P8160 - Project 3
Baysian modeling of hurrican trajectories

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

2022-05-09

Motivation

Climate researchers are interested in modeling the hurricane trajectories to forecast the wind speed.

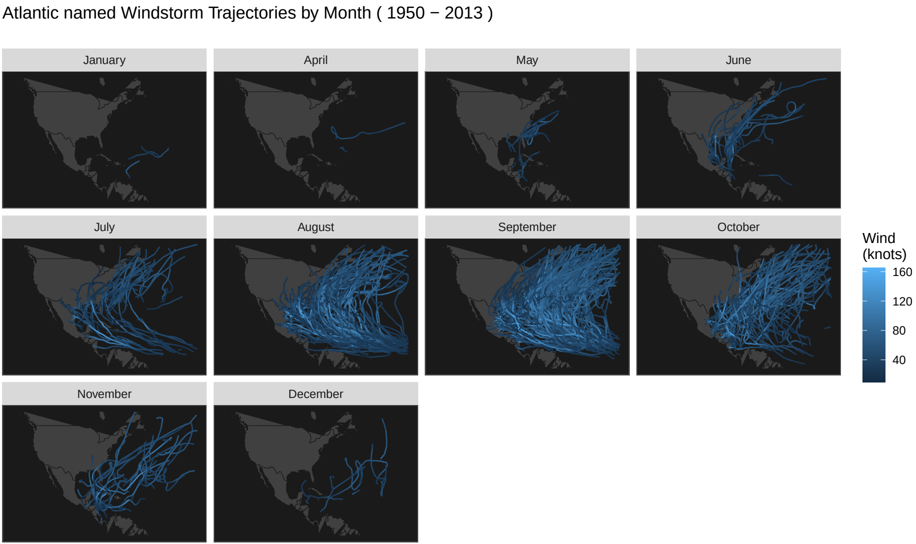
Data

- **ID**: ID of hurricanes
- **Year**: In which the hurricane occurred
- **Month**: In which the hurricane occurred
- **Nature**: Nature of the hurricane
  - ET: Extra Tropical
  - DS: Disturbance
  - NR: Not Rated
  - SS: Sub Tropical
  - TS: Tropical Storm
- **Time**: dates and time of the record
- **Latitude** and **Longitude**: The location of a hurricane check point
- **Wind.kt**: Maximum wind speed (in Knot) at each check point

Outline

1. Exploration into the Data
2. Bayesian modeling of hurricane wind speed
   - Model Equation
   - Posterior Derivation
   - MCMC Algorithm
3. How Month, Year, and the Nature of the hurricane affect wind speed
   - Explore seasonal differences
   - Explore if wind speeds is increasing over years
4. Exploring wind speeds impact on death and damages
5. As well as the characteristic of a hurricane associated with damages and deaths

Data


Atlantic named Windstorm Trajectories by Month (1950 − 2013)
January
April
May
June
July
August
September
October
November
December
Wind
(knots)
160
120
80
40

Data Cleaning

- We are only concerned about observations that occurred on 6 hour intervals. hour 0, 6, 12, and 18.
- In addition we will exclude all hurricane IDs that have less then 7 observations.
- We used the lag difference ($t-6$ to $t-12$) for latitude, longitude and wind speed to build $\Delta_{i,1}(t), \Delta_{i,2}(t), \Delta_{i,3}(t)$ and lag of wind speed as $Y_i(t-6)$

Through this process we remove 460 observations so we are left with 21578 observations and 681 unique hurricanes.

Bayesian Model for Hurricane Trajectories

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 $t-12$ and $t-6$
- $\epsilon_i(t) \sim N(0, \sigma^2)$ 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 $\Sigma \in \mathbb{R}^{d\times d}$.

