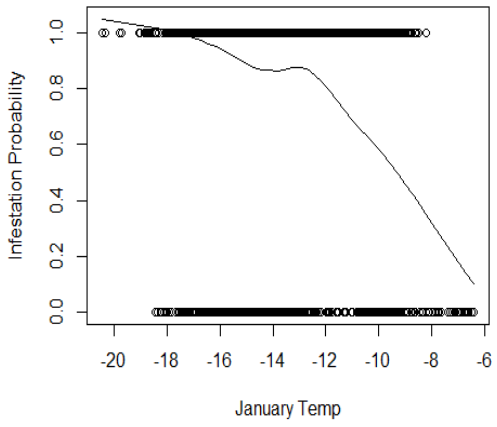
Jake Szendre

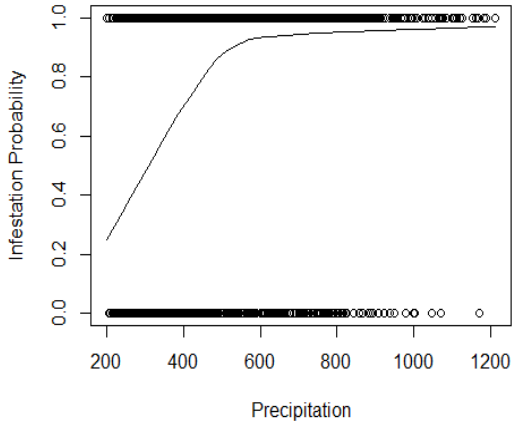
PINE BEETLE DAMAGE REPORT

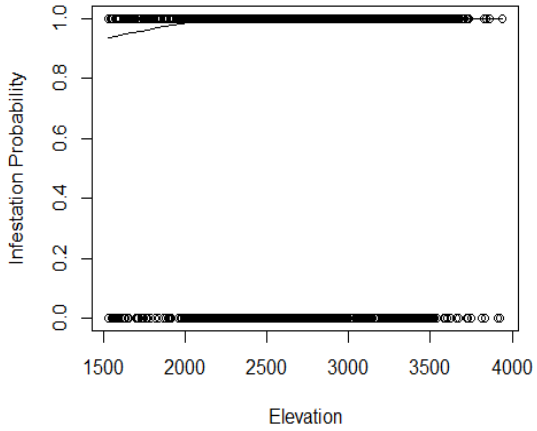
Introduction and Problem Background

Pine beetles have recently been a major problem for forests in the western United States. Pine beetles feed on trees, and in normal temperature they are good for the ecosystem because they typically take care of dead trees. But with higher temperatures in recent years, pine beetles have been able to survive for longer, and are eating live trees as well. To stop them, we first need to know where they likely will be. We wish to find places of infestation of pine beetles and, most importantly, predict probabilities of future places the pine beetles will infest next, so we can be a step ahead and control them.

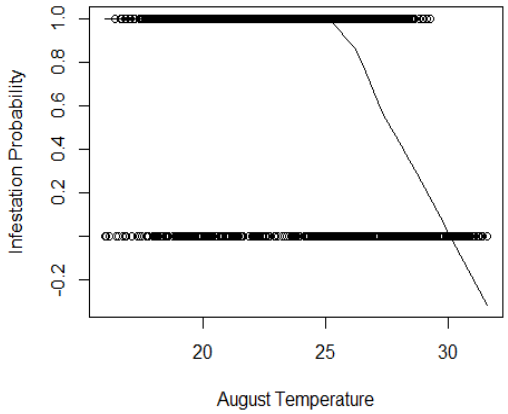
To address this issue, we are going to analyze a data set that has values of potential factors of different locations in the western United States that do and do not have pine beetle infestations. The factors that have been chosen include the following: minimum January temperature, maximum August temperature, slope of the mountain, elevation, precipitation, and what region the location is in. We are going to determine which of these factors contribute to higher chances of pine beetle infestations. On the next page are graphs showing each of the numerical factors and the likelihood of an infestation as the factors increase.

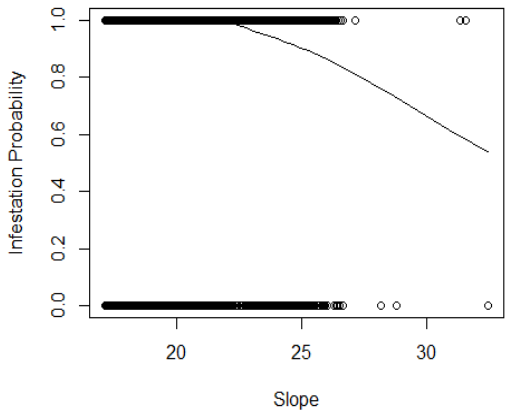






Based on these graphs, it appears that as elevation and precipitation increase, the probability of an infestation also increases, on average. As August/ January temperature and slope of the mountains increase, the probability of an infestation decreases, on average.





