

# P8160 Group Project 3: Bayesian Modeling of Hurricane Trajectories

Anyu Zhu | Haotian Wu | Wenhan Bao | Yiming Li | Qihang Wu

5/6/2022

## Objective

In this project, we firstly built a Bayesian model based on the track data of 703 hurricanes in the North Atlantic area since 1950. A Markov Chain Monte Carlo (MCMC) algorithm was designed to generate the distribution of corresponding parameters. With the start time and type of each hurricane, the estimated coefficients from the Bayesian model were used to explore the seasonal differences and wind speed changes over years. Finally, we explored the characteristics of hurricanes associated with the damage and deaths.

## Background

Hurricanes are large rotating tropical storms with winds in excess of 119 kilometers per hour (74 mph). They usually form between June 1 and November 30 in the Atlantic Ocean but can develop in other oceans as well. They are known as typhoons in the western Pacific and cyclones in the Indian Ocean[1].

When a hurricane approaches land, tremendous damage can occur to the nearby cities. Therefore, scientists continue to improve their ability to forecast hurricanes. The sooner they can access accurate information about a hurricane's location and intensity, the better the chances to minimize the its impacts.

## Data Description and Preprocessing

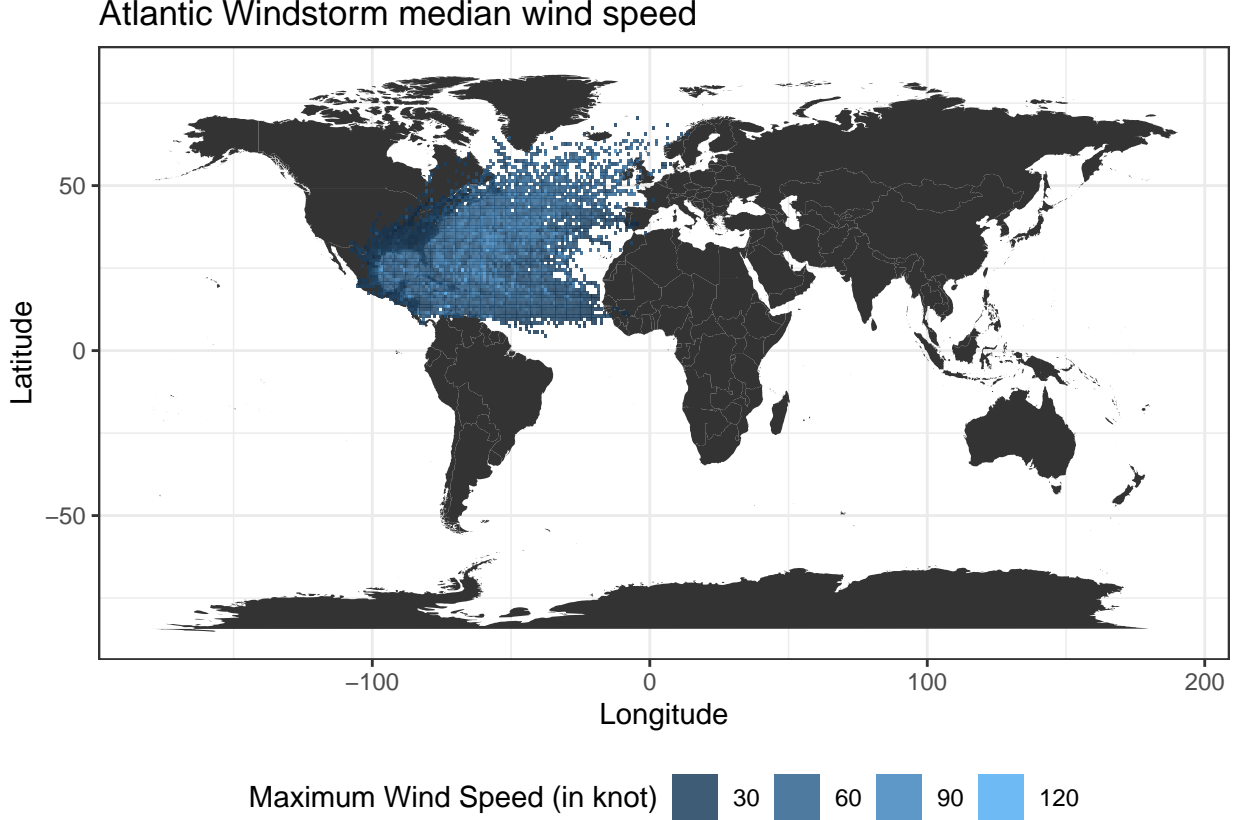
The first data `hurricane703.csv` collected the track data of 703 hurricanes in the North Atlantic area from 1950 to 2013. For all the hurricanes, their locations (longitude & latitude) and maximum wind speed were recorded every 6 hours. The variables include `ID`, `Season`, `Month`, `Nature`, `time`, `Latitude`, `Longitude`, and `Wind.kt`. Main data processing steps for this data are listed as follows:

