Daniel Witt

**REGRESSION ANALYSIS - PINE BEETLE DAMAGE**

**Section 1: Introduction and Problem Background**

The mountain pine beetle (MPB) is a species that, in the past, has helped develop many forests in the western United States. They tend to feed on weaker trees, which allow the forests to grow more quickly. However, in light of recent climate changes and higher temperatures, the MPB population has skyrocketed and started to have a negative effect on the health of the forests.

Data has been collected from a series of aerial detection surveys that observes the amount of MPB damage form the southern Rocky Mountains to southern Wyoming to the Black Hills of South Dakota. Many areas were surveyed and recorded as to whether they had MPB damage contained in them, along with other variables that are believed to help explain the cause of so much MPB damage. In order to reduce MPB damage, the goal of this study is to determine which variables significantly affect MPB infestation and to predict where future MPB infestations will take place. We plan to use the data collected from the aerial detection surveys, which has been compiled into a CSV file, to address these problems. The explanatory variables we will observe include:

Infested - Is the region infested with pine beetle?

January - average January minimum temperature in degrees C

August\_max - average August maximum temperature

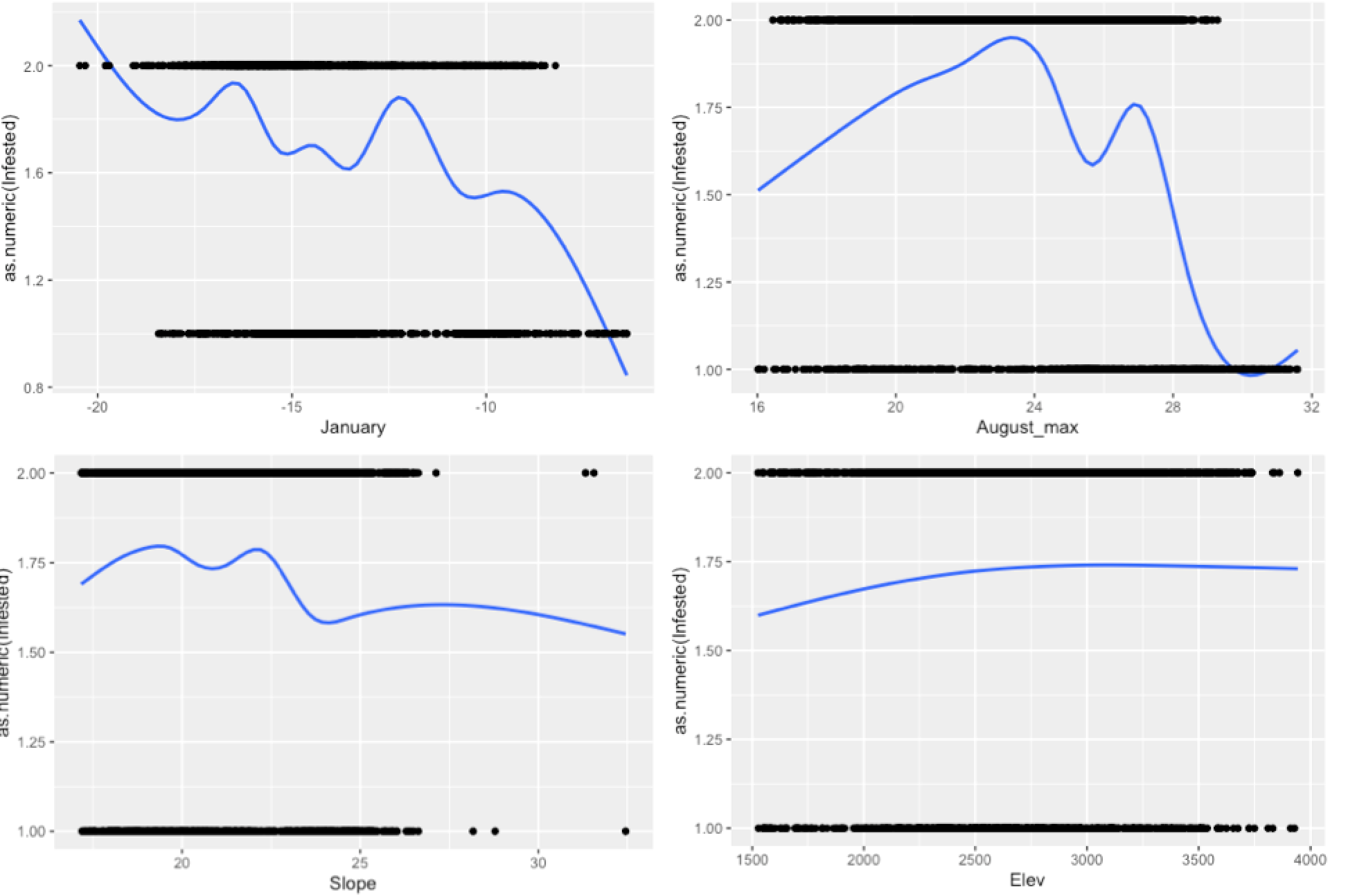
Slope - angle of mountain slope

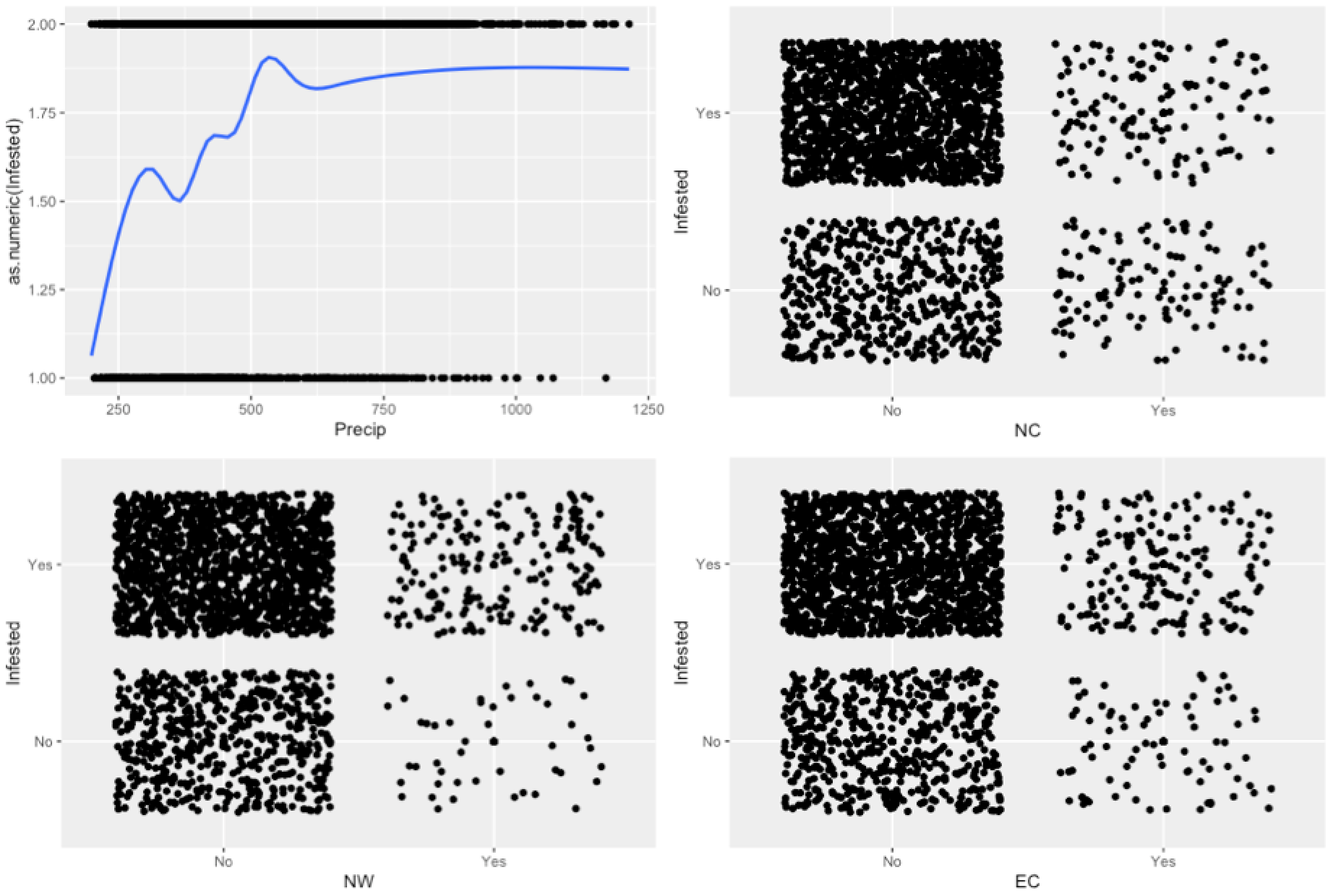
Elev - elevation in feet

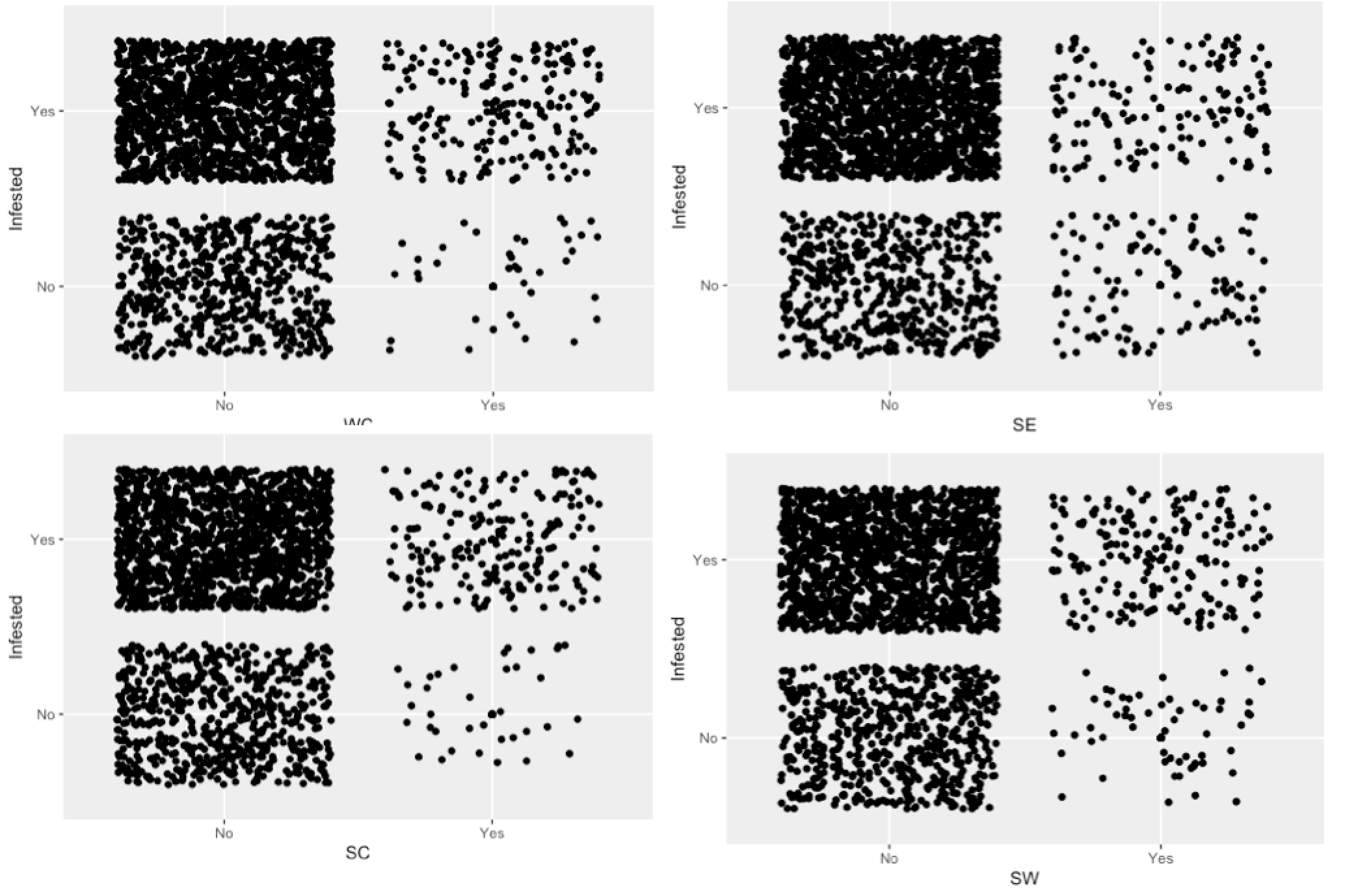
Precip - mean annual precipitation in inches

