137 Project

Quynh Trinh,

2025-05-30

```r
# Load required libraries
library(tidyverse)      # For data manipulation and visualization
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```r
library(forecast)       # For Box-Cox transformation and time series analysis
```

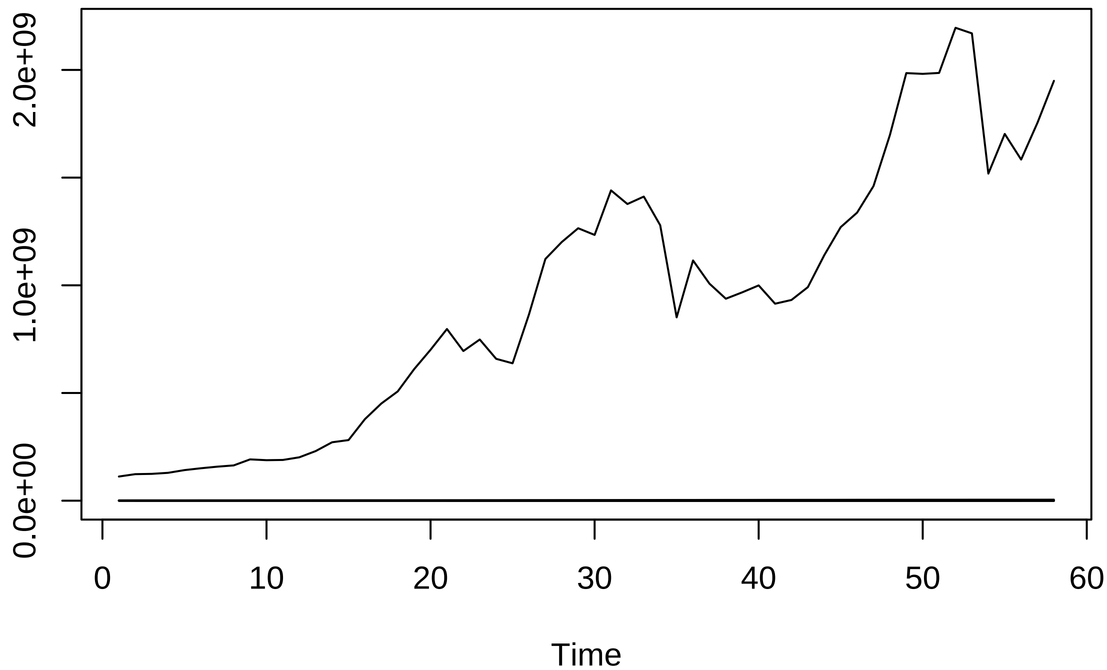
```
## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo
```

```r
library(tseries)        # For additional time series functions
library(ggplot2)        # For plotting
library(carData)
library(car)
```

```
##
## Attaching package: 'car'
##
## The following object is masked from 'package:dplyr':
##
##     recode
##
## The following object is masked from 'package:purrr':
##
##     some
```

Time Series Plot for the Data

```r
load("finalproject.Rdata")
ts.plot(finalPro_data)
```



Quick look:

- Upward trend
- Some fluctuations and short-term volatility
- Non-stationary

GDP Model

Summary GDP

```r
# 1. Diagnose model for GDP
model_gdp <- lm(GDP ~ Year + Growth + CPI + Imports + Exports + Population, data = finalPro_data)
summary(model_gdp)
```

```
##
## Call:
## lm(formula = GDP ~ Year + Growth + CPI + Imports + Exports +
##     Population, data = finalPro_data)
```

