Catalogue

1.Research background and purpose

The concept of sharing has long existed. In traditional society, borrowing books or sharing a piece of information between friends, including borrowing things between neighbors, is a form of sharing.

The term shared economy was first proposed by Marcus Felson (Marcus Felson), professor of sociology at Texas State University, and Joel.Spaeth, professor of sociology at the University of Illinois, in a paper (Community Structureand Collaborative Consumption:ARoutine Activity Approach) published in 1978.

After 2000, with the advent of the Internet web2.0 era, a variety of online virtual communities, BBS and forums began to appear, and users began to express their views and share information with strangers in cyberspace. Around 2010, with the emergence of a series of physical sharing platforms such as Uber and Airbnb, sharing began to move from pure free sharing and information sharing to a "sharing economy" with the main purpose of getting a certain reward, based on strangers and the temporary transfer of the right to the use of goods.

In recent years, as governments have begun to take active measures to reduce carbon emissions, the bike-sharing market has made remarkable development and become one of the hottest directions in the sharing economy.

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 is able to easily rent a bike from a particular position and return back at another position. Currently, there are about over 500 bike-sharing programs around the world which is composed of over 500 thousands bicycles. Today, there exists great interest in these systems due to their important role in traffic, environmental and health issues.

Apart from interesting real world applications of bike sharing systems, the characteristics of data being generated by these systems make them attractive for the research. Opposed to other transport services such as bus or subway, the duration of travel, departure and arrival position is explicitly recorded in these systems. This feature turns bike sharing system into a virtual sensor network that can be used 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.

But not all shared bikes can be successful. we see more and more companies getting involved in the field of shared bikes, as well as some failed attempts like OFO.

Compared with traditional travel services, shared travel emphasizes the short-term purchase of use rights rather than ownership as needed to achieve "on-demand travel". The emergence of shared bikes has effectively solved the problem of "the last kilometer", so it is of great significance to establish optimal scheduling to improve the service efficiency of shared bikes and ensure the maximum utilization rate of shared bikes. This is also a very important aspect that bike-sharing companies can maintain bike usage and profitability. Because if people have demand for bicycles and there are no bikes they want to ride nearby, then people will definitely choose other means of transportation or other brands of bikes; on the contrary, if people have little

demand for shared bikes but have a large number of bikes, there will be "placement costs", and there may even be problems such as maintenance costs caused by damage to bikes, or occupying space to affect urban traffic. Therefore, releasing more cars when the demand is high and less when the demand is low can not only meet the needs of customers and improve the profit level, but also has a certain social significance, which is beneficial to all parties.

In the scheduling of shared bikes, it is certain that the same scheduling frequency is not maintained in all time periods and in all cases, which varies with people's demand and frequency of use of shared bikes. Some studies have also pointed out that "bike-sharing enterprises still have shortcomings in cycling preferences, user personal information protection, vehicle delivery location distribution and bicycle information level" (Jiang Yujie, Zhang Bin, 2020). For example, during the peak period of every day, there is a high demand for shared bikes, so we need to increase the scheduling frequency to meet people's needs, while the low peak time can reduce scheduling to save costs. For example, some weather is sunny and cool, bicycles are used frequently, so it is necessary to speed up dispatching, while sometimes the weather is cold and people seldom choose bicycles as a means of travel, they can reduce dispatching.

This paper explores the factors that affect the frequency of shared bikes, hoping to explore the relationship between the usage of shared bikes and different factors. Thus, it is predicted in which cases the scheduling of shared bikes should be accelerated to ensure supply, and in which cases the scheduling of shared bikes should be reduced to control costs, meet people's demand for shared bikes, and improve their own profits at the same time.

Bike sharing began to enter the Chinese market about five years ago, but it appeared earlier in the United States. our data is between 2011 and 2012, the data of shared bike usage in Washington, USA.

2.Data introduction

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. The core data set is related to the two-year historical log corresponding to years 2011 and 2012 from Capital Bikeshare system, Washington D.C., USA which is publicly available in http://capitalbikeshare.com/system-data. We aggregated the data on two hourly and daily basis and then extracted and added the corresponding weather and seasonal information. Weather information are extracted from http://www.freemeteo.com.

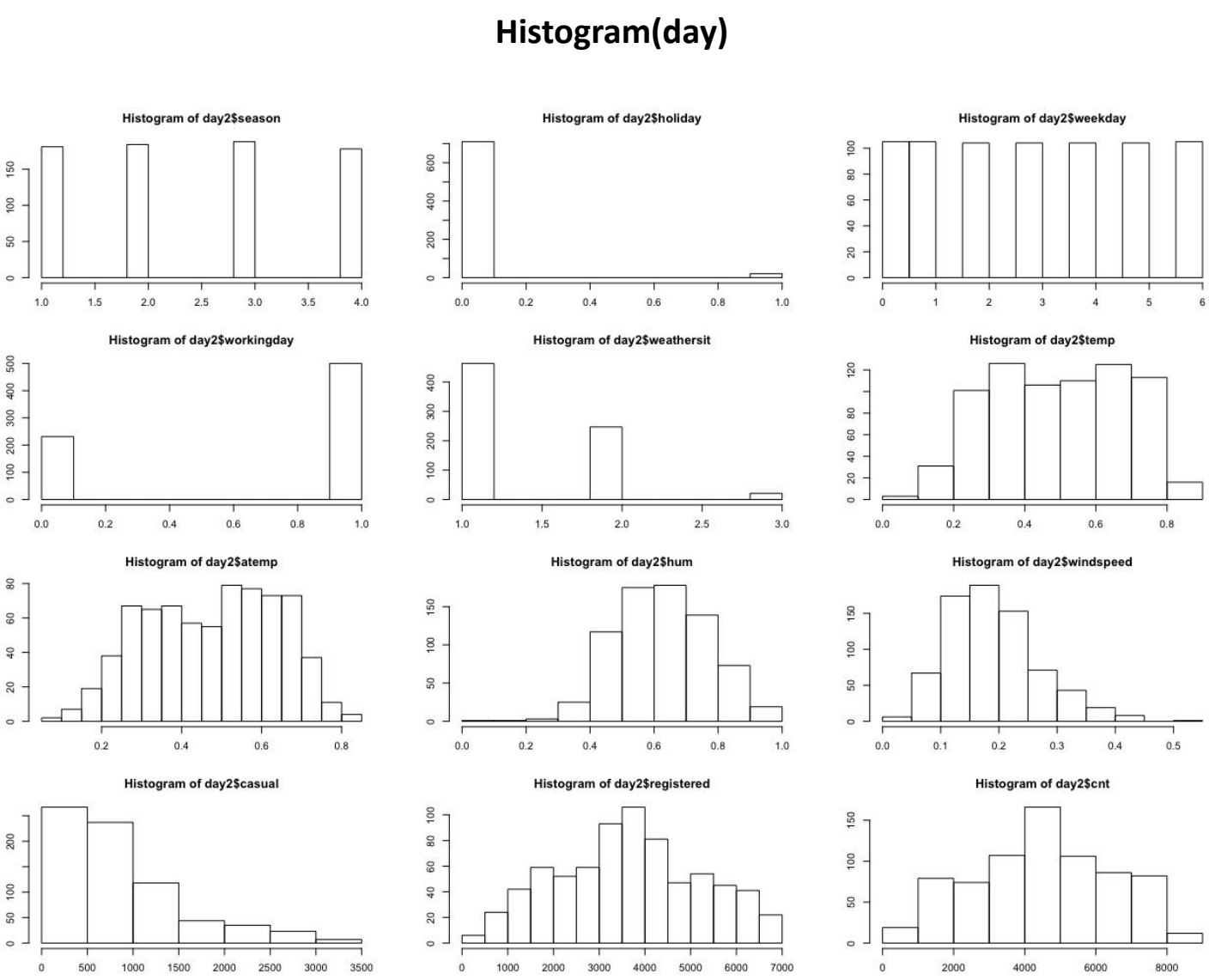
The data is divided into two parts: "day" and "hour". The first part statistics the external environment and the number of bicycle users between different days, and the second part makes detailed statistics on the environment and the number of bicycles in 24 hours every day. The number of users is divided into three categories: "registered users", " casual users" and "total rental bikes".

Interpretation of variable names

instant: record index
dteday : date
season : season (1:springer, 2:summer, 3:fall, 4:winter)
yr : year (0: 2011, 1:2012)
mnth : month (1 to 12)
hr : hour (0 to 23)
holiday : weather day is holiday or not (extracted from
http://dchr.dc.gov/page/holiday-schedule)
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. The values are divided to 41 (max)
atemp: Normalized feeling temperature in Celsius. The values are divided to 50 (max)
hum: Normalized humidity. The values are divided to 100 (max)
windspeed: Normalized wind speed. The values are divided to 67 (max)
casual: count of casual users
registered: count of registered users
cnt: count of total rental bikes including both casual and registered