- 1) Created 3 new variables including the changes of latitude and longitude, as well as the wind speed between the time  $t$  and  $t - 6$ ;
- 2) Removed 9 hurricanes with observations less than 5 to ensure the data partition;
- 3) To explore the seasonal differences, we converted the start month for each hurricane into the variable `season`, which includes Spring, Summer, Fall, and Winter. Finally, we have totally **691** hurricanes in the updated dataset.

The second data `hurricaneoutcome2.csv` recorded the damages and death caused by 46 hurricanes in the United States, and some features extracted from the above hurricane records. To better explore the characteristics related with death and damage, we combined this data with the coefficients obtained from the first model by the hurricane ID. For this data, we also converted different start months into the corresponding seasons.

## Exploratory Data Analysis

To generally understand the distribution of wind speed across the North Atlantic area, we created the following figure to show the mean maximum wind speed within each knot based on the longitude and latitude from the original data `hurricane703.csv`.



## Statistical Methods

### Likelihood

For each hurricane  $i$  and  $k_i$ 's time points, we have the following Bayesian model:

$$Y_i(t+6) = \beta_{0i} + \beta_{1i}Y_i(t) + \beta_{2i}\Delta_{i1}(t) + \beta_{3i}\Delta_{i2}(t) + \beta_{4i}\Delta_{i3}(t) + \varepsilon_i(t),$$

where  $Y_i(t)$  is the wind speed at time  $t$ ,  $\Delta_{i1}$ ,  $\Delta_{i2}$ , and  $\Delta_{i3}$  are the changes of latitude, longitude, and the wind speed between time point  $t$  and  $t-6$ , respectively.  $\varepsilon_i(t)$  follows a normal distributions with mean zero and variance  $\sigma^2$ . The above Bayesian model can be simplified as:

$$Y_i(t+6) = x_i(t) + \varepsilon_i(t),$$

where  $\beta_i = (\beta_{0i}, \beta_{1i}, \dots, \beta_{4i}) \sim N(0, \sigma^2)$ . Based on the property of the multivariate linear regression model, for each hurricane  $i$ , we have:

$$Y_i \mid X_i \sim N_{k_i}(x_i\beta_i, \sigma^2 I_{k_i}),$$

where  $I_{k_i}$  is an identity matrix with  $k_i$  dimensions.

Thus, we can consider the following distribution of each hurricane  $i$ :

$$f(y_i \mid \beta_i, \sigma^2) = [(2\pi)^{k_i} \cdot \det(\sigma^2 I_{k_i})]^{-\frac{1}{2}} \cdot \exp \left\{ -\frac{1}{2} (y_i - x_i\beta_i)^\top (\sigma^2 I_{k_i})^{-1} (y_i - x_i\beta_i) \right\}$$

From above, we derive the following likelihood function:

$$\begin{aligned} f(y | B, \sigma^2) &= \prod_{i=1}^n f(y_i | \beta_i, \sigma^2) \\ &= \prod_{i=1}^n \left( [(2\pi)^{k_i} \cdot \det(\sigma^2 I_{k_i})]^{-\frac{1}{2}} \cdot \exp \left\{ -\frac{1}{2} (y_i - x_i \beta_i)^\top (\sigma^2 I_{k_i})^{-1} (y_i - x_i \beta_i) \right\} \right) \end{aligned}$$

