HW5

Ben Tankus

4/28/2021

```r
library(arm)
```

```
## Warning: package 'arm' was built under R version 4.0.5

## Loading required package: MASS

## Warning: package 'MASS' was built under R version 4.0.3

## Loading required package: Matrix

## Loading required package: lme4

## Warning: package 'lme4' was built under R version 4.0.3

##
## arm (Version 1.11-2, built: 2020-7-27)

## Working directory is C:/Users/tanku/OneDrive/CodeRoot/RootRCode/ST518
```

```r
library(Sleuth3)
```

```
## Warning: package 'Sleuth3' was built under R version 4.0.3
```

```r
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.0.3

## -- Attaching packages ------------------------------------------------------------------------------

## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.0.3     v dplyr   1.0.2
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.0

## Warning: package 'ggplot2' was built under R version 4.0.3

## Warning: package 'tidyr' was built under R version 4.0.3
## Warning: package 'readr' was built under R version 4.0.3
```

```
## Warning: package 'purrr' was built under R version 4.0.3
```

```
## Warning: package 'dplyr' was built under R version 4.0.3
```

```
## Warning: package 'forcats' was built under R version 4.0.3
```

```
## -- Conflicts ---------------------------------------------------------------------------------------
## x tidyr::expand() masks Matrix::expand()
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## x tidyr::pack()   masks Matrix::pack()
## x dplyr::select() masks MASS::select()
## x tidyr::unpack() masks Matrix::unpack()
```

```r
library(vcdExtra)
```

```
## Warning: package 'vcdExtra' was built under R version 4.0.4
```

```
## Loading required package: vcd
```

```
## Warning: package 'vcd' was built under R version 4.0.4
```

```
## Loading required package: grid
```

```
## Loading required package: gnm
```

```
## Warning: package 'gnm' was built under R version 4.0.4
```

```
##
## Attaching package: 'vcdExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     summarise
```

```r
library(magrittr)
```

```
##
## Attaching package: 'magrittr'
```

```
## The following object is masked from 'package:purrr':
##
##     set_names
```

```
## The following object is masked from 'package:tidyr':
##
##     extract
```

R Questions 1. (2 points) Recall the obesity problem from Homework 1. The data are as follows:

```
CVD Death
 Yes No
```

Obese 16 2045 Not obese 7 1044

```
cvdDeath <- expand.grid(
  obesity = factor(c('Obese', 'Not obese'), levels = c('Obese', 'Not obese')),
  death = factor(c('Death', 'No Death'), levels = c('Death', 'No Death')))
cvdDeath$Freq <- c( 16, 7, 2045, 1044)

cvdDeath_tab <- xtabs(data = cvdDeath, Freq ~ obesity + death)

head(cvdDeath_tab)
```

```
##            death
## obesity     Death No Death
##   Obese        16     2045
##   Not obese     7     1044
```

