STA457 Final Report

Shu Wang

04/13/2022

1 Abstract

Air travelling has become a popular choice for many travellers around the world. In this report, I try to use SARIMA model to fit the number of air passengers and make predictions on the number of air passengers for the next ten months. I managed to make predictions that fits the trend of the data. I conclude that this time series can be fitted into SARIMA model with well-behaved patterns and parameters.

Key words: time series, air passengers, SARIMA model, predictions, dominant frequency

2 Introduction

An airline company needs to know how many passengers it will have in a given term and the profit it will generate.[1]To obtain predictions,the airline

passenger data from Kaggle is used.[2]This data set has 2 columns:time and number of air passengers.It is a monthly-recorded time series ranging from January 1949 to December 1960 with a total of 144 observations.The purpose of this report is to make predictions on number of air passengers and help airlines to manage ahead.This method can be used in other transportation areas like trains and subways.

3 Statistical Methods

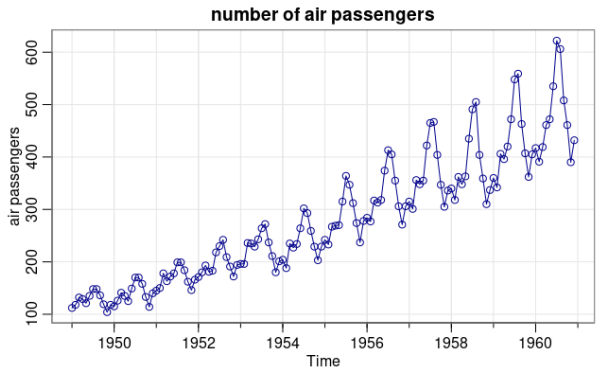

Figure 1: time series plot of number of air passengers

We use rstudio cloud to perform all analysis.[3]From the tsplot and ACF plot,it is obvious that there is a trend in the time series and the variance is non-constant.We can use box-cox transformation and data differencing to make the time series stationary and apply ARIMA model.

The 95 percent confidence interval for λ includes 0. For convenience,we choose 0 to perform a log-transformation.

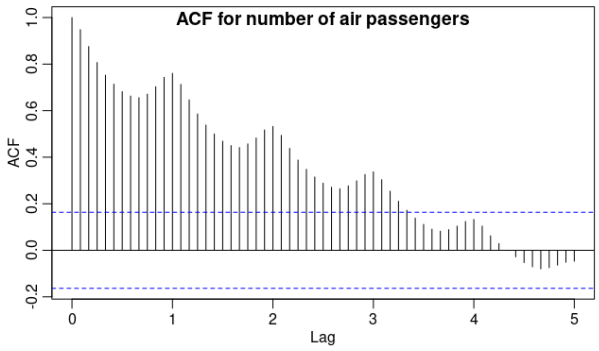

Figure 2: ACF of time series of number of air passengers

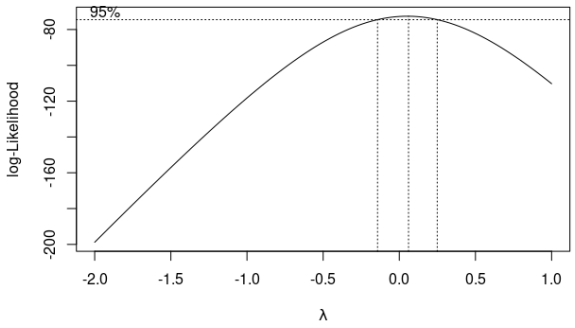

Figure 3: MLE of λ for box-cox transformation

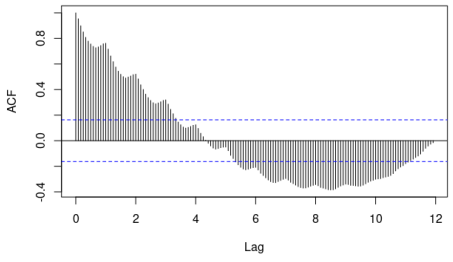

Figure 4: ACF of log of number of air passengers

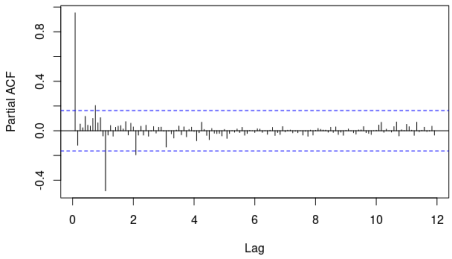

Figure 5: PACF of log of number of air passengers

We plot the decomposition of the transformed data,the ACF and PACF after having confirmed that the Box-Cox transformation was warranted.We also see that there is a high amount of seasonality at about the end of each year. This is also observed in the ACF of the data. The ACF takes on a shape that suggests that there is seasonality in our data. The ACF seems to suggest that we have seasonality at lag=12 as we suspected from the plot of our time series at the beginning of our analysis. We begin by differencing at lag=12 and then address the trend in our data by further differencing.

