Time Series Time Series for Finance and Macroeconomics (1001)

Group 20

Instructor: Dr. Peng JIN

Xinyi LI	1930018036
Ruisi BI	1930024003
Yuetong ZHAO	1930016057

1. Data description

We use BJsales from R database. It contains 150 observations of sales data (for more information consult the R documentation).

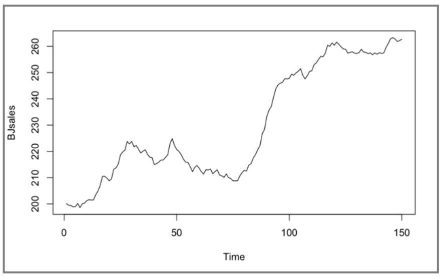
Basic Information of BJsales	
Graph	
Length	150
Mean	229.978
Median	220.65
Variance	461.3769

Table 1 Basic Information of BJsales

2. Stationary Test and Transformation

2.1. Stationarity Test

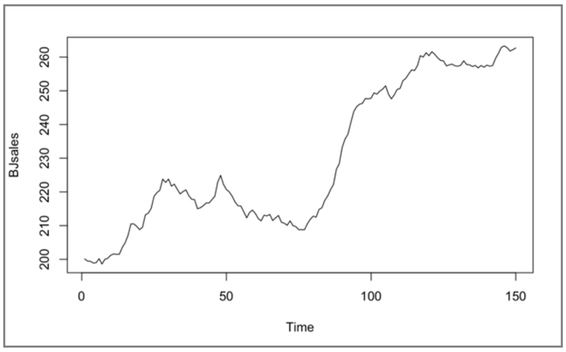


Figure 1 BJsales

The series shows an upward trend, with higher sales over time. Hence, the stationary model does not seems reasonable.

```
Warning in kpss.test(BJsales) : p-value smaller than printed p-value

        KPSS Test for Level Stationarity

data:  BJsales
KPSS Level = 2.6244, Truncation lag parameter = 4, p-value = 0.01
```

Figure 2 KPSS Test for Original Data

In addition we also used the KPSS method for unit root tests. In this method, the NULL and alternative hypotheses are first determined.

H0: Sequence does not have a unit root (series is stationary)

H1: Sequence has a unit root (series is non-stationary).

Calculating by KPSS test, p-value is less than 0.01. Therefore, the NULL hypothesis was rejected at the 99% confidence level. That is, we have 99% confidence that the series is not stationary.

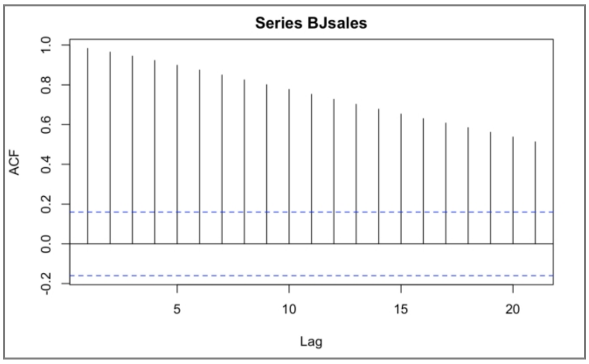


Figure 3 ACF for BJsales

From the ACF plot we can clearly tell that there is a slow, downward trend with a large amount of values lying outside the two approximate standard errors $\pm\frac{2}{\sqrt{150}}$. Therefore, we can conclude that the series is non-stationary.

If we want to further analyze the data, we need to transform it.

2.2 Stationarity Through Differencing

We consider the most common way -- differencing.

We plot the graph of BJsales after the first difference. Now the data looks more stationary.

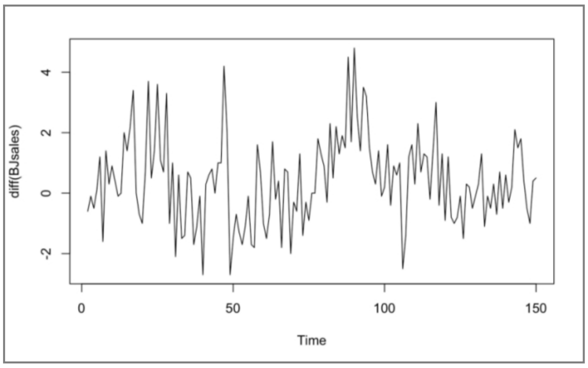


Figure 4 BJsales after the First Differencing

Figure 5 is the result of KPSS unit root test for the series after one difference. P-value is 0.1 which is larger than 0.05. Therefore, under 95% confidence level, the series is stationary this time.

```
Warning in kpss.test(diff(BJsales)) :
  p-value greater than printed p-value

        KPSS Test for Level Stationarity

data:  diff(BJsales)
KPSS Level = 0.13437, Truncation lag parameter = 4, p-value = 0.1
```

Figure 5 KPSS Test after the First Differencing

The ACF and PACF also shows that the model has improved a lot after the first differencing.

