

# **Eradicating the Asian Longhorned Beetle in South Carolina**

## **Researcher:**

Abigail Ratcliff - Forestry and Environmental Resources Department

## **Consultants:**

Ashley Ko

Joshua Burrows

## **1. Introduction**

### **1.1 Background Information**

The Asian Longhorned Beetle (ALB) is an invasive pest that was first detected in the United States in 1996. Adult ALB lay eggs in hardwood trees and, after the eggs hatch, the juvenile beetles feed on the surround tree. Once a beetle reaches maturity, it bores a hole through the wood and exits the tree. An ALB infestation often results in the death of the host tree.

The ALB poses a risk to North American forests. Conservationists want to control infested zones and eliminate the beetle. Pesticides are costly and ineffective against the ALB because the larvae live deep within trees. Eradication is usually achieved using heavy machinery to cut down the trees and chop them into small pieces. This approach has been successful at eradicating the ALB in parts of the Northern United States.

Recently, the ALB was detected in South Carolina swamps and bottomland. Heavy machinery cannot be used in these areas because of the terrain. Therefore, other methods of eradication need to be explored. The ALB is not known to infest dead trees, so killing infested trees might kill the beetles. The goal of this study is to test the effectiveness of three potential eradication methods, each of which involves killing the infested tree. These three methods are:

- Herbicide (Hack & Squirt or H&S)
- Tree cut down and left whole on the ground (Cut & Leave or C&L)
- Tree cut down, chopped into logs, and left on the ground (Cut, Section & Leave or CS&L)
- Tree cut down, chopped into logs, and left on the ground (Cut, Section & Leave or CS&L)

### **1.2 Research Questions**

This study focuses on the following research questions:

- Are the eradication methods effective?
- Is one method more effective than the others?
- Does the size of the tree affect the efficacy of the eradication methods?

### 1.3 Statistical Questions

To answer the research questions, several statistical questions need to be addressed. These statistical questions are below. The section in which each question is addressed is given in parentheses.

- How is the response variable distributed? (Section 3)
- Are there any outliers? If there are, what if anything should be done about them? (Section 3)
- What type of model is appropriate for the ALB data? (Sections 4 and 7)
- What is the justification for using the chosen model? (Sections 4 and 7)
- Using statistical inference based on the model, what can be said in answer to the research questions? (Sections 5 and 6)

## 2. Data

### 2.1 Study Design

To investigate the research questions, Abigail Ratcliff conducted a study using 48 red maple trees located in a South Carolina bottomland site. Red maple was selected as it is a hardwood species. Individual trees were selected by the existence egg sites on the tree and if the tree had 1 - 10 ALB exit holes. Due to practical limitations such as funding, it was not possible to use a larger sample.

The trees were split into size categories based on diameter at breast height (DBH). Breast height is defined as approximately 4.5 feet from the ground. Trees with DBH measurements between 2 and 8 inches were categorized as small, trees with measurements between 8.1 and 12 inches were categorized as medium, and trees that measured between 12.1 and 20 inches were categorized as large.

The trees were randomly assigned to one of four treatment groups (H&S, C&L, CS&L, or control). Randomization was done within size groupings.<sup>1</sup> Furthermore, within each size-treatment group, the trees were randomized to one of four harvest dates: three-, six-, nine-, or twelve-months after treatment. The trees were treated in May 2021. Harvests for data collection were carried out the following August, November, February, or May.

It is worth noting that the experiment is balanced. There are 16 trees for each level of Size, 12 trees for each level of Treatment, and 12 trees for each level of Harvest Cycle. There is one tree for each Size/Treatment/Harvest Cycle combination.

### 2.2 Variables

This section discusses the variables used in the study.

The response variable is Survivors. Survivors is a count of all the living ALB detected at the time of data collection. There were three ways of detecting a living ALB. The first was to find a larva and the second was to find a pupa.<sup>2</sup> The third way was to find an exit hole that was created by a mature ALB

---

<sup>1</sup> One of the random assignments had to be changed. This will be discussed in the limitations section below.

<sup>2</sup> It is possible that not all the larvae were ALB larvae. They were sent for genetic testing to confirm, but the genetic results have not been received yet. This is discussed in greater detail in section 2.3.

emerging from the tree. If the exit hole was made after the treatments were applied, it should be counted as an ALB that survived treatment.

Since there were three ways to detect a living ALB, the response variable Survivors is the sum of three other variables: Candidate Larvae, Pupae, and Exit New. More information about these variables is provided in the following table:

**Table 1:** Dependent Variables

Variable Name	Description
Candidate Larvae	A count of the larvae found at harvest that could potentially be ALB larvae
Pupae	A count of the ALB pupae found at harvest
Exit New	A count of ALB exit holes that were made after treatment
Survivors (response variable)	The sum of Candidate Larvae, Pupae, and Exit New. Exit holes are made by adult ALB that are past the larvae and pupae stages, so Survivors does not double count any beetles.

There are three predictor variables: Treatment, Size, and Harvest Cycle. More information about these variables is provided in the following table:

**Table 2:** Independent Variables

Predictor	Description
Treatment	A factor representing the treatment applied. Levels: control, H&S, C&L, CS&L
Size	A factor representing the size of the tree. Levels: small, medium, large
Harvest Cycle	A factor representing the month of data collection. Levels: August, November, February, May

## 2.3 Limitations

Four limitations associated with the study design should be noted. First, it is possible that not all the Candidate Larvae are ALB larvae as there are other species which look similar. The collected larvae have been sent for genetic analysis to determine their species, but the results are still outstanding. For logistic reasons, analysis had to proceed without the pending genetic results, and the variable Candidate Larvae will be used.

Second, a tree that was originally randomized to control (tree 19) was switched with one of the trees that was originally randomized to H&S (tree 16). This switch was made due to concerns over the

proximity of tree 16 to another control tree and possible root sharing. Root sharing could result in the transfer of herbicide from a H&S tree to the control tree and thus, the switch was made. Strictly speaking, because of this switch, the assumption of randomization is not met. However, if the trees had not been switched, the assumption of independent errors would have been dubious. Since it was the one or the other, we feel that the decision to make the switch was reasonable. Given that only one switch was made, we believe that the impact on the experiment was minimal.

Third, several trees received extra herbicide because the initial treatment did not kill them. In May 2021, each H&S tree received 0.69 milliliters (mL) of herbicide per inch of DBH. In August, it was discovered that seven of the H&S trees did not die. These trees were re-treated with a second round of herbicide. The second dose was 0.69 mL of herbicide per inch of DBH. Despite the re-treatment, two H&S trees were still partially alive at the time of harvest. It is possible that the re-treatment of trees introduced additional variability.

Fourth, due to practical limitations, it was not possible to take a random sample of all the infested trees in the South Carolina quarantine site. Therefore, this study does not make claims about that population. The statistical tests in this study assess the level of compatibility between a given null model and the sample data, but further claims about a broader population are not made. Despite this limitation, when considered in conjunction with other information about the ALB, the results of this study could still be useful.

