Project#2

me

4/23/2017

Project 2

1. Data Loading

```
#reading csv files

train = read.csv("train.csv", header = T)
test = read.csv("test.csv" ,header = T)

#create a seperate variable called "time" which includes afternoon(12pm-6pm), evening(6pm-12am), midnig

time.ori = train$datetime
time.string = as.character(time.ori)
list = c()


#this function detects whether a certain hourly time shows up in the datetime column and group them int
time.func = function(x){
  for (i in x){
    if (grepl("00:00:00",i) |grepl("01:00:00",i) |grepl("02:00:00",i)|grepl("03:00:00",i)|grepl("04:00:
      list = c(list,"midnight")
    }

    if (grepl("06:00:00",i) |grepl("07:00:00",i) |grepl("08:00:00",i)|grepl("09:00:00",i)|grepl("10:00:
      list = c(list,"sunrise")
    }

    if (grepl("12:00:00",i) |grepl("13:00:00",i) |grepl("14:00:00",i)|grepl("15:00:00",i)|grepl("16:00:
      list = c(list,"afternoon")
    }

    if (grepl("18:00:00",i) |grepl("19:00:00",i) |grepl("20:00:00",i)|grepl("21:00:00",i)|grepl("22:00:
      list = c(list,"evening")
    }
  }
  return(list)
}

#add the time.list into the train table and name it as "time"
time.list = time.func(time.string)
train$time = time.list
head(train)
```

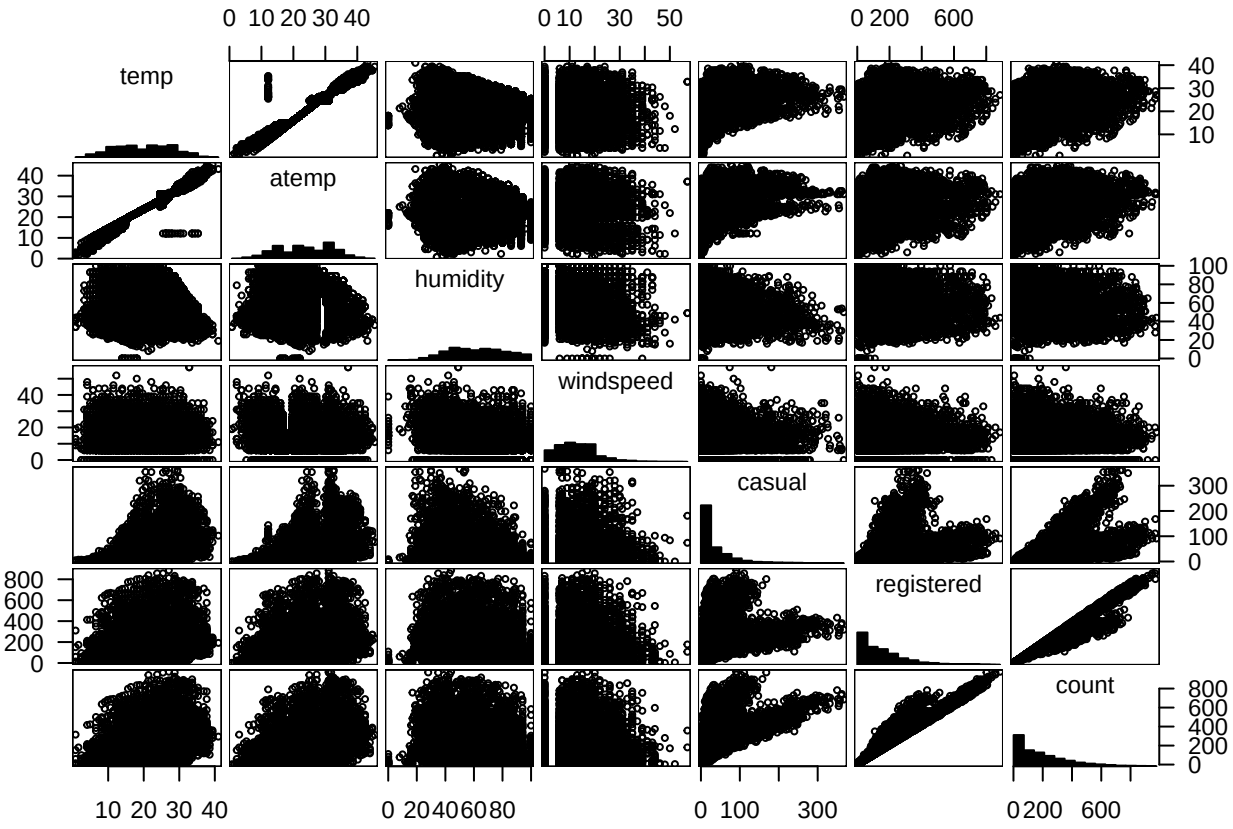
```
##              datetime season holiday workingday weather temp  atemp
## 1 2011-01-01 00:00:00      1       0          0       1 9.84 14.395
## 2 2011-01-01 01:00:00      1       0          0       1 9.02 13.635
```

```
## 3 2011-01-01 02:00:00      1       0          0       1 9.02 13.635
## 4 2011-01-01 03:00:00      1       0          0       1 9.84 14.395
## 5 2011-01-01 04:00:00      1       0          0       1 9.84 14.395
## 6 2011-01-01 05:00:00      1       0          0       2 9.84 12.880
##   humidity windspeed casual registered count     time
## 1       81    0.0000      3         13    16 midnight
## 2       80    0.0000      8         32    40 midnight
## 3       80    0.0000      5         27    32 midnight
## 4       75    0.0000      3         10    13 midnight
## 5       75    0.0000      0          1     1 midnight
## 6       75    6.0032      0          1     1 midnight
```

2. Exploratory Data Analysis

Pairs Scattered Plots

```
# pairplots
library(gpairs)
gpairs(train[,-c(1,2,3,4,5,13)])
```




```
#gpairs(train, c(2,3,4,5,13,12))
```

From the pairs plots above, we can see that the temp and atemp has a strong positive correlation and we might want to use only one of these two variables for our prediction.

Also, we can see that the correlation between registered counts and total counts is stronger than the correlation

between casual counts and total counts. We might want to investigate these two variables more to determine whether we want to predict them separately or together. As we can see in the scattered plots above, there are too many data points and the scattered plot just turnded out to be a huge blurb providing us very few information. Thus, I am going to plot 2D density smoothing plots to see where do most data points line.

2D Density Smooth Plots

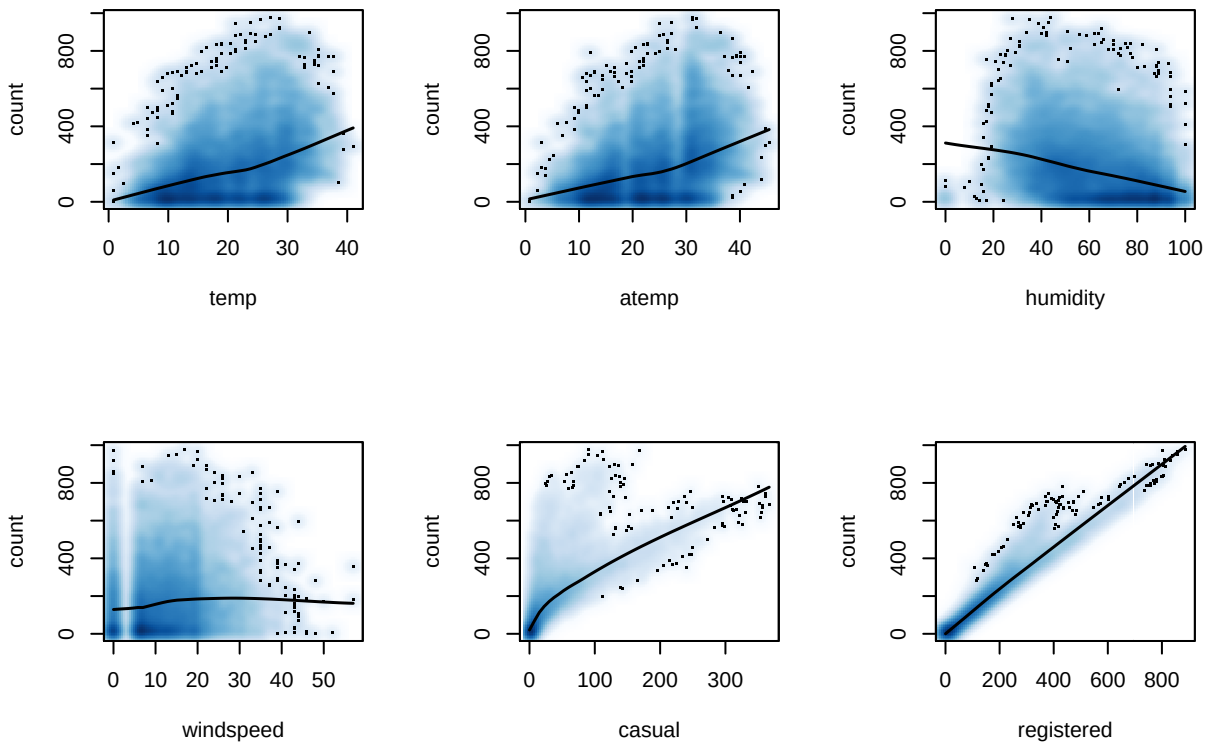
```
#create a smoothing plot function that loops over all pairs plots between "count" and other continous v

scatter.func = function(){
  for (i in c(6,7,8,9,10,11)) {
    x.lab = names(train)[i]

      smoothScatter(x = train[, x.lab], y = train$count, xlab = x.lab, ylab = "count")
      loess = loess.smooth(x = train[, x.lab], y = train$count)
      lines(loess,lwd = 1.5)

    }
  }

par(mfrow = c(2,3))
scatter.func()
```



Now, after plotting out the 2D density plots and add the loess smoothing curve on top, we have a much better idea of the relationship of these variables. As we can see from the plots, temp, atemp, casual and

