Statistical Methods & Applications (2001) 10: 3–9

SMA


# A family of repeated measurements models

**J.K. Lindsey**

Biostatistics, Limburgs Universitair Centrum, 3590 Diepenbeek, Belgium (e-mail: jlindsey@luc.ac.be)

**Abstract.** A general family of multivariate distributions for repeated measures can be obtained by applying the Laplace transform of a gamma distribution to the integrated intensity function of any continuous distribution on the positive real line. Both clustering and serial dependence can be handled. The response variable may be counts, durations between events, or any continuous positive-valued measurements.

**Key words:** Clustering, counts, event history, gamma distribution, integrated intensity, Kalman filter, Laplace transform, longitudinal study, negative binomial process, Pareto distribution, random effects, serial dependence.

## 1 Introduction

Two basic types of repeated measurements arise from clustered and from longitudinal studies (Lindsey, 1993). Consider the latter first. Suppose that each observed individual has some process operating over continuous time, $t$, although observations will necessarily occur at discrete points to be indexed by $i$. Associated with each of these will be some recorded value of a random variable; this may be a count of the number of events between $t_{i-1}$ and $t_i$, say $n_i$, the measurement of some continuous positive-valued variable, say $y_i$, or the time, itself, since the last event, $y_i = t_i - t_{i-1}$.

Let us consider first these latter two cases where $Y_i$ is the random variable. Suppose that a cumulative distribution of interest is

$$F(y_i; \boldsymbol{\theta}) = 1 - \exp\{-H(y_i; \boldsymbol{\theta})\} \tag{1}$$

where, in the context of survival distributions, $H(y_i; \boldsymbol{\theta})$ is the corresponding integrated intensity function. We seek multivariate generalizations that will allow us to account for serial dependence in longitudinal data or intra-class (frailty) dependence under clustering.

From Equation (1), the integrated intensity function $H(y_i;\boldsymbol{\theta})$ has a unit exponential distribution. Thus, any distribution on the positive real line can be transformed to an exponential distribution by transforming the response variable using $H(y_i;\boldsymbol{\theta})$. As a member of the exponential family, this latter distribution has a number of powerful and simple properties (see Lindsey, 1996, Ch. 2), including those related to its conjugate distribution and to its Laplace transform, that will be used below. The interesting point is that these properties are transferred to *any* distribution on the positive real line once the transformation of the response variable has been performed.

Here and at various points below, I recall some elementary results from probability theory and model construction. In order to handle count data as well as continuous variables, we require the following results showing the relationship between the two.

If a sequence of counts of events over time, $n_i$, follows a Poisson process with intensity $\lambda$ and probability

$$\Pr(N_i = n_i;\lambda) = \frac{\mathrm{e}^{-\lambda z_i}(\lambda z_i)^{n_i}}{n_i!}$$

over the observation times $z_i$, then

$$\begin{aligned}\Pr(N_i = 0;\lambda) &= \mathrm{e}^{-\lambda z_i}\\ &= \Pr(Z_i > z_i;\lambda)\end{aligned}$$

is the survival function of the exponential distribution so that the times between events have this distribution. Then, from the results above, the times between events, $y_i$, from *any* such cumulative distribution $F(y_i;\boldsymbol{\theta})$ can be transformed to a Poisson process using the time transform, $z_i = H(y_i;\boldsymbol{\theta})$, with $\lambda = 1$.

If the Poisson intensity parameter $\lambda(= s)$ is allowed to vary stochastically following a gamma distribution,

$$f(s) = \frac{\beta^{\alpha}s^{\alpha-1}\mathrm{e}^{-\beta s}}{\Gamma(\alpha)}$$

which is its conjugate distribution, then the resulting mixture distribution is negative binomial with probability

$$\Pr(N_i = n_i;\alpha,\beta) = \frac{\Gamma(\alpha+n_i)}{\Gamma(\alpha)n_i!}\left(\frac{z_i}{\beta+z_i}\right)^{n_i}\left(\frac{\beta}{\beta+z_i}\right)^{\alpha} \tag{2}$$

