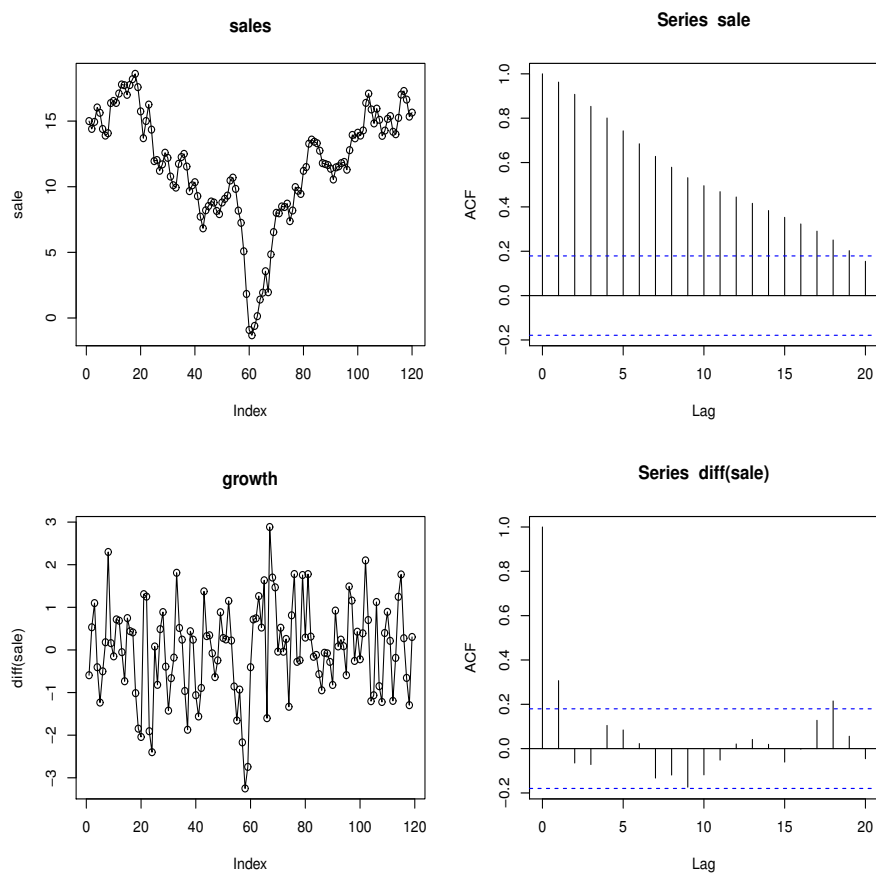


Eco 672: Forecasting Weekly Paper Towel Sale with ARIMA Model

This short note shows how to forecast the weekly sale of paper towel using ARIMA model.

1 Big Picture

Our analysis starts with plotting the sales data. There are three signals for nonstationarity: (1) the series is smooth (not mean-reverting); (2) there are local trends (long swing); (3) the ACF is slow-decaying. After we take difference, the sale growth indicates stationarity—the growth series is mean-reverting, choppy, and has fast-decaying ACF.



Below are the R codes

```
setwd("C:/Users/chnlj/Dropbox/400")
data = read.table("672_papertowel_data.txt", header=F)
sale = data[,1]
```

```
par(mfrow=c(2,2))
plot(sale, type="o", main="sales")
acf(sale)
plot(diff(sale), type="o", main="growth")
acf(diff(sale))
```

2 Unit Root Test

Next we conduct the Augmented Dickey-Fuller unit root test and we find that unit root cannot be rejected for sale (ADF Test is -1.7698, p-value is 0.3278); while unit root can be rejected for sale growth (ADF Test is -7.588). Thus we conclude that sale is difference-stationary, and taking difference once is able to achieve stationarity (i.e., sale is integrated of order one).

```
> library(tseries)
> adf.test(sale, alternative=c("e"), k=1)
Augmented Dickey-Fuller Test
data:  sale
Dickey-Fuller = -1.7698, Lag order = 1, p-value = 0.3278
alternative hypothesis: explosive

> adf.test(diff(sale), alternative=c("s"), k=1)
Augmented Dickey-Fuller Test
data:  diff(sale)
Dickey-Fuller = -7.588, Lag order = 1, p-value = 0.01
alternative hypothesis: stationary
```

3 ARIMA(1,1,0) Model

Given that the ACF for sale growth implies weak correlation, we try ARIMA(1,1,0) model for sale, which is the same as AR(1) model for sale growth

```
> arima(sale,order=c(1,1,0))
```

