Jacob Andros

STAT 330, Section 1

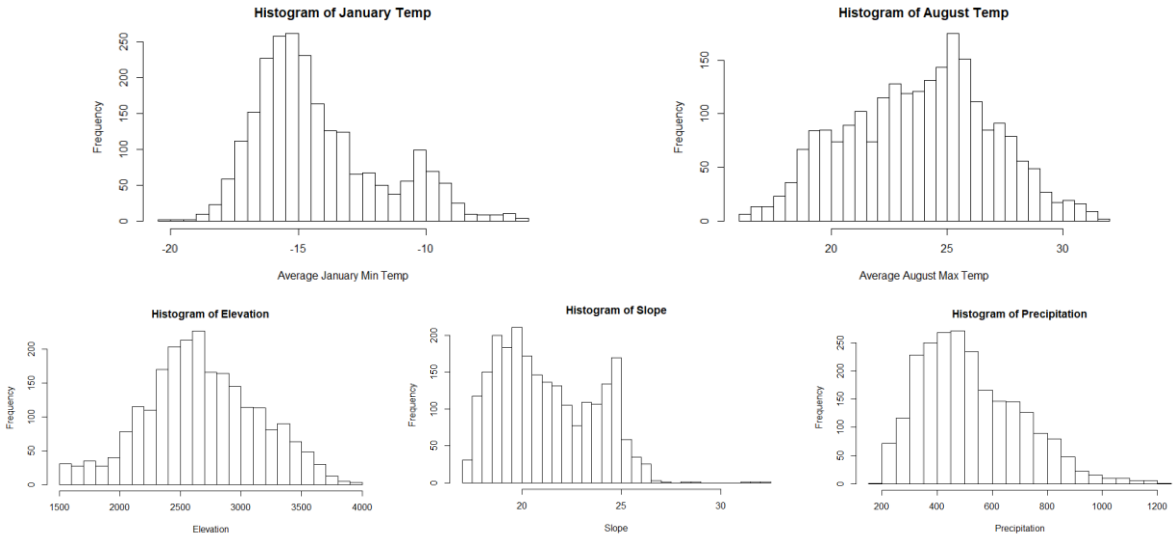
Exam 3 Analysis – Pine Beetles

1. Introduction and Problem Background

Due to climate change and various environmental factors, the presence of the pine beetle has become increasingly detrimental in coniferous forests in mountainous areas of the Western United States. In order to better protect and preserve these areas and the wildlife therein, it is important to better understand the nature of the relationship between environmental factors and pine beetle infestation. One goal of this study is inference - to better understand the relationship between infestations and these factors, such as temperature, elevation, precipitation, and more. The other goal is to produce a model that allows us to effectively predict whether or not an area will become infested in the coming years so that the forest service can concentrate their efforts on protecting those more vulnerable forests. Statistical modeling will give us probabilities of being infested for each forest, which will allow us to carry out this prediction process.

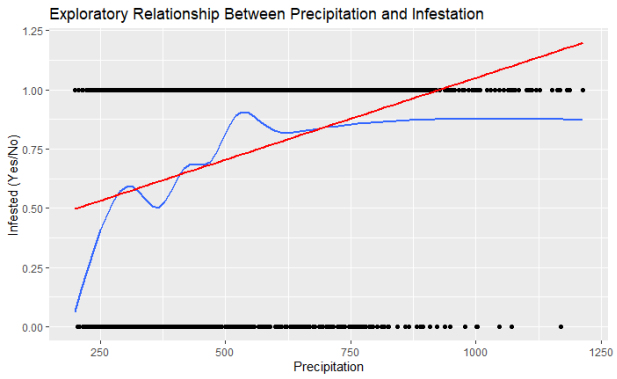
The dataset provided contains observations for over two thousand forests. For each forest, the dataset gives its average January minimum temperature, average August maximum temperature, elevation, slope (of the mountain), precipitation levels, and region. The "region" variable is categorical and is divided into seven columns – one for each region – with a corresponding "Yes" or "No" to indicate if the forest fell in that area. It is assumed that any observation with a "No" to all seven of these regions is located in another region not specified.

