Homework 1
Yanran Li (yl5465)
October 7, 2023

1 Problem 1

1.1 Introduction

Heat warnings are issued in advance of forecast extreme heat events, yet little evidence is available regarding their effectiveness in reducing heat-related illness and death[1]. In this homework 1, I implemented a stochastic MAP estimation model (Bayesian linear regression) on a real-world data set (heat warning related) to estimate the association between causes (heat index etc.) and heat alerts issued by the United States National Weather Service in 2,817 counties, 2006–2016.

1.2 Methods

1.2.1 Data

- **Description of the dataset:** I used the daily time series data[2], acquired during the warm months (April-October) of 2006-2016 for 2837 U.S. counties. For each county, we obtained 1) daily maximum heat index (an index that combines air temperature and relative humidity to posit a human-perceived equivalent temperature); 2) daily issuance of heat alerts (binary).

- **Description of the response variable:** We obtained text files containing records of all non-precipitation alerts issued by the US NWS during the warm months of 2006 –2016 from the National Oceanic and Atmospheric Administration. Our dataset is a daily time series containing a binary variable for heat alert exposure for each county.

- **Description of the features:**

  Heat Index: We obtained gridded (4-km resolution) estimates of daily maximum temperature and vapor-pressure deficit for the contiguous US from the Parameter-elevation Regressions on Independent Slopes Model. From these variables, we generated time series of population-weighted daily maximum heat index for each county[1]. To visualize the "Heat Index" varaible over time, I plotted 3 states' (Arizona, California, and New York) data over time (2006-2016), where we can see their general trends during this time (Fig 1).

  Weekday: Since we already have the information of "Date", we use the `wday()` function from R package `lubridate` to extract the weekday information of each day. And we denote weekdays "1" and weekends "0".

  Month: Similarly, we use the `month()` function from `lubridate` to extract the month information of each day. We utilized this feature as characters in our model.

  Longitude/Latitude: Since our heat alert data already has the FIP code information, I expanded the data frame with each counties' longitude and latitude. This data comes from

the 2013 Cartographic Boundary Shapefiles (https://www.census.gov/geographies/mapping-files/2013/geo/carto-boundary-file.html).

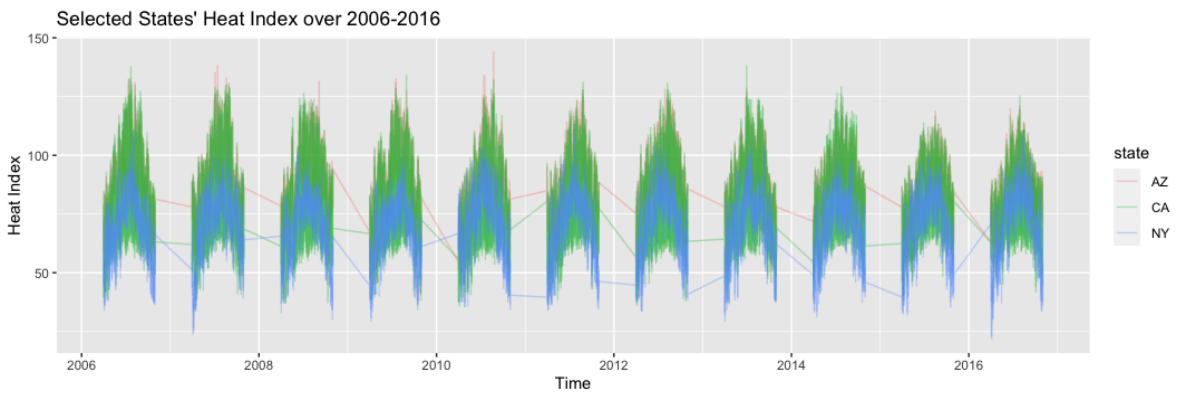

Figure 1: Selected States' Heat Index over 2006-2016

1.2.2 Model

Our goal is to predict the binary heat alert implementation (1/0) from all other features we choose. Since the response is a binary variable, it makes sense to use a Bayesian logistic regression. To compute the posterior coefficients of our Bayesian logistic regression, we examine the log of the posterior as a function of the coefficients:

$$\log p(\beta|x,y) = -\frac{1}{2\lambda^2}\sum_{k=1}^{p}\beta_k^2 + \sum_{i=1}^{n}(y_i \log \sigma(\beta \cdot x_i) + (1-y_i)\log \sigma(-\beta \cdot x_i)) + const$$

