Computational Statistics & Data Analysis 13 (1992) 373–384
North-Holland

Fitting and comparing probability distributions with log linear models

J.K. Lindsey and G. Mersch
Université de Liège, 4000 Liège, Belgium

Abstract: Probability distributions in the exponential family can be fitted directly as log linear models and the usual maximum likelihood estimates of parameters obtained. If a composite distribution is constructed from several competing candidates, these may also be compared within a log linear model, and those not acceptable eliminated. The approach also applies to truncated distributions, as well as when independent variables are present (generalized linear models). Here, in the most general case, the different sub-populations may even have different distributions within the same model. With additional iterations, estimation may be extended to certain models outside the generalized linear framework.

Keywords: Comparison of models, Generalized linear models, Goodness of fit, Likelihood inference, Log linear models, Probability distributions, Truncated distributions.

1. Introduction

Several approaches have been suggested for the comparison of probability models. In the context of classical hypothesis testing, Cox [4], [5] and Atkinson [2] introduce methods of imbedding several alternative models in a combined probability distribution. They concentrate especially on exponential combinations of the form

$$c\prod_{i=1}^{N}\left\{f_1(y_j,\theta_1)\right\}^{\lambda}\left\{f_2(y_j,\theta_2)\right\}^{1-\lambda} \tag{1.1}$$

for variable vector, $\boldsymbol{y}$, parameter vectors, $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$, and normalizing constant, c, with N observations and j indexing individual observations. A value of λ near *1* indicates that the suitable distribution is proportional to the probability density, $f_1(.)$, while one near *0* indicates that it is $f_2(.)$.

A second approach, using likelihood inference but not equation (1.1), was proposed by Lindsey [11][12], who imbeds probability models within a more general multinomial distribution. Here, with n_k observations for value y_k of the variable, where k indexes distinct values of y, so that $\Sigma n_k = N$, the likelihood function is

$$L(\boldsymbol{p}) \propto \prod p_k^{n_k}, \quad \text{where} \tag{1.2}$$

$$p_k = F(y_k;\, \boldsymbol{\theta}), \quad y \text{ discrete} \tag{1.3a}$$

$$= \int_{y_k - \Delta y_k/2}^{y_k + \Delta y_k/2} f(y_k;\, \boldsymbol{\theta})\, \mathrm{d}y, \quad y \text{ continuous} \tag{1.3b}$$

$$\cong f(y_k;\, \boldsymbol{\theta})\, \Delta y_k,$$

with $F(.)$ a probability function, $f(.)$ a probability density function, and Δy_k the unit of measurement. The general multinomial model places no constraints on the probabilities, p_k, and hence fits the data exactly. Then, for each model of interest, a different vector, $\boldsymbol{p}$, is estimated from equations (1.2) and (1.3), and that giving a larger value of the likelihood function (1 2) and hence a smaller deviance, i.e., that closest to the unconstrained multinomial model, is considered more plausible.

This second approach will be pursued in the present paper, where we shall show how many models, specifically those which are members of the exponential family, can be fitted and compared in this way as log linear models.

2. Likelihood functions for grouped data

For discrete models, the form of the likelihood function is not a problem for fitting distributions. For continuous data, some kind of grouping must be performed, although note that any empirically observed continuous data are already grouped. For the present, we restrict attention to the exponential family, which has a density of the form

$$f(y_k;\, \boldsymbol{\theta}) = \exp\left\{ \sum_{j=1}^{p} t_j(y_k)\theta_j + c(\boldsymbol{\theta}) + d_1(y_k) \right\}, \tag{2.1}$$

where c(.) is the normalizing constant and $t_j(.)$ are the sufficient statistics for the parameters. For any empirically observed data, a probability function must be used

$$F(y_k;\, \boldsymbol{\theta}) = \exp\left\{ \sum_{j=1}^{p} t_j(y_k)\theta_j + c(\boldsymbol{\theta}) + d_2(y_k) \right\}, \tag{2.2}$$

where $d_2(.)$ is a function of y_k depending on both $d_1(.)$ and the width of the grouping intervals associated with each observed y_k. This is just the approximation to an integral shown in equation (1.3b). We may now consider discrete and continuous data on the same footing.

