Anthropology Statistical Model Building

Hanna Grossman

Contents

1 Introduction

The recovery and analysis of archaeological plant remains, known as paleoethnobotany, can yield insight into past human cultures (Van der Veen 2007). This data can also allow us to analyze how populations changed their agricultural, food preparation and consumption, social, and cultural practices over time, as well as how one location may vary from another and why this may be (Farahani in press: 17-18). At archaeological sites, deposits are sampled either through a "bulk" or "scatter" sampling strategy. If the "bulk" strategy is used, samples are collected from separate locations and analyzed independently, while if the "scatter" strategy is used, smaller samples are collected throughout each deposit and then combined for analysis. These samples are then processed, most commonly through flotation, a process which separates archaeological plant remains from the surrounding sediment through water sumbersion and agitation (Farahani in press: 12). This ultimately yields paleoethnobotanical data, which are quantified through counts of each type of plant remain found at a particular archaeological site in a given deposit and sample. As each sample may vary in volume, these counts may be analysed as densities if divided by the volume of each sample (Van der Veen 2007: 969). Here, volume refers to the amount of sediment collected in the field, most commonly in liters (L). These density analyses of plant remains can give insight into what types of plants were consumed and used in various ways and why. However, this data must be collected and analysed in a similar fashion across sites and time periods in order to be able to compare and contrast data found in different areas (Lee 2012: 654). In addition, one must take into consideration how these plant remains came to be preserved and how this preservation may differ from one set of data to another (Lee 2012: 651).

2 Literature Review

Paleoethnobotanical count data is often strongly skew right because many of the counts for any given sample are zero despite the size of the sample. This is because the plant remains are not evenly distributed throughout any given site (Van der Veen 2007: 971). Because of this uneven distribution of plant remains, the collected data may not follow a normal, or gaussian, distribution. If the data is not normally distributed, it will likely not meet the assumptions needed for linear modeling, and therefore will not be able to make accurate predictions about the number of archaeological plant remains found in a given time period. In addition, a linear model may not fit this data well if the variation in the volume, or amount of sediment collected in the field, does not account for a large amount of the variation in seed density. It is important to find a model that properly fits the data, allowing us to analyze how the number of archaeological plant remains found varies by period and volume and therefore make archaeological interpretations about past humans.

Overdispersion occurs in count data when there is extra variation present, or in other words when the variance is greater than the mean (Ver Hoef and Boveng 2007: 2766). This problem is often prominent in paleoethnobotanical data, as the variance of the seed densities is far greater than the mean of the seed densities. The data are overdispersed because, similar to the problem with paleoethnobotanical data being skewed right, many samples have zero or very low counts of plant remains regardless of the volume of the sample. Once again, this is because the plant remains are not evenly distributed throughout any given site and therefore there will be areas without any samples present, regardless of volume (Van der Veen 2007: 971). This uneven distribution of plant remains is caused by the actions of past humans who affect where these plant remains are used, how they are disposed of, and how they come to be preserved (Lee 2012: 651). These low values bring the mean down, while the samples with high counts keep the variance high. Because of this overdispersion, we first consider transforming our plant seed counts using the log +1 transformation (Ives 2015: 828). We also can consider the quasi-poisson and negative binomial regression models which are shown through past studies to fit overdispersed data well (Ver Hoef and Boveng 2007: 2766). We can also consider variations of these models such as the poisson, zero inflated quasi-poisson, and zero inflated negative binomial to see which models fit this type of data best (UCLA: Statistical Consulting Group 2019).

3 Methods and Materials

In this paper, we analyze two paleoethnobotanical count data sets, one from Dhiban and another from Las Capas to analyze the various models discussed above on real world data. The Dhiban data set has 211 samples and 204 variables, or columns. Out of these 204 variables, 127 represent species or taxa found at the archaeological site. In addition, the Dhiban data set contains samples from seven different time periods, spanning almost 2500 years: Iron I, Iron II, Nabataean-Roman, Late Byzantine, Late Antique Transitional, Middle Islamic I, and Middle Islamic II. This Dhiban data was collected from an excavation project in Dhiban, Jordan, by Alan Farahani (Farahani 2018). Below, the number of samples for each time period are displayed.

Period	Samples
Iron I	4
Iron II	22
Nabataean-Roman	7
Late Byzantine	39
Late Antique Transitional	61
Middle Islamic I	18
Middle Islamic II	60

The Las Capas data set has 1324 samples, and 51 variables. Out of these 51 variables, 45 represent species or taxa found at the archaeological site. The Las Capas data set contains samples from 3 time periods in the Late Archaic period: 504, 505, and 506. The Las Capas data was collected from the Las Capas site in southern Arizona by Desert Research Inc and represents material dating to 1000 BCE to 700 BCE (Sinensky and Farahani 2018). The number of samples for each time period are displayed below.

Period	Samples
504	618
505	206
506	458

In order to analyze these data sets, we first read the data into R and then create new data frames that include the total plant seed counts, grouped by Volume and Period. Volume refers to the amount of sediment collected in the field, and Period represents the various time periods at the given site. Here, the Total Plant Seed Counts include the counts of all of the plant seeds, ignoring fragments and other miscellaneous data. In the tables below, the total number of seeds for each time period represented in the Dhiban data, followed by the Las Capas data, are displayed.

Period	Total
Iron I	4
Iron II	223
Nabataean-Roman	556
Late Byzantine	924
Late Antique Transitional	867
Middle Islamic I	3927
Middle Islamic II	6484

Period	Total
504	39099
505	16233
506	15492

After visualizing the data as a preliminary step, we determine that Plant Seed Counts over 2,500 for the Las Capas data set were outliers, and thus create a new data set to use in our future models without these observations. The Plant Seed Counts for each time period for the Las Capas data free of outliers, can also be seen below.

