EC 361–001

Problem Set 3

Prof. Santetti

Spring 2024

`INSTRUCTIONS`: Carefully read all problems. You must submit a single R script with your *first name(s)* (mine would be `marcio.R`). In case you submit your files with different names, you will lose 1 point.

You can find templates for your answer R script on `theSpring`, under the "Templates" module. Please consider using it.

I should be able to fully replicate your code to answer the questions, as well as fully understand your written interpretations to the proposed problems.

Avoid using unnecessary code in your submission files. It is totally fine to do other things by yourself that may help you better understand the data and the problems. However, for grading purposes, I am only interested in the commands and interpretations that actually answer the questions. You may keep a separate file for yourself with your additional explorations.

Recall that group work is *strongly encouraged* for going over this and all Problem Sets. For submission purposes, you may work *either* individually or in pairs.

`Assignment due Apr 26 (Fr), 12:20 PM`.
`Points Possible: 40`

- You have 2 weeks to complete this assignment. In accordance with our `course syllabus`, no late submissions will be accepted.
- Be honest. Don't cheat.
- As a Skidmore student, always recall your votes of academic integrity, and the **`Honor Code`** you have abided by:

> *"I hereby accept membership in the Skidmore College community and, with full realization of the responsibilities inherent in membership, do agree to adhere to honesty and integrity in all relationships, to be considerate of the rights of others, and to abide by the college regulations."*

Have fun!

Problem 1

For this first problem, we will work with **simulated data**. Hopefully, this exercise will provide some insights about the *structure* of ARMA models.

(a) The following code will generate an AR(1) model with $\varphi_1 = 0.6$:

```
set.seed(1903)     # a "seed" number to ensure reproducibility.

y ← numeric(100)  # the "y" series.
epsilon ← rnorm(n = 100, mean = 0, sd = 1)  # a random stochastic term,
                                        # normally distributed with mean 0
                                        # and standard deviation 1.

for(i in 2:100){
  y[i] ← 0.6*y[i-1] + epsilon[i]
  }

sim ← tsibble(date_index = seq_len(100),
               y = y, index = date_index)
```

After spending some time understanding the above simulation, produce a *time plot* of the simulated series you have created.

(b) Write your *own code* to create an AR(1) model for y_t with $\varphi_1 = -0.2$. Use the same "seed" number as in (a). *Hint*: Make sure not to use the name `sim` again for this new simulation.

(c) Produce a *time plot* for your simulation from part (b). What is (are) the difference(s) you notice from the other AR(1) model?

(d) Inspired by parts (a) and (b), write your own code to simulate the following MA(1) model: $y_t = \theta_1 \varepsilon_{t-1} + \varepsilon_t$ with $\theta_1 = 0.8$. Keep using the same "seed" number as in (a).

(e) Produce a *time plot* for your simulation from part (d).

Problem 2

In the same spirit as **Problem 1**, use a similar approach to simulate the following ARMA(2, 1) model:

$$y_t = 100 + 0.2y_{t-1} - 0.65y_{t-2} + 0.05\varepsilon_{t-1} + \varepsilon_t$$

where ε_t is a white noise series. Keep using the same "seed" number as in Problem 1. *Hint*: In your simulation code, make sure to set the "for" loop to `3:100` instead of `2:100`. Otherwise, you will get an error message. Then, answer the following questions:

(a) Plot the y series over time.

(b) Without running any tests, do you believe the y_t series is stationary? Explain your reasoning.

(c) Run a KPSS test on your simulated time series. What is your inference? Explain.

(d) Using the same "seed" number, simulate a *random walk* process. Do not include a drift. *Hint*: This time around, use `2:100` in your "for" loop.

(e) Graph your series from part (d), and test for stationarity. Did the result from the KPSS surprise you? Explain.

Problem 3

For this problem, you will work with the `chinese_gdp.csv` files (available on `theSpring`). It contains data on the Chinese economy's real Gross Domestic Product (GDP) from 1960 to 2017. The data come from the World Bank's economic indicators. After importing the data set and converting it into a time series tibble (`tsibble`) object, answer the following questions:

(a) Plot the data over time. What relevant *feature*(s) do you observe?

(b) Do you believe applying a *transformation* may be helpful for modeling and forecasting purposes? If so, plot your *transformed series* over time.

(c) Estimate the following exponential smoothing methods to your data: (*i*) ETS with *damped trend* and (*ii*) the automatic selection by the `{fable}` package. What are the differences between these two? Also, which model would you select, based on the *corrected Akaike Information Criterion* (AICc)? **Explain your answers**. *Hint*: In case you have applied transformations to your data, make sure to include such transformations in this step.

(d) From your part (c)'s selected model, plot (*i*) its *residuals* (if you've used a transformation, the correct column to select is the `.innov` one); (*ii*) its ACF plot; and (*iii*) histogram. In addition, run a Ljung-Box portmanteau test for autocorrelation using the appropriate number of lags. What is your inference from the test? Explain.

(e) Using 95% prediction intervals, plot a forecast for the future 10 years of the series. Make sure to have a clear graph.

Problem 4

The file `natural_gas.csv` (available on `theSpring`) contains data on U.S. annual natural gas consumption (in million cubic feet) for each state between 1997 and 2019. The data was retrieved from the `U.S. Energy Information Administration (EIA)`. After importing the data set into your R environment, answer the following questions:

(a) Convert your data object into a time series tibble (`tsibble`) object. Recall that you have different states here.

(b) Create a separate `tsibble` object containing only data for the state of New York. Plot NY's natural gas consumption over time.

(c) Produce the *ACF* and *PACF* plots of New York's natural gas consumption over time. Also, run a *KPSS* test on the series. What is your inference from the test?

(d) Based on your answer to the previous part, estimate the *ARIMA* model that minimizes the corrected Akaike Information criterion (AICc). What is the *ARIMA* model specification?

(e) Based on your model from part (d), run the appropriate Ljung-Box test for residual serial correlation. Do not forget to include the appropriate number of degrees-of-freedom. What is your inference from this test? Then, produce a 5-year ahead forecast for NY's natural gas consumption. Make sure to have a clear graph.