ST 540: Midterm 2

Lee Pixton

April 26, 2021

Introduction

The American Black Bear is the most common and widely populated bear species in North America. It can be found in many places, from northeastern Mexico all the way into Alaska. According to the International Union for Conservation of Nature, there are estimated to be twice as many American Black Bears as all other bear species combined on the continent. Despite this, they are rarely seen in populated areas and mostly reside in forests around North America. Using a random sample (n=927) of data collected by citizen scientists in Western North America, we will examine four specific ecoregions and build a statistical model to test for local adaptation of bears by region. These ecoregions are the Marine West Coast Forest (n=123), Mediterranean California (n=171), North American Deserts (n=284), and Northwestern Forested Mountains (n=349). We will examine data showing how often black bears are reported in each of these regions, which includes ecological data such as the proportion of land that is forest, annual average temperature, and human population. This analysis will examine these differences between locations and attempt to determine what niche habitat black bears prefer in the different regions.

Models and Hypothesis

For this analysis, we will build multiple models and then compare the results in order to find the best fitting model. In all cases, the data will be modeled as:

$$Y_i \sim Binomial(N_i, p_i)$$

where Y_i is the number of observed black bears in a specific location in an ecoregion, N_i is the number of observations in that location, and p_i is the true probability of observing a black bear in that location. It should be noted that only 100 of the locations in the 927 sampled recorded one or more instance of sighting a black bear (10.79%). The other 827 observations include a Y value of 0.

There are $p = 8$ covariates included in the models:

- Forest - The proportion of the region that is forest
- Grassland - The proportion of the region that is grassland
- Cropland - The proportion of the region that is cropland
- Temp - Annual average temperature of the region
- Precip - Annual average precipitation in the region
- Population - Human population in the region

- Protected - Indicates whether the region includes protected land
- Ecoregion - One of the four ecoregions named above

We will test a total of four models in this analysis. Each model is a logistic regression model and will be described below. The first model includes *p = 10* covariates, with covariates 8, 9, and 10 being indicator variables for three of the regions and will be modeled as:

$$logit(p_i) = \beta_0 + \sum_{j=1}^{k} X_i\beta_j + \varepsilon_i$$

We select uninformative priors $\beta_j \sim Normal(0, 10^2)$.

For the remaining three models, we look at every region in ecoregion *r* and assume the logistic model:

$$logit(p_i) = \beta_{0r} + \sum_{j=1}^{k} X_i\beta_{jr} + \varepsilon_i$$

where β_{jr} is the effect of covariate *j* in ecoregion *r*. We compare three separate models for β_{jr}

1. Constant slopes where $\beta_{jr} \equiv \beta_j$ for all regions
2. Fixed effects with varying slopes using uninformative priors: $\beta_{jr} \sim Normal(0, 10^2)$
3. Random effects with varying slopes using informative priors: $\beta_{jr} \sim Normal(\mu_j, \sigma_j^2)$

The third model here includes means and variances that are estimated from the data. Each model will be compared with both DIC and WAIC and then a best-fit model will be selected. Once a model is chosen, we will analyze the posterior distribution to test for local adaptation by ecoregion for black bears and discover niche habitats that may exist within ecoregions.

Computation

The models described above were run using Markov Chain Monte Carlo (MCMC) sampling in the R programing language. The software JAGS was used and integrated into R with the library ‘rjags’, which was used to facilitate the MCMC. Two chains were run for each model specified, including a discarded burn-in of 10,000. Each chain then ran for 50,000 iterations with a thinning of 5. Different priors were tested in a sensitivity analysis, and the outcome from the models were not shown to be sensitive to the prior. Following the sampling, the trace plots were examined, and each indicated convergence for the different models. The effective sample sizes were large and the Gelman and Rubin's Convergence Diagnostic were less than 1.01 for all models further pointing to convergence.

Model Comparisons

For each model, both DIC and WAIC were computed and are shown in Table 1. Because each model converged well, we will use DIC and WAIC as the main indicators for model fit. Both the DIC and WAIC were lowest for Model 4, indicating that this is the best of the four models that were fit to the data.

