Institute of Mathematical Statistics

LECTURE NOTES — MONOGRAPH SERIES

EXTENDED QUASI-LIKELIHOOD AND ESTIMATING EQUATIONS

J. A. Nelder
Imperial College, London

Y. Lee
Seoul National University, Korea

1 INTRODUCTION

This paper compares the estimation of mean and dispersion parameters using

(i) extended quasi-likelihood (EQL) and

(ii) optimum estimating equations (OEE).

1.1 Extended quasi-likelihood (EQL)

Quasi-likelihood (QL) was introduced by Wedderburn (1974) as a way of weakening the distributional assumptions of GLMs (McCullagh and Nelder, 1989) by specifying only the form of the linear parameters β and the variance function $V(\mu)$, which expressed the variance as a function of μ. The linear score function

$$dl/d\mu = (y - \mu)/V(\mu)$$

of GLMs was preserved in the wider class of models, and estimates obtained from maximizing the QL have many properties analogous to ML estimators. The importance of QLs is that they allow us to extend the ideas of GLMs where no exponential family exists to supply an error structure of the original form.

EQL was introduced by Nelder and Pregibon (1987) to allow comparison of different variance functions on the same data. It is the key to methods for the joint modelling of mean and dispersion based on QL ideas.

A QL model defines a deviance component d_i for observation i with mean μ_i given by

$$d_i = 2\int_{\mu_i}^{y_i} \frac{y_i - u}{V(u)} du. \tag{1}$$

The deviance $D = \Sigma d_i$. EQL is most simply stated in terms of its extended deviance

$$D^+ = \sum_i \frac{d_i}{\phi_i} + \sum_i \log(2\pi\phi_i V(y_i)). \tag{2}$$

Here the dispersion parameter ϕ is allowed to vary over the observations. For distributions of the GLM family, the EQL can be derived as a saddle-point approximation in which all the factorials are replaced by their Stirling approximations. It is thus exact for the Normal and inverse gamma distributions, which have no factorials, and for the gamma differs only in the normalizing factor. Note that the EQL corresponds to an exact distribution of the kind Jorgensen (1996) calls a *regular proper dispersion model* whenever the normalizing factor is exactly one. The quantity D^+ can be used as a criterion to fit models when both the mean and dispersion are assumed functions of explanatory variables.

1.2 Optimum estimating equations (OEE)

If a function $g(Y;\theta)$ of a random variable Y, having a distribution depending on θ, has zero mean for all θ, then $g(Y;\theta)$ is an estimating function. If θ is a scalar then the optimum estimating equation (OEE) is based on

$$h = g.E(dg/\partial\theta)/E(g^2)$$

and for a sample of n independent Ys the OEE is given by

$$\sum_{i=1}^{n} h_i = 0.$$

More generally if θ is a vector, $\boldsymbol{V}$ is the conditional covariance matrix of g, given a model matrix A and $E(g|A) = 0$, the optimum estimating function is given by

$$U(\theta;y) = D^T V^{-1} g,$$

where

$$D_{ir} = -E\left(\frac{\partial g_i}{\partial\theta_r}\,|A\right).$$

If U is linear (quadratic) in y we have an optimum linear (quadratic) estimating function (Godambe, 1991).

It may happen that we require more than one estimating function, and this occurs if we require to estimate both mean and dispersion parameters in a model. Godambe and Thompson (1989) derive such joint estimating equations, and use components $(y-\mu)$ and $(y-\mu)^2$, orthogonalizing the second with respect to the first. We discuss this case further in Section 2.2 below.

2 JOINT ESTIMATION OF MEAN AND DISPERSION

2.1 The EQL criterion

For ϕ_i given, the estimates of β obtained from (2) are just those given by the Wedderburn QL equations

$$\sum \frac{y_i - \mu_i}{\phi_i V_i} \frac{\partial \mu}{\partial \beta} = 0$$

while for given μ, D^+ is the same as would be obtained from a QL model with response variable d and variance function ϕ^2, i.e. that of the gamma distribution. That this may be a good approximation even when y has a non-normal distribution may be seen from the result of Pierce and Schafer (1986) that the deviance function is nearly an optimal normalizing transform.

