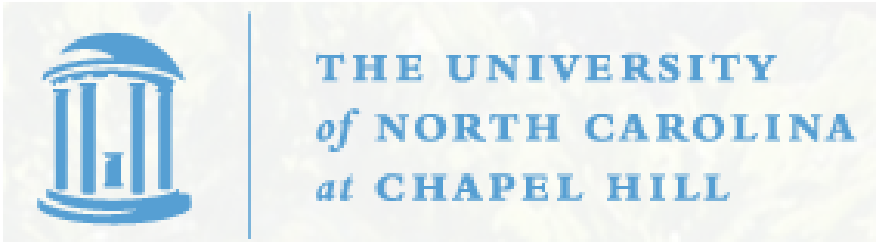
STOR 556: ADV METH DATA ANAL
Instructor: Richard L. Smith

Class Notes #13:

February 21, 2019



Schedule for the Take-home Midterm

- Midterm, posted noon Feb 24, will be set as an assignment
- FIRM deadline: 6pm Feb 25, load through Assignments tab of sakai
- If anyone needs to request any special arrangements please let me know **TODAY**.

Homework 5

- Chapter 5, Problems 2 and 5 (pages 99–101)
- Due date: Tuesday March 5
- I have posted the TA's solutions to HW1–4. Please treat these as *for personal use only* and do not pass on to anyone outside the class!

Guidelines for the Midterm

- There will be 3 questions. Each should take you about an hour including the write-up.
- All questions will involve practical data analysis that you are expected to complete in R.
- Any techniques from chapters 1,2,3,5 that we have covered in class could be needed
- The emphasis will be on statistical modeling and its interpretation. There will be no questions specifically asking you to produce some complicated graphic, but you can earn extra credit by using graphics appropriately and imaginatively
- You don't have to give long-winded answers but I do expect you to *explain your reasoning* and make sure you answer the question as asked! You should use R code, tables and graphics to the extent that they help explain your answer, but page after page of R output without much explanation won't get a lot of credit

Resources Allowed

- You may use the text, all class handouts and all resources contained within R or R-Studio.
- You are not allowed to consult with each other or any outside person besides me.
- You may email me if you don't understand the question or think there might be a mistake. However, I won't give you hints how to do it.
- Use of web resources:
  - Using google, wikipedia or other web resources to make generic queries about R or the statistical methods you are using is allowed and encouraged
  - If you use some resource that we have not covered in class (e.g. some other R package), this is allowed but you should give full citation.
  - If by any chance you find yourself on a webpage that specifically discusses one of the datasets on the exam, *close the page* and do not make further use of the material. If I suspect you are directly copying something unattributed, I will penalize that.
- The Honor Code applies to all aspects of this exam!

Other things

- Resources in R that you have learned about in other UNC courses are also allowed, but please give full citation and say in which other course you learned about it (in case I have to check up with the instructor)
- You are *allowed* but not *required* to submit your solution in R-Markdown. If you do use it, make sure you insert detailed comments to explain what you are doing.

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

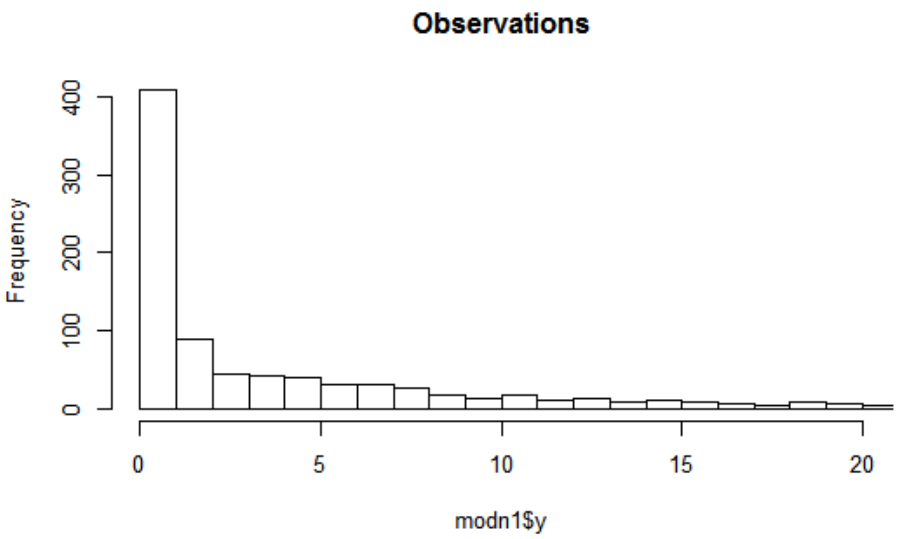
Venables-Ripley method with k fixed:

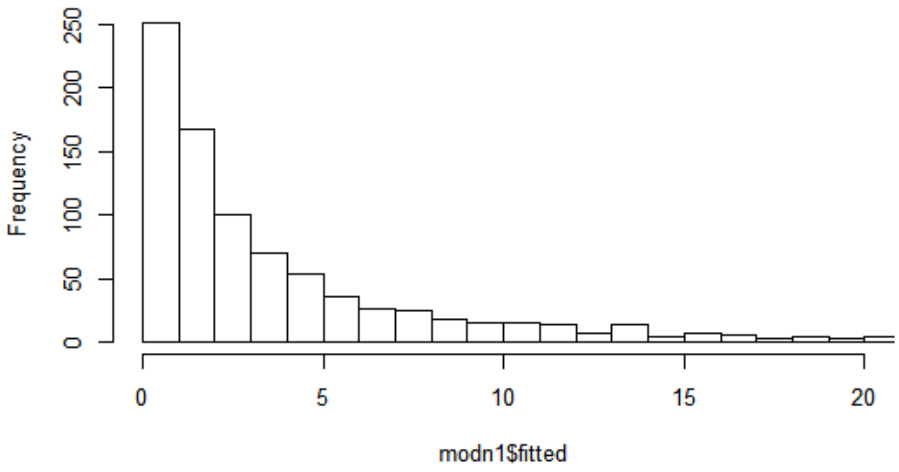
```
library(MASS) modn=glm(skips .,negative.binomial(k),solder)
```

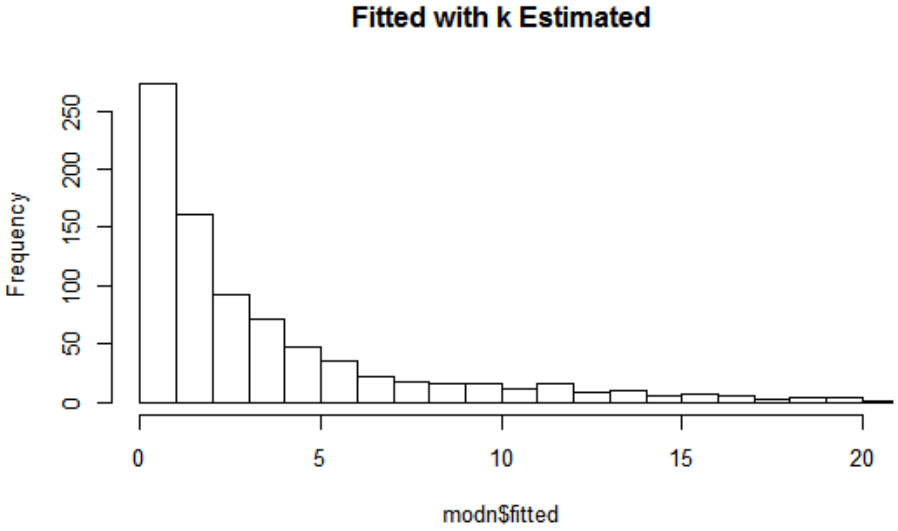
Alternative: determine k by maximum likelihood

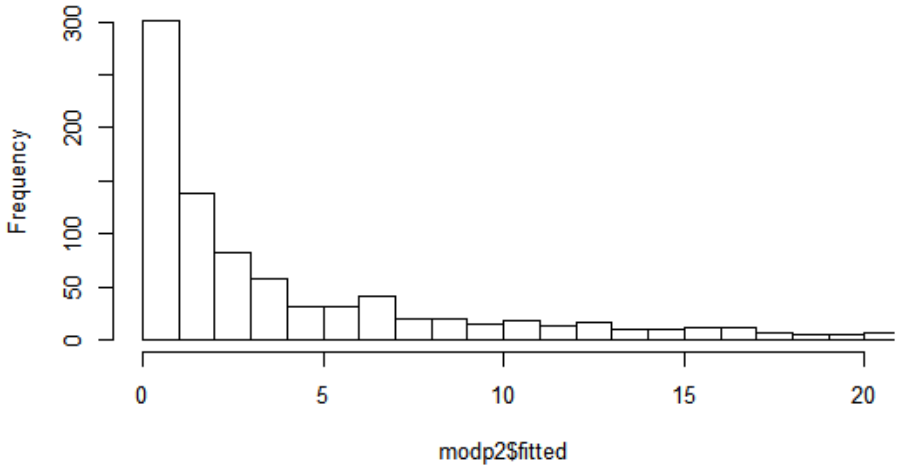
```
modn=glm.nb(skips .,solder)
```

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:

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

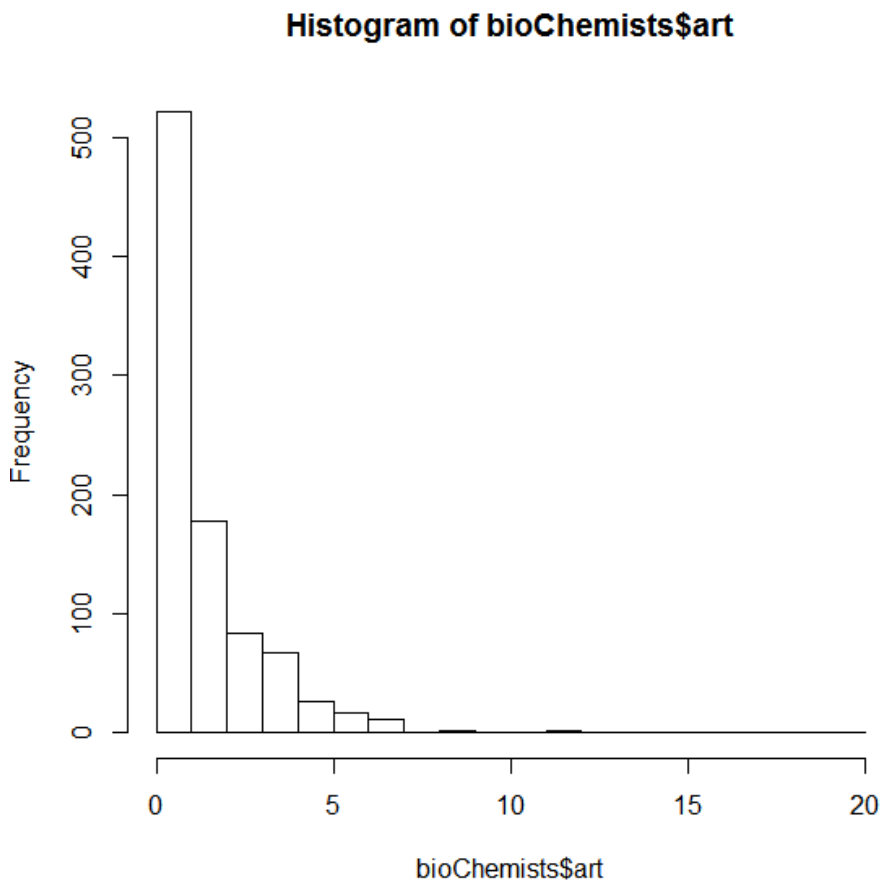
  where (by default) f_2 is Poisson. Fit in R: `hurdle(y~.,data=dataframe)`

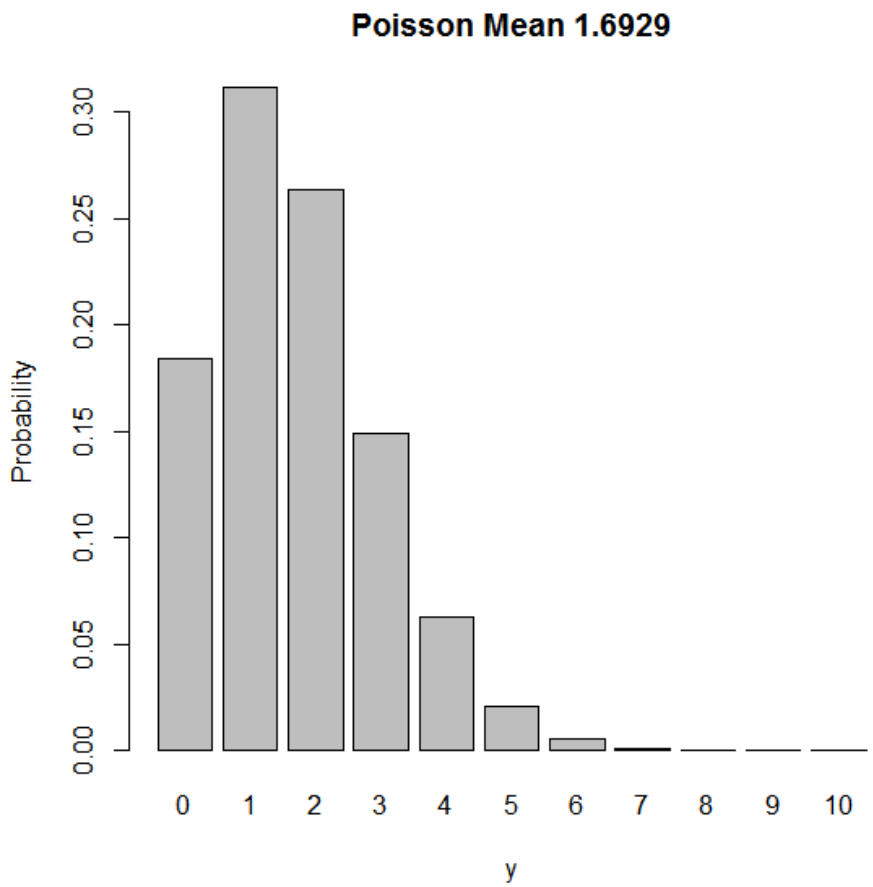
- ZIP model:

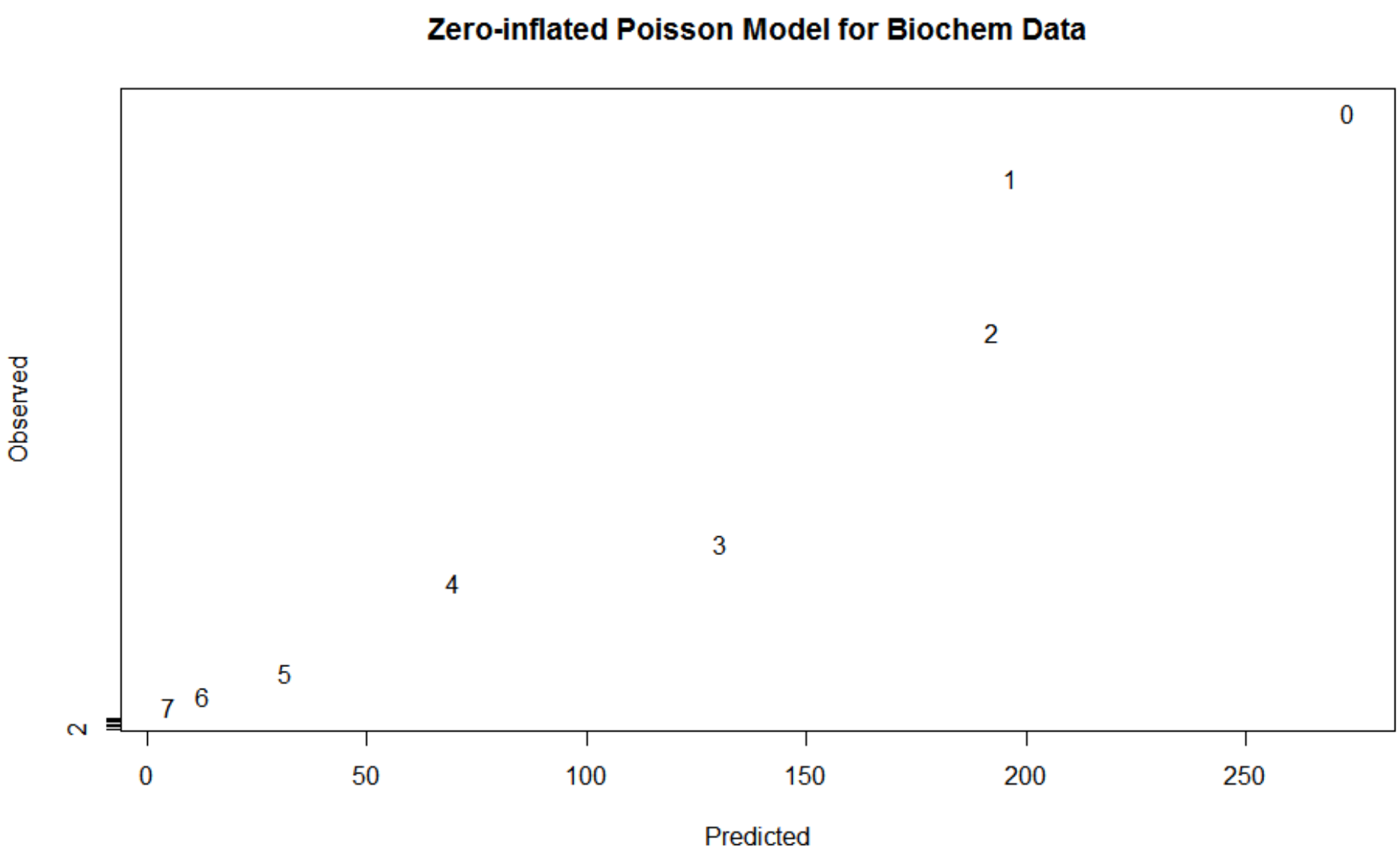
$$\begin{aligned} P(Y=0) &= \phi + (1-\phi) f(0), \\ P(Y=j) &= (1-\phi) f(j),\ j>0. \end{aligned}$$

  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