Then, in a similar way to the development for the Poisson process, if a sequence of overdispersed counts follows a negative binomial process,

$$\begin{aligned}\Pr(N_i = 0;\alpha,\beta) &= \left(\frac{\beta}{\beta+z_i}\right)^{\alpha} \qquad (3)\\ &= \Pr(Z_i > z_i;\alpha,\beta)\end{aligned}$$

is the survival function for some variable $Z_i$, arising from a special case of the Pareto distribution. Here, the average intensity over the interval $z_i$ is $\alpha/\beta$. Below I shall develop a relationship between this equation and Equation (1).

With such results, we can easily pass from times to counts, depending on which of them is random. But recall that $Y_i$ in Equation (1) may also be any positive-valued random variable, not necessarily time.

To obtain a mixture distribution for the transformed responses that are varying proportionally in the population according to a gamma distribution, take them now to be $sH(y_i;\boldsymbol{\theta})$. Following Hougaard (1986) and Aalen and Husebye (1991), we can use a Laplace transform, $\mathrm{E}[\exp\{H(y_i;\boldsymbol{\theta})s + \log(s)\}]$, of the gamma distribution. Applied to $sH(y_i;\boldsymbol{\theta})$, this gives

$$f(y_i;\boldsymbol{\theta},\alpha_{i-1},\beta_{i-1}) = \frac{\alpha_{i-1}\beta_{i-1}^{\alpha_{i-1}}}{\{\beta_{i-1}+H(y_i;\boldsymbol{\theta})\}^{\alpha_{i-1}+1}}h(y_i;\boldsymbol{\theta}) \tag{4}$$

where the intensity function $h(y_i;\boldsymbol{\theta}) = dH(y_i;\boldsymbol{\theta})/dy_i$ is the Jacobian of the transformation, $H(\cdot)$, of $y_i$, and the meaning of the subscripts on $\alpha$ and $\beta$ will be made clear in the next section. This procedure works because, in the exponential family, the Laplace transform and the mixture distribution derived from a conjugate mixing distribution both have closed forms and, in the present case, are closely related.

This construction has yielded the form of Pareto distribution whose survival function was given in Equation (3) when $z_i = H(y_i;\boldsymbol{\theta})$. Thus, it can be used to construct models for counts, based on Equation (2), as well as for continuous variables. It can then form one possible basis for the construction of multivariate distributions appropriate for repeated measures. The parameters, $\alpha$ and $\beta$, will be used to model the dependence among the repeated observations.

However, notice that, for arbitrary transformation, $z = H(\cdot)$, Equation (2) is just a 'transformed' negative binomial distribution and Equation (4) a transformed Pareto distribution. When $H(\cdot)$ arises from Equation (1), useful interpretations results but the following development does not depend on this.

## 2 Multivariate dependency

A standard procedure for constructing multivariate distributions suitable for modelling dependence among repeated measurements is to allow some parameter in a univariate distribution to vary stochastically (Lindsey, 1993, Ch. 2). Thus a random effect applied to the mean of a univariate normal distribution gives a multivariate normal distribution with a uniform dependence covariance structure. A procedure closer to that to be used here occurs with autoregression. In a simple autoregression model, the conditional mean parameter, say $\mu_i$, depends stochastically on the previous response value, which however has already been observed and is fixed at the time of the new observation:

$$\mathrm{E}(Y_i) = \mu_t = \rho y_{i-1}$$

where $\rho$ is the autocorrelation. Starting from a univariate normal distribution, this yields a multivariate normal distribution with an autoregressive covariance structure. However, both of these techniques for model construction are much more widely applicable than simply to the normal distribution.

In a similar way to the autoregression construction, we can make $\alpha_i$ and $\beta_i$ in the Pareto distribution of Equation (4) functions of previous observation in time. A first simple possibility, based on the conjugate structure of distributions in the exponential family, is to set

$$\begin{aligned}\alpha_i &= \alpha_{i-1} + n_i \\ \beta_i &= \beta_{i-1} + H(y_i;\boldsymbol{\theta})\end{aligned} \tag{5}$$

where, for discrete observation times, $n_i$ is the number of identical tied events observed at that time point (or, more exactly, occurring since the previous observation point). This will generally be unity except for count data; for a right-censored time interval, it will be zero.