We are going to use logistical regression for this data set, because the response variable (what we want to predict) is categorical (yes or no), not a quantitative number. If the response was quantitative, we would consider using multiple linear regression. We will use logistic regression, as opposed to Poisson regression, because Poisson regression requires the response variable to be a count, but we are not counting the predicted number of pine beetles. We just want to know the chances of there being an infestation of pine beetles in a certain area, so we will use logistical regression.

Statistical Modeling

Running the bestglm function in R, the variables selected to use for our logistical model were: January minimum temperature, August maximum temperature, Slope of mountain, Elevation, Precipitation amount, and if the region is NC, SW, or SE. I chose the "AIC" metric, because we want to predict the probability of there being an infestation. I chose the "exhaustive" algorithm because of the relatively small amount of variables that we have to choose from.

Our model comes to be:

Y_i ~(ind.) Bern(p_i); $\log(p_i/(1-p_i)) = \beta_0 + \beta_1 x_i(\text{January}) + \beta_2 x_i(\text{August}) + \beta_3 x_i(\text{Slope}) + \beta_4 x_i(\text{Elevation}) + \beta_5 x_i(\text{Precipitation}) + \beta_6 x_i(\text{NC-yes}) + \beta_7 x_i(\text{SW-yes}) + \beta_8 x_i(\text{SE-yes}) + \varepsilon_i$

β_1 = Holding all other factors constant, as minimum January temperature increases by one degree, we expect an increase in the log-odds ratio by β_1, on average. Another way to interpret this value is: Holding all other factors constant, as minimum January temperature increases by one degree, the likelihood of there being an infestation of pine beetles is increased by $100*((e^{\beta_1})-1)\%$, on average.

β_7 = Holding all other factors constant, if you go from any other region to the southeast region, we expect an increase in the log-odds ratio by β_7, on average. Another way to interpret this value is: Holding all other factors constant, as you go from any other region to the southeast region, the likelihood of there being an infestation of pine beetles is increased by 100*((e^β_7)-1)%, on average.

For logistic regression, we have two assumptions: linearity in log-odds (monotone in probability), and independence. If we take a look at the graphs back on the second page, all are monotone in probability, but January temperature is the only factor that may be of concern. While it eventually flattens out to a consistent slope, there is a slight bump in the beginning towards the middle. For now, we are going to ignore this and go ahead and assume linearity in our relationships. The other assumption, independence of data gathered, is also a point of concern. The temperature of a region will definitely be related to the temperature of a region next to it. But again, we are going to ignore this for now and assume independence of data.

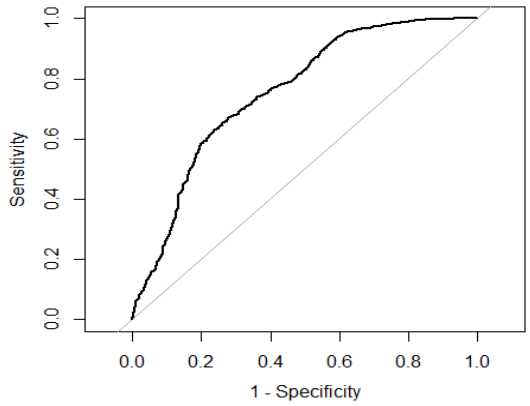
Results

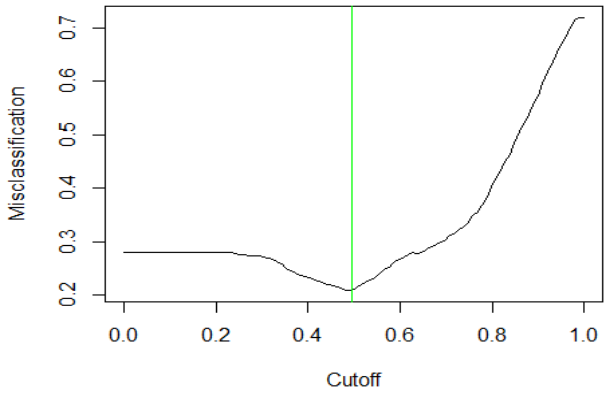
The estimated effects of each factor from our model and uncertainty of each can be found in the table below:

Variable	2.5 %	Actual Estimate	97.5 %
(Intercept)	-4.993678e+00	-2.2897076	0.3957892948
January	-2.192101e-01	-0.1645551	-0.1106598404
August_max	-1.332574e-01	-0.0861046	-0.0391276391
Slope	-1.299132e-02	0.0550539	0.1243144334
Elevation	9.136559e-06	0.0002287	0.0004494225
Precipitation	2.184840e-03	0.0029940	0.0038241824
NCYes	-1.720940e+00	-1.3290251	-0.9421374810
SEYes	-1.064886e+00	-0.7301014	-0.3954468665
SWYes	7.637815e-02	0.4974778	0.9285117849

To interpret this, let's take a look at the confidence interval and estimate for "Elevation." With 95% confidence, we believe the true effect of increasing elevation of a mountain by one foot will increase the log-odds ratio of having an infestation somewhere between (9.136559e-06, 0.0004494225), on average.