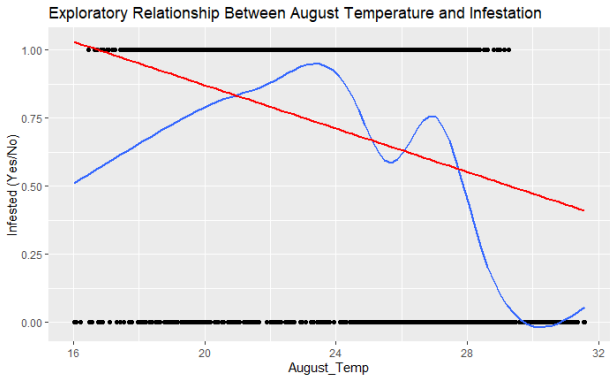
The histograms below show the distribution of the forests for each quantitative variable. None of them show any extreme skewness except for the slope variable.



After converting the yes/no variable for infestation status to a numeric variable (1 for yes and 0 for no), we can make rough plots of the relationships between this status and each quantitative variable. Some of the plots show a linear or at least a monotonic relationship (such as precipitation – it appears that

wetter areas are more likely to become infested), and some show no clear relationship (such as the August temperature). More of these plots are included later in the document.

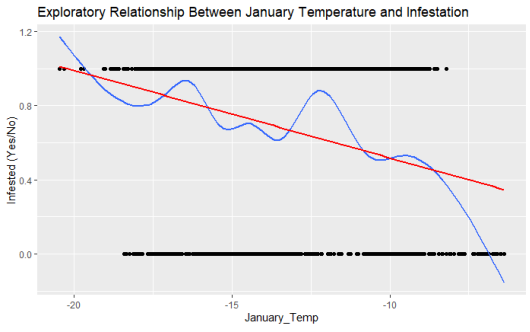


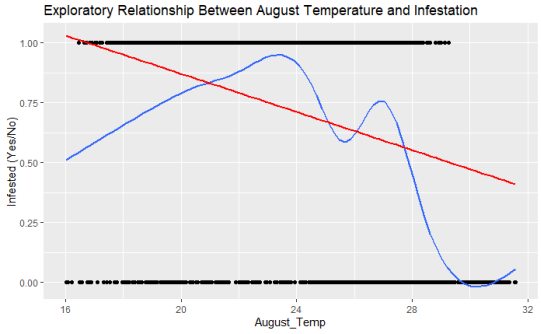


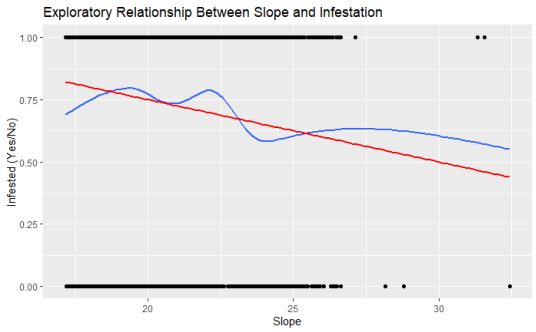
A logistic regression model will be used to accomplish the goals mentioned above. Logistic regression is a more correct tool for this situation because the response variable (the thing we want to predict – whether or not a forest will be infested) – is a dichotomous, categorical variable. This method of regression will yield for each forest a probability of being infested (that is, the dichotomous response variable being positive). Linear regression would not be able to accommodate this.