NC,SW,... - Region indicators

We can explore the data by graphing it to look for patterns, and to get a better idea of what analysis we should perform:







By looking at our various scatter smooth graphs for our quantitative variables, we can see some patterns. It looks as though as the minimum average January temperature increases, the amount of areas recorded as infested with MPBS decreased. This also seemed to be the pattern for the angle of the mountain slope. As the slope increased, the amount of MPB infestation seemed to drop as well. There were also positively correlated explanatory variables as well, like the mean annual precipitation, which seemed to have more MPB infestation when there were higher recorded precipitation rates. This was the same for elevation.

We can observe some patterns in the affect that different areas has on the amount of MPB infestation by looking at the jittered scatterplots of the different regions. From looking at our graphs, it seems that southern areas tended to have more infestation than the northern areas.

Based on our data, we will use a logistic regression model to answer the problems from our data. This is because our response variable, whether or not the area is infested with MPBs, is categorical, which means we cannot use multiple linear regression. Logistic regression will allow us to determine the explanatory variables that most significantly influence whether or not an area is infested with MPBs. It will also allow us to predict probabilities for each area on how likely they are to be infested, which will help us to predict future areas that will develop MPB infestation.

**Section 2: Statistical Modeling**

In order to determine which explanatory variables are important to include in our logistic regression model for our MPB infestation data, we will use variable selection using best subset selection (exhaustive) algorithm, and a BIC metric. We want to use the exhaustive algorithm because it will look at every possible combination of variables and return the best model fit for our data. It will give us a model that best explains the probability that an area will develop MPB infestation or not. For this particular model, we want to use BIC for our metric because we are mainly interested in determining what factors lead to MPB infestation. Using BIC technique will give us a model that helps us determine the most important variables, whereas AIC is more oriented towards prediction, and, although we want to be able to predict which areas will be infested in the future, it is more important to understand what factors lead to infestation so that the infestation can be combatted more effectively.

When we perform variable selection with these parameters, we identify these variables as the most important to use in our model for our data:

January - average January minimum temperature in degrees Celsius

August\_max - average August maximum temperature in degrees Celsius

Precip - mean annual precipitation in inches

NCYes

SEYes

Our logistic regression model can be written in this way:

log(*pi /1–pi)* = **β0 + ΣJj=1 *xij*βj**

***yi*** *ind~Bern(pi)*

where:

*pi*  = the probability that an area will be infested by MPBs

**β0** =y-intercept = when an area has an average minimum January temperature of 0 degrees Celsius, a max August temperature of 0 degrees Celsius, has 0 inches of precipitation, and is not located in the NC or SE areas, the log odds ratio will be to **β0**

*Xij* = each explanatory variable value, or the value of January, August\_max, Precip, NC, and SE

**Βj** = For every unit increase in Xj, the log-odds ratio increases by **Βj**, or As Xj increases by 1, a area is exp(**Βj)** times more likely to have infestation, or as Xj increases by 1, an area is more 100 \* (exp(**Βj** ) – 1)% more likely to have infestation

Our model follows a Bernoulli distribution.

