MATH 640 Final Project

Jason Michaels (jam521), Niko Paulson (ndp32), Miranda Seitz-McLeese (mgs85)

1 Introduction

This analysis will be done on a data set of a variety of measurements about severe weather in the United States. The data set contains a variety of measures from severe weather events in the United States from 1996-2016. For this project we focused on the deaths directly attributable to the event. Understanding how and at what rate severe weather events become lethal in the United States has tremendous public health impacts. In this paper we compare four possible models for the deaths: The traditional Poisson and negative binomial distributions, as well as the zero inflated variant of each.

The remainder of this analysis is organized as follows: Section 2 discusses and derives the models. Section 3 describes the results of the analysis. And Section 4 contains the conclusions.

2 Methods

The deaths attributed to a severe weather event is 'count' data. The most common model used for count data is the Poisson distribution. However for some weather events, the negative binomial model is a better fit, because the Poisson distribution assumes that the events being counted occur independently.

Fortunately, the vast majority of severe weather events in the United States involve no deaths, therefore we wanted to also account for the possibility of structural zeros, therefore we also considered zero inflated variants. These distributions are created by returning 0 with probability σ and sampling from the original distribution with probability $(1-\sigma)$.

We will derive and fit a model for each of the four distributions and see if there is a difference in our results and evaluate to determine which model best fits the data.

2.1 Poisson

The Poisson model has one parameter, λ represents the expected number of occurances of the event of interest. For a single random variable x, the probability density is:

$$p(x|\lambda) = \frac{\lambda^x e^{-\lambda}}{x!}.$$

Thus the likelihood for a sum of Poisson random variables can be written as follows:

$$\mathcal{L}(X|\lambda) \propto \lambda^{n\bar{X}} e^{-n\lambda}.$$

Since we want the data to speak for itself, we will use a non-informative random variable, namely, Jeffreys' prior. For the Poisson distribtion this is given as follows:

$$\pi(\lambda) \propto \lambda^{1/2-1} e^{-0\cdot\lambda}.$$

We recognize this as the kernel of an improper gamma distribution. Combining the likelihood and the prior distribtion yields the following posterior distribution:

$$p(\lambda|X) = \lambda^{n\bar{X}+1/2-1} e^{-n\lambda}.$$

We recognize this as the kernel of a gamma distribution, namely

$$\lambda|X \sim \mathcal{G}amma(n\bar{X}+1/2, n).$$

2.2 Negative Binomial

The Negative Binomial model has two parameters, r, p represents the expected number of occurances of the event of interest. For a single random variable x, the probability density is:

$$p(X|r,p) = \frac{\Gamma(r+x)}{\Gamma(r)x!} p^x (1-p)^r.$$

Thus the likelihood for a sum of Negative Binomial random variables can be written as follows:

$$\mathcal{L}(X|r,p) = \left[\prod_{i=1}^{n} \frac{\Gamma(r+x_i)}{\Gamma(r)x_i!}\right] p^{n\bar{X}}(1-p)^{nr}.$$

Since we want the data to speak for itself, we will use a non-informative random variable, namely, Jeffreys' prior. For the Poisson distribtion this is given as follows:

$$\pi(r,p) = r^{1/2}p^{-1}(1-2)^{-1/2}.$$

Combining the likelihood and the prior distribtion yields the following posterior distribution:

$$p(r,p|X) = \left\{\left[\prod_{i=1}^{n} \frac{\Gamma(r+x_i)}{\Gamma(r)x_i!}\right] p^{n\bar{X}}(1-p)^{nr}\right\}\left\{r^{1/2}p^{-1}(1-2)^{-1/2}\right\}$$

From this posterior we obtain the full conditionals. First consider $p|r,X$:

$$p(p|r,X) \propto p^{n\bar{x}-1}(1-p)^{nr+1/2-1}$$

We recognize this as the kernel of a beta distribution, namely

$$p|r,X \sim \mathcal{B}eta(n\bar{X}, nr+1/2).$$

Next consider $r|p,X$:

$$p(r|p,X) \propto \left[\prod_{i=1}^{n} \Gamma(r+x_i)\right]\Gamma(r)^{-n}(1-p)^{nr}r^{1/2}$$

This is not a recognized distribtion. So if we wish to make inferences on r we must use a Metropolis algorithm to sample from it.

