**Laura Hendley and Rose Wilson Time Series Analysis 11/20/2020 Final Project 2020**

**Question 2:** Load the library(fpp2) package. Please use the dataset(mens400)

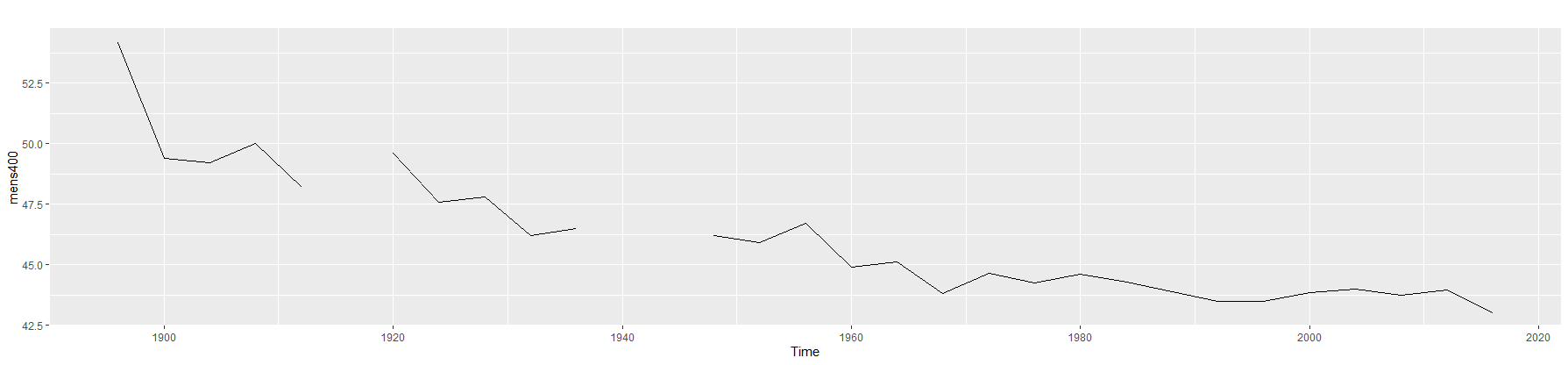
**2a: Find a way to export this data back into Excel and explain your code. Copy and paste will not be accepted.**

write.csv(mens400,"mens400.csv")

Write.csv function takes in an object to be written, in this case the mens400 data, and then writes the data to a file named within quotation marks, in this case “mens400.csv”.

**2b: Graph a picture of this data set.**

autoplot(mens400)

****

**2c: Explain the data set and what patterns are in the set.**

This data set is the winning times of the men’s 400m track final in the Olympics from 1896-2016. There are gaps in the data set corresponding to the years of the World Wars. There is a decreasing trend in the data. For many years, the times were decreasing rapidly but the decreasing trend has slowed and become more horizontal.

**2d: What is the frequency and does it make sense with this data.**

The frequency is .25 which makes sense because the olympics only happen every four years.

**Please use the year 2016 as a test value so do not include it in any of the models below.**

mensTest <-window(mens400, end=c(2012,1))

**2e: In this part, use all the data available to build your model except 2016. Please give three forecasting options and explain which one you would choose as the best option.**

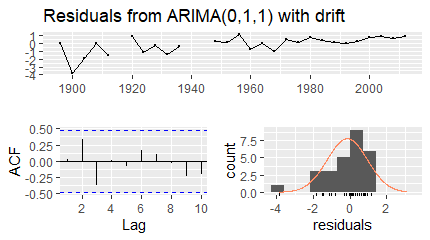
mod1 <- auto.arima(mensTest)

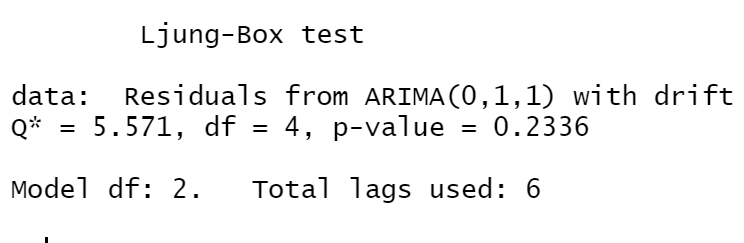
checkresiduals(mod1)

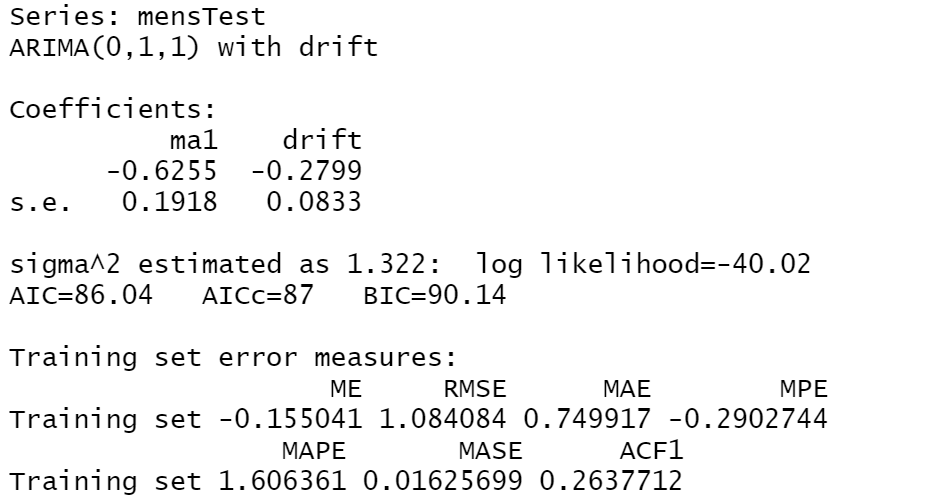
summary(mod1)

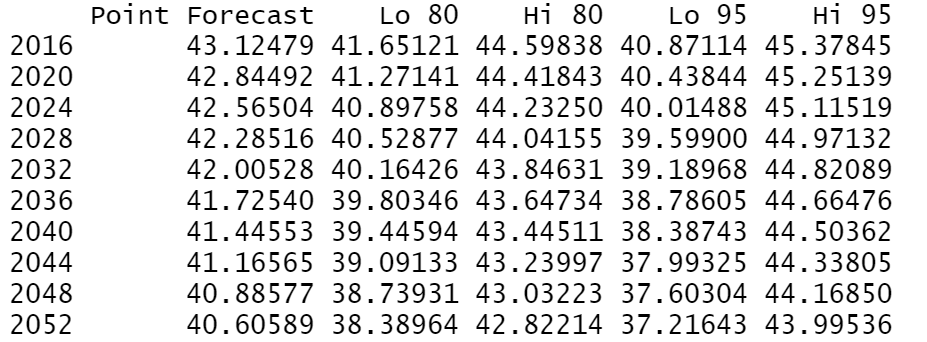
forecast(mod1)

