P8160 - Project 3
We're Not in Kansas Anymore:
An Application of the MCMC Algorithm to Hurricane Trajectory Data

Group 6

Amy Pitts, Jiacheng Wu, Jimmy Kelliher, Ruiqi Yan & Tianchuan Gao

2022-05-09

Abstract

This project studies the factors that impact the wind speed, death and damage of hurricanes in the North Atlantic area since 1950. Using Bayesian model fitted via an MCMC algorithm to estimate the effects of hurricane trajectory on wind speed, we find that increases in lagged wind speed and lagged change in wind speed correspond to a larger current wind speed, while increases in lines of latitude and longitude result in a decrease in wind speed. By regressing estimated random effects against the nature, month and year of a hurricane, we observe no significant seasonal difference in the effects, but the effect of lagged wind speed has decreased over time, and the wind speed of extratropical hurricanes tends to be more strongly correlated with lagged wind speed. We also find that estimated random effects, along with poverty, year, month, and nature, tend to be strong, statistically significant predictors of fatalities. Conversely, only some random effects, season, and month are associated with damage incurred by a hurricane. Generally, the predictability of a hurricane is a strong predictor of deaths. The estimated model based on the MCMC algorithm works well in terms of tracking most hurricanes, but the performance worsens when the sample size is not large enough. The methodology for tracking hurricanes with fewer records needs further investigation.

1. Introduction

1.1. Background

Hurricanes are tropical storms that reach a wind speed of 74 miles (119 kilometers) per hour or greater. They also bring heavy rain, thunder storms, and other severe meteorological phenomena. Throughout history, hurricanes have brought about huge death tolls and financial loss, especially in coastal cities and countries. Therefore, it is crucial for researchers to predict the behaviors of hurricanes, take appropriate counter measures, and minimize casualties. There is some literature [1] on forecasting hurricane trajectory, but there is an ongoing need to understand and employ these forecast in predicting and mitigating loss.

1.2. Objectives

For this project, we aim to examine the factors that could predict the wind speed of hurricanes, using the track data of 703 hurricanes in the North Atlantic area since 1950. The data includes each hurricane's geographical coordinates, maximum speeds recorded every 6 hours, the year (season) and month in which the hurricane took place, and its nature. We build a Bayesian model for each hurricane that predicts its wind speed at a given time based on its last recorded speed, and the change in longitude, latitude, and speed from the last observation. After deriving the posterior distributions, we use the Markov Chain Monte-Carlo Method to estimate the parameters. We then explore whether the season, month, and nature of the hurricane affect the wind speed and the impact of wind speed on death counts and financial loss.

2. Methods

2.1. Data Cleaning and Exploratory Analysis

There are 702 unique hurricanes in this dataset all the occurred in the North American region between the years 1950 to 2013. Data on the storm's location (longitude & latitude) and maximum wind speed were recorded every 6 hours. The number of observations we have for each storm range from 1 to 118 with a mean value of 31 observations per hurricane. Data is also collected on the storm's month, and the nature of the hurricane; (Extra-Tropical (ET), Disturbance (DS), Not Rated (NR), Sub Tropical (SS), and Tropical Storm (TS)).

To conduct our analysis we require at least 7 observations for each unique hurricane. In addition, we are only concerned about observations that occurred at 6-hour intervals. The dataset includes a couple of observations between the 6-hour periods. For our analysis, we are only going to include observations that are recorded on hours 0, 6, 12, and 18. In addition, we will exclude all hurricane IDs that have less than 7 observations. Through this process we remove 460 observations so we are left with 21578 observations and 681 unique hurricanes. In addition we also created variables of lag difference ($t-6$ to $t-12$) for latitude, longitude and wind speed as $\Delta_{i,1}(t), \Delta_{i,2}(t), \Delta_{i,3}(t)$ and lag of wind speed as $Y_i(t-6)$. The first two observations for each hurricane left are removed, because the change during previous time interval is not available for those observations, such that each hurricane has at least 5 observations.

Table 1: Data Characteristics shown by the Nature of the Hurricane variable

Characteristic	**Overall**, N = 21,578	**DS**, N = 963	**ET**, N = 2,149	**NR**, N = 96	**SS**, N = 750	**TS**, N = 17,620
Wind.kt	45 (30, 65)	25 (20, 30)	40 (30, 50)	25 (20, 25)	40 (30, 50)	50 (35, 70)
Latitude	26 (19, 34)	24 (17, 31)	44 (38, 50)	18 (14, 22)	32 (30, 36)	25 (18, 31)
Longitude	-64 (-78, -48)	-65 (-81, -45)	-46 (-62, -28)	-64 (-69, -55)	-65 (-74, -51)	-65 (-80, -51)
Month						
January	69 (0.3%)	5 (0.5%)	0 (0%)	0 (0%)	19 (2.5%)	45 (0.3%)
February	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
March	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)

Characteristic	**Overall**, N = 21,578	**DS**, N = 963	**ET**, N = 2,149	**NR**, N = 96	**SS**, N = 750	**TS**, N = 17,620
April	53 (0.2%)	0 (0%)	23 (1.1%)	0 (0%)	17 (2.3%)	13 (<0.1%)
May	274 (1.3%)	38 (3.9%)	18 (0.8%)	0 (0%)	57 (7.6%)	161 (0.9%)
June	795 (3.7%)	34 (3.5%)	113 (5.3%)	0 (0%)	58 (7.7%)	590 (3.3%)
July	1,478 (6.8%)	95 (9.9%)	103 (4.8%)	19 (20%)	41 (5.5%)	1,220 (6.9%)
August	5,101 (24%)	290 (30%)	319 (15%)	14 (15%)	108 (14%)	4,370 (25%)
September	8,810 (41%)	249 (26%)	854 (40%)	39 (41%)	160 (21%)	7,508 (43%)
October	3,717 (17%)	163 (17%)	547 (25%)	11 (11%)	146 (19%)	2,850 (16%)
November	1,047 (4.9%)	58 (6.0%)	158 (7.4%)	13 (14%)	86 (11%)	732 (4.2%)
December	234 (1.1%)	31 (3.2%)	14 (0.7%)	0 (0%)	58 (7.7%)	131 (0.7%)

2.2. Bayesian Model for Hurricane Trajectories

Climate researchers are interested in modeling the hurricane trajectories to forecast the winds peed. To model the wind speed of the i^{th} hurricane at time t we will use

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

Where $\Delta_{i,1}(t)$, $\Delta_{i,2}(t)$ and $\Delta_{i,3}(t)$ are changes in latitude longitude and wind speed respectively between $t-6$ and t. The $\epsilon_i(t) \sim \mathcal{N}(0, \sigma^2)$ are independent across t. Let $\beta_i = (\beta_{0,i}, \beta_{1,i}, \beta_{2,i}, \beta_{3,i}, \beta_{4,i})^T \sim \mathcal{N}(\mu, \Sigma)$ be multivariate normal distribution where $\mu \in \mathbb{R}^d$ and $\Sigma \in \mathbb{R}^{d \times d}$. For this model we will be assuming non-informative or weak prior distributions for our unknown parameters σ^2, μ and Σ .

$$\pi(\sigma^2) \propto \frac{1}{\sigma^2}, \qquad \pi(\mu) \propto 1, \qquad \pi\left(\Sigma^{-1}\right) \propto |\Sigma|^{-(d+1)} \exp\left\{-\frac{1}{2}\Sigma^{-1}\right\}$$

Our goal is to estimate $\Theta = (\boldsymbol{B}, \boldsymbol{\mu}, \Sigma^{-1}, \sigma^2)$. To do this we need to establish our likelihood and prior functions. Since this is a Bayesian model we have that the likelihood will be expressed as

$$L(\boldsymbol{Y} \mid \boldsymbol{\Theta}) \propto \prod_{i=1}^{m} \left(\sigma^2\right)^{-\frac{n_i}{2}} \exp\left\{-\frac{1}{2\sigma^2}\left(\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta}_i\right)^T\left(\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta}_i\right)\right\}$$

where m is the number of hurricane and n_i is the number of observations for i^{th} hurricane. The joint prior distribution is expressed as

$$\pi\left(\boldsymbol{\Theta}\right) \propto \left(\sigma^2\right)^{-1}\left|\Sigma^{-1}\right|^{d+1} \exp\left\{-\frac{1}{2}\operatorname{tr}\left(\Sigma^{-1}\right)\right\} \prod_{i=1}^{m}\left|\Sigma^{-1}\right|^{\frac{1}{2}} \exp\left\{-\frac{1}{2}\left(\boldsymbol{\beta}_i - \boldsymbol{\mu}\right)^T \Sigma^{-1}(\boldsymbol{\beta}_i - \boldsymbol{\mu})\right\}$$

where d is the dimension of $\boldsymbol{\mu}$. Using the likelihood and priors derived above we can calculate the posterior distribution.

