Homework 4

Noah Johnson

March 7, 2018

Bolstad 5.8

Let Y have the Poisson($\mu = 3$) distribution

(a) Calculate P(Y=2).

$$P(Y = 2|\mu = 3) = \frac{3^2 e^{-3}}{2!} \approx 0.224$$

Or alternatively, we can use R.

```
dpois(x = 2, lambda = 3)
```

```
## [1] 0.2240418
```

(b) Calculate P($Y \leq 2$).

```
ppois(2, 3)
```

```
## [1] 0.4231901
```

(c) Calculate P($1 \leq Y < 4$).

This is the same as $P(1 \leq Y \leq 3) = P(Y \leq 3) - P(Y = 0)$.

```
ppois(3, 3) - dpois(0, 3)
```

```
## [1] 0.5974448
```

BYSH 4.15.1 Conceptual Exercise 1

What are features of inferential OLS models that make them less suitable for count data?

We could try to fit a simple linear model to the expected mean, such as $\lambda_{X_i} = \beta_0 + \beta_1 X_i$, and then fit this model by minimizing the residual sum of squares $\sum_i (y_i - \lambda_{X_i})^2$.

But count data can never be negative, which this simple linear model allows. Also, to perform inference with OLS estimates, we have to assume equal variance of the response across all predictor values. I'm not sure why count data in particular would be unlikely to satisfy this requirement.

BYSH 4.15.1 Conceptual Exercise 2

Models of the form $Y_i = \beta_0 + \beta_1 X_i, \epsilon \sim iidN(0, \sigma)$ are fit using the method of least squares [note that least squares estimates for linear regression are also MLEs]. What method is used to fit Poisson regression models?

Poisson regression models of the form $log(\lambda_{X_i}) = \beta_0 + \beta_1 X_i$ are fit using the method of maximum likelihood estimation. The joint probability of the observed data occurring is simply the product of each Poisson random variable, since the observations are assumed to be independent. The log of this likelihood is taken, and numerical methods for convex optimization such as gradient descent are used to maximize this function

with respect to the coefficients. This is equivalent to choosing the generating model which maximizes the probability of observing the data which we did observe.

BYSH 4.15.2 Guided Exercise 1 Elephant Mating

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 is found in elephant.csv, and relevant R code can be found under elephantMating.R.

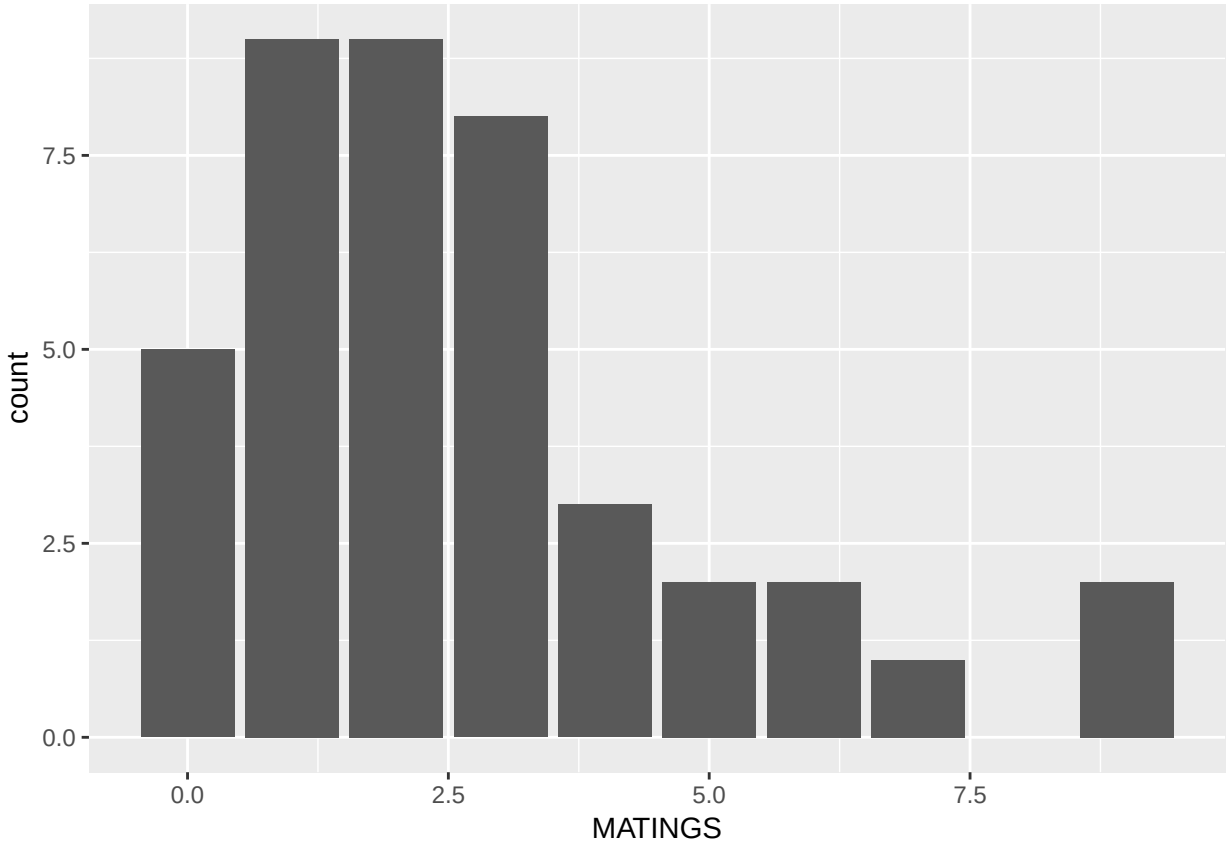
The variables are:
MATINGS: the number of matings in a given year
AGE: the age of the elephant in years.

```
elephant_matings <- read.csv('elephant.csv')
str(elephant_matings)
```

```
## 'data.frame':    41 obs. of  2 variables:
##  $ AGE    : num  27 28 28 28 28 29 29 29 29 29 ...
##  $ MATINGS: num  0 1 1 1 3 0 0 0 2 2 ...
```

