Bike Sharing in Washington D.C.

MA 575 Final Project
Professor Kolaczyk
Dec 12th, 2017

Author: Ruochen Ji, Zijian Wang, Ruiheng Jiang, Mingda Han, Yubei Ding

Contents:

Footnote:

Author Contributions: Yubei Ding performed research; Ruochen Ji, Yubei Ding, Mingda Han analyzed data; Ruochen Ji, Zijian Wang developed the structure model; Ruochen Ji, Zijian Wang provided project development and support; Zijian Wang, Yubei Ding, Mingda Han, Ruiheng Jiang wrote the paper.

Abstract:

In this project, we used Bike Sharing Dataset from UCI Machine Learning Repository[1]. Our goal is to build a linear regression model to predict the trend of bike rental in 2012 based on the data of year 2011. We would also like to see the prediction of exact daily count of bike rental in 2012, but it is quite hard to fulfill since we are using only one year's data to build our model. To train a model, we need to do data cleaning, investigate and adjust leverage points, transform model, and select valid variables. Our final model successfully captured the trend and indicated that there is a positive jump of total count of bike rental between 2011 and 2012. The peak of total bike rental occurs in early summer(April-May) and late summer(August-September).

Introduction:

In the modern age, with the rapid growth of human population and car population, traffic has become a serious and inevitable issue. What is more, the pollution caused by tremendous transportations has been a threat to our health. Thus, an alternative transportation for our daily traveling is required.

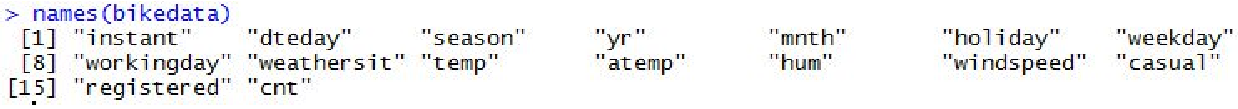
The idea of bike rental and sharing was first emerged in Germany, and later became a worldwide thing. A company named Capital Bikeshare officially carried out the bike sharing system in Washington, DC in 2010. Unlike the traditional bike rentals, the whole process of renting, using, and returning for this new generation has become automatic. Users have access to rentals and returns at any stations in the network 24 hours a day for 365 days of a year . Another great thing that comes with this alternative transportation is that it helps reducing the amount of pollution by increase the number of people ride and decrease the number of people drive.

Bike sharing system is highly correlated to the environmental and seasonal settings. So, for this project, we are planning on to use the data (weather conditions, precipitation, day of week, season, hour of the day, etc.) collected by Hadi Fanaee-T to analyze the daily count of bike rental to better understand the the rental behavior.

Background:

The data we used for this project was compiled by Hadi Fanaee-T, at University of Porto. The core data set(rental time, date, etc.) is related to the two-year historical log corresponding to years 2011 and 2012 from Capital Bikeshare system, Washington D.C., USA. Corresponding details on weather and holiday status was gathered from Freemeteo and the DC Department of Human Resources, respectively.

The objective of this project is to build a model to predict the total number of bike rentals daily from Capital Bikeshare based on the environmental and seasonal settings. Following the project requirements, we will train a model on the data of year 2011 and test it on the data of year 2012. Prediction could be challenging because using one year's data to predict another whole year is not sufficient and it is hard to capture the time trend in just two years.



```
> names(bikedata)
 [1] "instant"    "dteday"     "season"     "yr"         "mnth"       "holiday"    "weekday"
 [8] "workingday" "weathersit" "temp"       "atemp"      "hum"        "windspeed"  "casual"
[15] "registered" "cnt"
```

Figure 1: Response Variable and Predictors

Figure 1 above shows the response variable and predictors of our model. As you can see, date, season, year, month, holiday, and all other variables more or less play certain roles in deciding the daily counts of bike rental. Just for your convenience, "weathersit" is a dummy variable represents the condition of weather(1: clear or partly cloudy, 2: mist and cloudy, 3: light snow or light rain, 4: heavy snow, heavy rain, ice pallet or thunderstorm); "temp" represents normalized temperature in degree Celsius, and the values are divided to 41(max); "atemp" represents normalized feeling temperature in Celsius, and the values are divided to 50(max); "hum" represents normalized humidity, and the values are divided to 100(max); "casual" represents the count of casual users; "registered" represents the count of registered users; "cnt" represents the total count of rental bikes. In the data, we decided to create only one model to count the total number of bike rentals including both casual and registered users.

