project_2

Jessica Chen

2022-11-21

Abstract:

Environmental factors are always considered to be contributing to species' living habits.In this project, we are specifically interested in how environmental variables affect the number of species present in the locations. Specifically, we are going to study three species - gazelleThomsons, topi and zebra.

The data set we used comes from a long-term research program run by Wake Forest University in the Serengeti Park in Tanzania. The data image was collected from cameras installed at 19 different sites in the Serengeti. We are only going to use the subset of this data. In the report, we performed analysis on every potential variables in the data, and fitted the models which can explain statistically how environmental variables affect the counts of species. We then compared the results of each species. It turned out that the count of species has different relationships with our predicted environmental variables.

Section1: Introduction

We are interested in the question: Is there a relationship between the environmental variables and the counts for each species? The data we used comes from the research program run by Wake Forest University in the Serengeti Park in Tanzania. The data image was collected from cameras installed at 19 different sites in the Serengeti.

To answer the research question, we use visualizations and zero inflated poisson models to explore the relationship between the counts of three species and different environmental factors such as risk to lion predation and distance to the nearest river..etc. We find that, the counts of the species are related to environmental factors, and also different environmental factors have different impacts on different species.

Section2 Data

Part1: Data Preparation

There are overall 966 rows and 15 columns in the serengeti dataset, which means that there are 966 independent observations. The observations are taken in 19 different sites in Serengeti in different 8-day time periods. Specifically, each observation contain 15 key informations: siteID, date, site.date, ndvi,gazelleThomsons.count,gazelleThomsons.present, topi.count, topi.present, zebra.count, zebra.present, fire, amRiDist, TM100, LriskDry, which corresponds to 15 column variables in the datatable.

No missing value exists in the data, which is good. But as we assume the very first three column variables in the dataset siteID, date, site.date, which records the site location and the time period, function mostly as the identifiers of each observation, and therefore should not be count as a potentially statistically significant factor in our research question. And the three indicators in the dataset: gazelleThomson.present, topi.present, zebra.present, we do not want to use them in the analysis because the logistic regression part of the model we choose (zero inflated poisson) is givng the similar output. We remove these six columns and create a subset of dataset serengeti named as serengeti_new. The new dataset contains 996 rows and 9 columns.

To help with further analysis, we will specify the meaning of each 9 variables in the data set. There are three continuous numerical data: ‘ndvi’ is a measure of how “green” the location is. Locations have high values of ndvi if they have a lot of vegetation, or if the vegetation is of high quality and very green. We may refer to the situation as “greenness” in the later text. And ‘amRivDist’ is the distance from the location to the nearest river. ‘LriskDry’ is a measure of how risky the location is to lion predation.

And there are overall 5 discrete numerical variables in the dataset. ‘gazelleThomsons.count’ refers to the count of Thomson’s gazelles that were spotted over the 8-day period. ‘topi.count’ refers to the count of topi that were spotted over the 8-day period. ‘zebra.count’ refers to the count of zebras that were spotted over the 8-day period. ‘TM100’ counts the number of termite mounds within 100 meters of the location. ‘T50’ counts the number of trees within 50 meters of the location.

And there is one binary response variables exist in the dataset: ‘fire’ refers to the presence of a detected wildfire in the 30 days preceding the time period.And Termite mounds are an indicator of soil quality.

Part2: EDA

a. EDA on three response variables: gazelleThomsons.count, topi.count,zebra.count

gazelleThomsons.count, topi.count and zebra.count are count data, so we can draw a histogram of each to visualize its distribution. From Figure2.1, Figure2.2, and Figure3, we can see that there are excess of zero count in the distribution.We want to model these excess zeros independently so our model can be a better fit. We may use zero inflated Poisson model later to investigate the relationship between the species count and the environmental factors.

Figure 2.1 : gazelleThomsons.count

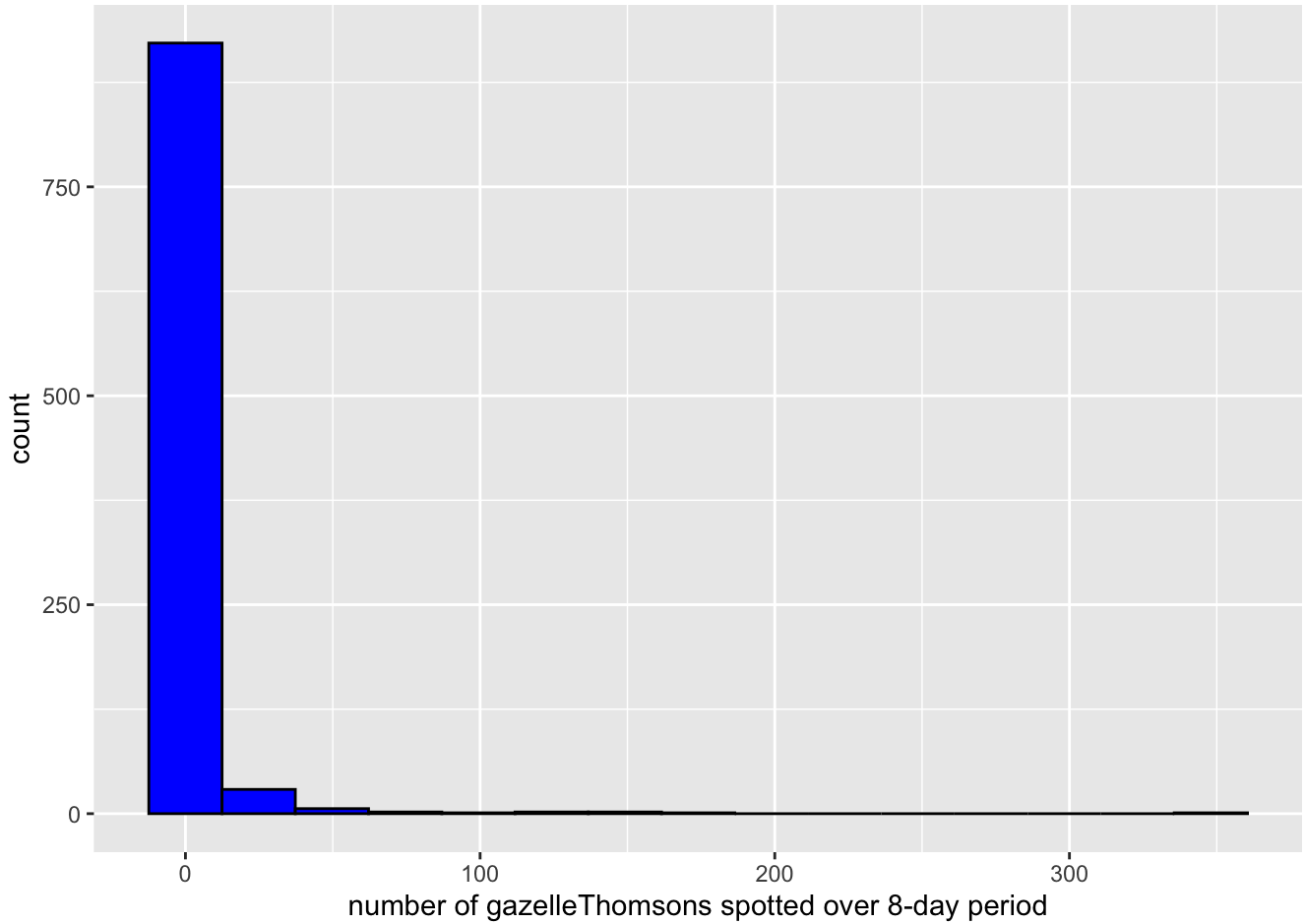

Figure 2.1: Histogram of the number of gazelleThomsons spotted over the 8-day period

