A strategy for modelling count data which may have extra zeros

Alan Welsh

Centre for Mathematics and its Applications
Australian National University

The Data

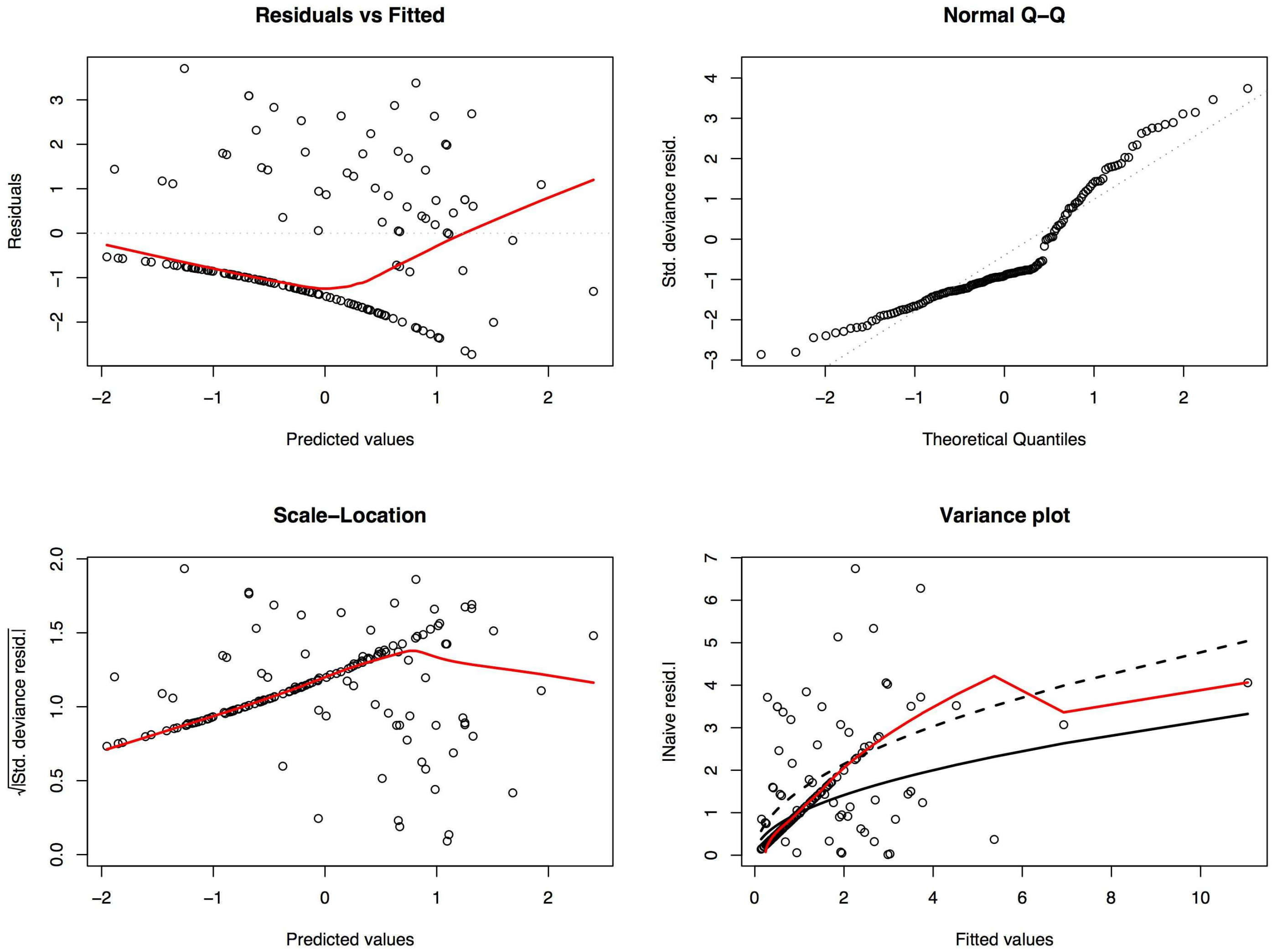
Response is the number of Leadbeater's possum on 151 3-ha sites in montane ash forests in Victoria, SE Australia.

