EBC4222
Descriptive and Predictive Analytics
Tutorial 1 Exercises: ARMA Models

Roselinde Kessels
r.kessels@maastrichtuniversity.nl

23 April 2024

Questions

1. This question relates to the unconditional and conditional expectations in ARMA models. For the derivations, use the following properties of expectations.
   For two random variables x, y and scalar a, the following equalities hold for unconditional expectations:

$$
\begin{aligned}
E(a \times x) &= a \times E(x) \\
E(x + y) &= E(x) + E(y)
\end{aligned}
$$

   where $E()$ is the expectation operator.

   (a) Calculate the unconditional expectation, hence the long-run mean, of an AR(2) model below:

$$y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 y_{t-2} + \varepsilon_t,$$

   assuming that ε_t is a white noise series (mean 0 and variance σ^2), $\beta_0, \beta_1, \beta_2$ are known model parameters and data y_t are stationary.

   (b) To expand on part (a), what are the three conditions for a time series model to be stationary?

   (c) Calculate the conditional expectation of $y_t | y_{t-1}, y_{t-2}$ in the AR(2) model $y_{t-1} = 0.1$, $y_{t-2} = 0.2$ and the estimates of model parameters $\hat{\beta}_0 = 1, \hat{\beta}_1 = 0.5$, $\hat{\beta}_2 = 0$, $\hat{\sigma}^2 = 0.1$.

   (d) Consider the below random walk model:

$$y_t = y_{t-1} + \varepsilon_t, \;\; \varepsilon_t \sim NID(0, \sigma^2).$$

We want to show that the 'shocks' in this model do not die out. I.e. the effect of a change in y_t in the past affects all following time periods. For this purpose, show that the conditional expectation of y_t given y_{t-k} is the following:

$$E(y_t|y_{t-k}) = y_{t-k}.$$

2. The purpose in this question is to analyze time series properties of monthly beer sales. Monthly beer sales are provided in *R* package `TSA` in millions of barrels, 01/1975 - 12/1990.

   (a) Plot the beer sales data, and the ACF and PACF of the beer sales data. Comment on the patterns such as trends and seasonality.
   Hint: For monthly sales data, one of the first patterns to suspect is monthly autocorrelation (seasonality). Comment on the ACF and PACF figures based on this expectation.

   (b) We next test the beer sales data for stationarity. For this, use two tests, the Augmented Dickey Fuller test (see `adf.test` in *R*) and the KPSS test (see `kpss.test` in *R*). Carefully study the properties of ADF and KPSS tests. What do you conclude about the stationarity of the data. Do you think these results are affected by possible seasonal patterns in the data?

   (c) Create two dummy variables to test for seasonal patterns in the data. Specifically, create a variable `dJan` that takes the value of 1 in January, and 0 in other months. In addition, create a variable `dDec` that takes the value of 1 in December, and 0 in other months. Estimate a linear regression model (see `lm` in *R*) by regressing beer sales on the two dummy variables. This regression provides the 'seasonal patterns' with respect to the January and December months. Remove the fitted values of this regression from the original beer sales data to 'remove seasonality'.
   Hint: Define a variable `lm_seasonal` that is the output of the `lm` regression. Then define `beersalesSeasonAdj` as the difference between `beersales` and `lm_seasonal$fitted.values`

   (d) Plot the seasonally adjusted data in part (c), and the ACF and PACF of these data. Do you think that the January and December dummies removed the seasonal patterns in the data?

   (e) Apply the sine-cosine method for smoothing (de-trending) the data in addition to the January and December dummy variables by using the code below, adjusting the appropriate 'frequency'. Remove this trend from beer sales, plot the ACF and PACF of the detrended data. Does this method remove correlation patterns from the data?

```r
t <- 1:length(beersales)
cos.t <- cos(2*pi*t/frequency)
sin.t <- sin(2*pi*t/frequency)
trend <- lm(beersales ~ t + cos.t + sin.t + dDec + dJan)$fitted.values
```

(f) Apply four ARMA models to apply to the original (not de-trended) beer sales data. Specifically, estimate an ARMA(1,1) model, an AR(12) model, an ARMA(1,1) model with a deterministic trend, and an AR(12) model with a deterministic trend. Define the deterministic trends 'within a year' as below, and add `time` as an explanatory variable for the models with a deterministic trend.

```r
time <- rep(1:12, length(beersales)/12)
```

(g) The next purpose is to choose the most appropriate models in part (f) for the data. Calculate the Bayesian Information Criteria (BIC) for the four models in part (f). Based on these criteria, which model performs best, and which models seem to 'overfit' the data? In addition, compare the Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) of the alternative models. Which model performs best according to these criteria?

(h) Take the 'best performing model' according to the BIC in part (g). Test the significance of the model coefficients using t-tests. Comment on these tests of significance. Inspect the residuals of this model. Do you think that the residuals have (remaining) autocorrelation?
Hint: Recall the formula and critical values of a t-test. In addition, use `coef()` and `sqrt(diag(vcov()))` to extract the coefficients and standard errors of ARIMA model estimates to calculate the t-statistics. For the t-statistics, you can use the following quantile function:

```r
  t_value <- qt(p = c(0.05, 0.95), df = (T-n))
```

where T is the number of observations in the data, and n is the number of coefficients in the model (including the error variance parameter).