# 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 function 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 $\mathrm{Cov}(\mathbf{Y}) = \sigma^2 \mathbf{b}''(\boldsymbol{\theta}) = \sigma^2 \mathbf{V}(\boldsymbol{\mu})$.

**In fact,** the $r$th 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
   ↳ $E(\mathbf{Y}) = \boldsymbol{\mu} = \boldsymbol{\mu}(\boldsymbol{\beta})$ or $E\{(\mathbf{h}(\mathbf{Y}))\} = \boldsymbol{\psi}(\boldsymbol{\beta})$ (implicitly involving $\sigma^2$).

# A class of likelihood functions

**If $\sigma^2$ 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 $\beta$ through weighted least squares

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

↳ $\beta$ and $\sigma^2$ also orthogonal;

**If $\sigma^2$ is unknown,** log-likelihood is not generally an exponential family

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

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

# 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 $\sigma^2$.

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

# Properties of quasi-likelihood functions

*Very similar to those of ordinary likelihoods except that the nuisance parameter, $\sigma^2$, when unknown, is treated separately from $\beta$ and is not estimated by weighted least squares.*

The principal results fall into three classes:

1. Those concerning the score function $\mathbf{U}_\beta = \partial\ell/\partial\beta$;

$$\mathbf{U}_\beta = \mathbf{D}^\top \mathbf{V}^- (\mathbf{Y} - \boldsymbol{\mu}(\beta)) \quad \text{has zero mean and covariance matrix} \quad \sigma^2 \mathbf{i}_\beta = \sigma^2 \mathbf{D}^\top \mathbf{V}^- \mathbf{D}$$

   where $\mathbf{i}_\beta$ is the expected second derivative matrix of $\ell(\boldsymbol{\mu}(\beta); \mathbf{Y})$.

2. Those concerning the estimator $\hat{\beta}$;

$$\text{There exists a } \hat{\beta} \text{ satisfying } \hat{\beta} - \beta = \mathbf{I}_{\beta^*}^{-1} \mathbf{U}_\beta \quad \text{(minimum asymptotic variance)}$$

   where $\mathbf{I}_\beta$ is the observed matrix of second derivatives
   and $\beta^*$ is a point lying on the line segment joining $\hat{\beta}$ and $\beta$.

# Properties of quasi-likelihood functions

3. Those concerning the distribution of the quasi-likelihood-ratio statistic.

$$2\ell(\hat{\beta};\mathbf{Y}) - 2\ell(\beta;\mathbf{Y}) = \mathbf{U}_\beta^\top \mathbf{i}_\beta^{-1}\mathbf{U}_\beta + O_p(N^{-1/2}) \quad \text{is asymptotically} \quad \sigma^2\chi^2_p$$

The **asymptotic optimality** follows the same line as the optimal property of Gauss-Markov estimators.

The conclusions (asymptotic unbiasedness),

- while they apply more widely than the Gauss-Markov theorem, are inevitably a little weaker being asymptotic rather than exact.

# Estimation of $\sigma^2$

In the absense of information beyond the second moments:

$$\tilde{\sigma}^2 = \frac{(\mathbf{y} - \hat{\mu})^\top \mathbf{V}^- (\mathbf{y} - \hat{\mu})}{N - p} = \frac{X^2}{N - p}$$

where $X^2$ is a generalized form of Pearson's statistic.

If the log-likelihood is in the exponential family,
an estimate can be obtained by equating the observed deviance

$$d(\mathbf{y}; \hat{\mu}) = 2\ell(\mathbf{y}; \mathbf{y}) - 2\ell(\hat{\mu}; \mathbf{y}), \text{ to its approximate expected value.}$$

**Advantage:** asymptotically independent of $\hat{\beta}$;

**Disadvantage:** the expectation of $d(\mathbf{Y}; \hat{\mu})$ is often difficult to compute.

# Examples of quasi-likelihood functions

**Least squares.** $\mathbf{V}(\mu) = \mathbf{V}$;

**Uncorrelated observations.** $\mathbf{V}(\mu) = \text{diag}\{\mathbf{v}(\mu_1), \mathbf{v}(\mu_2), \ldots, \mathbf{v}(\mu_N)\}$;

**Invariance under linear transformations.**

$$\mathbf{Y}_L = \mathbf{L}\mathbf{Y}, \quad \mu_L = \mathbf{L}\mu, \quad \mathbf{V}_L = \mathbf{L}\mathbf{V}\mathbf{L}^\top,$$

with $\mathbf{L}$ being a nonsingular matrix of order $N$.

- Weaker than the corresponding result for log-likelihoods which are invariant under all invertible transformations;

**Multinomial response models.** Logistic regression, e.g.;

**Models with constant coefficient of variation.**

**Over-dispersion.** Quasi-likelihood for over-dispersed discrete observations.

# A higher order theory

- Basically, there are problems if the exponential family is curved, conditioning on ancillary statistics.

**Acknowledgments.** McCullagh says thanks to Professor D. R. Cox and **Dr. B. Jorgensen** (patrão's PhD advisor, nice!).

## THANKS FOR WATCHING AND HAVE A GREAT DAY


henriquelaureano.github.io