TU Dortmund

Introductory Case Studies

Project 3: Regression Analysis

Lecturers:
Prof. Dr. Katja Ickstadt
M. Sc. Zeyu Ding
M. Sc. Yassine Talleb

Author: Vishesh Srivastava

Group number: 1

Group members: Arindam Pal, Jaimin Prashantkumar Oza

July 6, 2023

Contents

1 Introduction

Bike sharing has become a popular mode of transportation in many urban cities due to its mobility comfort, which offers convenience and an eco-friendly way of commuting. Understanding the factors influencing the demand for bike rentals is crucial in ensuring a steady supply, reducing waiting times, and making them frequently accessible to people. The project studies factors influencing the demand for bicycle rental in the Seoul Bike Sharing System (Seoul Bike Sharing Demand, 2020).

The project aims to create a regression model to predict the number of bikes rented each hour based on weather-related information. The significant factors responsible for bike demand are analyzed to improve the bike-sharing system, ultimately improving bike allocation and the user experience. A descriptive analysis uses central tendency measures and a correlation plot followed by fitting a linear regression model using all independent variables to quantify the influence of each variable on the bike rental. Subsequently, significant independent variables are selected using the best subset selection method, considering the Akaike information criterion (AIC) and Bayesian information criterion (BIC). Finally, residual plots are created to assess model performance and identify the patterns of linearity, heteroscedasticity, and normality, whereas the variance inflation factor checks for multicollinearity among the independent variables.

Descriptive analysis reveals the highest variability in *visibility* while the lowest in *snowfall*. The variables *temperature* and *humidity* correlate positively and negatively, respectively, with *log rented bike count*. The regression model formed with all the independent variables shows *hour*, *temperature*, and *holiday[no holiday]* positively influence the bike rentals. In contrast, all the remaining variables negatively influence bike rentals. According to the best subset selection method with the lowest AIC score, it is observed that the variables *visibility*, *solar radiation*, and *snowfall* are not included in the model due to their non-significance. The model maintains linearity, homoscedasticity, and normality assumptions with no multicollinearity among independent variables.

The data set used in this project is described in detail with the project's objectives in Section 2. The statistical analysis techniques, such as linear regression, the goodness of fit, hypothesis testing and confidence interval, best subset selection method, residual plot, and variance inflation factor, are explained in Section 3. Section 4 analyzes and interprets the results achieved by the statistical methods. Section 5 summarizes the central findings and discusses possible suggestions for further exploration.

2 Problem Statement

2.1 Description of Data Set

The data set used in this project is a small extract of the 'Seoul Bike Sharing Demand Data Set' obtained from the official website of the South Korean government. The data set contains information on various factors responsible for the bike rental each hour. The original data was collected through an observational study by routinely recording the number of bicycles rented each hour and weather-related information (Seoul Bike Sharing Demand, 2020). Furthermore, entries with zero are eliminated, and the dependent variable is logarithmically transformed to improve the statistical quality of the data. Table 1 provides a description and type of each variable.

Table 1: Description of all variables with their respective types

Variable	**Description**	**Type**
log rented bike count	Logarithm of bike rental count per hour	Continuous
hour	Hour of the day	Discrete
temperature	Temperature in degree celsius	Continuous
humidity	Humidity in percentage	Discrete
wind speed	Speed of wind in meter per second	Continuous
visibility	Visibility in ten meters intervals	Discrete
solar radiation	Solar radiation in megajoules per square meter	Continuous
rainfall	Rainfall in millimeters	Continuous
snowfall	Snowfall in centimeters	Continuous
seasons	Classified into winter, spring, summer, autumn	Nominal
holiday	Classified into holiday and no holiday	Nominal

The data set contains 11 variables and 2,905 observations. The dependent variable *log rented bike count* expresses the logarithm of the number of bikes rented each hour and is numerical. The variable *hour* representing the bike rental time is measured using a numerical scale. The variable *temperature* describes the temperature at the time of rental and is measured on a numerical scale in degrees Celsius (°C). The variable *humidity* describes the percentage of humidity and is also measured numerically. The *wind speed* and *visibility* are measured numerically in meters per second (m/s) and meters (m), respectively, where visibility is measured in intervals of ten meters. Additional factors include the continuous variables *solar radiation*, *rainfall* and *snowfall* with units of megajoules per square meter (MJ/m^2), millimeters (mm), and centimeters (cm),

respectively. The variable *seasons* is classified into *winter*, *spring*, *summer*, and *autumn* and is of nominal category. The variable *holiday* is also nominal and indicates whether the day on which the rental took place is a *holiday* or *no holiday*.

The data set used in this project can be considered of good quality as it contains no missing values, and the dependent variable is modified to enhance the statistical properties of the data. In addition, the zero values are also removed. Furthermore, the data set shows excellent quality, characterized by a wide range of variables and a substantial number of accurate observations, assuring its suitability for our analysis.

