STAT455 - Mini-Investigation 3

Dereje Pollock

How does age affect male elephant mating patterns? An article by Poole(1989) investigated whether mating success in male elephants increases with age and whether there is a peak age for mating success. To address this question, the research team followed 41 elephants for one year and recorded both their ages and their number of matings. The data (Ramsey and Schafer) is found in `elephant.csv`, and the variables are: - `MATINGS` = the number of matings in a given year - `AGE` = the age of the elephant in years.

We want to model the number of matings, using age as the explanatory variable.

a) Which type of model is most appropriate for these data? Explain your reasoning.

We could use a Poisson regression model for elephant matings, but it might have trouble handling many zeros that could show up for elephants that don't mate in the year the study was conducted. If this is the case, we could use a Zero-Inflated Poisson model. This would deal with the zeros separately, and help us understand the likelihood of elephants not mating as well as the effects of age on mating frequency for elephants that do mate. We can check to see if the distribution of elephant Matings follow a poisson distribution, or if it would be better to use the ZIP model.

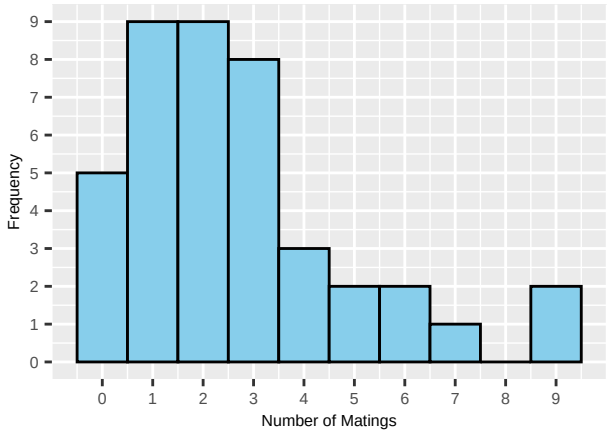


Figure 1: Distribution of matings counts.

From the histogram, we can see that it roughly follows the Poisson distribution, so for simplicity I will use the Poisson model.

b) Fit the model you chose, with a linear term for AGE. Display the model summary. Also display confidence intervals associated with the model parameters.

M1

Consider Y_i to be the number of matings for the i-th elephant, where $i = 1, 2, \ldots, n$, and n.

$$Y_i \sim \text{Poisson}(\lambda_i)$$

The relationship between the expected number of matings and the age of the elephant is log-linear, given by:

$$\log(\lambda_i) = \beta_0 + \beta_1 \text{AGE}_i$$

```
MP1 = glm(MATINGS ~ AGE, family = poisson, data = elephants)
```

```
# Call:
# glm(formula = MATINGS ~ AGE, family = poisson, data = elephants)
#
# Coefficients:
#             Estimate Std. Error z value Pr(>|z|)
# (Intercept) -1.58201    0.54462  -2.905  0.00368 **
# AGE          0.06869    0.01375   4.997 5.81e-07 ***
# ---
# Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
#
# (Dispersion parameter for poisson family taken to be 1)
#
#     Null deviance: 75.372  on 40  degrees of freedom
# Residual deviance: 51.012  on 39  degrees of freedom
# AIC: 156.46
#
# Number of Fisher Scoring iterations: 5
#
# (Intercept)         AGE
#   0.2055619   1.0711071
# Waiting for profiling to be done...
#                 2.5 %    97.5 %
# (Intercept) 0.0694813 0.5892357
# AGE         1.0425585 1.1003602
```

Exponentiated Coefficients for M1:

```
(Intercept)         AGE
  0.2055619   1.0711071
```

Confidence Intervals for M1

```
confint(MP1)
```

```
                 2.5 %      97.5 %
(Intercept) -2.66669764 -0.52892903
AGE          0.04167776  0.09563762
```

Exponentiated Confidence Intervals:

```
                2.5 %    97.5 %
(Intercept) 0.0694813 0.5892357
AGE         1.0425585 1.1003602
```

c) Write 1-2 sentences interpreting the estimate and p-value associated with the age coefficient. In your interpretations, write rounded numbers (i.e. say 3.22 instead of $e^{1.17}$).

The exponentiated coefficient for AGE is approximately 1.07, indicating that for each one-year increase in age, the expected number of matings increases by about 7%. This effect is statistically significant, as indicated by the very small p-value (5.81e-07), suggesting strong evidence against the null hypothesis that age has no effect on the number of matings.

d) For each age, calculate the mean number of matings. Take the log of each mean and plot it by AGE. Explain what the plot tells us about the appropriateness of the linearity assumption in Poisson regression.