	DIC	DIC penalty	WAIC	WAIC penalty
Model 1	1370	11.01	1532.41	124.2
Model 2	1464	7.98	1609.98	113.25
Model 3	1168	31.13	1340.99	138.13
Model 4	1167	28.15	1337.12	132.45

Table 1: Model comparison values and penalties across the four models

The third model was very close to Model 4 in both measures but had a higher penalty for each. Model 3 is also more complex than Model 4, so by both simplicity and diagnostic measures Model 4 is the best model to analyze the data. It is interesting to note that Model 1 had a lower DIC and WAIC than Model 2, indicating that there is a benefit to including the ecoregion variable even in a basic logistic regression. We will briefly examine Model 1 before turning to Model 4 to complete the analysis.

Results

Although Model 1 did not provide the best fit for the data, it does provide insight into the four ecoregions in a very simple manner. Table 2 shows the beta values for the different ecoregions, using Northwestern Forested Mountains as a baseline. We can see that the probability of seeing a black bear, while holding all other variables constant in this model, remains about the same in both the Marine West Coast Forest and Mediterranean California ecoregions compared to the Northwestern Forested Mountains. Where it differs

	Mean	SD	95% Interval
Northwestern Forested Mountains	0	0	(0, 0)
Marine West Coast Forest	0.016	0.086	(-0.150, 0.188)
Mediterranean California	0.052	0.121	(-0.184, 0.289)
North American Deserts	-1.831	0.397	(-2.724, -1.175)

Table 2: Beta values for ecoregions in Model 1. NW Forested Mountains as baseline

most from the other regions is the North American Deserts. There is a significantly lower chance to observe a black bear in this ecoregion than in any of the others.

Model 4 provides further insight into the analysis. Table 3 shows the beta values for each ecoregion from Model 4. Variables whose 95% credible intervals do not include 0 are considered significant, and are marked with an asterisk (*). Every variable was significant in the Mediterranean California ecoregion, while none were significant in the North American Deserts. Figure 1 shows the variance between the effects of the covariates across ecoregions. The effect of average annual temperature varied the most, while population had comparatively little variance in its effect. Studying the variables appears to show evidence of local adaption among black bears. For example, protected areas in the Marine West Coast Forest ecoregion have an increased likelihood of observing a black bear, while protected regions of the North American Deserts decrease the likelihood. In addition, as temperature increases in the Mediterranean California ecoregion, the probability of observing black bears increases significantly, while in the Marine West Coast Forests as temperature increases that same probability decreases significantly. These large variations within the variables indicates that habitat preference does indeed vary across ecoregion.

	Marine WC Forest	NW Forested Mountains	Mediterranean California	NA Deserts
Intercept	-4.904	-3.176	-6.747	-7.046
Forest	-0.424*	0.247	-0.791*	1.001
Grassland	-0.71*	0.18	1.136*	-0.272
Cropland	0.048	-0.384	-1.116*	-0.205
Temp	-1.371*	0.39*	2.001*	-1.316
Precip	0.429*	-0.139	1.596*	1.307
Population	-0.705*	-0.771	-0.248*	-0.323
Protected	0.975*	-0.046	1.675*	-1.137

Table 3: Beta values for Model 4 by ecoregion. Significance denoted by asterisk (*)

