Bike Sharing Between Registered and Casual Group in Washing D.C.

Our project focuses on the effect of environmental factors on bike sharing behavior of two user groups, namely registered users and casual users. We try to identify the effect of several indicators of weather and other conditions, e.g. holiday, working day and explore the similarity and difference between the two groups.

Background and Data Source

Nowadays bike sharing systems are playing an important role in traffic. Currently, there are about over 500 bike-sharing programs around the world which is composed of over 500 thousands bicycles. The popularity of bike sharing in China has been growing all the time. In the United States, bike sharing dates back to early 20th century and those bike sharing systems are owned by local governments to promote an environment-friendly way of traffic.

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. Furthermore, different type of users may react to the environment differently. As registered users are those who register for an annual or 30-day membership and rent more frequently, they may be less vulnerable to environment change than casual users who are usually short-term users.

Therefore, our project focuses on the impact of environmental factors on sharing bike usage. By regressing against count of registered users and casual users respectively, we aim to identify the differences in behavior pattern of the two group of users.

The dataset we use is from UCI Machine Learning Repository (http://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset). It contains daily count of rental bikes usage between years 2011 and 2012 in Capital bikeshare system with the corresponding weather and seasonal information. We separate the data into two groups, registered and causal. Using count of registered users and count of causal users as explained variables and other variables as potential explanatory variables, we analyzed the two groups separately and made a comparison.

The response variables of the two regressions are:

Variable	Description	Observations
casual	count of casual users in a given day	731
registered	count of registered users in a given day	731
cnt	count of total rental bikes including both casual and registered in a given day	731

The potential explanatory variables are:

Variable	Description	Variable	Description
season	1:springer, 2:summer, 3:fall, 4:winter	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
year	0:2011 1:2012	temp	Normalized temperature in Celsius
month	1 to 12	atemp	Normalized feeling temperature in Celsius
holiday	0: not a holiday, 1: holiday	hum	Normalized humidity
weekday	day of the week	windspeed	Normalized wind speed
workingday	1: working day 0: otherwise		

Data exploration and preprocessing

Refer to totally sixteen variables mentioned above, we initially remove instant, dteday, season and weekday due to the following considerations: instant contains actually no information; dteday is useless for analysis given year and month data; as for season, since we have more interest in the effect of weather on the usage of sharing bikes, we think the effect of weather can be reflected by other weather variables based on our common sense; working day makes more sense rather than detailed weekday.

Then we transform the data into observed scale according to the transformed method used by the data provider because we think the normalized way is not so appropriate and it might be the case that normalization is unnecessary in our analysis.

After simple exploration of each continue variables with histograms and summary statistics (see Appendix 1), it seems that the distribution of each variable is quite good. Refer to the correlation matrix, we can see that weather condition and environment do have some effect on the usage frequency of both registered and casual users. And temperature has really high linear correlation with feeling temperature which can be seen from Figure 1 as well. It gives us incentive to only keep feeling temperature to avoid multicollinearity. Furthermore, Figure 1 conveys some kind of interesting patterns: there are two linear trend in the relationship of casual and registered users which arose our interest to figure out the hidden reason.

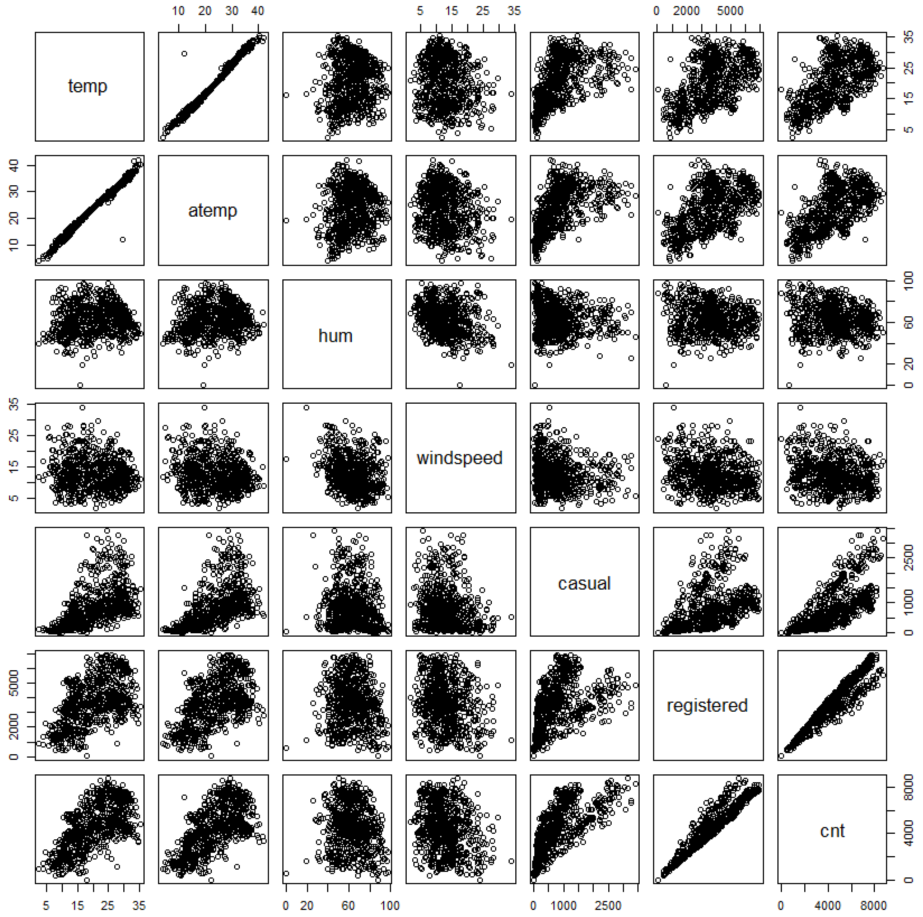


Figure 1 Scatter plot

Creating dummy variables for weathersit

While it comes to discrete variable weathersit, it doesn't have extreme weather record which is acceptable since weather condition in Washington D.C. is quite stable. So we create two dummy variables to denote the weather state:

$$\text{good} = \begin{cases} 1 & weathersit = 1 \\ 0 & otherwise \end{cases} \qquad \text{fair} = \begin{cases} 1 & weathersit = 2 \\ 0 & otherwise \end{cases}$$

Creating quadratic and cross term

It is reasonable to assume that there will be an optimum temperature which means usage will increase as temperature increases at first and after the turning point, the effect of temperature will be negative. In addition, we can also see the similar pattern in Figure1, 2, 3. Thus, we introduce the quadratic term of temperature into our model. And so does the humidity in spite of no obvious pattern of humidity in Figure 1.

To figure out which result in the obvious apart trend in Figure 1, we make an assumption that the effect of weather condition on the usage might depend on whether it is a working day nor not. Therefore, we add workingday×atemp (hum, windspeed).

After all these data preprocessing, we have totally 17 variables.

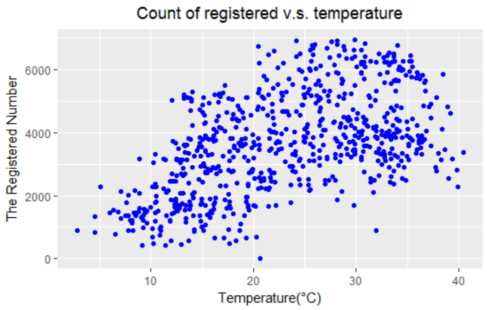


Figure 2 Apparent temp on register

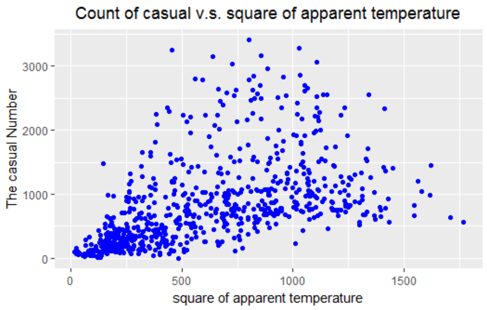


Figure 3 Apparent temp on casual

Analysis of Casual Group

Casual group are those who only use bicycle someday. We think whether they choose to ride a bicycle largely depends on weather condition. Since they do not stick to usage, time might not be the important factor. There is no significant increasing pattern over time in Figure 2, which support us to remove year and month. For analysis of casual group, the number of variables is 13.

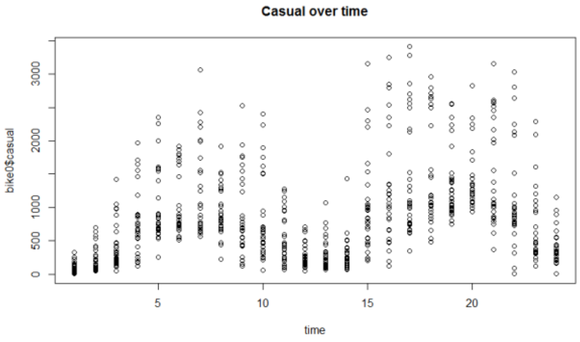


Figure 4 Casual usage overtime

Model Selection and Data Transformation

We hope to use linear regression model to fit the response. There are 12 potential explainable variables in total and based on the first try of regression, some regression coefficient turns out not to be significant. Model selection help us to choose the most appropriate one. Since the number of predictors is not too large, exhaustive search based on specific criteria is sufficient.

We use Mallow's C_p as the main criteria and refer to PRESS, AIC and BIC at the same time. For each subset size, we choose best five models. The visualization of the selected models seems to violate the basic assumption that residuals should be normally distributed and have constant variance. But there is an obvious "horn" in the residual plot as seen in Figure 3. Boxcox method helps us to solve this problem as shown in Figure 4.

