GLMs Homework 6

Paul Harmon

March 30, 2017

## Problem 1

1. The data frame is displayed below.

Defendant <- c("W","W","W","W","B","B","B","B")  
Victim <- c("W","W","B","B","W","W","B","B")  
Frequency <- c(19, 151-19,0,9,11,63-11,6,103-6)  
Penalty <- c("Y","N","Y","N","Y","N","Y","N")  
   
death <- data.frame(cbind(Defendant,Victim,Penalty,Frequency))   
death

## Defendant Victim Penalty Frequency  
## 1 W W Y 19  
## 2 W W N 132  
## 3 W B Y 0  
## 4 W B N 9  
## 5 B W Y 11  
## 6 B W N 52  
## 7 B B Y 6  
## 8 B B N 97

Def <- c("W", "W", "B", "B")  
Vic <- c("W", "B", "W", "B")  
Freq <- c(19, 0, 11, 6)  
Total <- c(151, 9, 63,103)  
  
death.logit <- data.frame(cbind(Def, Vic, Freq, Total))  
death.logit

## Def Vic Freq Total  
## 1 W W 19 151  
## 2 W B 0 9  
## 3 B W 11 63  
## 4 B B 6 103

1. **Construct partial tables needed to study the conditional association between the defendant's race and the death penalty verdict, conditional on the victim's race. Find and interpret the sample conditional odds ratios, adding 0.5 to each cell to reduce the impact of the zero cell count.** The conditional partial table is given below. We are conditioning on the victim's race, so we are looking at each column in the table. The conditional odds ratios are given below:

Given that the victim is black, the estimated conditional odds that a black defendant gets the death penalty verdict are 1.27 times higher than the odds of a white defendant getting the death penalty. Given that the victim is white, the estimated conditional odds that a black defendant gets the death penalty verdict are 1.4 times the odds of a white defendant getting the death penalty.

t1 <- xtabs(Frequency ~ Defendant + Penalty + Victim ) + 0.5  
t1

## , , Victim = B  
##   
## Penalty  
## Defendant N Y  
## B 97.5 6.5  
## W 9.5 0.5  
##   
## , , Victim = W  
##   
## Penalty  
## Defendant N Y  
## B 52.5 11.5  
## W 132.5 19.5

#conditional that victim is black  
(t1[[2]]\*t1[[3]])/(t1[[1]]\*t1[[4]])

## [1] 1.266667

#conditional that victim is white  
(t1[[6]]\*t1[[7]])/(t1[[5]]\*t1[[8]])

## [1] 1.4884

1. **Sample marginal odds ratio. Do the data exhibit Simpson's paradox?** The partial marginal table is given below. The marginal odds that a black defendant gets the death penalty are 16 percent less than the marginal odds that a white defendant gets the death penalty.

To put it succinctly, the conditional tables indicate that white defendants are more likely to get the death penalty but the marginal tables indicate that white people are more likely to get the death penalty. Because the conditional and marginal odds appear to say different things, we would say that these data do indeed exhibit Simpson's paradox.

t2 <- xtabs(Frequency ~ Defendant + Penalty)  
t2

## Penalty  
## Defendant N Y  
## B 149 17  
## W 141 19

(t2[2]\*t2[3])/(t2[1]\*t2[4])

## [1] 0.8466973

d.Logistic Regression Model: + i. What is the estimated marginal odds ratio between defendant's race and death penalty verdict? The estimated marginal odds ratio is calculated as follows, using the model:

The odds of a white defendant getting the death penalty vs the odds of a black defendant getting the death penalty look like the following since the Defendant race variable is 1 for whites (in the numerator) and 0 for blacks (denominator). Then, we have:

* 1. Calculate and interpret a 95% CI for the marginal odds ratio for the death penalty comparing white defendants to black ones. We are 95 % confident that the true marginal odds for the death penalty for white defendants is between 0.59 and 2.36 times the odds of the death penalty for black defendants.
  2. Is there significant statsitical evidence to suggest that the defendant's race has an effect on the odds of the death penalty? No. The 95% confidence interval includes a value of 1, indicating no evidence that the defendant's race influences the odds of the death penalty.

glm.1 <- glm(cbind(Freq,Total - Freq) ~ Def, family = binomial(link = "logit"))  
summary(glm.1)