Figure 2.2 : topi.count

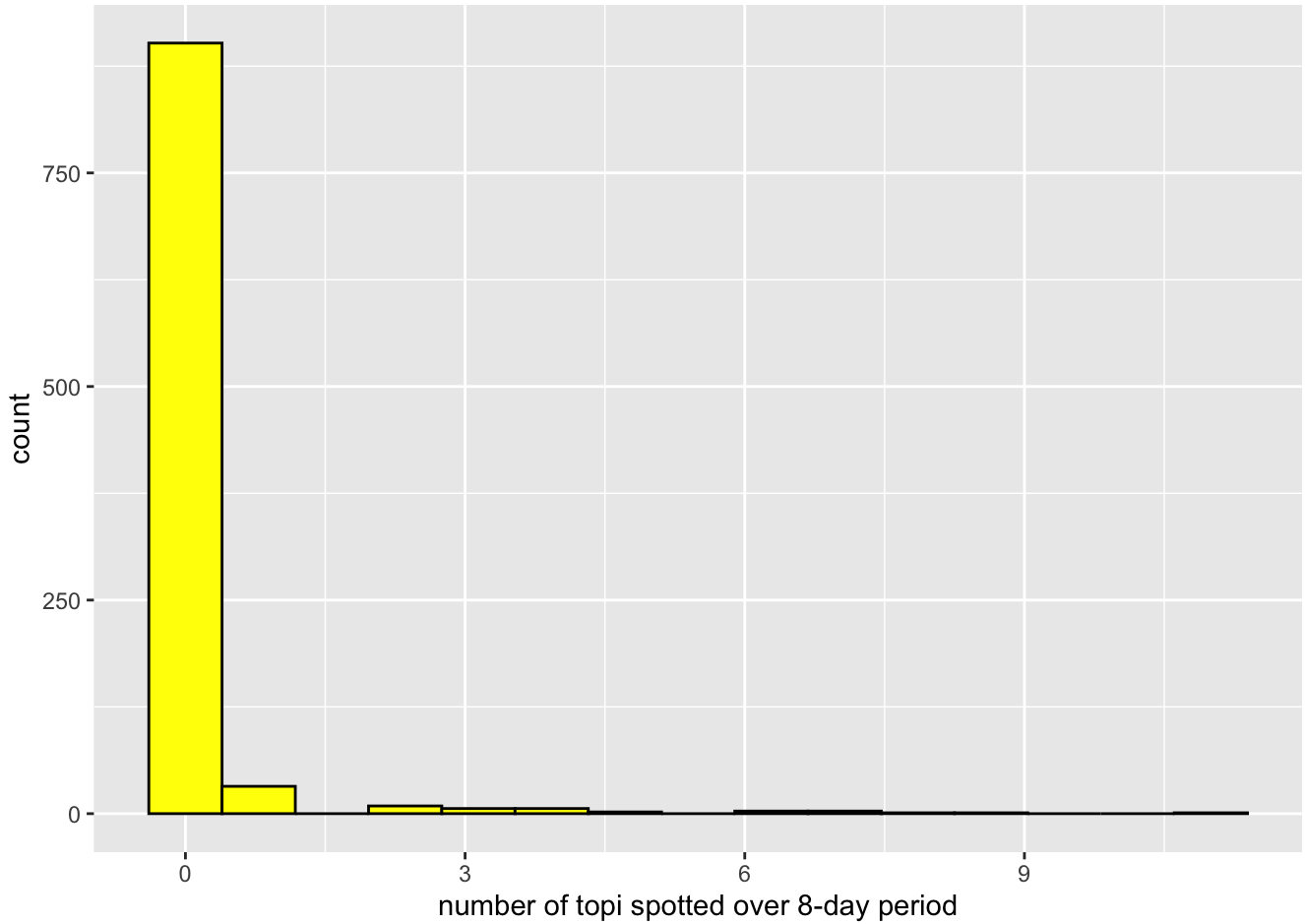

Figure 2.2: Histogram of the number of topi spotted over the 8-day period

Figure 2.3 : zebra.count

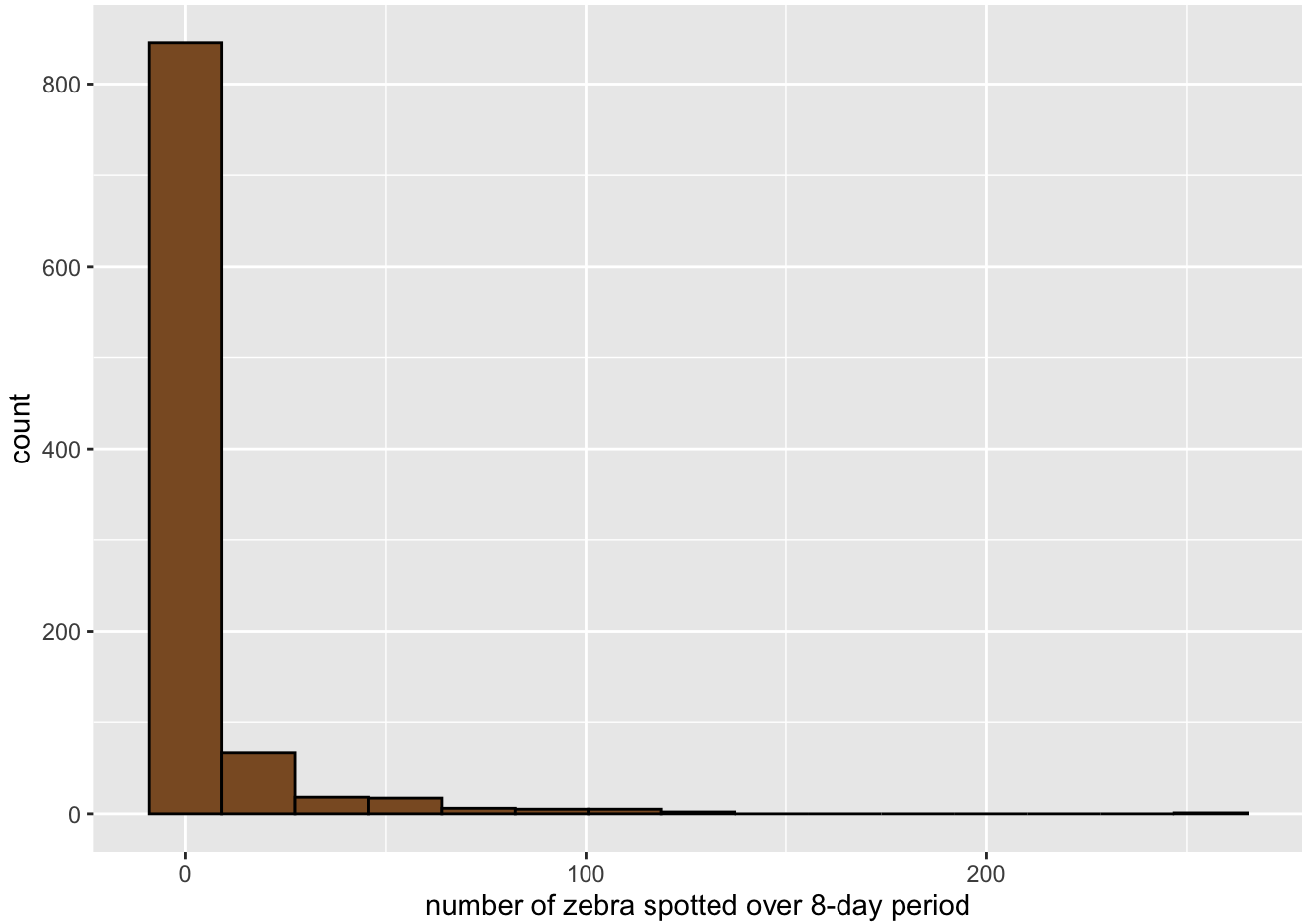

Figure 2.3: Histogram of the number of zebra spotted over the 8-day period

b. EDA on bivariate data

We then performed data analysis on 3 continuous quantitative variables - ndvi, amRivDist, LriskDry - to each of the three response variables to check the shape of their relationships. (From Figure 2.4 to Figure 2.12)

The response variables are count data, therefore, we will use the log mean plots to visualzie the its relationship with other quantitative variables.

Figure 2.4 gazelleThomsons.count vs. ndvi

Figure2.4 indicates the relationship between ndvi and the count of gazelleThomsons over the 8-day period, with ndvi on the x-axis and log mean of gazelleThomsons.count on the y-axis. As shown from the figure, the log mean of gazelleThomsons.count shows up a generally negative linear relationship with ndvi.

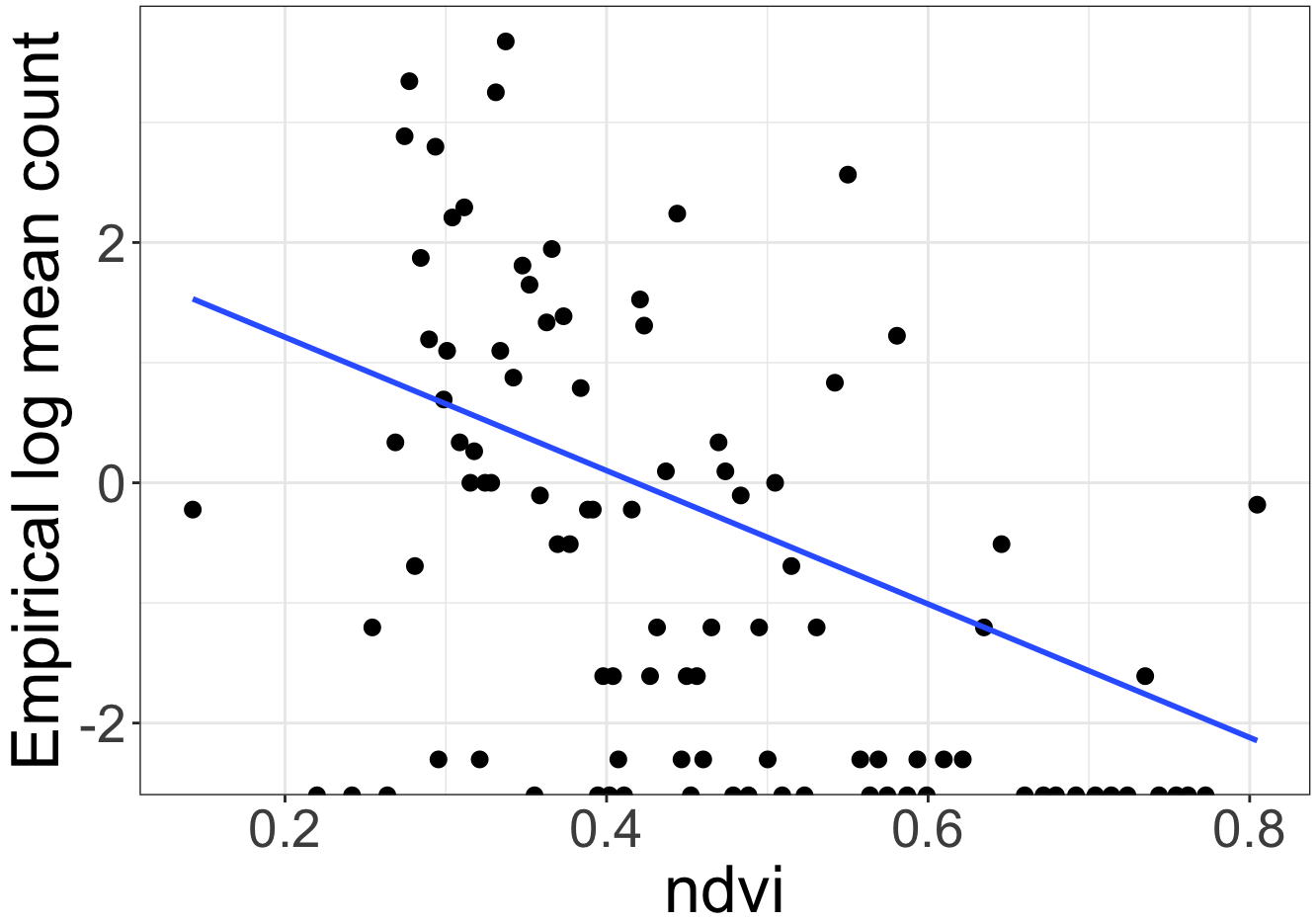

Figure 2.4: The relationship between ndvi and the count of gazelleThomsons over the 8-day period

Figure 2.5 gazelleThomsons.count vs. amRivDist

As indicated by Figure 2.5, amRivDist and the log mean of gazelleThomsons.count show up a generally positive linear shape relationship.

