TU Dortmund

Introductory Case Studies

Project III: Regression Analysis

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

Author: Mohammad Sakhawat Hossain

Matriculation No: 231838

Group number: 10

Group member: Md Shahabub Alam

July 12, 2023

Contents

1 Introduction

Bike-sharing has emerged as an important topic of interest in contemporary urban transportation, as cities around the world are looking for sustainable, eco-friendly, and efficient mobility solutions. Bike sharing system present an environmental friendly, affordable and healthy alternative to traditional mode of transportation. By analyzing various factors such as weather, season, temperature and socio-economic factors, researchers try to gain insights into the determinants of bike sharing demand. By thoroughly understanding these factors policymakers, urban planners and bike sharing operators can optimize system operations, enhance infrastructure and promote sustainable transportation options (Eren and Uz, 2020).

The purpose of this project is to develop a linear regression model that can explain the relationship between the number of rented bikes and ten other independent variables. Initially, a descriptive analysis is performed on all continuous variables. Subsequently, a linear regression model is constructed using all the variables. To select the most appropriate set of explanatory variables, a model selection technique like AIC is used. The statistical model is then analysed using the selected explanatory variables. Finally, the model is evaluated along with the multicollinearity checking.

In Section 2, a detailed explanation is given regarding the dataset and the objectives of the project. Section 3 describes statistical methods, including statistical models, criteria for model selection, statistical tests to assess the significance of parameters, confidence intervals, goodness of fit and use of the variance inflation factor (VIF) to examine multicollinearity. In Section 4, these statistical methods are applied to the given dataset, and the obtained results are analyzed. Section 5 presents the outcomes and provides a comprehensive summary of the project, followed by a discussion of potential future research directions related to this dataset.

2 Problem statement

2.1 Dataset and data quality

This report deals with the analysis of a dataset that contains logarithmic transformed rented bike counts consisting of 2905 observations and 11 variables. The data set is provided by the instructor of the course and sourced from South Korean government's

official website (Seoul Bike Sharing Data, 2023). The original data set contains 13 independent variables. However, 3 variables are removed by the instructors and therefore our data set contains only 10 independent variables along with one dependent variable. The dependent variable *log.Rented.Bike.Count* is a logarithmic transformation of the count of bikes rented every hour. The independent variable *Hour* refers to the hour of the day which is a discrete variable. As the name suggests *Humidity* refers to the humidity which is also a discrete variable. Similarly, *Visibility* is also discrete variable. On the other hand, *Temperature*, *Wind.Speed*, *Solar.Radiation*, *Rainfall*, *Snowfall* are continuous variables. *Seasons* is a categorical variable that refers to the four seasons such as winter, summer, autumn and spring. *Holiday* is also a categorical variable with two categories holiday or no holiday. The dependent variable *log.Rented.Bike.Count* is renamed as *LogBC* and used this name in our entire analysis. There is no missing value in the dataset.

2.2 Project objectives

The primary aim of this report is to establish a linear regression model that can effectively explain the relationship between the number of rented bikes and 10 other independent variables. Initially, a descriptive analysis is performed on continuous variables using scatter plots. Subsequently, a model is built by using all 10 independent variables. To identify the most suitable set of variables for the linear regression model, the best selection method is applied using the Akaike Information Criterion (AIC). The significance, confidence intervals, and goodness of fit of the coefficients in this model are interpreted. Finally, an assessment of the model has been conducted by creating a residual plot, where linearity, heteroskedasticity, and normality assumptions are investigated. Additionally, the variance inflation factor (VIF) is also used to analyze multicollinearity.

3 Statistical methods

In this section, several statistics methods are introduced which are later used in our analysis. The software R (R Core Team, 2022), ggpubr (Kassambara, 2023), ggplot (Wickham, 2016) are used for data analysis and visualization. To do data manipulation and transformation playr (Wickham, 2011) and car (Fox and Weisberg, 2019) packages are utilized. Finally, to merge the multiple plots gridextra (Auguie, 2017) has been used.

3.1 Linear regression and its assumptions

Linear regression is a statistical method used to establish a mathematical model that captures the association between a response variable and explanatory variables. If the response variable is y and k explanatory variables (also known as covariates) $x_1, x_2, ...x_k$ then the relationship can be defined as follows:

$$y = f(x_1, x_2...x_k) + \epsilon. \tag{1}$$

Where $f(x_1, x_2, ...x_k)$ is a linear function and ϵ represent an error term. ϵ accounts for the failure of the model to fit the data exactly. The linear combination of explanatory variables can be represented as follows:

$$f(x_1, x_2...x_k) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ...\beta_k x_k. \tag{2}$$

Where β_0 is intercept and $\beta_0, \beta_1...\beta_k$ are coefficients also called model parameters which need to be calculated. We can combine all response variables and error terms into a single column. In addition, all explanatory variables can also be grouped as design matrix $\mathbf{X}$ where first column reflects the effect of the intercept.

