Elizabeth Mitchell

Stat 330

17 December 2019

**Final Exam**

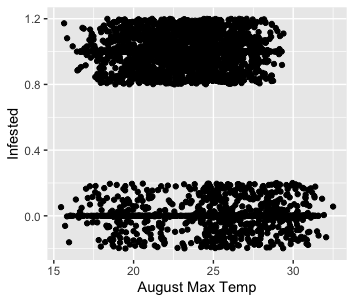
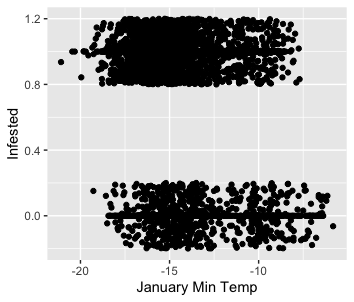
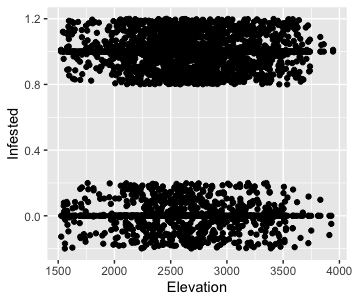
Mountain Pine Beetle Analysis

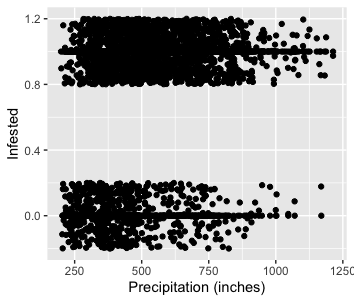
**Introduction**

The mountain pine beetle (MPB), once a helpful agent in the forest life cycle has become a harmful epidemic in the Rocky Mountains. In the last few decades milder weather conditions haven’t been enough to control the size of the beetle population. With the surplus population, beetles that in the past burrowed into weakened or dead trees are now attacking healthy trees as well and damaging entire forests. We want to determine which factors contribute to risk of infestation as well as predict if areas are at risk for future infestation.

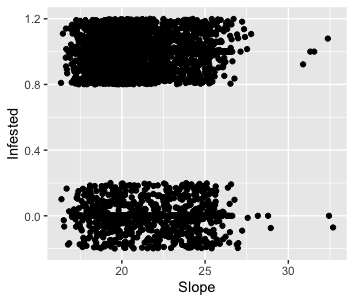
This dataset includes information on many areas, including whether or not they are infested, what region it is found in, the lowest temperature in January, highest temperature in August, the angle of the mountain, elevation and precipitation. We will determine which of these factors have a significant effect on the odds of an area being infested and how those factors impact the odds.

For an initial exploration of the data we will look at jitter-plots of the quantitative explanatory variables vs. “Infested” as well as a cross tabulation of region with “Infested.” From these plots we see that there is a slight concentration of not infested sections that have higher minimum temperatures in January and higher maximums in August. As far as elevation goes, there seem to be a sprinkling of areas at higher elevations that aren’t infested with fewer at those elevations that are infested. With more precipitation there seem to be more infested than not while slope doesn’t have any notable patterns.



The cross tabulation of region with infestation shows that overall there are many more infested regions than not. The north central region (NC) seems notable for having a lower portion of infected areas, only 129 out of 252. The south central region, on the other hand, has the highest portion of infested areas with 216 out of 252.

|  |  |  |  |
| --- | --- | --- | --- |
|  | No | Yes | Sum |
| EC | 80 | 186 | 266 |
| NC | 123 | 129 | 252 |
| NW | 46 | 206 | 252 |
| SC | 36 | 216 | 252 |
| SE | 113 | 153 | 266 |
| SW | 58 | 194 | 252 |
| WC | 38 | 214 | 252 |
| Sum | 494 | 1298 | 1792 |



We will use a logistic regression model because the response variable in our data is categorical– whether the area is infested or not. A linear regression model would not be appropriate for this data because with a categorical response variable the assumptions of linear regression would be violated and our predictions wouldn’t be limited to a binary response as would be appropriate.

**Statistical Modeling**

We determined the best model using best subset selection with AIC criteria. This is appropriate because we don’t have enough variables to prevent us using best subset selection and AIC criteria is suitable for predictions. Since a primary goal of this analysis is to predict which areas will become infested in the next ten years, prediction is a priority. The result of this selection is a model including the following response variables: January max temperature, August min temperature, slope, elevation, precipitation, and if it is in the north central, south east or south west regions.