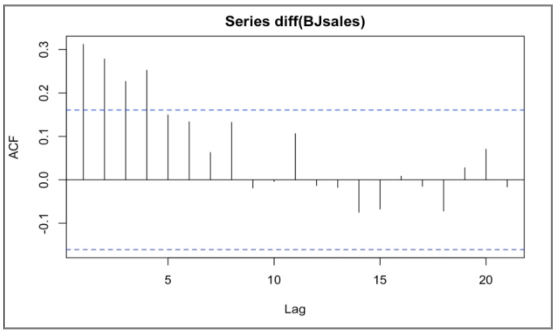


Figure 6 ACF after the First Differencing

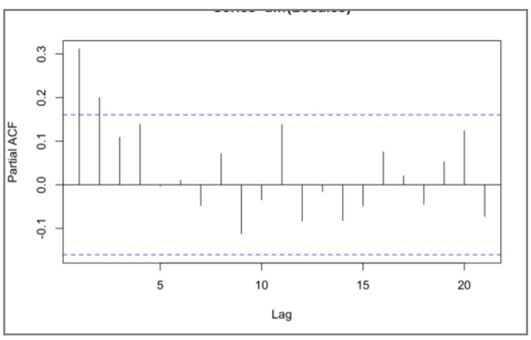


Figure 7 PACF after the First Differencing

3. Model Specification

3.1 Model Specification for the Series after the First Difference

According to ACF plot (figure 6), ACF is significantly 0 when the lag greater than 4. Thus, it is similar to the MA(4) model. However, the ACF drops very slowly.

Then we draw PACF graph (figure 7), when the lag greater than 2, PACF is significantly 0. Thus, it is similar to AR(2) model.

The ACF and PACF plot can only imply pure MA or AR models. In this case, they do not depict the same model. So we need to introduce EACF diagram.

```
AR/MA
  0 1 2 3 4 5 6 7 8 9 10 11 12 13
0 x x x x o o o o o o o  o  o  o
1 x o o o o o o o o o o  o  o  o
2 x o o o o o o o o o o  o  o  o
3 x x o o o o o o o o o  o  o  o
4 o o x x o o o o o o o  o  o  o
5 x o o o x o o o o o o  o  o  o
6 x x o o o o o o o o o  o  o  o
7 x x o o o o o o o o x  o  o  o
```

Figure 8 EACF

The triangular region of zeros shown in the sample EACF clearly indicates that an ARMA(p,q) model with p = 1 and with q = 1 is reasonable.

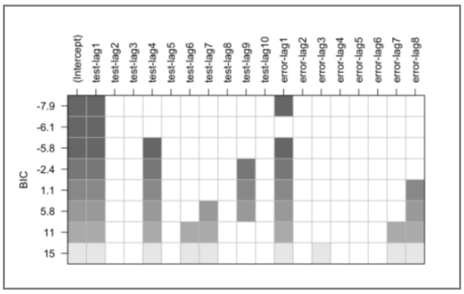


Figure 9 Best Subset

Moreover, by analyzing the selection of the optimal subset based on BIC, the model of ARMA(1,1) would be the best.

The best model contains only lag 1 of the observed time series and lag 1 of the error process. The next best model contains only lag 1 of the time series. The third best model: Contains lags 1 and 4 of the time series and lag 1 of the errors. For our data, the lag 1 and lag 1 of the error process appear most frequently in the subset model, which indicates that they may be the important variables. As we already know, they do matter.

3.2 Over Differencing

To check whether we should do differencing agian, we calculated the variance, log likelihood, AIC and BIC of model ARIMA(1,1,1) and model ARIMA(1,2,1). The result is shown in table 2. We can see that ARIMA(1,1,1) has less variance, AIC, BIC and larger log likelihood.

	1st Differencing	2nd Differencing
Variance	2.085 *	2.882
Log likelihood	-254.37 *	-256.49
AIC	512.74 *	516.97
BIC	523.748 *	527.965

Table 2 Comparition of 1st and 2nd Differencing

All those values can indicate one difference is greater than two. Thus the results of both methods above indicate that second order difference is over differencing.

