Mountain Pine Beetles in Colorado:

Logistic Regression Analysis

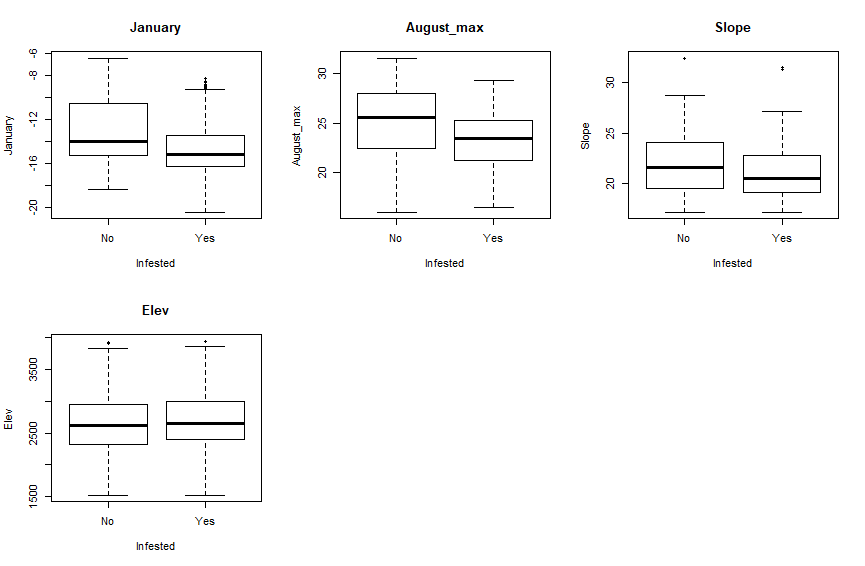
Carly Lundgreen

STAT 330-Brigham Young University

12/18/18

**Introduction**

In recent years, ecologists have been worried about the increasing amount of forest damage caused by the mountain pine beetle (MPB). Epidemics of these beetle infestations have led to a significant weakening of conifer forests. Ecologists are interested in determining specific environmental and climate-related factors that play a role in whether an area experiences MPB infestation. The goals of this study are to establish which of these factors have significant effects on pine beetle infestation and predict whether a region will experience infestation in the future.

An aerial detection survey conducted by the Colorado State Forest Service focusing on the North Central Rocky Mountains in Colorado. As surveyors flew over this region, they analyzed whether pine beetle damage could be seen, which would indicate that the location has experienced an MPB infestation. While also recording whether the location is infested, surveyors took note of various environmental factors for the area such as the average minimum January temperature, average maximum August temperature, annual precipitation rates, angle of mountain slope, which mountain region the location is part of, and the elevation at that location. A matrix of boxplots for each of the quantitative explanatory variables with respect to the variable ‘Infested’ (our response) is shown to the right. It seems that that, on average, the infestation rates decrease as the minimum temperature in January decreases, which aligns with our knowledge environmental stamina of MPBs—they usually die out at colder temperatures. There seems to be a few slopes measured that are outliers, and there is less variability of the slope of mountainsides where beetle infestations have occurred.

I will be using a logistic regression model because the response variable (whether the location is infested with MPBs) is categorical and thus follows a Bernoulli distribution. Using this model, values of the explanatory variables may be “plugged in” and predictions for probabilities of infestations can be calculated.

