Journal of Applied Statistics
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/cjas20

Quasi-likelihood and pseudo-likelihood are not the same thing

J. A. Nelder [a]
[a] Department of Mathematics , Imperial College , UK
Published online: 02 Aug 2010.

To cite this article: J. A. Nelder (2000) Quasi-likelihood and pseudo-likelihood are not the same thing, Journal of Applied Statistics, 27:8, 1007-1011, DOI: 10.1080/02664760050173328

To link to this article: http://dx.doi.org/10.1080/02664760050173328

Journal of Applied Statistics, Vol. 27, No. 8, 2000, 1007–1011

Quasi-likelihood and pseudo-likelihood are not the same thing

J. A. NELDER, *Department of Mathematics, Imperial College, UK*

ABSTRACT *Models described as using quasi-likelihood (QL) are often using a different approach based on the normal likelihood, which I call pseudo-likelihood. The two approaches are described and contrasted, and an example is used to illustrate the advantages of the QL approach proper.*

1 Introduction

A recent paper published in this journal (Mvoi & Lin, 2000, to be referred to as [ML]), discusses what the authors call quasi-likelihood (QL) and an extension, which they call AQL. In my view they are discussing a related topic, which I shall call pseudo-likelihood (PL). In this paper I shall distinguish between the two approaches, and give reasons for preferring QL to PL in defining classes of models for analysing data.

2 Generalized linear models (GLMs)

Wedderburn (1974) derived QL as an extension of GLMs, so it is necessary to give a brief description of the latter first. In classical normal models, a response y is assumed to have normal i.i.d. errors for the random part and, for the systematic effects, a set of explanatory variables $x_1, x_2, \ldots, x_k$ assumed related to the mean μ of y by $\mu = \sum \beta_j x_j$, where β_j are parameters to be estimated. GLMs involve two extensions to this formulation. First, the class of error distributions is extended to one-parameter exponential families, which include Poisson, binomial, gamma and inverse Gaussian distributions, and secondly the systematic part assumes additivity of x effects on some transformed scale, given by $\eta = g(\mu)$ where $\eta = \sum \beta_j x_j$. The quantity η is known as the linear predictor and $g()$ as the link function. In this framework, classical normal models have error distribution normal and link identity,

Correspondence: Department of Mathematics, Imperial College, London, UK.

ISSN 0266-4763 print; 1360-0532 online/00/081007-05

log-linear models have error Poisson and link function log, logistic regression has error binomial and link logit, and so on. GLMs have a common algorithm for fitting them by maximum likelihood, namely iterative weighted least squares, acting on an adjusted dependent variable z, defined by

$$z = \eta + (y - \mu)(\mathrm{d}\eta/\mathrm{d}\mu)$$

in place of y, and a weight W given by

$$1/W = \mathrm{var}(z) = \mathrm{var}(y)(\mathrm{d}\eta/\mathrm{d}\mu)^2$$

The goodness-of-fit statistic (more strictly the badness-of-fit statistic) is derived from the log-likelihood-ratio statistic and is called the *deviance*. It generalizes the residual sum of squares of normal models. For a detailed account of GLMs, see McCullagh & Nelder (1989).

The form of the variance in GLMs is important: it has the general form $\mathrm{var}(y) = \phi V(\mu)$, where $V(\mu)$, which depends on μ, is called the variance function and ϕ, which is independent of μ, is called the dispersion parameter. This form expresses algebraically what Box (1988) has termed the separation principle in models for the joint modelling of both mean and dispersion (Lee & Nelder, 1998).

3 Quasi-likelihood

In his original derivation Wedderburn (1974) began with the score equation for GLMs, which takes the simple form

$$\mathrm{d}l/\mathrm{d}\mu = (y - \mu)/(\phi V(\mu))$$

where l is the log likelihood. For some GLMs, in particular those based on the Poisson and binomial distributions, the dispersion parameter is fixed *a priori* at unity. With count data it often happens that the mean deviance (a measure of dispersion), after fitting a suitable model with Poisson errors, is considerably in excess of unity. We should like to be able to use a variance of the form $\phi\mu$ with $\phi > 1$. However, there is no distribution of the GLM family having this property. Wedderburn's solution was to define a quasi-likelihood (more strictly a quasi-log-likelihood) q, by the analogous relation

$$\mathrm{d}q/\mathrm{d}\mu = (y - \mu)/(\phi V(\mu))$$

and to derive estimates of β by maximizing q. He showed that maximum-quasi-likelihood estimates (MQLEs) had many properties in common with MLEs. More recently, Lee & Nelder (1999) have shown that MQLEs have an optimum property with respect to the class of all distributions with a given variance function.

