2022 | By: Jinushki Saluwadana (AS2018317)

BIKE SHARING DEMAD

ASP 459 2.0 Advanced Regression Analysis

Bike Sharing Demand

Linear Regression Analysis

Contents

Introduction

Background of the problem

Bike sharing systems are new generation of traditional bike rentals where whole process from membership, rental and return back has become automatic. Through these systems, user is able to easily rent a bike from a particular position and return back at another position. Currently, there are about over 500 bike-sharing programs around the world which is composed of over 500 thousands bicycles. Today, there exists great interest in these systems due to their important role in traffic, environmental and health issues.

Apart from interesting real-world applications of bike sharing systems, the characteristics of data being generated by these systems make them attractive for the research. Opposed to other transport services such as bus or subway, the duration of travel, departure and arrival position is explicitly recorded in these systems. This feature turns bike sharing system into a virtual sensor network that can be used for sensing mobility in the city. Hence, it is expected that most of important events in the city could be detected via monitoring these data.

The bicycle sharing process is closely linked to environmental and seasonal settings. For example, weather conditions, rainfall, day of the week, time and other factors can influence rent behavior. A researcher is interested to know how these environmental factors affect the number of bicycle rentals per day. In addition to that results can be used to make future predictions about daily bicycle rental used.

Description of Data

A total number of 731 observations were considered for the study under 14 variables consisting of both quantitative and qualitative types. Given below are those variables;

- **instant** : record index
- **dteday** : date
- **season** : season
  1: spring 2: summer 3: fall 4: winter
- **yr** : year
  0: 2011 1: 2012

- **mnth** : month
  1 : January to 12 : December

- **holiday** : whether day is holiday or not (extracted from http://dchr.dc.gov/page/holiday-schedule)
- **weekday** : day of the week
- **workingday** : if day is neither weekend nor holiday - 1, otherwise - 0.
- **weathersit** : (weather situation)

  1: Clear, Few clouds, 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** : Normalised temperature in Celsius. The values are divided to 41 (max)
- **atemp** : Normalised feeling temperature in Celsius. The values are divided to 50 (max)
- **hum** : Normalised humidity. The values are divided to 100 (max)
- **windspeed** : Normalised wind speed. The values are divided to 67 (max)
- **cnt** : Count of total rental bikes

Objectives

The main objective of this study is to investigate how the above-mentioned environmental and seasonal factors affect the number of bicycle rentals per day. Furthermore, the author intends to improve the model to make future predictions about daily bicycle rental used.

The study will ultimately be beneficial for businesses in achieving the demand for sharing bikes by maintaining the supply.

Methodology

This section of the report explains the methodology adopted for the study.

The original data set consists of 14 variables including the response variable “Count of total rental bikes (cnt)”. But based on the simplicity and appropriateness of the variables in terms of interpretability and to avoid problems of multicollinearity, some variables were removed before the analysis.

Similar Variables	Correlation with cnt	Selected Variable
temp	0.4320092	atemp
atemp	0.4321852	
holiday	-0.06834772	holiday
weekday	0.06744341	
workingday	0.06115606	
mnth	0.2799771	season
season	0.4061004	

Furthermore “instant”, “dteday” and “yr” variables were removed based on their low importance to the model.

Since missing values affect the accuracy of model, it was planned to exclude them from the dataset next. By evaluating dataset, it was found that there are no missing values containing in dataset.

Initially the data set was partitioned into two parts as a training set and testing set based on the year and 2012 year data was used to build the model while 2011 year data was considered for validation. Thereafter, all possible selection procedure was used to select the appropriate model for the study. The model derived was then evaluated based on the model assumptions and the final model was obtained after performing necessary transformations with the assumption violations. Model significance and validation were done afterwards to check for its appropriateness towards fulfilling the research objectives.

All the computations have been performed mainly using the statistical programming language R.

