Modelling the position and shape of regression curves when the distribution is skewed, possibly with censoring and autocorrelation

J.K. Lindsey* and T. Ring†

*Faculty of Economics, University of Liège, Belgium

†Department of Nephrology, Aalborg Hospital, Denmark

Email: jlindsey@luc.ac.be

www.luc.ac.be/∼jlindsey

Abstract

Regression models based on the normal distribution are easy to interpret because the distribution is symmetric and, with constant variance, always has the same shape about the regression function. The same is not true of skewed distributions. In simple cases, such as the gamma and inverse Gauss distributions within the family of generalised linear models, an easily interpretable parameter, the mean, is still available for regression modelling. This is not more widely true, even with as common a distribution as the Weibull.

We describe the problems that have arisen in the analysis of a two-treatment cross-over trial involving use of vitamin B12 for hemodialysis patients. The natural parameter for regression modelling in the best fitting generalised Weibull distribution does not appropriately describe the changing position of this distribution. The problem is further aggravated when autocorrelation is included.

KEYWORDS: Autocorrelation, censoring, generalised Weibull distribution, location parameter regression.

1 Introduction

In classical normal-theory linear regression modelling, a plot of the fitted regression line fairly represents the position of the model because the normal distribution is symmetric, with constant variance, and this line describes changes in the conditional mean. However, such a plot, alone, does not describe the stochastic variability about that line, as given by the variance of that distribution.

The location–scale family of distributions has the location (μ) and scale (σ) parameters related to the response by $(Y-\mu)/\sigma$. Such distributions are often symmetric, such as the normal, Cauchy, logistic, Laplace, and Student t distributions, so that μ is the mode, as well as the mean (when it

exists) and the median. In this sense, with constant scale parameter, they are ideal for constructing regression models to describe changes in the position of a distribution with changing covariates.

In generalised linear models, the regression line describes changes in the mean of the chosen distribution (for example, gamma or inverse Gauss). However, this may no longer be a suitable representation of the model if the distribution is very skewed. For example, the mode may be more appropriate. In addition, the variance is no longer sufficient to describe the variability about the regression line because of the skewness. Indeed, the variance is not even constant in these models. In other words, the complete shape of the distribution needs to be taken into consideration.

In the above families, the mean or some other location parameter determines the position of the distribution. However, in other distributions, when such a parameter is not available, another parameter may play a similar role with respect to the position. Thus, outside the generalised linear and location–scale families, no parameter may directly represent the mean so that the situation becomes even more complex.

Consider, for example, the Weibull distribution,

$$f(y;\nu,\phi) = \frac{\phi y^{\phi-1} \mathrm{e}^{-\left(\frac{y}{\nu}\right)^{\phi}}}{\nu^{\phi}}$$

In regression models involving this distribution, $\log(\nu)$ is usually allowed to depend on various covariates. Note that $\nu = \sigma$ is a scale, not a location, parameter. However, it does have an interpretation in terms of the position of the distribution: the scale parameter is proportional both to the mean

$$\mu = \Gamma\left(1 + \frac{1}{\phi}\right)\nu$$

and to the mode

$$\text{mode}(Y) = \left(\frac{\phi - 1}{\phi}\right)^{\frac{1}{\phi}} \nu$$

In some other distributions, such as the gamma, the mean is a scale parameter. In both of these distributions, the scale parameter indicates position; in contrast, for the symmetric members of the location–scale family mentioned above, the scale parameter is independent of position.

In many modelling situations, it is also necessary to allow the shape parameter to depend on some covariates. Thus, for the Weibull distribution, $\log(\phi)$ would depend on covariates, in which case ν is no longer proportional to the mean and the mode. It then becomes essential to know how the shape of the distribution is changing with the covariates. One example of such changing shape was given by Lambert and Lindsey (1999) for the four-parameter family of stable distributions.