Explanatory variables are site variables including:

- *lstags*: log(number of trees with hollows + 1)
- *baa*: basal area of *Acacia* species (m^2/ha)
- *bark*: score for the degree of decorticating bark
- *nos*: a score based on *bark* and the number of shrubs
- *slope*: the slope of the site

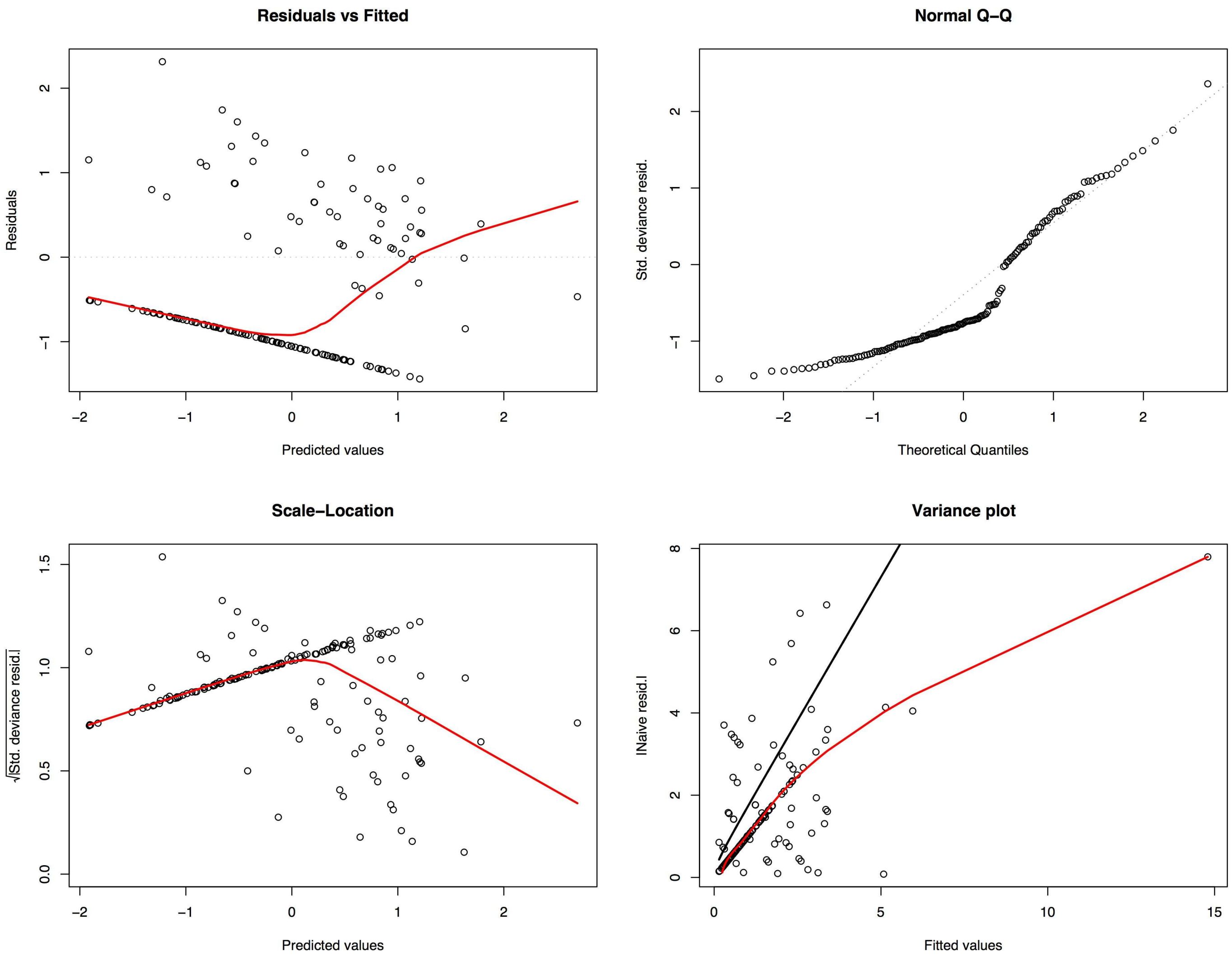
More details in Lindenmayer (1989), Welsh et al (1996.)

