Homework 3: Bayesian and Extreme Value Statistics

BEE 4850/5850, Fall 2024

Due Date

Friday, 3/22/24, 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.

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 DataFramesMeta # API which can simplify chains of DataFrames
 ↪  transformations
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 Rahmstorf (2007) sea-level rise model from Homework 2:

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

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

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

Suppose that we wanted to develop a Bayesian probability model for this problem, assuming independent normal residuals:

$$y(t) = F(t) + \varepsilon_t$$
$$\varepsilon_t \sim \mathcal{N}(0, \sigma^2)$$

We might specify the following priors (assuming independence across parameters):

- $T_0 \sim \mathcal{N}(-0.5, 0.1)$;
- $\alpha \sim \mathcal{TN}(0, 5; 0, \infty)$ (truncated normal between 0 and infinity);
- $H_0 \sim \mathcal{N}(-150, 25)$;
- $\sigma \sim \mathcal{TN}(0, 5; 0, \infty)$

In this problem:

- Historical and RCP 8.5 global mean temperatures from NOAA can be found in `data/NOAA_IPCC_RCPtempsscenarios.csv` (use the fourth column for the temperature series).
- 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).
- Simulate from the prior predictive distribution. What do you think about the priors?
- Would you propose new priors? If so, what might they be and why?

Problem 2

Following from Problem 1, compare the maximum likelihood and maximum *a posteriori* estimates for the model.

In this problem:

- Find the MLE and MAP parameter values using the prior distributions given in Problem 1.
- Plot the median and 95% credible intervals for the hindcasts and the projections until 2100 under RCP 8.5 (using `data/NOAA_IPCC_RCPtempsscenarios.csv`; make sure T_0 and H_0 have the same meaning as in Problem 1!).
- What differences do you observe? What do you attribute these differences to? What conclusions can you draw about the Bayesian model?

Problem 3

Let's look at how (modeled) daily maximum temperatures have (or have not) increased in Ithaca from 1850–2014. Model output from NOAA's GFDL-ESM4 climate model (one of the models used in the latest Climate Model Intercomparison Project, CMIP6) is available in `data/gfdl-esm4-tempmax-ithaca.csv`. While this model output has not been bias-corrected, we won't worry about that for the purposes of this assignment.

In this problem:

- Load and plot the temperature maxima data from `data/gfdl-esm4-tempmax-ithaca.csv`.
- Suppose that we were interested in looking at temperature exceedances over 28°C. Decluster these occurrences and plot the number of exceedances by year. Have they increased over time?
- Fit a stationary GPD model for the exceedances. What does this distribution look like?

Problem 4

GRADED FOR 5850 STUDENTS ONLY

In class, we modeled the annual maxima of the San Francisco tide gauge data using a stationary GEV distribution. We could also hypothesize that the tide extremes are influenced by the Pacific Decadal Oscillation (PDO), which is a climate pattern related to the sea-surface temperatures in the Pacific Ocean (similar to the El Niño-Southern Oscillation), in the following fashion (where only the GEV location is variable):

$$y_t \sim \text{GEV}(\mu_0 + \mu_1 p_t, \sigma, \xi)$$

In this problem:

- Load the San Francisco tide gauge data (`data/h551.csv`) and the PDO index dataset (`data/errst.v5.pdo.dat`; this file is a space-delimited file, versus the comma-delimited `.csv` files, which can be loaded in Julia with `CSV.read(data/errst.v5.pdo.dat, DataFrame; delim=" ", header=2, ignorerepeated=true)`). The PDO data is given as monthly values; convert these to yearly indices by taking the mean. You should also drop 2023 due to the incomplete record. You can use the function at the bottom of these instructions to load the data, or adapt accordingly to a different language.
- Find the MLE of the non-stationary GEV model and for a stationary GEV (constant μ; we did this in class).
- Discuss the difference(s) between the two fitted models based on the coefficient values (you can also bring to bear the range(s) of PDO values from the data), the 100- and 500-year return periods in 2022, and plotted hindcasts.

```
## load the data from the file and return a DataFrame of DateTime values and
 ↪  gauge measurements

function load_pdo(fname)
    # This uses the DataFramesMeta.jl package, which makes it easy to string
     ↪  together commands to load and process data
    df = DataFrame(CSVFiles.load(File(format"CSV", fname), spacedelim=true,
 ↪  skiplines_begin=1))
    # take yearly average
    @transform!(df, :PDO = mean(AsTable(names(df)[2:13])))
    @select!(df, $[:Year, :PDO])
    @rsubset!(df, :Year != 2023)
    return df
end

pdo = load_pdo("data/ersst.v5.pdo.dat")
# subset for years that match the tide gauge data
years = pdo[!, :Year]
@rsubset!(pdo, :Year in years)
```

	Year	PDO
	Int64	Float64
1	1854	-0.818333
2	1855	-0.743333
3	1856	-0.2075
4	1857	-0.0266667
5	1858	0.196667
6	1859	-1.74333
7	1860	-0.954167
8	1861	-0.955
9	1862	-0.226667
10	1863	-0.095
11	1864	-0.281667
12	1865	-0.553333
13	1866	-0.9425
14	1867	-0.225
15	1868	-0.1025
16	1869	0.944167
17	1870	-0.319167
18	1871	-0.455
19	1872	-0.7425
20	1873	-1.125
21	1874	-0.264167
22	1875	-0.905833
23	1876	-1.08833
24	1877	-0.304167
25	1878	0.236667
26	1879	-0.385833
27	1880	-1.36083
28	1881	-0.266667
29	1882	-1.36917
30	1883	-1.4225
...	...	...