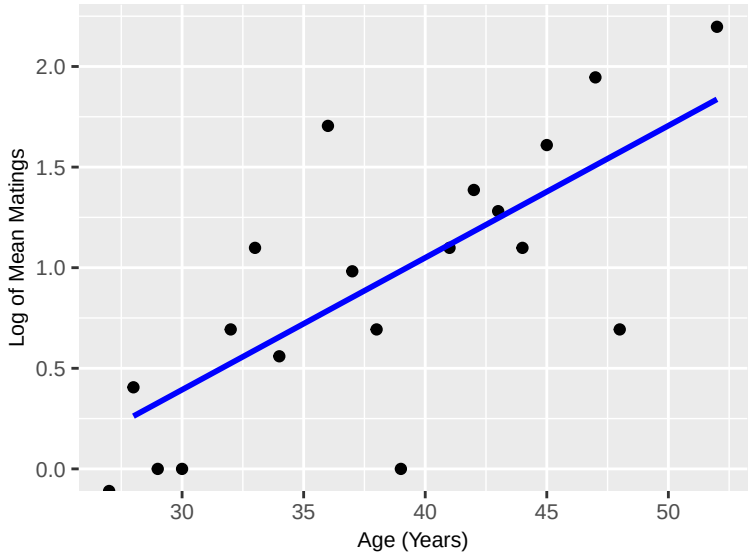


Figure 2: Log of Mean Matings by Age.

The points seem to follow a linear trend, suggesting that the log-linear relationship assumed in Poisson regression is reasonable. Overall, the plot supports the linearity assumption and the simple linear model seems to be appropriate for now.

e) Group the elephants by age, with age ranges of (25-30], (30-35], (30-40], (40-45], (45-50], (50-55]. Create histograms displying the number of matings by elephants in each age group. (Hint: use the `cut` function). Create a table displaying the average number of matings in each age range, as well as the variance. Does this table raise any concerns about any assumptions related to the Poisson regression model? If so, which?

Table 1: Grouped by Age

AgeGroups	Mean_Matings	Var_Matings	n
(25,30]	1.083333	0.9924242	12
(30,35]	2.400000	0.9333333	10
(35,40]	3.142857	5.8095238	7
(40,45]	3.666667	6.0000000	9
(45,50]	4.500000	12.5000000	2
(50,55]	9.000000	NA	1

The table raises concerns about the mean=variance assumption. Some groups, like (25,30] and (30,35], show underdispersion while (35,40] and (45,50], show overdispersion. This suggests that a standard Poisson model may not be appropriate, and we might consider using a quasi-Poisson model. Additionally, the small sample sizes in some groups (45,50] and (50,55], make their variance estimates unreliable. We might want to combine these into one group.

f) Perform a goodness of fit test for the model. Are your results consistent with your observations in (d) and (e)? Explain why or why not.

```
MP1$deviance
```

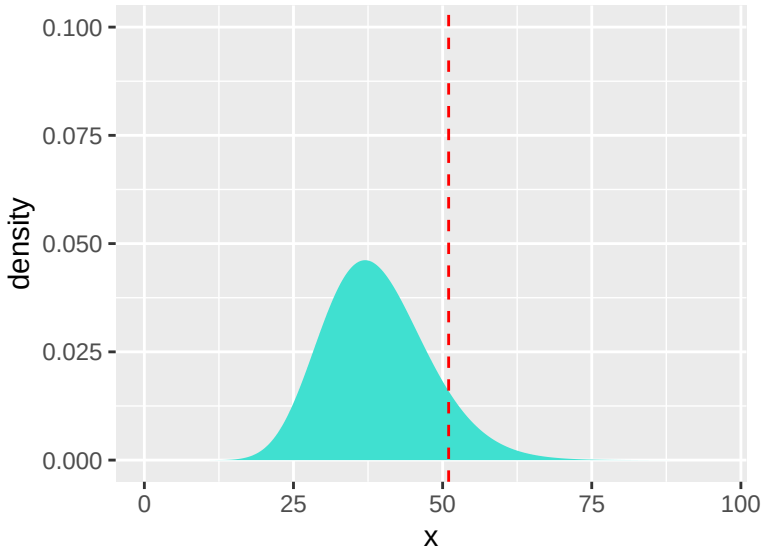
```
[1] 51.01163
```

```
MP1$df.residual
```

```
[1] 39
```

```
1-pchisq(MP1$deviance, MP1$df.residual)
```

```
[1] 0.09426231
```



The goodness-of-fit test results show a residual deviance of 51.01 with 39 degrees of freedom, and a p-value of 0.094, which is not statistically significant. This suggests that the Poisson model is an adequate fit for the data overall. The visualization further supports this, as the observed deviance falls within the expected range of the chi-square distribution.