Poisson Regression (all covariates)



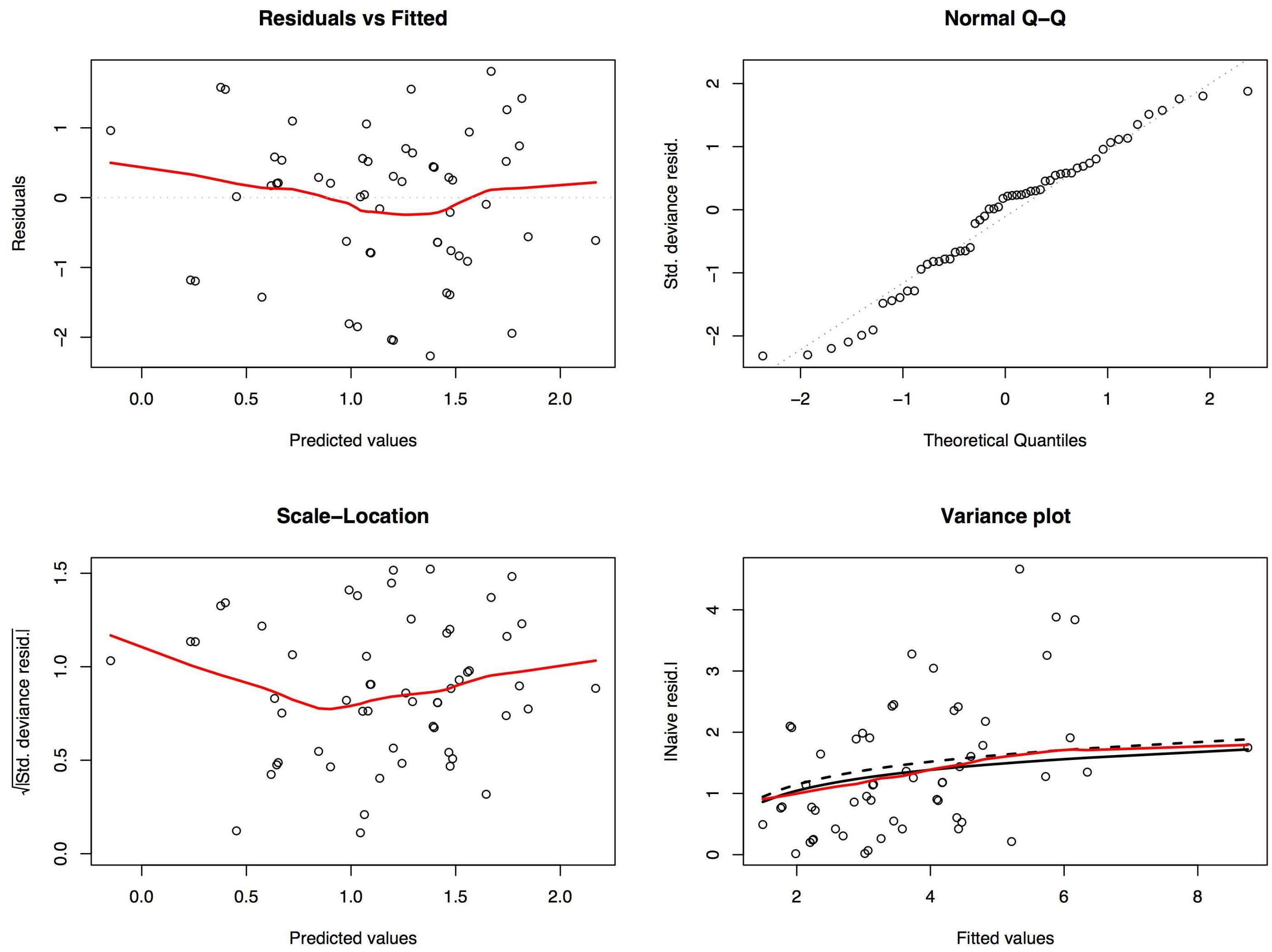
Residual mean deviance = 2.30; $Z_e = 0.234$.

Negative Binomial Regression (all covariates)



Residual mean deviance = 0.89; $Z_e = 0.046$.

Truncated Poisson Regression (all covariates)



Residual mean deviance = 1.20

Overdispersion versus Extra Zeros

For the truncated Poisson model,

- the variance function looks reasonable
- the residual mean deviance is 1.20
- the truncated negative binomial model reduces to the truncated Poisson

suggesting no overdispersion.

(This makes sense for both models.)

The parametric bootstrap based on the fitted model estimates the null sampling distribution of Z_e and gives a percentile p-value of zero.

(The bootstrap distribution is essentially normal so a direct asymptotic argument will give the same result.)

There is significant zero inflation in the data.

In the initial Poisson regression model, the apparent overdispersion was in fact due to zero-inflation.

Separated and Overlapping Models

- Separated Models (also Two-part, Hurdle or Conditional models)

$$\Pr(Y_i = y|\boldsymbol{x}_i) = \begin{cases} 1 - \pi_i & y = 0 \\ \pi_i \frac{g_i(y)}{1-g_i(0)} & y = 1, 2, \ldots \end{cases}$$

  Let the Poisson parameter be θ_i.