## Prior distributions

We assume the following non-informative prior distributions:

$$\beta_i = (\beta_{0i}, \beta_{1i}, \dots, \beta_{4i}) \sim N_5(\mu, \Sigma),$$

where  $B = (\beta_1^\top, \beta_2^\top, \dots, \beta_n^\top)^\top$  and  $n$  is the number of hurricanes. So,

$$\pi(B | \mu, \Sigma) = \prod_{i=1}^n f(\beta_i) \propto \det(\Sigma)^{-n/2} \cdot \exp \left\{ -\frac{1}{2} \sum_i [(\beta_i - \mu)^\top (\Sigma)^{-1} (\beta_i - \mu)] \right\}.$$

Also,  $\pi(\sigma^2) \propto \frac{1}{\sigma^2}$ ;  $\pi(\mu) \propto 1$ ;  $\pi(\Sigma^{-1}) \propto |\Sigma|^{-(d+1)} \cdot \exp(-\frac{1}{2}\Sigma^{-1})$ .

## Conditional posteriors

The posterior distribution is the product of the likelihood and the prior:

$$g(B, \sigma^2, \mu, \Sigma^{-1} | y) \propto f(y | B, \sigma^2) \cdot \pi(B | \mu, \Sigma^{-1}) \cdot \pi(\sigma^2) \cdot \pi(\mu) \cdot \pi(\Sigma^{-1}),$$

so we have:

For  $\sigma^2$ ,

$$\begin{aligned} \pi(\sigma^2 | \cdot) &\propto \prod_{i=1}^n \det(\sigma^2 I_{k_i})^{-\frac{1}{2}} \cdot \exp \left\{ -\frac{1}{2} \sum_i [(y_i - x_i \beta_i)^\top (\sigma^2 I_{k_i})^{-1} (y_i - x_i \beta_i)] \right\} \cdot \sigma^{-2} \\ &= (\sigma^2)^{-\frac{1}{2} \sum_i k_i} \cdot \exp \left\{ -\frac{1}{2\sigma^2} \sum_i [(y_i - x_i \beta_i)^\top (y_i - x_i \beta_i)] \right\} \cdot \sigma^{-2} \\ &= (\sigma^2)^{-1 - \frac{1}{2} \sum_i k_i} \cdot \exp \left\{ -\frac{1}{2\sigma^2} \sum_i \sum_{t_i} (y_{i,t} - x_{i,t} \beta_i)^2 \right\} \end{aligned}$$

Therefore,  $\sigma^2 \sim \text{Inverse Gamma} \left( \frac{1}{2} \sum_i k_i, \frac{1}{2} \sum_i \sum_{t_i} (y_{i,t} - x_{i,t} \beta_i)^2 \right)$ .

For  $\Sigma$ ,

$$\begin{aligned} \pi(\Sigma^{-1} | \cdot) &\propto \det(\Sigma)^{-n/2} \cdot \exp \left\{ -\frac{1}{2} \sum_i (\beta_i - \mu)^\top \Sigma^{-1} (\beta_i - \mu) \right\} \cdot \det(\Sigma)^{-(d+1)} \cdot \exp \left\{ -\frac{1}{2} \Sigma^{-1} \right\} \\ &= \det(\Sigma)^{-(n/2 + d + 1)} \cdot \exp \left\{ -\frac{1}{2} \left[ \Sigma^{-1} + \sum_i (\beta_i - \mu)^\top \Sigma^{-1} (\beta_i - \mu) \right] \right\} \\ &\propto \det(\Sigma^{-1})^{(n + 2d + 2)/2} \cdot \exp \left\{ -\frac{1}{2} \text{tr} \left[ \Sigma^{-1} \cdot \left( I + \sum_i (\beta_i - \mu) (\beta_i - \mu)^\top \right) \right] \right\} \\ &\propto \det(\Sigma^{-1})^{(n + 3d + 3 - d - 1)/2} \cdot \exp \left\{ -\frac{1}{2} \text{tr} \left[ \Sigma^{-1} \cdot \left( I + \sum_i (\beta_i - \mu) (\beta_i - \mu)^\top \right) \right] \right\} \end{aligned}$$

Thus  $\Sigma^{-1} \sim \text{Wishart} \left( n + 3d + 3, [I + \Sigma_i (\beta_i - \mu) (\beta_i - \mu)^\top]^{-1} \right)$ , that is:

$$\Sigma \sim \text{Inverse Wishart} \left( n + 3d + 3, I + \Sigma_i (\beta_i - \mu) (\beta_i - \mu)^\top \right)$$

