Lab2_BikeShared_Model1_LinearR

```
library(tinytex)
```

```
library(ggplot2)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(Hmisc)
```

```
## Loading required package: lattice

## Loading required package: survival

## Loading required package: Formula

##
## Attaching package: 'Hmisc'

## The following objects are masked from 'package:dplyr':
##
##     src, summarize

## The following objects are masked from 'package:base':
##
##     format.pval, units
```

Get the Date

```
bike <- read.csv("hour.csv", header= TRUE )
```

Clear Data

```
bike$instant<- NULL
```

the temp in raw dataset have been divided by 100 in fahrenheit. Let's convert them by times 100 and then to Celsius.

```
library(weathermetrics)
```

```
bike <- mutate(bike, temp = fahrenheit.to.celsius(bike$temp *100 ))
```

```
bike <- mutate(bike, atemp = fahrenheit.to.celsius(bike$atemp *100 ))
```

```
bike$dteday <- as.POSIXct(bike$dteday)
```

the hum in the raw dataset have been divided by 100, let's convert them back.

```
bike <- mutate(bike, hum = bike$hum*100 )
```

the windspeed in the dataset have been divided by 67. let's convert them back.

```
bike <- mutate(bike, windspeed = bike$windspeed*67)
```

```
head(bike)
```

```
##       dteday season yr mnth hr holiday weekday workingday weathersit  temp
## 1 2011-01-01      1  0    1  0       0       6          0          1 -4.44
## 2 2011-01-01      1  0    1  1       0       6          0          1 -5.56
## 3 2011-01-01      1  0    1  2       0       6          0          1 -5.56
## 4 2011-01-01      1  0    1  3       0       6          0          1 -4.44
## 5 2011-01-01      1  0    1  4       0       6          0          1 -4.44
## 6 2011-01-01      1  0    1  5       0       6          0          2 -4.44
##   atemp hum windspeed casual registered cnt
## 1 -1.78  81    0.0000      3         13  16
## 2 -2.63  80    0.0000      8         32  40
## 3 -2.63  80    0.0000      5         27  32
## 4 -1.78  75    0.0000      3         10  13
## 5 -1.78  75    0.0000      0          1   1
## 6 -3.47  75    6.0032      0          1   1
```

use one year dataset for doing lenear Regression model

```
bike2012<- bike[8646:17379, ]
```

```
numeric_cols = sapply(bike2012, is.numeric)
bike_num_only = bike2012[, numeric_cols]
```

```
colnames(bike_num_only)
```

```
##  [1] "season"     "yr"         "mnth"       "hr"         "holiday"
##  [6] "weekday"    "workingday" "weathersit" "temp"       "atemp"
## [11] "hum"        "windspeed"  "casual"     "registered" "cnt"
```

```
bike2012_sub= bike2012[ , c("season", "mnth", "hr" ,  "holiday","weekday", "workingday","weathersit", "
```

Determine correlation of each attributes to the counts of bike rent.

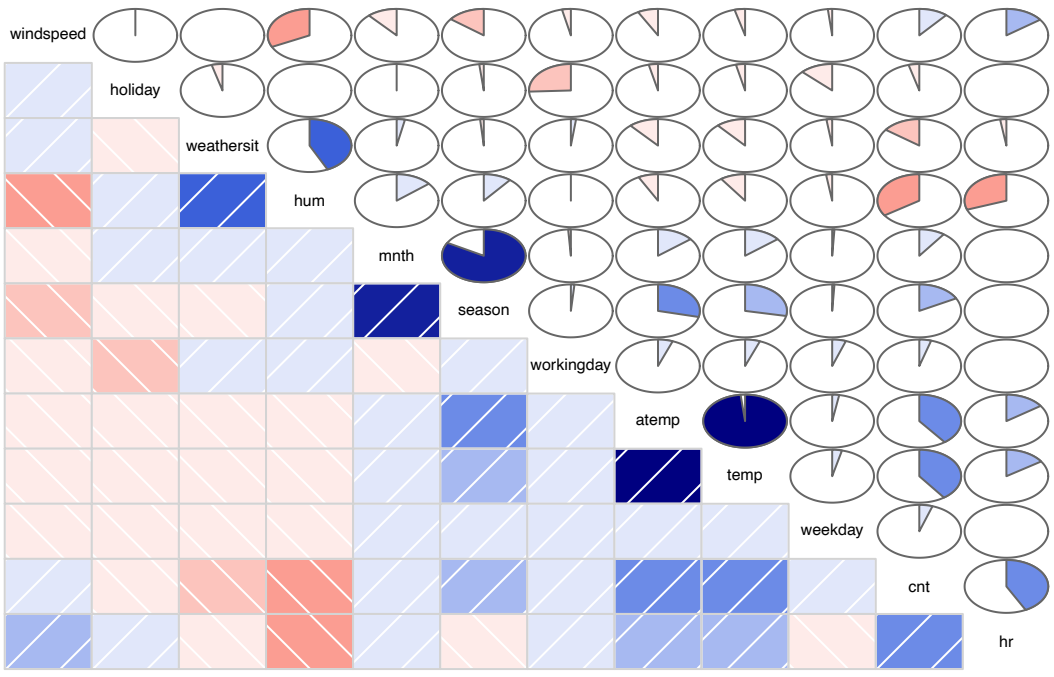
```
library(corrplot)
```

```
## corrplot 0.84 loaded
```

```
library(corrgram)
```

```
## Registered S3 method overwritten by 'seriation':
##   method         from
##   reorder.hclust gclus

##
## Attaching package: 'corrgram'

## The following object is masked from 'package:lattice':
##
##     panel.fill
```

```
corrgram(bike2012_sub,order=TRUE, lower.panel=panel.shade,
  upper.panel=panel.pie, text.panel=panel.txt)
```




```
cor(bike2012_sub)
```

