

Design and Analysis of Count Data

Mani Lakshminarayanan
Merck & Co, Inc

Biopharmaceutical Section Webinar
June 18, 2009



OUTLINE

- Count data as an Endpoint
- Poisson Distribution
- Introduction to GLM
- Over-dispersion in Counts
 - Negative Binomial
 - Zero-inflated Poisson
- Examples
- Sample Size
- Conclusion
- Simulation Results (Backup)

Count Data as an Endpoint



- Examples
 - Number of new enhancing lesions seen on monthly MRI of the brain
 - Number of hospitalizations over a period of time
 - Blood cells in a blood sample (space=volume)
 - Number of patients “at risk” with a differing years of exposure
 - Pregnancy rates measured in terms of Pearl Index that is calculated as the number of pregnancies that would occur per 100 women years of being exposed to the risk of pregnancy.

3

Other examples



- Additional examples include
 - Number of flying-bomb hits in the south of London during World War II (Feller, 1957)
 - Radioactive disintegrations
 - Chromosome interchanges in cells
 - Number of telephone connections to a wrong number
 - Number of bacteria in different areas of a Petri plate

4



ORIGINAL ARTICLE

A Controlled Trial of Natalizumab for Relapsing Multiple Sclerosis

David H. Miller, M.D., Omar A. Khan, M.D., William A. Sheremata, M.D., Lance D. Blumhardt, M.D., George P.A. Rice, M.D., Michele A. Libonati, M.S., Allison J. Willmer-Hulme, Ph.D., Catherine M. Dalton, M.B., Katherine A. Miskiel, M.B., and Paul W. O'Connor, M.D., for the International Natalizumab Multiple Sclerosis Trial Group[‡]

2003, Vol 348, p 15-23

5

Table 2. Overall MRI Findings during the Six Months of Treatment and during Six Months of Follow-up.[‡]

Finding	Placebo (N=71)	3 mg of Natalizumab/kg (N=68)	6 mg of Natalizumab/kg (N=74)	P Value	
				Placebo vs. 3 mg of Natalizumab	Placebo vs. 6 mg of Natalizumab
Treatment					
No. of new enhancing lesions/patient					
Mean	9.6	0.7	1.1	<0.001	<0.001
Median	2.0	0	0		
New enhancing lesions — no. of patients (%)					
No lesions	23 (32)	51 (75)	48 (65)		
1-3 Lesions	18 (25)	14 (21)	20 (27)		
4-6 Lesions	13 (18)	1 (1)	5 (7)		
7-9 Lesions	0	0	0		
10-12 Lesions	3 (4)	1 (1)	0		
>12 Lesions	14 (20)	1 (1)	1 (1)		
No. of persistent enhancing lesions/patient					
Mean ±SD	3.6±6.4	0.8±1.9	1.3±2.6	<0.001	0.005
Median	1	0	0		
No. of new active lesions/patient					
Mean ±SD	9.7±27.4	0.8±2.2	1.1±3.0	<0.001	<0.001
Median	2.0	0	0		
Scans showing activity — %	39	9	11	<0.001	<0.001
Volume of enhancing lesions — mm ³					
Mean ±SD	1168.8±2665.5	156.2±349.3	278.6±632.5	0.005	0.01
Median	266.0	0	0		
Follow-up[†]					
No. of new enhancing lesions/patient					
Mean ±SD	2.4±4.2	2.5±4.5	2.3±5.2	0.89	0.98
Median	1.0	1.0	1.0		
No. of persistent enhancing lesions/patient					
Mean ±SD	0.2±0.46	0.1±0.33	0.1±0.32	0.17	0.09
Median	0	0	0		
No. of new active lesions/patient					
Mean ±SD	2.7±4.3	2.8±5.7	2.5±5.3	0.72	0.82
Median	1.0	1.0	1.0		
Scans showing activity — %	40	41	36	0.96	0.35
Volume of enhancing lesions — mm ³					
Mean ±SD	442.3±813.1	306.0±524.2	234.3±700.7	0.28	0.16
Median	85.0	31.0	25.0		

[‡] Because of rounding, percentages may not total 100.
[†] Values obtained at month 9 and month 12 were combined.

6



Contraceptive Trials



Evaluation of Contraceptive Efficacy and Cycle Control of a Transdermal Contraceptive Patch vs an Oral Contraceptive: A Randomized Controlled Trial

Marie-Claude Audet, MD
 Michèle Moreau, MD
 William D. Kolman, MD
 Arthur S. Waldbaum, MD
 Gary Shargold, MD
 Alan C. Fisher, DrPH
 George W. Cooney, MD
 for the ORTHO EVRA/EVRA 004 Study Group

WORLDWIDE MORE THAN 100 million women choose hormonal contraception for family planning,¹ with more than 1.2 million users in the United States alone.^{2,3} Combined oral contraceptives (OCs) are widely used because of the efficacy demonstrated in clinical trials and the established safety from postmarketing surveillance.⁴ However, while clinical trials have shown that correct and consistent use of OCs results in a first-year failure rate of 0.1%,⁵ the 1995 National Survey of Family Growth (US data) estimated that actual first-year failure rates during typical use of OCs range from 7.3% to as high as 8.5%.⁶ Noncompliance is the primary reason cited to explain this difference.^{7,8} There is clearly a need for reversible contraceptives with a more convenient dosing schedule that would enhance patient compliance and achieve high contraceptive efficacy in typical use. The transdermal contraceptive patch has been evaluated as a new method of contraception in several trials. Ther-

Context Oral contraceptive (OC) pills are effective, but poor compliance increases rates of pregnancy during treatment.

Objective To compare the contraceptive efficacy, cycle control, compliance, and safety of a transdermal contraceptive patch and an OC.

Design Randomized, open-label, parallel-group trial conducted October 1997 to June 1999.

Setting Forty-five clinics in the United States and Canada.

Participants A total of 1417 healthy adult women of child-bearing potential.

Interventions Participants were randomly assigned to receive a transdermal contraceptive patch (n=812) vs an OC (n=605) for 6 or 13 cycles. Patch treatment consisted of application of 3 consecutive 7-day patches followed by 1 patch-free week.

Main Outcome Measures Overall and method-failure Pearl indexes (number of pregnancies/100 person-years of use) and life-table estimates of the probability of pregnancy were calculated. Cycle control, compliance, patch adhesion, and adverse events were also assessed.

Results Overall and method-failure Pearl indexes were numerically lower with the patch (1.24 and 0.99, respectively) vs the OC (2.18 and 1.25, respectively); this difference was not statistically significant (P=.57 and .80, respectively). The incidence of breakthrough bleeding and/or spotting was significantly higher only in the first 2 cycles in the patch group, but the incidence of breakthrough bleeding alone was comparable between treatments in all cycles. The mean proportion of participants' cycles with perfect compliance was 88.2% (81.1 total participants, 214.1 total cycles) with the patch and 77.2% (60.9 total participants, 413.4 total cycles) with the OC (P<.001). Only 1.8% (3/167) of patches completely detached. Both treatments were similarly well tolerated; however, application site reactions, breast discomfort, and dysmenorrhea were significantly more common in the patch group.

Conclusion The contraceptive patch is comparable to a combination OC in contraceptive efficacy and cycle control. Compliance was better with the weekly contraceptive patch than with the OC.

Author Affiliations: Centre Médical des Hôpitaux de Ste-Foy, Ste-Foy, Québec, Dr Audet's Department of Obstetrics and Gynecology, Université de Montréal, Montréal, Québec, Dr Moreau's Medical Center for Women's Clinical Research, San Diego, Calif, Dr Kolman's Department of Clinical Research, The R. W. Johnson Pharmaceutical Research Institute, Raritan, NJ, Dr Shargold, Fisher, and Cooney, Dr Waldbaum is in private practice in Denver, Colo. Reprints: Dr Audet's Department of Obstetrics and Gynecology, Université de Montréal, 3841 Avenue Lacombe, Ste-Foy, Québec, G1Z 3G4, Canada. Group size listed at the end of the article.

Corresponding Author and Reprints: George W. Cooney, MD, The R. W. Johnson Pharmaceutical Research Institute, 920 Route 202, PO Box 300, Raritan, NJ 08869 (e-mail: gcooney@jprc.com).

www.jama.com

(Reprinted) JAMA, May 9, 2001—Vol 285, No. 18 2347

©2001 American Medical Association. All rights reserved.



European Medicines Agency
 Pre-authorisation Evaluation of Medicines for Human Use

London, 27 July 2005
 Doc. Ref. EMEA/CPMP/EWP/519/98 Rev 1



COMMITTEE FOR MEDICINAL PRODUCTS FOR HUMAN USE
 (CHMP)

GUIDELINE ON CLINICAL INVESTIGATION OF STEROID CONTRACEPTIVES
 IN WOMEN

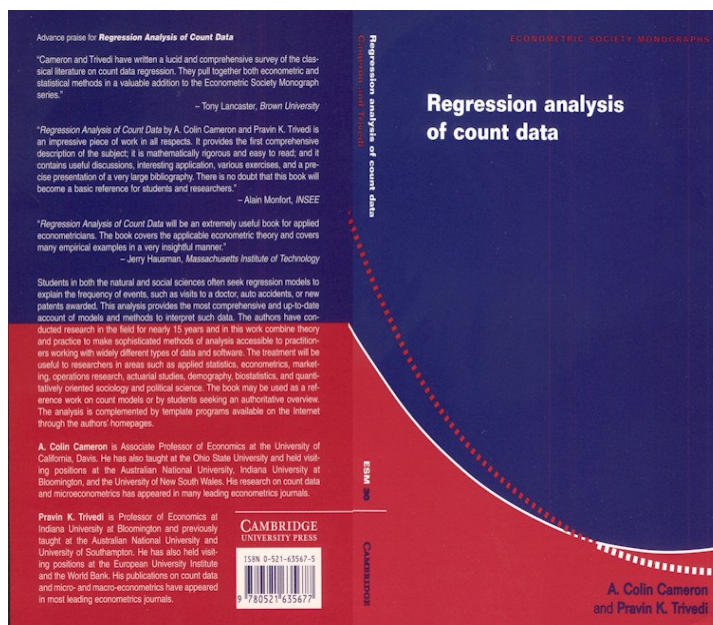
The key studies, carried out in a sufficiently representative population, should normally be at least large enough to give the overall Pearl Index (number of pregnancies per 100 woman years) with a two-sided 95% confidence interval such that the difference between the upper limit of the confidence interval and the point estimate does not exceed 1 (pregnancies per 100 woman years).

Poisson model for Pearl Index



- In the Poisson model, we assume a Poisson distributed random variable X with parameter $\theta=100\lambda T$
 - where λ is the incidence rate per year and T is the total exposure time in 100 woman years.
- X represents the number of pregnancies occurred during a total exposure time of $100T$ woman years which is modeled as Poisson.

9



10

Count data can be modeled using Poisson distribution



Poisson Properties



- **Poisson** distribution for the Count Data can be motivated via a counting stochastic process (aka Poisson process)
 - The expected number of events occurring in an interval of time is proportional to the size of the interval
 - The probability that two events occurring in an infinitesimally small interval of space-time is 0
 - The number of events occurring in separate intervals of space-time are independent
- **Poisson sampling**
 - Distribution of observed counts in the contingency table
 - Thus the counting process has a Poisson due to the Poisson approximation to the binomial distribution

Poisson Distribution



- The probability density function and its central moments are

$$P[Y = y] = \frac{e^{-\lambda} \lambda^y}{y!}, \quad y = 0, 1, 2, \dots$$

$$\lambda_{r+1} = r \lambda \lambda_{r-1} + \pi \frac{\partial \lambda_r}{\partial \lambda}, \quad r = 1, 2, \dots$$

where $\lambda_0 = 0$

- Mean and variance are equal
 - *Equidispersion* property of Poisson
 - Frequently violated in real-life resulting in *overdispersion* (*underdispersion*) due to variance exceeding (or less than) the mean

13

Poisson Distribution



- Use of Poisson as the “law of rare event”
- Poisson counting process
- Use as a waiting time distribution
- Characterization that treats the number of events as repetitions of a binomial outcome, with the number of repetitions taken as a Poisson (aka Poisson-stopped binomial)
 - Useful if the count is generated by randomly repeating a binary outcome

14

Poisson and Counting Process



- A counting process is a stochastic process $\{N(t), t \geq 0\}$ that satisfies the following:
 - $N(t) \geq 0$ and integer valued
 - If $s < t$, then $N(s) \leq N(t)$, and
 - For $s < t$, $N(t) - N(s)$ is the number of events that occur in the time interval $(s, t]$
- The counting process $\{N(t), t \geq 0\}$ in turn is called a Poisson process with event rate $\lambda > 0$ if
 - $N(0) = 0$
 - The process has independent increments, i.e. the numbers of events that occur in disjoint time intervals are independent
 - The process has stationary increments, i.e. the distribution of the number of events $N(t_2 + s) - N(t_1 + s)$ has the same distribution as $N(t_2) - N(t_1)$ for all $t_2 > t_1$ and $s > 0$ and
 - The number of events in any interval of length t follows a Poisson distribution with mean λt , i.e. for all $s, t \geq 0$,

$$P[N(t+s) - N(s) = n] \\ = e^{-\lambda t} (\lambda t)^n / n!, \quad n = 0, 1, \dots$$

15

Note



- The process $N(t)$ has a Poisson due to the Poisson approximation to the binomial distribution (for large n and for small p , so that $\lambda = np$ is finite)

16

Poisson as the “Law of Rare Events”



- Assume that the total number of events will follow, approximately, a Poisson distribution when an event may occur in any large number of trials but the chance of its occurrence may be small in any given trial.
- Formalizing, let $Y_{n,\pi}$ be the total number of successes in a large number of independent Bernoulli trials with π being the success probability in each trial that is small

$$P[Y_{n,\pi} = k] = \binom{n}{k} \pi^k (1 - \pi)^{n-k}, \quad k = 0, 1, \dots, n$$