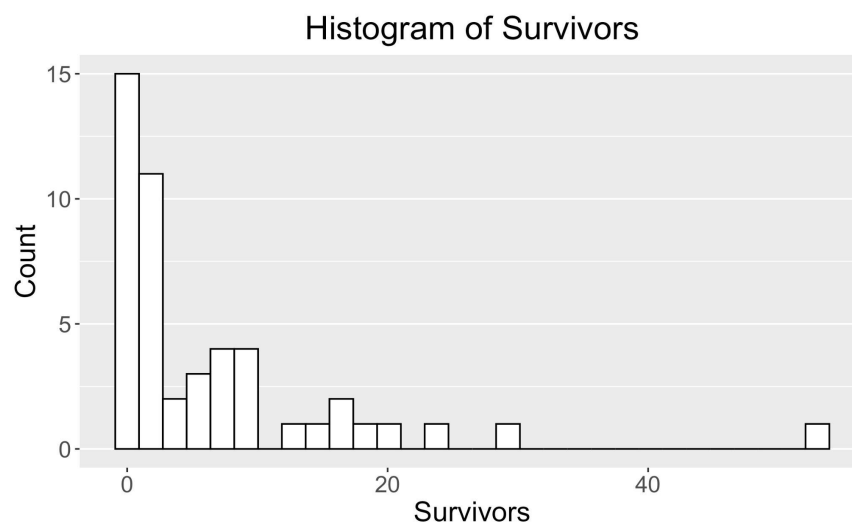
### 3. Exploratory Data Analysis

This section covers the data exploration process. One important observation is that Survivors is not normally distributed. Another is that Survivors is often zero.

#### 3.1 The Distribution of Survivors

The following histogram of Survivors is right skewed with the highest concentration of counts at zero. This indicates that distribution of the variable Survivors is not normally distributed.

**Figure 1:**

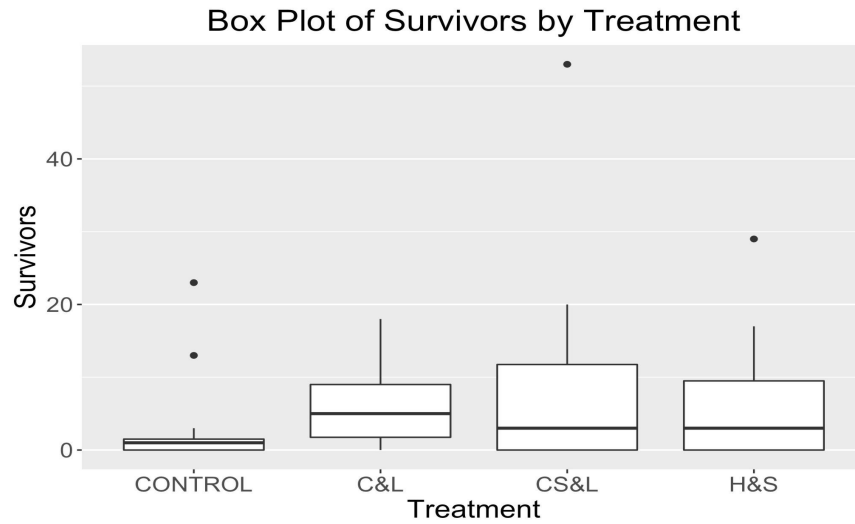


### 3.2 The Distribution of Survivors by Each Predictor

This section examines the center and spread of Survivors by each predictor variable.

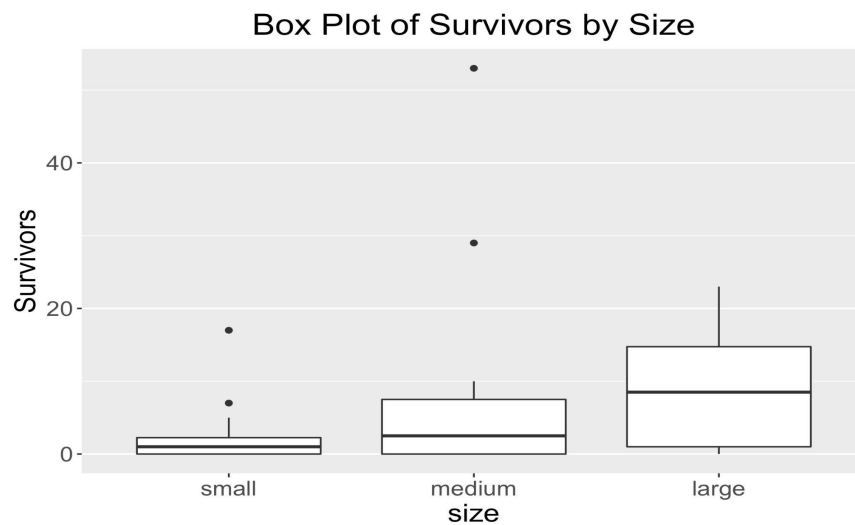
Surprisingly, the control trees had relatively few Survivors. Also, there is not a striking difference between the three treatments. It is worth noting that one tree had over 50 survivors. That tree is clearly an outlier, and it will be discussed in more detail in section 3.3.

**Figure 2:**



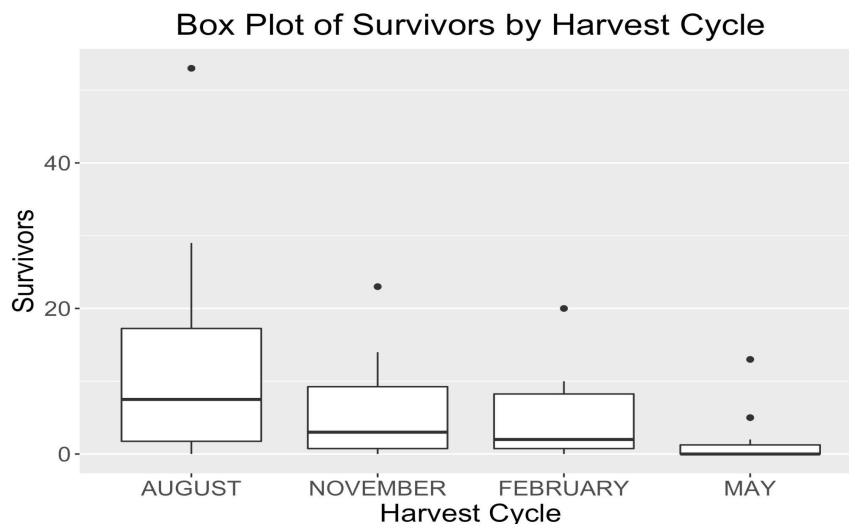
The median and interquartile range of Survivors both increase as Size increases. There appears to be a relationship between Size and Survivors.

**Figure 3:**



There is a distinct decline in the number of survivors with each subsequent harvest cycle. Survivors is highest in August and lowest in May. This pattern is consistent with the life cycle of the ALB. Because August is mid-season for the ALB, it is unsurprising that many ALB were found during that month. Adult ALB emerge from their trees in May, so very few ALB remain in trees in May.<sup>3</sup> Overall, there appears to be a relationship between Harvest Cycle and Survivors.

**Figure 4:**



### 3.3 Tree 47

Tree 47 is an outlier. It had 53 Survivors, which is 24 more than the next highest survivor count. Even though it is an outlier, we do not know of a compelling reason to exclude tree 47 from the study. As far as we know, nothing out of the ordinary happened to it.

It is worth noting that tree 47 was harvested in August, and trees harvested in August tended to have more Survivors. Also, it is a medium CS&L tree. If it was a small control tree, such a high number of Survivors would be difficult to square with the rest of the data. In a medium CS&L tree, however, it is not as surprising.