In most applications, the latter part of the probability in (2.2), the function of the grouping width, is a constant independent of the parameters to be estimated, so that equations (2.1) and (2.2) yield the same likelihood function. If we maximize this likelihood with respect to $\boldsymbol{\theta}$ in the usual way, we obtain the usual maximum likelihood estimates.

However, the multinomial likelihood of equation (1.2) can also be maximized as a log linear model for categorical data, for example with GLIM, if we ignore that c(.) is a function of $\boldsymbol{\theta}$ and treat it is a global constant parameter, the intercept. Thus, we propose to maximize the equivalent Poisson likelihood

$$\prod \lambda_k^{n_k} \mathrm{e}^{-\lambda_k}/n_k!,$$

where

$$\lambda_k = \exp\left\{\theta_0 + \sum_{j=1}^{p} t_j(y_k)\theta_j + d_2(y_k)\right\}.$$

θ_0 replaces $c(\boldsymbol{\theta})$, and the total number of observations is fixed. For a member of the exponential family, the sufficient statistics for the parameters are fitted as explanatory variables in a Poisson regression. Now, the grouping intervals are no longer independent of the parameters, since we are, in fact, estimating the probabilities of observations falling in the different grouping intervals and the probability parameters are a function of these intervals. The full likelihood for our model is the multinomial likelihood given by equation (1.2) with $p_k = F(y_k; \boldsymbol{\theta})$.

In log linear models for categorical data, conditioning on the total number of observations in the Poisson likelihood ensures that the total multinomial probability of all categories included in the model equals one. Here, as defined so far, this is not what we require, since, in a probability model to be fitted, the sum of probabilities of all possible values of y_k must equal one. The normalizing constant, $c(\boldsymbol{\theta})$, will only normalize the probabilities contained in the model. This may simply be accomodated by including zero frequencies for all unobserved values of y_k, since these are sampling zeroes, which might be nonzero in another sample, and not structural zeroes (see Lindsey [13] p. 78). In practice, we can only include a finite subset, but parameter estimates with any desired degree of accuracy can be obtained by only excluding intervals with small enough probabilities. In this way, θ_0 provides an estimate of $c(\boldsymbol{\theta})$, hence a second estimate of (a function of) $\boldsymbol{\theta}$.

If we purposely exclude certain grouping intervals which have relatively high probabilities under our model, but have zero observed frequencies, we are fitting a truncated distribution. In terms of categorical log linear models, we are saying that these are structural zeroes which are impossible to observe.

Three points may be noted. The function $d_2(y_k)$ does not involve $\boldsymbol{\theta}$ and, thus, forms a known constant term in the log linear regression model, something which is known as an offset in GLIM terminology. Secondly, as Lindsey [11] shows, Δy_k in equation (1.3b), which is included in the offset, may be allowed to vary within a certain objective range without greatly modifying the results. Thus, it may not necessarily be the unit of measurement, but may be chosen to reduce the number of small frequencies, if a measure of goodness of fit is required. Thirdly, the deviance does not remain unchanged as zeroes are added. In fact, it increases as the parameter estimates become more accurate, since the best estimates are for a truncated model with just the observed non-zero frequencies included. On the other hand, if the data have been regrouped, so that Δy_k is larger than the unit of measurement, $\boldsymbol{\theta}$ will not be accurately estimated, as is always the case with grouped data.

This, then, is a procedure which gives the usual maximum likelihood estimates of all parameters, with the same standard errors when they are in canonical form. The statistical properties of these models are well known; see, for example, Haberman [7] and Silvapulle and Burridge [15]. If the parameter estimates exist, they are unique. For our present case, the exponential family, these estimates do exist.

The deviance of this Poisson regression model from the saturated model provides a measure of goodness of fit of (1.3); see Lindsey [11]. This involves a comparison of the estimated probabilities for the model (1.3) to the observed relative frequencies. As always with inferences for goodness of fit, the problem of small and zero frequencies must be taken into account. Obviously, if the fitting procedure involves a lot of zeroes, deviance values should be interpreted in terms of relative plausibility and the corresponding test statistics as a guide to selecting a model. Conditional tests may be more appropriate in such cases.

