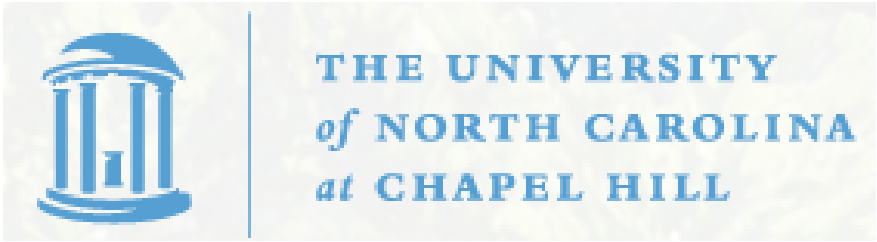
STOR 590:
ADVANCED LINEAR MODELS
Instructor: Richard L. Smith

Class Notes:

September 21, 2020



1

CLASS ANNOUNCEMENTS

- HW4: New deadline is Monday September 21 (late deadline Wednesday)
- From next week, we will revert to regular office hours, but look out for announced changes
- Take-home Midterm: Posted September 26, 6pm, to be returned by September 28, 6pm
- Spring 2020 midterm and final exams have been posted
- Final exam — still planning take-home exam, will update plans after the Midterm

Use of Offset in R

"dicentric" example

```
rmod=glm(ca∼offset(log(cells))+log(doserate)*dosef,
family=poisson,dicentric)
```

Negative Binomial Model

- Number of Bernoulli trials needed to get k'th success
- $\Pr\{Z = z\} = \binom{z-1}{k-1} p^k (1-p)^{z-k}$.
- Alternative: $Y = Z - k$, $p = \frac{1}{1+\alpha}$ so $\Pr\{Y = y\} = \binom{y+k-1}{k-1} \frac{\alpha^y}{(1+\alpha)^{y+k}}$.
- $\mathrm{E}(Y) = \mu = k\alpha$ and $\mathrm{Var}(Y) = k\alpha + k\alpha^2 = \mu + \frac{\mu^2}{k}$.
- Log likelihood is

$$\ell = \sum_{i=1}^{n} \left(y_i \log \frac{\alpha}{1+\alpha} - k \log(1+\alpha) + \sum_{j=0}^{y_i - 1} \log(j+k) - \log(y_i!) \right).$$

- $\eta = \sum_{j=0}^{p} x_{ij}\beta_j = \log \frac{\alpha}{1+\alpha} = \log \frac{\mu}{\mu+k}$

Fitting in R

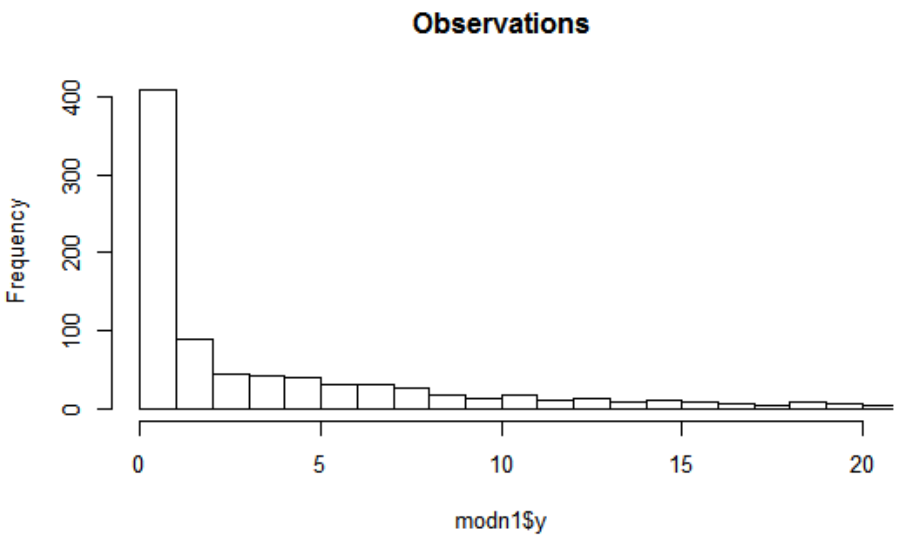
Venables-Ripley method with k fixed:

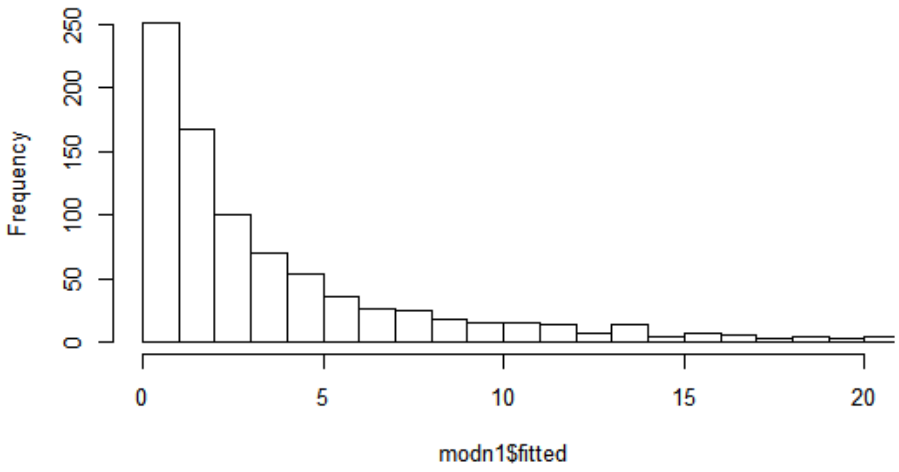
```
library(MASS) modn=glm(skips∼.,negative.binomial(k),solder)
```

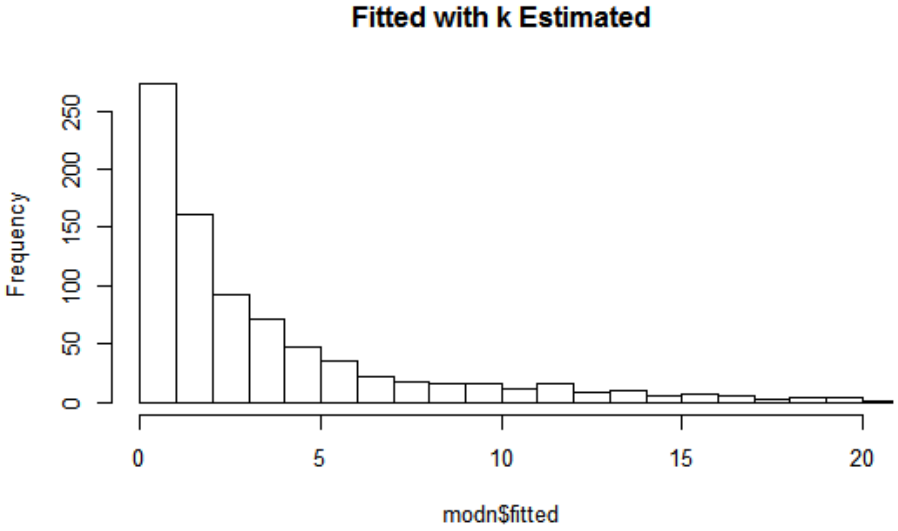
Alternative: determine k by maximum likelihood