Period	Total
504	39099
505	11295
506	15492

Next, were create our models for both the Dhiban and Las Capas data, which included linear, log-linear, negative binomial, and poisson models. We first choose a linear model to see if as Volume increases, the Plant Seed Count will also increase in a linear fashion. However, as discussed in the introduction, this tends to not be the case with paleoethnobotanical count data as the variation in the Plant Seed Count is not largely explained by Volume, or amount of sediment collected in the field. Many samples have zero or very low counts of plant remains regardless of the volume of the sample, as plant remains are not spread evenly throughout a site. Because of this, we next include the log-linear model, as transforming non-Gaussian data via the log transformation is a traditional approach to create a predictive model that satisfies parametic test assumptions (Ives 2015: 828). One negative to the log linear model is the fact that 1 must be added to each Plant Seed Count in order to avoid taking the log of any zeros. In addition, transformations have been seen to perform poorly on count data, as demonstrated by O'Hara (2010: 118). Because of this, O'Hara argues the use of generalized linear models, such as the Negative Binomial and Poisson to model count data (2010: 118). Because of this, we also model both of our data sets using the Poisson and Negative Binomial generalized linear models.

- Model 1 - Linear Model - Plant Seed Count ~ Volume + Period
- Model 2 - Log Linear Model - Log(Plant Seed Count + 1) ~ Log(Volume) + Period
- Model 3 - Poisson Model - Plant Seed Count ~ Volume + Period
- Model 4 - Negative Binomial Model - Plant Seed Count ~ Volume + Period

After using these four models on both the Dhiban and Las Capas data sets, we then create predictive data sets for each of the models. Each predictive data set has every possible combination of Volume and Period, allowing us to apply each model to the data, revealing the Plant Seed Counts each model predicts for any given combination of Volume and Period. This then allows us to graph Volume by Predicted Seed Count, separated by Period for each model, thus revealing how well each model predicts the Plant Seed Counts.

Finally, we use the function lrtest() from the package lmtest in R to compare the linear models to the log linear models, and the poisson models to the negative binomial models for each given data set, concluding which out of each set of models best fits the two data sets. From here, we then compare the residuals to the predicted counts of the two models that best fit each data set (Colin Cameron and Trivedi 2013). By comparing these plots, we ultimately conclude which models best fit both the Dhiban and Las Capas data.

4 Results

4.1 Fit of the Four Models on the Dhiban and Las Capas Data Sets

After fitting the four models to both the Dhiban and Las Capas data sets, we then analyze each model to see how well the model fits the data and thus how well the model is able to predict the Plant Seed Counts from the Volume and Period. The linear model when fit to the Dhiban data produces an R-squared of 46.26%, meaning the variation in Volume and Period account for about 46% of the variation in Plant Seed Counts. Next, the log linear model for the Dhiban data produces an R-squared of 69.41%, and is therefore able to account for a greater amount of the variation in the Plant Seed Counts.

For the poisson and negative binomial generalized linear models, rather than using the R-squared to measure how well the model fits the data, we instead look at null and residual deviance. The null deviance indicates how well the model predicts the response variable, while the residual deviance is similar to the residual sum of squares seen in the linear model. A small residual deviance indicates a good fit, while a large residual deviance indicates a poor fit (Zuur 2011: 83). The poisson model for the Dhiban data produces a null deviance of 22118.3, and a residual deviance of 7765.4, indicating a strong fit as the null deviance is much larger than the residual deviance. The negative binomial model for the Dhiban data produces a null deviance of 592.86, and a residual deviance of 232.19, once again indicating a strong fit as the null deviance is much larger than the residual deviance.

For the Las Capas data, the linear model produces an R-squared of only 1.09%, meaning Volume and Period account for only about 1% of the variation in Plant Seed Counts. The log linear model for the Las Capas data produces an R-squared of 1.78%, meaning it fits the data slightly better, but still does not perform well. The Las Capas poisson model produces a null deviance of 148222 and a residual deviance of 142880, and the

Las Capas negative binomial model produces a null deviance of 1615.2 and a residual deviance of 1549.4. Both of these models have null deviances that are only slightly larger than the residual deviances, indicating once again that the model does not fit the data as well as we would like.

For all of the models discussed above, the Period was a significant predictor, and therefore an important inclusion in the model. This can be seen below in the summary of the negative binomial model for the Dhiban data set. Here we see that every period is a significant predictor in this model.

```
##
## Call:
## glm.nb(formula = plant_seed_counts ~ Vol_L + Period, data = dhiban_count2,
##     init.theta = 1.378117357, link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.8026  -0.9962  -0.4076   0.1691   4.7809
##
## Coefficients:
##                                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)                    -0.219647   0.657497  -0.334 0.738331
## Vol_L                           0.037232   0.007168   5.194 2.05e-07 ***
## PeriodIron II                   2.128906   0.684786   3.109 0.001878 **
## PeriodLate Antique Transitional  2.466687   0.666233   3.702 0.000214 ***
## PeriodLate Byzantine            3.069972   0.670990   4.575 4.76e-06 ***
## PeriodMiddle Islamic I          4.742129   0.693248   6.840 7.89e-12 ***
## PeriodMiddle Islamic II         4.241278   0.671124   6.320 2.62e-10 ***
## PeriodNabataean-Roman           3.867663   0.735293   5.260 1.44e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(1.3781) family taken to be 1)
##
##     Null deviance: 592.86  on 210  degrees of freedom
## Residual deviance: 232.19  on 203  degrees of freedom
## AIC: 1906.8
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  1.378
##           Std. Err.:  0.135
##
##  2 x log-likelihood:  -1888.841
```

