Topics of the lab: different methods for estimating the parameters, modelling real time series data, predictions in the case of fitting models to transformed data.

There are several ways to estimate parameters of a given model. If we make an assumption about the distribution of innovations (and assume that the innovations correspond to iid random variables), it is possible to find parameters by maximizing the likelihood of the observed series, and this is the default method used in R (by assuming that the innovations are normally distributed). An alternative, which does not require making assumptions about the distribution of the innovations, is to find the parameters by minimizing a measure of accuracy like MSE, MAD, MAPE or some other measure. The most common approach is to minimize the MSE (or, equivalently, the sum of squared one-step prediction errors, SSE). Since for computing prediction errors for all time moments, it is necessary to give values starting values for some variables and the computed errors are valid if we assume that the starting values are correct, the sum of squared errors is called "conditional sum of squared errors", which is abbreviated to CSS. This is other possibility implemented in the arima() command. This is computationally much faster than maximizing the full likelihood, so for very long time series with long term dependencies it may speed up the fitting process considerably. Finally, maximizing the likelihood requires starting values for parameters and since the likelihood function may have several local maxima, the final result may depend on the choice of those values. A quite natural choice is to find starting values by minimizing CSS, and such approach is also implemented in the arima() command (option method can take values ML,CSS,CSS−ML, the default is ML).

A remark about comparing different models. If models with different number of parameters are fitted **to the same data** (all models to the original data or all models to differences of the data or all models to seasonal differences of the data or all models to a transformed version of data and so on), we can compare their relative performance on the same data only if we take into account the number of parameters. The only question is how exactly to balance the effect of added parameters and the increase of likelihood (or the reduction of a performance measure). Two common measures are Akaike information criterion (it is recommended to use AICc version if the number of parameters is relatively large compared to the number of observations) and Bayesian information criterion

$$AIC = 2k - 2\ln(L),\ AICc = AIC + \frac{2k(k+1)}{n-k-1},\ BIC = k\ln n - 2\ln(L).$$

Smaller values correspond to models which are more consistent with the data (in certain sense). If you plan to base your decisions on the values of those criteria, please try to read the information aobut them and try to understand, what is the logic behind them.

Finally, let us consider the question of predicting future values of the original series $y_t,\ t = 1, \dots, n$ in the case when we have fitted a model to $z_t = f(y_t)$, where f is an invertible function with the inverse function g. Let $\hat{z}_{t+1}$ be our forecast for z_{t+1} at time t. Then

$$z_{t+1} = \hat{z}_{t+1} + A_{t+1},$$

where A_{t+1} is a centered (with zero mean) random variable and we also know (an estimate for) it's standard deviation. Then $g(\hat{z}_{t+1})$ is usually not an unbiased estimate for y_{t+1} and we have to correct the estimate to reduce bias (if we want to have unbiased estimate). The exact form of correction depends on f and the distribution of A_{t+1}. For example, in the case $f(y) = \ln y$ we have

$$\hat{y}_{t+1} = e^{\hat{z}_{t+1}} E(e^{A_{t+1}})$$

and the value of $E(e^{A_{t+1}})$ is known for many distributions (it is the value of the moment generating function at 1). If we assume that innovations are normally distributed, then

$$\hat{y}_{t+1} = e^{\hat{z}_{t+1} + \frac{\sigma_A^2}{2}}.$$

Here the standard deviation of A_t is usually found by model fitting method. If the distribution of A_{t+1} is not known or the exact bias is difficult to find, we can remove the main part of the bias by defining

$$\hat{y}_{t+1} = g(\hat{z}_{t+1}) + \frac{\sigma_A^2}{2} g''(\hat{z}_{t+1}).$$

Exercises:

1. Find suitable ARIMA or SARIMA models for time series in the file `lab11.csv` (available from Moodle).
2. Find suitable ARIMA or SARIMA models for time series data used in the first and second lab