In this way, the distribution of the current response, either $y_i$ or $n_i$, is conditional on the previous values of these two parameters: $\alpha_{i-1}$ and $\beta_{i-1}$ which, in turn, are a function of previous values of the response. Notice that such a substitution into Equation (4) or (2) yields a proper density or probability function, now conditional, because the two parameters only depend on information in the *previous* history of the individual.

Now let the initial conditions $\alpha_0 = \beta_0 = \delta$ be an unknown parameter. Then, depending on whether $y_i$ or $n_i$ is random, Equation (4), with $n_i = 1$ fixed, yields the conditional distribution,

$$\begin{aligned}f(y_i|y_1,\cdots,y_{i-1};\boldsymbol{\theta}) &= \frac{\alpha_{i-1}\beta_{i-1}^{\alpha_{i-1}}}{\{\beta_{i-1}+H(y_i;\boldsymbol{\theta})\}^{\alpha_{i-1}+1}}h(y_i;\boldsymbol{\theta}) \\ &= \frac{\alpha_{i-1}\beta_{i-1}^{\alpha_{i-1}}}{\beta_i^{\alpha_i}}h(y_i;\boldsymbol{\theta})\end{aligned} \tag{6}$$

or, with $y_i$ fixed, Equation (2) gives

$$\begin{aligned}\Pr(n_i|n_1,\cdots,n_{i-1};\boldsymbol{\theta}) &= \frac{\Gamma(\alpha_{i-1}+n_i)}{\Gamma(\alpha_{i-1})n_i!}\frac{H(y_i;\boldsymbol{\theta})^{n_i}\beta_{i-1}^{\alpha_{i-1}}}{\{\beta_{i-1}+H(y_i;\boldsymbol{\theta})\}^{\alpha_{i-1}+n_i}} \\ &= \frac{\Gamma(\alpha_i)}{\Gamma(\alpha_{i-1})n_i!}\frac{H(y_i;\boldsymbol{\theta})^{n_i}\beta_{i-1}^{\alpha_{i-1}}}{\beta_i^{\alpha_i}}\end{aligned} \tag{7}$$

From the identity,

$$f(y_1,\cdots,y_N) = f(y_1)\prod_{i=2}^{N} f(y_i|y_1,\cdots,y_{i-1})$$

for responses ordered in time, the resulting multivariate distribution is

$$\begin{aligned}f(y_1,\cdots,y_N;\boldsymbol{\theta},\delta) &= \frac{\delta^\delta}{\beta_N^{\alpha_N}}\prod_{i=1}^{N}\alpha_{i-1}h(y_i;\boldsymbol{\theta}) \\ &= \frac{\Gamma(\alpha_N)\delta^\delta}{\Gamma(\delta)\beta_N^{\alpha_N}}\prod_{i=1}^{N}h(y_i;\boldsymbol{\theta})\end{aligned} \tag{8}$$

for random $y_i$ and $n_i = 1$ fixed or

$$\Pr(n_1, \cdots, n_N; \boldsymbol{\theta}, \delta) = \frac{\Gamma(\alpha_N)\delta^\delta}{\Gamma(\delta)\beta_N^{\alpha_N}} \prod_{i=1}^{N} \frac{H(y_i;\boldsymbol{\theta})^{n_i}}{n_i!} \tag{9}$$

for random counts.

Note that, although this particular family of models was derived from an ordered sequence of observations, the resulting multivariate distributions are invariant to reordering and hence suitable for modelling clusters. Each individual response has a distribution that is conditional on all other responses in the same cluster through the sum of integrated intensities within that cluster.