There are two common misunderstandings about QL that need discussion. The first relates to the fact that we appear to have lost the probability distribution for the model when we use a QL that cannot be interpreted as one deriving from a GLM distribution. However, by normalizing the quasi-likelihood proper, $\exp(q)$, we can make it into a distribution, which we call a QL distribution. The normalizing factor of such a distribution will now depend on μ, whereas in GLMs it does not. This means that the MQL estimators and those derived from the QL distribution will not coincide. However, if, as is common, the normalizing factor changes only slowly with the mean, differences between the two estimators will be small. Thus, MQL estimates may be regarded as good approximations to ML estimates from the QL distribution.

The second misunderstanding about QL is to assert that, because the definition depends only on the first two moments (which is true), it implies that we are saying nothing about higher cumulants (which is not true). The normal distribution is defined by its first two moments, but in assuming it we are also assuming that all the higher cumulants are zero. In GLMs the cumulants are obtained from the derivatives of the function $b(\theta)$ where θ is the canonical parameter occurring in the kernel $(y\theta - b(\theta))$ of the log-likelihood of one-parameter exponential families. The third and fourth cumulants of a QL distribution are often approximated well by the formulae derived from $b(\theta)$ that would hold if there was an exact GLM distribution. In particular, by choosing a non-normal QL we shall be assuming that the distribution of the data errors is skew.

Important QLs are those for overdispersed Poisson and binomial distributions, and those of the form $\text{var}(y) = \phi\mu^{\alpha}$ for continuous data with non-constant variance.

4 Pseudo-likelihood (PL)

PL models arise when the normal likelihood is extended to allow the variance σ^2 to become a function of the mean μ. This is what happens with [ML]'s models on p. 348, derived from Muller & Zhao (1995). Some of these models are of the strict GLM form, and some also involve parameters in the variance function. There are two snags about this generalization: the first is that the assumption of a symmetric distribution for the errors is being retained. Now, while there is no reason in theory why errors should not be symmetric when the variance is a function of the mean, in practice I have found that they are always skew in such circumstances. There is no problem with QL models in this respect, since skewness is built into the model. The second snag is that now the normalizing factor for the distribution, $1/\sqrt{(2\pi\sigma^2)}$ is also a function of the mean and so should be included in the ML equations for beta. Without the inclusion of this factor, the estimators may not be consistent. No such problem arises with GLMs.

The algorithm for PL models, as given by Montgomery & Myers (1997), allows for a link function and uses a linearization of μ with the unadjusted response y and weights based the inverse of σ^2. The quantity being minimized is a Pearson-type weighted sum of squares, giving rise to Pearson-type residuals. These latter are known in non-normal GLMs to be inferior to deviance residuals in their distributional properties. Readers of that paper should not be misled by the title; the authors believe that they are describing GLMs, but in fact are describing PL models. For an elementary account of GLMs proper see the following paper (Hamada & Nelder, 1997).

Note that the split between dispersion parameter and variance function, so useful in GLMs, is blurred in [ML]'s account, which moves between models where the variance is a function of μ and where it is a function of x. GLMs that have ϕ as a function of covariates with parameters needing to be estimated occur in quality-improvement experiments where we seek to minimize the variance while keeping the mean fixed. This leads to joint modelling of mean and dispersion, the algorithm for which can be reduced to two interconnected GLMs (Lee & Nelder, 1998). The GLM for the mean uses prior weights derived from the reciprocals of the fitted values from the dispersion GLM and the GLM for the dispersion uses as a response the standardized deviance components derived from the mean GLM. The method does not assume normality for the errors, and allows an arbitrary GLM for the mean.

TABLE 1. Fits of PL and QL models

Method	parameters	deviance	d.f.
PL	β_0,β_1	6.948	82
PL	$\beta_0 = 0$, β_1	6.906	83
QL	$\beta_0\beta_1$	6.347	82
QL	$\beta_0,\beta_1 = 1$	6.456	83

