Vector Differential Calculus in Statistics

M. P. Wand

Many statistical operations benefit from differential calculus. Examples include optimization of likelihood functions and calculation of information matrices. For multiparameter models differential calculus suited to vector argument functions is usually the most efficient means of performing the required calculations. We present a primer on vector differential calculus and demonstrate its application to statistics through several worked examples.

KEY WORDS: Best linear prediction; Generalized linear model; Generalized linear mixed model; Information matrix; Matrix differential calculus; Maximum likelihood estimation; Penalized quasi-likelihood; Score equation.

1. INTRODUCTION

Matrix notation is an indispensable tool in statistics. In particular, the storage of model parameters in a vector allows for economical expression of models that relate the parameters to the data. Functions used for estimation of the parameter vector, such as likelihood functions and posterior densities, are also conveniently expressed using matrix notation. But when it comes to optimization of these functions through differential calculus the statistician usually reverts to scalar subscript notation and applies ordinary univariate calculus to the entries of the parameter vector. Subsequent calculus steps, such as those required to obtain the information matrix, are also usually done with an element-wise approach—sometimes followed by conversion back to matrix notation.

This article demonstrates that it is often possible to perform the entire operation using matrix algebra. Apart from being more streamlined, it saves conversion between matrix notation and subscript notation. The key is a differential calculus suited to vector argument and scalar-valued functions.

The book by Magnus and Neudecker (1988) describes an elegant approach to differential calculus in statistics for general *matrix*-argument functions. However, the simplifications that arise for scalar-valued *vector*-argument functions are somewhat opaque in that reference. This article aims to redress this deficiency, while otherwise using the tools of Magnus and Neudecker. Other contributions to matrix differential calculus may be found in Dwyer (1967), Searle (1982), Basilevsky (1983), and Harville (1997) but each of these references have similar shortcomings for vector-argument functions.

Section 2 begins with a simple illustrative example. Section 3 presents the fundamentals of vector differential calculus, and Section 4 illustrates their use in statistics. Section 5 describes an extension of the methodology to matrix arguments.

2. SIMPLE ILLUSTRATIVE EXAMPLE

Let (x_i, y_i), $1 \le i \le n$, be a set of measurements on two variables x and y, and consider the problem of fitting a line $y = \beta_0 + \beta_1\, x$ to the data. The homoscedastic Gaussian regression model is

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I}),$$

where

$$\mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \quad \text{and} \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}.$$

For σ^2 known, the log-likelihood for estimation of $\boldsymbol{\beta}$ is (up to an additive constant)

$$\ell(\boldsymbol{\beta}) = -\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}). \tag{1}$$

The scalar differential calculus approach to maximization of $\ell(\boldsymbol{\beta})$ involves rewriting (1) as

$$\ell(\boldsymbol{\beta}) = -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1\, x_i)^2$$

and then obtaining

$$\begin{aligned} \tfrac{\partial \ell(\boldsymbol{\beta})}{\partial \beta_0} &= (1/\sigma^2)\textstyle\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1\, x_i), \\ \tfrac{\partial \ell(\boldsymbol{\beta})}{\partial \beta_1} &= (1/\sigma^2)\textstyle\sum_{i=1}^{n} x_i(y_i - \beta_0 - \beta_1\, x_i). \end{aligned} \tag{2}$$

Setting this to zero we obtain the unique stationary point of $\ell(\boldsymbol{\beta})$ occurring at

$$\begin{aligned} \widehat{\beta}_0 &= \overline{y} - \widehat{\beta}_1\overline{x} \\ \widehat{\beta}_1 &= \sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) \Big/ \sum_{i=1}^{n}(x_i - \overline{x})^2. \end{aligned}$$

As is well-known, (2) is algebraically equivalent to

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}. \tag{3}$$

The Hessian matrix of second-order partial derivatives is

$$\begin{aligned} &\begin{bmatrix} \frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \beta_0^2} & \frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \beta_0 \partial \beta_1} \\ \frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \beta_0 \partial \beta_1} & \frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \beta_1^2} \end{bmatrix} \\ &\quad = -(1/\sigma^2)\begin{bmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{bmatrix} = -(1/\sigma^2)\mathbf{X}^\top\mathbf{X} \end{aligned}$$

which can be shown to be negative definite. Therefore $\widehat{\boldsymbol{\beta}} = [\widehat{\beta}_0\ \widehat{\beta}_1]^\top$ as defined in (2) is the maximizer of $\ell(\boldsymbol{\beta})$ and returns the least squares line. The information matrix of $\boldsymbol{\beta}$ is

$$I(\boldsymbol{\beta}) = -E\{-(1/\sigma^2)\mathbf{X}^\top\mathbf{X}\} = (1/\sigma^2)\mathbf{X}^\top\mathbf{X}$$

M. P. Wand is Associate Professor, Department of Biostatistics, School of Public Health, Harvard University, 665 Huntington Avenue, Boston, MA 02115 (E-mail: mwand@hsph.harvard.edu). I am grateful to Mahlet Tadesse, Misha Salganik, Bhaswati Ganguli, and Yannis Jemiai for comments on the article.

so the covariance matrix of $\widehat{\boldsymbol{\beta}}$ is

$$I(\boldsymbol{\beta})^{-1} = \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}.$$

Vector differential calculus achieves the same result more directly:

$$\begin{aligned}
\ell(\boldsymbol{\beta}) &= -\frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^\top(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) \\
\Longrightarrow \quad d\,\ell(\boldsymbol{\beta}) &= -\frac{1}{2\sigma^2}[\{d\,(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\}^\top(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) \\
&\quad +(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^\top\{d\,(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\}] \\
&= -\frac{1}{2\sigma^2}[-(\mathbf{X}\,d\,\boldsymbol{\beta})^\top(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) \\
&\quad -(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^\top\mathbf{X}d\boldsymbol{\beta}] \\
&= (1/\sigma^2)(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^\top\mathbf{X}\,d\boldsymbol{\beta} \\
\Longrightarrow \quad \mathsf{D}\ell(\boldsymbol{\beta}) &= (1/\sigma^2)(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^\top\mathbf{X} = \mathbf{0} \\
&\text{iff} \quad (\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^\top\mathbf{X} = \mathbf{0} \\
\Longrightarrow \quad \boldsymbol{\beta} &= (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}.
\end{aligned}$$

Also,

$$d^2\,\ell(\boldsymbol{\beta}) = (d\boldsymbol{\beta})^\top\left(-\frac{1}{\sigma^2}\mathbf{X}^\top\mathbf{X}\right)(d\,\boldsymbol{\beta})$$

so

$$\mathsf{H}\,\ell(\boldsymbol{\beta}) = -\frac{1}{\sigma^2}\mathbf{X}^\top\mathbf{X} \quad \Longrightarrow \quad I(\boldsymbol{\beta})^{-1} = \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}.$$

The symbols "$\mathsf{D}\,\ell(\boldsymbol{\beta})$" and "$\mathsf{H}\,\ell(\boldsymbol{\beta})$" may be unfamiliar to some readers, and even "$d\boldsymbol{\beta}$" may require clarification. The latter is a vector of differentials, analogous to the numerator and denominator of dy/dx. In vector differential calculus these differentials are often vectors, so cannot be divided according to the usual rules of algebra. The D and H notation is used throughout Magnus and Neudecker (1988) and, for scalar-valued functions of vectors, is defined by the following

Definitions: Let f be a scalar-valued function with argument $\mathbf{x} \in \mathbb{R}^p$. The *derivative vector* of f, $\mathsf{D}f(\mathbf{x})$, is the $1 \times p$ vector whose ith entry is

$$\frac{\partial f(\mathbf{x})}{\partial x_i}$$

The *Hessian matrix* of f is

$$\mathsf{H}\,f(\mathbf{x}) = \mathsf{D}\{\mathsf{D}f(\mathbf{x})^\top\}$$

and is, alternatively, the $p \times p$ matrix with (i, j) entry equal to

$$\frac{\partial^2 f(\mathbf{x})}{\partial x_i \partial x_j}.$$

Many of our examples are concerned with score and information calculations. If the data vector $\mathbf{y}$ has density $f(\mathbf{y};\boldsymbol{\beta})$ then the score vector and information matrices are, respectively,

$$\mathsf{D}_{\boldsymbol{\theta}} \log f(\mathbf{y};\boldsymbol{\theta}) \quad \text{and} \quad I(\boldsymbol{\beta}) \equiv -E\{\mathsf{H}_{\boldsymbol{\theta}} \log f(\mathbf{y};\boldsymbol{\theta})\}.$$

Note the necessity of a subscript on the D and H in this instance to specify the argument of $\log f(\mathbf{y};\boldsymbol{\theta})$ being differentiated. We use such subscript notation liberally in the examples of Section 4.