2.3 Zero Inflated Poisson

The Zero Inflated Poisson (ZIP) model has two parameters. The parameter p is the probability of a structural zero, and λ corresponds to the parameter in a typical Poisson model. For a single observation x, the probability density is:

$$p(x|p,\lambda) = pI_{x=0}(x) + (1-p)\frac{e^{-\lambda}\lambda^x}{x!}$$

We can write the likelihood as follows:

$$L(p,\lambda|X) = \prod_{x_i=0}\left[p + (1-p)\frac{e^{-\lambda}\lambda^{x_i}}{x_i!}\right]\prod_{x_i\neq 0}\left[(1-p)\frac{e^{-\lambda}\lambda^{x_i}}{x_i!}\right]$$

Bayarri, Berger, and Datta (2008) suggest using the prior distribution $\pi(\lambda,p) \propto \frac{1}{\sqrt{\lambda}}I(0<p<1)$. This gives us the following posterior

$$\prod_{x_i=0}\left[p + (1-p)\frac{e^{-\lambda}\lambda^{x_i}}{x_i!}\right]\prod_{x_i\neq 0}\left[(1-p)\frac{e^{-\lambda}\lambda^{x_i-1/2}}{x_i!}\right]$$

In obtaining our full conditionals, we can simplify this slightly to obtain the following:

$$p(\lambda|X,p) \propto \prod_{x_i=0}\left[p + (1-p)\frac{e^{-\lambda}\lambda^{x_i}}{x_i!}\right]\prod_{x_i\neq 0}\left[e^{-\lambda}\lambda^{x_i-1/2}\right]$$

$$p(p|X,\lambda) \propto \prod_{x_i=0}\left[p + (1-p)\frac{e^{-\lambda}\lambda^{x_i}}{x_i!}\right]\prod_{x_i\neq 0}\left[(1-p)\right]$$

Neither of these distributions is recognizable. We can use a Metropolis-Hastings algorithm to sample from both of them. We will use a beta distribution as a proposal for p, and a gamma for λ. We will tune them to obtain a better acceptance rate.

2.4 Zero Inflated Negative Binomial

The Zero Inflated Negative Binomial (ZINB) model has three parameters σ, the probability of a structural zero, and p, r the usual negative binomial parameters. For a single X the probability density is:

$$p(X|\sigma, p, r) = \sigma I_{X=0}(X) + (1-\sigma)\frac{\Gamma(r+X)}{\Gamma(r)X!}.$$

We take the uniform priors for σ and p as well as the non-informative gamma for r which is $r^{-1/2}$. For a full derivation, see B.1. My posterior is:

$$p(r, \sigma, p|X) \propto (\sigma + (1-\sigma)(1-p)^r)^Z (1-\sigma)^{N-Z}(1-p)^{(N-Z)r} p^{\sum_{i=1}^N X_i} r^{-1/2} \prod_{i=1}^N \left(\frac{\Gamma(r+X_i)}{\Gamma(r)}\right)$$

This distribution does not factor nicely, so I will use the Metropolis algorithm to sample from it. Because this posterior does not suggest any obvious proposal distributions I will sample each independently from a normal distribution centered at θ^*, and with a variance that is tuned to yeild an appropriate acceptance rate.

3 Results

3.1 Poisson

3.2 Negative Binomial

3.3 Zero Inflated Poisson

In tuning the parameters of the two models, slightly different values were chosen for the two different event types. The proposal distribution selected for λ was a gamma(2, 2) for tornados, and a gamma(1, 2) for flash floods. The proposal for p within the tornado model was a beta(1940, 60), whereas it was a beta(2945, 55) in the flash flood model.

