STAT4158
Min Yang

Final Project Report

Introduction

The dataset collected the average electricity consumption of the residence. I choose "global_active_power" to be my dependent variable and "sub_metering_1", "sub_metering_2", "sub_metering_3" be my three covariates, and they are agree in the unit.
I choose these three to be my covariates because each variable measure the usage of active energy in certain region as recorded below.

1. onem—sub_metering_1: energy sub-metering No. 1 (in watt-hour of active energy). It corresponds to the kitchen, containing mainly a dishwasher, an oven and a microwave (hot plates are not electric but gas powered).
2. twom—sub_metering_2: energy sub-metering No. 2 (in watt-hour of active energy). It corresponds to the laundry room, containing a washing-machine, a tumble-drier, a refrigerator and a light. 3.
3. threem—sub_metering_3: energy sub-metering No. 3 (in watt-hour of active energy). It corresponds to an electric water-heater and an air-conditioner.

The data contains every minute usage of active power for four years from December 2006 to November 2010, a total of 48 months or 207 weeks. So I think it is reasonable to analyze it based on a weekly usage. Since daily does not make much sense as usage of electricity generally shows a seasonality according to weather, season, and monthly or yearly would provide too few data to analyze.

Data Cleaning

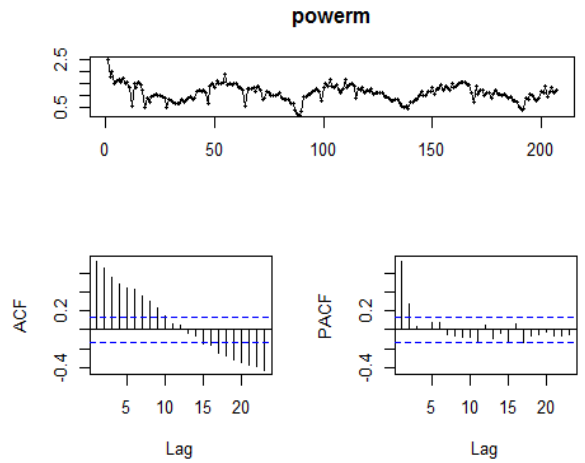
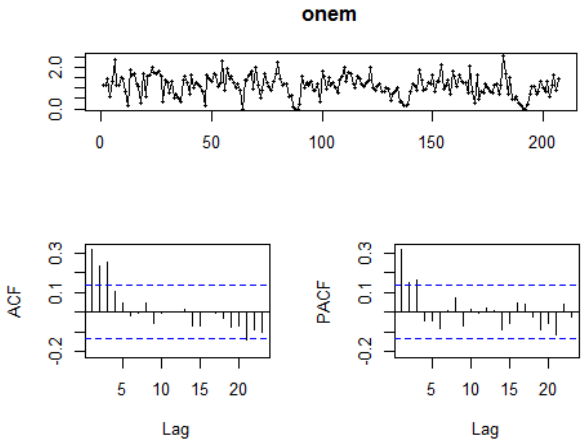
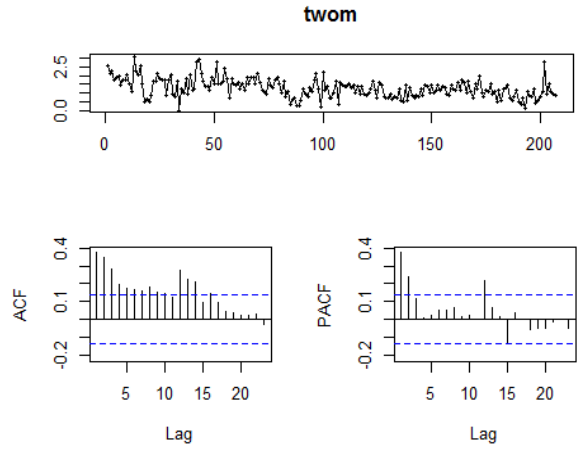
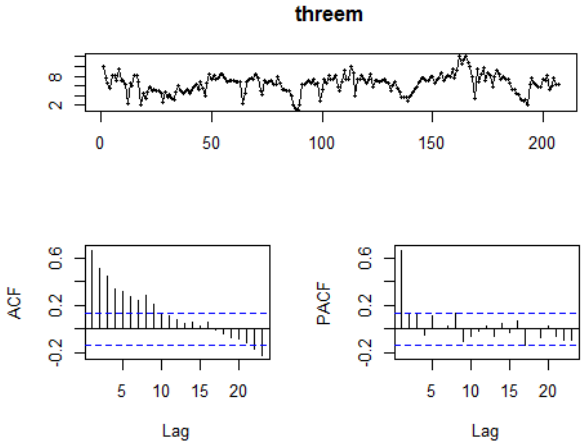
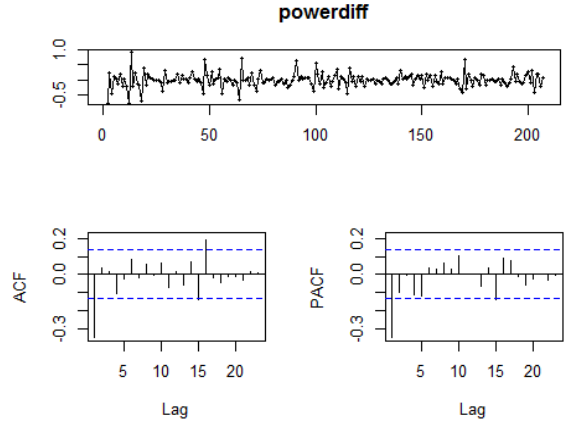
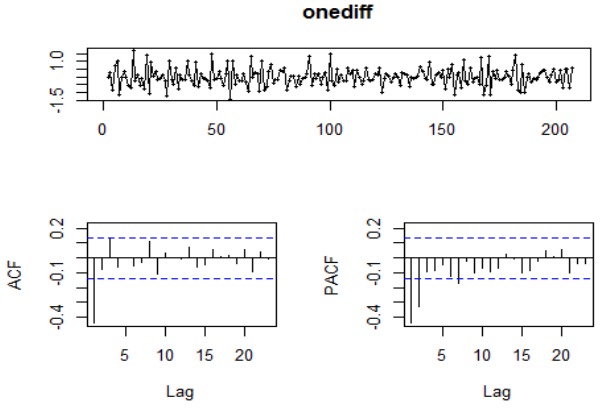
The raw data is generated in minutes, so I convert it to weekly first.
First, there are some missing value, so I fit them using pervious data.

