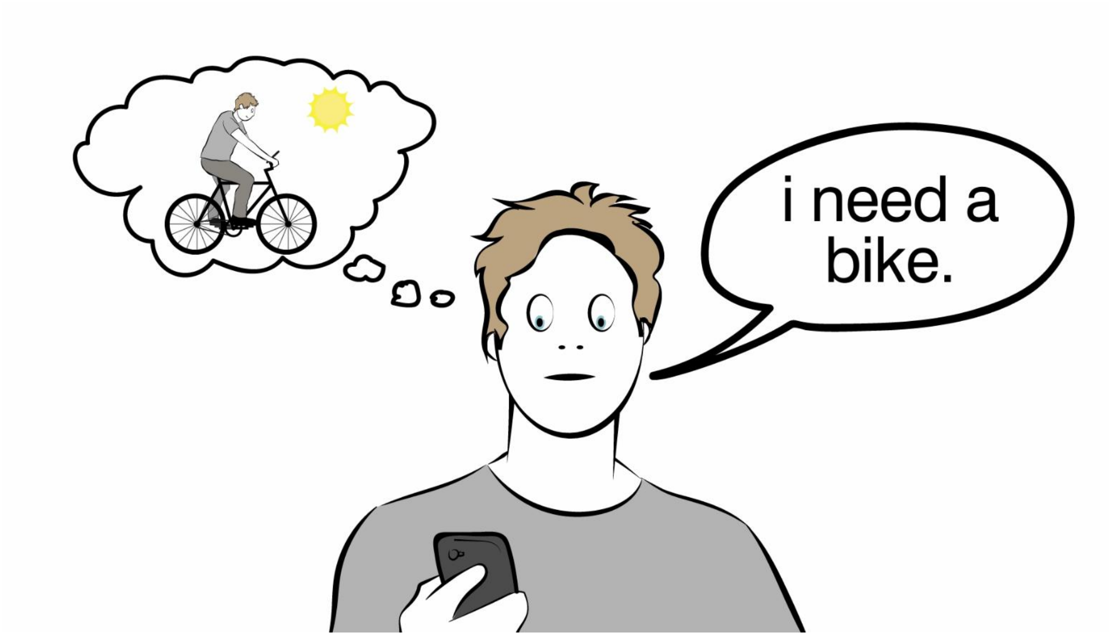
ST451: Bayesian Machine Learning
Analysis of Biking Sharing

Candidate Number: 40039



Analysis of biking sharing

1. Introduction

1.1 Background

Public bike sharing is a service in which bikes are made available for users to rent and return whenever and wherever an individual is in need. Due to its cheapness and convenience, biking sharing services are widely established across the world. Through rental and return track system, user information of each bike is automatically recorded. More issues about how to make the full use of public sharing and the pattern of number of users are of interest and in this study we would like to dig out information behind it.

1.2 Dataset Description

The datsets we used in this study contain both information on the basis of day and hour and include the following variables:

Response variable : the count of total rental bikes including both casual and registered users per day/hour, the count of causal users per day/hour, the count of registered users per day/hour

Explanatory variables:
instant -- record index
dteday -- date , Yr -- year, Mnth -- month, Holiday -- whether a day is holiday or not
Weekday -- day of the week, Workingday -- if day is neither weekend nor holiday is 1, otherwise is 0
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
Atemp -- normalized feeling temperature in Celsius
Hum -- humidity
Windspeed -- wind speed

1.3 Problem Statement

In this study, I intend to treat the response variable from perspectives of continuous response variables and sequential data respectively and employ the corresponding approaches. The targets of this study is shown as following:

1. Through building regression models, I would like to find out the most important factors in interpreting the number of rentals.
2. To find out the pattern of the daily sequential data of number of rentals using Gaussian process approach and Vector Autoregression model
3. Infer the trend and predict the future number of rentals
4. Finally propose strategies based on the findings above

