Final Exam: Pine Beetles: Kevin Toney

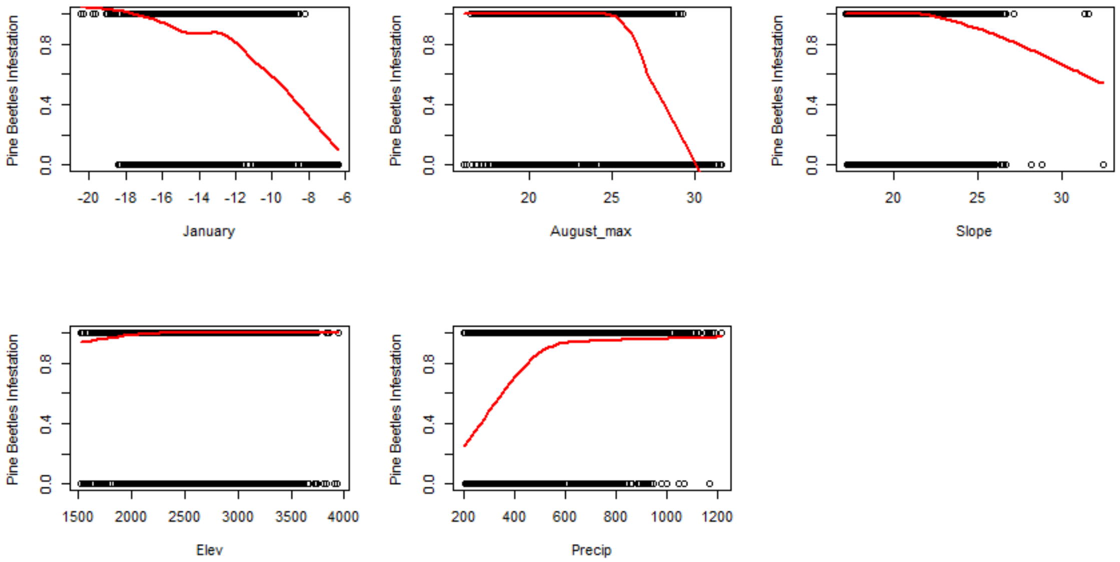
December 16, 2017

1 Research Problem and Background

For many years, pine beetles (MPB) have destroyed old and weak trees by eating nutrients underneath the bark. Once winter begins, the beetles usually die because of the harsh conditions. Recent years have had warmer summers and drier conditions. Therefore, more beetles survived, thus causing wide spread tree mortality in conifer forests all around the North Central Rocky Mountains in Colorado. The overarching purpose of this analysis is to infer which covariate variables, if any, significantly affect the infestation of pine beetles in this portion of the mountain range. Also, I will predict the probability of future occurances of pine beetle infestions at a location in the mountain range.

To fulfill these goals, I used a dataset of 2,310 locations, which were at risk for pine beetle infestion. The response variable is the presence of pine beetles. The outcome of this variable is ”Yes” or ”No”. The statistical term for this kind of response variable is categorical, or binary. The covariate variables are the average January minimum temperature in degrees celcius, the average August maxiumum temperature, the angle of mountain slope, the elevation in feet, the mean annual percipitation in inches, and region indicators (NC, NW, EC, etc...).

There are scatter plots below that show the relationship the infestion of beetles has with the quanititative variables. In other words, the region indicators are not shown in the scatter plots.



The quantitative variables seem to have an approximately linear relationship with the presence of a pine beetle infestation. We can see this relationship by noting that the scatter plot line shows a constant trend for each variable. Either the line trends upward throughout all the data or downward. Therefore, some sort of regression is appropriate for this analysis.

The type of regression that is suitable for this analysis is logistic regression. This method is optimal for this analysis because it does the best job analyzing a binary response variable. Moreover, the assumptions

I needed to make in order to use this type of regression are all justified. The assumptions for traditional multiple linear regression or Poisson regression are not. Logistic regression uses the log-odds ratio, $log(p_i/(1-p_i))$, to create a mathematical linear equation with coefficients I can test for significance. Also, I can plug in values for each data variable, use the coefficients, and transform the result to receive a predicted probability of an MPB infestation.

