number: 2

Group members: Sarmistha Bhattacharyya , Jatin Rattan , Vikas Kumar

July 6, 2023

Contents

1 Introduction

In recent years, bike-sharing systems have received a lot of attention as an environmentally friendly and cost-effective alternative to traditional transportation methods, especially in urban areas. Analyzing bike rental data can help decision-makers understand user behavior, optimize the allocation of resources, and improve the overall efficiency of the system. Variables such as hour, temperature, humidity, wind speed, visibility, solar radiation, rainfall, snowfall, seasons, and holidays can affect the demand for rental bikes. This information is very valuable for city planners, policy makers, and bike-sharing service providers to improve infrastructure, optimise service operations, and effectively meet users' needs (Julio et al., 2022).

This project aims to predict the number of bikes required at each hour for the stable supply of rental bikes using regression analysis. The goal of this project is to develop a formula that predicts the number of bike rented in one hour from independent variables and find the best combination of explanatory variables. The first step involves fitting a linear regression model to understand the relationship between the number of rented bikes and the given variables. The selection of the optimal set of explanatory variables is determined using the selection criteria (AIC, BIC). Subsequently, model assumptions are verified, the coefficients of the model are interpreted, and their statistical significance is assessed. Finally, confidence intervals for the regression parameters are computed, and the goodness of fit is evaluated.

The second section describes the structure and quality of the dataset in more detail. Additionally, the goals of the project are stated in the second section. The third section explains the Multiple linear regression, Backward Selection, with the criteria AIC and BIC, Goodness-of-fit (adjusted $R-$squared), VIF and Residual plot. The fourth section focuses on the application of these methods and the interpretation of the graphs and results. Finally, the fifth section summarizes the most important findings.

2 Problem statement

2.1 Dataset and Data Quality

This report deals with the analysis of a small sample of the dataset "Seoul Bike Sharing Demand" taken from the official website of the South Korean government pertains to

bicycle sharing rentals. The South Korean government oversees public bike rental services in several cities across the country. These services are managed by the respective city governments and allow bicycles to be rented (Seoul Bike Sharing Demand, South Korean Govt., 2020).

The dataset contains information about the count of public bikes rented per hour in the Seoul Bike Sharing System, along with corresponding time, weather and holiday information. It consists of 13 independent variables along with one dependent variable. However, due to multicollinearity and the inclusion of a date variable, three variables have been excluded by the lecturers of the course, leaving us with 10 independent variables. Moreover, because the original dependent variable does not follow a normal distribution and includes a significant number of zeros, the course lecturers made modifications to enhance the data's statistical properties. Specifically, entries containing zeros are removed, and a logarithmic transformation is 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 with 2905 observations. Table 1 shows the description of all variables:

Variable	Type	Description
log.Rented.Bike.Count	Numeric	Logarithm of the count of bikes rented in each hour
Hour	Numeric	Hour of the day
Temperature	Numeric	Temperature in Celsius (°C)
Humidity	Numeric	Humidity in percentage (%)
Windspeed	Numeric	Windspeed (m/s)
Visibility	Numeric	Visibility $(10m)$
Solar radiation	Numeric	Megajoules per square meter MJ/m^2
Rainfall	Numeric	Millimeter (mm)
Snowfall	Numeric	Centimeter (cm)
Seasons	Nominal	Winter, Spring, Summer, Autumn
Holiday	Nominal	Holiday, No Holiday

Table 1: Data Description

The dataset is sourced from the South Korean government and is available on the UCI Machine Learning Repository. As an official government dataset, it has a certain level of credibility and reliability. The dataset is also released under a Creative Commons Attribution 4.0 International (CC BY 4.0) License. Overall, the Korean government's data on bike-sharing rentals appears to be of good data quality, with completeness, accuracy, consistency, relevance, and a trusted source.

2.2 Project Objectives

The goal of this project is to find a model that can effectively explain number of bike rentals and determine the best set of explanatory variables using backward selection method. The model's coefficients will be interpreted, their statistical significance will be assessed, and confidence intervals for the regression parameters will be provided. Additionally, the goodness of fit will be evaluated.

To achieve this, first, a complete linear regression model will be fitted using all the explanatory variables that are relevant in explaining bike rentals. The statistical significance of the parameters will be evaluated, and a linear regression equation will be developed. Next, the subset of explanatory variables will be determined by the AIC and the BIC as the selection criteria and we compare the included explanatory variables of the two resulting models. Then, one selection criteria will be selected, the coefficients of the model will be interpreted, and their statistical significance will be assessed. Confidence intervals for the regression parameters will also be provided. The model assumptions will be verified by examining the Q-Q plot and residual plot of the estimated model. This step ensures that the model is appropriate and accurately represents the data. Finally, the goodness of fit of the model will be evaluated, specifically by assessing the adjusted R-squared. Which will help determine how well the model fits the data.

3 Statistical methods

3.1 Multiple Linear Regression

The linear regression model is a statistical modelling method used to model a relation between a continuous variable of interest Y called the response variable and a set of explanatory variables $x_1, ..., x_k$, which can be continuous or categorical. This relation between Y and $x_1, ..., x_k$ is modeled using a function denoted as $f(x_1, ..., x_k)$ in the presence of additive errors, leading to the following equation:

$$Y = f(x_1, ..., x_k) + \epsilon$$

Where ϵ is a random variable, and the function f is a linear combination of covariates, given by:

$$f(x_1, ..., x_k) = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k$$

The parameters $\beta_0, \beta_1, ..., \beta_k$ are unknown and need to be estimated. The parameter β_0 corresponds to the intercept term. By combining the covariates and the unknown parameters into $p = k + 1$ dimensional vectors, denoted as $\mathbf{x_i} = (1, x_{i1}, ..., x_{ik})'$ and $\boldsymbol{\beta} = (\beta_0, ..., \beta_k)'$, the relationship can be expressed:

$$\hat{y}_i = \mathbf{x_i}'\boldsymbol{\beta} + \epsilon_i$$

Considering a total of n observations, the model can be represented using vector notation. We define the vectors $\mathbf{y}, \boldsymbol{\epsilon}$, and the design matrix $\mathbf{X}$ as follows:

$$\mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \quad \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1k} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{pmatrix} = \begin{pmatrix} \mathbf{x}_1' \\ \vdots \\ \mathbf{x}_n' \end{pmatrix}$$

Then we can write the model as (Fahrmeir et al., 2022, p. 74-75):

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

3.1.1 Model Assumptions

The following assumptions about the ϵ are made (Fahrmeir et al., 2022, p. 75-76):

Expectation of the errors: The errors are assumed to have a mean of zero, $E[\epsilon_i] = 0$.

