Homework 2: Probability Models

BEE 4850/5850, Fall 2024

Due Date

Friday, 2/23/24, 9:00pm

Tip

To do this assignment in Julia, you can find a Jupyter notebook with an appropriate environment in the homework's Github repository. Otherwise, you will be responsible for setting up an appropriate package environment in the language of your choosing. Make sure to include your name and NetID on your solution.

Overview

Instructions

The goal of this homework assignment is to practice developing and working with probability models for data.

- Problem 1 asks you to fit a sea-level rise model using normal residuals and to assess the validity of that assumption.
- Problem 2 asks you to model the time series of hourly weather-related variability at a tide gauge.
- Problem 3 asks you to model the occurrences of Cayuga Lake freezing, and is only slightly adapted from Example 4.1 in Statistical Methods in the Atmospheric Sciences by Daniel Wilks.
- Problem 4 (**graded only for graduate students**) asks you to revisit the sea-level model in Problem 1 by including a model-data discrepancy term in the model calibration.

Learning Outcomes

After completing this assignments, students will be able to:

- develop probability models for data and model residuals under a variety of statistical assumptions;
- evaluate the appropriateness of those assumptions through the use of qualitative and quantitative evaluations of goodness-of-fit;
- fit a basic Bayesian model to data.

Load Environment

The following code loads the environment and makes sure all needed packages are installed. This should be at the start of most Julia scripts.

```julia
import Pkg
Pkg.activate(@__DIR__)
Pkg.instantiate()
```

The following packages are included in the environment (to help you find other similar packages in other languages). The code below loads these packages for use in the subsequent notebook (the desired functionality for each package is commented next to the package).

```julia
using Random # random number generation and seed-setting
using DataFrames # tabular data structure
using CSVFiles # reads/writes .csv files
using Distributions # interface to work with probability distributions
using Plots # plotting library
using StatsBase # statistical quantities like mean, median, etc
using StatsPlots # some additional statistical plotting tools
using Optim # optimization tools
```

Problems (Total: 30 Points for 4850; 40 for 5850)

Problem 1

Consider the following sea-level rise model from Rahmstorf (2007):

$$\frac{dH(t)}{dt} = \alpha(T(t) - T_0),$$

where T_0 is the temperature (in $^\circ C$) where sea-level is in equilibrium ($dH/dt = 0$), and α is the sea-level rise sensitivity to temperature. Discretizing this equation using the Euler method and using an annual timestep ($\delta t = 1$), we get

$$H(t+1) = H(t) + \alpha(T(t) - T_0).$$

In this problem:

- Load the data from the `data/` folder
  - Global mean temperature data from the HadCRUT 5.0.2.0 dataset (https://hadobs.metoffice.gov.uk/hadcrut5/data/HadCRUT.5.0.2.0/download.html) can be found in `data/HadCRUT.5.0.2.0.analysis.summary_series.global.annual.csv`. This data is averaged over the Northern and Southern Hemispheres and over the whole year.
  - Global mean sea level anomalies (relative to the 1990 mean global sea level) are in `data/CSIRO_Recons_gmsl_yr_2015.csv`, courtesy of CSIRO (https://www.cmar.csiro.au/sealevel/sl_data_cmar.html).
- Fit the model under the assumption of normal i.i.d. residuals by maximizing the likelihood and report the parameter estimates. Note that you will need another parameter H_0 for the initial sea level. What can you conclude about the relationship between global mean temperature increases and global mean sea level rise?
- How appropriate was the normal i.i.d. probability model for the residuals? Use any needed quantitative or qualitative assessments of goodness of fit to justify your answer. If this was not an appropriate probability model, what would you change?

Problem 2

Tide gauge data is complicated to analyze because it is influenced by different harmonic processes (such as the linear cycle). In this problem, we will develop a model for this data using NOAA data from the Sewell's Point tide gauge outside of Norfolk, VA from `data/norfolk-hourly-surge-2015.csv`. This is hourly data (in m) from 2015 and includes both the observed data (`Verified (m)`) and the tide level predicted by NOAA's sinusoidal model for periodic variability, such as tides and other seasonal cycles (`Predicted (m)`).

In this problem: * Load the data file. Take the difference between the observations and the sinusoidal predictions to obtain the tide level which could be attributed to weather-related variability (since for one year sea-level rise and other factors are unlikely to matter). Plot this data. * Develop an autoregressive model for the weather-related variability in the Norfolk tide gauge. Make sure to include your logic or exploratory analysis used in determining the model specification. * Use your model to simulate 1,000 realizations of hourly tide gauge observations. What is the distribution of the maximum tide level? How does this compare to the observed value?

Problem 3

As of 2010, Cayuga Lake has frozen in the following years: 1796, 1816, 1856, 1875, 1884, 1904, 1912, 1934, 1961, and 1979. Based on this data, we would like to project whether Cayuga Lake is likely to freeze again in the next 25 years.

In this problem:

- Assuming that observations began in 1780, write down a Bayesian model for whether Cayuga Lake will freeze in a given year, using a Bernoulli distribution. How did you select what prior to use?
- Find the maximum *a posteriori* estimate using your model.
- Generate 1,000 realizations of Cayuga Lake freezing occurrences from 1780 to 2010 and check the simulations against the occurrance data.
- Using your model, calculate the probability of Cayuga Lake freezing at least once in the next 10 years.
- What do you think about the validity of your model, both in terms of its ability to reproduce historical data and its use to make future projections? Why might you believe or discount it? What changes might you make (include thoughts about the prior)?

Problem 4

GRADED FOR 5850 STUDENTS ONLY

For the sea-level model in Problem 1, model the model-data discrepancy using an AR(1) process, with observation error modeled as normally distributed with standard deviation given by the uncertainty column in the data file.

In this problem:

- Find the maximum likelihood estimate of the parameters with this discrepancy structure. How does the parameter inference change from the normal i.i.d. estimate in Problem 1?
- Generate 1,000 traces, plot a comparison of the hindcasts to those from Problem 1, and compare the surprise indices.
- Determine whether you have accounted for autocorrelation in the residuals appropriately (hint: generate realizations of just the discrepancy series, compute the resulting residuals from the model fit + discrepancy, and look at the distribution of autocorrelation values).
- Which model specification would you prefer and why?

:::