TU Dortmund

Introductory Case Studies

Project 3: Regression Analysis

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

Author: Abdul Muqsit Farooqi

December 17, 2023

Contents

1 Introduction

Rental bikes have a huge importance in transport so the demand for the usage of rental bikes in many urban cities for mobility comfort has been introduced. To reduce the waiting time, the availability of rental bikes at the time of requirement is crucial. Thus, it become a huge concern to provide rental bikes to the city with a stable supply.

The dataset contains the number of bikes rented per hour and weather information that includes Temperature, Humidity, Windspeed, Visibility, Solar radiation, Snowfall, Rainfall, and seasons and holiday. The data "Seoul Bike Sharing Demand Data Set" is taken from the official website of the South Korean government contains 10 independent variables and a dependent variable where the independent variables are hour, temperature, humidity, wind speed, visibility, solar radiation, snowfall, rainfall, seasons, and holiday and the dependent variable log.Rented.Bike.Count.

The goal of this project is to fit the best possible regression model on the rented bike that could possibly clarify the relationship between rented bike count based on hours and weather information. Initially, the dataset is subjected to descriptive analysis by using a correlation plot to understand the distribution of variables involved in the study. Afterward, a regression model is constructed using log.Rented.Bike.Count as the response variable and other variables as the explanatory variables. By employing the best subset selection technique, various models are compared, and a model with the lowest AIC and increased R-Square is selected as the best model which is a model of seven features.

In section 2, an overview of the dataset is presented, including a discussion on the data's quality. Section 3 delves into the comprehensive discussion of the statistical methods employed for dataset analysis. It encompasses an in-depth exploration of the assumptions, and formulas of the linear regression models, alongside outlining the focal points of Best Subset Selection, Akaike Information Criterion, adjusted r square, and multicollinearity. In section 4, we apply the methods discussed in section 3 to the dataset and interpret the results. Lastly in section 5, the Summary section provides a brief overview of the findings and suggests potential areas for improvement in this study.

2 Problem Statement

2.1 Data Set and Data Quality

The data "Seoul Bike Sharing Demand Data Set" is taken from the official website of the South Korean government. The dataset contains weather information (temperature, humidity, wind speed, visibility, solar radiation, snowfall, rainfall), hours, seasons, holidays,

and the number of bikes rented. The dataset contains 10 independent variables and a dependent variable where the independent variables are hour, temperature, humidity, wind speed, visibility, solar radiation, snowfall, rainfall, seasons, and holiday and the dependent variable is the natural logarithm of the number of bike rentals log.Rented.Bike.Count. For categorical variables holidays and seasons, we have dummy coding HolidayNo Holiday, SeasonsSpring, SeasonsAutumn, SeasonsSummer, and SeasonsWinter. As the dataset with missing values are already removed therefore no requirement of cleaning up the data.

Table 1: Variable types and their description.

Variable Name	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 (%)
Wind.speed	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	Categorical	Winter, Spring, Summer, Autumn
Holiday	Categorical	Holiday/No holiday

2.2 Project Objectives

In this project, there's no need for data pre-processing as the provided data is already processed. The primary objective of the project is to have the best-fitted model of the given data and to achieve it there are some tasks defined. In the first task, a correlation plot is used to describe the relationship between the variables. In the second task, the linear regression model of the log(Rented.Bike.Count) is determined based on all other independent variables which are temperature, humidity, wind speed, visibility, solar radiation, snowfall, rainfall, hours, seasons, and holidays. In the third task, a suitable subset of explanatory variables for the log(Rented.Bike.Count) is performed and results by its reduced AIC and increased R^2. In the fourth task, the residual plots for model evaluation are created to check for patterns of linearity, heteroskedasticity, and normality. Also, the multicollinearity is checked using the variance inflation factor (VIF), and the problems faced in the final model are discussed.

3 Statistical Methods

In this Section, various statistical methods are discussed for the analysis of the data. For the calculation and graphical representation of statistical measures R Software version, 4.2.1 (Core-Team, 2022) is used with the package ggplot2 (Wickham, 2016), dplyr (Wickham et al., 2021), and olsrr (Hebbali, 2020).

3.1 Multiple Linear Regression

Regression is widely recognized as a valuable tool for aiding scientists and researchers in making informed decisions on intricate matters. Such decisions may encompass predicting future events or comprehending the influence of one or more variables on each other in a given relationship. Multiple linear regression is a regression model specifically designed to elucidate the relationship between multiple independent variables (both categorical and continuous) and a single continuous dependent variable. The general equation for multiple linear regression can be expressed as follows:

$$y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \ldots + \beta_k x_{i,k} + \epsilon_i$$

where, for $i = n$ observations and k number of variables:

y_i represents the dependent variable for the i^{th} observation, $x_{i,k}$ represents the value of the k^{th} explanatory variable for the ith observation, β_0 represents the intercept, and $\beta_1, \ldots, \beta_k$ represent the coefficients for each explanatory variable, and ϵ_i represents the error or noise in the model.

