Predictive Analytics - Exercise Sheet 3

(graded)

Prof. Dr. Benjamin Buchwitz

2024

Task 1

a) What does stationarity mean? Describe it in your own words.

b) What is the consequence of stationarity concerning the mean (μ), variance (σ^2) and autocorrelation function (ACF) of a time series?

Task 2

a) What does *differencing* time series data mean? Why and when do you do that?

b) What *forms* of differencing exist and what do they mean?

c) Consider the time series `datasets::AirPassengers` (R Core Team 2021).

c.1) Plot the time series after transforming it to a `tsibble` object (Wang, Cook, and Hyndman 2020) and describe its features.

c.2) Then, stabilize its variance by using a Box-Cox transformation and plot the modified data.

c.3) Afterwards, stabilize the mean. How often do you have to difference the data to make it stationary and why? Make use of the *Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test* to test if and how often *differencing* is necessary. What are the hypotheses of the KPSS test? Plot the (modified) time series **each time** when the time series has been *differenced* and save the values you get after each *differencing* in (a) separate column(s) in your `tsibble`.

d) Formulate an equation which reflects the number and forms of differencing you did in task 2.c.3). Then convert this equation by using the backshift notation. Then multiply out the brackets.

Task 3

a) How does a random walk (RW) model look like? Describe it and write down its equation.

b) Implement a random walk model in R by writing a function named `randomWalk()` that is dependent on

- the number of time points (argument `times`) and
- on a trend (drift term) c (argument `trend_c`).

Your function should return the generated times series (argument `y`). Assume that ϵ_t is white noise and that the process starts with $y_1 = 0$. Since the way of your implementation will be graded, find a smart and short solution. Avoid loops!

c) Then, simulate five random walk series with 100 time points each: three of them with no drift, one with a positive drift and one with a negative drift. Present all five random walk series in one plot. Add a legend in which c is defined.

Task 4

In the course and in some of the previous exercise sheets you have worked with the drift method: You have seen that the drift method can be applied to data with y as the dependent variable based on the function `fable::RW()` which creates a random walk model (O'Hara-Wild, Hyndman, and Wang 2021) as follows:

```
data %>%
model(
  Drift = RW(y ~ drift()),
)
```

Write down the equations of a random walk model and the drift method and show and explain why the drift method can be implemented by a random walk model as described in the previous code chunk.

Task 5

a) What is the difference between a moving average you got to know from exercise sheet 1 and from chapter 3 in Hyndman and Athanasopoulos (2021) and the moving average component of an ARIMA model?

b) How does a non-seasonal ARIMA model look like? Write down its general equation. Describe and explain the components of an ARIMA model in your own words.

c) What is the difference between an ARMA and an ARIMA model?

d) Write down the equation of an ARIMA(0,1,0) and show that a random walk (RW) model (see task 3.a)) corresponds to it. Can you formulate a random walk model as an ARMA model? Why is it (not) possible?

Task 6

In task 2 of this exercise sheet you have already visualized and *differenced* the time series `datasets::AirPassengers` (R Core Team 2021). In this task, you will need those results again.

a) Based on your results from task 2.c), do you need an ARIMA(p, d, q) or an ARIMA$(p, d, q)(P, D, Q)_m$ to model the time series `AirPassengers`? Give an explanation.

b) Create a time series plot, an acf plot and a pacf plot of your *modified* `AirPassengers` time series data (after stabilizing the variance and mean).

b.1) Based on the acf and pacf plot, which ARIMA models (mention at least two models) do you assume to be appropriate to describe the `AirPassengers` time series and why? Give an explanation for your choice. Then define your chosen ARIMA models by using `fable::ARIMA()` (O'Hara-Wild, Hyndman, and Wang 2021). Look at the AIC, AICc and BIC of your defined models.
Can you find a *better* (AICc) ARIMA model? Which ARIMA model minimizes the AICc and is consequently the *best*? *Hint*: To search for the *best* ARIMA model set `stepwise = FALSE`, `approx = FALSE` and `greedy = FALSE`.

b.2) Check the residuals of the ARIMA model you have chosen as the *best* one in task 6.b.1). Do the residuals resemble white noise? Examine the residuals by visualizing them (time series plot, acf plot and histogram of the residuals) and by using a Ljung-Box test. Describe and interpret the results.

c) Use the Backshift notation to write down the ARIMA model you have chosen as the *best* one in task 6.b.1). *Hint:* If your chosen ARIMA model has the form ARIMA$(p, d, q)(P, D, Q)_m$, start with using the backshift notation for ARIMA(p, d, q) and then expand your equation by adding the second factor $(P, D, Q)_m$ to your equation.

d) Expand your equation from task 6.c) by multiplying out the brackets so that an equation of the form $y_t = ...$ results.

e) Which values do the parameters of the ARIMA model - you have chosen as the *best* one in task 6.b.1) - have? Substitute the parameters by the concrete values in your equation from task 6.d).

f) Based on the ARIMA model you have chosen in task 6.b.1) as the *best* one, produce forecasts for the next three years with confidence intervals of 80 % and 95 % coverage probability and plot them together with the time series `AirPassengers` in one plot and add a legend. Do you think your forecasts are reliable? Why or why not?

References

Hyndman, Rob J, and George Athanasopoulos. 2021. *Forecasting: Principles and Practice*. 3rd ed. Springer-Lehrbuch. Melbourne, Australia: OTexts.

O'Hara-Wild, Mitchell, Rob Hyndman, and Earo Wang. 2021. *Fable: Forecasting Models for Tidy Time Series*. https://CRAN.R-project.org/package=fable.

R Core Team. 2021. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Wang, Earo, Dianne Cook, and Rob J Hyndman. 2020. "A New Tidy Data Structure to Support Exploration and Modeling of Temporal Data." *Journal of Computational and Graphical Statistics* 29 (3): 466–78. https://doi.org/10.1080/10618600.2019.1695624.