Infant Mortality Rate Time Series Analysis, Group Code: 17415

Viral Sanjiv Desai, Fabian Colin, Catherine Chang, Jordan Jiada Low, Xianda Gao

3/14/2019

Contents

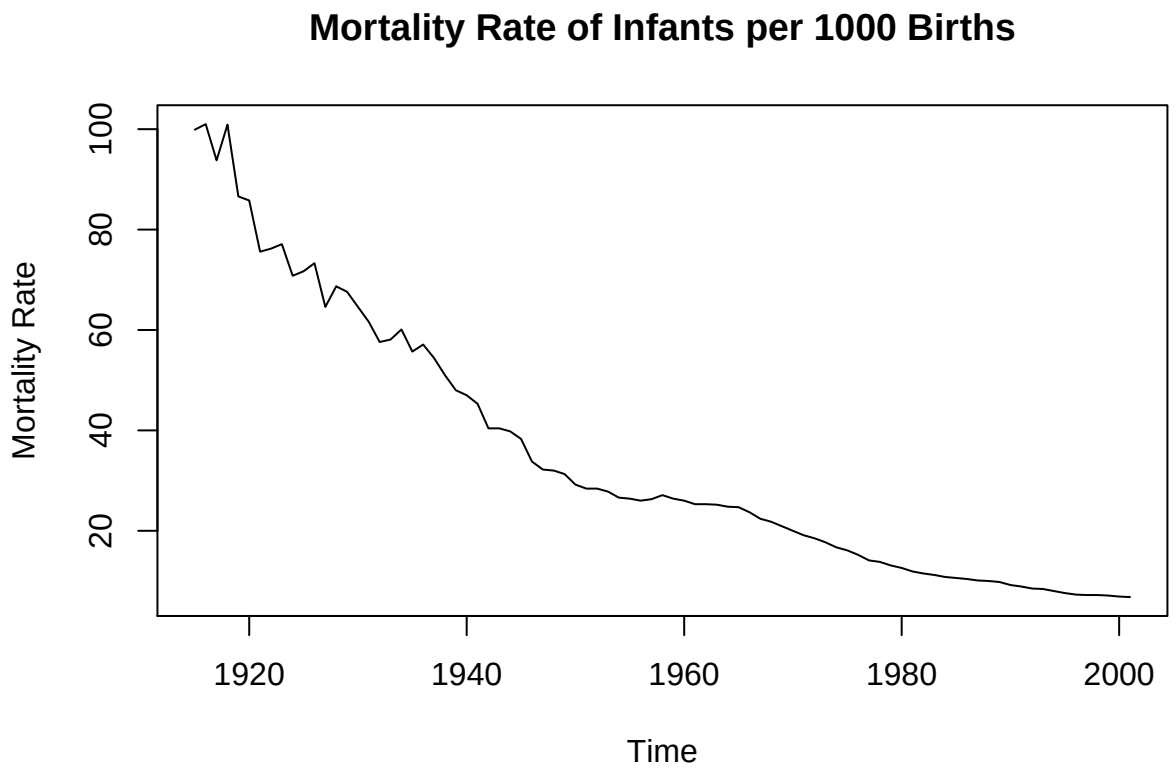
1. Introduction

One of the most important aspects of life is survival rate and/or the mortality rates of newborns. There have been increasing concerns over the past few years about the overpopulation of humans on Earth. Many people fear that at the rate huamns are growing, there simply won't be enough resources to sustain all of us. At the same time, people still choose to have children for a variety of personal reasons so its important if a family decides to have a child that they can have one and a healthy one at that. One important measure of childbirth is survival rates and the mortality rates of new born infants. Its important to keep these rates so families can happily receive a new child. Our data comes from the Center of Disease Control and Prevention in the United States and the mortality rate of infants(under 1 year) and neonatal(under 28 days) from 1915-2013 per 1000 births. This means that if our mortality rate is 100, then that implies 100 infants died out of 1000 births that year. Our total dataset orginally 198, but for our purposes since we will only be looking at infant mortality rate we have a subset of n = 99 observations to analyze.

Our goal with this dataset will be to find a suitable ARIMA/SARIMA model that best suits our data. Our ultimate goal will be to predict future mortality rates. In order to test how accurate our predictions are to the actual values recorded, we will create a training set and testing set to build our model and compare to respectively. We will subset roughly the last 12 years worth of observations into a test set so that we can forecast 12 years starting from 2002-2013 and compare our ARIMA/SARIMA model vs the actual observations. So to get started we will first do some Exploratory Data Analysis to get a better understanding of our data and see if we can see some obvious areas of interest we might want to dive deeper into.

2. Exploratory Data Analysis

2.1 Initial Plot Diagnostics



From our initial diagnostic of our plot there seems to be a very obvious downwards trend, something most people would expect. This decreasing trend could be due to technology advancements in medicine, increasing

the chances of infant survival during birth. Another thing that we notice from this data is that there is no clear seasonality. In general, the mortality rate seems to decrease but at certain points we see upward movement. We would expect something like this to be the case because the mortality rate is already per 1000 births so even if there are higher births during certain seasons of the year our data is standardized in a sense.

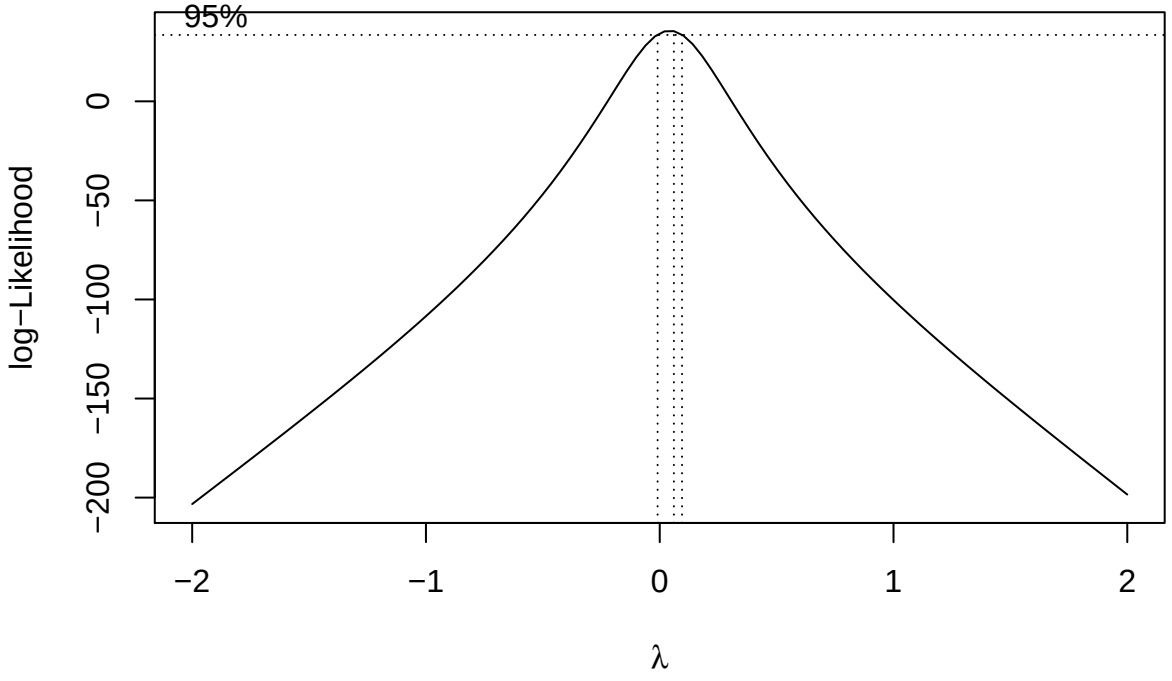
Another thing we notice is that the decline between 1960-present is much smoother than the decline in the early 1920s. From this we can conclude that there is a difference in variances depending on different time frames. This can be possibly explained by the fact that technology does not exactly get better on a linear trend, so we would expect some sort of change of variance. In addition, illnesses and diseases can cause some set backs in the improvement of infanty mortality rates.