Our model calculated these coefficients:

|  |  |
| --- | --- |
| Variable | Coefficient |
| Intercept | -0.157 |
| January | -0.146 |
| August\_max | -0.085 |
| Precip | 0.002 |
| NCYes | -1.218 |
| SEYes | -0.919 |

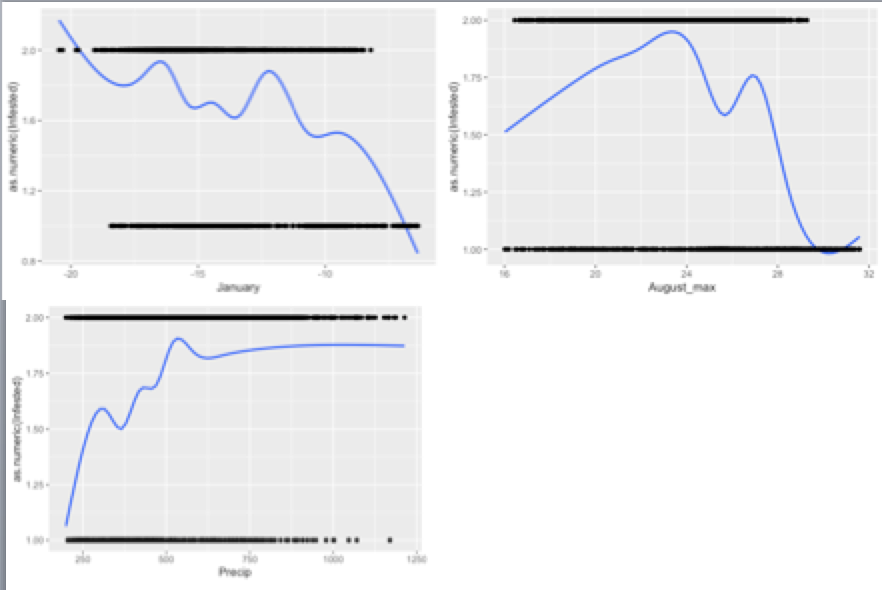
Our quantitative coefficients can be interpreted in this way:

For every additional inch of precipitation an area receives, the log odds that an area will be infested increases by 0.002, is 1.002 times more likely to be infested, or is 0.287% more likely to be infested, holding all else constant.

Our categorical coefficients can be interpreted in this way:

If an area is located in the NC region, the log odds that the area will be infested increases by -1.218, or is 0.295 times more likely to be infested, or 70.425% LESS likely to be infested, holding all else constant.

In order to use our logistic regression model, we need to check our assumptions of linearity and independence. We can check linearity by looking at our scatter smooth plots from earlier:



For the most part, our scatter-smooth plots seem to be relatively linear in shape. The linearity assumption that is most worrying is the August\_max because it seems to start with an upward trend and then move down to a negative trend. However, the overall trend seems to be relatively downwards, so we can assume this assumption is met for our data.

When considering independence, we can assume that each area is independent of each other. That is, each area does not influence the other area in a significant way, so we can assume that our model meets the independence assumption.

**Section 3: Results**

We can observe the effects on infestation of our selected variables, with a measure of uncertainty by using confidence intervals:

|  |  |  |  |
| --- | --- | --- | --- |
| Variable | Fit | Confidence Interval Lower | Confidence Interval Upper |
| Intercept | -0.157 | -1.886 | 1.568 |
| January | -0.146 | -0.193 | -0.100 |
| August\_max | -0.085 | -0.132 | -0.038 |
| Precip | 0.002 | 0.002 | 0.003 |
| NCYes | -1.218 | -1.512 | -0.923 |
| SEYes | -0.919 | -1.218 | -0.619 |

