Eagle data analysis

Maike Holthuijzen

June 11, 2018

1 Model

The model used to describe the observance of eagle decoys was a logistic regression model of the form:

$$logit(p) = log(p/(1-p)) = \beta_0 + \beta_1 * Year + \beta_2 * Age\_class + \beta 3 * dist\_to\_shoreline \quad (1)$$

Age was a dummy variable, where "adult" was coded as a 1 and "juvenile" was coded as a 0. Year was also a dummy variable; 2003 was coded as 0 and 2004 was coded as 1. Distance to the shoreline (dist_to_shoreline) was centered prior to fitting the model. The model was fit using the `lm` function in R. The variable "substrate" was not included because there were not enough observations in the four substrate classes, and preliminary analyses showed the variable did not contribute to the model (also, see frequency tables below). However, if some of the classes were condensed to result in fewer classes (e.g. if "shore" was combined with "ground"), it might be worth considering its contribution to the model fit. Frequency tables are shown below for Year vs Observed, substrate vs observed, and age vs observed:

```
> with(eagles, table(Year, Observed))
      Observed
Year    0  1
  2003  8 23
  2004 28 32
> with(eagles, table(substrate, Observed))
         Observed
substrate 0  1
  GROUND 23 24
  ROCK   10 20
  SHORE   1  7
  TREE    2  4
> with(eagles, table(age, Observed))
   Observed
age  0  1
  0 25 23
  1 11 32
```

2 Frequentist analysis

The freqentist analysis for a logistic regression analysis assumes that there is no multicollinearity among independent variables, that the observations are independent, and that there is a linear relationship between the independent variables and the log odds. Assumptions for the model were met (only dist_to_shoreline was included as a distance measure as it was correlated with the distance to thalweg).

The final model is given in equation 2:

logit$(\hat{y}) = log(\hat{y}/(1-\hat{y})) = \hat{\beta_0} + \hat{\beta_1} * dist_ + \hat{\beta_2} * age +$ to_shoreline + $\hat{\beta_3} 0 * Year$
$= 0.49631 - 0.05258 * dist\_to\_shoreline + 1.05527 * age - 0.78274 * Year$, (2)where age = 0 was the baseline.

The results of coefficients, their standard errors, and p-values are shown in table 1.

Table 1: Coefficients for frequentist logistic regression model

Estimate	**Estimate**	**Std. error**	**Z val.**	**Pr(>—z—)**
(Intercept)	0.49631	0.47215	1.051	0.2932
dist.s	-0.05258	0.23158	-0.227	0.8204
age	1.05527	0.46546	2.267	0.0234
year01	-0.78274	0.50604	-1.547	0.1219

The only significant predictor was *age* ($p = 0.02$). *dist_to_shoreline* and *year* were not significantly associated with the odds of observing a decoy. The results can be interpreted in the following way:

- Holding *dist_to_shoreline* and *year* at a fixed value, the odds of observing decoys for adults (age = 1) over the odds of observing a decoy for immature (age = 0) is $exp(1.05527) = 2.87$. In terms of percent change, the odds for observing an adult is 187% higher than the odds for immature decoys.
- The coefficient for *Year* says that, holding *dist_to_shoreline* and *age* at a constant value, the odds of observing a decoy for 2004 (year = 1) over the odds of observing a decoy for 2003 (Year = 0) is $exp(-0.78) = 0.45$, which is a 55% decrease. (Note that though there were more decoys overall in 2004, the success rate for observing a decoy in 2004 was only 53%, compared to 2003, when it was 74%.)
- Finally, holding *year* and *age* constant, we could expect a 5% decrease in the odds of observing a decoy for a one-unit increase in distance to shoreline (since $exp(-0.05258) = 0.95$).

3 Bayesian Analysis

The Bayesian analysis differs from the frequentist analysis in that the parameters we wish to estimate ($\beta_0, \beta_1, \beta_2$, and $\beta 3$) are given as *posterior distributions*. This means that each of these parameters has a posterior distribution, which is the result of the data and the prior distributions given in the model fitting process. For this model, the all coefficients except the intercept had the double exponential with location parameter of 1 and $\sqrt(2)$ for the inverse scale parameter (this corresponds to a variance of 1). The intercept had a normal prior distribution mean a mean of 0 and precision of $\frac{1}{25}$ (the precision is the inverse of the variance). The Bayesian model was run with 4 chains. Diagnostics (Gelman diagonal and autocorrelation plots) were inspected to make sure chains adequately converged and that autocorrelation in chains was low.