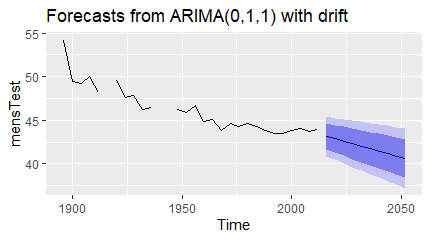
autoplot(forecast(mod1))









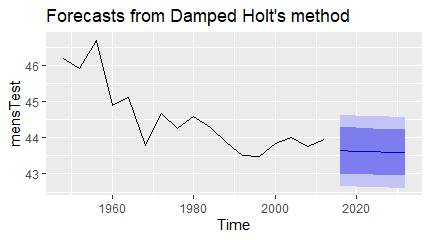


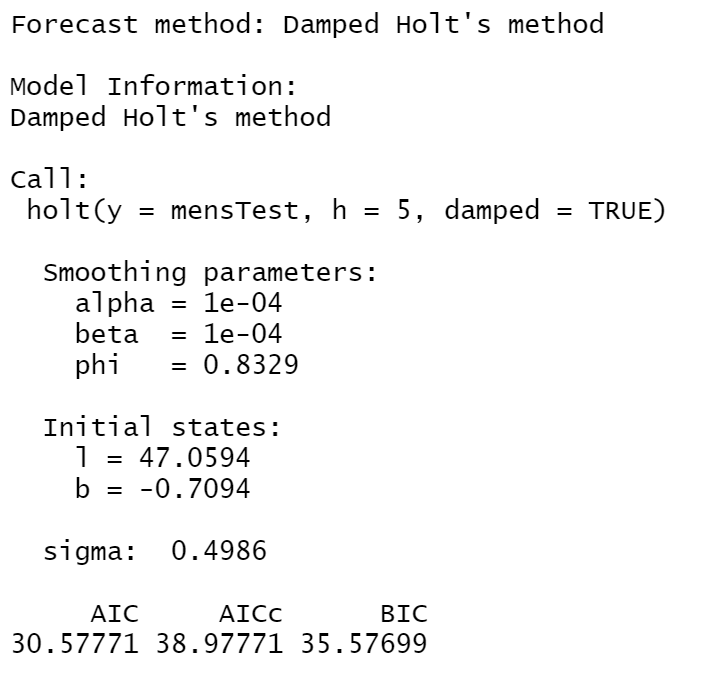
mod2 <- holt(mensTest,h=5,damped=TRUE)

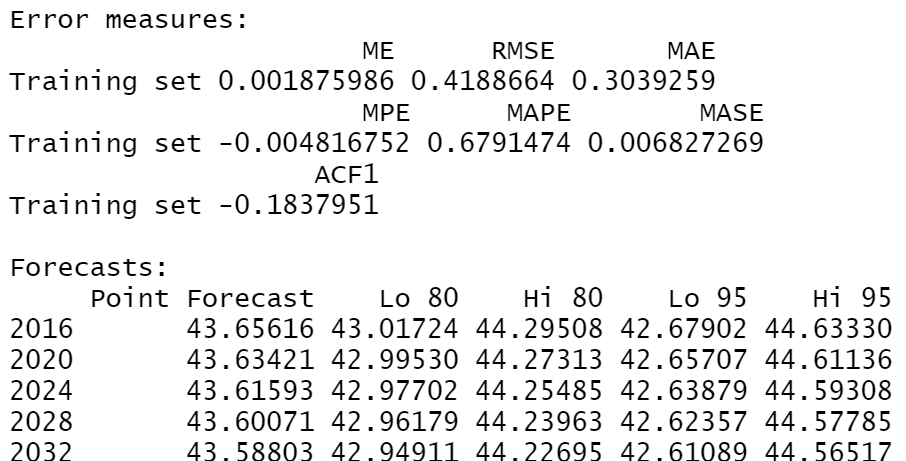
autoplot(mod2)

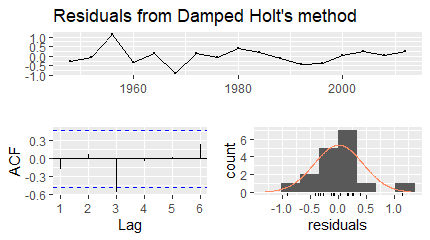
checkresiduals(mod2)

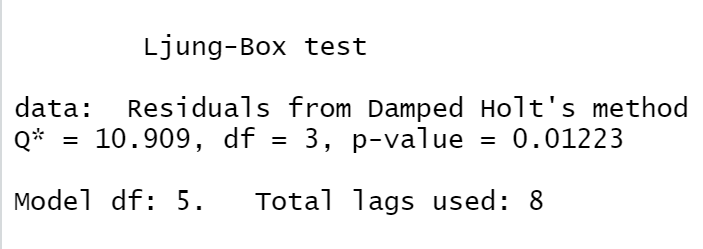
summary(mod2)









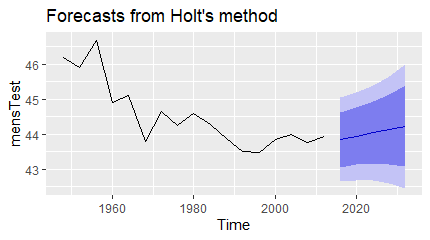


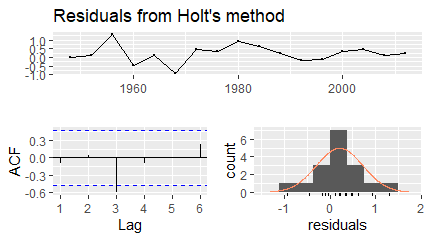
mod3 <- holt(mensTest, h=5)

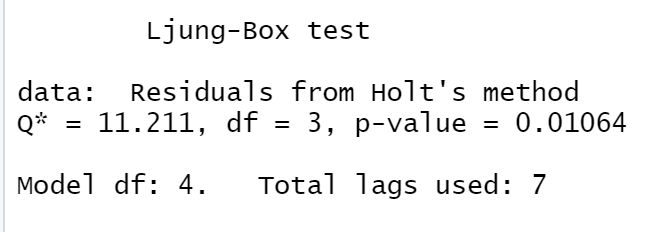
autoplot(mod3)