17

Properties - Summary



- For a Poisson random variable with parameter λ , $E(Y) = \text{Var}(Y) = \lambda$
- Since mean equals the variance, any factor that affects one will also affect the other
 - Usual assumption of homoscedasticity is not appropriate for Poisson data
- Poisson distribution provides an approximation to the binomial for rare events, where p is small and n is large
- Sum of independent Poisson is also a Poisson (additivity property)

$$S_Y = \sum Y_i \sim \text{Poisson}[\sum \lambda_i]$$

18



Properties - Summary

- Normal approximation is satisfactory for $\lambda > 5$ and the proportion λ/n is not too close to zero or one
 - Transformations (logarithmic or arcsine square root) may also be used
- Excessive zero counts may make the variance $>$ mean, overdispersion
 - Failure to observe an event during the period
 - Inability ever to experience an event

19



Multinomial and Poisson

- Using the additivity and conditional properties,

$$\begin{aligned}
 & P[Y_1 = y_1, Y_2 = y_2, \dots, Y_n = y_n \mid S_Y = s] \\
 &= \frac{\left[\prod_{j=1}^n \frac{e^{-\lambda_j} \lambda_j^{y_j}}{y_j!} \right]}{\left[\frac{(\sum \lambda_j)^s e^{-\sum \lambda_j}}{s!} \right]} \\
 &= \frac{s!}{y_1! y_2! \dots y_n!} \left(\frac{\lambda_1}{\sum \lambda_j} \right)^{y_1} \left(\frac{\lambda_2}{\sum \lambda_j} \right)^{y_2} \dots \left(\frac{\lambda_n}{\sum \lambda_j} \right)^{y_n} \\
 &= \frac{s!}{y_1! y_2! \dots y_n!} \pi_1^{y_1} \pi_2^{y_2} \dots \pi_n^{y_n}, \quad \pi_j = \frac{\lambda_j}{\sum \lambda_j}
 \end{aligned}$$

Multinomial distribution

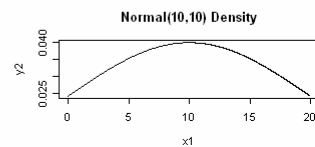
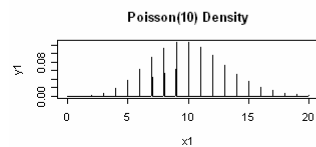
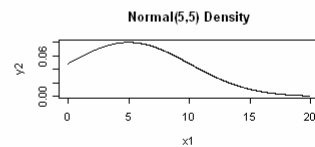
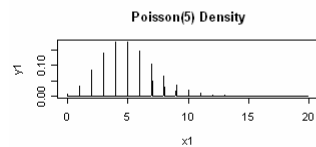
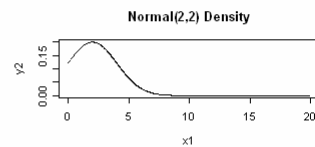
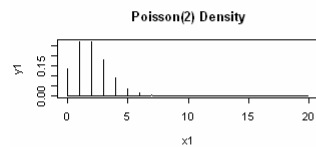
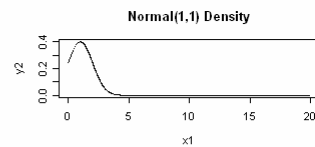
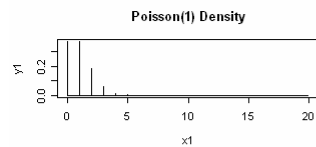
20

Note on multinomial sampling



- Agresti, Chapter 3
- Unusual feature of Poisson sampling is that the total sample size, S_Y , is random rather than fixed.
 - If we start with a Poisson model but condition on $S_Y, \{Y_i\}$
 - No longer is Poisson since Y_i cannot exceed S_Y
 - No longer independent since the value of one affects the possible range of others

21



22



Note

- Recall the pmf of Poisson

$$p(x, \lambda) = e^{-\lambda} \frac{\lambda^x}{x!}$$

- With this, note that, with $p(0, \lambda)$ as the probability of zero

$$\frac{1}{\lambda} \ln[p(0, \lambda)] + 1 = 0$$

- The left side of the above equation, called the zero inflation index, is also used as a measure of departure from Poisson.
 - This index and the coefficient of variation have been used for building GOF procedures

23



Katz system

- Katz(1963) system of discrete distributions are based on the recursive relationship

$$P[y + 1] = P[y] \frac{\omega + \gamma y}{1 + y}, \quad \omega + \gamma y \geq 0, \mu > 0, \gamma < 1$$

- Which has mean $\mu = \omega / (1 - \gamma)$ and variance $\omega / (1 - \gamma)^2$
- Note that Katz family of distributions include, as special cases, Poisson ($\gamma = 0$) and the NB ($0 < \gamma < 1$)
 - The ratio of the variance to the mean is $r = 1 / (1 - \gamma)$
 - Poisson has $r = 1$ and is said to exhibit equi-dispersion
 - The binomial and NB are underdispersed ($r < 1$) and overdispersed ($r > 1$), respectively.

24

Model Details



General Linear Model (GLM)



- Definition

- A vector of observations y of length N is assumed to be a realization of a vector of iid random variables Y with mean μ . A set of covariates x_1, x_2, \dots, x_p defines a linear predictor:

$$\eta = \sum \beta_j x_j$$

- The classical linear model can be written in a tripartite form:

- Random Component

- Y has an independent normal distribution with constant variance σ^2 and mean μ .

- Systematic Component

- Covariates x_1, x_2, \dots, x_p produce a linear predictor

$$\eta = \sum \beta_j x_j$$

- The Link $\eta = \mu$



Generalized Linear Model

- Consider exponential family
 - Normal distribution can be expressed in terms of an exponential family

$$f(y_i : \mu_i) = a(\mu_i)b(y_i) \exp[y_i g(\mu_i)]$$

- Replace the identity link by a monotonic differentiable function

$$\eta = g(\mu)$$

27



Exponential family

- Write Poisson as an exponential family

$$\begin{aligned} f(y : \mu) &= e^{-\mu} \frac{\mu^y}{y!} \\ &= \exp(-\mu) \left(\frac{1}{y!} \right) \exp(y \log \mu) \end{aligned}$$

\uparrow \uparrow \uparrow
 $a(\mu)$ $b(y)$ $g(\mu)$

- Natural parameter = $\log(\mu)$

$$\longrightarrow \eta = \log(\mu)$$

28



Poisson Models

- Link Function

$$\ln \mu_i = \eta = \mathbf{x}_i^T \boldsymbol{\beta}$$

- Log link ensures $\tilde{\mu} \in [0, \infty)$
- Poisson model with log link is sometimes called log-linear model
- Also, this is the canonical link

29



ML Estimation

- Assuming the data independent

$$f(y) = e^{-\mu} \mu^y / y!$$

$$\Rightarrow \ln f(y) = -\exp(\underline{x}' \underline{\beta}) + y \underline{x}' \underline{\beta} - \ln y!$$

$$\Rightarrow \ln L(\underline{\beta}) = \sum_{i=1}^n \{-\exp(\underline{x}_i' \underline{\beta}) + y_i \underline{x}_i' \underline{\beta} - \ln y_i!\}$$

$$\Rightarrow \frac{\partial \ln L(\underline{\beta})}{\partial \underline{\beta}} = \sum_{i=1}^n \{-\exp(\underline{x}_i' \underline{\beta}) \underline{x}_i + y_i \underline{x}_i\}$$

- ML estimates are solutions to

$$\sum_{i=1}^n [y_i - \exp(\underline{x}_i' \underline{\beta})] \underline{x}_i = \underline{0}$$

30



Fitting the Model

- Use IRWLS (or Fisher's scoring procedure or Newton-Raphson)

- One iteration of the algorithm

$$\beta = (X^T W X)^{-1} X^T W Z$$

- An adjusted dependent variate

$$z = \eta + (y - \mu) \frac{d\eta}{d\mu}$$

- A (quadratic) weight W

$$W^{-1} = \left[\frac{d\eta}{d\mu} \right]^2 V(\mu)$$

31



Fitting the Model

- Because $\eta = \log \mu$, it follows that

$$\text{so that } \frac{d\eta}{d\mu} = \frac{1}{\mu}$$

$$W = \text{Diag}(\mu_i)$$

$$z_i = \eta_i + (y_i - \mu_i) / \mu_i$$

- Given the current estimate for β , we

- calculate $\eta_i = X_i^T \beta$, $\mu_i = \exp(\eta_i)$, and

$$z_i = \eta_i + (y_i - \mu_i) / \mu_i$$

- regress z on X using the μ_i 's as weights to get the new estimate of β

- Repeat until the value of β converges

- Ref: *Categorical Data Analysis*, Alan Agresti

32



Diagnostics

- The LR test statistic is the deviance G^2 for comparing the fit to the saturated model that fits separate mean to each y_i and makes no assumption about its relationship to the covariates.

$$G^2 = 2 \sum_{i=1}^N \left\{ y_i \log \frac{y_i}{\mu_i} - (y_i - \mu_i) \right\}$$

$$= 2 \sum_{i: y_i=0} \left\{ y_i \log \frac{y_i}{\mu_i} - (y_i - \mu_i) \right\} + 2 \sum_{i: y_i=0} \mu_i$$

- Other quantities that may be of interest are:

- Estimated covariance matrix

$$(X^T W X)^{-1} \text{ for } \hat{\beta}$$

- Pearson residuals and Pearson GOF statistic

$$r_i = \frac{y_i - \mu_i}{\sqrt{\mu_i}}$$

$$X^2 = \sum_{i=1}^N r_i^2$$



Poisson vs Multinomial - Note

- Note that the deviance reduces to

$$G^2 = 2 \sum_{i=1}^N y_i \log \frac{y_i}{\mu_i}$$

- For a loglinear model with an intercept
- The above is also the deviance statistic for a multinomial model for the following reason:
 - We can factor the independent Poisson observations $y_i \sim \text{Poisson}(\mu_i)$, $i=1,2,\dots,N$ into
 - a Poisson distribution for $y_+ = \sum y_i$ with mean $\mu_+ = \sum \mu_i$ and
 - a multinomial distribution for $y = (y_1, \dots, y_N)^T$,

$$y | y_+ \sim \text{Multinomial}(y_+, \pi) \text{ where}$$

$$\pi = (\pi_1, \dots, \pi_N)^T \text{ and } \pi_i = \mu_i / \mu_+$$

- Assume that the observations are from multinomial as y_+ is fixed but let us pretend that the sample is from Poisson with means

$$\mu_i = \mu_+ \pi_i$$

- Under this assumption, we can fit a multinomial model as a Poisson with a log link as long as the model contains an intercept
 - When log is applied, the term μ_+ is absorbed into the intercept.
 - More: Agresti's book



Poissonness Plot

- Hoaglin and Tukey(1985), Johnson and Kotz (1969)
- Graphical display to assess whether data comes from a Poisson distribution
- Suppose Y follows a Poisson distribution

$$P(Y = k) = \frac{e^{-\lambda} \lambda^k}{k!}, \quad k = 0, 1, 2, \dots$$

- Suppose there are N observations that are iid Poisson(λ), the distribution of expected frequencies is

$$\eta_k = N.P(Y = k) = N e^{-\lambda} \lambda^k / k!, \quad k = 0, 1, 2, \dots$$

35



Poissonness Plot

- Take logarithms (to base e) on both sides

$$\log_e \eta_k = \log_e N - \lambda + k \log_e \lambda - \log_e k!$$

- A plot of

$$\log_e \eta_k + \log_e k! = \log_e (\eta_k k!) \quad \text{versus} \quad k$$

- Will be a straight line with slope $\log_e \lambda$ and intercept $\log_e N - \lambda$
- This linear relationship is exploited graphically to show whether a set of data comes from a Poisson distribution
- If the data do not correspond to a straight line, it is reasonable to say that Poisson assumption is inappropriate
 - Each unusual data count affects only its own point in the plot, so that the Poissonness plot is resistant

36

Poisson regression in SAS



- In SAS, several procedures in both STAT and ETS modules are available for Poisson regression.
 - GENMOD, GLIMMIX, COUNTREG are easy to use with standard MODEL statement
 - NLMIXED, MODEL, NLIN provide great flexibility to model count data where one can specify the log likelihood function explicitly

37

Number of Tumors



Example



- Development of mammary cancer in rats
- Ref: M.H. Gail et al, Biometrics, June 1980
- Animals injected with carcinogen and given retinyl acetate to prevent cancer for 2 months
- Those that remain cancer-free were randomized to retinoid prophylaxis (treatment, 23 rats) or control (25 rats)
- Number of tumors developed over 4 months were reported

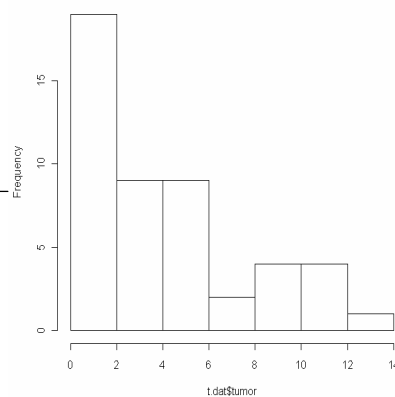
39

Summary



```
>
>
> summary(t.dat)
      trt      tumor
Min. :1.000 Min. : 0.000
1st Qu.:1.000 1st Qu.: 1.000
Median :2.000 Median : 3.500
Mean   :1.521 Mean   : 4.417
3rd Qu.:2.000 3rd Qu.: 6.000
Max.   :2.000 Max.   :13.000
> var(t.dat$tumor)
[1] 12.63121
> |
```

Histogram of t.dat\$tumor



Var > Mean

40

Poisson fit



The SAS System 16:57 Saturday, June 6, 2009 4
 The COUNTREG Procedure
 Model Fit Summary

Dependent Variable tumor
 Number of Observations 48
 Data Set WORK.DAT1
 Model Poisson
 Log Likelihood -122.63371
 Maximum Absolute Gradient 3.0518E-10
 Number of Iterations 4
 Optimization Method Newton-Raphson
 AIC 249.26743
 SBC 253.00983

Algorithm converged.

Parameter Estimates

Parameter	DF	Estimate	Standard Error	t Value	Approx Pr > t
Intercept	1	0.152355	0.268694	0.57	0.5707
trt	1	0.823024	0.151710	5.42	<.0001

```
proc countreg data=data1;
model tumor= trt/dist=poisson;
ods output ParameterEstimates=pe;
run;
```

41

PROC GENMOD



```
proc genmod data=data1;
model tumor=trt/dist=poisson;
run;
```

The GENMOD Procedure

Model Information

Data Set WORK.DAT1
 Distribution Poisson
 Link Function Log
 Dependent Variable tumor

Number of Observations Read 48
 Number of Observations Used 48

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	46	102.0922	2.2194
Scaled Deviance	46	102.0922	2.2194
Pearson Chi-Square	46	93.1254	2.0245
Scaled Pearson X ²	46	93.1254	2.0245
Log Likelihood		-122.6337	
Full Log Likelihood		-122.6337	
AIC (smaller is better)		249.2674	
AICC (smaller is better)		249.5441	
BIC (smaller is better)		253.0098	

Algorithm converged.

