Final Project

Feng Qiao Lin

4/14/2021

Abstract

The goal of this report is to identify a ARIMA or SARIMA model that will best fit the data “salmon” from the *astsa* package in R. Using the method of maximum likelihood to estimate the parameters, forecasting can be done using said estimated parameters. The prices of farm bred Norwegian salmon, from September 2003 to June 2017, can be modelled using a seasonal ARIMA model: SARIMA(0,1,1)(0,0,1)12. A 10-month forecast shows that that price of salmon will remain at approximately $8/kilogram, subject to the seasonal price changes that previous salmon prices went through. Salmon buyers are encouraged to inform themselves of the seasonal swings of the salmon prices, before making an informed decision when purchasing salmon.

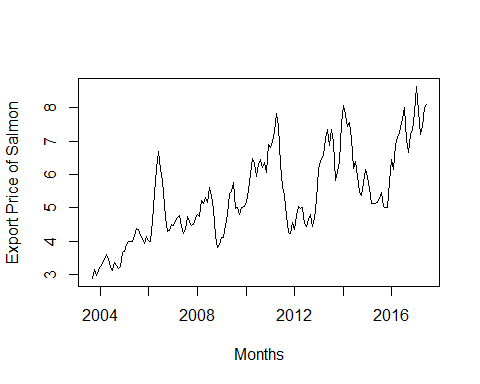
Introduction

Salmon is a commodity that is enjoyed globally, whether it be eaten smoked, grilled, baked, or raw. With such popular demand comes popular supply, and our aim is to fit a time series model using the data set “salmon” in the astsa package in R. This data set includes prices of farm bred Norwegian salmon in dollars per kilogram in the US, from September 2003 all the way to June 2017. After some inspection of the data, there seems to be hints of seasonality in the pricing of salmon, leading to a seasonal version of the ARIMA model to be fit. Salmon prices tend to fall during the summer every year (June, July, August and September), and begin to rise again during winter and spring (e.g November through April). This is due to growth cycles on salmon farms leading to faster salmon growth during certain periods of time in a year. Another factor in the seasonality of salmon is the nature of salmon breeding and the opening of the wild salmon season. Salmon season typically brings in an abundant diversity in the type of salmon, increasing the supply and driving the salmon prices down. After fitting a seasonal ARIMA model, forecasting and spectral analysis of the salmon prices will be done to give a better understanding of the trends of the salmon price and give a 10 month forecast based on the model.

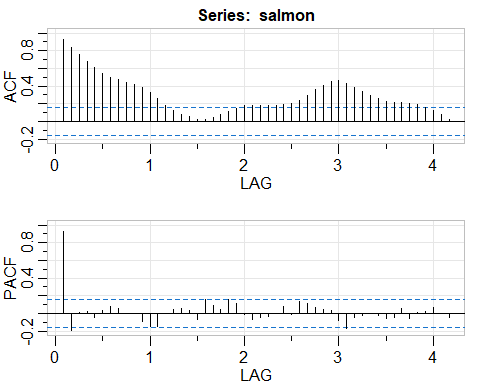
Statistical Methods Employed

The data was first plotted on a graph to assess the stationarity of the data, and whether transformations or differences would be needed to make the data stationary. After a 12 month seasonal difference and another regular difference on top of that, the ACF and PACF of the differenced data were examined and 3 models were proposed: SARIMA(0,1,1)(0,0,1)[12], SARIMA(1,1,0)(2,0,0)[12], and SARIMA(1,1,1)(2,0,1)[12]. The first model focuses on the MA terms, the second model focuses on the AR terms, while the last model incorporates both. Both the ACF and the PACF is cutting off at lag 1, while the seasonal lags cut off at 1 for the ACF and 2 for the PACF. This is the reasoning behind the models that were proposed. There is much more statistical method explanations in greater detail scattered throughout the report, because it made more sense to include the explanation with the data visualizations for better understandings. Results and data visualizations start below.

##Load the required libraries  
library(astsa)  
library(forecast)  
#plot the data  
plot.ts(salmon, xlab = "Months", ylab = "Export Price of Salmon")

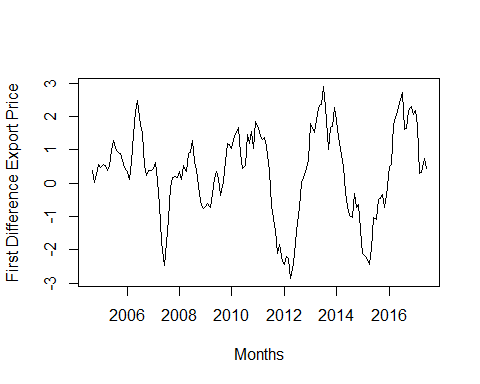


#check the acf  
acf2(salmon, 50)