Here, the problem of describing the changing position and shape of a skewed distribution over time has arisen in the analysis of a two-treatment cross-over trial involving seven repeated measurements within each of three periods. The study involved use of vitamin B12 for hemodialysis patients with treatments being either the conventional (HD) procedure or hemo-diafiltration (HDF). One mg of B12 was injected intramuscularly at the beginning of each period and B12 was

subsequently measured at weeks 1, 2, 3, 6, 8, 10, and 12. The two sequences were HDF/HD/HDF and HD/HDF/HD. A washout period could not be used because the patients require dialysis at least several times a week; there is no neutral treatment; either HD or HDF has to be used. There were 26 patients involved, aged between 23 and 74 years, six of whom were female.

After injection, B12 was expected to decline in a nonlinear fashion. In an earlier study, Moelby *et al.* (2000) had found that, on conventional HD, vitamin B12 declined following such an injection whereas methyl malonic acid (MMA) increased. This reciprocal movement of B12 and MMA indicated that the perturbations of B12 concentration could be of biological significance. The new study, presented here, was designed to address this issue. The HDF treatment mentioned above involved the use of a filter with higher porosity to see if indeed B12 declined more and MMA also increased more. Here, we look only at the changes in B12.

This study, thus, is somewhat similar to that of propoxyphene analysed by Lindsey (1999, pp. 157–162 and 2001, pp. 131–135). However, an additional complication in the analysis of these data is the fact that the B12 concentration was censored at an upper limit of 1500 pmol/l. This occurred at the beginning of each period just after injection. As well, there is strong indication of autocorrelation among the responses of each individual within each period.

Analysis showed that the generalised Weibull distribution

$$f(y;\nu,\phi,\theta) = \frac{\theta\phi y^{\phi-1}[1-\mathrm{e}^{-\left(\frac{y}{\nu}\right)^{\phi}}]^{\theta-1}\mathrm{e}^{-\left(\frac{y}{\nu}\right)^{\phi}}}{\nu^{\phi}} \tag{1}$$

best described the distributional shape when ν depended on covariates. Note that, when $\theta = 1$, this is the usual Weibull distribution. Unfortunately, in general, when $\theta \neq 1$, this distribution does not have a simple relationship between the mean/mode and ν.

Here, we only consider a model for concentration of vitamin B12 depending on time and patient weight at the end of each dialysis (dry weight):

$$\nu_t = \beta_0 \mathrm{e}^{\beta_1 t' + \beta_2 x_{2t}} \tag{2}$$

where t is total time from the beginning of the trial, t' time from the beginning of a period (that is, from injection of B12), and x_{2t} is the dry weight at time t. This latter covariate is time-varying; it can change both within and between periods. Treatment (x_{1t}) is also a time-varying covariate between periods, but was found not to be necessary in this model. As well, the ϕ shape parameter in Equation (1) was found to depend on dry weight but not directly on time:

$$\phi_t = \beta_0' \mathrm{e}^{\beta_2' x_{2t}} \tag{3}$$

For simplicity and clarity of the analysis here, we shall ignore the effects specific to the cross-over design, such as period and carry-over. A clinical interpretation of these data will be published elsewhere.

2 Censoring

Let us first ignore the dependencies in the repeated measurements and concentrate on the shape of the distribution about the regression curve. We fit the above model based on the generalised Weibull distribution of Equation (1) with two of the parameters depending on covariates as described by Equations (2) and (3). If we use the density for the uncensored response values and the survivor function for the censored values, we obtain a negative log likelihood of 2400.9 whereas, if we set the latter values to 1500 and use the density for all responses, we obtain 3287.7. This clearly indicates that censoring is important. In the former model, $\hat{\theta} = 7.94$, very far from a standard Weibull distribution. Thus, we may suspect that ν_t may not follow closely the mode of the distribution.

The corresponding curves for the scale parameter ν_t, with the observed values, for one typical subject are plotted in Figure 1. Clearly, accounting for censoring influences the whole curve and not just the estimates at the censored values. Although the likelihood indicates that the model allowing for censoring fits better, the plot seems to show that the fitted regression curve is closer to the observations when censoring is ignored. However, this is a plot of the scale parameter which, with such a large value of $\hat{\theta}$, may not a be a good measure of position.