Analysis Of Maximum Likelihood Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits	Wald Chi-Square	Pr > ChiSq
Intercept	1	0.1524	0.2687	-0.3743 0.6790	0.32	0.5707
trt	1	0.8230	0.1517	0.5257 1.1204	29.43	<.0001
Scale	0	1.0000	0.0000	1.0000 1.0000		

NOTE: The scale parameter was held fixed.

42

Overdispersion tests



The SAS System 16:57 Saturday, June 6, 2009 5

The COUNTREG Procedure

Model Fit Summary

Dependent Variable	tumor
Number of Observations	48
Data Set	WORK.DAT01
Model	NegBin1
Log Likelihood	-114.96684
Maximum Absolute Gradient	6.0005E-9
Number of Iterations	5
Optimization Method	Newton-Raphson
AIC	235.01367
SBC	241.42728

Algorithm converged.

Parameter Estimates

Parameter	DF	Estimate	Standard Error	t Value	Approx Pr > t
Intercept	1	0.256405	0.375229	0.79	0.4286
trt	1	0.738463	0.213157	3.46	0.0005
_alpha	1	1.186654	0.499490	2.38	0.0175

```

proc countreg data=data1;
model tumor= trt/dist=negbin(p=1);
ods output ParameterEstimates=pe;
run;

```

Linear

The SAS System 16:57 Saturday, June 6, 2009 6

The COUNTREG Procedure

Model Fit Summary

Dependent Variable	tumor
Number of Observations	48
Data Set	WORK.DAT01
Model	NegBin
Log Likelihood	-113.97288
Maximum Absolute Gradient	2.09615E-7
Number of Iterations	5
Optimization Method	Newton-Raphson
AIC	232.94595
SBC	239.55955

Algorithm converged.

Parameter Estimates

Parameter	DF	Estimate	Standard Error	t Value	Approx Pr > t
Intercept	1	0.152955	0.359436	0.42	0.6717
trt	1	0.820204	0.212742	3.87	0.0001
_alpha	1	0.266453	0.107613	2.48	0.0133

```

proc countreg data=data1;
model tumor= trt/dist=negbin(p=2);
ods output ParameterEstimates=pe;
run;

```

Quadratic

PROC NL MIXED



```

proc nlmixed data=data1;
  parms b0=0 b1=0;
  mu=exp(b0+b1*trt);
  ll=-mu+tumor*log(mu)-log(fact(tumor));
  model tumor ~ general(ll);
  predict mu out=poi_out (rename=(pred=Yhat));
run;

proc print data=poi_out;
run;

```

The SAS System 11:00 Monday, June 15, 2009 2

The NL MIXED Procedure

Parameter Estimates

Parameter	Estimate	Standard Error	DF	t Value	Pr > t	Alpha	Lower	Upper	Gradient
b0	0.1524	0.2607	48	0.57	0.5733	0.05	-0.3873	0.6926	-4.88E-6
b1	0.8220	0.1517	48	5.42	<.0001	0.05	0.5180	1.1261	-1.58E-7

Seizure Counts in Epileptics



Seizure counts in Epileptics



- Thall, P. F. and Vail, S. C. (1990) Some covariance models for longitudinal count data with over-dispersion. *Biometrics* **46**, 657–671
- Data set on two-week seizure counts for 59 epileptics. The number of seizures was recorded for a baseline period of 8 weeks, and then patients were randomly assigned to a treatment group or a control group. Counts were then recorded for four successive two-week periods (response: count for the 2 week period, 236 observed counts). The subject's age is the only covariate.

Data



Format

This data frame has 236 rows and the following 9 columns:

y the count for the 2-week period.
trt treatment, "placebo" or "progabide".
base the counts in the baseline 8-week period.
age subject's age, in years.
V4 a indicator variable of period 4.
subject subject number, 1 to 59.
period period, 1 to 4.
lbase log-counts for the baseline period, centred to have zero mean.
lage log-ages, centred to have zero mean.

47

Analysis in R



```
> #load package MASS and get data epil
> data(epil)
> attach(epil)
> names(epil)
[1] "y"      "trt"    "base"   "age"    "V4"     "subject" "period"
[8] "lbase"  "lage"
> summary(epil)
      y          trt          base          age
Min.   : 0.000  placebo :112   Min.   : 6.00   Min.   :18.00
1st Qu.: 2.750  progabide:124   1st Qu.: 12.00  1st Qu.:23.00
Median : 4.000                Median : 22.00  Median :28.00
Mean   : 8.254                Mean   : 31.22   Mean   :28.34
3rd Qu.: 9.000                3rd Qu.: 41.00  3rd Qu.:32.00
Max.   :102.000               Max.   :151.00  Max.   :42.00

      V4          subject          period          lbase
Min.  :-0.000  Min.   : 1   Min.   :1.00   Min.  :-1.362e+00
1st Qu.:0.000  1st Qu.:15   1st Qu.:11.75  1st Qu.:-6.693e-01
Median :0.000  Median :30   Median :12.50  Median :-6.321e-02
Mean   :0.25   Mean   :30   Mean   :12.50   Mean  :-2.296e-16
3rd Qu.:0.25   3rd Qu.:45   3rd Qu.:13.25  3rd Qu.: 5.593e-01
Max.   :1.00   Max.   :59   Max.   :14.00   Max.   : 1.863e+00

      lage
Min.  :-4.294e-01
1st Qu.:-1.843e-01
Median : 1.242e-02
Mean   : 1.054e-16
3rd Qu.: 1.460e-01
Max.   : 4.179e-01

> var(y)
[1] 152.4457
> |
```

Histogram of y

48

Poisson fit



```
>
> summary(glm(y~lbase*trt+lage+V4,family=poisson,data=epil),cor=FALSE)

Call:
glm(formula = y ~ lbase * trt + lage + V4, family = poisson,
    data = epil)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-5.0915  -1.4126  -0.2739   0.7580  10.7711

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.89791    0.04260  44.552 < 2e-16 ***
lbase         0.94862    0.04360  21.759 < 2e-16 ***
trtprogabide -0.34588    0.06100  -5.670 1.42e-08 ***
lage         0.88760    0.11650   7.619 2.56e-14 ***
V4          -0.15977    0.05458  -2.927 0.00342 **
lbase:trtprogabide 0.56154    0.06352   8.841 < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Dispersion parameter for poisson family taken to be 1)
```

49

Overdispersion



Overdispersion for Poisson



- Count data are often over-dispersed relative to Poisson
- Variance $>$ mean [\neq Poisson]
- Overdispersion may be a result of
 - higher incidence of zero counts
 - subject heterogeneity
- Overdispersion is not an issue in ordinary regression when $Y \sim \text{Normal}$
 - normal distribution has a separate variance parameter

51

Overdispersion for Poisson



- Over-dispersion is common in counts
- Correct model for the mean with distribution \neq Poisson
 - \Rightarrow ML estimates are consistent
 - Standard Errors are incorrect

52



A Simple GOF test

- Use Fisher's dispersion test (~ Pearson GOF Test)

$$X^2 = \frac{\hat{\mu}_2}{\hat{\mu}_1} = \frac{\sum_{i=1} (y_i - \bar{y})^2}{\bar{y}}$$

- X^2 is approximately distributed as χ^2 with n-1 degrees of freedom
- Software available (SAS, S-Plus)

53



Overdispersion Tests

- Data are overdispersed if the conditional variance exceeds the conditional mean
 - Compare sample mean and variance to investigate overdispersion or underdispersion
 - In Poisson regression, (Cameron and Trivedi)
 - If the sample variance is less than the sample mean, the data are even underdispersed once covariates are included
 - If the sample variance is more than twice the sample mean, then data are likely to remain overdispersed even after covariates are included

54



Overdispersion

- Poisson is a special case of negative binomial (dispersion parameter $\kappa=0$)
 - For a given data, try both Poisson and negative binomial models
 - Test the hypothesis $H_0:\kappa=0$ using either LR or Wald test

55



Lack of Fit test

- Recall the ZIP distribution is defined as

$$f_Y(y | \lambda, p) = pI_{\{y=0\}} + (1-p) \frac{e^{-\lambda} \lambda^y}{y!}$$

with $y = 0, 1, 2, \dots$; and $\lambda > 0, 0 \leq p < 1$

- With count data, our interest is to investigate if a regular Poisson model gives an adequate fit or if the lack of fit is due to excessive zeros.
- Suppose we are interested in testing this with Poisson versus ZIP based on the hypotheses:
 $H_0:p=0$ versus $H_1:p>0$

56



Lack of Fit test

- Suppose we have a random sample $(y=y_1, y_2, \dots, y_n)$ from the ZIP distribution
- The null and the alternative hypotheses correspond to the two models

$$M_0 : Y_i \stackrel{iid}{\sim} \text{Poisson}(\lambda), \quad i = 1, 2, \dots, n$$

$$M_1 : Y_i \stackrel{iid}{\sim} f_Y(y | \lambda, p), \quad i = 1, 2, \dots, n$$

57



LOF test

- A score test proposed by van Broek (1995) is given by

$$T_S = \frac{\left(\frac{\sum_{i=1}^n I_{[y_i=0]}}{e^{-\hat{\lambda}}} - n \right)^2}{n \left[\frac{1 - e^{-\hat{\lambda}}}{e^{-\hat{\lambda}}} - \hat{\lambda} \right]}$$

- Where $\hat{\lambda}$ is MLE calculated under the null hypothesis.
- This test does not perform well when λ is large.
- A Bayesian objective testing for Poisson vs ZIP was provided by Bayyari et al (2008)

58

Comparing two Poisson Means



- Suppose consider two independent random samples as:

$$X_1 = \sum_{i=1}^{n_1} X_{1i} \sim \text{Poisson}(n_1 \lambda_1)$$

$$X_2 = \sum_{i=1}^{n_2} X_{2i} \sim \text{Poisson}(n_2 \lambda_2)$$

- Let k_1 and k_2 be the observed values of X_1 and X_2 , respectively. We want to test

$$H_0: \lambda_1 - \lambda_2 \leq d \quad \text{vs} \quad H_a: \lambda_1 - \lambda_2 > d \quad \text{OR}$$

$$H_0: \frac{\lambda_1}{\lambda_2} \leq c \quad \text{vs} \quad H_a: \frac{\lambda_1}{\lambda_2} > c$$

based on (n_1, k_1, n_2, k_2)

- Przyborowski and Wilenski (1940)

- Based on the fact that the conditional distribution of X_1 given $X_1 + X_2$ follows binomial with success probability defined as

$$p(\lambda_1 / \lambda_2) = (n_1 / n_2) (\lambda_1 / \lambda_2) / [1 + (n_1 / n_2) (\lambda_1 / \lambda_2)]$$

- Therefore, hypothesis testing and interval estimation can be readily developed using exact methods

- Other more powerful tests have been proposed since then

- Ref. Krishnamoorthy and Thomson, J of Stat Planning and Inference (2004), 23-35.

59

Modeling Overdispersion (McCullagh and Nelder, 1989)



- Overdispersion is said to exist if
 - $\text{Var}(Y_i) > E(Y_i)$
- Mixing the distributions would help with modeling overdispersion
- Assume

$$Y_i | Z_i \sim \text{Poisson}(Z_i), \quad Z_i \geq 0,$$

$$E(Z_i) = \mu_i$$

independently

60

Overdispersion



- Suppose Z_i is distributed as gamma with mean μ_i and index $\phi\mu_i$

$$f_{Z_i}(z_i) = \frac{1}{\Gamma(\phi\mu_i)} \left(\frac{\phi\mu_i z_i}{\mu_i} \right)^{\phi\mu_i} \exp\left(-\frac{\phi\mu_i z_i}{\mu_i}\right) \frac{1}{z_i}$$

$$\Rightarrow E(Z_i) = \mu_i \quad \text{Var}(Z_i) = \frac{\mu_i^2}{\phi\mu_i} = \frac{\mu_i}{\phi}$$

61

Overdispersion



- Can show that the pdf of Y as

$$P(Y = y) = \frac{\Gamma(y + \phi\mu)}{\Gamma(\phi\mu)\Gamma(y + 1)} \left(\frac{\phi}{1 + \phi} \right)^{\phi\mu} \left(\frac{1}{1 + \phi} \right)^y$$

- That is,

$$Y_i \sim \text{negbin}(a_i, b_i) \text{ with } a_i = \phi\mu_i, b_i = 1/\phi$$

$$E(Y_i) = \frac{\phi\mu_i}{\phi} = \mu_i \text{ and } \text{Var}(Y_i) = \frac{\phi\mu_i}{\phi} (1 + 1/\phi) = \mu_i (1 + 1/\phi)$$

linear

62

Overdispersion



- Other choices:

$$Z_i \sim \text{Gamma}(\mu_i, \nu)$$

$$\Rightarrow Y_i \sim \text{Negbin}(\nu, \mu_i / \nu)$$

$$E(Y_i) = \mu_i, \quad \text{Var}(Y_i) = \mu_i + \mu_i^2 / \nu$$

$\Rightarrow \text{var}(Y)$ is a quadratic function in μ

- Multiplicative heterogeneity

$$Y_i | Z_i \sim \text{Poisson}(\mu_i Z_i), \quad E(Z_i) = 1$$

$$Z_i \sim \text{Gamma}(1, \nu)$$

$$\Rightarrow Y_i \sim \text{Negbin}(\nu, \mu_i / \nu)$$

$$E(Y_i) = \mu_i \quad \text{Var}(Y_i) = \mu_i + \mu_i^2 / \nu$$

quadratic ↑

63

Test for Overdispersion



- Dean (1992), JASA
- In this, we assume

$$Y_i \sim \text{Poisson}(\mu_i) \text{ with } \mu_i = \exp(x_i^t \beta)$$

$$\Rightarrow \delta_i = \ln(\mu_i) = x_i^t \beta$$

- For modeling overdispersion, Dean assumed that the canonical parameters δ_i are not fixed but random δ_i^* with

$$E(\delta_i^*) = \delta_i$$

$$\text{Var}(\delta_i^*) = \tau k_i(\delta_i) > 0 \text{ for } \tau \geq 0 \text{ and } k_i(\delta_i) \text{ differentiable}$$