After analyzing the fit of these models on our data, we then use the lrtest() function in R's lmtest package to perform Likelihood Ratio Tests, thus allowing us to compare the given models. As seen by the output below for the dhiban data set, the log linear model performs better than the linear model, and the negative binomial model performs better than the negative binomial model. The Likelihood Ratio Tests for the Las Capas data sets yield the same conclusion.

```
## Likelihood ratio test
##
## Model 1: plant_seed_counts ~ Vol_L + Period
## Model 2: log_plant_seed_counts ~ log_Volume + Period
##   #Df   LogLik Df Chisq Pr(>Chisq)
## 1   9 -1219.08
## 2   9  -254.59  0  1929  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## Likelihood ratio test
##
## Model 1: plant_seed_counts ~ Vol_L + Period
## Model 2: plant_seed_counts ~ Vol_L + Period
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   8 -4397.0
## 2   9  -944.4  1 6905.2  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Next, we compare the residuals to the predicted counts for the log linear and negative binomial models of each data set to further determine which model provides the best fit (Colin Cameron and Trivedi 2013). Below the Residual vs Fitted plots for the Dhiban data (**Figure 1**), followed by the plots for the Las Capas data (**Figure 2**), are displayed. For both data sets, the log linear and negative binomial plots are broadly comparable, but upon closer inspection, the log linear models provide a slightly better fit. Because this difference is slight, the final determination may be due to other factors, and therefore, the best fitting model depends on the underlying data generating mechanism (Ver Hoef and Boveng 2007).

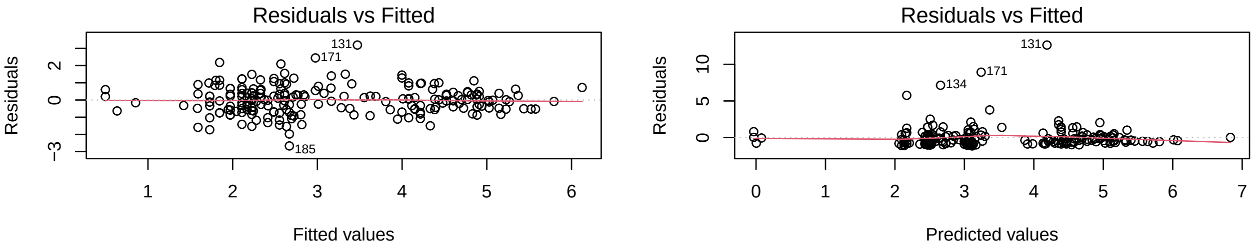