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \quad \text{and} \quad \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix} \quad , \quad \mathbf{X} = \begin{pmatrix} 1 & x_{11} & \dots & x_{1k} \\ 1 & x_{21} & \dots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \dots & x_{nk} \end{pmatrix} = \begin{pmatrix} \boldsymbol{x_1'} \\ \boldsymbol{x_2'} \\ \vdots \\ \boldsymbol{x_n'} \end{pmatrix}. \tag{3}$$

Then linear regression formula can be represented as follows in matrix notation.

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}. \tag{4}$$

Before proceeding with the linear regression model some assumptions need to be checked. The expectation or mean of the errors is zero, i.e., $E(\boldsymbol{\epsilon}) = 0$. Error terms are homoscedastic (constant variance) and uncorrelated to each other such as $Cov(\boldsymbol{\epsilon}) = \sigma^2 \boldsymbol{I}$. Error term $\boldsymbol{\epsilon}$ has to be normally distributed: $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2 \boldsymbol{I})$. It's assumed that $\boldsymbol{X}$ is full rank matrix which means all columns are linearly independent and the observation is required to be equal or higher than the number of regression coefficients (Fahrmeir et al., 2013, p. 73-75)

As it's nearly impossible to calculate the true population values $\boldsymbol{\beta}$ and σ^2, therefore least square (LS) method is applied (Fahrmeir et al., 2013, p. 77). The formula of LS is given below.

$$LS = \sum_{i=1}^{n}(y_i - \boldsymbol{x_i'}\boldsymbol{\beta}) = \sum_{i=1}^{n} \boldsymbol{\epsilon_i}^{\boldsymbol{2}} = \boldsymbol{\epsilon'}\boldsymbol{\epsilon} \tag{5}$$

To find the estimate of $\boldsymbol{\beta}$, we need to do first derivative of the LS formula with respect to $\boldsymbol{\beta}$ and set it to zero. The second derivative of this equation is calculated positively. $\hat{\boldsymbol{\beta}}$ is found after some simplification as follows:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X'X})^{-1}\boldsymbol{X'y}. \tag{6}$$

Least square method is used to estimate the regression parameters while minimizing the residuals (Fahrmeir et al., 2013, p. 105). Residuals can be represented as follows:

$$\hat{\boldsymbol{\epsilon}} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{y} - \hat{\boldsymbol{y}} \tag{7}$$

where $\boldsymbol{y}$ is actual value and $\hat{\boldsymbol{y}}$ is an estimated value. Since error term is assumed to be normally distributed i.e; $\boldsymbol{\epsilon} \sim \mathcal{N}(0,\, \sigma^2\boldsymbol{I})$, therefore, normality of error term indicates that $\boldsymbol{y}$ is normally distributed : $\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{X\beta},\, \sigma^2\boldsymbol{I})$.

Covariance of $\hat{\boldsymbol{\beta}}$ can be derived as follows where variance of y_i is σ^2.

$$Cov(\hat{\boldsymbol{\beta}}) = Cov((\boldsymbol{X'X})^{-1}\boldsymbol{X'y}) = \sigma^2(\boldsymbol{X'X})^{-1}. \tag{8}$$

Since true value of σ^2 is unknown, therefore, true value of covariance of $\hat{\boldsymbol{\beta}}$ is not estimated. However, using σ^2, covariance of $\hat{\boldsymbol{\beta}}$ can be calculated and formula is as follows.

$$\hat{\sigma}^2 = \frac{\hat{\boldsymbol{\epsilon}}'\hat{\boldsymbol{\epsilon}}}{n-p}. \tag{9}$$

Here p is the rank of the design matrix. By combining 8 and 9 we get,

$$\widehat{Cov(\hat{\boldsymbol{\beta}})} = \hat{\sigma^2}(\boldsymbol{X'X})^{-1} = \frac{\hat{\boldsymbol{\epsilon}}'\hat{\boldsymbol{\epsilon}}}{n-p}(\boldsymbol{X'X})^{-1}. \tag{10}$$

$\widehat{Var(\hat{\beta}_j)}$ can be found using the covariance matrix and it can be used to derive statistical testing and confidence interval (Fahrmeir et al., 2013, p. 117).

