

Ecological Statistics

Models for non-normal data: an introduction

Non-linear module – part 4

C. Staudhammer - Eco Stats (fall 2008)

Lecture outline

- Overview of Generalized linear models (GzLMs)
- Models for count data
 - Poisson regression
 - Zero-inflated models

C. Staudhammer - Eco Stats (fall 2008)

Generalized linear models (GzLMs) versus General linear models (GLMs)

- GzLMs are statistical models that combine elements of linear and nonlinear models
- GzLMs apply if the *response* variable is in the exponential family (e.g., Bernoulli, Binomial, Poisson, Normal, etc.)
 - The Normal (Gaussian) linear regression (GLM) and ANOVA models are special cases of GzLMs
- If you know the probability distribution of your response variable (e.g., count data ~ Poisson), it is often more appropriate to use a GzLM

GzLM components

- Systematic component
 - Linear combination of covariates
 - Additive, systematic part of model
- Link function
 - Transformation of the mean
 - Maps the mean onto a scale where the covariate effects are additive
 - Ensures range restrictions
- Random component
 - Distribution of the response chosen from the exponential family

The Random component

- A function is a member of the exponential family if it can be written as:

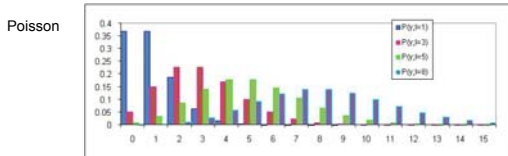
$$f(y) = \exp\left(\frac{1}{\psi} [y\theta - b(\theta)] + c(y, \psi)\right)$$

- Binomial, Poisson, and Normal are members of this family
- For all members, the mean is $b'(\theta)$
- For all members except Normal, the mean is a function of the variance

Examples of the Random component

- Binomial $P(y) = \binom{n}{y} \pi^y (1-\pi)^{n-y}, y = 0, 1, \dots, n$
- Negative Binomial $P(y) = \binom{k+y-1}{y} \pi^k (1-\pi)^y, y = 0, 1, \dots$
- Poisson $P(y) = \frac{\lambda^y e^{-\lambda}}{y!}, y = 0, 1, \dots$
- Gamma $P(y) = \frac{1}{\beta^\alpha \Gamma(\alpha)} y^{\alpha-1} \exp(-y/\beta), y > 0$

The Poisson and Negative Binomial Distributions



The Systematic component

- A linear combination of covariates and/or fixed effects parameters

- The linear predictor for the i th observation is:

$$\mathbf{x}'_i \boldsymbol{\beta} = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_k x_{ki}$$

- The linear predictor in a GzLM is equal to *some transformation $g(\cdot)$ of the mean*:

$$g(E[Y_i]) = \mathbf{x}'_i \boldsymbol{\beta}$$

The Link function

- The *transformation $g(\cdot)$ of the mean*
 - $g(\cdot)$ transforms the mean onto a scale where the covariate effects are additive
 - the link function is a linearizing transformation, and the GzLM is intrinsically linear
 - For example, if the mean function is:

$$\mu_i = E[Y_i] = \exp(\beta_0 + \beta_1 x)$$
 Then, the linearizing transformation is $\ln(\mu_i)$
 - $g(\cdot)$ also may serve another purpose: to confine predictions to an appropriate range

Link functions for count data

- Characteristics of count data:
 - Whole numbers only
 - Bounded below at zero, but no upper bound ($y \rightarrow \infty$)
- Link function should map from $[0, \infty]$ to $[-\infty, \infty]$:
 - Log link: $\ln(E[Y_i]) = \mathbf{x}'_i \boldsymbol{\beta}$
 - Effects are multiplicative, rather than additive

Multiplicative effect of Log Link function

- The log link implies a multiplicative impact for independent variables
 - E.g., consider the univariate case, where x is continuous:

$$E[Y | X] = \exp(\beta_0 + \beta_1 X)$$

$$E[Y | X + 1] = \exp(\beta_0 + \beta_1 (X + 1))$$

$$= \exp(\beta_0 + \beta_1 X + \beta_1) = \exp(\beta_0 + \beta_1 X) \exp(\beta_1)$$

$$\Rightarrow \frac{E[Y | X + 1]}{E[Y | X]} = \frac{\exp(\beta_0 + \beta_1 X) \exp(\beta_1)}{\exp(\beta_0 + \beta_1 X)} = \exp(\beta_1)$$

