ST 540 Midterm 2

Chelsea Robalino

April 17, 2019

1. Introduction

The data consists of meteorological data from the past 50 years. The response is the number of tropical storms that made landfall on the US Atlantic Coast in that year. The covariates are the sea surface temperatures (SST) from ten available location in the six months priors to the beginning of the hurricane season. The goal is to identify the months and locations of SST that are most predictive of the number of storms. Three models will be fit to the data, and the best fitting model will be chosen to forecast the number of storms in the final year of the data set.

2. Methods

The response variable is count data, so a Poisson likelihood function with a log link to the mean was chosen. Uninformative priors were used allowing the likelihood overwhelms the prior. Scaling the independent variables was not necessary as all are measuring SST and thus are on the same scale. Initially, the model was fit using all 60 covariates but the model did not converge as sample size increased or priors were changed. The dimensionality of the covariates was reduced so the number of predictors would be less than the sample size. This reduction was done in three ways:

1. Averaging over latitude and having 12 predictors — one for each combination of longitude and month.

2. Averaging over longitude and having 30 predictors — one for each combination of latitude and month.

3. Creating two groups for the months. Group 1 for months 1-3 and group 2 for months 4-6. Then averaging over groups and having 20 predictors — one for each combination of location and group.

The months were divided in this way as the average SST of months 1-3 was positive and the average

SST of months 4-6 was negative. Since the latitude and longitude of the locations was given, further north or south and closer to the US Atlantic coast were tested as useful predictors.

The likelihood and priors used for the models are below (the X's and p's are different for the three models as defined above):

$$Y_i \sim Poisson(\lambda_i)$$

$$log(\lambda_i) = \beta_0 + \sum_{j=1}^{p} X_{ij}\beta_j$$

$$\beta_0 \sim N(0, 1000)$$

$$\beta_j \sim N(0, 1000)$$

$$i = 1, 2, .., n = 49$$

$$j = 1, 2, .., p$$

3. Computation

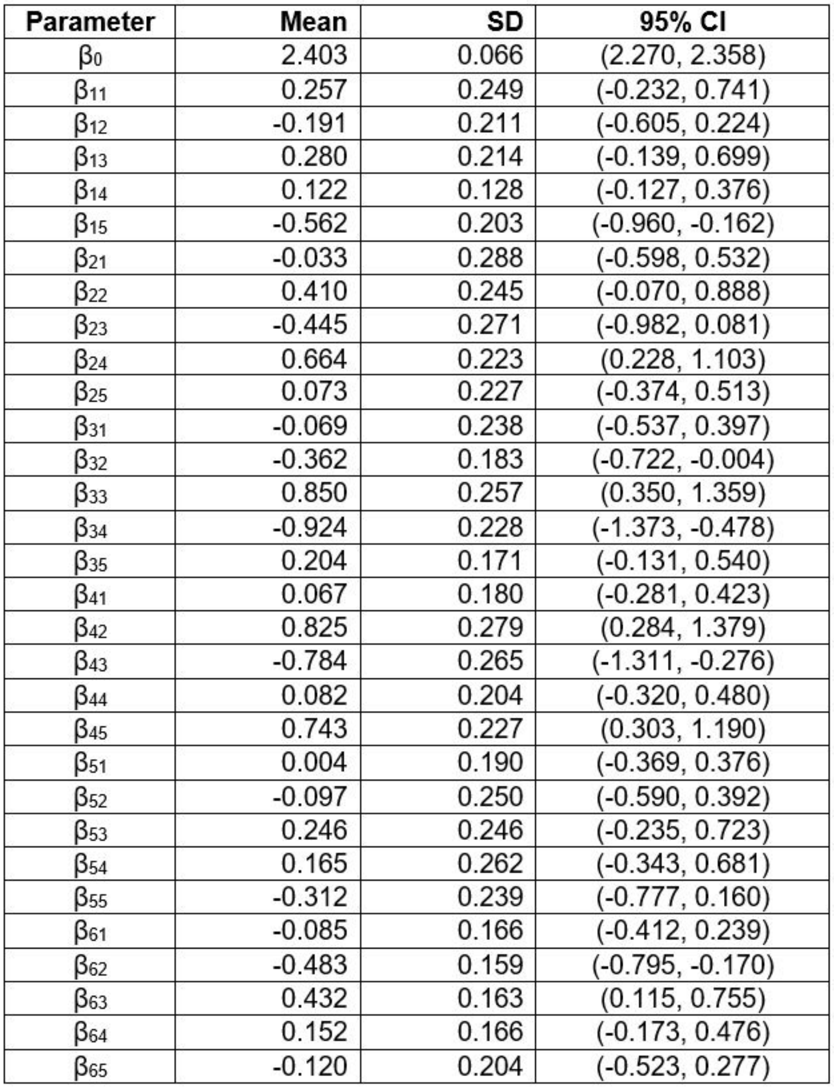
The models were run using MCMC sampling by utilizing the package 'rjags' in R. Two chains were run with the first 10,000 iterations discarded as burn-in and thinning set to 5. 250,000 posterior samples for each chain were retained. A sensitivity analysis of the prior was conducted, and the results were not sensitive to the prior. The trace plots indicated convergence. The effective sample sizes were greater than 1,000 and Gelman and Rubin's Convergence Diagnostic were less than 1.1, further implying convergence.