2.2 Project Objectives

The content-related objectives seek to explore and comprehend the relationship between the different independent variables and dependent variable. It aims to understand how weather-related information such as *temperature*, *humidity*, *wind speed*, *visibility*, *solar radiation*, *snowfall*, *rainfall* along with *hour*, *seasons*, and *holiday* influence the demand for bike rentals in Seoul. Additionally, the project aims to find the significant variables that influence the number of bike rentals demand.

The project's statistical objectives focus on developing a linear regression model that can forecast the logarithm of the number of bike rentals using the provided independent variables. The goal is to find a subset of independent variables that sufficiently explain the variation in the dependent variable. The project evaluates the statistical significance of the parameter estimates to measure the strength of the relationships between the independent and dependent variables. It also assesses the regression model's goodness-of-fit to see how well it captures the variation in the logarithm of bike rentals. The residual plot checks for linearity, heteroskedasticity, and normality patterns. In contrast, the variance inflation factor assesses multicollinearity among the selected independent variables. By achieving these objectives and building a solid regression model for predicting rental counts based on the given variables, the project seeks to analyze the factors influencing bike rentals in Seoul comprehensively.

3 Statistical Methods

This section presents the statistical methods used in the analysis, including their mathematical formulas and definitions. All analyses were performed using R software (Version

4.2.1, R Core Team, 2022), utilizing packages such as corrplot (Wei and Simko, 2021), car (Fox and Weisberg, 2019), and leaps (Lumley and Miller, 2020).

3.1 Linear Regression Model

Linear regression is a statistical modeling technique used to describe the linear relationship between the dependent variable or response variable represented as y, and the independent variables represented as $x_1, \ldots, x_k$, where k is the number of independent variables present in the regression model (Fahrmeir et al., 2013, p. 20).

3.1.1 Model Definition and Assumptions

The relationship between the dependent variable and the independent variables is not always a deterministic function $f(x_1, \ldots, x_k)$ of $x_1, \ldots, x_k$. Still, it exhibits some random error (ϵ), implying that the dependent variable is a random variable whose distribution depends on the independent variables. It is modeled using the conditional expected value $E(Y \mid x_1, \ldots, x_k)$ of Y given the independent variables and given as:

$$E(Y \mid x_1, \ldots, x_k) = f(x_1, \ldots, x_k)$$

The response variable is decomposed as;

$$Y = E(Y \mid x_1, \ldots, x_k) + \epsilon = f(x_1, \ldots, x_k) + \epsilon$$

The function $f(x_1, x_2, \ldots, x_k)$ expresses a linear combination of the regression model's independent variables. It is written as follows:

$$f(x_1, x_2, \ldots, x_k) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$$

Where the coefficients $\beta_0, \ldots, \beta_k$ are unknown and are estimated, β_0 is the intercept, and it denotes the expected value of the dependent variable considering all other variables to be zero (Fahrmeir et al., 2013, p. 21-22).

The response variable can be expressed using matrix notation as follows:

$$Y = X\beta + \varepsilon$$

Where $Y_{n\times 1}$ represents the response vector $(y_1, y_2, \ldots, y_n)'$, β is the $(k+1)$ dimensional unknown parameter vector represented as $(\beta_0, \ldots, \beta_k)'$, ε is the n-dimensional error vector given by $(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)'$, and $X_{n\times(k+1)}$ is the design matrix defined as:

$$X = \begin{bmatrix} 1 & x_{11} & \ldots & x_{1k} \\ 1 & x_{21} & \ldots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \ldots & x_{nk} \end{bmatrix}$$

Where the design matrix X is supposed to have a full column rank, indicated by $rk(X) = k + 1 = p$, where p represents the total number of coefficients including intercept, implying the linear independence of its columns, x_{ij} represents the value of the respective independent variable where $1 \leq i \leq n$ and $1 \leq j \leq k$ (Fahrmeir et al., 2013, p. 73-75).

The following assumptions must hold to model the data using linear regression correctly:

- There exists no multicollinearity among the independent variables.
- The errors must have constant variance across all levels of the independent variables, implying the assumption of homoscedasticity.
- The independent variables and dependent variable must have a linear relationship.
- The errors must follow a normal distribution, $\varepsilon \sim N(0, \sigma^2 I_n)$ with constant variance, where 0 represents the n-dimensional zero vector, σ^2 is the overall variance of ε, and I_n denotes the identity matrix of order n. It can be written as:

$$E(\varepsilon) = E\begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix} = 0, \mathrm{Var}(\varepsilon) = \begin{bmatrix} \sigma^2 & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & \sigma^2 \end{bmatrix} = \sigma^2 I_n.$$

(Fahrmeir et al., 2013, p. 75-77).