Coefficients:

ar1

0.3051

s.e. 0.0869

sigma^2 estimated as 1.095: log likelihood = -174.29, aic = 352.57

The AR(1) coefficient is statistically significant (t value is 3.511), and the estimated model is

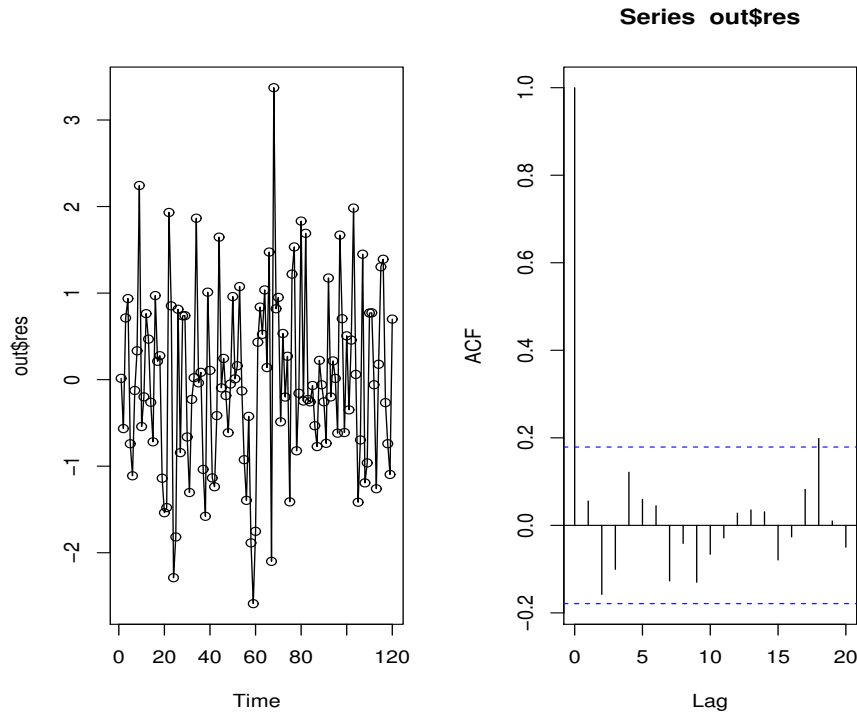
$$(sale_t - sale_{t-1}) = 0.3051(sale_{t-1} - sale_{t-2}) + u_t \quad (1)$$

or equivalently

$$sale_t = 1.3051sale_{t-1} - 0.3051sale_{t-2} + u_t \quad (2)$$

4 Diagnostic Checking

ARIMA(1,1,0) model seems to work well since we spot no pattern in its residual, and there is no correlation left in the residual



Formally, the adequacy of the ARIMA(1,1,0) model is indicated by the fact that the hypothesis of no serial correlation cannot be rejected for its residual (i.e., the residual behaves like a white noise, so there is no need to improve the ARIMA(1,1,0) model)

```
> out=arima(sale,order=c(1,1,0))  
> Box.test(out$res, lag = 1, type = "Ljung")
```

Box-Ljung test

data: out\$res

X-squared = 0.38138, df = 1, p-value = 0.5369

5 Out of sample forecast

The sample ends at the 120th week. The out-of-sample forecast is based on (2). For instance, the one-step-ahead forecast for the 121th week is

$$sale_{121} = 1.3051sale_{120} - 0.3051sale_{119} = 15.73814$$

Other forecasts can be computed in a recursive manner. R has **forecast** package that can generate the forecast.

```
> T = length(sale)
> yhat1 = 1.3051*sale[T]-0.3051*sale[T-1]
> yhat2 = 1.3051*yhat1-0.3051*sale[T]
> yhat3 = 1.3051*yhat2-0.3051*yhat1
> cbind(yhat1,yhat2,yhat3)
      yhat1    yhat2    yhat3
[1,] 15.73814 15.76647 15.77511

> library(forecast)
> forecast(Arima(sale,order=c(1,1,0),include.drift=F),h=3)
      Point Forecast    Lo 80    Hi 80    Lo 95    Hi 95
121      15.73813 14.39159 17.08467 13.67878 17.79749
122      15.76645 13.55255 17.98035 12.38058 19.15232
123      15.77509 12.86895 18.68124 11.33053 20.21966
```