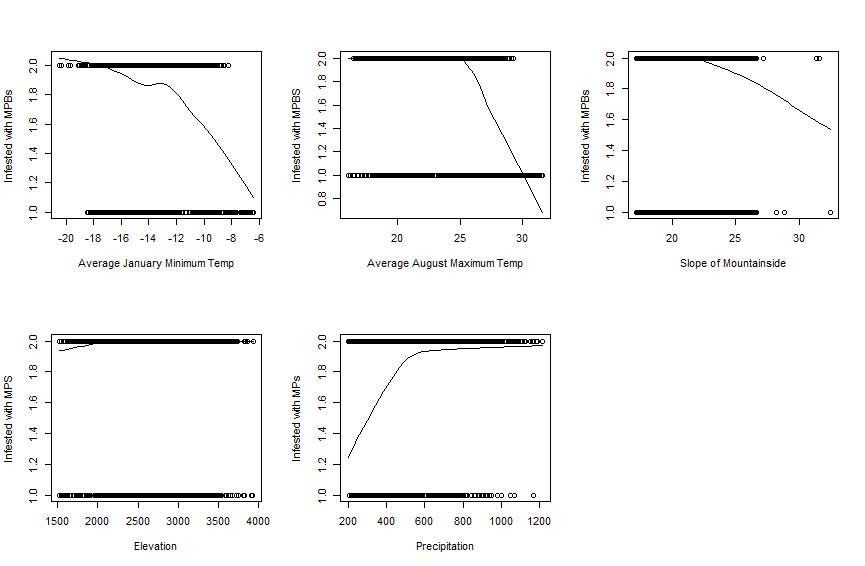
**Statistical Modeling**

We will use an exhaustive best subset variable selection method to fit this model. Our information criteria will be AIC because of the ecologists’ interest in predicting where future beetle infestations may occur. An exhaustive algorithm is appropriate because there are ony 15 variables in the dataset, so R can reasonably run through each possible combination. The following model will be utilized in this analysis:

*log(p/(1-p)) = + (January) +(August\_max) + (Slope) + (Elevation) + (Precipitation) +I(NC) +I(SE) + I(SW)*

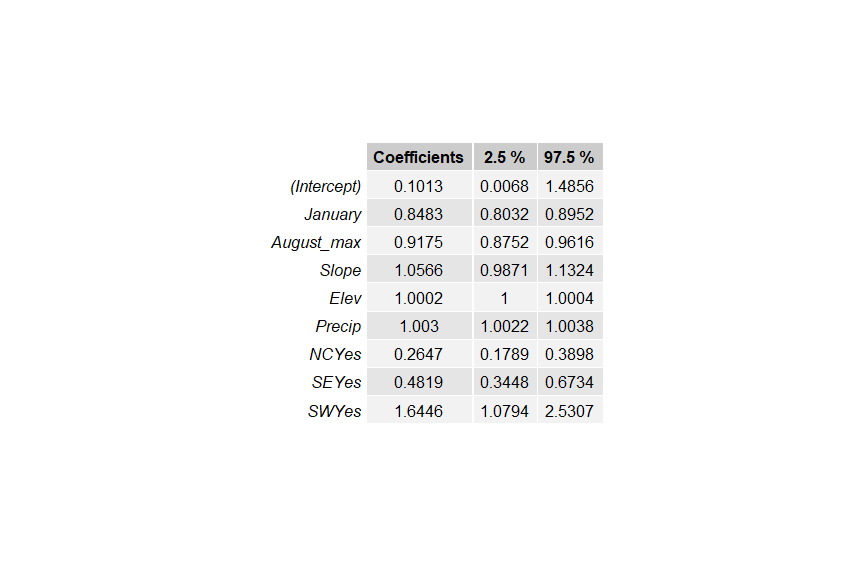
Holding all other variables constant, when the average minimum temperature in January increases by 1 degree Celsius, it is e^times more likely for that location in the Colorado mountain region to become infested by pine beetles. Similarly, a location in the North Central region is e^ times more likely to be infested by pine beetles than an area outside the North Central region—holding all other variables constant.