$$\begin{aligned}
\pi(\boldsymbol{\Theta} \mid \boldsymbol{Y}) \propto{} & \left(\sigma^2\right)^{-1}\left|\Sigma^{-1}\right|^{d+1} \exp\left\{-\frac{1}{2} + \operatorname{tr}\left(\Sigma^{-1}\right)\right\} \\
& \times \prod_{i=1}^{m}\left(\sigma^2\right)^{-\frac{n_i}{2}}\left|\Sigma^{-1}\right|^{\frac{1}{2}} \exp\left\{-\frac{1}{2\sigma^2}\left(\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta}_i\right)^T\left(\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta}_i\right)\right\} \exp\left\{-\frac{1}{2}\left(\boldsymbol{\beta}_i - \boldsymbol{\mu}\right)^T\Sigma^{-1}\left(\boldsymbol{\beta}_i - \boldsymbol{\mu}\right)\right\} \\
& = \left(\sigma^2\right)^{-\left(1+\frac{\sum_{i=1}^{m} n_i}{2}\right)}\left|\Sigma^{-1}\right|^{d+1+\frac{m}{2}} \exp\left\{-\frac{1}{2}\operatorname{tr}\left(\Sigma^{-1}\right)\right\} \exp\left\{-\frac{1}{2}\sum_{i=1}^{m}\left(\boldsymbol{\beta}_i - \boldsymbol{\mu}\right)^T\Sigma^{-1}\left(\boldsymbol{\beta}_i - \boldsymbol{\mu}\right)\right\} \\
& \times \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{m}\left(\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta}_i\right)^T\left(\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta}_i\right)\right\}
\end{aligned}$$

Due to the number of unknown parameters and the complexity we will need to use method to approximate this distribution. Before describing that process we first need to calculate the conditional posterior distribution for each unknown parameter.

$$
\begin{aligned}
&\boldsymbol{\beta}_i : \pi(\boldsymbol{\beta}_i \mid \boldsymbol{\Theta}_{(-\boldsymbol{\beta}_i)}\boldsymbol{Y}) \propto \exp\left\{-\frac{1}{2}\left(\boldsymbol{\beta}_i - \boldsymbol{\mu}\right)^T \Sigma^{-1}\left(\boldsymbol{\beta}_i - \boldsymbol{\mu}\right) - \frac{1}{2\sigma^2}\left(\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta}_i\right)^T\left(\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta}_i\right)\right\} \\
&\boldsymbol{\mu} : \pi\left(\boldsymbol{\mu} \mid \boldsymbol{\Theta}_{(-\boldsymbol{\mu})}, \boldsymbol{Y}\right) \sim N(\bar{\boldsymbol{\beta}}, \Sigma/m) \\
&\sigma^2 : \pi\left(\sigma^2 \mid \boldsymbol{\Theta}_{(-\sigma^2)}, \boldsymbol{Y}\right) \propto \left(\sigma^2\right)^{-\left(1+\frac{\sum_{i=1}^{m} n_i}{2}\right)} \times \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{m}\left(\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta}_i\right)^T\left(\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta}_i\right)\right\} \\
&\Sigma^{-1} : \pi\left(\Sigma^{-1} \mid \boldsymbol{\Theta}_{(-\Sigma^{-1})}, \boldsymbol{Y}\right) \sim \text{Wishart}\left(3d+3+m, \left(\boldsymbol{I} + \sum_{i=1}^{m}\left(\boldsymbol{\beta}_i - \boldsymbol{\mu}\right)\left(\boldsymbol{\beta}_i - \boldsymbol{\mu}\right)^T\right)^{-1}\right)
\end{aligned}
$$

To see a more detailed description of each conditional posterior please see Appendix A.

2.3. Markov Chain Monte Carlo (MCMC) Algorithm

Due to the complexity of the above posterior distribution, we will use a Markov chain Monte Carlo (MCMC) process. Since we could generate full conditional posterior distribution for some parameters but not all we will instead apply a hybrid algorithm consisting of Metropolis-Hastings (MH) steps and Gibbs steps. We will perform a MH step for $\beta_{j,i}$, σ^2 and sample $\Sigma^{-1^{(t+1)}}$ from Wishart distribution. We will also describe a Gibbs step for μ using the $\beta_{j,i}$ gathered.

For each iteration, we start from random walk MH step for $\beta_{j,i}$. Sampling proposed $\beta'_{j,i}$, $j \in 0, 1...4$ for i^{th} hurricane from proposal distribution $U\left(\beta_{j,i}^{(t)} - a_{j,i}, \beta_{j,i}^{(t)} + a_{j,i}\right)$, where $a_{j,i}$ is the search window for $\beta_{j,i}$. Since the proposed distribution is symmetric, the accepting or rejecting the proposed $\beta'_{j,i}$ depends on the ratio of posterior distribution. Some of the parameters in $\boldsymbol{\Theta}$ could be cancelled out and the ratio simplified to be:

$$
\frac{\pi\left(\boldsymbol{\beta}'_i, \boldsymbol{\Theta}^{(t)}_{(-\boldsymbol{\beta}_i)} \mid \boldsymbol{Y}\right)}{\pi\left(\boldsymbol{\beta}^{(t)}_i, \boldsymbol{\Theta}^{(t)}_{(-\boldsymbol{\beta}_i)} \mid \boldsymbol{Y}\right)} = \frac{\exp\left\{-\frac{1}{2\sigma^{2(t)}}\left(\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta}'_i\right)^T\left(\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta}'_i\right) - \frac{1}{2}\left(\boldsymbol{\beta}'_i - \boldsymbol{\mu}^{(t)}\right)^T \Sigma^{-1^{(t)}}\left(\boldsymbol{\beta}'_i - \boldsymbol{\mu}^{(t)}\right)\right\}}{\exp\left\{-\frac{1}{2\sigma^{2(t)}}\left(\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta}^{(t)}_i\right)^T\left(\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta}^{(t)}_i\right) - \frac{1}{2}\left(\boldsymbol{\beta}^{(t)}_i - \boldsymbol{\mu}^{(t)}\right)^T\Sigma^{-1(t)}\left(\boldsymbol{\beta}^{(t)}_i - \boldsymbol{\mu}^{(t)}\right)\right\}}
$$

where $\boldsymbol{\beta}'_i$ consisting with $\beta'_{j,i}$, $\beta^{(t)}_{k,i}$ for $k > j$ and $\beta^{(t+1)}_{k,i}$ for $k < j$. The log of the ratio is

$$
\begin{aligned}
\log\frac{\pi\left(\boldsymbol{\beta}'_i, \boldsymbol{\Theta}^{(t)}_{(-\boldsymbol{\beta}_i)} \mid \boldsymbol{Y}\right)}{\pi\left(\boldsymbol{\beta}^{(t)}_i, \boldsymbol{\Theta}^{(t)}_{(-\boldsymbol{\beta}_i)} \mid \boldsymbol{Y}\right)} = &-\frac{1}{2}\left(\frac{\left(\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta}'_i\right)^T\left(\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta}'_i\right)}{\sigma^{2(t)}} + \left(\boldsymbol{\beta}'_i - \boldsymbol{\mu}^{(t)}\right)^T \Sigma^{-1^{(t)}}\left(\boldsymbol{\beta}'_i - \boldsymbol{\mu}^{(t)}\right)\right) \\
&+\frac{1}{2}\left(\frac{\left(\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta}^{(t)}_i\right)^T\left(\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta}^{(t)}_i\right)}{\sigma^{2(t)}} + \left(\boldsymbol{\beta}^{(t)}_i - \boldsymbol{\mu}^{(t)}\right)^T \Sigma^{-1^{(t)}}\left(\boldsymbol{\beta}^{(t)}_i - \boldsymbol{\mu}^{(t)}\right)\right)
\end{aligned}
$$

Then we randomly sample u from $U(0,1)$ and compare $\log(u)$ with the log ratio. If the $\log(u)$ is smaller, we accept $\beta'_{j,i} = \beta^{(t+1)}_{j,i}$, otherwise we reject $\beta'_{j,i}$ and $\beta^{(t)}_{j,i} = \beta^{(t+1)}_{j,i}$.

Then, Gibb step for $\boldsymbol{\mu}$: Sample $\boldsymbol{\mu}^{(t+1)}$ from $\mathcal{N}\left(\bar{\boldsymbol{\beta}}^{(t+1)}, \Sigma^{(t)}/m\right)$, where $\bar{\boldsymbol{\beta}}^{(t+1)}$ is the average $\boldsymbol{\beta}^{(t+1)}_i$ over all hurricanes.

Next, is the random walk MH step to generate $\sigma^{2'}$ with step size from $U\left(-a_{\sigma^2}, a_{\sigma^2}\right)$. Firstly, check whether $\sigma^{2'}$ is positive, if not, we reject $\sigma^{2'}$. Then, we randomly sample u from $U(0,1)$ and compare $\log(u)$ with the log posterior ratio. The log posterior ratio

$$\log \frac{\pi\left(\sigma^{2'}, \boldsymbol{\Theta}^{(t)}_{(-\sigma^2)} \mid \boldsymbol{Y}\right)}{\pi\left(\sigma^{2^{(t)}}, \boldsymbol{\Theta}^{(t)}_{(-\sigma^2)} \mid \boldsymbol{Y}\right)} = -\left(1+\frac{M}{2}\right)\log(\sigma^{2'}) - \frac{1}{2\sigma^{2'}}\sum_{i=1}^{m}\left(\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta}_i^{(t+1)}\right)^T\left(\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta}_i^{(t+1)}\right)$$

$$+\left(1+\frac{M}{2}\right)\log(\sigma^{2^{(t)}}) + \frac{1}{2\sigma^{2^{(t)}}}\sum_{i=1}^{m}\left(\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta}_i^{(t+1)}\right)^T\left(\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta}_i^{(t+1)}\right)$$

where M is total number of observation for all hurricanes. If the $\log(u)$ is smaller, we accept $\sigma^{2'} = \sigma^{2^{(t+1)}}$, otherwise, $\sigma^{2^{(t+1)}} = \sigma^{2^{(t)}}$

Finally, we sample $\Sigma^{-1^{(t+1)}}$ from Wishart $\left(n = 3d+3+m, V = \left(\boldsymbol{I} + \sum_{i=1}^{m}\left(\boldsymbol{\beta}_i^{(t+1)} - \boldsymbol{\mu}^{(t+1)}\right)\left(\boldsymbol{\beta}_i^{(t+1)} - \boldsymbol{\mu}^{(t+1)}\right)^T\right)^{-1}\right)$.

