Assignment 2

Andy Yuan

2025-03-08

I used AI in this assignment to debug some of my code. All written responses are written by myself, with reference to course content and general online searches.

Question 1

a)

Establish that $SSE = \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p} X_{ij}\beta_j)^2$

Showing the deviance of a linear regression model is proportional to the residual sum of squares

$$
\begin{aligned}
D &= 2[logL_{Sat} - logL_M] \\
&= 2\Bigg[log\Bigg(\prod_{i=1}^{n}\frac{1}{\sigma\sqrt{2\pi}}exp\Bigg(-\frac{(y_i - y_i)^2}{2\sigma^2}\Bigg)\Bigg) - log\Bigg(\prod_{i=1}^{n}\frac{1}{\sigma\sqrt{2\pi}}exp\Bigg(-\frac{(y_i - \beta_0 - \sum_{j=1}^{p} X_{ij}\beta_j)^2}{2\sigma^2}\Bigg)\Bigg)\Bigg] \\
&= 2\Bigg[log\Bigg(\prod_{i=1}^{n}\frac{1}{\sigma\sqrt{2\pi}}exp(0)\Bigg) - log\Bigg(\prod_{i=1}^{n}\frac{1}{\sigma\sqrt{2\pi}}exp\Bigg(-\frac{(y_i - \beta_0 - \sum_{j=1}^{p} X_{ij}\beta_j)^2}{2\sigma^2}\Bigg)\Bigg)\Bigg] \\
&= 2\Bigg[-nlog(\sigma\sqrt{2\pi}) + \sum_{i=1}^{n}\frac{(y_i - \beta_0 - \sum_{j=1}^{p} X_{ij}\beta_j)^2}{2\sigma^2} + nlog(\sigma\sqrt{2\pi})\Bigg] \\
&= 2[SSE/2\sigma^2] \\
&= SSE/\sigma^2
\end{aligned}
$$

b)

$$P(Y_i = y_i) = \frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!}$$

$$D = 2logL_{Sat} - 2logL_m$$

$$
\begin{aligned}
logL_{Sat} &= \sum_{i=1}^{n} log\Bigg(\frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!}\Bigg) \\
&= \sum_{i=1}^{n}(-\lambda_i + y_i log\lambda_i - log(y_i!)) \\
&= \sum_{i=1}^{n}(-y_i + y_i logy_i - log(y_i!))
\end{aligned}
$$

$$logL_m = \sum_{i=1}^{n}(-\lambda_i + y_i log\lambda_i - log(y_i!))$$

$$\begin{aligned} logL_{Sat} - logL_m &= \sum_{i=1}^{n}\left[(-y_i + y_i logy_i - log(y_i!)) - (-\lambda_i + y_i log\lambda_i - log(y_i!))\right] \\ &= \sum_{i=1}^{n}(-y_i + \lambda_i + y_i log(\frac{y_i}{\lambda_i})) \end{aligned}$$

$$D = 2(logL_{Sat} - logL_m) = 2\sum_{i=1}^{n}\left[y_i log(\frac{y_i}{\hat{\lambda}_i}) - (y_i - \hat{\lambda}_i)\right]$$

c)

```
data_smoking <- read.csv("smoking.csv")
data_smoking$location <- factor(data_smoking$location,levels=c(0,1),
labels=c("Home","Work"))
```

```
fit_P <- glm(cigarettes ~ location,
             data = data_smoking,
             family = poisson)
coef(fit_P)
```

```
##  (Intercept) locationWork
##    0.6931472   -1.7917595
```

```
deviance(fit_P)
```

```
## [1] 3.243721
```

1. For the data set where **Home = 0, Work = 1**, $x_i^T\beta$ will be 0.6931472 if at home, then $0.6931472 - 1.7917595 = -1.0986123$ if at work.
2. We use the formula from part b) to calculate each observation, then sum them and multiply by 2 to get the deviance.
   - Obs 1: $3log(3/e^{0.6931472}) - 3 + e^{0.6931472} = 0.216395$
   - Obs 2 and 3: Since the y_i of these two cases is 0, we will get an error if we erroneously input 0 into $log(0)$. However, since the limit of $xlog(x)$ is 0 as x goes towards 0, we will only focus on the latter portion of the formula.
     - $e^{-1.0986123} = 0.3333333$ for both obs 2 and 3
   - Obs 4: $log(1/e^{-1.0986123}) - 1 + e^{-1.0986123} = 0.431945629$
   - Obs 5: $2log(2/e^{0.6931472}) - 2 + e^{0.6931472} = 0$
   - Obs 6: $log(1/e^{0.6931472}) - 1 + e^{0.6931472} = 0.306852838$
3. Summing them up and multiplying by two: $= 3.24372025$, which is equivalent to the deviance from the code above.

Question 2

```
suppressMessages(library(MASS))
suppressMessages(library(faraway))
bacteria$ap <- factor(as.character(bacteria$ap),levels=c("p","a"),
                      labels = c("Placebo","Active"))
str(bacteria)
```

```
## 'data.frame':    220 obs. of  6 variables:
##  $ y   : Factor w/ 2 levels "n","y": 2 2 2 2 2 2 1 2 2 2 ...
##  $ ap  : Factor w/ 2 levels "Placebo","Active": 1 1 1 1 2 2 2 2 2 2 ...
##  $ hilo: Factor w/ 2 levels "hi","lo": 1 1 1 1 1 1 1 1 2 2 ...
##  $ week: int  0 2 4 11 0 2 6 11 0 2 ...
##  $ ID  : Factor w/ 50 levels "X01","X02","X03",..: 1 1 1 1 2 2 2 2 3 3 ...
##  $ trt : Factor w/ 3 levels "placebo","drug",..: 1 1 1 1 3 3 3 3 2 2 ...
```

a)

```
logismodel <- glm(y ~ ap + week + week:ap, family = binomial, data=bacteria)
sumary(logismodel)
```

```
##                Estimate Std. Error z value  Pr(>|z|)
## (Intercept)    2.361584   0.522406  4.5206 6.167e-06
## apActive      -0.630132   0.631785 -0.9974    0.3186
## week          -0.083181   0.076681 -1.0848    0.2780
## apActive:week -0.047143   0.093743 -0.5029    0.6150
##
## n = 220 p = 4
## Deviance = 204.69715 Null Deviance = 217.37591 (Difference = 12.67875)
```

b)

$$log\left(\frac{p_i}{1-p_i}\right) = 2.361584 - 0.630132x_{i1} - 0.083181x_{i2} - 0.047143x_{i2}x_{i1}$$

c)

For the variable 'week', the odds are $exp(-0.083181) = 0.9201846$. This is under when 'ap' is placebo. This means that the odds the bacteria is present will decrease by $1 - 0.9201846 = 0.0798154$ for every 1 unit increase in the variable 'week' when 'ap' is passive.