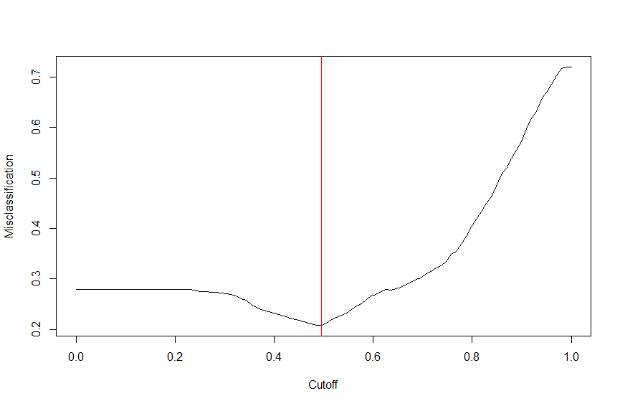
We assume that the log-odds ratio is linear when fitting this model, which means we assume that the responses are monotone in probability. To assess this condition, we plot each covariate against our response variable. The following scatterplots, with corresponding smoother curves, illustrate that the probabilities of areas becoming infested with pine beetles seem to be either strictly increasing or decreasing for all the following plots (for each quantitative explanatory variable).

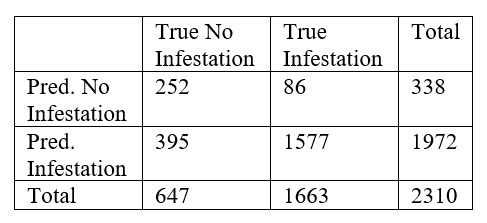


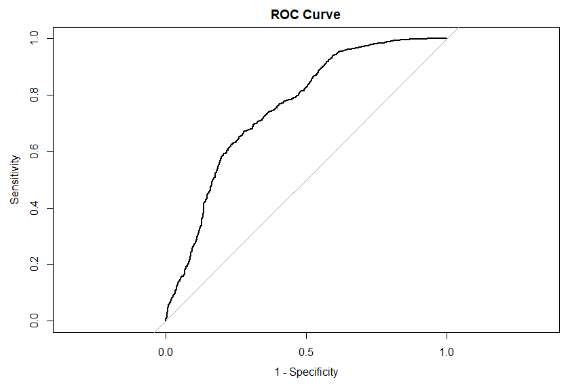
We can see from these scatterplots that the assumption of normality is met for this model. The second assumption made when fitting this model is that each of the responses is independent from each other. Because of the way the data were collected—by aerial survey—there may be instances of spatial dependence between areas. If one location has a high rate of beetle infestation, it is more probable that an area near that location also has a high rate of infestation, simply because of their proximity to each other. However, we will proceed with the analysis, keeping in mind this potential problem with the conditions of a logistic regression model.

**Results**

The table to the right displays estimates for each with corresponding 95% confidence intervals obtained after fitting the regression model. We are 95% confident that a Colorado mountain location is between 0.803 and 0.895 times more likely, on average, to become infested with MPBs for a 1oC increase in average minimum January temperature, holding all else constant.

We calculate a threshold of 0.495 that minimizes the misclassification rate for areas in the Colorado mountain region. The curve below shows various threshold probabilities and the corresponding misclassification rates for our model. It is clear that 0.495 is the threshold that leads to the lowest rate of MPB misclassification. The confusion matrix calculated across all the datapoints is also displayed below.



The sensitivity rating for the model across all the data is found to be 1577/1663=0.948. The specificity is found to be 252/647=0.389. The percent of positive predicted values is found to be 1577/1972=0.80, and the percent of negative predicted values is found to 252/338=0.476. It seems that the model fits well across all of the data, as the percent of true positives and percent of positive predicted values are relatively high. The following ROC curve indicates how well the model predicts across all cutoffs, plotting 1-specificity (how many “mistakes” the model is making) against sensitivity (how well the model classifies MPB-infested areas). The area under the curve is found to be 0.756, so the model seems to predict well across all cutoff values.

The average sensitivity rating across 250 cross-validation studies is found to be 0.949, while the average specificity is found to be 0.386. Further, the average PPV rating is found to be 0.799 and the average NPV rating is found to be 0.746. These values are similar to the values observed across all the data, and the percent of true positives and positive predicted values are high enough that we can say the model is predicting probabilities well. The percent of negative predicted values is also relatively high, indicating the model accurately predicts that an area will not become infested with MPBs.

