Time Series Analysis of Abortion Rates in the Unites States from 1973 to 2016

by Veronica Capco (capco@ucsb.edu)

Abstract

This project investigates the trends in abortion rates in the United States using data from 1973 to 2016, using Seasonal ARIMA modeling (SARIMA) and Threshold Autoregressive modeling (TAR). It can be seen through the time series plots that abortion rates for women in the United States have been on a decline, and that the ACF plots show that these rates are non-stationary. This project investigates possible reasons for that decline as well as predicting future values of abortion rates. Using the statistical methods mentioned earlier, I have found that abortions rates are likely to stay on this decline for the coming years. The results from the TAR modeling using a threshold = 0.05 show that the trend is consistent with the original data. With these findings, we can use the information obtained from the analysis and graphs to determine possible sudden changes with abortion rates and look deeper into what threshold parameters can say about the future of abortion in the United States.

1 Introduction

Abortion has had a long history in the United States, from significant court cases such as Roe v. Wade to the more recent Pro-Life movement. Abortion has existed for a very long time, but it wasn't until the mid-1800s when attitudes around abortion in the United States had changed. Since each state in the United States controls their own laws to abortion, some states have passed their own anti-abortion laws which have caused a lot of uproar considering there exists other states with less restrictions on abortion. Today, abortion remains a highly-controversial political and religious topic, and it is likely to remain that way in the next several years.

My project examines the time series of abortion rates in the United States from the years 1973 to 2016. Of course, the data I used only tracked abortions done in a professional medical setting, but there is no denying the fact that abortions under the table have occurred and continue to occur. I chose this data set because I have always been interested in statistics related to health and I thought it would be interesting to investigate data about a heavily-debated topic in the United States. With the overturn of Roe v. Wade in 2022, reproductive rights remains a hot topic in the US.

The purpose is to develop a better understanding of abortion trends in the United States and how external factors may affect these trends. In the past, my chosen data set has been used to track rates of pregnancy, birth, and abortion rates across the states and has been analyzed reproductive trends within age groups of women. An important discovering from the data set is the significant decrease in abortion rates over the past few decades, for women ages 15-44. I will be applying the following methods: SARIMA modeling and Threshold Autoregressive modeling. I decided on these methods because they will determine future predicted values of the abortion rates in the United States, as well as analyzing what thresholds trigger sudden changes in the abortion rates, which can be seen as having a non-linear relationship with time.

2 Data

My data set contains data about pregnancy, birth, and abortion rates in the United States from 1973 to 2016. It contains 5 variables with 23172 observations. The five variables are Year, State, Metric, Age Range, and Events per 1000 women with the pregnancy, birth, and abortion rates recorded in occurrences per 1000 women. The data was collected by several national data collection sources such as the US Census Bureau, National Center for Health Statistics, the CDC, and the Guttmacher Institute Abortion Provider Census. Birth and age data were collected from birth certificates and abortion rates were calculated from surveys conducted by the abortion census. For some years, surveys were not conducted so abortion rates were calculated using linear interpolation. Authors Isaac Maddow-Zimmet, Kathryn Kost, and Sean Finn of Guttmacher Institute used those sources to compile estimates of the data.

I found this data set on data.world but the original source is actually from OSF. I chose this data set because it had a good number of observation and had exactly what I was looking for: abortion rates by state in the United States. This data set is important because it gives us a lot of information about trends in pregnancy, birth, and abortion. Studying it can help us make better decisions about reproductive care and education.

webpage links:
https://data.world/vizwiz/pregnancy-birth-abortion-rates-in-the-united-states
https://www.guttmacher.org/report/pregnancies-births-abortions-in-united-states-1973-2016