##   
## Call:  
## glm(formula = cbind(Freq, Total - Freq) ~ Def, family = binomial(link = "logit"))  
##   
## Deviance Residuals:   
## 1 2 3 4   
## 0.2666 -1.5085 1.7371 -1.5957   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -2.1707 0.2560 -8.48 <2e-16 \*\*\*  
## DefW 0.1664 0.3539 0.47 0.638   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 8.1316 on 3 degrees of freedom  
## Residual deviance: 7.9102 on 2 degrees of freedom  
## AIC: 24.224  
##   
## Number of Fisher Scoring iterations: 4

library(boot)  
inv.logit(coef(glm.1)) #would give probability scale, not really what we want here

## (Intercept) DefW   
## 0.1024096 0.5415073

1/exp(coef(glm.1)) #gives us the odds

## (Intercept) DefW   
## 8.7647059 0.8466973

#confidence intervals:   
exp(coef(glm.1)[2] + 1.96 \* 0.3539)

## DefW   
## 2.363293

exp(coef(glm.1)[2] - 1.96 \* 0.3539)

## DefW   
## 0.5902365

e.Fit a logistic regression model which allows for conditional associations:

glm.2 <- glm(cbind(Freq,Total - Freq) ~ Def + Vic, family = binomial(link = "logit"))  
summary(glm.2)

##   
## Call:  
## glm(formula = cbind(Freq, Total - Freq) ~ Def + Vic, family = binomial(link = "logit"))  
##   
## Deviance Residuals:   
## 1 2 3 4   
## 0.0803 -0.8145 -0.1072 0.1394   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -2.8421 0.4203 -6.761 1.37e-11 \*\*\*  
## DefW -0.4402 0.4009 -1.098 0.2722   
## VicW 1.3242 0.5193 2.550 0.0108 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 8.13161 on 3 degrees of freedom  
## Residual deviance: 0.70074 on 1 degrees of freedom  
## AIC: 19.015  
##   
## Number of Fisher Scoring iterations: 4

#Condition on White Victim  
exp(coef(glm.2)[2] + coef(glm.2)[3])

## DefW   
## 2.42054

#Condition on Black Victim  
exp(coef(glm.2)[2])

## DefW   
## 0.6438933

library(boot)  
exp(confint(glm.2))

## Waiting for profiling to be done...

## 2.5 % 97.5 %  
## (Intercept) 0.02276198 0.1217834  
## DefW 0.29543951 1.4371323  
## VicW 1.41648864 11.1990289

* The estimated conditional odds ratio between defendant's race and the death penalty for black victims? For White victims? Show your work. If we condition on black victims, several indicators go to 0. We then have:

For white victims, we have to include a few more terms. However, we get the same conditional odds ratio.

* Calculate and interpret an approximate 95% confidence interval for the conditional odds ratio of the death penalty comparing white defendants to black defendants conditioned on victim's race.

We are 95% confident that, conditioning no the defendant's race, the true odds of the death penalty for white defendants are between .295 and 1.43 times the odds of the death penalty for black defendants.

* Is there significant statistical evidence to suggest that the defendant's race has an effect on the odds of the death penalty after we control for the victim's race? Justify you answer.

Since the confidence interval for the conditional odds includes 1, we conclude that the defendant's race does not have an effect on the odds of the death penalty after controlling for the victim's race.

1. I fit the four loglinear models. The fitted conditional odds ratios between Defendant and Penalty, conditioned on Victim type, are given below.

library(MASS)  
mod1.ll <- glm(Frequency ~ Defendant + Victim + Penalty, family = poisson(link = "log"))  
mod2.ll <- glm(Frequency ~ Defendant \* Victim + Penalty \* Victim, family = poisson(link="log"))  
mod3.ll <- glm(Frequency ~ Defendant \*Victim + Victim\*Penalty + Defendant\*Penalty, family = poisson(link = "log"))  
mod4.ll <- glm(Frequency ~ Defendant \* Victim \* Penalty, family = poisson(link = "log"))  
  
#Conditional Odds Ratios:  
exp(coef(mod1.ll)[2]) #Conditional odds ratio for white defendant

## DefendantW   
## 0.9638554

exp(coef(mod2.ll)[2] )#Conditional odds for white defendant, black victim

## DefendantW   
## 0.08737864

exp(coef(mod2.ll)[2] + coef(mod2.ll)[5]) #conditional odds for white defendant, white vic

## DefendantW   
## 2.396825

#black vic, no dp  
exp(coef(mod3.ll)[2])

## DefendantW   
## 0.08912716

#white vic, no dp  
exp(coef(mod3.ll)[2] + coef(mod2.ll)[5])

