Time Series Project #2

Anne Vierse
December 20, 2018
STA 9701, Professor Zeda Li

Table of Contents

Data Overview

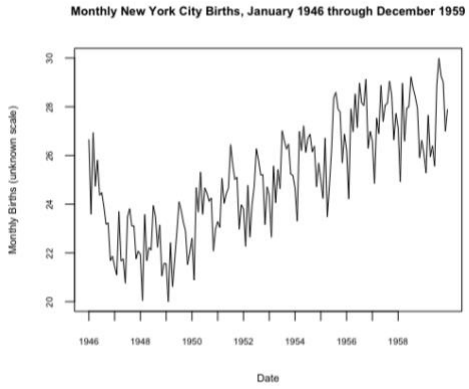
For Project #2, I chose to re-analyze my data from Project #1, a count of Monthly New York City births from Jan 1946 to December 1959[1], using a seasonal ARIMA model. In Project #1, I determined that a seasonal model would probably be a better fit, but I chose to try and fit an AR(12) model instead. In this project, I will pursue my curiosity and compare the AR(12) fit to a SARIMA model.

Additionally, my data has an early inconsistency that was determined, with some research, to possibly be due to a New York City blackout in the mid 1940's. In Project #1 I removed this data and chose to create a model without the intervention. For Project #2, after fitting the seasonal model and comparing that forecast with my earlier AR(12) model, I will also perform intervention analysis with the original data and my seasonal model. I am interested in continuing this analysis to hopefully arrive at more answers to questions raised by my historical background and now statistical investment in Project #1.

Plotting the Data

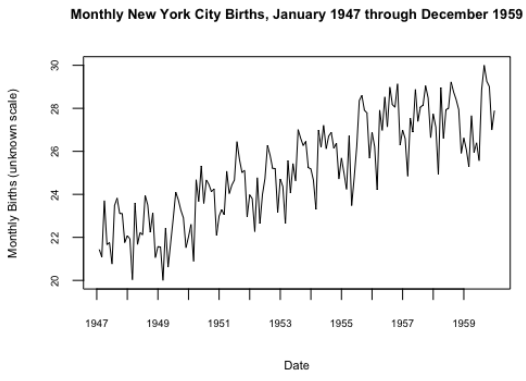
Below is a plot of the original data (Figure 1).

Figure 1:



As noted in Project #1, there is an obvious positive linear trend. However, the steep drop in births during the first year (1946) throws off the overarching trend. I will remove the twelve data points from 1946 to fit my seasonal model, and then use that model to perform intervention analysis. See the newly plotted data below in Figure 2 and the associated ACF and PACF plots in Figures 3 and 4.

Figure 2:



[1] SOURCE WEBSITE: https://datamarket.com/data/set/22nv/monthly-new-york-city-births-unknown-scale-jan-1946-dec-1959#!ds=22nv&display=line

Figure 3:

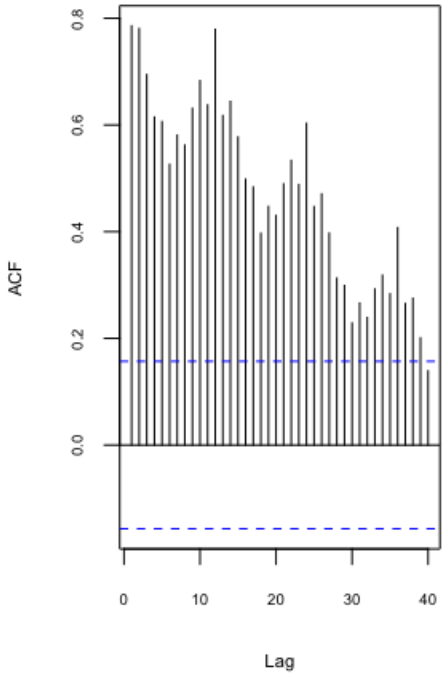
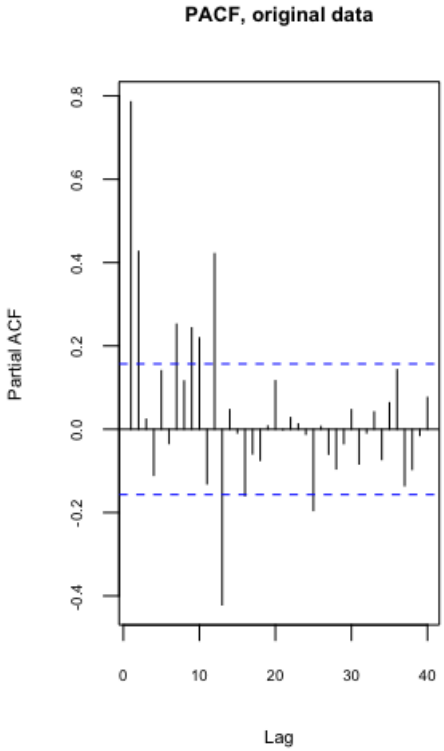


Figure 4:



Looking at Figure 2, it can be seen that variance is fairly stable. A transformation to stabilize variance is therefore not necessary, although, as noted before, the linear trend must be removed to stabilize the mean. This is confirmed by the ACF and PACF plots where it is noted that the data is not stationary (Figures 3 and 4).

