Predicting the Probability and Abundance of Finding the New Orchid in Glimdonia Using Logistic and Negative-Binomial Regression

Jiaqi Bi[a,1,*]

[a] *Western University,*
Schulich School of Medicine & Dentistry,
Department of Epidemiology and Biostatistics

Abstract

Along with a newly discovered island, Glimdonia, scientists have been working hard in finding the new orchid species residing on the island, which they believe it contains medical components that may increase people's lifetimes. The paper suggests logistic regression and negative-binomial regression statistical modeling to predict the probability and abundance of the orchid on the island. The study shows that there is a higher probability of finding the orchid when the elevation is not too high ($p < 0.001$), and when the river is present ($p = 0.037$). The analysis also suggests the soil terrain may have more residences for the orchid ($p < 0.001$), and the lowland forest has the most abundance compared to others ($p = 0.002$). The upland forest has plentiful orchids but less than the lowland forest ($p = 0.002$). Many abundances can be found near the river in the prediction results ($p < 0.001$). The probability and abundance maps are produced from the statistical modelling as shown in Figure 4 and Figure 5.

Keywords: Logistic Regression, Negative Binomial Regression, Statistical Modelling, Model Prediction

[*] Corresponding author
Email address: `jbi23@uwo.ca` (Jiaqi Bi)

[1] Study Affiliated with Department of Statistical and Actuarial Sciences

1. Introduction

Glimdonia is a recently discovered island located at 0° N/S and 157° W within the South Pacific, and it positions on the equator, directly south of the Hawaiian Islands. following extensive surveys by biologists and ecologists, a novel orchid species has been uncovered. Preliminary investigations have suggested that this species may contain a crucial component with potential implications for life-extending elixirs. We implemented logistic regression and negative-binomial regression to model the probability and orchid abundance for each grid cell of the entire island based on the survey data, in order to provide support for scientists in further investigation of this new species. The survey data were completed by our cooperating scientists through their first investigation on the islands that included 200 sites of the research.

2. Methods

2.1. Data Preparation & Exploration

The scientists have conducted an initial exploration of the island and randomly selected 200 sites, and have provided the necessary characteristics of all sites including the east/west and north/south coordinates at the middle of the grid cell, the elevation in meters of the grid cell, the terrain and vegetation categories, and the existence of the grid cell, as well as the count of each site. The orchid is unlikely to survive in the ocean or the beach due to the high concentration of salt level, and is unlikely to grow on the surface of the glacial as it would not be able to gain enough nutritional ingredients. Thus, the statistical analysis was only based on the observations that do not have those mentioned terrain categories. That is, observations that the terrain is found as the ocean, glacial, or beach simply have no probabilities or abundances. The reduced survey dataset contains 126 observations and the reduced island dataset contains 3390 observations. The data visualization was conducted based on all covariates including the coordinates of the site, the elevation, the river existence status, the vegetation and terrain categories. I included a new variable to indicate the orchid is found if the count of the observation is not 0.

2.2. Statistical Modelling

2.2.1. Probability Prediction: Logistic Regression

The probability of each grid cell of the entire island can be predicted using logistic regression. I proposed 4 models to identify the probability for each observation. Note that I have $Y_1, ..., Y_{126}$ observed proportion of finding the orchid in separate surveys of the study, such that the link function

$$\text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \eta_i \tag{1}$$

The first model (Model 1) accounts for the second-order polynomial of x coordinate and y coordinate because the probability has a quadratic trend based on the Figure 1, as well as the interaction of x and y, while the terrain has an interaction effect with other characteristics including the river status, elevation, and the vegetation:

$$\begin{aligned}\eta_i &= \beta_0 + \beta_1 x_{i,1}^2 + \beta_2 x_{i,2}^2 + \beta_3 x_{i,1} + \beta_4 x_{i,2} + \beta_5 x_{i,1} x_{i,2} + \beta_6 I(x_{i,3}) \\ &\quad + \beta_7 x_{ij,4} + \beta_8 x_{i,5} + \beta_9 I(x_{i,3}) x_{ij,4} + \beta_{10} x_{i,5}\end{aligned} \tag{2}$$

where $x_{i,1}$ and $x_{i,2}$ are x and y coordinates of the grid, $I(x_{i,3})$ is the indicator of the terrain for ith observation is soil, $x_{ij,4}$ is the categorical variable vegetation that ith observation is j such that j can be lowland forest, upland forest, and none (considering the baseline is alpine meadow), $x_{i,5}$ is the covariate of the elevation of the site. The second model (Model 2) does not consider the interaction of characteristics but the interaction of x and y coordinates remains in the equation:

$$\begin{aligned}\eta_i &= \beta_0 + \beta_1 x_{i,1}^2 + \beta_2 x_{i,2}^2 + \beta_3 x_{i,1} + \beta_4 x_{i,2} + \beta_5 x_{i,1} x_{i,2} + \beta_6 I(x_{i,3}) \\ &\quad + \beta_7 x_{ij,4} + \beta_8 x_{i,5}\end{aligned} \tag{3}$$

In order to test the necessity of the interaction between x and y coordinates in latter analysis, the third model (Model 3) does not include the effect of the interaction between these two:

$$\begin{aligned}\eta_i &= \beta_0 + \beta_1 x_{i,1}^2 + \beta_2 x_{i,2}^2 + \beta_3 x_{i,1} + \beta_4 x_{i,2} + \beta_5 I(x_{i,3}) \\ &\quad + \beta_6 x_{ij,4} + \beta_7 x_{i,5}\end{aligned} \tag{4}$$

The fourth model (Model 4) was constructed based on the understanding of the potential high correlation between elevation and vegetation as shown in Figure 10, that this model does not include the covariate of the vegetation while considering the interaction between x and y coordinates:

$$\begin{aligned}\eta_i &= \beta_0 + \beta_1 x_{i,1}^2 + \beta_2 x_{i,2}^2 + \beta_3 x_{i,1} + \beta_4 x_{i,2} + \beta_5 x_{i,1} x_{i,2} + \beta_6 I(x_{i,3}) \\ &\quad + \beta_7 x_{i,5}\end{aligned} \tag{5}$$

