Project: Regression Analysis

Author: Opeyemi Ayanwale

July 6, 2023

Contents

1 Introduction1

2 Problem Statement2

2.1 Dataset and Data Quality 2
2.2 Project Objectives 2

3 Statistical Methods3

3.1 Multiple Linear Regression and Assumptions 3
3.1.1 Estimating the regression coefficient 4
3.1.2 Significance of parameter estimates and confidence interval 4
3.2 Variable Selection with Akaike Information Criterion (AIC) 5
3.3 Goodness-of-Fit: R^2 or Adjusted R^2 6
3.4 Residual Plot and Model Diagnostics 7

4 Statistical Analysis8

4.1 Descriptive Statistics 8
4.2 Regression Model of Bike Count Based on All Variables. 11
4.3 Subset Selection Based on Stepwise AIC 12
4.4 Residual Plot for Model Evaluation 14

5 Summary14

Bibliography16

Appendix18

A Additional figures 18
B Additional tables 19

1 Introduction

Regression analysis is a statistical method for investigating and explaining the relation- ship between dependent and independent variables. It can be used to describe, estimate, predict, and control the effect of one or more independent variables on the dependent variable. The dependent variable also known as the response variable is the outcome that we want to predict and the independent or explanatory variables are variables that influence the dependent variable (Ali and Younas, 2021).

Regression analysis will be used in this report to predict bike rental demand in Seoul based on weather and holiday data. Bike sharing systems are systems that allow people to rent public bikes to get to their destinations without having to own a bike themselves. Seoul, South Korea's capital with many other urban cities has introduced rental bikes to improve mobility comfort. The original dataset of Seoul Bike Sharing Demand which is obtained from the South Korean government, is available on the UCI machine learning repository website (UC Irvine ML Repository, 2020).

The primary goal of this report is to examine the relationship between the dependent and independent variables and to find a suitable subset of explanatory/independent variables from the bicycle sharing rentals dataset, which includes 13 independent variables and 1 dependent variable (UC Irvine ML Repository, 2020). For the purpose of this report, only 10 independent variables will be considered for the analysis.

Descriptive statistics were used to describe the distribution of the variables. Scatter plots are used to visualize the data, correlation analysis is used to provide insight into the relationship between the variables, and regression analysis is used to determine the strength of the relationship between the variables.

A suitable subset of the independent variables in the regression model is found using the stepwise selection method with the best fit based on Akaike information criterion (AIC). To check for patterns of linearity, heteroskedasticity, and normality, residual plots for model evaluation were plotted. The variance inflation factor (VIF) was also used to investigate multicollinearity.

The second section provides a more detailed overview of the dataset, variables descrip- tion, and information on data quality. The statistical analysis methods are presented and explained in the third section. And in the fourth section, the presentation, analysis, and interpretation of the results are presented. Finally, in the fifth section, the main findings are summarized.

2 Problem Statement

2.1 Dataset and Data Quality

The dataset used in this report is an extract of the Seoul Bike Sharing Demand dataset. It includes 13 independent variables along with 1 dependent variable. However, 3 variables have been excluded due to mul- ticollinearity and the inclusion of a date variable, reducing the independent variable to 10, two of which are dummy variables. Further changes were made to improve the data properties as the original dependent variable was not normally distributed and contained a significant number of zeros. These zero-value entries were removed, and a logarithmic transformation was used on the dependent variable.

The dependent variable is the log.Rented.Bike.Count (natural logarithm of the number of bike rentals). The independent variables are Hour (Hour of the day), Temperature (in Celsius ($^{\circ}C$)), Humidity (in percentage(%)), Windspeed (m/s), Visibility (10m), Solar radiation (Megajoules per square meter MJ/m^2), Rainfall (in Millimeter (mm)), Snowfall (in Centimeter (cm)), Seasons (Winter, Spring, Summer, Autumn) and Holiday (Holiday/No holiday).

All variables of interest are numerical, with the exception of seasons and holidays, which are nominal variables. There are no missing data in the dataset. The original data with corresponding weather data, number of bikes rented per hour, and holiday information are available for educational research on the data website (UC Irvine ML Repository, 2020).

2.2 Project Objectives

The primary goal of this report is to examine the relationship between the depen- dent/response and independent/explanatory variables, as well as to select a suitable subset of independent variables from the bicycle sharing rentals dataset for the best subset model. The second goal is to evaluate the model, create a residual plot and examine it for patterns of linearity, heteroskedasticity, and normality. The variance in- flation factor (VIF) will be used to test for multicollinearity. The statistical method session will go over the methods used to meet the objectives of this report in detail.

3 Statistical Methods

This section presents several statistical methods that will be used to analyze the dataset based on the objectives of this report. All analyses and visualizations were performed using the statistical software R, version 4.2.3 (R Development Core Team, 2023), uti- lizing the basic r packages with DescTools (Signorell, 2023), olsrr (Hebbali, 2020a), car (Fox and Weisberg, 2019), pysch (William Revelle, 2023), mass (Venables and Ripley, 2002), broom (Robinson, 2023), knitr (Xie, 2023), and xtable (Scott, 2019) packages respectively.