Uploading Packages

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

##
## Attaching package: 'forecast'

## The following object is masked from 'package:astsa':
##
##     gas

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

Cleaning the Data

```
##   Year State        Metric Age.Range Events.per.1.000.women
## 1 1988    AL Abortion Rate     15-17                   24.0
## 2 1992    AL Abortion Rate     15-17                   19.1
## 3 1996    AL Abortion Rate     15-17                   12.5
## 4 2000    AL Abortion Rate     15-17                    9.3
## 5 2005    AL Abortion Rate     15-17                    7.0
## 6 2006    AL Abortion Rate     15-17                    7.1
```

```
## # A tibble: 6 x 2
##    Year Total
##   <int> <dbl>
## 1  1973  16.3
## 2  1974  19.3
## 3  1975  21.7
## 4  1976  24.2
## 5  1977  26.4
## 6  1978  27.7
```

3 Methodology

For the first method, I am going to be fitting a seasonal (S)ARIMA model for abortion rates in the United States from 1973 to 2016. The abortion rates are in recorded in events per 1000 women. I will be generating new data for each month of each year from the years 1973 to 2016 in order to predict abortion rates by month for the next 12 months. After doing so, I will be creating a time series for this new generated data and then fitting a seasonal ARIMA model using that time series.

For the second method, I am going to use a Threshold Autoregressive model and use the SETAR (Self Exciting Threshold Autoregression) function to perform the analysis. A Threshold Autoregressive model (TAR) is used in various fields, particularly with fickle data, because this model is uses thresholds to determine sudden changes in data. Using the `setar` function from the tsDyn library, I can generate a summary of the low regimes, high regimes, thresholds, and residuals which are necessary for analysis.

3.1 SARIMA (p, d, q) x (P, D, Q) Model

The fitted Seasonal ARIMA model is ARIMA(4,2,3)(1,0,1)[12].

```
## Series: ts_abortion_monthly
## ARIMA(4,2,3)(1,0,1)[12]
##
## Coefficients:
##           ar1     ar2     ar3     ar4      ma1     ma2     ma3    sar1     sma1
##       -0.2020  0.2077  0.1509  0.0729  -1.5831  0.2607  0.3337  0.6459  -0.4771
## s.e.   0.3977  0.0833  0.0834  0.0537   0.3964  0.7673  0.3748  0.1456   0.1632
##
## sigma^2 = 0.4979:  log likelihood = -550.65
## AIC=1121.31   AICc=1121.74   BIC=1163.75
```

3.2 Threshold Autoregressive Model

```
##
## Non linear autoregressive model
##
## SETAR model ( 2 regimes)
## Coefficients:
## Low regime:
##     const.L      phiL.1      phiL.2      phiL.3      phiL.4
## -0.17752192  0.28475377  0.31633772  0.04946762 -0.20237791
##
## High regime:
##    const.H     phiH.1     phiH.2     phiH.3     phiH.4
## -0.5764089  0.1827485  0.7652179 -0.2033106  0.0561145
##
## Threshold:
## -Variable: Z(t) = + (1) X(t)+ (0)X(t-1)+ (0)X(t-2)+ (0)X(t-3)
## -Value: 0.05 (fixed)
## Proportion of points in low regime: 82.05%    High regime: 17.95%
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.648981 -0.215473 -0.039707  0.125632  0.955006
```