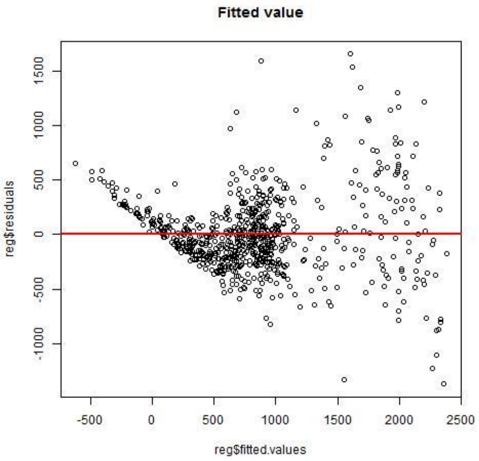


Figure 5 Residuals without transformation

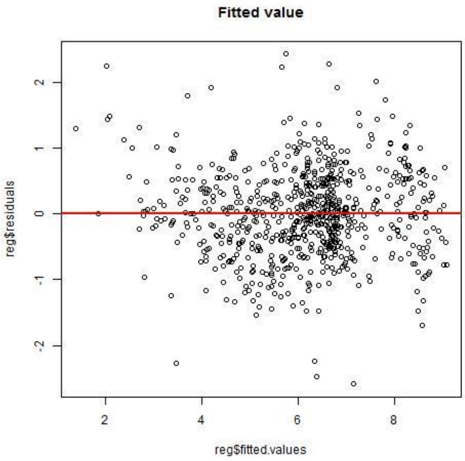


Figure 6 Residuals with transformation

Outliers Detection

The quality of dataset is quite high and even there exist outliers, they should be some extreme reality cases rather than record error. From this perspective, we should not remove any outliers. However, what we want to investigate is the cause for usage behind the casual and registered group. It seems reasonable to remove outliers for better comparison. According to Cook's distance, leverage, standardized deleted residual and DFBETAS, we remove 12 cases (see Appendix 2). Then, we redo the model selection and the "horn" pattern still exists, then we redo the boxcox transformation and find out best $\lambda = 0.257$.

Result and Interpretation

$$Y = \beta_0 + \beta_1 holiday + \beta_2 workingday + \beta_3 atemp + \beta_4 atemp2 + \beta_5 windspeed + \beta_6 good + \beta_7 fair + \beta_8 hum2 + \beta_9 wkTemp$$

Intercept	holiday	workingday	atemp	atemp2
1.051***	-0.3485**	-0.8654***	0.3723***	-0.005517***
(0.253)	(0.124)	(0.131)	(0.0151)	(0.000315)
windspeed	good	fair	hum2	wkAtemp
-0.02606***	0.7961***	0.7052***	-0.0001594***	-0.01698**
(0.00415)	(0.151)	(0.138)	(0.00001632)	(0.00525)

Given other predictors in the model, holiday has negative effect on causal users' usage of sharing bikes (notated as CUU). The possible reason might be that most American celebrate traditional holidays at home or travel far away rather than go outside within a short distance. Causal users tend to use less on working because they don't use bikes as their working transportation. The effect of feeling temperature on CUU is consistent with our assumption that there exists an optimum point and there is a difference between weekday and weekend. The result might be that when casual users need to use bikes on working day, they could be in a hurry to use sharing bikes and the effect of feeling temperature is moderated. The better the overall weather condition, the more the causal users' usage. It is consistent with the reason mentioned above that casual users mostly ride sharing bikes for a short trip, hiking for instance. The effect of squared humidity is negative but the regression coefficient is relatively small, which can almost be ignored.

Analysis of Registered Group

In this part, the response variable is the number of registered users on a daily basis.

Unlike casual group, the count of registered seem to increase as time goes on so we add this variable in our model as a dummy variable. In total, there are 13 potential explanatory variables including cross terms and quadratic terms.

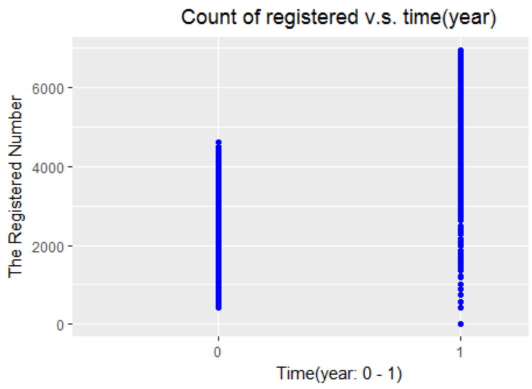


Figure 7 usage of registered against year

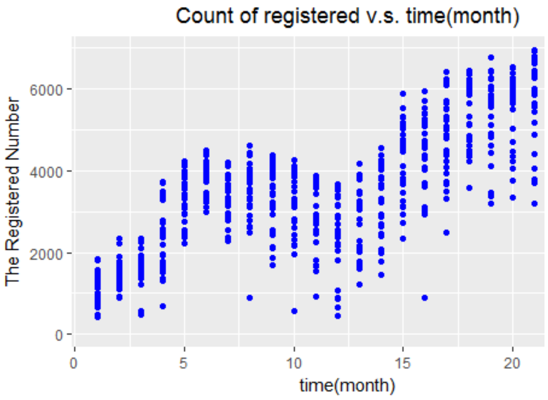


Figure 8 usage of registered against month

Model Selection and Data Transformation

We have considered 4 criterion in the model selection process: Adjusted $R_{a,p}^2$, Mallow's Cp, AIC, SBC. From the table (in appendix), we consider the first model that have 11 variables is a nice choice. Its adjusted $R_{a,p}^2$ is relatively high and AIC and SBC are relatively small. Meanwhile its Cp is the smallest of all the models and is close to p which is 11. The model is:

$$\begin{aligned}\mathrm{Y} = \beta_0 + \beta_1 year + \beta_2 workingday + \beta_3 atemp + \beta_4 atemp2 + \beta_5 windspeed \\ + \beta_6 hum + \beta_7 hum2 + \beta_8 good + \beta_9 fair + \beta_{10} wkhum \\ + \beta_{11} wkTemp\end{aligned}$$

We further plot residuals against fitted values and find a horn-like shape, which indicates a need for some transformation of response variable. Using Box-Cox transformation we select λ=0.77, and then the residual plots become more satisfactory.

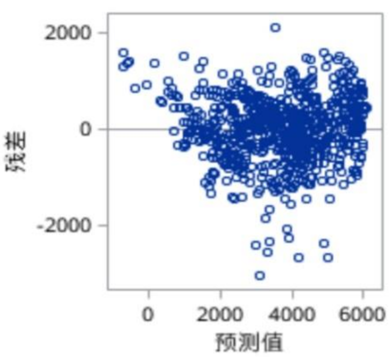


Figure 8 residual plots before transformation

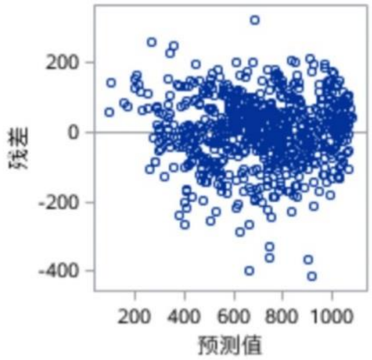


Figure 9 residual plots after transformation

Outlier Detection

After transformation we fitted the model and use several methods to detect outliers. The figure of Cook's D indicates the existence of evident outliers. We judged from each case's leverage and DIFFITS to test if they should be deleted. We throw 10 or so cases and make a new fit, and the result is shown in the appendix. Now in the new figure of Cook's D, each case is within the permitted scope. What's more, the Adjusted $R_{a,p}^2$ has increased slightly from 0.8029 to 0.8174.

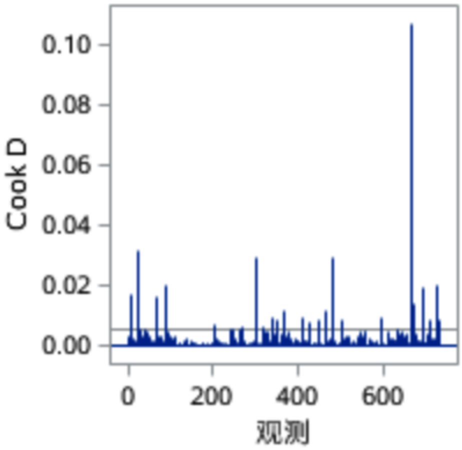


Figure 10 Cook's distance before transformation

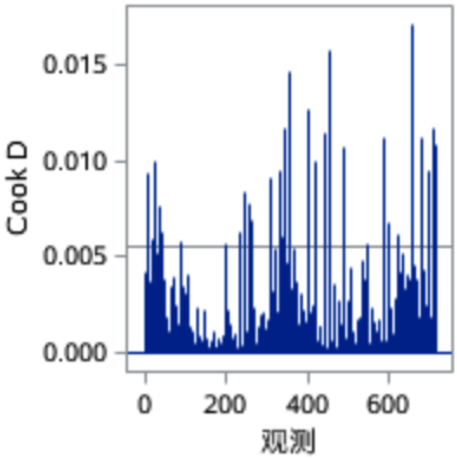


Figure 11 Cook's distance after transformation

Result and Interpretation

$$\begin{aligned} \mathrm{Y} = \beta_0 + \beta_1 year + \beta_2 workingday + \beta_3 atemp + \beta_4 atemp2 + \beta_5 windspeed \\ + \beta_6 hum + \beta_7 hum2 + \beta_8 good + \beta_9 fair + \beta_{10} wkhum \\ + \beta_{11} wkTemp \end{aligned}$$

Intercept	yr	workingday	atemp	atemp2	hum
-555.800***	257.6984***	176.9191***	56.5244***	-0.9460***	10.1506***
(80.0)	(7.59)	(41.4)	(2.96)	(0.0605)	(2.25)
windspeed	good	fair	hum2	wktemp	wkhum
-5.7311***	74.2399*	115.5507**	-0.0898***	2.4552*	-1.4627*
(0.785)	(30.4)	(32.6)	(0.0176)	(1.00)	(0.591)

At first, 'holiday'and 'workingday*windspeed' have low explanatory power and has been deleted from the model. A plausible explanation for holiday is that the number of registered don't change violently when in holiday, probably because holidays usually don't last too long, so people don't usually register to use the bikes in a long run just for the casual of holiday, i.e. people register to use bike mostly on other purposes, mostly not for the fun of holiday. As for workingday*windspeed, since workingday and windspeed are all influential to registered, so maybe we delete it because of the multicollinearity.