Since our model has no appearent seasonal component, we can write our prospective model as: $Y_t = m_t + S_t$. Our model will look something like that where m_t is the trend and S_{t} is our stationary series. In addition, because our model contains trend, we can conclude our data is non-stationary which leads us to the first step. We can move into transformations to help us remove trend which is important to obtain stationary data.

3. Data Transformation

3.1 Box-Cox Transformations

From looking at our initial data's time series plot, we can see that there is a very clear decreasing trend in our data and that variance may not be constant. In order to remove the trend and potentially unstable variance, we should first transform our data utilizing the boxcox function to find the optimal transformation for our data.



According to our plot above, optimal lambda value produced lying within a 95% confidence interval is around 0. Based on the result, we will apply a log() transformation. In order to check if our transformation removed the trend, we will check the variance before and after.

After taking the log of our data we see it turns out our variance decreased from 689.8548 to 0.6397329 implying this was a successful transformation. Now we will look to difference the data either at lag 12 or lag 1. Since there is no clear seasonality we will first difference our data to see if the variance decreases. So right

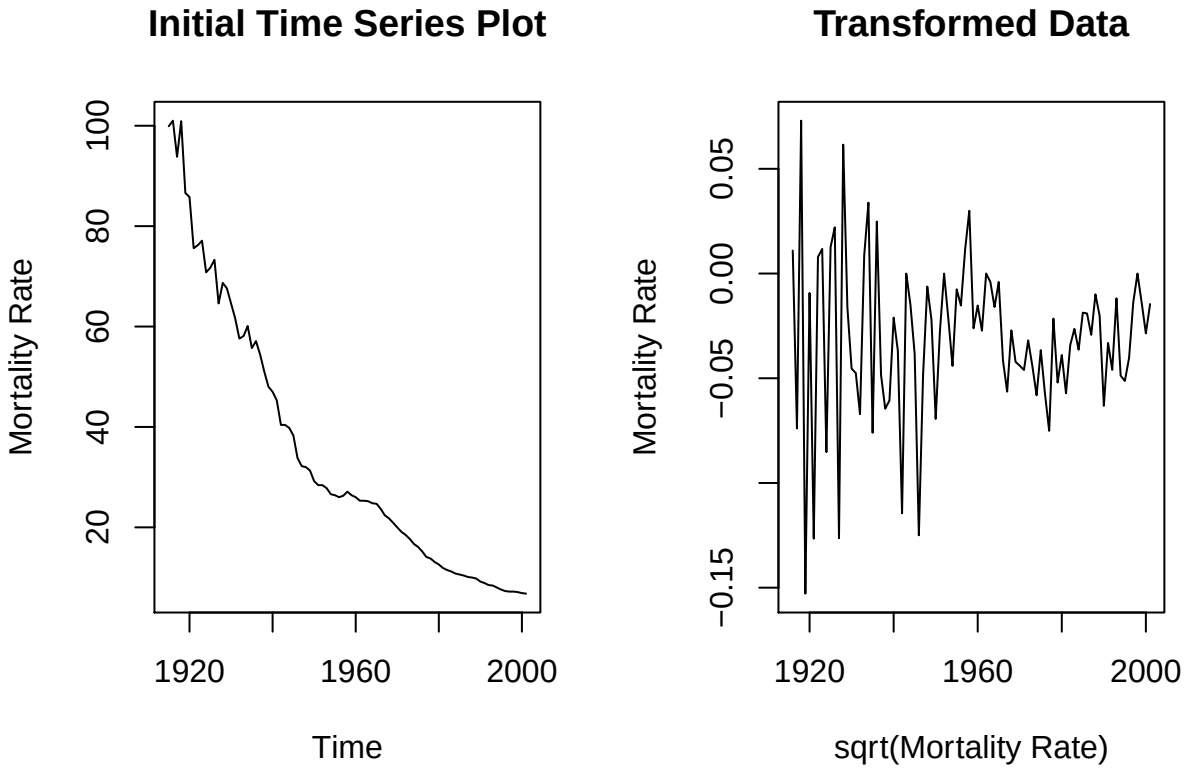
now our model looks like:

$$Y_t = \ln(X_t)$$

3.2 Differencing

Now we will difference at lag 1, and notice the variance further decreases (0.6397329 to 0.001461909) and our graph looks closer to stationary. Our new plot no longer has a trend to it but there still remains an uneven variance. Applying a second difference at lag 1 once more to see if our variance decreases further.

After differencing at lag 1 once again, our variance increased from 0.001461909 to 0.003777939. This result implies that we over differenced. As an extra measure, we will also try differencing at lag 4 and 12 just in case our data has some seasonality that is not visible in our plots. After differencing at lag 4 and 12 respectively for both cases, our variance increased (0.001461909 to (difference at lag 4 variance) 0.002539089 or (difference at lag 12 variance) 0.002920633). We conclude differencing once at lag one as this provides us with the most decrease in variance, in addition to stationary results. We want to void over differencing as this results in higher variance which affects the model building process.