Two simple examples may be used to illustrate this approach. Take first a discrete case, the Poisson distribution. With $\theta_1 = \log(\mu)$, equation (1.3a) becomes

$$p_k = \mathrm{e}^{-\mu + y_k \log(\mu) - \log(y_k!)} \tag{2.3}$$

so that we fit a log linear model with an intercept and a term in y_k with offset $-\log(y_k!)$. If we are not fitting a truncated model, the intercept will yield the negative of the mean plus $\log(N)$ and the parameter for y_k the log of the mean, and these two estimates of the mean will be identical to the degree of accuracy determined by the zero frequencies included.

The equivalent continuous case is the exponential distribution. Here, equation (1.3b) becomes

$$p_k \cong \mathrm{e}^{-\log(\mu) - y_k/\mu} \Delta y_k, \tag{2.4}$$

so that again we fit a log linear model with an intercept and a term in y_k but with offset $\log(\Delta y_k)$. Now $\theta_1 = -1/\mu$, so that the intercept yields the negative of the log mean plus $\log(N)$ and the parameter for y_k the negative reciprocal of the mean, and again these two estimates of the mean will be identical, if we have an untruncated model.

3. Comparing probability distributions

Suppose now that we do not know the correct form of the probability model (1.3), but have several possible candidates. As long as the competitors are all members of the exponential family, we may construct a composite log linear model which encompasses all of these candidates. Since each simple model is a log linear regression on some functions of the variable, the sufficient statistics, we can combine these functions as terms in a composite model. We are, thus, imbedding the several distributions of interest within a more general distribution which takes the form of a log linear model. We have two levels of imbedding: several specific probability distributions within a more general composite one, and the latter within the unconstrained multinomial distribution. In this way, we avoid the restrictions of the alternatives implied by using model (1.1) to compare distributions.

Suppose we wish to compare a gamma distribution

$$p_k \cong \mathrm{e}^{\alpha\log(\mu) + (\alpha - 1)\log(y_k) - \mu y_k - \log(\Gamma(\alpha))}\Delta y_k, \tag{3.1}$$

with the log normal distribution

$$p_k \cong \mathrm{e}^{\mathrm{const}-\log(\sigma^2)/2 - \mu^2/(2\sigma^2) + \mu\log(y_k)/\sigma^2 - \log^2(y_k)/(2\sigma^2)}\Delta y_k. \tag{3.2}$$

Here, $\theta_1 = -\mu$ and $\theta_2 = \alpha - 1$ for equation (3.1) and $\theta_1 = \mu/\sigma^2$, $\theta_2 = -1/(2\sigma^2)$, and const $= -\log(y_k) - \log(2\pi)/2$ in equation (3.2).

For the latter distribution by itself, we would fit our log linear model with an intercept, $\log(y_k)$, and $\log^2(y_k)$, and constant term (offset)

$$-\log(y_k) - \log(2\pi)/2 + \log(\Delta y_k). \tag{3.3}$$

From this model, we obtain estimates of both the mean and the variance using the parameter estimates for the two terms involving y_k.

To compare the two distributions, we must also include y_k, thus fitting the intercept, y_k, $\log(y_k)$, and $\log^2(y_k)$, with a combined offset, which here contains the same $-\log(y_k)$ from both distributions. We, then, test which terms in the log linear model are significant using standard techniques for log linear models. If the term for y_k can be eliminated, we have a log normal distribution, and if the term for $\log^2(y_k)$, we have a gamma distribution. However, it is always possible that the parameter estimates for the two distributions are such that it is impossible to choose between the models. In such cases, this method will have the same problems as any other.

With this procedure, we may discover acceptable probability distributions which are not among those usually considered. For example, the above composite distribution, in fact, incorporates four common distributions: the exponential (y_k), the Pareto ($\log(y_k)$), the gamma (y_k, $\log(y_k)$), and the log normal ($\log(y_k)$, $\log^2(y_k)$). However, only $\log(y_k)$ might fall out, leaving y_k and $\log^2(y_k)$ which corresponds empirically to the definition of a probability distribution, although it does not correspond analytically to any known distribution.

Note that, in such comparisons using composite distributions, we may be limited in the number of simple distributions which may be combined at one time, especially if each has several parameters. As usual, the total number of estimable parameters will depend on the number of categories of the variable, y, with non-zero frequencies, i.e., on the degrees of freedom available.