Figure 1: Comparison of Log Linear Model (left) to Negative Binomial Model (right) for Dhiban

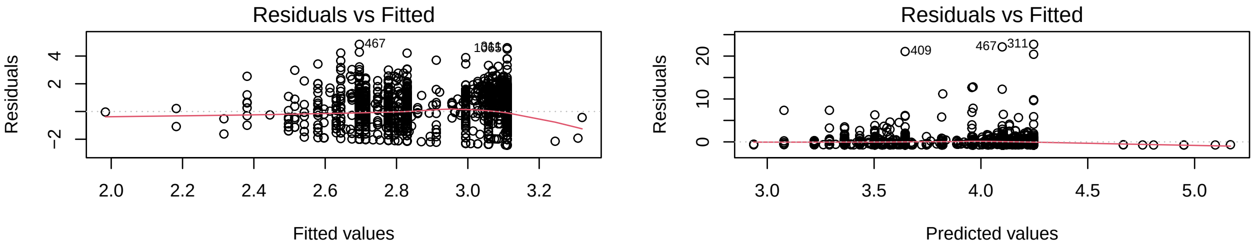

Figure 2: Comparison of Log Linear Model (left) to Negative Binomial Model (right) for Las Capas

4.2 Analysis of the Graphs of Predicted Data for Each Model

The graphs below show both the observed data, along with the Predicted Plant Seed Counts for each of the four models for both the Dhiban and Las Capas data sets. Each of these predicted models is broken up by Period, allowing for easy visualization of how the models behave, how well they fit the true Plant Seed Counts, and which Periods have the greatest influence on these counts.

As seen in the predictive graphs below, as Volume increases, Expected Plant seed Counts are seen to increase as well, in various fashions depending on the given model. In addition, the various Periods produce vastly different predictions, revealing the importance of Period as a predictor. For example, in both the log linear model for the Dhiban data (**Figure 5**), and the negative binomial for the Dhiban data (**Figure 7**), we see the strong increase in the Expected Seed Counts with the increase in Volume in the Nabataean-Roman, Middle Islamic I, and Midle islamic II periods. Meanwhile, we see very little, if any, increase in Expected Plant Seed Counts with respect to Volume for the Iron I, Iron II, and Late Antique Transitional periods.

When comparing the observed data for both the Dhiban data (**Figure 3**) and the Las Capas data (**Figure 8**) to the various predictions, it is clear that the log linear models (**Figure 5, 10**) and the negative binomial models (**Figure 7, 12**) best fit the observed data, explaining the results we observe from the Likelihood Ratio Tests above.