Let us look more closely at the first (week 1) and last (week 12) values in the first period for this individual. The estimated values of B12 concentration given by $\widehat{\nu_t}$ are, at week 1, 839.3 pmol/l allowing for censoring and 976.6 pmol/l ignoring it; they are, respectively, 230.0 and 386.1 pmol/l at week 12. To see how well these parameter values represent the position of the distribution, consider now the shapes of the distribution at these time points, as shown in Figure 2. Although ν_t is estimated to be quite different by the two models at week 12, the distributions are rather close, centred near 500 pmol/l where the observations lie. On the other hand, the shapes of the distributions at week 1 are quite different, as might be expected because the observation at this time point is censored. Although $\hat{\nu}_1 = 839.3$ pmol/l for the censoring model, the mode lies over 1500 pmol/l.

We can now use this approach to plot the regression curve using the mode instead of ν_t to indicate the position. The improvement in interpretability is clear in Figure 3. The superior fit of the model taking into account censoring is now evident, especially for the censored values. We may note that, for the complete set of observations on all subjects, the values of ν_t are between 156.27 and 912.93 pmol/l less than the corresponding modes, clearly not proportional. This is primarily due to the large value of $\hat{\theta}$, although ϕ_t is also varying over time with dry weight.

3 Autocorrelation

One important way of allowing for dependence in such longitudinal data is autocorrelation. If an observed response y_{it} for an individual i is some distance from the fitted regression function for some appropriate parameter ν_t describing the position of the distribution at a given time point t,