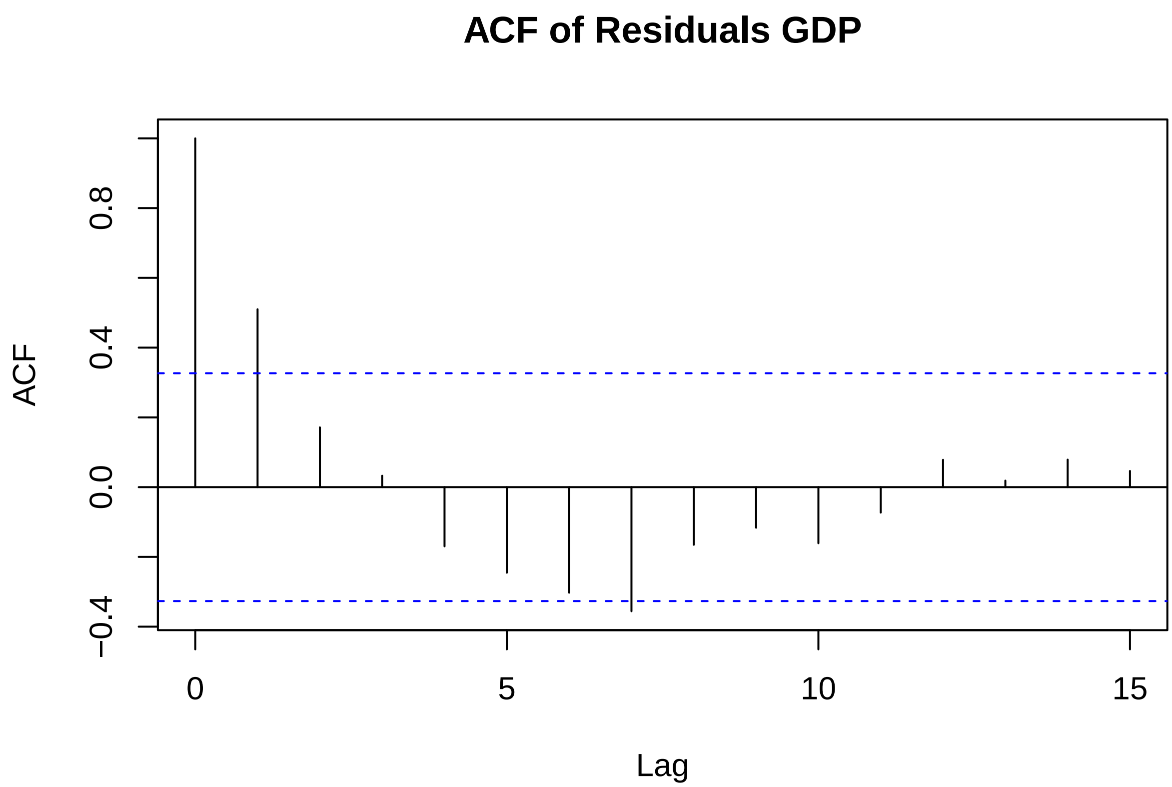
```
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -355647077 -123917222   18740212  114338509  328824034
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.704e+11  1.190e+11  -2.272   0.0307 *
## Year         1.394e+08  6.097e+07   2.286   0.0297 *
## Growth       6.543e+06  4.960e+06   1.319   0.1974
## CPI         -9.881e+06  4.121e+06  -2.398   0.0231 *
## Imports      1.590e+07  1.088e+07   1.461   0.1548
## Exports     -7.174e+07  1.240e+07  -5.784 2.89e-06 ***
## Population  -1.470e+03  7.621e+02  -1.928   0.0636 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 194300000 on 29 degrees of freedom
##   (22 observations deleted due to missingness)
## Multiple R-squared:  0.8316, Adjusted R-squared:  0.7967
## F-statistic: 23.87 on 6 and 29 DF,  p-value: 5.499e-10
```

- Significant variables: Year, Exports

Residuals Diagnostic

ACF and Ljung-Box

```r
#ACF GDP
acf(residuals(model_gdp), main="ACF of Residuals GDP")
```




```r
# L-jung-box
Box.test(residuals(model_gdp), lag = 20, type = "Ljung-Box")
```

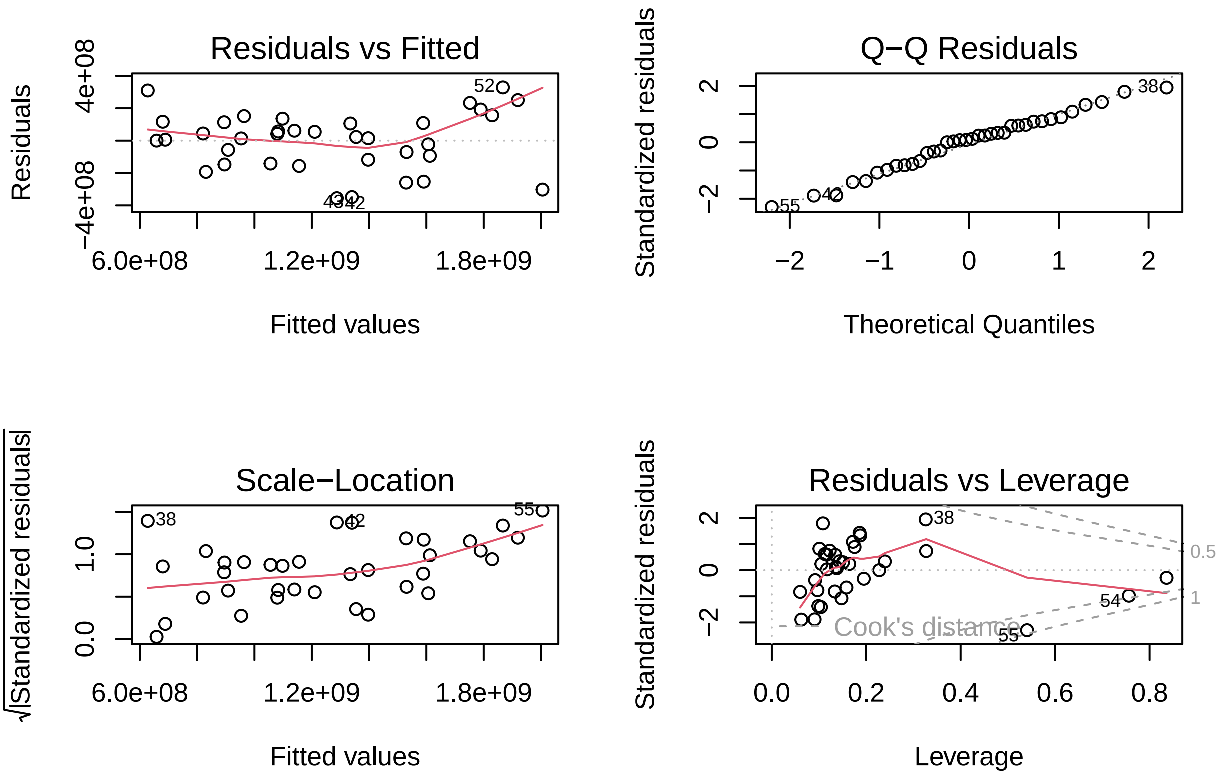
```
##
##  Box-Ljung test
##
## data:  residuals(model_gdp)
## X-squared = 32.013, df = 20, p-value = 0.04316
```

