CHAPTER 9

Quasi-likelihood functions

9.1 Introduction

One of the few points on which theoretical statisticians of all persuasions are agreed is the importance of the role played by the likelihood function in statistical inference. This role has been illustrated, chiefly from the frequentist viewpoint, in the developments of Chapters 2–8. In order to construct a likelihood function it is usually necessary to posit a probabilistic mechanism specifying, for a range of parameter values, the probabilities of all relevant samples that might possibly have been observed. Such a specification implies either knowledge of the mechanism by which the data were generated or substantial experience of similar data from previous experiments.

Often there is no theory available on the random mechanism by which the data were generated. We may, however, be able to specify the range of possible response values (discrete, continuous, positive, ...), and past experience with similar data is usually sufficient to specify, in a qualitative fashion, a few additional characteristic features of the data, such as

1. how the mean or median response is affected by external stimuli or treatments;
2. how the variability of the response changes with the average response;
3. whether the observations are statistically independent;
4. whether the response distribution under fixed treatment conditions is skewed positively, negatively or is symmetric.

Often interest attaches to how the mean response or other simple functional is affected by one or more covariates. Usually there is substantial prior information on the likely nature of this relationship, but rather little about the pattern of higher-order cumulants

or moments.

The purpose of this chapter is to show how inferences can be drawn from experiments in which there is insufficient information to construct a likelihood function. We concentrate mainly on the case in which the observations are independent and where the effects of interest can be described by a model for $E(Y)$.

9.2 Independent observations

9.2.1 *Covariance functions*

Suppose that the components of the response vector $\mathbf{Y}$ are independent with mean vector $\boldsymbol{\mu}$ and covariance matrix $\sigma^2\mathbf{V}(\boldsymbol{\mu})$, where σ^2 may be unknown and $\mathbf{V}(\boldsymbol{\mu})$ is a matrix of known functions. It is assumed throughout this section that the parameters of interest, $\boldsymbol{\beta}$, relate to the dependence of $\boldsymbol{\mu}$ on covariates $\boldsymbol{x}$. The nature of this relationship need not concern us for the moment, so we write $\boldsymbol{\mu}(\boldsymbol{\beta})$, thereby absorbing the covariates into the regression function. An important point is that σ^2 is assumed constant—in particular that σ^2 does not depend on $\boldsymbol{\beta}$.

Since the components of $\mathbf{Y}$ are independent by assumption the matrix $\mathbf{V}(\mu)$ must be diagonal. Thus we write

$$\mathbf{V}(\boldsymbol{\mu}) = \text{diag}\{V_1(\boldsymbol{\mu}), \ldots, V_n(\boldsymbol{\mu})\}.$$

One further assumption is required concerning the functions $V_i(\boldsymbol{\mu})$, namely that $V_i(\boldsymbol{\mu})$ must depend only on the ith component of $\boldsymbol{\mu}$. In principle it is possible, even under independence, for $V_1(\boldsymbol{\mu})$ to depend on several components of $\boldsymbol{\mu}$. However it is difficult to imagine a plausible physical mechanism that would produce such dependence in the variance function, while at the same time keeping the random variables statistically independent.

The above assumption of functional independence, namely that

$$\mathbf{V}(\boldsymbol{\mu}) = \text{diag}\{V_1(\mu_1), \ldots, V_n(\mu_n)\}, \tag{9.1}$$

although sensible physically, has been made for technical mathematical reasons that will become apparent in section 9.3. It is no more than a happy accident that such a technical mathematical requirement should coincide with what is sensible on external physical or scientific grounds.

In the majority of applications the functions $V_1(\cdot), \ldots, V_n(\cdot)$ may be taken to be identical, though their arguments, and hence their values, are different. However, this assumption is not required in the algebra that follows.

9.2.2 *Construction of the quasi-likelihood function*

Consider first a single component of the response vector $\mathbf{Y}$, which we write as Y or y without subscripts. Under the conditions listed above, the function

$$U = u(\mu; Y) = \frac{Y - \mu}{\sigma^2 V(\mu)}$$

has the following properties in common with a log-likelihood derivative:

$$\begin{aligned} E(U) &= 0, \\ \operatorname{var}(U) &= 1/\{\sigma^2 V(\mu)\}, \\ -E\Big(\frac{\partial U}{\partial \mu}\Big) &= 1/\{\sigma^2 V(\mu)\}. \end{aligned} \tag{9.2}$$

Since most first-order asymptotic theory connected with likelihood functions is founded on these three properties, it is not surprising that, to some extent, the integral

$$Q(\mu; y) = \int_y^\mu \frac{y - t}{\sigma^2 V(t)}\, dt \tag{9.3}$$

if it exists, should behave like a log-likelihood function for μ under the very mild assumptions stated in the previous two sections. Some examples of such quasi-likelihoods for a number of common variance functions are given in Table 9.1. Many, but not all, of these quasi-likelihoods correspond to real log likelihoods for known distributions.

We refer to $Q(\mu; y)$ as the quasi-likelihood, or more correctly, as the log quasi-likelihood for μ based on data y. Since the components of $\mathbf{Y}$ are independent by assumption, the quasi-likelihood for the complete data is the sum of the individual contributions

$$Q(\boldsymbol{\mu}; \mathbf{y}) = \sum Q_i(\mu_i; y_i).$$

Table 9.1. *Quasi-likelihoods associated with some simple variance functions*

Variance function $V(\mu)$	*Quasi-likelihood* $Q(\mu; y)$	*Canonical parameter* θ	*Distribution name*	*Range restrictions*
1	$-(y-\mu)^2/2$	μ	Normal	—
μ	$y\log\mu - \mu$	$\log\mu$	Poisson	$\mu > 0,\ y \geq 0$
μ^2	$-y/\mu - \log\mu$	$-1/\mu$	Gamma	$\mu > 0,\ y > 0$
μ^3	$-y/(2\mu^2) + 1/\mu$	$-1/(2\mu^2)$	Inverse Gaussian	$\mu > 0,\ y > 0$
μ^ζ	$\mu^{-\zeta}\left(\frac{\mu y}{1-\zeta} - \frac{\mu^2}{2-\zeta}\right)$	$\frac{1}{(1-\zeta)\mu^{\zeta-1}}$	—	$\mu > 0,\ \zeta \neq 0, 1, 2$
$\mu(1-\mu)$	$y\log\left(\frac{\mu}{1-\mu}\right) + \log(1-\mu)$	$\log\left(\frac{\mu}{1-\mu}\right)$	Binomial/m	$0 < \mu < 1,\ 0 \leq y \leq 1$
$\mu^2(1-\mu)^2$	$(2y-1)\log\left(\frac{\mu}{1-\mu}\right) - \frac{y}{\mu} - \frac{1-y}{1-\mu}$	—	—	$0 < \mu < 1,\ 0 < y < 1$
$\mu + \mu^2/k$	$y\log\left(\frac{\mu}{k+\mu}\right) + k\log\left(\frac{k}{k+\mu}\right)$	$\log\left(\frac{\mu}{k+\mu}\right)$	Negative binomial	$\mu > 0,\ y \geq 0$

By analogy, the quasi-deviance function corresponding to a single observation is

$$D(y;\mu) = -2\sigma^2 Q(\mu;y) = 2\int_{\mu}^{y} \frac{y-t}{V(t)}\,dt, \tag{9.4}$$

which is evidently strictly positive except at $y = \mu$. The total deviance $D(\mathbf{y};\boldsymbol{\mu})$, obtained by adding over the components, is a computable function depending on $\mathbf{y}$ and $\boldsymbol{\mu}$ alone: it does not depend on σ^2.

9.2.3 *Parameter estimation*

The quasi-likelihood estimating equations for the regression parameters $\boldsymbol{\beta}$, obtained by differentiating $Q(\boldsymbol{\mu};\mathbf{y})$, may be written in the form $\mathbf{U}(\hat{\boldsymbol{\beta}}) = \mathbf{0}$, where

$$\mathbf{U}(\boldsymbol{\beta}) = \mathbf{D}^T\mathbf{V}^{-1}(\mathbf{Y} - \boldsymbol{\mu})/\sigma^2 \tag{9.5}$$

is called the quasi-score function. In this expression the components of $\mathbf{D}$, of order $n \times p$, are $D_{ir} = \partial\mu_i/\partial\beta_r$, the derivatives of $\boldsymbol{\mu}(\boldsymbol{\beta})$ with respect to the parameters.