Data Exploration

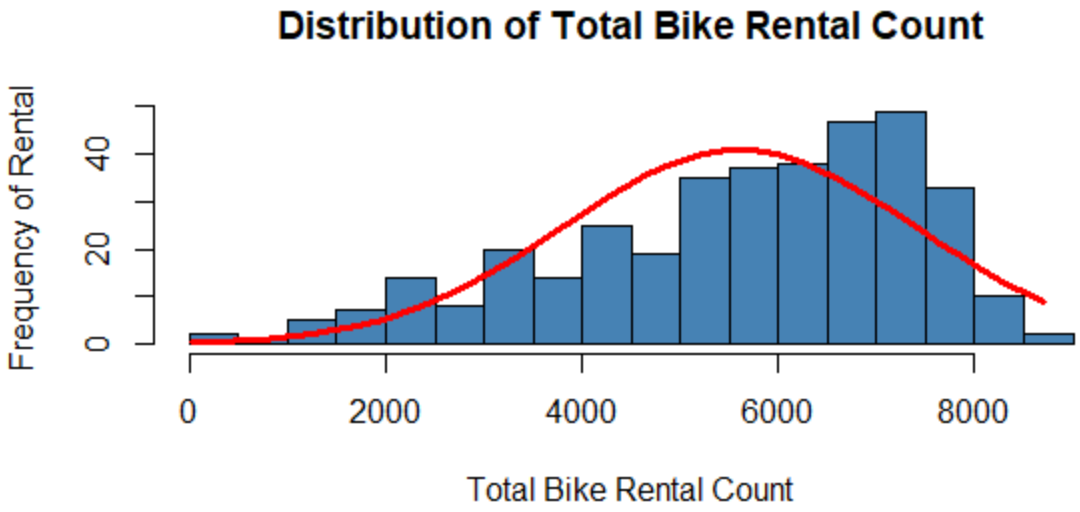
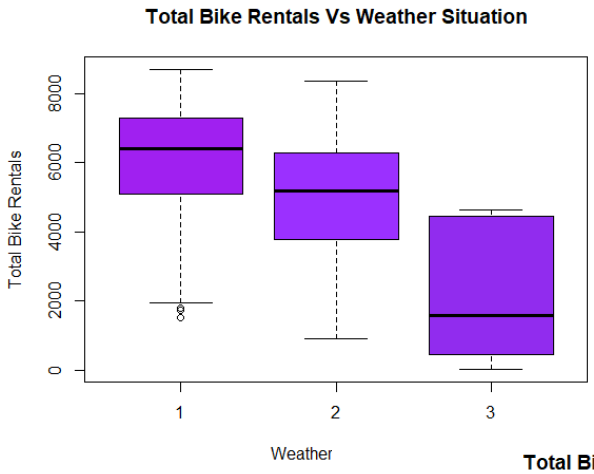
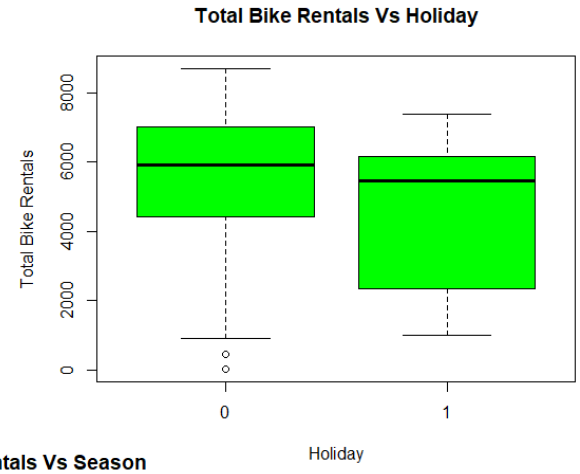


Figure 01: Distribution of total bike rental count in 2012

From Figure 1, it seems that the number of total rented bikes in 2012 follow a negatively skewed distribution. The mean of the total bike rental count is around 5600 while the median is around 6000. This graph indicates the fact that the people have an overall tendency towards using rented bikes throughout the year despite of external factors.





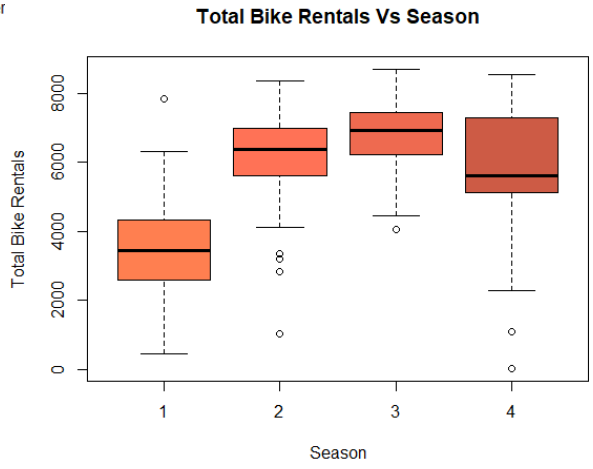


Figure 02: Distribution of bike rental counts with respect to categorical variables

The 1st boxplot demonstrates that the lowest number of rents is typical for the 3rd weather type (rain, snow, thunderstorm etc.) while the highest mean value of rentals have days with the 1st weather type (clear, partly cloudy etc.).There is a clearly decreasing trend of bike rentals when weather is bad.

