Analysis of the Sharing Bike Usage

Jing Ma, Xuan Wen, Zhaoyu Lu, Liyan Fan, Shangran Qiu

December 9, 2017

1 Abstract

In this project, our goal was to analyze the past data of bike sharing and predict the future usage of rental bikes. At the beginning, we transformed some of our variables in order to obtain non-skewed histograms and approximately linear scatter plot matrix. After that, GLS method was used to reduce the autocorrelation. Then, when we received an valid full linear model, variable selection was performed for a better model. In the end, the final ideal model contains only four predictors out of nine original variables we were given: seasons, weather, adjusted temperature and square root of wind speed. This final model appeared to be quite efficient, when it was used to predict the estimated count number in 2012.

2 Introduction

Nowadays, people can easily sell data to make a fortune, because the business secret is covered under these seemingly boring and trivial numbers. However, numbers won?t tell stories by themselves. The well-educated analysts, like us, have to interpret the meaning behind the puzzles, in order to get what we need. Therefore, the knowledge of using previous data to predict the future trend of some activities is a critical skill in the world of technology we are living in.

As Bike Sharing is gradually growing into a multi-billion-dollar business around the globe, people might begin to wonder what is the key to success behind this industry. Bike sharing not only provides people with an ?eco-friendly? way of transportation, it also gives a glimpse of what could potentially be part of the future generation?s way of life. In this project, our goal is to build a linear model that is able to predict the number of bikes being

rented on a daily basis, use the data from year 2011 and test the model on the data from year 2012.

3 Background

According to Hadi Fanaee-T and Joao Gama, the authors of ?Event labeling combining ensemble detectors and background knowledge?, the operation software will collect usage data during the rental periods [1]. Because the data we are facing is composed by environment status on daily basis, some variables are not as representative as others. Therefore, the most basic problem we faced during the analysis is to select the most useful variables from the redundant data we are given.

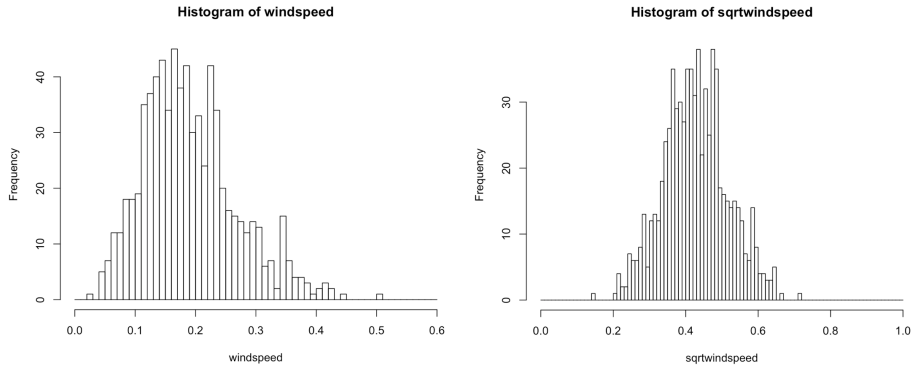
4 Modeling and Analysis

4.1 Goal and Data Splitting

In order to predict the number of bike rentals during the whole 2012 year, we built our linear models by performing various linear regressions based on the data of 2011 year. The data in 2011 has been split into the training part and validation part. Each part contains 50 % data of 2011 and is randomly sampled.

4.2 Data Transformation

Before performing linear regression, we firstly checked the distribution of each predictor. For the predictors with skewed distribution, transformations have been done on these variables to adjust these distributions as close to normal as possible. Among these variables, we applied Box-Cox transformation on the 'windspeed' variable and found the Box-Cox parameter lambda for 'windspeed' variable is 0.414. Thus we chose to round this number to 0.5 which corresponds to the square root transformation. The transformed variable is renamed 'sqrtwindspeed'.



