P8160 - Bayesian Modeling of Hurricane Trajectories

Paulina Han, Waveley Qiu, Yida Wang, Lin Yang, Jibei Zheng, Haolin Zhong

2022-05-09

Abstract

In light of the devastation incurred by hurricanes in recent years, interest has been taken into developing models to assist in forecasting the impact of these natural disasters. In this project, we developed and generated values from a Gibbs sampler to estimate the parameters in a hierarchical Bayesian model predicting the future wind speed of a hurricane, which is an important metric used in estimating the severity of the storm. These estimated parameters were then used in two sub-analyses; the first analysis investigated whether the derived coefficients have any underlying covariates and if the model constructed to examine those covariates indicate that wind speed increases over time, and the second investigated whether the coefficients have any predictive effect on estimating the impact of a hurricane. Ultimately, the parameter estimates obtained from the MCMC algorithm performed quite well and predicted the future wind speeds (at least for hurricanes with enough observations) with impressive accuracy. The seasonal analysis that was conducted caused us to conclude that there is not enough evidence to conclude that wind speed increases over time. Additionally, it was determined that many of the derived coefficients, when used as autoregressive variables, are significant predictors of the impact of a hurricane. Finally, though the chain of values generated from the sampler did not achieve perfect convergence for some parameters, we hypothesized this to be reflective of the error in the prior distribution that was assumed, particularly as the parameter space of the posterior distribution is large.

1 Introduction

1.1 Background

Against the backdrop of several seasons of severe hurricane impact, many efforts have been made to advance the technology available to measure different characteristics of these tropical storms. The data that has become available through these tools has become of interest to both meteorologists and natural disaster response agencies, as aspirations to forecast both the trajectory and impacts of these storms seem achievable.

One metric that is greatly associated with the impact of a hurricane is its wind speed. In fact, the standard measure of the severity of a hurricane, the Saffir-Simpson Hurricane Wind Scale, categorizes hurricanes into levels of predicted damage solely based on their maximum wind speeds. Constructing a method to predict the winds speed of a hurricane, particularly as this data is more easily accessible, then becomes a natural research interest.

1.2 Objectives

In this study, we first endeavored to design and execute a Markov chain Monte Carlo (MCMC) simulation to estimate parameters from a model that would employ information about the trajectory of a hurricane's velocity to predict its wind speed. The following hierarchical Bayesian model was proposed to predict the wind speed of the i^{th} hurricane at time $t+6$:

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

where $Y_i(t)$ is the wind speed at time t, $\Delta_{i,1}(t)$, $\Delta_{i,2}(t)$, $\Delta_{i,3}(t)$ are the changes in latitude, longitude, and wind speed between times t and $t-6$, $\varepsilon_i(t)$ is the random error associated with each $Y_i(t+6)$, and $\beta_i = (\beta_{0,i},\ \beta_{1,i},\ \beta_{2,i},\ \beta_{3,i},\ \beta_{4,i})$ are random coefficients associated with the i^{th} hurricane. The prior distributions for each of these parameters are assumed to be as follows:

$$\begin{aligned}
&\epsilon_i(t) \sim N(0,\sigma^2), \text{ which are independent across } t\\
&P\left(\sigma^2\right) \propto \frac{1}{\sigma^2}\\
&P(\boldsymbol{\mu}) \propto 1\\
&P\left(\Sigma^{-1}\right) \propto |\Sigma|^{-(d+1)} \exp\left(-\frac{1}{2}\text{tr}(\Sigma^{-1})\right), \text{ where } d \text{ is the dimension of } \beta_i\\
&\beta_i \sim MVN(\boldsymbol{\mu}, \Sigma)
\end{aligned}$$

Our goal in using an MCMC algorithm was to estimate $\Theta = (\boldsymbol{B}, \boldsymbol{\mu}, \sigma^2, \Sigma)$. We determined that a Gibbs sampler would be an adequate algorithm to utilize and its methodology, as well as the derived conditional distributions necessary for its use, are discussed in subsequent sections.

After generating the Markov chain and evaluating its performance in sampling from our desired posterior distribution, we then aimed to examine the properties of the derived coefficients, β_i, more closely. We first conducted an analysis to investigate whether any of the derived linear coefficients, β_i, experienced seasonal changes. Then, we also sought to evaluate whether these coefficients can be used as predictors themselves to forecast the impact of a hurricane directly.

2 Markov Chain Monte Carlo

2.1 Data

The original dataset contains 22038 rows of records for 703 hurricanes. Each row recorded the latitude, longitude, and wind speed of the hurricane at a specific time point, along with the hurricane's type and ID.

Self-join was performed on the original data to map the hurricane record at time t with the hurricane records at $t-6$ and $t+6$. The data was then partitioned into a training dataset and a test dataset. Sampling was stratified by hurricane, from which 80% of records for each hurricane were randomly assigned to the training dataset, and the remaining were reserved for the test dataset. Hurricanes with less than 7 records were discarded to ensure that at least 5 records appeared in the training dataset for each hurricane.

2.2 Methods

The posterior distribution of parameters involved in this hierarchical Bayesian model is $\pi(\Theta|\mathbf{Y}) \propto f(\mathbf{Y}|\Theta)\pi(\Theta)$. As it is difficult to derive the closed form for the posterior Bayesian estimator, we applied an Markov chain Monte Carlo (MCMC) algorithm to sample from this desired distribution. We decided to use a Gibbs sampler to draw from the posterior distribution, as it is possible to derive the condition distributions for each parameter of interest.

2.2.1 Derivations

As our MCMC algorithm of choice is Gibbs Sampling, we need to first derive the joint posterior distribution of all parameters and also the conditional distributions of each parameter of interest.

Joint Posterior Distribution of Θ

Let $\Theta = (\boldsymbol{B}^T, \boldsymbol{\mu}^T, \boldsymbol{\sigma}^2, \boldsymbol{\Sigma})$, the posterior distribution can be written as:

$$P(\Theta \mid Y) \propto f(Y|\Theta)P(\Theta) = f(Y \mid \boldsymbol{B}, \boldsymbol{\mu}, \sigma^2, \boldsymbol{\Sigma})f(\boldsymbol{B} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})P(\boldsymbol{\mu})P(\sigma^2)P(\boldsymbol{\Sigma}^{-1})$$

Denote $Y = (Y_1, Y_2, ..., Y_N)$, where N is the total number of hurricanes. Let $E(Y_i) = X_i\beta_i^T = \beta_{0,i} + \beta_{1,i}Y_i(t) + \beta_{2,i}\Delta_{i,1}(t) + \beta_{3,i}\Delta_{i,2}(t) + \beta_{4,i}\Delta_{i,3}(t)$ and we can see $Y_i \sim MVN(X_i\beta_i^T, \sigma^2 I_{n_i})$, where n_i is the number of observation of the i^{th} hurricane.

$$\begin{aligned} f(Y \mid \boldsymbol{B}, \boldsymbol{\mu}, \sigma^2, \boldsymbol{\Sigma}) &= \prod_{i=1}^{N} f(Y_i|B, \mu, \Sigma, \sigma^2) \\ &= \prod_{i=1}^{N} (2\pi)^{-\frac{n_i}{2}} \left|\sigma^2 I_{n_i}\right|^{-\frac{1}{2}} \exp(-\frac{1}{2}(Y_i - X_i\beta_i^T)^T(\sigma^2 I_{n_i})^{-1}(Y_i - X_i\beta_i^T)) \end{aligned}$$