Homoscedastic error variances: It is assumed that the errors exhibit a constant variance σ^2 across observations, referred to as homoscedastic errors.

Uncorrelated errors: The assumption of uncorrelated errors implies that the covariance between errors for different observations is zero. These assumptions lead to the covariance matrix $\text{Cov}(\boldsymbol{\epsilon}) = E(\boldsymbol{\epsilon}\boldsymbol{\epsilon}') = \sigma^2\mathbf{I}$, where $\mathbf{I}$ is the identity matrix. This suggests that the errors are stochastically independent of the covariates.

Full column rank of the design matrix X: The assumption is made that the design matrix possesses a full column rank, denoted as $\mathbf{rk}(\mathbf{X}) = k + 1 = p$. This condition guarantees the linear independence of the columns in the design matrix $\mathbf{X}$.

Gaussian errors: It is assumed that the errors follow a normal distribution to facilitate the construction of confidence intervals and hypothesis tests for the regression coefficients. Given the additional assumptions of homoscedastic error variance and un-

correlated errors, we can state that $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ or, in matrix notation, $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$. Consequently, the response variable $\mathbf{Y}$ follows a normal distribution of $\mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$.

3.1.2 Dummy Coding

A dummy variable is a numeric variable used to represent categorical data, such as gender or region. In regression analysis, usually the linear effect of continuous covariates on the response variable is examined. Dummy coding is a technique used to incorporate categorical variables into regression models. When modeling the effect of a covariate x with c categories using dummy coding, $c-1$ dummy variables are defined as follows:

$$x_{i1} = \begin{cases} 1, & \text{if } x_i = 1 \\ 0, & \text{otherwise} \end{cases} \quad ... \quad x_{i,c-1} = \begin{cases} 1, & \text{if } x_i = c-1 \\ 0, & \text{otherwise} \end{cases}$$

for $i = 1, ..., n$. A modified linear function can be formulated as follows:

$$f(x_1, ..., x_k) = \beta_0 + \beta_1 x_{i1} + ... + \beta_{i,c-1} x_{i,c-1} + ... + \beta_k x_k$$

To ensure identifiability, one of the dummy variables is omitted. This category is referred to as the reference category. By directly comparing the estimated effects to the reference category, we can interpret them (Fahrmeir et al., 2022, p. 95-97).

3.1.3 Estimation

The estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\sigma}^2$ are represented as $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\sigma}}^2$, respectively.

Regression parameters estimation: Estimation of regression parameters usually involves the use of **least squares**. According to this principle, the unknown regression coefficient $\boldsymbol{\beta}$ is estimated by minimizing the sum of squared deviations between the true response value y_i and the predicted value $\hat{y}_i = \mathbf{x}_i'\hat{\beta}$.

$$\text{LS}(\beta) = \sum_{i=1}^{n} (y_i - \mathbf{x}_i'\beta)^2$$

in Matrix notation: $(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$. The estimation of $\boldsymbol{\beta}$ involves minimizing the LS($\boldsymbol{\beta}$) by setting its first derivative to zero and solving for $\boldsymbol{\beta}$. This estimator is (Fahrmeir et al., 2022, p. 105)

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

Predicted values and residuals: Estimator for the mean of y_i can be derived as:

$$\widehat{E[y_i]} = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + ... + \hat{\beta}_k x_{ik} = \mathbf{x}_i' \hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

Typically, $\widehat{E[y_i]}$ is referred to as $\hat{y}_i$. The difference between the true value y_i and the estimated value $\hat{y}_i$, expressed as $\hat{\epsilon}_i$, is called the residual (Fahrmeir et al., 2022, p. 107):

$$\hat{\boldsymbol{\epsilon}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}$$

Estimation of the error variance: The maximum likelihood estimator of the error variance, as described in (Fahrmeir et al., 2022, p. 108), is given by:

$$\hat{\sigma}^2 = \frac{\hat{\boldsymbol{\epsilon}}\hat{\boldsymbol{\epsilon}}'}{n}$$

However, this estimator is biased. As an alternative, the Restricted Maximum Likelihood Estimator (REML) is commonly used for estimating σ^2, which is expressed as:

$$\hat{\sigma}^2 = \frac{1}{n-p}\hat{\boldsymbol{\epsilon}}\hat{\boldsymbol{\epsilon}}'$$

3.1.4 Hypothesis Testing and Confidence Intervals for $\hat{\beta}_j$

By assuming $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ that errors are independently and identically normally distributed, the relationship between the independent and dependent variables can be assessed through hypothesis tests of the regression coefficients $\hat{\beta}_j$, where $j = 1, ..., k$. In these tests, we evaluate the significance of each covariate's effect on the response variable. The null hypothesis H_0 asserts that the j^{th} covariate has no impact on the response variable, while the alternative hypothesis H_1 posits that the j^{th} covariate does have a significant effect on the response variable (Fahrmeir et al., 2022, p. 125-139).

$$H_0 : \hat{\beta}_j = 0 \quad \text{and} \quad H_1 : \hat{\beta}_j \neq 0, \qquad j = 1, ..., k$$

t-test: In multiple linear regression, the t-test examines the significance of each coefficient by comparing it to a t-distribution with $n - p$ degrees of freedom, where n is the number of observations and p is the number of predictors. The test statistic t_j is:

$$t_j = \frac{\hat{\beta}_j}{\widehat{se}_j}$$

Here, $\hat{\beta}_j$ is the estimated regression coefficient for the jth covariate, and $\widehat{se}_j$ denotes the estimated standard deviation or standard error of $\hat{\beta}_j$. To determine the significance of the regression coefficient, the p-value of t_j is calculated. The p-value is the probability of observing a test statistic as extreme as t_j, assuming the null hypothesis is true. If the p-value is less than a predetermined significance level α, typically 0.05, the null hypothesis H_0 is rejected (Rasch et al., 2020, p. 333-334).

