Mean and Variance Modeling of Under and Over Dispersed Count Data

D.M.Smith*

Abstract
A family of models based on the extended Poisson process that can flexibly handle both under- and over-dispersion compared to the Poisson and negative binomial distributions will be described. Many sets of count data display such under- or over-dispersion and although there are a number of distributional models that can handle over-dispersion, there is a lack of models that can handle under-dispersion. Models with mean and variance related to covariates can also be constructed within this family using a generalized linear model formulation; estimation of parameters being by maximum likelihood. An R package for fitting such models will be described, and its use to analyze health outcomes and other types of health related data illustrated.

Key Words: Poisson distribution, negative binomial distribution, under-dispersion, overdispersion.

1. Introduction

- Discrete probability distributions having very general dispersion properties can be modeled using extended Poisson process models (EPPMs), Faddy (1997).
- Poisson & negative binomial distributions are special cases including both under-dispersion and over-dispersion relative to the Poisson, the negative binomial having the most extreme level of over-dispersion.
- Covariate dependences in the mean & variance are incorporated via re-parameterisations using approximate forms, Faddy & Smith (2011).
- A software package for EPPMs has been developed in R; comments of Hilbe (2014, p274).

2. Extended Poisson Process Models (EPPMs)

2.1 Definition

$$\text{The defining equation is: } \boldsymbol{p} = (1\,0\,\cdots\,0)\exp(\boldsymbol{Q}) \tag{1}$$

where $\boldsymbol{p}$ is a row vector of probabilities and $\boldsymbol{Q}$ is the matrix:

$$\boldsymbol{Q} = \begin{bmatrix} -\lambda_0 & \lambda_0 & 0 & \cdots & 0 \\ 0 & -\lambda_1 & \lambda_1 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & -\lambda_n \end{bmatrix}$$

with the λ_i parameters rates of an extended Poisson process, and $\exp(\boldsymbol{Q})$ the matrix exponential function.

- Constant λ's correspond to the simple Poisson process, the distribution of the number of events in a time interval of length 1 being Poisson with mean λ.

*Truven Health Analytics, 7700 Old Georgetown Road, Suite 650, Bethesda, MD 20814

- The extension has the λ's depending on i, the cumulative number of events occurring.
- The probability p_i of obtaining a count of size i $(i = 0, 1, \ldots, n)$ in the time interval is the $(i + 1)$th element of $\boldsymbol{p}$, so discrete probability distributions can be constructed from a sequence of λ's.

2.2 λ_i sequence

A three parameter function for the λ_i sequence:

$$\lambda_i \;=\; a(b \;+\; i)^c, \text{ for } i = 0, 1, 2, \ldots, \text{ where } a > 0,\; b > 0 \text{ and } c \leq 1 \tag{2}$$

produces discrete distributions including the Poisson ($c = 0$) and negative binomial ($c = 1$) as special cases with the probability mass function for the negative binomial distribution being in the form:

$$\mathrm{P}\{r = i\} \;=\; \left(\begin{array}{c} b+i-1 \\ b-1 \end{array} \right) \{1 \;-\; \exp(-a)\}^i \{\exp(-a)\}^b, \;\; \text{for } i = 0, 1, \ldots \infty.$$

Generally, $c > 0$ in equation (2) results in distributions over-dispersed relative to the Poisson distribution, and $c < 0$ distributions under-dispersed relative to the Poisson.

2.3 Means and variances

The mean and variance of distributions from EPPMs defined by equation (2) generally have to be determined numerically directly from the probability distribution of equation (1) using a suitable truncation (n); however, approximations are available:

$$\text{mean} \;\approx\; m \quad = \quad b \left[\left(1 + \frac{a(1-c)}{b^{1-c}} \right)^{\frac{1}{1-c}} \;-\; 1 \right] \tag{3}$$

$$\text{variance} \;\approx\; v \quad = \quad \frac{b\left(\frac{m}{b} \;+\; 1\right) \left[1 \;-\; \left(\frac{m}{b} \;+\; 1\right)^{(2c-1)} \right]}{1 \;-\; 2c} \;. \tag{4}$$

For $c = 1$ the variance v of equation (4) is that of the NB2 distribution of Hilbe (2014).