Denote $B = (\beta_1^T, \beta_2^T, ..., \beta_N^T)^T$, and $\beta_i \sim \text{MVN}(\mu, \Sigma)$.

$$f(B|\mu, \Sigma) = \prod_{i=1}^{N} (2\pi)^{-\frac{5}{2}} \left|\sigma\right|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\beta_i - \mu)^T\Sigma^{-1}(\beta_i - \mu)\right)$$

Therefore, the posterior distribution of Θ is as follows:

$$\begin{aligned}
&P(\Theta \mid Y) \propto f(Y \mid \boldsymbol{B}, \boldsymbol{\mu}, \sigma^2, \boldsymbol{\Sigma}) f(\boldsymbol{B} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) P(\boldsymbol{\mu}) P(\sigma^2) P(\boldsymbol{\Sigma}^{-1}) \\
&= \prod_{i=1}^{N} (2\pi)^{\frac{n_i}{2}} \left|\sigma^2 I_{n_i}\right|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(Y_i - X_i\beta_i)^T (\sigma^2 I_{n_i})^{-1} (Y_i - X_i\beta_i^T)\right) |\Sigma|^{-\frac{n}{2}} \\
&\qquad \exp\left(-\frac{1}{2}(\beta_i - \mu)^T \Sigma^{-1} (\beta_i - \mu)\right) |\Sigma|^{-(d+1)} \exp(-\frac{1}{2} tr(\Sigma^{-1})) \frac{1}{\sigma^2} \\
&= \prod_{i=1}^{N} \sigma^{-n_i - 2} |A|^{d + \frac{N}{2} + 1} \exp\left[-\frac{1}{2}\left[(Y_i - X_i\beta_i^T)^T (\sigma^2 I_{n_i})^{-1} (Y_i - X_i\beta_i^T) + (\beta_i - \mu)^T A (\beta_i - \mu)\right] - \frac{1}{2} tr(A)\right] \\
&= \sigma^{-\sum_{i=1}^{N} n_i - 2} |A|^{d + \frac{N}{2} + 1} \exp\left[-\frac{1}{2}\sum_{i=1}^{N}\left[(Y_i - X_i\beta_i^T)^T (\sigma^2 I_{n_i})^{-1} (Y_i - X_i\beta_i^T) + (\beta_i - \mu)^T A (\beta_i - \mu)\right] - \frac{1}{2} tr(A)\right],
\end{aligned}$$

where $A = \Sigma^{-1}$

Conditional Distribution of $\boldsymbol{B}$

The conditional distribution of $\boldsymbol{B}$ is derived as follows:

$$\begin{aligned}
f(\beta_i | \mu, \sigma^2, \Sigma,) &\propto \exp(-\frac{1}{2}[\big(Y_i - X_i\beta_i^T)^T (\sigma^2 I_{ni})^{-1} (Y_i - X_i\beta_i^T) + (\beta_i - \mu)^T \Sigma^{-1} (\beta_i - \mu)]) \\
&= R + \beta_i^T V \beta_i - 2M\beta_i \\
&\propto (\beta_i - V^{-1}M)^T V (\beta_i - V^{-1}M) \\
\implies & \beta_i \sim MVN(V^{-1}M, V^{-1}),
\end{aligned}$$

where $V = \Sigma^{-1} + X_i^T \sigma^{-2} I_{ni} X_i$, $R = Y_i^T \sigma^{-2} I_{ni} Y_i + \mu^T \Sigma^{-1} \mu$, $M = Y_i^T \sigma^{-2} I_{ni} X_i + \mu^T \Sigma^{-1}$

Conditional Distribution of $\boldsymbol{\mu}$

The conditional distribution of $\boldsymbol{\mu}$ is derived as follows:

$$\begin{aligned}
f(\mu | B, \Sigma, \sigma^2, Y) &\propto \exp(-\frac{1}{2}\sum_{i=1}^{N}(\beta_i - \mu)^T \Sigma^{-1} (\beta i - \mu)) \\
&= \exp(-\frac{1}{2}(\sum_{i=1}^{N} \beta_i^T \Sigma^{-1} \beta_i + \mu^T N \Sigma^{-1} \mu - 2\sum_{i=1}^{N} \beta_i^T \Sigma^{-1} \mu)) \\
&\propto \exp(R + \mu V \mu - 2M\mu) \\
&= \exp((\mu - V^{-1}M)^T V (\mu - V^{-1}M)) \\
\implies & \mu \sim MVN(V^{-1}M, V^{-1}),
\end{aligned}$$

where $V = N\Sigma^{-1}$, $R = \sum_{i=1}^{N} \beta_i^T \Sigma^{-1} \beta_i$, and $M = \sum_{i=1}^{N} \Sigma^{-1} \beta_i$.

Conditional Distribution of σ^2

The conditional distribution of σ^2 is derived as follows:

$$\begin{aligned}
f(\sigma^2|B,\mu,\Sigma,Y) &\propto \sigma^{-\sum_{i=1}^N n_i - 2} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^N \left[(Y_i - X_i\beta_i^T)^T I_{n_i}(Y_i - X_i\beta_i^T)\right]\right] \\
&= \left(\frac{1}{\sigma^2}\right)^{\frac{\sum_{i=1}^N n_i}{2}+1} \exp\left[-\frac{1}{2}\sum_{i=1}^N\sum_{t=1}^{n_i}\left[(y_{it} - x_{it}\beta_i^T)^2\right]\frac{1}{\sigma^2}\right] \\
\text{Let } W = \sigma^2 : & \\
&= \left(\frac{1}{W}\right)^{\frac{\sum_{i=1}^N n_i}{2}+1} \exp\left[-\frac{1}{2}\sum_{i=1}^N \left[(Y_i - X_i\beta_i^T)^T(Y_i - X_i\beta_i^T)\right]\frac{1}{W}\right] \\
&\implies W \sim \text{inverse Gamma}(\frac{\sum_{i=1}^N n_i}{2}, \frac{1}{2}\sum_{i=1}^N\sum_{t=1}^{n_i}(y_{it} - x_{it}\beta_i^T)^2)
\end{aligned}$$

Conditional Distribution of $\mathbf{\Sigma}^{-1}$

The conditional distribution of $\mathbf{\Sigma}^{-1}$ is derived as follows:

Let $A = \Sigma^{-1}$. Then:

$$\begin{aligned}
f(A|B,\mu,\sigma^2,Y) &\propto |A|^{d+\frac{N}{2}+1} \exp\left[-\frac{1}{2}\sum_{i=1}^N (\beta_i - \mu)^T A(\beta_i - \mu)\right] \exp(-\frac{1}{2}tr(A)) \\
&= |A|^{d+\frac{N}{2}+1} \exp\left[-\frac{1}{2}tr\left(A + \sum_{i=1}^N (\beta_i - \mu)^T A(\beta_i - \mu)\right)\right] \\
&= |A|^{d+\frac{N}{2}+1} \exp\left[-\frac{1}{2}tr\left(A\left(I + \sum_{i=1}^N (\beta_i - \mu)(\beta_i - \mu)^T\right)\right)\right] \\
&\implies A \sim \text{Wishart}\left(n^*,\ (I + \sum_{i=1}^N (\beta_i - \mu)(\beta_i - \mu)^T)^{-1}\right),
\end{aligned}$$

where $n^* = 3d + N + 3$

2.2.2 Gibbs Sampling

After deriving the necessary conditional distributions, implementing the Gibbs sampling algorithm is straightforward. As our target posterior distribution to sample from is $f(\mathbf{B}, \boldsymbol{\mu}, \sigma^2, \mathbf{\Sigma}|\mathbf{Y})$, we first set $\mathbf{B}, \boldsymbol{\mu}, \sigma^2, \mathbf{\Sigma}$ to some initial values, and then cycle through each of the conditional distribution and sample from each in turn, always conditioning on the most recent values of the other parameters. More precisely,

