

Module 09 HW

Diana McSpadden

10/27/2020

Stat 6021: Homework Set 9

H. Diana McSpadden

UID: hdm52

Date: 10/30/2020

Attended group with: Bernhardt, Cheu, Nam, Hiatt, Newman

Question 1. In an experiment testing the effect of a toxic substance, ...

1,500 insects were divided at random into six groups of 250 each.

The insects in each group were exposed to a fixed dose of the toxic substance. A day later, the insects were observed. Death from exposure was recorded as 1, and survival 0.

The file "insects.txt" contains the data from the experiment.

- The first column denotes the dose level on a logscale, x ;
- the 2nd column denotes the sample size for each group, n_i
- ; and the 3rd column denotes the number of insects that died in the group, y_i

You will probably want to read the data in and give appropriate names for the columns.

```
library(dplyr)

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##   filter, lag

## The following objects are masked from 'package:base':
##
##   intersect, setdiff, setequal, union

insectsData <- read.table("insects.txt", header=FALSE, sep="")
```

```

insectsData <- insectsData %>% rename(logdose = V1, size = V2, died = V3)

#names(insectsData)

head(insectsData)

##   logdose size died
## 1      1   250   28
## 2      2   250   53
## 3      3   250   93
## 4      4   250  126
## 5      5   250  172
## 6      6   250  197

attach(insectsData)

#knowns
n <- 1500
groupn <- 250
groups <- 6

```

Q1 (a) Plot the sample log odds against the log dose, x.

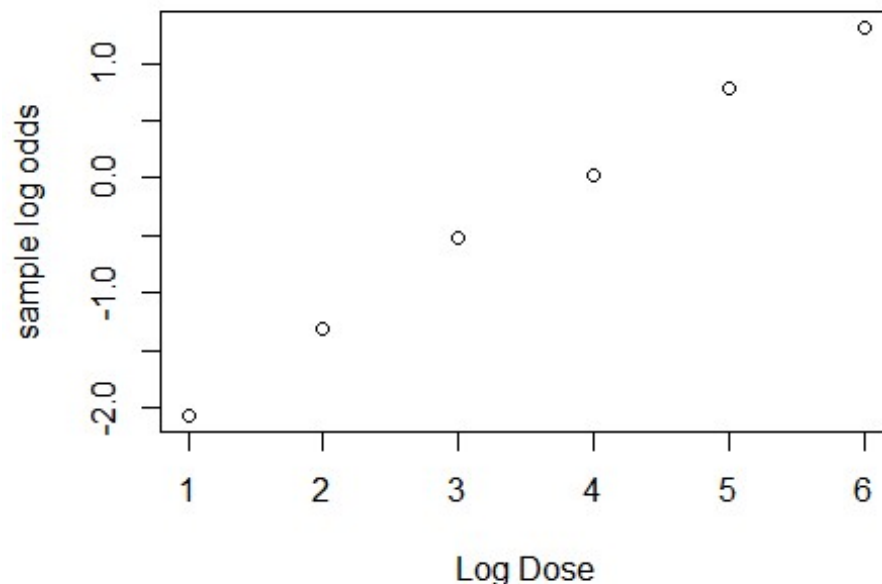
Does this plot suggest fitting a logistic regression model to be appropriate?

```

prop<-died/size

plot(logdose, log(prop/(1-prop)), xlab= "Log Dose", ylab="sample log odds")

```



Answer Q1a: Yes, this plot looks perfect for a logistic regression model because the relationship is almost perfectly linear.

Q1 (b) Use R to fit the logistic regression model, and write the estimated logistic regression equation.

```
modelInsects <- glm(prop~logdose, family="binomial", weights=size) # fit
model. The weight is the sample/group size
summary(modelInsects)
```

```
##
## Call:
## glm(formula = prop ~ logdose, family = "binomial", weights = size)
##
## Deviance Residuals:
##      1       2       3       4       5       6
## -0.5092 -0.1115  0.7461 -0.2869  0.4744 -0.5599
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.64367    0.15610  -16.93  <2e-16 ***
## logdose      0.67399    0.03911   17.23  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
```

```
##      Null deviance: 383.0695  on 5  degrees of freedom
## Residual deviance:   1.4491  on 4  degrees of freedom
## AIC: 39.358
##
## Number of Fisher Scoring iterations: 3
```

Answer Q1b

$$E(y) = \exp(-2.64 + (0.67 * \text{logdose})) / (1 + \exp(-2.64 + (0.67 * \text{logdose})))$$

or

$$\log(\pi / (1 - \pi)) = -2.64 + (0.67 * \text{logdose})$$

Q1 (c) Interpret the estimated coefficient β^1 in context.

```
exp(0.67)
```

```
## [1] 1.954237
```

The predicted odds of an insect dying is multiplied by $\exp(0.67) \approx 1.95$ for a 1 unit increase in log dose.

or

The predicted odds ratio of an insect dying for a one unit increase in log dose is 1.95.