Table 1
Simulated log normal data

y	0.5	1.5	2.5	3.5	4.5	5.5	6.5	7.5	8.5	9.5
Frequency	1	14	8	12	15	14	6	10	5	11
y	10.5	11.5	12.5	13.5	14.5	15.5	16.5	17.5	18.5	19.5
Frequency	6	9	6	7	7	1	8	5	5	5
y	20.5	21.5	22.5	23.5	24.5	25.5	26.5	27.5	28.5	30.5
Frequency	2	1	1	1	2	4	4	3	4	2
y	31.5	32.5	33.5	34.5	36.5	37.5	38.5	43.5	44.5	50.5
Frequency	1	1	1	2	1	1	1	1	1	1
y	51.5	53.5	65.6	74.5	95.5	97.5	104.5	165.5		
Frequency	2	1	2	1	1	1	1	1		

On the other hand, although not noted in the previous section, where it would have had limited relevance, models may be fitted by this method even when all values of the frequency vector, ***n*** are very small, even one. Regrouping is not required for parameter estimation and comparing models, but ***n*** must be appropriately filled with zeroes. However, in such a case of sparse observations, measures of goodness of fit have little or no meaning.

Example 1. Table 1 provides a set of 200 simulated log normal values. The procedure described above was used to compare the gamma and log normal distributions, with the vector ***n*** extended with zeroes to $y_k = 200$. The intercept, y_k, $\log(y_k)$, and $\log^2(y_k)$ were fitted. Table 2 gives the analysis from GLIM, with parameter values, standard errors, and changes in deviance due to removing each term in turn from the full composite model. In this case, the offset does not contain $-\log(y_k)$ of expression (3.3), since this is not present in the exponential distribution.

From these results, it is clear that the term for y_k may be eliminated, but not the other two. This strongly indicates that the gamma distribution is rejected in favour of the log normal distribution, as would be expected. At the same time, the Pareto and exponential are also clearly unacceptable. The deviance for goodness of fit of the log normal distribution is 98.93 with 197 d.f., but note that 152 of the frequencies are zero and that many others are very small. The mean and variance calculated from the parameter estimates, with the y_k term re-

Table 2
Comparison of the gamma and log normal models for the simulated data of Table 1 using GLIM

Term	Estimate	s.e.	Change in Deviance
1	2.660	0.4686	
y_k	−0.0002538	0.01313	0.00
$\log(y_k)$	1.290	0.3188	25.73
$\log^2(y_k)$	−0.4967	0.1046	25.23

Table 3
Returns from a postal survey for the number of occupants in each house

Number of Occupants	1	2	3	4	5	6	7
Number of Houses	436	133	19	2	1	0	1

moved, are: 2.302 and 1.003 respectively. The values calculated directly from the data are 2.296 and 0.997. If we extend n with, for example, 50 more zeroes, to $y_k = 250$, we obtain 2.299 and 0.997, even closer to the correct values.

Example 2. Consider the data in Table 3 from a postal survey giving the number of occupants in each house, kindly supplied by A.J. Scallan. We may wish to fit a Poisson distribution truncated at zero which has

$$p_k = \mathrm{e}^{-\mu - \log[1-\exp(-\mu)] + y_k \log(\mu) - \log(y_k!)}.$$

If we fit the log linear model, using GLIM, with the frequency vector extended with zeroes to $y_k = 15$, we obtain an intercept of 6.632 and a slope of -0.5505. The slope should be equal to $\log(\mu)$ which gives an estimate of the mean parameter equal to 0.5766. The intercept should be equal to $\log(N) - \mu - \log[1 - \exp(-\mu)]$. If we substitute in the value for μ just obtained, we have 6.632, identical to three decimals with our estimated intercept.

4. Comparing statistical models

Suppose now that we have a more complicated model, one with independent variables. Parameters and variables indexed by k will be taken, as above, to refer to the dependent variable defining the probability distribution, while i indicates values of the independent variable(s). We restrict attention to one independent variable without loss of generality.

Consider a composite model containing appropriate terms for all distributions of interest. If this can be reduced to a model of the form

$$\lambda_{ik} = \exp\left\{\theta_0 + \sum_{j=1}^{p} t_j(y_{ik})\theta_j + d_2(y_{ik})\right\},$$

then we have an identical distribution for all values of the independent variable, x. On the other hand, if the model must be

$$\lambda_{ik} = \exp\left\{\theta_{i0} + \sum_{j=1}^{p} t_j(y_{ik})\theta_{ij} + d_2(y_{ik})\right\},$$

where different elements of the vector, $\boldsymbol{\theta}_i$, are possible zero for various values of i, the probability distribution will change in form depending on what value x takes. Here we must distinguish two important cases.