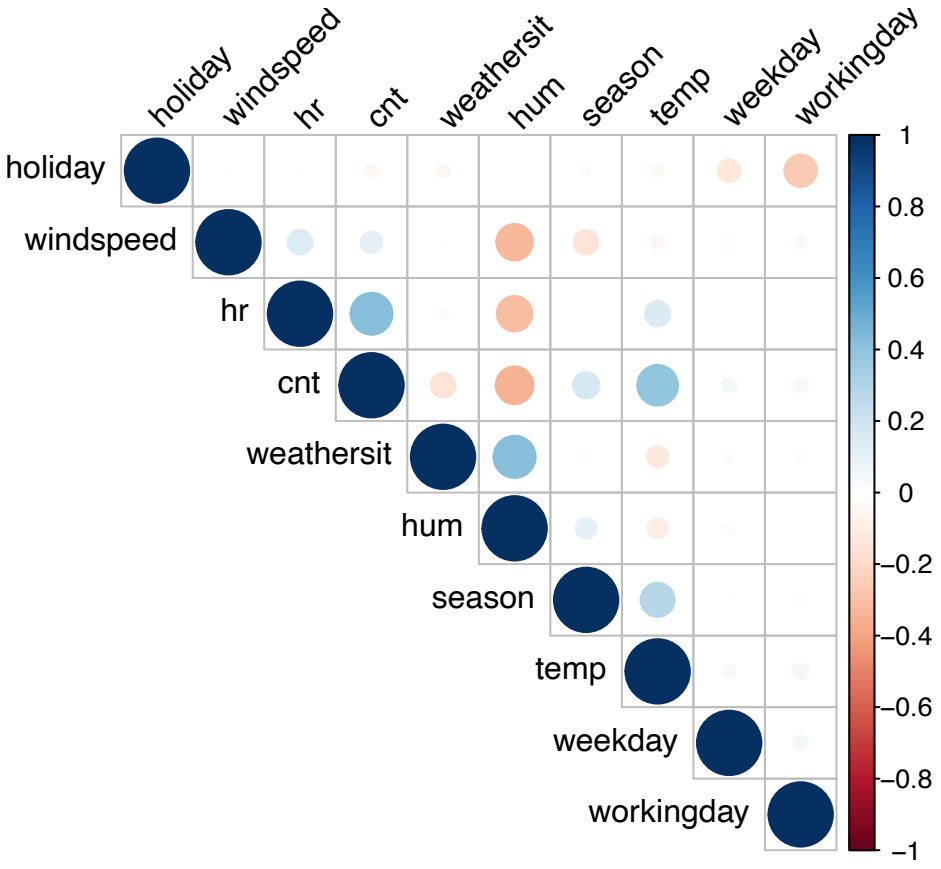
```
##                  season          mnth            hr       holiday       weekday
## season      1.0000000000  8.316577e-01 -2.270054e-04 -0.017499999  0.008735100
## mnth        0.8316577130  1.000000e+00  7.990218e-05  0.007797492  0.008710095
## hr         -0.0002270054  7.990218e-05  1.000000e+00  0.001869484 -0.001436206
## holiday    -0.0174999989  7.797492e-03  1.869484e-03  1.000000000 -0.126642035
## weekday     0.0087350995  8.710095e-03 -1.436206e-03 -0.126642035  1.000000000
## workingday  0.0136783682 -1.044709e-02  7.820535e-04 -0.256852423  0.053122187
## weathersit -0.0141023076  3.184429e-02 -2.418098e-02 -0.041803723 -0.022220696
## temp        0.2808095057  1.418881e-01  1.561488e-01 -0.035357509  0.037599087
## atemp       0.2870251337  1.421059e-01  1.526109e-01 -0.035731363  0.026853925
## hum         0.1075657222  1.396473e-01 -3.053484e-01  0.002675235 -0.023588523
## windspeed  -0.1450312963 -1.153985e-01  1.493244e-01  0.007474693 -0.016778200
## cnt         0.1717043364  9.937272e-02  4.240457e-01 -0.041315974  0.051170585
##               workingday   weathersit        temp       atemp          hum
## season      0.0136783682 -0.014102308  0.28080951  0.28702513  0.107565722
## mnth       -0.0104470890  0.031844287  0.14188811  0.14210589  0.139647261
## hr          0.0007820535 -0.024180977  0.15614877  0.15261088 -0.305348400
## holiday    -0.2568524234 -0.041803723 -0.03535751 -0.03573136  0.002675235
## weekday     0.0531221871 -0.022220696  0.03759909  0.02685393 -0.023588523
## workingday  1.0000000000  0.019844039  0.05780063  0.05720205  0.005520906
## weathersit  0.0198440392  1.000000000 -0.11286550 -0.11410042  0.429349584
## temp        0.0578006296 -0.112865503  1.00000000  0.98279969 -0.096547758
## atemp       0.0572020512 -0.114100422  0.98279969  1.00000000 -0.076590599
## hum         0.0055209055  0.429349584 -0.09654776 -0.07659060  1.000000000
## windspeed  -0.0342522129  0.002583501 -0.04066481 -0.07902589 -0.321941772
## cnt         0.0458205736 -0.148346392  0.39962599  0.39639356 -0.344065069
##               windspeed         cnt
## season     -0.145031296  0.17170434
## mnth       -0.115398519  0.09937272
## hr          0.149324399  0.42404571
## holiday     0.007474693 -0.04131597
## weekday    -0.016778200  0.05117058
## workingday -0.034252213  0.04582057
## weathersit  0.002583501 -0.14834639
## temp       -0.040664811  0.39962599
## atemp      -0.079025891  0.39639356
## hum        -0.321941772 -0.34406507
## windspeed   1.000000000  0.11115494
## cnt         0.111154939  1.00000000
```

```
bike2012_sub = bike2012[ , c("season", "hr" ,  "holiday","weekday", "workingday","weathersit", "temp",
```

```
cor_result = rcorr(as.matrix(bike2012_sub))
```

```
corrplot(cor_result$r, type = "upper", order = "hclust", tl.col = "black", tl.srt = 45)
```

