# Homework 2 Solutions

**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.

```
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).

```
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

Random.seed!(1)
```

## 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 $\alpha$ 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?

***Solution***:

First, let's load the data and implement the model with a function.

```
# load data files
slr_data = DataFrame(load("data/CSIRO_Recons_gmsl_yr_2015.csv"))
gmt_data =
 ↪  DataFrame(load("data/HadCRUT.5.0.2.0.analysis.summary_series.global.annual.csv"))
slr_data[:, :Time] = slr_data[:, :Time] .- 0.5; # remove 0.5 from Times
dat = leftjoin(slr_data, gmt_data, on="Time") # join data frames on time
select!(dat, [1, 2, 3, 4])  # drop columns we don't need
first(dat, 6)
```

| | Time | GMSL (mm) | GMSL uncertainty (mm) | Anomaly (deg C) |
|---|---|---|---|---|
| | Float64 | Float64 | Float64 | Float64? |
| 1 | 1880.0 | -158.7 | 24.2 | -0.315832 |
| 2 | 1881.0 | -153.1 | 24.2 | -0.232246 |
| 3 | 1882.0 | -169.9 | 23.0 | -0.29553 |
| 4 | 1883.0 | -164.6 | 22.8 | -0.346474 |
| 5 | 1884.0 | -143.7 | 22.2 | -0.49232 |
| 6 | 1885.0 | -145.2 | 21.9 | -0.471124 |

```
# slr_model: function to simulate sea-level rise from global mean temperature
 ↪  based on the Rahmstorf (2007) model

function slr_model( , T , H , temp_data)
    temp_effect =   .* (temp_data .- T )
    slr_predict = cumsum(temp_effect) .+ H
    return slr_predict
end
```

```
slr_model (generic function with 1 method)
```

Now, let's fit the model under the normal residual assumption.

```
# split data structure into individual pieces
years = dat[:, 1]
sealevels = dat[:, 2]
temp = dat[:, 4]

# write function to calculate likelihood of residuals for given parameters
# parameters are a vector [ , T , H ,  ]
function llik_normal(params, temp_data, slr_data)
    slr_out = slr_model(params[1], params[2], params[3], temp_data)
    resids = slr_out - slr_data
    return sum(logpdf.(Normal(0, params[4]), resids))
end

# set up lower and upper bounds for the parameters for the optimization
lbds = [0.0, -50.0, -200.0, 0.0]
ubds = [10.0, 1.0, 0.0, 20.0]
p0 = [5.0, -1.0, -100.0, 5.0]
p_mle = Optim.optimize(p -> -llik_normal(p, temp, sealevels), lbds, ubds,
 ↪  p0).minimizer
```