TU Dortmund

Introductory Case Studies

Project 3: Regression Analysis

July 27, 2023

Contents

1 Introduction **1**

2 Problem statement **1**

2.1 Dataset description 1

2.2 Project objective 2

3 Statistical methods **2**

3.1 Multiple linear regression and Collinearity assumptions 3

3.2 Parameter estimation in linear regression 5

3.3 T-test 5

3.4 Confidence interval 6

3.5 Best subset selection 7

3.5.1 Akaike information criterion (AIC) 7

3.6 Coefficient of determination 8

3.7 Residuals vs fitted plot 9

3.8 Dummy coding 9

4 Statistical analysis **10**

4.1 Descriptive analysis of data 10

4.2 Linear regression model based on all parameters 12

4.3 Best subset selection with AIC 13

4.4 Model evaluation and multicollinearity assessment 13

5 Summary **15**

Bibliography **16**

Appendix **17**

A Additional figures 17

B Additional tables 18

1 Introduction

Shared mobility is a modern mode of transportation that combines the use of automobiles, bicycles, and other vehicles. This enables individuals to utilize short-term transportation for urgent needs. Bike-sharing systems on a global scale have revolutionized urban transportation by enabling sustainable mobility alternatives. Understanding the factors, barriers, and incentives that motivate individuals to cycle and utilize bike-sharing services is essential to their success (Torrisi et al., 2021).

This project aims to analyze the "Seoul Bike Sharing Demand Data Set" and construct a regression model to comprehend the correlation between various variables and the number of bike rentals. Our specific objective is to identify the primary variables that have a significant impact on bike rentals and assess the individual effects of these variables. This analysis will offer valuable insights for stakeholders, including city planners and bike-sharing operators. It will empower them to make well-informed decisions to enhance system efficiency and address the increasing demand for bike rentals. The dataset initially comprised 13 independent variables and one dependent variable. However, three variables are eliminated to mitigate the effects of multicollinearity and data-related concerns. The ultimate dataset comprises ten independent variables, encompassing two dummy variables alongside one dependent variable (Seoul Bike Sharing Data, 2023).

In Section 2, we provide a comprehensive analysis of the dataset, including its characteristics and the project's primary goals. Section 3 explores the concept of multiple linear regression, detailing its underlying assumptions and discussing the other statistical techniques used to analyze the dataset. The linear regression models developed by incorporating all variables and those generated using the AIC method are presented in Section 4. We elaborate on the model analysis results and multicollinearity evaluation with visual aids and suitable statistical techniques to facilitate comprehension. Section 5 discusses the key findings of our analysis and suggests possible paths for future research.

2 Problem statement

2.1 Dataset description

This report analyzes the "Seoul Bike Sharing Demand Data Set" from the South Korean government website. An observational study yields 10 independent variables and one

dependent variable *log.Rented.Bike.Count*. Here, we assume this dependent variable as LogBC for the entire report. Variables in the dataset have different data types and measurement scales. The LogBC variables, including *Hour, Temperature, Humidity, Wind.speed, Visibility, Solar.Radiation, Rainfall*, and *Snowfall* are numeric and continuous. Numerical values allow precise measurement and analysis of these variables. Here *Seasons* is a categorical variable representing Winter, Spring, Summer, and Autumn. Finally, the *Holiday* variable categorizes whether an observation is made on a holiday or a regular day. It labels observations "No Holiday" or "Holiday." Importantly, this dataset contains all data points without missing values (Seoul Bike Sharing Data, 2023).

2.2 Project objective

This objective of the report aims to perform a comprehensive descriptive dataset analysis and construct a linear regression model to predict the logarithm of rented bikes. The objective is to determine the optimal subset of explanatory variables for the LogBC from the dataset and assess the performance of the selected model. The present study investigates relationships and patterns within the data using descriptive analysis techniques. The construction of the linear regression model will involve utilizing suitable variable selection criteria. Subsequently, the model outcomes will be concisely presented, encompassing parameter estimates, statistical significance, confidence intervals, and goodness-of-fit measures. To evaluate the effectiveness of the model, residual plots will be generated. These plots will be used to examine the linearity, heteroscedasticity, and normality of the residuals. Additionally, the plots will be utilized to identify and address any potential concerns related to multicollinearity and the interpretation of parameters. The primary objective of this project is to offer significant insights and develop a dependable regression model to predict bike rentals.

3 Statistical methods

This section provides a comprehensive overview of various statistical methods of linear regression. The above methods are then employed to analyze the dataset in line with the defined problem statements. We utilize the R software (R Core Team, 2022) and the following additional packages: ggplot2 (Wickham, 2016), ggpubr (Kassambara, 2023), gridExtra (Auguie, 2017), plyer (Wickham, 2011) and car (Fox and Weisberg, 2019).