64



Test for Overdispersion

- To test for overdispersion, now we want to test

$$H_0: \tau=0 \quad \text{vs} \quad H_1: \tau>0$$

- Example

$$\begin{aligned} \delta_i^* &= x_i^t \beta + Z_i \text{ with } Z_i \text{ i.i.d. } E(Z_i) = 0, \text{Var}(Z_i) = \tau < \infty \\ \Rightarrow E(\delta_i^*) &= \delta_i \end{aligned}$$

- That is $\text{Var}(\delta_i^*) = \tau \quad k_i(\delta_i) = 1$

$$Y_i | \delta_i^* \sim \text{Poisson}(\exp(\delta_i^*)) \quad \delta_i = \ln \mu_i = x_i^t \beta$$

65



Poisson model with random effects

- Now we can show that

$$E(Y_i) = E_{\delta_i^*} [E(Y_i | \delta_i^*)] = E_{\delta_i^*} (e^{\delta_i^*})$$

$$\begin{aligned} &\approx E_{\delta_i^*} (1 + \delta_i^*) = 1 + \delta_i \\ \text{Taylor series} &\nearrow \\ \text{approx} &\rightarrow \\ &= 1 + \ln \mu_i \approx 1 + \mu_i - 1 = \mu_i \end{aligned}$$

66



With random effects

- Also we can show that

$$\begin{aligned} \text{Var}(Y_i) &= E_{\delta_i^*} [\text{Var}(Y_i | \delta_i^*)] + \text{Var}_{\delta_i^*} [E(Y_i | \delta_i^*)] \\ &= E(Y_i) + [e^{\delta_i}]^2 \text{Var}(e^{Z_i}) \\ &\approx \mu_i + \mu_i^2 \tau \end{aligned}$$

67



Test for Overdispersion

- Dean (1992) derived test statistics of the form for the quadratic and linear function variance function, respectively,

$$T_q = \frac{\sum_{i=1}^n [(y_i - \hat{\mu}_i)^2 - \hat{\mu}_i]}{\sqrt{2 \sum_{i=1}^n \hat{\mu}_i^2}}$$
$$T_l = \frac{\sum_{i=1}^n \frac{(y_i - \hat{\mu}_i)^2 - y_i}{\hat{\mu}_i}}{\sqrt{2n}}$$

- Cameron and Trivedi (1986) had proposed a similar statistic for Poisson model vs negative binomial

68

Models for Overdispersion



- Negative Binomial
- Zero-inflated Poisson
- Others..

69

Negative Binomial Model



- Suppose
 - $y|\lambda \sim \text{Poisson}(\lambda)$
 - $\lambda \sim \text{Gamma}(\alpha, \beta)$
 - $y \sim \text{Negative Binomial}$
- Probability mass function

$$\Rightarrow P(y) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha) y!} \left(\frac{\beta}{1 + \beta} \right)^y \left(\frac{1}{1 + \beta} \right)^\alpha \quad y = 0, 1, \dots$$

70

Negative Binomial Properties



- For this,

$$\begin{aligned} E(y) &= E[E(y|\lambda)] \\ &= E(\lambda) = \alpha\beta \\ \text{Var}(y) &= E[\text{Var}(y|\lambda)] + \text{Var}[E(y|\lambda)] \\ &= \text{Var}(\lambda) + E(\lambda) \\ &= \alpha\beta^2 + \alpha\beta \end{aligned}$$

71

Properties (contd)



- Reparameterizing $\mu = \alpha\beta$, $\kappa=1/\alpha$
 $E(y) = \mu$, $\text{Var}(y) = \mu + \kappa\mu^2$
- Note that variance function is quadratic
- Density can be rewritten as

$$P(y) = \frac{\Gamma(\kappa^{-1} + \kappa\mu)}{\Gamma(\kappa^{-1})y!} \left(\frac{\kappa\mu}{1 + \kappa\mu}\right)^y \left(\frac{1}{1 + \kappa\mu}\right)^{1/\kappa}$$

- Also, note that
 $\text{Var}[Y]/E[Y] = 1 + \kappa E[Y]$

72



Properties (contd)

- The index κ is called dispersion parameter
- For a given κ , negative binomial is in the natural exponential family
- The natural parameter is

$$\theta = \log \left(\frac{\kappa\mu}{1 + \kappa\mu} \right)$$

73



Remarks

- The greater κ , the greater the over-dispersion compared to the ordinary Poisson GLM
- If κ were known, we could fit the model by IRWLS to obtain the MLE of β
- Estimating κ is problematic, need to jointly estimate κ and β

74

When there are excessive zeros?



Zero-Inflated Poisson (ZIP)

- Model count data with excess zero (overdispersion)
- Model y_i as a mixture:

$$Y = \begin{cases} 0, & \text{with prob } p \\ \text{Poisson } (\mu) & \text{with prob } 1-p \end{cases}$$

76

Zero-Inflated Poisson (ZIP)



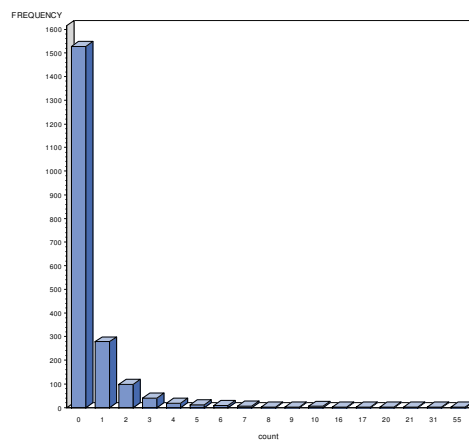
- The Y has a ZIP distribution given by

$$\Pr(Y = y) = \begin{cases} p + (1-p)\exp(-\mu) & y=0 \\ (1-p)\exp(-\mu)\mu^y / y! & y>0 \end{cases}$$

- For ease of presentation, let $p=1-\pi$

77

ZIP distribution



78



ZIP Properties

- For the ZIP model,

$$E[Y] = \pi\mu$$

$$\text{Var}[Y] = \pi\mu + \pi\mu^2(1 - \pi)$$

$$\text{Var}[Y]/E[Y] = 1 + [(1-\pi)/\pi] E[Y]$$

- Problem in distinguishing the ZIP model from NB as the source of the over-dispersion

79



ZIP Parameter Estimation

- Method of Moments
 - Solving the equations

$$E[Y] = \bar{Y}$$

$$S^2 = E(Y)[1 + \mu - E(Y)]$$

we get

$$\hat{\mu}_{MO} = S^2 / \bar{Y} - 1 + \bar{Y}$$

$$\hat{\pi}_{MO} = \bar{Y} / \hat{\mu}_{MO}$$

80

ZIP Parameter Estimation



- Maximum Likelihood
 - Let n_i be the number of i 's in the sample; in particular, n_0 is the number of zeros in the sample. Write the log-likelihood as

$$L(\pi, \mu) = n_0 \log[1 - \pi + \pi e^{-\mu}] + \sum_{y=1}^m n_y \log[\pi P_0(y, \mu)]$$

81

ZIP Parameter Estimation



- Score vector is

$$\left[n_0 \frac{e^{-\mu} - 1}{1 - \pi + \pi e^{-\mu}} + (n - n_0) / \pi, n_0 \frac{-\pi e^{-\mu}}{1 - \pi + \pi e^{-\mu}} - (n - n_0) + n \bar{Y} / \mu \right]^T$$

82

ZIP Parameter Estimation



- As a result, the score equations are

$$\pi = \frac{1 - n_0/n}{1 - e^{-\mu}} \quad \mu = \bar{Y} / \pi$$

- Which can be written in one equation as

$$\mu = \frac{\bar{Y}}{(1 - n_0/n) / (1 - e^{-\mu})} \equiv G(\mu)$$

83

Estimation



- Need iteration to solve this equation
- Since $G'(\mu) > 0$, $\mu_{j+1} = G(\mu_j)$ converges for any initial value μ_0 to the MLE satisfying the equation $\mu = G(\mu)$.
- Choose the moment estimate as the initial value
- Can use EM Algorithm
 - Ref: Lambert, D. (1992), Technometrics

84

Note



- Note that ZIP is a special case of mixture model and overdispersion is usually modeled through random effects or hierarchical models.
- Efron (1986) introduced a double exponential family of distributions to account for overdispersion.
 - It extends a regular one-parameter exponential family by introducing a second parameter that controls variance independently of the mean.
 - It provides a greater flexibility over one-parameter exponential distribution where variance is a function of the mean

85

Zero-inflated negative binomial (ZINB)



- Like ZIP, ZINB is also a mixture distribution that assigns a mass of p to 'extra' zeros and a mass of $(1-p)$ to a negative binomial distribution.
 - Recall that a negative binomial distribution is a continuous mixture of Poisson distributions, which allows the Poisson mean λ to be gamma distributed.
- Recall the negative binomial distribution

$$P(Y = y) = \frac{\Gamma(y + \tau)}{y! \Gamma(\tau)} \left(\frac{\tau}{\lambda + \tau} \right)^\tau \left(\frac{\lambda}{\lambda + \tau} \right)^y, \quad y = 0, 1, \dots; \lambda, \tau > 0$$

86

Zero-inflated NB



- Summary NB

$$E(Y)=\lambda, \text{Var}(Y)=\lambda+\lambda^2/\tau$$

τ : shape parameter that quantifies the amount of overdispersion

NB \rightarrow Poisson when $\tau \rightarrow \infty$ (no overdispersion)

- ZINB can now be defined as

$$P(Y=y)=\begin{cases} p+(1-p)(1+\lambda/\tau)^{-\tau}, & y=0 \\ (1-p)\frac{\Gamma(y+\tau)}{y!\Gamma(\tau)}(1+\lambda/\tau)^{-\tau}(1+\tau/\lambda)^{-y}, & y=1,2,\dots \end{cases}$$

87

Summary of ZINB



- The mean and variance of ZINB are

$$E(Y) = (1-p)\lambda$$

$$\text{Var}(Y) = (1-p)\lambda(1+p\lambda + \lambda/\tau)$$

- Note

- ZINB approaches ZIP and NB as $\tau \rightarrow \infty$ and $p \rightarrow 0$
- If both $1/\tau$ and $p \approx 0$, then ZINB reduces to Poisson

- Also, note that

- ZINB also arises in Bernoulli trials with non-equal success probabilities
- The overdispersed data are characterized by “excess zeros”, “excess large outcomes” or both.
 - ZINB model therefore accounts for “excess zeros” and also for extra heterogeneity in a positive outcome

88

Other models



- 'Hurdle' or Zero-Altered Poisson (ZAP)
 - ZIP generally depends heavily on the assumed distributional shape for the non-zero component.
 - Instead, let us separate the zero part and define

$$P(Y=y) = \begin{cases} \pi_0, & y=0 \\ \frac{(1-\pi_0)e^{-\mu}\mu^y}{(1-e^{-\mu})^y}, & y>0 \end{cases}$$

- 'Hurdle' models is just a reparametrization of the ZIP model with

$$\pi_0 = p + (1-p)e^{-\mu}$$

- Truncated Poisson
- Same as the second part for $y>0$

$$P(Y=y) = (e^\lambda - 1)^{-1} \lambda^y / y!, \quad y=1,2,\dots$$

- Example: number of cholera cases in a household where the event $Y=0$ is not observable since the observational apparatus (ie, diagnosis) is activated only when $Y>0$
- Shanmugam (Biometrics, 1985) describes the following scenario:
 - Assuming that various preventive treatments are used for the cholera incidence could result in the effect λ changing from one incidence to another
 - Assume that such effects change from λ to $\rho\lambda$

89

Other models



- So, we can have $E(Z)=\rho\lambda$, $0 \leq \rho < 1$ is the intervention parameter and Z is the number of cases that occurred after the prevention is applied.
 - Z is Poisson with mean $\rho\lambda$, and Y and Z are stochastically independent

- Suppose the apparatus keeps only the total number of cholera cases that occurred all together, ie, the r.v $X=Y+Z$

$$P(X=x) = \sum_{l=0}^{x-1} P(Y=x-l)P[Z=l|Y=x-l]$$

$$= [e^{\rho\lambda}(e^\lambda - 1)]^{-1} [(1+\rho)^\lambda - \rho^\lambda] \lambda^x / x!, \quad x=1,2,\dots$$

- The above is called the intervened Poisson distribution (IPD)

- The mean and variance of IPD are

$$\mu = E(X) = \lambda[\rho + 1 + (e^\lambda - 1)^{-1}]$$

$$\sigma^2 = Var(X) = \mu - e^\lambda \left(\frac{\lambda}{e^\lambda - 1} \right)^2$$

- One of the characteristics of IPD is that the variance is less than the mean, different from Poisson.

Will not be discussed further.

90

Example



Australian Health Survey



- Reference
 - A.C. Cameron and Pravin K. Trivedi (1998), REGRESSION ANALYSIS OF COUNT DATA, Chapter 3
 - Dataset (name: dvisits) available in the **R Package**, *faraway*
- Description
 - The data come from the Australian Health Survey of 1977-78 and consist of 5190 single adults where young and old have been oversampled.