Modeling and Analysis:

In our analysis, we loaded the data and plotted the scatter plot matrix first. From the data. We noticed that instant and date are representing the same thing. Therefore we did not include date in our scatter plot. We also exclude the year factor since we only use data of 2011 for our training and validation set and use data of 2012 for our testing data set. Other than that, casual and registered also been excluded by us since count represents casual plus register. The graph below is the corresponding scatter plot:

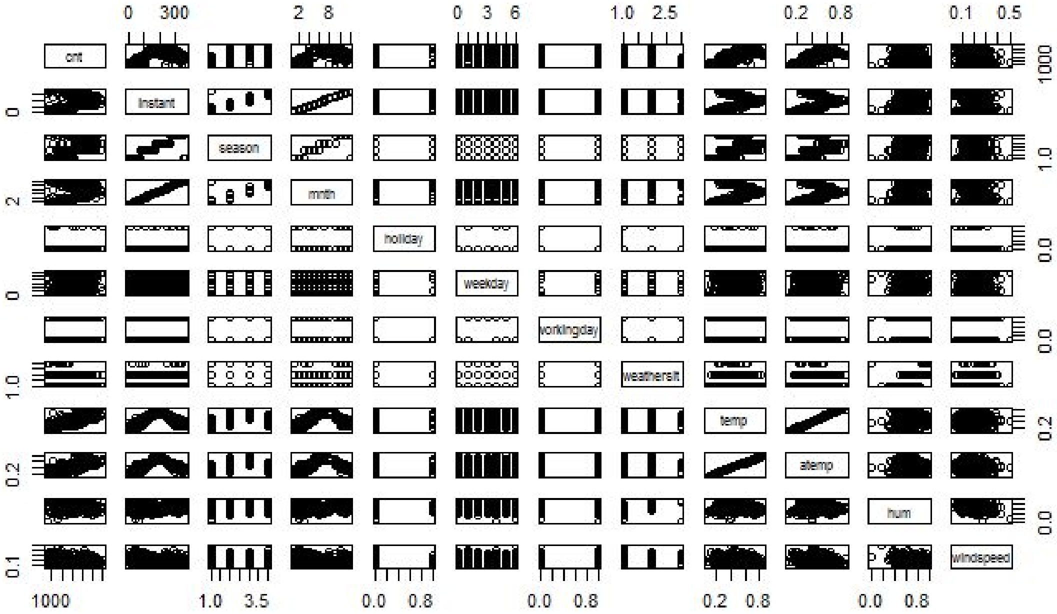


Figure 2: Scatterplot Matrix of all predictors and response variable.

A lot of information can be observed from the scatterplot matrix (Figure 2). Temp and atemp appears to be highly correlated. Instant, season and month also appears to have a positive correlation. We decided to keep one of the predictors from each high correlated variables group because having them all in a model will affect the estimation of parameters due to collinearity.

Then we analyzed the relationship between response variable (cnt) and predictors. The categorical variables such as month should be factorized because January and December suggested a low number of count of rentals while in the middle of the year the rentals reach the max amount. Other numeric predictors like temperature and humidity etc. have shown that there is a trend in the plot. However, some graph such as temperature vs counts appears to have a quadratic pattern which we may need some transformation in future modeling and analysis.

We began by fitting our very first OLS model. We include instant in the model hoping that it could capture some time effects because number of users may increase along with time. However, the parameter of instant is -4.414, which is quite small relative to other parameters and it has a large P-value, 0.223. We believe this is because instant is getting too large by the end of the year (from 1 to 365) so it is not rational to use it as an estimator of bike daily rentals, and it is excluded from our model.

In order to choose using month (mnth) or season, we found that residual standard error for the model with month which is smaller than for the model with season. Also from the "Count vs Fitted value" plots we found that month is giving better prediction. The p-value of each month is significant(less than 0.05), however, the p-values seasons are not significant. Month can totally explain the trend of season while season cannot explain the trend of each month relatively. Even the season has smaller inflation factor in this situation, we still choose to keep month due to the reasons above. Therefore, we conclude that month is better and more accurate. Among temperature and feeling temperature(atemp), we choose to keep temperature. Both of these predictors make sense, however, we believe most people will check the next day's temperature before they choose their next day transportation tool.