The results show that the effect of year is significant. It is reasonable because the popularity of bike sharing grows as time goes on so more people registered in for a long-time membership and use bikes more frequently.

The apparent temperature and quadratic of apparent temperature are also significant. The sign of temperature is positive and the sign of quadratic temperature is negative , indicating a inverted U shape relationship. In other words, there may be a optimal temperature for people to go out and use bikes as a transportation mean. To be specific, people tend to use bike less if it's too cold and as temperature goes up, registered users rent bikes more frequently. However after reaching a threshold, the temperature's effect become negative. The same logic applies to humidity as well.

As for the general indicators for weather condition, fair day and good day are helpful in increasing the number of registered people. Above all, these factors all have effect on people's feeling of weather, thus influencing the number of registered.

The cross terms in the predictors, workingday times temperature and workingday times humidity shows that working day affects the response variable by affecting people's reaction to weather conditions. To be specific, if it's working day, the temperature's effect strengths while the humidity's influence become weaker. In a warmer day, if people have to go to work, they are more inclined to register, and in a wetter day, people are more likely not in the mood of riding a bike to go to work. But as the p values are relatively large, these factors only account for a little of the reason of usage of registered users.

Overall regression

Finally we aggregate count of registered and casual users as a response variable to see the effect as a whole. Using similar method as the former two parts, we get the final Model:

$$\mathrm{Y} = \beta_0 + \beta_1 year + \beta_2 windspeed + \beta_3 atemp + \beta_4 atemp2 + \beta_5 good + \beta_6 fair + \beta_7 spring + \beta_8 summer$$

Intercept	yr	windspeed	atemp	atemp2
-78.3512***	115.2921***	-1.4737***	20.9246***	-0.3334***
(19.0)	(3.07)	(0.308)	(1.23)	(0.0250)
good	fair	spring	summer	
148.9304***	109.1599***	-79.6812***	-17.0833***	
(10.6)	(10.7)	(4.96)	(3.87)	

From the final model, we can see all the coefficients are significant. From 2011 to 2012, we can observe an apparent increase in the *cnt*, which is corresponding yr has a positive influence on the *cnt*. The atemp has a positive effect on the cnt, while the temperature second order term atemp2 has a negative influence, which indicates there may be an optimal temperature for bike rent, atemp lower than this temperature has a positive effect, otherwise a negative effect. The larger the windspeed is, the smaller amount of bike rent is. The bike rent is highly correlated with the weather. Good weather can increase the bike use. The *cnt* is also different in difference seasons. The coefficients indicate that the there's a significant increase of *cnt* from Spring to Fall, but bike rent in fall and winter has little difference.

Discussion

Comparing the environmental effect on the registered and casual group’s usage of sharing bike, we can see quite significant difference in both the predictors and the relative regression coefficient.

As for working day, the effect for registered group is positive while for casual group is negative. The result means that some of registered group might use sharing bike as their working transportation while casual group have high mobility. The holiday is not included in the registered group and the reason might be that weekend seems to have similar effect with holiday for them.

The result tells some story but it is not fancy to support our assumption at the very beginning. There are more predictors directly related to detailed weather condition for the registered group and this means their usage frequency is significantly affected by the weather, which might due to American life style of using sharing bike.

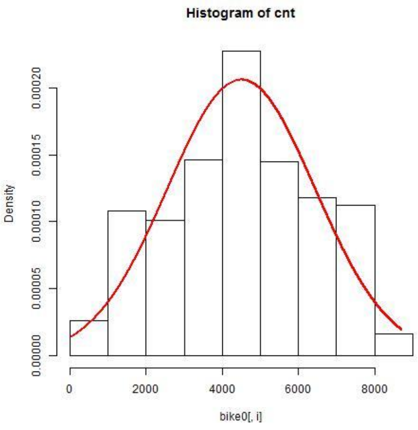
Further work

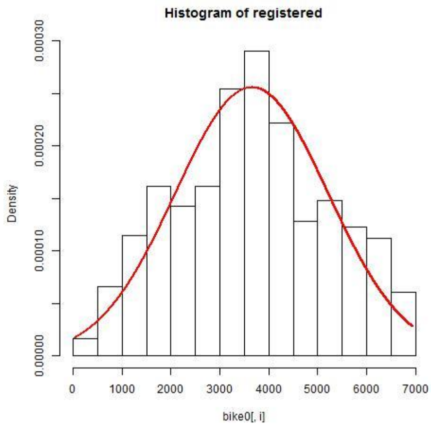
We firstly remove season because we mainly focus on the effect of weather condition on the usage of bike sharing and assume that the effect of season can be explained by other predictors, like feeling temperature and humidity. But the overall regression seems that season might have other effects since different season have different scenery and people might tend to go outside by bike in some season.

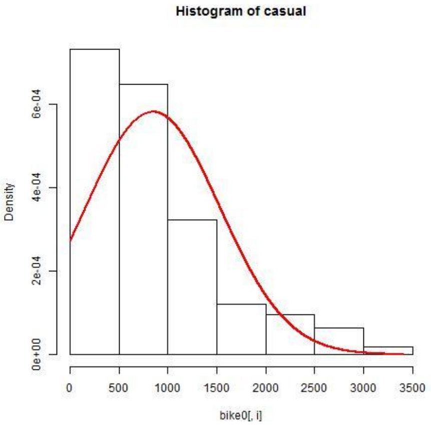
We can further use Chinese bike sharing data to explore the specific condition in China because the result turns out not to be explainable in China.

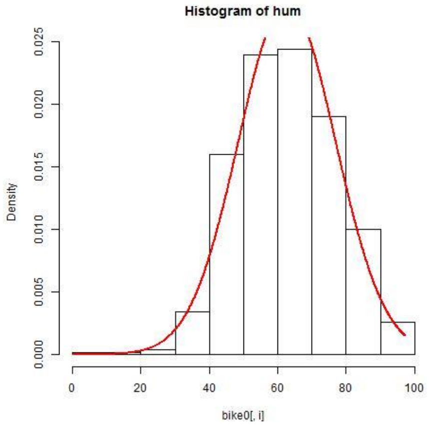
Appendix:

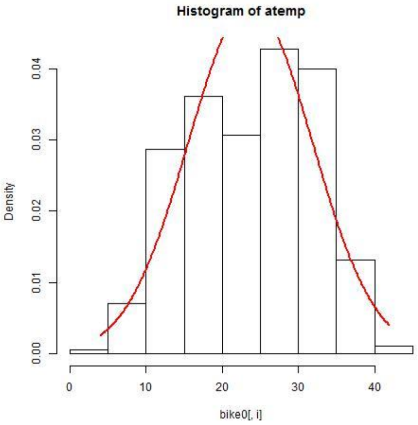
(A) Data exploration

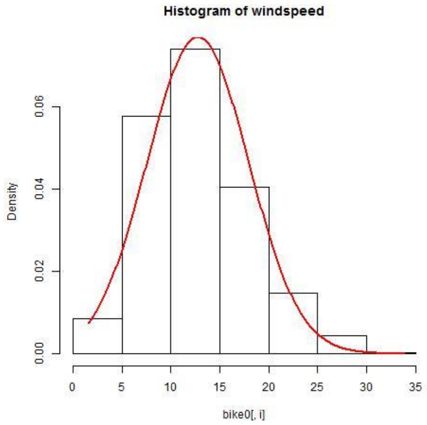


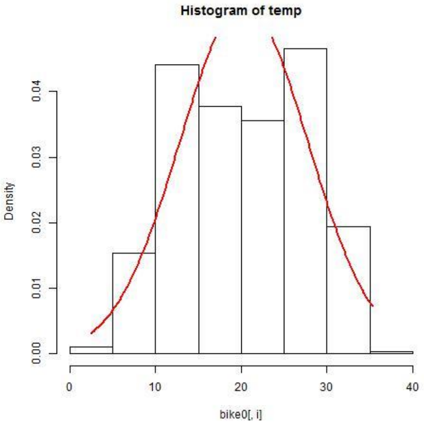




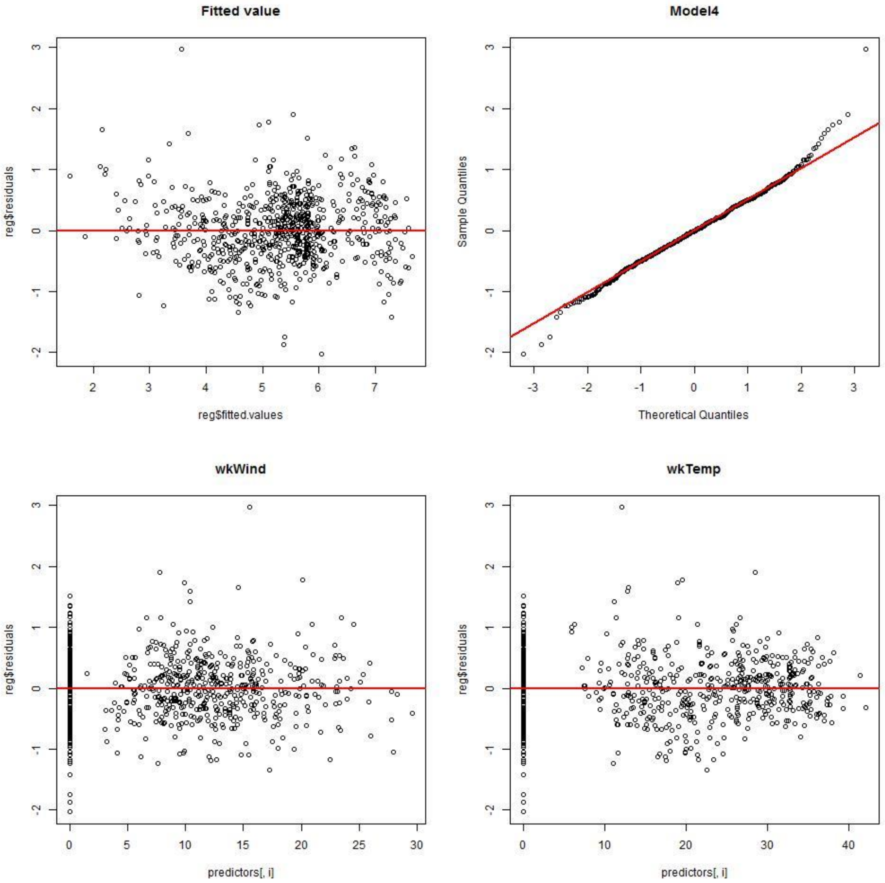


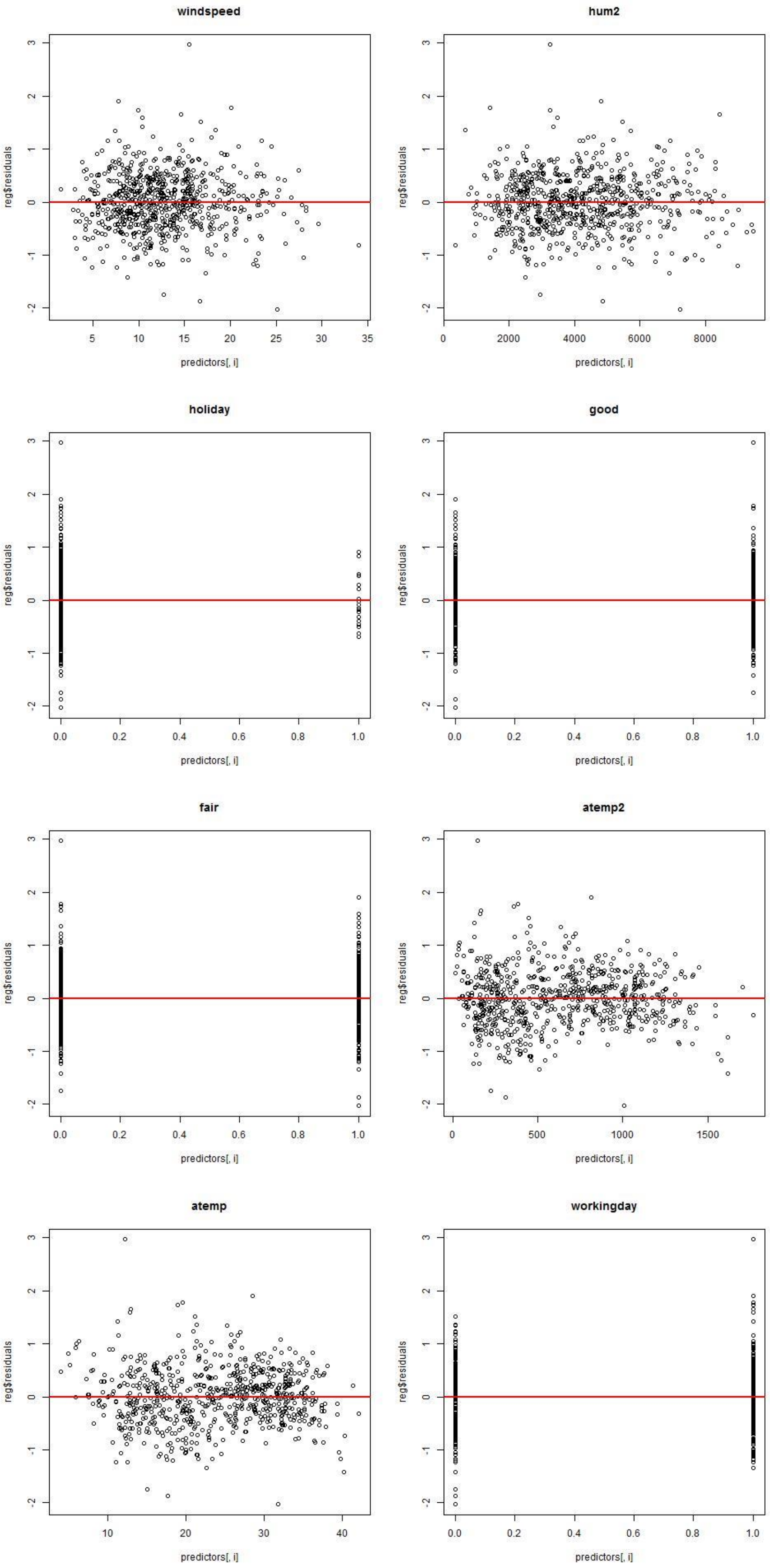






(B) For analysis of casual users:

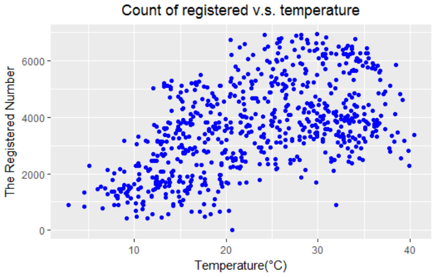


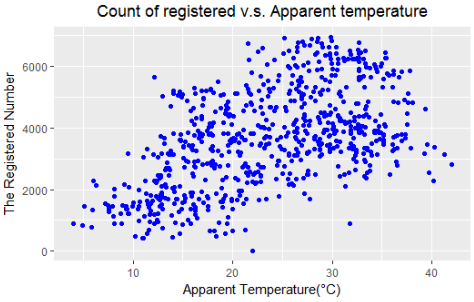

windspeed
reg$residuals
predictors[, i]
3
2
1
0
-1
-2
5
10
15
20
25
30
35
hum2
reg$residuals
predictors[, i]
0
2000
4000
6000
8000
holiday
reg$residuals
predictors[, i]
0.0
0.2
0.4
0.6
0.8
1.0
good
reg$residuals
predictors[, i]
fair
reg$residuals
predictors[, i]
atemp2
reg$residuals
predictors[, i]
0
500
1000
1500
atemp
reg$residuals
predictors[, i]
10
20
30
40
workingday
reg$residuals
predictors[, i]

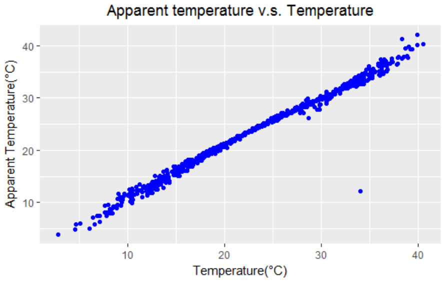
(C) For analysis of registered users:

Descriptive statistics:

a) Temperature



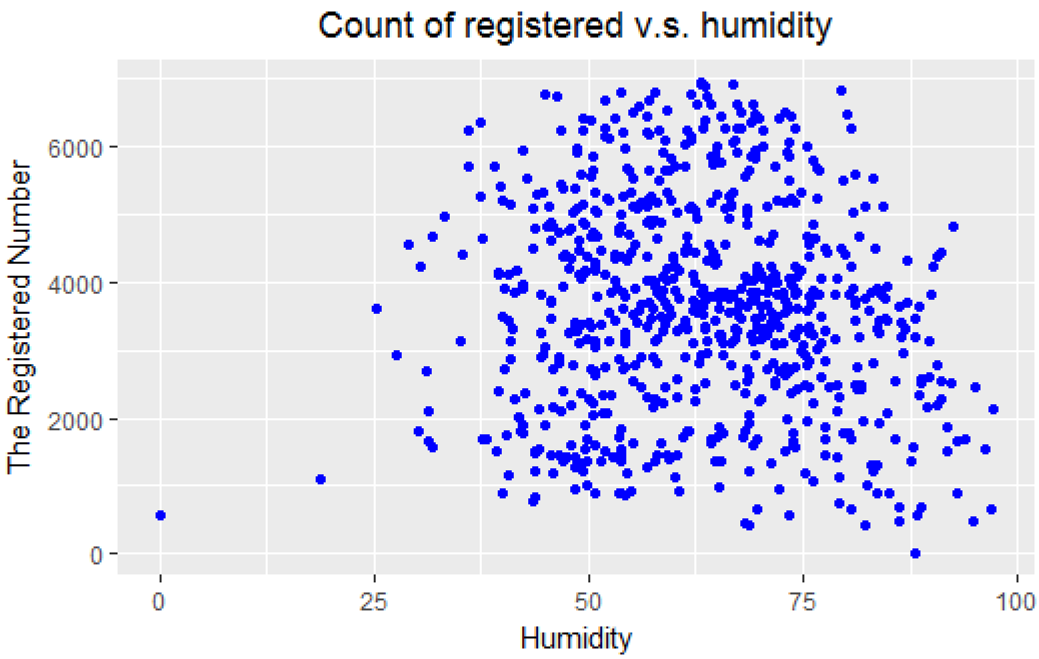




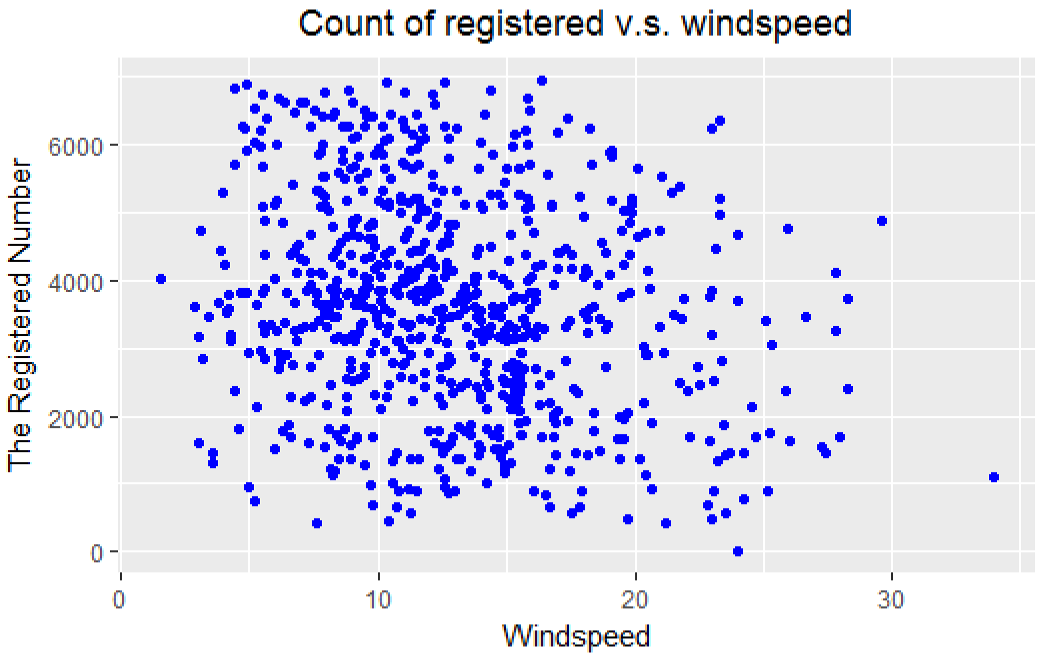
From the plots, apparent temperature has similar distribution on the plot. Thus, we assume there exists a strong correlation between temperature and apparent temperature. To justify our assumption, we study the relationship between those two variables.

From the plot above, it is quite obvious that those two variables are linear dependent. Though these variables are both thought to contribute to the response variable, we have to choose one from them because the linearity between them. Intuitively, we think the apparent temperature may be more important for users to make decision, which means the apparent temperature should be retained and the temperature should be removed temporarily.

b) Humidity



c) Wind Speed

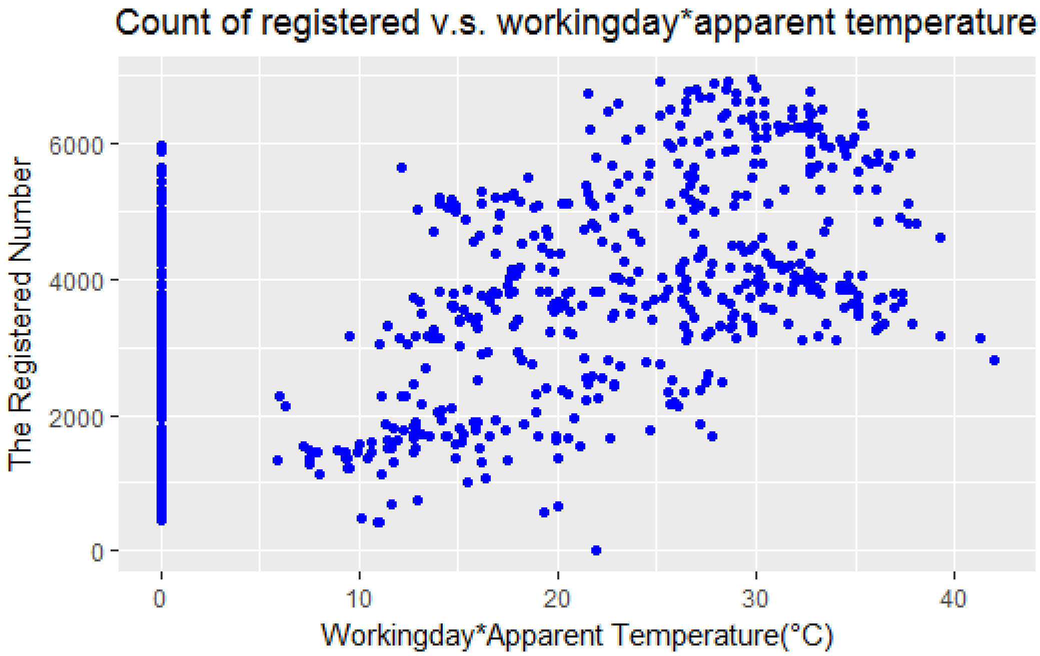


As for the humidity and wind speed variables, though we don't see some apparent correlation, we still cannot remove them easily as removing temperature. We choose to reserve them and test them in the model.

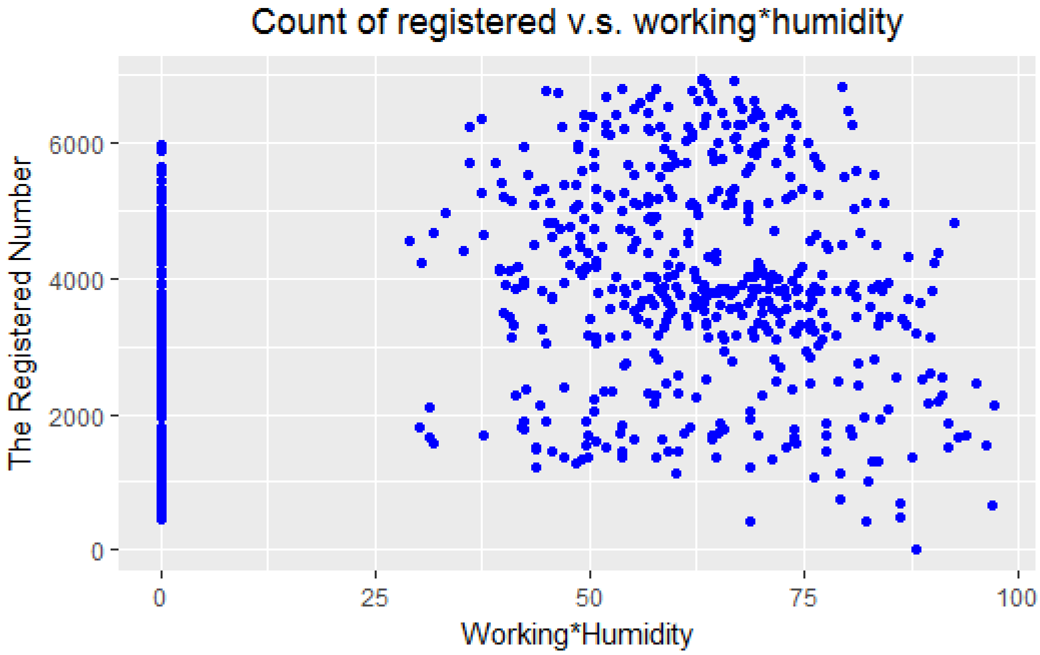
1) Cross Terms

It can be easily assumed that some of the variables have influence on each other and be influential to the response variable as a whole. For example, the temperature may have difference impacts on the number of users with regard to the fact whether it is a working day or not.