2.4 Obtaining parameters a, b, c

The mean approximation is used to obtain a in equation (2) as:

$$a \;=\; \frac{(m+b)^{1-c} - b^{1-c}}{1-c} \quad \left[\;=\; \log\left(1 \;+\; \frac{m}{b}\right) \text{ for } \; c = 1 \right] \;. \tag{5}$$

Covariates $\boldsymbol{x}$ are incorporated using a log link:

$$\log(m) \;=\; \text{linear predictor}(\boldsymbol{x}) \;. \tag{6}$$

Covariate dependence in the variance has v as a function of another linear predictor, then solving the equation:

$$\frac{v}{m \;+\; b} \;=\; \frac{\exp\{(2c \;-\; 1)\log(m/b \;+\; 1)\} \;-\; 1}{2c \;-\; 1} \tag{7}$$

for c in terms of v, m and b.The right-hand-side of equation (7) is a convex increasing function of c, so Newton-Raphson iteration (initial value $c_0 = 1$) converges to the solution.

2.5 Limiting form

Poor convergence to the maximum can occur for under-dispersed data if the parameter b becomes large and the hessian at an apparent maximum is poorly conditioned. In such circumstances a limiting form of the λ_i's of equation (2) for $b \to \infty$ with $a = \alpha b^{-b/\beta}$ and $c = b\beta$, and a special case of equation (2) can be used (Faddy & Smith, 2011):

$$\lambda_i = \alpha \exp(\beta i),\ i = 0, 1, 2, \ldots, \tag{8}$$

with $\beta < 0$, corresponding to $c < 0$ in (2) and under-dispersion; the approximate mean and variance are now given by:

$$m = \frac{-\log(1 - \alpha\beta)}{\beta} \tag{9}$$

$$\text{and } v = \frac{\exp(2\beta m) - 1}{2\beta} \tag{10}$$

with solutions of equations (9) and (10) achieving the re-parameterisation in terms of the approximate mean and variance.

2.6 Comments on models

- Covariate dependence in the variance permits modelling of datasets where some subsets exhibit under-dispersion and others over-dispersion.
- If a solution $c > 1$ is indicated, that parameterisation is inadmissible since the resulting probability distribution will be improper.
- The approximations are only used to derive the re-parameterisation of equations (5) and (7); exact calculation of means and variances are done numerically from the probability distribution equation (1) to minimise the effect of this approximation.
- For $c \neq 0, 1$ the relationship between log-mean and covariates $\boldsymbol{x}$ is not quite linear, but any departure from linearity is often almost imperceptible.
- Parameter b in equation (2) is a nuisance parameter, and can be poorly estimated; it is better estimated as $\log(b)$.
- The model is specified by the two parameters a and c in equation (2) dependent on covariates with b always a single scalar-valued constant.
- When fitting mean and variance models of equations (3) and (4) (or equations (9) and (10)), the use of a log link function for both mean and variance means that the scale factor (variance/mean) can be modelled rather than the variance by simply including log(mean) as an offset in the linear predictor for the variance.
- An alternative formulation only involving the mean model is available Faddy & Smith (2005), where only the proportionality constant a in equation (2) (or α in equation (8)) is dependent on the covariates $\boldsymbol{x}$; this does not actually quantify the mean but is based on increasing (decreasing) values of a leading to increased (decreased) rates λ_i's of the extended Poisson process and hence increased (decreased) counts on average.

3. R user contributed package CountsEPPM

- Focused on models with two covariate dependences linked to the mean and variance.
- Input is a formula involving a single dependent variable and one or two linear predictors related to the mean and variance models.
- The link function between both dependent variables (mean, variance) and linear predictor is log.
- Log of parameter b is used but parameter c of equation (2) is untransformed.

```
CountsEPPM(formula,data,model.type,model,offset,initial,
     ltvalue,utvalue, optimization.method,control,scale.factor.model)
```

Output is statements of type of data input; optimization method used with return code; the number of iterations; and warning messages. An object summarizing the model fit is returned.

3.1 Other functions linked to CountsEPPM

Two other functions are available.

