**Abstract**

This report summarizes the statistical modelling and forecast results associated with the sales data. After the remainder analysis, seasonal component analysis, and comparison of diagnostic plots and AIC, AICc, and BIC criteria, ARIMA is evaluated to be the best model, and the forecast sales for 2017 are 3.477666, 3.275568, 3.632091, 3.531044, 3.637497, 3.458033, 3.456673, 3.400179, 3.258374, 3.367593, 3.294429, and 3.434967

**Univariate Time Series Data Analysis**

As can be seen from figure 1, there is clear upward trend in this time series data, as the sales increase overall over time. The mean is not constant over time. Therefore, the data is not stationary, and seasonal component may be present. The variance of the data is constant over time, and no obvious outliers can be identified. With 216 data points, it is difficult to judge whether there is seasonality. As a result, the monthly means of data are calculated, and then compared to see if there is a monthly difference. Figure 2 indicates that there is a monthly difference (seasonality). In addition, based on the time series plot of the data (figure1), the seasonal variation is roughly constant in size over time, and the random fluctuations seem constant over time. With seasonal component present, decomposition is applied to the time series data.

**Model Selection**

Additive model ) is an appropriate decomposition model since the seasonal variation seems to be about the same magnitude across time. To decompose the series, the function stl () is used. Figure 3 shows the plot. The decomposition divides the time series into three components namely: the trend, seasonality, and remainder. The estimated additive seasonal effects are 0.026597548, -0.178516189, 0.132703441, 0.069823277, 0.153609760, 0.010078603, 0.023102982, 0.023102982, 0.007476214, -0.114206180, -0.018514685, -0.110878765, -0.001276026 for respective month starting in January. The seasonal effect for each month is constant over the years. Both the output and the seasonal plot of Figure 3 confirm that additive composition is a suitable description of the data.

After decomposition, remainder is analyzed. The remainder data is isolated, and then acf2() function is used to show its ACF and PACF plot. As can be seen from Figure 4, there is no obvious trend in remainder. Although variance is not constant through time, it is acceptable; the mean of the remainder is 0.0008288457. Figure 5 shows the acf2() plot. Acf2() plot shows that the time series lags are correlated with time since there are many significant spikes outside the significant band. ARMA process is used to model remainder. Figure 5 shows that the ACF has one clearly significant value at lag 1 and a barely significant value at lag 2. The PACF also has a clearly significant value at lag1, so MA (1), MA (2), or AR (1) are all possible to model the remainder. MA (2) is evaluated to be the best fit as it has the lowest BIC -792.7. Then we take the PACF and ACF plots of the MA (2) model residuals. As is shown in Figure 6, despite some spikes, most spikes are within the significance band which shows that MA (2) is appropriate for modeling remainder.

We have identified the non-seasonal MA (2) term. The next step is to find the seasonal component. With linear trend and monthly seasonality, both a non-seasonal first and a seasonal twelfth difference are taken. Figure 9, 10, 11 show the ACF and PACF plots of twelfth difference, first difference, and the twelfth and the first difference. As can be seen from the figure 11, spikes taper in the PACF graph at lag 12, 24, 36, while there is a significant spike at lag 12 in the ACF graph, so it suggests that MA (1) could be potential model for seasonal part. In addition, AR (1) could also be a possible guess for the seasonal part of the data as there is one significant spike at lag 12 in the PACF graph. In summary, ARIMA and ARIMA are selected.