- Overlapping models (also Zero-Inflated Models)

$$\Pr(Y_i = y|\boldsymbol{x}_i) = \begin{cases} 1 - p_i + p_i g_i(0) & y = 0 \\ p_i g_i(y) & y = 1, 2, \ldots \end{cases}$$

  Let the Poisson parameter be λ_i.

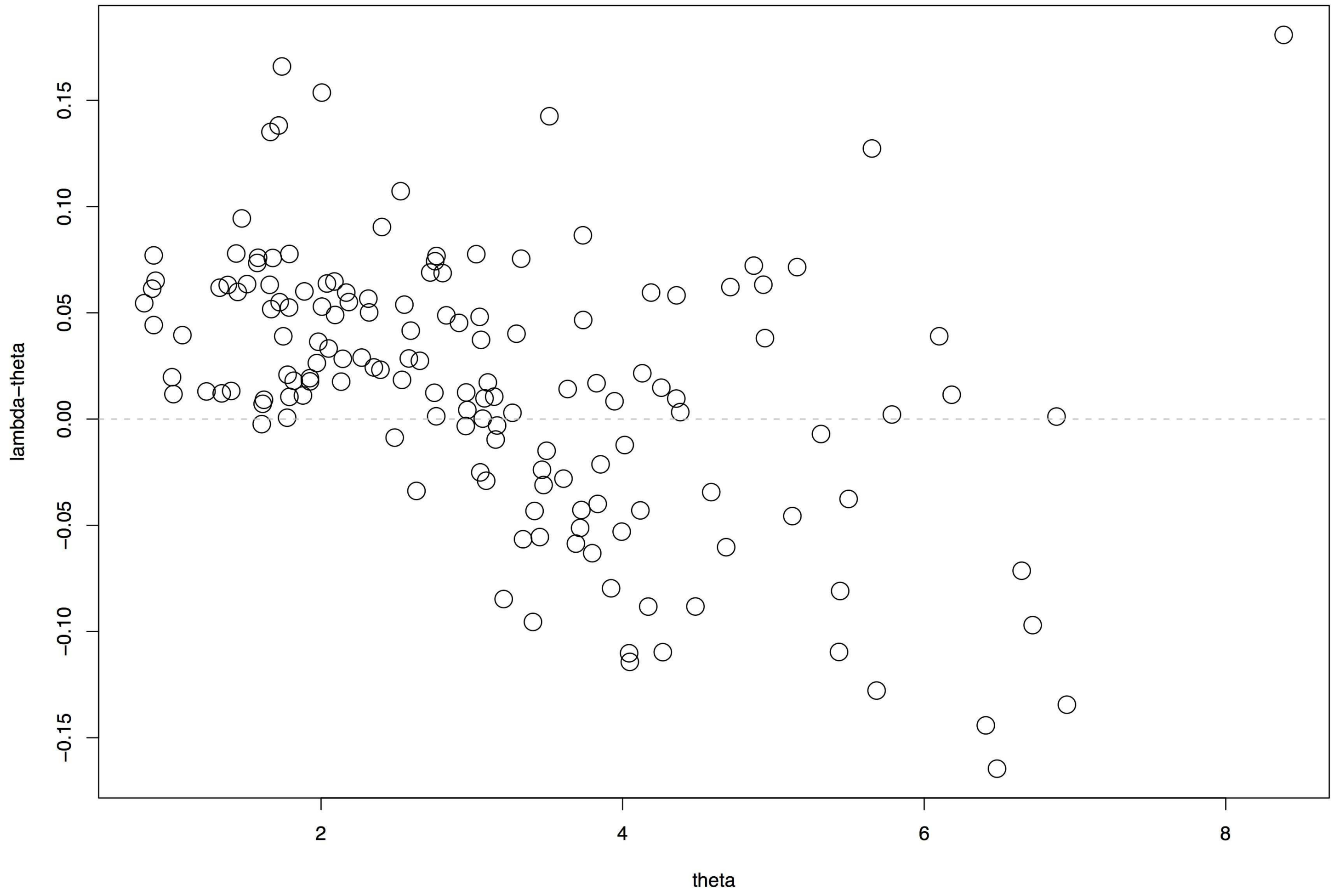
The models would be the same if $g_i(0) = 0$ so the difference arises when $g_i(0) > 0$: this induces different interpretations to π_i and p_i.