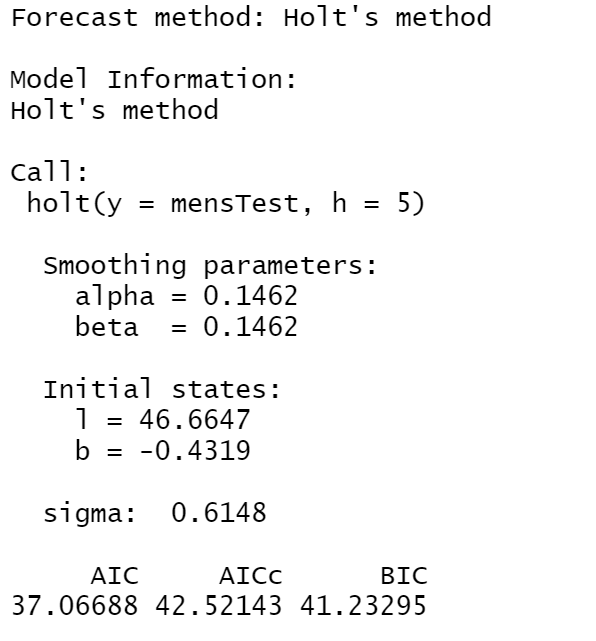
checkresiduals(mod3)

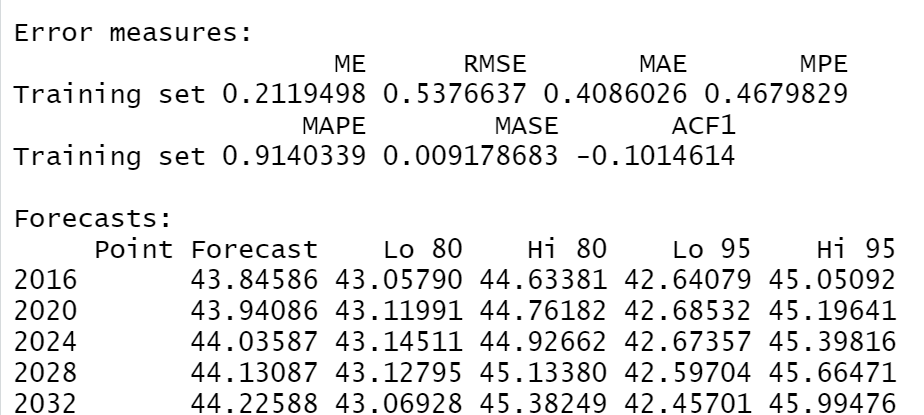
summary(mod3)











The best model from these three is mod1 which is the arima model. This is the best model because, of the three, it is the only one which passes the residual checks. The residuals are centered roughly at 0 and there do not appear to be any patterns in the ACF. The residuals do not fit the normal curve exactly, but that doesn’t indicate that our model is bad. For the Ljung-Box Test the null hypothesis is that autocorrelation is not present in the model and the alternative hypothesis is that autocorrelation is present. The p-value from this test is 0.2336, which is well above the level of significance of 0.05, indicating that we do not reject the null hypothesis and that autocorrelation is not present.

**2f: How good was your model of choice from part 2d in predicting the actual value of 2016?**

Our model of choice predicted that the value in 2016 would be 43.12479. The actual value for 2016 was 43.03. The difference between forecasted and actual was 0.09479, which means the forecast was quite close to the actual value.

**2g: In this part, use only the values from 1980 to 2012 to build your model. Please give three forecasting options and explain which one you would choose as the best option.**

mensTest2 <- window(mens400,start=c(1980,1), end=c(2012,1))

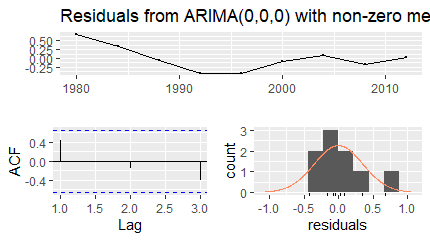
mod1a <- auto.arima(mensTest2)

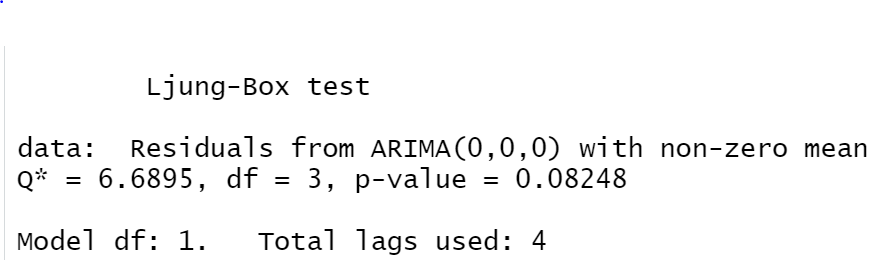
checkresiduals(mod1a)

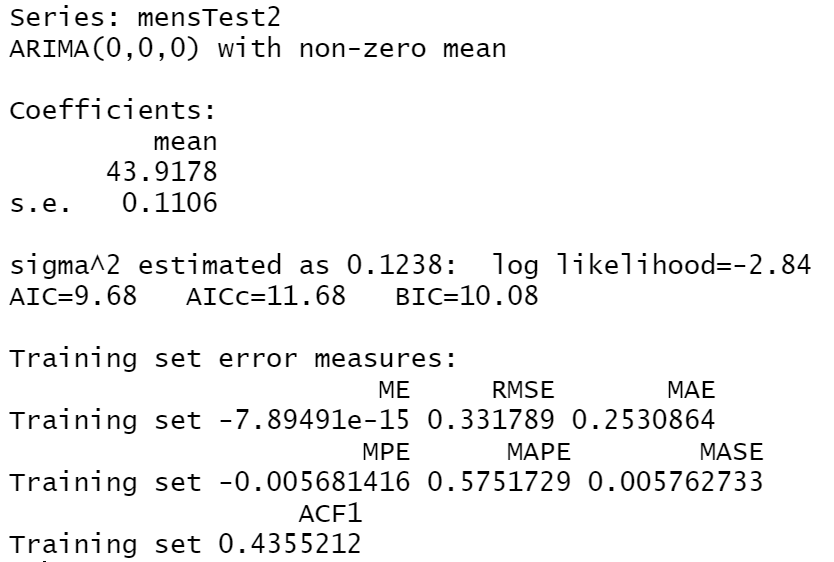
summary(mod1a)