Using Poisson log linear regression, test for independence between obesity and CVD death outcome. (Hint: this is equivalent to testing that the interaction term in the model is zero.) How does your answer compare to a chi-square test on the same data?

```
fit <- glm(Freq ~ obesity + death + obesity:death , data = cvdDeath_tab, family = poisson)
summary(fit)
```

```
##
## Call:
## glm(formula = Freq ~ obesity + death + obesity:death, family = poisson,
##     data = cvdDeath_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 ***
## obesityNot obese              -0.8267     0.4532  -1.824   0.0681 .
## deathNo Death                  4.8506     0.2510  19.327   <2e-16 ***
## obesityNot obese:deathNo Death   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: 5.3246e-13  on 0  degrees of freedom
## AIC: 34.678
##
## Number of Fisher Scoring iterations: 3
```

```r
anova(fit, test = ’Chisq’)
```

```
## Analysis of Deviance Table
##
## Model: poisson, link: log
##
## Response: Freq
##
## Terms added sequentially (first to last)
##
##
##               Df Deviance Resid. Df Resid. Dev Pr(>Chi)
## NULL                              3     4376.5
## obesity        1    333.8         2     4042.7   <2e-16 ***
## death          1   4042.6         1        0.1   <2e-16 ***
## obesity:death  1      0.1         0        0.0   0.7319
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

Using poisson log linear regression these two factors (obesity and death) are independent, with an interaction term pvalue of 0.7343 (95% level confidence interval). This is the same as the chi-sq test.

2. (3 points) Although male elephants are capable of reproducing by 14 to 17 years of age, younger adult males are usually unsuccessful in competing with their larger elders for the attention of receptive females. Since male elephants continue to grow throughout their lifetimes, and since larger males tend to be more successful at mating, the males most likely to pass their genes to future generations are those whose characteristics enable them to live long lives. Joyce Poole studied a population of African elephants in Amboseli National Park, Kenya, for 8 years. You can explore this data set in case2201 in the Sleuth3 library. This data frame contains the number of successful matings and ages (at the study’s beginning) of 41 male elephants.

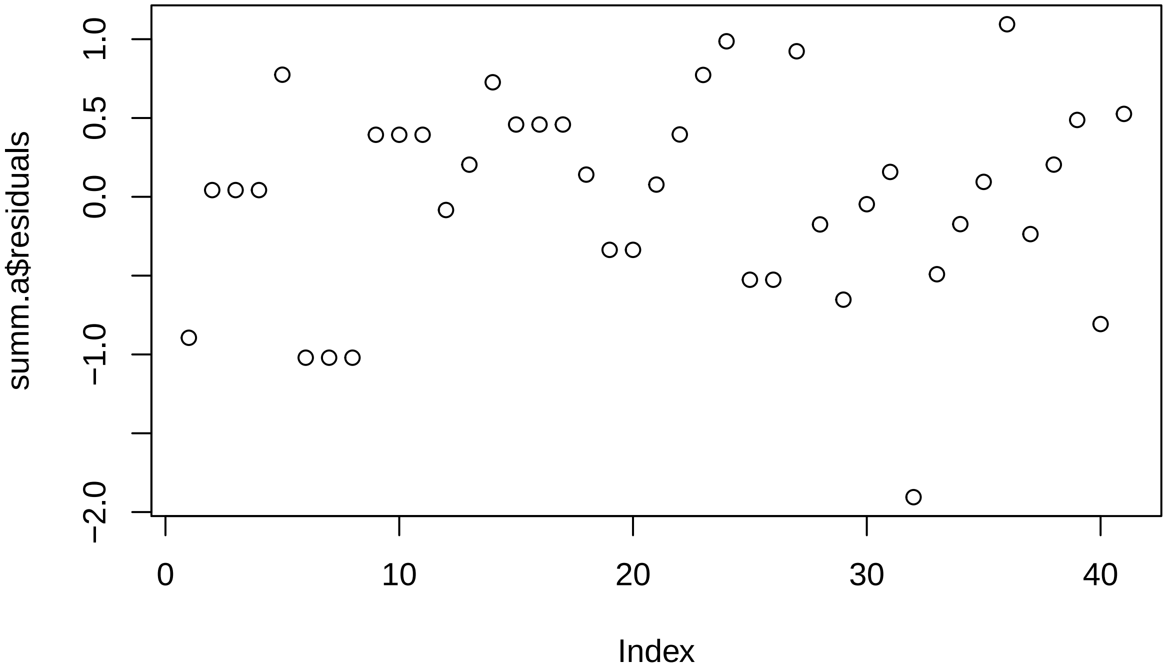
Give an estimated model for describing the number of successful matings as a function age, using

```r
df <- case2201

df[’sqrtMatings’] <- sqrt(df$Matings)
fit.a <- lm(sqrtMatings ~ Age, data = df)
summ.a <- summary(fit.a)
print(summ.a)
```

```
##
## Call:
## lm(formula = sqrtMatings ~ Age, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.90532 -0.33654  0.07767  0.45871  1.09468
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.81220    0.56867  -1.428 0.161187
## Age          0.06320    0.01561   4.049 0.000236 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6493 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
```