3. A PRIMER ON VECTOR DIFFERENTIAL CALCULUS

Consider $x \in \mathbb{R}$ and the function

$$y = x^3 \sin x.$$

Then scalar differential calculus leads to

$$\begin{aligned}
\frac{dy}{dx} &= \left(\frac{d}{dx}x^3\right)\sin x + x^3\left(\frac{d}{dx}\sin x\right) \\
&= 3x^2 \sin x + x^3 \cos x.
\end{aligned}$$

An alternative piece of calculus is

$$\begin{aligned}
dy &= d(x^3 \sin x) \\
&= (dx^3)\sin x + x^3(d\sin x) \\
&= (3x^2\,dx)\sin x + x^3(\cos x\,dx) \\
&= (3x^2 \sin x + x^3\cos x)dx.
\end{aligned}$$

Hence

$$\frac{dy}{dx} = 3x^2 \sin x + x^3 \cos x$$

as before. This second derivation has used the rules:

1. $d(uv) = (du)v + u(dv)$
2. $du^3 = 3u^2\,du$
3. $d\sin u = \cos u\,du.$

Each of these rules seems reasonable given the product rule and rules for differentiation of polynomials and trigonometric functions. The only difference is that they do not involve division by differentials such as du. In scalar differential calculus such division is "allowable," but in the vector world division is not standard. Therefore, for vector differential calculus, it is more appropriate to work with products and then compute the derivative vector using (Magnus and Neudecker 1988):

First Identification Theorem: If $\mathbf{a}$ *is a* $1 \times p$ *vector for which*

$$d\,f(\mathbf{x}) = \mathbf{a}\,d\,\mathbf{x},$$

then

$$\mathbf{a} = \mathsf{D}f(\mathbf{x}).$$

The Hessian matrix can be found from

Second Identification Theorem: If $\mathbf{A}$ *is a* $p \times p$ *matrix for which*

$$d^2\,f(\mathbf{x}) = (d\,\mathbf{x})^\top\mathbf{A}\,d\,\mathbf{x},$$

then

$$\mathbf{A} = \mathsf{H}f(\mathbf{x}).$$

It follows from these theorems that all that is required are rules for expressing $df(\mathbf{x})$ in the forms

$$\mathbf{a}\,d\mathbf{x} \quad \text{and} \quad (d\mathbf{x})^\top \mathbf{A}\,d\mathbf{x}.$$

3.1 Notation

For vectors

$$\mathbf{a} = \begin{bmatrix} a_1 \\ \vdots \\ a_p \end{bmatrix} \qquad \text{and} \qquad \mathbf{b} = \begin{bmatrix} b_1 \\ \vdots \\ b_p \end{bmatrix}$$

we will introduce the notation (used by modern programming languages such as S-Plus and Matlab)

$$\mathbf{a} \odot \mathbf{b} = \begin{bmatrix} a_1 b_1 \\ \vdots \\ a_p b_p \end{bmatrix},$$

$$\mathbf{a}/\mathbf{b} = \begin{bmatrix} a_1/b_1 \\ \vdots \\ a_p/b_p \end{bmatrix},$$

and

$$s(\mathbf{a}) = \begin{bmatrix} s(a_1) \\ \vdots \\ s(a_p) \end{bmatrix}.$$

where $s : \mathbb{R} \to \mathbb{R}$ is a scalar function. We will use $\mathbf{1}$ to denote a vector of ones, with dimension clear from the context. Another useful notation is

$$\text{diag}(\mathbf{a}) = \begin{bmatrix} a_1 & 0 & \cdots & 0 \\ 0 & a_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & a_p \end{bmatrix}.$$

For a $p \times q$ matrix $\mathbf{A}$ we define vec$(\mathbf{A})$ to be the $pq \times 1$ vector obtained by stacking the columns of $\mathbf{A}$ underneath each other, in order from left to right.

If $\mathbf{v}$ is a random vector then we use $E(\mathbf{v})$ to denote is expectation and cov$(\mathbf{v})$ to denote its covariance matrix.

3.2 Some Matrix Algebraic Rules

The following matrix algebraic relationships are useful in vector differential calculus. The first one is particularly crucial.

$$\begin{aligned} \text{diag}(\mathbf{a})\mathbf{b} &= \mathbf{a} \odot \mathbf{b}, \\ \text{diag}(\mathbf{a})\mathbf{1} &= \mathbf{a}, \\ \text{tr}(\mathbf{AB}) &= \text{tr}(\mathbf{BA}), \\ \text{tr}(\mathbf{A}^\top\mathbf{B}) &= \text{vec}(\mathbf{A})^\top \text{vec}(\mathbf{B}). \end{aligned}$$

3.3 Rules for Differentials

Let $\mathbf{u}$ and $\mathbf{v}$ be vector functions and $\mathbf{U}$ and $\mathbf{V}$ be matrix functions. $\mathbf{A}$ will denote a constant matrix and s a scalar function.

3.3.1 Rules for Scalar Functions

$$\begin{aligned} du^\alpha &= \alpha u^{\alpha-1}\,du, \\ d\log u &= u^{-1}du, \\ de^u &= e^u du. \end{aligned}$$

3.3.2 Chain Rule

$$d\{s(\mathbf{u})\} = s'(\mathbf{u}) \odot (d\mathbf{u}) = \text{diag}\{s'(\mathbf{u})\}d\mathbf{u}.$$

3.3.3 Rules Involving Linear Functions

$$\begin{aligned} d(\mathbf{AU}) &= \mathbf{A}d\mathbf{U}, \\ d(\mathbf{U}+\mathbf{V}) &= d\mathbf{U} + d\mathbf{V}, \\ d\,\text{diag}(\mathbf{u}) &= \text{diag}(d\,\mathbf{u}), \\ d\mathbf{U}^\top &= (d\mathbf{U})^\top, \\ d\text{vec}\,\mathbf{U} &= \text{vec}(d\mathbf{U}), \\ d(\text{tr}\mathbf{U}) &= \text{tr}(d\mathbf{U}), \\ d(E\mathbf{U}) &= E(d\mathbf{U}). \end{aligned}$$

3.3.4 Product and Quotient Rules

$$\begin{aligned} d(\mathbf{U} \odot \mathbf{V}) &= (d\mathbf{U}) \odot \mathbf{V} + \mathbf{U} \odot (d\mathbf{V}), \\ d(\mathbf{UV}) &= (d\mathbf{U})\mathbf{V} + \mathbf{U}(d\mathbf{V}), \\ d(\mathbf{u}/\mathbf{v}) &= \frac{(d\mathbf{u}) \odot \mathbf{v} - \mathbf{u} \odot (d\mathbf{v})}{\mathbf{v} \odot \mathbf{v}}. \end{aligned}$$