The table below is arranged in descending order by Survivors. It illustrates tree 47 relative to some of the other trees.

---

<sup>3</sup> Since Exit New is included in Survivors, shouldn't adults that emerged still be counted? If so, why would Survivors be lower in May? There are two possible explanations. The first is that not all ALB survive to adulthood, so Survivors could be higher for months when there are more larvae and pupae. The second is that the tree bark might have been damaged before data collection, which would reduce the count of new exit sites. These explanations are not competing. They could both be contributing factors.

**Table 3:** The Top Ten Observations by Survivors

Trees With Highest Survivor Count

ID	Treatment	Size	Harvest Cycle	Survivors
47	CS&L	medium	AUGUST	53
36	H&S	medium	AUGUST	29
12	CONTROL	large	NOVEMBER	23
4	CS&L	large	FEBRUARY	20
29	C&L	large	AUGUST	18
25	CS&L	small	AUGUST	17
37	H&S	large	AUGUST	17
40	H&S	large	NOVEMBER	14
51	CONTROL	large	MAY	13
1	CS&L	medium	FEBRUARY	10

### 3.4 Trees with Zero Survivors

Fifteen trees had zero Survivors. The following tables (Tables 4, 5, and 6) count the number of zeroes by each predictor variable. It is striking that seven of the twelve trees harvested in May had zero survivors.

**Table 4:**

Survivor Counts of Zero by Treatment

Treatment	Number of Zeros
CONTROL	5
C&L	2
CS&L	4
H&S	4

**Table 5:**

Survivor Counts of Zero by Size

Size	Number of Zeros
small	7
medium	5
large	3

**Table 6:**

Survivor Counts of Zero by Harvest Cycle

Harvest Cycle	Number of Zeros
AUGUST	2
NOVEMBER	3
FEBRUARY	3
MAY	7

## 4. Methodology

A zero inflated negative binomial (ZINB) model is used to analyze the ALB data. The ZINB model was selected from a field of five candidate models. The model selection process is discussed in detail in the appendix.

### 4.1 Response and Predictor Variables

The response variable is Survivors, which is the sum of Candidate Larvae, Pupae, and Exit New. Survivors is non-normal, and it is often zero. There are three predictor variables: Treatment, Size, and Harvest Cycle.

### 4.2 Model Specification

The idea behind the ZINB model is that there are two separate processes at work. First, there is a process that generates excess zeros. Then, there is a process that generates all the other observations. The ZINB model uses logistic regression to model the first process. Then, observations that are not excess zeros are modeled using a negative binomial model. In a sense, then, a ZINB model is two models put together. The ZINB model is appropriate for the ALB data because it does not assume a normal response, it accounts for the large number of zeros in the data, and it performs better than the alternatives based on model diagnostics.

In the logistic regression half of the ZINB model, Harvest Cycle is the only predictor variable. Seven out of the twelve trees harvested in May had zero Survivors. There are 15 zeros total, so 47% of the zeros



occurred in May. This is consistent with the life cycle of the ALB, so it is reasonable to model excess zeros based on Harvest Cycle. In the negative binomial half of the model, the predictors are Treatment, Size, and Treatment by Size interaction.

The model consists of four equations:

**Equation 1:**

$$P(Y = 0) = \psi_m + (1 - \psi_m)g(0|\mu_{ij}, k)$$

**Equation 2:**

$$P(Y > 0) = 1 - P(Y = 0) = (1 - \psi_m)g(y|\mu_{ij}, k)$$

**Equation 3:**

$$\log\left(\frac{\psi_m}{1 - \psi_m}\right) = \beta_0 + \tau_m$$

**Equation 4:**

$$\log(\mu_{ij}) = \beta_1 + \alpha_i + \gamma_j + (\alpha\gamma)_{ij}$$

where

$\psi_m$  is the probability of an inflated zero

$g(\mu_{ij}, k)$  is the negative binomial probability function

$y > 0$

$\tau_m$  is the effect of the  $m^{th}$  level of Harvest Cycle

$\alpha_i$  is the effect of the  $i^{th}$  level of Treatment

$\gamma_j$  is the effect of the  $j^{th}$  level of Size

$(\alpha\gamma)_{ij}$  is the interaction effect between Treatment and Size

$\beta_0, \beta_1$  are intercept terms

$m = 1, 2, 3, 4$

$i = 1, 2, 3, 4$

$j = 1, 2, 3$

$\mu_{ij} = E(\text{Survivors}|\alpha_i, \gamma_j)$ , when Survivors is not an excess zero

**Survivors**  $\sim$  **NB**( $\mu_{ij}, k$ ), when Survivors is not an excess zero

$k$  is a dispersion parameter

Equation 1 describes the probability of observing zero survivors. The probability of observing zero survivors is the sum of the probability of an excess zero plus the probability of a “non-excess” zero. A non-excess zero is a zero generated by the negative binomial process.

Equation 2 describes the probability of observing more than zero survivors. The probability of observing more than zero survivors is one minus the probability of observing zero survivors. However, it

can also be understood as the probability that the observation is not an excess zero times the probability that the negative binomial process generates a positive number.

Equation 3 models the probability of an excess zero. More specifically, the log-odds of an excess zero are understood to depend on Harvest Cycle.

Equation 4 models the mean of the negative binomial distribution. More specifically, the log of the mean is understood to be a linear function of Treatment, Size, and Treatment by Size interaction.  $k$  is a parameter that describes the dispersion of the negative binomial distribution.

### **4.3 Analytical Method**

The results of this study are based on p-values. P-values are used to assess the degree to which observed data is compatible with a null model. Typically, the null model “postulates the absence of an effect”.<sup>4</sup> For example, our most important goal is to assess the effectiveness of the treatment methods. An appropriate null model, then, is a model that does not include the effect of Treatment. The lower the p-value, the less compatible the data is with the null model.

We provide p-values for Type 3 tests. Type 3 tests are “partial” tests, which means that the null model they test contains all the factors except the one under consideration. For example, the p-value for Treatment tests the null hypothesis that the negative binomial portion of the model contains only Size, Treatment by Size interaction, and an intercept term.

Additionally, we use Tukey’s method to compare different factor levels. The goal is to determine whether the mean for one factor level is significantly different from the mean for another. For example, Tukey’s method can be used to determine whether the mean of Survivors in H&S trees is significantly different from the mean of Survivors in C&L trees. In other words, Tukey’s method can help determine whether there is evidence that one treatment method was more effective than the others. For Tukey’s method, an alpha level of 0.05 is used.

### **4.4 Statistical Software**

R is used for Exploratory Data Analysis. SAS is used for modeling.

## **5. Results**

### **5.1 Hypothesis Tests**

The observed p-value for Harvest Cycle is 0.0306. Harvest Cycle is the only factor in the logistic regression portion of the ZINB model, so this p-value tests the null hypothesis that the excess zeros were generated by a model with no factors in it. In other words, a model that only contains an intercept term. Based on the p-value of 0.0306, it is reasonable to conclude that the observed data is incompatible with this null model. A reasonable alternative, based on the life cycle of the ALB, is to include Harvest Cycle in the model.

