Analyzing Count Data with Count Distributions Using SAS

Monsanto Regulatory Statistics Technology Center

Outline

1. A motivating example
2. Concept of Poisson mixtures
3. Regression analysis using SAS
4. Simulations for model selection
5. Analysis of NTA data of a GM maize
6. Summary and conclusion

Motivating example:

Purpose of the study is the environment and ecological risk assessment (ERA) of a GM maize, expressing an insecticidal double-stranded RNA, on NTA. Field studies are conducted under diverse geographic and environmental conditions to assess potentially adverse effects of the GM maize relative to the conventional control.

Region	US	Argentina	Brazil
# of sites	4	4	6
# of replicates, RCBD	4	4	4
# of collections, random	5	5-6	5

Arthropod collection methods

Sticky Trap

Visual Count

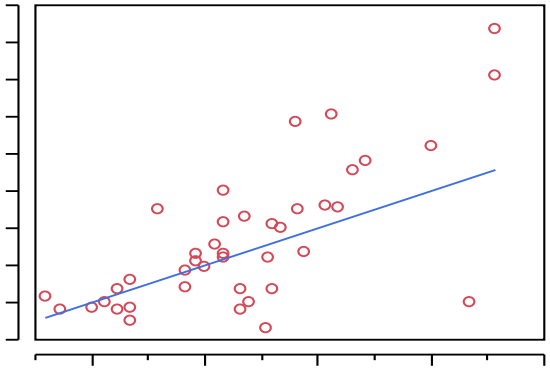
Data: Arthropod count for each taxa by Region x Site x Rep x Line x Collection

Statistical task: Modeling variance-mean relationship

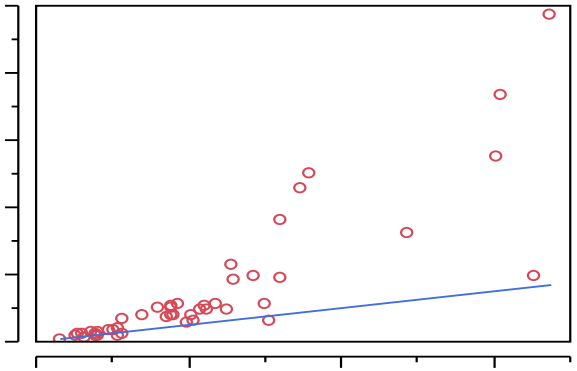
Evidence of different types of over-dispersion

By taxa Var-vs-Mean plots in square-root scale (blue line for Var = Mean as a reference line for no over-dispersion):

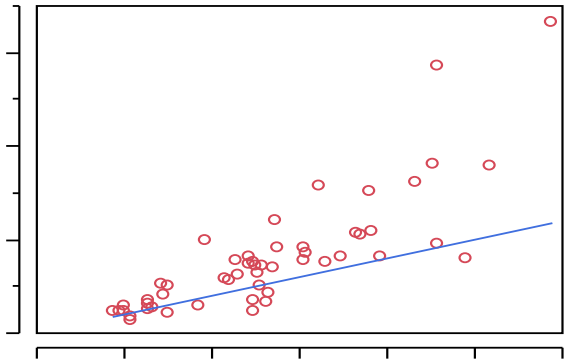
Spiders (No disp.)



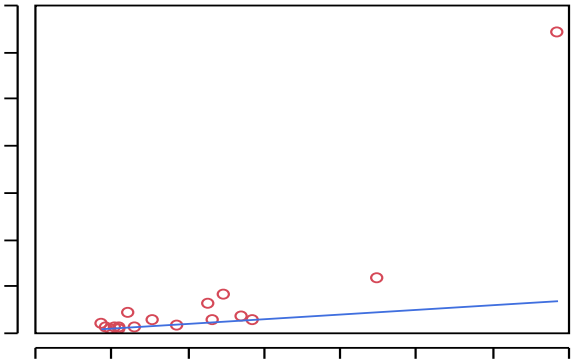
Leafhoppers (Quad.disp.)



Minute pirate bugs (Linear disp.)



Aphids(Het. Disp. Or GP)



What is the Poisson mixture?

Probability Mass Function (pmf) of Poisson model:

$$f_P(y;\lambda) = \frac{e^{-\lambda}\lambda^y}{y!} \quad y = 0, 1, 2, \dots, \ \lambda > 0$$

Regression: $\log(\lambda) = X'\beta$; restriction: $\mathrm{E}(y) = Var(y) = \lambda$.

