QUASI-LIKELIHOOD FUNCTIONS

By Peter McCullagh, 1983

Henrique Laureano (.github.io)
LEG @ UFPR

July 23, 2021

The Annals of Statistics
1983, Vol. 11, No. 1, 59–67

QUASI-LIKELIHOOD FUNCTIONS

BY PETER MCCULLAGH

Imperial College, London and University of British Columbia

The connection between quasi-likelihood functions, exponential family models and nonlinear weighted least squares is examined. Consistency and asymptotic normality of the parameter estimates are discussed under second moment assumptions. The parameter estimates are shown to satisfy a property of asymptotic optimality similar in spirit to, but more general than, the corresponding optimal property of Gauss-Markov estimators.

1. Distinguished Professor
   in the Department of Statistics @ University of Chicago;
2. Completed his PhD at Imperial College London,
   supervised by David Cox and Anthony Atkinson;
3. Also at Imperial College London,
   was the PhD supervisor of Gauss Cordeiro.

Introduction

1. Likelihood fucntion with **exponential family form**
   ↳ MLE through **weighted least squares**
   - **variance (assumed) constant**: we minimize a sum of squared residuals;
   - **variance not constant**:
     estimating equations can be thought as a generalization of the scoring method.
2. Likelihood function without exponential family form
   ↳ In some cases: weighted least squares
   ↳ Jorgensen, B. (1983). Maximum likelihood estimation and large sample inference for generalized linear and non-linear regression models. *Biometrika* 70

Paper purposes

1. Maximize the likelihood function through weighted least squares
   ↳ In which class of problems;
2. Weighted least squares under 2nd moment assumptions (**quasi-likelihood**).

A class of likelihood functions

Log likelihood written in the form : $\sigma^{-2}\{\mathbf{y}^\top\boldsymbol{\theta} - \mathbf{b}(\boldsymbol{\theta}) - \mathbf{c}(\mathbf{y}, \sigma)\}$

The first two cumulants

By differentiating it and assuming that the support does not depend on $\boldsymbol{\theta}$

↳ $E(\mathbf{Y}) = \boldsymbol{\mu} = \mathbf{b}'(\boldsymbol{\theta})$ and $\text{Cov}(\mathbf{Y}) = \sigma^2\mathbf{b}''(\boldsymbol{\theta}) = \sigma^2\mathbf{V}(\boldsymbol{\mu})$.

In fact, the rth order cumulants of $\mathbf{Y}$ are given by $\kappa_r = \sigma^{2r-2}\mathbf{b}^{(r)}(\boldsymbol{\theta})$.

1. The first two cumulants describe the random component of the model;
2. However, in applications it is usually the systematic or nonrandom variation that is of primary importance

   ↳ $E(\mathbf{Y}) = \boldsymbol{\mu} = \boldsymbol{\mu}(\beta)$ or $E\{(\mathbf{h}(\mathbf{Y}))\} = \psi(\beta)$ (implicitly involving σ^2).

A class of likelihood functions

If σ^2 is known, log-likelihood is an exponential family

↳ variance and all higher order cumulants of $\mathbf{Y}$ are functions of the mean vector alone

↳ exponential, Poisson, multinomial, noncentral hypergeometric and partial likelihoods (survival analysis)

↳ MLE of $\boldsymbol{\beta}$ through weighted least squares

↳ $\boldsymbol{\mu}$ and σ^2 are orthogonal

↳ $\boldsymbol{\beta}$ and σ^2 also orthogonal;

If σ^2 is unknown, log-likelihood is not generally an exponential family

↳ However, MLE of $\boldsymbol{\beta}$ still through weighted least squares

↳ If $E(\mathbf{Y})$ does not involve σ^2.

A class of likelihood functions

Least square equations

$$\mathbf{D}^\top \mathbf{V}^- \{\mathbf{y} - \boldsymbol{\mu}(\hat{\beta})\} = \mathbf{0}, \qquad \text{for the parameters in} \quad E(\mathbf{Y}) = \boldsymbol{\mu}(\beta)$$

- $\mathbf{D} = d\boldsymbol{\mu}/d\beta, \quad N \times p;$
- $\mathbf{V}^-$ is a generalized inverse of $\mathbf{V}$.

Why its name?

1. **Geometrical interpretation**: projections of the residual vector $\mathbf{y} - \boldsymbol{\mu}(\hat{\beta}_0)$ on to the tangent space of the solution locus $\boldsymbol{\mu}(\beta)$;
2. These equations **do not** depend on σ^2.

Newton-Raphson method

We replace the second derivative matrix by its expected value, $\mathbf{D}^\top \mathbf{V}^- \mathbf{D}$

$$\hat{\beta}_1 - \hat{\beta}_0 = (\mathbf{D}^\top \mathbf{V}^- \mathbf{D})^{-1} \mathbf{D}^\top \mathbf{V}^- (\mathbf{y} - \hat{\mu}_0).$$

Quasi-likelihood functions

Reversing the natural order of assumptions

1. Instead of taking the log-likelihood to be of the exponential family form and then deriving its moments;
2. We begin with the moments and then attempt to reconstruct the log-likelihood.

