

# W271 Lab 3

*April 17, 2016*

## Part 1

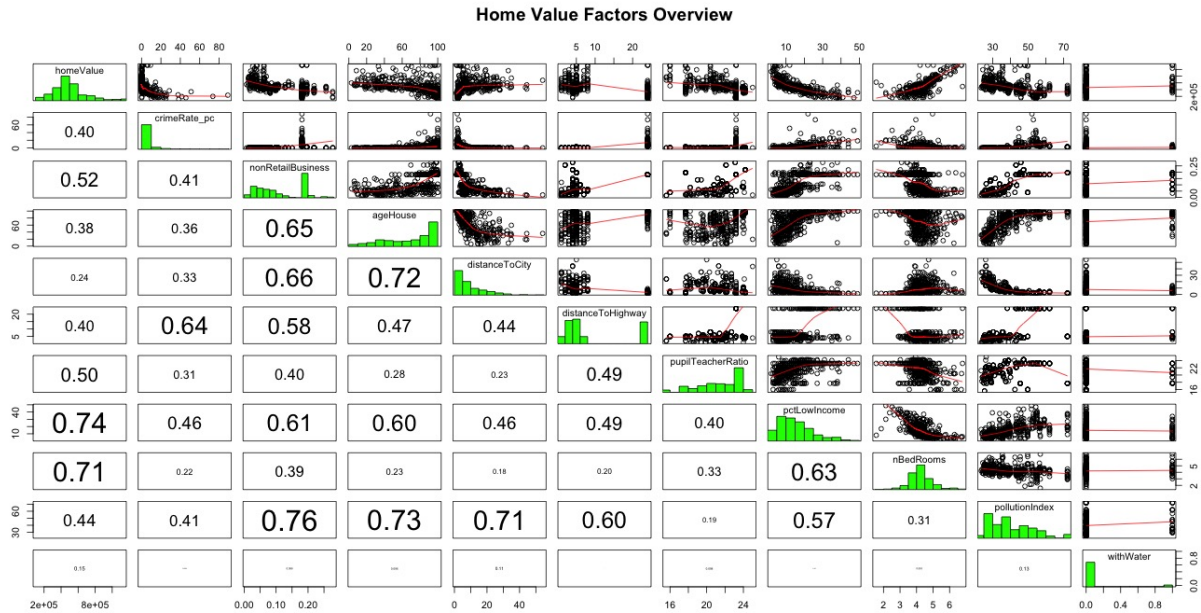
Load data and display some basic statistics:

```
## Loading required package: zoo
##
## Attaching package: 'zoo'
##
## The following objects are masked from 'package:base':
##
##      as.Date, as.Date.numeric
##
## Loading required package: survival
## Loading required package: splines
## Loading required package: timeDate
## Loading required package: timeSeries
##
## Attaching package: 'timeSeries'
##
## The following object is masked from 'package:zoo':
##
##      time<-
##
## Loading required package: fBasics
##
##
## Rmetrics Package fBasics
## Analysing Markets and calculating Basic Statistics
## Copyright (C) 2005-2014 Rmetrics Association Zurich
## Educational Software for Financial Engineering and Computational Science
## Rmetrics is free software and comes with ABSOLUTELY NO WARRANTY.
## https://www.rmetrics.org --- Mail to: info@rmetrics.org
##
## Attaching package: 'fBasics'
##
## The following object is masked from 'package:car':
##
##      densityPlot
##
##
## Please cite as:
##
## Hlavac, Marek (2015). stargazer: Well-Formatted Regression and Summary Statistics Tables.
## R package version 5.2. http://CRAN.R-project.org/package=stargazer
##
## Loading required package: forecast
## This is forecast 6.2
##
## 'data.frame':   400 obs. of  11 variables:
```

```
## $ crimeRate_pc      : num  37.6619 0.5783 0.0429 22.5971 0.0664 ...
## $ nonRetailBusiness: num   0.181 0.0397 0.1504 0.181 0.0405 ...
## $ withWater         : int    0 0 0 0 0 0 0 0 0 0 ...
## $ ageHouse          : num   78.7 67 77.3 89.5 74.4 71.3 68.2 97.3 92.2 96.2 ...
## $ distanceToCity    : num    2.71 4.12 7.82 1.95 5.54 ...
## $ distanceToHighway: int    24 5 4 24 5 5 5 5 3 5 ...
## $ pupilTeacherRatio: num   23.2 16 21.2 23.2 19.6 23.9 22.2 17.7 20.8 17.7 ...
## $ pctLowIncome      : int    18 9 13 41 8 9 12 18 5 4 ...
## $ homeValue         : int  245250 1125000 463500 166500 672750 596250 425250 483750 852750 1125000 .
## $ pollutionIndex    : num   52.9 42.5 31.4 55 36 37 34.9 72.1 33.8 45.5 ...
## $ nBedRooms         : num    4.2 6.3 4.25 3 4.86 ...
```

```
## crimeRate_pc      nonRetailBusiness  withWater      ageHouse
## Min.   : 0.00632   Min.   :0.0074   Min.   :0.0000   Min.   : 2.90
## 1st Qu.: 0.08260   1st Qu.:0.0513   1st Qu.:0.0000   1st Qu.: 45.67
## Median : 0.26600   Median :0.0969   Median :0.0000   Median : 77.95
## Mean   : 3.76256   Mean   :0.1115   Mean   :0.0675   Mean   : 68.93
## 3rd Qu.: 3.67481   3rd Qu.:0.1810   3rd Qu.:0.0000   3rd Qu.: 94.15
## Max.   :88.97620   Max.   :0.2774   Max.   :1.0000   Max.   :100.00
## distanceToCity    distanceToHighway pupilTeacherRatio pctLowIncome
## Min.   : 1.228     Min.   : 1.000     Min.   :15.60     Min.   : 2.00
## 1st Qu.: 3.240     1st Qu.: 4.000     1st Qu.:19.90     1st Qu.: 8.00
## Median : 6.115     Median : 5.000     Median :21.90     Median :14.00
## Mean   : 9.638     Mean   : 9.582     Mean   :21.39     Mean   :15.79
## 3rd Qu.:13.628     3rd Qu.:24.000     3rd Qu.:23.20     3rd Qu.:21.00
## Max.   :54.197     Max.   :24.000     Max.   :25.00     Max.   :49.00
## homeValue         pollutionIndex    nBedRooms
## Min.   : 112500    Min.   :23.50     Min.   :1.561
## 1st Qu.: 384188    1st Qu.:29.88     1st Qu.:3.883
## Median : 477000    Median :38.80     Median :4.193
## Mean   : 499584    Mean   :40.61     Mean   :4.266
## 3rd Qu.: 558000    3rd Qu.:47.58     3rd Qu.:4.582
## Max.   :1125000    Max.   :72.10     Max.   :6.780
```

We first generate the matrix plot to have an overview of all variables.



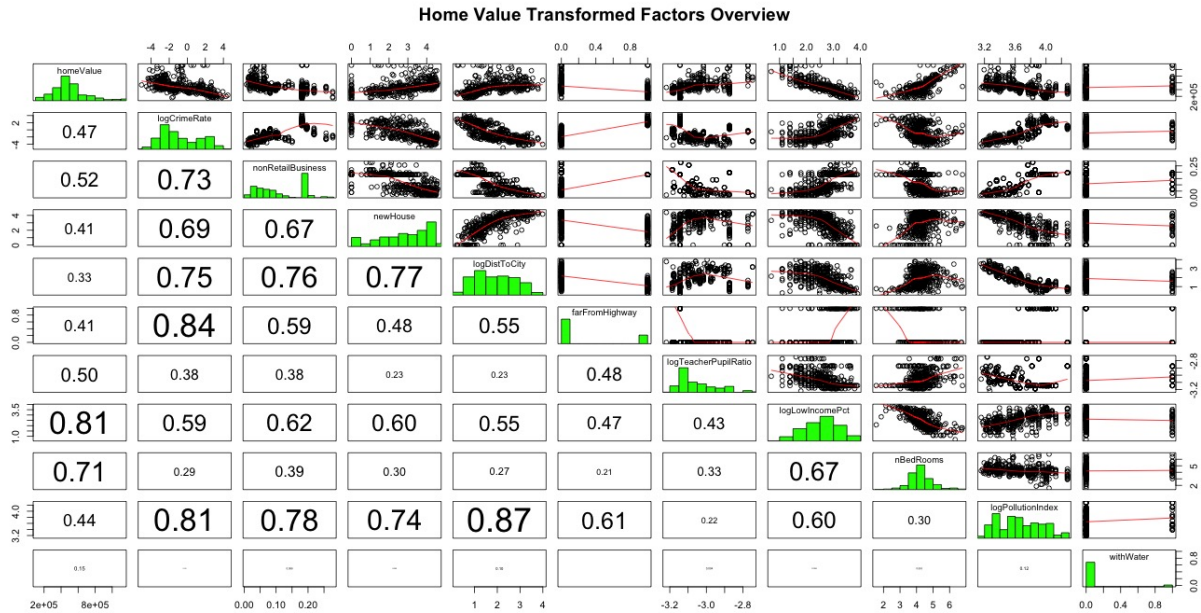
Upon first glance, two things stand out: no highly-correlated pair of variables, thus collinearity won't be a concern of our analysis, in addition, the majority of the distributions are skewed and non-normal. More specifically:

- crime rate, distance to city, low income percentage, and pollution index are negatively skewed.
- age of house, pupil teacher ratio are positively skewed
- non retail business, and distance to highway have bi-modal distribution
- home value, number of bedroom are approximately normal

we then do some transformation on the variables:

- take log of the negatively skewed variables
- convert distance to highway to a binary variable, **farFromHighway**, if it's bigger than 10
- for positive skewness, we "reverse" the variable first then take log, and the interpretation of coefficients in the model need to adjust accordingly. Specifically:
  - a. take the reciprocal of pupilTeacherRatio, it becomes teacherPupilRatio
  - b. take 100 - ageHouse, it becomes proportion of house built **after** 1950

Let's evaluate matrix plot again with the transformed variables:



Based on correlation coefficients, we propose a hypothesis of house value:

*House value is significantly affected by factors from crime rate, education quality (represented by teacher pupil ratio), low income percentage, bedroom number, and pollution index.*

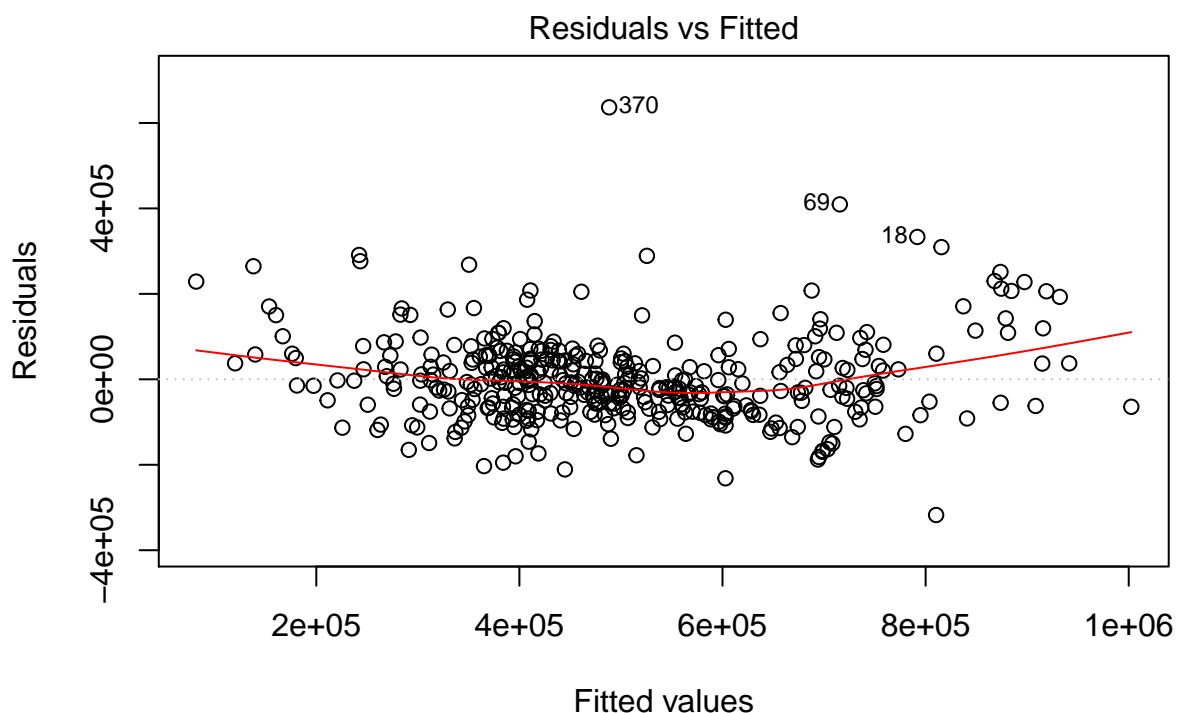
We build a linear model first with those variables:

```
##
## Call:
## lm(formula = homeValue ~ logCrimeRate + logTeacherPupilRatio +
##     logLowIncomePct + nBedRooms + logPollutionIndex, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -317589  -64311  -10894   46801  636599
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1554893.2  234001.9   6.645 1.01e-10 ***
## logCrimeRate      961.9    4320.6   0.223  0.824
## logTeacherPupilRatio 314867.8  55323.6   5.691 2.46e-08 ***
## logLowIncomePct  -173745.0  13708.0 -12.675 < 2e-16 ***
## nBedRooms        78593.6   9713.1   8.092 7.38e-15 ***
## logPollutionIndex  5615.7   32473.7   0.173  0.863
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 101100 on 394 degrees of freedom
## Multiple R-squared:  0.7374, Adjusted R-squared:  0.7341
## F-statistic: 221.3 on 5 and 394 DF, p-value: < 2.2e-16
```