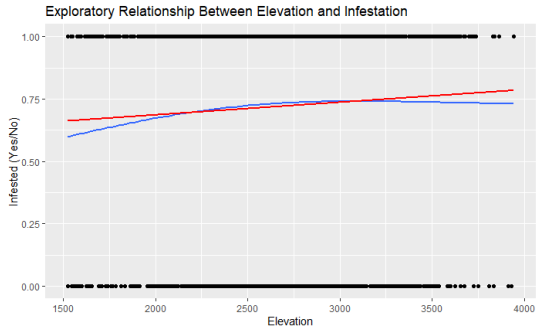
2. Statistical Modeling

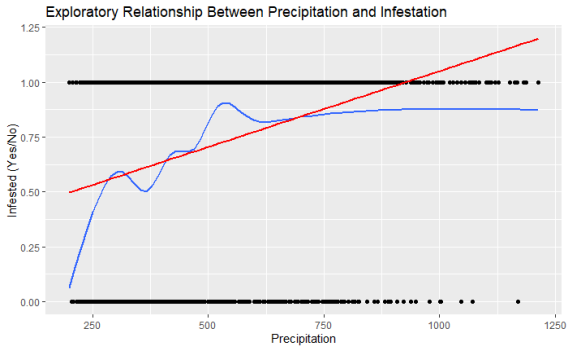
Before selecting the variables that will be used in the model, there are two assumptions that we need to justify. First, we need to be confident that the observations contained in this dataset are independent of each other – meaning that the infestation condition of one forest is not dependent on or connected with the infestation status of another forest in the dataset. Since we are not made aware of the details of the data collection methods used, we will assume that the researchers surveyed forests that were geographically far enough apart to be independent of each other. Secondly, there needs to be a somewhat linear, or at least a mostly monotonic relationship between the log-odds ratio and each explanatory variable that is to be included. The following five figures illustrate this for us (and two of these were already included above):











While some of these plots are somewhat borderline in terms of meeting the assumption, the only plot that raises concern in meeting the mostly monotonic trend is that of the August temperature. We could either remove that variable and select the next best fitting model, or we could proceed with the August variable in the model regardless of our concerns. Since the smooth curves shown above only examine one variable at a time and do not account for the effects of other variables in the process (they don't "hold all else constant"), we will assume that the assumption is still reasonably met.

The next step in building the model is to decide which variables are most important to include. R can do this for us, but we need to specify a few things to help it in reaching its decision. First, we must choose to minimize either Akaike information criterion (AIC) or Bayesian information criterion (BIC). Minimizing AIC leads to a more effective predictive model, and minimizing BIC leads to a tighter-fitting model for better inference (understanding the relationship). Inference and prediction are both important goals in this study, so either method would be justifiable here. Since BIC yields a much simpler model and the first request in the prompt involves the understanding of the relationship, we will use BIC-minimization.

Next, we need to decide what method of selection to use. We will use the best subsets method here because there are only 12 explanatory variables, even when the region variable is divided up over seven columns. The best subsets method works as well as any other method when there are few enough explanatory variables because it doesn't take more than a minute to run in R.

The variable selection under BIC and best subsets yields a model with January (average minimum) temperature, August (average maximum) temperature, annual precipitation, the north central region indicator, and the southeast region indicator as the most significant variables.

A mathematical expression for the model just described is:

$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 * \text{JanuaryTemp} + \beta_2 * \text{AugustTemp} + \beta_3 * \text{Precipitation} + \beta_4 * \text{I(North Central)} + \beta_5 * \text{I(Southeast)} \quad y_i \sim iid \text{ Bernoulli}(p_i)$

To interpret the meaning of these parameters, we can use precipitation and the southeast indicator as examples:

- β_3 represents the average increase in the log-odds ratio for each additional inch of annual precipitation, holding all other variables in the model constant. Since the results of this study will not be in terms of the log-odds ratio, we can instead say that holding all else constant, we expect a forest to become e^{β_3} times more likely to be infested for each additional inch of annual precipitation it receives.
- β_5 represents the average difference in the log-odds ratio for forests in the southeast region compared to forests outside of the southeast region, holding all else constant. Again, we can also say that holding all else constant, we expect a forest in the southeast region to be e^{β_5} times more likely to be infested than those outside of the southeast region.