Therefore we will use the following model:

Bern

log =

Yi refers to each observation’s value for being infested or not. This will follow a Bernoulli distribution of pi where the log of pi divided by 1 - pi equals a value for the intercept and each subsequent as the correlation coefficient for the respective variable. In this case we have to for each of the included covariates. For example, refers to the slope of which is the minimum January temperature. So for every increase in the minimum January temperature, the log odds ratio will increase by . refers to the slope associated with whether the area is in the north central region or not. If the area is in the north central region the log odds ratio will increase by .

In order to use this model we assume that there is a monotonic log-odds relationship between infestation and the following quantitative explanatory variables: January max temperature, August min temperature, slope, elevation, and precipitation. We can see that this assumption is met in the following scatterplots. Since all of the smoothed lines are monotonic we are safe to move forward with the assumption of a monotonic log-odds relationship. The other assumption to consider is independence. We will assume independence for the sake of the analysis but realistically, it is unlikely that the infestation of one area is independent from the infestation in another since the regions may be touching.

A screenshot of a cell phone

Description automatically generated

A close up of a logo

Description automatically generatedA screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated

**Results**

The adjacent table provides slope estimates and associated confidence intervals for each of the coefficients included in the model. These estimates should be untransformed to be appropriately interpreted. For example, by raising to the interval for August max, subtracting 1 and multiplying by 100 we get the interval (-.0125, -10.476). This means that we are 95% confident that as the as the maximum temperature in august increases by 1 the chance of an area being infested goes down by between .0125 and 10.476 percent. Likewise, we are 95% confident that, all else held equal, if an area is in the north central region the likelihood that it is infested decreases by between .821 and 61.021 percent. Overall it appears that an increase in the variables January min, August max, north central and south east all lessen the odds of a region being infested. On the other hand, being in the southwest region and an increase in precipitation increase the odds of infestation.

I chose a cut off value of .4873 because that is the value that minimizes misclassification, see graph below.

A screenshot of a cell phone

Description automatically generated

This cutoff value in conjunction with the fitted model gives us a sensitivity of .956; this means that of the areas that of the infested areas our model correctly predicted 95.6%. The specificity is .379, meaning that of non-infested areas our model correctly predicted 37.9%. We have a positive predicted value of .798 (79.8% of the areas that were predicted to be infested actually were) and a negative predicted value of .768 (76.8% of the areas predicted to not be infested were not infested.) This is a reasonably good fit, especially if catching areas that are likely to be infested is a priority.

To assess the model’s predictive capabilities, I ran 1,000 cross validations reporting the sensitivity, specificity, PPV, and NPV for each. This resulted in a mean sensitivity of 0.952, specificity of 0.369, PPV of 0.796 and NPV of 0.751. Therefore, we can conclude that this model will correctly predict approximately 95% of areas likely to be infested. On the other hand, it will only correctly predict approximately 37% of areas that are not in fact at risk.

For the given area in the south east I recommend that the forest service concentrate on preventing infestation in this region over the next ten years as it is likely to become infested. The predicted likelihoods for infestation over each of the next ten years are between .617 and .895 with a mean likelihood of .818. Since these are all well above the established cutoff of .4873 it is reasonable to predict that this area will become infested in the next 10 years.

**Conclusion**

From this analysis we found that the increases in minimum temperature in January, maximum temperature in august, and being in the north central or south east region decrease the chances that an area will have a beetle infestation. Additionally, we found that increases in precipitation, and being in the north central region increase the chance of an area having a beetle infestation. In the future it would be useful to perform a similar analysis but taking into account the challenges that questionable independence presented us. It would be interesting to look at different areas in the world where there are mountain pine beetle infestations if any exist. It would also be interesting to look at beetle infestation as a quantitative variable, examining how infested an area is rather than just whether it is infested or not.

Code:

library(ggplot2)

library(gmodels)

library(bestglm)

library(pROC)

library(stargazer)

library(dplyr)

library(car)

library(stargazer)

pine.beetles <- read.csv("PineBeetle.csv", sep = ",", head = TRUE)

pine.beetles$Infested <- as.numeric(pine.beetles$Infested)

pine.beetles$Infested[pine.beetles$Infested == 1] <- 0

pine.beetles$Infested[pine.beetles$Infested == 2] <- 1

pine.beetles <- mutate(pine.beetles, region = ifelse(NW == "Yes", "NW", ifelse(SC == "Yes", "SC", ifelse(SE == "Yes", "SE", ifelse(SW == "Yes", "SW", ifelse(WC == "Yes", "WC", ifelse(NC == "Yes", "NC", ifelse(EC == "Yes", "EC", NA))))))))

ggplot(data = pine.beetles, mapping = aes(x = January, y = Infested)) + geom\_point() + geom\_jitter(height = .2, width = 1) + xlab("January Min Temp")