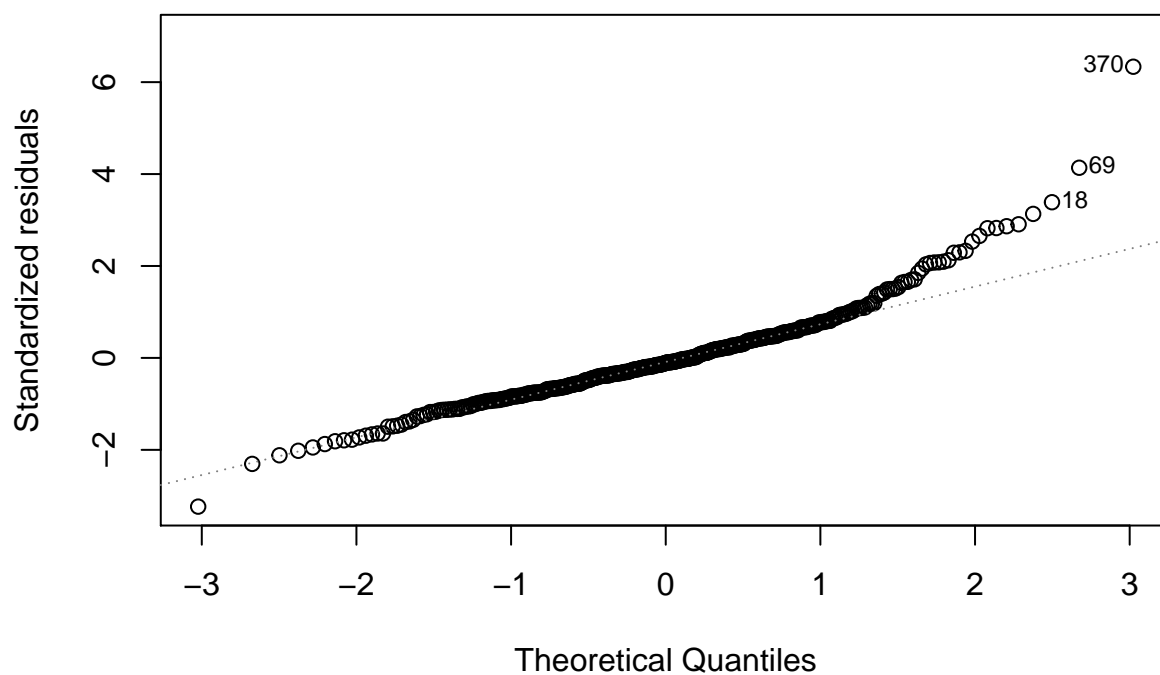
We can see that education quality, low income percentage, and number of bedrooms have significant impact on house value. On average, one more bedroom will increase the value by \$78.6k, one percent increase in the

low income percentage will reduce house value by \$173.7k, and one percent increase in teacher pupil ratio will increase house value by \$314.9k. Surprisingly here crime rate is not a significant factor.

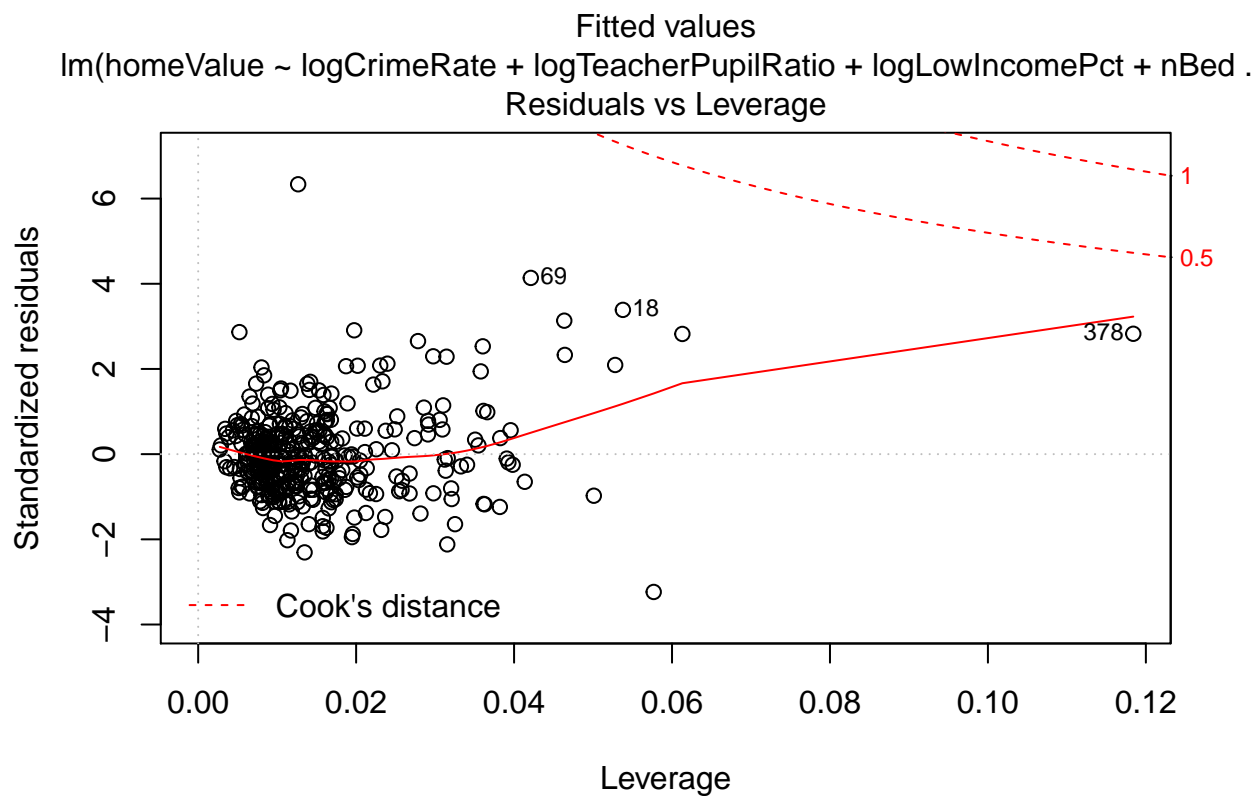
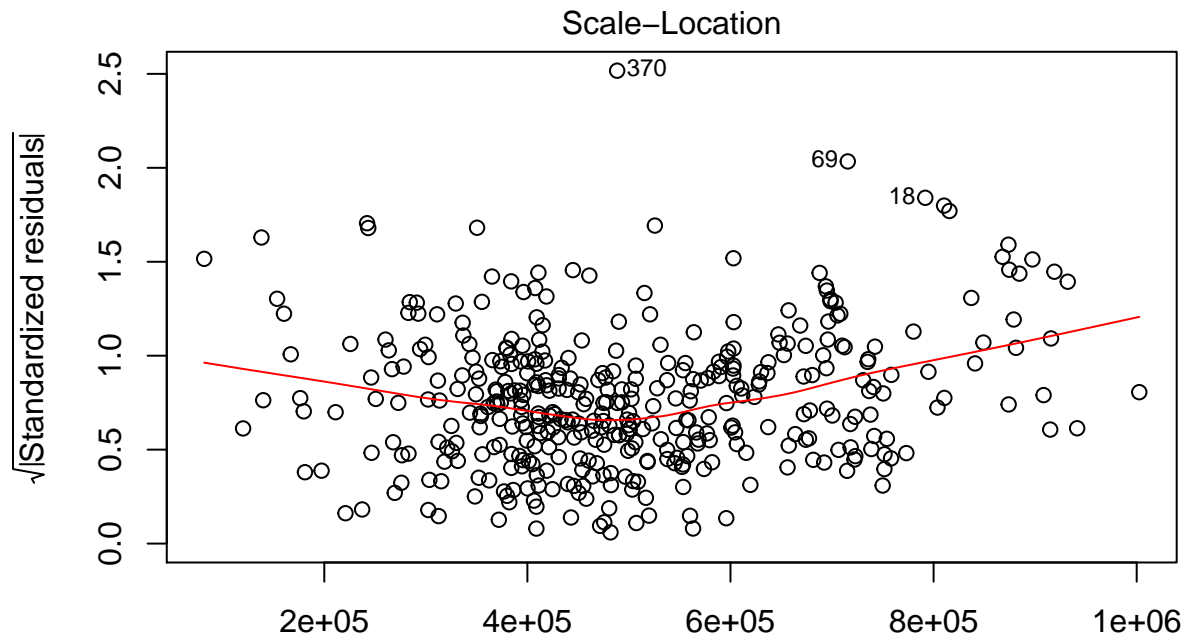
Next, we do model diagnostics:



$\text{lm}(\text{homeValue} \sim \log\text{CrimeRate} + \log\text{TeacherPupilRatio} + \log\text{LowIncomePct} + \text{nBed})$   
Normal Q-Q



$\text{lm}(\text{homeValue} \sim \log\text{CrimeRate} + \log\text{TeacherPupilRatio} + \log\text{LowIncomePct} + \text{nBed})$



$\text{lm}(\text{homeValue} \sim \log\text{CrimeRate} + \log\text{TeacherPupilRatio} + \log\text{LowIncomePct} + \text{nBed})$

From the chart we can see, the model doesn't violate homoscedasticity assumption, and there is no concern of outliers in the data. However, the normality and zero-conditional mean assumptions are questionable towards the high value house.