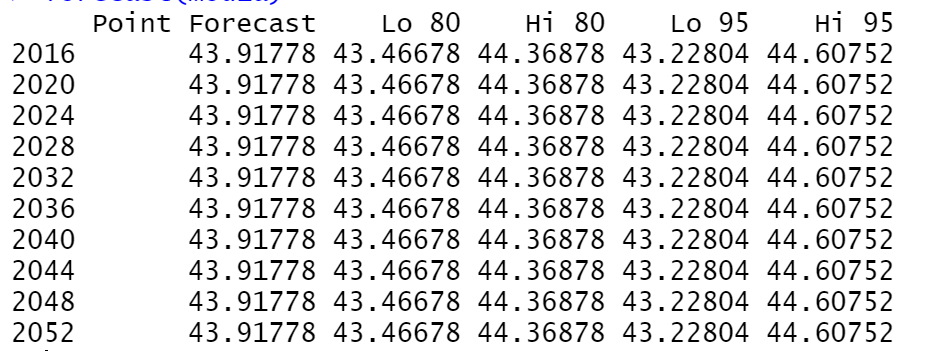
forecast(mod1a)

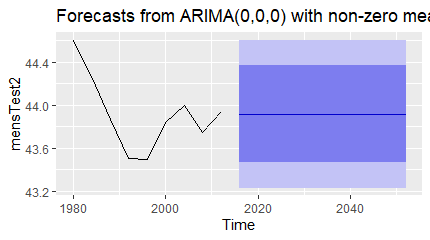
autoplot(forecast(mod1a))









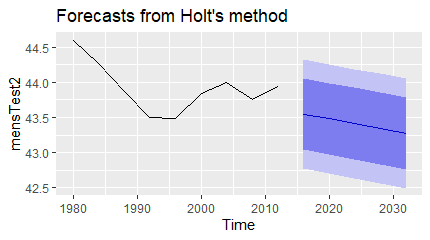


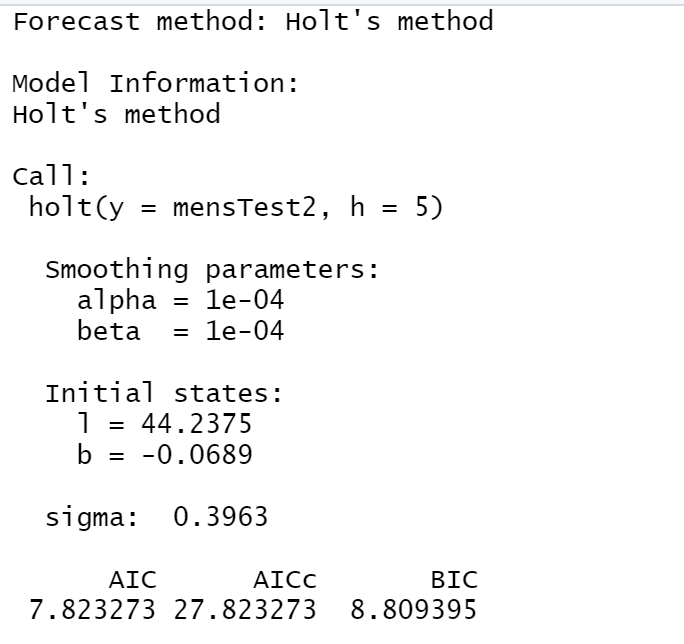
mod2a <- holt(mensTest2, h=5)

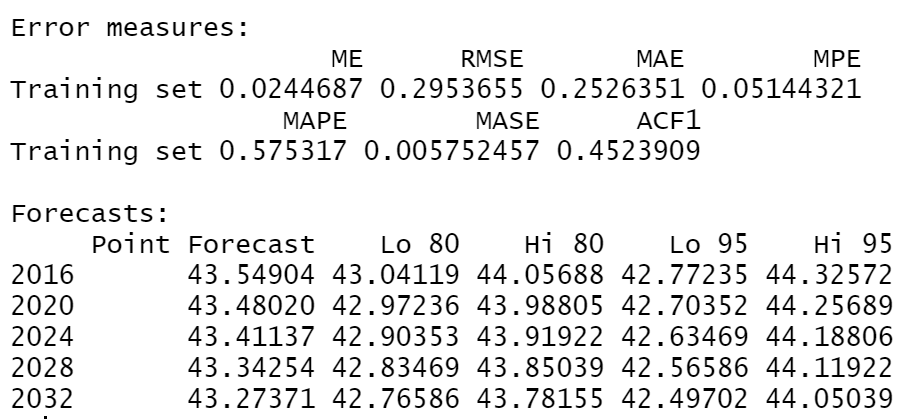
autoplot(mod2a)

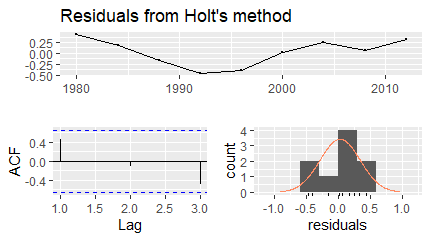
summary(mod2a)

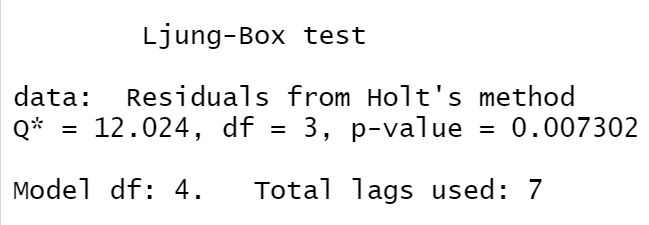
checkresiduals(mod2a)









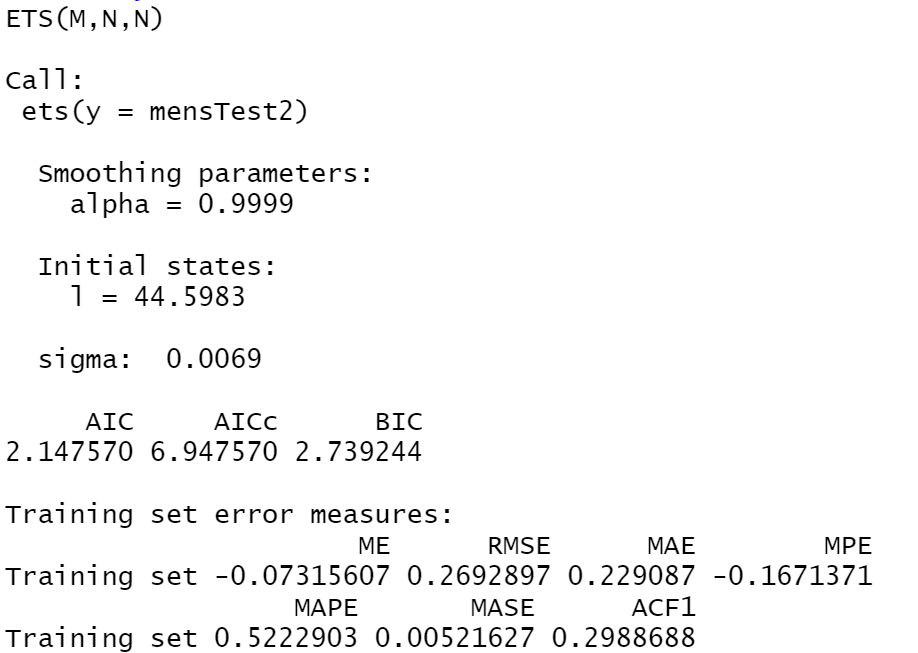


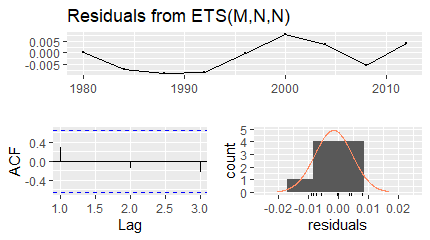
mod3a <- ets(mensTest2)