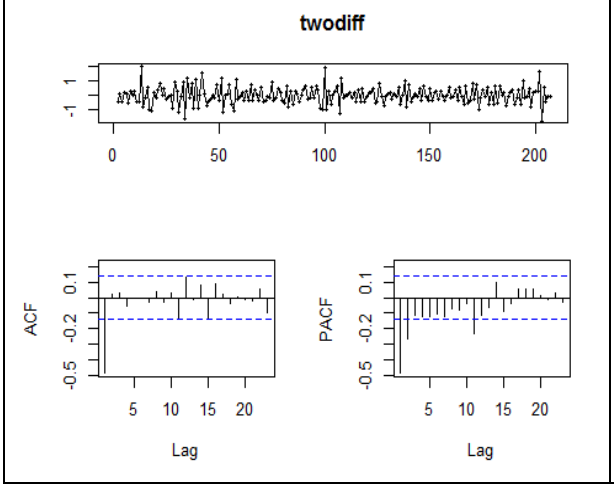
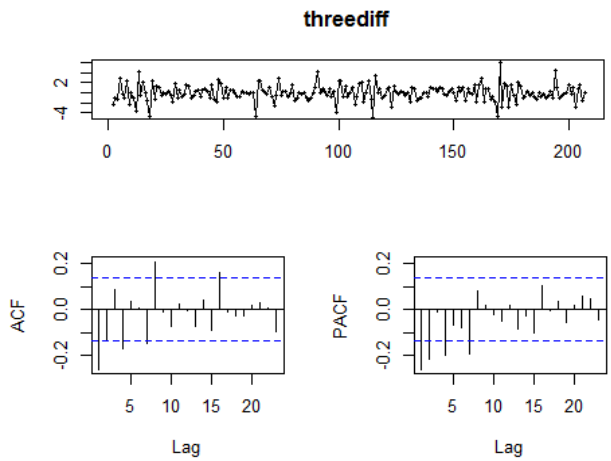
6838	21/12/20...	11:21:00	0.242	0.000	241.670	1.000	0.000
6839	21/12/20...	11:22:00	0.244	0.000	242.290	1.000	0.000
6840	21/12/20...	11:23:00	?	?	?	?	?
6841	21/12/20...	11:24:00	?	?	?	?	?

Since my data involves date, I convert them to zoo first, ordered by date, then to xts which is easier for future analysis.
And then I use apply.weekly and take the mean of the measurement to gather my data of 207 groups of weekly measurements.

