Tyler Cann

Stat 330 – Heaton

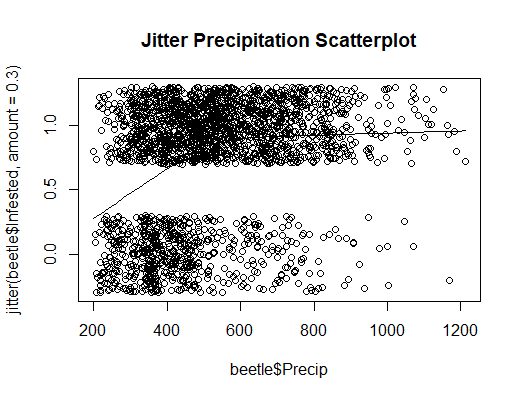
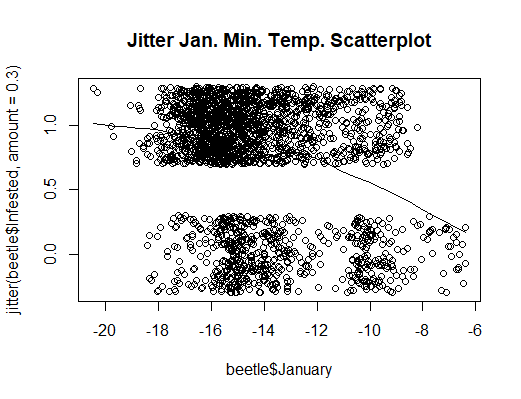
**Pine Beetle Damage**

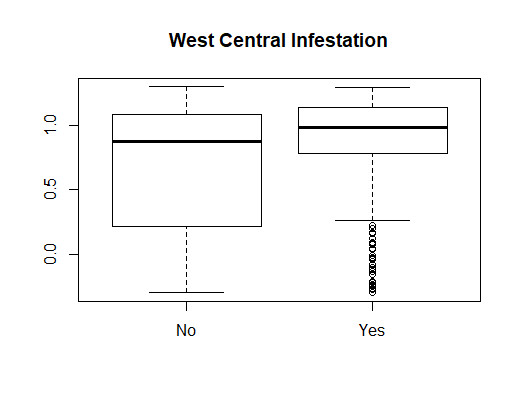
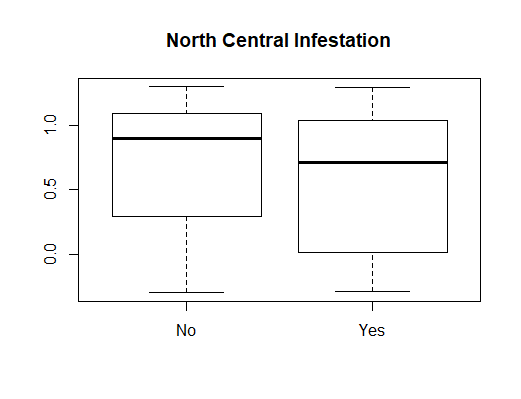
**Section 1: Introduction and Problem Background**

The mountain pine beetle has infested the forests of the western United States. Because of recent warm summers and dry conditions due to climate change, the pine beetles haven’t been killed off in the winter to the extent that they previously have been. To estimate the extent of pine beetle damage in Northern Colorado, data has been collected by the Colorado State Forest Service using Aerial Detection Surveys. We will analyze this data to determine which variables significantly affect pine beetle infestation and predict where future occurrences of pine beetle will take place.

We are going to use the PineBeetle.csv dataset we obtained from the CSFS to achieve our goals. We want to look at the relationship between whether the region is infested and the other variables (January temperatures, August temperatures, mountain slope, elevation, mean precipitation, and relative regional location).

First we want to analyze the data using some simple graphics and summaries.





The above plots are a summary of the data we were given. The scatterplots above are representations of the January Min. and August Max. Temperature variables in the Beetle dataset. Looking at these scatterplots, we can assume that that lower minimum temperatures in January appears to be an indicator of pine beetle infestation. The same can be said when there are higher levels of precipitation. The infestation levels appear to even out after around 500 inches though. It’s important to keep these variables in mind as we continue our analysis.

The boxplots are also representations of infestation levels in select regions, showing lower infestation levels in the north central region, and higher infestation levels in the west central region. These plots represented the most extreme disparities between “yes” and “no” results, but other less-extreme examples were also observed.