4. Choose parameter

4.1. Calculation of Parameter

When selecting parameters, use the CSS-ML method, which first uses the least squares method to select the starting point, and then use the maximum likelihood method to calculate it, which is a better method of ARIMA. Additionally, non-stationary ARIMA doesn't contain constant term. (μ =0). To assess the significance, calculate the var($\overline{Y}$) = 1.1875, and $\overline{Y}$ = 0.42. Because μ=0 $\in$ ($\overline{Y}$ $\pm$ $1.96\sqrt{\text{var}(\overline{Y})}$). So we can choose constant term equal to 0 at the 95% confidence level.

```
```{r}
arima(BJsales,order=c(1,1,1),method='CSS')
arima(BJsales,order=c(1,1,1),method='ML')
arima(BJsales,order=c(1,1,1),method='CSS-ML')
```
```

Figure 10 R Code for Parameter Chosen

```
Call:
arima(x = BJsales, order = c(1, 1, 1), method = "CSS")

Coefficients:
         ar1      ma1
      0.8809  -0.6374
s.e.  0.0652   0.1020

sigma^2 estimated as 1.788:  part log likelihood = -254.71

Call:
arima(x = BJsales, order = c(1, 1, 1), method = "ML")

Coefficients:
         ar1      ma1
      0.8799  -0.6415
s.e.  0.0644   0.1035

sigma^2 estimated as 1.775:  log likelihood = -254.37,  aic = 512.74

Call:
arima(x = BJsales, order = c(1, 1, 1), method = "CSS-ML")

Coefficients:
         ar1      ma1
      0.8800  -0.6415
s.e.  0.0644   0.1035

sigma^2 estimated as 1.775:  log likelihood = -254.37,  aic = 512.74
```

Figure 11 Result of Different Methods

Detection of parameters:

$\hat{\varphi}$= 0.8800 $\hat{\theta}$ = -0.6451 and $\hat{\sigma}_e^2$ = 1.775

4.2 Test the Significance of Parameters

Whether $\hat{\varphi}$ or $\hat{\theta}$ can be ignored?

1. According to the theory: if $|\Phi|$ $\leq$ $2\sqrt{\text{var}(\Phi)}$, then Φ = 0.

Because 0.88 > 0.1288 and |-0.6451| > 0.207. So both $\hat{\varphi}$ and $\hat{\theta}$ are significantly not equal to 0.

2. Calculate $\hat{\varphi}$/ s and $\hat{\theta}$ /s , the result is in figure 14.

```
t1=0.8800/0.0644
pt(t1,df=148,lower.tail=F)
t2=-0.6415/0.1035
pt(t2,df=148,lower.tail=T)
```

```
[1] 2.584248e-28
[1] 2.698431e-09
```

Figure 12 t-test of $\widehat{\varphi}$ / s and $\hat{\theta}$ /s

The t-test also reveals that the p values are extremely small, so the null hypothesis can be rejected, that is, the coefficients are significantly non-zero. (H0: coefficient = 0, H1: coefficient $\neq$ 0)

4.3 The Estimated Model: ARIMA(1,1,1)

$Y_t - Y_{t-1} = 0.88(Y_{t-1} - Y_{t-2}) + e_t + 0.6451e_{t-1}$

$Y_t = 1.88Y_{t-1} - 0.88Y_{t-2} + e_t + 0.6451e_{t-1}$

The model is non-stationary and invertible.

5. Model Diagnostics

5.1 Residual Analysis

Residual = true value - predict value = $e_t + 0.6451e_{t-1}$

5.2 Plot of the residuals

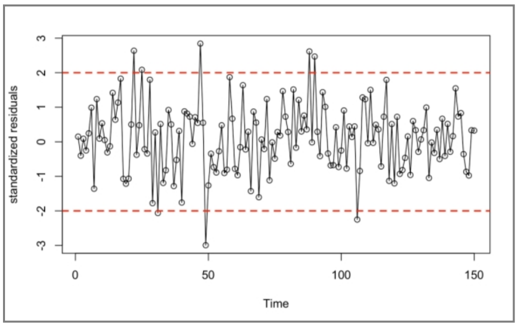


Figure 13 Standardized Residuals for ARIMA(1,1,1) Model

We can see from figure 13 that it suggest an approximate rectangular scatter around a zero horizontal level with no trends.