Confidence intervals for $\hat{\beta}_j$: Under the assumption of normally distributed errors, confidence intervals for the $\hat{\beta}_j$ provide a range of plausible values for β_j with a specified level of confidence. The confidence interval, using standard errors $\widehat{se}_j$ of $\hat{\beta}j$, and the critical value $t_{n-p}\left(1-\frac{\alpha}{2}\right)$, which corresponds to the $(1-\alpha)$ quantile of the t-distribution with $n-p$ degrees of freedom, is given by (Fahrmeir et al., 2022, p. 125-139):

$$\left[\hat{\beta}_j - t_{n-p}\left(1-\frac{\alpha}{2}\right).\,\widehat{se}_j, \hat{\beta}_j + t_{n-p}\left(1-\frac{\alpha}{2}\right).\,\widehat{se}_j\right]$$

3.2 Backward Selection Method

Backward selection starts with the full model including all predictors and eliminates the least important variables until a stopping criterion is met. The stopping criteria can be based on a predetermined significance level, a specific number of predictors to retain, or a chosen model evaluation criterion (e.g., AIC). The process continues until the stopping criteria is met, resulting in a final model with the most impactful predictors (Fahrmeir et al., 2022, p. 151).

3.3 The Selection Criteria

The Akaike information criterion (AIC) is a widely used criterion for model selection in the context of likelihood-based inference. It provides a measure of the trade-off between the goodness of fit of the model, and the complexity of the model and defined as:

$$\text{AIC} = -2l(\hat{\beta}_M, \hat{\sigma}^2) + 2(|M| + 1)$$

Where M is number of the covariates, $\hat{\beta}_M$, and $\hat{\sigma}^2$ are the ML estimators, $l(\hat{\beta}_M, \hat{\sigma}^2)$ represents the maximum value of the log-likelihood. A lower AIC value indicates a better fit of the model. The Bayesian Information Criterion (BIC) is similar to that of

the AIC. It is defined as:

$$\text{BIC} = -2.l(\hat{\beta}_M, \hat{\sigma}^2) + log(n)(|M| + 1)$$

Again, smaller values indicate better model fit, with the main difference being that BIC penalises complex models more than AIC. Hence, the resulting models are generally more parsimonious when using BIC rather than AIC (Fahrmeir et al., 2022, p. 148-149).

3.4 The Coefficient of Determination (Adjusted R-squared)

R-squared represents the proportion of explained variance in the regression model.

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{\sum_{i=1}^{n}\hat{\epsilon}^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

A higher coefficient (closer to 1) of determination indicates a better fit. Adjusted R^2 penalizes adding independent variables in the model (Fahrmeir et al., 2022, p. 112).

$$\tilde{R}^2 = 1 - \frac{(1 - R^2)(n - 1)}{n - k - 1}$$

The main difference between R^2 and $\tilde{R}^2$ is that adding an independent variable to the model increases R^2 even if the independent variable is insignificant. The $\tilde{R}^2$ increases only when significant independent variables have an impact on the dependent variable.

3.5 The Variance Inflation Factor (VIF)

The Variance Inflation Factor (VIF) measures multicollinearity in regression. Multicollinearity can inflate the variance of the regression coefficients and thus negatively affect the reliability of the regression coefficients. The formula for the VIF for the jth predictor is:

$$\text{VIF}_j = \frac{1}{1 - R_j^2}$$

where R_j^2 is the R-squared value when the jth predictor is regressed on the other predictors. A VIF value of 1 suggests no correlation between the jth predictor and the other predictors, indicating that the coefficient's variance is not inflated. Typically, a VIF greater than 5 or 10 indicates high multicollinearity (Fahrmeir et al., 2022, p. 158).

3.6 Interpretation of Parameter Estimates

The sign of a linear regression coefficient indicates the direction of the relationship between the explanatory variable and the dependent variable. A positive coefficient indicates a positive relationship in which an increase in the explanatory variable corresponds to an increase in the dependent variable. Conversely, a negative coefficient indicates an inverse relationship, where an increase in the explanatory variable leads to a decrease in the dependent variable. The coefficient value is the magnitude of the change in the dependent variable associated with a one-unit shift in the independent variable, while keeping other variables in the model constant (Fahrmeir et al., 2022, p. 107).

3.7 Residual Plot

A residual plot plots residuals $\hat{\epsilon}$ on the y-axis against the predicted values $\hat{y}_i$ on the x-axis. The ideal residual plot shows a random scatter of points around zero with a constant variance. If the points exhibit an increasing, decreasing or non-constant fluctuation, then the variance is not constant (heteroscedastic variance). The linearity of a model function can also be checked through residual plots. If the points form a pattern then the model function is incorrect. If points are s cattered equally above and below the average line then $E[\boldsymbol{\epsilon}] = \mathbf{0}$ (Fahrmeir et al., 2022, p. 79-80).

