Real Disposable Income

STA 567
Xander and Kyle

Introduction

For our project we are using a time series data set containing real disposable income, or after tax income, from the Federal Reserve of Economic Research. The data was collected from 1929-2011, with 83 years of data regarding disposable income. Real disposable income is how much money the population of the United States has to spend after considering taxes and inflation. The data is measured in billions and was measured at a frequency of once per year. So for example, the 83rd observation is 12,775.26 is $12.78 Trillion. We are analyzing this data to gain an understanding of economic trends regarding real disposable income. As well as make predictions to further our understanding of how real disposable income has changed in the subsequent years of the data being collected.

Methods

For this project we will analyze the data's time series plot and ACF to check for trends. If we identify a stochastic trend we will apply differencing to the data and evaluate the plot and ACF again. If we identify a deterministic trend we will apply detrending then evaluate the plot and ACF again.

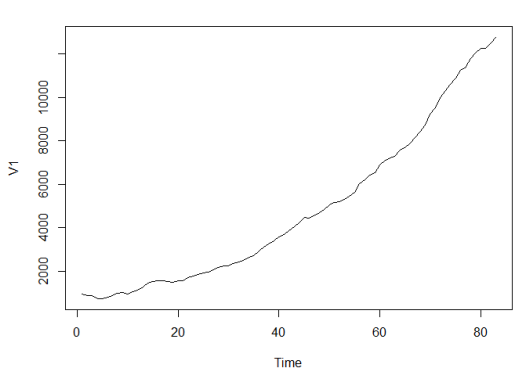
After evaluating the plots and identifying the right trends and eliminating them in the appropriate way we began to identify what underlying model captures the data. This involved looking at the ACF, PACF and EACF. If we see a sharp cut off in the ACF that suggests an MA component, but if we see a gradual decay in the ACF that suggests an AR component. In the PACF if we see a sharp cut off that suggests an AR component and if it slowly decays that suggests an MA component. The EACF will give us several potential candidates for the model. After we evaluate the ACF, PACF and EACF and identify the potential models for the process we will begin evaluating the AIC/BIC as well as residual diagnostics.

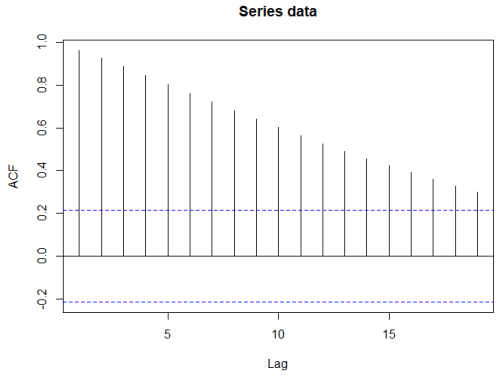
Our selection of the best fit will be based on AIC/BIC, the Shapiro Wilk test of normality on the standardized residuals, a plot of the standardized residuals, and the ACF of the residuals and finally the Ljung-Box Statistics. We expect to see the plot of the standardized residuals to show it is normally distributed, and this will be reflected in the Shapiro Wilkes test for normality. We will also check for only 1 significant lag in the ACF of the residuals, and this significant lag should be at lag 0. Finally we expect to see no significant p-values in the Ljung-Box statistics. Once we identify a model, or models, that satisfy the residual diagnostics, we will choose the model with the lowest AIC/BIC.

Once we have a satisfactory model we will begin forecasting. Forecasts will be made for 10 time points of the model we chose. We will look at the forecasts and how they change over time, as well as the standard error and how it increases over time. Then we will plot our predictions to the known values with the confidence intervals. Due to this data being collected until 2011, we will be able to verify our predictions and compare them to the known real disposable income. This will give us an idea of how accurate our predictions are, and how well the confidence intervals capture the variability of the data.