3.3.5 Rules for Determinant and Matrix Inverse

$$\begin{aligned} d|\mathbf{U}| &= |\mathbf{U}|\text{tr}(\mathbf{U}^{-1}d\mathbf{U}), \\ d\mathbf{U}^{-1} &= -\mathbf{U}^{-1}(d\mathbf{U})\mathbf{U}^{-1}. \end{aligned}$$

Quadratic forms, particularly those involving symmetric matrices, are very common in statistics, so the following results are worth learning:

3.3.6 Rules Involving Quadratic Forms

$$\begin{aligned} d\mathbf{u}^\top\mathbf{A}\mathbf{u} &= \mathbf{u}^\top(\mathbf{A}+\mathbf{A}^\top)d\mathbf{u}, \\ d\mathbf{u}^\top\mathbf{A}\mathbf{u} &= 2\mathbf{u}^\top\mathbf{A}d\mathbf{u}, \quad \mathbf{A} \text{ symmetric}. \end{aligned}$$

4. EXAMPLES

4.1 Generalized Linear Models

Let $\mathbf{y}$ be a vector of responses and $\mathbf{X}$ be a corresponding design matrix. The one-parameter exponential family model, with canonical link, is characterized by the joint density

$$f(\mathbf{y};\boldsymbol{\beta}) = \exp\{\mathbf{y}^\top(\mathbf{X}\boldsymbol{\beta}) - \mathbf{1}^\top b(\mathbf{X}\boldsymbol{\beta}) + \mathbf{1}^\top c(\mathbf{y})\}$$

where $\boldsymbol{\beta}$ is the vector of coefficients (e.g., McCullagh and Nelder 1990). For example, $b(x) = \log(1+e^x)$ corresponds to binary regression with a logit link function.

The log-likelihood of $\boldsymbol{\beta}$ is

$$\ell(\boldsymbol{\beta}) = \mathbf{y}^\top\mathbf{X}\boldsymbol{\beta} - \mathbf{1}^\top b(\mathbf{X}\boldsymbol{\beta}) + \mathbf{1}^\top c(\mathbf{y})$$

$$\begin{aligned} \Longrightarrow \quad d\ell(\boldsymbol{\beta}) &= \mathbf{y}^\top\mathbf{X}\, d\boldsymbol{\beta} - \mathbf{1}^\top\, d\, b(\mathbf{X}\boldsymbol{\beta}) \\ &= \mathbf{y}^\top\mathbf{X}\, d\boldsymbol{\beta} - \mathbf{1}^\top \text{diag}\{b'(\mathbf{X}\boldsymbol{\beta})\} d(\mathbf{X}\boldsymbol{\beta}) \\ &= \mathbf{y}^\top\mathbf{X}\, d\boldsymbol{\beta} - b'(\mathbf{X}\boldsymbol{\beta})^\top\mathbf{X}\, d\,\boldsymbol{\beta} \\ &= \{\mathbf{y} - b'(\mathbf{X}\boldsymbol{\beta})\}^\top\mathbf{X} d\,\boldsymbol{\beta}. \end{aligned}$$

From the first identification theorem the score is

$$\mathsf{D}\,\ell(\boldsymbol{\beta}) = \{\mathbf{y} - b'(\mathbf{X}\boldsymbol{\beta})\}^\top\mathbf{X}.$$

The information matrix of $\boldsymbol{\beta}$ is obtained from

$$\begin{aligned} d^2\ell(\boldsymbol{\beta}) &= d\{\mathbf{y} - b'(\mathbf{X}\boldsymbol{\beta})\}^\top\mathbf{X}\, d\,\boldsymbol{\beta} \\ &= -\{\text{diag}\{b''(\mathbf{X}\boldsymbol{\beta})\}\mathbf{X} d\,\boldsymbol{\beta}\}^\top\mathbf{X}\, d\,\boldsymbol{\beta} \\ &= (d\,\boldsymbol{\beta})^\top\mathbf{X}^\top[-\text{diag}\{b''(\mathbf{X}\boldsymbol{\beta})\}]\mathbf{X}(d\,\boldsymbol{\beta}) \end{aligned}$$

which leads to

$$\mathsf{H}\,\ell(\boldsymbol{\beta}) = -\mathbf{X}^\top\text{diag}\{b''(\mathbf{X}\boldsymbol{\beta})\}\mathbf{X}$$

and the information matrix being

$$\begin{aligned} I(\boldsymbol{\beta}) &= -E\{\mathsf{H}\,\ell(\boldsymbol{\beta})\} \\ &= \mathbf{X}^\top\text{diag}\{b''(\mathbf{X}\boldsymbol{\beta})\}\mathbf{X} = \mathbf{X}^\top\text{cov}(\mathbf{y})\mathbf{X}. \end{aligned}$$

Therefore, the asymptotic covariance matrix of $\widehat{\boldsymbol{\beta}}$ is

$$I(\boldsymbol{\beta})^{-1} = [\mathbf{X}^\top\text{diag}\{b''(\mathbf{X}\boldsymbol{\beta})\}\mathbf{X}]^{-1} = [\mathbf{X}^\top\text{cov}(\mathbf{y})\mathbf{X}]^{-1}.$$

4.2 Kriging

Let $(\mathbf{x}_i, y_i)$ be a set of observations with $\mathbf{x}_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$. The simple kriging model for estimation of $E(y|\mathbf{x}_0)$, $\mathbf{x}_0 \in \mathbb{R}^d$, is

$$y_i = \mu + S(\mathbf{x}_i) + \varepsilon_i$$

where the random vectors

$$\mathbf{S} = \begin{bmatrix} S(\mathbf{x}_1) \\ \vdots \\ S(\mathbf{x}_n) \end{bmatrix} \quad \text{and} \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

have the moment structure

$$E(\mathbf{S}) = E(\boldsymbol{\varepsilon}) = \mathbf{0}, \quad \text{cov}(\mathbf{S}) = \mathbf{C} \quad \text{and} \quad \text{cov}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}$$

(e.g., Cressie 1993). Kriging involves determination of the *best linear predictor*, $\widehat{\mathbf{S}}(\mathbf{x}_0)$, of $\mathbf{S}(\mathbf{x}_0)$ in the sense that

$$E[\{\widehat{\mathbf{S}}(\mathbf{x}_0) - \mathbf{S}(\mathbf{x}_0)\}^2]$$

is minimized among all linear combinations of the form

$$\widehat{\mathbf{S}}(\mathbf{x}_0) = \mathbf{a}^\top\mathbf{y} + b.$$

The function to be minimized over $\mathbf{a} \in \mathbb{R}^d$ and $b \in \mathbb{R}$ is then

$$\begin{aligned} \mathcal{C}(\mathbf{a}, b) &= E[\{\mathbf{a}^\top\mathbf{y} + b - S(\mathbf{x}_0)\}^2] \\ &= (\mu\mathbf{1}^\top\mathbf{a} + b)^2 + \mathbf{a}^\top(\mathbf{C} + \sigma^2\mathbf{I})\mathbf{a} \\ &\quad -2\mathbf{c}_0^\top\mathbf{a} + E\{S(\mathbf{x}_0)^2\}, \end{aligned}$$

where

$$\mathbf{c}_0 = [\text{cov}\{S(\mathbf{x}_0), S(\mathbf{x}_1)\}, \ldots \text{cov}\{S(\mathbf{x}_0), S(\mathbf{x}_n)\}]^\top.$$