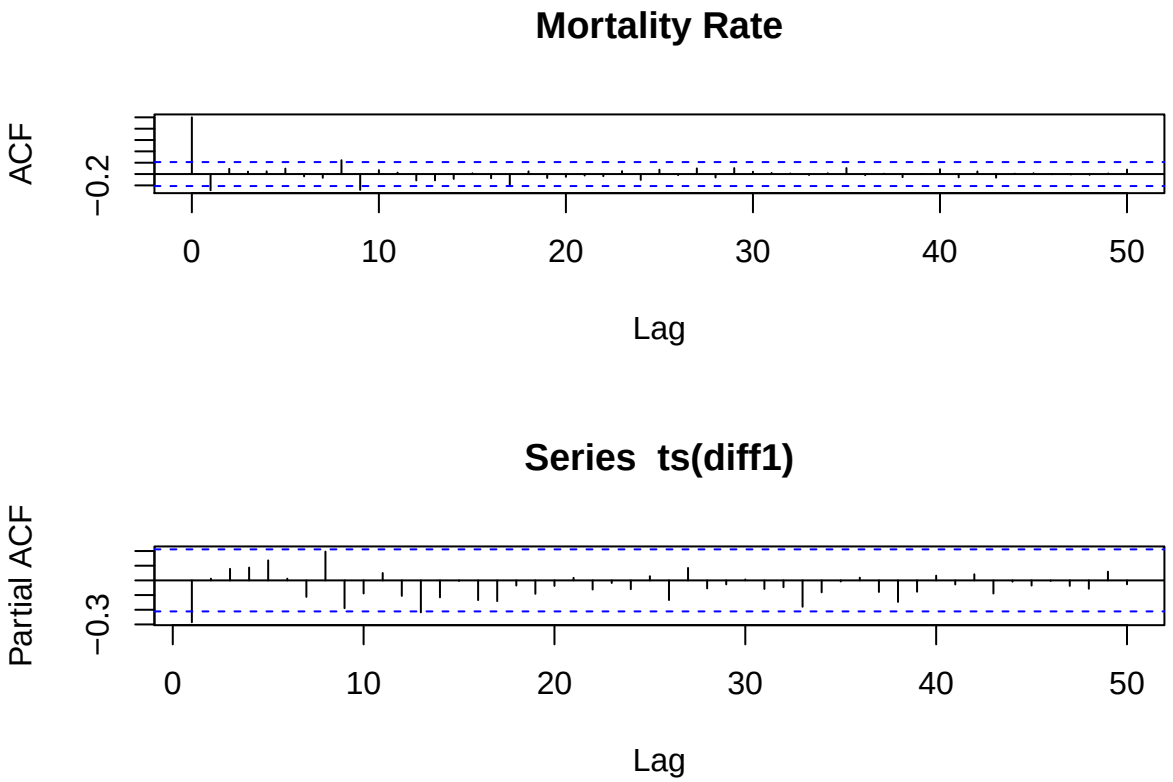
This means that our final model will look something like:

$$\bigtriangledown X_t$$

Where:

$$X_t = \ln(Y_t)$$

3.3 Model Identification



From looking at our ACF/PACF on a grand scale, we can see that our ACF/PACF both tail off after a certain point meaning this is more than likely an $ARIMA$ model. This confirms our diagnostics from differencing since we did not difference seasonally due to our data having no seasonality. But from looking at the plot, our graph seems to tail off at lag 1 for both PACF/ACF plot. There are some spikes in the ACF plot that are outside the bounds after lag 1 but for the most part they are tailing off within the 95% bound and not as significant as the first spike. Since we lack seasonality and we only differenced once at lag 1, our predicted model looks like:

$$ARIMA(1,1,1)$$

This means that our model p,q can either be 0 or 1 for both. We can use AICc/BIC to obtain potential models and do diagnostic checks on multiple models to find the optimal model.

4. Model Identificaiton and Estimation

We have classified our data to not having a seasonal component which means the appropriate models to consider follow the $ARIMA(p,d,q)$ model.

The p represents the order of non-seasonal AR process, d represents non-seasonal differencing, and q represents the order of non-seasonal MA process. By looking at our ACF/PACF plots, we are given that the p and q variable can take the values of either 0 or 1.

