ST1131 Introduction to Statistics and Statistical Computing

Individual Assignment

Tan Kin Ru A0277174W

Introduction

The Capital Bikeshare system in Washington D.C., USA, has provided invaluable data spanning the years 2011 and 2012.[1] Additionally, supplementary weather and seasonal information[2] have been integrated into the dataset to enrich the analysis. This dataset captures the daily count of rental bikes (*cnt*) along with various predictor variables such as *season*, *workingday*, *weathersit* (weather situation), *temp* (temperature), *hum* (humidity), and *windspeed*.[3]

The primary objective of this research is to identify the variables that significantly influence response variable - the daily count of rental bikes. Through statistical analysis, we aim to construct a linear regression model that effectively predicts the rental bike count based on the selected predictors.

Questions of Interest:

- Which variables, among *season*, *workingday*, *weathersit*, *temp*, *hum*, and *windspeed*, have a significant impact on the daily count of rental bikes?
- Can we develop a reliable linear regression model that accurately predicts the daily count of rental bikes based on the identified predictors?

Summary of Statistical Findings

The histogram and QQ-plot of the response variable *cnt* were plotted to check for normality. If not symmetrical, transformations may be considered to satisfy this assumption or else the model would not be adequate for any conclusions. From Figures 1 and 2, the histogram and QQ-plot analyses of *cnt* suggest normality assumption is violated as both figures suggest an under-dispersed dataset.

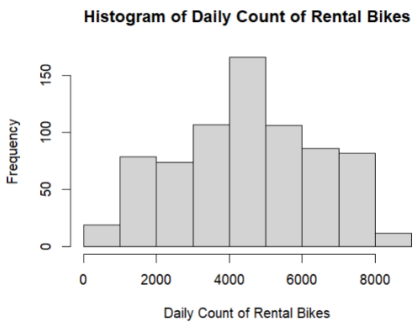

Figure 1 Histogram of Daily Count of Rental Bikes (cnt)

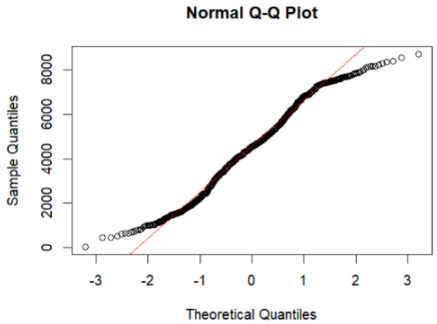

Figure 2 QQ-Plot of Daily Count of Rental Bikes (cnt)

[1] http://capitalbikeshare.com/system-data
[2] http://www.freemeteo.com
[3] season : season (1:springer, 2:summer, 3:fall, 4:winter)
workingday : if day is neither weekend nor holiday is 1, otherwise is 0.
weathersit : 1: Clear, Few clouds, Partly cloudy, Partly cloudy, 2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist, 3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds, 4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog
temp : Normalized temperature in Celsius. The values are divided to 41 (max)
hum: Normalized humidity. The values are divided to 100 (max)
windspeed: Normalized wind speed. The values are divided to 67 (max)
cnt: count of total rental bikes including both casual and registered

Several transformations have been made to the response variable e.g. taking ln(*cnt*), square-root of *cnt* and the reciprocal of *cnt*. Through analysis, the square-root of the response variable was chosen as it produced a more symmetrical and bell-shaped histogram and with more points on the QQ-line for the QQ-plot. However, there were some noticeable outliers after transformation which will be removed during model building.

Before building a model, boxplots were generated to visualize the relationship between the response variable and categorical predictors (*season*, *workingday*, *weathersit*). Differences in the median and range among the levels of categorical predictors suggest potential associations with the response variable. Scatterplots of the response variable against quantitative predictors (*temp*, *hum*, *windspeed*) were created to identify potential linear associations. Additionally, correlation coefficients were calculated to quantify the strength and direction of the associations. Significant correlation coefficients and clear patterns in scatterplots indicate potential predictors for the linear model. Through scatterplots, we also assess the variability of the response variable as the quantitative predictors change. If the variability appears unstable, it suggests a potential violation of the constant variance assumption, which will be further examined in subsequent analyses. To address this issue, transformations may be performed on the quantitative predictors.