Traditional multiple linear regression wouldn’t be suitable for this problem because our response variable consists of a success (1 = the region is infested) or a failure (0 = the region is NOT infested). If we tried running regular multiple regression, then our predictions would fall outside of the prescribed {0,1} range that the response variable requires. For that reason, Logistic regression would be the most suited for the data that we’ve collected.

Not to mention, the traditional L-I-N-E (minus independence) assumptions would likely be violated:

-Linear assumptions might be violated (as we saw in the curved precipitation scatterplot)

-The errors won’t follow a normal curve (because we will assume a Bernoulli distribution)

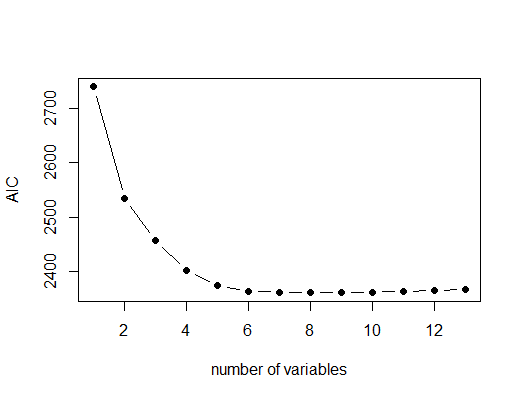
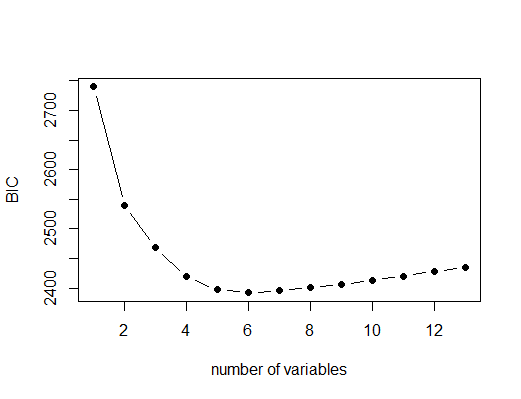
-Equal variance is likely going to be violated (The probability of beetle infestation [p(1-p)] changes from region

to region, therefore the variance isn’t equal between regions in our data).

**Section 2: Statistical Modeling**

The variable selection procedure that I chose was exhaustive selection. This was because the exhaustive method is the most thorough in scanning a data frame. Previous training I’ve received has said to use the exhaustive method unless the computer program won’t run it. (However, if I had to choose between forward and backward, I would choose forward because forward selection is best used for large datasets like this one.)

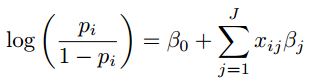
The model comparison criterion I chose was Bayesian Information Criterion. This was because the BIC only keeps the most important variables, and I only want to know which variables are the most important in finding out what affects beetle infestation rates. If we’re going to find a way to stop the pine beetles, we want to pinpoint the most direct cause, and simplifying which variables are included will help us do so. The BIC gave me 5 variables to use, while the AIC gave me 8, so I went with the BIC. The ones that only slightly affect the pine beetle infestation rates aren’t important to me, so I only want to see the most important variables.