2.3. Abundance Prediction: Negative-Binomial Regression

The most common statistical model for the count data is the Poisson regression, but it requires the assumption that the equality of mean and variance. Our data does not satisfy this property as shown in Figure 11. The variance is much larger in many groups. Therefore, an alternative approach of negative-binomial regression was used. Suppose that $Y_i|\mu_i \sim \text{Pois}(\mu_i)$ that Y_i is the observed count and μ_i is the mean for observation i. Introducing the extra variability into μ_i, we assume

$$\mu_i|\gamma, x_i \sim \Gamma(1/\gamma, \gamma\mu(x_i)) \tag{6}$$

such that the link function is log-linear,

$$\log(\mu(x_i)) = \eta_i \tag{7}$$

Based on the distribution of Y_i and μ_i, we can derive the marginal distribution of $Y|\gamma, x_i$ is negative-binomial (The derivation is referred to Section 6.1):

$$Y|\gamma, x_i \sim \text{Negative-Binomial}\left(1/\gamma, \frac{\gamma\mu(x_i)}{1+\gamma\mu(x_i)}\right) \tag{8}$$

This allows excessive variation for groups because $E(Y_i) = \mu(x_i)$ and $V(Y_i) = \mu(x_i) + \gamma\mu(x_i)^2$ for negative-binomial distribution. The first model is similar to the first model for logistic regression, we considered the quadratic trend as shown in Figure 2 and Figure 3, such that

$$\begin{aligned}\eta_i = \beta_0 + \beta_1 x_{i,1}^2 + \beta_2 x_{i,2}^2 + \beta_3 x_{i,1} + \beta_4 x_{i,2} + \beta_5 x_{i,1}x_{i,2} + \beta_6 I(x_{i,3}) \\ + \beta_7 x_{ij,4} + \beta_8 x_{i,5} + \beta_9 I(x_{i,3})x_{ij,4} + \beta_{10} x_{i,5}\end{aligned} \tag{9}$$

The notation is the same. The rest of the models are based on the same model settings as the logistic regression in the previous section. The model selection criterion is implemented using the AIC since both the logistic and negative-binomial regression are likelihood-based models. The equation for calculating the AIC is

$$AIC = 2k - 2\log(\hat{L})$$

where k is the number of parameters in the model and $\hat{L}$ is the maximum value of the likelihood function of the model.

3. Results

3.1. Explanatory Analysis

To visualize the probability of finding the habitat of the newly discovered orchid, I plotted the sites versus the x & y coordinates of the map grid as shown in Figure 1. Most of the habitats were found approximately between $(25, 25)$ to $(75, 75)$. In Figure 2 and Figure 3, it is quite visually significant that the orchid is in favor of the terrain soil and residing near the river, while mostly in the middle of the island as well. The orchid was found in the initial research mostly in the forest, and there was more probability to find them in the lowland forest than the upland forest as shown in Figure 6. The soil has a higher probability to find the orchid than the rock terrain as shown in Figure 7. Figure 8 shows that there are more lands that are not near the river, but there is a higher probability to find the orchid near the river. Most of the abundances of the orchid are below 400 meters of elevation on the island as shown in Figure 9. The elevation and the vegetation have a very significant correlation that can be concluded by Figure 10, that these two characteristics can influence each other.

3.2. Statistical Analysis

3.2.1. Logistic Regression Model

Using the AIC, Model 2 was selected to be the final model for the prediction of probability in each grid for Glimdonia Island as shown in Table 3. Table 1 shows the coefficients of the final logistic regression model. The second order of the polynomial term in the x-coordinate and y-coordinate shows a statistically significant result (p = 0.019 and p = 0.002), indicating that the quadratic effect of the probability is significant. The terrain soil has approximately 0.704 more log odds than the terrain rock holding other factors constant with no statistical evidence (p = 0.224, 95% CI: -0.454 to 1.840). When there is 1-meter increase in the elevation adjusting other covariates unchanged, the average log odds of finding the orchid decrease by 0.01 with strong evidence (p < 0.001, 95%CI: -0.016 to -0.005). The average log odds ratio of finding the orchid for the river-side site compared to the non-river-side site is 2.562 with moderate statistical significance (p = 0.037, 95%CI: 0.530 to 5.428). Figure 12 and Figure 13 have shown the x and y coordinates versus the standardized residual plot, that most of the points have lied within -2 and 2 with no significant trend, indicating a good fit of these covariates. Figure 15 shows a binned residual plot for the logistic regression, that most of the fitted values are within the standard error bounds indicating a good of fit for our model. Figure 16 to Figure 18 have shown that covariates of terrain, river and elevation versus the standardized residual that most of the values lied within -2 to 2 as well. The elevation residual plot does not concern me since the survey data contains more points from elevation 0 to 200. The Cook's distance in Figure 14 shows there are several influential points, that could be caused by some categories containing fewer data points than others. Using Hosmer-Lemeshow Test (HL Test), either I chose the group to be 10 or 100, I obtained insignificant C statistic and H statistic as shown in Table 4. This indicates the goodness of fit was not rejected. Finally, Figure 4 shows the predicted probability of finding the orchid on the island, using x and y coordinates to visualize the aerial view of the probability distribution in the geographic data.