---

<sup>4</sup> Wasserstein, R. L., & Lazar, N. A. (2016), “The ASA Statement on p-Values: Context, Process, and Purpose” *The American Statistician*, 70(2), 129-133.

The observed p-value for Treatment is 0.0098. This p-value tests the null hypothesis that the negative binomial portion of the model contains only Size, Treatment by Size interaction, and an intercept term. Based on the p-value of 0.0098, it is reasonable to conclude that the observed data is incompatible with this null model. A reasonable alternative is to include Treatment in the model.

The observed p-value for Size is 0.0074. This p-value tests the null hypothesis that the negative binomial portion of the model contains only Treatment, Treatment by Size interaction, and an intercept term. Based on the p-value of 0.0074, it is reasonable to conclude that the observed data is incompatible with this null model. Bigger trees have more wood for the beetles to infest, so a reasonable alternative is to include Size in the model.

The observed p-value for Treatment by Size interaction is 0.1234. This p-value tests the null hypothesis that the negative binomial portion of the model contains only Treatment, Size, and an intercept term. Based on the p-value of 0.1234, there is no motivation to reject the null model. In other words, there is no motivation to think that Treatment and Size interact.

The SAS output for these tests is below. P-values can be found in the column labeled 'Pr > ChiSq'.

**Table 7:**

LR Statistics For Type 3 Analysis			
Source	DF	Chi-Square	Pr > ChiSq
Treatment	3	11.40	0.0098
Size	2	9.82	0.0074
Treatment*Size	6	10.03	0.1234

LR Statistics For Type 3 Analysis of Zero Inflation Model			
Source	DF	Chi-Square	Pr > ChiSq
Harvest_Cycle	3	8.90	0.0306

## 5.2 Tukey's Honestly Significant Differences

Since it is reasonable to think that both Treatment and Size have a relationship with Survivors, it is natural to ask what that relationship is. The answer is a little bit disappointing.

To begin with Treatment, according to Tukey's method, the control trees have significantly fewer Survivors than the rest of the trees. However, the means of C&L, CS&L, and H&S trees are not significantly different from each other. In other words, the control trees are significantly different from the treated trees, but the three treatments may very well all have the same effect.

We did not expect the control trees to have significantly fewer Survivors than the rest of the trees. The hope was that treatment would reduce Survivors, but there is evidence that it had the opposite effect. It is possible that the reason why there are fewer Survivors in the control trees is because Survivors includes Candidate Larvae, and there is no guarantee that all the Candidate Larvae are ALB. If other insects are being mistaken for ALB, then they could be throwing off the numbers, especially if they are more likely than ALB to infest dead trees. This issue should be cleared up when the genetic results are received.

Tukey's method finds a significant difference between the mean of Survivors in small trees and the mean of survivors in large trees. However, it does not find a significant difference between medium trees and small trees, and it does not find a significant difference between medium trees and large trees. These results are unsurprising. Bigger trees have more wood for insects to infest, so you would expect them to have more Survivors.

SAS outputs with estimates of the means and Tukey's comparisons are below. To understand the outputs, it is important to know that the ZINB model does not model the mean directly. Instead, it models the log transformation of the mean. To get estimates of the means, the log transformation needs to be reversed. In tables 8 and 9, estimates for the transformed means can be found in the 'Estimate' column, and estimates for the means themselves can be found in the 'Mean' column.

Figures 5 and 6 present the results of Tukey's method in chart format. A bar covers the factor levels whose means are not significantly different from each other. In figure 5, H&S, C&L, and CS&L are all covered by one bar because their means are not significantly different from each other. Control is covered by a separate bar because the mean for control is significantly different from the others. In figure 6, one bar covers small and medium, and another bar covers medium and large. This signifies that the means for small and medium are not significantly different, and the means for medium and large are not significantly different, but the means for small and large are significantly different.

Output for Harvest Cycle is not included because the nature of the ZINB model does not allow Tukey's method to be used for Harvest Cycle.

**Table 8:**

Treatment Least Squares Means						
Treatment	Estimate	Standard Error	z Value	Pr >  z	Mean	Standard Error of Mean
C&L	1.7747	0.3287	5.40	<.0001	5.8984	1.9389
CONTROL	0.3527	0.4741	0.74	0.4569	1.4229	0.6746
CS&L	2.3558	0.3854	6.11	<.0001	10.5465	4.0645
H&S	1.8717	0.3625	5.16	<.0001	6.4996	2.3561

Figure 5:

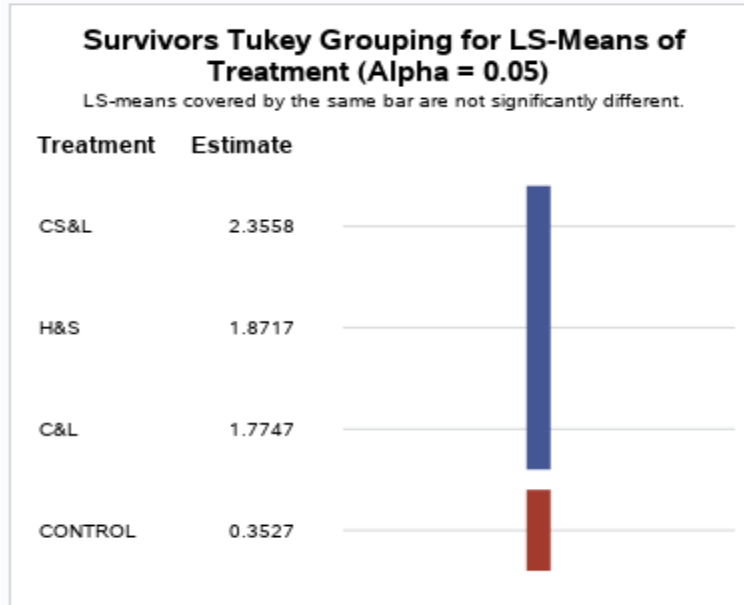
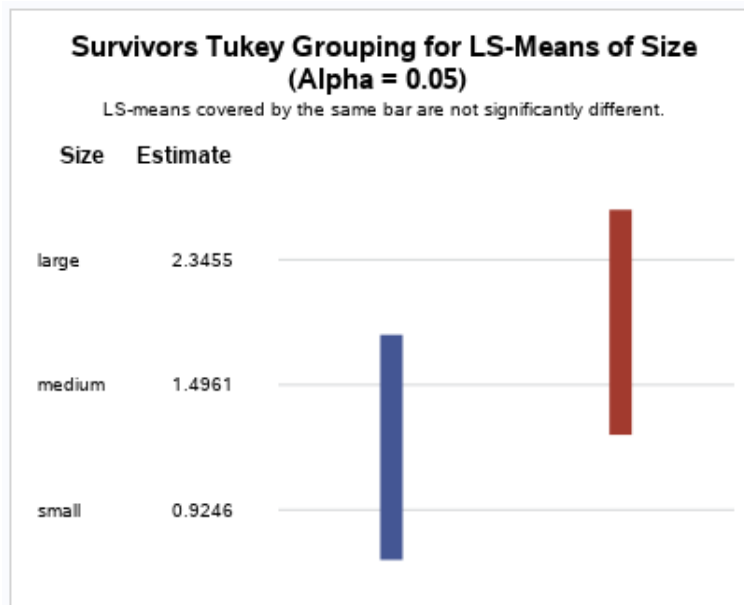


Table 9:

Size Least Squares Means						
Size	Estimate	Standard Error	z Value	Pr >  z	Mean	Standard Error of Mean
large	2.3455	0.2779	8.44	<.0001	10.4385	2.9006
medium	1.4961	0.3693	4.05	<.0001	4.4642	1.6484
small	0.9246	0.3830	2.41	0.0158	2.5208	0.9654

Figure 6:



## 6. Conclusion

### 6.1 Addressing the Research Questions

In conclusion, we will return to the research questions.