Since errors are assumed to be normally distributed therefore the likelihood function can be written as follows:

$$L(\boldsymbol{\beta}, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} exp(-\frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})) \tag{11}$$

and the log-likelihood function can be obtained as follows:

$$l(\boldsymbol{\beta}, \sigma^2) = -\frac{n}{2} log(2\pi) - \frac{n}{2} log(\pi^2) - \frac{1}{2\pi^2}\sigma^2(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}). \tag{12}$$

We only consider maximization of last part as it includes $\boldsymbol{\beta}$. Maximization of this part coincides with the minimization of least square method. Therefore, least square estimates is equivalent to maximum likelihood estimate of $\boldsymbol{\beta}$ (Fahrmeir et al., 2013, p. 107).

3.2 Categorical covariates

When a dataset has categorical variables, prior to merging them into the design matrix $\boldsymbol{X}$, these variables need to be encoded. Here, if x_i represents the categorical explanatory variable with m categories, then it necessitates the inclusion of an additional $m - 1$ covariates. These are also referred to as dummy variables. The category that is not included becomes the reference category, which is captured by β_0.

We would then express the coding of these as follows:

$$x_{i,1} = \begin{cases} 1 & \text{if } x_i = 1 \\ 0 & \text{if } x_i \neq 1 \end{cases} \qquad ... \qquad x_{i,m-1} = \begin{cases} 1 & \text{if } x_i = m-1 \\ 0 & \text{if } x_i \neq m-1 \end{cases}. \tag{13}$$

Here, m represents the reference category and the new dummy variables are $x_{i,1}$ through $x_{i,m-1}$. It does not matter which category is picked as reference as the results will be consistent (Fahrmeir et al., 2013, p. 97).

3.3 Best subset selection

In a model with p explanatory variables, a stepwise procedure called best subset selection is often used to find the best model. This process involves fitting all possible combinations of models with k explanatory variables, ranging from 1 to p, using least

squares regression. The model with the highest coefficient of determination (R^2) is considered the best model and denoted as M_k. This selection process is repeated for each value of k. Ultimately, we end up with a total of $p+1$ models: M_0, M_1, M_2, ..., M_p. To choose the final model from these options, various selection criteria can be employed, such as cross-validated prediction error, Akaikes Information Criterion (AIC), Bayesian Information Criterion (BIC), or Mallow's Cp statistic.

However, best subset selection has a potential drawback in terms of computational limitations. As the number of explanatory variables (p) increases, the total number of models grows substantially, reaching 2^p. This exponential growth in the number of models becomes computationally infeasible when p exceeds a certain threshold. For instance, when p becomes larger than 40, the computational algorithm becomes impractical to execute effectively (James et al., 2013, p. 227-228).

3.3.1 Akaike information criterion

The Akaike Information Criterion (AIC) serves as a selection criterion for determining the optimal model. It is represented by the following formula:

$$AIC = -2l(\hat{\boldsymbol{\beta}}_{\boldsymbol{M}}, \hat{\sigma}^2) + 2(|M| + 1) \tag{14}$$

.

Here, $\hat{\boldsymbol{\beta}}_{\boldsymbol{M}}$ and $\hat{\sigma}^2$ are estimators that maximize the log-likelihood function $l(\hat{\boldsymbol{\beta}}_{\boldsymbol{M}}, \hat{\sigma}^2)$, and $M+1$ represents the total number of parameters. In our report, since we are utilizing a linear regression model with $\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I})$ (where $\boldsymbol{y}$ denotes the response variable, $\boldsymbol{X}$ represents the design matrix, and $\boldsymbol{\beta}$ is the coefficient vector), we can simplify the log-likelihood as follows:

$$l(\hat{\boldsymbol{\beta}}_{\boldsymbol{M}}, \sigma^2) = -\frac{n}{2}\log(\hat{\sigma}^2) - \frac{1}{2\hat{\sigma}^2}(\boldsymbol{y} - \boldsymbol{X}_{\boldsymbol{M}}\hat{\boldsymbol{\beta}}_{\boldsymbol{M}})'(\boldsymbol{y} - \boldsymbol{X}_{\boldsymbol{M}}\hat{\boldsymbol{\beta}}_{\boldsymbol{M}}) \tag{15}$$

From this, we derive the following AIC formula:

$$AIC = n\log(\hat{\sigma}^2) + \frac{1}{\hat{\sigma}^2} + 2(|M| + 1) \tag{16}$$

AIC takes into account the maximum likelihood (ML) estimator $\hat{\sigma}^2 = \frac{\hat{\boldsymbol{\epsilon}}'\hat{\boldsymbol{\epsilon}}}{n}$ instead of an unbiased variance estimator.

