Project 2

Alex Zhang

2022-11-11

```r
logodds_plot <- function(data, num_bins, bin_method,
                         x_name, y_name, grouping = NULL, reg_formula = y ~ x){

  if(is.null(grouping)){
    dat <- data.frame(x = data %>% pull(x_name),
                      y = data %>% pull(y_name),
                      group = 1)
  } else {
    dat <- data.frame(x = data %>% pull(x_name),
                      y = data %>% pull(y_name),
                      group = data %>% pull(grouping))
  }

  if(bin_method == "equal_size"){
    logodds_table <- dat %>%
      drop_na() %>%
      arrange(group, x) %>%
      group_by(group) %>%
      mutate(obs = y,
             bin = rep(1:num_bins,
                       each=ceiling(n()/num_bins))[1:n()]) %>%
      group_by(bin, group) %>%
      summarize(mean_x = mean(x),
                prop = mean(c(obs, 0.5)),
                num_obs = n()) %>%
      ungroup() %>%
      mutate(logodds = log(prop/(1 - prop)))
  } else {
    logodds_table <- dat %>%
      drop_na() %>%
      group_by(group) %>%
      mutate(obs = y,
             bin = cut(x,
                       breaks = num_bins,
                       labels = F)) %>%
      group_by(bin, group) %>%
      summarize(mean_x = mean(x),
                prop = mean(c(obs, 0.5)),
                num_obs = n()) %>%
      ungroup() %>%
      mutate(logodds = log(prop/(1 - prop)))
  }
  if(is.null(grouping)){
    logodds_table %>%
      ggplot(aes(x = mean_x,
                 y = logodds)) +
      geom_point(size=2) +
      geom_smooth(se=F, method="lm", formula = reg_formula) +
      theme_bw() +
      labs(x = x_name,
           y = "Empirical log odds") +
      theme(text = element_text(size=25))
  } else {
    logodds_table %>%
      ggplot(aes(x = mean_x,
                 y = logodds,
                 color = group,
                 shape = group)) +
      geom_point(size=2) +
      geom_smooth(se=F, method="lm", formula = reg_formula) +
      theme_bw() +
      labs(x = x_name,
           y = "Empirical log odds",
           color = grouping,
           shape = grouping) +
      theme(text = element_text(size=25))
  }

}
```

```
logmean_plot <- function(data, num_bins, bin_method,
                         x, y, grouping = NULL, reg_formula = y ~ x){

  if(is.null(grouping)){
    dat <- data.frame(x = data[,x],
                      y = data[,y],
                      group = 1)
  } else {
    dat <- data.frame(x = data[,x],
                      y = data[,y],
                      group = data[,grouping])
  }

  if(bin_method == "equal_size"){
    log_table <- dat %>%
      drop_na() %>%
      arrange(group, x) %>%
      group_by(group) %>%
      mutate(obs = y,
             bin = rep(1:num_bins,
                       each=ceiling(n()/num_bins))[1:n()]) %>%
      group_by(bin, group) %>%
      summarize(mean_x = mean(x),
                mean_y = mean(obs),
                num_obs = n()) %>%
```

```r
      ungroup() %>%
      mutate(log_mean = log(mean_y))
  } else {
    log_table <- dat %>%
      drop_na() %>%
      group_by(group) %>%
      mutate(obs = y,
             bin = cut(x,
                       breaks = num_bins,
                       labels = F)) %>%
      group_by(bin, group) %>%
      summarize(mean_x = mean(x),
                mean_y = mean(obs),
                num_obs = n()) %>%
      ungroup() %>%
      mutate(log_mean = log(mean_y))
  }

  if(is.null(grouping)){
    log_table %>%
      ggplot(aes(x = mean_x,
                 y = log_mean)) +
      geom_point(size=2.5) +
      geom_smooth(se=F, method="lm", formula = reg_formula) +
      theme_bw() +
      labs(x = x,
           y = "Empirical log mean count") +
      theme(text = element_text(size=25))
  } else {
    log_table %>%
      ggplot(aes(x = mean_x,
                 y = log_mean,
                 color = group,
                 shape = group)) +
      geom_point(size=2.5) +
      geom_smooth(se=F, method="lm", formula = reg_formula) +
      theme_bw() +
      labs(x = x,
           y = "Empirical log mean count",
           color = grouping,
           shape = grouping) +
      theme(text = element_text(size=25))
  }

}
```

Abstract

Section 1: Introduction

Section 2: Data

```
serengeti <- serengeti %>%
  drop_na()
glimpse(serengeti)
```

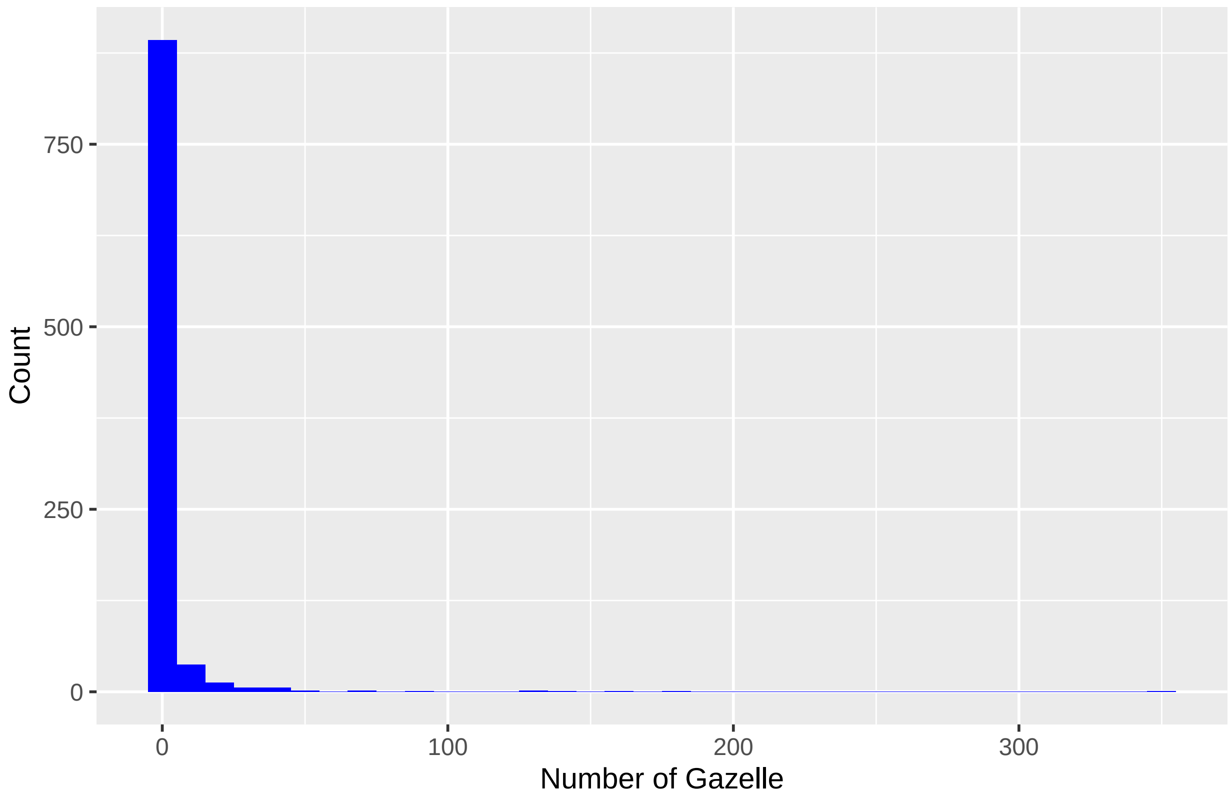
```
## Rows: 966
## Columns: 15
## $ siteID                  <chr> "B06", "B06", "B06", "B06", "B06", "B06", "B06~
## $ date                    <chr> "2012-01-01", "2012-01-09", "2012-01-17", "201~
## $ site.date               <chr> "B062012-01-01", "B062012-01-09", "B062012-01-~
## $ ndvi                    <dbl> 0.7553498, 0.6792881, 0.5774416, 0.5225141, 0.~
## $ gazelleThomsons.count   <int> 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 5~
## $ gazelleThomsons.present <int> 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1~
## $ topi.count              <int> 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0~
## $ topi.present            <int> 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0~
## $ zebra.count             <int> 0, 0, 0, 0, 16, 47, 118, 26, 6, 2, 1, 0, 1, 1,~
## $ zebra.present           <int> 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 0~
## $ fire                    <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ amRivDist               <dbl> 4135.915, 4135.915, 4135.915, 4135.915, 4135.9~
## $ TM100                   <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ LriskDry                <dbl> 0.001216255, 0.001216255, 0.001216255, 0.00121~
## $ T50                     <int> 500, 500, 500, 500, 500, 500, 500, 500, 500, 5~
```

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, gazalle's count and appearance, topi count, zebra count, and other environmental factors like whether having wildfire or having enough vegetation.

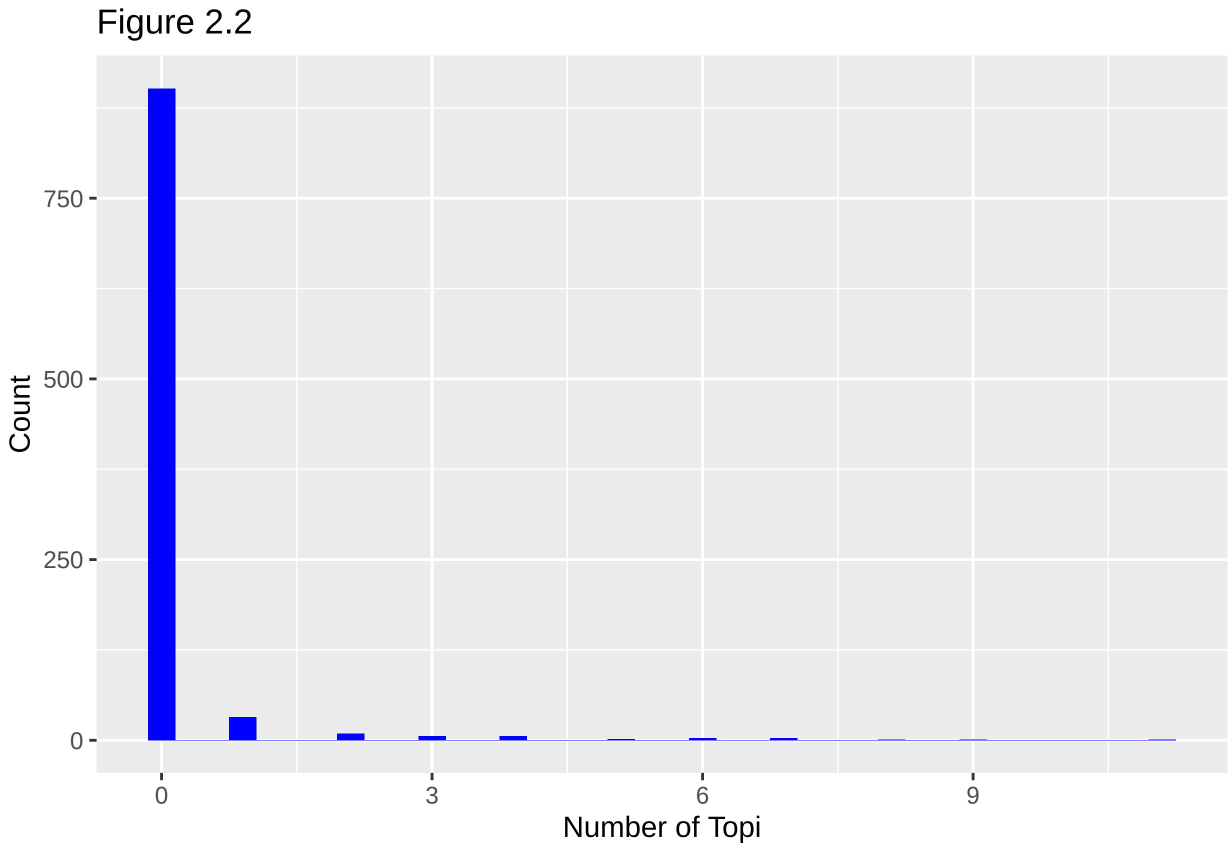
```
ggplot(data = serengeti, aes(x = gazelleThomsons.count)) + geom_histogram(binwidth = 10, fill = "blue")
```