Therefore

$$d_{\mathbf{a}}\mathcal{C}(\mathbf{a}, b) = 2\{(\mu\mathbf{1}^\top\mathbf{a} + b)\mu\mathbf{1}^\top + \mathbf{a}^\top(\mathbf{C} + \sigma^2\mathbf{I}) - \mathbf{c}_0^\top\}d\,\mathbf{a}$$

and

$$d^2_{\mathbf{a}}\mathcal{C}(\mathbf{a}, b) = 2(d\mathbf{a})^\top(\mu^2\mathbf{1}\mathbf{1}^\top + \mathbf{C} + \sigma^2\mathbf{I})d\mathbf{a}$$

leading to

$$\mathsf{D}_{\mathbf{a}}\mathcal{C}(\mathbf{a}, b) = 2\{(\mu\mathbf{1}^\top\mathbf{a} + b)\mu\mathbf{1}^\top + \mathbf{a}^\top(\mathbf{C} + \sigma^2\mathbf{I}) - \mathbf{c}_0^\top\}, \quad (4)$$

and

$$\mathsf{H}_{\mathbf{a}}\mathcal{C}(\mathbf{a}, b) = 2(\mu^2\mathbf{1}\mathbf{1}^\top + \mathbf{C} + \sigma^2\mathbf{I}) > \mathbf{0} \quad \text{for all} \quad b \in \mathbb{R}. \quad (5)$$

Since

$$\frac{d}{db}\mathcal{C}(\mathbf{a}, b) = 2(\mu\mathbf{1}^\top\mathbf{a} + b),$$

and

$$\frac{d^2}{db^2}\mathcal{C}(\mathbf{a}, b) = 2 > 0 \quad \text{for all} \quad \mathbf{a} \in \mathbb{R}^d$$

it follows from (4) and (5) that the unique minimizers of $\mathcal{C}(\mathbf{a}, b)$ are

$$\begin{aligned} b &= -\mu\mathbf{1}^\top\mathbf{a} \\ \mathbf{a} &= (\mathbf{C} + \sigma^2\mathbf{I})^{-1}\mathbf{c}_0 \end{aligned}$$

and the best linear predictor of $\mathbf{y}$ at $\mathbf{x}_0$ is

$$\mu + \mathbf{c}_0^\top(\mathbf{C} + \sigma^2\mathbf{I})^{-1}(\mathbf{y} - \mu\mathbf{1}).$$

In practice, μ and σ^2 are replaced by estimators $\widehat{\mu}$ and $\widehat{\sigma}^2$, and $\mathbf{C}$ parameterized and estimated by $\widehat{\mathbf{C}}$, leading to the kriging formula

$$\widehat{y}(\mathbf{x}_0) = \widehat{\mu} + \mathbf{c}_0^\top(\mathbf{C} + \widehat{\sigma}^2\mathbf{I})^{-1}(\mathbf{y} - \widehat{\mu}\mathbf{1}).$$

4.3 Variance Regression Models

The general Gaussian variance regression model is

$$\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\beta}, \mathbf{V}_{\boldsymbol{\theta}})$$

where $\mathbf{V}_{\boldsymbol{\theta}}$ represents a parameterization of the covariance matrix of $\mathbf{y}$ in terms of $\boldsymbol{\theta}$. An important special case is the Gaussian mixed model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon}$$

with the random effects vector $\mathbf{u}$ and the error $\boldsymbol{\varepsilon}$ having joint distribution

$$\begin{bmatrix} \mathbf{u} \\ \boldsymbol{\varepsilon} \end{bmatrix} \sim N\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \mathbf{G}_{\boldsymbol{\zeta}} & \mathbf{0} \\ \mathbf{0} & \mathbf{R}_{\boldsymbol{\tau}} \end{bmatrix} \right)$$

for some parameterized matrices $\mathbf{G}_{\boldsymbol{\zeta}}$ and $\mathbf{R}_{\boldsymbol{\tau}}$ (e.g., Robinson 1991). In this instance

$$\boldsymbol{\theta} = \begin{bmatrix} \boldsymbol{\zeta} \\ \boldsymbol{\tau} \end{bmatrix} \quad \text{and} \quad \mathbf{V}_{\boldsymbol{\theta}} = \mathbf{Z}\mathbf{G}_{\boldsymbol{\zeta}}\mathbf{Z}^\top + \mathbf{R}_{\boldsymbol{\tau}}.$$

The log-likelihood of $(\boldsymbol{\beta}, \boldsymbol{\theta})$ is

$$\ell(\boldsymbol{\beta}, \boldsymbol{\theta}) = -\frac{1}{2}\left\{\log|\mathbf{V}_{\boldsymbol{\theta}}| + (\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^{\top}\mathbf{V}_{\boldsymbol{\theta}}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\right\} \\ -\frac{n}{2}\log(2\pi).$$

First,

$$\begin{aligned} d_{\boldsymbol{\beta}}\ell(\boldsymbol{\beta}, \boldsymbol{\theta}) &= -(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^{\top}\mathbf{V}_{\boldsymbol{\theta}}^{-1}d(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) \\ &= (\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^{\top}\mathbf{V}_{\boldsymbol{\theta}}^{-1}\mathbf{X}d\boldsymbol{\beta}, \end{aligned}$$

and

$$\begin{aligned} d^2_{\boldsymbol{\beta}}\ell(\boldsymbol{\beta}, \boldsymbol{\theta}) &= (-\mathbf{X}d\boldsymbol{\beta})^{\top}\mathbf{V}_{\boldsymbol{\theta}}^{-1}\mathbf{X}d\boldsymbol{\beta} \\ &= (d\boldsymbol{\beta})^{\top}(-\mathbf{X}^{\top}\mathbf{V}_{\boldsymbol{\theta}}^{-1}\mathbf{X})d\boldsymbol{\beta} \end{aligned}$$

so, from the Second Identification Theorem,

$$\mathsf{H}_{\boldsymbol{\beta}}\ell(\boldsymbol{\beta}, \boldsymbol{\theta}) = -\mathbf{X}^{\top}\mathbf{V}_{\boldsymbol{\theta}}^{-1}\mathbf{X},$$

and the $\boldsymbol{\beta}$ block of $I(\boldsymbol{\beta}, \boldsymbol{\theta})$ is

$$-E\{\mathsf{H}_{\boldsymbol{\beta}}\ell(\boldsymbol{\beta}, \boldsymbol{\theta})\} = \mathbf{X}^{\top}\mathbf{V}_{\boldsymbol{\theta}}^{-1}\mathbf{X}.$$

To see that the maximum likelihood estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ are asymptotically uncorrelated, we start with

$$d_{\boldsymbol{\theta}}\{d_{\boldsymbol{\beta}}\ell(\boldsymbol{\beta}, \boldsymbol{\theta})\} = (\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^{\top}(d\mathbf{V}_{\boldsymbol{\theta}}^{-1}\mathbf{X})d\boldsymbol{\beta}$$

and then note that

$$E[d_{\boldsymbol{\theta}}\{d_{\boldsymbol{\beta}}\ell(\boldsymbol{\beta}, \boldsymbol{\theta})\}] = \mathbf{0}.$$

