ST 540 Exam 2, Hemant Kumar, April 16, 2018

1. Introduction

The number of tropical storms that make landfall on US Atlantic Coast is related to the sea surface temperatures in the six months preceding the hurricane season. This paper identifies the locations and months which are most predictive of number of storms.

2. Methods

The response variable (number of storms) is a non-negative discrete random variable and Poisson distribution was chosen to calculate the data likelihood. Further uninformative priors have been used for the predictors since no information on predictive power of any given location or month is available a priori.

The following Bayesian model with 60 predictors has been used.

Y_i : Number of storms in the ith year where $i = 1, 2, \ldots, 50$.

$Y_i \sim Poisson(\lambda_i = \exp(\alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{60} X_{i60}))$

$\beta_j = \gamma_j \times \delta_j$; $\gamma_j \sim Bernoulli(q)$; $\delta_j \sim Normal(0, \sigma^2)$; $\alpha \sim Normal(0, \sigma^2)$;

$q \sim Beta(1,1)$; $\sigma^2 \sim InGamma(0.1, 0.1)$;

$Y_i \in \mathbb{N} \cup \{0\}$; $\alpha, \beta_j, \delta_j, X_{ij} \in \mathbb{R}$; $q \in [0,1]$; $\gamma_j \in \{0,1\}$; $\lambda_i > 0$; $j = 1, 2, \ldots, 60$

The γ_i variable works as an identifier to include/exclude predictors.

The model probabilities have been compared to choose the most important predictors. Further, the effect of prior on selection of these predictors has been studied using two additional priors: $Beta(2,1)$ and $Beta(1,10)$ for q.

Notation: The location (`l`) is referred to as 1, 2… 10 based on the `S` variable of provided dataset. The months (`m`) have been referred to as 1, 2… 6 based on the `X` variable of supplied dataset. Thus, `l8.m3` refers to 3rd month of 8th location (x = 2, y = 1).

3. Computation

The model was fitted in R using the `rjags` package. The jags model code is given in the appendix. The models were run with following specifications: burn = 10000, 3 chains, 100000 iterations per chain, and no thinning. The model successfully converged for all three priors as judged from trace plots (see Figure 1 for representation). Most of the predictor coefficients have spike-and-slab histograms.

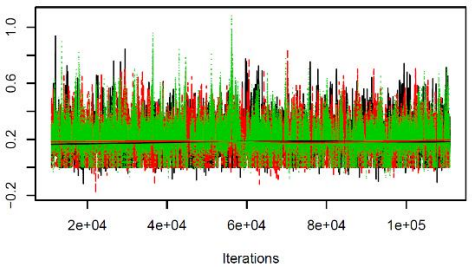


Figure 1: Trace plot of coefficient (βj) of 2nd month of location 7 (x =1, y=1) using Beta(1,1) prior.

4. Model comparisons

The decision to which variables should be included in the model was based on calculating model probabilities. First, the predictors with inclusion probability higher than 0.5 were identified (see Table 1). Then the model probabilities were obtained by calculating the fraction of a given model's presence in the drawn samples (based on the example Dr. Reich's example using Gambia data[1]) with a minor difference. The current problem has 60 predictors and resulting in very large number of combinations of models. I just looked at the combinations of predictors with inclusion probability greater than 0.5 to reduce number of combinations (see Table 2).

Four predictors amongst total 60 have inclusion probability more than 0.5. As can be seen in Table 1, 4th month of 10th location (x=4, y=1) [l10.m4] has inclusion probability nearly one and is clearly the most important predictor. 5th month of 8th location (x=4, y=1) [l8.m5] is the next important predictor. 6th month of location 9 (x=3, y=1) [l9.m6] and 2nd month of location 7 (x=1, y=1) [l7.m2] are the other two important predictors but their importance has to be gauged from model probabilities.

Table 1: Inclusion probability for βj for 3 different priors. Only cases with inclusion probability more than 0.5 have been shown.

Predictor	Prior : $q \sim Beta(a,b)$		
	(1,1)	(2,1)	(1,10)
l10.m4	0.999	0.999	1
l8.m5	0.923	0.906	0.952
l9.m6	0.838	0.834	0.884
l7.m2	0.738	0.711	0.775

[1] Variable selection for the Gambia data (https://www4.stat.ncsu.edu/~reich/BSMdata/SSVS.html)

Table 2 gives the comparison of models with different set of predictors and shows that the model with l7.m2, l8.m5, l9.m6 & l10.m4 occurs most often followed by model with l8.m5, l9.m6 & l10.m4. It becomes clear that l9.m6 and l10.m4 have to be included in the model while inclusion of l7.m2 and l8.m5 is not so clear which agrees with observation made in Table 1.

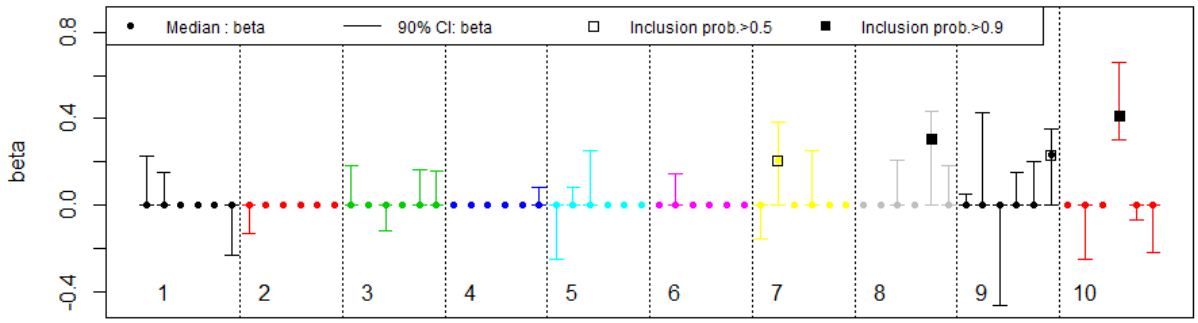
Table 2: Model probabilities of models with different predictor sets for three different priors.