For  $\mu$ ,

$$\begin{aligned} \pi(\mu | \cdot) &\propto \exp \left\{ -\frac{1}{2} \Sigma_i \left[ (\beta_i - \mu)^\top \Sigma^{-1} (\beta_i - \mu) \right] \right\} \\ &= \exp \left\{ -\frac{1}{2} \Sigma_i \left( \beta_i^\top \Sigma^{-1} \beta_i + \mu^\top \Sigma^{-1} \mu - 2\beta_i^\top \Sigma^{-1} \mu \right) \right\} \\ &= \exp \left\{ -\frac{1}{2} \left( \Sigma_i \beta_i^\top \Sigma^{-1} \beta_i + \mu^\top n \Sigma^{-1} \mu - 2 \Sigma_i \beta_i^\top \Sigma^{-1} \mu \right) \right\} \\ &= \exp \left\{ -\frac{1}{2} \left( \mu^\top n \Sigma^{-1} \mu - 2 \Sigma_i \beta_i^\top \Sigma^{-1} \mu + \Sigma_i \beta_i^\top \Sigma^{-1} \beta_i \right) \right\} \\ &= \exp \left\{ -\frac{1}{2} \left( \underbrace{\mu^\top n \Sigma^{-1} \mu}_M - 2 \mu^\top \underbrace{\Sigma_i \Sigma^{-1} \beta_i}_N + \Sigma_i \beta_i^\top \Sigma^{-1} \beta_i \right) \right\} \\ &= \exp \left\{ -\frac{1}{2} \left[ (\mu - M^{-1} N)^\top M (\mu - M^{-1} N) \right] \right\}. \end{aligned}$$

Therefore,  $\mu \sim \text{MVN} (M^{-1} N, M^{-1})$ .

And for  $B$ ,

$$\begin{aligned} \pi(B | \cdot) &\propto \exp \left\{ -\frac{1}{2} \Sigma_i \left[ (y_i - x_i \beta_i)^\top (\sigma^2 I_{k_i})^{-1} (y_i - x_i \beta_i) \right] \right\} \cdot \exp \left\{ -\frac{1}{2} \Sigma_i \left[ (\beta_i - \mu)^\top (\Sigma)^{-1} (\beta_i - \mu) \right] \right\} \\ &= \exp \left\{ -\frac{1}{2} \Sigma_i \left[ (y_i - x_i \beta_i)^\top (\sigma^2 I_{k_i})^{-1} (y_i - x_i \beta_i) + (\beta_i - \mu)^\top \Sigma^{-1} (\beta_i - \mu) \right] \right\} \\ &= \exp \left\{ -\frac{1}{2} \Sigma_i \left[ y_i^\top \sigma^{-2} I_{k_i} y_i + \beta_i^\top x_i^\top \sigma^{-2} I_{k_i} x_i \beta_i - 2 y_i^\top \sigma^{-2} I_{k_i} x_i \beta_i + \beta_i^\top \Sigma^{-1} \beta_i + \mu^\top \Sigma^{-1} \mu - 2 \mu^\top \Sigma^{-1} \beta_i \right] \right\} \\ &= \exp \left\{ -\frac{1}{2} \Sigma_i \left[ y_i^\top \sigma^{-2} I_{k_i} y_i + \mu^\top \Sigma^{-1} \mu + \beta_i^\top (\Sigma^{-1} + x_i^\top \sigma^{-2} I_{k_i} x_i) \beta_i - 2 (y_i^\top \sigma^{-2} I_{k_i} x_i + \mu^\top \Sigma^{-1}) \beta_i \right] \right\} \end{aligned}$$

We can define the following terms:

$$\begin{aligned} R &= y_i^\top \sigma^{-2} I_{k_i} y_i + \mu^\top \Sigma^{-1} \mu \\ V &= \Sigma^{-1} + x_i^\top \sigma^{-2} I_{k_i} x_i \\ M &= \sigma^{-2} x_i^\top y_i + \Sigma^{-1} \mu \end{aligned}$$

Thus,  $\pi(B | \cdot) \propto (\beta_i - V^{-1} M)^\top V (\beta_i - V^{-1} M) \sim \text{MVN} (V^{-1} M, V^{-1})$

## Gibbs Sampling

Since directly generating the above parameters from the jointly density is rather complicated, we implement the gibbs sampler to sample each variable in turn. Based on the previous conditional distributions, we updated the four parameters including  $B$ ,  $\sigma^2$ ,  $\Sigma$ , and  $\mu$  in sequence. Finally, we determined 10,000 iterations. Note that we ignored the first 8,000 iterations so that the stationary distribution of the Markov chain was reached.

