Homework 4: The Bootstrap

BEE 4850/5850, Fall 2025

Due Date

Friday, 3/28/25, 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 simulation-based uncertainty quantification, focusing on the bootstrap.

- Problem 1 asks you to use the non-parametric bootstrap to estimate uncertainty in a Poisson regression model.
- Problem 2 asks you to use the bootstrap (through resampling residuals) to estimate uncertainty in a semi-empirical sea-level rise model.
- Problem 3 (only required for students in BEE 5850) asks you to use a moving block bootstrap to estimate the sampling distribution of the median of extreme water level 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 CSV # 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
```

Problems

Scoring

- Problem 1 is worth 10 points;
- Problem 2 is worth 10 points;
- Problem 3 is worth 5 points;

Problem 1

Revisit the salamander model from Homework 2, using percent groundcover as a predictor in the Poisson regression. Use the non-parametric bootstrap to estimate bias and confidence intervals for the model parameters.

In this problem:

- Load the data from `data/salamanders.csv`.
- Fit a Poisson regression model for salamander counts using the percentage of ground cover.
- Use 1,000 non-parametric bootstrap samples to obtain estimates of bias and the 90% confidence interval for the intercept and coefficient in the Poisson regression.

Problem 2

Revisit the sea-level rise model from Homework 2:

$$\frac{dS}{dt} = \frac{S_{\text{eq}} - S}{\tau}$$
$$S_{\text{eq}} = aT + b,$$

where

- $S(t)$ is the global mean sea level (in mm) at time t;
- τ is the response time of sea level (in yrs);
- S_{eq} is the equilibrium sea-level (in mm) at temperature T (in °C);
- a is the sensitivity of S_{eq} to T (in mm/°C);
- b is the intercept of S_{eq}, or the S_{eq} when $T = 0$°C (in mm).

We would like to quantify uncertainty in the model parameters using the bootstrap.

In this problem:

- Load the data from the `data/` folder and, following Grinsted et al (2010), normalize both datasets to the 1980-1999 mean (subtract that mean from the data).
  - 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). The standard deviation of the estimate is also added for each year.
- Write a function to simulate global mean sea levels under a set of model parameters after discretizing the equations above with a timestep of $\delta t = 1$ yr. You will need to subset the temperature data to the years where you also have sea-level data and include an initial sea-level parameter S_0. This will be similar to the model from Homework 2. Fit this model to the data with AR(1) residuals.
- Use your fitted model and the AR(1) residual process to generate 1,000 parametric bootstrap samples. Refit the model to each. Plot histograms of the bootstrap samples for each parameter. What is the 90% confidence interval for the sensitivity of sea level to global mean temperature?

Problem 3

Let's revisit the 2015 Sewell's Point tide gauge data, which consists of hourly observations and predicted sea-level based on NOAA's harmonic model.

```
function load_data(fname)
    date_format = "yyyy-mm-dd HH:MM"
    # this uses the DataFramesMeta package -- it's pretty cool
    return @chain fname begin
        CSV.File(; dateformat=date_format)
        DataFrame
        rename(
            "Time (GMT)" => "time", "Predicted (m)" => "harmonic", "Verified
            ↪  (m)" => "gauge"
        )
        @transform :datetime = (Date.(:Date, "yyyy/mm/dd") + Time.(:time))
        select(:datetime, :gauge, :harmonic)
        @transform :weather = :gauge - :harmonic
        @transform :month = (month.(:datetime))
    end
end

dat = load_data("data/norfolk-hourly-surge-2015.csv")

plot(dat.datetime, dat.gauge; ylabel="Gauge Measurement (m)",
 ↪  label="Observed", legend=:topleft, xlabel="Date/Time", color=:blue)
plot!(dat.datetime, dat.harmonic, label="Prediction", color=:orange)
```

```
UndefVarError: UndefVarError(:Date)
UndefVarError: `Date` not defined
Stacktrace:
  [1] (::var"#1#2")(233::PooledArrays.PooledVector{String15, UInt32,
 ↪  Vector{UInt32}}, 234::PooledArrays.PooledVector{String7, UInt32,
 ↪  Vector{UInt32}})
    @ Main.Notebook ~/.julia/packages/DataFramesMeta/1Y7m8/src/parsing.jl:303
  [2] _transformation_helper(df::DataFrame, col_idx::Vector{Int64},
 ↪  ::Base.RefValue{Any})
    @ DataFrames
 ↪  ~/.julia/packages/DataFrames/kcA9R/src/abstractdataframe/selection.jl:605
  [3] select_transform!(::Base.RefValue{Any}, df::DataFrame,
 ↪  newdf::DataFrame, transformed_cols::Set{Symbol}, copycols::Bool,
 ↪  allow_resizing_newdf::Base.RefValue{Bool}, column_to_copy::BitVector)
```

```
    @ DataFrames
↪   ~/.julia/packages/DataFrames/kcA9R/src/abstractdataframe/selection.jl:805
  [4] _manipulate(df::DataFrame, normalized_cs::Vector{Any}, copycols::Bool,
↪   keeprows::Bool)
    @ DataFrames
↪   ~/.julia/packages/DataFrames/kcA9R/src/abstractdataframe/selection.jl:1783
  [5] manipulate(::DataFrame, ::Any, ::Vararg{Any}; copycols::Bool,
↪   keeprows::Bool, renamecols::Bool)
    @ DataFrames
↪   ~/.julia/packages/DataFrames/kcA9R/src/abstractdataframe/selection.jl:1703
  [6] manipulate
    @
↪   ~/.julia/packages/DataFrames/kcA9R/src/abstractdataframe/selection.jl:1693
↪   [inlined]
  [7] select
    @
↪   ~/.julia/packages/DataFrames/kcA9R/src/abstractdataframe/selection.jl:1303
↪   [inlined]
  [8] transform
    @
↪   ~/.julia/packages/DataFrames/kcA9R/src/abstractdataframe/selection.jl:1383
↪   [inlined]
  [9] macro expansion
    @ ~/.julia/packages/DataFramesMeta/1Y7m8/src/macros.jl:1465 [inlined]
 [10] load_data(fname::String)
    @ Main.Notebook ~/Teaching/BEE4850/sp24/assignments/hw04/hw04.qmd:162
 [11] top-level scope
    @ ~/Teaching/BEE4850/sp24/assignments/hw04/hw04.qmd:169
```

We detrend the data to isolate the weather-induced variability by subtracting the predictions from the observations; the results (following the Julia code) are in `dat[:, :weather]`.

```
plot(dat.datetime, dat.weather; ylabel="Gauge Weather Variability (m)",
↪   label="Detrended Data", linewidth=1, legend=:topleft, xlabel="Date/Time")
```

```
UndefVarError: UndefVarError(:dat)
UndefVarError: `dat` not defined
Stacktrace:
 [1] top-level scope
   @ ~/Teaching/BEE4850/sp24/assignments/hw04/hw04.qmd:181
```

We would like to understand the uncertainty in an estimate of the median sea level.

In this problem:

- Construct 1,000 bootstrap replicates by adding a moving block bootstrap replicate from the weather-induced variability series (with block length 20) to the harmonic prediction. Use these replicates to compute a 90% confidence interval. What is the bias of the estimator?
- Repeat the analysis with block length 50. How does this affect the confidence intervals and estimate of bias?
- Why do you think using different block lengths produced the results that they did?