MTH783P - Time Series Analysis for Business - Final Project

Can Guvener

2025-05-06

Abstract

This report analyzes U.S. quarterly economic indicators to forecast personal consumption using time series models. The dataset spans from 1970 Q1 to 2016 Q3, is split into training (1970 Q1 to 2014 Q4) and test (2015 Q1 to 2016 Q3) periods. The data is explored and a regression with ARIMA errors is selected after evaluation. The final model's residuals and forecast accuracy are assessed and discussed.

Question (a): Load and Split the Data

The dataset used in this project is `uschange`, is included in the R package fpp2. It includes quarterly percentage changes for five economic variables in the US. These are consumption, income, production, savings, and unemployment. The dataset spans from 1970 Q1 to 2016 Q3.

The objective is to forecast Consumption during the last 7 quarters (2015 Q1 to 2016 Q3). To do this, the dataset is split into training and test sets, via window() function:

- **Training set**: from 1970 Q1 to 2014 Q4
- **Test set**: from 2015 Q1 to 2016 Q3

Question (b): Exploratory Data Analysis

Exploratory analysis is conducted to assess the characteristics of the variables, and their relationships with the target variable, consumption.

Time Series Behavior

Consumption:

The Consumption series shows limited fluctuations around the mean of 0.75, with no obvious trend. There is some volatility in the early parts. No visible seasonality is present. The behavior suggests that the series is stationary in variance.

Income:

Income shows a pattern similar to Consumption. Fluctuations are limited. Slightly higher volatility is present in the early years. No visible seasonality is present.

Production:

Production demonstrates significant volatility in the early period. After around 1985, the fluctuations become smaller and more stable. There is no clear trend or seasonality.

Savings:

Savings is more volatile compared to the other variables. Between 2000 and 2010, there are large spikes and dips with high frequency. There is no obvious trend or seasonality visible.

Unemployment:

Unemployment shows downward trend from 1980 to 2000, followed by some quick increases around 2008. The trend is volatile overall with frequent increases and drops. Some periods are relatively flat and some periods have very high variation. There is no visible seasonality present.

To sum up, the statistics for the training set show that all variables are approximately centered around zero.

- Consumption, Income, and Production exhibit moderate variation, with ranges roughly between –2% and +2%, -5% and 5%, -7.5% and 5% respectively.

Savings has a higher span, which can mean the presence of outliers or extreme fluctuations.

Unemployment has the smallest range, with values between –0.9% and +1.4%, which shows relatively low variation.

Stationarity Assessment

All five variables passed stationarity tests, which indicates that they are suitable for modeling. The Augmented Dickey-Fuller (ADF) test rejected the null hypothesis of a unit root for each series (p-value < 0.01),

which indicates stationarity. The KPSS test returned non-significant results (p-value around 0.1), which suggests no strong evidence against stationarity. These results confirm that the series can be weakly stationary, as its first differenced type suggests.

Autocorrelation Structures

Consumption shows significant autocorrelation that tails off after lag 3, which suggests short-term dependence. AR(3) structure.

Income remains entirely within the confidence bounds. This suggests no significant autocorrelation and behaves similarly to white noise.

Production displays notable autocorrelation that tails off after lag 3, which suggests short-term dependence. AR(3) structure.

Savings shows a cut-off after lag 1, suggesting an MA(1) structure.

Unemployment displays strong autocorrelation that tails off after lag 3. This indicates short-term dependence. AR(3) structure.

Interrelationships Between Variables

Correlation and scatterplot matrix analysis demonstrate that:

- Consumption is positively correlated with both Income and Production, which complies with the economic theory.
- Unemployment is negatively correlated with Consumption, as higher unemployment prevents spending.
- The relationship between Savings and Consumption is negative, as people usually choose between one over the other.

Question (c): Model Fitting and Forecasting

Model Selection and Rationale

Several modeling approaches were considered:

- Univariate ARIMA model for Consumption alone.
- Multiple linear regression model using the remaining variables as predictors.
- Regression with ARIMA errors, combining the benefits of explanatory modeling with autoregressive error correction.

The univariate ARIMA model was quickly ruled out due to limited predictive power; it fails to involve valuable information from leading indicators. The multiple regression model improved the explanatory strength but left autocorrelation in residuals unaddressed.

The most effective model was a regression model with ARIMA errors, implemented via the auto.arima() function. This approach allows for the inclusion of relevant macroeconomic predictors, while modeling the error structure as an ARIMA process.

The selected model for Consumption was:

ARIMA(3,1,0)(1,0,0)[4] with 4 remaining exogenous variables.

The seasonal component at lag 4 is selected because of the quarterly data structure.

Model Assumptions and Residual Diagnostics

The regression with ARIMA errors assumes:

- Stationarity of the time series (already achieved via differencing)
- Linearity in the relationship between Consumption and independent variables
- Independence and white-noise residuals: residuals should show no autocorrelation, constant variance, and normal-like distribution

Diagnostic checks were conducted using residual plots, autocorrelation function, and Ljung-Box test:

- Residuals appeared random and centered around zero, with no visible trend or heteroskedasticity.
- The ACF plot showed only mild autocorrelation at lag 4, understandable in quarterly data.
- The Ljung-Box test at lag 8 produced a p-value of 0.0009, suggesting mild autocorrelation that shouldn't impact the validity of the model.

While it is not perfectly white noise, the residuals seem sufficient.