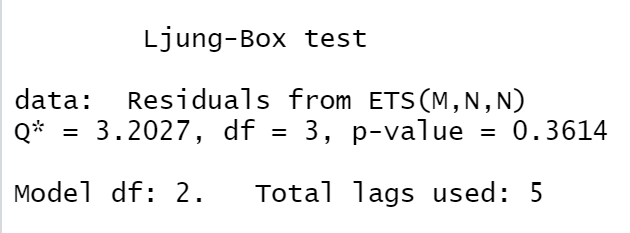
summary(mod3a)

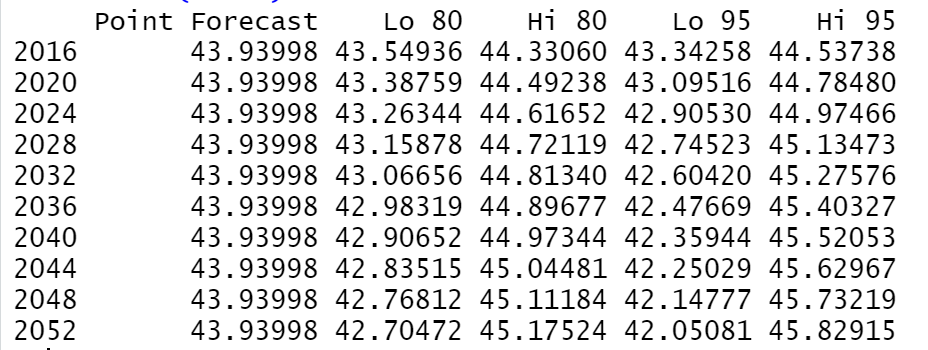
checkresiduals(mod3a)

forecast(mod3a)









The best model from these three is mod3a, which is an ETS model. This model is best because it has the lowest RMSE of 0.2692897. Additionally, the residual checks are the best of the three models. The residuals are centered approximately at 0, which is necessary. The lags don’t appear to form a pattern, however it is difficult to tell with only 3, and they somewhat fit a normal curve. Neither of these are necessities for a good model. For the Ljung-Box Test the null hypothesis is that autocorrelation is not present in the model and the alternative hypothesis is that autocorrelation is present. The p-value from this test is 0.3614, which is well above the level of significance of 0.05, indicating that we do not reject the null hypothesis and that autocorrelation is not present.

**2h: How good was your model of choice from part 2g in predicting the actual value of 2016?**

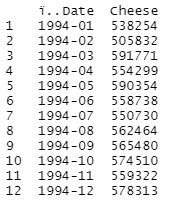
The value predicted from mod3a, the ETS model, for 2016 was 43.93998. The actual value for 2016 was 43.03. The difference between forecasted and actual was 0.90998. The prediction could have been better, as the time is close to 1 second different than the actual value, which is significant in the spread of times in this data.

**2i: Overall, in your opinion, which model would be best to use and why?**

The best model to use for this data is the ARIMA(0,1,1). This model passed all of the necessary checks, making it a valid model. Additionally, the model has the closest forecast to the actual value for 2016 and provides a forecast that continues to decrease into the future that is consistent with what we would believe to be true based on the data set.

**Question 1: Your data**

**Part a: I want 20 years of data minimum. Need annually and then either quarterly or monthly. Turn this dataset into a CSV file and load it into RStudio as (mydata1). Please show me ONLY the first 12 lines of your data. I will need you to upload your data set in Excel format for me to check.**

****

**Part b: Why did you pick this data set and write a paragraph on this data set.**

This dataset is monthly cheese production from 1994 to 2019. We chose this dataset because we thought that cheese production can have very interesting analysis associated. Such as it would be interesting to see if there is seasonality in cheese production because I would imagine that cows always produce milk. Additionally, with the rise of popularity of going vegan, vegetarian, and non-dairy, would that impact cheese production? Lastly, COVID-19 has impacted many industries, it would be interesting to see if it has also impacted cheese production.

**Part c: How did you collect the data and if you needed to clean the data explain how you did it.**

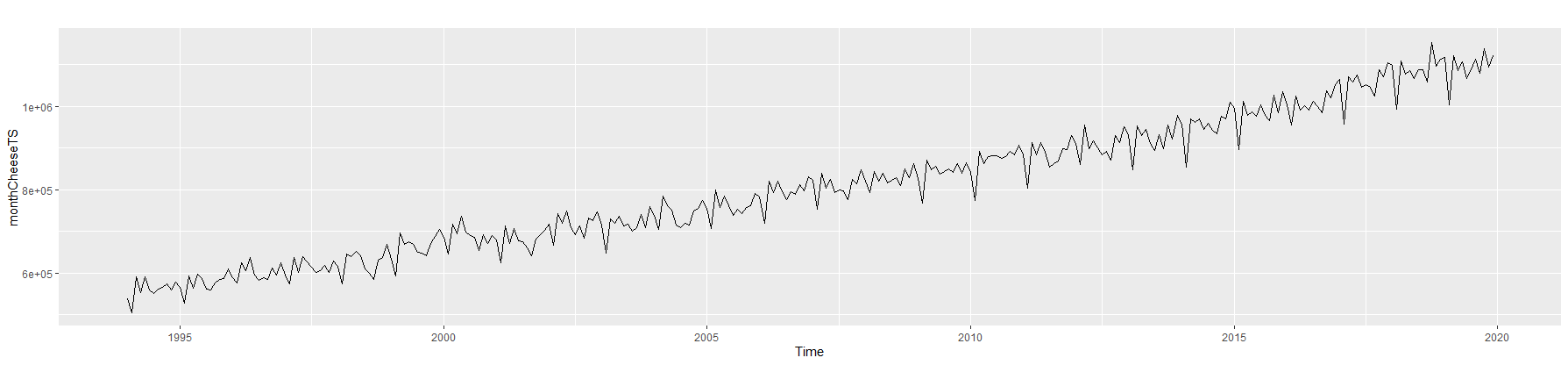
The data was collected from The USDA National Agricultural Statistics Service It was collected by selecting February’s pdf report and copying the previous year’s cheese production from the report into a csv file.