3.1.2 Parameter Estimation and Residuals

The least squares and maximum likelihood methods are frequently used in linear regression models to estimate the regression coefficient (β). The least squares approach aims to minimize the sum of squared errors, which is the overall discrepancy between the

predicted values and the real values of the dependent variable in a regression model. It can be formulated as follows:

$$LS(\beta) = \sum_{i=1}^{n}(y_i - x_i'\beta)^2 = \sum_{i=1}^{n}\varepsilon_i^2 = \sum_{i=1}^{n}\varepsilon'\varepsilon = (y - X\beta)'(y - X\beta).$$

To minimize the method of least squares, the parameter values are determined by setting the first derivative to zero and ensuring the second derivative is positive and definite. This results in the equation:

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

Assuming $\varepsilon \sim N(0, \sigma^2 I_n)$, and $Y = X\beta + \varepsilon$, thus $Y \sim N(X\beta, \sigma^2 I_n)$. Now the log-likelihood can be expressed as:

$$l(\beta, \sigma^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}(y - X\beta)'(y - X\beta).$$

Maximizing the log-likelihood is equivalent to minimizing $(y-X\beta)'(y-X\beta)$ with respect to β. The expectation of Y can be formulated as follows:

$$E(Y) = Y = X\hat{\beta} = X(X'X)^{-1}X'y = HY$$

Where H is a symmetric and idempotent hat matrix given as $H = X(X'X)^{-1}X'$ which is a $n \times n$ matrix having the rank equal to its trace, i.e., $rk(H) = Tr(H)$. The unbiased estimator of the error variance denoted as $\hat{\sigma}^2$ is given as $\hat{\sigma}^2 = \hat{\varepsilon}'\hat{\varepsilon}/(n-p)$. The estimated residuals, which is the difference between the observed value and the predicted value and is written as $\hat{\varepsilon} = Y - \widehat{Y}$ (Fahrmeir et al., 2013, p. 105-109).

3.1.3 Dummy Encoding

In situations where the independent variables are categorical, dummy encoding is commonly employed. For categorical variable with c categories defined as $x_i \in 1, \ldots, c$ using dummy encoding, where x_i is the i^{th} dummy variable, $c-1$ dummy variables are defined. This can be formally written as:

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

Where $i = 1, \ldots, n$ denote the different categories of the respective observation, the dummy variable for category c is omitted for identifiability, making it the reference category. The estimated effects are compared to the reference category to facilitate their interpretation (Fahrmeir et al., 2013, p. 97).

3.1.4 Logarithmically Transformed Response Variable

In case when the dependent variable is logarithmically transformed given as $\ln(y)$ with mean $\mu = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k$ and variance σ^2 where k is the number of independent variables, the correct estimate of the mean $\hat{\mu}$ is obtained by exponentiating it as follows:

$$y = e^{\hat{\mu}} = e^{\hat{\beta_0} + \hat{\beta_1} x_1 + ... + \hat{\beta_k} x_k}$$

(Fahrmeir et al., 2013, p.62).

3.2 Goodness-of-Fit Measure: R-squared

The coefficient of determination, commonly known as R-squared, measures goodness of fit. It measures the percentage of the dependent variable's variation that the model can account for. The R-squared values range from zero to one, with a value close to one indicating a strong fit and a value close to zero indicating a poor fit. The R-squared (R^2) is defined as:

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

Where y_i and $\hat{y}_i$ represent the observed response value and the predicted response value, respectively, with $1 \leq i \leq n$ (n is the number of observations), and $\bar{y}$ represents the mean value of the response variable calculated from all the observations in the data set. In the R^2 formula, the $\sum_{i=1}^{n} (\hat{y}_i - y_i)^2$ represents the sum of squared errors, quantifying the difference between the predicted and the actual values of the dependent variable. On the other hand, $\sum_{i=1}^{n} (y_i - \bar{y})^2$ corresponds to the total sum of squares and captures the overall variance in the dependent variable.

While interpreting R^2, it is very crucial to consider all independent variables present in the regression model. An increased number of variables can artificially inflate R^2 and potentially result in an overfitting situation (Fahrmeir et al., 2013, p. 112-115).

3.3 Hypothesis Testing and Confidence Interval

In linear regression analysis, the null hypothesis (H_0) states that the estimated regression parameter for j^{th} covariate is zero. In contrast, the alternative hypothesis (H_1) states that the estimated regression parameter for j^{th} covariate is non-zero. The H_0 and H_1 for $\hat{\beta}_j$ ($0 \leq j \leq k$, where k is the number of regression parameters) can be written as:

$$H_0 : \hat{\beta}_j = 0 \quad \text{vs} \quad H_1 : \hat{\beta}_j \neq 0$$

In linear regression, the significance of individual regression coefficients is assessed using the t-test, given by:

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

Where $\hat{\beta}_j$ represents the estimated coefficient's value of the independent variable for the j^{th} category, and se_j is the standard error related to the coefficient estimate and it is given as $\text{se}_j = \widehat{Var(\hat{\beta}_j)}^{\frac{1}{2}}$ where $\widehat{Var(\hat{\beta}_j)}^{\frac{1}{2}}$ is the diagonal element of the covariance matrix $\text{Cov}(\hat{\beta}) = \hat{\sigma}(X'X)^{-1}$.

The calculated t_j value is compared to the critical value to determine whether to reject the H_0. The critical value is obtained from the $(1 - \alpha/2)$ quantile of t-distribution with $(n - p - 1)$ degrees of freedom, where n and p represents the size of the sample and the number of independent variables, respectively. The H_0 is rejected if:

$$P_{H_0}(|t_j| > t_{1-\alpha/2}(n - p)).$$

The decision to reject the H_0 or not can also be made by comparing the calculated p-value to the chosen significance level (α). If the p-value is less than α, H_0 is rejected. Otherwise, if the p-value is greater than or equal to α, there is no sufficient evidence to reject H_0. The p-value is typically obtained using the t-table or other statistical methods (Fahrmeir et al., 2013, p. 125-131).

