Sarah Karamarkovich
ST 540 – Midterm 2
Spring 2019

Introduction

Tropical storms can be detrimental to towns and the people who live in them. Being able to predict the number of tropical storms that make landfall on the US Atlantic Coast would help towns and people prepare and reduce the amount of damage these storms can cause. To predict the number of tropical storms that make landfall, sea surface temperatures are taken six times throughout the year at ten different locations. Our goal is to identify when and where sea surface temperatures are the most predictive of the number of tropical storms that hit the US Atlantic Coast.

Methods and Computations

I will compare three different models, each using a Bayesian Poisson Regression. The dependent variable—number of tropical storms that hit the US Atlantic Coast—is a count variable and thus not normally distributed, which is why I used a Poisson regression. Each model follows the same general format—$Y_i \sim \text{Poisson}(\lambda_i)$ and $\log(\lambda_i) = \gamma + X_{1i}\,\alpha_1 + \ldots + X_{ni}\,\alpha_n$—where X_{ni} are the covariates, α_n are the regression coefficients, and γ is the intercept. I will use uninformative priors for γ and the αs ~ Normal(0, 0.1).

Model Comparisons

Three Bayesian Poisson regressions and three Maximum Likelihood Estimation regressions will be compared. The Bayesian Poisson regressions are described here. The first model takes the average sea surface temperature across all months and locations and uses these averages to predict the number of tropical storms: $Y_i \sim \text{Poisson}(\lambda_i)$ and $\log(\lambda_i) = \gamma + \overline{X}_i\,\alpha$. The second model looks at how the monthly average (i.e., average for each month across all ten locations) is associated with the number of tropical storms: $Y_i \sim \text{Poisson}(\lambda_i)$ and $\log(\lambda_i) = \gamma + M_{1i}\,\alpha_1 + \ldots + M_{6i}\alpha_6$. The third model uses the average sea surface temperature by location (i.e., average for each location across all six months) to predict the number of tropical storms: $Y_i \sim \text{Poisson}(\lambda_i)$ and $\log(\lambda_i) = \gamma + L_{1i}\,\alpha_1 + \ldots + L_{10i}\alpha_{10}$.

To compare the Bayesian models the Watanabe-Akaike information criteria (WAIC) will be used. I chose to use WAIC over the Bayesian information criteria (BIC) because the three models do not have the same likelihood. In addition to WAIC, posterior predictive checks were run to determine Bayesian *p*-values. Although no formal comparison between the MLE and Bayesian models were conducted, both were run to see if the same covariates were statistically significantly associated with the number of tropical storms.

Results

Model One

The MLE determined that the average sea surface temperature across all months and locations was a statistically significant predictor of the number of tropical storms (b = 1.454, $p < .001$). The Bayesian Poisson regression for this model converged and also indicated that the average sea surface temperature was statistically significant ($\alpha = 1.453$, 95% credible set [1.212, 1.694]).

Model Two

Looking at the monthly average sea surface temperature, the MLE model indicated that three months were statistically significantly associated with the number of tropical storms that year—Month 2 (b = 0.769, $p = .005$), Month 4 (b = 0.782, $p = .003$), and Month 6 (b = 0.535, $p = .010$). The same months were found to be statistically significant in the Bayesian Poisson as well—Month 2 ($\alpha = 0.764$, 95% credible set [0.228, 1.307]), Month 4 ($\alpha = 0.772$, 95% credible set [0.248, 1.297]), Month 6 ($\alpha = 0.532$, 95% credible set [0.121, 0.937]). On average, higher temperatures in these months resulted in more storms that year.

