Regression Analysis Based on BikeSharing Data

April 13, 2020

Introduction

After exploratory data analysis and data processing, Stepwise is used to select variables. Then, two candidate model are selected under AIC and BIC. Cross-validation is further introduced to get the final model. Finally, some useful information is interpreted based on coefficients and basic statistics of the model.

Description of data

- EDA

We have 16 explanatory variables to be considered. First, it is obvious that variable *casual* and *registered* should not include in the regression model, because they have mathematical relationship with our response variable, cnt, rather than statistical relationship. Secondly, variables *instant* and *dteday* will not provide any useful information about *cnt*.

Excluding variable *casual* , *registered* , *instant* and *dteday*, we have 7 variables about date and 5 variables about weather condition, but most of them are categorical variables. In other words, there are only 4 continuous varibles, *temp* , *atemp* , *hum* and *windspeed*. So here we explore the linear correlation relationship [1]between these 4 continuous variables and the response variable *cnt*.

According to Figure 1, firstly, *temp* and *atemp* are highly correlated, which means one of them should be excluded from our regression model. Secondly, according to the scatterplot of *temp* and *atemp* and the scatterplot of *hum* and *atemp*, there are some outliers. Thirdly, although windspeed is described as a continuous variable, but it seems to only have several possible values according to its density plot, or histogram plot. It shows the bars are very "sparse". What's more, the distribution of our response variable, *cnt*, seems to be a little right-skewed, which may cause heteroscedasticity.

- Data processing

Based on EDA, I deleted outliers and employed Box-Cox[2] data transformation in response variable, *cnt*. The lambda of Box-Cox transformation is a hyper-parameter determined by the optimal result of transformation.[3]

[1]See Figure 1

[2]They reason why I did not use log transformation is that cnt is mildly right-skewed. It will become left-skewed after log transformation

[3]See Figure2

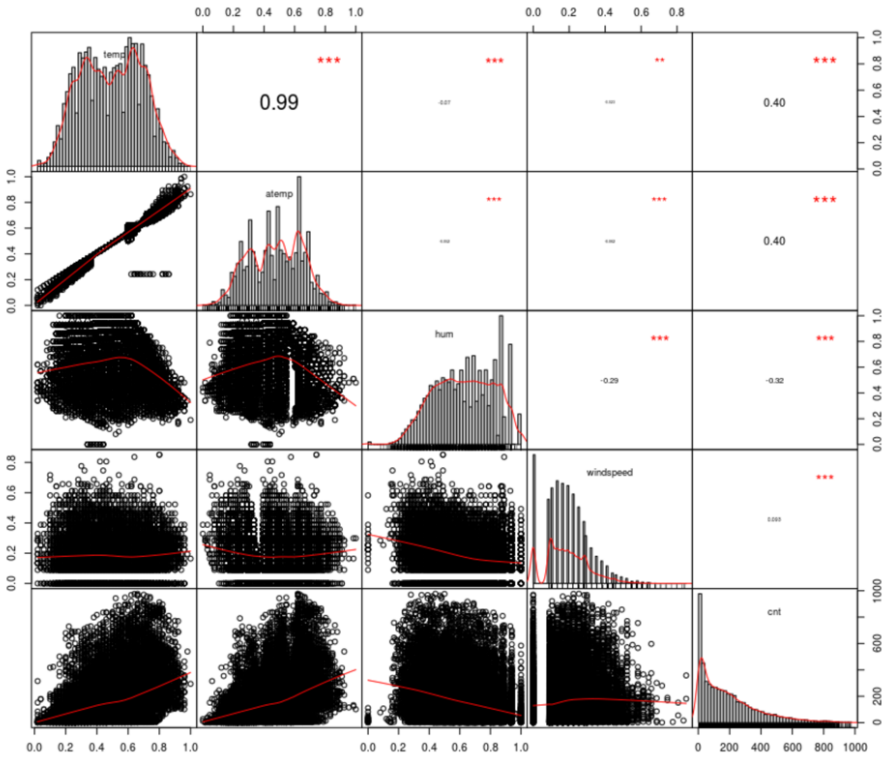

Figure 1: Correlation

Fitting models

- Candidate variables

Theoretically, we would better pick up those variables who are independent to other explanatory variables and highly correlated to the response variable. Based on EDA, I chose *atemp* , *hum* and *windspeed* as part of candidate variables. About categorical variables, considering their meanings, **mnth** contains more detail than *season*. However, it will cause more loss of degree of freedom. So I excluded *mnth* from candidate variables. About *holiday*, *weekday* and *workingday*, it is obvious that we cannot include them all in our model. Considering minimizing the loss of degree of freedom, I excluded *holiday* and *weekday*.