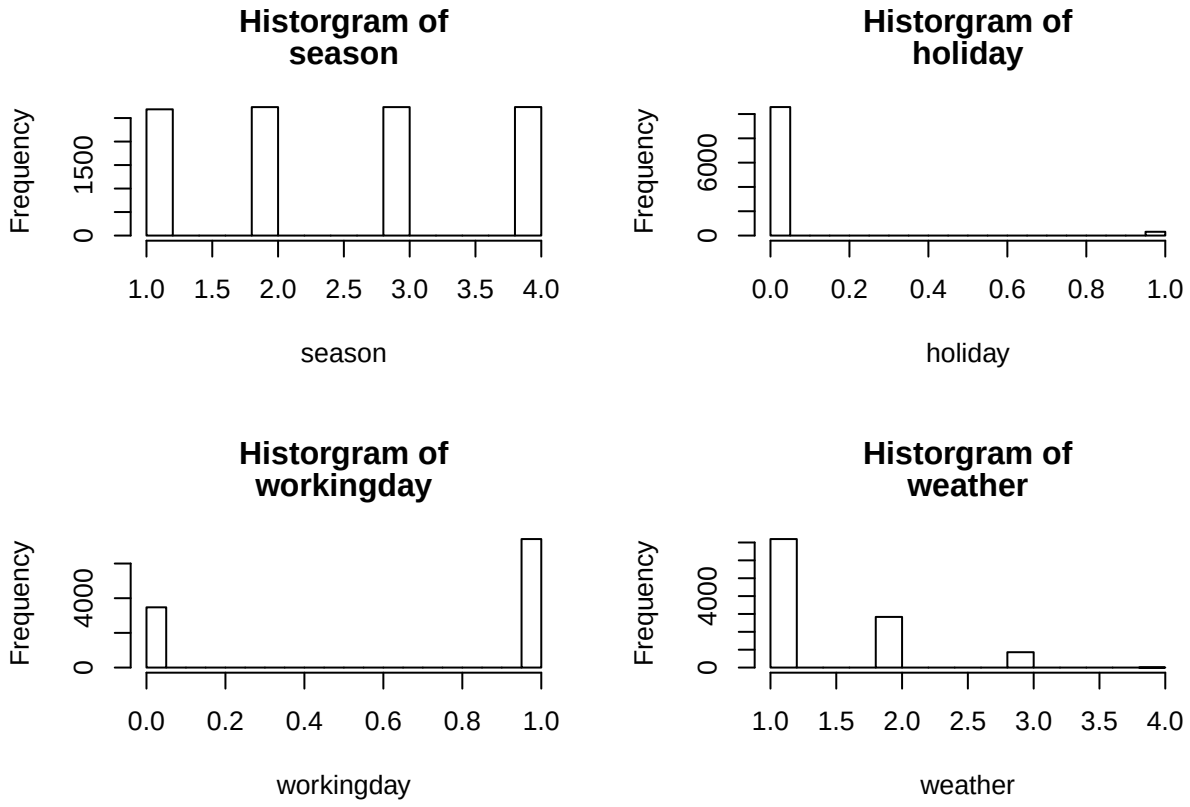
registered all have a positive relationship with count. Humidity has a negative relationship with count, which is reasonable. The loess line between windspeed and count is almost flat, which indicates that changes in windspeed doesn't really change the bike rental. Therefore, we could conclude that there is only a very weak relationship between windspeed and bike rental counts.

Histograms

Besides all these continuous variables above, we still haven't investigated the relationships between categorical variables. I am going to include some histograms below to further explore these relationships.

```
hist.func = function(){
  for (i in c(2,3,4,5)) {
    lab = names(train)[i]
    hist(x = train[, lab], xlab= lab, main = c("Historgram of", lab))
  }
}

par(mfrow= c(2,2))
hist.func()
```



As we can see from the histograms above, the bike rental counts is evenly distributed across all seasons. There are more bikes being rented on workingday and most bikes are rented on weather 1 (Clear, Few clouds, Partly cloudy, Partly cloudy).

3. Choice of Response in Regression

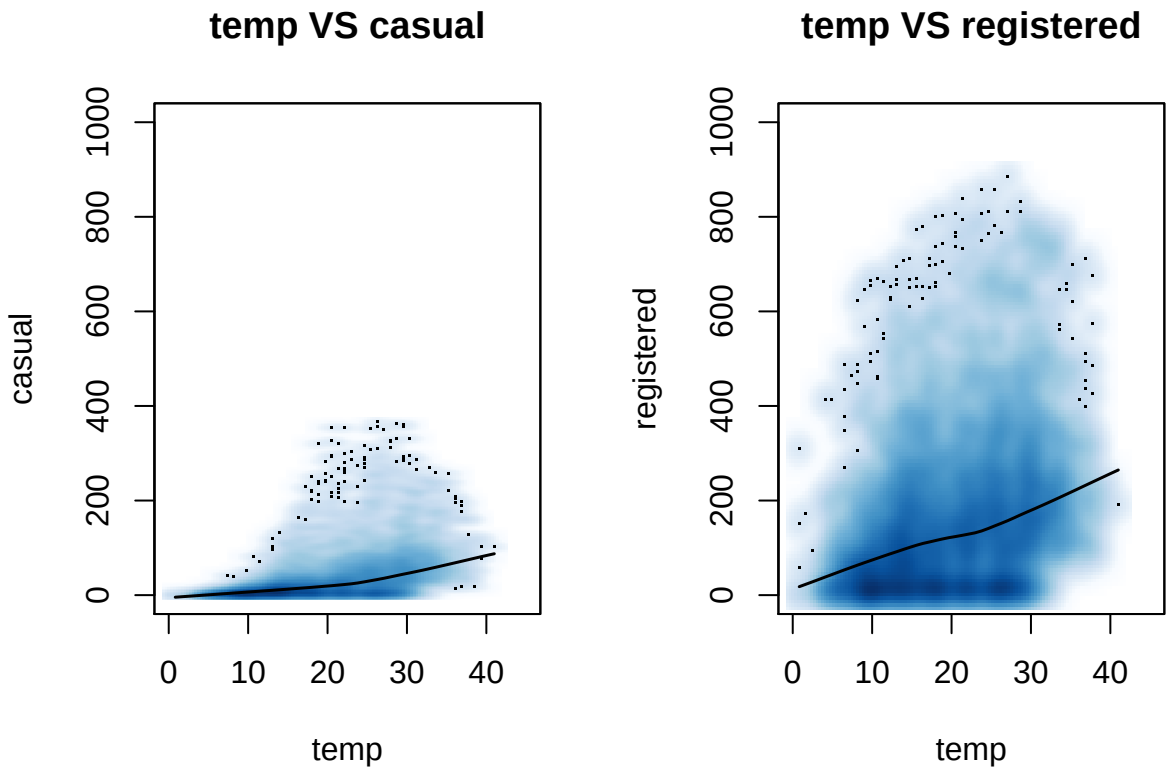
I make the choice base on whether casual users and registered users show different renting patterns, if so, then it would be more accurate for us to fit a linear regression for each group.

We can find out through visualization. As we can see the in the pairs plot at the very beginning. "casual VS temp" pair plot is noticeably different from "registered VS temp" pair plot. I would like to draw a 2D density plot and fit a loess curve to better visualize the differences.

```
par(mfrow = c(1,2))

smoothScatter(x = train$temp, y = train$casual, ylim = c(0,1000), xlim = c(0,45), xlab = "temp", ylab =
loess.casual = loess.smooth(x = train$temp, y = train$casual)
lines(loess.casual, lwd = 1.5)

smoothScatter(x = train$temp, y = train$registered, ylim = c(0,1000),xlim = c(0,45), xlab = "temp", ylal
loess.registered = loess.smooth(x = train$temp, y = train$registered)
lines(loess.registered, lwd = 1.5)
```



As we can see from the comparison above, the casual bikers are very different from the registered bikers when they react to temperature changes. Therefore, to better predict the bike rental count, we should keep them separate and make two linear regressions.

Also, from the plot aove, we can see that the moajority of data is located in low values, thus, it would be a good idea for us to take the log of counts to better fit the linear regression.

4. Regression Analysis

Basic Linear Regression Model

```
# create a "time" variable for test data as well

time.string.test = as.character(test$datetime)
time.list.test = time.func(time.string.test)
test$time = time.list.test
```

Here, I am going to first combine weather3(light rain) and weather4 (heavy rain& snow) before I run any linear regression because there's only one data point for weather4 = 1 and if I add the interaction term between weather4 and other variables, singularity issues will occur.

```
#predict using casual.lm and registered.lm.

#combine weather 3 and 4 by changing the row that contains weather = 4 into weather = 3

train$weather[5632] = 3

modify = function(){

}


#because I need to take the log of both casual and registered bikers, I need to take out the rows where

train.omit = train[train$casual != 0 & train$registered != 0, ]


# I create my own test dataset
sample.num = sample(c(1:nrow(train.omit)), 200)
my.test = train[sample.num,]


# now I fit a linear regression model for casual bikers and registered bikers separetely
casual.lm = lm(log(casual + 1) ~ as.factor(season) + as.factor(holiday) + as.factor(workingday) + as.fa

summary(casual.lm)
```

```
##
## Call:
## lm(formula = log(casual + 1) ~ as.factor(season) + as.factor(holiday) +
##     as.factor(workingday) + as.factor(weather) + temp + atemp +
##     humidity + windspeed + as.factor(time), data = train[-sample.num,
##     -c(11, 12)])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5186 -0.4748  0.0638  0.5298  3.1337
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)              2.672735   0.047575  56.180  < 2e-16 ***
## as.factor(season)2       0.367188   0.027716  13.248  < 2e-16 ***
```

```
## as.factor(season)3       0.059748   0.035589   1.679   0.0932 .
## as.factor(season)4       0.442748   0.023013  19.239  < 2e-16 ***
## as.factor(holiday)1     -0.240524   0.046364  -5.188 2.17e-07 ***
## as.factor(workingday)1  -0.730948   0.016657 -43.883  < 2e-16 ***
## as.factor(weather)2      0.009463   0.018392   0.515   0.6069
## as.factor(weather)3     -0.469589   0.031053 -15.122  < 2e-16 ***
## temp                     0.048723   0.006106   7.980 1.61e-15 ***
## atemp                    0.038018   0.005346   7.111 1.23e-12 ***
## humidity                -0.011556   0.000513 -22.528  < 2e-16 ***
## windspeed               -0.001284   0.001013  -1.268   0.2048
## as.factor(time)evening  -0.508510   0.022082 -23.029  < 2e-16 ***
## as.factor(time)midnight -2.090701   0.024794 -84.324  < 2e-16 ***
## as.factor(time)sunrise  -0.596941   0.023309 -25.610  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7747 on 10671 degrees of freedom
## Multiple R-squared:  0.7297, Adjusted R-squared:  0.7293
## F-statistic:  2057 on 14 and 10671 DF,  p-value: < 2.2e-16
```

```
registered.lm = lm(log(registered + 1) ~ as.factor(season) + as.factor(holiday) + as.factor(workingday)

summary(registered.lm)
```

```
##
## Call:
## lm(formula = log(registered + 1) ~ as.factor(season) + as.factor(holiday) +
##     as.factor(workingday) + as.factor(weather) + temp + atemp +
##     humidity + windspeed + as.factor(time), data = train[-sample.num,
##     -c(10, 12)])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.7289 -0.4946 -0.0118  0.5246  2.7096
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)              4.4539424  0.0491224  90.670  < 2e-16 ***
## as.factor(season)2       0.1946775  0.0286176   6.803 1.08e-11 ***
## as.factor(season)3       0.1141121  0.0367464   3.105  0.00191 **
## as.factor(season)4       0.5333576  0.0237616  22.446  < 2e-16 ***
## as.factor(holiday)1     -0.0508183  0.0478726  -1.062  0.28847
## as.factor(workingday)1   0.0735724  0.0171987   4.278 1.90e-05 ***
## as.factor(weather)2      0.0141194  0.0189899   0.744  0.45718
## as.factor(weather)3     -0.3861362  0.0320629 -12.043  < 2e-16 ***
## temp                     0.0292559  0.0063042   4.641 3.51e-06 ***
## atemp                    0.0106664  0.0055204   1.932  0.05337 .
## humidity                -0.0072029  0.0005297 -13.599  < 2e-16 ***
## windspeed               -0.0023347  0.0010456  -2.233  0.02558 *
## as.factor(time)evening  -0.1137063  0.0227998  -4.987 6.22e-07 ***
## as.factor(time)midnight -2.3906792  0.0256003 -93.385  < 2e-16 ***
## as.factor(time)sunrise  -0.1662078  0.0240668  -6.906 5.26e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 0.7999 on 10671 degrees of freedom
## Multiple R-squared:  0.6727, Adjusted R-squared:  0.6722
## F-statistic:  1566 on 14 and 10671 DF,  p-value: < 2.2e-16
```

```
count.lm = lm(log(count + 1) ~ as.factor(season) + as.factor(holiday) + as.factor(workingday) + as.facto

summary(count.lm)
```

```
##
## Call:
## lm(formula = log(count + 1) ~ as.factor(season) + as.factor(holiday) +
##     as.factor(workingday) + as.factor(weather) + temp + atemp +
##     humidity + windspeed + as.factor(time), data = train[-sample.num,
##     -c(10, 11)])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9640 -0.4792  0.0113  0.5134  2.8258
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)               4.6649693  0.0480838  97.017  < 2e-16 ***
## as.factor(season)2        0.2314935  0.0280126   8.264  < 2e-16 ***
## as.factor(season)3        0.1157881  0.0359695   3.219  0.00129 **
## as.factor(season)4        0.5248170  0.0232593  22.564  < 2e-16 ***
## as.factor(holiday)1      -0.0850019  0.0468605  -1.814  0.06972 .
## as.factor(workingday)1   -0.0755174  0.0168351  -4.486 7.34e-06 ***
## as.factor(weather)2       0.0106062  0.0185884   0.571  0.56829
## as.factor(weather)3      -0.4114003  0.0313851 -13.108  < 2e-16 ***
## temp                      0.0324668  0.0061710   5.261 1.46e-07 ***
## atemp                     0.0154123  0.0054037   2.852  0.00435 **
## humidity                 -0.0080247  0.0005185 -15.478  < 2e-16 ***
## windspeed                -0.0024031  0.0010235  -2.348  0.01889 *
## as.factor(time)evening   -0.1896753  0.0223178  -8.499  < 2e-16 ***
## as.factor(time)midnight  -2.4278772  0.0250591 -96.886  < 2e-16 ***
## as.factor(time)sunrise   -0.2250052  0.0235580  -9.551  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.783 on 10671 degrees of freedom
## Multiple R-squared:  0.695,  Adjusted R-squared:  0.6946
## F-statistic:  1737 on 14 and 10671 DF,  p-value: < 2.2e-16
```