Parameter estimates

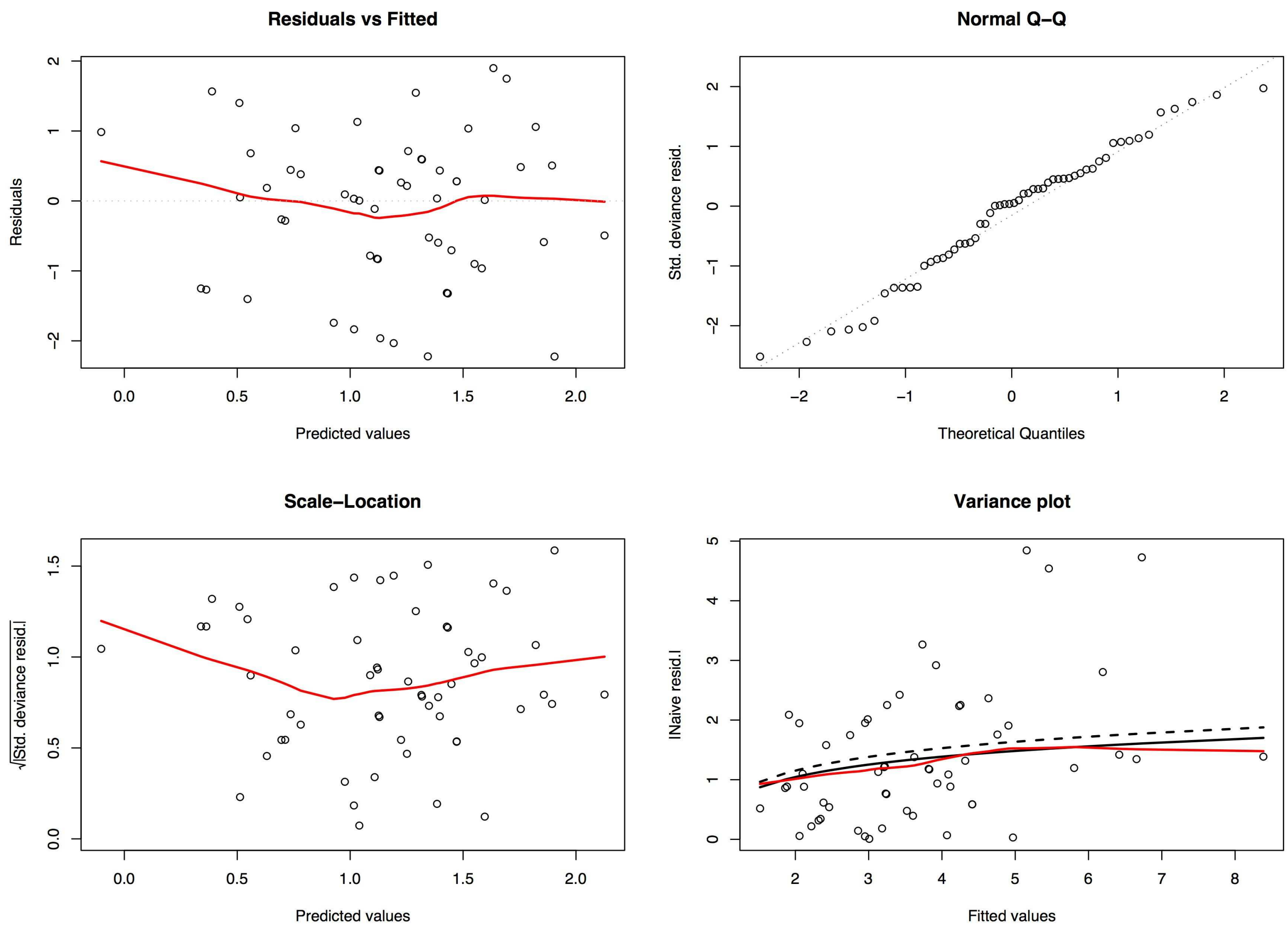
	Separated model			Overlapping model		
	Estimate	se	t-ratio	Estimate	se	t-ratio
(Intercept)	-3.288	0.645	-5.10	-2.987	0.678	-4.41
lstags	0.841	0.259	3.24	0.768	0.274	2.81
baa	0.093	0.024	3.93	0.090	0.0247	3.66
(Intercept)	1.130	0.333	3.39	1.080	0.329	3.28
lstags	0.246	0.111	2.21	0.249	0.111	2.25
bark	0.037	0.015	2.57	0.038	0.014	2.70
nos	-0.099	0.028	-3.48	-0.095	0.027	-3.52
slope	-0.031	0.013	-2.46	-0.028	0.012	-2.33

The similarity in the models for the probabilities is astonishing!