(a) Create a histogram of MATINGS. Is there preliminary evidence that number of matings could be modeled as a Poisson response? Explain.

```
elephant_matings %>% ggplot(aes(MATINGS)) +
  geom_bar()
```

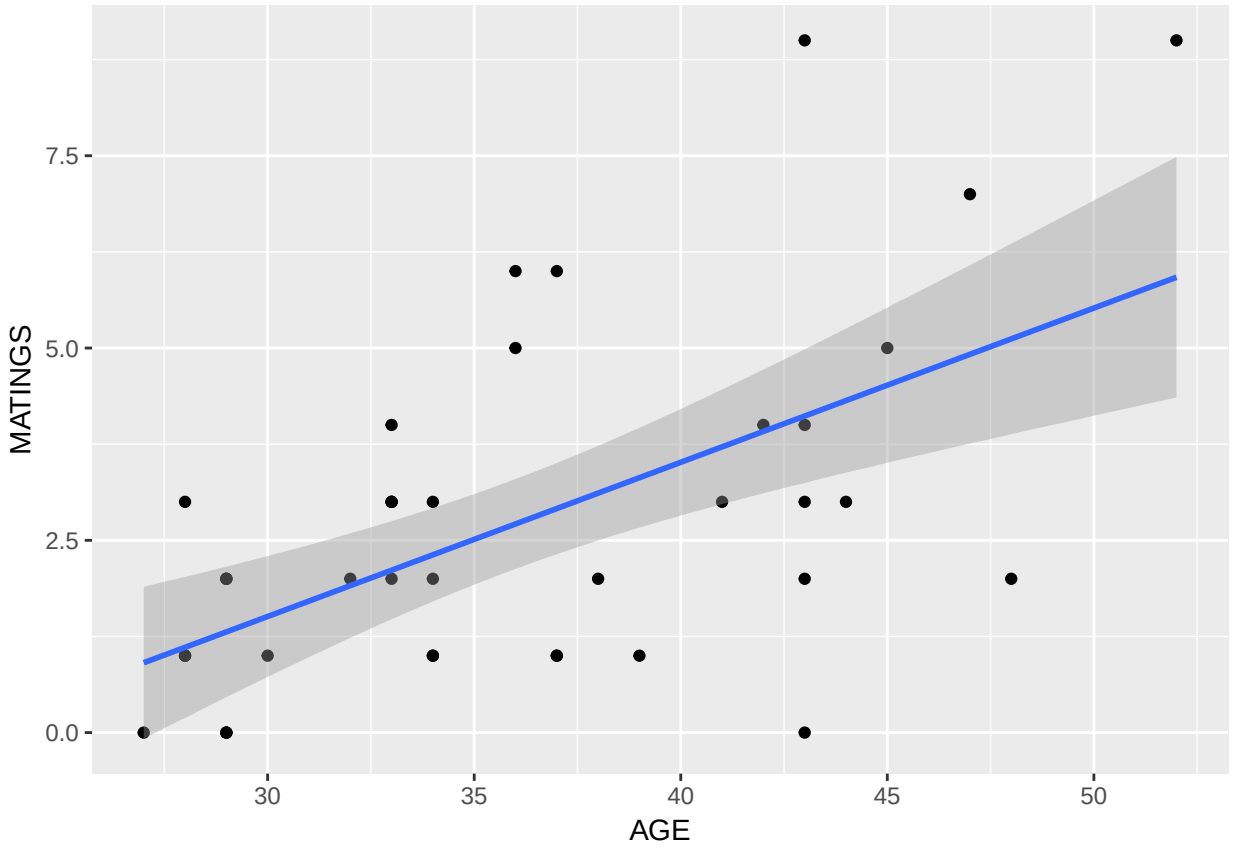


The histogram suggests that the distribution of the number of matings is skewed. It looks decidedly not normally distributed, since it is sharply cut off at 0. Perhaps if the mean were higher a normal distribution would be an acceptable approximation, but here it will not work.

The fact that the response is discrete and >= 0 indicates that we can model the response as a Poisson random variable.

(b) Plot MATINGS by AGE. Add a least squares line. Is there evidence that modeling matings using a linear regression with age might not be appropriate? Explain.

```
elephant_matings %>% ggplot(aes(x = AGE, y = MATINGS)) +
  geom_point() +
  geom_smooth(method='lm')
```



There seems to be some visual evidence of heteroscedasticity - there is more variance in the number of matings for higher values of age. This may make a linear regression inappropriate.

(c) For each age, calculate the mean number of matings. Take the log of each mean and plot it by AGE.