3.1 Multiple Linear Regression and Assumptions

Simple linear regression is an effective method for predicting a response with a single predictor variable. Most of the time, there is more than one predictor (independent variable), and a simple linear regression must be extended to accommodate multiple predictors, this extension is known as **multiple linear regression**. Multiple linear regression is a statistical method that predicts the outcome of a dependent variable by using several independent variables, such as how weather and holiday affect the bike rental demand (James et al., 2013, p. 71).

Mathematical formula:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots \beta_p X_p + \epsilon$$

where Y = dependent variable, β_0 = intercept, β_p = slope coefficient, X =independent variable and ϵ = residual/model error

Assumption of multiple linear regression

- There is a linear relationship between the dependent variable and the independent variables.
- The residuals should be normally distributed with a mean of zero.
- The independent variables are not overly correlated with one another.

3.1.1 Estimating the regression coefficient

Coefficient estimates β are the values that minimize the sample's sum of squared residuals/errors. We can use the formula below to make predictions given the estimate $\hat{\beta}_0$, $\hat{\beta}_1$, ..., $\hat{\beta}_p$.

Mathematical formula:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_p x_p$$

where $\hat{y}$ = a prediction of Y based on $X = x$, $\hat{\beta}_0$-$\hat{\beta}_p$ = estimate for the model coefficient/parameters, x_i for i = 1,..,p is ith value of x (James et al., 2013, p. 61). The parameters are estimated using the least squares coefficient estimates and we use β_0, β_1, ..., β_p tominimizethesumofsquaredresiduals(RSS)(Jamesetal.,2013,p.72).

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y_i})^2$$

$$= \sum_{i=1}^{n} (y_i \quad \beta_0 \quad \beta_1 x_{i1} \quad \beta_2 x_{i2} \quad \ldots \quad \beta_p x_{ip})^2$$

The residual error is calculated as $\epsilon_i = y_i - \hat{y}_i$, the difference between an actual and a predicted value of y (James et al., 2013, p. 62).

3.1.2 Significance of parameter estimates and confidence interval

In a multiple regression model, we want to know if a specific x variable can help us predict or explain the y variable; for example, if the model has three x variables, we can test if variable x_1 is a useful predictor variable in the model.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$$

The standard hypothesis test compares the null hypothesis to the alternative hypothesis. This is equivalent to:

$$H_0 : \beta_1 = 0 \qquad versus \qquad H_a : \beta_1 \neq 0$$

To test the null hypothesis, we need to see if $\hat{\beta}_1 x_1$, our estimate for β_1 is sufficiently far from zero which is determined by the standard error of estimate ($SE(\hat{\beta}_1)$). If β_1 = 0, the model is reduced to $y = \beta_0 + \epsilon$ and X does not have a relationship with Y. If $SE(\hat{\beta}_1)$ is small, $\hat{\beta}_1$ may provide strong evidence that $\beta_1 \neq 0$ hence there is a relationship between X and Y. Alternatively, if $SE(\hat{\beta}_1)$ is large, then $\hat{\beta}_1$ must be large in absolute value to reject the null hypothesis (James et al., 2013, p. 67). Statistical software willgenerate p-values for all coefficients in the model to perform the test, which measures the number of standard deviations that $\hat{\beta}_1$ is away from zero. Each p-value will be calculated using the *t-statistic* (James et al., 2013, p. 67).

$$t = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)}$$

Confidence interval

Confidence intervals can be computed using standard errors (SE). A 95% confidence interval is defined as a range of values with a 95% probability that contains the true unknown value of the parameter. The 95% confidence interval for β_1 and β_0 roughly takes the form below, the factor "2" slightly varies depending on the number of observations(n) in the linear regression (James et al., 2013, p. 66).

$$\hat{\beta}_1 \pm 2.\, SE(\hat{\beta_1}) \qquad or \qquad \hat{\beta}_0 \pm 2.\, SE(\hat{\beta_0})$$

3.2 Variable Selection with Akaike Information Criterion (AIC)

Variable selection refers to the process of determining which predictors are associated with the response in order to fit a single model that only includes those predictors. There are four major variable selection methods: forward selection, backward selection, stepwise selection, and best subset selection. Forward selection begins with no variables in the model and gradually adds variables to the model, whereas backward selection begins with a full model that takes into account all of the variables to be included in the model, and stepwise selection is a combination of forward and backward selection (James et al., 2013, p. 78-79). The best subset selection method involves testing every possible combination of predictor variables and then selecting the best model (James et al., 2013, p. 205).

Stepwise Selection with AIC: In this report, the stepwise selection method will be used to find the suitable subset from the model because it is easy to conduct automati- cally in most statistical packages, improves generability, and yields a simple model that is easy to understand. This method is a combination of forward and backward selection procedures that allow moving in both directions, adding and removing variables at dif- ferent steps with AIC as the selection criteria. The AIC criterion is defined for a large class of models fit by maximum likelihood. In general, a small value indicates a model with a low test error, therefore, the model with the lowest AIC value is selected as the best model (James et al., 2013, p. 212).