4 Statistical analysis

This section presents a descriptive analysis of the dataset using statistical measures and plots described in the previous section. For calculation of all statistical measures and graphical representations R software (R Core Team, 2022) Version, 4.2.1 is used with additional packages **ggpubr** (Kassambara, 2022), **dplyr** (Wickham et al., 2022), **ggplot2** (Wickham, 2016), **GGally** (Schloerke et al., 2021), **reshape2** (Wickham, 2007), **car** (Fox and Weisberg, 2019), **MASS** (Venables and Ripley, 2002), **RColorBrewer** (Neuwirth, 2022).

4.1 Descriptive Analysis

Table 2 in the appendix summarizes the descriptive statistics about the data. The mean of the log-transformed counts of rented bikes is 6.09 with a standard deviation of 1.16. It

implies that the average number of bikes rented in one hour is around $\exp(6.09) = 441$. The median (6.30) of the log-transformed counts of rented bikes is slightly larger than the mean, indicating a left-skewed distribution. This observation is further supported by Figure 2(a), which displays the frequency distribution of the dependent variable, confirming the left-skewed pattern.

Figure 1 consists of scatter plots showing the relationship between the log-transformed counts of rented bikes and each of the other covariates. As can be seen from these plots, wind speed, visibility, snowfall, and rainfall have very weak correlations with the log-transformed number of rental bikes. Figure 2(b) presents a correlation heatmap, revealing the correlation coefficients between the log-transformed counts of rented bikes and the covariates. Notably, the correlation coefficient for temperature is 0.55, indicating a stronger positive correlation compared to the other covariates.

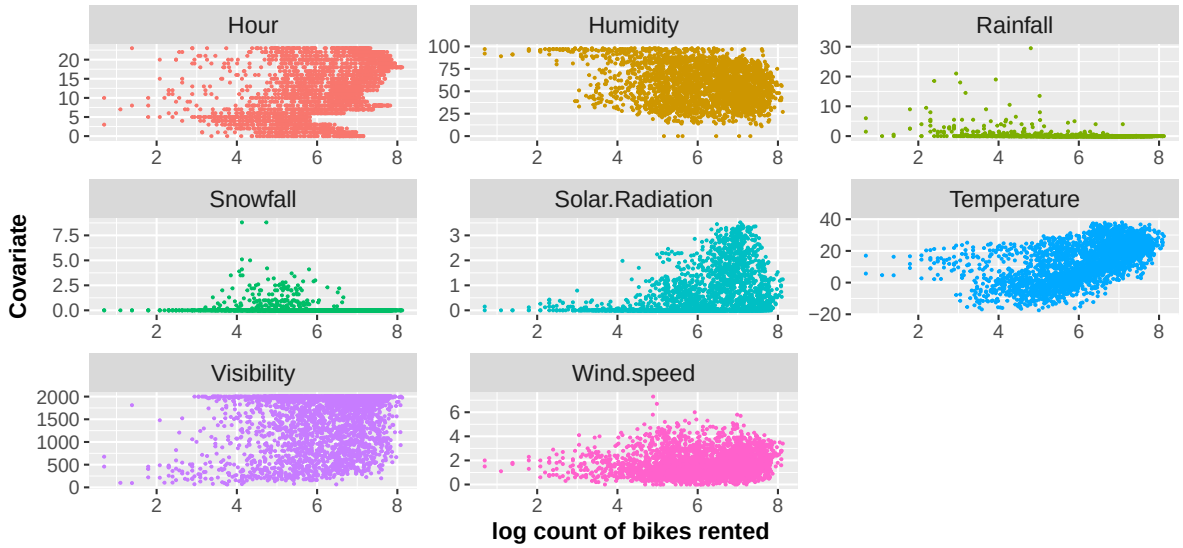
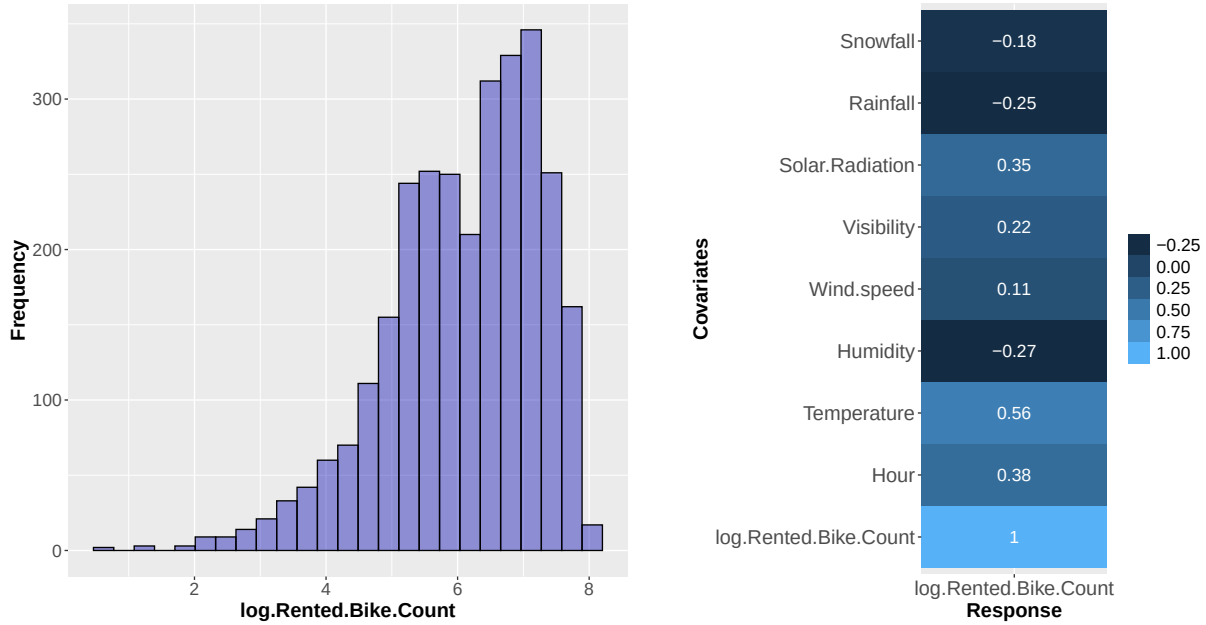