Prior Distributions Assumptions:

- For σ^2 we assume $\pi(\sigma^2) \propto \frac{1}{\sigma^2}$
- For μ we assume $\pi(\mu) \propto 1$
- For Σ we assume $\pi\left(\Sigma^{-1}\right) \propto |\Sigma|^{-(d+1)} \exp\left\{-\frac{1}{2}\Sigma^{-1}\right\}$

Goal: Estimate $\Theta = (B, \mu, \Sigma^{-1}, \sigma^2)$

Likelihood & Prior Function

Likelihood Let $X_i = \big(\mathbf{1}, Y_i(t-6), \Delta_{i,1}(t), \Delta_{i,2}(t), \Delta_{i,3}(t)\big)$

$$L(Y \mid \Theta) \propto \prod_{i=1}^{m} \left(\sigma^2\right)^{-\frac{n_i}{2}} \exp\left\{-\frac{1}{2\sigma^2}\left(Y_i - X_i\beta_i\right)^T\left(Y_i - X_i\beta_i\right)\right\}$$

where m is the number of hurricane and n_i is the number of observations for i^{th} hurricane

Prior Let $\Theta = (B, \mu, \Sigma^{-1}, \sigma^2)$

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

where d is the dimension of μ

Posterior Calculation

Posterior

$$\pi(\Theta \mid Y) \propto \left(\sigma^2\right)^{-\left(1+\frac{\sum_{i=1}^{m} n_i}{2}\right)} \left|\Sigma^{-1}\right|^{d+1+\frac{m}{2}} \exp\left\{-\frac{1}{2}\operatorname{tr}\left(\Sigma^{-1}\right)\right\}$$
$$\times \exp\left\{-\frac{1}{2}\sum_{i=1}^{m}\left(\beta_i-\mu\right)^T \Sigma^{-1}\left(\beta_i-\mu\right)\right\} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{m}\left(Y_i - X_i\beta_i\right)^T\left(Y_i - X_i\beta_i\right)\right\}$$

Conditional Posterior

$$\beta_i : \pi(\beta_i \mid \Theta_{(-\beta_i)} Y) \propto \exp\left\{-\frac{1}{2}\left(\beta_i-\mu\right)^T \Sigma^{-1}\left(\beta_i-\mu\right) - \frac{1}{2\sigma^2}\left(Y_i - X_i\beta_i\right)^T\left(Y_i - X_i\beta_i\right)\right\}$$

$$\mu : \pi\left(\mu \mid \Theta_{(-\mu)}, Y\right) \sim \mathcal{N}(\bar{\beta}, \Sigma/m), \bar{\beta} = \left(\bar{\beta}_{0,.}, \bar{\beta}_{1,.}, \bar{\beta}_{2,.}, \bar{\beta}_{3,.}, \bar{\beta}_{4,.}\right)^T$$

$$\sigma^2 : \pi\left(\sigma^2 \mid \Theta_{(-\sigma^2)}, Y\right) \propto \left(\sigma^2\right)^{-\left(1+\frac{\sum_{i=1}^{m} n_i}{2}\right)} \times \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{m}\left(Y_i - X_i\beta_i\right)^T\left(Y_i - X_i\beta_i\right)\right\}$$

$$\Sigma^{-1} : \pi\left(\Sigma^{-1} \mid \Theta_{(-\Sigma^{-1})}, Y\right) \sim \text{Wishart}\left(3d+3+m, \left(I + \sum_{i=1}^{m}\left(\beta_i-\mu\right)\left(\beta_i-\mu\right)^T\right)^{-1}\right)$$

MCMC Algorithm

We apply hybrid algorithm consisting with Metropolis-Hastings steps and Gibbs steps.

Update component wise:

- Sampling proposed $\beta'_{j,i}$, $j = 0, 1...4$, 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 proposals are symmetry, the accepting or rejecting the proposed $\beta'_{j,i}$ depends on the ratio of posterior distribution.
- Then, Gibb step for μ: Sample $\mu^{(t+1)}$ from $\mathcal{N}\left(\bar{\beta}^{(t+1)}, \Sigma^{(t)}/m\right)$, where $\bar{\beta}^{(t+1)}$ is the average of $\beta_i^{(t+1)}$ over all hurricanes.
- Next, MH step to generate $\sigma^{2'}$ from $U\left(\sigma^{2^{(t)}} - a_{\sigma^2}, \sigma^{2^{(t)}} + a_{\sigma^2}\right)$.
- Finally, we sample $\Sigma^{-1^{(t+1)}}$ from $\text{Wishart}\left(3d + 3 + m, \left(I + \sum_{i=1}^{m}\left(\beta_i^{(t+1)} - \mu^{(t+1)}\right)\left(\beta_i^{(t+1)} - \mu^{(t+1)}\right)^T\right)^{-1}\right)$