2 Statistical Modeling

I used best subset selection and the AIC criterion to decide which statistical model would be best. I decided the statistical program should test the model with all combinations of variables, if possible. Testing the model with all combinations of variables is widely known as the most reliable method for picking the best model. This method exemplifies best model selection. The computer performed this selection method quickly. Therefore, best subset selection was an appropriate model selection method to use.

I decided to use the information criterion AIC, because the criterion only excludes variables that do not make significant contributions to the strength of the best model. Most of the time, selection using the AIC criterion tends to have more variables than other criteria. If some of the covariates are excluded from the best model, we can be rest assured they would not be helpful for this analysis. Therefore, the angle of mountain slope, and the region indicators NW, EC, WC, and SC are not useful in describing the presence of pine beetle infestations. Average January minimum temperature (in degrees C), average August maximum temperature, elevation, mean annual precipitation and region indicators NC, SE and SW are important in explaining the presence of these infestations.

A justifiable logistic regression model, and the best statistical model, is below:

$$log(p_i/(1-p_i)) = \beta_0 + \beta_1 * (January_i) + \beta_2 * (AugustMax_i) + \beta_3 * (Elevation_i) + \beta_4 * (Precip_i) + \beta_5 * (NC_i) + \beta_6 * (SE_i) + \beta_7 * (SW_i)$$

The interpretation of a few of the mathematical terms is below:

p_i = the probability of a location having a pine beetle infestation.

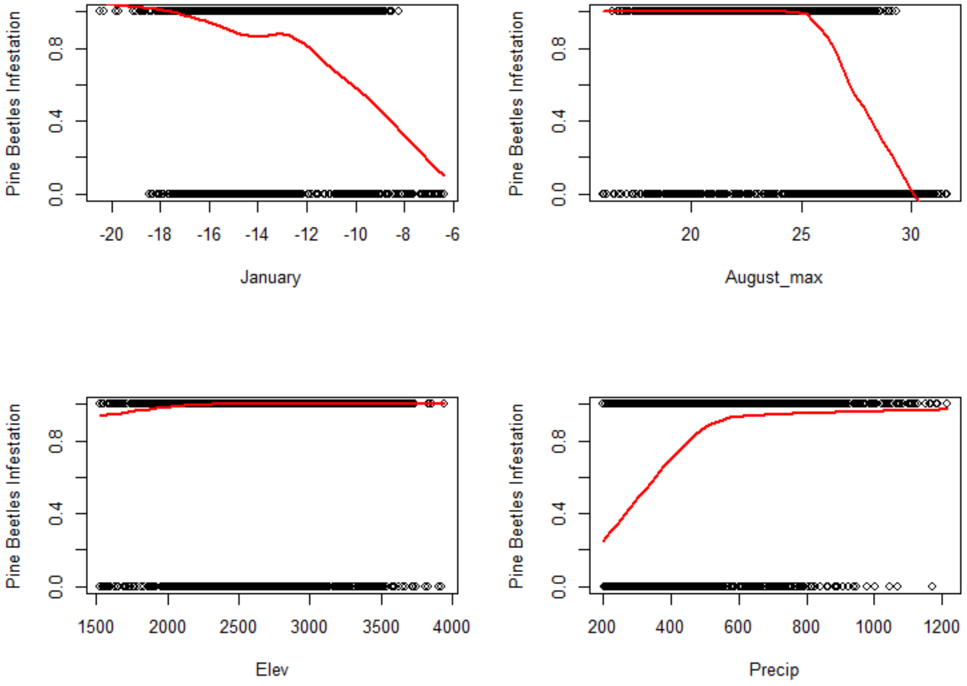
β_0 = the probability of a location having a pine beetle infestation, if the minimum average January temperature is 0, the average August maximum temperature is 0, the elevation is 0, the mean annual precipitation is 0, and the location isn't in the NC, SE, or SW region, is $e^{\beta_0}/(1+e^{\beta_0})$.