*Were the eradication methods effective?*

The study found no evidence that the eradication methods were effective. In fact, there is reason to think that the treatments increased the number of Survivors.

*Was one method more effective than the others?*

The study found no evidence that one eradication method was more effective than the others. For one thing, the study found no evidence that any of the methods were effective. For another, although there is evidence that control trees had fewer Survivors than the rest of the trees, there is no evidence of differences among the C&L, CS&L, and H&S trees.

*Did the size of the tree affect the efficacy of the eradication methods?*

The study found no evidence of a relationship between the size of the tree and the efficacy of the eradication methods. There was evidence of a relationship between Size and the number of Survivors, but there was no evidence of an interaction between Size and Treatment.

### 6.2 Limitations

It is worth reiterating the limitations mentioned in section 2.3. First, it is possible that the reason why there are fewer Survivors in the control trees is because Survivors includes Candidate Larvae, and there is no guarantee that all the Candidate Larvae are ALB. Second, one of the trees that was originally randomized to control was switched with one of the trees that was originally randomized to H&S. Therefore, strictly speaking, the assumption of randomization is not met. Third, some of the H&S trees were treated twice because the first round of treatment did not kill them. This could have introduced additional variability. Fourth, the statistical tests in this study assess the level of compatibility between a given null model and the sample data. Further claims about a broader population are not made.

### 6.3 Future Directions

Once the genetic data for the Candidate Larvae is received, it is possible that the results of this study will change. However, if the results do not change, then alternative methods for eradicating the ALB may need to be investigated. The three methods examined by this study do not look promising.

## 7. Appendix: Model Selection

The zero inflated negative binomial model was selected from a field of five different candidate models. The candidate models were chosen because they are appropriate for a response variable that is a non-normal count. The other four candidates were: square root transformation, log transformation, Poisson with overdispersion, and negative binomial. This section discusses our reasons for preferring the zero inflated negative binomial model over the other candidates.

Every model other than the zero inflated negative binomial model includes Treatment, Size, Harvest Cycle, and all two-way interactions. The three-way interaction was not included because, given the sample size, there were not enough degrees of freedom to include it.

### 7.1 Square Root Transformation

This model applies a square root transformation to Survivors in the hope that the result will be approximately normal. A standard ANOVA analysis is then performed on the square root of Survivors.

The diagnostics for this model suggest that it is not a good fit for the ALB data. The following issues can be observed in the diagnostic plots:

- Even after the transformation, the response does not appear normal. The histogram of square root transformed Survivors makes the non-normality clear.
- A straight line of points is visible in the lower left quadrant of the Residuals vs. Predicted plot.
  - These points are the trees for which Survivors is zero.
  - In a Residuals vs. Predicted plot, a random scatter of points indicates a good model fit. Patterns such as a straight line indicate a poor model fit.
  - Due to this line, the plot exhibits increasing variance, so the assumption of homoscedasticity is dubious.

Figure 7:

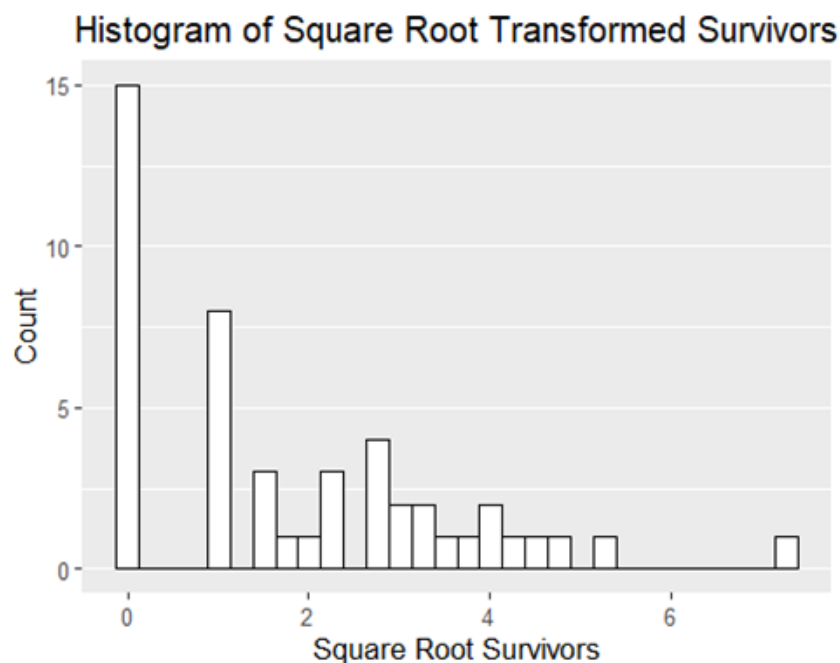


Figure 8:

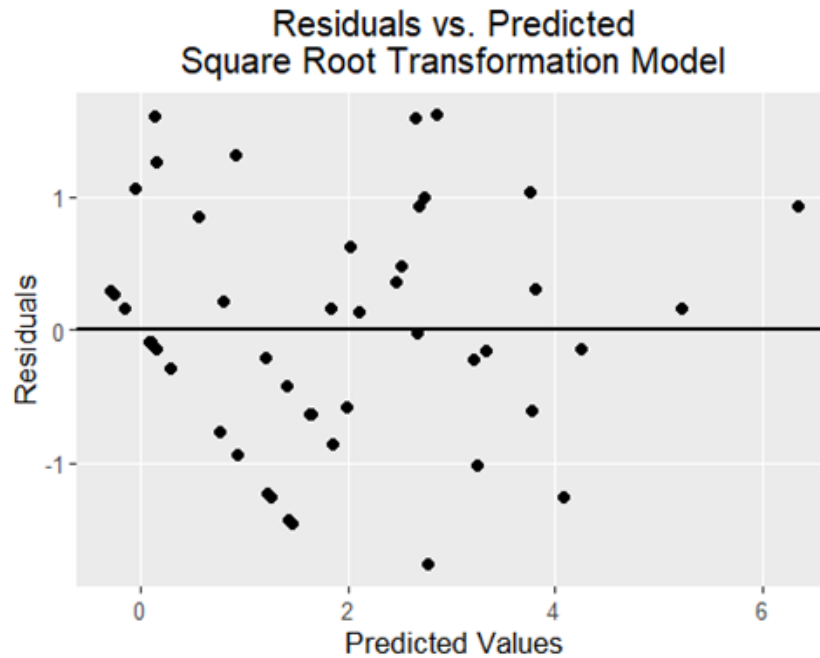
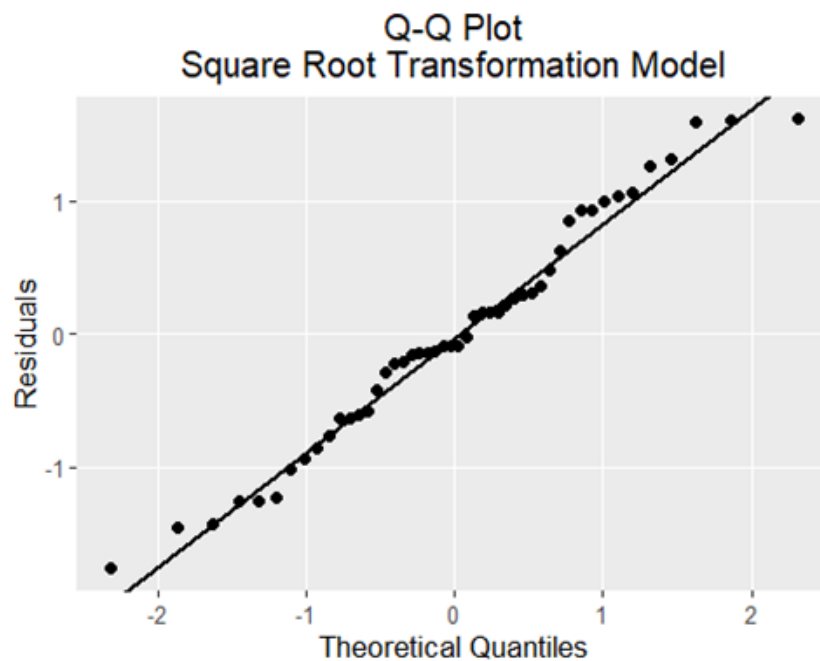