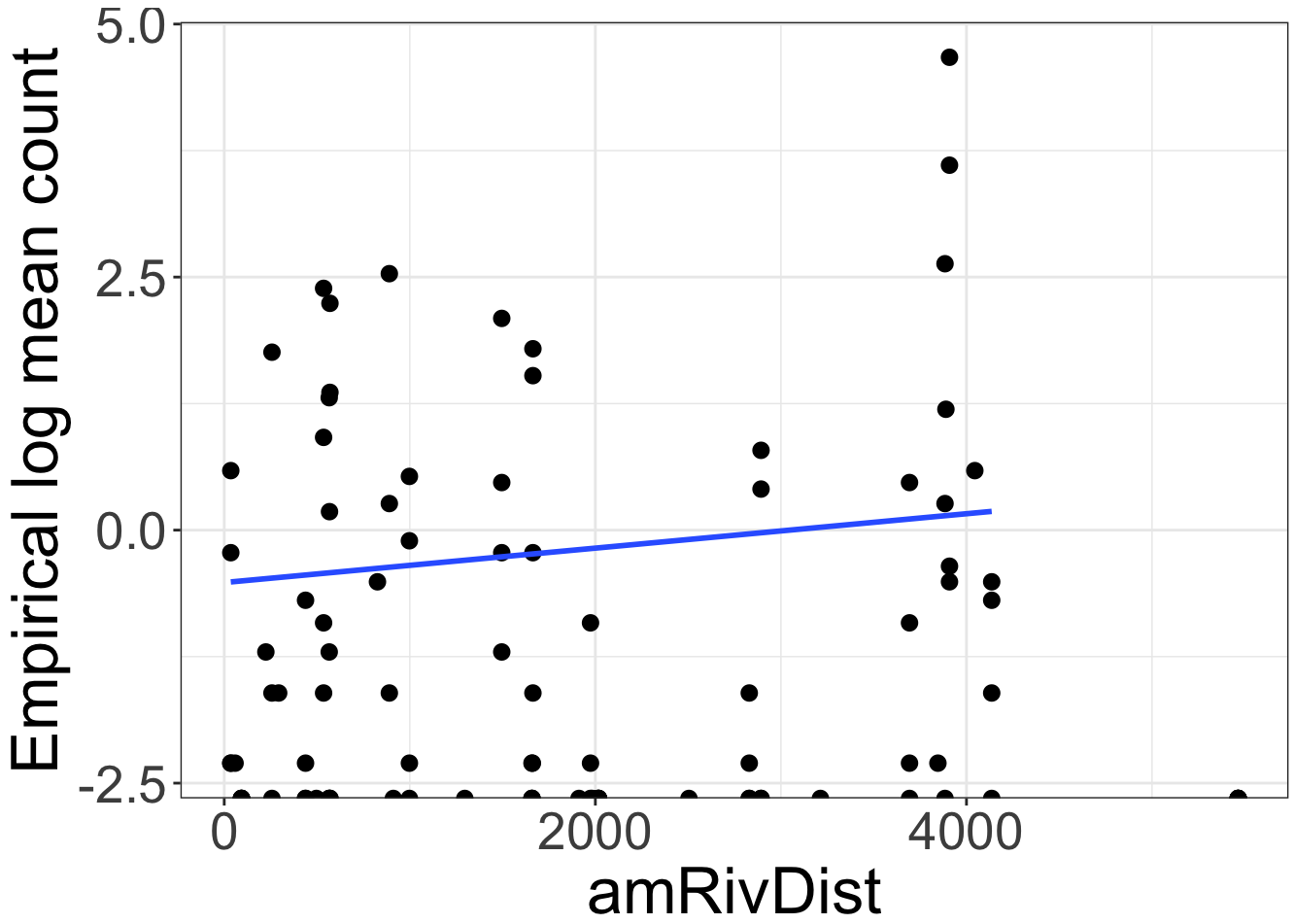

Figure 2.5: The relationship between amRivDist and the count of gazelleThomsons over the 8-day period

Figure 2.6 gazelleThomsons.count vs. LriskDry

As indicated by Figure 2.6, LriskDry and the log mean of gazelle Thomsons.count show up a generally positive linear shape relationship.

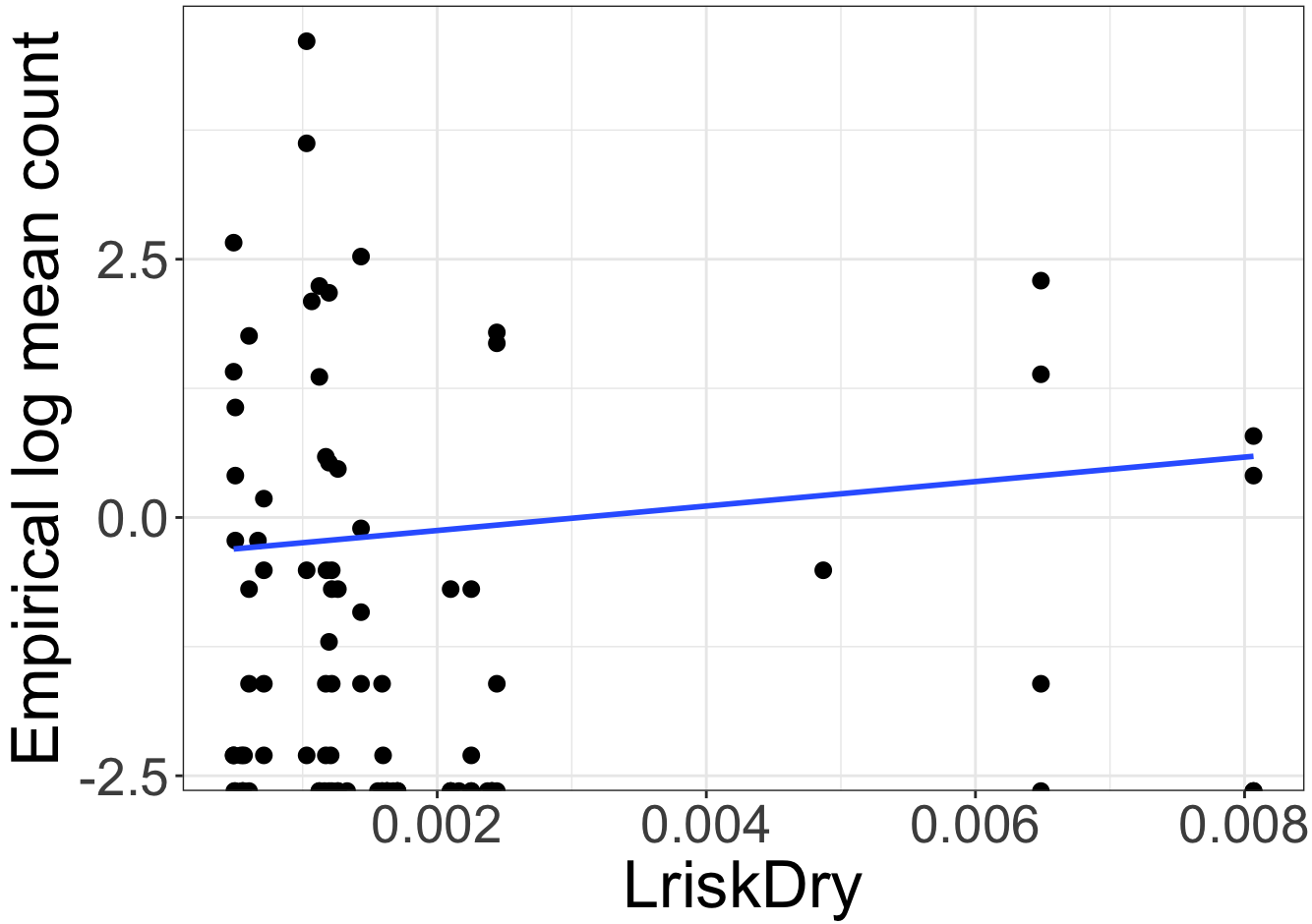

Figure 2.6: The relationship between LriskDry and the count of gazelleThomsons over the 8-day period

Figure 2.7 topi.count vs. ndvi

Figure2.7 indicates the relationship between ndvi and the count of topi over the 8-day period, with ndvi on the x-axis and log mean of topi.count on the y-axis. As shown from the figure, the log mean of topi.count shows up a generally positive linear relationship with ndvi. The relationship is not too obvious, the coeficient for the slope should be relatively small.