- Variable Selection

I used Stepwise technique to add and drop variables of candidate regression models. The best model based on AIC and BIC are (1) and (2) respectively:

$$cnt = hum + atemp + windspeed + fac(season) + fac(yr) + fac(hr) + fac(workingday) + fac(weathersit) \quad (1)$$

$$cnt = hum + atemp + windspeed + fac(season) + fac(yr) + fac(hr) + fac(weathersit) \quad (2)$$

\

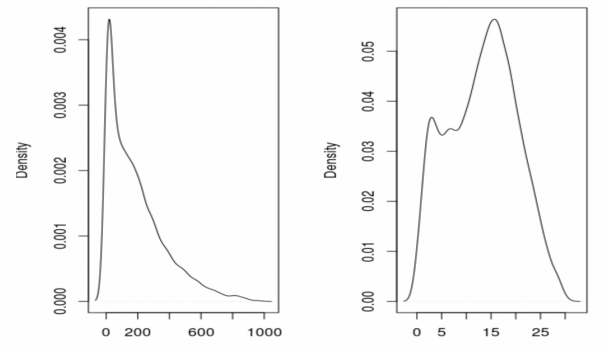

Figure 2: Left: before Box-Cox transformation; Right: after

- Model Selection

Once we got the best model under AIC and BIC, we could use Cross-Validation to select the best regression model, which has the smallest average MSE.

The Best Model Under	Average MSE
AIC	9.342
BIC	9.343

We finally picked up the best model under BIC as out final regression model.

Model Diagnosis

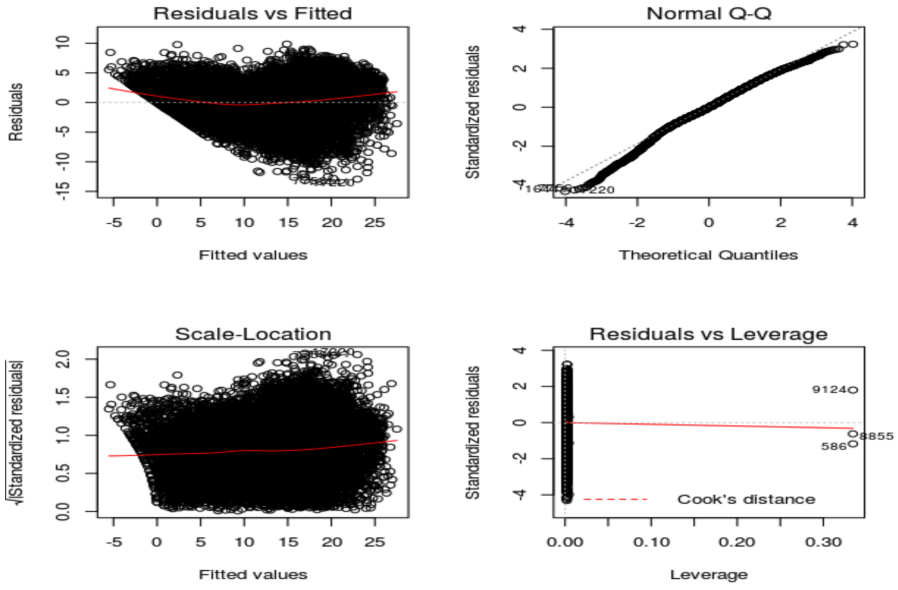

Figure 3: Regression Diagnosis Figures

The Figure 3 shows model diagnosis plots. The upper-left one is called "Residuals vs Fitted" plot. It shows if residuals have any kind of linear or non-linear patterns. In this case, the fitted red line lies around 0, which means residuals are randomly distributed and has homoscedasticity. The plot at the top-right corner shows if residuals are normally distributed. In this case, residuals follow a straight line well. I would not be concerned about this assumption too much. At the bottom left of

Final Model

The best regression model are demonstrated below:

$$cnt' = hum + atemp + windspeed + fac(season) + fac(yr) + fac(hr) + fac(workingday) + fac(weathersit) \tag{3}$$

$$where \;\; cnt' = \frac{cnt^{\lambda} - 1}{\lambda} \;\; , \lambda = 0.36 \tag{4}$$