Estimated θ_i and λ_i

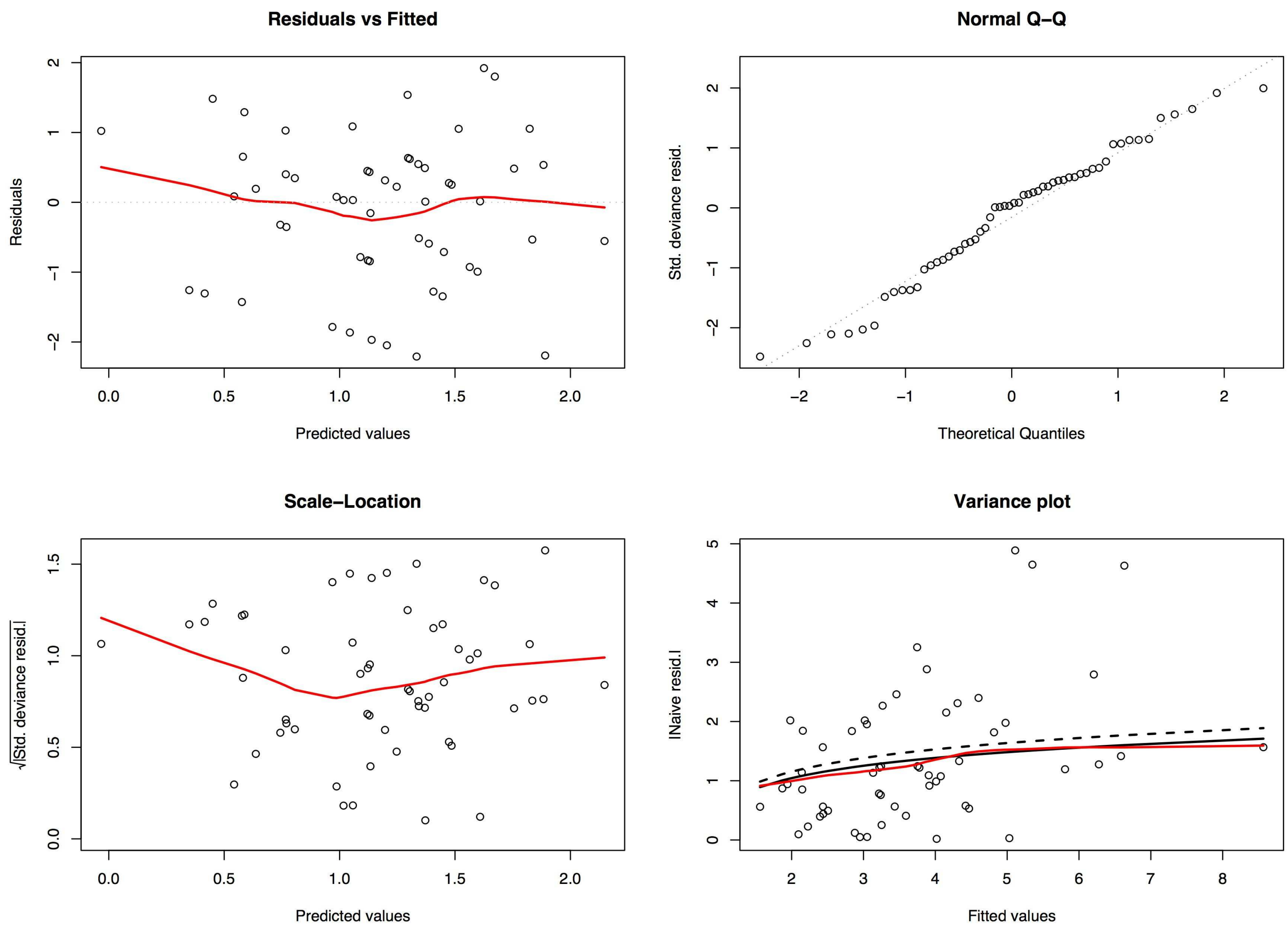


Separated Model - Abundance (selected model)



