EBC4222
Descriptive and Predictive Analytics
Tutorial 3 Exercises: ARIMA and seasonal ARIMA Models

Roselinde Kessels
r.kessels@maastrichtuniversity.nl

30 April 2024

Questions

1. This question relates to forecasting in ARIMA (particularly AR) models.

   (a) Calculate the mean and variance of the one-step ahead forecast of y_t in the following AR(1) model:

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

   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.
   Hint: For this you will need to calculate the conditional expectation and variance of $y_{T+1}|y_{1:T}$, where T is the sample size and $y_{1:T} = \{y_1, \ldots, y_T\}$. To compute the variance, start from the definition:
   $Var(y_{T+1}|y_{1:T}) = E[(y_{T+1}|y_{1:T} - E(y_{T+1}|y_{1:T}))^2]$.

   (b) Calculate the mean and variance of the two-step ahead forecast, $y_{T+2}|y_{1:T}$, where $y_{1:T} = \{y_1, \ldots, y_T\}$.
   Hint: The easiest method is to calculate the mean and variance of $y_{T+1}|y_{1:T}$ first (as in part (a)), and then calculate the two-step ahead forecast based on this.

   (c) Suppose that the AR(1) model in the question is used for $T = 200$ observations, and the data has a breakpoint at time $t = 100$. Write down the AR(1) model with a dummy variable that indicates a change in the mean starting from observation $t = 100$.

   (d) Calculate the mean and variance of the one-period ahead forecasts for the model in part (c). Assume that the coefficient of the break dummy is known.

(e) In many applications, it is argued that a 'random walk' model is very hard to beat in terms of the forecast performance. The point in this exercise is to see that this model might not have a good performance for multiple period ahead forecasts. In order to see this, calculate the variance of the k-period ahead forecasts of the below random walk model:

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

How does the forecast variance change with k? Relate your finding to the claim 'this model might not have a good performance for multiple period ahead forecasts'.

2. The purpose in this question is to understand how a regression with non-stationary data can lead to a spurious regression using simulated (artificial) data.

(a) Use the script below to simulate two independent random walk series. Plot the two simulated series, and inspect this plot to see if there is any 'co-movement' between the series.

```
    # Simulate two independent random walks
set.seed(12345)
rwalk1 = c(cumsum(rnorm(1000)))
rwalk1.ts = ts(rwalk1)
rwalk2 = c(cumsum(rnorm(1000)))
rwalk2.ts = ts(rwalk2)
seqplot.ts(rwalk1.ts, rwalk2.ts)
```

(b) First fit an AR model to one of the series. For this, use the R function `ar.ols`, with the option to select the number of lags using Akaike Information Criteria (AIC). Comment on the parameter estimates. Do these estimates indicate a stationary data series?

(c) Obtain the 'spurious regression results' by estimating a linear regression model regressing the first series on the second. Comment on the significance of the obtained coefficients and relate your finding to the issue of a spurious regression.

(d) To remove the problem of a spurious regression, test whether both time series are stationary and if not, difference them until they become stationary. Obtain the new regression results with the (differenced or integrated) time series and compare with part (c).

3. The purpose in this question is to understand how a regression with non-stationary data can lead to a spurious regression in a real data example and understand the intuition of unit root tests (tests for a stochastic trend). For the following exercises, use the Google trends data file, `correlate-historical_data.csv`. This file includes the weekly number of Google searches for a set of keywords. Define two variables `histData` as the 'historical data' and `computational` as the 'computational' variables

in the csv file.
Source: https://trends.google.com

(a) Estimate a linear regression model (using `lm`) where the number of searches for 'historical data' are explained by an intercept and the number of searches for 'computational'. Print the summary of this regression and comment on the causality: Do you think the searches for 'computational' explain the searches for 'historical data'? Did you expect these two variables to be correlated?

(b) Estimate an AR(3) model for `histData` and add `computation` as an explanatory variable for this model. Comment on the estimation results using a 5% significance level. Compare the effect of 'computation' on 'historical data' with the one obtained in part (a).

In order to remove the problem of a spurious regression, we should have first tested the series for stationarity. It is often difficult to decide on a deterministic or stochastic trend in the data both visually and statistically. In several applications deterministic and stochastic trend detection is done in two steps:

1 Removing the deterministic trend from the data.
2 Testing the de-trended series for a stochastic trend using unit root tests, e.g. ADF or KPSS test provided in the *R* package *aTSA*.

The following exercises aim to set up the theoretical reason for following this approach, and to apply it to a real dataset.

(c) Consider the random walk model for x_t (stochastic trend):

$$x_t = x_{t-1} + \varepsilon_t$$

Consider also a deterministic trend for each time period:

$$\texttt{trend}_t = \texttt{trend}_{t-1} + 1$$

Define a new random variable $y_t = x_t - \beta \times \texttt{trend}_t$. Show that this variable is a random walk (with a drift if $\beta \neq 0$):

$$y_t = y_{t-1} - \beta + \varepsilon_t.$$

Motivate the use of the two-step trend detection method based on the above derivations.

(d) De-trend the `histData` series using a linear regression with a trend. Apply an ADF stationarity test to test for a stochastic trend in the series. What is your conclusion about the stationarity of `histData`? Do you find a deterministic and/or stochastic trend in `histData`?