When 'ap' is active, the odds are $exp(-0.083181 - 0.047143) = 0.877811$. This means that the odds will decrease by $1 - 0.877811 = 0.122189$ for every 1 unit increase in 'week' when 'ap' is active.

For the intercept 2.361584, when week is 0 and ap is placebo, then the interpretation is that the odds of bacteria being present is $exp(2.361584) = 10.60774$.

When week is 0 and ap is active, the interpretation is the odds of bacteria being present is: $exp(2.361584 - 0.630132) = 5.64885$.

I would say there could be a practical interpretation since if the treatment works, then there would be less bacteria in children with middle-ear infection the longer after the treatment. Furthermore, if the drug is the active and not the placebo, then the odds of the bacteria being present becomes lower if the drug is effective.

d)

Standard errors can be overestimated so the z-value is too small and the significance of a predictor could be missed. Instead, we can use the deviance-based likelihood ratio test, which follows a chi-squared distribution. Additionally, we can create a confidence interval that also makes use of the estimate $\pm$ the standard error.

e)

```
# a likelihood ratio test using deviance
difference <- logismodel$null.deviance - logismodel$deviance
pchisq(difference,3,lower.tail = FALSE)
```

```
## [1] 0.005385431
```

Question 3

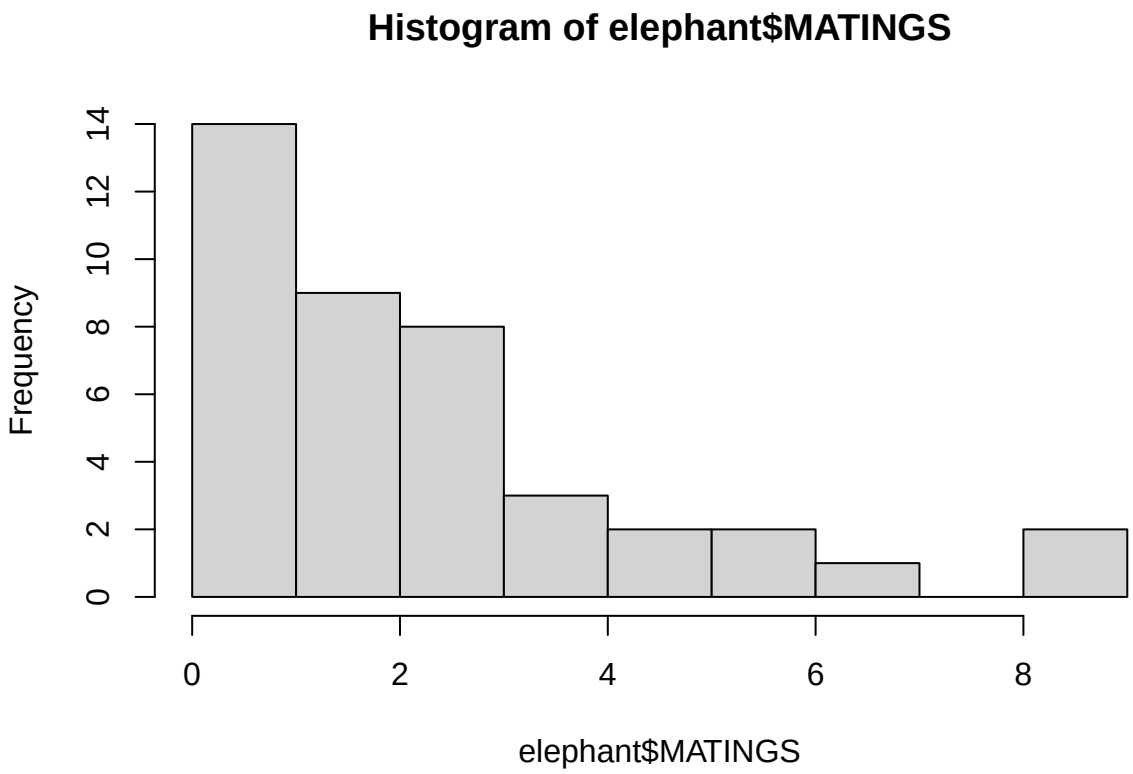
```
elephant <- read.csv("elephant.csv")
str(elephant)
```

```
## '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)

The shape of the histogram suggests that there may be a poisson distribution to the data. There is a bias towards values of 0, which is characteristic of the poisson distribution. The data is also independent at a fixed interval (years).

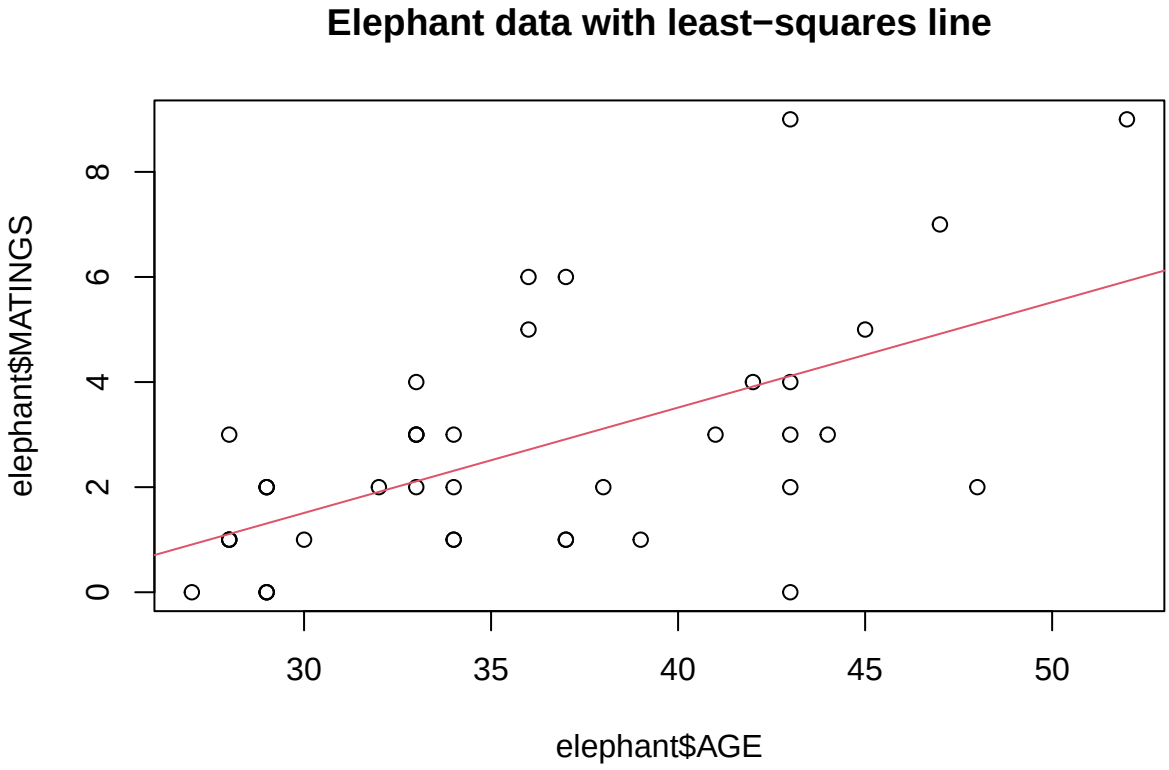
```
hist(elephant$MATINGS)
```



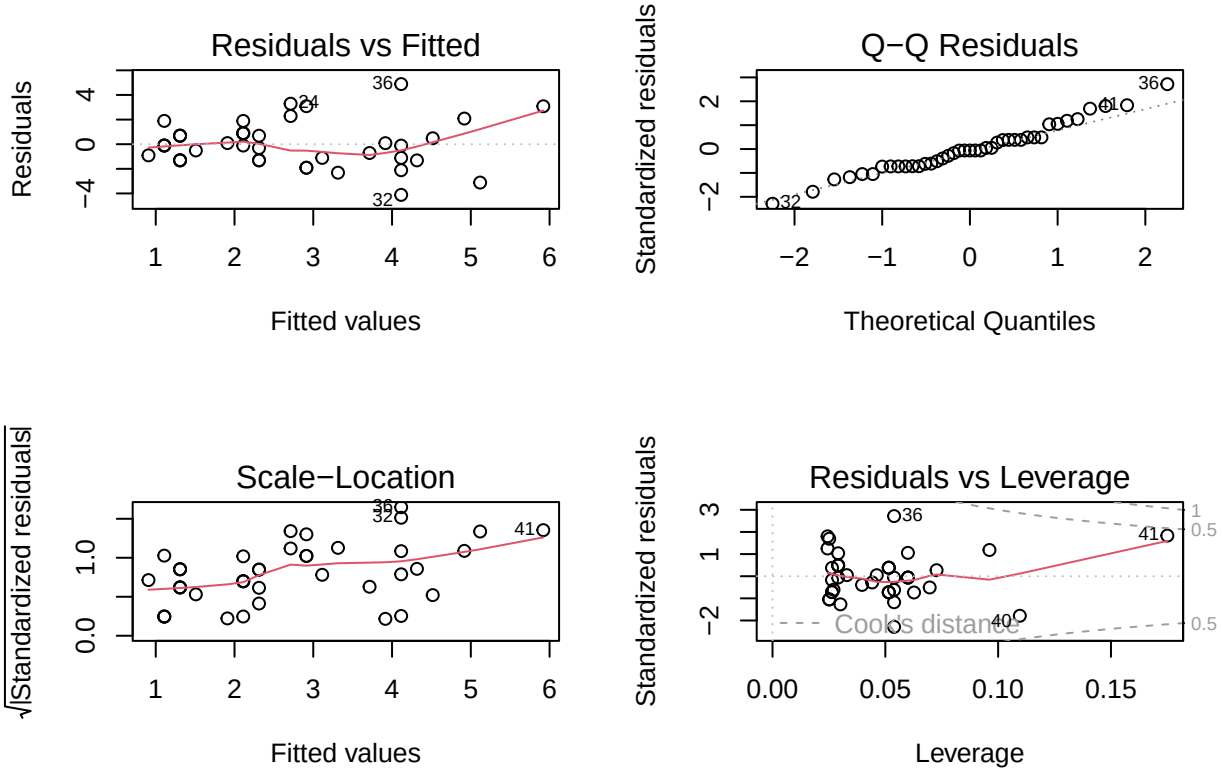
b)

In the residual plots, the Q-Q residual plots show a potential departure from the normal assumption, especially at the tail ends of the data. In the fitted vs residuals plot, residuals are large and the general shape of the data is not linear and does not appear to bounce randomly around the horizontal line. There appears to be a non-random shape to the data residuals.