Figure 9:



## 7.2 Log Transformation

This model applies a log transformation in the hope that the result will be approximately normal. Because Survivors is often zero, the log transformation is applied to Survivors plus one. The result is modeled using a standard ANOVA model.



The diagnostics for this model suggest that it is not a good fit for the ALB data. The following issues can be observed in the diagnostic plots:

- Even after the transformation, the response does not appear normal. The histogram of log transformed Survivors makes the non-normality clear.
- The same straight line of points is visible in the lower left quadrant of the Residuals vs. Predicted plot.

**Figure 10:**

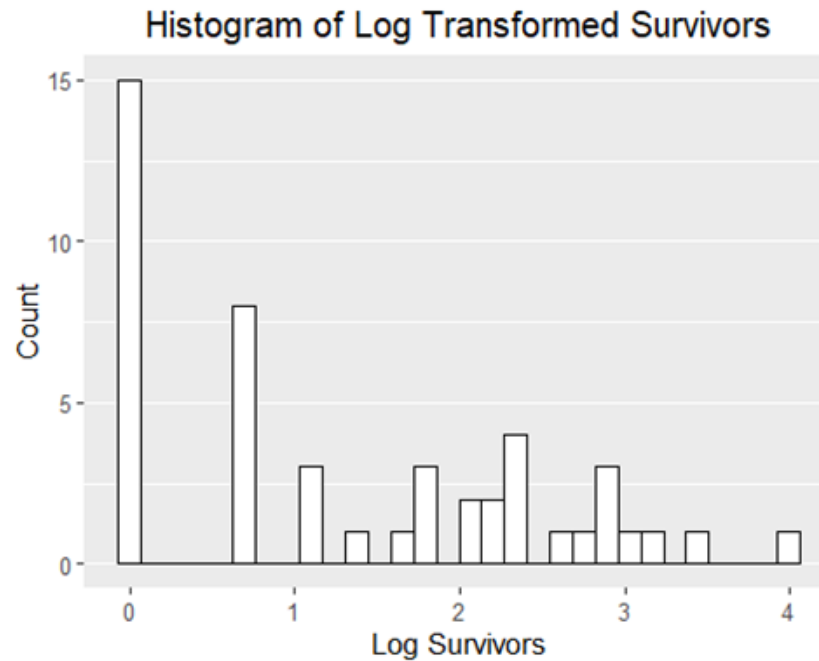


Figure 11:

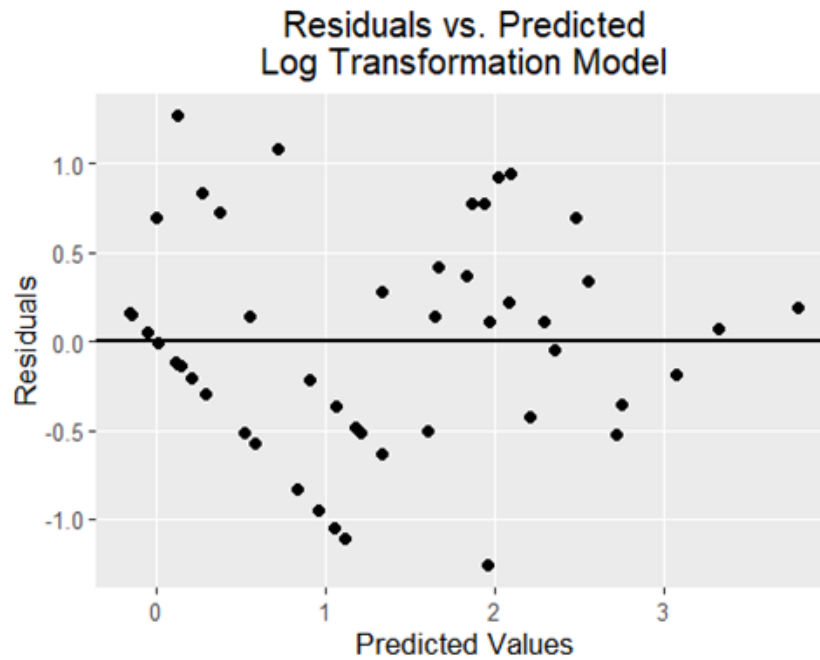
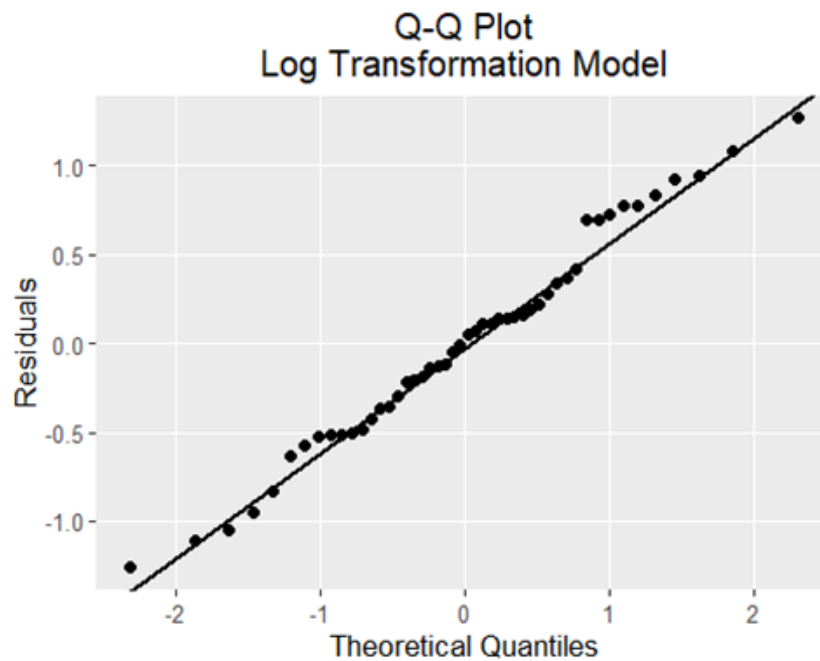