These packages provide various statistical methods for analyzing and visualizing the data.

3.1 Multiple linear regression and Collinearity assumptions

Multiple linear regression is a statistical method used to examine the influence of a set of explanatory variables (covariates) on a response variable Y. A function $f(x_1, x_2, ..., x_k)$ captures the relationship between the covariates and the response, which is affected by random noise ϵ. Here, the objective is to estimate the parameters that define the relationship between the covariates and the response, incorporating random variation in the data into account.

In the context of multiple linear regression, the relationship between the response variable Y and a set of covariates $x_1, x_2, ..., x_k$ is expressed by the equation:

$$Y = f(x_1, x_2, ..., x_k) + \epsilon \ ,$$

where the unknown function f is aimed to estimate and separate from the random noise. This unknown function, $f(x_1, x_2, ..., x_k)$, which is a linear combination of the covariates with unknown parameters $\beta_0, \beta_1, ..., \beta_k$ is subject to estimation. The objective of the regression model is to estimate these unknown parameters. The intercept coefficient β_0 corresponds to the estimated value of the response variable when all the explanatory variables are set to zero. By defining the covariate vector $\boldsymbol{x} = (1, x_1, ..., x_k)'$ and the parameter vector $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_k)'$, both of dimension $p = k+1$, the linear combination $f(x_1, x_2, ..., x_k) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$ can be written as $f(x) = \boldsymbol{x'\beta}$. Thus, the equation $Y = f(x_1, x_2, ..., x_k) + \epsilon$ can be represented as $Y = \boldsymbol{x'\beta} + \epsilon$.

Considering a dataset with n observations, for each observation $i = 1, ..., n$, the equation can be written as:

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + \epsilon_i$$

To summarize the equations for all n observations, we define the design matrix $\boldsymbol{X}$ as:

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & \dots & x_{1k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \dots & x_{nk} \end{bmatrix} = \begin{bmatrix} \boldsymbol{x_1'} \\ \vdots \\ \boldsymbol{x_n'} \end{bmatrix}$$

the response vector $\boldsymbol{Y}$ as:

$$\boldsymbol{Y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix},$$

and the error vector $\boldsymbol{\epsilon}$ as:

$$\boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}.$$

With these definitions, the equations for all n observations can be expressed as (Fahrmeir et al., 2013, p. 73-75):

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

In the context of linear regression, several assumptions are made about the errors in the model (Fahrmeir et al., 2013, p. 76). Firstly, the errors are assumed to have an expected value or mean of zero, denoted as $E(\boldsymbol{\epsilon}) = 0$. Secondly, the errors are assumed to exhibit homoscedasticity, meaning that their variance is constant across different values of the covariates. This is represented by the equation $Var(\epsilon_i) = \sigma^2$, where σ^2 represents the error variance. Additionally, the errors are assumed to be uncorrelated, implying that the covariance between different errors is zero, except for cases where i is equal to j. Mathematically, this can be expressed as $\text{Cov}(\boldsymbol{\epsilon}) = E(\boldsymbol{\epsilon}\boldsymbol{\epsilon'}) = \sigma^2\boldsymbol{I}$, where $\boldsymbol{I}$ is the identity matrix.

Furthermore, it is assumed that the design matrix $\boldsymbol{X}$ has a full column rank. The full column rank is denoted as $\text{rk}(\boldsymbol{X}) = k + 1 = p$, meaning there is no multicollinearity among the covariates. Multicollinearity occurs when two or more covariates are highly correlated, leading to difficulty in estimating their individual effects. It arises when a variable can be expressed as a linear combination of one or more additional variables.

The presence of multicollinearity can be assessed using the variance inflation factor (VIF), calculated as $\text{VIF}_j = \frac{1}{1-R_j^2}$, where R_j^2 is the coefficient of determination for the j-th covariate. The VIF_j is calculated using a linear model with x_j as the target variable. A VIF value exceeding 10 indicates a significant multicollinearity problem (Fahrmeir et al., 2013, p. 158).

Finally, the errors are assumed to follow a normal distribution, with a mean of zero and a covariance matrix of $\sigma^2\boldsymbol{I}$. This assumption ensures that the errors capture the random variation in the data and allow for appropriate statistical inference.

3.2 Parameter estimation in linear regression

Several methods are available for estimating the unknown regression parameters β. These estimated parameters are $\hat{\beta}$. One commonly used method is least squares estimation which minimizes the sum of squared errors involving the formula:

$$\hat{\beta} = (\boldsymbol{X'X})^{-1}\boldsymbol{X'Y} \quad ,$$

here, $\boldsymbol{X}$ in this equation represents the design matrix of the covariates, and $\boldsymbol{Y}$ is the vector of response variables.

