StatsHw9

Trevor Freeland

May 14, 2018

```
library(tidyverse)
library(MASS)
library(stargazer)
library(pscl)
library(pander)
```

```
med <- read.csv("http://math.carleton.edu/Chihara/Stats345/medvisits.cs")
beatles <- read.csv("http://math.carleton.edu/Chihara/Stats345/BeetleEggs.csv")
model.counts <- source("http://math.carleton.edu/Chihara/Stats345/Count_Function.R")$value
```

71

	X = 0	X = 1
W = 0	$\exp(\beta_0)$	$\exp(\beta_0 + \beta_X)$
W = 1	$\exp(\beta_0 + \beta_W)$	$\exp(\beta_0 + \beta_X + \beta_W + \beta_{WX})$

The odds ratio is ad/bc =

$$\frac{e^{\beta_0+\beta_0+\beta_X+\beta_W+\beta_{WX}}}{e^{\beta_0+\beta_0+\beta_X+\beta_W}} = e^{\beta_{WX}}$$

72

(A)

Our goodness of fit test gives us an extremely small p-value, which indicates that our model is not adequate.

(B)

The goodness of fit tests gives us a p-value of about .5 which indicates that our model is a good fit. There appears to be independence for gender and president voted for.

(C)

The odds that a female owns a gun is 1.6 ($e^{.475}$) times higher than the odds that a male favors the death penalty.

```
gender <- rep(c("Female","Male"),c(4,4))
gun <- rep(rep(c("No","Yes"), c(2,2)),2)
pres <- rep(c("Bush","Kerry"),4)
y <- c(207, 274, 155, 100, 141, 144, 153, 90)
pres <- data.frame(y,gender,gun,pres)

pres.glm1 <- glm(y~gender+gun+pres, data = pres, family = poisson)
summary(pres.glm1)
1-pchisq(52,4)
```

```
pres.glm2 <- glm(y~(gender+gun+pres)^2, data = pres, family = poisson)
summary(pres.glm2)
1-pchisq(.49,1)
```

```
pres.glm3 <- glm(y~gender*gun+pres*gun, data = pres, family = poisson)
summary(pres.glm3)
```

73

(A)

The reason why we can't use AIC or BIC or other similar model comparative measures is because those approaches "depend on a distributional form and a likelihood" and quasi models, like quasi-poisson, do not neccesarily have a distribution form since they are only characterized by their mean and variance.

(B)

They want to estimate overall abundance and in equations 4 and 5 they take into account the fact that negative binomial models give smaller sites more weight relative to quasi-Poisson models, and so they account for that before comparing the two different model types.

74

(A)

From our summary table we can see that the mean and median for number of visists sits around 2-3, but that there also was someone with 60 visits, and we may want to be aware of that possible outlier going forward. Looking at a histogram of the number of visits split by before and after that reform we see more zeros after the reform but also the tail appears to get longer. In both phases we notice a very large uptick at 0, and so that might be a sign we want to use a negative binomial model or a zip model.

```
pander(summary(med))
```