3.2.2. Negative-Binomial Model

After applying the AIC model selection criterion, the final model was chosen to be Model 4 as shown in Table 3. Table 2 shows the coefficients of the final negative binomial model (exponentiated from the log scale). The quadratic term effect of the x-coordinate is not significant in the negative-binomial model ($p = 0.183$). The model suggests the soil terrain contains 5.677 times more orchids than the terrain rock considering other covariates constant with strong evidence ($p < 0.001$, 95%CI: 5.403 to 5.964). When there is a 1-meter increase in the elevation, the average abundance of orchids is 1.001 times higher with very little evidence ($p = 0.820$, 95%CI: 1.000 to 1.001). The lowland forest has the average abundance of orchids of 87.930 times more than the alpine meadow with statistical significance ($p = 0.003$, 95%CI: 67.581 to 114.406). Due to the lack of counts in the survey data, the model does not converge when the vegetation is none. The upland forest has average 28.457 times more orchids than the alpine meadow with strong evidence ($p = 0.002$, 95%CI: 23.459 to 34.519). The orchid is 14.762 times more discovered near the river than sites with no rivers with very strong evidence ($p < 0.001$, 95%CI: 14.221 to 15.323). Figure 19 to Figure 22 have shown the covariates of coordinates are well fitted in the model since the points are within -2 to 2 and no significant outliers exist. Figure 23 to Figure 26 have shown the covariate of vegetation, terrain, river and elevation are well fitted. Based on Figure 27, there are several influential points affecting the model a lot. Figure 28 has shown that the negative-binomial model solves the unequal mean and variance problem that could be encountered in the

Poisson model by applying the overdispersion term, though there could be more flexible adjustments by using quasipoisson model. However, quasipoisson model should be avoided as much as possible. Finally, Figure 5 shows the predicted abundance on the island. Most of the abundance are near the river.

4. Conclusion

In summary, considering the quadratic effect of coordinates is important in predicting the probability of finding the orchid. The terrain has a minimal effect in predicting the probability of where the orchid is residing. The factor of elevation has a negative relationship with the log odds of finding the orchid, which suggests scientists focus on the lower sea level. The orchid is likely to be found near the river as well. The negative binomial model has given the predicted abundance of the orchid, that there are more orchids residing in the soil terrain. The elevation in this model does not really reflect how many orchids are residing. However, there are most orchids reside in the lowland forest, followed by upland forests and alpine meadows. There is no evidence of habitat for orchids when there is no vegetation at all. The prediction also suggests that there are more orchids near the river. Figure 4 and Figure 5 have shown the predicted probability and abundance across the entire island.

5. Discussion and Limitation

The prediction used supervised learning from statistical inference. There is a limitation of sample size, that scientists only provided 200 observations of the survey, while there are 10000 observations in the entire map. The small sample size might have given some biased estimates of the prediction. Moreover, the negative-binomial model solved the excessive variance in the sample, but the mean-variance relationship is still not optimized.

6. Appendix

6.1. Derivation of Negative Binomial Distribution

We have $Y_i|\mu_i \sim \text{Pois}(\mu_i)$ and $\mu_i|\gamma, x_i \sim \text{Gamma}(\frac{1}{\gamma}, \gamma\mu(x_i))$. Note that μ_i and $\mu(x_i)$ is not the same. For simplicity and convenience, I will use $y|\mu$, $\mu|\gamma, x_i$ as the notation. The PMF of $y|\mu$ is

$$\pi(y|\mu) = \frac{\mu^y e^{-\mu}}{y!} \tag{10}$$

The PDF of $\mu|\gamma, x_i$ is

$$\pi(\mu|\gamma, x_i) = \frac{1}{\Gamma(1/\gamma)(\gamma\mu(x_i))^{1/\gamma}}\mu^{1/\gamma-1}e^{-\mu/\gamma\mu(x_i)} \tag{11}$$

The marginal PDF for $y|\gamma, x_i$ is then

$$\begin{align}
\pi(y|\gamma, x_i) &= \int_0^\infty \pi(y|\mu)\pi(\mu|\gamma, x_i)d\mu \tag{12}\\
&= \int_0^\infty \frac{\mu^y e^{-\mu}}{y!}\left(\frac{1}{\Gamma(1/\gamma)(\gamma\mu(x_i))^{1/\gamma}}\mu^{1/\gamma-1}e^{-\mu/\gamma\mu(x_i)}\right)d\mu \tag{13}\\
&= \frac{1}{y!\Gamma(1/\gamma)(\gamma\mu(x_i))^{1/\gamma}}\int_0^\infty \mu^y e^{-\mu}\mu^{1/\gamma-1}e^{-\mu/\gamma\mu(x_i)}d\mu \tag{14}\\
&= \frac{1}{y!\Gamma(1/\gamma)(\gamma\mu(x_i))^{1/\gamma}}\int_0^\infty \mu^{y+1/\gamma-1}e^{-\mu\left(\frac{1+\gamma\mu(x_i)}{\gamma\mu(x_i)}\right)}d\mu \tag{15}\\
&= \frac{1}{y!\Gamma(1/\gamma)(\gamma\mu(x_i))^{1/\gamma}}\frac{\Gamma(y+\frac{1}{\gamma})}{\left(\frac{1+\gamma\mu(x_i)}{\gamma\mu(x_i)}\right)^{y+1/\gamma}} \tag{16}\\
&= \frac{\Gamma(y+1/\gamma)}{\Gamma(1/\gamma)y!}(\gamma\mu(x_i))^{-1/\gamma}\left(\frac{\gamma\mu(x_i)}{1+\gamma\mu(x_i)}\right)^{y+1/\gamma} \tag{17}\\
&= \frac{(y+1/\gamma-1)!}{(1/\gamma-1)!y!}(\gamma\mu(x_i))^{-1/\gamma}(\gamma\mu(x_i))^{y+1/\gamma}\left(\frac{1}{1+\gamma\mu(x_i)}\right)^{y+1/\gamma} \tag{18}\\
&= \binom{y+1/\gamma-1}{y}(\gamma\mu(x_i))^y\left(\frac{1}{1+\gamma\mu(x_i)}\right)^{y+1/\gamma} \tag{19}\\
&= \binom{y+1/\gamma-1}{y}\left(\frac{\gamma\mu(x_i)}{1+\gamma\mu(x_i)}\right)^y\left(1-\frac{\gamma\mu(x_i)}{1+\gamma\mu(x_i)}\right)^{1/\gamma} \tag{20}
\end{align}$$

Note that Equation (16) can be obtained using the property that $\int_0^\infty x^b e^{-ax} = \frac{\Gamma(b+1)}{a^{b+1}}$.