**Diagnostics**

Diagnostics are then applied to compare these two models. Figure 12 is the diagnostics plot for ARIMA, and figure 13 is for ARIMA. As can be seen from figure 12, the standard residuals plot is good; residuals are around zero with no trend, and variance is almost the same. Although there are two significant spikes in the ACF graph of residuals, the ACF graph is good as most spikes are inside the significance band. The third plot shows that residuals follow normal distribution; Some P value for the Ljung-Box statistics are below 0.05. The AIC, AICc, and BIC for ARIMA are -5.050201, -5.040064, and -6.003322 respectively. Its estimated variance is 0.002293, and the non-seasonal MA (2) term is non-significant as P value is greater than 0.05. Figure 13 shows that the standard residuals plot of ARIMA is also good; residuals are around zero, and variance is almost the same. There are several significant spikes in the ACF graph of residuals. Residuals follow normal distribution; most P value for the Ljung-Box statistics are below 0.05. The AIC, AICc, and BIC are -4.858433, -4.848296, and -5.811554 respectively. Its estimated variance is 0.002777. The MA (2) term is also not significant. By comparison, ARIMA evaluates to be better than ARIMA as it has better diagnostic plots and lower AIC, AICc, BIC, and variance. Since MA (2) term is evaluated to be non-significant, ARIMA is reduced to ARIMA, and diagnostics plot is evaluated. As can be seen from Figure 14, its diagnostic plot is similar to the plot of ARIMA; standard residuals plot is good, and ACF of residuals is good. Residuals follow normal distribution, and less P values of Ljung-Box statistics are below 0.05. Its AIC, AICc, and BIC are -5.5057554, -5.04777, and -6.026301. Its estimated variance is 0.002297. ARIMA is better than ARIMA because it has lower AIC, AICc, and BIC, given that they have similar diagnostic plot and estimated variance. ARIMA is selected for forecast.

**Forecast**

R command sarima.for () is used to forecast the sales revenue for the next 12 months. The forecast results are 3.477666, 3.275568, 3.632091, 3.531044, 3.637497, 3.458033, 3.456673, 3.400179, 3.258374, 3.367593, 3.294429, and 3.434967. Figure 15 shows the forecast plot.

**Evaluation**

In this case, I use ARIMA model, but there are other ways to forecast the sales. For example, we can consider trend, seasonal, and remainder separately, form a model to forecast each of them, and then use for sales forecasts. For seasonality part, since we know the seasonal effect for each month for the previous years, we can just use the seasonal effect for each month. For the trend, we can use exponential smoothing method/time series regression model to forecast, and for the remainder we can form ARMA model to forecast. Then, we add each component forecasts together to get the final forecasts.

**Appendix**

>install.packages("astsa")

>library(astsa)

>install.packages("TSA")

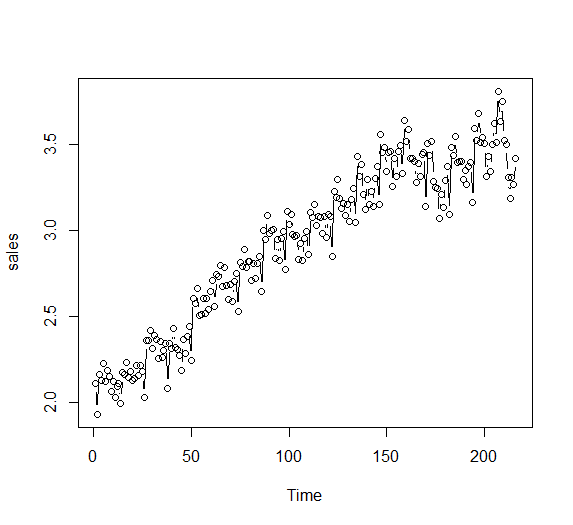
>library(TSA)

>salesdata=read.table("SalesData.csv",sep = ",", header=TRUE)

>salesdata=ts(salesdata) #this makes sure R knows that x is a time series #time series plot of x with points marked as "o"

>sales=salesdata[,3]

>plot(sales, type="b")



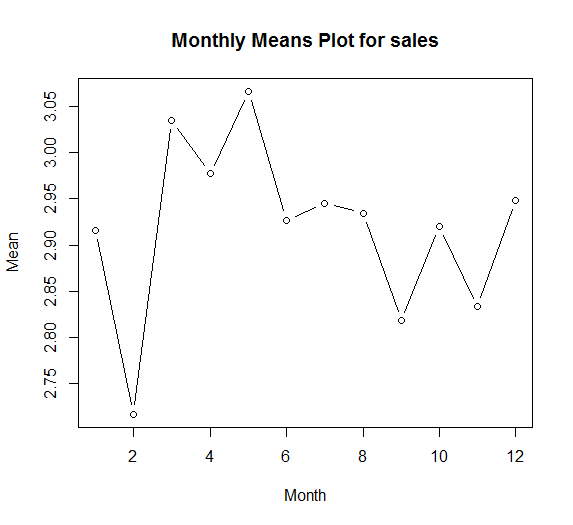
Figure

|  |
| --- |
|  |
|  |
| |  | | --- | |  | |

>salesm = matrix(sales, ncol=12,byrow=TRUE)