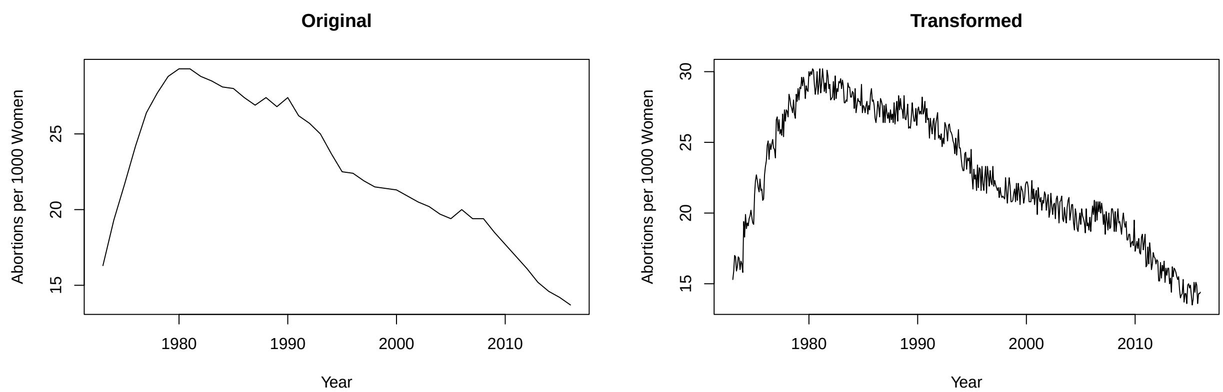
```
##
## Fit:
## residuals variance = 0.1186,  AIC = -72, MAPE = 76.73%
##
## Coefficient(s):
##
##          Estimate  Std. Error  t value Pr(>|t|)
## const.L -0.177522    0.137291  -1.2930  0.20498
## phiL.1   0.284754    0.223067   1.2765  0.21068
## phiL.2   0.316338    0.157639   2.0067  0.05303 .
## phiL.3   0.049468    0.152214   0.3250  0.74724
## phiL.4  -0.202378    0.133467  -1.5163  0.13896
## const.H -0.576409    0.472123  -1.2209  0.23078
## phiH.1   0.182749    0.516450   0.3539  0.72570
## phiH.2   0.765218    0.630018   1.2146  0.23314
## phiH.3  -0.203311    0.329236  -0.6175  0.54113
## phiH.4   0.056114    0.428767   0.1309  0.89667
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Threshold
## Variable: Z(t) = + (1) X(t) + (0) X(t-1)+ (0) X(t-2)+ (0) X(t-3)
##
## Value: 0.05 (fixed)
```

4 Results

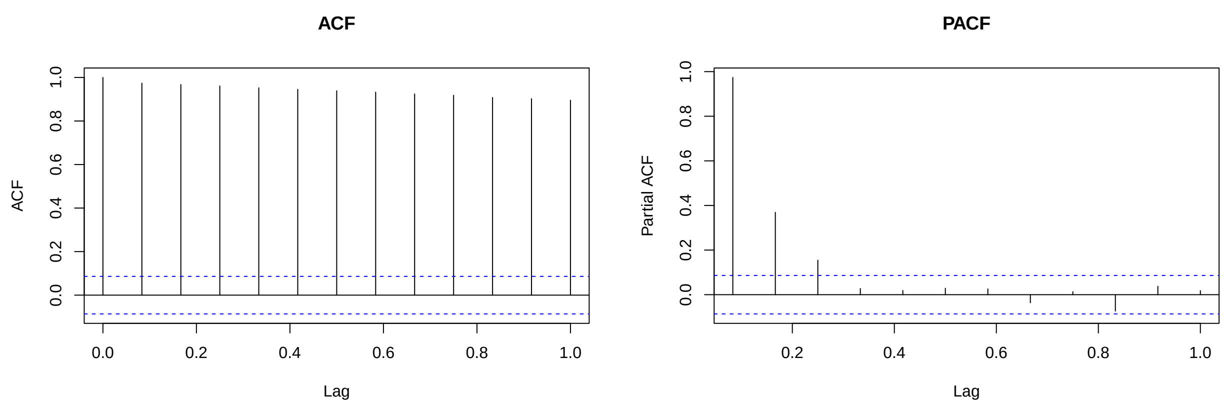
Below are the results of my applied methods.

4.1 Results from SARIMA (p, d, q) x (P, D, Q)

These are the plots of the original and transformed data.



Below are the plots of the ACF and PACF.