In linear regression, the confidence interval (CI) provides an estimated range of values for the β parameter with a specified confidence level of $(1 - \alpha) \times 100\%$. Construction of a CI for the coefficient β_j under the assumption of normally distributed errors, the test statistic $t_j = \frac{\hat{\beta}_j - d_j}{\text{se}_j}$ corresponding to the test $H_0 : \beta_j = d_j$ is used where d_j represents the hypothesized value of the coefficient β_j for the j^{th} independent variable. As the H_0 is rejected when $|t_j| > t_{n-p}(1 - \frac{\alpha}{2})$, thus, the test is constructed in such a way that the

probability of rejecting H_0 when H_0 holds is equal to α. Thus, the (1- α) CI for β_j which is the probability of not rejecting H_0 given H_0 is true is written as:

$$P(|t_j| < t_{n-p}(1 - \tfrac{\alpha}{2})) = 1 - \alpha$$

Thus, we obtain:

$$[\hat{\beta}_j - t_{n-p}(1 - \alpha/2) \cdot \mathrm{se}_j, \hat{\beta}_j + t_{n-p}(1 - \alpha/2) \cdot \mathrm{se}_j]$$

(Fahrmeir et al., 2013, p. 136).

3.4 Best Subset Selection

The best subset selection method is used in linear regression to identify the optimal subset of k predictor variables. This approach involves fitting $2^k - 1$ (since the null model is not regarded) linear models and utilizing the least squares method to determine the subset of variables that yield the best fit. The selection process includes three steps. Firstly, a null model (M_0) without any predictors is used to predict the sample means for each observation. Subsequently, all possible models containing exactly k predictors are fitted, and the best model (M_r) is chosen based on the highest R^2 value. Finally, the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) are employed to select the best model from $M_0, ..., M_k$ (James et al., 2013, p. 227).

The Akaike Information Criterion (AIC) is a likelihood-based model selection criterion. It considers both the model's goodness of fit and its complexity. In the best subset selection method, lower AIC values indicate a better trade-off between model fit and complexity, suggesting a more optimal subset of predictors. The AIC is defined as:

$$\mathrm{AIC} = -2 \cdot l(\hat{\beta}_k, \hat{\sigma}^2) + 2 \cdot (|k| + 1)$$

Where $l(\hat{\beta}_k, \hat{\sigma}^2)$ refers to the highest value attained by the log-likelihood function given the estimated values of the regression coefficients or the maximum likelihood estimator ($\hat{\beta}_k$) and the estimated error variance ($\hat{\sigma}^2$) which is estimated using maximum likelihood estimation, and $(|k| + 1)$ represents the total number of parameters present in the regression model (Fahrmeir et al., 2013, p. 148).

3.5 Linearity and Heteroskedasticity Analysis Using Residual Plot

In a linear regression model, the patterns of linearity and heteroskedasticity can be examined using the residual vs. fitted plot. It is possible to evaluate the linear model assumptions and spot extreme values by plotting the projected values on the x-axis and the residuals on the y-axis. The residuals are computed as the difference between the observed values (Y) and the predicted values ($\hat{Y}$) for each observation. If the underlying assumptions of the linear model are satisfied then the residuals exhibit random distribution around zero without any noticeable patterns. Any pattern suggests that the assumptions of linearity and homoscedasticity are not fulfilled. The residual vs. fitted plot also facilitates the evaluation of the residuals' heteroscedasticity and the normality of errors. Heteroscedasticity occurs when there is a spread of residuals as fitted values increase. In addition to it, the residuals are not randomly scattered around zero if errors follow a non-normal distribution (Dodge, 2008, p. 6).

3.6 Multicollinearity Analysis Using Variance Inflation Factor

Multicollinearity occurs when the independent variables show a strong correlation with each other. The regression coefficients cannot be calculated correctly if multicollinearity exists because the inverse of the design matrix $(X'X)^{-1}$ does not exist; thus, the least squares method cannot be used. Multicollinearity can be verified using the formula:

$$\mathrm{Var}(\hat{\beta}_j) = \frac{\sigma^2}{(1 - R_j^2) \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2}$$

Where R_j^2 measures the linear dependence of an independent variable x_j on the remaining independent variables. A high value of $\mathrm{Var}(\hat{\beta}_j)$ indicates a stronger linear dependence between x_j and the other independent variables. The variance inflation factor (VIF) quantifies the extent to which $\mathrm{Var}(\hat{\beta}_j)$ increases due to the linear dependence between x_j and other independent variables. The VIF is written as $\mathrm{VIF}_j = 1/(1 - R_j^2)$. A VIF exceeding 10 indicates multicollinearity, while a VIF close to 1 indicates the absence of multicollinearity (Fahrmeir et al., 2013, p. 157-158).