In the situation to which we are accustomed in generalized linear models, and its special case, classical normal theory models, the parameter vector, $\boldsymbol{\theta}_i$, simply takes on different values as i varies. For example, if we take Poisson model (2.3) and introduce an independent variable, we obtain

$$p_{ik} = e^{-\mu_i + y_{ik}\log(\mu_i) - \log(y_{ik}!)},$$

when x is discrete. If x is continuous, it must be cut into discrete segments to form a contingency table for this method to be applied. (Unfortunately, little is known about the effects of grouping in independent variables; see Heitjan [8].) Thus, any computer program capable of fitting fairly general log linear models can be used to fit generalized linear models with discrete explanatory variables, although certainly not necessarily in the simplest manner possible!

Consider now the second, less usual situation, where we begin from a composite probability model, such as that illustrated above or, for example, by combining equations (2.4) and (3.2). We are hoping that certain terms will not be significant so that we can choose one of our alternative models. Here, this means that the same term(s) must be non-significant for all values of the independent variable(s). If such is the case, we fall back to the first situation, just discussed.

However, it is now possible to detect and fit different probability distributions for different values of the independent variable(s), i.e. for different sub-populations under study. Such a situation is easily imaginable, for example, in a mortality study, where test and control sub-populations have different hazard functions or in a medical study, where healthy and ill people have very different distributions of a substance in the blood. Now, different terms of the composite probability model are significant for (some of) the different sub-populations.

Example 3. We shall apply these procedures to the distributions of successive quarterly losses from two groups of staff recruited to the Post Office in the first quarter of 1973, presented in Table 4 of Burridge [3]. The two groups correspond to two different grades. In his presentation, Burridge uses a gamma distribution which would require the terms, y_k and $\log(y_k)$, in our procedure. In this example, we shall fit truncated distributions in order to accomodate the large number of survivors after 24 quarters.

When we fit the truncated gamma distribution, ignoring group differences, we obtain a deviance of 76.98 with 44 d.f. This indicates a large lack of fit (although 14 of the 48 cells are zero). (Adding group differences reduces the deviance to 76.79 with a loss of 2 d.f., perhaps indicating that, for this model, such differences are probably not required.) We try a composite model with the following additional terms: $1/y_k$, $1/y_k^2$, y_k^2, and $\log^2(y_k)$. In this way, we cover the normal, log normal, inverse Gauss, Pareto, exponential, and gamma distributions, as well as several extensions. We discover that, although several complex models may be possible, one of the simplest still has four terms required: y_k, $\log(y_k)$, $1/y_k$, and $1/y_k^2$. This model has a deviance of 39.79 with 42 d.f. This now appears to be an acceptable goodness of fit. Again, when we add differ-