Forecast Generation and Interpretation

Using the fitted model, forecasts were generated for the seven quarters from 2015 Q1 to 2016 Q3, with the corresponding exogenous variables provided as inputs. The resulting forecast series closely tracked the actual Consumption values.

Key observations:

- Forecast intervals widen over time, reflecting growing uncertainty, yet the model successfully captures turning points in the Consumption trend.
- Out of seven quarters, five actual values fell within the 95% prediction intervals, indicating well-calibrated uncertainty estimates.
- Performance metrics were satisfactory with the values of **MAE = 0.074**, **RMSE = 0.097**, and **MAPE = 10.5%**.

Overall, the model demonstrates sufficient forecasting capability.

Strengths, Limitations, and Potential Improvements

Strengths:

The model uses both economic theory (through predictive variables), and time series analysis. All predictors are reliable and valid within the economic theory. The modeling process also included residual validation, which contributes to the validity of the report.

Limitations:

Residual autocorrelation might suggest possible remaining structure. The model assumes linear relationships and fixed coefficients, which can be too restrictive. Extenuating circumstances such as the 2008 crisis are not controlled for in the model. The forecast is only for a short amount of time, which can limit generalizability.

Possible Improvements:

Extreme economic conditions like crises can be flagged via dummy variables for the regression models. Different independent variables can be added through trial and error.

Recall the maximum length of the report: 3 pages, excluding the titlepage and the "R code" sections.

[END of the REPORT]

R code

Q a)

```
# loading the library
library(fpp2)
```

```
## Warning: package 'fpp2' was built under R version 4.4.3
```

```
## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo
```

```
## -- Attaching packages --------------------------------------------- fpp2 2.5 --
```

```
## v ggplot2   3.5.1      v fma       2.5
## v forecast  8.23.0     v expsmooth 2.3
```

```
## Warning: package 'fma' was built under R version 4.4.3
```

```
## Warning: package 'expsmooth' was built under R version 4.4.3
```

```
##
```

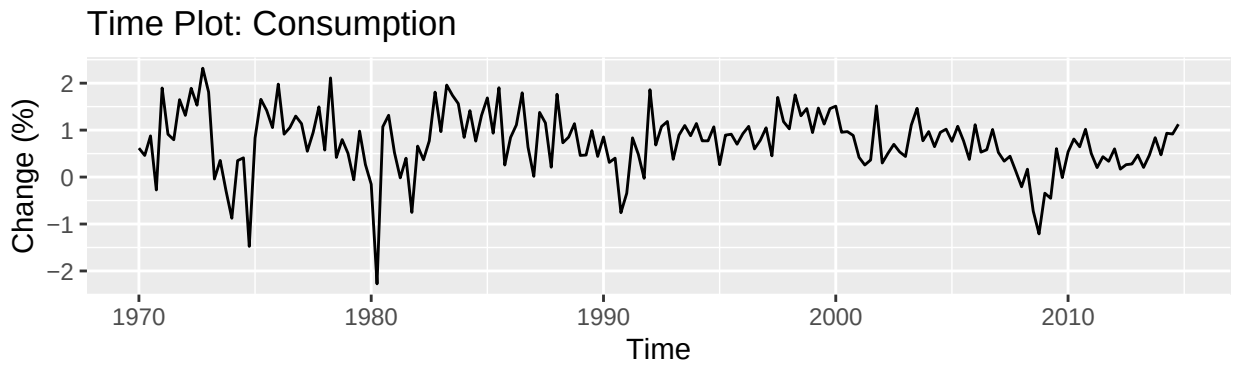
```
# Loading the dataset
data("uschange")

# Splitting into training and test sets
train <- window(uschange, end = c(2014, 4))
test <- window(uschange, start = c(2015, 1))
```

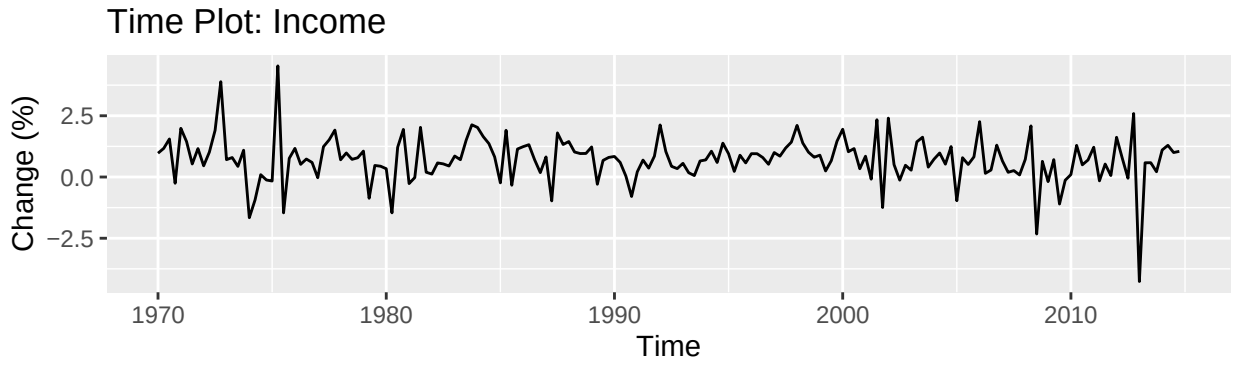
Q b)