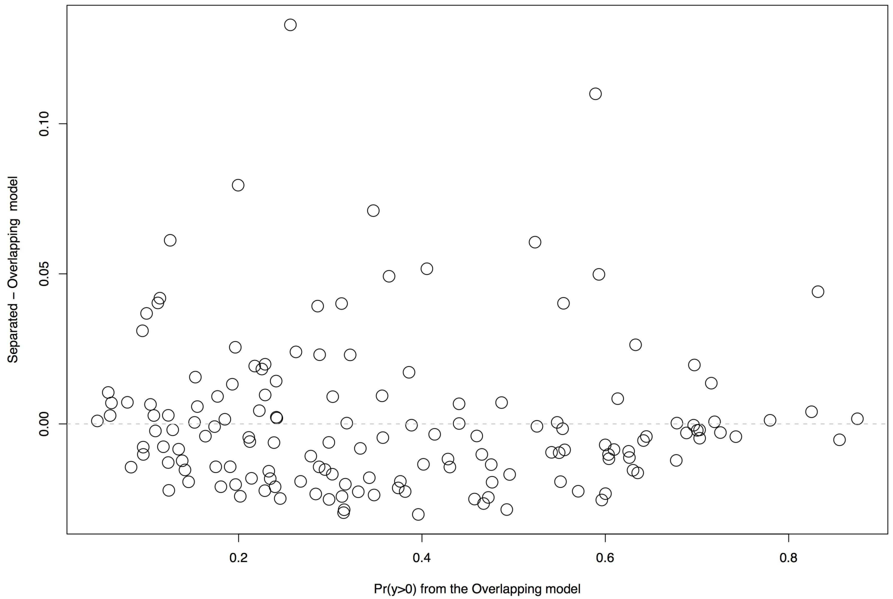
Residual mean deviance = 1.193

Overlapping Model - Abundance (selected model)



Residual mean deviance = 1.195

MODELLING $Pr(y_i = 0|\boldsymbol{x}_i)$



Measure	Separated	Overlapping
Deviance	169.1	169.6
AIC	175.1	185.6
BIC	184.1	209.8

The fits are actually essentially the same. In AIC and BIC we count parameters which don't contribute.

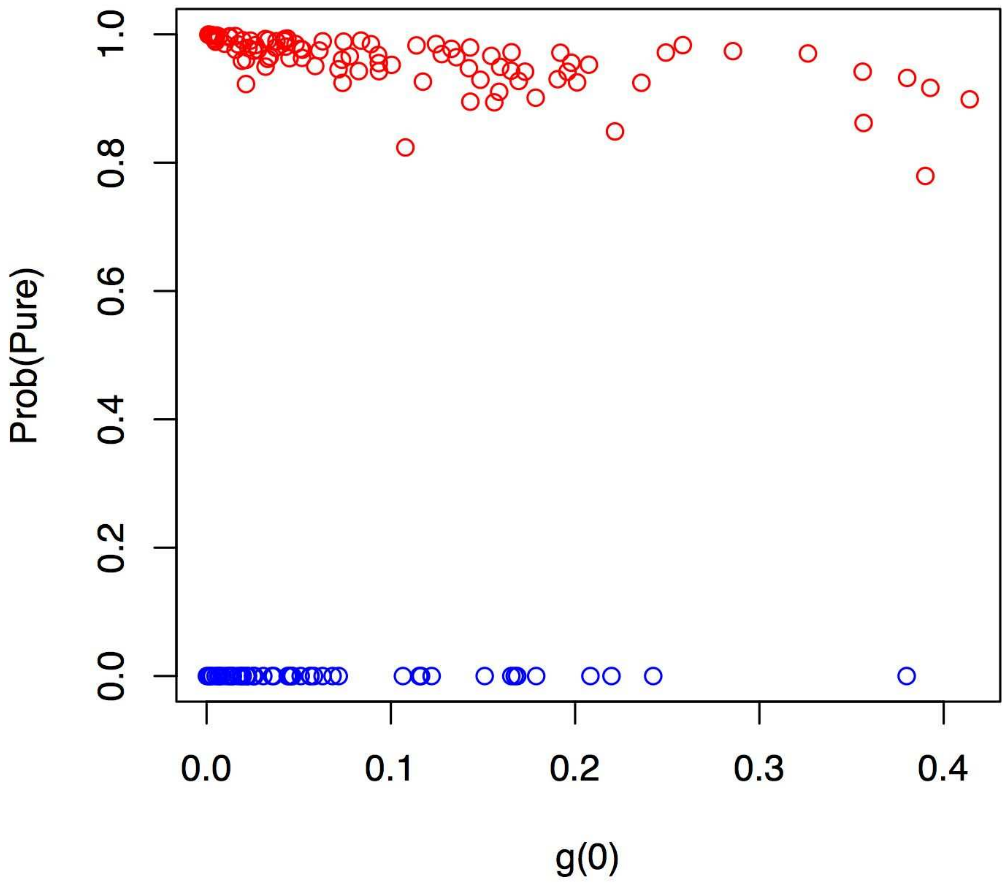
Identifying the Poisson zeros?