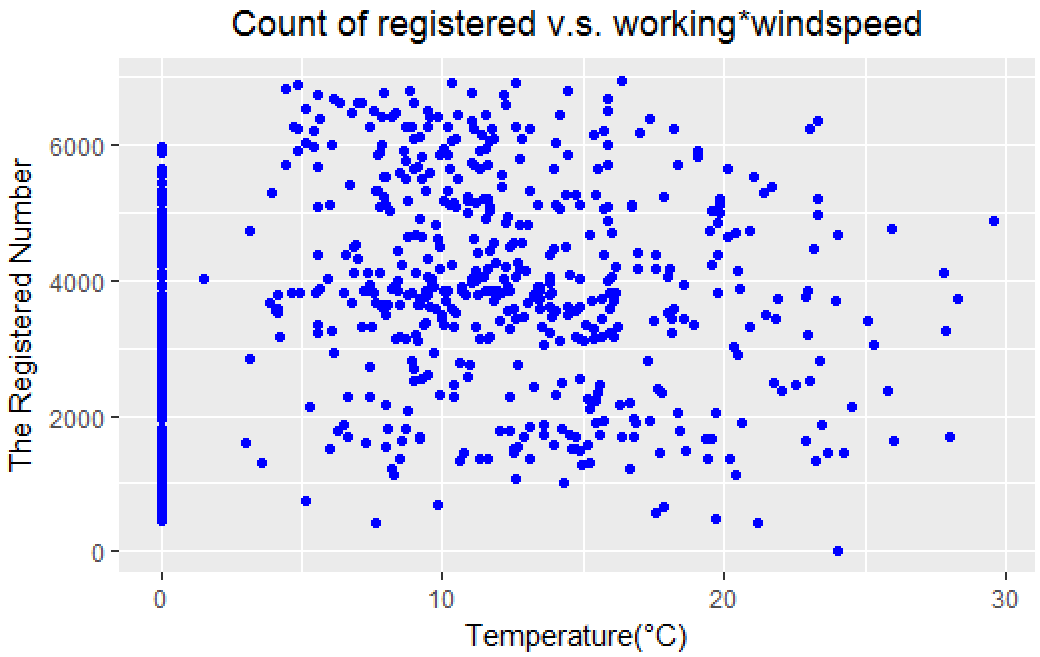
a) Working day × Apparent Temperature



b) Working day × Humidity



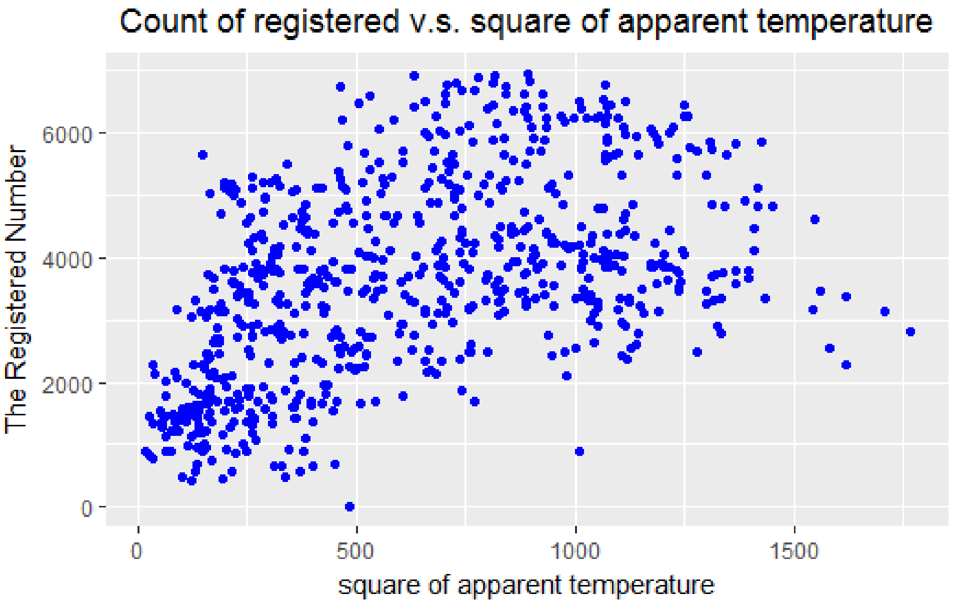
c) Working day × Wind Speed

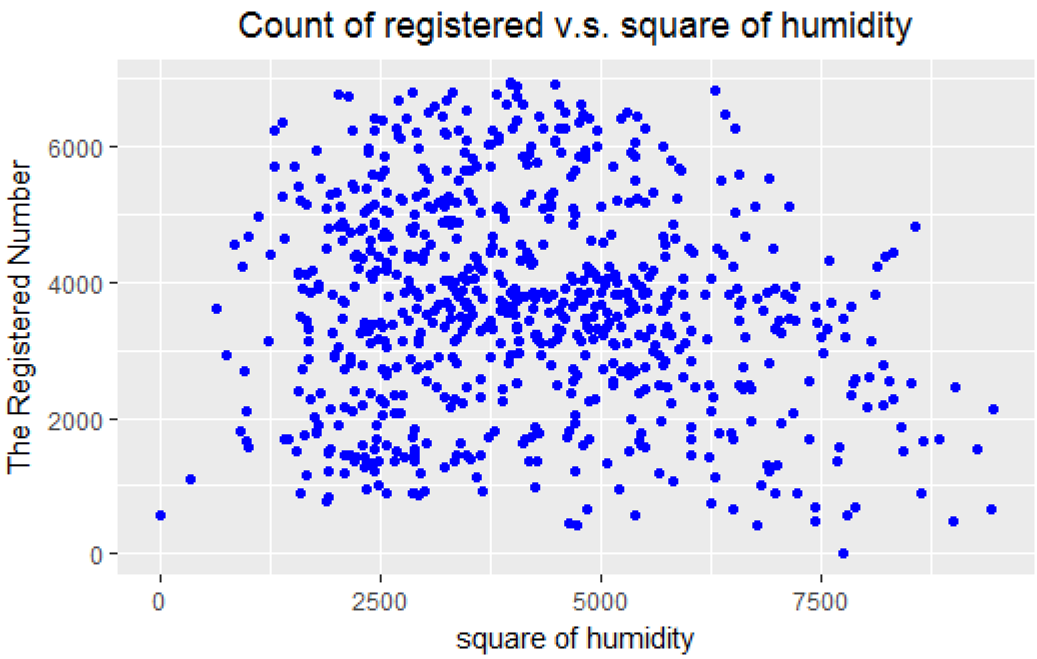


2) Quadratic Terms

Besides the raw variables and the cross terms, we think up some additional transformed variables that may be useful.

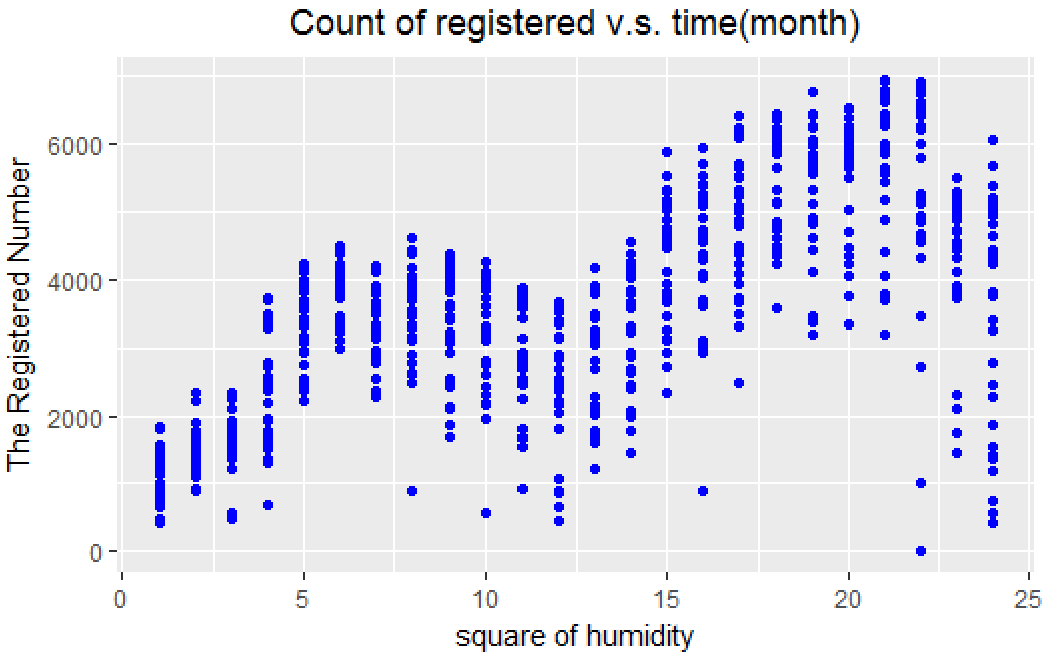
It is an empirical fact that there exists the most comfortable temperature and humidity for human. Since we are studying the impacts which the weather condition has on users' behavior, we assume there are some quadratic terms in the model. Based on the intuition, we add the square of apparent temperature and the square of humidity. The wind speed doesn't seem to fit in the situation above, so no quadratic term of wind speed is included. We also do some data visualization, and whether the variables are significant remains to be seen by later analyses.