The results for the Bayesian model were similar to those obtained via the frequentist method. The means and standard deviations of parameter coefficients are given in table 2, and the quantiles are given in table 3. Age was the only variable for which the 2.5% and 97.5% quantiles were greater than 0 (if 0 is contained within these quantiles, then the probability that the posterior distribution of a parameter contains 0 is 0.95). Plots of the posterior distributions for all coefficients are shown in figure 1.

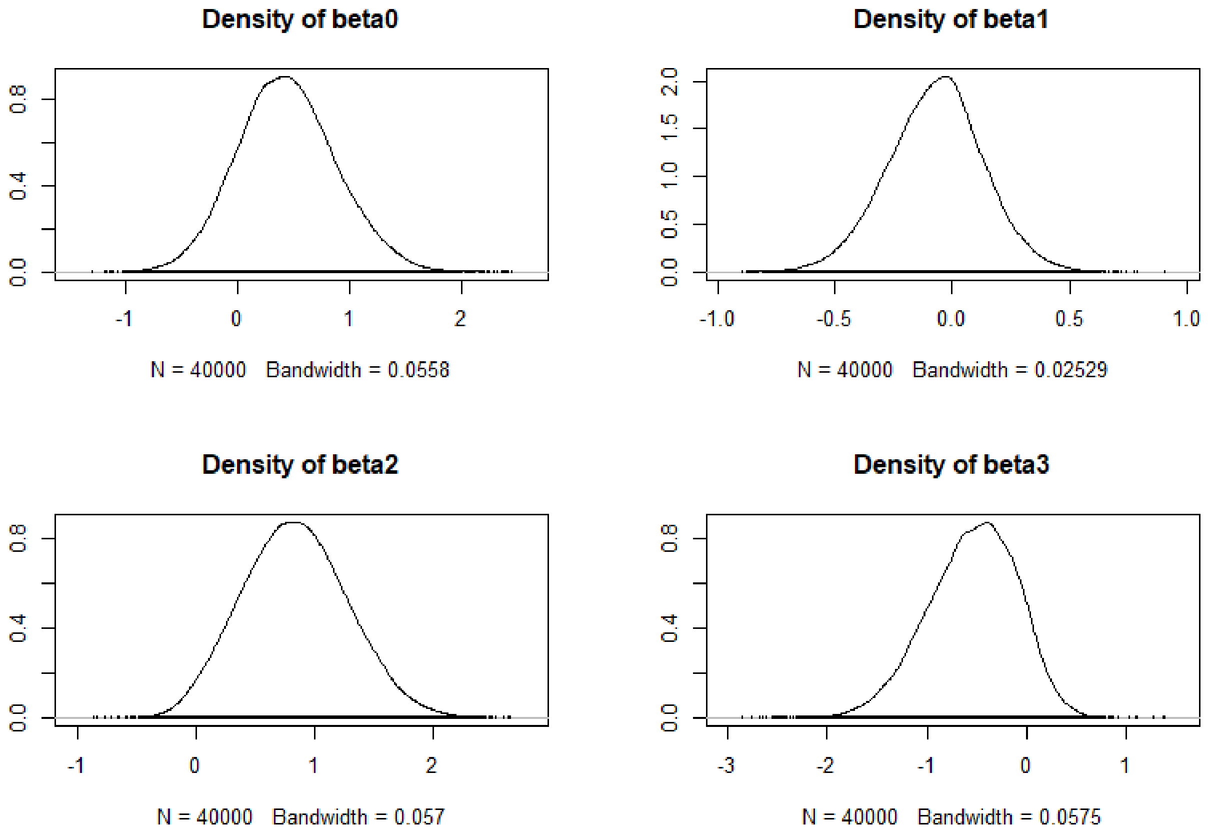

Figure 1: Posterior distributions of regression coefficients

Table 2: Mean, standard deviation and standard error for coefficient values based on posterior distributions of parameters