Poisson regression model

- the dependent variable Y is the number of occurrences of an event, e.g., seedling counts
- $Y \sim \text{Poisson}$, given the independent variables X_1, X_2, \dots, X_k , which may be continuous (e.g., stand density) or categorical (e.g., forest type) variables

$$P(Y = y | X_1, X_2, \dots, X_k) = \frac{\mu^y e^{-\mu}}{y!}, \quad y = 0, 1, \dots$$

- the log of the mean μ is assumed to be a linear function of the independent variables:

$$\ln(\mu) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k$$

- μ is the exponential function of the independent variables:

$$\mu = \exp[\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k]$$

Examples of Poisson Regression Models

- Y is the number of longleaf pine seedlings in area, some of which was subject to prescribed burning
 - the mean μ is a function of the distance to the nearest adult ($dist_{ad}$), the distance from the center of a gap opening ($dist_{gap}$), and a categorical variable indicating burning (B)

$$\ln(\mu) = \beta_0 + \beta_1 dist_{ad} + \beta_2 dist_{gap} + \beta_k B$$

- Y is the number of shade leaves produced by ivy growing in trees
 - the mean μ is a function of day length (dl), rainfall ($rain$), and a categorical variable indicating location (inside vs. outside canopy: L)

$$\ln(\mu) = \beta_0 + \beta_1 dl + \beta_2 rain + \beta_k L$$

We make these log functions because we want $\mu > 0$

Fitting a Poisson regression model

- These kinds of models **cannot** be fit with PROC GLM or MIXED
 - PROC GENMOD
 - fits generalized linear models
 - allows the response probability distribution to be any member of an exponential family of distributions
 - PROC GLIMMIX
 - fits statistical models to data with random effects, correlations or non-constant variability
 - Allows the response probability distribution to be non-normally distributed
 - With version 9.1, you must download it from the SAS website as an add-in

Example

- Regeneration data was collected on Brazil nut (*Bertholletia excelsa*) seedlings in two communities in an effort to understand the demographic structure of Western Amazonian populations that have different histories and levels of nut exploitation
 - To what extent are *B. excelsa* seedling/sapling densities explained by fruit fate (collection intensity) and/or forest structure?



Data

- 144 – 25x25 m subplots were sampled in each site
 - Within each subplot, all Brazil nut seedlings and saplings were measured
 - (x,y) coordinates,
 - height and diameter at base of seedlings
 - height and diameter at breast height of saplings
 - Number and basal area of all tree species ≥ 10 cm dbh
 - Fruits on the ground were counted
 - Number of unopened, intact fruits
 - Number of fruits opened by rodents

Forest Structure

The number of closed plus opened fruits per reproductively mature tree per ha indicates collection intensity

The proportion of opened fruits indicates predator/disperser activity

How do we answer the research question?

- What is our dependent variable?
- What is/are our independent variable(s)?
- What distribution do the data have?
 - What model is appropriate for this distribution?

PROC GENMOD syntax

```
PROC GENMOD DATA=bnutdata;
  CLASS site;
  MODEL num_sdgs = site num_fruit tph
    ba / dist=poisson link=log type3;
RUN;
```

This specifies that a Poisson distribution will be used; that is, $num_sdgs \sim \text{Poisson}$

This specifies that the log of the mean of num_sdgs is a linear function of the independent variables

This requests a test for all levels together for any class variables in the model.

PROC GENMOD output: fit

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	283	248.8100	0.8792
Scaled Deviance	283	248.8100	0.8792
Pearson Chi-Square	283	362.5891	1.2812
Scaled Pearson X2	283	362.5891	1.2812
Log Likelihood		-166.1534	

The Deviance and Pearson Chi-Square $\sim \chi^2_{(DF)}$. Scaled Deviance and Scaled Pearson X2 are the above divided by the dispersion parameter (which is 1 for the Poisson).