2. Regression

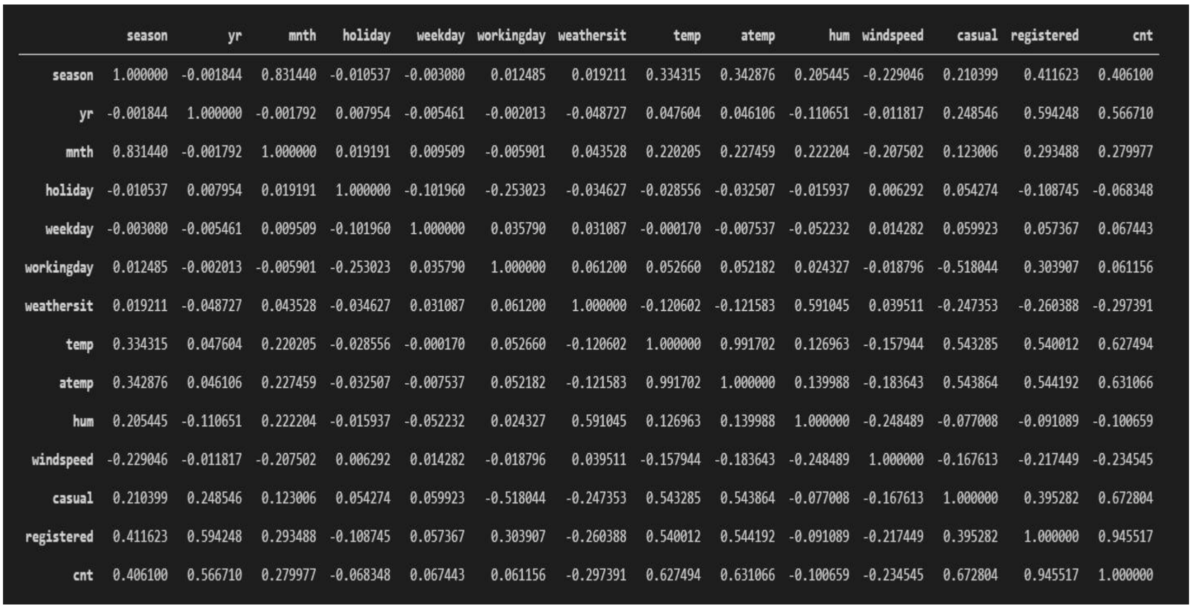
2.1 Frequentist linear regression

Firstly, a simple linear regression is built in order to find out the relationship between each explanatory factors and the count of total rental of bikes every day. General assumptions which are required to fit such a linear model:

1) All the observations in the dataset are required to be independent and normally distributed, i.e. no autocorrelation between residuals
2) The contributing factors should be negligibly dependent on time.
3) The trend of bike rentals should not exhibit a significant trend with time.
4) There exists no perfect multicollinearity between explanatory factors.

2.1.1 Detection of the multicollinearity

According to the table 1 below, it can be observed that there exists a highly linear dependent relationship between temp(measured temperature) and atemp(feeling temperature). To avoid this issue leading to the inconsistent estimates of standard error, one of them should be removed. Here, since we are concerned about human behavior in renting bikes, I prefer to keep the variable *feeling temperature* and drop the other off the model.