Another approach is maximum likelihood estimation, which assumes that the errors follow a normal distribution, specifically $\boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{I})$. This assumption leads to the likelihood function which follows $\boldsymbol{Y} \sim N(\boldsymbol{X\beta}, \sigma^2\boldsymbol{I})$:

$$L(\boldsymbol{\beta}, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{(\boldsymbol{Y}-\boldsymbol{X\beta})'(\boldsymbol{Y}-\boldsymbol{X\beta})}{2\sigma^2}\right)$$

The log-likelihood function can be represented by the following expression (Fahrmeir et al., 2013, p. 107):

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

3.3 T-test

The t-test is a statistical method used to assess the significance of a variable, determining whether it should be included as a covariate in the regression model. The hypotheses

for the t-test are:

$$H_0 : \beta_j = 0 \quad \text{against} \quad H_1 : \beta_j \neq 0$$

The null hypothesis (H_0) suggests that the coefficient β_j has no significance, while the alternative hypothesis (H_1) states that it is significantly different from zero. The t-value, t_j, is computed as the ratio of the estimated coefficient $\hat{\beta}_j$ to its estimated standard deviation, se_j. The formula can be defined by:

$$t_j = \frac{\hat{\beta}_j}{\text{se}_j} = \frac{\hat{\beta}_j}{\sqrt{\widehat{Var(\hat{\beta}_j)}}} \sim t_{n-k} \quad , \tag{1}$$

here, se_j can be written as $\sqrt{\widehat{Var(\hat{\beta}_j)}}$. The uncertainty in the estimation is reflected by this number, which is the standard deviation of the estimated coefficient ($\hat{\beta}_j$). The null hypothesis is rejected if the absolute value of t_j exceeds the critical value $t_{1-\alpha/2}(n-k)$, mathematically it is $|t_j| > t_{1-\alpha/2}(n-k)$. Here, se_j represents the estimated standard error of β_j, and $t_{n-k(1-\alpha/2)}$ is the (1-α/2) quantile of the t-distribution with $n-k$ degrees of freedom (Fahrmeir et al., 2013, p. 131).

3.4 Confidence interval

The estimation of regression coefficients is subject to the influence of the particular sample of data employed for analysis. The estimation of coefficients can vary across different samples collected from the population.

A confidence interval is a statistical technique employed to approximate the range of values in which the true population coefficient β_j is expected to be present with a specified confidence level. The confidence interval is determined by α, such as $\alpha = 0.05$, which corresponds to a 95 % confidence interval. If the hypothesis test is conducted at a significance level of 5 %, the interval will not reject the null hypothesis. Additionally, the confidence interval has a 95 % probability of containing the true value of the coefficient β_j. The confidence interval for the parameter β_j is determined by the following equation:

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

where the symbol in this context $\hat{\beta}_j$ denotes the estimated coefficient, se_j corresponds to the estimated standard error of $\hat{\beta}_j$, and $t_{n-k(1-\alpha/2)}$ represents the critical value derived from the t-distribution with $n-k$ degrees of freedom (Fahrmeir et al., 2013, p. 137).

3.5 Best subset selection

Best subset selection is a valuable method to identify the optimal subset of covariates for enhancing a regression model. This technique is particularly advantageous in scenarios with a large number of covariates. The procedure involves generating and assessing all possible subsets of covariates. A set of k covariates generates $2^k - 1$ subsets. Initially, for each subset size p ranging from 1 to k, all $\binom{k}{p}$ models are constructed. The best model M_p is then selected from these $\binom{k}{p}$ models based on the highest adjusted coefficient of determination. The null model, denoted as M_0, represents the absence of any covariate and predicts the sample mean. Ultimately, the best model is chosen from the collection $M_0, \ldots, M_k$ using a model selection criterion such as the Akaike information criterion (AIC) or the Bayesian information criterion (BIC) (James et al., 2013, p. 227).

3.5.1 Akaike information criterion (AIC)

The Akaike information criterion (AIC) is widely employed in selecting models. The approach relies on the principles of maximum likelihood inference and assigns a numerical value to each evaluated model. The model displaying the lowest AIC score is generally considered the most suitable fit. The formula for the Akaike Information Criterion (AIC) is provided as follows:

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

where $l(\hat{\beta}_M, \hat{\sigma}^2)$ represents the maximum log-likelihood of the estimated parameters $\hat{\beta}_M$ and $\hat{\sigma}^2$, and M denotes the total number of parameters included in the model. The error variance is considered an additional parameter, thus leading to an addition of $(|M| + 1)$ in the formula. The AIC formula for a linear regression model with Gaussian errors after ignoring the constant n can be simplified to:

$$\text{AIC} = n \cdot \log(\hat{\sigma}^2) + 2(|M| + 1) \ ,$$

where n represents the sample size (Fahrmeir et al., 2013, p. 148).