We now add the omitted variables to our model and compare the results:

We can see that in model 3 pollution index becomes significant. In addition, distance to city and water

Table 1: House Value Model Summary

	<i>Dependent variable:</i>		
	House Value		
	(1)	(2)	(3)
logCrimeRate	961.901 (−7,506.387, 9,430.188)	8,156.263 (−3,607.645, 19,920.170)	1,666.250 (−9,613.651, 12,946.150)
logTeacherPupilRatio	314,867.800*** (206,435.600, 423,300.000)	276,554.600*** (163,014.900, 390,094.300)	274,726.300*** (165,000.700, 384,451.900)
logLowIncomePct	−173,745.000*** (−200,612.200, −146,877.800)	−172,090.700*** (−198,877.200, −145,304.200)	−181,403.400*** (−208,434.500, −154,372.300)
nBedRooms	78,593.580*** (59,556.310, 97,630.850)	78,980.880*** (60,093.840, 97,867.920)	69,215.170*** (50,660.260, 87,770.080)
logPollutionIndex	5,615.722 (−58,031.470, 69,262.920)	−14,073.550 (−78,298.340, 50,151.240)	−182,025.200*** (−264,518.800, −99,531.650)
farFromHighway		−37,459.410 (−82,239.580, 7,320.766)	−14,017.560 (−57,147.040, 29,111.930)
withWater		53,643.820*** (13,550.510, 93,737.120)	54,161.730*** (16,438.880, 91,884.590)
nonRetailBusiness			−297,234.800** (−540,375.300, −54,094.240)
ageHouse			393.526 (−237.781, 1,024.833)
logDistToCity			−81,172.700*** (−105,548.500, −56,796.860)
Constant	1,554,893.000*** (1,096,258.000, 2,013,528.000)	1,515,577.000*** (1,056,843.000, 1,974,311.000)	2,339,513.000*** (1,827,152.000, 2,851,874.000)
Observations	400	400	400
R <sup>2</sup>	0.737	0.744	0.777
Adjusted R <sup>2</sup>	0.734	0.739	0.771
Residual Std. Error	101,125.200	100,125.200	93,770.050
F Statistic	221.330***	162.682***	135.630***

*Note:*

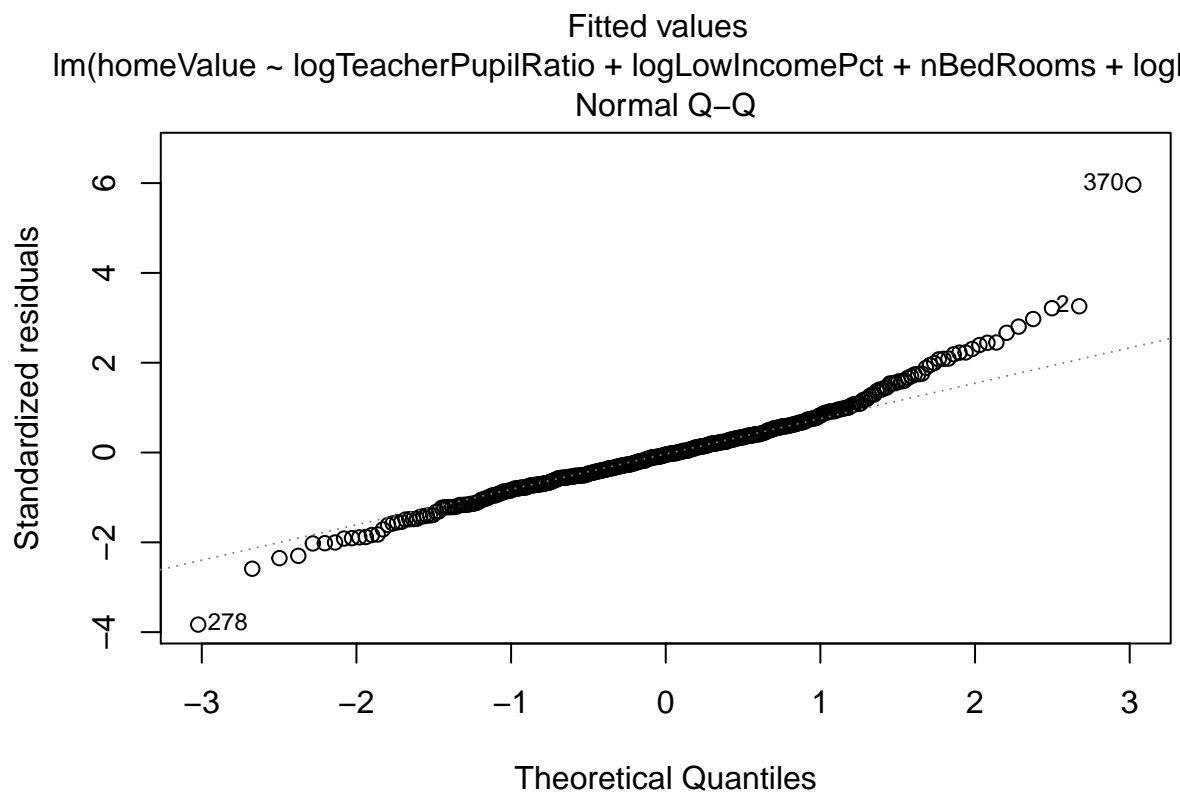
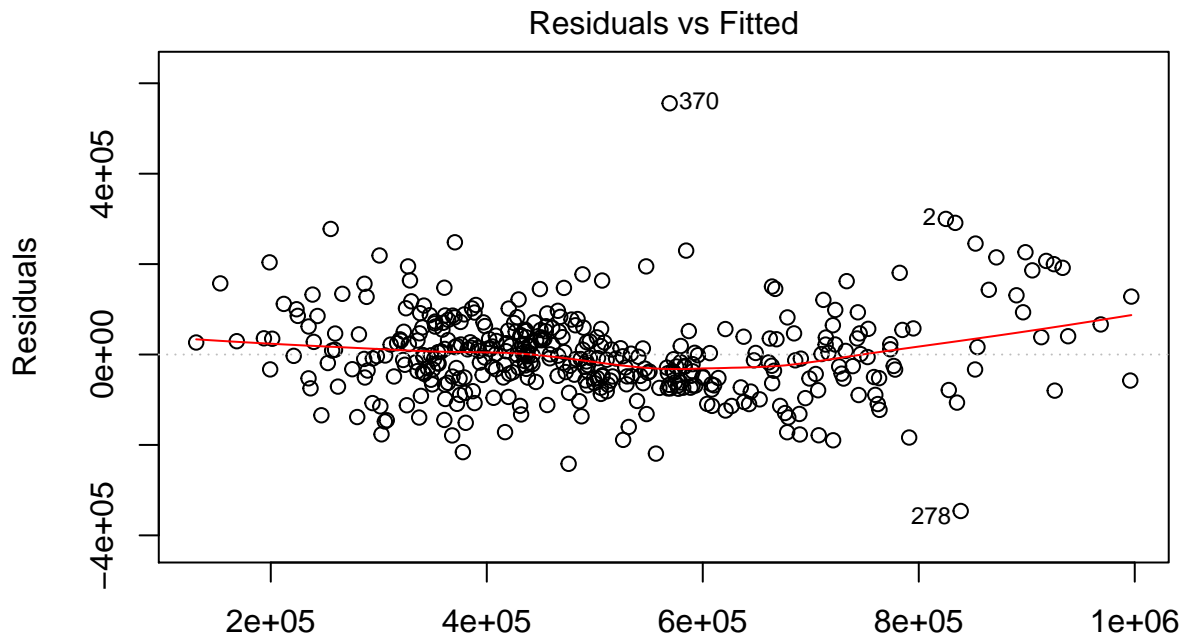
\*p&lt;0.1; \*\*p&lt;0.05; \*\*\*p&lt;0.01

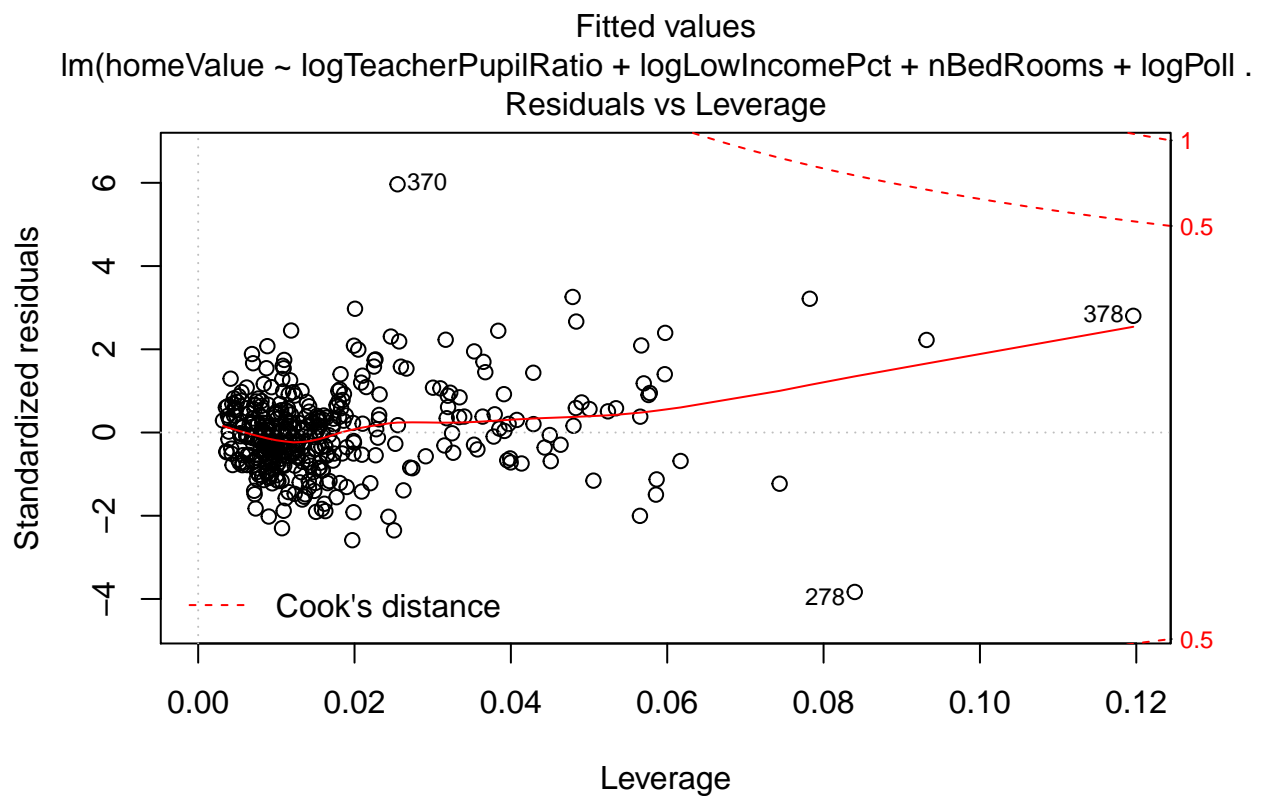
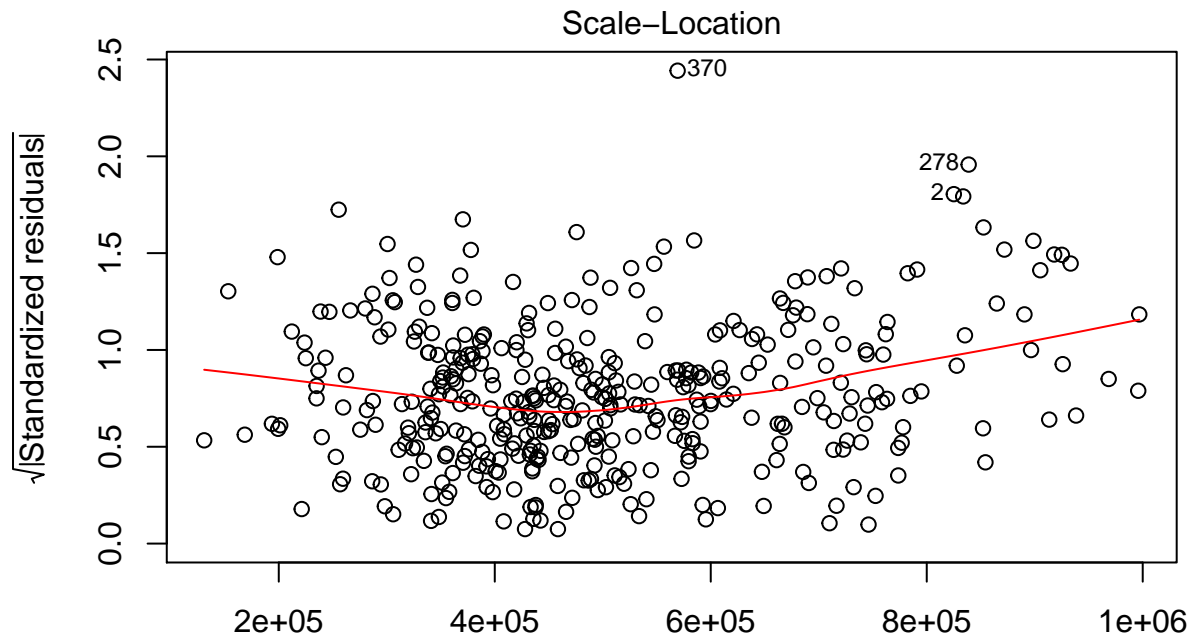
proximity are also significantly affecting house value. Finally, we build the linear model with the significant predictors identified above:

```
##
## Call:
## lm(formula = homeValue ~ logTeacherPupilRatio + logLowIncomePct +
##      nBedRooms + logPollutionIndex + withWater + logDistToCity,
##      data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -346067  -53036   -4417    46708   555679
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    2517544    216732  11.616 < 2e-16 ***
## logTeacherPupilRatio    318860    49288   6.469 2.93e-10 ***
## logLowIncomePct    -178453    12737 -14.011 < 2e-16 ***
## nBedRooms           73823     9028   8.177 4.06e-15 ***
## logPollutionIndex   -204869    35696  -5.739 1.90e-08 ***
## withWater          52940     19262   2.749 0.00626 **
## logDistToCity       -79854     11138  -7.169 3.77e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 94370 on 393 degrees of freedom
## Multiple R-squared:  0.7719, Adjusted R-squared:  0.7685
## F-statistic: 221.7 on 6 and 393 DF,  p-value: < 2.2e-16
```

we see that being further away from city will reduce house value, while having a body of water closeby will increase the value. Finally we diagnose this model







lm(homeValue ~ logTeacherPupilRatio + logLowIncomePct + nBedRooms + logPoll .