The average predicted probability that this sight will become infested within the next ten years is 0.817. The forecasts for each year were run through the logistic regression model, and the average probability over the next ten years was calculated. Because this probability is above the cutoff value of 0.495, we can conclude that the area will become infested in the next ten years.

**Conclusion**

Using statistical modeling, we have concluded that an area in the SE region with a slope of 18.07 and an elevation of 1901.95 has, on average, a 0.817 probability of becoming infested with MPBs within the next 10 years. The model seems to fit the data well, and as a high level of predictive accuracy, so this prediction is reliable. Based on a selection algorithm that selected the optimum variables to include in the analysis, the model took into account the average minimum January temperature, average maximum August temperature, the angle of the mountain slope, elevation (in feet), average annual precipitation (in inches), and the NW, SE, and SW regions.

The ecologists should also consider other variables that may affect pine beetle infestations, such as the temperatures of other months of the year. If they include these other variables, that will make predictions of probabilities of infestation more accurate. They may also want to consider a larger area than the smaller subset of the Colorado mountain region which they gathered data from. There is also a possible dependence among infestations in neighboring locations, so they should look into solving that potential violation of one of our model assumptions.

R Code: Appendix

beetles <- read.csv(file="C:/Users/Carly/Documents/Semester 6/Stat 330/PineBeetle2.csv",sep=",",header=TRUE)  
###Explore data####  
par(mfrow=c(2,3))  
myPlot<-function(index) {plot(beetles[,index] ~ beetles$Infested, main=names(beetles[index]),pch=16,xlab="Infested",ylab=names(beetles[index]))}  
lapply(1:4,FUN=myPlot)

## [[1]]  
## NULL  
##   
## [[2]]  
## NULL  
##   
## [[3]]  
## NULL  
##   
## [[4]]  
## NULL

#create a scatterplot in order to check conditions  
par(mfrow=c(2,3))

scatter.smooth(beetles$January,beetles$Infested,xlab="Average January Minimum Temp",ylab="Infested with MPBs")  
scatter.smooth(beetles$August\_max,beetles$Infested,xlab="Average August Maximum Temp",ylab="Infested with MPBS")  
scatter.smooth(beetles$Slope,beetles$Infested,xlab="Slope of Mountainside",ylab="Infested with MPBs")  
scatter.smooth(beetles$Elev,beetles$Infested,xlab="Elevation",ylab="Infested with MPS")  
scatter.smooth(beetles$Precip,beetles$Infested,xlab="Precipitation",ylab="Infested with MPs")  
##1 means NO, 2 means YES, the area is infested  
  
#best GLM   
vs.res <- bestglm(beetles,IC="AIC",method="exhaustive",family=binomial)

## Morgan-Tatar search since family is non-gaussian.

par(mfrow=c(1,1))

plot(vs.res$Subsets$AIC,type="b",pch=19,xlab="# of Vars",ylab="AIC")

beetle.mod <- vs.res$BestModel  
  
##assess model and get confidence intervals (transformed)  
summary(beetle.mod)

##   
## Call:  
## glm(formula = y ~ ., family = family, data = Xi, weights = weights)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.7503 -0.9311 0.5133 0.7860 1.4743   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -2.2897076 1.3740592 -1.666 0.095637 .   
## January -0.1645551 0.0276759 -5.946 2.75e-09 \*\*\*  
## August\_max -0.0861046 0.0240017 -3.587 0.000334 \*\*\*  
## Slope 0.0550539 0.0350013 1.573 0.115739   
## Elev 0.0002287 0.0001123 2.037 0.041645 \*   
## Precip 0.0029940 0.0004180 7.162 7.94e-13 \*\*\*  
## NCYes -1.3290251 0.1985342 -6.694 2.17e-11 \*\*\*  
## SEYes -0.7301014 0.1706632 -4.278 1.89e-05 \*\*\*  
## SWYes 0.4974778 0.2171866 2.291 0.021989 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 2739.8 on 2309 degrees of freedom  
## Residual deviance: 2344.5 on 2301 degrees of freedom  
## AIC: 2362.5  
##   
## Number of Fisher Scoring iterations: 5

confint(beetle.mod)

