Katherine Olson

998992204

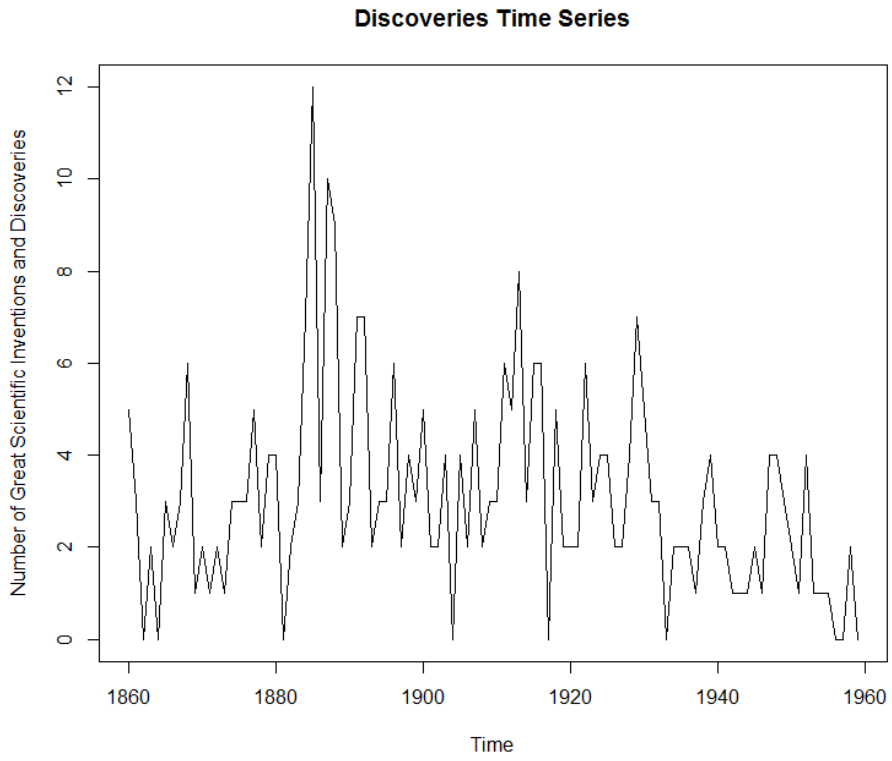
Using the Great Invention of Time Series Exploratory Data Analysis to Analyze the Number of Great Inventions and Scientific Discoveries from 1860 to 1959

Introduction

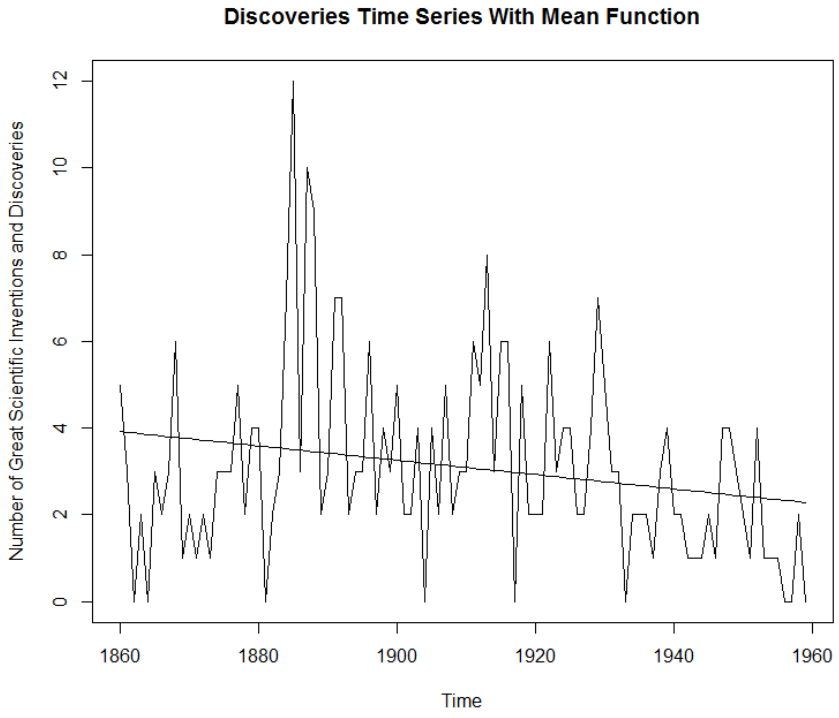
I am using the discoveries data from the datasets R package. The data contains the numbers of "great" inventions and scientific discoveries in each year from 1860 to 1959. It contains 100 entries. This is a time series data set because it contains data taken over a period of time. It is important to analyze this data set to see how the number of scientific discoveries changed in a span of 100 years. This can be interesting because there were a lot of technology improvements from 1860 to 1959 that lead to the thought that the time series would have an upward trend. But, there reaches a point where it seems like so many things have already been discovered and invented which leads to the thought that the time series might not have an upward trend. There are also a lot of other factors to consider, so it will be interesting to see what pattern the data has.