Similarly, the normality and zero-conditional mean assumption are questionable as price increases. Therefore we will use robust error to compensate:

```
##
## Call:
## lm(formula = homeValue ~ logTeacherPupilRatio + logLowIncomePct +
```

```
##      nBedRooms + logPollutionIndex + withWater + logDistToCity,
##      data = data)
##
## Residuals:
##      Min        1Q    Median        3Q        Max
## -346067  -53036   -4417    46708   555679
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)      2517544      216732  11.616 < 2e-16 ***
## logTeacherPupilRatio    318860       49288   6.469 2.93e-10 ***
## logLowIncomePct     -178453       12737 -14.011 < 2e-16 ***
## nBedRooms           73823        9028   8.177 4.06e-15 ***
## logPollutionIndex  -204869       35696  -5.739 1.90e-08 ***
## withWater           52940       19262   2.749 0.00626 **
## logDistToCity       -79854       11138  -7.169 3.77e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 94370 on 393 degrees of freedom
## Multiple R-squared:  0.7719, Adjusted R-squared:  0.7685
## F-statistic: 221.7 on 6 and 393 DF,  p-value: < 2.2e-16

## [1] "Robust Standard Errors"

##              (Intercept) logTeacherPupilRatio      logLowIncomePct
##              231450.28      55184.01      18941.97
##              nBedRooms      logPollutionIndex      withWater
##              15365.45      36594.87      23188.05
##              logDistToCity
##              14596.04
```

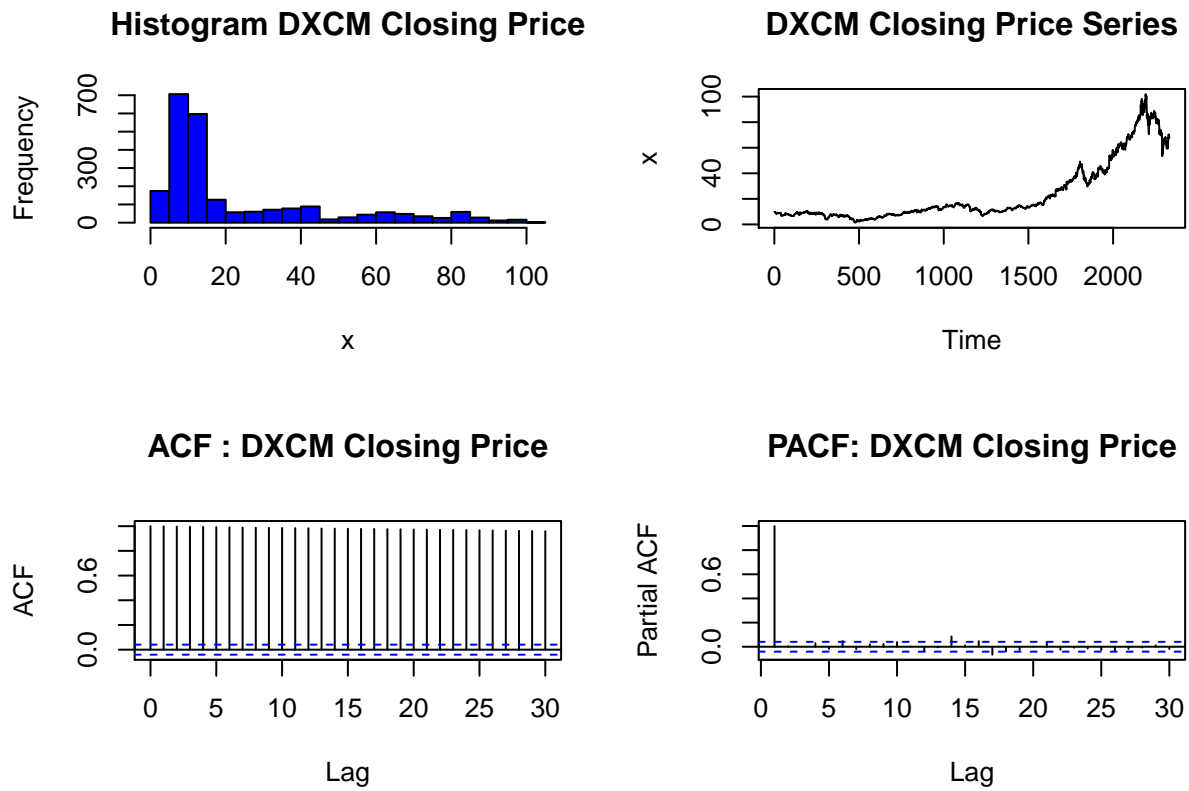
## Part 2

Load data, package, and show descriptive statistics:

```
## 'data.frame':   2332 obs. of  2 variables:
## $ X          : int  1 2 3 4 5 6 7 8 9 10 ...
## $ DXCM.Close: num  9.88 9.79 9.68 9.64 9.42 9.47 9.16 8.99 8.6 8.81 ...

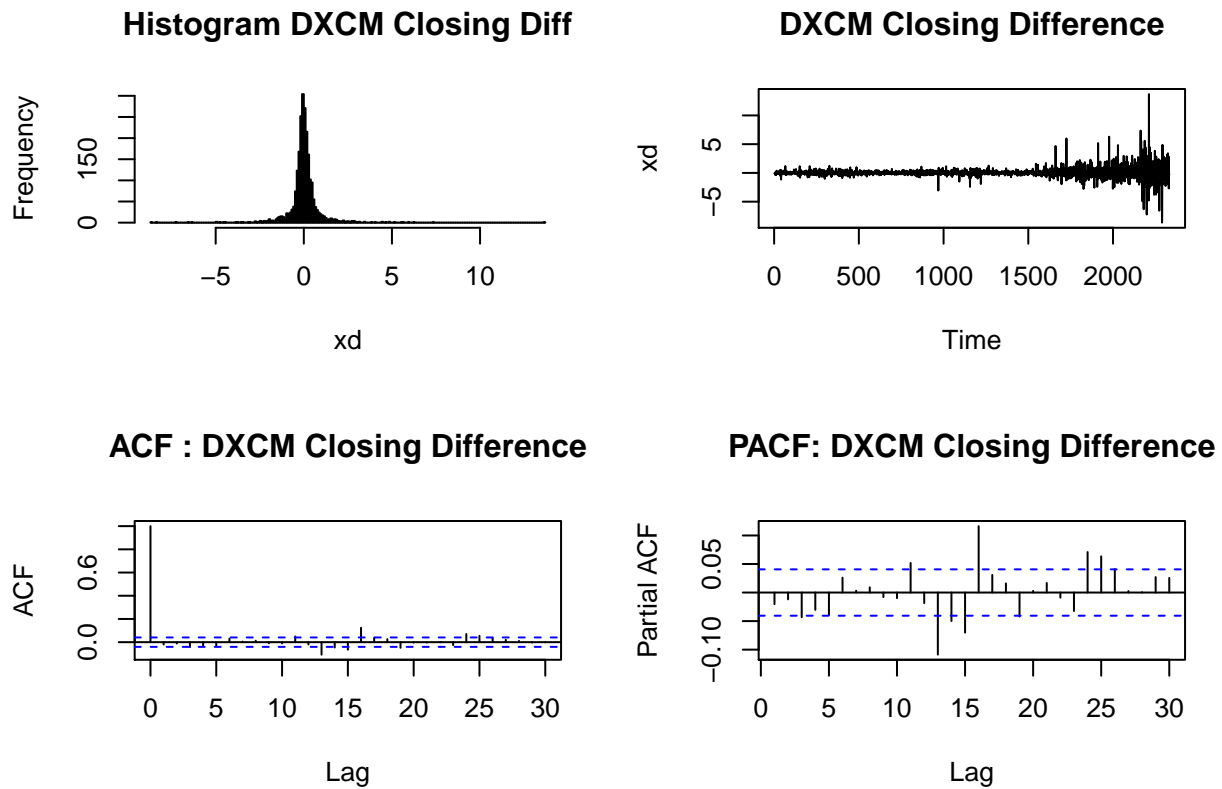
##      Min. 1st Qu.  Median      Mean 3rd Qu.      Max.
##      1.390   8.188  12.360  23.210  32.560 101.900
```

Let's evaluate the time series plot, histogram, ACF and PACF of the data:



```
##
## Box-Ljung test
##
## data: x
## X-squared = 2327.388, df = 1, p-value < 2.2e-16
```

The Box test indicates that our original series  $x$  is **not** a stationary series, and we can observe a upward trend, thus simple ARMA model won't be adequate and we further evaluate the difference of the  $x$ ,  $x_d$ :



```
##
## Box-Ljung test
##
## data:  xd
## X-squared = 0.9615, df = 1, p-value = 0.3268
```

Box test now indicates  $x_d$  is stationary, however we can see that the variance of  $x_d$  is time-varying, as such, we **cannot** apply ARIMA alone. To address that, we use ARIMA to fit the original series  $x$ , then apply GARCH model on the ARIMA residue to estimate conditional variance and obtain the final prediction by integrating GARCH results into the ARIMA prediction. To obtain the best ARIMA model, we run the following procedure to identify the optimal order  $(p, d, q)$ :

