Midterm 2

Robin Baldeo

ST 540

Introduction

The data consist of metrological data over the past 50 years, each year has 10 measurements of sea surface temperature taken at 10 spatial location in the Atlantic Ocean within the past 6 months. The response is the number of tropical storms that make landfall in the US Atlantic Coast. The goal of my investigation is to model the data using three models then pick the best model and use this model to predict the number of storms to make landfall in the 50th year and the locations most predictive of these tropical storms.

Methods

For my experiment I used three models. All models are Poisson regression models with three different shrinkage priors. Poisson was used because the response consists of count data. Shrinkage or informative priors were used because the numbers of covariates(p) exceeds the numbers of observations(n) and because of this I think some of the covariates may not be significant, and contributed noise to the overall model. Let Y_i be the number of tropical storms i = 1, . . . , 50 .

First model:

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

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

$$\beta_j \sim DoubleExpo(0, \sigma^2),\ \sigma^2 \sim InvGamma(.01, .01)$$

Second model:

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

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

$$\beta_j \sim Norm(0, \sigma^2),\ \sigma^2 \sim InvGamma(.01, .01)$$

Third model:

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

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

$$\beta_j \sim Cauchy(0, \sigma^2),\ \sigma^2 \sim InvGamma(.01, .01)$$

Computation

Models were computed in R and JAGS. First the raw data containing the covariates was transformed into a 50 x 60 matrix. The data was divided into training and testing data. The first 44 rows were considered training data and the remaining testing. For all three models this data along with response was packaged into a list and passed into JAGS. JAGS was set up to run 2 chains with 20000 iterations for each chain where 10000 were cast away as burn-ins. Resulting in a total of 20000 posterior samples for all parameters with 10 thinning. All three of my models parameters converge where the Gelman test for all parameters is less than 1.1 for both the point estimates and upper CI. Samples size are large with model 3 showing the minimum sample size of 549 for parameter with month = 4 with location latitude = 4 and longitude= 1. As a result of the Gelman test and the large sample size, there is good convergence.

Comparisons

Model Comparisons

Models	**Mean deviance**	**Penalty**	**Penalized deviance**
Model 1	229.4	26.97	256.3
Model 2	225	30.84	255.8
Model 3	224	32.64	256.6