Figure 1 shows for the Poisson distribution the expectation and variance of d, which should be 1 and 2 respectively, and also the correlation of $(y - \mu)$ and d, which should be zero. For $\mu \geq 3$ the approximations are acceptable, but, not surprisingly, break down for small μ.

We are thus led to form two interlinked GLMs, one for the mean and one for the dispersion. That for the mean has response y, prior weight ϕ^{-1}, a linear predictor $\eta = \Sigma \beta_j x_j$, and some suitable link and variance functions. For the dispersion we have response d, a linear predictor $\xi = \Sigma \gamma_k u_k$ with explanatory variables $u_1, u_2, ...$, a variance function ϕ^2 and a suitable link, usually chosen as log.

An obvious algorithm for fitting the joint model is a 'see-saw' one, in which the model for the mean is fitted first with current estimates of ϕ_i, then the model for the dispersion is fitted given current estimates of μ_i and hence of d_i. Three cycles, starting with $\phi = 1$, are often sufficient to fit the joint model. Standard GLM techniques may be used to check both models for internal consistency.

2.2 The OEE criterion

Godambe and Thompson (1989) derive what they call an extended quasi-score function which gives estimating equations of the form

$$\sum \{(y_i - \mu_i)/V(\mu_i) - a_i h_i\} \frac{\partial \mu_i}{\partial \beta_r} = 0, \tag{3}$$

for the mean parameters β where,

$$h_i = (y_i - \mu_i)^2 - \phi V(\mu_i) - \gamma_{1i}(\phi V(\mu_i))^{1/2}(y_i - \mu_i) \tag{4}$$

and

$$\Sigma h_i = 0,$$

for the dispersion parameter ϕ. In these equations γ_{1i} and γ_{2i} are the standardized third and fourth cumulants and

$$a_i = (\gamma_{1i} - \gamma_{1i}^*)/\{\phi^{1/2} V(\mu_i)^{3/2} (\gamma_{2i} + 2 - \gamma_{1i})^2\},$$

where

$$\gamma_{1i}^* = (\phi/V(\mu_i))^{1/2} \frac{\partial V}{\partial \mu}$$

is the exponential skewness. For simplicity these equations are given assuming a constant dispersion. If ϕ is structured they can be extended in a standard way.

The form in which a_i is defined shows immediately that for distributions of the GLM family a_i is identically zero, because $\gamma_1 = \gamma_1^*$, so that (3) reduces to the QL equations for β. Thus for GLM models, the optimum linear, the optimum quadratic and the quasi-likelihood estimating equations for β are all the same. In (4) the last term is identically zero for normal errors, or for gamma errors with a log link, and is usually much smaller than the first; thus, approximately, (4) is equivalent to equating the Pearson X^2 to its expectation, uncorrected for d.f. lost in fitting β.

3 COMPARISON OF EQL AND OEE CRITERIA

There are two basic (but connected) differences in the equations for the joint estimation of mean and dispersion produced by the use of EQL and OEE. The first concerns the response variable for the dispersion; the use of the function $(y - \mu)^2$ in OEE equations leads to the Pearson X^2, whereas the EQL equations use the deviance component. The second difference, which follows from the first, is that the OEE equations, in general, require knowledge of γ_1 and γ_2, or equivalently the third and fourth cumulants, whereas the EQL equations do not.

For normal models $\gamma_1 = \gamma_2 = 0$, and the EQL and OEE equations are identical. For GLM non-normal models the equations for the mean parameters are the same, but those for the dispersion differ, while for non-GLM models both sets of equations are different. What reasons are there for preferring one method to the other?

3.1 Pseudo-likelihood v quasi-likelihood

Nelder and Lee (1992) compared estimates from the EQL having

$$D^+ = \sum d_i/\phi_i + \Sigma \log(2\pi\phi_i V(y_i)),$$

with pseudo-likelihood (PL) estimates obtained from minimizing

$$D_p = \Sigma X_i^2/\phi_i + \Sigma \log(2\pi\phi_i V(\mu_1))$$

for three non-GLM models. Model (1) involved the NBα distribution, a form of negative binomial distribution obtained mixing the Poisson with a gamma where the shape parameter ν varies with μ, instead of the scale parameter α. The response y has