Figure 1: Scatter plots of log-transformed count of rented bikes vs all other covariates

The correlation coefficients for windspeed (0.11), visibility (0.22), solar radiation (0.35), and hour (0.38) suggest weak positive correlations. In contrast, snowfall (-0.18), rainfall (-0.25), and humidity (-0.27) exhibit weak negative correlations. Figure 4 in the appendix, shows boxplots of log-transformed counts of rented bikes in different seasons. The 'Winter' category exhibits the lowest median and range, while 'Summer' has the highest median. Outliers are present in 'Autumn' and 'Spring' with lower values than 'Winter'. In Figure 5 in the appendix, boxplots of log-transformed counts of rented bikes

are shown for holiday categories. The median for 'No Holiday' is larger than 'Holiday', and there are fewer outliers in 'Holiday' compared to 'No Holiday'.



(a) Frequency Distribution of Body Height

(b) Correlation Heatmap

Figure 2: Frequency Distribution Histogram and Correlation Heatmap of the response

4.2 Full Linear Regression Model

A linear regression model that explains the log.Rented.Bike.Count based on all other covariates is fit to the given data set. Since Seasons and Holiday are categorical variables, dummy coding is used. The 'winter' and the 'holiday' are taken as the reference categories. The following model equation is formulated:

$$\begin{aligned}\hat{y}_{log.Rented.Bike.Count} = \hat{\beta}_{Intercept} + \hat{\beta}_{hour}.x_{hour} + \hat{\beta}_{humidity}.x_{humidity} + \hat{\beta}_{rainfall}.x_{rainfall} \\ + \hat{\beta}_{snowfall}.x_{snowfall} + \hat{\beta}_{solarRadiation}.x_{solarRadiation} + \hat{\beta}_{temperature}.x_{temperature} \\ + \hat{\beta}_{visibility}.x_{visibility} + \hat{\beta}_{windSpeed}.x_{windSpeed} + \hat{\beta}_{seasonsAutumn}.x_{seasonsAutumn} \\ + \hat{\beta}_{seasonsSpring}.x_{seasonsSpring} + \hat{\beta}_{seasonsSummer}.x_{seasonsSummer} + \hat{\beta}_{noHoliday}.x_{noHoliday}\end{aligned}$$

Table 3 in the appendix shows the output of the model. At the significance level $\alpha = 0.05$, the coefficient p-values for all categories in hour, temperature, humidity, rainfall, season, and holiday are less than α, indicating that these variables are statistically significant.

Table 4 in the appendix shows an overview of the model, adjusted R-squared is 0.5922, AIC = 6525.976 and BIC = 6609.615.

4.3 Model Selection

The backward selection/elimination method is employed to identify the optimal subset of independent variables for the log.Rented.Bike.Count. This approach starts with a complete model and progressively eliminates the covariate with the least impact on the model. It is computationally advantageous to use backward selection since the full linear regression model has already been fitted. Alternatively, forward or best subset selection methods can also be used.

AIC and BIC are used for model selection. AIC favors more complex models and is better for prediction accuracy, while BIC is more parsimonious and helps avoid overfitting. Backward selection with AIC results in 7 selected covariates, including "wind speed," which was insignificant in the full model. The adjusted R-squared is 0.5925, which is close to the full model and AIC is 6521.455, while BIC is 6587.171. Backward selection with BIC chooses 6 covariates, excluding "wind speed," with an adjusted R-squared of 0.5919, AIC is 6524.577 and BIC is 6584.319 which is lowest among all other models. Finally, BIC is chosen as the selection criterion for its stricter criteria to prevent overfitting. In the appendix, Table 4 shows an overview of all models' results.

4.4 Fitting the Model Using Backward Selection With BIC

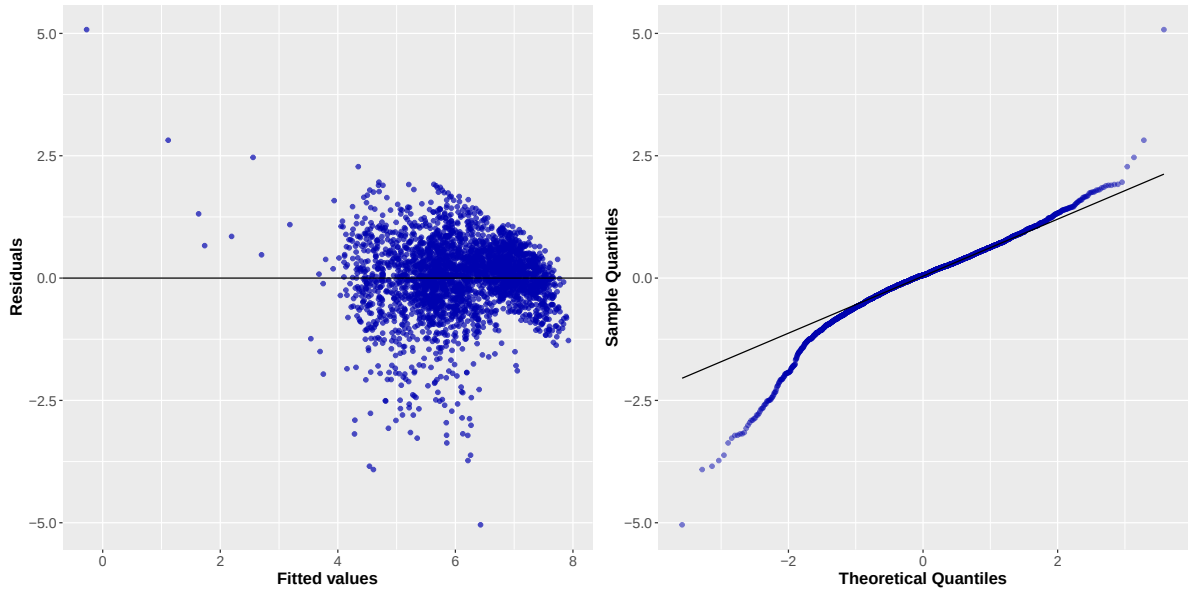
Table 5 in appendix shows the summary of model fit using backward selection method with BIC. We get the following model equation:

$$\begin{aligned}\hat{y}_{log.Rented.Bike.Count} = 5.2862 + 0.0437.x_{hour} + 0.0398.x_{temperature} - 0.0168.x_{humidity} \\ -0.227.x_{rainfall} + 0.7959.x_{seasonAutumn} + 0.5146.x_{seasonSpring} + \\ 0.6187.x_{seasonSummer} + 0.3335.x_{noHoliday}\end{aligned}$$

