Project 2

Alex Zhang

2022-11-25

Abstract

Finding how different environmental causes would affect species's appearance and number is critical on creating wild life reservations. In this project, I will try to use different techniques to analyze whether there is a relationship between the environmental variables and the number of species. I will use the data collected through a long tern research program runs by Wake Forest University professors in Serengeti Park. I applied Zero-Inflated-Poisson model (ZIP) to fit variables and the count for different species. I also used hypothesis tests to check whether there is evidence showing that the relationships actually exist. After all these processing, I got the result that for Thomson's gazelle and Zebra, the hypothesis tests indicate that overall there is strong evidence showing that there are relationships between environmental variables and the number of these species. For topi, the hypothesis test illustrate that there is weak or no evidence showing that there is a relation between environmental variables and the number of topi. This is probably that the number of topi is too small to be representative. In the discussion section, I also mentioned that the small size of the number of topi may be one of the limitation for this report.

Section 1: Introduction

The motivation for research question is checking how some environmental effects could impact whether species would present of the number of species with given recorded sites. The data set is gained by an long term program done by Wake Forest University professors in Serengeti Park in Tanzania. The data is collected by using cameras shooting motions and heat when animals passing by. The researchers then analyze the photos and classify different species and count the number of appearances in different sites. In this report, I will first do the exploratory data analysis on showing the number of different species and their relation between different environmental variables. Further I will then do hypothesis tests that check whether there is evidence showing there are some relationships between the environmental effect and the number of species. In this analysis, I used the ZIP model to fit the data, and through hypothesis tests, I know that p-value for gazelle are very small, but p-value for topi are relatively large. Later I will compare environmental effect on different species like how same variable could has different impact on different species. In the last, I will discuss some limitations for this report.

Section 2: Data

After trying to drop the missing data, it still exists 966 rows and 15 columns which indicate there is no missing data.

Based on the summary of the data, each row it represents the information for a site with specific date. Each row also contains the the count and appearance of certain species and some other environmental factors. There are total 966 rows and 15 columns. The data set contains information about the site id, recorded date, gazelle's count and appearance, topi count, zebra count, and other environmental factors like whether having wildfire or having enough vegetation.