Frequently, the connection between a dependent variable and explanatory variables is represented using a function f, i.e.,

$$y = f(x_1, \ldots, x_k) + \epsilon$$

where,

$$f(x_1, \ldots, x_k) = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \ldots + \beta_k x_{i,k}$$

The components f and ϵ in the equation are commonly known as the systematic and stochastic components of the model, respectively. The stochastic component captures the random noise or error present in the model, while the systematic component represents the linear combination of the covariates. To obtain an accurate estimation of the function f, it is essential to separate the random noise from the systematic component and estimate the unknown parameters $\beta_0, \beta_1, \ldots, \beta_k$ using the available data collected for x_i and y_i (where i ranges from 1 to n) (Fahrmeir et al., 2013, p. 22).

For n-observations, the above equation, in matrix notation, is summarized as

$$y = \mathbf{X}\beta + \epsilon,$$

where y and ϵ represent n-dimensional vectors, and β represents the unknown parameters with $p = k + 1$ dimensions.

$$y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_n \end{pmatrix}, \quad \text{and} \quad \epsilon = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

and $\mathbf{X}$ represents the $n \times p$ design matrix,

$$\mathbf{X} = \begin{bmatrix} 1 & x_{1,1} & \dots & x_{1,k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n,1} & \dots & x_{n,k} \end{bmatrix}$$

Additionally, the model makes certain assumptions about the vector ϵ. It assumes that the mean of all the errors is equal to zero, i.e., $\mathbb{E}(\epsilon) = 0$, constant error variance across observations, i.e., $\text{Var}(\epsilon_i) = \sigma^2$, errors are uncorrelated, i.e., $\text{Cov}(\epsilon_i, \epsilon_j) = 0$ for $i \neq j$, which leads to the covariance matrix $\text{Cov}(\epsilon) = \mathbb{E}(\epsilon\epsilon^\top) = \sigma^2 I$. The design matrix X is assumed to be of full column rank. Lastly, the errors are assumed to be normally distributed, i.e., $\epsilon_i \sim \mathcal{N}(0, \sigma^2 I)$ (Fahrmeir et al., 2013, p. 75).

3.1.1 Dummy Coding for Nominal Covariates

Regression models typically assume that the independent variables are measured on a numeric scale and that the relationship between the independent and dependent variables is linear. However, in real-world scenarios, this assumption may not always hold true. Regression models frequently include categorical variables to account for different categories or groups. Dummy coding is a widely used technique to incorporate categorical variables into regression analysis. Dummy coding involves representing a categorical variable with "c" categories using a set of binary variables. Each binary variable corresponds to one category and indicates its presence or absence in a particular observation.

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

Dummy coding is a technique used to create binary variables for categorical variables with 'c' categories, where one category is chosen as the reference category (Fahrmeir et al., 2013, p. 97). The reference category is typically selected based on ease of interpretation and its frequency of occurrence in the dataset.

3.1.2 Assumptions

The assumptions of linear regression are as follows:

1. The errors have zero mean and can be written as $\mathrm{E}(\epsilon) = 0$.
2. The errors have constant variance among them and they are normally distributed.
3. The design matrix X is assumed to be full rank.

(Fahrmeir et al., 2013, p. 74-76)
Then n equations for Eq.1 can be summarized as:

$$y = X\beta + \varepsilon$$

3.1.3 Parameter Estimators and Residuals

Various techniques for estimating parameters are employed in regression analysis, depending on the model assumptions and data characteristics. The method of least squares is the most commonly used technique for estimating regression parameters. It assumes a linear relationship between the dependent and independent variables. To distinguish between the model parameters and their estimated values, the "hat" symbol ($\hat{\beta}_k$) is used (Fahrmeir et al., 2013, p. 77). This distinction is necessary because obtaining the true parameter value without any error is not feasible. In light of this distinction, the estimator for the mean ($E(y_i)$) of the dependent variable (y_i) is given by:

$$E(y_i) = \hat{\beta}_0 + \hat{\beta}_1 x_{i,1} + \ldots + \hat{\beta}_k x_{i,k}$$

and the estimated error referred to as residuals, is defined as the difference between the true value and the predicted value

$$\varepsilon_i = y_i - \hat{y}_i$$

The least squares method minimizes the sum of squared deviations to approximate the unknown regression parameters (Fahrmeir et al., 2013, p. 105).

$$LS(\beta) = \sum_{i=1}^{n}(y_i - x_i'\beta)^2$$

The unbiased estimate of the β coefficients, assuming that the design matrix X is linearly independent and $X'X$ is positive definite (Fahrmeir et al., 2013, p. 107), can be calculated as follows:

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

3.1.4 Significance of Parameter Estimates

The significance of the parameter estimates is commonly assessed to determine the statistical significance of the estimated parameters in the regression model. This test aims to evaluate whether the sample estimates of the parameters differ significantly from zero. The assessment is conducted based on the following null hypothesis (H_0) and alternative hypothesis (H_1):

$$H_0 : \beta_j = 0$$
$$H_1 : \beta_j \neq 0$$

A p-value below a specified threshold, such as 0.05, is commonly employed to indicate the statistical significance of the parameter estimate. It suggests that the relationship between the independent variable and the dependent variable is unlikely to be attributed to chance. The t-test is frequently employed to assess this significance, wherein the estimated parameter value is compared to zero using the t-distribution (Fahrmeir et al., 2013, p. 132). The test statistic utilized under the null hypothesis is given by:

$$t_j = \frac{\hat{\beta}_j}{\sqrt{\widehat{Var(\widehat{\beta_j})^{1/2}}}} \sim t_{n-p}$$

where j = (0, ..., k),n denotes the number of observations, p denotes the length of the covariates vector and $\sqrt{\widehat{Var(\widehat{\beta_j})^{1/2}}}$ denotes the estimated standard deviation of the estimator $\hat{\beta}_j$.

3.2 Confidence Intervals for Regression Coefficients

The uncertainty arising from sampling influences the determination of coefficients in a regression model. This uncertainty arises due to the fact that the sample used for

estimation is a random subset of the overall population, and the estimates are derived from sample statistics. Based on sample data, confidence intervals offer a range of values within which the parameter of a random variable's distribution can be expected to lie with a certain probability. These intervals are computed using relevant sample statistics and a selected confidence level, typically 95% (equivalent to 100(1 - α)), which represents the likely range of values encompassing the population parameter.

Provided that the errors are normally distributed or given that the sample size is large (Fahrmeir et al., 2013, p. 137), the $(1 - \alpha)$ confidence interval for the estimate of $\hat{\beta}_j$ is represented by:

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

The standard error se_j of the jth coefficient is computed using the t-distribution's quantile level $(1 - \frac{\alpha}{2})$ with $(n - p)$ degrees of freedom. Here, n represents the total number of observations, and p denotes the length of the covariates vector.

3.3 Best Subset Selection

Best Subset Selection is a method that is usually used in the context of multiple linear regression, aiming to find a subset of p independent variables that best explain the outcome. The initial step of this algorithm begins by finding a null model, which contains no predictors (i.e., $k = 0$). In this case, it simply predicts the average mean for each observation. In the next step, the algorithm considers all possible subsets of the independent variables, starting with $k = 1$ until all the variables are used, i.e., for each subset size $k = 1, 2, \ldots, p$. It then fits the model for all $\binom{p}{t}$ using the least squares method and looks for the model with the smallest residual sum of squares. Finally, the algorithm selects the single best model out of all the models considered by using a chosen criterion such as the Akaike information criterion (AIC), or adjusted R^2 (Heiberger Holland, 2015, p. 639).

3.4 Akaike Information Criterion

The Akaike Information Criterion (AIC) is a statistical method employed for model selection, aiding in the comparison and identification of the most suitable models. AIC incorporates a penalty term to account for model complexity, assigning higher scores to models with more parameters. The preferred model is determined by selecting the one with the lowest AIC value. The AIC is calculated as follows:

$$\text{AIC} = -2ln(\hat{L}) + 2k$$

where k represents the number of parameters required to estimate the model, and $\hat{L}$ denotes the maximum log-likelihood value of the model, which serves as an indicator of the model's goodness of fit (Heiberger Holland, 2015, p. 639).

3.5 Model Accuracy

The coefficient of determination, commonly known as R^2, is a widely used measure in regression models to assess the goodness of fit by evaluating how well the model aligns with the data. A higher R^2 value ($R^2 = 1$) indicates a better fit, while a lower R^2 value ($R^2 = 0$) suggests a poorer fit. The concept of how well the model captures the data is quantified by the proportion of the dependent variable's variation explained by the model relative to the total variation in the dependent variable (Fahrmeir et al., 2013, p. 115).

The formula for R^2 is given by:

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

A limitation of R-squared (R^2) is that it is influenced by the number of independent variables in the model, meaning that it does not decrease even when additional variables are included. This can make it challenging to compare models using R-squared as a measure of goodness of fit (Fahrmeir et al., 2013, p. 114).

A more robust measure for comparing model goodness of fit is the adjusted R-squared (R^2 adjusted). The adjusted R^2 is computed as follows:

$$\bar{R}^2 = 1 - \frac{(n-1)}{(n-p)}(1 - R^2)$$

where n represents the number of observations and p denotes the number of independent variables in the model.

3.6 Multicollinearity

Multicollinearity is a statistical condition where there is a high correlation between two or more predictor variables in a regression model. This correlation creates challenges in determining the independent impact of each variable on the dependent variable. It can result in unstable coefficient estimates, inflated standard errors, and difficulties in interpreting the significance and relevance of the variables within the model.