For the MCMC process we will run 10,000 iterations using the code shown in Appendix B.

2.4. Initial Starting Values for MCMC

To initiate the MCMC process we need to specify the starting values. The choice of starting value is important to help with the convergence time of our algorithm. For each parameter we will chose the start values to be:

- $\boldsymbol{\beta}_i$: Fit OLS multivariate linear regression (MLR) for i^{th} hurricane and use the coefficients as $\boldsymbol{\beta}_i^{(0)}$
- $\boldsymbol{\mu}$: Average over all $\boldsymbol{\beta}_i^{(0)}$ as $\boldsymbol{\mu}^{(0)}$
- σ^2: $\hat{\sigma}_i^2$ is the mean square residuals of the OLS model for i^{th} hurricane. Take the mean over all $\hat{\sigma}_i^2$ as $\sigma^{2^{(0)}}$
- Σ^{-1}: Generate the covariance matrix of $\boldsymbol{\beta}_i^{(0)}$ and take the inverse of the matrix as $\Sigma^{-1^{(0)}}$

There is a MH step in the MCMC algorithm so the choice of search window is important. The acceptance rate of the MH step is around 0.317 to 0.637, so the search window of our algorithm is appropriate. We tune the search windows multiple times to achieve this result. Table 1 demonstrates the range of search windows a and associated acceptance rates for $\boldsymbol{\beta}$ and σ^2.

Table 2: Range of Search Window and Acceptance Rate for paraemters used MH step

	Search Window	Acceptance Rate (%)
$\boldsymbol{\beta}_{0,}$	1.1	45.87 - 51.36
$\boldsymbol{\beta}_{1,}$	(0.04, 0.1)	31.67 - 63.68
$\boldsymbol{\beta}_{2,}$	(0.8, 1.0)	38.60 - 45.60
$\boldsymbol{\beta}_{3,}$	(0.5, 0.6)	33.20 - 61.32
$\boldsymbol{\beta}_{4,}$	(0.3, 0.4)	34.95 - 60.45
σ^2	2.0	44.83

2.5. MCMC Model Performance

To assess the quality of the proposed model we will assess the overall adjusted R^2 value from the Bayesian model. In addition the model performance will be assessed for each hurricane by the adjusted R^2 value as well as a goodness of fit test. The goodness of fit test we will use residuals of Bayesian estimates, r_{ik}, of the k^{th} observation in i^{th} hurricane, to calculate the test statistics. $\chi^2_{stat} = \frac{\sum_{j=1}^{n_i} r_{ij}^2}{\sigma^2}$, where σ^2 is the estimate σ^2 from MCMC. Based on the normal assumption in intro, $\chi^2_{stat} \sim \chi^2_{n_i}$, where n_i is the number of observation for i^{th} hurricane. Visual inspection can also be used for each individual hurricane plotting their observed wind speed with the model's predicted wind speed.

2.6. Models to Explore Seasonal Differences and Yearly Trends

In the dataset information about the hurricanes start month, year, and the nature of the hurricane is recorded. It is interesting to explore how these three variables affect wind speed. Specifically, we are interested in exploring the seasonal differences, and if there is any evidence supporting the statement that "the hurricane wind speed has been increasing over years". To do visual inspection can first be used to help us understand how these three variables impact the β estimates. Then we can see if there is a linear trend given the three variables using each β estimate gained from the Bayesian model described above. For each β value we fit a different linear model. We can first fit a model using the three variables (month, year, and nature) as predictors and the 5 different β values as the outcome. This will result in 5 different linear models, one for each β value.

$$Y_{ji} = \alpha_{0j} + \alpha_{1j} \times \text{Decade}_i + \alpha_{(k+1)j} I(\text{Nature} = k)_i + \alpha_{(l+5)j} I(\text{Month} = l)_i + \epsilon_{ji}$$

Where i is the hurricane, j is the Beta model (0 through 4), $k \in (ET,\ NR,\ SS,\ TS)$ making DS the reference group. Let l in (April - December) where January is the reference group. We chose to include nature and month as categorical variables and year (transformed into decade) as a continuous variable.

We are also going to consider 5 different models just using decade as a predictor. $Y_i = \alpha_{0i} + \alpha_{1i} \times \text{Decade}$ where Y_i is each β_i and $i \in (0, \ldots, 4)$.

2.7. Models to Explore Death and Damages

Using cross-sectional data for $n = 43$ distinct hurricanes pertaining to damage (in billions of US dollars), deaths, and possible features of interest, we seek to build a model that can predict the destruction of a hurricane and offer insight into inferential relationships between selected features, damage, and deaths.

Features include the season and nature of each hurricane, the total population of the affected area, the fraction of poor residents in the affected area, as well as coefficients $\boldsymbol{\beta}$ estimated via our MCMC algorithm. Because both outcomes (i.e., damage and deaths) are non-negative, right-skewed, and subject to an exposure characterized by population, we naturally consider Poisson regression with a log link and a population offset. In particular, for each hurricane $i \in \{1, \ldots, n\}$, we assume outcome $y_i \sim \text{Poisson}(\theta_i)$, such that

$$\log(\theta_i) = \log(m_i) + \mathbf{x}_i' \boldsymbol{\gamma} + \boldsymbol{\beta}_i' \boldsymbol{\delta},$$

where m_i is the total population of the affected area, $\mathbf{x}_i \in \mathbb{R}^p$ is a vector of observed features, and $\boldsymbol{\beta}_i \in \mathbb{R}^5$ is a vector of estimated random effects from the MCMC algorithm.

Because $p = 8$ in the dataset provided, the dimension of our parameter space is large relative to n, which is small. As such, we consider a penalized lasso Poisson regression to select only the most important features. To identify the optimal penalty, we employ a leave-one-out cross-validation procedure, which is feasible for small n. This procedure is also desirable in that there is only one unique way to assign the n folds, and hence the model selection is non-random. Our selection criterion will be the minimizer of the average Poisson deviance, as more familiar metrics like the linear correlation coefficient and the AUC are not appropriate for Poisson regression.

Finally, we employ a bootstrap smoothing procedure due to Efron [2] in order to consider post-selection inference. Generally, inference is not tenable when considering penalized regression, but we can estimate standard errors for our constrained point estimates via bootstrapping. In particular, the algorithm proceeds as follows.

1. Estimate the *unconstrained* $\hat{\theta} \equiv (\hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\delta}})$ via Poisson regression for the full model.
2. For each bootstrap sample $b \in \{1, \ldots, B\}$ with $B = 5{,}000$, randomly draw $y_{bi} \sim \text{Poisson}\big(\log(m_i) + \mathbf{x}_i' \hat{\gamma} + \boldsymbol{\beta}_i' \hat{\boldsymbol{\delta}}\big)$ for each hurricane $i \in \{1, \ldots, n\}$.
3. For each bootstrap sample $b \in \{1, \ldots, B\}$, estimate the *constrained* $\check{\theta}_b \equiv (\check{\boldsymbol{\gamma}}_b, \check{\boldsymbol{\delta}}_b)$ via Poisson lasso regression using $\mathbf{y}_b \equiv (y_{b1}, \ldots, y_{bn})$ as the smoothed outcome.

4. Compute the empirical standard deviation of the bootstrap estimates $\check{\theta}_1, \ldots \check{\theta}_B$.

Implicitly, the component-wise empirical standard deviation of our bootstrap estimates is an application of the delta method. From these bootstrapped standard errors, we can then construct p-values and confidence intervals around our point estimates from the previous analysis.

3. Results

3.1 MCMC Convergence and Estimates

Using the starting values describe above, we get 10000 samples for each parameters. We are particular interested in the parameter convergence. Below we can see on Figure 1 tracing 10000 samples for selected parameters. Each parameter reaches a convergence after 5,500 runs. Therefore, we will use 5500 as our burn-in period. Figure 2 displays the histogram of the selected parameters after burn in. We see that each of the distributions is relatively normal with some skewed results for Σ^{-1} and heavy tails in μ_0.

Then, we can take the average estimate excluding the burn in period. We could visualize estimate of $\beta_{j,i}$ for each hurricane by each j on Figure 3. j represents the index of co-variate in the model. Figure 4 displays the μ_j estimates, σ^2 estimates, estimated Σ^{-1} matrix as well as the correlation matrix ρ derived from estimated Σ^{-1}. The estimated correlation is not greater 0.1, implying that the correlation among β_j, is weak. The average of β_j. estimate is same as μ_j estimate displayed by the vertical line of the corresponding histogram on Figure 3.

The μ_0 estimate is positive and represents the average wind speed at the start of the hurricane, so the positive value is sensible. The μ_1 estimate is positive, so it indicated that a larger previous wind speed causes a higher wind speed for the next time point. The μ_2 estimate represents the average impact of the change in latitude, and the μ_3 estimate represents the average impact of the change in longitude. Both of the impact of change in latitude and longitude have negative values indicating that traveling further to north and east in direction corresponds with a decrease in wind speeds. The μ_4 estimate is 0.481 which represents the average impact of the change in wind speed. This positive value indicated that a larger change between previous time slots is associated with a larger current wind speed.