Figure 2.1

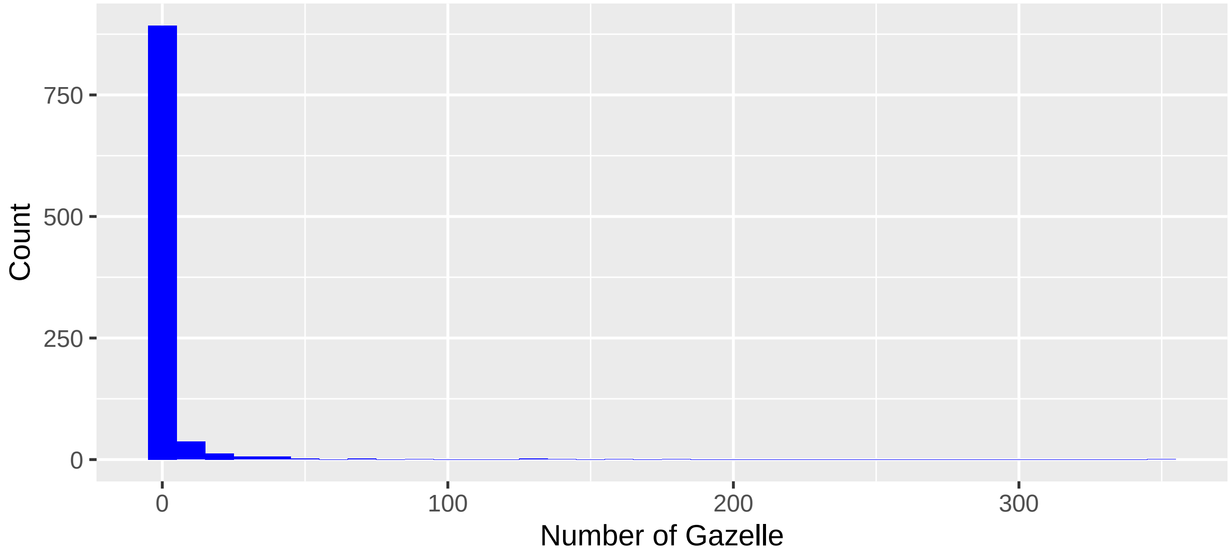

Figure 1: Thomsons's Gazelle's count

Figure 2.2

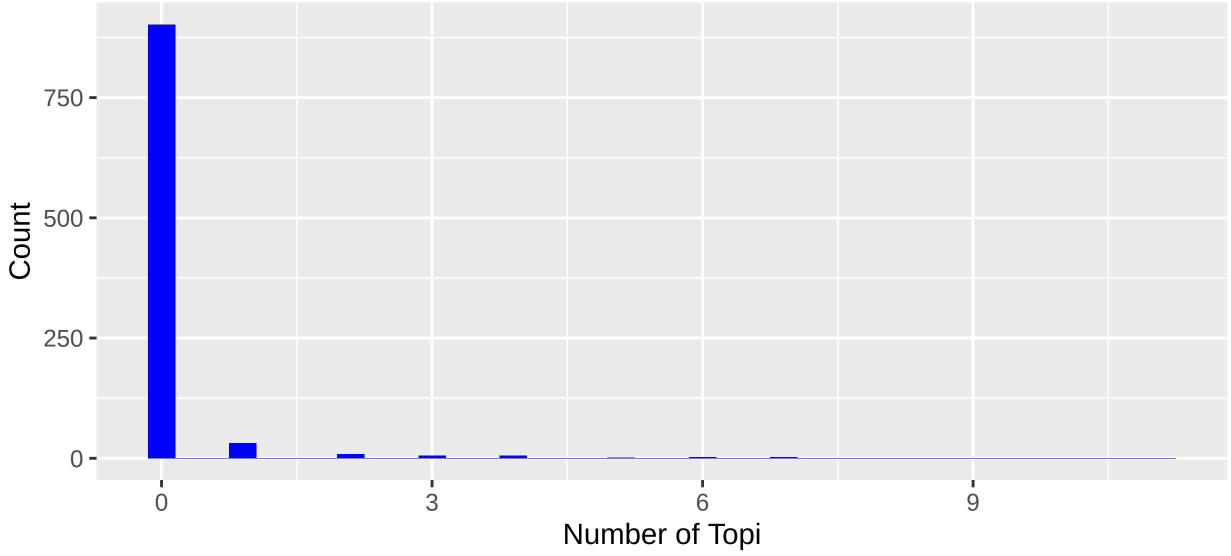

Figure 2: Zebra's count

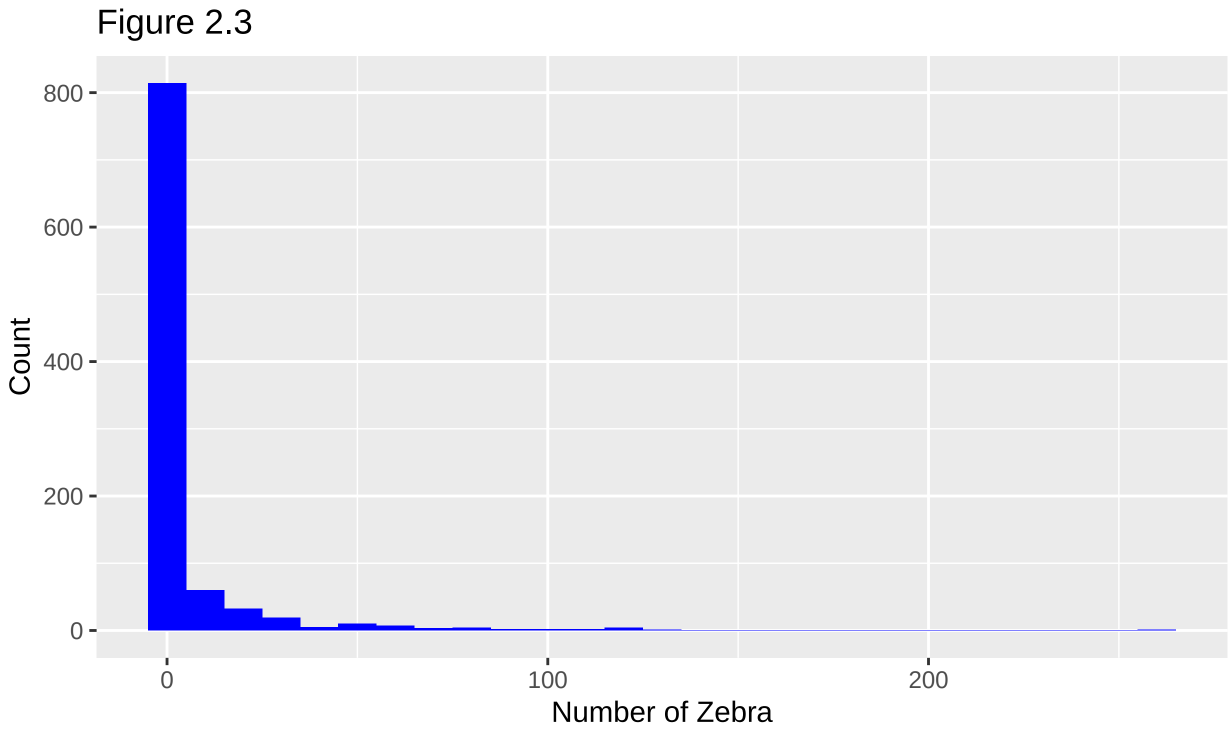

Figure 3: Topi's count

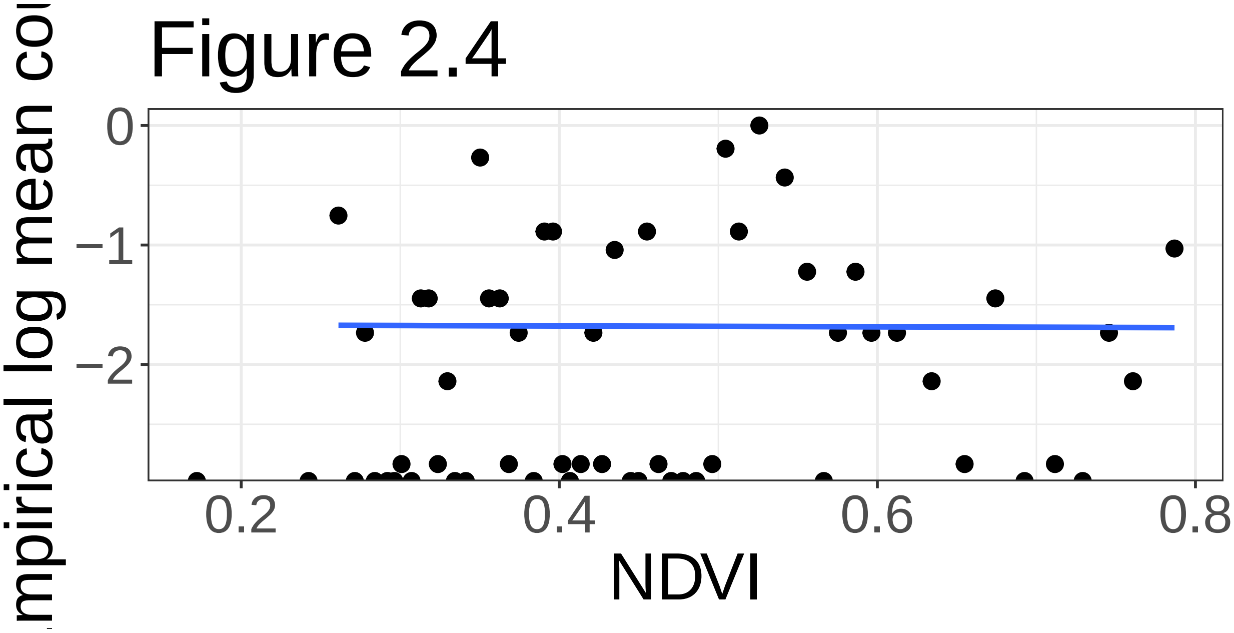

Figure 4: Log mean plot for topi.count vs. ndvi

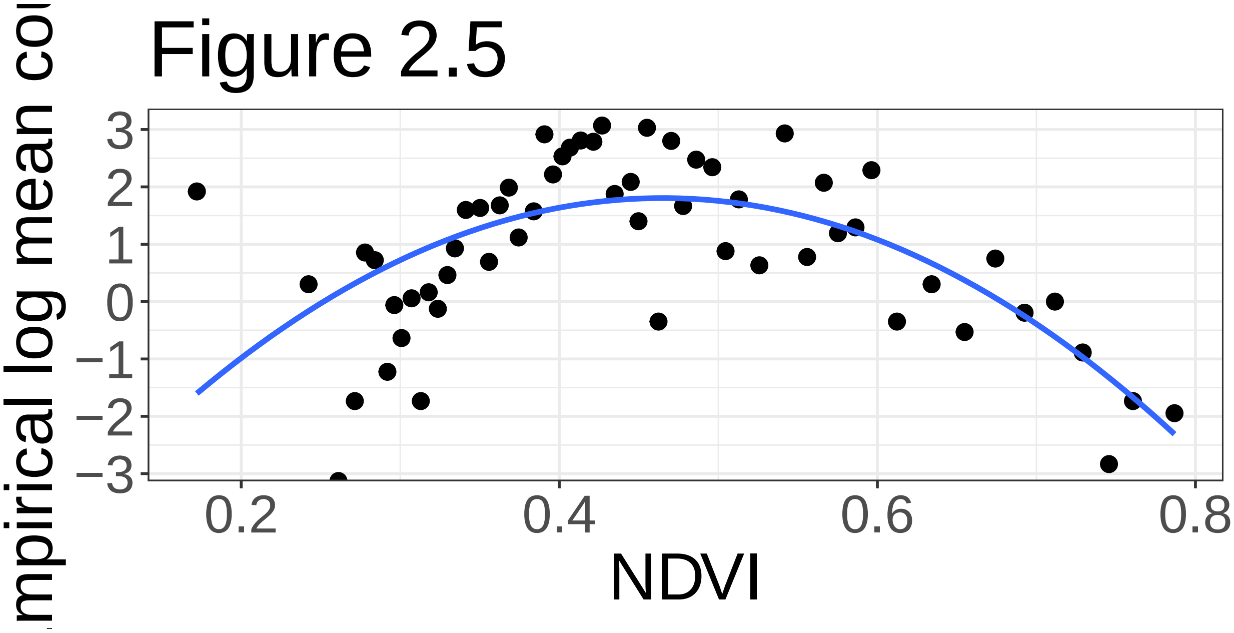

Figure 5: Log mean plot for zebra.count vs. ndvi

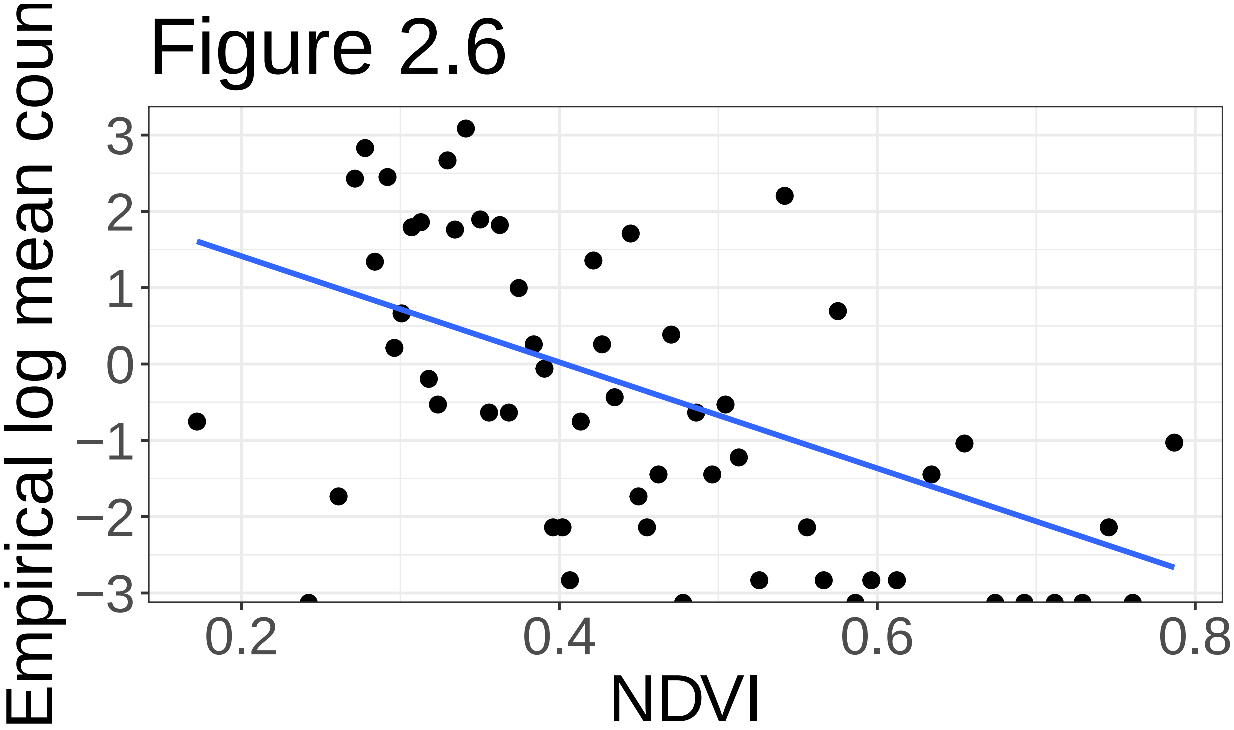

Figure 6: Log mean plot for gazelleThomsons.count vs. ndvi

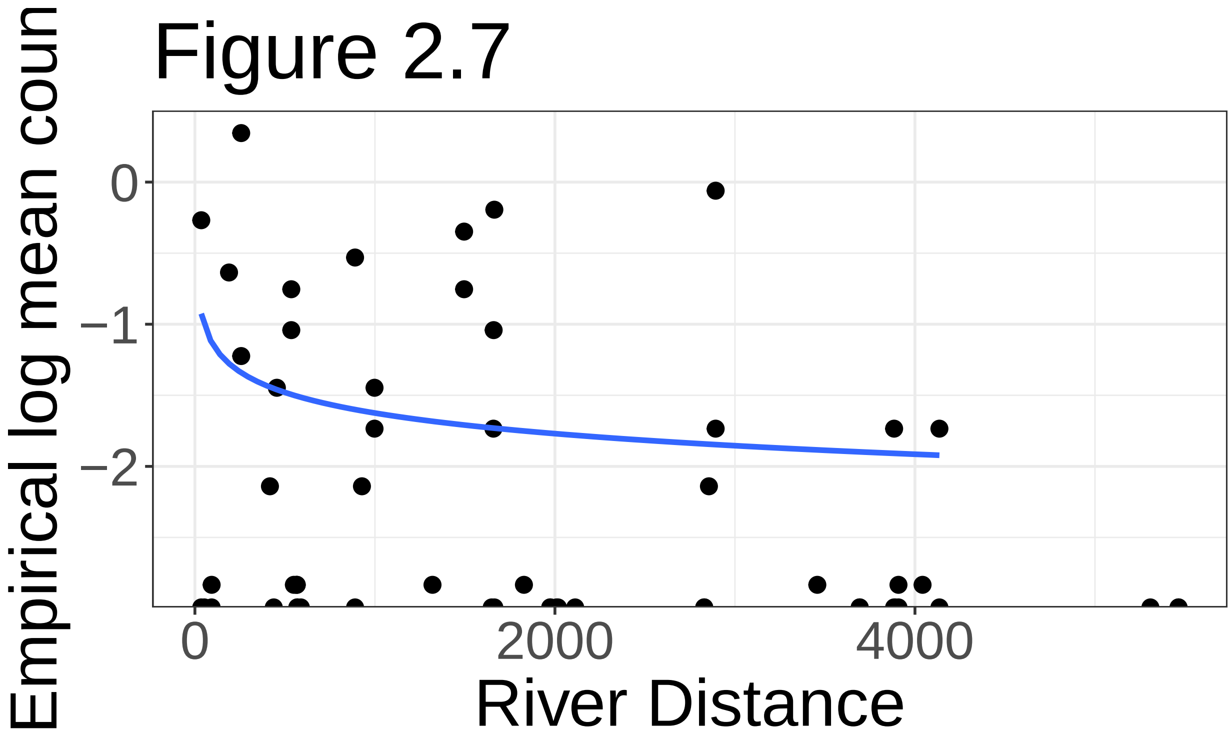

Figure 7: Log mean plot for log(amRivDist) vs. topi.count

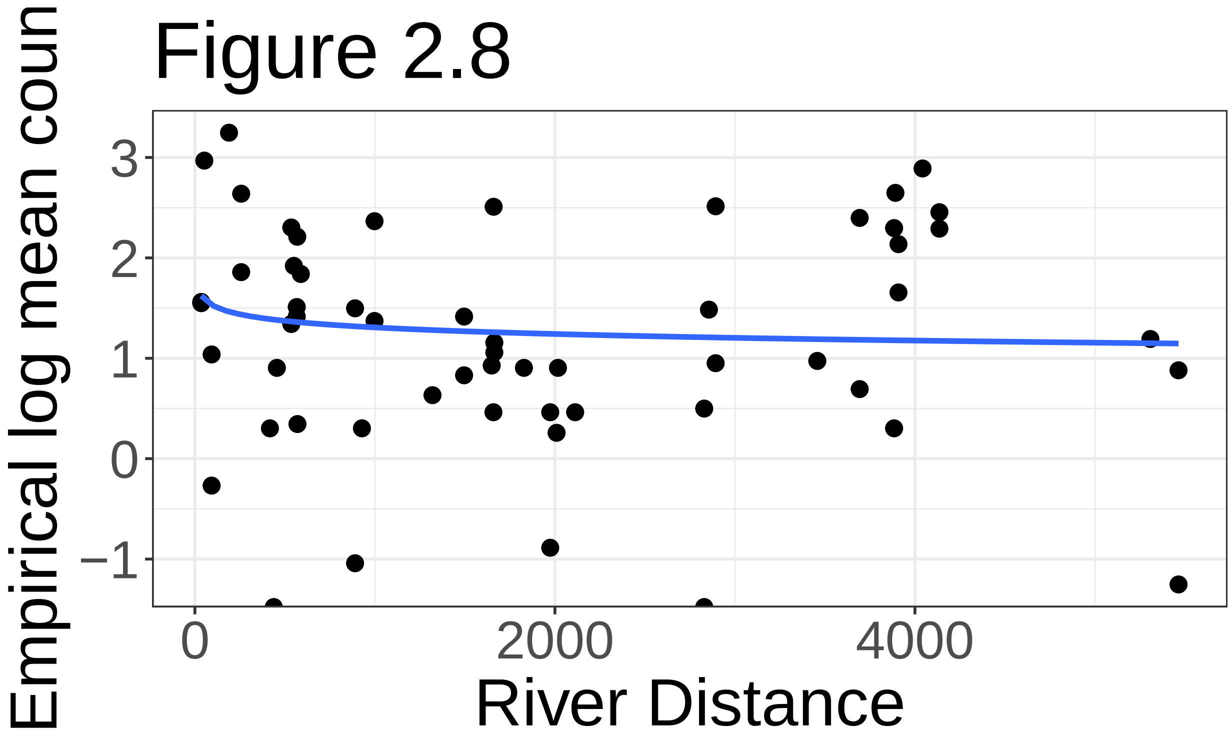

Figure 8: Log mean plot for log(amRivDist) vs zebra.count

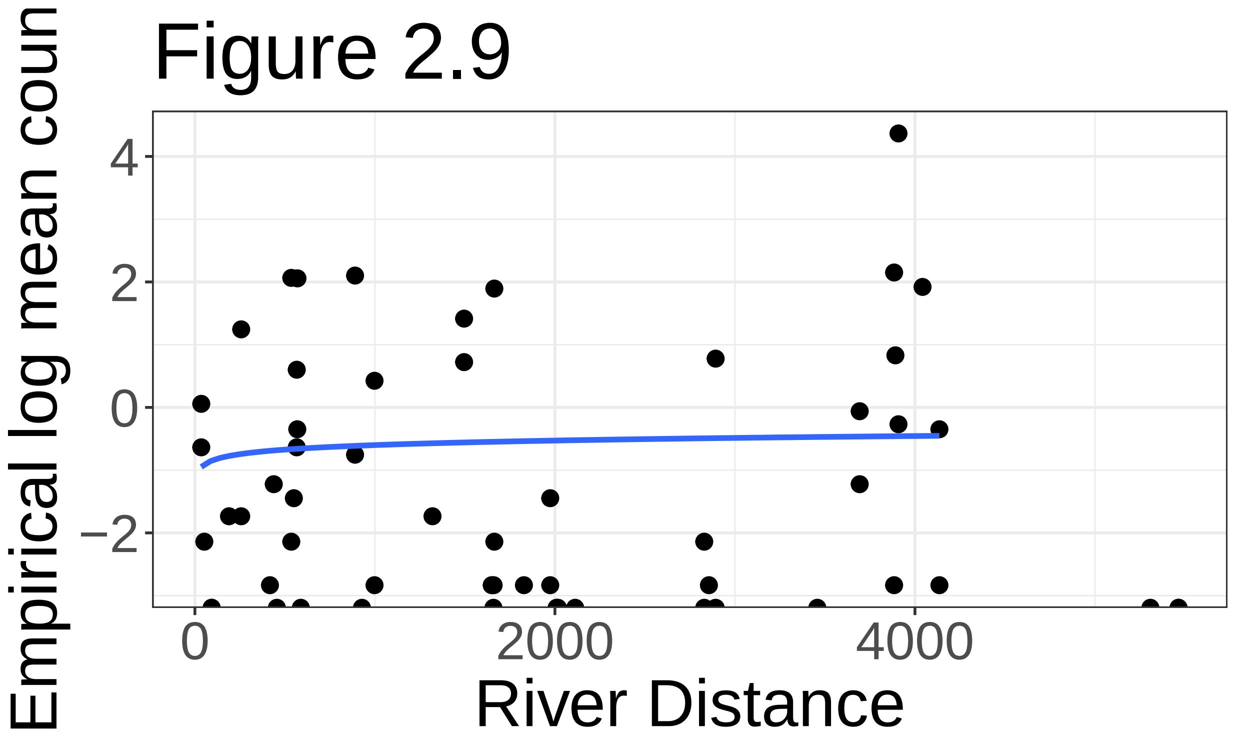

Figure 9: Log mean plot fot log(amRivDist) vs. gazelleThomsons.count

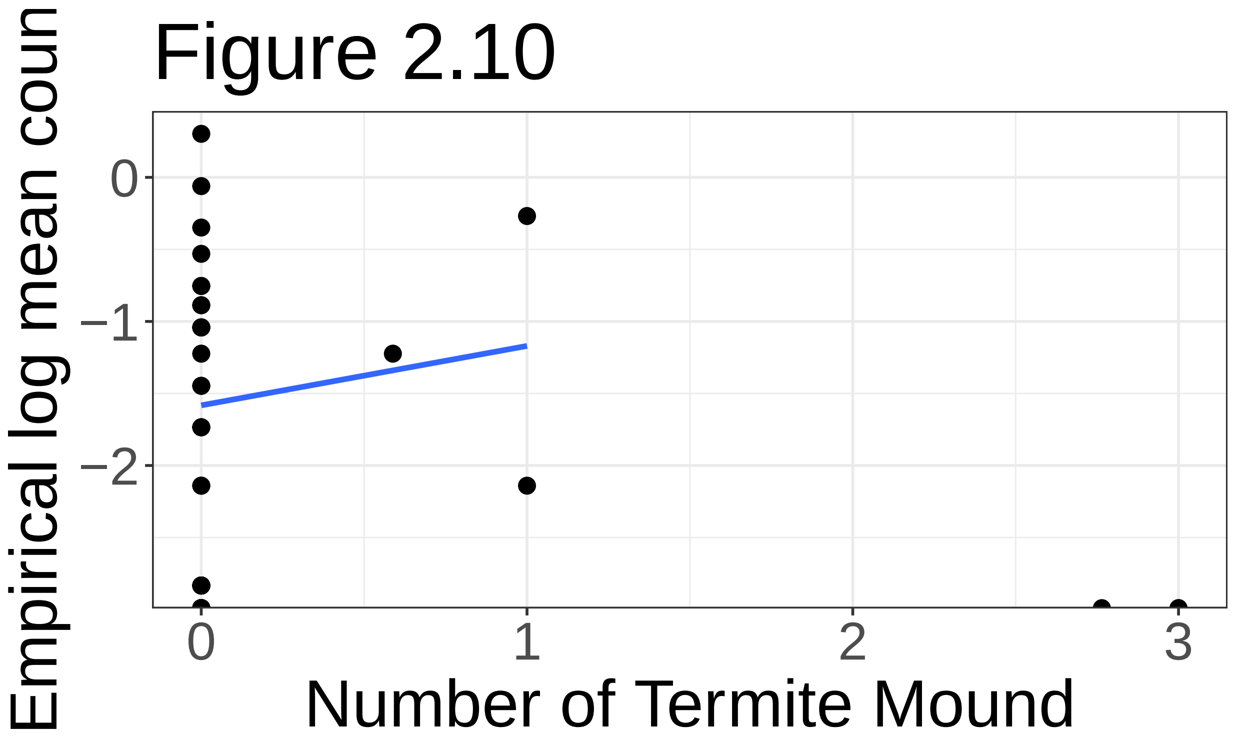

Figure 10: Log mean plot fot TM100 vs. topi.count

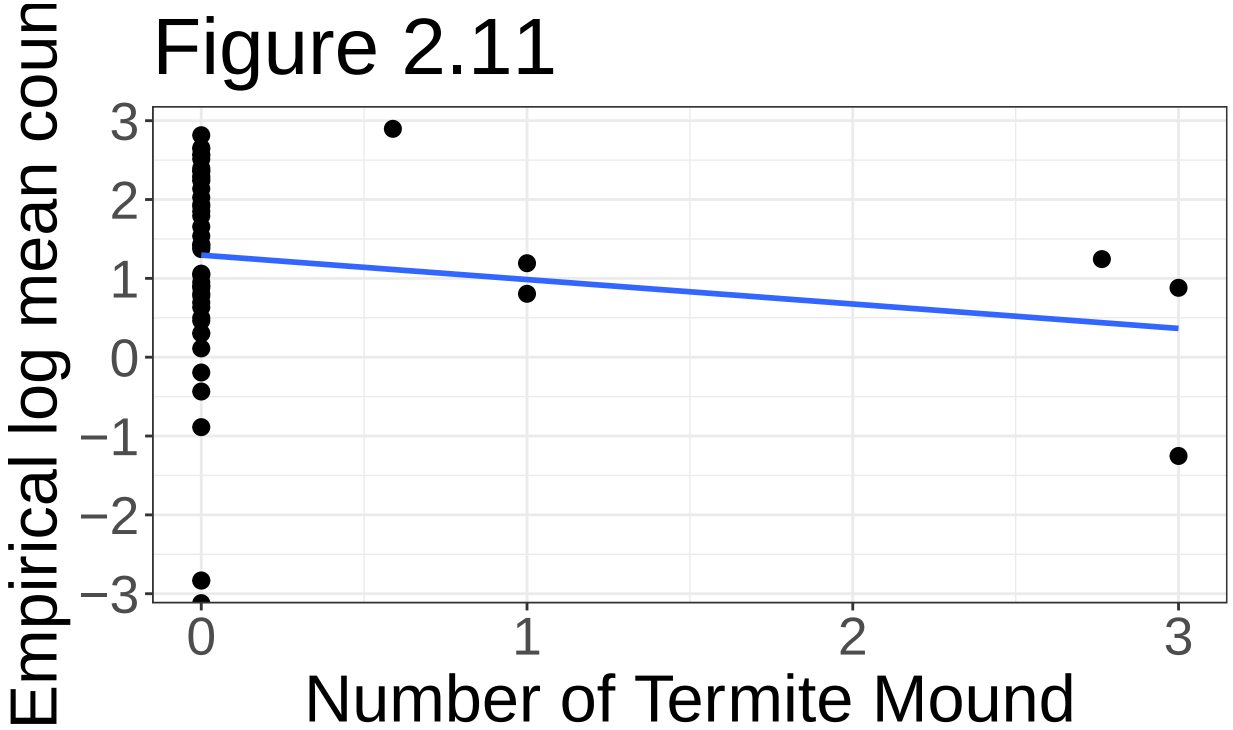

Figure 11: Log mean plot fot TM100 vs. zebra.count

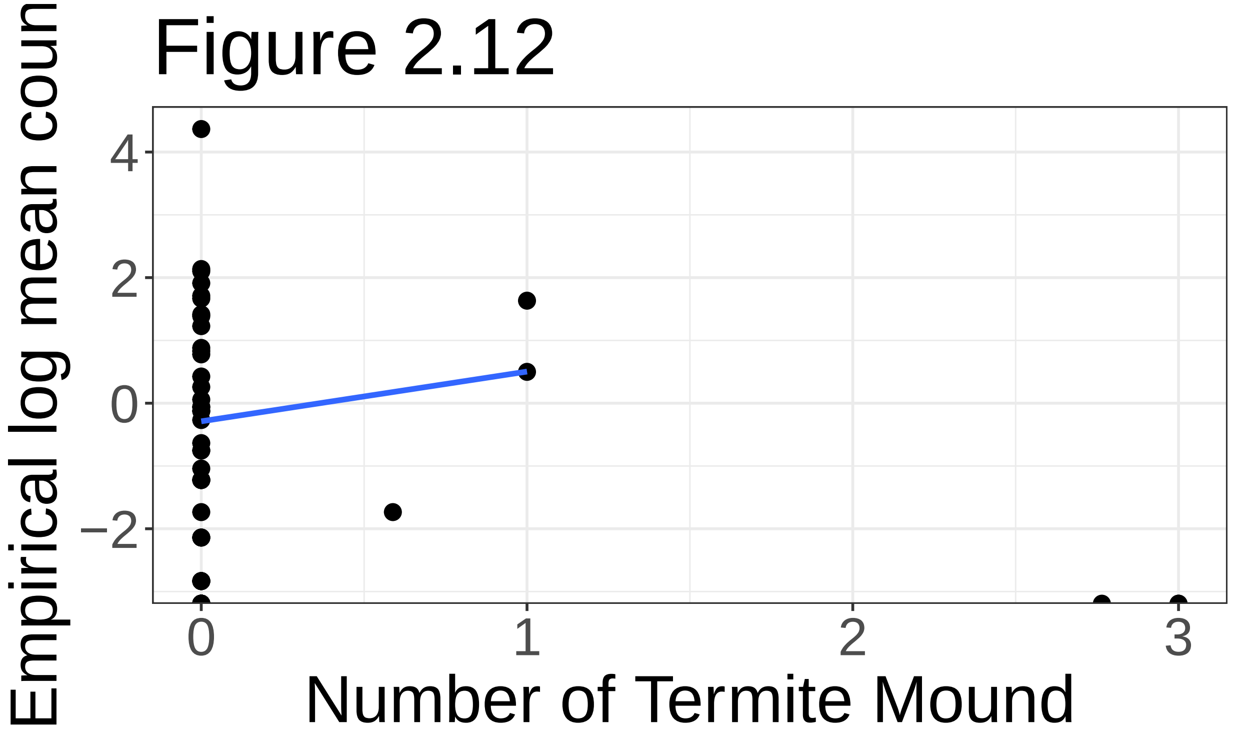

Figure 12: Log mean plot fot TM100 vs. gazelleThomsons.count

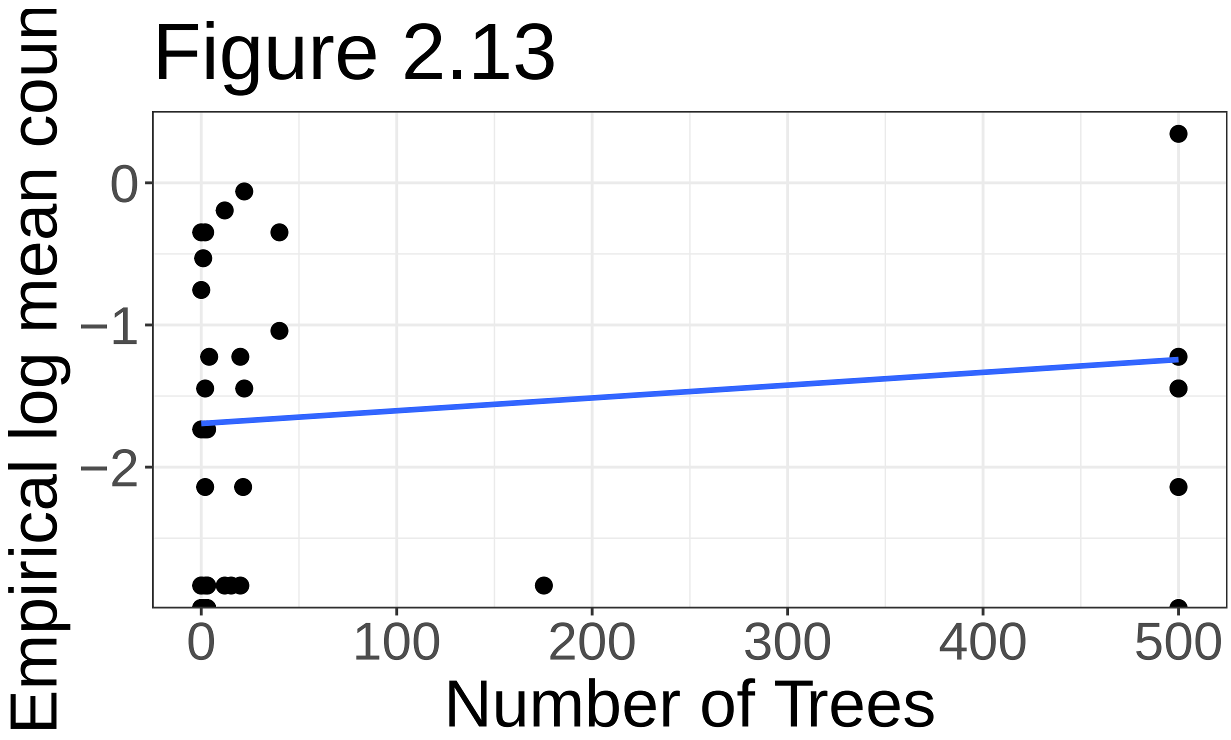

Figure 13: Log mean plot for T50 vs. topi.count

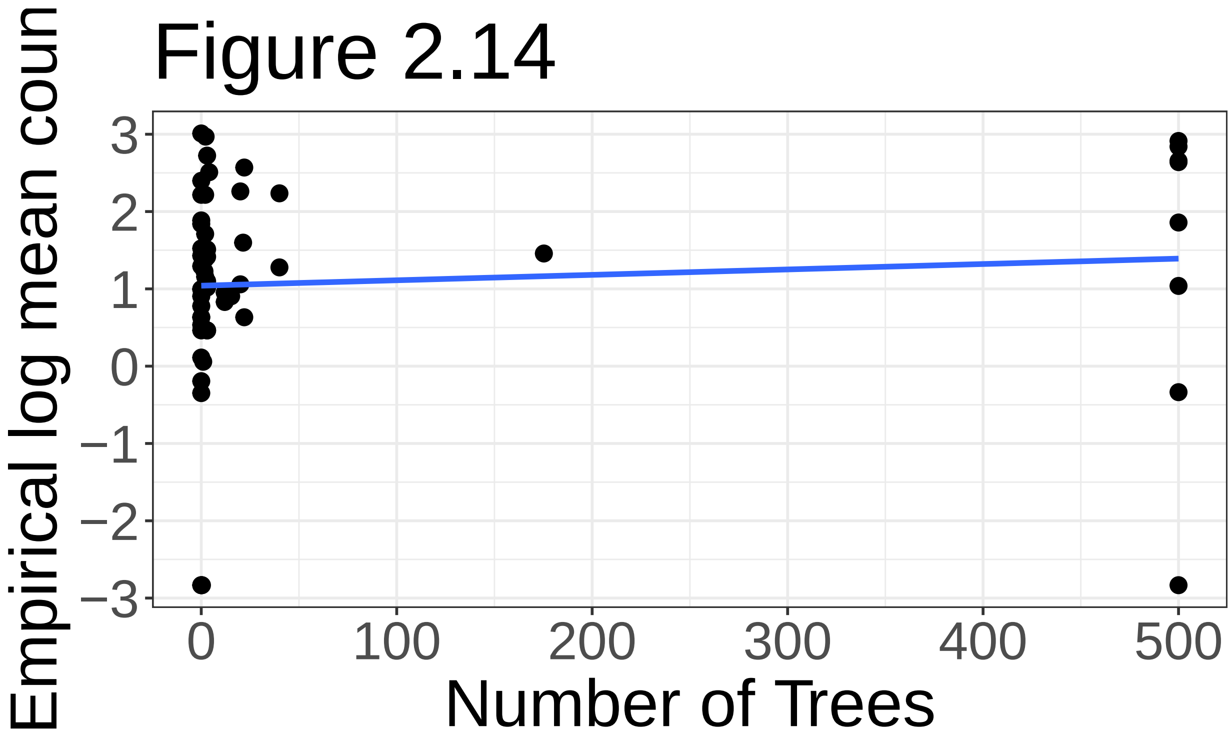

Figure 14: Log mean plot for T50 vs. zebra.count

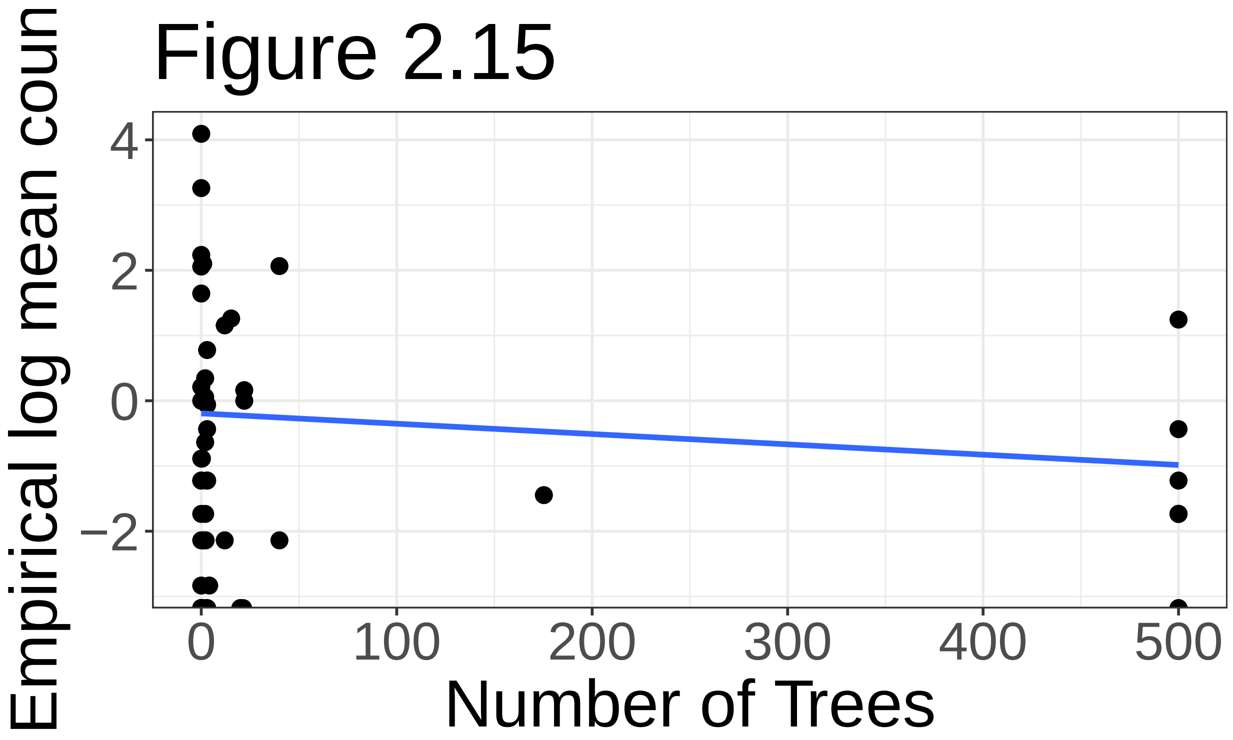

Figure 15: Log mean plot for T50 vs. gazelleThomsons.count

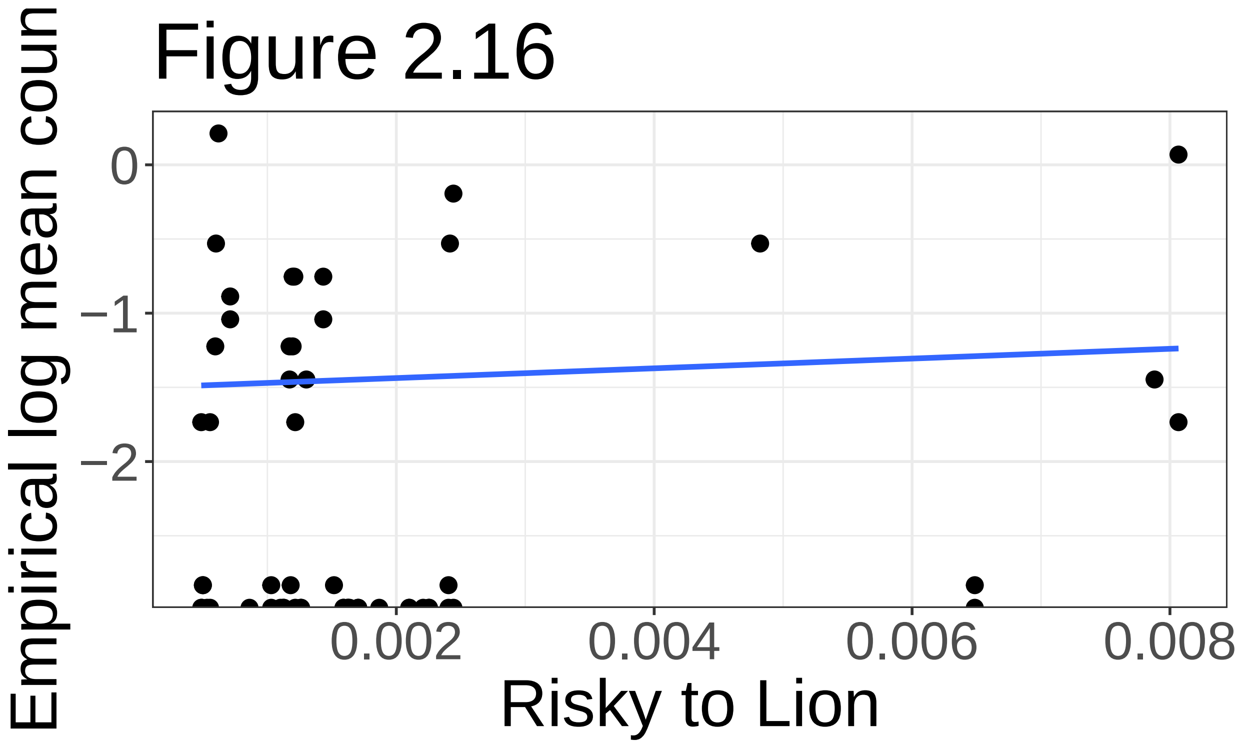

Figure 16: Log mean plot for LriskDry vs. topi.count

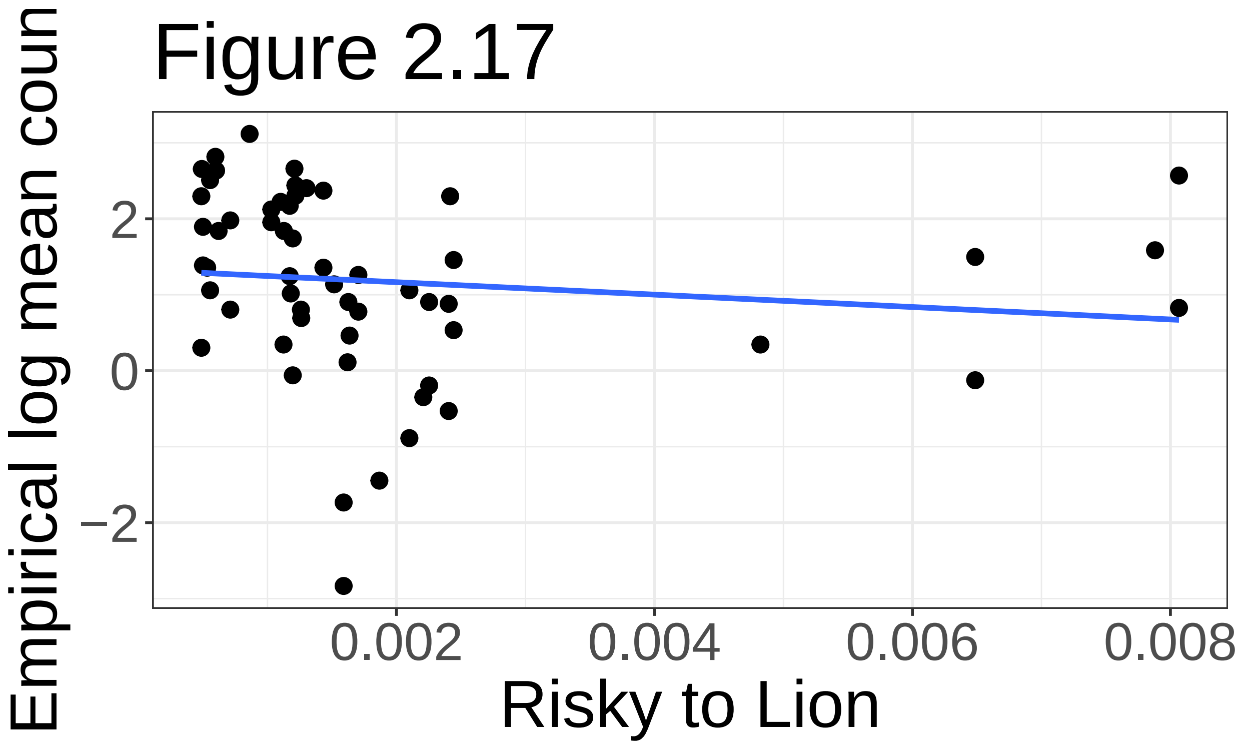

Figure 17: Log mean plot for LriskDry vs. zebra.count