$$var(y) = \mu(1 + \alpha),$$

i.e. looks like an overdispersed Poisson. We did a 5-factor simulation in a 2^{5-1} fractional factorial, the factors being aspects of the configuration of the means we thought might be important. Model (2) was a mixture of two Poisson samples with a ratio of means of 4:1; the variance function was assumed to have the form $\phi\mu^\alpha$, and estimates of α were of interest. Model (3) was a Poisson-Inverse Gaussian mixture with the inverse Gaussian distribution parameterized so that the variance function had the form $\mu+\alpha\mu^2$ (the IG-2 distribution). The experimental factors were sample size and value of α. For detailed results see our paper. The main conclusion was that though the bias of the maximum EQL (MEQL) estimator was usually larger than that of the maximum PL (MPL) estimator, this was more than offset in moderate sample sizes by the larger variance of the latter; the result was that in terms of MSE, the MEQL estimator was never appreciably inferior to the MPL estimator, and was often much better. An interesting result was that in finite samples the MEQL estimate was frequently better than the ML estimate.

3.2 The value of knowing γ_1 and γ_2

The OEE for mean and dispersion require knowledge of γ_1 and γ_2 for non-GLM models, and knowledge of these can improve estimates. For example, Lee and Nelder (unpublished) consider a model with log link and NBα errors, and two groups having means μ and 2μ. Table 1 shows the asymptotic variance ratios for estimates of the group difference for various μ and ϕ. The ML estimates are equivalent to those derived by knowing all the cumulants, the estimates based on quadratic estimating functions (QEF) to knowing the third and fourth cumulants, and the MEQL estimates to assuming an exponential-family pattern for the cumulants. For large overdispersion ($\phi = 5$) there is considerable loss of efficiency of MEQL relative to the ML estimates, with the QEF results showing that a considerable part of this loss can be recovered if γ_1 and γ_2 are known. Similar results were found with another example using the IG-2 distribution for errors.

While these results are interesting, they assume that the distribution is known at least up to the fourth cumulant. In practice this is almost never so.

Attempts to estimate γ_1 and γ_2 from moderate amounts of data may lead to estimates with considerable errors; for example, for a sample of 10 from a Poisson distribution with $\mu = 5$, 100 simulated samples gave 33% of estimates of γ_1 with negative values, i.e. of the wrong sign. We need to know, therefore, what loss of information arises with various estimators when we assume values ρ_1 and ρ_2 for the unknown γ_1 and γ_2. We shall be mainly concerned with the optimum quadratic estimating function, the ML estimator and the MEQL estimator. For simplicity we restrict the argument to independent Y_i.

The QL equations assume a GLM pattern of cumulants. The optimum QEF equations improve efficiency by using information from the third and fourth cumulants and the ML equations from the all the cumulants. So the ML estimator will be most informative if the true model is known. But it is often not consistent if a model is wrongly chosen.

Consider the trace or determinant of normalized asymptotic variance $n^{-1/2}cov(\hat{\beta})$ as the risk. For a given $V(\mu)$ with μ fixed, the MEQL estimator is a mini-max estimator among QEF estimators, since its risk remains constant, *i.e.* does not depend upon the true values of γ_1 and γ_2; it attains the minimum risk under the GLM skewness. The MEQL estimator is also the ML estimator if a GLM family exists. If so it is again a mini-max estimator among ML estimators for the class of distributions with a given $V(\mu)$. Therefore, the MEQL estimator would be most conservative (in the sense that the possible maximum risk is minimal) against a possible misspecification of either the likelihood or cumulants of the model.

Godambe and Thompson's joint QEFs with $\rho_1 = \rho_2 = 0$ lead to the Normal ML estimator for β. For Normal heteroscedastic linear models, assuming only the first moment is correctly specified, Carroll and Ruppert (1982) showed that the MEQL estimator is robust against a small variance-function mis-specification compared with the Normal ML estimator. Under mild regularity conditions, the consistency of the MEQL estimator depends only upon the correct specification of the regression while that of the optimum QEF estimator requires also the correct specification of $V(\mu)$; see Crowder (1987). The QEF equations (2) become the QL equations when $a_i = 0$. When $V(\mu)$ is mis-specified, the QEF estimator is no longer consistent unless $a_i = 0$; see (2). So the MEQL estimator is most robust among QEF estimators against a mis-specification of $V(\mu)$.

