STA 319 2.0 Advanced Regression Analysis
2022

Multiple Linear Regression

Project Report

Predication of Bike Rental Daily Count Based On the Environmental and Seasonal Settings

Department Of Statistics
Faculty Of Applied Sciences
University Of Sri Jayewardenepura

Kulugammana K.D.P

AS2019421

04 December 2022

Contents

Abstract

The problem of unused bikes being left at the docking stations as trash is addressed by forecasting bike rental based on weather and seasons. By using the information from one docking station, we will make a prediction about the total number of people who will register. In order to address the issue of abuse of the bikes present in the docking stations, we will collect the raw data from the geographic data, weather report, and bike-sharing data. Using all of this data, we will utilize the multiple linear regression approaches. We will successfully forecast the future daily counts for bike rentals using this regression technique. We will remove duplicate and unnecessary data from this model, clean the data and We will obtain a prediction line and an accurate score from these data points.

1 Introduction

1.1 Background of the study

Ridesharing businesses are excellent business models because they offer clients who wish to travel without the inconvenience of owning or maintaining a vehicle quick, easy, and economical transportation options. But with more cars on the road, carpooling is no longer practical, especially in densely populated places like central cities. Bike sharing is a fantastic idea since it gives people another short-distance transportation choice, allowing them to go without worrying about getting stuck in traffic and perhaps even take in some city views or work out.

Beginning in 2010, this study will look at the rental information for the Capital Bikeshare bike share system, which serves Washington, D.C., and the neighboring areas. These systems' data properties open up fascinating new areas for research. In contrast to other types of transportation, these systems explicitly record departure and arrival positions. This feature transforms the bike-sharing program into an imaginary sensor network that can track urban motion. As a result, it is anticipated that monitoring this data will enable the detection of the town's most important occurrences.

In order to estimate rentals based on the data and models I have available, I must first identify the determining factor that drives the demand for bike share rentals, which is tightly tied to environmental and seasonal conditions. R will be used for my data analysis and study.

1.2 Description of data

This Capital Bikeshare bike share rental data solely includes entries from Washington, D.C., and spans the two years from January 1, 2011, to December 31, 2012. The weather data for the appropriate day and time is also linked to the dataset. Complete data were split into a training set that only included entries from 2012 and a testing set that only included items from 2011. I will use the training set to look for unique properties of the predictor variable and the response variable when exploring and analyzing the data.

The variable description is as follows:

1. instant: record index
2. dteday: date
3. season: season (1: spring, 2: summer, 3: fall, 4: winter)
4. yr: year (0: 2011, 1: 2012)
5. mnth: month (1 to 12)
6. holiday: weather day is holiday or not
7. weekday: day of the week
8. workingday: if day is neither weekend nor holiday - 1, otherwise - 0.
9. weathersit: (weather situation)

    - Clear, Few clouds, Partly cloudy
    - Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist
    - Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds
    - Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog

10. temp: Normalized temperature in Celsius. The values are divided to 41 (max)

11. atemp: Normalized feeling temperature in Celsius. The values are divided to 50 (max)

12. hum: Normalized humidity. The values are divided to 100 (max)

13. windspeed: Normalized wind speed. The values are divided to 67 (max)

14. cnt: Count of total rental bike

the first six rows of data set is as follows

```
# A tibble: 6 x 14
  instant dteday              season    yr  mnth holiday weekday workingday
    weathe~1  temp atemp
    <dbl> <dttm>              <fct>  <dbl> <dbl>   <dbl>   <dbl> <fct>      <
    fct>    <dbl> <dbl>
1       1 2011-01-01 00:00:00 1          0     1       0       6 0          2
           0.344 0.364
2       2 2011-01-02 00:00:00 1          0     1       0       0 0          2
           0.363 0.354
3       3 2011-01-03 00:00:00 1          0     1       0       1 1          1
           0.196 0.189
4       4 2011-01-04 00:00:00 1          0     1       0       2 1          1
           0.2   0.212
5       5 2011-01-05 00:00:00 1          0     1       0       3 1          1
           0.227 0.229
6       6 2011-01-06 00:00:00 1          0     1       0       4 1          1
           0.204 0.233
# ... with 3 more variables: hum <dbl>, windspeed <dbl>, cnt <dbl>, and
    abbreviated variable
#   name 1: weathersit
# i Use ‘colnames()‘ to see all variable names
```

Figure 1: First six rows of data set

1.3 Objectives

1. Predication of bike rental count daily based on the environmental and seasonal settings.

2. Event and Anomaly Detection

2 Methodology

2.1 Model Selection

We have used R programming running on the Rstudio platform for all analysis purposes and to obtain results.

The backward elimination search strategy is employed for the model selection. After filtering variables using exploratory data analysis, it starts with the Model including all potential X variables. First, there are the category variables Season, Working Day, Whether situation and the quantitative variables Temperature and Wind speed. Using type III Extra Sum of Squares, we determined which P-value in the first model was the most significant. None of the variable is eliminated if there P-values are less than a predefined threshold (0.15). Then the Model with the all X variables is fitted, and the next candidate for dropping is not found since all of the p values are small to be significant to carry out the procedure to reject the null hypothesis. The model then goes on to consider temperature, wind speed ,season,working day and the weather situation.