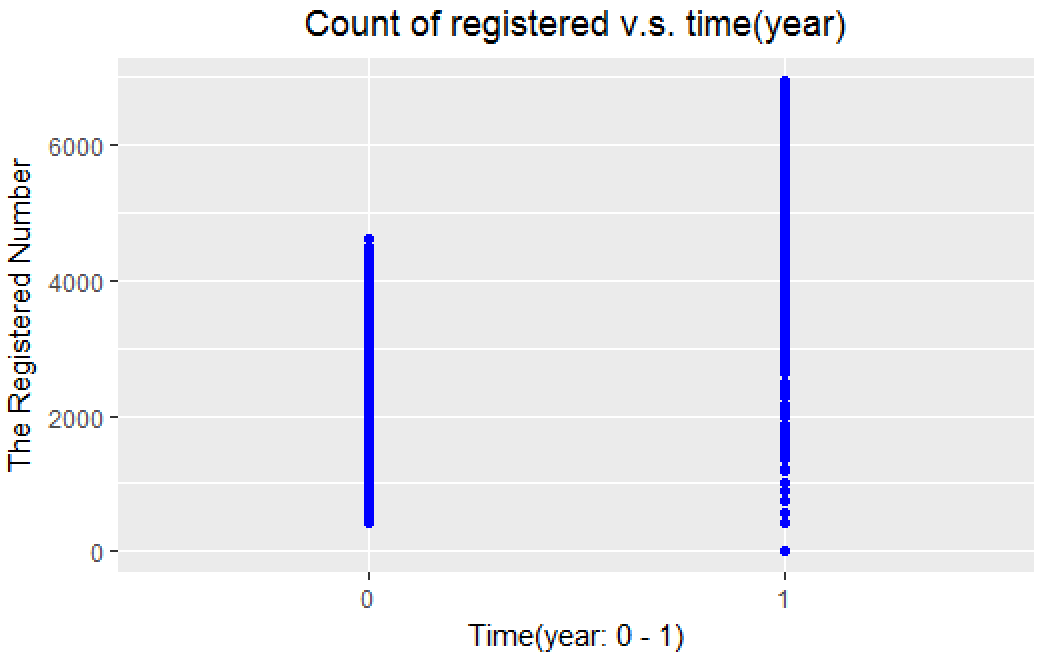


3) Method to deal with time

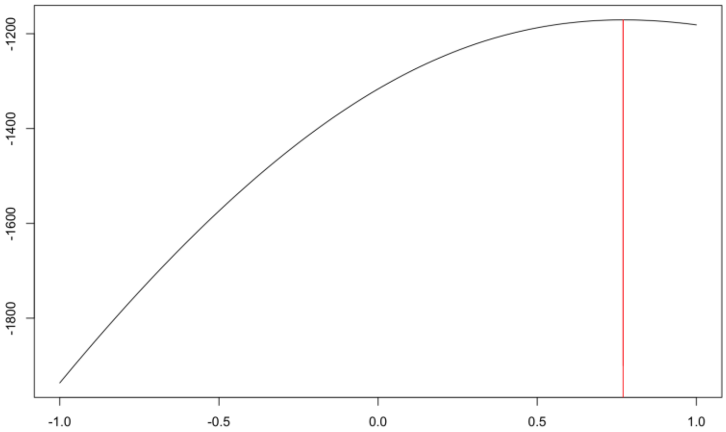
The variables include the year and month. And we think time also kind of contributes to the change of the number of registered users. To validate our thoughts, some plots are shown below.



This plot shows the number of registered users in different months (0~24 corresponds to 24 months). From the plots, we can see that the number of users fluctuates. But we can still notice that the number has been larger in the second year. It is an overall trend.



If we plot in different years, the trend can be seen more apparently. It means that time surely has influence on the response variable. And we decide include variable 'year' into the model and test whether it is significant later.



(D) For analysis of Cnt (total amount of bike rent)

1) Model setup

Our first research is to study the factors that affects the daily rent of sharing bikes in Washington D.C. So we use *cnt* (count of total rental bikes including both casual and registered) as dependent variable. We set up a suitable model in the following steps:

Firstly, we recover temperature and humidity from norm form, so we can study the direct effect of temperature and humidity on the bike rent amount. There are two variables represents temperature, *temp* and *atemp*. Corr(*temp* , *atemp*)=0.99, so we only use one variable, *atemp*.

Secondly, we include the variables which we are interested in, the reason why we choose these variables has been discussed. Here we want to introduce the season dummies to our model. There's three season dummies, *spring*, *summer* and *fall* (winter is the base) which are transferred from *season*.

The initial regression:
cnt~ yr+ holiday+ workingday+ atemp+ hum+ atemp2 + hum2+ windspeed+ goodTRUE+ fairTRUE+ springTRUE + summerTRUE + fallTRUE + wkAtemp + wkHum + wkWind。
In the output, there are three variables, fallTRUE, wkHum and wkWind are not significant so we exclude them in the next steps of regression.

Variable	Estimate	Std.error	T value	P(>|t|)
fallTRUE	-73.4300	119.5466	-0.614	0.539254
wkHum	-4.6420	4.4163	-1.051	0.293571
wkWind	-3.4370	12.3227	-0.279	0.780390

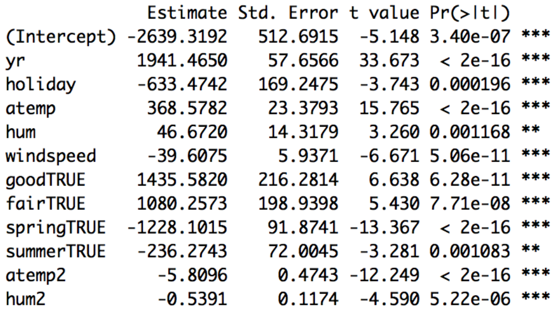
Then we go to formal model selection:

Number	rsquared	radjust	cp	aic	bic	cvss
12	0.84800895	0.84546870	17.29680	11790.22	11854.54	435941200
11	0.84751009	0.84517714	17.66748	11790.61	11850.34	435935311
12	0.84760166	0.84505461	19.23232	11792.17	11856.50	437127340

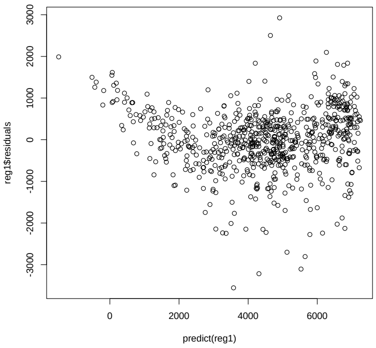
After comprehensive evaluation, we choose the underscored model, which eliminates *workingday* and *wkAtemp*.

2) Model Selection

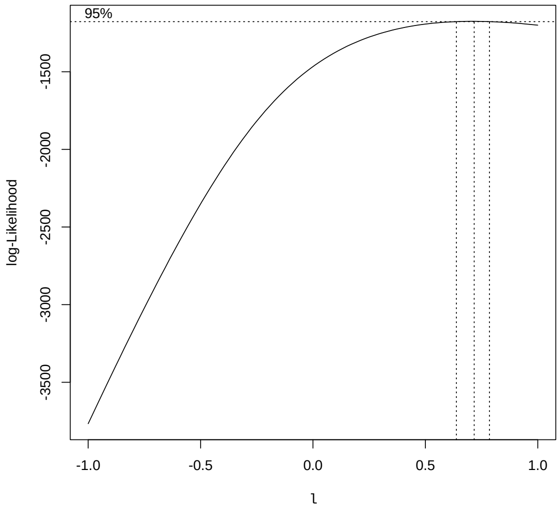
The second model goes to:
cnt~ yr+ holiday+ atemp+ hum+ atemp2 + hum2+ windspeed+ goodTRUE+ fairTRUE+ springTRUE + summerTRUE + fallTRUE。
Here's the regression coefficient (R-squared=0.8475) and residuals plot



```
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -2639.3192   512.6915  -5.148 3.40e-07 ***
yr           1941.4650    57.6566  33.673  < 2e-16 ***
holiday      -633.4742   169.2475  -3.743 0.000196 ***
atemp         368.5782    23.3793  15.765  < 2e-16 ***
hum            46.6720    14.3179   3.260 0.001168 ** 
windspeed     -39.6075     5.9371  -6.671 5.06e-11 ***
goodTRUE     1435.5820   216.2814   6.638 6.28e-11 ***
fairTRUE     1080.2573   198.9398   5.430 7.71e-08 ***
springTRUE  -1228.1015    91.8741 -13.367  < 2e-16 ***
summerTRUE   -236.2743    72.0045  -3.281 0.001083 ** 
atemp2         -5.8096     0.4743 -12.249  < 2e-16 ***
hum2           -0.5391     0.1174  -4.590 5.22e-06 ***
```

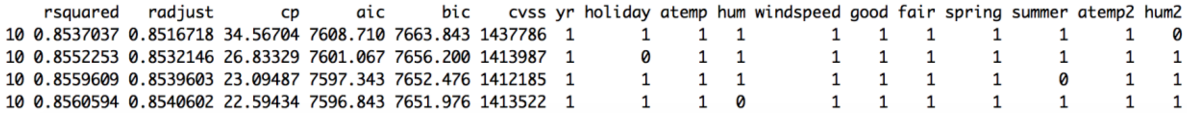


Because the residuals look like nonconstant and not asymptotically normal. So we try to do Box-Cox Transformation to adjust dependent variable for nonconstant variance and non-normality.



We choose λ=0.7 to transfer Y into Y^{λ}

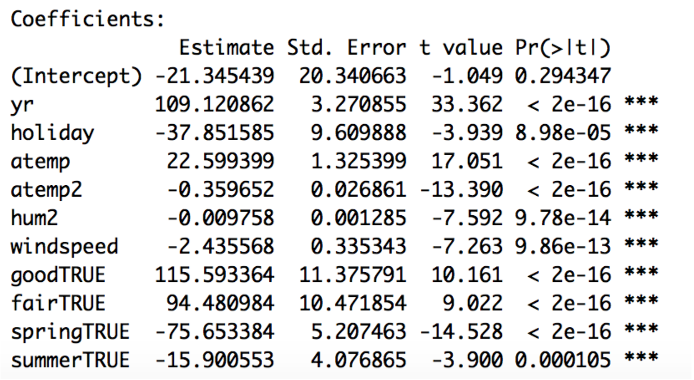
Then we go back to the model selection procedure and get the results below.



```
     rsquared    radjust       cp      aic      bic    cvss yr holiday atemp hum windspeed good fair spring summer atemp2 hum2
10 0.8537037 0.8516718 34.56704 7608.710 7663.843 1437786  1       1     1   1         1    1    1      1      1      1    0
10 0.8552253 0.8532146 26.83329 7601.067 7656.200 1413987  1       0     1   1         1    1    1      1      1      1    1
10 0.8559609 0.8539603 23.09487 7597.343 7652.476 1412185  1       1     1   1         1    1    1      1      0      1    1
10 0.8560594 0.8540602 22.59434 7596.843 7651.976 1413522  1       1     1   0         1    1    1      1      1      1    1
```

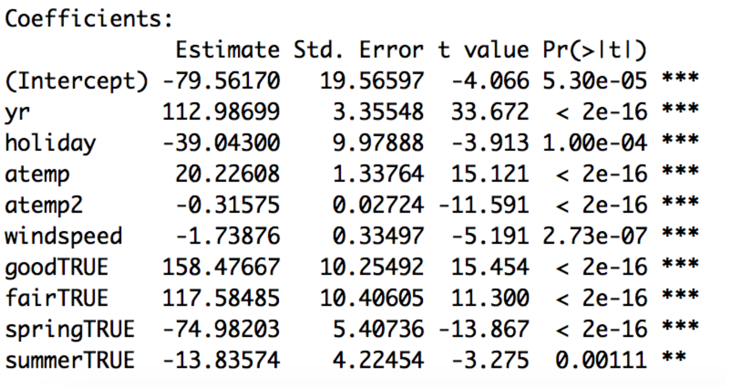
So we choose the last model, which excludes *hum*.