Mathematical formula:

$$AIC = \frac{1}{n\hat{\sigma}^2}(RSS + 2d\hat{\sigma}^2)$$

where n = number of observations, σ^2 = estimated variance of the residuals, RSS = Residual Sum of Squares and d = number of parameters (coefficients) in the model.

3.3 Goodness-of-Fit: R^2 or Adjusted R^2

Another popular method for choosing the best model with varying numbers of variables is the adjusted R^2 statistic.

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

Where TSS = total sum of squares for the response. R^2 cannot be used to choose between a set of models with varying numbers of variables because it does not measure the goodness of fit or predictive error but rather measure how strongly two variables are correlated hence the introduction of adjusted R^2 which considers and tests different independent variables against the model (James et al., 2013, p. 210)

$$Adjusted\ R^2 = 1 - \frac{RSS/(n - d - 1)}{TSS/(n - 1)}$$

A large adjusted R^2 value indicates a model with a small test error and correct variables that are free of noise (James et al., 2013, p. 212).

3.4 Residual Plot and Model Diagnostics

Residuals plots are used to check if the model works well on our data; they may reveal patterns in the data that are not explained by the fitted model. Model diagnostics is a procedure for evaluating the results of regression analysis and detecting any deviation from regression assumptions (UVA Library StatLab, 2015).

Check Non-linearity: Using Residual versus Fitted plot

Non-linearity, unequal error variances, and outliers can all be detected using a scatter plot of residuals on the y-axis and fitted values on the x-axis (Hebbali, 2020c). The residual versus fitted plot shows whether or not the residuals have non-linear patterns (UVA Library StatLab, 2015).

Check Heteroskedasticity: Using Scale versus Location plot

This plot shows whether residuals are distributed evenly across predictor ranges. It is used to check whether the assumption of equal variance (homoscedasticity) or unequal variance (heteroskedasticity) is correct. Fitted values are shown on the x-axis, and the square root of standardised residuals is shown on the y-axis of the plot (UVA Library StatLab, 2015).

Check Normality: Using Quantile-Quantile plots of residual

A "quantile-quantile" plot (QQ-plot) can be used to check if a sample violates the normality assumption. This plot shows whether the residuals are normally distributed by checking if the points lie approximately on the straight reference line in the plot (UVA Library StatLab, 2015).

Check Multicollinearity: Check the correlation between covariates or using the Variance Inflation Factor (VIF)

Multicollinearity occurs when two or more variables in a regression model have a high correlation with one another, and the regression estimate becomes unstable with a high standard error. To detect multicollinearity in a dataset, we can compute the Variance Inflation Factor (VIF) for each independent variable. We can compute VIF by regressing the K^{th} predictor on the rest of the predictors in the model and compute the R^2 (Hebbali, 2020b)

$$VIF = \frac{1}{1 - R_k^2} \qquad where \qquad 1 - R_k^2 = Tolerance$$

4 Statistical Analysis

4.1 Descriptive Statistics

Descriptive statistics are used to describe the distribution of the Seoul bike-sharing demand dataset. From Table 1, the average log-transformed number of rented bikes is 6.09, with a standard deviation of 1.16. The minimum value is 0.69, indicating that there were no bike rentals during those hours, and the maximum value is 8.12. The average hour of the day is around 11.58, ranging from 0 to 23 and the average temperature is
12.81 degrees Celsius, with a standard deviation of 12.22. Temperature ranges from
-17.50 to 38.00. The average humidity is 57.73%, with a standard deviation of 20.57. The humidity levels range from 0 to 98, indicating a wide range of humidity levels. The average wind speed is 1.73 m/s, with a standard deviation of 1.03 while the average visibility is 1440.73 meters, with a standard deviation of 607.94, and ranges from 63 to 2000 meters. The average solar radiation is 0.58 MJ/m^2, with a standard deviation of
0.87 and the values range from 0 to 3.52. The average amount of rain is 0.15 mm, with a standard deviation of 1.16, and this ranges from 0 to 29.50, with the distribution being highly positively skewed, indicating that the majority of observations had little or no rain. The average amount of snowfall is 0.08 cm, with a standard deviation of 0.46 cm.

Table 1:Descriptive Statistics of Seoul Bike Sharing Demand Dataset

	n	mean	sd	skew	min	max
log.Rented.Bike.Count	2905	6.09	1.16	-0.83	0.69	8.12
Hour	2905	11.58	6.87	-0.01	0.00	23.00
Temperature	2905	12.81	12.22	-0.18	-17.50	38.00
Humidity	2905	57.73	20.57	0.07	0.00	98.00
Wind.speed	2905	1.73	1.03	0.91	0.00	7.30
Visibility	2905	1440.73	607.94	-0.71	63.00	2000.00
Solar.Radiation	2905	0.58	0.87	1.50	0.00	3.52
Rainfall	2905	0.15	1.16	14.45	0.00	29.50
Snowfall	2905	0.08	0.46	9.01	0.00	8.80
Seasons*	2905	2.55	1.10	-0.06	1.00	4.00
Holiday*	2905	1.95	0.22	-4.08	1.00	2.00

Similar to rainfall, the distribution of snowfall is highly positively skewed. Season is a categorical variable and values range from 1 to 4, representing Winter, Spring, Summer, and Autumn. Holiday is another categorical variable and values range from 1 to 2, indicating whether or not the day is a holiday (see Table 1).

