# I590 - Time Series Analysis - Final Report

*James Provost*

*April 28, 2019*

## Data Description

This report will explore the dataset birth from the astsa library (Applied Statistical Time Series Analysis - see https://www.rdocumentation.org/packages/astsa/versions/1.8/topics/birth for additional documentation).

The dataset contains number of live births (in thousands) per month for the United States between January 1948 and January 1979.

This table shows some basic statistics of these births.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   238.0   284.0   310.0   310.9   336.0   399.0
```

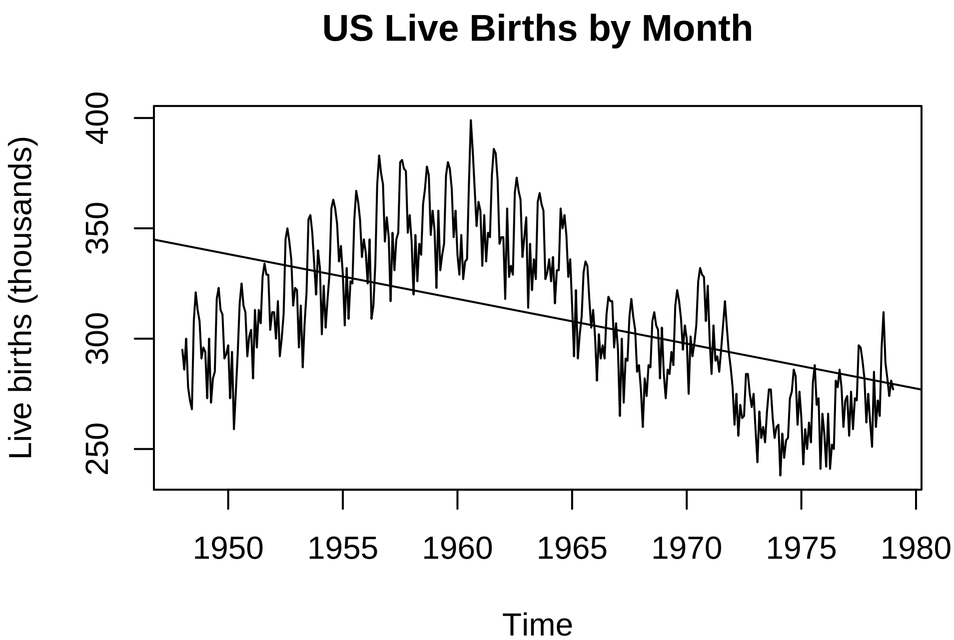
In addition, the standard deviation of live births is 35.3. To confirm the contents of the time series, here are the start, end and frequency of the dataset, respectively.

```
## [1] 1948    1
```

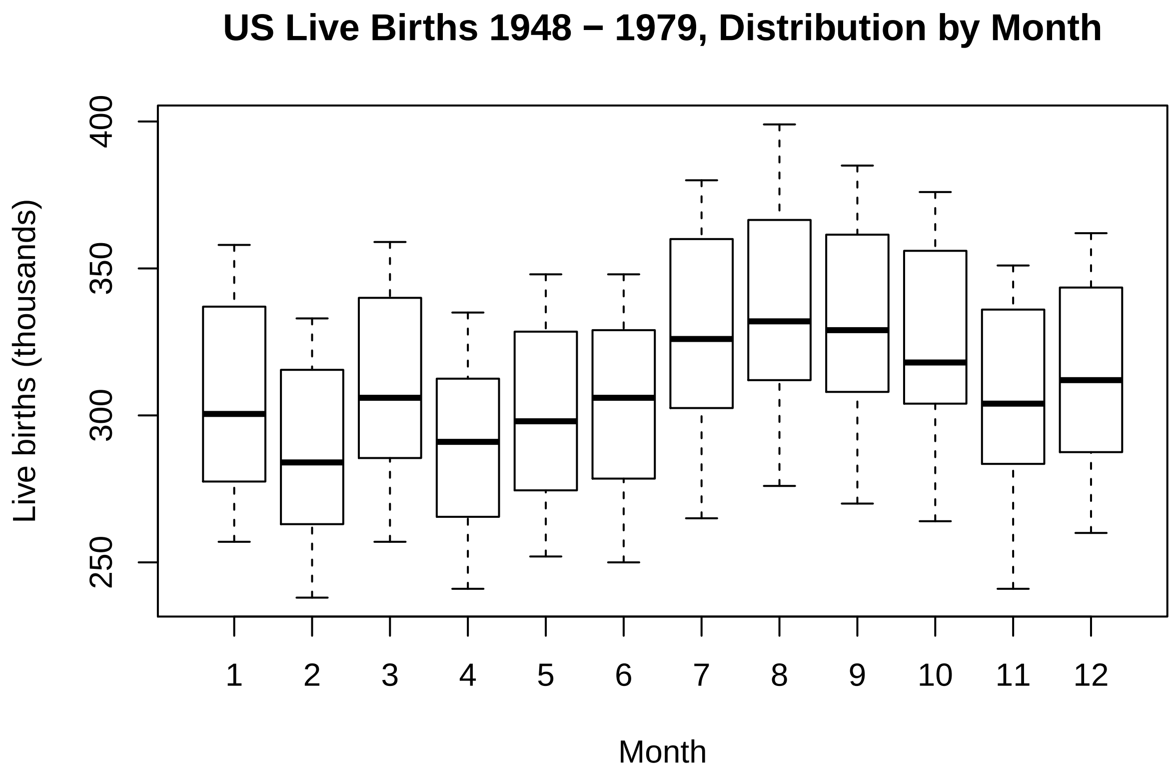
```
## [1] 1979    1
```

```
## [1] 12
```

**Data Exploration**




This plot shows the overall time series across months. Clearly there is a seasonal aspect that is repeated throughout and a trend that rises through the 1950s, peaking in the early 1960s before dropping, with a couple of upward trends around the late 1960s and again in the late 1970s. The overall linear trend line shows as negative across the entire time series.

