Data Analysis & Graphics Using R, 2^{nd} edn – Solutions to Exercises (December 13, 2006)

Preliminaries

```
> library(DAAG)
```

Exercise 1
A time series of length 100 is obtained from an AR(1) model with $\sigma = 1$ and $\alpha = -.5$. What is the standard error of the mean? If the usual $\sigma/\sqrt{n}$ formula were used in constructing a confidence interval for the mean, with σ defined as in Section 9.1.3, would it be too narrow or too wide?

If we know σ, then the usual $\sigma/\sqrt{n}$ formula will give an error that is too narrow; refer back to Subsection 9.1.3 on pp. 288-289.

The need to estimate σ raises an additional complication. If σ is estimated by fitting a time series model, e.g., using the function `ar()`, this estimate of σ can be plugged into the formula in Subsection 9.1.3. The note that now follows covers the case where σ^2 is estimated using the formula

$$\hat{\sigma^2} = \frac{\sum(X_i - \bar{X})^2}{n-1}$$

The relevant theoretical results are not given in the text. Their derivation requires a kmowledge of the algebra of expectations.

Note 1: We use the result (proved below)

$$E[(X_i - \mu)^2] = \sigma^2/(1-\alpha^2) \tag{1}$$

and that

$$E[\sum(X_i - \bar{X})^2] = \frac{1}{1-\alpha^2}(n-1-\alpha)\sigma^2 \simeq \frac{1}{1-\alpha^2}(n-1)\sigma^2 \tag{2}$$

Hence, if the variance is estimated from the usual formula $\hat{\sigma^2} = \frac{\sum(X_i-\bar{X})^2}{n-1}$, the standard error of the mean will be too small by a factor of approximately $\sqrt{\frac{1-\alpha}{1+\alpha}}$.

Note 2: We square both sides of

$$X_t - \mu = \alpha(X_{t-1} - \mu) + \varepsilon_t$$

and take expectations. We have that

$$E[(X_t - \mu)^2] = (1-\alpha^2)E[(X_t - \mu)^2] + \sigma^2$$

from which the result (eq.1) follows immediately. To derive $E[\sum(X_i - \bar{X})^2]$, observe that

$$E[\sum(X_i - \bar{X})^2] = E[(X_t - \mu)^2] - n(\bar{X} - \mu)^2$$

Exercise 2
Use the `ar` function to fit the second order autoregressive model to the Lake Huron time series.

```
> ar(LakeHuron, order.max = 2)

Call:
ar(x = LakeHuron, order.max = 2)

Coefficients:
      1        2
 1.0538  -0.2668

Order selected 2  sigma^2 estimated as  0.5075
```

It might however be better not to specify the order, instead allowing the `ar()` function to choose it, based on the AIC criterion. For this to be valid, it is best to specify also `method="mle"`. Fitting by maximum likelihood can for long series be very slow. It works well in this instance.

```
> ar(LakeHuron, method = "mle")

Call:
ar(x = LakeHuron, method = "mle")

Coefficients:
      1        2
 1.0437  -0.2496

Order selected 2  sigma^2 estimated as  0.4788
```

The AIC criterion chooses the order equal to 2.

Exercise 3
Repeat the analysis of Section 9.2, replacing `avrain` by: (i) `southRain`, i.e., annual average rainfall in Southern Australia; (ii) `northRain`, i.e., annual average rainfall in Northern Australia.

The following functions may be used to automate these calculations. First, here is a function that gives the time series plots.

```
> bomts <- function(rain = "NTrain") {
+     plot(ts(bomsoi[, c(rain, "SOI")], start = 1900), panel = function(y,
+         ...) panel.smooth(bomsoi$Year, y, ...))
+ }
```