Material and Methods / Results

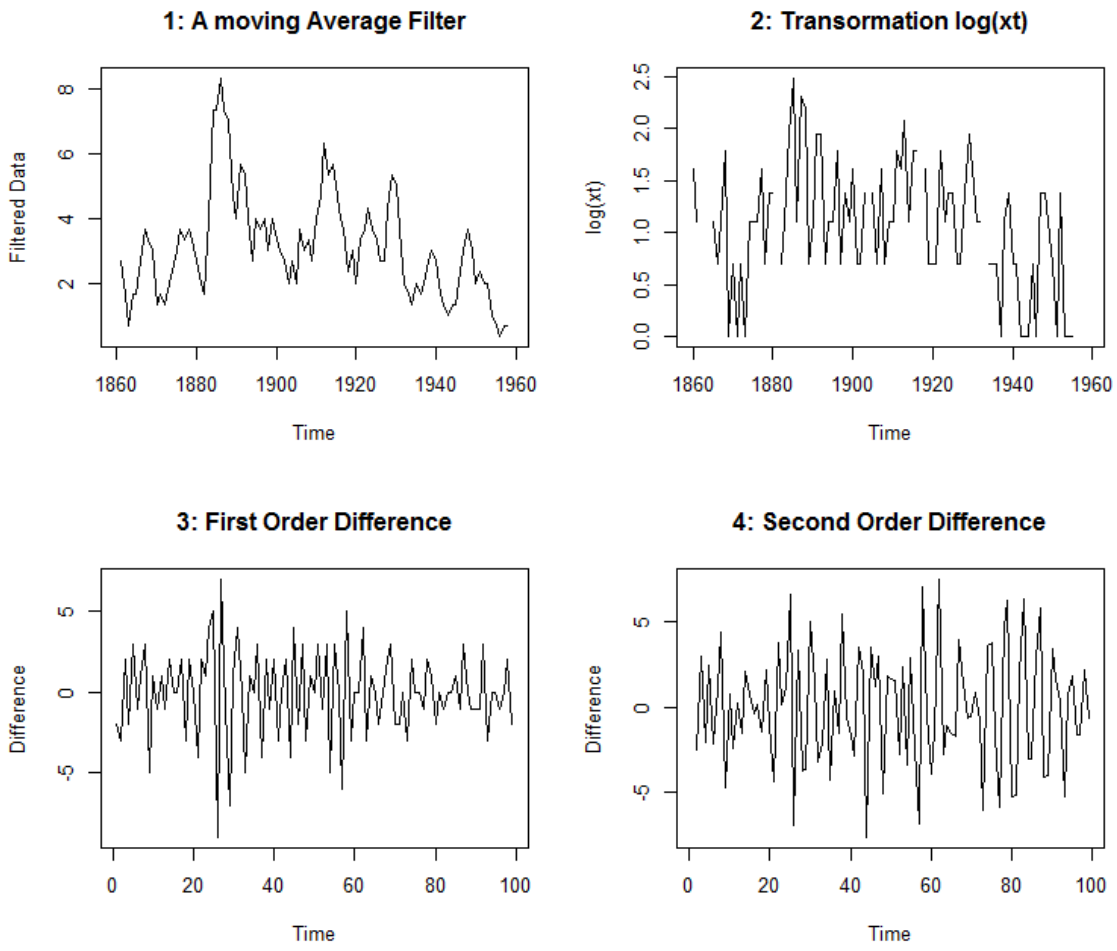
I started by plotting the data.