Attribute "mnth" has noticeable correlation with "season","Holiday" and "hr". Remove mnth as an attrubute. Attibute"temp" and "atemp" have very high correlation at 0.98. Remove



"atemp" to avoid intercollinearity.

Create Linear Regression Model

```
library(caTools)
```

```
bike2012.model <- lm(cnt ~ season  + temp + hum +  weathersit , data= bike2012_sub)
```

```
summary(bike2012.model)
```

```
##
## Call:
## lm(formula = cnt ~ season + temp + hum + weathersit, data = bike2012_sub)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -374.01 -124.50  -38.54   88.24  667.88
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 324.0092     7.9222  40.899  < 2e-16 ***
## season       21.3966     1.8207  11.752  < 2e-16 ***
## temp          6.8297     0.1944  35.124  < 2e-16 ***
## hum          -3.7695     0.1141 -33.039  < 2e-16 ***
## weathersit   12.4860     3.4103   3.661 0.000253 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 179 on 8729 degrees of freedom
## Multiple R-squared:  0.2663, Adjusted R-squared:  0.266
## F-statistic: 792.2 on 4 and 8729 DF,  p-value: < 2.2e-16
```

We get pretty low Adjustered R_squared, which is 0.266. Let add "hr", "holiday".

```
colnames(bike2012_sub)
```

```
##  [1] "season"     "hr"         "holiday"    "weekday"    "workingday"
##  [6] "weathersit" "temp"       "hum"        "windspeed"  "cnt"
```

```
bike2012.model <- lm(cnt ~ season  + temp + hum +  weathersit + hr + holiday , data= bike2012_sub)
```

```
summary(bike2012.model)
```

```
##
## Call:
## lm(formula = cnt ~ season + temp + hum + weathersit + hr + holiday,
##     data = bike2012_sub)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -400.17 -112.97  -38.57   71.41  665.80
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 174.7804     8.7412  19.995  < 2e-16 ***
## season       21.2724     1.7146  12.407  < 2e-16 ***
## temp          5.9548     0.1850  32.186  < 2e-16 ***
## hum          -2.5543     0.1134 -22.519  < 2e-16 ***
## weathersit   -2.8016     3.2463  -0.863 0.388147
## hr            9.2813     0.2788  33.293  < 2e-16 ***
## holiday     -35.8185    10.6140  -3.375 0.000742 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 168.5 on 8727 degrees of freedom
## Multiple R-squared:  0.3496, Adjusted R-squared:  0.3491
## F-statistic: 781.7 on 6 and 8727 DF,  p-value: < 2.2e-16
```

Adjusted R-squared increase to 0.3491."Weathersit" have shown to be not significant as p>0.05.