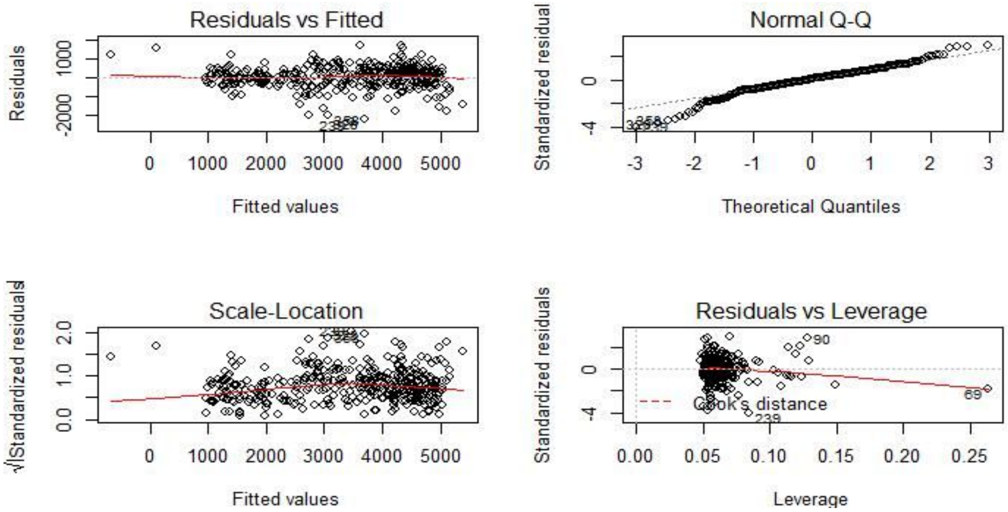


Figure 3: Diagnostic Plot

During the first modeling, from the diagnostic plot (Figure 3), we observed that some points are away from other points which could be leverage points or outliers. So we go deep to investigate them. Instant 69 appeared to be a high leverage point and we found that the humidity has a 0 value in that row. Given the fact, weathersit is 3 (Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds) on that day which humidity should be relatively high. So we conclude that instant 69 has a missing data of humidity. We found humidity index are relatively high when weathersit= 3. So we created a subset data with weathersit=3 and used spline to predict missing data in humidity. We got a value of 0.8931057 which is rational in a raining or snowing day. So we fix the data with the predicted value.

From the residual plot, points 90, 185, 239, 328, 359 have relative higher leverage. After investigation, we find out that point 185 is fourth of July, people may ride bike to attend celebrating activities. 328 is for the Thanksgiving Day and point 359 is for the Christmas day. Since people stay at home with their families during these national holidays, the two points can be forgiven. For point 239, it is on 2011/8/27, at that day hurricane Irene hit the northeastern part of America, which will lead people stay at home. Point 90 has a relatively low number in count which could because of the bad weather, we decide to continuing modeling with point 90 since the cnt number on this day is not too far from the rest of the data and is within an acceptable range.

After data cleaning and our very first modeling, we went a few tracks to improve our model:

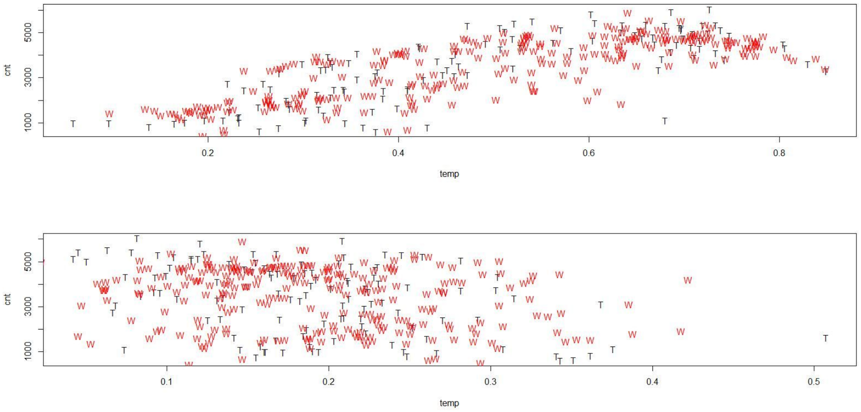


Figure 4: Interaction plot

We checked the interaction effect of working day and other predictors. From Figure 4 above, the pattern shows there is no difference when working day= 0 or 1. So we concluded that there is no need to add interaction terms. However, the working day variable in the model has relative low parameter estimate and high p-value(0.214). It is not significant factor for the model. We tried to change it to holiday and the result gets better. During holiday, people may stay at home or go to travel, so less people will rent bikes, however, the difference of people who rent bike does not differ much on working=0 or 1. This analysis result matches with the 2011 member survey report of Capital Bike Sharing Program[2], which indicated that more than 56% of the bike trips are for recreational purpose, such as meeting a friend. In other words, holiday have a larger effect on rental count than working day.

We tried to run generalized least square with AR(1) model which generally is in use to model time-varying processes in which one term correlates with its previous term to check the effect of time. The P-value of a few parameter increased and our residual standard did not get better, so we did not use GLS to fit our model.