These results are partially consistent with the observations from (d) and (e). In (d), the log-linear plot suggested that a Poisson model was reasonable in terms of linearity. However, in (e), we observed overdispersion and underdispersion in different age groups, which could suggest some deviation from Poisson assumptions. While the overall goodness-of-fit test does not indicate a poor fit, the variance patterns in (e) still suggest that it might be be worth using a quasi-Poisson model.

Question 2:

An article in the *Journal of Animal Ecology* by Bishop(1972) investigated whether moths provide evidence of "survival of the fittest" with their camouflage traits. Researchers glued equal numbers of light and dark morph moths in lifelike positions on tree trunks at 7 locations from 0 to 51.2 km from Liverpool. They then recorded the number of moths removed after 24 hours, presumably by predators. The hypothesis was that, since tree trunks near Liverpool were blackened by pollution, light morph moths would be more likely to be removed near Liverpool.

Data (Ramsey and Schafer, 2002) can be found in `moth.csv` and contains the variables below.

- `` `MORPH` = light or dark ``
- `` `DISTANCE` = kilometers from Liverpool ``
- `` `PLACED` = number of moths of a specific morph glued to trees at that location ``
- `` `REMOVED` = number of moths of a specific morph removed after 24 hours ``

We want to model the number of moths removed out of the total number placed, using morph and distance as explanatory variables.

a) Which type of model is most appropriate for these data? Explain your reasoning.

The response variable represents the proportion of moths removed out of the total placed. Since each moth is either removed or not, the data follows a binomial outcome. The most appropriate model for this data is Binomial Logistic Regression.

b) Fit the model you chose. Display the summary output.

I will make a new column called REMAINING using PLACED - REMOVED

MORPH	DISTANCE	PLACED	REMOVED	REMAINING
light	0.0	56	17	39
dark	0.0	56	14	42
light	7.2	80	28	52
dark	7.2	80	20	60
light	24.1	52	18	34
dark	24.1	52	22	30

```
MB1 <- glm(cbind(REMOVED, REMAINING) ~ DISTANCE + MORPH , family = binomial(link="logit"), data = mo
```

MB1

Consider Y_i to represent the number of moths removed for the i-th observation, where each Y_i follows a binomial distribution.

$$Y_i \sim \text{Binomial}(n_i, p_i)$$

Here, n_i is the total number of moths placed, and p_i is the probability of a moth being removed. The relationship between the probability of removal and the explanatory variables is given by:

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 \text{DISTANCE} + \beta_2 \text{MORPH}$$

```
# Call:
# glm(formula = cbind(REMOVED, REMAINING) ~ DISTANCE + MORPH, family = binomial(link = "logit"),
#     data = moth)
#
# Coefficients:
#             Estimate Std. Error z value Pr(>|z|)
# (Intercept) -0.732690   0.151221  -4.845 1.27e-06 ***
# DISTANCE     0.005314   0.004002   1.328  0.18422
# MORPHlight  -0.404052   0.139377  -2.899  0.00374 **
# ---
# Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
#
# (Dispersion parameter for binomial family taken to be 1)
#
```

```
#     Null deviance: 35.385  on 13  degrees of freedom
# Residual deviance: 25.161  on 11  degrees of freedom
# AIC: 93.836
#
# Number of Fisher Scoring iterations: 4
```

Exponetiated MB1 Coefficients:

```
(Intercept)    DISTANCE  MORPHlight
  0.4806144   1.0053278   0.6676093
```

c) Write sentences interpreting the coefficients associated with the `DISTANCE` and `MORPH` variables.

The DISTANCE coefficient (1.0053) suggests that for each 1 km increase, the odds of removal increase by 0.53%, though this effect is not statistically significant ($p = 0.184$). The MORPH coefficient (0.668) shows that light morph moths have 33.2% lower odds of being removed compared to dark morphs ($p = 0.0037$), suggesting that dark moths are more likely to survive, supporting the camouflage hypothesis.

d) Calculate the probability of a moth being removed assuming it is 15 km from Liverpool and is light `MORPH`. Then, calculate the probability of a moth being removed assuming it is 35 km from Liverpool and is dark `MORPH`.