https://usda.library.cornell.edu/concern/publications/m326m1757?locale=en&page=37#release-items

**Part d: Load your data as an annual time series and other (quarter/monthly) and plot the two time plots. Explain any important points of these graphs.**

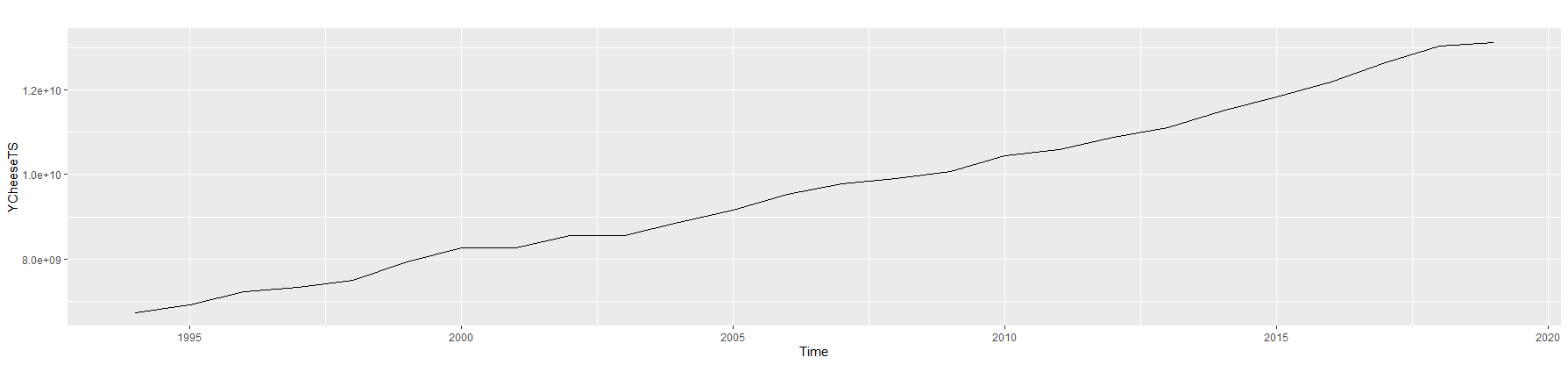
monthCheeseTS <- ts(myData1$Cheese,frequency=12,start=c(1994,1))

autoplot(monthCheeseTS)



YCheeseTS <- ts(myData2$CheeseY,frequency=1,start=c(1994,1))

autoplot(YCheeseTS)

****

There is a steadily increasing trend visible in both the monthly and yearly data. However, a seasonal pattern is apparent in the monthly data with a large downward spike occurring in february of each year.

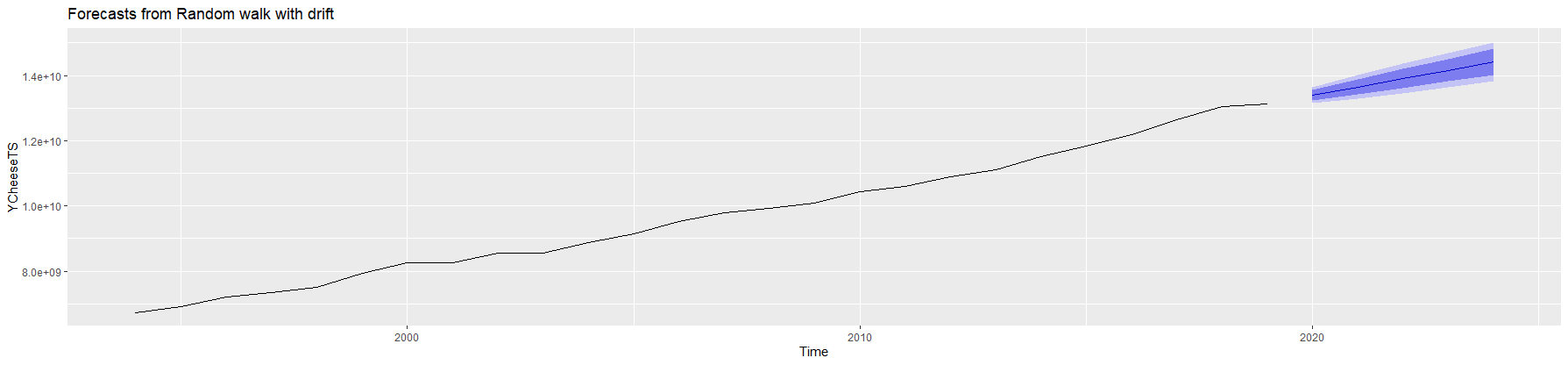
**Part e: Project the next five years using two techniques for annual data. Showing all expected details.**

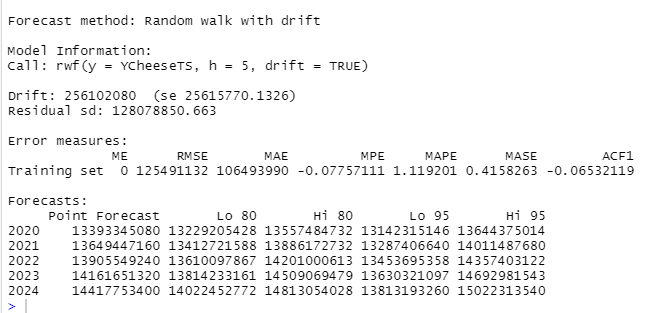
mod1 <- rwf(YCheeseTS,h=5, drift=TRUE)

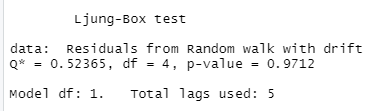
summary(mod1)

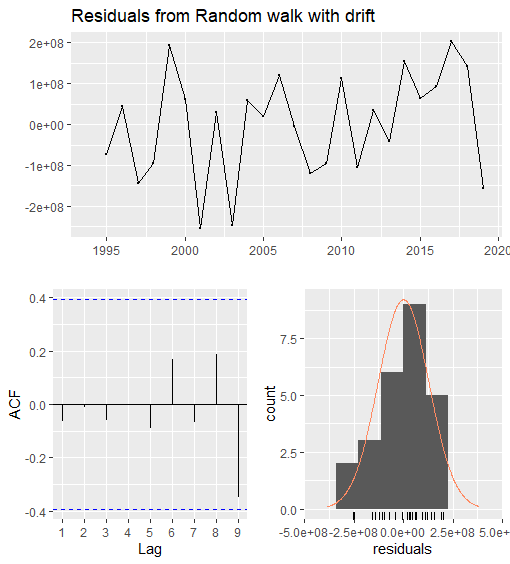
autoplot(mod1)

checkresiduals(mod1)









