BIOL 520 Assignment 2

Portiaa McGonigal & Sarah Lyons

10/28/2019

library(readxl)  
library(ggplot2)  
library(nlme)  
library(lme4)  
library(tidyverse)  
library(tigerstats)  
library(visreg)  
library(emmeans)  
library(lmerTest)

The dataset we selected to analyze was obtained through Dryad (<https://datadryad.org/stash>). We chose to analyze raw data from “Lower bumblebee colony reproductive success in agricultural compared with urban environments” (Samuelson et al., 2018). In this study, the authors investigated whether the degree of urbanization impacted the reproductive success of bumblebees (*Bombus terrestris*) by monitoring the production of sexual offspring in separately-located colonies.

To monitor this, the researchers reared colonies from wild-caught bumblebee queens and placed them in different environments with varying degrees of urbanization and agricultural development. The colonies were monitored for many different variables in order to assess reproductive success and overall survivorship.

The hypotheses we are testing are:

1. the level of urbanization will have an effect on the reproductive success of bumblebee colonies.
2. parasitic invasion will decrease the reproductive success of bumblebee colonies, resulting in fewer sexual offspring being produced.

The variables we selected from the data are:

1. The number of reproductive individuals produced in each colony **(Tot\_rep)**, which includes the total number of males (Tot\_male) and total number of gynes (sexually reproductive female caste of insects, Tot\_gyne). This discrete fixed variable is our chosen response variable.
2. The 500m radius clustered land-use (surrounding the colony) variable **(LU500)**, a categorical fixed factor that includes three different categorical levels: Agriculture, City, and Village. This variable was chosen as one of our explanatory variables.
3. Whether *Bombus vastalis* invaded the colony or not **(Cu\_bin)**, a binary categorical variable (two levels, yes-1, or no-0). This was chosen as our second explanatory variable.

First we must read in and examine the data:

ab <- read\_excel("Ubran\_Bumblebee\_ColonyData.xlsx", sheet = "Ubran\_Bumblebee\_ColonyData")

## New names:  
## \* `` -> ...13

ab$Cu\_bin <- as.factor(ab$Cu\_bin)  
str(ab)

## Classes 'tbl\_df', 'tbl' and 'data.frame': 38 obs. of 13 variables:  
## $ our\_colony\_id: num 11 12 18 23 32 1 3 4 5 8 ...  
## $ Col : chr "41" "43" "60" "76" ...  
## $ Site : chr "WIS" "LAR" "WEY" "WEL" ...  
## $ Countave : num 22.67 68 64 73.33 9.67 ...  
## $ LU750 : chr "Rural" "Rural" "Rural" "Rural" ...  
## $ LU500 : chr "Agricultural" "Agricultural" "Agricultural" "Agricultural" ...  
## $ LU250 : chr "Open" "Built\_up" "Open" "Open" ...  
## $ LU100 : chr "Fields" "Sparse\_housing" "Fields" "Fields" ...  
## $ Tot\_male : num 0 3 0 0 0 23 41 67 0 4 ...  
## $ Tot\_gyne : num 0 0 0 0 0 0 1 12 0 0 ...  
## $ Cu\_bin : Factor w/ 2 levels "0","1": 2 2 1 2 2 1 1 1 1 1 ...  
## $ Tot\_rep : num 0 3 0 0 0 23 42 79 0 4 ...  
## $ ...13 : num 0.6 NA NA NA NA ...

To visualize the data, we created a stripchart showing the relationship between the population of the reproductive class of each colony and land use in a 500 meter radius around each hive. We used ggplot for this step (don’t be mad we are doing our best).

ggplot(ab, aes(LU500, Tot\_rep)) +  
 geom\_jitter(color = "firebrick", size = 3, height = 0.2, width = .05, alpha = 0.5) +  
 labs(x = "Location Description", y = "Population of Reproductive Class") +   
 theme\_classic()

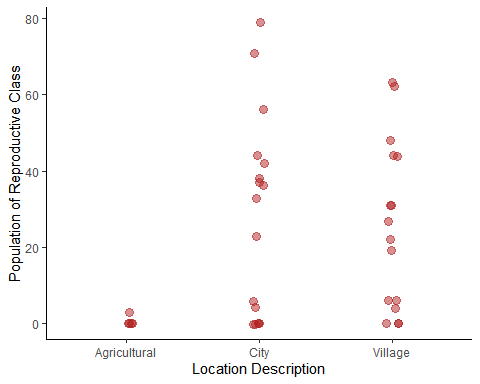


Figure 1: Strip chart showing numbers of individuals in reproductive class for each colony broken down by surroundings of the hive (500m radius, n = 38)

xtabs(~ LU500 + Cu\_bin, data = ab)