## DefendantW   
## 2.444788

#white vic dp  
exp(sum(coef(mod3.ll)[c(2:7)]))

## [1] 0.191875

#black vic dp  
exp(coef(mod3.ll)[2] + coef(mod3.ll)[4] + coef(mod3.ll)[6])

## DefendantW   
## 0.0195343

coef(mod4.ll)

## (Intercept) DefendantW   
## 4.5747110 -2.3774864   
## VictimW PenaltyY   
## -0.6234673 -2.7829515   
## DefendantW:VictimW DefendantW:PenaltyY   
## 3.3090446 -21.7168582   
## VictimW:PenaltyY DefendantW:VictimW:PenaltyY   
## 1.2296031 21.3318437

#black vic no dp  
exp(coef(mod4.ll)[2])

## DefendantW   
## 0.09278351

# white vic no dp  
exp(coef(mod4.ll)[2] + coef(mod4.ll)[3] + coef(mod4.ll)[5])

## DefendantW   
## 1.360825

# black vic dp  
exp(coef(mod4.ll)[2] + coef(mod4.ll)[4] + coef(mod4.ll)[6])

## DefendantW   
## 2.124901e-12

# white vic dp  
exp(sum(coef(mod4.ll)[c(2:8)]))

## [1] 0.1958763

1. I assessed the goodness-of-fit for each model type. I would choose either model two or model three as they have large p-values. In either case, we do not have evidence of a lack of fit. The saturated model technically fits the best (model 4); however, in the interest of parsimony I think that either model 2 or 3 are better. Model 3 requires estimation of more parameters; however, I think it may be more useful in this context.

pchisq(mod1.ll$dev,mod1.ll$df.residual,lower.tail=FALSE) #small p-value

## [1] 7.83249e-29

pchisq(mod2.ll$dev,mod2.ll$df.residual,lower.tail=FALSE)

## [1] 0.3902578

pchisq(mod3.ll$dev,mod3.ll$df.residual,lower.tail=FALSE)

## [1] 0.4025351

#Model 4 is the saturated model so it has deviance 0. No test needed here; would result in p-value of 1.   
#I'd pick model 2 since it has the largest p-value and is more parsimonious than .

h.A 95% confidence interval for the conditional odds ratios between Victim and Penalty, conditioned on Defendant, is given below. We have to calculate the standard error (for whites) using the variance-covariance matrix from the model.

In the first model, if we condition on having a black defendant, everything in the odds ratio will cancel out except the last interaction term, the interaction between victim and penalty. We are 95% confident that the true odds of a black defendant getting the death penalty with a white victim is between 1.97 and 3.78 times the odds of getting the death penalty with a black victim.

In the second model, conditioning on a white defendant, we will end up keeping the defendant/victim interaction as well in the model. We are 95% confident that the true odds of getting the death penalty for a white defendant with white victim are between 24 and 255 times larger than the odds of the death penalty with a black victim. This seems unusually large, but I am pretty sure I have the right values; however, there are a number of moving parts involved in these odds ratios so my intutition may just be wrong.

# picked model 2  
vc <- vcov(mod2.ll)  
  
# Odds ratios for black defendant  
exp(coef(mod2.ll)[6]) + c(1,-1)\*1.96\*.4635

## [1] 3.788895 1.971975

# Odds ratio for white defendant: needs covariance matrix  
#We would have   
standard.error <- sqrt(vc[5,5] + vc[6,6] + 2\*vc[5,6])  
#  
exp(sum(coef(mod2.ll)[c(5,6)]) + c(1,-1)\*1.96\*standard.error)

## [1] 255.3464 24.4483

## Problem 2

1. The marginal pmf of y can be found as follows. First we need to get the joint distribution, then integrate out the $\labmda$ piece.
2. Where and k and are both > 0.

Carrying out some algebra and changing the way we write a few terms, we obtain the joint:

We now integrate the joint over the support of :

As much fun as it would be to try and directly integrate this, we can move some terms around to see that there's a kernel of a gamma distribution sitting in this thing. Moving some terms around, we can see that this is equal to the following:

We can multiply in the correct constants to get:

The integral is a PDF integrated over its entire support so it is equal to 1. Then we are left with the terms on the left. This is the PMF of a negative binomial distribution; we can transform some terms to quickly see that it is equal to the desired form:

1. We can then show that this is a member of the exponential dispersion family:

First, recall the form of an exponenetial dispersion family:

I think it is a bit easier to go back to the previous form to try and get the negative binomial PDF into the exponential dispersion family form. Starting here, we have:

We take the exponential of the log to get in the right form:

Simplifying, we obtain:

More algebra follows:

We can organize some terms to obtain:

Here, I have the right form for the natural parameter in the first term; I still have a few pieces floating around that need to be cleaned up. Using properties of logs, I can move the and to see that the terms in the log end up cancelling out.