It follows that $I(\boldsymbol{\beta}, \boldsymbol{\theta})$ is block-diagonal. It remains to find the $\boldsymbol{\theta}$ block:

$$\begin{aligned} d_{\boldsymbol{\theta}}\ell(\boldsymbol{\beta}, \boldsymbol{\theta}) &= -\frac{1}{2}|\mathbf{V}_{\boldsymbol{\theta}}|^{-1}d|\mathbf{V}_{\boldsymbol{\theta}}| \\ &\quad -\frac{1}{2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^{\top}d\mathbf{V}_{\boldsymbol{\theta}}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) \\ &= -\frac{1}{2}\mathrm{tr}\{\mathbf{V}_{\boldsymbol{\theta}}^{-1}(d\mathbf{V}_{\boldsymbol{\theta}})\} \\ &\quad +\frac{1}{2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^{\top}\mathbf{V}_{\boldsymbol{\theta}}^{-1}(d\mathbf{V}_{\boldsymbol{\theta}})\mathbf{V}_{\boldsymbol{\theta}}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) \end{aligned}$$

$$\begin{aligned} & d^2_{\boldsymbol{\theta}}\ell(\boldsymbol{\beta}, \boldsymbol{\theta}) \\ &= \frac{1}{2}\mathrm{tr}\{\mathbf{V}_{\boldsymbol{\theta}}^{-1}(d\mathbf{V}_{\boldsymbol{\theta}})\mathbf{V}_{\boldsymbol{\theta}}^{-1}(d\mathbf{V}_{\boldsymbol{\theta}})\} \\ &\qquad -\frac{1}{2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^{\top}\mathbf{V}_{\boldsymbol{\theta}}^{-1}(d\mathbf{V}_{\boldsymbol{\theta}})\mathbf{V}_{\boldsymbol{\theta}}^{-1}(d\mathbf{V}_{\boldsymbol{\theta}})\mathbf{V}_{\boldsymbol{\theta}}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) \\ &\qquad -\frac{1}{2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^{\top}\mathbf{V}_{\boldsymbol{\theta}}^{-1}(d\mathbf{V}_{\boldsymbol{\theta}})\mathbf{V}_{\boldsymbol{\theta}}^{-1}(d\mathbf{V}_{\boldsymbol{\theta}})\mathbf{V}_{\boldsymbol{\theta}}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}). \end{aligned}$$

Hence,

$$\begin{aligned} -E\{d^2_{\boldsymbol{\theta}}\ell(\boldsymbol{\beta}, \boldsymbol{\theta})\} &= -\frac{1}{2}\mathrm{tr}\{\mathbf{V}_{\boldsymbol{\theta}}^{-1}(d\mathbf{V}_{\boldsymbol{\theta}})\mathbf{V}_{\boldsymbol{\theta}}^{-1}(d\mathbf{V}_{\boldsymbol{\theta}})\} \\ &\quad +\mathrm{tr}\{\mathbf{V}_{\boldsymbol{\theta}}\mathbf{V}_{\boldsymbol{\theta}}^{-1}(d\mathbf{V}_{\boldsymbol{\theta}})\mathbf{V}_{\boldsymbol{\theta}}^{-1}(d\mathbf{V}_{\boldsymbol{\theta}})\mathbf{V}_{\boldsymbol{\theta}}^{-1}\} \\ &= \frac{1}{2}\mathrm{tr}\{\mathbf{V}_{\boldsymbol{\theta}}^{-1}(d\mathbf{V}_{\boldsymbol{\theta}})\mathbf{V}_{\boldsymbol{\theta}}^{-1}(d\mathbf{V}_{\boldsymbol{\theta}})\} \qquad (6) \end{aligned}$$

where the rule

$$E(\mathbf{u}^{\top}\mathbf{A}\mathbf{u}) = (E\mathbf{u})^{\top}\mathbf{A}(E\mathbf{u}) + \mathrm{tr}\{\mathrm{cov}(\mathbf{u})\mathbf{A}\}$$

has been used. It is tricky to write down an explicit expression for the Hessian for general $\mathbf{V}_{\boldsymbol{\theta}}$, but a common class of submodels that allows for considerable simplification is

$$\mathbf{V}_{\boldsymbol{\theta}} = \sum_{i=1}^{c}\theta_i\,\mathbf{K}_i, \quad \boldsymbol{\theta} = [\theta_1, \ldots, \theta_c]^{\top} \qquad (7)$$

for a set of $n \times n$ matrices $\mathbf{K}_1, \ldots, \mathbf{K}_c$. For example, variance components models (e.g. Searle, Casella, and McCulloch 1992) have this structure. An alternative representation of (7) is

$$\mathrm{vec}(\mathbf{V}_{\boldsymbol{\theta}}) = \mathcal{K}\boldsymbol{\theta}, \qquad \mathcal{K} = [\mathrm{vec}(\mathbf{K}_1)|\ldots|\mathrm{vec}(\mathbf{K}_c)].$$

Let $\otimes$ denote Kronecker product (see, e.g., Magnus and Neudecker 1988, p. 27). From the matrix algebraic result

$$\mathrm{tr}(\mathbf{ABCD}) = \mathrm{vec}(\mathbf{D})^{\top}(\mathbf{A}\otimes\mathbf{C}^{\top})\mathrm{vec}(\mathbf{B}^{\top})$$

we arrive at, via the Second Identification Theorem,

$$\mathsf{H}_{\boldsymbol{\theta}}\,\ell(\boldsymbol{\beta}, \boldsymbol{\theta}) = -\frac{1}{2}\mathcal{K}^{\top}(\mathbf{V}_{\boldsymbol{\theta}}^{-1}\otimes\mathbf{V}_{\boldsymbol{\theta}}^{-1})\mathcal{K}$$

and the full information matrix being

$$I(\boldsymbol{\beta}, \boldsymbol{\theta}) = \begin{bmatrix} \mathbf{X}^{\top}\mathbf{V}_{\boldsymbol{\theta}}^{-1}\mathbf{X} & \mathbf{0} \\ \mathbf{0} & \frac{1}{2}\mathcal{K}^{\top}(\mathbf{V}_{\boldsymbol{\theta}}^{-1}\otimes\mathbf{V}_{\boldsymbol{\theta}}^{-1})\mathcal{K} \end{bmatrix}.$$

However, since $\mathbf{V}_{\boldsymbol{\theta}}^{-1}\otimes\mathbf{V}_{\boldsymbol{\theta}}^{-1}$ is an $n^2 \times n^2$ matrix a preferable expression for practical use is

$$I(\boldsymbol{\beta}, \boldsymbol{\theta}) = \begin{bmatrix} \mathbf{X}^{\top}\mathbf{V}_{\boldsymbol{\theta}}^{-1}\mathbf{X} & \mathbf{0} \\ \mathbf{0} & \{\frac{1}{2}\mathrm{tr}(\mathbf{V}_{\boldsymbol{\theta}}^{-1}\mathbf{K}_i\mathbf{V}_{\boldsymbol{\theta}}^{-1}\mathbf{K}_j)\}_{1\le i,j\le c} \end{bmatrix}$$

which follows from (6) and the fact that

$$\frac{d}{d\theta_i}\mathbf{V}_{\boldsymbol{\theta}} = \mathbf{K}_i, \quad 1 \le i \le c.$$

This matches equation (38) on page 239 of Searle, McCulloch, and Casella (1992).

4.4 Generalized Linear Mixed Models

Generalized linear mixed models extend generalized linear models by allowing for the incorporation of random effects (e.g., McCulloch and Searle 2000). The exponential family model of Section 4.1 can be extended by writing

$$f(\mathbf{y}|\mathbf{u}) = \exp\{\mathbf{y}^{\top}(\mathbf{X}\boldsymbol{\beta}+\mathbf{Z}\mathbf{u}) - \mathbf{1}^{\top}\,b(\mathbf{X}\boldsymbol{\beta}+\mathbf{Z}\mathbf{u}) + \mathbf{1}^{\top}c(\mathbf{y})\},$$

where the random effects vector $\mathbf{u}$ has density $f(\mathbf{u};\boldsymbol{\theta})$. Mostly common, the random effects distribution is Gaussian: $N(\mathbf{0}, \mathbf{G}_{\boldsymbol{\theta}})$, for some positive definite matrix $\mathbf{G}_{\boldsymbol{\theta}}$ parameterized by $\boldsymbol{\theta}$. In this case, with q denoting the dimension of $\mathbf{u}$,

$$f(\mathbf{u};\boldsymbol{\theta}) = (2\pi)^{-q/2}|\mathbf{G}_{\boldsymbol{\theta}}|^{-1/2}\exp\left(-\frac{1}{2}\,\mathbf{u}^{\top}\mathbf{G}_{\boldsymbol{\theta}}^{-1}\mathbf{u}\right)$$