We illustrate the nature of the minimax property of the MQL estimator by two of the examples from Section 3.1 (Lee and Nelder, unpublished).

In the first the unknown true distribution is the NBα distribution. For this $\gamma_1 = \theta_1/\mu^{1/2}$ and $\gamma_2 = \theta_2/\mu$, where $\theta_1 = (1 + 2\alpha)/(1 + \alpha)^{1/2}$ and $\theta_2 = (1 + 6\alpha + 6\alpha^2)/(1 + \alpha)$. Suppose we assume values $\rho_1 = \lambda_1/\mu^{1/2}$ and $\rho_2 = \lambda_2/\mu$ with λ_1 and λ_2 being values of θ_2 and θ_2 for which $\alpha = 0.2$, 1 and

5, corresponding to small, moderate and large amounts of overdispersion. Figure 2 shows the lower bound of the variance ratio for the QEF estimator with respect to the MEQL estimator over the θ_1 scale, when $\mu = 1$, an unfavourable value where the MEQL estimator is known from simulation to have low efficiency. It is clear that having the assumed value of θ_1 too high can lead to much greater loss of efficiency than the corresponding gain in efficiency when the correct value is chosen. For larger μ this effect would be even more marked.

In the second example the unknown true distribution is the IG-2 distribution. Here $\gamma_1 = 3/\nu^{1/2}$ and $\gamma_2 = 15/\nu$, where $var(Y) = \mu + \mu^2/\nu$. Let assumed values ρ_1 and ρ_2 be γ_1 and γ_2 values at $\nu = 0.5$, 1, and 2, so that the corresponding values of ρ_1 are 4.243, 3, and 2.121 respectively. Figure 3 shows similar curves to Figure 2 over γ_1, the true skewness scale. Here the losses from using the MEQL estimator when ρ_1 and ρ_2 are nearly right are much less than the gains when they are too large.

4 CONCLUSION

We can derive the EQL equations for mean and dispersion from the OEEs by replacing the h_i of equation (4) by $d_i - \phi_i$, where d is the deviance component and ϕ the dispersion parameter, which may be structured. If we make three approximations, namely that $E(d) = \phi$, $var(d) = 2\phi^2$ and $\text{corr}(y - \mu; d) = 0$, the resulting estimating equations are given by

$$\Sigma(y - \mu)\frac{\partial \mu}{\partial \beta}\Big/[\phi V(\mu)] = 0$$

$$\Sigma(d - \phi)\frac{\partial \phi}{\partial \gamma}\Big/\phi^2 = 0$$

and these are just those obtained from the use of EQL. Note that the three approximations become exact as $\mu \to \infty$; however, their joint effects for small μ need further investigation.

The properties of the MEQL estimator give it a special place among estimators in the joint estimation of mean and dispersion.

References

CARROLL, R. J. and RUPPERT, D. (1982). A comparison between maximum likelihood and generalized least squares in a heteroscedastic linear model. *J. Am Statist. Assoc.*, **77**, 878-882.

COX, D. R. (1993). Some remarks on overdispersion. *Biometrika*, **70**, 279-274.

CROWDER, M. (1986). On consistency and inconsistency of estimating equations. *Econometrics Theory*, **3**, 305-330.

CROWDER, M. (1987). On linear and quadratic estimating functions. *Biometrika*, **74**, 591-597.

FIRTH, D. (1987). On the efficiency of quasi-likelihood estimation. *Biometrika*, **74**, 2333-245.

FIRTH, D. (1988). Multiplicative errors: lognormal or gamma? *J. R. Statist. Soc. B*, **50**, 266-268.

GODAMBE, V. P. and THOMPSON, M. E. (1989). An extension of quasi-likelihood estimation. *J. Statist. Plann. Inference*, **22**, 137-152.

GODAMBE, V. P. (Ed.) (1991). *Estimating Functions.* Oxford: Clarendon Press.

JORGENSEN, B. (1996). Proper dispersion models. *Braz. J. Probab. Statist.* (to appear).

McCULLAGH, P. (1983). Quasi-likelihood functions. *Ann. Statist.*, **11**, 59-67.