3.6 Coefficient of determination

Goodness-of-fit refers to how well the data aligns with a given model. In regression analysis, the coefficient of determination, denoted as R^2, measures goodness-of-fit. The coefficient of determination quantifies the proportion of total variability in the target variable that can be accounted for by the model. The formula for R^2 is defined as follows:

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

here, y_i represents the actual response, $\hat{y}_i$ represents the predicted response, $\bar{y}$ denotes the mean of all actual responses, and $\sum_{i=1}^{n}\hat{\epsilon}_i^2$ represents the residual sum of squares.

The value of R^2 ranges between 0 and 1, where a value closer to 1 indicates a better fit of the model to the data. When R^2 is close to 1, the residual sum of squares $\sum\hat{\epsilon}_i^2$ is small, suggesting that the model captures a significant portion of the variability in the data. Conversely, if R^2 is close to 0, the fit of the model is poor, indicating that the model does not sufficiently explain the variability in the response variable.

When comparing models using the coefficient of determination (R^2), it is essential to ensure that the models being compared possess the same response variable and an equal number of parameters. Additionally, all models should include the intercept (β_0) to maintain consistency (Fahrmeir et al., 2013, p. 113-114).

One limitation of R^2 is that it tends to increase when additional covariates are added to the model, even if those covariates are unrelated to the response variable. The revised coefficient of determination, denoted as $\bar{R^2}$, is utilized to tackle this concern. The adjusted R^2 penalizes the inclusion of new covariates and only increases if the additional variable contributes to the prediction. For a dataset with n observations and k covariates (where $p = k+1$), the formula for adjusted $\bar{R^2}$ is given by (Fahrmeir et al., 2013, p. 148):

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

3.7 Residuals vs fitted plot

Residuals represent the discrepancies between actual and predicted or fitted values in a statistical model. Residuals vs fitted plots can be generated to assess the assumption of homoscedastic error variances in a linear model. This plot displays the fitted values on the x-axis and the corresponding residuals on the y-axis, including a horizontal reference line.

By visually inspecting the residuals vs fitted plot, it is possible to assess the validity of the homoscedasticity assumption, which is essential for accurate model inference and interpretation. The formula for the residuals is given by:

$$\epsilon_i = y_i - \boldsymbol{x_i'}\hat{\boldsymbol{\beta}}$$

Mathematically, the residuals (ϵ_i) are computed as the differences between the observed response (y_i) and the predicted response ($\boldsymbol{x_i'}\hat{\boldsymbol{\beta}}$), where x_i represents the covariate values and $\hat{\boldsymbol{\beta}}$ denotes the estimated regression coefficients.

The purpose of the residuals vs fitted plot is to examine the pattern and dispersion of the points around the reference line. If the points are randomly scattered around the reference line with a consistent spread (variance), this suggests homoscedastic error variances. In other words, the assumption of equal error variances holds. However, suppose the plot reveals a systematic pattern or a changing spread of residuals. In that case, it indicates the presence of heteroscedasticity, where the error variances differ across the range of the fitted values (Fahrmeir et al., 2013, p. 183).

3.8 Dummy coding

The application of dummy coding is a statistical technique employed in linear regression models for the purpose of incorporating categorical covariates. The process entails the utilization of binary values (1/0) to represent categorical variables, indicating the specific category to which an observation applies. In the case of a variable x with k categories, selecting one category as the reference category is common. The remaining $k-1$ dummy variables for variable x can be formally defined as follows:

$$x_{i1} = \begin{cases} 1, & \text{if } x_i = 1 \\ 0, & \text{otherwise} \end{cases}, \quad x_{i2} = \begin{cases} 1, & \text{if } x_i = 2 \\ 0, & \text{otherwise} \end{cases}, \quad \ldots, \quad x_{i,k-1} = \begin{cases} 1, & \text{if } x_i = k-1 \\ 0, & \text{otherwise} \end{cases},$$

where i ranges from 1 to n for n observations. In the regression model, these dummy variables are included as predictors:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_{i,k-1} x_{i,k-1} + \ldots + \epsilon_i$$

All dummy variables are assigned a value of 0 in the reference category. The reference category is typically selected as the most frequently observed category in the observations (Fahrmeir et al., 2013, p. 97).

4 Statistical analysis

4.1 Descriptive analysis of data

The correlation analysis uncovers several noteworthy findings about the associations between LogBC and other variables. It is clear from our observations that there is a tendency for bike rentals to increase as the day progresses, exhibiting a positive correlation with the variable of *Hour*.