To effectively assess multicollinearity, an alternative approach involves computing the variance inflation factor (VIF). The VIF measures the ratio between the variance of $\hat{\beta}_j$ obtained from fitting the complete model and the variance of $\hat{\beta}_j$ if it were fitted independently. A VIF value of 1 indicates the complete absence of collinearity. In practice, a certain degree of collinearity among predictors is typically present. As a general rule, a VIF value exceeding 10 indicates a problematic level of collinearity. You can calculate the VIF for each variable using the following formula:

$$VIF(\hat{\beta}_j) = \frac{1}{1 - R^2_{X_j|X_{-j}}}$$

where $R^2_{X_j|X_{-j}}$ represents the coefficient of determination obtained from regressing the predictor X_j onto all the other predictors. When $R^2_{X_j|X_{-j}}$ approaches one, it indicates the presence of collinearity, leading to a large value for the VIF (Gareth James, 2013, p. 243.).

3.7 Residual Plot

In regression analysis, residual plots are utilized to identify potential issues related to the non-linear relationship between the dependent variable and independent variables. A properly fitting regression model should exhibit residuals (the discrepancies between actual values and estimated values) that are evenly distributed around zero. The presence of any discernible pattern in the residuals could indicate a problem with the linear model. In situations involving non-linear associations in the data, employing techniques such as non-linear transformations of predictors or incorporating interaction terms into the model can be beneficial (Gareth et al., 2013, p. 93).

3.8 Scale vs. Location Plot

A "scale vs. location plot" is a visual tool commonly employed in statistical analysis to evaluate the assumptions of homoscedasticity (constant variance) and location (mean or median) in a dataset. Figure 1 involves plotting the standardized residuals against the predicted values or a measure of location, such as the predicted means or medians on the x-axis. On the y-axis, the standardized residuals are the residuals divided by their standard deviation. By observing this plot, analysts can visually inspect the patterns or trends in the dispersion (scale) and central tendency (location) of the residuals. The red line in the plot is not horizontal indicating heteroscedasticity in variance (Weisberg, S., 1999.).

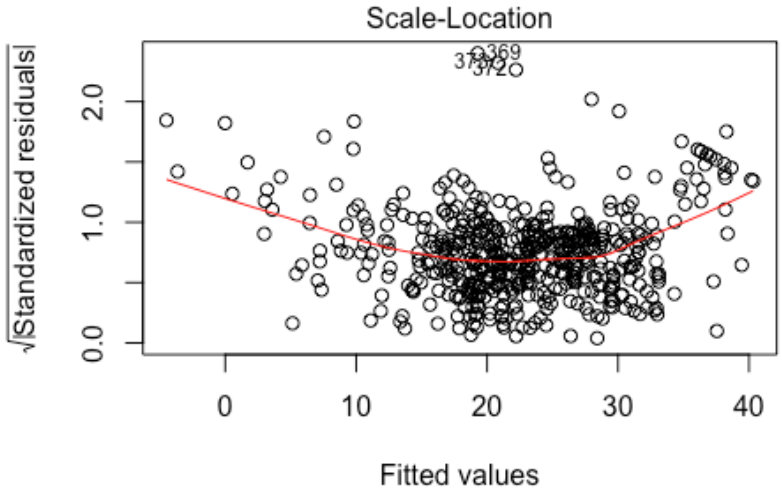


Figure 1: Scale-location plot.

4 Statistical Analysis

The statistical methods discussed earlier are applied in this section to analyze the provided dataset and interpret the obtained results.

4.1 Descriptive Analysis of all Variables

In this section, we are performing the descriptive analysis of the eleven variables which are hour, temperature, humidity, wind speed, visibility, solar radiation, snowfall, rainfall, seasons, holiday, and log.Rented.Bike.Count. From the dataset, we have the details of the rented bikes count in specified hours by the change in weather or season. Table 2 shows the data summary of the change in rented bike count to the change in weather and season.

For the below summary, the Min column represents the minimum temperature and that is -17.50. The 1st quartile is a measure of central tendency that indicates the value below which 25% of the data falls indicating the lower range of values. For instance, in the category "solar.radiation" then the first quartile is "0.0000". For this data, the median of a "log.rented.bike.count" is 6.2971 whereas the mean of a "log.rented.bike.count" is 6.0909. The 3rd quartile is the value below which 75% of the data falls indicating the upper range of values. For example, in the category "visibility" then the third quartile is 2000. The Max column where the variable "visibility" have the maximum value of 2000.

