Multivariate Analysis of Capital Bike Share Ridership Data

Final Report BIA 652

Team B:

Alex Pollara
Danrong Liu
Naman Jain
Rui Xue

Introduction:

Bike sharing systems are a means of renting bicycles where the process of obtaining membership, rental, and bike return is automated via a network of kiosk locations throughout a city. Using these systems, people are able rent a bike from a one location and return it to a different place on an as-needed basis. Currently, there are over 500 bike-sharing programs around the world.

The data generated by these systems makes them attractive for researchers because the duration of travel, departure location, arrival location, and time elapsed is explicitly recorded. Bike sharing systems therefore function as a sensor network, which can be used for studying mobility in a city. In this competition, participants are asked to combine historical usage patterns with weather data in order to forecast bike rental demand in the Capital Bikeshare program in Washington, D.C.

You are provided hourly rental data spanning two years. For this competition, the training set is comprised of the first 19 days of each month, while the test set is the 20th to the end of the month. You must predict the total count of bikes rented during each hour covered by the test set, using only information available prior to the rental period.

Data Fields

Independent Variables:-

datetime - hourly date + timestamp

season - 1 = spring, 2 = summer, 3 = fall, 4 = winter

holiday - whether the day is considered a holiday

workingday - whether the day is neither a weekend nor holiday

weather -

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 - temperature in Celsius
atemp - "feels like" temperature in Celsius
humidity - relative humidity
windspeed - wind speed

casual - number of non-registered user rentals initiated
registered - number of registered user rentals initiated

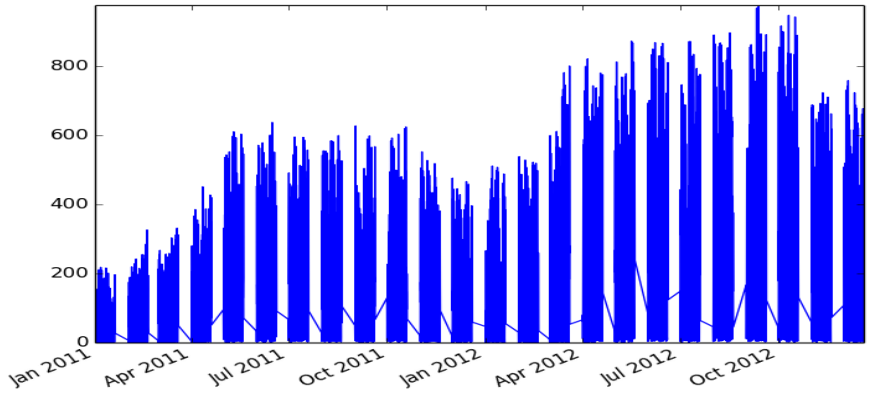
count - number of total rentals

NOTE : *Principal Component Analysis* was also performed but it did not yield significant results, so it was dropped. However, the results of PCA are attached in the pdf file.

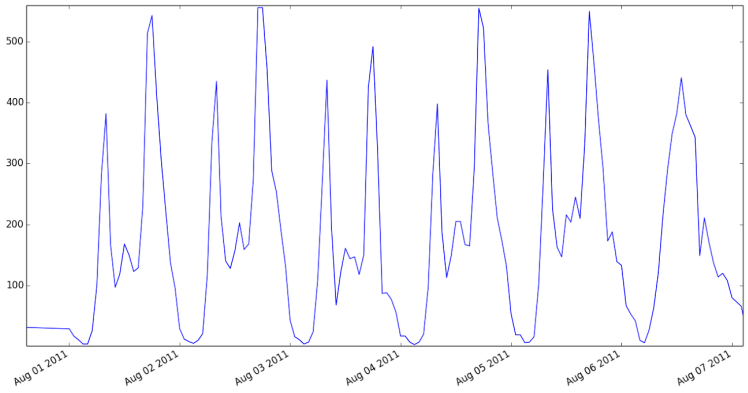
Initial Data Processing:

Examining the Data:

The data set in its initial form consisted of several metric and non-metric variables sampled on an hourly basis 24/7 from 2011-2013 with gaps at roughly two week intervals where data had been removed to form a test data set. Figure 1 shows the number of riders by hour over the entire span of the data set. Several things are immediately apparent: First there is a linear positive relationship between the average number of riders and time. Second that season clearly has some impact on usage of the system.



Looking at a smaller time span like Figure 2 we can see that ridership varies significantly between nighttime and daytime, and that there is a significant difference between usage patterns on weekdays and weekends. Most apparent is the impact of rushour on use of the system with large spikes in the number of riders during both the morning (06:00-10:00) and evening (16:00-19:00) rush hours. These peaks are absent on the weekends, when usage of the system appears to be more recreational.