In Figure 1, a moderate positive correlation exists between warmer temperatures and bike rental counts, suggesting that increased temperatures are linked to increased bike rentals. In contrast, we also observe a negative correlation between elevated levels of humidity and the number of bike rentals. Although wind speed and snowfall have minimal impact, there is a positive correlation between bike rentals, improved visibility, and higher solar radiation levels. The findings above offer significant contributions to understanding the various factors that impact the demand for bike rentals. However, it is essential to acknowledge that correlation does not necessarily indicate causality, implying that an alteration in one variable does not directly lead to a change in another variable. To achieve a comprehensive understanding of the underlying relationships and influential factors, an additional examination is required.

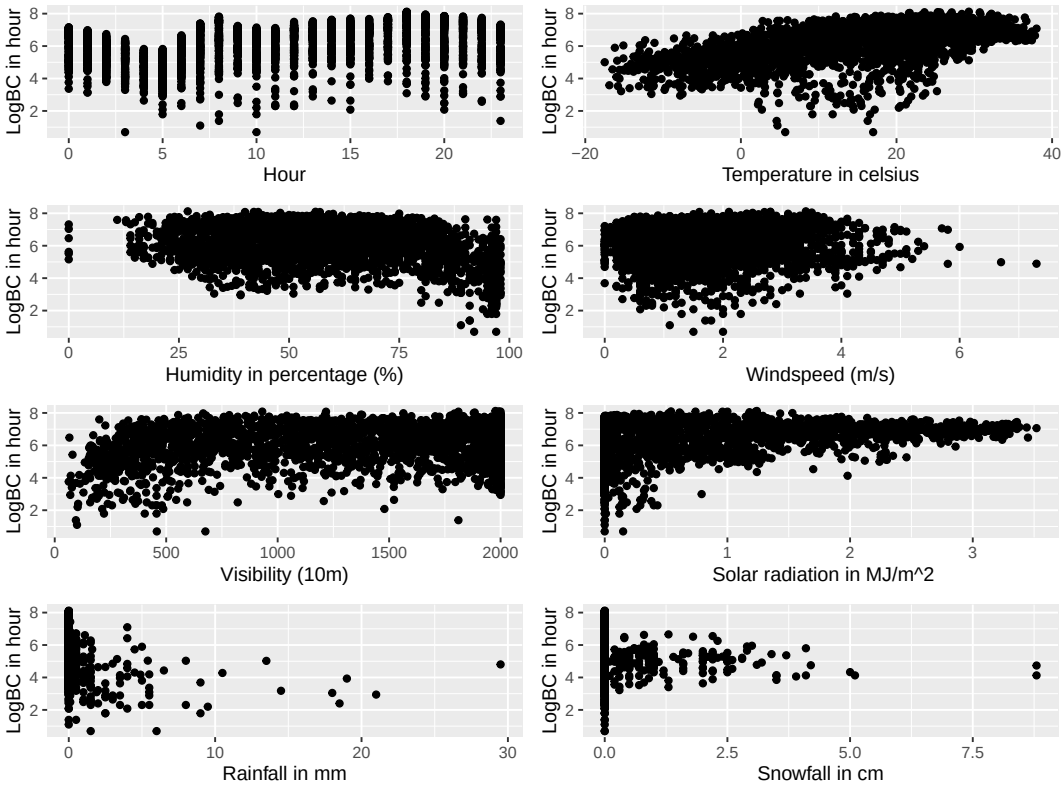

Figure 1: Scatterplots of LogBC in an hour against variables

In Table 1, the variable LogBC (*log.Rented.Bike.Count*) represents the natural logarithm of the hourly count of bike rentals. The table is given as follows:

Table 1: A summary table of continuous variables

	Variable	Count	Min	Max	Mean	Median	IQR	SD
1	log.Rented.Bike.Count	2905	0.69	8.12	6.09	6.30	1.63	1.16
2	Hour	2905	0	23	11.6	12	11	6.87
3	Temperature	2905	-17.5	38	12.8	13.4	20	20.2
4	Humidity	2905	0	98	57.7	57	32	20.6
5	Wind.speed	2905	0	7.3	1.73	1.5	1.4	1.03
6	Visibility	2905	63	2000	1441	1703	1060	608
7	Solar.Radiation	2905	0	3.52	o.58	0.02	0.93	0.87
8	Rainfall	2905	0	29.5	0.15	0	0	1.16
9	Snowfall	2905	0	8.8	0.08	0	0	0.46

The dataset displays a moderate level of variability, as shown by a mean of 6.09 and a standard deviation of 1.16. The remaining variables, namely *Hour, Temperature, Humidity, Wind.speed, Visibility, Solar.Radiation, Rainfall*, and *Snowfall*, are accompanied by

their respective ranges and descriptive statistics. The calculated mean value of LogBC (6.09) estimates the average count of bike rentals per hour. Similarly, the standard deviation (1.16) measures the degree to which the data points deviate from the mean value.