(a) 1a (b) 1b

Figure 1

Then, from the scatter matrix, we found that the count versus month or adjtemp doesn't have a linear relationship, and not even monotonic. This is intuitive since the best time for biking is some month in the middle, and the best temperature should be neither too cold nor too hot. We dealt with this problem by setting the distance between the real month/adjtemp and the best-for-biking month/adjtemp as predictors. To be more specific, we adjusted variable ?month? by subtracting 7 (this is the month number with the highest monthly average bike rentals in the year 2011) and used the absolute value of the outcome as a new variable called ?absmonth?. Similarly, adjtemp was substracted by 0.665417, which is the temperature when count is at the maximum of 2011.

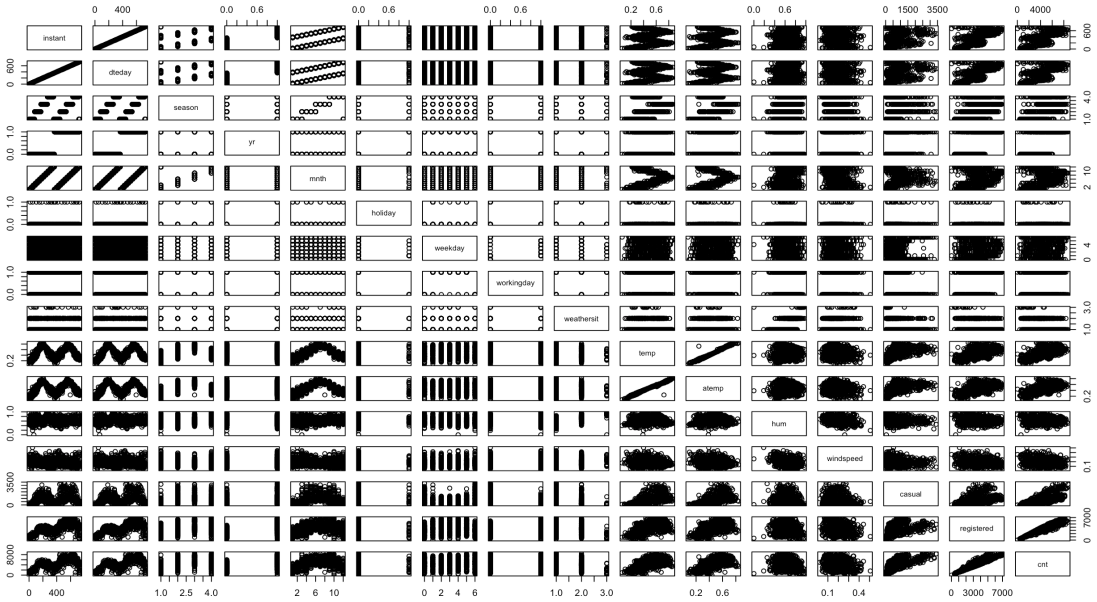
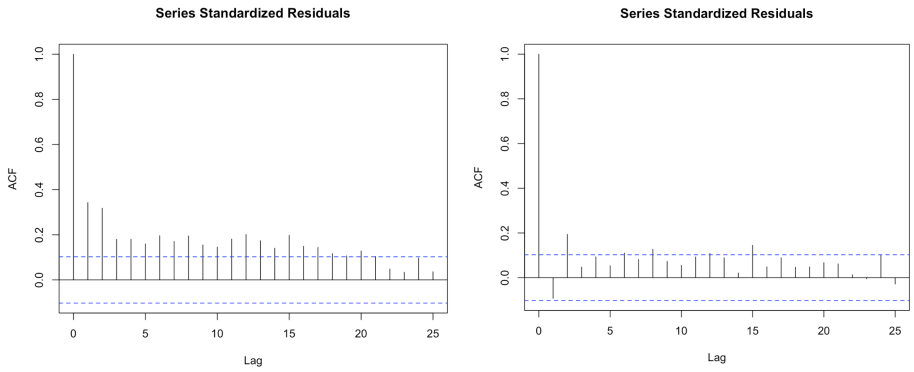


Figure 2: 1a

4.3 Reduction of Autocorrelation

Clearly these data are collected over time, and we could see from ACF that several lags of autocorrelation function of residuals exceed the normal two standard error cut-off value. Therefore, we decided to use the method based on generalized least squares with AR(1) errors on our 2011 data. And then we transformed our GLS model back to LS model with uncorrelated errors.