4.4.1 Verifying the Assumptions

In this section the assumptions of the classical linear model are validated for our dataset. **Expectation of Errors:** To verify $E[\boldsymbol{\epsilon}] = \mathbf{0}$, Figure 3(a) is analyzed which shows the residual plot where the selected model's fitted values are plotted against residuals. Data

points are not scattered almost equally above and below the average line and forming a decreasing pattern, it suggests the presence of a non-zero mean error. **Homoscedastic Error Variances:** Figure 3(a) also shows that the spread of residuals varies systematically (e.g., narrowing/funnel shape) as the predicted values $\hat{y}$ change, it indicates the heteroscedasticity. **Gaussian Errors:** Figure 3(b) shows a Q-Q plot, the points deviate significantly from a straight line, showing pronounced departures from linearity, it suggests non-normality of the residuals.



(a) Residual Plot (b) Q-Q Plot

Figure 3: Residual and Q-Q Plots of fitted linear regression model using BIC

Absence of Multicollinearity: To verify the absence of multicollinearity, Variance Inflation Factor (VIF) is calculated. Table 6 in the appendix shows the VIF values. All values fall between 1-5, which shows moderate multicollinearity. Temperature (4.48) and Seasons (4.62) show relatively high VIF values.

4.4.2 Parameters Interpretation

The intercept (5.2862) is the estimated average log-transformed count of rented bikes when all predictors are zero. Since the predictors include dummy variables, this is the average log count when Hour, Temperature, Humidity, Rainfall are zero and the season is Winter and it is a Holiday. In other words, when all other variables are zero and it is a holiday day in winter, the predicted log of the number of rented bikes is 5.2862 and

the actual count of rented bikes (since we are dealing with log-transformed counts) is expected to be exp(5.2862). The confidence interval is $[5.1305, 5.4419]$. This means we are 95% confident that the true population value of the Intercept lies between 5.1305 and 5.4419. The coefficient for Hour (0.0437) represents that for each unit increase in the Hour variable, holding all other variables constant, the predicted log of the number of rented bikes is expected to increase by 0.0437. In terms of the actual count of rented bikes, this means that for each additional hour, the count of rented bikes is expected to be multiplied by exp(0.0437), assuming all other variables in the model are held constant. The confidence interval is $[0.0395, 0.0479]$. This means we are 95% confident that the true population value of the coefficient for Hour lies between 0.0395 and 0.0479. The coefficient for Temperature (0.0398) shows that for each unit increase in the temperature variable, holding all other variables constant, the predicted log of the number of rented bikes is expected to increase by 0.0398. The confidence interval is $[0.0352, 0.0445]$.

The coefficient for Humidity (-0.0168) with CI $[-0.0182, -0.0154]$ tells that for each unit increase in the humidity variable, holding all other variables constant, the predicted log of the number of rented bikes is expected to decrease by 0.0168. The coefficient for Rainfall (-0.2270) with CI $[-0.2510, -0.2030]$ means that for each unit increase in the rainfall variable, holding all other variables constant, the predicted log of the number of rented bikes is expected to decrease by 0.2270. Coefficients for the dummy variables seasonAutumn (0.7959) with CI $[0.6847, 0.9071]$, seasonSpring (0.5146) with CI $[0.4068, 0.6224]$, and seasonSummer (0.6187) with CI $[0.4586, 0.7788]$ represent the difference in the log-transformed count of rented bikes between the respective season, and the reference season (Winter), holding all other predictors constant. For example, seasonAutumn (0.7959) suggests that, all other covariates being constant, the log count of rented bikes in Autumn is expected to be higher by 0.7959 units than in Winter. In terms of the actual count, this means that the count of rented bikes in Autumn is expected to be exp(0.7959) times higher than in Winter. The coefficient for the dummy variable noHoliday (0.3335) with CI $[0.2089, 0.4581]$ suggests that, holding all other predictors constant, the log count of rented bikes on non-holiday days is expected to be higher by 0.3335 units than on holidays.

4.5 Summary

The analysis of rented bikes has played a vital role in various areas such as urban planning, sustainability efforts, transportation optimization, public health promotion, and

tourism development. The objective of this project was to identify the most suitable model for explaining the number of rented bikes. The dataset used in this study comprised 2905 observations and 11 variables, including the response variable. A linear regression model was fitted to understand the relationship between the log-transformed count of bikes rented and the explanatory variables: Hour, Temperature, Humidity, Wind speed, Visibility, Solar radiation, Rainfall, Snowfall, Seasons, and Holiday.

The goal was to find the best linear regression model to predict the number of rented bikes. Backward selection with AIC and BIC was used to determine the optimal set of covariates. The final choice of BIC is to prevent overfitting. The models were evaluated based on adjusted R-squared, AIC, and BIC values.