The scaled Pearson χ^2 is one indicator of fit, but not usually sufficient on its own. If these ratios were far from 1, we would conclude that the fit of the Poisson model is not adequate.

Over- and under-dispersion

- Deviance and Pearson Chi-Square divided by the degrees of freedom detect over- or under-dispersion
 - Since in the Poisson distribution the mean and the variance are equal, the deviance and the Pearson statistic divided by the degrees of freedom should be ~ 1
 - $> 1 \rightarrow$ over-dispersion
 - $< 1 \rightarrow$ underdispersion
- To accommodate under- or over-dispersion indicates inadequate fit of the Poisson model:
 - Add a dispersion parameter instead of setting it to 1 (options DSCALE and PSCALE in the MODEL statement)
 - For over-dispersion, try the negative binomial regression instead of the Poisson regression (options DIST=NB instead of DIST=POISSON in the MODEL statement)

PROC GENMOD output: significance

Analysis Of Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits	Chi-Square	Pr>ChiSq
Intercept	1	-0.9801	3.4920	-7.8242 5.8641	0.08	0.7790
Site	2	-1.4647	0.7224	-2.8806 -0.0488	4.11	0.0426
Site	3	0	0.0000	0.0000 0.0000	.	.
num_fruit	1	-0.0476	0.0193	-0.0854 -0.0098	6.08	0.0137
tph	1	-0.0102	0.0072	-0.0244 0.0039	2.01	0.1567
ba	1	0.2399	0.0890	0.0654 0.4143	7.26	0.0070
Scale	0	1.0000	0.0000	1.0000 1.0000		

There is a significant difference (at the 0.05 level) between the two sites, as well as a significant effect of basal area and numbers of fruits on the ground

PROC GENMOD output: estimates

Analysis Of Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits	Chi-Square	Pr>ChiSq
Intercept	1	-0.9801	3.4920	-7.8242 5.8641	0.08	0.7790
Site	2	-1.4647	0.7224	-2.8806 -0.0488	4.11	0.0426
Site	3	0	0.0000	0.0000 0.0000	.	.
num_fruit	1	-0.0476	0.0193	-0.0854 -0.0098	6.08	0.0137
tph	1	-0.0102	0.0072	-0.0244 0.0039	2.01	0.1567
ba	1	0.2399	0.0890	0.0654 0.4143	7.26	0.0070
Scale	0	1.0000	0.0000	1.0000 1.0000		

The number of seedlings is lowest in site 2

To get predicted numbers of seedlings for site 2 with basal area=25 and 50 open+closed fruits per reproductive adult:
 Predicted = $\exp(-0.98 - 1.46 + 50 \cdot -0.0476 + 25 \cdot 0.24)$
 = 3.23

PROC GENMOD output: test of fixed effects

LR Statistics For Type 3 Analysis

Source	DF	Chi-Square	Pr > ChiSq
Site	1	4.06	0.0438
num_fruit	1	6.34	0.0118
tph	1	2.08	0.1492
ba	1	6.97	0.0083

Density (tph=trees per ha) is not a significant predictor of number of seedlings

Or use PROC GLIMMIX

- For vers. 9.1, download the Add-in from SAS at: <http://support.sas.com/rnd/app/da/glimmix.html>

```
PROC GLIMMIX DATA=bnutdata;
  CLASS site;
  MODEL num_sdlgs = site num_fruit tph
    ba /dist=poisson link=log SOLUTION;
RUN;
```

Use SOLUTION option to get predictive equation

PROC GLIMMIX output: fit and test of fixed effects

Fit Statistics	
-2 Log Likelihood	367.10
AIC (smaller is better)	377.10
AICC (smaller is better)	377.31
BIC (smaller is better)	395.41
CAIC (smaller is better)	400.41
HOIC (smaller is better)	384.43
Pearson Chi-Square	362.59
Pearson Chi-Square / DF	1.28

Recall GENMOD output:
Pearson chi-sq = 1.2812

Recall GENMOD P-values:
Site: 0.04
fruits: 0.01
Ba: 0.008

Type III Tests of Fixed Effects				
Effect	Num DF	Den DF	F Value	Pr > F
Site	1	283	4.11	0.0435
num_fruit	1	283	6.08	0.0143
tph	1	283	2.01	0.1578
ba	1	283	7.26	0.0075