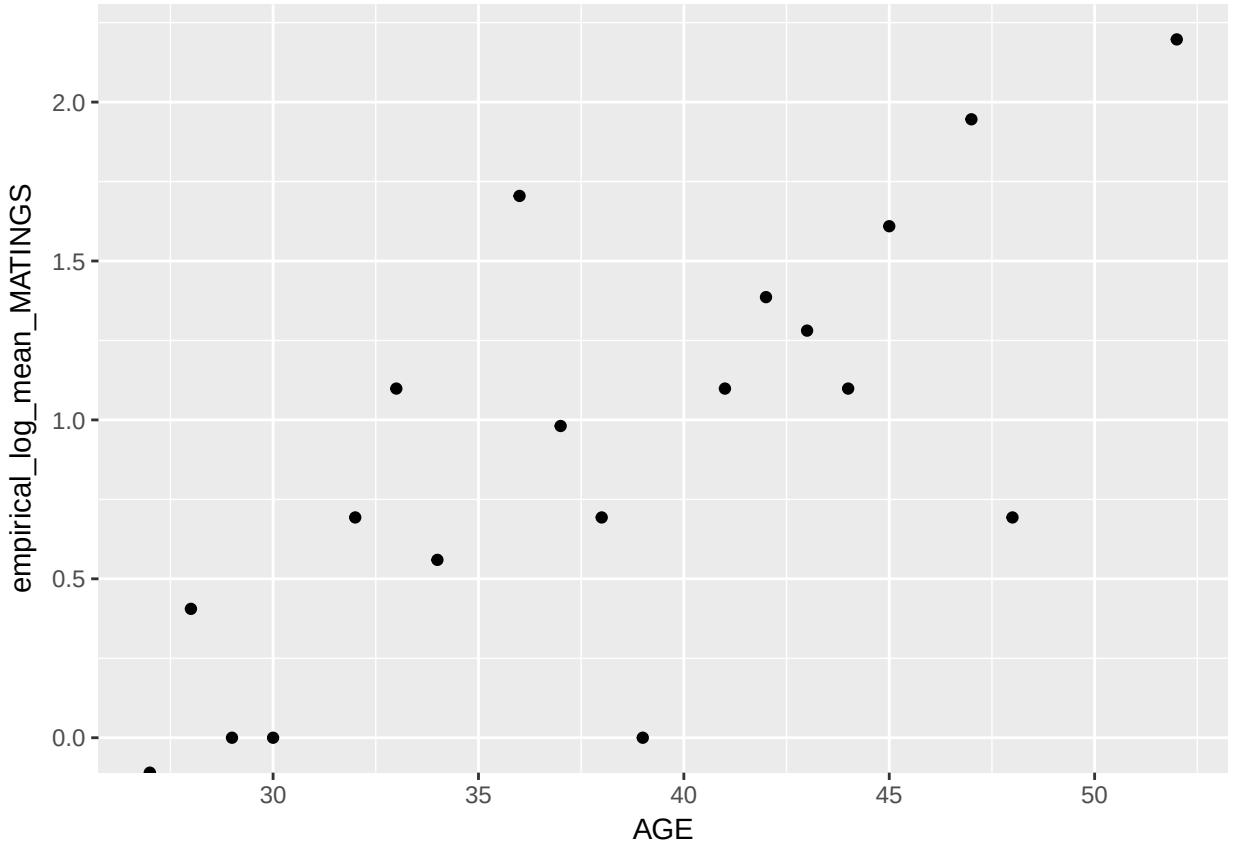
```
empirical_means <- elephant_matings %>% group_by(AGE) %>% summarise(mean(MATINGS))
print(empirical_means)
```

```
## # A tibble: 19 x 2
##      AGE `mean(MATINGS)`
##    <dbl>           <dbl>
##  1    27        0.000000
##  2    28        1.500000
##  3    29        1.000000
##  4    30        1.000000
```

```
##  5    32        2.000000
##  6    33        3.000000
##  7    34        1.750000
##  8    36        5.500000
##  9    37        2.666667
## 10    38        2.000000
## 11    39        1.000000
## 12    41        3.000000
## 13    42        4.000000
## 14    43        3.600000
## 15    44        3.000000
## 16    45        5.000000
## 17    47        7.000000
## 18    48        2.000000
## 19    52        9.000000
```

```
empirical_log_means <- empirical_means %>% mutate(empirical_log_mean_MATINGS=log(`mean(MATINGS)`))

empirical_log_means %>% ggplot(aes(x = AGE, y = empirical_log_mean_MATINGS)) +
  geom_point()
```



i. What assumption can be assessed with this plot?

We can check the assumption that $log(\lambda_X)$ is linear in X.

ii. Is there evidence of a quadratic trend on this plot?

There doesn't seem to be an obvious quadratic trend on this plot.

(d) Fit a Poisson regression model with a linear term for AGE. Interpret the coefficient for

AGE. Exponentiate and interpret the result.

```