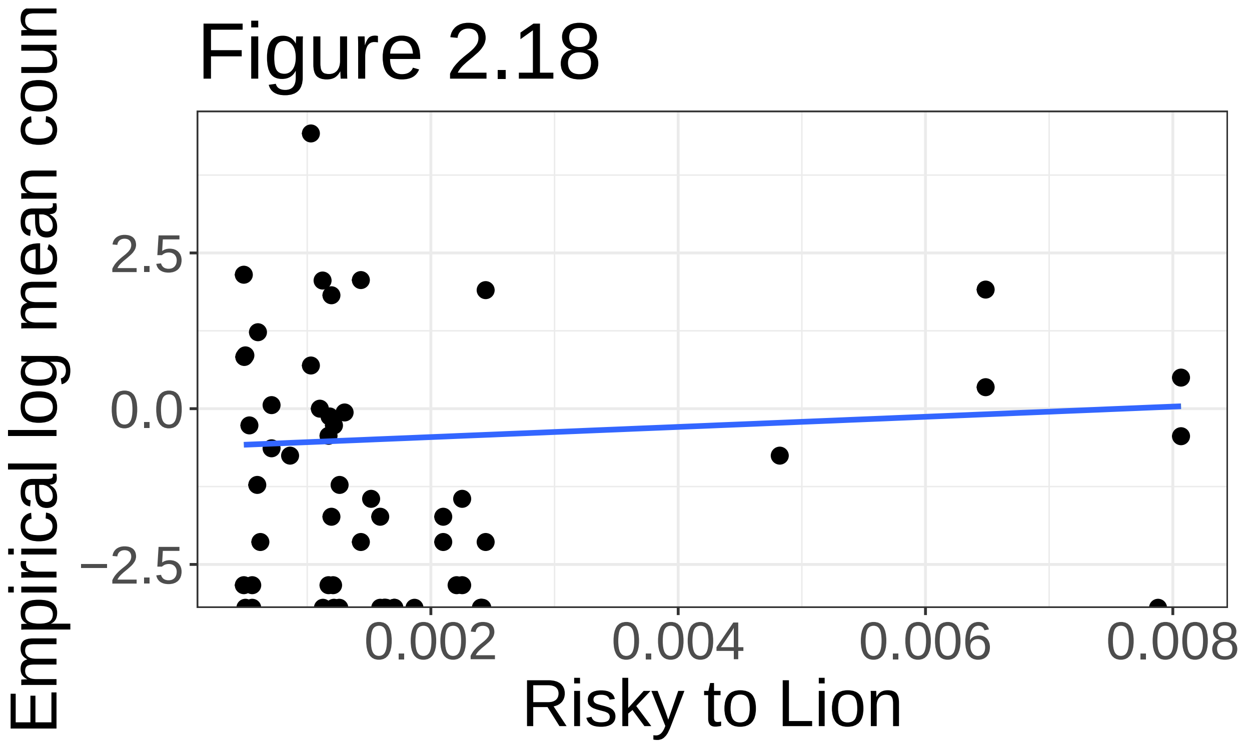

Figure 18: Log mean plot for LriskDry vs. gazelleThomsons.count

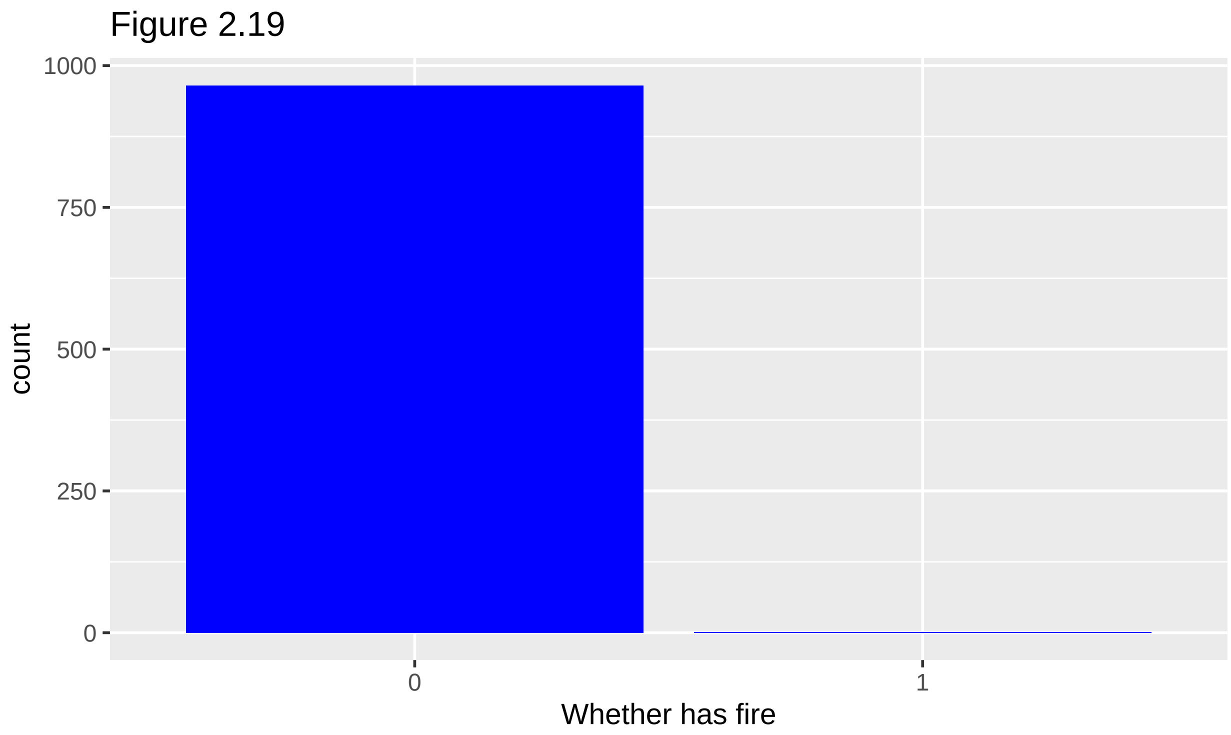

Figure 19: Distribution of Wild Fire

The interpretations for different plots are added in the end of this section.

As showed from Figure 2.4, the relationship between the number of topi and the vegetation condition is also linear since points are bouncing around the linear which means no transformation needed.

Based on Figure 2.5, the relationship between the number of zebra and the vegetation condition is nonlinear. So the polynomial transformation with a power of two is needed.

Based on Figure 2.6, the relationship between vegetation condition and the number of gazelle can be defined as linear since there is no explicit shape. No transformation is required.

Based on Figure 2.7, ideally no transformation is needed for the relation between river distance and the number of topi. However, since the numbers for river distance are really large, outweighed other variables, I choose to do the log transformation on it.

As Figure 2.8 shows, it is also reasonable to have a log transformation with the same reason for the relation between river distance and the number of topi.

Based on Figure 2.9, having a log transformation is reasonable for the relation between the number of gazelle and the river distance.

Based on Figure 2.10, no transformation is needed since it is hard to describe the shape.

Based on Figure 2.11, there is also no transformation needed for the relationship between the number of zebra and the number of termite mounds since most of the points is about around the line.