From Figures 3 to 5, positive linear association with stable variability is observed between *cnt* and *temp* (correlation coefficient = 0.627), as well as potential associations between *cnt* with *season* and *weathersit*.

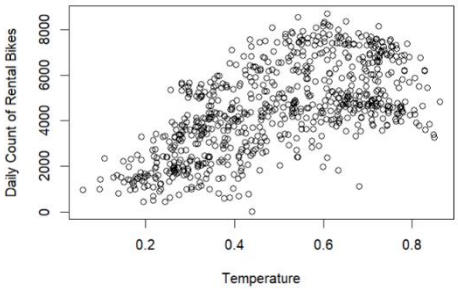

Figure 3 Scatterplot of Daily Count of Rental Bikes (cnt) against Temperature (temp)

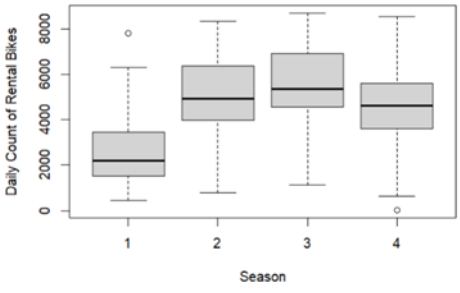

Figure 4 Boxplot of Daily Count of Rental Bikes (cnt) by Season (season)

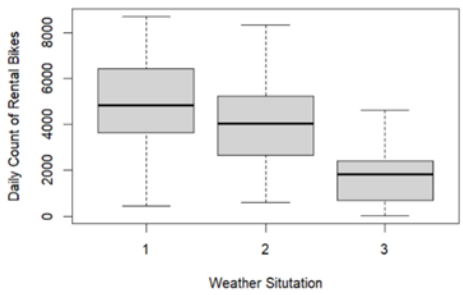

Figure 5 Boxplot of Daily Count of Rental Bikes (cnt) by Weather Situation (weathersit)

However, when *season* = "2", "3" or "4", the boxplots are similar to one another in terms of range and median as compared to when *season* = "1". Hence, we have decided to split *season* into two categories instead of four (*season* = "1" or "2-4"). Similarly, *weathersit* was split into two categories instead of three (*weathersit* = "1-2" or "3"). The boxplots after combination are shown in Figures 6 and 7.

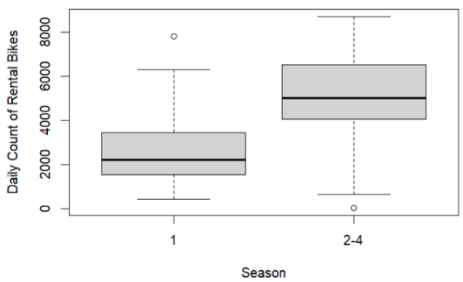

Figure 6 Boxplot of Daily Count of Rental Bikes (cnt) by Season (season)

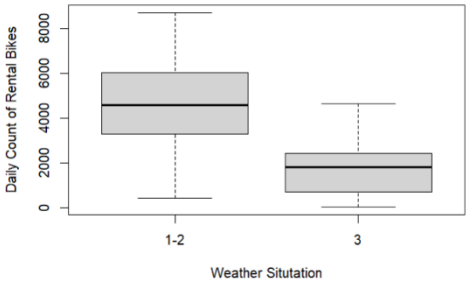

Figure 7 Boxplot of Daily Count of Rental Bikes (cnt) by Weather Situation (weathersit)

Initial Model - Model 1

For the initial model, we aimed to capture potential associations among variables and the response variable based on preliminary findings as these variables are most likely significant in the model. Recognizing that certain variables might also be correlated with each other, we constructed a model encompassing all interaction terms involving *temp*, *season*, or *weathersit* and all other variables mentioned in them. This resulted in all individual variables being included and an equation with 18 coefficients (excluding intercept), which though not included in the report due to its length, can be referenced in the accompanying R code.