Table 2: Data summary

Variable name	Min	1st Quartile	Median	Mean	3rd Quartile	Max
log.Rented.Bike.Count	0.6931	5.3660	6.2971	6.0909	7.0003	8.1212
Hour	0.00	6.00	12.00	11.58	17.00	23.00
Temperature	-17.50	2.80	13.40	12.81	22.80	38.00
Humidity	0.00	42.00	57.00	57.73	74.00	98.00
Wind.speed	0.000	0.900	1.500	1.734	2.300	7.300
Visibility	63	940	1703	1441	2000	2000
Solar.Radiation	0.0000	0.0000	0.0200	0.5753	0.9300	3.5200
Rainfall	0.0000	0.0000	0.0000	0.1456	0.0000	29.5000
Snowfall	0.00000	0.00000	0.00000	0.08296	0.00000	8.80000

4.1.1 Correlation Plot

Figure 3 shows a correlation plot matrix for the numeric variables. There is a negative correlation between log.rented.bike.count and humidity having a linear correlation of -0.27. This result provides an understanding that decrease in the humidity increases the log.rented.bike.count. Similarly, the comparison between log.rented.bike.count, rainfall and snowfall have a negative relationship where rainfall has -0.25 and snowfall has -0.18. Additionally, a positive linear correlation between log.rented.bike.count, hour, temperature, wind speed, visibility, and solar radiation is found with correlations of 0.38, 0.56, 0.11, 0.22, and 0.35 for each of the variables respectively.

4.2 Linear Regression Analysis

In this section, we present the findings of a regression model that examines the correlation between the response variable, log.rented.bike.count, and a set of explanatory variables including hour, temperature, humidity, wind speed, visibility, solar radiation, snowfall, rainfall, seasons, and holiday. The estimated coefficients and corresponding p-values are displayed in Table 3. Here, the interpretation of the dependent variable log.rented.bike.count taking logarithm changes, and calculated by the $100 \times (\exp(\beta_j) - 1)\%$. The regression model adopts SeasonsAutumn as the reference category, and the intercept of $(\exp(6.213)-1)\times 100 = 49819.66$ percent represents the mean rented.bike.count for this particular season, assuming all other variables are held at zero. Our results reveal that a one-hour increase in the day is associated with an average increase of $(\exp(4.448 \times 10^{-2}) - 1) \times 100 = 4.55$ percent in the rented.bike.count, with all other variables held constant. Moreover, our analysis demonstrates that each unit(%) increase in humidity corresponds to a decrease of $(\exp(-1.805 \times 10^{-2}) - 1) \times 100 = 1.79$ percent in the rented.bike.count, assuming the other variables remain unchanged. Importantly, the

obtained p-values in Table 3 are under 0.05 suggesting that out of the twelve considered variables including dummy coded categorical variables, only nine exhibit significant explanatory power for the variation in log.rented.bike.count, underscoring their significance within the model.

Table 3: Regression Coefficients

Variable	Estimate	P-value
(Intercept)	6.213×10^{0}	$< 2 \times 10^{-16}$
Hour	4.448×10^{-2}	$< 2 \times 10^{-16}$
Temperature	4.094×10^{-2}	$< 2 \times 10^{-16}$
Humidity	-1.805×10^{-2}	$< 2 \times 10^{-16}$
Wind.speed	-2.858×10^{-2}	0.062492
Visibility	-1.734×10^{-5}	0.551694
Solar.Radiation	$-2.472e-02$	0.261291
Rainfall	-2.259×10^{-1}	$< 2 \times 10^{-16}$
Snowfall	-6.272×10^{-3}	0.841810
HolidayNo Holiday	3.354×10^{-1}	1.41×10^{-7}
Seasons (Spring)	-2.735×10^{-1}	5.24×10^{-11}
Seasons (Summer)	-1.764×10^{-1}	0.000511
Seasons (Winter)	-7.835×10^{-1}	$< 2 \times 10^{-16}$

4.3 Model Selection

When developing a model, it is essential to include only pertinent variables that have a significant impact on the response variable. Incorporating irrelevant variables not only introduces unnecessary complexity to the model but also raises the risk of overfitting. To determine the optimal model, we assess the Akaike Information Criterion (AIC) values across different models. Table 4 showcases the AIC values for various models, and based on this assessment, the most favorable model encompasses seven covariates: hour, temperature, humidity, wind speed, rainfall, holiday, and seasons. Opting for a simpler model with fewer parameters is recommended, and in this instance, the model with seven covariates serves as a subset of the other model, making it a more desirable choice.