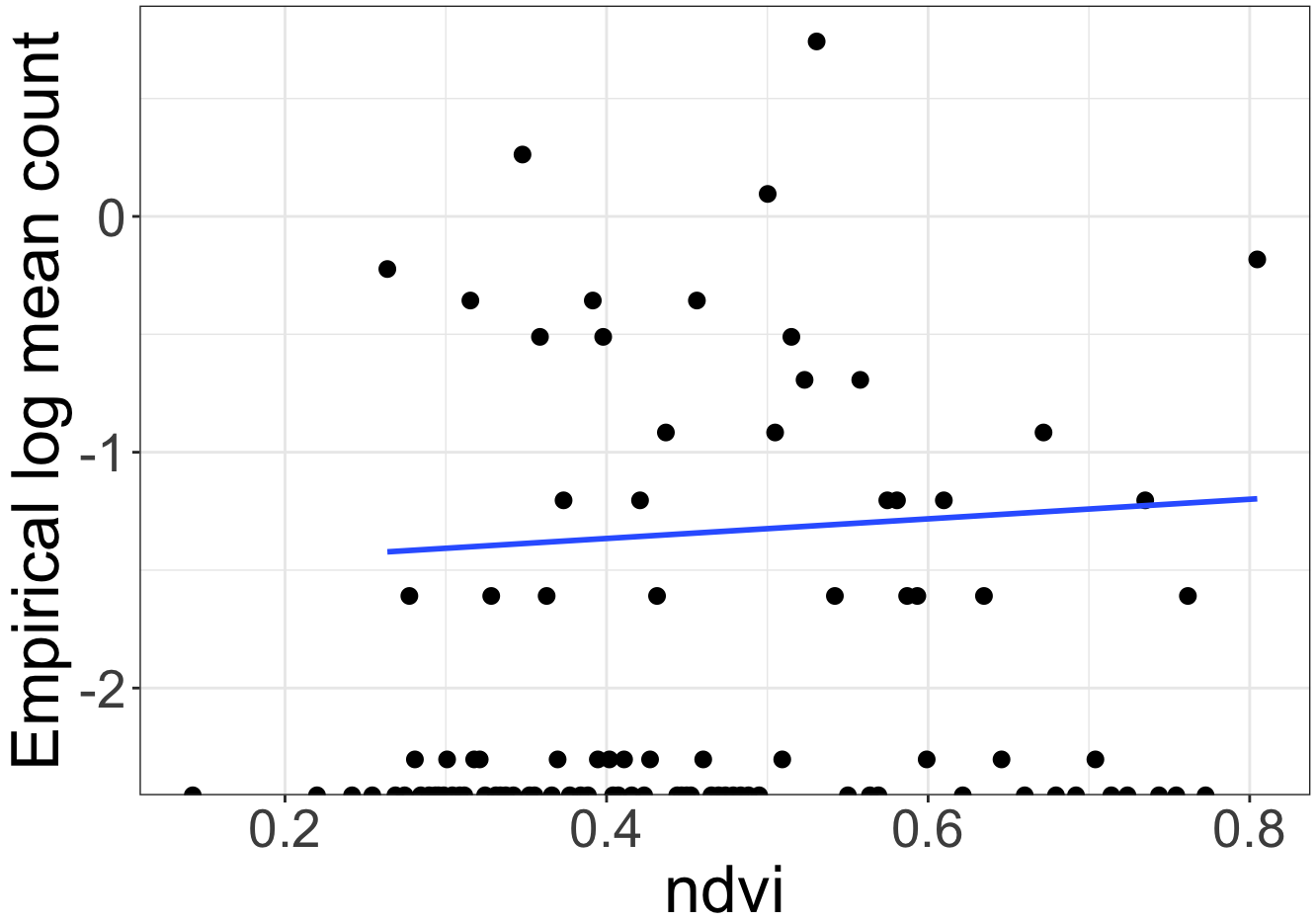

Figure 2.7: The relationship between ndvi and the count of topi over the 8-day period

Figure 2.8 topi.count vs. amRivDist

As shown from Figure 2.8, a linear relationship between amRivDist and topi.count looks reasonable, so we conclude that there is a generally negative relationship between amRivDist and topi.count.

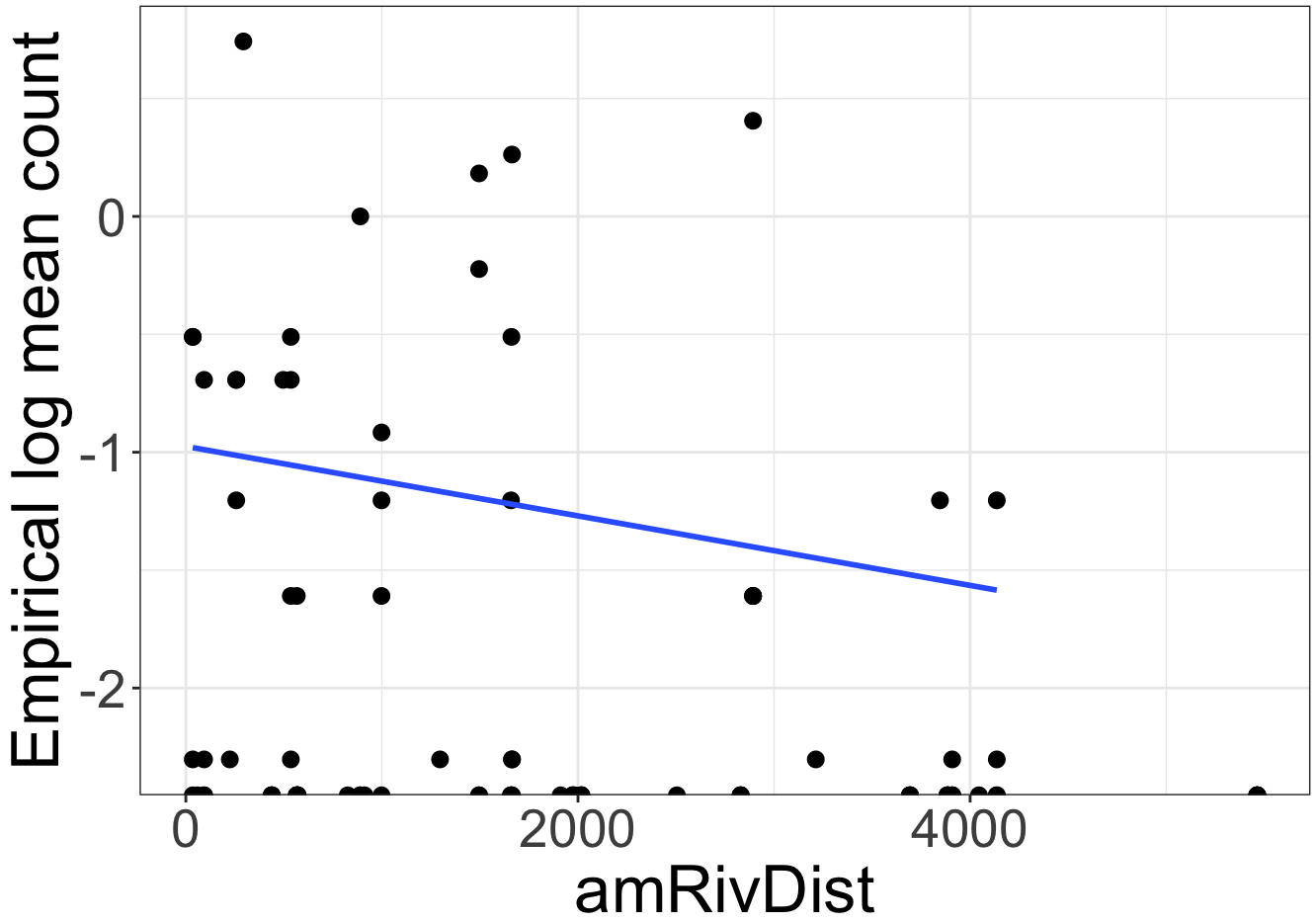

Figure 2.8: The relationship between amRivDist and the count of topi over the 8-day period

Figure 2.9 topi.count vs. LriskDry

As shown from Figure 2.9, we applied log transformation on the x variable lriskDry. The figure indicates a slightly positive linear log relationship. However, the relationship is not so obvious, we may need to test the significance in the later text.

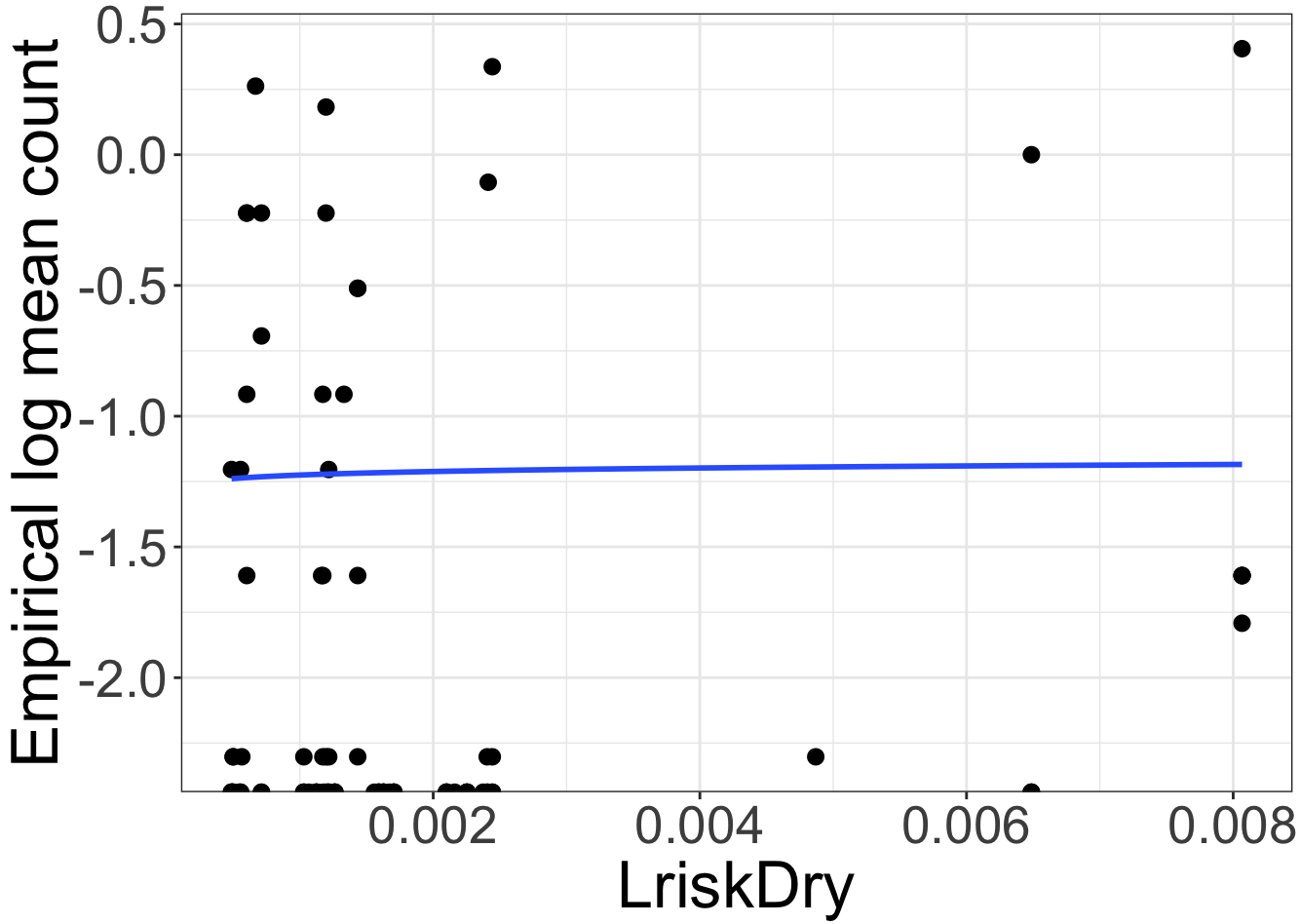

Figure 2.9: The relationship between LriskDry and the count of topi over the 8-day period

Figure 2.10 zebra.count vs. ndvi

Indicated by Figure2.10, we applied the second order polynomial transformation to the variable ndvi to rationalize the its relationship with the log mean of zebra.count.