Scatterplots and boxplots are further used to visualize the distribution and relationship between independent variables and the dependent variable. Figure 1 depicts scatter plots that provide insights into the relationship between the rented bike count and various factors. The rented bike count shows a distinct pattern with darker linear bands, indicating that certain hours of the day have higher bike rentals. The rented bike count forms a cloud-like distribution across a wide temperature range (-20 to 40 degrees Celsius), suggesting that temperature influences bike rentals but without a clear linear trend. The bike count distribution shows a prominent cloud up to a wind speed of around 5 meters per second, after which the distribution becomes more scattered. Higher wind speeds may have a smaller impact on bike rentals.

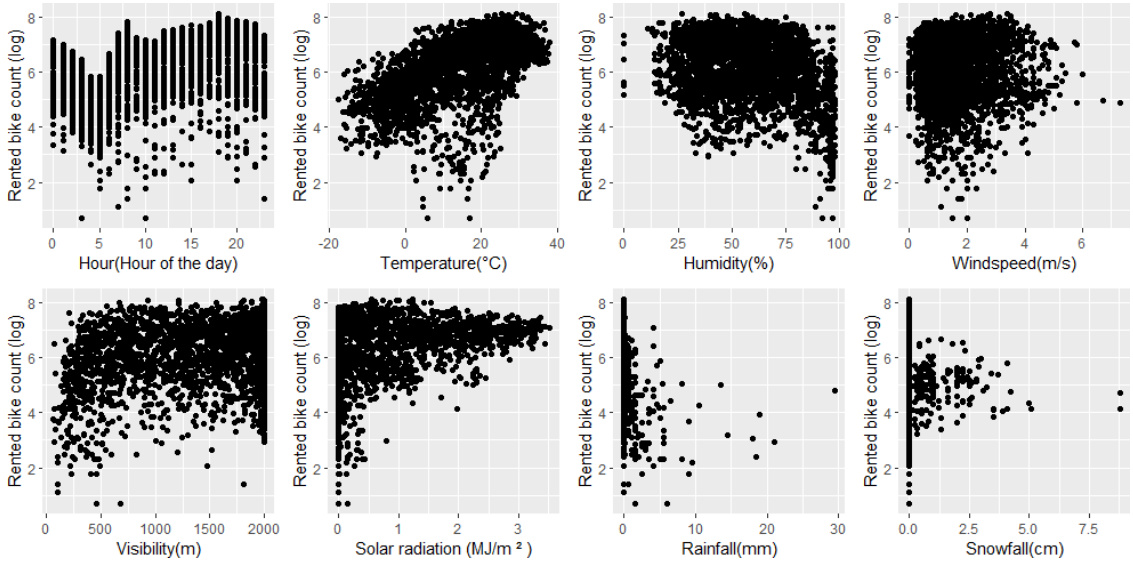

Figure 1:Scatterplot of Bike Sharing Demand Data.

The rented bike count is spread across the humidity range of 20 to 100 percent. The data points are more concentrated at higher humidity levels, indicating some relationship between humidity and bike rentals. The bike count distribution forms a cloud-like pattern with visibility ranging from 0 to 2000 meters. There is no clear linear trend, suggesting that visibility alone may not be a strong predictor of bike rentals. A considerable num- ber of bike rentals occur on days with low solar radiation (0 MJ/mš), and the count decreases as solar radiation increases. There may be an inverse relationship between solar radiation and bike rentals. The majority of bike rentals occur on days with no or very low rainfall (0 mm). As rainfall increases, the number of rentals becomes scattered, indicating a possible impact of rainfall on bike rentals. Similar to rainfall, most bike

rentals occur on days with no snowfall (0 cm). The count decreases as snowfall increases, with a few data points above 5cm suggesting potential outliers (see Figure 1).

Many other factors influence bike rental demand, including weather and holidays. Figure 2 shows that rental demand is higher in the summer, autumn, and spring than in the winter. We can deduce that people prefer to cycle more during warmer weather, such as summer and whereas bikes are rented at the lowest rates during winter. This may be due to the cold weather and snowfall that occurred during the winter. There are also a few outliers throughout the season, which could be due to other underlying factors. Similarly, bike demand is higher in autumn than in spring, despite our expectation that demand would be higher in spring which might be because Spring has more rainy days than Autumn. Table 2 summarizes the median values for each of the four seasons, with summer having the highest median value of 6.86 and autumn at 6.69 (see Table 2 and Figure 2).