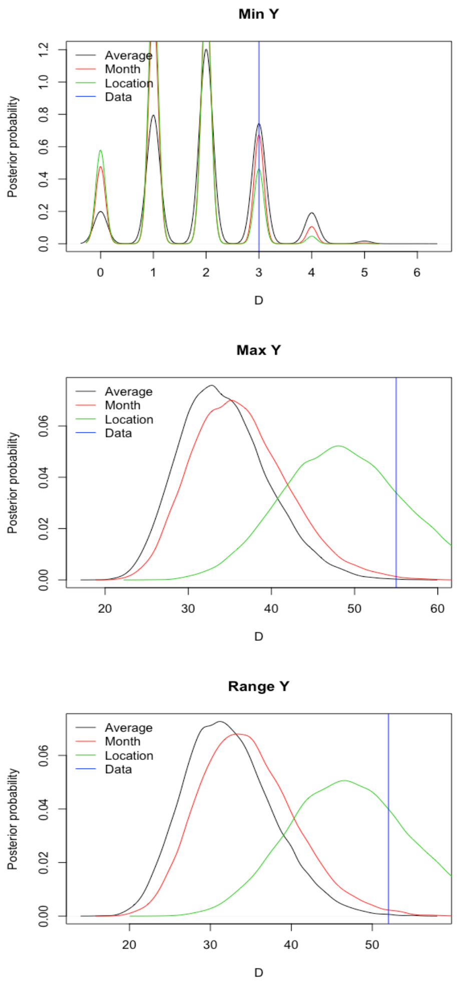
Model Three

Model three looked at the average sea surface temperature for the ten locations. Half of the locations were statistically significantly associated with the number of tropical storms in both the MLE and the Bayesian Poisson regression. When temperatures were higher at Location 3 (b = 0.253, $p = .001$; $\alpha = 0.256$, 95% credible set [0.107, 0.401]), Location 7 (b = 0.300, $p < .001$; $\alpha = 0.302$, 95% credible set

[0.182, 0.424]), Location 9 (b = 0.338, *p* < .001; α = 0.339, 95% credible set [0.218, 0.460]), and Location 10 (b = 0.450, *p* < .001; α = 0.451, 95% credible set [0.344, 0.558]), that year had more tropical storms land on average than when temperatures were lower in those locations. Conversely, on average, when temperatures were low at Location 4, there were more tropical storms (b = -0.199, *p* = .017; α = -0.200, 95% credible set [-0.367, -0.035]).

Model Comparisons

When using WAIC to compare models, the lower the WAIC, the better the model fit. Model One had a WAIC of 420, Model Two had a WAIC of 436, and Model Three had a WAIC of 373. Therefore, Model Three was the best fitting model for the data, whereas Model Two was the worst fitting model.



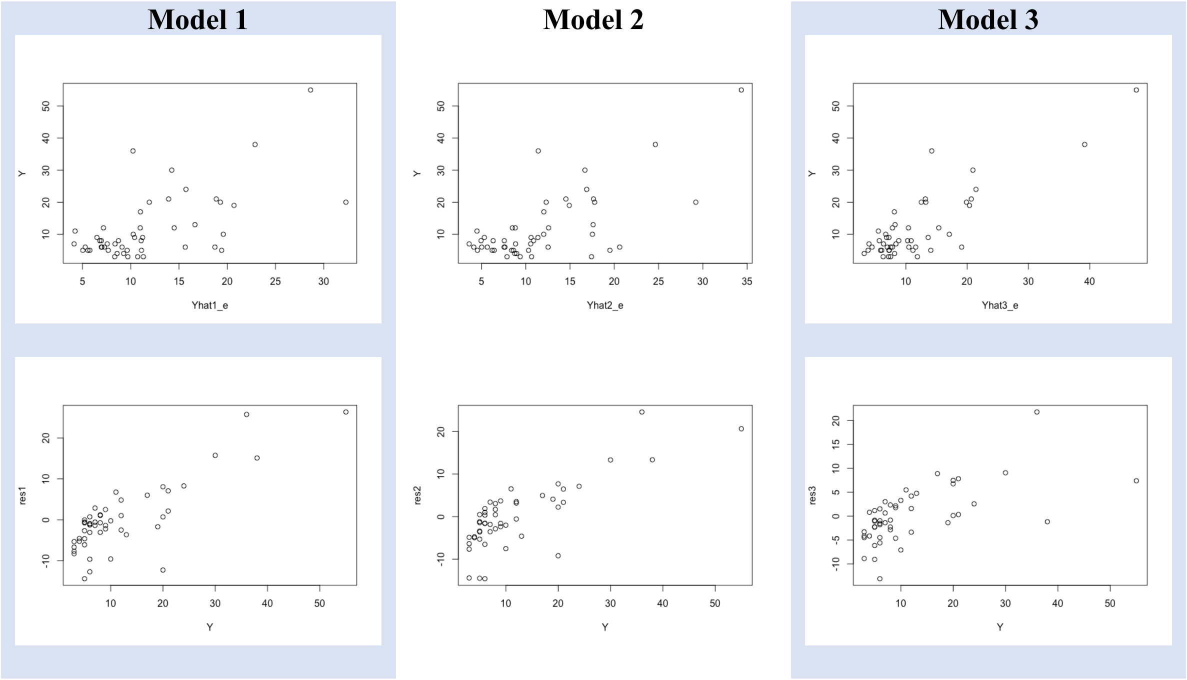
Bayesian *p*-values

	Model 1	*Model 2*	*Model 3*
Minimum	.0651	.0230	.0104
Maximum	.0002	.0025	.2000
Range	.0007	.0050	.2578

The above figure contains the plots for the posterior predictive checks and the Bayesian *p*-values. From these, we can confirm that Model Three is the best fitting of the three models because its Bayesian *p*-values are closest to .50. That being said, it does not appear to be a great model, especially for the smaller values of Y (the number of tropical storms).