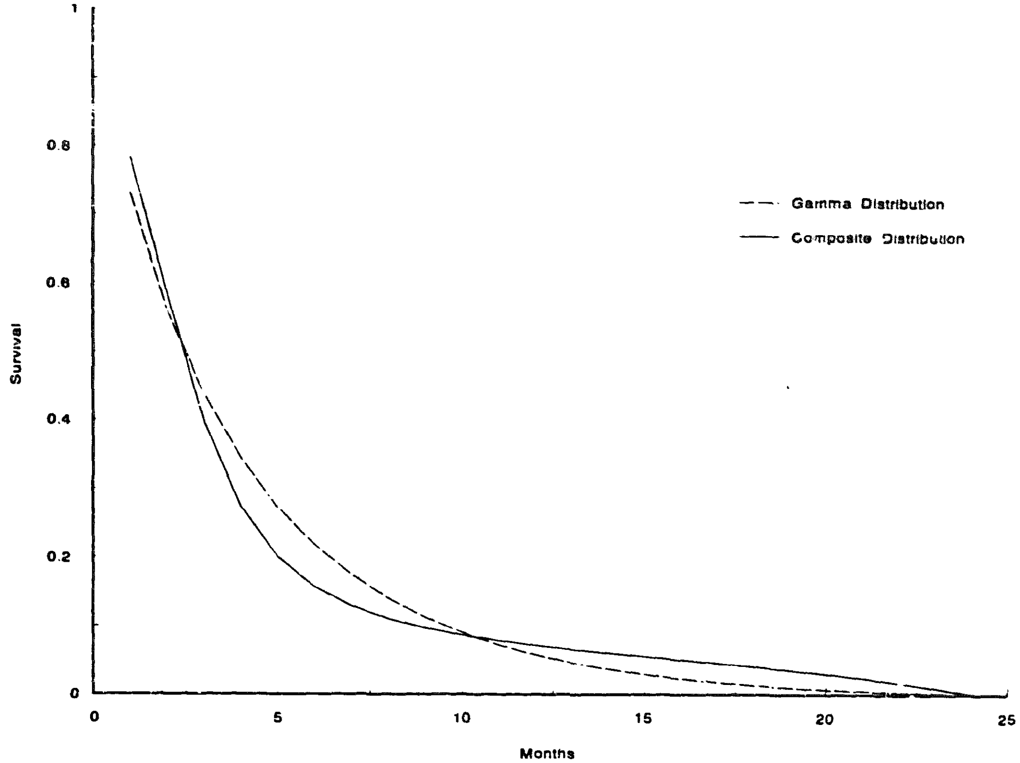

Fig. 1. Survivor functions for the gamma (dashed) and four term composite (solid) functions for the Post Office data (Burridge, [3]).

ences between groups, the resulting change is small: a deviance of 38.99 with 38 d.f. There is no indication that any individual term differs between the two groups.

We conclude that a relatively complex four parameter model is required to describe these data on staff leaving the Post Office, but that there exists no significant difference in the distribution of losses between the two grades. The common survivor curve for the two groups is plotted for the two models in Figure 1. Although the two curves may appear fairly similar, we have just seen that they are very significantly different. The composite model shows a function which drops more quickly in the early stages than the gamma survivor function, but then levels off at higher values.

Example 4. Consider now a much analyzed data set, that for time intervals between coal-mining disasters (Maguire et al. [14]). We use the corrected data of Jarrett [9]. We may test for a Poisson process in two ways. First let us group the data into intervals of 400 days and fit a Poisson distribution to the frequencies of disasters in these intervals. With a maximum frequency of 8 disasters per 400 days, we extend with zeroes to 14. The deviance is 21.127 with 13 d.f., indicating a poor fit. Second, let us look at the actual intervals between disasters, as an exponential distribution, regrouping them into intervals of width 20 days. The deviance is 106.20 with 123 d.f. By comparison, a truncated normal distribution has a deviance of 80.73 with 122 d.f.

From his Figure 1, Jarrett [9] suggests that the mean rate of disasters changed after about 125 events. We now cut our time series at this point and fit separate Poisson processes to each segment in the two ways just described. For the Poisson distribution, the deviance is now 5.89 with 26 d.f., while for the exponential distribution, it is 116.63 with 246 d.f., both indicating a very much improved fit.

Diggle and Marron [6] have applied density estimation to these data. Their Figure 1 clearly shows the presence of a mixture of two distributions. With our method, we have shown that this can be modelled as a 'mixture' of two Poisson processes, in fact separated in time.

5. Extensions

If we are prepared to go to non-linear Poisson regression, any probability distribution can be fitted by this method, but the software is not generally available. However, it is possible to fit some such models with existing software.

As we have seen, models for members of the exponential family can be fitted using the log linear framework requiring only the iterations necessary to fit any log linear model. If the statistical package used for fitting log linear models has some programming capabilities incorporated, as in the case of GLIM, a supplementary iteration procedure can be developed to estimate one or more parameters not following the exponential family. This could be done in a way similar to that already used for many survival distributions with GLIM, as proposed by Aitkin and Clayton [1], among others. Thus, for example, more general classes of survival distributions could be fitted and compared, with a wider range of different hazard functions possible for the various sub-populations. However, since the models are no longer linear in the parameters, their existence and uniqueness are no longer guaranteed.

Censored observations may also easily be treated in many cases. Consider one of the simplest, the exponential distribution. Equation (2.4) becomes

$$p_k \cong \mathrm{e}^{-c_k \log(\mu) - y_k/\mu} \Delta y_k,$$

where c_k is an indicator variable with one for uncensored observations and zero for censored. We fit c_k and y_k, with the usual offset and no intercept.

