Math 558 Final Project

Authors:
Anthony RORING
Jia li DONG
Wen CUI

May 10, 2019

Abstract

This project focuses on exploring and establishing a model structure for deciding which variables analyzed are important in determining the average number of ticks on red grouse chicks and which contribute to the variability of the number of ticks. Specifically, there are two variable groups, one is fixed variables, and one is random variables. We implemented different mixed models and settled on a log transformed mixed model without any interaction terms. The fixed effects are both significant and random effects explained over half of the variability of the number of ticks.

Introduction

In ecology, the most widely used statistical approach for aggregation of parasites among hosts is the negative binomial (Poisson-gamma) distribution. While in *Analysis of aggregation, a worked example: Number of ticks on red grouse chicks*[1], the authors choose to use Poisson-lognormal model because the approach has the advantage that it can be fitted as a generalized linear mixed model, thereby quantifying the sources of aggregation in terms of both fixed and random effects[1].

The authors take a regression approach to determine the influence each of the variables have on the number of ticks found on the red grouse chicks (chicks). This is different from the approach we took as our area of study utilizes the mean comparisons and analysis of variance. The authors did put together six models that analyze the variances of the random effects by including/excluding many of the variables to see how this affected the variance within each model. None of these models contained an interaction effect.

The baseline model only includes brood and individual chick nested in brood as random effect. Each of the following models includes this baseline model and adds other variables to create a new model. Model two adds year as fixed effect. Model three adds altitude as the only fixed effect. Model four has both year and altitude as fixed effect. Model five includes year as the only fixed effect and adds location to the random effect. Model six has both year and altitude as fixed effects and includes location as a random effect.

Compared to the baseline model, the model with year, altitude and location effects significantly reduced the variance of brood. However, the variance of the chick nested in brood remains the same. Hence they conclude that the year effect and altitude effect contributes most of the variance of brood while none of the variance of chick nested in brood was explained by those two effects.

In this project, we aim to evaluate several approaches on *grouseticks* dataset and find a good model that measures the importance of the fixed effects included as well as determining which random effects contribute the most to the variability of the number of ticks.

Data and Exploratory Data Analysis

The *grouseticks* dataset contains 5 variables: brood, year, index (chick), location, and height (grouped into three buckets), and one response variable: ticks. Below is a summary of each of the variables.

- Tick is the response variable. It represents the total number of ticks found on each chick's head. This ranged from zero to 85 with an average of a little over 6 ticks.
- Height is a fixed effect that represents the altitude in which each chick was caught, ranging from 403 to 533 meters. Because this was a continuous variable, we divided height into three groups called grouped heights. Each group represents a 50 meter range. For example, each chick that was caught in an altitude ranging from 400-450 meters was classified as Group one. Group two ranged from 450-500 meters and Group three from 500-550 meters. Therefore, grouped height has three levels: Group 1, Group 2, and Group 3. Height is considered a fixed effect because we are interested in determining whether the mean number of ticks in one of the height groups is different from another.

- Index represents a chick. Each chick was examined and the number of ticks were counted. A total of 403 chicks were caught from 118 broods from 63 locations over a three year period. Index is considered a random effect that is nested within brood.
- Brood represents a family of chicks. Brood is considered a random effect because we are interested in understanding how a brood effects the variability of the number of ticks on a chick. Brood has 118 levels.
- Location is the geographic location where the chick was caught. Location is considered a random effect because we are interested in understanding how the location in which a chick was caught effects the variability of the number of ticks on a chick. Location has 63 levels.
- Year represents a the years: 1995, 1996, and 1997. The data was gathered over the period of three years. Year is considered a fixed effect because we are interested in determining whether the mean number of ticks in a certain year is different from another year. Year has three levels: 1995, 1996 and 1997.

The summary statistics of *grouseticks* is presented in Table 1.

Variable	**Number of Records**	**Mean**	**Std Dev**	**Minimum**	**Maximum**
Ticks	403	6.37	13.14	0	85
Height	403	462.24	35.96	403	533
Grouped Height	403	1.74	0.72	1	3
Brood	403	625.22	78.44	501	743
Year	403	96.03	0.78	95	97
Index	403	202.00	116.48	1	403
Location	403	30.68	18.75	1	63

Table 1: Summary Statistics of *grouseticks*

In the histogram (Figure 1), we can see that the data is very skewed to the right. Most of the ticks found per chick range from zero to 20, but there are a few chicks with more than 75 ticks on them.