β_2 = holding all other covariate variables constant, the probability of a location having a pine beetle infestation, if the average August maximum temperature increases by one, will be e^{β_2} times as much as before.

β_3 = holding all other covariate variables constant, the probability of a location having a pine beetle infestation, if the location's elevation increases by one foot, will be e^{β_3} times as much as before.

β_7 = holding all other covariate variables constant, the probability of a location having a pine beetle infestation, if the location is in the south west region, will be e^{β_7} times as much as regions not in the southwest.

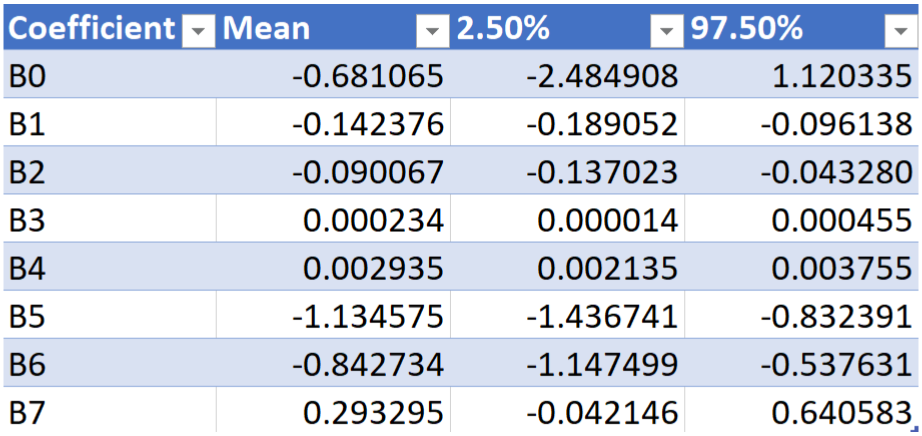
In order to use logistic regression, I assumed the response variable measurements, which are the presence of pine beetles, are independent of each other. This assumption is justified enough for our analysis. One location having a pine beetle infestation may not affect another location as much as the environmental and seasonal factors. Another assumption I made was the relationship between the presence of pine beetles and each of the covarites is linear in the log-odds scale. The assumption is justified by the following scatter plots:



Each plot shows an upward or downward trend throughout the data. Since the trends appear in only one direction, the assumption is justified. The relationships between the presence of pine beetles and each of the covariates in our model is indeed linear in the log odds scale. As a reminder, the log-odds is the following: $log(p_i/(1-p_i))$.

3 Results

I fit the logistic regression model to the pine beetles dataset. The model gave me estimated average coefficients, which are the covariate effects, and their 95% confidence intervals, which are found in the table below:



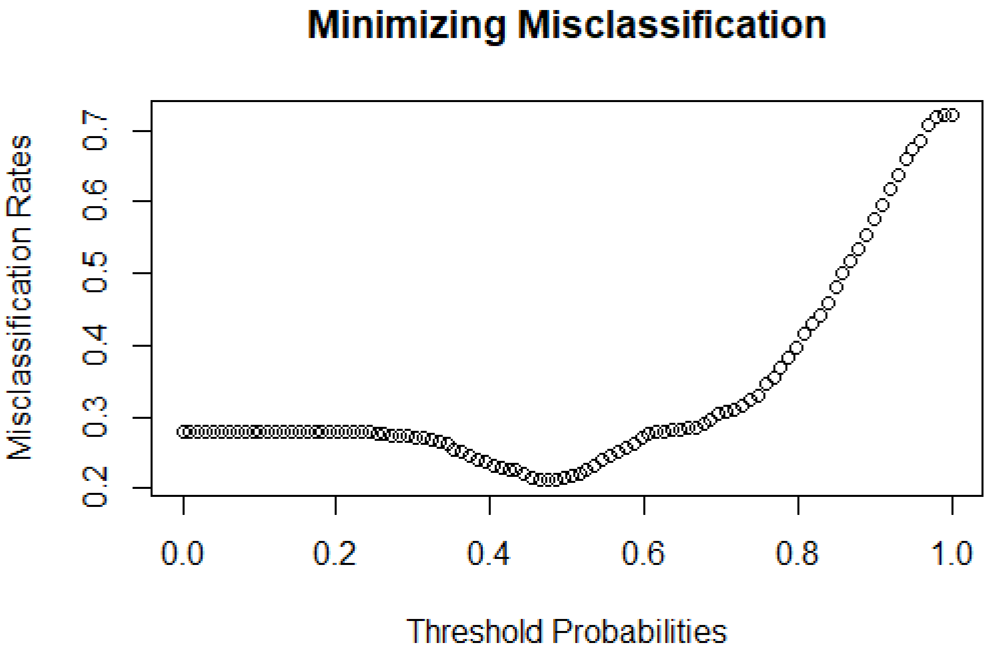
Coefficient	Mean	2.50%	97.50%
B0	-0.681065	-2.484908	1.120335
B1	-0.142376	-0.189052	-0.096138
B2	-0.090067	-0.137023	-0.043280
B3	0.000234	0.000014	0.000455
B4	0.002935	0.002135	0.003755
B5	-1.134575	-1.436741	-0.832391
B6	-0.842734	-1.147499	-0.537631
B7	0.293295	-0.042146	0.640583