## Waiting for profiling to be done...

## 2.5 % 97.5 %  
## (Intercept) -4.993678e+00 0.3957892948  
## January -2.192101e-01 -0.1106598404  
## August\_max -1.332574e-01 -0.0391276391  
## Slope -1.299132e-02 0.1243144334  
## Elev 9.136559e-06 0.0004494225  
## Precip 2.184840e-03 0.0038241824  
## NCYes -1.720940e+00 -0.9421374810  
## SEYes -1.064886e+00 -0.3954468665  
## SWYes 7.637815e-02 0.9285117849

exp(confint(beetle.mod))

## Waiting for profiling to be done...

## 2.5 % 97.5 %  
## (Intercept) 0.006780678 1.4855563  
## January 0.803152969 0.8952432  
## August\_max 0.875239814 0.9616280  
## Slope 0.987092707 1.1323719  
## Elev 1.000009137 1.0004495  
## Precip 1.002187229 1.0038315  
## NCYes 0.178897884 0.3897938  
## SEYes 0.344767167 0.6733791  
## SWYes 1.079370664 2.5307401

#make a nice table  
Coefficients <- exp(beetle.mod$coefficients)  
coefs <- data.frame(Coefficients)  
coefs2 <- round(coefs, 4)  
int <- exp(confint(beetle.mod))

## Waiting for profiling to be done...

int <- round(int,4)  
estimates <- tableGrob(coefs2)  
intervals <- tableGrob(int,rows=NULL)  
grid.newpage()  
grid.arrange(gtable\_combine(estimates,intervals),ncol=1)

###find threshold and verify it's the minimum   
pred.probs <- predict.glm(beetle.mod,type="response")  
thresh <- seq(0,1,length=100)  
y <- beetles$Infested  
misclass <- rep(NA,length=length(thresh))   
for(i in 1:length(thresh)) {  
 my.classification <- ifelse(pred.probs>thresh[i],'Yes','No')  
 misclass[i] <- mean(my.classification!=y)  
}  
  
threshold <- thresh[which.min(misclass)]  
threshold

## [1] 0.4949495

plot(thresh, misclass, pch=20,type="l",xlab="Cutoff",ylab="Misclassification")  
abline(v=threshold,col="red")

##confusion matrix for ALL the data  
true.class <- beetles$Infested   
pred.class <- ifelse(pred.probs>threshold, 'Yes', 'No')  
table(true.class,pred.class)

## pred.class  
## true.class No Yes  
## No 252 395  
## Yes 86 1577

addmargins(table(pred.class,true.class))

## true.class  
## pred.class No Yes Sum  
## No 252 86 338  
## Yes 395 1577 1972  
## Sum 647 1663 2310

##create ROC curve  
pred.probs <- predict.glm(beetle.mod,type="response")   
obs.y <- beetles$Infested  
a.roc <- roc(obs.y,pred.probs)  
plot(a.roc,legacy.axes=TRUE,main="ROC Curve")

auc(a.roc) ##area under the ROC curve

## Area under the curve: 0.7555

my.roc <- roc(obs.y,pred.probs)  
##cross-validation study for sensitivity, specificity, etc.  
## Choose number of CV studies to run in a loop & test set size  
n.cv <- 500  
n.test <- round(.1\*nrow(beetles))  
  
## Set my threshold for classifying  
cutoff <- threshold  
## Initialize matrices to hold CV results  
sens <- rep(NA,n.cv)  
spec <- rep(NA,n.cv)  
ppv <- rep(NA,n.cv)  
npv <- rep(NA,n.cv)  
auc <- rep(NA,n.cv)  
  