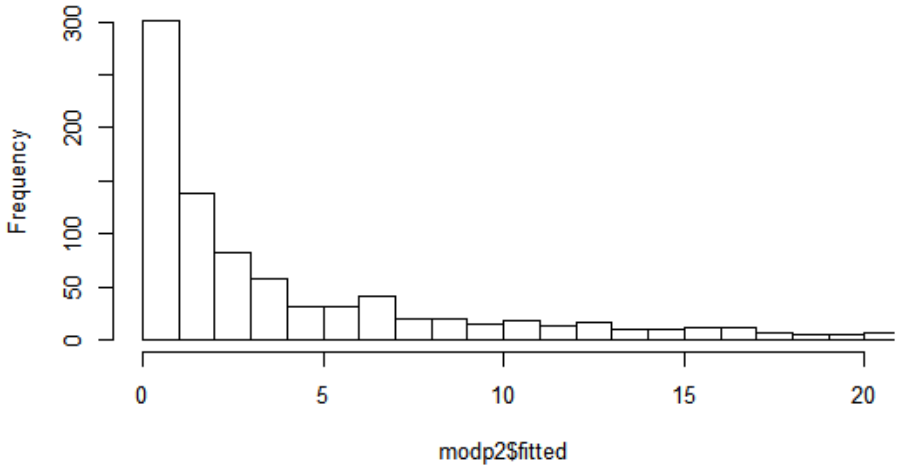
```
modn=glm.nb(skips∼.,solder)
```

The next slide shows histograms of the original data and fitted values under both versions of the negative binomial model. The fit is still not too great.


Observations
Frequency
0
100
200
300
400
0
5
10
15
20
modn1$y


Fitted with k=1
Frequency
0
50
100
150
200
250
0
5
10
15
20
modn1$fitted


Fitted with k Estimated
Frequency
0
50
100
150
200
250
0
5
10
15
20
modn$fitted


Second Poisson Fit
Frequency
0
50
100
200
300
0
5
10
15
20
modp2$fitted

Zero-inflated counts models

- First load package "pscl"
- Hurdle model:

$$\begin{aligned} P(Y=0) &= f_1(0), \\ P(Y=j) &= \frac{1-f_1(0)}{1-f_2(0)} f_2(j),\ j>0, \end{aligned}$$

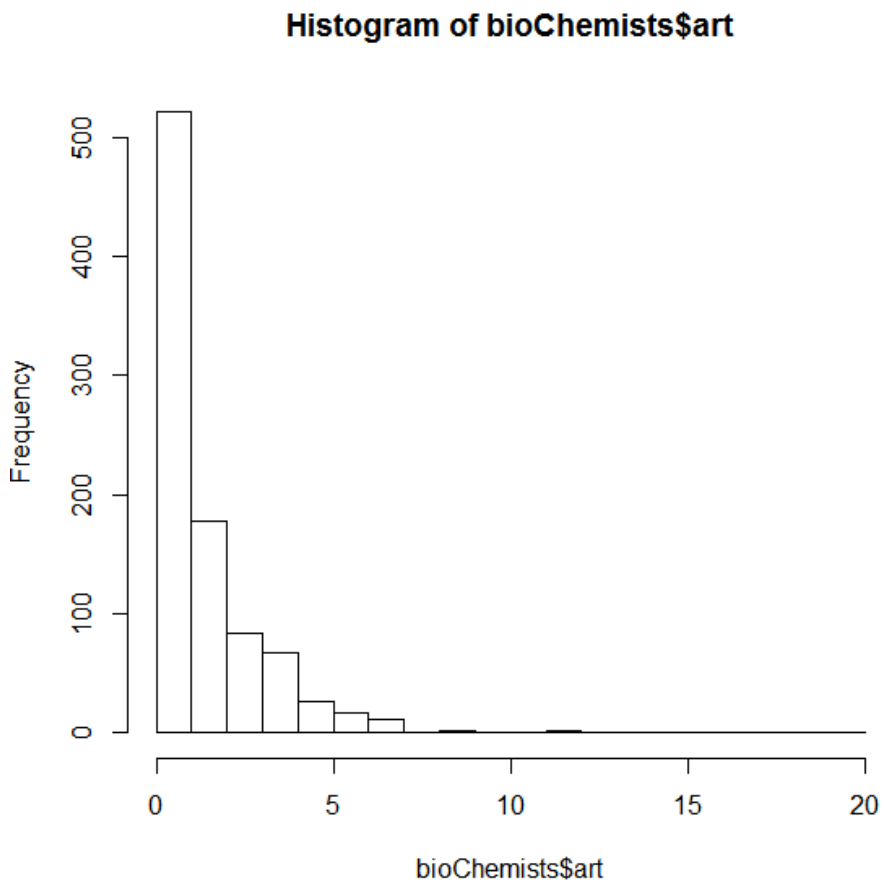
  where (by default) f_2 is Poisson. Fit in R: `hurdle(y~.,data=dataframe)`