Creating Stationary Data

Looking at the ACF and PACF plots, the possible seasonality at lag 12 is already evident. First, however, I will stabilize the mean using regular differencing. Figure 5 displays the plot of the differenced data, along with ACF and PACF plots.

Figure 5:

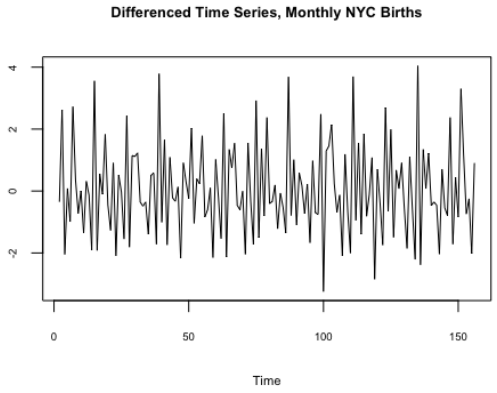


Figure 6:

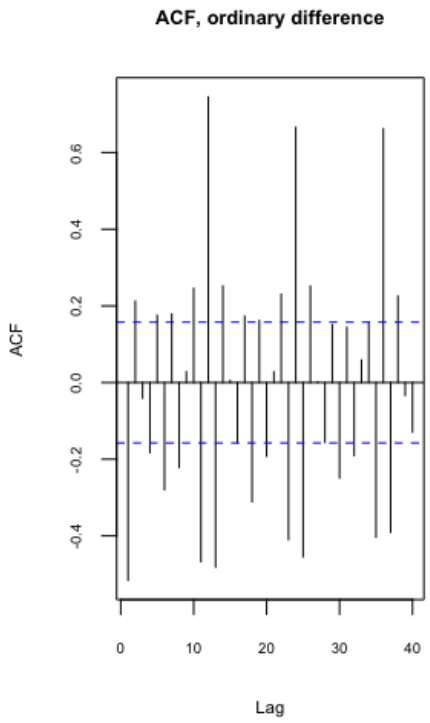
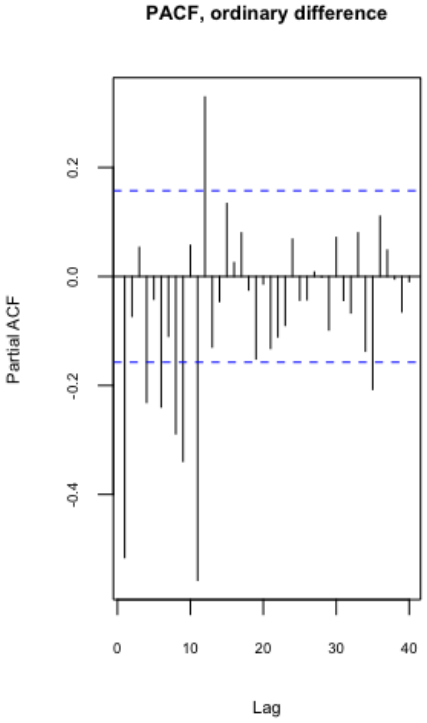


Figure 7:



Looking at the ACF and PACF plots, the spike at lag 12 in the ACF signals seasonality. This would make sense, if we consider that different months might yield more births, due to the seasons and various other cultural phenomenon. Figures 8 and 9 below show this data then differenced at lag 12.

Figure 8:

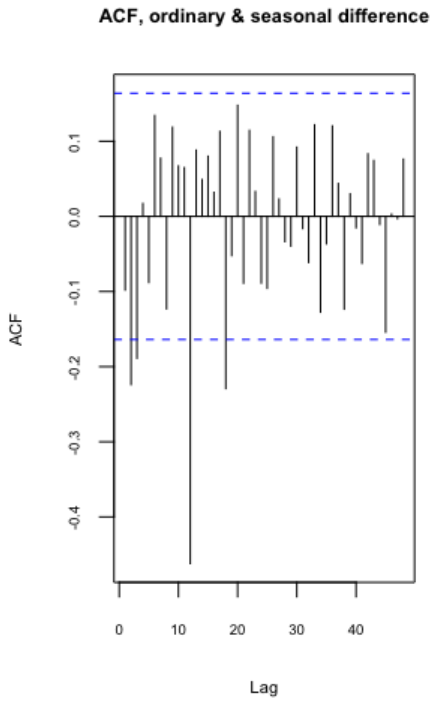
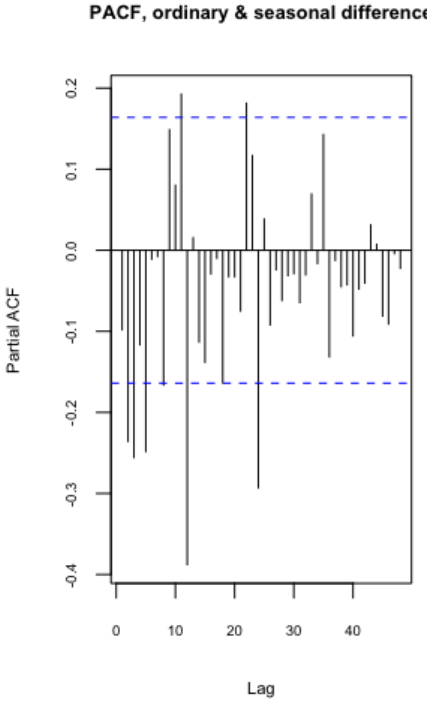


