Time Series and Forecasting Methods
-
Assignment: November 2024

Overview of the variables in the dataset

The dataset you need to analyze is available in the eclass file "data-assignment.txt" or the Excel file "data-assignment.xls". The dependent variables, for which you will construct the requested models, are the monthly returns of nine alternative investment forms (Y1 = HFRI, Y2 = EH, Y3 = M, Y4 = RVA, Y5 = ED, Y6 = CA, Y7 = DS, Y8 = EMN, Y9 = MA) for the period 4/1990 - 12/2005.
The independent variables you will use in the models are the monthly values/returns for the variables x1 = RUS-Rf, x2 = RUS(-1)-Rf(-1) lagged Russel index, x3 = MXUS-Rf, x4 = MEM-Rf, x5 = SMB, x6 = HML, x7 = MOM, x8 = SBGC-Rf, x9 = SBWG-Rf, x10 = LHY-Rf, x11 = DEFSPR, x12 = FRBI-Rf, x13 = GSCI-Rf, x14 = VIX, x15 = Rf, also for the period 4/1990 - 12/2005.
You are asked to perform an analysis of the dependent variables based on the data for the period 4/1990 - 12/2005.

Overview of the variables
1. HFRI (Hedge Fund Research Index)
HFRI indexes track the performance of different hedge fund strategies. These indexes are widely used benchmarks for the hedge fund industry, with sub-categories for strategies like equity hedge, event-driven, relative value arbitrage, and macro.
2. EH (Equity Hedge)
A strategy within the hedge fund industry that focuses on long and short positions in equities, aiming to profit from rising (long) and falling (short) prices in the stock market.
3. M (Macro)
Macro strategies involve investing based on global economic trends, such as interest rates, currency rates, or commodity prices. Macro funds often use a top-down approach, assessing broad economic factors to make investment decisions across asset classes.
4. RVA (Relative Value Arbitrage)
A strategy that seeks to capitalize on price discrepancies between related financial instruments, betting on the convergence of their values. This can involve bond arbitrage, volatility arbitrage, or other market inefficiencies.
5. ED (Event Driven)
Event-driven strategies focus on corporate events like mergers, acquisitions, restructurings, or bankruptcies. These funds aim to profit from price changes resulting from these specific events.
6. CA (Convertible Arbitrage)
A form of relative value arbitrage that focuses on convertible securities, such as bonds that can convert to stocks. Managers typically go long on the convertible bond and short the underlying stock to exploit price differentials.

7. DS (Distressed Securities)
Investing in the debt of companies in financial distress or near bankruptcy, often with the goal of restructuring the company or profiting from price rebounds. It’s a higher-risk, potentially high-reward strategy within the event-driven category.
8. EMN (Equity Market Neutral)
A strategy that aims to eliminate market risk by maintaining balanced long and short positions in equity markets. The goal is to generate returns regardless of the overall market direction.
9. MA (Merger Arbitrage)
A type of event-driven strategy that seeks to profit from mergers and acquisitions. For example, an investor might go long on a target company's stock (likely to rise) and short the acquiring company's stock (likely to fall or stay flat).
10. RUS-Rf
Represents the excess return of the Russell Index (RUS) over the risk-free rate (Rf), where Rf is typically a benchmark yield, such as that of a U.S. Treasury bond. This measures the performance of the Russell Index adjusted for risk.
11. RUS(-1)-Rf(-1)
The lagged excess return of the Russell Index, where both the index return and the risk-free rate are taken from the previous period. It’s often used in predictive models.
12. MXUS-Rf
The excess return of the MSCI US Index (MXUS) over the risk-free rate. This index represents the U.S. equity market performance, adjusted by the risk-free rate.
13. MEM-Rf
The excess return of the MSCI Emerging Markets Index (MEM) over the risk-free rate, offering insight into emerging market performance beyond what could be earned with a risk-free investment.
14. SMB (Small Minus Big)
A factor in the Fama-French model representing the size premium, or the return difference between small-cap and large-cap stocks. SMB captures the tendency for small companies to outperform larger companies over time.
15. HML (High Minus Low)
Another Fama-French factor, representing the value premium by comparing stocks with high book-to-market values (value stocks) to those with low book-to-market values (growth stocks).
16. MOM (Momentum)
A factor that captures the tendency of assets with high recent returns to continue performing well in the short term, and vice versa. Momentum is often used as a strategy that bets on past winners to continue their performance.
17. SBGC-Rf
The excess return of the S&P/BGCantor U.S. Treasury Bond Index over the risk-free rate. This measures the performance of the Treasury bond index, adjusted for a baseline return from risk-free assets.
18. SBWG-Rf
The excess return of the S&P/BGCantor U.S. Government Bond Index over the risk-free rate, focusing on U.S. government bonds.
19. LHY-Rf
The excess return of the U.S. High Yield Bond Index over the risk-free rate, representing returns of high-yield (often called "junk") bonds beyond a risk-free investment.

20. DEFSPR (Default Spread)
The difference between yields on high-yield bonds and safer, lower-yielding investment-grade bonds. A wider default spread indicates higher credit risk in the market, while a narrow spread suggests lower perceived credit risk.
21. FRBI-Rf
The excess return of a foreign bond index over the risk-free rate, providing insight into the returns of foreign bonds adjusted by the risk-free rate.
22. GSCI-Rf
The excess return of the Goldman Sachs Commodity Index (GSCI) over the risk-free rate, which measures the performance of a broad basket of commodities beyond what a risk-free asset would yield.
23. VIX (Volatility Index)
Also known as the "Fear Index," the VIX measures market expectations for volatility over the next 30 days. A higher VIX implies more anticipated market volatility.
24. Rf (Risk-Free Rate)
The theoretical rate of return of an investment with zero risk, often proxied by the return on short-term U.S. Treasury bills. The risk-free rate is used as a baseline to assess risk-adjusted returns.

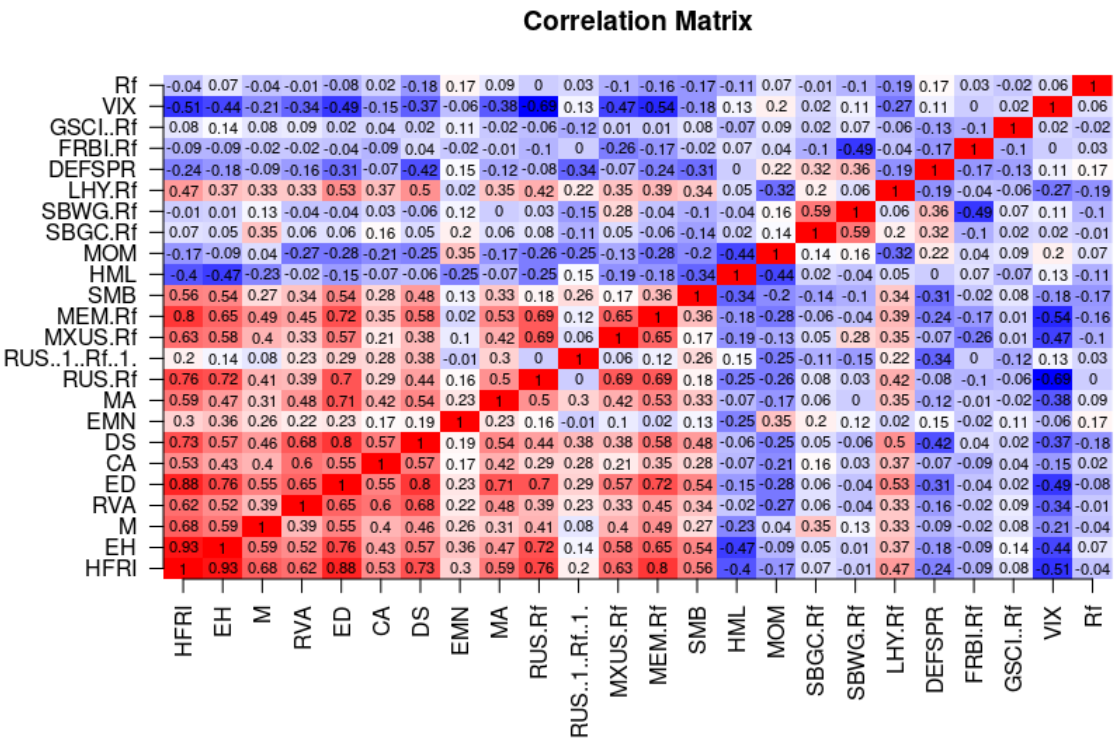
Reading the data

```
# Load the CSV file with tab as the separator and no headers
data <- read.csv("data_assignment.txt", header = FALSE, sep = "\t")

# Set custom column names
colnames(data) <- c("HFRI", "EH", "M", "RVA", "ED", "CA", "DS", "EMN",
"MA", "RUS-Rf",
                    "RUS(-1)-Rf(-1)", "MXUS-Rf", "MEM-Rf", "SMB",
"HML", "MOM",
                    "SBGC-Rf", "SBWG-Rf", "LHY-Rf", "DEFSPR",
"FRBI-Rf",
                    "GSCI--Rf", "VIX", "Rf")

# Create a sequence of dates from April 1990 to December 2005
date_sequence <- seq(as.Date("1990-04-01"), as.Date("2005-12-01"), by =
"months")

# Ensure the length of date_sequence matches the number of rows in data
if (length(date_sequence) == nrow(data)) {
  # Add date_sequence as a new column
  data <- data.frame(Date = date_sequence, data)  # Add Date as the
first column
}
```

```r
# Display the shape of the data
table_shape <- dim(data)
```

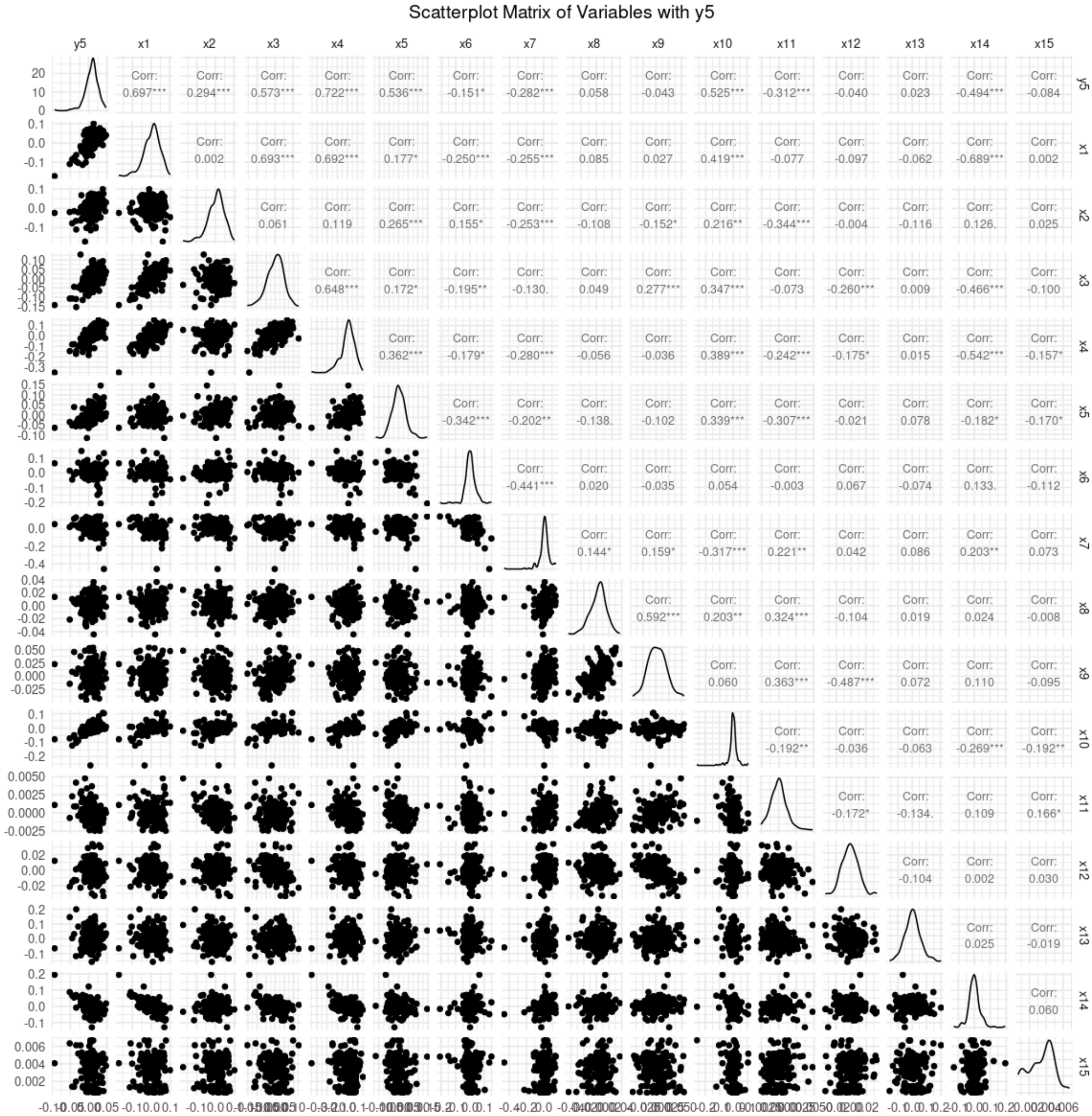
```
> str(data)
'data.frame':	189 obs. of  25 variables:
 $ Date          : Date, format: "1990-04-01" "1990-05-01" "1990-06-01" "1990-07-01" ...
 $ HFRI          : num  -0.00711 0.02383 0.01433 0.00856 -0.04085 ...
 $ EH            : num  -0.0154 0.0525 0.0185 0.0136 -0.0251 ...
 $ M             : num  -0.01411 0.03313 0.00983 0.03186 -0.04415 ...
 $ RVA           : num  0.00829 0.00903 0.00293 0.00706 -0.01095 ...
 $ ED            : num  0.00489 0.01263 0.00703 0.00496 -0.05165 ...
 $ CA            : num  0.00809 0.01083 0.01053 0.00506 -0.00815 ...
 $ DS            : num  -0.00021 0.00153 0.02983 0.01376 -0.02535 ...
 $ EMN           : num  0.00059 -0.00168 0.00703 0.00126 0.01165 ...
 $ MA            : num  0.00309 0.01613 0.00063 -0.00624 -0.01455 ...
 $ RUS.Rf        : num  -0.0356 0.0772 -0.0133 -0.0192 -0.1094 ...
 $ RUS..1..Rf..1.: num  0.0163 -0.0356 0.0772 -0.0133 -0.0192 ...
 $ MXUS.Rf       : num  -0.0195 0.0982 -0.0169 0.0066 -0.1074 ...
 $ MEM.Rf        : num  0.056 0.0707 0.0259 0.0755 -0.146 ...
 $ SMB           : num  -0.0038 -0.0271 0.013 -0.0319 -0.0358 -0.0368 -0.0562 0.005 0.0093 0.0389 ...
 $ HML           : num  -0.0264 -0.0377 -0.022 0.0006 0.0146 0.0062 0.0032 -0.0331 -0.0181 -0.0181 ...
 $ MOM           : num  0.0042 0.0418 0.0177 0.0308 0.0169 0.0366 0.0615 -0.0061 0.0629 -0.0946 ...
 $ SBGC.Rf       : num  -0.01541 0.02069 0.00952 0.00661 -0.02136 ...
 $ SBWG.Rf       : num  -0.0098 0.026 0.0115 0.0244 -0.0142 ...
 $ LHY.Rf        : num  -0.26108 0.00308 0.00804 0.01497 -0.07189 ...
 $ DEFSPR        : num  -0.0011 0.0014 0.0009 -0.0001 -0.0007 0.0009 0.0027 0.0021 0.0012 0.0001 ...
 $ FRBI.Rf       : num  -0.00069 -0.01255 -0.00311 -0.02199 -0.02144 ...
 $ GSCI..Rf      : num  -0.0172 -0.0141 -0.0223 0.0729 0.1472 ...
 $ VIX           : num  -0.0021 -0.0215 -0.0187 0.0561 0.0879 -0.0079 0.0093 -0.0788 0.0422 -0.0547 ...
 $ Rf            : num  0.00671 0.00668 0.00667 0.00644 0.00635 0.00614 0.00613 0.00604 0.00553 0.00532 ...
```

```
> summary(data)
      Date                 HFRI                 EH                  M                  RVA                  ED                  CA
 Min.   :1990-04-01   Min.   :-0.091080   Min.   :-0.08058   Min.   :-0.066870   Min.   :-0.062080   Min.   :-0.093080   Min.   :-0.035980
 1st Qu.:1994-03-01   1st Qu.:-0.003980   1st Qu.:-0.00626   1st Qu.:-0.005190   1st Qu.: 0.001020   1st Qu.:-0.000290   1st Qu.: 0.000430
 Median :1998-02-01   Median : 0.009860   Median : 0.01036   Median : 0.006630   Median : 0.006010   Median : 0.010330   Median : 0.006840
 Mean   :1998-01-30   Mean   : 0.007879   Mean   : 0.01014   Mean   : 0.009129   Mean   : 0.005991   Mean   : 0.008263   Mean   : 0.004785
 3rd Qu.:2002-01-01   3rd Qu.: 0.019220   3rd Qu.: 0.02450   3rd Qu.: 0.020680   3rd Qu.: 0.010890   3rd Qu.: 0.018560   3rd Qu.: 0.011430
 Max.   :2005-12-01   Max.   : 0.072150   Max.   : 0.10445   Max.   : 0.076290   Max.   : 0.053920   Max.   : 0.047510   Max.   : 0.029520
       DS                 EMN                MA                RUS.Rf            RUS..1..Rf..1.        MXUS.Rf               MEM.Rf
 Min.   :-0.089080   Min.   :-0.02078   Min.   :-0.060980   Min.   :-0.172070   Min.   :-0.172070   Min.   :-0.1529700   Min.   :-0.35059
 1st Qu.:-0.000520   1st Qu.:-0.00166   1st Qu.: 0.000820   1st Qu.:-0.022360   1st Qu.:-0.022360   1st Qu.:-0.0302000   1st Qu.:-0.02645
 Median : 0.008130   Median : 0.00319   Median : 0.006430   Median : 0.010140   Median : 0.010620   Median : 0.0033600   Median : 0.00853
 Mean   : 0.008437   Mean   : 0.00390   Mean   : 0.004981   Mean   : 0.003693   Mean   : 0.003799   Mean   : 0.0004348   Mean   : 0.00333
 3rd Qu.: 0.017130   3rd Qu.: 0.00927   3rd Qu.: 0.011120   3rd Qu.: 0.031950   3rd Qu.: 0.031950   3rd Qu.: 0.0325800   3rd Qu.: 0.04447
 Max.   : 0.067320   Max.   : 0.03218   Max.   : 0.023550   Max.   : 0.100140   Max.   : 0.100140   Max.   : 0.1288200   Max.   : 0.14974
      SMB                 HML                 MOM                SBGC.Rf             SBWG.Rf              LHY.Rf              DEFSPR
 Min.   :-0.116000   Min.   :-0.207900   Min.   :-0.458800   Min.   :-0.044280   Min.   :-0.044580   Min.   :-0.261080   Min.   :-2.500e-03
 1st Qu.:-0.018600   1st Qu.:-0.016700   1st Qu.:-0.012600   1st Qu.:-0.004950   1st Qu.:-0.009800   1st Qu.:-0.010720   1st Qu.:-8.000e-04
 Median : 0.001200   Median : 0.002100   Median : 0.010900   Median : 0.003830   Median : 0.002830   Median :-0.001160   Median : 0.000e+00
 Mean   : 0.002505   Mean   : 0.002241   Mean   : 0.007633   Mean   : 0.002742   Mean   : 0.002741   Mean   :-0.003456   Mean   : 1.217e-05
 3rd Qu.: 0.023700   3rd Qu.: 0.018100   3rd Qu.: 0.032800   3rd Qu.: 0.010310   3rd Qu.: 0.014600   3rd Qu.: 0.008980   3rd Qu.: 6.000e-04
 Max.   : 0.146200   Max.   : 0.149200   Max.   : 0.137500   Max.   : 0.036880   Max.   : 0.054370   Max.   : 0.106390   Max.   : 4.800e-03
    FRBI.Rf            GSCI..Rf               VIX                  Rf
 Min.   :-0.034030   Min.   :-0.156520   Min.   :-0.1290000   Min.   :0.000700
 1st Qu.:-0.008720   1st Qu.:-0.030850   1st Qu.:-0.0202000   1st Qu.:0.002470
 Median :-0.001290   Median : 0.003090   Median :-0.0015000   Median :0.003730
 Mean   :-0.001145   Mean   : 0.003233   Mean   :-0.0004053   Mean   :0.003405
 3rd Qu.: 0.006230   3rd Qu.: 0.033760   3rd Qu.: 0.0163000   3rd Qu.:0.004410
 Max.   : 0.034050   Max.   : 0.200380   Max.   : 0.1948000   Max.   :0.006710
```