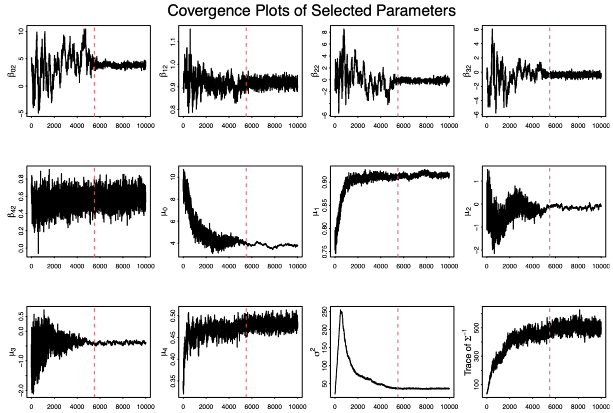


Figure 1: Convergence Plot of the Selected Parameters after 10,000 Iterations

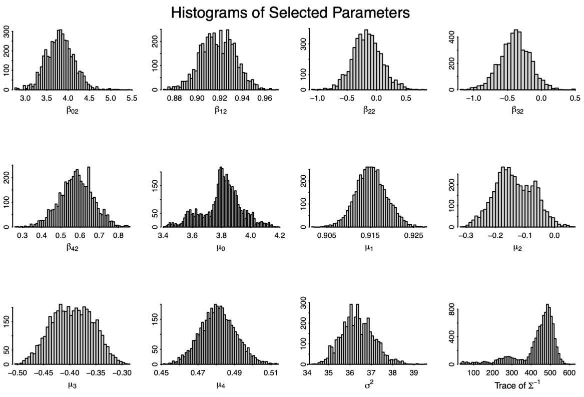


Figure 2: Histogram Plot of the Selected Parameters not including the 5,500 burn in period results

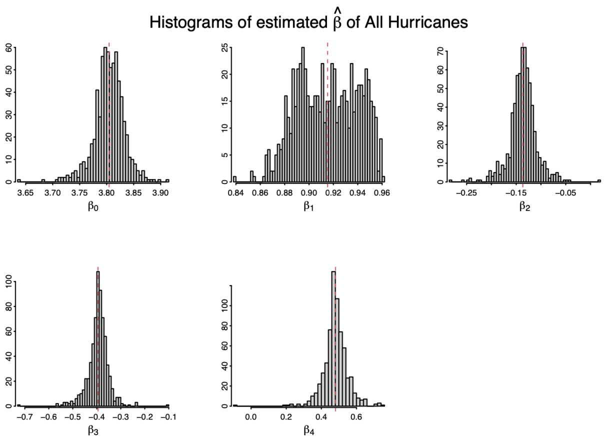


Figure 3: Bayesian Estiamtes for the β values for all Hurricane

	μ_0	μ_1	μ_2	μ_3	μ_4	σ^2	σ^2_{00}	σ^2_{11}	σ^2_{22}	σ^2_{33}	σ^2_{44}
Estimates	3.8	0.92	-0.14	-0.39	0.48	36.36	0.063	0.003	0.047	0.042	0.018

$$\hat{\Sigma}^{-1} = \begin{bmatrix} 16.07 & 7.63 & 0.34 & -1.78 & 0.47 \\ 7.63 & 381.32 & 5.39 & 3.96 & -6.32 \\ 0.34 & 5.39 & 21.43 & 1.48 & 1.14 \\ -1.78 & 3.96 & 1.48 & 24.35 & -3.3 \\ 0.47 & -6.32 & 1.14 & -3.3 & 55.12 \end{bmatrix} \quad \hat{\rho} = \begin{bmatrix} 1 & -0.101 & -0.018 & 0.094 & -0.011 \\ -0.101 & 1 & -0.057 & -0.043 & 0.043 \\ -0.018 & -0.057 & 1 & -0.067 & -0.042 \\ 0.094 & -0.043 & -0.067 & 1 & 0.089 \\ -0.011 & 0.043 & -0.042 & 0.089 & 1 \end{bmatrix}$$

Figure 4: Bayesian Estiamtes for μ and σ^2

3.2. MCMC Model Performance

The overall adjusted R^2 of the estimated Bayesian model is 0.9524156. We can also look at the adjusted R^2 value for each hurricane. Table 2 displays the values grouped by the adjusted R^2 showing the number of hurricanes and the percentage. We see that most hurricanes do well with an adjusted R^2 above 0.6. Even though most of the estimated models track hurricanes well with great adjusted R^2, some estimated models track the hurricanes extremely bad. 23.3% of the estimated models do not track the hurricane well and have adjusted R^2 less than 0.6. Few models have negative adjusted R^2. We also perform the goodness-of-fit test for the individual estimated model. The estimated models of 88 hurricanes have a p-value less than 0.05, implying that those models do not fit well.

Figure 5 shows the adjusted R^2 for each hurricane by the number of observations. This plot shows that with more observations the better our model with do. The to the right of the vertical line shows that 40 or more observations give us adjusted R^2 values above 0.5.

Figure 6 displays 9 randomly selected hurricanes actual wind speeds and predicted wind speeds. We see that for all of them the lines are very similar with some slight derivations away when the number of observations are small. In DOG 1950 we see very good model prediction and for this particular hurricane we have a lot of observations. However, for JERRY 2007 we do not have as many observations and see deviations at the begging and end of the storm.

Table 3: R^2_{adj} for each hurricane

R^2_{adj}	Count of Hurricanes	Percentage(%)
0.6-1	522	76.7
0.2-0.6	79	11.6
< 0.2	80	11.7

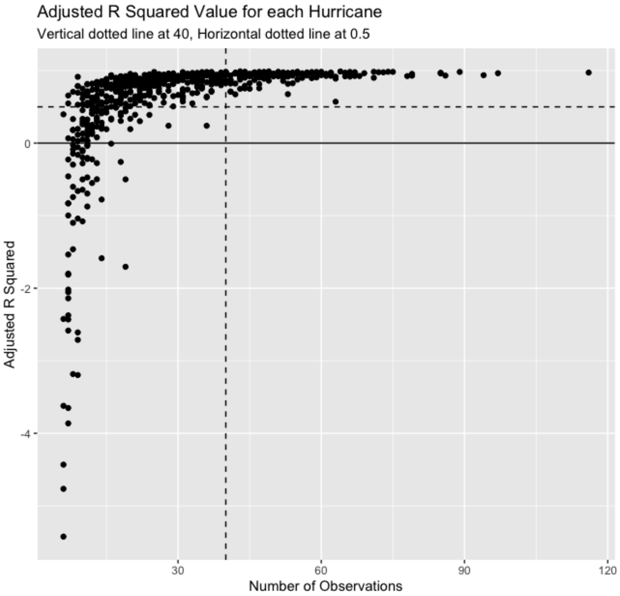


Figure 5: Adjusted R squared for each Hurricane plotted by the number of observations.

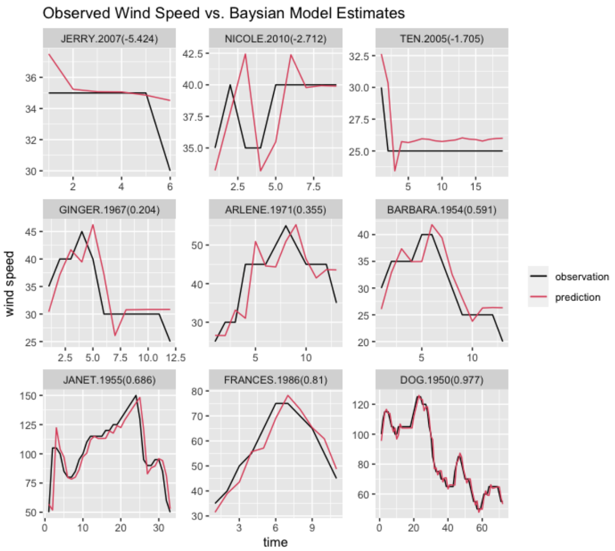


Figure 6: The predicted wind speeds verses the actual observed wind speeds for 9 randomly selected hurricanes.

3.3. Seasonal Differences and Yearly Trends Results

Using the model output obtained in the Bayesian MCMC algorithm we can now explore how the hurricane's start month, nature, and year affect the wind speed. Particularly we are interested in if there are any seasonal differences and if there is evidence to support the statement that "the hurricane wind speed has been increasing over years".

We can first explore if there are the seasonal difference. Visually inspecting the β values obtained from the MCMC process with respect to year, month and nature we can look at Figure 7 and 8. Looking at the beta estimates over the years grouped by month and the βs, Figure 7, we can see that the majority of the hurricanes fall within the June through November months. We can also see that looking at December in the later years there seem to be more Hurricanes. This time spent is what is conventionally accepted as hurricane season. Within each month there does not seem to be any obvious change in the β estimates over the years. Looking at the Nature of the hurricane by the β's box plot, Figure 8, we see that all the box plots for the nature seem to be relatively similar within each β estimate. This indicates that there don't seem to be drastic differences between the 5 different hurricane natures. However, looking at the β_1 section we see that the Extra-Tropical (ET) box seems to be much larger than the disturbance (DS), not rated (NR), and subtropical (SS) nature types. This visually implies that the extratropical hurricane has increased the importance of previous time wind speed impacting the next wind speed prediction.

Fitting the linear model using decade, month, and nature as predictors results in 5 different models, one for each β vector. The model output can be seen in Appendix C, due to the length it was not included in the main body. To summarize the results no nature, or month indicators were significant. The β_1 model had the decade estimates as significant -0.003 (-0.002, -0.004). Thus month and nature don't have a linear association within each $\boldsymbol{\beta}_j$ vector when adjusting for the year.