Figure 2.1



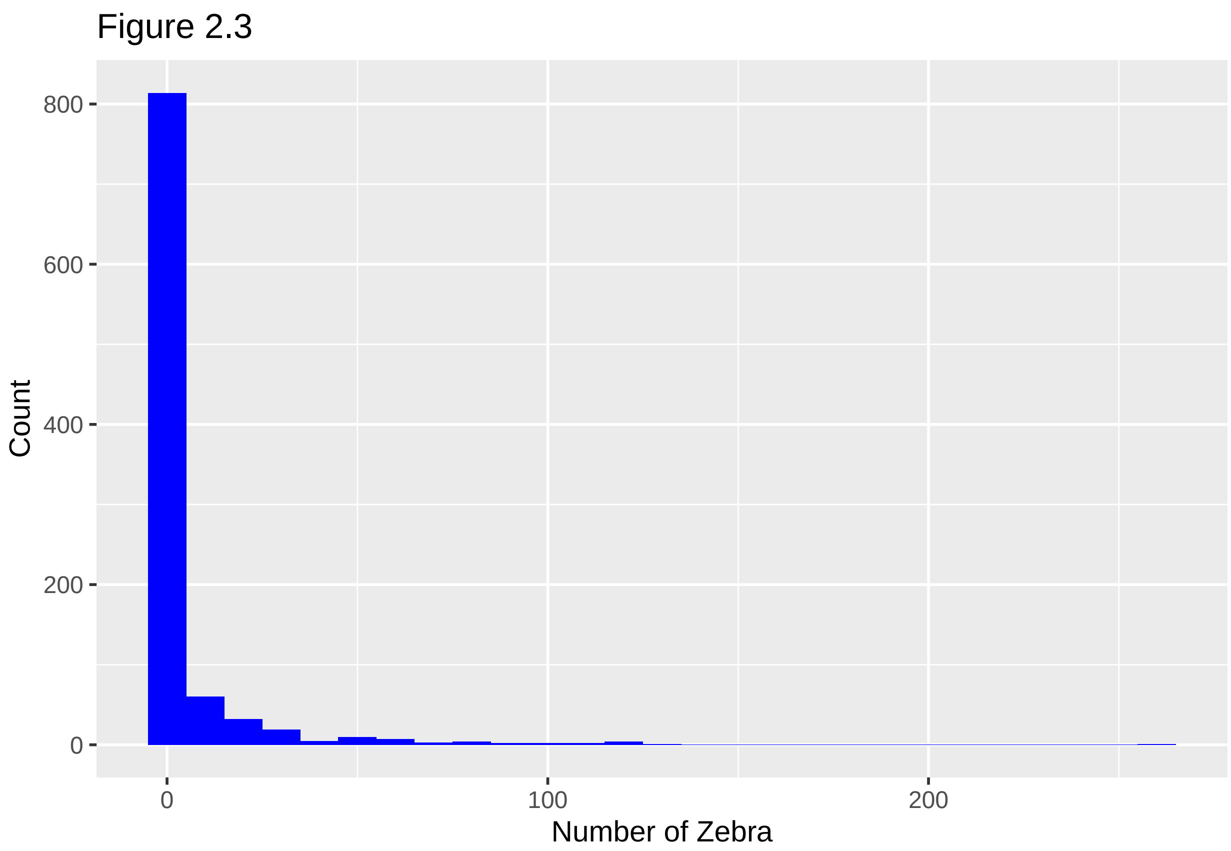
Based on Figure 2.1, we can observe that the number of gazelle Thomsons in these sites ar 0 in most sites. There are also hundreds of gazelle Thomsons in some specific areas.

```
ggplot(data = serengeti, aes(x = topi.count)) + geom_histogram(binwidth = 0.3, fill = "blue") + labs(ti
```



Based on Figure 2.2, it shows that in most recorded sites, the number of topi is 0 and there are few of them in some specific sites.

```
ggplot(data = serengeti, aes(x = zebra.count)) + geom_histogram(binwidth = 10, fill = "blue") + labs(ti
```

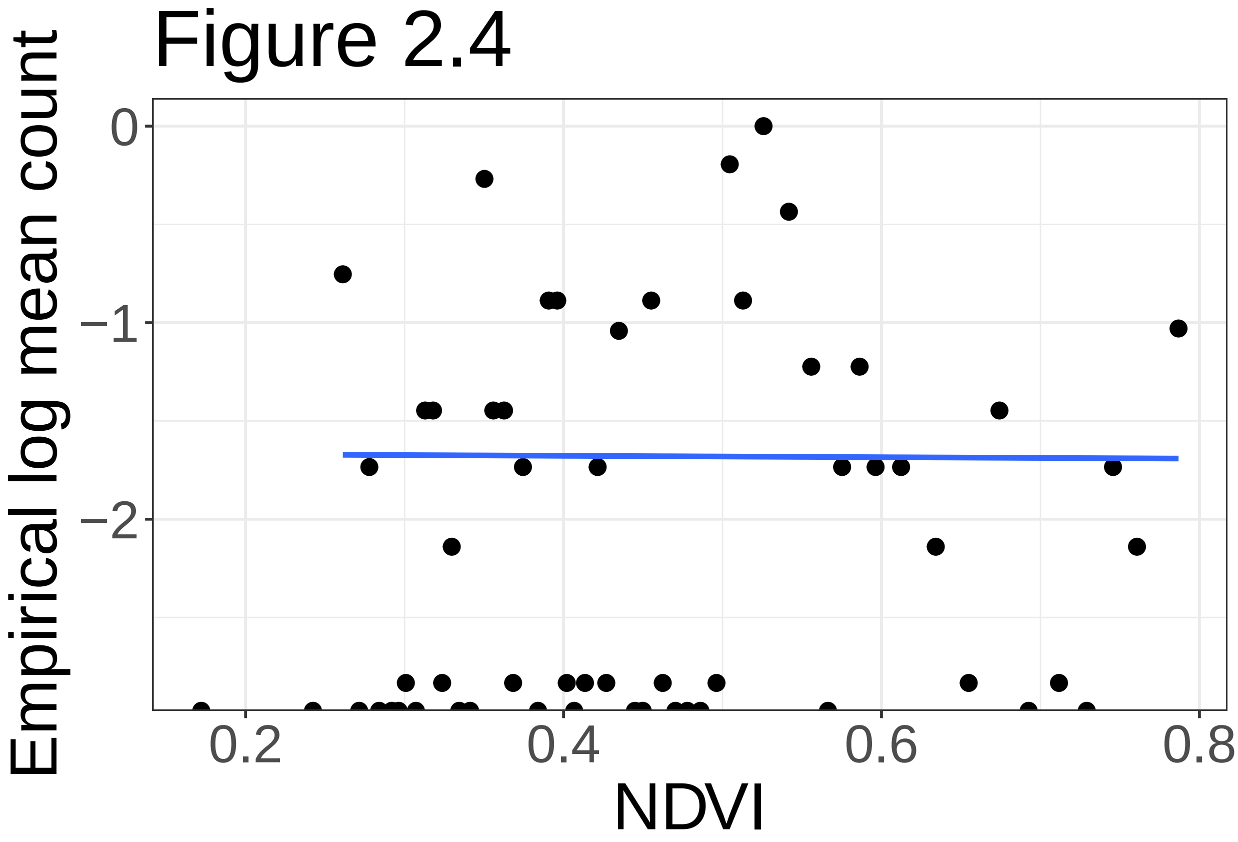


Based on Figure 2.3, we can also conclude that the number of zebra in most sites is 0.

```
logmean_plot(data = serengeti, 60, "equal_size","ndvi", "topi.count", reg_formula = y ~ x) + labs(title
```

```
## ‘summarise()‘ has grouped output by ’bin’. You can override using the ‘.groups‘
## argument.
```

```
## Warning: Removed 19 rows containing non-finite values (stat_smooth).
```

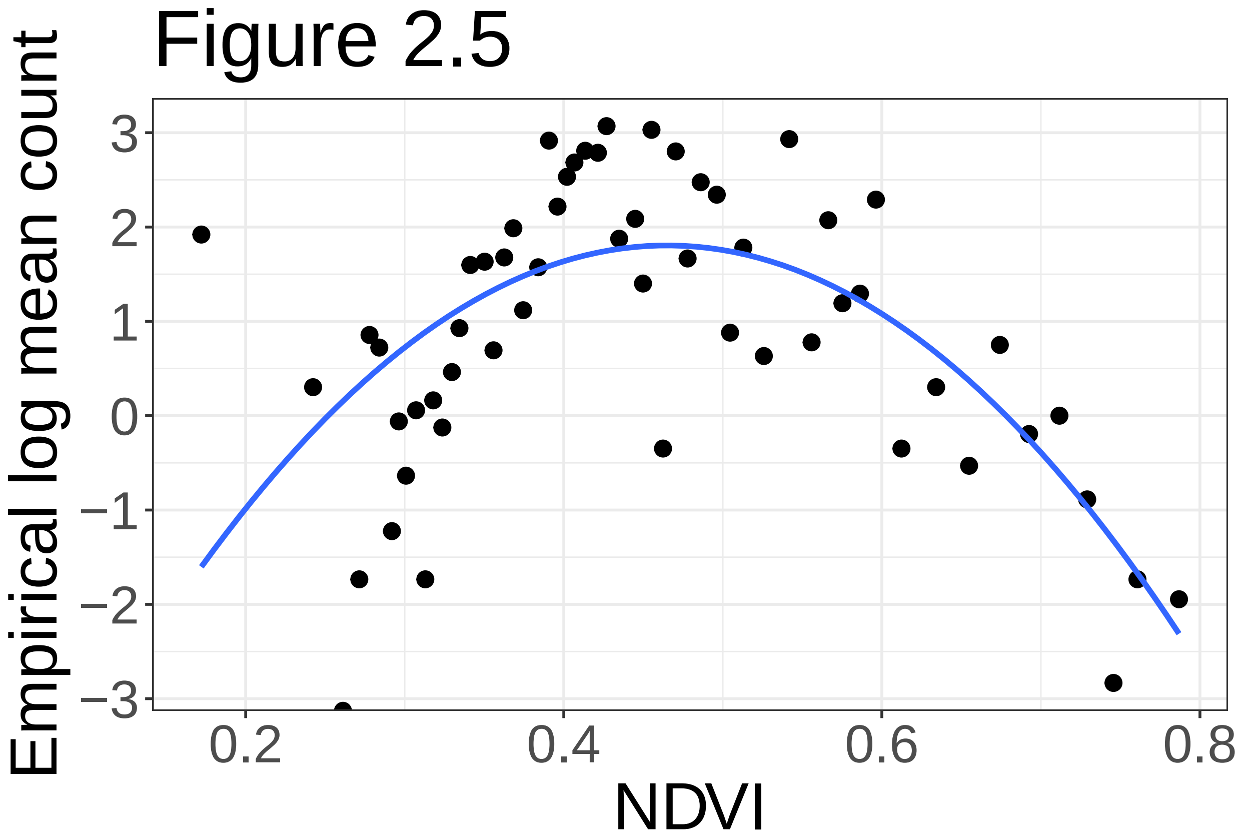


As showed from Figure 2.4, the relationship between the number of topi and the vegetation condition is also linear since points are bouncing around the linear which means no transformation needed.

```
logmean_plot(data = serengeti, 60, "equal_size","ndvi", "zebra.count", reg_formula = y ~ poly(x,2)) + la
```

```
## ‘summarise()‘ has grouped output by ’bin’. You can override using the ‘.groups‘
## argument.
```

```
## Warning: Removed 1 rows containing non-finite values (stat_smooth).
```