5 An example

I shall use the assay data modelled by [ML] and given in their Table 1. The response y is the result from a test method of a hormone assay; there is a single explanatory variate x, a reference method for the same assay. The scatter plot of y against x, shown in their Fig. 1, reveals that the variance is increasing with the mean. It also shows that y is roughly proportional to x and that the origin lies on the line. There are thus two possible approaches to modelling the systematic part: one uses an identity link with $\mu = \beta_0 + \beta_1 x$ and the other uses a log link with $\eta = \log(\mu) = \beta_0 + \beta_1 \log(x)$. For the first model $\beta_0 = 0$ is a special point for checking whether the fitted line goes through the origin, and for the second model $\beta_1 = 1$ checks for linearity. The last point in the data (#85) has the largest y and x and is somewhat isolated from the rest, so [ML] decide to discard it *a priori*; for comparison's sake we shall follow them in this, although model checking shows that it is not different from the rest. Note that [ML] do not consider a non-identity link function so that $E(Y)$ is assumed linear on x. If we follow them in assuming that $\text{var}(y) = \phi\mu^2$, the appropriate GLM is one using the gamma distribution for the errors, which fits naturally with a log link.

For the pseudo-likelihood fit we use normal errors and a weight inversely proportional to the square of μ, iterating as required. For the QL model we use a log link and variance function proportional to μ^2, which is equivalent to using a GLM with a gamma distribution. The fits are shown in Table 1.

Note that (i) the fits as judged by the deviance are slightly better for the QL models and (ii) that the fit after dropping a parameter is better for PL, whereas that for QL is worse, as would be expected. The deviance in a QL model never decreases as parameters are dropped, whereas the chi-squared-like criterion for PL models does not have this desirable property. The histogram for the residuals from the first PL fit is markedly skewed to the right, as would also be expected, showing that a model that assumes symmetry in the errors is contradicted by the data.

The inclusion of point #85 in the fit increases the first and third deviances in the table to 7.000 and 6.420 respectively, showing that there is no reason not to include it. Although this point is acceptable, the QL plot of residuals versus fitted values shows a group of five points with over-large positive residuals. This group is also shown clearly in the top left-hand corner of Fig. 4 of [MV]. It has an effect when we try to check if our assumption of μ^2 for the variance function is right for the QL model. When these points are included the optimum power for α in the variance function μ^α is about 1.74, and this is marginally different from the prior value of 2, as judged by the extended quasi-deviance (Nelder & Pregibon, 1987). However, when the points are omitted, the optimum value of α is very close to 2.

6 Conclusions

I contend that the PL approach shows several disadvantages over the QL approach for modelling non-normal data. It assumes symmetry of errors when the variance changes with the mean, although this seems never to occur in practice. The fitting criterion does not have the property that it always increases if a parameter is omitted from the model. It does not naturally encourage the split, characteristic of GLMs, of the variance into dispersion parameter and variance function, leading in QL models to a general algorithm for models where both components depend on covariates. Lastly, the PL models in [MV] require special code, whereas the analogous QL models can all be fitted by any decent statistical system that supports GLMs. The analyses in this paper were done with Genstat (Payne *et al.*, 1993), and the K system, a set of Genstat procedures developed by myself; it supports highly interactive analysis of GLMs, including QL models.

REFERENCES

Box, G. E. P. (1988) Signal-to-noise ratios, performance criteria and transformations, *Technometrics*, 30, pp. 1–17.

Lee, Y. & Nelder, J. A. (1988) Generalized linear models for the analysis of quality-improvement experiments, *Canad. J. of Statistics*, 26, pp. 95–105.

Lee, Y. & Nelder, J. A. (1999) The robustness of the quasi-likelihood estimator, *Canad. J. of Statistics*, 27, pp. 321–327.

McCullagh, P. & Nelder, J. A. (1989) *Generalized Linear Models* (New York, Chapman & Hall).

Montgomery, D. L. & Myers, R. H. (1997) A tutorial on generalized linear models, *J. Quality Technology*, 29, pp. 274–291.

Muller, H. G. & Zhao, P.-L. (1995) On a semiparametric variance function model and a test for heteroscedasticity, *The Annals of Statistics*, 23, pp. 946–967.

Mvoi, S. & Lin, Y.-X. (2000) Criteria for estimating the variance function used in the asymptotic quasi-likelihood approach, *J. Appl. Statist.*, 27, pp. 347–362.

Hamada, M. & Nelder, J. A. (1997) Generalized linear models for quality-improvement experiments, *J. Quality Technology*, 29, pp. 292–304.

Nelder, J. A. & Pregibon, D. (1987) An extended quasi-likelihood function, *Biometrika*, 74, pp. 221–232.

Payne, R. W. (Ed.) (1993) *Genstat 5 Release 3 Reference Manual* (Oxford, Clarendon Press).

Wedderburn, R. W. M. (1974) Quasilikelihood functions, generalized linear models and the Gauss–Newton method, *Biometrika*, 61, pp. 439–447.