Just exploring the decade variable we can address the question of does "the hurricane wind speed has been increasing over years"? To do this we can do a visual inspection as well as a linear association test. Figure 9 explore this question by showing a visual inspection on the left and a linear model on the right. The model output just shows the estimates related to the decade using each $\boldsymbol{\beta}_j$ vector as an outcome. The results are the β_1 and β_3 decade estimates are the only significant ones. The β_1 decade estimates indicate a decrease in the change of wind speed over years. The β_3 decade estimates Indicate an increase in the impact of change in longitude over years. None of the other decade estimates are significant.

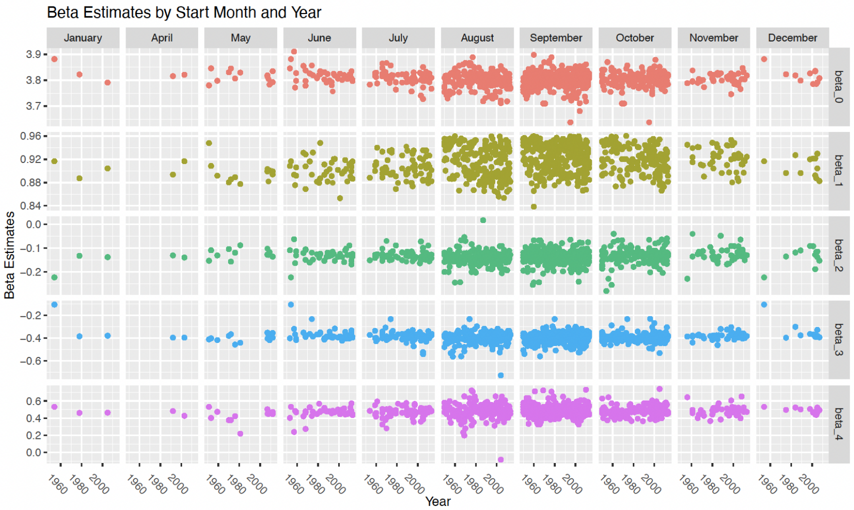


Figure 7: Exploring the Beta estimates by month and year.

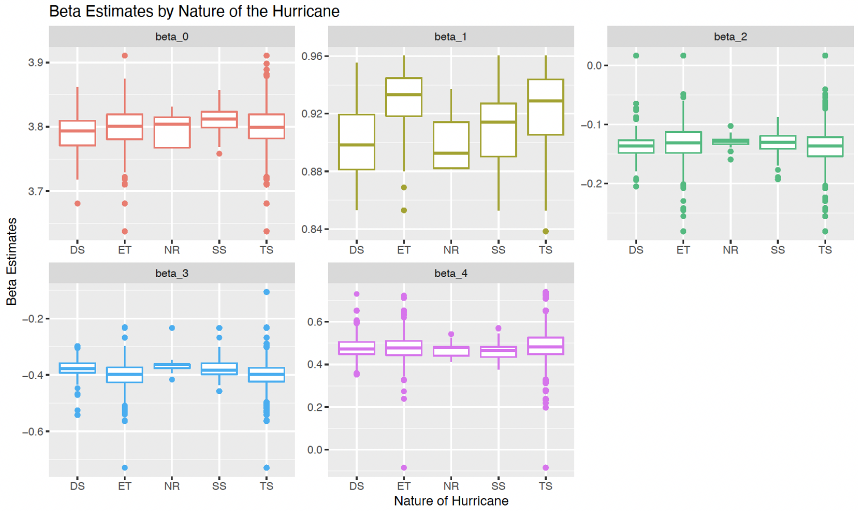
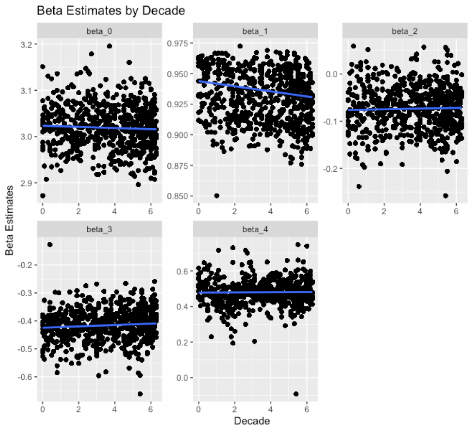
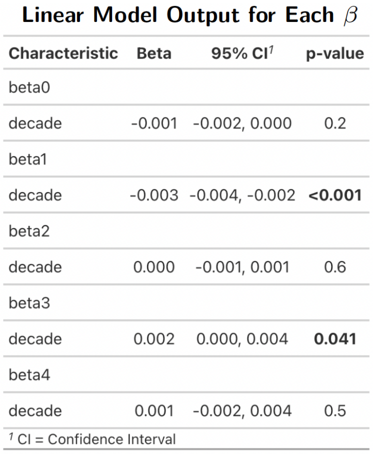


Figure 8: Exploring the Beta estimates by nature of the hurricanes





Linear Model Output for Each β

Characteristic	Beta	95% CI[1]	p-value
beta0			
decade	-0.001	-0.002, 0.000	0.2
beta1			
decade	-0.003	-0.004, -0.002	**<0.001**
beta2			
decade	0.000	-0.001, 0.001	0.6
beta3			
decade	0.002	0.000, 0.004	**0.041**
beta4			
decade	0.001	-0.002, 0.004	0.5

[1] CI = Confidence Interval

Figure 9: Exploring how wind speed has changed over the years

3.5 Death and Damages Results

Upon executing our cross-validated penalized Poisson regression for the death outcome, we select thirteen features. The deviance of this model and that of other specifications is summarized in the figure below.

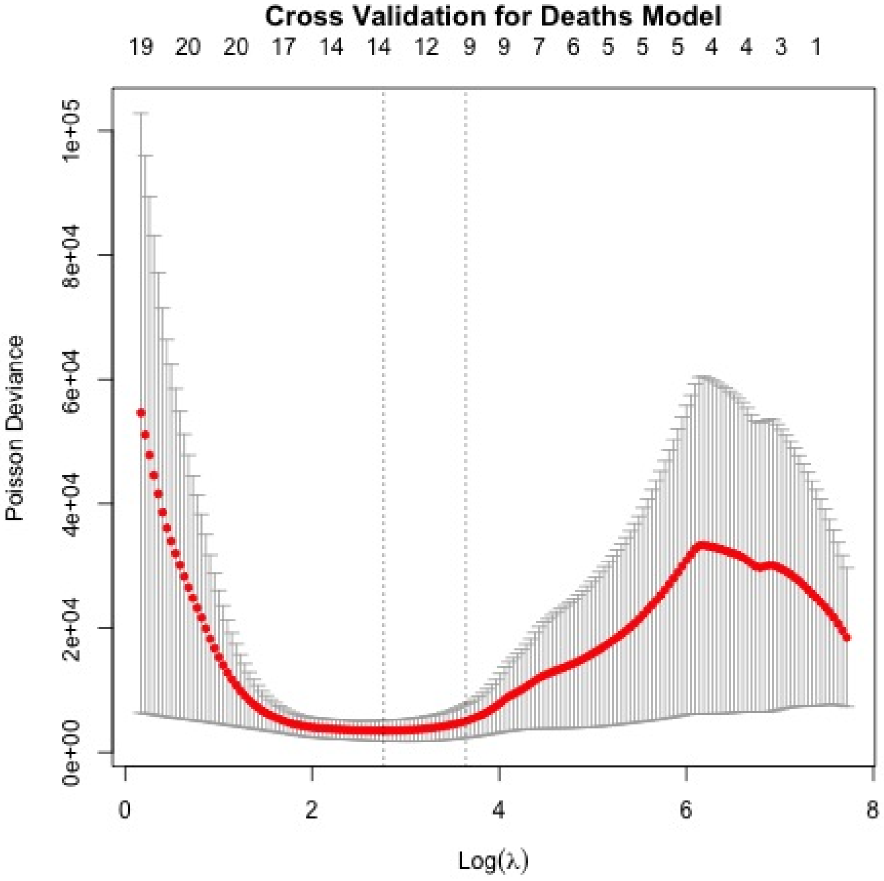


Figure 10: Poisson Deviance against Lasso Penalty with Deaths as Outcome

The selected model (the leftmost vertical dotted line) enjoys a 94% reduction in deviance relative to that of the full model, which can be thought of as a pseudo-estimate of the adjusted R^2.

For the damage outcome, we select only seven features. The deviance of this selected model and that of other specifications is summarized in the figure below.