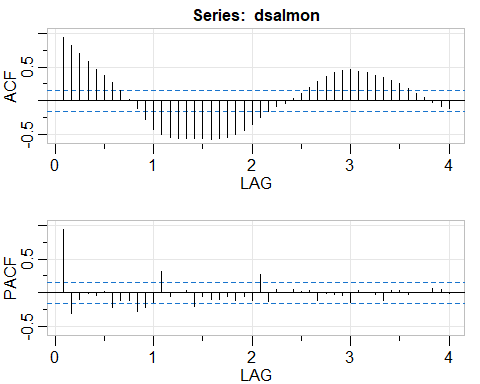


Since the sample acf does not decay to zero fast as h increases, this is a sign that differencing may be needed. There seem to be signs of a possible seasonal factor to the price of salmon, so we will difference it seasonally.

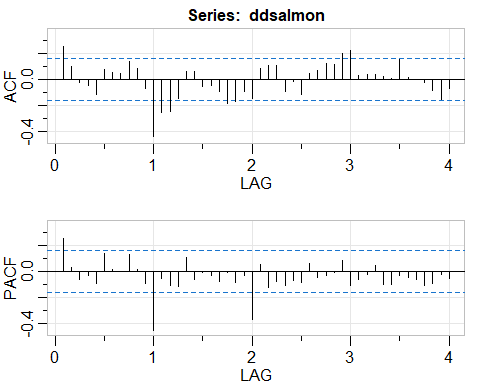
#seasonally difference and plot the data  
dsalmon = diff(salmon, 12)  
plot.ts(dsalmon, xlab = "Months", ylab = "First Difference Export Price")



#check the acf again  
acf2(dsalmon)

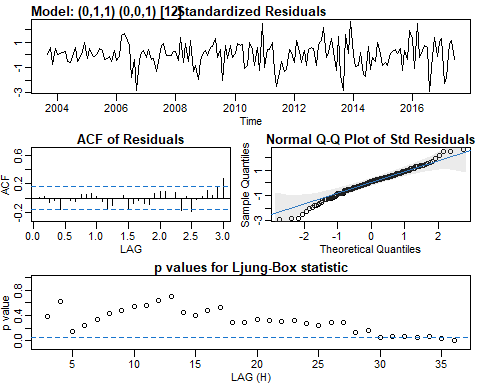


#difference it again since there is a trend  
ddsalmon = diff(dsalmon)  
acf2(ddsalmon)



Both the ACF and the PACF is cutting of at lag 1, while the seasonal lags cut off at 1 for the ACF and 2 for the PACF. We will propose 3 models: SARIMA(0,1,1)(0,0,1)12, SARIMA(1,1,0)(2,0,0)12, and SARIMA(1,1,1)(2,0,1)12. The first model focuses on the MA terms, the second model focuses on the AR terms, while the last model incorporates both.

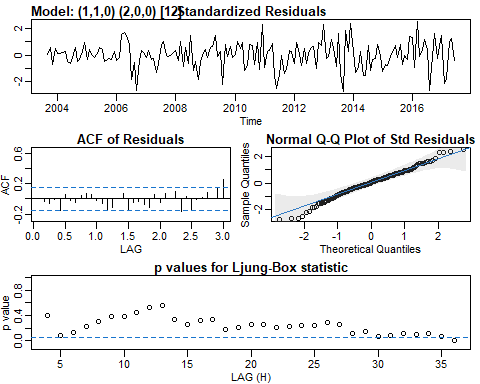
#fitting the 3 models  
sarima(salmon, 0, 1, 1, 0, 0, 1, 12)



The SARIMA(0,1,1)(0,0,1)[12] seems the fit the data well based off the model diagnostics. Not many residuals that are 3 standard deviations away, most of the residuals do not have significant spikes in the ACF, the QQ plot shows approximate normality, and the p-values for the Ljung-Box statistic are above the blue line for the majority of the lags. Overall, this model is a good fit and can be considered for the forecasting model. In addition, the p-values for the coefficients shown below are significant at all significant levels, but the constant is not significant at any significance level.

