Stat 463, Lab 3

Sept 14, 2007

1 In Class

In this section, we will go over how to use R to produce the output and graphs presented in class for the Yule data set. First, create a folder "lab2" and download the files "yule1.dat" and "yule2.dat" from the course website.

1. We need to read the data into R.

   ```
   mortality=scan("yule1.dat", skip=12)
   marriages=scan("yule2.dat", skip=12)
   time=1866:1911
   ```

2. Now, we should plot the data to see what is going on.

   ```
   par(mfrow=c(2,1))
   plot.ts(marriages)
   plot.ts(mortality)
   dev.off()
   par(mfrow=c(1,1)) #return plot to regular
   plot(marriages,mortality)
   ```

3. We will now perform a regression with time as the indpendent variable and `marriages` as the dependent variable. The following will perform the regression and output some results. The fit is saved in the variable `timefit` and then output different aspects of the fit with commands like `anova()`.

   ```
   timefit=lm(marriages~time)
   anova(timefit)
   summary(timefit)
   AIC(timefit)
   ```

4. We can now plot some of the diagnostic plots to help determine the quality of the fit and whether assumptions of the model are satisfied.

   ```
   timeres=residuals(timefit)
   plot(timeres)
   acf(timeres)
   qqnorm(timeres)
   ```

5. As we did in class, we can see if there is also a quadratic trend in `marriages`. We will first need to declare a new variable consisting of `time` squared. Then, we will perform regression just as we did before.

   ```
   timesq=timeˆ2
   timefitsq=lm(marriages˜time+timesq)
   anova(timefitsq)
   summary(timefitsq)
   AIC(timefitsq)
   ```

   Go ahead and plot the residual diagnostics as you did before.

6. Now, we would like to regress `marriages` on `mortality` and plot diagnostics.

   ```
   comparefit=lm(marriages˜mortality)
   anova(comparefit)
   summary(comparefit)
   AIC(comparefit)
   ```

2 Homework

For the homework, be sure to give full explanations where required and to turn in any relevant plots.

1. The file “berkeley.dat” contains average yearly temperatures for the cities of Berkeley and Santa Barbara. Import the data into R using the following commands

   ```
   berk=scan("berkeley.dat", what=list(double(0),double(0),double(0)))
   time=berk[[1]]
   berkeley=berk[[2]]
   stbarb=berk[[3]]
   ```

(a) Plot the variables `berkeley` and `stbarb` versus `time`. Also, plot `berkeley` versus `stbarb`.

(b) Perform a regression of `berkeley` on `time`. What do you think about this fit? Be sure to make diagnostic plots (including ACF) of the residuals. If there are any violations of the assumptions for a linear regression model, make sure to comment on them.

(c) Perform a regression of `berkeley` on `stbarb`. Comment on the fit and the residuals.

(d) Make a plot of the variable `berkeley` and an ACF plot of the data. Does the time series appear to be stationary? Explain. Interpret the ACF plot in this situation.

(e) Difference the data. Plot this differenced data, and make an ACF plot. What is your opinion of whether the series is stationary after differencing?

(f) Now, we have detrended this series by using linear regression and with differencing. The result of detrending via regression was a model that fit rather well and residuals that had no apparent dependency. Let us assume then that the true model for this data is

$$x_t = \beta_0 + \beta_1 t + w_t$$

where $w_t, t = 1, ..., T$ is normal white noise with variance σ^2. (This is the same as assuming that this data follows the standard regression assumptions.) Assuming this model, describe the model (i.e. write out a formula) for the differenced time series, ∇x_t. Use this to explain the apparent dependency in the differenced data from 1e above.

2. Load the data in "dailyibm.dat" using the command `ibm=scan("dailyibm.dat", skip=1)`. This series is the daily closing price of IBM stock from Jan 1, 1980 to Oct 8, 1992.

(a) Make a plot of the data and an ACF plot of the data. Does the time series appear to be stationary? Explain. Interpret the ACF plot in this situation.

(b) Difference the data. Plot this differenced data, and make an ACF plot. What is your opinion of whether the series is stationary after differencing?

(c) Another option for attempting to obtain stationary data when there is something similar to an exponential trend is to take the logarithm. Use the R command `log()` to take the logarithm of the data. Plot this transformed data. Does the transformed data appear stationary? Explain.

(d) Perhaps some combination of differencing and the logarithmic transform will give us stationary data. Why would `log(diff(ibm))` not be a very good idea? Try the opposite, difference the log transformed data `difflogibm=diff(log(ibm))`. Except for a few extreme outliers, does this transformation succeed in creating stationary data ?

(e) Delete the extreme outliers using the following command:

```
difflogibm=difflogibm[difflogibm> -0.1]
```

Plot this data and the ACF for this data. Sometimes with very long time series like this one, portions of the series exhibit different behavior than other portions. Break the series into two parts using the following commands:

```
difflogibm1= difflogibm[1:500]
difflogibm2= difflogibm[501:length(difflogibm)]
```

Plot both of these and create ACF plots of each. Do you notice a difference between these two sections of the larger time series?

(f) Assume the model for the data that we have called `difflogibm2` is of the following form:

$$d_t = \delta + w_t$$

where $w_t, t = 1, ..., T$ is normal white noise with variance σ^2. Is this reasonable from what you now know of this time series? How would you estimate δ and σ? Give the estimates.

3. (No R required.) Show that the $MA(3)$ model is weakly stationary. You need to show that the mean is zero and the covariance function depends only on distance. This will be very similar to what was done in class for the $MA(2)$ model.