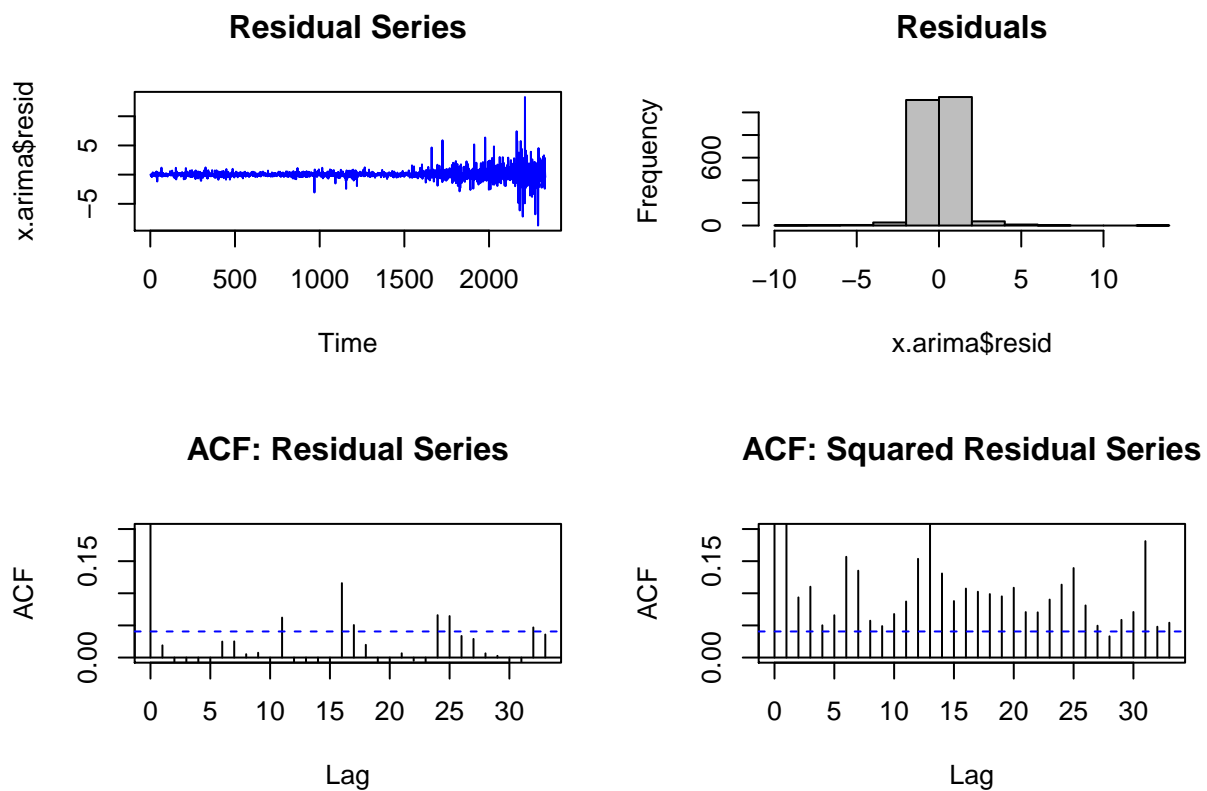
```
# procedure to get best ARIMA order
get.best.arma <- function(x.ts, maxord = c(1,1,1)) # don't change any of this code
{
  best.aic <- 1e8
  n <- length(x.ts)
  for (p in 0:maxord[1]) for(d in 0:maxord[2]) for(q in 0:maxord[3])
  {
    fit <- arima(x.ts, order = c(p,d,q))
    fit.aic <- -2 * fit$loglik + (log(n) + 1) * length(fit$coef)
    if (fit.aic < best.aic)
    {
      best.aic <- fit.aic
      best.fit <- fit
      best.model <- c(p,d,q)
    }
  }
}
```

```
list(best.aic, best.fit, best.model)
}
# model selection
#x.best = get.best.arma(x, maxord = c(2,2,2))[[3]]
```

the best order of ARIMA is (0,1,0):

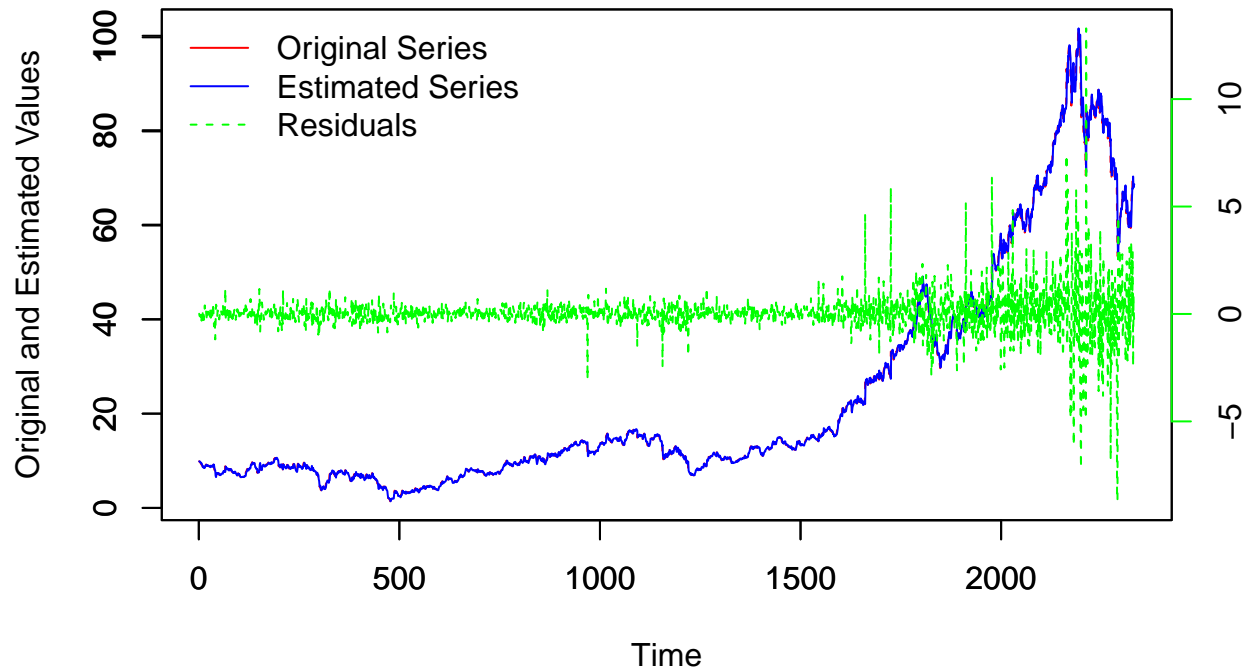
```
##      Min.   1st Qu.   Median     Mean   3rd Qu.    Max.
## -8.716000 -0.211900  0.003836  0.028970  0.256700 13.290000
```

```
##
## Box-Ljung test
##
## data:  x.arma$resid
## X-squared = 0.8374, df = 1, p-value = 0.3601
```



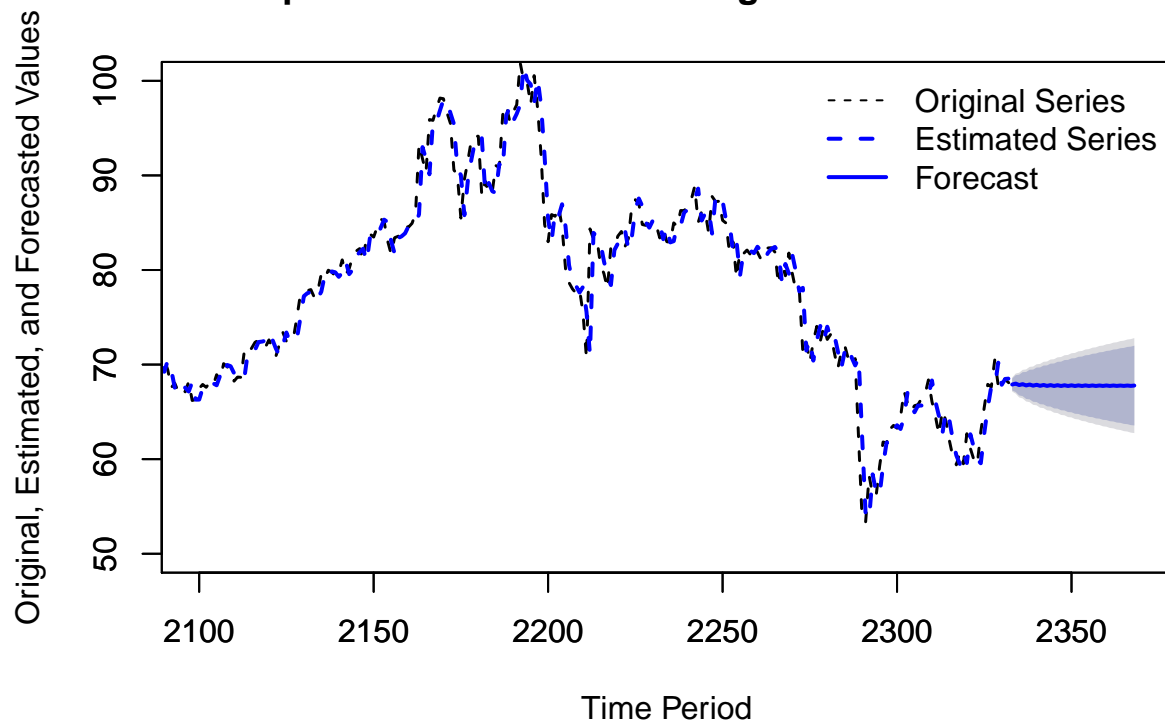
we can see although the ACF of residual indicates insignificant autocorrelation, ACF of squared residual showed otherwise. Let further check the in-sample fit of the model:

## Original vs ARIMA(0,1,0) Estimated Series with Residuals



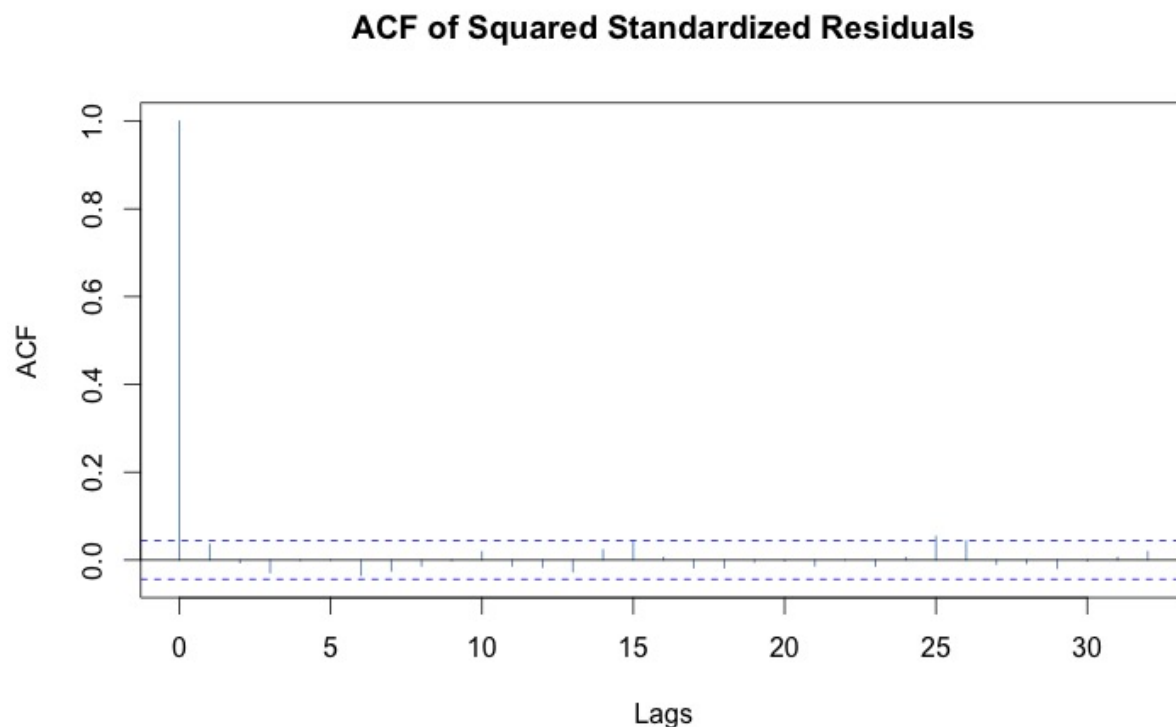
the in-sample fit closely follow the original series. However the residuals demonstrate varying variance, we apply GARCH model on the residuals to correct the prediction uncertainty:

## 36-Step Ahead Forecast and Original & Estimated Series



The prediction of  $ARIMA(0,1,0)$  gives a flat prediction, due to the fact that the model doesn't have AR and MA coefficient. Thus the prediction is equivalent to a random walk without noise, which has become a constant of the last observation. However, after correct the prediction variance, we find the confidence

interval is reduced compared with that given by ARIMA model. Finally, the ACF of squared residuals of GARCH model shows the variance of residual is no long varying by time.