>col.means=apply(salesm,2,mean)

>plot(col.means,type="b", main="Monthly Means Plot for sales", xlab="Month", ylab="Mean")

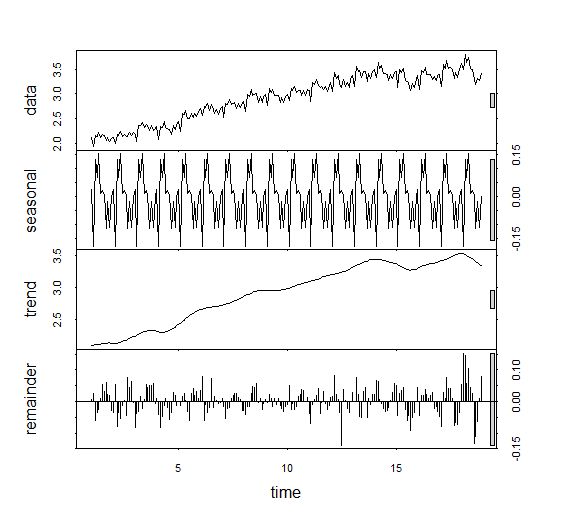


Figure

>sales=ts(sales, freq =12)

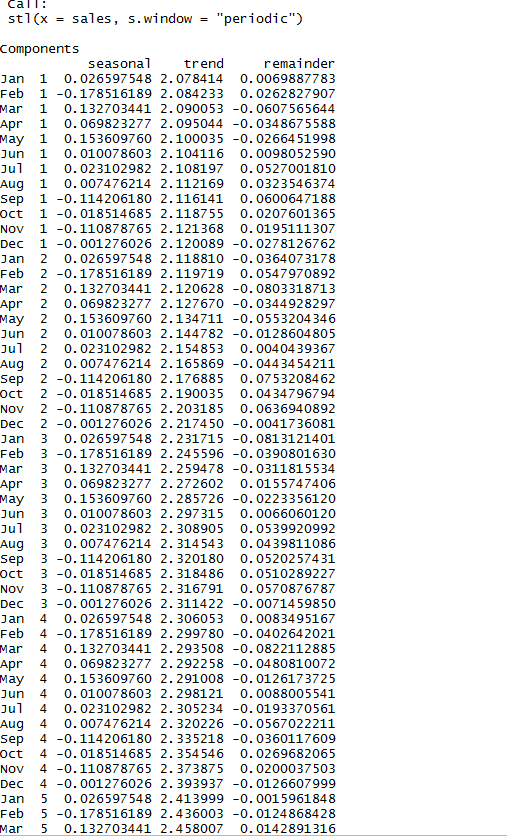
>stl.sales=stl(sales, "periodic")

>plot(stl(sales, "periodic"))



Figure

>stl.sales



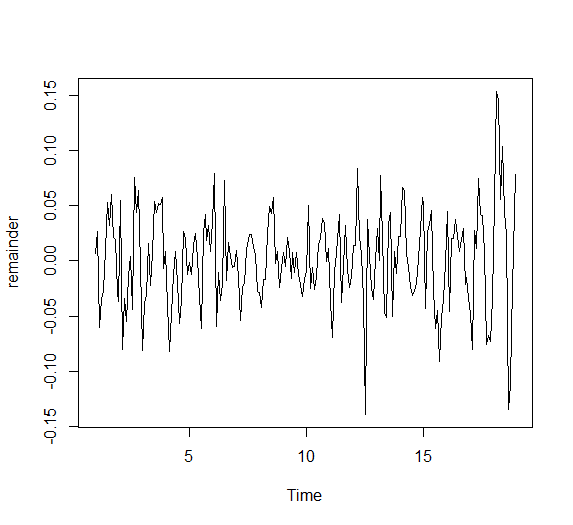
> results=stl(sales,"periodic")