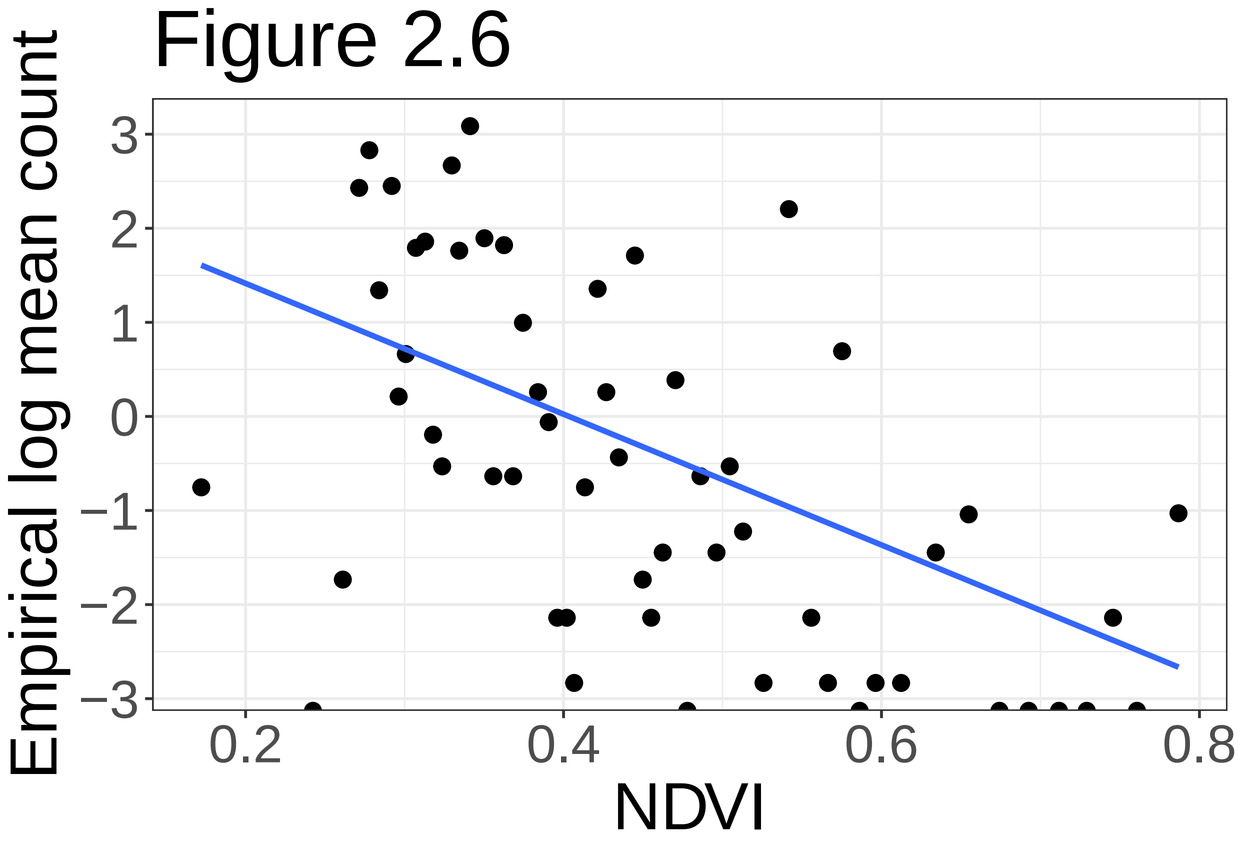


Based on Figure 2.5, the relationship between the number of zebra and the vegetation condition is nonlinear. So the polynomial transformation with a power of two is needed.

```
logmean_plot(data = serengeti, 60, "equal_size","ndvi", "gazelleThomsons.count", reg_formula = y ~ x) +
```

```
## ‘summarise()‘ has grouped output by ’bin’. You can override using the ‘.groups‘
## argument.
```

```
## Warning: Removed 8 rows containing non-finite values (stat_smooth).
```

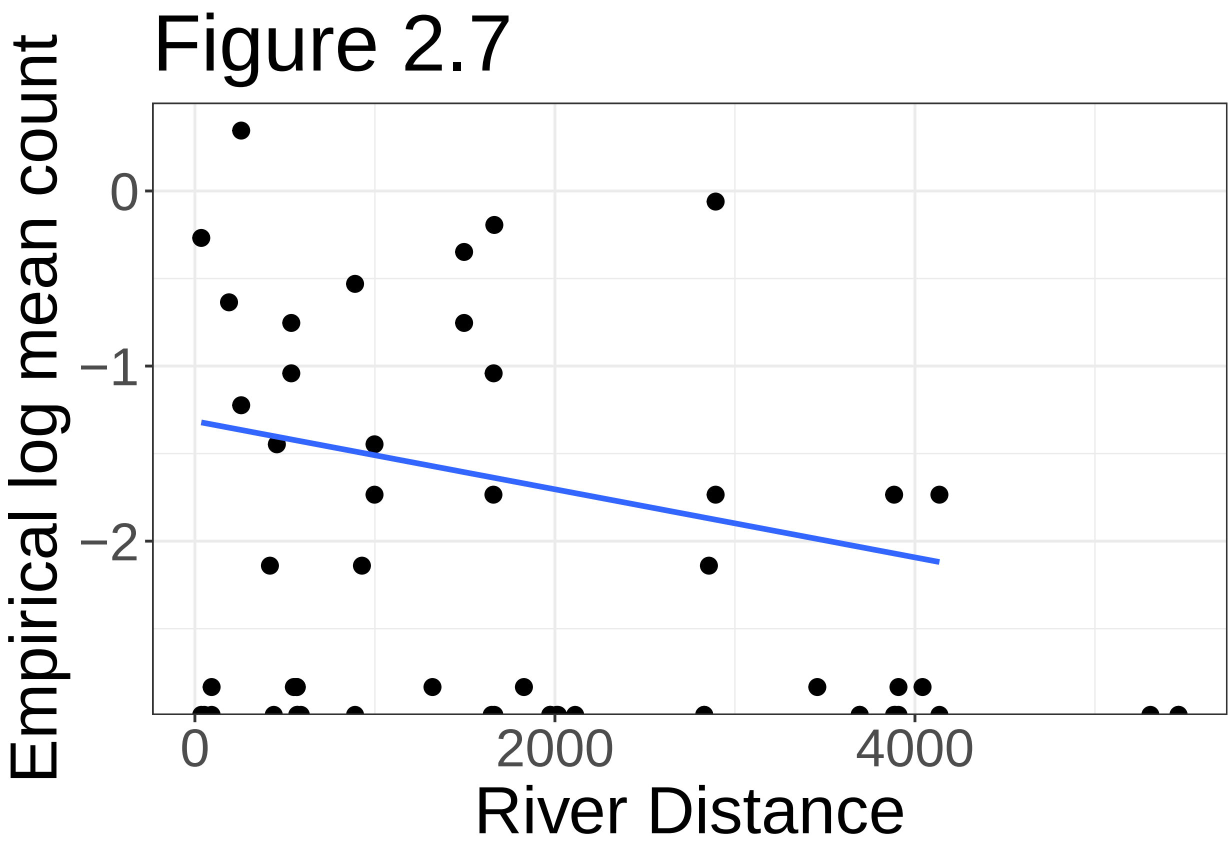


Based on Figure 2.6, the relationship between vegetation condition and the number of gazelle can be defined as linear since there is no explicit shape. No transformation is required.

```
logmean_plot(data = serengeti, 60, "equal_size",x = "amRivDist",  y = "topi.count", reg_formula = y ~ (
```

```
## `summarise()` has grouped output by 'bin'. You can override using the `.groups`
## argument.
```

```
## Warning: Removed 26 rows containing non-finite values (stat_smooth).
```