1. Initialize $\Theta_0 = (\mathbf{B}_0, \boldsymbol{\mu}_0, \sigma_0^2, \mathbf{\Sigma}_0)$
2. for iteration i = 1,2,...K

Sample $\mathbf{B}_i \sim f(\mathbf{B}|\boldsymbol{\mu}_{i-1}, \sigma^2_{i-1}, \boldsymbol{\Sigma}_{i-1}, \mathbf{Y})$.
Sample $\boldsymbol{\mu}_i \sim f(\boldsymbol{\mu}|\mathbf{B}_i, \sigma^2_{i-1}, \boldsymbol{\Sigma}_{i-1}, \mathbf{Y})$.
Sample $\sigma^2_i \sim f(\sigma^2|\mathbf{B}_i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_{i-1}, \mathbf{Y})$.
Sample $\boldsymbol{\Sigma}_i^{-1} \sim f(\boldsymbol{\Sigma}^{-1}|\mathbf{B}_i, \boldsymbol{\mu}_i, \sigma^2_i, \mathbf{Y})$ then take inverse.

We started the algorithm off with the following initial values: $\mathbf{B}_0$ is a $678 * 5$ matrix of ones; $\boldsymbol{\mu}_0$ is a 5-dimensional vector of ones; $\sigma^2_0 = 1$; and $\boldsymbol{\Sigma}_0 = \boldsymbol{I}_5$. We set the algorithm to run a total of K=10000 iterations and decided to incorporate a burn-in of 5000 values. A number of different starting values were tested, but the chain behaved similarly, no matter what values it was initially given.

2.2.3 Results

To assess the convergence of the generated chain of values, we constructed auto-correlation function (ACF) and time series (TS) plots for each (or a sample of each) parameter, and defined non-convergence to occur when the correlations between two values with $lag > 1$ is significant. In addition, we also constructed histograms to visualize the sampling distributions.

To examine the sampled values for $\mathbf{B}$, we randomly selected the 33th hurricane "DOLLY.1953" as representative (Figure 1). Overall, it appears that the parameters have been estimated quite well, and all five coefficients have approximately normal distributions. The vertical blue lines in the histograms indicate the sample means and the two vertical red lines indicate the lower and upper boundaries of an approximate 95% credible interval. $\beta_1, \beta_2, \beta_3, \beta_4$ seem to achieve convergence, but the ACF plot indicates that there is some correlation ($\rho < 0.2$) between the $lag > 1$ values in the chain for intercept term, β_0, and we are less confident about its convergence.

The ACF plot for the sampled values in $\boldsymbol{\mu}$ (Figure 2) reported more correlation between $lag > 1$ terms, and obvious fluctuations in the trace plot indicates non-convergence, particularly for μ_0. There was also no indication of convergence improving for longer chains. Nonetheless, because the generated values centered around the same horizontal line, remained within a range, and the distributions appear approximately bell-shaped, we felt comfortable accepting the distribution as is and drawing parameter estimates from it.

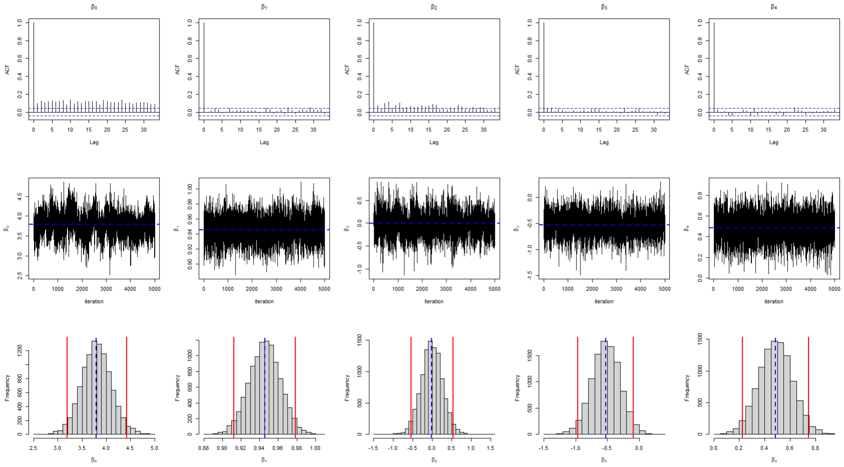


Figure 1: Auto-correlation functions, time series plots and histograms of $\boldsymbol{\beta}_{33}$

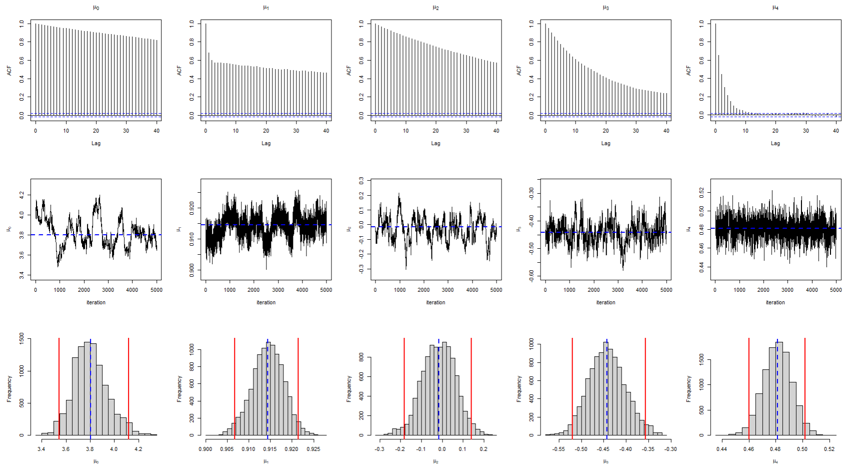


Figure 2: Auto-correlation functions, time series plots and histograms of $\boldsymbol{\mu}$

Finally, we examined the convergence of the variance and covariance parameters, σ^2 and $\boldsymbol{\Sigma}^{-1}$, whose convergence plots are depicted in Figure 3. In order to represent the values of $\boldsymbol{\Sigma}^{-1}$ in these plots, we have used the trace of $\boldsymbol{\Sigma}^{-1}$, which effectively summarizes the information in the matrix. From the plots, we see that σ^2 and $\boldsymbol{\Sigma}^{-1}$ converged.