```
linmod <- lm(MATINGS ~ AGE, data = elephant)
plot(elephant$AGE, elephant$MATINGS, main =
       "Elephant data with least-squares line")
abline(linmod, col= 2)
```

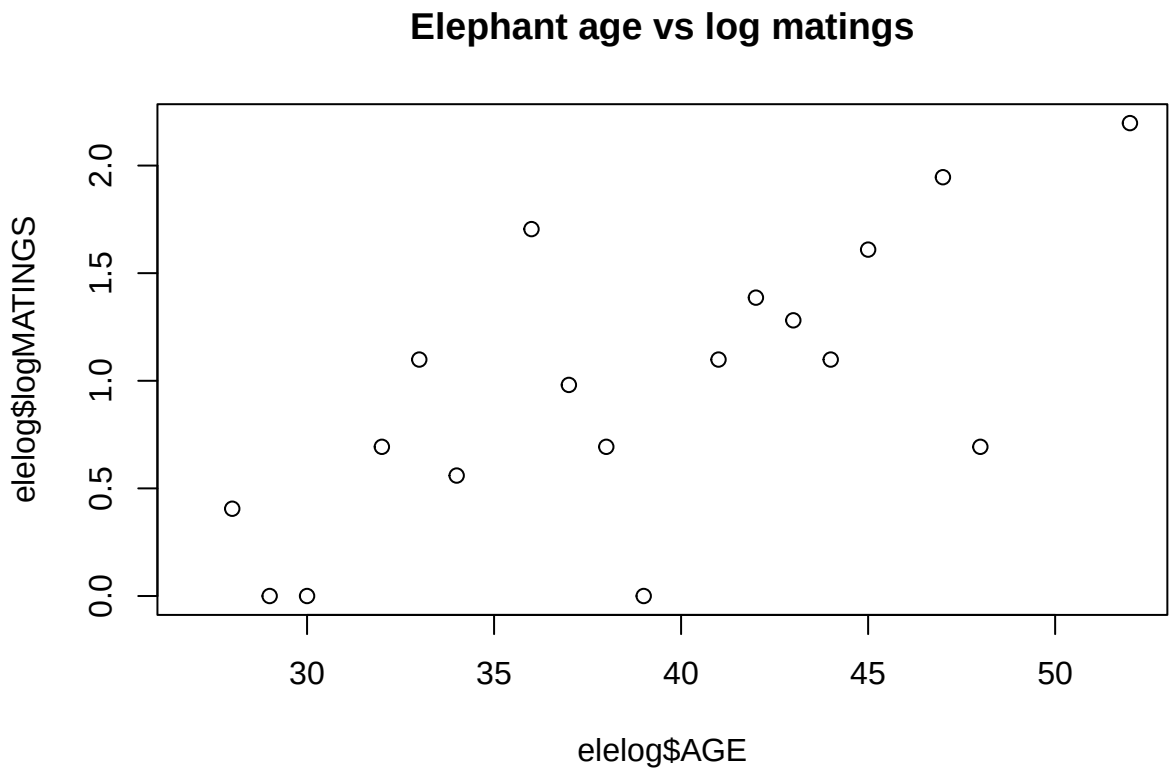



```
par(mfrow=c(2,2))
plot(linmod)
```



c)

```
elelog <- aggregate(MATINGS ~ AGE, data = elephant, FUN = mean)
elelog$logMATINGS <- log(elelog$MATINGS)
plot(elelog$AGE, elelog$logMATINGS, main="Elephant age vs log matings")
```



i)

If the data was exponential, where the number of matings would increase exponentially through age, then this plot would be approximately linear. This is an assumption of poisson data since we see a term that exponentiates the mean λ in the probability distribution.

ii)

There is no evidence of a quadratic trend in this plot since it is largely linear, rather than a U-shape that is indicative of a quadratic relationship.

d)

```
library(faraway)
poismod <- glm(MATINGS ~ AGE, family = poisson, data = elephant)
sumary(poismod)
```

```
##              Estimate Std. Error z value  Pr(>|z|)
## (Intercept) -1.582008   0.544621 -2.9048  0.003675
## AGE          0.068693   0.013746  4.9974 5.812e-07
##
## n = 41 p = 2
## Deviance = 51.01163 Null Deviance = 75.37174 (Difference = 24.36011)
```

Since the link function of the poisson model is $log(\lambda) = \beta_0 + \beta_1 x$, the coefficient for AGE is interpreted by exponentiation. $exp(0.068693) = 1.071107$, which means that for each one unit (year) increase in age, the mean MATINGS increase by 7.1107%.

e)

This confidence means that in 95% of cases, the true effect of a one year increase in AGE should fall between a $exp(0.04167776) = 1.042558 \rightarrow 4.2558\%$ and $exp(0.09563762) = 1.10036 \rightarrow 10.036\%$ increase in mean MATINGS.

```
confint(poismod, "AGE", level = 0.95)
```

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

```
##      2.5 %     97.5 %
## 0.04167776 0.09563762
```

f)

The predicted MATINGS for a 31 year old elephant is 1.728872 with a standard error of 0.2517927. The upper bound is 2.222377 MATINGS and the lower bound is 1.235367.

```
pred <- predict.glm(poismod, newdata = data.frame(AGE = c(31)), type = "response", se.fit = TRUE)
fit <- pred$fit
sefit <- pred$se.fit
z_val <- qnorm(0.975)
data.frame(LB = c(fit-z_val*sefit), UB = c(fit+z_val*sefit))
```

```
##         LB       UB
## 1 1.235367 2.222377
```

g)

According to the test below, the number of matings is significantly related to age. I ran a likelihood ratio test between the model with AGE and without, which showed that the AGE predictor is significant.

```
drop1(poismod,test = "Chi")
```

```
## Single term deletions
##
## Model:
## MATINGS ~ AGE
```

```
##        Df Deviance    AIC   LRT  Pr(>Chi)
## <none>      51.012 156.46
## AGE     1   75.372 178.82 24.36 7.991e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

h)

By the chi-squared test, there does not seem to be a preference for a quadratic model to a linear model. The p-value is not significant at 0.6693 in the full model, and continues to be not significant when the chi-squared test is applied with a p-value at 0.6667.

```
poismodquad <- glm(MATINGS ~ I(AGE**2) + AGE, family = poisson, data = elephant)
sumary(poismodquad)
```

```
##                Estimate  Std. Error z value Pr(>|z|)
## (Intercept) -2.85740597  3.03563826 -0.9413   0.3466
## I(AGE^2)    -0.00085951  0.00201239 -0.4271   0.6693
## AGE          0.13595444  0.15800953  0.8604   0.3896
##
## n = 41 p = 3
## Deviance = 50.82618 Null Deviance = 75.37174 (Difference = 24.54556)
```

```
drop1(poismodquad,test = "Chi")
```

```
## Single term deletions
##
## Model:
## MATINGS ~ I(AGE^2) + AGE
##          Df Deviance    AIC     LRT Pr(>Chi)
## <none>        50.826 158.27
## I(AGE^2)  1   51.012 156.46 0.18544   0.6667
## AGE       1   51.590 157.04 0.76332   0.3823
```

i)