Initial Starting Values

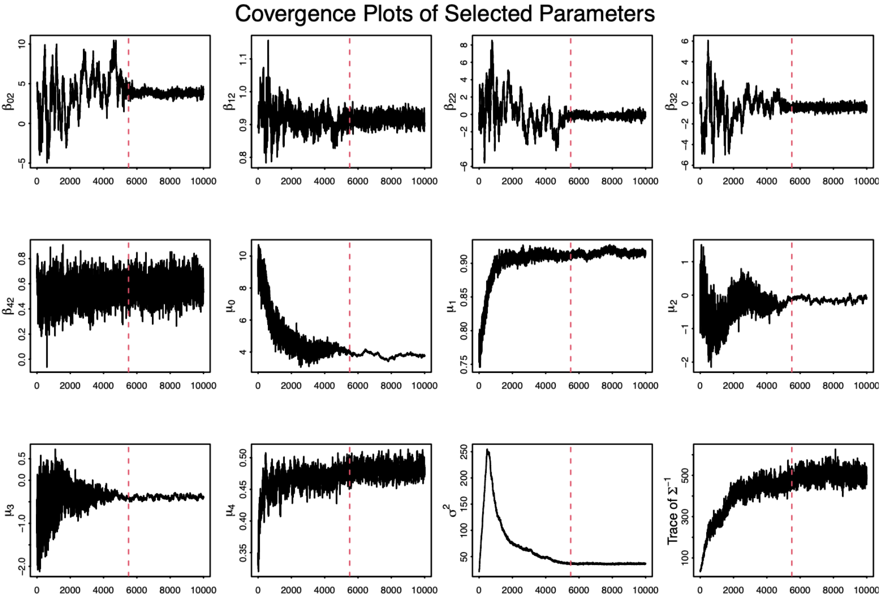
Initial Values:

- β_i: Fit OLS multivariate linear regression (MLR) for i^{th} hurricane and use the coefficients as $\beta_i^{(0)}$
- μ: Average over all $\beta_i^{(0)}$ as $\mu^{(0)}$
- 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 $\beta_i^{(0)}$ and take the inverse of the matrix as $\Sigma^{-1^{(0)}}$

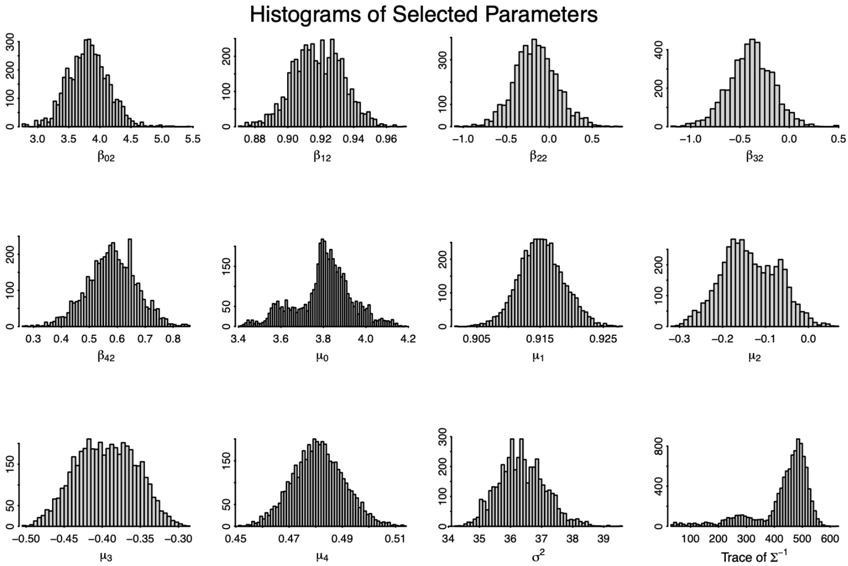
Table 1: Range of Search Window and Acceptance Rate for paraemters used MH step

	Search Window	Acceptance Rate (%)
β_0	1.1	45.87 - 51.36
β_1	(0.04, 0.1)	31.67 - 63.68
β_2	(0.8, 1.0)	38.60 - 45.60
β_3	(0.5, 0.6)	33.20 - 61.32
β_4	(0.4, 0.5)	34.95 - 60.45
σ^2	2.0	44.83