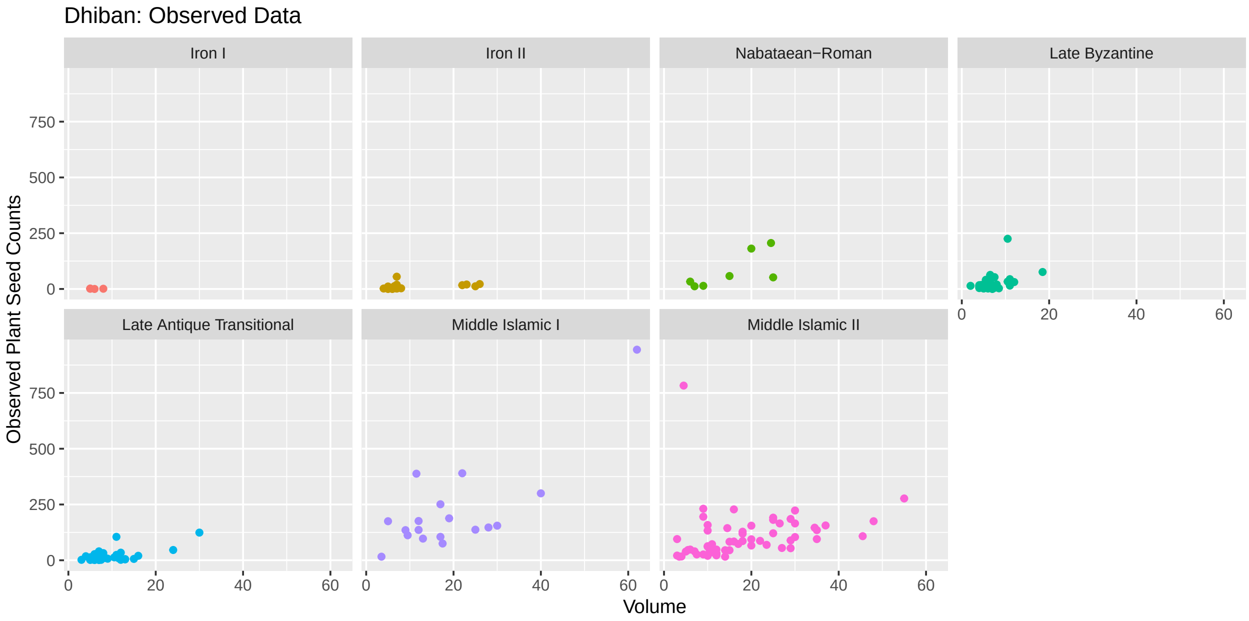

Figure 3: Observed Data for Dhiban

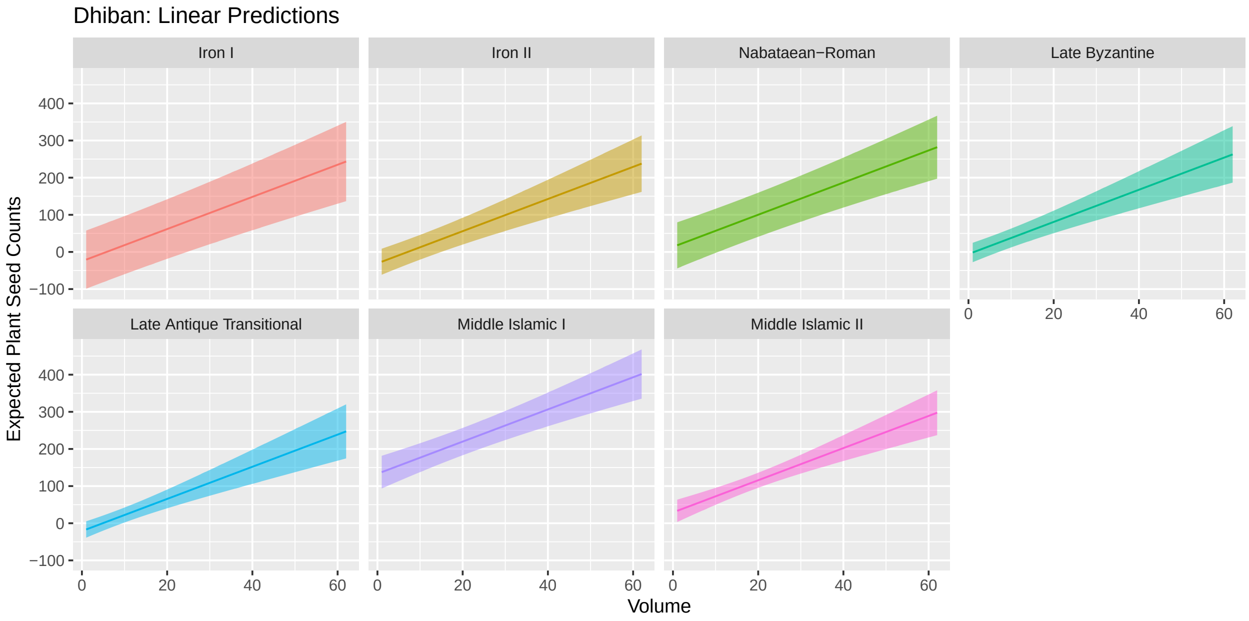

Figure 4: Linear Predictions for the Dhiban Data Set

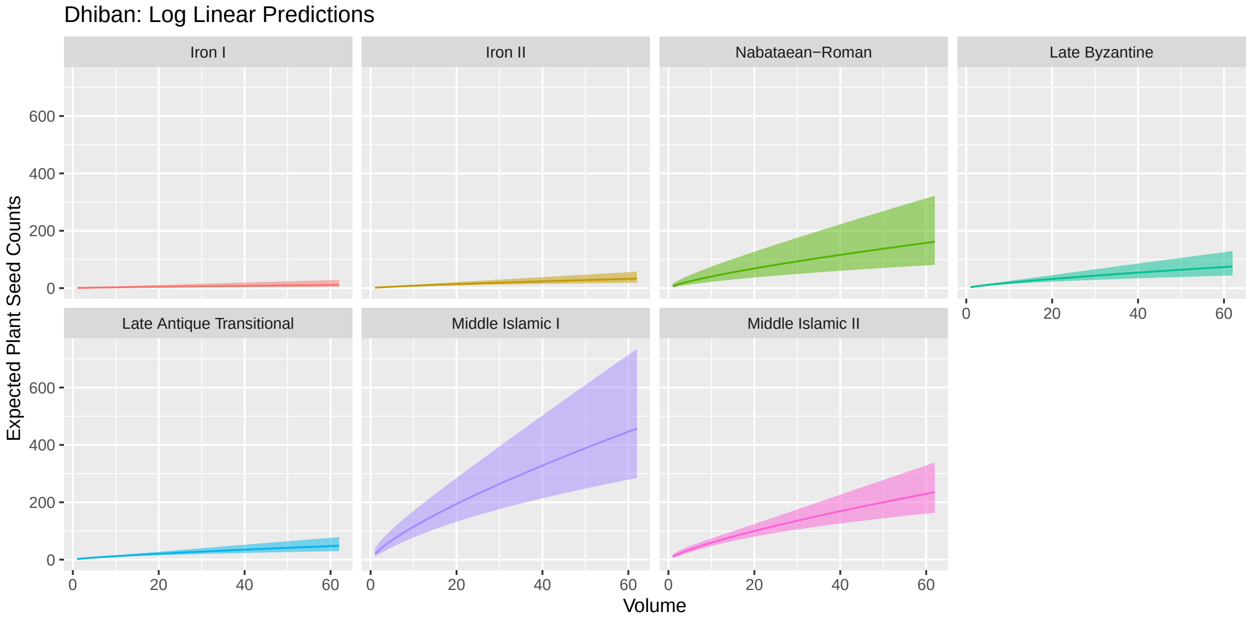

Figure 5: Log Linear Predictions for the Dhiban Data Set

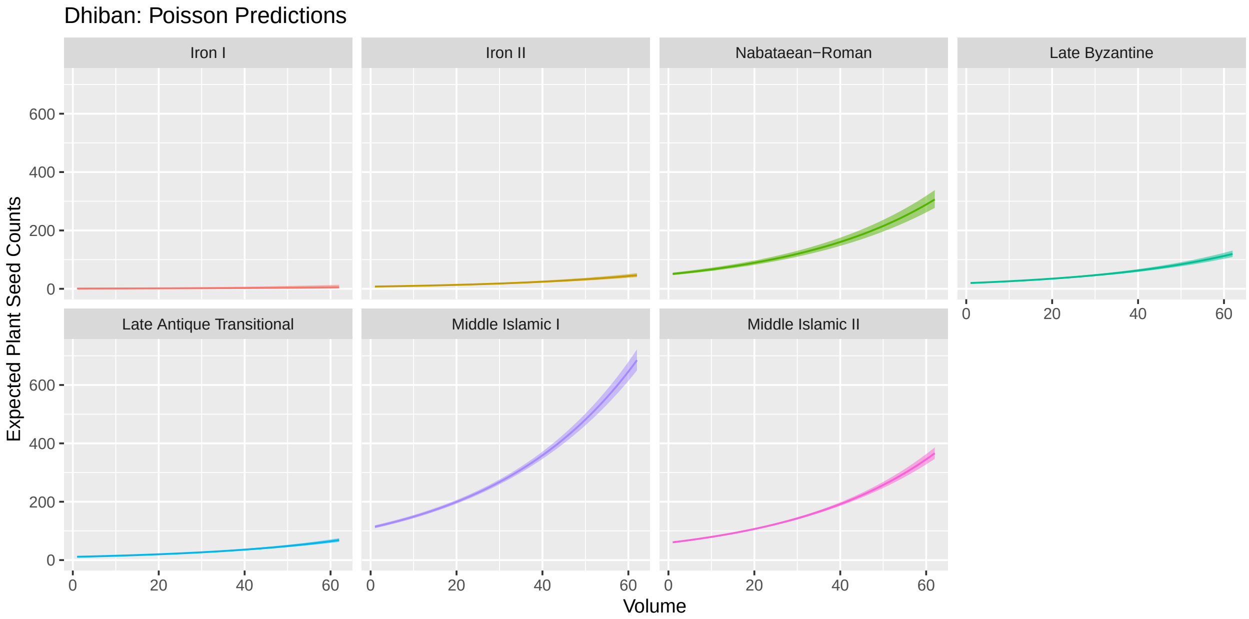

Figure 6: Poisson Predictions for the Dhiban Data Set

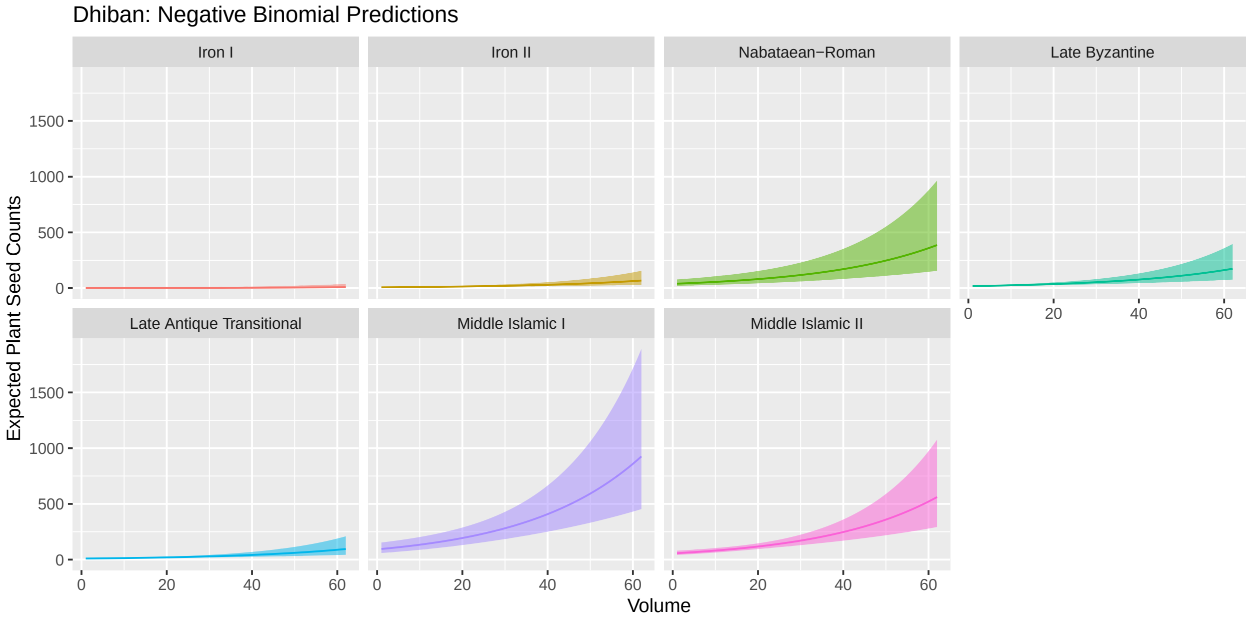

Figure 7: Negative Binomial Predictions for the Dhiban Data Set

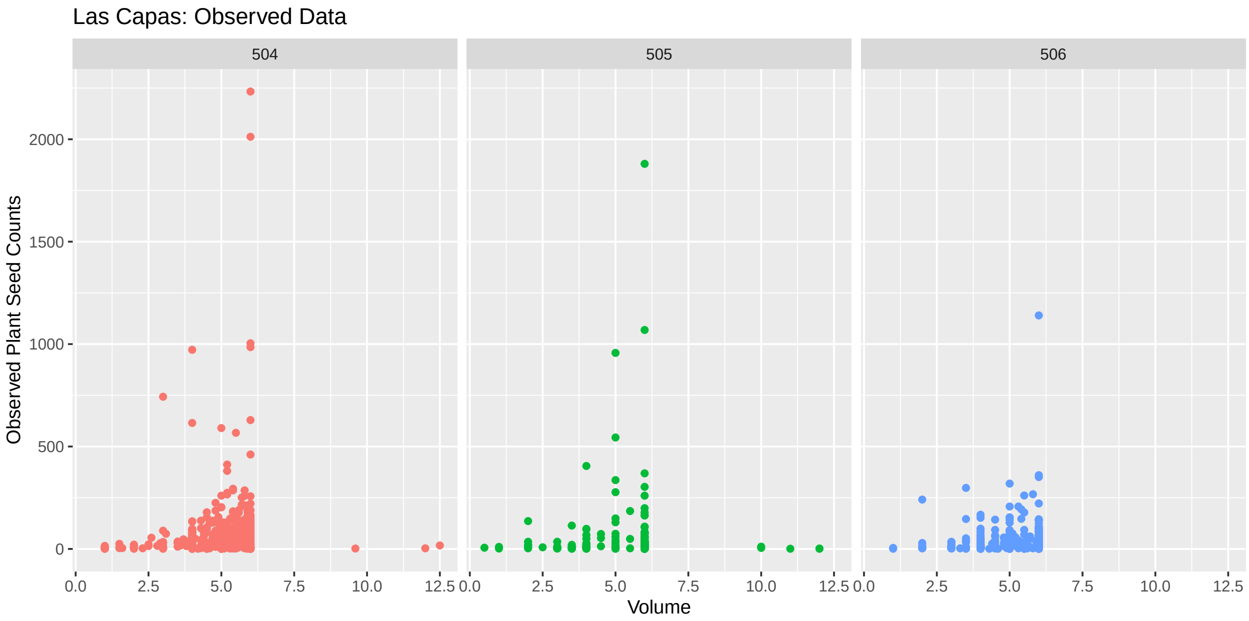

Figure 8: Observed Data for Las Capas

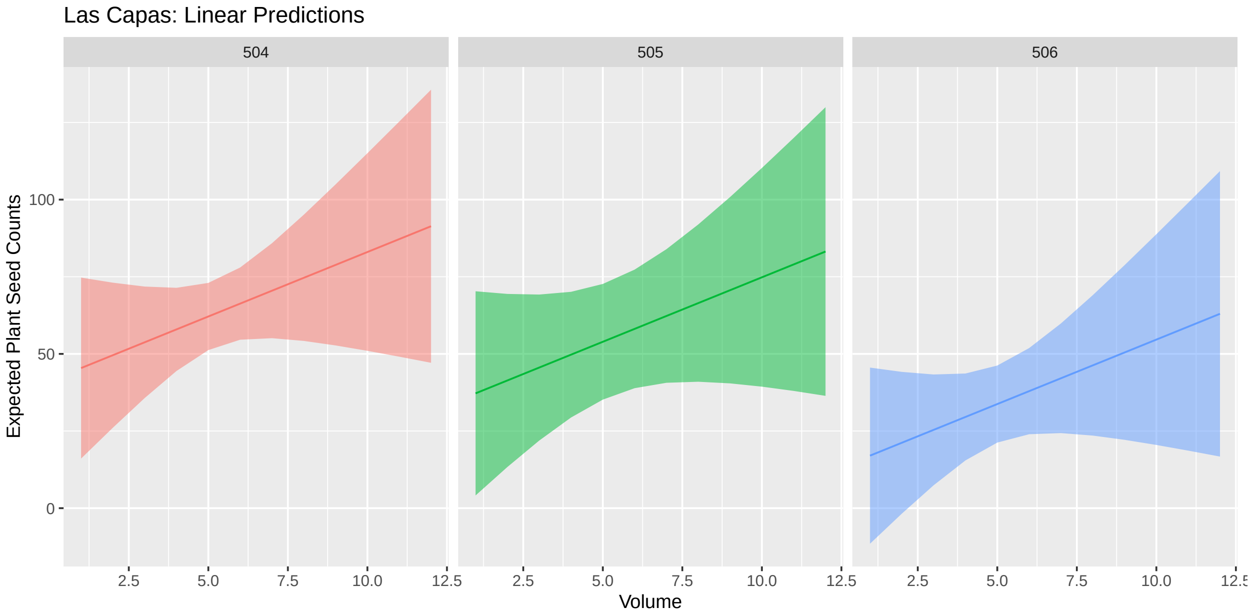

Figure 9: Linear Predictions for the Las Capas Data Set

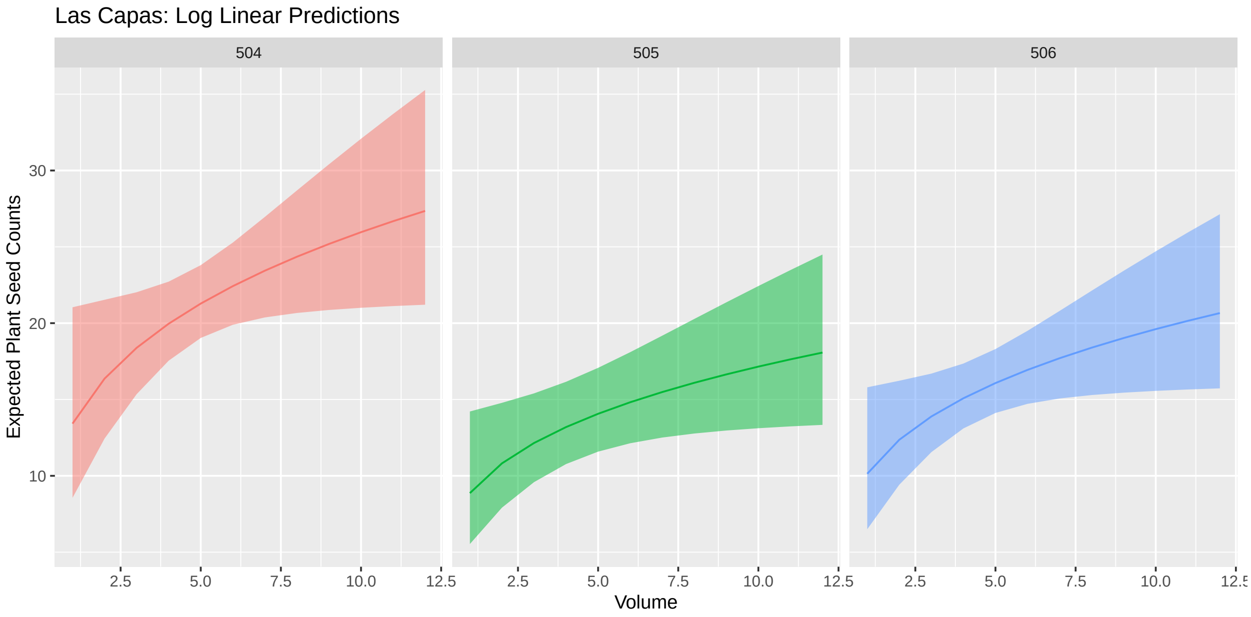

Figure 10: Log Linear Predictions for the Las Capas Data Set

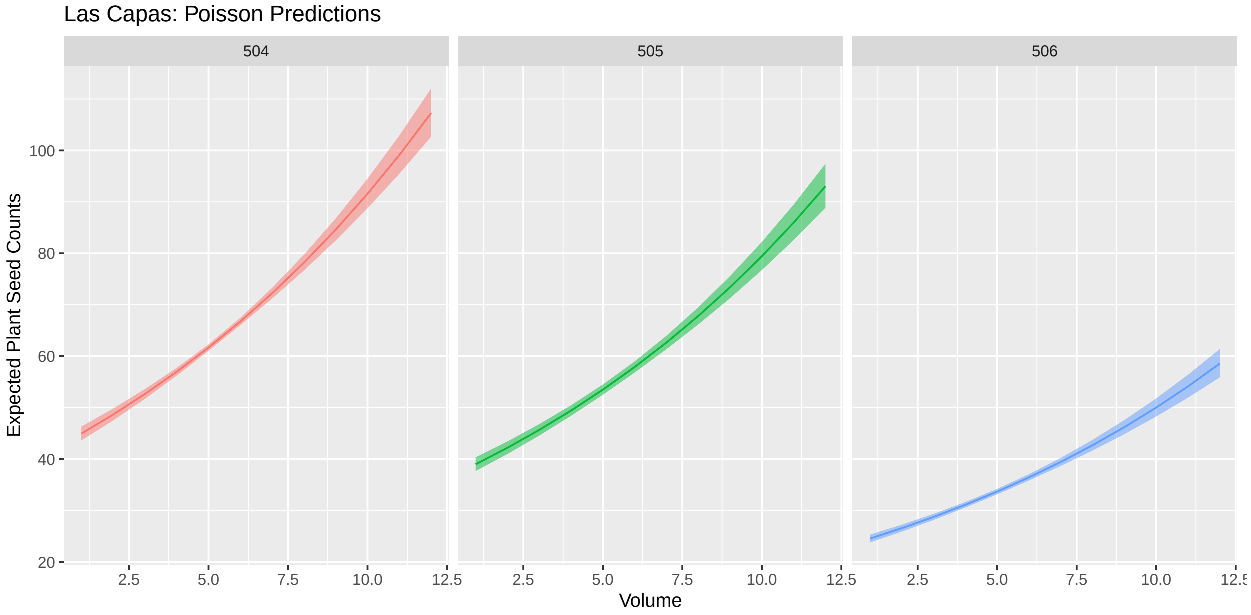

Figure 11: Poisson Predictions for the Las Capas Data Set

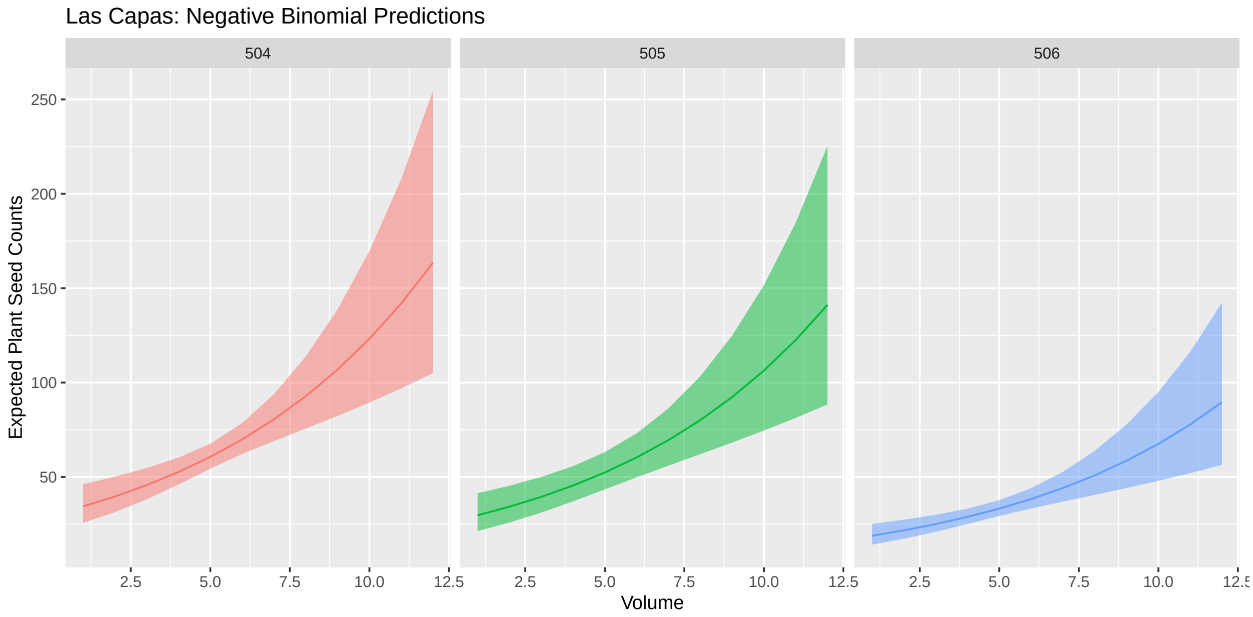

Figure 12: Negative Binomial Predictions for the Las Capas Data Set

5 Discussion

As seen above in the Results, Period is an extremely important predictor, as it is significant and therefore accounts for a large amount of the variation in the Plant Seed Counts for both the Dhiban and Las Capas sites. For example, this is seen in the Dhiban data's negative binomial model (**Figure 7**) where we see a strong increase in the Expected Seed Counts with the increase in Volume in the Nabataean-Roman, Middle Islamic I, and Midle islamic II periods, but very little increase in the Iron I, Iron II, and Late Antique Transitional periods.This can also be seen in the Las Capas data's negative binomial model (**Figure 12**) as well, which predicts a higher number of Plant Seed Counts for Period 504, a lower amount for Period 505, and an even lower amount for Period 506 as the Volume increases.