## Regression Models

To explore the seasonal and annual difference in the wind speed, we took the  $\beta$  coefficients obtained from the Bayesian model as the response and built the following linear regression model:

$$\beta_j \sim Season + Year + Nature, j = 0, \dots, 4,$$

where season and nature are categorical variables, and year is continuous.

To further predict the hurricane-induced damage, we incorporated the coefficients from the Bayesian model and the new predictors in the second data, selected some features by lasso, and finally fit a regression model. The variables  $\beta_0$ , Year, max speed, total affected population, and the affected population reside in the United States were selected in the model. That is,

$$Damage \sim \beta_0 + Year + Maxspeed + Total.Pop + Percent.USA$$

Then we also select some variables by lasso and built the following poisson regression model to evaluate the characteristics related to deaths.

$$Death \sim \beta_4 + MonthSummer + Maxspeed + Maxpressure \\ + Meanpressure + Hours + Total.Pop + Percent.Poor,$$

where the above  $\beta_0$  and  $\beta_4$  are the coefficients from the Bayesian model.

## Results

### MCMC-Gibbs Sampling

The initial values for  $\mu$ ,  $\sigma^2$  and  $\Sigma$  we set are (50,0,0,0,0), 0.5 and `diag(0.5, 5, 5)` respectively. In the case of Gibbs results, we set 10000 iterations and record the  $\beta_{i,j}^n$  in each iteration with a linear model (from the least square method) to demonstrate how similar the coefficients are for each hurricane's train data and  $\mu^n$ .

To minimize the variance generated randomly by Gibbs sampling and avoid the non convergent situation at the beginning of the iteration, we only take the coefficients value from 8000 to 10000 times where we take the mean value for each coefficients  $\beta_{i,j}$  as the final results. To evaluate how well the estimated Bayesian model tracks the individual hurricanes, we divided the observations within each hurricane ID by 8:2 randomly to create a training and testing set. Then we calculated RMSE for each different hurricane.

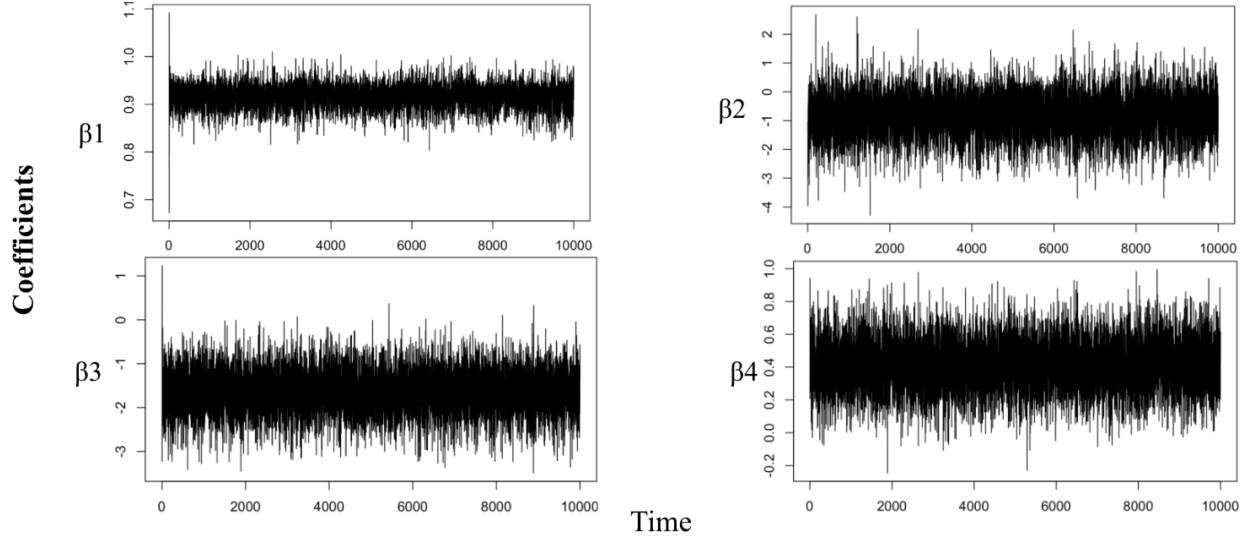
For our final results obtained from Gibbs sampling, you could jump to our [github](#)(Column2: RMSE for the test dataset, Column3-7: Mean coefficients  $\beta_{i,j}$  for each hurricane by Gibbs sampling, Column8-12: coefficients  $\beta_{i,j}$  for each hurricane by `lm()` function, Column13-17: Mean  $\mu$  for  $B$ )