The covariance matrix of $\mathbf{U}(\boldsymbol{\beta})$, which is also the negative expected value of $\partial\mathbf{U}(\boldsymbol{\beta})/\partial\boldsymbol{\beta}$, is

$$\mathbf{i}_\beta = \mathbf{D}^T\mathbf{V}^{-1}\mathbf{D}/\sigma^2. \tag{9.6}$$

For quasi-likelihood functions, this matrix plays the same role as the Fisher information for ordinary likelihood functions. In particular, under the usual limiting conditions on the eigenvalues of $\mathbf{i}_\beta$, the asymptotic covariance matrix of $\hat{\boldsymbol{\beta}}$ is

$$\mathrm{cov}(\hat{\boldsymbol{\beta}}) \simeq \mathbf{i}_\beta^{-1} = \sigma^2(\mathbf{D}^T\mathbf{V}^{-1}\mathbf{D})^{-1},$$

as can be seen from the argument given below.

Beginning with an arbitrary value $\hat{\boldsymbol{\beta}}_0$ sufficiently close to $\hat{\boldsymbol{\beta}}$, the sequence of parameter estimates generated by the Newton-Raphson method with Fisher scoring is

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

The quasi-likelihood estimate $\hat{\boldsymbol{\beta}}$ may be obtained by iterating until convergence occurs. An important property of the sequence so generated is that it does not depend on the value of σ^2.

For theoretical calculations it is helpful to imagine the iterations starting at the true value, $\boldsymbol{\beta}$. Thus we find

$$\hat{\boldsymbol{\beta}}_1 = \boldsymbol{\beta} + (\mathbf{D}^T\mathbf{V}^{-1}\mathbf{D})^{-1}\mathbf{D}^T\mathbf{V}^{-1}(\mathbf{y} - \boldsymbol{\mu}), \tag{9.7}$$

showing that the one-step estimate is a linear function of the data. Provided that the eigenvalues of $\mathbf{i}_\beta$ are sufficiently large, subsequent iterations produce asymptotically negligible adjustments to $\hat{\boldsymbol{\beta}}_1$. Thus, for a first-order asymptotic theory, we may take $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_1$, even though $\hat{\boldsymbol{\beta}}_1$ is not a computable statistic. Approximate unbiasedness and asymptotic Normality of $\hat{\boldsymbol{\beta}}$ follow directly from (9.7) under the second-moment assumptions made in this chapter.

In all of the above respects the quasi-likelihood behaves just like an ordinary log likelihood. For the estimation of σ^2, however, $Q(\cdot; y)$ does not behave like a log likelihood. The conventional estimate of σ^2 is a moment estimator based on the residual vector $\mathbf{Y} - \hat{\boldsymbol{\mu}}$, namely

$$\tilde{\sigma}^2 = \frac{1}{n-p}\sum_i (Y_i - \hat{\mu}_i)^2/V_i(\hat{\mu}_i) = X^2/(n-p),$$

where X^2 is the generalized Pearson statistic.

9.2.4 *Example: incidence of leaf-blotch on barley*

The data in Table 9.2, taken from Wedderburn (1974), concerns the incidence of *Rhynchosporium secalis*, or leaf blotch, on 10 varieties of barley grown at nine sites in 1965. The response, which is the percentage leaf area affected, is a continuous proportion in the interval $[0, 1]$. For convenience of discussion we take Y to be a proportion in $[0, 1]$ rather than a percentage. Following the precedent set in section 6.3.1, we might attempt an analysis, treating the data as pseudo-binomial observations, taking the variances, at least initially to be $\sigma^2\mu(1-\mu)$. A linear logistic model with main effects appears to describe adequately the site and variety effects.

This analysis is certainly reasonable as a first step. The usual residual plots and additivity tests indicate no significant departures

Table 9.2. *Incidence of* R. secalis *on the leaves of ten varieties of barley grown at nine sites: response is the percentage of leaf affected*

	Variety										
Site	1	2	3	4	5	6	7	8	9	10	*Mean*
1	0.05	0.00	0.00	0.10	0.25	0.05	0.50	1.30	1.50	1.50	0.52
2	0.00	0.05	0.05	0.30	0.75	0.30	3.00	7.50	1.00	12.70	2.56
3	1.25	1.25	2.50	16.60	2.50	2.50	0.00	20.00	37.50	26.25	11.03
4	2.50	0.50	0.01	3.00	2.50	0.01	25.00	55.00	5.00	40.00	13.35
5	5.50	1.00	6.00	1.10	2.50	8.00	16.50	29.50	20.00	43.50	13.36
6	1.00	5.00	5.00	5.00	5.00	5.00	10.00	5.00	50.00	75.00	16.60
7	5.00	0.10	5.00	5.00	50.00	10.00	50.00	25.00	50.00	75.00	27.51
8	5.00	10.00	5.00	5.00	25.00	75.00	50.00	75.00	75.00	75.00	40.00
9	17.50	25.00	42.50	50.00	37.50	95.00	62.50	95.00	95.00	95.00	61.50
Mean	4.20	4.77	7.34	9.57	14.00	21.76	24.17	34.81	37.22	49.33	20.72

Source: Wedderburn (1974) taken from an unpublished Aberystwyth Ph.D thesis by J.F. Jenkyn.

from the linear logistic model. The residual deviance is 6.13 on 72 degrees of freedom and Pearson's statistic is equal to 6.39. Thus the estimate of σ^2 is $\tilde{\sigma}^2 = 6.39/72 = 0.089$. Since the data do not involve counts there is no reason to expect σ^2 to be near 1.0.

The estimated variety effects together with their standard errors, are shown below:

Variety									
1	2	3	4	5	6	7	8	9	10
0.00	0.15	0.69	1.05	1.62	2.37	2.57	3.34	3.50	4.25
(0.00)	(0.72)	(0.67)	(0.65)	(0.63)	(0.61)	(0.61)	(0.60)	(0.60)	(0.60)

Since these are simple contrasts with variety 1, the correlations among the estimates are approximately equal to 1/2. The actual correlations range from 0.68 to 0.83, which are larger than expected because variety 1 has a larger variance on the logistic scale than the other varieties. Evidently varieties 1–3 are most resistant to leaf blotch and varieties 8–10 least resistant.

In fact, however, as is shown in Fig. 9.1, the variance function $\mu(1-\mu)$ is not a satisfactory description of the variability in these data for very small or very large proportions. The variability observed in these plots is smaller at the extreme proportions than that predicted by the binomial variance function. Following Wedderburn's suggestion, we try an alternative variance function of the form $\mu^2(1-\mu)^2$ to mimic this effect. The resulting quasi-likelihood function can be obtained in closed form as

$$Q(\mu; y) = (2y-1)\log\left(\frac{\mu}{1-\mu}\right) - \frac{y}{\mu} - \frac{1-y}{1-\mu}.$$

Unfortunately this function is not defined for $\mu = 0$ or $\mu = 1$. A deviance function cannot be defined in the usual way for the data in Table 9.1 because some of the observed proportions are zero.

A linear logistic model with the new variance function gives the following estimated variety effects and standard errors:

Variety									
1	2	3	4	5	6	7	8	9	10
0.00	−0.47	0.08	0.95	1.35	1.33	2.34	3.26	3.14	3.89
(0.00)	(0.47)	(0.47)	(0.47)	(0.47)	(0.47)	(0.47)	(0.47)	(0.47)	(0.47)

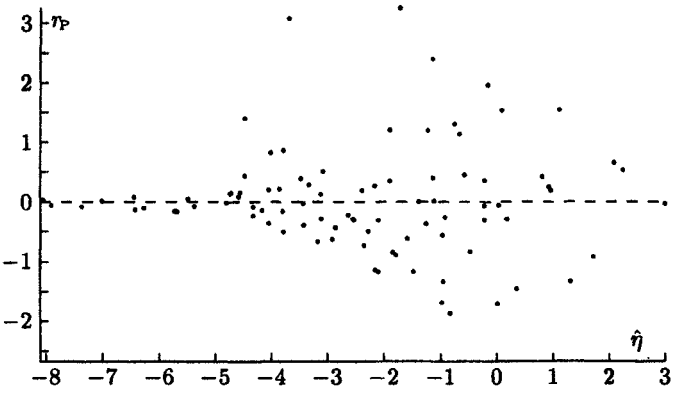

Fig. 9.1a. *Pearson residuals plotted against the linear predictor* $\hat{\eta} = \log(\hat{\pi}/(1-\hat{\pi}))$ *for the 'binomial' fit to the leaf-blotch data.*