Table 4: AIC values for different combinations of variables

n	Variables	AIC
1	Temperature	8013.4791
2	Temperature Humidity	7337.4915
3	Hour Temperature Humidity	7081.2672
4	Hour Temperature Humidity Rainfall	6780.4554
5	Hour Temperature Humidity Rainfall Seasons	6550.0708
6	Hour Temperature Humidity Rainfall Holiday Seasons	6524.5772
7	Hour Temperature Humidity Wind.speed Rainfall Holiday Seasons	6521.4547
8	Hour Temperature Humidity Wind.speed Solar.Radiation Rainfall Holiday Seasons	6522.3604
9	Hour Temperature Humidity Wind.speed Visibility Solar.Radiation Rainfall Holiday Seasons	6524.0159
10	Hour Temperature Humidity Wind.speed Visibility Solar.Radiation Rainfall Snowfall Holiday Seasons	6525.9759

4.4 Best Linear Regression Model

This section examines the parameter estimates obtained from the model with the lowest AIC. Table 5 displays the estimated parameter values, along with their corresponding p-values and confidence intervals. The model employs SeasonsAutumn as the reference category, and we interpret the results as follows: the intercept $((\exp(6.1447703) - 1) \times 100 = 46527.25)$ percent represents the average rented.bike.count when all other variables are held at zero. With the other variable of SeasonsSpring, there is a decrease in the rented.bike.count compare to the reference category SeasonsAutumn. Additionally, for the temperature variable, its coefficient suggests an increase in the rented.bike.count compared to variations in other weather conditions. Regarding the parameter estimate for SeasonsSpring, a one-unit increase in SeasonsSpring corresponds to a decrease of $(\exp(-0.2698005) - 1) \times 100 = 23.65$ percent in the rented.bike.count, assuming all other variables remain constant. Similar interpretations can be made for the remaining parameter estimates.

Based on the 95% Confidence interval, it is clear that none of the parameters contain zero within their intervals. If a parameter were to encompass zero within its interval, it would suggest that one or the other variable is not statistically significant in determining the log.rented.bike.count, which does not appear to be true in this scenario.

To ensure accurate and dependable estimates of the relationship between the independent variable and dependent variable, regression models rely on several assumptions that must be validated. Firstly, regression models assume that observations are independent. In the case of the "Seoul Bike Sharing Demand Data Set", it is confirmed that each recorded observation pertains to a unique bike, eliminating any duplicates or repetitions. In Figure 4(a), the scatter of data points is away from zero, with a relatively consistent

spread. This suggests that the assumption of homoscedasticity is not satisfied. Additionally, Figure 4(b) illustrates that the residuals do not exhibit a normal distribution, as the majority of points do not align along a straight line. This further validates the assumption of normality is not being met. In Figure 2, the blue line in the plot is not horizontal indicating heteroscedasticity in variance and when the line in a scale vs location plot is not horizontal, it indicates that the variability of the data points changes as the predicted values increase or decrease. In other words, the spread of residuals or errors is not constant across different predicted values. This is known as heteroscedasticity in variance, and it implies that the variability of the data is not uniform throughout the entire range of predicted values. Regarding multicollinearity, it can be argued that the weather measurements in the dataset are measured on a continuous scale with high precision. For the assumption of multicollinearity, there is no multicollinearity since each VIF value of the independent variable has a value below 10.

Apart from the model meeting the assumptions, it yields an adjusted R-squared value of 0.5925, suggesting that 59.25% of the variability in the dependent variable can be accounted for by the independent variables. Since this value is relatively low, it implies that the model may not exhibit linearity and homogeneity.

Table 5: Linear regression analysis on seven parameters

Variables	Estimates	P-values	Confidence Interval
(Intercept)	6.1447703	$< 2e-16$	[5.955, 6.335]
Hour	0.0448769	$< 2e-16$	[0.041, 0.049]
Temperature	0.0399991	$< 2e-16$	[0.035, 0.045]
Humidity	-0.0172877	$< 2e-16$	[-0.019, -0.016]
Wind.speed	-0.0334258	0.023870	[-0.062, -0.004]
Rainfall	-0.2259778	$< 2e-16$	[-0.250, -0.202]
Holiday No Holiday	0.3344850	1.49e-07	[0.210, 0.459]
SeasonsSpring	-0.2698005	2.43e-11	[-0.349, -0.191]
SeasonsSummer	-0.1733308	0.000591	[-0.272, -0.075]
SeasonsWinter	-0.7842994	$< 2e-16$	[-0.896, -0.673]

5 Summary

The dataset examined in this project comprises information on the hourly count of public bicycles rented in the Seoul Bike Sharing System, along with corresponding weather data and holiday information. It consists of 2905 observations. The objective of the analysis was to identify the most suitable regression model that elucidates the relationship between log.rented.bike.count and various features, such as hour, temperature, humidity, wind speed, visibility, solar radiation, snowfall, rainfall, seasons, and holiday. The dataset was

complete, with no missing values. A logarithmic transformation was applied, resulting in the current dataset where the dependent variable is the natural logarithm of the number of bike rentals, log.Rented.Bike.Count. Descriptive analysis was conducted, and it was found that the average log.rented.bike.count was 6.0909

Following the descriptive analysis, a regression model was fitted using all variables in the dataset. Subsequently, the Best Subset method was employed for model selection, using the AIC measure to determine that a model with seven variables provided a better fit to the data. A linear regression model was then constructed for the dependent variables based on the AIC, and the coefficients of the estimated parameters were interpreted. The model was assessed for adherence to the assumptions of a regression model using a Q-Q plot, residual plot, and scale vs location plot. A goodness-of-fit measure was calculated, suggesting that the model was a suitable fit for the data.