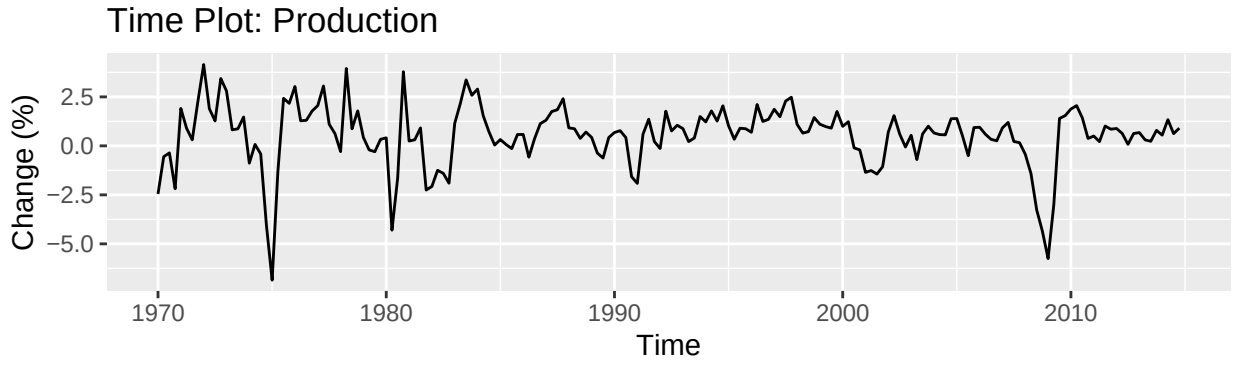
```
# General info of the dataset
summary(train)
```

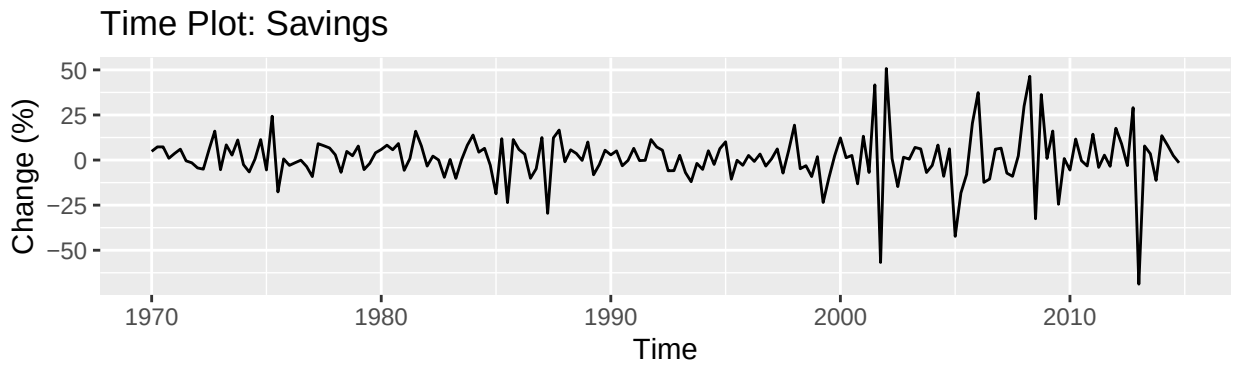
```
##    Consumption          Income          Production         Savings
##  Min.   :-2.2741   Min.   :-4.2652   Min.   :-6.8510   Min.   :-68.788
##  1st Qu.: 0.4159   1st Qu.: 0.2833   1st Qu.: 0.1429   1st Qu.: -4.820
##  Median : 0.7888   Median : 0.7232   Median : 0.6979   Median :  1.133
##  Mean   : 0.7493   Mean   : 0.7185   Mean   : 0.5377   Mean   :  1.215
##  3rd Qu.: 1.1083   3rd Qu.: 1.1727   3rd Qu.: 1.3420   3rd Qu.:  7.065
##  Max.   : 2.3183   Max.   : 4.5365   Max.   : 4.1496   Max.   : 50.758
##    Unemployment
##  Min.   :-0.90000
##  1st Qu.:-0.20000
##  Median : 0.00000
##  Mean   : 0.01167
##  3rd Qu.: 0.10000
##  Max.   : 1.40000
```

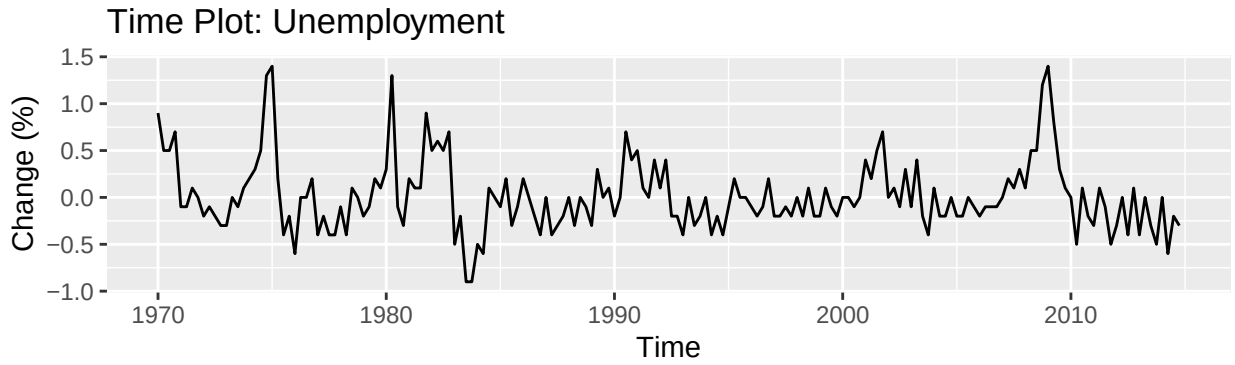
```r
# generating time plots for each variable

for (var in colnames(train)) {
  autoplot(train[, var]) +
    ggtitle(paste("Time Plot:", var)) +
    xlab("Time") + ylab("Change (%)") -> p
  print(p)
    }
```