BIC Coefficients AIC Coefficients

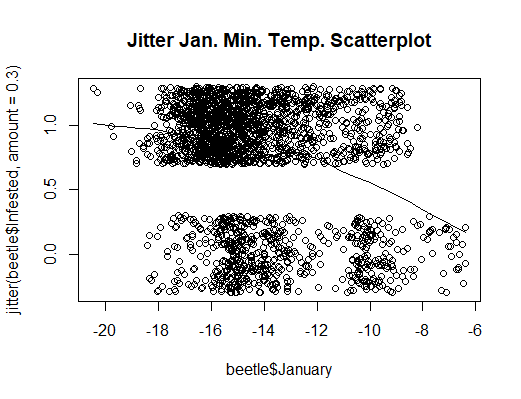
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Estimate** | **Std. Error** | **z value** | **Pr(>|z|)** |  |  | **Estimate** | **Std. Error** | **z value** | **Pr(>|z|)** |  |  |  |  |  |  |
| **Intercept** | -0.16 | 0.88 | -0.18 | 0.86 |  | **Intercept** | -2.29 | 1.37 | -1.67 | 0.95 |  |  |  |  |  |  |
| **January** | -0.15 | 0.02 | -6.19 | 0.00 |  | **January** | -0.16 | 0.03 | -5.95 | 0.00 |  |  |  |  |  |  |
| **August\_**  **Max** | -0.09 | 0.02 | -3.56 | 0.00 |  | **August\_**  **Max** | -0.09 | 0.02 | -3.59 | 0.00 |  |  |  |  |  |  |
| **Precip** | 0.003 | 0.0004 | 6.97 | 0.00 |  | **Slope** | 0.06 | 0.04 | 1.57 | 0.12 |  |  |  |  |  |  |
| **NCYes** | -1.22 | 0.15 | -8.12 | 0.00 |  | **Elev** | 0.00 | 0.00 | 2.04 | 0.04 |  |  |  |  |  |  |
| **SEYes** | -0.92 | 0.15 | -6.03 | 0.00 |  | **Precip** | 0.00 | 0.00 | 7.16 | 0.00 |  |  |  |  |  |  |
|  |  |  |  |  |  | **NCYes** | -1.32 | 0.20 | -6.69 | 0.00 |  |  |  |  |  |  |
|  |  |  |  |  |  | **SEYes** | -0.73 | 0.17 | -4.28 | 0.00 |  |  |  |  |  |  |
|  |  |  |  |  |  | **SWYes** | 0.50 | 0.22 | 2.29 | 0.02 |  |  |  |  |  |  |

As a result of selecting BIC/exhaustive procedure, we got the following mathematical Logistical Regression model:



The assumption that we make when using this model is that the data follows a linear pattern in log-odds (It’s monotone in probability). Basically, this means that if we draw a scatterplot, it will always be going down or up. An example of this can be seen in the January Min. Temperature Scatterplot below. We also assume independence, meaning that that one region being infested isn’t going to affect another region’s being infested.

We also assume this follows a Bernoulli distribution: **Yi ~ Bern(pi)**



Interpretation of variables and parameters:

**Yi:** Whether the region is infested or not

**log**: logarithmic transformation

**pi:** Probability of the region being infested

**pi/(1-pi):** Odds Ratio

**β0**: (This is the intercept) When each of the x’s (January min. temp., August max. temp., Precipitation, and Regional Location) are at their average, we would expect a region’s log-odds of being infested to be β0.

**xij**: Various measurements of variables related to being infested (January min. temp., August max. temp., Precipitation, and Regional Location)

**βj**: (This is one of the slopes) Holding all other x’s constant, as xij(January min. temp., August max. temp., Precipitation, and Regional Location) goes up by one, the region is exp{βj} times more likely to be infested.

**β1**: Holding all other x’s (August max. temp., Precipitation, and Regional Location) constant, as x1(January) goes up by one, a region is exp{-0.147} times more likely to be infested on average (0.86 times more likely on average).

**β4**: Holding all other x’s (January, August max. temp., Precipitation, and Regional Location) constant, as x4(NCYes) goes up by one, a region is exp{-1.218} times more likely to be infested on average (0.296 times more likely on average).

As stated previously, the assumption that we make when using this model is that the data follows a linear pattern in log-odds (It’s monotone in probability). We also assume independence, and that it follows a Bernoulli distribution: **Yi ~ Bern(pi)**

**Section 3: Results**

Coefficients Table:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variable | Coefficient | 95% Confidence Interval | 95% Multiple  Confidence Interval | 95% Percentage Confidence Interval |
| β0 (intercept) | -0.157 | (-1.89, 1.56) | (0.15x, 4.79x) | (-84.84%, 379.74%) |
| β1 January | -0.147 | (-0.19, -0.10) | (0.82, 0.90) | (-17.58, -9.55) |
| β2 August Max | -0.085 | (-0.13, -0.04) | (0.88, 0.96) | (-12.37, -3.76) |
| β3 Precipitation | 0.0029 | (0.002, 0.003) | (1.002, 1.003) | (0.21, 0.37) |
| β4  NCYes | -1.218 | (-1.51, -0.92) | (0.22, 0.39) | (-77.97, -60.30) |
| β5  SEYes | -0.919 | (-1.22, -0.62) | (0.29, 0.54) | (-70.42, -46.19) |