Processing the Data:

In order to extract the maximum value from the time series data the following non-metric boolean variables were created:

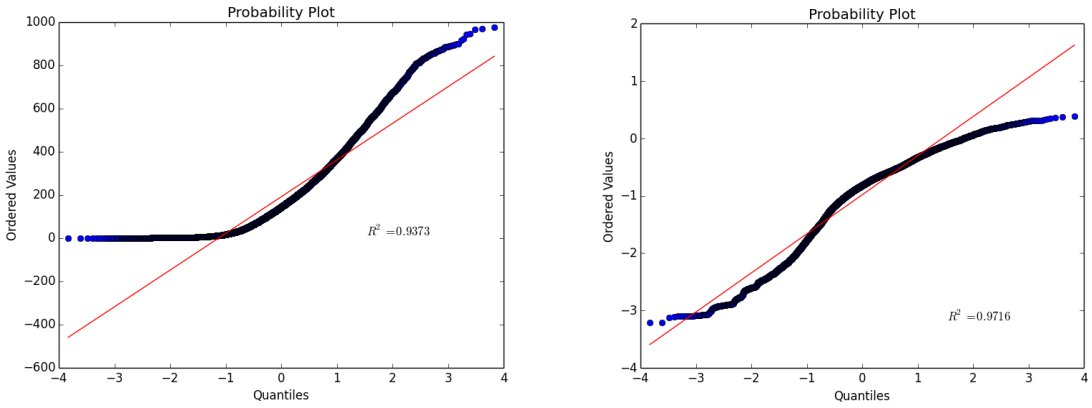
- AM/PM- a variable dividing the day based on whether it was AM or PM
- Daytime- a variable dividing the day based on whether it was daytime or night
- Rushhour- a variable which divided the day based on whether it was rush hour
- Weekday- a variable dividing time into weekdays and weekends

Since many multivariate techniques require only boolean non-metric variables the two categorical variables weather and season were then broken up into dummy variables by creating 3 boolean variables for each categorical variable to account for the 4 possible categories.

The dependent variable “count” or ridership presented another problem. Due to long term trend of increased use the number of people using the system was exponentially distributed with a majority of samples being relatively low values. As the system grew the number of people using it grew and therefore so did the distribution of the number of riders. In order to solve this a method for accounting for the overall grown of the system was needed. Through the bike share program’s website historical data on the number of bicycles available for use on any given day was obtained. This information was used to transform the count variable according to the following formula:

$$C_{new} = \log\left(\frac{\left(\frac{C_{old}}{B_{available}}\right)}{1-\left(\frac{C_{old}}{B_{available}}\right)}\right)$$

Where C_{new} is the new count variate, now nondimensionalized, and continuous. C_{old} and $B_{available}$ are the original count variable and the number of bicycles available respectively. The results of this transformation may be seen from the following two figures:



Normal probability before and after transformation respectively

Finally in order to facilitate the use of logistic regression a second dependent variable was created by dividing the dependent variable “count” into two classes, high and low usage.

Multiple Regression Analysis

Objective:
Multiple regression analysis may be useful in forecasting the count of the number of bikes in the Capital Bikeshare program in Washington D.C. This analysis also helped identify the variables that have the most influence on the count of the bikes in the system.
Analysis:
An examination of the correlation matrix between 17 variables *“holiday-weather3”* showed that (daytime) has the highest correlation of 0.621 with the target variable (count). Stepwise regression was then performed to detect variables (daytime, AM/PM, atemp, rushour, weather3, season1, humidity, weekend, season3, season2, windspeed, workingday, holiday, count) as the best-fit for dependent variable (count). These 13 predictor variables together accounted for 68.03% of variance in the dependent variable. This means that these 13 variables together are able to predict (count) with 68.03% accuracy, with Adjusted R-Square value of 68.00%. It was evident that with more independent variables the coefficient of determination increases very slightly to 68.04%, but still the Adj-Rsquare value remained the same, which indicates the inclusion of several independent variables that were non-significant in the regression equation.