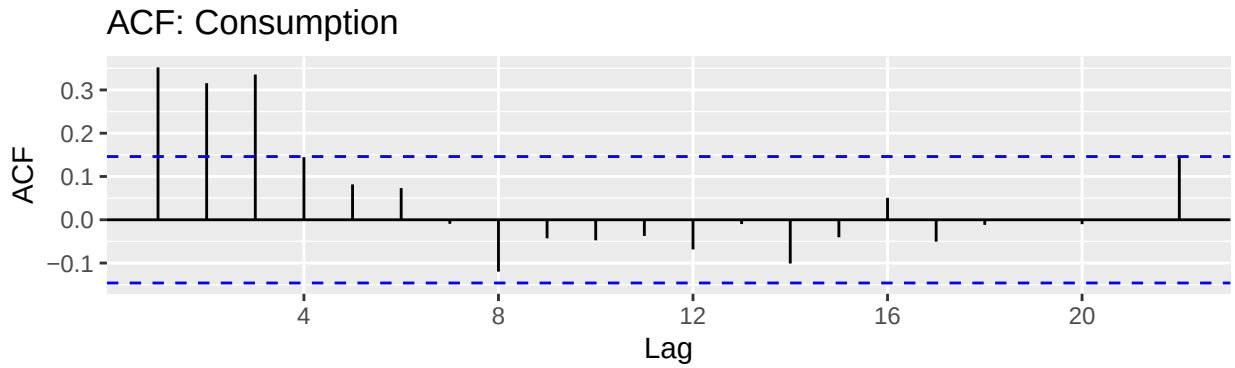




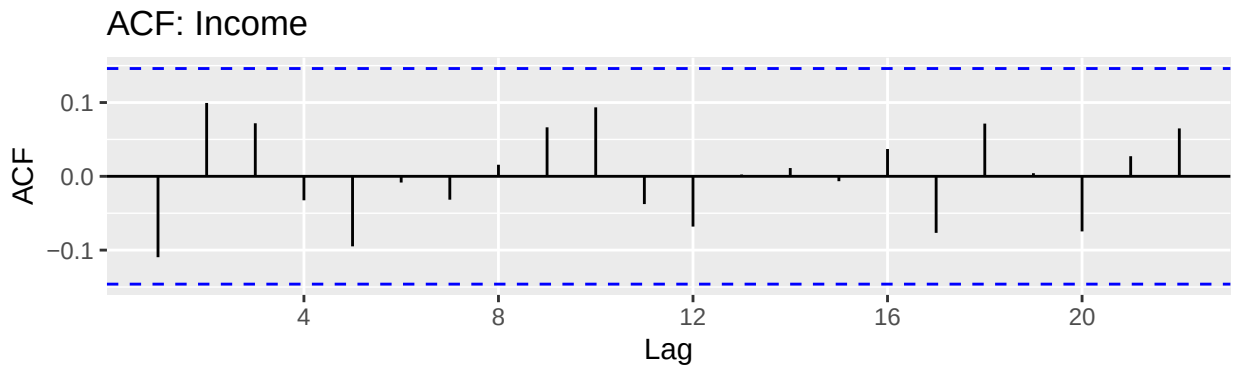



```
# generating ACF plots

library(ggplot2)
lapply(colnames(train), function(var) ggAcf(train[, var]) + ggtitle(paste("ACF:", var)))
```