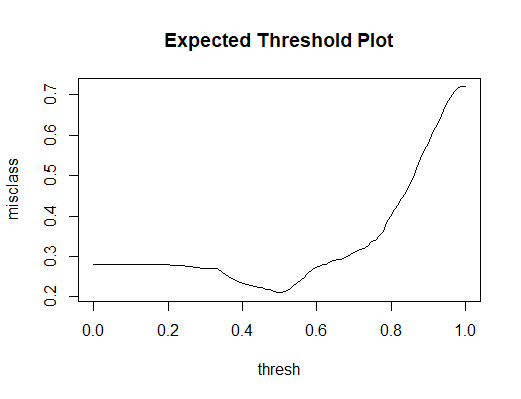
According to the model, our average region of the forest should be about 0.85 times (exp{-0.157}) likely to be infested, ignoring all other variables. So the good news is that the average region of the forest is more likely to NOT be infested than the other way around.

(For β1) Holding all other x’s (August max. temp., Precipitation, and Regional Location) constant, as x1(January) goes up by one, a region is exp{-0.147} times more likely to be infested on average (0.86 times more likely on average). I am 95% confident that the mean likelihood of a region being infested increases 0.82 to 0.90 times for each degree warmer it’s minimum temperature in January is.

(For β4) Holding all other x’s (January, August max. temp., Precipitation, and Regional Location) constant, as x4(NCYes) goes up by one, a region is exp{-1.218} times more likely to be infested on average (0.296 times more likely on average). I am 95% confident that the mean likelihood of a region being infested increases 0.22 to 0.39 times if it’s in the North Central Region.

An appropriate threshold for classification that minimizes the misclassification rate is 0.495. This is the value we’ll use as a cutoff to determine whether a region is infested or not.

The following is a plot demonstrating the threshold cutoff we should use after running the data through a computer program:



Non Cross-Validated Confusion Matrix:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Predicted Non-Infested | Predicted Infested | Total |
| True Non-Infested | 245 | 402 | 647 |
| True Infested | 82 | 1581 | 1663 |
| Total | 327 | 1983 | 2310 |

-Sensitivity (% of true positives): 1581/1663 = 0.951

-Specificity (% of true negatives): 245/647 = 0.379

-Positive Predictive Value (% Correctly Predicted Yes): 1581/1983 = 0.797

-Negative Predictive Value (% Correctly Predicted No): 245/327 = 0.749

After building a confusion matrix from all of the data, we calculated a sensitivity level of 95%,

a specificity level of 38%, a positive predictive value of 80%, and a negative predictive value of 75%. Judging from these values alone, it’s fair to say our model is able to correctly classify infested regions in the dataset fairly well. The model correctly predicts that a region is infested about 80% of the time, and correctly predicts that they *aren’t* infested about 75% of the time, both of which are very good. The sensitivity level of 95% also helps back up this claim extremely well. However, the specificity level of 38% may make us more hesitant.

The model appears to fit the data well for the most part. The low specificity levels can be forgiven due to the high levels in the sensitivity, positive predictive value, and negative predictive value categories.

The only hiccup is found in our pseudo r-squared score, which only came out to be 0.141. This means only 14.1% of variation in our “log-odds” [log(p/(1-p)] can be explained by the model that we’ve created.

Overall, I still think the model is worth using, despite the r-squared being so low.

Next, to test the predictive ability of our model, we ran a cross-validation study. This basically means we ran our model through a simulation hundreds of times after taking out part of the data and reinserting it. This verifies whether our previous assumptions of the model were true or not.

After running a cross-validation study, our model got the following mean values:

- average sensitivity: 0.9652 (about 97%)

- average specificity: 0.2921 (about 29%)

- average positive predictive value: 0.7797 (about 78%)

- average negative predictive value: 0.7660 (about 76%)

These aren’t far from the values that were calculated without the cross-validation, so the predictive ability of our model appears to be very good.

Using our model, we’ve helped ecologists predict pine beetle infestations for the next 10 years. For a South East region with a slope of 18.07 and elevation of 1901.95, we’ve created the table below. We used data collected regarding future temperatures and precipitation rates in the region to fit the model.