pm <- glm(MATINGS ~ AGE, family = poisson, data = elephant_matings)
pm
```

```
##
## Call:  glm(formula = MATINGS ~ AGE, family = poisson, data = elephant_matings)
##
## Coefficients:
## (Intercept)          AGE
##    -1.58201      0.06869
##
## Degrees of Freedom: 40 Total (i.e. Null);  39 Residual
## Null Deviance:       75.37
## Residual Deviance: 51.01     AIC: 156.5
```

Interpreting just the coefficient without exponentiating it is difficult. It is the expected gain in $log(\lambda)$ when age increases by 1 year. Thanks to log rules, this means it is also the log of the ratio of means.

$log(\frac{\lambda_{X+1}}{\lambda_X}) = 0.0686928$

But this coefficient is best interpreted after exponentiating.

$\frac{\lambda_{X+1}}{\lambda_X} = e^{\beta_1} = 1.0711071$ tells us that the expected number of matings increases by 7.1% for every additional year old the male elephant is.

(e) Construct a 95% confidence interval for the slope and interpret in context (you may want to exponentiate endpoints).

```
confint(pm)
```

```
## Waiting for profiling to be done...
```

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

```
exp(confint(pm))
```

```
## Waiting for profiling to be done...
```

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

So we are 95% confident that the expected number of matings increases by between 4.3 and 10 percent for every additional year old the male elephant is.