4 Statistical Analyses

This section applies the statistical methods defined in Section 3 on the data set to address the problem statements. In addition, the results are interpreted to provide insights and draw conclusions. The null and alternate hypotheses are denoted as H_0 and H_1, respectively, and a significance level (α) uses a value of 0.05 for the analysis.

4.1 Descriptive Analysis of the Variables

The descriptive analysis of the variables using the central tendency measures is shown in Table 2 on page 17 in the Appendix. The mean of *log rented bike count* is 6.091 and a variance of 1.351. The variable *hour* center around 11.583, with moderate variability throughout the day with a value of 47.198. The variable *temperature* averages at 12.807 °C, and a large variance of 149.294 implies substantial temperature variations. The Variable *humidity* averages 57.735% with a variance of 422.948, indicating varying moisture levels. The variable *wind speed* is consistent at 1.734 m/s with low variance. The *visibility* fluctuates widely, averaging at 1440.729 with a large variance of 369590.128. The variables *solar radiation*, *rainfall*, and *snowfall* have low variation among the data points compared to other variables.

The correlation plot provides relationships between the numeric variables in the data set, and Figure 1 depicts this relationship. The variable *log rented bike count* shows the highest correlation coefficient value of 0.56 with *temperature* followed by *hour*, *solar radiation*, *visibility*, and *wind speed* with correlation coefficients of 0.38, 0.35, 0.22, and 0.11, respectively. The results suggest an increase in bike demand as these variables increase. On the other hand, negative correlation coefficient values are observed for *humidity*, *rainfall*, and *snowfall* with a value of -0.27, -0.25, and -0.18, respectively. These values suggest that *humidity* has a strong negative influence on the dependent variable *log rented bike count*. In addition, the increase in rainfall and snowfall also adversely affects bike rentals.

The independent variables show a strong negative correlation between the pairs *'visibility - humidity'* and *'solar radiation - humidity'* with the correlation coefficients of -0.53 and -0.45, respectively. On the other hand, the independent variable pairs *'solar radiation - temperature'* and *'solar radiation - wind speed'* show the highest value of correlation coefficient, i.e., 0.34, suggesting that with increasing solar radiation an increase in temperature and wind speed is observed and vice versa.

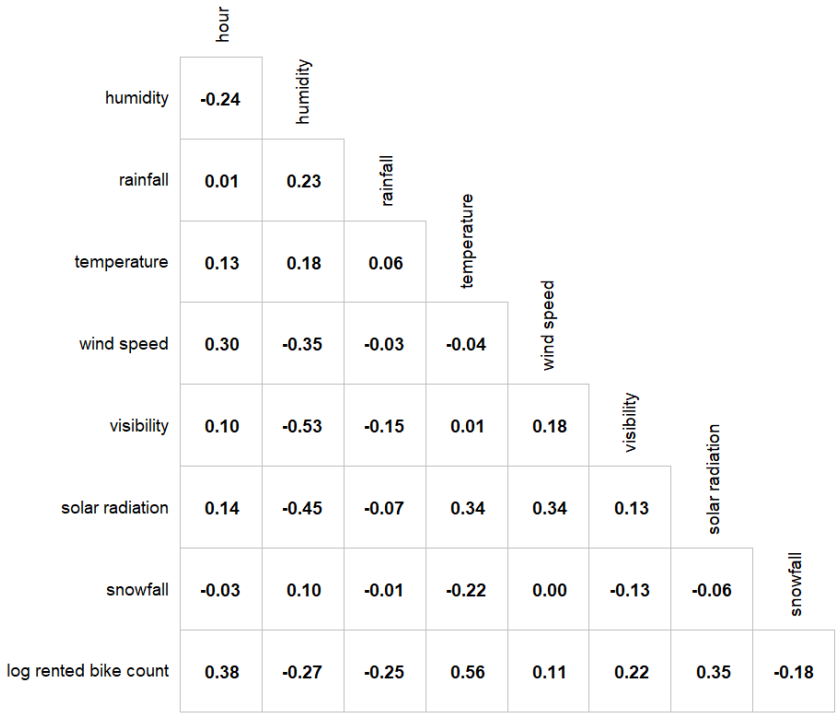


Figure 1: Descriptive analysis using the correlation plot

4.2 Linear Regression Model with All Independent Variables

This section describes the linear regression model for *log rented bike count* based on all independent variables present in the data set, and the results are shown in Table 3 on page 17 in the Appendix. The estimated intercept value of 6.213 indicates that when all independent variables are zero, the predicted value for the *log rented bike count* is 6.213. The standard error of the intercept is the lowest, with a value of 1.289×10^{-1}, suggesting the lowest uncertainty associated with the estimate. The *hour* variable significantly affects the dependent variable as the estimate differs significantly from zero. In the original model, each one-unit increase in the hour of the day corresponds to an approximate $\exp(4.448 \times 10^{-2})$ unit increase in the number of bike rentals while holding other variables constant. The *temperature* variable also significantly affects the dependent variable where a one-unit increase in temperature corresponds to an approximate 4.094×10^{-2} unit increase in the logarithm of the number of bike rentals while holding other variables constant. In the real variable model, a one-unit increase in temperature is associated with an approximately $\exp(4.094 \times 10^{-2})$ increase in bike rentals while keeping other variables constant. The holiday variable is a dummy variable representing the average difference in the logarithm of the number of bike rentals between

holiday and non-holiday periods. The positive parameter estimate of 0.335 and p-value of 1.410×10^{-7} suggests that during holidays, the log rented bike count is approximately 3.354×10^{-1} unit higher on average compared to non-holiday periods while keeping all other variables constant. The variable is also statistically significant, rejecting the H_0.