The table shows the estimated coefficients, and a 95% confidence interval for each parameter. I am 95% confident the true coefficients lie in their corresponding confidence interval. The coefficients, which have intervals that don't include zero, significantly affect where an infestation of pine beetles occurs. The effects that qualify as significant are minimum January temperatures, August max temperature, elevation, precipitation, NC, and SE.

Here are some examples of interpreting these intervals. Regarding β_2, I am 95% confident, holding all other covariate variables constant, the probability of a location having a pine beetle infestation, if the average August maximum temperature increases by one, will be between $e^{-0.137}$ and $e^{-0.043}$ times as much as before.

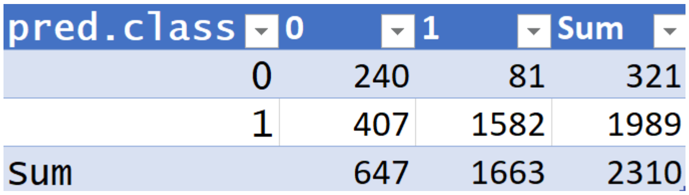
Another example of how to interpret these intervals is regarding β_6. I am 95% confident, holding all other covariate variables constant, the probability of a location having an MPB infestation, if the location is in the south east region, will be between $e^{-1.147}$ and $e^{-0.538}$ times as much as other locations not in the region.

Next, I ran predictions on the presence of pine beetle infestations in a location, given certain information. In order to minimize the probability of misclassifying the location as having the beetles, I made predictions for each data point and compared them to 100 different threshold probabilities. Then, I found the misclassification rate, meaning the chance of making an incorrect prediction, at each proabability. The threshold probabilities and the misclassification rates are plotted below:



I chose a threshold probability of 0.4848485. After testing other probabilities, this cutoff had the lowest misclassification rate. In other words, using this cutoff probability helps my predictions have the highest accuracy.

Using the classification threshold I found in the last part, I created a confusion matrix, which shows the correct predictions for the presence of MPB and the absence of it. The table is below:



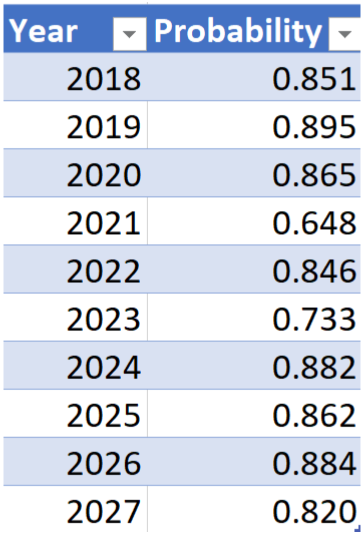
pred.class	0	1	Sum
0	240	81	321
1	407	1582	1989
Sum	647	1663	2310

According to this table, the sensitivity of these predictions is 0.7954. Therefore, the percentage of a location having an infestation when I predicted it would (true positives) was 79.54%. The specificity, or the percentage of true negatives, is 74.77%. The positive predictive value is 0.9513, meaning I correctly predicted the presence of pine beetle infestations 95.13% of the time. The negative predictive value is 0.3710. I correctly predicted the absence of pine beetles 37.1% of the time.