Figure 9:



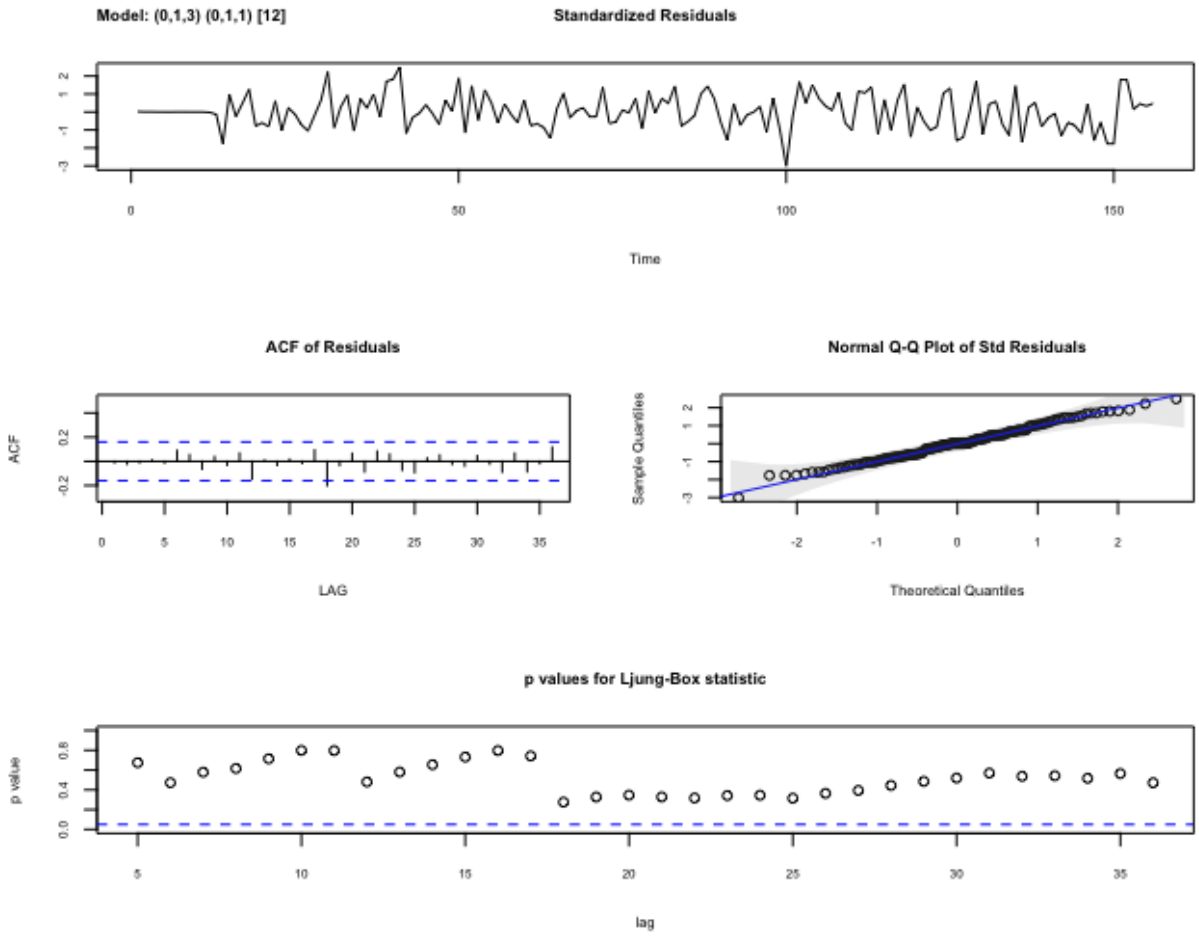
It is noted that the null was rejected using the Augmented Dickey-Fuller Test meaning this transformation is not a random walk. The ACF and PACF plots look stationary and can be used to fit a seasonal ARIMA model.

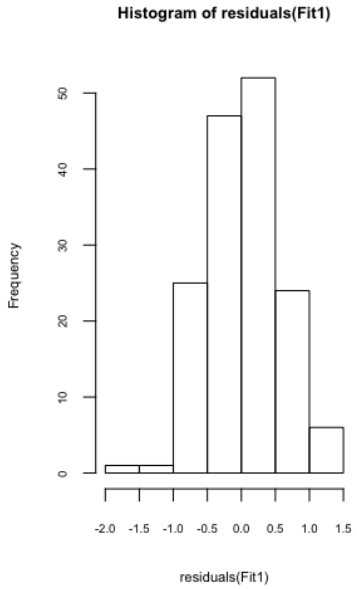
Fitting the Model

Based on the ACF and PACF plots, there is a possible seasonal spike at lag 12. The non-seasonal ACF cuts off at lag 2 or 3 and the non-seasonal PACF slowly decreases. The seasonal trend cuts off at 12 in the ACF plot and slowly decreases in the PACF plot. This signifies possible MA models for both the seasonal and non-seasonal ARMIA components, so I first fit an SARIMA $(0, 1, 3) \times (0, 1, 1)_{12}$. The residual plots do not look too bad for this fit, although there is some slight heteroskedasticity in the Q-Q plot. I tried several other models to compare, including adding an AR term because the non-seasonal ACF has a spike at lag 18, which may not necessarily correlate to an MA term. The model SARIMA $(1, 1, 2) \times (0, 1, 2)_{12}$ has comparable diagnostic plots.