Quick look:

- Autocorrelation remains positive before approaching zero
- The pattern shows strong autocorrelation and persistence in the residuals
- Slow decay –> the data is non-stationary or trending series

Plot GDP

```r
# Diagnostic plots for GDP model
par(mfrow = c(2,2))
plot(model_gdp)
```




```r
# Shapiro-Wilks for GDP
ei_gdp = model_gdp$residuals
the.SWtest_gdp = shapiro.test(ei_gdp)
the.SWtest_gdp
```

```
##
##  Shapiro-Wilk normality test
##
## data:  ei_gdp
## W = 0.97508, p-value = 0.5794
```

```r
# Brown/Levene test
# Extract residuals from your model
res <- residuals(model_gdp)

# Make a grouping variable (e.g., you can split by median fitted value)
fit <- fitted(model_gdp)
group <- ifelse(fit > median(fit), "High", "Low")

# Levene's Test (Brown-Forsythe is a median-centered version of Levene's test)
leveneTest(res ~ group, center=median) # Brown-Forsythe test
```

```
## Warning in leveneTest.default(y = y, group = group, ...): group coerced to
## factor.
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value  Pr(>F)
## group  1  6.5555 0.01507 *
##       34
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Quick note:

- Residuals vs fitted: looks good, no obvious pattern suggests linearity
- QQ plot: residuals are close to the line, suggests approximately normally distributed
- Scale-Location: slightly upward trend, meaning the variance of errors is not constant
- Residuals vs Leverage: most points have low leverage so this is good
- Shapiro-Wilk test: W = 0.98882, p-value = 0.8709 –> met normality assumption
- Brown test: p-value = 0.002518, so constant variance assumption is not met

Recommendations:

- Transforming the dependent variable; considering Box-Cox or log transformation

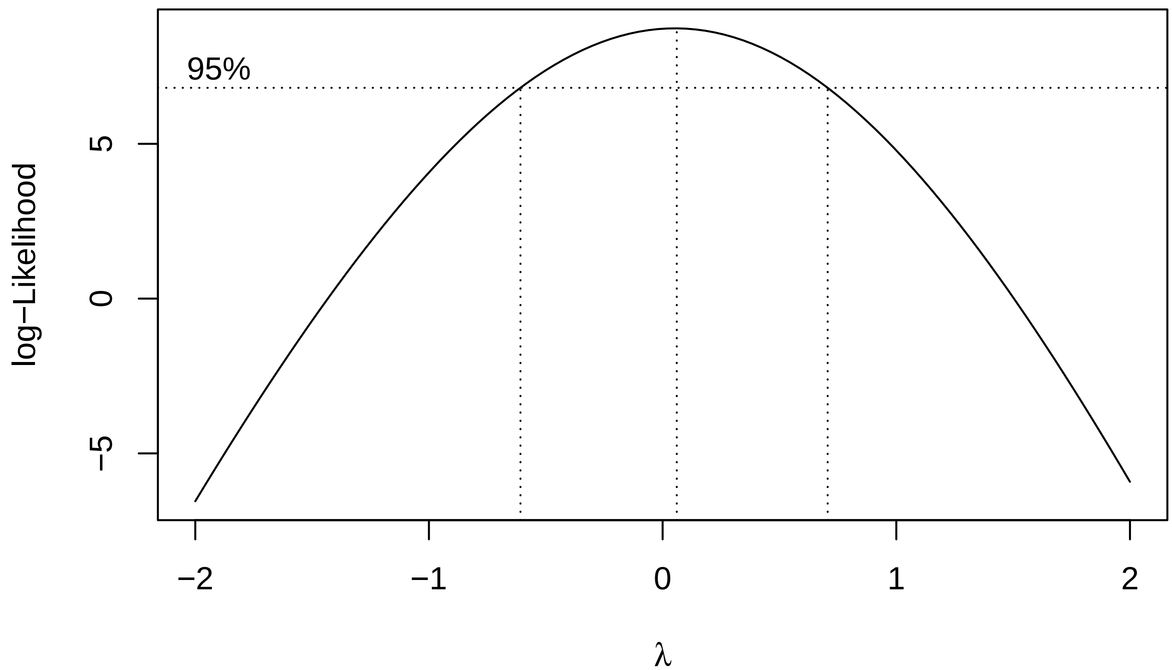
Box-Cox for GDP (Draft)