The pseudo-R^2 for my logistic regression model is 0.1434. Therefore, my model explains 14.34 percent of the variation found in the log odds of the presence of pine beetles.

I ran 500 cross validation studies where I classified, or predicted, the presence of pine beetles using the threshold probability of 0.4848485. The predictive ability of my logistic regression model will be best described by stating the average sensitivity, specificity, positive predictive value and negative predictive value. My average sensitivity was 0.9473. The average specificity was 0.3630. The average positive predictive value was 0.7941. The average negative predictive value was 0.7290.

Given the forecasts you gathered for the next 10 years of the location that is in the southeast region and has an elevation of 1,901.95 feet, the predicted probabilities of a pine beetle infestation occuring there are in the following table:



Year	Probability
2018	0.851
2019	0.895
2020	0.865
2021	0.648
2022	0.846
2023	0.733
2024	0.882
2025	0.862
2026	0.884
2027	0.820

I think the region will become infested in the next 10 years. I conclude this because the predicted probabilities are high; most of the probabilities are above 80%. Despite the probabilities being high, I suggest the forest service concetrate their efforts in preventing the pine beetles from infesting the trees.

4 Conclusion

I analyzed the pine beetle dataset to infer what environmental factors significantly affect the probability of pine beetle infestations and predict where pine beetle infestations will happen in the North Central Rocky Mountains in Colorado during the next 10 years. I fulfilled this purpose by creating a statistical model, which gives a predicted probability of infestation as the output of a mathematical equation, given some information and coefficients. Using this model, I concluded average minimum January temperatures, average maximum August temperatures, elevation, precipitation and region indicators north central, SE and SW regions significantly impact the presence of a pine beetle infestation. The model fit the data fairly well because the sensitivity, or the percentage of predicting an area will have pine beetles when it actually did was 0.7954. The specificity of my predictions, or the percentage of predicting the absence of pine beetles when there was an absence, was 0.7477. The percentage of correctly predicting pine beetle infestations (positive predictive value) was 0.9513 and the probability of correctly predicting no infestation (negative predictive value) was 0.3710. I made predictions 500 more times and found the average sensitivity, specificity, positive predictive value and negative predictive value. Given the forecasts for the next 10 years of a location in the southeast region, with an elevation of 1,901.95, the probability of an infestation is high. I expect this location to have an infestation during that time.

I encourage the forest service to focus on preventing an infestation there. Moreover, I recommend the forest service keep track of the data of the first snow and record the date of the last snow. These two factors may affect the pine beetle population in Colorado.