### Part 3

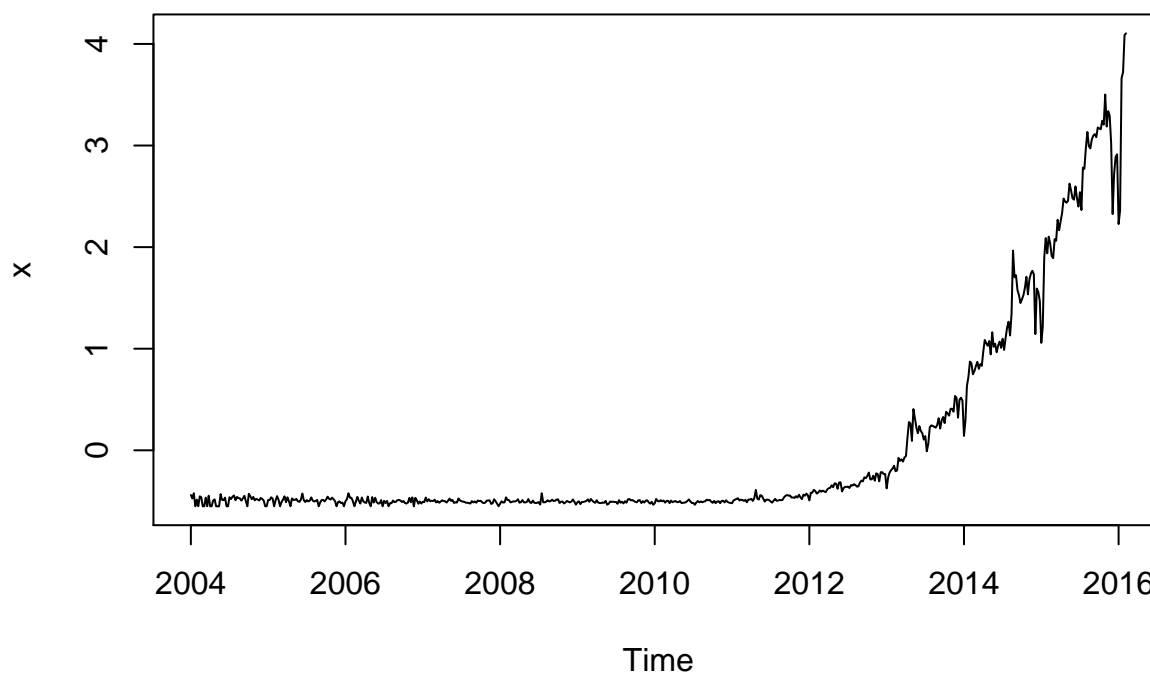
Load data and show descriptive statistics:

```
## 'data.frame': 630 obs. of 2 variables:
## $ Date : Factor w/ 630 levels "1/1/06","1/1/12",...: 47 5 18 33 215 260 226 239 251 311 ...
## $ data.science: num -0.44 -0.474 -0.423 -0.551 -0.486 -0.551 -0.453 -0.462 -0.551 -0.551 ...

## Min. 1st Qu. Median Mean 3rd Qu. Max.
## -0.551000 -0.506000 -0.485000 0.000038 -0.200000 4.104000
```



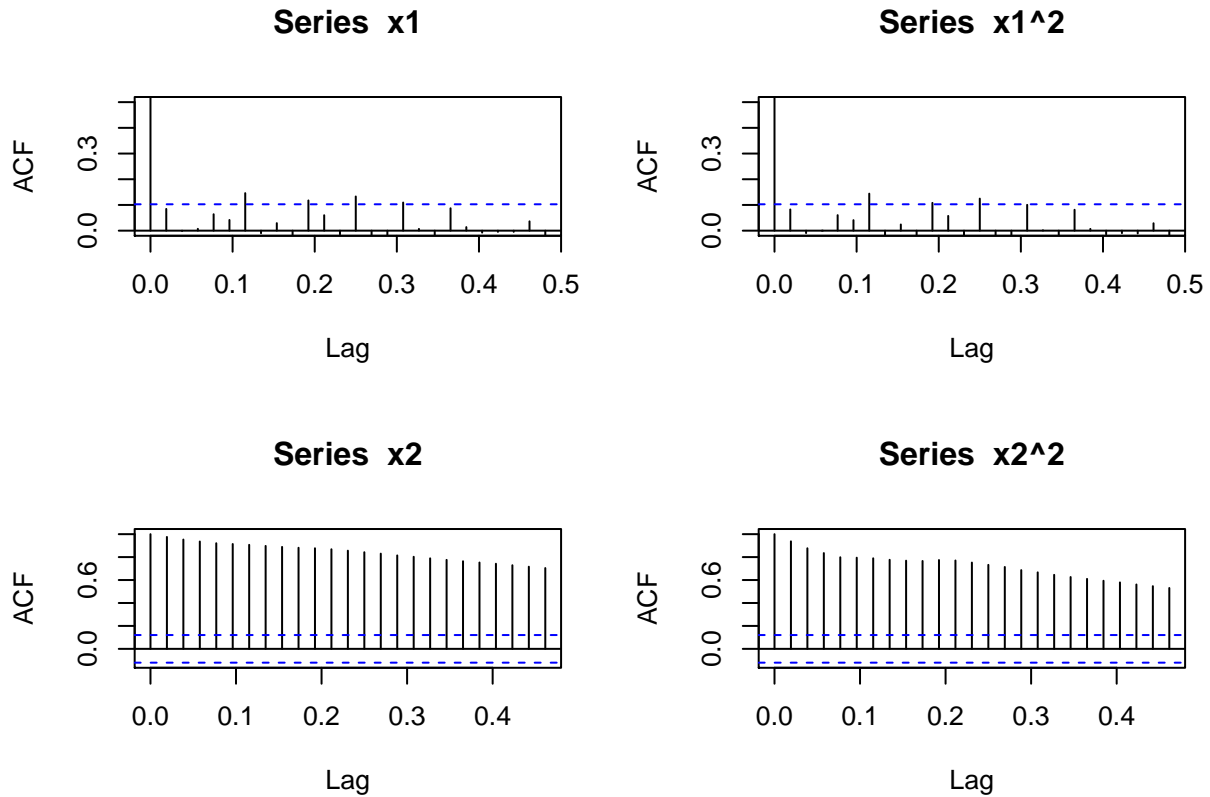
## Web Search Rate of Global Warming



upon close look, it appears the majority of the dynamics of the series appears after 2011, which prior to that it's mostly flat without too much changes. Let's then cut the series into two: before and after 2011

```
cutoff = 365  
x1 <- ts(data$data.science[1:cutoff], start=c(2004,1,4), frequency = 52)  
x2 <- ts(data$data.science[(cutoff+1):630], start=c(2011,1,2), frequency = 52)
```

we then evaluate ACF of both series, and conduct Box test:

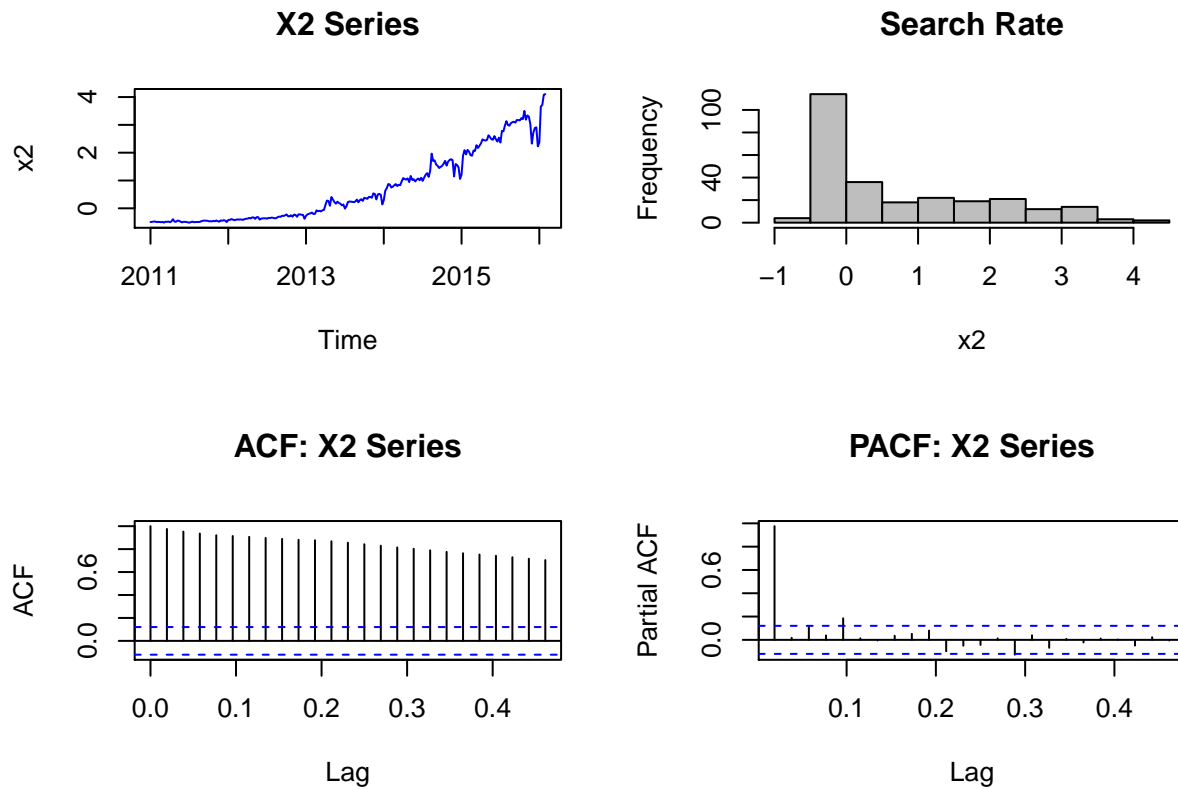


```
##
## Box-Ljung test
##
## data:  x1
## X-squared = 2.6322, df = 1, p-value = 0.1047

##
## Box-Ljung test
##
## data:  x2
## X-squared = 255.178, df = 1, p-value < 2.2e-16
```

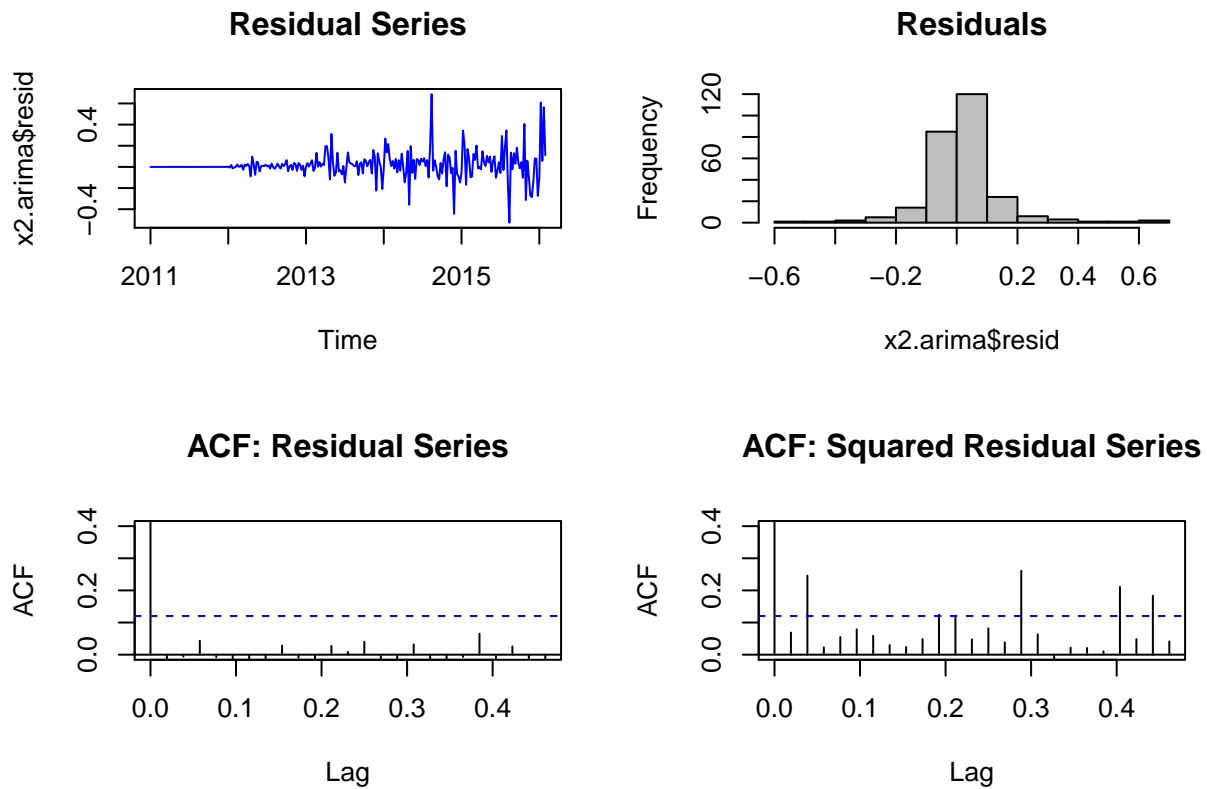
For series  $x_1$ , ACF of both original and squared value don't show significant autocorrelation, while  $x_2$  shows strong autocorrelation for both. In addition, the Box test indicates that we can't reject the null hypothesis that  $x_1$  is white-noise, while  $x_2$  is significantly autocorrelated.