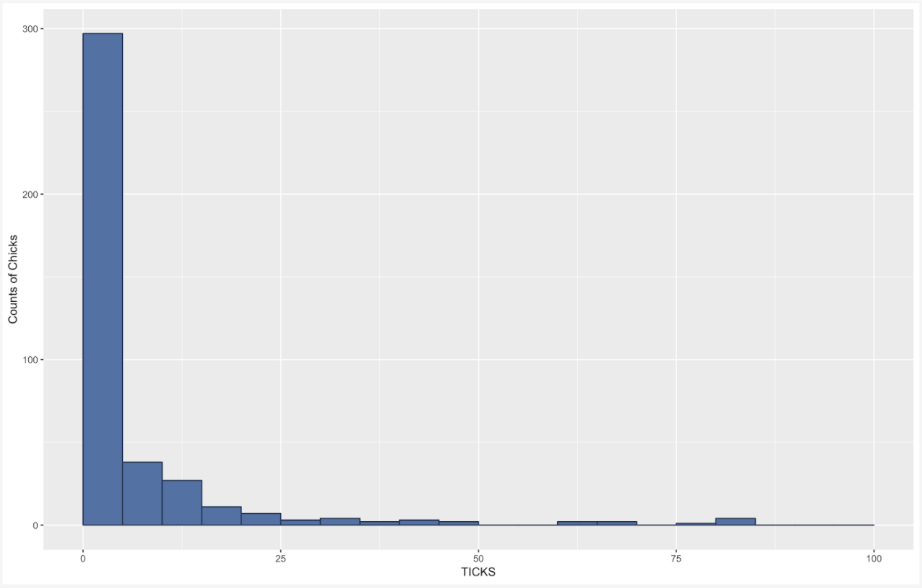

Figure 1: Frequency Histogram of Ticks

We also created a boxplot of ticks found on chicks over the three years period shown in Figure 2. The plot shows that, generally, the number of ticks found on the chicks in 1996 was much higher than in the other two years. In 1997, the number of ticks found was quite small. It suggests that the factor year may have play an important effect on the numbers of ticks.

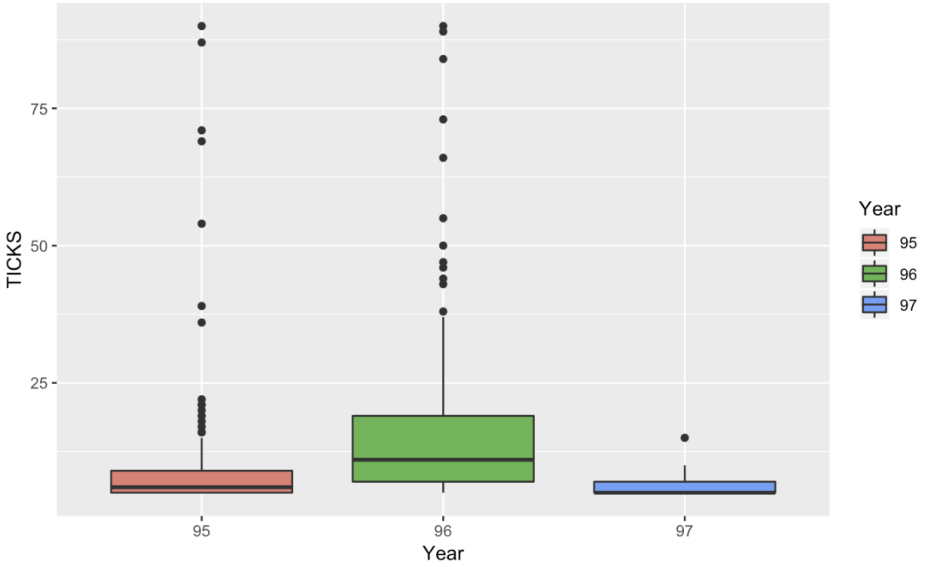

Figure 2: Boxplot of Ticks over 3 years

For the fixed effects, year and height, the goal was to determine whether these variables had an impact on the mean number of ticks found on the chicks. The null hypotheses for the year is the that the mean number of ticks does not change year over year. Whereas the alternative hypothesis is that at least one mean tick count in one year is different from the others.