- CountsEPPM.summary prints out a summary in GLM like form of, model information, estimates of parameters with their standard errors, log likelihood and Akaike's Information Criterion (AIC).
- CountsEPPM.distribution produces an object consisting of the fitted means, variances and, optionally, total probabilities and/or parameters (a, b, c) of the distributions of equation (2).

Both functions have the output object from CountsEPPM, named output.fn here, as an input object.

3.2 Example

These data on takeover bids are from Cameron & Trivedi (2013) and are used as example data in Sáez-Castillo & Conde-Sánchez (2013). The data are included in the data sets of Version 2.0 of CountsEPPM, but not in those of Version 1.0. Three sets of inputs and outputs are shown, one for input of data as a list of frequency distributions for Version1.0; one for input of data as a list of frequency distributions for Version 2.0; and one for input of data as a data.frame of individual count data for Version 2.0.

3.2.1 Version 1.0 of CountsEPPM

Input of data as a list for version 1.0. The same data can not be entered as a data.frame in this version.

```
> takeover.bids.frequency <- list(covariates=data.frame(
+    LEGLREST=factor(c(0,1,0,1),levels=c(0,1)),
+    REALREST=factor(c(0,0,1,1),levels=c(0,1))),
+    list.counts=list(c(3,36,16,6,1,0,1),
+    c(1,18,11,5,3,1,1),c(3,4,1,1),
+    c(2,5,3,0,2,0,0,1,0,0,1)))
```

Function call

```
> output.fn  <- CountsEPPM(
+    mean.obs | scalef.obs~LEGLREST+REALREST |
+    LEGLREST+REALREST,data=takeover.bids.frequency,
+    optimization.method='nlm',scale.factor.model='yes')
```

Output from function call

```
optimization method nlm:
code        5
Maximum step size stepmax exceeded five consecutive times.
Either the function is unbounded below, becomes asymptotic
to a finite value from above in some direction, or stepmax
is too small
iterations  45
```

Function call

```
> CountsEPPM.summary(output.fn)
```

Output from function call

```
Model type: mean and variance
Model     : general
link      : log
scale factor model fitted
Parameter estimates and se's
        name    estimates          se
 (Intercept)  -0.16154702  0.17067110
   LEGLREST1   0.57425571  0.21100182
   REALREST1   0.72713232  0.26872650
 (Intercept)  -0.11598728  0.08637971
   LEGLREST1   0.03840492  0.04644361
   REALREST1   0.15213170  0.12343674
   log(beta) -21.00042698 16.66060084

log likelihood  -181.7724

AIC 377.5449
```

3.2.2 Version 2.0 of CountsEPPM

Input of data as a list for version 2.0.

```
> takeover.bids.frequency <- list(
+   LEGLREST=factor(c(0,1,0,1),levels=c(0,1)),
+   REALREST=factor(c(0,0,1,1),levels=c(0,1)),
+   NUMBIDS=list(c(3,36,16,6,1,0,1),
+   c(1,18,11,5,3,1,1),c(3,4,1,1),
+   c(2,5,3,0,2,0,0,1,0,0,1)))
```

Function call

```
> output.fn  <- CountsEPPM(NUMBIDS~LEGLREST+REALREST |
+   LEGLREST+REALREST,data=takeover.bids.frequency,
+   optimization.method='nlm',scale.factor.model='yes')
```

Output from function call

```
Dependent variable is a list of frequency distributions
of counts

optimization method nlm:
iterations   44
return code  5
Maximum step size stepmax exceeded five consecutive times.
Either the function is unbounded below, becomes asymptotic
to a finite value from above in some direction, or stepmax
is too small
```

Function call

```
> CountsEPPM.summary(output.fn)
```

Output from function call

```
Model type: mean and variance
Model     : general
Link for mean         : log
Link for scale factor : log
scale factor model model fitted
Parameter estimates and se's
        name    Estimates          se
 (Intercept)  -0.16405142  0.17503584
   LEGLREST1   0.57888308  0.21248704
   REALREST1   0.73263975  0.27050074
 (Intercept)  -0.11667630  0.10066490
   LEGLREST1   0.03828664  0.04874788
   REALREST1   0.15300618  0.14208048
      log(b) -20.99072940 19.36595466

log likelihood  -181.7717
AIC 377.5435
```