PROC GLIMMIX output: estimates

Parameter Estimates						
Effect	Site	Estimate	Standard Error	DF	t Value	Pr > t
Intercept		-0.9801	3.4920	283	-0.28	0.7792
Site	2	-1.4647	0.7224	283	-2.03	0.0435
Site	3	0
num_fruit		-0.04758	0.01930	283	-2.47	0.0143
tph		-0.01022	0.007216	283	-1.42	0.1578
ba		0.2399	0.08900	283	2.69	0.0075

Predicted numbers of seedlings for site 2 with basal area=25 and 50 open+closed fruits per reproductive adult:
Predicted = $\exp(-0.98 - 1.46 + 50 \cdot -0.0476 + 25 \cdot 0.24) = 3.23$
(same as GENMOD)

Also, a 1-unit increase in basal area reduces the number of seedlings by 0.2399

Evaluating model fit

- As with other models, check p-values associated with predictor values
- Examine the probability plot of the *Pearson residuals*
 - These are normalized so that when the model is a reasonable fit to the data, they have roughly a standard normal distribution
 - Note: non-Pearson residuals will have different variances!
- Check the value of the Deviance

C. Staudhammer - Eco Stats (fall 2008)

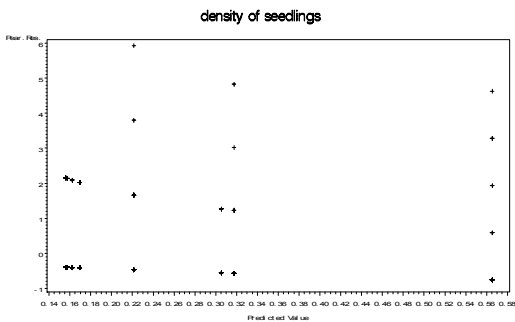
Obtaining Pearson residuals

```
PROC GENMOD DATA=bnutdata;
  CLASS site;
  MODEL num_sdlgs = site num_fruit tph ba /
    dist=poisson link=log type3;
  OUTPUT OUT=out P=pred STDRESCHI=pearson;
RUN;
```

OR

```
PROC GLIMMIX DATA=bnutdata;
  CLASS site;
  MODEL num_sdlgs = site num_fruit tph ba /
    dist=poisson link=log SOLUTION;
  OUTPUT OUT=out P(ILINK)=p PEARSON(ILINK)=prs;
RUN;
```

Pearson residuals - example



Comparing model fits with “Type I analysis”

- Likelihood ratio test
 - Allows comparison of nested models
 - Available in GENMOD
 - Procedure:
 - Specify `type1` in model options
 - Specify model statement with variables in order of desired succession in test
 - GENMOD computes the LR statistic for the inclusion of each additional variable, which is asymptotically $\sim \chi^2$

C. Staudhammer - Eco Stats (fall 2008)

Type 1 analysis - example

The GENMOD Procedure

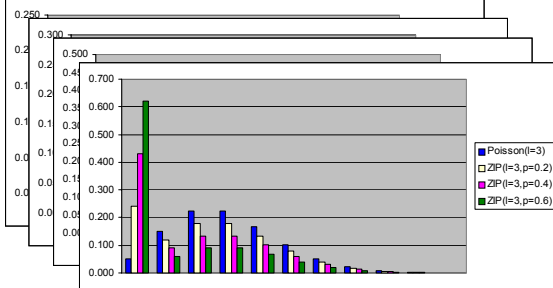
LR Statistics For Type 1 Analysis

Source	Deviance	DF	Chi-Square	Pr > ChiSq
Intercept	265.6199			
Site	262.1329	1	3.49	0.0619
Num_fruit	256.1442	1	5.99	0.0144
tph	255.7798	1	0.36	0.5461
ba	248.8100	1	6.97	0.0083

Too many zeroes?