Through R, the pseudo-R^2 value came to be 0.1442902. This tells us the percent of variation in log(p/(1-p)) due to the factors in our model. This isn't any concern because the upper bound is not 1. Also, our AUC (area under the curve) has been determined to be 0.7555. Area under the curve tells us how well we classify across all thresholds. 0.7555 is a high amount, so we know the model is a good fit. The AUC's graph is at the top left of the next page. The threshold probability value has been calculated to be 0.4949495. This is shown (in the graph at the top right of the next page) to be the value at which the smallest amount of misclassification occurs.





The following confusion matrix was also developed in R:

	Predicted No	Predicted Yes	Sum
Actual No	252	395	647
Actual Yes	86	1577	1663
Sum	338	1972	2310

Specificity: 252/647 = approx. 38.95%. This is the percent of locations we predicted not to have an infestation, of all that did not have an infestation

Sensitivity: 1577/1663 = approx. 94.83%. This is the percent of locations we predicted had an infestation, of all that had infestations.

Positive predictive value: 1577/1972 = approx. 79.97%. This is the percent we correctly predicted had an infestation, of all places we predicted had an infestation.

Negative predictive value: 252/338 = approx. 74.56%. This is the percent we correctly predicted did not have an infestation, of all places that we predicted did not have an infestation.

To see how well our model correctly predicts new locations of pine beetle infestations, I ran cross-validation studies in R. The average specificity came to be around 0.3839261, the average sensitivity came to be around 0.9442728, the average positive predicted value came to be around 0.7981496, and the average negative predicted value came to be around 0.7300411. Our specificity is not very high, however our sensitivity is really high at 0.9442728. This means we are great at correctly predicting which location will have an infestation, but not great at saying which places will not have an infestation. One way to look at this model is that it takes on a “better safe than sorry” approach. This model leans towards predicting there to be an infestation, on average.

Looking at the projected data for the next ten years and applying our model to it, I am led to believe that there will likely be an infestation in the region during the next ten years. The probabilities for an infestation for each year are much greater than the threshold probability value, found earlier. I would strongly recommend concentrating on this area now.

Conclusions

Our logistical regression model for detecting future pine beetle infestations is trustworthy. It is great at correctly predicting future infestations, and this is proven true because of cross-validation simulations we conducted. The only concern we may have is in what we assumed, in order to even make a model. Temperatures will be related to each other between regions close together. To address this, you will need to talk to someone who has been taught how to deal with this. Finally, with the provided projected data ten years from now, our model predicts an infestation each year in this location.

To better predict infestations in the future, having more data always helps. You may also consider different potential factors that contribute to infestations. For instance, the prevalence of a certain type of tree may be a factor. Pine beetles may prefer Douglas fir trees as opposed to Aspen trees.