3. Results

The table below shows estimates and 95% confidence intervals for each slope coefficient, as well as intervals for the exponentiated values that we can use to interpret each slope.

Variable	**Symbol**	**Estimate**	**Interval for β_i**	**Interval for e^{β_i}**	**Interval for $100*(e^{\beta_i} - 1)$**
Intercept	β_0	-0.16	(-1.89, 1.57)	(0.15, 4.80)	(-84.84, 379.75)
January Temp	β_1	-0.15	(-0.19, -0.10)	(0.82, 0.90)	(-17.58, -9.55)
August Temp	β_2	-0.85	(-0.13, -0.038)	(0.88, 0.96)	(-12.37, -3.76)
Precipitation	β_3	0.0029	(.0021, .0037)	(1.002, 1.004)	(0.21, 0.37)
I (North Central)	β_4	-1.22	(-1.51, -0.92)	(0.22, 0.4)	(-77.97, -60.30)
I (Southeast)	β_5	-0.92	(-1.22, -0.62)	(0.30, 0.54)	(-70.42, -46.19)

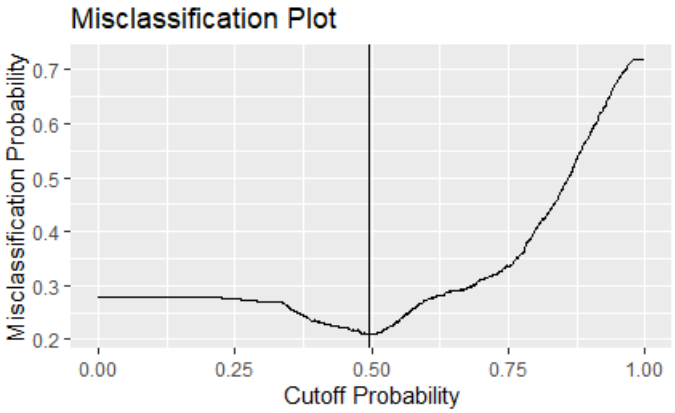
Using January temperature as an example, interpretations of these intervals are as follows:

- We are 95% confident that, holding all else constant, as a forest's minimum January temperature increases by 1 degree Celsius, that forest's log-odds ratio of being infested decreases by between 0.10 and 0.19.
- We are 95% confident that, holding all else constant, as a forest's minimum January temperature increases by 1 degree Celsius, that forest is between .82 and .90 times as likely to become infested.
- We are 95% confident that, holding all else constant, as a forest's minimum January temperature increases by 1 degree Celsius, that forest is between 9.55% and 17.58% less likely to become infested.

It is interesting to note that almost all the variables in the model (besides precipitation) have a negative slope coefficient estimate. That is to say, as any of those variables increases, the probability of the forest being infested decreases.

To assess the fit of the model, there are a few statistics we can examine. The first thing we will examine is the cutoff probability. In logistic regression, a probability of becoming infested is generated for each

forest based on its observed characteristics. The question is, how low does that probability need to be for us to safely say that the forest is not going to become infested? To do this, we create a misclassification plot, which is a plot that shows the misclassification rate (based on the data we have) for every possible cutoff probability. Naturally, we want that misclassification rate to be as low as possible.



The vertical line represents the cutoff probability that minimizes the misclassification of an infested forest as safe from infestation, which is at about 0.495, and results in a misclassification rate of just over 0.2. This means that if the model determines that a given forest has an infestation probability of 0.495 or higher, we will classify it as a forest that is going to become infested and needs attention. If a forest receives a probability of less than 0.495, we would classify it as being safe from infestation for the time being. Using this cutoff will minimize the misclassifying of infestation-vulnerable forests as being safe from infestation.