	season	yr	mnth	holiday	weekday	workingday	weathersit	temp	atemp	hum	windspeed	casual	registered	cnt
season	1.000000	-0.001844	0.831440	-0.010537	-0.003080	0.012485	0.019211	0.334315	0.342876	0.205445	-0.229046	0.210399	0.411623	0.406100
yr	-0.001844	1.000000	-0.001792	0.007954	-0.005461	-0.002013	-0.048727	0.047604	0.046106	-0.110651	-0.011817	0.248546	0.594248	0.566710
mnth	0.831440	-0.001792	1.000000	0.019191	0.009509	-0.005901	0.043528	0.220205	0.227459	0.222204	-0.207502	0.123006	0.293488	0.279977
holiday	-0.010537	0.007954	0.019191	1.000000	-0.101960	-0.253023	-0.034627	-0.028556	-0.032507	-0.015937	0.006292	0.054274	-0.108745	-0.068348
weekday	-0.003080	-0.005461	0.009509	-0.101960	1.000000	0.035790	0.031087	-0.000170	-0.007537	-0.052232	0.014282	0.059923	0.057367	0.067443
workingday	0.012485	-0.002013	-0.005901	-0.253023	0.035790	1.000000	0.061200	0.052660	0.052182	0.024327	-0.018796	-0.518044	0.303907	0.061156
weathersit	0.019211	-0.048727	0.043528	-0.034627	0.031087	0.061200	1.000000	-0.120602	-0.121583	0.591045	0.039511	-0.247353	-0.260388	-0.297391
temp	0.334315	0.047604	0.220205	-0.028556	-0.000170	0.052660	-0.120602	1.000000	0.991702	0.126963	-0.157944	0.543285	0.540012	0.627494
atemp	0.342876	0.046106	0.227459	-0.032507	-0.007537	0.052182	-0.121583	0.991702	1.000000	0.139988	-0.183643	0.543864	0.544192	0.631066
hum	0.205445	-0.110651	0.222204	-0.015937	-0.052232	0.024327	0.591045	0.126963	0.139988	1.000000	-0.248489	-0.077008	-0.091089	-0.100659
windspeed	-0.229046	-0.011817	-0.207502	0.006292	0.014282	-0.018796	0.039511	-0.157944	-0.183643	-0.248489	1.000000	-0.167613	-0.217449	-0.234545
casual	0.210399	0.248546	0.123006	0.054274	0.059923	-0.518044	-0.247353	0.543285	0.543864	-0.077008	-0.167613	1.000000	0.395282	0.672804
registered	0.411623	0.594248	0.293488	-0.108745	0.057367	0.303907	-0.260388	0.540012	0.544192	-0.091089	-0.217449	0.395282	1.000000	0.945517
cnt	0.406100	0.566710	0.279977	-0.068348	0.067443	0.061156	-0.297391	0.627494	0.631066	-0.100659	-0.234545	0.672804	0.945517	1.000000

Table 1

After removing linear dependent variables, variables such as *season* and *weather type* were converted into dummy variables. The result is shown in the table 2 below. It can be concluded that seasonality is an important factor in explaining the total number of rents. The baseline category is winter and the model suggests that the number of bike sharing in all seasons is greater than that of winter. The peak appears in fall and the coefficient can be interpreted that the figure in autumn is nearly twice as that in winter.

Besides, humidity and windspeed is negatively correlated with biking sharing cases. Users rent significantly less in more humid and and windy days. Poor weather also reduces the number of rentals greatly in the sense that both the coefficient of mist/cloudy variable and rain/snow are significantly negative and the latter one exhibits a large absolute magnitude. However, it turns out that holiday and working days are not deterministic factors in affecting renting bikes.

Summary of linear regression model				
	Coefficient	P-value	Lower bound	Upper bound
intercept	2647.8591	0.000	1926.948	3368.770
holiday	-489.1849	0.101	-1074.343	95.974
weekday	62.4684	0.010	15.076	109.861
workingday	107.0555	0.317	-102.858	316.968
atemp	6610.7400	0.000	5579.759	7641.721
hum	-2534.6641	0.000	-3444.042	-1625.286
windspeed	-2946.6667	0.000	-4270.759	-1622.574
spring	969.3500	0.000	621.116	1317.584
summer	626.5057	0.006	178.773	1074.239
fall	1496.7324	0.000	1197.358	1796.107
mist/cloudy	-255.3381	0.047	-506.712	-3.964
rain/snow	-1958.0093	0.000	-2600.732	-1315.286

Table 2

2.2 Bayesian linear regression

In contrast, a Bayesian linear regression model has been built. Since we have no prior information about the distribution of parameters, an uninformative uniform prior distribution was employed and the result was similar to the linear regression.