Q1 (d) What are the estimated odds of death at log dose level $x = 2$?

```
x_logdose <- 2
```

```
oddsOfDeath <- exp(-2.64 + (0.67 * x_logdose))
print(paste("Odds of Death: ", oddsOfDeath))
```

```
## [1] "Odds of Death: 0.272531793034013"
```

Q1 (e) Predict the probability of death from exposure at log dose level $x = 2$.

```
#probOfDeathStart <- exp(-2.64 + (0.67 * x_Logdose)) / (1 + exp(-2.64 +
(0.67 * x_Logdose)))
```

```
probDeath <- oddsOfDeath / (1 + oddsOfDeath)
print(paste("Probability of Death: ", probDeath))
```

```
## [1] "Probability of Death: 0.214165016957441"
```

Q1 (f) Carry out both the Pearson's χ^2 and deviance goodness of fit tests to check the fit ...

of this logistic regression model. Clearly state the null and alternative hypotheses, test statistic, and conclusion.

H0: The model fits well Ha: The model does not fit well

First, Pearson's

```
#Pearsons
pearson<-residuals(modelInsects,type="pearson")
X2<-sum(pearson^2)
X2 # test statistic is 1.45

## [1] 1.451786

print(paste("Pearson test statistic",X2))

## [1] "Pearson test statistic 1.4517864549869"

n= 6
p=2

pValue <- 1-pchisq(X2,n-p) # df = n-p, this gives p value
print(paste("pvalue: ",pValue))

## [1] "pvalue: 0.835146184673398"
```

Second, Deviance GOF

```
1 - pchisq(modelInsects$deviance, n-p)

## [1] 0.8356191
```

Answer q1f: We fail reject the null hypothesis, the model is a good fit because p value > .05 for both GOF tests. The model fits well which agrees with our analysis of the plot of sample log odds against log dose.

Question 2. (No R needed) A health clinic sent ...

fliers to its clients to encourage everyone to get a flu shot. In a follow-up study, 159 elderly clients were randomly selected and asked if they received a flu shot.

A client who received a flu shot was coded $y = 1$, and a client who did not receive a flu shot was coded $y = 0$.

Data were also collected on their

- age, x_1 ,
- health awareness rating on a 0-100 scale (higher values indicate greater awareness), x_2 ,
- and gender, x_3 , where males were coded $x_3 = 1$ and females were coded $x_3 = 0$.

A first order logistic regression model was fitted and the output is displayed below. ...

Q2 (a) Interpret the estimated coefficient for x3, gender, in context.

Work on q2a:

$B_3 = 0.43397$.

```
exp(0.43397)
```

```
## [1] 1.543373
```

Answer q2a: When age, and awareness rating are held constant, the odds of an elderly male patient receiving the flu shot is $\exp(0.43397)$, or 1.543373 times the odds of an elderly female patient receiving the flu shot.

Q2 (b) Conduct the Wald test for B3. ...

State the null and alternative hypotheses, calculate the test statistic, and make a conclusion in context.

Work on q2b

H_0 : The coef for gender, B_3 , equals 0 H_a : The coef for gender, B_3 , does not equal 0

Wald test statistic: $B_j\text{-hat} - 0 / \text{se}(B_j\text{-hat})$

```
Bjhat <- 0.43397
```

```
seBjhat <- 0.52179
```

```
wTestStatistic <- Bjhat / seBjhat
```

```
print(paste("Wald test statistic for B3: ", wTestStatistic))
```

```
## [1] "Wald test statistic for B3: 0.831694743095882"
```

```
#compare to pnorm
```

```
wPValue <- (1-pnorm(wTestStatistic))*2
```

```
print(paste("Wald p value for B3: ", wPValue))
```

```
## [1] "Wald p value for B3: 0.405581268667017"
```

Answer q2b: We do not reject the null hypothesis, gender is not statistically significant in our model.