For each variable of interest, 20,000 samples were taken. As convergence was not immediately achieved, the first 10,000 samples were discarded as a burn-in. The results are as follows:

Table 1: Results of taking 20,000 samples from the posterior distributions of λ and p, compared between the two event types. The second column relays the mean of the sample for λ, with the 95 percent credible interval in parentheses. The third does the same for p. All means and credible intervals are taken after discarding the first 10,000 samples as a burn-in.

Event Type	λ	p
Tornado	1.872 (1.030, 3.710)	0.971 (0.968, 0.973)
FlashFlood	0.562 (0.293, 0.972)	0.982 (0.981, 0.983)

3.4 Zero Inflated Negative Binomial

4 Discussion

References

[1] NOAA's Severe Weather Data Inventory, `https://www1.ncdc.noaa.gov/pub/data/swdi/stormevents/csvfiles/`. Accessed April 2017.

[2] Bayarri, M., Berger, J., Datta, G. (2008). Objective testing of Poisson versus inflated Poisson models. IMS Collections, 3, 105-121.

A Code

This appendix includes the code used to implement the models.

B Derivations

This appendix will include details on the calculations required to derive our models.

B.1 Zero Inflated Negative Binomial Derivation

The likelihood for the ZINB is

$$\mathcal{L}(X|\sigma,p,r) = \prod_{i=1}^{N} \sigma I_{X=0}(X_i) + (1-\sigma)\frac{\Gamma(r+X_i)}{\Gamma(r)X_i!}$$

For ease of notation let Z be the number of zero values in X, and N be the total number of observations.

$$\begin{aligned}
&= \prod_{X_i=0} \left(\sigma + (1-\sigma)p^{X_i}(1-p)^r \frac{\Gamma(r+X_i)}{\Gamma(r)X_i!}\right) \prod_{X_i\neq 0} \left((1-\sigma)p^{X_i}(1-p)^r \frac{\Gamma(r+X_i)}{\Gamma(r)X_i!}\right) \\
&= \left(\sigma + (1-\sigma)(1-p)^r\right)^Z \prod_{X_i\neq 0} \left((1-\sigma)p^{X_i}(1-p)^r \frac{\Gamma(r+X_i)}{\Gamma(r)X_i!}\right) \\
&\propto \left(\sigma + (1-\sigma)(1-p)^r\right)^Z \prod_{X_i\neq 0} \left((1-\sigma)p^{X_i}(1-p)^r \frac{\Gamma(r+X_i)}{\Gamma(r)}\right) \\
&\propto \left(\sigma + (1-\sigma)(1-p)^r\right)^Z (1-\sigma)^{N-Z}(1-p)^{(N-Z)r} p^{\sum_{i=1}^N X_i} \prod_{i=1}^{n} \left(\frac{\Gamma(r+X_i)}{\Gamma(r)}\right)
\end{aligned}$$

As mentioned in section 2.4 we take the uniform priors for σ and p as well as the non-informative gamma for r which is $r^{-1/2}$. Therefore my joint posterior is:

$$p(r,\sigma,p|X) \propto \left(\sigma + (1-\sigma)(1-p)^r\right)^Z (1-\sigma)^{N-Z}(1-p)^{(N-Z)r} p^{\sum_{i=1}^N X_i} r^{-1/2} \prod_{i=1}^{N} \left(\frac{\Gamma(r+X_i)}{\Gamma(r)}\right)$$

I am now going to take the log of the posterior because it helps with computation

$$\begin{aligned}
\ln\left(p(r,\sigma,p|X)\right) \propto & Z\ln\left(1+(1/\sigma-1)(1-p)^r\right) + Z\ln(\sigma) + (N-Z)\ln(1-\sigma) + (N-Z)r\ln(1-p) \\
& + \sum_{i=1}^{N} X_i \ln(p) - \ln(r)/2 - N\ln(\Gamma(r)) + \sum_{i=1}^{N} \ln(\Gamma(r+X_i))
\end{aligned}$$