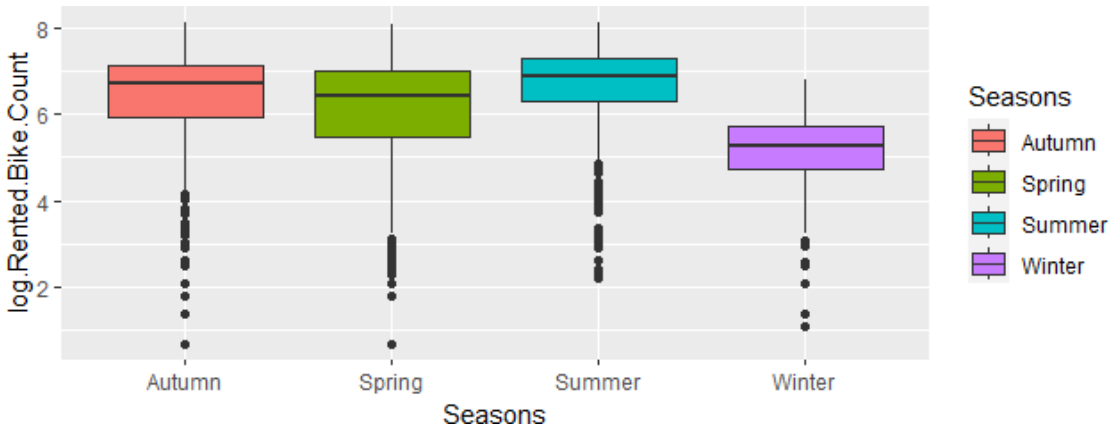

Figure 2:Boxplot of Seasonal Bike Count.

Table 2:Medians Values for the Seasons

Seasons	log.Rented.Bike.Count
Autumn	6.6970
Spring	6.4118
Summer	6.8627
Winter	5.2832

Figure 3 depicts the holiday boxplot, which shows how the holiday/no holiday affects the demand for bike rentals. Comparing the median values of the two categories, bike count tends to be higher on days without a holiday (6.322) compared to days with a holiday (5.6185) (see Figure 3 and Table 3).

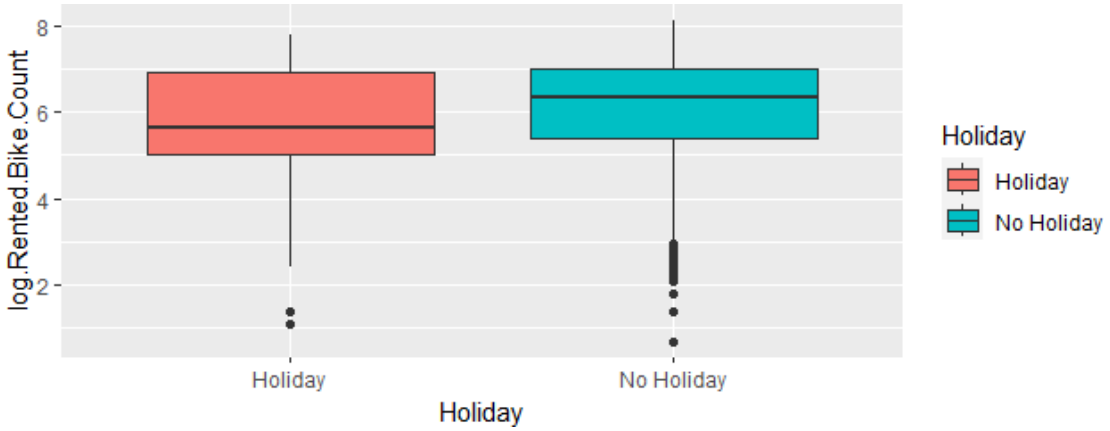

Figure 3:Boxplot of Holiday Bike Count.

Table 3:Medians Values for Holiday

Holiday Status	log.Rented.Bike.Count
Holiday	5.6185
No Holiday	6.3225

Table 4 depicts the independent variables and their corresponding correlation coeffi- cients. Hour, Temperature, Windspeed, Visibility, and Solar-Radiation have positive correlations with the response variable (log.Rented.Bike.Count) while Humidity, Rain- fall, and Snowfall have negative correlations (see Table 4).

Table 4:Correlation Coefficient Table

Covariate	Correlation_Coefficient
Hour	0.3802643
Temperature	0.5632
Humidity	-0.2670
Wind.speed	0.1103
Visibility	0.2243
Solar.Radiation	0.3478
Rainfall	-0.2516
Snowfall	-0.1783

4.2 Regression Model of Bike Count Based on All Variables.

The "Intercept" which is the estimated value of the response variable (log.Rented.Bike.Count) when all other predictors are zero has a significant positive effect on the response, with

an estimated coefficient of 6.2132 ($p < 0.0001$). The predictor variables "Hour," "Temperature," and "Holiday/No Holiday" has a significant positive effect on the response variable, whereas "Humidity", and "Rainfall" have a significant negative effect on the response variable. The predictor variables "Wind-speed," "Visibility," "Solar-Radiation," and "Snowfall" do not have a significant effect on the response as they have p-values
> 0.05. Similarly, the predictor variables related to different seasons ("SeasonsSpring," "SeasonsSummer," and "SeasonsWinter") have a significant negative effect on the response variable compared to the reference season(Autumn) (see Table 5).