Remove weathersit and add weekday, windseppd and workingday for adjusting the model

```
bike2012.model <- lm(cnt ~ season  + temp + hum + hr + holiday + weekday + workingday + windspeed, data=
```

```
summary(bike2012.model)
```

```
##
## Call:
## lm(formula = cnt ~ season + temp + hum + hr + holiday + weekday +
##     workingday + windspeed, data = bike2012_sub)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -413.57 -113.02  -39.03   71.62  657.48
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 143.8417    10.8331  13.278  < 2e-16 ***
## season       21.7921     1.7192  12.676  < 2e-16 ***
## temp          5.9477     0.1847  32.200  < 2e-16 ***
## hum          -2.5173     0.1060 -23.746  < 2e-16 ***
## hr            9.2175     0.2764  33.343  < 2e-16 ***
## holiday     -23.7160    11.0243  -2.151 0.031484 *
## weekday       3.2183     0.9067   3.549 0.000388 ***
## workingday   10.0245     4.0090   2.500 0.012422 *
## windspeed     0.5842     0.2363   2.472 0.013449 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 168.3 on 8725 degrees of freedom
## Multiple R-squared:  0.3513, Adjusted R-squared:  0.3507
## F-statistic: 590.7 on 8 and 8725 DF,  p-value: < 2.2e-16
```

The adjustered R-squred increases to 0.3507.P-values are all less than 0.05.

###let have a look if we build the model base on the original dataset, "hour.csv" (renamed as "bike"), for 2011 and 2012.

```
bike_subset = bike[ , c("season", "hr" ,  "holiday", "weekday", "workingday", "temp",  "hum", "windspeed
```

```
bike.model <- lm(cnt ~ season  + temp + hum + hr + holiday + weekday + workingday + windspeed, data= bik
```

```
summary(bike.model)
```

```
##
## Call:
## lm(formula = cnt ~ season + temp + hum + hr + holiday + weekday +
##     workingday + windspeed, data = bike_subset)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -335.82  -96.77  -30.46   54.01  691.14
```

```
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 132.53273    6.50831  20.364  < 2e-16 ***
## season       19.63438    1.08840  18.040  < 2e-16 ***
## temp          5.27017    0.11199  47.060  < 2e-16 ***
## hum          -2.22902    0.06339 -35.162  < 2e-16 ***
## hr            7.47914    0.17018  43.947  < 2e-16 ***
## holiday     -21.62282    6.95442  -3.109  0.00188 **
## weekday       1.57787    0.56159   2.810  0.00496 **
## workingday    3.49010    2.48953   1.402  0.16096
## windspeed     0.23108    0.14396   1.605  0.10848
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 147.6 on 17370 degrees of freedom
## Multiple R-squared:  0.3384, Adjusted R-squared:  0.3381
## F-statistic:  1110 on 8 and 17370 DF,  p-value: < 2.2e-16
```

We get less Adjusted R-squared which is 0.3381. Attributes "workingday" and "windspeed" have shown to be not significant as P-value <0.05.We will continue use bike2012.model to train and test the model.

Interpret bike2012.model

```
coef(bike2012.model)
```

```
## (Intercept)      season        temp         hum          hr     holiday
## 143.8417033  21.7920637   5.9476594  -2.5172664   9.2175063 -23.7160157
##     weekday  workingday   windspeed
##   3.2183018  10.0244554   0.5841525
```

Explaine the model manually —-If every attribute are "zero", the intercept is 143.842. —-add one if it is in winter,(times 2 if it is in spring, times 3 if it in autumns). —- add 5.948 if temp increase one Celsius; —- minus 2.217 if humidity increase 1/100 measurement; —- add 9.218 or minus 9.218 base on the time period and if it is on
workingday or not. for example if it is on Tuesday 8am, then add 8*9.218, which is 73.744; —- minus 23.716 if it is on holiday(people might drive more in holidays); —- add one unit of 3.218 from Monday to Friday then reduce to Monday; —-add 10.024 if it is workingday (minus 10.024 if it is none working day); —-The values of windspeed have been divided by 67 in the raw dataset. add 0.5841 if wind speed increase 1/100 k/h degree, or minus 0.5841 if it is reduced 1/100 k/h.

Base on the data visualization , we adjust the number accordinly.

Model prediction

Because this dataset is base on datetime. It is not necessary to train the model. We just use the coefficient result to predict the future numbers of bike rented.

Let's say on April 12 2020, the Weather of Washington USA forecast for the time of 23pm: temp is 19C, Cloudy,hum 66%, wind speed is 27 k/h, Sunday.(We have not yet consider the growth rate of the number of bike rental from 2012 to 2019)

We can estimate that the number of bike will be rent are:

143.84 + 21.792*2 + 5.948*19 - 2.517*66 + (10*9.217- 8*90217) - 23.716 + (3.218*4-3.218*2) -10.024 + (0.584 27) =141.212 ### The median of the residuals is -30.46 Estimate numbers will be 110.752.

```
143.84 + (21.792*2) + (5.948*19) -( 2.517*66) + (10*9.217- 8*9.217) - 23.716 + (3.218*4 - 3.218*2) -10.
```

caculate manually:

```
## [1] 141.212
```

```
141.212 - 30.46
```

```
## [1] 110.752
```