On the other hand, the variables *humidity*, *wind speed*, *visibility*, *solar radiation*, *rainfall*, *snowfall* and the dummy variable *seasons[spring]*, *seasons[summer]*, *seasons[winter]* are having a negative estimate value. The variable *visibility* has the lowest estimate value of -1.734×10^{-5}. A one-unit increase in these variables, while keeping all other variables constant, leads to a decrease in the logarithm of the number of bike rentals. The variables *wind speed*, *visibility*, *solar radiation*, and *snowfall* have a p-value greater than 0.05, suggesting these variables are not statistically significant leads in failure to reject H_0. The included independent variables explain about 59.4% of the variation in the *log rented bike count* as observed from the R-squared value.

4.3 Regression Model Selection and Results Summary

The best subset selection method using the AIC criterion is chosen to build the final model as it considers the goodness of fit and the number of parameters in the model. Table 4 on page 18 in the Appendix shows the AIC and BIC scores of the $model_1$ and $model_2$. The $model_1$ has the lowest BIC score while $model_2$ has the lowest AIC score. Table 5 on page 18 in the Appendix shows the regression results for the model formed with the selected variables from the AIC criteria. The estimated intercept of 6.145 represents the expected *log rented bike count* when all other variables are zero or at their reference levels and serve as the baseline value for the model. In addition, the intercept also has the largest size for the confidence interval ranging from 5.955 to 6.335. The variable *hour* has an estimate of 0.045, indicating that for each additional hour, there is an expected increase of 0.045 in the natural logarithm of bike rentals, assuming all other variables remain constant suggesting that bike rentals increase as the day progresses. The *temperature* variable also exhibits a positive coefficient of 0.040, implying that higher temperatures are associated with a higher *log rented bike count*. Additionally, the variable *holiday[no holiday]* has a positive coefficient of 0.334, indicating that bike rentals are higher on non-holiday days than on holidays.

The variable *humidity* has a coefficient of -0.017, suggesting that higher humidity levels are associated with decreased bike rentals. High humidity may discourage individuals

from engaging in outdoor activities like bike riding, leading to lower demand for bike rentals. It has the smallest size for the confidence interval ranging from -0.019 to -0.016. The variable *wind speed* is also negatively correlated with the *log rented bike count*, with a coefficient of -0.033. Higher wind speeds and rainfall discourage bike rentals, possibly due to the discomfort and additional effort required to ride against strong winds and rainfall. Furthermore, the variables related to seasons and holidays also significantly impact bike rentals. The dummy variables *seasons[spring]*,*seasons[summer]*, and *seasons[winter]* all have negative coefficients, indicating that bike rentals tend to be lower during these specific seasons compared to the reference season *autumn*.

4.4 Residual Analysis and Model Evaluation

The analysis of the residual plot, as shown in Figure 2 (a) on page 19 in the Appendix, reveals that the residuals exhibit a random pattern as they bounce around the estimated regression line, indicating the reasonableness of the assumption of a linear relationship. The residuals form a relatively horizontal band around the estimated regression line, suggesting approximately equal variances of the error terms, meeting the assumption of homoscedasticity. Furthermore, the absence of deviations in standout residuals from the general random pattern suggests the presence of only a few extreme values in the data. These findings indicate that the model adequately captures the underlying trends in the data and has almost equal variability across different levels of the independent variables. The analysis of the QQ plot as shown in Figure 2 (b) on page 19 in the Appendix, reveals that most of the points lie on the diagonal line, suggesting that the residuals approximately follow a normal distribution. However, some points at the tail of the plot deviate from the diagonal line suggesting the presence of extreme values or a heavy-tailed distribution in the residuals.

The VIF results, as shown in Table 6 on page 19 in the Appendix, indicate varying levels of multicollinearity among the independent variables. As all the variables have a VIF value of less than 10, there is no multicollinearity among the independent variables. However, the *temperature* and *seasons* variables exhibit marginal multicollinearity, with VIF values of 4.481 and 4.619, respectively. Further investigation is necessary to understand these relationships and their impact on the model's performance. On the other hand, no multicollinearity is observed for the variables *hour*, *humidity*, *rainfall*, and *holiday*, making them very reliable in the model.

5 Summary

This project analyzed bike rental demand factors using a linear regression model in the subset of Seoul Bike Sharing Demand Data. The data set consisted of ten independent variables and one dependent variable. The research question focused on understanding the underlying relationship between different independent variables and the dependent variable. This data set was provided by the Introductory Case Studies course faculty at TU Dortmund in the summer term of 2023.