Prediction

As we can see in the figure below, none of the models were particularly accurate in predicting the number of tropical storms that landed per year. All models' residuals appear to be heteroscedastic and non-linear.



```
###Model 1###
poisson_model1 <- "model{
#Likelihood
for(i in 1:n){
Y[i] ~ dpois(lambda[i])
log(lambda[i]) <- gamma + beta*X[i]
}
#Priors
gamma ~dnorm(0,.1)
beta ~dnorm(0,.1)
#WAIC calculations
for(i in 1:n){
like1[i] <- dpois(Y[i],lambda[i])
}
#Posterior preditive checks
  for(i in 1:n){
Y1_pred[i]    ~ dpois(lambda[i])
}
D1[1] <- min(Y1_pred[])
D1[2] <- max(Y1_pred[])
D1[3] <- max(Y1_pred[])-min(Y1_pred[])
}"

dat1 <- list(Y=Y, n=n, X=averages)
model1 <- jags.model(textConnection(poisson_model1),data = dat1,n.chains=3)
update(model1, 10000)
samp1 <- coda.samples(model1,
                   variable.names=c("gamma","beta"),
                   n.iter=20000)
waic1   <- coda.samples(model1,
                      variable.names=c("like1"),
                      n.iter=20000, progress.bar="none")
like1   <- waic1[[1]]
fbar1   <- colMeans(like1)
Pw1     <- sum(apply(log(like1),2,var))
WAIC1   <- -2*sum(log(fbar1))+2*Pw1
D1 <- coda.samples(model1,
                 variable.names=c("D1"),
                 n.iter=20000)
D1 <- D1[[1]]

###Model 2###
fulldata <- as.data.frame(melt(X))
colnames(fulldata)<-c("Month", "Location", "Year", "SST")
ave_month <- spread(fulldata, Month, SST)
ave_month <- aggregate(ave_month, by=list(ave_month$Year), FUN=mean)
Month <- as.matrix(ave_month[,4:9])

poisson_model2 <- "model{
# Likelihood
for(i in 1:n){
```

```
Y[i] ~ dpois(lambda[i])
log(lambda[i]) <- gamma + alpha[1]*Month[i,1] + alpha[2]*Month[i,2] +  alpha[3]*Month[i,3] +
        alpha[4]*Month[i,4] + alpha[5]*Month[i,5] + alpha[6]*Month[i,6]
like[i] <- dbin(Y[i],lambda[i],1)
}
#Priors
gamma ~dnorm(0,.1)
for(j in 1:6){
  alpha[j] ~ dnorm(0,0.1)
}
#WAIC calculations
for(i in 1:n){
like2[i] <- dpois(Y[i],lambda[i])
}
#Posterior preditive checks
 for(i in 1:n){
Y2_pred[i]   ~ dpois(lambda[i])
}
D2[1] <- min(Y2_pred[])
D2[2] <- max(Y2_pred[])
D2[3] <- max(Y2_pred[])-min(Y2_pred[])
}"
dat2 <- list(Y=Y, Month=Month, n=n)
model2 <- jags.model(textConnection(poisson_model2),data = dat2,n.chains=3)
update(model2, 10000)
samp2 <- coda.samples(model2,
             variable.names=c("gamma","alpha"),
             n.iter=20000)
waic2   <- coda.samples(model2,
               variable.names=c("like2"),
               n.iter=20000, progress.bar="none")
like2   <- waic2[[1]]
fbar2   <- colMeans(like2)
Pw2     <- sum(apply(log(like2),2,var))
WAIC2   <- -2*sum(log(fbar2))+2*Pw2
D2 <- coda.samples(model2,
            variable.names=c("D2"),
            n.iter=20000)
D2 <- D2[[1]]
```