In order to further elucidate the results from our Gibbs methods, we take `ABLE.1950` as an example. Firstly, we trace the Gibbs step for only extracting the  $\beta_{1,j}^n$  from  $B^n$  by plotting the value.



Then to look deeply into how the coefficients change with the increasing of iteration times, we also plot the autocorrelation plot(ACF) and time series plot (TS) where we could tell that the correlation for coefficient values between zero and blue line are not statistically significant.





For the case of ABLE.1950, the RMSE for the remaining 20% test dataset is 4.25 and the comparison of coefficients  $\beta_{1,j}$  is in the table below.

Table 1: Compare the Coefficients from Gibbs Sampling and lm function

| model          | Intercept | Beta1 | Beta2  | Beta3  | Beta4 |
|----------------|-----------|-------|--------|--------|-------|
| Gibbs sampling | 7.601     | 0.919 | -0.788 | -1.644 | 0.411 |
| lm             | 8.833     | 0.912 | -1.394 | -2.013 | 2.983 |

## Seasonal and Annual Difference

There are three significant variables in the regression model for  $\beta_1$ , which represent current wind speed. Season spring and summer, year all have negative coefficients, indicating wind speed tend to decrease in spring and summer. There is no evidence of increasing trend in  $\beta_1$  with the increase in year.

Table 2: Beta 1 Coefficients

|              | Estimate | Std. Error | t value | Pr(>  t ) |
|--------------|----------|------------|---------|-----------|
| (Intercept)  | 2.3394   | 0.3366     | 6.9498  | 0.0000    |
| seasonSpring | -0.0636  | 0.0195     | -3.2599 | 0.0012    |
| seasonSummer | -0.0161  | 0.0062     | -2.6198 | 0.0090    |
| seasonWinter | -0.0476  | 0.0303     | -1.5721 | 0.1164    |
| year         | -0.0008  | 0.0002     | -4.4804 | 0.0000    |
| natureET     | 0.0247   | 0.0233     | 1.0597  | 0.2897    |
| natureNR     | -0.0174  | 0.0367     | -0.4743 | 0.6355    |
| natureSS     | 0.0027   | 0.0160     | 0.1703  | 0.8648    |
| natureTS     | -0.0086  | 0.0125     | -0.6900 | 0.4904    |

There are no significant variables in the regression model for  $\beta_2$  and  $\beta_3$ . Season spring is the only significant variable in the regression model for  $\beta_4$ , indicating seasonal difference in acceleration of wind speed.

Table 3: Beta 4 Coefficients

|              | Estimate | Std. Error | t value | Pr(>  t ) |
|--------------|----------|------------|---------|-----------|
| (Intercept)  | 2.3357   | 1.3705     | 1.7043  | 0.0888    |
| seasonSpring | -0.2316  | 0.0794     | -2.9177 | 0.0036    |
| seasonSummer | -0.0024  | 0.0251     | -0.0938 | 0.9253    |
| seasonWinter | -0.0587  | 0.1233     | -0.4758 | 0.6343    |
| year         | -0.0009  | 0.0007     | -1.3816 | 0.1676    |
| natureET     | -0.0984  | 0.0947     | -1.0391 | 0.2991    |
| natureNR     | -0.1024  | 0.1495     | -0.6846 | 0.4938    |
| natureSS     | -0.0971  | 0.0650     | -1.4927 | 0.1360    |
| natureTS     | -0.0365  | 0.0509     | -0.7176 | 0.4733    |

## Hurricane-induced damage and deaths prediction

After variable selection by lasso regression, the linear regression model for hurricane-induced damage has positive coefficients for Year,  $\beta_0$ , Maxspeed, Total.Pop, and Percent.USA, indicating positive correlation between these hurricane characteristics and damage. Among these, year is the only significant variable.