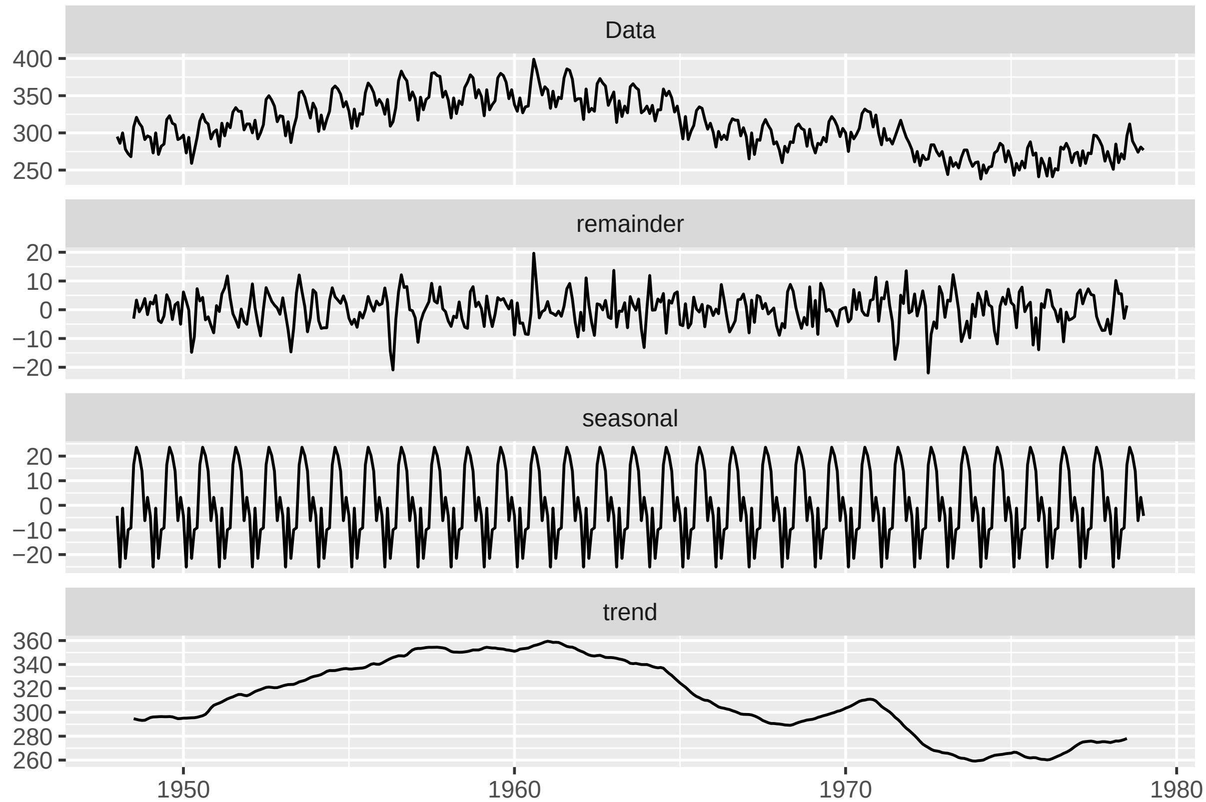



This plot shows the distribution of the values by each month of the year and gives a perspective of the seasonal aspect of the data. The lowest births appear to be early in the year, in February and then April, with the highest live births in the fall around September and October. The variance in each month is fairly consistent, with none of the months standing out significantly.

As we look at the data, the variance is not growing over time, nor is there a growing trend, so there does not seem like an obvious transformation to make to the data, so we will not do that here.