A lower AIC value indicates a better fit. Therefore, we aim to maximize the log-likelihood to minimize the AIC, indicating an improved model fit. However, we need to consider the trade-off between a good fit and model complexity. The addition of covariates enhances the log-likelihood, signifying a better fit. Conversely, an increase in the number of parameters incurs a penalty in the AIC. Consequently, our goal is to strike a balance by selecting the model that achieves the best fit while avoiding overfitting (Fahrmeir et al., 2013, p. 148).

3.4 The coefficient of determination

R-squared, also known as the coefficient of determination, measures the extent to which the independent variables explain the proportion of variance in the dependent variable. The formula for R-squared is:

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{\sum_{i=1}^{n}\epsilon^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}. \tag{17}$$

R-squared can take values between 0 and 1, where values closer to 1 indicate a better alignment between the model and the data. A perfect fit is represented by an R-squared value of 1. However, a limitation of R-squared is that it increases even when irrelevant additional variables are included in the model. To address this issue, the adjusted R-squared method is utilized. Adjusted R-squared does not increase when irrelevant explanatory variables are added to the model. The formula for adjusted R-squared is:

$$AdjustedR^2 = 1 - \frac{\frac{\sum_{i=1}^{n}\epsilon^2}{n-k-1}}{\frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{n-1}} = 1 - \frac{(1-R^2)(n-1)}{n-k-1}, \tag{18}$$

where n is the sample size and k represents the number of explanatory variables. The difference between R-squared and adjusted R-squared increases with the inclusion of more irrelevant variables in the model. Therefore, adjusted R-squared is generally preferred as it provides a more reliable measure of goodness of fit compared to R-squared. (Fahrmeir et al., 2013, p. 112-115).

3.5 t-test

The t-test is used to determine the statistical significance of regression coefficients. The null hypothesis, denoted as $\beta_j = 0$, is tested against the alternative hypothesis, $\beta_j \neq 0$, where β_j represents a specific regression coefficient (indexed by j ranging from 1 to k). The null hypothesis is rejected when the p-value is smaller than the predetermined significance level, typically set at 0.05 which is also assumed in our analysis. Rejecting the null hypothesis implies that the explanatory variable x_j has a significant influence on the response variable.

The t-statistic is used to express the test, and it is defined as follows:

$$t_j = \frac{\hat{\beta}_j}{\sqrt{\widehat{Var(\hat{\beta}_j)}}} = \frac{\hat{\beta}_j}{se_j} \sim t_{n-p}. \tag{19}$$

Here, se_j represents the standard error of $\hat{\beta}_j$, $\widehat{Var(\hat{\beta}_j)}$ is the estimated variance of the individual coefficient, and t_j follows a t-distribution with $n - p$ degrees of freedom. To assess the significance, the calculated t-value is compared to the critical t-value corresponding to the $(1-\frac{\alpha}{2})$-th quantile of the t-distribution with $n-p$ degrees of freedom. The null hypothesis is rejected when the absolute value of t_j exceeds $t_{1-\frac{\alpha}{2}}(n-p)$. A larger t-value corresponds to a smaller p-value, indicating stronger evidence to reject the null hypothesis. Therefore, it can be concluded that the coefficient β_j significantly differs from zero (Fahrmeir et al., 2013, p. 131).

3.6 Confidence interval

To construct a confidence interval for a specific parameter β_j (where j ranges from 0 to k), the test statistic t_j is used, corresponding to the hypothesis test $H_0 : \beta_j = 0$. As mentioned earlier, the null hypothesis is rejected when the absolute value of t_j exceeds $t_{1-\frac{\alpha}{2}}(n-p)$. The test is designed such that the probability of rejecting the null hypothesis when it is true is α. This can be expressed as:

$$P(|t_j| > t_{n-p}(1 - \frac{\alpha}{2})) = \alpha. \tag{20}$$

Conversely, the probability of not rejecting the null hypothesis when it is true is given by:

$$P(|t_j| < t_{n-p}(1 - \frac{\alpha}{2})) = 1 - \alpha. \tag{21}$$

After some calculations, the resulting $(1-\alpha)$ confidence interval for β_j can be expressed as:

$$[\hat{\beta}_j - t_{n-p}(1 - \frac{\alpha}{2})se_j, \hat{\beta}_j + t_{n-p}(1 - \frac{\alpha}{2})se_j]. \tag{22}$$

This interval provides an estimate of the range within which the true value of β_j is likely to fall with a confidence level of $(1 - \alpha)$ (Fahrmeir et al., 2013, p. 136).