3.Research process and analysis

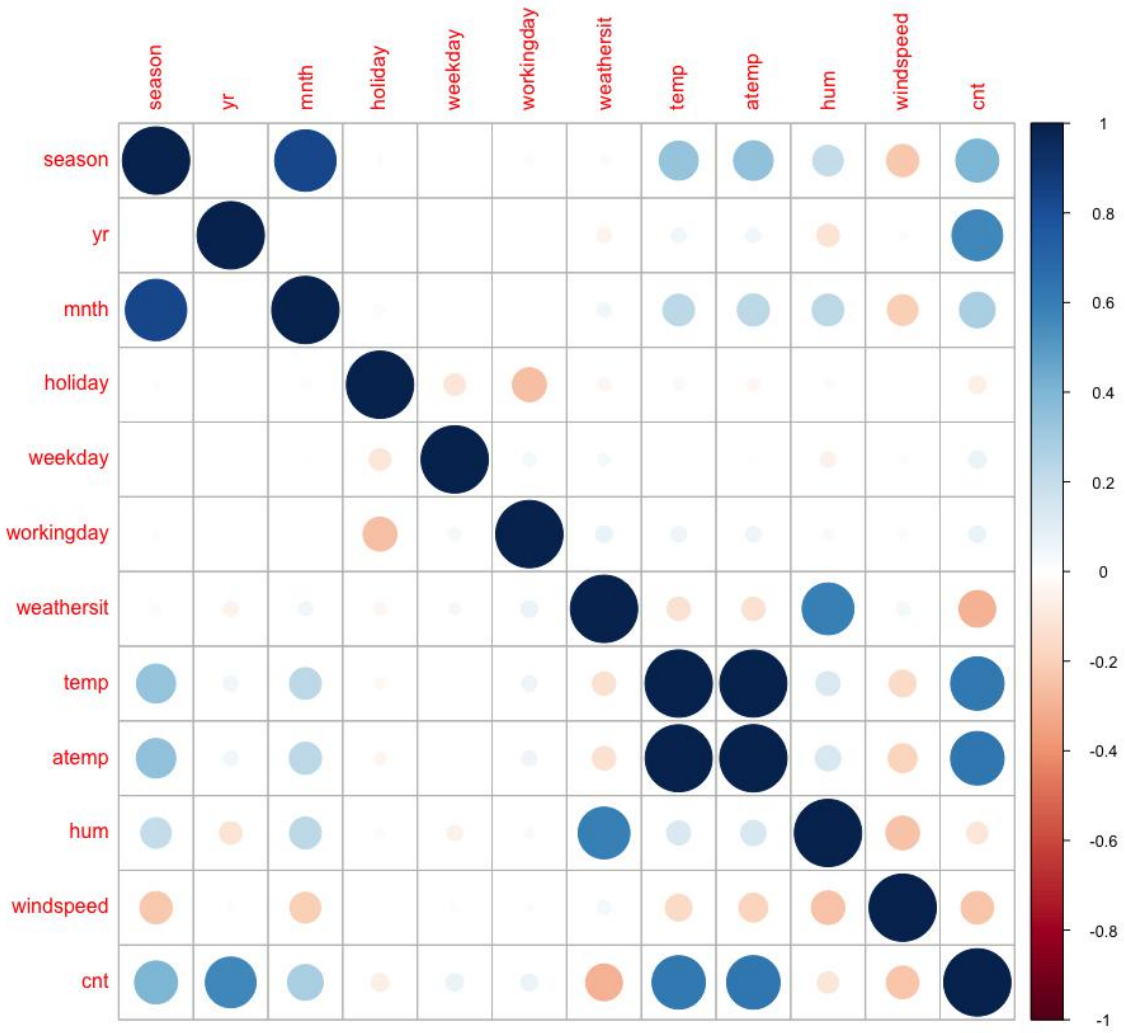
3.1Data Exploration



(Figure 1)

In the Histogram above, we can see the distribution of variables. It is worth noting that the distribution of "registered" and the "cnt" are basically in line with the normal distribution, but the number of "casual" is concentrated in a relatively low position.

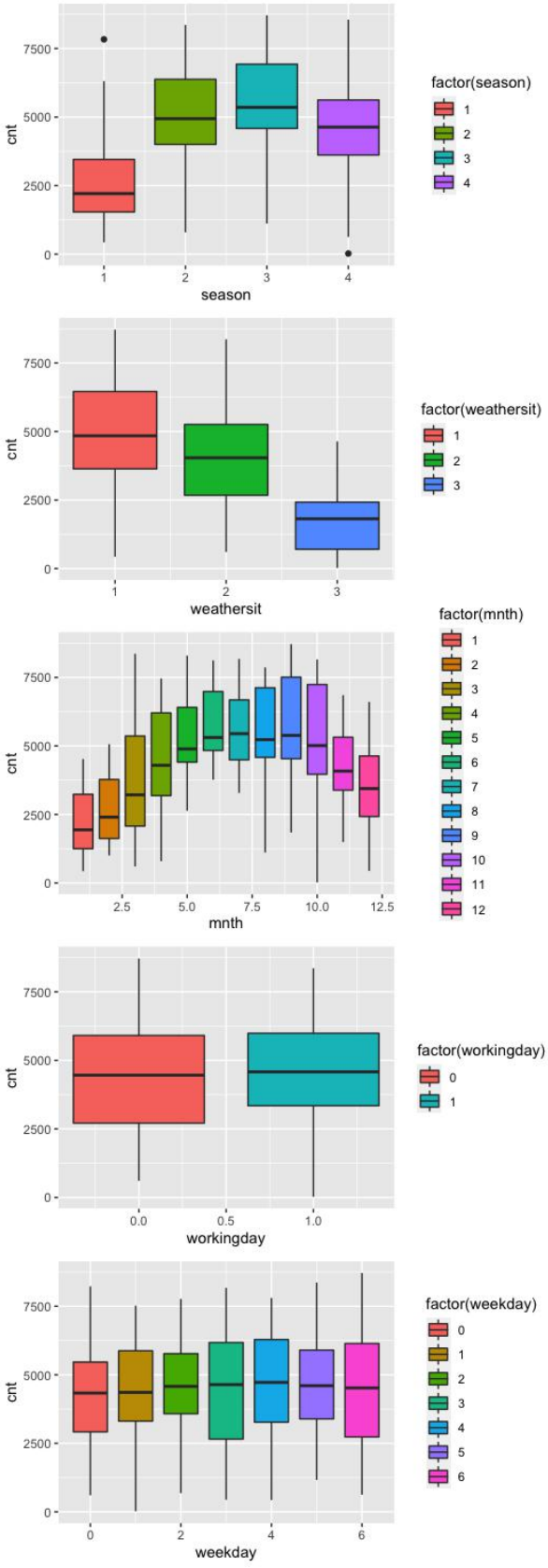
Correlation(day)



(Figure 2)

In the correlation analysis, we see that the correlation between the “cnt” and seasons, weather conditions and temperature is relatively high.

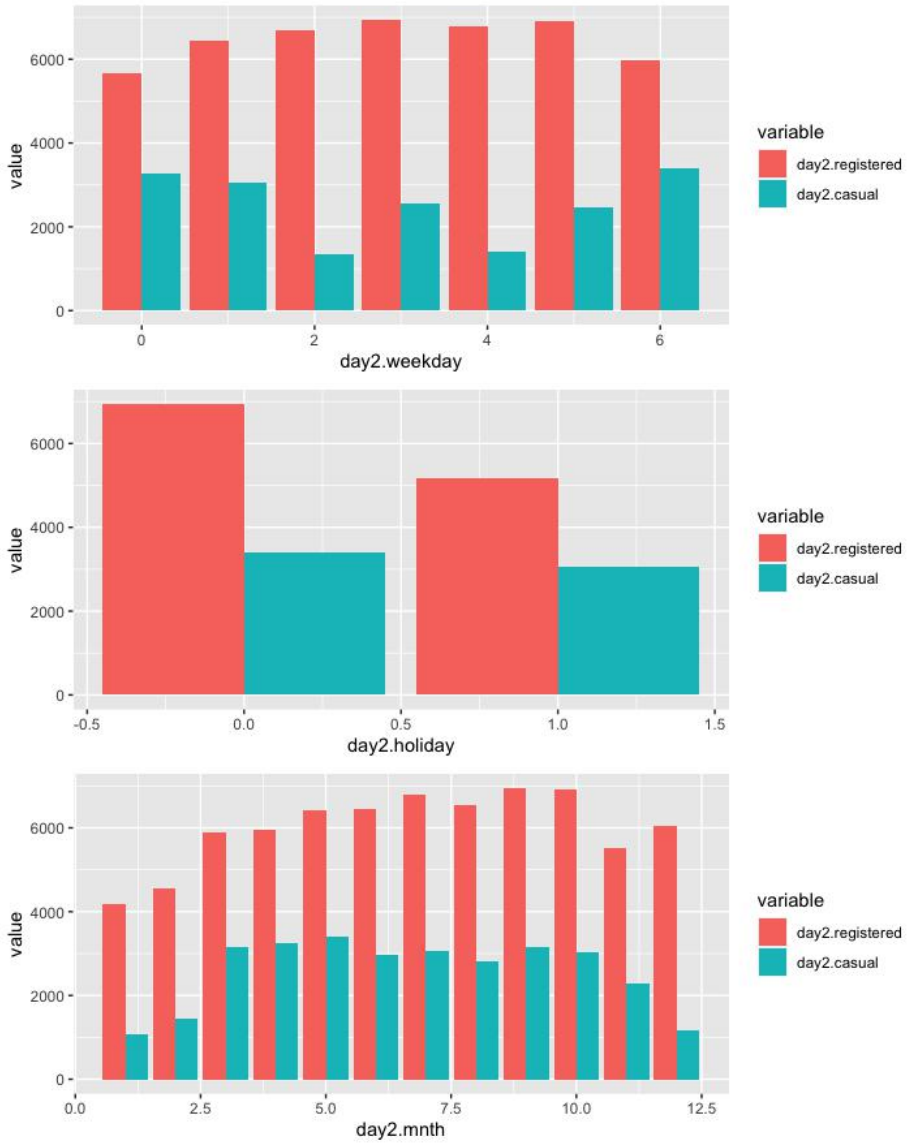
The Box Plot(day)



(Figure3)

From the box chart, we can see that the trend and volatility of the dependent variable cnt can be observed by the box chart analysis of the dependent variable under different independent variables. From the "weathersit"-"cnt" box chart, it can be observed that the mean value of cnt decreases obviously with the weather (1,2,3), indicating that the weather greatly affects the usage. The box chart "season"-"cnt" and "mnth"-"cnt" consistently shows the trend of bicycle usage with time: it rises in summer and autumn, decreases in spring and winter, and fluctuates greatly in late spring and late autumn. In addition, we can also see from the box chart "weekday"-"cnt" that usage is more volatile on Wednesdays and Saturdays.