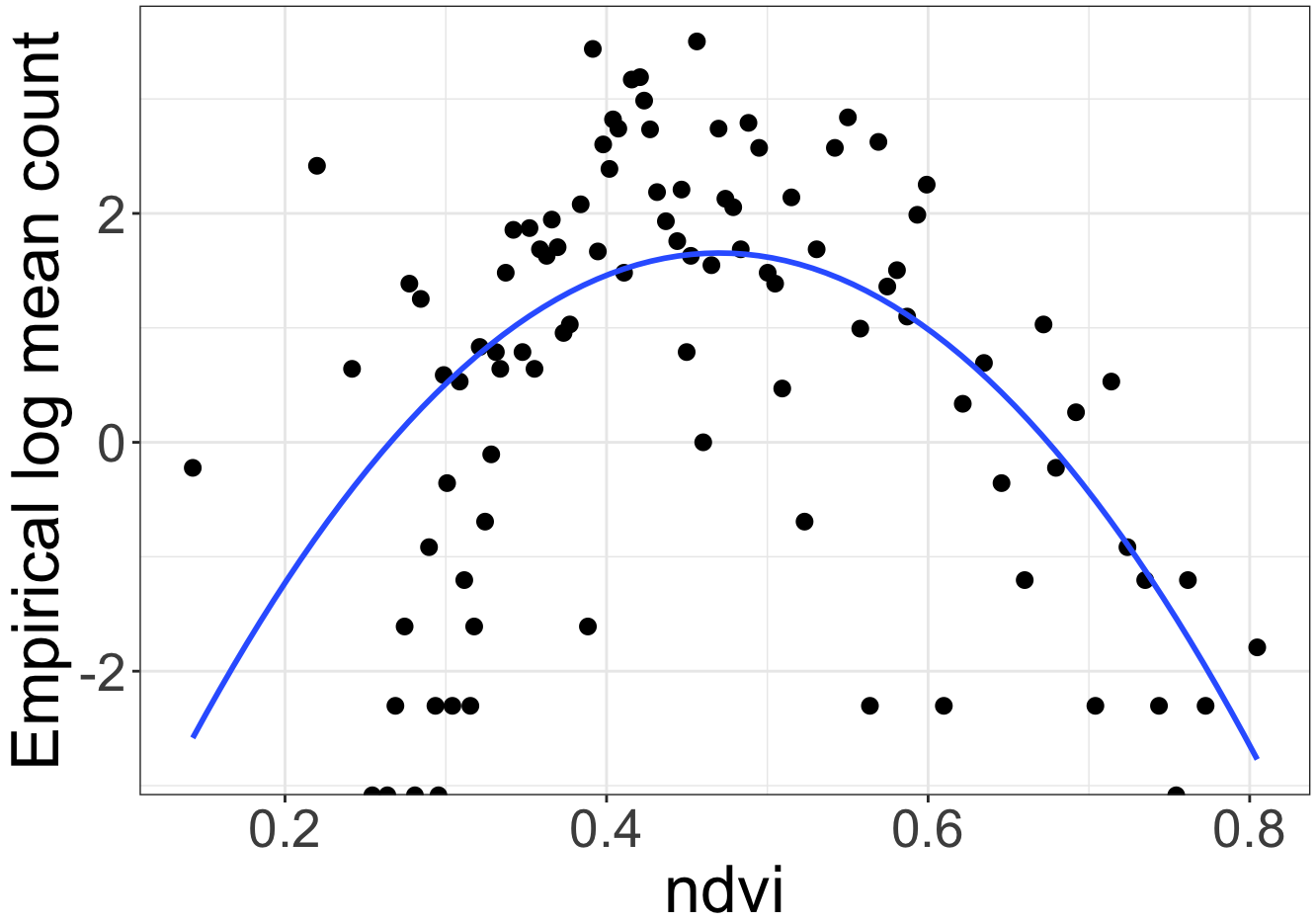

Figure 2.10: The relationship between ndvi and the count of zebra over the 8-day period

Figure 2.11 zebra.count vs. amRivDist

As indicated by Figure2.11, there is slightly a negative relationship between amRivDist and zebra.count. We can testify that in the later relationship.

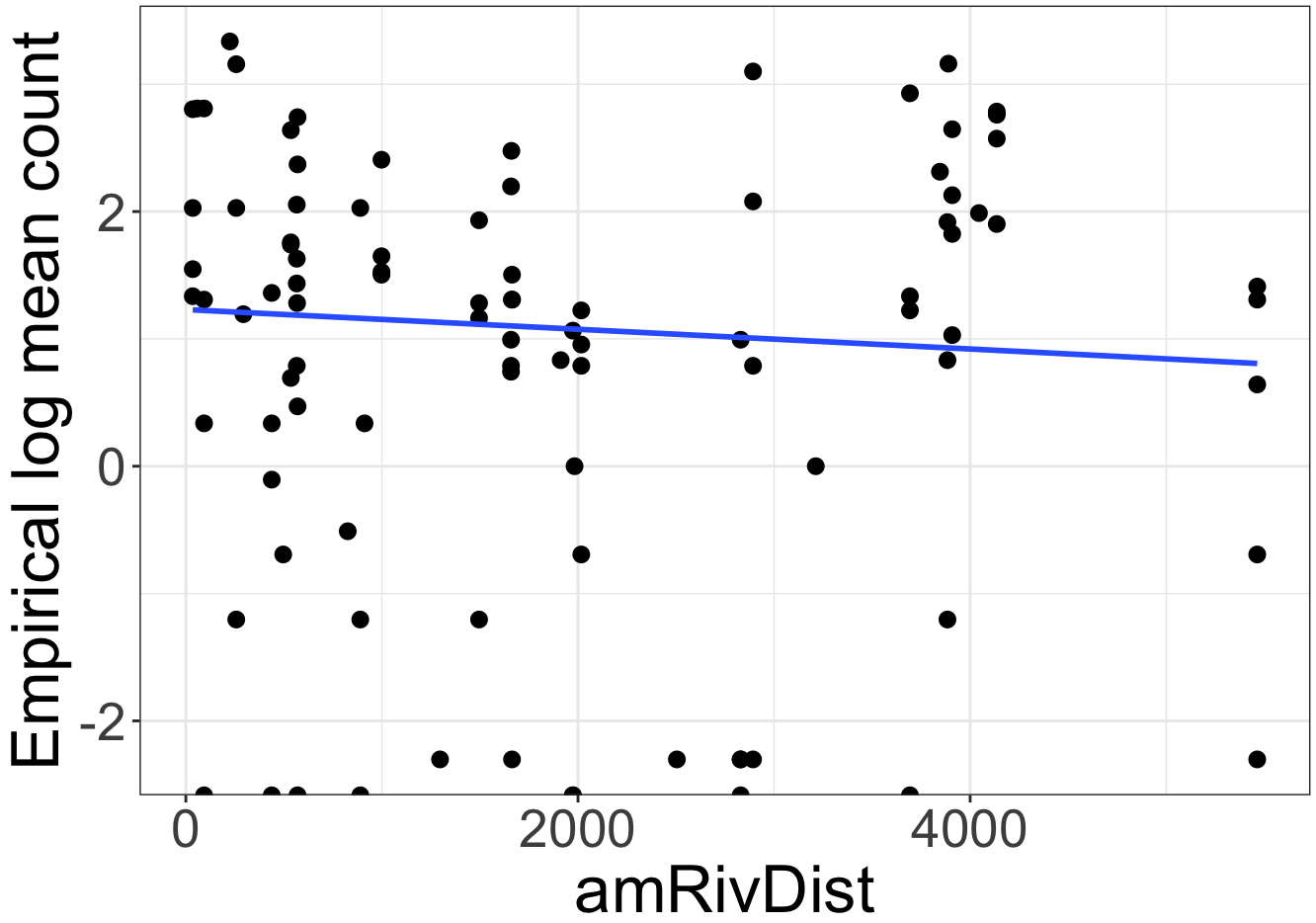

Figure 2.11: The relationship between amRivDist and the count of zebra over the 8-day period

Figure 2.12 zebra.count vs. LriskDry

From Figure 2.12, the log mean of zebra.count and LriskDry shows a negative linear relationship. But as the point mainly gathered at the left side of the Figure, the relationship can not be easily defined. So we will testify the relation further in the later hypothesis test.