> remainder=results$time.series[,3]

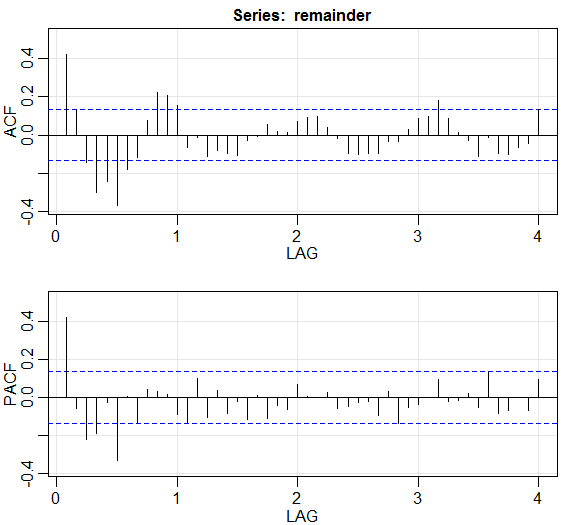
>mean(remainder)

[1] 0.0008288457

>plot(remainder)



Figure

>acf2(remainder)

Figure

> sales.arima1=arima(remainder,order=c(0,0,2), include.mean=F)

> sales.arima1

Call:

arima(x = remainder, order = c(0, 0, 2), include.mean = F)

Coefficients:

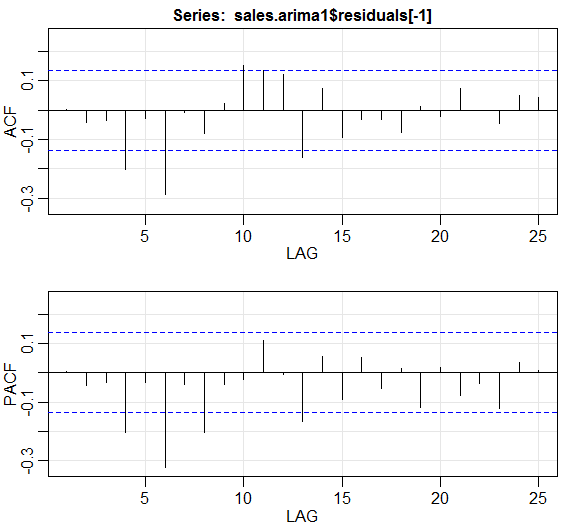
ma1 ma2

0.4381 0.3007

s.e. 0.0635 0.0754

sigma^2 estimated as 0.001462: log likelihood = 398.35, aic = -792.7

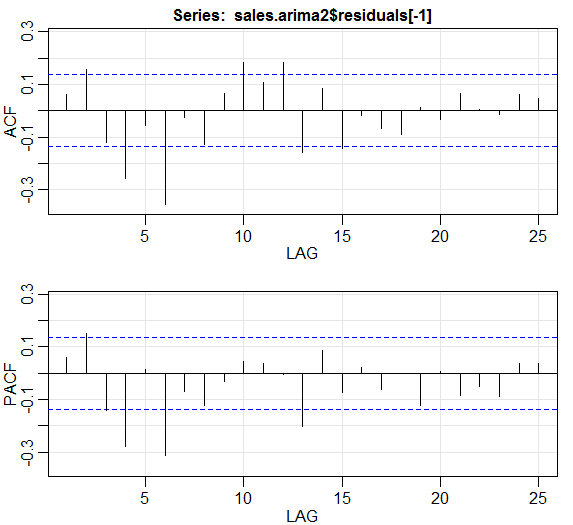
> acf2(sales.arima1$residuals[-1],col="sales ARIMA model residuals", is.df=F)> acf2(sales.arima1$residuals[-1],col="sales ARIMA model residuals", is.df=F)



Figure

|  |
| --- |
| > sales.arima2=arima(remainder,order=c(0,0,1), include.mean=F)  > sales.arima2  Call:  arima(x = remainder, order = c(0, 0, 1), include.mean = F)  Coefficients:  ma1  0.3787  s.e. 0.0580  sigma^2 estimated as 0.001546: log likelihood = 392.4, aic = -782.81 |

