Homework 2: Probability Models

BEE 4850/5850

Due Date

Friday, 2/20/26, 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 some linear regressions to Chicago mortality data.
- Problem 2 asks you to analyze the probability of Cayuga Lake freezing using a logistic regression model.
- Problem 3 asks you to use Poisson regression to predict salamander counts based on environmental data.
- Problem 4 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
using LaTeXStrings # latex formatting for plot strings
```

Problems

Problem 1 (6 points)

Let's revisit the `chicago` dataset from HW1 (found in `data/chicago.csv`). We will look at the relationship of the potential predictor variables `pm25median` (the median density of smaller pollutant particles), `o3median` (the median concentration of O_3), `so2median` (the median concentration of SO_2), and `tmpd` (mean daily temperature) with the variable we would like to predict, `death` (the number of non-accidental deaths on that day).

Problem 1.1

Plot each predictor variable against the response variable (make sure to include labels and titles for each plot). Describe what you observe about the bivariate relationships, if one exists (*e.g.* does it look like they exist, are they linear, positive/negative, etc.) Does a linear regression seem appropriate from any of these plots? Which predictor do you think is most appropriate for a linear regression model and why?

Problem 1.2

You decide to focus on the relationship between `tmpd` and `death`. Your classmate suggests that you use a Gaussian-error linear regression model. Based on this suggestion, calculate the regression coefficients and error variance. Explain the interpretation of the coefficients that you obtain **in this specific problem context**.

Problem 1.3

Add your fitted regression line and a 90% prediction interval to your scatterplot of `tmpd` and `death`. Using this and any other diagnostics, do you agree with the Gaussian-error linear model assumption? Why or why not?

Problem 2 (6 points)

This problem is adapted from Example 4.1 and 4.2 in Daniel Wilks' *Statistical Methods in the Atmospheric Sciences*. Updating the table in that book, Cayuga Lake has frozen in 1796, 1816, 1856, 1875, 1884, 1904, 1912, 1934, 1961, 1979, and 2015[2]. The probability of whether a severe enough cold spell to freeze the lake occurs is related to changes in global mean temperature. The file `data/HadCRUT5.1Analysis_gl.txt` contains estimates of monthly global mean temperature anomalies (in ° C, relative to the 1961–1990 mean) as well as the annual global temperature anomaly. You would like to build a model to predict the probability that Cayuga Lake freezes based on the winter (DJF) temperature anomaly.

Problem 2.1

Write down a logistic regression model for the probability that Cayuga Lake freezes in a given winter. Encode the occurrence of freezing as a 1, and non-freezing as 0. Load the temperature data and fit this model. The format of the data file is a little awkward, as the even rows are the number of stations (which should be ignored) and have uneven number of space delimiters. The first 13 columns are the year and the monthly means; the final column is the annual average (which we don't want to use). Another complication is that the (even-numbered) station rows don't have an entry in the 14th column. In Julia, you can load the file with

```
temp_dat = CSV.read("data/HadCRUT5.1Analysis_gl.txt", delim=" ",
    ignorerepeated=true, header=false,
    silencewarnings=true, DataFrame)
```

[2]Many of these dates, particularly prior to the 20th century, are subject to some debate, and would be different depending on the data source. Nevertheless, let's use this dataset.

To simplify the analysis, since we don't always know whether the lake froze in the January starting a year or the December ending a year, let's assume that the years in which the lake froze correspond to the winter starting the year (so the DJF mean corresponding to the freezing in 1884 would include the December from 1883 and the January and February from 1884). You will only be able to use freezing events between 1851 and the present day due to the temperature data. Interpret the coefficients from your model fit **in this specific problem context**.

Problem 2.2

Plot the modeled probability of freezing as a function of the DFJ temperature anomaly, and how this probability has changed over time. Does this model seem reasonable? How might you determine this?

Problem 2.3

What winter temperature anomaly would be required for less than a 1% probability of Cayuga Lake freezing?

Problem 3 (6 points)

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.

Problem 3.1

Write down and fit a Poisson regression model for salamander counts using the percentage of ground cover as a predictor. You may need to standardize the predictors as they are much larger than the counts, which you can do using the following function (or its equivalent):

```
stdz(x) = (x .- mean(x)) / std(x)                                    ①
```

① This function makes it convenient to standardize when we fit the models, rather than changing the data itself.

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

Problem 3.2

Plot the expected counts and 90% prediction intervals from your model (based on 1,000 simulations from the model) along with the data. How well does the model predict the observed counts? In what ways does it do a good or bad job?

Problem 3.3

Can you improve the model by including forest age as a predictor? How did you determine whether this helps or does not help with prediction?

Problem 4 (7 points)

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[3]. The file `data/Hurricanes.csv` contains the original data used in this analysis (note that this dataset excludes Hurricanes Katrina and Audrey because they were "outliers"). While we won't replicate the specific analysis in this paper, let's use the data to look at this hypothesis.

Problem 4.1

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 Poisson regression model that predicts deaths (`deaths`) using the gender of the name (`female`) and minimum pressure (you may need to standardize the pressure). We use `female` instead of the continuous measure of the `femininity` of the name as this is highly subjective and has some weird coding (such as "Sandy" being coded as one of the most female names).

Problem 4.2

Interpret the results by generating 10,000 counterfactual simulations for hurricanes with male and female names. 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?

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

Problem 4.3

Does the gender effect size resulting from this model seem reasonable to you? Draw on qualitative or quantitative assessments to justify your assessment of reasonableness.

Problem 4.4

We can now stop taking this hypothesis seriously. To what extent do you think the finding could be an artifact of the dataset (e.g. there is no actual effect, but there are coincidental features of the data that produce the result that female-named storms are more deadly than male-named storms)? Justify this conclusion with specific reference to an exploratory analysis of the data.

References