### Data Decomposition

Next we will decompose the data set into overall trend, seasonal trend and random.




From these graphs we can clearly see the overall trend is not linear and is similar to our earlier description. We also see that there is a significant seasonal trend. The remainder appears to be fairly stationary, with a mean around zero and fairly consistent variability.

## Regression

We'll now build a regression model based purely on the birth data, since there is no other data point with which to form a relationship, and show a summary of the coefficients.

```
##
## Call:
## lm(formula = birth ~ time(birth), na.action = NULL)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -78.810 -22.411  -1.172  21.602  82.170
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 4297.3111   342.2257   12.56   <2e-16 ***
## time(birth)   -2.0303     0.1743  -11.65   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 30.2 on 371 degrees of freedom
## Multiple R-squared:  0.2678, Adjusted R-squared:  0.2658
## F-statistic: 135.7 on 1 and 371 DF,  p-value: < 2.2e-16
```

We can see that there is a small negative coefficient for the time component, which indicates a downward trend. Note that this is the formula for the trend line that we plotted earlier and, as we noted earlier, that this data isn't well modeled by a linear trend, and therefore this model isn't a very good one at fully describing this data, especially if we wished to ultimately use it for prediction.

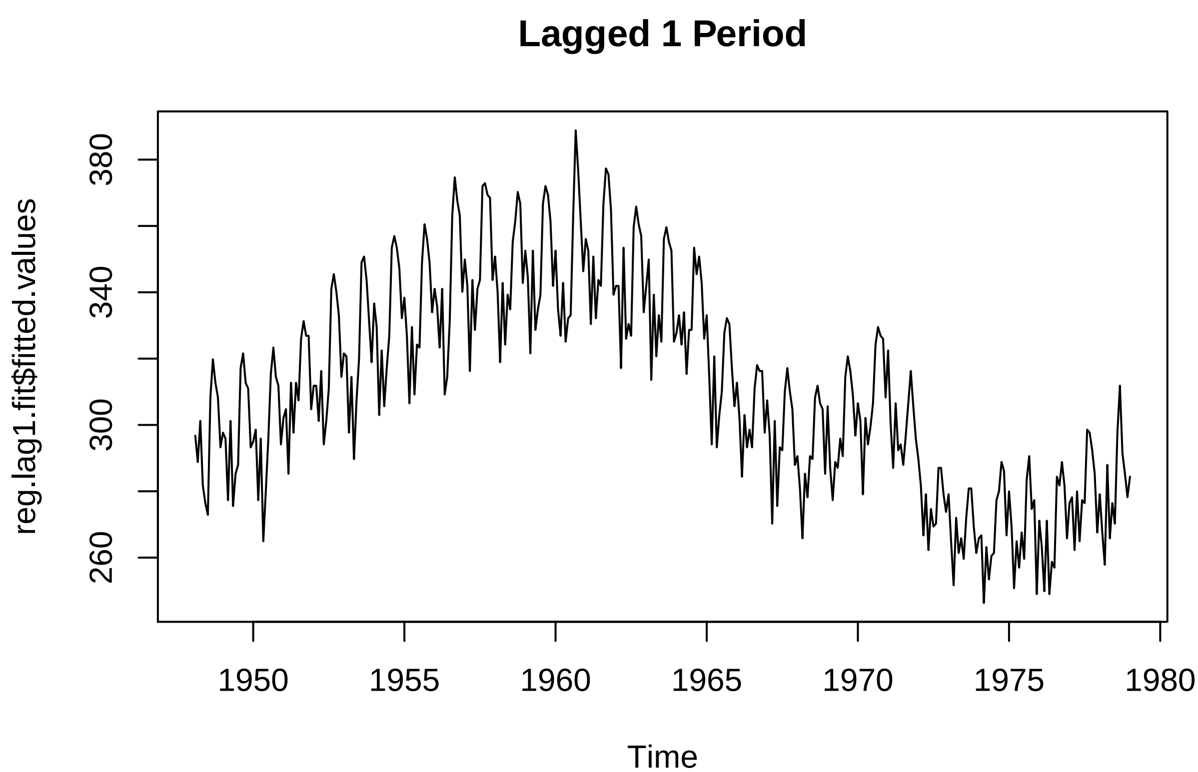
Because we know that this model has a seasonal component to it, we'll build three other models, one with a lag of one period, one with a lag of 12 periods and one with a lag of both one period and 12 periods (a multi seasonal regression).