R Code Appendix

```
library(pROC)
library(bestglm)
library(ggplot2)
library(MASS)
library(lmtest)
library(car)
library(tidyverse)

PB <- read.csv("~/Stat 123/Stat 330/Stat 330/PineBeetle2 (1).csv", header = TRUE)
View(PB)

#Explore Data

PB$Infested <- as.numeric(PB$Infested)-1

scatter.smooth(PB$January,PB$Infested, xlab = "January Temp", ylab = "Infestation Probability")
scatter.smooth(PB$Precip,PB$Infested, xlab = "Precipitation", ylab = "Infestation Probability")
scatter.smooth(PB$August_max,PB$Infested, xlab = "August Temperature", ylab = "Infestation Probability")
scatter.smooth(PB$Slope,PB$Infested, xlab = "Slope", ylab = "Infestation Probability")
scatter.smooth(PB$Elev,PB$Infested, xlab = "Elevation", ylab = "Infestation Probability")


#Best Model

vs.res <- bestglm(PB,IC = "AIC", method = "exhaustive", family = binomial)
vs.res
vs.res$BestModels

PBLM <- glm(data = PB, family = binomial, Infested ~ January + August_max + Slope + Elev + Precip + NC + SE + SW)
PBLM
summary(PBLM)

#Effect of each coefficient

confint(PBLM)

#Threshold

thresh <- seq(0,1,length=100)
misclass <- rep(NA,length=length(thresh))
pred.probs <- predict.glm(PBLM,type="response")
for(i in 1:length(thresh)) {
  #If probability greater than threshold then 1 else 0
  my.classification <- ifelse(pred.probs>thresh[i],1,0)
```

```
  # calculate the pct where my classification not eq truth
  misclass[i] <- mean(my.classification!=PB$Infested)
}
#Find threshold which minimizes miclassification
thresh[which.min(misclass)]
T <- thresh[which.min(misclass)]
plot(thresh, misclass, pch=20, type = "l", xlab = "Cutoff", ylab = "Misclassification")
abline(v=T, col = "green")

#Matrix

pred.class <- ifelse(pred.probs>T, 1,0)
table <- table(PB$Infested,pred.class)
addmargins(table)

deviance <- PBLM$deviance
nullDeviance <- PBLM$null.deviance
1-(deviance/nullDeviance)


n.cv <- 500
n.test <- round(.1*nrow(PB))
cutoff <- T
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)
for(cv in 1:n.cv){
  test.obs <- sample(1:nrow(PB),n.test)
  test.set <- PB[test.obs,]
  train.set <- PB[-test.obs,]
  train.model <- glm(Infested~January+August_max+Slope+Elev+Precip+NC+SE+SW,data=train.set,family=binomial)
  pred.probs <- predict.glm(train.model,newdata=test.set,
                            type="response")
  test.class <- ifelse(pred.probs>cutoff,1,0)
  conf.mat <- addmargins(table(test.set$Infested,test.class))
  sens[cv] <- conf.mat[2,2]/conf.mat[2,3]
  spec[cv] <- conf.mat[1,1]/conf.mat[1,3]
  ppv[cv] <- conf.mat[2,2]/conf.mat[3,2]
  npv[cv] <- conf.mat[1,1]/conf.mat[3,1]
  auc[cv] <- auc(roc(test.set$Infested,pred.probs))
}

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

summary(PBLM)

#Projected
PB2018 <- data.frame(SE = "Yes", NC = "No", SW = "No", Slope = 18.07, Elev = 1901.95, January = -13.98, August_max =
15.89, Precip = 771.13)
#0.8390319
PB2019 <- data.frame(SE = "Yes", NC = "No", SW = "No", Slope = 18.07, Elev = 1901.95, January = -17.80,August_max =
18.07, Precip = 788.54)
#0.8951125
PB2020 <- data.frame(SE = "Yes", NC = "No", SW = "No", Slope = 18.07, Elev = 1901.95, January =-17.27,August_max
=16.74, Precip = 677.63)
#0.8628658
```

```
PB2021 <- data.frame(SE = "Yes", NC = "No", SW = "No", Slope = 18.07, Elev = 1901.95, January =-12.52,August_max
=18.06, Precip = 522.77)
#0.6178305
PB2022 <- data.frame(SE = "Yes", NC = "No", SW = "No", Slope = 18.07, Elev = 1901.95, January =-15.99,August_max
=18.23, Precip = 732.32)
#0.8407897
PB2023 <- data.frame(SE = "Yes", NC = "No", SW = "No", Slope = 18.07, Elev = 1901.95, January =-11.97,August_max
=15.81, Precip = 615.96)
#0.703202
PB2024 <- data.frame(SE = "Yes", NC = "No", SW = "No", Slope = 18.07, Elev = 1901.95, January =-15.75,August_max
=16.85, Precip = 805.90)
#0.8769379
PB2025 <- data.frame(SE = "Yes", NC = "No", SW = "No", Slope = 18.07, Elev = 1901.95, January =-16.19,August_max
=16.51, Precip = 714.57)
#0.8571702
PB2026 <- data.frame(SE = "Yes", NC = "No", SW = "No", Slope = 18.07, Elev = 1901.95, January =-17.87,August_max
=17.84, Precip = 740.50)
#0.8840701
PB2027 <- data.frame(SE = "Yes", NC = "No", SW = "No", Slope = 18.07, Elev = 1901.95, January =-12.44,August_max
=16.96, Precip = 801.22)
#0.8014825



pred.log.odds <- predict.glm(PBLM,newdata=PB2027)
pred.prob <- exp(pred.log.odds)/(1+exp(pred.log.odds))
pred.prob <- predict.glm(PBLM,newdata=PB2027,type="response")
pred.prob

pred.probs <- predict.glm(PBLM,type="response")
a.roc <- roc(PB$Infested,pred.probs)
auc(a.roc)

plot(a.roc,legacy.axes=TRUE)
```