I decided to proceed with the SARIMA $(0, 1, 3) \times (0, 1, 1)_{12}$ model because adding the AR term in complicates the model – simpler is better. The diagnostic plots for this fit are below in Figure 10.

Figure 10:





$ttable

	Estimate	SE	t.value	p.value
ma1	-0.2130	0.0847	-2.5141	0.0131
ma2	-0.3507	0.0774	-4.5288	0.0000
ma3	-0.2038	0.0850	-2.3964	0.0179
sma1	-1.0000	0.1397	-7.1584	0.0000

Coefficients:

	ma1	ma2	ma3	sma1
	-0.2130	-0.3507	-0.2038	-1.0000
s.e.	0.0847	0.0774	0.0850	0.1397

sigma^2 estimated as 0.3289: log likelihood=-137.32
AIC=284.64 AICc=285.08 BIC=299.46

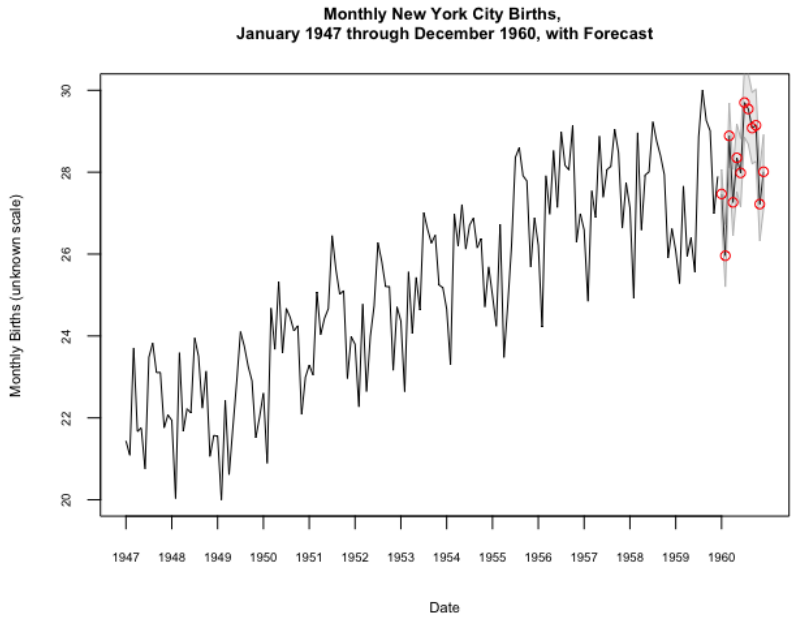
The resulting equation for this fitted model is:

$$(1-B)\,(1-B^{12})\,x_t = (1-0.2130B)\,(1-0.3507B^2)\,(1-0.2038B^3)\,(1-B^{12})\,\omega_t$$

Forecasting the Seasonal Model

Using the SARIMA $(0, 1, 3) \times (0, 1, 1)_{12}$ model, the next twelve values are forecasted. Below is the forecast plot and values table (Figure 11).

Figure 11:



1960 Forecast Values

Jan	Feb	Mar	Apr	May	June
27.46771	25.95984	28.8858	27.26243	28.35135	27.98282
July	**Aug**	**Sept**	**Oct**	**Nov**	**Dec**
29.70005	29.54374	29.07751	29.14420	27.22051	28.01081

Model Comparison

To test the model, I used it to forecast 1959, and then compared the actual values.

SARMIA (0, 1, 3) x $(0, 1, 1)_{12}$
1959 Forecast and Actual Values

	Jan	Feb	Mar	Apr	May	June
Prediction	26.9811	26.66217	26.71119	26.55183	26.69665	26.67937
Actual	26.076	25.286	27.66	25.951	26.398	25.565
Difference	*0.9051*	*1.37617*	*-0.94881*	*0.60083*	*0.29865*	*1.11437*
	July	Aug	Sept	Oct	Nov	Dec
Prediction	26.44601	26.5081	26.61537	26.53992	26.52019	26.65369
Actual	28.865	30	29.261	29.012	26.992	27.897
Difference	*-2.41899*	*-3.4919*	*-2.64563*	*-2.47208*	*-0.47181*	*-1.24331*

Although, compared to the forecasts made in Project #1 using the AR(12) model (below), the SARIMA prediction is closer to the actual for only 7 of the months, many of the differences less extreme. I would chose to use the SARIMA model over the AR(12) model to fit the data based on these forecasts and also because the SARIMA model is simpler in that it has fewer terms than the AR(12).