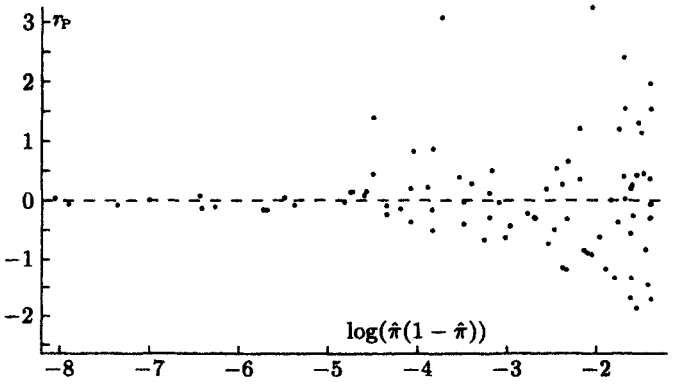

Fig. 9.1b. *Pearson residuals plotted against the logarithm of the variance function* $\hat{\pi}(1-\hat{\pi})$ *for the 'binomial' fit to the leaf-blotch data.*

The estimated dispersion parameter, obtained from the residual weighted mean square, is now $\tilde{\sigma}^2 = 71.2/72 = 0.99$, which differs very slightly from Wedderburn's value. The correlations among these estimates are exactly 1/2 because the iterative weights are exactly unity, and the analysis is effectively orthogonal. All variety contrasts in this model have equal standard error. Note that the ordering of varieties in the revised analysis differs slightly from the previous ordering. The principal difference between the two analyses, however, is that variety contrasts for which the incidence

is low are now estimated with greater apparent precision than in the previous model. Variety contrasts for which the incidence is high have reduced apparent precision.

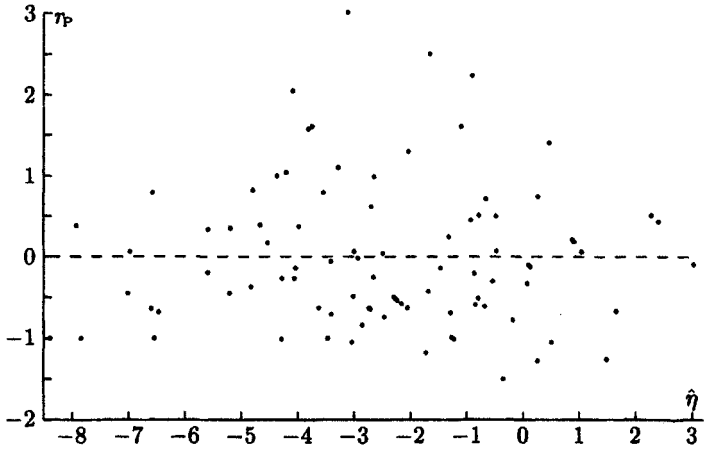

Fig. 9.2. *Pearson residuals using the variance function* $\pi^2(1-\pi)^2$ *plotted against the linear predictor* $\hat{\eta}$ *for the leaf-blotch data.*

The residuals are shown in Fig. 9.2 plotted against the linear predictor. To some extent the characteristic shape of Fig. 9.1a remains, though the effect is substantially diminished. Examination of individual residuals reveals three that are large and positive. These correspond, in decreasing order, to variety 4 at site 3 (3.01), variety 5 at site 7 (2.51), and variety 6 at site 8 (2.24). These residuals are computed by the formula $(y-\hat{\mu})/(\tilde{\sigma}\hat{\mu}(1-\hat{\mu}))$. There is no further evidence of systematic departures from the model.

9.3 Dependent observations

9.3.1 *Quasi-likelihood estimating equations*

Suppose now that $\text{cov}(\mathbf{Y}) = \sigma^2\mathbf{V}(\boldsymbol{\mu})$ where $\mathbf{V}(\boldsymbol{\mu})$ is a symmetric positive-definite $n\times n$ matrix of known functions $V_{ij}(\boldsymbol{\mu})$, no longer diagonal. The score function (9.5), with components $U_r(\boldsymbol{\beta})$, has the following properties:

$$E\{U_r(\boldsymbol{\beta})\} = 0,$$

$$\operatorname{cov}\{\mathbf{U}(\boldsymbol{\beta})\} = \mathbf{D}^T\mathbf{V}^{-1}\mathbf{D}/\sigma^2 = \mathbf{i}_\beta, \tag{9.8}$$
$$-E\Big(\frac{\partial U_r(\boldsymbol{\beta})}{\partial \beta_s}\Big) = \mathbf{D}^T\mathbf{V}^{-1}\mathbf{D}/\sigma^2.$$

Thus, for the reasons given in section 9.2.2, we may treat $\mathbf{U}(\boldsymbol{\beta})$ as if it were the derivative with respect to $\boldsymbol{\beta}$ of a log-likelihood function. Under suitable limiting conditions, the root $\hat{\boldsymbol{\beta}}$ of the estimating equation

$$\mathbf{U}(\hat{\boldsymbol{\beta}}) = \hat{\mathbf{D}}^T\hat{\mathbf{V}}^{-1}(\mathbf{Y} - \hat{\boldsymbol{\mu}}) = \mathbf{0}$$

is approximately unbiased for $\boldsymbol{\beta}$ and asymptotically Normally distributed with limiting variance

$$\operatorname{cov}(\hat{\boldsymbol{\beta}}) \simeq \sigma^2(\mathbf{D}^T\mathbf{V}^{-1}\mathbf{D})^{-1} = \mathbf{i}_\beta^{-1}.$$

The exact conditions required for consistency and asymptotic Normality of $\hat{\boldsymbol{\beta}}$ are rather complicated. Roughly speaking, however, it is necessary that as $n \to \infty$, $\mathbf{U}(\boldsymbol{\beta})$ should be asymptotically Normal, and the eigenvalues of $\mathbf{i}_\beta$ should tend to infinity for all $\boldsymbol{\beta}$ in an open neighbourhood of the true parameter point.

Block-diagonal covariance matrices arise most commonly in longitudinal studies, in which repeat measurements made on the same subject are usually positively correlated. Such applications are discussed by Liang and Zeger (1986) and Zeger and Liang (1986). These authors exploit the property that the quasi-likelihood estimate $\hat{\boldsymbol{\beta}}$ is often consistent even if the covariance matrix is misspecified. For a second example in which $\mathbf{V}$ is not block-diagonal, see section 14.5.

9.3.2 *Quasi-likelihood function*

Thus far, there is no essential difference between the discussion in section 9.2, for independent observations, and the more general case considered here. There is, however, one curious difference whose importance for inference is not entirely clear. If the score vector $\mathbf{U}(\boldsymbol{\beta})$ is to be the gradient vector of a log likelihood or quasi-likelihood it is necessary and sufficient that the derivative matrix of $\mathbf{U}(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$ be symmetric. In general, however, for $r \neq s$,

$$\frac{\partial U_r(\boldsymbol{\beta})}{\partial \beta_s} \neq \frac{\partial U_s(\boldsymbol{\beta})}{\partial \beta_r},$$

even though these matrices are equal in expectation to $-\mathbf{i}_\beta$. Consequently, unless some further conditions are imposed on the form of the matrix $\mathbf{V}(\boldsymbol{\mu})$, there can be no scalar function whose gradient vector is equal to $\mathbf{U}(\boldsymbol{\beta})$.