```r
library(MASS)
```

```
##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:dplyr':
##
##     select
```

```r
# Perform Box-Cox transformation
bc <- boxcox(model_gdp, lambda = seq(-2, 2, by = 0.1))
```

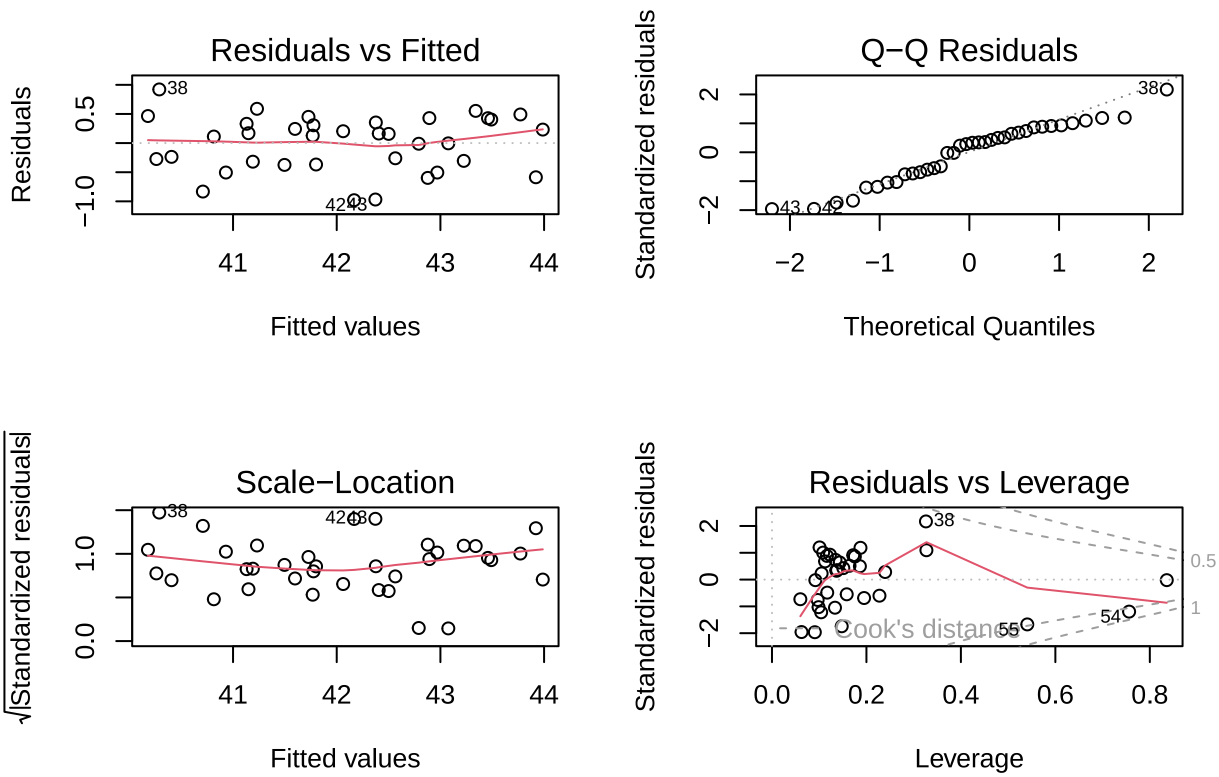



```
# Find the best lambda
best_lambda <- bc$x[which.max(bc$y)]
#cat("Best lambda:", best_lambda, "\n")

# Refit using the Box-Cox transformed GDP:
finalPro_data$GDP_boxcox <- (finalPro_data$GDP^best_lambda - 1) / best_lambda
model_gdp_boxcox <- lm(GDP_boxcox ~ Year + Growth + CPI + Imports + Exports + Population, data = finalP
summary(model_gdp_boxcox)
```

```
##
## Call:
## lm(formula = GDP_boxcox ~ Year + Growth + CPI + Imports + Exports +
##     Population, data = finalPro_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.9805 -0.3315  0.1426  0.3662  0.9210
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -8.288e+02  3.168e+02  -2.616  0.01398 *
## Year         4.474e-01  1.623e-01   2.757  0.00999 **
## Growth       1.540e-02  1.320e-02   1.166  0.25295
## CPI         -2.445e-02  1.097e-02  -2.229  0.03373 *
## Imports      1.188e-02  2.898e-02   0.410  0.68487
## Exports     -1.901e-01  3.302e-02  -5.755 3.12e-06 ***
```

```
## Population  -5.099e-06  2.029e-06  -2.513  0.01777 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5172 on 29 degrees of freedom
##   (22 observations deleted due to missingness)
## Multiple R-squared:  0.847,  Adjusted R-squared:  0.8154
## F-statistic: 26.76 on 6 and 29 DF,  p-value: 1.41e-10
```