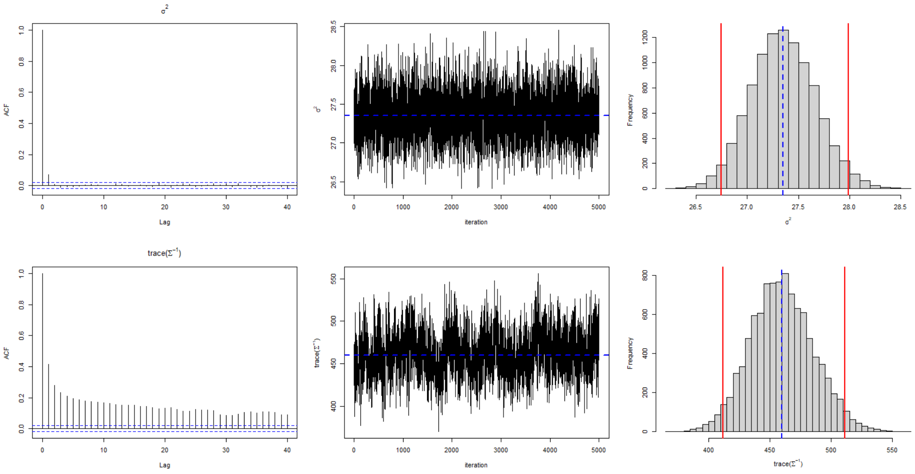


Figure 3: Auto-correlation functions, time series plots and histograms of σ^2 and $tr(\boldsymbol{\Sigma}^{-1})$

Our final parameter estimates are obtained by averaging all the post-burn-in samples. The estimates for $\mathbf{B}$ and $\boldsymbol{\mu}$ with 95% empirical credible intervals and the estimate for $\boldsymbol{\Sigma}^{-1}$ are shown in Table 1. The estimate for σ^2 is 27.36[26.76, 27.98].

Hurricane	β_0	β_1	β_2	β_3	β_4
ABLE.1950	3.79[3.19,4.42]	0.95[0.92,0.97]	-0.12[-0.65,0.4]	-0.53[-0.98,-0.1]	0.54[0.32,0.77]
BAKER.1950	3.79[3.2,4.4]	0.92[0.89,0.95]	-0.1[-0.65,0.44]	-0.39[-0.84,0.07]	0.68[0.49,0.87]
CHARLIE.1950	3.78[3.17,4.38]	0.94[0.92,0.97]	-0.01[-0.53,0.51]	-0.42[-0.85,0.04]	0.45[0.18,0.71]
DOG.1950	3.81[3.22,4.43]	0.96[0.94,0.97]	-0.06[-0.58,0.45]	-0.39[-0.79,-0.01]	0.53[0.31,0.76]
EASY.1950	3.8[3.19,4.43]	0.92[0.88,0.95]	-0.01[-0.53,0.53]	-0.43[-0.89,0.02]	0.54[0.33,0.74]
FOX.1950	3.79[3.16,4.4]	0.95[0.93,0.98]	-0.1[-0.66,0.42]	-0.56[-1.02,-0.11]	0.56[0.31,0.81]
GEORGE.1950	3.81[3.21,4.4]	0.95[0.93,0.98]	-0.03[-0.56,0.49]	-0.38[-0.78,0.02]	0.46[0.19,0.73]
HOW.1950	3.82[3.22,4.43]	0.89[0.82,0.96]	-0.02[-0.55,0.52]	-0.43[-0.89,0.03]	0.47[0.17,0.79]
ITEM.1950	3.83[3.23,4.45]	0.92[0.88,0.97]	-0.05[-0.57,0.49]	-0.45[-0.91,0.01]	0.5[0.3,0.71]
JIG.1950	3.83[3.23,4.45]	0.95[0.92,0.98]	-0.02[-0.55,0.5]	-0.47[-0.92,-0.03]	0.48[0.22,0.75]

(a)

μ_0	μ_1	μ_2	μ_3	μ_4
3.82[3.59,4.12]	0.91[0.91,0.92]	-0.03[-0.2,0.12]	-0.44[-0.52,-0.36]	0.48[0.46,0.5]

(b)

17.0148877	7.740045	-0.1368787	0.8033023	1.8493396
7.7400455	360.894502	5.5883769	3.7666486	-9.9431197
-0.1368787	5.588377	19.0227527	1.0456746	0.5528947
0.8033023	3.766649	1.0456746	22.7116754	-3.4877498
1.8493396	-9.943120	0.5528947	-3.4877498	40.9917065

(c)

Table 1: (a) Part of the estimates of $\mathbf{B}$ (b) Estimates for $\boldsymbol{\mu}$ (c) Estimates for $\boldsymbol{\Sigma}^{-1}$

Figure 4 depicts the prediction performance of our estimated Bayesian models on a few hurricanes. It appears that our model is able to track individual hurricanes very well.

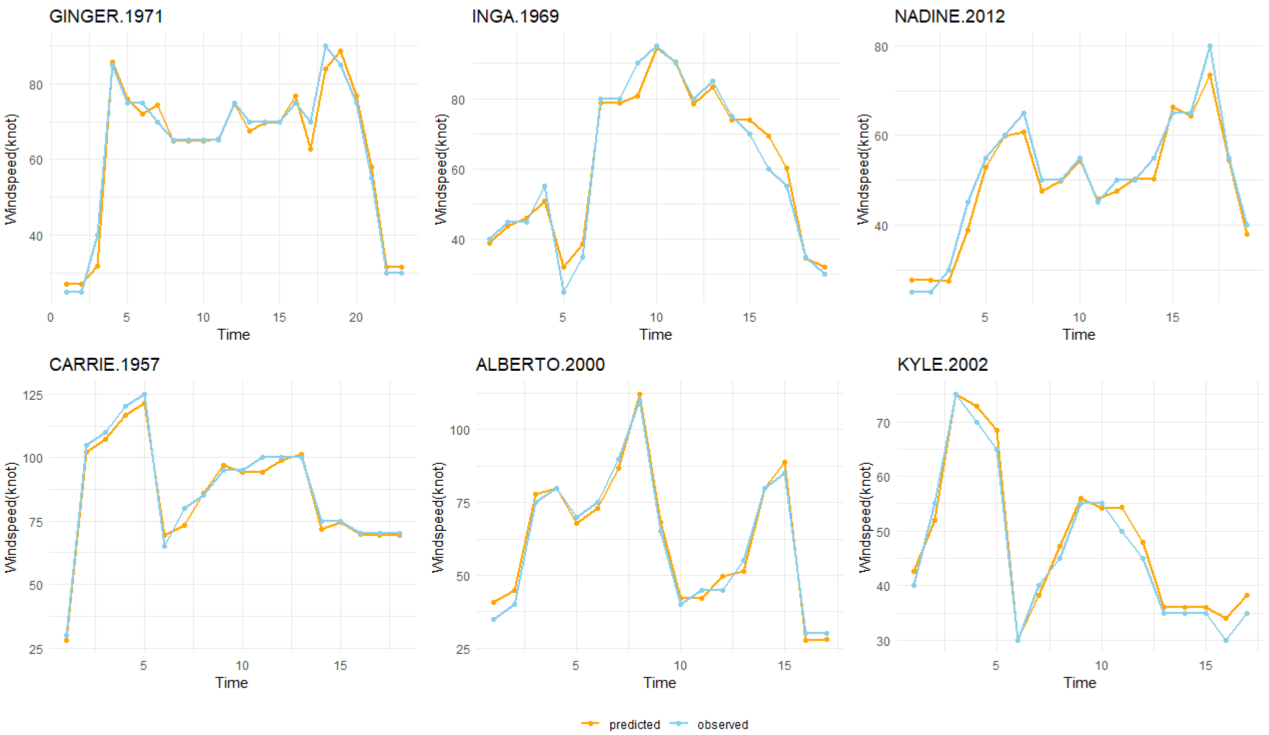


Figure 4: Predicted vs. observed values of windspeed

The RMSE obtained from the entire test dataset is 5.78. The RMSEs of some individual hurricanes are shown in Table 2. The density plot of individual RMSEs is shown in Figure 5. Most of the individual models have relatively small RMSEs (around 3 or 4), but there are a few models having abnormally large RMSEs, indicating poor predictive performance. This could be mostly due to a lack of observations for those particular hurricanes.