Analysis

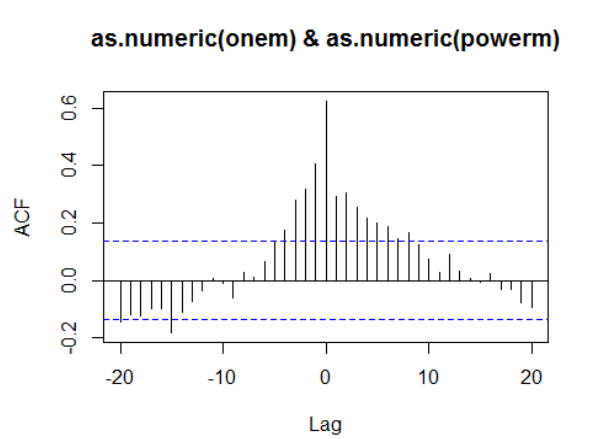
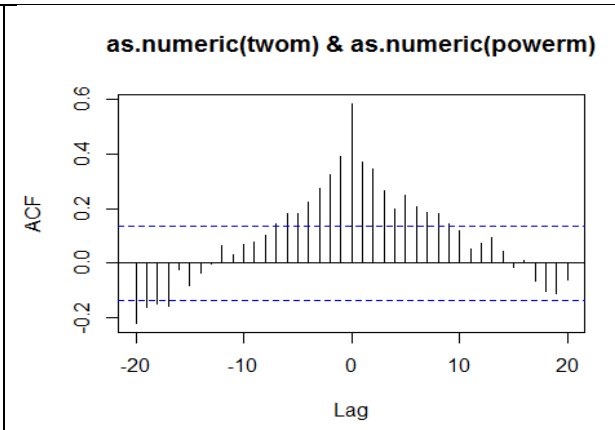
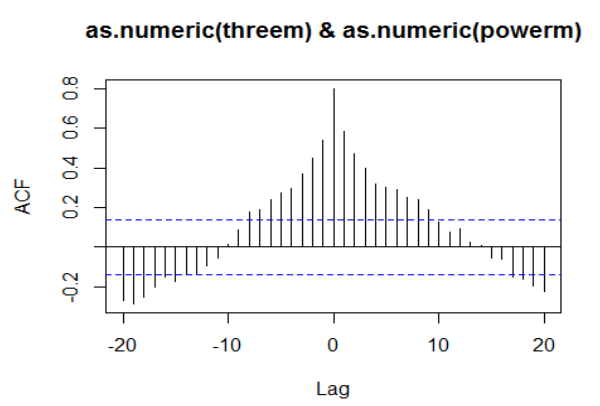
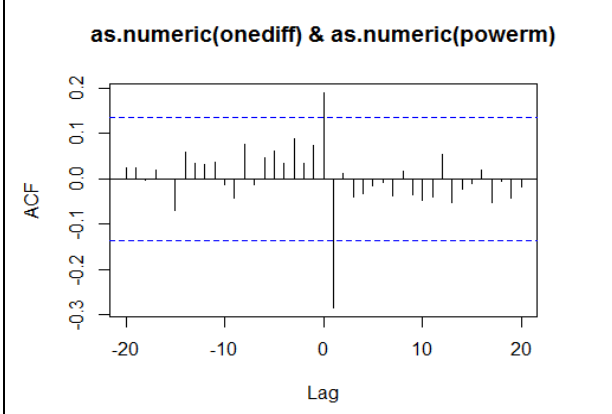
2. acf and pacf of 8 time series

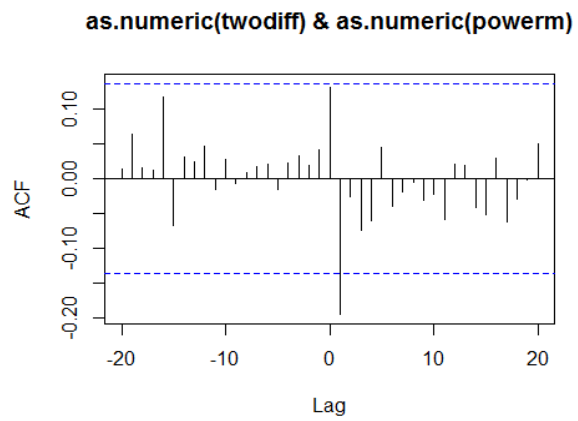
Weekly global active power	Weekly sub metering 1
	
Weekly sub metering 2	**Weekly sub metering 3**
	
Weekly global active power difference	**Weekly sub metering 1 difference**
	

Weekly sub metering 2 difference	Weekly sub metering 3 difference
	

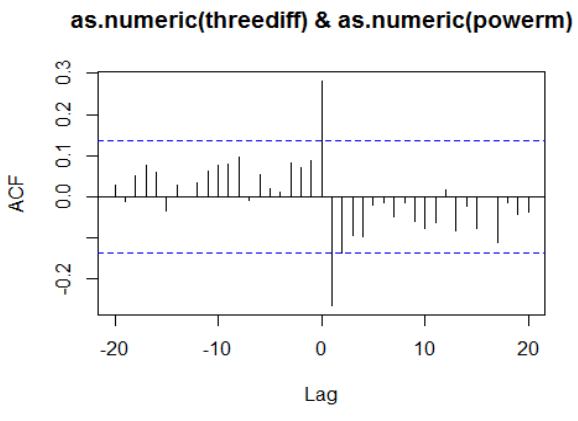
From these acf plots and pacf plots, the powerm, onem, twom, and threem have strong auto-correlation.