Table 5:Summary of the Regression Model Output (full model)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	6.2132	0.1288	48.2070	0.0000 ***
Hour	0.0444	0.0022	19.9400	0.0000 ***
Temperature	0.0409	0.0025	15.8126	0.0000 ***
Humidity	-0.0180	0.0010	-16.7961	0.0000 ***
Wind.speed	-0.0285	0.0153	-1.8635	0.0624
Visibility	-0.0000	0.0000	-0.5952	0.5516
Solar.Radiation	-0.0247	0.0220	-1.1235	0.2612
Rainfall	-0.2259	0.0122	-18.4069	0.0000 ***
Snowfall	-0.0062	0.0314	-0.1995	0.8418
SeasonsSpring	-0.2734	0.0415	-6.5891	0.0000 ***
SeasonsSummer	-0.1764	0.0507	-3.4786	0.0005 ***
SeasonsWinter	-0.7835	0.0580	-13.4897	0.0000 ***
Holiday/No Holiday	0.3353	0.0635	5.2769	0.0000 ***

4.3 Subset Selection Based on Stepwise AIC

The stepwise method with the AIC criterion is used to calculate the best model for the prediction. Table 6 summarizes regression results, including parameter estimates, statistical significance (p-values), and confidence intervals. The intercept term represents the expected log.Rented.Bike.Count when all other predictor variables are zero. In this case, the estimated intercept is 6.1448. For a one-unit increase in Hour and Temper- ature, holding all other variables constant, the log.Rented.Bike.Count is expected to increase by 0.0449 and by 0.0400. Similarly, for a one-unit increase in Humidity, Wind- Speed, and Rainfall, holding all other variables constant, the log.Rented.Bike.Count is expected to decrease by 0.0173, 0.0334, and 0.2260. Compared to the reference category, being in the Spring or Summer or Winter season is associated with a decrease in the

log.Rented.Bike.Count by 0.2698, 0.1733, and 0.7843 respectively. Lastly, compared to holidays, being a non-holiday is associated with an increase in the log.Rented.Bike.Count by 0.3345. In the provided summary table, all coefficients have a p-value less than 0.05, indicating that they are all statistically significant (see Table 6).

Table 6:Summary of Stepwise Regression

	Estimate	Std.Error	t-value	2.5%CI	97.5%CI	Pr(>/t/)
(Intercept)	6.1448	0.0968	63.47	5.9549	6.3345	0.0000 ***
Hour	0.0449	0.0022	20.39	0.0405	0.0491	0.0000 ***
Temperature	0.0400	0.0024	16.76	0.0353	0.0446	0.0000 ***
Humidity	-0.0173	0.0008	-22.43	-0.0187	-0.0157	0.0000 ***
Wind.speed	-0.0334	0.0148	-2.26	-0.0624	-0.0044	0.0239 *
Rainfall	-0.2260	0.0122	-18.46	-0.2499	-0.2019	0.0000 ***
SeasonsSpring	-0.2698	0.0402	-6.70	-0.3487	-0.1908	0.0000 ***
SeasonsSummer	-0.1733	0.0504	-3.44	-0.2721	-0.0745	0.0006 ***
SeasonsWinter	-0.7843	0.0569	-13.78	-0.8959	-0.6726	0.0000 ***
Holiday/No Holiday	0.3345	0.0635	5.27	0.2099	0.4590	0.0000 ***

Table 7 in the Appendix on page 19 summarizes the goodness-of-fit measure. The residual standard error (0.7419) is the measure of the model's accuracy in predicting the log-transformed Rented Bike Count. A lower residual standard error indicates that the model fits the data better and the R-Square value is the goodness-of-fit and is a statistical measure of how well the data fit the regression line (James et al., 2013, p. 68-69). The selected predictors account for approximately 59.37% of the variability in the log-transformed Rented Bike Count, and the adjusted R-squared 59.25% is very close to the multiple R-squared, indicating that the model's performance is consistent. The F-statistic (470.1) evaluates the overall significance of the model by comparing the explained variation (regression sum of squares) to the unexplained variation (residual sum of squares). A higher F-statistic indicates a better fit of the model. From the outcome of this analysis, F-statistic is highly significant with a p-value less than 2.2e- 16, indicating that the overall model is statistically significant in explaining the log- transformed Rented Bike Count (see Table 7 in Appendix on page 19).

The deviance analysis, which shows the stepwise removal of variables, is summarised in Table 8 in the Appendix on page 19. The initial model includes all the variables and in step 2, the variable "Snowfall" is removed from the model and the model's deviance decreases by 0.0219 and the AIC decreases from -1720.057 to -1722.017, indicating a slight improvement in model fit. In step 3, "Visibility" is removed, the deviance and AIC

value also decrease. In Step 4, "Solar-Radiation" is removed from the model, the deviance decreases by 0.6001 and AIC decreases to -1724.578, indicating the most significant improvement in model fit (see Table 8 in Appendix on page 19).