2.2 Model Adequacy

The Shapiro-Wilk normality test is used for the given model to verify the normality assumption. It demonstrated that the errors are not distributed regularly. Furthermore, we used a Q-Q plot and an error histogram to visually depict the non-normality behavior. There was a clear pattern in the residuals vs. fitted plot, which suggests that the error variance is not constant.

2.3 Event and Anomaly Detection

Outliers and significant moments were examined. The standardized residuals vs observation chart identifies outliers in the y-direction. We would rather not get rid of them right away. The Cooks distance also keeps an eye on crucial places. There are numerous situations that have significant influence. We frequently use Google to check for the causes and other events in the Washington town for each case based on the documented date.

2.4 Modification of the model

In order to create a model that predicts bike rental counts as the next step, we eliminated the influential points. All of the variables we previously included stayed the same when we re-fit the model using the same backward elimination method. The calculated coefficients, however, were altered.

Then, we use Adjusted R-squared to determine how well the new model fits the data compared to the old model. It showed that the new model has greater coverage of variability than the old model.

Next, the model's assumptions were examined. The errors are not regularly distributed, according to the Shapiro-Wilk and normal Q-Q tests. The Studentized Breusch-Pagan test demonstrates that the error variance is heteroscedastic. Additionally, the variance inflation factor is used to assess for and identify multicollinearity (V.I.F.).Results of Breusch-Godfrey indicated Test autocorrelation exists among the residuals also.

2.5 Model Validation and Model Building

However, with so many issues, the model was validated using R.M.S.E. , M.A.E and R^2 by the testing data set of 2011. It has proved that the model without influential cases can better predict the bike and daily rental counts.

3 Data Exploration

This chapter primarily focuses on analyzing and characterizing the structure of the data that is already accessible, finding trends, and detecting anomalies that can be helpful for creating a fresh approach to data experimentation and analysis. The total number of bicycle rentals is our response variable, and other variables were used as predictors in the initial analysis. The 2012 data subset is used for this data exploration because the model design is based on 2012.

To begin with, the distributions and relationships of the variables are examined in order to streamline the complex scenario caused by the large number of variables. Record index and recording date are initially disregarded because they cannot be regarded as predictors. They serve only as database guidance variables.

Since there are several predictors, correlation among them is checked to drop variables statistically.

The given data set was cleaned by identifying missing data and eliminating superfluous predictor variables as instant, date, day of the week, year, and month. Working day provided information on both weekends and holidays, and month was eliminated because it had 12 levels and confused the model.

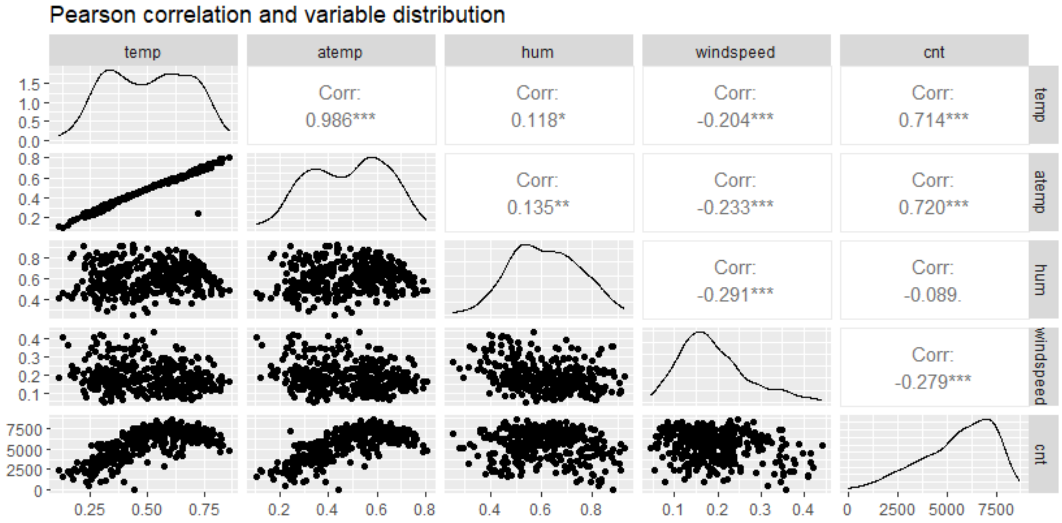

Figure 2: Pearson correlation and variable distribution

The correlation matrix demonstrates that the temp and atemp have an extremely high association. That is why the model only considers one variable out of temp and atemp.