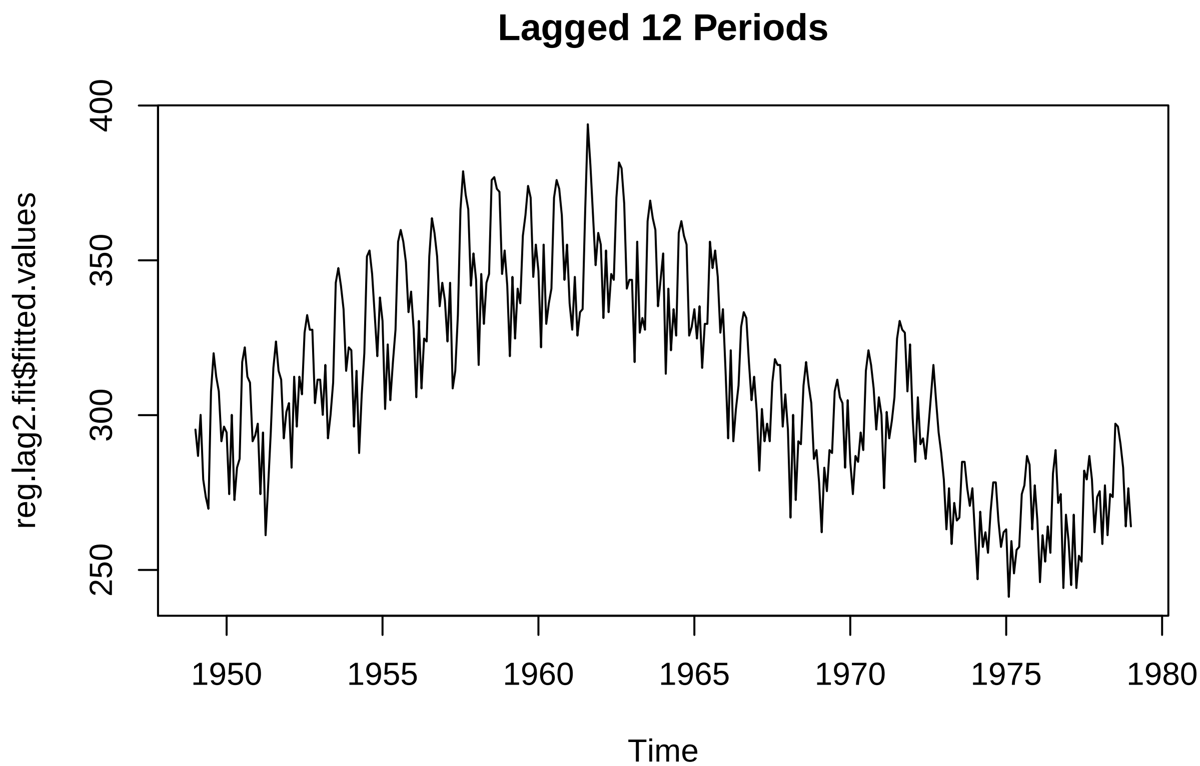
Then we'll print a summary of the models, plot the fitted values of the models and print the AIC values of the three models.

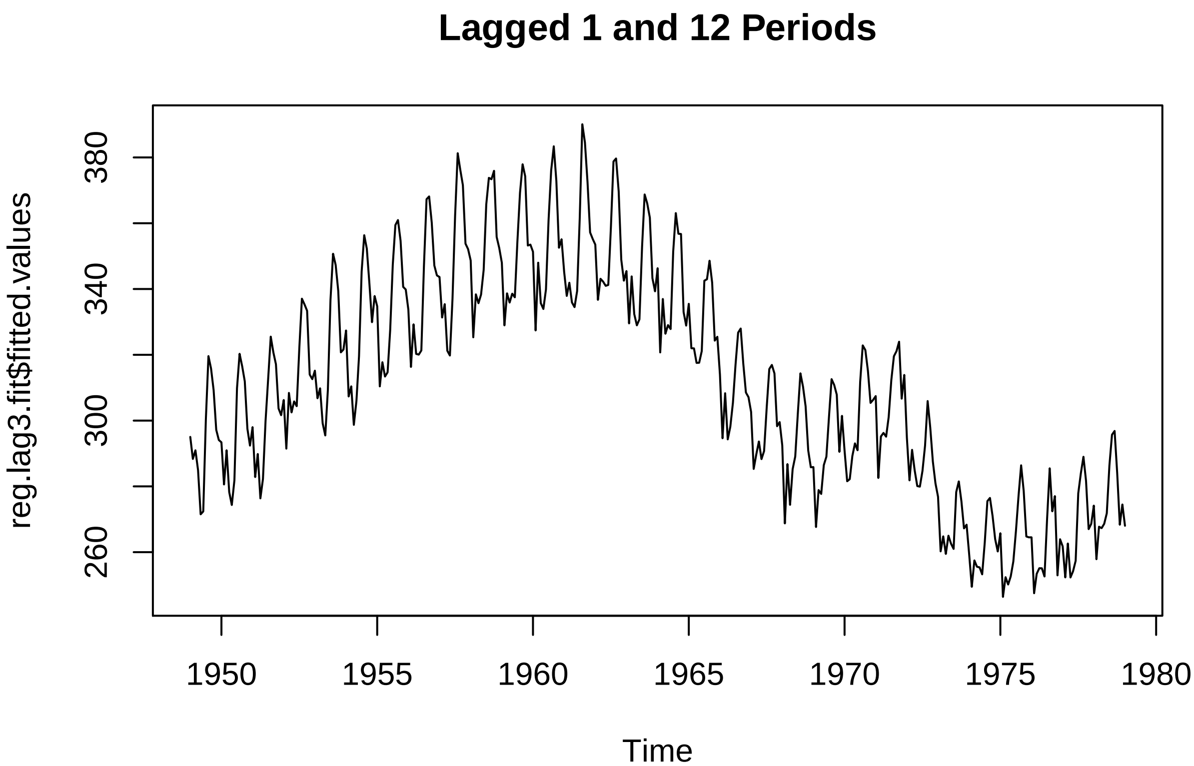
```
##
## Time series regression with "ts" data:
## Start = 1948(2), End = 1979(1)
##
## Call:
## dynlm(formula = birth ~ L(birth, 1))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -36.905 -12.171   0.112  10.754  41.853
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  35.6862     7.6068   4.691 3.83e-06 ***
## L(birth, 1)   0.8851     0.0243  36.418  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.5 on 370 degrees of freedom
## Multiple R-squared:  0.7819, Adjusted R-squared:  0.7813
## F-statistic:  1326 on 1 and 370 DF,  p-value: < 2.2e-16
```

```