4.4 Residual Plot for Model Evaluation

The final best model is used to generate residual plots to test the linearity, heteroskedas- ticity, and normality assumptions. The variance inflation factor (VIF) was used to fur- ther test for multicollinearity. The residual plot for model evaluation is shown in Figure 4 in the Appendix on page 18 and Table 9 in the Appendix on page 19. The residuals versus fitted plot shows there are some non-linear relationships between the indepen- dent and the dependent variable. The Scale-Location plot shows that residuals are not distributed evenly across the predictor which implies that it may not have constant vari- ance. The QQ plot also shows that not all of the residuals are on the dotted line; there is a slight deviation that may cause the normality test to fail which might be due to outliers, which Cook's distance plot also confirms (see Figure 4 in the Appendix on page 18). The model passes the multicollinearity VIF test, as all the GVIF values are less than 10, it can be assumed that there is no multicollinearity in the final Model (see Table 9 in the Appendix on page 19). Possible issues with the final model include: the model violating the normality and heteroskedasticity assumption which may affect the accurate prediction of bike demand; and the final model may be sensitive to outliers, which can have a disproportionate influence on the estimated coefficients and affect overall model performance.

5 Summary

The dataset used in this report is an extract of the Seoul Bike Sharing Demand dataset. The objective of this report is to examine the relationship between the dependent and independent variables, as well as to select a suitable subset of independent variables from the bicycle sharing rentals dataset for the best subset model. Secondly, to evaluate the model, create a residual plot and examine it for patterns of linearity, heteroskedasticity, normality, and test for multicollinearity.

There are so many factors that influence bike rental demand. From the result of the analysis, Seoul residents are more likely to rent a bike in the summer and autumn compared to the spring or winter season which may be due to cold during the winter period. They also use the bike more on non-holiday days than on holiday days. From the result of correlation analysis, Hour, Temperature, Windspeed, Visibility, and Solar-Radiation have positive correlations with the response/dependent variable while Humidity, Rain- fall, and Snowfall have negative correlations.

The result from the regression analysis also shows that variables Hour, Temperature, and Holiday/No Holiday have a significant positive impact on the log bike count while Humidity, Rainfall, Spring, Summer, and Winter have a significant negative impact on the log bike count when other variables are held constant. The variables Wind-speed, Visibility, Solar-Radiation, and Snowfall do not have a significant effect on the response variable as they have p-values > 0.05.

The stepwise method resulted in a final regression model that did not account for snow- fall, visibility, or solar radiation. The following variables are included in the final model: log Rented Bike Count Hour, Temperature, Humidity, Wind-speed, Rainfall, Seasons, and Holiday. The results of the stepwise regressions show that the selected predictors account for approximately 59.37% of the variability in the log-transformed Rented Bike Count, and the adjusted R-squared in this model is very close to the multiple R-squared, indicating that the model's performance is consistent. The final model explains 59.25% of the dataset and is statistically significant with a p-value less than 0.05. The residuals versus fitted plot show some non-linear relationships but the model passes the multi- collinearity VIF test.

The Scale-Location plot shows that residuals are not distributed evenly across the predictor, implying that the model does not have constant variance. The QQ plot also shows that the residuals are not normally distributed, as there is a slight deviation which could be due to outliers in the model's residuals, as confirmed by Cook's distance plot. The violation of these assumptions can have a disproportionate influence on the estimated coefficients and affect overall model performance. For accurate model performance, additional transformations, such as the box-cox transformation, can be applied to the model to deal with heteroskedasticity, deviation from normality, and to remove outliers for future research.

Bibliography

Ali, P. and Younas, A. (2021). Understanding and interpreting regression analysis. *Evidence-Based Nursing*, 24(4):116–118. https://ebn.bmj.com/content/24/4/116.

Fox, J. and Weisberg, S. (2019). *An R Companion to Applied Regression*. Thou- sand Oaks CA, third edition. https://socialsciences.mcmaster.ca/jfox/Books/ Companion/, (visited on 2nd July 2023).

Hebbali, A. (2020a). *Olsrr: Tools for Building OLS Regression Models*. https://cran. r-project.org/package=olsrr, (visited on 2nd July 2023).

Hebbali, A. (2020b). *Olsrr: Tools for Building OLS Regression Models (Collinear- ity Diagnostics)*. https://cran.r-project.org/web/packages/olsrr/vignettes/ regression_diagnostics.html, (visited on 2nd July 2023).

Hebbali, A. (2020c). *Olsrr: Tools for Building OLS Regression Models (Resid- ual Diagnostics)*. https://cran.r-project.org/web/packages/olsrr/vignettes/ residual_diagnostics.html, (visited on 2nd July 2023).

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An introduction to statis- tical learning with applications in R*. Springer, New York.

R Development Core Team (2023). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www. R-project.org/ (visited on 1st June2023).

Robinson, D. (2023). *Broom: Convert Statistical Objects into Tidy Tibbles*. https: //cran.r-project.org/package=broom, (visited on 2nd July 2023).

Scott, D. (2019). *xtable: Export Tables to LaTeX or HTML*. https://cran.r-project. org/package=xtable, (visited on 2nd July 2023).