Using this cutoff probability, we can examine the confusion matrix and ROC curve to determine the fit of our model to all available data. The confusion matrix compares our classifications (infested or not infested) based on the 0.495 probability to the actual observed infestation status of each forest.

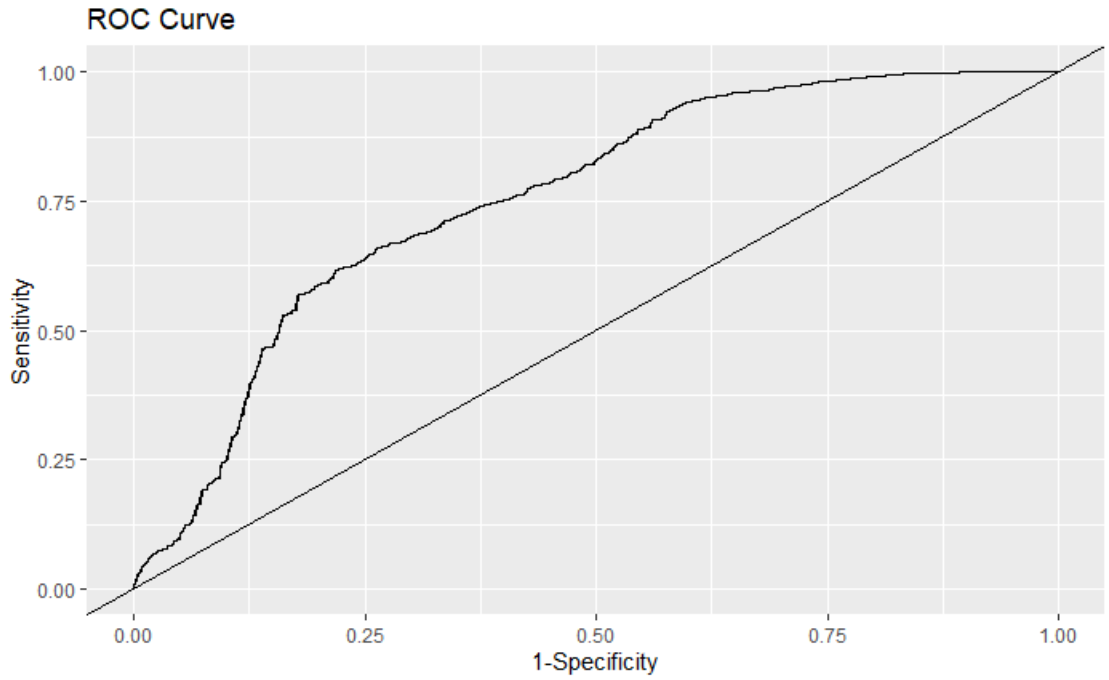
	Prediction: No	*Prediction: Yes*	*Total*
Not Infested (Observed)	246	401	647
Infested (Observed)	82	1581	1663
Total	328	1982	2310

There are four summary statistics that can be gathered from this confusion matrix (again, all of this using the cutoff probability of 0.495):

- The sensitivity of this model is 0.95, which is calculated as 1581 divided by 1663. In other words, this is the percentage of infested forests that were classified as infested in the dataset.
- The specificity is 0.38, which is calculated by doing 246 divided by 647. It is the percentage of non-infested forests that we classified as not being infested.
    - It should be noted from these two figures that the model is aggressive in classifying forests as vulnerable to infestation. It is much more likely to classify a safe forest as vulnerable than it is to classify a vulnerable forest as safe. It's an "over-approximation".

- The positive predictive value is 0.798, or 1581 divided by 1982. This is the percentage of forests classified by the model as infested that were in fact infested.
- The negative predictive value is 0.75, or 246 divided by 328. This is the percentage of forests classified by the model as safe that were in fact safe from infestation.