To state the same conclusion in a slightly more constructive way, the line integral

$$Q(\boldsymbol{\mu}; \mathbf{y}, \boldsymbol{t}(s)) = \sigma^{-2} \int_{\boldsymbol{t}(s)=\mathbf{y}}^{\boldsymbol{t}(s)=\boldsymbol{\mu}} (\mathbf{y} - \boldsymbol{t})^T \{\mathbf{V}(\boldsymbol{t})\}^{-1} \, d\boldsymbol{t}(s)$$

along a smooth path $\boldsymbol{t}(s)$ in R^n from $\boldsymbol{t}(s_0) = \mathbf{y}$ to $\boldsymbol{t}(s_1) = \boldsymbol{\mu}$, ordinarily depends on the particular path chosen. Evidently, if the integral is path-independent, the gradient vector of $Q(\boldsymbol{\mu}; \mathbf{y}, \cdot)$ with respect to $\boldsymbol{\mu}$ is $\mathbf{V}^{-1}(\mathbf{y} - \boldsymbol{\mu})/\sigma^2$, and the gradient vector with respect to $\boldsymbol{\beta}$ is given by (9.5). The derivative matrix of $\mathbf{U}(\boldsymbol{\beta})$ is then symmetrical. Conversely it can be shown that if the derivative matrix of $\mathbf{U}(\boldsymbol{\beta})$ is symmetrical, the integral is path-independent for all paths of the form $\boldsymbol{t}(s) = \boldsymbol{\mu}\big(\boldsymbol{\beta}(s)\big)$. Only if the line integral is independent of the path of integration does it make sense to use this function as a quasi-likelihood. Then, and only then, does quasi-likelihood estimation correspond to the maximization on the solution locus $\boldsymbol{\mu}(\boldsymbol{\beta})$ of a function defined pointwise for each $\boldsymbol{\mu} \in R^n$. We now investigate briefly the conditions on the covariance function that are required to make the integral path-independent.

The integral can be shown to be path-independent if the partial derivatives of the components of $\mathbf{V}^{-1}(\boldsymbol{\mu})$ with respect to $\boldsymbol{\mu}$ form an array that is symmetrical in all three directions—i.e. under permutations of the three indices. In other words, if $\mathbf{W} = \mathbf{V}^{-1}$, we require

$$\partial W_{ij}/\partial \mu_k = \partial W_{ik}/\partial \mu_j = \partial W_{jk}/\partial \mu_i.$$

A necessary and sufficient condition for this to hold is that $\mathbf{V}^{-1}(\boldsymbol{\mu})$ should be the second derivative matrix with respect to $\boldsymbol{\mu}$ of a scalar function $b^*(\boldsymbol{\mu})$, which is necessarily convex. The existence of a convex function $b^*(\boldsymbol{\mu})$ implies the existence of a canonical parameter $\boldsymbol{\theta}(\boldsymbol{\mu})$ and a cumulant function $b(\boldsymbol{\theta})$ defined on the dual space, such that

$$\boldsymbol{\theta} = b^{*\prime}(\boldsymbol{\mu}), \qquad \boldsymbol{\mu} = b'(\boldsymbol{\theta}) \quad \text{and} \quad \mathbf{V}(\boldsymbol{\mu}) = b''(\boldsymbol{\theta}). \tag{9.9}$$

These conditions exclude from consideration a large class of covariance functions that are physically unappealing for reasons discussed in section 9.2.1. However, some apparently physically sensible covariance functions are also excluded by the criterion.

In general it is not easy to construct covariance functions satisfying the above property even though the property itself is easy to verify. We now describe ways in which non-diagonal covariance functions satisfying (9.9) can be constructed. The integral can be shown to be path-independent if $\mathbf{V}^{-1}$ can be written as the sum of matrices, each having the form $\mathbf{A}^T\mathbf{W}(\boldsymbol{\gamma})\mathbf{A}$, in which $\mathbf{A}$ is independent of $\boldsymbol{\mu}$, $\boldsymbol{\gamma} = \mathbf{A}\boldsymbol{\mu}$, and $\mathbf{W}$ is diagonal of the form

$$\mathbf{W} = \operatorname{diag}\{W_1(\gamma_1), \ldots, W_n(\gamma_n)\}.$$

In other words we require a decomposition

$$\mathbf{V}^{-1}(\boldsymbol{\mu}) = \sum_{j=1}^{k} \mathbf{A}_j^T \mathbf{W}_j(\mathbf{A}_j\boldsymbol{\mu})\mathbf{A}_j \tag{9.10}$$

in which each $\mathbf{W}_j$ is a diagonal matrix of the required form as a function of its argument $\boldsymbol{\gamma}_j = \mathbf{A}_j\boldsymbol{\mu}$. In particular if $k = 1$ the covariance matrix can be diagonalized into the form (9.1). The latter condition can sometimes be verified directly for certain covariance functions—e.g. the multinomial covariance matrix.

In order to construct the quasi-likelihood function explicitly we may consider the straight-line path

$$\boldsymbol{t}(s) = \mathbf{y} + (\boldsymbol{\mu} - \mathbf{y})s$$

for $0 \leq s \leq 1$, so that $\boldsymbol{t}(0) = \mathbf{y}$ and $\boldsymbol{t}(1) = \boldsymbol{\mu}$. Provided that it exists, the quasi-likelihood function is given by

$$Q(\boldsymbol{\mu}; \mathbf{y}) = -(\mathbf{y} - \boldsymbol{\mu})^T \left\{ \sigma^{-2} \int_0^1 s \left\{ \mathbf{V}\big(\boldsymbol{t}(s)\big) \right\}^{-1} ds \right\} (\mathbf{y} - \boldsymbol{\mu}). \tag{9.11}$$

This integral is expressed directly in terms of the mean-value parameter and is sometimes useful for purposes of approximation. For example, if $\mathbf{V}^{-1}(\boldsymbol{t})$ is approximately linear in $\boldsymbol{t}$ over the straight-line path from $\boldsymbol{t} = \mathbf{y}$ to $\boldsymbol{t} = \boldsymbol{\mu}$, the integral may be approximated by

$$\begin{aligned} Q(\boldsymbol{\mu}; \mathbf{y}) \simeq & -\tfrac{1}{3}(\mathbf{y} - \boldsymbol{\mu})^T \mathbf{V}^{-1}(\boldsymbol{\mu})(\mathbf{y} - \boldsymbol{\mu})/\sigma^2 \\ & -\tfrac{1}{6}(\mathbf{y} - \boldsymbol{\mu})^T \mathbf{V}^{-1}(\mathbf{y})(\mathbf{y} - \boldsymbol{\mu})/\sigma^2. \end{aligned} \tag{9.12}$$

The alternative, and more familiar expression,

$$Q(\boldsymbol{\mu}; \mathbf{y}) = \sigma^{-2}\{\mathbf{y}^T\boldsymbol{\theta} - b(\boldsymbol{\theta}) - b^*(\mathbf{y})\}$$

presupposes the existence of the required functions, and is less useful for computational purposes unless those functions are available in a computable form.

The integral (9.11) can be evaluated whether or not $\mathbf{V}(\boldsymbol{\mu})$ satisfies the conditions in (9.9). However the use of $Q(\boldsymbol{\mu}; \mathbf{y})$ as a test statistic or as a quasi-likelihood function is then less compelling because an additional argument is required to justify the choice of the path of integration.

9.3.3 *Example: estimation of probabilities from marginal frequencies*

The following rather simplified example concerns the estimation of voter transition probabilities based only on the vote totals for each of two parties, C and L, say, in two successive elections. Suppose for simplicity of exposition that the same electorate votes in two successive elections and that we observe only the vote totals for C and L at each election. For each constituency, the 'complete data', which we do not observe, may be displayed as follows:

Votes cast		*Election 2*		
	Party	C	L	*Total*
Election 1	C	X_1	$m_1 - X_1$	m_1
	L	X_2	$m_2 - X_2$	m_2
	Total	$Y = X_.$	$m_. - X_.$	$m_.$

Only the row and column totals of the above Table are observed.

Since interest is focused on transition probabilities, we condition on the observed vote totals at election 1, regarding the entries in the body of the table as random variables. The simplest binomial model takes $X_1 \sim B(m_1, \pi_1)$ and $X_2 \sim B(m_2, \pi_2)$ as independent random variables. Thus π_1 is the probability that a voter who votes for party C in the first election also votes for C in the second election. Similarly π_2 is the probability that a voter who previously voted L subsequently switches to C.

Taking $Y = X_{\cdot}$, together with m_1, m_2 as the observed response in each constituency, we have that

$$E(Y) = m_1\pi_1 + m_2\pi_2 = \mu, \text{ say,}$$
$$\text{var}(Y) = m_1\pi_1(1-\pi_1) + m_2\pi_2(1-\pi_2).$$

Evidently var(Y) is not a function of $E(Y)$ alone, and hence it is not possible to construct a quasi-likelihood function along the lines described in sections 9.2.2 or 9.3.2. Nevertheless, given sufficient data from several constituencies, we may still use the score function (9.5) to estimate the parameters, which in this case are π_1 and π_2.