3. ccf plot

	
This one has a significant lag at lag = 0	This one has a significant lag at lag = 0
	
This one has a significant lag at lag = 0	This one has a significant lag at lag = 1



This one has a significant lag at lag = 1



This one has a significant lag at lag = 0, 1

4. reasonable linear model

From the above ccf plots, the reasonable linear model for y from X is with variable onem (lag = 0), twom (lag = 0) and onem (lag = 0).

For y from diff(X) is with variable onediff (lag = 1), twodiff (lag = 1), and threediff (lag= 0,1).

6. Acf and pacf of the residuals.

a) X model

```
Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)               0.059844   0.042814   1.398 0.163707
## L(as.numeric(onem), 0)    0.099631   0.029633   3.362 0.000924 ***
## L(as.numeric(twom), 0)    0.187640   0.024336   7.710 5.51e-13 ***
## L(as.numeric(threem), 0) 0.105563   0.007224  14.613  < 2e-16 ***
## Residual standard error: 0.1679 on 203 degrees of freedom
## Multiple R-squared:  0.7584, Adjusted R-squared:  0.7548
## F-statistic: 212.4 on 3 and 203 DF,  p-value: < 2.2e-16

AIC(weeklm)

## [1] -145.3661
```

All three coefficient are significant as their p-value are roughly 0, and all of them have a linear relationship with the electricity consumption.

b) Diff(X) model

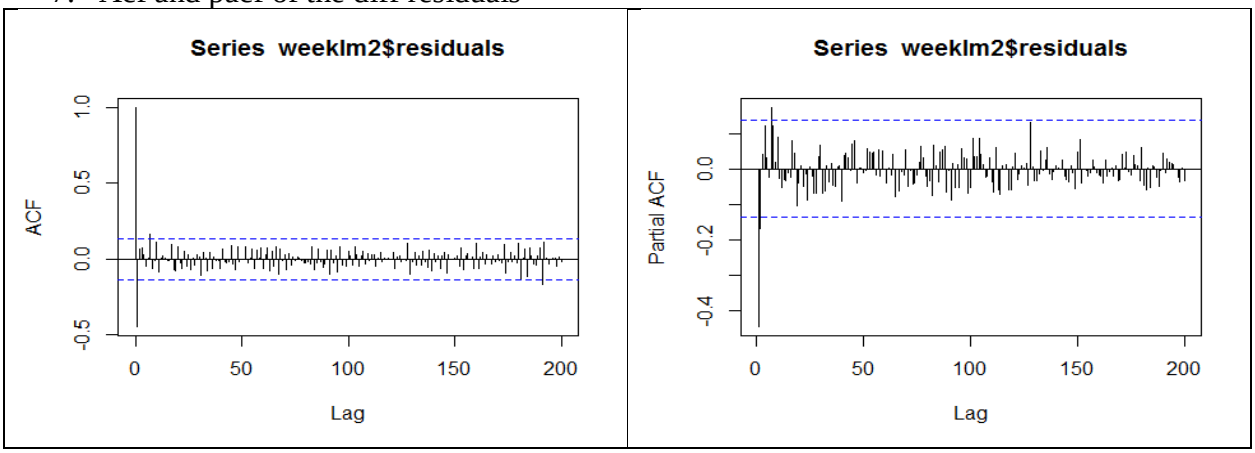
```
Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                -0.004662   0.008097  -0.576    0.565
## L(as.numeric(onediff), 1)   0.130380   0.016890   7.719 5.30e-13 ***
## L(as.numeric(twodiff), 1)   0.067467   0.014596   4.622 6.76e-06 ***
## L(as.numeric(threediff), 1) 0.080564   0.006344  12.698  < 2e-16 ***
## L(as.numeric(threediff), 0)       NA         NA      NA       NA
```

```
## Residual standard error: 0.1162 on 202 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.7534, Adjusted R-squared:  0.7497
## F-statistic: 205.7 on 3 and 202 DF,  p-value: < 2.2e-16
AIC(weeklm2)

## [1] -296.2608
```

All three coefficient are significant as their p-value are roughly 0, and all of them have a linear relationship with the electricity consumption.
Since the Diff(x) model has a smaller AIC, it is a better model than X model, either threediff with lag = 0 or lag = 1 gives the same coefficients and AIC, so we only need to include one of them, in this case, I choose lag = 1.