and the likelihood of $(\boldsymbol{\beta}, \boldsymbol{\theta})$ is then

$$\begin{aligned} \mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\theta}) &= f(\mathbf{y};\boldsymbol{\beta},\boldsymbol{\theta}) \\ &= \int_{\mathbb{R}^q} f(\mathbf{y}|\mathbf{u})f(\mathbf{u})\,d\mathbf{u} \\ &= (2\pi)^{-q/2}|\mathbf{G}_{\boldsymbol{\theta}}|^{-1/2}\exp\{\mathbf{1}^{\top}c(\mathbf{y})\}\,J(\boldsymbol{\beta},\boldsymbol{\theta}) \end{aligned}$$

where

$$J(\boldsymbol{\beta},\boldsymbol{\theta}) = \int_{\mathbb{R}^q} \exp\Big\{\mathbf{y}^\top(\mathbf{X}\boldsymbol{\beta}+\mathbf{Z}\mathbf{u}) \\ -\mathbf{1}^\top b(\mathbf{X}\boldsymbol{\beta}+\mathbf{Z}\mathbf{u}) - \frac{1}{2}\mathbf{u}^\top\mathbf{G}_{\boldsymbol{\theta}}^{-1}\mathbf{u}\Big\}\, d\mathbf{u}.$$

The integral $J(\boldsymbol{\beta},\boldsymbol{\theta})$ is, in general, irreducible so approximation methods are required for maximum likelihood estimation. The simplest is penalized quasi-likelihood (PQL) (Breslow and Clayton 1993; Wolfinger and O'Connell 1993). Note that

$$J(\boldsymbol{\beta},\boldsymbol{\theta}) = \int_{\mathbb{R}^q} \exp\{h(\mathbf{u})\}\, d\mathbf{u}, \tag{8}$$

where

$$h(\mathbf{u}) \;=\; \mathbf{y}^\top(\mathbf{X}\boldsymbol{\beta}+\mathbf{Z}\mathbf{u}) - \mathbf{1}^\top b(\mathbf{X}\boldsymbol{\beta}+\mathbf{Z}\mathbf{u}) - \frac{1}{2}\mathbf{u}^\top\mathbf{G}_{\boldsymbol{\theta}}^{-1}\mathbf{u}. \tag{9}$$

PQL is based on Laplace approximation of $J(\boldsymbol{\beta},\boldsymbol{\theta})$, which involves the Taylor series approximation

$$h(\mathbf{u}) \simeq h(\widetilde{\mathbf{u}}) + \mathsf{D}\,h(\widetilde{\mathbf{u}})(\mathbf{u}-\widetilde{\mathbf{u}}) + \frac{1}{2}(\mathbf{u}-\widetilde{\mathbf{u}})^\top\mathsf{H}\,h(\widetilde{\mathbf{u}})(\mathbf{u}-\widetilde{\mathbf{u}}).$$

We therefore need expressions for $\mathsf{D}\,h(\mathbf{u})$ and $\mathsf{H}\,h(\mathbf{u})$. First

$$d_{\mathbf{u}}h(\mathbf{u}) = \{\mathbf{y}-b'(\mathbf{X}\boldsymbol{\beta}+\mathbf{Z}\mathbf{u})\}^\top\mathbf{Z}d\mathbf{u} - \mathbf{u}^\top\mathbf{G}_{\boldsymbol{\theta}}^{-1}d\mathbf{u}$$

and so

$$\mathsf{D}_{\mathbf{u}}\,h(\mathbf{u}) = \{\mathbf{y}-b'(\mathbf{X}\boldsymbol{\beta}+\mathbf{Z}\mathbf{u})\}^\top\mathbf{Z} - \mathbf{u}^\top\mathbf{G}_{\boldsymbol{\theta}}^{-1}.$$

Second,

$$d^2_{\mathbf{u}}h(\mathbf{u}) = -(d\mathbf{u})^\top[\mathbf{Z}^\top\mathrm{diag}\{b''(\mathbf{X}\boldsymbol{\beta}+\mathbf{Z}\mathbf{u})\}\mathbf{Z}+\mathbf{G}_{\boldsymbol{\theta}}^{-1}](d\mathbf{u})$$

which leads to

$$\mathsf{H}_{\mathbf{u}}\,h(\mathbf{u}) = -\mathbf{Z}^\top\mathrm{diag}\{b''(\mathbf{X}\boldsymbol{\beta}+\mathbf{Z}\mathbf{u})\}\mathbf{Z} - \mathbf{G}_{\boldsymbol{\theta}}^{-1}.$$

The resulting log-likelihood approximation is then

$$\ell(\boldsymbol{\beta},\boldsymbol{\theta}) \;\simeq\; \mathbf{y}^\top(\mathbf{X}\boldsymbol{\beta}+\mathbf{Z}\widetilde{\mathbf{u}}) - \mathbf{1}^\top b(\mathbf{X}\boldsymbol{\beta}+\mathbf{Z}\widetilde{\mathbf{u}}) - \frac{1}{2}\widetilde{\mathbf{u}}^\top\mathbf{G}_{\boldsymbol{\theta}}^{-1}\widetilde{\mathbf{u}} \\ -\frac{1}{2}\log|\mathbf{I}-\mathbf{G}_{\boldsymbol{\theta}}\mathbf{Z}\mathrm{diag}\{b''(\mathbf{X}\boldsymbol{\beta}+\mathbf{Z}\widetilde{\mathbf{u}})\}\mathbf{Z}^\top|.$$

However, for ease of fitting, PQL uses one final approximation, based on the argument that $b''(\mathbf{X}\boldsymbol{\beta}+\mathbf{Z}\widetilde{\mathbf{u}})$ is relatively constant as a function of $\boldsymbol{\beta}$. This argument gives some justification for its omission from the log-likelihood to yield

$$\ell(\boldsymbol{\beta},\boldsymbol{\theta}) \simeq \mathbf{y}^\top(\mathbf{X}\boldsymbol{\beta}+\mathbf{Z}\widetilde{\mathbf{u}}) - \mathbf{1}^\top b(\mathbf{X}\boldsymbol{\beta}+\mathbf{Z}\widetilde{\mathbf{u}}) - \frac{1}{2}\widetilde{\mathbf{u}}^\top\mathbf{G}_{\boldsymbol{\theta}}^{-1}\widetilde{\mathbf{u}}.$$

The maximization of the right hand side of this expression can be carried out using standard generalized linear model software.

4.5 *t* Distribution Regression

Lange, Little, and Taylor (1989) described robust statistical modeling based on the t distribution. Here we concentrate on the regression model

$$y_i \sim t((\mathbf{X}\boldsymbol{\beta})_i, \psi^2, \nu),$$

where $t(\mu,\psi^2,\nu)$ denotes the t distribution with density function

$$f(t;\mu,\psi^2,\nu) \;\equiv\; \psi^{-1}c(\nu)\{1+(t-\mu)^2/(\nu\psi^2)\}^{-(\nu+1)/2}, \tag{10}$$

and $c(\nu) = \left(\Gamma\left(\frac{\nu+1}{2}\right)\right)/(\sqrt{\pi}\Gamma(\nu/2)\sqrt{\nu})$. The log-likelihood is

$$\ell(\boldsymbol{\beta},\psi^2,\nu) = n\log\{c(\nu)/\psi\} \\ -\frac{\nu+1}{2}\mathbf{1}^\top\log\{\mathbf{1}+(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^2/(\nu\psi^2)\}.$$