Suppose that data are available from each of n constituencies for which the transition probabilities may be assumed constant. Thus we have

$$E(\mathbf{Y}) = \mathbf{M}\boldsymbol{\pi},$$
$$\text{cov}(\mathbf{Y}) = \text{diag}\{m_{i1}\pi_1(1-\pi_1) + m_{i2}\pi_2(1-\pi_2)\} = \mathbf{V}(\boldsymbol{\pi}),$$

where $\mathbf{M}$ is an $n\times 2$ matrix giving the total votes cast for each party in the first election. The quasi-likelihood score function (9.5) is

$$\mathbf{U}(\boldsymbol{\pi}) = \mathbf{M}^T\mathbf{V}^{-1}(\boldsymbol{\pi})(\mathbf{Y} - \mathbf{M}\boldsymbol{\pi}). \tag{9.13}$$

The two components of this vector are

$$U_1(\boldsymbol{\pi}) = \sum_i m_{i1}(y_i - m_{i1}\pi_1 - m_{i2}\pi_2)/V_i(\boldsymbol{\pi})$$
$$U_2(\boldsymbol{\pi}) = \sum_i m_{i2}(y_i - m_{i1}\pi_1 - m_{i2}\pi_2)/V_i(\boldsymbol{\pi}).$$

Using these expressions it may be verified that

$$\frac{\partial U_1}{\partial \pi_2} \neq \frac{\partial U_2}{\partial \pi_1},$$

showing that (9.13) cannot be the gradient vector of any scalar function $Q(\boldsymbol{\pi})$. The information matrix $\mathbf{i}_{\boldsymbol{\pi}} = \mathbf{M}^T\mathbf{V}^{-1}\mathbf{M}$ has rank 2 provided that the vote ratios m_{i1}/m_{i2} are not all equal.

In order to compare the quasi-likelihood estimates with possible alternatives, we suppose that the following meagre vote totals are observed in three constituencies.

Y	m_1	m_2
7	5	5
5	6	4
6	4	6

After iteration in the usual way, the quasi-likelihood estimate obtained is $\hat{\boldsymbol{\pi}} = (0.3629, 0.8371)$. Thus the fitted values and the information matrix are

$$\hat{\boldsymbol{\mu}} = \mathbf{M}\hat{\boldsymbol{\pi}} = \begin{pmatrix} 6.000 \\ 5.526 \\ 6.474 \end{pmatrix} \quad \text{and} \quad \mathbf{i}_{\pi} = \begin{pmatrix} 41.4096 & 39.7904 \\ 39.7904 & 42.5357 \end{pmatrix}.$$

It is readily verified that these values satisfy the vector equation $\mathbf{U}(\hat{\boldsymbol{\pi}}) = \mathbf{0}$. The approximate standard errors of $\hat{\pi}_1$ and $\hat{\pi}_2$ are 0.489 and 0.482. The correlation, however, is given as -0.948, showing that the sum $\pi_1 + \pi_2$ is tolerably well estimated but there is little information concerning measures of difference such as $\pi_1 - \pi_2$, π_1/π_2 or $\psi = \pi_1(1-\pi_2)/\{\pi_2(1-\pi_1)\}$.

If all values in the above Table were increased by a factor of 100 the same parameter estimate and correlation matrix would be obtained. Standard errors of $\hat{\boldsymbol{\pi}}$ would be reduced by a factor of 10.

The likelihood function in this problem for a single observation y is

$$\sum_j \binom{m_1}{j}\binom{m_2}{y-j}\pi_1^j(1-\pi_1)^{m_1-j}\pi_2^{y-j}(1-\pi_2)^{m_2-y+j}.$$

The log likelihood for the full data, which is the sum of the logarithms of such factors, is numerically and algebraically unpleasant. It can, however, be maximized using the EM algorithm as described by Dempster *et al.* (1977). A simpler direct method is described in Exercise 9.2. We find $\hat{\boldsymbol{\pi}}_{\mathrm{ML}} = (0.2, 1.0)$, on the boundary of the parameter space and rather different from the quasi-likelihood estimate. In both cases, however, $\hat{\pi}_1 + \hat{\pi}_2 = 1.2$, a consequence of the identity $\sum y_i = \sum \hat{\mu}_i$.

The Fisher information matrix and its inverse, evaluated at the maximum quasi-likelihood estimate $\hat{\boldsymbol{\pi}} = (0.363, 0.837)$, are

$$\mathbf{I}_{\pi} = \begin{pmatrix} 62.18 & 4.57 \\ 4.57 & 102.24 \end{pmatrix} \quad \text{and} \quad \mathbf{I}_{\pi}^{-1} = \begin{pmatrix} 0.0161 & -0.0007 \\ -0.0007 & 0.0098 \end{pmatrix}.$$

Evidently in this case the maximum-likelihood estimator is considerably more efficient than the quasi-likelihood estimator, particularly for the estimation of differences. It is a curious fact that

the Fisher information matrix has rank 2 even when the matrix $\mathbf{M}$ has rank 1. Thus it is possible, in principle at least, for the quasi-likelihood estimates of certain contrasts to have negligible efficiency compared with the maximum-likelihood estimate. In both cases the estimated standard error of $\hat{\pi}_1 + \hat{\pi}_2$ is given as 0.1565. For further details see Exercises 9.2–9.3.

If all of the row totals m_{i1} and m_{i2} are equal across constituencies then $Y_1, \ldots, Y_n$ are identically distributed and $Y_\cdot$ is essentially sufficient for $\pi_\cdot$. It is still possible to estimate the odds ratio ψ because the variability of Y_i depends on ψ. It is this information, available in the likelihood function, that is discarded by the quasi-likelihood and accounts for the reduction in efficiency.

If all the observed values are increased by a factor of 100 the maximum-likelihood estimate changes to $\hat{\boldsymbol{\pi}}_{\mathrm{ML}} = (0.467, 0.733)$. The Fisher information matrix also changes in a moderately complicated way. Evidently the maximum-likelihood estimate is not a simple linear function of the data. This observation makes sense in the light of the comments in the previous paragraph.

9.4 Optimal estimating functions

9.4.1 *Introduction*

The quasi-score function (9.5) is a rather special case of what is known in Statistics as an estimating function. An estimating function $g(\mathbf{Y}; \boldsymbol{\theta})$ is a function of the data $\mathbf{Y}$ and parameter $\boldsymbol{\theta}$ having zero mean for all parameter values. Higher-order cumulants of $g(\cdot;\cdot)$ need not be independent of $\boldsymbol{\theta}$, so that $g(\mathbf{Y}; \boldsymbol{\theta})$ need not be a pivotal statistic. Provided that there are as many equations as parameters, estimates are obtained as the root of the equation $g(\hat{\boldsymbol{\theta}}; \mathbf{Y}) = \mathbf{0}$.

Usually it is fairly straightforward to construct estimating functions. For example, taking $\boldsymbol{\beta}$ to be the parameter of interest in the context described in sections 9.2 and 9.3, $\mathbf{Y} - \boldsymbol{\mu}(\boldsymbol{\beta})$ is a vector-valued estimating function. The difficult part is to reduce this n-vector to a p-vector with minimal sacrifice of information. The theory of optimal estimating equations can be used to demonstrate that the score function (9.5) is in fact the optimal combination within the class of linear estimating functions.

If the observations form a time-ordered sequence $Y_1, \ldots, Y_n$, it may be helpful to consider a corresponding sequence of elementary estimating functions of the form $g_t(Y_{(t)}; \boldsymbol{\theta})$, where $Y_{(t)} = (Y_1, \ldots, Y_t)$ is the process up to time t. If the conditional expectation of $g_t(\cdot\,;\cdot)$ given the past history of the process satisfies

$$E\{g_t(Y_{(t)}; \boldsymbol{\theta}) \mid Y_{(t-1)}\} = 0,$$

the cumulative sequence is said to form a martingale. Evidently $g_t(\cdot\,;\cdot)$ and all linear combinations of the gs are estimating functions. Thus there is a close connection between the theory of martingales and the theory of estimating functions (Godambe and Heyde, 1987).

In the linear regression and related contexts we usually require the elementary estimating functions to have zero mean conditionally on the values of the design points or covariates. This is a stronger condition than simply requiring zero unconditional mean. To underline the role played by such conditioning variables we write

$$E\{g_i(\mathbf{Y}; \boldsymbol{\theta}) \mid A_i\} = 0,$$

where $A_i \equiv A_i(\mathbf{y}; \boldsymbol{\theta})$, to cover both regression and time-series problems. In the regression context $A_i = A$, the set of covariates. More generally, however, the sequence A_i must be nested in the sense that $A_{i-1} \subseteq A_i$. Usually it is desirable to choose A to have maximum dimension.