These are the results of the forecast for the next 12 months. The predicted average abortions per 1000 women for the next 12 months is 13.92.

```
##           Jan      Feb      Mar      Apr      May      Jun      Jul      Aug
## 2016          14.12809 13.92238 13.78338 13.83995 14.13859 13.85841 14.03255
## 2017 13.86894
##           Sep      Oct      Nov      Dec
## 2016 14.00405 13.67460 13.89647 13.88096
## 2017
```

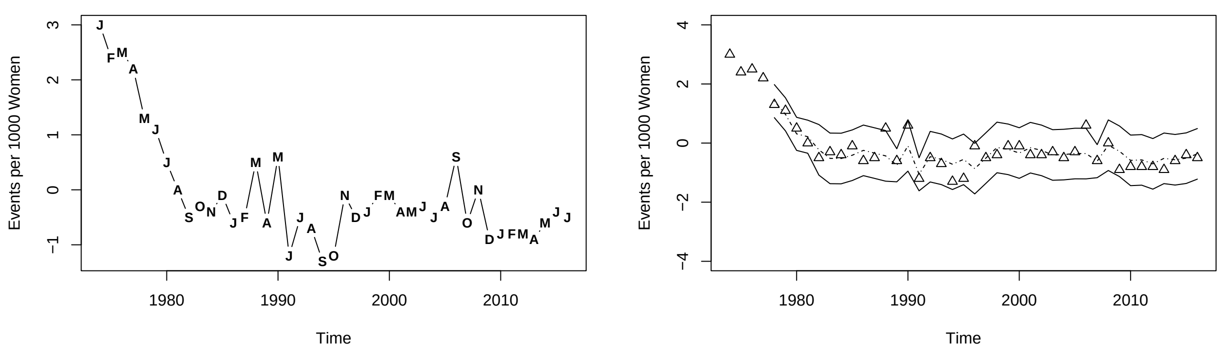
Below are the diagnostics and model selection.

```
## Series: ts_abortion_monthly
## ARIMA(4,2,3)(1,0,1)[12]
##
```

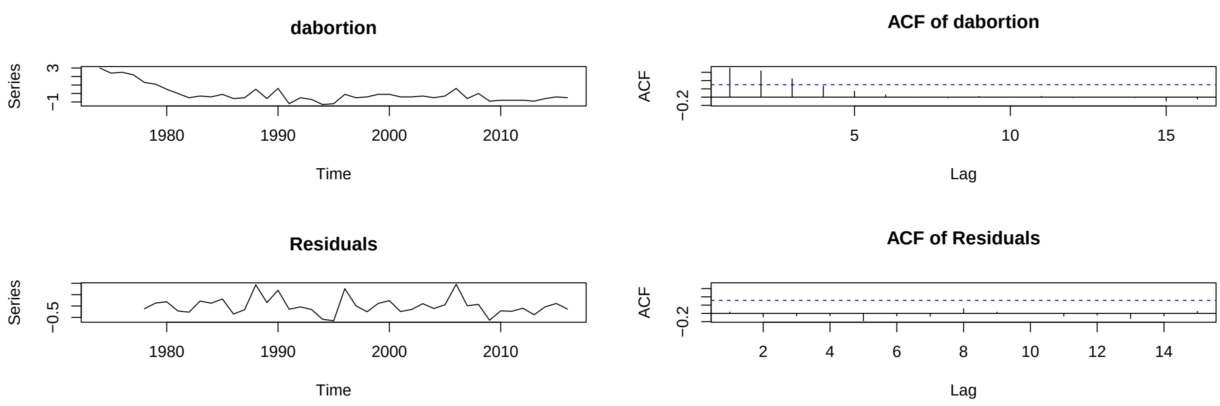
```
## Coefficients:
##           ar1     ar2     ar3     ar4      ma1     ma2     ma3    sar1     sma1
##       -0.2020  0.2077  0.1509  0.0729  -1.5831  0.2607  0.3337  0.6459  -0.4771
## s.e.   0.3977  0.0833  0.0834  0.0537   0.3964  0.7673  0.3748  0.1456   0.1632
##
## sigma^2 = 0.4979:  log likelihood = -550.65
## AIC=1121.31   AICc=1121.74   BIC=1163.75
##
## Training set error measures:
##                       ME      RMSE       MAE        MPE     MAPE      MASE
## Training set -0.02009519 0.6980633 0.5739672 -0.1471643 2.627296 0.5709148
##                      ACF1
## Training set -0.004370296
```