Australian Health Study



Variables:

sex = 1 if respondent is female, 0 if male
age = respondent's age in years divided by 100,
agesq = age squared
income = respondent's annual income in Australian dollars divided by 1000
levyplus = 1 if respondent is covered by private health insurance fund for private patient in public hospital (with doctor of choice), 0 otherwise
freepoor = 1 if respondent is covered by government because low income, recent immigrant, unemployed, 0 otherwise
freerepa = 1 if respondent is covered free by government because of old-age or disability pension, or because invalid veteran or family of deceased veteran, 0 otherwise
illness = number of illnesses in past 2 weeks with 5 or more coded as 5
actdays = number of days of reduced activity in past two weeks due to illness or injury
hscore = respondent's general health questionnaire score using Goldberg's method, high score indicates bad health.
chcond1 = 1 if respondent has chronic condition(s) but not limited in activity, 0 otherwise
chcond2 = 1 if respondent has chronic condition(s) and limited in activity, 0 otherwise
dvists = number of consultations with a doctor or specialist in the past 2 weeks
nondocco = number of consultations with non-doctor health professionals, (chemist, optician, physiotherapist, social worker, district community nurse, chiroprapist or chiropractor in the past 2 weeks
hospadmi = number of admissions to a hospital, psychiatric hospital, nursing or convalescent home in the past 12 months (up to 5 or more admissions which is coded as 5)
hospdays = number of nights in a hospital, etc. during most recent admission, in past 12 months
medicine = total number of prescribed and nonprescribed medications used in past 2 days
prescribe = total number of prescribed medications used in past 2 days
nonprescribe = total number of nonprescribed medications used in past 2 days
constant = 1 for all observations
id = ij

93

Summary

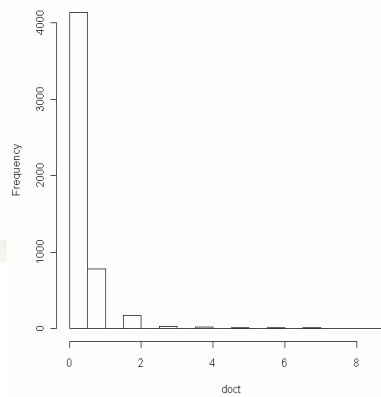


```

R >>> (www)
> table(doct)
doct
 0    1    2    3    4    5    6    7    8    9
4141 782 174  30  24   9  12  12   5   1
> |

```

Histogram of doct



94

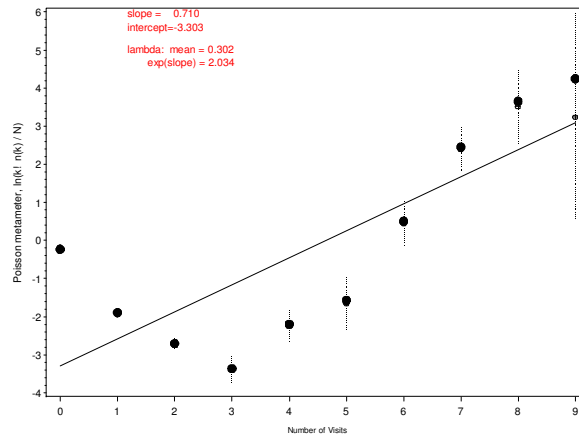
Summary



```
> summary(dvis)
      sex          age          agesq          inc
Min.   :0.0000   Min.   :0.1900   Min.   :0.0361   Min.   :0.0000
1st Qu.:0.0000   1st Qu.:0.2200   1st Qu.:0.0484   1st Qu.:0.2500
Median :1.0000   Median :0.3200   Median :0.1024   Median :0.5500
Mean   :0.5206   Mean   :0.4064   Mean   :0.2071   Mean   :0.5832
3rd Qu.:1.0000   3rd Qu.:0.6200   3rd Qu.:0.3844   3rd Qu.:0.9000
Max.   :1.0000   Max.   :0.7200   Max.   :0.5184   Max.   :1.5000
      lev1      freep      freer      ill
Min.   :0.0000   Min.   :0.00000   Min.   :0.0000   Min.   :0.000
1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0.0000   1st Qu.:0.000
Median :0.0000   Median :0.00000   Median :0.0000   Median :1.000
Mean   :0.4428   Mean   :0.04277   Mean   :0.2102   Mean   :1.432
3rd Qu.:1.0000   3rd Qu.:0.00000   3rd Qu.:0.0000   3rd Qu.:2.000
Max.   :1.0000   Max.   :1.00000   Max.   :1.0000   Max.   :5.000
      act          hscore          chi          ch2
Min.   :0.0000   Min.   :0.000   Min.   :0.00000   Min.   :0.0000
1st Qu.:0.0000   1st Qu.:0.000   1st Qu.:0.00000   1st Qu.:0.0000
Median :0.0000   Median :0.000   Median :0.00000   Median :0.0000
Mean   :0.8618   Mean   :1.218   Mean   :0.4031   Mean   :0.1166
3rd Qu.:0.0000   3rd Qu.:2.000   3rd Qu.:1.00000   3rd Qu.:0.0000
Max.   :14.0000   Max.   :12.000   Max.   :1.00000   Max.   :1.0000
      doct          nond          hospa          hospd
Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
Median :0.0000   Median :0.000   Median :0.00000   Median :0.0000
Mean   :0.3017   Mean   :0.2146   Mean   :0.1736   Mean   :1.334
3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:0.00000   3rd Qu.:0.0000
Max.   :9.0000   Max.   :11.0000   Max.   :5.00000   Max.   :80.000
      med          pres          nonp
Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
Median :1.0000   Median :0.0000   Median :0.0000
Mean   :1.218   Mean   :0.8626   Mean   :0.3557
3rd Qu.:2.000   3rd Qu.:1.0000   3rd Qu.:1.0000
Max.   :8.000   Max.   :8.0000   Max.   :8.0000
```

95

Poissonness plot



Doct: mean=0.3017, var=0.6241

96

NB fit



Parameter Estimates					
Parameter	DF	Estimate	Standard Error	t Value	Approx Pr > t
Intercept	1	-2.107751	0.232692	-9.06	<.0001
sex	1	0.169026	0.069777	2.42	0.0154
age	1	-0.013224	1.277089	-0.01	0.9917
agesq	1	-0.110280	1.404870	-0.08	0.9374
inc	1	-0.093596	0.107453	-0.87	0.3837
levy	1	0.062992	0.084574	0.74	0.4564
freep	1	-0.536811	0.206685	-2.60	0.0094
freer	1	0.092103	0.116371	0.79	0.4287
ill	1	0.164760	0.025515	6.46	<.0001
act	1	0.123367	0.008061	15.30	<.0001
hscore	1	0.028924	0.013854	2.09	0.0368
chl	1	0.037266	0.079028	0.47	0.6372
ch2	1	-0.000030611	0.107066	-0.00	0.9998

The COUNTREG Procedure

Parameter Estimates					
Parameter	DF	Estimate	Standard Error	t Value	Approx Pr > t
nond	1	0.021959	0.024523	0.90	0.3705
hospa	1	0.190922	0.056357	3.39	0.0007
hospd	1	0.001208	0.004558	0.26	0.7910
med	0	0	.	.	.
pres	1	0.166831	0.023327	7.15	<.0001
nonp	1	-0.110321	0.046011	-2.40	0.0165
_Alpha	1	0.964182	0.097004	9.94	<.0001

ZIP fit



Parameter Estimates					
Parameter	DF	Estimate	Standard Error	t Value	Approx Pr > t
Intercept	1	-0.593509	0.242177	-2.45	0.0143
sex	1	-0.062946	0.067859	-0.93	0.3536
age	1	1.184510	1.245520	0.95	0.3416
agesq	1	-1.648888	1.314903	-1.25	0.2098
inc	1	-0.212478	0.108118	-1.97	0.0494
levy	1	-0.138410	0.094047	-1.47	0.1411
freep	1	-0.278151	0.224995	-1.24	0.2164
freer	1	-0.245387	0.111968	-2.19	0.0284
ill	1	0.035695	0.024420	1.46	0.1438
act	1	0.075328	0.005766	13.06	<.0001
hscore	1	0.015379	0.010959	1.40	0.1605

The COUNTREG Procedure

Parameter Estimates					
Parameter	DF	Estimate	Standard Error	t Value	Approx Pr > t
chl	1	-0.171235	0.089178	-1.92	0.0548
ch2	1	-0.283552	0.100262	-2.83	0.0047
nond	1	0.011274	0.015782	0.71	0.4750
hospa	1	0.202633	0.032929	6.15	<.0001
hospd	1	-0.001956	0.003002	-0.65	0.5146
med	0	0	.	.	.
pres	1	0.077419	0.018993	4.09	<.0001
nonp	1	-0.105372	0.042348	-2.49	0.0156
Inf_Intercept	1	1.830012	0.638748	2.91	0.0036
Inf_sex	1	-0.208671	0.184298	-1.13	0.2575
Inf_age	1	5.204481	3.695789	1.41	0.1591
Inf_agesq	1	-6.484050	4.208518	-1.54	0.1234
Inf_inc	1	-0.340724	0.287680	-1.18	0.2363
Inf_levy	1	-0.585029	0.206913	-2.83	0.0047
Inf_freep	1	0.501557	0.491142	1.02	0.3072
Inf_freer	1	-1.244545	0.347508	-3.58	0.0003
Inf_ill	1	-0.228309	0.085088	-2.66	0.0081
Inf_act	1	1.027950	0.236582	4.35	<.0001
Inf_hscore	1	-0.081114	0.043261	-1.87	0.0609
Inf_chl	1	-0.031744	0.210691	-0.15	0.8802
Inf_ch2	1	-0.319126	0.359803	-0.89	0.3751
Inf_nond	1	-0.275644	0.200651	-1.37	0.1695
Inf_hospa	1	-0.002431	0.208514	-0.01	0.9907
Inf_hospd	1	-0.046830	0.031046	-1.51	0.1314
Inf_med	0	0	.	.	.
Inf_pres	1	-1.183970	0.186891	-6.33	<.0001
Inf_nonp	1	0.073534	0.128208	0.61	0.5407

ZINB fit



Parameter Estimates				
Parameter	DF	Estimate	Standard Error	Approx Pr > t
Intercept	1	-0.593509	0.242198	-2.45 0.0143
sex	1	-0.062946	0.067859	-0.93 0.3536
age	1	1.184514	1.245609	0.95 0.3416
agesq	1	-1.648893	1.315001	-1.25 0.2099
inc	1	-0.212478	0.108117	-1.97 0.0494
levy	1	-0.138410	0.094045	-1.47 0.1411
freep	1	-0.278151	0.224995	-1.24 0.2164
freer	1	-0.245387	0.111966	-2.19 0.0284
ill	1	0.035695	0.024420	1.46 0.1438
act	1	0.075328	0.005766	13.07 <.0001
hscore	1	0.015379	0.010959	1.40 0.1605

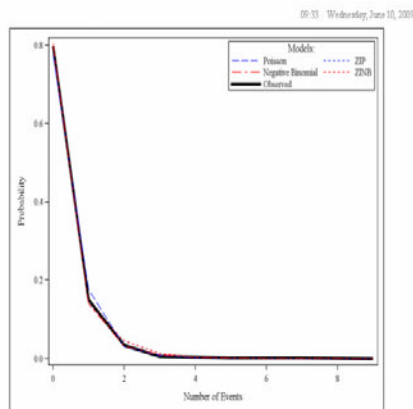
Parameter Estimates					
Parameter	DF	Estimate	Standard Error	t Value	Approx Pr > t
ch1	1	-0.171235	0.089179	-1.92	0.0548
ch2	1	-0.283552	0.100268	-2.83	0.0047
mond	1	0.011274	0.015782	0.71	0.4750
hoopa	1	0.202633	0.032929	6.15	<.0001
hoopd	1	-0.001958	0.003002	-0.65	0.5146
msd	0
psos	1	0.077419	0.018993	4.08	<.0001
msop	1	-0.102372	0.042348	-2.42	0.0156
Inf Intercept	1	1.030009	0.620804	2.91	0.0036
Inf_sex	1	-0.206671	0.104298	-1.13	0.2575
Inf_age	1	5.204506	3.696036	1.41	0.1591
Inf_agesq	1	-6.484079	4.208786	-1.54	0.1234
Inf_inc	1	-0.340723	0.287679	-1.18	0.2363
Inf_levy	1	-0.385020	0.205911	-2.83	0.0047
Inf_freep	1	0.501557	0.491141	1.02	0.3072
Inf_freer	1	-1.244545	0.347507	-3.58	0.0003
Inf_ill	1	-0.328309	0.085088	-3.86	0.0001
Inf_act	1	-1.027950	0.236583	-4.34	<.0001
Inf_hscore	1	-0.081114	0.043282	-1.87	0.0609
Inf_ch1	1	-0.031744	0.210691	-0.15	0.8802
Inf_ch2	1	-0.319126	0.359014	-0.89	0.3751
Inf_mond	1	-0.275644	0.200652	-1.37	0.1695
Inf_hoopa	1	-0.002431	0.200513	-0.01	0.9907
Inf_hoopd	1	-0.046830	0.031046	-1.51	0.1314
Inf_msd	0
Inf_psos	1	-1.183969	0.186951	-6.33	<.0001
Inf_msop	1	0.075334	0.120208	0.61	0.5407
Alpha	0	1.0536712E8	.	.	.
Restrict1	-1	1012.289132	.	.	.

Predicted zeros, fit



Obs	mu0	mu1	mu2	mu3	mu4	mu5	mu6	mu7	mu8	mu9
1	0.77638	0.17740	0.031245	0.007974	0.03177021	0.01565531	0.00857010	0.00306904	0.00210454	0.00207730
2	0.80472	0.14025	0.031555	0.010041	0.04454239	0.02435139	0.01549239	0.01039288	0.00780374	0.00566236
3	0.79452	0.14013	0.045595	0.013121	0.04104993	0.01486162	0.00823662	0.00211033	0.00116079	0.00055941
4	0.79452	0.14013	0.045595	0.013121	0.04104993	0.01486162	0.00823662	0.00211033	0.00116079	0.00055941

Predicted zero counts



Fit with a subset of variables



Model Fit Summary	
Dependent Variable	doct
Number of Observations	5190
Data Set	WORK.DOCVIS
Model	Poisson
Log Likelihood	-3675
Maximum Absolute Gradient	4.92654E-6
Number of Iterations	5
Optimization Method	Newton-Raphson
AIC	7360
SBC	7392

Algorithm converged.

Parameter Estimates					
Parameter	DF	Estimate	Standard Error	t Value	Approx Pr > t
Intercept	1	-1.855552	0.074545	-24.89	<.0001
sex	1	0.235583	0.054362	4.33	<.0001
inc	1	-0.242095	0.077829	-3.11	0.0019
ill	1	0.270326	0.017080	15.83	<.0001
hscore	1	0.096313	0.009089	10.60	<.0001

Model Fit Summary	
Dependent Variable	doct
Number of Observations	5190
Data Set	WORK.DOCVIS
Model	ZIP
ZI Link Function	Logistic
Log Likelihood	-3501
Maximum Absolute Gradient	3.37733E-9
Number of Iterations	5
Optimization Method	Newton-Raphson
AIC	7013
SBC	7061

Algorithm converged.