3.7 Multicollinearity analysis

Multicollinearity is a phenomenon that can occur when there is high correlation among the explanatory variables, although it is generally undesirable in regression analysis. This high correlation can lead to inaccurate estimates of regression coefficients. Multicollinearity is directly related to the assumption in linear regression that the design matrix must have full rank. In the presence of multicollinearity, the inverse of the design matrix, denoted as $(X'X)^{-1}$, does not exist. Therefore, the least squares method is ineffective. The extent of multicollinearity can be examined using the variance formula for $\hat{\beta}_j$:

$$Var(\hat{\beta}_j) = \frac{\sigma^2}{(1 - R_j^2)\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2}. \tag{23}$$

The correlation between an explanatory variable x_j and the other variables is measured by R_j^2. A higher value of $Var(\hat{\beta}_j)$ indicates a stronger linear dependence of x_j on the other explanatory variables. To assess the extent of the increase in $Var(\hat{\beta}_j)$ due to the linear dependence between x_j and other variables, the variance inflation factor (VIF) is utilized. The VIF is formulated as:

$$VIF_j = \frac{1}{1 - R_j^2}. \tag{24}$$

Since a large correlation between x_j and other variables corresponds to a high value of R_j^2, the VIF for variable x_j increases accordingly based on the aforementioned formula. A

VIF value exceeding 10 is typically considered as evidence of collinearity issues. According to R help page and original paper (Fox and Monette, 1992, p. 178-183), genralized variance inflation factor (GVIF) corresponds to VIF for more than one coefficient. One approach to address the problem of collinearity is to remove the affected explanatory variables from the analysis (Fahrmeir et al., 2013, p. 158).

3.8 Residuals vs. fitted plot

Residuals refer to the differences between the actual value and the predicted or fitted value. A Residuals vs. Fitted plot is a visual representation in the form of a scatter plot, where the fitted values and their associated residuals are plotted alongside a horizontal reference line. The formula for calculating residuals is $\epsilon_i = y_i - \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}$. This plot serves two purposes: to validate the assumption of homoscedastic error variances in a linear model and to evaluate the overall fit of the model. The x-axis of the plot represents the fitted axis where residuals are plotted on the y-axis. Should the data points exhibit a random and consistent spread around the reference line, it indicates homoscedastic error variances. Conversely, if the data points do not conform to this pattern, then the error variances are classified as heteroscedastic (Fahrmeir et al., 2013, p. 183).

4 Statistical analysis

4.1 Descriptive analysis

We conduct a descriptive analysis using all the continuous variables. Based on the correlation coefficients and Figure 2 on page 17 in the appendix section, we observe that LogBC and temperature have a moderately strong positive correlation (0.56), followed by LogBC and solar radiation. This suggests that as the temperature and solar radiation increase, the number of rented bike counts tends to increase as well. LogBC and wind speed have a weak positive correlation (0.11) and same scenario can be found for the Visibility variable. This suggests that there is a slight tendency for the number of rented bike counts to increase with higher wind speeds and visibility, although the correlation is not very strong. LogBC and rainfall have a moderate negative correlation (-0.25) followed by LogBC and humidity where LogBC and snowfall have a weak negative

correlation (-0.18). This indicates that as rainfall humidity and snowfall increase, the number of rented bike counts tends to decrease.

From Table 2 on page 18 in the appendix section, we observe that lowest temperature is -17.5 degree Celsius where highest is 38 degree Celsius. Highest visibility is 2000 followed by humidity which is 98%. Variance and IQR is also high for the visibility. Lowest average is recorded for the snowfall which is .08 cm. The median value for rainfall and snowfall are 0 where the mean value is 0.15 and 0.08 respectively which indicates mean is affected by the skewness of the distribution caused by the extreme values that can also be seen from the scatter plot.