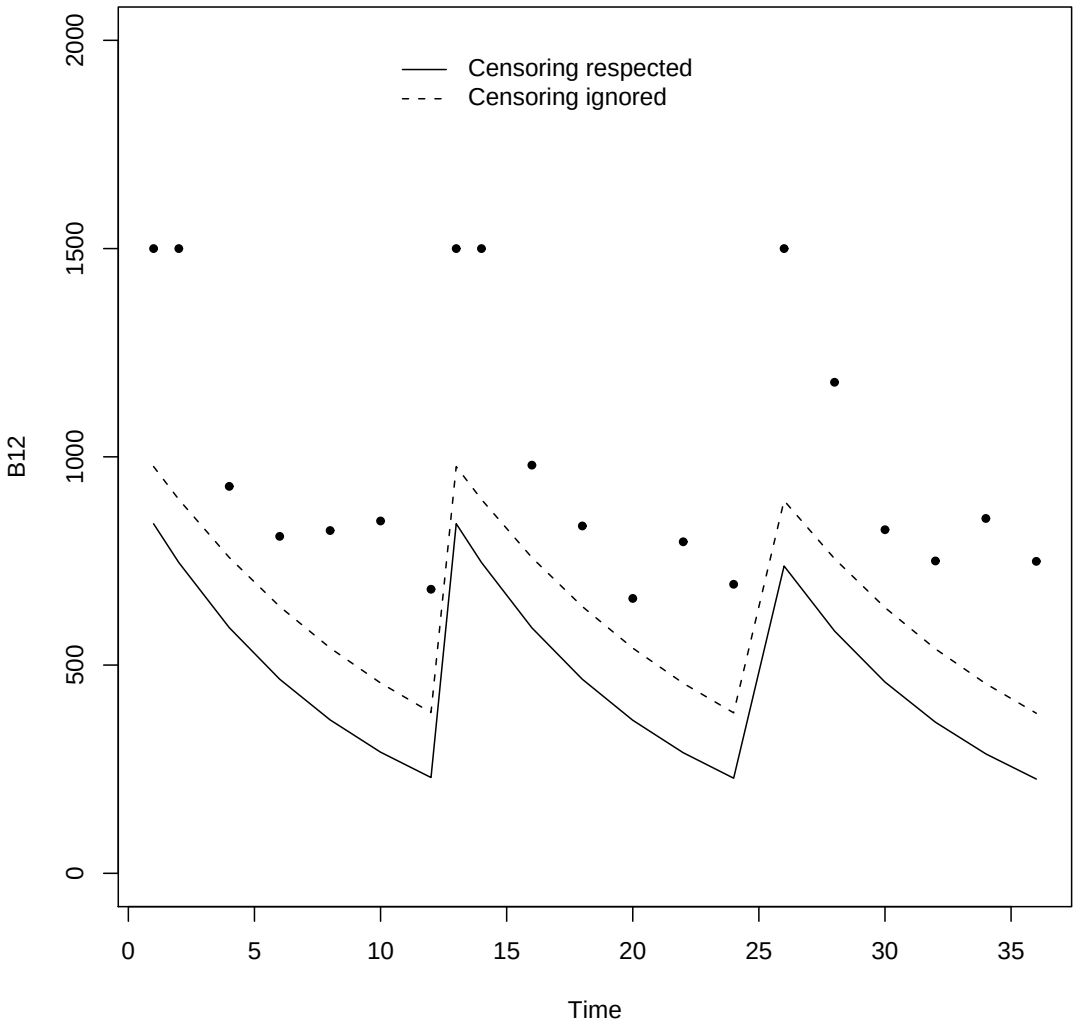

Figure 1: Fitted regression functions for ν in the generalised Weibull distribution, ignoring censoring and taking it into account, for one subject.

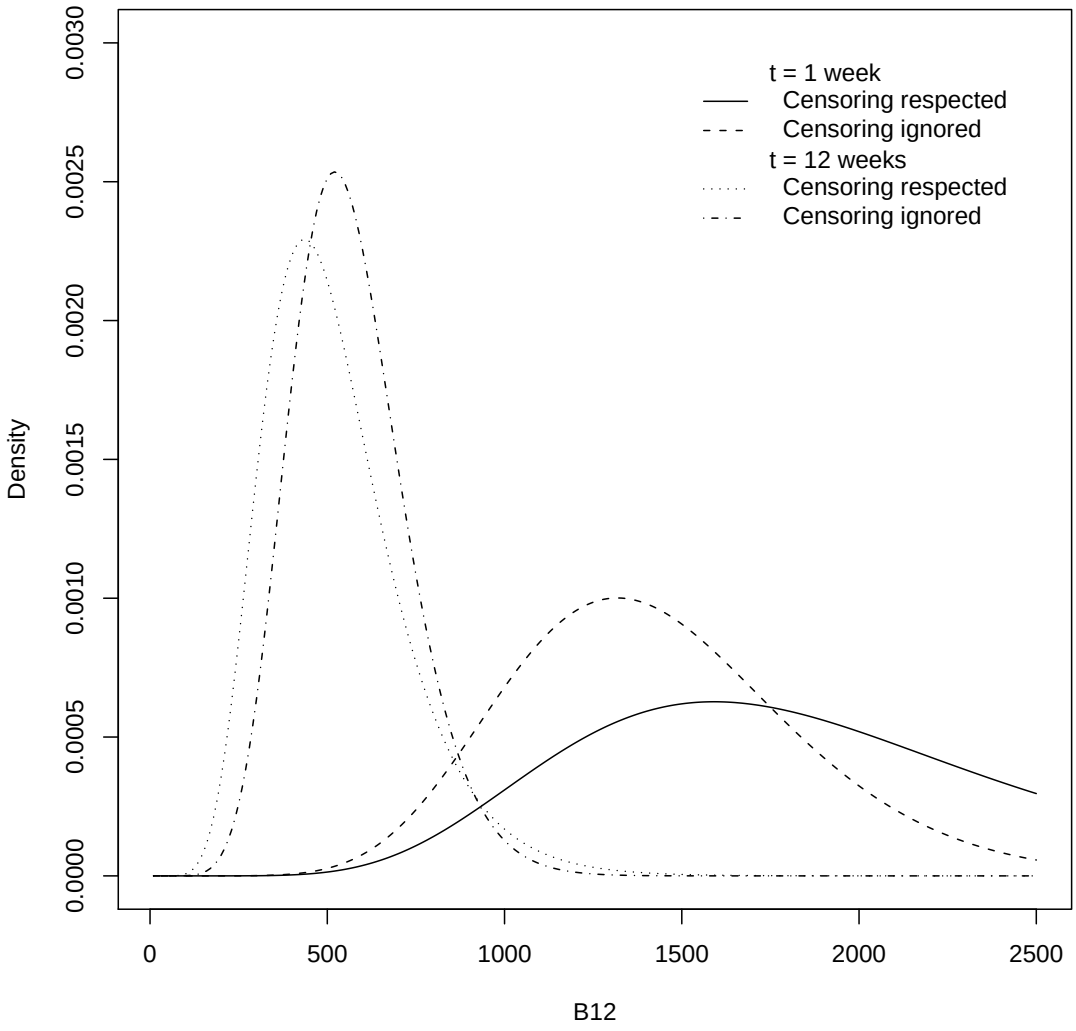

Figure 2: Fitted generalised Weibull distributions at times 1 and 12 weeks for the curves of Figure 1.