Figure 12:



### 7.3 Poisson with Overdispersion

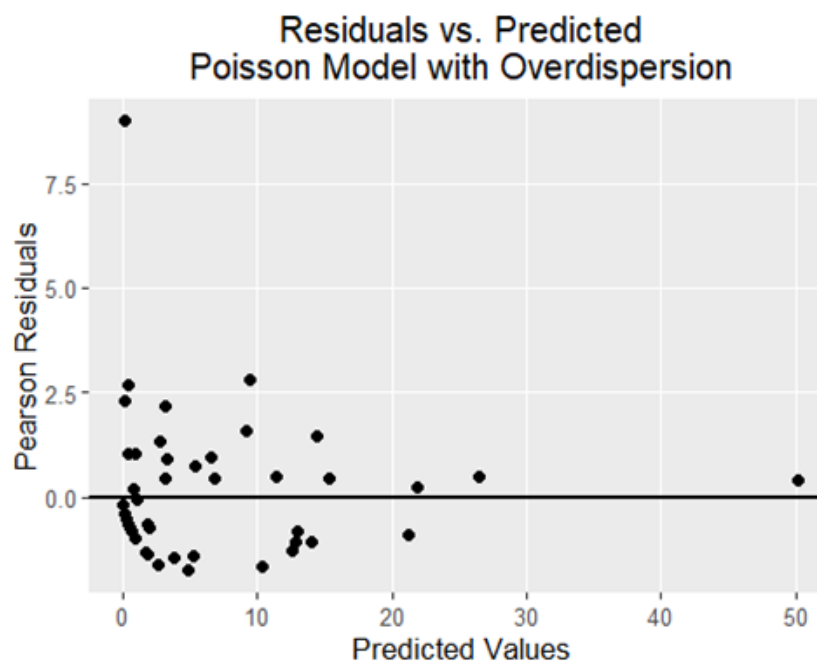
The Poisson model with overdispersion is different from the regular Poisson model. The regular Poisson model assumes that the response follows a Poisson distribution. If we assume that Survivors follows a Poisson distribution, we are assuming that the mean and variance of Survivors are equal. The sample mean of Survivors is 9.25, and the sample variance of Survivors is 256.74, so it does not appear

reasonable to assume that the mean and variance of Survivors are equal. The Poisson model with overdispersion adjusts for unequal mean and variance by introducing a “dispersion parameter”.

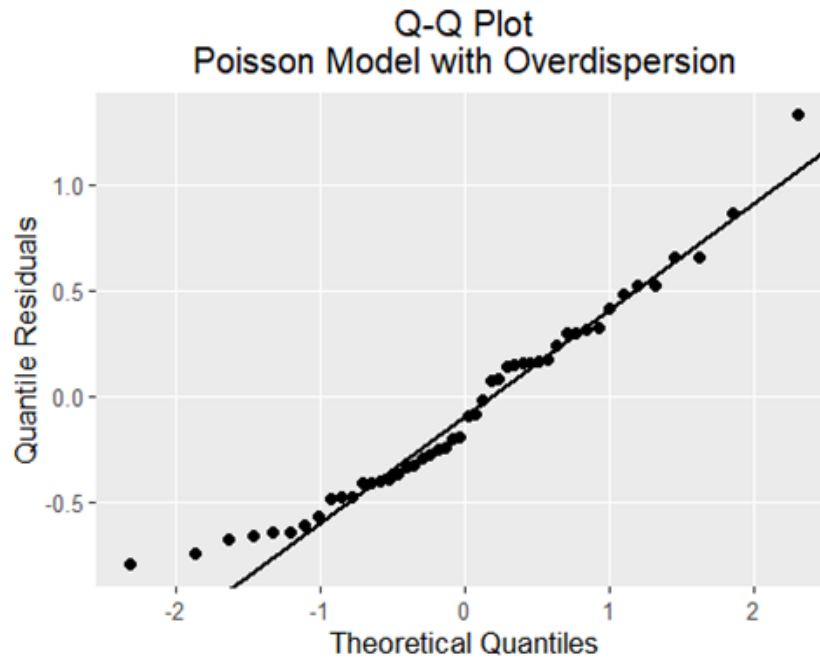
The diagnostic plots for the Poisson model with overdispersion are not bad, and they look comparable to the plots for the negative binomial model. Because the Poisson model with overdispersion and the negative binomial model have comparable diagnostic plots, goodness-of-fit statistics are also presented. Those statistics, however, do not clearly favor either model. Fortunately, it is not important to decide between the Poisson model with overdispersion and the negative binomial model because the zero inflated negative binomial model outperforms both.

Diagnostics for the Poisson model with overdispersion are below. It is worth noting that the Residuals vs. Predicted plot has Pearson residuals on the y-axis, and the Q-Q plot has quantile residuals on the y-axis. Because the Poisson with overdispersion model does not assume a normal response, the residuals need to be standardized before they are plotted. The Pearson residuals and the quantile residuals are appropriately standardized for the Residuals vs. Predicted and Q-Q plots, respectively.

**Figure 13:**



**Figure 14:**



**Table 10:** Model selection criteria for the Poisson model with overdispersion

Statistic	Value
AIC	263.04
AICC	372.45
BIC	319.18
Scaled Deviance	10.61
Pearson Chi-Square	147.13

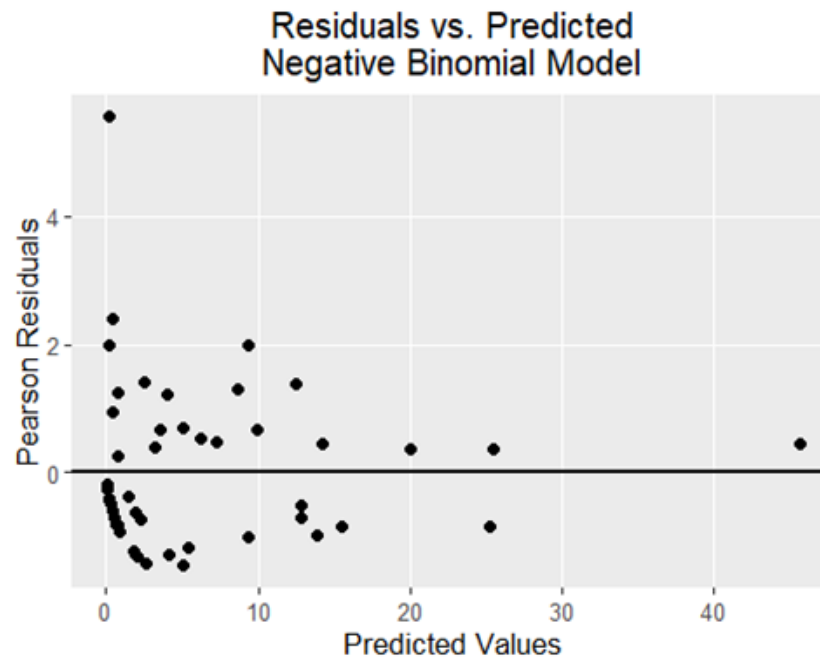
#### 7.4 Negative Binomial

The negative binomial model assumes that the response follows a negative binomial distribution. The negative binomial distribution allows the variance to be larger than the mean, which is reasonable for Survivors.