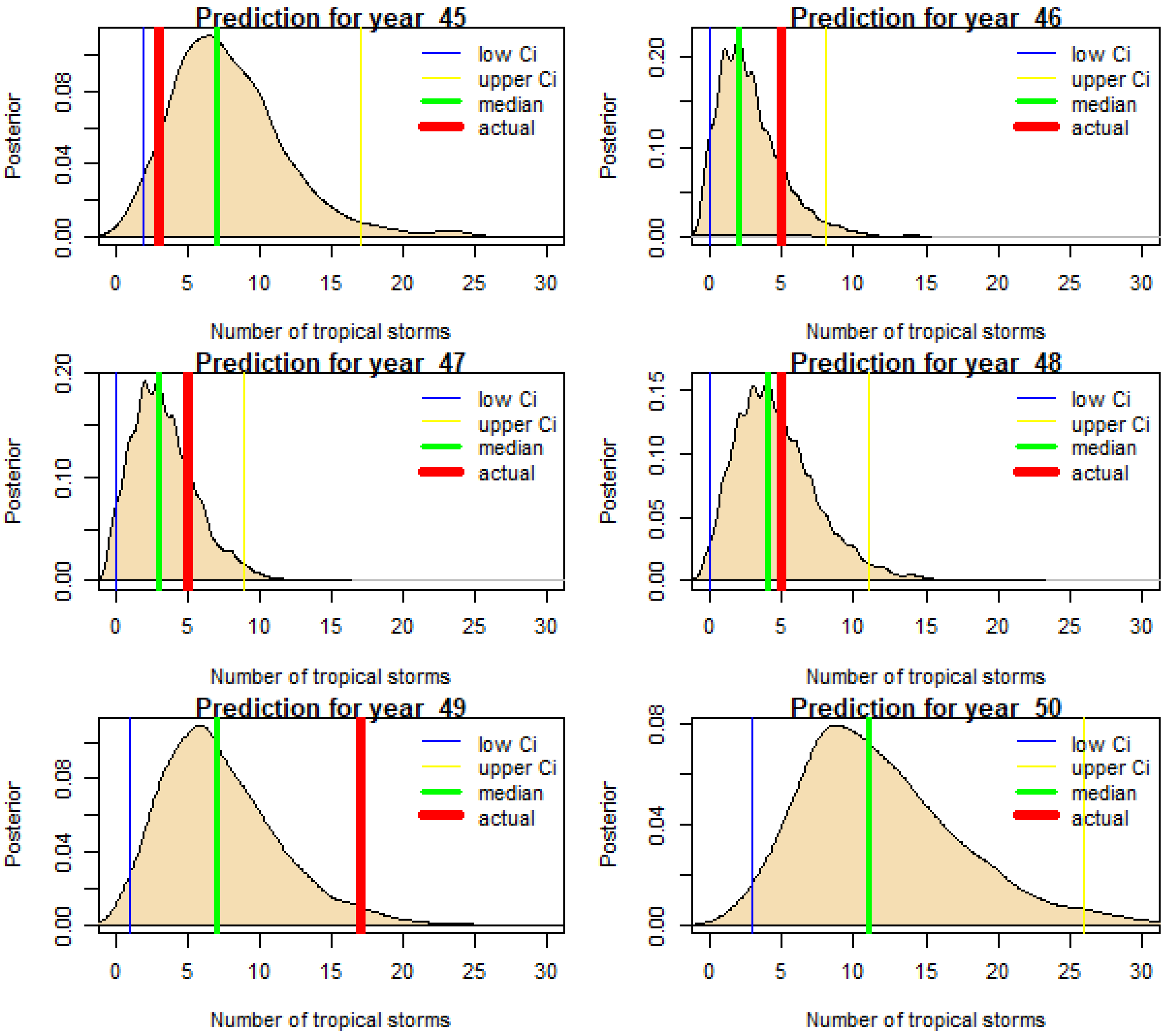
The table above shows the DIC calculation for each model. Model 1 has the largest mean deviance while model 3 has the smallest mean deviance indicating that model 3 fits the data better in comparison to the two other models. With penalty (PD), model 1 has the smallest followed by model 2 then model 3. Indicating that model 1 is the least complicated model out of the 3 proposed models. Penalized deviance (DIC) is the smallest for model 2 and largest for model 3. These differences are so small it’s not definitive. Because the DIC was not definitive, model 1 was selected as the best model because it was the least complicated models because of the low PD.

Results

Statistically Significant Covariates						
Alias	**Month**	**Latitude**	**Longitude**	**2.50%**	**50%**	**97.50%**
beta_58	4	4	1	0.02	0.25	0.49
beta_47	5	2	1	0.01	0.16	0.38
beta_27	3	4	0	0.01	0.15	0.38

Out of the 60 covariates only 3 were statistically significant because 0 was not in the 95% credible interval. These variables are shown in the table above, with the corresponding summary statistics. As a result, beta_58, beta_47 and beta_27 were indicated by my chosen model as most predictive of these tropical storms.

Prediction



The prediction for year 50 is 11 storms with a 95% credible interval of 3 and 26 (As shown in the plot above for year 50). I have also shown my prediction against the actual storm count for year 45,46,47,48, and 49. Years 45, 46 ,47 and 48 the predicted values are very close to the actual number of tropical storms. However, with year 49 my prediction is outside the credible set of 1 and 16 where the predicted values is 7 and the actual value is 17.

Appendix

Model 1 Code:

```
data <- list(n=n,p=p,Y=y,X=x, n_ = n_, X_= x_)

model_string <- textConnection("model{

 # Likelihood
 for(i in 1:n){
  Y[i] ~ dpois(pr[i])
  log(pr[i]) =  alpha +inprod(X[i,],beta[])
 }

 # Priors
 for(j in 1:p){
  beta[j] ~  ddexp(0,taub)
 }

   alpha ~ dnorm(0,0.001)
   taub ~ dgamma(0.1, 0.1)

 # Predictions
   for(i in 1:n_){
    Y_[i] ~ dpois(pr_[i])
    log(pr_[i]) =  alpha +inprod(X_[i,],beta[])
 }

 }")

model <- jags.model(model_string,data = data, n.chains=2,quiet=TRUE)
update(model, 10000, progress.bar="none")
params <- c("beta", "Y_")
samples <- coda.samples(model,variable.names=params,n.iter=20000,thin=10, progress.bar="none")

DIC<- dic.samples(model,n.iter=20000,n.thin = 10, progress.bar="none")
DIC

summary(samples)
gelman.diag(samples)
effectiveSize(samples)
```