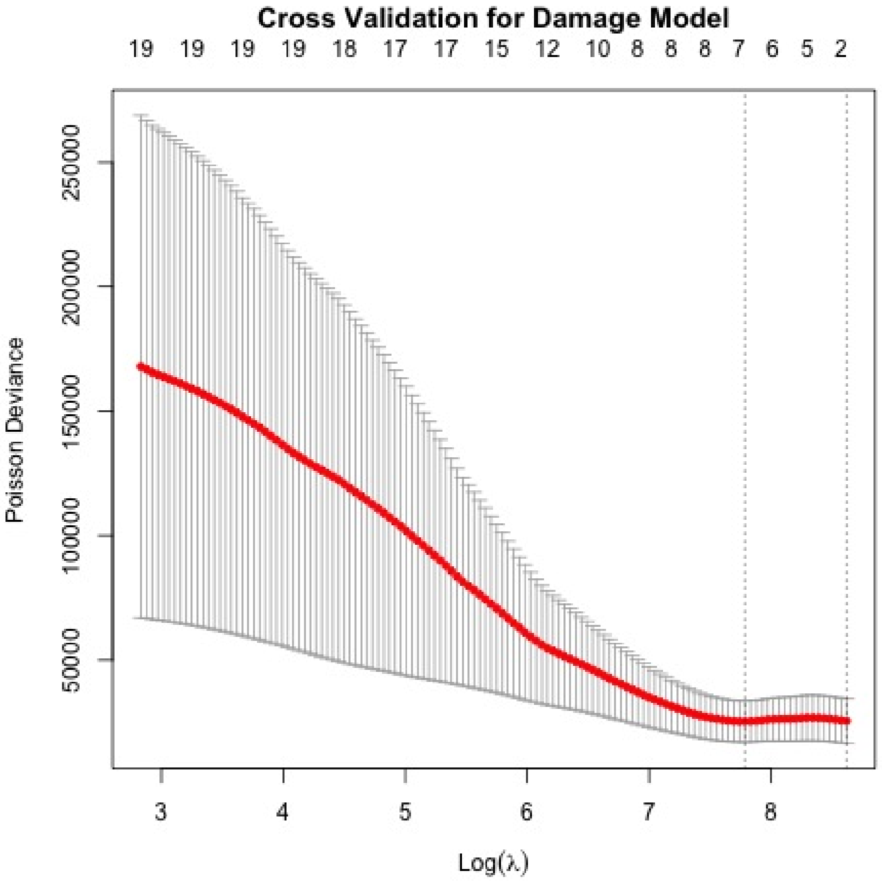


Figure 11: Poisson Deviance against Lasso Penalty with Damage as Outcome

The selected model enjoys an 85% reduction in deviance relative to that of the full model. Unlike in the death model, however, the null model - that which selects no features - does not perform substantially worse from the optimal model. As such, we should approach interpretations of the damage model more carefully. We now proceed to our main results.

Table 4: Inferential Statistics for Selected Features in Death Model

Covariate	Estimate	SE	p-value	Left CI	Right CI
season	0.0385	0.0013	0.0000	0.0359	0.0410
monthJuly	-3.7255	0.1095	0.0000	-3.9401	-3.5108
monthOctober	0.8073	0.0484	0.0000	0.7124	0.9022
natureNR	3.6402	0.1214	0.0000	3.4022	3.8782
natureTS	3.1078	0.0774	0.0000	2.9560	3.2596
maxpressure	-0.0379	0.0049	0.0000	-0.0474	-0.0283
hours	0.0054	0.0002	0.0000	0.0050	0.0057

Covariate	Estimate	SE	p-value	Left CI	Right CI
percent_poor	0.0568	0.0005	0.0000	0.0558	0.0578
percent_usa	-0.0007	0.0005	0.1436	-0.0016	0.0002
beta_0	23.0045	0.5643	0.0000	21.8985	24.1104
beta_1	-10.9904	2.0393	0.0000	-14.9873	-6.9935
beta_2	1.1669	0.7284	0.1092	-0.2608	2.5945
beta_3	17.3277	0.3498	0.0000	16.6421	18.0132

We find that the year (season), month, nature, pressure, and length of the hurricane, in addition to the percent of the affected area that is poor or in the US, can be used to predict deaths. Moreover, all random effects besides the acceleration effect of wind speed (i.e., β_4) are selected for prediction. However, upon obtaining bootstrapped standard errors, we find that the latitude random effect (β_2) and the percent of the affected area that is in the US are not statistically significant.

Interpreting the effects of the β coefficients can be challenging, as they are themselves random effects. For average wind speed β_0, the positive point estimate suggests that hurricanes with higher average wind speeds are more deadly. For the autoregressive random effect β_1, the negative point estimate suggests that hurricanes with more volatile wind speeds are more deadly. For the longitude effect β_3, the positive coefficient suggests that hurricanes covering more distance eastward tend to be more deadly.

We now proceed to selection and inference in the damage model.

Table 5: Inferential Statistics for Selected Features in Damage Model

Covariate	Estimate	SE	p-value	Left CI	Right CI
season	0.0399	0.0004	0.0000	0.0392	0.0406
monthJuly	-0.5762	0.0192	0.0000	-0.6137	-0.5386
monthOctober	0.4680	0.0074	0.0000	0.4536	0.4824
percent_usa	0.0001	0.0001	0.1052	0.0000	0.0003
beta_1	10.5907	0.2533	0.0000	10.0943	11.0871
beta_2	-3.3668	0.1255	0.0000	-3.6128	-3.1208
beta_3	2.4291	0.0669	0.0000	2.2979	2.5603

We find that the year (season), month, and the percent of the affected area that is in the US can be used to predict damage. Moreover, random effects β_1, β_2, and β_3 are also selected for prediction. However, upon obtaining bootstrapped standard errors, we again find that the percent of the affected area that is in the US is not actually statistically significant.

For the autoregressive random effect β_1, the positive point estimate suggests that hurricanes with more correlated wind speeds are more damaging. For the latitude effect β_2, the negative coefficient suggests that hurricanes covering more distance southward tend to be more damaging. For the longitude effect β_3, the positive coefficient suggests that hurricanes covering more distance eastward tend to be more deadly.

4. Discussion

To be able to take appropriate actions and minimize casualties it is crucial for research to predict the behaviors of hurricanes. To do this we proposed a Bayesian model that predicts the wind speeds of hurricanes. Due to the complexity and number of parameters we implemented a Markov Chain Monte Carlo algorithm to estimate each parameter. Using the parameter estimates obtained in the MCMC process we are able to explore if wind speed differs between month, year and by the nature of the hurricane. We can also used the MCMC estimates to understand death and damages of hurricanes.

4.1. Summary of Findings

The MCMC process we developed reached convergence at around 5,500 iterations. Looking at all the estimates excluding the burn in period the distributions all look relatively normal. We are able to take the averages of the iteration for each of the parameters. Exploring the $\boldsymbol{\beta}_i$ estimates for each of the hurricanes we are able to see that the some $\beta_{j,i}$ values are negative and some are positive. Both the impact of change in latitude and longitude have negative means indicating that the wind is slower when the hurricane move further to North and East in direction. The positive μ_1 and μ_4 indicate that a higher previous wind speed has positive effect on the next time's wind speed, and that a larger in change also has a positive effect on the next time's wind speed. The correlation of $\beta_{j,i}$ and $\beta_{k,i}$ for $j \neq k$ is not greater than 0.1.

Overall, our model predicts wind speeds well when observing the adjusted R^2 value. Taking a look at a select few hurricanes we also see that our model does well comparing the actual wind speeds against the observed wind speeds. According to the goodness-fit-test, only 12.9221733% of β_i do not track the hurricane well. We also observed a trend in a higher number of observation corresponding to better predicting ability. Therefore, to have the best ability to predict hurricanes movements having a lot of observations is the best.

Exploring how the decade, month and nature of the hurricane play in role in wind speed it was found that month and nature do not have a significant effect on each of the β_j values when adjusting for the decade. When fitting a model with just the decade we do see that there does seem to be a linear trend for β_1 (corresponding to previous wind speed) and β_3 (corresponding to the impact of change in longitude). Because one of these is estimates is positive and the other is negative there does not seem to be conclusive evidence that wind speed has increased over the years.

In terms of predicting deaths and damage, there is a common theme in our findings: predictability is an important predictor of deaths, but it is less important for predicting total damage. In the deaths model, the autoregressive random effect β_1 had a negative association with deaths, suggesting that hurricanes that are less predictable lead to more fatalities. This makes sense, as people might not have time to prepare for a hurricane with rapidly changing wind speeds. This association completely reverses in thee damage model, however. This could be due to hurricanes with more constant wind speeds having higher (in magnitude) average wind speeds, thereby causing more damage over a sustained period of time. Intuitively, people can prepare for predictable hurricanes, whereas buildings and infrastructure cannot. We also find that poverty is significantly associated with deaths, which is an alarming public health outcome that demands further research.

4.2. Limitations

Due to some hurricanes only having a few observations we had to exclude them from the model. This might have introduced biased into our model due to the systematic exclusion. For example, the hurricanes with limited observations might have been smaller hurricanes that lasted for less amount of time. Therefore, our model might be more powerful at predicting bigger more destructive hurricanes that last for a longer amount of time. This should be kept in mind when using our algorithm. However, we would like to argue that being able to predict bigger more destructive hurricanes is of more interest to public health officials.

Another limitation is that we only have up to year 2013. As climate change become an ever increasing problem storms in years closer to 2022 might be more extreme compared to in 2013. To be able to have a full grasp on how hurricanes behave using data from years closer to 2022 is important to be able to predict this years hurricane behaviors.

Regarding the deaths and damage models, it is helpful to understand the impact of our estimated random effects on relevant outcomes. However, as is the case with random effects models in general, out-of-sample prediction is not possible, especially when trying to predict the destructive force of a hurricane as it is forming. One idea for future research is to use the estimated random effects center $\hat{\mu}$ and covariance structure $\hat{\Sigma}$ in order to generate a random effects model so that out-of-sample random effects $\hat{\beta}_i$ can be estimated.

4.3. Group Contributions

All team members collaborated on this project. Ruiqi Yan, Tianchuan Gao, and Jiacheng Wu worked collectively on tasks 1-4. Ruiqi designed the MCMC algorithm to obtain the parameter estimates and assessed the model performance. Tianchuan generated the latex expressions of the likelihood and posterior distribution of the parameters. Jiacheng created the convergence trace plots, histograms, and chart of the estimated parameters. Amy Pitts coordinated our meetings, arranged the github files, and completed task 5. Jimmy Kelliher completed task 6, designed the algorithm for post-selection inference, and compiled references. All team members worked on the presentation slides and the report write-up.

