Stat 5170: Assignment 12

due May 10, 5pm

1 R Tutorial

In this tutorial, we learn how to carry out a lagged regression in R. We will use the "bluebird.dat" dataset (same dataset as worked example in Unit 24). The data contains data on log-transformed sales (first column) and price (second column) of large packages of potato chips from Bluebird Foods Ltd. over a period of 104 weeks from September 20 1998 to September 10 2000. We will treat price as the input series. Read the data using the following

```
data<-read.table("bluebird.dat")
sales<-data$V1
price<-data$V2
```

To obtain stationarity for both series, we apply a first difference to both `sales` and `price`,

```
dy<-diff(sales)
dx<-diff(price)
```

1. CCF for prewhitened data. The `prewhiten()` function from the `TSA` package produces a CCF plot after prewhitening the data.

   ```
   library(TSA)
   xy<-ts.intersect(as.ts(dx),as.ts(dy))
   prewhiten(as.vector(xy[,1]),as.vector(xy[,2])) ##input series first
   ```

   What does this CCF plot tell us what lagged regression model we should consider? Sometimes, it may be difficult to visually which lag is significant. Type `prewhiten(as.vector(xy[,1]),as.vector(xy[,2]))$ccf` to extract the sample CCF with the lags.

2. Checking residuals. So we regress `dy` on `dx`, and then check the ACF and PACF plots of the residuals. What ARMA structure do we appear to have for the residuals of this regression?

3. Fitting lagged regression with ARMA errors. We use the `arima()` function to fit a lagged regression with ARMA errors. For example

   ```
   result.arma<-arima(dy,order=c(3,0,4),xreg=data.frame(dx))
   result.arma
   ```

   Notice that $\hat{\phi}_3$ and $\hat{\theta}_4$ are insignificant, so we can re-fit the model with an ARMA(2,3) structure for the errors. For this model, notice that $\hat{\phi}_1$ and $\hat{\beta}_0$ are insignificant. We can use `arima()` to fit the model and force these coefficients to be 0.

   ```
   result.arma3<-arima(dy,order=c(2,0,3),xreg=data.frame(dx),
   fixed=c(0,NA,NA,NA,NA,0,NA))
   result.arma3
   ```

   Before finalizing a model, we need to check the residuals. Make sure the ACF and PACF plots indicate the residuals are white. Once we are satisfied, how do we write this model?

2 Assignment

1. Use the “sales.dat” and “lead.dat” datasets for this part. Each file contains 150 consecutive monthly values. The sales variable is monthly sales of a product, and the input series is a company indicator (predictor) of future sales for measured in the same 150 months. For this problem, well treat sales as y_t and lead as x_t. After reading in the data, apply a differencing operation to x_t, and a differencing operation to y_t twice to obtain stationarity. For example,

   ```
   x<-ts(scan("lead.dat"))
   y<-ts(scan("sales.dat"))
   dx<-diff(x)
   dy<-diff(diff(y))
   ```

   (a) Plot the CCF between the two differenced variables without prewhitening, e.g. type `ccf(dx,dy)`. For which lags h are the significant correlations between the variables?

   (b) Carry out a regression with the twice differenced sales as the response and several lagged versions of the differenced lead as the predictor variables; choose the lags you observed as the strongest in Part 1a.

   R hint: for example, if you observe that lags $-1, -2, -3$ are significant then the following code can be used:

```
dx1<-lag(dx,-1)
dx2<-lag(dx,-2)
dx3<-lag(dx,-3)
a<-cbind(dy,dx1,dx2,dx3)
result<-lm(dy~dx1+dx2+dx3,data=a,na.action=na.omit)
##data and na.action specified because we apply
##different differencing operations to input
##and output, resulting in time series of
##different lengths.
summary(result)
```

i. Are the coefficient estimates significant? Comment on whether you find the results surprising (or not).

ii. Are the ACF and PACF plots of the residuals indicative of white noise?

(c) Prewhiten the data and produce the CCF plot on the prewhitened data. What lags appear to have strong correlations? Why should we prewhiten the data when we want to examine a CCF plot?

(d) Carry out a regression with the differenced sales as the response and several lagged versions of the twice differenced lead as the predictor variables; choose the lags you observed as the strongest in Part 1c. Are the ACF and PACF plots of the residuals from this model indicative of white noise?

R hint: for example, if you observe that lags $-2, -3, -4$ are significant based on the prewhitened data then the following code can be used:

```
dx4<-lag(dx,-4)
b<-cbind(dy,dx2,dx3,dx4)
result2<-lm(dy~dx2+dx3+dx4,data=b,na.action=na.omit)
##i suggest creating a new dataframe with the new
##lag. and then fit the regression model using
##this new dataframe.
```

(e) Using the `arima()` function, fit the model from part 1d, but with the appropriate ARMA structure for the residuals. Write out the estimated regression equation together with the estimated model for the noise terms. Be sure to include ACF and PACF plots of the residuals from your final model, making sure they are white.

R hint: if you type `b` to examine the dataframe, you will notice a lot of entries that are `"NA"`. This is due to the different values of lags we are considering. So, for example, if we want to regress `dy` on `dx2`, `dx3`, and `dx4`, we will have to consider indexes 4 to 148 for all of these columns. (R will complain if you have NAs in any column). So I suggest the following code

```
dy<-b[4:148,1]
dx2<-b[4:148,2]
dx3<-b[4:148,3]
```

```
dx4<-b[4:148,4]

result.arma<-arima(dy,order=c(1,0,1),xreg=data.frame(dx2,dx3,dx4))
result.arma
```