In the 2nd boxplot, there is a slight difference between the average number of bike rentals on nonholidays than on holidays.

The 3rd boxplot shows that the number of bike rentals in summer and fall is high but it's the same in winter while in the spring mean number of bike rentals is the smallest.

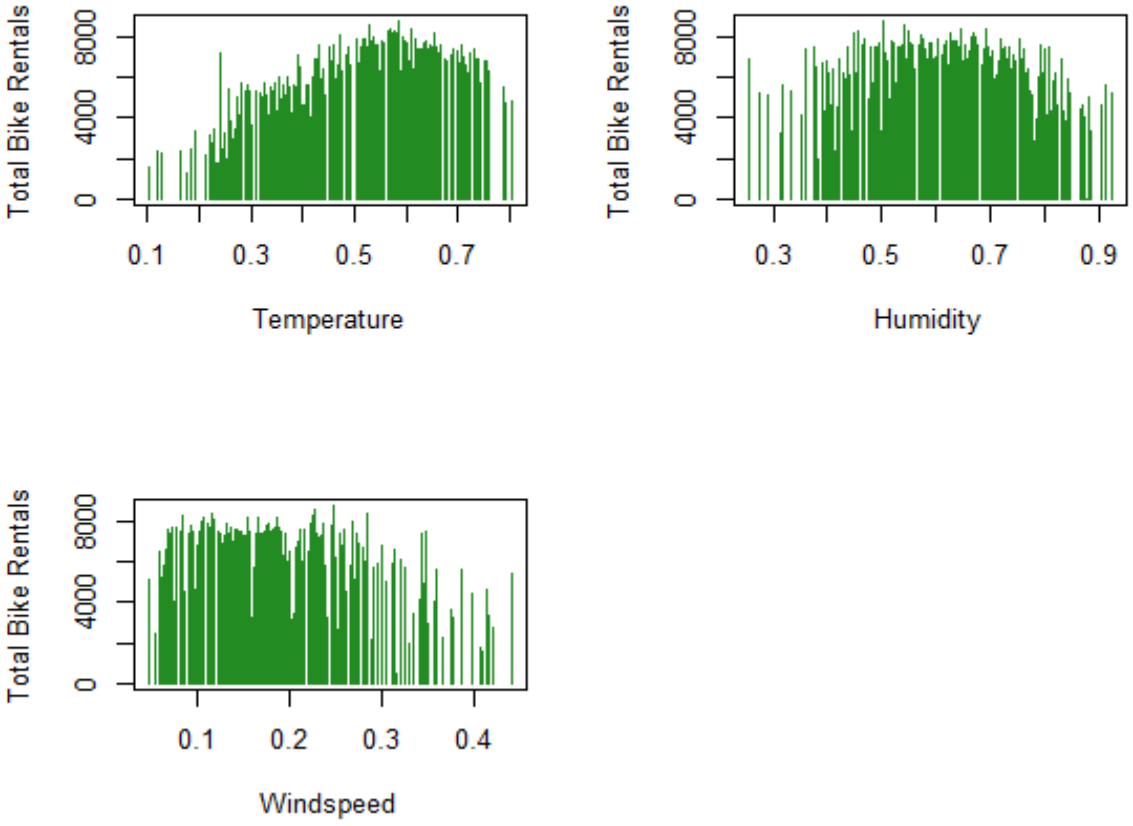


Figure 03: Distribution of quantitative predictor variables

It seems these numerical variables are distributed quite differently to each other with respect to bike rental counts. The above histograms indicate that the people tend to use more rented bikes in higher temperatures, mid humidity levels and lower windspeeds.

Figure 04: Correlation Plot

Figure 4 illustrates the correlations between each of the numeric variables. According to the above graph, normalised feeling temperature has a significant positive association with bike rental counts than other variables but humidity and windspeed also have slight negative correlations with bike rental counts. It is also an observable fact that each predictor variable has a correlation with other variables where a significant correlation can be identified in both humidity and windspeed.

Data Analysis and Results

Model Selection

As the first step of the data analysis, it was decided to find the best subset of predictor variables that are necessary and accountable for nearly as much of total variance of the response variable which bike rental count. In order to meet the objectives of the study, all-possible selection procedure was performed for model selection.