- The expected number of zeroes under the Poisson model is $\exp(-\lambda)$
 - The number of counts (e.g., # seedlings surviving) may be controlled by one factor (e.g., rainfall), while whether or not regeneration occurs is controlled by another factor (e.g., soil fertility)
- More zeroes than expected under the Poisson (or negative binomial model) causes over-dispersion
 - Tests of significance and CI's are then incorrect

Poisson versus zero-inflated Poisson (ZIP)



C. Staudhammer - Eco Stats (fall 2008)

ZIP Model

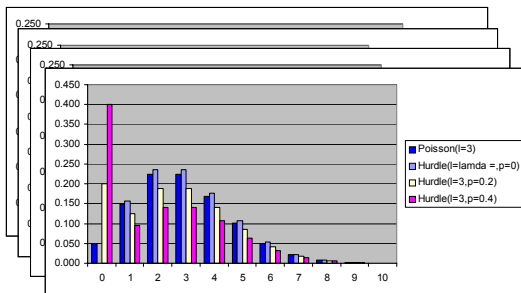
- Zero-Inflated Poisson probability density function

$$P(X = x) = \begin{cases} \pi + (1 - \pi)P(X = 0 | X \sim \text{Poisson}), & x = 0 \\ (1 - \pi)P(X > 0 | X \sim \text{Poisson}), & x > 0 \end{cases}$$

$$= \begin{cases} \pi + (1 - \pi)e^{-\lambda}, & x = 0 \\ (1 - \pi) \frac{\lambda^x e^{-\lambda}}{x!}, & x > 0 \end{cases}$$

where: π = inflation probability

Poisson versus Hurdle Model



C. Staudhammer - Eco Stats (fall 2008)

Hurdle Model

- Hurdle probability density function

$$P(X = x) = \begin{cases} \pi, & x = 0 \\ (1 - \pi) \frac{P(X > 0 | X \sim \text{Poisson})}{1 - P(X = 0 | X \sim \text{Poisson})}, & x > 0 \end{cases}$$

$$= \begin{cases} \pi, & x = 0 \\ (1 - \pi) \frac{\lambda^x e^{-\lambda}}{x!(1 - e^{-\lambda})}, & x > 0 \end{cases}$$

where: π = inflation probability

Incorporating effects into the ZIP, Hurdle Models

- To add effects, we let the parameters λ and π be functions of covariates

$$\pi_i = \frac{1}{1 + \exp(f(\text{covariates}))}$$

$$\lambda_i = g(\text{covariates})$$

This is a logit transformation, which ensures that $\pi_i \in [0,1]$

Obtaining ZIP model parameter estimates

- We must use NLMIXED to do this, as we need to define a pdf (i.e., ZIP and Hurdle are not "built-in", like Normal or Poisson)
- First, define parameters
 - For example, let the inflation probability be a function of site and numbers of fruits, let lambda be a function of ba and tph
 - We use programming statements within PROC NLMIXED to make lambda and pi functions of these covariates

```
PROC NLMIXED DATA=bnutdata;
PARAMETERS a0=0 a1=0 a2=0 b0=0 b1=0 b2=0;
pi = 1/(1 + exp(-(a0 + a1*num_fruit + a2*site)));
lambda = exp(b0 + b1*tph + b2*ba);
...
...
```

Note: class variables are not allowed in NLMIXED, so define site as a dummy (0/1) variable

Obtaining ZIP model parameter estimates – cont'd

- Next, we construct a log-likelihood function (more on this later in the term)
- Since there are two distinct cases, we define LL for zero and non-zero counts

```
...
IF count = 0 THEN
  LL = log(pi + (1-pi)*exp(-lambda));
ELSE LL = log(1-pi) + count*log(lambda) -
  lgamma(count+1) - lambda;
MODEL count ~ general(LL);
ESTIMATE "inflation probability" pi;
ESTIMATE "lambda" lambda;
RUN;
```

lgamma is the log of the gamma function

general specifies a general log likelihood

ESTIMATE gives additional user-defined estimates

ZIP model PROC NLMIXED output

Fit Statistics									
				-2 Log Likelihood					361.3
				AIC (smaller is better)					371.3
				AICC (smaller is better)					371.5
				BIC (smaller is better)					389.6

Are these parameters necessary?