Following model construction, we conducted tests to validate key assumptions, including normality and constant variance, which were satisfied as seen in Figures 8 to 11. Normality can be assumed as even though not all the Standard Residuals (SRs) in the QQ-plot lie on the QQ-line, however, there seems to be an equal number of SRs that are not on the line for both tails, which indicates symmetricity - both tails are shorter than normal. In addition, the SRs fall within the interval $(-3, 3)$ on the histogram and a bell-shaped and symmetrical histogram further supports the normality assumption. The scatterplot of SRs against the predicted values and the scatterplots of SRs against the quantitative regressors exhibit random dispersion of points and the points are within the interval $(-3, 3)$ and hence satisfy the constant variance assumption. Due to space constraints, only one graph of SRs against regressors is shown in the report. However, upon scrutinizing influential points, which were identified through Cook's distance exceeding a threshold of 1, and outlier points, which were characterized by SRs beyond ±3, one influential (69^{th} row) and two outlier points (87^{th} and 239^{th} row) were identified. To gauge their impact on the model, we proceeded to exclude these points from Model 1 and re-ran the analysis. The observed changes in adjusted R-squared values underscored the influence of these data points, affirming their removal from subsequent models (Adjusted R^2 value for Model 1: 0.617 and Model 2 without outlier and influential points: 0.623). We believe that these outliers and influential points may be due to wrong recording or a specific event happening that day which caused model to be influenced by it or not being able to account for them.

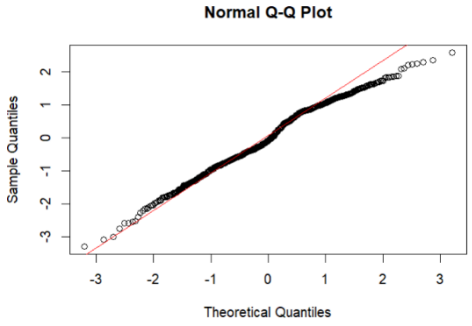

Figure 8 QQ-Plot of Standard Residuals for Model 18

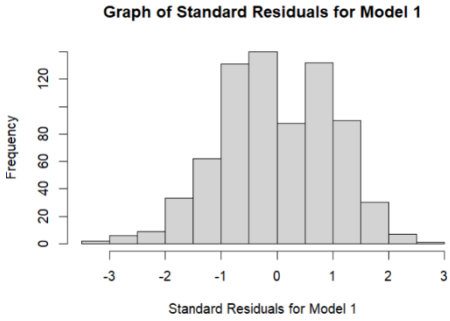

Figure 9 Histogram of Standard Residuals for Model 1

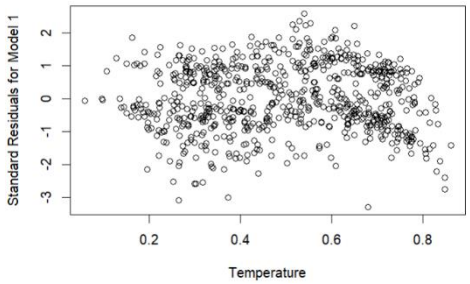

Figure 10 Scatterplot of Standard Residuals for Model 1 against Temperature (temp)

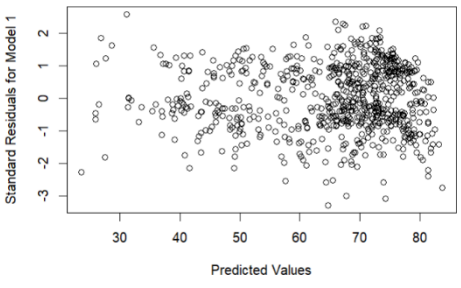

Figure 11 Scatterplot of Standard Residuals for Model 1 against Predicted Values

While the R-squared value for Model 1 may appear substantial (R^2 value = 0.626), it's crucial to recognize that its size is primarily a reflection of its complexity rather than its efficiency in explaining the variability of the data. Thus, it's imperative to explore additional models to ensure that our chosen model is not just overly complex but also effective in capturing the underlying relationships within the data.