Evaluate the Model

```
library(MASS)
```

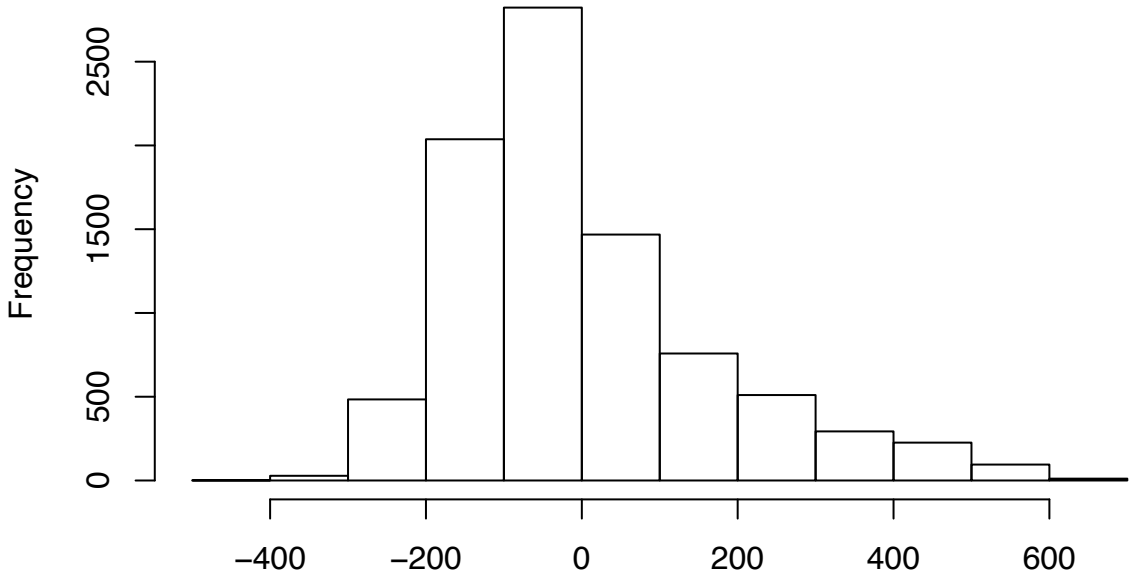
```
##
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':
##
##     select
```

```
fit <- lm(cnt ~ season  + temp + hum + hr + holiday + weekday + workingday + windspeed, data= bike2012_
```

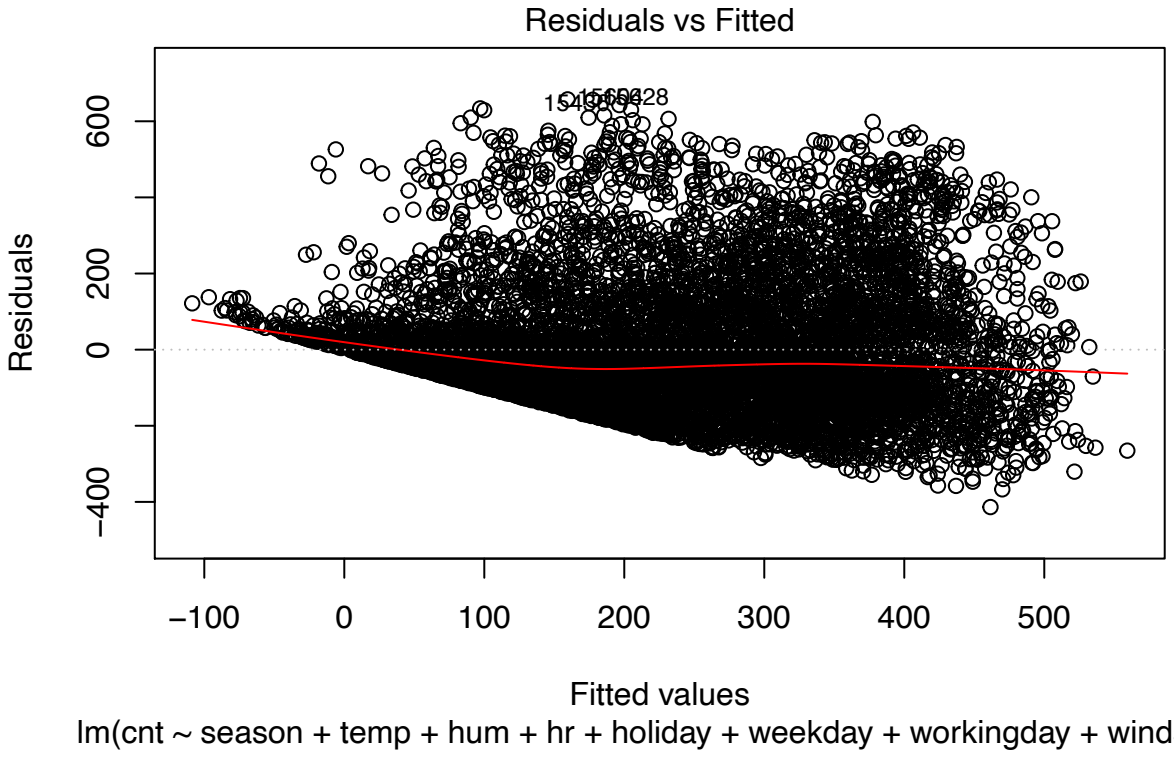
```
step <- stepAIC(fit, direction = "both", trace = FALSE)

step$anova
```