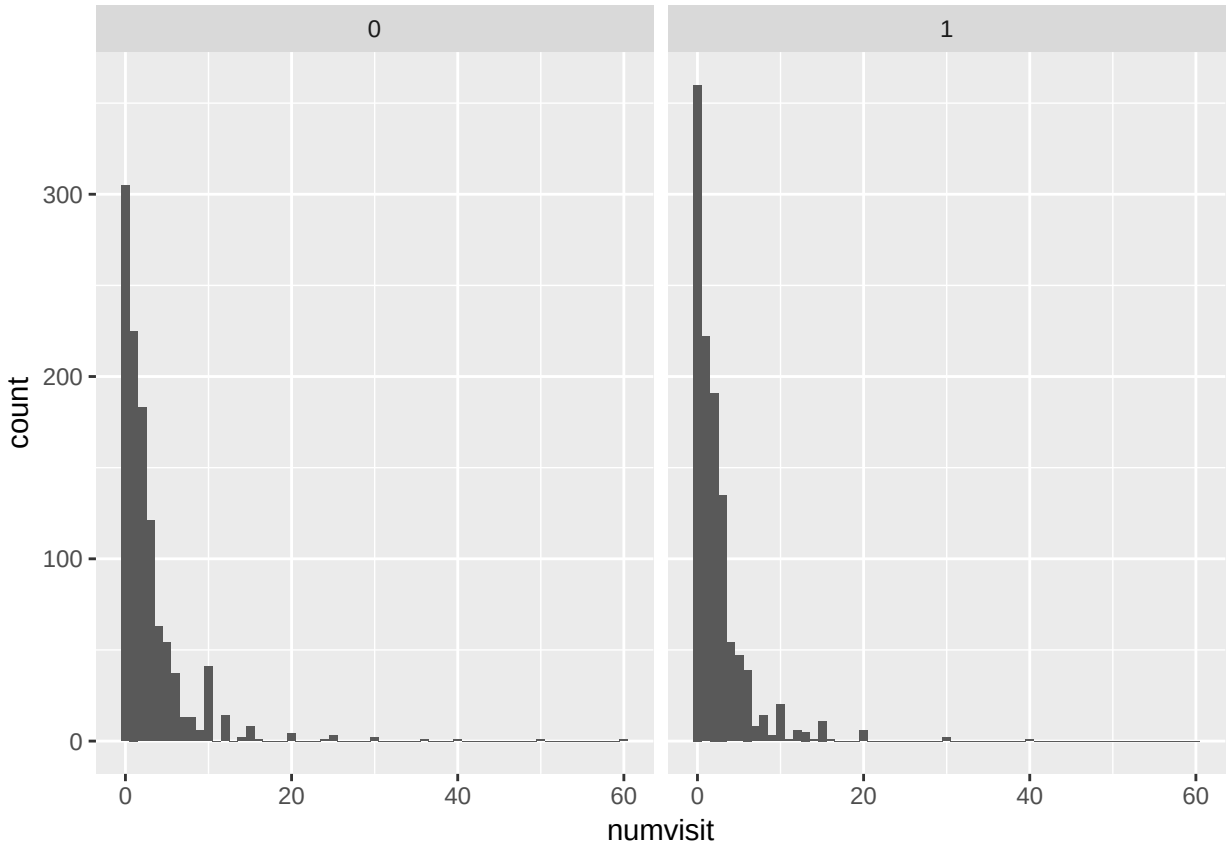
Table 1: Table continues below

numvisit	reform	badh	age
Min. : 0.000	Min. :0.0000	Min. :0.0000	Min. :20.00
1st Qu.: 0.000	1st Qu.:0.0000	1st Qu.:0.0000	1st Qu.:28.00
Median : 2.000	Median :1.0000	Median :0.0000	Median :34.00
Mean : 2.589	Mean :0.5061	Mean :0.1136	Mean :36.78
3rd Qu.: 3.000	3rd Qu.:1.0000	3rd Qu.:0.0000	3rd Qu.:46.00
Max. :60.000	Max. :1.0000	Max. :1.0000	Max. :60.00

educ	loginc
Min. :1.000	Min. :5.832
1st Qu.:2.000	1st Qu.:7.484
Median :2.000	Median :7.725
Mean :2.091	Mean :7.713
3rd Qu.:3.000	3rd Qu.:7.949
Max. :3.000	Max. :9.333

educ	loginc

```
ggplot(data = med, aes(x = numvisit)) + geom_histogram(binwidth = 1) + facet_wrap(~reform)
```



(B)

The mean number of visits after the reform was instituted is .87 ($e^{-.138}$) times smaller than the mean number of visits before the reform was instituted holding everything else constant.

```
med.glm <- glm(numvisit~reform+badh+age+educ+loginc, data= med, family = poisson)
summary(med.glm)
```