4. Model Comparisons

The DIC was 368.6, 313.9, and 327.8 for Model 1, Model 2, and Model 3 respectively. The WAIC was 406.2, 342.1, and 364.2 for model 1, model 2, and model 3 respectively. DIC and WAIC were lowest for Model 2, so it is the model that best fits the data. DIC and WAIC are appropriate to use because the posterior densities of the parameters are approximately normal and the models have the same likelihood.

5. Results

The locations and times that are most predictive of the number of hurricanes were found by looking at the 95% credible intervals. In Table 1, β_0 is the intercept and β_{ij} are the 30 independent variables with i = 1, 2, . . . , 6 as the months and j = 1, 2, . . . , 5 as the latitudes.



Parameter	Mean	SD	95% CI
β_0	2.403	0.066	(2.270, 2.358)
β_{11}	0.257	0.249	(-0.232, 0.741)
β_{12}	-0.191	0.211	(-0.605, 0.224)
β_{13}	0.280	0.214	(-0.139, 0.699)
β_{14}	0.122	0.128	(-0.127, 0.376)
β_{15}	-0.562	0.203	(-0.960, -0.162)
β_{21}	-0.033	0.288	(-0.598, 0.532)
β_{22}	0.410	0.245	(-0.070, 0.888)
β_{23}	-0.445	0.271	(-0.982, 0.081)
β_{24}	0.664	0.223	(0.228, 1.103)
β_{25}	0.073	0.227	(-0.374, 0.513)
β_{31}	-0.069	0.238	(-0.537, 0.397)
β_{32}	-0.362	0.183	(-0.722, -0.004)
β_{33}	0.850	0.257	(0.350, 1.359)
β_{34}	-0.924	0.228	(-1.373, -0.478)
β_{35}	0.204	0.171	(-0.131, 0.540)
β_{41}	0.067	0.180	(-0.281, 0.423)
β_{42}	0.825	0.279	(0.284, 1.379)
β_{43}	-0.784	0.265	(-1.311, -0.276)
β_{44}	0.082	0.204	(-0.320, 0.480)
β_{45}	0.743	0.227	(0.303, 1.190)
β_{51}	0.004	0.190	(-0.369, 0.376)
β_{52}	-0.097	0.250	(-0.590, 0.392)
β_{53}	0.246	0.246	(-0.235, 0.723)
β_{54}	0.165	0.262	(-0.343, 0.681)
β_{55}	-0.312	0.239	(-0.777, 0.160)
β_{61}	-0.085	0.166	(-0.412, 0.239)
β_{62}	-0.483	0.159	(-0.795, -0.170)
β_{63}	0.432	0.163	(0.115, 0.755)
β_{64}	0.152	0.166	(-0.173, 0.476)
β_{65}	-0.120	0.204	(-0.523, 0.277)

Table 1: Posterior summaries of parameters

In Table 1, there are variables with 95% credible intervals that do not span 0 and are thus significant effects. These variables are β_{15}, β_{24}, β_{32} , β_{33} , β_{34} , β_{42} , β_{43} , β_{45} , β_{62}, and β_{63}. Notice there are no significant effects for latitude 1, the latitude closest to the US Atlantic Coast. Additionally, most of the significant effects happen in months 3-6, the months closest to hurricane season.

6. Prediction

The predicted number of tropical storms in the final year was found by checking the predictive posterior distribution of Y_{50} using Model 2. The posterior mean is 17.6 with standard deviation of 10.6 and 95% credible interval (4, 44). The model does predict tropical storms to hit the US Atlantic Coast in the final year of the dataset. Figure 1 shows the posterior predictive distribution.

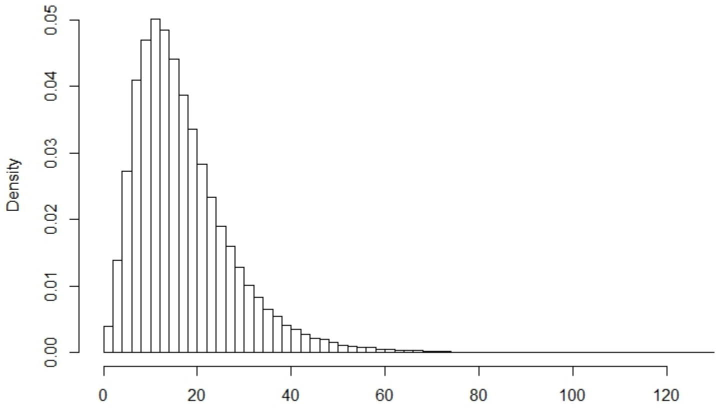


Figure 1: Posterior predictive distribution of number of tropical storms for year = 50

Jags Code for Model 2 (the final model):

```
#Model 2 average over longitude. finallat is the X matrix

Y <-Y[1:49]

n <-49

p<-30

data <- list(Y=Y,p=p,n=n,finallat=finallat)

model_string <- textConnection("model{

          # Likelihood

          for(i in 1:n){

          Y[i] ~ dpois(lambda[i])

          lambda[i] <- exp(beta0+inprod(beta[],finallat[i,]))

          }

          # Priors

          for(i in 1:p){

          beta[i] ~ dnorm(0,0.001)

          }

          beta0 ~ dnorm (0,0.001)

         #Predictions

          ypred~dpois(exp(beta0+inprod(finallat[50,],beta[]))

          }")

model <- jags.model(model_string,data = data, n.chains=2)

update(model, 10000)

params  <- c("beta0", "beta","ypred")

samples3 <- coda.samples(model, variable.names=params, thin=5, n.iter=250000)
```