6.2. Tables and Figures

Table 1: Final Logistic Regression Model (Log-Odds Scale)

	Coefficient	p-value	95% Confidence Interval
β_0	-23.070	0.002	(-38.738, -9.109)
x^2	-0.004	0.019	(-0.008, -0.001)
y^2	-0.005	0.002	(-0.009, -0.002)
x	0.492	0.007	(0.149, 0.866)
y	0.578	0.001	(0.249, 0.947)
Terrain (Soil)	0.704	0.224	(-0.454, 1.840)
Elevation	-0.010	<0.001	(-0.016, -0.005)
River (True)	2.562	0.037	(0.530, 5.428)
$x \times y$	-0.001	0.427	(-0.003, 0.001)

Table 2: Final Negative-Binomial Regression Model (Converted From Natural Log Scale)

	Coefficient	p-value	95% Confidence Interval
β_0	2.804×10^{-6}	< 0.001	$(4.483 \times 10^{-6}, 1.754 \times 10^{-6})$
x^2	0.999	0.183	(0.999, 0.999)
y^2	0.998	0.033	(0.998, 0.999)
x	1.138	0.075	(1.124, 1.153)
y	1.205	0.012	(1.190, 1.221)
Terrain (Soil)	5.677	< 0.001	(5.403, 5.964)
Elevation	1.001	0.820	(1.000, 1.001)
Vegetation (Lowland Forest)	87.930	0.003	(67.581, 114.406)
Vegetation (None)	0	1.000	Not Convergent
Vegetation (Upland Forest)	28.457	0.002	(23.459, 34.519)
River (True)	14.762	< 0.001	(14.221, 15.323)
$x \times y$	1.000	0.251	(1.000, 1.000)

Table 3: AIC Model Comparison

	AIC (Logistic)	AIC (Negative Binomial)
Model 1	130.306	534.335
Model 2	128.084	530.916
Model 3	129.732	532.455
Model 4	123.900	534.102

Table 4: Hosmer-Lemeshow Goodness of Fit Test for Logistic Regression

	χ^2 value	p-value
10-Group C Statistic	6.895, 8 degrees of freedom	0.548
10-Group H Statistic	11.3, 8 degrees of freedom	0.185
100-Group C Statistic	83.807, 98 degrees of freedom	0.846
100-Group H Statistic	54.813, 98 degrees of freedom	0.999