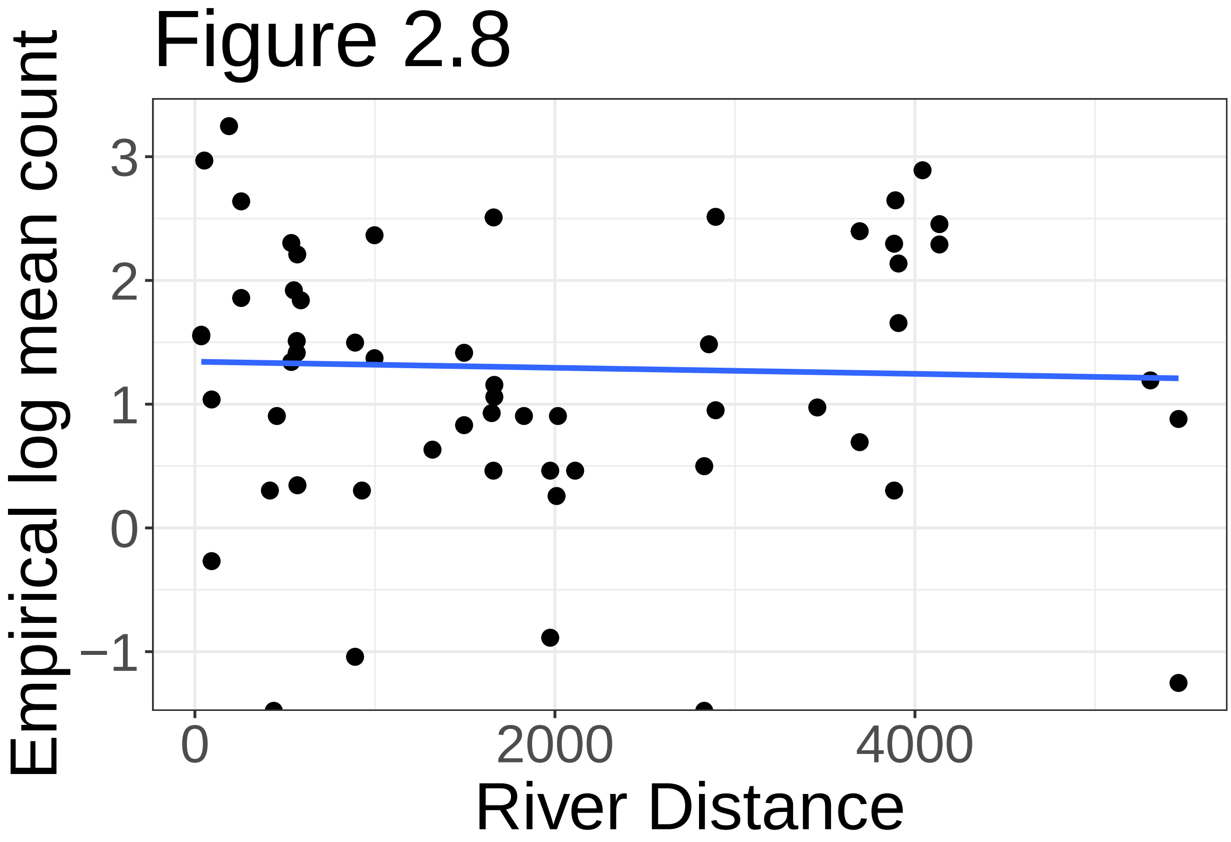


Based on Figure 2.7, ideally no transformation is needed for the relation between river distance and the number of topi. However, since the numbers for river distance are really large, outweighted other variablers, I choose to do the log transformation on it.

```
logmean_plot(data = serengeti, 60, "equal_size",x = "amRivDist",  y = "zebra.count", reg_formula = y ~
```

```
## ‘summarise()‘ has grouped output by ’bin’. You can override using the ‘.groups‘
## argument.
```

```
## Warning: Removed 2 rows containing non-finite values (stat_smooth).
```

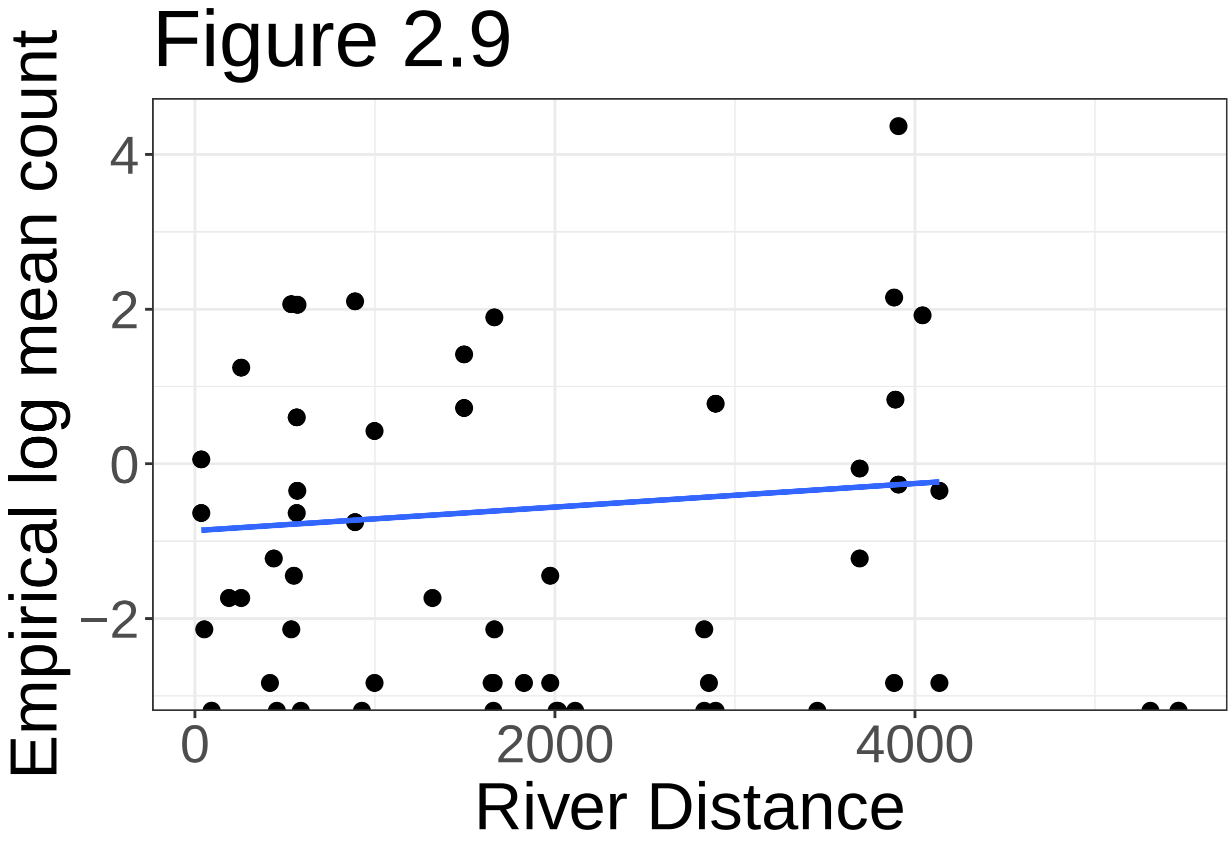


As Figure 2.8 shows, it is also reasonable to have a log transformation with the same reason for the relation between river distance and the number of topi.

```
logmean_plot(data = serengeti, 60, "equal_size", x = "amRivDist", y = "gazelleThomsons.count", reg_formu
```

```
## ‘summarise()‘ has grouped output by ’bin’. You can override using the ‘.groups‘
## argument.
```

```
## Warning: Removed 15 rows containing non-finite values (stat_smooth).
```

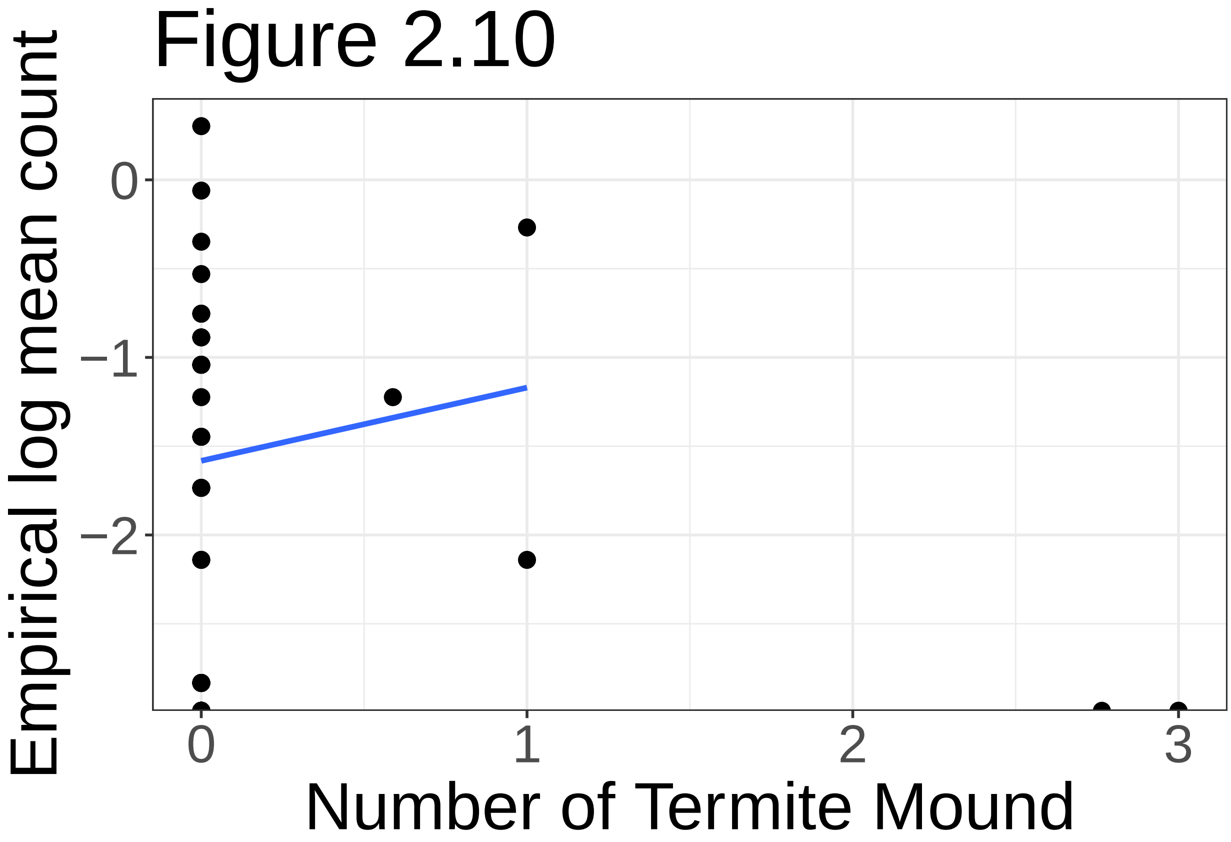


Based on Figure 2.9, having a log transformation is reasonable for the relation between the number of gazelle and the river distance.

```
logmean_plot(data = serengeti, 60, "equal_size", x = "TM100", y = "topi.count", reg_formula = y ~ x) + l
```

```
## `summarise()` has grouped output by 'bin'. You can override using the `.groups`
## argument.
```

```
## Warning: Removed 26 rows containing non-finite values (stat_smooth).
```

