Regression Analysis of the Bike Sharing Demand

Chance Robinson, Jayson Barker and Neha Dixit

Master of Science in Data Science, Southern Methodist University, USA

Introduction

[Intro]

Data Description

The dataset we chose for this project was a publicly shared, hourly bike sharing dataset made available through Kaggle in csv format.

This data set is divided into two distinct sets – a train and test set. The train set consists of 10,886 rows (titled "train.csv") and the test set consists of 6,493 rows (titled "test.csv"). Within the training set, the first 19 days of each month are captured whereas in the test data set, the 20th day to the end of each month is present. The entirity of the data spans from 1/1/2011 through 12/31/2012 - encompassing two full years of bike sharing data. Interstingly, the time component of this analysis is captured in hours of each day meaning we have a calendar date represented 24 times (for each hour of that day) in the data, along with it's associated attribute values.

In train data set, there are a total of 12 attributes which capture multiple variables related to bike rentals. Some of these attributes are categorical, and others are continuous. All attributes are summarized in the table below:

Column Name	Type	Description
1. datetime	Date	YYYY-MM-DD HH24 (example: 2011-01-01 04:00:00)
2. season	Integer	(1-4)
3. holiday	Integer	(0 or 1)
4. workingday	Integer	(0 or 1)
5. weather	Integer	(1-4)
6. temp	Float	temparture in Celcius
7. atemp	Float	"feels like" temperature in Celsius
8. humidity	Integer	relative humidity
9. windspeed	Float	wind speed
10. casual	Integer	count of casual users
11. registered	Integer	count of registered users
12. count	Integer	count of total users `response variable`

** The test data set lacks the casual, registered and count variables.

Exploratory Data Analysis

Plotting Explanatory vs. Response Variables

Categorical Variables

Several numeric variables were found to be better suited to categorization and were converted to factors.

- Season
- Holiday
- Working Day
- Weather

Season	Label	Description
1	Spring	Dec 21 ~ Mar 20
2	Summer	Mar 21 ~ Jun 20
3	Fall	Jun 21 ~ Sep 20
4	Winter	Sep 21 ~ Dec 20

The datetime column was broken out into multiple factors as well so that we could visualize the components of each date and aggregate by different dimensions of the timestamp. We felt this was also necessary due to the nature of how the train/ test data sets had been pre-split. (i.e. . . . with the first 19 days of the month holding the only true counts to validate our models against.)

- Year
- Month
- Day
- Hour

Continuous Variables

- Count by Temperature
- Count by "Feels like" Temperature
- Count by Wind Speed

Correlation Matrix

There are several variables with a relatively high level of covariance. The following columns should therefore be removed so as not to be picked up by any automated techniques.

- atemp
- causual
- registered

Objective I Analysis

Question of Interest

Question of Interest

Model Selection

[Lasso, Forward, Backward, Stepwise, Manual]

Model Assumptions Assessment

- One
- Two
- Three

Comparing Competing Models

Comparing Competing Models

Parameters

Parameters

Model Interpretation

Model Interpretation

Conclusion

Conclusion

Objective II Analysis

Question of Interest

[Discuss time series method]

Time-Series Analysis

[Details goe here]

Model Assumption Assessment

Model Assumption Assessment

Comparing Competing Models

Conclusion

Conclusion

Appendix

Code

References