The Compare Histogram(day)



(Figure4)

From the Compare Histogram, we can see that no matter it is a working day or a rest day, or at any time of the year, the frequency of use of members is much higher than that of non-members. The total number of ”registered” is about four times that of “casual”.

In the first compare histogram above, members use bicycles more in the middle of the week and less on weekends. On the contrary, there are fewer non-members in the middle of the week and more on weekends.

In the second compare histogram above, it shows that the use of bicycles increases during the holidays, both for members and non-members.

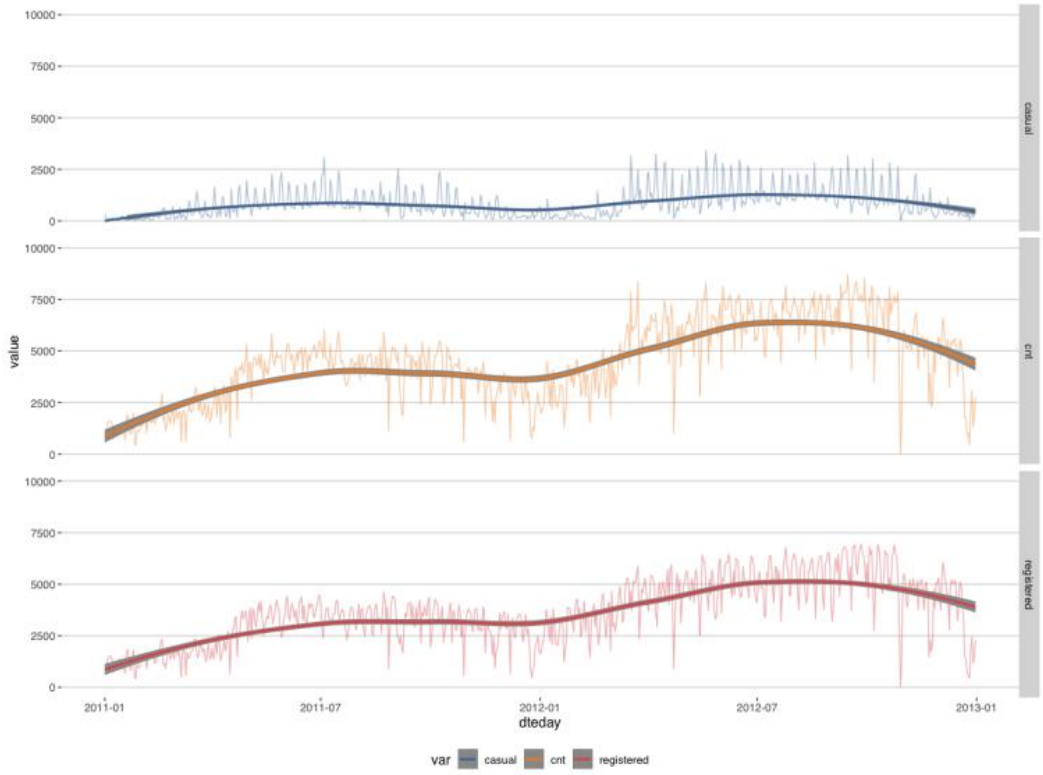
In the third compare histogram above, it shows that both members and non-members have a seasonal trend in the use of bicycles. At the height of summer, there will be a slight decline.

Time Series Plot(day)

cnt,registered,casual

Legend: blue=casual orange=cnt red=registered

X=day Y=value

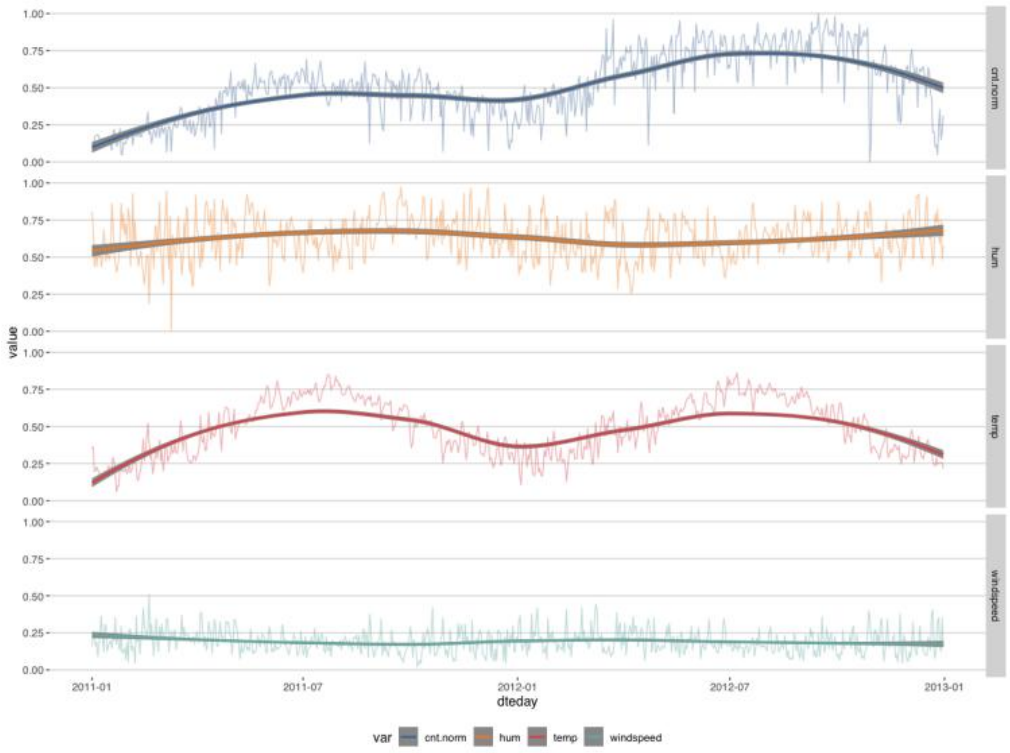


(Figure5)

Different Variable : temp, hum, windspeed, cnt

Legend: dark blue=cnt.norm orange=hum red=temp light blue=windspeed

X=day Y=value

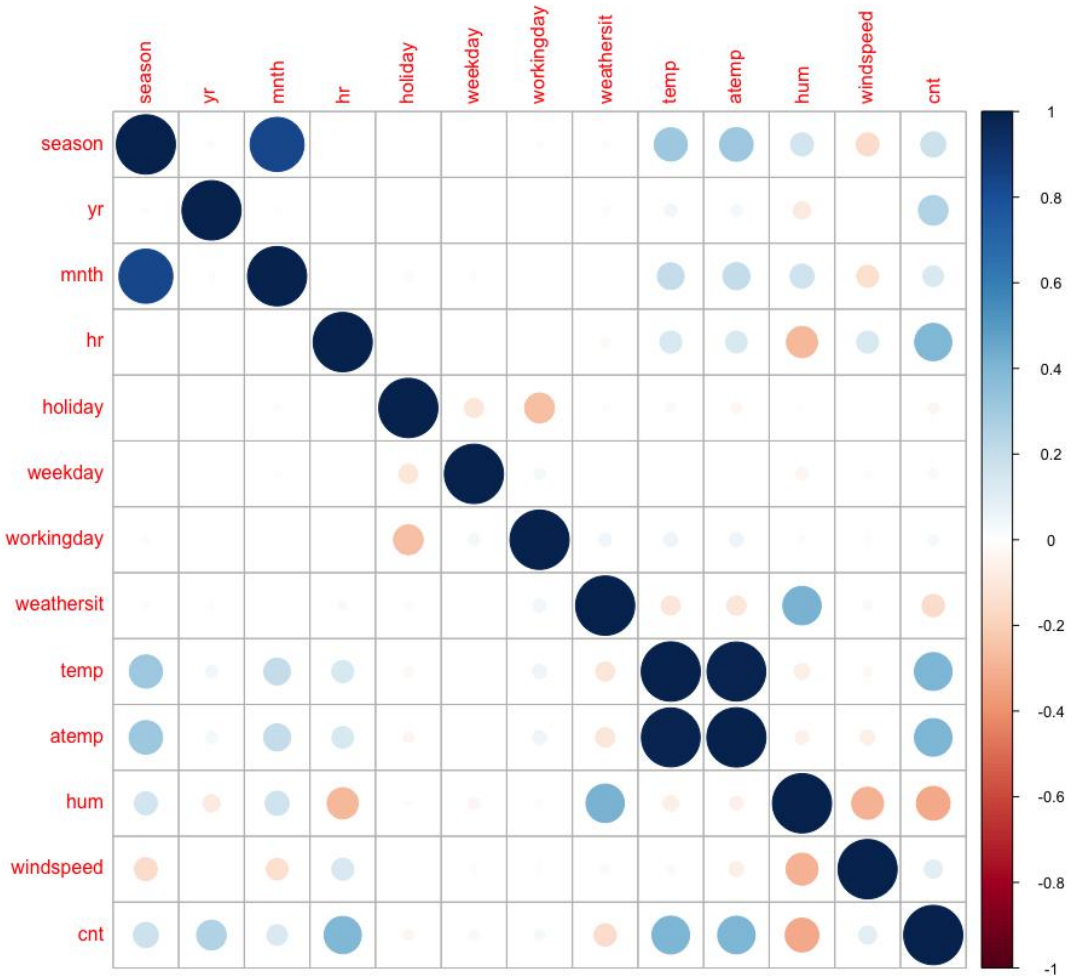


(Figure6)

The time series diagram gives us an intuitive understanding of the overall trend of bicycle usage, as well as the detailed trends of the two groups (members and non-members). From figure 5, we can see that the bicycle usage of both non-members and members shows a seasonal trend, in which the volatility of non-member (casual) is more stable than that of members. In terms of member usage (registered), in addition to the seasonal trend, the number of members continues to rise.