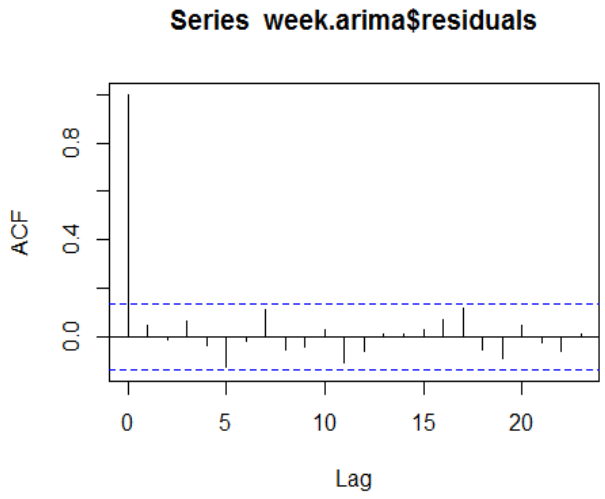
7. Acf and pacf of the diff residuals



The plots do not show correlation, so the model is useful.

8. Fit the best ARIMA model for the residuals

```
Series: weeklm2$residuals
## ARIMA(3,0,2) with non-zero mean
##
## Coefficients:
##          ar1     ar2      ar3      ma1     ma2      mean
##       0.9738  0.0312  -0.1268  -1.6078  0.7413  -0.0006
## s.e.  0.1278  0.1182   0.1062   0.1042  0.1066   0.0074
##
## sigma^2 estimated as 0.009707:  log likelihood=187.68
## AIC=-361.36   AICc=-360.79   BIC=-338.06
```



The acf of this model shows no auto-correlation.
The auto.arima function fits the best model is ARMA (3,2). Thus, there is no further fitting needed, we can apply it to fit a gls model.

9. Fit a gls model

```
## Generalized least squares fit by REML
##   Model: powerm ~ onem + twom + threem
##   Data: weekly
##         AIC       BIC   logLik
##   -328.0576 -294.9256 174.0288
##
## Correlation Structure: ARMA(3,2)
##  Formula: ~1
##  Parameter estimate(s):
##        Phi1        Phi2        Phi3      Theta1      Theta2
##  0.68379837  0.08770059  0.22550313 -0.29932174  0.20526056
##
## Coefficients:
##                 Value Std.Error   t-value p-value
## (Intercept) 0.6644577 0.8542027  0.777869  0.4376
## onem        0.1072522 0.0166055  6.458850  0.0000
## twom        0.0803659 0.0147900  5.433799  0.0000
## threem      0.0902733 0.0053887 16.752182  0.0000
```

10.

The model is y=0.66+0.107X1+0.08X2+0.09X3, all the coefficients are positive related, and all three variates' p-values are roughly at 0, which is smaller than the critical level of 0.01, which is very significant.
From above analysis, R square adjusted=0.7497, so 74.97% of the data is explained by this model.

I also tried two other combinations to see if the model is even more significant.

a) With variable onediff and threediff

```
## Series: weeklm3$residuals
## ARIMA(3,0,2) with zero mean

## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)

## (Intercept)                -0.005116   0.008493  -0.602    0.548

## L(as.numeric(onediff), 1)    0.141767   0.017528   8.088 5.42e-14
***
## L(as.numeric(threediff), 1)  0.087369   0.006474  13.496  < 2e-16
***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1219 on 203 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.7273, Adjusted R-squared:  0.7246
## F-statistic: 270.7 on 2 and 203 DF,  p-value: < 2.2e-16

AIC(weeklm3)

## [1] -277.5486
```

b) With variables twodiff and threediff

```
Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                -0.003902   0.009191  -0.425    0.672
## L(as.numeric(twodiff), 1)    0.083899   0.016392   5.118 7.13e-07 ***
## L(as.numeric(threediff), 1)  0.105404   0.006207  16.981  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1319 on 203 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.6807, Adjusted R-squared:  0.6775
## F-statistic: 216.3 on 2 and 203 DF,  p-value: < 2.2e-16

AIC(weeklm4)

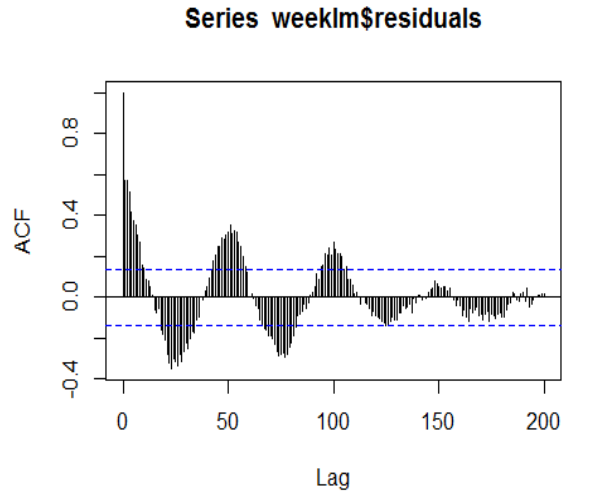
## [1] -245.0095
```

