Regression Analysis for prediction of bike rental count hourly or daily based on the environmental and seasonal settings

Viren Bhanushali, Master in Industrial Engineering, Arizona State University

Abstract. 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 can easily rent a bike from a position and return back at another position. Today, there exists great interest in these systems due to their important role in traffic, environmental and health issues. The main objective behind this project is to find out the important parameters affecting the Bike rental count in Washington, D.C. The Data is analyzed and interpreted on SPSS software to perform Model generation and regression analysis. The interpretations are showcased in the form of result tables and graphs. Bike-sharing rental process is highly correlated to the environmental and seasonal settings. For instance, weather conditions, precipitation, day of week, season, hour of the day, etc. can affect the rental behaviors. Bike Sharing system can be a real time sensor network that can be used to 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**.** Comment on the translational aspect of the work presented in the paper and its potential clinical impact. Detailed discussion of these aspects should be provided in the main body of the paper.

Keywords- Bike rental, regression analysis, weather, environmental factors
(Note that the organization of the body of the paper is at the authors' discretion; the only required sections are Introduction, Methods and Procedures, Results, Conclusion, and References. Acknowledgements and Appendices are encouraged but optional.)

1. INTRODUCTION

A BICYCLE-SHARING SYSTEM, IS A SERVICE IN WHICH BICYCLES ARE MADE AVAILABLE FOR SHARED USE TO INDIVIDUALS ON A VERY SHORT-TERM BASIS FOR A PRICE. BIKE SHARE SCHEMES ALLOW PEOPLE TO BORROW A BIKE FROM A "DOCK" AND RETURN IT AT OTHER DOCK IN THE CITY, AS LONG AS THE TWO DOCKS BELONG TO THE SAME SYSTEM. People use bike-share for various reasons. Some who would otherwise use their own bicycle have concerns about theft or vandalism, parking or storage, and maintenance. E-bike sharing is becoming more popular. The e-bikes are generally recharged upon parking them at their station. E-bikes extend the range of the bikes and make cities with more difficult topographies more accessible to biking. People are more concerned towards the environmental crisis and thus are more interested by this. In this project, I analyzed Rental Bikeshare's data with the aim of predicting daily average demand using past demand, day, date and weather conditions as independent variables The dataset contains daily data for 2011-12 with three types of information that vary by their availability *w.r.t.* a day

_ **Day** variables: Characteristics of a day that are fixed and known before the start of a day, *e.g.* weekend, holiday.

_ **Weather** variables: weather information pertaining to a day that becomes available when the day starts, *e.g.* temperature, humidity, windspeed.

_ **Demand** variables: Demand count gathered at the end of a day. *e.g.* casual, registered and total bicycle demand.

2. Data Processing and parameters selection

Dataset contains information on 16 variables for two years, 2011 and 2012. This results in 731 rows and 16 columns. Out of 16, Instant data is just a record ID and data is transformed into numeric data type. The dataset includes 2 data: Day based and hour based. The characteristics of other variables are listed below in Table 1.

Table 1: Characteristics of data variables (Range)

Regressors	Description
Instant	record index
Dtedaydate	Date
Season	1: Spring, 2: summer, 3 : fall, 4: winter
Year	0:2011, 1: 2012
Mnth	month (1 ..12)
hr	hour (0 .. 23)
Holiday	1: Yes, 0: No
Weekday	day of the week (0-6)
workingday	1: if day is neither weekend nor holiday 0: otherwise
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)(0.06-0.85)
atemp	Normalized feeling temperature in Celsius. The values are divided to 50 (max)(0.07-0.90)
hum	Normalized humidity. The values are divided to 100 (max)(0.18-0.97)
windspeed	Normalized wind speed. The values are divided to 67 (max) (0.02-0.51)
casual	count of Unregistered users (9-3065)
registered:	count of registered users (416-4614)
cnt	count of total rental bikes including both casual and registered (431-6043)

2.1 Data exploratory analysis

Selection of the data range has varied just for year 2011, not considering the data of year 2012. From this assumption, the Year column would be 0, thus no analysis is done on that variable. Also, the Date variable is transformed into individual variables as Month and Day, thus no action required. Statistical analysis is done on all the other variables.

2.2 Transformation:

1. Converting categorical data into binary variables:
We are not using the month variables and instead using season variable. We have converted the season variable into binary by creating dummy variables.

2. Weekday: We have grouped 6 weekdays as a single group stating 1 when it is a weekday and 0 weekend.

3. Weathersit: Converted the variable into 4 different binary variables.

4. Removing a temp variable
We are using only atemp variable instead of both (temp and atemp) because there is high multicollinearity between the two data sets. We chose atemp because has a wider range of data thus giving us broad output.

5. Removing Holiday variable
We found out that the data is very less and figured out the dates of the holidays mentioned in year 2011.It directly doesn't affect the bike count, but I have made a graph showing the impact of holiday on bike count.

6. There were some missing values in the Humidity dataset, which were replaced by the average humidity value in year 2011.

2.3 Assumption Testing:

2.3.1 Hypothesis-The values of the residuals are normally distributed.

i. P-P plot: Checking the P-P plot for each value of the dependent variables, we found out that all the variables follow normal distribution except in casual regressor. (input P-P plot)

ii. Kurtosis and Skewness: Looking at the values from the descriptive table gives an estimate of non normality. The actual value of skewness and Kurtosis should lie within in the range of ±1.96. The data for all the variables are although not perfectly zero but are approximately linear and well within the range.
From the Histograms generated, I found that Casual and Windspeed are Positive skewed, therefore there is a need for transformation. I have applied Variance Stabilization technique for these variables. Transforming the data by applying square root, followed by analysis reduces the skewness values of Casual and Windspeed from 1.266 to 0.355 and 0.677 to 0.101.
(input histogram and descriptive table)

2.3.2 Hypothesis: Check Heteroscedasticity- The variance of the residual is constant.

Plotting the graph of standardized residuals versus predicted residuals. These graph does not show any formation of patterns. Also checking the plot between Standardized residuals versus individual independent variable, we found a rectangular pattern which shows there is no Heteroscedasticity in the model.