McCULLACH, P. and NELDER, J. A. (1989). *Generalized Linear Models*, 2nd edn. London: Chapman and Hall.

NELDER, J. A. (1989). Discussion of the paper by Godambe and Thompson. *J. Statist. Plann. Inference*, **22**, 158-160.

NELDER, J. A. and LEE, Y. (1992). Likelihood, quasi-likelihood and pseudo-likelihood: some comparisons. *J. R. Statist. Soc. B*, **54**, 273-284.

NELDER, J. A. and PREGIBON, D. (1987). An extended quasi-likelihood function. *Biometrika*, **74**, 221-231.

PIERCE, D. A. and SCHAFER, D. W. (186). Residuals in generalized linear models. *J. Am. Statist. Assoc.*, **81**, 977-986.

WEDDERBURN, R. W. M. (1974). Quasi-likelihood functions, generalized linear models and the Gauss-Newton method. *Biometrika*, **61**, 439-447.

Table 1: Asymptotic variance ratios of the ML, MQL and optimum QEF estimators of group difference for NBa distribution

	$\phi = 2$				$\phi = 5$			
μ	1	5	10	50	1	5	10	50
ML/MQL	.875	.963	.980	.996	.608	.805	.885	.974
QEF/MQL	.900	.966	.981	.996	.768	.865	.911	.976

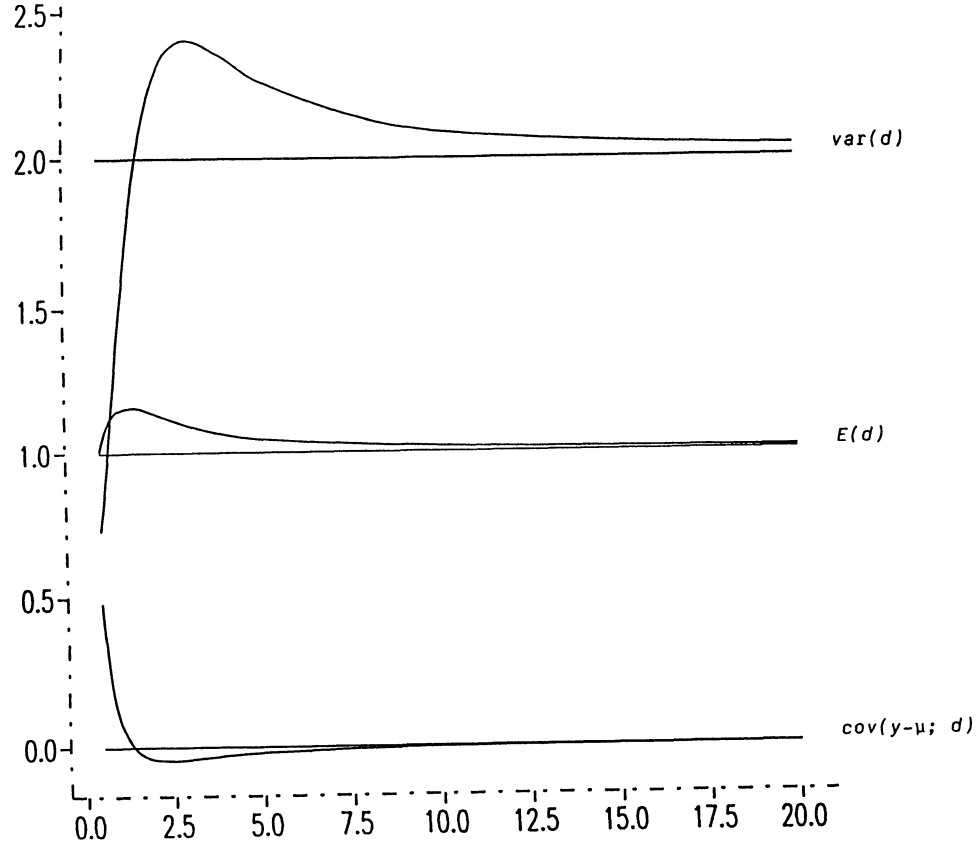

Figure 1: Properties of the Poisson deviance as a function of μ.