Correlation Matrix
Rf
VIX
GSCI..Rf
FRBI.Rf
DEFSPR
LHY.Rf
SBWG.Rf
SBGC.Rf
MOM
HML
SMB
MEM.Rf
MXUS.Rf
RUS..1..Rf..1.
RUS.Rf
MA
EMN
DS
CA
ED
RVA
M
EH
HFRI

Question 1




```
> # Display summary of the full model
> summary(full_model)

# Full model for y5 including only x1 through x15 as predictors
full_model <- lm(y5 ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9 + x10 + x11 +
x12 + x13 + x14 + x15, data = data)


Residuals:
      Min        1Q    Median        3Q       Max 
-0.021119 -0.003606 -0.000189  0.003049  0.028772 

Coefficients:
```

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -4.301e-03  5.464e-03  -0.787 0.432261
Date         3.491e-07  4.059e-07   0.860 0.391046
y1           5.128e-01  1.476e-01   3.473 0.000658 ***
y2          -6.238e-02  6.448e-02  -0.967 0.334785
y3          -1.097e-04  3.401e-02  -0.003 0.997429
y4           3.941e-02  7.263e-02   0.543 0.588159
y6           7.802e-02  6.597e-02   1.183 0.238660
y7           2.144e-01  5.586e-02   3.838 0.000177 ***
y8          -2.817e-02  6.883e-02  -0.409 0.682909
y9           3.221e-01  6.248e-02   5.155 7.21e-07 ***
x1           8.468e-02  3.015e-02   2.808 0.005584 **
x2          -1.375e-02  1.434e-02  -0.959 0.338993
x3           1.881e-02  1.766e-02   1.065 0.288498
x4          -2.024e-02  1.535e-02  -1.319 0.189052
x5           7.715e-02  2.400e-02   3.215 0.001572 **
x6           7.722e-02  1.923e-02   4.015 9.01e-05 ***
x7           1.296e-02  1.192e-02   1.087 0.278479
x8           2.987e-02  5.887e-02   0.507 0.612515
x9          -2.890e-02  4.528e-02  -0.638 0.524280
x10         -5.749e-03  1.854e-02  -0.310 0.756868
x11         -9.334e-01  5.096e-01  -1.832 0.068839 .
x12         -2.765e-02  5.237e-02  -0.528 0.598243
x13         -5.536e-03  9.175e-03  -0.603 0.547081
x14          1.603e-02  1.992e-02   0.805 0.422118
x15          2.878e-01  4.721e-01   0.610 0.542896
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.006287 on 164 degrees of freedom
Multiple R-squared:  0.8956,	Adjusted R-squared:  0.8803
F-statistic: 58.59 on 24 and 164 DF,  p-value: < 2.2e-16
```

Statistical Findings

Model Fit
The model has a very strong overall fit with an R-squared of 0.8956 (Adjusted R-squared: 0.8803)
This means approximately 89.56% of the variance in ED is explained by the predictors
The model is highly significant overall (F-statistic: 58.59, p-value < 2.2e-16)

Highly Significant Predictors (p < 0.001 '*'):**

- y1 (HFRI): β = 0.5128, indicating a strong positive relationship
- y7 (DS): β = 0.2144, showing a moderate positive relationship
- y9 (MA): β = 0.3221, demonstrating a strong positive relationship
- x6 (HML): β = 0.0772, showing a positive relationship

Significant Predictors (p < 0.01 ''):**

- x1 (RUS-Rf): β = 0.0847
- x5 (SMB): β = 0.0772

Marginally Significant Predictor (p < 0.1 '.'):

- x11 (DEFSPR): β = -0.9334, suggesting a potentially important negative relationship

Residual Analysis:

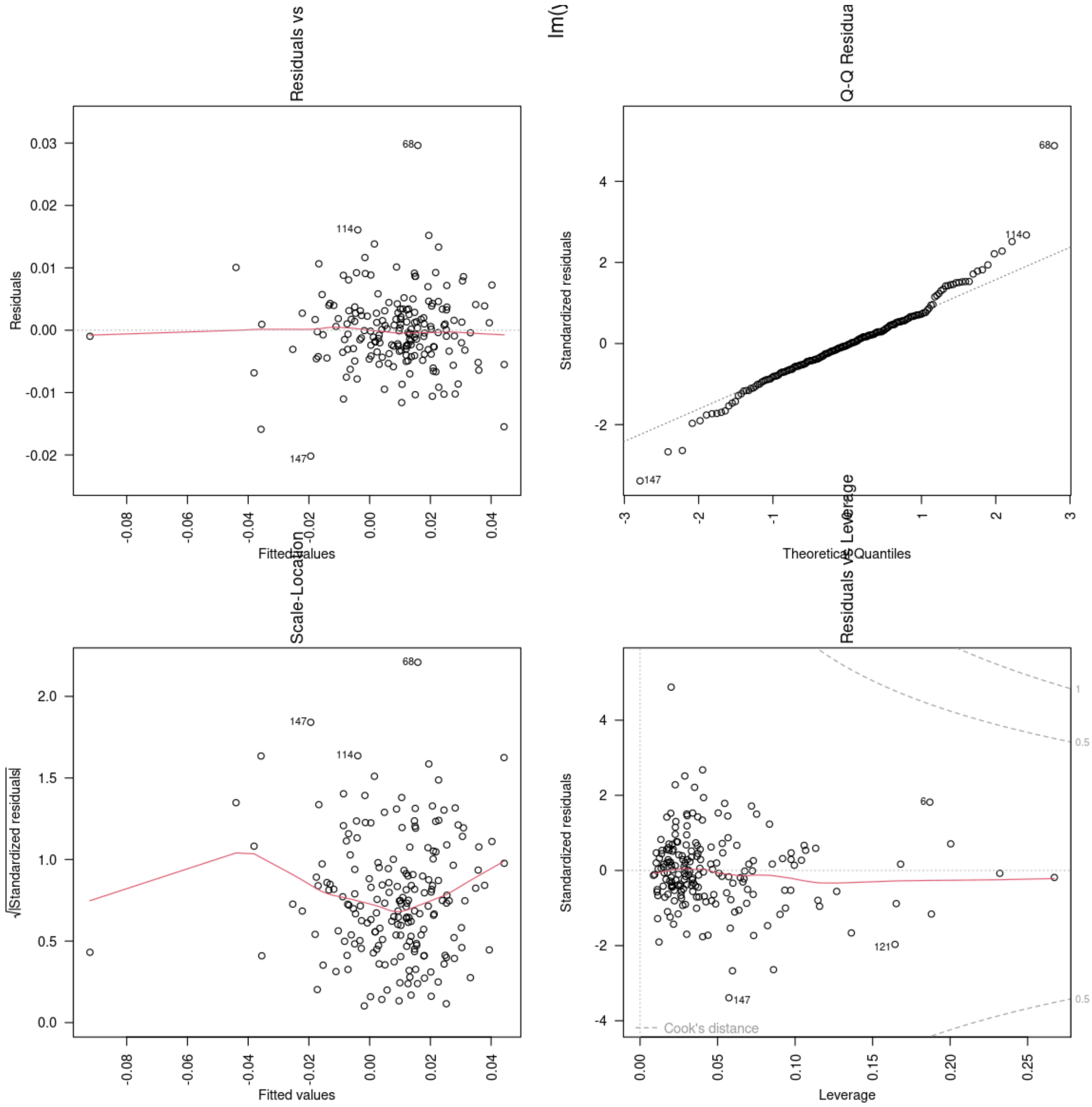
- The residuals range from -0.021119 to 0.028772
- The residual standard error is 0.006287, indicating relatively small prediction errors
- The distribution of residuals appears fairly symmetric around zero, suggesting good model fit

Most other variables in the model, including Date and various market factors, show no statistically significant relationship with ED.

Stepwise selection of regressors for y5

```
# Stepwise selection of regressors for y5
stepwise_model <- step(full_model, direction = "both")  # "both" adds or
removes variables
summary(stepwise_model)  # View the final model

Call:
lm(formula = y5 ~ y1 + y6 + y7 + y9 + x1 + x5 + x6 + x11, data = data)

Residuals:
       Min         1Q     Median         3Q        Max
-0.0201817 -0.0033633 -0.0002468  0.0031117  0.0296270

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.0006419  0.0005629   1.140 0.255659
y1           0.4000246  0.0645099   6.201 3.74e-09 ***
y6           0.0839679  0.0585560   1.434 0.153314
y7           0.2239428  0.0465491   4.811 3.17e-06 ***
y9           0.3252089  0.0543899   5.979 1.18e-08 ***
x1           0.0739171  0.0200607   3.685 0.000303 ***
x5           0.0624734  0.0186338   3.353 0.000976 ***
x6           0.0610610  0.0139307   4.383 1.98e-05 ***
x11         -0.6580328  0.4220025  -1.559 0.120679
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.006134 on 180 degrees of freedom
```