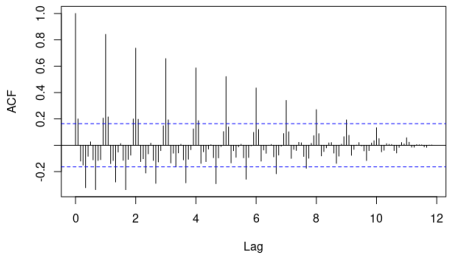

Figure 6: ACF of log of number of air passengers after differencing

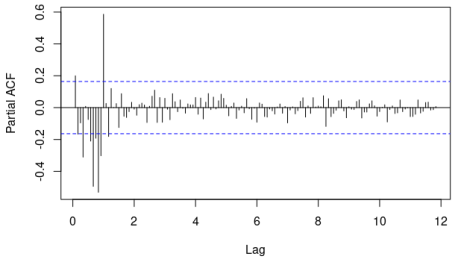

Figure 7: PACF of log of number of air passengers afetr differencing

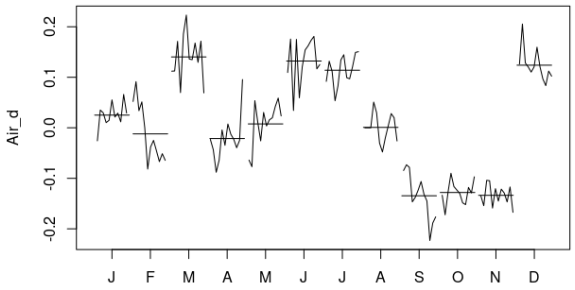

Figure 8: month plot of the differencing time series

The time series is non-stationary,with some seasonality.So we take a seasonal difference with lag = 12.The seasonally differenced time series are shown in the following figure.

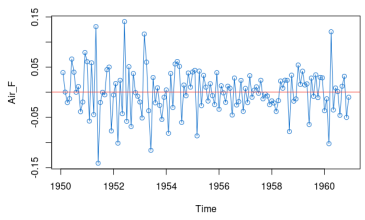

Figure 9: plot of final transformed time series

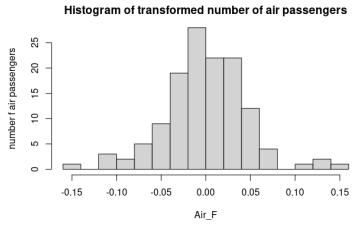

Figure 10: histogram of final transformed time series

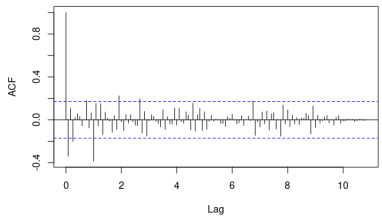

Figure 11: ACF of final transformed time series

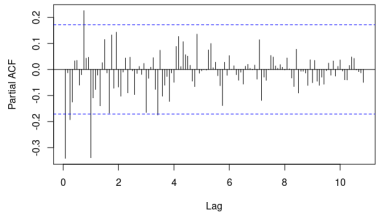

Figure 12: PACF of final transformed time series

As we can see from below,our box-cox method and differencing stabilizes the variance.Now,we perform an adf test on the final transformed data,and test the normality of this data.We observe that the normality is held and condition of stationary is satisfied.

Our aim is to find an appropriate ARIMA model based on the ACF and PACF shown in the above ACF and PACF.The significant spike at lag 1 in the ACF suggests a seasonal MA(1) component.$Q = 1, P = 0$.It is also reasonable to believe that PACF cuts off at lag 1,and ACF tails off.So we

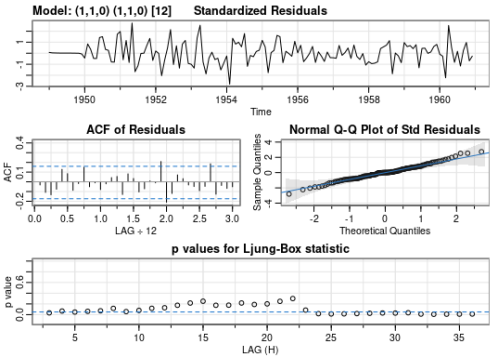

Figure 13: diagnostics of model 1

coefficient	Estimate	SE	t.value	p.value
ar1	-0.3745	0.0808	-4.6319	0
sar1	-0.4637	0.0808	-5.7314	0