Here's the output of the new regression model.



```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -21.345439  20.340663  -1.049 0.294347    
yr          109.120862   3.270855  33.362  < 2e-16 ***
holiday     -37.851585   9.609888  -3.939 8.98e-05 ***
atemp        22.599399   1.325399  17.051  < 2e-16 ***
atemp2       -0.359652   0.026861 -13.390  < 2e-16 ***
hum2         -0.009758   0.001285  -7.592 9.78e-14 ***
windspeed    -2.435568   0.335343  -7.263 9.86e-13 ***
goodTRUE    115.593364  11.375791  10.161  < 2e-16 ***
fairTRUE     94.480984  10.471854   9.022  < 2e-16 ***
springTRUE  -75.653384   5.207463 -14.528  < 2e-16 ***
summerTRUE  -15.900553   4.076865  -3.900 0.000105 ***
```

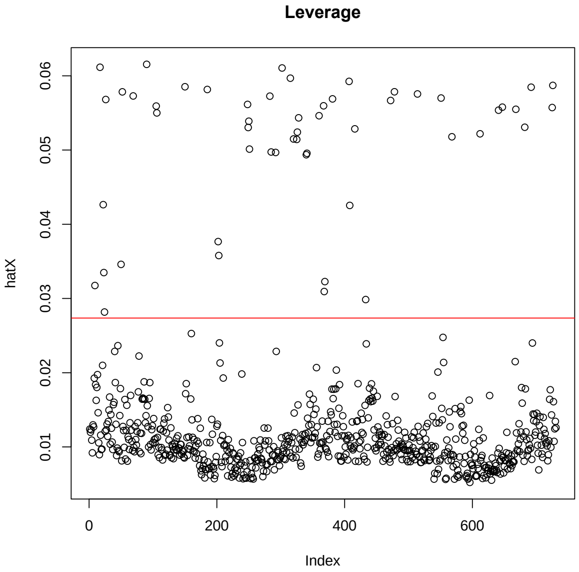
Because the coefficient on *hum2* is very small, which is hard to explain, so we'll eliminate it in this part to modify this model (Result is below):



```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -79.56170   19.56597  -4.066 5.30e-05 ***
yr          112.98699    3.35548  33.672  < 2e-16 ***
holiday     -39.04300    9.97888  -3.913 1.00e-04 ***
atemp        20.22608    1.33764  15.121  < 2e-16 ***
atemp2       -0.31575    0.02724 -11.591  < 2e-16 ***
windspeed    -1.73876    0.33497  -5.191 2.73e-07 ***
goodTRUE    158.47667   10.25492  15.454  < 2e-16 ***
fairTRUE    117.58485   10.40605  11.300  < 2e-16 ***
springTRUE  -74.98203    5.40736 -13.867  < 2e-16 ***
summerTRUE  -13.83574    4.22454  -3.275  0.00111 ** 
```

3) Outlier Detection

Then we enter into the outlier detection part, here we use leverage, studentized deleted residuals and cook's distance to detect outliers.



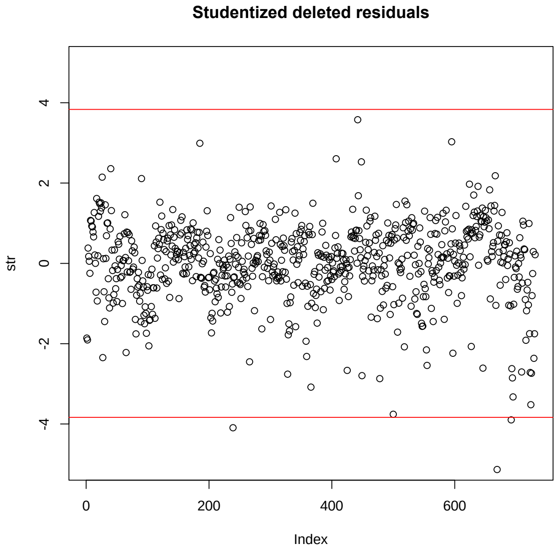
Because there are too many points above the criterion, we make hat.outlier>0.06 as the outlier bottom-line to shrink the scope of outliers.

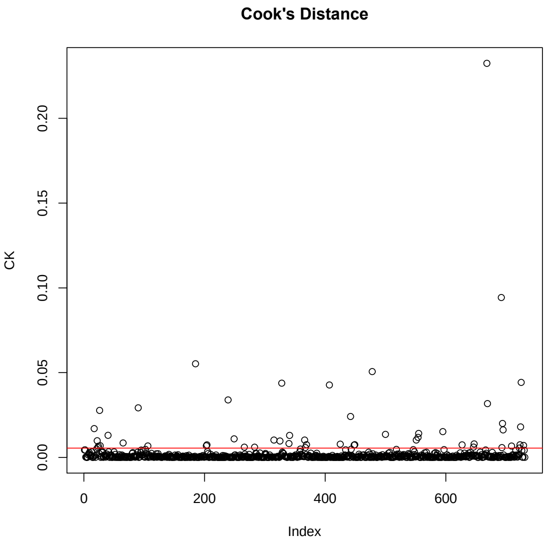
17	90	302
0.0612	0.0615	0.0611

Here are the outliers detected by calculating studentized deleted residuals:

239	668	669	692
-4.0935	-6.2897	-5.1341	-3.8969

Graph of studentized residuals and Cook's distance:





Cook’s distance outliers:

185	239	328	407	478	668	669	692	725
0.055	0.034	0.044	0.043	0.051	0.232	0.032	0.094	0.044

4) Model modification

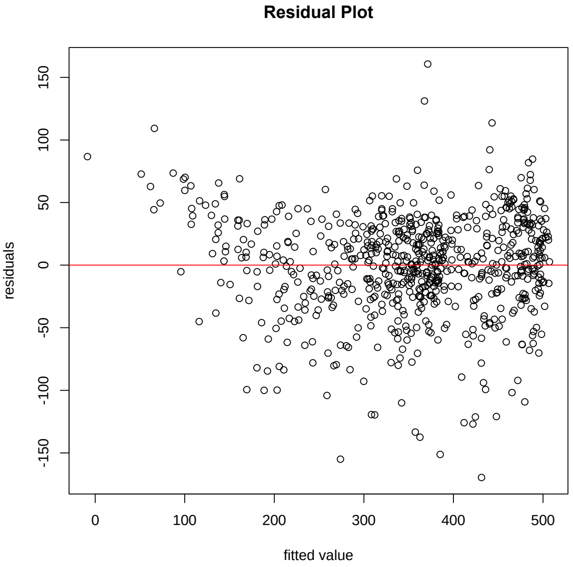
Then we delete the outliers from the dataset bike7, and perform model selection procedure.

```
   rsquared   radjust        cp      aic      bic    cvss yr holiday atemp windspeed good fair spring summer atemp2
7 0.8581820 0.8567858 56.56827 7424.550 7465.751 1284411  1       1     1         0    1    1      1      0      1
7 0.8620667 0.8607087 35.76219 7404.580 7445.781 1250366  1       0     1         0    1    1      1      1      1
7 0.8627022 0.8613504 32.35858 7401.260 7442.461 1244976  1       0     1         1    1    1      1      0      1
8 0.8635067 0.8619687 30.04965 7399.035 7444.813 1239455  1       1     1         0    1    1      1      1      1
8 0.8639679 0.8624351 27.57954 7396.601 7442.380 1235747  1       1     1         1    1    1      1      0      1
8 0.8663727 0.8648670 14.69963 7383.777 7429.556 1215256  1       0     1         1    1    1      1      1      1
```

We choose the last model which excludes *holiday*.

Intercept	yr	windspeed	atemp	atemp2
-78.3512***	115.2921***	-1.4737***	20.9246***	-0.3334***
(19.0)	(3.07)	(0.308)	(1.23)	(0.0250)
good	fair	spring	summer	
148.9304***	109.1599***	-79.6812***	-17.0833***	
(10.6)	(10.7)	(4.96)	(3.87)	

Then check the residuals again. The residual plot is suitable, so we’ve done all the part of building model.



We get the final model of *cnt*:

$$Y = \beta_0 + \beta_1 year + \beta_2 windspeed + \beta_3 atemp + \beta_4 atemp2 + \beta_5 good + \beta_6 fair + \beta_7 spring + \beta_8 summer$$