For future studies, the methodology employed in this project can be extended to a larger scale, encompassing a greater number of observations and variables like the strategic placement of rental stations, automated bike redistribution, regular maintenance, and quick repairs can be used to make studies better. For instance, the model could be trained on a dataset to predict the required bike count for each hour, ensuring a stable supply of rental bikes to make them readily available and accessible to the public, thereby reducing waiting times.

References

[1] Core-Team, R. 2022. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

[2] Fahrmeir, Ludwig, Kneib, Thomas, Lang, Stefan, & Marx, Brian. 2013. *Regression: Models, Methods, and Applications*. First edn. Springer Berlin, Heidelberg.

[3] Gareth, James, Daniela, Witten, Trevor, Hastie, & Robert, Tibshirani. 2013. *An Introduction to Statistical Learning: With Applications in R*. Springer.

[4] Hebbali, Aravind. 2020. *olsrr: Tools for Building OLS Regression Models*. R package version 0.5.3.

[5] Kassambara, Alboukadel. 2020. *ggpubr: 'ggplot2' Based Publication Ready Plots*. R package version 0.4.0.

[6] Roser, Max, Appel, Cameron, & Ritchie, Hannah. 2013. *Human Height*. Our World in Data. `https://ourworldindata.org/human-height`.

[7] Wickham, Hadley. 2016. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

[8] Wickham, Hadley, François, Romain, Henry, Lionel, & Müller, Kirill. 2021. *dplyr: A Grammar of Data Manipulation*. R package version 1.0.6.

Appendix

A Additional figures

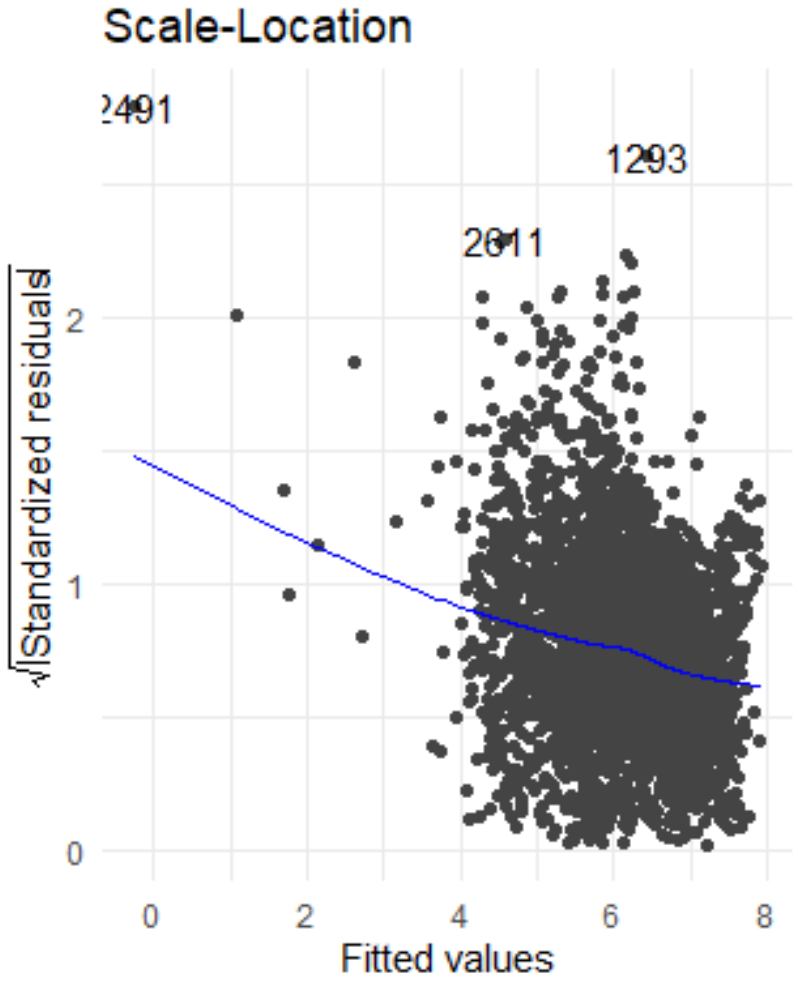


Figure 2: Scale-location plot.