Input of data as a data.frame for version 2.0.

```
    NUMBIDS LEGLREST REALREST
1         2        1        0
2         0        0        0
3         1        1        0
...     ...      ...      ...
124       2        1        0
125       0        0        1
126       1        1        0

takeover.bids.case <- data.frame(
                            NUMBIDS,LEGLREST,REALREST)
```

Function call

```
> output.fn  <- CountsEPPM(NUMBIDS~LEGLREST+REALREST |
+  LEGLREST+REALREST,data=takeover.bids.case,
+  optimization.method='nlm',scale.factor.model='yes')
```

Output from function call

```
Dependent variable is a vector of single counts.

optimization method nlm:
iterations   29
return code  5
Maximum step size stepmax exceeded five consecutive times.
Either the function is unbounded below, becomes asymptotic
to a finite value from above in some direction, or stepmax
is too small
```

Function call

```
> output.fn$estses[[1]] <- c('(Intercept) mean',
+     'LEGLREST mean','REALREST mean',
+     '(Intercept) variance','LEGLREST variance',
+     'REALREST variance','log(b)')
> CountsEPPM.summary(output.fn)
```

Output from function call

```
Model type: mean and variance
Model     : general
Link for mean         : log
Link for scale factor : log
scale factor model model fitted
Parameter estimates and se's
                 name    Estimates         se
     (Intercept) mean  -0.13365633 0.14723721
        LEGLREST mean   0.56599546 0.20130003
        REALREST mean   0.70554506 0.26041163
 (Intercept) variance  -0.15806731 0.03709528
    LEGLREST variance   0.05333941 0.04854162
    REALREST variance   0.21406529 0.06820072
               log(b) -15.15881410 2.20814777

log likelihood  -181.8567
AIC 377.7134
```

3.2.3 Comment

This data set has proved to be difficult to fit models to as illustrated in Sáez-Castillo & Conde-Sánchez (2013), and by the variation in the outputs given here for the different versions and forms of data. What is presented here should be considered as an illustration of how to use these R functions, which are still in development rather than as definitive analyses of this data set.

4. Concluding remarks

- The R package CountsEPPM uses EPPMs to model the means & variances of count data that exhibit under- or over-dispersion relative to the Poisson distribution.
- The mean and variance of binary data can be modeled using EPPMs in a similar way to that described here for count data, Faddy & Smith (2012). A R function to do this is in development.
- Version 1.0 is currently on CRAN as a user contributed package, Smith & Faddy (2013). Version 2.0 is still being developed. An article for the Journal of Statistical Software is also under revision.

5. Acknowledgement

Much of the theoretical development and promotion of EPPMs for the analysis of count and binary data has been by my collaborator and friend Dr Malcolm Faddy. His support for, and assistence with, this article is much appreciated.

REFERENCES

Cameron, A., Trivedi, P. (2013), *Regression Analysis of Count Data* (2nd ed.), Cambridge, UK: Cambridge University Press.

Faddy, M. (1997)," Extended Poisson Process Modelling and Analysis of Count Data", *Biometrical Journal*, **39**, 431-440.

Faddy, M., Smith, D. (2005), "Modelling the Dependence between the Number of Trials and the Success Probability in Binary Trials", *Biometrics*, **61**, 1112-1114.

Faddy, M., Smith, D. (2011), "Analysis of Count Data with Covariate Dependence in Both Mean and Variance", *Journal of Applied Statistics*, **38**, 2683-2694.

Faddy, M., Smith, D. (2012), "Extended Poisson Process Modeling and Analysis of Grouped Binary Data", *Biometrical Journal*, **53**, 426-435.

Hilbe, J. (2014), *Modeling Count Data*, Cambridge, UK: Cambridge University Press.

Sáez-Castillo, A.J., Conde-Sánchez,A. (2013), "A hyper-Poisson regression model for overdispersed and underdispersed count data", *Computational Statistics and Data Analysis*, **61**, 148-157.

Smith, D.M., Faddy, M.J. (2013), 'CountsEPPM'.
R package version 1.0, URL http://CRAN.R-project.org/package=CountsEPPM .