Poisson mixture model: Assume the mixing density function $g(.)$

$$f_{PM}(y;\theta) = \int \frac{e^{-\lambda}\lambda^y}{y!} \cdot g(\lambda|\theta) d\lambda$$

where $g(\lambda|\theta)$ provides flexibility to account for over-dispersion, θ can be 2-dimensional or multidimensional.

Six Poisson mixture models: Conditional mean and variance with fixed line effect *A* and random collection effect *B*.

Distribution Model Abbreviation	**Conditional Variance as Function of $\mu_{M\|B}$**	**Var. Function Type or Over-dispersion Type**
P	$\mu_{M\|B}$	Constant (1)
PN	$\mu_{M\|B} + \left(e^{\sigma_\varepsilon^2} - 1\right)\mu_{M\|B}^2$	Quadratic (2)
NB	$\mu_{M\|B} + \alpha\mu_{M\|B}^2$	Quadratic (2)
NB1	$[(1+\delta)/\delta]\mu_{M\|B}$	Linear (3)
GP	$[1/(1-\xi)^2]\mu_{M\|B}$	Linear (3)
NB1N	$\left(\frac{1+\delta}{\delta}\right)\mu_{M\|B} + \left(e^{\sigma_\varepsilon^2} - 1\right)\mu_{M\|B}^2$	Linear Regression (4)
GPN	$\left(\frac{1}{(1-\xi)^2}\right)\mu_{M\|B} + \left(e^{\sigma_\varepsilon^2} - 1\right)\mu_{M\|B}^2$	Linear Regression (4)

Challenges

- How to do it in SAS?
- Which Poisson mixture to select?
- Impact on estimate of treatment effect?

How to specify the different distribution models in SAS

- SAS procedures GLIMMIX and MCMC
- Only Poisson can be directly called in both procedures
- NB can be directly called in GLIMMIX but not in MCMC
- Using likelihood functions: NB1, GP, NB1N, GPN
- Using numerical integration in likelihood calculation for NB1N, GPN.

Difficulties in using GLIMMIX:

- Often failure to converge even assuming the simpler model without any interaction effects

  NOTE: Did not converge.

  ERROR: QUANEW Optimization cannot be completed.

  Converge, but the fixed effect estimate has 0 StdErr.
- Method = Laplace or Quad is required using AIC BIC for model comparison
- No flexibility for # of dispersion parameters > 1

  Such as heterogeneous over-dispersion
- Could not use NB1N and GPN with self-defined likelihood

Modeling flexibility of MCMC:

- No convergence problem even with self-defined likelihood function using numerical integration
- Modeling over-dispersion with more than one parameters
- DIC is always provided for model comparison

```
ods output DIC = fitGPN;
```

Simulation for model selection

Mean function for simulated data: $log(\mu_{klm}) = \mu + B_k + L_l + C_m$

Model selection and misspecification (with criterion AIC diff. < 6 for GLIMMIX and DIC diff. < 6 for MCMC) from 100 simulations.

Model for Data Generation	Portion of Selected Model						
	P	PN	NB	NB1	GP	NB1N	GPN
SAS GLIMMIX analyses results							
L-01, PN	39	100	99	58	62		
H-01, PN	0	100	99	5	8		
SAS MCMC analyses results							
L-01, PN	64	90	91	35	38	23	11
H-01, PN	0	100	92	4	7	40	49

Mean difference in proportion assuming the test mean being 50% less of the control.

Model for Data	Mean Difference and SE or SD						
	P	PN	NB	NB1	GP	NB1N	GPN
Analyses results from SAS GLIMMIX							
L-01, PN	-0.49 (0.073)	-0.49 (0.089)	-0.49 (0.092)	-0.48 (0.081)	-0.48 (0.088)		
H-01, PN	-0.49 (0.032)	-0.50 (0.060)	-0.50 (0.059)	-0.47 (0.058)	-0.47 (0.060)		
Analyses results from SAS MCMC							
L-01, PN	-0.49 (0.073)	-0.49 (0.099)	-0.49 (0.097)	-0.48 (0.090)	-0.48 (0.093)	-0.48 (0.090)	-0.48 (0.105)
H-01, PN	-0.49 (0.032)	-0.50 (0.062)	-0.50 (0.061)	-0.47 (0.058)	-0.47 (0.060)	-0.47 (0.058)	-0.50 (0.064)

Analysis of NTA data of a GM maize

Distribution model selection:

Method: MCMC (GLIMMIX did not work)

Four models: P, NB, GP, GPN (other four models would also work well)

Analysis results on nine arthropod taxa presented in at least two countries.