A useful property of estimating functions is that very often they are rather simple functions of the data. For example (9.5) is linear in $\mathbf{Y}$. Statistical properties of the estimate, $\hat{\boldsymbol{\theta}}$, which is a non-linear function of $\mathbf{Y}$, can frequently be deduced from the properties of the estimating function. We now give a very brief outline of the theory of non-linear estimating functions, concentrating on ways of combining elementary estimating functions.

9.4.2 *Combination of estimating functions*

Suppose that the observed random variables $\mathbf{Y}$ have a distribution that depends on $\boldsymbol{\theta}$ and that, for each $\boldsymbol{\theta}$, $g_i(\mathbf{Y}; \boldsymbol{\theta})$ is a sequence of independent random variables having zero mean for all $\boldsymbol{\theta}$. For example if the Ys are generated by the autoregressive process

$$Y_t = \theta Y_{t-1} + \epsilon_t, \qquad Y_0 = \epsilon_0,$$

where ϵ_t are *i.i.d.* $N(0,1)$, we could take

$$g_t = Y_t - \theta Y_{t-1}$$

or, if $\theta \neq 0$,

$$g_t^* = Y_t/\theta - Y_{t-1}.$$

A second example in which the g_i are non-linear functions of $\mathbf{Y}$ is given in the following section.

In order to combine the n elementary estimating functions into a single p-dimensional optimal estimating equation for $\boldsymbol{\theta}$, we define the following $n \times p$ matrix, $\mathbf{D}$, with components D_{ir}, which depend on both $\boldsymbol{\theta}$ and $\mathbf{y}$.

$$D_{ir} = -E\Big\{\frac{\partial g_i(\mathbf{Y};\boldsymbol{\theta})}{\partial \theta_r} \mid A_i\Big\}. \tag{9.14}$$

If $\mathbf{V}$ is the (diagonal) conditional covariance matrix of g_i given A_i, we take as our estimating function for $\boldsymbol{\theta}$

$$\mathbf{U}(\boldsymbol{\theta};\mathbf{y}) = \mathbf{D}^T\mathbf{V}^{-1}\boldsymbol{g}. \tag{9.15}$$

Since D_{ir} and $\mathbf{V}$ are functions of the conditioning variables, and g_i has zero mean conditionally, it follows that $\mathbf{U}(\boldsymbol{\theta};\mathbf{y})$ also has zero conditional mean.

The above estimating function is unaffected by linear transformation of the elementary estimating functions $\boldsymbol{g}$ to $\boldsymbol{g}^* = \mathbf{B}\boldsymbol{g}$, where $\mathbf{B} \equiv \mathbf{B}(\boldsymbol{\theta})$ is a full-rank $n \times n$ matrix whose components may depend on A. Under this transformation we have

$$\boldsymbol{g}^* = \mathbf{B}\boldsymbol{g}, \qquad \mathbf{V}^* = \mathbf{B}\mathbf{V}\mathbf{B}^T, \qquad \mathbf{D}^* = \mathbf{B}\mathbf{D},$$

so that $\mathbf{D}^{*T}\mathbf{V}^{*-1}\boldsymbol{g}^* = \mathbf{D}^T\mathbf{V}^{-1}\boldsymbol{g}$ as claimed. Note that the components of $\boldsymbol{g}^*$ are not independent.

In the autoregressive process described above, this procedure applied to the sequence g_t gives

$$U(\theta;\mathbf{y}) = \sum_t Y_{t-1}g_t = \sum_t Y_{t-1}(Y_t - \theta Y_{t-1}).$$

When applied to the sequence g_t^* the same estimating function, which is also the log-likelihood derivative, is obtained by a slightly

more circuitous route. Note that, although g_t is linear in $\mathbf{Y}$, the final estimating function is quadratic in $\mathbf{Y}$.

By a modification of the argument given in section 9.2.3, the asymptotic variance of $\hat{\boldsymbol{\theta}}$ is

$$\mathbf{i}_{\theta}^{-1} = \text{cov}(\hat{\boldsymbol{\theta}}) = (\mathbf{D}^T\mathbf{V}^{-1}\mathbf{D})^{-1}.$$

Usually, and particularly so in the autoregressive model, it is better to use the observed value of this matrix rather than its expected value. The same recommendation applies in the example below.

Example: Fieller-Creasy problem

The following rather simple example is chosen to illustrate the fact that elementary estimating functions can frequently be chosen in such a way that incidental parameters are eliminated. Suppose that the observations come in pairs $\mathbf{Y}_i = (Y_{i1}, Y_{i2})$, which are independent with means $(\mu_i, \mu_i/\theta)$ and variances σ^2. It is assumed that $\theta = E(Y_{i1})/E(Y_{i2})$ is the parameter of interest. From the stated assumptions it follows that

$$g_i(\mathbf{Y}_i; \theta) = Y_{i1} - \theta Y_{i2}$$

has mean zero and variance $\sigma^2(1 + \theta^2)$. The derivative of g_i with respect to θ is $-Y_{i2}$. Application of (9.14), taking $A_i(\theta) = Y_{i2} + \theta Y_{i1}$, gives the residual derivative

$$\begin{aligned} D_i &= Y_{i2} + (Y_{i1} - \theta Y_{i2})\theta/(1 + \theta^2) \\ &= (Y_{i2} + \theta Y_{i1})/(1 + \theta^2). \end{aligned}$$

The estimating function for θ is therefore

$$U(\theta) = \sum_i \frac{(Y_{i2} + \theta Y_{i1})(Y_{i1} - \theta Y_{i2})}{\sigma^2(1 + \theta^2)^2},$$

which is identical to the conditional log likelihood score statistic (7.3).

This score function ordinarily has two roots, one at $\hat{\theta}$, the other at $-1/\hat{\theta}$. One of these corresponds to what would be considered a maximum of the log likelihood: the other corresponds to a minimum. In any case, Normal approximation for $\hat{\theta}$ may be unsatisfactory unless the information is large. The alternative method

of generating confidence sets directly from the score statistic is preferable. In other words we take the set

$$\{\theta : |U(\theta)/i^{1/2}(\theta)| \leq k^*_{\alpha/2}\},$$

where the observed information for θ may be taken to be

$$i(\theta) = \sum \frac{(y_{i2} + \theta y_{i1})^2}{\sigma^2(1+\theta^2)^3},$$

and $\Phi(k^*_\alpha) = 1 - \alpha$.

For an alternative argument leading to the same result see section 7.2.2.

9.4.3 *Example: estimation for megalithic stone rings*

Suppose that $\mathbf{Y}_i = (Y_{i1}, Y_{i2})$, $i = 1, \ldots, n$, are the observed Cartesian coordinates of points in a plane, assumed to lie on or near the circumference of a circle with centre at (ω_1, ω_2) and radius ρ. We have in mind archaeological applications in which the 'points' are standing stones forming a curve that is assumed here to be an arc of a circle rather than a complete circle. For a statistical model we assume that the points are generated in the following way:

$$\begin{aligned} Y_{i1} &= \omega_1 + R_i \cos \epsilon_i \\ Y_{i2} &= \omega_2 + R_i \sin \epsilon_i \end{aligned} \tag{9.16}$$

in which $R_1, \ldots, R_n$ are positive independent and identically distributed random variables with mean ρ. The quantities $\epsilon_1, \ldots, \epsilon_n$ may be regarded either as fixed (non-random) nuisance parameters or as random variables having a joint distribution independent of $\mathbf{R}$. It is undesirable and unnecessary in the archaeological context to assume that $\epsilon_1, \ldots, \epsilon_n$ are identically distributed or mutually independent. Usually the stones are roughly equally spaced around the circle, or what remains of the circle, so the angles cannot be regarded as independent.

In order to construct an estimating equation for the parameters $(\omega_1, \omega_2, \rho)$ we observe first that, under the assumptions stated above,

$$\begin{aligned} g_i &= \{(Y_{i1} - \omega_1)^2 + (Y_{i2} - \omega_2)^2\}^{1/2} - \rho \\ &= R_i(\omega_1, \omega_2) - \rho \end{aligned}$$

are independent and identically distributed with mean zero conditionally on $A = (\epsilon_1, \ldots, \epsilon_n)$.