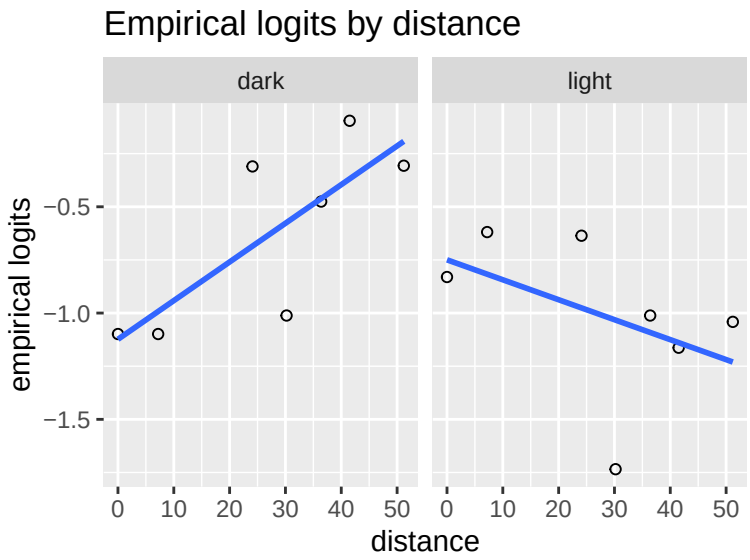
```
[1] "Probability of light moth being removed with distance of 15km from Liverpool"
```

```
[1] 0.2578771
```

```
[1] "Probability of dark moth being removed with distance of 35km from Liverpool"
```

```
[1] 0.3666304
```

A logit is the log of the odds of a moth being removed within 24 hours. The following code will create an empirical logit plot of logits vs. distance, faceted by morph.



e) What should we conclude from the plots in (d)? What do they say about the possibility of an interaction between morph and distance?

The empirical logit plots suggest a potential interaction between morph and distance. For dark moths, the log-odds of removal increase with distance, meaning they are more likely to be removed further from Liverpool. In contrast, for light moths, the log-odds decrease, suggesting they are less likely to be removed at greater distances. This opposite trend indicates that morph type influences how distance affects removal probability, supporting the need for an interaction term in the model.

f) Create a model with `DISTANCE`, `MORPH`, and the interaction between both variables. Display the summary output.

MB2

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 \text{DISTANCE} + \beta_2 \text{MORPH} + \beta_3 \text{DISTANCE} \times \text{MORPH}$$

```
MB2 <- glm(cbind(REMOVED, REMAINING) ~ DISTANCE * MORPH,
                              family = binomial(link = "logit"),
                              data = moth)
```

```
#
# Call:
# glm(formula = cbind(REMOVED, REMAINING) ~ DISTANCE * MORPH, family = binomial(link = "logit"),
#     data = moth)
#
# Coefficients:
#                      Estimate Std. Error z value Pr(>|z|)
# (Intercept)         -1.128987   0.197906  -5.705 1.17e-08 ***
```

```
# DISTANCE             0.018502   0.005645   3.277 0.001048 **
# MORPHlight           0.411257   0.274490   1.498 0.134066
# DISTANCE:MORPHlight -0.027789   0.008085  -3.437 0.000588 ***
# ---
# Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
#
# (Dispersion parameter for binomial family taken to be 1)
#
#     Null deviance: 35.385  on 13  degrees of freedom
# Residual deviance: 13.230  on 10  degrees of freedom
# AIC: 83.904
#
# Number of Fisher Scoring iterations: 4
```

Exponentiated Coefficients

```
       (Intercept)            DISTANCE          MORPHlight DISTANCE:MORPHlight
         0.3233608           1.0186745           1.5087133           0.9725935
```

g) As distance gets farther from the city, do light moths become more or less likely to be removed? What about dark moths? Cite values from your model output in (f) to justify your answer.