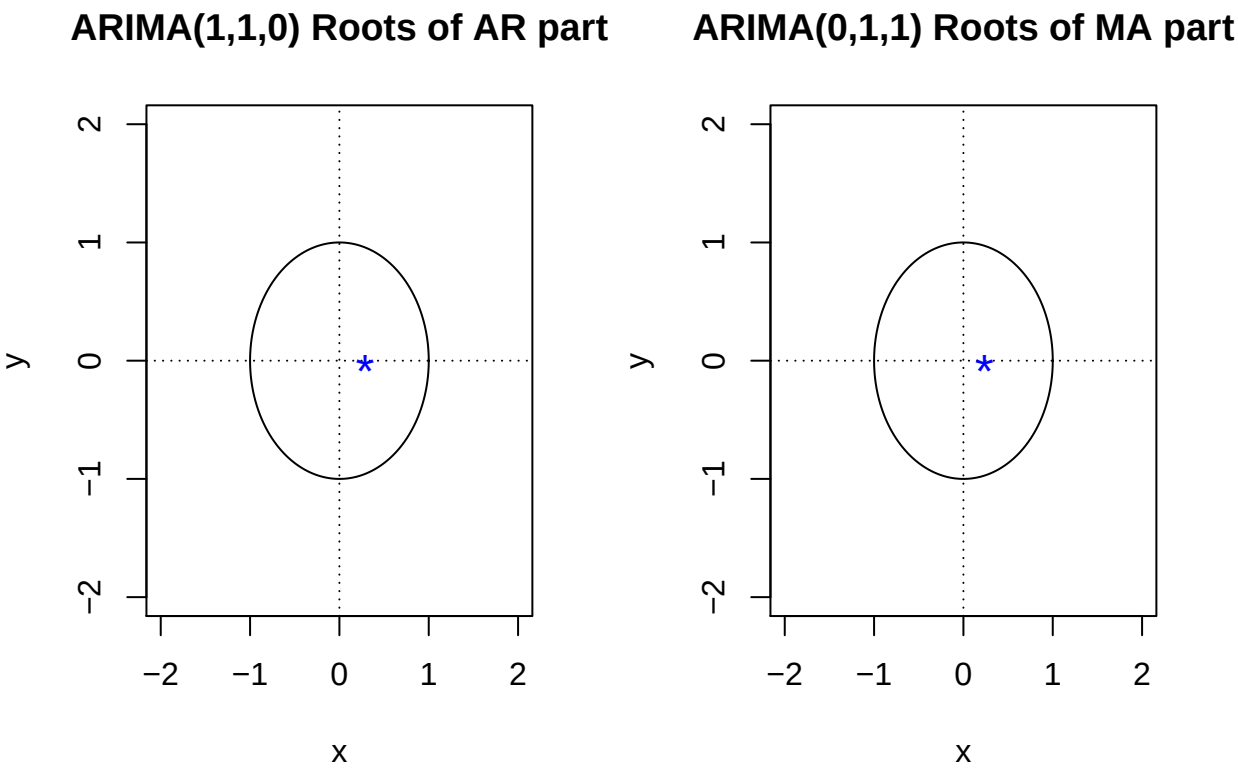
4.1 Model Selection

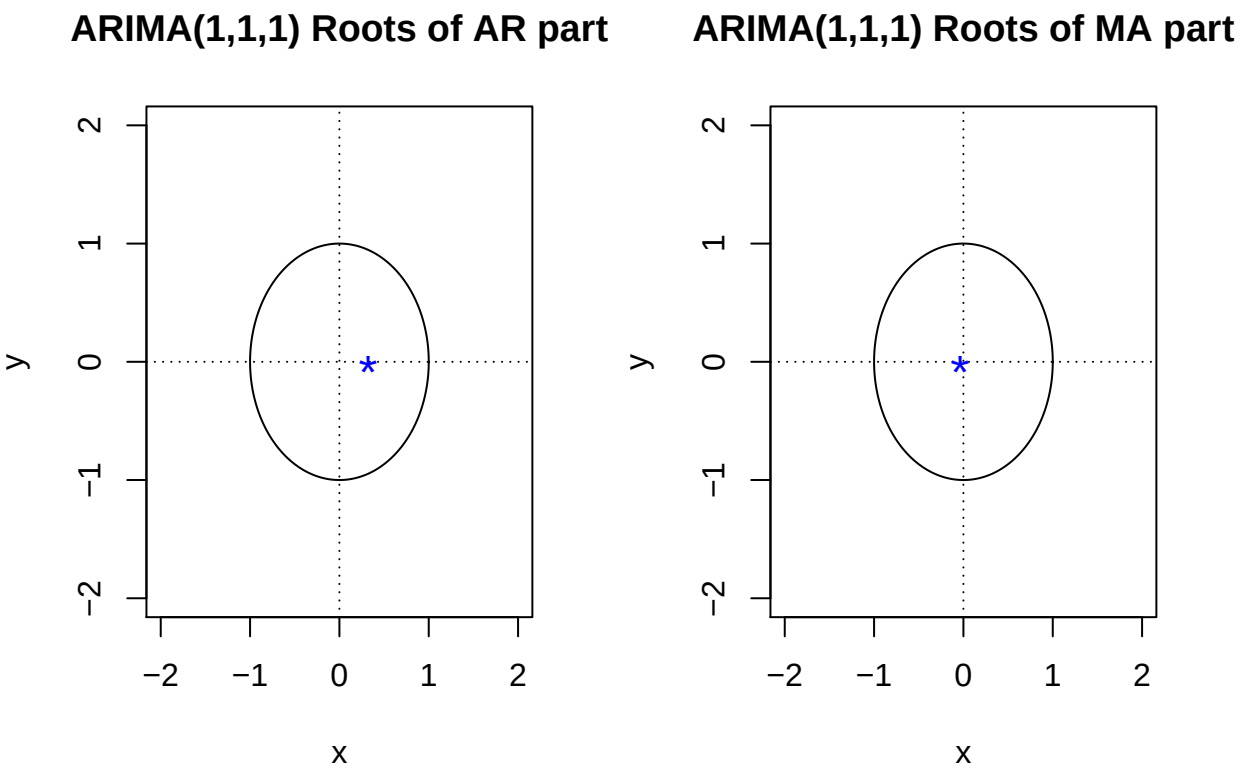
After identifying the preliminary model by looking at the ACF/PACF plots, we will use information such as AICc and BIC to gather the best model or models for this specific dataset. By choosing the model with the lowest AICc and BIC, we will have our best model.

Model	AICc	BIC
ARIMA(0,1,0)	-271.9597	-269.5054
ARIMA(0,1,1)	-272.7291	-267.8204
ARIMA(1,1,0)	-274.8896	-269.9809
ARIMA(1,1,1)	-307.2559	-299.8928

We will go ahead and test four different models. We choose these specific models to test because $ARIMA(0,1,0)$ has the lowest AIC, Akaaike information criterion, which is used to estimate the likelihood of a specific model in order to predict future values. We also picked $ARIMA(0,1,1)$ to test because it obtains the lowest BIC, Bayesian information criterion, which measures the trade-off between a model's fit and the model's complexity. The model with minimum AIC and BIC shows a better fit. We also tested $ARIMA(1,1,0)$ because it was the model that the function auto.arima gave us. This function returns the best fit ARIMA model according to either AIC, AICc or BIC value. Lastly, we picked $ARIMA(1,1,1)$ as a model to test because of our diagnostics from the ACF and PACF graphs. But in the end, we can automatically forfeit the ARIMA(0,1,0) model from our diagnostics since we cannot estimate coefficients for that model. Thus we only have 3 remaining models.

4.2 Unit Root Plotting





From these plots we can see that all 3 of our models have roots outside the unit circle. Since the blue dots represent the inverse of our roots (or 1/root), and our blue points are < 0.5, we conclude our red points will be > 2 which is outside the bounds for the plot. Because the roots of our AR components land outside of the unit circle, this implies stationarity. Similarly, for the MA process, roots lying outside the unit circle implies invertibility. From this we can move on to diagnostic checks to find the best model.

5. Diagnostics

Now that we have identified 3 models and their parameters, we will conduct diagnostic checks in order to identify the optimal model for forecasting. We will validate the following assumptions for the remaining 3: error normality, error independence, and constant error variance.

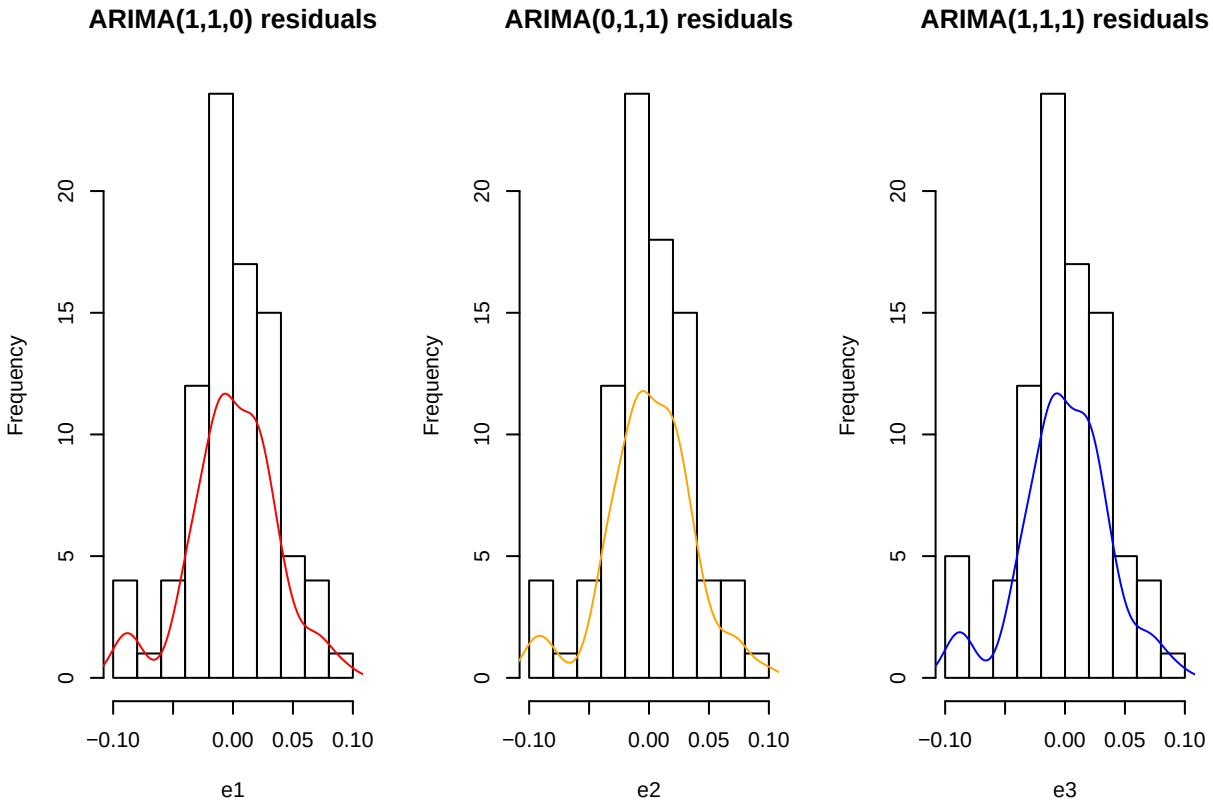
Our 3 models include the following:

$$ARIMA(1,1,0)$$
$$ARIMA(0,1,1)$$
$$ARIMA(1,1,1)$$

5.1 Normality in residuals

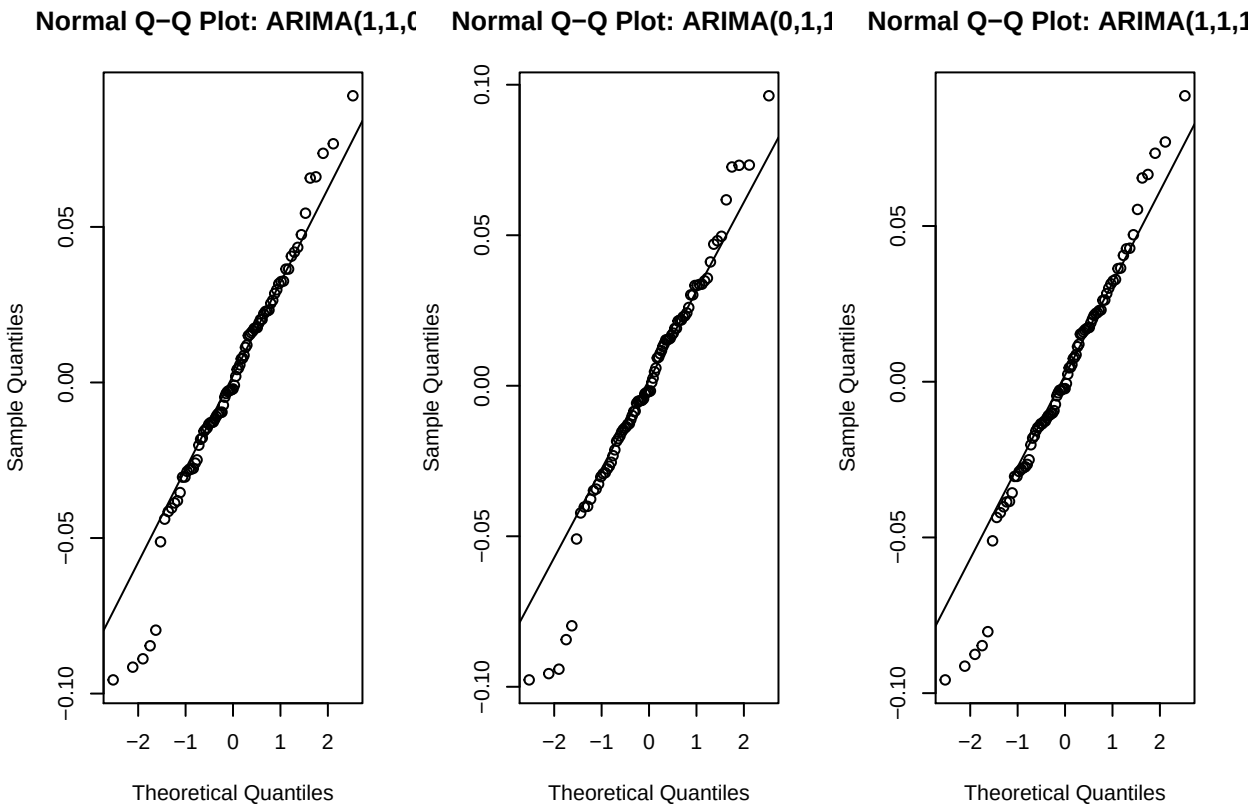
We will use histograms and Q-Q plots of residuals in order to check whether if our normal assumptions hold.

a. Histograms of residuals



We can see from the histograms provided above that our residuals appear to be normally distributed. We can see that Model 2 appears to have a more complete normally distributed histogram without gaps. We can move onto Q-Q plots which will further help us check this assumption.

b. Q-Q Plots of residuals



From our Q-Q Plots, we can see again that normality does not appear to be an issue. While tails seem to be a bit of an issue for all 3 graphs, there is no appearant issues for any individual graph. In conclusion, residuals appear to be normal.

c. Shapiro-Wilk Test

Performing the Shapiro-Wilk test can provide us a more empirical testimonial about the assumtion of normality. Under this test we have the following testing hypothesis ($\alpha = 0.05$):

H_0 : residuals are normal vs H_a : residuals are not normal

	W Statistic	P-Value
ARIMA(0,1,1)	0.9758836	0.1035541
ARIMA(1,1,0)	0.9707369	0.0457646
ARIMA(1,1,1)	0.9760776	0.1067898

We can see from the Shapiro-Wilk test results that model 1 and model 3 both have p-values greater than $\alpha = 0.05$ From our hypothesis, we conclude that model 1 and model 3 have residuals that follow normal. Model 2 rejects H_0, which implies its residuals are not normal.

5.2 Serial Correlation Check

In this section, we will check for correlation among the residuals. It is essential that we must ensure the residuals are uncorrelated with lagged values to ensure independence and randomness. The two tests we will apply are the Ljung-Box Test and the Box-Pierce Test.