## Cu\_bin  
## LU500 0 1  
## Agricultural 1 4  
## City 17 0  
## Village 13 3

We also visualized the response varaible cateogorized by our secondary explanatory variable of *B. vastalis* invasion.

ggplot(ab, aes(as.factor(Cu\_bin), Tot\_rep)) +  
 geom\_jitter(color = "firebrick", size = 3, height = 0.2, width = 0.05, alpha = 0.5) +  
 labs(x = "B. vastalis invasion", y = "Population of Reproductive Class") +   
 theme\_classic()

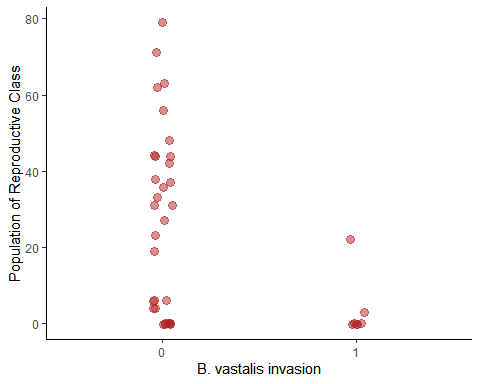


Figure 2: Strip chart showing numbers of individuals in reproductive class for each colony broken down by presence of *B. vastalis* (n = 38)

This data indicates that the presence of *B. vastalis* appears to reduce the reproductive success of the colonies.

Next, we set out to determine if our data is normally distributed using the hist( ) function to plot a histogram of the given values of the response variable.

par(mfrow = c(1,2))  
hist(ab$Tot\_rep, main = "Histogram A", ylab = "Frequency of Values", xlab = "Count of Reproducibles")  
hist(log(ab$Tot\_rep), main = "Histogram B", ylab = "Frequency of Values", xlab = "Count of Reproducibles (log scale)")

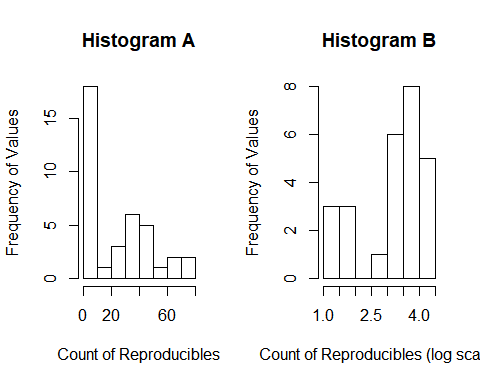


Figure 3: Histogram A shows the distribution of the counts of individuals in the reproductive class on the orignal scale. Histogram B is the same data on a log scale (n=38).

As seen in the above histograms, our count data is not normally distributed even after log transformation and as such we are not able to use a simple linear model. Therefore we will use a generalized linear model (glm) to examine the relationship between the counts of reproductive individuals in each colony, the type of land-use in a 500 meter radius surrounding the colony, as well as the presence of parastic invasion by B. vastalis. By using a glm, we are transforming the expected values as apposed to the response variable itself.

A poission distribution is likely appropriate to model the response variable because our data is count data of individuals in a colony. This is because the Poisson distribution will generate only whole numbers, in line with the counts of bees we are modeling. The Poisson distribution has only one parameter describing the mean and the variance, which is also its expected value. In this case, we would use the log function to link the transformed expected values to the linear model. The log function is the link function for the parameter of the poisson distibution.

The poisson distribution assumes that the varaince and the mean are equal within each treatment group. Therfore we need to test that the variance of our count data in each land-use category is similar to the mean.

mean(Tot\_rep ~ LU500 + Cu\_bin, data = ab)

## Agricultural.0 City.0 Village.0 Agricultural.1 City.1   
## 0.000000 27.588235 29.615385 0.750000 NaN   
## Village.1   
## 7.333333

var(Tot\_rep ~ LU500 + Cu\_bin, data = ab)

## Agricultural.0 City.0 Village.0 Agricultural.1 City.1   
## NA 684.8824 473.9231 2.2500 NA   
## Village.1   
## 161.3333

