Homework 2: Probability Models

BEE 4850/5850

Due Date

Friday, 2/21/25, 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 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 using an autoregressive model.
- Problem 3 asks you to use Poisson regression to predict salamander counts based on environmental data.
- Problem 4 (only required for students in BEE 5850) asks you to look at the impact of the gender of hurricane names on deaths[1].

[1] Yes, seriously. Ish. Trust me, I know.

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 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
using Optim # optimization tools
```

Problems

Scoring

- Problem 1 is worth 7 points.
- Problem 2 is worth 6 points.
- Problem 3 is worth 7 points.
- Problem 4 is worth 5 points.

Problem 1

Consider the following sea-level rise model from Grinsted et al (2010), which models sea-level rise based on a linear relationship between global mean temperature and an "equilibrium" sea level:

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

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).

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.
- Fit the model under the assumption of Gaussian i.i.d. residuals (include both an uncertain model error term and the standard deviation of the observations in the probability model specification) by maximizing the likelihood. Report the parameter estimates. Note that you will need another parameter S_0 for the initial sea level. What can you conclude about the relationship between global mean temperature increases and global mean sea level rise rates?
- How appropriate was the Gaussian 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 (AR) 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 by adding simulations from your AR model back to the predicted sinusoidal trend. What is the distribution of the maximum tide level? How does this compare to the observed value?

Problem 3

The file `data/salamanders.csv` contains counts of salamanders from 47 different plots of the same area in California, as well as the percentage of ground cover and age of the forest in the plot. You would like to see if you can use these data to predict the salamander counts with a Poisson regression.

In this problem:

- Load the data. You may need to standardize the predictors as they are much larger than the counts.
- Fit a Poisson regression model for salamander counts using the percentage of ground cover.
- Plot the expected counts and 90% prediction intervals from your model. How well does the model predict the observed counts? In what ways does it do a good or bad job?
- Can you improve the model by including forest age? Why do you think this helps or does not help with prediction?

Problem 4

GRADED FOR 5850 STUDENTS ONLY

In 2014, a paper was published in a prestigious journal which claimed that hurricanes with more feminine names are deadlier than hurricanes with more masculine names because people take warnings about female-named hurricanes less seriously[2]. The file `data/Hurricanes.csv` contains the original data used in this analysis. While we won't replicate the specific analysis in this paper, let's use the data to look at this hypothesis.

In this problem:

[2]This paper has become a bit of a joke among statisticians, but let's take the hypothesis seriously for this problem's sake.

- One might interpret the hypothesis to claim that the impact of the name is strengthened by the the more powerful. A measure of hurricane strength is its minimum pressure (`min_pressure` in the dataset). Fit a model that predicts deaths ('deaths') using the femininity of the name (`femininity`) and minimum pressure (you may need to standardize the pressure).
- Interpret the results by generating counterfactual simulations for hurricanes with the most feminine and masculine name scores. Plot the expected values and 90% prediction intervals from these two sets of simulations and compare with the observed storm deaths. Where does the model do well or not well? Does the effect size of the gender of the name seem plausible?
- Conclude with a summary of your conclusions about the impact of the gender of a hurricane's name on deaths. How might you change the approach in this problem to keep exploring this hypothesis, if at all?

References