

Statistical Methods for Discrete Response, Time Series, and Panel Data (W271): Lab 2

Instructions (Please Read Carefully):

- **Due Date: Sunday March 15 2020, 11:59pm**
- No page limit, but be reasonable
- Do not modify fontsize, margin or line-spacing settings
- One student from each group should submit the lab to their student github repository by the deadline; submission and revisions made after the deadline will not be graded
- Answers should clearly explain your reasoning; do not simply ‘output dump’ the results of code without explanation
- Submit two files:
 1. A pdf file that details your answers. Include all R code used to produce the answers. Do not suppress the codes in your pdf file
 2. The R markdown (Rmd) file used to produce the pdf file

The assignment will not be graded unless **both** files are submitted

- Name your files to include all group members names. For example the students’ names are Stan Cartman and Kenny Kyle, name your files as follows:
 - StanCartman_KennyKyle_Lab2.Rmd
 - StanCartman_KennyKyle_Lab2.pdf
- Although it sounds obvious, please write your name on page 1 of your pdf and Rmd files
- All answers should include a detailed narrative; make sure that your audience can easily follow the logic of your analysis. All steps used in modelling must be clearly shown and explained
- If you use libraries and functions for statistical modeling that we have not covered in this course, you must provide an explanation of why such libraries and functions are used and reference the library documentation
- For mathematical formulae, type them in your R markdown file. Do not e.g. write them on a piece of paper, snap a photo, and use the image file
- Incorrectly following submission instructions results in deduction of grades
- Students are expected to act with regard to UC Berkeley Academic Integrity.