Summary of Bayesian linear regression			
	posterior mean	Lower bound	Upper bound
intercept	2628.3130	-1161.4451	6383.8571
holiday	-495.8488	-3616.3368	2609.5926
weekday	63.1088	-191.1258	308.3391
workingday	117.9243	-986.5708	1239.2526
atemp	6554.3181	1180.8752	11950.6315
hum	-2513.404	-7313.5886	2258.9999
windspeed	-2938.0653	-10056.5709	4041.4598
spring	982.8536	-827.7686	2772.4047
summer	648.6921	-1691.9903	2951.8816
fall	1497.5901	-97.0289	3035.7265
mist/cloudy	-255.0855	-1561.7789	1081.6790
rain/snow	-1948.1902	-5333.2043	1450.7332

Table 3

2.3 Best Subset Selection Model

To further improve the model efficiency and look for a better tradeoff between the number of parameters and the performance of data fitting, I would like to find out the most important factors among all potential variables. Since there are only 11 variables in the dataset, it's feasible to conduct the best subset selection method rather than simple forward/backward selection.

For each time, one possible combination of explanatory variables was selected and used to fit the linear model. Keeping the number of parameters fixed for each round, the model with highest R^2 and lowest RSS are chosen to be the candidate of the best model. Finally, the model which has both a relatively fitting performance and a simple form was considered to be the best model.

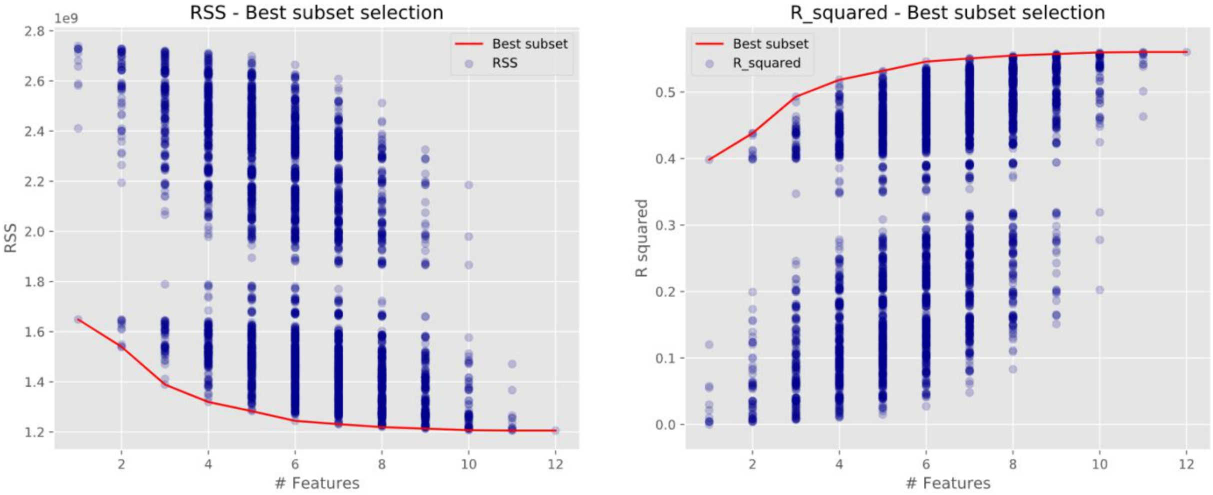


Figure 1

Figure 1 shows the scatter plot of RSS and R-squared of each model respectively. The red line depicts the trend of best candidate for each fixed number of parameters. Through observing the location of the elbow point, the best model which both captures simplicity and goodness of fit lies at the point where the number of parameters equal 5.

The variables included are *feeling temperature, humidity, spring, fall, rain/snow.*

Besides, the trend of R-squared and RSS suggests around 50% of variability in response variable is explained by explanatory factors. Since this figure is close to the limit of goodness of fitness under given information, it propels us to think about other ways in dealing with this problem.