The null hypothesis for the height is that the mean number of ticks does not change with the altitude in which the chick is living. Whereas the alternative hypothesis for height is that at least one mean tick count for a grouped height is different from the others.

For the random effects, brood, index, and location, we are interested understanding how much they contribute to the overall variance of the number of ticks. The null hypothesis for each of these effects is that the variance each contributes is zero or negligible. Whereas the alternative hypothesis is that an effect (brood, index, or location) contributes to a good portion of the overall variance.

Methodology

In order to run our analysis, we wanted to ensure the assumptions that the errors are independent and normally distributed with mean zero and variance sigma squared. However, after looking over the residual plot and a Q-Q plot, we deemed it is necessary to transform the data in order to make these assumptions.

In the residual plot (Figure 3), we see that moving from left to right, the data seems to spread out. This is a good indication of heteroscedasticity, and transformations can often help the data to be more homoscedastic.

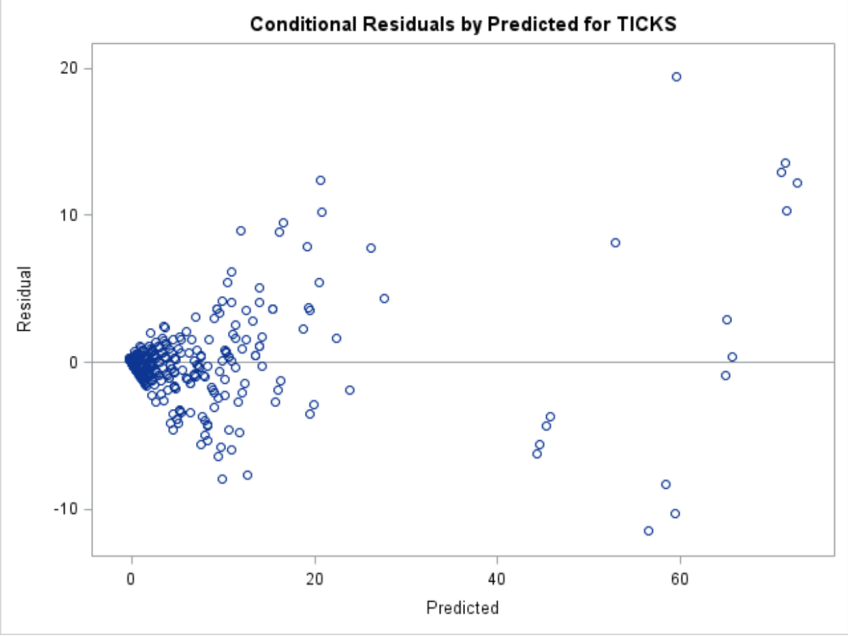

Figure 3: Conditional Residuals by Predicted for Ticks

The Q-Q plot (Figure 4) shows that the data is very skewed and does not follow the normal distribution very well.