Arthropod Taxa	**P**	**NB**	**GP**	**GPN**	**NBH**	**GPH**	**Range (Max – Min)**
Aphids	691.45	647.49	619.49	635.63	648.12	618.74	72.71
Leafhoppers	4438.99	3039.56	3207.75	2997.09			1441.90

Posterior density distribution of difference: Test and control difference in proportion of control, and 95% high probability density (HPDL, HPDU).

Insect	Selected Model	Diff. in Proportion (SD)	HPDL	HPDU
Aphids	GP	0.47 (1.05)	-0.50	3.81
Dermaptera	NB	-0.08 (0.23)	-0.43	0.46
Lacewings	P	-0.04 (0.43)	-0.58	1.08
Ladybird beetles	P	0.09 (0.38)	-0.45	1.00
Leafhoppers	GPN	0.05 (0.26)	-0.34	0.62
Minute pirate bugs	GP	-0.04 (0.24)	-0.40	0.50
Parasitic wasps	NB	0.04 (0.35)	-0.42	0.97
Sap beetles	GP	-0.00 (0.24)	-0.39	0.57
Spiders	P	0.17 (0.38)	-0.32	1.21

By-site summary of test and control comparisons for Aphids

Arthropod	Country	Method	Site	GM Maize (Mean SE)	Control (Mean SE)	Ref Range (Min – Max)
Aphids	ARG	Sticky	Site 1	5.0 (0.61)	5.5 (1.06)	5.2 - 6.1
Aphids		Visual	Site 2	15.7 (8.90)	6.5 (2.88)	2.2 - 41.0
Aphids	US	Sticky	Site 3	3.2 (0.95)	1.2 (0.34)	2.1 - 5.1
Aphids		Sticky	Site 4	5.1 (1.03)	4.1 (1.41)	4.9 - 7.5

Posterior density distribution estimate of dispersion parameters for six arthropod taxa with the selected distribution model

Insect	Selected Model	Linear Comp. Coef. Mean (SD)	Quadratic Comp. Coef. Mean (SD)
Aphids	GP	3.00 (0.61)	
Dermaptera	NB		0.18 (0.02)
Leafhoppers	GPN	11.16 (1.22)	0.10 (0.01)
Minute pirate bugs	GP	2.03 (0.18)	
Parasitic wasps	NB		0.08 (0.01)
Sap beetles	GP	2.69 (0.27)	

Conclusion and summary

- Over-dispersion in count data are common
- Dispersion type likely can be only determined by model fitting.
- Model misspecification had limited impact on treatment comparisons.
- GLIMMIX presents serious convergence problem, while Bayesian analysis with MCMC showed no problem.
- No evidence of the GM maize effect on NTA abundance

References

Ahmad, A., Oliveira, W., Brown, C., Asiimwe, P., Sammons, B., Horak, M., Jiang, C., Carson, D. (2015), "Transportable data from non-target arthropod field studies for the environmental risk assessment of genetically modified maize expressing an insecticidal double-stranded RNA," *Transgenic Research*, 24, DOI 10.1007/s11248-015-9907-3.

Cameron, A. C. and Trivedi, P. K. (1998), *Regression Analysis of Count Data*, New York: Cambridge University Press.

Consul, P. C. and F. Famoye (1992), "Generalized Poisson regression model," *Communications in statistics – Theory and Method* 21: 89-109.

Efron, B. (1986), "Double Exponential Families and Their Use in Generalized Linear Regressions," *Journal of the American Statistical Association*, 81, 709-721.

FAO/WHO. (2001), Food and Agriculture Organization of the United Nations/World Health Organization. "Evaluation of allergenicity of genetically modified foods." *Report of a joint FAO/WHO Expert Consultation on Allergenicity of Foods Derived from Biotechnology*, January 22–25, Rome, Italy.

Hilbe, J. M. (2011), *Negative Binomial Regression*, New York: Cambridge University Press.

Joe, H. and Zhu, R. (2005), "Generalized Poisson Distribution: The Property of Mixture of Poisson and Comparison with Negative Binomial Distribution," *Biometrical Journal*, 47, 219–229.

McCulloch, C. E., Shayle R. Searle, and John M. Neuhaus (2008), *Generalized, Linear, and Mixed Models*, 2nd Edition. A John Wiley & Sons, Inc. Publication.

SAS (2012), *Bayesian Analysis Using SAS*. SAS Publishing.

Sileshi, G. (2006), "Selecting the right statistical model for analysis of insect count data by using information theoretic measures," *Bulletin of Entomological Research* (2006) 96, 479–488