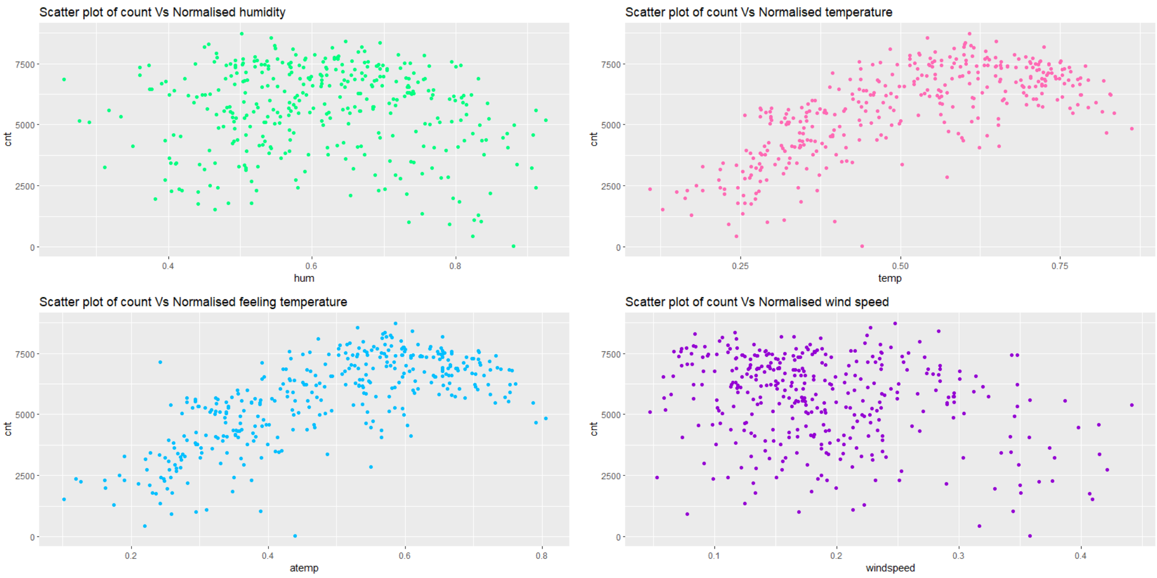

Figure 3: Scatter plots of quantitative pairs

It can be seen as positive linear relationships in both temp and atemp. So those variables should be included in the model, but because of the high correlation between them as mentioned temp was taken to build the model.

It can be seen a negative relationship between cnt and windspeed. So windspeed should also be included in the model.

Figure 03 shows that the humidity variable may be eliminated from the model we are developing because it has no effect on our response variable (cnt). The similar humidity range for all cnt values can plainly be noticed. That is what led us to make the conclusion we did.

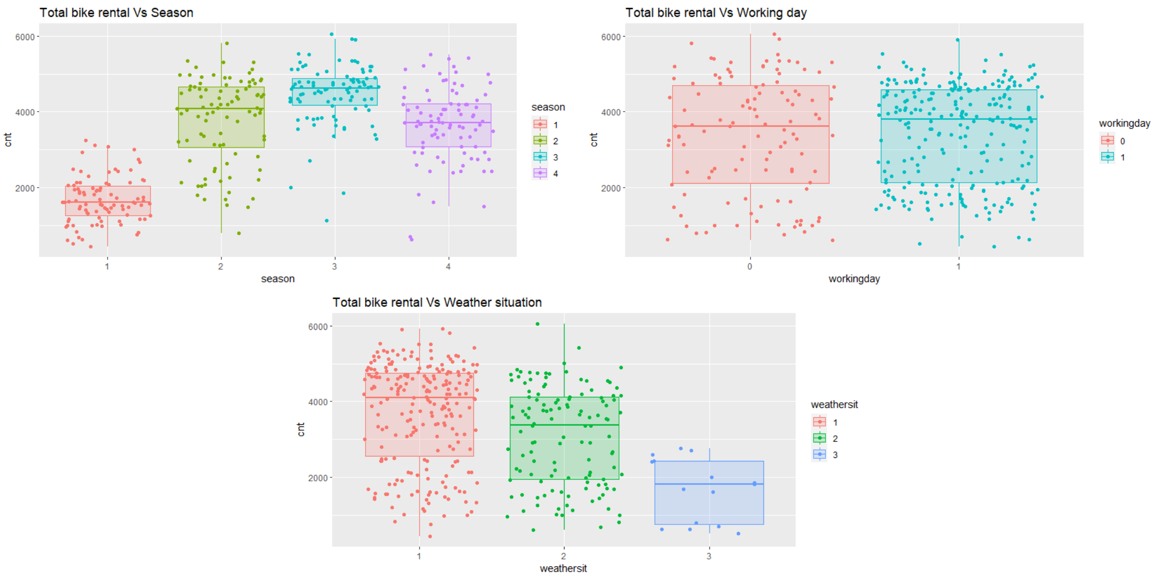

Figure 4: Box plots of qualitative pairs

While the other two categorical variables did not exhibit any outlier situations, the Fall and Winter values of the season variable did.

The fall season has the highest bike rental rates, while spring has the lowest rates. It demonstrates a high bike rental trend compared to the other two weather situations when the weather is calm (weather scenario 1: Clear+Few Clouds+Partly cloudy). Here, the fourth weather condition is disregarded because it did not occur in any of the dataset's examples (from 2011 or 2012). (Heavy rain, thunder, mist, snow, and fog)

As our response variable total bike rental count is also not normally distributed. Log transformation has led to a severe normality problem. So we take pure complete counts as it is without any modification.