Based on the Figure 2.12, the relation between number of termite and the number of gazelle does not need to do any transformation.

Based on the Figure 2.13, there is no transformation needed for the number of trees and the number of topi.

Based on Figure 2.14, no transformation is required for the number of trees and number of zebra.

As Figure 2.15 shows, only doing linear transformation for the relationship between number of gazelle and number of trees is fine.

As shown in Figure 2.16, it is reasonable to not have any transformation between the risk to lion and the number of topi.

Based on Figure 2.17, no transformation is needed for the relationship between the number of zebra and the risk to lion in certain site.

From the Figure 2.18, it shows that just having linear transformation for the relation between the risk to lion and number of gazelle.

Based on Figure 2.19, the $fire$ variable is very discrete and there is only one row showing that there is a wild fire. So we may not be able to add this variable into the model.

Section 3: Modeling

In the modeling section, since based on the Figure 2.1, 2.2, and 2.3 showed, three ZIP models for each species are required since there are excessive number of 0 in each species counts.

Section 3.1: Importance of enviornment variables

For topi and gazelle, the population model will be the same. The population model for logistic regression part will be:

$$Z_i \sim Bernoulli(\alpha_i)$$

$$\log\left(\frac{\alpha_i}{1-\alpha_i}\right) = \gamma_0 + \gamma_1 ndvi_i + \gamma_2 log(amRivDist_i) + \gamma_3 TM100_i + \gamma_4 LriskDry_i + \gamma_5 T50_i$$

The Poisson part for topi will be:

$$Topi_i \sim Poisson(\lambda_i)$$

$$\log(\lambda_i) = \beta_0 + \beta_1 ndvi_i + \beta_2 log(amRivDist_i) + \beta_3 TM100_i + \beta_4 LriskDry_i + \beta_5 T50_i$$

The Poisson part for gazelle will be:

$$Gazelle_i \sim Poisson(\lambda_i)$$

$$\log(\lambda_i) = \beta_0 + \beta_1 ndvi_i + \beta_2 log(amRivDist_i) + \beta_3 TM100_i + \beta_4 LriskDry_i + \beta_5 T50_i$$

For zebra, the population model will be:

$$Z_i \sim Bernoulli(\alpha_i)$$

$$\log\left(\frac{\alpha_i}{1-\alpha_i}\right) = \gamma_0 + \gamma_1 ndvi_i + \gamma_2 ndvi_i^2 + \gamma_3 log(amRivDist_i) + \gamma_4 TM100_i + \gamma_5 LriskDry_i + \gamma_6 T50_i$$

The Poisson part will be:

$$Zebra_i \sim Poisson(\lambda_i)$$

$$\log(\lambda_i) = \beta_0 + \beta_1 ndvi_i + \beta_2 ndvi_i^2 + \beta_3 log(amRivDist_i) + \beta_4 TM100_i + \beta_5 LriskDry_i + \beta_6 T50_i$$

Topi's Model

After fitting the model for topi, the fitted regression line for topi will be:

$$\log\left(\frac{\alpha_i}{1-\alpha_i}\right) = 1.889 - 1.252 ndvi_i + 0.241 log(amRivDist_i) - 0.0227 TM100_i - 154.3 LriskDry_i - 0.0011 T50_i$$

$$\log(\lambda_i) = 1.779 - 0.241 ndvi_i - 0.152 log(amRivDist_i) - 0.204 TM100_i + 62.511 LriskDry_i + 0.0006 T50_i$$

Based on the fitted line, we predict that with one unit increase in ndvi, the log mean number of topi will decrease by 0.241. With one unit increase in log river distance, we predict that the log mean for topi count will decrease by 0.152. With one unit increase