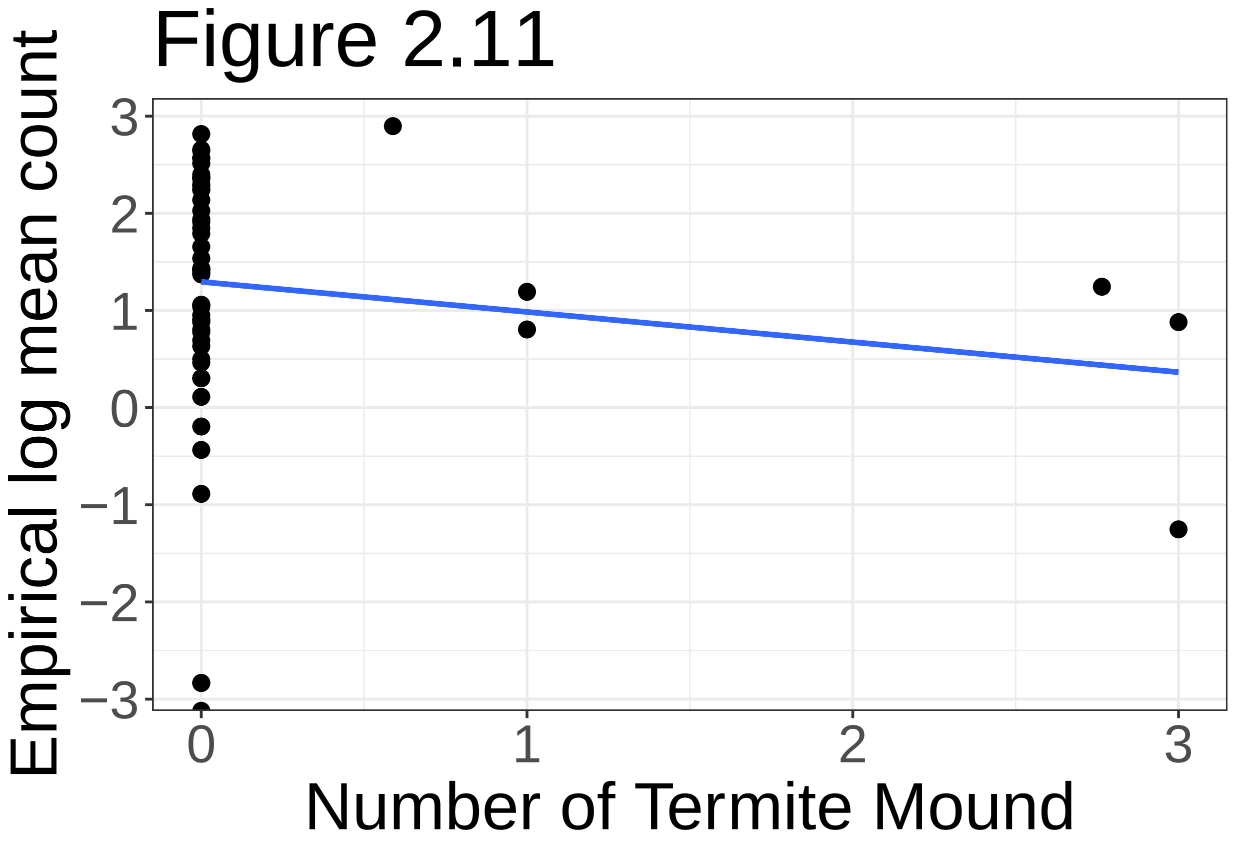


Based on Figure 2.10, no transformation is needed since it is hard to describe the shape.

```
logmean_plot(data = serengeti, 60, "equal_size", x = "TM100", y = "zebra.count", reg_formula = y ~ x) +
```

```
## `summarise()` has grouped output by 'bin'. You can override using the `.groups`
## argument.
```

```
## Warning: Removed 1 rows containing non-finite values (stat_smooth).
```

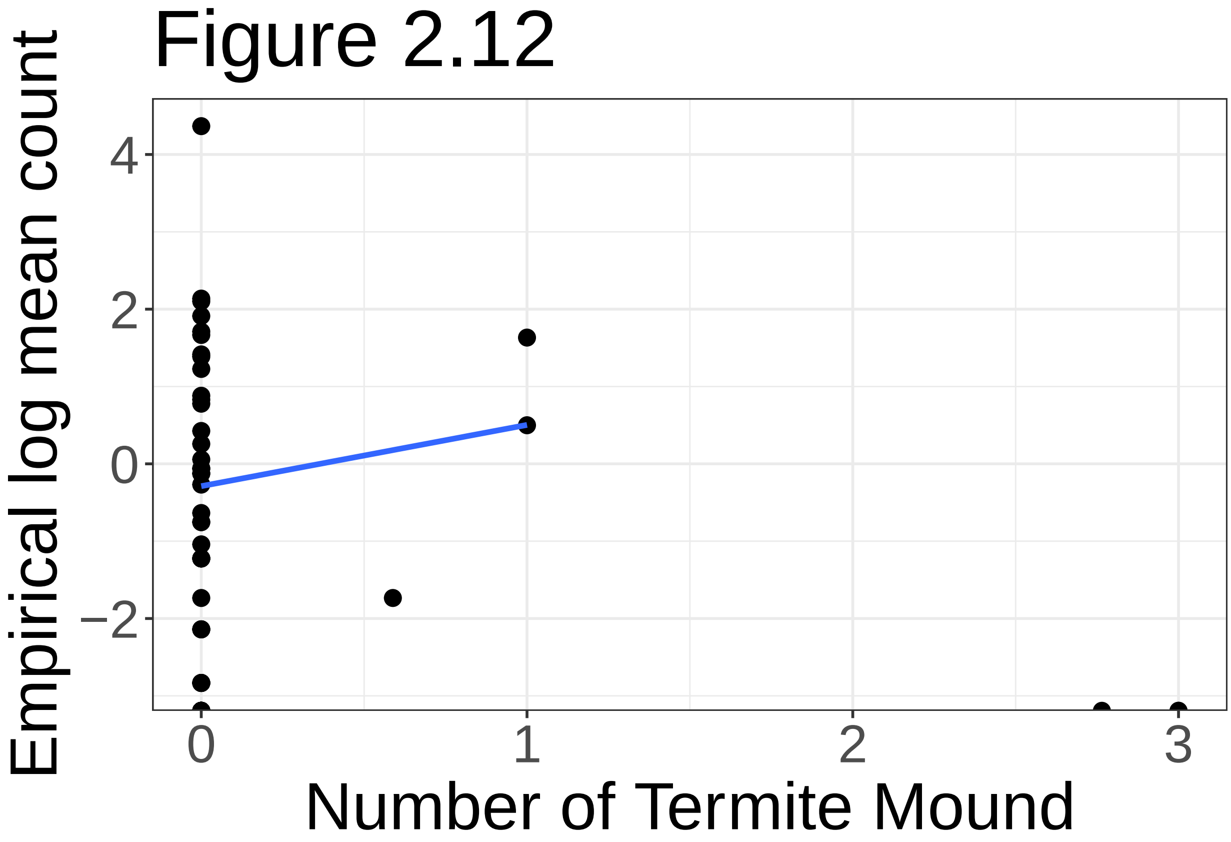


Based on Figure 2.11, there is also no transformation needed for the relationship between the number of zebra and the number of termite mounds since most of the points is about around the line.

```
logmean_plot(data = serengeti, 60, "equal_size", x = "TM100", y = "gazelleThomsons.count", reg_formula
```

```
## ‘summarise()‘ has grouped output by ’bin’. You can override using the ‘.groups‘
## argument.
```

```
## Warning: Removed 17 rows containing non-finite values (stat_smooth).
```

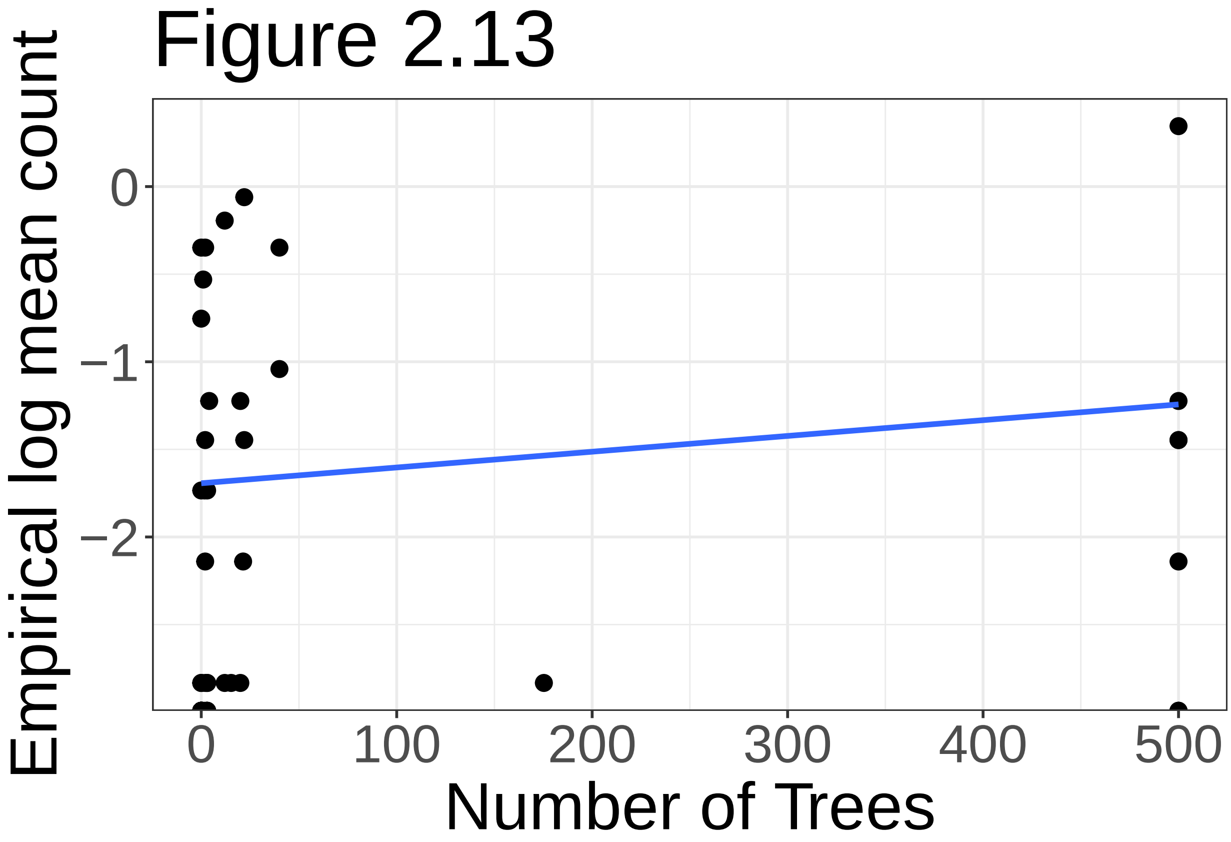


Based on the Figure 2.12, the relation between number of termite and the number of gazelle does not need to do any transformation.

```
logmean_plot(data = serengeti, 60, "equal_size", x = "T50", y = "topi.count", reg_formula = y ~ x) + lal
```

```
## `summarise()` has grouped output by 'bin'. You can override using the `.groups`
## argument.
```

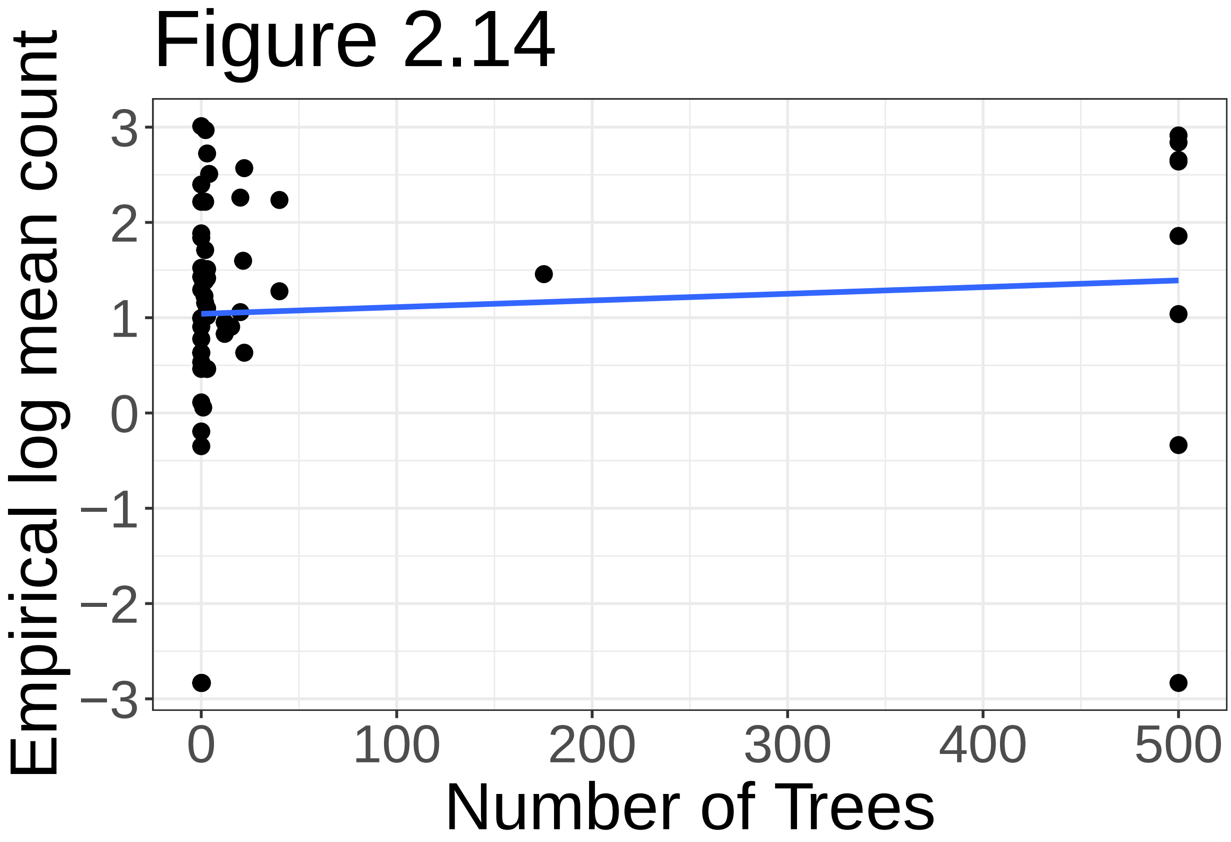
```
## Warning: Removed 26 rows containing non-finite values (stat_smooth).
```