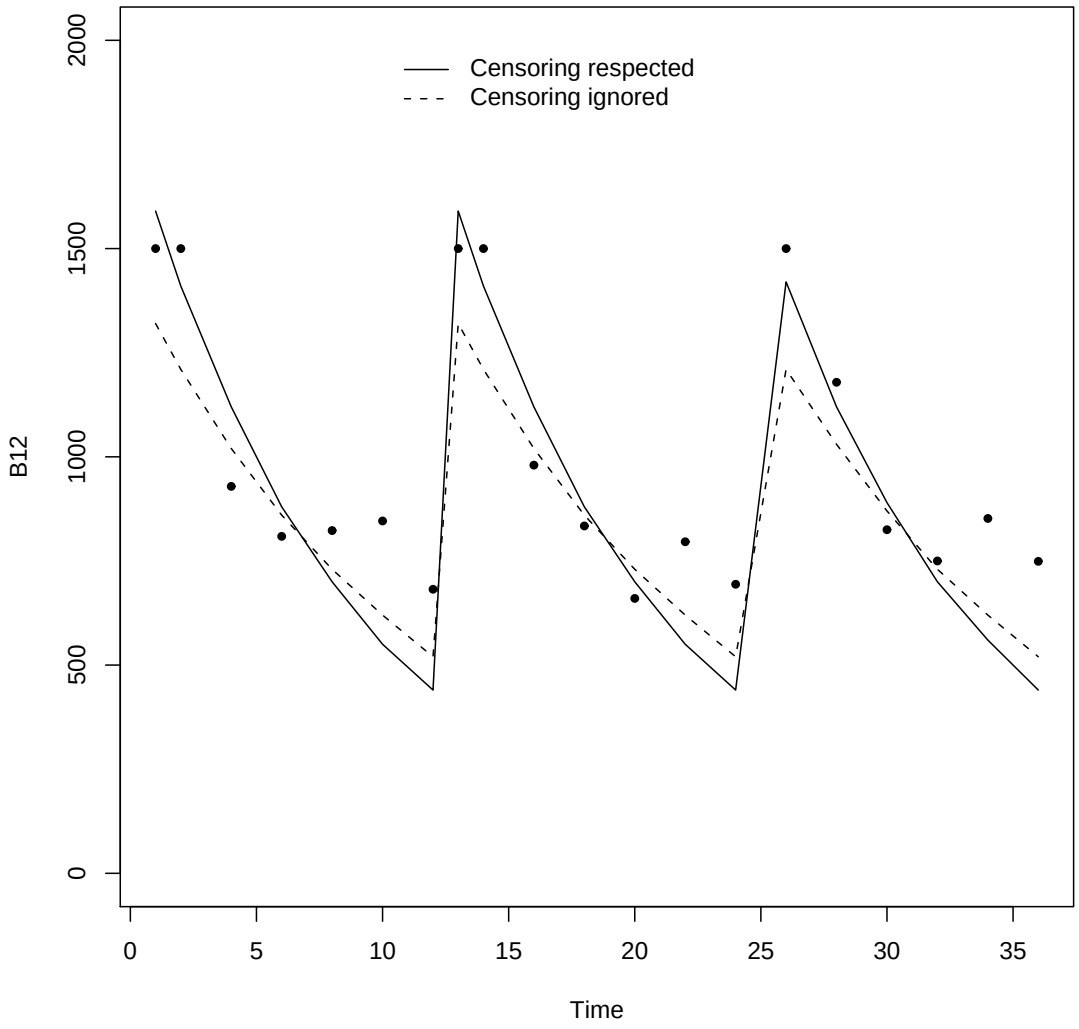

Figure 3: Fitted regression functions for the mode in the generalised Weibull distribution, ignoring censoring and taking it into account, for one subject.