in termite mound, we predict that the log mean topi will decrease by 0.204. With one unit increase risky to lion, we predict that the log mean topi count will increase by 62.511. With on unit increases in T50, we predict that the log topi number mean will increase by 0.0006. All the interpretations are done when other variables holding constant.

Zebra's Model

For zebra model, the fitted regression line will be:

$$\log\left(\frac{\alpha_i}{1-\alpha_i}\right) = 0.127 + 2.4925 ndvi_i + 19.34 ndvi_i^2 + 0.0663 log(amRivDist_i) + 0.26 TM100_i + 83.7532 LriskDry_i - 0.0009 T50_i$$

$$\log(\lambda_i) = 2.426 - 8.638 ndvi_i - 21.32 ndvi_i^2 - 0.0051 log(amRivDist_i) - 0.1599 TM100_i - 37.82 LriskDry_i + 0.0004 T50_i$$

From the fitted regression line, we predict that with one unit increase in log(amRivDist), the expected log mean for zebra will decrease by 0.0051. With on unit increase in Tm100, we predict that the log mean for zebra number will decrease by 0.1599. With one unit increase in LriskDry, we predict that the log mean zebra number will decrease by 37.82. With one unit increase in T50, we predict that the log mean for zebra number will increase by 0.0004.

Gazelle's Model

After fitted the model for gazelle, The fitted regression line for gazelle will be:

$$\log\left(\frac{\alpha_i}{1-\alpha_i}\right) = 0.5805 + 5.235 ndvi_i - 0.1815 log(amRivDist_i) + 0.6865 TM100_i + 43.7 LriskDry_i + 0.00003 T50_i$$

$$\log(\lambda_i) = -0.0422 - 2.816 ndvi_i + 0.5595 log(amRivDist_i) - 0.3943 TM100_i - 102.3 LriskDry_i - 0.0026 T50_i$$