Coefficient	**Mean**	**SD**	**Std. Error**
beta0	0.45728	0.4431	0.002215
beta1	-0.07084	0.2062	0.001031
beta2	0.82944	0.4458	0.002229
beta3	-0.56526	0.4565	0.002283

Table 3: Quantiles for posterior distributions for all regression coefficients

coefficient	2.50%	25%	50%	75%	97.5%
beta0	-0.37496	0.1535	0.44576	0.74288	1.3654
beta1	-0.48814	-0.2031	-0.06582	0.06182	0.3302
beta2	0.00274	0.5178	0.82072	1.1248	1.7307
beta3	-1.51582	-0.8633	-0.53893	-0.23515	0.2334

Based on the results, holding *dist_to_shoreline* and *year* at fixed values, the odds of observing decoys for adults (age = 1) over the odds of observing decoys for immature eaglse (age = 0) was $exp(0.8294) = 2.29$. In terms of percent change, the odds for observing an adult was 129% greater than odds for observing immature eagle decoys. The coefficient for *Year* meant that, holding *dist_to_shoreline* and *age* at constant values, the odds of observing a decoy for 2004 (year = 1) over the odds of observing a decoy for 2003 (Year = 0) was $exp(-0.566) = 0.57$, which represents a 43% decrease in odds. Note that though there were more decoys overall in 2004, the success rate for observing a decoy in 2004 was only 53%, compared to 2003, when it was 74%. Finally, holding *year* and *age* constant, we could expect a 7% decrease in the odds of observing a decoy for a one-unit increase in *dist_to_shoreline* since $exp(-.07) = 0.93$.

In addition, the Highest Posterior Density Interval (HPD) can help in showing how far or close to 0 the means for coefficients are. The HPD is the shortest possible interval that contains the specified probability (usually 0.95 or 0.90). The 90% HPD intervals below clearly show that *age* was the only coefficient that age had a 0.90 probability of being greater than 0.

```
           lower     upper
beta0 -0.2707582 1.1824916
beta1 -0.4151352 0.2558572
beta2  0.1080028 1.5783104
beta3 -1.2691483 0.1779567
```

Finally, we can use the posterior distributions for each of the parameters to determine how certain we are of our parameter estimates. For instance, the probability that the coefficient for *age* is greater than 1 is 0.77, but the probability that is greater than 1.5 is 0.07. Similarly, the probability that the coefficient for *Year* is less than -0.3 is 0.69, but the probability that it is less than -0.9 is 0.21.