Recognizing that , we can write this as:

This would seemingly put us in a bit of a conundrum. The inside of the second log term cannot be broken up into two pieces. I have to credit Kenny with with coming up with a devious mathematical trick that solves this issue. It relies on the fact that the need only be some *function* of . We can simply take the exponential of the log of the inside, as follows.

Now we have the correct form! This gives us natural parameter and dispersion parameter 1.

## Problem 3

1. If we model the number of reinfections observed in a group, the variable that should be considered the offset term is the amount of time that each group was followed for. A quick look at the "years followed" variable shows that some people were followed for more than 55 years whereas others were followed only briefly; we should account for this difference when modeling the number of reinfections observed.
2. The Poisson regression model that shows how the rate of reinfection varies with race, education, initial infection and condom use is given below. Note that we are not directly modeling the total number of reinfections because each group was observed for a different amount of time. Instead, we model the *rate* of reinfections that occurs over a given constant period of time.

The (additive) theoretical model takes the following form with as the total number of reinfections per group and the as the amount of time the individuals in the group were observed. Technically, the way I have this written is as a model of the rate .

We can directly model the mean, then, by expanding the log and moving the offset term to the right-hand side of the equation.

1. I fit the model below. The summary is included. We can interpret the coefficients for condom usage and race.

The estimated count of STD reinfections for folks who always use condoms is 32 percent lower than the estimated count of reinfections for those who either sporadically use them or do not use them at all, holding other covariates constant.

The estimated count of STD reinfections for white people is 25.5% lower than for black people, assuming that their other characteristics remain roughly the same.

setwd("C:/Users/Paul/Downloads")  
std <- read.table("std.txt", header = TRUE)  
  
std.mod <- glm(n.reinfect ~ factor(white) + factor(edugrp) + factor(inftype) + factor(condom) , offset = log(yrsfu), data = std, family = poisson(link = "log"))  
  
summary(std.mod)

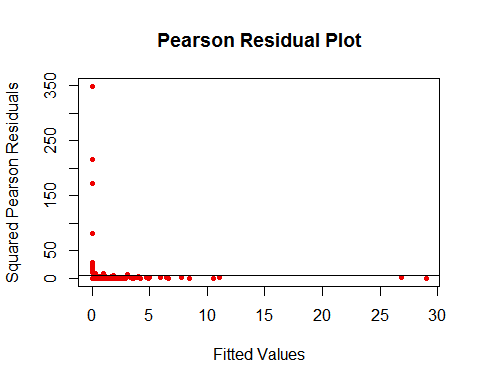
##   
## Call:  
## glm(formula = n.reinfect ~ factor(white) + factor(edugrp) + factor(inftype) +   
## factor(condom), family = poisson(link = "log"), data = std,   
## offset = log(yrsfu))  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.4764 -0.9200 -0.2389 0.7529 3.1191   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -0.5020 0.1465 -3.427 0.00061 \*\*\*  
## factor(white)1 -0.2935 0.1296 -2.264 0.02357 \*   
## factor(edugrp)(12.9,18] -0.3450 0.1922 -1.795 0.07265 .   
## factor(edugrp)[6,11.9] 0.2226 0.1165 1.910 0.05610 .   
## factor(inftype)2 -0.3747 0.1478 -2.535 0.01124 \*   
## factor(inftype)3 -0.2968 0.1452 -2.044 0.04092 \*   
## factor(condom)1 -0.3724 0.1150 -3.238 0.00121 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for poisson family taken to be 1)  
##   
## Null deviance: 364.88 on 257 degrees of freedom  
## Residual deviance: 324.98 on 251 degrees of freedom  
## AIC: 686.55  
##   
## Number of Fisher Scoring iterations: 6

#for coefficients  
exp(coef(std.mod))

## (Intercept) factor(white)1 factor(edugrp)(12.9,18]   
## 0.6053149 0.7456474 0.7082351   
## factor(edugrp)[6,11.9] factor(inftype)2 factor(inftype)3   
## 1.2493044 0.6875241 0.7431784   
## factor(condom)1   
## 0.6890760

1. To examine the data for overdispersion we can plot the pearson residuals. The plot is given below. Given the mean of the squared residuals is roughly 5, it might be worth looking into a quasi-likelihood method that allows us to estimate an overdispersion parameter.