Note: AICC was 371 with Poisson model

Parameter Estimates									
Parameter	Estimate	Standard Error	DF	t Value	Pr > t	Alpha	Lower	Upper	Gradient
a0	-0.4739	1.5055	288	-0.31	0.7532	0.05	-3.4370	2.4893	0.00068
a1	0.02070	0.0316	288	0.66	0.5125	0.05	-0.04143	0.08283	0.002496
a2	0.4521	0.8550	288	0.53	0.5974	0.05	-1.2308	2.1351	-1.59E-6
b0	-6.9462	3.4853	288	-1.99	0.0472	0.05	-13.8061	-0.08623	-0.00009
b1	0.01007	0.0053	288	1.92	0.0564	0.05	-0.00028	0.02042	-0.03141
b2	0.09446	0.0810	288	1.17	0.2446	0.05	-0.06499	0.2539	-0.00241

Additional Estimates								
Label	Estimate	Standard Error	DF	t Value	Pr > t	Alpha	Lower	Upper
inflation prob	0.6241	0.1252	288	4.99	<.0001	0.05	0.3778	0.870

How to improve the ZIP model?

- Examine the parameters
 - Try a different formulations, e.g., all covariates in λ

```
PROC NLMIXED DATA=bnutdata;
PARAMETERS a0=0 b0=0 b1=0 b2=0 b3=0 b4=0;
pi = 1/(1 + exp(-a0));
lambda = exp(b0 + b1*tph + b2*ba + b3*num_fruit + b4*site);
IF count = 0 THEN LL = log(pi + (1-pi)*exp(-lambda));
ELSE LL = log(1-pi) + count*log(lambda) -
  lgamma(count+1) - lambda;
MODEL count ~ general(LL);
ESTIMATE "inflation probability" pi;
ESTIMATE "lambda" lambda;
RUN;
```

Revised ZIP model PROC NLMIXED output

Fit Statistics									
				-2 Log Likelihood					350.9
				AIC (smaller is better)					362.9
				AICC (smaller is better)					363.2
				BIC (smaller is better)					384.9

Note: testing if a0=0 is equivalent to testing $\pi=1/(1+\exp(0))=0.5$

Note: AICC was 371 with Poisson model

Parameter Estimates									
Parameter	Estimate	Standard Error	DF	t Value	Pr > t	Alpha	Lower	Upper	Gradient
a0	0.2948	0.3153	288	0.93	0.3520	0.05	-0.3277	0.9174	-0.00002
b0	-2.1810	3.8405	288	-0.57	0.5705	0.05	-9.7401	5.3781	0.000128
b1	-0.01061	0.008376	288	-1.27	0.2063	0.05	-0.02710	0.00588	0.052954
b2	0.2791	0.1047	288	2.67	0.0081	0.05	0.07300	0.4851	0.003099
b3	-0.05570	0.02257	288	-2.47	0.0142	0.05	-0.1001	-0.01127	0.004169
b4	1.5371	0.8929	288	1.72	0.0863	0.05	-0.2204	3.2946	0.000085

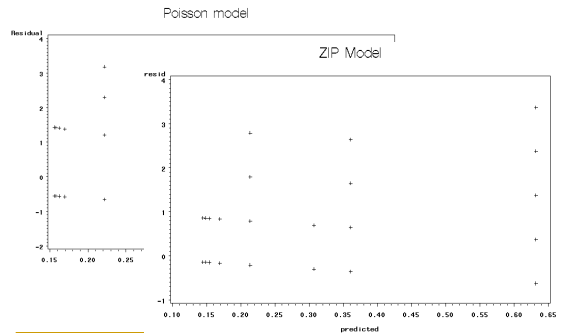
Additional Estimates								
Label	Estimate	Standard Error	DF	t Value	Pr > t	Alpha	Lower	Upper
inflation prob	0.5732	0.07738	288	7.41	<.0001	0.05	0.4209	0.7255
lambda	0.3955	0.1326	288	2.98	0.0031	0.05	0.1346	0.6564

Comparison of Poisson and ZIP models for seedling counts

- Poisson model:
 - Site, number of fruits on the ground, and basal area were significant ($p < 0.05$)
 - AIC=371
- ZIP model:
 - Number of fruits on the ground and basal area were significant ($p < 0.05$), site significant ($p < 0.10$)
 - Inflation probability estimate is 0.57
 - AIC=363
- Residuals?