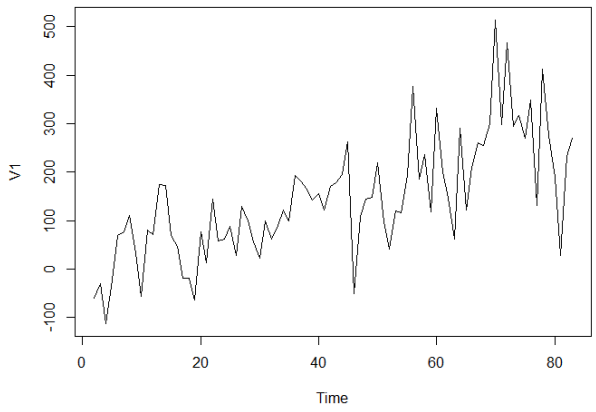
Results and discussion

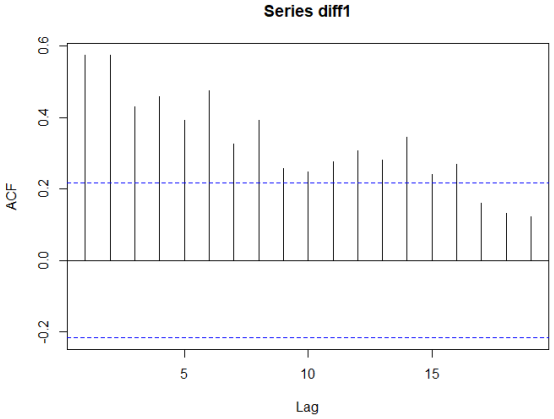
We began analyzing the data with a time series plot.



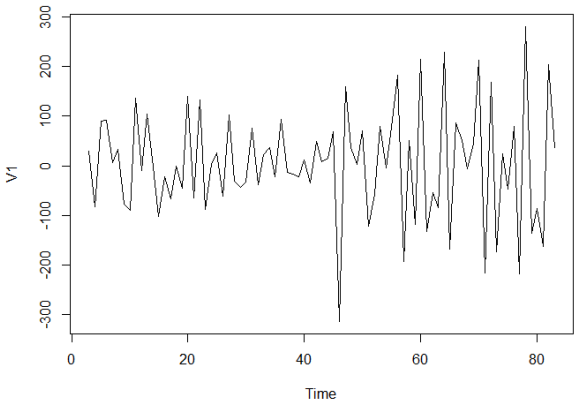


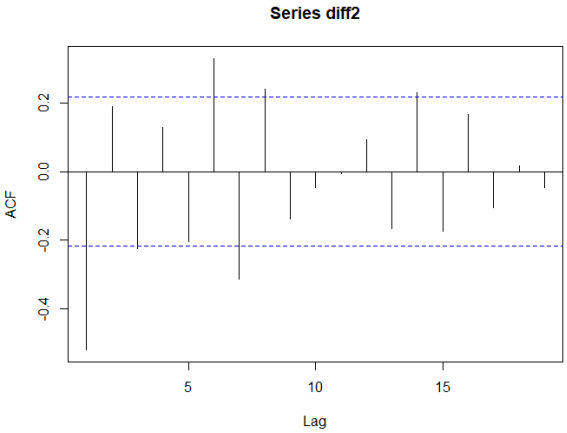
After evaluating the plot it was immediately obvious that there existed a stochastic trend in the data. This is seen in the thread-like trend of the plot. This is reflected in the ACF of the data, with a very slow decay in significance of the lag. We continued by applying first order differencing to the time series data.

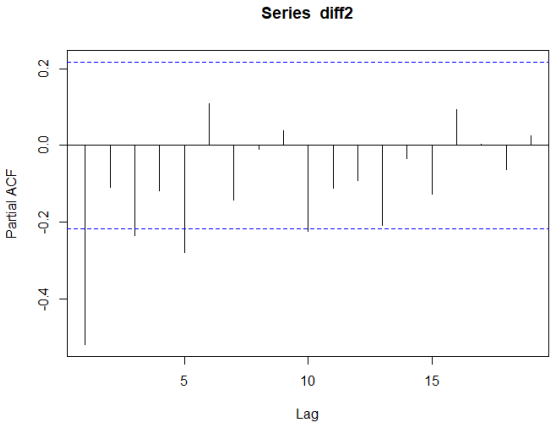




After applying first order differencing the plot appears to be a deterministic trend, however the ACF suggests that there is still a stochastic trend. So we applied second order differencing to the time series data.