The values of coefficients are showed in Figure 4:

```
Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          2.73954    0.17242  15.889  < 2e-16 ***
hum                 -2.26673    0.16508 -13.731  < 2e-16 ***
atemp                9.66861    0.23172  41.725  < 2e-16 ***
windspeed           -0.10575    0.02505  -4.221 2.44e-05 ***
factor(season)2      1.73178    0.08371  20.688  < 2e-16 ***
factor(season)3      1.33697    0.10603  12.609  < 2e-16 ***
factor(season)4      2.57395    0.07283  35.341  < 2e-16 ***
factor(yr)1          2.68933    0.04686  57.388  < 2e-16 ***
factor(hr)1         -1.95808    0.16060 -12.192  < 2e-16 ***
factor(hr)2         -3.31267    0.16115 -20.557  < 2e-16 ***
factor(hr)3         -4.79187    0.16221 -29.541  < 2e-16 ***
factor(hr)4         -5.54712    0.16234 -34.170  < 2e-16 ***
factor(hr)5         -2.92257    0.16132 -18.117  < 2e-16 ***
factor(hr)6          1.63090    0.16088  10.137  < 2e-16 ***
factor(hr)7          6.86104    0.16060  42.720  < 2e-16 ***
factor(hr)8         10.82161    0.16047  67.435  < 2e-16 ***
factor(hr)9          7.83191    0.16067  48.746  < 2e-16 ***
factor(hr)10         5.81127    0.16124  36.042  < 2e-16 ***
factor(hr)11         6.58868    0.16223  40.612  < 2e-16 ***
factor(hr)12         7.79179    0.16339  47.687  < 2e-16 ***
factor(hr)13         7.61582    0.16428  46.358  < 2e-16 ***
factor(hr)14         7.09639    0.16502  43.004  < 2e-16 ***
factor(hr)15         7.42775    0.16524  44.950  < 2e-16 ***
factor(hr)16         9.23001    0.16491  55.968  < 2e-16 ***
factor(hr)17        12.48683    0.16415  76.069  < 2e-16 ***
factor(hr)18        11.83778    0.16331  72.485  < 2e-16 ***
factor(hr)19         9.51747    0.16203  58.740  < 2e-16 ***
factor(hr)20         7.40212    0.16131  45.887  < 2e-16 ***
factor(hr)21         5.79202    0.16075  36.031  < 2e-16 ***
factor(hr)22         4.33405    0.16050  27.003  < 2e-16 ***
factor(hr)23         2.34876    0.16039  14.644  < 2e-16 ***
factor(workingday)1  0.10910    0.04995   2.184    0.029 *
factor(weathersit)2 -0.25626    0.05762  -4.447 8.76e-06 ***
factor(weathersit)3 -2.68381    0.09855 -27.232  < 2e-16 ***
factor(weathersit)4 -1.51214    1.76583  -0.856    0.392
```

Figure 4: Values of Coefficients

Some basic statistics of fitted model:

Basic Statistics of Model			
Adjusted-R-square	Residual Standard Error	F-stat	P-value of F-stat
0.8037	3.053	2088	<2e-16

interpretation

- Interpretation of Coefficients of Environmental-condition variables

First, we looked at all environmental variables included in our model. Those are continuous variables: *hum*, *atemp* and *windspeed*, and categorical variables: *weathersit*.

About *hum*, the estimated coefficient is -3.09, which means when humidity increase 1 unit, the *cnt* will decrease 1 unit averagely. That is, people prefer to ride bikes when humidity is low. About *atemp*, feeling temperature in Celsius, the coeffecient is -10, which means people tend to ride bikes when weather is relatively cool. The coefficient of *windspeed* also make sense. It is a negtive value, -0.12, which means windspeed has slightly negative influence on the number of bikes rented by users. About *weathersit*, according to Figure 5, we could interpret that when weather condition is clear, few clouds, people tend to ride bikes. All other weather conditions have negative influence on *cnt*.