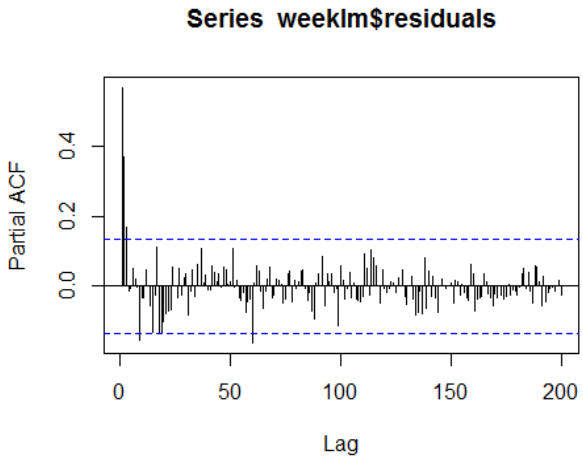
Neither of them perform better than the model with all three variables onediff, twodiff, and threediff.

11. Redo step 7 and 8

Reconsider the y with X model since its coefficients are significant.

Acf and pacf of the residuals

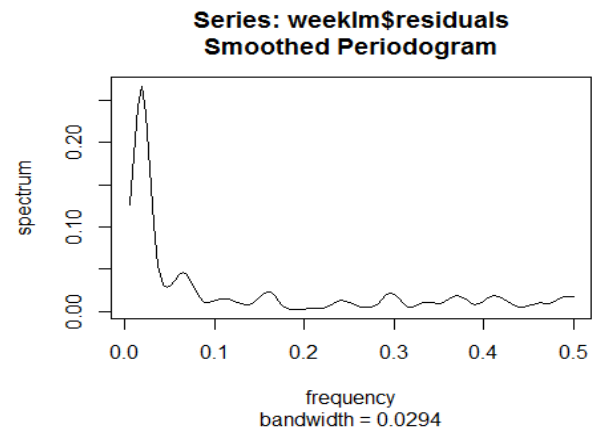




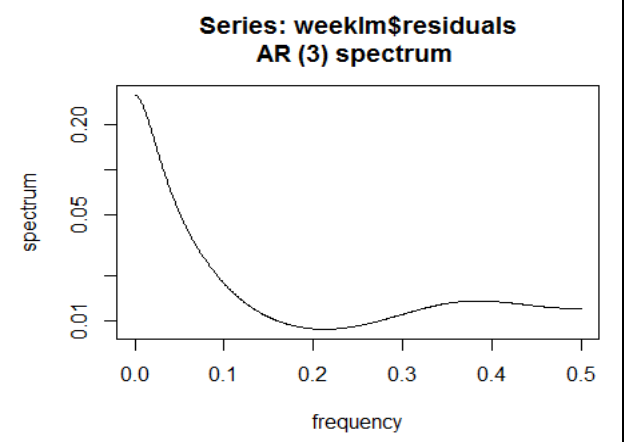

```
## Series: weeklm$residuals
## ARIMA(1,1,1)
```

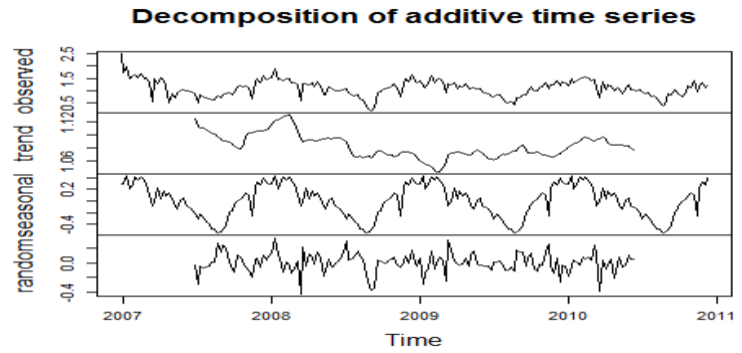
Since the residual acf plots shows seasonality, we fit a seasonal ARIMA model instead.