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

```
logmean_plot(data = serengeti, 60, "equal_size", x = "T50", y = "zebra.count", reg_formula = y ~ x)  +
```

```
## ‘summarise()‘ has grouped output by ’bin’. You can override using the ‘.groups‘
## argument.
```

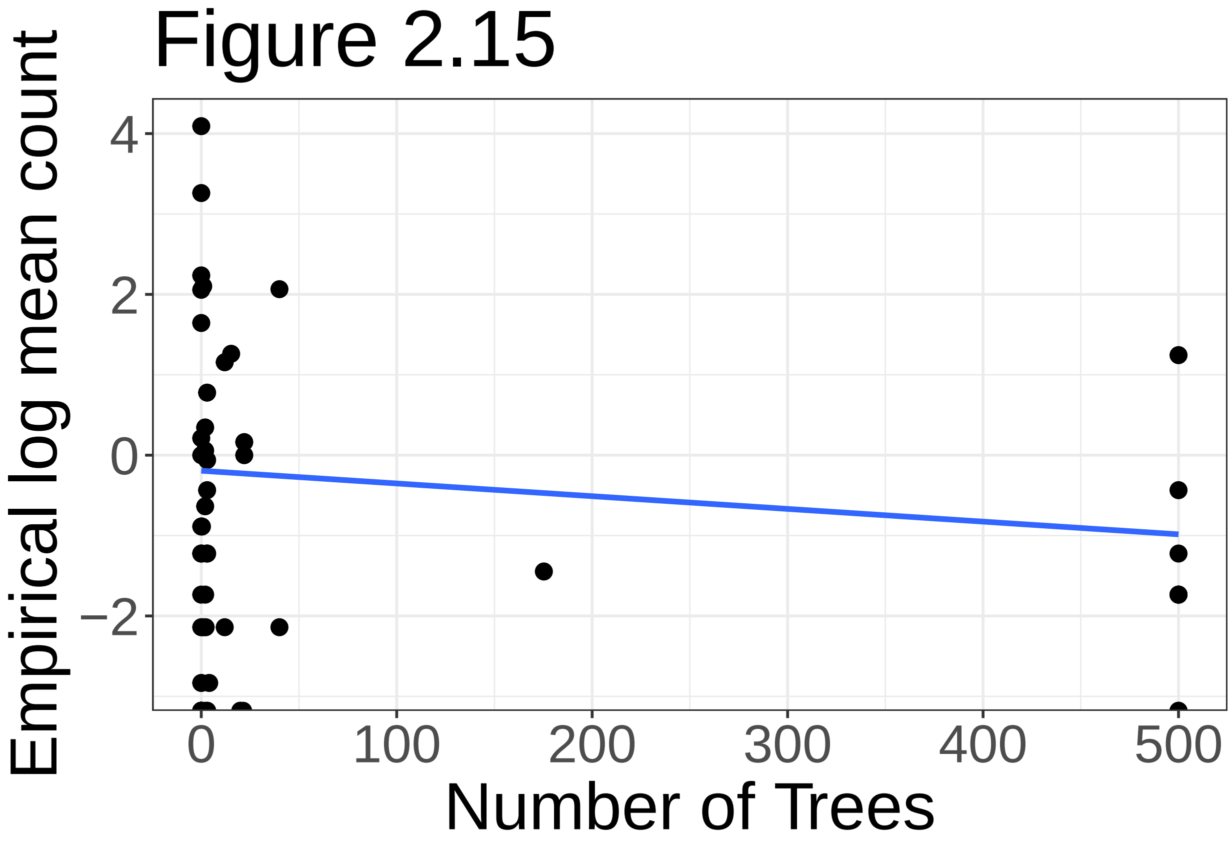


Based on Figure 2.14, no transformation is required for the number of trees and number of zebra.

```
logmean_plot(data = serengeti, 60, "equal_size", x = "T50", y = "gazelleThomsons.count", reg_formula =
```

```
## ‘summarise()‘ has grouped output by ’bin’. You can override using the ‘.groups‘
## argument.
```

```
## Warning: Removed 16 rows containing non-finite values (stat_smooth).
```

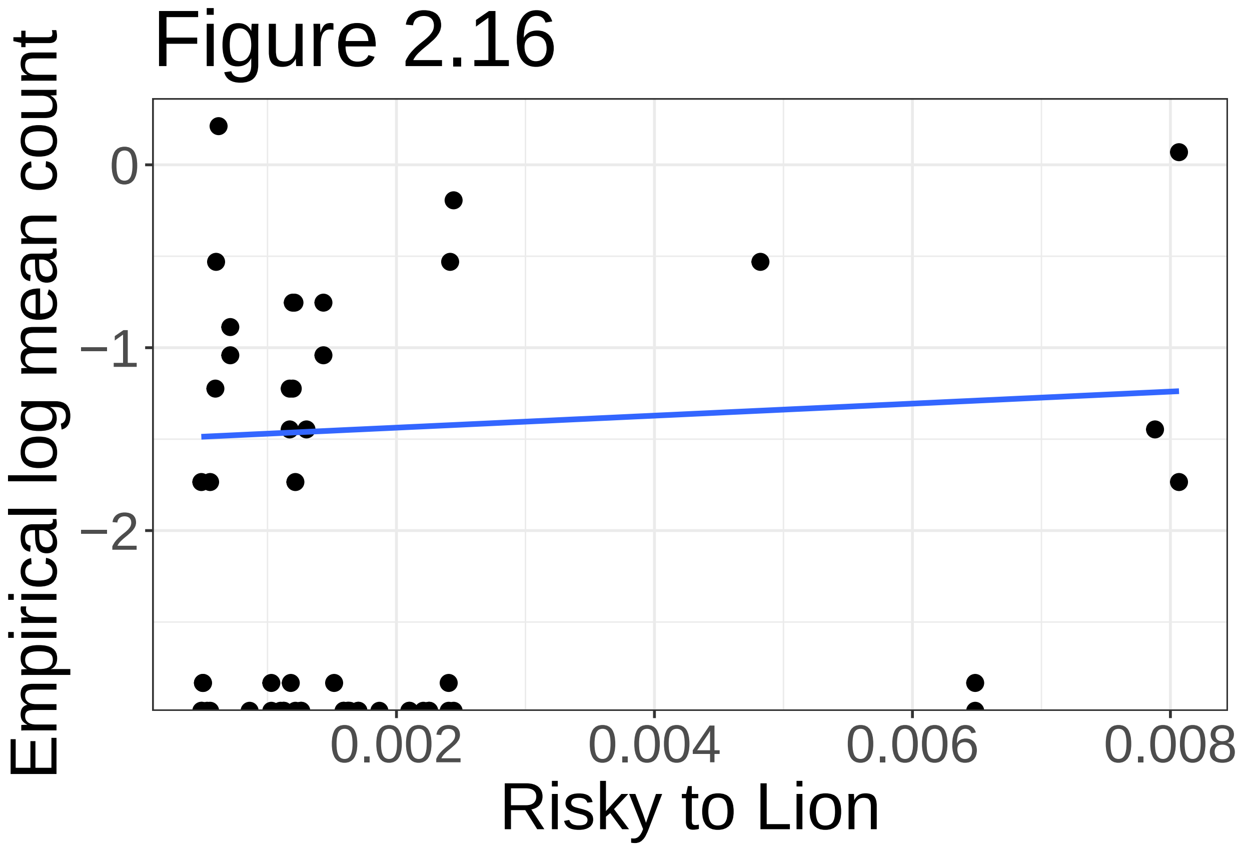


As Figure 2.15 shows, only doing linear transformation for the relationship between number of gazelle and number treesis fine.

```
logmean_plot(data = serengeti, 60, "equal_size", x = "LriskDry", y = "topi.count", reg_formula = y ~ x)
```

```
## ‘summarise()‘ has grouped output by ’bin’. You can override using the ‘.groups‘
## argument.
```

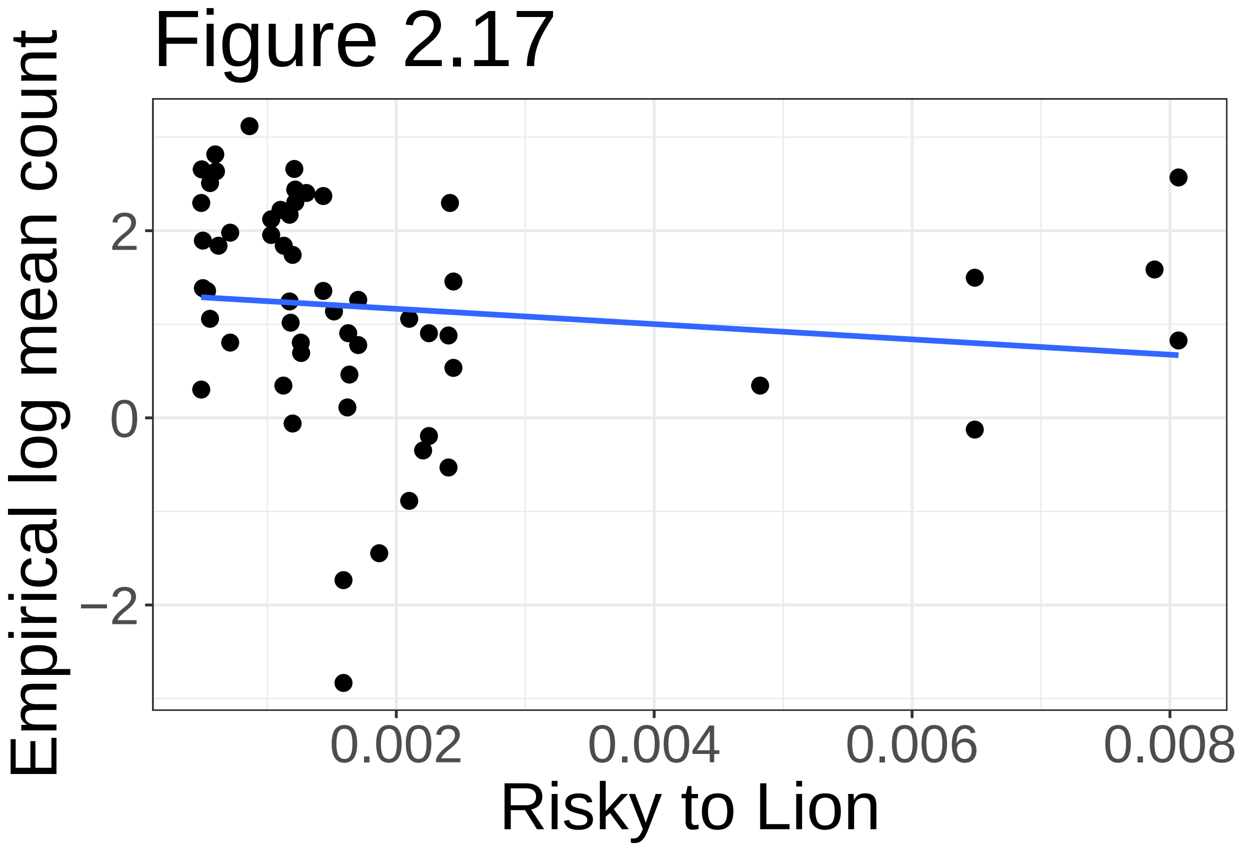
```
## Warning: Removed 28 rows containing non-finite values (stat_smooth).
```