We can also observe the trend relationship between dependent variables and other variables in time series diagrams. From figure 6, we can see that the trend change of bicycle usage is consistent with the temperature change. The sudden and sharp decline of on the another hand, from the dependent variable cnt, we can speculate that it is extremely vulnerable to unexpected factors such as extreme weather, holidays and other established factors.

Correlation(hour)

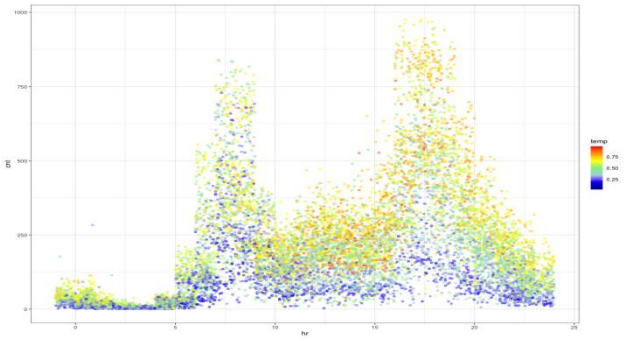


(Figure7)

From the correlation matrix diagram above, we can see that there is a correlation between the number of cyclists and many factors in the "Hour" data set, such as season, hour, temperature, humidity, etc., but there are slight differences in "cnt", "casual" and "registered".

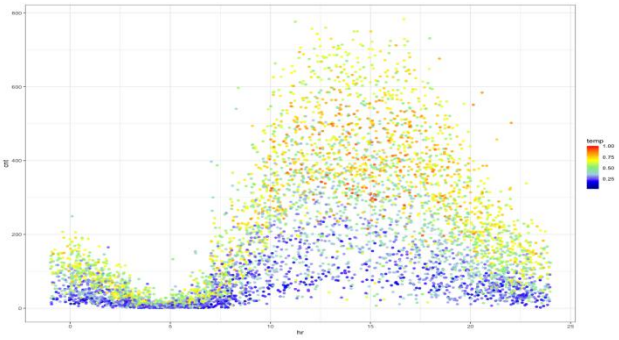
Hour Distribution

working day = 1



(Figure8)

working day = 0



(Figure9)

From the figure above, "the number of hours of cnt distributed in the day", we can see that there are two obvious peaks on weekdays, 8: 00 in the morning and 6: 00 in the evening, which shows that commuters are mainly used in daily at work. But there is no obvious peak in non-working days. They are used more frequently in the afternoon during the non-working days. We can speculate that people take more rest on non-working days, get up later and go out less in the morning. The main outing activities are concentrated in the afternoon.

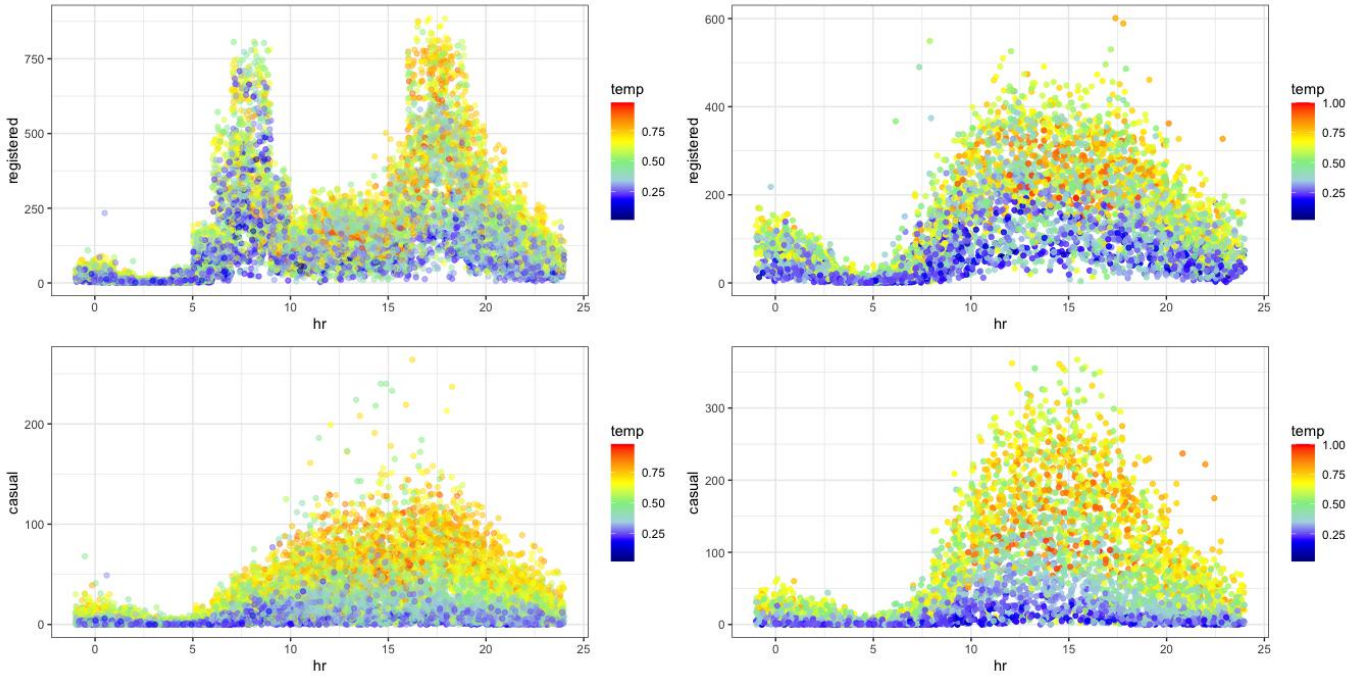
At the same time, the color reflects the real-time temperature, and we can see that there are more cyclists when the temperature is high than when the temperature is low, but when the temperature is particularly high, the number of cyclists decreases, especially on non-working days.

And we can see that the frequency of use of non-members is more obvious than that of members affected by temperature. Because members use it as a fixed means of transportation (perhaps out of habit or membership fee), they will basically insist on using it as long as the weather is not particularly extreme. On the other hand, non-members have not paid the "sunk cost" before, so once the temperature makes them feel unsuitable for cycling, they will not hesitate to choose other means of transportation as an alternative. As a result, non-members show greater temperature sensitivity.

And through figure 10, we can see that when we look at the distribution characteristics of the use of “registered” alone, it is basically the same as “cnt”. But when we look at ”casual”, we find that there is no obvious peak in both working days and non-working days, which are concentrated in the afternoon. We can understand that most of the non-members use it not as a means of commuting to and from work, but for occasional travel purposes.

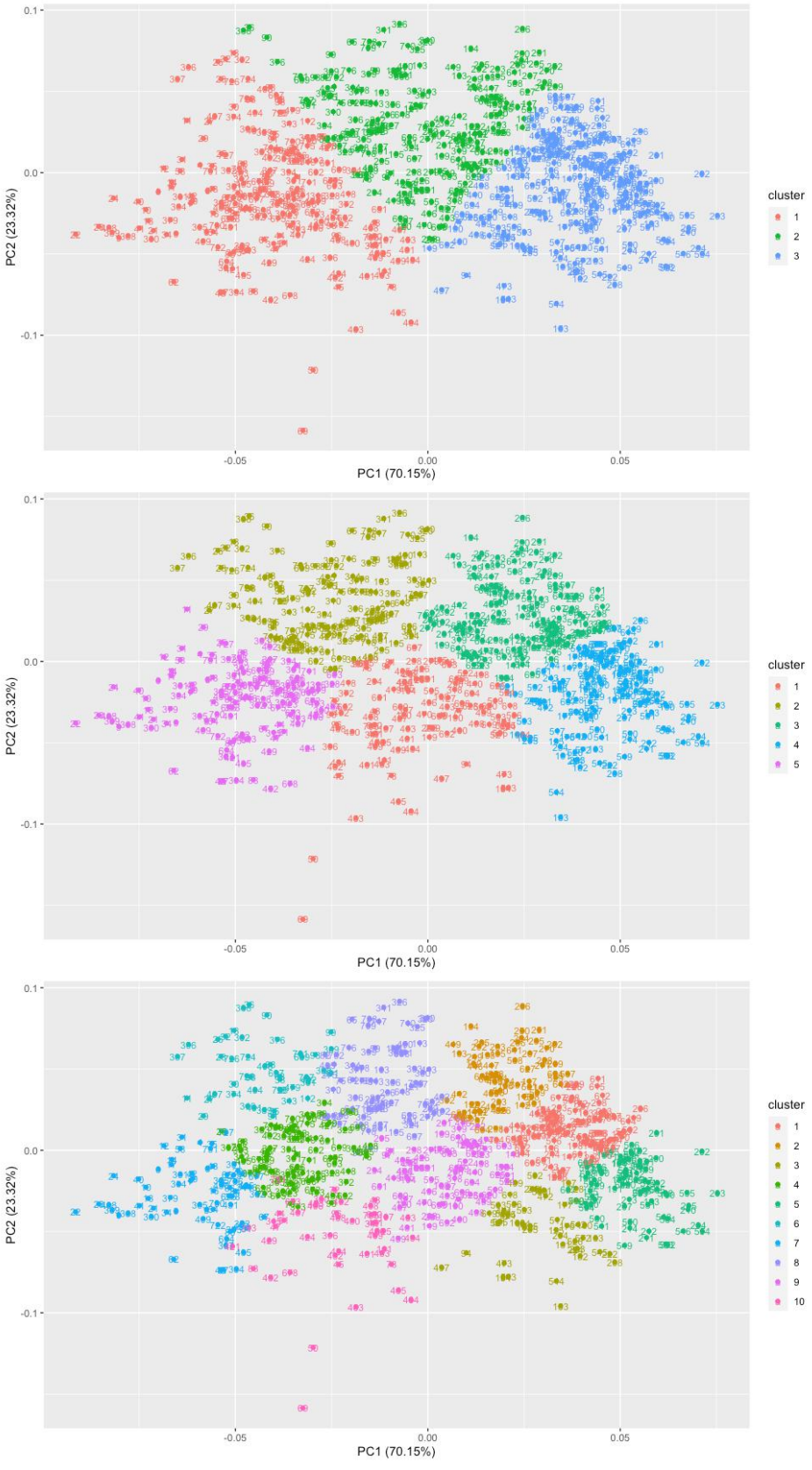
working day = 1

working day = 0



(Figure10)

Clustering(day)



(Figure11)

Using environmental factors (feeling temperature, humidity and wind speed) as variables to do cluster analysis of "day" to verify whether the above variables can reflect the change of bicycle usage. The results of manova analysis show that there are differences in mean vectors of "day" whether they are divided into 3 categories, 5 categories or 10 categories. The types of weather are varied, but only when the "day" is divided into three categories, the mean value of cnt calculated by classification (3218, 4292, 5846 respectively) will be significantly different. Therefore, cluster analysis shows that the above three variables do reflect the impact of environmental factors on cnt, and it is reasonable for weathersit to classify the weather into three categories of data.