Under this test we have the following testing hypothesis ($\alpha = 0.05$):

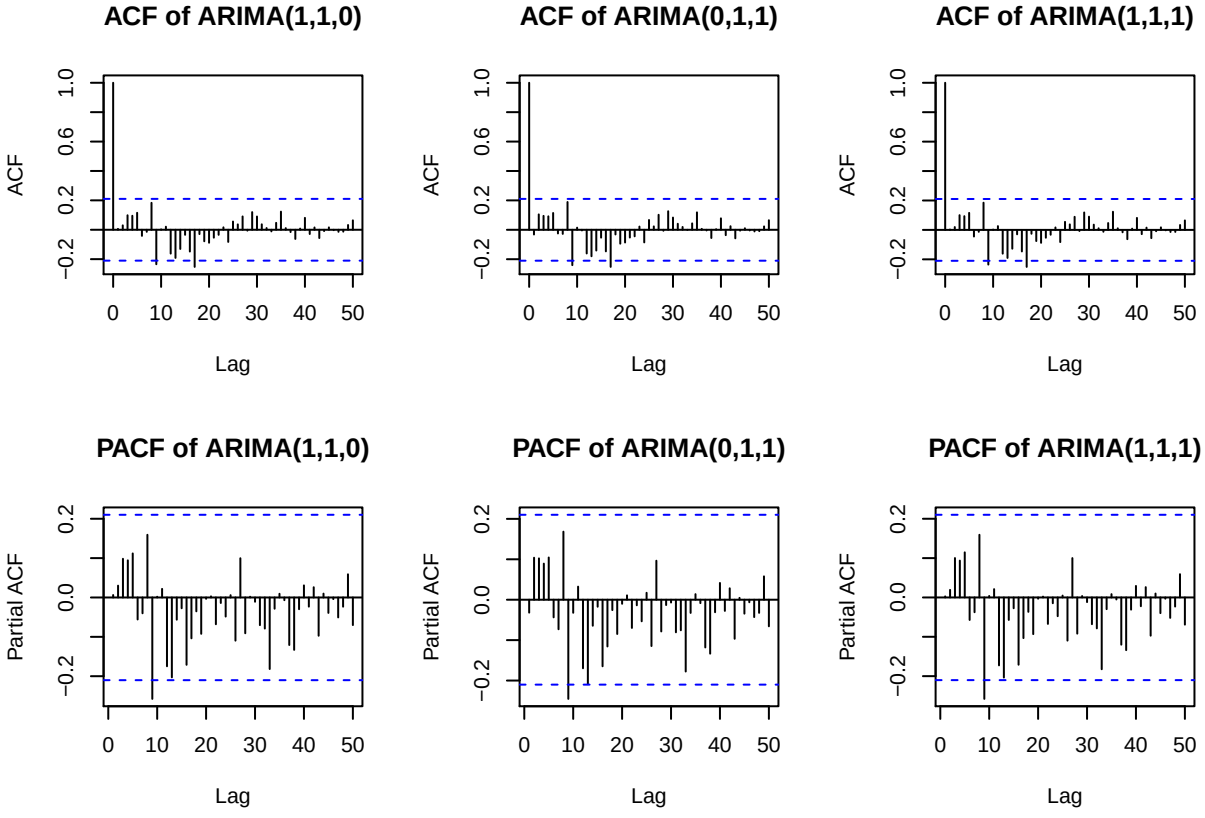
H_0 : residuals are serially uncorrelated vs H_a : residuals are not serially uncorrelated

	ARIMA(1,1,0)	ARIMA(0,1,1)	ARIMA(1,1,1)
Box-Pierce	0.9536220	0.7651247	0.9796628
Ljung-Box	0.9528209	0.7611856	0.9793112

From the calculated p-values, we can conclude that all p-values are above $\alpha = 0.05$. This implies that all three models fail to reject the null hypothesis: Residuals are serially uncorrelated.

5.3 Constant Variance Check

In order for us to continue with model estimation and forecasting, we must ensure that our variance within the errors is constant. By analyzing ACF and PACF plots of our models, we are able to check for heteroskedasticity. Heteroskedasticity is simply a fancy way of saying our constant variance assumption is violated. Below are our plots.



From our ACF and PACF plots, we can see that our values stay within the 95% white noise limits. Thus we can conclude that heteroskedasticity does not exist.

5.3 Final Model

From the previous diagnostic checks, we can decide which model to continue with into forecasting. Because our model 1 did have the lowest AIC, in addition to holding our assumptions of normality and constant variance, we will continues with it. The final model will follow:

$$ARIMA(1,1,0)$$

Final Model:

$$(1-\phi_1 B)X_t = (1-\theta_1 B)Z_t$$

Where:

$$\phi_1 = -0.3214$$
$$\theta_1 = 0.0375$$

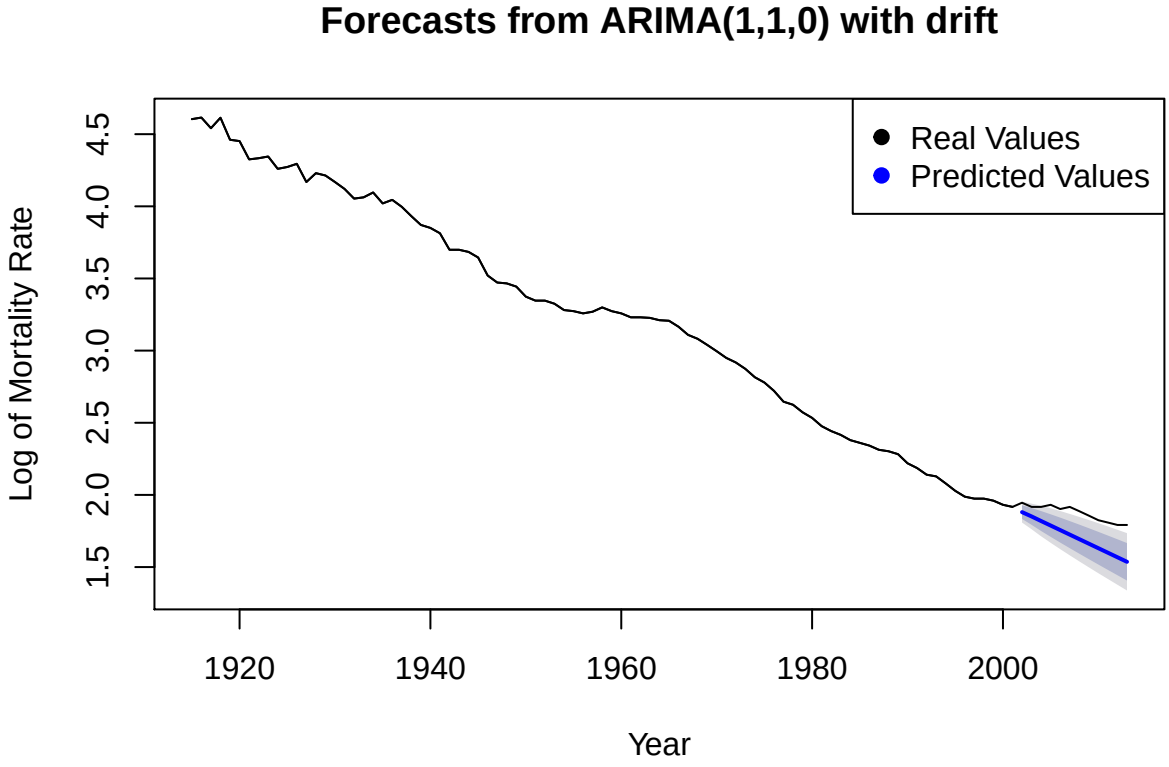
Hench, we have:

$$Y_t = -.3214Y_{t-1} + Z_t + 0.0375Z_{t-1}$$

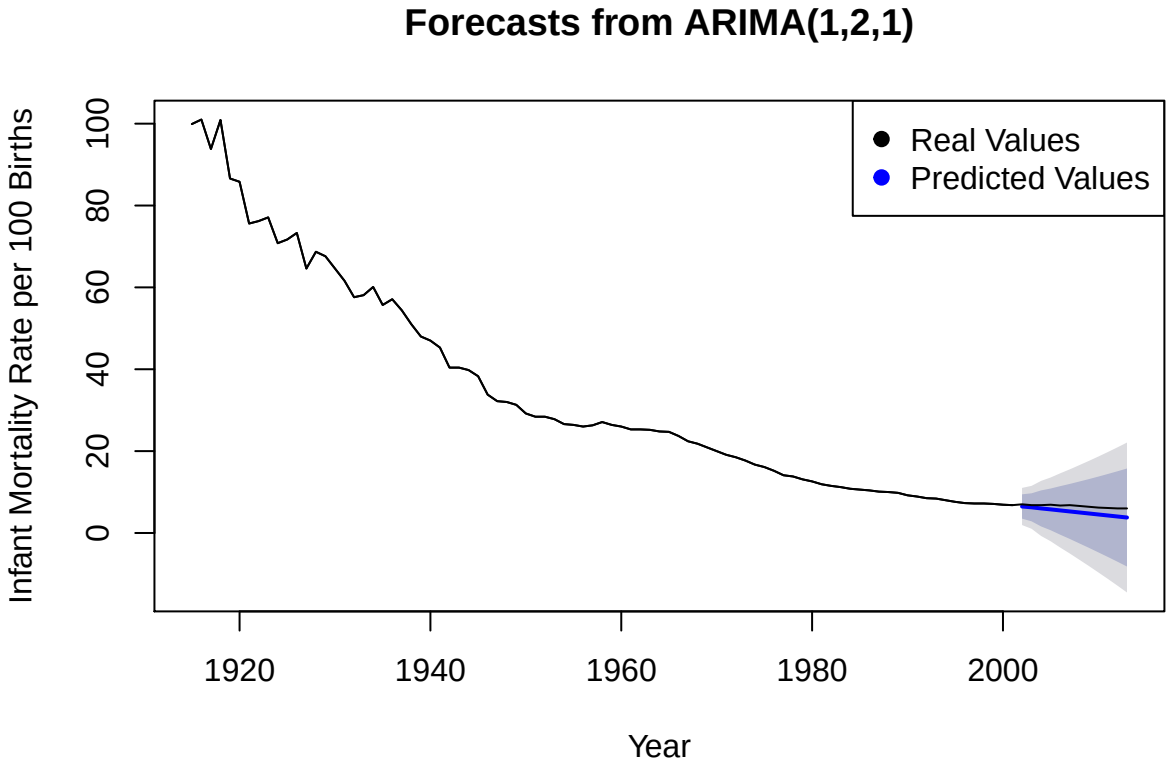
6. Forecasting

6.1 Forecast

So now that we have our final model we will try to predict the future values. In the beginning, we built our model under a training set with 87 observations with 12 years of data left to forecast on. We did this so we can compare our models accuracy to understand how well our model performed. First, we forcast on the transformed data and check our results.



Firstly, we decided to apply the forecast function to forecast the future 12 years values based on the logarithmic model. We can see from the graph below that the forecast shows a general decreasing trend with a gradient similar to the one from 1960s to 2000. However, when we compare the predicted values with the actual values during the same time period, the actual values do not coincide with the actual values exactly. However, the difference in forecasted values could be due to the simplicity of the model where we do not account for other variables and factors that could affect the mortality rate in real life. This can be seen in the plot below where there is a significant change in the rate of decrease as seen from the change in gradient from 2000 onwards. Therefore, the model may not be the best model and we decided to transform the data back and to conduct another forecast.



For this forecast, we decided to exclude the last 12 years from the model so that we are able to compare our forecast with the actual values (the blue line is the projected line while the black one is the actual line). From the figure above, we can see how the predicted amounts and the 95% confidence region. Although the forecast does not coincide exactly with the actual values, but we can see that the actual values do lie within the confidence interval. Therefore, we can conclude that the forecast is relatively accurate and that the chosen model is a suitable model. It is also interesting to note that the confidence band increases as time goes on which suggests that there is an increasing uncertainty for longer term forecast. Therefore, we may require a more stable model or more data.

7. Conclusion

Our objective was to produce a time series model that would allow us to predict and forecast future infant mortality rates. With the original data provided, we found that there was a clear decreasing trend that could be explained by the medical advancements and practices transitioning from the 20th to 21st century. With the use of a Box-Cox transformation, we were able to find the optimal lambda to remove this trend which resulted in a stationary data set. Although seasonality was not notably present, it was important to apply differencing to ensure there was no hidden effects towards non-constant variance. After model identification with the utilization of AIC, and diagnostic check to ensure the residuals assumptions of normality and constant variance, we were able to arrive at our final model:

Table 4: True Values vs Predicted Values

x	Mortality.Rate
6.472197	7.0
6.268363	6.8
5.997066	6.8
5.762482	6.9
5.507919	6.7
5.264228	6.8
5.014621	6.6
4.768234	6.4
4.520094	6.2
4.272908	6.1
4.025203	6.0
3.777780	6.0

This final model was then able to forecast infant mortality rates for 12 years ahead starting in 2002 to 2013 based on the 87 previous years recordied. Taking a look at table 4, we can compare the forecasted values (x), in comparison to the actual mortality rate. We were able to successfully predict future mortality rates for infants, and hope our study sheds light into improving birth care and practices.

We want to thank Professor Bapat for the knowledge and aid with the project he provided us during the course of PSTAT 174.