ggplot(data = pine.beetles, mapping = aes(x = August\_max, y = Infested)) + geom\_point() + xlab("August Max Temp")+ geom\_jitter(height = .2, width = 1)

ggplot(data = pine.beetles, mapping = aes(x = Slope, y = Infested)) + geom\_point() + geom\_jitter(height = .1, width = 1)

ggplot(data = pine.beetles, mapping = aes(x = Elev, y = Infested)) + geom\_point() + geom\_jitter(height = .2, width = 1) + xlab("Elevation")

ggplot(data = pine.beetles, mapping = aes(x = Precip, y = Infested)) + geom\_point() + geom\_jitter(height = .2, width = 1) + xlab("Precipitation (inches)")

addmargins(table(pine.beetles$region, pine.beetles$Infested))

pine.beetles$region <- NULL

beetle\_model <- bestglm(pine.beetles, IC = "AIC", method = "exhaustive", family = binomial)

beetle\_model <- beetle\_model$BestModel

summary(beetle\_model)

scatter.smooth(pine.beetles$January, pine.beetles$Infested, xlab = "January Min Temp", ylab = "Infested")

scatter.smooth(pine.beetles$August\_max, pine.beetles$Infested, xlab = "August Max Temp", ylab = "Infested")

with(pine.beetles, lines(loess.smooth(August\_max, Infested), col = "red"))

scatter.smooth(pine.beetles$Slope, pine.beetles$Infested, xlab = "Slope", ylab = "Infested")

plot(pine.beetles$Elev, pine.beetles$Infested, xlab = "Elevation", ylab = "Infested")

with(pine.beetles, lines(loess.smooth(Elev, Infested), col = "red"))

scatter.smooth(pine.beetles$Precip, pine.beetles$Infested, xlab = "Preciptation", ylab = "Infested")

(exp(confint(beetle\_model))- 1)\*100

stargazer(beetle\_model, out = "beetles.html", ci = T)

pred.probs <- predict.glm(beetle\_model, type="response")

thresh <- seq(from=0, to=1, length=10000)

misclass <- rep(NA,length=length(thresh)) #Empty vector to hold misclassification rates

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!=pine.beetles$Infested)

}

(c <- thresh[which.min(misclass)])

ggplot(mapping = aes(x = thresh, y = misclass)) + geom\_point() + geom\_vline(xintercept =c) + xlab("Threshold") + ylab("Missclassifications")

my.roc <- roc(pine.beetles$Infested, pred.probs)

ggplot() + geom\_line(aes(x=1-my.roc[["specificities"]], y=my.roc[["sensitivities"]])) + geom\_abline(intercept=0, slope=1)

(auc(my.roc))

pred.class <- ifelse(pred.probs > c, 1, 0)

conf.matrix <- addmargins(table(pred.class, pine.beetles$Infested))

(sens <- conf.matrix[2,2]/conf.matrix[3, 2])

(spec <- conf.matrix[1, 1]/conf.matrix[3, 1])

(PPV <- conf.matrix[2,2]/conf.matrix[2, 3])

(NPV <- conf.matrix[1, 1]/conf.matrix[1, 3])

conf.matrix

n.cv <- 1000

n.test <- round(.1\*nrow(pine.beetles))

## Set my threshold for classifying

cutoff <- c

## 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){

test.obs <- sample(1:nrow(pine.beetles), n.test)

test.set <- pine.beetles[test.obs,]

train.set <- pine.beetles[-test.obs,]

## Fit best model to training set

train.model <- glm(Infested ~ January + August\_max + Slope + 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")

## Classify according to threshold

test.class <- ifelse(pred.probs>cutoff, 1, 0)

## Create a confusion matrix

conf.mat <- addmargins(table(factor(test.set$Infested, levels = c(0, 1)), 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]

}

(mean(sens))

(mean(spec))

(mean(ppv))

(mean(npv))

ten.years <- data.frame(January = c(-13.98, -17.8, -17.27, -12.52, -15.99, -11.97, -15.75, -16.19, -17.87, -12.44), August\_max = c(15.89, 18.07, 16.74, 18.06, 18.23, 15.81, 16.85, 16.51, 17.84, 16.96), Slope = rep(18.07, 10), Elev = rep(1901.95, 10), Precip = c(771.13, 788.54, 677.63, 522.77, 732.32, 615.96, 805.9, 714.57, 740.5, 801.22), SE = rep("Yes", 10), NC = rep("No", 10), SW = rep("No", 10))

predict.glm(beetle\_model, newdata = ten.years, type = "response")

mean(predict.glm(beetle\_model, newdata = ten.years, type = "response"))