```
Multiple R-squared:  0.8909,	Adjusted R-squared:  0.886 
F-statistic: 183.7 on 8 and 180 DF,  p-value: < 2.2e-16
```



Model Efficiency

- The stepwise selection has reduced the model from 24 to 8 predictors while maintaining similar explanatory power
- R-squared remains high at 0.8909 (only slightly lower than the full model's 0.8956)
- Adjusted R-squared of 0.886 is actually higher than the full model, indicating better model efficiency
- The model remains highly significant (F-statistic: 183.7, p-value < 2.2e-16)

Highly Significant Predictors (p < 0.001 '***'):

- y1 (HFRI): β = 0.4000, strongest positive relationship
- y7 (DS): β = 0.2239, substantial positive effect

- y9 (MA): β = 0.3252, strong positive relationship
- x1 (RUS-Rf): β = 0.0739, moderate positive effect
- x5 (SMB): β = 0.0625, positive size factor effect
- x6 (HML): β = 0.0611, positive value factor effect

Non-Significant but Retained Variables:

- y6 (CA): β = 0.0840, p = 0.153
- x11 (DEFSPR): β = -0.6580, p = 0.121 These variables were retained by the stepwise procedure, suggesting potential importance despite not reaching statistical significance

Residual Analysis:

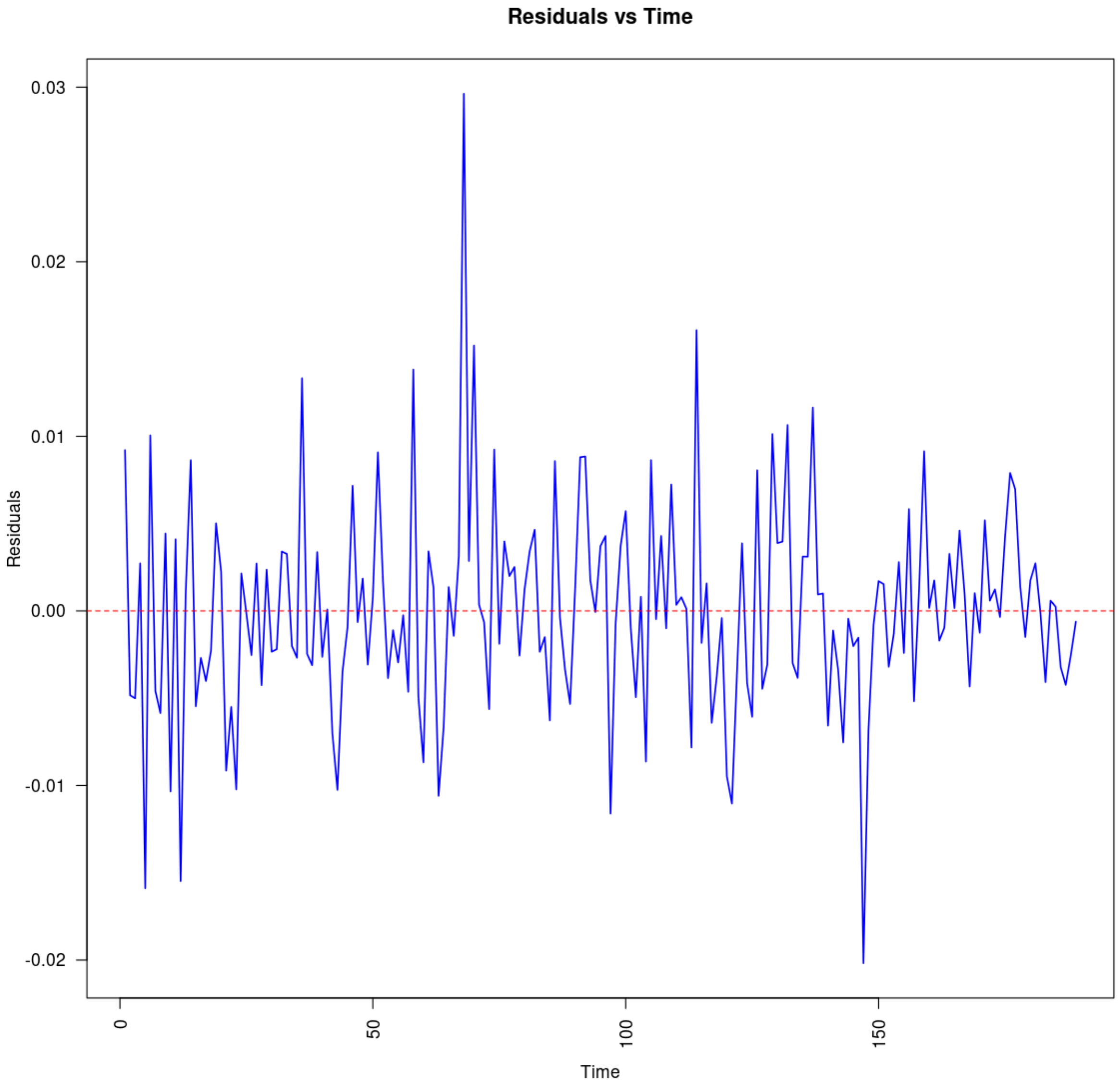
- Residuals range from -0.0202 to 0.0296, similar to the full model
- Residual standard error (0.006134) is slightly better than the full model (0.006287)

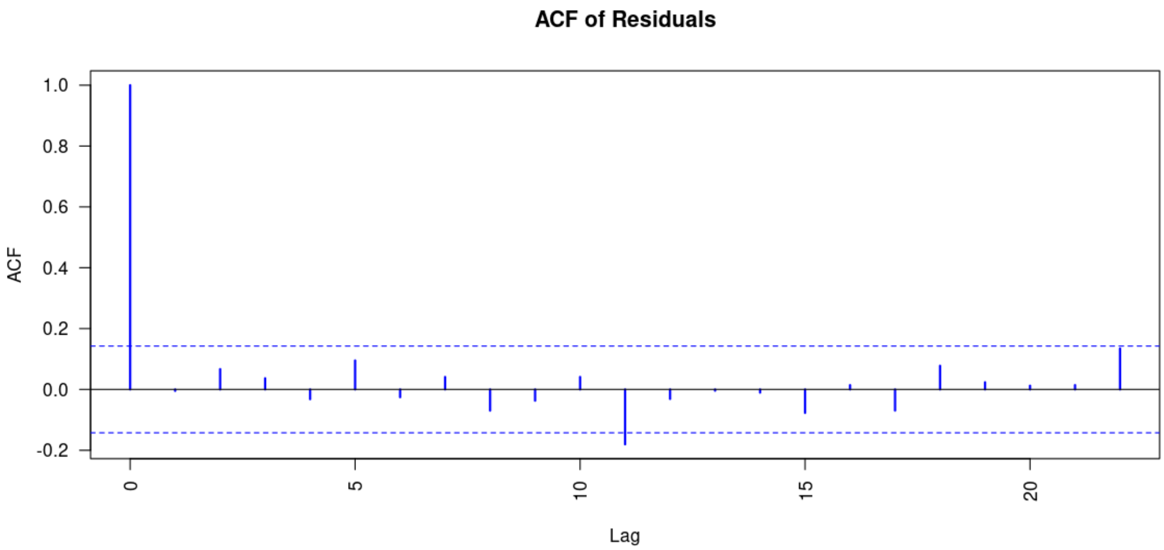
Model Improvement:

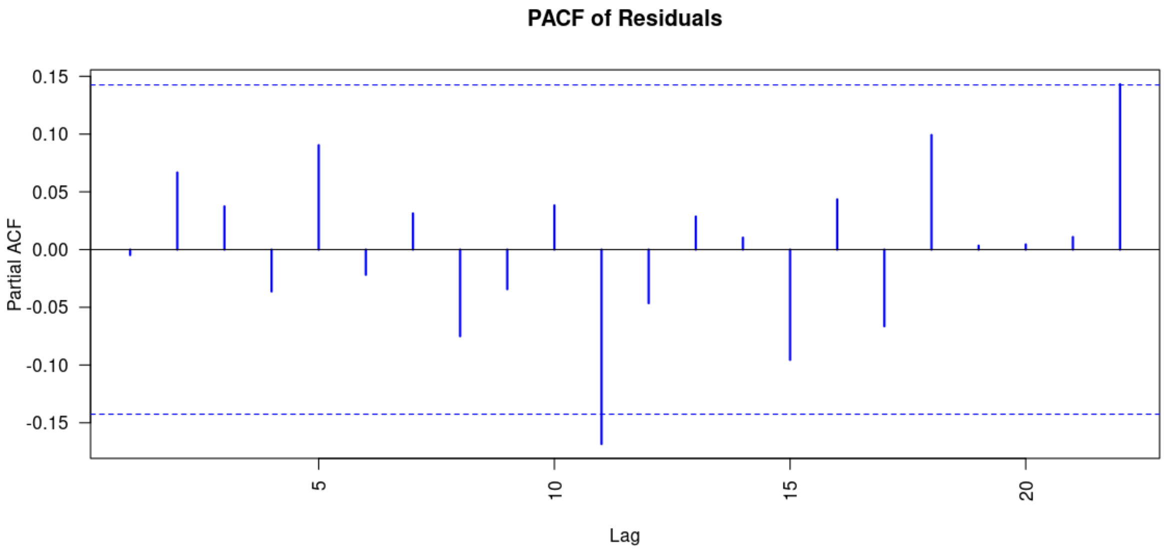
- The stepwise selection has created a more parsimonious model
- Removed 16 non-contributing variables while maintaining similar predictive power
- Higher F-statistic (183.7 vs 58.59) indicates better overall model efficiency

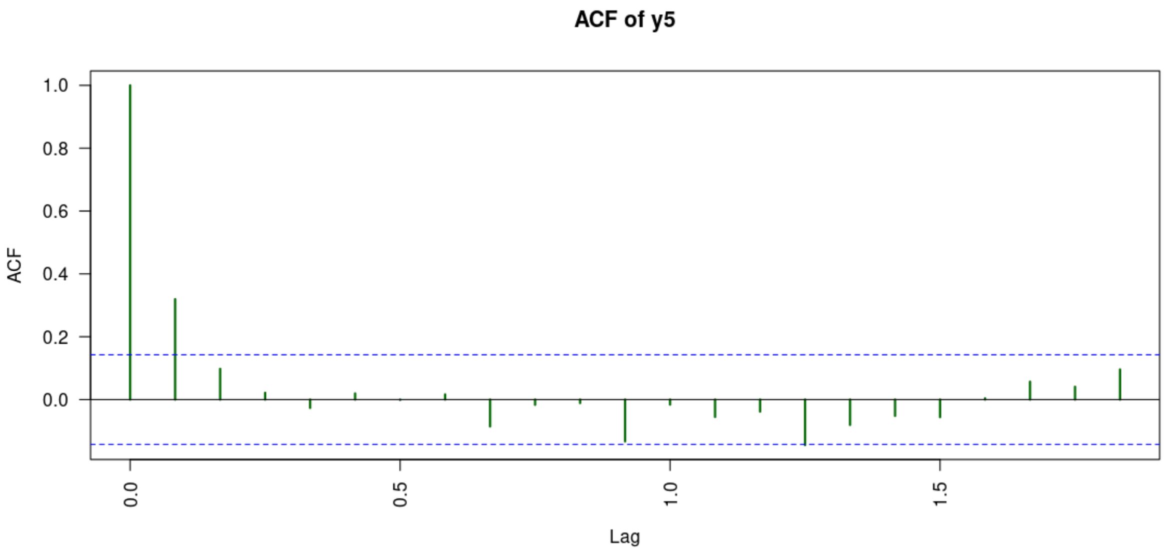
Question 2

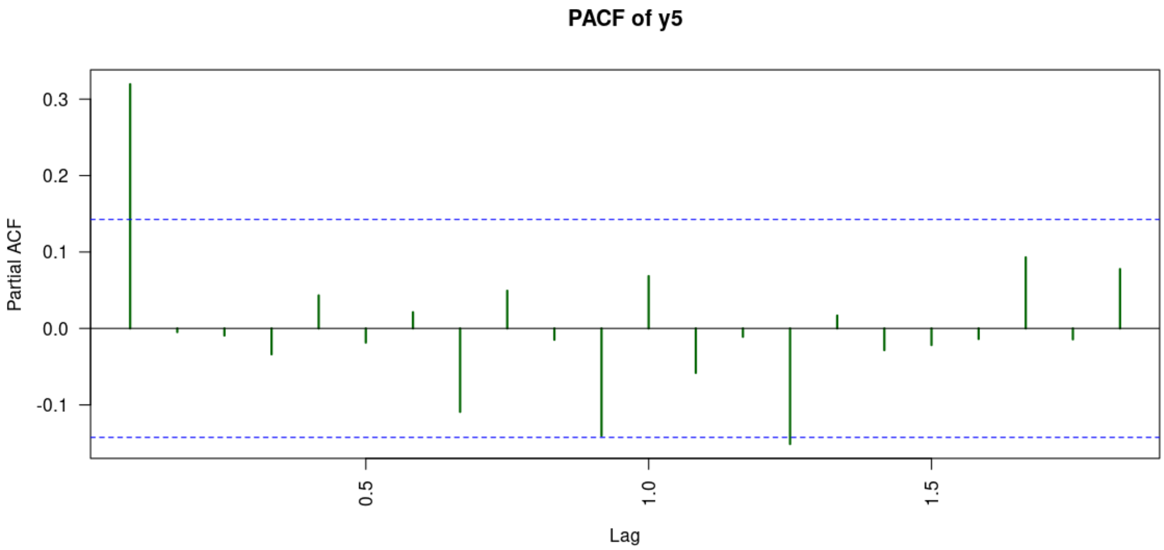
```
# Set up single plot layout for residuals vs. time plot
par(mfrow = c(1, 1))
# Calculate and plot residuals
residuals_stepwise <- residuals(stepwise_model)
plot(residuals_stepwise, type = "l", main = "Residuals vs Time", ylab =
"Residuals", xlab = "Time", col = "blue", lwd = 1.5)
abline(h = 0, col = "red", lty = 2)  # Add horizontal line at zero for
reference
```


Residuals vs Time
0.03
0.02
0.01
0.00
-0.01
-0.02
Residuals
0
50
100
150
Time


ACF of Residuals
ACF
Lag
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
0
5
10
15
20


PACF of Residuals
Partial ACF
Lag
0.15
0.10
0.05
0.00
-0.05
-0.10
-0.15
5
10
15
20





Autocorrelation

The output displays the results of the Box-Pierce and Ljung-Box tests applied to the time series data y5. Both tests measure the presence of autocorrelation within the series. For the Box-Pierce test, the test statistic (X-squared) is 19.292 with 1 degree of freedom and a p-value of approximately 1.122×10−51.122 \times 10^{-5}1.122×10−5. For the Ljung-Box test, the X-squared value is 19.6 with the same degrees of freedom and a p-value of about 9.546×10−69.546 \times 10^{-6}9.546×10−6. In both cases, the very low p-values indicate significant evidence against the null hypothesis of no autocorrelation, suggesting that the time series y5 exhibits autocorrelation.

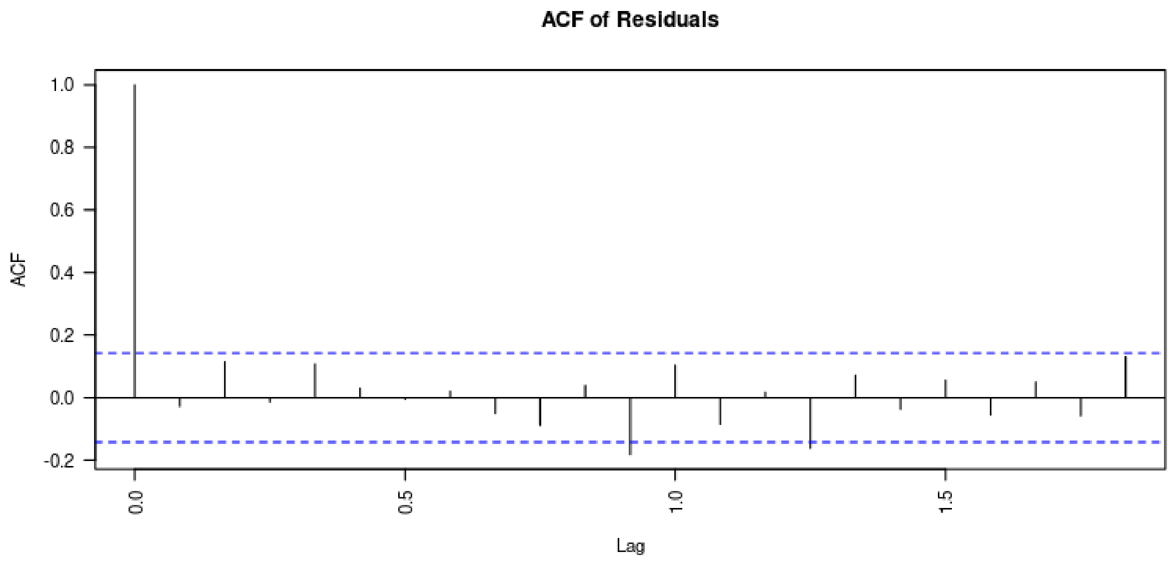
```
[1] "Box-Pierce Test Result for y5:"
> print(box_pierce_test)

	Box-Pierce test

data:  y5
```

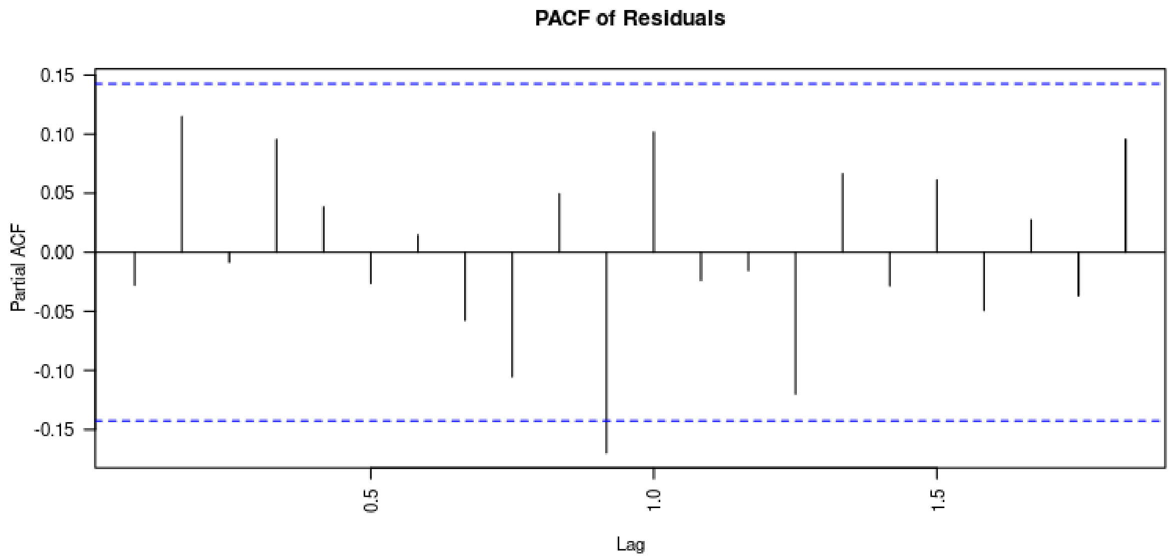
```
X-squared = 19.292, df = 1, p-value = 1.122e-05

> print("Ljung-Box Test Result for y5:")
[1] "Ljung-Box Test Result for y5:"
> print(ljung_box_test)