I created two very basic linear regression models above. Now I am going to explore some interactions between different variables and pick the best model before I use step function and cross validation to further select my variables.

I am going to first incluse the interactions between numerical and categorical variables, if the RSS decreases significantly, I will continue to include interactions between all numericals & numericals, numericals & categorical and catrgorical & categorical. If not, I am going to stop and start variable selection.

Explore the interactions between different explanatory variables

Include interactions between numerical & categorical

```r
casual.lm.1 = lm(log(casual + 1) ~ as.factor(season) + as.factor(holiday) + as.factor(workingday) + as.
as.factor(holiday):temp + as.factor(holiday):humidity + as.factor(holiday):windspeed + as.factor(holida
as.factor(workingday):temp + as.factor(workingday):humidity + as.factor(workingday):windspeed + as.facto
as.factor(weather):temp + as.factor(weather):humidity + as.factor(weather):windspeed + as.factor(weathe
as.factor(time):temp + as.factor(time):humidity + as.factor(time):windspeed + as.factor(time):atemp
  , data = train[-sample.num,-c(11,12)])

summary(casual.lm.1)
```

```
##
## Call:
## lm(formula = log(casual + 1) ~ as.factor(season) + as.factor(holiday) +
##     as.factor(workingday) + as.factor(weather) + temp + atemp +
##     humidity + windspeed + as.factor(time) + as.factor(season):temp +
##     as.factor(season):humidity + as.factor(season):windspeed +
##     as.factor(season):atemp + as.factor(holiday):temp + as.factor(holiday):humidity +
##     as.factor(holiday):windspeed + as.factor(holiday):atemp +
##     as.factor(workingday):temp + as.factor(workingday):humidity +
##     as.factor(workingday):windspeed + as.factor(workingday):atemp +
##     as.factor(weather):temp + as.factor(weather):humidity + as.factor(weather):windspeed +
##     as.factor(weather):atemp + as.factor(time):temp + as.factor(time):humidity +
##     as.factor(time):windspeed + as.factor(time):atemp, data = train[-sample.num,
##     -c(11, 12)])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.3785 -0.4447  0.0551  0.5007  3.0531
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      2.1819940  0.1217850  17.917  < 2e-16
## as.factor(season)2               1.3840125  0.1374099  10.072  < 2e-16
## as.factor(season)3               2.5786588  0.2007694  12.844  < 2e-16
## as.factor(season)4               1.0134918  0.1207885   8.391  < 2e-16
## as.factor(holiday)1             -0.5089380  0.2446543  -2.080 0.037528
## as.factor(workingday)1          -0.6740940  0.0907256  -7.430 1.17e-13
## as.factor(weather)2             -0.0263271  0.0982049  -0.268 0.788640
## as.factor(weather)3             -0.6758423  0.1767763  -3.823 0.000133
## temp                             0.1408066  0.0260839   5.398 6.88e-08
## atemp                           -0.0095136  0.0224024  -0.425 0.671087
## humidity                        -0.0052545  0.0013048  -4.027 5.69e-05
## windspeed                       -0.0169200  0.0030419  -5.562 2.73e-08
## as.factor(time)evening          -1.1966610  0.1078810 -11.092  < 2e-16
## as.factor(time)midnight         -1.6893850  0.1183408 -14.276  < 2e-16
## as.factor(time)sunrise          -0.2719335  0.1129374  -2.408 0.016065
## as.factor(season)2:temp         -0.2225236  0.0271238  -8.204 2.59e-16
## as.factor(season)3:temp         -0.1809362  0.0229791  -7.874 3.77e-15
## as.factor(season)4:temp         -0.1599842  0.0286958  -5.575 2.53e-08
## as.factor(season)2:humidity     -0.0099653  0.0011457  -8.698  < 2e-16
## as.factor(season)3:humidity     -0.0148988  0.0014384 -10.358  < 2e-16
## as.factor(season)4:humidity     -0.0111778  0.0011797  -9.475  < 2e-16
## as.factor(season)2:windspeed     0.0181728  0.0030002   6.057 1.43e-09
## as.factor(season)3:windspeed     0.0204129  0.0031213   6.540 6.44e-11
## as.factor(season)4:windspeed     0.0121899  0.0031793   3.834 0.000127
```

```
## as.factor(season)2:atemp           0.1492237  0.0238961   6.245 4.41e-10
## as.factor(season)3:atemp           0.0803692  0.0191640   4.194 2.77e-05
## as.factor(season)4:atemp           0.1214190  0.0248570   4.885 1.05e-06
## as.factor(holiday)1:temp          -0.0098582  0.0523046  -0.188 0.850507
## as.factor(holiday)1:humidity       0.0051576  0.0028196   1.829 0.067402
## as.factor(holiday)1:windspeed     -0.0041740  0.0060140  -0.694 0.487670
## as.factor(holiday)1:atemp          0.0116711  0.0482395   0.242 0.808830
## as.factor(workingday)1:temp        0.0421240  0.0170473   2.471 0.013489
## as.factor(workingday)1:humidity    0.0051625  0.0008847   5.835 5.52e-09
## as.factor(workingday)1:windspeed   0.0037502  0.0021674   1.730 0.083611
## as.factor(workingday)1:atemp      -0.0542119  0.0155793  -3.480 0.000504
## as.factor(weather)2:temp           0.0100975  0.0142844   0.707 0.479650
## as.factor(weather)3:temp           0.0157303  0.0236532   0.665 0.506041
## as.factor(weather)2:humidity      -0.0001623  0.0010207  -0.159 0.873679
## as.factor(weather)3:humidity       0.0004688  0.0015823   0.296 0.767010
## as.factor(weather)2:windspeed     -0.0042502  0.0022894  -1.856 0.063413
## as.factor(weather)3:windspeed     -0.0031759  0.0032386  -0.981 0.326786
## as.factor(weather)2:atemp         -0.0045766  0.0131028  -0.349 0.726883
## as.factor(weather)3:atemp         -0.0051888  0.0217150  -0.239 0.811150
## temp:as.factor(time)evening        0.0544293  0.0157919   3.447 0.000570
## temp:as.factor(time)midnight       0.0380685  0.0160548   2.371 0.017750
## temp:as.factor(time)sunrise        0.0241660  0.0150579   1.605 0.108551
## humidity:as.factor(time)evening   -0.0026443  0.0011446  -2.310 0.020893
## humidity:as.factor(time)midnight   0.0004747  0.0012875   0.369 0.712356
## humidity:as.factor(time)sunrise   -0.0096551  0.0011940  -8.086 6.82e-16
## windspeed:as.factor(time)evening  -0.0019759  0.0025854  -0.764 0.444734
## windspeed:as.factor(time)midnight -0.0045223  0.0027770  -1.628 0.103455
## windspeed:as.factor(time)sunrise   0.0029773  0.0025902   1.149 0.250387
## atemp:as.factor(time)evening      -0.0122164  0.0144860  -0.843 0.399064
## atemp:as.factor(time)midnight     -0.0546814  0.0145790  -3.751 0.000177
## atemp:as.factor(time)sunrise      -0.0114826  0.0137048  -0.838 0.402133
##
## (Intercept)                       ***
## as.factor(season)2                ***
## as.factor(season)3                ***
## as.factor(season)4                ***
## as.factor(holiday)1               *
## as.factor(workingday)1            ***
## as.factor(weather)2
## as.factor(weather)3               ***
## temp                              ***
## atemp
## humidity                          ***
## windspeed                         ***
## as.factor(time)evening            ***
## as.factor(time)midnight           ***
## as.factor(time)sunrise            *
## as.factor(season)2:temp           ***
## as.factor(season)3:temp           ***
## as.factor(season)4:temp           ***
## as.factor(season)2:humidity       ***
## as.factor(season)3:humidity       ***
## as.factor(season)4:humidity       ***
## as.factor(season)2:windspeed      ***
```

```
## as.factor(season)3:windspeed      ***
## as.factor(season)4:windspeed      ***
## as.factor(season)2:atemp          ***
## as.factor(season)3:atemp          ***
## as.factor(season)4:atemp          ***
## as.factor(holiday)1:temp
## as.factor(holiday)1:humidity      .
## as.factor(holiday)1:windspeed
## as.factor(holiday)1:atemp
## as.factor(workingday)1:temp       *
## as.factor(workingday)1:humidity   ***
## as.factor(workingday)1:windspeed  .
## as.factor(workingday)1:atemp      ***
## as.factor(weather)2:temp
## as.factor(weather)3:temp
## as.factor(weather)2:humidity
## as.factor(weather)3:humidity
## as.factor(weather)2:windspeed     .
## as.factor(weather)3:windspeed
## as.factor(weather)2:atemp
## as.factor(weather)3:atemp
## temp:as.factor(time)evening       ***
## temp:as.factor(time)midnight      *
## temp:as.factor(time)sunrise
## humidity:as.factor(time)evening   *
## humidity:as.factor(time)midnight
## humidity:as.factor(time)sunrise   ***
## windspeed:as.factor(time)evening
## windspeed:as.factor(time)midnight
## windspeed:as.factor(time)sunrise
## atemp:as.factor(time)evening
## atemp:as.factor(time)midnight     ***
## atemp:as.factor(time)sunrise
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7324 on 10631 degrees of freedom
## Multiple R-squared:  0.7593, Adjusted R-squared:  0.7581
## F-statistic: 621.1 on 54 and 10631 DF,  p-value: < 2.2e-16
```

From the summaries above, we can see that for our very basic model without any interactions (causual.lm), the multiple r square is 0.6855. After adding all interactions between numerical and categorical variables, the multiple r square becomes 0.7198.