plot(fitted(std.mod),resid(std.mod,"pearson")^2,ylab = "Squared Pearson Residuals",xlab = "Fitted Values", main ="Pearson Residual Plot", pch = 20, col = "red2")  
abline(h = mean(resid(std.mod,"pearson")^2))



mean(resid(std.mod,"pearson")^2)

## [1] 5.029057

I refit my model using quasi-poisson below. After accounting for overdispersion, the conclusions are pretty boring by comparison to the original model. The first model had standard errors which were too small, so nearly every predictor had a small associated p-value. After accounting for problem that the model assumed too small a variance for the data, none of the p-values were very small. In this case, accounting for overdispersion changes the results in a pretty big way.

std.mod.2 <- glm(n.reinfect ~ factor(white) + factor(edugrp) + factor(inftype) + factor(condom) , offset = log(yrsfu), data = std, family = quasipoisson(link = "log"))  
  
summary(std.mod.2)

##   
## Call:  
## glm(formula = n.reinfect ~ factor(white) + factor(edugrp) + factor(inftype) +   
## factor(condom), family = quasipoisson(link = "log"), data = std,   
## offset = log(yrsfu))  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.4764 -0.9200 -0.2389 0.7529 3.1191   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) -0.5020 0.3330 -1.507 0.133  
## factor(white)1 -0.2935 0.2947 -0.996 0.320  
## factor(edugrp)(12.9,18] -0.3450 0.4370 -0.789 0.431  
## factor(edugrp)[6,11.9] 0.2226 0.2649 0.840 0.402  
## factor(inftype)2 -0.3747 0.3360 -1.115 0.266  
## factor(inftype)3 -0.2968 0.3301 -0.899 0.369  
## factor(condom)1 -0.3724 0.2615 -1.424 0.156  
##   
## (Dispersion parameter for quasipoisson family taken to be 5.16931)  
##   
## Null deviance: 364.88 on 257 degrees of freedom  
## Residual deviance: 324.98 on 251 degrees of freedom  
## AIC: NA  
##   
## Number of Fisher Scoring iterations: 6

library(pander)  
exp(coef(std.mod.2))

## (Intercept) factor(white)1 factor(edugrp)(12.9,18]   
## 0.6053149 0.7456474 0.7082351   
## factor(edugrp)[6,11.9] factor(inftype)2 factor(inftype)3   
## 1.2493044 0.6875241 0.7431784   
## factor(condom)1   
## 0.6890760

dat1 <- data.frame(cbind(exp(coef(std.mod)),exp(coef(std.mod.2))))  
colnames(dat1) <- c("Poisson","QuasiPoisson")  
pander(dat1)

|  |  |  |
| --- | --- | --- |
|  | Poisson | QuasiPoisson |
| **(Intercept)** | 0.6053 | 0.6053 |
| **factor(white)1** | 0.7456 | 0.7456 |
| **factor(edugrp)(12.9,18]** | 0.7082 | 0.7082 |
| **factor(edugrp)[6,11.9]** | 1.249 | 1.249 |
| **factor(inftype)2** | 0.6875 | 0.6875 |
| **factor(inftype)3** | 0.7432 | 0.7432 |
| **factor(condom)1** | 0.6891 | 0.6891 |

1. The grouped reinfection counts do not appear to folllow a true Poisson distribution because their variance is larger than their mean. If they were truly distributed Poisson, the variance and the mean would be equal. Because they are overdispersed, these data are likely something similar to a Poisson but are better modeled by something slightly less restrictive.
2. We conduct a likelihood ratio test to see whether the use of condoms has an effect on the rate of reinfection. We showed in class that the LRT test fitted on a quasi-Poisson model actually uses the correct standard errors. In this case, we have only weak (at best) evidence that condom usage actually reduces the chance of re-infection.

anova(std.mod.2, test = "LRT")

## Analysis of Deviance Table  
##   
## Model: quasipoisson, link: log  
##   
## Response: n.reinfect  
##   
## Terms added sequentially (first to last)  
##   
##   
## Df Deviance Resid. Df Resid. Dev Pr(>Chi)  
## NULL 257 364.88   
## factor(white) 1 9.5263 256 355.35 0.1746  
## factor(edugrp) 2 13.7200 254 341.63 0.2653  
## factor(inftype) 2 5.7952 252 335.84 0.5709  
## factor(condom) 1 10.8545 251 324.98 0.1473