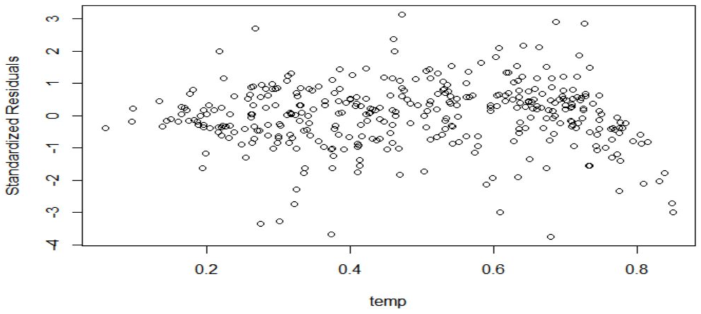


Figure 5: Residual Plot of Predictor Temp

We took a look at the residual plots with respect to each chosen variable. We noticed that the standardized residual plot of temperature (Figure 5) shows a quadratic pattern. Therefore we decide to add a quadratic term of temp to our model. Compare with our first model, we got a better R-squared value and P-value in parameters.

After transformation in temperature, we analyzed the response plot. As shown in Figure 6 below, R gave the best transformation on Y. However, the λ is not that small, and it would be easier to interpret our model so we decided to keep our original model.

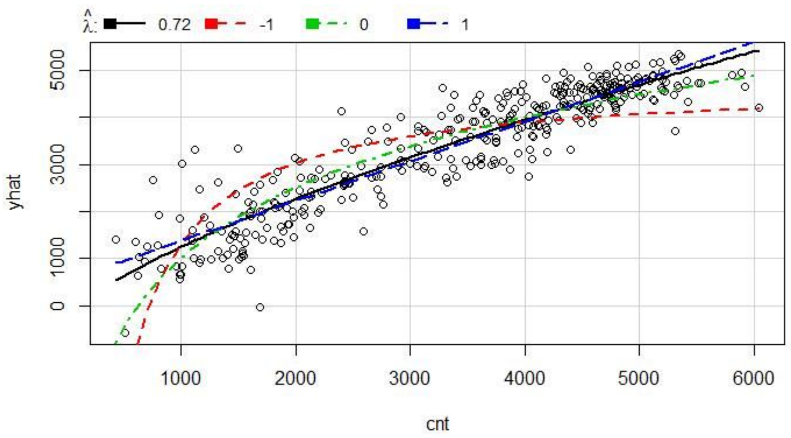


Figure 6: Y-Transformation Response Plot

Next, we used stepwise selection by AIC to select our predictors. Through comparing the value of AIC, R actually did not eliminate any predictors which indicates that there are no redundant variables in the model.