There are two other figures that should be mentioned in discussing the fit of the model. One of these is the pseudo-r-squared (in this case, about 0.14), which is supposed to be an approximation of the percentage of variation in the log-odds ratio for infestation explained by the explanatory variables fitted in the model. However, this pseudo-r-squared is often an underapproximation of the true fit, so despite 14% being very low, it is likely that our model accounts for more than just 14% of the variation in the log-odds ratio. The other important figure is AUC, or the area under the ROC curve, which is 0.7549 in this scenario. The ROC curve for the full dataset is pictured below:



The AUC is a measure that typically falls between 0.5 and 1. If the AUC is closer to 0.5, that implies the model is not predicting much more effectively than a basic "coin-flip" prediction system would work (guessing). The closer the AUC is to 1, the more effective the model is in predicting based on our dataset. The straight line above represents the coin-flip system, and the curve above it represents the model we are using. Since the AUC is about 0.75, we can conclude that although far from perfect, it is predicting the vulnerability of these forests much more correctly than guessing would.

In summary, while the fit of this model could be better, it is clearly fitting the data better than most other models would, and the model allows us to identify vulnerable forests approximately 95% of the time.

To validate the model fit statistics from above and also assess the predictive capability of the model, we can use R to run a cross-validation study, where we repeatedly remove a small piece of the dataset and test our predictions on those observations using the remainder of our dataset. The cross-validation study yields the following figures:

Sensitivity	0.943
Specificity	0.304
Positive Predictive Value	0.778
Negative Predictive Value	0.677
Area Under the Curve (AUC)	0.749

Again, these figures con be interpreted as follows, with the word "classify" being synonymous with "predict":

- We will classify approximately 94% of infested forests as being infested (so we only classify about 6% of infested forests as being clean).
- We will classify approximately 30% of clean forests as being infested (so the downside of the model is that it incorrectly labels 70% of the clean forests as infested).
- Approximately 78% of forests that we classify as infested will actually be infested.
- Approximately 68% of forests that we classify as clean will actually be free of infestation.
- Since the AUC is well above 0.5, the model is predicting much more effectively than we would on our own.

The fact that these five values for the cross-validation are all quite close to their correspondents in the confusion matrix for the whole data is a good sign. With the exception of specificity, it can be said from the cross-validation study that this model has excellent predictive capacity.

The model can be used to predict yearly infestation probabilities for this forest in the southeast region that ecologists are examining. Using their predicted measurements of January/August temperatures and annual precipitation, our model shows that this forest will consistently have infestation probabilities of 0.67 and higher for the next ten years (see table below). Since these probabilities are all well above the cutoff probability of 0.495, we predict that this forest will become infested if nothing else is done to help it. For that reason, it is recommended that the forest service concentrate their efforts on this forest (and others similar to it) to prevent infestation.

Year	**Probability of Infestation**	**Year**	**Probability of Infestation**
2018	0.86	2023	0.75
2019	0.91	2024	0.89
2020	0.88	2025	0.87
2021	0.67	2026	0.90
2022	0.86	2027	0.83

4. Conclusions

The logistic regression model discussed in this document reveals that the most important quantitative variables in determining whether a forest will be infested are the January average minimum temperature, the August average maximum temperature, and the annual precipitation. It also indicates that the North Central and Southeast regions' forests are less vulnerable to infestation, on average. If we predict any forest with an infestation probability of at least 0.495 to be a vulnerable forest, then we will be able to correctly identify around 94% of those forests, but at the expense of worrying about many forests that are not actually vulnerable.

The next step for the forest service should be to assess which forests are likely to become infested in the next few years and deploy measures to fight the pine beetle infestations in those areas. This may include cutting down trees so that there is more space between trees, or using pesticides and other chemicals to combat the pine beetle population. Meanwhile, ecologists should continue searching for variables that might be useful to add to the model, like humidity or average tree height, so that another logistic regression can be carried out with improved fit and accuracy. Lastly, statisticians should further examine the monotonicity assumption of the August temperature variable and determine if that variable needs to be removed from the model or not.