Best Subsets Regression

Model Index	Predictors
1	atemp
2	season atemp
3	season weathersit atemp
4	season holiday weathersit atemp
5	season holiday weathersit atemp windspeed
6	season holiday weathersit atemp hum windspeed

Subsets Regression Summary

Model	R-Square	Adj. R-Square	Pred R-Square	C(p)	AIC	SBIC	SBC	MSEP	FPE	HSP	APC
1	0.5189	0.5175	0.5133	249.7365	6258.0163	5217.4238	6269.7242	564949922.5321	1552013.7537	4252.2835	0.4864
2	0.6141	0.6099	0.6021	130.5879	6183.2412	5138.9255	6206.6570	454318558.5346	1258405.9546	3447.9973	0.3922
3	0.6932	0.6880	0.6744	32.1317	6103.3833	5058.2178	6134.6043	362289933.0177	1009028.1814	2764.8747	0.3136
4	0.7035	0.6977	0.6832	20.9591	6092.8120	5047.8565	6127.9357	351029745.3678	980324.4962	2686.4238	0.3047
5	0.7115	0.7050	0.6904	12.8573	6084.8700	5040.1790	6123.8964	342574405.9044	959304.8445	2629.0590	0.2982
6	0.7200	0.7129	0.6981	4.0000	6075.8746	5031.6033	6118.8036	333364365.7684	936038.2320	2565.5638	0.2909