Hurricane	RMSE
ABLE.1950	3.142189
BAKER.1950	6.665961
CHARLIE.1950	2.420988
DOG.1950	3.352041
EASY.1950	7.954826
FOX.1950	3.360563
GEORGE.1950	3.966812
HOW.1950	3.212678
ITEM.1950	15.515327
JIG.1950	2.198730

Table 2: RMSEs of some individual hurricanes

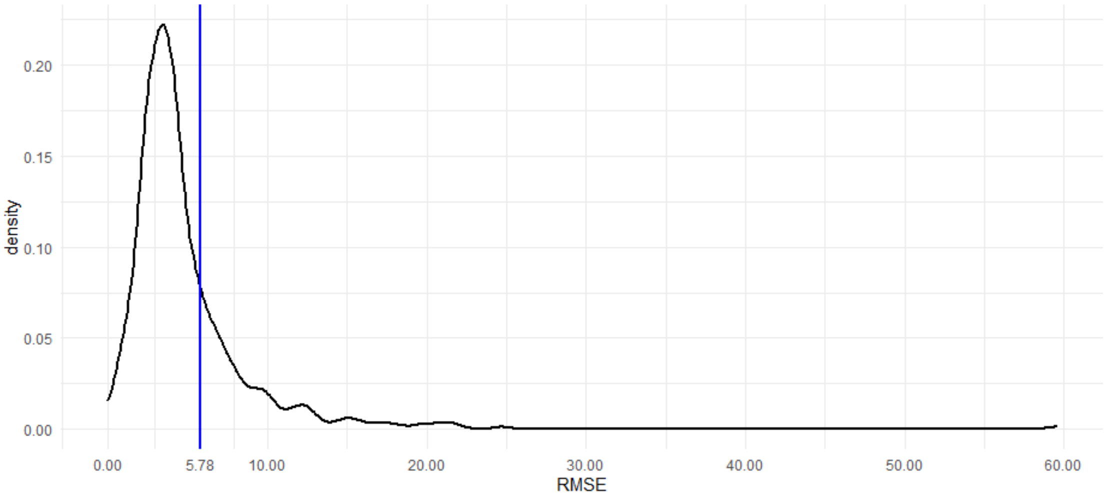


Figure 5: Density plot of RMSEs

3 Seasonal Analysis

3.1 Exploratory Analysis

In order to investigate whether the coefficients derived from the MCMC method may depend on any other covariates, we first examined whether any seasonal patterns exist within the coefficients. These coefficients $\boldsymbol{\beta}_i$ were plotted against the years that the hurricanes occurred in (Appendix C), the months they started in (Appendix B), and also according to their type (Appendix E).

It seems that hurricanes occurred mostly from July to October corresponding to the seasons summer and fall (Appendix D). We can also see that the relationships between $\boldsymbol{\beta}_i$ and hurricane occurred year are very steady, except $\boldsymbol{\beta}_1$ has a slightly decrease trend over the year. Based on the boxplot of hurricane types and $\boldsymbol{\beta}_i$, we can draw a conclusion that tropical storm hurricanes are the most changeable hurricane for the biggest interval of beta. Besides, an important fact is that $\boldsymbol{\beta}_2$ have both negative and positive value, $\boldsymbol{\beta}_3$ are always below zero, while others are all positive.

3.2 Modeling

We first used the year, starting month, and hurricane type as predictors and fit five linear models, where each of the $\boldsymbol{\beta}_i (i = 0, 1, 2, 3, 4)$ was an outcome variable. Starting month and hurricane type were both taken to be categorical variables, while year was treated as a continuous predictor. The constructed model can be defined as follows:

$$\begin{aligned} E(\beta_i) = \ & \alpha_0 + \alpha_{1m} I(\text{Month} = M) + \alpha_2 \times \text{Year} + \alpha_{3n} I(\text{Type} = N) \\ & \text{M: April-December, January(reference)} \\ & \text{N: ET, DS, NR, SS, DS(reference)} \end{aligned}$$

All months excluding February and March were represented in the dataset, with January being set as the reference month in our model. The hurricane types represented in this dataset are extra tropical (ET), disturbance (DS), not rated (NR), sub tropical (SS) and tropical storm (TS), with disturbance (DS) taken as the reference type in our model.

We also built linear models for different seasons by defining a new variable that would take on the value "Spring" if the hurricane occurred between March to May, "Summer" if it occurred between June through August, "Fall" if it occurred from September to November, and "Winter" if it occurred from December to February. With the exclusion of the hurricane's starting month, all other predictors from the previous model were also used in this model.

$$\begin{aligned} E(\beta_i) = \ & \alpha_0 + \alpha_{1s} I(\text{Season} = S) + \alpha_2 \times \text{Year} + \alpha_{3n} I(\text{Type} = N) \\ & \text{S: Spring, Summer, Winter, Fall (reference)} \\ & \text{N: ET, DS, NR, SS, DS (reference)} \end{aligned}$$

3.2.1 Results

The coefficients from the first set of models constructed (Appendix F) indicates that the starting year of a hurricane is a significant predictor for β_1, at the 0.05 significance level. Since the coefficient associated with year in this model is negative, this model indicates that an increase in year negatively affects the autoregressive change in wind speed. As β_1 is a positive coefficient in our Bayesian model, we conclude that as year increases, the change of wind speed tends to slow down.

The coefficients in the second set of models (Appendix G) indicate that the spring season is a significant predictor for $\beta_0, \beta_1, \beta_4$. Spring is positively correlated with β_0, indicating that the baseline wind speed is higher in the spring than it is in the fall. In addition, wind speed in the spring tends to change slower in spring than fall, as spring has negative coefficients for both the β_1 and β_4 models. Similarly, the summer season produced a negative coefficient in the β_1 model.

The tropical storm (TS) type of hurricane was a significant predictor of β_3 in the second set of models. As β_3 is a negative value, we can say that there is a positive influence on the rate of wind speed change that is attributable to the change in longitude for tropical storm (TS) hurricanes. Similar to the findings for the first set of models, we found that year is a significant predictor of β_1 in the second set of models, and also produces a negative coefficient.

According to these results, though we have found some significant predictors of the estimated Bayesian model coefficients, we do not have enough evidence to conclude that hurricane wind speed has been increasing over the years.

4 Forecasting Hurricane Impact

We constructed models to predict the hurricane-induced damage and deaths, utilizing another dataset that contains summary features of 43 hurricanes and the estimated coefficients from the Bayesian model.

4.1 Data

This secondary dataset provides us with variables summarizing the trajectory of 43 of the hurricanes in the first dataset. These variables include measures such as the reported financial loss caused by the hurricane (damage), the reported number of deaths caused by the hurricane (deaths), the maximum and average recorded wind speed of the hurricane (max/meanspeed), the maximum and average recorded central pressure of the hurricane (max/meanpressure), the total population affected by the hurricane (total_pop), and the percentage of the affected population that resides in the United States (percent_usa). We are interested in using these variables, as well as the coefficients obtained in the MCMC process, to forecast the financial loss (damage) and loss of life (deaths) that a particular hurricane may cause.

4.2 Predicting Damages

We constructed a linear regression model to predict the hurricane-induced damage, and LASSO was used to select significant predictors for damage.

$$E(Y_i) = \gamma_0 + \gamma_1 \text{Season} + \gamma_2 \text{Maxspeed} + \gamma_3 \text{TotalPop} + \gamma_4 \text{PercentUSA}$$

4.2.1 Results

From the coefficients produced by the optimal LASSO model predicting hurricane damage (Appendix H (a)), it was determined that the variable, season, maxspeed, total_pop, and percent_usa were selected to be significantly associated with damage. After refitting the linear regression with the selected predictors, all predictors were found to be positively associated with damage (Appendix H (b)). This indicates that hurricanes with higher maximum wind speed, larger total affected population, or larger affected population that reside in the United States tend to result in larger financial loss, and hurricanes occurring in recent years tend to have larger financial impact.