Table 1: coefficient of model 1

can also purpose $P = 1, Q = 0$.For the non-seasonal components,since ACF cuts off at first lag,I purpose $q = 1, p = 0$.Also,the reversed argument is that $p = 1, q = 0$.In conclusion,I have 2 choices (0 and 1) for p,q,P,Q respectively.I begin with an ARIMA $(1, 1, 0) * (1, 1, 0)_{12}$

The following are possible suggested models:

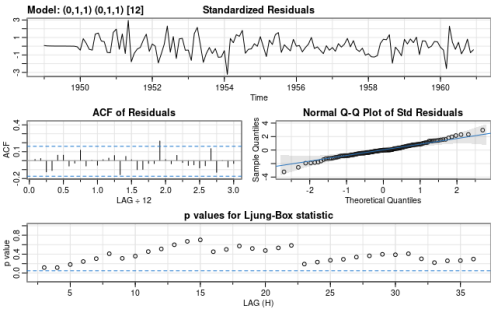

Figure 14: diagnostics of model 2

coefficient	Estimate	SE	t.value	p.value
ma1	-0.4018	0.0896	-4.4825	0
sma1	-0.5569	0.0731	-7.6190	0

Table 2: coefficient of model 2

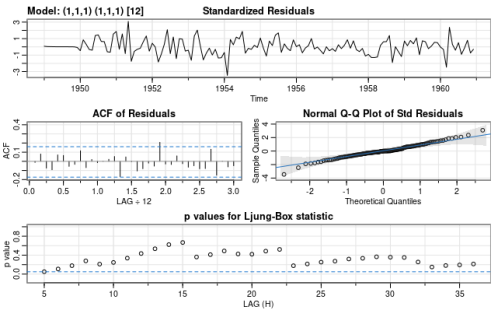

Figure 15: diagnostics of model 3

coefficient	Estimate	SE	t.value	p.value
ar1	0.1666	0.2459	0.6777	0.4992
ma1	-0.5615	0.2116	-2.6542	0.0090
sar1	-0.0990	0.1540	-0.6430	0.5214
sma1	-0.4973	0.1360	-3.6577	0.0004

Table 3: coefficient of model 3

4 Results

4.1 Parameters for Proposed Models

For the proposed models,we have four final SARIMA model:

model 1:SARIMA $(1,1,0)*(1,1,0)_{12}$ model 2:SARIMA $(0,1,1)*(0,1,1)_{12}$

model 3:SARIMA $(1,1,1)*(1,1,1)_{12}$

Which are all in the form of SARIMA$(p,d,q)*(P,D,Q)_S$ where p is the order of the non-seasonal AR(p) model,d is the order of ordinary differencing which $=1$ in our models.q is the order of the non-seasonal MA(q) model.P is the order of the seasonal AR(P) model,which is 0 in our models.D is the order of seasonal differencing,which is 1 in our model.Q is the order of the seasonal MA(Q) model,S is the order of seasonal differencing,which is 12.

4.2 Interpretation of Parameters

The mathematical formula is $\Phi_P(B^S)\phi(B){\nabla_S}^D\nabla^d x_t = \delta + \Theta_Q(B^S)\theta(B)\omega_t$,where $\nabla^d = (1-B)^d$ and ${\nabla_s}^D = (1-B^s)^D$,where x_t is $log(AirP)$

$\Phi_P(B^S)$ and $\Theta_Q(B^S)$ are the polynomial for seasonal component in AR(P) and MA(Q) respectively.$\phi(B)$ and $\theta(B)$ are the polynomial for non-seasonal component in AR and MA(q).${\nabla_S}^D$ is the differencing order for seasonal component,which is 12 in our case. ∇^d is the differencing order for non-seasonal part,which is 1.

4.3 Significance of Parameter Estimates

Now we examine these models.For parameters,at 5% significance level,for model 1 and 2,all the coefficients in these models have p value ≤ 0.05.Hence,we have strong evidence against the null hypothesis that these coefficients are 0 in model 1 and 2.In model 3,the p value for AR(1) and SAR(1) are ≥ 0.05,which suggests that this coefficient is not significantly different from 0. We first rule our model 3.