First, using the chain rule (Section 3.3.2) with $s(u) = \log\{1+u^2/(\nu\psi^2)\}$,

$$d_{\boldsymbol{\beta}}\,\ell(\boldsymbol{\beta},\psi^2,\nu) \;=\; \tfrac{(\nu+1)}{2}\mathbf{1}^\top\mathrm{diag}\left\{\tfrac{2(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})/(\nu\psi^2)}{\mathbf{1}+(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^2/(\nu\psi^2)}\right\}\mathbf{X}d\boldsymbol{\beta} \\ \;=\; \frac{\nu+1}{\nu\psi^2}\left(\frac{\mathbf{y}-\mathbf{X}\boldsymbol{\beta}}{\mathbf{1}+\frac{1}{\nu\psi^2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^2}\right)^\top\mathbf{X}d\boldsymbol{\beta}$$

and so using the chain rule again with $s(u) = u/\{1+u^2/(\nu\psi^2)\}$,

$$d^2_{\boldsymbol{\beta}}\,\ell(\boldsymbol{\beta},\psi^2,\nu) \\ = -(d\boldsymbol{\beta})^\top\mathbf{X}^\top\frac{\nu+1}{\nu\psi^2}\mathrm{diag}\left[\frac{\mathbf{1}-\frac{1}{\nu\psi^2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^2}{\{\mathbf{1}+\frac{1}{\nu\psi^2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^2\}^2}\right]\mathbf{X}d\boldsymbol{\beta}.$$

It is easily shown that for T having density (10)

$$E\left[\frac{1-\frac{1}{\nu\psi^2}(T-\mu)^2}{\{1+\frac{1}{\nu\psi^2}(T-\mu)^2\}^2}\right] = \frac{\nu}{\nu+3}.$$

So the $\boldsymbol{\beta}$ block of the information matrix of $(\boldsymbol{\beta},\psi^2,\nu)$ is

$$\frac{\nu+1}{(\nu+3)\psi^2}\mathbf{X}^\top\mathbf{X}$$

which agrees with the results derived in Lange, Little, and Taylor (1989). Similar arguments can be used to establish the orthogonality of $\boldsymbol{\beta}$ and (ψ^2,ν). The (ψ^2,ν) block of the information matrix can be obtained using scalar calculus.

4.6 Negative Binomial Regression

The negative binomial regression model (e.g., Lawless 1987) is

$$f(y_i;\boldsymbol{\beta},\kappa) = \frac{\Gamma(y_i+\kappa)}{\Gamma(\kappa)\Gamma(y_i+1)} \\ \times\left(\frac{\exp{(\mathbf{X}\boldsymbol{\beta})_i}}{\kappa+\exp{(\mathbf{X}\boldsymbol{\beta})_i}}\right)^{y_i}\left(\frac{\kappa}{\kappa+\exp{(\mathbf{X}\boldsymbol{\beta})_i}}\right)^{\kappa}$$

The log-likelihood for independent data from this model is

$$\ell(\boldsymbol{\beta},\kappa) \;=\; \mathbf{y}^\top[\mathbf{X}\boldsymbol{\beta}-\log\{\kappa\mathbf{1}+\exp(\mathbf{X}\boldsymbol{\beta})\}] \\ -\kappa\mathbf{1}^\top\log\{\kappa\mathbf{1}+\exp(\mathbf{X}\boldsymbol{\beta})\} \\ +n\kappa\log(\kappa)+\mathbf{1}^\top\log\Gamma(\mathbf{y}+\kappa\mathbf{1})-n\log\Gamma(\kappa).$$

Now, using the chain rule (Section 3.3.2) with $s(u) = \log(\kappa+e^u)$,

$$d_{\boldsymbol{\beta}}\ell(\boldsymbol{\beta},\kappa) \\ \;=\; \mathbf{y}^\top\left(\mathbf{X}-\mathrm{diag}[\exp(\mathbf{X}\boldsymbol{\beta})/\{\kappa\mathbf{1}+\exp(\mathbf{X}\boldsymbol{\beta})\}]\mathbf{X}\right)d\boldsymbol{\beta}$$

$$
\begin{aligned}
&-\kappa\mathbf{1}^\top \text{diag}[\exp(\mathbf{X}\boldsymbol{\beta})/\{\kappa\mathbf{1}+\exp(\mathbf{X}\boldsymbol{\beta})\}]\mathbf{X}d\boldsymbol{\beta}\\
&= \kappa\left(\frac{\mathbf{y}-\exp(\mathbf{X}\boldsymbol{\beta})}{\kappa\mathbf{1}+\exp(\mathbf{X}\boldsymbol{\beta})}\right)^\top \mathbf{X}d\boldsymbol{\beta}
\end{aligned}
$$

and so, using the chain rule with $s(u)=(y-e^u)/(\kappa+e^u)$,

$$d^2_{\boldsymbol{\beta}}\ell(\boldsymbol{\beta},\kappa)=\kappa\left[\text{diag}\left\{-\frac{\exp(\mathbf{X}\boldsymbol{\beta})\odot(\kappa\mathbf{1}+\mathbf{y})}{\{\kappa\mathbf{1}+\exp(\mathbf{X}\boldsymbol{\beta})\}^2}\right\}\mathbf{X}\right]^\top \mathbf{X}\boldsymbol{\beta}.$$

Therefore

$$
\begin{aligned}
&-E\{d^2_{\boldsymbol{\beta}}\ell(\boldsymbol{\beta},\kappa)\}\\
&\quad=\kappa\left\{\text{diag}\left(\frac{\exp(\mathbf{X}\boldsymbol{\beta})}{\kappa\mathbf{1}+\exp(\mathbf{X}\boldsymbol{\beta})}\right)\mathbf{X}d\boldsymbol{\beta}\right\}^\top \mathbf{X}d\boldsymbol{\beta}\\
&\quad=(d\boldsymbol{\beta})^\top\mathbf{X}^\top\text{diag}\left(\frac{\kappa\exp(\mathbf{X}\boldsymbol{\beta})}{\kappa\mathbf{1}+\exp(\mathbf{X}\boldsymbol{\beta})}\right)\mathbf{X}d\boldsymbol{\beta}
\end{aligned}
$$

so the $\boldsymbol{\beta}$ block of $I(\boldsymbol{\beta},\kappa)$ is

$$\mathbf{X}^\top\text{diag}\left(\frac{\kappa\exp(\mathbf{X}\boldsymbol{\beta})}{\kappa\mathbf{1}+\exp(\mathbf{X}\boldsymbol{\beta})}\right)\mathbf{X}.$$

For the $(\boldsymbol{\beta},\kappa)$ block

$$
\begin{aligned}
&d_\kappa\{d_{\boldsymbol{\beta}}\ell(\boldsymbol{\beta},\kappa)\}\\
&\quad=\left\{(\mathbf{y}-\exp(\mathbf{X}\boldsymbol{\beta}))\odot d_\kappa\left(\frac{\kappa\mathbf{1}}{\kappa\mathbf{1}+\exp(\mathbf{X}\boldsymbol{\beta})}\right)\right\}^\top\mathbf{X}d\boldsymbol{\beta}
\end{aligned}
$$

and so

$$-E[d_\kappa\{d_{\boldsymbol{\beta}}\ell(\boldsymbol{\beta},\kappa)\}]=\mathbf{0}.$$

Using ordinary differential calculus we can establish

$$
\begin{aligned}
\frac{d^2}{d\kappa^2}\ell(\boldsymbol{\beta},\kappa)=\sum_{i=1}^n\left\{\text{trigamma}(y_i+\kappa)-\frac{1}{\exp(\mathbf{X}\boldsymbol{\beta})_i+\kappa}\right\}\\
+n\{1/\kappa-\text{trigamma}(\kappa)\}
\end{aligned}
$$

so the information matrix for negative binomial regression is

$$I(\beta,\kappa)=\begin{bmatrix}\mathbf{X}^\top\text{diag}\left(\frac{\kappa\exp(\mathbf{X}\boldsymbol{\beta})}{\kappa\mathbf{1}+\exp(\mathbf{X}\boldsymbol{\beta})}\right)\mathbf{X} & \mathbf{0}\\ \mathbf{0} & \begin{matrix}\sum_{i=1}^n\left[\frac{1}{\kappa+\exp(\mathbf{X}\boldsymbol{\beta})_\mathrm{i}}\right.\\ \left.-E\{\text{trigamma}(y_i+\kappa)\}\right]\\ +n\{\text{trigamma}(\kappa)-1/\kappa\}\end{matrix}\end{bmatrix}.$$

This matches the results given in Lawless (1987).