After the second order differencing was applied the data appeared to be stationary, while the variance appears to be increasing over time it did not seem significant enough to conclude non-stationarity. The ACF appears to be suggestive of an MA component in the data, while the PACF appears to be suggestive of an AR component. This could indicate that a mixed ARIMA model will be necessary.

```
AR/MA
  0 1 2 3 4 5 6 7 8 9 10 11 12 13
0 x o x o o x x x o o o  o  o  o
1 o x o o o o o o o o o  o  o  o
2 x x o o o o o o o o o  o  o  o
3 x x o x o o o o o o o  o  o  o
4 x x x x o o o o o o o  o  o  o
5 x o o o o o o o o o o  o  o  o
6 x x o o o o o o o o o  o  o  o
7 o x o o o o o o o o o  o  o  o
```

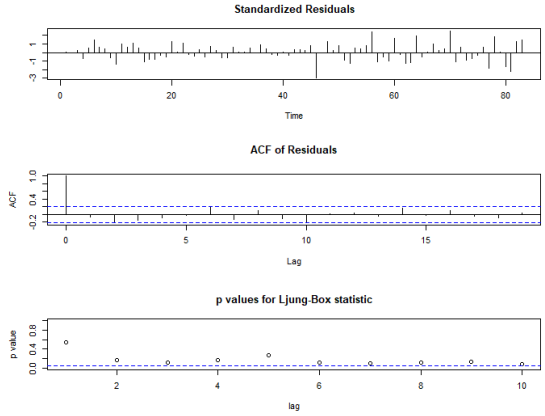
This is the EACF we got from the second order differencing of the time series data. Giving us the potential candidates of ARI(1,2), IMA(2,1), ARIMA(1,2,2) and ARIMA(2,2,2). We then fit the models and took a look at the AIC/BIC.

```
          Model      AIC      BIC
1 ARIMA(1,2,2) 969.8040 974.5929
2     IMA(2,1) 963.4839 968.2728
3 ARIMA(1,2,2) 963.3649 972.9427
4 ARIMA(2,2,2) 962.4542 974.4265
> 
```

It looks like ARIMA(2,2,2) has the lowest AIC, however IMA(2,1) has the lowest BIC. Residual diagnostics will help in narrowing down the candidates. When performing residual diagnostics we took a look at the Shapiro-Wilkes test as well as the standardized residuals, ACF of Residuals and the Ljung-Box statistics.

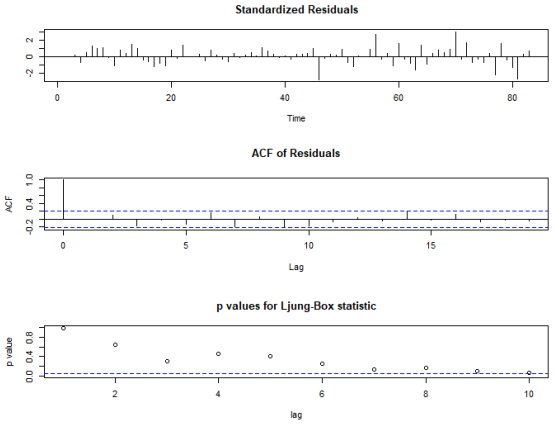
For ARI(1,2)

```
        Shapiro-Wilk normality test

data:  rstandard(data_ari12)
W = 0.9928, p-value = 0.9307

> 
```



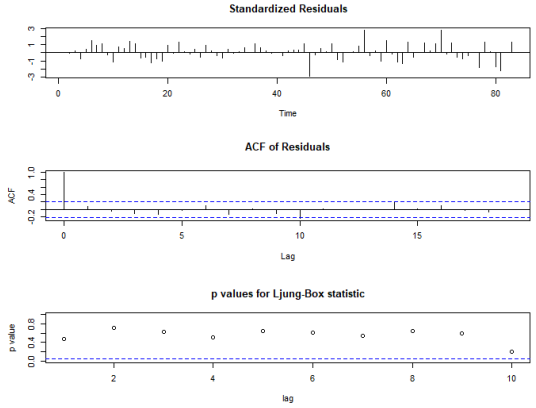
For IMA(2,1)

```
        Shapiro-Wilk normality test

data:  rstandard(data_ima21)
W = 0.97559, p-value = 0.1151
```