Signorell, A. (2023). *DescTools: Tools for Descriptive Statistics*. https://CRAN. R-project.org/package=DescTools, (visited on 2nd July 2023).

UC Irvine ML Repository (2020). Seoul Bike Sharing Demand. https://archive.ics. uci.edu/dataset/560/seoul+bike+sharing+demand.

UVA Library StatLab (2015). Understanding diagnostic plots for linear regression anal- ysis. https://data.library.virginia.edu/diagnostic-plots/ (visited on 28th June 2023).

Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S*. Springer, New York, fourth edition. https://www.stats.ox.ac.uk/pub/MASS4/, (visited on 2nd July 2023).

William Revelle (2023). *Psych: Procedures for Psychological, Psychometric, and Per- sonality Research*. Northwestern University, Evanston, Illinois. https://CRAN. R-project.org/package=psych, (visited on 2nd July 2023).

Xie, Y. (2023). *knitr: A General-Purpose Package for Dynamic Report Generation in R*. https://yihui.org/knitr/, (visited on 2nd July 2023).

Appendix

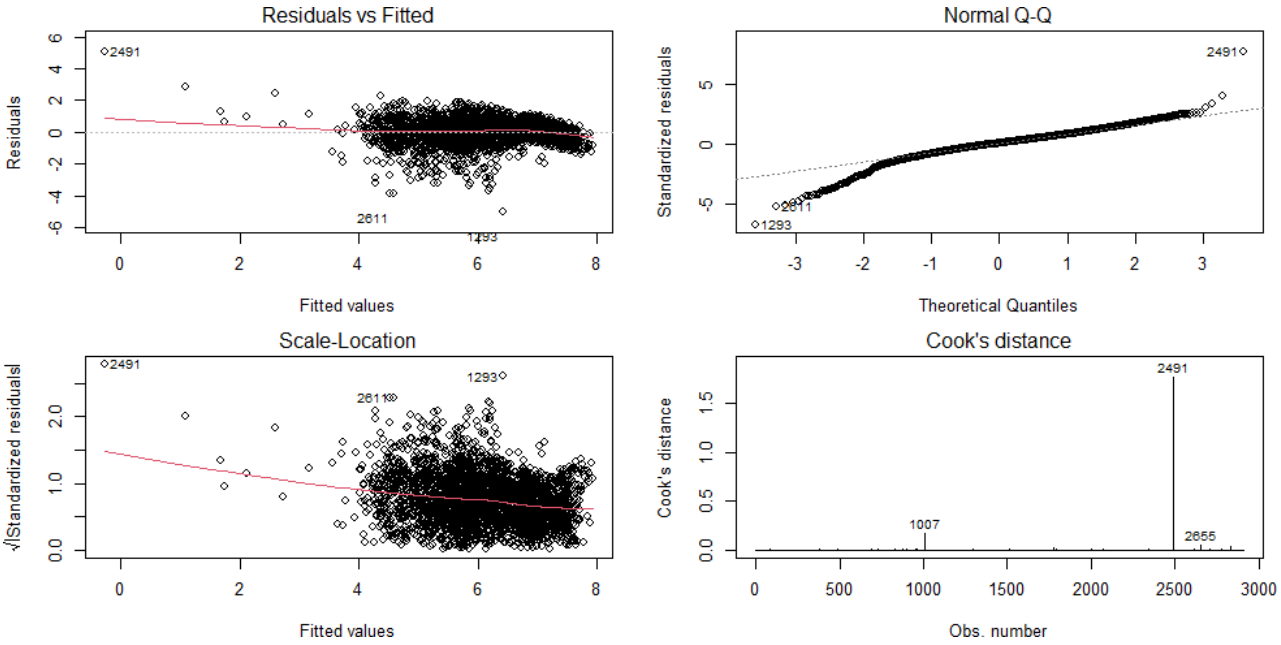

Figure 4:Residual Plot for Model Evaluation

Table 7:Summary of Goodness of Fit

Residual standard error	0.7419 on 2895 degrees of freedom
Multiple R-squared	0.5937
Adjusted R-squared	0.5925
F-statistic	470.1 on 9 and 2895 df
p-value	< 2.2e-16
logLik	-3249.72
AIC	6521.45
BIC	6587.17
deviance	1593.43
nobs	2905
Signif	0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Table 8:Analysis of Deviance Table

Step	Df	Deviance	Resid. Df	Resid. Dev	AIC
	NA	NA	2892	1592.621	-1720.057
- Snowfall	1	0.0219393	2893	1592.643	-1722.017
- Visibility	1	0.1888638	2894	1592.832	-1723.672
- Solar.Radiation	1	0.6001088	2895	1593.432	-1724.578

Table 9:Checking for Multicollinearity using VIF.

	GVIF	Df	GVIF^(1/(2*Df))
Hour	1.206633	1	1.098468
Temperature	4.484710	1	2.117713
Humidity	1.326116	1	1.151571
Wind.speed	1.231455	1	1.109709
Rainfall	1.062999	1	1.031019
Seasons	4.702495	3	1.294359
Holiday	1.029207	1	1.014498