The derivative vector of g_i with respect to $(\omega_1, \omega_2, \rho)$ is equal to $(\cos\epsilon_i, \sin\epsilon_i, 1)$, which is independent of g_i by assumption. Consequently the estimating functions for the three parameters are

$$\begin{aligned} &\sum \frac{Y_{i1} - \omega_1}{\rho^2 R_i}(R_i - \rho) = \cos\epsilon_i \times (R_i - \rho)/\rho^2 \\ &\sum \frac{Y_{i2} - \omega_2}{\rho^2 R_i}(R_i - \rho) = \sin\epsilon_i \times (R_i - \rho)/\rho^2 \qquad (9.17) \\ &\sum (R_i - \rho)/\rho^2, \end{aligned}$$

where we have taken $\mathrm{var}(R_i) = \sigma^2\rho^2$. Under the assumption that the ϵs are independent of $\mathbf{R}$, or at least that $\cos\epsilon_i$ and $\sin\epsilon_i$ are uncorrelated with $\mathbf{R}$, these three functions all have zero mean. These equations are in fact the ordinary least-squares equations obtained by minimizing the sum of squares of the radial errors ignoring the angular errors.

We take as our estimate of $\rho^2\sigma^2$ the mean squared radial error, namely

$$\hat{\rho}^2\tilde{\sigma}^2 = \sum (\hat{R}_i - \hat{\rho})^2/(n-3).$$

Note in this case that it would be a fatal mistake to use the unconditional expected value of the derivatives of g_i in the definition of the coefficient matrix (9.14). If the angles are identically distributed, not necessarily uniformly, around the circle then $E(\cos\epsilon)$ and $E(\sin\epsilon)$ are both constant. The resulting estimating equations then have rank 1.

To illustrate this method of estimation we consider the data in Table 9.3, taken from Angell and Barber (1977). The Avebury ring has been studied by a number of authors including Thom (1967) and Thom, Thom and Foord (1976) who have divided the stones into four major groups, labelled A, B, C and W. From the diagram in Fig. 9.3 it can be seen that each of the individual arcs is quite shallow and that arc W is not immediately distinguishable from a straight line.

Table 9.3 gives the fitted centres and radii for each of the arcs considered separately. The final line gives the residual mean squared radial error, using degrees of freedom rather than sample

Table 9.3 *Stone number and position in the Avebury ring*[†]

Arc C			*Arc W*			*Arc A*			*Arc B*		
No.	x	y	*No.*	x	y	*No.*	x	y	*No.*	x	y
1	733.7	44.0	9	445.3	23.4	30	19.3	624.4	40	146.8	936.9
3	659.7	28.0	10	413.8	46.2	31	24.9	663.0	41	175.2	962.4
4	624.2	19.3	11	377.9	74.1	32	33.3	698.3	42	206.7	984.7
5	588.4	13.9	12	357.1	94.1	33	43.7	731.3	43	237.6	1002.9
6	551.6	12.3	13	327.7	112.4	34	55.5	764.4	44	270.3	1022.5
7	515.1	9.5	14	300.6	136.2	35	62.9	790.1	45	292.5	1031.2
8	478.0	16.6	15	272.0	158.8	36	69.2	815.0	46	315.8	1042.0
			16	243.5	183.0	37	85.0	849.8			
			17	216.3	205.0	38	98.5	884.6			
			18	188.9	229.8	39	123.6	910.5			
			19	163.5	255.5						
			20	140.0	285.0						
			21	120.6	305.7						
			22	103.1	323.1						
			23	85.9	344.0						
			24	61.8	371.3						
$\hat{\omega}_1 = 530.8$			$\hat{\omega}_1 = 1472.0$			$\hat{\omega}_1 = 795.0$			$\hat{\omega}_1 = 512.7$		
$\hat{\omega}_2 = 651.0$			$\hat{\omega}_2 = 1553.4$			$\hat{\omega}_2 = 516.5$			$\hat{\omega}_2 = 533.1$		
$\hat{\rho} = 638.8$			$\hat{\rho} = 1840.4$			$\hat{\rho} = 782.8$			$\hat{\rho} = 545.4$		
$\hat{\rho}^2\tilde{\sigma}^2 = 5.60$			$\hat{\rho}^2\tilde{\sigma}^2 = 3.78$			$\hat{\rho}^2\tilde{\sigma}^2 = 9.00$			$\hat{\rho}^2\tilde{\sigma}^2 = 0.72$		

[†]Data taken from Angell and Barber (1977).

size as divisor. In order to test whether the data are consistent with a single circle for arcs A, B and C, we fitted a circle to these three arcs together. The position of the fitted centre is shown in Fig. 9.3. The residual sum of squared radial errors is 878.8 on 21 degrees of freedom, whereas the pooled residual sum of squares from the separate fits is

$$4 \times 5.60 + 7 \times 9.00 + 4 \times 0.72 = 88.3$$

on 15 degrees of freedom. The increase in residual sum of squares is clearly statistically highly significant, showing that the three arcs are not homogeneous.

When models are fitted to shallow arcs, the parameter estimates tend to be highly correlated. For example the standard errors and

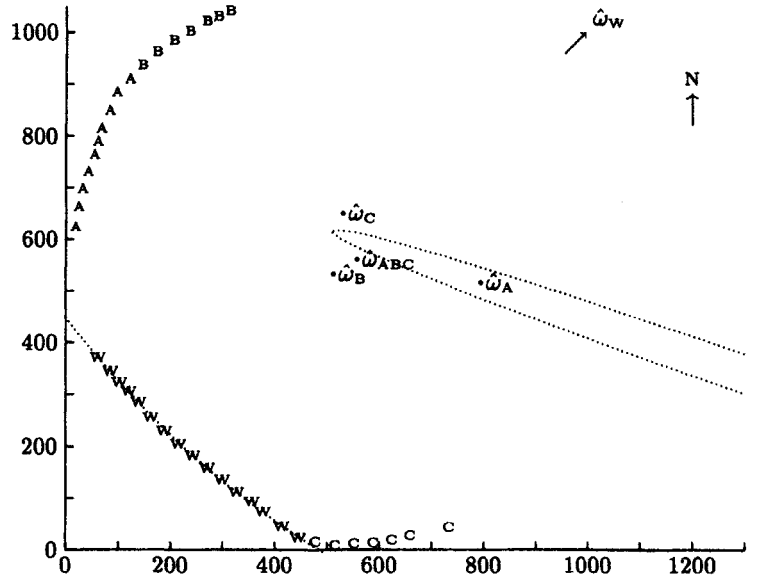

Fig. 9.3 *Diagram of the Avebury ring showing the fitted centres of arcs A, B and C, together with the joint fit and the fitted arc W. An approximate 99% confidence set is shown for the centre of arc A. Distances in feet.*

correlation matrix for arc C are

$$\begin{matrix} \text{s.e.}(\hat{\rho}) = 108.4 \\ \text{s.e.}(\hat{\omega}_1) = 14.57 \\ \text{s.e.}(\hat{\omega}_2) = 108.6 \end{matrix} \quad \text{and} \quad \begin{pmatrix} 1.0 & & \\ -0.87883 & 1.0 & \\ 0.99995 & -0.87584 & 1.0 \end{pmatrix}.$$

A confidence set for the position of the centre would be approximately elliptical with major axis almost vertical and very much larger than the minor axis. Such a 99% confidence set for ω_A, corresponding to RSS $\leq$ 181, is shown in Fig. 9.3. The correlations are considerably smaller when the three arcs are combined. If the arc were a complete circle the correlations would be nearly zero.

This example has been chosen by way of illustration. In the archaeological context it is difficult to take the standard errors literally, but they may be useful in a qualitative way. The main statistical assumption, that the radial errors are independent, is not supported by residual plots of fitted residuals against stone

number. There is a noticeable but complicated residual pattern of positive serial dependence for arc W.

A version of model (9.16) for fitting ellipses is described in Exercise 9.5.

The estimates given here are the same as those obtained by Freeman (1977), but different from the two sets of estimates given by Thom, Thom and Foord (1976) and Angell and Barber (1977). For a comparison of various alternative methods of estimation from the viewpoint of consistency, efficiency and so on, see Berman and Griffiths (1985), Berman and Culpin (1986) and Berman (1987). A related, but less tractable, model for fitting circles is discussed by Anderson (1981).