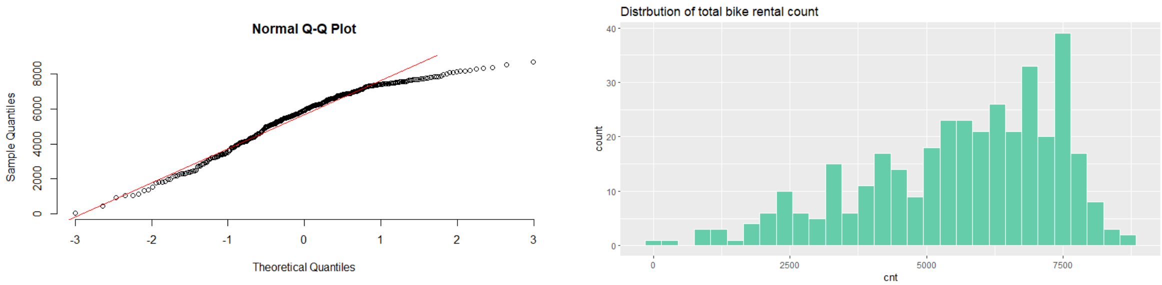

Figure 5: Distribution and the Normal Q-Q plot of Total bike rental counts

So at the end of the data exploration, initial candidate variables to the fresh Model are Temperature, Windspeed, Season, Working day and finally, weather situation.

4 Data Analysis and Results

4.1 Model Selection

As a starting point, a first-order regression model based on all predictor variables is fitted. In based on the above mentioned data and results, my first Model(model 1) can be expressed as

$$\begin{aligned} counts\ of\ total\ rental\ bikes = \beta_0 + \beta_1 Temperature + \beta_2 Wind\ speed + \beta_3 season + \beta_4 Working\ day \\ + \beta_5 Weather\ situation + \epsilon_i \end{aligned}$$

By Backward Elimination method, we begin with the Model containing all potential predictors and identify predictors of least significance considering α to remove (α_R) as 0.15.

In this case we have three categorical variables,namely ,

1. season: season(1: spring (Reference level), 2: summer, 3: fall, 4: winter)
2. working day(workingday): if day is neither weekend nor holiday - 1, otherwise - 0 (Reference level)
3. weather situation(weathersit):
    - 1: Clear, Few clouds, Partly cloudy (Reference level)
    - 2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist
    - 3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds

```
  Call:
lm(formula = cnt ~ temp + windspeed + season + workingday + weathersit,
    data = training_set)

Residuals:
    Min      1Q  Median      3Q     Max
-3688.5  -433.7    70.1   537.6  3520.6

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2364.1      277.5   8.521 4.52e-16 ***
temp          5869.2      537.9  10.912  < 2e-16 ***
windspeed    -2484.3      690.2  -3.599 0.000364 ***
season2       1197.8      190.0   6.303 8.61e-10 ***
season3        742.4      252.1   2.945 0.003446 **
season4       1610.1      155.8  10.335  < 2e-16 ***
workingday1    263.5      110.1   2.393 0.017234 *
weathersit2   -791.5      110.6  -7.160 4.64e-12 ***
weathersit3  -2992.7      407.7  -7.341 1.45e-12 ***
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .    0.1
        1

Residual standard error: 975.1 on 357 degrees of freedom
Multiple R-squared:  0.7093,  Adjusted R-squared:  0.7028
F-statistic: 108.9 on 8 and 357 DF,  p-value: < 2.2e-16
```

Figure 6: Summary of model 1

Model 1's adjusted R-squared value is 0.7028, which indicates that it adequately for 70.28 % of the variability.

Regression analysis use the technique of backward elimination to pick a subset of explanatory variables for the model. Backward elimination is used in the model to incorporate the beginning and all explanatory variables. The variable with the p-value which grater than α to remove is next taken out of the model.

All p-values for all levels in each variable are less than 0.15.Therefore none of the variables is going to be dropped.

model 1 without influential cases,

$$\begin{aligned} counts\ of\ total\ rental\ bikes = \beta_0 + \beta_1 Temperature + \beta_2 Wind\ speed + \beta_3 season + \beta_4 Working\ day \\ + \beta_5 Weather\ situation + \epsilon_i \end{aligned}$$