4.3 Predicting Deaths

Since the number of deaths is a count variable, a Poisson regression was fit to predict the number of hurricane-induced deaths. An offset term, $log(hours)$, was incorporated in the Poisson model to adjust for exposure.

$$E(log(\lambda_i)) = X_i^T \gamma + log(hours_i)$$

4.3.1 Results

From the coefficients produced from the Poisson regression (Appendix I), all variables, with the exception of maxpressure and $\boldsymbol{\beta}_3$, have p values less than 0.05, indicating that these are all significantly associated with the number of deaths. For example, the coefficient of monthNovember is positive, meaning compared to August (the reference month), the average number of hurricane-induced deaths in November is larger, holding other variables fixed. The negative coefficient of natureTS suggests that the average number of deaths caused by Tropical Storm is lower than the average number of deaths caused by Disturbance (the reference nature), holding other variables fixed. Most β coefficients from the Bayesian model are found to be significant predictors for deaths. For example, β_0 is positively associated with deaths, indicating that an increase in β_0 is expected to cause an increase in the number of deaths. This also means that the larger the baseline wind speed of a hurricane, the larger the number of deaths caused by that hurricane. In addition, β_1 is also positively associated with deaths, suggesting that hurricanes that accelerate faster tend to cause more deaths, as affected populations may have less time to prepare for the hurricanes.

5 Discussion

5.1 Summary of Findings

In this analysis, we utilized Gibbs sampling, a Markov Chain Monte Carlo method, to numerically generate values from a given distribution. Though our final model does not reach perfect convergence for every parameter, we find that it produces decently accurate predictions, with an RMSE of 5.78 on the testing dataset.

We also analyzed the effect of hurricane type and starting time on Bayesian model's coefficients and found that year, season and hurricane type are significant predictors for some coefficients.

A linear model with lasso penalty and a Poisson model were further constructed to predict the hurricane-induced damage and deaths. We found year, maximum speed, total affected population and the proportion of US citizens in affected population are important variables in predicting financial loss induced by hurricanes. In addition, several coefficients of the Bayesian model, which can be regarded as characteristics of hurricanes, are significantly associated with mortality prediction.

5.2 Limitations and Strength

The Bayesian model with MCMC method allows for flexibility in modeling and the expression of our prior beliefs on quantities of interest. However, one critical limitation to the MCMC approach is that it is often computationally expensive, since it involves thousands of rounds of sampling and updating. In addition, convergence to a stationary distribution is not always guaranteed. Given inaccurate assumptions, it can be hard for the MCMC chain to converge. In our case, the assumption $\boldsymbol{\beta}_i \sim MVN(\boldsymbol{\mu}, \Sigma)$ might be too strong. In fact, the coefficients estimated by OLS for each hurricane shows a skewed normal pattern, which is in conflict with the above assumption (Appendix J).

5.3 Future Work

We expect that selecting a more adequate distributional assumption that would be able to account for skewness, such as a Lognormal distribution or a Gamma distribution, would allow our algorithm to perform better than it currently does under the multivariate normal distribution assumption on $\boldsymbol{\beta}_i$. In addition, the number of iterations in our Gibbs sampler was somewhat arbitrarily set to 10000. In the future we will try to use numerical standard errors to determine whether the loop should be terminated.

5.4 Group Contribution

All aspects of this project were conducted with contribution from every group member, through frequent in-person meetings and discussion. Paulina and Waveley directed discussions about each component of the project at a high-level. Paulina worked on deriving the distributions required for the MCMC algorithm. Haolin established code to clean, prepare, and split the datasets for analysis purposes. Jibei and Lin developed the code for our Gibbs sampler and Jibei took on the task of generating the chain. Lin and Yida established the models for the seasonal and impact forcasting analysis. Haolin worked on generating predictions of wind speed from the MCMC parameter estimates. Waveley and Haolin conducted an analysis comparing the parameter estimates derived from the MCMC process with parameter estimates derived from OLS. Everyone contributed to the report and presentation equally.

6 Reference

1. Livingston, I. (2021, September 3). Ida's impact from the Gulf Coast to northeast - by the numbers. The Washington Post. https://www.washingtonpost.com/weather/2021/09/03/hurricane-ida-numbers-surge-wind-pressure-damage/

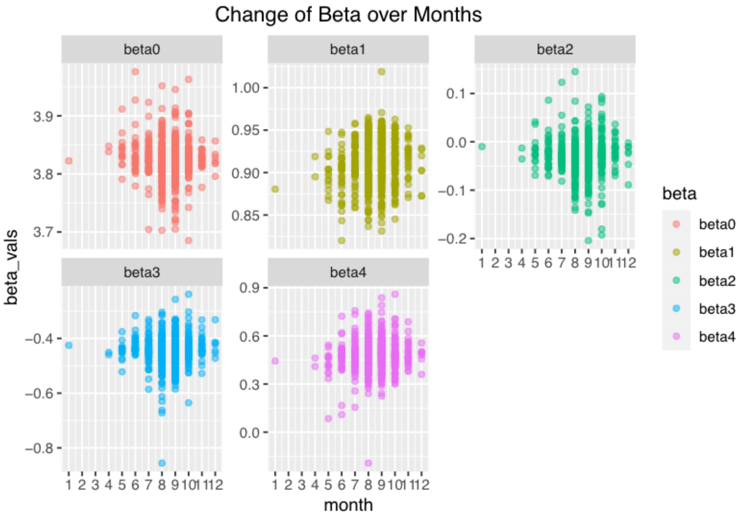
2. Saffir-Simpson Hurricane Wind Scale. (n.d.). https://www.nhc.noaa.gov/aboutsshws.php

7 Appendices

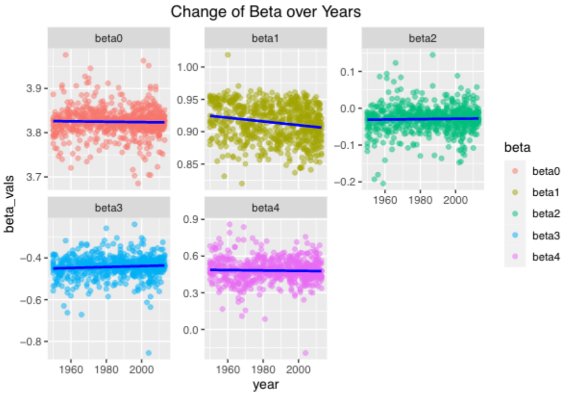
7.1 Appendix A: GitHub Repository

All code and associated files for this project can be found in this GitHub repository.