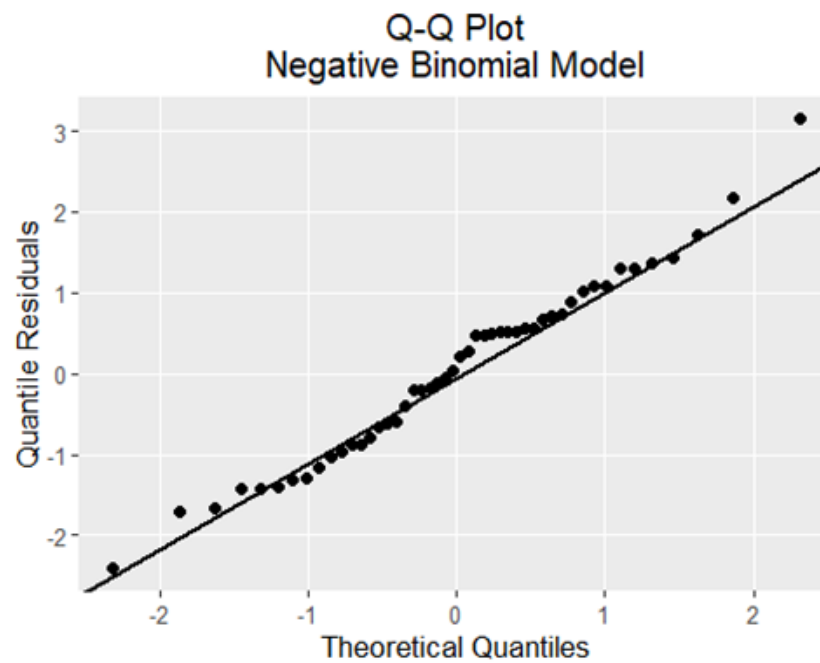
The diagnostic plots for the negative binomial model are comparable to the diagnostic plots for the Poisson model with overdispersion. AIC and Pearson Chi-Square favor the negative binomial model. However, AICC, BIC, and scaled deviance favor the Poisson model with overdispersion. Overall, the Poisson model with overdispersion and the negative binomial model perform comparably, but the zero inflated negative binomial model outperforms both.

Diagnostics for the negative binomial model are below. Again, Pearson residuals and quantile residuals are used.

**Figure 15:**



**Figure 16:**



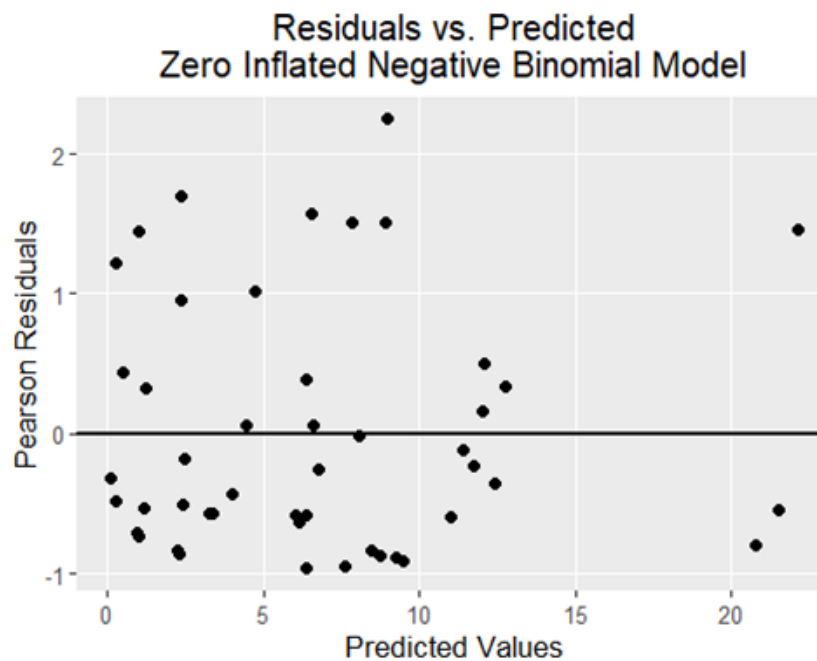
**Table 11:** Model selection criteria for the negative binomial model

Statistic	Value
AIC	262.82
AICC	386.82
BIC	320.82
Scaled Deviance	64.78
Pearson Chi-Square	78.53

### 7.5 Zero Inflated Negative Binomial

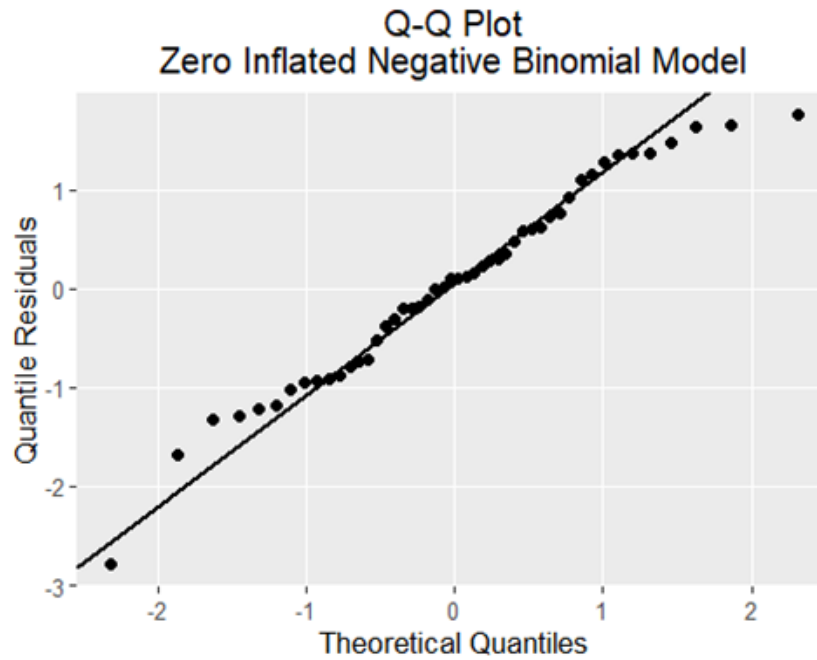
The Residuals vs. Predicted plot for the zero inflated negative binomial model is far better than the Residuals vs. Predicted plots for the other models, and the Q-Q plot is reasonable.<sup>5</sup> AICC, BIC, and Pearson chi-square favor the zero inflated negative binomial model. However, AIC and scaled deviance do not. Overall, the zero inflated negative binomial model is the clear choice.

**Figure 17:**



<sup>5</sup> The Q-Q plot for the zero inflated negative binomial model was created using version 0.2-1 of the countreg package in R. The countreg package is not available on CRAN because it is still in development. It is available on R-Forge [here](#).

**Figure 18:**



**Table 12:** Model selection criteria for the zero inflated negative binomial model

Statistic	Value
AIC	272.59
AICC	292.99
BIC	304.4
Scaled Deviance	238.59
Pearson Chi-Square	35.4

## 8. References

Wasserstein, R. L., & Lazar, N. A. (2016), "The ASA Statement on p-Values: Context, Process, and Purpose" *The American Statistician*, 70(2), 129-133.