2.3.3. Hypothesis: At every value of the dependent variable the expected value of the residuals is zero.
The relationship between the Independent variable and dependent variable should be Zero.

Looking at the scatterplot between each Independent variable and Dependent variable, we found that there is linear relationship except in humidity variable. It shows that the relationship is Quadratic in nature, thus there is a need for transformation in the data.

I have created a new variable which is square of humidity values, after testing the plot there is increase in the R^2 square values.

2.3.4. Hypothesis: The expected correlation between residual for any cases is Zero. There is lack of autocorrelation.

Durbin-Watson Test: for sample size > 100, alpha = 0.05, number of regressors >5, we get dl=1.57 and du=1.78. The d value is 1.248, which is near to 2. So, there is no evidence of positive or negative order autocorrelation.

Also, looking at the graph between plot ID and residuals, there is no sign of evidence. There is no heteroscedasticity either.

2.3.5. Hypothesis: No independent variables are a perfect linear function of other independent variables.
Check for Multicollinearity:
While examining the data, we found that there is high correlation between casualsqrt and atemp, atemp and seasonfall and atemp and registered values.

Checking at the VIF scores, we find slightly high values for casualsqrt , atemp and seasonfall. Athough the tolerance values are nearly very high from the threshold values of tolerance (0.10 or 0.20).

For solving the Correlation between these variables, we add an interaction term between them.
a. atemp*registered
b. atemp*seasonfall
c. atemp*casualsqrt

After performing analysis, we found out that there is increase in R^2 value upto 0.997 i.e 99.7%. There is no requirement of having new interaction term atemp*seasonfall hence it is removed.

2.3.6. Hypothesis Testing: There are no influential cases biasing your Model.
Cook's Distance Statistic: None of the observations are above 1.0. Testing the value 4/(n-k-1), we get value as 0.00549. Checking the cook's distance from the dataset, we get 52 potential outliers.

Plotting a graph between Cook's distance and Centered Leverage Value, we get all the potential outliers. I removed the data points 463,685,595,69 followed by analysis. After removal of some outliers, R^2 value increased to 99.8%.

Thus here, we can conclude that the model is free of all the outliers and have linearity in it.

3.0 Procedures

3.1 Hypothesis Analysis: The test for significance of regression is a test to determine if there is a linear relationship
between the response y and any of the regressor variables $X_1, X_2, X_3, \ldots\ldots, X_8$.

Null hypothesis:
$H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = \beta_8 = \beta_9 = \beta_{10} = \beta_{11} = 0$

Alternate hypothesis:
$H_1: \beta_j \neq 0$ for at least one j

Critical Value Testing:
$f_{cv} = f_{\alpha/2,\ k,\ n-k-1}$
k: number of regressors
n: number of observations.
$f_{cv} = f_{0.05,11,713} = 1.83$(We assume a constant α value of 0.05).
We reject null hypothesis if the value of f_0 significance test is greater than the f_{cv} critical value otherwise we fail to reject the null hypothesis.

3.2 Initial Output:
The Initial output obtained by regression analysis gave us a linear model equation as

$$\hat{Y} = 2878.728 + 6762.093(atemp) - 1712.967(weatherLightsnow) + 1512.877(seasonwinter) - 3117.639(humidity) + 967.938(season\ summer) - 3150.184(windspeed) + 632.085(seasonfall) + 292.201(weekday\ yes)$$

This Initial Model gave us R^2 value = 0.553 = 55.3%. This explains the variability present in the model which can be further improved by performing assumption testing.

All the regression analysis is done by using stepwise variable input method because it gave me a higher R^2 and

R^2 adjusted values. Also, it helped me in finding the significance level for each regressor.

Final Output Model:
After performing all the assumption testing analysis and removing the outliers, the final output model linear equation for transformed data is:

$\hat{Y}$ = = *-46.531 + 1.021(registered) + 28.991(casualsqrt) -1924.231(atemp)+74.334(atemp_casualsqrt) + 212.842(weatherLightsnow) + 90.701(humidity_square) -0.075(atemp_registered) – 66.622(Season fall) + 43.499(weekday yes) – 23.967(Season summer) +18.640 (weather mist)*

This model gives us an R^2 value = 99.7%. This model explains approximately all the necessary parameters affecting the Bike rental count.

All the results are showed in the Tables and Graphs developed on SPSS softwares.

4.0. Graph and Tables: The results are presented in a different word file name “INSTANT”

4.1: References:
1.) Data Citation:
Use of this dataset in publications must be cited to the following publication:

[1] Fanaee-T, Hadi, and Gama, Joao, "Event labeling combining ensemble detectors and background knowledge", Progress in Artificial Intelligence (2013): pp. 1-15, Springer Berlin Heidelberg, doi:10.1007/s13748-013-0040-3.

2.) wikipedia: Bike rental sharing

3.) http://datasciencejourney.com/bike_share

4.2: Future Work:
The Data can be more accurate if the analysis is further conducted for year 2012. The data can also be interpreted on Hourly BASIS.

4.3 Conclusion:
The Model shows that the there is increase in the bIke count during fall with lightsnow and it does not depend on Humidity. Thus putting up the logic, the Bike share company should come up with more demand during the following seasons. The Data results is shown in the other file.
NOTE: THE TABLES AND GRAPHS ARE PRESENTED IN A DIFFERENT FILE DUE TO PROBLEM IN EDITING.