References

[1] Elsner, James B., et al. "Detecting Shifts in Hurricane Rates Using a Markov Chain Monte Carlo Approach." *Journal of Climate*, vol. 17, no. 13, 2004, pp. 2652–66, http://www.jstor.org/stable/26251818. Accessed 9 May 2022.

[2] Efron, Bradley. "Estimation and Accuracy After Model Selection." *Journal of the American Statistical Association*, vol. 109, no. 507, 2014, pp. 991–1007, http://www.jstor.org/stable/24247426. Accessed 9 May 2022.

Appendices

Appendix A

Calculating the conditional Priors for each paramater.

$$\boldsymbol{\beta}_i : \pi(\boldsymbol{\beta}_i \mid \boldsymbol{\Theta}_{(-\boldsymbol{\beta}_i)} \boldsymbol{Y}) \propto \exp\left\{ -\frac{1}{2} \left(\boldsymbol{\beta}_i - \boldsymbol{\mu}\right)^T \Sigma^{-1} \left(\boldsymbol{\beta}_i - \boldsymbol{\mu}\right) - \frac{1}{2\sigma^2} \left(\boldsymbol{Y}_i - \boldsymbol{X}_i \boldsymbol{\beta}_i\right)^T \left(\boldsymbol{Y}_i - \boldsymbol{X}_i \boldsymbol{\beta}_i\right) \right\}$$

$$\boldsymbol{\mu} : \pi\left(\boldsymbol{\mu} \mid \boldsymbol{\Theta}_{(-\boldsymbol{\mu})}, \boldsymbol{Y}\right) \propto \exp\left\{ -\frac{1}{2} \sum_{i=1}^{m} \left(\boldsymbol{\beta}_i - \boldsymbol{\mu}\right)^T \Sigma^{-1} \left(\boldsymbol{\beta}_i - \boldsymbol{\mu}\right) \right\}$$

$$\begin{aligned}
(\boldsymbol{\beta}_i - \boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{\beta}_i - \boldsymbol{\mu}) &= \operatorname{tr}\left(\left(\boldsymbol{\beta}_i - \boldsymbol{\mu}\right)^T \Sigma^{-1} \left(\boldsymbol{\beta}_i - \boldsymbol{\mu}\right) \right) \\
&= \operatorname{tr}\left(\Sigma^{-1} \left(\boldsymbol{\beta}_i - \boldsymbol{\mu}\right) \left(\boldsymbol{\beta}_i - \boldsymbol{\mu}\right)^T \right) \\
\pi\left(\boldsymbol{\mu} \mid \boldsymbol{\Theta}_{(-\boldsymbol{\mu})}, Y\right) &\propto \exp\left\{ -\frac{1}{2} \operatorname{tr}\left(\Sigma^{-1} \sum_{i=1}^{m} \left(\boldsymbol{\beta}_i - \boldsymbol{\mu}\right) \left(\boldsymbol{\beta}_i - \boldsymbol{\mu}\right)^T \right) \right\} \\
&= \exp\left\{ -\frac{1}{2} \operatorname{tr}\left(\Sigma^{-1} m \left(\boldsymbol{\mu} - \bar{\boldsymbol{\beta}}\right) \left(\boldsymbol{\mu} - \bar{\boldsymbol{\beta}}\right)^T \right) \right\} \\
&= \exp\left\{ -\frac{1}{2} (\boldsymbol{\mu} - \bar{\boldsymbol{\beta}})^T \Sigma^{-1} m (\boldsymbol{\mu} - \bar{\boldsymbol{\beta}}) \right\} \\
&\Rightarrow \boldsymbol{\mu} \mid \boldsymbol{\Theta}_{(-\boldsymbol{\mu})}, \boldsymbol{Y} \sim N(\bar{\boldsymbol{\beta}}, \Sigma/m)
\end{aligned}$$

$$\sigma^2 : \pi\left(\sigma^2 \mid \boldsymbol{\Theta}_{(-\sigma^2)}, \boldsymbol{Y}\right) \propto \left(\sigma^2\right)^{-\left(1 + \frac{\sum_{i=1}^{m} n_i}{2}\right)} \times \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{m} \left(\boldsymbol{Y}_i - \boldsymbol{X}_i \boldsymbol{\beta}_i\right)^T \left(\boldsymbol{Y}_i - \boldsymbol{X}_i \boldsymbol{\beta}_i\right) \right\}$$

$$\begin{aligned}
\Sigma^{-1} : \pi\left(\Sigma^{-1} \mid \boldsymbol{\Theta}_{(-\Sigma^{-1})}, \boldsymbol{Y}\right) &\propto \left|\Sigma^{-1}\right|^{d+1+\frac{m}{2}} \exp\left\{ -\frac{1}{2} \operatorname{tr}\left(\Sigma^{-1}\right) \right\} \exp\left\{ -\frac{1}{2} \sum_{i=1}^{m} \left(\boldsymbol{\beta}_i - \boldsymbol{\mu}\right)^T \Sigma^{-1} \left(\boldsymbol{\beta}_i - \boldsymbol{\mu}\right) \right\} \\
&= \left|\Sigma^{-1}\right|^{d+1+\frac{m}{2}} \exp\left\{ -\frac{1}{2} \operatorname{tr}\left(\Sigma^{-1}\right) - \frac{1}{2} \operatorname{tr}\left(\Sigma^{-1} \sum_{i=1}^{m} \left(\boldsymbol{\beta}_i - \boldsymbol{\mu}\right) \left(\boldsymbol{\beta}_i - \boldsymbol{\mu}\right)^T \right) \right\} \\
&= \left|\Sigma^{-1}\right|^{d+1+\frac{m}{2}} \exp\left\{ -\frac{1}{2} \operatorname{tr}\left(\Sigma^{-1} \left(I + \sum_{i=1}^{m} \left(\beta_i - \mu\right) \left(\beta_1 - \mu\right)^T \right) \right) \right\} \\
&\Rightarrow \Sigma^{-1} \mid \boldsymbol{\Theta}_{(-\Sigma^{-1})}, \boldsymbol{Y} \sim \text{Wishart}\left(3d + 3 + m, \left(\boldsymbol{I} + \sum_{i=1}^{m} \left(\boldsymbol{\beta}_i - \boldsymbol{\mu}\right) \left(\boldsymbol{\beta}_i - \boldsymbol{\mu}\right)^T \right)^{-1} \right)
\end{aligned}$$

Appendix B

The code associated with the MCMC algorithm.

Cleaning the data

```
# load the data
dt = read.csv("hurrican703.csv")

# clean the original data
df_temp = dt %>%
  mutate( # Factoring the Nature and Months Variable
```

```r
    Nature = factor(Nature),
    Month = factor(Month, levels = c(month.name)),
  ) %>%  # Seperating the day and time to usable info
  tidyr::separate(
    time, c('Date', 'Time'),
                       sep = ' ', extra = 'merge') %>%
  tidyr::separate(
    Date, c('year_num', 'month_num', 'day'),
                       sep = '-', extra = 'merge') %>%
  tidyr::separate(
    Time, c('hour', 'min', 'sec'),
                       sep =':', extra = 'merge')  %>%
  select(ID, Nature, Latitude, Longitude, Wind.kt,
         Season, Month, day, hour, min) %>% # selecting relevent columns
  filter(min == "00") %>%
  filter(hour %in% c("00", "06", "12", "18")) %>% # observations on every 6 hours
  mutate(
    hour = as.numeric(hour),
    day = as.numeric(day)
  )

# find the id's with limited observations
temp = table(df_temp$ID) %>% as.data.frame()

# eliminate anyone that has less or equal to 7
id_get_rid <- temp %>% arrange(Freq) %>% filter(Freq < 7) %>% pull(Var1)
df_rm <- df_temp %>% filter(!(ID %in% c(id_get_rid)))

# figuring out the percentage we are removing
(dim(dt)[1]-dim(df_rm)[1]) / dim(dt)[1]
```