4. This question is related to ARMA model selection based on forecast comparisons. There are several applications where we are interested in obtaining good forecasts, but not necessarily good in-sample fit metrics such as R^2, AIC or BIC. For the illustrations, use the data on monthly beer sales provided in the *R* package `TSA` in millions of barrels, 01/1975 - 12/1990. See https://cran.r-project.org/web/packages/TSA/TSA.pdf

   (a) Split the data into an estimation (training) sample and a forecast (test) sample. Specifically, leave out the last 92 observations of the data for forecast comparisons.

   (b) Estimate two ARMA models for the estimation data: estimate an ARMA(1,1) model, and an ARMA(1,1) model with a deterministic trend. Define the deterministic trend 'within a year' as below, and add `time` as an explanatory variable for the model with a deterministic trend.

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

   (c) Forecast/predict the beer sales data for the forecast sample for both models in part (b). Compare the Mean Squared Forecast Error (MSFE) and Mean Absolute Forecast Error (MAFE) of the models. Which model performs best in terms of these forecast comparison metrics?

   Hint: You can use the *R* package *forecast* (not *aTSA*) for this purpose. See the example code to obtain forecasts from this package, where `beersales_est` is the estimation data and `beersales_for` is the forecast data.

```
# ARMA(1,1) model estimation
fit_arma11 <- Arima(beersales_est, order = c(1, 0, 1))
# include forecast sample
for_arma11 <- forecast(fit_arma11, h = 92, level = c(0.05, 0.95))
plot(for_arma11)
MSFE_arma11 <- mean((beersales_for - for_arma11$mean)^2)
MAFE_arma11 <- mean(abs(beersales_for - for_arma11$mean))
```

   (d) Plot the forecasts and the forecast intervals of the ARMA(1,1) model as given in the previous question's hint. Discuss why the forecast intervals are large. To give an answer to this, also think about the variance derivations in question 1.b.

5. We will use the `SARIMA` function in the accompanying *R* package `astsa` of Shumway and Stoffer (2016). As the authors note, do not install this package from CRAN but use the more up-to-date `github` version:

```
# install R packages for examples in the book.
# see: https://github.com/nickpoison/astsa/astsa_build
 install.packages("devtools")     # only need to do this once
 devtools::install_github("nickpoison/astsa/astsa_build")
```

We will analyze the monthly beer sales data provided in the *R* package `TSA` in millions of barrels, 01/1975 - 12/1990. Because there is a significant overall increasing trend, work with the 'de-trended' beer sales data for this question.

(a) Define the trend as

```
time <- c(1:length(beersales))
```

Plot the de-trended data, ACF and PACF, and the monthly averages using function `monthplot`. Apply an ADF test to check for stationarity of the data. What do you conclude about data stationarity (after de-trending) and about possible seasonality in these data. In other words, do you think a SARMA, ARIMA or SARIMA model could be more useful than an ARMA model?

(b) **General to specific model selection**
The purpose in this model selection is to start with a relatively large model. Based on this large model, a more parsimonious model is selected based on coefficient tests. Estimate an ARMA(10, 0) model with 1 seasonal lag using `sarima` (p = 10, P = 1, S = 12). Based on the coefficient tests (t-tests and p-values of these tests), define and estimate smaller models step-by-step. What is the final model you achieve?

Hint: Model selection based on t-tests have to be made sequentially. I.e. you remove one variable and re-estimate the model again. Use the t-tests on the highest lag coefficient at each time.

(c) **Specific to general model selection**
The purpose here is to start with a relatively small model, and increase the number of coefficients based on residual diagnostics. To inspect the residuals, you can use the ACF of the residuals, Q-Q plot of the residuals and the p-values of the Ljung-Box test which are plotted automatically using `sarima`.

Start with an ARMA(2, 0) model with 1 seasonal lag. Increase the AR lags to higher numbers, until you think that the residual plots do not show major problems.

(d) **Model choice based on information criteria**
The two model choice tools in parts (a) and (b) can lead to different models or be inconclusive particularly in small samples. A further tool to compare models and select a model is using information criteria. Consider all models you estimated in parts (a) and (b). Report the Bayesian Information Criteria (BIC) for all models. Report the best performing model in terms of the BIC.

(e) **Model choice for non-nested models using information criteria**
So far we have compared 'nested models', i.e. one model could be defined in terms of the other model only with one or more coefficients fixed to 0. Coefficient tests, such as the t-test, can be used to compare and choose between such

nested models, as we have seen in parts (a) and (b) of the question.
One advantage of information criteria is that they can also be used to compare non-nested models. To see this, consider the sine-cosine estimation from Tutorial 1. Apply the model to de-trended beer sales:

```
t <- 1:length(beer_detrended)
cos.t <- cos(2*pi*t/12)
sin.t <- sin(2*pi*t/12)
outSinCos <- lm(beer_detrended ~ cos.t + sin.t)
```

Report the BIC of this regression, and compare the result with that in part (d). Which model do you now choose for modeling de-trended beer sales?

(f) Explain why the models below are nested. Which model is the nested model, and which model the nesting model?

$$y_t = \beta_0 + \beta_1 y_{t-1} + \varepsilon_t.$$

$$y_t = \alpha_0 + \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + \epsilon_t.$$

(g) **Comparing results for data with and without a trend:** Explain why we cannot use the above model comparison tools to compare a model for beer sales (not de-trended) and a model for de-trended beer sales. Could you use forecast comparisons to discriminate between these models?