It appears that the model with age as the sole predictor is a better fit model than the one with a quadratic term. The addition of the quadratic term is not significant according to the chi-squared test done in h). Additionally, as shown below, the dispersion is not very different, so adding the quadratic term has a small effect on predicting the response. A single AGE predictor performs better at predicting the log(mean) of the response than the addition of a quadratic term.

```
sum(residuals(poismod, type = "pearson")**2) / poismod$df.residual
```

```
## [1] 1.157334
```

```
# dispersion for the linear predictor
sum(residuals(poismodquad, type = "pearson")**2) / poismodquad$df.residual
```

```
## [1] 1.175194
```

```
# dispersion for the quadratic and linear predictor
```

j)

```
quaspoismod <- glm(MATINGS ~ AGE, family = quasipoisson, data = elephant)
sumary(quaspoismod)
```

```
##              Estimate Std. Error t value  Pr(>|t|)
## (Intercept) -1.582008   0.585901 -2.7001    0.0102
## AGE          0.068693   0.014788  4.6453 3.809e-05
##
## Dispersion parameter = 1.15733
## n = 41 p = 2
## Deviance = 51.01163 Null Deviance = 75.37174 (Difference = 24.36011)
```

i)

The estimates coefficients did not change from the original poisson model

ii)

The standard errors did become larger for both the intercept and coefficient.

iii)

The estimated dispersion parameter is 1.157334

```
dp <- sum(residuals(quaspoismod, type = "pearson")**2) / quaspoismod$df.residual
dp
```

```
## [1] 1.157334
```

iv)

When adjusting for overdispersion, you are less likely to obtain a significant result when testing coefficients because you increase standard errors and thus make z-values smaller, which leads to a lesser probability that coefficients are significant.

Question 4

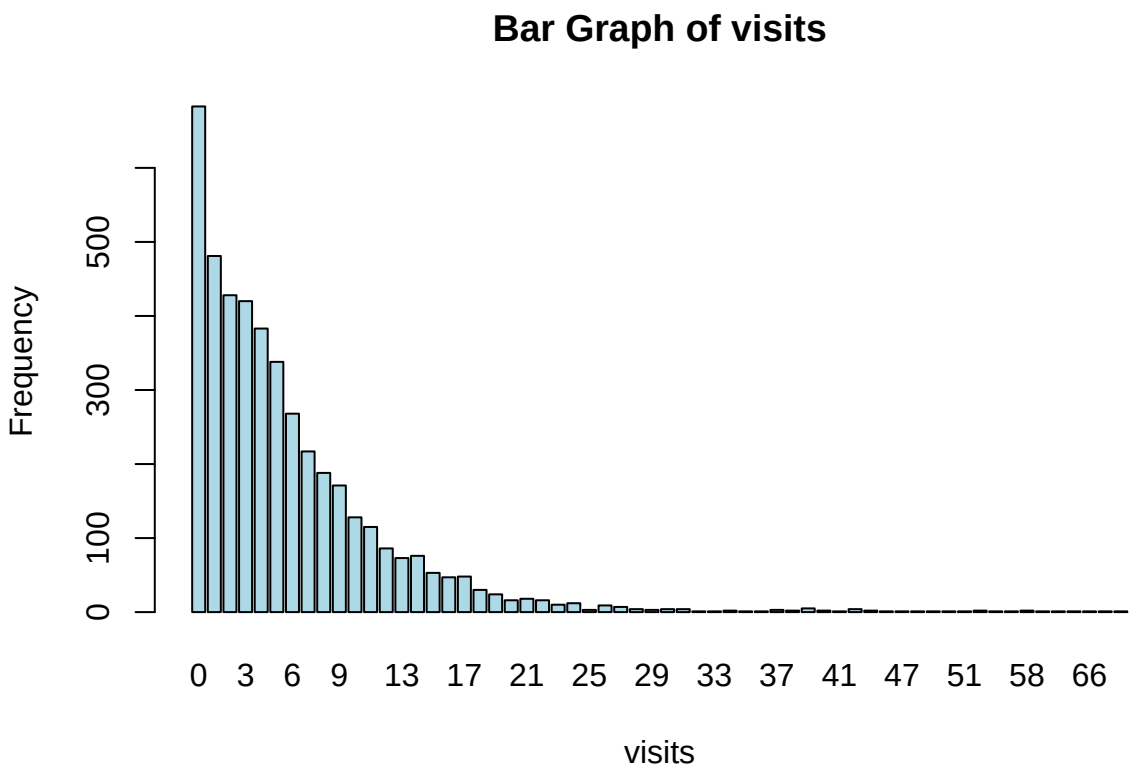
```
suppressMessages(library(AER))
data(NMES1988)
str(NMES1988)
```

```
## 'data.frame':    4406 obs. of  19 variables:
##  $ visits   : int  5 1 13 16 3 17 9 3 1 0 ...
##  $ nvisits  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ ovisits  : int  0 2 0 5 0 0 0 0 0 0 ...
##  $ novisits : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ emergency: int  0 2 3 1 0 0 0 0 0 0 ...
##  $ hospital : int  1 0 3 1 0 0 0 0 0 0 ...
##  $ health   : Factor w/ 3 levels "poor","average",..: 2 2 1 1 2 1 2 2 2 2 ...
##   ..- attr(*, "contrasts")= num [1:3, 1:2] 1 0 0 0 0 1
##   .. ..- attr(*, "dimnames")=List of 2
##   .. .. ..$ : chr [1:3] "poor" "average" "excellent"
##   .. .. ..$ : chr [1:2] "poor" "excellent"
##  $ chronic  : int  2 2 4 2 2 5 0 0 0 0 ...
##  $ adl      : Factor w/ 2 levels "normal","limited": 1 1 2 2 2 2 1 1 1 1 ...
##  $ region   : Factor w/ 4 levels "northeast","midwest",..: 4 4 4 4 4 4 2 2 2 2 ...
##   ..- attr(*, "contrasts")= num [1:4, 1:3] 1 0 0 0 0 1 0 0 0 0 ...
##   .. ..- attr(*, "dimnames")=List of 2
##   .. .. ..$ : chr [1:4] "northeast" "midwest" "west" "other"
##   .. .. ..$ : chr [1:3] "northeast" "midwest" "west"
##  $ age      : num  6.9 7.4 6.6 7.6 7.9 6.6 7.5 8.7 7.3 7.8 ...
##  $ afam     : Factor w/ 2 levels "no","yes": 2 1 2 1 1 1 1 1 1 1 ...
##  $ gender   : Factor w/ 2 levels "female","male": 2 1 1 2 1 1 1 1 1 1 ...
##  $ married  : Factor w/ 2 levels "no","yes": 2 2 1 2 2 1 1 1 1 1 ...
##  $ school   : int  6 10 10 3 6 7 8 8 8 8 ...
##  $ income   : num  2.881 2.748 0.653 0.659 0.659 ...
##  $ employed : Factor w/ 2 levels "no","yes": 2 1 1 1 1 1 1 1 1 1 ...
##  $ insurance: Factor w/ 2 levels "no","yes": 2 2 1 2 2 1 2 2 2 2 ...
##  $ medicaid : Factor w/ 2 levels "no","yes": 1 1 2 1 1 2 1 1 1 1 ...
```