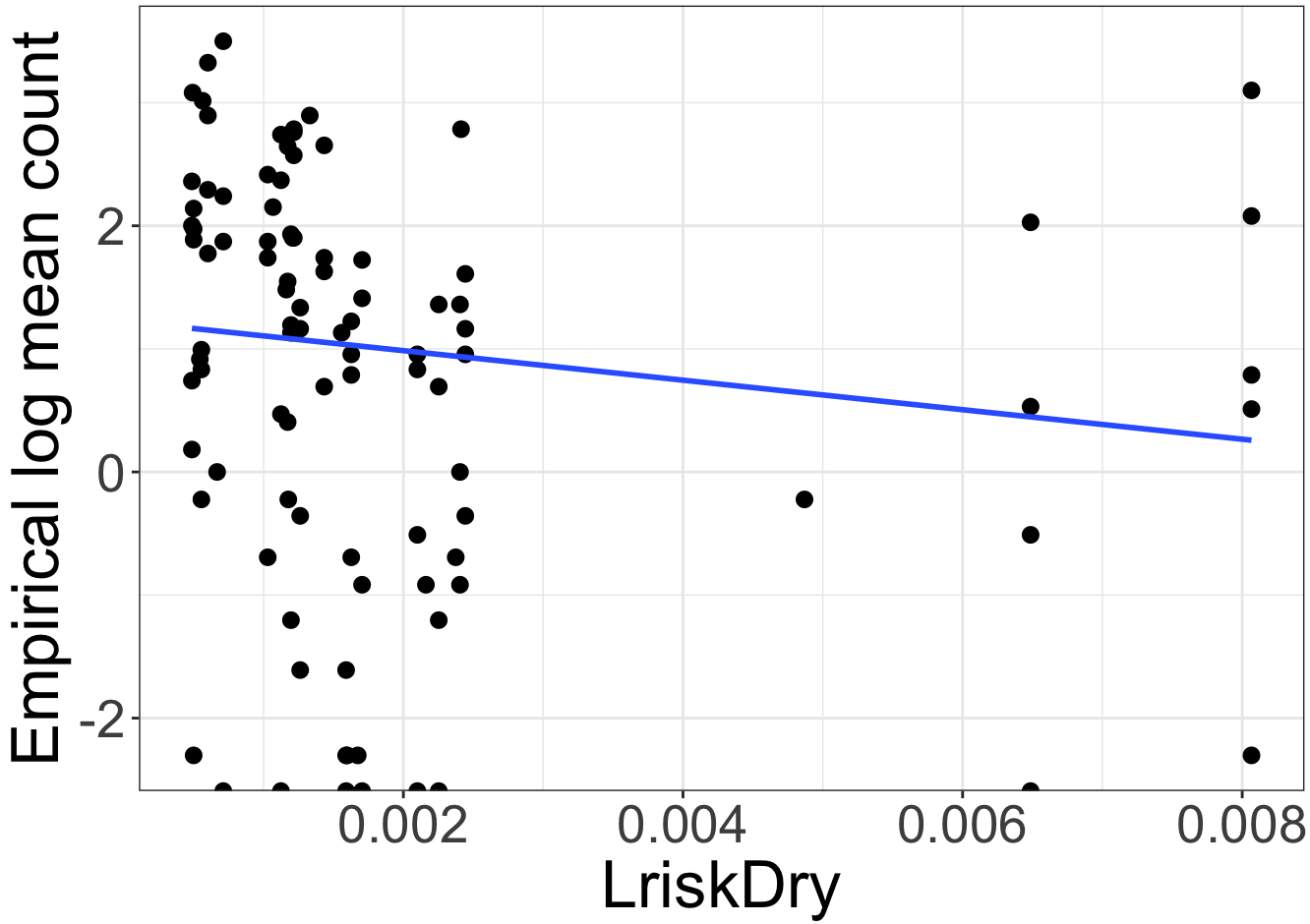

Figure 2.12: The relationship between LriskDry and the count of zebra over the 8-day period

EDA summary

For the convenience of modeling, we sum up all the transformations needed for each response variable in this section. For gazelleThomsons.count, it has linear relationship with ndvi, amRivDist, and LriskDry. For topi.count, is has linear relationship with ndvi, amRivDist and we applied log transformation on LriskDry factor. For zebra.count, it has a lienar relationship with amRivDist, and we applied second degree of polynomial transformation on variables ndvi and LriskDry.

Section 3: Modeling

Section 3.1 Importance of Environmental Variables.

As what we observed in EDA on the three response variables, there are excess zeros in the data. In order to build models where we can independently analyzed these ecxess zeros, we use ZIP models.

We will break this section into three parts where part A we build model on gazelleThomsons.count, and part B we build modell on topi.count, and part C we build model on zebra.count.

Part A Model on gazelleThomsons.count

$$P(Y_{gazelleThomson} = y) = \begin{cases} e^{-\lambda_{gazelleThomson}}(1-\alpha_{gazelleThomson})+\alpha_{gazelleThonson} & y=0 \\ \frac{e^{-\lambda_{gazelleThomson}}\lambda_{gazelleThomson}^{y}}{y!}(1-a_{gazelleThomson}) & y=0 \end{cases}$$

$$Y_{gazelleThomson} \text{ represents the count of gazelleThomson over the } 8 \text{ day time period.}$$

$$\lambda_{gazelleThomsons} = \text{average number of gazelleThomsons over the } 8 \text{ day time period.}$$

$$a_{gazelleThomsons} = \text{probability the gazelleThomsons were never present in this site.}$$

There are two parts of the ZIP model: the logistic regression model and the poisson model.From EDA, we already knew that gazelleThomsons.count has linear relationship with ndvi, amRivDist, LriskDry.And we consider ndvi, amRivDist, and LriskDry to be significant variables handling the excess zeros. Therefore, we add these factors and built model A1.

Except continuous quantitative variables, we also have other discrete and categorical variables in the dataset, so we need to add them into the model. However, as we glimpse through these variable T50, TM100, fire, we found out that are bunch of zeros in these three variables, which may have possibility to affect the performance of the fitted model, we will only add TM100 into modelA2 and perform further evaluation on T50 and fire in modelA3 and modelA4.

Further, we also believe that T50, TM100, and fire might be also reasonable variables to be considered affecting the probability of gazelleThomsons never being present at a specific site. Therefore, in modelA5, we add these three variables into the logistic regression model.

Population Model of ModelA1 to ModelA4.

$$ModelA1 : log\frac{\alpha_{gazelleThomson}}{1-\alpha_{gazelleThomson}} = \gamma_0 + \gamma_1 ndvi + \gamma_2 amRivDist + \gamma_3 LriskDry$$

$$log(\lambda_{gazelleThomson}) = \beta_0 + \beta_1 ndvi + \beta_2 amRivDist + \beta_3 LriskDry$$

$$ModelA2: log\frac{\alpha_{gazelleThomson}}{1-\alpha_{gazelleThomson}} = \gamma_0 + \gamma_1 ndvi + \gamma_2 amRivDist + \gamma_3 LriskDry$$
$$log(\lambda_{gazelleThomson}) = \beta_0 + \beta_1 ndvi + \beta_2 amRivDist + \beta_3 LriskDry + \beta_4 TM100$$

After fitting the model, we performed hypothesis test between ModelA1 and ModelA2. The drop in deviance G = 2logL(ModelA2)-2logL(Model1) = 2(-2980-(-2980)) = 0. There is no improvements in the model performance. Therefore, we will keep choice with ModelA1.

$$ModelA3: log\frac{\alpha_{gazelleThomson}}{1-\alpha_{gazelleThomson}} = \gamma_0 + \gamma_1 ndvi + \gamma_2 amRivDist + \gamma_3 LriskDry$$
$$log(\lambda_{gazelleThomson}) = \beta_0 + \beta_1 ndvi + \beta_2 amRivDist + \beta_3 LriskDry + \beta_4 T50$$

Just as we did between ModelA1 and ModelA2, after fitting ModelA3, we performed hypothesis test between ModelA2 and ModelA3. The drop in deviance G = 2logL(ModelA3)-2logL(ModelA1) = 2(-2878-(-2980)) = 204. G follows a chi-squared distribution with degrees freedom of 1. The probability after performing the hypothesis test is is 2.798881e-46. We therefore have strong evidence that T50 is a valid variable adding into the model. We will then build ModelA4 by adding new variable fire into ModelA3.

$$ModelA4: log\frac{\alpha_{gazelleThomson}}{1-\alpha_{gazelleThomson}} = \gamma_0 + \gamma_1 ndvi + \gamma_2 amRivDist + \gamma_3 LriskDry$$
$$log(\lambda_{gazelleThomson}) = \beta_0 + \beta_1 ndvi + \beta_2 amRivDist + \beta_3 LriskDry + \beta_4 T50 + \beta_5 fire$$

The drop in deviance G = 2logL(ModelA4)-2logL(ModelA3) = 2(-2871-(-2878)) = 14. G follows a chi-squared distribution with degrees freedom of 1. The probability of getting the drop in deviance 16 if ModelA3 is more significant is 0.0001828106, which smaller than 0.005. We therefore have prefer ModelA4 over ModelA3. And finally, in ModelA5, we want to add TM100 and fire into the logistic regression model.

$$ModelA5: log\frac{\alpha_{gazelleThomson}}{1-\alpha_{gazelleThomson}} = \gamma_0 + \gamma_1 ndvi + \gamma_2 amRivDist + \gamma_3 LriskDry + \gamma_4 TM100 + \gamma_5 T50 + \gamma_6 fire$$
$$log(\lambda_{gazelleThomson}) = \beta_0 + \beta_1 ndvi + \beta_2 amRivDist + \beta_3 LriskDry + \beta_4 T50 + +\beta_5 fire$$

The drop in deviance G = 2logL(ModelA5)-2logL(ModelA4) = 2(-2869-(-2871)) = 4. G follows a chi-squared distribution with degrees freedom of 2.

The probability of getting the drop in deviance 4 if ModelA3 is more significant is 0.1353353. We therefore will still choose ModelA4 as our model.

Conclusion on gazelleThomsonFinal model:

Our Final model for the response variable gazelleThomsons.count will be:

$$gazelleThomsonFinal: log\frac{\alpha_{gazelleThomson}}{1-\alpha_{gazelleThomson}} = -0.3900 + 5.162ndvi - 0.00008381amRivDist + 0.3286LriskDry$$

$$log(\lambda_{gazelleThomson}) = 3.081 - 2.969ndvi + 0.0004118amRivDist - 0.6517LriskDry - 0.002676T50 + 0.5012fire$$

ndvi, LriskDry have a positive relationship with the count of gazelleThomson over the 8 day time period to be 0, while amRivDist has a negative relationship. And the log mean of the count of gazelleThomson over the day time period is positively associated with RivDist (the distance from the site to the nearest river) and fire (the presence of wildfire over the past 30 days in the site), and is negatively associated with ndvi(the 'greenness' of the site) and LriskDry(the risk of lion predation), T50(the number of trees within 50 meters of the location). Specifically, ndvi, LriskDry, and fire impact the log mean of gazelleThomsons.count the most.