(a) 1a, before applying GLS (b) 1b, after applying GLS

Figure 3: autocorrelation function of the GLS residuals from model **

4.4 Variable Selection

After reducing the inference of autocorrelation, we obtained a valid full model as follows:

$$\begin{aligned}\text{count} = \beta_0 + \beta_1\text{season} + \beta_2\text{absmonth} + \beta_3\text{holiday} + \beta_4\text{weekday} + \beta_5\text{workingday} \\ + \beta_6\text{weather} + \beta_7\text{absadjtemp} + \beta_8\text{humidity} + \beta_9\text{sqrtwindspeed}\end{aligned}$$

We searched over all possible subsets using our training dataset to determine the best predictor subsets for different number of predictors. And these selected subsets are then tested on the validation set, showing that the model with 4 predictors gives a most satisfying result in terms of the information criteria without overfitting:

$$\text{count} = \beta_0 + \beta_1\text{season} + \beta_6\text{weather} + \beta_7\text{absadjtemp} + \beta_9\text{sqrtwindspeed}$$

	adjr2	AIC	AICc	BIC
1	0.5386858	2450.909	2450.942	2460.098
2	0.6151575	2434.723	2434.778	2448.506
3	0.6986540	2358.220	2358.302	2376.597
4	0.7287369	2352.660	2352.776	2375.632
5	0.7368152	2354.631	2354.785	2382.197
6	0.7439687	2355.959	2356.157	2388.119
7	0.7474445	2337.180	2337.429	2373.935
8	0.7472181	2335.395	2335.700	2376.745

Figure 4: information criteria for the best subset of each size

The following figure shows the diagnostic plots of the model with 4 predictors. We can see that the residuals are pretty random, without showing any patterns.