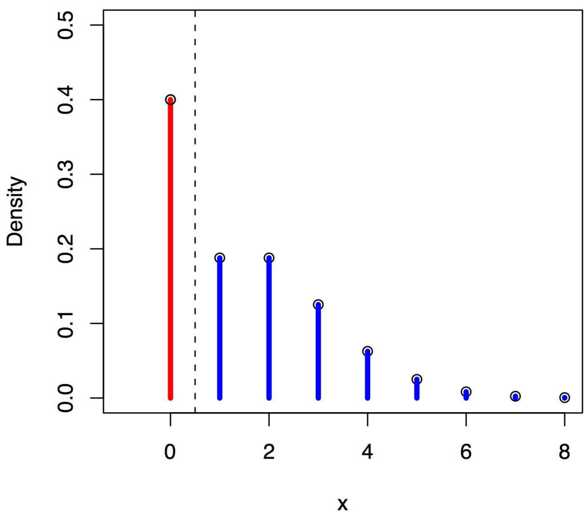
The data contain 95 zeros out of 151 observations. In the overlapping model, which of these are from the Poisson distribution?

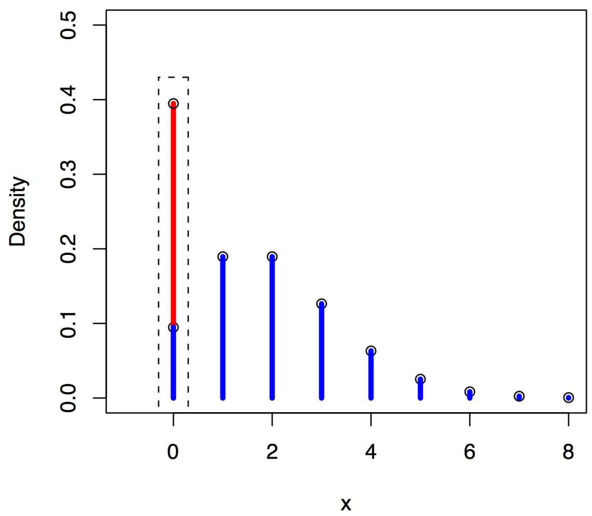
Empirical best prediction identifies none.

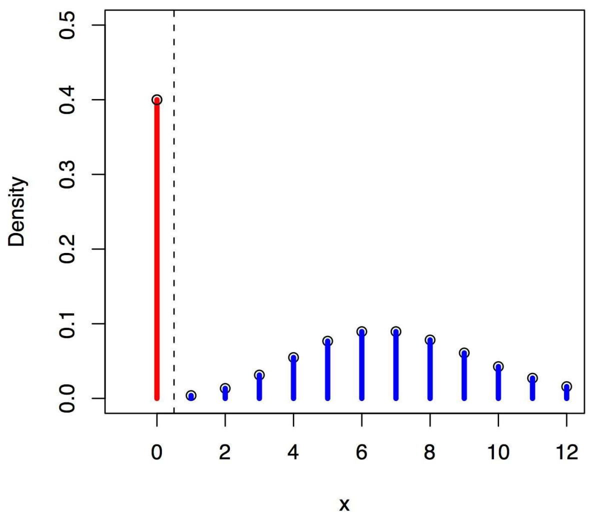
The fits from the two models are the same because there is NO overlap in the overlapping model. i.e. $g_i(0) = 0$.

This was suggested by and explains the the fact that the fitted binary models are the same.




Density
x
0.0
0.1
0.2
0.3
0.4
0.5
0
2
4
6
8


Density
x
0.0
0.1
0.2
0.3
0.4
0.5
0
2
4
6
8


Density
x
0.0
0.1
0.2
0.3
0.4
0.5
0
2
4
6
8
10
12

Comments

- The models can be made to give similar treatment of non-zero observations but differ in their treatment of zeros.
- The models are the same when $g_i(0) = 0$. i.e. for continuous non-zero data or when the non-zero counts are large. Thus the issues arise for data with extra zeros an small counts.
- Extra-zeros can induce overdispersion (relative to the Poisson model). This effect should be distinguished from overdispersion in the non-zero data (which requires modification to h_i or g_i).
- The presence-absence and abundance components can involve different covariates.

Acknowledgements