Part B Model on topi.count

$$P(Y_{topi} = y) = \begin{cases} e^{-\lambda_t topi}(1-\alpha_{topi})+\alpha_{topi} & y=0 \\ \frac{e^{-\lambda_{topi}}\lambda_{topi}^{y}}{y!}(1-a_{topi}) & y=0 \end{cases}$$

$$Y_{topi}\ represents\ the\ count\ of\ topi\ over\ the\ 8\ day\ time\ period.$$

$$\lambda_{topi} = average\ number\ of\ topi\ over\ the\ 8\ day\ time\ period.$$

$$a_{topi} = probability\ the\ topi\ were\ never\ present\ in\ this\ site.$$

From Section2 EDA, we knew that the log mean of topi.count has a linear relationship with ndvi and amRivDist, and log transformation has applied to LriskDry. However, as shown from Figure 2.7, the relationship between log mean of topi.count and ndvi is not too obvious. Therefore, we will only add amRivDist, log(LriskDry) to ModelB1, and add ndvi to ModelB2 to test the significance of variable ndvi.Also, we want to test whether TM100,T50,fire impact the log mean of topi in ModelB3, ModelB4, ModelB5, and test the necessary variables adding to the logistic regression model.

Population Model of ModelB1 to ModelB5.

$$ModelB1: log\frac{\alpha_{topi}}{1-\alpha_{topi}} = \gamma_0 + \gamma_1 amRivDist + \gamma_2 log(LriskDry)$$

$$log(\lambda_{topi}) = \beta_0 + \beta_1 amRivDist + \beta_2 log(LriskDry)$$

$$ModelB2: log\frac{\alpha_{topi}}{1-\alpha_{topi}} = \gamma_0 + \gamma_1 ndvi + \gamma_2 amRivDist + \gamma_3 log(LriskDry)$$
$$log(\lambda_{topi}) = \beta_0 + \beta_1 ndvi + \beta_2 amRivDist + \beta_3 log(LriskDry)$$

We then performed the same hypothesis test on ModelB1 and ModelB2 after fitting in the data. The drop in deviance is G = 2logL(ModelB2)-2logL(ModelB1) = 2(-357.5-(-358.7)) = 2.4. G follows a chi-squared distribution with degrees freedom of 1. The probability of getting the drop in deviance 2.4 if ModelB1 is more significant is 0.1213353, which is not significant enough to add the new variable ndvi. Therefore, we will keep our choice with ModelB2. And we will build ModelB3 based on ModelB1 by adding new variable TM100.

$$ModelB3: log\frac{\alpha_{topi}}{1-\alpha_{topi}} = \gamma_0 + \gamma_1 amRivDist + \gamma_2 log(LriskDry)$$
$$log(\lambda_{topi}) = \beta_0 + \beta_2 amRivDist + \beta_3 log(LriskDry) + TM100$$

Comparing ModelB3 with ModelB1,the drop in deviance is G = 2logL(ModelB3)-2logL(ModelB1) = 2(-356.3-(-358.7)) = 4.8 G follows a chi-squared distribution with degrees freedom of 1. The probability of getting the drop in deviance 0.8 if ModelB1 is more significant is 0.3710934. Therefore, we will keep our choice with ModelB1. Then, we build ModelB4 based on ModelB1 by adding variable T50.

$$ModelB4: log\frac{\alpha_{topi}}{1-\alpha_{topi}} = \gamma_0 + \gamma_1 amRivDist + \gamma_2 log(LriskDry)$$
$$log(\lambda_{topi}) = \beta_0 + \beta_2 amRivDist + \beta_3 log(LriskDry) + T50$$

Comparing ModelB4 with ModelB1, the drop in deviance is G = 2logL(ModelB4)-2logL(ModelB1) = 2(-356.3-(-358.7)) = 4.8 G follows a chi-squared distribution with degrees freedom of 1. The probability of getting the drop in deviance 2.6 if ModelB1 is more significant is 0.3710934, which is still too small to be in favour of ModelB4. Therefore, we will keep our choice with ModelB1. Then, we build ModelB5 based on ModelB1 by adding variable fire.

$$ModelB5: log\frac{\alpha_{topi}}{1-\alpha_{topi}} = \gamma_0 + \gamma_1 amRivDist + \gamma_2 log(LriskDry)$$
$$log(\lambda_{topi}) = \beta_0 + \beta_2 amRivDist + \beta_3 log(LriskDry) + fire$$

Comparing ModelB5 with ModelB1, the drop in deviance is G = 2logL(ModelB5)-2logL(ModelB1) = 2(-358.7 -(-358.7)) = 0. There is no improvements in by adding the variable fire in the drop in deviance. Therefore, our final model for the response variable topi.count will be ModelB1.

Conclusion on topiFinal model:

Our Final model for the response variable topi.count will be:

$$topiFinal: log\frac{\alpha_{topi}}{1-\alpha_{topi}} = 1.2399062 + 0.0002155amRivDist - 0.1365214log(LriskDry)$$
$$log(\lambda_{topi}) = 2.517228 - 0.000266amRivDist + 0.203947log(LriskDry)$$

amRivDist has a strong negative relationship with the log mean of count of topi over the 8 day time period to be 0, while log(LriskDry) has a negative impact. And the log mean of the count of topi over the day time period is positively associated with log(LriskDry) and negatively associated with RivDist. The impact of amRivDist given to the log mean of topi.count is relatively small.

Part C Model on zebra.count

$$P(Y_{zebra} = y) = \begin{cases} e^{-\lambda_{zebra}}(1-\alpha_{zebra}) + \alpha_{zebra} & y=0 \\ \frac{e^{-\lambda_{zebra}}\lambda_{zebra}^{y}}{y!}(1-a_{zebra}) & y=0 \end{cases}$$

$$Y_{zebra}\ represents\ the\ count\ of\ zebra\ over\ the\ 8\ day\ time\ period.$$

$$\lambda_{zebra} = average\ number\ of\ zebra\ over\ the\ 8\ day\ time\ period.$$

$$a_{zebra} = probability\ the\ zebra\ were\ never\ present\ in\ this\ site.$$

From EDA, we applied second degree of polynomial transformation to ndvi, and we also see a seemingly linear relationship between log mean of zebra.count and amRivDist, LriskDry. So we will build ModelC1 with only variable ndvi, and add amRivDist, LriskDry, TM100, T50, and fire consecutively into the different model to testify their statistical significance, and use several other models to test what variables should be put in the logistic regression model.

Population Model of ModelC1 to ModelC6.

$$ModelC1: log\frac{\alpha_{zebra}}{1-\alpha_{zebra}} = \gamma_0 + \gamma_1 ndvi$$
$$log(\lambda_{zebra}) = \beta_0 + \beta_1 ndvi + \beta_2 ndvi^2$$

$$ModelC2 : log\frac{\alpha_{zebra}}{1-\alpha_{zebra}} = \gamma_0 + \gamma_1 ndvi$$
$$log(\lambda_{zebra}) = \beta_0 + \beta_1 ndvi + \beta_2 ndvi^2 + \beta_3 amRivDist$$

We then perform hypothesis test on ModelC1 and ModelC2. The drop in deviance is G = 2logL(ModelC2)-2logL(ModelC1) = 2(-5437-(-5466)) = 58. G follows a chi-squared distribution with degrees freedom of 1.Use G value of 58 and df of 1, we get the probability equals to 2.621178e-14,. Therefore, we have strong evidence that there is a linear relationship between amRivDist and log mean of zebra.count. ModelC2 is chosen over ModelC1. We then build ModelC3 by adding variable LriskDry to ModelC2.

$$ModelC3 : log\frac{\alpha_{zebra}}{1-\alpha_{zebra}} = \gamma_0 + \gamma_1 ndvi$$
$$log(\lambda_{zebra}) = \beta_0 + \beta_1 ndvi + \beta_2 ndvi^2 + \beta_3 amRivDist + \beta_4 LriskDry$$

Same as all the previous hypothesis test, we calculate the drop in deviance from ModelC2 to ModelC3. G = 2logL(ModelC3)-2logL(ModelC2) = 2(-5416-(-5437)) = 42.G follows the chi squared distribution with degree freedom of 1. The probability we get is 9.127342e-11. Therefore, we have strong evidence that there is a linear relationship between log mean of zebra.count and LriskDry.

Then, we build and perform hypothesis test on ModelC4, ModelC5, ModelC6 to decide whether to add the discrete variable TM100, T50, and fire into the model.

$$ModelC4 : log\frac{\alpha_{zebra}}{1-\alpha_{zebra}} = \gamma_0 + \gamma_1 ndvi$$
$$log(\lambda_{zebra}) = \beta_0 + \beta_1 ndvi + \beta_2 ndvi^2 + \beta_3 amRivDist + \beta_4 LriskDry + TM100$$

G = 2logL(ModelC4)-2logL(ModelC3) = 2(-5386-(-5416)) = 60.It follows chi-squared distribution of 1 degree freedom. The calculated probability is 9.485738e-15. Therefore, there is a relationship between log mean of zebra.count and TM100. We then choose ModelC4 over ModelC3, and add T50 to ModelC4 to build ModelC5.

$$ModelC5 : log\frac{\alpha_{zebra}}{1-\alpha_{zebra}} = \gamma_0 + \gamma_1 ndvi$$
$$log(\lambda_{zebra}) = \beta_0 + \beta_1 ndvi + \beta_2 ndvi^2 + \beta_3 amRivDist + \beta_4 LriskDry + TM100 + T50$$

G = 2logL(ModelC5)-2logL(ModelC4) = 2(-5373-(-5386)) = 26.It follows chi-squared distribution of 1 degree freedom. The calculated probability is 3.414174e-07. Therefore, there is a relationship between log mean of zebra.count and T50. We then choose ModelC5 over ModelC4. Finally, we add fire to ModelC5 to test the significance of variable fire.

$$ModelC6: log\frac{\alpha_{zebra}}{1-\alpha_{zebra}} = \gamma_0 + \gamma_1 ndvi$$

$$log(\lambda_{zebra}) = \beta_0 + \beta_1 ndvi + \beta_2 ndvi^2 + \beta_3 amRivDist + \beta_4 LriskDry + TM100 + T50 + fire$$

G = 2logL(ModelC6)-2logL(ModelC5) = 2(-5373-(-5373)) = 0.There is no improvement in the model performance by adding variable fire. Therefore, we will choose ModelC5. And then, we are going to add variables amRivDist, LriskDry, TM100, and T50 into ModelB5 to test whether we should put these variables into the logistic regression model.

$$ModelC7: log\frac{\alpha_{zebra}}{1-\alpha_{zebra}} = \gamma_0 + \gamma_1 ndvi + \gamma_2 amRivDist + \gamma_3 LriskDry + \gamma_4 TM100 + \gamma_5 T50$$

$$log(\lambda_{zebra}) = \beta_0 + \beta_1 ndvi + \beta_2 ndvi^2 + \beta_3 amRivDist + \beta_4 LriskDry + TM100 + T50$$

G = 2logL(ModelC7)-2logL(ModelC6) = 2(-5363-(-5373)) = 20.G follows the chi squared distribution of degree freedom 4. The probability statistic is 0.0004993992, which is smaller than 0.005. Therefore,we will add these variables into the logistic regression model, and our final model should be ModelC7.