Calculating the log posterior functions

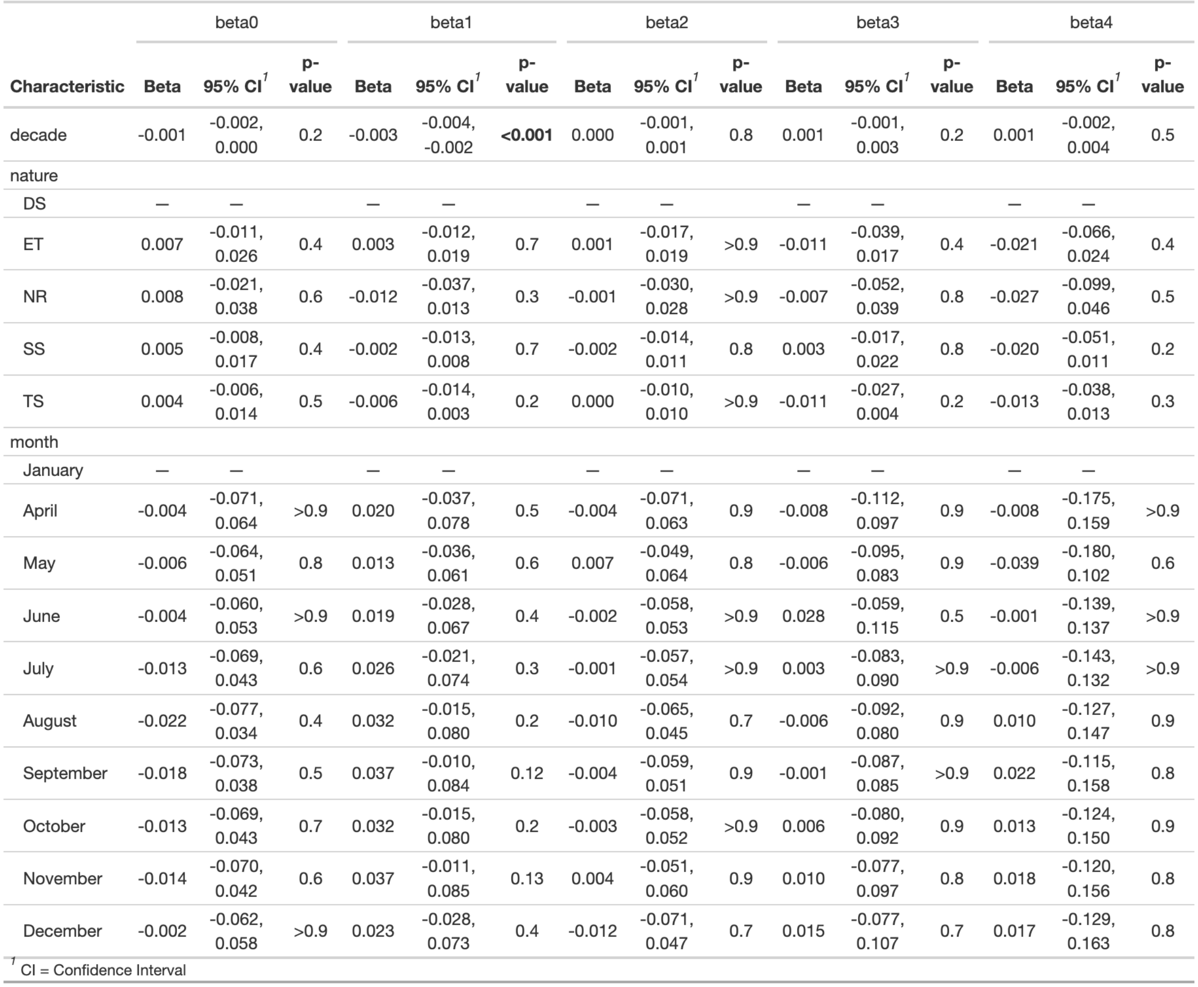
```r
## log posterior for beta_j of jth hurricane

log_posterior_beta = function(beta, sigma, mu, sigma_p,j){
  x = obs_list$x_matrix[[j]]
  y = obs_list$y_obs[[j]]
  return(-t(y - x %*% beta) %*% (y - x %*% beta)/(2*sigma)-(1/2)*t(beta-mu)%*%sigma_p%*%(beta-mu))
}

## log posterior for sigma
log_posterior_sigma = function(beta_frame, sigma){
  ## make sure sigma is positive
  if(sigma <= 0){
    return(-Inf)
  } else{
    log_lik = -(n_obs/2)*log(sigma)-sum((dt$wind_kt-beta_frame[,1]-dt$wind_lag*beta_frame[,2]-dt$lat_shi
    log_prior = -log(sigma)
    return(log_lik + log_prior)
  }
}
```

The MCMC algorithm

```
componentwisemixedMH=function(a, a_sigma, beta_0,sigma_0,cov_0, nrep=10000){
  sigma_p = cov_0 #inverse covariance matrix
  beta = beta_0
  mu = colMeans(beta)
  sigma = sigma_0
  theta_chain_mix = matrix(0, nrow = nrep, ncol = n_theta)
  for(i in 1:nrep){
    sigma_m = solve(sigma_p) #current covariance matrix
    # sampling beta for jth hurricane
    for(j in 1:m){
      #component_wise for each beta_kj in beta_j
      for(k in 1:5){
        prop = beta[j,]
        # random walk with uniform step size for beta_k,j
        prop[k] = prop[k]+2*(runif(1)-0.5)*a[(k-1)*681+j]
        # accepting or rejecting
        if(log(runif(1)) < (log_posterior_beta(prop, sigma, mu, sigma_p,j) - log_posterior_beta(beta[j,]
          beta[j,k] = prop[k]
        }
      }
    }
    # sampling mu from mvn
    beta_mean = colMeans(beta) #average over current beta_i
    mu = mvrnorm(1,beta_mean,Sigma = sigma_m/m)
    # random walk for sigma^2 with uniform step size
    beta_frame = matrix(rep(c(beta),times = rep_time), ncol = d)
    prop = sigma
    prop = prop+2*(runif(1)-0.5)*a_sigma
    # accepting or rejecting
    if(log(runif(1)) < (log_posterior_sigma(beta_frame, prop) - log_posterior_sigma(beta_frame, sigma))
      sigma = prop
    }
    # sampling inverse sigma matrix from wishart
    beta_mu = t(beta)-mu
    S = solve(diag(d)+beta_mu%*%t(beta_mu)) # V matrix in wishart
    sigma_p = rWishart(1,df = 3*d+3+m,Sigma = S)
    sigma_p = apply(sigma_p,2,c) # extract the inverse covariance matrix sample
    #store new samples
    theta_chain_mix[i,] = c(c(beta),mu,sigma,sigma_p[lower.tri(sigma_p, diag = T)])
  }
  return(theta_chain_mix)
}
```

Appendix C

The Model Estimates from Section 3.3.



	beta0			beta1			beta2			beta3			beta4		
Characteristic	**Beta**	**95% CI**[1]	**p-value**	**Beta**	**95% CI**[1]	**p-value**	**Beta**	**95% CI**[1]	**p-value**	**Beta**	**95% CI**[1]	**p-value**	**Beta**	**95% CI**[1]	**p-value**
decade	-0.001	-0.002, 0.000	0.2	-0.003	-0.004, -0.002	**<0.001**	0.000	-0.001, 0.001	0.8	0.001	-0.001, 0.003	0.2	0.001	-0.002, 0.004	0.5
nature															
DS	—	—		—	—		—	—		—	—		—	—	
ET	0.007	-0.011, 0.026	0.4	0.003	-0.012, 0.019	0.7	0.001	-0.017, 0.019	>0.9	-0.011	-0.039, 0.017	0.4	-0.021	-0.066, 0.024	0.4
NR	0.008	-0.021, 0.038	0.6	-0.012	-0.037, 0.013	0.3	-0.001	-0.030, 0.028	>0.9	-0.007	-0.052, 0.039	0.8	-0.027	-0.099, 0.046	0.5
SS	0.005	-0.008, 0.017	0.4	-0.002	-0.013, 0.008	0.7	-0.002	-0.014, 0.011	0.8	0.003	-0.017, 0.022	0.8	-0.020	-0.051, 0.011	0.2
TS	0.004	-0.006, 0.014	0.5	-0.006	-0.014, 0.003	0.2	0.000	-0.010, 0.010	>0.9	-0.011	-0.027, 0.004	0.2	-0.013	-0.038, 0.013	0.3
month															
January	—	—		—	—		—	—		—	—		—	—	
April	-0.004	-0.071, 0.064	>0.9	0.020	-0.037, 0.078	0.5	-0.004	-0.071, 0.063	0.9	-0.008	-0.112, 0.097	0.9	-0.008	-0.175, 0.159	>0.9
May	-0.006	-0.064, 0.051	0.8	0.013	-0.036, 0.061	0.6	0.007	-0.049, 0.064	0.8	-0.006	-0.095, 0.083	0.9	-0.039	-0.180, 0.102	0.6
June	-0.004	-0.060, 0.053	>0.9	0.019	-0.028, 0.067	0.4	-0.002	-0.058, 0.053	>0.9	0.028	-0.059, 0.115	0.5	-0.001	-0.139, 0.137	>0.9
July	-0.013	-0.069, 0.043	0.6	0.026	-0.021, 0.074	0.3	-0.001	-0.057, 0.054	>0.9	0.003	-0.083, 0.090	>0.9	-0.006	-0.143, 0.132	>0.9
August	-0.022	-0.077, 0.034	0.4	0.032	-0.015, 0.080	0.2	-0.010	-0.065, 0.045	0.7	-0.006	-0.092, 0.080	0.9	0.010	-0.127, 0.147	0.9
September	-0.018	-0.073, 0.038	0.5	0.037	-0.010, 0.084	0.12	-0.004	-0.059, 0.051	0.9	-0.001	-0.087, 0.085	>0.9	0.022	-0.115, 0.158	0.8
October	-0.013	-0.069, 0.043	0.7	0.032	-0.015, 0.080	0.2	-0.003	-0.058, 0.052	>0.9	0.006	-0.080, 0.092	0.9	0.013	-0.124, 0.150	0.9
November	-0.014	-0.070, 0.042	0.6	0.037	-0.011, 0.085	0.13	0.004	-0.051, 0.060	0.9	0.010	-0.077, 0.097	0.8	0.018	-0.120, 0.156	0.8
December	-0.002	-0.062, 0.058	>0.9	0.023	-0.028, 0.073	0.4	-0.012	-0.071, 0.047	0.7	0.015	-0.077, 0.107	0.7	0.017	-0.129, 0.163	0.8

[1] CI = Confidence Interval

Figure 12: The Linear estimates for each beta vector model to explore the effect of nature, month and decade on the wind speed. This corresponds with section 3.3