## $fit  
##   
## Call:  
## stats::arima(x = xdata, order = c(p, d, q), seasonal = list(order = c(P, D,   
## Q), period = S), xreg = constant, transform.pars = trans, fixed = fixed,   
## optim.control = list(trace = trc, REPORT = 1, reltol = tol))  
##   
## Coefficients:  
## ma1 sma1 constant  
## 0.2680 0.2273 0.0313  
## s.e. 0.0715 0.0817 0.0454  
##   
## sigma^2 estimated as 0.1443: log likelihood = -74.75, aic = 157.51  
##   
## $degrees\_of\_freedom  
## [1] 162  
##   
## $ttable  
## Estimate SE t.value p.value  
## ma1 0.2680 0.0715 3.7479 0.0002  
## sma1 0.2273 0.0817 2.7815 0.0061  
## constant 0.0313 0.0454 0.6896 0.4914  
##   
## $AIC  
## [1] 0.9545793  
##   
## $AICc  
## [1] 0.9554827  
##   
## $BIC  
## [1] 1.029875

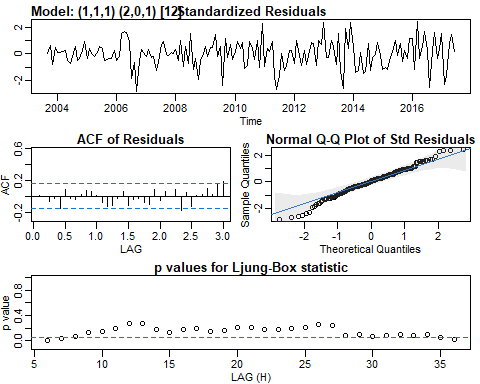
sarima(salmon, 1, 1, 0, 2, 0, 0, 12)



Once again, the SARIMA(1,1,0)(2,0,0)[12] model diagnostics shows many similarities to the previous model. The only major difference being the standardized residuals having a smaller standard deviation than the previous model. As shown below, the first two coefficients are significant at all significance levels, but the seasonal AR term 2 and the constant are not.

## $fit  
##   
## Call:  
## stats::arima(x = xdata, order = c(p, d, q), seasonal = list(order = c(P, D,   
## Q), period = S), xreg = constant, transform.pars = trans, fixed = fixed,   
## optim.control = list(trace = trc, REPORT = 1, reltol = tol))  
##   
## Coefficients:  
## ar1 sar1 sar2 constant  
## 0.2383 0.2218 0.1379 0.0356  
## s.e. 0.0761 0.0799 0.0825 0.0566  
##   
## sigma^2 estimated as 0.1403: log likelihood = -72.79, aic = 155.58  
##   
## $degrees\_of\_freedom  
## [1] 161  
##   
## $ttable  
## Estimate SE t.value p.value  
## ar1 0.2383 0.0761 3.1304 0.0021  
## sar1 0.2218 0.0799 2.7740 0.0062  
## sar2 0.1379 0.0825 1.6702 0.0968  
## constant 0.0356 0.0566 0.6286 0.5305  
##   
## $AIC  
## [1] 0.9429392  
##   
## $AICc  
## [1] 0.9444543  
##   
## $BIC  
## [1] 1.037059

sarima(salmon, 1, 1, 1, 2, 0, 1, 12)

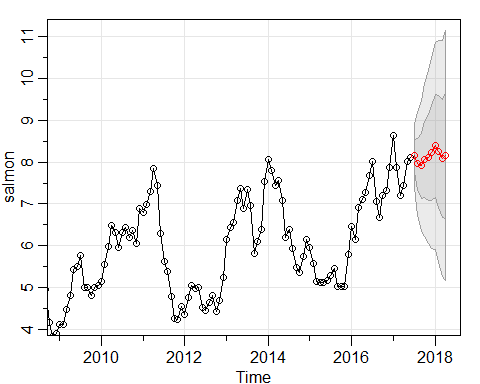


This model shows a much worse Ljung-Box statistic when compared to the previous two models, having much lower p-values and many more p-values below the blue line. The normality is more off than the other two as well. Furthermore, the AR1, MA1, SAR1 and the constant are all not statistically significant at any significance level. This is a bad sign and we will most likely not use this model over the other two.

## $fit  
##   
## Call:  
## stats::arima(x = xdata, order = c(p, d, q), seasonal = list(order = c(P, D,   
## Q), period = S), xreg = constant, transform.pars = trans, fixed = fixed,   
## optim.control = list(trace = trc, REPORT = 1, reltol = tol))  
##   
## Coefficients:  
## ar1 ma1 sar1 sar2 sma1 constant  
## 0.1274 0.1058 0.8490 0.0374 -0.6976 0.0374  
## s.e. 0.2715 0.2688 0.1935 0.1141 0.1871 0.0704  
##   
## sigma^2 estimated as 0.1321: log likelihood = -68.87, aic = 151.74  
##   
## $degrees\_of\_freedom  
## [1] 159  
##   
## $ttable  
## Estimate SE t.value p.value  
## ar1 0.1274 0.2715 0.4695 0.6394  
## ma1 0.1058 0.2688 0.3935 0.6945  
## sar1 0.8490 0.1935 4.3869 0.0000  
## sar2 0.0374 0.1141 0.3278 0.7435  
## sma1 -0.6976 0.1871 -3.7290 0.0003  
## constant 0.0374 0.0704 0.5310 0.5961  
##   
## $AIC  
## [1] 0.9196562  
##   
## $AICc  
## [1] 0.9228783  
##   
## $BIC  
## [1] 1.051424

We will propose the first model for prediction since all of its parameters other than the constant are significant and it satisifies all of the model assumptions. The SARIMA(1,1,0)(2,0,0)[12] has one coefficient that is not significant.