	Box-Ljung test

data:  y5
X-squared = 19.6, df = 1, p-value = 9.546e-06
```

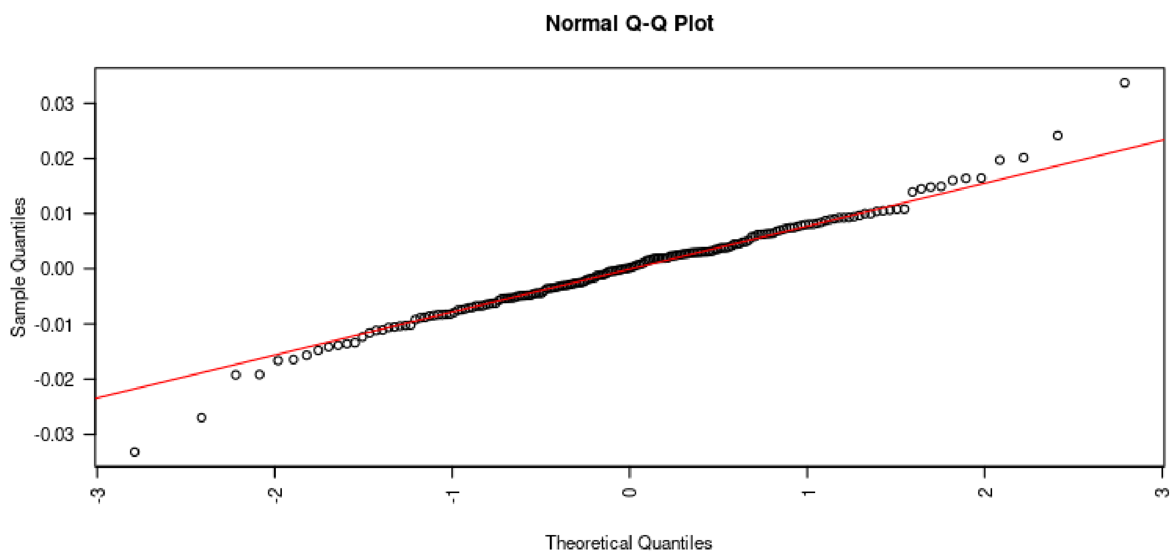
Sub question 2A

The Augmented Dickey-Fuller (ADF) test checks if a time series is stationary, meaning its statistical properties don't change over time. Here, the test statistic appears as -8.7749, which, if lower (more negative) than the ADF critical values, would indicate that we can reject the null hypothesis of non-stationarity, suggesting the series is stationary. The other values (25.6695 and 38.5024) may represent additional test metrics or critical values, though they are not in a standard ADF output format. Assuming -8.7749 is the main test statistic, it likely indicates that the time series is stationary.

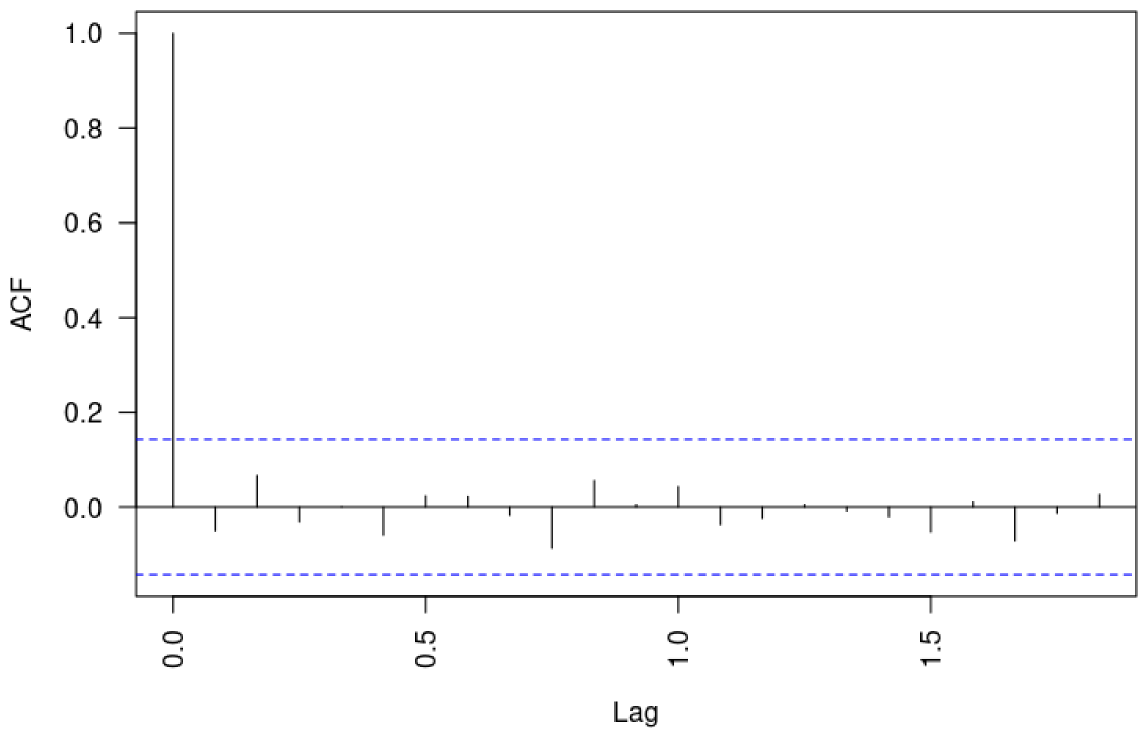
```
Call:
arima(x = y5, order = c(1, 0, 0), xreg = x_reg)

Coefficients:
      ar1  intercept  x_reg1  x_reg2  x_reg3  x_reg4  x_reg5  x_reg6  x_reg7  x_reg8
x_reg9
    0.2455     0.0046  0.2230  0.0569  0.0739  0.2129  0.0798  0.0314  0.0091  0.0085
0.4221
s.e.  0.0747     0.0022  0.0271  0.0178  0.0147  0.0245  0.0251  0.0135  0.0116  0.0239
0.5853

sigma^2 estimated as 7.607e-05:  log likelihood = 628.02,  aic = -1232.03


###############################################################
# Augmented Dickey-Fuller Test Unit Root / Cointegration Test #
###############################################################

The value of the test statistic is: -8.7749 25.6695 38.5024
```


ACF of Residuals
ACF
Lag
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
0.0
0.5
1.0
1.5

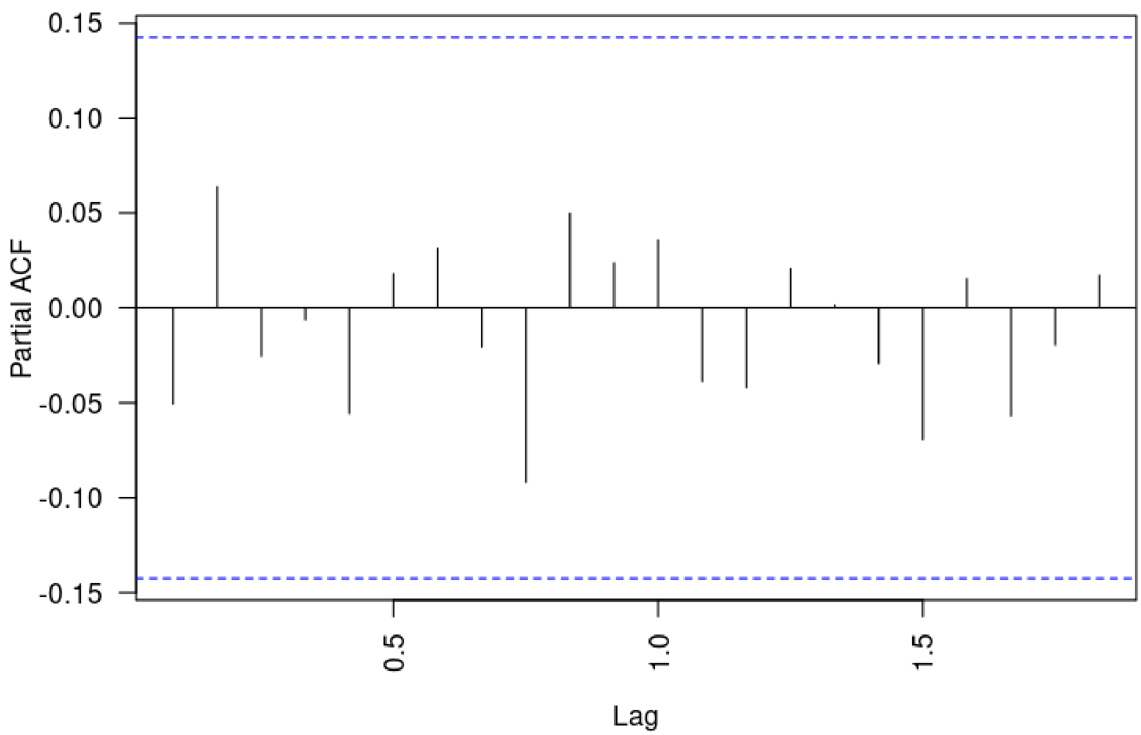

PACF of Residuals
Partial ACF
Lag
0.15
0.10
0.05
0.00
-0.05
-0.10
-0.15
0.5
1.0
1.5


Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles
0.03
0.02
0.01
0.00
-0.01
-0.02
-0.03
-3
-2
-1
0
1
2
3

ACF of Squared Residuals



PACF of Squared Residuals



```
Box.test(residuals_ar, lag = 10, type = "Ljung-Box")

	Box-Ljung test

data:  residuals_ar
X-squared = 7.6754, df = 10, p-value = 0.6605

> Box.test(residuals_ar^2, lag = 10, type = "Ljung-Box")  # Check for ARCH
effects

	Box-Ljung test

data:  residuals_ar^2
X-squared = 4.6004, df = 10, p-value = 0.9162

>
> # Reset plotting layout to single plot
> par(mfrow = c(1, 1))
```

AR(2) Model

Model Fitting

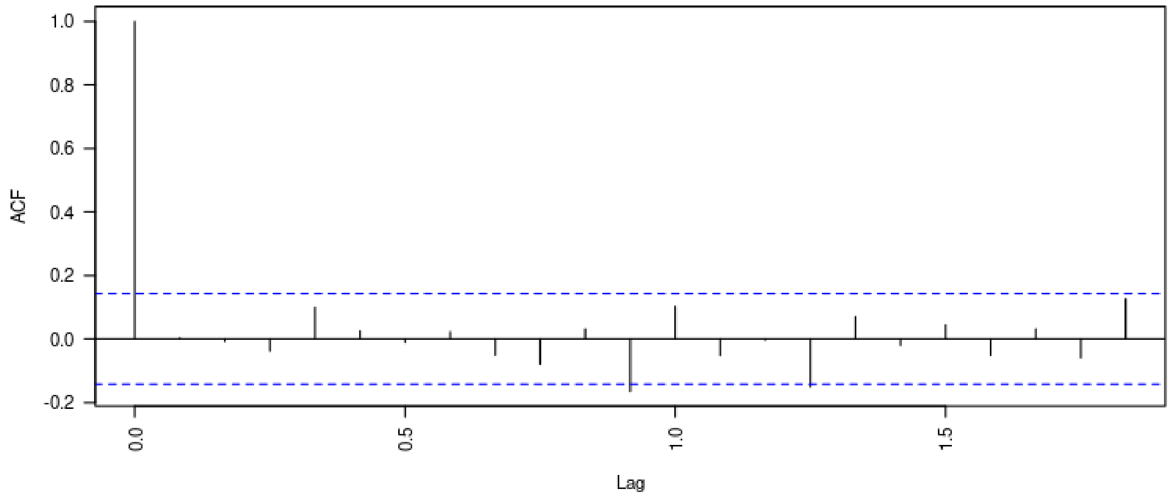
```
Call:
arima(x = y5, order = c(2, 0, 0), xreg = x_reg)

Coefficients:
         ar1     ar2  intercept  x_reg1  x_reg2  x_reg3  x_reg4  x_reg5  x_reg6  x_reg7  x_reg8  x_reg9
      0.2112  0.1290     0.0045  0.2203  0.0581  0.0750  0.2131  0.0811  0.0368  0.0098  0.0067  0.4305
s.e.  0.0758  0.0752     0.0025  0.0270  0.0175  0.0148  0.0247  0.0250  0.0139  0.0113  0.0239  0.6530

sigma^2 estimated as 7.49e-05:  log likelihood = 629.47,  aic = -1232.94
```

Augmented Dickey-Fuller Test

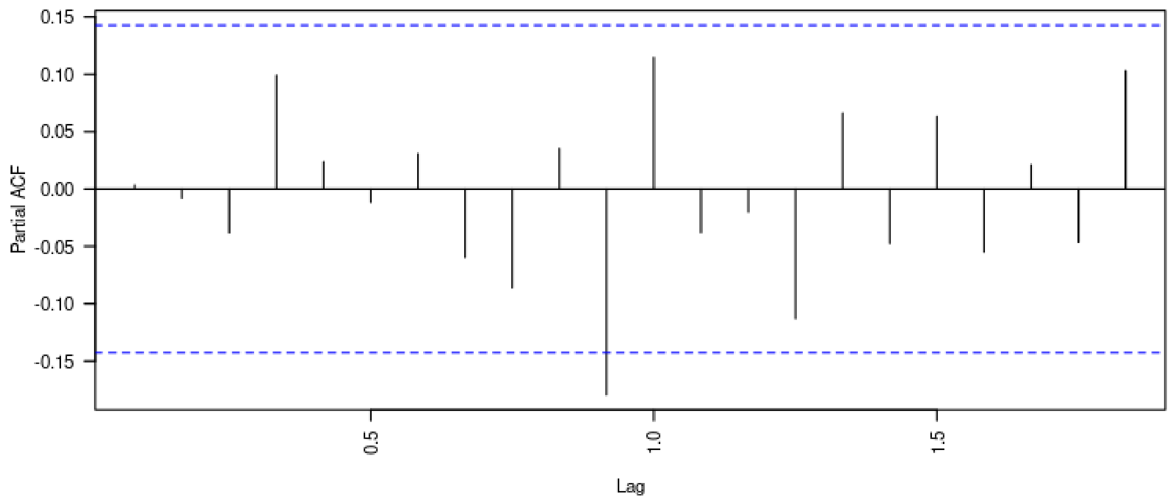
```
############################################################### 
# Augmented Dickey-Fuller Test Unit Root / Cointegration Test # 
############################################################### 