```r
par(mfrow = c(2,2))
plot(model_gdp_boxcox)
```




```r
# Check variance again
res_boxcox <- residuals(model_gdp_boxcox)

# Make a grouping variable
fit_boxcox <- fitted(model_gdp_boxcox)
group_boxcox <- ifelse(fit_boxcox > median(fit_boxcox), "High", "Low")

# Levene's Test (Brown-Forsythe is a median-centered version of Levene's test)
leveneTest(res_boxcox ~ group_boxcox, center=median) # Brown-Forsythe test
```

```
## Warning in leveneTest.default(y = y, group = group, ...): group coerced to
## factor.
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  1  0.3973 0.5327
##       34
```

The variance for model GDP improved after transforming

(Draft) Use auto.arima()

```r
# GDP time series covers years 1960-2017:
GDP_ts <- ts(finalPro_data$GDP, start = 1960, frequency = 1)

# Use auto.arima to find the best ARIMA model
best_arima_gdp <- auto.arima(GDP_ts)

# Show model summary
summary(best_arima_gdp)
```

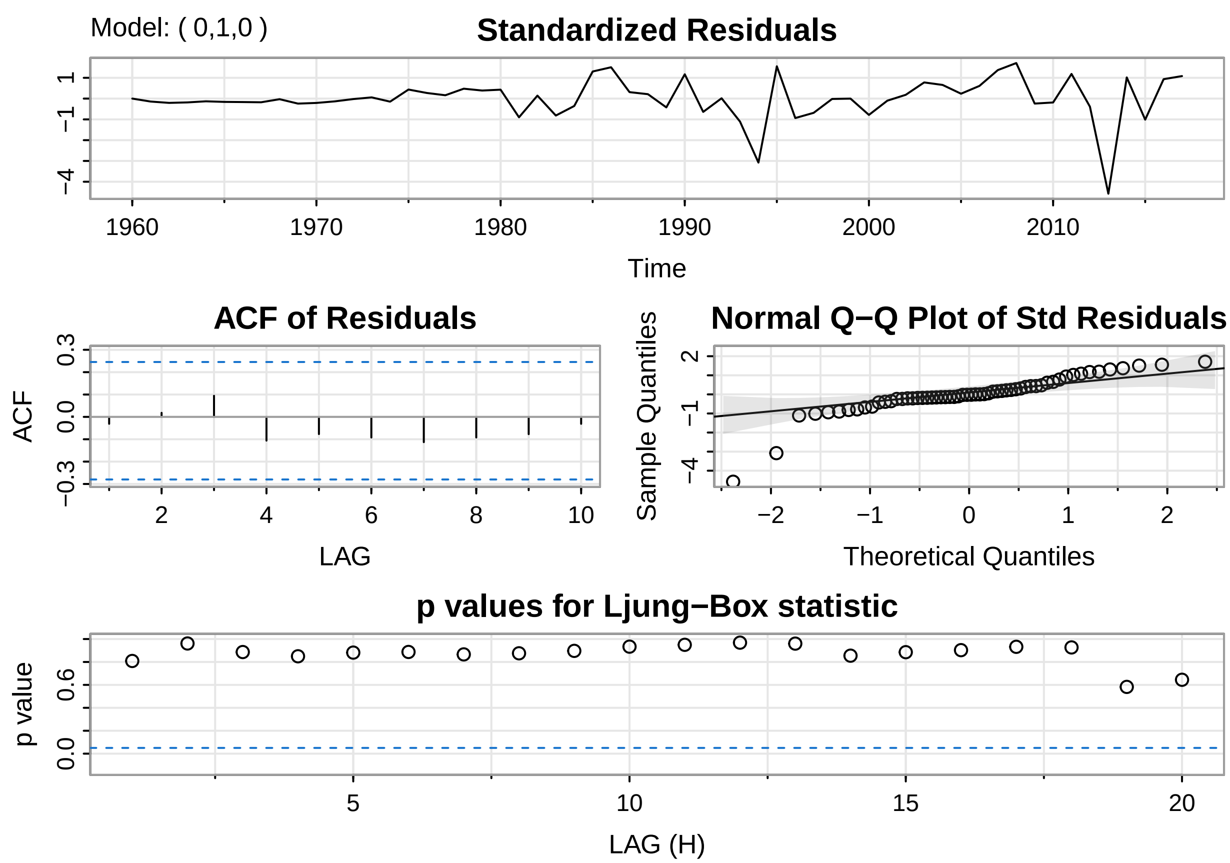
```
## Series: GDP_ts
## ARIMA(0,1,0) with drift
##
## Coefficients:
##          drift
##       32232562
## s.e.  19852873
##
## sigma^2 = 2.269e+16:  log likelihood = -1153.71
## AIC=2311.42   AICc=2311.64   BIC=2315.51
##
## Training set error measures:
##                    ME      RMSE      MAE       MPE     MAPE      MASE
## Training set 1377.983 148026421 93710161 -2.966937 11.05622 0.9363027
##                     ACF1
## Training set -0.03099395
```