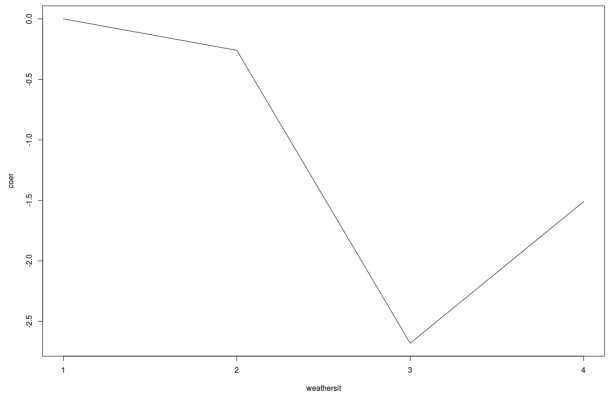

Figure 5: Coefficients of weathersit

- Interpretation of Coefficients of Time variables

Then, we focus on interpreting Time variables. They are *season* , *yr* , *hr* and *workingday*. About *season*, based on Figure 6, people seems to be more likely to ride bikes in winter. Averagely, when season is winter, the intercept will add 2.57 compared to spring, which is the base case. About *yr*, the coefficient is 2.68, it seems that the bike sharing company is expanding, and they have more customers years after years. About *hr*, we could interpret some patterns from Figure 7. There are two peaks, 9 o'clock and 18 o'clock. Intuitively speaking, they are exactly when people go to office and leave office. Besides, at 5 o'clock, which is midnight, people are not very likely to ride bikes around that time period. Finally, about *workingday*, averagely, *cnt'* is 0.1 unit higher in workingdays compared with holidays and weekends. That also makes sense.

- Interpretation of Basic Statistics of Regression Model

First, we looked at the p-values of coefficients. The p-value for each term tests the null hypothesis that the coefficient is equal to zero. A low p-value indicates that we can reject the null hypothesis. In other words, a predictor that has a low p-value is likely to be a meaningful addition to your model because changes in the predictor's value are related to changes in the response variable. In this case, only the p-value of the coefficient of weathersit when it equals to 4 is larger than 0.05, which is 0.392. So, I am confident enough that most of the coefficients are significantly not 0.

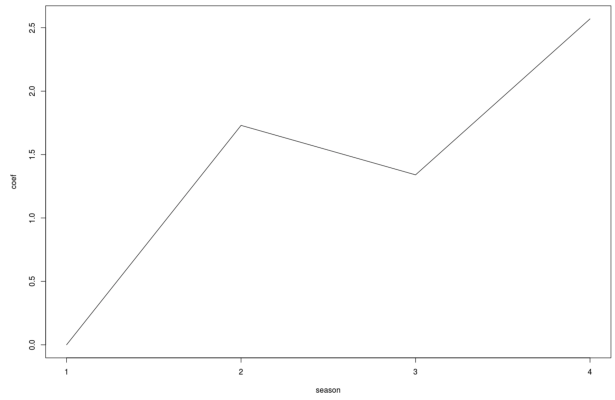

Figure 6: Coefficients of factor(season)

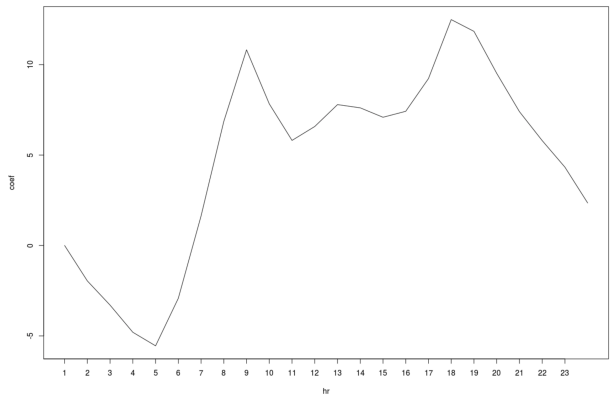

Figure 7: Coefficients of factor(hr)

Secondly, the R-square is the percentage of the response variable variation that is explained by a linear model. In this case, adjusted r-square is 0.8, which indicates that the model explains most the variability of the response data around its mean. The model fits data very well.

Thridly, the P value of F-test is smaller than 2e-16. We could not tell anything wrong about this model from F-test.

In summary, we could be confident to the result of this model.

Discussion

- Casual Analysis

In this case, I think we could view our fitted parameters as causal effects, because they are time variables and weather-conditon variables. There is no way to influence or change those variable by changing the number of bikes rented by users. In other words, theoretically, weather and time could cause the change of the number of rented bikes. Conversely, it is not possible. More importantly, in reality, when people decided if they want to ride bikes, they do concern about the weather. And people tend to ride bikes when they go to work and go back home. This casuality makes sense.

- Guidance to further study

In my model, the effect of categorical variables on the intercept of the regression model is significant. However, they may effect the slope of regression. There may be interaction terms in the "true model".

Conclusion

According to the interpretation of the regression model, the bike sharing company could deploy more bikes around 9AM and 6PM on working days with good weather condition. They should also increase the number of their available bikes in winter and invest more on next year.