Thus we will focus on modeling  $x_2$  to predict the search rate of Global Warming.



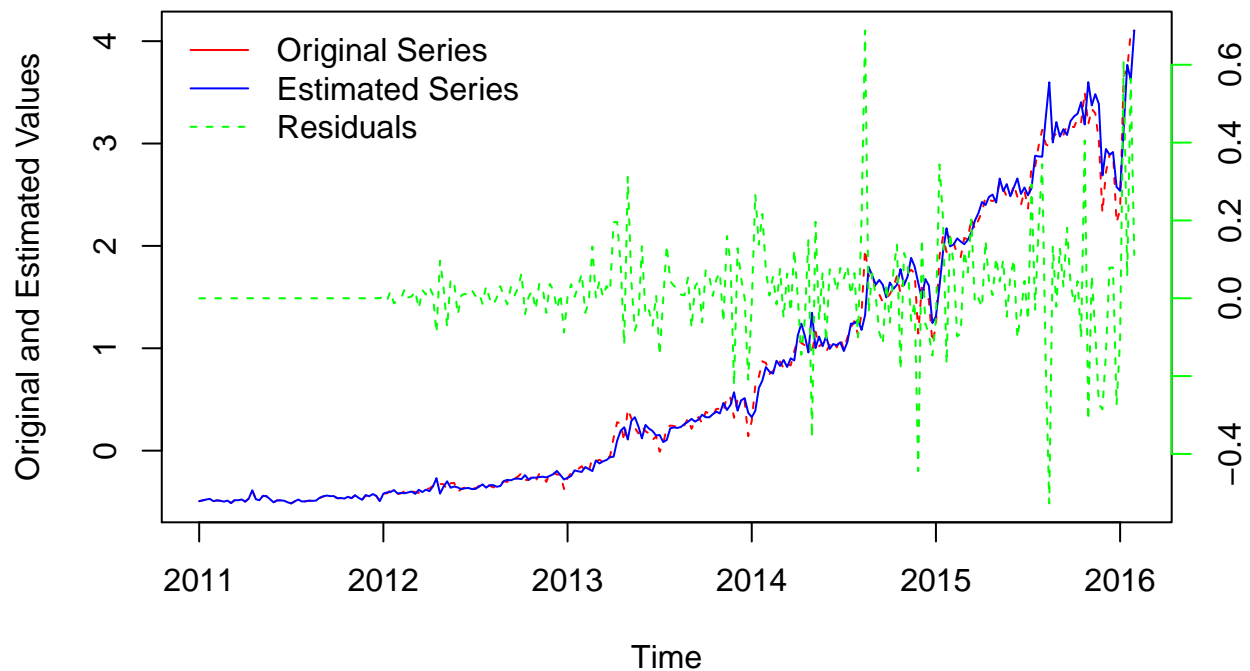
From ACF we see that it's a non-stationary series, which show strong persistent upward trend. We then use ARIMA model to fit it.

```
## Series: x2
## ARIMA(1,1,1)(0,1,1)[52]
##
## Coefficients:
##      ar1      ma1      sma1
##      0.4550 -0.7844 -0.1681
## s.e.  0.1423  0.1072  0.0723
##
## sigma^2 estimated as 0.01957:  log likelihood=115.26
## AIC=-222.52  AICc=-222.32  BIC=-209.09
##
##
## Box-Ljung test
##
## data:  x2.arima$resid
## X-squared = 0.4348, df = 1, p-value = 0.5096
```



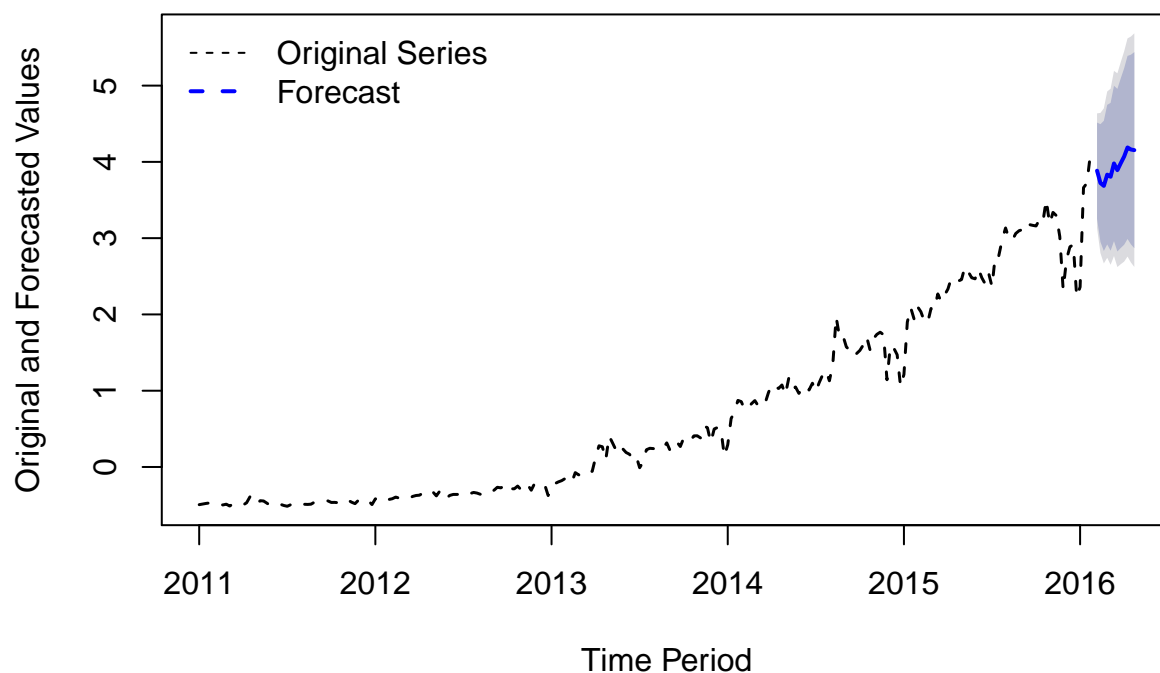
The best model is ARIMA(1,1,1)(0,1,1)[52], and the ACF of squared residual also indicate its variation is time-varying.

### Original vs ARIMA(1,1,1)(0,1,1)[52] Estimated Series with Residuals



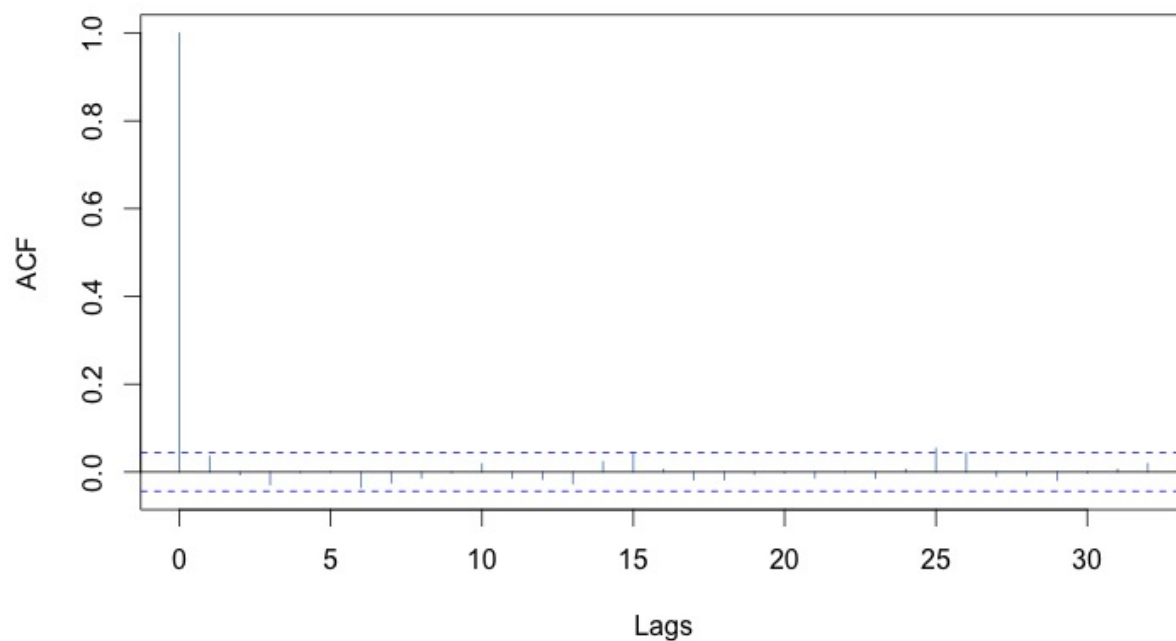
Similar with part 2, we fit a GARCH model on the residuals, and integrate the conditional variance from GARCH into the prediction of ARIMA model:

## 12-Step Ahead Forecast and Original & Estimated Series



Finally, the ACF of squared residuals of GARCH model shows the variance of residual is no long varying by time.

## ACF of Squared Standardized Residuals



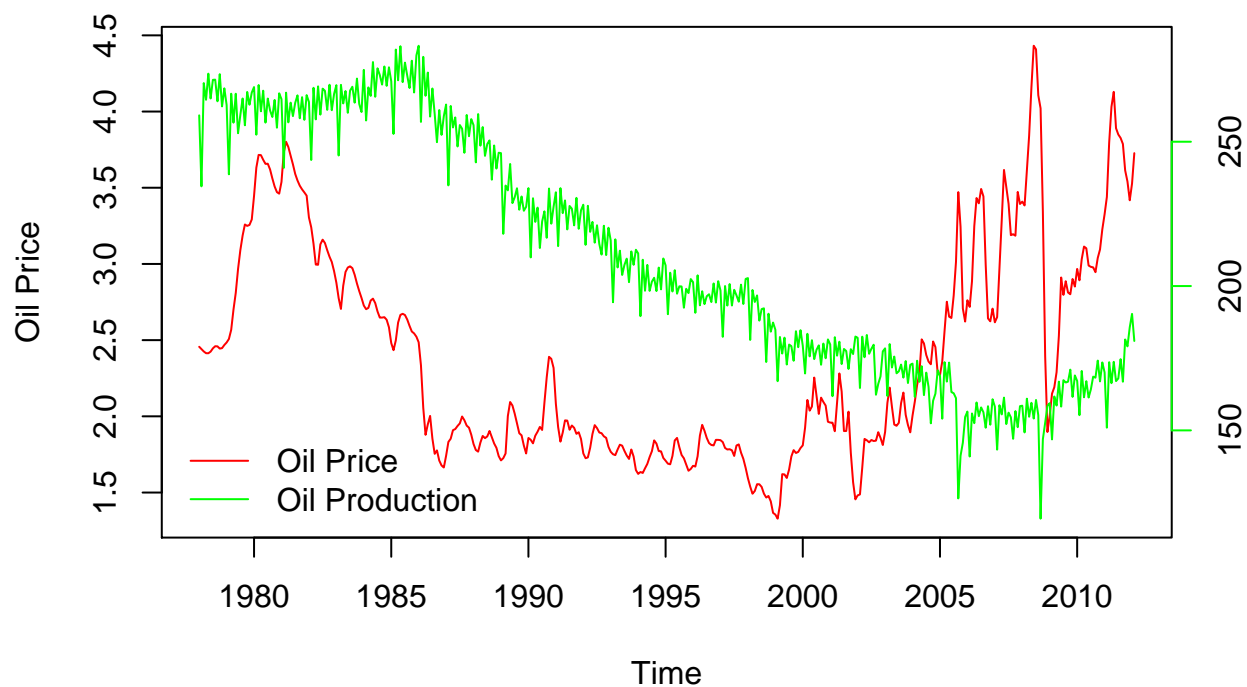
## Part 4

Load data and show descriptive statistics:

```
## 'data.frame': 410 obs. of 3 variables:
## $ Date : chr "1978-01-01" "1978-02-01" "1978-03-01" "1978-04-01" ...
## $ Production: num 259 235 270 265 274 ...
## $ Price : num 2.46 2.44 2.43 2.41 2.41 ...
```