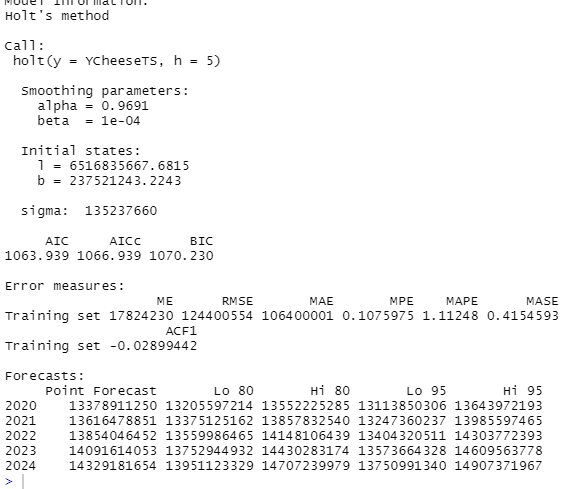
The random walk function with drift is a good model for the yearly data. It passes the check residuals such as, the residuals are centered around zero and random, normally distributed, and does not pass the blue lines in the acf graph. Additionally there is no autocorrelation based of the ljung-box test with H0 is that there is no autocorrelation and H1 is that there is autocorrelation; The p-value is .9712 which is greater than .05 meaning that we do not reject the null.

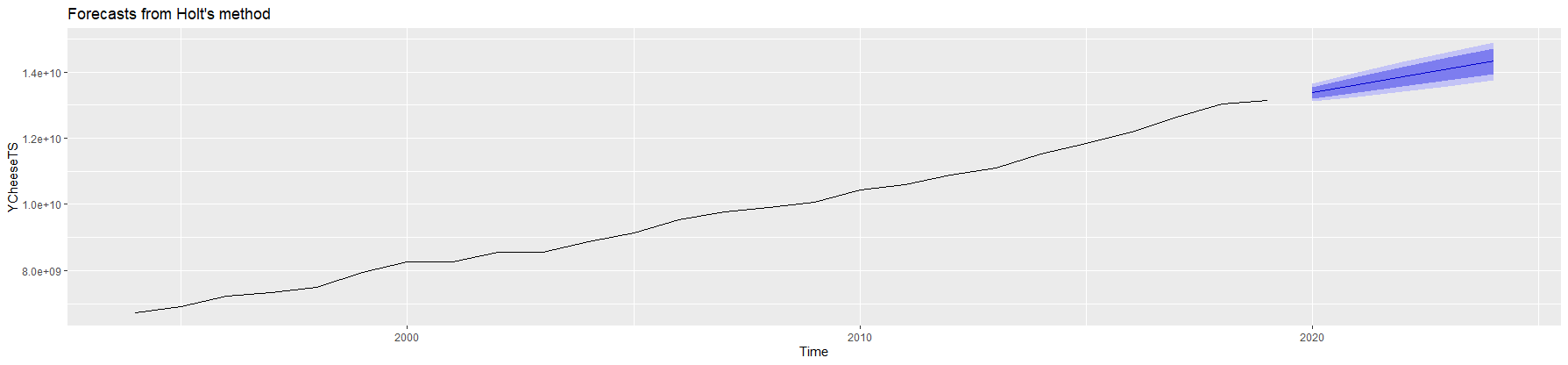
mod2<-holt(YCheeseTS,h=5)

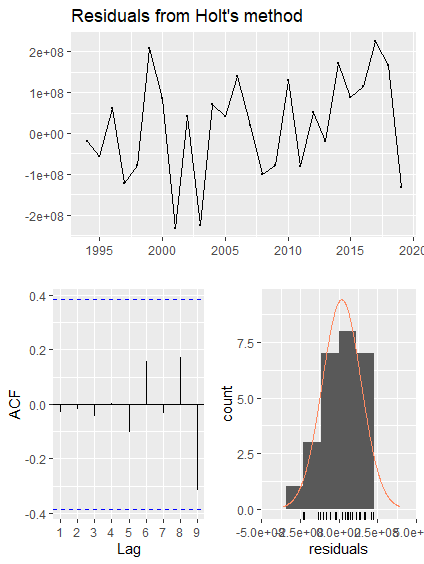
summary(mod2)

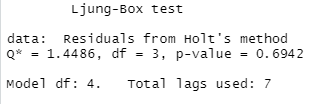
autoplot(mod2)

checkresiduals(mod2)









Holt's method is also a good model for the yearly data. It passes the check residuals such as, the residuals are centered around zero and random, normally distributed, and does not pass the blue lines in the acf graph. Additionally there is no autocorrelation based of the ljung-box test with H0 is that there is no autocorrelation and H1 is that there is autocorrelation; The p-value is .6942 which is greater than .05 meaning that we do not reject the null.

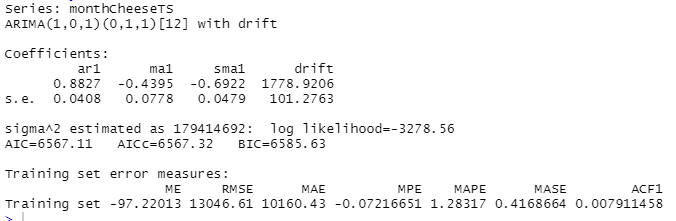
**Part f: Project the next five years using two techniques for the (quarter/monthly) data. Showing all expected details.**

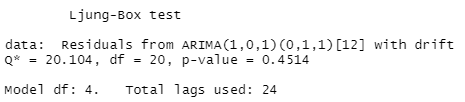
mod3<- auto.arima(monthCheeseTS)

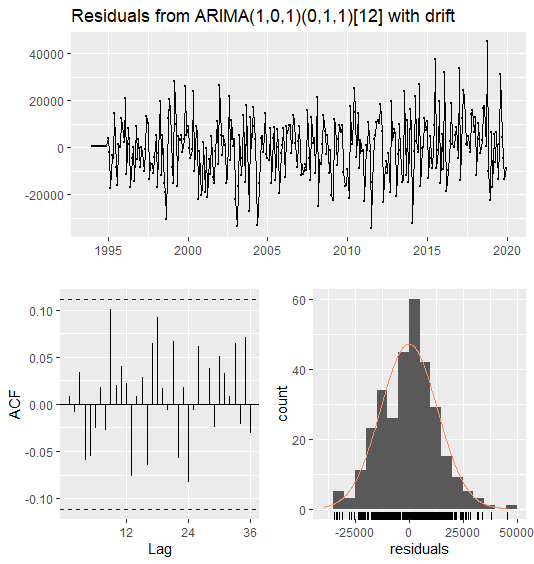
summary(mod3)