Independence occurs when

$$\begin{aligned}\lim_{\delta\to\infty} f(y_1, \cdots, y_N; \boldsymbol{\theta}, \delta) &= \exp\left\{-\sum_{i=1}^{N} n_i H(y_i;\boldsymbol{\theta})\right\} \prod_{i=1}^{N} h(y_i;\boldsymbol{\theta}) \\ &= \prod_{i=1}^{N} f(y_i;\boldsymbol{\theta})\end{aligned}$$

This is equivalent to letting the variance of the underlying gamma distribution (from the Laplace transform) go to zero.

In the context of serial dependence over time, the problem with the functions of time used for updating $\alpha_i$ and $\beta_i$ in Equations (5) above is that, by recursion, each new observation depends on all of the preceding ones to the same extent. One among many other possible ways to update these parameters in a serial fashion is

$$\begin{aligned}\alpha_i &= \rho^{t_i - t_{i-1}}\alpha_{i-1} + (1 - \rho^{t_i - t_{i-1}})\delta + n_i \\ \beta_i &= \rho^{t_i - t_{i-1}}\beta_{i-1} + (1 - \rho^{t_i - t_{i-1}})\delta + H(y_i;\boldsymbol{\theta})\end{aligned}$$

a nonstationary dependence, with $0 < \rho < 1$ a serial dependence parameter. Markov dependence can be obtained by changing the second equation to

$$\beta_i = \delta + \rho^{t_i - t_{i-1}} H(y_{i-1};\boldsymbol{\theta}) + H(y_i;\boldsymbol{\theta})$$

(Notice that this latter equation is no longer recursive.)

When either of these sets of updates is used, the conditional distributions of Equations (6) and (7) remain unchanged, but the multivariate distributions no longer collapse to simple forms. For example, Equation (8) is replaced by

$$f(y_1, \cdots, y_N; \boldsymbol{\theta}, \delta, \rho) = \prod_{i=1}^{N} \frac{\alpha_{i-1}\beta_{i-1}^{\alpha_{i-1}}}{\{\beta_{i-1} + H(y_i;\boldsymbol{\theta})\}^{\alpha_{i-1}+1}} h(y_i;\boldsymbol{\theta}) \tag{10}$$

When $\rho = 1$ in the nonstationary update, we obtain the cluster (frailty) model of Equations (8) and (9). With the Markov update, independence occurs when $\rho = 0$.

## 3 Special cases

Various special forms of this family have been studied and many applications using them have appeared in the literature; only a few of the relevant ones will be noted here. Not surprisingly, most are connected with survival or repeated event data, although the applications to repeated measurements data are much wider.

Suppose first that the observation times, $t_i$, are fixed, and not random, and that the corresponding response values, $y_i$, are continuous positive-valued measurements observed at these unequally-spaced times. Then, Equations (8) and (10) will yield appropriate models for multivariate dependency, providing a wide range of models, depending on the choice of $F(y_i;\boldsymbol{\theta})$ in Equation (1), a welcome complement to the multivariate normal distribution. Scallan (1987) has proposed the multivariate logistic distribution for repeated measures; this is obtained by using an extreme value intensity with Equation (8).

Next, suppose that the observation times are random, or more exactly, correspond to the occurrences of some event of interest. Then, let the response value, $y_i = t_i - t_{i-1}$, be the time passed since the previous event (or possibly since the start of observation, if this has a special significance). In this case, Equations (8) and (10) will yield models for event histories. With $F(y_i;\boldsymbol{\theta})$ an exponential distribution, we obtain the state space model of Smith and Miller (1986); they also discuss the possibility of transformations as used here. When we start with a Weibull distribution, the frailty Equation (8) gives a multivariate Burr distribution, also known as a multivariate log logistic distribution obtained by using the log response in the model mentioned in the previous paragraph. Examples are the repeated failure time model of Crowder (1985) and the frailty model of Aalen and Husebye (1991). Clayton (1988) uses a non-parametric intensity function to develop a proportional hazards frailty model. Yue and Chan (1997) use a piece-wise exponential distribution with Equation (10), but a different update than those suggested above.