Next, here is a function that automates the calculations and resulting plots, for the analysis used for all-Australian rainfall data. The parameter choices may for some areas need to be varied, but output from this function should be a good start.

```
> bomplots <- function(loc = "NTrain") {
+     oldpar <- par(fig = c(0, 0.5, 0.5, 1), mar = c(3.6, 3.6,
+         1.6, 0.6), mgp = c(2.25, 0.5, 0))
```

```
+     on.exit(par(oldpar))
+     rain <- bomsoi[, loc]
+     xbomsoi <- with(bomsoi, data.frame(SOI = SOI, cuberootRain = rain^0.33))
+     xbomsoi$trendSOI <- lowess(xbomsoi$SOI)$y
+     xbomsoi$trendRain <- lowess(xbomsoi$cuberootRain)$y
+     rainpos <- pretty(rain, 5)
+     par(fig = c(0, 0.5, 0.5, 1), new = TRUE)
+     with(xbomsoi, {
+         plot(cuberootRain ~ SOI, xlab = "SOI", ylab = "Rainfall (cube root scale)",
+             yaxt = "n")
+         axis(2, at = rainpos^0.33, labels = paste(rainpos))
+         lines(lowess(cuberootRain ~ SOI))
+         lines(lowess(trendRain ~ trendSOI), lwd = 2, col = "gray40")
+     })
+     xbomsoi$detrendRain <- with(xbomsoi, cuberootRain - trendRain +
+         mean(trendRain))
+     xbomsoi$detrendSOI <- with(xbomsoi, SOI - trendSOI + mean(trendSOI))
+     par(fig = c(0.5, 1, 0.5, 1), new = TRUE)
+     plot(detrendRain ~ detrendSOI, data = xbomsoi, xlab = "Detrended SOI",
+         ylab = "Detrended rainfall", yaxt = "n")
+     axis(2, at = rainpos^0.33, labels = paste(rainpos))
+     with(xbomsoi, lines(lowess(detrendRain ~ detrendSOI)))
+     attach(xbomsoi)
+     xbomsoi.ma12 <- arima(detrendRain, xreg = detrendSOI, order = c(0,
+         0, 12))
+     xbomsoi.ma12s <- arima(detrendRain, xreg = detrendSOI, seasonal = list(order = c(0,
+         0, 1), period = 12))
+     print(xbomsoi.ma12)
+     print(xbomsoi.ma12s)
+     par(fig = c(0, 0.5, 0, 0.5), new = TRUE)
+     acf(resid(xbomsoi.ma12))
+     par(fig = c(0.5, 1, 0, 0.5), new = TRUE)
+     pacf(resid(xbomsoi.ma12))
+     par(oldpar)
+     detach(xbomsoi)
+ }
```

Data for further regions of Australia are available from the websites noted on the help page for `bomsoi`.

Exercise 4
In the calculation

```
Box.test(resid(lm(detrendRain ~ detrendSOI, data = xbomsoi)),
         type="Ljung-Box", lag=20)
```

try the test with `lag` set to values of 1 (the default), 5, 20, 25 and 30. Comment on the different results.

The calculation for a lag of 20 was given on page 296. Here are the results for the other suggested lags:

```
> if (!exists("xbomsoi")) {
+     xbomsoi <- with(bomsoi, data.frame(SOI = SOI, cuberootRain = avrain^0.33))
```

```
+      xbomsoi$trendSOI <- lowess(xbomsoi$SOI)$y
+      xbomsoi$trendRain <- lowess(xbomsoi$cuberootRain)$y
+ }
> xbomsoi$detrendRain <- with(xbomsoi, cuberootRain - trendRain +
+      mean(trendRain))
> xbomsoi$detrendSOI <- with(xbomsoi, SOI - trendSOI + mean(trendSOI))
> Box.test(resid(lm(detrendRain ~ detrendSOI, data = xbomsoi)),
+      type = "Ljung-Box", lag = 15)

        Box-Ljung test

data:  resid(lm(detrendRain ~ detrendSOI, data = xbomsoi))
X-squared = 32.8614, df = 15, p-value = 0.004905

> Box.test(resid(lm(detrendRain ~ detrendSOI, data = xbomsoi)),
+      type = "Ljung-Box", lag = 25)

        Box-Ljung test

data:  resid(lm(detrendRain ~ detrendSOI, data = xbomsoi))
X-squared = 38.436, df = 25, p-value = 0.04192

> Box.test(resid(lm(detrendRain ~ detrendSOI, data = xbomsoi)),
+      type = "Ljung-Box", lag = 30)

        Box-Ljung test

data:  resid(lm(detrendRain ~ detrendSOI, data = xbomsoi))
X-squared = 46.4144, df = 30, p-value = 0.02836
```

The p-values are:

n=15	n=20	n=25	n=30
0.005	0.023	0.042	0.028

Notice that the indication of sequential correlation is much stronger for n=15 than for larger values of n. As the number of possibilities that are canvassed increases (a greater number of lags at which there may be autocorrelations) the probability of detection of autocorrelation decreases.The small p-value for n=30 may thus seem surprising.