Based on the fitted regression line, with one unit increase in the ndvi, we predict that the log mean for gazelle's count will decrease by 2.816. With one unit increase in log(amRivDist), we predict that the log mean for the number of gazelle will increase by 0.5595. With one unit increase in TM100, we predict that the log mean for gazelle count will decrease by 0.3943. With one unit increase in LriskDry, we predict that the log mean of gazelle number will decrease by 102.3. With one unit increase in number of trees, we predict that the log mean number of gazelle will decrease by 0.0026. With other variables holding constant.

With the given research question, I will first test whether environmental variables have relationship with the number of topi in data set. Basically, I decided to do the hypothesis test for variables "ndvi", "TM100", and "T50" together since the measurement on how "green" the location is related to the number of trees in location and also relates to the soil quality about termite mounds. I also will give separate test for the rest to variables for each different species.

Topi

Test 1 The first test is testing whether there is a relation between the vegetation condition (ndvi), number of trees (T50), and the number of termite mound (TM100) and topi count. The hypothesis test will be

$H_0 : \beta_1 = \beta_3 = \beta_5 = 0$

H_a : At least one of the β are not zero

The log likelihood for the full model will be -355.7883

The log likelihood for the reduced model will be -356.9112

The test statistic G will be 2(-355.7883-(-356.9112)) = 2.2458, and the degree of freedom is 3 (ndvi, TM100, T50).