```
plot(summ.a$residuals)
```

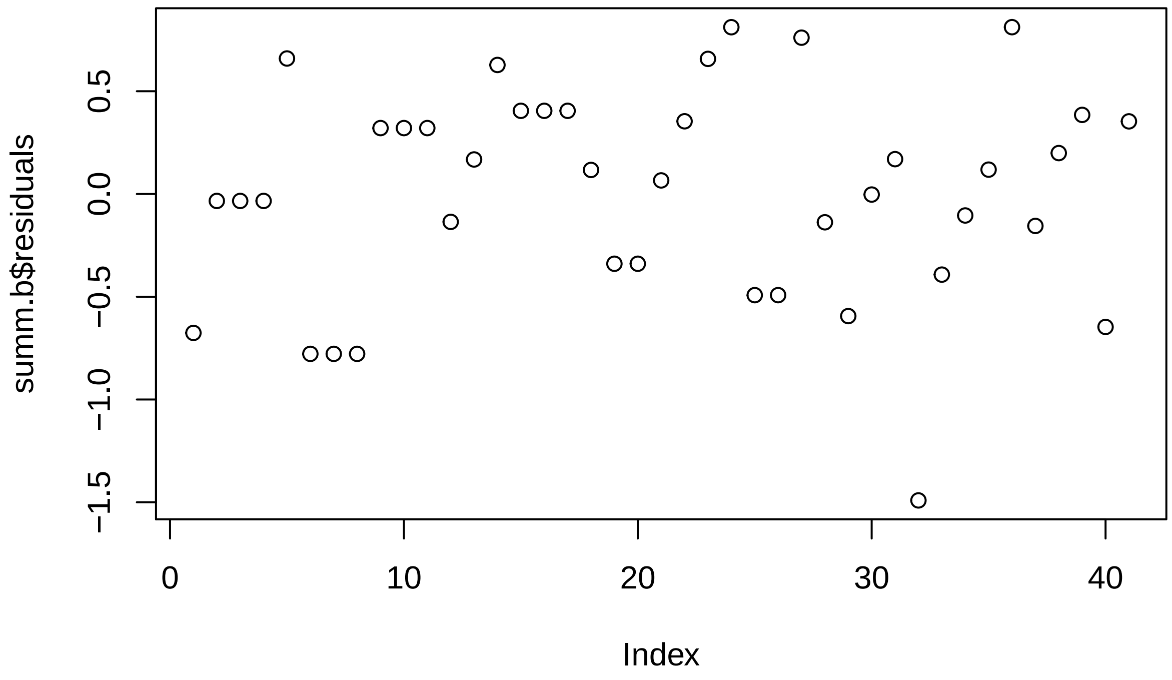


```
df['MatingsLog'] <- log(df$Matings + 1)
fit.b <- lm(MatingsLog ~ Age, data = df)
summ.b <- summary(fit.b)
print(summ.b)
```

```
##
## Call:
## lm(formula = MatingsLog ~ Age, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.49087 -0.33939  0.06607  0.35376  0.81171
##
## Coefficients:
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.69893    0.45861  -1.524 0.135567
## Age          0.05093    0.01259   4.046 0.000238 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.5237 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
```