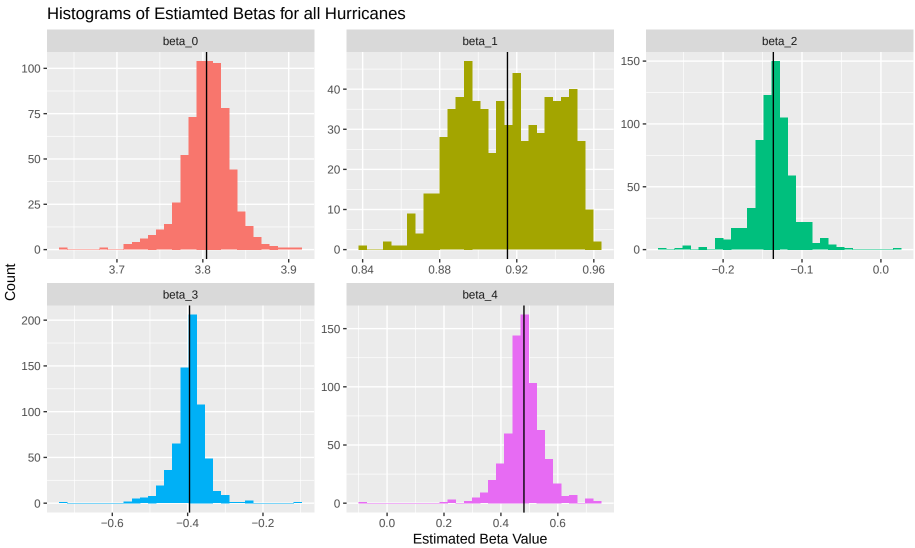
MCMC Model Convergence



MCMC Model Convergence



B Estimates


Histograms of Estiamted Betas for all Hurricanes
beta_0
beta_1
beta_2
beta_3
beta_4
Count
Estimated Beta Value

The μ and σ^2 Estiamtes

Table 2: Bayesian Estiamtes for μ and σ^2

	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

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

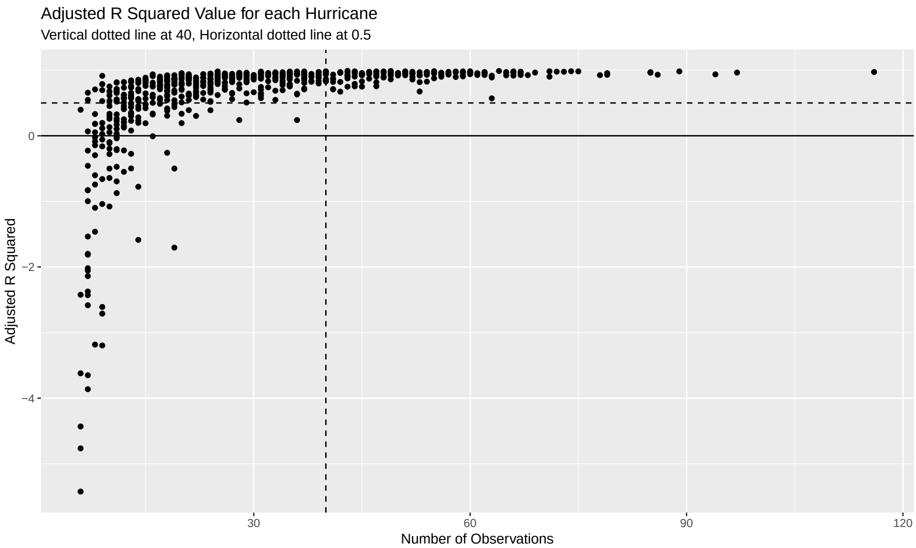
Model Performance

The overall adjusted R^2 of the estimated Bayesian model is 0.9524156.