AIC: Akaike Information Criteria

SBIC: Sawa's Bayesian Information Criteria

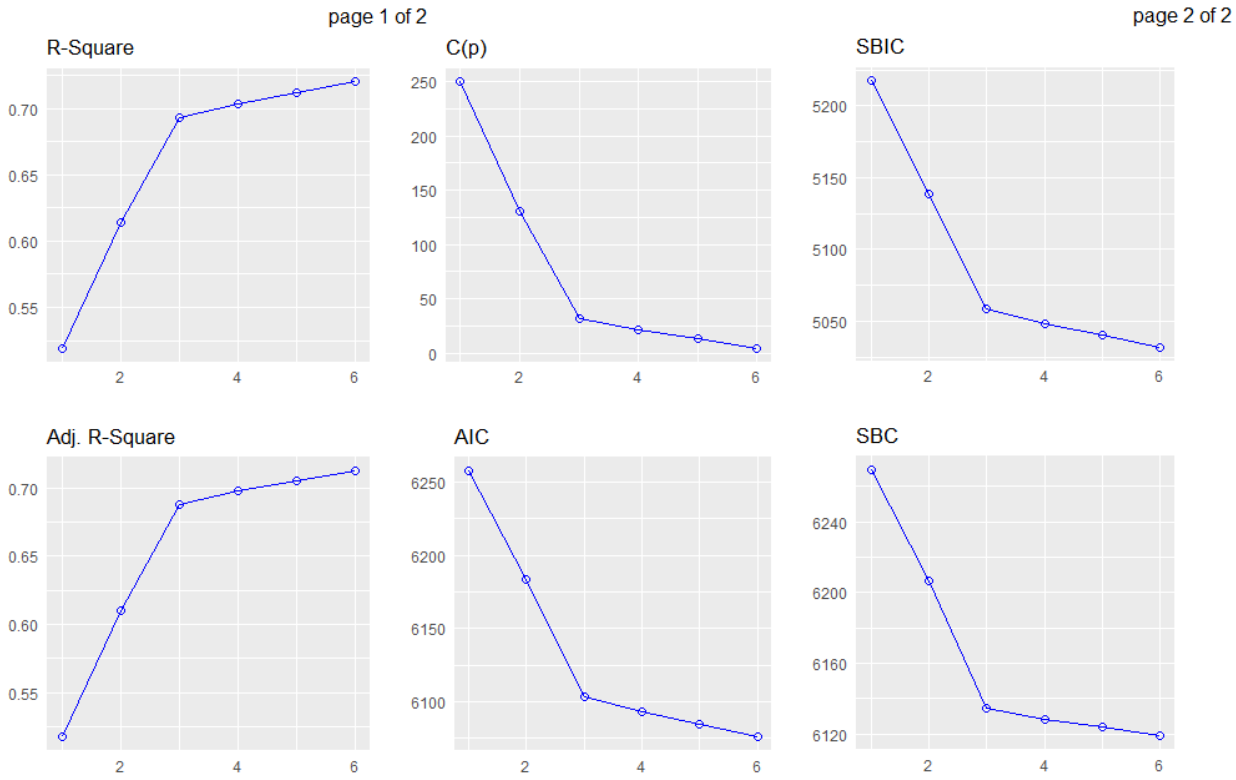
SBC: Schwarz Bayesian Criteria

MSEP: Estimated error of prediction, assuming multivariate normality

FPE: Final Prediction Error

HSP: Hocking's Sp

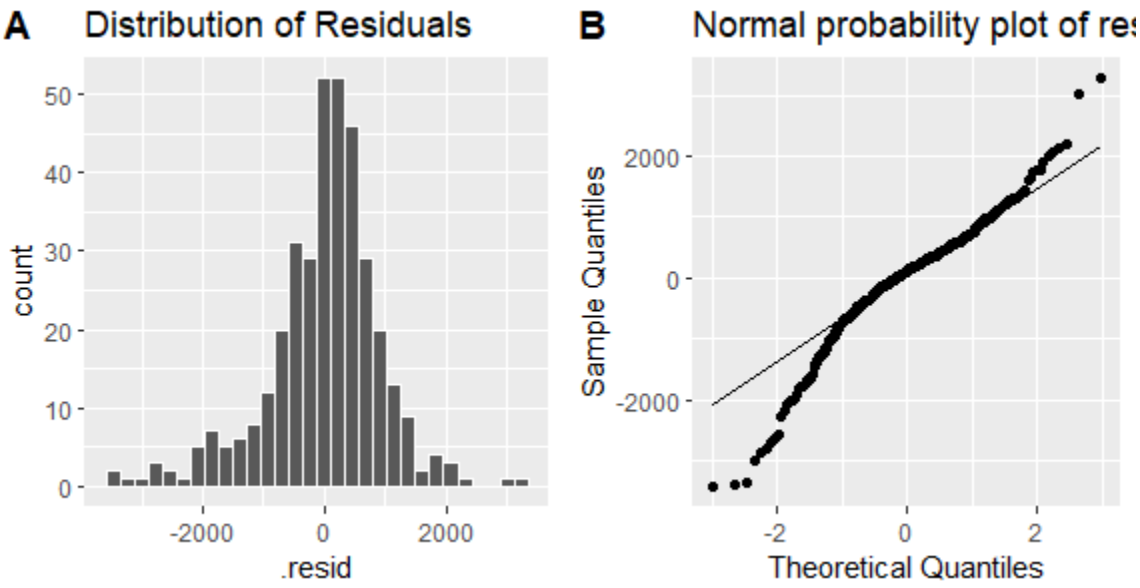
APC: Amemiya Prediction Criteria



According to the Adj. R square values in each model, model 6 which consists of all the predictor variables has the highest value. Moreover, model 6 has the minimum AIC and SBIC values which confirms the suitability of the model for the study. Therefore, full model was chosen as the best model.

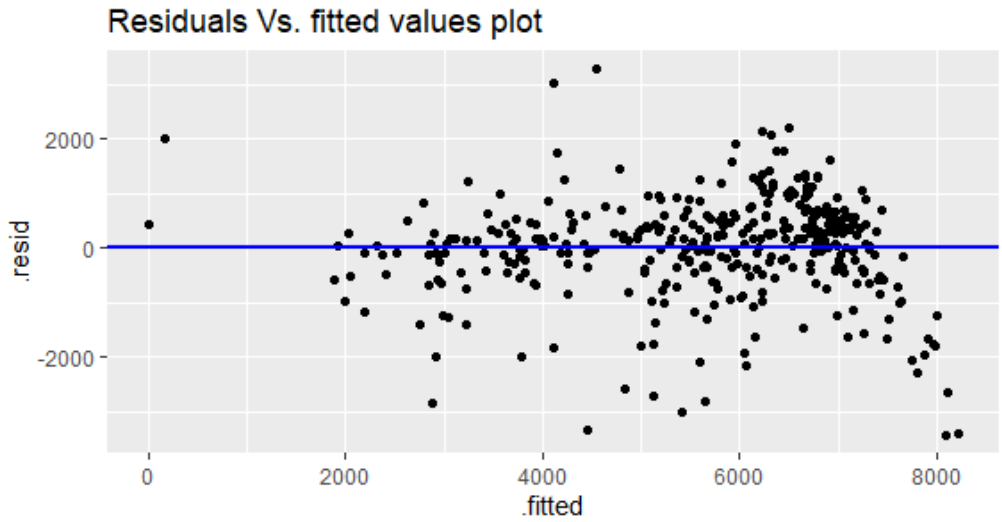
Model Assumptions

Normally distributed errors



According to the histogram, it is visible that the distribution approximately normally distributed. As per the normal probability plot of residuals, most of the points are along the straight line. Therefore, it can be concluded that the residuals are normally distributed.

