Robust analysis of Poisson and binary data using a mixture model approach

Ken Beath
Department of Statistics, Macquarie University, Sydney, Australia

Introduction

- A common problem in analysis is the presence of outliers or unusual observations, which may lead to incorrect inference
- Simplest technique is to remove the observations that are unusual
- This has a number of problems
  - When more than a small number of outliers they may affect the fitted model sufficiently so that it is difficult to determine which observations are outliers
  - Must choose a threshold to identify an outlier
  - An observation is either in the data set, or is removed in which case it provides no further information
- Method that avoids some of the problems are *M*-estimators, which downweight the observations based on how unusual they are, and possibly how influential they are. A method for generalized linear models is described in Cantoni and Ronchetti [2001].
- Problems:
  - Not likelihood based, so can no longer use related theory, for example methods for model selection, and are less efficient.
  - There are many options. Have choice of function used for down-weighting and tuning constants. Alternatives can give quite different results.

Robust mixture models

- A method for robust linear models was the model of Box and Tiao [1968] and Abraham and Box [1978] for Bayesian methods, but still applicable to frequentist methods.
- This has an underlying statistical model, so is suitable for Bayesian methods.
- Assumes a mixture model where observations are in either of two classes, standard or outliers.
- Both have the same underlying model, except that the error variance is greater in the outlier group.
- Aitkin and Wilson [1980] showed how models can be fitted as mixtures using the EM algorithm.
- Has some similarities to the Forward Selection method, which starts by choosing a homogeneous group and adds observations that are most similar.

Mixture models for Poisson and binary models

- For Poisson and binary it is not possible to modify the error variance directly, as it has a fixed relationship to the mean.
- Instead assume two classes:
  - Standard which is the usual Poisson or binary model.
  - Outlier which is an overdispersed Poisson or binary model, with the overdispersion achieved by incorporating an observation level random effect in the model.
- The only change to the generalized linear model is to modify the linear predictor. In the following we have class $c_i = 1$ for standard and $c_i = 2$ for outliers, and $\lambda_i \sim N\left(0, \tau^2\right)$

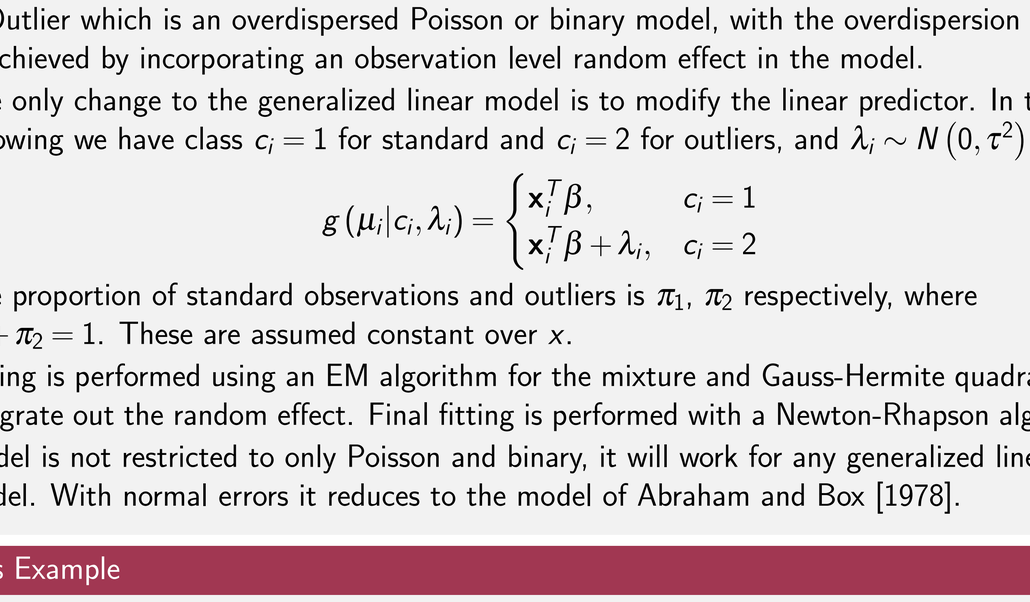
$$g\left(\mu_i | c_i, \lambda_i\right) = \begin{cases} \mathbf{x}_i^T \beta, & c_i = 1 \\ \mathbf{x}_i^T \beta + \lambda_i, & c_i = 2 \end{cases}$$

- The proportion of standard observations and outliers is π_1, π_2 respectively, where $\pi_1 + \pi_2 = 1$. These are assumed constant over x.
- Fitting is performed using an EM algorithm for the mixture and Gauss-Hermite quadrature to integrate out the random effect. Final fitting is performed with a Newton-Rhapson algorithm.
- Model is not restricted to only Poisson and binary, it will work for any generalized linear model. With normal errors it reduces to the model of Abraham and Box [1978].

Possums Example

- Study on the diversity of possums [Lindenmayer et al., 1990] used as an example in Cantoni and Ronchetti [2001]
- No outliers, so one introduced for observation 110, by changing response value from 2 to 5.
- Model has predictors Stags, Bark, Habitat and BAcacia and aspect. I have $\log(1+x)$ transformed Stags, Bark and BAcacia.