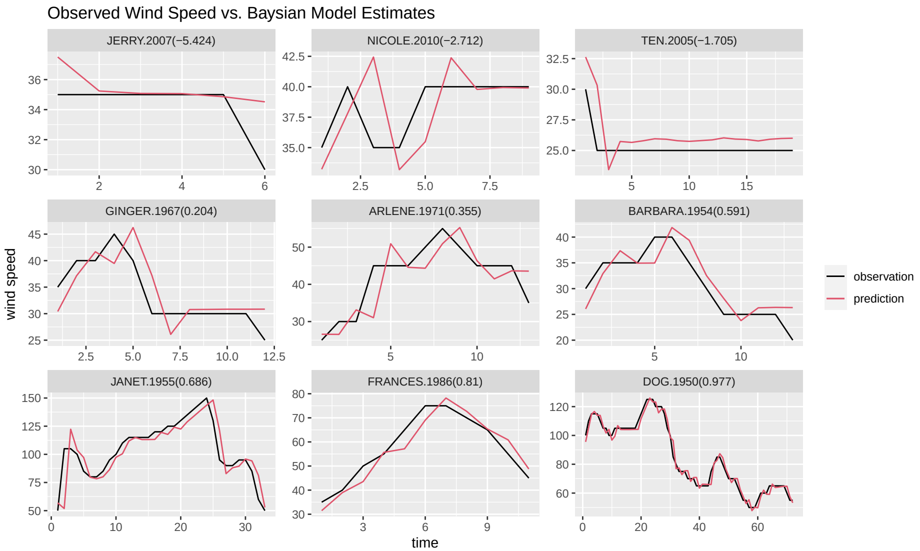
Table 3: R^2_{adj} for each hurricane

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

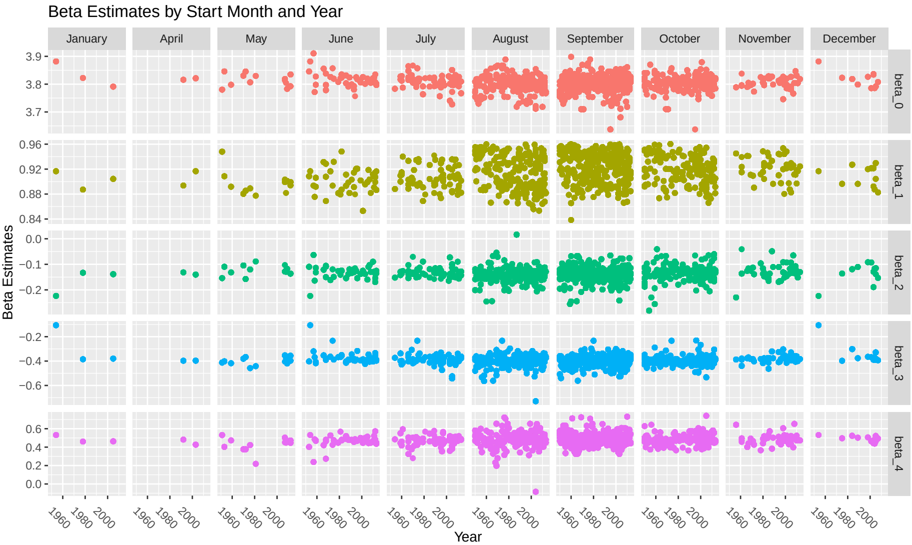
Model Performance



Model Performance

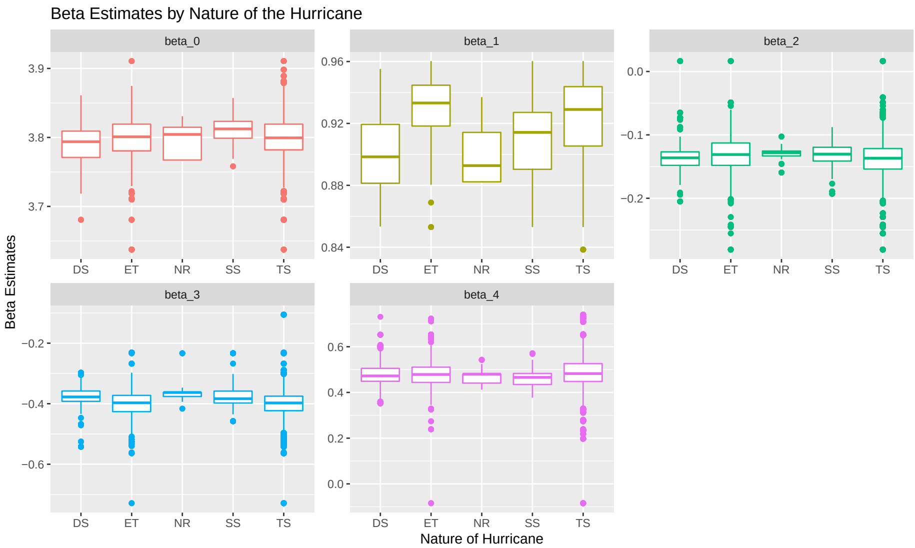


Understanding Seasonal Differences



Typical Hurricane season is June to November

Understanding Nature of Hurricane Differences



Modeling Seasonal, and Nature Difference

Model:

For each β value we fit a different linear model.

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

Where i is the hurricane, j is the Beta model, $k \in (ET,\ NR,\ SS,\ TS)$ making DS the reference group. Let $l \in$(April - December).

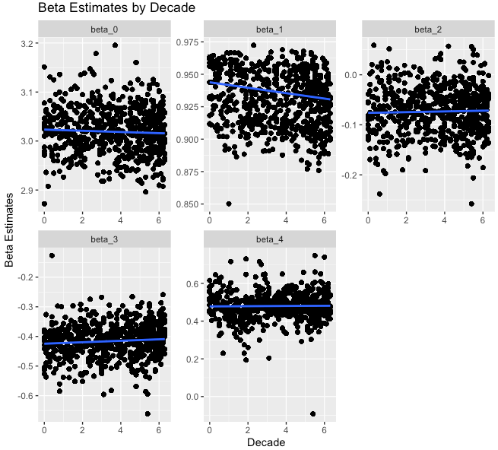
Results:

- Due to length each model estimate were omitted
- No Nature Indicators were significant
- No Month Indicators were significant
- For the β_1 model the decade estimated is significant -0.003 (-0.002, -0.004)