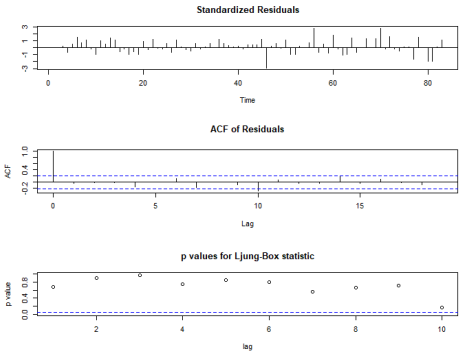
For ARIMA(1,2,2)

```
        Shapiro-Wilk normality test

data:  rstandard(data_arima122)
W = 0.97867, p-value = 0.1832
```



For ARIMA(2,2,2)

```
        Shapiro-Wilk normality test

data:  rstandard(data_arima222)
W = 0.97922, p-value = 0.1987
```



All of the models appear to have satisfied the Shapiro-Wilk normality test. However when it comes to the residual diagnostics, all of them appeared to satisfy normality in the residuals, and no significant lag points in the ACF, except some of them had significant p-values in the Ljung-Box Statistics. The only models that satisfied the Shapiro-Wilk normality test, and the residual diagnostics, are ARIMA(1,2,2) and ARIMA(2,2,2). Considering that ARIMA(2,2,2) has the lower AIC of the two, we came to an agreement that ARIMA(2,2,2) was the better option.

After we settled on ARIMA(2,2,2) we began doing some forecasting with the fitted model. The model we are using is ARIMA(2,2,2) with the parameters $\Phi 1 = -0.6541$ $\Phi 2 = 0.3013$ $\Theta 1 = 0.0265$ and $\Theta 2 = -0.7315$

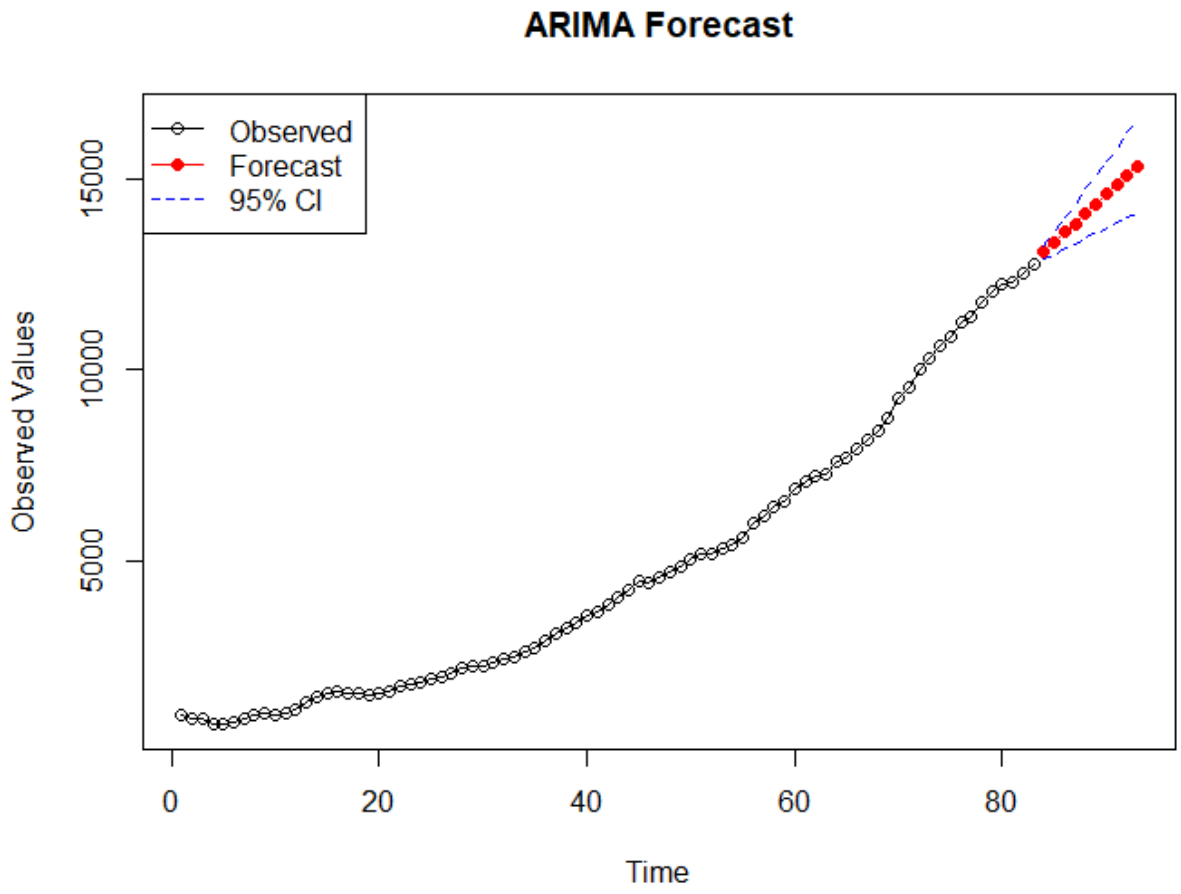
So our fitted model was as follows:

$\nabla^2 Y_t = -0.6541 Y_{t-1} + 0.3013 Y_{t-2} - 0.0265 e_{t-1} + 0.7315 e_{t-2}$

Once we had the fitted model we were satisfied with we began forecasting. We did forecasts for 10 time points beyond the recorded data.

	Forecast	SE	Lower	Upper
1	13079.30	85.86889	12911.00	13247.61
2	13302.29	145.81346	13016.49	13588.08
3	13588.61	207.85853	13181.21	13996.01
4	13809.07	264.32723	13290.99	14327.15
5	14091.69	324.26839	13456.13	14727.26
6	14313.81	381.86883	13565.35	15062.28
7	14594.23	443.40186	13725.17	15463.30
8	14818.29	503.88699	13830.67	15805.91
9	15096.79	568.23611	13983.04	16210.53
10	15322.69	632.14734	14083.68	16561.70

Based on these forecasts, the United States in the year 2012 would have expected to have around $13.08 Trillion of disposable income. We see this rise to $15.32 Trillion by the year 2021. While these forecasts seem to align with the trend of the data quite well. The SE also spikes significantly in the 10 years of forecasts. So the true disposable income for 2021 should lie somewhere in the range of $14.08 Trillion to $16.56 Trillion. However the accuracy of the higher points is questionable, due to the very high standard error.



The plot of the forecasts fitted to the time series data shows that our predictions align quite closely to the known real disposable income. However, how well this accounts for inflation is still in question. The blue lines show that the range for our confidence intervals expand significantly in time. This is certainly a result of the increase in standard error in the higher lag predictions. The first few predictions have tighter confidence intervals, and should be more accurate predictions to the true real disposable income of those subsequent years. However the later predictions have wide ranges in the confidence intervals.

For the sake of discussion we decided to compare the 10th prediction, for the year 2021, to the actual real disposable income recorded for 2021. Our prediction is $15.322 Trillion with the 95% confidence interval (14,083.68, 16,561.70). The true measure of real disposable income in December 2021 was recorded to be 17.17 Trillion (*US real disposable personal income (I:USRDPINY)*) . This does not fall into our confidence interval, but is a couple of trillion dollars short. However if we compare the first prediction, for the year 2012, we get the estimate of 13.08 Trillion with a 95% confidence interval of (12911, 13247.61). The true measure of real disposable income in December 2012 was 13.13 Trillion (*US real disposable personal income (I:USRDPINY)*) . So for the first prediction it was quite accurate in terms of the confidence interval, and the predicted didn't come too short of the true disposable income of 2012.

Summary

Overall, this analysis was successful. Our completed model has a reasonable degree of accuracy, so the model will be useful overall. While we did have to difference the data twice for it to be usable, it ultimately does not hurt our predictions. One interesting detail is that the dataset had so many candidate models from the EACF. While we were able to narrow it down, there were a plethora of possible underlying processes to worry about. The forecast seems reasonable, and its predictions were within reason for the first few predictions past the collected data.