The descriptive analysis revealed that the variable *snowfall* and *solar radiation* show the least variation while *visibility* and *humidity* shows the most variation among the data points. The correlation analysis revealed that the higher *temperature* and *humidity* were associated with increased bike rentals. Conversely, an increase in *humidity*, *rainfall*, and *snowfall* saw a decrease in bike rentals. On fitting the linear regression model with all the independent variables, it is found that the increase in the hour of the day, temperature, and non-holiday result in increased bike rentals. In contrast, the remaining independent variables were responsible for decreased bike rentals. The result suggested that on non-holidays, the supply of bikes must be adequate with the increase in hour of the day and temperature.

The final model is selected using the best subset selection method using AIC criteria. It is observed that all independent variables present in the model significantly influence the *log rented bike count*. As the day progressed, the bike rental increased, possibly due to the increased commuting and recreational activities during daytime hours, while an increase in temperature also saw an increase in bike rentals due to pleasant weather conditions. Conversely, *humidity*, *wind speed*, *rainfall*, and the *seasons* namely *spring*, *summer*, and *winter* negatively affect bike count. Additionally, non-holiday days exhibit higher bike counts compared to holidays. The residual analysis indicates that the regression model satisfies all the underlying assumptions, minimizing any potential bias in the model.

In conclusion, the descriptive analysis revealed information on the central tendency measure and the possible correlation among the variables. A linear regression model helped us to predict the dependent variable more effectively. Selecting the appropriate subset of independent variables and removing insignificant ones from the model helped in predicting the dependent variable accurately. These results can help policymakers and bike rental companies understand and forecast demand. Additional research might investigate new variables or look at different models to improve prediction accuracy.

Bibliography

Dodge, Y. (2008): *The Concise Encyclopedia of Statistics*. Springer Science+Business Media, LLC. Available at: https://doi.org/10.1007/978-0-387-32833-1

Fahrmeir, L., Kneib, T., Lang, S., & Marx, B. D. (2013). *Regression: Models, Methods, and Applications*. Springer-Verlag GmbH.

Fox, J., and Weisberg, S. (2019). car:*An R Companion to Applied Regression*, Third edition. Sage, Thousand Oaks, CA. https://socialsciences.mcmaster.ca/jfox/Books/Companion/.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning with Applications in R*. Springer, New York.

Miller, A. & Miller, T., Lumley, T. (2020). *leaps: Regression Subset Selection* (Version 3.1). Available from https://CRAN.R-project.org/package=leaps.

R Core Team (2023). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria. url: https://www.R-project.org/.

Seoul Bike Sharing Demand. (2020). UCI Machine Learning Repository. https://doi.org/10.24432/C5F62R.

Wei, T., & Simko, V. (2021). *R package 'corrplot': Visualization of a Correlation Matrix* (Version 0.92). Available from https://github.com/taiyun/corrplot.

Appendix

Table 2: Descriptive analysis using the central tendency measures

Variable	**Mean**	**Variance**
log rented bike count	6.091	1.351
hour	11.583	47.198
temperature	12.807	149.294
humidity	57.735	422.948
wind speed	1.734	1.067
visibility	1440.729	369590.128
solar radiation	0.575	0.751
rainfall	0.146	1.344
snowfall	0.083	0.215

Table 3: Linear regression model for *log rented bike count* with all independent variables

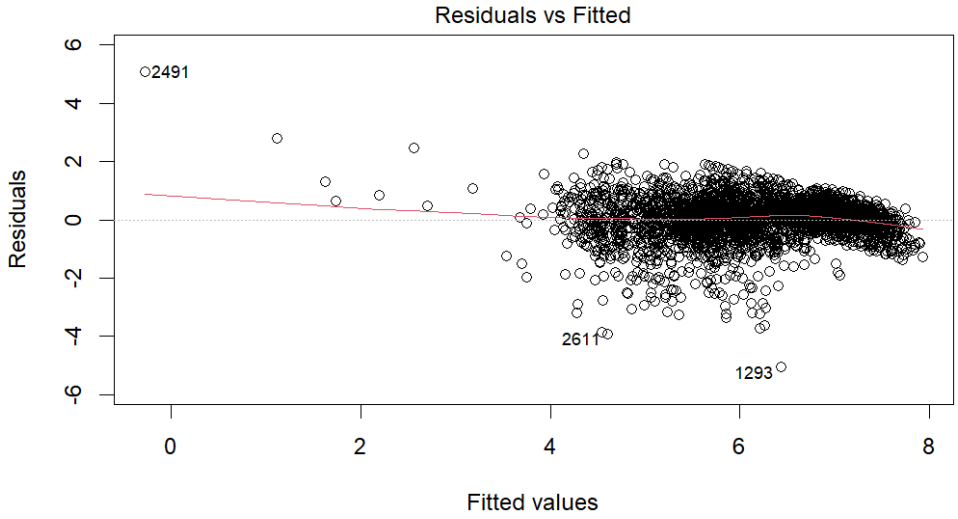
Variable	**Estimate**	**Standard Error**	***t*-Value**	***p*-Value**
(Intercept)	6.213	1.289×10^{-1}	48.207	$< 2.000 \times 10^{-16}$
hour	4.448×10^{-2}	2.231×10^{-3}	19.940	$< 2.000 \times 10^{-16}$
temperature	4.094×10^{-2}	2.589×10^{-3}	15.813	$< 2.000 \times 10^{-16}$
humidity	-1.805×10^{-2}	1.074×10^{-3}	-16.796	$< 2.000 \times 10^{-16}$
wind speed	-2.858×10^{-2}	1.534×10^{-2}	-1.864	0.062
visibility	-1.734×10^{-5}	2.912×10^{-5}	-0.595	0.552
solar radiation	-2.472×10^{-2}	2.200×10^{-2}	-1.124	0.261
rainfall	-2.259×10^{-1}	1.227×10^{-2}	-18.407	$< 2.000 \times 10^{-16}$
snowfall	-6.272×10^{-3}	3.142×10^{-2}	-0.200	0.841
seasons[spring]	-2.735×10^{-1}	4.150×10^{-2}	-6.589	5.240×10^{-11}
seasons[summer]	-1.764×10^{-1}	5.072×10^{-2}	-3.479	0.001
seasons[winter]	-7.835×10^{-1}	5.808×10^{-2}	-13.490	$< 2.000 \times 10^{-16}$
holiday[no holiday]	3.354×10^{-1}	6.356×10^{-2}	5.277	1.410×10^{-7}
R^2				0.594