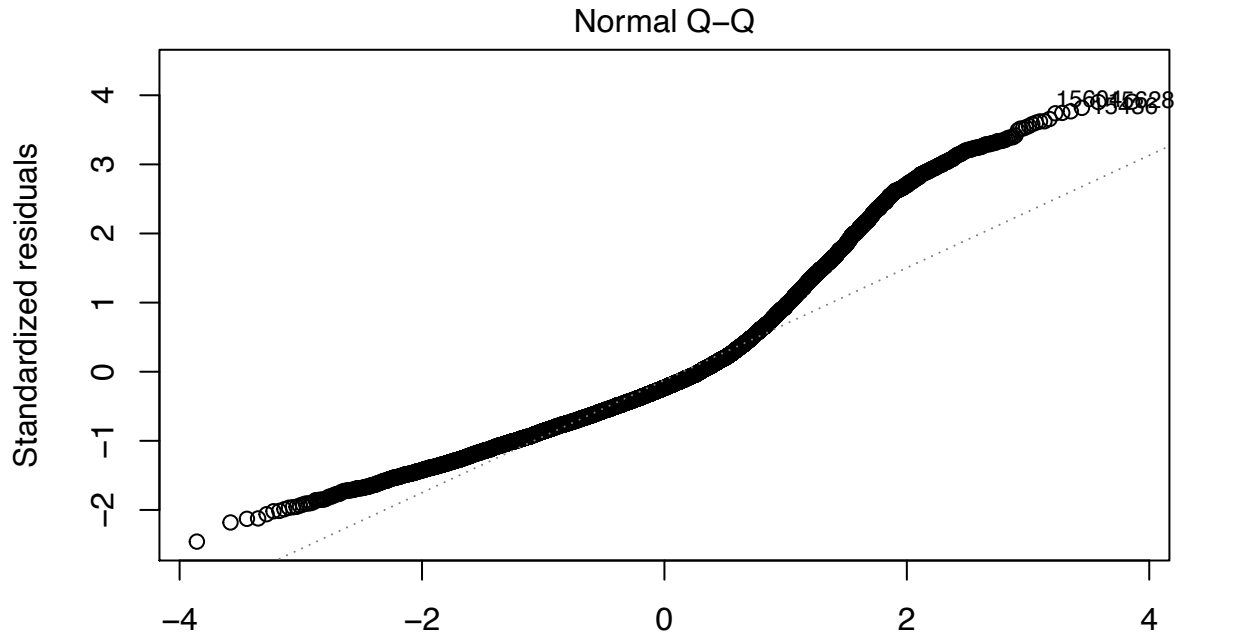
```
## Stepwise Model Path
## Analysis of Deviance Table
##
## Initial Model:
## cnt ~ season + temp + hum + hr + holiday + weekday + workingday +
##     windspeed
##
## Final Model:
## cnt ~ season + temp + hum + hr + holiday + weekday + workingday +
##     windspeed
##
##
##   Step Df Deviance Resid. Df Resid. Dev      AIC
## 1                       8725  247233352 89549.05
```

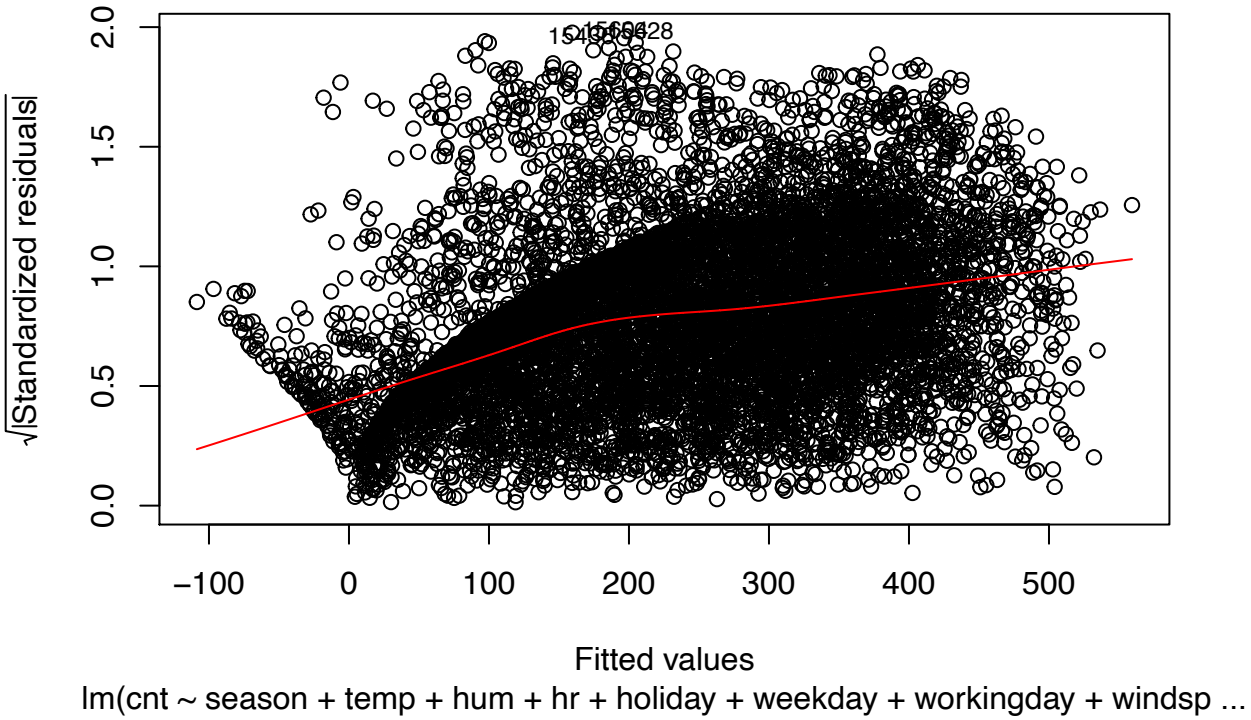
```
hist(residuals(bike2012.model), xlab = "", main = "")
```