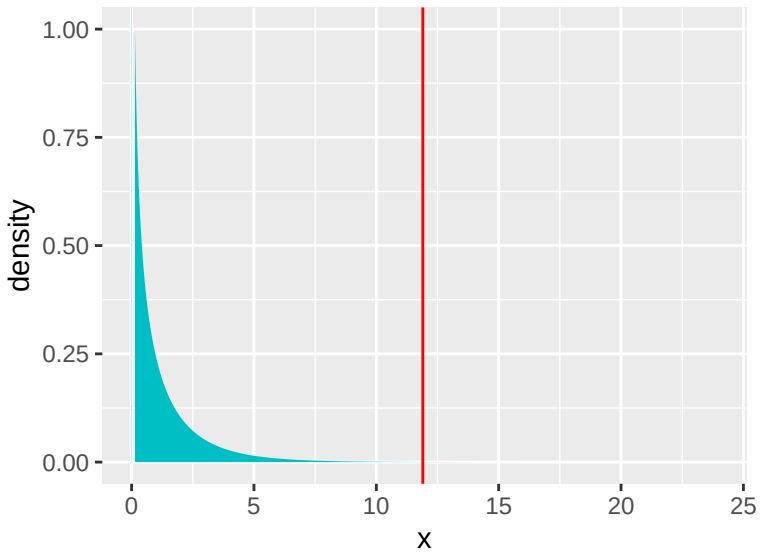
The exponentiated coefficients show that dark moths become more likely to be removed as distance increases, with the odds of removal increasing by 1.87% per kilometer (1.0187). In contrast, light moths become less likely to be removed further from Liverpool, as indicated by the interaction term (0.9726), meaning their odds of removal decrease by 2.74% per kilometer. At 0 km (Liverpool), light moths initially have 50.87% higher odds of being removed than dark moths (1.5087), but this effect diminishes with distance. These results support the idea that light moths gain a survival advantage in less polluted areas, where tree trunks are lighter, while dark moths are more vulnerable when further from the city.

h) Perform a drop-in-deviance test whether there is evidence of an interaction between distance and morph. Explain your conclusion in context.

```
anova(MB1, MB2, test = "Chisq")
```

```
Analysis of Deviance Table

Model 1: cbind(REMOVED, REMAINING) ~ DISTANCE + MORPH
Model 2: cbind(REMOVED, REMAINING) ~ DISTANCE * MORPH
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1        11     25.161
2        10     13.230  1   11.931 0.0005519 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



The drop-in-deviance test shows a significant reduction in deviance (11.931, p = 0.00055) when adding the interaction term between distance and morph, indicating that the interaction significantly improves model fit. This suggests that the effect of distance on moth removal differs between light and dark morphs.

The visualization of the chi-square distribution further supports this conclusion. The observed statistic, indicates a very low probability that the improvement in model fit is due to chance.

i) Test the goodness-of-fit for the interaction model. What can we conclude about this model?

```
MB2$deviance
```

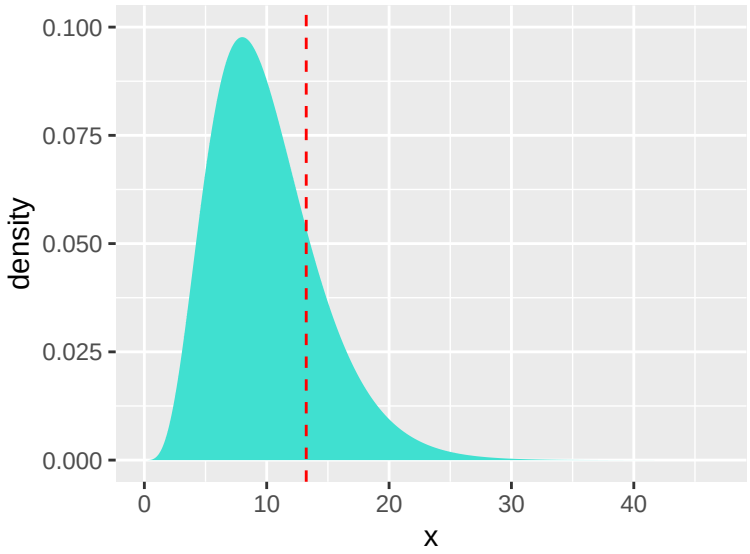
```
[1] 13.2299
```

```
MB2$df.residual
```

```
[1] 10
```

```
1-pchisq(MB2$deviance, MB2$df.residual)
```

```
[1] 0.2111003
```



The goodness-of-fit test for the interaction model shows a residual deviance of 13.23 with 10 degrees of freedom, resulting in a p-value of 0.211. Since the p-value is greater than 0.05, we do not have strong evidence to reject the model, indicating that it fits the data well. The chi-square distribution visualization further supports this conclusion, as the observed deviance falls within an expected range, suggesting that the model adequately captures the variability in moth removal.