Multivariate spectral matrix estimate



AR fit:





From the decompose, we can see there is a clear seasonality, which agreed with the seasonality observed from residual.
Univariate and multivariate pectral estimation:

```
## Generalized least squares fit by REML
##   Model: powerm ~ onem + twom + threem
##   Data: weekly
##         AIC      BIC   logLik
##   -278.5932 -258.714 145.2966
##
## Correlation Structure: AR(1)
##  Formula: ~1
##  Parameter estimate(s):
##       Phi
## 0.8688455
##
## Coefficients:
##                 Value  Std.Error   t-value p-value
## (Intercept) 0.3613295 0.06819133  5.298761       0
## onem        0.1273742 0.01755098  7.257382       0
## twom        0.0710257 0.01516284  4.684194       0
## threem      0.0825515 0.00653347 12.635170       0
```

AR fit:

```
## Generalized least squares fit by REML
##   Model: powerm ~ onem + twom + threem
##   Data: weekly
##         AIC     BIC   logLik
##   -330.9356 -304.43 173.4678
##
## Correlation Structure: ARMA(3,0)
##  Formula: ~1
##  Parameter estimate(s):
##      Phi1      Phi2      Phi3
## 0.3773302 0.3951394 0.2254525
##
## Coefficients:
##                 Value Std.Error   t-value p-value
## (Intercept) 0.6808517 1.0755207  0.633044  0.5274
## onem        0.1062807 0.0168136  6.321127  0.0000
## twom        0.0811366 0.0149558  5.425099  0.0000
## threem      0.0900718 0.0054774 16.444223  0.0000
```

Both models have all coefficients are significant as their p-values are smaller than critical level 0.001, especially for SARIMA AR(1)'s intercept is also significant.
The better model is ARMA(3,0) since it has smaller AIC. The model is y=0.68+0.1X1+0.08X2+0.09X3. The coefficients are positive, so these covariates are positively correlated with the dependent variable.

Conclusion

The decompose of global active power shows seasonal pattern, and we estimate from its acf and pacf that there are strong auto-correlation for the variable and covariates. Furthermore, from ccf, variable sub_metering_1, sub_metering_2, and sub_metering_3 all have significant lag. And their difference's ccf to the dependent variable also shows strong correlations. Both linear model for regular and diff model are significant, as their p-values are smaller than critical level 0.001. The diff model fit a gls model with an ARMA(3,2) . The regular model's residual shows seasonality, so we fit a SARIMA with ARMA(3,0) to its gls model. The best model for this dataset is the gls model with ARMA(3,2) since it has the smallest AIC, but logically, we know the data has seasonality, it seems that SARIMA should be a better fit. For both models, the coefficients are positive, so the three covariate has positive relationship with the global active power.

Appendix: R Code

````
---
title: "STAT4181 Project Min Yang"
output:
  word_document: default
  html_notebook: default
  pdf_document: default
---

```{r}
library(zoo)
library(xts)
library(lubridate)
library(tidyverse)
library(forecast)
```

```{r}
electricity <-read.table("C:/Users/yangm/Desktop/data.txt",
sep=";",header=T,stringsAsFactors=FALSE)
```

```{r}
data<-electricity[,c(1,3,7:9)]
```

```{r}
str(data)
```

```{r}
#fit missing data with previous data
NAs <- data == "?"
is.na(data)[NAs] <- TRUE
data$Global_active_power <- na.locf(data$Global_active_power, fromLast = FALSE)
data$Sub_metering_1 <- na.locf(data$Sub_metering_1, fromLast = FALSE)
data$Sub_metering_2 <- na.locf(data$Sub_metering_2, fromLast = FALSE)
data$Sub_metering_3 <- na.locf(data$Sub_metering_3, fromLast = FALSE)
```

```{r}
data$Date <-dmy(data$Date)
data$Global_active_power <-as.numeric(data$Global_active_power)
data$Sub_metering_1 <-as.numeric(data$Sub_metering_1)
data$Sub_metering_2 <-as.numeric(data$Sub_metering_2)
````
````

data$Sub_metering_3 <-as.numeric(data$Sub_metering_3)
```

```{r}
str(data)
```
```{r}
power <- zoo(data$Global_active_power,order.by = data$Date)
str(power)
```

```{r}
powerxts <-as.xts(power)
str(powerxts)
```

```{r}
powerm<- apply.weekly(powerxts,FUN=mean)
head(powerm)
str(powerm)
length(powerm)
```

```{r}
one <- zoo(data$Sub_metering_1,order.by = data$Date)
onexts <-as.xts(one)
onem<- apply.weekly(onexts,FUN=mean)
head(onem)
```

```{r}
two<- zoo(data$Sub_metering_2,order.by = data$Date)
twoxts <-as.xts(two)
twom<- apply.weekly(twoxts,FUN=mean)
head(twom)
```

```{r}
three <- zoo(data$Sub_metering_3,order.by = data$Date)
threexts <-as.xts(three)
threem<- apply.weekly(threexts,FUN=mean)
head(threem)
```

#1.
```{r}

```
powerdiff <-diff(powerm)
onediff <-diff(onem)
twodiff <- diff(twom)
threediff <- diff(threem)
```