#forecasting the price of salmon 10 months ahead using MA model  
pred = sarima.for(salmon, 10, 0, 1, 1, 0, 0, 1, 12)

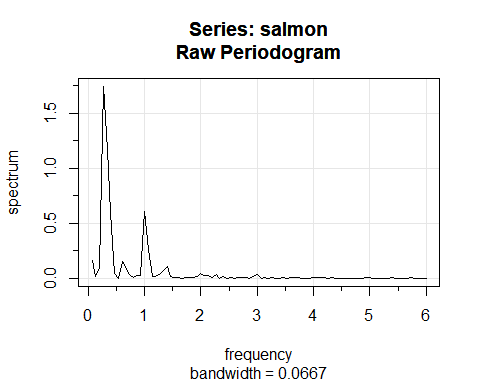


The prediction seems to follow the trends, and the confidence interval is very wide for these predictions.

#calculate the confidence interval for the prediction  
upper = pred$pred + qnorm(0.975) \* pred$se  
lower = pred$pred - qnorm(0.975) \* pred$se  
data.frame("Prediction" = pred$pred, "95% Lower Bound" = lower, "95% Upper Bound" =   
 upper)

## Prediction X95..Lower.Bound X95..Upper.Bound  
## 1 8.159462 7.415020 8.903905  
## 2 7.956584 6.754423 9.158745  
## 3 7.913200 6.384742 9.441659  
## 4 8.059593 6.263157 9.856029  
## 5 8.113936 6.084607 10.143266  
## 6 8.224325 5.986206 10.462443  
## 7 8.389811 5.960785 10.818838  
## 8 8.255066 5.649080 10.861053  
## 9 8.082074 5.310403 10.853746  
## 10 8.151525 5.223530 11.079521

#spectral analysis for the salmon data  
salmon.per = mvspec(salmon, log='no')



#identifying the dominant frequencies for the salmon data  
dom = salmon.per$details[order(salmon.per$details[,3], decreasing = TRUE),]  
dom[1,]

## frequency period spectrum   
## 0.2667 3.7500 1.7437

dom[2,]

## frequency period spectrum   
## 0.3333 3.0000 1.1697

dom[3,]

## frequency period spectrum   
## 1.0000 1.0000 0.6074

#identifying the plotting the 95% confidence interval  
library(MASS)  
salmon.u1 = 2\*dom[1,3]/qchisq(0.025,2)  
salmon.l1 = 2\*dom[1,3]/qchisq(0.975,2)  
salmon.u2 = 2\*dom[2,3]/qchisq(0.025,2)  
salmon.l2 = 2\*dom[2,3]/qchisq(0.975,2)  
salmon.u3 = 2\*dom[3,3]/qchisq(0.025,2)  
salmon.l3 = 2\*dom[3,3]/qchisq(0.975,2)  
result = data.frame(Dominant.Freq=c(dom[1,1], dom[2,1], dom[3,1]), Spec=c(dom[1,3], dom[2,3], dom[3,3]), Lower=c(salmon.l1, salmon.l2, salmon.l3), Upper=c(salmon.u1, salmon.u2, salmon.u3))  
result

## Dominant.Freq Spec Lower Upper  
## 1 0.2667 1.7437 0.4726910 68.87247  
## 2 0.3333 1.1697 0.3170882 46.20068  
## 3 1.0000 0.6074 0.1646570 23.99102

Conclusions and Limitations

Overall, the model fit was pretty good, but it was not perfect. The QQ plot had a significant number of outliers and the Ljung-Box statistic could have been better because the model fit was not good past lags 30+. This begs the question: “What kind of model would have been better?”. Perhaps a different seasonal difference such as 7 or 8 would have been better instead of the 12, aligning with the salmon farm growing cycles.

This report showed that salmon pricing has a seasonal trend behind it, and used a SARIMA(0,1,1)(0,0,1)[12] model to fit the data and forecast 10 months into the future with it. Salmon prices are expected to stay within the $8 dollar range, but seasonal pricing will occur according to the growing cycles of salmon farms and breeding habits of the wild salmon.