The value of the test statistic is: -8.2657 22.7823 34.172
```

ACF, PACF and Normal Q-Q plots

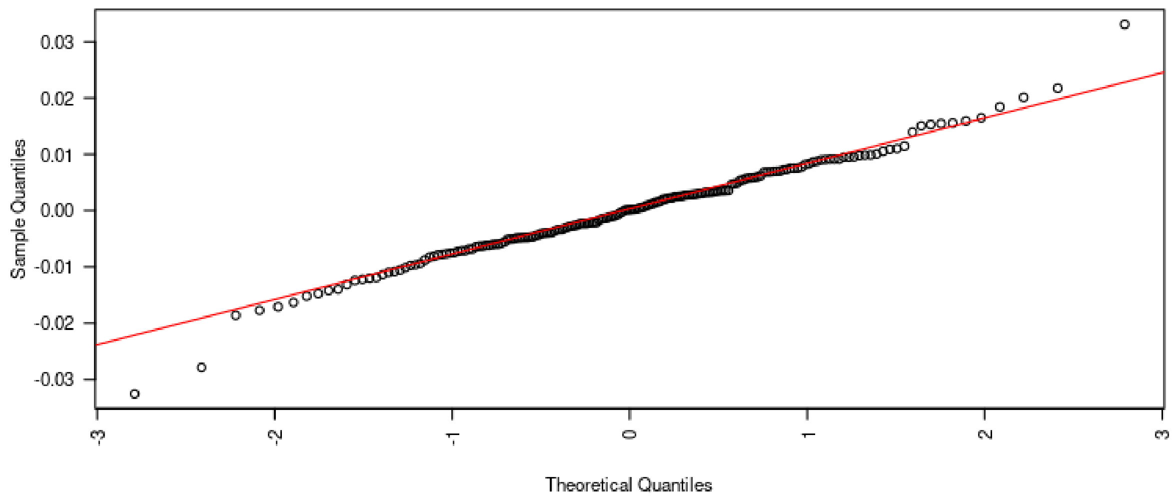
ACF of Residuals - AR(2)



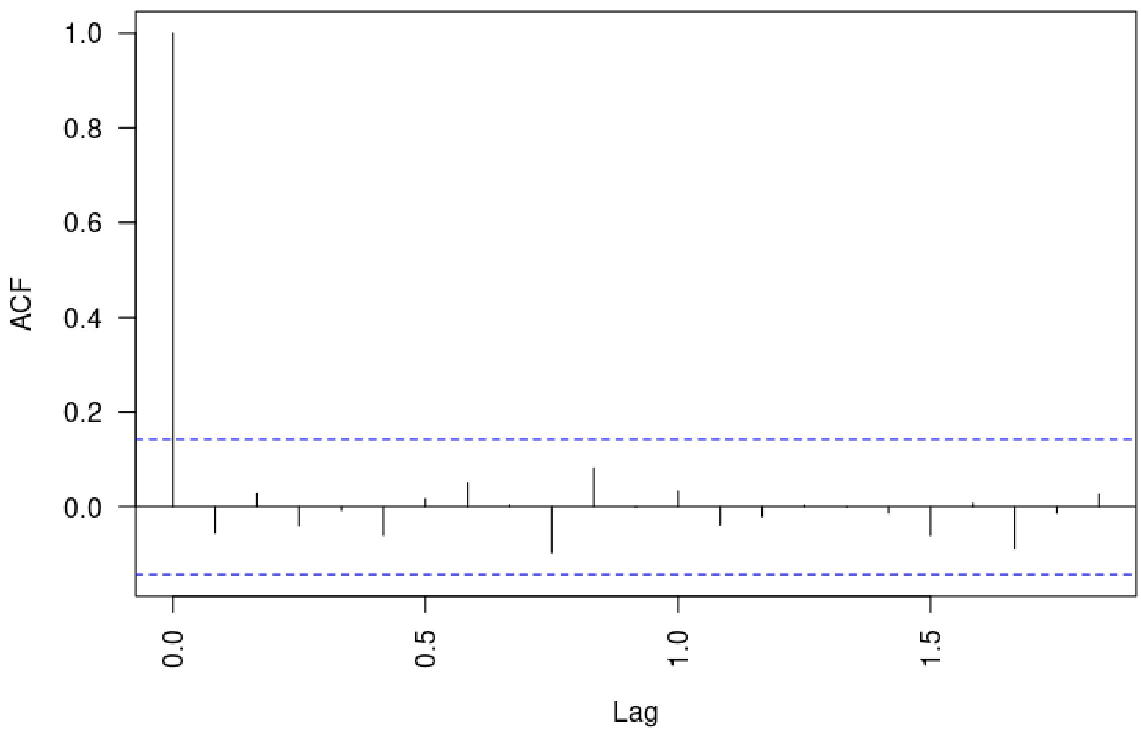
PACF of Residuals - AR(2)



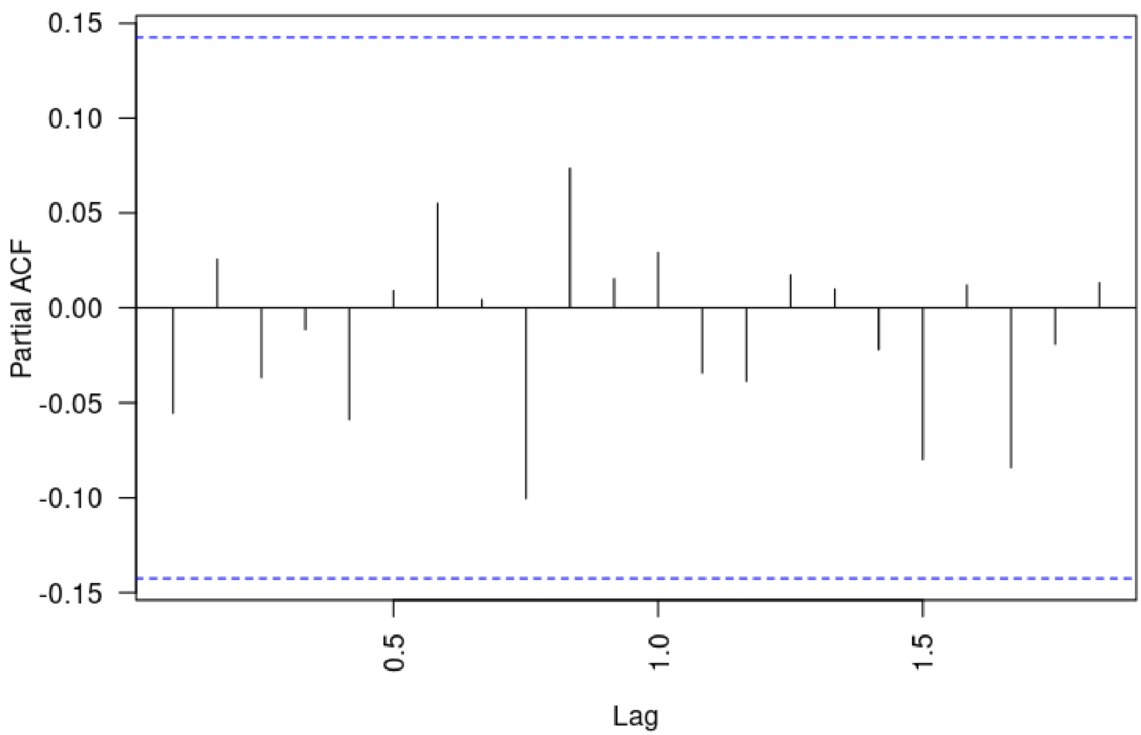
Normal Q-Q Plot - AR(2)



ACF of Squared Residuals - AR(2)



PACF of Squared Residuals - AR(2)



The Ljung-Box test, a statistical test used to check for the presence of autocorrelation in residuals, is particularly useful in time series analysis to evaluate if a model's residuals resemble white noise (i.e., lack significant autocorrelation). In the given example, the test is applied to the squared residuals of an AR(2) model using 10 lags (lag = 10) with type = "Ljung-Box", which specifies the Ljung-Box variant of the test. The output shows a test statistic (X-squared) of 5.5679, a degrees of freedom (df) of 10, and a p-value of 0.8502. This high p-value (significantly above a typical alpha level of 0.05) suggests that we fail to reject the null hypothesis, implying there is no significant autocorrelation within these residuals up to the 10th lag. Thus, the model's residuals appear to behave like white noise, supporting the adequacy of the AR(2) model in capturing the data's structure.

Ljung-Box tests

```
> Box.test(residuals_ar2^2, lag = 10, type = "Ljung-Box")

	Box-Ljung test

data:  residuals_ar2^2
X-squared = 5.5679, df = 10, p-value = 0.8502
```

MA(1) Model

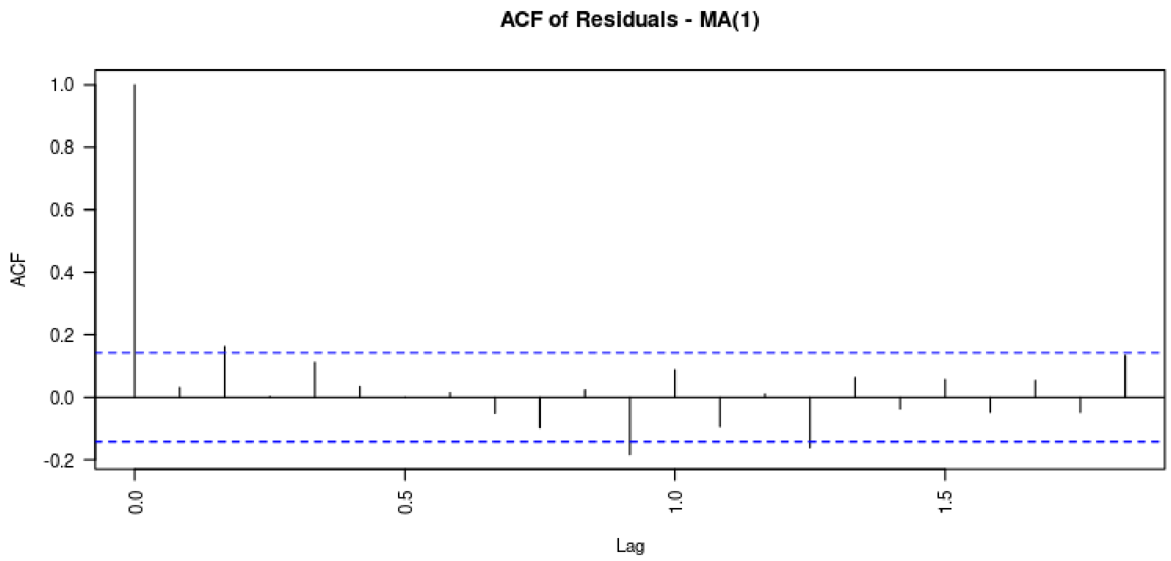
Model Fitting

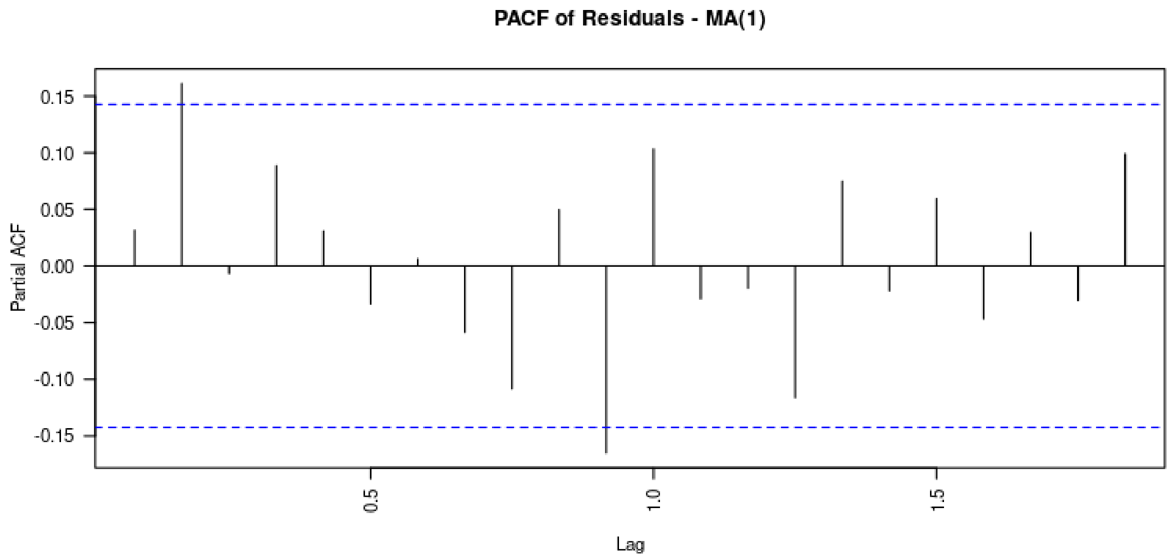
```
Call:
arima(x = y5, order = c(0, 0, 1), xreg = x_reg)

Coefficients:
         ma1  intercept  x_reg1  x_reg2  x_reg3  x_reg4  x_reg5  x_reg6  x_reg7  x_reg8  x_reg9
      0.1874     0.0045  0.2274  0.0596  0.0737  0.2141  0.0846  0.0321  0.0103  0.0098  0.4292
s.e.  0.0653     0.0020  0.0271  0.0178  0.0147  0.0248  0.0250  0.0140  0.0118  0.0245  0.5341

sigma^2 estimated as 7.711e-05:  log likelihood = 626.74,  aic = -1229.49
```

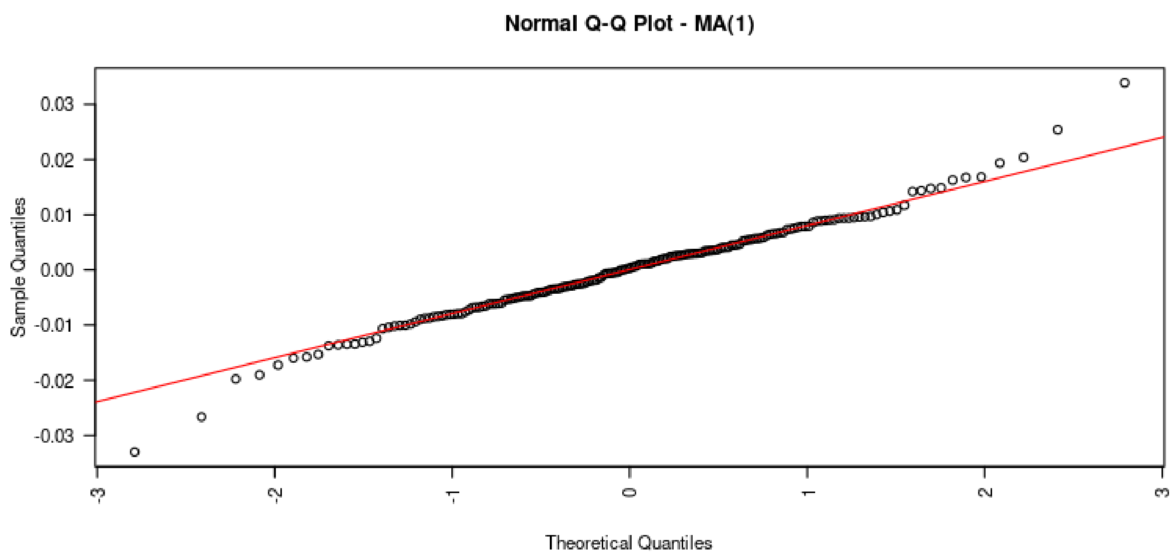
Augmented Dickey-Fuller Test

```
###############################################################
# Augmented Dickey-Fuller Test Unit Root / Cointegration Test #
###############################################################

The value of the test statistic is: -8.1345 22.0604 33.0895
```

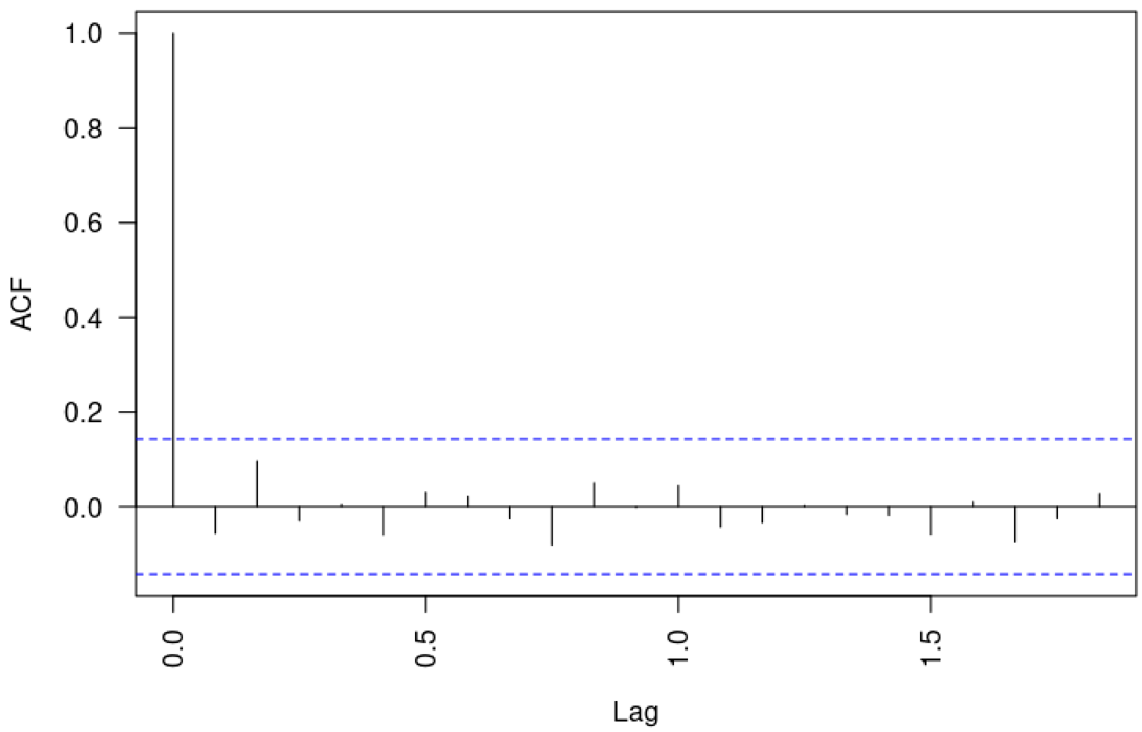
ACF, PACF and Normal Q-Q plots


ACF of Residuals - MA(1)
ACF
Lag
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
0.0
0.5
1.0
1.5

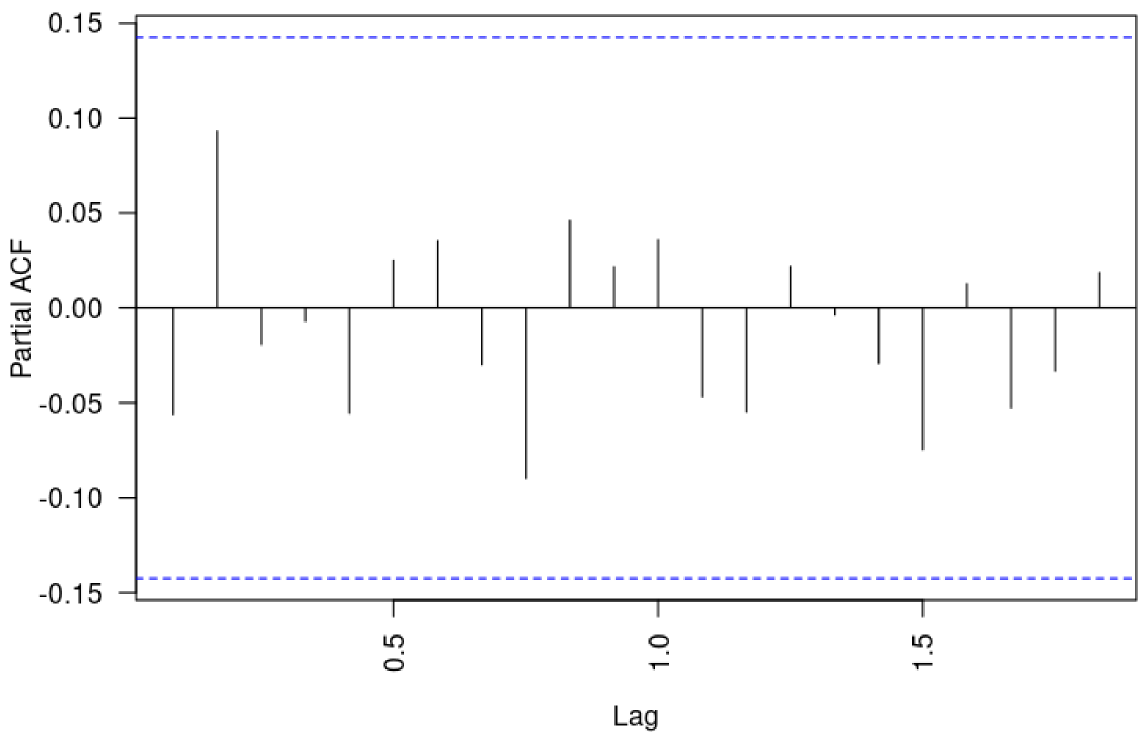

PACF of Residuals - MA(1)
Partial ACF
Lag
0.15
0.10
0.05
0.00
-0.05
-0.10
-0.15
0.5
1.0
1.5


Normal Q-Q Plot - MA(1)
Sample Quantiles
Theoretical Quantiles
0.03
0.02
0.01
0.00
-0.01
-0.02
-0.03
-3
-2
-1
0
1
2
3

ACF of Squared Residuals - MA(1)



PACF of Squared Residuals - MA(1)