> acf2(sales.arima2$residuals[-1],col="sales ARIMA model residuals", is.df=F)



Figure

> sales.arima3=arima(remainder,order=c(1,0,0), include.mean=F)

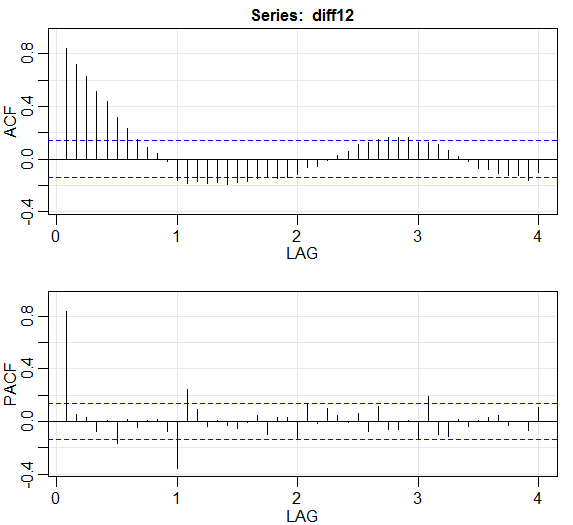
> sales.arima3

> acf2(sales.arima3$residuals[-1],col="sales ARIMA model residuals", is.df=F)

Figure

> diff12=diff(sales,12)

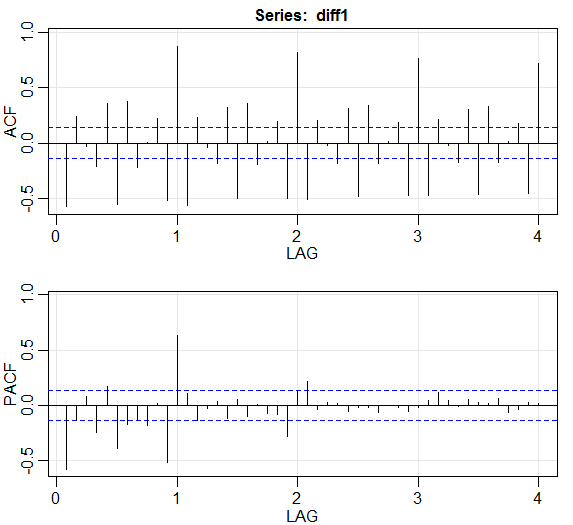
> acf2(diff12)



Figure

>diff1=diff(sales,1)

>acf2(diff1)



Figure

>diff12and1=diff(diff12,1)

>acf2(diff12and1)

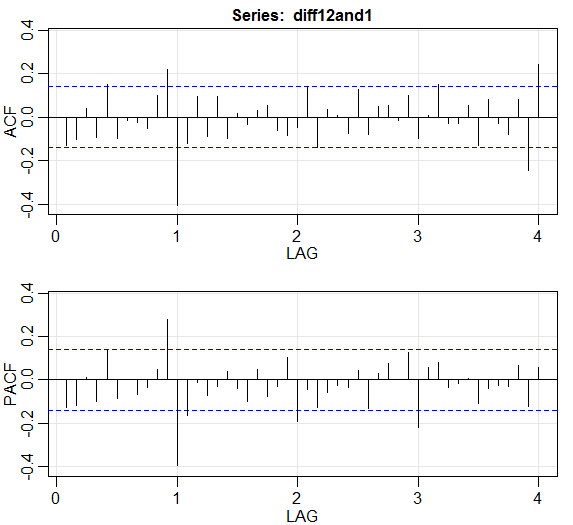


Figure 11

> sarima(sales,0,1,2,0,1,1,12)

initial value -2.822082

iter 2 value -2.973322

iter 3 value -3.013732

iter 4 value -3.026593

iter 5 value -3.028533

iter 6 value -3.031839

iter 7 value -3.032106

iter 8 value -3.032121

iter 9 value -3.032121

iter 9 value -3.032121

iter 9 value -3.032121

final value -3.032121

converged

initial value -3.017868

iter 2 value -3.018035

iter 3 value -3.018163

iter 4 value -3.018164

iter 4 value -3.018164

iter 4 value -3.018164

final value -3.018164

converged

$fit

Call:

stats::arima(x = xdata, order = c(p, d, q), seasonal = list(order = c(P, D,

Q), period = S), include.mean = !no.constant, optim.control = list(trace = trc,

REPORT = 1, reltol = tol))