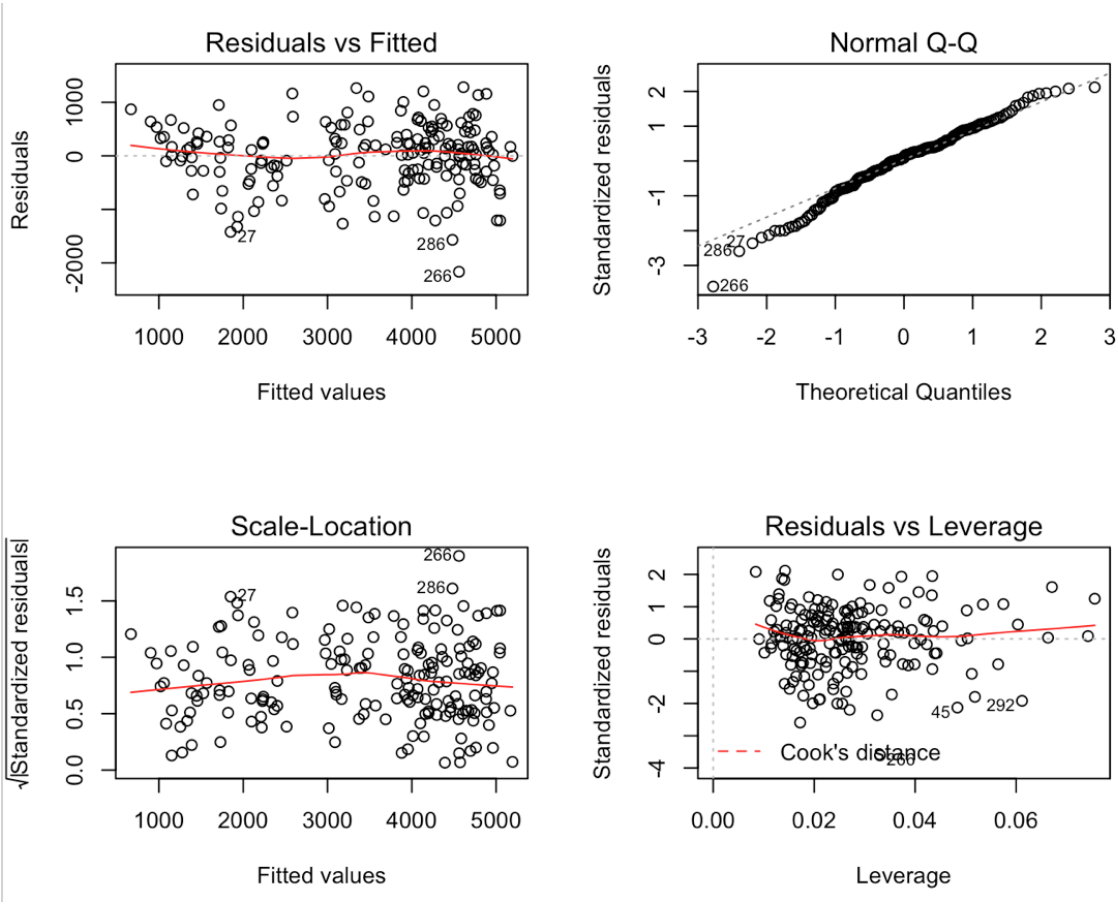


Figure 5: diagnostic plots for the model with four predictors

5 Prediction

To visualize the accuracy of our prediction, we plotted both the real and predicted bike rental data against dates over the year. In above figure, the significant overlap between real 2012 data and predicted 2012 data illustrates the capability of our linear model, trained on 2011 data, in predicting next year?s bike rental amount. Both the real data and predicted data consistently show the upward trend of bike rental in the spring and summer seasons and the downward trend in the fall and winter seasons. However, through above figure,

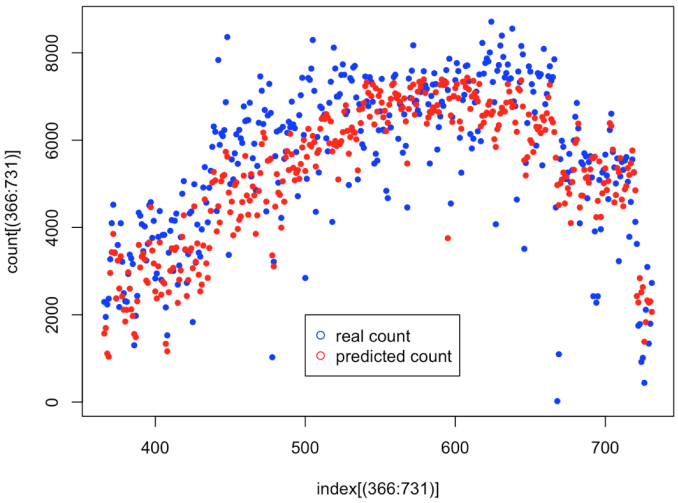


Figure 6: the comparison between real bike rentals and predicted results in 2012

we can also see the real data is slightly higher than the predicted data over the year. To quantify this difference, we also plotted real 2012 data against predicted 2012 data. In the ideal case, where the model can always predict accurately, the plot of real data against predicted data should have a slope of 1, with a zero intercept. In the following figure, the slope of our linear model is 0.93 which is slightly lower than 1. And there is an intercept of about 733, which is an indicator of the development of bike sharing industry.

Thus, our prediction is good enough to show some main behavior of the data in year 2012. The slight bias is the result of us having only one year of data.