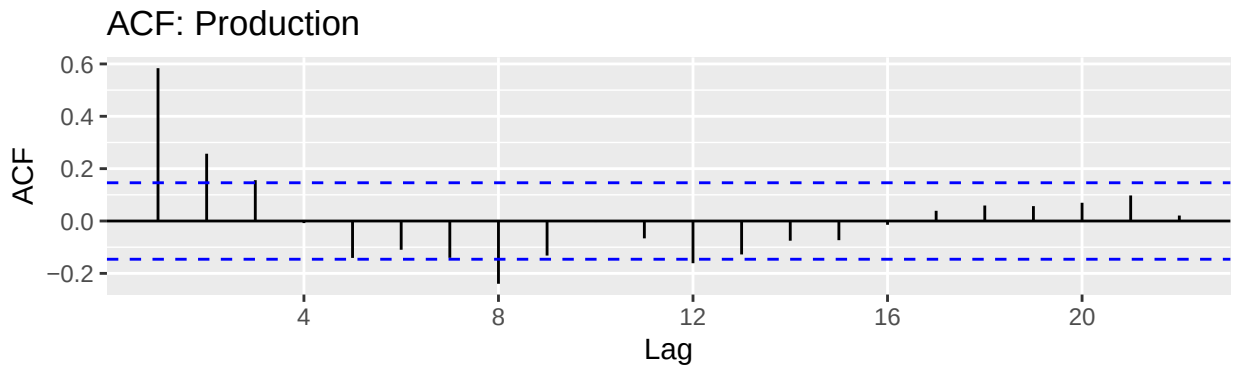
```
## [[1]]
```



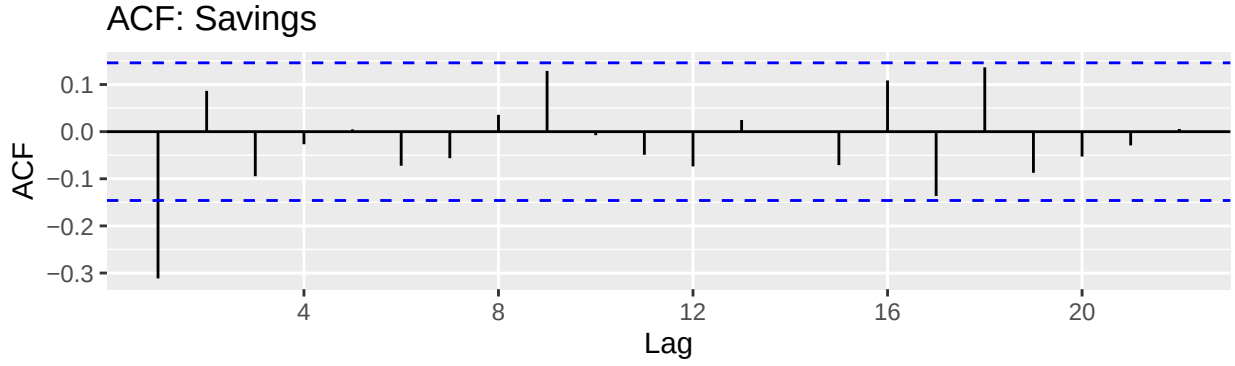

```
##
## [[2]]
```


ACF: Income
ACF
0.1
0.0
−0.1
4
8
12
16
20
Lag

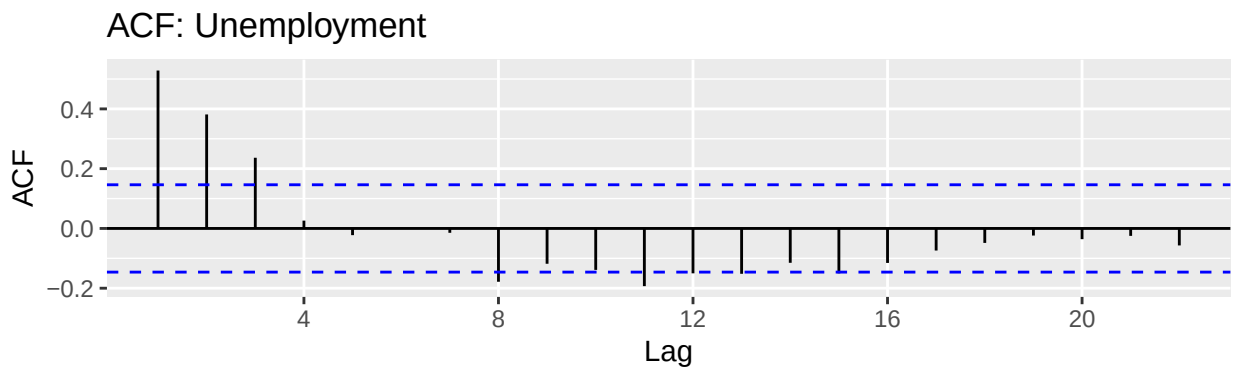
```
##
## [[3]]
```


ACF: Production
ACF
0.6
0.4
0.2
0.0
−0.2
4
8
12
16
20
Lag

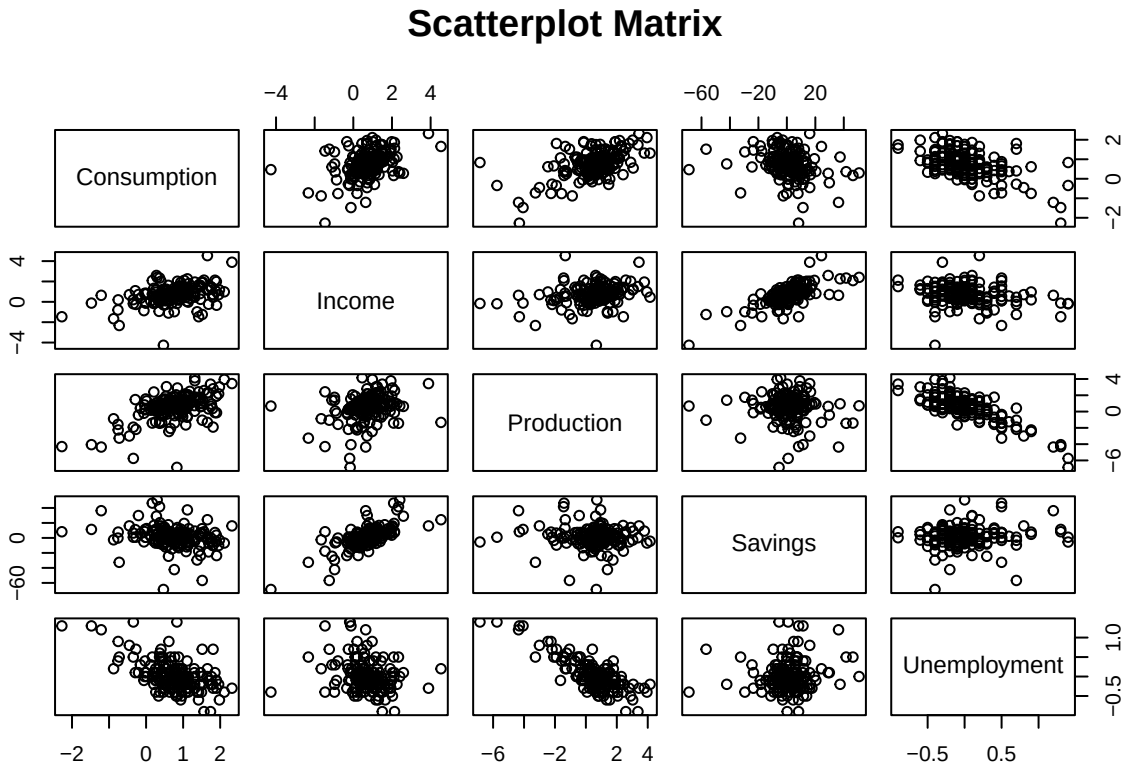
```
##
## [[4]]
```


ACF: Savings
ACF
0.1
0.0
−0.1
−0.2
−0.3
4
8
12
16
20
Lag

```
##
## [[5]]
```




```
# Correlation matrix with every variable
pairs(train, main="Scatterplot Matrix")
```




```
# Stationarity tests for each variable
library(tseries)
adf.test(train[, "Consumption"])
```

```
## Warning in adf.test(train[, "Consumption"]): p-value smaller than printed
## p-value

##
##  Augmented Dickey-Fuller Test
```

```
##
## data:  train[, "Consumption"]
## Dickey-Fuller = -4.4219, Lag order = 5, p-value = 0.01
## alternative hypothesis: stationary
```

```
kpss.test(train[, "Consumption"])
```

```
## Warning in kpss.test(train[, "Consumption"]): p-value greater than printed
## p-value
```

```
##
##  KPSS Test for Level Stationarity
##
## data:  train[, "Consumption"]
## KPSS Level = 0.27652, Truncation lag parameter = 4, p-value = 0.1
```

```
adf.test(train[, "Income"])
```

```
## Warning in adf.test(train[, "Income"]): p-value smaller than printed p-value
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  train[, "Income"]
## Dickey-Fuller = -6.0046, Lag order = 5, p-value = 0.01
## alternative hypothesis: stationary
```