Thus Month and Nature don't have a linear association with each beta value.

Exploring if wind speeds have increased over the years

Model: $Y_i = \alpha_{0i} + \alpha_{1i} \times \text{Decade}$ where Y_i is each β_i and $i \in (0, \ldots, 4)$.



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

- β_1: Indicates a decrease in the change of wind speed over years

Deaths and Damages - Data Exploration

Distribution of Continuous Variables (Standardized)

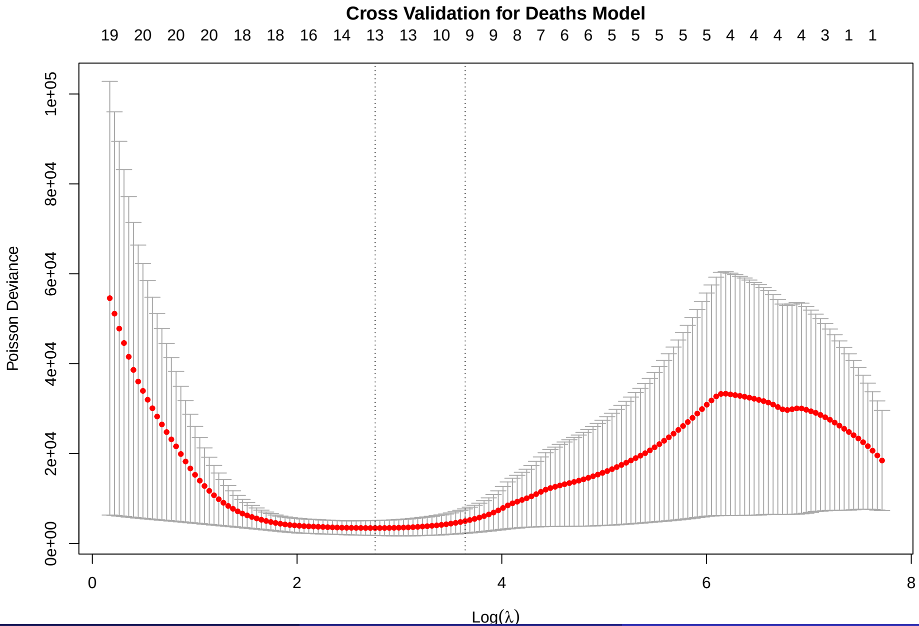
Characteristic	**N = 43**
damage	(-0.46, -0.43, -0.33, -0.09, 5.19)
deaths	(-0.29, -0.28, -0.27, -0.22, 5.81)
maxspeed	(-1.94, -0.81, 0.14, 0.89, 1.64)
meanspeed	(-1.67, -0.82, -0.02, 0.74, 2.18)
maxpressure	(-3.88, -0.57, 0.10, 0.32, 2.09)
meanpressure	(-3.94, 0.24, 0.31, 0.36, 0.41)
hours	(-1.87, -0.85, 0.00, 0.74, 2.50)
total_pop	(-0.98, -0.69, -0.37, 0.13, 2.74)
percent_poor	(-0.30, -0.30, -0.30, -0.29, 3.64)
percent_usa	(-1.22, -1.22, 0.14, 0.96, 1.14)