5.3 Normality of Residuals

1. We assume that the standardized residuals follow the standard normal distribution. Therefore, there should be 95% of the standardized residuals, lies in the interval [-2,2]. In our model, 142 of 150 data (approximately 95%) are inside the boundary ± 2. According to figure 13, we have 7 exceptions.

2. Shapiro-Wilk Test

```
        Shapiro-Wilk normality test

data:  residuals(m2)
W = 0.99258, p-value = 0.6301
```

Figure 14 Result of Shapiro-Wilk Test

The closer the value of W is to 1, the better the data fits the normal distribution. The p-value is large enough so that we do not reject the null hypothesis that the residual is normally distributed. The result shows that the w-value is 0.99 and p-value is 0,63. (H0: the sample follows normal distribution H1: the sample does not follow normal distributions)

3. Quantile-quantile plot

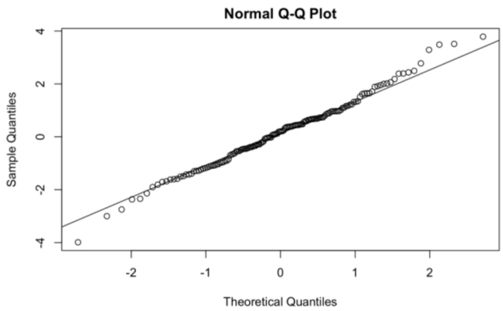


Figure 15 QQ Plot for Residuals

The points are near a straight line, so the residuals are from normal distribution.

Above all, all three methods indicate that residuals follow a normal distribution.

5.4 Autocorrelation of the Residuals

1. ACF

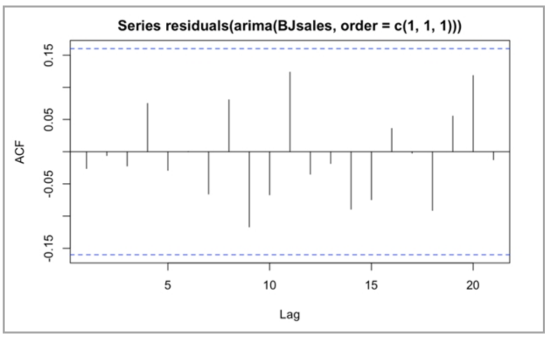


Figure 16 ACF of the Residuals

The ACF does not show statistically significant evidence of nonzero autocorrelation in the residuals.

2. Ljung-Box Test

After calculating Ljung-Box test statistic for different values of k from 1 to 20, we plot the p-value of according to k.

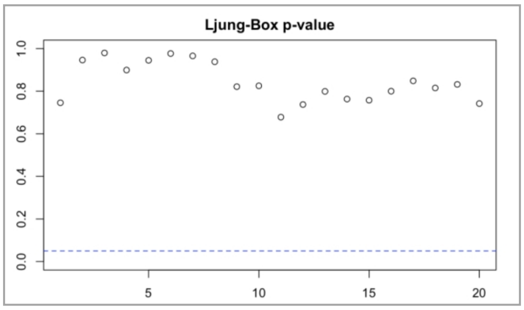


Figure 17 Ljung-Box P-value

As all the p-values are much larger than 0.05, we have no evidence to reject the null hypothesis that the error terms are uncorrelated. The estimated ARIMA(1,1,1) model seems to be capturing the dependence structure of the time series very well.

Therefore, AR or MA processes are no longer included in the residuals, the residuals satisfy non-autocorrelation.

All in all, the residual is white noise. All useful information in the time series has been extracted and all that is left is random perturbation, which cannot be predicted or used so the modelling can be terminated

5.5 Overfitting

The model we choose is ARIMA(1,1,1). After specifying, it is necessary to fit a slightly more general mode that contains the original model as a special case. Furthermore, ARMA model with the additional parameters equal to zero can be seen as a more general model as well. To check whether the model is overfitting, we have tried different models ARIMA(2, 1, 1), ARIMA(1,1,2), ARIMA(1,2,1), IMA(1,1) and ARI(1,1).

```
Call:
arima(x = BJsales, order = c(1, 1, 1))

Coefficients:
         ar1      ma1
      0.8800  -0.6415
s.e.  0.0644   0.1035

sigma^2 estimated as 1.775:  log likelihood = -254.37,  aic = 514.74

Call:
arima(x = BJsales, order = c(1, 1, 0))

Coefficients:
         ar1
      0.3647
s.e.  0.0759

sigma^2 estimated as 1.945:  log likelihood = -261.06,  aic = 526.13

Call:
arima(x = BJsales, order = c(0, 1, 1))

Coefficients:
         ma1
      0.2562
s.e.  0.0653

sigma^2 estimated as 2.042:  log likelihood = -264.63,  aic = 533.27
```