it is often useful to predict that the response at the following observation point for that individual ($\nu_{i,t+1}$) may also be away from the regression function in the same direction. This can be written as

$$\nu_{i,t+1} = \nu_{t+1} + \rho(y_{it} - \nu_t)$$

where ν_t gives the underlying prediction curve if there were no dependence over time and $\nu_{i,t+1}$ is the predicted value of the parameter for the subject i if that individual was at a distance $y_{it} - \nu_t$ from the underlying prediction curve at the previous time point. Here, $y_{it} - \nu_t$ is supposed to be a measure of the previous residual or innovation. (For the censored values in our data, the censoring point will be used.)

However, the question for skewed distributions is what to use for the prediction ν_t. This is further complicated if there is no natural location or other position parameter. As we have seen in the previous section, for our data, the models based on the generalised Weibull distribution have an estimated ν_t, a scale parameter, systematically underestimating the position of the distribution as compared to the mode. In other words, $y_{it} - \nu_t$ will generally be positive even when the responses are following the regression function. However, these differences should nevertheless vary in an appropriate way as y_{it} moves away from the main mass of the density of the distribution.

Indeed, introducing the autocorrelation parameter, based on the generalised Weibull scale parameter (ν_t), greatly improves the above model for these data, reducing the negative log likelihood from 2400.9 above to 2152.7 here. Now, $\hat{\theta} = 2.16$, closer to the standard Weibull distribution but still significantly different from it. The regression functions for this model, based on ν, are plotted in Figure 4 for the same subject as previously. As in Figure 1, the ν scale parameter of the generalised Weibull distribution does not go through the data. But the problem is much more serious for ν_t than for ν_{it}, for the reason just explained. As expected, ν_{it} lies consistently above ν_t.

Here, the situation is not as simple as for the independence case. The parameter of the generalised Weibull distribution is now ν_{it}, not ν_t. As for the independence case, the mode can be plotted. This is shown in Figure 5 along with the 10% probability density contours. Here, because $\hat{\theta}$ is much smaller than in the models of the previous section, ν_{it} follows the mode very well. In contrast to the independence case given above, here for all observations, the values of ν_{it} are only between 27.25 and 99.8 less than the corresponding modes. Most of the variation is now due to the dependence of ϕ_t on dry weight.

On the other hand, this procedure cannot be used to correct the estimated values of ν_t; they are depressed because of the systematically positive values of $y_{it} - \nu_t$. One solution would be to replace ν_t by the mode in the calculation of the residuals for prediction. However, this would be very computationally expensive and hence would make estimation of the model parameters extremely slow. Another possibility would be to introduce a true location parameter into the generalised Weibull distribution, at the expense of added complexity. These possibilities will be the subject of further research.