These archaeological plant remains from both Dhiban and Las Capas reveal important information about the day to day routine of past humans (Van der Veen 2007: 968). The fact that Period is a significant predictor for both the Dhiban and Las Capas sites reveals that the agricultural, social, or cultural practices of past humans may have varied from one time period to another in each given site (Farahani in press: 17-18). For example, past humans may have used or disposed of these plant remains differently in the different time periods, thus affecting how these plant remains are preserved in each time period (Lee 2012: 651).

After determining the importance of both Volume and Period as predictors and creating our various models, we are able to analyze each model's ability to predict the Plant Seed Counts for each data set. As seen above, the log linear and negative binomial models best fit the data. It is important to use these models over others, as the ability to predict our data well allows us to analzye trends in the data, thus better understanding the day to day lives of past humans.

From this study of the Dhiban and Las Capas data sets, we reveale the difficulties in modeling paleoethnobotanical count data, including the fact that the data is often skew right and overdispersed. We then predict the Plant Seed Counts using linear, log-linear, poisson, and negative binomial models to see which model or models were able to correclty address these issues in the data and best predict the plant seed counts. After creating and plotting these models, we then use the lrtest() function in R to compare the four models and find which model performs the best on the given data. From this test, we observe the log linear models and negative binomial models best fit both of our data sets. Next, we compare the residuals to the predicted counts for the log linear and negative binomial models of each data set, ultimately concluding that the best fitting model depends on the underlying data generating mechanism, and other factors. The log linear and negative binomial models' ability to fit the data will be a useful tool in predictive modeling for paleoethnobotanical data sets as a whole, allowing us to better understand how both Volume and Period affect the number of Plant Seed Counts in a given area. Through this modeling, we are able to make conclusions about the day to day activities of humans at these given locations and time periods, thus gaining a stronger understanding of our past.

6 Works Cited

Colin Cameron, A. and P. Trivedi (2013). Regression Analysis of Count Data. Cambridge: Cambridge University Press.

Farahani, A. (2018). A 2500-Year Historical Ecology of Agricultural Production under Empire in Dhiban, Jordan. Journal of Anthropological Archaeology (52), 137–155.

Farahani, A. (in press). Paleoethnobotany and Ancient Agriculture. In T. Howe and D. Hollander (Eds.), A Companion to Ancient Agriculture. New York: Wiley-Blackwell.

Ives, A. R. (2015). For testing the significance of regression coefficients, go ahead and log-transform count data. Methods in Ecology and Evolution 6 (7), 828–835.

Lee, G.-A. (2012). Taphonomy and sample size estimation in paleoethnobotany. Journal of Archaeological Science 39, 648–655.