Expected value of residuals should be zero



In the Residuals Vs. fitted values plot, residuals are distributed randomly around 0. This let us conclude that expected value is approximately equals to zero.

Multicollinearity

Multicollinearity is the occurrence of high intercorrelations among two or more independent variables in a multiple regression model. Multicollinearity can lead to skewed or misleading results when a researcher or analyst attempts to determine how well each independent variable can be used most effectively to predict or understand the dependent variable in a statistical model.

A variance inflation factor (VIF) provides a measure of multicollinearity among the independent variables in a multiple regression model. To ensure the model is functioning properly without multicollinearity issues, VIF was used as a measuring tool in this study

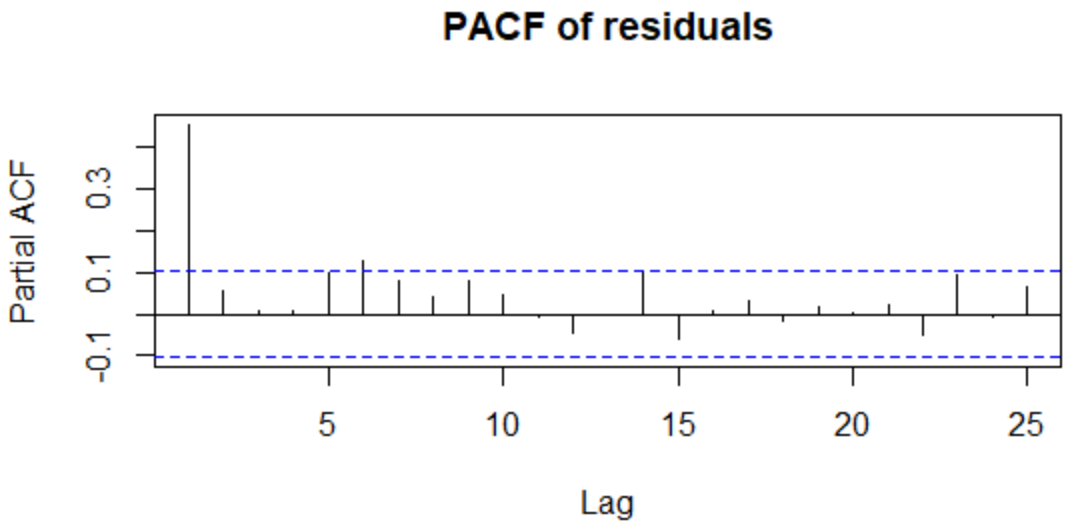
	GVIF	Df	GVIF^(1/(2*Df))
season	**3.261206**	**3**	**1.217764**
holiday	**1.012254**	**1**	**1.006108**
weathersit	**1.826079**	**2**	**1.162465**
atemp	**3.333197**	**1**	**1.825705**
hum	**1.928398**	**1**	**1.388668**
windspeed	**1.216834**	**1**	**1.103102**

Since all the VIF values are less than 10, any harmful multicollinearity is not found to be present in the data and this assumption is therefore regarded as confirmed.

Autocorrelation

Autocorrelation represents the degree of similarity between a given time series and a lagged version of itself over successive time intervals. Autocorrelation can also be referred to as lagged correlation or serial correlation, as it measures the relationship between a variable's current value and its past values.

Following PACF plot of the residuals helps to identify the lag values.



There is a lag of 1 is clearly visible since the lag-1 partial autocorrelation is so large and way beyond the "5% significance limits" which is shown by the blue lines. The partial autocorrelations at lags 6 is only slightly beyond the limits and would lead to an overly complex model at this stage of the analysis.

We can also perform Durbin-Watson test in detecting autocorrelation.

Hypothesis:

*H*0: There is no correlation among the residuals.

*H*1: The residuals are autocorrelated.

lag	Autocorrelation	D-W Statistic	p-value
1	0.453116	1.083706	0

Alternative hypothesis: rho != 0

Since the p-value is less than 0.05, we can reject null hypothesis and conclude that autocorrelation exists among the residuals.

To address this issue, Cochrane-Orcutt Estimation is used which is an interactive method using to solve first order autocorrelation problems.