4.2 Linear regression model based on all parameters

The intercept coefficient is estimated to be 6.21, representing the expected value of the log-transformed rented bike counts when all predictor variables in the model are set to zero. The coefficient for each predictor variable represents a change in LogBC variable when the corresponding variable increases by one unit while keeping all other variables unchanged. The equation for it is as follows:

$$\begin{aligned}\text{LogBC} = 6.21 + 0.04 \cdot \textit{hour} + 0.04 \cdot \textit{temperature} - 0.01 \cdot \textit{humidity} - 0.02 \cdot \textit{wind speed} \\ -1.73 \times 10^{-5} \cdot \textit{visibility} - 0.02 \cdot \textit{solar radiation} - 0.22 \cdot rainfall - 0.006 \cdot snowfall \\ -0.27 \cdot \textit{season spring} - 0.18 \cdot \textit{season summer} - 0.78 \cdot \textit{season winter} + 0.34 \cdot \textit{no holiday}\end{aligned}$$

Positive coefficients for the variables *hour* and *temperature* imply that a spike in both *hour* and *temperature* is associated with a corresponding rise in the logarithm of the count of rented bikes. This indicates a tendency for bike rentals to increase as the day progresses and temperatures increase.

The presence of negative coefficients related to *humidity, wind speed, visibility, solar radiation, rainfall, snowfall*, and the *spring, summer*, and *winter* seasons shows a correlation between an increase in these variables and a decrease in the logarithm of the rented bike count (LogBC). This implies that increased levels of *humidity, wind speed, snowfall*, and specific seasons (*spring, summer*, and *winter*) negatively impact the demand for bike rentals. Reduced *visibility* and diminished *solar radiation* are additional factors that contribute to a decline in bicycle rentals.

The positive coefficient associated with the *Holiday* variable suggests that there is a tendency for bike rentals to be higher on days designated as holidays because people may have more leisure time or can engage in recreational activities more during holiday periods.

4.3 Best subset selection with AIC

Table 3 in the Appendix on page 18, the linear regression analysis is conducted on a subset of chosen variables based on the Akaike information criterion (AIC). The variables chosen for analysis are *Hour, Temperature, Humidity, Rainfall, Seasons (Summer, Winter)*, and *Holiday*, display statistically significant coefficients ($p < 0.05$), except for *Wind.speed*, which demonstrates only marginal significance close to 0.05. According to the model, it is clear that variables such as *Hour, Temperature*, and *Holiday* exhibit positive correlations with LogBC. In contrast, variables include *Humidity, Wind.speed, Rainfall*, and *Seasons (Spring, Summer, Winter)* demonstrate negative correlation with LogBC. The model explains approximately 59.37 % of the variance observed in the log-transformed count of rented bikes.

Furthermore, the model demonstrates statistical significance, with a p-value less than 2.2×10^{-16}. The residual standard error, with a value of 0.7419, serves as a measure of the average magnitude of the residuals, thereby providing insight into the quality of fit of the model. The AIC value of 6521.455 also serves as a criterion for the selection of models.

4.4 Model evaluation and multicollinearity assessment

This section focuses on evaluating the assumptions of linear models for the AIC model and interpreting the parameter estimates. Figure 2 in the Appendix on page 17 illustrates the Q-Q plot and the scatter plot of residuals versus fitted values.

The Q-Q plot is a graphical tool used to assess the normality assumption in statistical models. It accomplishes this by plotting the theoretical quantiles on the x-axis and the residuals of the model on the y-axis. The presence of deviations from a linear trend indicates a departure from the expected state of normality. In the present study, it is observed that specific data points in the Q-Q plot fail to show a close alignment with the anticipated linear trend, thus suggesting a departure from the assumption of normality.

The scatter plot of residuals and fitted values is used to test the homoscedasticity assumption, which says that the spread of errors stays the same at all levels of the independent variables. The scatter plot illustrates the relationship between the residuals on the y-axis and the fitted values on the x-axis. Our observations show that most data points are distributed within the range of 4 to 8 on the x-axis, suggesting a consistent

level of variability. Because the residuals are evenly distributed around the horizontal line, which serves as a positive indication of the linear relationship between the response variable and the covariates. This observation indicates that the assumption of homoscedasticity is valid, as there is no observable pattern in the residuals concerning the predicted values.

The variance inflation factor (VIF) is applied in Table 2 to evaluate the presence of multicollinearity within the overall model. The provided table displays the VIF values corresponding to each covariate. It is important to note that lower VIF values indicate reduced levels of multicollinearity.

Based on the data presented in Table 2, it can be observed that the variables of *Hour, Humidity, Wind.speed, Rainfall*, and *Holiday* show relatively low VIF values. This suggests a minimal correlation between these variables and the other factors under consideration. These variables can be regarded as displaying minimal or negligible multicollinearity concerns.