As an example of a model combining a distribution outside the exponentially family and censored observations, consider the Weibull distribution with

$$p_k \cong \mathrm{e}^{c_k \log(\alpha/\mu) + (\alpha - 1) c_k \log(y_k) - y_k^\alpha/\mu} \Delta y_k .$$

With an initial value of $\alpha = 1$ in y_k^α, we fit y_k^α, $c_k \log(y_k)$, and c_k, again with the usual offset and no intercept. From the second term, we obtain a new estimate of α, which we use in the first term in the iteration, continuing until convergence, if a solution exists. An advantage of this approach for linear models, with independent variables, is that the α parameter may be allowed to vary with the independent variables as easily as the usual parameters of the exponentially family.

Similar iterative methods can easily be developed for other distributions close to the exponential family. One such possibility is the delta algorithm, available in GLIM: see Jorgensen [10].

6. Discussion

The method described in this paper may prove useful in a number of contexts for several reasons. With a single algorithm, Poisson regression, available in many statistical packages, a large number of probability distributions may be fitted. This contrasts with classical methods which require a distinct algorithm for each distribution. The hypotheses are very weak: independent observations following a multinomial distribution or, equivalently, a Poisson distribution with fixed total number of observations. The integration constant need not be known in advance, but is obtained numerically. This allows us to estimate distributions which otherwise might be excluded because of their analytic complexity. For example, parameter estimates for truncated distributions may easily be obtained. A good model may often be rapidly chosen using the well-known methods of stepwise regression and nested linear models. Distinct distributions with different forms may easily by identified as a function of discrete independent variables. This can often provide a powerful replacement for such classical approaches as discriminant analysis.

Acknowledgements

The authors would like to thank Bent Jorgensen for helpful comment on an early draft and Dr. J. Burridge and A.J. Scallan for comments and suggestions which substantially clarified the presentation.

References

[1] M. Aitkin and D. Clayton, The fitting of exponential, Weibull and extreme value distributions to complex censored survival data using GLIM. *Appl. Statist.* **29** (1980) 156–163.

[2] A.C. Atkinson, A method of discriminating between models. *J. Roy. Statist. Soc. Ser. B* **32** (1970) 323–353.

[3] J. Burridge, Empirical Bayes analysis of survival time data. *J. Roy. Statist. Soc. Ser. B* **43** (1981) 65–75.

[4] D.R. Cox, Tests of separate families of hypotheses. *Proc. 4th Berkeley Symp.* **1** (1961) 105–123.

[5] D.R. Cox, Further results on tests of separate families of hypotheses. *J. Roy. Statist. Soc. Ser. B* **24** (1962) 406–424.

[6] P. Diggle and J.S. Marron, Equivalence of smoothing parameter selectors in density and intensity estimation. *J. Amer. Statist. Assoc.* **83** (1988) 793–800.

[7] S.J. Haberman, *The Analysis of Frequency Data.* (University of Chicago Press, Chicago, 1974).

[8] D.F. Heitjan, Inference from grouped continuous data: a review. *Statistical Science* **4** (1989) 164–183.

[9] R.G. Jarett, A note on the intervals between coal-mining disasters. *Biometrika* **66** (1979) 91–193.

[10] B. Jorgensen, The delta algorithm and GLIM. *Int. Statist. Rev.* **52** (1984) 283–300.

[11] J.K. Lindsey, Comparison of probability distributions. *J. Roy. Statist. Soc. Ser. B* **36** (1974) 38–47.

[12] J.K. Lindsey, Construction and comparison of statistical models. *J. Roy. Statist. Soc. Ser. B* **36** (1974) 418–425.

[13] J.K. Lindsey, *The Analysis of Categorical Data Using GLIM.* (Springer Verlag, New York, 1989).

[14] B.A. Maguire, E.S. Pearson, and A.H.A. Wynn, The time intervals between industrial accidents. *Biometrika* **40** (1952) 212–213.

[15] M.J. Silvapulle and J. Burridge, Existence of maximum likelihood estimates in regression models for grouped and ungrouped data. *J. Roy. Statist. Soc. Ser. B* **48** (1986) 100–106.