I used Gaussian prior to shrink the coefficients towards zero. Before training the model, I standardized all features and split the data into train/valid/test sets with a 80%/10%/10% proportion.

1.3 Discussion

- **The influence of the prior on the model and how to select the hyperparameters**

  The coefficients (β_k) of the logistic regression have Gaussian prior. I used λ to determine the variance of the Gaussian prior on the coefficients. λ value means that the standard deviation is inversely proportional to the value of λ, making the prior tighter around zero. The penalty term represents the log of the Gaussian prior, which ensures that the weights don't stray too far from zero. To select the value of the priors, I tried different value of λ ($\lambda = 0.5, \frac{1}{\sqrt{2}}, 1, \sqrt{2}, 5, 10$) and plot the influence on priors when epoch increases on Figure 2.

- **Interpretation of the posterior coefficients**

  Referring to Figure 2, we observe that a prior with $\lambda = 1$ yields a sufficiently high log likelihood while maintaining a relatively low penalty value. Consequently, we select $\lambda = 1$

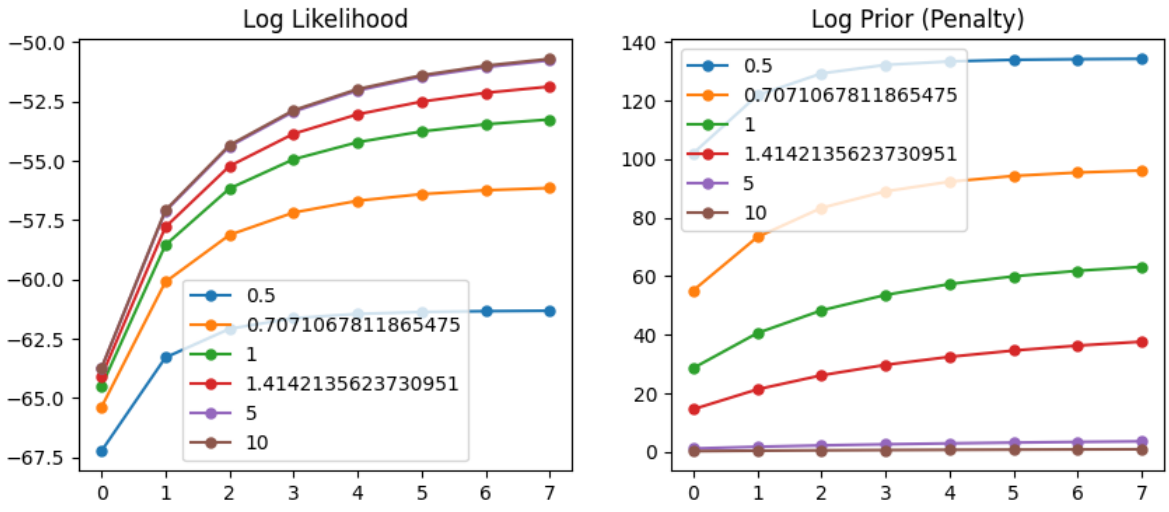

Figure 2: Prior's influence on the Log Likelihood and Penalty term as training epoch increases

for subsequent experiments. With this choice of λ, the derived posterior results are shown in Table 1. The coefficient for Longitude is 0.040602 and the coefficient for Latitude is 0.563841. This suggests that latitude has a stronger influence on the response than longitude. For 'HImaxFPopW', this variable has a coefficient of 2.957884, which is notably higher than most other coefficients in the table. It suggests that this variable has a significant positive influence on the response. For every unit increase in 'HImaxFPopW', the log-odds of the response variable increases by roughly 2.957884. Weekdays and weekends do not have too much different impact on their influences. For the months, July, August and October shows the strongest impact.