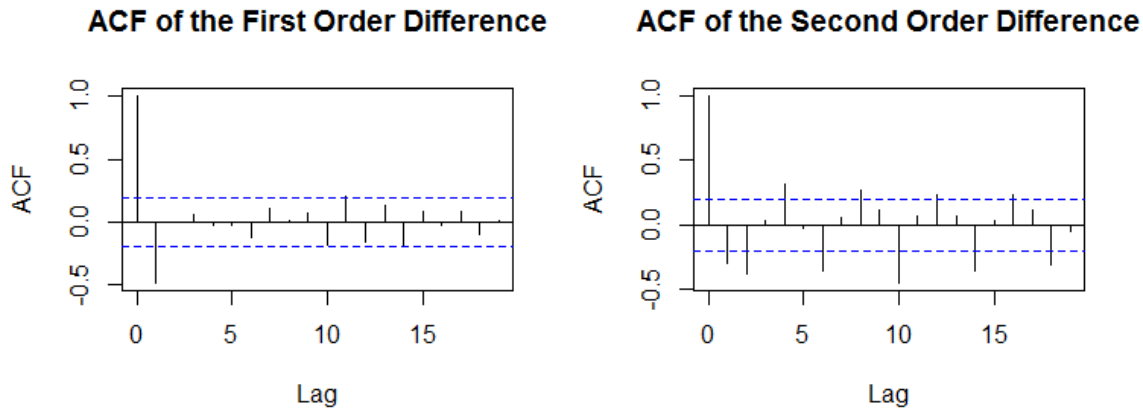
Looking at the data, it does definitely not look stationary. It appears to rise in the middle and the variance does not seem constant. There is a point around 1880 where the variance gets large. So, next I decided to look at the mean function by using linear regression on the data and time to see how dependent the mean function is on time. I plotted the fitted line with the data and got the result below. The mean function is almost horizontal, so it is not very dependent on time, but still a small amount dependent. This means that the reason the plot of the time series does not look stationary probably has more to do with the variance.