2.4 Ridge Regression

Furthermore, ridge regression is another method in shrinking the coefficients and is widely used to reduce the complexity of the model. Compared with the full model including all the variables, the coefficients are shrunk towards zero. Judged from the table 4 below which compares the mean squared error between the full model, ridge regression model and full subset selected model, it can be observed that the performance of ridge model performs nearly as good as the full model in the sense that MSE of each differs slightly. Since it is of a much simpler form, we prefer the ridge regression to the full model.

	Mean squared error
Full model	1729720.1322
Ridge regression model	1732739.2828
Best subset selected model	1711748.0120

Table 4

An interesting finding is that the best subset selected model has a even smaller MSE than that of the full model. For this part, best subset selection proves to be the best method in building regression models.

Conclusion:

Regression method is the most straightforward way in finding out the relationship between the number of daily users and potential factors. Models with interaction effects also have been considered but seems it didn't improve the performance of regression model very much.

Due to the limited explanatory capability of contributing factors and some degree of violation of the basic assumption such as normality and independence between observations, it's worthwhile considering introducing more variables other than weather and date such as location and population density to further increase the prediction power of the model.

3. Sequential Data

3.1 Gaussian process Model

In this part, we treated the every-day number of users as sequential data. Thus, the analysis is conducted from perspective of finding the pattern from itself rather than draw information from contributing factors.

The model considered here is Gaussian process model. It assumes every linear combination of the collection of individuals are multivariate normally distributed and the most important advantage is its flexibility.

Kernel has been set beforehand and it captures the similarity between neighboring observations.

The overall pattern is presented in the figure 2 below.

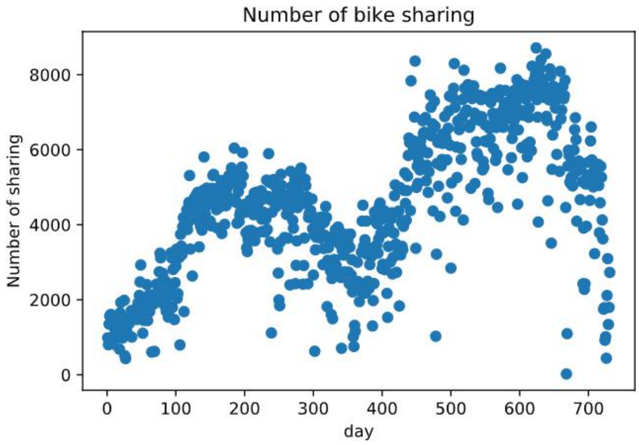


Figure 2

The overall trend reflects some seasonal pattern and generally neighboring observations share some similarity and tend to take close values with each other. The whole dataset is split into two parts for use of training and testing separately. Observations whose index are even numbers have been selected as the testing data.

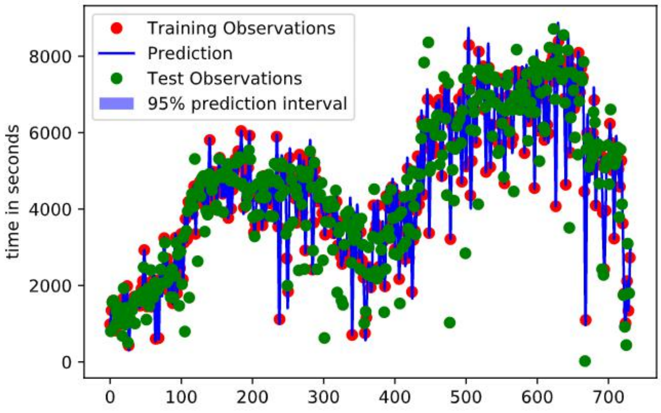


Figure 3

Conclusion and Comparison with regression method:

It turns out that Gaussian process method gives better performance in fitting the training data and allows the model to be much more flexible. On the basis of detection of figure 3, performance on testing data is fairly good and did not show many large deviations from the true value. This suggests the assumption of multivariate normal distribution holds to some extent.