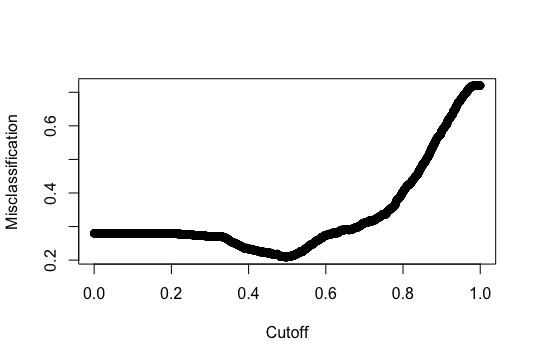
Our quantitative effects can be interpreted in this way:

We are 95% confident that, holding all else constant, for every additional inch of precipitation an area receives, the log odds that an area will be infested increases by between 0.002 and 0.003, is between 1.002 and 1.003 times more likely to be infested, or is between 0.207% and 0.369% more likely to be infested.

Our categorical effects can be interpreted in this way:

We are 95% confident that, holding all else constant, if an area is located in the NC region, the log odds that the area will be infested increases by between -1.512 and -0.923, or is between 0.220 and 0.396 times more likely to be infested, or is between 77.969% and 60.300% LESS likely to be infested.

We can now assess how well our model fits our data by assessing our ROC curve, developing a cutoff classification probability, and observing our sensitivity, specificity, and positive predicted value, and negative predicted value. First we will look develop a threshold cutoff probability so that we can classify if an area is predicted to be infested with MPBs or not based on their probability of being infested. We pick a cutoff value that minimizes the misclassification rate of whether or not an area is infested. The value that does this is 0.497. This means that any probability above this cutoff is predicted to be infested and any probability below this value is predicted to not be infested. . We can see from this graph of the different misclassification rates that this is indeed the threshold that minimizes the number of misclassifications:



Now we can create a confusion matrix in order to calculate the sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV):

|  |  |  |  |
| --- | --- | --- | --- |
|  | Predicted Yes | Predicted No | Sum |
| True Yes | 1577 | 86 | 1663 |
| True No | 396 | 251 | 647 |
| Sum | 1973 | 337 |  |

Sensitivity (Percent of true positives) = 0.948

This is the ability of our model to correctly identify those areas that are infested with MPBs.

Specificity (Percent of true negatives) = 0.387

This is the ability of our model to correctly identify those areas that are not infested with MPBs.

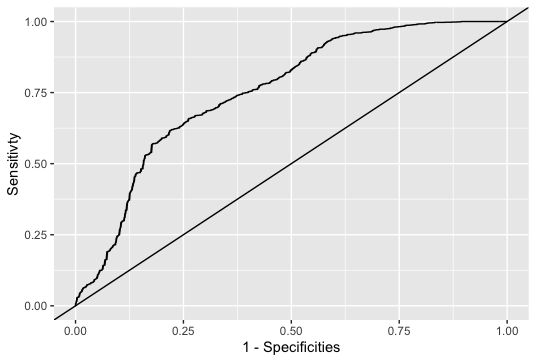
PPV (Percent of correctly predicted yes’s) = 0.799

This is the proportion of areas that are actually infested with MPBs that were predicted to have infestation.

NPV (Percent of correctly predicted no’s) = 0.744

This is the proportion of areas that are actually not infested with MPBs that were predicted to not have infestation.

Now we can calculate our ROC curve and determine the AUC:



The area under the curve (AUC) of our model tells us how well we classify across all thresholds. Our AUC was calculated as 0.7549, which shows that using our model gives a decent increase of accuracy in predicting if an area is infested or not, over simply guessing if it is infested or not.

By looking at all of our values, we can say that our model fits the data well. We have relatively high sensitivity, PPV, and NPV, which shows that our model is, for the most part, correctly predicting if an area is infested or not. However, our specificity rate, which calculates the percent of true negatives, is rather low. This means that our model is failing to correctly identify those areas without infestation. This means that more areas are being predicted as having infestation than is actually the case, so we must keep that in mind when using our model to predict whether or not an area is infested with MPBs.