```
kpss.test(train[, "Income"])
```

```
## Warning in kpss.test(train[, "Income"]): p-value greater than printed p-value
```

```
##
##  KPSS Test for Level Stationarity
##
## data:  train[, "Income"]
## KPSS Level = 0.23245, Truncation lag parameter = 4, p-value = 0.1
```

```
adf.test(train[, "Production"])
```

```
## Warning in adf.test(train[, "Production"]): p-value smaller than printed
## p-value
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  train[, "Production"]
## Dickey-Fuller = -5.1571, Lag order = 5, p-value = 0.01
## alternative hypothesis: stationary
```

```r
kpss.test(train[, "Production"])
```

```
## Warning in kpss.test(train[, "Production"]): p-value greater than printed
## p-value

##
##  KPSS Test for Level Stationarity
##
## data:  train[, "Production"]
## KPSS Level = 0.07317, Truncation lag parameter = 4, p-value = 0.1
```

```r
adf.test(train[, "Savings"])
```

```
## Warning in adf.test(train[, "Savings"]): p-value smaller than printed p-value

##
##  Augmented Dickey-Fuller Test
##
## data:  train[, "Savings"]
## Dickey-Fuller = -6.7917, Lag order = 5, p-value = 0.01
## alternative hypothesis: stationary
```

```r
kpss.test(train[, "Savings"])
```

```
## Warning in kpss.test(train[, "Savings"]): p-value greater than printed p-value

##
##  KPSS Test for Level Stationarity
##
## data:  train[, "Savings"]
## KPSS Level = 0.060873, Truncation lag parameter = 4, p-value = 0.1
```

```r
adf.test(train[, "Unemployment"])
```