a)

The expected number of 0's in a poisson distribution is $Ne^{-\lambda}$, which in this case, depending on the value of lambda, could easily be less than 500 if lambda exceeds 3, which graphically, appears to be the case. Therefore, for there to be more than 500 cases of 0 is abnormal and thus we should use a zero-inflated model.

```
suppressMessages(library(pscl))
barplot(table(NMES1988$visits), main = "Bar Graph of visits",
        xlab = "visits", ylab = "Frequency",
        col = "lightblue", border = "black")
```



b)

```
modz <- zeroinfl(visits ~ chronic + health + insurance | chronic + insurance, data=NMES1988)
summary(modz)
```

```
##
## Call:
## zeroinfl(formula = visits ~ chronic + health + insurance | chronic +
##     insurance, data = NMES1988)
##
## Pearson residuals:
##     Min      1Q  Median      3Q     Max
## -3.9221 -1.2195 -0.4316  0.5598 24.1031
##
## Count model coefficients (poisson with log link):
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)      1.55878    0.01762  88.448   <2e-16 ***
## chronic          0.11868    0.00462  25.691   <2e-16 ***
## healthpoor       0.29470    0.01729  17.043   <2e-16 ***
## healthexcellent -0.30482    0.03115  -9.786   <2e-16 ***
## insuranceyes     0.14467    0.01631   8.870   <2e-16 ***
##
## Zero-inflation model coefficients (binomial with logit link):
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -0.37426    0.09213  -4.062 4.86e-05 ***
## chronic      -0.56112    0.04334 -12.948  < 2e-16 ***
## insuranceyes -0.88314    0.09464  -9.332  < 2e-16 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Number of iterations in BFGS optimization: 14
## Log-likelihood: -1.651e+04 on 8 Df
```

- The coefficient of chronic in the poisson part of the model is 0.11868, which when exponentiated, is $exp(0.11868) = 1.12601$. This means that for every 1 unit increase in chronic, the mean visits increase by 12.601%.
- The coefficient of poor health in the poisson part of the model is 0.29470, which when exponentiated, is $exp(0.29470) = 1.342723$. This means that when health is poor, the mean visits increase by 34.2723%
- The intercept in the logistic part of the model is -0.37426, which when exponentiated, is $exp(-0.37426) = 0.6877981$. This means that baseline odds of never visiting is 0.6877981 when chronic is zero and insurance is no.
- The coefficient of insurance in the logistic part of the model is -0.88314, which when exponentiated, is $exp(-0.88314) = 0.4134825$. This means that when insurance is yes, then the odds of never visiting decrease by $1 - 0.4134825 = 58.65175\%$

c)

Under the specified components, the value of the intercept and coefficients is $log(pi/1 - pi) = -0.37426 - 0.56112 * 2 - 0.88314 = -2.37964$. Exponentiating to get the odds, $exp(-2.37964) = 0.0925839$, you can now find the probability by doing some basic algebra. $0.0925839/(1 + 0.0925839) = 0.08473848$, which represents the probability of never going to the doctor.

```
predict(modz, type = "zero", newdata = data.frame(chronic = 2,
                                                   insurance = "yes", health = "average"))
```

```
##          1
## 0.08473874
```

d)

The poisson part of the model is $log(\lambda) = 1.55878 + 0.11868 * 4 + 0.29470 * 1 + 0.14467 * 0 = 2.3282$. Exponentiating to get the mean, $exp(2.3282) = 10.25946$. The person is expected to visit the doctor just over 10 times. Then, getting the logistic part of the model: $-0.37426 - 0.56112 * 4 = -2.61874$, $exp(-2.61874) = 0.07289465$, probability of never going $= 0.07289465/(1 + 0.07289465) = 0.06794204$. If the person is in the population that does visit the doctor, then their probability is $1 - 0.06794204 = 0.932058$. Therefore, the probability of exactly 5 visits is the probability of visiting the doctor multiplied by the poisson distribution with lambda as 10.25946 and x as 5, which is $0.932058 * 0.03318 = 0.03092568$. Therefore the probability of exactly 5 visits with all of the coefficients is 3.092568%.

```
lambda <- predict(modz, type = "count",
                  newdata = data.frame(chronic = 4,
                                       insurance = "no", health = "poor"))
probab <- predict(modz, type = "zero",
                  newdata = data.frame(chronic = 4,
                                       insurance = "no", health = "poor"))
(1-probab)*dpois(5,lambda)
```

```
##          1
## 0.03092156
```

Question 5

```
suppressMessages(library(faraway))
str(UCBAdmissions)
```

```
##  'table' num [1:2, 1:2, 1:6] 512 313 89 19 353 207 17 8 120 205 ...
##  - attr(*, "dimnames")=List of 3
##   ..$ Admit : chr [1:2] "Admitted" "Rejected"
##   ..$ Gender: chr [1:2] "Male" "Female"
##   ..$ Dept  : chr [1:6] "A" "B" "C" "D" ...
```

a)