Coefficients:

ma1 ma2 sma1

-0.1535 -0.0592 -0.7101

s.e. 0.0703 0.0769 0.0543

sigma^2 estimated as 0.002293: log likelihood = 324.64, aic = -641.29

$degrees\_of\_freedom

[1] 200

$ttable

Estimate SE t.value p.value

ma1 -0.1535 0.0703 -2.1844 0.0301

ma2 -0.0592 0.0769 -0.7703 0.4420

sma1 -0.7101 0.0543 -13.0798 0.0000

$AIC

[1] -5.050201

$AICc

[1] -5.040064

$BIC

[1] -6.003322

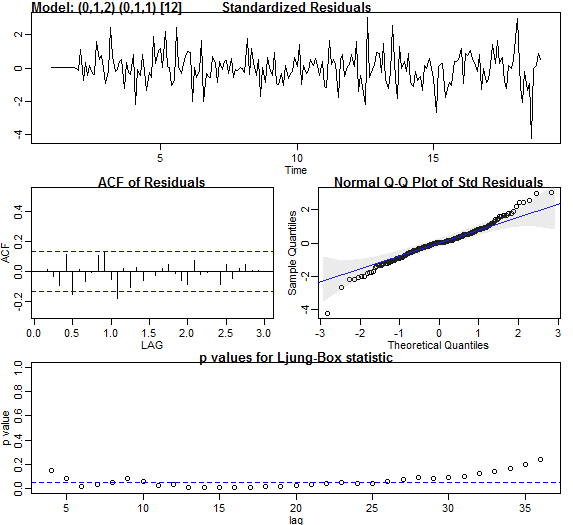


Figure 12

> sarima(sales,0,1,2,1,1,0,12)

initial value -2.808234

iter 2 value -2.926397

iter 3 value -2.928318

iter 4 value -2.928418

iter 5 value -2.928419

iter 5 value -2.928419

iter 5 value -2.928419

final value -2.928419

converged

initial value -2.936121

iter 2 value -2.936294

iter 3 value -2.936312

iter 3 value -2.936312

iter 3 value -2.936312

final value -2.936312

converged

$fit

Call:

stats::arima(x = xdata, order = c(p, d, q), seasonal = list(order = c(P, D,

Q), period = S), include.mean = !no.constant, optim.control = list(trace = trc,

REPORT = 1, reltol = tol))

Coefficients:

ma1 ma2 sar1

-0.1518 -0.0618 -0.4506

s.e. 0.0704 0.0751 0.0664

sigma^2 estimated as 0.002777: log likelihood = 308.03, aic = -608.05

$degrees\_of\_freedom

[1] 200

$ttable

Estimate SE t.value p.value

ma1 -0.1518 0.0704 -2.1563 0.0323

ma2 -0.0618 0.0751 -0.8220 0.4121

sar1 -0.4506 0.0664 -6.7900 0.0000

$AIC

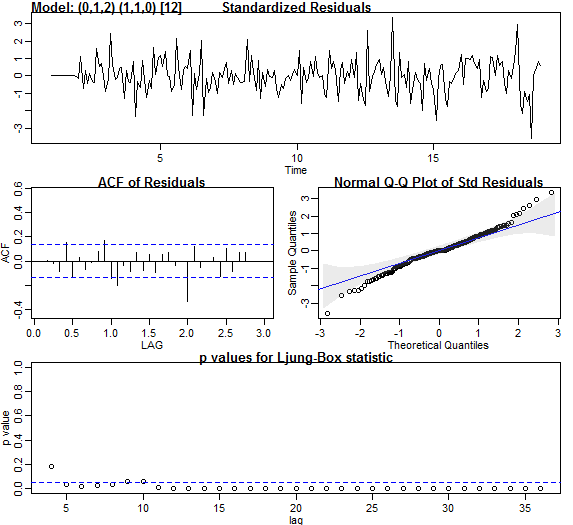
[1] -4.858433

$AICc

[1] -4.848296

$BIC

[1] -5.811554



Figure

> sarima(sales,0,1,1,0,1,1,12)