However, in terms of each individual prediction, there exists some points have large variability. The reason behind is that the Gaussian process model didn't take any causing factors into account and thus failed to predict those observations with extremely high or low values. For instance, if a weather condition is extremely terrible and largely different from previous days, the number of users could be quite low. In this case, the regression model will give a more reliable prediction than Gaussian process model.

3.2 Vector Autoregression Model

As we have tried the fitting result of Gaussian process model, here I intended to analyse the sequential data using time series model. VAR (vector autoregression model) is an extension of time series algorithm which deals with multivariate sequential data and allow for influence between each other.

Since the total count of rentals is composed with both casual users and registered users, the VAR is an appropriate method in forecasting such bivariate sequential data.

Generally speaking, VAR is a system of equations which allow each variable depends on past lags. The formula of VAR is as follows:

$$y_{1,t} = c_1 + a_{1,1} y_{1,t-1} + a_{1,2} y_{2,t-1} + e_{1,t}$$

$$y_{2,t} = c_2 + a_{2,1} y_{1,t-1} + a_{2,2} y_{2,t-1} + e_{2,t}.$$

Based on the trend of time-series plot, casual users and registered users share the same pattern and correlates to each other to some extent.

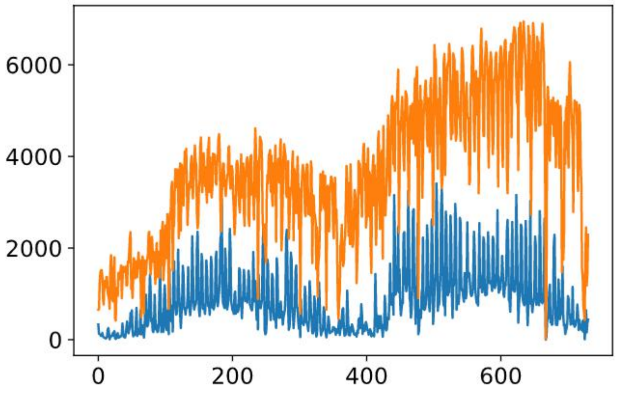


Figure 4

3.2.1 Check for causation using Granger's Causality Test

We have seen the correlated pattern between two variables across the time lag. Here, I aim to use quantitative method to test the causality.
The outcome of the test suggests there indeed exists a significant association between the current measure and past lags and it verifies the idea of introducing VAR.

Granger Causality					
Number of lags 1			Number of lags 2		
SSR-F test	F=65.89	P=0.000	SSR-F test	F=83.72	P=0.000
SSR-chi2test	Chi2=66.17	P=0.000	SSR-chi2test	Chi2=168.60	P=0.000
LR test	Chi2=63.34	P=0.000	LR test	Chi2=151.67	P=0.000
F test	F=65.89	P=0.000	F test	F=83.72	P=0.000

Table 5

3.2.2 Check for Stationarity Assumption

After splitting the training and testing dataset, the next step is to check whether the stationarity holds for measures of casual and registered users. Stationarity is one of the key assumptions in employing time series model in the sense that variables should have constant mean and variance across all time lags.

There are multiple ways to test whether the data is stationary or not and the method used here is so-called Augmented Dickey-Fuller Test. According to the test statistic, there's insufficient evidence rejecting the null hypothesis that two measures are stationary. Therefore, taking the one order difference is being implemented. Perform the procedure above again and the new outcome suggests the one-order difference data has been made stationary.

ADF Test on "casual"		ADF Test on "registered"	
Test Statistic = -8.3982	Number of Lags Chosen = 20	Test Statistic = -7.7014	Number of Lags Chosen = 20
P-Value = 0.0. Rejecting Null Hypothesis. Series is Stationary.		P-Value = 0.0. Rejecting Null Hypothesis. Series is Stationary.	

Table 6

3.2.3 Select the best order of the lags

Multiple trials have been conducted with maximum order of the lags equal to 9. It turns out that the best order of the lags is 7 as the model under the given condition has the lowest BIC. This finding aligns with the fact that the sequential data runs each circle on the basic unit of weeks.