AR(12)
1959 Forecast and Actual Values

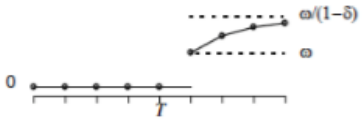
	Jan	Feb	Mar	Apr	May	June
Prediction	27.1193	24.8766	30.0528	24.8930	28.0842	26.4909
Actual	26.0760	25.2860	27.6600	25.9510	26.3980	25.5650
Difference	*1.0433*	*-0.4094*	*2.3928*	*-1.0580*	*1.6862*	*0.9259*
	July	Aug	Sept	Oct	Nov	Dec
Prediction	27.4657	26.0568	26.1670	26.6262	24.9675	27.2648
Actual	28.8650	30.0000	29.2610	29.0120	26.9920	27.8970
Difference	*-1.3993*	*-3.9432*	*-3.0940*	*-2.3858*	*-2.0246*	*-0.6322*

Intervention Analysis

Now, I will go back to the original data, including the 1946 year of data, and perform intervention analysis. Based on the original data plot and the way that I would predict effects might be felt from a sudden spike in births, I determined to use the below step function for my intervention term:

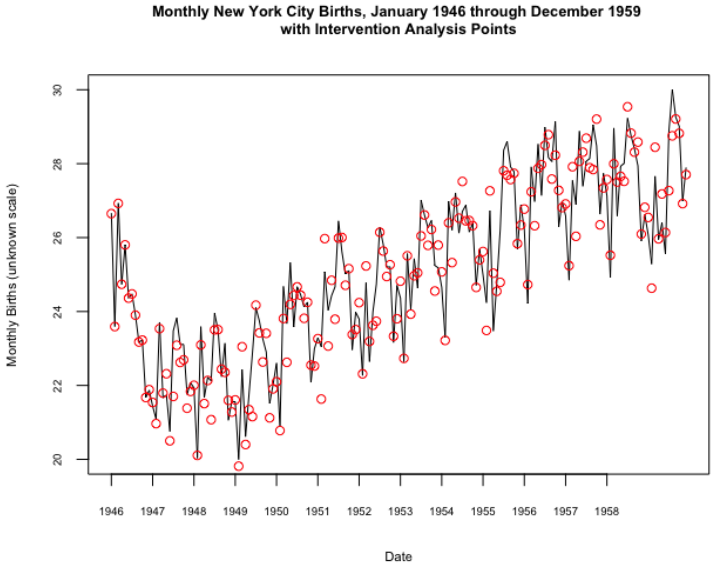
$$\frac{\omega B}{1-\delta B} S_t^{(T)}$$

This function can be used to represent the spike in births before the trend falls back, as modeled below. The step function will then take place directly after the intervention.



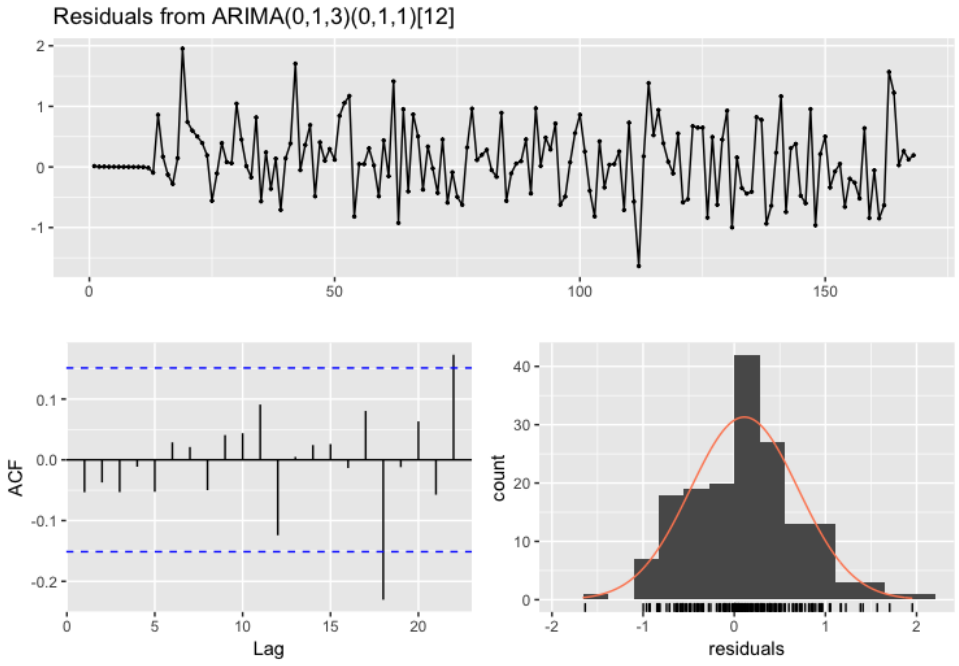
Using arimax in R, I fit my SARIMA (0, 1, 3) x $(0, 1, 1)_{12}$ model with the intervention term. The resulting plot is below (Figure 12). The majority of the points fit along the plot of the original data.