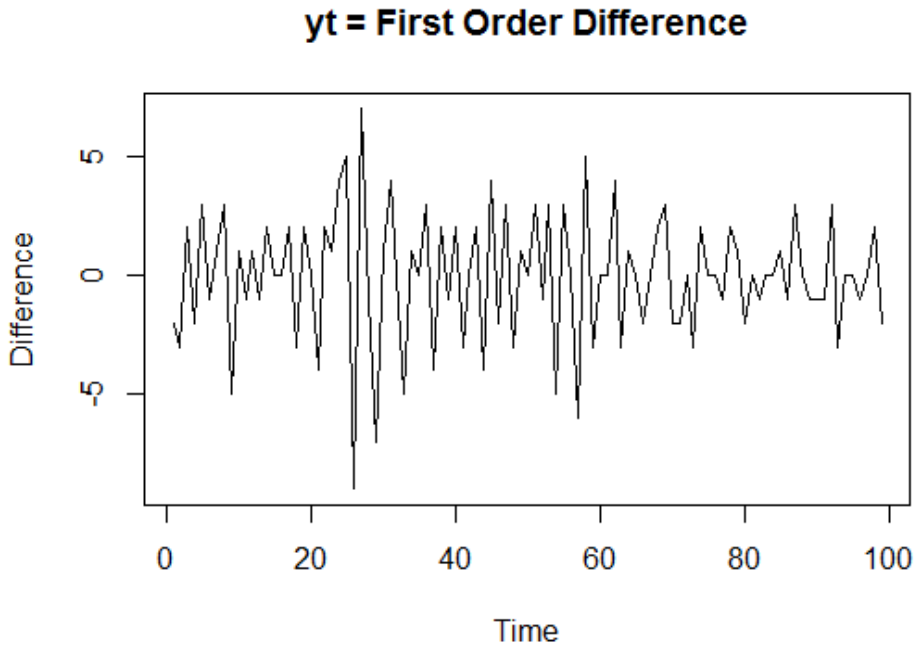
Next, I tried four techniques to make the data stationary. The plots of each of these can be seen on the next page. The first thing I tried was a moving average filter where the data x_t is replaced by $\frac{1}{3}(x_{t+1} + x_t + x_{t-1})$. The new series is constructed by taking the average of the current, one past and one future observation of the original series. Looking at the plot of this new series, it still does not look stationary. It resulted in a smoothed version of the time series, which did not solve any of the problems with stationarity. So next I tried a log transformation where x_t is replaced by $\log(x_t)$. Looking at this plot, it also does not look stationary and has some strange looking gaps. The middle section of the transformed time series is much higher than the outer chunks. So for my last two techniques, I used differencing. I tried a first order difference ($x_t - x_{t-1}$) and a second



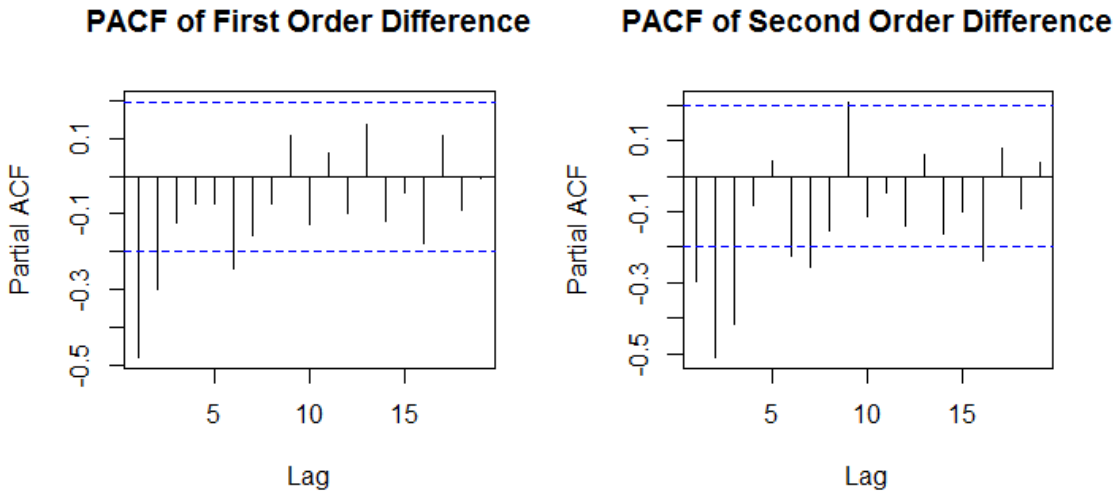
order difference(x_t - x_{t-1} - x_{t-2}). These plots look much closer to stationary, but I am not sure which one is more stationary, so I looked at their ACF plots. Keeping in mind that the goal for a stationary ACF is a value of 1 at lag 0 and no value for ACF for the rest of the lags, the ACF for a first order difference has smaller lines after lag 0. So I am picking a first order difference, referred to as y_t from now on, to use for the rest of the project as a stationary difference of the discoveries data.