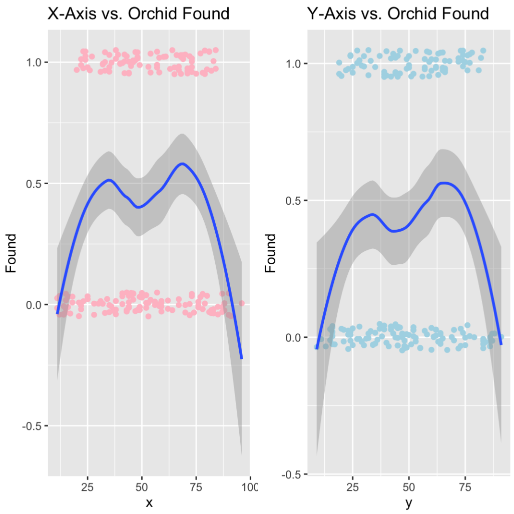

Figure 1: Coords vs. Number of Orchid Found Sites

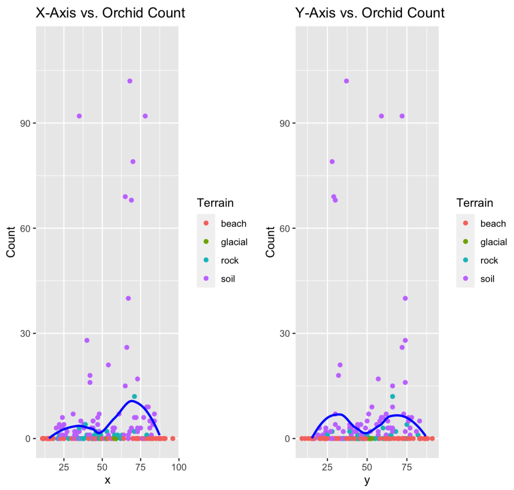

Figure 2: Coords vs. Count by Terrain

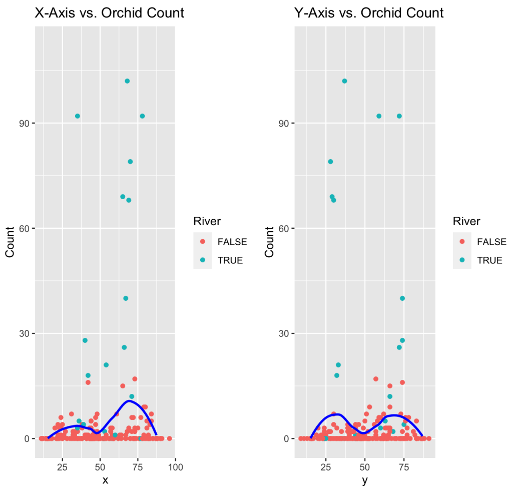

Figure 3: Coords vs. Count by River

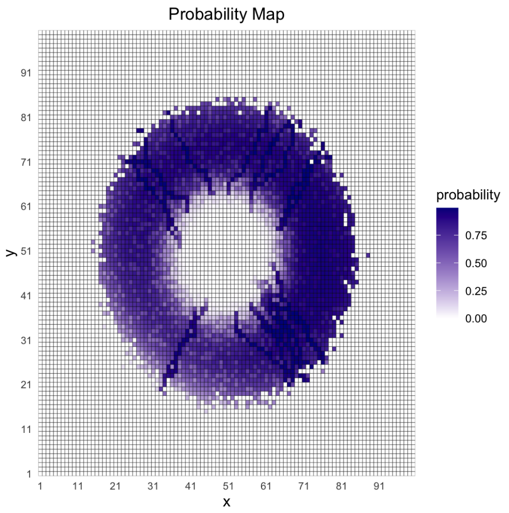

Figure 4: Predicted Probability Map

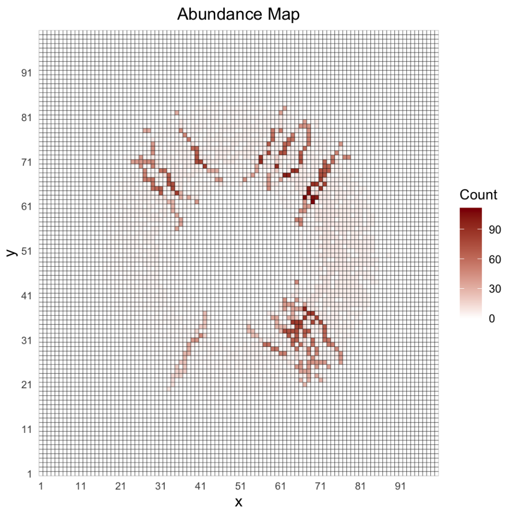

Figure 5: Predicted Abundance Map

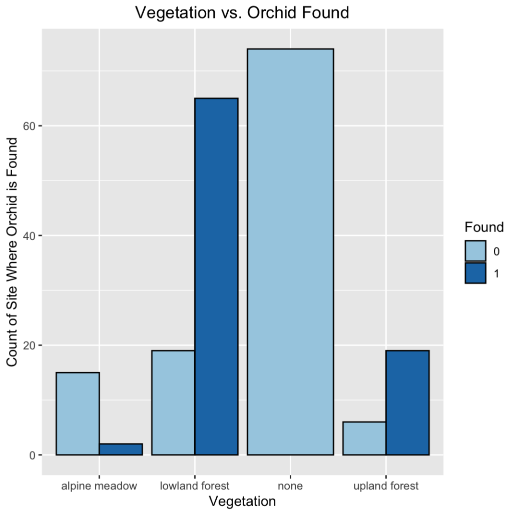

Figure 6: Vegetation vs. Sites Found

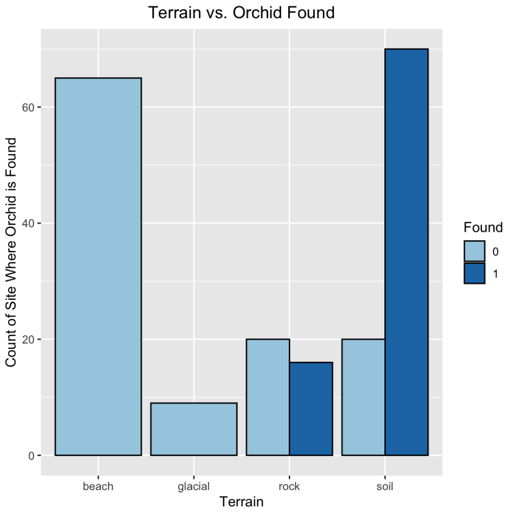

Figure 7: Terrain vs. Sites Found

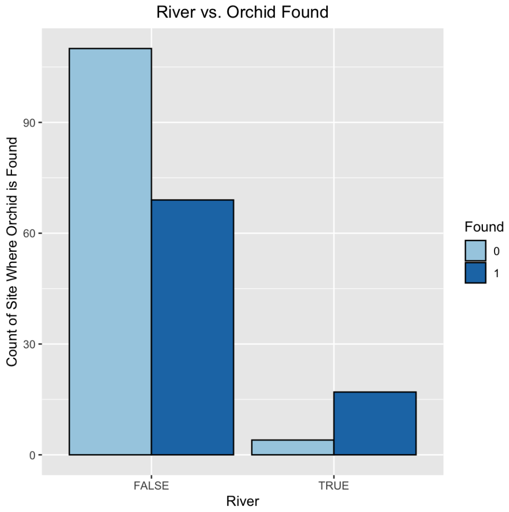

Figure 8: River vs. Sites Found