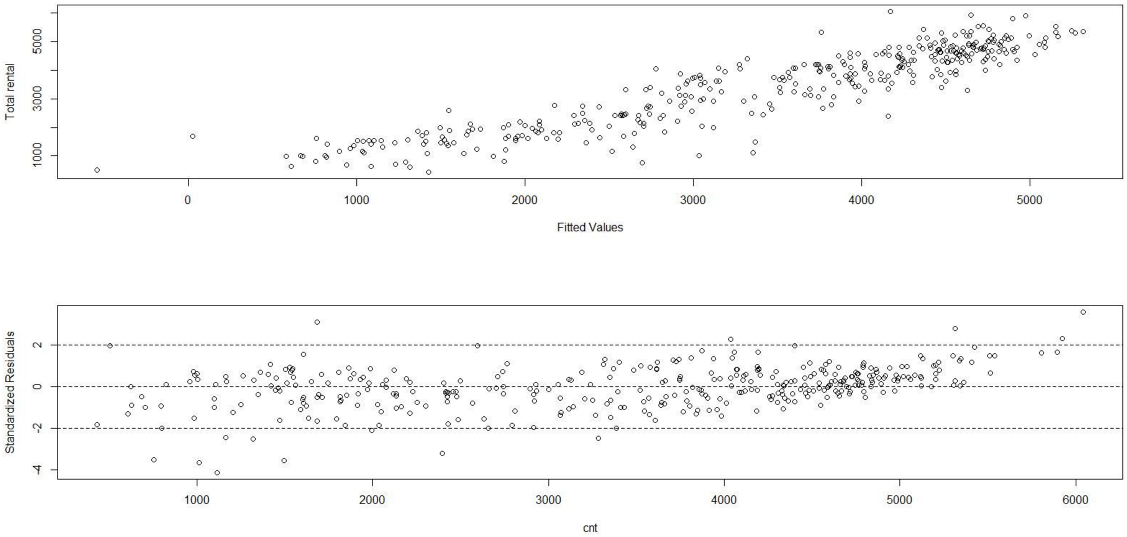
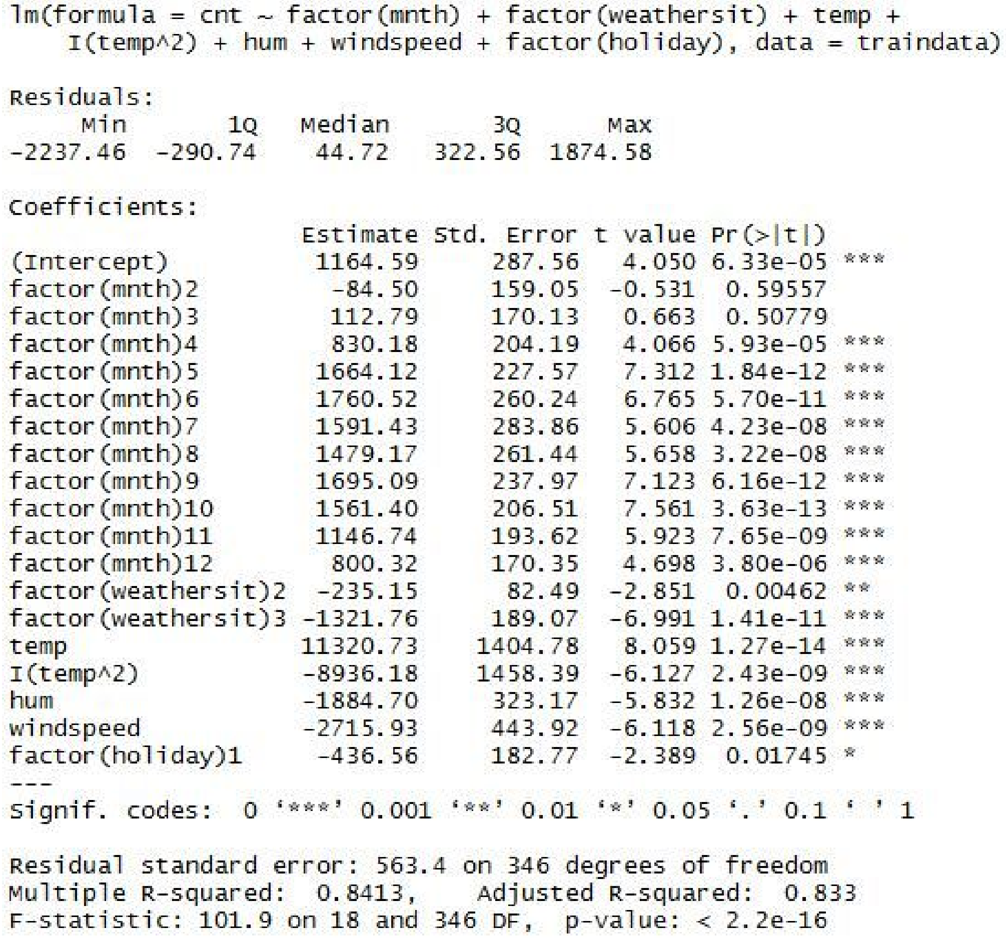


Figure 7: Plot of cnt vs. Fitted Values and Residual Plot of cnt



```
lm(formula = cnt ~ factor(mnth) + factor(weathersit) + temp +
    I(temp^2) + hum + windspeed + factor(holiday), data = traindata)

Residuals:
     Min       1Q   Median       3Q      Max
-2237.46  -290.74    44.72   322.56  1874.58

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          1164.59     287.56   4.050 6.33e-05 ***
factor(mnth)2         -84.50     159.05  -0.531  0.59557
factor(mnth)3         112.79     170.13   0.663  0.50779
factor(mnth)4         830.18     204.19   4.066 5.93e-05 ***
factor(mnth)5        1664.12     227.57   7.312 1.84e-12 ***
factor(mnth)6        1760.52     260.24   6.765 5.70e-11 ***
factor(mnth)7        1591.43     283.86   5.606 4.23e-08 ***
factor(mnth)8        1479.17     261.44   5.658 3.22e-08 ***
factor(mnth)9        1695.09     237.97   7.123 6.16e-12 ***
factor(mnth)10       1561.40     206.51   7.561 3.63e-13 ***
factor(mnth)11       1146.74     193.62   5.923 7.65e-09 ***
factor(mnth)12        800.32     170.35   4.698 3.80e-06 ***
factor(weathersit)2  -235.15      82.49  -2.851  0.00462 **
factor(weathersit)3 -1321.76     189.07  -6.991 1.41e-11 ***
temp                11320.73    1404.78   8.059 1.27e-14 ***
I(temp^2)           -8936.18    1458.39  -6.127 2.43e-09 ***
hum                 -1884.70     323.17  -5.832 1.26e-08 ***
windspeed           -2715.93     443.92  -6.118 2.56e-09 ***
factor(holiday)1     -436.56     182.77  -2.389  0.01745 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 563.4 on 346 degrees of freedom
Multiple R-squared:  0.8413,    Adjusted R-squared:  0.833
F-statistic: 101.9 on 18 and 346 DF,  p-value: < 2.2e-16
```

Figure 8: Final Model Summary

Since many parameters appear in our model, in order to make the model more precise, we should consider about the parameter penalty scheme. One scheme is to apply the cross validation, which includes LASSO and Ridge Regression. However, the penalized model (ridge model) have a significant larger mean squared error(=435589.2) compared to that of the OLS model(=322983.8).