Infestation Likelihood Over the Next 10 Years:

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Year | 2018 | 2019 | 2020 | 2021 | 2022 | 2023 | 2024 | 2025 | 2026 | 2027 |
| Probability | 0.86 | 0.91 | 0.87 | 0.67 | 0.86 | 0.75 | 0.89 | 0.88 | 0.90 | 0.83 |
| Infested? | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |

According to our model, the region will most certainly become infested in the next 10 years. Each probability value we observed was above the 0.495 cutoff value that we set, so we can be fairly certain that the land will be infested.

The forest service should definitely concentrate on this area to prevent future infestation. Not only was each year a ‘yes’ when it came to likely being infested, but the South East region was one of two that was actually included in the bestglm model that we created. The forest service should probably look at the North Central region too if they have the time and budget.

**Section 4: Conclusions**

We found that the factors that help determine whether a forest region is infested with pine beetles the most are the average January minimum temperature, the average August maximum temperature, the mean annual precipitation levels, and whether the region is in the South East or North Central part of the forest. The model built off of these factors wasn’t extremely strong, but it had enough solid elements to be useful. By finding this relationship and exploring the data, we were able to create a mathematical model that would let us predict whether a forest region was infested or not using a logistical approach. The model’s predictions won’t always be accurate, but will give us a ballpark idea of what to expect.

1) The Colorado State Forest Service should gather more data. According to our math, only about 14% of the infestation occurred as a result of the variables in our model. If we really want to pinpoint how to stop the pine beetle, more factors should be taken into account.

2) Until more data can be gathered, the CSFS should use this data to combat the pine beetles in the South East and North Central regions of the forest. Those seem to be the most affected at the moment.

**R CODE**:

setwd("C:/Users/thety/Desktop/330/Final Exam")

beetle <- read.csv("PineBeetle2.csv", header = TRUE)

beetle$Infested<-ifelse(beetle$Infested=='Yes', 1,0)

head(beetle)

summary(beetle)

#1b

scatter.smooth(beetle$January,jitter(beetle$Infested,amount=0.3),main="Jitter Jan. Min. Temp. Scatterplot")

#strong negative correlation (infestation numbers go down as it gets warmer from the super cold)

scatter.smooth(beetle$August\_max,jitter(beetle$Infested,amount=0.3),main="Jitter Aug. Max. Temp. Scatterplot")

#strong negative correlation after 25 (infestation numbers go down after it gets too hot)

scatter.smooth(beetle$Slope,jitter(beetle$Infested,amount=0.3),main="Jitter Mtn. Slope Scatterplot")

#weak negative correlation (the steeper the slope, the less beetles infesting)

scatter.smooth(beetle$Elev,jitter(beetle$Infested,amount=0.3),main="Jitter Elevation Scatterplot")

#weak positive correlation (about the same infestation all the way through)

scatter.smooth(beetle$Precip,jitter(beetle$Infested,amount=0.3),main="Jitter Precipitation Scatterplot")

#strong positive correlation until 500 inches (more infestation until 500 inches)

plot(beetle$NC,jitter(beetle$Infested,amount=0.3),main="North Central Infestation") #mostly no

plot(beetle$NW,jitter(beetle$Infested,amount=0.3)) #mostly yes

plot(beetle$EC,jitter(beetle$Infested,amount=0.3)) #mostly no

plot(beetle$WC,jitter(beetle$Infested,amount=0.3),main="West Central Infestation") #mostly yes

plot(beetle$SE,jitter(beetle$Infested,amount=0.3)) #mostly no

plot(beetle$SC,jitter(beetle$Infested,amount=0.3)) #mostly yes

plot(beetle$SW,jitter(beetle$Infested,amount=0.3)) #even but mostly yes

#2a

#fortunately diabetes is already the last variable

library(bestglm)

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

vs.res$BestModels

vs.res$Subsets

plot(vs.res$Subsets$BIC,type="b",pch=19,xlab="number of variables",ylab="BIC")

best.lm <- vs.res$BestModel

summary(best.lm)

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

vs.res2$BestModels #true means its included, false means it isnt

vs.res2$Subsets #minimize AIC maximize Rsquared

plot(vs.res2$Subsets$AIC,type="b",pch=19,xlab="number of variables",ylab="AIC")

best.lm2 <- vs.res2$BestModel

summary(best.lm2)