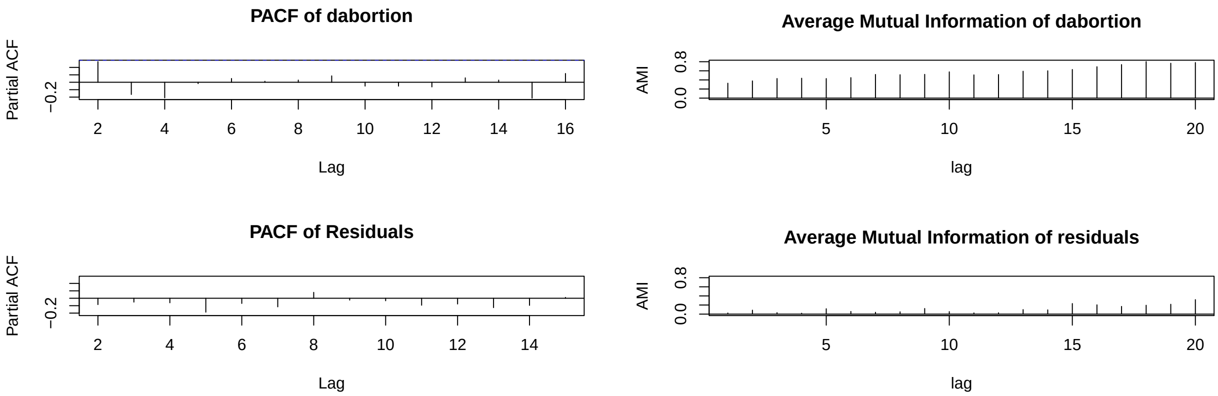
4.2 Results from Threshold Model

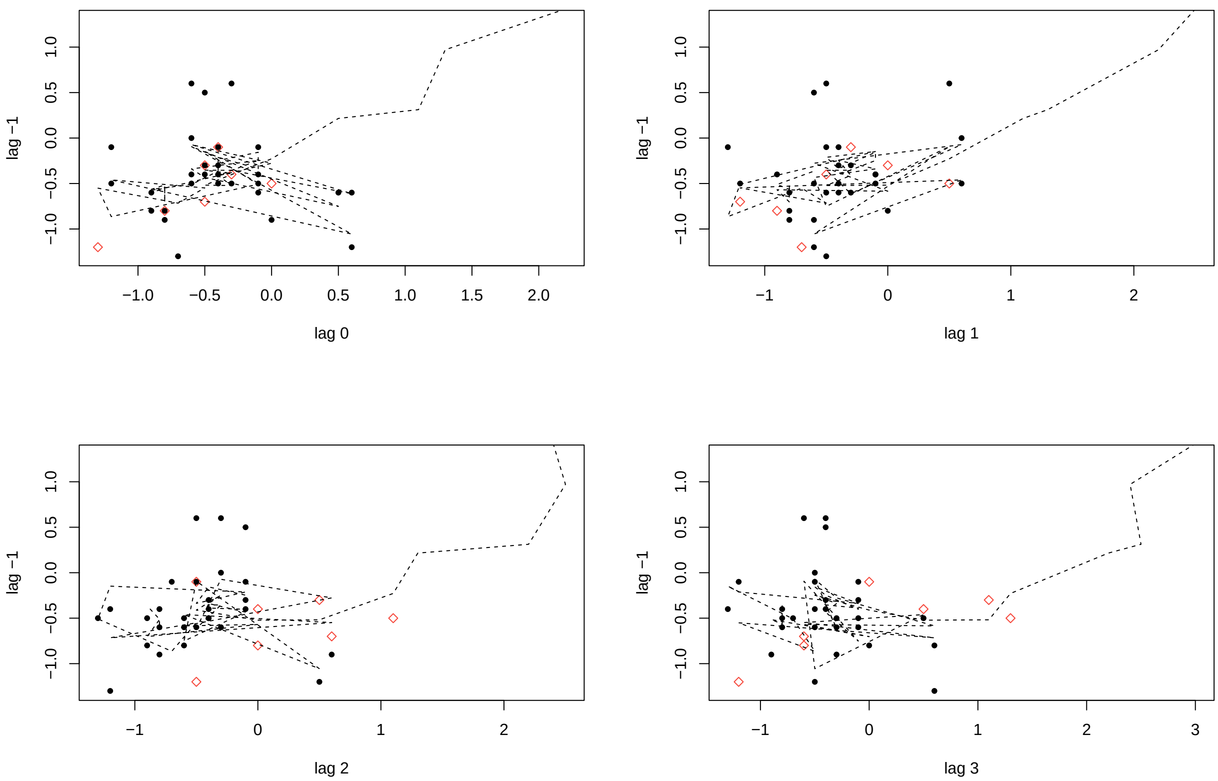
These are the plots of the original and transformed data.