At first, descriptive analysis revealed a left-skewed distribution for the response variable. The summary statistics indicated that the median (6.30) of the log-transformed counts of rented bikes was slightly higher than the mean (6.09). Subsequently, a full linear regression model incorporating all the given covariates was constructed. Six covariates, namely Hour, Temperature, Humidity, Rainfall, all Seasons categories, and Holiday, were found to be statistically significant. The adjusted R-squared value for this model was 0.5922. Using the backward selection method with AIC and BIC as selection criteria, two optimal models were identified. The AIC-based model consisted of 7 explanatory variables, yielding an adjusted R-squared of 0.5925. Conversely, the BIC-based model included 6 covariates with an adjusted R-squared of 0.5919.

The selected linear model, determined by the BIC criteria with 6 covariates, indicated positive relationships for Hour, Temperature, all Seasons categories, and Holiday, while Humidity and Rainfall showed negative associations with the number of rented bikes. The confidence intervals for the parameter estimates were reasonably narrow. However, it is important to note that the fitted model exhibited violations of key assumptions in linear regression, including non-zero expectation of errors, heteroscedasticity, non-normality of residuals, and the presence of multicollinearity. Consequently, the coefficient estimates may be biased or inefficient, the predictions may be less accurate, and the standard errors may not be reliable.

In addition, it is recommended to conduct further diagnostic tests, such as outlier analysis and identification of influential observations, to gain more insights into the performance of the model and identify potential sources of the violations observed. Addressing these violations and improving the overall performance of the model will require additional analysis and refinement.

Bibliography

Ludwig Fahrmeir, Thomas Kneib, Stefan Lang, and Brian D. Marx. *Regression*. Springer-Verlag GmbH, 2022. 2nd Edition.

John Fox and Sanford Weisberg. *An R Companion to Applied Regression*. Sage, Thousand Oaks CA, third edition, 2019. URL `https://socialsciences.mcmaster.ca/jfox/Books/Companion/`.

Raky Julio, Andres Monzon, and Yusak O. Susilo. Identifying key elements for user satisfaction of bike-sharing systems: a combination of direct and indirect evaluations. *Journal Name*, 2022. doi: 10.1007/s11116-022-10335-3. URL `https://link.springer.com/article/10.1007/s11116-022-10335-3`.

Alboukadel Kassambara. *ggpubr: ggplot2' Based Publication Ready Plots*, 2022. URL `https://rpkgs.datanovia.com/ggpubr/`. R package version 3.4.

Erich Neuwirth. *RColorBrewer: ColorBrewer Palettes*, 2022. R package version 1.1-3.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2022. URL `https://www.R-project.org/`.

Dieter Rasch, Rob Verdooren, and Jürgen Pilz. *Applied Statistics*. WILEY, 2020. 1st Edition.

Barret Schloerke, Di Cook, Joseph Larmarange, Francois Briatte, Moritz Marbach, Edwin Thoen, Amos Elberg, and Jason Crowley. *GGally: Extension to 'ggplot2'*, 2021. https://ggobi.github.io/ggally/, https://github.com/ggobi/ggally.

Seoul Bike Sharing Demand, South Korean Govt. Seoul Bike Sharing Demand Dataset. UCI Machine Learning Repository, 2020. DOI: https://doi.org/10.24432/C5F62R.

W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. Springer, New York, fourth edition, 2002. URL `https://www.stats.ox.ac.uk/pub/MASS4/`. ISBN 0-387-95457-0.

Hadley Wickham. Reshaping data with the reshape package. *Journal of Statistical Software*, 21(12):1–20, 2007. URL `http://www.jstatsoft.org/v21/i12/`.

Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*, 2016. URL `https://ggplot2.tidyverse.org`.

Hadley Wickham, Romain François, Lionel Henry, and Kirill Müller. *dplyr: A Grammar of Data Manipulation*, 2022. https://dplyr.tidyverse.org, https://github.com/tidyverse/dplyr.

Appendix

A Additional figures

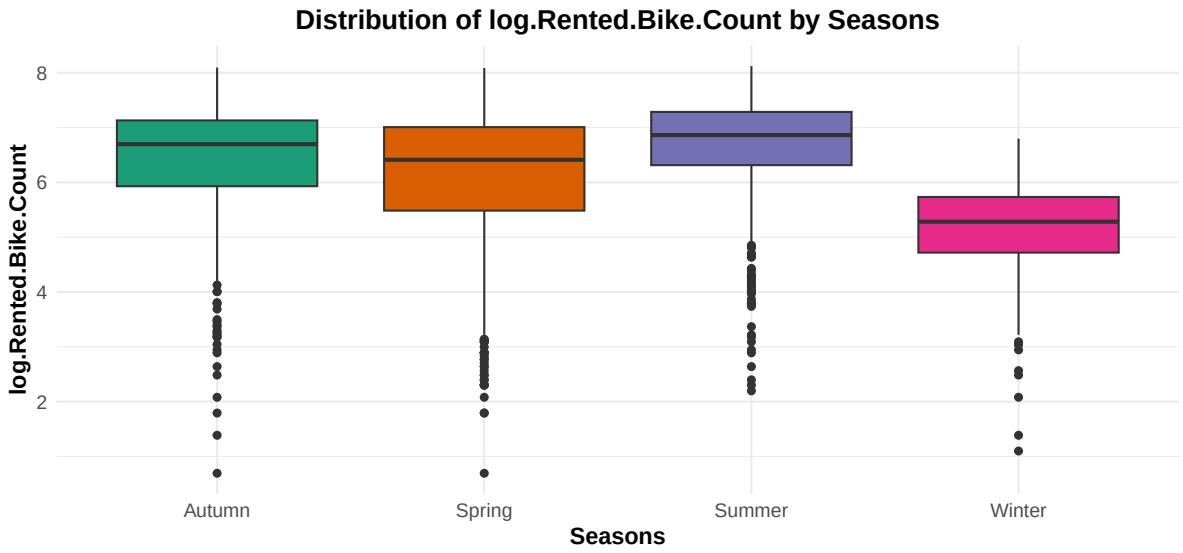


Figure 4: Box plots of log-transformed counts of rented bikes vs all categories of Seasons