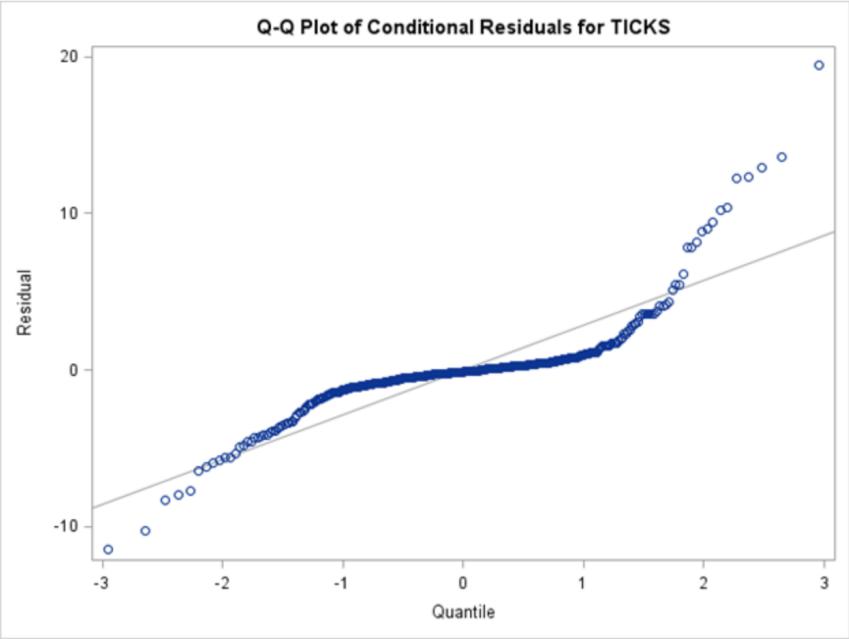

Figure 4: Q-Q Plot of Conditional Residuals for Ticks

A log transformation will often help with both skewed and non-normal data. The response variable in this dataset, the number of ticks, ranges from zero to 85. In order to avoid taking the log of zero and ending up with negative infinity for several of the responses, we shifted the data. This was done by simply adding one to each response. So a chick that previously had no ticks would now have one tick. Then by taking the log of this shifted response, a chick with no ticks originally, now has the log of one tick, which is actually zero. By making this shift, we are able to keep the relative difference between tick count among the chicks the same while also transforming the data.

Note that with this shift in the data, the variability will remain the same, however the mean number of ticks will increase slightly. Because we are taking the log of the shifted data, this increase is minor and will be ignored.

The resulting residual plot of the log of the shifted ticks is presented in Figure 5. Moving from left to right, there does still appear to be some heteroscedasticity, specifically when the residuals are below zero and the predicted values are between zero and one and between two and four. However, this looks much better than the non-transformed data, and the assumption of homoscedasticity is generally met.

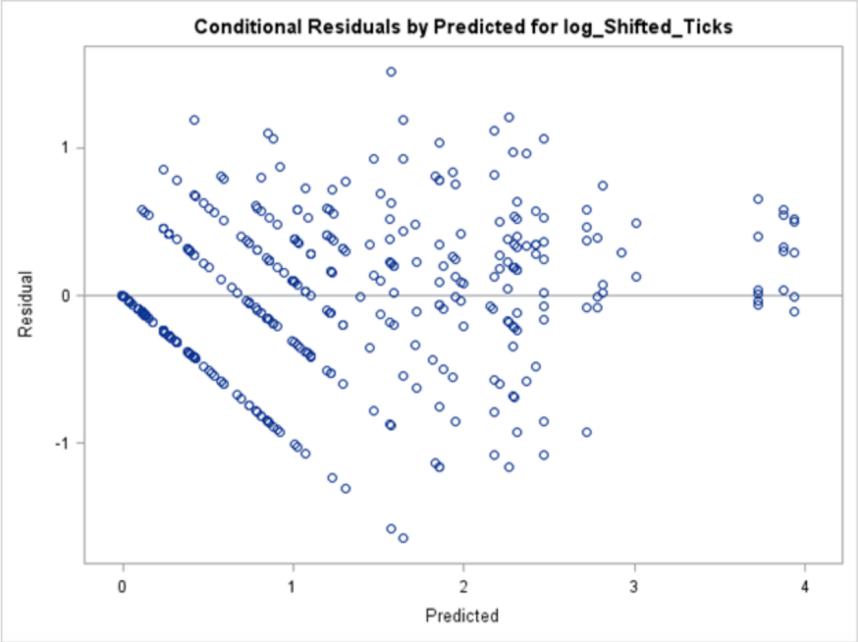

Figure 5: Conditional Residuals by Predicted for log transformed Ticks

The Q-Q plot of the log of the shifted tick count is presented in Figure 6. This plot looks much better than the non-transformed data, and we feel confident that our assumption of normality is met.