3.2Regression analysis

Multiple collinearity

There is a high correlation between the variables temp and atemp, and they obviously share a common time trend. When temp and atemp exist at the same time, the result of calculating the condition number (Kappa) shows that the variance expansion factor is more than 100 (VIF=265), and there is serious multicollinearity. After excluding the variable temp, the VIF value is 1.86, so temp is excluded in the following regression analysis.

We carried out regression analysis on the two data sets of "day" and "hour" respectively.

Day

Glm(day):

The results of linear regression show that it is not suitable for this packet, so we show generalized linear regression (glm) here.

Model	X	Y	Goodness	MSE
Model1	mnth, atemp, hum, windspeed	cnt	0.486	-
Model2	as.factor(mnth), as.factor(weathersit), atemp, hum, windspeed	cnt	0.561	-
Model3	as.factor(season), as.factor(mnth), as.factor(weathersit), atemp, hum, windspeed	cnt	0.581	-
Model4	as.factor(season), as.factor(yr), as.factor(holiday), as.factor(weekday), as.factor(workingday), as.factor(mnth), as.factor(weathersit), atemp, hum, windspeed	cnt	0.848	639885.6
Model5	as.factor(yr)+as.factor(weekday)+as.factor(workingday)+as.factor(mnth)+as.factor(weathersit)+atemp+hum+windspeed	cnt	0.829	700190.3

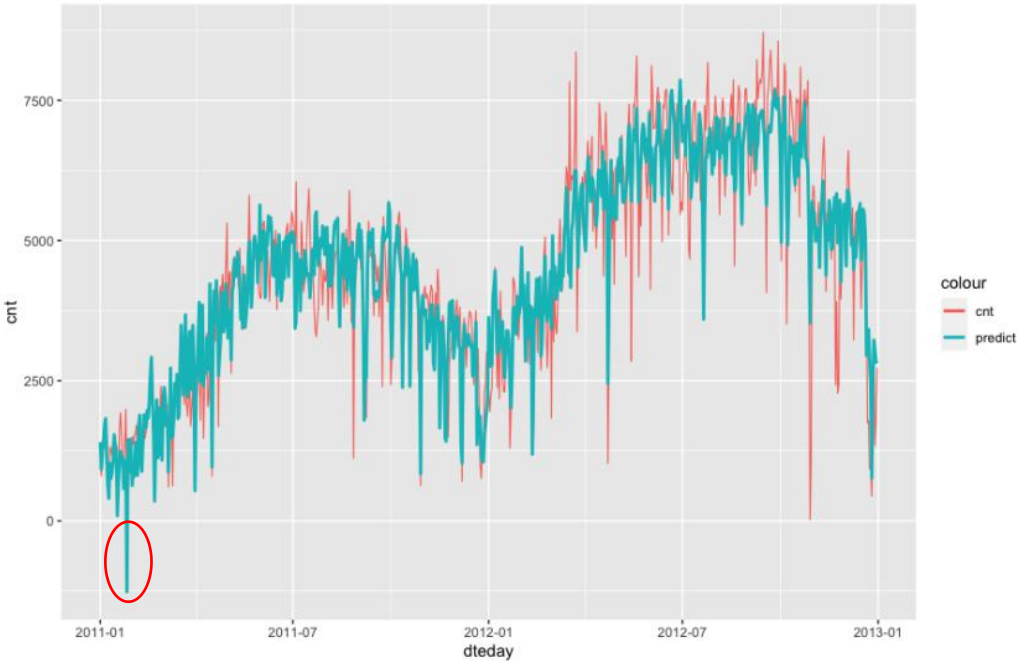
（Table1）

From the above table, we can see that when using glm for regression, the established model 4 is the most suitable model, with the highest degree of Goodness(0.848) and small MSE(639885.6).

The following is Model 4：

Coefficients:	Estimate	Std.Error	t value	Pr(>|t|)	
(Intercept)	1420	238.1	5.96	3.90E-09	***
as.factor(season)2	876.4	179.9	4.87	1.40E-06	***
as.factor(season)3	848.8	213.5	3.98	7.70E-05	***
as.factor(season)4	1578.8	181.4	8.7	<2E-16	***
as.factor(yr)1	2027.9	58.2	34.83	<2E-16	***
as.factor(holiday)1	-561.2	180.2	-3.11	1.92E-03	**
as.factor(weekday)1	209.7	109.7	1.91	5.64E-02	.
as.factor(weekday)2	315.1	107.3	2.94	3.44E-03	**
as.factor(weekday)3	392.5	107.6	3.65	2.80E-04	***
as.factor(weekday)4	394.1	107.7	3.66	2.70E-04	***
as.factor(weekday)5	453.9	107.4	4.23	2.70E-05	***
as.factor(weekday)6	444.3	106.8	4.16	3.60E-05	***
as.factor(mnth)2	137.4	144.1	0.95	3.41E-01	
as.factor(mnth)3	577.6	164.9	3.5	4.90E-04	***
as.factor(mnth)4	501	246.9	2.03	4.28E-02	*
as.factor(mnth)5	839.6	263.3	3.19	1.49E-03	**
as.factor(mnth)6	660.7	273.9	2.41	1.61E-02	*
as.factor(mnth)7	177.8	306.1	0.58	5.62E-01	
as.factor(mnth)8	607.5	293.2	2.07	3.86E-02	*
as.factor(mnth)9	1112.1	260.4	4.27	2.20E-05	***
as.factor(mnth)10	566.6	241	2.35	1.90E-02	*
as.factor(mnth)11	-104.1	231.2	-0.45	6.53E-01	
as.factor(mnth)12	-84.5	182.6	-0.46	6.44E-01	
as.factor(weathersit)2	-467.1	77.2	-6.05	2.40E-09	***
as.factor(weathersit)3	-1955.8	197.4	-9.91	<2E-16	***
atemp	4639.4	431.5	10.75	<2E-16	***
hum	-1516.8	293	-5.18	2.90E-07	***
windspeed	-2647.1	406.4	-6.51	1.40E-10	***