4 R code

```
#bald eagle for Pappa
eagles= read.csv("Maike_Logistic_Data.csv")
#drop obs that do not have values for substrate
eagles = eagles[complete.cases(eagles), ]

head(eagles)
library(plyr)
#change 'other' to 'ground'
sub = mapvalues(eagles$Substrate, from = "OTHER", to = "GROUND")
eagles$sub = sub
```

```r
#look at scatter matrix
library(gclus)
df <- eagles[, c(1,2,5,6)]# get data
str(df)
dfcorrs <- abs(cor(df)) # get correlations
dta.col <- dmat.color(dfcorrs) # get colors
# reorder variables so those with highest correlation
# are closest to the diagonal
#looks like dist shoreline and dist thalweg have a logistic assocation.
dforder <- order.single(dfcorrs)
cpairs(df, dforder, panel.colors=dta.col, gap = .5,
       main="Variables Ordered and Colored by Correlation" )

#convert age to numeric
eagles$Age_Class = as.numeric(eagles$Age_Class)
age = eagles$Age_Class
ifelse(age == 1, 1, ifelse(age == 2, 2, 3))
age=ifelse(age == 1, 1, 0)
eagles$age = age

#convert year to 0/1
Year = eagles$Year
year=ifelse(Year == 2003, 0, 1)
eagles$year01 = year

#standardize distance to shorline
eagles$dist.s = (eagles$Dist_Shoreline - mean(eagles$Dist_Shoreline)) / sd(eagles$Dist_S


#look at distribution of obs/not observed over years
#table looks OK...not so many 0's in year 2003
with(eagles, table(Year, Observed))
with(eagles, table(sub, Observed))
with(eagles, table(age, Observed))
with(eagles, table(age, Year))


eaglemod = glm(Observed ~   dist.s + age  + year01, family = binomial(link = 'logit'), d
summary(eaglemod)


#random forest just for fun...
library(randomForest)
head(eagles)
str(eagles)
eagles$factObs = as.factor(as.character(eagles$Observed))
str(eagles)
rfeagles = randomForest(factObs ~ Year + sub + Age_Class + log_Dist_Sh , importance = TR
varImpPlot(rfeagles)
```

```
imp = rfeagles$importance
plot(rfeagles)

impvar <- rownames(imp)[order(imp[, 1], decreasing=TRUE)]
op <- par(mfrow=c(2, 2))

for (i in seq_along(impvar)) {
  partialPlot(rfeagles, eagles, impvar[i], xlab=impvar[i],
              main=paste("Partial Dependence on", impvar[i]))
}

par(mfrow = c(1,1))
partialPlot(rfstream, train, "AirMwmt", xlim = c(35, 40), ylim= c(16, 17.5))

##############################################################################
#bayesian model
##############################################################################
newdf = as.data.frame(cbind(eagles$Observed, eagles$dist.s, eagles$age, eagles$year01))
names(newdf) = c("Observed",'DistShorlineS', 'age', 'year')
head(newdf)

#set variables for bayesian model
Observed = newdf$Observed
DistShorelineS = newdf$DistShorlineS
age = newdf$age
year = newdf$year

library(rjags)
modelstring2 = "
model {
#likelihood

for (i in 1:N) {

Observed[i] ~ dbern(mu[i])

logit(mu[i]) <- beta0 + beta1 * DistShorelineS[i] + beta2*age[i] + beta3*year[i]

}
#priors for beta0 and beta1
beta0 ~ dnorm(0, 1/25)
beta1 ~ ddexp(0, sqrt(2))
beta2 ~ ddexp(0, sqrt(2))
beta3 ~ ddexp(0, sqrt(2))
}"


#library(rjags)
writeLines(modelstring2, con='code.jags.txt')
```

```
library(rjags)
jagsLR <- jags.model('code.jags.txt',
                     data = list('Observed' = Observed,
                                 'DistShorelineS' = DistShorelineS,
                                 "age" = age,
                                 "year" = year,
                                 'N' = length(DistShorelineS)),
                     inits<-list(
                       list('beta0' = 0.1, 'beta1' = .1, 'beta2' = -0.2, 'beta3' = -0.1)
                       list('beta0' = 0.01, 'beta1' = 0.2, 'beta2' = 0.05, 'beta3' = 0.1
                       list('beta0' = 0.1, 'beta1' = 0.2, 'beta2' = 1, 'beta3' = -0.2),
                       list('beta0' = -0.2, 'beta1'= -0.1, 'beta2' = 0.1, 'beta3' = 0.2)
                     n.chains = 4,
                     n.adapt = 100)

update(jagsLR, 1000)

jags.samples(jagsLR,
             c('beta0', 'beta1'),
             1000)

codaSamplesLR = coda.samples(jagsLR, c('beta0','beta1', 'beta2' , 'beta3'), 10000, 1)
jagsLR_csim = as.mcmc(do.call(rbind, codaSamplesLR))

#diagnostic plots and posterior densities
plot(codaSamplesLR, density = TRUE)

#summaries

summary(codaSamplesLR)

#trace plots
traceplot(codaSamplesLR)

#gelman diags -ok
gelman.diag(codaSamplesLR)

#autocorrelation -ok
autocorr.plot(codaSamplesLR)
par(mfrow = c(2,2))
densplot(jagsLR_csim)
HPDinterval(jagsLR_csim, prob = 0.90)
head(jagsLR_csim)


#separate coeffients
beta1 = jagsLR_csim[,2]
beta2 = jagsLR_csim[,3]
beta3 = jagsLR_csim[,4]
```

```r
#look at posterior probabilities...more can be added--these are just some
head(jagsLR_csim)
mean(beta2 > .5)
mean(beta2 > 1)
mean(beta2 > 1.5)
mean(beta3 < -.3)
mean(beta3 < -.9)
#get coefficients (posterior means)
(pm_coef = colMeans(jagsLR_csim))
```