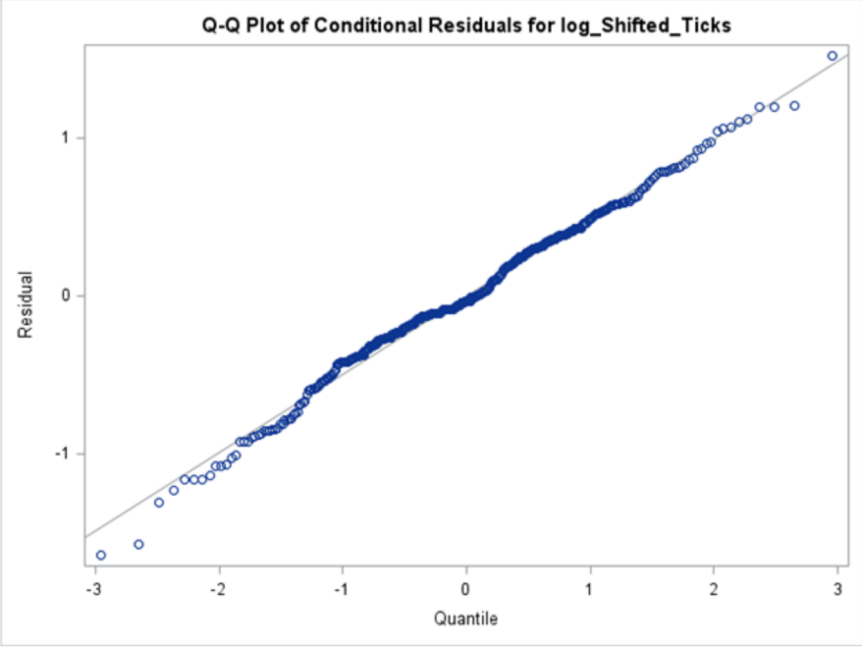

Figure 6: Q-Q Plot of Conditional Residuals for Log transformed Ticks

We analyzed several models to determine which one would best fit this data utilizing the designs we learned over the course of the semester. These models are each discussed below with the last model being the one we settled on for our analysis.

Interaction with the Fixed Effects

The first approach was to include all variables (height, year, index, brood, and location) along with an interaction term between the two fixed effects, height and year. After running this model, the interaction was not significant at an alpha level of .05 (Table 2), but would have been at an alpha level of .1. This made us contemplate whether or not to include this term in the final model.

Effect	Num DF	Den DF	F Value	Pr> F
Grouped Height	2	48.5	19.38	<.0001
Year	2	108	24.53	<.0001
Grouped Height*Year	4	106	2.30	0.0632

Table 2: ANOVA table for fixed effects

After looking at an interaction plot between year and height (Figure 7), we feel that this does not need to be included. The interaction plot shows that as we move from year 1995 to 1996 (left to right), the slope is increasing in each case, although the slope for the grouped height one is smaller, nearly level. Then moving from 1996 to 1997, all of the slopes are decreasing with grouped height one having the largest negative slope. Grouped height one is likely why the p-value is showing up to be borderline significant. But due to the p-value not actually being significant at a .05 alpha level and the evidence from this plot, we did not include this interaction term.

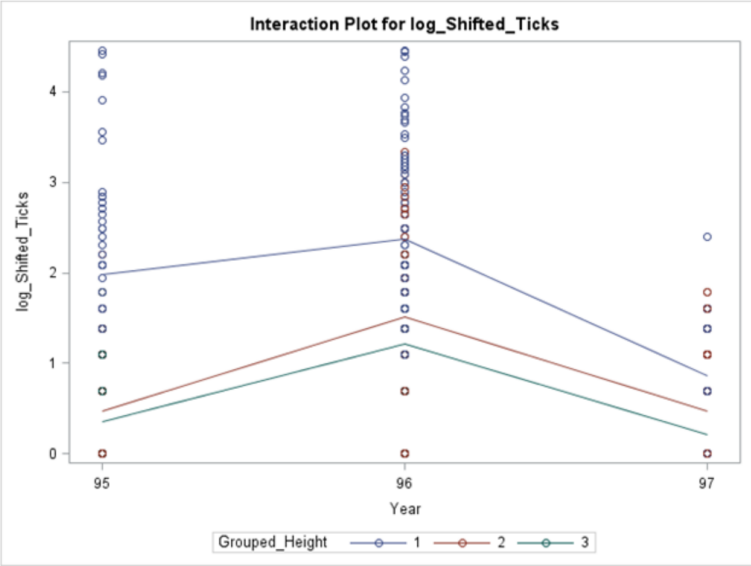

Figure 7: Interaction plot for Log transformed Ticks

Interaction with the Random Effects

We tried including a couple of interactions between a few of the random effects. The first was an interaction between the index (chick) and location. This interaction resulted in contributing to none of the variance. Therefore, this interaction was not included in our model.