Cochrane-orcutt estimation for first order autocorrelation

Call:

lm(formula = cnt ~ season + holiday + weathersit + atemp + hum + windspeed, data = bike_train)

number of interaction: 8

rho 0.498746

Durbin-Watson statistic

(original): 1.08371 , p-value: 9.155e-20

(transformed): 2.05021 , p-value: 6.451e-01

coefficients:

(Intercept) season2 season3 season4 holiday weathersit2 weathersit3

3923.9486 1131.5811 875.8317 1557.0963 -507.2699 -472.1028 -2066.0068

atemp hum windspeed

6568.3644 -2724.7990 -2597.7907

The above output brings out the coefficients of the transformed model after solving for multicollinearity.

Heteroscedasticity

The unequal error variance of a multiple linear regression model is called heteroscedasticity. To formally test for heteroscedasticity of the model, Whites' General Heteroscedasticity Test was performed.

Studentized Breusch-Pagan test

data: model2

BP = 1.2381, df = 2, p-value = 0.5385

Hypothesis:

*H*0: Variance of the error term is constant (homoscedasticity)

*H*1: Variance of the error term is not a constant (heteroscedasticity)

Since the p-value is greater than 0.05, we do not have enough evidences to reject the null hypothesis. Therefore, heteroscedasticity is not present in the model. Hence, we can use the model derived in the previous step.

Results Interpretation

Call:

lm(formula = cnt ~ season + holiday + weathersit + atemp + hum +

windspeed, data = bike_train)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3923.95	403.53	9.724	< 2.2e-16 ***
season2	1131.58	278.84	4.058	6.087e-05 ***
season3	875.83	330.61	2.649	0.008431 **
season4	1557.10	251.77	6.185	1.714e-09 ***
holiday	-507.27	228.27	-2.222	0.026896 *
weathersit2	-472.10	112.01	-4.215	3.174e-05 ***
weathersit3	2066.01	342.61	-6.030	4.108e-09 ***
atemp	6568.36	697.86	9.412	< 2.2e-16 ***
hum	-2724.80	502.17	-5.426	1.069e-07 ***
windspeed	-2597.79	613.97	-4.231	2.963e-05 ***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 835.1595 on 358 degrees of freedom

Multiple R-squared: 0.5646 , Adjusted R-squared: 0.5573

F-statistic: 51.2 on 6 and 358 DF, p-value: < 7.29e-59

Durbin-Watson statistic

(original): 1.08371 , p-value: 9.155e-20

(transformed): 2.05021 , p-value: 6.451e-01

After selecting the most suitable variables and applying necessary remedies for assumption violations, following model was derived as the best fitted model for bike rental counts.

$$\boldsymbol{Y_i = 3923.95 + 1131.58X_{i1} + 875.83X_{i2} + 1557.10X_{i3} - 507.27X_{i4} - 472.10X_{i5} + 2066.01X_{i6} + 6568.36X_{i7} - 2724.80X_{i8} - 2597.79X_{i9}}$$

where $\varepsilon \sim N\ (0, \sigma^2\ )$; ε's are independent.

$\boldsymbol{Y_i}$: Count of total rental bikes

$\boldsymbol{X_{i1}}$, $\boldsymbol{X_{i2}}$, $\boldsymbol{X_{i3}}$: dummy variables of season

$\boldsymbol{X_{i4}}$: holiday variable

$\boldsymbol{X_{i5}}$, $\boldsymbol{X_{i6}}$: dummy variables of weather situation.

$\boldsymbol{X_{i7}}$: Normalized temperature in Celsius.

$\boldsymbol{X_{i8}}$: Normalized humidity.

$\boldsymbol{X_{i9}}$: Normalized wind speed.

Significance of the Model

The F-test of overall significance indicates whether the derived linear regression model provides a better fit to the data than a model that contains no independent variables.

Hypothesis to be tested:

H_0: $\beta_1 = \beta_2 = \ ... \ = \beta_9 = 0$

H_1: At least one $\beta_i \neq 0 \quad ; \ i = 1, 2, ..., 9$

The reported F-statistic is 51.2 with a p-value < 0.05. Therefore, we reject the null hypothesis. Hence, the full model with all the predictors is statistically significant at 5% level of significance.

Adjusted R^2 is a corrected goodness-of-fit (model accuracy) measure for linear models. It identifies the percentage of variance in the response that is explained by the input or inputs. Since the Adjusted R^2 value is 0.5573 for this model, it can only explain 55.73% of the variability in the bike rental count which isn't a much satisfactory amount.

Significance of the Model Parameters

The t-test conducted for each model parameter is obtained from the model summary output above.