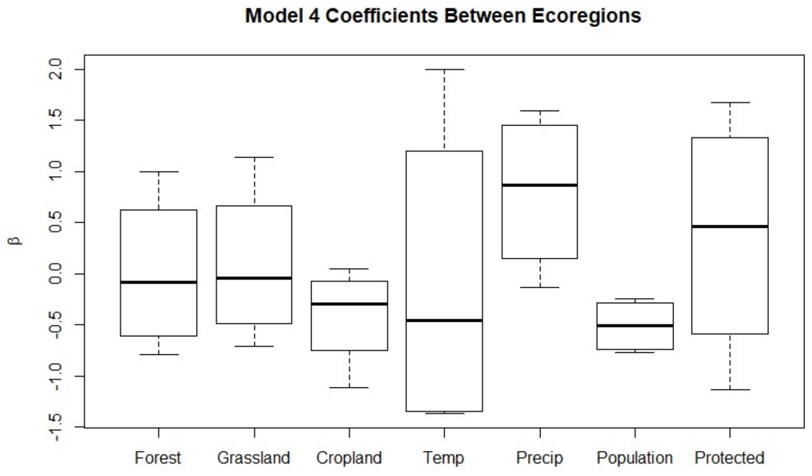

Figure 1: Beta values boxplot for Model 4. Intercept excluded

Discussion

The objective in this analysis was to build a statistical model for the ecological niche of American Black Bears. As discussed above, there is evidence from this model of local adaptation across ecoregions. The only definitive similarity between the ecoregions is that higher populations decrease the probability of observing a black bear. Outside of this covariate, each ecoregion differs in black bear niche and preference. In the Marine West Coast Forests, areas with lower proportion of both forest and grassland and larger amounts of annual average precipitation increase the likelihood of observing black bears. In the Northwestern Forested Mountains, unprotected locations with increased forest proportion, decreased cropland proportion, and higher average temperatures appear to be more likely to allow for black bear observations. Black bears are more likely to be spotted in protected areas with high proportions of grasslands and higher average temperatures in the Mediterranean California ecoregion. Finally, within the North American Deserts, unprotected locations with higher proportions of forests and higher average annual precipitation lead to a higher likelihood of observing these bears.

One limitation to this study is the lack of consistent data across the ecoregions. The Northwestern Forested Mountains had the most data points, followed by the North American Deserts. These regions held a much higher proportion of the data than the Marine West Coast Forests and Mediterranean California. Having a more equitable sample size from each location would be beneficial. In addition, the data is only binary indicators of whether black bears were observed in the outing or not. By obtaining data that is a count of black bears observed rather than binary, a Poisson model could be fit. This in turn could better pinpoint the habitats of the black bears, by estimating the number of black bears in a location rather than a model of the probability of observing one in a given location. It would be interesting and beneficial to the research question to model how many black bears are in a specific location with the given variables within an ecoregion.

Code Appendix

```
### 540 Midterm

# Model 1 - Ecoregion as binary variables
model_string <- textConnection("model{
  # Likelihood
  for(i in 1:n){
    Y[i] ~ dbinom(p[i],N[i])
    logit(p[i]) <- inprod(X[i,],beta[])
  }

  # Priors
  for(j in 1:k){beta[j] ~ dnorm(0,0.01)}

  # WAIC calculations
  for(i in 1:n){
    like[i] <- dbin(Y[i],p[i],N[i])
  }
}")

# Initialize JAGS Requirements
data   <- list(Y=Y,n=n,N=N,X=X,k=k)
params <- c("beta")

# Run model
model <- jags.model(model_string,data = data, n.chains=2,quiet=TRUE)
update(model, 10000, progress.bar="none")
samp <- coda.samples(model, variable.names=params, thin=5, n.iter=50000, progress.bar="none")

# Model 4 - Slopes as Random Effects
Model4_string <- textConnection("model{
  # Likelihood
  for(i in 1:n){
    Y[i] ~ dbinom(p[i],N[i])
    logit(p[i]) <- inprod(X[i,],beta[id[i],])
  }

  # Priors
  for(j in 1:k){
    for(i in 1:4){
      beta[i,j] ~ dnorm(mu[j],tau[j])
    }
    mu[j] ~ dnorm(0,0.01)
    tau[j] ~ dgamma(0.1,0.1)
  }

  # WAIC calculations
  for(i in 1:n){
    like[i] <- dbin(Y[i],p[i],N[i])
  }
}")
```

```
# Initialize JAGS Requirements
data   <- list(Y=Y,n=n,N=N,X=X,k=k,id=id)
params <- c("beta")

# Run model
model4 <- jags.model(model4_string,data = data, n.chains=2,quiet=TRUE)
update(model4, 10000, progress.bar="none")
samp4 <- coda.samples(model4, variable.names=params, thin=5, n.iter=50000, progress.bar="none")
```