Fitting Extra-Poisson Variation in Poisson Regression Models using *Arc*

Luca Scrucca
Department of Statistics, Università degli Studi di Perugia, Italy
luca@stat.unipg.it

March 21, 2000

Abstract

This document describes the fitting of overdispersed Poisson log-linear models using *Arc*, a computer program for the analysis of regression data, as described in Cook and Weisberg (1999), which can be obtained from the Web at `http://www.stat.umn.edu/arc`.

The code is written in *Xlisp-Stat* and is contained in the file `poisson-extra-var.lsp`, available on the Web at the address `http://www.stat.umn.edu/arc/addons` or at `http://www.stat.unipg.it/~luca/xlispstat`. To use it, download the file and put it in your `Extras` directory.

1 Introduction

Breslow (1984) proposed an iterative algorithm for fitting a dispersion parameter in overdispersed Poisson regression models. The method is similar to that proposed by Williams (1982) for handling overdispersion in logistic regression models. The lisp code for the latter method is already available in the standard distribution of *Arc*, and the corresponding documentation may be found at the address `http://www.stat.umn.edu/arc/addons`.

Suppose we observe n independent responses such that

$$y_i|\lambda_i \sim \text{Poisson}(\lambda_i n_i) \qquad i = 1 \ldots, n$$

The response variable y_i may be an event counts variable observed over a period of time (or in the space) of length n_i, whereas λ_i is the rate parameter. Then,

$$\text{E}(y_i|\lambda_i) = \mu_i = \lambda_i n_i = \exp(\log(n_i) + \log(\lambda_i))$$

where $\log(n_i)$ is an offset and $\log(\lambda_i) = \boldsymbol{\beta}^\top \mathbf{x}_i$ expresses the dependence of the Poisson rate parameter on a set of, say p, predictors. If the periods of time are all of the same length, we can set $n_i = 1$ for all i so the offset is zero. MLE for $\boldsymbol{\beta}$ are usually obtained by iterative methods, such as Newton-Raphson or Fisher scoring, also called IRWLS (*Arc* uses the latter iterative procedure).

The Poisson distribution has $\text{E}(y_i|\lambda_i) = \text{Var}(y_i|\lambda_i)$, but it may happen that the actual variance exceeds the nominal variance under the assumed probability model. Suppose now that $\theta_i = \lambda_i n_i$ is a random variable distributed according to

$$\theta_i \sim \text{Gamma}(\mu_i, 1/\phi)$$

where $\mathrm{E}(\theta_i) = \mu_i$ and $\mathrm{Var}(\theta_i) = \mu_i^2\phi$. Thus, it can be shown that the unconditional mean and variance of y_i are given by

$$\mathrm{E}(y_i) = \mu_i$$

and

$$\mathrm{Var}(y_i) = \mu_i + \mu_i^2\phi = \mu_i(1 + \mu_i\phi)$$

Hence, for $\phi > 0$ we have overdispersion. It is interesting to note that the same mean and variance arise also if we assume a negative binomial distribution for y.

The method proposed by Breslow uses an iterative algorithm for estimating the dispersion parameter ϕ and hence the necessary weights $(1 + \mu_i\hat{\phi})^{-1}$ (for details see Breslow, 1984).

2 Fitting using *Arc*

To fit extra-Poisson variation in *Arc* you need to load the file `poisson-extra-var.lsp`. If the file is in the `Extras` directory, then it is automatically loaded at the start-up.

Thus, every time you fit a Poisson log-linear model a new item called **Fit extra Poisson variance** appears on the model's menu. Selecting the item, the iterative fitting procedure for the estimation of the dispersion parameter starts. At convergence, the usual summary for a GLM regression model is shown with a new information, the estimate $\hat{\phi}$. In addition, the final estimated weights used in the fitting procedure are added to the dataset for using in subsequent modelling stages.

The variable `*poisson-extra-variance*` (by default set at T) controls the appearance of the item on the model's menu for all Poisson models. If you do not want the item to appear, you may change the value of `*poisson-extra-variance*` to NIL, using the **Settings** item from the **Arc** menu.

Example: Ames Salmonella reverse mutagenicity assay The following table shows the numbers of revertant colonies of TA98 Salmonella (`Salm`) observed on each of three replicate plates tested at each of six dose levels of quinoline (`Dose`):

Doses of quinoline					
0	10	33	100	333	1000
15	16	16	27	33	20
21	18	26	41	38	27
29	21	33	60	41	42

The model proposed for the expected rate $\lambda(x)$ of revertants as a function of dose x is the log-linear model:

$$\log(\lambda(x)) = \alpha + \beta\log(x + 10) - \gamma x$$

where β is the mutagenic effects of dose, and γ is the toxic effect.

We start fitting the above model.

```
Data set = SalmonellaTA98, Name of Fit = P1
Poisson Regression
Kernel mean function = Exponential
Response      = salm
Terms         = (log[dose+10] dose)
Coefficient Estimates
```

```
Label          Estimate        Std. Error     Est/SE
Constant       2.17277         0.218309        9.953
log[dose+10]   0.319825        0.0569740       5.614
dose          -0.00101303      0.000245118    -4.133
Scale factor:                  1.
Number of cases:              18
Degrees of freedom:           15
Pearson X2:               46.229
Deviance:                 43.716
```

Then, we select the item **Fit extra Poisson variance** on the corresponding model's menu.

```
Extra-Poisson Variation in Log-linear Models
Fitting . . .
Converged after 3 iteration(s)

Data set = SalmonellaTA98, Name of Fit = P1
Poisson Regression
Kernel mean function = Exponential
Response      = salm
Terms         = (log[dose+10] dose)
Weights       = Extra-poisson-var
Coefficient Estimates
Label          Estimate        Std. Error     Est/SE
Constant       2.20302         0.363361        6.063
log[dose+10]   0.310977        0.0989867       3.142
dose          -0.000974122     0.000436848    -2.230

Scale factor:                  1.
Number of cases:              18
Degrees of freedom:           15
Pearson X2:               14.992
Deviance:                 14.219
Extra variation:           0.072
```

Thus, the estimated dispersion parameter is $\hat{\phi} = 0.072$, and consequently the standard errors of the regression coefficients increase, whereas the values of coefficient estimates are almost the same. As for overdispersed binomial models, the goodness of fit statistics are no longer meaningful, and we must rely on graphical checking, such as residuals plots and marginal model plots.

References

Breslow N.E. (1984), Extra-Poisson variation in log-linear models, *Applied Statistics*, 33, pp. 38-44.

Cook R.D. and Weisberg S. (1999), *Applied Regression Including Computing and Graphics*. New York: Wiley.

McCullagh P. and Nelder J.A. (1989), *Generalized Linear Models*. (2nd ed.), London: Chapman and Hall.

Tierney L. (1990), *Lisp-Stat: An Object-Oriented Environment for Statistical Computing and Dynamic Graphics*. New York: Wiley.

Williams D.A. (1982), Extra-binomial variation in logistic linear models, *Applied Statistics*, 31, pp. 144-148.