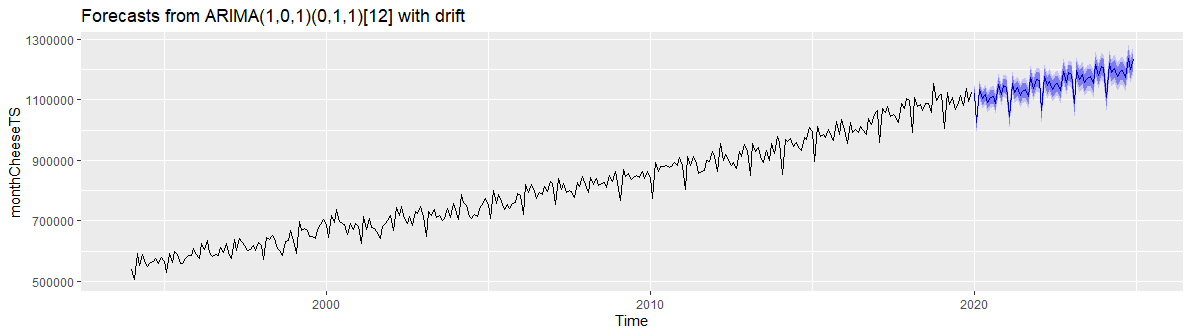
checkresiduals(mod3)

autoplot(forecast(mod3, h=60))









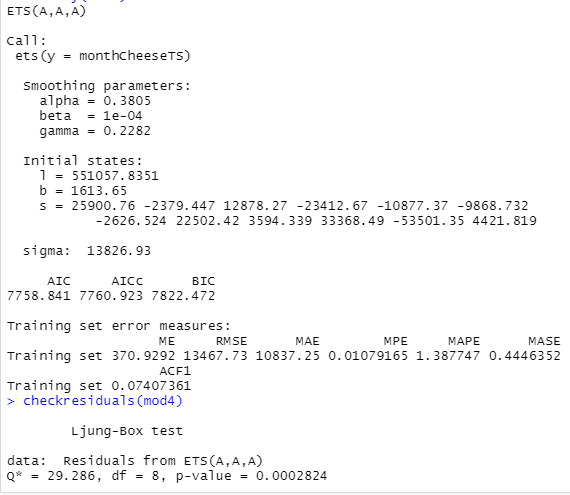
Arima model is also a model for the monthly data. It passes the check residuals such as, the residuals are centered around zero and random, normally distributed, and does not pass the blue lines in the acf graph. Additionally there is no autocorrelation based of the ljung-box test with H0 is that there is no autocorrelation and H1 is that there is autocorrelation; The p-value is .4514 which is greater than .05 meaning that we do not reject the null.

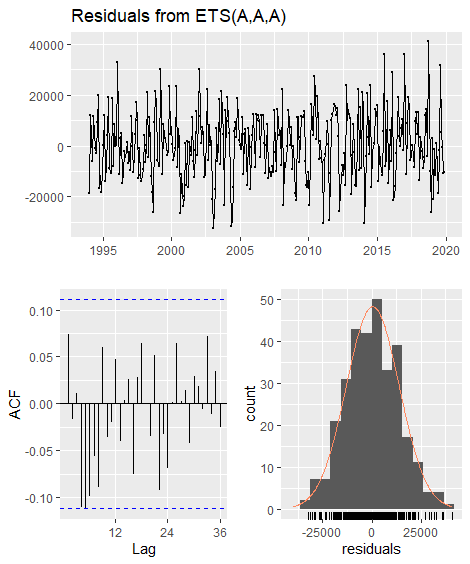
mod4<- ets(monthCheeseTS)

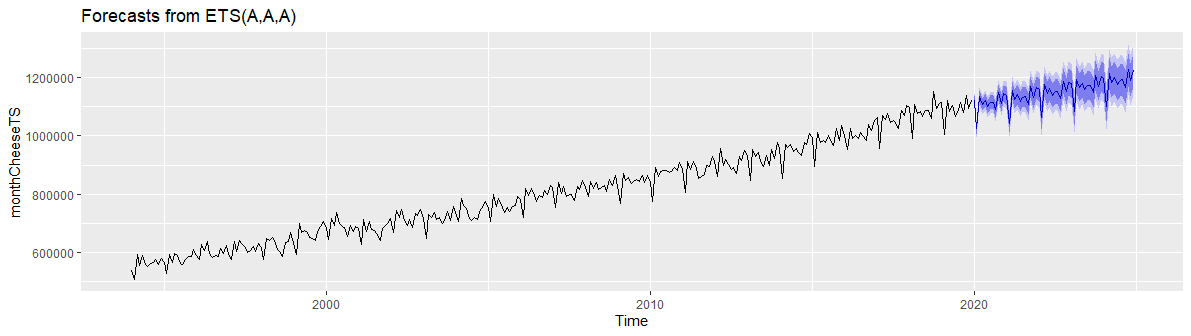
summary(mod4)

checkresiduals(mod4)

autoplot(forecast(mod4, h=60))







ETS model is not a good model for the monthly data. It passes some of

the check residuals such as, the residuals are centered around zero and random, normally distributed, and does not pass the blue lines in the acf graph. Additionally there is autocorrelation based of the ljung-box test with H0 is that there is no autocorrelation and H1 is that there is autocorrelation; The p-value is .0002824 which is less than .05 meaning that we reject the null.

**Part g: Given all the previous parts, please write a conclusion on the forecasts of your data set.**

Provided all of the previous parts, the best forecast for the cheese production dataset would be Arima(1,0,1)(0,1,1)[12] because it passes all of the residual checks, has the best RMSE, MAE, and MAPE, and accounts for seasonality. Additionally to answer the question of, “Has COVID-19 impacted cheese production?” The Arima model accurately predicted the 2020 cheese production regardless of the current pandemic with an exception with April which has a decrease in sales outside of the predictive model. In may, American Cheese Society (ACS) surveyed 980 artisan cheese community members including cheesemakers, cheesemongers, retailers, buyers, and other industry stakeholders done by the American Cheese Society, 73% of respondents claimed to have used a pause to be able to improve their business operations, which accurately reflects the drop in production in April.

Actual 2020 monthly data