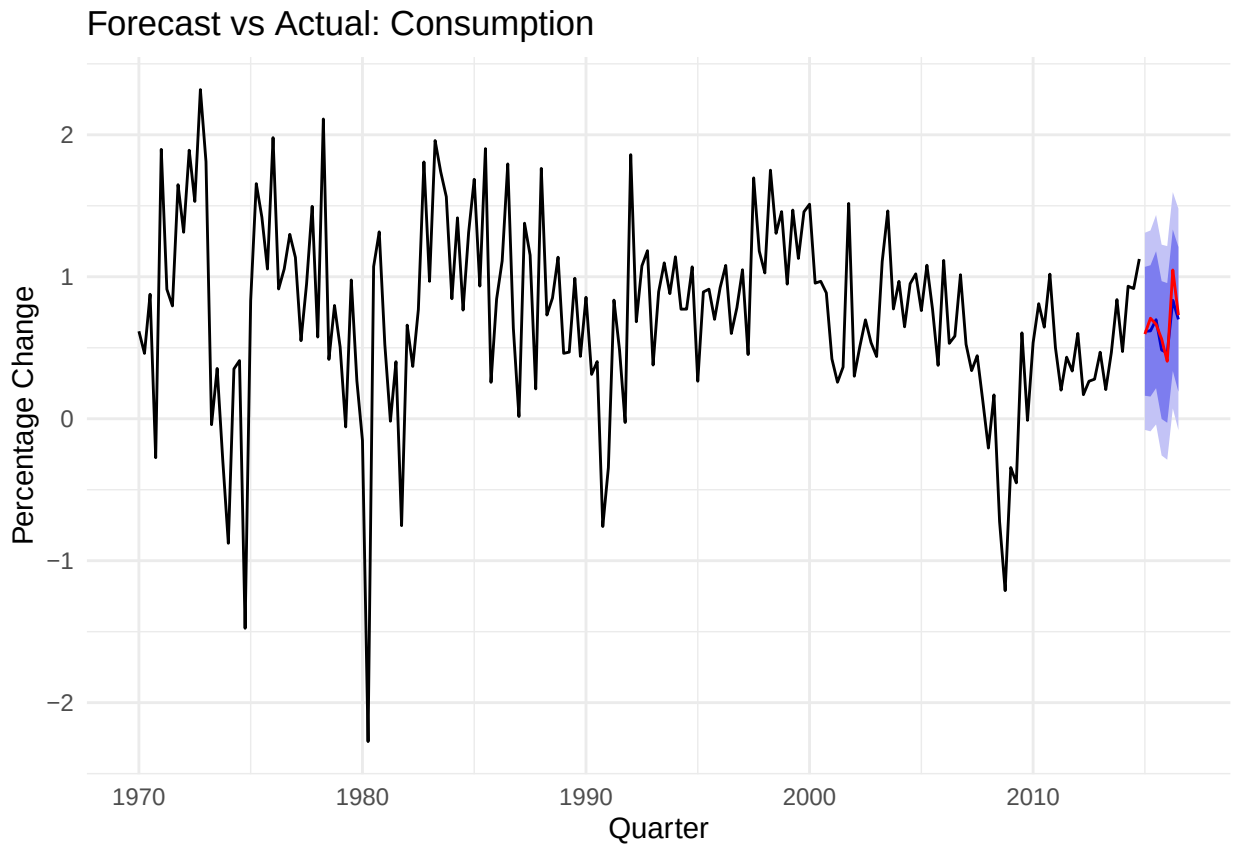
```
## Warning in adf.test(train[, "Unemployment"]): p-value smaller than printed
## p-value

##
##  Augmented Dickey-Fuller Test
##
## data:  train[, "Unemployment"]
## Dickey-Fuller = -4.3548, Lag order = 5, p-value = 0.01
## alternative hypothesis: stationary
```

```r
kpss.test(train[, "Unemployment"])
```

```
## Warning in kpss.test(train[, "Unemployment"]): p-value greater than printed
## p-value

##
##  KPSS Test for Level Stationarity
##
## data:  train[, "Unemployment"]
## KPSS Level = 0.095659, Truncation lag parameter = 4, p-value = 0.1
```

Q c)

```
library(forecast)

# Fitting and Assigning regression with ARIMA errors

fit <- auto.arima(train[, "Consumption"], xreg = train[, c("Income", "Production", "Savings", "Unemploy

# generating forecasts for the test period
fc <- forecast(fit, xreg = test[, c("Income", "Production", "Savings", "Unemployment")])

# Forecasting the plot

autoplot(fc, PI = TRUE) +
  autolayer(test[, "Consumption"], series = "Actual", color = "red") +
  ggtitle("Forecast vs Actual: Consumption") +
  ylab("Percentage Change") + xlab("Quarter") +
  theme_minimal()
```

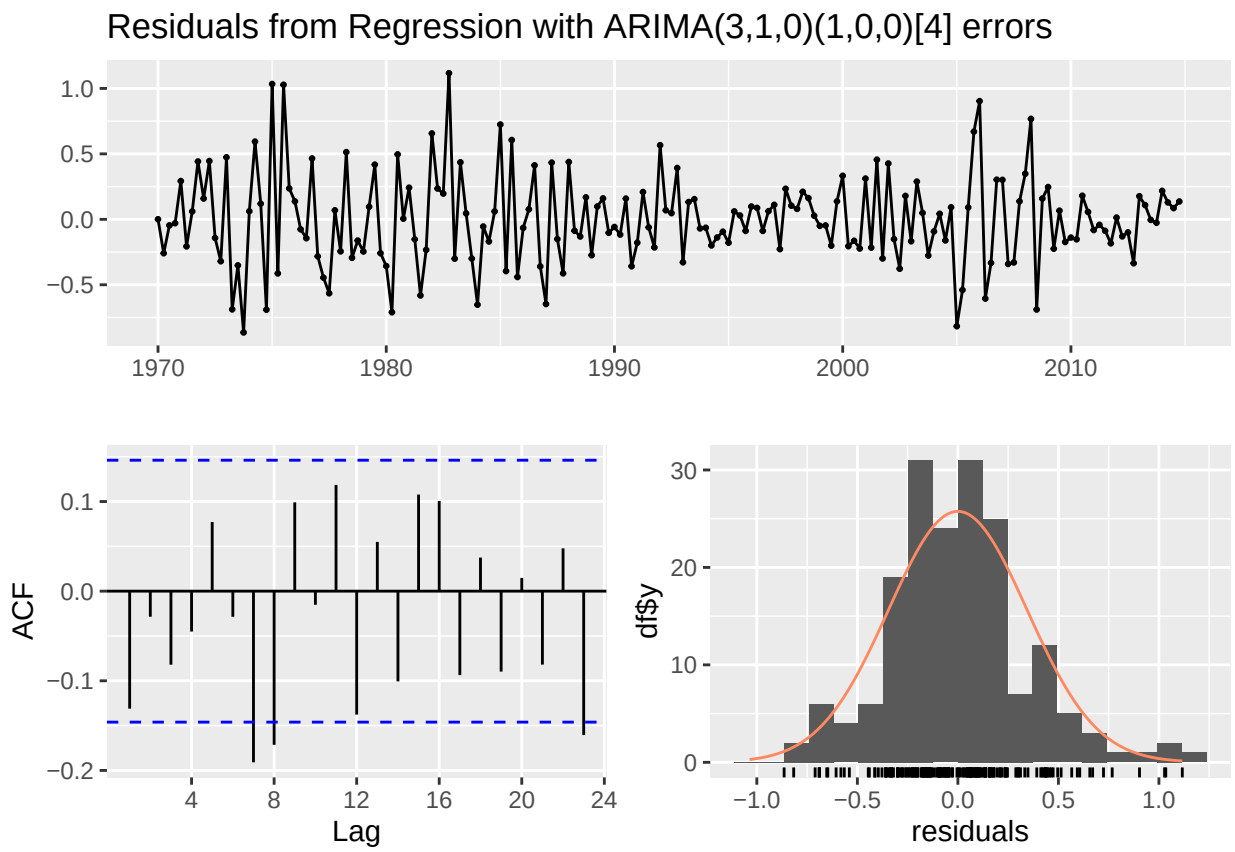



```
# printing all accuracy values
accuracy(fc, test[, "Consumption"])
```

```
##                       ME       RMSE        MAE      MPE     MAPE      MASE
## Training set -0.0004764331 0.34422720 0.26338713 6.178704 92.59988 0.4029595
```

