ST 518 - Homework 5

Nora Quick

R Question:

1. Using Poisson log linear regression, test for independence between obesity and CVD death outcome. How does your answer compare to a chi-square test on the same data?

```
obesity <- expand.grid(
  Obesity = factor(c("Obese", "Notobese"), levels = c("Obese", "Notobese")),
  Dead = factor(c("Deaths", "NonDeaths"), c("Deaths", "NonDeaths")))
obesity$Freq <- c(16, 7, 2045, 1044)
obesity_tab <- xtabs(data = obesity, Freq ~ Obesity + Dead)

ftable(obesity_tab)
```

```
##          Dead Deaths NonDeaths
## Obesity
## Obese              16      2045
## Notobese            7      1044
```

```
mod1 <- glm(data = obesity_tab, Freq ~ Obesity + Dead, family = poisson)
summary(mod1)
```

```
##
## Call:
## glm(formula = Freq ~ Obesity + Dead, family = poisson, data = obesity_tab)
##
## Deviance Residuals:
##        1         2         3         4
##  0.19508  -0.28018  -0.01697   0.02377
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)       2.7234     0.2089   13.04   <2e-16 ***
## ObesityNotobese  -0.6734     0.0379  -17.77   <2e-16 ***
## DeadNonDeaths     4.9001     0.2093   23.41   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
```

```
##     Null deviance: 4376.49698  on 3  degrees of freedom
## Residual deviance:    0.11741  on 1  degrees of freedom
## AIC: 32.796
##
## Number of Fisher Scoring iterations: 3
```

```
mod2 <- glm(data = obesity_tab, Freq ~ Obesity + Dead + (Obesity * Dead), family = poisson)
summary(mod2)
```

```
##
## Call:
## glm(formula = Freq ~ Obesity + Dead + (Obesity * Dead), family = poisson,
##     data = obesity_tab)
##
## Deviance Residuals:
## [1]  0  0  0  0
##
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                    2.7726     0.2500  11.090   <2e-16 ***
## ObesityNotobese               -0.8267     0.4532  -1.824   0.0681 .
## DeadNonDeaths                  4.8506     0.2510  19.327   <2e-16 ***
## ObesityNotobese:DeadNonDeaths  0.1543     0.4548   0.339   0.7343
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 4.3765e+03  on 3  degrees of freedom
## Residual deviance: 4.7296e-13  on 0  degrees of freedom
## AIC: 34.678
##
## Number of Fisher Scoring iterations: 3
```

```
chisq.test(obesity_tab)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  obesity_tab
## X-squared = 0.014031, df = 1, p-value = 0.9057
```

As we can see from both models above the obesity and death values indicate that they are significant on their own but together (as seen in the second model) together there is no significance. We can also see this in the chi-squared test with a large p-value indicating the variables are independent of each other.

2. Be sure to examine residuals from each of these models. How do the models compare? Please be specific. Is there evidence of over dispersion? If so, fit another model and report results from that model. If not, why not?

```
work <- case2201
work
```

```
##    Age Matings
## 1   27       0
## 2   28       1
## 3   28       1
## 4   28       1
## 5   28       3
## 6   29       0
## 7   29       0
## 8   29       0
## 9   29       2
## 10  29       2
## 11  29       2
## 12  30       1
## 13  32       2
## 14  33       4
## 15  33       3
## 16  33       3
## 17  33       3
## 18  33       2
## 19  34       1
## 20  34       1
## 21  34       2
## 22  34       3
## 23  36       5
## 24  36       6
## 25  37       1
## 26  37       1
## 27  37       6
## 28  38       2
## 29  39       1
## 30  41       3
## 31  42       4
## 32  43       0
## 33  43       2
## 34  43       3
## 35  43       4
## 36  43       9
## 37  44       3
## 38  45       5
## 39  47       7
## 40  48       2
## 41  52       9
```

a. simple linear regression after taking a square root transformation of the number of successful mat-ings

```
work$sq_mat <- sqrt(work$Matings)
```

```r
mod3 <- lm(data = work, Age ~ sq_mat)
summary(mod3)
```

```
##
## Call:
## lm(formula = Age ~ sq_mat, data = work)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.1574 -4.1574 -0.0961  3.8426 13.9549
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   29.045      1.895  15.331  < 2e-16 ***
## sq_mat         4.684      1.157   4.049 0.000236 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.59 on 39 degrees of freedom
## Multiple R-squared:  0.296,  Adjusted R-squared:  0.2779
## F-statistic:  16.4 on 1 and 39 DF,  p-value: 0.0002362
```

b. simple linear regression after taking a logarithmic transformation (after adding 1)

```r
plus <- work$Matings + 1

work$log_mat <- log(plus)

mod4 <- lm(data = work, Age ~ log_mat)
summary(mod4)
```

```
##
## Call:
## lm(formula = Age ~ log_mat, data = work)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.3595 -4.3595 -0.3107  3.6649 13.6893
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   29.311      1.838  15.949  < 2e-16 ***
## log_mat        5.806      1.435   4.046 0.000238 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.591 on 39 degrees of freedom
## Multiple R-squared:  0.2957, Adjusted R-squared:  0.2776
## F-statistic: 16.37 on 1 and 39 DF,  p-value: 0.0002385
```

c. log-linear regression.

```
mod5 <- glm(data = work, Age ~ Matings, family = poisson)
summary(mod5)
```

```
##
## Call:
## glm(formula = Age ~ Matings, family = poisson, data = work)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4166  -0.6962  -0.4461   0.6703   2.1524
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  3.45374    0.04148  83.273  < 2e-16 ***
## Matings      0.04492    0.01105   4.066 4.79e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 47.279  on 40  degrees of freedom
## Residual deviance: 31.380  on 39  degrees of freedom
## AIC: 257.03
##
## Number of Fisher Scoring iterations: 4
```

Interestingly they all appear to be fine models for looking at the interaction between age and successful mating. Overall, they all show a significant correlation between the age of an elephant and the succuss of mating. I would say that the log linear regression is the best fit but the other models do show similar outcomes. Each model has a small p-value for successful mating in relation to age.

Conceptual Questions:

3. What is the difference between a log-linear model and a linear model after the log transformation of the response?

The difference can come down to what is being asked/what is being looked for. We use log-linear models when we want to know the outcome of the overall mean of a variable. We use a linear model after the log transformation to look at the mean outcome of something that has been log transformed.

4. his question refers to the elephant mating data from question 2 above.

a. Both the binomial and the Poisson distributions provide probability models for random counts. Which distribution is more appropriate to model the number of successful matings for male African elephants, and why?

The Poisson distribution model because we want to know the mean of all males. The poisson distribution model gives the outcome of successfully matings for all males and not just the ones that successfully mated.

In addition the response is count data and we want the probability of an eveny which makes it a random event.

b. In the following plot, we see that the spread of responses is larger for larger values of the mean response. Is this something to be concerned about if we perform a Poisson log-linear regression?

Yes, becuase to use the Poisson log-linear regression model we assume linearity and from the graph we can see that it is not linear and the variation is not even.

c. Performing a Poisson log-linear regression results in the following output: What are the mean and variance of the distribution of counts of successful matings (in 8 years) for elephants who are aged 25 years at the beginning of the observation period? What are the mean and variance for elephants who are aged 45 years?

For 25 year old elephants they have a mean/variance of 1 successful mating. For 45 year old elephants they have a mean/variance of 5 successful matings.