Parameter Estimates					
Parameter	DF	Estimate	Standard Error	t Value	Approx Pr > t
Intercept	1	-1.033387	0.096973	-10.66	<.0001
sex	1	0.122511	0.062566	1.96	0.0502
inc	1	-0.143945	0.087810	-1.64	0.1012
ill	1	0.237478	0.019997	11.88	<.0001
hscore	1	0.088386	0.010043	8.80	<.0001
Inf_Intercept	1	0.986537	0.131339	7.51	<.0001
Inf_age	1	-2.090924	0.270580	-7.73	<.0001

101

Estimate probabilities in various age groups



- Estimate of probabilities

$$20 \text{ yrs} : \frac{\exp(0.99 - 2.09 * 20)}{1 + \exp(0.99 - 2.09 * 20)} = 0.64$$

$$50 \text{ yrs} : \frac{\exp(0.99 - 2.09 * 50)}{1 + \exp(0.99 - 2.09 * 50)} = 0.49$$

$$70 \text{ yrs} : \frac{\exp(0.99 - 2.09 * 70)}{1 + \exp(0.99 - 2.09 * 70)} = 0.38$$

- Estimated probability of belonging to the low-risk group for a 20-year old person is about 0.64 etc.

102

Poisson vs Negative Binomial



Parameter Estimates					Parameter Estimates						
Parameter	DF	Estimate	Standard Error	t Value	Approx Pr > t	Parameter	DF	Estimate	Standard Error	t Value	Approx Pr > t
Intercept	0	-1.954725	.	.	.	Intercept	1	-2.107751	0.232632	-9.06	<.0001
sex	1	0.127584	0.034422	3.71	0.0002	sex	1	0.163026	0.063777	2.42	0.0154
age	0	-0.249180	.	.	.	age	1	-0.013224	1.277089	-0.01	0.9317
agesq	0	0.215886	.	.	.	agesq	1	-0.110280	1.404870	-0.08	0.9374
inc	1	-0.058453	0.025043	-2.33	0.0196	inc	1	-0.093536	0.107453	-0.87	0.3837
levy	1	0.075249	0.015880	4.74	<.0001	levy	1	0.062392	0.084574	0.74	0.4564
freep	1	-0.277249	0.161328	-1.71	0.0859	freep	1	-0.536811	0.206685	-2.60	0.0094
freer	1	0.064043	0.060920	1.05	0.2931	freer	1	0.092103	0.116371	0.79	0.4287
ill	1	0.135396	0.018716	7.27	<.0001	ill	1	0.164760	0.025515	6.46	<.0001
act	1	0.110723	0.005434	20.38	<.0001	act	1	0.123367	0.008061	15.30	<.0001
hscore	1	0.018253	0.010214	1.79	0.0739	hscore	1	0.028924	0.013854	2.09	0.0368
chl	1	0.067572	0.058553	1.15	0.2485	chl	1	0.037266	0.079028	0.47	0.6372
ch2	1	-0.028626	0.080854	-0.35	0.7233	ch2	1	-0.000030611	0.107066	-0.00	0.9998
nond	1	0.022442	0.015203	1.48	0.1399	nond	1	0.021953	0.024523	0.90	0.3705
hospa	1	0.159421	0.031224	5.11	<.0001	hospa	1	0.130922	0.056357	3.39	0.0007
hospd	1	0.000398	0.002908	0.14	0.8912	hospd	1	0.001208	0.004558	0.26	0.7910
med	0	0	.	.	.	med	0	0	.	.	.
pres	1	0.134586	0.015655	8.60	<.0001	pres	1	0.166831	0.023327	7.15	<.0001
nonp	1	-0.086398	0.037003	-2.35	0.0187	nonp	1	-0.110321	0.046011	-2.40	0.0165
						Alpha	1	0.964182	0.037004	9.34	<.0001

```

proc countreg data=docvis;
model doct= sex age agesq inc levy freep freer ill act hscore chl ch2 nond hospa hospd med pres nonp/dist=negbin(p=2);
ods output ParameterEstimates=pe;
run;

```

Sample Size Considerations



Sample Size for Discrete Distributions



- If the endpoint is count data then that should be taken into consideration for sample size calculation
- Most commonly used sample size software (nQuery, PASS 2002) do not have the options for discrete data. Poisson regression is available in PASS 2002.
- If normal approximation is used then the sample size estimates might be too high
 - Increases cost and time of subject recruitment
- Ref: Gerald van Belle, *Statistical Rules of Thumb*, Wiley

105

Sample Size



- The usual formula for sample size required to compare two population means with a common variance is

$$n = \frac{2(z_{1-\alpha/2} + z_{1-\beta})^2}{\left(\frac{\mu_0 - \mu_1}{\sigma}\right)^2}$$

- Note that for $\alpha=0.05$, $\beta=0.20$, the numerator of the equation is

$$2(1.96 + 0.84) = 15.68 \approx 16$$

- Thus the equation is approximated to

$$n = \frac{16}{\Delta^2} \text{ where } \Delta = \frac{\mu_0 - \mu_1}{\sigma} \leftarrow \text{Standardized difference}$$

106

Sample Size



- Now, suppose

$$Y_i \sim \text{Poisson}(\lambda_i), \quad i = 0, 1$$

$$\Rightarrow \sqrt{Y_i} \overset{\text{approx}}{\sim} N(\mu_i, \sigma^2) \quad \text{where}$$

$$\mu_i = \sqrt{\lambda_i}, \quad \sigma^2 = 0.25$$

- Variance stabilizing transformation (*McCullagh and Nelder*)
- Using this in the sample size formula, we get

$$n = \frac{4}{(\sqrt{\lambda_0} - \sqrt{\lambda_1})^2}$$

107

Sample Size



- The denominator can also be written as

$$\frac{\lambda_0 + \lambda_1}{2} - \sqrt{\lambda_0 \lambda_1}$$

- Which is always positive!
 - Difference between arithmetic mean (AM) and geometric mean (GM)
 - $AM > GM$ (Jensen's inequality)

108



Rate per unit time

- It is also common (eg, epidemiology) to consider the Poisson rates as rates per unit time, where we observe the data for a period of time T.
- Under this scenario,

$$Y_i \sim \text{Poisson}(\lambda_i T), \quad i = 0, 1$$

$$\Rightarrow n = \frac{4}{T(\sqrt{\lambda_0} - \sqrt{\lambda_1})^2}$$

Increasing the observation period T reduces the sample size proportionately!

109



With background rate

- Suppose we are interested in discussing rates over and above certain background rate, say, λ^* .
- Now the sample size formula can be appropriate changed to

$$n = \frac{4}{(\sqrt{\lambda^* + \lambda_0} - \sqrt{\lambda^* + \lambda_1})^2}$$

- Example
 - With means 1 and 2, you would need n=24. With background rate of 1.5, this sample size would be increased to 48.
 - Sample size is doubled when the background rate is halfway between two rates

110

Sample Size – Other options



- Instead of normal approximation, some authors have suggested to use non parametric approach.
- Example: publications about Multiple Sclerosis trials suggest sample size calculation based on Wilcoxon rank sum test. Software is available for calculating these estimates.

111

Table: Sample size per treatment group for selected power and treatment effect



Means (var)	Power 80%		Power 90%	
	Normal	Non Parametric	Normal	Non Parametric
13, 7 (21)	194	205	259	274
13, 4 (21)	87	91	116	122
8, 4 (10)	100	109	133	145
8, 2 (45)	45	52	60	69

112

Sample Size Using Negative Binomial Approach



- Since Multiple Sclerosis (MS) Phase II trials deal with discrete endpoint, authors have explored more new approaches.
- A parametric simulation procedure based on negative binomial distribution was used by M.P. Sormani and et al (2001).
- Authors provided sample size for several designs and types of MS for future studies.

113

Sample Size Using Negative Binomial Approach (cont)



- Methods used
 - Sormani et. al. used “Mixed Poisson Model” model.
 - Negative Binomial model is a Poisson process where the mean is random, follows a Gamma distribution.
 - Treatment effect was expressed as percentage difference between treated and untreated subjects.

114



Sormani's model

- As we have mentioned before,
 $y|\lambda \sim \text{Poisson}(\lambda)$
 $\lambda \sim \text{Gamma}(\alpha, \beta)$
 $y \sim \text{Negative Binomial}$
- Probability mass function of y is

$$P(y) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)} \frac{\beta^y}{(1 + \beta)^y} \left(\frac{1}{1 + \beta} \right)^\alpha \quad y = 0, 1, \dots$$

115



Sormani's model (cont)

- Therefore,
 $E(y) = E[E(y|\lambda)]$
 $= E(\lambda) = \alpha\beta$
 $= \mu$
 $\text{Var}(y) = E[\text{Var}(y|\lambda)] + \text{Var}[E(y|\lambda)]$
 $= \text{Var}(\lambda) + E(\lambda)$
 $= \alpha\beta^2 + \alpha\beta$
 $= \mu + \kappa\mu^2, \quad \kappa = 1/\alpha$

116

Sormani's model (cont)



- Data obtained from the placebo arms of large scale trials where subjects are monthly scanned.
- We present the case of RRMS population, parallel group design study where subjects are not selected for MRI lesion activity at baseline.
- 66 placebo subjects received monthly MRI scans and followed up to 6 months.

117

Sormani's model (cont)



- From the placebo subjects:
 - N= 66, Mean = 13.0, SD = 20.6, Range 0-122.
- The parameters μ and κ are estimated from this data
$$\mu = 13.0, \quad \alpha = 0.52$$
- For simulation, the untreated (placebo) group was obtained by sampling from this Negative Binomial distribution

118

Sormani's model (cont)



- The treated group was obtained from randomly sampling from the distribution with mean $\mu (1 - \text{treatment effect})$ and same κ ; as κ was not supposed to change.
- Sample sizes were presented for various combinations of treatment effect and power.

119

Table: Sample size per treatment group for selected power and treatment effect



Treatment Effect	Power 80%		Power 90%	
	Discrete (Sormani)	Continuous (Normal)	Discrete (Sormani)	Continuous (Normal)
50%	118	164	140	220
60%	65	114	85	153
70%	40	80	58	112
80%	24	65	35	86

120

Sample Size - Summary



- Literature suggests to use Nonparametric approach for analyzing count data instead of using the usual normal distribution approach
- However, sample size calculation by nonparametric approach is not helpful – requires “larger” sample size
- Mixed Poisson Model produces “significantly” smaller sample sizes!

121

Bayesian Analysis of Poisson



- Closed form posterior distribution exists for Poisson model without covariates
- Suppose $\{y_i\}$, $i=1,2,\dots,n$ is a random sample from $\text{Poisson}(\lambda)$
 - Prior of λ is $\text{gamma}(\alpha, \beta)$
 - Mean of prior is given by $E_0(\lambda)=\alpha/\beta$
 - Gamma is a conjugate prior for Poisson

$$g(\lambda | y) \propto \left(\prod_{i=1}^n e^{-\lambda} \lambda^{y_i} \right) \frac{\alpha^\beta}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\lambda\beta}$$
$$\propto e^{-\lambda(\beta-n)} \lambda^{\alpha+n\bar{y}-1}$$

122

Bayesian Analysis



- Hence, the posterior of λ is a gamma with parameters

$$\tilde{\alpha} = \alpha + n\bar{y}, \quad \tilde{\beta} = \beta + n$$

- With posterior mean given by

$$\begin{aligned} E_{\pi}(\lambda | y, \alpha, \beta) &= \frac{\alpha + n\bar{y}}{\beta + n} \\ &= \frac{\beta}{\beta + n} E_0(\lambda) + \frac{n}{\beta + n} \bar{y} \end{aligned}$$

- Posterior mean is a weighted average of prior mean and sample mean
- Weight associated with sample mean is an increasing function of sample size

123

Paired count data



- Bayesian analysis of paired count data (correlated Poisson counts) was provided by Karlis and Ntzoufras, Stat Medicine, 2006
 - Based on a bivariate Poisson distribution, they showed that the distribution of the difference of two correlated Poisson variables takes the same form as that of two independent Poisson variates (Skellam, JRSS-A, 1946)
 - They extended the result to ZIP model and presented Bayesian inference of the parameters including developing a MCMC algorithm
 - Not discussed further

124



Illustrative Example

- Hypoglycemia in the DCCT
 - Episode of hypoglycemia where the blood glucose level falls too low
 - Data from secondary cohort of 715 patients
 - Variables used:
 - Group assignment (intensive, conventional)
 - Number of events (Response)
 - Number of months duration of diabetes on entry into the study
- Biostatistical Methods – The Assessment of Relative Risks” by John Lachin

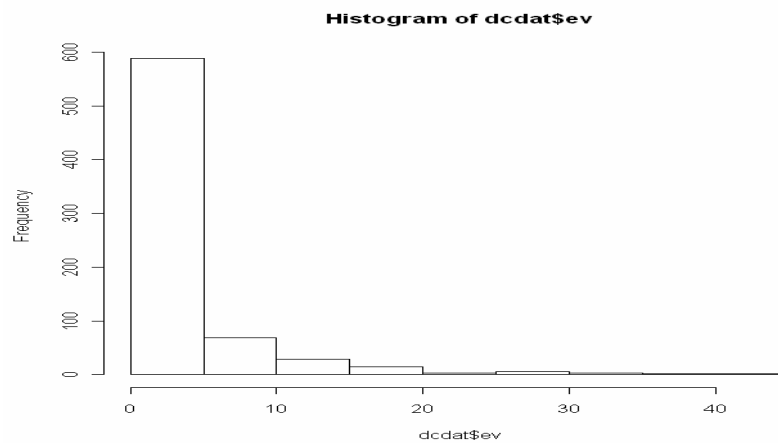
<http://www.bsc.gwu.edu/jml/biostatmethods/datasets/hypoglycemia/dccthyppo.dat>