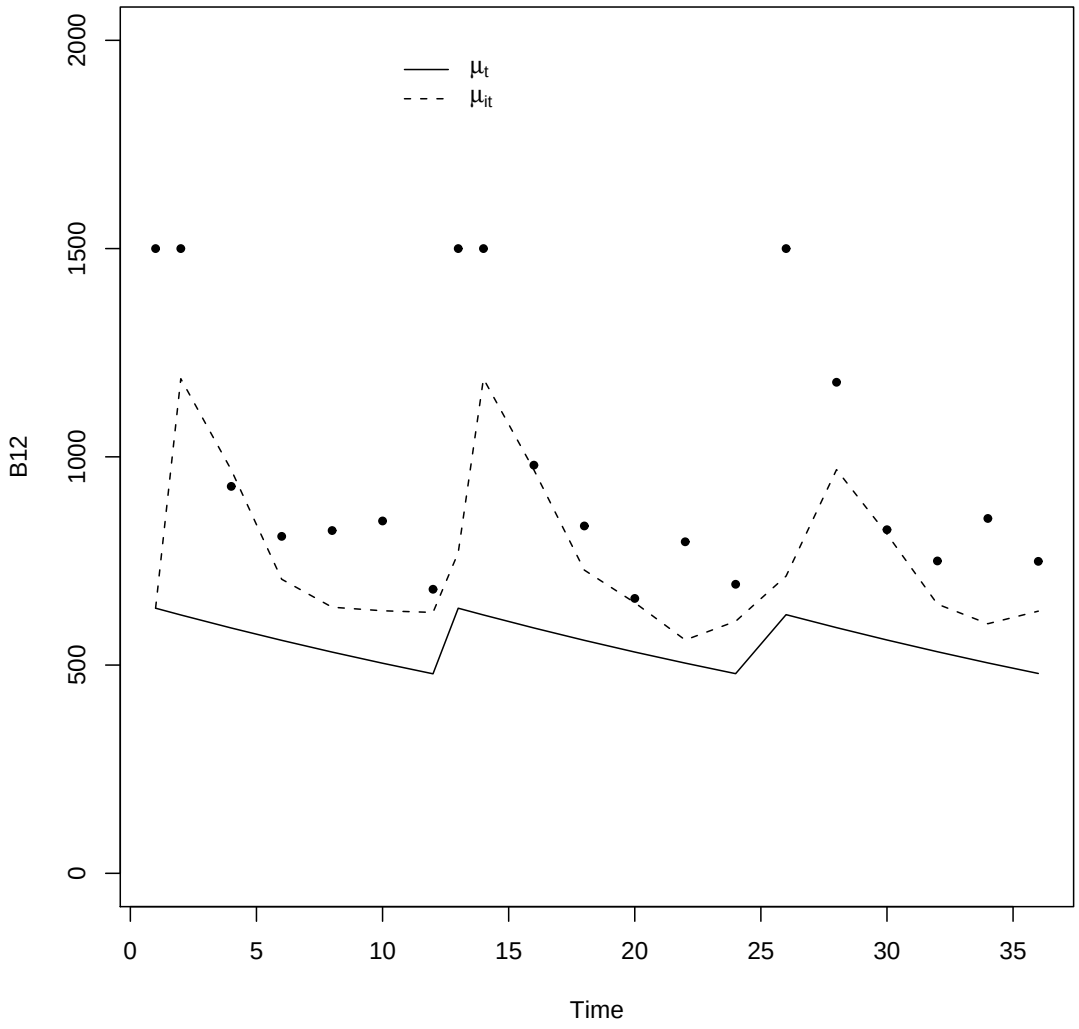

Figure 4: Fitted regression functions for ν_t and ν_{it} in the generalised Weibull distribution, taking censoring into account, for one subject.