- **Influence of batch size and learning rate on the optimization**

  Referring to Figure 3, within the span of 8 training epochs, a larger batch size tends to enhance the log likelihood as well as the penalty. Employing an overly small batch size appears to be ineffective. From Figure 4, an excessively large learning rate is counterproductive.

2 Problem 2

2.1 Variables in the data

I plan to utilize the data-set mentioned in Problem 1, supplementing it with Medicare data, which will serve as our primary outcome of interest.[1] This data will encompass the total number of all-cause deaths among all Medicare beneficiaries. Additional variables in our analysis will encompass factors such as heat alerts, heat index, population, geographical coordinates (latitude/longitude), designation of weekdays or weekends, the month of the year, air pollution indices, and more.

[1]Medicare is a federal health insurance program in the US, covering a significant majority of older adults. As of 2016, out of the 49.2 million individuals aged 65 and older in the US, approximately 47.8 million were enrolled.

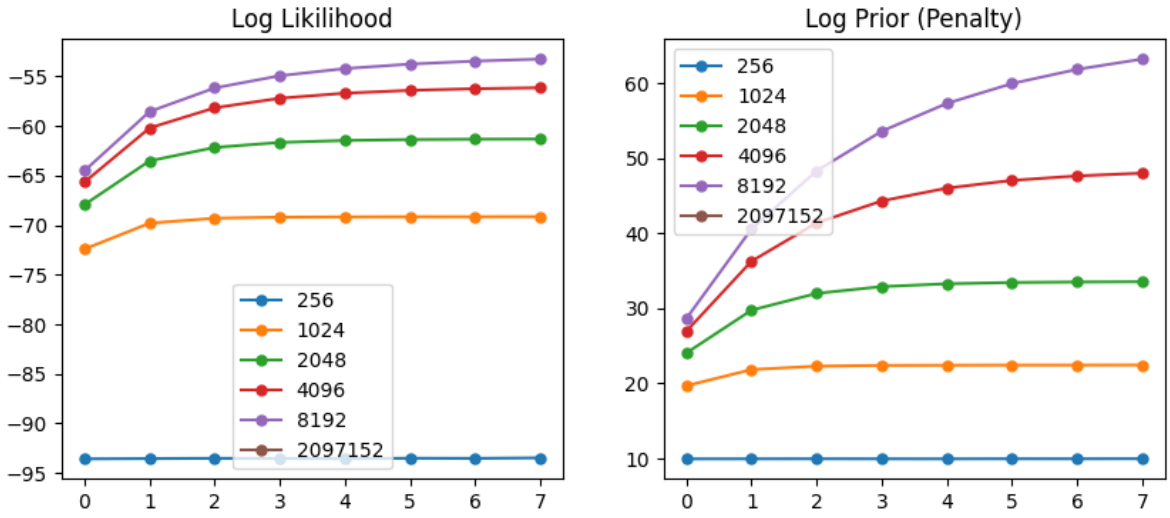

Figure 3: Influence of batch size on the optimization

2.2 Latent Variables

1. Socio-Economic Status can influence Medicare beneficiaries' health outcomes. Areas with lower SES might have higher hospital admissions due to limited access to preventive care or health resources. 2. Healthcare Accessibility: Even if Medicare is available, the quality and accessibility of healthcare services can vary across counties. 3. Environmental Factors: Apart from the heat index and air pollution indices, other environmental factors like humidity, wind speed, or even green space availability in counties can play a role in health outcomes.