Now I am going to do the same thing for registered bikers.

```
registered.lm.1 = lm(log(registered+1) ~ as.factor(season) + as.factor(holiday) + as.factor(workingday)
as.factor(holiday):temp + as.factor(holiday):humidity + as.factor(holiday):windspeed + as.factor(holida
as.factor(workingday):temp + as.factor(workingday):humidity + as.factor(workingday):windspeed + as.facto
as.factor(time):temp + as.factor(time):humidity + as.factor(time):windspeed + as.factor(time):atemp , da



summary(registered.lm.1)
```

```
##
```

```
## Call:
## lm(formula = log(registered + 1) ~ as.factor(season) + as.factor(holiday) +
##     as.factor(workingday) + as.factor(weather) + temp + atemp +
##     humidity + windspeed + as.factor(time) + as.factor(season):temp +
##     as.factor(season):humidity + as.factor(season):windspeed +
##     as.factor(season):atemp + as.factor(holiday):temp + as.factor(holiday):humidity +
##     as.factor(holiday):windspeed + as.factor(holiday):atemp +
##     as.factor(workingday):temp + as.factor(workingday):humidity +
##     as.factor(workingday):windspeed + as.factor(workingday):atemp +
##     as.factor(time):temp + as.factor(time):humidity + as.factor(time):windspeed +
##     as.factor(time):atemp, data = train[-sample.num, -c(10, 12)])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.7916 -0.4757 -0.0126  0.4990  2.6528
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      4.4274853  0.1282305  34.528  < 2e-16
## as.factor(season)2               0.4212761  0.1453620   2.898 0.003762
## as.factor(season)3               1.5487540  0.2133672   7.259 4.18e-13
## as.factor(season)4               1.2914756  0.1291319  10.001  < 2e-16
## as.factor(holiday)1             -0.4528031  0.2618925  -1.729 0.083844
## as.factor(workingday)1          -0.1410292  0.0970140  -1.454 0.146059
## as.factor(weather)2              0.0141319  0.0188696   0.749 0.453920
## as.factor(weather)3             -0.3682107  0.0320785 -11.478  < 2e-16
## temp                             0.1297084  0.0278287   4.661 3.19e-06
## atemp                           -0.0627329  0.0238952  -2.625 0.008669
## humidity                        -0.0044928  0.0013236  -3.394 0.000690
## windspeed                       -0.0159139  0.0032147  -4.950 7.52e-07
## as.factor(time)evening          -0.3663878  0.1147608  -3.193 0.001414
## as.factor(time)midnight         -2.2567877  0.1255447 -17.976  < 2e-16
## as.factor(time)sunrise          -0.2774994  0.1200048  -2.312 0.020775
## as.factor(season)2:temp         -0.1169679  0.0290968  -4.020 5.86e-05
## as.factor(season)3:temp         -0.1244147  0.0246584  -5.046 4.60e-07
## as.factor(season)4:temp         -0.0742486  0.0307280  -2.416 0.015695
## as.factor(season)2:humidity     -0.0052540  0.0011833  -4.440 9.08e-06
## as.factor(season)3:humidity     -0.0087707  0.0014537  -6.033 1.66e-09
## as.factor(season)4:humidity     -0.0074378  0.0012499  -5.951 2.76e-09
## as.factor(season)2:windspeed     0.0085687  0.0032134   2.667 0.007676
## as.factor(season)3:windspeed     0.0157419  0.0033267   4.732 2.25e-06
## as.factor(season)4:windspeed     0.0047205  0.0033975   1.389 0.164731
## as.factor(season)2:atemp         0.0890957  0.0256334   3.476 0.000511
## as.factor(season)3:atemp         0.0631298  0.0205668   3.069 0.002150
## as.factor(season)4:atemp         0.0387764  0.0266223   1.457 0.145273
## as.factor(holiday)1:temp        -0.1088055  0.0561354  -1.938 0.052617
## as.factor(holiday)1:humidity     0.0023434  0.0030222   0.775 0.438135
## as.factor(holiday)1:windspeed    0.0176885  0.0064377   2.748 0.006013
## as.factor(holiday)1:atemp        0.0970108  0.0517719   1.874 0.060983
## as.factor(workingday)1:temp     -0.0214485  0.0182856  -1.173 0.240833
## as.factor(workingday)1:humidity  0.0049279  0.0009436   5.223 1.80e-07
## as.factor(workingday)1:windspeed 0.0096302  0.0023250   4.142 3.47e-05
## as.factor(workingday)1:atemp     0.0094420  0.0167113   0.565 0.572079
## temp:as.factor(time)evening      0.0229471  0.0156373   1.467 0.142281
```

```
## temp:as.factor(time)midnight       0.0537238  0.0171932   3.125 0.001785
## temp:as.factor(time)sunrise        0.0162840  0.0161388   1.009 0.312997
## humidity:as.factor(time)evening   -0.0062285  0.0012240  -5.089 3.67e-07
## humidity:as.factor(time)midnight  -0.0040337  0.0013698  -2.945 0.003239
## humidity:as.factor(time)sunrise   -0.0000263  0.0012741  -0.021 0.983533
## windspeed:as.factor(time)evening   0.0004072  0.0027612   0.147 0.882755
## windspeed:as.factor(time)midnight -0.0122060  0.0029739  -4.104 4.08e-05
## windspeed:as.factor(time)sunrise  -0.0018272  0.0027738  -0.659 0.510068
## atemp:as.factor(time)evening       0.0045692  0.0143244   0.319 0.749748
## atemp:as.factor(time)midnight     -0.0360549  0.0156231  -2.308 0.021029
## atemp:as.factor(time)sunrise      -0.0105769  0.0146899  -0.720 0.471532
##
## (Intercept)                        ***
## as.factor(season)2                 **
## as.factor(season)3                 ***
## as.factor(season)4                 ***
## as.factor(holiday)1                .
## as.factor(workingday)1
## as.factor(weather)2
## as.factor(weather)3                ***
## temp                               ***
## atemp                              **
## humidity                           ***
## windspeed                          ***
## as.factor(time)evening             **
## as.factor(time)midnight            ***
## as.factor(time)sunrise             *
## as.factor(season)2:temp            ***
## as.factor(season)3:temp            ***
## as.factor(season)4:temp            *
## as.factor(season)2:humidity        ***
## as.factor(season)3:humidity        ***
## as.factor(season)4:humidity        ***
## as.factor(season)2:windspeed       **
## as.factor(season)3:windspeed       ***
## as.factor(season)4:windspeed
## as.factor(season)2:atemp           ***
## as.factor(season)3:atemp           **
## as.factor(season)4:atemp
## as.factor(holiday)1:temp           .
## as.factor(holiday)1:humidity
## as.factor(holiday)1:windspeed      **
## as.factor(holiday)1:atemp          .
## as.factor(workingday)1:temp
## as.factor(workingday)1:humidity    ***
## as.factor(workingday)1:windspeed   ***
## as.factor(workingday)1:atemp
## temp:as.factor(time)evening
## temp:as.factor(time)midnight       **
## temp:as.factor(time)sunrise
## humidity:as.factor(time)evening    ***
## humidity:as.factor(time)midnight   **
## humidity:as.factor(time)sunrise
## windspeed:as.factor(time)evening
```

```
## windspeed:as.factor(time)midnight ***
## windspeed:as.factor(time)sunrise
## atemp:as.factor(time)evening
## atemp:as.factor(time)midnight     *
## atemp:as.factor(time)sunrise
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7865 on 10639 degrees of freedom
## Multiple R-squared:  0.6845, Adjusted R-squared:  0.6832
## F-statistic: 501.9 on 46 and 10639 DF,  p-value: < 2.2e-16
```

From the summaries above, we can see that for our very basic model without any interactions (registered.lm), the multiple r square is 0.6016. After adding all interactions between numerical and categorical variables, the multiple r square becomes 0.6184.

Predict on test data and compare RSS

To see the effect of adding interaction on our prediction. I am going to predict the hourly bike rental counts using both models and compare the RSS.

```
predicted.casual.1 = predict.lm(casual.lm.1, newdata = my.test[,-c(10,11,12)])
predicted.registered.1 = predict.lm(registered.lm.1, newdata = my.test[,-c(10,11,12)])
predicted.sum.1 = predicted.casual.1 + predicted.registered.1
true.value = log(my.test$count)
RSS.1 = sum((true.value-predicted.sum.1)^2)



RSS.1
```

```
## [1] 1718.551
```

```
predicted.casual = predict.lm(casual.lm, newdata = my.test[,-c(10,11,12)])
predicted.registered= predict.lm(registered.lm, newdata = my.test[,-c(10,11,12)])
predicted.sum = predicted.casual + predicted.registered
RSS = sum((true.value-predicted.sum)^2)

RSS
```

```
## [1] 1772.943
```

```
(RSS - RSS.1)/RSS
```

```
## [1] 0.03067887
```

As we can see from the RSS values above. The RSS only decreases 2% after I include all interaction terms. This does not seem to be a significant decrease. Therefore,I decide to stop here and not include the interactions between all numericals & numericals, numericals & categorical and catrgorical & categorical.