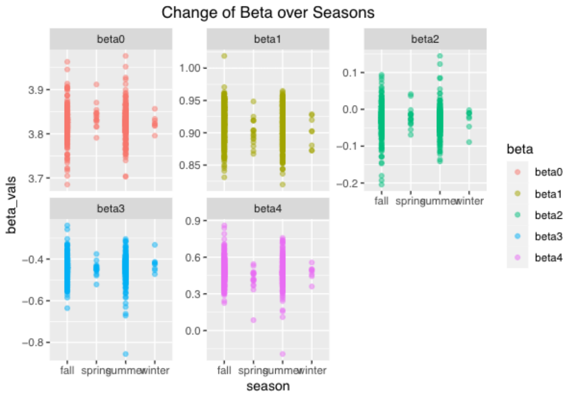
7.2 Appendix B: $\hat{\beta}_i$ by Month



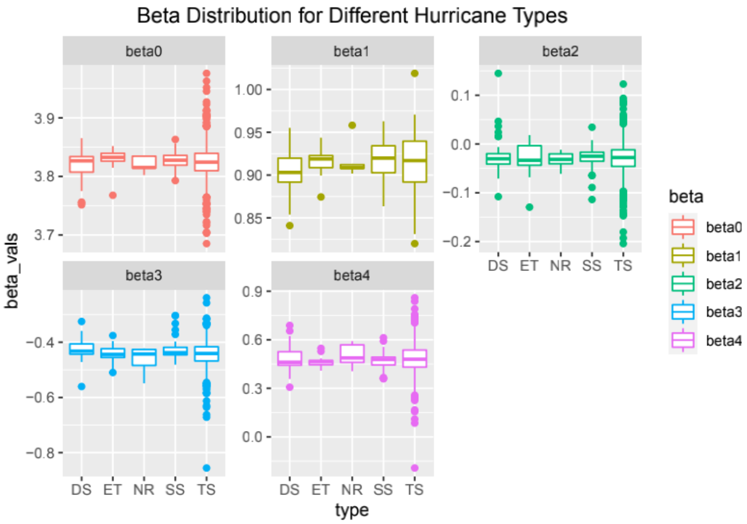
7.3 Appendix C: $\hat{\beta}_i$ by Year



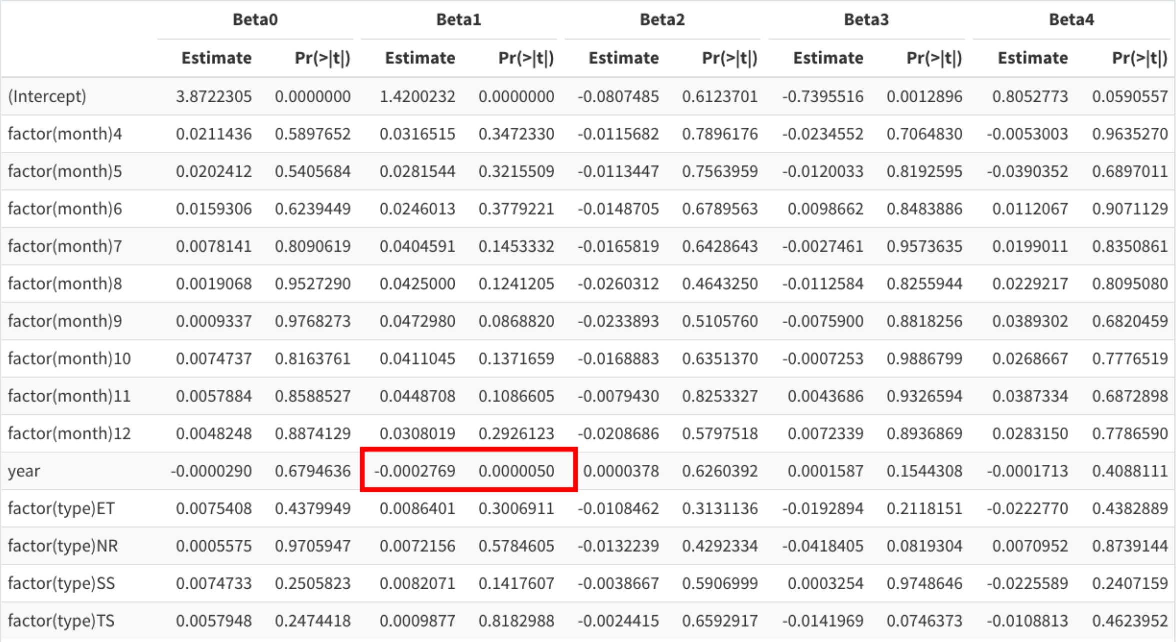
7.4 Appendix D: $\hat{\beta}_i$ by Season