Note that while performing the stepwise regression, two predictor variables viz., *temp* and *atemp* were showing high multicollinearity, so one of the variable (temp) with high multicollinearity was deleted.
The screenshots of the coefficient of determination and Parameter estimates are as shown,

Root MSE	0.39620	R-Square	0.6803
Dependent Mean	-0.98426	Adj R-Sq	0.6800
Coeff Var	-40.25365		

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate	Tolerance	Variance Inflation
Intercept	1	-1.25255	0.02768	-45.25	<.0001	0	.	0
daytime	1	0.50491	0.01056	47.80	<.0001	0.35513	0.53265	1.87741
AMPM	1	-0.60372	0.00863	-69.93	<.0001	-0.43103	0.77380	1.29232
atemp	1	0.01749	0.00075348	23.21	<.0001	0.21161	0.35369	2.82736
rushour	1	0.40668	0.01103	36.87	<.0001	0.27133	0.54290	1.84194
weather3	1	-0.25676	0.01513	-16.97	<.0001	-0.09885	0.86687	1.15357
season1	1	-0.16053	0.01164	-13.79	<.0001	-0.09882	0.57243	1.74693
humidity	1	-0.00229	0.00024573	-9.32	<.0001	-0.06291	0.64482	1.55083
weekend	1	0.09877	0.01471	6.72	<.0001	0.04967	0.53757	1.86024
season3	1	-0.06854	0.01424	-4.81	<.0001	-0.04244	0.37796	2.64577
season2	1	-0.03056	0.01189	-2.57	0.0102	-0.01892	0.54247	1.84341
windspeed	1	-0.00145	0.00050338	-2.89	0.0039	-0.01695	0.85380	1.17124
workingday	1	-0.05806	0.01098	-5.29	<.0001	-0.03865	0.55070	1.81586
holiday	1	-0.05730	0.02460	-2.33	0.0198	-0.01363	0.85877	1.16445

Finally, the training data set was scored, and the scoring coefficients were used to predict the count of the bikes in the test dataset.

Generalized Linear Model

Objective:

Generalized linear model (GLM) can be used to analysis how classified variables affect dependent variables, which is new count variable in this case. Independent variables include holiday or not, rush hour or not, weekend or not, daytime or not, AM or PM. Because we have only one dependent variable, and cell size distributed to different classes is not equal, we choose generalized linear model (GLM) to see those impacts of independent variables and interactions between them, instead of ANOVA. And also, GLM can be applied in a more general rage.

Analysis:

According to results from GLM, new count variable (X7) reflect significant difference when applied to a generalized linear model. And this model include the following independent variables: holiday (X1), rush hour (X8), weekend (X9), daytime (X10), AM or PM (X11). The model is quite fit, F value is 1453.91 ($p<0.001$), can be viewed in Figure 3:

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	15	3562.991658	237.532777	1453.91	<.0001
Error	10870	1775.888539	0.163375		
Corrected Total	10885	5338.880197			

R-Square	Coeff Var	Root MSE	X7 Mean
0.667367	-41.06593	0.404197	-0.984264

Figure 3

When turn to the specific effect of every independent variable in Figure 4, we can find that all rush hour (X8), weekend (X9), daytime (X10), and AM or PM (X11) have significant impacts on new count variable (X7). And their interactions give a significant influence to new count variable (X7) as well.

Looking at independent variables, then their interactions can be easily understood. Because almost all the independent variables are relative to time series, actually they may not be independent and have interactions.

Although new count variable (X7) on holiday (X1) does not show a significant difference, it still may have a real influence to X7. Because interactions between variables may decrease the effect of independent variables themselves, and interaction between holiday (X1) and rush hour (X8) shows an impact on new count variable (X7). In this case, afterwards multiple tests are needed to find if X1 really acts an important role in this model.

Source	DF	Type I SS	Mean Square	F Value	Pr > F
X1	1	0.008415	0.008415	0.05	0.8205
X8	1	945.196838	945.196838	5785.44	<.0001
X1*X8	1	3.906831	3.906831	23.91	<.0001
X9	1	79.191423	79.191423	484.72	<.0001
X1*X9	0	0.000000	.	.	.
X8*X9	0	0.000000	.	.	.
X1*X8*X9	0	0.000000	.	.	.
X10	1	1069.655719	1069.655719	6547.23	<.0001
X1*X10	1	1.934584	1.934584	11.84	0.0006
X8*X10	0	0.000000	.	.	.
X1*X8*X10	0	0.000000	.	.	.
X9*X10	1	11.890724	11.890724	72.78	<.0001
X1*X9*X10	0	0.000000	.	.	.
X8*X9*X10	0	0.000000	.	.	.
X1*X8*X9*X10	0	0.000000	.	.	.
X11	1	1111.181976	1111.181976	6801.41	<.0001
X1*X11	1	0.341453	0.341453	2.09	0.1483
X8*X11	1	84.677399	84.677399	518.30	<.0001
X1*X8*X11	1	1.050091	1.050091	6.43	0.0113
X9*X11	1	16.809607	16.809607	102.89	<.0001
X1*X9*X11	0	0.000000	.	.	.
X8*X9*X11	0	0.000000	.	.	.
X1*X8*X9*X11	0	0.000000	.	.	.

Figure 4