Q2 (c) Calculate a 95% confidence interval for B3, and interpret the interval in context.

Work on q2c

```
zValue <- 1.96 # for 95% CI
```

```
CILOW <- Bjhat - (zValue * seBjhat)
```

```

CIHigh <- Bjhat + (zValue * seBjhat)

print(paste("The 95% CI for B3 is: ", CIHigh, " - ", CIHigh))
## [1] "The 95% CI for B3 is:  -0.5887384  -  1.4566784"

print(paste("Low odds ratio: ", exp(CILow)))
## [1] "Low odds ratio:  0.555027065365734"

print(paste("High odds ratio: ", exp(CIHigh)))
## [1] "High odds ratio:  4.2916805807803"

```

When age, and awareness rating are held constant, the odds of an elderly male patient receiving the flu shot is between -0.589 times and 1.46 time the odds of a an elderly female receiving the flu shot. 0 is within the CI, which supports the conclusion that gender is not a significant predictor.

Q2 (d) Comment on whether your conclusions from parts 2b and 2c are consistent.

Answer q2d: 0 is within the CI, which supports the conclusion that gender is not a significant predictor.

Q2 (e) Suppose you want to drop the coefficients for age and gender, B1 and B3.

A logistic regression model for just awareness was fitted, and the output is shown below. ...

Carry out the appropriate hypothesis test to see if the coefficients for age and gender can be dropped.

Work on q2e

This is analogous to a partial F test in linear regression, but comparing deviance of the partial and full model.

H0: The coefs B1 and B3 are equal to 0. Ha: At least one of the coef B1 or B3 is not equal to 0.

```

partialDeviance <- 113.2

fullDeviance <- 105.09

df <- 2 # considering removing 2 predictors

deltaG2 <- partialDeviance - fullDeviance

1-pchisq(deltaG2, df)

```

```
## [1] 0.01733548  
qchisq(.025, df) # critical value  
## [1] 0.05063562
```

Conclusion q2e: We reject the null hypothesis because p-value, 0.017, is < .05. We must use the full model when comparing to the awareness-only model.

Q2 (f) Based on your conclusion in question 2e, what are the estimated odds of a client receiving the ...

flu shot if the client is 70 years old, has a health awareness rating of 65, and is male?

What is the estimated probability of this client receiving the flu shot?

Work on q2f

Use the reduced model with only B2

```
x_age <- 70  
x_gender <- 1  
x_awareness <- 65  
  
linearPart <- -1.17716 + (0.07279 * x_age) - (0.09899 * x_awareness) +  
(0.43397 * x_gender)  
print(paste("log odds: ", linearPart))  
## [1] "log odds: -2.08224"  
  
print(paste("odds: ", exp(linearPart)))  
## [1] "odds: 0.124650681714281"  
  
Ey1 <- exp(linearPart) / (1 + exp(linearPart))  
Ey1 #prob  
## [1] 0.110835  
  
print(paste("prob: ", Ey1))  
## [1] "prob: 0.110835020812221"
```

Answer q2f: I estimate the odds as .125 that this patient received a flu shot.

Q2 (g) Based on your conclusion in question 2e, ...

interpret the estimated coefficient for age, in context.

```
exp(0.07279)  
## [1] 1.075505
```


Interpretation of the coefficient for age: When awareness and gender are held constant, each increase in 1 year of age multiplies the predicted odds of receiving a flu shot by 1.076

What does this interpretation tell us about the estimated probability of a client receiving the flu shot if the client is 71 years old, has a health awareness rating of 65, and is male, compared to the estimated probability found for a similar client who is a year younger?

Answer q2g:

I would expect the odds of the similar patient, except 1 year older than 70 to be $12.5 + (12.5 * .076)\% \approx 13.4\%$

And to compare probabilities we exam the odds ratio $= \exp(\hat{B}) = 1.075505$ and is the increase in probability of flu shot, when gender and awareness are constant, of a 1 year increase in age.

And to confirm

```
x_age <- 71
x_gender <- 1
x_awareness <- 65

linearPart71 <- -1.17716 + (0.07279 * x_age) - (0.09899 * x_awareness) +
(0.43397 * x_gender)
exp(linearPart71)

## [1] 0.1340624
```