Hypothesis:

H_0: $\beta_i = 0$

$H1$: $\beta_i \neq 0 \;\; ; \; i = 1, 2, \ldots, 9$

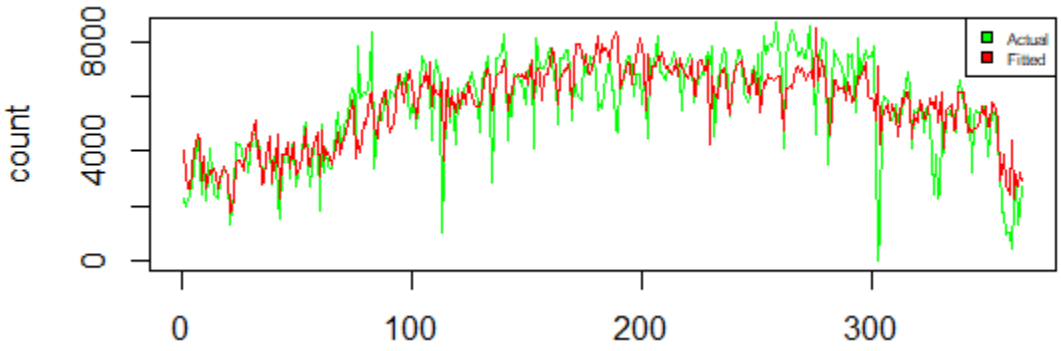
All the variables were reported to be statistically significant when other variables are already in the model at a significance level of 5%.

Model Validation

Model Validation is the final step of model building process which involves checking the fitted model against independent data.

Since the full data set contains data of both 2011 and 2012 years, the dataset was divided by the year where 2012 data was taken as a training dataset to build the model while 2011 data was chosen for validation of the model.

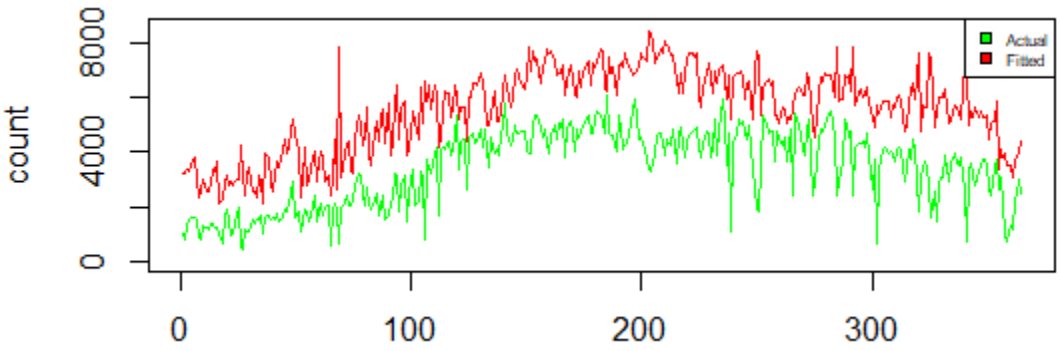
Mean Squared Error was calculated at first using the actual and fitted bike rental counts for 2012 and the following graph and the error value were obtained.



MSE of the fitted Model	1223342

By observing the graph above, it can be seen that the model obtained using 2012-year data approximately fits with the actual data.

The second tool used to validate the model was mean of squared prediction error (MSPR). This tool basically tests the predictive ability of the fitted model. Furthermore, actual values were also plotted with the fitted values for the year 2011 as given below.



MSPR	5513102

Since MSPR is much higher than MSE, the MSE of the fitted model is seriously biased and gives an inappropriate indication of the predictive ability of the model.

According to MSPR, predicted testing data is very far from the observed testing data which indicates a large error. This gives the idea that the model derived is best when it comes to the particular year the model was built but its predictive ability is very low.

Conclusion and Discussion

The goal of the study was to investigate how environmental and seasonal factors affect the number of bicycle rentals per day. Initially, the exploration done through graphical illustrations indicated an approximate idea that the variables considered as predictors have an impact towards the response variable which is bike rental counts but when considering the correlations, apart from the normalized feeling temperature, others had slightly low association with the response variable.

A perfect model to predict the number of bike rental counts is very hard to achieve, since it has to include all aspects of what makes the count valuable. The results of our study showed that our final fitted model explains about 56% of variability in the response variable which was not within the satisfactory level. However, we found that the MSE of the fitted model is seriously biased and gives an inappropriate indication of the predictive ability of the model. Moreover, with the MSPR value, the model derived has a low predictive ability and it suits only for the particular year the model was built.

And also, significant autocorrelation structure in the residuals from regression models suggests that future investigation could use Time Series analysis methods to predict the number of bike rental users.