4.2 Linear regression based on all variables

A linear regression model is constructed with "LogBC" as the reference variable and the remaining 10 variables serving as covariates. The equation for this model incorporates estimated parameters and can be expressed as follows:

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

The intercept term in the regression model is estimated as 6.21, representing expected value of the count of bikes rented when all predictor variables in the model are held constant. This coefficient is significantly different than zero as corresponding p-value is less than assumed significance level 0.05. The coefficient for hour and temperature is 0.04 which means for each unit increase in the hour, the count of rented bikes is expected to increase by approximately 4% when all other variables are considered constant. Both coefficients are significantly different than zero at significance level 0.05. In contrast, the coefficient for humidity, wind speed, visibility, solar radiation, rainfall, snowfall is negative where the values are - 0.018, - 0.028, -1.73×10^{-5}, - 0.025, - 0.226, - 0.006 respectively indicating for each unit increase in each of these variables, the count of rented bikes is expected to decrease by approximately 1.8%, 2.8%, 0.001734%, 2.5%, 22.6%, 0.6%. Holiday and seasons are two categorical variables where holiday and season autumn are reference categories respectively. The coefficients for seasons spring, seasons summer, and season winter are negative, indicating a decrease in the count of rented bikes compared to the reference season (autumn).

4.3 Best subset selection with AIC

To identify a suitable combination of explanatory variables that can effectively explain the response variable *LogBC* the best subset selection method is applied. With 10 explanatory variables in consideration, a total of $2^{10} - 1 = 1023$ models are evaluated, subtracting the null model that only includes an intercept term. All models are compared using AIC. Lowest AIC is 6521.45 which belongs to the model that contains a set of 9 explanatory variables which are *hour*, *temperature*, *humidity*, *wind speed*, *rainfall*, *season spring*, *season winter*, *season summer* and *no holiday*. Two variables *visibility* and *solar radiation* variables are not selected by AIC.

4.3.1 Analysis and interpretation of AIC model

Table 1 shows the summary of the AIC model where the parameter estimates, their t-values, p-values and confidence intervals for $\alpha = 0.5$ are shown. The estimate of the intercept is 6.14 which is significantly different than 0 as corresponding p-value is lower than 0.05. All other coefficient of the variables *hour*, *temperature*, *humidity*, *wind speed*, *rainfall*, *season spring* , *season winter*, *season summer* and *no holiday* also have smaller p-value which is less than 0.05 and thus significantly different than 0. The logarithmic transformation of the coefficient for hours and temperature is 0.04. This suggests that, when all other variables are held constant, a 4% increase in the count of rented bikes can be expected.

Covariates	Estimate	t-value	Pr(>|t|)	Confidence Interval	
				2.5%	97.5%
(Intercept)	6.14	63.47	<2e-16	5.95	6.33
Hour	0.04	20.38	<2e-16	0.04	0.05
Temperature	0.04	16.763	<2e-16	0.0353	0.0446
Humidity	-0.017	-22.426	<2e-16	-0.018	-0.0157
Wind speed	-0.033	-2.26	.024	-0.0624	-0.0044
Rainfall	-0.226	-18.455	<2e-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	.0005	-0.272	-0.0745
Season winter	-0.784	-13.77	<2e-16	-0.896	-0.673
No holiday	0.334	5.267	1.49×10^{-7}	0.210	0.459

Table 1: Coefficients of the AIC model: Estimate, t-value (t-statistic), and Pr(>|t|) (p-value) and Confidence intervals

The remaining coefficients indicate that for each unit increase in any of them, while keeping the other variables constant, the count of rented bikes decreases. Here, season autumn is reference category. Season spring, summer and winter decrease count of bike rent by 27%, 17% and 78% respectively. Holiday is reference category where during working day (no holiday) rented bike count increases by approximately 33% when all other coefficient are considered constant. Last two columns of Table 1 shows confidence interval of coefficients while the second last column denoted as 2.5% is lower confidence limit and last column 97.5% represent upper confidence limit. It indicates that 95% of the values of each variable lie in the range of lower and upper limitations. For instance when temperature increases 1 degree celsius the value of logBC increases between 3% and 4%. All the variables are statistically significant as before since none of them contain 0 in their 95% confidence interval. Adjusted R^2 is used to measure the goodness of fit. We get the value of adjusted R^2 as 0.59. It implies that explanatory variables can explain approximately 59% percent of the variance of response variable.