2.3 Research question

I will explore:1) understanding the correlation between heat alerts and overall mortality rates among Medicare beneficiaries and identifying counties potentially more vulnerable to heat-induced health complications. 2) determine if counties with lower socio-economic status experience a pronounced increase in hospital admissions during heat alerts compared to affluent areas and how the accessibility and quality of healthcare services influence health outcomes for older adults during these periods. 3) assess whether the repercussions of heat on health outcomes demonstrate any monthly or seasonal variations and pinpointing months where the relationship between heat index and health metrics is especially robust, taking into consideration other environmental factors.

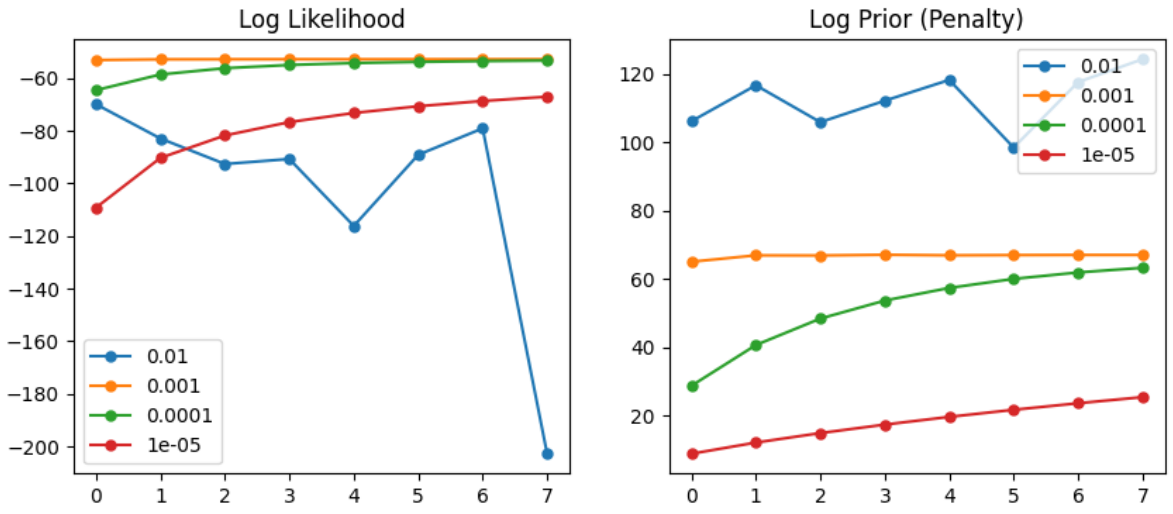

Figure 4: Influence of learning rate on the optimization

References

[1] Kate R. Weinberger, Xiao Wu, Shengzhi Sun, Keith R. Spangler, Amruta Nori-Sarma, Joel Schwartz, Weeberb Requia, Benjamin M. Sabath, Danielle Braun, Antonella Zanobetti, Francesca Dominici, and Gregory A. Wellenius. Heat warnings, mortality, and hospital admissions among older adults in the united states. *Environment International*, 157: 106834, 2021. ISSN 0160-4120. doi: https://doi.org/10.1016/j.envint.2021.106834. URL `https://www.sciencedirect.com/science/article/pii/S0160412021004591`.

[2] Xiao Wu, Kate R Weinberger, Gregory A Wellenius, Francesca Dominici, and Danielle Braun. Assessing the causal effects of a stochastic intervention in time series data: are heat alerts effective in preventing deaths and hospitalizations? *Biostatistics*, page kxad002, 02 2023. ISSN 1465-4644. doi: 10.1093/biostatistics/kxad002. URL `https://doi.org/10.1093/biostatistics/kxad002`.

Table 1: Coefficients of the Model

Variable	Longitude	Latitude	HImaxF_PopW	Weekday
Coefficient	0.040602	0.563841	2.957884	-4.331589
Variable	Weekend	Is April	Is May	Is June
Coefficient	-4.533930	-1.288872	-1.308641	-1.229269
Variable	Is July	Is August	Is September	Is October
Coefficient	-0.699266	-0.834644	-1.453653	-0.858467