Figure 18 Comparison with General Models

```
Call:
arima(x = BJsales, order = c(2, 1, 1))

Coefficients:
         ar1     ar2     ma1
      0.8305  0.0360  -0.607
s.e.  0.1774  0.1178   0.160

sigma^2 estimated as 1.774:  log likelihood = -254.32,  aic = 516.64

Call:
arima(x = BJsales, order = c(1, 1, 2))

Coefficients:
         ar1      ma1     ma2
      0.8705  -0.6483  0.0286
s.e.  0.0743   0.1101  0.0904

sigma^2 estimated as 1.774:  log likelihood = -254.32,  aic = 516.64

Call:
arima(x = BJsales, order = c(1, 2, 1))

Coefficients:
         ar1      ma1
      0.0528  -0.7801
s.e.  0.1313   0.1004

sigma^2 estimated as 1.863:  log likelihood = -256.49,  aic = 518.97
```

Figure 19 Comparison with General Models

Compared with other models, ARIMA(1,1,1) has the approximately smallest variance and the largest log likelihood along with the definitely smallest AIC at the same time. Moreover, $\widehat{\varphi}$ and $\hat{\theta}$ are statistically different from 0 but either of two estimated parameters from other models is quite close to 0, which also would support the choice of model ARIMA (1,1,1). It is obvious that there is no parameter redundancy.

6. Prediction

6.1 Forecasting

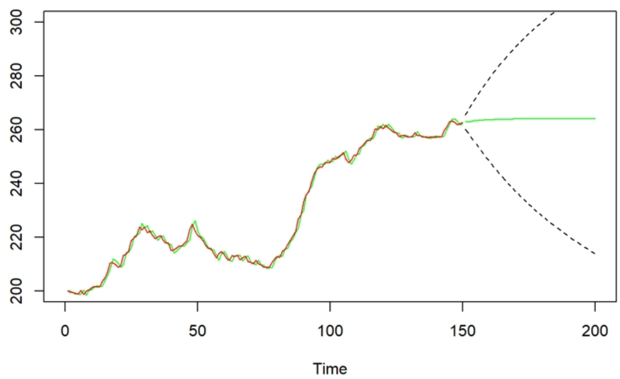


Figure 20 Forecasting Graph

```
#ARIMA(1,1,1)
model=Arima(BJsales,order=c(1,1,1))
#95%confidence level
forecast1<-forecast(model,h=50,level=c(95))
plot(forecast1$fitted,ylim=c(200,300),xlim=c(0,200),col="green")
lines(BJsales,col="red")
lines(forecast1$mean,col="green")#predicted value
lines(BJsales,col="red")#true value corresponding to the predicted value
lines(forecast1$upper,lty=2)#upper bound of forecasting
lines(forecast1$lower,lty=2)#lower bound of forecasting
```

Figure 21 The Code of R

The graph shows the visuals of BJsales without forecasting in red and with forecasting in green predicted by the ARIMA model of BJsales dataset. The line graph also displays that the series with forecasts out to lead time 50 with the upper and lower 95% prediction limits for those forecasts. Furthermore, a horizontal line for the process mean which is approximately 264 is demonstrated. Moreover, as the lead time increases, the forecasts are close to the mean exponentially and the prediction limits increase in width.

6.2 Prediction evaluation indicators

ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
0.2655281	0.9417852	0.7011216	0.1264694	0.2948326	0.007915889	-0.3628648

Figure 22 Prediction Evaluation Indicators

The RMSE is about 0.9. It means the degree of variability in the data is small and the predictive model describes the experimental data with accuracy. The other prediction evaluation indicators are all smaller than 1. Therefore, our model is appropriate.

References

Cryer, J. D. (n.d.). Time Series Analysis with Applications in R (2nd ed.). China Machine Predd.

Wang, Y. (2005). Apply Time Series Analysis (1st ed.).

Yang, F. Y. (2020). Information Theory Method Research of Data Mining. 06. https://doi.org/10.27307/d.cnki.gsjtu.2019.002428

Guo, Q. S. (2021). Time Series Forecasting Model Based on Point-in-Time Processes. Chinese Core Journal of PKU.