```r
# Plot diagnostics:
#checkresiduals(best_arima_gdp)

library(astsa)
```

```
##
## Attaching package: 'astsa'

## The following object is masked from 'package:forecast':
##
##     gas
```

```r
arima_010 = sarima(GDP_ts, 0, 1, 0)
```

```
## initial  value 18.821597
## iter   1 value 18.821597
## final  value 18.821597
## converged
## initial  value 18.821597
## iter   1 value 18.821597
## final  value 18.821597
## converged
## <><><><><><><><><><><><><><>
##
## Coefficients:
##          Estimate       SE t.value p.value
## constant 32232562 19852873  1.6236  0.1101
##
## sigma^2 estimated as 2.229624e+16 on 56 degrees of freedom
##
## AIC = 40.55125  AICc = 40.55252  BIC = 40.62293
##
```



Summary: - ARIMA(0,1,0) model fits the GDP time series pretty well - Residuals show no significant autocorrelation - Errors appear approximately normal - Good baseline so far

Imports Model

Summary Imports

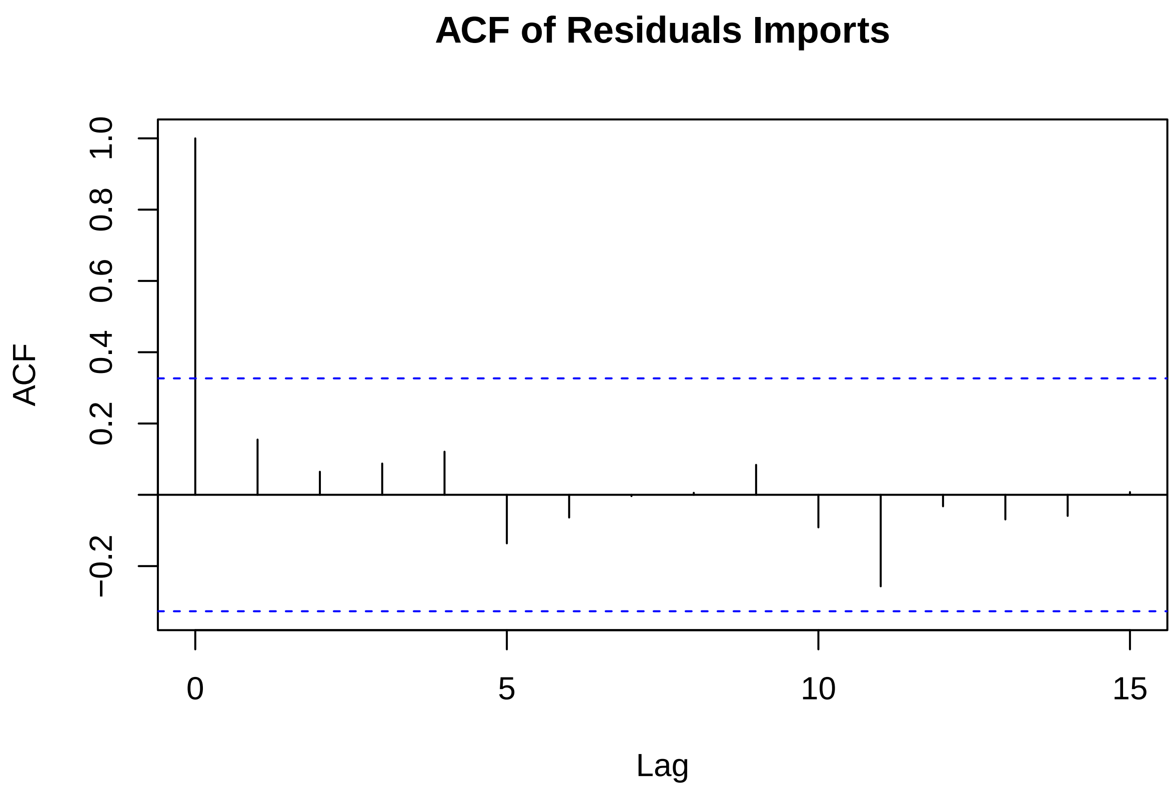
```r
# 2. Diagnose model for Imports
model_imports <- lm(Imports ~ Year + GDP + Growth + CPI + Exports + Population, data = finalPro_data)
summary(model_imports)
```

```
##
## Call:
## lm(formula = Imports ~ Year + GDP + Growth + CPI + Exports +
##     Population, data = finalPro_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.8591 -1.7526 -0.4215  1.2413  8.9778
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.762e+02  2.121e+03   0.413  0.68251
## Year        -4.330e-01  1.088e+00  -0.398  0.69347
## GDP          4.311e-09  2.951e-09   1.461  0.15484
## Growth      -5.806e-02  8.339e-02  -0.696  0.49179
## CPI          2.359e-01  5.998e-02   3.933  0.00048 ***
## Exports      5.924e-01  2.788e-01   2.125  0.04226 *
## Population  -5.022e-06  1.330e-05  -0.378  0.70842
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.199 on 29 degrees of freedom
##   (22 observations deleted due to missingness)
## Multiple R-squared:  0.7213, Adjusted R-squared:  0.6636
## F-statistic: 12.51 on 6 and 29 DF,  p-value: 6.306e-07
```

- Significant variables: Year, CPI, Exports, Population

Residuals Diagnostics