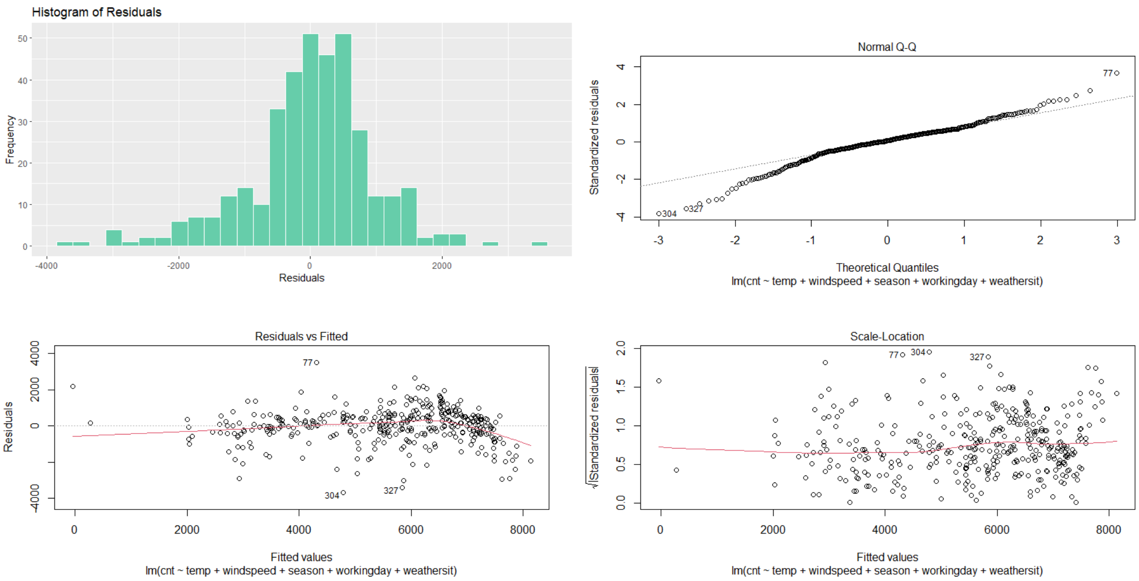

Figure 7: Plots of residuals for model 1

The Fitted vs Residuals plot shows that the error variance is not constant because it clearly shows a pattern.

```

    Shapiro-Wilk normality test

data:  model1_aug$.resid
W = 0.96479, p-value = 1.03e-07
```

Figure 8: Shapiro-Wilk normality test for model 1

With the Shapiro-Wilk normality test the normality assumption of the error term is not verified since p-value less than 0.05 (0.96479, p-value = 1.03e-07). A similar situation has shown by the histogram and the Normal Q-Q plot above.

```
                GVIF Df GVIF^(1/(2*Df))
temp        3.444909  1        1.856047
windspeed   1.118267  1        1.057481
season      3.493812  3        1.231828
workingday  1.010565  1        1.005269
weathersit  1.066348  2        1.016190
```

Figure 9: VIF values for model 1

A general guideline is that a V.I.F. larger than 5 or 10 is large, indicating that the Model has problems estimating the coefficient. However, this, in general, does not degrade the quality of predictions.

If the V.I.F. is larger than 1/(1-R^2), where R^2 is the Multiple R-squared of the regression, then that predictor is more related to the other predictors than the response. But in our Model Multiple R-squared =0 .7093. Thus, 1/(1-R^2) = 3.43 In our Model, VIF values of temp and season are grater than 3.43 .This indicates the multicollinearity of the variables.

```
  studentized Breusch-Pagan test

data:  model1
BP = 31.592, df = 8, p-value = 0.0001102
```

Figure 10: studentized Breusch-Pagan test for model 1

```
Breusch-Godfrey test for serial correlation of order up to 1

data:  model1
LM test = 75.667, df = 1, p-value < 2.2e-16
```

Figure 11: Breusch-Godfrey test for model 1

Constant error variance assumption is not satisfied.It can be confirmed by BP test (figure 10). It has a small p-value. Error variance is not constant hence its heterscedastic.

Since p-value of BG test is less than 0.05, we can reject the null hypothesis and conclude that autocorrelation exists among the residuals at some order less than or equal to 1.

4.2 Event and Anomaly Detection

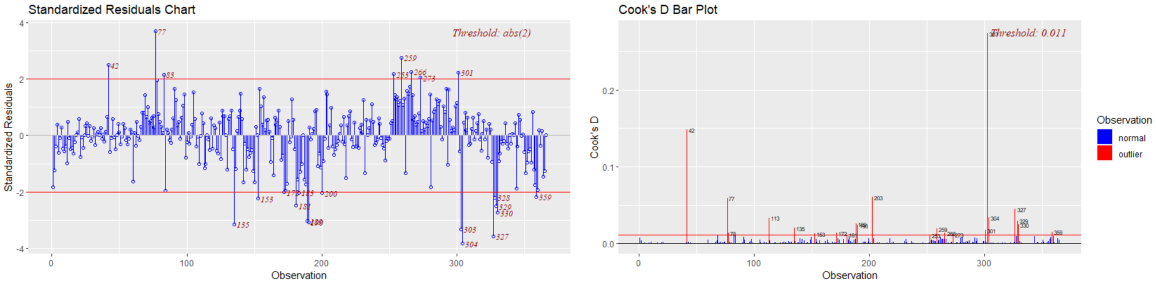

Figure 12: Plots for identifying outliers

Under this situation, we often focus on outliers and significant points. The standardized residuals vs observation chart identifies outliers in the y-direction.23 points were detected.

In respect to the sample size, there aren't many outliers, and deleting them could cause non-normality. They are therefore left alone. The Cook's distance is used to observe influential cases.

Here we use the threshold of 0.011 to exclude influential cases from the Model.

According to the Cook's distance bar plot 22 influential observations were detected. Some of the detected outline observation numbers are 47,135,113,203,253,209.