Conclusion on zebraFinal model:

Our Final model for the response variable zebra.count will be:

$$zebraFinal: log\frac{\alpha_{zebra}}{1-\alpha_{zebra}} = 0.4317 - 0.01434 ndvi + 0.00005262 amRivDist + 0.7283 LriskDry + 0.2101 TM100 - 0.001077 T50$$

$$log(\lambda_{zebra}) = -3.342 + 0.2789 ndvi - 0.3080 ndvi^2 + 0.00008309 amRivDist - 0.4423 LriskDry - 0.2336 TM100 + 0.0003457 T50$$

amRivDist and LriskDry haVE a strong negative relationship with the log mean of count of topi over the 8 day time period to be 0, while ndvi has a negative impact. the log mean of zebra.count forms a positive relationshiop with ndvi, amRivDist and T50, but have negative relationship with the second order of ndvi, and LriskDry and TM100. Overall, LriskDry and TM100 has relatively large impact on the log mean of zebra.count.

```
pchisq(20, 4,lower.tail = FALSE)
```

```
## [1] 0.0004993992
```

Interpretation on the relevant coefficient:

A. gazelleThomson

$$gazelleThomsonFinal: log\frac{\alpha_{gazelleThomson}}{1-\alpha_{gazelleThomson}} = -0.3900 + 5.162ndvi - 0.00008381amRivDist + 0.3286LriskDry$$

$$log(\lambda_{gazelleThomson}) = 3.081 - 2.969ndvi + 0.0004118amRivDist - 0.6517LriskDry - 0.002676T50 + 0.5012fire$$

logistic Regression Model - ndvi:

coefficient : 5.162

For every unit increase in ndvi, the odds of gazelleThomsons never being present in the site is e^5.162 times higher for the site.

amRivDist:

coefficient : -0.00008381

For every unit increase in amRivDist, the odds of gazelleThomsons never being present in the site is e^0.00008381 times lower for the site.

lriskDry:

coefficient : 0.3286

For every unit increase in LriskDry, the odds of gazelleThomsons never being present in the site is e^0.3286 times higher for the site.

Poisson Regression - ndvi

coefficient : -2.969

Among sites that have gazelleThomsons present before, for every unit increase in the ‘greenness’ of the site,the expected number of count of gazelleThomsons is lowered by a factor of e^2.969 , holding other variables constant.

amRivDist

coefficient : 0.0004118

Among sites that have gazelleThomsons present before, for every unit increase in the distance from the site to that nearest river, the expected number of count of gazelleThomsons is increased by a factor of e^0.0004118, holding other variables constant.

LriskDry

coefficient : -0.6517

Among sites that have gazelleThomsons present before, for every unit increase in the risk of lion predation, the expected number of count of gazelleThomsons is lowered by a e^0.6517 , holding other variables constant.

T50

coefficient : -0.002676

Among sites that have gazelleThomsons present before, for every unit increase in the number of trees within 50 meters of the site, the expected number of count of gazelleThomsons is lowered by a factor of e^0.002676, holding other variables constant.

fire

coefficient : 0.5012

Among sites that have gazelleThomsons present before, the expected number of count of gazelleThomsons is e^0.5012 times higher for the site has detected wildfire in the past 30 days, holding other variables constant.

$$topiFinal : log\frac{\alpha_{topi}}{1-\alpha_{topi}} = 1.2399062 + 0.0002155amRivDist - 0.1365214log(LriskDry)$$
$$log(\lambda_{topi}) = 2.517228 - 0.000266amRivDist + 0.203947log(LriskDry)$$

Logistic Regression - amRivDist

coefficient : 0.0002155

For every unit increase in amRivDist, the odds of gazelleThomsons never being present in the site is e^0.0002155 times higher for the site.

log(LriskDry)

coefficient : -0.1365214

For every unit increase in log(LriskDry), the odds of gazelleThomsons never being present in the site is e^0.3286 times lower for the site.

poisson Regression - amRivDist

coefficient: -0.000266

Among sites that have topi present before, for every unit increase in the distance from the site to that nearest river, the expected number of count of topi is decreased by a factor of e^0.000266, holding other variables constant.

LriskDry

coefficient : 0.203947

Among sites that have topi present before, for every unit increase in the log risk of lion predation, the expected number of count of topi is increased by a e^0.203947, holding other variables constant.

$$zebraFinal: log\frac{\alpha_{zebra}}{1-\alpha_{zebra}} = 0.4317 - 0.01434ndvi + 0.00005262amRivDist + 0.7283LriskDry + 0.2101TM100 - 0.001077T50$$

$$log(\lambda_{zebra}) = -3.342 + 0.2789ndvi - 0.3080ndvi^2 + 0.00008309amRivDist - 0.4423LriskDry - 0.2336TM100 + 0.0003457T50$$

logistic Regression Model - ndvi:

coefficient : -0.01434

For every unit increase in ndvi, the odds of gazelleThomsons never being present in the site is e^0.01434 times lower for the site.

amRivDist:

coefficient : 0.00005262

For every unit increase in amRivDist, the odds of gazelleThomsons never being present in the site is e^0.00005262 times higher for the site.

lriskDry:

coefficient : 0.7283

For every unit increase in LriskDry, the odds of gazelleThomsons never being present in the site is e^0.7283 times higher for the site.

TM100

coefficient : 0.2101

For every unit increase in TM100, the odds of gazelleThomsons never being present in the site is e^0.2101 times higher for the site.

T50

coefficient : -0.001077

For every unit increase in T50, the odds of gazelleThomsons never being present in the site is e^0.001077 times lower for the site.

Poisson Regression -ndvi

coefficient : 0.2789

Among sites that have zebra present before, for every unit increase in the ‘greenness’ of the site,the expected number of count of zebra is increased by a factor of e^0.2789, holding other variables constant.

ndvi^2

coefficient : -0.3080

Among sites that have zebra present before, for every unit increase in ndvi^2 of the site,the expected number of count of zebra is decreased by a factor of e^0.3080, holding other variables constant.

amRivDist

coefficient : 0.00008309

Among sites that have zebra present before, for every unit increase in the distance from the site to that nearest river, the expected number of count of gazelleThomsons is increased by a factor of e^0.00008309, holding other variables constant.

LriskDry

coefficient : -0.4423

Among sites that have zebra present before, for every unit increase in the risk of lion predation, the expected number of count of gazelleThomsons is lowered by a e^0.4423, holding other variables constant.

TM100

coefficient : -0.2336

Among sites that have zebra present before, for every unit increase in the number of termite mounds within 100 meters of the location, the expected number of count of zebra is lowered by a e^0.0.2336, holding other variables constant.

T50

coefficient : 0.0003457T50

Among sites that have zebra present before, for every unit increase in the number of trees within 50 meters of the site, the expected number of count of zebra is increased by a factor of e^0.0003457T50, holding other variables constant.

Section 3.2 Comparing species

We will write the final models here again for the three species for the convenience of comparison.

$$gazelleThomsonFinal: log\frac{\alpha_{gazelleThomson}}{1-\alpha_{gazelleThomson}} = -0.3900 + 5.162ndvi - 0.00008381amRivDist + 0.3286LriskDry$$

$$log(\lambda_{gazelleThomson}) = 3.081 - 2.969ndvi + 0.0004118amRivDist - 0.6517LriskDry - 0.002676T50 + 0.5012fire$$

$$topiFinal: log\frac{\alpha_{topi}}{1-\alpha_{topi}} = 1.2399062 + 0.0002155amRivDist - 0.1365214log(LriskDry)$$

$$log(\lambda_{topi}) = 2.517228 - 0.000266amRivDist + 0.203947log(LriskDry)$$

$$zebraFinal: log\frac{\alpha_{zebra}}{1-\alpha_{zebra}} = 0.4317 - 0.01434ndvi + 0.00005262amRivDist + 0.7283LriskDry + 0.2101TM100 - 0.001077T50$$

$$log(\lambda_{zebra}) = -3.342 + 0.2789ndvi - 0.3080ndvi^2 + 0.00008309amRivDist - 0.4423LriskDry - 0.2336TM100 + 0.0003457T50$$

We can first compare the logistic regression part of three models. LriskDry has great impact on the probability of the presence of the species. While it is positively related to the odds of gazelle and zebra not being present, it is negatively related to the odds of topi not being present. In another word, LriskDry gives a negative impact on the gazelleThomsons.count and zebra.count, but contributes negatively to the topi.count. Also, amRivDist is a common factor affecting the counts of three species, but we can see the coefficient is a relatively small number. So, a reasonable assumption is that amRivDist may not make a big difference on the species counts. Another interesting observation is that ndvi is positively related to zebra.count, which matches what we expect, while it is negatively related to gazelleThomsons.count. We may want to conduct further study and repetitive researches to test the accuracy of the results. And the zebra.count is also affected by TM100 and T50.

Section 4 Discussion

Environmental factors definitely have some relationship with the number of species count in one location. And according to our analysis, the counts of these three species - gazelle Thomsons, topi, and zebra are impacted by different environmental variables with different coefficient. For some variables, they even give an opposite correlation. For example, ndvi is positively related to the counts of zebra but it is negatively related to the counts of gazelleThomsons. To find out the reason that explains the difference, we may need to consider add in more data to see whether the result is an occasion.

As we view through the subset data, the value of the species count is relatively small, which makes the analysis hard in some way. So in the further study, we can probably increase the observing time period, and add in more environmental variables to improve the performance of the model.