The reconstituted function is called a **quasi-likelihood**.

The log-quasi-likelihood, function of $\boldsymbol{\mu}$,
is given by the system of partial differential equations

$$\frac{\partial \ell(\boldsymbol{\mu}; \mathbf{y})}{\partial \boldsymbol{\mu}} = \mathbf{V}^{-}(\boldsymbol{\mu})(\mathbf{y} - \boldsymbol{\mu}).$$

Which extends **Wedderburn's (1974)** definition.

1. We get $\hat{\beta}$ from $\mathbf{D}^{\top}\mathbf{V}^{-}\{\mathbf{y} - \boldsymbol{\mu}(\hat{\beta})\} = \mathbf{0}$ (**generalized least squares equations**);
2. There is no guarantee that $\hat{\beta}$ is the MLE.

Properties of quasi-likelihood functions

Risk model

Latent effects only on the risk level i.e.,

$$\Sigma = \begin{bmatrix} \sigma^2_{u_1} & \mathrm{cov}_{u_1,u_2} \\ & \sigma^2_{u_2} \end{bmatrix}.$$

Time model

Latent effects only on the failure time trajectory level i.e.,

$$\Sigma = \begin{bmatrix} \sigma^2_{\eta_1} & \mathrm{cov}_{\eta_1,\eta_2} \\ & \sigma^2_{\eta_2} \end{bmatrix}.$$

Block-diag model

Latent effects on the risk and time levels without cross-correlations i.e.,

$$\Sigma = \begin{bmatrix} \sigma^2_{u_1} & \mathrm{cov}_{u_1,u_2} & 0 & 0 \\ & \sigma^2_{u_2} & 0 & 0 \\ & & \sigma^2_{\eta_1} & \mathrm{cov}_{\eta_1,\eta_2} \\ & & & \sigma^2_{\eta_2} \end{bmatrix}.$$

Complete model

A *complete* latent effects structure i.e.,

$$\Sigma = \begin{bmatrix} \sigma^2_{u_1} & \mathrm{cov}_{u_1,u_2} & \mathrm{cov}_{u_1,\eta_1} & \mathrm{cov}_{u_1,\eta_2} \\ & \sigma^2_{u_2} & \mathrm{cov}_{u_2,\eta_1} & \mathrm{cov}_{u_2,\eta_2} \\ & & \sigma^2_{\eta_1} & \mathrm{cov}_{\eta_1,\eta_2} \\ & & & \sigma^2_{\eta_2} \end{bmatrix}.$$

Simulation study setup

Four latent effects structures:

1. Risk model;
2. Time model;
3. Block-diag model;
4. Complete model.

Two CIF configurations:

Low max incidence ≈ 0.15; **High** max incidence ≈ 0.60.

For each of those $4 \times 2 = 8$ scenarios, we vary the sample and cluster sizes:

5000 data points	*30000 data points*	*60000 data points*
• 2500 clusters of **size 2**;	• 15000 clusters of **size 2**;	• 30000 clusters of **size 2**;
• 1000 clusters of **size 5**;	• 6000 clusters of **size 5**;	• 12000 clusters of **size 5**;
• 500 clusters of **size 10**.	• 3000 clusters of **size 10**.	• 6000 clusters of **size 10**.

Totalizing, $8 \times 3 \times 3 = 72$ scenarios.
For each scenario, we simulate **500** samples, totalizing $72 \times 500 = 36000$ model fittings.

Simulation study results

First of all, the **time**.

- The *non-complete* models (2D Laplace aprox.) are kind of fast, taking always **less than 5 min**.
- In the most expensive scenarios (30K 4D Laplaces), **the complete model takes 30 min**.
  In a **full R** implementation with 10K 4D Laplaces, it **took 30hrs**. **TMB is fast**.
- We also did a Bayesian analysis via Stan/NUTS-HMC [@RStan].
  - **1 week of parallelized processing** for a 2500 size 2 clusters scenario with tuned NUTS.
    This just reinforces the MCMC impracticability for some complex models.

Parameters estimation.

- The *non-complete* models fail to learn the data.
  They appear to be *not structured enough* to capture the data characteristics.

Take-home message

The complete model works. It's not magnificent, but it works.

1. It works better in the high CIF scenarios;
2. As expected, as the sample size increases the results get better;
3. We do not see any considerable performance difference between cluster/family sizes;
4. Satisfactory full likelihood analysis under the maximum likelihood estimation framework (the estimates bias-variance could be smaller).

What else can we do?

1. Instead of a conditional approach (latent effects model), we can try a marginal approach e.g., an McGLM [@mcglm];
2. We can also try a copula [@copulas], on maybe two fronts: 1) for a full specification; 2) to accommodate the within-cluster dependence.

For more read @laurence master thesis.

Thanks for watching and have a great day

Special thanks to