5. Variable Selection

```
casual.lm.short = step(casual.lm.1, direction = "backward")
```

```
## Start:  AIC=-6601.13
## log(casual + 1) ~ as.factor(season) + as.factor(holiday) + as.factor(workingday) +
```

```
##     as.factor(weather) + temp + atemp + humidity + windspeed +
##     as.factor(time) + as.factor(season):temp + as.factor(season):humidity +
##     as.factor(season):windspeed + as.factor(season):atemp + as.factor(holiday):temp +
##     as.factor(holiday):humidity + as.factor(holiday):windspeed +
##     as.factor(holiday):atemp + as.factor(workingday):temp + as.factor(workingday):humidity +
##     as.factor(workingday):windspeed + as.factor(workingday):atemp +
##     as.factor(weather):temp + as.factor(weather):humidity + as.factor(weather):windspeed +
##     as.factor(weather):atemp + as.factor(time):temp + as.factor(time):humidity +
##     as.factor(time):windspeed + as.factor(time):atemp
##
##                                    Df Sum of Sq    RSS     AIC
## - as.factor(weather):humidity       2     0.075 5702.6 -6605.0
## - as.factor(weather):atemp          2     0.078 5702.6 -6605.0
## - as.factor(weather):temp           2     0.398 5702.9 -6604.4
## - as.factor(holiday):temp           1     0.019 5702.5 -6603.1
## - as.factor(holiday):atemp          1     0.031 5702.5 -6603.1
## - as.factor(holiday):windspeed      1     0.258 5702.8 -6602.6
## - as.factor(weather):windspeed      2     2.054 5704.6 -6601.3
## <none>                                    5702.5 -6601.1
## - as.factor(workingday):windspeed   1     1.606 5704.1 -6600.1
## - windspeed:as.factor(time)         3     3.901 5706.4 -6599.8
## - as.factor(holiday):humidity       1     1.795 5704.3 -6599.8
## - as.factor(workingday):temp        1     3.275 5705.8 -6597.0
## - temp:as.factor(time)              3     7.237 5709.7 -6593.6
## - atemp:as.factor(time)             3     7.840 5710.3 -6592.5
## - as.factor(workingday):atemp       1     6.495 5709.0 -6591.0
## - as.factor(workingday):humidity    1    18.266 5720.8 -6569.0
## - as.factor(season):atemp           3    23.653 5726.1 -6562.9
## - as.factor(season):windspeed       3    27.271 5729.8 -6556.2
## - as.factor(season):temp            3    40.404 5742.9 -6531.7
## - humidity:as.factor(time)          3    46.366 5748.9 -6520.6
## - as.factor(season):humidity        3    78.531 5781.0 -6461.0
##
## Step:  AIC=-6604.99
## log(casual + 1) ~ as.factor(season) + as.factor(holiday) + as.factor(workingday) +
##     as.factor(weather) + temp + atemp + humidity + windspeed +
##     as.factor(time) + as.factor(season):temp + as.factor(season):humidity +
##     as.factor(season):windspeed + as.factor(season):atemp + as.factor(holiday):temp +
##     as.factor(holiday):humidity + as.factor(holiday):windspeed +
##     as.factor(holiday):atemp + as.factor(workingday):temp + as.factor(workingday):humidity +
##     as.factor(workingday):windspeed + as.factor(workingday):atemp +
##     as.factor(weather):temp + as.factor(weather):windspeed +
##     as.factor(weather):atemp + temp:as.factor(time) + humidity:as.factor(time) +
##     windspeed:as.factor(time) + atemp:as.factor(time)
##
##                                    Df Sum of Sq    RSS     AIC
## - as.factor(weather):atemp          2     0.091 5702.7 -6608.8
## - as.factor(weather):temp           2     0.435 5703.0 -6608.2
## - as.factor(holiday):temp           1     0.018 5702.6 -6607.0
## - as.factor(holiday):atemp          1     0.030 5702.6 -6606.9
## - as.factor(holiday):windspeed      1     0.268 5702.8 -6606.5
## <none>                                    5702.6 -6605.0
## - as.factor(weather):windspeed      2     2.189 5704.8 -6604.9
## - as.factor(workingday):windspeed   1     1.609 5704.2 -6604.0
```

```
## - windspeed:as.factor(time)        3     3.881 5706.5 -6603.7
## - as.factor(holiday):humidity      1     1.786 5704.4 -6603.6
## - as.factor(workingday):temp       1     3.282 5705.9 -6600.8
## - temp:as.factor(time)             3     7.208 5709.8 -6597.5
## - atemp:as.factor(time)            3     7.822 5710.4 -6596.3
## - as.factor(workingday):atemp      1     6.502 5709.1 -6594.8
## - as.factor(workingday):humidity   1    18.630 5721.2 -6572.1
## - as.factor(season):atemp          3    23.674 5726.2 -6566.7
## - as.factor(season):windspeed      3    27.281 5729.9 -6560.0
## - as.factor(season):temp           3    40.433 5743.0 -6535.5
## - humidity:as.factor(time)         3    46.555 5749.1 -6524.1
## - as.factor(season):humidity       3    80.276 5782.8 -6461.6
##
## Step:  AIC=-6608.82
## log(casual + 1) ~ as.factor(season) + as.factor(holiday) + as.factor(workingday) +
##     as.factor(weather) + temp + atemp + humidity + windspeed +
##     as.factor(time) + as.factor(season):temp + as.factor(season):humidity +
##     as.factor(season):windspeed + as.factor(season):atemp + as.factor(holiday):temp +
##     as.factor(holiday):humidity + as.factor(holiday):windspeed +
##     as.factor(holiday):atemp + as.factor(workingday):temp + as.factor(workingday):humidity +
##     as.factor(workingday):windspeed + as.factor(workingday):atemp +
##     as.factor(weather):temp + as.factor(weather):windspeed +
##     temp:as.factor(time) + humidity:as.factor(time) + windspeed:as.factor(time) +
##     atemp:as.factor(time)
##
##                                   Df Sum of Sq    RSS     AIC
## - as.factor(holiday):temp          1     0.018 5702.7 -6610.8
## - as.factor(holiday):atemp         1     0.030 5702.7 -6610.8
## - as.factor(holiday):windspeed     1     0.274 5702.9 -6610.3
## - as.factor(weather):windspeed     2     2.111 5704.8 -6608.9
## <none>                                       5702.7 -6608.8
## - as.factor(workingday):windspeed  1     1.600 5704.3 -6607.8
## - windspeed:as.factor(time)        3     3.898 5706.6 -6607.5
## - as.factor(holiday):humidity      1     1.786 5704.4 -6607.5
## - as.factor(workingday):temp       1     3.322 5706.0 -6604.6
## - as.factor(weather):temp          2     4.777 5707.4 -6603.9
## - atemp:as.factor(time)            3     7.748 5710.4 -6600.3
## - temp:as.factor(time)             3     8.619 5711.3 -6598.7
## - as.factor(workingday):atemp      1     6.558 5709.2 -6598.5
## - as.factor(workingday):humidity   1    18.726 5721.4 -6575.8
## - as.factor(season):atemp          3    23.683 5726.3 -6570.5
## - as.factor(season):windspeed      3    27.327 5730.0 -6563.7
## - as.factor(season):temp           3    40.581 5743.2 -6539.0
## - humidity:as.factor(time)         3    46.578 5749.2 -6527.9
## - as.factor(season):humidity       3    80.191 5782.9 -6465.6
##
## Step:  AIC=-6610.79
## log(casual + 1) ~ as.factor(season) + as.factor(holiday) + as.factor(workingday) +
##     as.factor(weather) + temp + atemp + humidity + windspeed +
##     as.factor(time) + as.factor(season):temp + as.factor(season):humidity +
##     as.factor(season):windspeed + as.factor(season):atemp + as.factor(holiday):humidity +
##     as.factor(holiday):windspeed + as.factor(holiday):atemp +
##     as.factor(workingday):temp + as.factor(workingday):humidity +
##     as.factor(workingday):windspeed + as.factor(workingday):atemp +
```

```
##     as.factor(weather):temp + as.factor(weather):windspeed +
##     temp:as.factor(time) + humidity:as.factor(time) + windspeed:as.factor(time) +
##     atemp:as.factor(time)
##
##                                    Df Sum of Sq    RSS     AIC
## - as.factor(holiday):atemp          1     0.153 5702.8 -6612.5
## - as.factor(holiday):windspeed      1     0.331 5703.0 -6612.2
## - as.factor(weather):windspeed      2     2.107 5704.8 -6610.8
## <none>                                        5702.7 -6610.8
## - as.factor(workingday):windspeed   1     1.584 5704.3 -6609.8
## - windspeed:as.factor(time)         3     3.885 5706.6 -6609.5
## - as.factor(holiday):humidity       1     1.947 5704.6 -6609.1
## - as.factor(workingday):temp        1     3.674 5706.4 -6605.9
## - as.factor(weather):temp           2     4.787 5707.5 -6605.8
## - atemp:as.factor(time)             3     7.740 5710.4 -6602.3
## - temp:as.factor(time)              3     8.613 5711.3 -6600.7
## - as.factor(workingday):atemp       1     7.167 5709.8 -6599.4
## - as.factor(workingday):humidity    1    18.744 5721.4 -6577.7
## - as.factor(season):atemp           3    23.668 5726.3 -6572.5
## - as.factor(season):windspeed       3    27.355 5730.0 -6565.6
## - as.factor(season):temp            3    40.563 5743.2 -6541.0
## - humidity:as.factor(time)          3    46.561 5749.2 -6529.9
## - as.factor(season):humidity        3    80.179 5782.9 -6467.6
##
## Step:  AIC=-6612.5
## log(casual + 1) ~ as.factor(season) + as.factor(holiday) + as.factor(workingday) +
##     as.factor(weather) + temp + atemp + humidity + windspeed +
##     as.factor(time) + as.factor(season):temp + as.factor(season):humidity +
##     as.factor(season):windspeed + as.factor(season):atemp + as.factor(holiday):humidity +
##     as.factor(holiday):windspeed + as.factor(workingday):temp +
##     as.factor(workingday):humidity + as.factor(workingday):windspeed +
##     as.factor(workingday):atemp + as.factor(weather):temp + as.factor(weather):windspeed +
##     temp:as.factor(time) + humidity:as.factor(time) + windspeed:as.factor(time) +
##     atemp:as.factor(time)
##
##                                    Df Sum of Sq    RSS     AIC
## - as.factor(holiday):windspeed      1     0.419 5703.3 -6613.7
## - as.factor(weather):windspeed      2     2.127 5705.0 -6612.5
## <none>                                        5702.8 -6612.5
## - as.factor(workingday):windspeed   1     1.561 5704.4 -6611.6
## - windspeed:as.factor(time)         3     3.861 5706.7 -6611.3
## - as.factor(holiday):humidity       1     2.115 5704.9 -6610.5
## - as.factor(workingday):temp        1     3.628 5706.5 -6607.7
## - as.factor(weather):temp           2     4.869 5707.7 -6607.4
## - atemp:as.factor(time)             3     7.735 5710.6 -6604.0
## - temp:as.factor(time)              3     8.590 5711.4 -6602.4
## - as.factor(workingday):atemp       1     7.170 5710.0 -6601.1
## - as.factor(workingday):humidity    1    18.665 5721.5 -6579.6
## - as.factor(season):atemp           3    23.595 5726.4 -6574.4
## - as.factor(season):windspeed       3    27.300 5730.1 -6567.5
## - as.factor(season):temp            3    40.494 5743.3 -6542.9
## - humidity:as.factor(time)          3    46.533 5749.4 -6531.7
## - as.factor(season):humidity        3    80.471 5783.3 -6468.8
##
```

```
## Step:  AIC=-6613.71
## log(casual + 1) ~ as.factor(season) + as.factor(holiday) + as.factor(workingday) +
##     as.factor(weather) + temp + atemp + humidity + windspeed +
##     as.factor(time) + as.factor(season):temp + as.factor(season):humidity +
##     as.factor(season):windspeed + as.factor(season):atemp + as.factor(holiday):humidity +
##     as.factor(workingday):temp + as.factor(workingday):humidity +
##     as.factor(workingday):windspeed + as.factor(workingday):atemp +
##     as.factor(weather):temp + as.factor(weather):windspeed +
##     temp:as.factor(time) + humidity:as.factor(time) + windspeed:as.factor(time) +
##     atemp:as.factor(time)
##
##                                    Df Sum of Sq    RSS     AIC
## - as.factor(weather):windspeed      2     2.072 5705.3 -6613.8
## <none>                                          5703.3 -6613.7
## - windspeed:as.factor(time)         3     3.886 5707.1 -6612.4
## - as.factor(workingday):windspeed   1     2.148 5705.4 -6611.7
## - as.factor(holiday):humidity       1     3.405 5706.7 -6609.3
## - as.factor(workingday):temp        1     3.572 5706.8 -6609.0
## - as.factor(weather):temp           2     4.855 5708.1 -6608.6
## - atemp:as.factor(time)             3     7.796 5711.1 -6605.1
## - temp:as.factor(time)              3     8.619 5711.9 -6603.6
## - as.factor(workingday):atemp       1     7.102 5710.4 -6602.4
## - as.factor(workingday):humidity    1    19.260 5722.5 -6579.7
## - as.factor(season):atemp           3    23.417 5726.7 -6575.9
## - as.factor(season):windspeed       3    27.011 5730.3 -6569.2
## - as.factor(season):temp            3    40.240 5743.5 -6544.6
## - humidity:as.factor(time)          3    46.480 5749.7 -6533.0
## - as.factor(season):humidity        3    80.451 5783.7 -6470.0
##
## Step:  AIC=-6613.83
## log(casual + 1) ~ as.factor(season) + as.factor(holiday) + as.factor(workingday) +
##     as.factor(weather) + temp + atemp + humidity + windspeed +
##     as.factor(time) + as.factor(season):temp + as.factor(season):humidity +
##     as.factor(season):windspeed + as.factor(season):atemp + as.factor(holiday):humidity +
##     as.factor(workingday):temp + as.factor(workingday):humidity +
##     as.factor(workingday):windspeed + as.factor(workingday):atemp +
##     as.factor(weather):temp + temp:as.factor(time) + humidity:as.factor(time) +
##     windspeed:as.factor(time) + atemp:as.factor(time)
##
##                                    Df Sum of Sq    RSS     AIC
## <none>                                          5705.3 -6613.8
## - windspeed:as.factor(time)         3     3.808 5709.1 -6612.7
## - as.factor(workingday):windspeed   1     2.087 5707.4 -6611.9
## - as.factor(holiday):humidity       1     3.147 5708.5 -6609.9
## - as.factor(workingday):temp        1     3.680 5709.0 -6608.9
## - as.factor(weather):temp           2     4.826 5710.2 -6608.8
## - atemp:as.factor(time)             3     7.844 5713.2 -6605.2
## - temp:as.factor(time)              3     8.644 5714.0 -6603.7
## - as.factor(workingday):atemp       1     7.288 5712.6 -6602.2
## - as.factor(workingday):humidity    1    19.182 5724.5 -6580.0
## - as.factor(season):atemp           3    23.123 5728.4 -6576.6
## - as.factor(season):windspeed       3    26.568 5731.9 -6570.2
## - as.factor(season):temp            3    40.080 5745.4 -6545.0
## - humidity:as.factor(time)          3    46.461 5751.8 -6533.2
```

```
## - as.factor(season):humidity       3    80.830 5786.2 -6469.5
```

```
registered.lm.short = step(registered.lm.1, direction = "backward")
```

```
## Start:  AIC=-5086.45
## log(registered + 1) ~ as.factor(season) + as.factor(holiday) +
##     as.factor(workingday) + as.factor(weather) + temp + atemp +
##     humidity + windspeed + as.factor(time) + as.factor(season):temp +
##     as.factor(season):humidity + as.factor(season):windspeed +
##     as.factor(season):atemp + as.factor(holiday):temp + as.factor(holiday):humidity +
##     as.factor(holiday):windspeed + as.factor(holiday):atemp +
##     as.factor(workingday):temp + as.factor(workingday):humidity +
##     as.factor(workingday):windspeed + as.factor(workingday):atemp +
##     as.factor(time):temp + as.factor(time):humidity + as.factor(time):windspeed +
##     as.factor(time):atemp
##
##                                   Df Sum of Sq    RSS     AIC
## - as.factor(workingday):atemp      1     0.197 6580.9 -5088.1
## - as.factor(holiday):humidity      1     0.372 6581.1 -5087.8
## - as.factor(workingday):temp       1     0.851 6581.6 -5087.1
## <none>                                       6580.7 -5086.4
## - atemp:as.factor(time)            3     4.393 6585.1 -5085.3
## - as.factor(holiday):atemp         1     2.172 6582.9 -5084.9
## - as.factor(holiday):temp          1     2.324 6583.1 -5084.7
## - temp:as.factor(time)             3     6.159 6586.9 -5082.4
## - as.factor(holiday):windspeed     1     4.670 6585.4 -5080.9
## - as.factor(season):atemp          3     8.604 6589.3 -5078.5
## - as.factor(workingday):windspeed  1    10.612 6591.3 -5071.2
## - windspeed:as.factor(time)        3    13.090 6593.8 -5071.2
## - as.factor(season):windspeed      3    15.347 6596.1 -5067.6
## - as.factor(season):temp           3    17.179 6597.9 -5064.6
## - as.factor(workingday):humidity   1    16.871 6597.6 -5061.1
## - humidity:as.factor(time)         3    22.551 6603.3 -5055.9
## - as.factor(season):humidity       3    32.299 6613.0 -5040.1
## - as.factor(weather)               2    93.339 6674.1 -4939.9
##
## Step:  AIC=-5088.13
## log(registered + 1) ~ as.factor(season) + as.factor(holiday) +
##     as.factor(workingday) + as.factor(weather) + temp + atemp +
##     humidity + windspeed + as.factor(time) + as.factor(season):temp +
##     as.factor(season):humidity + as.factor(season):windspeed +
##     as.factor(season):atemp + as.factor(holiday):temp + as.factor(holiday):humidity +
##     as.factor(holiday):windspeed + as.factor(holiday):atemp +
##     as.factor(workingday):temp + as.factor(workingday):humidity +
##     as.factor(workingday):windspeed + temp:as.factor(time) +
##     humidity:as.factor(time) + windspeed:as.factor(time) + atemp:as.factor(time)
##
##                                   Df Sum of Sq    RSS     AIC
## - as.factor(holiday):humidity      1     0.385 6581.3 -5089.5
## <none>                                       6580.9 -5088.1
## - atemp:as.factor(time)            3     4.393 6585.3 -5087.0
## - as.factor(holiday):atemp         1     1.981 6582.9 -5086.9
## - as.factor(holiday):temp          1     2.131 6583.1 -5086.7
## - temp:as.factor(time)             3     6.156 6587.1 -5084.1
## - as.factor(holiday):windspeed     1     4.544 6585.5 -5082.8
```

```
## - as.factor(season):atemp          3     9.009 6589.9 -5079.5
## - windspeed:as.factor(time)        3    12.973 6593.9 -5073.1
## - as.factor(workingday):windspeed  1    10.647 6591.6 -5072.9
## - as.factor(season):windspeed      3    15.282 6596.2 -5069.3
## - as.factor(workingday):temp       1    15.873 6596.8 -5064.4
## - as.factor(season):temp           3    18.436 6599.4 -5064.2
## - as.factor(workingday):humidity   1    17.106 6598.0 -5062.4
## - humidity:as.factor(time)         3    22.546 6603.5 -5057.6
## - as.factor(season):humidity       3    32.286 6613.2 -5041.8
## - as.factor(weather)               2    93.408 6674.3 -4941.5
##
## Step:  AIC=-5089.5
## log(registered + 1) ~ as.factor(season) + as.factor(holiday) +
##     as.factor(workingday) + as.factor(weather) + temp + atemp +
##     humidity + windspeed + as.factor(time) + as.factor(season):temp +
##     as.factor(season):humidity + as.factor(season):windspeed +
##     as.factor(season):atemp + as.factor(holiday):temp + as.factor(holiday):windspeed +
##     as.factor(holiday):atemp + as.factor(workingday):temp + as.factor(workingday):humidity +
##     as.factor(workingday):windspeed + temp:as.factor(time) +
##     humidity:as.factor(time) + windspeed:as.factor(time) + atemp:as.factor(time)
##
##                                   Df Sum of Sq    RSS     AIC
## <none>                                         6581.3 -5089.5
## - atemp:as.factor(time)            3     4.426 6585.7 -5088.3
## - as.factor(holiday):atemp         1     2.511 6583.8 -5087.4
## - as.factor(holiday):temp          1     2.646 6584.0 -5087.2
## - temp:as.factor(time)             3     6.229 6587.5 -5085.4
## - as.factor(holiday):windspeed     1     4.159 6585.5 -5084.8
## - as.factor(season):atemp          3     9.055 6590.4 -5080.8
## - as.factor(workingday):windspeed  1    10.395 6591.7 -5074.6
## - windspeed:as.factor(time)        3    13.073 6594.4 -5074.3
## - as.factor(season):windspeed      3    15.386 6596.7 -5070.5
## - as.factor(workingday):temp       1    15.987 6597.3 -5065.6
## - as.factor(season):temp           3    18.494 6599.8 -5065.5
## - as.factor(workingday):humidity   1    16.806 6598.1 -5064.2
## - humidity:as.factor(time)         3    22.357 6603.7 -5059.3
## - as.factor(season):humidity       3    32.096 6613.4 -5043.5
## - as.factor(weather)               2    93.444 6674.8 -4942.8
```

```
summary(casual.lm.short)
```

```
##
## Call:
## lm(formula = log(casual + 1) ~ as.factor(season) + as.factor(holiday) +
##     as.factor(workingday) + as.factor(weather) + temp + atemp +
##     humidity + windspeed + as.factor(time) + as.factor(season):temp +
##     as.factor(season):humidity + as.factor(season):windspeed +
##     as.factor(season):atemp + as.factor(holiday):humidity + as.factor(workingday):temp +
##     as.factor(workingday):humidity + as.factor(workingday):windspeed +
##     as.factor(workingday):atemp + as.factor(weather):temp + temp:as.factor(time) +
##     humidity:as.factor(time) + windspeed:as.factor(time) + atemp:as.factor(time),
##     data = train[-sample.num, -c(11, 12)])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -3.3662 -0.4443  0.0545  0.5017  3.0432
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      2.2032940  0.1179726  18.676  < 2e-16
## as.factor(season)2               1.3997802  0.1363432  10.267  < 2e-16
## as.factor(season)3               2.5692755  0.2001933  12.834  < 2e-16
## as.factor(season)4               1.0168764  0.1201764   8.462  < 2e-16
## as.factor(holiday)1             -0.5499089  0.1626462  -3.381 0.000725
## as.factor(workingday)1          -0.6728014  0.0884378  -7.608 3.03e-14
## as.factor(weather)2             -0.0988971  0.0489037  -2.022 0.043172
## as.factor(weather)3             -0.6888921  0.0898938  -7.663 1.97e-14
## temp                             0.1394919  0.0255137   5.467 4.67e-08
## atemp                           -0.0082600  0.0218813  -0.377 0.705817
## humidity                        -0.0053912  0.0012446  -4.332 1.49e-05
## windspeed                       -0.0182098  0.0029604  -6.151 7.97e-10
## as.factor(time)evening          -1.1972941  0.1072067 -11.168  < 2e-16
## as.factor(time)midnight         -1.7009386  0.1174494 -14.482  < 2e-16
## as.factor(time)sunrise          -0.2753464  0.1121104  -2.456 0.014064
## as.factor(season)2:temp         -0.2215846  0.0270780  -8.183 3.08e-16
## as.factor(season)3:temp         -0.1792939  0.0229281  -7.820 5.79e-15
## as.factor(season)4:temp         -0.1555891  0.0285573  -5.448 5.20e-08
## as.factor(season)2:humidity     -0.0100982  0.0011292  -8.942  < 2e-16
## as.factor(season)3:humidity     -0.0149099  0.0014188 -10.509  < 2e-16
## as.factor(season)4:humidity     -0.0112096  0.0011733  -9.554  < 2e-16
## as.factor(season)2:windspeed     0.0178888  0.0029907   5.981 2.28e-09
## as.factor(season)3:windspeed     0.0198899  0.0030936   6.429 1.34e-10
## as.factor(season)4:windspeed     0.0115958  0.0031567   3.673 0.000241
## as.factor(season)2:atemp         0.1482698  0.0238569   6.215 5.33e-10
## as.factor(season)3:atemp         0.0795102  0.0191290   4.157 3.26e-05
## as.factor(season)4:atemp         0.1180095  0.0247512   4.768 1.89e-06
## as.factor(holiday)1:humidity     0.0061081  0.0025214   2.422 0.015433
## as.factor(workingday)1:temp      0.0431711  0.0164791   2.620 0.008812
## as.factor(workingday)1:humidity  0.0052430  0.0008766   5.981 2.29e-09
## as.factor(workingday)1:windspeed 0.0041175  0.0020873   1.973 0.048562
## as.factor(workingday)1:atemp    -0.0555835  0.0150773  -3.687 0.000228
## as.factor(weather)2:temp         0.0052057  0.0022857   2.278 0.022775
## as.factor(weather)3:temp         0.0101452  0.0043167   2.350 0.018780
## temp:as.factor(time)evening      0.0567811  0.0145639   3.899 9.73e-05
## temp:as.factor(time)midnight     0.0380521  0.0160234   2.375 0.017577
## temp:as.factor(time)sunrise      0.0240850  0.0150260   1.603 0.108989
## humidity:as.factor(time)evening -0.0025902  0.0011407  -2.271 0.023190
## humidity:as.factor(time)midnight 0.0006437  0.0012786   0.503 0.614654
## humidity:as.factor(time)sunrise -0.0095680  0.0011874  -8.058 8.59e-16
## windspeed:as.factor(time)evening -0.0021344  0.0025723  -0.830 0.406699
## windspeed:as.factor(time)midnight -0.0045518  0.0027697  -1.643 0.100329
## windspeed:as.factor(time)sunrise  0.0027903  0.0025837   1.080 0.280186
## atemp:as.factor(time)evening    -0.0142637  0.0133456  -1.069 0.285185
## atemp:as.factor(time)midnight   -0.0546733  0.0145509  -3.757 0.000173
## atemp:as.factor(time)sunrise    -0.0113922  0.0136764  -0.833 0.404872
##
## (Intercept)                      ***
## as.factor(season)2               ***
## as.factor(season)3               ***
```

```
## as.factor(season)4                ***
## as.factor(holiday)1               ***
## as.factor(workingday)1            ***
## as.factor(weather)2               *
## as.factor(weather)3               ***
## temp                              ***
## atemp
## humidity                          ***
## windspeed                         ***
## as.factor(time)evening            ***
## as.factor(time)midnight           ***
## as.factor(time)sunrise            *
## as.factor(season)2:temp           ***
## as.factor(season)3:temp           ***
## as.factor(season)4:temp           ***
## as.factor(season)2:humidity       ***
## as.factor(season)3:humidity       ***
## as.factor(season)4:humidity       ***
## as.factor(season)2:windspeed      ***
## as.factor(season)3:windspeed      ***
## as.factor(season)4:windspeed      ***
## as.factor(season)2:atemp          ***
## as.factor(season)3:atemp          ***
## as.factor(season)4:atemp          ***
## as.factor(holiday)1:humidity      *
## as.factor(workingday)1:temp       **
## as.factor(workingday)1:humidity   ***
## as.factor(workingday)1:windspeed  *
## as.factor(workingday)1:atemp      ***
## as.factor(weather)2:temp          *
## as.factor(weather)3:temp          *
## temp:as.factor(time)evening       ***
## temp:as.factor(time)midnight      *
## temp:as.factor(time)sunrise
## humidity:as.factor(time)evening   *
## humidity:as.factor(time)midnight
## humidity:as.factor(time)sunrise   ***
## windspeed:as.factor(time)evening
## windspeed:as.factor(time)midnight
## windspeed:as.factor(time)sunrise
## atemp:as.factor(time)evening
## atemp:as.factor(time)midnight     ***
## atemp:as.factor(time)sunrise
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7323 on 10640 degrees of freedom
## Multiple R-squared:  0.7592, Adjusted R-squared:  0.7582
## F-statistic: 745.4 on 45 and 10640 DF,  p-value: < 2.2e-16
```

```
summary(registered.lm.short)
```

```
##
## Call:
## lm(formula = log(registered + 1) ~ as.factor(season) + as.factor(holiday) +
```

```
##     as.factor(workingday) + as.factor(weather) + temp + atemp +
##     humidity + windspeed + as.factor(time) + as.factor(season):temp +
##     as.factor(season):humidity + as.factor(season):windspeed +
##     as.factor(season):atemp + as.factor(holiday):temp + as.factor(holiday):windspeed +
##     as.factor(holiday):atemp + as.factor(workingday):temp + as.factor(workingday):humidity +
##     as.factor(workingday):windspeed + temp:as.factor(time) +
##     humidity:as.factor(time) + windspeed:as.factor(time) + atemp:as.factor(time),
##     data = train[-sample.num, -c(10, 12)])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.7987 -0.4745 -0.0123  0.4986  2.6518
##
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       4.404e+00  1.256e-01  35.072  < 2e-16
## as.factor(season)2                4.225e-01  1.453e-01   2.908 0.003641
## as.factor(season)3                1.544e+00  2.133e-01   7.242 4.73e-13
## as.factor(season)4                1.289e+00  1.291e-01   9.984  < 2e-16
## as.factor(holiday)1              -2.945e-01  1.856e-01  -1.587 0.112557
## as.factor(workingday)1           -1.087e-01  8.913e-02  -1.220 0.222574
## as.factor(weather)2               1.406e-02  1.886e-02   0.745 0.456039
## as.factor(weather)3              -3.685e-01  3.207e-02 -11.487  < 2e-16
## temp                              1.226e-01  2.458e-02   4.990 6.15e-07
## atemp                            -5.616e-02  2.070e-02  -2.714 0.006666
## humidity                         -4.397e-03  1.310e-03  -3.355 0.000797
## windspeed                        -1.556e-02  3.184e-03  -4.886 1.05e-06
## as.factor(time)evening           -3.676e-01  1.147e-01  -3.203 0.001363
## as.factor(time)midnight          -2.261e+00  1.254e-01 -18.028  < 2e-16
## as.factor(time)sunrise           -2.796e-01  1.200e-01  -2.331 0.019792
## as.factor(season)2:temp          -1.175e-01  2.908e-02  -4.040 5.38e-05
## as.factor(season)3:temp          -1.270e-01  2.431e-02  -5.226 1.77e-07
## as.factor(season)4:temp          -7.492e-02  3.072e-02  -2.439 0.014746
## as.factor(season)2:humidity      -5.280e-03  1.183e-03  -4.464 8.14e-06
## as.factor(season)3:humidity      -8.732e-03  1.453e-03  -6.011 1.90e-09
## as.factor(season)4:humidity      -7.407e-03  1.249e-03  -5.928 3.15e-09
## as.factor(season)2:windspeed      8.555e-03  3.213e-03   2.663 0.007768
## as.factor(season)3:windspeed      1.577e-02  3.325e-03   4.743 2.13e-06
## as.factor(season)4:windspeed      4.772e-03  3.397e-03   1.405 0.160082
## as.factor(season)2:atemp          8.959e-02  2.563e-02   3.496 0.000474
## as.factor(season)3:atemp          6.546e-02  2.020e-02   3.241 0.001195
## as.factor(season)4:atemp          3.933e-02  2.661e-02   1.478 0.139441
## as.factor(holiday)1:temp         -1.100e-01  5.317e-02  -2.069 0.038613
## as.factor(holiday)1:windspeed     1.593e-02  6.142e-03   2.593 0.009522
## as.factor(holiday)1:atemp         9.858e-02  4.892e-02   2.015 0.043938
## as.factor(workingday)1:temp      -1.123e-02  2.209e-03  -5.084 3.76e-07
## as.factor(workingday)1:humidity   4.793e-03  9.195e-04   5.213 1.90e-07
## as.factor(workingday)1:windspeed  9.126e-03  2.226e-03   4.100 4.17e-05
## temp:as.factor(time)evening       2.304e-02  1.564e-02   1.474 0.140637
## temp:as.factor(time)midnight      5.401e-02  1.719e-02   3.142 0.001682
## temp:as.factor(time)sunrise       1.628e-02  1.614e-02   1.009 0.313169
## humidity:as.factor(time)evening  -6.183e-03  1.223e-03  -5.056 4.36e-07
## humidity:as.factor(time)midnight -3.975e-03  1.367e-03  -2.907 0.003658
## humidity:as.factor(time)sunrise   1.525e-05  1.273e-03   0.012 0.990444
```

```
## windspeed:as.factor(time)evening   4.810e-04  2.760e-03   0.174 0.861650
## windspeed:as.factor(time)midnight -1.215e-02  2.970e-03  -4.092 4.31e-05
## windspeed:as.factor(time)sunrise  -1.827e-03  2.773e-03  -0.659 0.510166
## atemp:as.factor(time)evening       4.411e-03  1.432e-02   0.308 0.758103
## atemp:as.factor(time)midnight     -3.629e-02  1.562e-02  -2.323 0.020172
## atemp:as.factor(time)sunrise      -1.059e-02  1.469e-02  -0.721 0.471125
##
## (Intercept)                        ***
## as.factor(season)2                 **
## as.factor(season)3                 ***
## as.factor(season)4                 ***
## as.factor(holiday)1
## as.factor(workingday)1
## as.factor(weather)2
## as.factor(weather)3                ***
## temp                               ***
## atemp                              **
## humidity                           ***
## windspeed                          ***
## as.factor(time)evening             **
## as.factor(time)midnight            ***
## as.factor(time)sunrise             *
## as.factor(season)2:temp            ***
## as.factor(season)3:temp            ***
## as.factor(season)4:temp            *
## as.factor(season)2:humidity        ***
## as.factor(season)3:humidity        ***
## as.factor(season)4:humidity        ***
## as.factor(season)2:windspeed       **
## as.factor(season)3:windspeed       ***
## as.factor(season)4:windspeed
## as.factor(season)2:atemp           ***
## as.factor(season)3:atemp           **
## as.factor(season)4:atemp
## as.factor(holiday)1:temp           *
## as.factor(holiday)1:windspeed      **
## as.factor(holiday)1:atemp          *
## as.factor(workingday)1:temp        ***
## as.factor(workingday)1:humidity    ***
## as.factor(workingday)1:windspeed   ***
## temp:as.factor(time)evening
## temp:as.factor(time)midnight       **
## temp:as.factor(time)sunrise
## humidity:as.factor(time)evening    ***
## humidity:as.factor(time)midnight   **
## humidity:as.factor(time)sunrise
## windspeed:as.factor(time)evening
## windspeed:as.factor(time)midnight ***
## windspeed:as.factor(time)sunrise
## atemp:as.factor(time)evening
## atemp:as.factor(time)midnight      *
## atemp:as.factor(time)sunrise
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 0.7864 on 10641 degrees of freedom
## Multiple R-squared:  0.6845, Adjusted R-squared:  0.6832
## F-statistic: 524.7 on 44 and 10641 DF,  p-value: < 2.2e-16
```

```
length(coef(casual.lm.short))
```

```
## [1] 46
```

```
predicted.casual.short = predict.lm(casual.lm.short, newdata = my.test[,-c(10,11,12)])
predicted.registered.short = predict.lm(registered.lm.short, newdata = my.test[,-c(10,11,12)])
predicted.sum.short = predicted.casual.short + predicted.registered.short
true.value = log(my.test$count)
RSS.short = sum((true.value-predicted.sum.short)^2)



RSS.short
```

```
## [1] 1718.734
```

```
RSS.1
```

```
## [1] 1718.551
```