Note: For each characteristic, values denote the minimum, 25th percentile, median, 75th percentile, and maximum, respectively.

Deaths and Damages - Prediction and Inference

- Model Selection
    - Poisson regression with total population as offset
        - $y_i \sim \text{Poisson}(\mu_i)$
        - $\log(\mu_i) = \mathbf{x}_i'\gamma$
    - penalized regression via lasso
    - optimal λ selected via leave-one-out cross validation
        - feasible with small n
        - non-random
- Post-selection Inference
    - bootstrap smoothing as proposed by Efron (2014)
        - fit full model via Poisson-family GLM to obtain $\hat{\mu}$
        - for single bootstrap b, draw $\mathbf{y}_b^* \sim \text{Poisson}(\hat{\mu})$
        - execute lasso with $\mathbf{y}_b^*$ as output to obtain distribution of $\hat{\gamma}_b^*$
        - compute empirical standard error of $\hat{\gamma}_b^*$

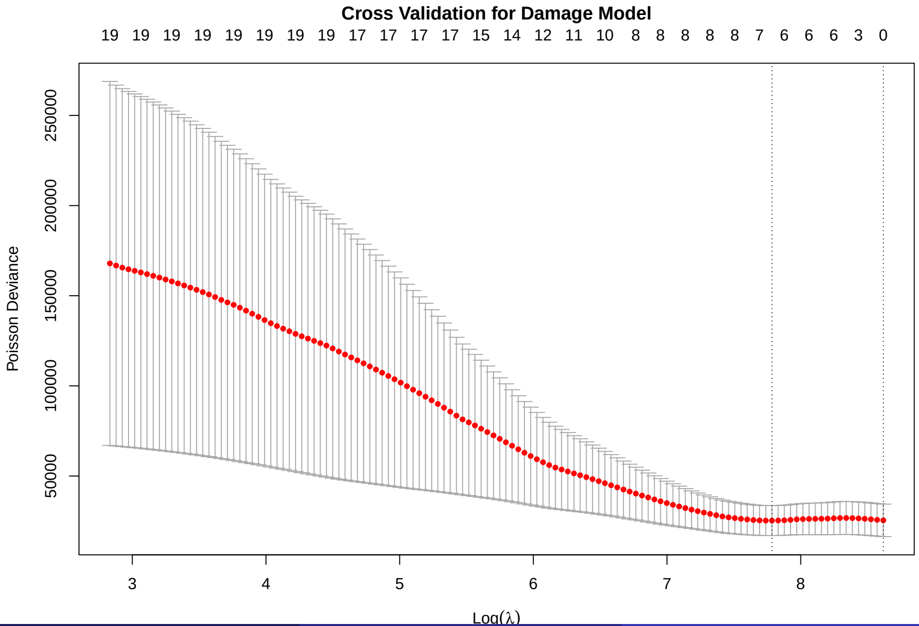
Hurricane Deaths - Model Selection



Hurricane Deaths - Inference

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
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

Hurricane Damages - Model Selection



Hurricane Damages - Inference

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

Conclusions

- Hurricanes with more observations enjoy better MCMC estimates
- With such a high-dimensional parameter space:
  - burn-in takes longer than expected
  - very sensitive to starting values
- Estimated β coefficients generally prove useful in predicting damage and deaths
  - particularly, $(\beta_1, \beta_2, \beta_3)$ are always selected
  - however, β_2 is only inferentially significant in the damage model