Figure 12:



The residual plots look fairly normal:

Figure 13:



The estimated parameters (s.e.) are:

$\widehat{\theta}_1$ = -0.0751 (0.083)
$\widehat{\theta}_2$ = -0.2443 (0.078)
$\widehat{\theta}_3$ = -0.174 (0.0841)
$\widehat{\Theta}$ = -0.9048 (0.1159)
ω = 1.2034 (0.5059)
δ = -0.6247 (0.1363)

The resulting equation for this model is:

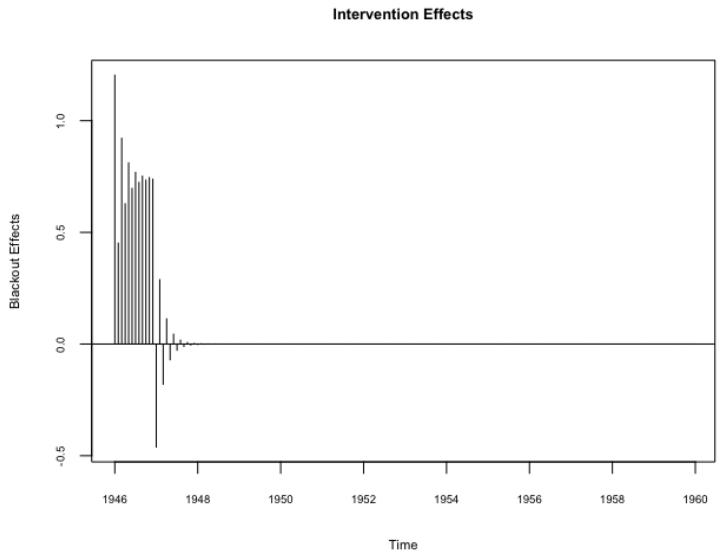
$$y_t = y_{t-1} + y_{t-12} + \varepsilon_t - 0.0751\varepsilon_{t-1} - 0.2443\varepsilon_{t-2} - 0.174\varepsilon_{t-3} - 0.9048\varepsilon_{t-12} + [(1.2034B)/(1+0.6247B)]*S_t^{(T)}$$

As stated earlier in the project, the spike in births was hypothesized to be due to a blackout event in New York City. The effects of the blackout in births are plotted in Figure 14. As can be interpolated from the plot of the original data, the effects of the blackout in number of births faded out by 1948. This makes sense when looking at how the original plotted data behaves and considering that logically, birth rates will eventually normalize, barring any other cultural abnormalities.

Births k months later were lowered by:

$$\{1-\exp(1.2034 \times -0.6247^k)\} \times 100\%$$

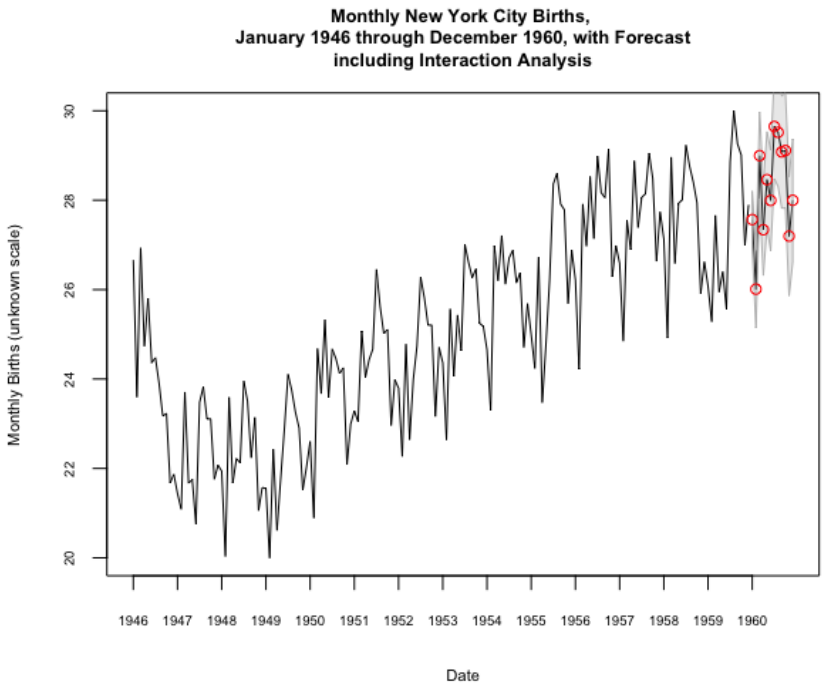
Figure 14:



Forecasting after Intervention Analysis

Using the SARIMA model and incorporating the intervention term, the updated forecast plot and values are below in Figure 15.

Figure 15:



1960 Forecast Values

Jan	**Feb**	**Mar**	**Apr**	**May**	**June**
27.56666	26.011	28.99712	27.33989	28.46005	27.99494
July	**Aug**	**Sept**	**Oct**	**Nov**	**Dec**
29.64875	29.51842	29.08068	29.11198	27.19642	27.99877

Performing the forecast test again to compare 1959 predicted values with the actual, the values are, for the most part, closer to the actual values and with less extreme differences then my previous two models.