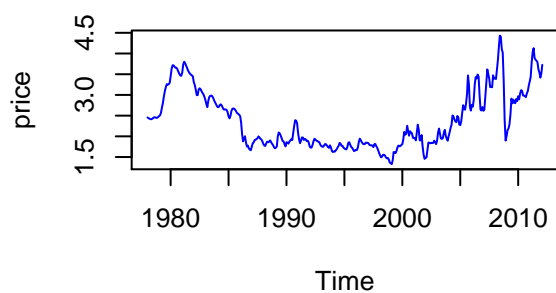
```
##      Date      Production      Price
## Length:410      Min. :119.4      Min. :1.329
## Class :character 1st Qu.:173.0      1st Qu.:1.823
## Mode :character  Median :201.4      Median :2.096
##                      Mean :210.0      Mean :2.391
##                      3rd Qu.:255.8      3rd Qu.:2.909
##                      Max. :283.2      Max. :4.432
```

### U.S. Oil Production & Price

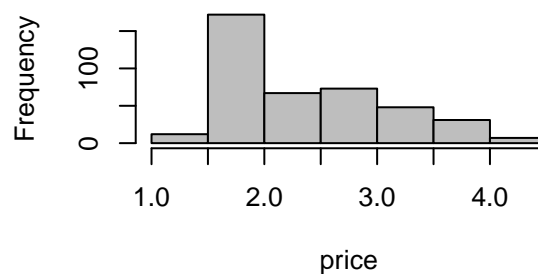


## Task 2 - Forecast the inflation-adjusted gas prices from 2012 to 2016

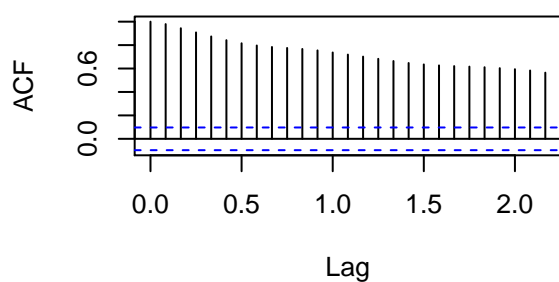
**U.S. Oil Price Series**



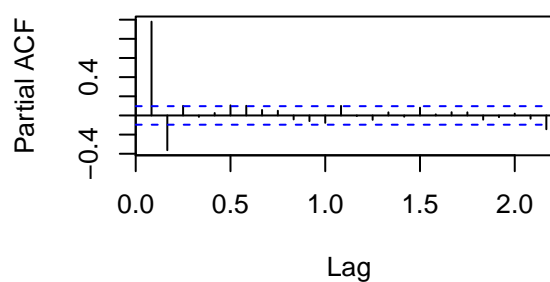
**U.S. Oil Price**



**ACF: U.S. Oil Price Series**

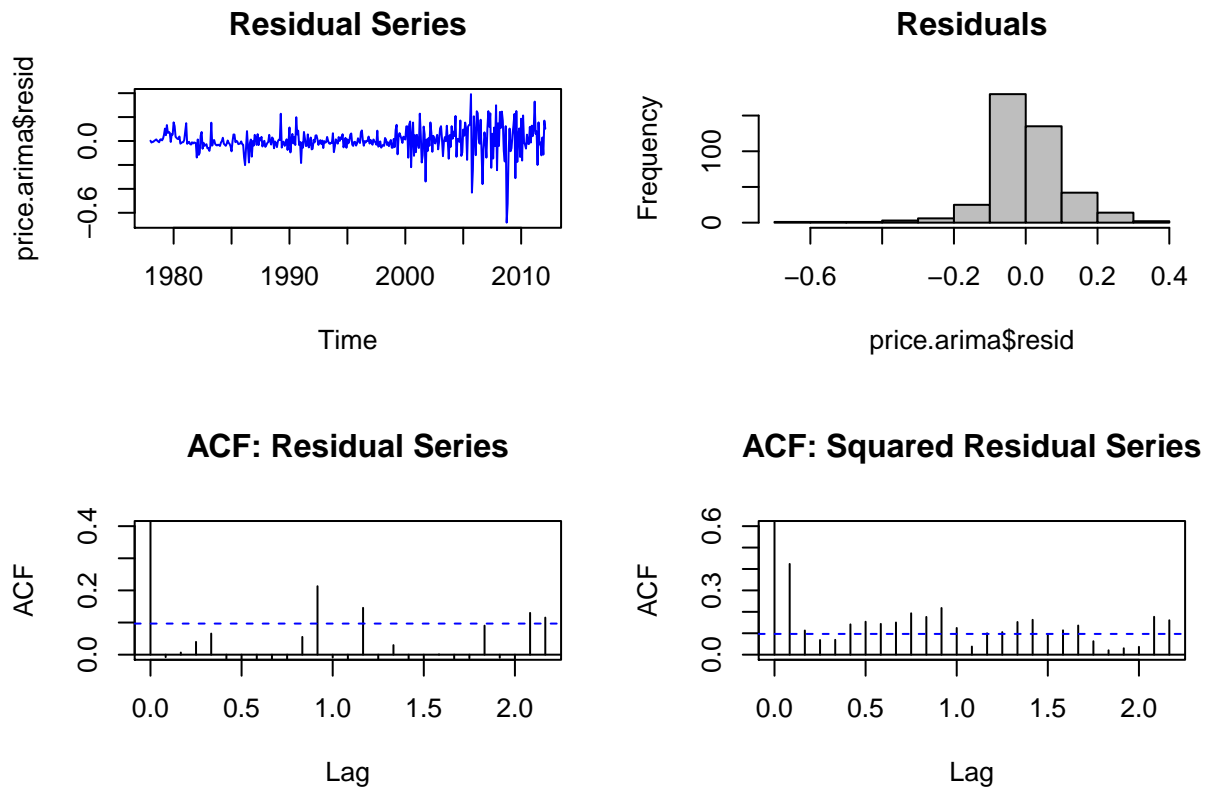


**PACF: U.S. Oil Price Series**



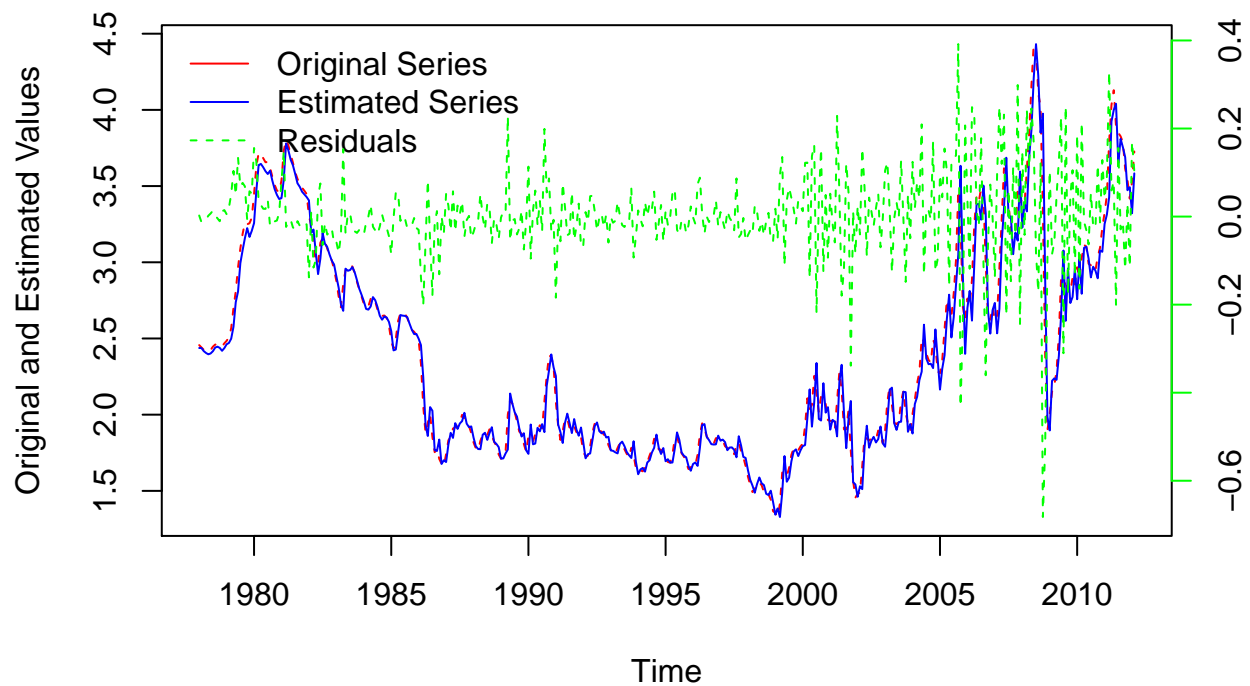
From ACF we see that it's a non-stationary series, which show strong persistent upward trend. We then use ARIMA model to fit it.

```
##
## Box-Ljung test
##
## data: price.arima$resid
## X-squared = 0.0245, df = 1, p-value = 0.8755
```



The best model is ARIMA(1,1,3), and the ACF of squared residual also indicate its variation is time-varying.

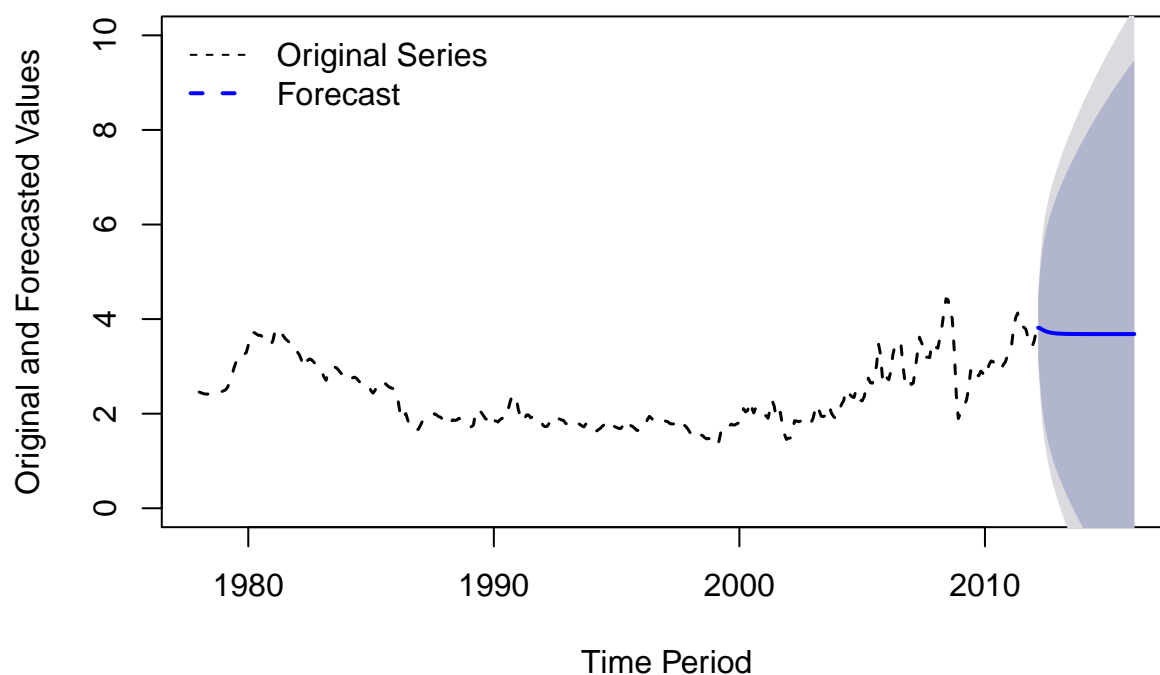
### Original vs ARIMA(1,1,3) Estimated Series with Residuals



Similar with part 2, we fit a GARCH model on the residuals, and integrate the conditional variance from GARCH into the prediction of ARIMA model:



### 4-year Ahead Forecast and Original & Estimated Series



Finally, the ACF of squared residuals of GARCH model shows the variance of residual is no long varying by time.

### ACF of Squared Standardized Residuals