ID	Predictors	Prior : $q \sim Beta(a,b)$		
		(1,1)	(2,1)	(1,10)
A	Intercept, l7.m2, l8.m5, l9.m6, l10.m4	0.579	0.552	0.664
B	Intercept, l8.m5, l9.m6, l10.m4	0.217	0.230	0.193
C	Intercept, l7.m2, l8.m5, l10.m4	0.122	0.119	0.091
D	Intercept, l9.m6, l10.m4	0.032	0.040	0.021

5. Results

Bayesian approach has been used to identify the most important predictors for number of annual storms. The model successfully converges for all three priors and is not very sensitive to prior. Predictors: l10.m4, l8.m5, l9.m6, and l7.m2 are the most important predictors in decreasing order of importance (see Methods for notation). Further, the model with all these four predictors has the highest model probability and this does not depend on priors. The number of storms if positively correlated with the four selected predictors (note the positive value of median of coefficients in Figure 2).

Figure 2: Coefficient quantiles (90% credible interval, median) and inclusion probability of predictors with Beta(1,1) prior. Each subpanel denotes a location indicated by the text label. Each subpanel has 6 months starting from 1st month on the left to 6th on the right. For example, the rightmost panel represents 10th location and the 4th month has inclusion probability greater than 0.9 (shown with filled square).



The final model with all important predictors with Beta(1,1) prior is:

$$Y_i = 2.331 + 0.205 * l7.m2 + 0.305 * \text{l8.m5} + 0.232 * \text{l9.m6} + 0.411 * \text{l10.m4}$$

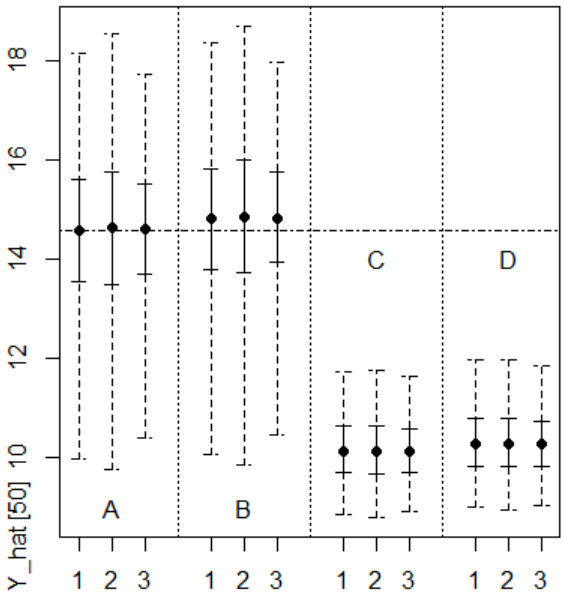
It may appear that 3rd month of 9th location [l9.m3] is unjustly left out despite having a strong negative coefficient along with few others (Figure 2). However, l9.m3 has a very strong spike in in its histogram at 0 and hence it inclusion probability is less than 0.5 and many other predictors also show this behavior. It should be noted that the selected predictors have non-zero median unlike l9.m3.

6. Prediction

The four models given in Table 2 have been used to predict the expected value of number of storms in 50th year. Using the model from equation given in results section, the estimated number of storm is 14.57 with 95% CI of [9.97, 18.15].

We can note from Figure 3 that effect of prior is minimal on model prediction but that of model is impactful. Also, the predicted value by models A & B [14.57 & 14.81 with 1st prior: Beta(1,1)] is drastically different from that of models C & D [10.14 & 10.80 with 1st prior: Beta(1,1)]. This is perhaps due to absence of predictor important l8.m5 in models C and D.

Figure 3: Estimate of number of storms in 50th year ($\widehat{E[Y_{50}]}$) using 4 different models (A, B, C, D) and 3 different priors (1, 2, 3). The filled circles show the median value and solid and dotted bars show 50% and 95% CI respectively. Priors: $q \sim Beta(a, b)$ where 1: (1,1), 2: (2,1), 3: (1,10). See Table 2 for definitions of A, B, C, D.



Appendix

Notation: n = 49 (number of years for which data is available); p=60 (number of predictors: 10 locations and 6 months); Y: response variable; X: predictor matrix

Prior 1

```
library("rjags")
n.chains = 3; burn = 10000; n.iter = 10*burn;
data = list(X = SST, n = nrow(SST), Y = Y, p = ncol(SST))
model_str = textConnection("model{
                           # Likelihood
                           for (i in 1:n){
                           Y[i] ~ dpois(lamb[i])
                           log(lamb[i])=alpha+ inprod(X[i,],beta[])}

                           # Priors
                           for (j in 1:p){
                           beta[j] = gamma[j]*delta[j]
                           gamma[j] ~ dbern(q)
                           delta[j] ~ dnorm(0,taue)    }

                           alpha ~ dnorm(0, taue)
                           q ~ dbeta (1,1)
                           taue ~ dgamma(0.1,0.1)
                           }")

model = jags.model(model_str, data = data, n.chains = n.chains)
update(model, burn)
params = c("alpha","beta","gamma","q","delta","taue")
samples_2 <- coda.samples(model, variable.names=params,n.iter=n.iter)
```

Prior 2:

```
q ~ dbeta (2,1)
```

Prior 3

```
q ~ dbeta (1,10)
```

Software used: RStudio 1.1.463; R 3.5.1; Package rjags version 4-8