initial value -2.822082

iter 2 value -2.969883

iter 3 value -3.013888

iter 4 value -3.026731

iter 5 value -3.027685

iter 6 value -3.030852

iter 7 value -3.031024

iter 8 value -3.031112

iter 9 value -3.031113

iter 9 value -3.031113

iter 9 value -3.031113

final value -3.031113

converged

initial value -3.016424

iter 2 value -3.016550

iter 3 value -3.016683

iter 4 value -3.016684

iter 4 value -3.016684

iter 4 value -3.016684

final value -3.016684

converged

$fit

Call:

stats::arima(x = xdata, order = c(p, d, q), seasonal = list(order = c(P, D,

Q), period = S), include.mean = !no.constant, optim.control = list(trace = trc,

REPORT = 1, reltol = tol))

Coefficients:

ma1 sma1

-0.1560 -0.7165

s.e. 0.0739 0.0538

sigma^2 estimated as 0.002297: log likelihood = 324.34, aic = -642.68

$degrees\_of\_freedom

[1] 201

$ttable

Estimate SE t.value p.value

ma1 -0.1560 0.0739 -2.1113 0.036

sma1 -0.7165 0.0538 -13.3260 0.000

$AIC

[1] -5.057554

$AICc

[1] -5.04777

$BIC

[1] -6.026301

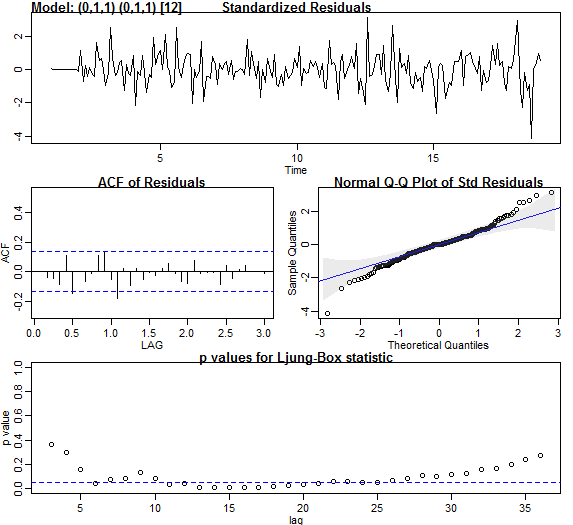


Figure 14

> sarima.for(sales,12,0,1,1,0,1,1,12)

$pred

Jan Feb Mar Apr May Jun Jul Aug

19 3.477666 3.275568 3.632091 3.531044 3.637497 3.458033 3.456673 3.400179

Sep Oct Nov Dec

19 3.258374 3.367593 3.294429 3.434967

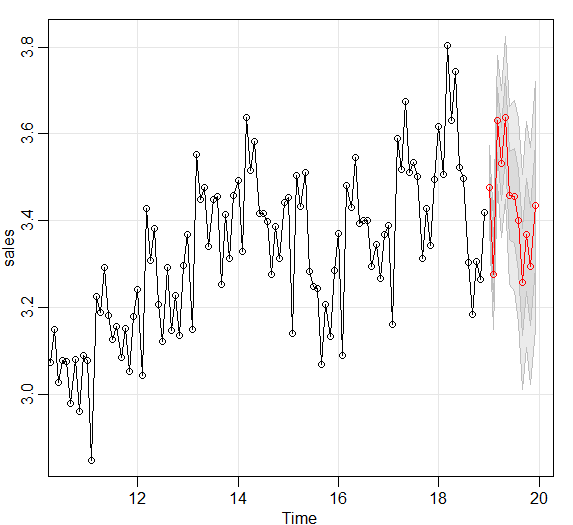
$se

Jan Feb Mar Apr May Jun

19 0.04792906 0.06271868 0.07463291 0.08489117 0.09403695 0.10236888

Jul Aug Sep Oct Nov Dec

19 0.11007191 0.11727004 0.12405121 0.13048042 0.13660739 0.14247111



Figure