```r
#ACF Imports
acf(residuals(model_imports), main="ACF of Residuals Imports")
```




```
# L-jung-box
Box.test(residuals(model_imports), lag = 20, type = "Ljung-Box")
```

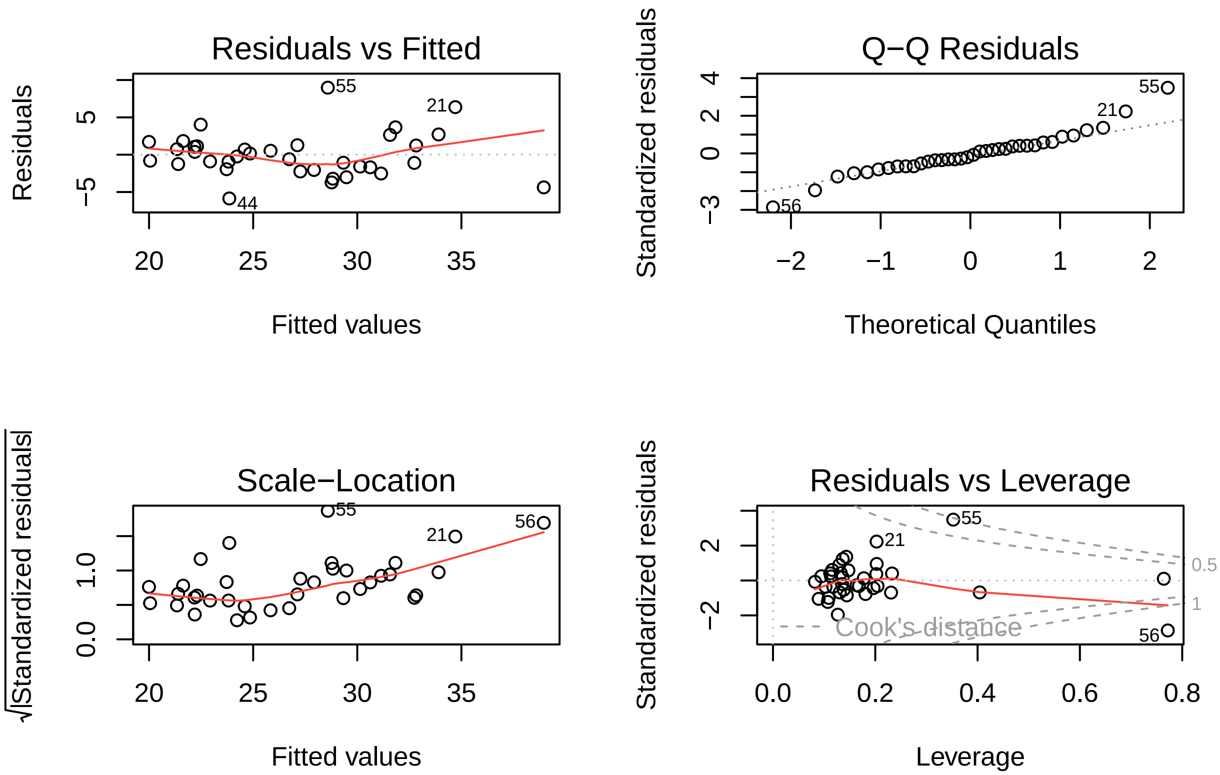
```
##
##  Box-Ljung test
##
## data:  residuals(model_imports)
## X-squared = 12.965, df = 20, p-value = 0.8789
```

Quick look:

- ACF plot shows no significant autocorrelation at higher lags
- Possible AR(1) model
- Ljung-box p-value > 0.1, so residuals are likely white noise

Plot Imports