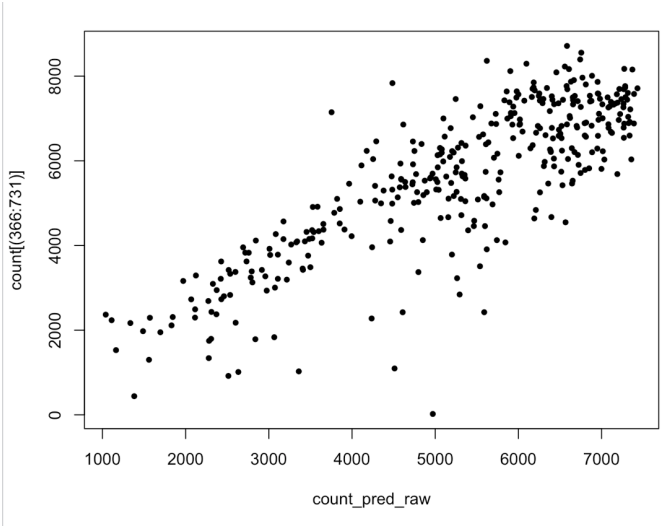


Figure 7: real bike rentals versus predicted results in 2012

6 Discussion

To revisit our initial goal of this project, which is to predict the number of bike rentals on a daily basis by building a linear model from the past data, we as a group think that we have mostly achieved our goal. Although our models are not perfect, we are still able to successfully predict most of the behavior of future bike usage, i.e. in the year 2012. Given that bike sharing is a newly formed but fast growing business, there still remain some uncertain factors in the future. However, due to the growing trend of bike-sharing business, it?s really hard to predict data of a brand new year, with only one year of data. The only solution is to have more data, e.g. in year 2009, 2010, etc. Also, from Figure of autocorrelation function, the errors are not just simply AR(1) since there are more than one lags exceeding the cut-off value. So simply applying GLS with AR(1) errors did not eliminate the autocorrelations entirely.

A ?perfect model? does not exist in the real world, because no one can be completely certain about what will happen in the future. But we always want to try our best to make the most likely predictions about the future by using the past experience. Nonetheless, from a business prospect of view, any dramatic event will change our model significantly . For example, though the bike-sharing business is now blooming around the world, once it reaches to a peak in the market, our current model will no longer be valid and we should

change our model to best fit the situation at that time.

References

[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.

Appendices

```
### to determine the transformation for windspeed
library(MASS)
Box = boxcox(windspeed~1,lambda=seq(-1,1,0.05))
Cox = data.frame(Box$x,Box$y)
Cox2 = Cox[with(Cox,order(-Cox$Box.y)),]
lambda = Cox2[1,'Box.x']
print(lambda)
T_box = (windspeed^lambda-1)/lambda
hist(T_box)

### to split all data into train (50% of 2011), validation (the other 50%
### of 2011), and test sets (2012)
data_2011 = data[(1:365),]
data_2012 = data[(366:731),]
set.seed(1)
train = sample(1:nrow(data_2011), nrow(data_2011)/2)
train_2011 = data_2011[c(train),]
val = (-train)
val_2011 = data_2011[val,]


### to select variables from several predictor subsets using validation
### data set
om1<- lm(count~ X[,8],data=val_2011)
om2<- lm(count~ X[,8]+X[,7],data=val_2011)
om3<- lm(count~ X[,8]+X[,7]+X[,2],data=val_2011)
om4<- lm(count~ X[,8]+X[,7]+X[,2]+X[,10],data=val_2011)
om5<- lm(count~ X[,8]+X[,7]+X[,2]+X[,10]+X[,1],data=val_2011)
om6<- lm(count~ X[,8]+X[,7]+X[,2]+X[,10]+X[,1]+X[,4],data=val_2011)
om7<- lm(count~ X[,8]+X[,7]+X[,2]+X[,10]+X[,1]+X[,4]+X[,9],data=val_2011)
om8<- lm(count~ X[,8]+X[,7]+X[,2]+X[,10]+X[,1]+X[,4]+X[,9]+X[,5],data=val_2011)
```

```
n <- nrow(data)
lm.stat <- function(l) {
  npar = length(coef(l)) + 1
  AIC = extractAIC(l, k = 2)[2]
  BIC = extractAIC(l, k = log(n))[2]
  return( c(rs$adjr2[npar - 2], #adjr2
            AIC,
            AIC + 2 * npar * (npar + 1) / (n - npar + 1), # AICc
            BIC))
}
matrix(unlist(lapply(list(om1, om2, om3, om4, om5, om6, om7,om8),lm.stat)),
byrow = TRUE, ncol = 4, dimnames = list(1:8, c("adjr2", "AIC", "AICc", "BIC")))

### to generate the diagnostic plots using chosen predictor subsets
plot(om4)

### to give a comparison of real bike rentals and predicted results in 2012
plot(index[(366:731)],count[(366:731)],col='blue',pch=20)
points(index[(366:731)],count_pred_raw,col='red',pch=20)
legend(500,2000, legend=c('real count','predicted count'), col=c('blue','red'),
lty=0, cex=1, pch=1)

plot(count_pred_raw,count[(366:731)],pch=20)
lm(count[(366:731)]~count_pred_raw)
```