The assumptions are not met for either explanatory variable. Therefore we must use the quasi-poisson distribution. In this distribution, we calculate the relationship between the mean and the variance as apposed to a probabilty distribution. This produces a “quasi-likelihood” estimate similar to a maxiumum likelihood estimate.

glm\_1<-glm(Tot\_rep ~ LU500 + Cu\_bin, data =ab, family = quasipoisson(link="log"))  
  
summary(glm\_1)

##   
## Call:  
## glm(formula = Tot\_rep ~ LU500 + Cu\_bin, family = quasipoisson(link = "log"),   
## data = ab)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -7.6816 -5.2078 -0.6731 2.5306 7.9625   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 0.3759 2.5711 0.146 0.885  
## LU500City 2.9415 2.5790 1.141 0.262  
## LU500Village 3.0086 2.5711 1.170 0.250  
## Cu\_bin1 -1.3280 0.9158 -1.450 0.156  
##   
## (Dispersion parameter for quasipoisson family taken to be 19.16613)  
##   
## Null deviance: 1064.35 on 37 degrees of freedom  
## Residual deviance: 784.02 on 34 degrees of freedom  
## AIC: NA  
##   
## Number of Fisher Scoring iterations: 6

We can see from the estimates of the coefficients that in comparison to agriculturally predominant land-use, cities and villages have a positive effect on the reproductive success of colonies while B. vestalis infection has a negative effect.

Deviance values relate to the “goodness of fit” of the glm applied to the data. Higher values correspond to a worse fit and vice versa. The null deviance shows how well the response variable is predicted by a model that includes only the intercept (grand mean). The residual deviance shows that the inclusion of our independent variables decreased the deviance from 1064.35 on 37 degrees of freedom to 784.02 on 34 degrees of freedom, a large reduction. The Fischer’s scoring algorithm is derived from Newton’s algorithm for maximum likelihood. Our summary indicates that Fischer’s scoring algorithm required six iterations in order to perform the fit.

The dispersion parameter is the mean square error (the average squared difference between the estimated values and the actual values) of the glm, which in our data is quite high. This indicates within our treatments our means and variances are not equal. Therefore it is appropriate that we used a quasipoisson distribution and not a Poisson distribution.

To visualize the model, we will use the visreg() function.

visreg(glm\_1, by = "Cu\_bin", xvar = "LU500", main = "")

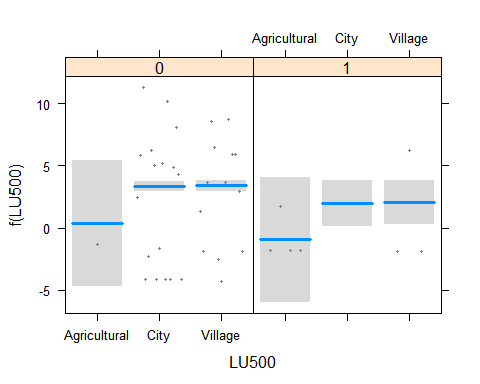


Figure 4: shows the estimated mean number of reproductive individuals in colonies for each treatment on the transformed log scale as well as working values for the response variable (n = 38).

visreg(glm\_1, by = "Cu\_bin", xvar = "LU500", scale = "response", rug = FALSE, main = "")

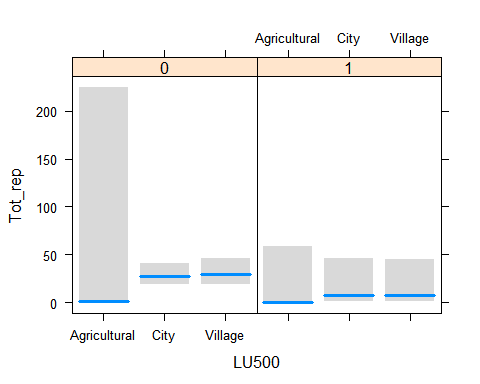


Figure 5: shows the estimated mean of reproductive individuals in colonies for each treatment on the original scale (n = 38). Standard errors are shown as grey bars around estimates.

In figure 4, the points are not the transformed data, but rather are “working values” for the response variable from the last iteration of model fitting, which glm() uses behind the scenes to fit the model on the transformed scale. The fitted values are transformed means that were estimated using the log scale. The estimated means in figure 5, as well as the fitted values below, are displayed in the original scale.

