Improving the Precision of Estimation by fitting a Generalized Linear Model, and Quasi-likelihood.

P.M.E.Altham, Statistical Laboratory, University of Cambridge

June 27, 2006

This article was published in the GLIM newsletter No 23, 1994 (ISSN 0269-0772). It is given on this web-page as I have now added some conjectures, and a little numerical example.

1. Introduction

Altham (1984) proved a result for maximum likelihood estimators, showing that one of the purposes of fitting a parsimonious model is to improve the precision of estimation of those parameters that remain. Here this result is extended to quasi-likelihood and generalized linear models.

2. Statement and Proof of Result

Altham (1984) proved the following result. Suppose that a random sample of size n has log-likelihood function $L_n(p)$, where p is a k-dimensional unknown parameter. Let ω be the hypothesis

$$\omega: \ p_i = p_i(\theta) \text{ for } i = 1, \cdots, k$$

where $p_i(\cdot)$ are known functions, θ is an f-dimensional parameter, and $f < k$. Suppose that $\hat{p}$ maximises $L_n(p)$, and p^* maximises $L_n(p)$ subject to $p \epsilon \omega$, and let $\phi(p)$ be an arbitrary function of p. Then, for large n, under ω,

$$var(\phi(p^*)) \leq var(\phi(\hat{p})). \tag{1}$$

This is essentially proved from a particular matrix inequality.

The result remains true when we recast it in terms of quasi-likelihood functions and the Generalized Linear Model: thus it is not necessary to know $L_n(p)$ exactly. Following the notation of Chapter 9 of McCullagh and Nelder

(1989), let Y be the n-dimensional response vector, with expectation μ and covariance matrix $\sigma^2 V$, where $V = V(\mu)$ is a matrix of known functions and σ^2 is unknown.
Assume $\mu = \mu(\beta)$, where β is a vector of dimension p. The the quasi-likelihood estimator for β is $\hat{\beta}$, where $\hat{\beta}$ is the solution of

$$U(\hat{\beta}) = 0,$$

and

$$U(\beta) = D^T V^{-1}(Y - \mu(\beta))/\sigma^2,$$

and D is the $n \times p$ matrix of derivatives of μ, thus

$$D_{ir} = \frac{\partial \mu_i}{\partial \beta_r},\ 1 \leq i \leq n,\ 1 \leq r \leq p.$$

We assume D is of rank p.
As shown in McCullagh and Nelder,

$$cov(\hat{\beta}) \simeq \sigma^2 (D^T V^{-1} D)^{-1}.$$

Now suppose that in fact $E(Y)$ can be represented by a simpler model than $\mu(\beta)$: specifically suppose that

$$E(Y) = \mu(\beta) \text{ and } \beta = \beta(\gamma),$$

wher γ is a q-dimensional vector, $q < p$, and $\beta(\gamma)$ is a known function. Then the quasi-likelihood estimating equations for γ must be

$$W(\hat{\gamma}) = 0$$

where

$$W(\gamma) = E^T V^{-1}(Y - \mu)/\sigma^2$$

and

$$E = (E_{is}) = (\frac{\partial \mu_i}{\partial \gamma_s}).$$

Hence

$$cov(\hat{\gamma}) \simeq (E^T V^{-1} E)^{-1}.$$

Let γ_0 be the true value of γ. Then, to a first approximation,

$$\beta(\hat{\gamma}) \simeq \beta(\gamma_0) + A(\gamma_0)(\hat{\gamma} - \gamma_0)$$

where

$$A = (A_{rs}) = (\frac{\partial \beta_r}{\partial \gamma_s}),$$

a $p \times q$ matrix, assumed to be of rank q.
Hence

$$cov(\beta(\hat{\gamma})) \simeq A\ cov(\hat{\gamma}) A^T.$$

We prove below that, if $\beta = \beta(\gamma)$, then

$$cov(\beta(\hat{\gamma})) \leq cov(\hat{\beta}) \tag{2}$$

i.e. that

$$\sigma^2 A (E^T V^{-1} E)^{-1} A^T \leq \sigma^2 (D^T V^{-1} D)^{-1}. \tag{3}$$

Now the chain rule for derivatives shows us that

$$E = DA$$

so we need to prove that

$$\sigma^2 A (A^T D^T V^{-1} DA)^{-1} A^T \leq \sigma^2 (D^T V^{-1} D)^{-1}. \tag{4}$$

Put $\Sigma^{-1} = D^T V^{-1} D$; this is a $p \times p$ positive definite matrix of full rank, and we see that (4) is equivalent to

$$A (A^T \Sigma^{-1} A)^{-1} A^T \leq \Sigma. \tag{5}$$

This inequality, which was of course also the key to the proof given in Altham (1984), is familiar from least squares theory.
Thus we have shown that approximately

$$cov(\beta(\hat{\gamma})) \leq cov(\hat{\beta}) \tag{6}$$

so that for any real-valued function $\phi(\cdot)$, the following inequality is approximately true:

$$var(\phi(\beta(\hat{\gamma})) \leq var(\phi(\hat{\beta})). \tag{7}$$

It appears that quasi-likelihood function is special in this respect. If we use a more general linear estimating function (see p347 of McCullagh and Nelder 1989) and have, say

$$H^T (Y - \mu(\hat{\beta})) = 0 \tag{8}$$

where H is an $n \times p$ matrix, then for $\beta = \beta(\gamma)$ the natural estimating equation for γ appears to be

$$(HA)^T(Y - \mu(\beta(\hat{\gamma}))) = 0, \tag{9}$$

with the matrix A defined above. However, it appears that we cannot expect the inequality (2) to hold for a general $n \times p$ matrix H.