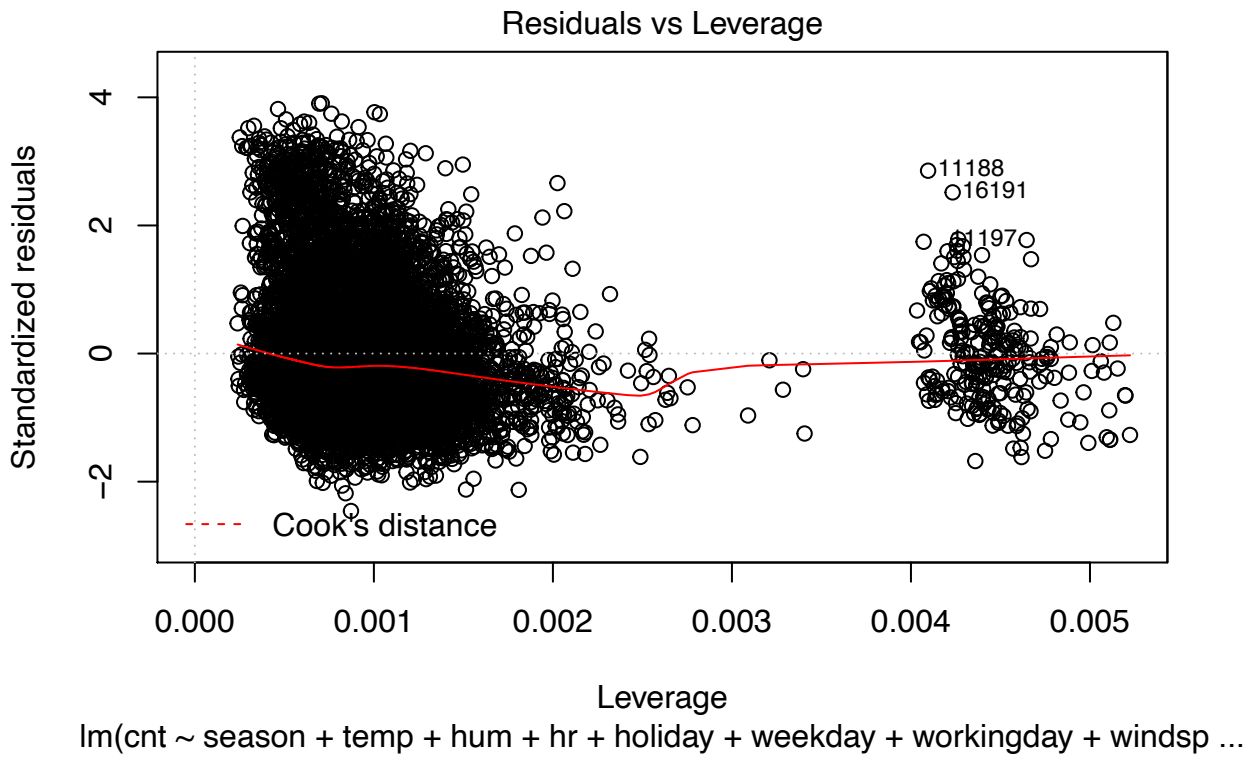



```
plot(bike2012.model)
```




Normal Q–Q
Standardized residuals
Theoretical Quantiles
lm(cnt ~ season + temp + hum + hr + holiday + weekday + workingday + windsp ...


Scale–Location
√|Standardized residuals|
Fitted values
lm(cnt ~ season + temp + hum + hr + holiday + weekday + workingday + windsp ...