We also tried an interaction between location and brood. Table 3 lists the estimate of each effect and its contribution to the overall variance. We see that the interaction of location and brood makes up a little over ten percent of the overall variance. While this is not zero, it is still pretty small when compared with the other effects. In an effort to keep the model simple, we felt that this interaction was not necessary to include.

Effect	Estimate	Percent of Variance
INDEX(BROOD)	0	0.0%
BROOD	0.1875	25.9%
LOCATION	0.1458	20.1%
LOCATION*BROOD	0.07653	10.6%
Residual	0.314	43.4%

Table 3: Estimate of random effects

How to Include Height in the Model

Because the variable height is continuous and we are not running a regression model, we had to break it up into different groups to include it in our ANOVA analysis. But how many groups should we use? Because height ranged from 400-550 meters, it naturally breaks into three groups, and that seemed like a good place to start.

Understanding that an arbitrary cutoff for the groups could potentially mask what is happening in the data, the more groups we include, the easier it would be to spot differences the height effects have on the number of ticks found on each chick. However, there were fewer chicks sampled at the higher altitude range. Because of this, if we grouped too granularly, we would run into an issue of having some groups from higher altitudes sparsely represented. We also wanted to ensure that each group contained the same length in height so as not to introduce any kind of bias based on having varying lengths for the different groups. This forced us to find a balance between having enough groups to not mask the effect of height, but also not so many that some groups were not well represented from the sample.

It turns out that three groups seemed to provide a good balance. Table 4 shows the frequencies for the three groups, group one being in the lowest altitude of 400-450 meters. We see that the higher altitude group, group three, only has 66 observations, whereas the other two groups have almost triple this number.

Grouped Height	Frequency
1	171
2	166
3	66

Table 4: Frequency for 3 height groups

Final Model

We used a mixed model with 3 random effects and 2 fixed effect. One random effect, index or chick, was nested within another random effect, brood. The following is the model we used:

$$\log(Y_{ijklmn}) = \mu + H_i + Z_j + L_k + B_l + I(B)_{(l)m} + \varepsilon_{ijklmn}$$

Where,

μ: Overall mean of the number of ticks

H: Height altitude where the chick was caught (grouped into 3 buckets of 50 meters), fixed effect

Z: Year in which a chick was caught and examined, fixed effect

L: Location in which the chick was caught and examined, random effect

B: Brood to which a chick belongs, random effect

I: Index or chick that was caught and examined (this is nested within brood), random effect

ε: Remaining error

Results

Fixed Effects

By taking the shifted log of the response variable, ticks, we believe the assumptions of normality and homoscedasticity are met. Therefore, the results of the model may be relied upon. From our model, we found that both the height and year had p-values less than .0001, see Table 5. This is strong evidence to reject the null hypothesis that height or year do not play a significant role in the mean number of ticks found on a chick in favor of the alternative hypothesis that the elevation where the chicks live and the particular year in which a chick lived play a significant role in the mean number of ticks that are found on each chick.

Source	DF	Den DF	F Value	Pr>F
Height	2	52.6	18.02	<.0001
Year	2	114	28.18	<.0001

Table 5: ANOVA table for fixed effects

Since both the height and year are significant effects, we ran pairwise comparisons for each to see which effect level, specifically, is different from the others. Note that because we shifted the data by one and took the log transformation, the estimates found below are estimating each level's contribution to the shifted log of the total number of ticks found on each chick rather than just each level's contribution to the total number of ticks found on each chick.

Table 6 includes each effects estimate and associated precision measures. Without yet looking at the pairwise differences, we see that group height level one has an estimated value of 1.7, which is much higher than levels two and three. Year 1996 looks to have an estimated value decently higher than 1995, and much higher than 1997.