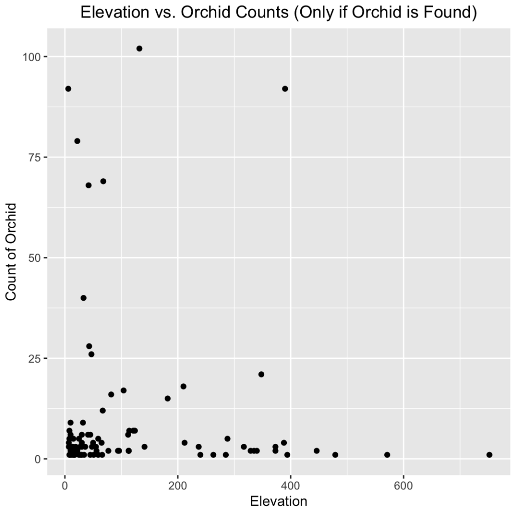

Figure 9: Elevation vs. Count

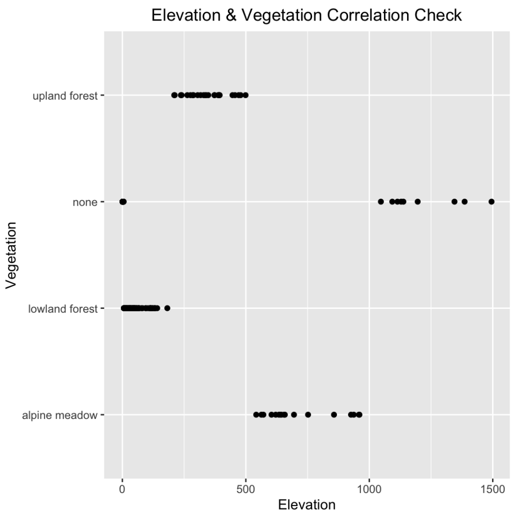

Figure 10: Elevation Vegetation Correlation

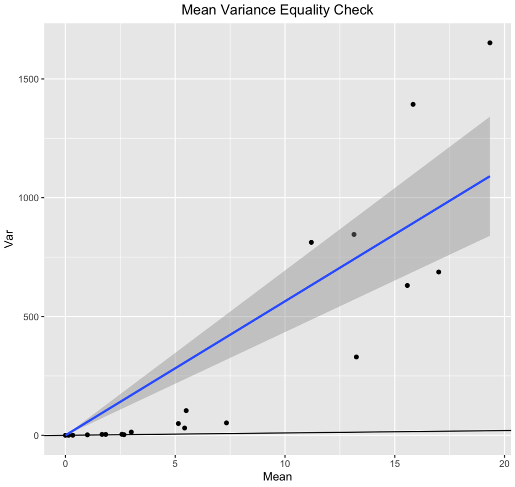

Figure 11: Poisson Mean vs. Variance

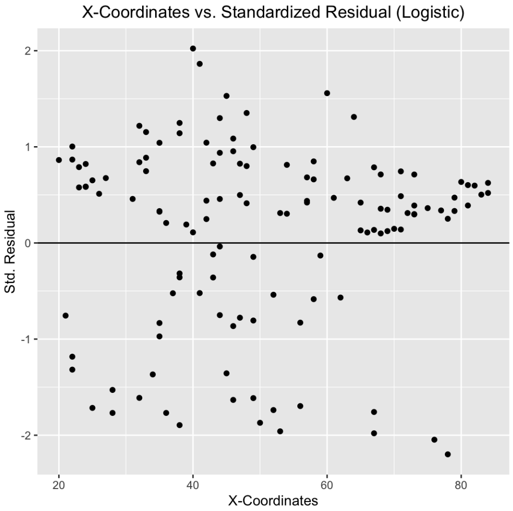

Figure 12: X-Coords vs. Residual (Logistic)

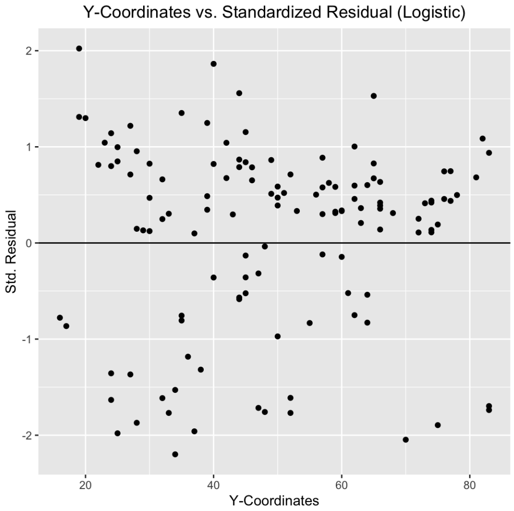

Figure 13: Y-Coords vs. Residual (Logistic)

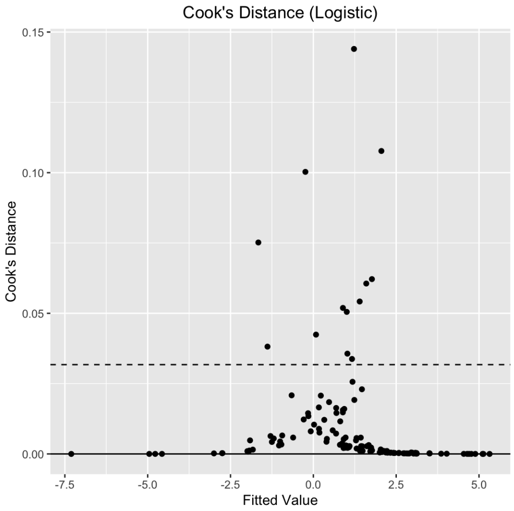

Figure 14: Cook's Distance (Logistic)

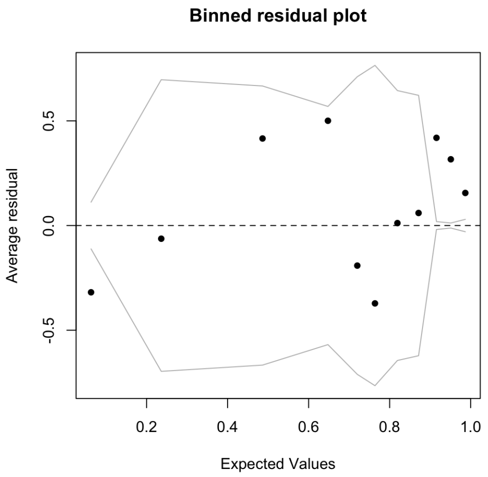

Figure 15: Binned Residual

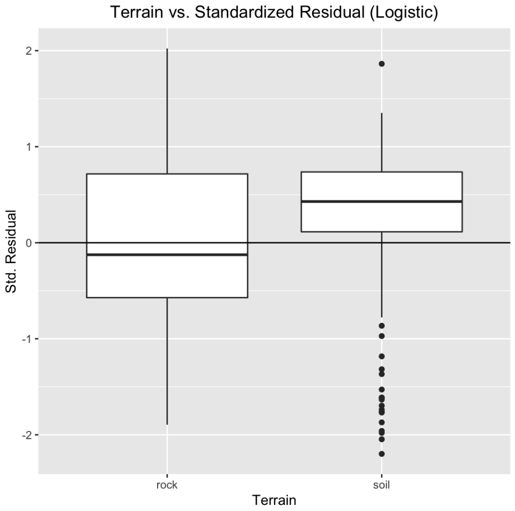

Figure 16: Terrain vs. Residual (Logistic)