The Simpson's paradox is when data from two or more groups appear to trend one way, but can either stop trending or reverse trends when combined. This is illustrated below with the 7 contingency tables built: the first from all the data, then the others from specific departments. We can see that although the total acceptance rate for males is greater than females, across nearly every department listed the female acceptance rate was higher.

```
UCBAdmissionsdataframe <- data.frame(UCBAdmissions)
# percentage of males admitted is greater than females
(ct <- xtabs(Freq ~ Admit+Gender, UCBAdmissionsdataframe))
```

```
##           Gender
## Admit      Male Female
##   Admitted 1198    557
##   Rejected 1493   1278
```

```
# 89 accepted vs 19 rejected for females in department A = 82.4% accept rate
# 512 accepted vs 313 rejected for males in department A = 62.1% accept rate
(cta <- xtabs(Freq ~ Admit+Gender, UCBAdmissionsdataframe,
              subset = (Dept == "A")))
```

```
##           Gender
## Admit      Male Female
##   Admitted  512     89
##   Rejected  313     19
```

```
# 68% acceptance for females in department B
# 63% acceptance for males in department B
(ctb <- xtabs(Freq ~ Admit+Gender, UCBAdmissionsdataframe,
              subset = (Dept == "B")))
```

```
##           Gender
## Admit      Male Female
##   Admitted  353     17
##   Rejected  207      8
```

```r
# 34% acceptance for females in department C
# 37% acceptance for males in department C
(ctc <- xtabs(Freq ~ Admit+Gender, UCBAdmissionsdataframe,
              subset = (Dept == "C")))
```

```
##           Gender
## Admit      Male Female
##   Admitted  120    202
##   Rejected  205    391
```

```r
# 35% acceptance for females in department D
# 33% acceptance for males in department D
(ctd <- xtabs(Freq ~ Admit+Gender, UCBAdmissionsdataframe,
              subset = (Dept == "D")))
```

```
##           Gender
## Admit      Male Female
##   Admitted  138    131
##   Rejected  279    244
```

```r
# 24% acceptance for females in department E
# 28% acceptance for males in department E
(cte <- xtabs(Freq ~ Admit+Gender, UCBAdmissionsdataframe,
              subset = (Dept == "E")))
```

```
##           Gender
## Admit      Male Female
##   Admitted   53     94
##   Rejected  138    299
```

```r
# 7.1% acceptance for females in department F
# 5.9% acceptance for males in department F
(ctf <- xtabs(Freq ~ Admit+Gender, UCBAdmissionsdataframe,
              subset = (Dept == "F")))
```

```
##           Gender
## Admit      Male Female
##   Admitted   22     24
##   Rejected  351    317
```

b)

```r
# creating a poisson model of all combinations
modi <- glm(Freq ~ Admit*Gender*Dept,
            UCBAdmissionsdataframe, family = poisson)

# the three way interaction term should not be dropped
drop1(modi, test="Chi")
```

```
## Single term deletions
##
## Model:
## Freq ~ Admit * Gender * Dept
##                   Df Deviance    AIC    LRT Pr(>Chi)
## <none>                  0.000 207.06
## Admit:Gender:Dept  5   20.204 217.26 20.204 0.001144 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# creating a poisson model with 2 way interactions
modu <- glm(Freq ~ (Admit+Gender+Dept)^2,
            UCBAdmissionsdataframe, family=poisson)
# admit: gender can be dropped
drop1(modu, test="Chi")
```

```
## Single term deletions
##
## Model:
## Freq ~ (Admit + Gender + Dept)^2
##               Df Deviance     AIC     LRT Pr(>Chi)
## <none>              20.20  217.26
## Admit:Gender  1    21.74  216.80    1.53   0.2159
## Admit:Dept    5   783.61  970.67  763.40   <2e-16 ***
## Gender:Dept   5  1148.90 1335.96 1128.70   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# creating a poisson model with the terms alone
modt <- glm(Freq ~ Admit + Gender + Dept,
            UCBAdmissionsdataframe, family=poisson)
# all predictors alone are significant
drop1(modt, test="Chi")
```

```
## Single term deletions
##
## Model:
## Freq ~ Admit + Gender + Dept
##        Df Deviance    AIC    LRT  Pr(>Chi)
## <none>      2097.7 2272.7
## Admit   1   2327.7 2500.8 230.03 < 2.2e-16 ***
## Gender  1   2260.6 2433.6 162.87 < 2.2e-16 ***
## Dept    5   2257.2 2422.2 159.52 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the three tests done above, the optimal model is as follows

```
modfinal <- glm(Freq ~ Admit:Gender:Dept + Admit:Dept + Gender:Dept +
                  Gender + Admit + Dept, UCBAdmissionsdataframe, family = poisson)
```

c)

```
ybin <- matrix(UCBAdmissionsdataframe$Freq, ncol=2)
binmod <- glm(ybin ~ Gender*Dept,
              UCBAdmissionsdataframe[1:12,],
              family = binomial)
drop1(binmod,test="Chi")
```

```
## Single term deletions
##
## Model:
## ybin ~ Gender * Dept
##             Df Deviance     AIC    LRT  Pr(>Chi)
## <none>           476.34  555.97
## Gender:Dept  2  1090.51 1166.14 614.17 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
binmodr <- glm(ybin ~ Gender + Dept,
               UCBAdmissionsdataframe[1:12,],
               family = binomial)
drop1(binmodr,test="Chi")
```

```
## Single term deletions
##
## Model:
## ybin ~ Gender + Dept
##        Df Deviance    AIC     LRT  Pr(>Chi)
## <none>      1090.5 1166.1
## Gender  1   1383.7 1457.4 293.217 < 2.2e-16 ***
## Dept    2   1140.3 1212.0  49.815 1.523e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The ideal model using the binomial is:

```
binmodfinal <- glm(ybin~ Gender + Dept + Gender:Dept,
                   UCBAdmissionsdataframe[1:12,],
                   family = binomial)
```

This is equal to the poisson regression model because we have all the terms in the previous model minus the ones with Admit in them. In the binomial model, a success corresponds to the level "Admitted" in Admit. In other words, the main effects of the relationship continue in both the binomial and poisson models.