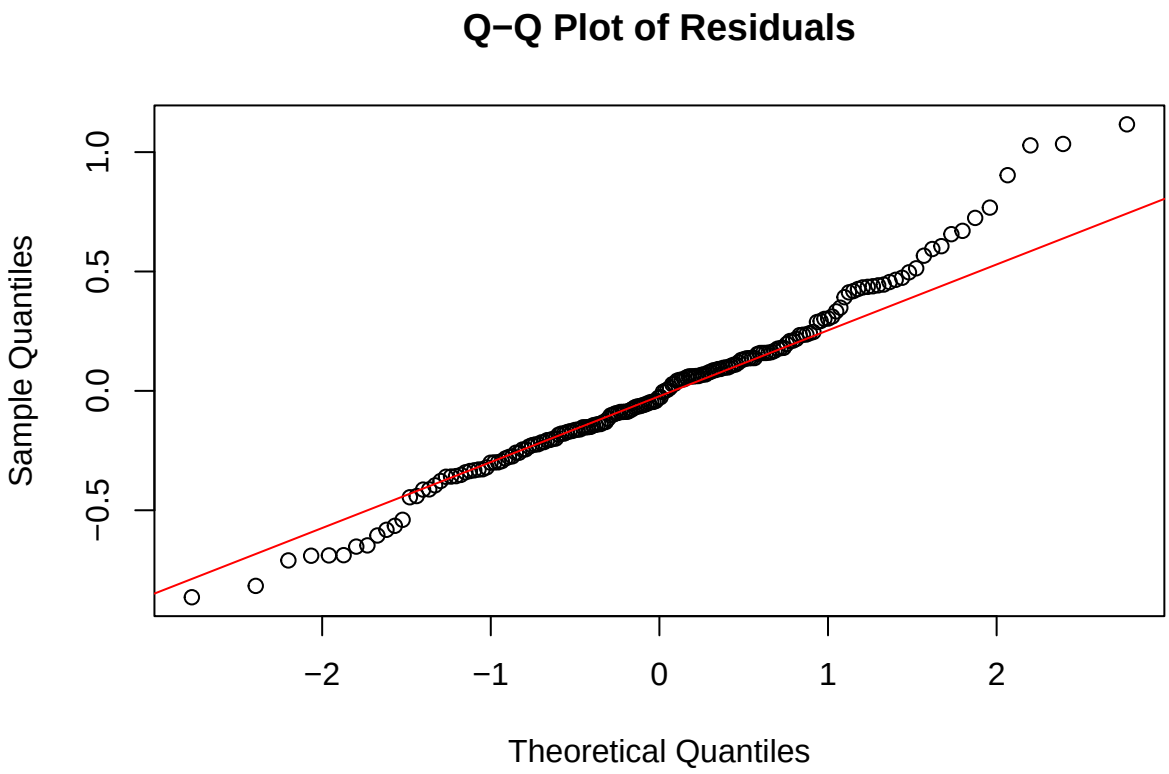
```
## Test set      0.0427189688 0.09658557 0.07448284 4.022171 10.52587 0.1139523
##                     ACF1 Theil's U
## Training set -0.1310828        NA
## Test set     -0.6131077 0.3446035
```

```
# generating residual graphs
checkresiduals(fit)
```




```
##
##  Ljung-Box test
##
## data:  Residuals from Regression with ARIMA(3,1,0)(1,0,0)[4] errors
## Q* = 18.69, df = 4, p-value = 0.0009042
##
## Model df: 4.   Total lags used: 8
```

```
# Base R Q-Q plot of residuals
residuals_arima <- residuals(fit)
qqnorm(residuals_arima, main = "Q-Q Plot of Residuals")
qqline(residuals_arima, col = "red")
```


Q−Q Plot of Residuals
Sample Quantiles
−0.5
0.0
0.5
1.0
−2
−1
0
1
2
Theoretical Quantiles