We can assess how well our model does at predicting new locations where MPB infestations will occur by using cross validation. When we do this, we remove part of the data set and then create a model with the data and our variable selection, then we use the model to predict whether the areas that are we left out of our model are infested or not. This way we can see how well our model does at predicting MPB infestation. When we performed 1000 cross validations, we received these average values for sensitivity, specificity, PPV, and NPV:

Sensitivity (Percent of true positives) = 0.932

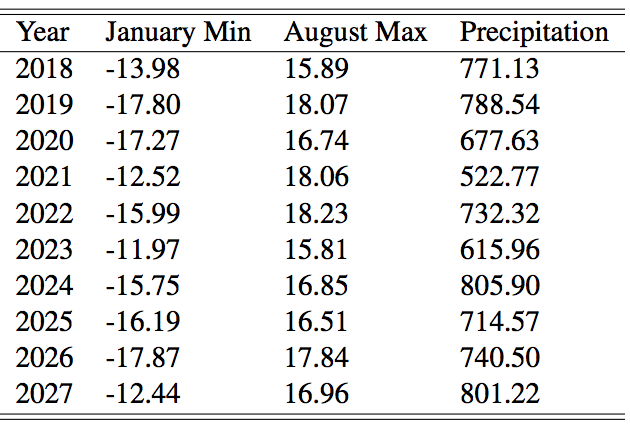
Specificity (Percent of true negatives) = 0.333

PPV (Percent of correctly predicted yes’s) = 0.766

NPV (Percent of correctly predicted no’s) = 0.723

Based on these values, it seems that our suspicions from earlier are the same. We are classifying sensitivity very well and our PPV and NPV are decently high, however, our model is having trouble classifying the specificity, or percent of true negatives. This means that we are often predicting that an area is infested with MPBs, when in fact it is not infested. This could be troublesome because much effort could be put into defending these areas from MPBs when in reality, focusing on other areas could be more important.

We can use our model to predict if a specific area will develop MPB infestation in the next 10 years. The area is located in the South East region and has the following forecasted values:



We can plug these values into our model to determine the probability that this specific area will develop an infestation for each year:

|  |  |
| --- | --- |
| Year | Probability of Infestation of MPBs |
| 2018 | 0.862 |
| 2019 | 0.905 |
| 2020 | 0.878 |
| 2021 | 0.673 |
| 2022 | 0.860 |
| 2023 | 0.750 |
| 2024 | 0.892 |
| 2025 | 0.874 |
| 2026 | 0.895 |
| 2027 | 0.832 |

Since these probabilities are very high for each year, and since they are all above our cutoff threshold of 0.497, we can predict that this particular area will become infested with MPBs in the next 10 years. We would suggest that the forest service should focus on this area to prevent infestation, since the probability is so high.

**Section 4: Conclusions**

From our analysis, we have developed a model that can predict whether or not an area of forest will develop a MPB infestation. The main goals we wanted to achieve in developing this model were to understand which factors played a significant role in determining if an area would become infested, and we wanted to be able to predict if an area would become infested. We developed a logistic regression model that helped us achieve both of these goals. We learned that the most significant variables that led to MPB infestation were the Minimum January Temperature, the Max August Temperature, the amount of Precipitation in inches, and whether the region was located in the NC or SE regions. Using this model we were able to develop decent predictions on whether an area would become infested with MPBs.

We would advise the ecologists to identify additional variables and collect data on those variables to better understand and predict MPB infestation. Although our model was fairly good at predicting infestation, the ability of the model to correctly predict that an area was not infested was low. If more variables were collected, we could again develop a model that could more accurately predict infestation.

R-Code:

# read in data

pine <- PineBeetle2

head(pine)

tail(pine)