4.4 Diagnostics for Proposed Models

For the remaining three models.If the model fits well,the standardized residuals should behave as an $i.i.d$ sequence.The ACF of residuals in model 1 shows a significant spike at lag 1,which suggests that the residuals might not be independent.The normality of the residuals are satisfied in all 3 models. For the Ljung-box test,model 1 suggests that approximately half of the lags fails the Ljung-box test,with p value ≤ 0.05.Model 2 suggests that all lags passes the Ljung-box test.

4.5 Model Selection

Combine all these diagnostics together,we choose model 2 as our final model.

4.6 Estimation

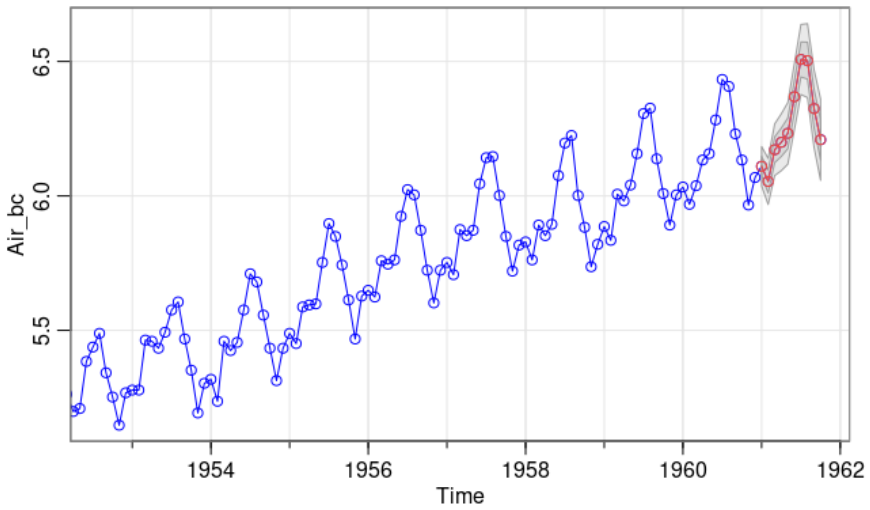

Figure 16: month plot of the differencing time series

This estimate fits perfectly the trend as we observed from the data.

We cannot establish the significance of the first peak since the periodogram ordinate is 0.0676,which lies in the confidence intervals of the second and third peak. We cannot establish the significance of the second peak,since the periodogram ordinate is 0.0201,which lies in the confidence intervals of the third peak. We cannot establish the significance of the third peak,since the periodogram ordinate is 0.0068,which lies in the confidence interval of the second peak.

Month	Predict of log AirP	CI_lower	CI_upper
Jan 1961	6.110186	6.038224	6.182147
Feb 1961	6.053775	5.969922	6.137628
Mar 1961	6.171715	6.077459	6.265971
Apr 1961	6.199300	6.095680	6.302920
May 1961	6.232556	6.120351	6.344761
Jun 1961	6.368779	6.248600	6.488957
Jul 1961	6.507294	6.379639	6.634949
Aug 1961	6.502906	6.368189	6.637623
Sep 1961	6.324698	6.183271	6.466125
Oct 1961	6.209008	6.061176	6.356841

Table 4: Prediction of log_AirP and 95% prediction intervals

Series	Dominant.Freq	Spec	Lower	Upper
log_AirP	1.0000	0.0676	0.0226	1.3179
log_AirP	2.0000	0.0201	0.0067	0.3919
log_AirP	0.0833	0.0068	0.0023	0.1326

Table 5: Three dominant frequencies and 90% confidence intervals

5 Discussion

It is possible to predict number of air passengers for a given term using SARIMA model with certain transformations on original data.I use SARIMA model based on the log-transformation of the original data.$\lambda = 0$ is not the exact value.Hence,we need to test whether this transformation actually stabilizes the variance.Also,this time series is based on data from 1949-1960,which is 60 years from now,so the prediction on the following 10-month period might not be useful in practical.In conclusion,this model can be used to predict number of air passengers with high accuracy.

6 Reference

[1]OptiWisdom. (2020, February 12). Predicting number of passenger on next month with ai. Medium.
Retrieved April 17, 2022, from https://optiwisdom.medium.com/predicting-number-of-passenger-on-next-month-with-ai-8d34dce80f6d

[2]Rocha, G. (2021, May 2).Airpassengers Time series. Kaggle.
Retrieved April 14, 2022, from https://www.kaggle.com/code/georgerocha/airpassengers-notebook/data.

[3]RStudio Team (2020). RStudio: Integrated Development for R. RStudio, PBC, Boston, MA URL http://www.rstudio.com/.