6 Additional Thoughts

1. We also try ARIMA(2,1,0) model, and the result is

```
> arima(sale,order=c(2,1,0))
Coefficients:
          ar1      ar2
      0.3612  -0.1781
s.e.  0.0902   0.0904
sigma^2 estimated as 1.06:  log likelihood = -172.38,  aic = 350.75
```

Despite of the smaller AIC, we still prefer ARIMA(1,1,0) over ARIMA(2,1,0) for three reasons: (1) the AR(2) coefficient -0.1781 is barely significant; (2) we already showed ARIMA(1,1,0) model is adequate; (3) the principle of parsimony states that we should use the simpler ARIMA(1,1,0) model if the goal is forecasting

2. The weekly sale of paper towel shows no seasonality. Otherwise, dummy variables or seasonal AR model can be used
3. The sale data shows no global trend. Otherwise we should try command

```
arima(diff(sale),order=c(1,0,0))
```

which allows for a nonzero drift term.

4. If there is structural change or break, using only recent data may be more appropriate than using the whole sample. That may be the case for the sale because of the turning point at the 60th week. Using only the last 60 weeks data (subsample) we get

```
> arima(diff(sale)[61:120],order=c(1,0,0))
Coefficients:
          ar1  intercept
      0.0686    0.2881
s.e.  0.1290    0.1368
sigma^2 estimated as 0.9597:  log likelihood = -82.51,  aic = 171.01
```

which implies

$$(\Delta sale_t - 0.2881) = 0.0686(\Delta sale_{t-1} - 0.2881) + u_t \quad (3)$$

Note that the reported intercept is actually the drift term (mean for the growth). Alternatively, given that AR(1) coefficient 0.0686 is insignificant, we can drop it and the re-estimated model is a random walk with drift

$$sale_t = 0.2875 + sale_{t-1} + u_t \quad (4)$$

```
> Arima(sale[61:120],order=c(0,1,0),include.drift=T)
Coefficients:
      drift
      0.2875
s.e.    0.1278
sigma^2 estimated as 0.981; log likelihood=-82.65; aic=169.29
```

The out-of-sample forecasts based on (3) are

```
> forecast(Arima(sale[61:120],order=c(1,1,0),include.drift=T),h=3)
      Point Forecast    Lo 80    Hi 80    Lo 95    Hi 95
61      15.93447 14.65718 17.21176 13.98102 17.88792
62      16.22260 14.35322 18.09199 13.36362 19.08159
63      16.51066 14.19243 18.82890 12.96523 20.05610
```

The out-of-sample forecasts based on (4) are

```
> forecast(Arima(sale[61:120],order=c(0,1,0),include.drift=T),h=3)
      Point Forecast    Lo 80    Hi 80    Lo 95    Hi 95
61      15.93280 14.66348 17.20212 13.99155 17.87406
62      16.22030 14.42522 18.01539 13.47496 18.96565
63      16.50781 14.30928 18.70633 13.14546 19.87015
```

5. We may also try ARIMA(0,1,1) model

```
> arima(sale,order=c(0,1,1))
Coefficients:
```



```

          ma1
          0.3518
s.e.    0.0800
sigma^2 estimated as 1.071:  log likelihood = -172.99,  aic = 349.98

```

The estimated model is

$$\Delta sale_t = u_t + 0.3518u_{t-1} \quad (5)$$

The forecasts are

$$sale_{121} = sale_{120} + 0.3518\hat{u}_{120}$$

$$sale_{121+j} = sale_{121} \quad j = 1, 2, \dots$$

```
> forecast(Arima(sale,order=c(0,1,1),include.drift=F),h=3)
```

Point Forecast	Lo 80	Hi 80	Lo 95	Hi 95
121	15.88729	14.55557	17.21902	13.85059
122	15.88729	13.64798	18.12661	12.46256
123	15.88729	13.01405	18.76054	11.49304

6. We may average the different out-of-sample forecasts. For instance, $(15.73813 + 15.93447 + 15.88729)/3 = 15.8533$ can serve as the final forecast for the 121th week sale
7. We haven't used any external information. If we predict, say, the COVID pandemic is worsening, then we may use information such as positive cases of covid to help improve the forecast of paper towel.
8. Finally, remember that ARIMA model is based on the idea that *history will repeat itself*, so **cannot predict black swan**. All the forecasts for the 121th week sale are wrong if all of sudden the company is destroyed by an explosion or the company goes to bankrupt.