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

```
logmean_plot(data = serengeti, 60, "equal_size", x = "LriskDry", y = "zebra.count", reg_formula = y ~ (
```

```
## ‘summarise()‘ has grouped output by ’bin’. You can override using the ‘.groups‘
## argument.
```

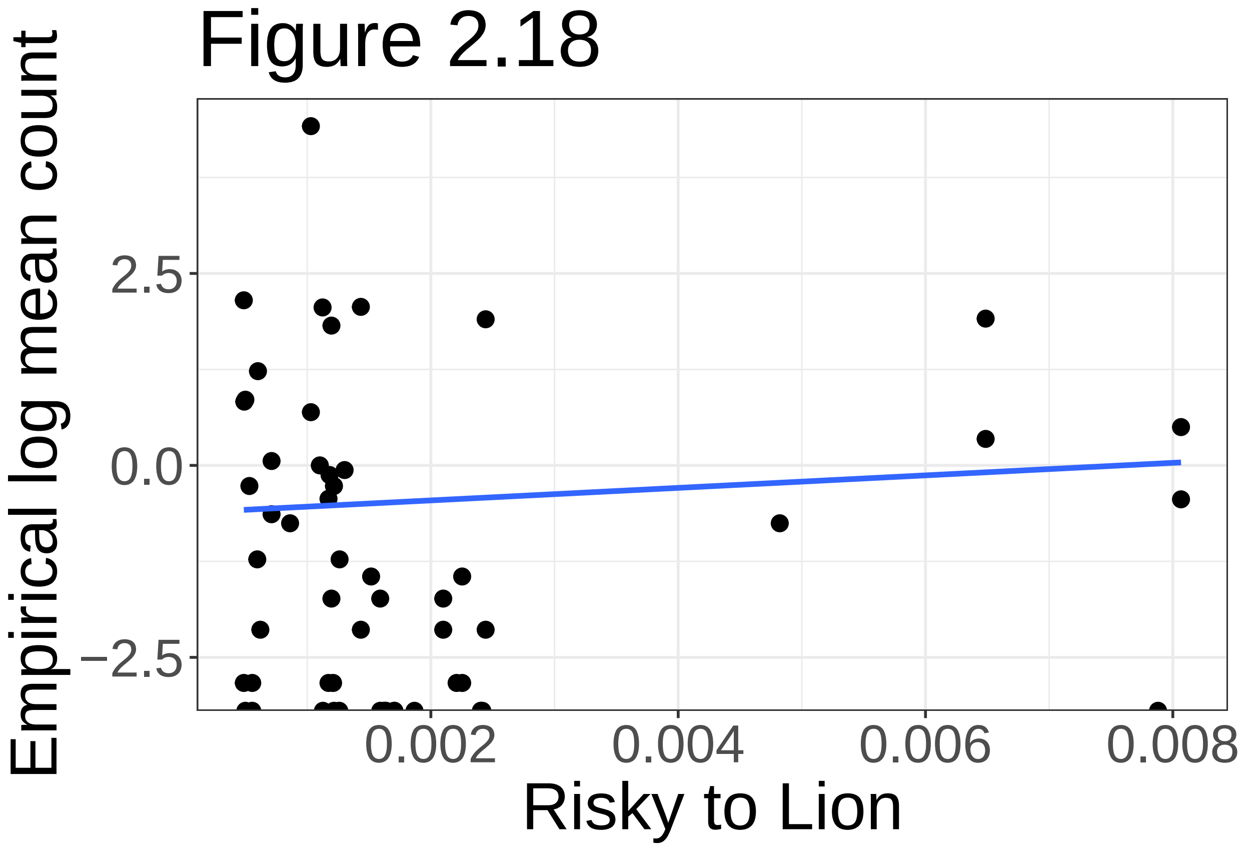


Based on Figure 2.17, no transformation is needed for the relationship between the number of zebra and the risk to lion in certain site.

```
logmean_plot(data = serengeti, 60, "equal_size", x = "LriskDry", y = "gazelleThomsons.count", reg_formul
```

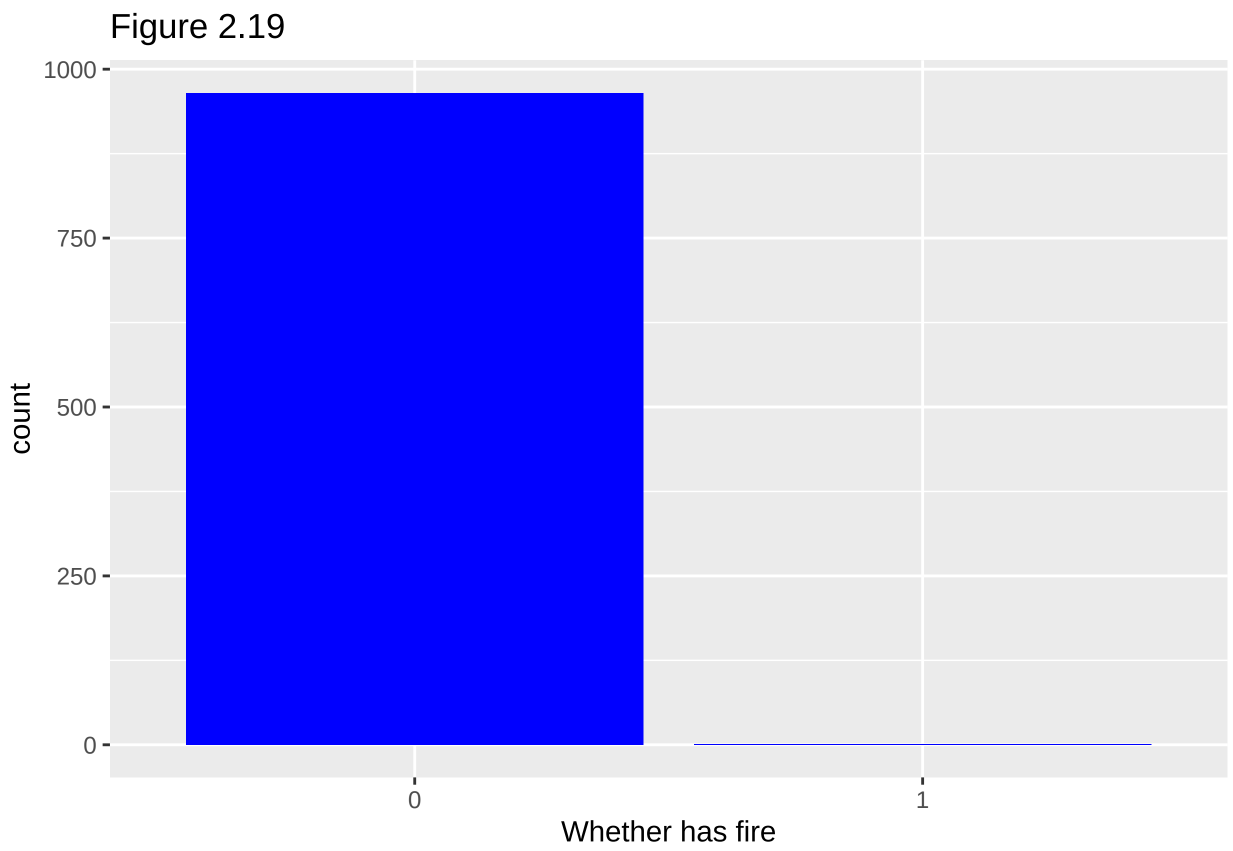
```
## ‘summarise()‘ has grouped output by ’bin’. You can override using the ‘.groups‘
## argument.
```

```
## Warning: Removed 16 rows containing non-finite values (stat_smooth).
```



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

```
serengeti <- serengeti %>%
  mutate(fire = as.factor(fire))
ggplot(data = serengeti, aes(x = fire)) + geom_bar(fill = "blue") + labs(title = "Figure 2.19", x = "Whe
```



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 this variable will not be chosen to include in the models since there is rarely relation between the number of species and whether has a wild fire.

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 three different species, 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 will be:

$$Y_i \sim Poisson(\lambda_i)$$

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

For specie topi, the summary for fitted regression model will be:

```
zip_m1 <- zeroinfl(topi.count ~ ndvi + log(amRivDist) + TM100 + LriskDry + T50 | ndvi  + log(amRivDist)
summary(zip_m1)
```

```
##
## Call:
## zeroinfl(formula = topi.count ~ ndvi + log(amRivDist) + TM100 + LriskDry +
##     T50 | ndvi + log(amRivDist) + TM100 + LriskDry + T50, data = serengeti)
##
## Pearson residuals:
##     Min      1Q  Median      3Q     Max
## -0.3873 -0.2630 -0.2072 -0.1744 14.8260
##
## Count model coefficients (poisson with log link):
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)     1.7787610  0.5482199   3.245  0.00118 **
## ndvi           -0.2410404  0.7838326  -0.308  0.75845
## log(amRivDist) -0.1518015  0.0924268  -1.642  0.10051
## TM100          -0.2039337  0.4361713  -0.468  0.64010
## LriskDry       62.5114507 43.3208101   1.443  0.14902
## T50             0.0006309  0.0004607   1.369  0.17090
##
## Zero-inflation model coefficients (binomial with logit link):
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)     1.889e+00  7.686e-01   2.457   0.0140 *
## ndvi           -1.252e+00  9.255e-01  -1.353   0.1761
## log(amRivDist)  2.414e-01  1.069e-01   2.258   0.0239 *
## TM100          -2.273e-01  3.864e-01  -0.588   0.5564
## LriskDry       -1.543e+02  6.399e+01  -2.411   0.0159 *
## T50            -1.106e-03  7.216e-04  -1.533   0.1252
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Number of iterations in BFGS optimization: 51
## Log-likelihood: -355.8 on 12 Df
```