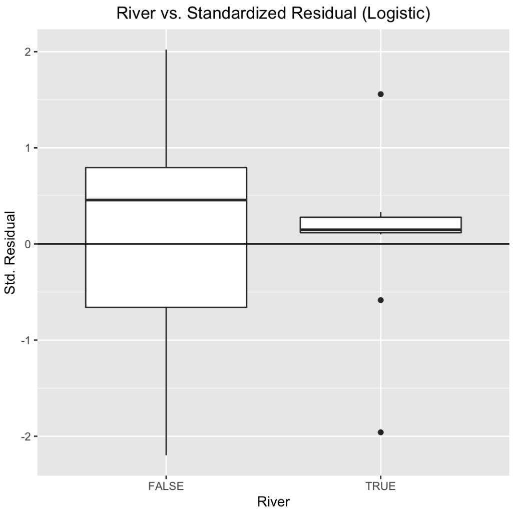

Figure 17: River vs. Residual (Logistic)

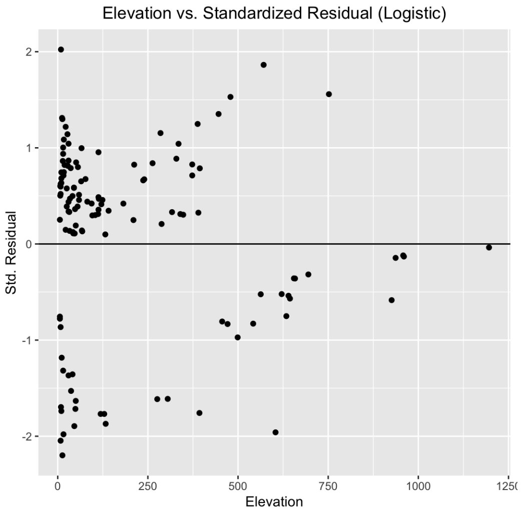

Figure 18: Elevation vs. Residual (Logistic)

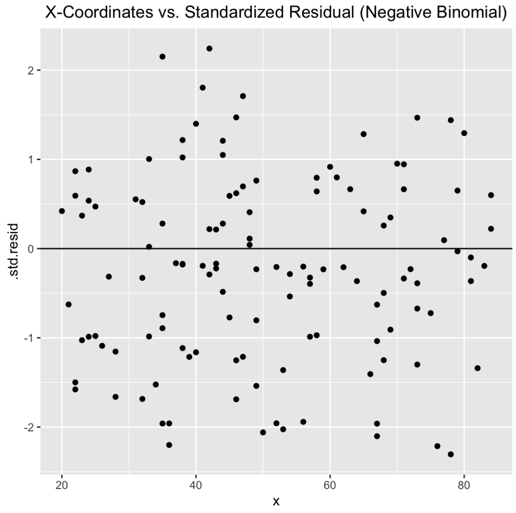

Figure 19: x Coords vs. Residual (Neg-Bin)

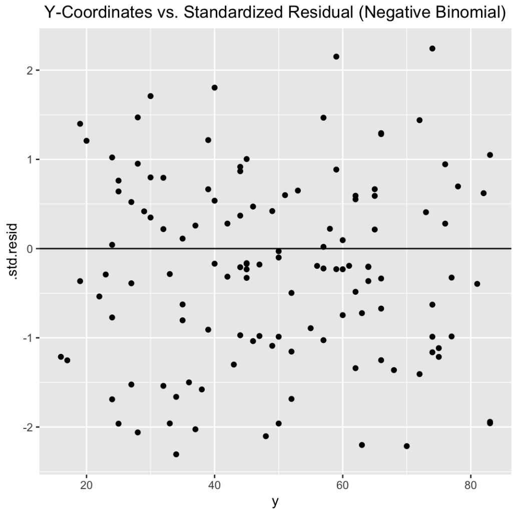

Figure 20: y Coords vs. Residual (Neg-Bin)

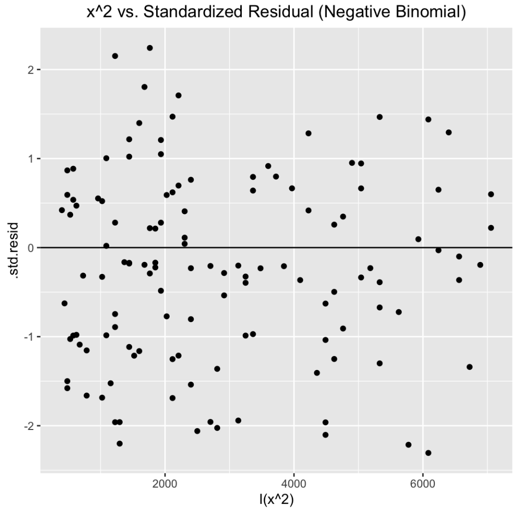

Figure 21: x^2 vs. Residual (Neg-Bin)

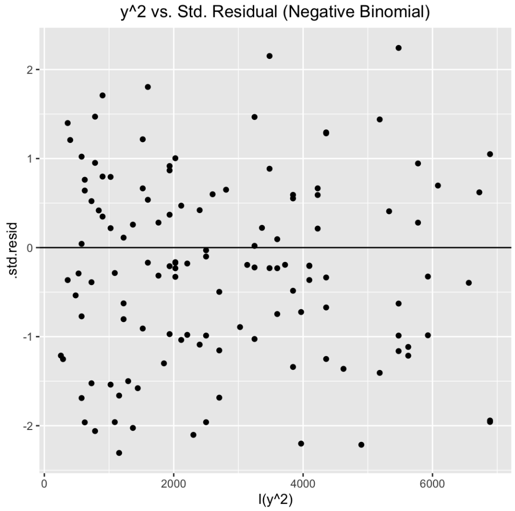

Figure 22: y^2 vs. Residual (Neg-Bin)