4.4 Model evaluation and mulitcollinearity

To determine whether the error terms exhibit homoscedasticity, lower part of Figure 1 shows a plot of residuals versus fitted values. The plot indicates that the variance remains approximately constant around the zero line. A roughly consistent pattern is observed both below and above the horizontal axis, suggesting homogeneity and supporting the assumption of homoscedasticity to some extent as a few points are also scattered on the top left side. The presence of an overall horizontal pattern in the plot indicates that the assumption of linearity can be considered satisfied. Additionally, the fact that the residuals are centered around zero, despite some deviations in the lower pattern, implies that the assumption of an average error term of zero $E(\epsilon) = 0$ is satisfied. Furthermore, no trend is observed in the plot, indicating that the error terms are uncorrelated. The normality assumption of the error term can be assessed using a Q-Q plot, which is also presented in the upper part of Figure 1. The Q-Q plot shows that few points on the lower and upper sides of the plot don't align with the reference line suggesting that normality assumptions don't hold ultimately.

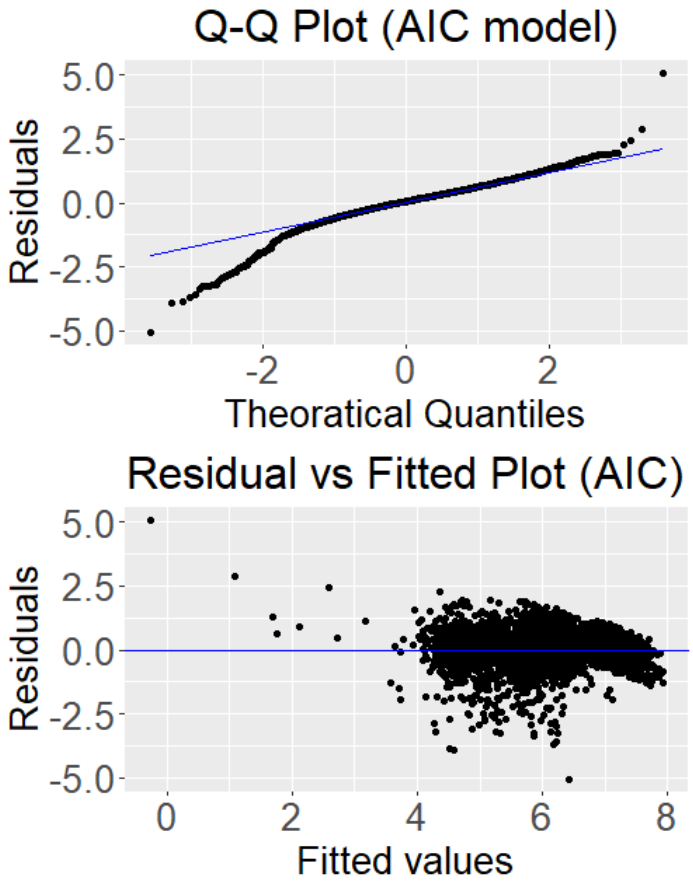

Figure 1: QQ Plots and residual vs fitted plot for AIC model

Considering these observations, it is noticeable that the response variable *LogBC* can be selected for further analysis using a linear regression model. This variable doesn't satisfy the key assumptions of linear regression to a reasonable extent.

As mentioned earlier, high correlation among the explanatory variables leads to incorrect estimation of regression coefficient. Therefore, it is necessary to check if there exists any high correlation between the variables. GVIF (corresponds to VIF for more than one coeffient) is used to detect multicollinearity. Table 3 on page 18 in Appendix shows the calculated GVIF for all variables. The maximum GVIF is observed for the variable *temperature*. Since no GVIF exceeds 10, it can be said that there is no collinearity between the variables and there is no need to omit any variable.

5 Summary

This project aims to find a linear regression model which can estimate number of rented bike in Seol city of South Korea based on some influencing variables. The dataset used in this analysis contains 2905 observations, 10 explanatory variables or covariates and one response variable. Some descriptive analysis was made on the dataset using scatter plots before building the regression models. Then, a regression model was built which included all explanatory variables. The best subset selection method was applied to find the best subset of the covariates which fitted the data better. This method was applied based on Akaike information criterion (AIC). The best model found based on AIC included *temperature, humidity windspeed, rainfall, snowfall, seasons* and *holiday* where *windspeed*, *snowfall* and *solar radiation* were not included in the AIC model. This model was analyzed further and the assumptions of a linear regression model were verified. The Q-Q plot showed the residuals don't come from a normal distribution and the residual vs. fitted plot didn't perfectly show homoscedasticity of the errors. In addition, no multicollinearity observed in the data. The estimated coefficient of *seasons* showed that during autumn people rented bike more than summer, spring or winter seasons. Similarly, increases of *humidity*, *wind speed* and *rainfall* has negative effect on renting bike. In contrast, as time increases during the day and temperature rises the percentage of renting bike increases by 4%. The p-value of all estimated coefficients were less than the significance level of 0.05. This proved that all of the coefficients were significant. Also, the estimated coefficients were in their 95% confidence intervals. The coefficient of determination of the model was 0.59 which indicates that the AIC model fitted the data well.