3.2.4 Results of the model and prediction of testing data

The coefficients of each lag are presented in the figure 5 in appendix. The correlation coefficient between the number of casual users and registered users is 0.304.

Fit the selected model into the testing data and table 6 below compares the difference between parts of the predicted value and the true value.

ID	Predicted casual	Predicted registered	True casual	True registered
728	566.935373	1722.003830	159	1182
729	453.506012	1272.595720	364	1432
730	502.105111	1137.894075	439	2290

Table 7

Conclusion:

In summary, the VAR model fits sequential data problem due to its characteristic in capturing the dependence relationship between lags. However, the predicting ability remains to be improved because of the large variability of each observation. There exists some patterns across the time lags but in terms of predicting each specific individual, it’s inevitable to deviate from the true value.

4. Strategies and Improvement:

In short, this report revolved around analyzing the number of users for biking sharing business. In conclusion, the overall cases exhibits an increasing trend which means public bikes are getting more and more popular among citizens. However, the specific number of daily users vary largely and depend on lots of factors.

Both sequential data analysis and regression models play important roles in detecting some characteristics of the response variables. Time series model focuses on itself past trajectory and captures the seasonality trend while regression model draws information from strong explanatory variables. For further predicting purpose, with the help of accurate weather forecasting data and pre-specified information such as national holiday, a roughly reliable forecast could be generated on the basis of models above.

To dig out more extensions of this study, additional information such as geographical data and social-economic factors will be of great help. Multivariate sequential data analysis can also be conducted to find out the mutual relationship between bikes sharing services and public transportation development.

Reference:

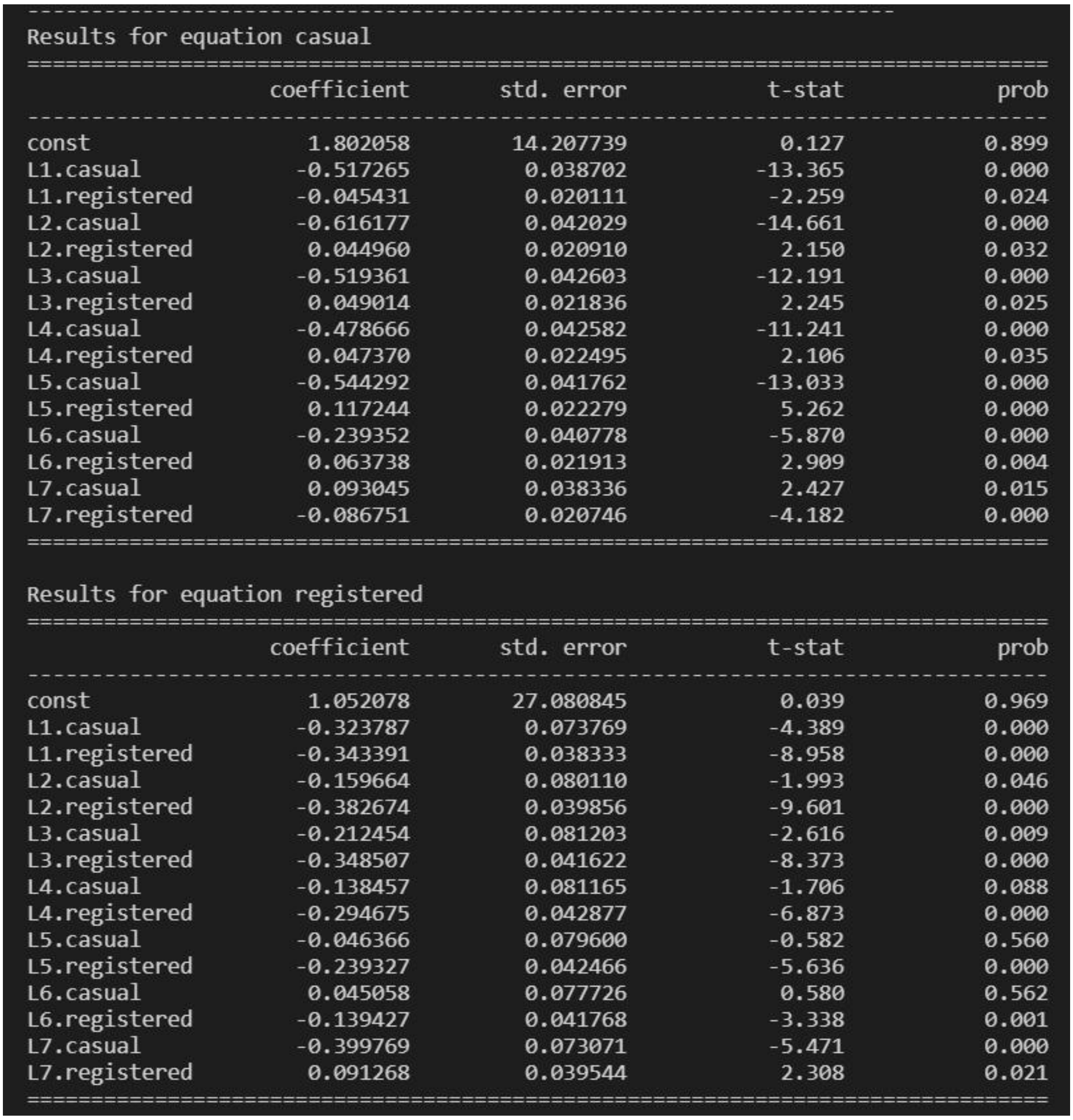
UCI machine learning dataset: Bike Sharing Dataset Data Set