#2.
```{r}
library(astsa)
```
```{r}
tsdisplay(powerm)
tsdisplay(onem)
tsdisplay(twom)
tsdisplay(threem)
tsdisplay(powerdiff)
tsdisplay(onediff)
tsdisplay(twodiff)
tsdisplay(threediff)
```

#3.ccf
```{r}
tsmonth <-ts(powerm,frequency=52,start=c(2006,52))
dd<-decompose(tsmonth)
plot(dd)
```

```{r}
ccf(as.numeric(onem), as.numeric(powerm))
ccf(as.numeric(twom), as.numeric(powerm))
ccf(as.numeric(threem), as.numeric(powerm))
ccf(as.numeric(onediff), as.numeric(powerm),na.action = na.pass)
ccf(as.numeric(twodiff), as.numeric(powerm),na.action = na.pass)
ccf(as.numeric(threediff), as.numeric(powerm),na.action = na.pass)
```

#5.
```{r}
weekly<-merge(powerm,onem,twom,threem,all=c(FALSE,FALSE))
str(weekly)
```

```{r}
library(dynlm)
```

```{r}
weeklm <-
dynlm(as.numeric(powerm)~L(as.numeric(onem),0)+L(as.numeric(twom),0)+L(as.numeric(threem),0),data=weekly)
summary(weeklm)
AIC(weeklm)
```

```{r}
weeklm2 <-
dynlm(as.numeric(powerdiff)~L(as.numeric(onediff),1)+L(as.numeric(twodiff),1)+L(as.numeric(threediff),1),data=weekly)
summary(weeklm2)
AIC(weeklm2)
```

#7.
```{r}
acf(weeklm2$residuals,lag.max=200)
pacf(weeklm2$residuals,lag.max=200)
```

#8. auto.arima
```{r}
library(forecast)
auto.arima(weeklm2$residuals,max.p=5,max.q=5)
```

```{r}
week.arima <-arima(weeklm2$residuals,order=c(3,0,2))
acf(week.arima$residuals)
```

#9.
```{r}
library(nlme)
```

```{r}
week.gls<- gls(powerm~onem+twom+threem,data=weekly,correlation =
corARMA(p=3,q=2))
week.gls
```

```{r}
```

summary(week.gls)
```

##step4-7 2nd
```{r}
weeklm3 <-
dynlm(as.numeric(powerdiff)~L(as.numeric(onediff),1)+L(as.numeric(threediff),1),data=wee
kly)
summary(weeklm3)
AIC(weeklm3)
```

```{r}
acf(weeklm3$residuals,lag.max=200)
pacf(weeklm3$residuals,lag.max=200)
```
```{r}
auto.arima(weeklm3$residuals,max.p=5,max.q=5)
```
```{r}
week.arima1 <-arima(weeklm3$residuals,order=c(3,0,2))
acf(week.arima1$residuals)
```
```{r}
week.gls1<- gls(powerm~onem+threem,data=weekly,correlation = corARMA(p=3,q=2))
week.gls1
summary(week.gls1)
```

#step4-7 3rd

```{r}
weeklm4 <-
dynlm(as.numeric(powerdiff)~L(as.numeric(twodiff),1)+L(as.numeric(threediff),1),data=we
ekly)
summary(weeklm4)
AIC(weeklm4)
```
```{r}
acf(weeklm4$residuals,lag.max=200)
pacf(weeklm4$residuals,lag.max=200)
```
```{r}
auto.arima(weeklm4$residuals,max.p=5,max.q=5)
```
```{r}

```
week.arima2 <-arima(weeklm4$residuals,order=c(1,0,1))
acf(week.arima2$residuals)
```

```{r}
week.gls2<- gls(powerm~twom+threem,data=weekly,correlation = corARMA(p=1,q=1))
week.gls2
summary(week.gls2)
```

#redo step 7,8 with X model
```{r}
acf(weeklm$residuals,lag.max=200)
pacf(weeklm$residuals,lag.max=200)
```

```{r}
auto.arima(weeklm$residuals,max.p=5,max.q=5)
```

Since the residaul shows seasonality, we fit a SARIMA model instead.

```{r}
library(astsa)
spec.pgram(weeklm$residuals)
spectrum(weeklm$residuals)
mvspec(weeklm$residuals,spans=c(5,5),log="no")
spec.ar(weeklm$residuals)
```

```{r}
week.gls3<- gls(powerm~onem+twom+threem,data=weekly,correlation = corAR1(0.0294))
week.gls3
summary(week.gls3)
```

```{r}
week.gls4<- gls(powerm~onem+twom+threem,data=weekly,correlation = corARMA(p=3))
week.gls4
summary(week.gls4)
```