4.3 Fitting the Model without Outliers

model 2:Without influential cases

$$counts\ of\ total\ rental\ bikes = \beta_0 + \beta_1 Temperature + \beta_2 Wind\ speed + \beta_3 season + \beta_4 Weather\ situation + \epsilon_i$$

```
 Call:
lm(formula = cnt ~ temp + windspeed + season + workingday + weathersit,
    data = .)

Residuals:
     Min       1Q   Median       3Q      Max
-2312.18  -394.82    31.37   489.60  1978.34

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2310.40     226.30  10.209  < 2e-16 ***
temp         5670.06     444.97  12.743  < 2e-16 ***
windspeed   -2570.46     562.58  -4.569 6.90e-06 ***
season2      1396.43     154.91   9.014  < 2e-16 ***
season3       870.06     208.59   4.171 3.86e-05 ***
season4      1819.52     128.14  14.199  < 2e-16 ***
workingday1   384.28      91.07   4.220 3.15e-05 ***
weathersit2  -819.99      89.07  -9.206  < 2e-16 ***
weathersit3 -2888.21     546.58  -5.284 2.28e-07 ***
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .    0.1
       1

Residual standard error: 765.4 on 335 degrees of freedom
Multiple R-squared:  0.7982,  Adjusted R-squared:  0.7934
F-statistic: 165.7 on 8 and 335 DF,  p-value: < 2.2e-16
```

Figure 13: Summary of model 2

4.4 Goodness of the Model

```
# A tibble: 1     12
  r.squared adj.r.squared sigma  statis    p.value    df logLik   AIC   BIC
    devia   df.re     nobs
      <dbl>         <dbl> <dbl>    <dbl>    <dbl> <dbl>  <dbl> <dbl> <dbl>   <
   dbl>   <int> <int>
1     0.709         0.703  975.     109.  5.78e-91     8 -3034.  6088.  6127.
   3.39e8     357   366
#     with abbreviated variable names     statistic    ,     deviance    ,    df    .
   residual
```

Figure 14: Measuring the strength of the fit model 1

```
# A tibble: 1    12
  r.squared adj.r.squared sigma  s t a t i     p.value    df logLik   AIC   BIC
    d e v i a   df. r e     nobs
      <dbl>         <dbl> <dbl>   <dbl>      <dbl> <dbl>  <dbl> <dbl> <dbl>   <
    dbl>   <int> <int>
1     0.798         0.793  765.    166.  1.50e-111     8 -2768. 5556. 5594.
    1.96e8     335   344
#     with abbreviated variable names    statistic   ,    deviance   ,   df   .
    residual
```

Figure 15: Measuring the strength of the fit model 2

We can clearly see that the value of Adjusted R-square has increased from 0.703 to 0.793 (79.3%). Therefore, approximately 79% of the variation of counts of total rental bikes is explained by the predictor variables in the new Model without influential cases.

```
Analysis of Variance Table

Response: cnt
            Df    Sum Sq   Mean Sq F value    Pr(>F)    
temp         1 549228204 549228204 937.407 < 2.2e-16 ***
windspeed    1  18429946  18429946  31.456 4.274e-08 ***
season       3 137664233  45888078  78.320 < 2.2e-16 ***
workingday   1   8382238   8382238  14.307 0.0001839 ***
weathersit   2  62874266  31437133  53.656 < 2.2e-16 ***
Residuals  335 196276975    585901                      
---
Signif. codes:  0   ***   0.001   **   0.01   *   0.05   .   0.1   
        1
```

Figure 16: ANOVA for the model 2

All the variables have a significant effect on cnt at 0.05 significance level

4.5 Model Assumptions

- Random error ϵ normally distributed with mean zero and constant error variance $\mathcal{N}(0, \sigma^2)$
- No autocorrelation

```
  Shapiro-Wilk normality test

data:  model2_aug$.resid
W = 0.98627, p-value = 0.002378
```

Figure 17: Shapiro-Wilk normality test for model 2

Shapiro-Wilk normality test (W = 0.98627, p-value = 0.002378 less than 0.05) mathematically proved that the errors are not normally distributed in the Model without influential cases at a 5% significance level. A similar situation has shown the Normal Q-Q plot below.

In four different methods, the diagnostic plots display residuals:

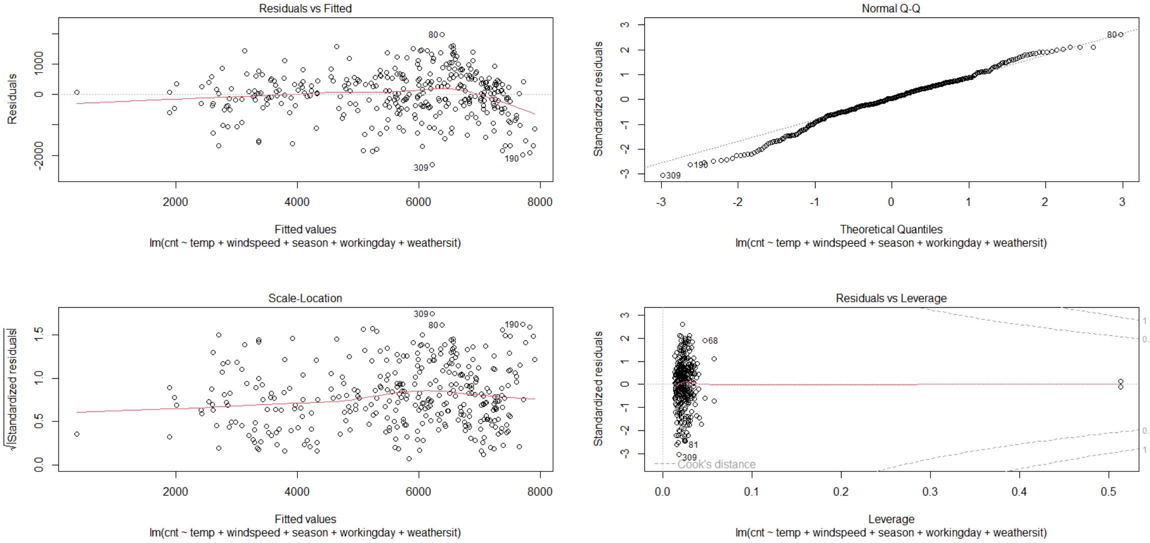

Figure 18: Plots of residuals for model 2

```
    studentized Breusch-Pagan test

data:  model2
BP = 24.309, df = 8, p-value = 0.002034
```

Figure 19: Studentized Breusch-Pagan test for model 2

Use of Scale-Location (or Spread-Location) to examine the homogeneity of the residual's variance (homoscedasticity). Homoscedasticity is best demonstrated by a horizontal line with evenly spaced points, which is not the case in this instance.

It can be seen that the constant error variance assumption was not satisfied. Therefore constant error variance assumption is not satisfied.It can be further confirmed by BP test (figure 18). It has a small p-value. Error variance is not constant hence its heterscedastic.

Here, residuals vs. fitted is utilized to test the assumptions of a linear relationship. A linear relationship is good if it roughly resembles a horizontal line without any obvious patterns.

```
              GVIF Df GVIF^(1/(2*Df))
temp       3.539670  1        1.881401
windspeed  1.112097  1        1.054560
season     3.660669  3        1.241443
workingday 1.009836  1        1.004906
weathersit 1.048723  2        1.011964
```

Figure 20: VIF values for model 2

A predictor is more closely related to the other predictors than the response if the V.I.F. is greater than 1/(1-R^2), where R^2 is the Multiple R-squared of the regression. However, Multiple R-squared = 0.7982 in this model. Thus, 1/(1-R^2) = 4.95.This demonstrates the none of the variable shows multicollinearity.

```
  Breusch-Godfrey test for serial correlation of order up to 1

data:  model2
LM test = 36.979, df = 1, p-value = 1.194e-09
```

Figure 21: Breusch-Godfrey test for model 2

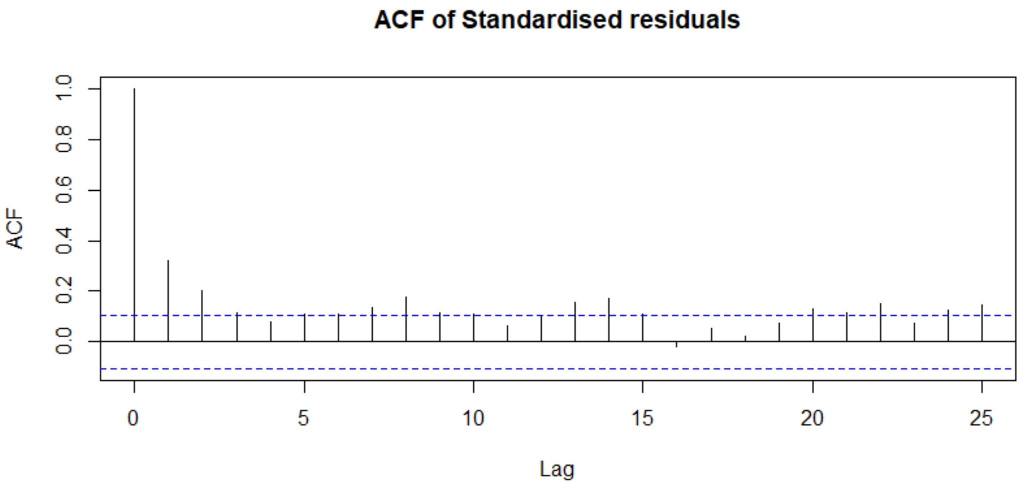

Figure 22: ACF of standardised residuals of model 2

Acf of standardized residuals(figure 22) shows some significant lags since the beginning. There are four significant lags, and we can see a clear cut off at lag 3.

The output shows that the test statistic is $X^2 = 36.979$ with 1 degrees of freedom. The corresponding p-value is 1.194e-09. Since this p-value is less than 0.05, we can reject the null hypothesis and conclude that autocorrelation exists among the residuals at some order less than or equal to 1.

4.6 Validating the Model

The validation set method randomly divides the data into two sets, one for training the model and the other for testing it. Here, the model is validated using data from 2011. The preferred model is the one that, when two models are compared, yields the lowest test sample R.M.S.E. and highest test sample R^2.

On the same scale as the outcome variable, the R.M.S.E. and the M.A.E. are measured. The prediction error rate, which we can calculate by dividing the R.M.S.E. by the average value of the outcome variable, should be as low as possible.

model	RMSE	MAE	R^2
Model with influential cases (Model 1)	2146.654	2015.373	0.7093
Model without influential cases (Model 2)	2194.417	2058.925	0.7982

Table 1: RMSE & MAE values for model 1 & model 2

Therefore, the Model with Influential Cases provides the lowest RMSE and MAE in this situation (Model 1).