Furthermore, upon reviewing the model summary, it became evident that certain terms lacked statistical significance. In response, we refined the model by retaining only those terms significant at the 0.05 significance level by looking at their p-value from their t-test from the model summary, which tests for the significance of one regressor or one coefficient, for the next few iterations. An interaction term is dropped only if all p-value of the interaction terms containing both variables are bigger than 0.05. A variable is dropped only if all of its interaction terms are dropped and its individual p-value from the t-test is not significant.

Model 2 – Model 5

Model 2 was an iteration of Model 1 but without the outlier and influential points. Model 3 and 4 presented herein as iterations where terms were gradually removed based on their p-values obtained from the t-test from the model summary with a significance level at 0.05 (Model 2 to Model 3) and at 0.01 (Model 3 to Model 4). Ultimately, *workingday* and *weathersit* were eliminated as none of its interaction terms proved significant, nor was it individually significant. Model 4 and Model 5 are iterations with the same coefficients but with different data points excluded in each model as through analyses of both models, there appears to be several outlier and influential points in both iterations. Hence, these points were removed to observe the changes to their adjusted R-squared value.

Final Model - Model 6

In the end, all outlier and influential points observed from Model 4 and Model 5 were removed as they significantly impacted the adjusted R-squared value of each model (Adjusted R^2 value for Model 4: 0.588 and Model 5: 0.593).

Hence, the final model – Model 6 - comprises regressors including *season*, *temp*, *hum* and *windspeed*, alongside the interaction term between *season* and *temp*. The regression equation is formulated accordingly:[4]

$$\sqrt{\hat{y}} = 47.619 + 34.073I(X_1 = 2) + 95.350X_2 - 31.888X_3 - 36.912X_4 - 66.276I(X_1 = 2) \times X_2$$

yielding an R-squared value of 0.599 and adjusted R-squared value of 0.596. The p-value from the F test substantially lower than 0.01, thus indicating the significance of the model. The Anova function analysis indicates the significance of the variables in the model, with all p-values substantially below 0.01. In addition, all confidence intervals for the coefficients do not include 0 and the p-value from their t-test is smaller than 0.01 hence showing significance to the model.

[4] $\hat{y}$: predicted value of daily count of rental bikes
$I(X_1 = 2)$: Indicator variable for season when season = 2, 3 or 4
X_2: temp
X_3: hum
X_4: windspeed

Following model construction, we conducted tests to validate key assumptions, including normality and constant variance, which were satisfied as seen from Figures 12 to 15. Normality can be assumed as even though not all SRs in the QQ-plot lie on the QQ-line, however, there seems to be an equal number of SRs that are not on the line for both tails which indicates symmetricity. In addition, the SRs fall within the interval $(-3, 3)$ on the histogram and a bell-shaped and symmetrical histogram further supports the normality assumption. The scatterplot of SRs against the predicted values and the scatterplots of SRs against the quantitative regressors exhibit random dispersion of points and the points are within the interval $(-3, 3)$ and hence satisfy the constant variance assumption. Due to space constraints, only one graph of standard residuals against regressors is shown in the report. There were no influential or outlier points.