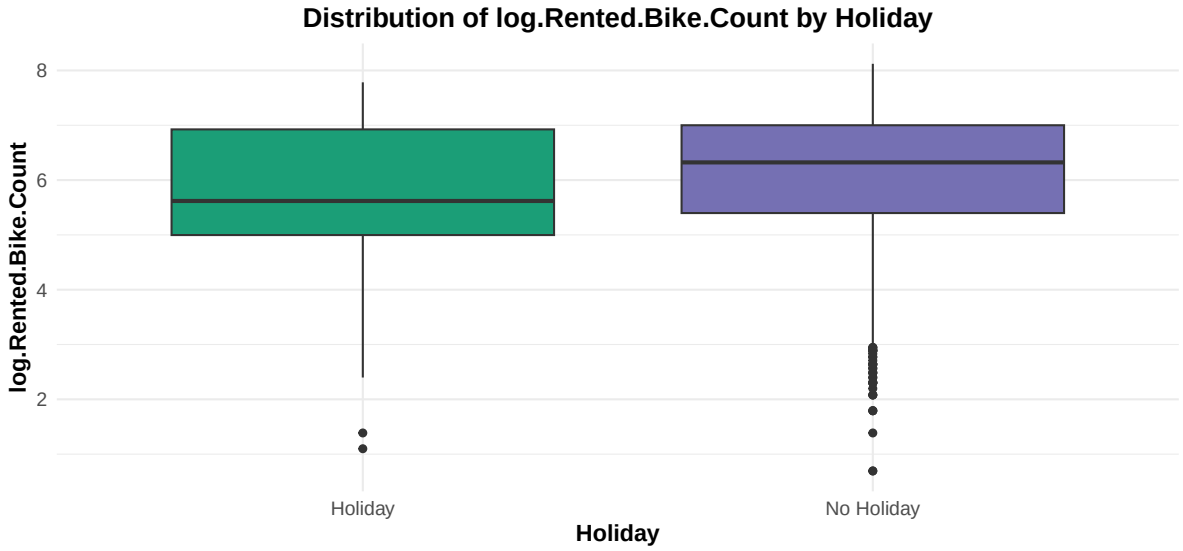


Figure 5: Box plots of log-transformed counts of rented bikes vs all categories of Holiday

B Additional tables

	Min	Q1	Median	Mean	Q3	Max	SD
log.Rented.Bike.Count	0.69	5.37	6.30	6.09	7.00	8.12	1.16
Hour	0.00	6.00	12.00	11.58	17.00	23.00	6.87
Temperature	-17.50	2.80	13.40	12.81	22.80	38.00	12.22
Humidity	0.00	42.00	57.00	57.73	74.00	98.00	20.57
Wind.speed	0.00	0.90	1.50	1.73	2.30	7.30	1.03
Visibility	63.00	940.00	1703.00	1440.73	2000.00	2000.00	607.94
Solar.Radiation	0.00	0.00	0.02	0.57	0.93	3.52	0.87
Rainfall	0.00	0.00	0.00	0.15	0.00	29.50	1.16
Snowfall	0.00	0.00	0.00	0.08	0.00	8.80	0.46

Table 2: Summary Statistics

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	5.4297	0.1156	46.97	<0.0001
Hour	0.0445	0.0022	19.94	<0.0001
Temperature	0.0409	0.0026	15.81	<0.0001
Humidity	-0.0180	0.0011	-16.80	<0.0001
Wind.speed	-0.0286	0.0153	-1.86	0.0625
Visibility	-0.0000	0.0000	-0.60	0.5517
Solar.Radiation	-0.0247	0.0220	-1.12	0.2613
Rainfall	-0.2259	0.0123	-18.41	<0.0001
Snowfall	-0.0063	0.0314	-0.20	0.8418
SeasonsAutumn	0.7835	0.0581	13.49	<0.0001
SeasonsSpring	0.5101	0.0553	9.22	<0.0001
SeasonsSummer	0.6071	0.0833	7.28	<0.0001
HolidayNo Holiday	0.3354	0.0636	5.28	<0.0001

Table 3: Full Linear Regression Model

Model	Selected Covariates	Adjusted.R^2	AIC	BIC
full model	Hour, Temperature, Humidity, Wind speed, Visibility, Solar radiation, Rainfall, Snowfall, Seasons, Holiday	0.5922	6525.976	6609.615
backward selection with AIC	Hour, Temperature, Humidity, Wind speed, Rainfall, Seasons, Holiday	0.5925	6521.455	6587.171
backward selection with BIC	Hour, Temperature, Humidity, Rainfall, Seasons, Holiday	0.5919	6524.577	6584.319

Table 4: Models Overview: Selected Covariates, Adjusted R^2, AIC and BIC

Parameters	Estimated Coefficients	Confidence Interval	Pr(>|t|)
(Intercept)	5.2862	[5.1305 , 5.4419]	<0.0001
Hour	0.0437	[0.0395 , 0.0479]	<0.0001
Temperature	0.0398	[0.0352 , 0.0445]	<0.0001
Humidity	-0.0168	[-0.0182 , -0.0154]	<0.0001
Rainfall	-0.2270	[-0.2510 , -0.2030]	<0.0001
SeasonsAutumn	0.7959	[0.6847 , 0.9071]	<0.0001
SeasonsSpring	0.5146	[0.4068 , 0.6224]	<0.0001
SeasonsSummer	0.6187	[0.4586 , 0.7788]	<0.0001
HolidayNo-Holiday	0.3335	[0.2089 , 0.4581]	<0.0001

Table 5: Backward Model Selection with BIC

	VIF	Df	VIF^(1/(2*Df))
Hour	1.14	1.00	1.07
Temperature	4.48	1.00	2.12
Humidity	1.22	1.00	1.11
Rainfall	1.06	1.00	1.03
Seasons	4.62	3.00	1.29
Holiday	1.03	1.00	1.01

Table 6: Variance Inflation Factor (VIF)