7.5 Appendix E: $\hat{\beta}_i$ by Hurricane Type

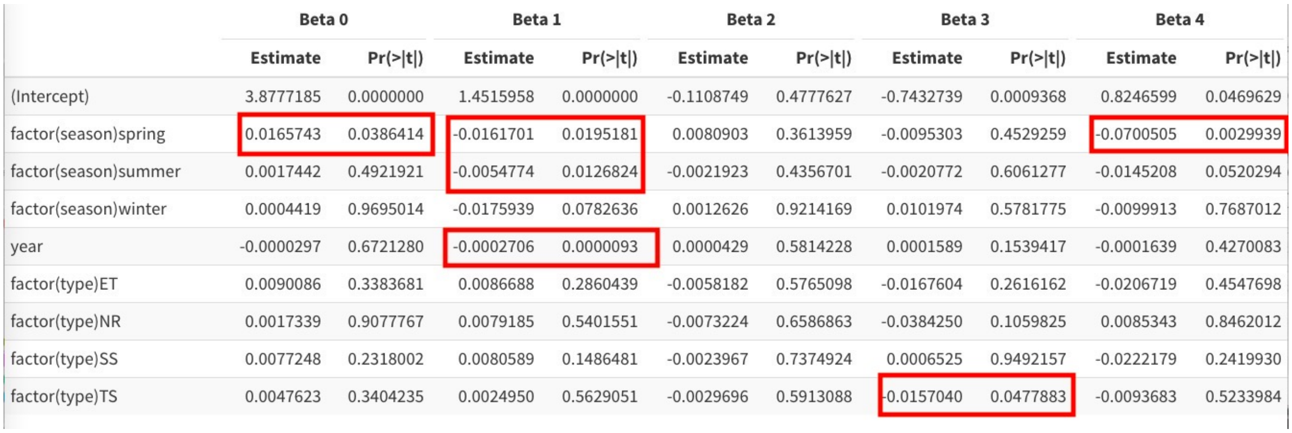


7.6 Appendix F: Coefficients for our first set of linear models



	Beta0		Beta1		Beta2		Beta3		Beta4	
	Estimate	Pr(>|t|)	Estimate	Pr(>|t|)	Estimate	Pr(>|t|)	Estimate	Pr(>|t|)	Estimate	Pr(>|t|)
(Intercept)	3.8722305	0.0000000	1.4200232	0.0000000	-0.0807485	0.6123701	-0.7395516	0.0012896	0.8052773	0.0590557
factor(month)4	0.0211436	0.5897652	0.0316515	0.3472330	-0.0115682	0.7896176	-0.0234552	0.7064830	-0.0053003	0.9635270
factor(month)5	0.0202412	0.5405684	0.0281544	0.3215509	-0.0113447	0.7563959	-0.0120033	0.8192595	-0.0390352	0.6897011
factor(month)6	0.0159306	0.6239449	0.0246013	0.3779221	-0.0148705	0.6789563	0.0098662	0.8483886	0.0112067	0.9071129
factor(month)7	0.0078141	0.8090619	0.0404591	0.1453332	-0.0165819	0.6428643	-0.0027461	0.9573635	0.0199011	0.8350861
factor(month)8	0.0019068	0.9527290	0.0425000	0.1241205	-0.0260312	0.4643250	-0.0112584	0.8255944	0.0229217	0.8095080
factor(month)9	0.0009337	0.9768273	0.0472980	0.0868820	-0.0233893	0.5105760	-0.0075900	0.8818256	0.0389302	0.6820459
factor(month)10	0.0074737	0.8163761	0.0411045	0.1371659	-0.0168883	0.6351370	-0.0007253	0.9886799	0.0268667	0.7776519
factor(month)11	0.0057884	0.8588527	0.0448708	0.1086605	-0.0079430	0.8253327	0.0043686	0.9326594	0.0387334	0.6872898
factor(month)12	0.0048248	0.8874129	0.0308019	0.2926123	-0.0208686	0.5797518	0.0072339	0.8936869	0.0283150	0.7786590
year	-0.0000290	0.6794636	-0.0002769	0.0000050	0.0000378	0.6260392	0.0001587	0.1544308	-0.0001713	0.4088111
factor(type)ET	0.0075408	0.4379949	0.0086401	0.3006911	-0.0108462	0.3131136	-0.0192894	0.2118151	-0.0222770	0.4382889
factor(type)NR	0.0005575	0.9705947	0.0072156	0.5784605	-0.0132239	0.4292334	-0.0418405	0.0819304	0.0070952	0.8739144
factor(type)SS	0.0074733	0.2505823	0.0082071	0.1417607	-0.0038667	0.5906999	0.0003254	0.9748646	-0.0225589	0.2407159
factor(type)TS	0.0057948	0.2474418	0.0009877	0.8182988	-0.0024415	0.6592917	-0.0141969	0.0746373	-0.0108813	0.4623952

7.7 Appendix G: Coefficients for our second set of linear models



	Beta 0		Beta 1		Beta 2		Beta 3		Beta 4	
	Estimate	Pr(>|t|)	Estimate	Pr(>|t|)	Estimate	Pr(>|t|)	Estimate	Pr(>|t|)	Estimate	Pr(>|t|)
(Intercept)	3.8777185	0.0000000	1.4515958	0.0000000	-0.1108749	0.4777627	-0.7432739	0.0009368	0.8246599	0.0469629
factor(season)spring	0.0165743	0.0386414	-0.0161701	0.0195181	0.0080903	0.3613959	-0.0095303	0.4529259	-0.0700505	0.0029939
factor(season)summer	0.0017442	0.4921921	-0.0054774	0.0126824	-0.0021923	0.4356701	-0.0020772	0.6061277	-0.0145208	0.0520294
factor(season)winter	0.0004419	0.9695014	-0.0175939	0.0782636	0.0012626	0.9214169	0.0101974	0.5781775	-0.0099913	0.7687012
year	-0.0000297	0.6721280	-0.0002706	0.0000093	0.0000429	0.5814228	0.0001589	0.1539417	-0.0001639	0.4270083
factor(type)ET	0.0090086	0.3383681	0.0086688	0.2860439	-0.0058182	0.5765098	-0.0167604	0.2616162	-0.0206719	0.4547698
factor(type)NR	0.0017339	0.9077767	0.0079185	0.5401551	-0.0073224	0.6586863	-0.0384250	0.1059825	0.0085343	0.8462012
factor(type)SS	0.0077248	0.2318002	0.0080589	0.1486481	-0.0023967	0.7374924	0.0006525	0.9492157	-0.0222179	0.2419930
factor(type)TS	0.0047623	0.3404235	0.0024950	0.5629051	-0.0029696	0.5913088	-0.0157040	0.0477883	-0.0093683	0.5233984

7.8 Appendix H: LASSO Coefficients and Refit Model

	Coefficients
(Intercept)	-533.5099174
season	3.3837824
deaths	0.0000000
monthJuly	0.0000000
monthJune	0.0000000
monthNovember	0.0000000
monthOctober	0.0000000
monthSeptember	0.0000000
natureNR	0.0000000
natureTS	0.0000000
maxspeed	1.2117851
meanspeed	0.0000000
maxpressure	0.0000000
meanpressure	0.0000000
hours	0.0000000
total_pop	0.3187361
percent_poor	0.0000000
percent_usa	0.7073409
beta0	0.0000000
beta1	0.0000000
beta2	0.0000000
beta3	0.0000000
beta4	0.0000000

(a)

	Coefficients
(Intercept)	-1316.7386136
season	0.6485139
maxspeed	0.1968674
total_pop	0.0000033
percent_usa	0.1356486

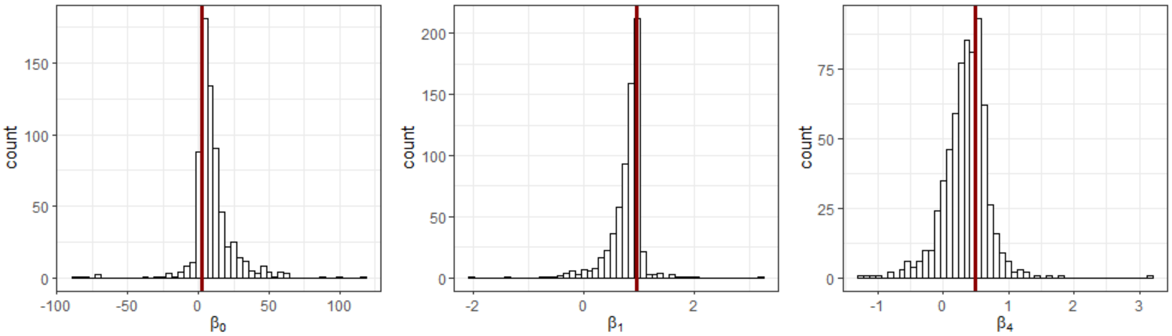
(b)

(a) Coefficients of the optimal LASSO model for hurricane-induced damage. (b) Coefficients of the refitted linear regression model for damage using selected predictors from LASSO.

7.9 Appendix I: Coefficients of the Poisson model for hurricane-induced deaths.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-199.5331639	11.8792784	-16.7967411	0.0000000
season	-0.0404185	0.0028048	-14.4104991	0.0000000
damage	0.0220163	0.0005679	38.7649762	0.0000000
monthJuly	-10.2286750	0.1604645	-63.7441688	0.0000000
monthJune	0.3928062	0.0989170	3.9710698	0.0000716
monthNovember	1.8733767	0.1664682	11.2536625	0.0000000
monthOctober	-1.6041896	0.0787720	-20.3649754	0.0000000
monthSeptember	1.2490015	0.0575033	21.7205350	0.0000000
natureNR	2.0903864	0.1371766	15.2386495	0.0000000
natureTS	-1.1903051	0.1118619	-10.6408484	0.0000000
maxspeed	0.0035207	0.0013988	2.5168778	0.0118400
meanspeed	-0.1978651	0.0039977	-49.4953412	0.0000000
maxpressure	0.0048106	0.0075485	0.6372945	0.5239331
meanpressure	0.0021204	0.0001759	12.0515409	0.0000000
total_pop	0.0000009	0.0000000	31.4237737	0.0000000
percent_poor	0.0873434	0.0010058	86.8433730	0.0000000
percent_usa	-0.0080185	0.0004884	-16.4173000	0.0000000
beta0	41.3531048	0.5634443	73.3934206	0.0000000
beta1	132.8784572	1.9305164	68.8305252	0.0000000
beta2	-10.7339527	0.5001340	-21.4621524	0.0000000
beta3	-0.4736994	0.5091748	-0.9303277	0.3522014
beta4	4.4919244	0.1971025	22.7897893	0.0000000

7.10 Appendix J: Distribution of each hurricane's coefficients obtained by OLS


count
β0
count
β1
count
β4