Table 4: Coefficients of regression model for Damage

|              | Estimate | Std. Error | t value | Pr(>  t ) |
|--------------|----------|------------|---------|-----------|
| (Intercept)  | 11.155   | 3.067      | 3.638   | 0.000833  |
| Year         | 7.918    | 3.229      | 2.452   | 0.019035  |
| beta0        | 3.728    | 3.158      | 1.181   | 0.245323  |
| Maxspeed     | 4.787    | 3.257      | 1.470   | 0.150046  |
| Total.Pop    | 2.128    | 3.240      | 0.657   | 0.515240  |
| Percent. USA | 5.913    | 3.205      | 1.845   | 0.073077  |

The variables selected by lasso regression for hurricane-induced deaths include Summer,  $\beta_4$ , Maxspeed, Maxpressure, Hours, Total.Pop, and Percent.Poor. All the selected variables shows significance in the poisson regression model.  $\beta_4$ , hours, and Percent.Poor has positive correlation with deaths while the other variables have negative coefficients.

Table 5: Coefficients of regression model for Deaths

|              | Estimate     | Std. Error  | z value | Pr(>  z )  |
|--------------|--------------|-------------|---------|------------|
| (Intercept)  | 5.539e + 01  | 2.775e + 00 | 19.958  | < 2e - 16  |
| MonthSummer  | -1.097e + 00 | 2.250e - 02 | -48.747 | < 2e - 16  |
| beta4        | 3.641e + 00  | 6.543e - 02 | 55.645  | < 2e - 16  |
| Maxspeed     | -8.392e - 04 | 4.104e - 04 | -2.045  | 0.0409     |
| Maxpressure  | -5.160e - 02 | 2.738e - 03 | -18.850 | < 2e - 16  |
| Hours        | 1.903e - 03  | 9.169e - 05 | 20.754  | < 2e - 16  |
| Total.Pop    | -8.973e - 08 | 1.371e - 08 | -6.547  | 5.87e - 11 |
| Percent.Poor | 3.452e - 02  | 1.514e - 04 | 227.951 | < 2e - 16  |

## Discussion and Limitations

MCMC algorithm could return converged results whatever the initial values are, but some initial values that allow for faster convergence. Since our likelihood are actually based on normal assumption, `lm` regression coefficients might be warm start. At the same time, MCMC returns consistent results with hypothesis. We supposed coefficients from 700 hurricanes follow multivariate normal distribution with common mean  $\mu$ . In our results, mean of coefficients beta are very close to  $\mu$ .



There are some limitations for our project. For mcmc, when calculating the variance of beta coefficients, we apply the sum square of residuals based on current beta values and corresponding x values, i.e. previous wind speed, latitude, longitude and previous wind speed variation. However, some specific hurricanes only have records no more than 10 and almost zero wind speed variation, which might cause anomalies in iteration. To be specific, zero predictor values force the product of coefficient and predictor to be zero and this is an extremely strong regularization on coefficients. For these types of hurricanes, mcmc might not work well and return large RMSE, which also depended on how to split training and testing data in cross validation.

We hope mcmc and ordinary linear regression have similar results, but the amount of some specific hurricanes is too small and some predictors values are highly correlated within such a limited time span. If more than two predictors of one hurricane are highly correlated, the determinant of the predictors matrix for this particular hurricane seems to be zero which indicates that coefficients are not unique. In this case, the lm function in R simply returns NA for constant predictors, so we could not make a comparison with mcmc predictors.

## Contributions

We contributed to this project evenly.

## Reference

[1] [https://www.whoi.edu/know-your-ocean/ocean-topics/ocean-human-lives/natural-disasters/hurricanes/?gclid=Cj0KCQjwsdiTBhD5ARIsAIPW8CIO0bvwlol6Pisxe\\_WV7owA4KdbMfpAhMxqa-z8Ir3fsQyfiHPa2MAaAlmZEALw\\_wcB](https://www.whoi.edu/know-your-ocean/ocean-topics/ocean-human-lives/natural-disasters/hurricanes/?gclid=Cj0KCQjwsdiTBhD5ARIsAIPW8CIO0bvwlol6Pisxe_WV7owA4KdbMfpAhMxqa-z8Ir3fsQyfiHPa2MAaAlmZEALw_wcB)

## Appendix

For codes please click [here](#)