Since the normality assumption of the regression model is violated, it is important to consider this important implication. The violation could indicate the presence of outliers or the need for non-linear transformations of the variables. Although there is some extent of homoscedasticity, it is essential to assess the overall pattern of heteroscedasticity in the model which also demonstrates an important limitation of the overall data. In addition, the data set needs to be validated over an independent dataset to assess the robustness and generalizability of the results.

Bibliography

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

Ezgi Eren and Volkan Emre Uz. A review on bike-sharing: The factors affecting bike-sharing demand. *Sustainable cities and society*, 54:101882, 2020.

Ludwig Fahrmeir, Thomas Kneib, Stefan Lang, Brian Marx, Ludwig Fahrmeir, Thomas Kneib, Stefan Lang, and Brian Marx. *Regression models*. Springer, 2013.

John Fox and Georges Monette. Generalized collinearity diagnostics. *Journal of the American Statistical Association*, 1992. doi: 10.1080/01621459.1992.10475190. URL https://www.tandfonline.com/doi/abs/10.1080/01621459.1992.10475190.

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, Robert Tibshirani, et al. *An introduction to statistical learning*, volume 112. Springer, 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].

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 figure

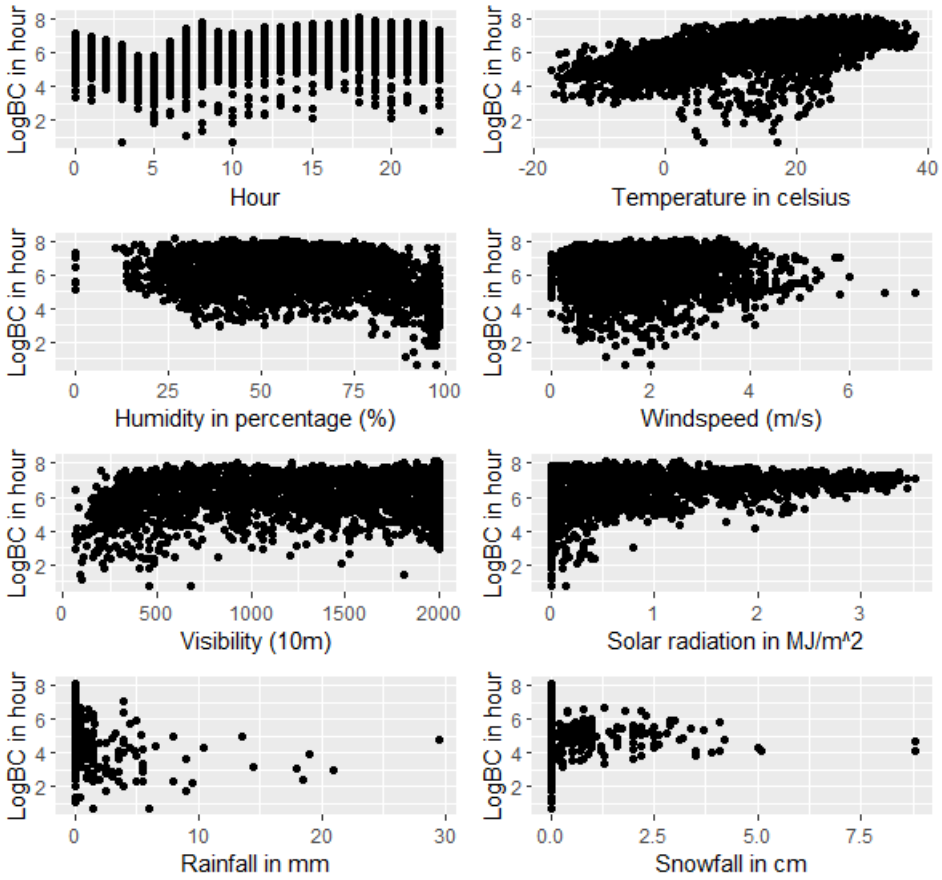

Figure 2: Scatter plot of all continuous variables with respect to log-transformed rented bike count

B Additional tables

	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

Table 2: Summary table of all continuous variables

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

Table 3: Variance inflation factor for multicollinearity