Our final model:

lm(cnt~factor(mnth)+factor(weathersit)+temp+I(temp^2)+hum+windspeed+factor(holiday)

Prediction:

Since the Capital Bikesharing program just started in 2010, we assume that it expands at a consistent rate in the first several years. Therefore, we decide to shift our model by 1317 units, which is the difference between the cnt on 2011 and our predicted value of first day of 2012, in purpose of simulating the expansion of the program.

In the graph below, black line is the actual data in 2012, green line is the predicted value given by our model with 2012's data. Red line is the model that after we shift our original up. We can see that the our predicted data has the same trend as 2012' data does. The difference between two lines can be explained by the new expanded rental stations. Since we do not have the data about rental station in our model, our prediction result might be less than the actual 2012 result.

In order to seeking why our prediction values are still below the actual values, we read through the 2012 member survey report of Capital Bikesharing Program[3]. We found that from 2011 to 2012, due to (1) increasing demand (2) positive feedbacks from users and (3) increasing using time per user, this program expanded their bikesharing service from D.C and Arlington County to City of Alexandria and increases 10 rental stations (or 150 bicycles) in total. At the same time, the total number of users increases from 18000 to 22000 (about 23%). This fact proves our assumption about the expansion. The actual expansion scale might be larger than what we assumed in our model, which can explain our small prediction values. However, the overall trend we predict does match most of the actual data's trend. If we can use 2012's data to reshape our model, we might have a better way to reshape our model and estimate the increasing intercept between each year.

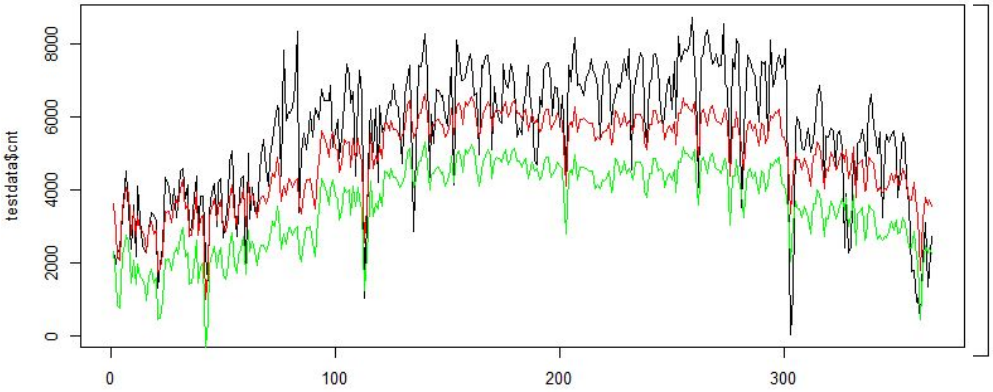
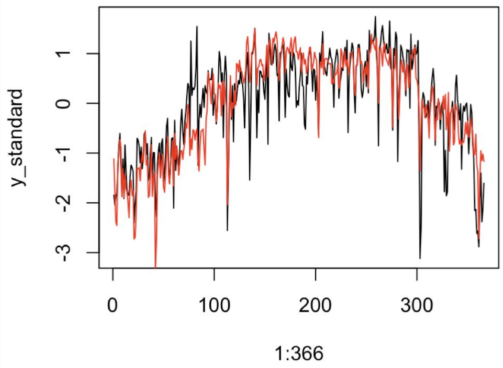


Figure 9: Plot of Actual vs. Prediction Values

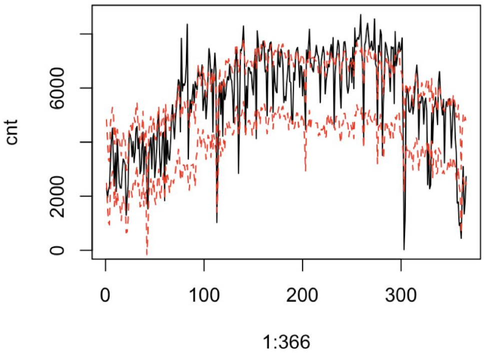
The MSE of our final model: 5069935
The MSE of our final model after shift: 1512536
We can observe that after we shift our model, the MSE largely decreases, which means our model fits better than before on 2012's data. Figure 10 is a plot after we standardized our predicted value and Y value:



We can see that our model still can't catch some extreme values in 2012's data, and 2011 and 2012 are total different years. We should assume that this situation is normal, because the variance between 2011 and 2012 might be different.