5. EXTENSION TO MATRIX ARGUMENT FUNCTIONS

In the case where the parameter of interest is a matrix rather than a vector then the rules of Section 3 can be applied to the vectorized version of the matrix parameter. We will give but one illustration of this. Suppose that

$$\mathbf{x}_1,\ldots,\mathbf{x}_n\quad\text{iid}\quad N(\boldsymbol{\mu},\boldsymbol{\Sigma}),$$

the p variate normal distribution. In this case $\boldsymbol{\Sigma}$ is a general $p\times p$ matrix. The log-likelihood for $(\boldsymbol{\mu},\boldsymbol{\Sigma})$ is

$$
\begin{aligned}
\ell(\boldsymbol{\mu},\boldsymbol{\Sigma})=-\frac{np}{2}\log(2\pi)-\frac{n}{2}\log|\boldsymbol{\Sigma}|\\
-\frac{1}{2}\sum_{i=1}^n(\mathbf{x}_i-\boldsymbol{\mu})\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i-\boldsymbol{\mu}).
\end{aligned}
$$

Ordinary vector differential calculus leads to

$$d_\mu\ell(\boldsymbol{\mu},\boldsymbol{\Sigma})=\sum_{i=1}^n(\mathbf{x}_i-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}d\boldsymbol{\mu}$$

from which it is easily established that

$$\widehat{\boldsymbol{\mu}}=\overline{\mathbf{x}}\equiv\frac{1}{n}\sum_{i=1}^n\mathbf{x}_i$$

is the maximizer of $\ell(\boldsymbol{\mu},\boldsymbol{\Sigma})$ for every $\boldsymbol{\Sigma}$. The maximum likelihood estimate of $\boldsymbol{\Sigma}$ then maximizes

$$
\begin{aligned}
\ell(\overline{\mathbf{x}},\boldsymbol{\Sigma})=\frac{-np}{2}\log(2\pi)\\
-\frac{n}{2}\log|\boldsymbol{\Sigma}|-\frac{1}{2}\sum_{i=1}^n(\mathbf{x}_i-\overline{\mathbf{x}})^\top\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i-\overline{\mathbf{x}}).
\end{aligned}
$$

The calculus required for this minimization can be performed using the results of Section 3, but with Σ replaced by $\text{vec}(\boldsymbol{\Sigma})$:

$$
\begin{aligned}
d_{\boldsymbol{\Sigma}}\ell(\overline{\mathbf{x}},\boldsymbol{\Sigma}) &= -\frac{n}{2}|\boldsymbol{\Sigma}|^{-1}d|\boldsymbol{\Sigma}|\\
&\quad-\frac{1}{2}\sum_{i=1}^n(\mathbf{x}_i-\overline{\mathbf{x}})^\top(d\boldsymbol{\Sigma}^{-1})(\mathbf{x}_i-\overline{\mathbf{x}})\\
&= -\frac{n}{2}\text{tr}(\boldsymbol{\Sigma}^{-1}d\boldsymbol{\Sigma})\\
&\quad+\frac{1}{2}\sum_{i=1}^n(\mathbf{x}_i-\overline{\mathbf{x}})^\top\boldsymbol{\Sigma}^{-1}(d\boldsymbol{\Sigma})\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i-\overline{\mathbf{x}})\\
&= -\frac{n}{2}\text{tr}(\boldsymbol{\Sigma}^{-1}d\boldsymbol{\Sigma})\\
&\quad+\frac{1}{2}\sum_{i=1}^n\text{tr}\{\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i-\overline{\mathbf{x}})(\mathbf{x}_i-\overline{\mathbf{x}})^\top\boldsymbol{\Sigma}^{-1}(d\boldsymbol{\Sigma})\}\\
&= -\frac{n}{2}\text{vec}(\boldsymbol{\Sigma}^{-1})^\top d\text{vec}(\boldsymbol{\Sigma})\\
&\quad+\frac{1}{2}\sum_{i=1}^n\text{vec}\{\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i-\overline{\mathbf{x}})\\
&\quad\times(\mathbf{x}_i-\overline{\mathbf{x}})^\top\boldsymbol{\Sigma}^{-1}\}^\top d\text{vec}(\boldsymbol{\Sigma}),
\end{aligned}
$$

where we have used the result $\text{tr}(\mathbf{A}^\top\mathbf{B})=\text{vec}(\mathbf{A})^\top\text{vec}(\mathbf{B})$. Therefore, the derivative vector of $\ell(\overline{\mathbf{x}},\boldsymbol{\Sigma})$ with respect to $\text{vec}(\boldsymbol{\Sigma})$ is

$$
\begin{aligned}
\mathsf{D}\ell(\overline{\mathbf{x}},\boldsymbol{\Sigma})=\text{vec}\Bigg[-\frac{n}{2}\boldsymbol{\Sigma}^{-1}\\
+\frac{1}{2}\sum_{i=1}^n\{\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i-\overline{\mathbf{x}})(\mathbf{x}_i-\overline{\mathbf{x}})^\top\boldsymbol{\Sigma}^{-1}\}\Bigg]^\top.
\end{aligned}
$$

Clearly, this is zero if and only if

$$\boldsymbol{\Sigma}=\frac{1}{n}\sum_{i=1}^n(\mathbf{x}_i-\overline{\mathbf{x}})(\mathbf{x}_i-\overline{\mathbf{x}})^\top$$

which is the sample covariance matrix with an n divisor.

[Received TKK. Revised TKK.]

REFERENCES

Basilevsky, A. (1983), *Applied Matrix Algebra in the Statistical Sciences*, New York: North-Holland.

Breslow, N. E., and Clayton, D. G. (1993), "Approximated Inference in Generalized Linear Mixed Models," *Journal of the American Statistical Association*, 88, 9–25.

Cressie, N. (1993), *Statistics for Spatial Data*, New York: Wiley.

Dwyer, P. S. (1967), "Some Applications of Matrix Derivatives in Multivariate Analysis," *Journal of the American Statistical Association*, 62, 607–625.

Harville, D. A. (1997), *Matrix Algebra From a Statistician's Perspective*, New York: Springer.

Lange, K. L., Little, R. J. A., and Taylor, J. M. G. (1989), "Robust Statistical Modeling Using the t-Distribution," *Journal of the American Statistical Association*, 84, 881–896.

Lawless, J. F. (1987), "Negative Binomial and Mixed Poisson Regression," *The Canadian Journal of Statistics*, 15, 209–225.

Magnus, J. R., and Neudecker, H. (1988), *Matrix Differential Calculus With Applications in Statistics and Econometrics*, Chichester: Wiley.

McCullagh, P., and Nelder, J. A. (1989), *Generalized Linear Models* (2nd ed.), London: Chapman and Hall.

McCulloch, C. E., and Searle, S. R. (2000), *Generalized, Linear, and Mixed Models*, New York: Wiley.

Robinson, G. K. (1991), "That BLUP is a Good Thing: The Estimation of Random Effects," *Statistical Science*, 6, 15–51.

Searle, S. R. (1982), *Matrix Algebra Useful for Statistics*, New York: Wiley.

Searle, S. R., Casella, G., and McCulloch, C. E. (1992), *Variance Components*, New York: Wiley.

Wolfinger, R., and O'Connell, M. (1993), "Generalized Linear Mixed Models: A Pseudo-Likelihood Approach," *Journal of Statistical Computation and Simulation*, 48, 233–243.