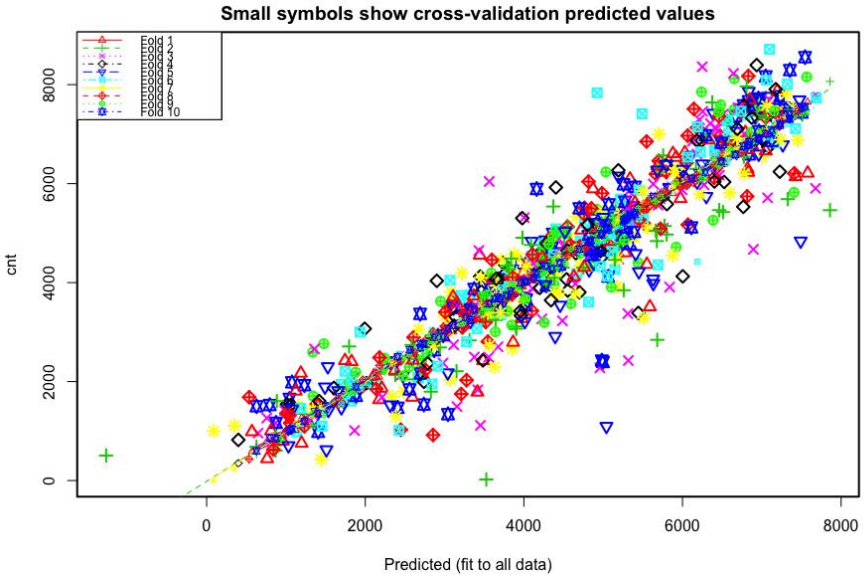
（Table2）



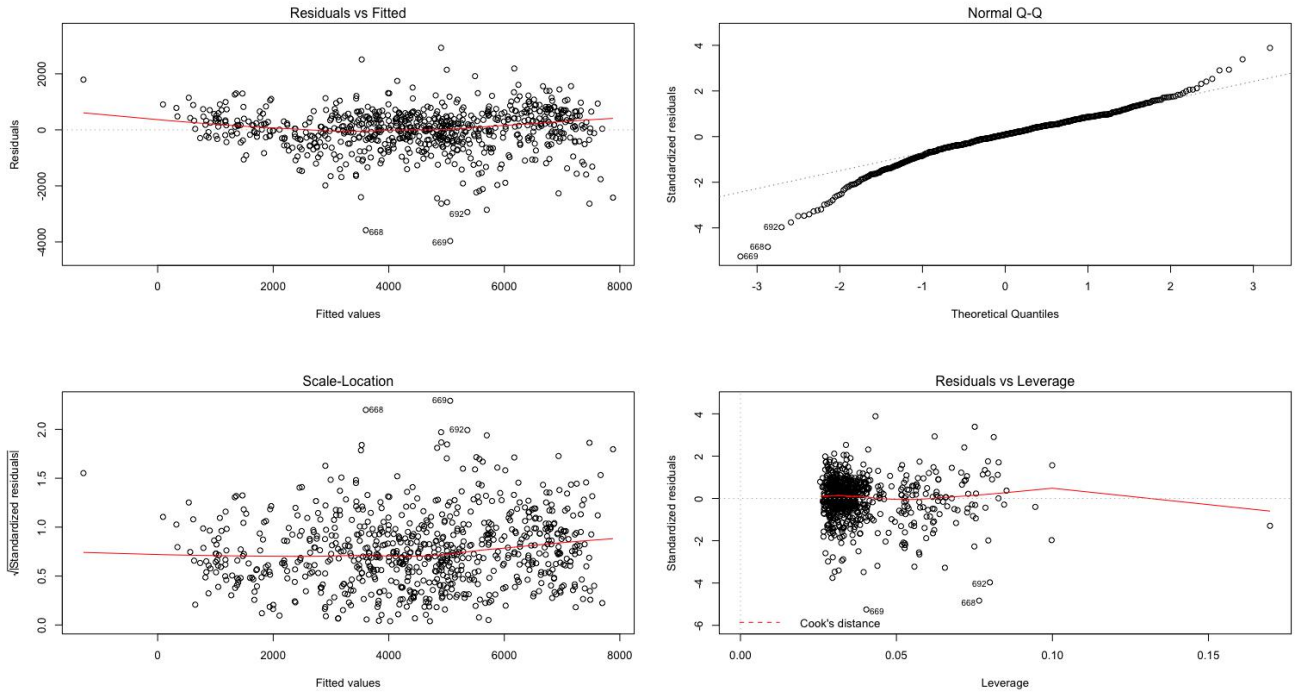
(Figure12)

By observing model 4 (Table2 & Figure 12), we can see that the degree of fitting between the predicted value and the actual value is good. But we also found that there is a place (about February 2011) where the forecast is negative, which is not in line with the actual situation. Therefore, although the fitting degree of the model is good, it is not very accurate, so next we use other methods to improve the model.

Dummy(day):Model6:



(Figure13)



(Figure14)

Model 6 sets season,year,month,holiday, weekday, workingday, weathersit as a linear variable and atemp,humid,windspeed as a continuous variable for linear regression. There is no essential difference between model 6 and model 4, but it is convenient for further analysis of the optimal linear model in model 6.

Figure 14 shows the k-fold cross-validation results of model 6 with a combined mean square error of 630989.

The residual data of model 6 is normally distributed, while the result of shapiro test is P = 0.0053, which rejects the original hypothesis. The upper-right small image of figure 14 shows that there is an outlier value of 692,668,669.

The non-horizontal curve in the lower left of figure 14 implies heteroscedasticity. The result of BP test is a small p value (0.018), which indicates that the original hypothesis is rejected, and the original hypothesis is that the variance is equal, so the residual has heteroscedasticity.

The upper-left small image of figure 14 shows that the scatters are distributed near the curve, indicating that there is no good linear relationship.

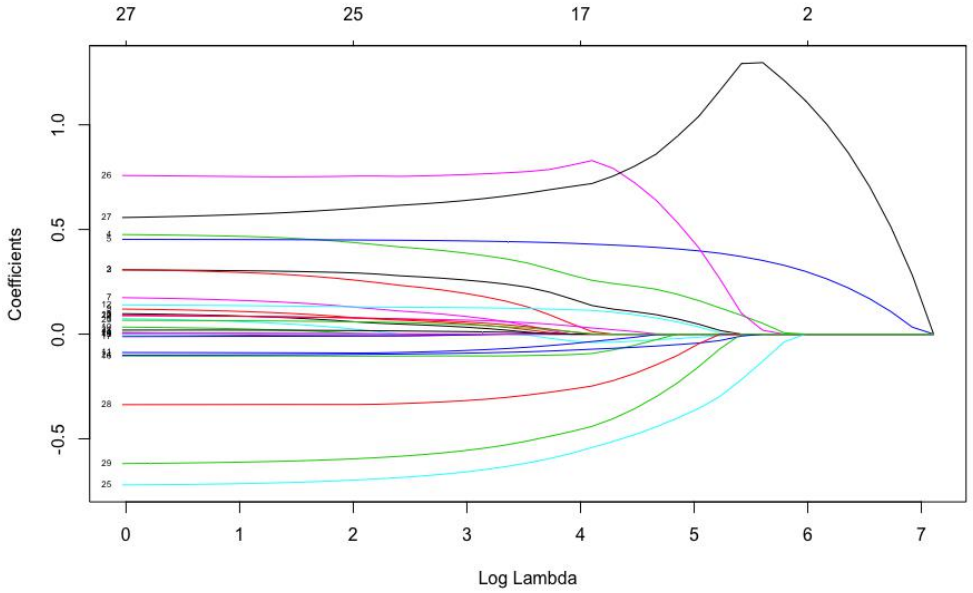
Lasso regression(day):

The predicted value of the above generalized linear regression will be not only negative, but also a lot of decimals, which is obviously not consistent with the actual situation. In order to solve the disadvantages in generalized linear regression, the Poisson regression model is used to show that the dependent variable is a non-negative positive integer and obeys Poisson distribution. Because there are discrete independent variables in the data, such as month, weather and other factors, One-hot encoding has been adopted. We set 50 different lambda values for analysis.

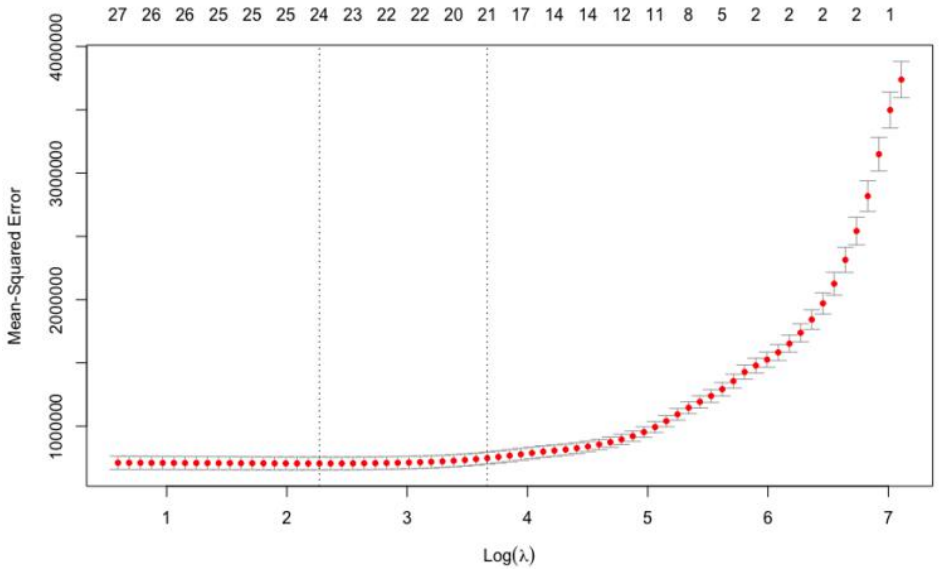
From figure 15, we can see how the coefficient of the model varies with lambda and gradually tends to zero, which indicates the result of variable selection, that is, to invalidate the

variable in order to avoid overfitting. Each curve in the graph represents the changing track of each independent variable coefficient, the ordinate is the value of the coefficient, the lower Abscissa is log (λ), and the upper Abscissa is the number of non-zero coefficients in the model at this time.

From figure 16, we can see the process of lambda reduction and MSE reduction. The dotted line in the figure shows a model with excellent performance but the least number of independent variables. So we can specify the lambda value and then predict it. From the results of regression analysis, it is concluded that the best λ value is 35.61, the degree of freedom is 21, the goodness of fit of the regression model under the above parameters is 81.3%, and the predicted values are non-negative positive integers.



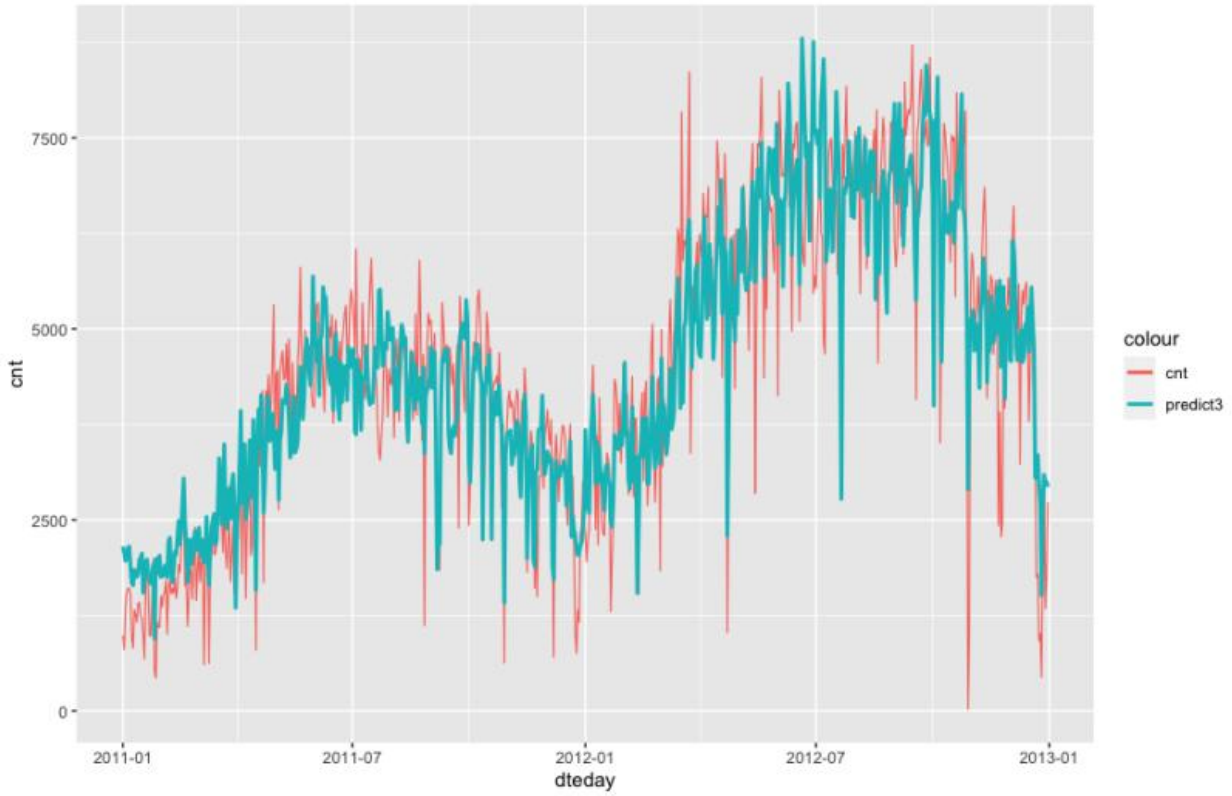
(Figure15)