Effect	Grouped Height	Year	Estimate	Standard Error	DF	t Value	Pr$>\|t\|$
Grouped Height	1		1.7046	0.1275	49	13.37	$<.0001$
Grouped Height	2		0.7340	0.1188	51.6	6.18	$<.0001$
Grouped Height	3		0.6286	0.1888	55.8	3.33	0.0015
Year		95	0.9360	0.1272	107	7.36	$<.0001$
Year		96	1.7027	0.1215	104	14.02	$<.0001$
Year		97	0.4276	0.1265	102	3.38	0.0010

Table 6: Effect estimate and corresponding precision measure

Table 7 looks specifically at the differences between the three levels of grouped height and the three levels of year. The p-values were adjusted using the Bonferroni adjustment. For grouped height, we see that level one is significantly different from level two and level three. However, the difference between level two and level three is not significant. Level one is the elevation between 400 and 450 meters. In the above table, we see that the estimate for this level was about 1.7, over double what the other two levels were estimated to be. The conclusion we draw from this is that chicks living at a lower altitude tend to have more ticks on average.

Looking at the differences between the years in the table below, we see that the year 1996 has a p-value less than .0001 when comparing against 1995 and 1997. This is very strong evidence that this year is significantly different from the other two years. We also note that the difference between 1995 and 1997 also had a small p-value of .0055, which is also strong evidence that there is a significant difference between these two years. Therefore, each year in which the data was collected played a significant role in determining the mean number of ticks. We see from the table above, the estimate for 1996 was the highest at about 1.7. In general more ticks were found on chicks in this year as compared to 1995 where the estimate was about .9 or to 1997 where the estimate is about .4.

Effect	Grouped Height	Year	Grouped Height	Year	Estimate	Std Error	DF	t Value	Pr$>\|t\|$	Adj	Adj P
Grouped Height	1		2		0.9706	0.1747	50.3	5.56	$<.0001$	Bon	$<.0001$
Grouped Height	1		3		1.0760	0.2281	53.8	4.72	$<.0001$	Bon	$<.0001$
Grouped Height	2		3		0.1054	0.2230	54.6	0.47	0.6384	Bon	1.0000
Year		95		96	-0.7677	0.1592	114	-4.82	$<.0001$	Bon	$<.0001$
Year		95		97	0.5084	0.1592	110	3.19	0.0018	Bon	0.0055
Year		96		97	1.2761	0.1562	108	8.17	$<.0001$	Bon	$<.0001$

Table 7: Pairwise comparison for fixed effects

Random Effects

We analyzed 3 random effects to see how much each contributed to the total variability of the number of ticks found on the chicks. The results of the model indicate that the chick, itself, does not contribute to the variability. This is a good indication that the chicks are homogeneous. The location contributed to 20.1% of the total variability. The brood contributed to 36.5% of the total variability, which is a large portion. The residuals made up 43.4% of the total variability. All in all, over half of the variability came from the brood and location. So the brood into which a chick was born and the location in which it was raised play a significant role in the number of ticks it contracts. Table 8 shows a summary of the variables and their contribution to the total variability.

Variable	Estimate	Percent of Variability
LOCATION	0.1458	20.1%
BROOD	0.2641	36.5%
INDEX(BROOD)	0	0.0%
Residual	0.314	43.4%
Total	0.7239	100.0%

Table 8: Estimate of random effects

Discussion/Conclusion

The model used in the previous research done by Elston fits the assumption necessary to analyze counts in ecology, and the Poisson-lognormal distribution accounts for overdispersion beyond Poisson distribution. In this paper, we used an analysis of variance approach to test for the significant random effects and fixed effects, and we succeeded in identifying those factors.

Because our approach is quite different from the authors approach, it is difficult to make direct comparisons between the models. However as discussed earlier, in Table 1 of their paper, they do analyze the variance components of six models. The author's model six is close to the model we used in the fact that it includes the same variables, both fixed and random, as we included in our model. From this standpoint, we can make a comparison of which variables had high variability in our model and which had a high variability in their model.

Excluding the residuals In our model, brood had the highest contribution to the overall variability, and it also had the highest variance component in their model. Location had the next highest contribution to the overall variability in our model, and it had the second highest variance component in their model. With this rough comparison, we can see that we are coming to a similar conclusion for these variables.

Other similar prior research aside from Elston detailed using Poisson-lognormal distribution in other ecological analysis, but did not conduct analysis on the same data as Elston's research.

The final model we reached, which is the mixed model with 3 random effects and 2 fixed effect, is obtained using a process of adding all relevant factors and effects, then eliminating irrelevant factors from the model along with irrelevant interaction effects (both fixed and random) that are not significant in determining the

variability of the ticks data as a random effect nor significant as a fixed factor.