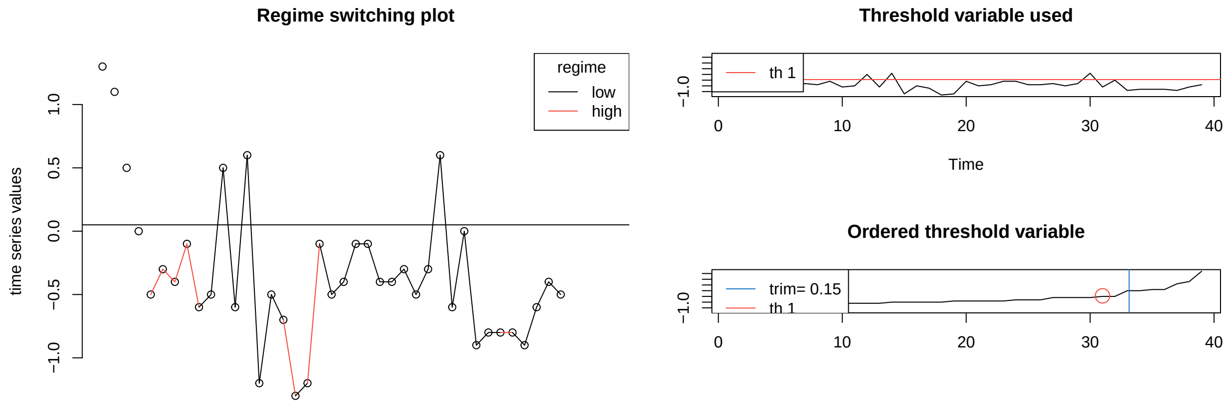


Below are the plots of the ACF, PACF, and more.




PACF of dabortion
Partial ACF
Lag
Average Mutual Information of dabortion
AMI
lag
PACF of Residuals
Partial ACF
Lag
Average Mutual Information of residuals
AMI
lag