<http://www.bsc.gwu.edu/jml/biostatmethods/pgmindex.html>

125

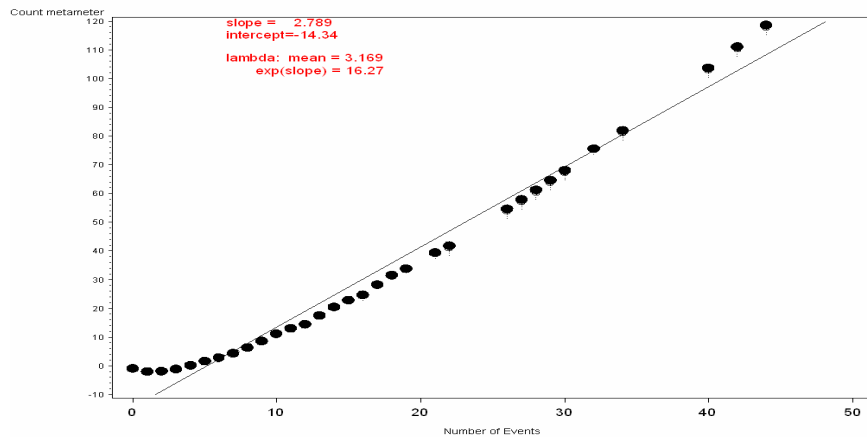


Number of Events



126

Poissonness Plot



127

Bayesian Poisson Model



```
ods graphics on;  
proc mcmc data=dcct seed=1181 nmc=100000 thin=10  
  propcov=quanew monitor=( _parms_ Pearson );  
  ods select Parameters PostSummaries PostIntervals tadpanel;  
  parms beta0 0 beta1 0 beta2 0;  
  prior beta: ~ normal(0, var=1000);  
  mu=exp(beta0+beta1*int*duration+beta2*conven*duration);  
  model nevents~poisson(mu);  
  if obs=1 then Pearson=0;  
  Pearson=Pearson+( (nevents-mu) **2/mu );  
run;  
ods graphics off;  
ods rtf close;
```

128

Output



The SAS System

The SAS System

The MCMC Procedure

The MCMC Procedure

Parameters			
Parameter	Sampling Method	Initial Value	Prior Distribution
beta0	N-Metropolis	0	normal(0,var=1000)
beta1	N-Metropolis	0	normal(0,var=1000)
beta2	N-Metropolis	0	normal(0,var=1000)

Posterior Summaries						
Parameter	N	Mean	Standard Deviation	Percentiles		
				25%	50%	75%
beta0	10000	1.0581	0.0552	1.0207	1.0588	1.0946
beta1	10000	0.00403	0.000470	0.00372	0.00403	0.00435
beta2	10000	-0.00479	0.000618	-0.00520	-0.00478	-0.00437
Pearson	10000	5943.4	128.7	3836.8	5941.9	6031.2

Posterior Intervals				
Parameter	Alpha	Equal-Tail Interval		HPD Interval
		Lower	Upper	
beta0	0.050	0.9499	1.1673	0.9533 1.1702
beta1	0.050	0.00309	0.00496	0.00307 0.00493
beta2	0.050	-0.00604	-0.00359	-0.00601 -0.00356
Pearson	0.050	5702.7	6209.1	5696.2 6199.5

129

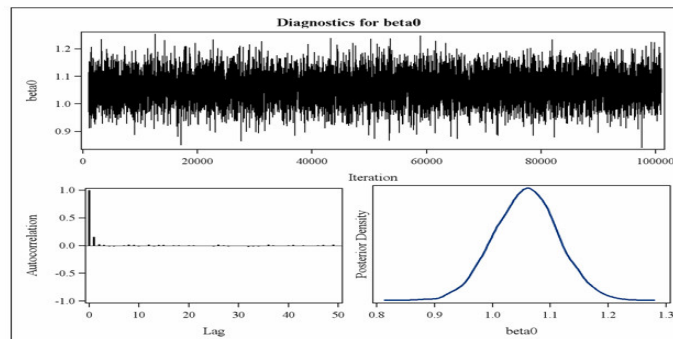
Output



12:41 Tuesday, May 26, 2009 3

The SAS System

The MCMC Procedure



130

Bayesian ZIP Model



```
ods graphics on;
proc mcmc data=dcct seed =1181 nmc=100000 thin=10
propcov=quanew monitor =(_parms_ Pearson);
ods select Parameters PostSummaries PostIntervals tadpanel;
parms beta0 0 beta1 0 beta2 0 eta .3;
prior beta: ~ normal(0,var=1000);
prior eta ~ uniform(0,1);
mu=exp(beta0 + beta1*int*duration + beta2*conven*duration);
llike=log(eta*(nevents eq 0) + (1-eta)*pdf("poisson",nevents,mu));
model general(llike);
if obs = 1 then Pearson = 0;
mean = (1 - eta)*mu;
sigma2 = (1 - eta)*mu*(1 + eta*mu);
Pearson = Pearson + ((nevents - mean)**2/sigma2);
run;
ods graphics off;
ods rtf close;
```

131

Summary



The SAS System
The FREQ Procedure

severe hypoglycemia episodes

nevents	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	313	43.78	313	43.78
1	105	14.63	418	58.46
2	53	7.25	477	65.71
3	42	5.87	519	72.59
4	37	5.17	556	77.76
5	33	4.62	589	82.38
6	18	2.52	607	84.90
7	12	1.68	619	86.57
8	11	1.54	630	88.11
9	12	1.68	642	89.79
10	15	2.10	657	91.89
11	9	1.26	666	93.15
12	9	1.26	675	94.41
13	5	0.70	679	95.11
14	7	0.98	686	96.09
15	5	0.70	691	96.79
16	2	0.28	693	97.07
17	4	0.56	697	97.63
18	6	0.84	703	98.47
19	3	0.42	706	98.89
20	1	0.14	707	99.03
21	2	0.28	709	99.31
22	1	0.14	710	99.45
23	1	0.14	711	99.59
24	1	0.14	712	99.73
25	1	0.14	713	99.87
26	1	0.14	714	100.00
27	1	0.14	715	100.00
28	1	0.14	716	100.00
29	1	0.14	717	100.00
30	1	0.14	718	100.00
31	1	0.14	719	100.00
32	1	0.14	720	100.00
33	1	0.14	721	100.00
34	1	0.14	722	100.00
35	1	0.14	723	100.00
36	1	0.14	724	100.00
37	1	0.14	725	100.00
38	1	0.14	726	100.00
39	1	0.14	727	100.00
40	1	0.14	728	100.00
41	1	0.14	729	100.00
42	1	0.14	730	100.00
43	1	0.14	731	100.00
44	1	0.14	732	100.00

The SAS System
The MEANS Procedure

Variable	Label	N	Mean	Std Dev
nevents	# severe hypoglycemia episodes	715	3.1692308	5.6415473
grp	tx group (1=Int, 0=Conv)	715	0.5076323	0.5002988
insulin	insulin units per kg weight	715	0.7126592	0.2380821
duration	months diabetes duration	715	104.7076323	44.7967636
female	female (1) or male (0)	715	0.4643357	0.4990756
adult	adult (1) or adolescent (0)	715	0.9020979	0.2973903
bcva15	C-peptide in pmol/L	715	0.0630007	0.0781259
hbae1	level of HbA1c at initial screening	715	3.1694825	1.5355919
hxcona	Prior history of coma/seizure	715	0.0517483	0.2216734

Variable	Label	Minimum	Maximum
nevents	# severe hypoglycemia episodes	0	44.0000000
grp	tx group (1=Int, 0=Conv)	0	1.0000000
insulin	insulin units per kg weight	0.1975009	2.0618557
duration	months diabetes duration	10.0000000	180.0000000
female	female (1) or male (0)	0	1.0000000
adult	adult (1) or adolescent (0)	0	1.0000000
bcva15	C-peptide in pmol/L	0.0100000	0.5000000
hbae1	level of HbA1c at initial screening	6.6000000	15.4200000
hxcona	Prior history of coma/seizure	0	1.0000000

132

Fit of the data



Parameter Estimates				
Parameter	DF	Estimate	Standard Error	Approx Pr > t
Intercept	1	1.217106	0.218753	5.56 <.0001
grp	1	1.098922	0.049289	22.30 <.0001
insulin	1	0.004156	0.100263	0.04 0.9669
duration	1	0.001410	0.000564	2.50 0.0123
female	1	0.196789	0.042345	4.65 <.0001
adult	1	-0.782375	0.065860	-11.88 <.0001
bcval5	1	-0.701629	0.371853	-1.89 0.0592
hbael	1	-0.039122	0.015184	-2.58 0.0100
hxcoma	1	0.684897	0.068641	9.98 <.0001

Poisson

Parameter Estimates					
Parameter	DF	Estimate	Standard Error	t Value	Approx Pr > t
Intercept	1	1.608803	0.651747	2.47	0.0136
grp	1	1.217992	0.124899	9.75	<.0001
insulin	1	0.278757	0.294584	0.95	0.3440
duration	1	0.000060202	0.001729	0.03	0.9722
female	1	0.280246	0.125037	2.24	0.0250
adult	1	-0.821468	0.219596	-3.74	0.0002
bcval5	1	-1.655617	1.004773	-1.65	0.0994
hbael	1	-0.090379	0.042992	-2.10	0.0355
hxcoma	1	0.594712	0.262574	2.26	0.0235
_Alpha	1	2.117071	0.160490	13.19	<.0001

Neg Binomial

133

Fit of the data



Parameter Estimates					
Parameter	DF	Estimate	Standard Error	t Value	Approx Pr > t
Intercept	1	1.133109	0.226076	5.01	<.0001
grp	1	0.680251	0.051962	13.09	<.0001
insulin	1	-0.052848	0.100738	-0.52	0.5999
duration	1	0.001991	0.000591	3.37	0.0008
female	1	0.215415	0.043704	4.93	<.0001
adult	1	-0.584156	0.066245	-8.82	<.0001
bcval5	1	0.518849	0.397188	1.31	0.1914
hbael	1	0.027962	0.015756	1.77	0.0759
hxcoma	1	0.424188	0.069277	6.12	<.0001
Inf_Intercept	1	-2.788875	0.898190	-3.10	0.0019
Inf_grp	1	-1.088274	0.166155	-6.55	<.0001
Inf_insulin	1	-0.157408	0.408176	-0.39	0.6998
Inf_duration	1	0.003400	0.002233	1.52	0.1278
Inf_female	1	0.039657	0.166887	0.24	0.8122
Inf_adult	1	0.756263	0.337011	2.24	0.0248
Inf_bcval5	1	3.716217	1.303137	2.85	0.0043

ZIP

Parameter Estimates					
Parameter	DF	Estimate	Standard Error	t Value	Approx Pr > t
Intercept	1	0.539934	0.663213	0.81	0.4156
grp	1	0.918404	0.152672	6.02	<.0001
insulin	1	0.394579	0.308247	1.28	0.2005
duration	1	0.003170	0.001813	1.75	0.0803
female	1	0.197026	0.134248	1.47	0.1422
adult	1	-0.690401	0.218453	-3.16	0.0016
bcval5	1	0.686592	1.145873	0.60	0.5490
hbael	1	-0.004412	0.048659	-0.09	0.9277
hxcoma	1	0.516430	0.240949	2.14	0.0321
Inf_Intercept	1	-11.603555	6.393404	-1.81	0.0695
Inf_grp	1	-2.094936	0.901570	-2.32	0.0201
Inf_insulin	1	1.447247	1.754393	0.82	0.4094
Inf_duration	1	0.021617	0.016980	1.27	0.2030
Inf_female	1	-0.326714	0.708204	-0.46	0.6446
Inf_adult	1	1.281047	1.057533	1.21	0.2258
Inf_bcval5	1	14.079566	8.352603	1.69	0.0919

ZINB

Parameter Estimates					
Parameter	DF	Estimate	Standard Error	t Value	Approx Pr > t
Inf_hbael	1	0.566251	0.248448	2.28	0.0227
Inf_hxcoma	0	-17.105373	.	.	.
_Alpha	1	1.620654	0.264734	6.12	<.0001

Predicted proportion of zero events



Obs	mn0	mn1	mn2	mn3	mn4	mn5	mn6	mn7	mn8	mn9	mn10
1	0.12930	0.19360	0.17731	0.14190	0.10970	0.080945	0.056097	0.036959	0.023887	0.015653	0.010577
2	0.42905	0.16018	0.09477	0.06375	0.04582	0.034375	0.026600	0.021082	0.017031	0.013975	0.011618
3	0.43813	0.03348	0.05679	0.07121	0.07539	0.071846	0.063338	0.052199	0.040516	0.029914	0.021280
4	0.43923	0.13781	0.09048	0.06471	0.04840	0.037280	0.029338	0.023479	0.019050	0.015635	0.012958

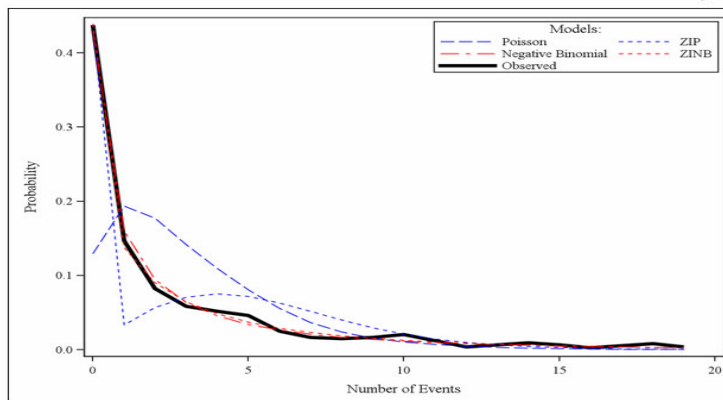
1

135

Average predicted count probability



07:08 Sunday, June 07, 2009 9



136

Poisson Fit



The SAS System

2

The COUNTREG Procedure

Model Fit Summary

```
Dependent Variable      nevents
Number of Observations      715
Data Set                  WORK.ALLEVT5
Model                      Poisson
Log Likelihood             -2557
Maximum Absolute Gradient   0.00115
Number of Iterations        6
Optimization Method        Newton-Raphson
AIC                        5132
SBC                        5174
```

Algorithm converged.