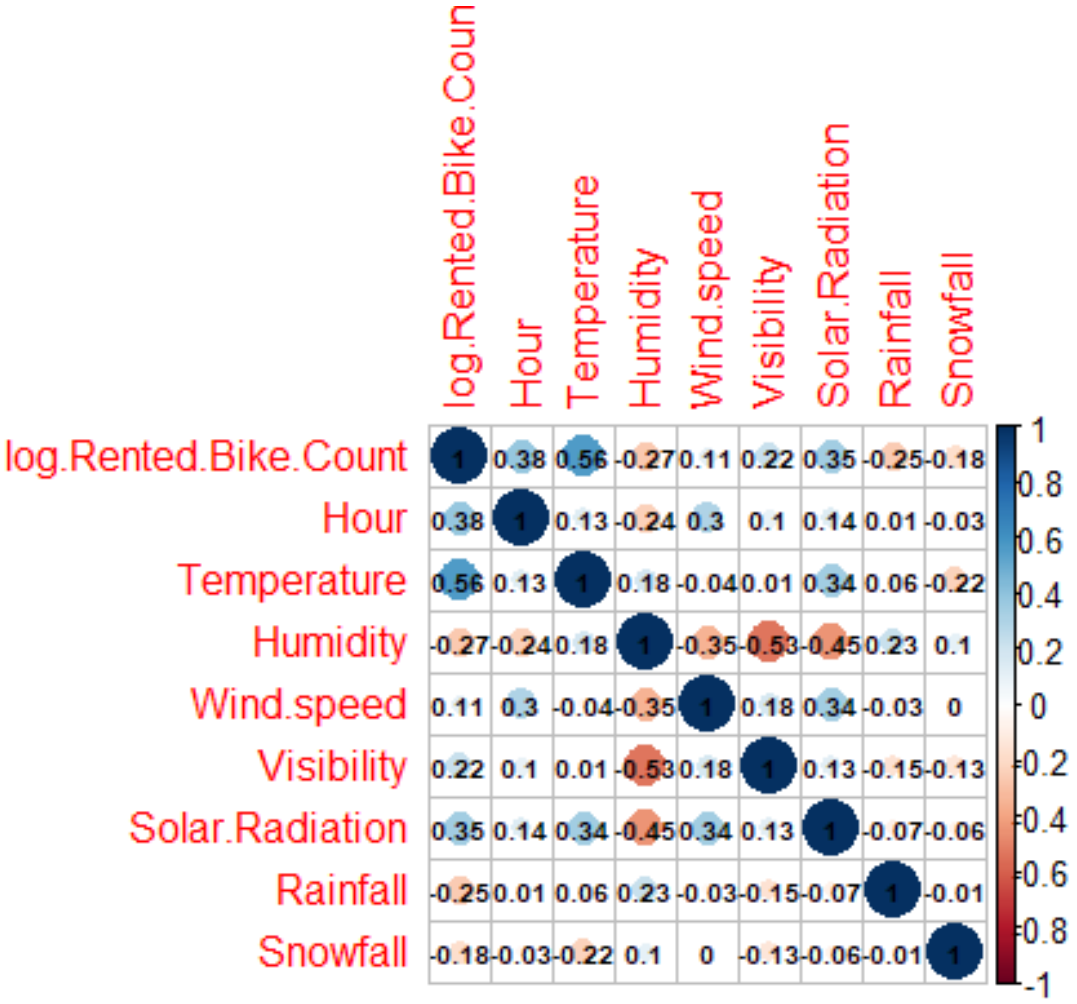
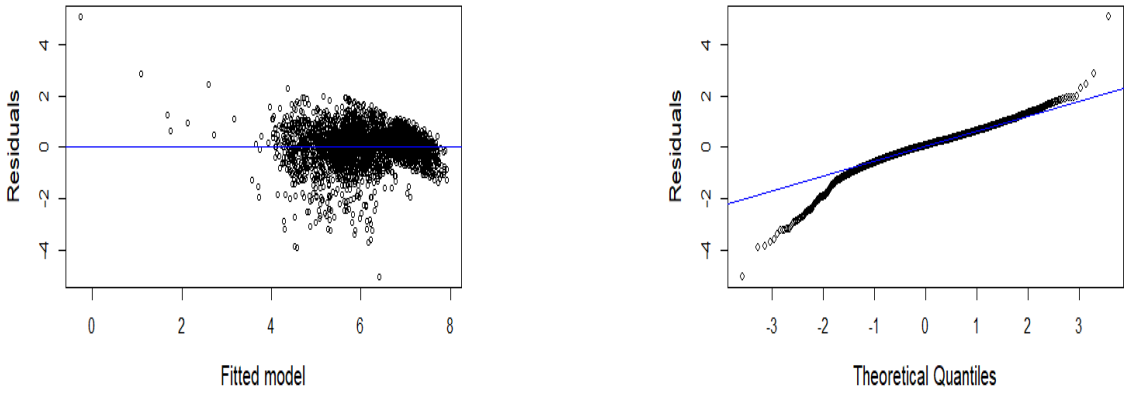


Figure 3: Correlation plot.



(a) Residual vs Fitted plot.

(b) Plot 2

Figure 4: Q-Q plot for normality assumption.