(Figure16)

From the figure above, we can see that the model processed by lasso has a good fit in the prediction, and there is no negative value, so we regard it as the most suitable model to predict the total amount of shared bikes used every day.

The model is as follows:



(Figure17)

Hour

Lm(hour):

The first is linear regression. Figures 18 to 20 show the results of lm regression for the cnt, casual and registered, respectively.

```
Residuals:
    Min      1Q  Median      3Q     Max
-388.01  -93.40  -27.58   60.68  642.06

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.575e+01  7.057e+00  -3.650 0.000263 ***
season       1.990e+01  1.819e+00  10.941  < 2e-16 ***
yr           8.109e+01  2.161e+00  37.463  < 2e-16 ***
mnth        -8.648e-03  5.672e-01  -0.015 0.987834
hr           7.671e+00  1.649e-01  46.513  < 2e-16 ***
holiday     -2.188e+01  6.694e+00  -3.268 0.001084 **
weekday      1.878e+00  5.407e-01   3.474 0.000514 ***
workingday   3.939e+00  2.395e+00   1.644 0.100126
weathersit  -3.432e+00  1.905e+00  -1.802 0.071558 .
temp         7.815e+01  3.695e+01   2.115 0.034478 *
atemp        2.332e+02  4.152e+01   5.616 1.99e-08 ***
hum         -1.982e+02  6.889e+00 -28.770  < 2e-16 ***
windspeed    4.157e+01  9.628e+00   4.317 1.59e-05 ***
```

(Figure18)

```
Residuals:
    Min      1Q  Median      3Q     Max
-95.644 -20.537  -3.665  13.641 272.152

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  18.97181    1.81005  10.481  < 2e-16 ***
season        1.64953    0.46649   3.536 0.000407 ***
yr           10.27258    0.55517  18.504  < 2e-16 ***
mnth         -0.03204    0.14547  -0.220 0.825676
hr            1.20460    0.04230  28.478  < 2e-16 ***
holiday     -11.65178    1.71703  -6.785 1.19e-11 ***
weekday       0.33061    0.13868   5.989 2.15e-09 ***
workingday   -3.52726    0.61446  -5.740  < 2e-16 ***
weathersit    1.99824    0.48850   4.091 4.32e-05 ***
temp         52.62079    9.47914   5.551 2.88e-08 ***
atemp        62.92738   10.64870   5.909 3.50e-09 ***
hum          69.04589    1.76687  39.087  < 2e-16 ***
windspeed     2.99422    2.46960   1.212 0.225365
```

(Figure19)

```
Residuals:
    Min      1Q  Median      3Q     Max
-305.33  -78.74  -25.16   46.54  620.24

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -44.72910    6.14037  -7.284 3.37e-13 ***
season       18.24981    1.58249  11.532  < 2e-16 ***
yr           70.81457    1.88333  37.601  < 2e-16 ***
mnth          0.02339    0.49318   0.047  0.96219
hr            6.46600    0.14349  45.061  < 2e-16 ***
holiday     -10.22744    5.82481  -1.756  0.07913 .
weekday       1.04774    0.47045   2.227  0.02595 *
workingday   39.21048    2.08447  18.811  < 2e-16 ***
weathersit   -5.43034    1.65719  -3.277  0.00105 **
temp         25.52899   32.15683   0.794  0.42727
atemp       170.22971   36.12441   4.712 2.47e-06 ***
hum        -128.23879    5.99389 -21.395  < 2e-16 ***
windspeed    38.57100    8.37783   4.604 4.18e-06 ***
```

(Figure20)

From the lm regression of Hour, we can see that most of the variables contained in the data have a positive impact on the cnt, casual and registered, respectively. But on the whole, the number of users of non-members has a greater impact on the environment, which is easy to understand, because the use of more members has become a habit and is less affected by the external environment. Secondly, we find that the feeling temperature has a more obvious effect on the usage of shared bicycles than the actual outdoor temperature.

Next up is glm analysis.

Glm(hour):

Model	X	Y	Goodness
Model1	mnth+hr+atemp+hum+windspeed	cnt	0.486
Model2	as.factor(mnth)+as.factor(hr)+as.factor(weathersit)+atemp+hum+windspeed	cnt	0.627
Model3	as.factor(season)+as.factor(mnth)+as.factor(hr)+as.factor(weathersit)+atemp+hum+windspeed	cnt	0.631
Model4	as.factor(season)+as.factor(yr)+as.factor(holiday)+as.factor(weekday)+as.factor(workingday)+as.factor(mnth)+as.factor(hr)+as.factor(weathersit)+atemp+hum+windspeed	cnt	0.686
Model5	as.factor(season)+as.factor(yr)+as.factor(weekday)+as.factor(workingday)+as.factor(mnth)+as.factor(hr)+as.factor(weathersit)+atemp+hum+windspeed	cnt	0.686
Model6	as.factor(hr)+as.factor(yr)+as.factor(weekday)+as.factor(workingday)+as.factor(mnth)+as.factor(weathersit)+atemp+hum+windspeed	cnt	0.682

(Table 3)

From Table 2, we can see that the goodness of fit of models 4 and 5（0.686） is relatively good, and the preliminary conclusion in glm is that it is more appropriate to use the independent variables contained in models 4 and 5 to predict the number of cyclists.

Next, let's take 80% of the data as the train set and 20% of the data as the test set to take a look at the predictive performance of lm and glm,

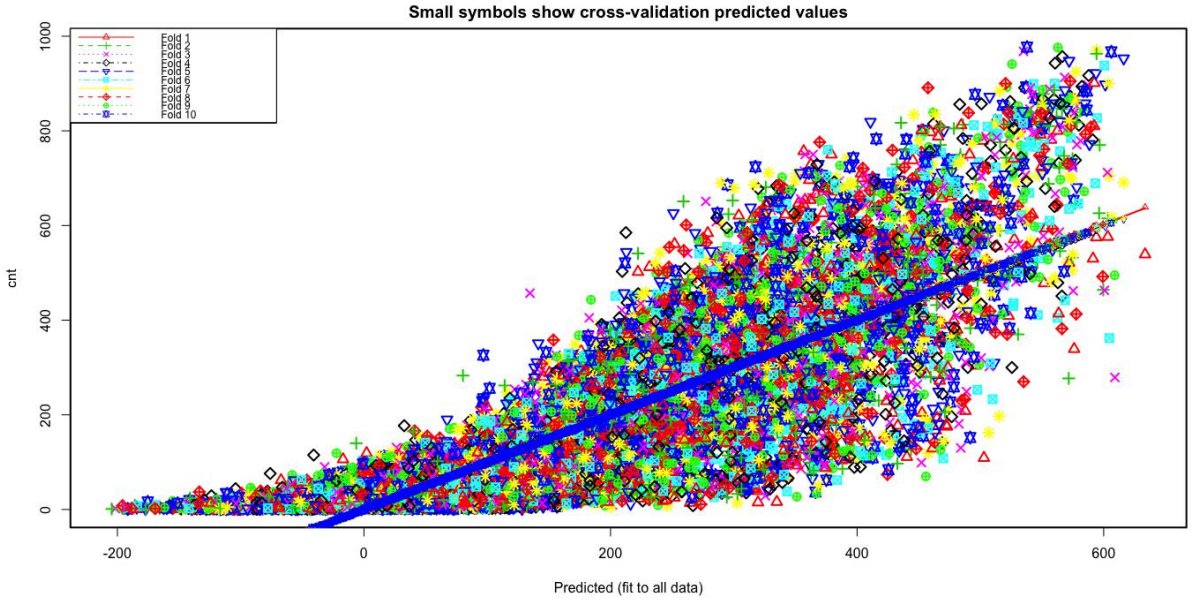
Y	回归方法	MSE
cnt	lm	20197.05
cnt	glm	10510.56
casual	lm	52029.16
casual	glm	50355.61
registered	lm	22667.43
registered	glm	13466.27

(Table4)