```
# Diagnostic plots for Imports model
par(mfrow = c(2,2))
plot(model_imports)
```



Quick interpretation:

- Residuals vs fitted: linearity assumption is reasonably met
- QQ plot: approximately normally distributed
- Scale-Location: variance of the residuals is roughly constant
- Residuals vs Leverage: most points have low leverage

(Draft) auto.arima() for Imports

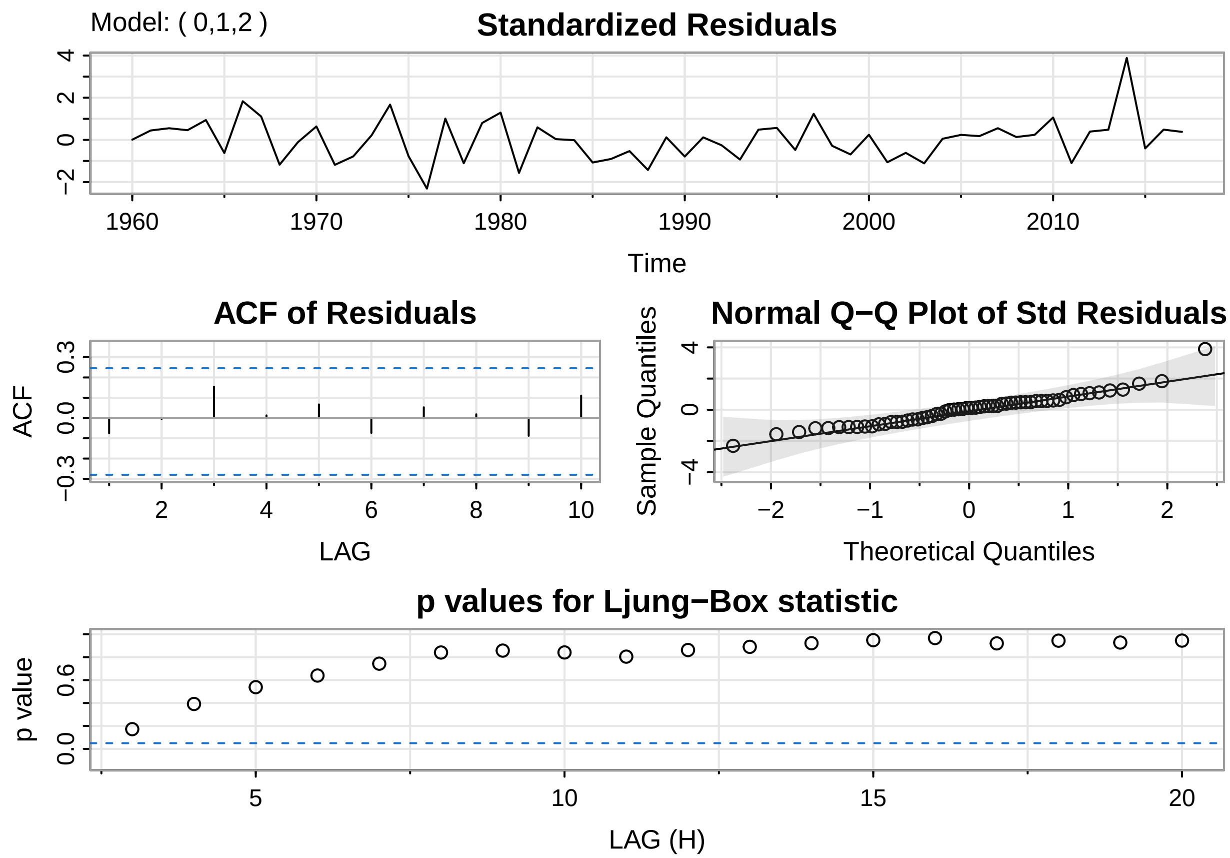
```
# Imports time series covers years 1960-2017:
Imports_ts <- ts(finalPro_data$Imports, start = 1960, frequency = 1)

# Use auto.arima to find the best ARIMA model
best_arima_imports <- auto.arima(Imports_ts)

# Show model summary
summary(best_arima_imports)
```

```
## Series: Imports_ts
## ARIMA(0,1,2)
##
## Coefficients:
##           ma1      ma2
##       -0.0463  -0.4473
```

```
## s.e.   0.1307   0.1361
##
## sigma^2 = 12.33:  log likelihood = -151.68
## AIC=309.37   AICc=309.82   BIC=315.5
##
## Training set error measures:
##                      ME     RMSE      MAE       MPE     MAPE     MASE
## Training set -0.1397995 3.419567 2.587628 -1.299382 8.423651 0.883748
##                     ACF1
## Training set -0.08094744
```

```r
# Plot diagnostics:
#checkresiduals(best_arima_imports)

arima_010 = sarima(Imports_ts, 0, 1, 2)
```

```
## initial  value 1.338143
## iter   2 value 1.244114
## iter   3 value 1.243039
## iter   4 value 1.240054
## iter   5 value 1.239916
## iter   6 value 1.239904
## iter   7 value 1.239904
## iter   7 value 1.239904
## iter   7 value 1.239904
## final  value 1.239904
## converged
## initial  value 1.240322
## iter   2 value 1.240244
## iter   3 value 1.240231
## iter   4 value 1.240162
## iter   5 value 1.240162
## iter   5 value 1.240162
## iter   5 value 1.240162
## final  value 1.240162
## converged
## <><><><><><><><><><><><><><>
##
## Coefficients:
##          Estimate     SE t.value p.value
## ma1       -0.0528 0.1300 -0.4065  0.6860
## ma2       -0.4622 0.1416 -3.2636  0.0019
## constant  -0.1157 0.2333 -0.4960  0.6219
##
## sigma^2 estimated as 11.84284 on 54 degrees of freedom
##
## AIC = 5.458552  AICc = 5.466496  BIC = 5.601924
##
```



Summary:

- ARIMA(0,1,2) model fits the Imports time series well
- Residual diagnostics shows good model fit with no significant autocorrelation left