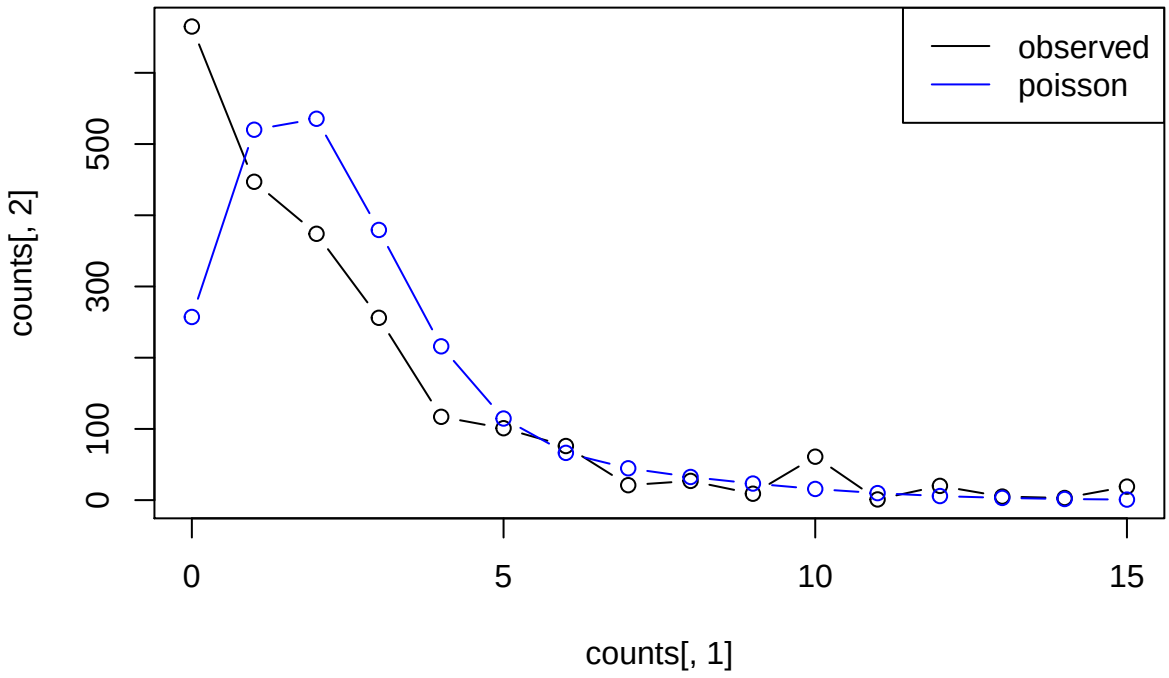
(C)

The estimated dispersion parameter is 4.36

(D)

```
counts <- model.counts(med.glm)
plot(counts[,1], counts[,2], type="b")
lines(counts[,1], counts[,3], type="b", col="blue")
```

```
legend("topright", legend=c("observed", "poisson"), lty=1,
col=c("black", "blue"))
```



(E)

Yes. Using a quasi-poission model education seems to no longer be significant and age now becomes right on the edge of significance.

```
sum(resid(med.glm, type="pearson")^2)/med.glm$df.residual

med.glm2 <- glm(numvisit~reform+badh+age+educ+loginc, data= med, family = quasipoisson)
summary(med.glm2)
```

75

(A)