lag −1
lag 0
lag −1
lag 1
lag −1
lag 2
lag −1
lag 3



Below are the diagnostics.

```
##
## Call:
## lm(formula = X1 ~ Z[, 2:5])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.6490 -0.2345 -0.0684  0.1143  0.9550
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       -0.17752    0.14943  -1.188   0.2452
## Z[, 2:5]stats::lag(dabortion, -1)  0.28475    0.24280   1.173   0.2511
## Z[, 2:5]stats::lag(dabortion, -2)  0.31634    0.17158   1.844   0.0762 .
## Z[, 2:5]stats::lag(dabortion, -3)  0.04947    0.16568   0.299   0.7675
## Z[, 2:5]stats::lag(dabortion, -4) -0.20238    0.14527  -1.393   0.1750
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4279 on 27 degrees of freedom
##   (7 observations deleted due to missingness)
## Multiple R-squared:  0.1783, Adjusted R-squared:  0.05653
## F-statistic: 1.464 on 4 and 27 DF,  p-value: 0.2405

##
## Call:
## lm(formula = X2 ~ Z[, 2:5])
##
## Residuals:
##       1       2       3       4      12      14      30
## -0.1191  0.1290  0.1876 -0.2159  0.1513 -0.1444  0.0115
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       -0.57641    0.33593  -1.716    0.228
## Z[, 2:5]stats::lag(dabortion, -1)  0.18275    0.36747   0.497    0.668
## Z[, 2:5]stats::lag(dabortion, -2)  0.76522    0.44828   1.707    0.230
## Z[, 2:5]stats::lag(dabortion, -3) -0.20331    0.23426  -0.868    0.477
## Z[, 2:5]stats::lag(dabortion, -4)  0.05611    0.30508   0.184    0.871
```

```
##
## Residual standard error: 0.2797 on 2 degrees of freedom
##   (32 observations deleted due to missingness)
## Multiple R-squared:  0.9703, Adjusted R-squared:  0.911
## F-statistic: 16.35 on 4 and 2 DF,  p-value: 0.05845
```

Conclusion and Future Study

By performing my chosen methods of analysis (SARIMA modeling and TAR modeling) I was able to find that the time series for abortion rates in the United States from 1973 to 2016 is non-linear and non-stationary. The Seasonal ARIMA model predicted a steady decline of abortion rates, with a predicted value of 13.92 abortions per 1000 women in the next 12 months after 2016. The ACF and PACF of the time series of the monthly abortion rates seem consistent with the original data. After fitting the Seasonal ARIMA model, I found the model to be ARIMA(4,2,3)(1,0,1)[12]. Now for TAR modeling, it can be seen from the plot of the original time series that there is a sharp decrease in the earlier years of the study with a few sharp increases throughout the following years. The transformed data, which took a first difference, matches these trends with a somewhat consistent downward trend. The plots of lag show a positive correlation but there does not seem to be any significant differences with the plots as the lag increases. The summary of the first case of the TAR modeling show that there is a 0.05 significance on the lag with k = -2. In the summaries of both cases, the p-values are 0.2405 for case 1 and 0.05845 for case 2; both p-values are greater than the threshold = 0.05 which mean that the alternative hypothesis is favored.

With the results of the analysis, I infer that changes in attitudes and laws surrounding abortions caused the downward trend over time, especially in the 1970s. It could also be attributed to the fact that sex education in American schools have improved over time, thus leading to overall less pregnancies, births, and abortions. As younger people become more educated about the logistics of reproduction, this will significantly affect their decision making surrounding pregnancy.

In a future study, we can use the same data set and focus on the pregnancy and birth rates to determine possible future trends in the United States. We can use the results of that study to make inferences about (possible) changing attitudes towards pregnancy and birth, especially with women belonging to a younger age group. Another possible focus could be how these rates differ throughout each state, taking the general political and religious affiliation of each state into consideration. Time series studies done on pregnancy, birth, and abortion rates are very important to the future of reproductive rights and sex education in the United States as well as the future of the US human population, which affects many other industries such as the job industry. With this in mind, we can use time series analysis to learn more about behavioral patterns and collective attitudes towards abortion, thus creating a safer and better environment for women and people who are able to reproduce.

References

Penn State University

https://online.stat.psu.edu/stat510/lesson/13/13.2

Wikipedia

https://en.wikipedia.org/wiki/SETAR_(model)

Statistics And Its Interface

https://users.ssc.wisc.edu/~bhansen/papers/saii_11.pdf

Hopkins Bloomberg Public Health Magazine

https://magazine.jhsph.edu/2022/brief-history-abortion-us

Time Series Analysis and Its Applications, Third Edition, by Shumway and Stoffer

the class textbook