To see how well we fit our shifted model on 2012's data, we drawed another 95 percent prediction interval plot, and we compare them to our 2012's data:

Figure 10: Plot of Standardized 2012 cnt VS. Standardized Predicted Value



From the plot we observe that our shifted model cover most points of 2012's data. By using R, we get the result that 219 points from 2012's data falls in our shifted model's 95% prediction interval, which means our model approximately cover 60% of 2012's data.

Again, the coverage rate is not ideal since we don't have have other years' data to predict the different variance between each year. If we can add 2012's data to our train dataset, our model will perform better in estimating the mean shift.

Figure 11: Plot of 2012 cnt vs. 95% Prediction Interval

Discussion:

We have conducted extensive OLS analysis for Bike data and gotten a relative good prediction for 2012 bike rental. We noticed an increasing trend for the first 3 months of a year. Then, the rental amount kept high through March to October and continued decreasing for the last two months. It is intuitive to believe that when days get cold, less people choose to bike. when days are cold, installing wind/rain shield device or handles with heating function might be an effective attempt that will please both program runners and users, since it can increase the number of bike rental at the same time that users can bike with comfort in cold days. From an investor view, investment on terms from April to September might receive higher payback with lower risk. And we may also predict the total bike rental in the future to balance the demand and supply. The investors would be happy if they know which time interval are more profitable, and which period they should increase their production level.

Even though our prediction trend is pretty close to the actual trend, the individual prediction values are not that close to the actual values. We still have large space and possibilities to improve the model and prediction. For example, once we see the plot of data with respect to time, it is intuitive to believe that the bike rental data has some time series features since the data shows regular up and down patterns along. Unfortunately, the only time series model AR(1) we learned did not turn out to be a better model. Instead of AR(1) model, there might be other time series models that will give us a more satisfying results.

Appendix:

Selected R code:

```
traindata<-bikedata[1:365,]
testdata<-bikedata[366:731,]

pairs(~cnt+mnth+season+holiday+weekday+workingday+weathersit+temp+atemp+hum+windspeed,data=traindata)

prototype <- lm(cnt ~ factor(mnth) + factor(workingday) +factor(weathersit)+temp+hum+windspeed,data=traindata)
summary(prototype)
prototype1 <- lm(cnt ~ factor(season) +factor(workingday) +factor(weathersit)+temp+hum+windspeed,data=traindata)
summary(prototype1)
plot(prototype)
plot(prototype$fitted.values,traindata$cnt,xlab="Fitted Values", ylab="Total rental")

StanRes <- rstandard(prototype)
par(mfrow=c(1,3))
```

```
plot(mnth,StanRes, ylab="Standardized Residuals"); abline(h=0, lty=2, col='red')
plot(temp,StanRes, ylab="Standardized Residuals");abline(h=0, lty=2, col='red')
vif(prototype)

plot(temp[workingday==0],cnt[workingday==0],pch=c("T"),col=c("black"),ylab="cnt",xlab="temp")
points(temp[workingday==1],cnt[workingday==1],pch=c("W"),col=c("red"))
plot(windspeed[factor(workingday)==0],cnt[factor(workingday)==0],pch=c("T"),col=c("black"),ylab="cnt",xlab="temp")
points(windspeed[factor(workingday)==1],cnt[factor(workingday)==1],pch=c("W"),col=c("red"))

prototype3 <- lm(cnt ~ factor(mnth)  +factor(weathersit)+temp+I(temp^2)+hum+windspeed+factor(holiday),data=traindata)
summary(prototype3)
plot(prototype3)
plot(prototype3$fitted.values,cnt,xlab="Fitted Values", ylab="Total rental")
StanRes3 <- rstandard(prototype3)
plot(cnt,StanRes3, xlab="cnt", ylab="Standardized Residuals")
abline(h=seq(-2,2, by =2),lty=2)

fit1 <- prototype3$fitted.values
m2 <- lm(cnt~fit1 + I(fit1^2))
plot(fit1,traindata$cnt,xlab="Fitted Values", col='gray75', pch = 19, cex=.75)
fitnew <- seq(0,6000,len=6000)
lines(fitnew,predict(m2,newdata=data.frame(fit1=fitnew)),col='red',lwd=2)
abline(lsfit(prototype3$fitted.values,cnt),lty=2, col='blue', lwd=2)
library(alr3)
inverse.response.plot(prototype3,key=TRUE)

# AR(1) model
library(nlme)
RefTime = seq(1,nrow(traindata),len=nrow(traindata))
prototype5 <- gls(cnt ~factor(mnth)  +factor(weathersit)+temp+I(temp^2)+hum+windspeed+factor(holiday)+ RefTime,
correlation=corAR1(form=~RefTime),method="ML")
summary(prototype5)

# Stepwise selction
library(MASS)
step <- stepAIC(prototype3, direction="both")
step$anova
summary(step)

# Ridge Regression
tempdataset <- data.frame(cnt,factor(mnth), factor(weathersit), temp,  I(temp^2),windspeed,hum,factor(holiday))
x <- model.matrix(cnt~., tempdataset)[,-1]
y <- tempdataset$cnt
lambda <- 10^seq(10, -2, length = 100)
library(glmnet)
```

```
set.seed(123)
train = sample(1:nrow(x), nrow(x)/2, replace= F)
ytest = y[-train];

ridge.mod <- glmnet(x[train,], y[train], alpha = 0, lambda = lambda) # alpha=0 means ridge model
cv.out <- cv.glmnet(x[train,], y[train], alpha = 0)
bestlamlasso <- cv.out$lambda.min; bestlamlasso
# Ridge MSE
ridge.predict <- predict(ridge.mod, s = bestlamlasso, newx = x[-train,])
mean((ridge.predict-ytest)^2)  # standardized mse

m.ols <- lm(cnt~., data = tempdataset, subset=train)
coef(m.ols)
ols.pred <- predict(m.ols, newdata = tempdataset[-train,])
mean((ols.pred-ytest)^2)
# Ridge coefficients
ridge.coef  <- predict(ridge.mod, type = 'coefficients', s = bestlamridge)[1:18,]
ridge.mod <- glmnet(x[train,], y[train], alpha = 0, lambda = bestlamridge)
summary(ridge.mod)

fit = glmnet(x[train,], y[train], alpha = 0, lambda = bestlamlasso)
fit$beta

# Sum of Squares Total and Error
sst <- sum((y - mean(y))^2)

##############################################################

testtest <- data.frame(cnt,factor(mnth), factor(weathersit), temp,
          I(temp^2),windspeed,hum,factor(holiday))
x <- model.matrix(testdata$cnt~., testtest)[,-1]
ridge.predict <- predict(ridge.mod, s = bestlamlasso, newx =x )
mean((ridge.predict-testdata$cnt)^2)

pred = predict(step, newdata = data.frame(testdata), se.fit=T)
lm.pred = pred$fit
mean((lm.pred-testdata$cnt)^2)

plot(1:366, testdata$cnt, type='l')
lines(1:366, lm.pred, col='green')
lines(1:366, lm.pred+1.5*(lm.pred[1]-traindata$cnt[1]) , col='red')

pred = predict(step, newdata = data.frame(testdata), se.fit=T)
lm.pred = pred$fit
sigma = sqrt(sum(step$residuals^2/step$df.residual));
```

```
se=sqrt(pred$se.fit^2+sigma^2)
pred.low = lm.pred - 2*se+1317
pred.up = lm.pred+2*se+1317
plot(1:366, cnt, type="l")
lines(1:366, pred.up, lty=2, col="red")
lines(1:366, pred.low, lty=2, col="red")
mean((lm.pred+1317-testdata$cnt)^2)
sum(cnt>=pred.low & cnt<=pred.up);


#Prediction Missing value (data cleaning):
hum[hum == 0.000000]<-NA
tempdataset <- data.frame(hum,weathersit)
tempdataset.sub <- subset(tempdataset, weathersit > 2)
RefTime = seq(1,nrow(tempdataset.sub),by=1)
tempdataset.sub$RefTime = RefTime
tempdataset2 = tempdataset.sub;
tempdataset.sub<-tempdataset.sub[complete.cases(tempdataset.sub), ]
h=0.5
localpoly<-locpoly(tempdataset.sub$RefTime, tempdataset.sub$hum,bandwidth = h,degree=1)
spline<-sm.spline(localpoly$x, localpoly$y,df=nrow(tempdataset2))
missing = is.na(tempdataset2$hum)
estimatedvalues=predict(spline,tempdataset2$RefTime[missing])
tempdataset2$hum[missing]=estimatedvalues
bikedata <- bikedata1
Estimatedvalues
```

References

[1] UCI Machine Learning Repository http://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset#

[2] Capital Bikeshare 2011 Member Survey Executive Summary, https://d21xlh2maitm24.cloudfront.net/wdc/Capital_Bikeshare_2011_Survey_Executive_Summary.pdf?mtime=20161206135934

[3] 2012 Capital Bikeshare Member Survey Report Executive Summary, https://d21xlh2maitm24.cloudfront.net/wdc/cabi-2012surveyreport-execsum-5-15-13-revtitle.pdf?mtime=20161206135941