As far as other changes or conclusions we could come to, we would be interested in taking a non-parametric approach because of its convenience in that we would not have to make any underlying assumptions of normality or homoscedasticity. If we truly felt that the homoscedasticity assumption was not sufficiently met, this would be a good approach to see what conclusions we could come to and whether these conclusions are similar to what we had in our results. If the conclusions were quite different, it would lead us to believe that perhaps the assumptions we made were just too far off. However, if we came to the same conclusions, it would help support the results of our study and the conclusions we drew.

References

[1] Elston, D., Moss, R., Boulinier, T., Arrowsmith, C., Lambin, X. (2001) *Analysis of aggregation, a worked example: Numbers of ticks on red grouse chicks*

1 Appendix

```
/*Final Project*/

proc import out = grouse
datafile = 'grouseticks.csv'
dbms = csv replace;
run;

proc contents data = grouse varnum;run;

data grouse2;
set grouse(rename = (INDEX = INDEX_temp BROOD = BROOD_temp
LOCATION = LOCATION_temp Year = Year_temp));
if HEIGHT < 450 then Grouped_Height = 1;
else if HEIGHT < 500 then Grouped_Height = 2;
else if HEIGHT < 550 then Grouped_Height = 3;
/*Check if brood number is unique by year*/
Unique_Brood = cat(BROOD_temp,Year_temp);
INDEX = INDEX_temp*1;
BROOD = BROOD_temp*1;
LOCATION = LOCATION_temp*1;
Year = Year_temp*1;
Shifted_Ticks = Ticks + 1;
log_Shifted_Ticks = log(Shifted_Ticks);
drop INDEX_temp BROOD_temp LOCATION_temp Year_temp;
run;

proc freq data = grouse2;table Grouped_Height;run;
proc freq data = grouse2;table Unique_Brood/out = check;run;
proc sort data = check;by Unique_Brood;run;
/*BROOD looks to be unique by year*/
/*500s for 95*/
/*600s for 96*/
/*700s for 97*/

proc contents data = grouse2 varnum;run;

proc means data = grouse2;
var TICKS HEIGHT Grouped_Height BROOD YEAR INDEX LOCATION;
run;

/*No transformation*/
proc mixed data=grouse2 cl plot=residualpanel (unpack);
class Grouped_Height YEAR LOCATION BROOD INDEX;
model Ticks=Grouped_Height YEAR/ddfm=kr;
```

```
random LOCATION BROOD INDEX(BROOD);
run;
quit;

/*With Shifted Log transformation -- FINAL MODEL USED*/
proc mixed data=grouse2 plot=residualpanel (unpack);
class Grouped_Height YEAR LOCATION BROOD INDEX;
model log_Shifted_Ticks=Grouped_Height YEAR/ddfm=kr;
random LOCATION BROOD INDEX(BROOD);
lsmeans Grouped_Height YEAR/pdiff=all adjust=BON;
run;
quit;

/*For interaction plot*/
ods graphics on;
proc glm data=grouse2 plot=diagnostics (unpack);
class INDEX YEAR Grouped_Height;
model log_Shifted_Ticks=Grouped_Height YEAR Grouped_Height*YEAR;
run;
quit;
ods graphics off;


/*With interaction of the fixed effects, height & year*/
proc mixed data=grouse2 cl plot=residualpanel (unpack);
class Grouped_Height YEAR LOCATION BROOD INDEX;
model log_Shifted_Ticks=Grouped_Height|YEAR/ddfm=kr;
random LOCATION BROOD INDEX(BROOD);
run;
quit;

/*With interaction of the random effects, index & location*/
proc mixed data=grouse2 cl plot=residualpanel (unpack);
class Grouped_Height YEAR LOCATION BROOD INDEX;
model log_Shifted_Ticks=Grouped_Height YEAR/ddfm=kr;
random BROOD INDEX(BROOD)|LOCATION;
run;
quit;

/*With interaction of the random effects, brood & location*/
proc mixed data=grouse2 cl plot=residualpanel (unpack);
class Grouped_Height YEAR LOCATION BROOD INDEX;
model log_Shifted_Ticks=Grouped_Height YEAR/ddfm=kr;
random INDEX(BROOD) BROOD|LOCATION;
run;
quit;
```