Parameter Estimates

Parameter	DF	Estimate	Standard Error	t Value	Approx Pr > t
Intercept	1	1.217106	0.218753	5.56	<.0001
grp	1	1.098922	0.049289	22.30	<.0001
insulin	1	0.004156	0.100263	0.04	0.9669
duration	1	0.001410	0.000564	2.50	0.0125
female	1	0.196789	0.042345	4.65	<.0001
adult	1	-0.782375	0.065860	-11.88	<.0001
bvval5	1	-0.701629	0.371853	-1.89	0.0592
hbae1	1	-0.039122	0.015184	-2.58	0.0100
hxcona	1	0.684897	0.068641	9.98	<.0001

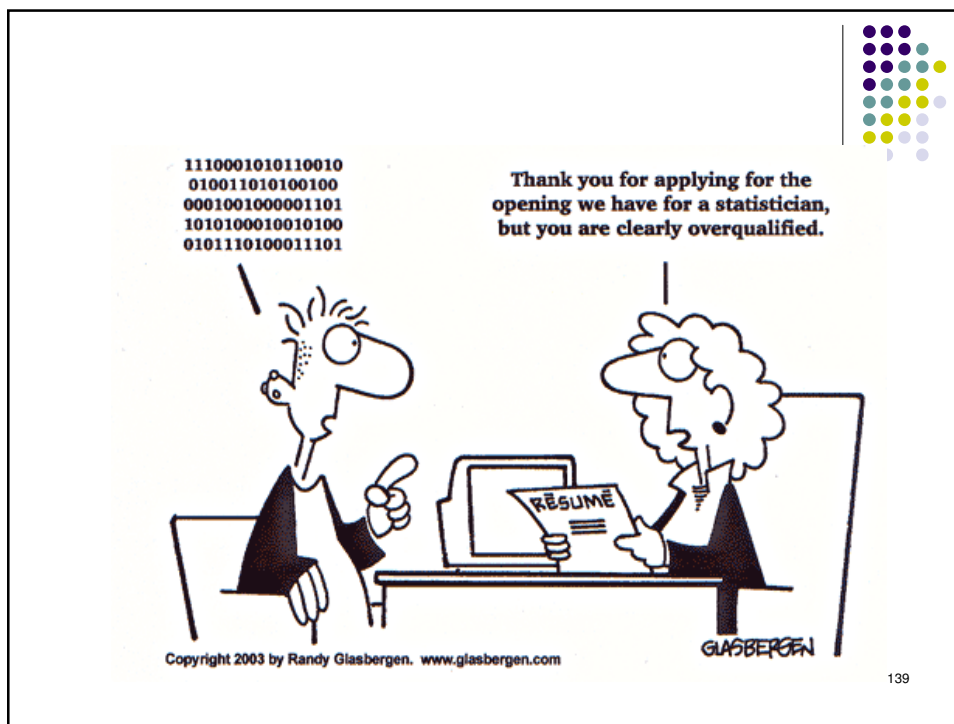
137

Conclusion



- Poisson models provide a standard framework for the analysis of count data.
- In practice, count data are often overdispersed relative to the Poisson distribution
- Alternative models such as NB, ZIP, ZINB and others are available when overdispersion is exhibited.
- Numerous software options are available for analyzing count data.

138



References

- Lambert, D. (1992). *Technometrics*, 34,1-14
- Thall, P. F. and Vail, S. C. (1990). *Biometrics* **46**, 657-671
- Agresti, A (2002). *Categorical Data Analysis, Second Edition*, Wiley
- Cameron and Trivedi (1998). *Regression Analysis of Count Data*, Cambridge University Press
- Ridout, Demetrio and Hinde (1998). *International Biometric Conference*, 1998
- Weller and Ryan (1998). *Biometrics*, 54, 762-773
- Sormani et. al (2001) *J of Neurol Neurosurg Psychiatry*, 70: 494-499 .

140

Backup Slides



Simulation Results



Simulation with Count Data



- Two independent samples from Poisson distributions are generated
 - Sample sizes 20, 30 per treatment group
 - Various Poisson parameters are considered
- Number of simulations is 500.

143

Simulation (cont)



- Following tests and model are considered to compare the population means:
 - Two sample T test
 - Nonparametric Wilcoxon Rank Sum
 - Poisson regression model
 - Negative Binomial model
 - ZIP model

144



	Power (20 subjects per treatment)				
Poisson Parameters	T-test	Nonparametric	Poisson	Neg. Binomial	ZIP
4, 3	0.39	0.36	0.41	0.36	0.31
4, 2	0.96	0.95	0.96	0.97	0.87
7, 6	0.24	0.22	0.26	0.23	0.24
7, 5	0.69	0.69	0.73	0.73	0.68

145



	Power (20 subjects per treatment)				
Poisson Parameters	T-test	Nonparametric	Poisson	Neg. Binomial	ZIP
10, 8	0.54	0.52	0.55	0.54	0.53
10, 7	0.90	0.88	0.92	0.92	0.91
12, 10	0.47	0.44	0.48	0.48	0.47
12, 9	0.83	0.80	0.84	0.84	0.83

146



	Power (30 subjects per treatment)				
Poisson Parameters	T-test	Nonparametric	Poisson	Neg. Binomial	ZIP
4, 2	0.98	0.98	0.98	0.98	0.96
7, 5	0.87	0.83	0.88	0.86	0.86
10, 7	0.96	0.96	0.97	0.96	0.97
12, 9	0.93	0.92	0.94	0.93	0.94

147

Simulation Summary



- For count data, without overdispersion, the powers of two sample t test, Wilcoxon rank sum tests have comparable powers based on discrete models.
- Both NB and ZIP models perform comparably well inspite of their complexity.
- For ease of use, nonparametric Wilcoxon rank sum test may be of interest to practitioners.

148

Simulation with Overdispersion



- Two independent samples from Poisson distributions are generated
- In each sample, approximately 20%, 40% zeros are added to create over-dispersion
- Number of simulations is 500
- Same tests and models are considered to compare the population means.

149



Poisson Parameters	Power (24 subjects per treatment)				
	T-test	Nonparametric	Poisson	Neg. Binomial	ZIP
4, 3	0.21	0.18	0.41	0.18	0.32
4, 2	0.90	0.84	0.96	0.87	0.88
7, 5	0.39	0.42	0.73	0.25	0.69
12, 9	0.33	0.55	0.84	0.06	0.83

150



	Power (28 subjects per treatment)				
Poisson Parameters	T-test	Nonparametric	Poisson	Neg. Binomial	ZIP
4, 2	0.82	0.64	0.96	0.76	0.89
6, 3	0.93	0.84	0.99	0.82	0.99
7, 4	0.80	0.65	0.99	0.49	0.98
12, 9	0.13	0.26	0.84	0.00	0.83

151



	Power (42 subjects per treatment)				
Poisson Parameters	T-test	Nonparametric	Poisson	Neg. Binomial	ZIP
4, 2	0.95	0.88	0.98	0.93	0.97
6, 3	0.98	0.97	0.99	0.97	0.99
7, 5	0.44	0.39	0.89	0.15	0.86
12, 9	0.34	0.53	0.94	0.00	0.94

152

Simulation of Over-dispersion Results



	Power (48 subjects per treatment)				
Poisson Parameters	T-test	Nonparametric	Poisson	Neg. Binomial	ZIP
4, 2	0.94	0.72	0.98	0.88	0.97
6, 3	0.98	0.88	0.99	0.91	0.99
7, 5	0.32	0.15	0.88	0.05	0.86
12, 9	0.20	0.24	0.94	0.00	0.94

153

Simulation Summary (overdispersion)



- In cases of overdispersion, Poisson regression model produces highest power consistently followed by ZIP model.
- When the means are lower (closer to zero), power based on t-test and nonparametric tests are not too low compared to the power from the Poisson regression model.

154

Simulation Summary (overdispersion)



- When the means are higher, Poisson regression and ZIP model are clearly the winners.
- Unless overdispersion is rather high (50% or more) and the means are high (i. e., the variability in the sample is very high) ZIP model may not be necessary!

155

Model Comparison for Zero-Inflated Data Score Statistic (Broek, 1995)



- Score statistic is calculated on the basis of a score test for $p=0$ in the inflated Poisson.
- H_0 : Poisson fits well vs H_1 : ZIP fits well

- The score statistic

$$S(\tilde{\gamma}) = \frac{\left\{ \sum_{i=1}^n 1_{(y_i = 0)} / e^{-\tilde{\lambda}_i} - 1 \right\}^2}{\left\{ \sum_{i=1}^n \left(\frac{1}{e^{-\tilde{\lambda}_i}} - 1 \right) \right\} - A},$$

$$A = \tilde{\lambda}^T X [X^T \text{diag}(\tilde{\lambda}) X]^{-1} X^T \tilde{\lambda},$$

where $\tilde{\gamma}$ and $\tilde{\lambda}_i$ are the estimates of γ and λ_i under the null hypothesis. Under null hypothesis, the statistic is asymptotically chi-squared distribution with 1 df.

Not shown further

156

Model Comparison for Zero-Inflated Data



- Akaike Information Criterion (AIC) (Akaike, 1974, 1975)

$$AIC = -2L(\tilde{\theta} | y) + 2k$$

- Bias adjustment approach to developing a selection criterion
- Estimates the expected overall Kullback-Leibler Discrepancy
- Predictive Divergence Criterion (PDC) (Davies and Cavanaugh, 2002)

$$PDC = \sum_{i=1}^N -2 \ln g_i(y_i | \tilde{\theta}_{-i})$$

- Cross validation approach to developing a selection criterion
- Development of a selection criterion which is not bias adjusted
- Estimates the expected overall PDC discrepancy

157

Simulation Schemes

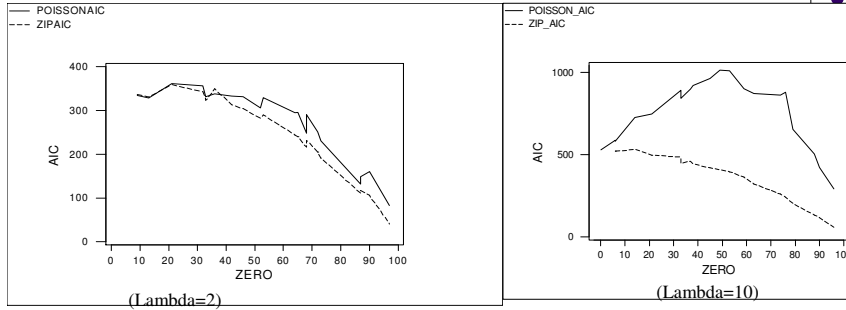


- Scheme I
 - No covariate
- Scheme II
 - Single covariate
- Scheme III
 - Multiple covariates
- Simulation details are not provided here
- Model comparisons were made between Poisson and ZIP

158

SIMULATION SCHEME I

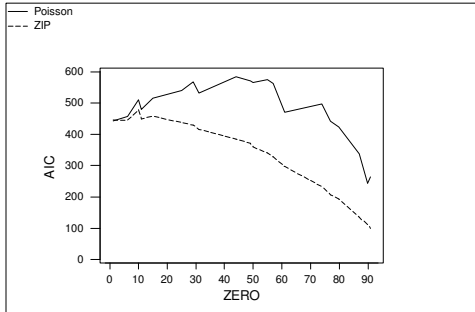
Plots of AIC vs ZERO (Without Covariate Case)



(Lambda=2)

(Lambda=10)

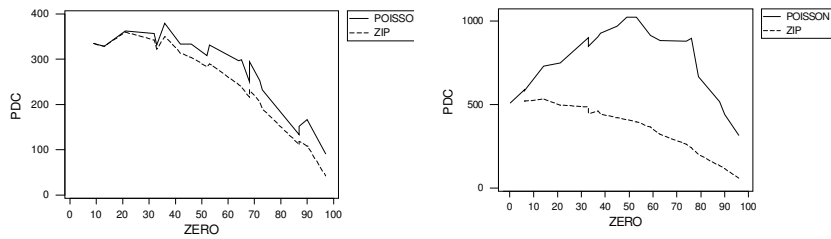
(Lambda=5)



159

SIMULATION SCHEME I

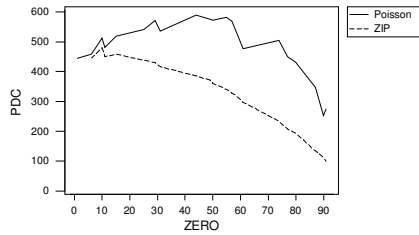
Plots of PDC vs ZERO (Without Covariate Case)



(Lambda=2)

(Lambda=10)

(Lambda=5)

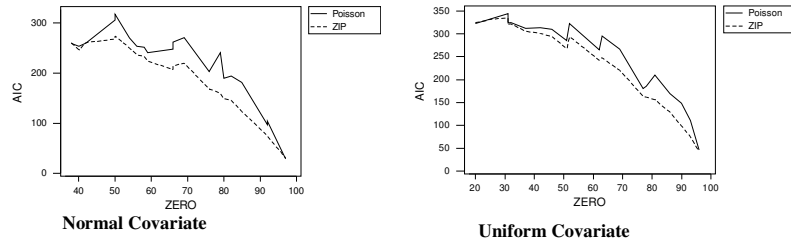


160

SIMULATION SCHEME II



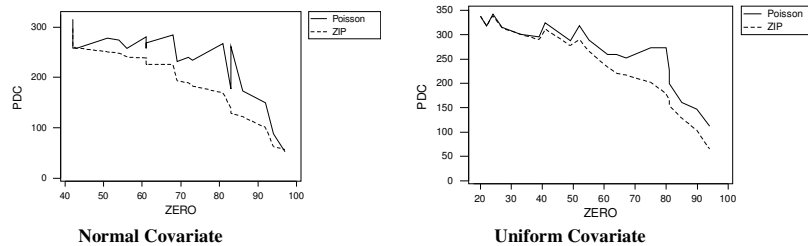
Plots of AIC vs Zero(%)



Normal Covariate

Uniform Covariate

Plots of PDC vs Zero(%)



Normal Covariate

Uniform Covariate

61

Summary of Model Comparison



- In case of without covariate model (simulation scheme I), more large the λ is, more certainty is to get “zero” from Bernoulli distribution only. ZIP is more preferable than Poisson model.
- In simulation scheme II, ZIP is better than Poisson. But it seems to the fact that incorporation of covariates increase the tendency of overdispersion irrespective of zero.
- When data simulated directly from Poisson, even for very large amount of zero, Poisson and ZIP are indistinguishable.

162

Summary of Model Comparisons



- In simulation III scheme (not shown here), we looked at single and two covariates in ZIP model (making it closer to Poisson Regression)
- In simulation scheme, though from the score statistics, Poisson is preferable compared to ZIP, AIC and PDC values for both models are not so much different.