###Model 3###

```
ave_loc <- spread(fulldata, Location, SST)
ave_loc <- aggregate(ave_loc, by=list(ave_loc$Year), FUN=mean)
Location <- as.matrix(ave_loc[,4:13])
poisson_model3 <- "model{
# Likelihood
for(i in 1:n){
Y[i] ~ dpois(lambda[i])
log(lambda[i]) <- gamma + alpha[1]*Location[i,1] + alpha[2]*Location[i,2] +
        alpha[3]*Location[i,3] + alpha[4]*Location[i,4] + alpha[5]*Location[i,5] +
        alpha[6]*Location[i,6] + alpha[7]*Location[i,7] + alpha[8]*Location[i,8] +
        alpha[9]*Location[i,9] + alpha[10]*Location[i,10]
like[i] <- dbin(Y[i],lambda[i],1)
}
#Priors
gamma ~dnorm(0,.1)
for(j in 1:10){
alpha[j] ~ dnorm(0,0.1)
}
#WAIC calculations
for(i in 1:n){
like3[i] <- dpois(Y[i],lambda[i])
}
#Posterior preditive checks
  for(i in 1:n){
Y3_pred[i]    ~ dpois(lambda[i])
}
D3[1] <- min(Y3_pred[])
D3[2] <- max(Y3_pred[])
D3[3] <- max(Y3_pred[])-min(Y3_pred[])
}"
dat3 <- list(Y=Y, Location=Location, n=n)
model3 <- jags.model(textConnection(poisson_model3),data = dat3,n.chains=3)
update(model3, 10000)
samp3 <- coda.samples(model3,
                  variable.names=c("gamma","alpha"),
                  n.iter=20000)

waic3   <- coda.samples(model3,
                     variable.names=c("like3"),
                     n.iter=20000, progress.bar="none")
like3   <- waic3[[1]]
fbar3   <- colMeans(like3)
Pw3     <- sum(apply(log(like3),2,var))
WAIC3   <- -2*sum(log(fbar3))+2*Pw3
D3 <- coda.samples(model3,
                variable.names=c("D3"),
                n.iter=20000)
D3 <- D3[[1]]

###Bayesian p-values###
D0   <- c(3, 55, 52)
Dnames <- c("Min Y", "Max Y", "Range Y")
pval1 <- rep(0,3)
names(pval1)<-Dnames
pval2 <- pval1
pval3 <- pval1

for(j in 1:3){
  plot(density(D1[,j]),
       xlab="D",ylab="Posterior probability",
      main=Dnames[j])
  lines(density(D2[,j]),col=2)
  lines(density(D3[,j]),col=3)
  abline(v=D0[j],col=4)
  legend("topleft",c("Average","Month","Location", "Data"),lty=1,col=1:4,bty="n")

  pval1[j] <- mean(D1[,j]>D0[j])
  pval2[j] <- mean(D2[,j]>D0[j])
  pval3[j] <- mean(D3[,j]>D0[j])
}

###Model Predictions###
beta1 <- samp1[[1]]
coeff1 <- apply(beta1, 2, mean)
Yhat1 <- rep(0, 50)
for(i in 1:n){
  Yhat1[i] <- coeff1[1]*averages[i] + coeff1[2]
}
Yhat1_e <- exp(Yhat2)
plot(Y ~ Yhat1_e)
res1 <- Y – Yhat1_e
plot(res1 ~ Y)

beta2 <- samp2[[1]]
coeff2 <- apply(beta2, 2, mean)
Yhat2 <- rep(0, 50)
for(i in 1:n){
  Yhat2[i] <- coeff2[1]*Month[i,1] + coeff2[2]*Month[i,2] + coeff2[3]*Month[i,3] +
    coeff2[4]*Month[i,4] + coeff2[5]*Month[i,5] + coeff2[6]*Month[i,6] + coeff2[7]
}
Yhat2_e <- exp(Yhat2)
plot(Y ~ Yhat2_e)
res2 <- Y - Yhat2_e
plot(res2 ~ Y)

beta3 <- samp3[[1]]
coeff3 <- apply(beta3, 2, mean)
Yhat3 <- rep(0, 50)
for(i in 1:n){
  Yhat3[i] <- coeff3[1]*Location[i,1] + coeff3[2]*Location[i,2] + coeff3[3]*Location[i,3] +
    coeff3[4]*Location[i,4] + coeff3[5]*Location[i,5] + coeff3[6]*Location[i,6] +
    coeff3[7]*Location[i,7] + coeff3[8]*Location[i,8] + coeff3[9]*Location[i,9] +
    coeff3[10]*Location[i,10] + coeff3[11]
}
Yhat3_e <- exp(Yhat3)
plot(Y ~ Yhat3_e)
res3 <- Y - Yhat3_e
plot(res3 ~ Y)
```