```
        Box-Ljung test

data:  residuals_ma1^2
X-squared = 5.4614, df = 10, p-value = 0.8583
```

ARMA(1,1)

Model Fitting

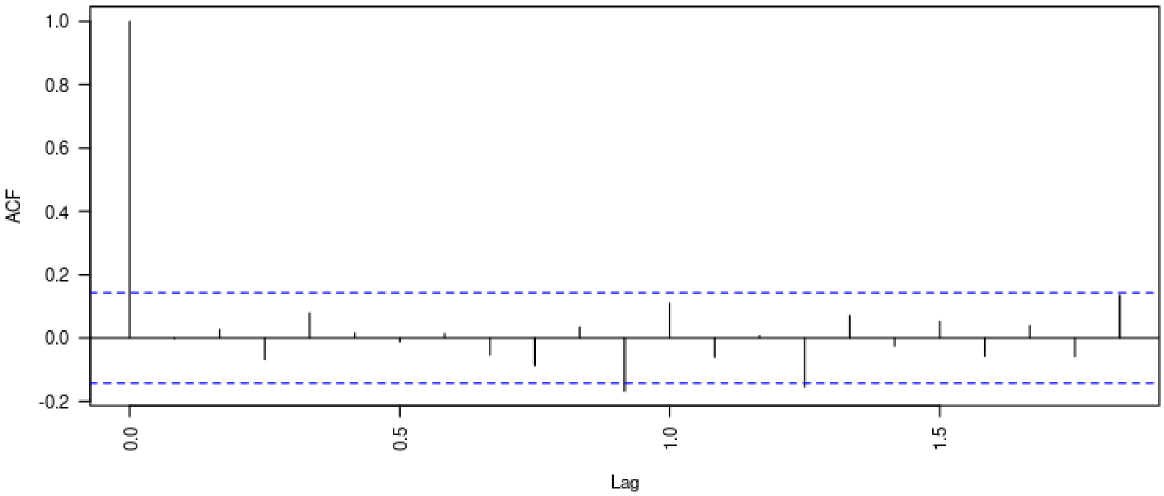
```
Call:
arima(x = y5, order = c(1, 0, 1), xreg = x_reg)

Coefficients:
         ar1      ma1  intercept  x_reg1  x_reg2  x_reg3  x_reg4  x_reg5  x_reg6  x_reg7  x_reg8  x_reg9
      0.6527  -0.4383     0.0045  0.2187  0.0577  0.0748  0.2106  0.0801  0.0345  0.0099  0.0064  0.4187
s.e.  0.1754   0.2078     0.0026  0.0272  0.0175  0.0147  0.0248  0.0250  0.0135  0.0114  0.0238  0.6862

sigma^2 estimated as 7.495e-05:  log likelihood = 629.41,  aic = -1232.81
```

Augmented Dickey-Fuller Test

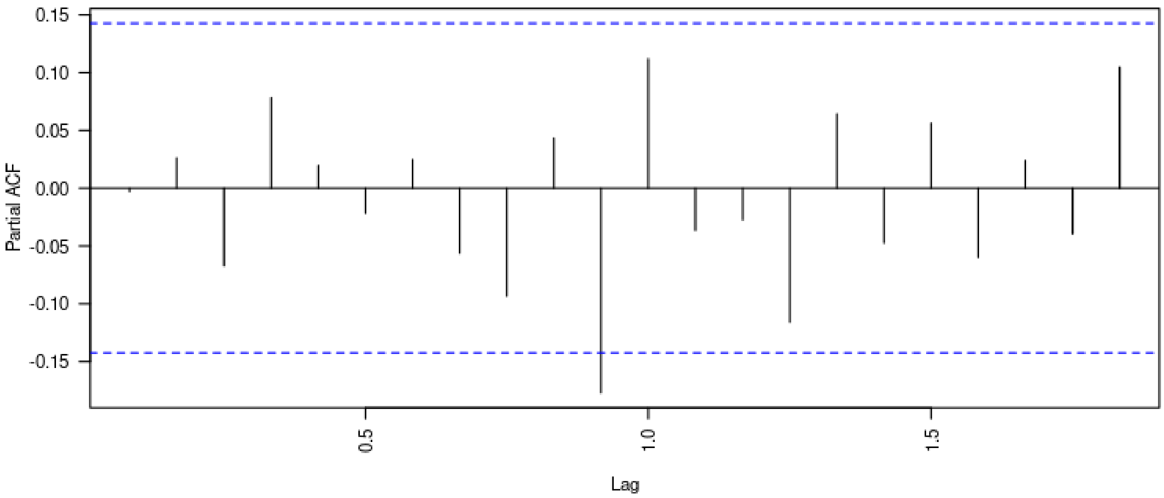
```
################################################################
# Augmented Dickey-Fuller Test Unit Root / Cointegration Test #
################################################################

The value of the test statistic is: -9.4425 29.7234 44.5823
```

ACF, PACF and Normal Q-Q plots

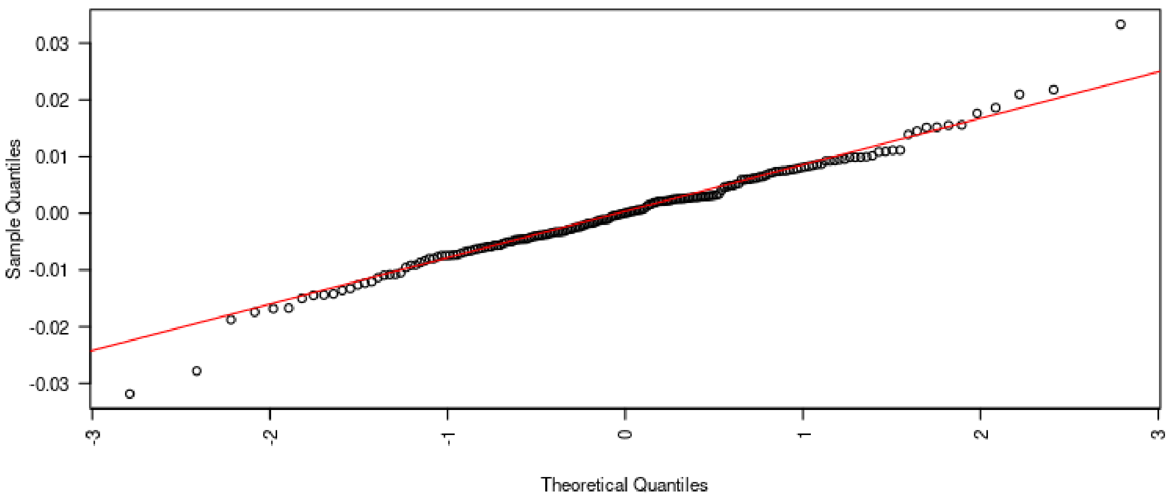
ACF of Residuals - ARMA(1,1)

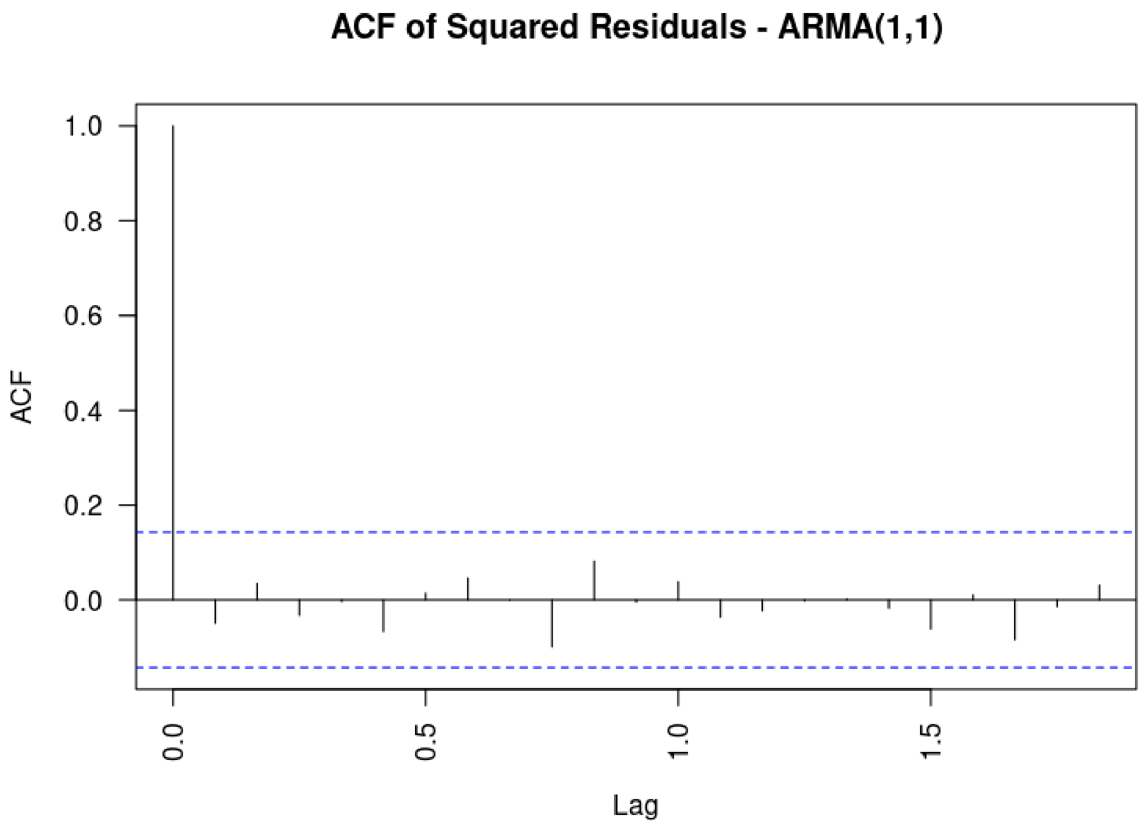


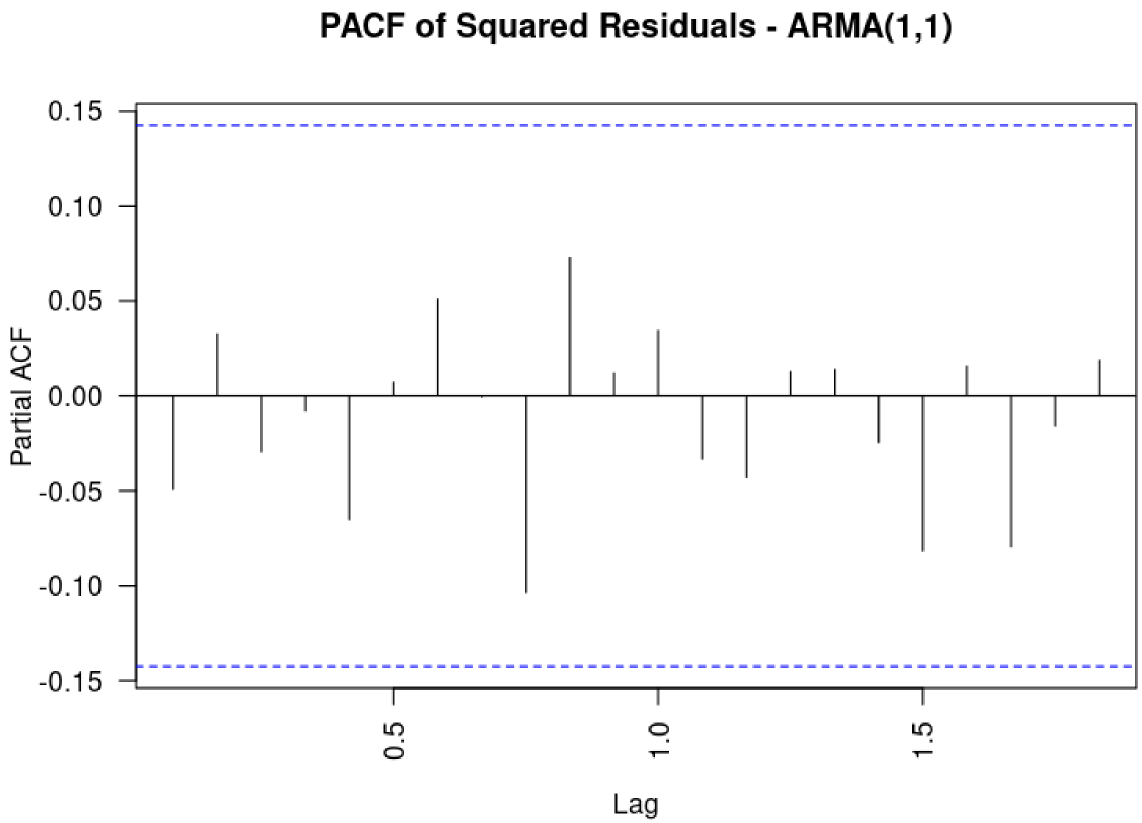
PACF of Residuals - ARMA(1,1)



Normal Q-Q Plot - ARMA(1,1)




ACF of Squared Residuals - ARMA(1,1)
ACF
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.5
1.0
1.5
Lag


PACF of Squared Residuals - ARMA(1,1)
Partial ACF
0.15
0.10
0.05
0.00
-0.05
-0.10
-0.15
0.5
1.0
1.5
Lag

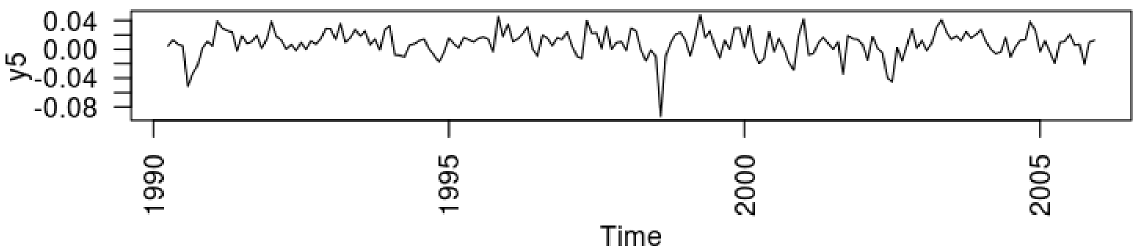
```
        Box-Ljung test