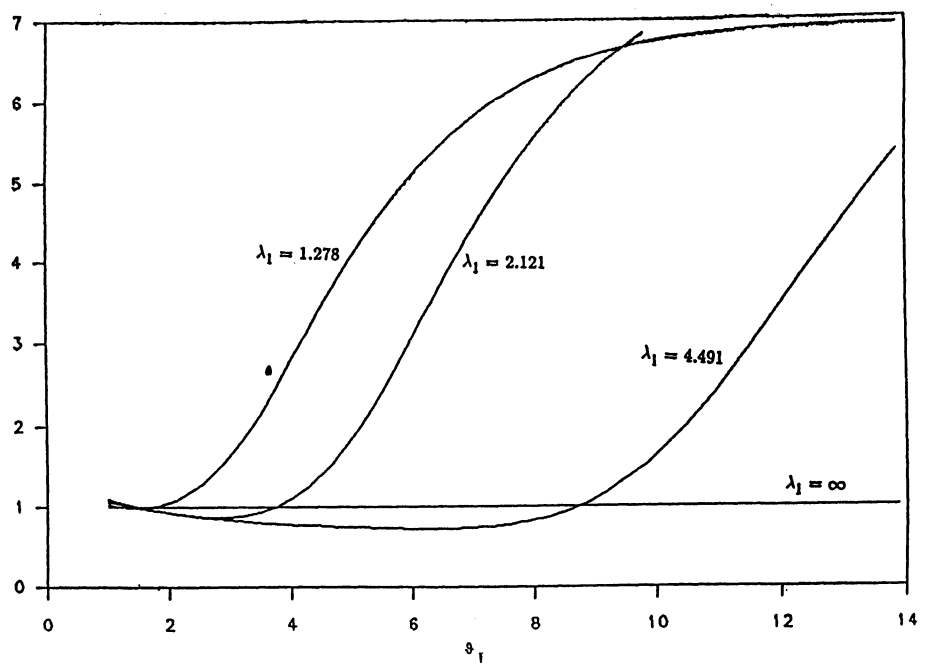

Figure 2: Asymptotic variance ratio of QEF/MEQL estimators as a function of θ ; example with NBα distribution. Curves with increasing λ_i correspond to increasing amounts of actual overdispersion. Abscissa is assumed amount of overdispersion.

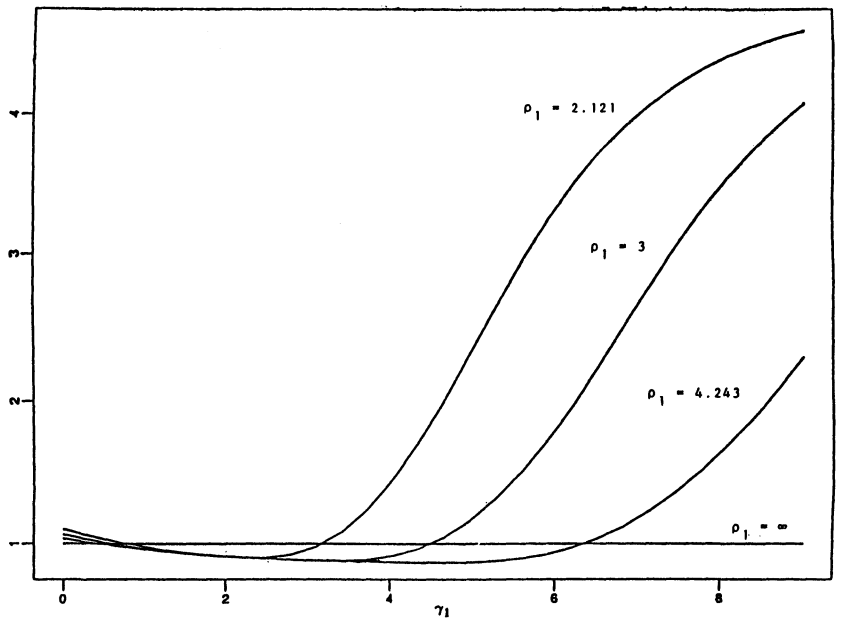

Figure 3: Asymptotic variance ratio of QEF/MEQL estimators as a function of γ_1; example with IG-2 distribution. Curves with increasing $\rho - 1$ correspond to increasing amounts of actual overdispersion. Abscissa is assumed amount of overdispersion.