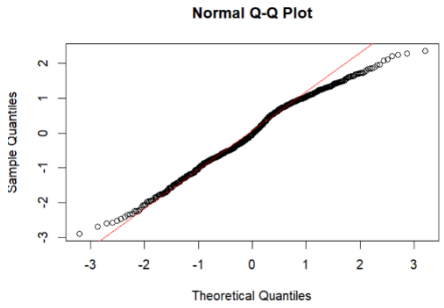

Figure 12 QQ-Plot of Standard Residuals for Model 6

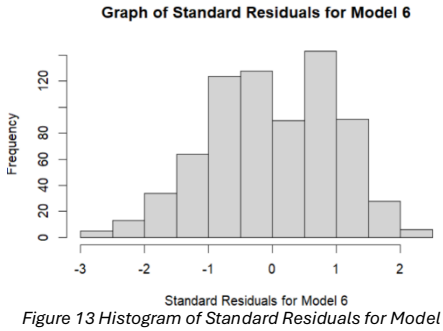

Figure 13 Histogram of Standard Residuals for Model 6

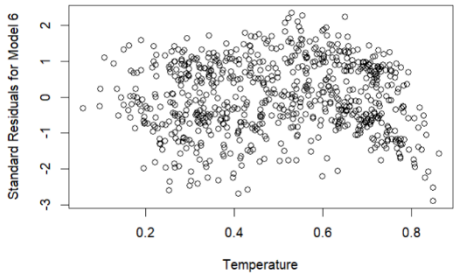

Figure 14 Scatterplot of Standard Residuals for Model 6 against Temperature (temp)

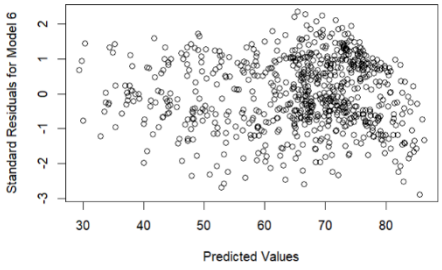

Figure 15 Scatterplot of Standard Residuals for Model 6 against Predicted Values

Summary of Statistical Findings

In summary, the iterative approach yielded a final regression model comprising key variables and interaction terms. With careful consideration of statistical significance and model fit, we arrived at a framework consisting of season, temperature, humidity and windspeed variables. From Table 1, even though the final model's adjusted R-squared value was not the largest, but it only contains variables and interaction terms that are most significant which is the aim as to make the model more meaningful instead of a complex model with insignificant terms in it.

Model	R-squared Value	Adjusted R-squared Value
Model 1	0.626	0.617
Model 2	0.633	0.623
Model 3	0.625	0.620
Model 4	0.591	0.588
Model 5	0.596	0.593
Model 6	0.599	0.596

Table 1 Table of R-squared Values and Adjusted R-squared Values of Models

From the R-squared value of the final model, 59.9% of the variation in the number of daily rental bikes is explained by this regression model. We believe that having a percentage around this range is expected as even with a complex model in Model 2, only 63.3% of the variation in the number of daily rental bikes can be explained by it. Hence, a possible reason for this low R-squared value may be because there may be other lurking variables affecting the response variable that we have yet to explore. Another possible reason for it may be that the model does not really fit a linear regression model to explain the relationship between *cnt* and the other variables.

This regression equation models the square root of the predicted daily count of rental bikes based on several predictor variables. The intercept term of 47.619 represents the expected square root of the daily count of rental bikes when all predictor variables are absent or zero. The effect of the indicator variable $I(X_1 = 2)$, which denotes *season* = 2, 3, or 4, is captured by the coefficient 34.073, indicating the change in the square root of the daily count of rental bikes when the season falls within this range compared to other seasons. Temperature, humidity and windspeed have effects of 95.350, -31.888 and 36.912 respectively, indicating the change in the square root of the daily count of rental bikes for each unit increase in temperature, humidity, and windspeed. Additionally, the interaction between the indicator variable of season when *season* = 2, 3, or 4 and temperature ($I(X_1 = 2) \times X_2$) is accounted by the term -66.276, indicating the combined effect of these variables on the square root of the daily count of rental bikes for each unit increase in temperature when both conditions are present. Hence, to obtain the predicted daily count of rental bikes, square both sides of the equation. This regression equation provides insights into how various factors, including seasonal variations, weather conditions, and their interactions, influence the predicted daily count of rental bikes.

Further study could explore the relationship between windspeed and the response variable as well as the relationship between humidity and the response variable as despite their initial lack of obvious association, they proved to be significant in the final model. In addition, as not all the points lie perfectly on the QQ-plot, hence making it slightly under-dispersed as both tails are still shorter than normal, a further study could include further exploring the transformation of the response variable to ensure normality. However, due to time constraints, we were not able to find the ideal transformation of the response variable.