So, a bigger plot of the first order difference I am choosing can be seen below. It is definitely not perfectly stationary. The variance is still a little off, especially from time 20 to 60. But, this is the closest to stationary I can get, and it is pretty good. Looking at the ACF of y_t on the last page, from the book we know that The ACF measures the linear predictability of the series at time t, say x_t , using only the value x_s. So, the ACF of the first order difference tells us that we have fairly good linear predictability. It is not as good for lag 1, but close to stationary for the rest of the lags.



Next, I looked at the PACF functions for the two differences to confirm I made a good choice.



The two PACF functions look very similar, but the second order difference has longer lines at a few lag points. So, picking the first order difference was a more stationary choice. The PACF function tells us the conditional correlation of x_t and x_{t-h}. The PACF of the first order difference is better than the second, but still not great. The conditional correlation is a bit too high at a good chunk of the points. Since I am just using one time series, computing the CCF is not applicable. Since I just have one variable with the one time series I have, I do not know how I would be able to try a bunch of models and use AIC or BIC, so I have done all I can do for EDA using the data set I picked. I do not have other variables, but if I did, I think it would be interesting to consider adding the variables: number of technological advances, gender of the inventor (probably in a percentage of male or female), education level of the world (maybe something like 1 = elementary school, 2 = middle school, 3 = high school, and so on), and much more to my model.

Conclusion and Discussion

In my exploratory data analysis, I looked at the discoveries data, decided it was not stationary, and tried the best options of all I could think of to make the data stationary. I used the ACF and PACF functions to help in my decision. I ended up picking a first order difference as the best option. I learned that in practice, making a time series data set stationary is not very easy. The result will most likely not be as perfect as you would like. I also learned that when you only have one time series with one variable, exploratory data analysis can only go so far. There is still much more that can be learned from the great scientific discoveries and inventions from 1860 to 1959. There are many more factors involved in what the number was each year that cannot be looked into with the given data.

Code Appendix

```
library(datasets)
xt = discoveries
plot.ts(xt, main = "Discoveries Time Series", ylab = "Number of Great Scientific Inventions and Discoveries")
plot.ts(xt, main = "Discoveries Time Series With Mean Function", ylab = "Number of Great Scientific Inventions and Discoveries")
fit = lm(xt~time(xt)) # mean function
```

```
lines(as.vector(time(xt)), as.vector(fitted(fit))) # add to plot
# Attempts to make data stationary:
par(mfrow = c(2, 2))
v = filter(xt, sides=2, rep(1/3,3)) # 1: Moving Average
plot.ts(v, main = "1: A moving Average Filter", ylab = "Filtered Data")
logxt = log(xt) # 2: Log
plot.ts(logxt, main = "2: Transormation log(xt)", ylab = "log(xt)")
xtDiff = c() # 3: First order difference
xtDiff[0] = 1
for (i in 1:length(xt))
{
  xtDiff[i] = xt[i] - xt[i-1]
}
xtDiff2 = xtDiff[-1] # Have to remove first element
plot.ts(xtDiff2, main = "3: First Order Difference", ylab = "Difference")
xtDiff3 = c() # 4: 2nd Order Difference
xtDiff3[0] = 1
for (i in 2:length(xt))
{
  xtDiff3[i] = xt[i] - xt[i-1] - x[i-2]
}
xtDiff4 = xtDiff3[-c(1, 2)] # Have to remove first element
plot.ts(xtDiff4, main = "4: Second Order Difference", ylab = "Difference")
# Comparing ACFs
par(mfrow = c(1, 2))
acf(xtDiff2, main = "ACF of the First Order Difference")
acf(xtDiff4, main = "ACF of the Second Order Difference")
# Plot of now (close to) stationary data:
par(mfrow = c(1, 1))
plot.ts(xtDiff2, main = "yt = First Order Difference", ylab = "Difference")
# Looking at pacf of the two differences
```

```
par(mfrow = c(1, 2))
pacf(xtDiff2, main = "PACF of First Order Difference")
pacf(xtDiff4, main = "PACF of Second Order Difference")
```