#plot data

library(ggplot2)

ggplot(pine, mapping = aes(x=January, as.numeric(Infested))) + geom\_smooth(se = FALSE) + geom\_point()

ggplot(pine, mapping = aes(x=August\_max, as.numeric(Infested))) + geom\_smooth(se = FALSE) + geom\_point()

ggplot(pine, mapping = aes(x=Slope, as.numeric(Infested))) + geom\_smooth(se = FALSE) + geom\_point()

ggplot(pine, mapping = aes(x=Elev, as.numeric(Infested))) + geom\_smooth(se = FALSE) + geom\_point()

ggplot(pine, mapping = aes(x=Precip, as.numeric(Infested))) + geom\_smooth(se = FALSE) + geom\_point()

ggplot(data=pine, mapping=aes(x=NC,y=Infested)) + geom\_point() + geom\_jitter(width=0.4,height=0.4)

ggplot(data=pine, mapping=aes(x=NW,y=Infested)) + geom\_point() + geom\_jitter(width=0.4,height=0.4)

ggplot(data=pine, mapping=aes(x=EC,y=Infested)) + geom\_point() + geom\_jitter(width=0.4,height=0.4)

ggplot(data=pine, mapping=aes(x=WC,y=Infested)) + geom\_point() + geom\_jitter(width=0.4,height=0.4)

ggplot(data=pine, mapping=aes(x=SE,y=Infested)) + geom\_point() + geom\_jitter(width=0.4,height=0.4)

ggplot(data=pine, mapping=aes(x=SC,y=Infested)) + geom\_point() + geom\_jitter(width=0.4,height=0.4)

ggplot(data=pine, mapping=aes(x=SW,y=Infested)) + geom\_point() + geom\_jitter(width=0.4,height=0.4)

scatter.smooth(y=as.numeric(pine$Infested), x=as.numeric(pine$NC))

scatter.smooth(y=pine$Infested, x=pine$SE)

#fit model

library(bestglm)

vs.res1 <- bestglm(pine, IC="AIC", method="exhaustive", family = binomial)

vs.res2 <- bestglm(pine, IC="BIC", method="exhaustive", family = binomial)

summary(vs.res1$BestModel)

summary(vs.res2$BestModel)

pineModel <- vs.res2$BestModel

pineModel

#confint

confint(pineModel)

exp(confint(pineModel))

100\*(exp(confint(pineModel))- 1)

#ROC curve

library(pROC)

pred.probs <- predict.glm(pineModel, type="response")

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

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

auc(my.roc)

#pseudo r-squared

1-(pineModel$deviance/pineModel$null.deviance)

#classification threshold

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], 'Yes', 'No')

# calculate the pct where my classification not eq truth

misclass[i] <- mean(my.classification!=pine$Infested)

}

#Find threshold which minimizes miclassification

thresh[which.min(misclass)]

plot(thresh, misclass, xlab = 'Cutoff', ylab = 'Misclassification')

#confusion matrix

probthresh <- ifelse(pred.probs>thresh[which.min(misclass)], 'Yes', 'No')

table(pine$Infested, probthresh)

addmargins(table(pine$Infested, probthresh))

#cross validation

## Choose number of CV studies to run in a loop & test set size

n.cv <- 1000

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

## Set my threshold for classifying

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

## 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){

# Choose which obs. to put in test set

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

## Separate into test and training sets

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

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

## Fit best model to training set

train.model <- glm(Infested~January+August\_max+Precip+NC+SE, 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, 'Yes', 'No')

## Create a confusion matrix

conf.mat <- addmargins(table(factor(test.set$Infested, levels= c('Yes','No')), factor(test.class, levels=c('Yes','No'))))

## Pull of sensitivity, specificity, PPV and NPV using bracket notation

sens[cv] <- conf.mat[1,1]/conf.mat[1,3]

spec[cv] <- conf.mat[2,2]/conf.mat[2,3]

ppv[cv] <- conf.mat[1,1]/conf.mat[3,1]

npv[cv] <- conf.mat[2,2]/conf.mat[3,2]

## Calculate AUC

auc[cv] <- auc(roc(test.set$Infested, pred.probs))

} #End for-loop

sens[cv]

spec[cv]

ppv[cv]

npv[cv]

auc[cv]

#predictions

newdframe <- data.frame(SE = c('Yes'), NC = c('No'), 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),

Precip = c(771.13, 788.54, 677.63, 522.77, 732.32, 615.96, 805.90, 714.57, 740.50, 801.22))

pred.prob <- predict.glm(pineModel, newdata=newdframe, type="response")

pred.prob