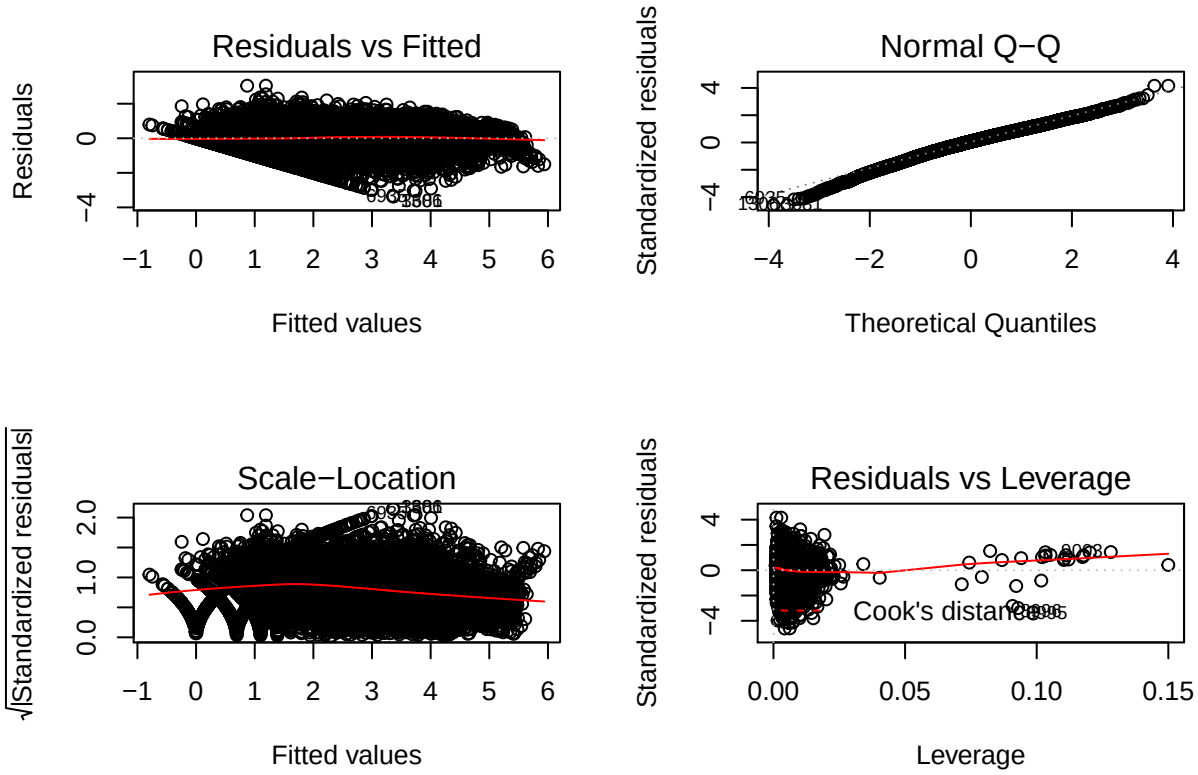
```
(RSS.short - RSS.1)/RSS.1
```

```
## [1] 0.0001062083
```