8. Appendix

```
library(readr)
library(MASS)
library(forecast)
#importing data
neonatal_2 <- read_csv("NCHS_-_Infant_and_neonatal_mortality_rates__United_States__1915-2013.csv")
#only taking infant mortality rate since data contains neonatal in half
infant <- neonatal_2[c(1:99),]
infant_train <- infant[c(1:87),] #subsetting into training set
infant_test <- infant[c(88:99)] #subsetting into testing set
#our timeseries object
ts_infant <- ts(infant_train[,3], start = c(1915), frequency = 1)
#initial plot
plot(ts_infant, main = "Mortality Rate of Infants per 1000 Births")
#variance check
var(ts_infant)
#checking what transformation to use to decrease variance
bcTransform <- boxcox(ts_infant ~ as.numeric(1:length(ts_infant)))
bcTransform$x[which(bcTransform$y == max(bcTransform$y))]
##So according to this 1/2 looks to be in the confidence interval so we will do log transformation
ts_log <- (log(ts_infant)) #our new transformed data
var(ts_log)
##variance decreased to 0.6397329
##So our data does not seem to have seasonality but does have a downwards trend
#so we will first difference at lag 1
diff1 <- diff(ts_log, lag = 1)
var(diff1)
##Variance decreased to 0.002832311
plot(diff1)
#ACF/PACF diagnostics
par(mfrow = c(2,1))
acf(ts(diff1), lag.max = 50)
pacf(ts(diff1), lag.max = 50)
par(mfrow = c(1,1))
##So our ACF/PACF definitely cuts off at lag 1 and tails off.
diff2 <- diff(diff1, lag = 1)
var(diff2)
##Differencing at lag 1 again increased our variance so we over differenced.
##In addition, differencing at lag 12 and lag 4 also over differenced.

######################################### MODEL ESTIMATION

auto.arima(ts_log) #auto.arima to find a potential model
#for loop to retrieve AICc of potential models
for (p in 0:1) {
  for (q in 0:1) {
    print(p)
    print(q)
    print(AICc(arima(ts_log, order = c(p,1,q), method = "ML")))
  }
}
#for loop to retrieve BIC of potential models
```

```r
for (p in 0:1) {
  for (q in 0:1) {
    print(p)
    print(q)
    print(BIC(arima(ts_log, order = c(p,1,q), method = "ML")))
  }
}

################################## POLY ROOT GRAPHS
model1
plot.roots(NULL, polyroot(c(1, -0.2867)), main = "ARIMA(1,1,0) Roots of AR part")
model2
plot.roots(NULL, polyroot(c(1, -0.2361)), main = "ARIMA(0,1,1) Roots of MA part")
model3
plot.roots(NULL, polyroot(c(1, -0.3214)), main = "ARIMA(1,1,1) Roots of AR part")
plot.roots(NULL, polyroot(c(1, 0.0375)), main = "ARIMA(1,1,1) Roots of MA part")

################################## MODEL DIAGNOSTICS
#ARIMA(1,1,0) model
model1 <- arima(ts_log, order = c(1, 1, 0), method = "ML", xreg=1 : length(ts_log))
#ARIMA(0,1,1) model
model2 <- arima(ts_log, order = c(0, 1, 1), method = "ML", xreg=1 : length(ts_log))
#ARIMA(1,1,1) model
model3 <- arima(ts_log, order = c(1, 1, 1), method = "ML", xreg=1 : length(ts_log))
#ARIMA(1,1,0) model residuals
e1 <- residuals(model1)
#ARIMA(0,1,1) model residuals
e2 <- residuals(model2)
#ARIMA(1,1,1) model residuals
e3 <- residuals(model3)
#plotting histograms of 3 models to check for normality
op <- par(mfrow = c(1, 3))
hist(e1, main = "ARIMA(1,1,0) residuals")
lines(density(e1), col = "red")
hist(e2, main = "ARIMA(0,1,1) residuals")
lines(density(e2), col = "orange")
hist(e3, main = "ARIMA(1,1,1) residuals")
lines(density(e3), col = "blue")
par(op)
#plotting QQ-plot to furthur confirm normality
op <- par(mfrow = c(1, 3))
qqnorm(e1, main = "Normal Q-Q Plot: ARIMA(1,1,0)")
qqline(e1)
qqnorm(e2, main = "Normal Q-Q Plot: ARIMA(0,1,1)")
qqline(e2)
qqnorm(e3, main = "Normal Q-Q Plot: ARIMA(1,1,1)")
qqline(e3)
#doing the shapiro wilk test to confirm that our residuals are indeed normal
shaptable <- matrix(c( shapiro.test(e1)$statistic, shapiro.test(e1)$p.value,
                       shapiro.test(e2)$statistic, shapiro.test(e2)$p.value,
                       shapiro.test(e3)$statistic, shapiro.test(e3)$p.value)
                    ,nrow = 3, byrow = T)
#turning into a nice matrix to output
```

```r
rownames(shaptable) <- c("ARIMA(1,1,0)", "ARIMA(0,1,1)", "ARIMA(1,1,1)") #rownames of shapiro results
colnames(shaptable) <- c("W Statistic", "P-Value") #colnames of shapiro results
#outputting matrix
shaptable
#using Box-Pierce test and Ljung-Box test for each model
p1 <- Box.test(e1, type = "Box-Pierce")$p.value
l1 <- Box.test(e1, type = "Ljung-Box")$p.value
p2 <- Box.test(e2, type = "Box-Pierce")$p.value
l2 <- Box.test(e2, type = "Ljung-Box")$p.value
p3 <- Box.test(e3, type = "Box-Pierce")$p.value
l3 <- Box.test(e3, type = "Ljung-Box")$p.value
#putting our outputs in a nice matrix for results
boxes <- matrix(c(p1, l1, p2, l2, p3, l3), nrow = 2, byrow = FALSE)
rownames(boxes) <- c("Box-Pierce", "Ljung-Box") #rownames of test results
colnames(boxes) <- c("ARIMA(1,1,0)", "ARIMA(0,1,1) P-value", "ARIMA(1,1,1) P-value") #colnames of test
#printing out our results
boxes
##Finally we plot our ACF/PACF for each models.
par(mfrow = c(2, 3))
# Model ACFs
acf(e1, main = "ACF of ARIMA(1,1,0)", lag.max = 50)
acf(e2, main = "ACF of ARIMA(0,1,1)", lag.max = 50)
acf(e3, main = "ACF of ARIMA(1,1,1)", lag.max = 50)
# Model PACFs
pacf(e1, main = "PACF of ARIMA(1,1,0)", lag.max = 50)
pacf(e2, main = "PACF of ARIMA(0,1,1)", lag.max = 50)
pacf(e3, main = "PACF of ARIMA(1,1,1)", lag.max = 50)
par(mfrow=c(1,1))

################################## Prediction

#getting log of actual values
ts_full <- ts(infant[,3], start = c(1915), frequency = 1)
#log transformation of all observations
ts_full_log <- log(ts_full)
#predicted values
plot(forecast(auto.arima(ts_log), h = 12))
par(new = TRUE)
lines(ts_full_log) #Actual values
#adding legend
legend("topright", legend = c("Real Values", "Predicted Values"), pch = rep(19,3), col = c(1,4))
#tranforming data back
plot(forecast(auto.arima(ts_infant), h = 12), , xlab = "Year", ylab = "Infant Mortality Rate per 100 Bi
par(new = TRUE)
lines(ts_full) #Actual values
#adding legend
legend("topright", legend = c("Real Values", "Predicted Values"), pch = rep(19,3), col = c(1,4))

########COMPARING VALUES
prediction_values <- forecast(auto.arima(ts_infant), h = 12)
#prediction
pred_values <- as.data.frame(prediction_values$mean)
#true values
```

```r
real.values <- infant_test[,3]
average_distance <- real.values - pred_values
#converting list to numeric to take mean
average_distance <- as.numeric(as.character(unlist(average_distance[[1]])))
#mean of distance between true and prediction values
##1.387409
```