Outlier Effect

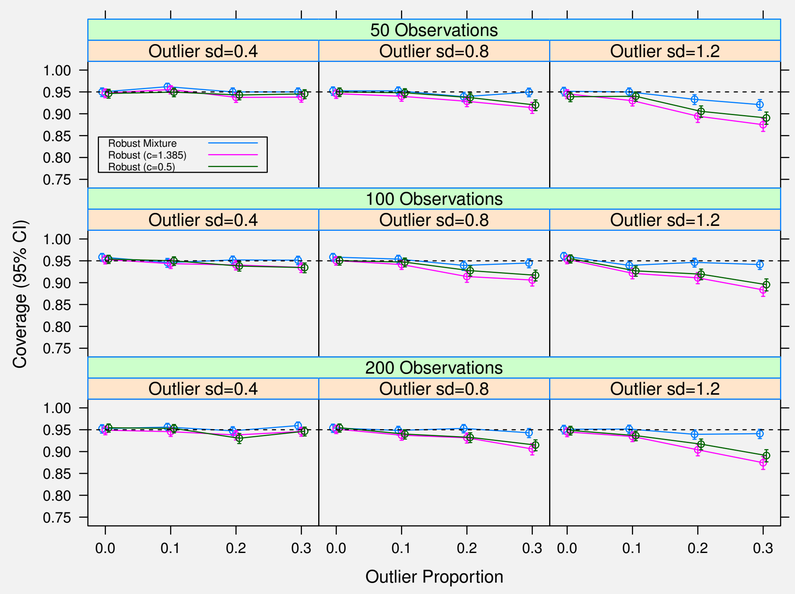


- Effect of robust mixture similar to that obtained using method of Cantoni and Ronchetti [2001]. In both cases estimates are similar to uncontaminated data.
- Robust mixture results in wider confidence intervals corresponding to increased coverage, which will be demonstrated in the simulations.

Simulations

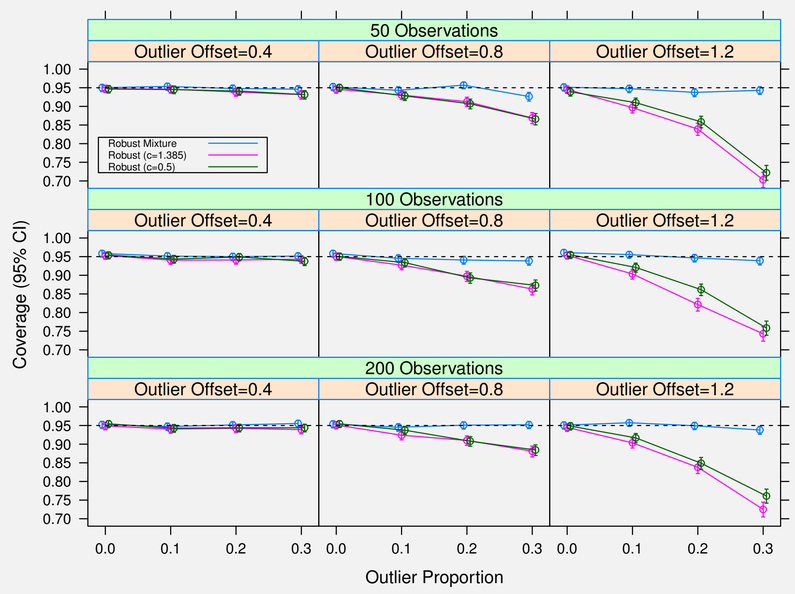
- Two simulations were performed to compare the robust Poisson based on an *M*-estimator (`glmrob` in the `robustbase` package in `R` without downweighting for leverage) and the robust mixture. `glmrob` was used with the default of $c = 1.385$ and $c = 0.5$. x is uniformly distributed between 0 and 1, and intercept $\beta_0 = 0.5$ and slope $\beta_1 = 0.0$. Coverage was calculated for β_1.
- Simulations are
  - Outliers are Poisson distributed with mean as for the standard data plus a normally distributed term. This is identical to the assumed mixture model.
  - Outliers are Poisson distributed with mean as for the standard data plus an offset term. This outlier distribution is different from that used for the robust mixture and is in only one direction.
- Each simulation consists of 2000 data sets.
- Varies proportion of outliers from 0.0 to 0.3 and standard deviation or offset of the outlier effect.
- Fitted using Latent GOLD with Syntax Module.

Coverage - Simulation 1



- Mixture model has superior coverage, especially for larger proportion of outliers. Slight improvement with `glmrob` using the lower value of c.

Coverage - Simulation 2



- No change in performance of robust mixture model from Simulation 1.
- Cantoni and Ronchetti [2001] worse than Simulation 1, as these are more extreme outliers.

Conclusions and Further Work

- Robust mixture is superior in simulations for the outlier patterns used.
- Robust mixture also allows some other features, the identification of outliers using posterior class probabilities and a test for presence of outliers using a parametric bootstrap.
- Method could be amended to allow downweighting for influential observations similar to what is available with *M*-estimators.
- Method also allows automatically for overdispersed data. Additional mixture components could also be included to allow for overdispersed data plus outliers.

References

B. Abraham and G. E. P. Box. Linear Models and Spurious Observations. *Applied Statistics*, 27 (2):131–138, 1978.

M. Aitkin and G. Wilson. Mixture models, outliers, and the EM algorithm. *Technometrics*, 22(3): 325–331, 1980.

G. Box and G. Tiao. A Bayesian Approach to some outlier problems. *Biometrika*, 55(1):119–129, 1968.

E. Cantoni and E. Ronchetti. Robust Inference for Generalized Linear Models. *Journal of the American Statistical Association*, 96(455):1022–1030, 2001.

D. Lindenmayer, R. Cunningham, M. Tanton, and A. Smith. The conservation of arboreal marsupials in the montane ash forests of the Central Highlands of Victoria, south-east Australia: II. The loss of trees with hollows and its implications for the conservation of leadbeater's possum Gymnobelideus leadbeateri. *Biological Conservation*, 54:133–145, 1990.