Tutorial 3

Begin by loading the packages `urca` and `fpp2`; the second of these contains most of the data sets we will be using.

1. **Efficiency when fitting an ARMA model** The aim of this exercise is to evaluate the commands for fitting an ARMA Time Series model. Firstly, data is simulated from a given ARMA model, then an ARMA model is fitted to the simulated data.

   Let us simulate an ARMA process

   $$X_t - 0.5X_{t-1} + 0.5X_{t-2} = Z_t + 0.8Z_{t-1} + 0.2Z_{t-2}.$$

```
ar <- arima.sim(n = 300, list(ar = c(0.5,-0.5), ma = c(0.8, 0.2), sd
= 1))
plot(ar)
```

   Now try to fit an ARMA to the simulated data:

```
?arima
arima(ar, order = c(2,0,2))

Call:
arima(x = ar, order = c(2, 0, 2))

Coefficients:
        ar1      ar2     ma1     ma2  intercept
      0.539  -0.5557  0.7536  0.2296    -0.1307
s.e.  0.093   0.0608  0.1092  0.0955     0.1076

sigma^2 estimated as 0.9131:  log likelihood = -413.25,  aic = 838.49
```

   The returned coefficients here seem close to those of the proposed model; each estimated coefficient is within one standard error of the true value.

   If the model is unknown and we try a larger model,

```
arima(ar, order = c(4,0,4))

Call:
arima(x = ar, order = c(4, 0, 4))

Coefficients:
         ar1      ar2     ar3      ar4      ma1      ma2     ma3
```

```
      2.1559  -2.1840  1.2721  -0.3949  -0.8934  -0.2174  0.2684
s.e.  0.4818   0.8121  0.7070   0.3043   0.4847   0.2384  0.1789
         ma4  intercept
      0.2224    -0.1290
s.e.  0.1480     0.1361

sigma^2 estimated as 0.8799:  log likelihood = -407.86,  aic = 835.72
```

The fitting is not so good; the coefficients are nothing like the true values.

- Plot the autocorrelation functions for the true model and the two estimated models. Are they similar? Even if the models look different, they may have similar autocorrelation functions.
- Try with different values of n and different models to see when the 'arima' fitting routine is successful.
- What does `auto.arima` produce?
- Simulate AR(p) processes. Does the fitting work better when the process does not have a MA component?
- **Estimation Methods** The package `itsmr` is useful; it contains a variety of estimation methods (covered later); for example, the Hannan-Rissainen method for general ARMA processes. Install and activate this package:

  ```
  install.packages("itsmr")
  library("itsmr")
  ```

  and use it to investigate the various methods for model building. For example, to fit an ARMA process by the Hannan-Rissainen algorithm, use:

  ```
  hannan(ar,3,3)
  ```

  Check the syntax in the documentation.

2. **Corticosteroid drug sales in Australia** We will try to forecast monthly corticosteroid drug sales in Australia. These are known as H02 drugs under the Anatomical Therapeutic Chemical classification scheme.

```
lh02 <- log(h02)
cbind("H02 sales (million scripts)" = h02,
      "Log H02 sales"=lh02) %>%
  autoplot(facets=TRUE) + xlab("Year") + ylab("")
```

What about the plots? Do you see seasonality? Trend?

The data are from July 1991 to June 2008. There is a small increase in the variance with the level, so we take logarithms to stabilise the variance.

The data are strongly seasonal and obviously non-stationary, so seasonal differencing will be used.

```
lh02 %>% diff(lag=12) %>%
  ggtsdisplay(xlab="Year",
    main="Seasonally differenced H02 scripts")
```

Does it seem advisable to difference twice?

In the plots of the seasonally differenced data, there are spikes in the PACF at lags 12 and 24, but nothing at seasonal lags in the ACF. This may be suggestive of a seasonal AR(2) term. In the non-seasonal lags, there are three significant spikes in the PACF, suggesting a possible AR(3) term. The pattern in the ACF is not indicative of any simple model.

Try various seasonal arima models. Which has the best AIC value?

```
(fit <- Arima(h02, order=c(3,0,1), seasonal=c(0,1,2),
              lambda=0))
```

Check the residuals for your fitted model. Do they look like white noise? Do they look normal?

```
checkresiduals(fit, lag=36)
```

Does the model pass the Ljung Box test?

Now try `auto.arima` with all arguments left as default. Which model does it choose? Does the model pass the Ljung Box test?

Now try to use the model for forecasting.

```
h02 %>%
  Arima(order=c(3,0,1), seasonal=c(0,1,2), lambda=0) %>%
  forecast() %>%
  autoplot() +
    ylab("H02 sales (million scripts)") + xlab("Year")
```

The last few observations appear to be different (more variable) from the earlier data. This may be due to the fact that data are sometimes revised when earlier sales are reported late.

Does taking logs make a serious difference when stabilising the residuals? Does it affect the results of the Ljung Box test?

3. **Comparing ARIMA modelling and Holt Winters (exponential smoothing)** Exponential smoothing can be done using the `ets()` function from **fpp2**.

   In this case we want to compare seasonal ARIMA and ETS models applied to the quarterly cement production data `qcement`. Because the series is relatively long, we can afford to use a training and a test set rather than time series cross-validation. The advantage is that this is much faster. We create a training set from the beginning of 1988 to the end of 2007 and select an ARIMA and an ETS model using the `auto.arima()` and `ets()` functions.

```
# Consider the qcement data beginning in 1988
cement <- window(qcement, start=1988)
# Use 20 years of the data as the training set
train <- window(cement, end=c(2007,4))

(fit.arima <- auto.arima(train))
checkresiduals(fit.arima)
```

   This looks reasonable. The exponential smoothing model also does reasonably well:

```
(fit.ets <- ets(train))
checkresiduals(fit.ets)
```

   Now evaluate the forecasting performance of the two competing models over the test set. Which seems better?

```
a1 <- fit.arima %>% forecast(h = 4*(2013-2007)+1) %>%
  accuracy(qcement)
a1[,c("RMSE","MAE","MAPE","MASE")]
a2 <- fit.ets %>% forecast(h = 4*(2013-2007)+1) %>%
  accuracy(qcement)
a2[,c("RMSE","MAE","MAPE","MASE")]
```

   Now generate and plot forecasts from an ETS model for the next 3 years.

```
cement %>% ets() %>% forecast(h=12) %>% autoplot()
```

4. Consider the data set `sheep`, the sheep population of England and Wales from 1867–1939.

   (a) Produce a time plot of the time series.

   (b) Assume you decide to fit the following model:

$$y_t = y_{t-1} + \phi_1(y_{t-1} - y_{t-2}) + \phi_2(y_{t-2} - y_{t-3}) + \phi_3(y_{t-3} - y_{t-4}) + \epsilon_t$$

   where ϵ_t is a white noise series. What sort of ARIMA model is this (i.e., what are p, d, and q)?

(c) By examining the ACF and PACF of the differenced data, explain why this model is appropriate.

(d) The last five values of the series are given below:

Year	1935	1936	1937	1938	1939
Millions of sheep	1648	1665	1627	1791	1797

The estimated parameters are: $\phi_1 = 0.42$, $\phi_2 = 0.20$, $\phi_3 = 0.30$, Without using the `forecast` funtion, compute the forecasts for the next three years (1940–1942).

(e) Now fit the model in R and obtain the forecasts using forecast. How are they different from yours? Why?