C. Staudhammer - Eco Stats (fall 2008)

Compare residuals



C. Staudhammer - Eco Stats (fall 2008)

Obtaining Hurdle model parameter estimates

- Again, use NL MIXED

```
PROC NL MIXED DATA=bnutdata;
PARAMETERS a0=0 b0=0 b1=0 b2=0 b3=0 b4=0;
pi = 1/(1 + exp(-a0));
lambda = exp(b0 + b1*tph + b2*ba +
             b3*num fruit + b4*site);
IF count = 0 THEN LL = log(pi);
ELSE LL = log(1-pi) - lambda - lgamma(count+1)
          + count*log(lambda) - log(1-exp(-lambda));
MODEL count ~ general(LL);
ESTIMATE "inflation probability" pi;
ESTIMATE "lambda" lambda;
RUN;
```

pi and lambda are defined as in the ZIP model; the log-likelihoods are different.

Hurdle model PROC NL MIXED output

		Fit Statistics							
		-2 Log Likelihood						349.6	
		AIC (smaller is better)						355.6	
		AICc (smaller is better)						355.9	
		BIC (smaller is better)						377.6	
		Parameter Estimates							
Parameter	Estimate	Standard Error	DF	t Value	Pr > t	Alpha	Lower	Upper	Gradient
a0	1.4663	0.1510	288	9.71	<.0001	0.05	1.1692	1.7635	-0.00014
b0	-26.8476	20.4652	288	-1.31	0.1906	0.05	-67.1299	13.4346	-0.00003
b1	0.04956	0.05512	288	0.90	0.3693	0.05	-0.05893	0.1581	-0.01172
b2	0.4611	0.1715	288	2.69	0.0076	0.05	0.1235	0.7986	-0.00051
b3	-0.1965	0.1452	288	-1.35	0.1768	0.05	-0.4822	0.08917	-0.00075
b4	-1.5389	2.9621	288	-0.52	0.6038	0.05	-7.3690	4.2912	-8.94E-6
		Additional Estimates							
Label	Estimate	Standard Error	DF	t Value	Pr > t	Alpha	Lower	Upper	
inflation prob	0.8125	0.02300	288	35.33	<.0001	0.05	0.7672	0.8578	
lambda	0.01683	0.05924	288	0.28	0.7766	0.05	-0.09977	0.1334	

Does anything look amiss?

Note: AICC was 371 with Poisson model

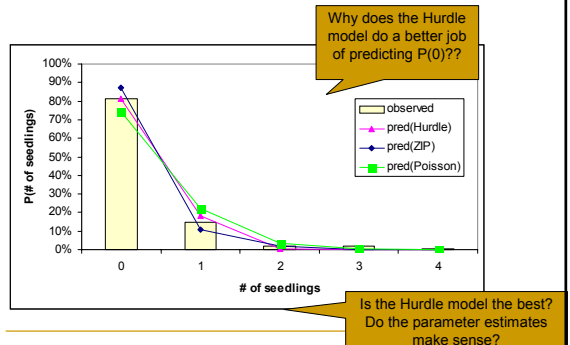
Which model is best?

- Fit statistics, parameter estimates, ...

Model	AICC	parameter estimates		
		pi	lambda	E(counts)
Poisson	371	-	0.296	0.296
ZIP	363	0.573	0.396	0.169
Hurdle	356	0.813	0.017	0.189
Observed				0.257

- Note: these do not tell you *the whole story!*
- Predicted vs. Observed...

Predicted vs. observed



Why does the Hurdle model do a better job of predicting P(0)??

Is the Hurdle model the best? Do the parameter estimates make sense?

Take home messages

- Poisson regression is a flexible way to model count data
 - When over-dispersion is observed in the data, zero-inflated models can improve model fit
 - Selecting the best model includes examining the AICC, as well as ensuring that parameter estimates are biologically reasonable
-