R Code

```
library(car)
library(RODBC)
library(MASS)
library(gplots)
library(plyr)
library(lattice)
library(visreg)
library(Hmisc)
library(ppcor)
library(TSA)
library(graphics)
library(astsa)
library(tseries)
library(forecast)
data = ts(read.csv("Project Data.csv", header = FALSE))
print(data)
plot.ts(data)
acf(data)
#The plot of the time series data has a clear thread-like structure suggesting a stochastic trend, this is reflected in the acf of the data
#showing a very slow decay
diff1 = diff(data)
plot.ts(diff1)
acf(diff1)
#The data still appears to not be stationary with an increasing trend. We attempted detrending but that appeared to result in the time series being white noise
#The slowly decaying acf seems suggestive that there is still a stochastic process in the data
#We will apply second order differencing
diff2 = diff(diff1)
plot.ts(diff2)
acf(diff2)
pacf(diff2)
eacf(diff2)
#After differencing the data appears to be stationary.
#Potential to be ARI(1,2), IMA(2,1), ARIMA(1,2,2) and ARIMA(2,2,2)
data_ari12 = arima(data, order = c(1,2,0), method = "ML", include.mean = F)
data_ima21 = arima(data, order = c(0, 2, 1), method = "ML", include.mean = F)
data_arima122 = arima(data, order = c(1, 2, 2), method = "ML", include.mean = F)
data_arima222 = arima(data, order = c(2,2,2), method = "ML", include.mean = F)
data_IMA22 = arima(data, order = c(0,2,2), method = "ML", include.mean = F)
aic_bic_values = data.frame(
  Model = c("ARIMA(1,2,2)", "IMA(2,1)", "ARIMA(1,2,2)", "ARIMA(2,2,2)"),
```

```
  AIC = c(AIC(data_ari12), AIC(data_ima21), AIC(data_arima122), AIC(data_arima222)),
  BIC = c(BIC(data_ari12), BIC(data_ima21), BIC(data_arima122), BIC(data_arima222))
)
print(aic_bic_values)
#ARIMA(2,2,2) appears to have the lowest AIC, but not the lowest BIC. The lowest BIC goes to
IMA21
#Residual Diagnostics
tsdiag(data_ari12)
shapiro.test(rstandard(data_ari12))
tsdiag(data_ima21)
shapiro.test(rstandard(data_ima21))
tsdiag(data_arima122)
shapiro.test(rstandard(data_arima122))
tsdiag(data_arima222)
shapiro.test(rstandard(data_arima222))
#All models appear to pass the shapiro wilk test, so they are all normally distributed in terms of
the standardized residuals.
#None of them appear to have significant acf in the residuals.
#However only ARIMA(2,2,2) appear to satisfy the Ljung-Box test
#Based on the residual diagnostics, and the AIC/BIC ARIMA(2,2,2) appears to be the best fit for
the data.
print(data_arima222)
fit2_pr <- predict(data_arima222, n.ahead = 10)
forecast <- fit2_pr$pred
se <- fit2_pr$se
l <- forecast - 1.96 * se
u <- forecast + 1.96 * se
forecast_results <- data.frame(Forecast = forecast, SE = se, Lower = l, Upper = u)
print(forecast_results)
par(mfrow = c(1, 1))
t <- 1:length(data)
plot(
  t, data, type = "o", xlim = c(1, length(data) + 10), ylim = range(c(data, l, u)),
  xlab = "Time", ylab = "Observed Values", main = "ARIMA Forecast"
)
forecast_time <- (length(data) + 1):(length(data) + length(forecast))
lines(forecast_time, forecast, col = "red", type = "o", pch = 16)
lines(forecast_time, u, col = "blue", lty = "dashed")
lines(forecast_time, l, col = "blue", lty = "dashed")
legend(
  "topleft",  # Change from "topright" to "topleft"
  legend = c("Observed", "Forecast", "95% CI"),
  col = c("black", "red", "blue"), lty = c(1, 1, 2), pch = c(1, 16, NA),
  bg = "white")
```

References

US real disposable personal income (I:USRDPINY). YCharts. (n.d.).https://ycharts.com/indicators/us_real_disposable_personal_income_yearly

U.S. Bureau of Economic Analysis, Real disposable personal income [A067RX1A020NBEA], retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/A067RX1A020NBEA, November 30, 2024