In this instance, our model 2 generates poor predictions (high RMSE), but they are consistently incorrect in that their bias is approximately constant (high R^2). Since the predictor has a significant influence on the observed value, there is still some hope despite the poor predictions. Therefore, we select model 1 for the model-building process, which do not eliminate the influential cases.

4.7 Model Building and Interpretations

By using Model 2 without influential cases (Model 2)

Regression equation as follows,

$$\begin{aligned} counts\,of\,total\,rental\,bikes = 2310.40 + 5670.06\,Temperature - 2570.46\,Wind\,speed + \\ 1396.43\,season2 + 870.06\,season3 + 1819.52\,season4 \\ +384.28\,working\,day1 - 819.99\,Weather\,situation2 \\ -2888.21\,Weather\,situation3 + \epsilon_i \end{aligned}$$

Here,

- Weathersit2 weather 1 - Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist
- Weathersit3 - Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds

And the reference level is considered a day with Clear+Few Clouds+Partly Cloudly in the spring season of weekend or holiday.

The average total number of rental bikes in the Summer season is 1396 times higher and in the Fall season is 870 times higher than that of the Spring season. The Winter season remarks the best average count of total rental bikes, which is 1820 times higher than the Spring Season.

The average number of rental bikes concerning a day of Clear+Few Clouds+Partly Cloudy, a weather condition like Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist has 820 times lower average count of rental bikes. And also, it's more severe in weather conditions like Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds which drops the mean total count 2888 times that in the day like Clear+Few Clouds+Partly Cloudy.

The average total number of rental bikes in the working day is 384 times higher than that of the holiday or weekend.

When all other quantitative variables are kept constant, and the reference of categorical levels is considered a day with weathersit1 and workingday0 in the spring season, the average total bike rentals is decreased nearly by 2570 when we increased windspeed by one unit.

When all other quantitative variables are kept constant, and the reference of categorical levels is considered a holiday or weekend with Clear+Few Clouds+Partly Cloudly(weathersit1) in the spring season, the average total bike rentals is increased nearly by 5670 when we increased temperature by one Celsius.

When all the quantitative variables are kept constant, and the reference of categorical levels is considered a holiday or weekend with Clear+Few Clouds+Partly Cloudly(weather -1) in the spring season, the average total bike rentals is nearly 2310.

5 Conclusion and General Discussion

5.1 Conclusion

- 22 influential points (outliers with respect to x-axis) were observed by cook's D-plot
- 23 outliers were there with respect to y-axis

Model with influential cases,

$$\begin{aligned} counts\,of\,total\,rental\,bikes = 2364.1 + 5869.2\,Temperature - 2484.3\,Wind\,speed \\ +1197.8\,season2 + 742.4\,season3 + 1610.1\,season4 + 263.5\,working\,day1 \\ -791.5\,Weather\,situation2 - 2992.7\,Weather\,situation3 + \epsilon_i \end{aligned}$$

Model without influential cases,

$$\begin{aligned} counts\,of\,total\,rental\,bikes = 2310.40 + 5670.06\,Temperature - 2570.46\,Wind\,speed + \\ 1396.43\,season2 + 870.06\,season3 + 1819.52\,season4 \\ +384.28\,working\,day1 - 819.99\,Weather\,situation2 \\ -2888.21\,Weather\,situation3 + \epsilon_i \end{aligned}$$

5.2 Limitation of the study and Recommendations

We instantly came to the conclusion that the 2012 rental bike total isn't normally distributed. It is not effective to perform log transformation. Here, 365 data points, or 50% of the sample, makes up the data set used to develop the model.

When the influence of one independent variable on the dependent variable varies depending on the value(s) of one or more other independent variables, this is known as an interaction effect in regression. When creating the model, all two-way and three-way interactions between independent variables are not considered. The interaction terms are probably required if they are considered and are statistically significant. In this case, it is advised to look for interactions and then logically and statistically include them into the model.

Alpha-to-remove is taken into consideration in the backward elimination process as 0.15. By not altering this alpha level, no one may obtain a different model in this case.

Model 1's influential points were eliminated and Model 2 hasn't confirmed the error term's normality assumptions. however the errors, did not conform to the constant error variance. Errors are heteroscedastic as a result. Therefore, measures should be implemented to prevent heteroscedasticity and non normality when developing the Model.

Therefore, it is advised to apply weighted least squares (WLS) when the transformed Model is estimated using the OLS method if the error variance is known. In cases where it is unknown, heteroscedasticity is then eliminated by examining a connection between the error variance and one of the explanatory factors. To reduce the variance of the estimators, it should be noted that White estimators are also advantageous when compared to OLS estimates.

The autocorrelation among the error components caused another significant issue to arise in this case. Investigating the deletion of a crucial predictor variable should be one of the first remedial measures taken when autocorrelated error terms are discovered to be present. This process is highly individualized. Certain changes on the variables can be carried out if such a predictor is unable to assist in lowering or eliminating the autocorrelation of the error terms.