Wikipedia: Vector autoregression

Multivariate time series forecasting

https://towardsdatascience.com/multivariate-time-series-forecasting-653372b3db36

Appendix:



```
-------------------------------------------------------------------
Results for equation casual
=================================================================================
                   coefficient       std. error           t-stat            prob
---------------------------------------------------------------------------------
const                 1.802058        14.207739            0.127           0.899
L1.casual            -0.517265         0.038702          -13.365           0.000
L1.registered        -0.045431         0.020111           -2.259           0.024
L2.casual            -0.616177         0.042029          -14.661           0.000
L2.registered         0.044960         0.020910            2.150           0.032
L3.casual            -0.519361         0.042603          -12.191           0.000
L3.registered         0.049014         0.021836            2.245           0.025
L4.casual            -0.478666         0.042582          -11.241           0.000
L4.registered         0.047370         0.022495            2.106           0.035
L5.casual            -0.544292         0.041762          -13.033           0.000
L5.registered         0.117244         0.022279            5.262           0.000
L6.casual            -0.239352         0.040778           -5.870           0.000
L6.registered         0.063738         0.021913            2.909           0.004
L7.casual             0.093045         0.038336            2.427           0.015
L7.registered        -0.086751         0.020746           -4.182           0.000
=================================================================================

Results for equation registered
=================================================================================
                   coefficient       std. error           t-stat            prob
---------------------------------------------------------------------------------
const                 1.052078        27.080845            0.039           0.969
L1.casual            -0.323787         0.073769           -4.389           0.000
L1.registered        -0.343391         0.038333           -8.958           0.000
L2.casual            -0.159664         0.080110           -1.993           0.046
L2.registered        -0.382674         0.039856           -9.601           0.000
L3.casual            -0.212454         0.081203           -2.616           0.009
L3.registered        -0.348507         0.041622           -8.373           0.000
L4.casual            -0.138457         0.081165           -1.706           0.088
L4.registered        -0.294675         0.042877           -6.873           0.000
L5.casual            -0.046366         0.079600           -0.582           0.560
L5.registered        -0.239327         0.042466           -5.636           0.000
L6.casual             0.045058         0.077726            0.580           0.562
L6.registered        -0.139427         0.041768           -3.338           0.001
L7.casual            -0.399769         0.073071           -5.471           0.000
L7.registered         0.091268         0.039544            2.308           0.021
=================================================================================
```

Figure 5