Appendix A

Some R code

```
############ data preparing #################

dat_o<-read.csv(file="BikeSharingDataset.csv",header=T)
dat<-dat_o[,-c(1:2)]

corM<-cor(dat[,c("temp","atemp","hum","windspeed","cnt")])
library(corrplot)
corrplot(corM, method="number")


library("PerformanceAnalytics")
chart.Correlation(dat[,c("temp","atemp","hum","windspeed","cnt")], histogram=TRUE, pch=19)


# temp & atemp, outliers

# atemp & hum, outliers
dat<-dat_o[-c(14132:14155,1552:1573),]
chart.Correlation(dat[,c("temp","atemp","hum","windspeed","cnt")], histogram=TRUE, pch=19)
chart.Correlation(dat[,c("atemp","hum","cnt")], histogram=TRUE, pch=19)


# data transformation, cnt Box-cox
library(EnvStats)
temp1<-boxcox(dat$cnt, lambda = c(-2,2), optimize=TRUE)

y<-dat$cnt
lambda<-0.36
```

```r
y<-(y^lambda-1)/lambda
par(mfrow = c(1, 1))
plot(density(dat$cnt))
plot(density(y))

dat$cnt<-y
# data transformation, atemp
x=dat$atemp
hh<-abs(0.6-x)
dat$atemp<-hh



### prepare data frame
dat1<-dat[,-c(1:2,15:16)]




################## chooose varibale by stepwise ##############

null.lm <- lm(cnt~ atemp + hum,dat1)

#summary(SignifReg(fit1))
#full.lm <- lm(cnt~hum+atemp+hr+yr+season+weathersit,dat1)
# AIC
step(null.lm, scope=cnt~ hum + atemp + windspeed +
       factor(season)+factor(yr)+
       factor(hr)+factor(workingday)+factor(weathersit), direction="both",k=2)
# BIC
step(null.lm, scope=cnt~ hum + atemp + windspeed +
       factor(season)+factor(yr)+
       factor(hr)+factor(workingday)+factor(weathersit), direction="both",k=log(dim(dat1)[1]))



#AIC,all

#BIC # excluded workingday
#lm(formula = cnt ~ atemp + hum + factor(hr) + factor(yr) + factor(season) +
#     factor(weathersit) + windspeed, data = dat1)


################ choose model by CV ###################

library(faraway)
library(caret)
```

```r
n <- dim(dat1)[1]
fold_n <- 10
folds<-createFolds(dat1$cnt, k=fold_n, list=T)


MSE1<-rep(NA, fold_n)
MSE2<-rep(NA, fold_n)

#AIC
for(i in seq(1,fold_n)){
  m1_1 <- lm(cnt~ hum + atemp + windspeed +
               factor(season)+factor(yr)+
               factor(hr)+factor(workingday)+factor(weathersit), data=dat1[-folds[[i]],])
  preds <- predict(m1_1, dat1[folds[[i]], ])
  yval <- dat1[folds[[i]], "cnt"]
  MSE1[i] <- 1 / length(folds[[i]]) * sum((preds - yval)^2)

}

# BIC
for(i in seq(1,fold_n)){
  m1_1 <- lm(formula = cnt~ hum + atemp + windspeed +
               factor(season)+factor(yr)+
               factor(hr)+factor(weathersit), data=dat1[-folds[[i]],])
  preds <- predict(m1_1, dat1[folds[[i]], ])
  yval <- dat1[folds[[i]], "cnt"]
  MSE2[i] <- 1 / length(folds[[i]]) * sum((preds - yval)^2)

}

print(
  c("MSE of Model_AIC", mean(MSE1),
    "MSE of Model_BIC", mean(MSE2) )
  )  # aic is slightly better



################# model diagnose ################

fit1 <- lm(formula = cnt~ hum + atemp + windspeed +
             factor(season)+factor(yr)+
             factor(hr)+factor(workingday)+factor(weathersit), data=dat1)
summary(fit1)
#par(mfrow = c(1, 1))
plot(fit1) #outliers without influence
```

```
################# make plots ########################
weathersit<-c(1,2,3,4)
coef<-c(0,-0.26,-2.68,-1.51)
df1<-data.frame(weathersit,coef)
plot(weathersit,coef,type="l",xaxt="n")
axis(1, at = seq(1, 4, by = 1))


season<-c(1,2,3,4)
coef<-c(0,1.73,1.34,2.57)
plot(season,coef,type="l",xaxt="n")
axis(1, at = seq(1, 4, by = 1))

hr<-c(seq(0:23))
coef<-c(0,-1.96,-3.3,-4.80,-5.55,-2.92,1.63,6.86,10.82,7.83,5.81,6.58,7.79,7.61,7.09,7.42,
        9.23,12.49,11.84,9.52,7.4,5.8,4.33,2.35)
plot(hr,coef,type="l",xaxt="n")
axis(1, at = seq(0, 23, by = 1))
```