Regression and Time Series Model
(MA31020)

Project on Regression and Time Series Analysis

Bike Rental Count Prediction

Submitted By:

Prerit Jain
(16IM10035)

The given dataset consists of the the count of bike rentals on an hourly basis along with various others enviormental and seasonal parameters on which the number of rentals is highly dependent. I have performed the analysis of both the daily as well as hourly data.
Regression analysis is applied to the daily data, with various transformations, and residual analysis.
On the other hand, both the time-series analysis and regression analysis is applied to the daily data. The following report shows some particular characterstics of the analysis done.

DATA PREPROCESSING AND TRANSFORMATIONS:

Various preprocessing steps are performed in order to improve the model adequacy and train a better regression model on the given dataset.

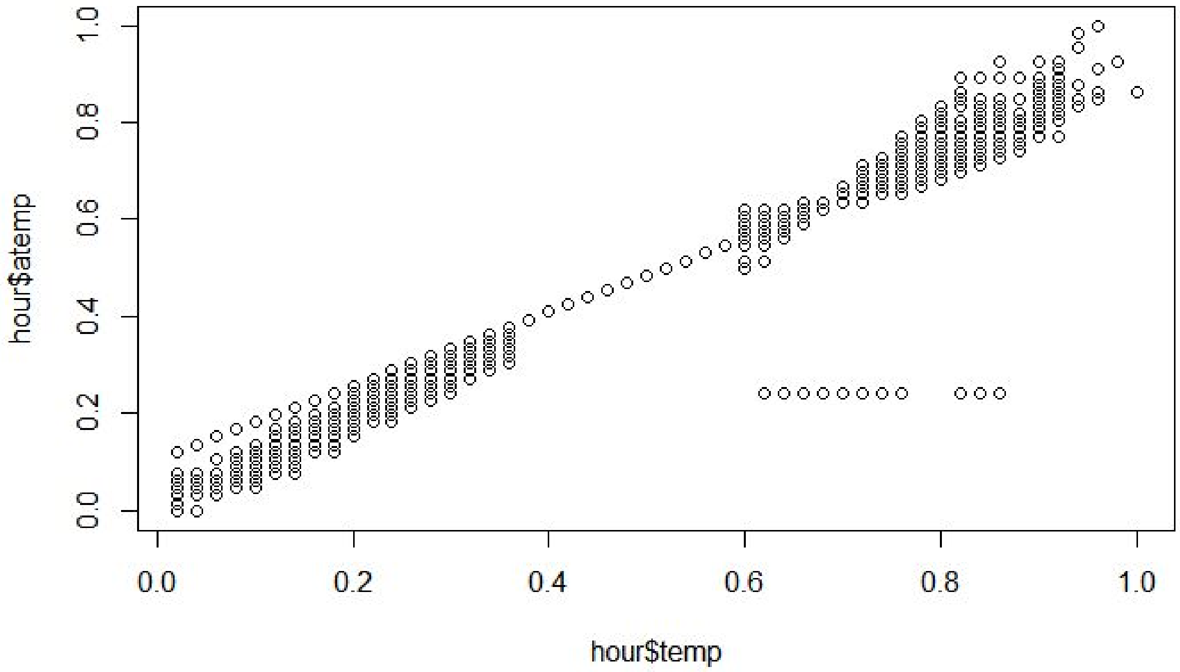
Creation of dummy variables:
Since most of the variables in given dataset are categorical in nature, hence in order to perform regression analysis on it, the creation of dummy variable corresponding to each instance of each categorical variable. In R the lm function automatically creates the dummy variables if we give input categorical variable as type factor. So, firstly all the categorical variables are converted to factor type.

Skewness removal and Normalization:
Skewness in statistics represents an imbalance and an asymmetry from the mean of a data distribution. Skewness creates a big problem while predicting the values of an outcome variable. As the data itself is skewed, then it means that it deviates from the regression line very much, so increasing the sum of squared error in predictions. So removing the skewness by applying a **suitable transformation** is very important. Here we checked the skewness value of the outcome variable for both hourly and daily data using the **function skewness** in the R package **‘e1071’.**
The skewness for the cnt variable i.e the outcome variable for **hourly data** came out to be **+1.277191**, which tells us that data is highly positively skewed. So, inorder to minimize the skewness and dispersion of the data, logarithmic transformation and normalization are performed on the outcome variable of hourly data.
On the other hand, cnt variable of the **daily data** had a skewness of **-0.04715862** which is an evidence of **weak skewness** on the negative side and hence only normalization is performed in this case.