Table 2: Assessment of multicollinearity using Variance inflation factor (VIF)

Covariates	GVIF	$GVIF^{(1/(2*DF))}$
Hour	1.207	1.098
Temperature	4.484	2.118
Humidity	1.326	1.152
Wind speed	1.231	1.110
Rainfall	1.063	1.031
Seasons	4.702	1.294
Holiday	1.029	1.014

Nevertheless, the variables of *Temperature* and *Seasons* indicate higher VIF values, implying the presence of moderate levels of multicollinearity. This observation suggests a potential relationship between these variables and the remaining predictors. While the presence of multicollinearity can impact the accuracy and interpretation of coefficient estimates, it is worth noting that the VIF values fall within acceptable thresholds. This suggests that the level of multicollinearity is not significant enough to compromise the overall integrity of the model.

5 Summary

The research project "Seoul Bike Sharing Demand DataSet" aimed to investigate the factors influencing bike rental demand and develop a linear regression model for estimating the count of rented bikes. Several key findings were uncovered through descriptive analysis, linear regression modeling, and multicollinearity assessment.

Descriptive statistics provided an overview of the dataset. The LogBC variable displayed moderate variability, with a mean of 6.09 and a standard deviation of 1.16. Summary statistics for the other variables offered additional information on their ranges, means, medians, and standard deviations. The correlation analysis revealed significant relationships between bike rentals and various variables. *Hour* exhibited a positive correlation, indicating higher rentals as the day progressed. *Temperature* showed a moderate positive correlation, suggesting increased rentals during warmer weather. *Humidity* displayed a negative correlation, indicating lower rentals during more humid conditions. Other variables, such as *Wind.speed*, *Visibility*, and *Solar.Radiation*, also showed a weakly positive correlation. *Rainfall* and *Snowfall* showed negative correlations with LogBC indicating decreased bike rentals. Guided by the AIC criterion, best subset selection identified a subset of variables with significant coefficients. *Hour, Temperature, Humidity, Wind.speed, Rainfall, Seasons*, and *Holiday* were selected, and the resulting model explained approximately 59.37 % of the variance in LogBC. The model evaluation revealed violations of the normality assumption but confirmed the homoscedasticity of errors. The assessment of multicollinearity using the Variance inflation factor (VIF) indicated low levels for several variables, while *Temperature* and *Seasons* showed moderate levels. Although multicollinearity was present, its impact on the model was deemed acceptable. To summarize, the data points indicate a departure from the normality assumption in the Q-Q plot, whereas the scatter plot provides evidence in favour of the assumption of homoscedasticity. Considering these findings when interpreting the parameter estimates of the AIC model and formulating conclusions is imperative.

In conclusion, it is important to acknowledge that the final model does not satisfy the normality and homoscedasticity assumptions. In addition, the model fails to achieve adequate levels of multicollinearity. Although a moderate correlation exists between *Temperature* and *Seasons* relative to other predictors, its effect on the model cannot invalidate it. In light of the potential influence of multicollinearity on the estimation precision of *Temperature* and *Seasons* coefficients, caution should be exercised when interpreting these coefficients.

Bibliography

Baptiste Auguie. *gridExtra: Miscellaneous Functions for "Grid" Graphics*, 2017. URL https://CRAN.R-project.org/package=gridExtra. R package version 2.3.

Ludwig Fahrmeir, Thomas Kneib, Stefan Lang, and Brian D Marx. Regression models. In *Regression: Models, methods and applications*. Springer, 2013.

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/.

Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning*. Springer, New York, NY, 1st edition, 2013.

Alboukadel Kassambara. *ggpubr: 'ggplot2' Based Publication Ready Plots*, 2023. URL https://CRAN.R-project.org/package=ggpubr. R package version 0.6.0.

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/.

Seoul Bike Sharing Data. South Korean goverment, 2023. URL https://archive.ics.uci.edu/dataset/560/seoul+bike+sharing+demand. [Visited on 01-07-2023].

Vincenza Torrisi, Matteo Ignaccolo, Giuseppe Inturri, Giovanni Tesoriere, and Tiziana Campisi. Exploring the factors affecting bike-sharing demand: evidence from student perceptions, usage patterns and adoption barriers. *Transportation Research Procedia*, 52:573–580, 2021. ISSN 2352-1465. doi: https://doi.org/10.1016/j.trpro.2021.01.068. URL https://www.sciencedirect.com/science/article/pii/S2352146521001095. [Online; accessed 12th July. 2023].

Hadley Wickham. The split-apply-combine strategy for data analysis. *Journal of Statistical Software*, 40(1):1–29, 2011. URL https://www.jstatsoft.org/v40/i01/.

Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. ISBN 978-3-319-24277-4. URL https://ggplot2.tidyverse.org.

Appendix

A Additional figures

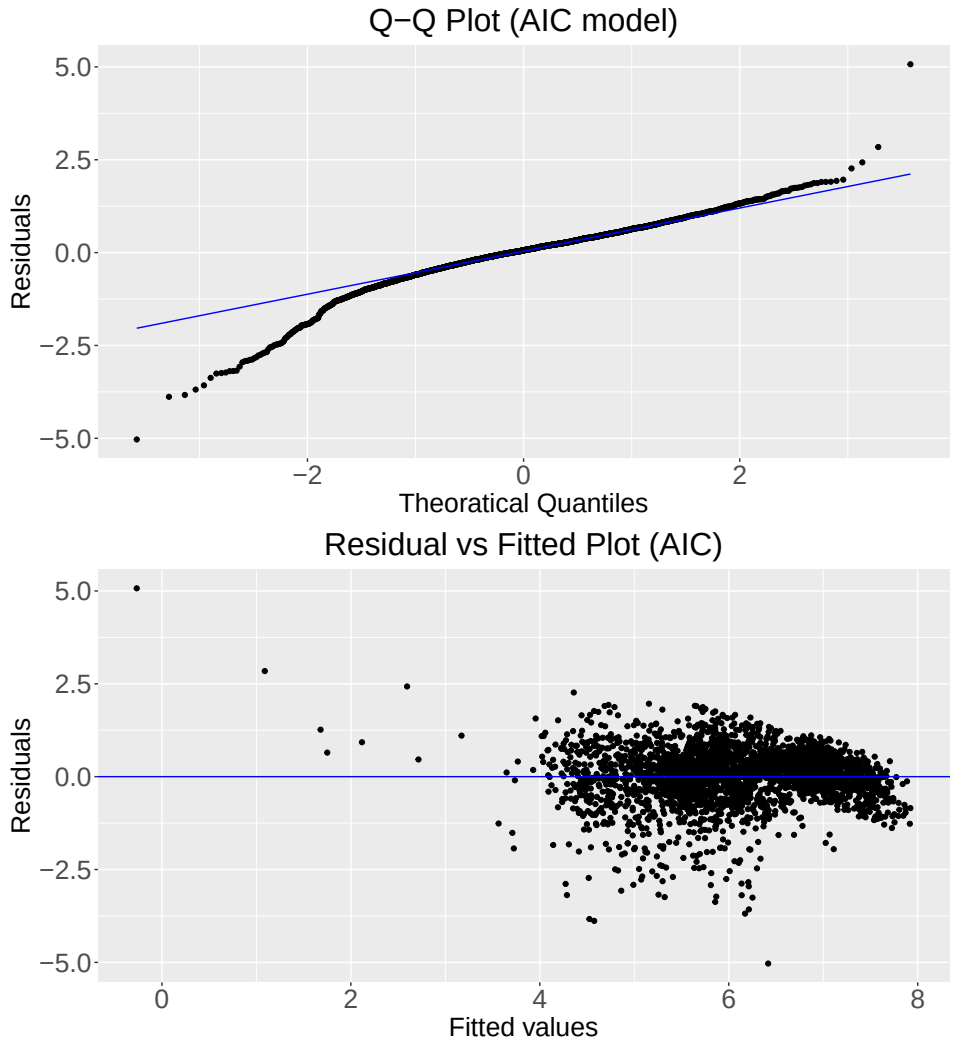

Figure 2: Assessment of Residuals in a LogBC Regression Model: Scatterplot and QQ Plot

B Additional tables

Table 3: Coefficients of the AIC Model: Estimates, t-values, p-values, and Confidence Intervals

Covariates	Estimate	t-value	Pr(>|t|)	Confidence Interval	
				2.5%	97.5%
(Intercept)	6.14	63.47	$< 2 \times 10^{-16}$	5.95	6.33
Hour	0.04	20.38	$< 2 \times 10^{-16}$	0.04	0.05
Temperature	0.04	16.763	$< 2 \times 10^{-16}$	0.0353	0.0446
Humidity	-0.017	-22.426	$< 2 \times 10^{-16}$	-0.018	-0.0157
Wind speed	-0.033	-2.26	0.024	-0.0624	-0.0044
Rainfall	-0.226	-18.455	$< 2 \times 10^{-16}$	-0.25	-0.202
Season spring	-0.27	-6.704	2.43×10^{-11}	-0.349	-0.190
Season summer	-0.173	-3.44	0.0005	-0.272	-0.0745
Season winter	-0.784	-13.77	$< 2 \times 10^{-16}$	-0.896	-0.673
No holiday	0.334	5.267	1.49×10^{-7}	0.210	0.459