SARMIA (0, 1, 3) x (0, 1, 1)$_{12}$ with Intervention Analysis
1959 Forecast and Actual Values

	Jan	Feb	Mar	Apr	May	June
Prediction	26.57267	25.08598	28.25108	26.60491	27.83698	27.41332
Actual	26.076	25.286	27.66	25.951	26.398	25.565
Difference	*0.49667*	*-0.20002*	*0.59108*	*0.65391*	*1.43898*	*1.84832*
	July	Aug	Sept	Oct	Nov	Dec
Prediction	28.85807	28.57603	28.20502	28.25795	26.3363	27.13487
Actual	28.865	30	29.261	29.012	26.992	27.897
Difference	*-0.00693*	*-1.42397*	*-1.05598*	*-0.75405*	*-0.6557*	*-0.76213*

R-Code

The R-code used for this project is copied below.

```
#load libraries for Project 1
library(astsa)
library(forecast)
library(TSA)
library(aTSA)

#https://datamarket.com/data/set/22nv/monthly-new-york-city-births-unknown-scale-jan-1946-dec-1959#!ds=22nv&display=line
#import data set - monthly nyc births, January 1946 through December 1959 (unknown scale)
NYC_births_o<-read_csv("~/Documents/Baruch/monthly_nyc_births.csv")
NYC_births_ots<-ts(NYC_births_o$Births)

#plot time series
par(cex.axis = 0.6, cex.lab = 0.7, cex.main = 0.8)
plot(NYC_births_ts, xaxt = "n",
    ylab = "Monthly Births (unknown scale)",
    xlab = "Date",
    main = "Monthly New York City Births, January 1946 through December 1959")
axis(side =1 , at= seq(1,156,12), labels = seq(1946, 1958))

#remove 1946
NYC_births_<-read_csv("~/Documents/Baruch/monthly_nyc_births.csv")
NYC_births_<-NYC_births[-c(1:12),]
NYC_births_ts<-ts(NYC_births$Births)

plot(NYC_births_ts, xaxt = "n",
    ylab = "Monthly Births (unknown scale)",
    xlab = "Date",
    main = "Monthly New York City Births, January 1947 through December 1959")
axis(side =1 , at= seq(0,144,12), labels = seq(1947, 1959))

acf(NYC_births$Births, lag.max = 40, main = "ACF, original data")
pacf(NYC_births$Births, lag.max = 40, main = "PACF, original data")


#fairly equal variance, go straight to differencing to remove trend
births_d<-diff(NYC_births_ts)
plot(births_d,
    ylab="",
    main = "Differenced Time Series, Monthly NYC Births")
acf(births_d, lag.max = 40, main = "ACF, ordinary difference")
pacf(births_d, lag.max = 40, main = "PACF, ordinary difference")
adf.test(births_d) #passes unit root test


births_ds<-diff(diff(NYC_births_ts,12))
plot(births_ds,
```

```
    ylab="",
    main = "Differenced Time Series, Monthly NYC Births")
acf(births_ds, lag.max = 48, main = "ACF, ordinary & seasonal difference")
pacf(births_ds, lag.max = 48, main = "PACF, ordinary & seasonal difference")
adf.test(births_ds) #passes unit root test

Fit1 <- Arima(NYC_births_ts, order=c(0,1,3), seasonal=list(order=c(0,1,1), period=12))
Fit1
AIC(Fit1); BIC(Fit1)
hist(residuals(Fit1))
sarima(NYC_births_ts, 0,1,3, P=0, D=1, Q=1, S=12)

Fit2 <- Arima(NYC_births_ts, order=c(0,1,2), seasonal=list(order=c(0,1,1), period=12))
Fit2
AIC(Fit2); BIC(Fit2)
hist(residuals(Fit2))
sarima(NYC_births_ts, 0,1,2, P=0, D=1, Q=1, S=12)

Fit3 <- Arima(NYC_births_ts, order=c(2,1,2), seasonal=list(order=c(0,1,1), period=12))
Fit3
AIC(Fit3); BIC(Fit3)
hist(residuals(Fit3))
sarima(NYC_births_ts, 2,1,2, P=0, D=1, Q=1, S=12)

Fit4 <- Arima(NYC_births_ts, order=c(0,1,2), seasonal=list(order=c(0,1,2), period=12))
Fit4
AIC(Fit4); BIC(Fit4)
hist(residuals(Fit4))
sarima(NYC_births_ts, 0,1,2, P=0, D=1, Q=2, S=12)

Fit5 <- Arima(NYC_births_ts, order=c(1,1,2), seasonal=list(order=c(1,1,2), period=12))
Fit5
AIC(Fit5); BIC(Fit5)
hist(residuals(Fit5))
sarima(NYC_births_ts, 1,1,2, P=1, D=1, Q=2, S=12)

Fit6 <- Arima(NYC_births_ts, order=c(0,1,2), seasonal=list(order=c(1,1,2), period=12))
Fit6
AIC(Fit6); BIC(Fit6)
hist(residuals(Fit6))
sarima(NYC_births_ts, 0,1,2, P=1, D=1, Q=2, S=12)

Fit7 <- Arima(NYC_births_ts, order=c(1,1,2), seasonal=list(order=c(0,1,2), period=12))
Fit7