Lindén, A. and S. Mantyniemi (2011). Using the negative binomial distribution to model overdispersion in ecological count data. Ecology 92 (7), 1414–1421.

Marston, J. M. (2014). Ratios and Simple Statistics in Paleoethnobotanical Analysis. In J. M. Marston, J. d'Alpoim Guedes, and C. Warinner (Eds.), Method and Theory in Paleoethnobotany, pp. 163–79. Boulder: University Press of Colorado.

O'Hara, R. B. and D. J. Kotze (2010). Do not log-transform count data. Methods in Ecology and Evolution 1 (2), 118–122.

Popper, V. S. (1988). Selecting quantitative measurements in paleoethnobotany. In C. A. Hastorf and V. S. Popper (Eds.), Current Paleoethnobotany: Analytical Methods and Cultural Interpretations of Archaeological Plant Remains, pp. 53 – 71. Chicago: University of Chicago Press.

Sinensky, R. J. and A. Farahani (2018). Diversity-Disturbance Relationships In The Late Archaic Southwest: Implications For Farmer-Forager Foodways. American Antiquity 83 (2), 281–301.

Van der Veen, M. (2007, June). Formation processes of desiccated and carbonized plant remains - the identification of routine practice. Journal of Archaeological Science 34 (6), 968–990.

Ver Hoef, J. M. and P. L. Boveng (2007). Quasi-Poisson vs. negative binomial regression: how should we model overdispersed count data? Ecology 88 (11), 2766–2772.

Zero-Inflated Negative Binomial Regression | R Data Analysis Examples. UCLA: Statistical Consulting Group. from https://stats.idre.ucla.edu/r/dae/zinb/ (accessed June 5, 2019).

Zero-Inflated Poisson Regression | R Data Analysis Examples. UCLA: Statistical Consulting Group. from https://stats.idre.ucla.edu/r/dae/zip/ (accessed June 5, 2019).

Zuur, A. F. et al. (2011). Generalised Linear Modelling. Analysing Ecological Data, Springer, 79-96.