Let's see if we use the train/test/split model to check any diffirent

```
set.seed(99)
sample<- sample.split(bike2012_sub$cnt, SplitRatio = 0.7)

train<- subset(bike2012_sub, sample==TRUE)
test <- subset(bike2012_sub,sample== FALSE)
```

```
biketrain.model <- lm(cnt ~ season  + temp + hum + hr + holiday + weekday + workingday + windspeed, data
```

```
summary(biketrain.model)
```

```
##
## Call:
## lm(formula = cnt ~ season + temp + hum + hr + holiday + weekday +
##     workingday + windspeed, data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -416.38 -114.57  -38.85   72.01  658.29
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 151.4166    12.8920  11.745  < 2e-16 ***
## season       20.7392     2.0775   9.983  < 2e-16 ***
## temp          6.0457     0.2227  27.152  < 2e-16 ***
```

```
## hum          -2.5475     0.1270 -20.057  < 2e-16 ***
## hr            9.1360     0.3343  27.328  < 2e-16 ***
## holiday     -20.2911    12.9535  -1.566 0.117294
## weekday       3.8218     1.0882   3.512 0.000448 ***
## workingday   11.5705     4.8133   2.404 0.016253 *
## windspeed     0.1913     0.2834   0.675 0.499587
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 169.7 on 6113 degrees of freedom
## Multiple R-squared:  0.3494, Adjusted R-squared:  0.3485
## F-statistic: 410.3 on 8 and 6113 DF,  p-value: < 2.2e-16
```

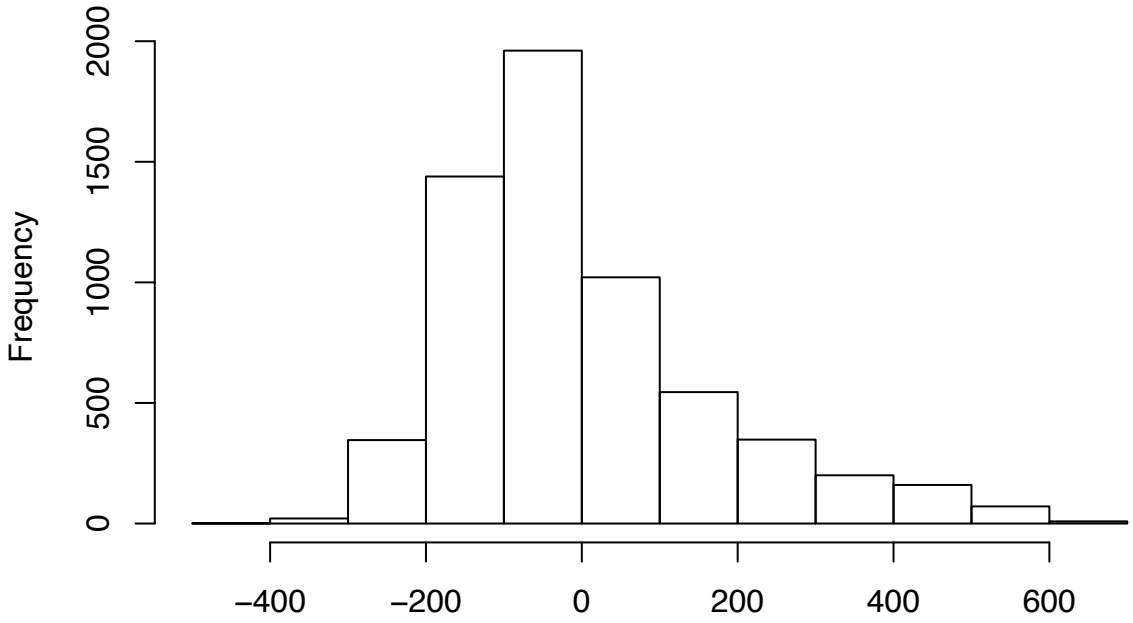
Adjustered R_squared is 0.3485. in this train model "holiday" and "windspeed" have no significant as p <0.05.

```
res <- residuals(biketrain.model)
```

```
head(as.data.frame(res))
```

```
##            res
## 8646 30.559900
## 8647 64.702289
## 8648 78.184470
## 8650 17.922822
## 8651  5.595775
## 8652 30.358775
```

```
hist(residuals(biketrain.model), xlab = "", main = "")
```



The residuals has the same as bike2012_sub.model, aslo shows normally distributaion.

Predition future bike counts by the bike.copy2012 model.

```
cnt.pred <- predict(biketrain.model,test)
```

```
results <- cbind(cnt.pred,test$cnt)
colnames(results) <- c("predicted", "actual")
results <- as.data.frame(results)

head(results)
```

```
##       predicted actual
## 8649 -13.496755     52
## 8653 -18.839447      7
## 8654   4.816501     14
## 8656  51.795132     70
## 8660 206.644692    267
## 8662 224.916784    215
```

```
to_zero <- function(x){
  if (x<0){
    return(0)
  }else{
    return(x)
  }
}
```

```
results$predicted <- sapply(results$predicted,to_zero)
head(results)
```

It is no sense if there is -13 or any minus zoro number of bike rented. Adjuste munus number to zero.

```
##       predicted actual
## 8649   0.000000     52
## 8653   0.000000      7
## 8654   4.816501     14
## 8656  51.795132     70
## 8660 206.644692    267
## 8662 224.916784    215
```

Basily the model is not that bad so far. However we will see if other kind of model will have a better results.