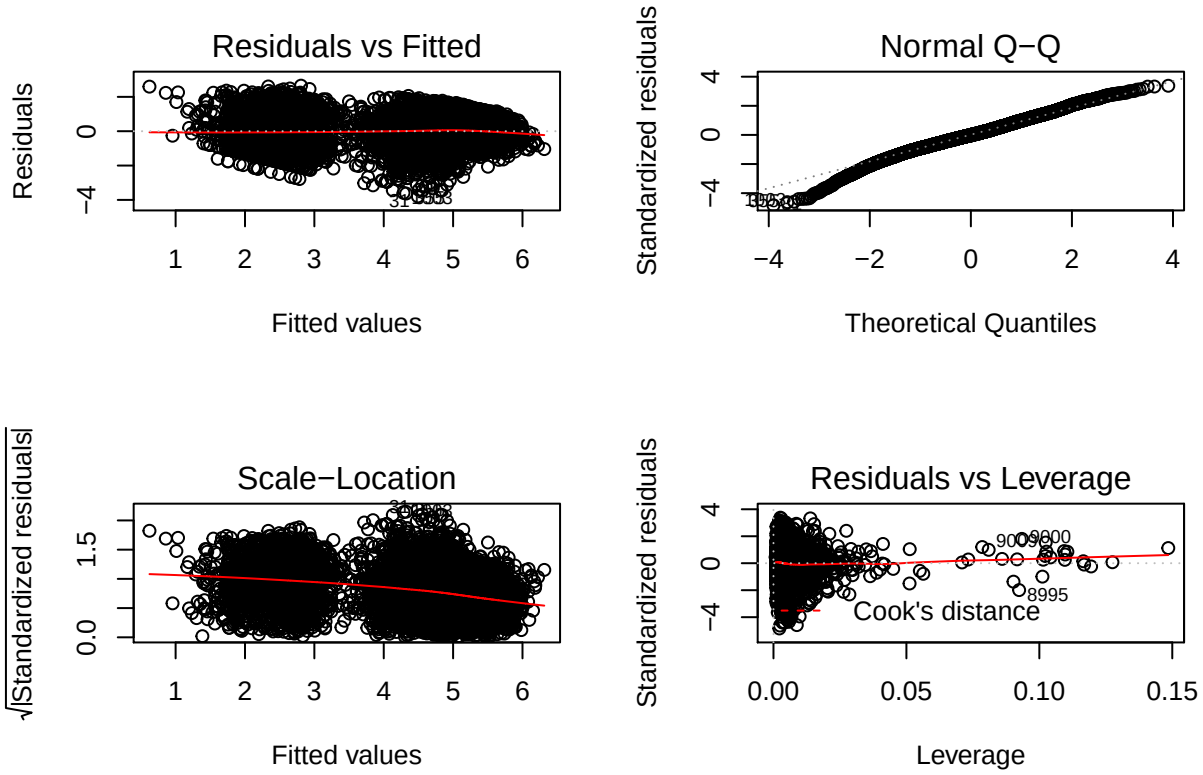
As shown in the comparison above, after I decrease the number of parameters through step function, the RSS of the prediction doesn't change much. Therefore, I will use the shorter version of the regression equation.