##
## Time series regression with "ts" data:
## Start = 1949(1), End = 1979(1)
##
## Call:
## dynlm(formula = birth ~ L(birth, 12))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -35.844  -7.949   0.585   8.559  35.663
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  15.67027    5.80130   2.701  0.00724 **
## L(birth, 12)  0.94807    0.01848  51.308  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.35 on 359 degrees of freedom
## Multiple R-squared:   0.88,  Adjusted R-squared:  0.8797
## F-statistic:  2632 on 1 and 359 DF,  p-value: < 2.2e-16
```

```
##
## Time series regression with "ts" data:
## Start = 1949(1), End = 1979(1)
##
## Call:
## dynlm(formula = birth ~ L(birth, 1) + L(birth, 12))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -29.9942  -7.3048  -0.1239   7.5561  30.7239
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.26995    5.22303   0.243    0.808
## L(birth, 1)   0.32316    0.03001  10.769   <2e-16 ***
## L(birth, 12)  0.67156    0.03030  22.166   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.74 on 358 degrees of freedom
## Multiple R-squared:  0.9094, Adjusted R-squared:  0.9088
## F-statistic:  1796 on 2 and 358 DF,  p-value: < 2.2e-16
```




Lagged 12 Periods
reg.lag2.fit$fitted.values
400
350
300
250
1950
1955
1960
1965
1970
1975
1980
Time

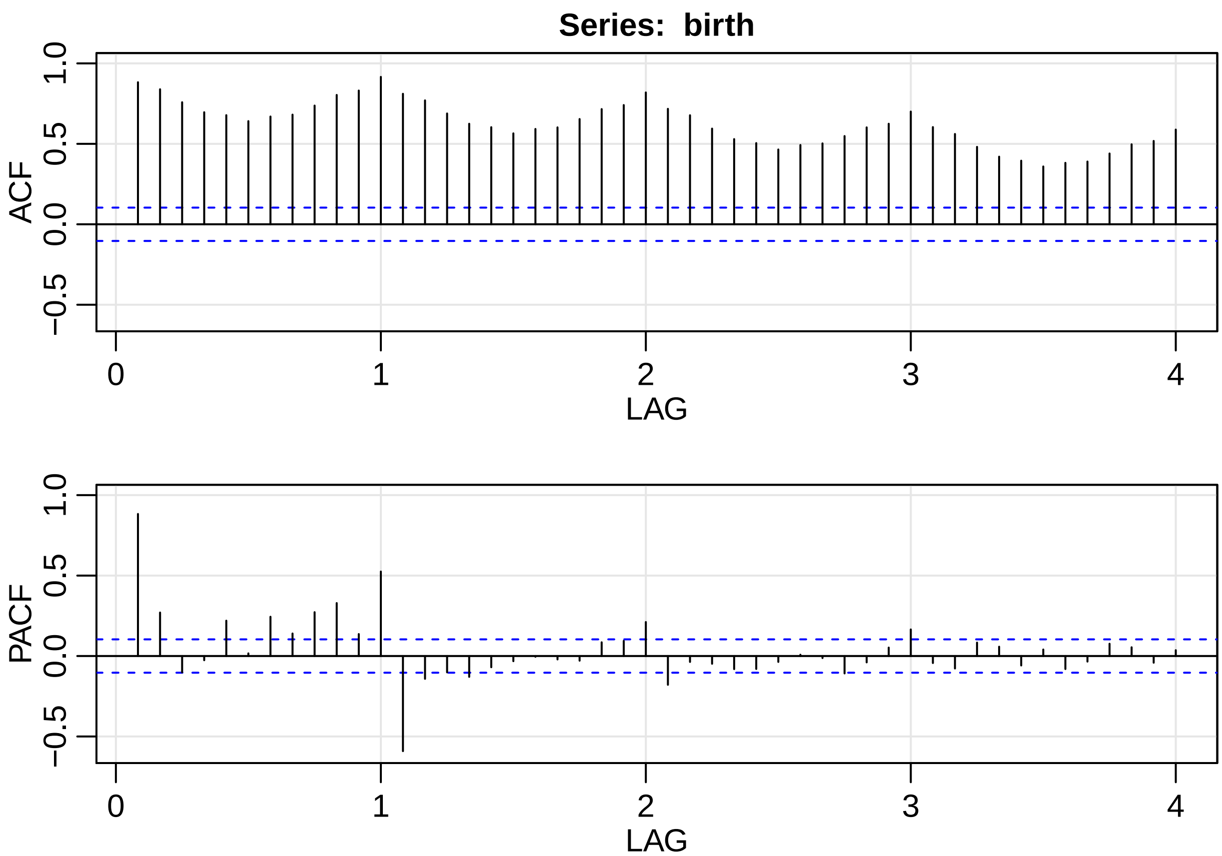



```
## [1] 3145.536
## [1] 2843.092
## [1] 2743.797
```

We can see that the R-squared value gets better and better with each model, as does the AIC value. This suggests the model with both a one period lag and a 12 period lag is the best model and suggests the data has a multi seasonal component.

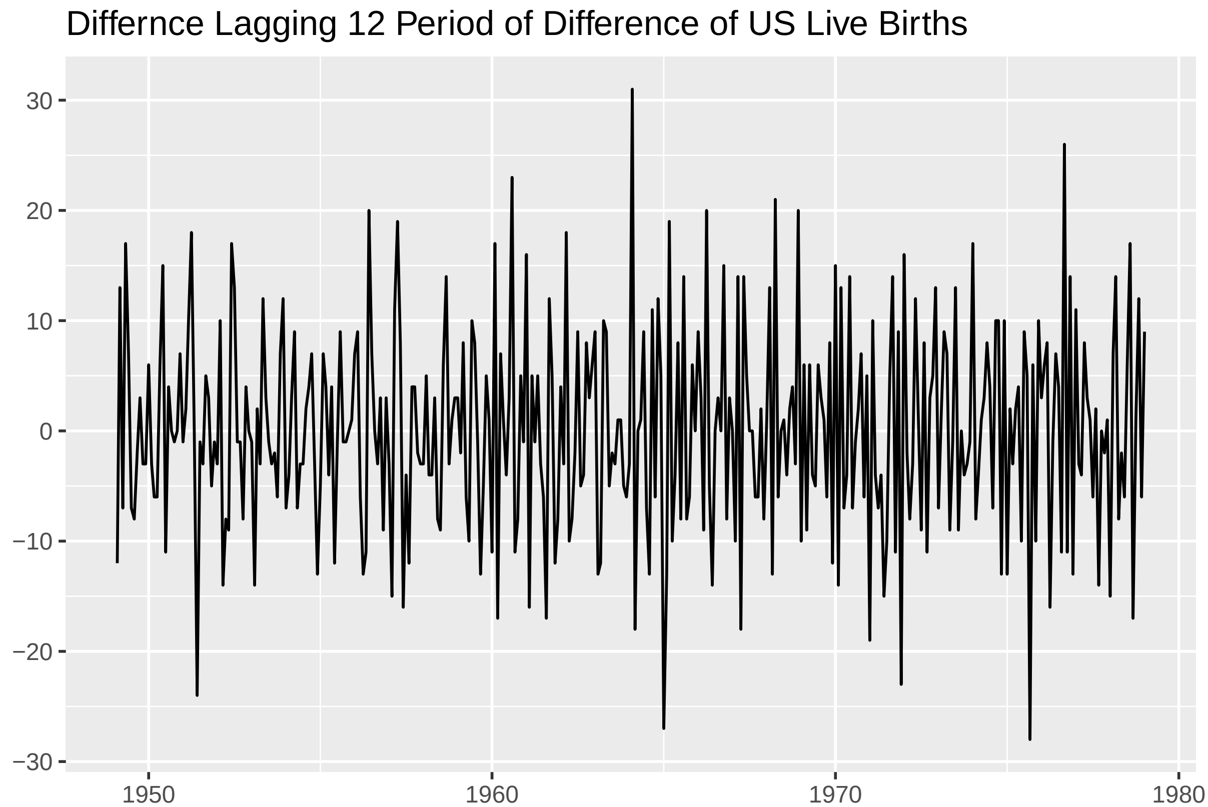
### ARIMA model

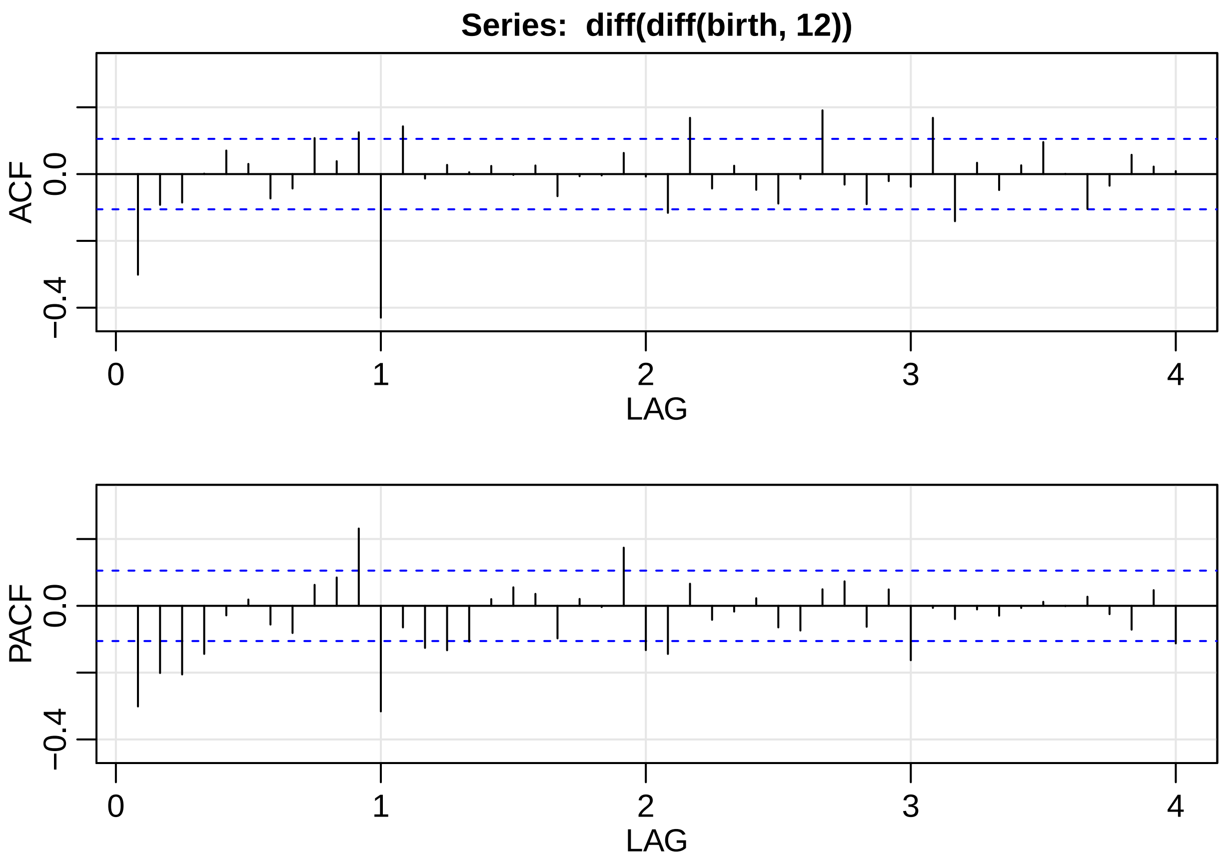
We'll run auto correlation and partial auto correlation functions on the raw dataset.




These graphs show a long trailing ACF with peaks around the yearly mark while the PACF cuts off more quickly while still having some response around the one year mark. This is clearly a time series with a significant seasonal component (as we saw earlier when we decomposed the data) and that yearly peak again suggests this might have a multi seasonal model.

We'll take the second difference of the data, lagging 12 periods for the second difference, plot the data and look at the ACF and PACF of the result.


Differnce Lagging 12 Period of Difference of US Live Births
30
20
10
0
−10
−20
−30
1950
1960
1970
1980




We see that the plot of the data looks fairly stationary. We also see that there is a significant ACF response at 1 year and a only a little bit later while the PACF has a short cut off and then a response again at one year. This is further evidence that a seasonal model might fit this data set.