data:  residuals_arma11^2
X-squared = 5.5478, df = 10, p-value = 0.8517
```

Comparison of models

The table compares three models—AR(2), MA(1), and ARMA(1,1)—based on their Akaike Information Criterion (AIC) and log-likelihood values. A lower AIC indicates a better model fit, and a higher log-likelihood suggests a more likely model given the data. The AR(2) model has the lowest AIC at -1232.94 and the highest log-likelihood at 629.47, making it the best choice among the three models, as it balances model fit and complexity most effectively. The ARMA(1,1) model is close in both AIC and log-likelihood but slightly underperforms compared to the AR(2) model.

Model	**AIC**	**Log Likeligood**
AR(2)	-1232.94	629.47
MA(1)	-1229.49	626.74
ARMA(1,1)	aic = -1232.81	629.41

Sub question 2B



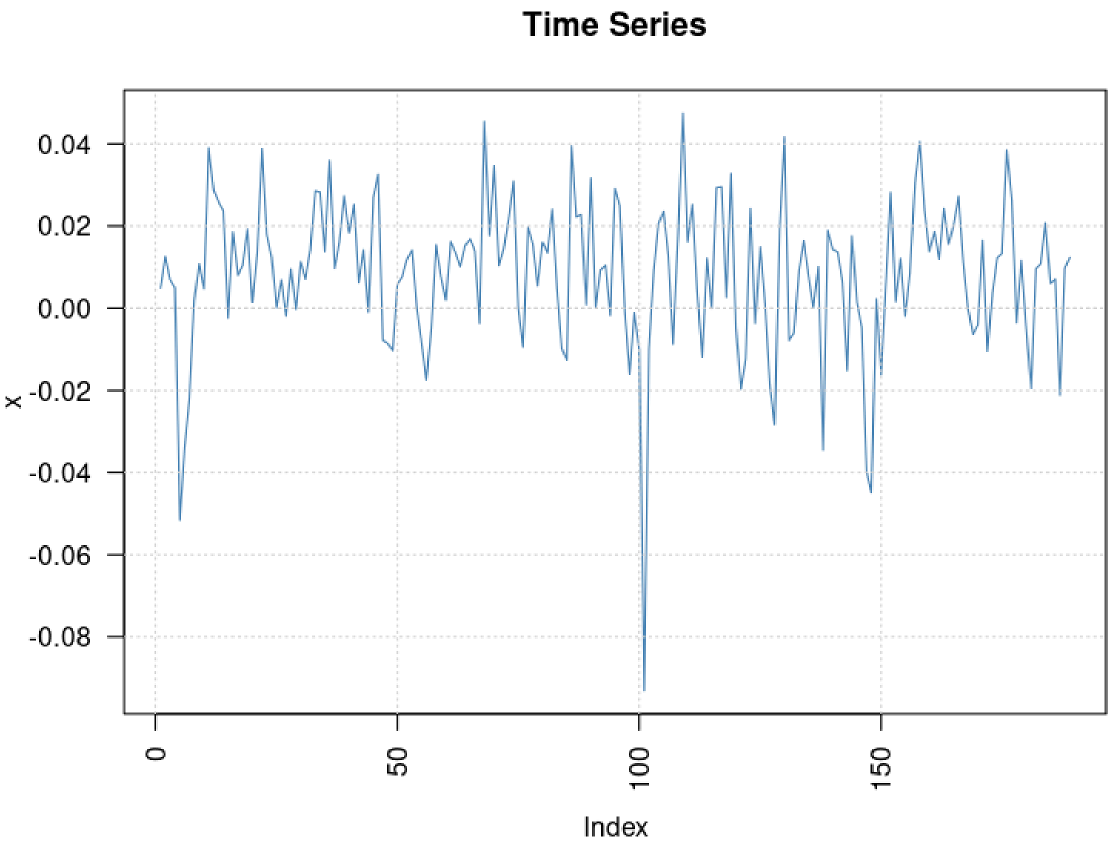

```
        Box-Ljung test

data:  y5
X-squared = 26.996, df = 12, p-value = 0.007737

> Box.test(y5^2, lag = 12, type = "Ljung")
```

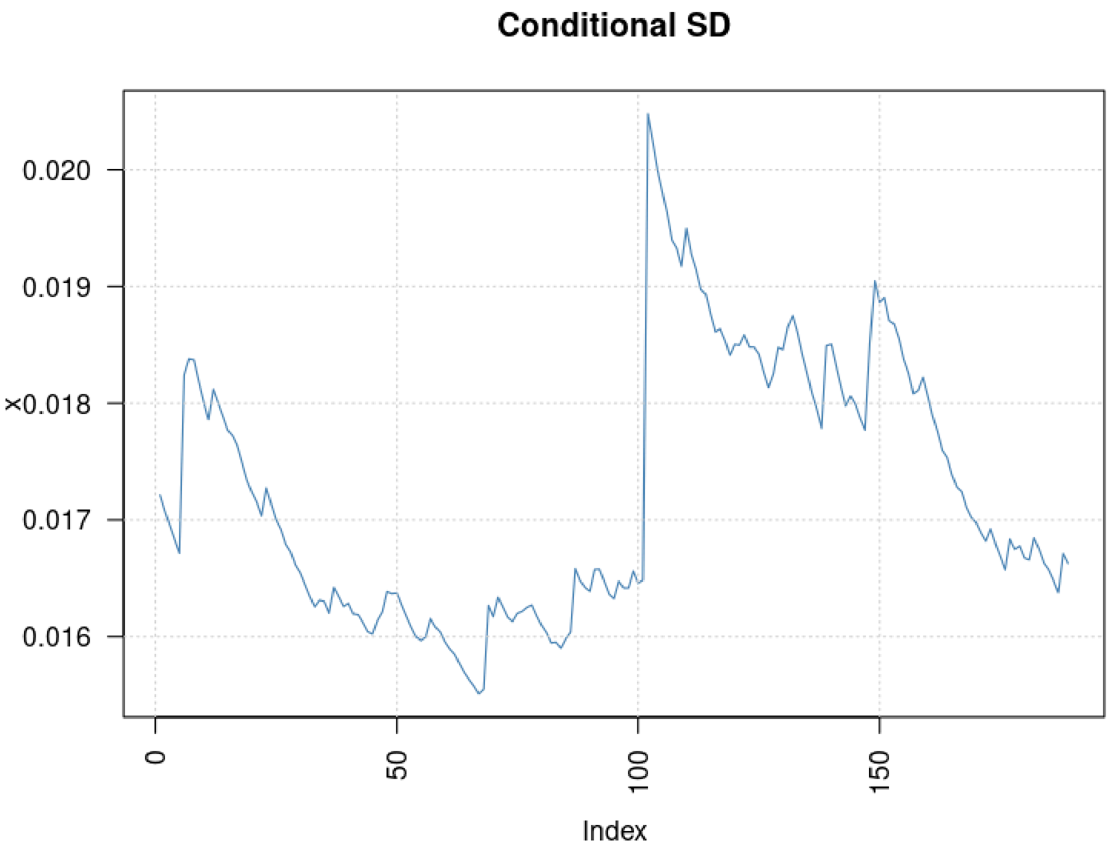
```
      Box-Ljung test

data:  y5^2
X-squared = 7.4373, df = 12, p-value = 0.8274
```

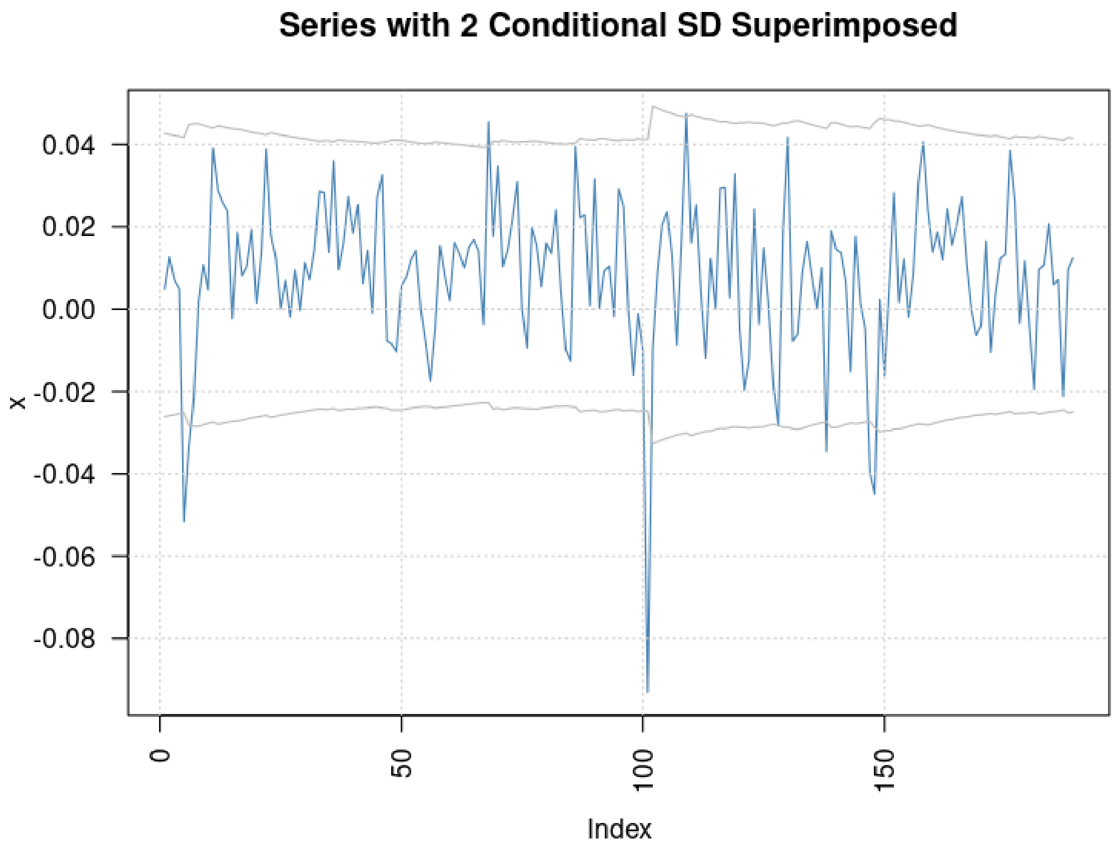
The ARCH model fitted to the data y5y5y5 with an ARMA(1,1) + GARCH(1,0) specification provides insights into the conditional mean and volatility of the time series. The model, specified using garchFit with a standardized Student's t-distribution for errors, estimates the mean parameter μ=0.0072\mu = 0.0072μ=0.0072, while the AR(1) and MA(1) coefficients are 0.2784 and -0.0168, respectively. The variance equation includes a constant (ω=0.0002542\omega = 0.0002542ω=0.0002542) and an ARCH(1) term (α1=0.1238\alpha_1 = 0.1238α1=0.1238), with the shape parameter of the distribution estimated at 5.3452, indicating heavy tails. Based on the t-values and p-values, the intercept and variance constant (ω\omegaω) are highly significant, while other parameters like the AR(1) and ARCH(1) terms are not statistically significant.
Diagnostics show robust model performance with a log-likelihood of 511.0237 and significant information criteria values (AIC = -5.3442, BIC = -5.2413), indicating a good model fit. Residuals analysis highlights normality issues, with significant Jarque-Bera and Shapiro-Wilk statistics, but the Ljung-Box tests for autocorrelation reveal no significant patterns, suggesting well-specified residuals in both raw and squared terms. The LM ARCH test result also supports no remaining ARCH effects, indicating that the model sufficiently captures the conditional heteroscedasticity in the data.

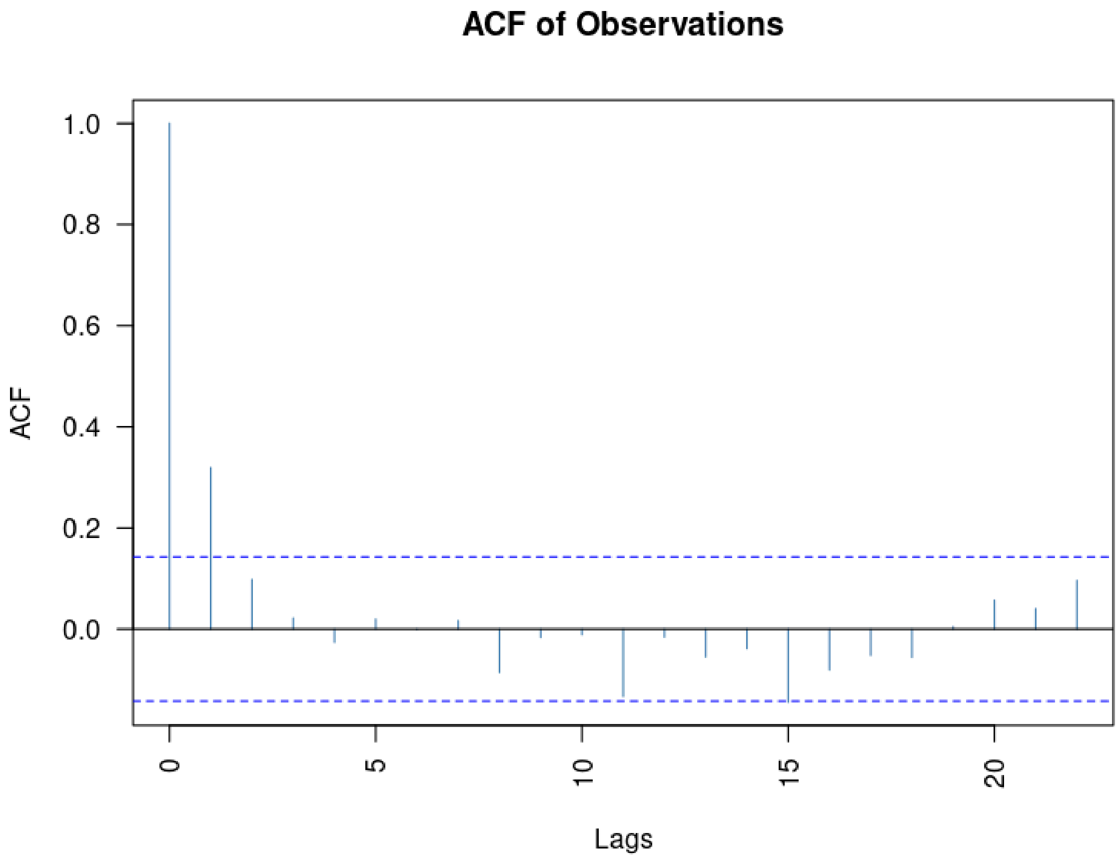
```
Title:
 GARCH Modelling

Call:
 garchFit(formula = ~arma(1, 1) + garch(1, 0), data = y5, cond.dist = 
"std", 
    trace = FALSE, xreg = x_reg)

Mean and Variance Equation:
 data ~ arma(1, 1) + garch(1, 0)
<environment: 0x565f9343bbd8>
 [data = y5]

Conditional Distribution:
 std

Coefficient(s):
        mu          ar1          ma1        omega       alpha1        shape  
 0.0071632    0.2783581   -0.0167956    0.0002542    0.1238481    5.3452209  

Std. Errors:
 based on Hessian 

Error Analysis:
         Estimate  Std. Error  t value Pr(>|t|)    
mu      0.0071632   0.0021340    3.357 0.000789 ***
ar1     0.2783581   0.1938401    1.436 0.150997    
ma1    -0.0167956   0.2012770   -0.083 0.933498    
omega   0.0002542   0.0000512    4.965 6.88e-07 ***
alpha1  0.1238481   0.1481131    0.836 0.403058    
shape   5.3452209   1.8512985    2.887 0.003886 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