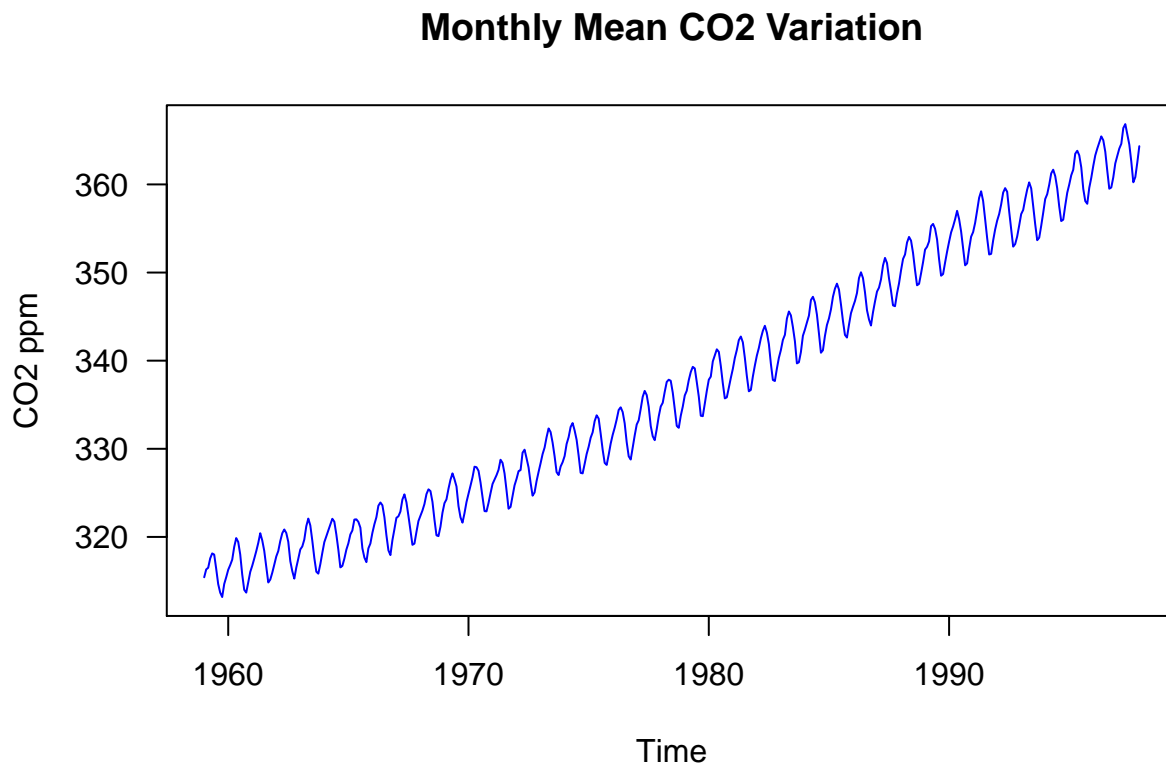
The Keeling Curve

In the 1950s, the geochemist Charles David Keeling observed a seasonal pattern in the amount of carbon dioxide present in air samples collected over the course of several years. He was able to attribute this pattern to the difference in the amount of land area and vegetation cover between the northern and southern hemispheres, and the resulting variation in global rates of photosynthesis as the hemispheres' seasons alternated throughout the year.

In 1958 Keeling began continuous monitoring of atmospheric carbon dioxide concentrations from the Mauna Loa Observatory in Hawaii and soon observed a trend increase carbon dioxide levels in addition to the seasonal cycle. He was able to attribute this trend increase to growth in global rates of fossil fuel combustion. This trend has continued to the present.

The `co2` data set in R's `datasets` package (automatically loaded with base R) is a monthly time series of atmospheric carbon dioxide concentrations measured in ppm (parts per million) at the Mauna Loa Observatory from 1959 to 1997. The curve graphed by this data is known as the 'Keeling Curve'.

```
plot(co2, ylab = expression("CO2 ppm"), col = 'blue', las = 1)
title(main = "Monthly Mean CO2 Variation")
```



Part 1 (4 points)

Conduct a comprehensive Exploratory Data Analysis on the `co2` series. This should include a thorough analysis of the trend, seasonal and irregular elements. Trends both in levels and growth rates should be discussed.

Part 2 (3 points)

Fit a linear time trend model to the `co2` series, and examine the characteristics of the residuals. Compare this to a quadratic time trend model. Discuss whether a logarithmic transformation of the data would be appropriate. Fit a suitable polynomial time trend model that incorporates seasonal dummy variables, and use this model to generate forecasts to the year 2020.

Part 3 (3 points)

Following all appropriate steps, choose an ARIMA model to fit to the series. Discuss the characteristics of your model and how you selected between alternative ARIMA specifications. Write your model (or models) using backshift notation. Use your model (or models) to generate forecasts to the year 2020.

Part 4 (4 points)

The file `co2_weekly_mlo.txt` contains weekly observations of atmospheric carbon dioxide concentrations measured at the Mauna Loa Observatory from 1974 to 2020, published by the National Oceanic and Atmospheric Administration (NOAA). Convert these data into a suitable time series object, conduct a thorough EDA on the data, and address the problem of missing observations. Describe how the Keeling Curve evolved from 1997 to the present and compare current atmospheric CO₂ levels to those predicted by your forecasts in Parts 2 and 3. Use the weekly data to generate a month-average series from 1997 to the present, and compare the overall forecasting performance of your models from Parts 2 and 3 over the entire period.

Part 5 (3 points)

Seasonally adjust the weekly NOAA data, and split both seasonally-adjusted (SA) and non-seasonally-adjusted (NSA) series into training and test sets, using the last two years of observations as the test sets. For both SA and NSA series, fit ARIMA models using all appropriate steps. Measure and discuss how your models perform in-sample and (psuedo-) out-of-sample, comparing candidate models and explaining your choice. In addition, fit a polynomial time-trend model to the seasonally-adjusted series and compare its performance to that of your ARIMA model.

Part 6 (3 points)

Generate predictions for when atmospheric CO₂ is expected to be at 420 ppm and 500 ppm levels for the first and final times (consider prediction intervals as well as point estimates in your answer). Generate a prediction for atmospheric CO₂ levels in the year 2100. How confident are you that these are accurate predictions?