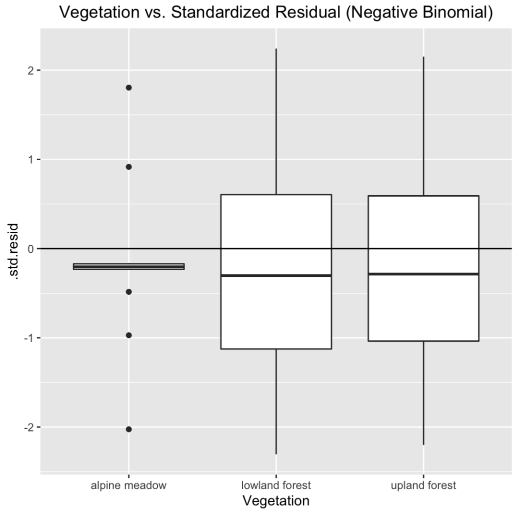

Figure 23: Vegetation vs. Residual (Neg-Bin)

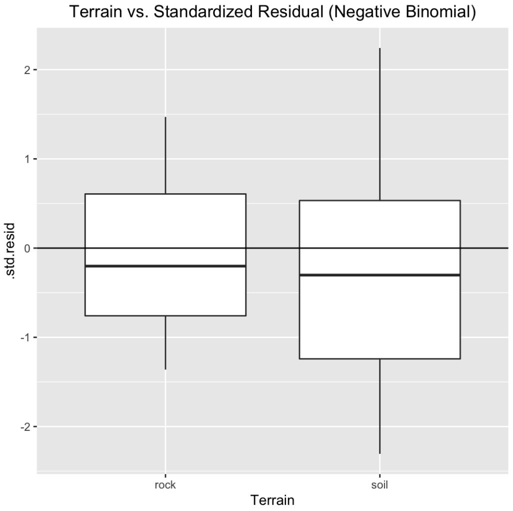

Figure 24: Terrain vs. Residual (Neg-Bin)

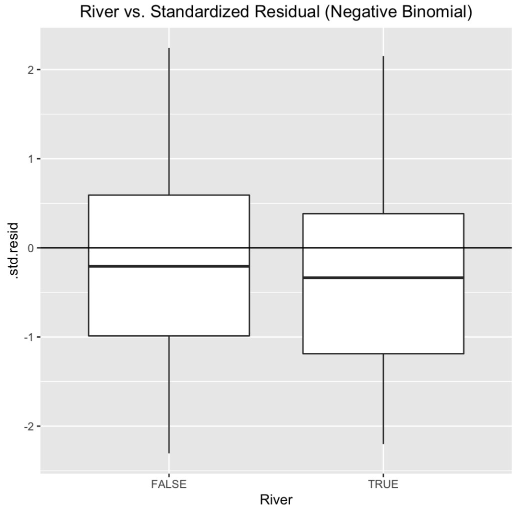

Figure 25: River vs. Residual (Neg-Bin)

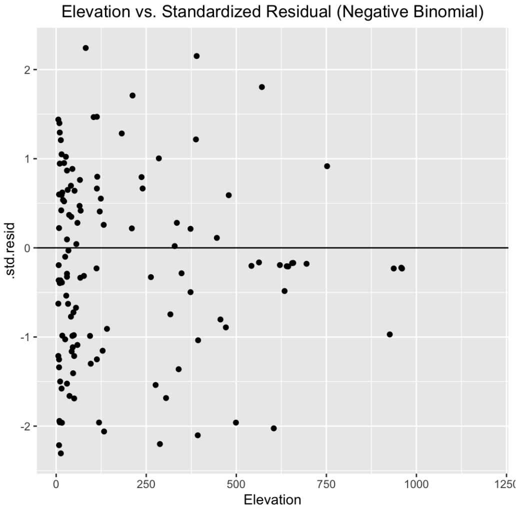

Figure 26: Elevation vs. Residual (Neg-Bin)

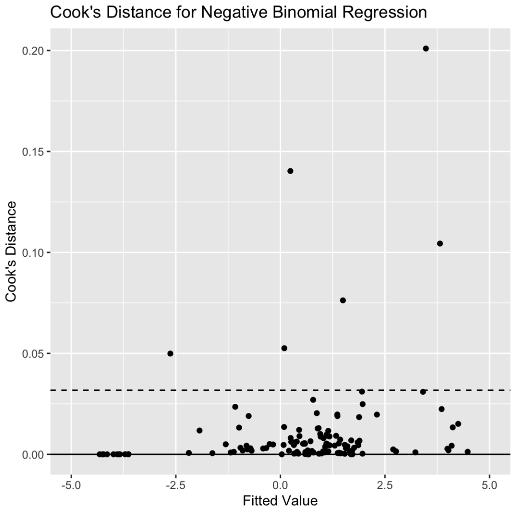

Figure 27: Cook's Distance (Negative-Binomial)

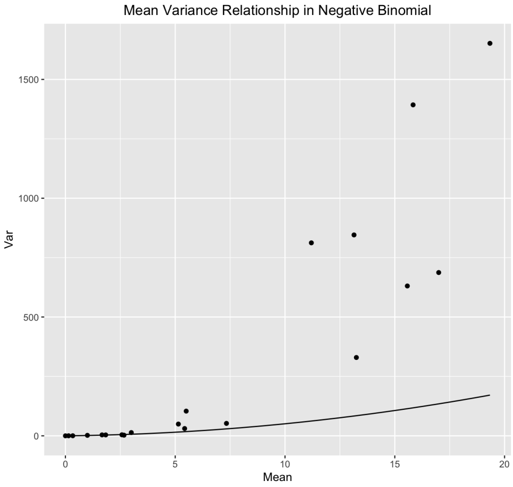

Figure 28: Mean Variance Relationship (Negative-Binomial)