Table 4: Subset of best model based on AIC and BIC criteria

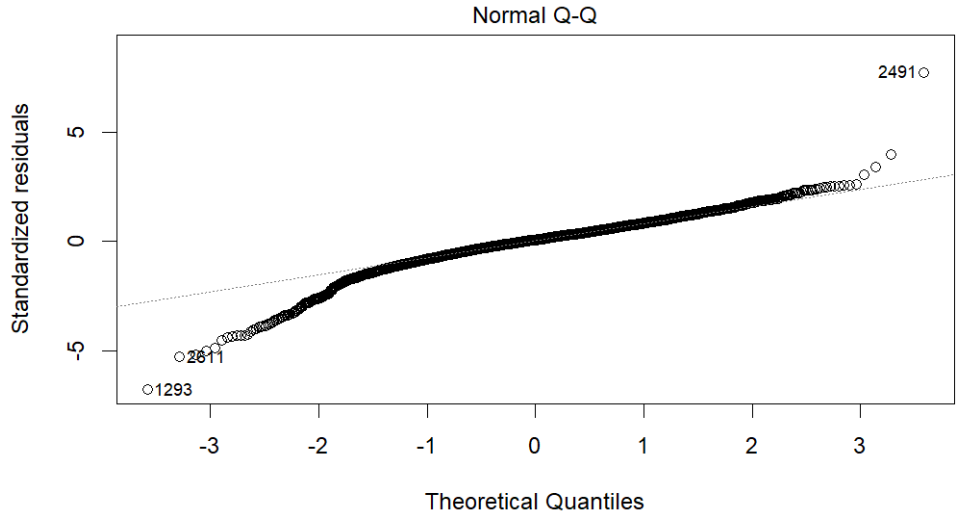
Model	**Variables**	**AIC**	**BIC**
model_1	hour, temperature, humidity, rainfall, seasons[spring], seasons[summer], seasons[winter], holiday[no holiday]	-2599.462	**-2539.721**
model_2	hour, temperature, humidity, wind speed, rainfall, seasons[spring], seasons[summer], seasons[winter], holiday[no holiday]	**-2602.585**	-2536.869

Table 5: Regression results for *log rented bike count* with selected explanatory variables

Variable	**Estimate**	***p*-Value**	**95% Confidence Interval**
(Intercept)	6.145	0.000×10^{0}	(5.955, 6.335)
hour	0.045	1.931×10^{-86}	(0.041, 0.049)
temperature	0.040	2.877×10^{-60}	(0.035, 0.045)
humidity	-0.017	7.792×10^{-103}	(-0.019, -0.016)
wind.speed	-0.033	2.387×10^{-2}	(-0.062, -0.004)
rainfall	-0.226	5.451×10^{-72}	(-0.250, -0.202)
seasons[spring]	-0.270	2.432×10^{-11}	(-0.349, -0.191)
seasons[summer]	-0.173	5.908×10^{-4}	(-0.272, -0.075)
seasons[winter]	-0.784	7.127×10^{-42}	(-0.896, -0.673)
holiday[no holiday]	0.334	1.489×10^{-7}	(0.210, 0.459)
R^2			0.592



(a) Residual plot for linearity and heteroskedasticity analysis



(b) Residual plot for normality analysis

Figure 2: Analyzing plots for linearity, heteroskedasticity, and normality patterns

Table 6: Multicollinearity analysis using variance inflation factor (VIF)

Variable	**VIF**
hour	1.144
temperature	4.481
humidity	1.225
rainfall	1.062
seasons	4.619
holiday	1.029