#3a

coef(best.lm)

confint(best.lm,level=.95)

multiple.conf <- exp(confint(best.lm))

multiple.conf

percentage.conf <- 100\*(multiple.conf-1)

percentage.conf

exp(-0.147)

exp(-1.218)

exp(-0.157)

#3b

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

thresh <- seq(0,1,length=100)

misclass <- rep(NA, length=length(thresh))

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!= beetle$Infested)

}

#Find threshold which minimizes miclassification

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

cutty

plot(thresh,misclass,type="l",main="Expected Threshold Plot")

table(beetle$Infested,pred.probs>cutty) #this is a confusion matrix

conf.mat <- addmargins(table(beetle$Infested,pred.probs>cutty))

conf.mat

sens <- conf.mat[2,2]/conf.mat[2,3]

sens

spec <- conf.mat[1,1]/conf.mat[1,3]

spec

ppv <- conf.mat[2,2]/conf.mat[3,2]

ppv

npv <- conf.mat[1,1]/conf.mat[3,1]

npv

###find pseudo r-squared###

up <- best.lm$deviance

down <- best.lm$null.deviance

pseudorsqr <- (1-(up/down))

pseudorsqr

#3c

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

n.cv <- 500

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

## Set my threshold for classifying

cutoff <- cutty #the value we found earlier after running "thresh"

## 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){

## Separate into test and training sets

obs.test <- sample(1:nrow(beetle),round(.1\*nrow(beetle)))

test.set <- beetle[obs.test,] #putting a blank after a comma will give you all of teh columns

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

train.set

## Fit best model to training set

train.model <- glm(Infested~January+August\_max+Precip+NC+SE,data=train.set)

## 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(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]

}

##End for-loop

mean(sens)

mean(spec)

mean(ppv)

mean(npv)

#3d

d2018 <- data.frame(NC="No",SE="Yes",January=-13.98,August\_max=15.89,Precip=771.13)

pred.log.odds <- predict.glm(best.lm,newdata=d2018)

pred.log.odds

pred.prob <- exp(pred.log.odds) / (1+exp(pred.log.odds))

pred.prob

#0.8625

pred.prob.alt <- predict.glm(best.lm, newdata=d2018, type="response")

pred.prob.alt #same as before

d2019 <- data.frame(NC="No",SE="Yes",January=-17.80,August\_max=18.07,Precip=788.54)

pred.prob.alt <- predict.glm(best.lm, newdata=d2019, type="response")

pred.prob.alt

d2020 <- data.frame(NC="No",SE="Yes",January=-17.27,August\_max=16.74,Precip=677.63)

pred.prob.alt <- predict.glm(best.lm, newdata=d2020, type="response")

pred.prob.alt

d2021 <- data.frame(NC="No",SE="Yes",January=-12.52,August\_max=18.06,Precip=522.77)

pred.prob.alt <- predict.glm(best.lm, newdata=d2021, type="response")

pred.prob.alt

d2022 <- data.frame(NC="No",SE="Yes",January=-15.99,August\_max=18.23,Precip=732.32)

pred.prob.alt <- predict.glm(best.lm, newdata=d2022, type="response")

pred.prob.alt

d2023 <- data.frame(NC="No",SE="Yes",January=-11.97,August\_max=15.81,Precip=615.96)

pred.prob.alt <- predict.glm(best.lm, newdata=d2023, type="response")

pred.prob.alt

d2024 <- data.frame(NC="No",SE="Yes",January=-15.75,August\_max=16.85,Precip=805.90)

pred.prob.alt <- predict.glm(best.lm, newdata=d2024, type="response")

pred.prob.alt

d2025 <- data.frame(NC="No",SE="Yes",January=-16.19,August\_max=16.51,Precip=714.57)

pred.prob.alt <- predict.glm(best.lm, newdata=d2025, type="response")

pred.prob.alt

d2026 <- data.frame(NC="No",SE="Yes",January=-17.87,August\_max=17.84,Precip=740.50)

pred.prob.alt <- predict.glm(best.lm, newdata=d2026, type="response")

pred.prob.alt

d2027 <- data.frame(NC="No",SE="Yes",January=-12.44,August\_max=16.96,Precip=801.22)

pred.prob.alt <- predict.glm(best.lm, newdata=d2027, type="response")

pred.prob.alt