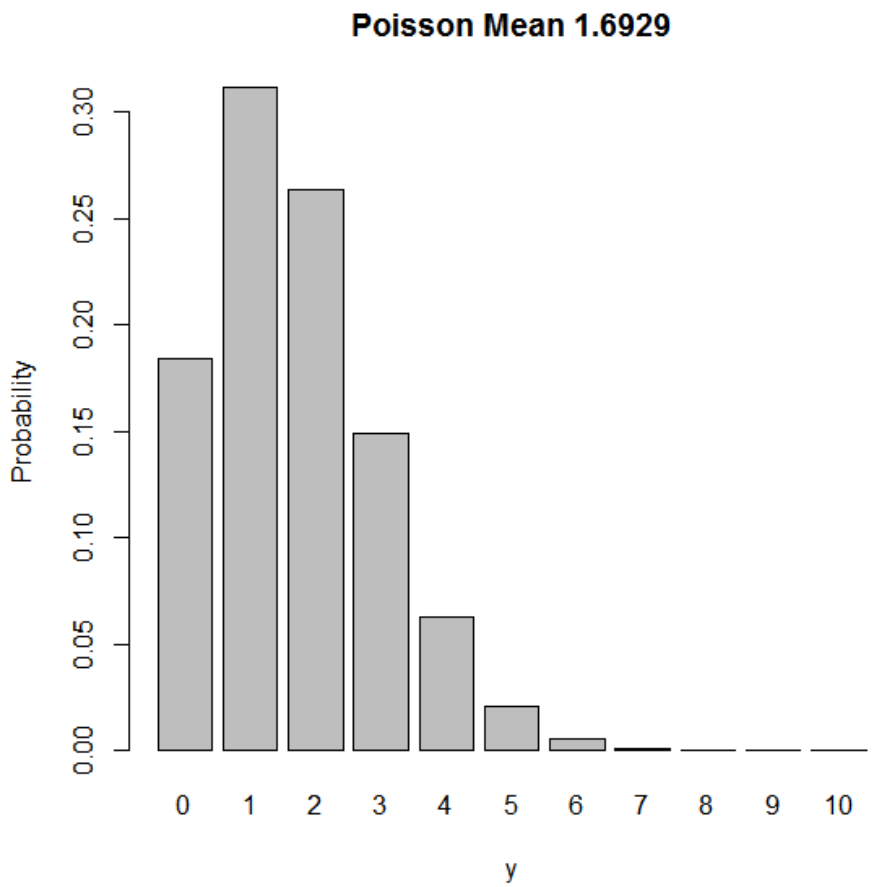
- ZIP model:

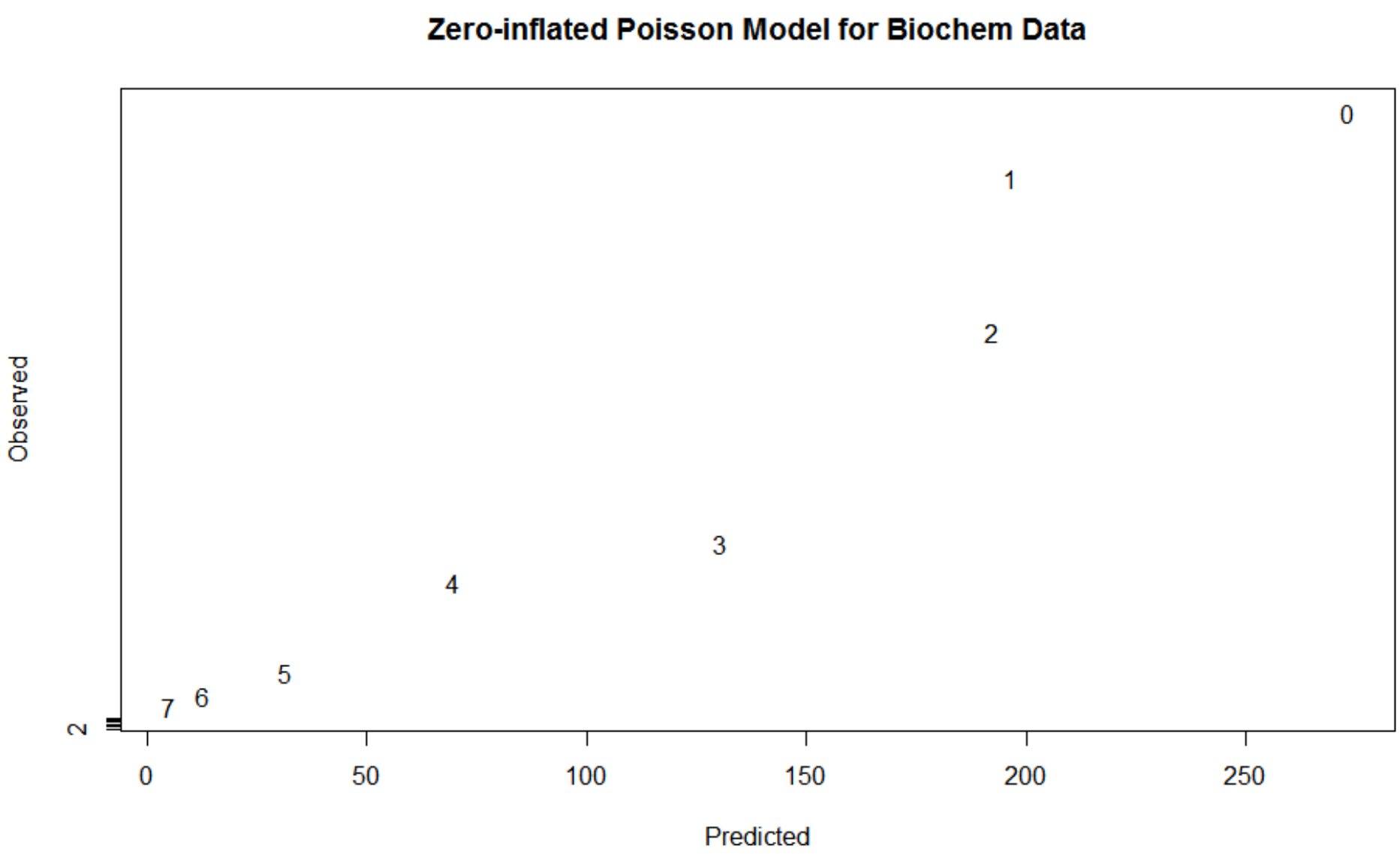
$$\begin{aligned} P(Y=0) &= \phi + (1-\phi) f(0), \\ P(Y=j) &= (1-\phi) f(j),\ j>0. \end{aligned}$$

  Fit in R: `zeroinfl(y~.,data=...)`

- Possibility of using different covariates for the two components, e.g. `zerolinf(y~x1+x2+x3|x4+x5,data=...)`


Histogram of bioChemists$art
Frequency
0
100
200
300
400
500
0
5
10
15
20
bioChemists$art


Poisson Mean 1.6929
Probability
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0
1
2
3
4
5
6
7
8
9
10
y


Zero-inflated Poisson Model for Biochem Data
Observed
Predicted
0
1
2
3
4
5
6
7
2
0
50
100
150
200
250