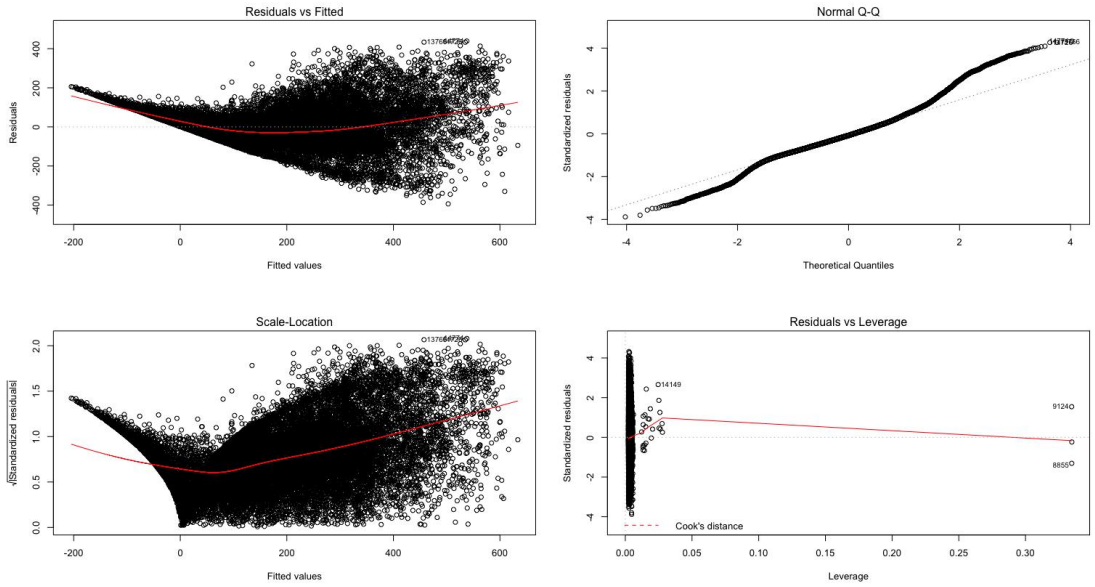
From the results in Table 3, we can see that no matter what kind of population it is, the test effect of glm is better than that of lm.

Dummy(hour):

We further analyze the advantages and disadvantages of linear regression according to dummy model. From the test result in Figure21 and Figure22, we can see that the linear model lacks good linear relationship and residual normality, and has heteroscedasticity, so it can not be used as an appropriate prediction model.



(Figure21)

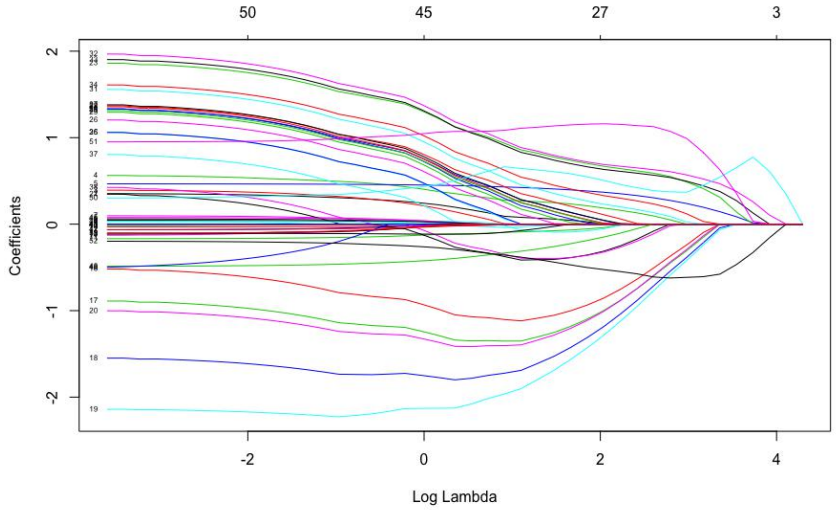


(Figure22)

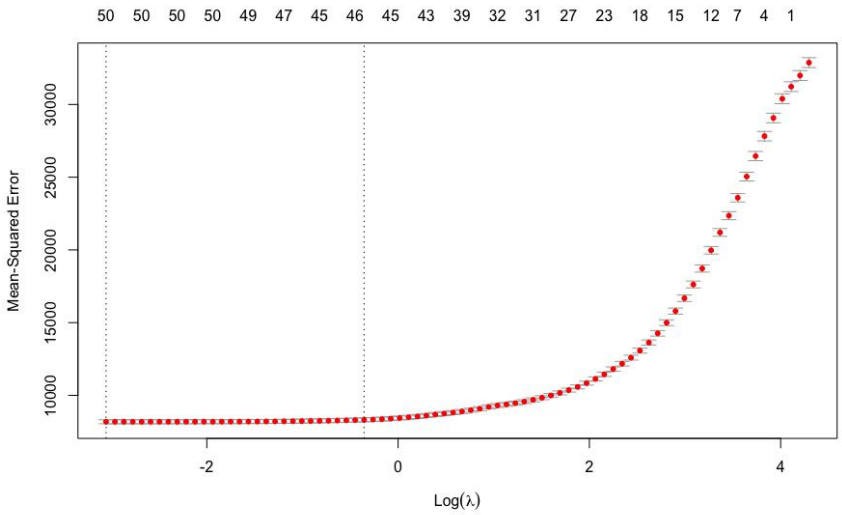
Lasso(hour):

In order to improve the prediction ability at different times of the day, we continue to use Lasso regression. Based on the Poisson regression model in the previous part, we add hour variables and use time series data sets divided into hours. We set 50 different lambda values for analysis. From the figure below, we can see the variation curve of the independent variable as the lambda increases, and the MSE growth curve. The lambda value with good prediction performance and avoiding overfitting is given at the dotted line of figure24. From the results of regression analysis, it is concluded that the best λ value is 0.700815, the degree of freedom is 45, the goodness of fit of the regression model under the above parameters is 79.52%, and the predicted values are non-negative positive integers.

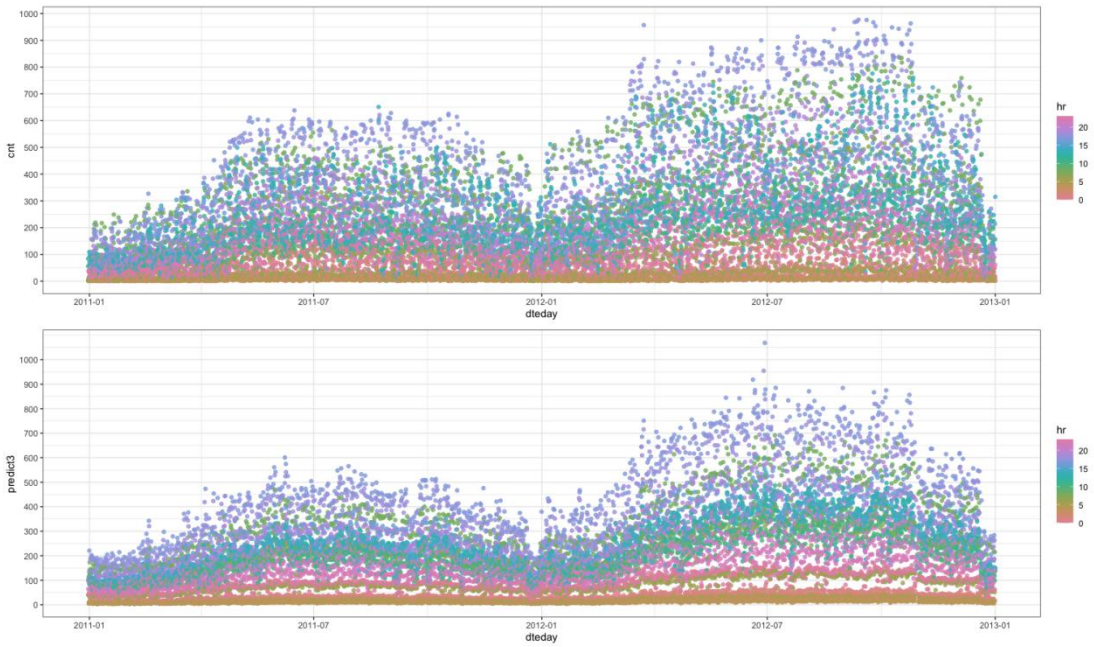
The fitting effect of the model is shown in figure 25, with each color representing an hour interval. Obviously, the fitting value of the prediction model is consistent with the actual value in the overall trend, but the distribution of the predicted value is more concentrated, thus forming obvious different layers, showing a strong generalization ability.



(Figure23)



(Figure24)



(Figure25)

4.Limitation

（1）In terms of data, there are still many factors that affect the number of cyclists that we have not included, such as the lack of users, geography and urban environment, which may affect the promotion of our analysis results in other cities.

（2）Although the analysis process is relatively complete, the regression methods we have mastered are still relatively limited. You can also try other regression methods.

（3）In the final prediction results, it is a great pity that the prediction accuracy is not improved to more than 90%.

（4）The report uses data on the use of shared bikes in Washington, USA, some of which are similar to but not exactly the same in China, and may change users' behavior habits, so there may be a deviation from the Chinese scenario in the detailed prediction analysis.

5.Suggestion In Management

We see that the two major variables that affect usage are time periods and external weather.

In the time period, the general trend is similar in all three cases, with a big peak at 8am and 6pm and a small peak at 12 noon, so bicycles must be deployed to the right place before the peak time, otherwise it will seriously affect the usage of the day.

We can see that compared with the actual temperature, the relationship between feeling temperature and use times is more obvious, even if the temperature is the same on the same day, we should put more bicycles in places with higher feeling temperature.

We can see that the correlation between the season and the number of times of use is very strong, because the season reflects a more comprehensive situation of the weather than the temperature or wind speed. In the season of less use, it is necessary to do a good job in the recycling and maintenance of bicycles.

The usage count of "registered" is about four times that of "casual". In the face of this huge gap, on the one hand, we should do a good job in the maintenance of existing members (such as good customer service), and on the other hand, we should strive to convert non-members into members. For example, the early launch of a lot of preferential activities.