6. Regression Diagnostics

```
par(mfrow = c(2,2))
plot(casual.lm.short)
```




```
par(mfrow = c(2,2))
plot(registered.lm.short)
```



From the regression diagnostic plots shown above. We could see that generally, the assumptions are satisfied.

For both casual.lm.short and registered.lm.short, the linearity assumptions holds true because the Residual vs Fitted plot shows a horizontal line, which means that the fitted values does not have a linear realtionship with residuals. From the Normal QQ plots of casual.lm.short and registered.lm.short, we see that the distribution of residuals are generally normal with slightly heavier tail at the ends. The homoscedasticity assumption is slightly violated since for both casual.lm.short and registered.lm.short, the variance of residual decreases as fitted values increases.

7. Predictions

As I mentioned above, I combined weather4 into weather3 in my training dataset. To make sure I correctly predict the test data, I am going to combine weather4 into weather3 in my test.csv correspondingly.

```
modify.func = function(){
  for (i in c(1: nrow(test))) {
    if ( test$weather[i] == 4) {
      return(i)
    }
  }
}

modify.func()
```

```
## [1] 155
```

```
test$weather[155] = 3
modify.func()
```

```
## [1] 3249
```

```
test$weather[3249] = 3
sum(test$weather == 4)
```

```
## [1] 0
```

Now I make sure that all weather4 has been included as weather3.

```
predict.log.casual =  predict(casual.lm.short, newdata = test)
predict.log.registered =  predict(registered.lm.short, newdata = test)
predict.norm.casual = expm1(predict.log.casual)
predict.norm.registered = expm1(predict.log.registered)
predict.norm.count = predict.norm.casual + predict.norm.registered


predict.test.1 = data.frame(datetime = test$datetime, count = predict.norm.count)




write.csv(predict.test.1, file = "prediction submission.csv", row.names = F)
```

8.Comparison

After submitting the results in kaggle, I got a error of 0.77, which is relatively high. I checked my prediction, it seems to be overall higher than the training dataset, I think it might be because the intercept is too high. Some methods that I could use to improve my prediction is that I could to reduce some insignificant categorical variables and run anova test. I could also fit a quadratic model to see whether it's a better fit. Using cross validation might also be a better way to reduce varaibles then stepwise function given more powerful platforms and more time.