REGRESSION ANALYSIS ON HOURLY DATA

Multiple linear regression model is trained over the hourly data as well as daily data. First of all, the model consisting of all the variables as independent variables except the outcome variable cnt is considered.

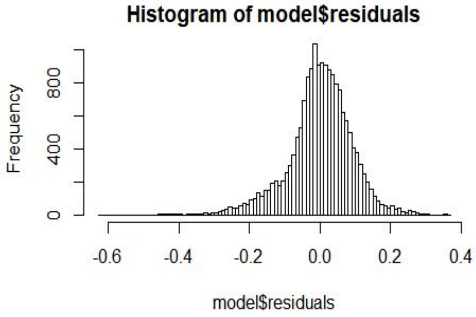
After checking the correlation matrix and analyzing some of the plots between variables we found out some of the variables were highly correlated and hence one of them is considered into the model in order to improve the quality of the model. For Example,

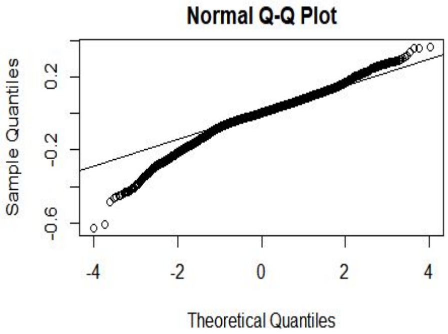


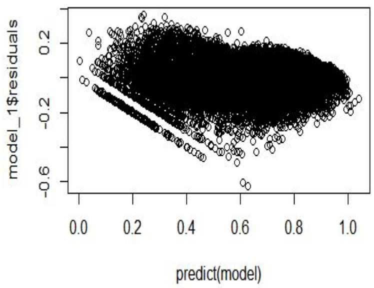
temp and atemp. Following curve shows the strong positive relationship between them. Some more variables are removed on the basis of the values of the t-statistic. **Best adjusted R-squared of 0.8235 and std residual value of 0.0905 is achieved. Following model is considered**

norm_log_cnt ~ season + yr + mnth + hr + holiday + weekday + weathersit + temp + atemp + hum + windspeed

Residual Analysis





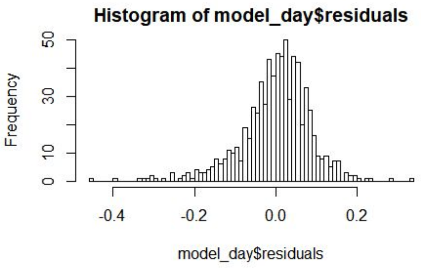


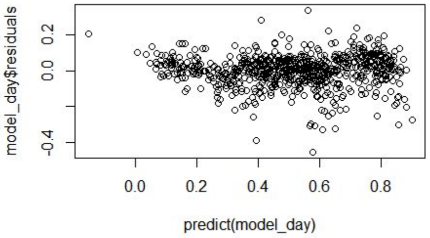
REGRESSION ANALYSIS ON DAILY DATA

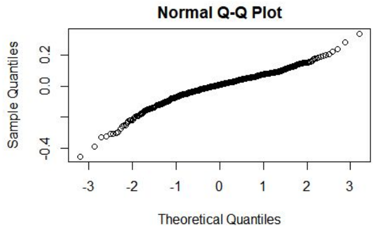
Similar analysis is performed for the daily data, and the results are as follows:

Adjusted R-squared = 0.8415
Residual std Error = 0.08873

Residual Analysis



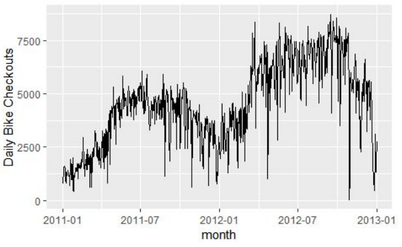




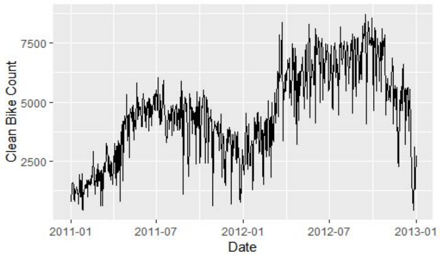
TIME-SERIES ANALYSIS OF DAILY DATA:

Dataset Examination:

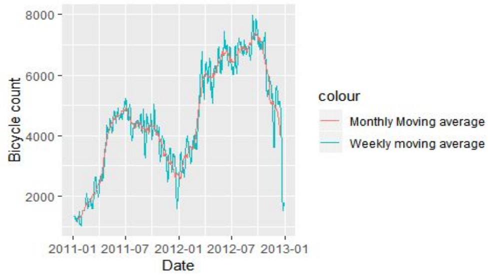
First of all the timeseries data is plotted and patterns and irregularities are examined.
The Data is cleaned using the tsclean function in the **'tseries' package of R.**



ORIGINAL DATA



CLEANED DATA

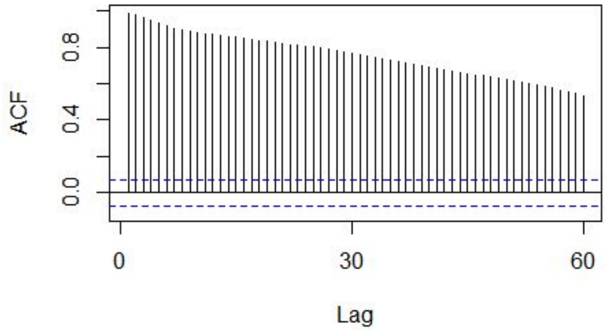


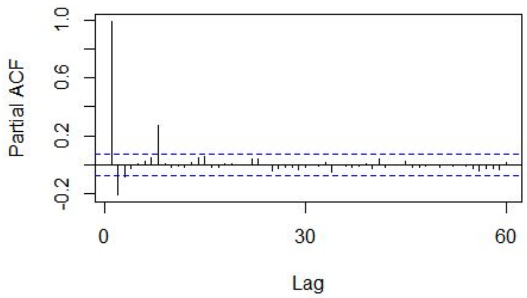
SMOOTHENING

The data is very much fluctuating in nature, which affects the results, hence the data is smoothed using Moving average with k = 3.

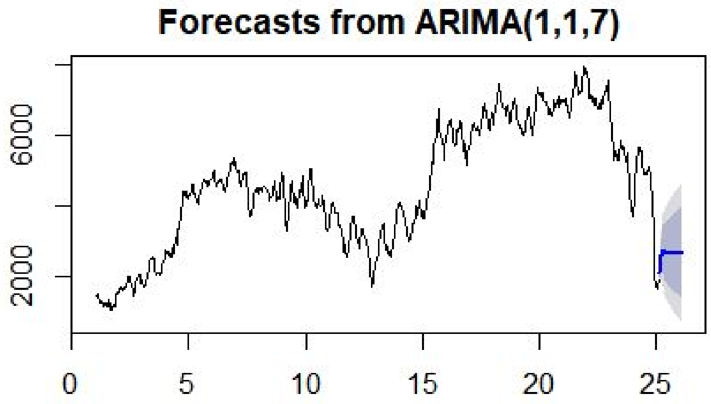
Now the data is decomposed into its components i.e seasonality is removed. In order to check the stationarity of the data points, the Augmented Dickey-Fuller Test is performed. The non-stationarity is removed using the differencing operator.

Finally, plots of ACF and PACF were plotted in order to determine the order of the ARIMA model which would be suitable for modeling the given time series data.





The final ARIMA Model selected is **ARIMA(1,1,7)** as it is evident from the above curves The forecasts from the ARIMA model is as follows:



--- **** ---