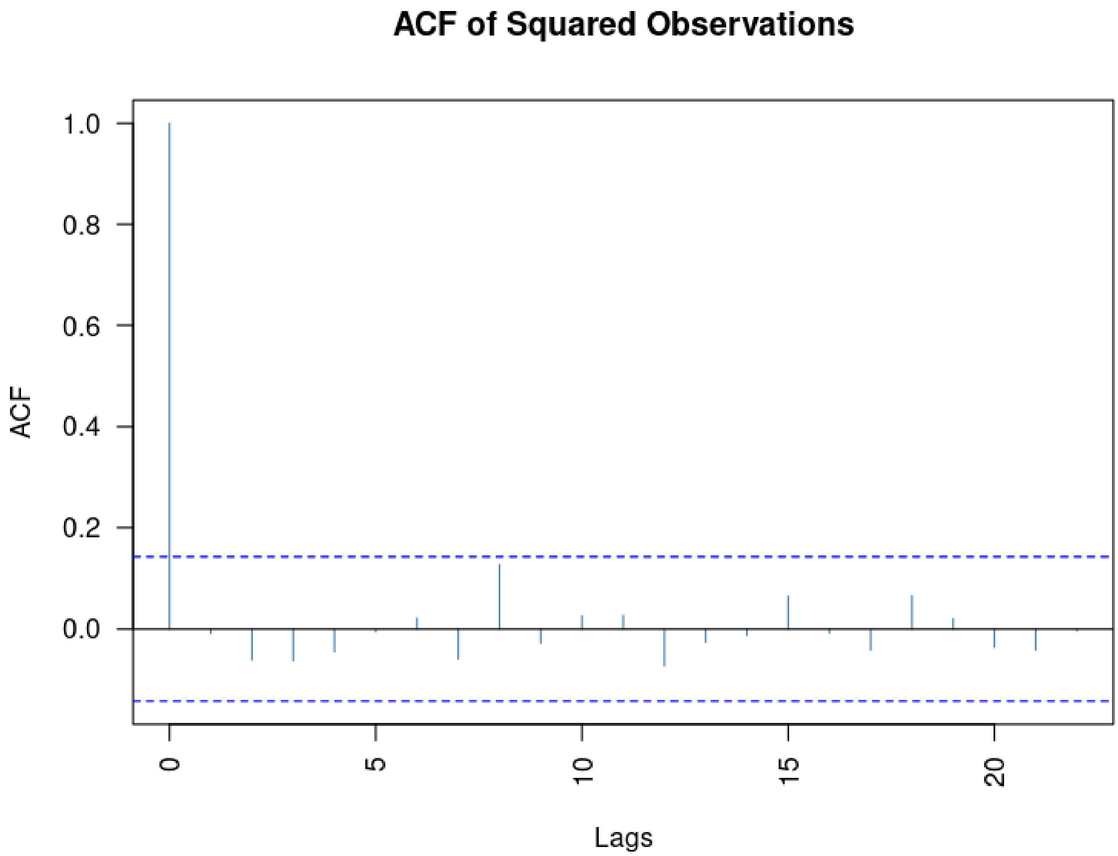
```
Log Likelihood:
 511.0237    normalized:  2.703829

Description:
 Thu Nov 14 01:46:10 2024 by user: jason


Standardised Residuals Tests:
                                  Statistic      p-Value
 Jarque-Bera Test   R    Chi^2  232.3243915 0.000000e+00
 Shapiro-Wilk Test  R    W        0.9371036 2.532369e-07
 Ljung-Box Test     R    Q(10)    3.3906384 9.706818e-01
 Ljung-Box Test     R    Q(15)   11.8273759 6.920456e-01
 Ljung-Box Test     R    Q(20)   13.1225492 8.720643e-01
 Ljung-Box Test     R^2  Q(10)    2.3687282 9.926537e-01
 Ljung-Box Test     R^2  Q(15)    5.1929370 9.903630e-01
 Ljung-Box Test     R^2  Q(20)    5.4968355 9.994263e-01
 LM Arch Test       R    TR^2     2.1406719 9.991579e-01

Information Criterion Statistics:
      AIC       BIC       SIC      HQIC
-5.344166 -5.241254 -5.346101 -5.302474
```

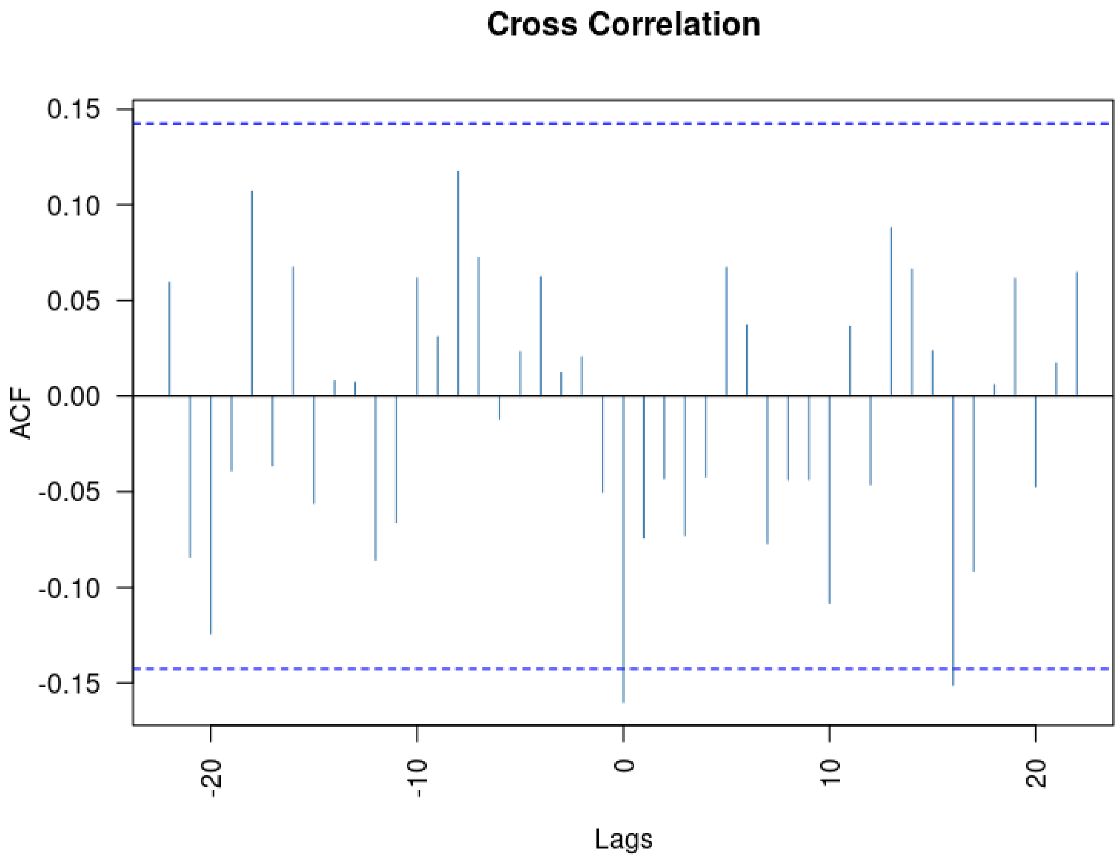

Time Series
0.04
0.02
0.00
-0.02
-0.04
-0.06
-0.08
x
0
50
100
150
Index

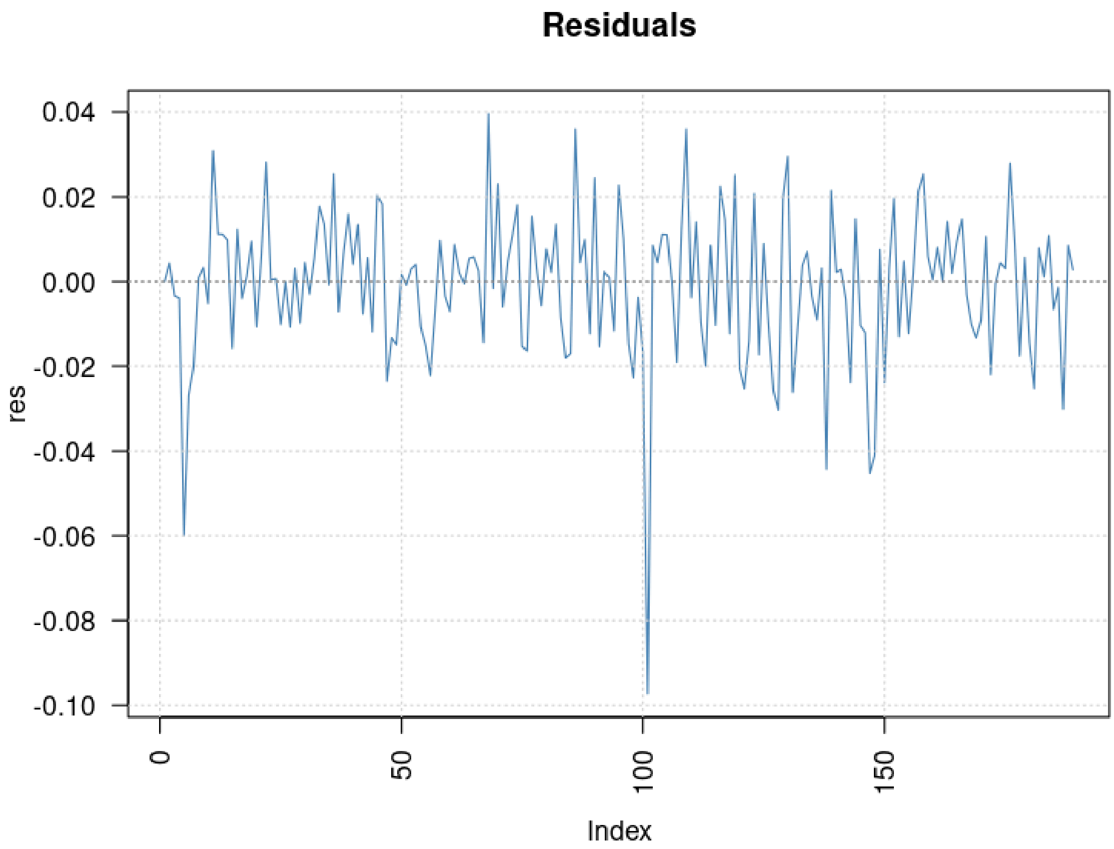

Conditional SD
0.020
0.019
0.018
0.017
0.016
x
0
50
100
150
Index


Series with 2 Conditional SD Superimposed
0.04
0.02
0.00
-0.02
-0.04
-0.06
-0.08
x
0
50
100
150
Index


ACF of Observations
ACF
1.0
0.8
0.6
0.4
0.2
0.0
0
5
10
15
20
Lags


ACF of Squared Observations
1.0
0.8
0.6
0.4
0.2
0.0
ACF
0
5
10
15
20
Lags


Cross Correlation
0.15
0.10
0.05
0.00
-0.05
-0.10
-0.15
ACF
-20
-10
0
10
20
Lags



The GARCH (Generalized Autoregressive Conditional Heteroskedasticity) model is employed here to analyze time series data with volatility clustering, allowing for conditional heteroscedasticity. The model specified in the garchFit function incorporates an ARMA(1, 1) process for the mean equation and a GARCH(1, 1) process for the variance equation, which is a commonly used structure to model financial time series with autoregressive and moving average components for the mean, and lagged variance for the volatility. The conditional distribution is assumed to be "std" (student's t-distribution) to account for heavy tails, and the regression includes external regressors (x_reg). The estimated coefficients include the intercept (mu), AR and MA terms (ar1, ma1), volatility parameters (omega, alpha1, beta1), and the shape parameter for the conditional distribution. Statistical significance of the coefficients is analyzed, with the beta1 parameter showing strong significance (p-value < 2e-16), indicating a high degree of persistence in the volatility process. In contrast, the ma1 and omega parameters are not significant. The model's goodness of fit is evaluated using various diagnostic tests, including the Jarque-Bera test and Shapiro-Wilk test, both of which reject the null hypothesis of normality, while the Ljung-Box tests indicate no autocorrelation in the standardized residuals. The Log Likelihood of 510.8883 suggests a reasonably good fit, and the information criteria (AIC, BIC, SIC, HQIC) provide further evidence of the model's adequacy. Overall, the results demonstrate that the GARCH model effectively captures the conditional variance in the data, with strong persistence of volatility over time.

```
Title:
 GARCH Modelling

Call:
 garchFit(formula = ~arma(1, 1) + garch(1, 1), data = y5, cond.dist =
"std",
    trace = FALSE, xreg = x_reg)

Mean and Variance Equation:
 data ~ arma(1, 1) + garch(1, 1)
<environment: 0x565f95d8d6e8>
 [data = y5]

Conditional Distribution:
 std

Coefficient(s):
        mu         ar1          ma1        omega       alpha1        beta1
shape
6.9840e-03  2.6627e-01  5.7345e-03  9.3387e-06  1.5964e-02  9.5255e-01
5.0219e+00

Std. Errors:
 based on Hessian

Error Analysis:
        Estimate  Std. Error  t value Pr(>|t|)
mu     6.984e-03   2.173e-03    3.213  0.00131 **
ar1    2.663e-01   1.998e-01    1.332  0.18274
ma1    5.735e-03   2.042e-01    0.028  0.97759
omega  9.339e-06   1.112e-05    0.839  0.40124
alpha1 1.596e-02   1.965e-02    0.812  0.41651
beta1  9.525e-01   4.455e-02   21.384  < 2e-16 ***
shape  5.022e+00   1.843e+00    2.725  0.00643 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Log Likelihood:
 510.8883    normalized:  2.703113

Description:
 Thu Nov 14 01:48:33 2024 by user: jason


Standardised Residuals Tests:
                                 Statistic      p-Value
 Jarque-Bera Test   R    Chi^2  308.138661 0.000000e+00
 Shapiro-Wilk Test  R    W        0.932362 1.048118e-07
 Ljung-Box Test     R    Q(10)    3.336391 9.723638e-01
 Ljung-Box Test     R    Q(15)   13.505436 5.633191e-01
 Ljung-Box Test     R    Q(20)   14.677591 7.945491e-01
 Ljung-Box Test     R^2  Q(10)    1.653779 9.983712e-01
 Ljung-Box Test     R^2  Q(15)    4.591055 9.950591e-01
 Ljung-Box Test     R^2  Q(20)    4.730739 9.998203e-01
 LM Arch Test       R    TR^2     1.427312 9.999001e-01

Information Criterion Statistics:
      AIC       BIC       SIC      HQIC
-5.332151 -5.212086 -5.334766 -5.283510
```

Question 3

- Based on the AIC (Akaike Information Criterion) and Log Likelihood values, the best model to select is the AR(2) model. It has the lowest AIC of -1232.94, which indicates the best balance between model fit and complexity among the three options. Additionally, its Log Likelihood value of 629.47 is the highest, further supporting its superior performance. The ARMA(1,1) model, with an AIC of -1232.81 and a Log Likelihood of 629.41, is very close in performance to the AR(2) model but slightly less optimal. The MA(1) model, with an AIC of -1229.49 and a lower Log Likelihood of 626.74, is not as good as the AR(2) or ARMA(1,1) models, making it the least favorable option. **Therefore, the AR(2) model is the preferred choice for this dataset.**
- Based on the results from the stepwise selection process, the final model for predicting y5y5y5 includes the following predictors: y1y1y1, y6y6y6, y7y7y7, y9y9y9, x1x1x1, x5x5x5, x6x6x6, and x11x11x11. Among these variables, y1y1y1, y7y7y7, y9y9y9, x1x1x1, x5x5x5, and x6x6x6 are statistically significant (p-values < 0.05), suggesting that they have a meaningful relationship with y5y5y5. While y6y6y6 and x11x11x11 are not statistically significant at the 0.05 level, they are retained in the model due to the stepwise procedure. The model has a high R2R^2R2 of 0.8909, indicating a strong fit. **Thus, the chosen model is robust for prediction, though it could be refined by potentially removing non-significant predictors such as y6y6y6 and x11x11x11.**
- Finally, due to heteroskedasticity suffering of the previous model, when investigating on ARCH and GARCH models, out findings are the following ones: The second model includes a GARCH(1,1) specification, which allows for more flexibility in modeling the conditional variance compared to the GARCH(1,0) model. The GARCH(1,1) model has a higher log-likelihood value (-5.332151 vs -5.344166) and lower information criteria (AIC, BIC, SIC, HQIC) compared to the GARCH(1,0) model, indicating it is a better fit for the data. Additionally, the estimated coefficients for the GARCH(1,1) model are more statistically significant, with the GARCH term (beta1) being highly significant at p < 0.00001. The diagnostic tests also show similar or slightly better performance for the GARCH(1,1) model, with the Ljung-Box and ARCH-LM tests suggesting no significant remaining autocorrelation or ARCH effects in the standardized residuals. **Therefore, based on the information provided, the GARCH(1,1) model appears to be the better choice for modeling the given time series data.**