3. The estimation of σ^2

Of course, the inequality (7) is one that is 'well known' to practical statisticians. The application is straightforward for the 'error functions' such as the binomial or the Poisson, where the scale parameter σ^2 is considered known. However, if σ^2 is not specified in advance the application is not entirely straightforward. This is because as we run through a sequence of dropping unnecessary parameters from a model, say in GLIM, we observe that the s.e.'s of the estimates of the parameters that remain decline satisfactorily, but of course our estimate of σ^2 also changes according to the model fitted. Thus in any standard generalized linear model package, if the scale parameter is unknown, we do not see a simple numerical example of the inequality (2), because the estimate of σ^2 used on the left-hand side of (3) (where the model $\beta = \beta(\gamma)$ is assumed) is in general smaller than the estimate of σ^2 used in the right-hand side of (3), provided that the model $\beta = \beta(\gamma)$ fits well. So the numerical estimates for the left- and right-hand sides of (3) actually enhance the inequality.

References

Altham, P.M.E. (1984) Improving the precision of estimation by fitting a model. *J. Roy. Statist. Soc, Ser, B* 46, 118-119.

McCullagh, P. and Nelder, J.A. (1989) *Generalized Linear models.* London: Chapman and Hall.

Queries/conjectures (June 2006)

1. Is the quasi-likelihood function special in this respect, ie in ensuring that the inequality (2) will always hold?
2. Is there a Bayesian formulation of this result?
3. Is there a formal link with the Akaike Information Criterion?
4. When I published this paper in 1994, I thought it obvious that we are assuming $n > p$. But now, of course, in these days of micro-array analysis, it may well be the case that $n << p$. Can we formulate and prove a version of the above inequality about precisions for this case of 'large p, small n'? To be able to do so simply for the case of constrained linear regression, eg

minimise $(Y - X\beta)^T(Y - X\beta)$ subject to $\beta^T\beta < s$
would seem a good way to start.
5. Here is an example, in R, illustrating the inequality about precision of estimates. I use the 'quine' dataset from Venables and Ripley, 'Modern Applied Statistics with S'. I'm using two models that are rather simple compared with the ones presented by Venables and Ripley: this is to give you a really straightforward application of the inequality in the context of quasilikelihood. (The fact that when I originally derived the inequality, it would be several **years** before I learnt how to illustrate it so easily with some actual data, is rather exciting to me, although I guess it was actually not so difficult in the 1980's to get GLIM to do this example.)
We follow the convention below in writing $\sigma^2 = \phi$ (and this quantity will be estimated from the fit of the model). The output has been slightly reduced.

```
>library(MASS)
>quine.qp = glm(Days ~ ., family = quasipoisson, data = quine)
>summary(quine.qp)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.7154     0.2347  11.569  < 2e-16 ***
EthN         -0.5336     0.1520  -3.511 0.000602 ***
SexM          0.1616     0.1543   1.047 0.296911
AgeF1        -0.3339     0.2543  -1.313 0.191410
AgeF2         0.2578     0.2265   1.138 0.256936
AgeF3         0.4277     0.2456   1.741 0.083830 .
LrnSL         0.3489     0.1888   1.848 0.066759 .

(Dispersion parameter for quasipoisson family taken to be 13.16677)

    Null deviance: 2073.5  on 145  degrees of freedom
Residual deviance: 1696.7  on 139  degrees of freedom

Number of Fisher Scoring iterations: 5
```

Thus we see that with the assumption $E(Y_i) = \mu_i, var(Y_i) = \phi\mu_i$, we find that $\hat{\phi} = 13.16677$. Here $\hat{\phi}$ is taken as X^2 divided by the df, as you can check (using the obvious quantities for y and e) from

```
> chisq= sum((y-e)*(y-e)/e); chisq/139
```

In this model the SexM term is not significant.
We therefore remove this from the model, to get the submodel

```
>quine.qp1 = glm(Days ~ . - Sex, family = quasipoisson, data = quine)
>summary(quine.qp1)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.8352     0.2021  14.025   <2e-16 ***
EthN         -0.5331     0.1518  -3.512   0.0006 ***
AgeF1        -0.3767     0.2511  -1.500   0.1359
AgeF2         0.2443     0.2271   1.076   0.2839
AgeF3         0.3800     0.2404   1.581   0.1162
LrnSL         0.3123     0.1862   1.677   0.0957 .

(Dispersion parameter for quasipoisson family taken to be 13.14278)

    Null deviance: 2073.5  on 145  degrees of freedom
Residual deviance: 1711.1  on 140  degrees of freedom

Number of Fisher Scoring iterations: 5
```

Observe that now $\hat{\phi} = 13.14278$, slightly less than before, and all the remaining parameter estimates have se's that are less than the corresponding se's given in the 'full' model.
You can see that I have used the standard (default in R) corner-point constraints for the factor parameters.
You might like to compare these results with what happens when you do

```
> anova(quine.qp1, quine.qp, test="F")
> summary(glm(Days ~ ., poisson, data = quine))
> summary(glm(Days ~ . - Sex, poisson, data = quine))
```