Finally, suppose again that the observation times are fixed and that the number of events, $n_i$, since the previous time point is counted. Then, Equation (9) and the analogue of Equation (10) are multivariate models for repeated counts. When an exponential intensity function is used in Equation (9), we obtain a multivariate negative binomial distribution. Thall (1988) uses $h(t_i;\boldsymbol{\theta}) = t_i^{\theta_1}\exp(\theta_2 t_i)$ with this equation, whereas Staniswalis *et al.* (1997) use a non-parametric intensity function.

## 4 Discussion

The family of models proposed here is closely related to the Kalman filtering procedure (see, for example, Smith and Miller, 1986, Harvey and Fernandes, 1989, and Lambert, 1996a, b, c) and more distantly related to copulas (see, for example, Joe, 1997). However, certain of those methods require moment approximations and/or computer-intensive numerical integration. In contrast, because of the special properties of the exponential family, the family discussed here can easily be programmed using the relationship, $H(y) = -\log\{1-F(y)\}$; the functions, $F(y)$ and $f(y)$, for a wide variety of distributions, are available in computer packages.

Only the Laplace transform of the gamma distribution has been considered here, because, as special cases, it leads to several models already presented in the literature. Other new classes of analogous families can be created by substituting the integrated intensity transformation corresponding to the distribution of interest into some density function for a positive random variable other than the Pareto distribution. However, the results will not be so tractable because the special properties of the exponential family cannot be used.

Besides the applications of this family of models reported in the articles referred to above, many other analyses using them have been successfully made. The results will be reported elsewhere. Three functions written by the author for the statistical language R (Ihaka and Gentleman, 1996) to fit the models in this family are available in the public R libraries called `repeated` and `event`, on CRAN (ftp.stat.math.ethz.ch).

*Acknowledgments.* Discussions with Philippe Lambert and Patrick Lindsey greatly aided in clarifying my ideas on this subject.

## References

Aalen OO, Husebye E (1991) Statistical analysis of repeated events forming renewal processes. Statistics in Medicine **10**, 1227–1240

Clayton DG (1988) The analysis of event history data: a review of progress and outstanding problems. Statistics in Medicine **7**, 819–841

Crowder MJ (1985) A distributional model for repeated failure time measurements. Journal of the Royal Statistical Society **B47**, 447–452

Harvey AC, Fernandes C (1989) Time series models for count or qualitative observations. Journal of Business and Economic Statistics **7**, 407–423

Hougaard P (1986) A class of multivariate failure time distributions. Biometrika **73**, 671–678

Ihaka, R. and Gentleman, R. (1996) R: a language for data analysis and graphics. Journal of Computational Graphics and Statistics **5**, 299–314

Joe H (1997) Multivariate Models and Dependence Concepts. London: Chapman and Hall

Lambert P (1996a) Modeling of nonlinear growth curve on series of correlated count data measured at unequally spaced times: a full likelihood based approach. Biometrics **52**, 50–55

Lambert P (1996b) Modelling irregularly sampled profiles of non-negative dog triglyceride responses under different distributional assumptions. Statistics in Medicine **15**, 1695–1708

Lambert P (1996c) Modelling repeated series of count data measured at unequally spaced times. Applied Statistics **45**, 31–38

Lindsey JK (1993) Models for Repeated Measurements. Oxford: Oxford University Press

Lindsey JK (1996) Parametric Statistical Inference. Oxford: Oxford University Press

Scallan AJ (1987) A GLIM model for repeated measurements. GLIM Newsletter **15**, 10–22

Smith RL, Miller JE (1986) A non-Gaussian state space model and application to prediction of records. Journal of the Royal Statistical Society **B48**, 79–88

Staniswalis JG, Thall PF, Salch J (1997) Semiparametric regression analysis for recurrent event interval counts. Biometrics **53**, 1334–1353

Thall PF (1988) Mixed Poisson likelihood regression models for longitudinal interval count data. Biometrics **44**, 197–209

Yue H, Chan KS (1997) A dynamic frailty model for multivariate survival data. Biometrics **53**, 785–793