```
plot(summ.b$residuals)
```

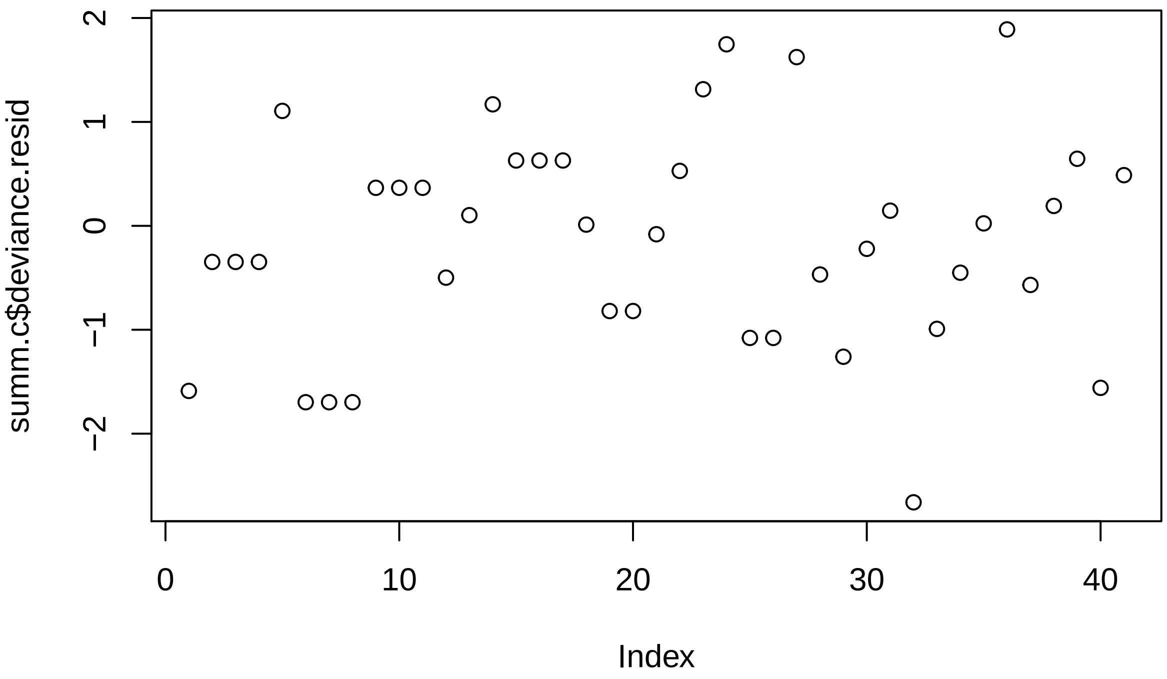


```
fit.c <- glm.nb(Matings ~ Age, data = df)
summ.c <- summary(fit.c)
print(summ.c)
```

```
##
## Call:
## glm.nb(formula = Matings ~ Age, data = df, init.theta = 16.48629005,
##     link = log)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.66032  -0.81966  -0.08079   0.52865   1.89021
##
```

```
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)    
## (Intercept) -1.58837    0.59393  -2.674  0.00749 ** 
## Age          0.06887    0.01519   4.534  5.8e-06 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
## 
## (Dispersion parameter for Negative Binomial(16.4863) family taken to be 1)
## 
##     Null deviance: 65.199  on 40  degrees of freedom
## Residual deviance: 44.498  on 39  degrees of freedom
## AIC: 157.92
## 
## Number of Fisher Scoring iterations: 1
## 
## 
##               Theta:  16.5 
##           Std. Err.:  25.7 
## 
##  2 x log-likelihood:  -151.923 
```

```
plot(summ.c$deviance.resid)
```



```
LRstats(fit.a,fit.b,fit.c)
```

```
## Likelihood summary table:
```

```
##            AIC     BIC LR Chisq Df Pr(>Chisq)
## fit.a  84.896  90.037   16.444 39     0.9994
## fit.b  67.256  72.397   10.695 39     1.0000
## fit.c 157.923 163.063   44.498 39     0.2514
```

(a) simple linear regression after taking a square root transformation of the number of successful matings;

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

(c) log-linear regression.

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?

The AIC and BIC values for the b model are best. The residual plot is also closest to zero in b, with a scale only extending to 0.5 as opposed to 2 in both A and C.

Conceptual Questions

3. (2 points) What is the difference between a log-linear model and a linear model after the log transformation of the response?

The difference is where we take the response average. a linear model takes the average of logged responses, where a log-linear model conducts the analysis, *then* transforms the mean to log-scale.

4. (3 points) This 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 appropriate to model the number of successful mating is male African elephants and why?

We should be using the poisson distribution because the probability of success is not the same for every trail (elephant mating). Older elephants have a better chance of mating, therefore, we should use a poisson to model age / mating.

According to the AIC and BIC values, SLR log transform model (b) is more appropriate to model the elephant mating pattern because the significantly smaller AIC and BIC values (67.26, 72.4) (b), (84.9, 90.0) (a), and (157.9, 44.5) (c) respectively, and the residual plot is closer to zero.

(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?

If the residual plot resembled this, with non-constant variance, I would be concerned. However since this is a direct measurement of the response, this variance should be handled within the log-linear transformation.

PLOT IN HW PDF

(c) Performing a Poisson log-linear regression results in the following output:

```
            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** * Residual Deviance: 51.01 on 39 degrees of freedom

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?

```
print(exp(-1.58 + 0.068*25))
```

```
## [1] 1.127497
```

```
print(exp(-1.58 + 0.068*45))
```

```
## [1] 4.392946
```

```
51.01/39
```

```
## [1] 1.307949
```