Using χ^2 distribution, we know that the p-value is kind of large (0.52), we may conclude that there is no evidence or very weak evidence showing there is a relationship between the number of topi and the vegetation condition, the number of trees, and the number of termite mounds after holding other variables constant.

Test 2 The second hypothesis test is testing whether there is a relationship between the number of topi and the log of river distance (log(amRivDist)). The hypothesis test will be:

$H_0 : \beta_2 = 0$

$H_a : \beta_2 \neq 0$

The log likelihood for the full model will be -355.7883

The log likelihood for the reduced model will be -357.0991

The test statistic G will be 2(-355.7883-(-357.0991)) = 2.2458, and the degree of freedom is 1.

Using χ^2 distribution, we know that p-value is about 10%, we can conclude that there is a moderate evidence showing there is a relationship between the number of topi and the log river distance over each site after holding other variables constant.

Test 3 The third hypothesis test is testing whether there is a relationship between the number of topi and the risky to lions. The hypothesis test will be:

$H_0 : \beta_4 = 0$

$H_a : \beta_4 \neq 0$

The log likelihood for the full model will be -355.7883

The log likelihood for the reduced model will be -356.8398

The test statistic G will be 2(-355.7883-(-356.8398)) = 2.103. The degree of freedom is 1.

Using χ^2 distribution, we know that p-value is about 14.7%, we can conclude that there is a moderate evidence showing that there is a relationship between the number of topi and the risky to lion after holding other variables constant.

Zebra

Test 1 Same as the tests in topi count, the first hypothesis test on zebra will be testing the relationship between the number of zebra and the vegetation condition (ndvi), number of trees (T50). The hypothesis test will be:

$H_0 : \beta_1 = \beta_2 = \beta_4 = \beta_6 = 0$

H_a : At least one of the β are not zero

The log likelihood for the full model will be 5376.208

The log likelihood for the reduced model will be -5786.12

The test statistic G will be 2(-5376.208-(-5786.12)) = 819.824. The degree of freedom will be 3.

Using χ^2 distribution, we know that this p-value is very close to zero, we can conclude that there is a strong evidence showing there is a relationship between the number of zebra and vegetation condition (ndvi), number of trees (T50), and the number of termite mound (TM100) in this data set after holding other variables constant.

Test 2 The second hypothesis test on zebra will be testing the relationship between the number of zebra and the log of the river distance (log(amRivDist)). The hypothesis test will be:

$H_0 : \beta_2 = 0$

$H_a : \beta_2 \neq 0$

The log likelihood for the full model will be -5376.208

The log likelihood for the reduced model will be -5376.331

The test statistic G will be 2(-5376.208-(-5376.331)) = 0.246, and the degree of freedom is 1.

Using χ^2 distribution, we know that the p-value is relatively large (0.62), we can conclude that there is a weak or even no evidence showing there is a relation between the number of zebra and the log river distance in this data set after holding other variables constant.

Test 3 The third hypothesis test on zebra will be testing the relationship between the number of zebra and the risky to lions (LriskDry) The hypothesis test will be:

$H_0 : \beta_6 = 0$

$H_a : \beta_6 \neq 0$

The log likelihood for the full model will be -5376.208

The log likelihood for the reduced model will be -5386.361

The test statistic G will be 2(-5376.208-(-5386.361)) = 20.306, and the degree of freedom will be 1.

Using χ^2 distribution, we know that the p-value is kind of close to zero. We can conclude that there is a strong evidence showing that there is a relationship between the number of zebra in sites and the risky to lions after holding other variables constant.

Gazelle

Test 1 The first test is testing whether there is a relation between the vegetation condition (ndvi), number of trees (T50), and the number of termite mound (TM100) and number of gazelles. The hypothesis test will be

$H_0 : \beta_1 = \beta_3 = \beta_5 = 0$

H_a : At least one of the β are not zero

The log likelihood for the full model will be -2963.71

The log likelihood for the reduced model will be -3196.036

The test statistic G will be 2(-2963.71 -(-3196.036)) = 464.652. The degree of freedom will be 3.

Using χ^2 distribution, we know that the p-value is very close to zero. We can conclude that there is a strong evidence for proving there is a relationship between the number of gazelle and the vegetation condition, number of trees and the termite mound after holding other variables constant.

Test 2 The second test is testing whether there is a relation between the number of gazelle and the log of river distance (log(amRivDist)). The hypothesis test will be:

$H_0 : \beta_2 = 0$

$H_a : \beta_2 \neq 0$

The log likelihood for the full model will be $-2963.71 $

The log likelihood for the reduced model will be -3290.404

The test statistic G will be 2(-2963.71 -(-3290.404)) = 653.388. The degree of freedom will be 1.

Using χ^2 distribution, we know that the p-value is very close to zero. We can conclude that there is a strong evidence showing that there is a relationship between the number of gazelle and the log river distance after holding other variables constant.

Test 3 The third test is testing whether there is a relation between the number of gazelle and the risky to lions (LriskDry). The hypothesis test will be:

$H_0 : \beta_5 = 0$

$H_a : \beta_5 \neq 0$

The log likelihood for the full model will be -2963.71

The log likelihood for the reduced model will be -3008.449

The test statistic G will be 2(-2963.71 -(-3008.449)) = 89.478. The degree of freedom will be 1.

Using χ^2 distribution, we know that the p-value is very close to zero. We can conclude that there is a strong evidence showing that there is a relationship between the number of gazelle and the risky to lion after holding other variables constant.

Section 3.2: Comparing Species

For the topi model, the fitted line will be:

$$\log\left(\frac{\alpha_i}{1-\alpha_i}\right) = 1.889 - 1.252 ndvi_i + 0.241 log(amRivDist_i) - 0.0227 TM100_i - 154.3 LriskDry_i - 0.0011 T50_i$$

$$\log(\lambda_i) = 1.779 - 0.241 ndvi_i - 0.152 log(amRivDist_i) - 0.204 TM100_i + 62.511 LriskDry_i + 0.0006 T50_i$$

For the zebra model, the fitted line will be:

$$\log\left(\frac{\alpha_i}{1-\alpha_i}\right) = 0.127 + 2.4925 ndvi_i + 19.34 ndvi_i^2 + 0.0663 log(amRivDist_i) + 0.26 TM100_i + 83.7532 LriskDry_i - 0.0009 T50_i$$

$$\log(\lambda_i) = 2.426 - 8.638 ndvi_i - 21.32 ndvi_i^2 - 0.0051 log(amRivDist_i) - 0.1599 TM100_i - 37.82 LriskDry_i + 0.0004 T50_i$$

For the gazelle model, the fitted line will be:

$$\log\left(\frac{\alpha_i}{1-\alpha_i}\right) = 0.5805 + 5.235 ndvi_i - 0.1815 log(amRivDist_i) + 0.6865 TM100_i + 43.7 LriskDry_i + 0.00003 T50_i$$

$$\log\left(\lambda_i\right) = -0.0422 - 2.816ndvi_i + 0.5595log(amRivDist_i) - 0.3943TM100_i - 102.3LriskDry_i - 0.0026T50_i$$

It can be observed that for ndvi variable, all three species' number have a negative relationship which the zebra has the most negative effect. For log river distance, topi and gazelle's count have a positive relationship, whereas the zebra's number has a negative relation. For variable TM100, all three species' count have a negative relation which gazelle has the most negative relation. For variable risky to lion, the topi's number has a positive relation with this variable. However, the rest species' number will both decrease if the risky to lion increase and the risky to lion seems to have the largest effect on gazelle number. For the number of trees variable, both topi and zebra's number have a positive relation with the number of trees but for gazelle's number, it has a negative relation with it.

For the p-value analysis, for species topi, many variables do not show a strong relationship, so there are two possibilities. The first one is that topi's have strong vitality, which with changing environment, the number of topi will not vary anyways. The other reason is that the number of topi is relatively smaller which the data set may not have enough data to get the correct conclusion. in the data set, the second possibility seems correct.

For species zebra, the p-value for the log river distance seems to be large which showing that the number of zebra may not related to the river distance. For the other variables, p-value indicates a strong relationship.

For species gazelle, all variables p-value are very small showing that the number of gazelle seems to have a strong relation between the environmental variables.

Section 4: Discussion

Based on the analysis, I think in this data set the environmental variables somehow have strong relationship for species count. However, for specie topi, it seems to have no relationship or weak relationships for different environmental variables. I think it is because the number of topi in this data set may not be enough to draw conclusion since based on Figure 2.2, the number of topi is much less than the number of zebra and the number of gazelle. For most variables, it seems that they have a negative relationship with the number of species, which I think is reasonable. For example, the number of species will obviously decrease if the vegetation condition is not good (ndvi is low). If a site is risky to lion, the number of species may also be small, and this also hold true for the sites which are far away from river. For the data set, there is one environmental variable that I did not use which is fire. Since by Figure 2.19, there is only one row with fire which make the fire variable not to be representative and I did not add it when fitting models. one limitation for the analysis is that the data set is not large enough. Because the data set is not large, I choose to not add fire variable into my model which I think this variable may affect the number of species somehow. Also the size of the data set also makes the result for specie topi not really convincing since I guess topi is an endangered animal which the number of topi is very small. Since there are not enough data for topi, the result may have some occasion like p-value is relatively large compared to the others.