5 Appendix

```
rm(list=ls())

########################
#Final Exam: Pine Beetles
########################
library(MASS)
library(lmtest)
library(car)
library(bestglm)


beetlesdat <- read.table(”C:/Users/kevin/Desktop/Fall 2017/STAT 330/
Final Exam/PineBeetle2.csv”, header = T, sep = ”,”)

beetlesdat$NC <- ifelse(beetlesdat$NC == ”Yes”, 1, 0)
beetlesdat$NW <- ifelse(beetlesdat$NW == ”Yes”, 1, 0)
beetlesdat$EC <- ifelse(beetlesdat$EC == ”Yes”, 1, 0)
beetlesdat$WC <- ifelse(beetlesdat$WC == ”Yes”, 1, 0)
beetlesdat$SE <- ifelse(beetlesdat$SE == ”Yes”, 1, 0)
beetlesdat$SC <- ifelse(beetlesdat$SC == ”Yes”, 1, 0)
beetlesdat$SW <- ifelse(beetlesdat$SW == ”Yes”, 1, 0)
beetlesdat$Infested <- ifelse(beetlesdat$Infested == ”Yes”, 1, 0)


windows()
par(mfrow=c(2,3))
for(i in 1:5) {
  plot(beetlesdat[,i], beetlesdat$Infested,
                 xlab = colnames(beetlesdat)[i],
                                ylab=”Pine Beetles Infestation”)

  lines(loess.smooth(beetlesdat[,i], beetlesdat$Infested), col=”red”, lwd=2)
}

#I decided to use best subset selection
#and the criterion AIC to decide which statistical model would be best.

best <- bestglm(beetlesdat, method = ”exhaustive”, IC=”AIC”)
best$BestModel
best.glm <- glm(best$BestModel, family = ”binomial”)


windows()
par(mfrow=c(2,2))
for(i in c(1,2,4,5)) {
  plot(beetlesdat[,i], beetlesdat$Infested,
                 xlab = colnames(beetlesdat)[i],
                                ylab=”Pine Beetles Infestation”)
  lines(loess.smooth(beetlesdat[,i], beetlesdat$Infested), col="red", lwd=2)
}


coefficients <- cbind(best.glm$coefficients, confint(best.glm))

write.csv(coefficients, file="C:/Users/kevin/Desktop/Fall 2017/
STAT 330/Final Exam/coefficients.csv")

summary(best.glm)


################### Really cool
#predicted probability for every point in my dataset
pred.probs <- predict.glm(best.glm, type="response")
#potential thresholds
thresh <- seq(0, 1, length=100)
misclass <- rep(NA, length=length(thresh))


#store misclassification in misclass
for(i in 1:length(thresh)) {
  #if probability is greater than threshold then 1, else 0
  my.classification <- ifelse(pred.probs > thresh[i],1,0)


  #calculate the pct where my classification doesn't equal the thruth
  misclass[i] <- mean(my.classification != beetlesdat$Infested)
}

#find the threshold, which minimizes misclassification
thresh[which.min(misclass)]
cutoff <- thresh[which.min(misclass)]

#plot the misclassification
plot(thresh, misclass, xlab = "Threshold Probabilities",
     ylab="Misclassification Rates",
     main="Minimizing Misclassification")
#################


#build a confusion matrix
pred.class <- ifelse(pred.probs > cutoff,1,0)
addmargins(table(pred.class, beetlesdat$Infested))

#find the pseudo R^2
best.glm$deviance
best.glm$null.deviance

pseudoR2 <- 1 - (best.glm$deviance/best.glm$null.deviance)
pseudoR2


#cross validation studies
```

```
## Choose number of CV studies to run in a loop & test set size
n.cv <- 500
n.test <- round(.1*nrow(beetlesdat))
## Set my threshold for classifying
cutoff <- 0.4848485
## 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)

## Begin for loop
for(cv in 1:n.cv){
  train <- sample(1:nrow(beetlesdat), 0.8*nrow(beetlesdat))
  train.set <- beetlesdat[train,]
  test.set <- beetlesdat[-train,]
 ## Separate into test and training sets
 ## Fit best model to training set
  train.model <- glm(Infested ~ January + August_max + Elev + Precip +
 NC + SE + SW,data=train.set, family = ”binomial”)
 ## 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,1,0)
 ## Create a confusion matrix
  conf.mat <- addmargins(table(test.set$Infested,test.class))
  factor(test.class, levels=c(0,1))
 ## 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]
} #End for-loop


mean(sens)
mean(spec)
mean(ppv)
mean(npv)


newdf <- read.csv(”C:/Users/kevin/Desktop/Fall 2017/STAT 330/Final Exam
/newdf.csv”,
                  header = T, sep=”,”)

newdf <- newdf[,-1]
newdf$Infested <- 0
pred.log.odds <- predict.glm(best.glm, newdata=newdf)
pred.prob <- exp(pred.log.odds)/(1+exp(pred.log.odds))

names(pred.prob) <- c(”2018”, ”2019”, ”2020”, ”2021”, ”2022”, ”2023”,
”2024”, ”2025”, ”2026”, ”2027”)
```

```
pred.prob
```