AIC(Fit7); BIC(Fit7)
hist(residuals(Fit7))
sarima(NYC_births_ts, 1,1,2, P=0, D=1, Q=2, S=12)

Fit8 <- Arima(NYC_births_ts, order=c(2,1,2), seasonal=list(order=c(0,1,2), period=12))
Fit8
AIC(Fit8); BIC(Fit8)
hist(residuals(Fit8))
```

```
sarima(NYC_births_ts, 2,1,2, P=0, D=1, Q=2, S=12)


#forecast with Fit1 to compare with AR12 from project #1
par(cex.axis = 0.6, cex.lab = 0.7, cex.main = 0.8)
fore12 <- predict(Fit1, 12)
ts.plot(NYC_births_ts, fore12$pred, gpars = list(xaxt="n"),
      ylab = "Monthly Births (unknown scale)",
      xlab = "Date",
      main = "Monthly New York City Births, \nJanuary 1947 through December 1960, with Forecast")
axis(side =1 , at= seq(1,168,12), labels = seq(1947, 1960))
U = fore12$pred+fore12$se; L = fore12$pred-fore12$se
xx = c(time(U), rev(time(U)));  yy = c(L, rev(U))
polygon(xx, yy, border = 8, col = gray(.6, alpha = .2))
lines(fore12$pred, type = "p", col = 2)

#values
fore12$pred

#forecast test
NYC_births_less<-NYC_births[-c(145:156),]
NYC_births_less_ts<-ts(NYC_births_less$Births)
births_test<-diff(diff(NYC_births_less_ts),12)
Fit1_test <- Arima(births_test, order=c(0,1,3), seasonal=list(order=c(0,1,1), period=12))
fore12_test<- predict(Fit1_test, 12)
fore12_test$pred+NYC_births_less_ts[144]


#working with intervention analysis

Fit1.i=arimax(NYC_births_ots,order=c(0,1,3),
          seasonal=list(order=c(0,1,1),period=12),
          xtransf=data.frame(blackout=(c(rep(1,12),rep(0,156))),
                    blackout=(c(rep(1,12),rep(0,156)))),transfer=list(c(1,0)),
          method='ML')

AIC(Fit1.i)
hist(residuals(Fit1.i))
checkresiduals(Fit1.i)

plot(NYC_births_ots, xaxt = "n",
     ylab = "Monthly Births (unknown scale)",
     xlab = "Date",
     main = "Monthly New York City Births, January 1946 through December 1959 \nwith Intervention
Analysis Points
      ")
axis(side =1 , at= seq(1,156,12), labels = seq(1946, 1958))
points(fitted(Fit1.i), col=2)

#plot intervention effects
blackout=(c(rep(1,12),rep(0,156)))
plot(ts(filter(blackout, filter=-0.6247, method='recursive', side=1)*
```

```
(1.2034),frequency=12,start=1946),ylab='Blackout Effects',main="Intervention Effects",
type='h'); abline(h=0)


#forecast with intervention term
#reference this online post: https://stats.stackexchange.com/questions/169564/arimax-prediction-using-forecast-package

tf<-filter(1*c(rep(1,12),rep(0,length(NYC_births_ts)+12)), filter = -0.6247, method='recursive',
side=1)*(1.2034)
new.forecast<-Arima(NYC_births_ots, order=c(0,1,3), seasonal=list(order=c(0,1,1), period=12),
xreg=tf[1:(length(tf)-12)])
new.forecast
newfore12<-predict(new.forecast,n.ahead = 12, newxreg=tf[169:length(tf)])
ts.plot(NYC_births_ots, newfore12$pred, gpars = list(xaxt="n"),
      ylab = "Monthly Births (unknown scale)",
      xlab = "Date",
      main = "Monthly New York City Births, \nJanuary 1947 through December 1960, with
Forecast\nincluding Interaction Analysis")
axis(side =1 , at= seq(1,168,12), labels = seq(1947, 1960))
U = newfore12$pred+newfore12$se; L = newfore12$pred-newfore12$se
xx = c(time(U), rev(time(U)));  yy = c(L, rev(U))
polygon(xx, yy, border = 8, col = gray(.6, alpha = .2))
lines(newfore12$pred, type = "p", col = 2)


#values
newfore12$pred

#forecast test
NYC_births_less2<-NYC_births_o[-c(157:168),]
NYC_births_less2_ts<-ts(NYC_births_less2$Births)
births_test2<-diff(diff(NYC_births_less2_ts),12)
tf2<-filter(1*c(rep(1,12),rep(0,length(NYC_births_less2_ts)+12)), filter = -0.6247, method='recursive',
side=1)*(1.2034)

new.forecast2<-Arima(NYC_births_less2_ts, order=c(0,1,3), seasonal=list(order=c(0,1,1), period=12),
xreg=tf2[1:(length(tf2)-12)])
new.forecast2
newfore122<-predict(new.forecast2,n.ahead = 12, newxreg=tf2[157:length(tf2)])
newfore122$pred
```