Based on the summary, the fitted regression line for topi will be:

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

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

The summary for the fitted regression line for zebra will be:

```
zip_m2 <- zeroinfl(zebra.count ~ poly(ndvi,2) + log(amRivDist) + TM100 + LriskDry + T50 | poly(ndvi,2)
summary(zip_m2)
```

```
##
## Call:
## zeroinfl(formula = zebra.count ~ poly(ndvi, 2) + log(amRivDist) + TM100 +
##     LriskDry + T50 | poly(ndvi, 2) + log(amRivDist) + TM100 + LriskDry +
##     T50, data = serengeti)
##
```

```
## Pearson residuals:
##     Min      1Q  Median      3Q     Max 
## -1.2351 -0.7478 -0.5553 -0.2738 55.6526 
## 
## Count model coefficients (poisson with log link):
##                  Estimate Std. Error z value Pr(>|z|)    
## (Intercept)     2.426e+00  7.437e-02  32.623  < 2e-16 ***
## poly(ndvi, 2)1 -8.638e+00  8.271e-01 -10.445  < 2e-16 ***
## poly(ndvi, 2)2 -2.132e+01  9.741e-01 -21.889  < 2e-16 ***
## log(amRivDist)  5.177e-03  1.044e-02   0.496     0.62    
## TM100          -1.599e-01  3.769e-02  -4.243 2.21e-05 ***
## LriskDry       -3.782e+01  8.615e+00  -4.390 1.13e-05 ***
## T50             4.540e-04  6.717e-05   6.759 1.39e-11 ***
## 
## Zero-inflation model coefficients (binomial with logit link):
##                  Estimate Std. Error z value Pr(>|z|)    
## (Intercept)     0.1270296  0.3951128   0.322   0.7478    
## poly(ndvi, 2)1  2.4924767  2.6355907   0.946   0.3443    
## poly(ndvi, 2)2 19.3404189  2.9404352   6.577 4.79e-11 ***
## log(amRivDist)  0.0663478  0.0568887   1.166   0.2435    
## TM100           0.2602937  0.1293838   2.012   0.0442 *  
## LriskDry       83.7532092 40.7796128   2.054   0.0400 *  
## T50            -0.0009661  0.0004083  -2.367   0.0180 *  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 
## 
## Number of iterations in BFGS optimization: 27 
## Log-likelihood: -5376 on 14 Df
```

For specie zebra, the fitted regression line will be:

$$\log\left(\frac{\alpha_i}{1-\alpha_i}\right) = -0.2919 + 0.467ndvi_i + 0.0946log(amRivDist_i) + 0.216TM100_i + 65.8486riskDry_i - 0.0011T50_i$$

$$\log(\lambda_i) = 3.157 - 0.5754ndvi_i - 0.0215log(amRivDist_i) - 0.1495TM100_i - 23.7LriskDry_i + 0.0006T50_i$$

The summary fitted regresion model for specie gazelle will be:

```
zip_m3 <- zeroinfl(gazelleThomsons.count ~ ndvi + log(amRivDist) + TM100 + LriskDry + T50  | ndvi + log
summary(zip_m3)
```

```
## 
## Call:
## zeroinfl(formula = gazelleThomsons.count ~ ndvi + log(amRivDist) + TM100 + 
##     LriskDry + T50 | ndvi + log(amRivDist) + TM100 + LriskDry + T50, 
##     data = serengeti)
## 
## Pearson residuals:
##     Min      1Q  Median      3Q     Max 
## -1.3405 -0.4588 -0.3266 -0.2035 29.1032 
## 
## Count model coefficients (poisson with log link):
##                  Estimate Std. Error z value Pr(>|z|)    
## (Intercept)    -4.223e-02  2.138e-01  -0.198    0.843    
## ndvi           -2.816e+00  2.269e-01 -12.411  < 2e-16 ***
## log(amRivDist)  5.595e-01  2.587e-02  21.626  < 2e-16 ***
## TM100          -3.943e-01  9.618e-02  -4.100 4.13e-05 ***
## LriskDry       -1.023e+02  1.191e+01  -8.589  < 2e-16 ***
## T50            -2.613e-03  2.306e-04 -11.328  < 2e-16 ***
##
## Zero-inflation model coefficients (binomial with logit link):
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)     5.805e-01  6.359e-01   0.913  0.36128
## ndvi            5.235e+00  7.768e-01   6.740 1.59e-11 ***
## log(amRivDist) -1.815e-01  8.136e-02  -2.231  0.02567 *
## TM100           6.865e-01  2.391e-01   2.872  0.00408 **
## LriskDry        4.370e+01  5.158e+01   0.847  0.39686
## T50             3.001e-05  5.528e-04   0.054  0.95671
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Number of iterations in BFGS optimization: 26
## Log-likelihood: -2964 on 12 Df
```

The fitted regression line for gazelle will be:

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

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

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:

```
zip_m1$loglik
```

```
## [1] -355.7883
```

The log likelihood for the reduced model will be:

```
zip_m1_re <- zeroinfl(topi.count ~ log(amRivDist) + LriskDry   | ndvi + log(amRivDist) + TM100 + LriskD
zip_m1_re$loglik
```

```
## [1] -356.9112
```

The test statistic G will be $2(-355.7883-(-356.9112)) = 2.2458 $

Using χ^2 distribution, the p-value will be

```r
pchisq(2.2458, df = 3, lower.tail = FALSE)
```

```
## [1] 0.5229836
```

Since the p-value is very close to 0, we can conclude that we there is a strong relationship between the topi count and the vegetation condition (ndvi), number of trees (T50), and the number of termite mound (TM100).

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:

```r
zip_m1$loglik
```

```
## [1] -355.7883
```

The log likelihood for the reduced model will be:

```r
zip_m1_re2 <- zeroinfl(topi.count ~ ndvi + TM100 + LriskDry + T50  | ndvi + log(amRivDist) + TM100 + Lr
zip_m1_re2$loglik
```

```
## [1] -357.0991
```

The test statistic G will be $2(-355.7883-(-357.0991)) = 2.2458 $

Using χ^2 distribution, the p-value will be

```r
pchisq(2.6216, df = 1, lower.tail = FALSE)
```

```
## [1] 0.1054181
```

Since p-value is relative small, we can conclude that there is a moderate relationship between the number of topi and the log river distance over each site.

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:

```
zip_m1$loglik
```

```
## [1] -355.7883
```

The log likelihood for the reduced model will be:

```
zip_m1_re3 <- zeroinfl(topi.count ~ ndvi + log(amRivDist) + TM100  + T50  | ndvi + log(amRivDist) + TM1
zip_m1_re3$loglik
```

```
## [1] -356.8398
```

The test statistic G will be $2(-355.7883-(-356.8398)) 2.103 $

Using χ^2 distribution, the p-value will be

```
pchisq(2.103, df = 1, lower.tail = FALSE)
```

```
## [1] 0.1470104
```

Since p-value is relative small, we can conclude that there is a moderate relationship between the number of topi and the risky to lions with these sites.

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:

```
zip_m2$loglik
```

```
## [1] -5376.208
```

The log likelihood for the reduced model will be:

```
zip_m2_re <- zeroinfl(zebra.count ~ log(amRivDist) +LriskDry  | poly(ndvi,2) + log(amRivDist) + TM100 +
zip_m2_re$loglik
```

```
## [1] -5786.12
```

The test statistic G will be $2(-5376.208-(-5786.12)) = 819.824 $

Using χ^2 distribution, the p-value will be

```r
pchisq(819.824, df = 4, lower.tail = FALSE)
```

```
## [1] 3.901478e-176
```

Since this p-value is very close to zero, we can conclude that there is a strong 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.

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:

```r
zip_m2$loglik
```

```
## [1] -5376.208
```

The log likelihood for the reduced model will be:

```r
zip_m2_re2 <- zeroinfl(zebra.count~poly(ndvi,2) + TM100+ T50 +LriskDry  | poly(ndvi,2) + log(amRivDist)
zip_m2_re2$loglik
```

```
## [1] -5376.331
```

The test statistic G will be \$2(-5376.208-(-5376.331)) = 0.246 \$

Using χ^2 distribution, the p-value will be

```r
pchisq(0.246, df = 1, lower.tail = FALSE)
```

```
## [1] 0.6199058
```

Based on the result, the p-value is relatively large, we can conclude that there is a weak or even no relation between the number of zebra and the log river distance in this data set.

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:

```r
zip_m2$loglik
```

```
## [1] -5376.208
```

The log likelihood for the reduced model will be:

```
zip_m2_re3 <- zeroinfl(zebra.count~poly(ndvi,2) + TM100+ T50  + log(amRivDist)  | poly(ndvi,2) + log(aml
zip_m2_re3$loglik
```

```
## [1] -5386.361
```

The test statistic G will be $2(-5376.208-(-5386.361)) = 20.306 $

Using χ^2 distribution, the p-value will be

```
pchisq(20.306, df = 1, lower.tail = FALSE)
```

```
## [1] 6.599331e-06
```

From the result, the p-value is kind of close to zero. We can conclude that there is a strong relationship between the number of zebra in sites and the risky to lions.

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:

```
zip_m3$loglik
```

```
## [1] -2963.71
```

The log likelihood for the reduced model will be:

```
zip_m3_re <- zeroinfl(gazelleThomsons.count ~ log(amRivDist) + LriskDry   | ndvi + log(amRivDist) + TM1(
zip_m3_re$loglik
```

```
## [1] -3196.036
```

The test statistic G will be $2(-2963.71 -(-3196.036)) = 464.652 $

Using χ^2 distribution, the p-value will be

```
pchisq(464.652, df = 3, lower.tail = FALSE)
```

```
## [1] 2.180398e-100
```

Based on the result, the p-value is very close to zero. We can conclude that there is a strong relationship between the number of gazelle and the vegetation condition, number of trees and the termite mound.

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:

```
zip_m3$loglik
```

```
## [1] -2963.71
```

The log likelihood for the reduced model will be:

```
zip_m3_re2 <- zeroinfl(gazelleThomsons.count ~ ndvi + TM100 + T50 + LriskDry   | ndvi + log(amRivDist) +
zip_m3_re2$loglik
```

```
## [1] -3290.404
```

The test statistic G will be \$2(-2963.71 -(-3290.404)) = 653.388 \$

Using χ^2 distribution, the p-value will be

```
pchisq(653.388, df = 1, lower.tail = FALSE)
```

```
## [1] 4.095348e-144
```

Based on the result, the p-value is very close to zero. We can conclude that there is a strong relationship between the number of gazelle and the log of river distance.

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:

```
zip_m3$loglik
```

```
## [1] -2963.71
```

The log likelihood for the reduced model will be:

```
zip_m3_re3 <- zeroinfl(gazelleThomsons.count ~ ndvi + TM100 + T50 +log(amRivDist)   | ndvi + log(amRivDi
zip_m3_re3$loglik
```

```
## [1] -3008.449
```

The test statistic G will be \$2(-2963.71 -(-3008.449)) = 89.478 \$

Using χ^2 distribution, the p-value will be

```r
pchisq(89.478, df = 1, lower.tail = FALSE)
```

```
## [1] 3.100674e-21
```

Based on the result, the p-value is very close to zero. We can conclude that there is a strong relationship between the number of gazelle and the risky to lion.

Section 3.2: Comparing species

Section 4: Discussion