## Begin for loop  
for(cv in 1:n.cv){  
 ## Separate into test and training sets  
 test.obs <- sample(1:nrow(beetles), n.test)  
 test.set <- beetles[test.obs,]  
 train.set <- beetles[-test.obs,]  
 ## Fit best model to training set  
 train.model <- beetle.mod  
 ## Use fitted model to predict test set  
 pred.probs <- predict.glm(train.model,newdata=test.set,type="response") #response gives probabilities  
 ## Classify according to threshold  
 test.class <- ifelse(pred.probs>cutoff,"Yes","No")  
 ## Create a confusion matrix  
 conf.mat <- addmargins(table(factor(test.set$Infested,levels=c("No","Yes")),  
 factor(test.class,levels=c("No","Yes"))))  
 ## Pull of sensitivity, specificity, PPV and NPV  
 ## using bracket notation  
 sens[cv] <- conf.mat[2,2]/conf.mat[2,3]  
 spec[cv] <- conf.mat[1,1]/conf.mat[1,3]  
 ppv[cv] <- conf.mat[2,2]/conf.mat[3,2]  
 npv[cv] <- conf.mat[1,1]/conf.mat[3,1]  
 ## Calculate AUC  
 auc[cv] <- auc(roc(test.set$Infested,pred.probs))  
} #End for-loop  
  
mean(sens)

## [1] 0.9482496

mean(spec)

## [1] 0.3899236

mean(npv)

## [1] 0.7471045

mean(ppv)

## [1] 0.7989635

mean(auc)

## [1] 0.7557357

#run through each row of the data frame given, with the slope and elevations as given for the specific area, to find the probability for each   
#of the next 10 years  
#I would've put this in a for loop, but I ran out of time  
year1 <- data.frame(January=-13.98,August\_max=15.89,Precip=771.13,Slope=18.07,Elev=1901.95,NC="No",SE="Yes",SW="No")  
year2 <- data.frame(January=-17.80,August\_max=18.07,Precip=778.54,Slope=18.07,Elev=1901.95,NC="No",SE="Yes",SW="No")  
year3 <- data.frame(January=-17.27,August\_max=16.74,Precip=677.63,Slope=18.07,Elev=1901.95,NC="No",SE="Yes",SW="No")  
year4 <- data.frame(January=-12.52,August\_max=18.06,Precip=522.77,Slope=18.07,Elev=1901.95,NC="No",SE="Yes",SW="No")  
year5 <- data.frame(January=-15.99,August\_max=18.23,Precip=732.32,Slope=18.07,Elev=1901.95,NC="No",SE="Yes",SW="No")  
year6 <- data.frame(January=-11.97,August\_max=15.81,Precip=615.96,Slope=18.07,Elev=1901.95,NC="No",SE="Yes",SW="No")  
year7 <- data.frame(January=-15.75,August\_max=16.85,Precip=805.90,Slope=18.07,Elev=1901.95,NC="No",SE="Yes",SW="No")  
year8 <- data.frame(January=-16.19,August\_max=16.51,Precip=714.57,Slope=18.07,Elev=1901.95,NC="No",SE="Yes",SW="No")  
year9 <- data.frame(January=-17.87,August\_max=17.84,Precip=740.50,Slope=18.07,Elev=1901.95,NC="No",SE="Yes",SW="No")  
year10 <- data.frame(January=-12.44,August\_max=16.96,Precip=801.22,Slope=18.07,Elev=1901.95,NC="No",SE="Yes",SW="No")  
year <- data.frame(cbind(year1,year2,year3,year4,year5,year6,year7,year8,year9,year10))  
  
get.probs<-function(index) {predict.glm(beetle.mod,newdata=year[index,],type="response")}  
  
#find the average probability of MPB infestation over the next 10 years for the specified location  
prob.vec <- as.numeric(lapply(1:10,FUN=get.probs))  
mean(prob.vec)

## [1] NA