## Model Diagnostics

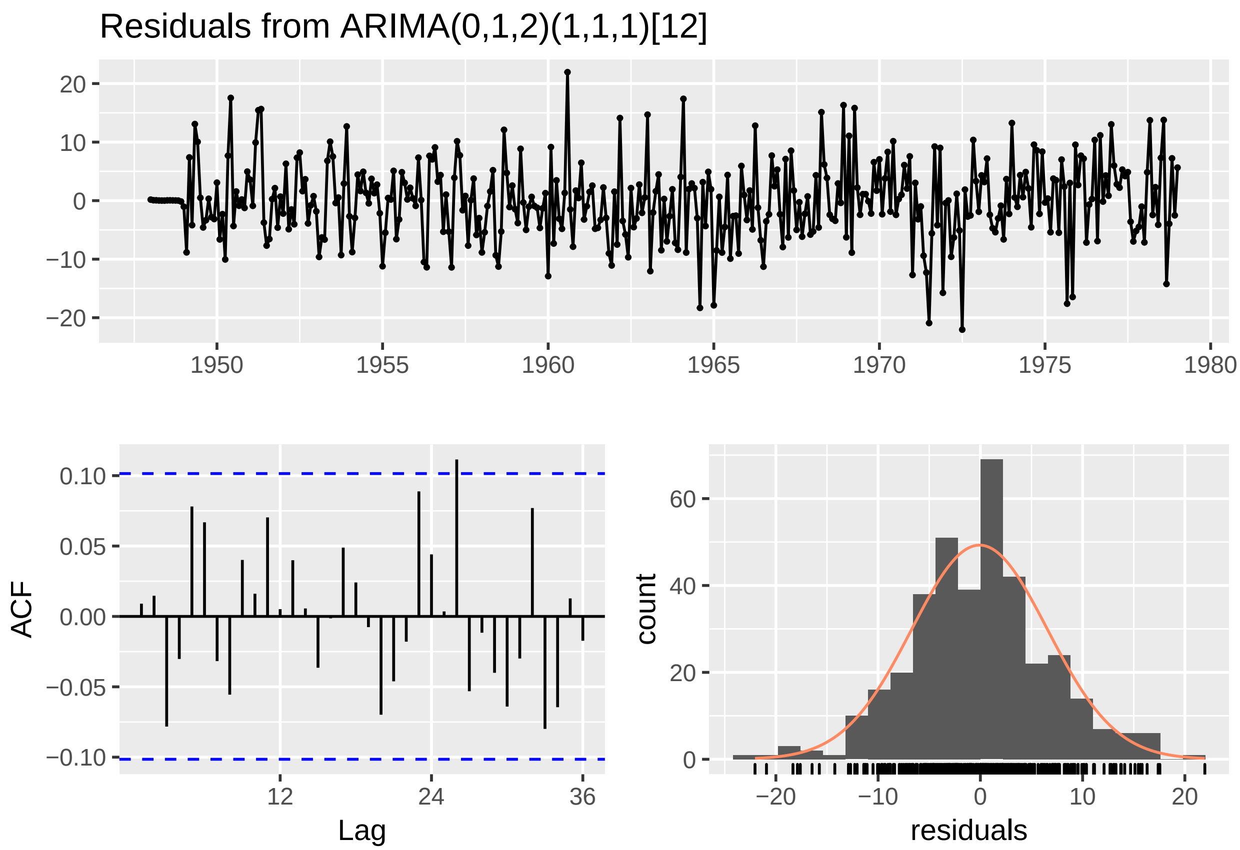
Let's attempt to build a seasonal ARIMA model based on the time series. We'll use auto.arima() to save us the computation time of choosing many different combinations of parameters.

```
## Series: birth
## ARIMA(0,1,2)(1,1,1)[12]
##
## Coefficients:
##           ma1      ma2    sar1     sma1
##       -0.3984  -0.1632  0.1018  -0.8434
## s.e.   0.0512   0.0486  0.0713   0.0476
##
## sigma^2 estimated as 46.1:  log likelihood=-1204.93
## AIC=2419.86   AICc=2420.03   BIC=2439.29
##
## Training set error measures:
##                       ME     RMSE      MAE         MPE     MAPE      MASE
## Training set -0.07998151 6.633018 5.048776 -0.02741433 1.656549 0.5145703
##                     ACF1
## Training set 0.009043143
```

The auto.arima() function suggested a mixed seasonal ARIMA model with Auto Regressive, Differencing and

Moving Average components.

Let's evaluate the residuals of this model to see if it appears to be a good model.




```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(0,1,2)(1,1,1)[12]
## Q* = 20.184, df = 20, p-value = 0.4465
##
## Model df: 4.   Total lags used: 24
```

The residuals appear to be reasonably around zero, and the ACF shows fairly tight response and the distribution is reasonably normal. Furthermore, the p-value is comfortably above 0.05 and suggests that the residuals are not correlated. Finally, the AIC score for this model beats our best regression model, so we can conclude that this is a reasonably good model.