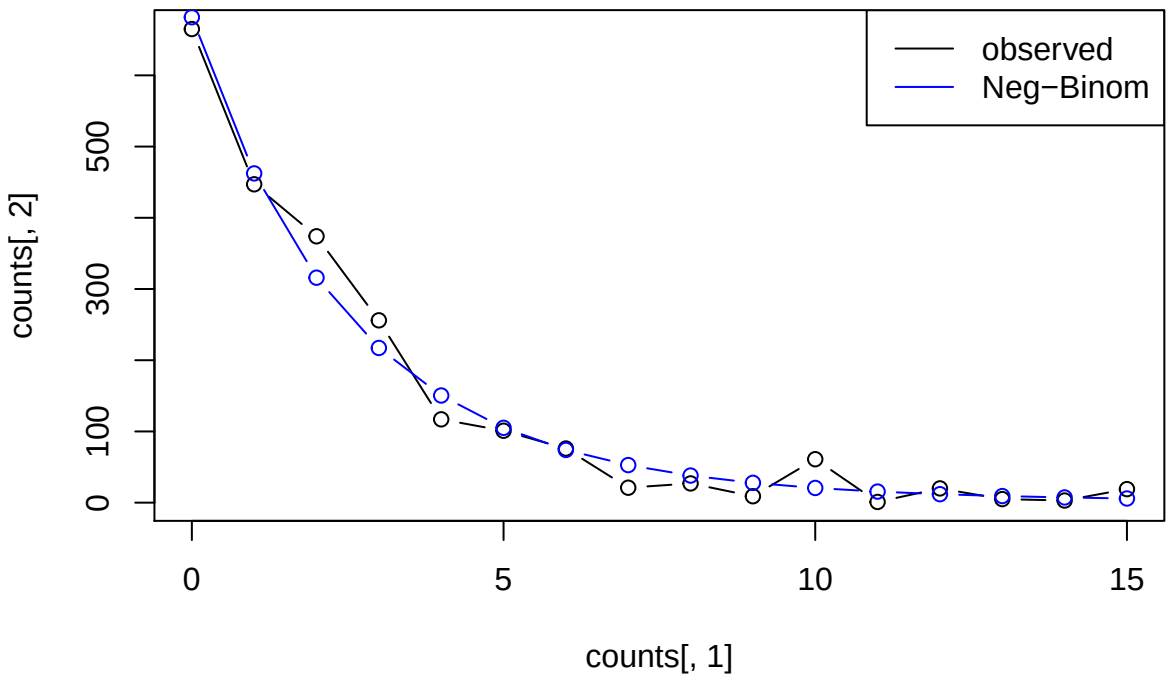
The mean number of visits after the reform was instituted is .87 ($e^{-.137}$) times smaller than the mean number of visits before the reform was instituted holding everything else constant.

```
med.nb <- glm.nb(numvisit~reform+badh+age+educ+loginc, data= med)
summary(med.nb)
med.nb2 <- glm.nb(numvisit~reform+badh+age+loginc, data= med)
summary(med.nb2)
anova(med.nb,med.nb2)
```

```
med.nb3 <- glm.nb(numvisit~reform+badh+age, data= med)
summary(med.nb3)
anova(med.nb2,med.nb3)
```

(B)

```
counts <- model.counts(med.nb3)
plot(counts[,1], counts[,2], type="b")
lines(counts[,1], counts[,3], type="b", col="blue")
legend("topright", legend=c("observed", "Neg-Binom"), lty=1,
col=c("black", "blue"))
```



(C)

Yes we can remove some variables. We removed reform age and loginc from the structural 0 part and the liklihood ratio test gave us a large p-value so we can stick with the smaller model.

(D)

We observed 665 0's and we predicted 664.80 0's.

```
med.zip <- zeroinfl(numvisit~reform+badh+age+educ+loginc, data= med)
summary(med.zip)
```

```
med.zip2 <- zeroinfl(numvisit~reform+badh+age+educ+loginc | educ + badh, data= med)
L0 <- logLik(med.zip2)
L1 <- logLik(med.zip)
D <- 2*(L1-L0)
1-pchisq(D, 3)
summary(med.zip2)
```