Question 6

```
suppressMessages(library(faraway))
str(melanoma)
```

```
## 'data.frame':    12 obs. of  3 variables:
##  $ count: num  22 16 19 11 2 54 33 17 10 115 ...
##  $ tumor: Factor w/ 4 levels "freckle","indeterminate",..: 1 4 3 2 1 4 3 2 1 4 ...
##  $ site : Factor w/ 3 levels "extremity","head",..: 2 2 2 2 3 3 3 3 1 1 ...
```

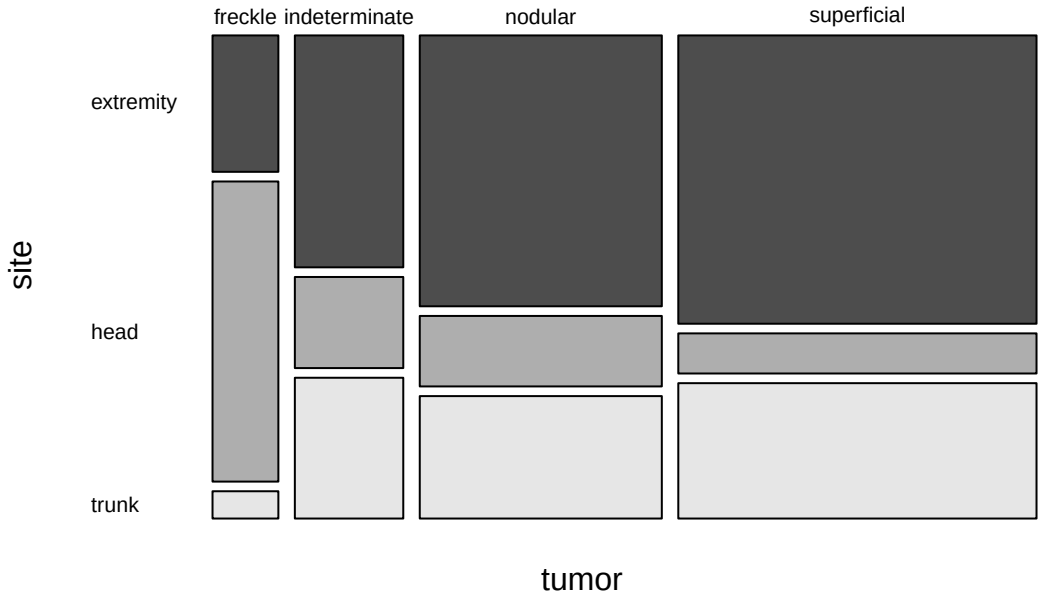
a)

Since the horizontal lines of the mosaic plot do not line up, then there is evidence of dependence between tumor and site.

```
y <- melanoma$count
tumor <- melanoma$tumor
site <- melanoma$site
melanomadata <- data.frame(y,tumor,site)
(mel <- xtabs(y ~ tumor+site))
```

```
##                site
## tumor           extremity head trunk
##   freckle              10   22     2
##   indeterminate        28   11    17
##   nodular              73   19    33
##   superficial         115   16    54
```

```
mosaicplot(mel,color=TRUE, main=NULL, las=1)
```



b)

According to the chi-squared test using `summary`, we will reject the hypothesis that the variables are independent. Thus we will conclude that the type of tumor is related to the site of the tumor.

```r
summary(mel)
```

```
## Call: xtabs(formula = y ~ tumor + site)
## Number of cases in table: 400
## Number of factors: 2
## Test for independence of all factors:
##  Chisq = 65.81, df = 6, p-value = 2.943e-12
```

Alternatively, we can use `chisq.test` on the table. This gives us the exact same result as the `summary` function.

```r
chisq.test(mel, correct=TRUE)
```

```
##
##  Pearson's Chi-squared test
##
## data:  mel
## X-squared = 65.813, df = 6, p-value = 2.943e-12
```

c)

The below code shows that the null model is not a good fit, thus the outcomes occurring at the same rate is not fit well to the response.

```r
contpoismod <- glm(count~tumor+site, data = melanoma, family = poisson)
sumary(contpoismod)
```

```
##                    Estimate Std. Error z value  Pr(>|z|)
## (Intercept)         2.95543    0.17702 16.6955 < 2.2e-16
## tumorindeterminate  0.49899    0.21741  2.2951   0.02173
## tumornodular        1.30195    0.19342  6.7312 1.682e-11
## tumorsuperficial    1.69400    0.18659  9.0786 < 2.2e-16
## sitehead           -1.20103    0.13831 -8.6834 < 2.2e-16
## sitetrunk          -0.75710    0.11772 -6.4312 1.266e-10
##
## n = 12 p = 6
## Deviance = 51.79501 Null Deviance = 295.20301 (Difference = 243.40799)
```

```r
pchisq(contpoismod$null.deviance,6,lower=FALSE) # null model is not a good fit
```

```
## [1] 8.720021e-61
```

The deviance of the fitted model is large for a chi-squared distribution with 6 degrees of freedom, showing lack of fit.

```r
pchisq(contpoismod$deviance, 6, lower=FALSE)
```

```
## [1] 2.050453e-09
```

However, the predictors are significant when considered independently, which implies that the lack of fit comes from a missing term. This missing term is the interaction term, but this would make the model the saturated model. In this case, as before, we can reject the hypothesis that the model with no interaction is a good fit and conclude that the tumor type is related to the site.

```
drop1(contpoismod, test="Chi")
```

```
## Single term deletions
##
## Model:
## count ~ tumor + site
##        Df Deviance    AIC     LRT  Pr(>Chi)
## <none>      51.795 122.91
## tumor   3  196.901 262.01 145.106 < 2.2e-16 ***
## site    2  150.097 217.21  98.302 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

d

- The relationship between freckle and the site have very large deviance residuals. This suggests that the freckle tumor type is not well fit to any specific site.
- The relationship between superficial and head has a very large deviance residual. This may suggest that the superficial tumor type does not appear often on the head.

```
devres <- residuals(contpoismod, type = "deviance")
(devcont <- xtabs(devres ~ tumor + site))
```

```
##                site
## tumor             extremity        head       trunk
##   freckle       -2.31583297  5.13537787 -2.82829426
##   indeterminate -0.66016102  0.46798432  0.54787007
##   nodular        0.28104581 -0.49711084 -0.02173229
##   superficial    1.00813975 -3.04533605  0.69899703
```