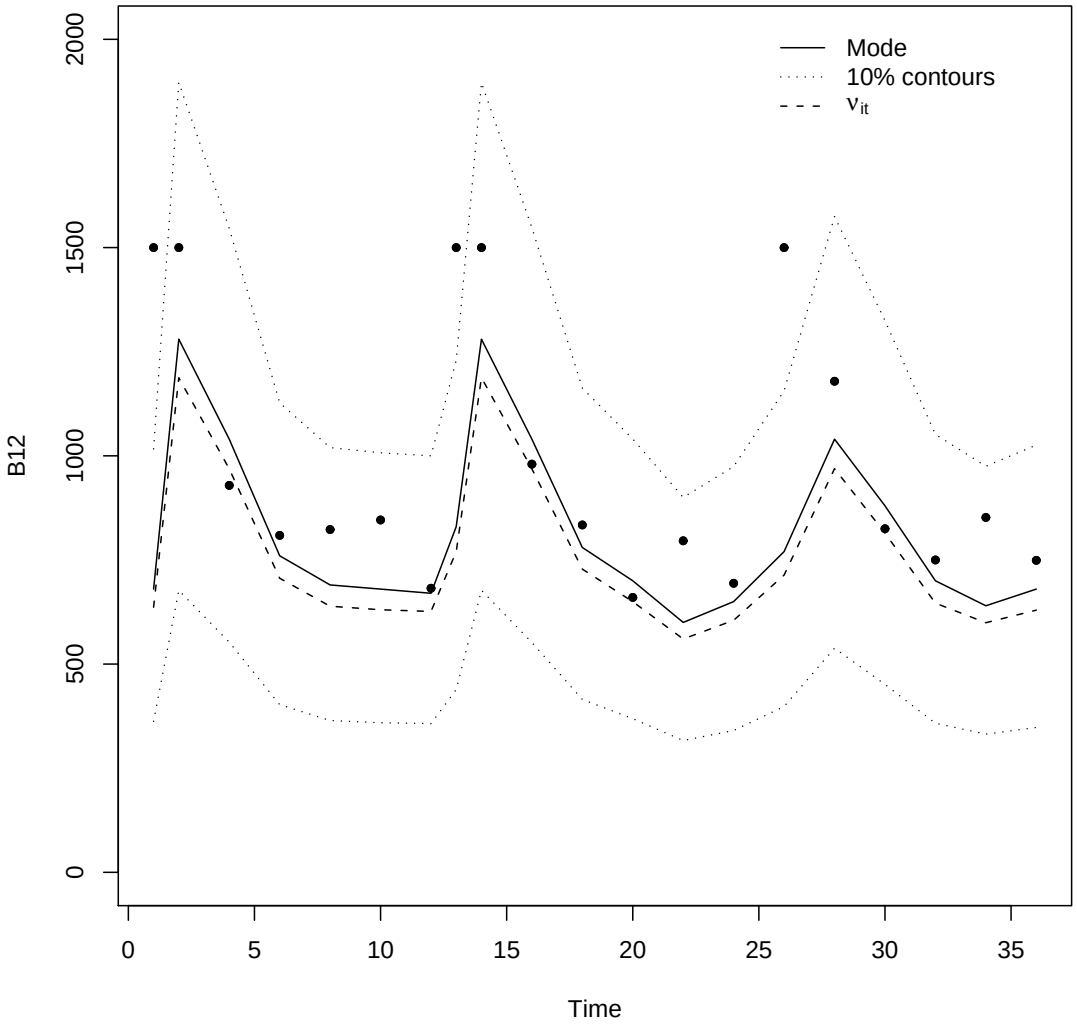

Figure 5: Mode of the fitted generalised Weibull model, taking censoring into account, for one subject, with the contours having 10% of the probability density at the mode, along with the fitted regression function for ν_{it} from Figure 4.

4 Discussion

Regression models based on the normal distribution are easy to interpret because the distribution is symmetric so that the regression function accurately represents its position. In addition, with constant variance, the distribution does not change shape. The same is not true of skewed distributions, as we have seen with the rather extreme case presented here.

Parameters that appear natural to model, such as ν in the generalised Weibull distribution of Equation (1), may sometimes have restricted interpretability in terms of the position of the distribution. The dependence of such a parameter on covariates does not have the same meaning as does the dependence of the mean of a generalised linear model. As we have seen, the difficulty can be compounded when dependence over time is also taken into account.

Often, as in the above study, more than one parameter of a distribution (for example, ϕ and/or θ in the generalised Weibull distribution) will require regression functions depending on covariates. Then, it is essential to study the changing *shape* of the distribution, not simply the regression *curve(s)*.

Acknowledgements The authors thank Patrick Lindsey for his valuable comments on an earlier version.

All of the examples were analysed using the R software with the functions `gnlr3` and `gar` in the first author's public libraries, respectively called `gnlm` and `repeated`, available at
`http://www.luc.ac.be/~jlindsey/rcode.html`

References

[1] Lambert, P. and Lindsey, J.K. (1999) Analysing financial returns using regression models based on non-symmetric stable distributions. *Journal of the Royal Statistical Society* **C48**, 409–424.

[2] Lindsey, J.K. (1999a, 2nd edn.) *Models for Repeated Measurements.* Oxford: Oxford University Press.

[3] Lindsey, J.K. (2001) *Nonlinear Models in Medical Statistics.* Oxford: Oxford University Press.

[4] Moelby, L., Rasmussen, K., Ring, T., and Nielsen, G. (2000) The relationship between methylmalonic acid and cobalamin in uremia. *Kidney Int.* **57**, 265–273.