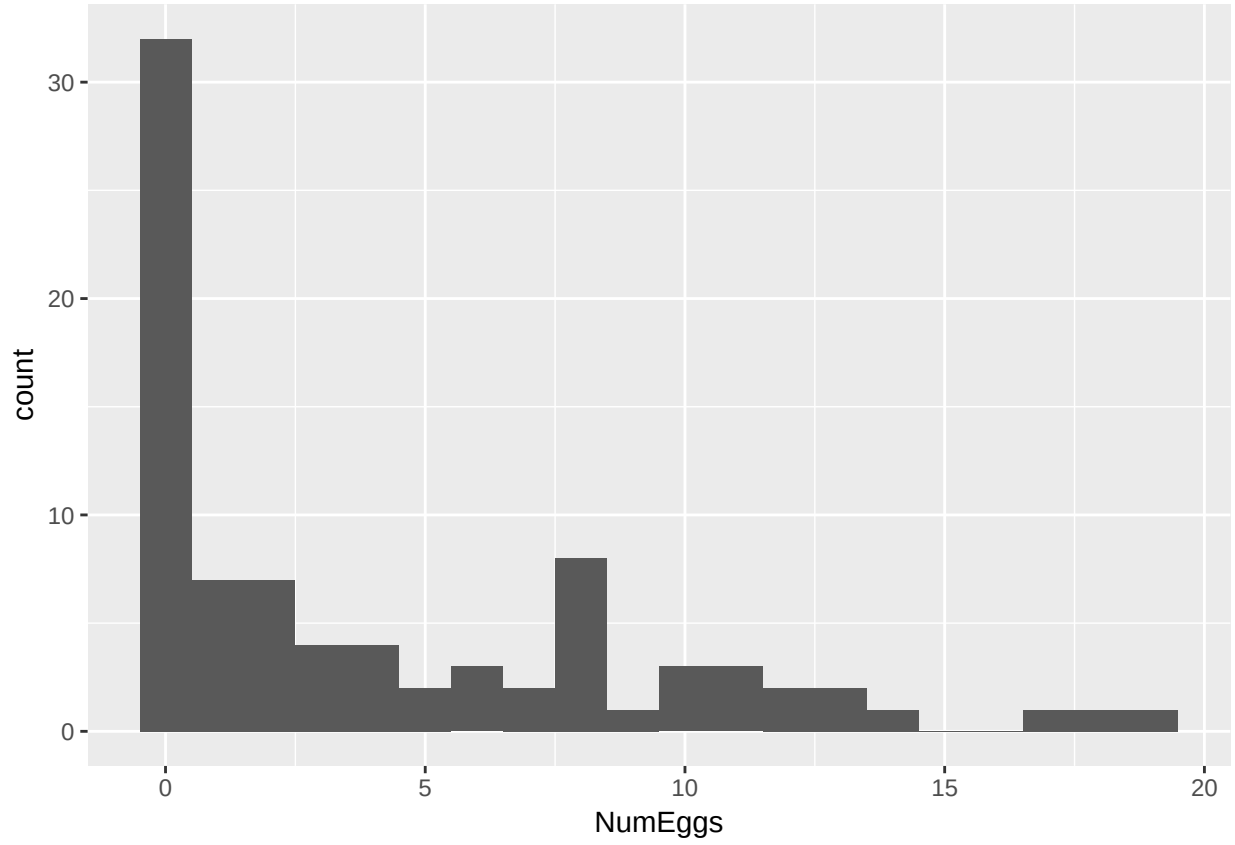
```
model.counts(med.zip2)
```

76

(A)

We can see from the histogram that there is a very very large number of 0's

```
ggplot(beatles, aes(x = NumEggs)) + geom_histogram(binwidth=1)
```



(B)

$\log(\mu)$ = -2.01 + .17(Temp)

$\log(\frac{\pi}{1-\pi})$ = -3.68 + .14(Temp)

(C)

The LRT gives us a p-value of about .35 which indicates that we want the smaller model.

(D)

The 95% confidence interval for temperature is between 1.11 and 1.28. This means for every 1 unit increase in temperature the average number of eggs laid by a female beetle not in diapause increase between 11% and 28% holding everything else constant.

(E)

Observed 32 zeros

Model 1: 32.00 zeros

Model 2: 32.02 zeros

```
beatles.zip <- zeroinfl(NumEggs~Temp, data= beatles)
summary(beatles.zip)
```

```
beatles.zip2 <- zeroinfl(NumEggs~Temp | 1, data= beatles)
summary(beatles.zip2)

L0 <- logLik(beatles.zip2)
L1 <- logLik(beatles.zip)
D <- 2*(L1-L0)
1-pchisq(D, 1)
```

```
exp(confint(beatles.zip2))
```

```
count1 <- model.counts(beatles.zip)
count2 <- model.counts(beatles.zip2)
```

77

(A)

I believe the model we should use in this situation is a zero-inflated model. It seems like this is a dataset that will have a lot of zeros, so we won't want a truncated poisson model. That leaves it to be between either a hurdle or a zero-inflated model and I don't think it makes sense to use a hurdle model because it doesn't seem like there would be a "structural source" and a "sampling source".

(B)

In this case I think we should use a truncated Poisson model because we shouldn't have any zero's because we are asking people who are attending a movie how may movies they have watched this year. Since they are at a movie it should be safe to assume that it would be at least 1 for a count.

(C)

I believe that we should us a hurdle model in this case. The people who say they haven't applied to any colleges are probably students who are not going to be attending college the next year, which means that they are of a different population than the students who are attending college.

(D)

I think this should be another zero-inflated model because this sounds like a dataset that will have a lot of zeros, but it doesn't sound like students would be from two "seperate" populations. I don't think there would be any clear cut seperation of students who have suspended to those who have not.