Series 5

1. In this exercise we investigate the connection between advertising expenditure and sales. Our example here is that of the annual advertising expenditure (`ADVERT`, in k$) for, and sales revenue (`SALES`, in k$) from, a particular brand of vegetable stock (Lydia Pinkham's Vegetable Compound, 1907 – 1960). These data are contained in the file `advert.dat`. Load this file and create a time series matrix:

```
> d.advert <- read.table("ftp://stat.ethz.ch/Teaching/Datasets/WBL/advert.dat",
+   header = TRUE)
> ts.advert <- ts(d.advert, start = 1907)
```

a) Make plots of both these time series, either simultaneously or below each other.
R hints: You can use `plot()` and `lines()` to do the plots, or, more elegantly,

```
> plot(ts.advert[...], plot.type = "single", lty = c(1, 2))
```

Note that the dataset has four columns, not only two; `SALES1` and `ADVERT1` indicate the sales revenue and advertising expenditure of the last year, resp.

b) Investigate the dependence of sales revenue (`SALES`) on advertising expenditure in the current year (`ADVERT`) as well as the last (`ADVERT1`) by means of a linear model. Investigate whether or not the residuals are independent. What are the consequences of these results?
R hints:

```
> lm.advert1 <- lm(SALES ~ ADVERT + ADVERT1, ...)
> summary(lm.advert1)
> acf(...)
```

c) Extend this model to include the previous year's sales revenue (variable `SALES1`), and analyse the residuals of this extended model (using Tukey-Anscombe and normal plots).

d) Another way to build a model is to look at the changes in sales revenue, $\texttt{D\_SALES}_t = \texttt{SALES}_t - \texttt{SALES}_{t-1}$, compared to the changes in advertising expenditure $\texttt{D\_ADVERT}_t = \texttt{ADVERT}_t - \texttt{ADVERT}_{t-1}$. What fitted model do you obtain in such a case? Look at the residuals, too.

e) Compare the models obtained in c) and d).

2. The dataset `voc2.dat` contains daily averages of measurements of volatile organic compound and other measurements in Wallisellen from 6.6.1996 till 31.8.1996.

`t`	Temperature (°C)
`rf`	Humidity
`O3`	Ozone
`X2mbu`	2-Methylbutan (ppbC)
`nbu`	Butan (ppbC)

We want to investigate different linear regression models for the ozone level as a function of temperature and humidity.

a) Read in the data and convert the relevant columns to a time series object:

```
> d.voc <- read.table("ftp://stat.ethz.ch/Teaching/Datasets/WBL/voc2.dat",
+   header = TRUE, sep = ";")
> ts.voc <- ts(d.voc[, c("t", "rf", "O3")])
```

Produce a pairs plot:

```
> pairs(data.frame(ts.voc))
```

Describe the influence of temperature and humidity on the ozone level.

b) Perform a linear least squares regression of the variable `O3` vs. `t` and `rf`. Which variables are significant? Are the assumptions for the errors satisfied? Perform a residual analysis.

c) Perform one iteration of the method of Cochrane and Orcutt to overcome the weakness of the linear model of part b). Again, perform a residual analysis. Is this model better than that of part b)?

R hint: the R function `lag()` can be helpful to do the required transformation of the time series. Use `colnames()` to make sure that the transformed model has the same column names as the initial one; this makes the call of `lm()` more clear.

d) Look at the untransformed time series again. Perform a *generalized* least squares regression of `O3` vs. `t` and `rf` using the R function `gls()` of the package `nlme`. Analyse the residuals and comment on them.

R hints:

```
> library(nlme)
> times <- time(ts.voc)
> gls.voc <- gls(..., data = ..., method = "ML",
+   correlation = corARMA(form = ~ times, p = ..., q = ....))
```

e) In parts **??** to d), you used three different methods to estimate the coefficients of the model

$$\texttt{O3} = \beta_0 + \beta_1 \cdot \texttt{t} + \beta_2 \cdot \texttt{rf} .$$

Compare the estimates $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\beta}_2$ for the three different estimation methods.

3. Let E_t be white noise with mean 0 and variance σ^2. Regard the following three processes:

i) $X_t = t + E_t$

ii) $Y_t = X_t - X_{t-1}$

iii) $Z_t = X_t - t$

a) Which of these three processes are stationary, and which are not? Why?

b) Compute the theoretical autocorrelation of the processes Y_t and Z_t and the cross-correlation between the two.

Hint: Write these processes using white noise E_t.

c) Simulate both Y_t and Z_t. To this end, assume that E_t follows a standard normal distribution $\mathcal{N}(0, 1)$. Simulate time series of length $n = 200$, and compare your empirical results to the theoretical ones of Part b).

R hint: suppose your simulated time series Y_t and Z_t are stored in R objects `ts.Y` and `ts.Z`, respectively; you can then plot auto- and cross-correlations using

```
> acf(ts.intersect(ts.Y, ts.Z), ylim = c(-1, 1), plot = TRUE)
```

Preliminary discussion: Monday, May 02.

Deadline: Monday, May 09.