We can now use the emmeans() function to see our predicted values of the model for each of our indepenedent variables. Estimated marginal means (EMMs) refer to the mean response for each factor, adjusted for any other variables in the model. We then compared the EMMs with one another using the pairs() function.

ab.emm <- emmeans(glm\_1, c("LU500","Cu\_bin"), type = "response")  
ab.emm

## LU500 Cu\_bin rate SE df asymp.LCL asymp.UCL  
## Agricultural 0 1.456 3.74 Inf 0.00944 224.8  
## City 0 27.588 5.58 Inf 18.56311 41.0  
## Village 0 29.503 6.59 Inf 19.04460 45.7  
## Agricultural 1 0.386 0.99 Inf 0.00252 59.0  
## City 1 7.311 6.86 Inf 1.16337 45.9  
## Village 1 7.819 6.96 Inf 1.36675 44.7  
##   
## Confidence level used: 0.95   
## Intervals are back-transformed from the log scale

The rate is the coefficient values for each combination of our predictive variables. *Estimated marginal means (EMMs) refer to the mean response for each factor, adjusted for any other variables in the model.*

In order to test our hypothesis, we will conduct and an ANOVA test against the null hypothesis that there is no difference in the number of individuals in the reproductive class between hives located in agricultural, city and village areas.

anova(glm\_1, test="F")

## Analysis of Deviance Table  
##   
## Model: quasipoisson, link: log  
##   
## Response: Tot\_rep  
##   
## Terms added sequentially (first to last)  
##   
##   
## Df Deviance Resid. Df Resid. Dev F Pr(>F)   
## NULL 37 1064.35   
## LU500 2 220.706 35 843.64 5.7577 0.007022 \*\*  
## Cu\_bin 1 59.621 34 784.02 3.1108 0.086762 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Counts of reproductive classes in colonies differ significantly among land use areas (ANOVA, F = 5.75, n = 38, df = 35, p = 0.007). There does not however appear to be a significant affect of *B. vesalis* invasion.

We then compared the EMMs with one another using the pairs() function.

pairs(ab.emm)

## contrast ratio SE df z.ratio p.value  
## Agricultural,0 / City,0 0.0528 0.136 Inf -1.141 0.8645   
## Agricultural,0 / Village,0 0.0494 0.127 Inf -1.170 0.8512   
## Agricultural,0 / Agricultural,1 3.7734 3.456 Inf 1.450 0.6961   
## Agricultural,0 / City,1 0.1992 0.576 Inf -0.558 0.9936   
## Agricultural,0 / Village,1 0.1862 0.534 Inf -0.586 0.9920   
## City,0 / Village,0 0.9351 0.282 Inf -0.223 0.9999   
## City,0 / Agricultural,1 71.4849 184.022 Inf 1.659 0.5594   
## City,0 / City,1 3.7734 3.456 Inf 1.450 0.6961   
## City,0 / Village,1 3.5285 3.220 Inf 1.382 0.7381   
## Village,0 / Agricultural,1 76.4472 197.625 Inf 1.678 0.5467   
## Village,0 / City,1 4.0354 4.088 Inf 1.377 0.7408   
## Village,0 / Village,1 3.7734 3.456 Inf 1.450 0.6961   
## Agricultural,1 / City,1 0.0528 0.136 Inf -1.141 0.8645   
## Agricultural,1 / Village,1 0.0494 0.127 Inf -1.170 0.8512   
## City,1 / Village,1 0.9351 0.282 Inf -0.223 0.9999   
##   
## P value adjustment: tukey method for comparing a family of 6 estimates   
## Tests are performed on the log scale

This is a post hoc test showing that there is not a significant difference between the treatment combinations that include the *B. vesalis* infections. This is further evidence that the presence of *B. vesalis* in the colonies does not significantly effect the numbers of individuals in the repoductive class.

This lack of significance could also be due to lack of datapoints as our data did not include many samples without in Agricultural areas and most colonies did not have *B. vesalis* infections.

xtabs(~ LU500 + Cu\_bin, data = ab)

## Cu\_bin  
## LU500 0 1  
## Agricultural 1 4  
## City 17 0  
## Village 13 3

**Refrences:**

Samuelson, A., Gill, R., Brown, M., & Leadbeater, E. (2018). Lower bumblebee colony reproductive success in agricultural compared with urban environments. Proceedings Of The Royal Society B: Biological Sciences, 285(1881), 20180807. doi: 10.1098/rspb.2018.0807

Link to the data:

<https://datadryad.org/stash/dataset/doi:10.5061/dryad.c68cj62>

Link to full paper:

<https://royalsocietypublishing.org/doi/pdf/10.1098/rspb.2018.0807>)