9.5 Optimality criteria

In order to justify the optimality claims made on behalf of the estimating functions (9.5) and (9.15) it is essential to state clearly the criterion used for making comparisons and also to state the class of estimators within which (9.5) and (9.15) are optimal. Evidently from the discussion in section 9.3.3 the quasi-likelihood estimate is sometimes less efficient than maximum likelihood, so that claims for optimality, even asymptotic optimality, cannot be global.

In keeping with the major theme of this chapter we consider first the class of linear estimating functions

$$\mathbf{h}(\mathbf{y};\boldsymbol{\beta}) = \mathbf{H}^T\big(\mathbf{y} - \boldsymbol{\mu}(\boldsymbol{\beta})\big) \tag{9.18}$$

where $\mathbf{H}$, of order $n \times p$, may be a function of $\boldsymbol{\beta}$ but not of $\mathbf{y}$. Clearly $\mathbf{h}(\mathbf{y};\boldsymbol{\beta})$ is linear in $\mathbf{y}$ for each $\boldsymbol{\beta}$. However the estimate $\tilde{\boldsymbol{\beta}}$, here assumed unique, satisfying $\mathbf{h}(\mathbf{y};\tilde{\boldsymbol{\beta}}) = \mathbf{0}$, is ordinarily non-linear in $\mathbf{y}$. We now demonstrate that, asymptotically at least, all linear functions of $\tilde{\boldsymbol{\beta}}$ have variance at least as great as the variance of the same linear function of $\hat{\boldsymbol{\beta}}$, i.e. $\mathrm{var}(\mathbf{a}^T\tilde{\boldsymbol{\beta}}) \geq \mathrm{var}(\mathbf{a}^T\hat{\boldsymbol{\beta}})$ where $\hat{\boldsymbol{\beta}}$ is the root of (9.5).

Under the usual asymptotic regularity conditions we may expand the estimating function in a Taylor series about the true parameter point giving

$$\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta} \simeq (\mathbf{H}^T\mathbf{D})^{-1}\mathbf{h}(\mathbf{y};\boldsymbol{\beta}),$$

where $\mathbf{H}$ and $\mathbf{D}$ are evaluated at the true parameter point. Thus the asymptotic covariance matrix of $\tilde{\boldsymbol{\beta}}$ is

$$\text{cov}(\tilde{\boldsymbol{\beta}}) \simeq \sigma^2(\mathbf{H}^T\mathbf{D})^{-1}\mathbf{H}^T\mathbf{V}\mathbf{H}(\mathbf{D}^T\mathbf{H})^{-1},$$

where $D_{ir} = \partial\mu_i/\partial\beta_r$. The claim for asymptotic optimality of the quasi-likelihood estimate rests on the fact that the matrix

$$\text{cov}(\tilde{\boldsymbol{\beta}}) - \text{cov}(\hat{\boldsymbol{\beta}}) \simeq \sigma^2(\mathbf{H}^T\mathbf{D})^{-1}\mathbf{H}^T\mathbf{V}\mathbf{H}(\mathbf{D}^T\mathbf{H})^{-1} - \mathbf{i}_{\beta}^{-1}$$

is non-negative definite for all $\mathbf{H}$. In order to demonstrate this claim we need only show that the difference between the precision matrices

$$\{\text{cov}(\hat{\boldsymbol{\beta}})\}^{-1} - \{\text{cov}(\tilde{\boldsymbol{\beta}})\}^{-1}$$

is non-negative definite (Exercise 9.7). This difference is equal to

$$\mathbf{D}^T\big(\mathbf{V}^{-1} - \mathbf{H}(\mathbf{H}^T\mathbf{V}\mathbf{H})^{-1}\mathbf{H}^T\big)\mathbf{D},$$

which is the residual covariance matrix of $\mathbf{D}^T\mathbf{V}^{-1}\mathbf{Y}$ after linear regression on $\mathbf{H}^T\mathbf{Y}$, and hence is non-negative definite. This completes the proof, that for large n, $\text{cov}(\tilde{\boldsymbol{\beta}}) \geq \text{cov}(\hat{\boldsymbol{\beta}})$ with the usual strong partial ordering on positive-definite matrices. The covariance matrices are equal only if $\mathbf{H}$ is expressible as a linear combination of the columns of $\mathbf{V}^{-1}\mathbf{D}$.

The proof just sketched is formally identical to a proof of the Gauss-Markov theorem for linear estimators. The strength of this proof is that it applies to a considerably wider class of estimators than does the Gauss-Markov theorem: its weakness is that it is an asymptotic result, focusing on $\hat{\boldsymbol{\beta}}$ rather than on the score function directly.

The proof just given applies equally to the so-called non-linear estimating function (9.15) provided that we agree to make all probability calculations appropriately conditional. Provided that $A_i = A$, the same for each i, (9.15) is conditionally linear in $\boldsymbol{g}$. In other words $\hat{\boldsymbol{\beta}}$ is asymptotically conditionally optimal given A within the class of estimating functions that are conditionally linear in $\boldsymbol{g}$. One difficulty with this criterion is that there may well be some ambiguity regarding the best choice of A. The theory offers little guidance in this respect.

A variety of optimality conditions, including both fixed-sample and asymptotic criteria, are discussed by Godambe and Heyde (1987).

9.6 Extended quasi-likelihood

The discussion in sections 9.2, 9.3 has been concerned entirely with the fitting and comparison of regression models in which the variance function is known. The quasi-likelihood function (9.3) or, if it exists, (9.11), cannot be used for the formal comparison of models having different variance functions or different dispersion parameters. The properties listed in section 9.3.1 refer only to derivatives with respect to $\boldsymbol{\beta}$ and not with respect to σ^2.

We have seen in section 9.2.4 how different variance functions can be compared graphically. The purpose of this section is to supplement and to formalize such comparisons via an extended quasi-likelihood. Then pairs of residual plots such as Figs 9.1a and 9.2 may be compared numerically via the extended quasi-likelihood as well as visually. The introduction of such an extended quasi-likelihood also opens up the possibility of modelling the dispersion parameter as a function of covariates (see Chapter 10).

In order to avoid difficulties of the kind encountered in section 9.3, we assume here that the observations are independent. The extended quasi-likelihood is then a sum over the n components of $\mathbf{y}$. For a single observation y we seek to construct a function $Q^+(\mu, \sigma^2; y)$ that, for known σ^2, is essentially the same as $Q(\mu; y)$, but which exhibits the properties of a log likelihood with respect to σ^2-derivatives. Thus we must have

$$\begin{aligned} Q^+(\mu, \sigma^2; y) &= Q(\mu; y) + h(\sigma^2; y) \\ &= -\frac{D(y; \mu)}{2\sigma^2} + h(\sigma^2; y) \end{aligned}$$

for some function $h(\sigma^2; y)$, which we take to be of the form

$$h(\sigma^2; y) = -\tfrac{1}{2} h_1(\sigma^2) - h_2(y).$$

In order for Q^+ to behave like a log likelihood with respect to σ^2 we must have $E(\partial Q^+/\partial \sigma^2) = 0$. Thus

$$0 = \frac{1}{2\sigma^4} E\{D(Y; \mu)\} - \tfrac{1}{2} h_1'(\sigma^2),$$

implying that

$$\sigma^4 h_1'(\sigma^2) = E\{D(Y; \mu)\}. \tag{9.19}$$

To a rough first order of approximation we have $E(D(Y;\mu)) = \sigma^2$, giving $h_1(\sigma^2) = \log(\sigma^2) + \text{const}$. Thus the extended quasi-likelihood function is given by

$$Q^+(\mu, \sigma^2; y) = -\tfrac{1}{2}D(y;\mu)/\sigma^2 - \tfrac{1}{2}\log\sigma^2. \tag{9.20}$$

This expression can be justified to some extent as a saddlepoint approximation for the log density provided that σ^2 is small and all higher-order cumulants are sufficiently small. To be explicit, suppose that the higher-order cumulants of Y are given by

$$\kappa_{r+1} = \kappa_r'\kappa_2, \quad \text{for } r \geq 2, \tag{9.21}$$

where $\kappa_2 = \sigma^2 V(\mu)$, and differentiation is with respect to μ. As shown in Exercise 2.7, (9.21) is a property of exponential-family distributions and of averages from such distributions in which σ^{-2} is an effective sample size. Thus $\kappa_3 = O(\sigma^4)$, $\kappa_4 = O(\sigma^6)$ and so on. The saddlepoint approximation for the log density is then

$$-\tfrac{1}{2}D(y;\mu)/\sigma^2 - \tfrac{1}{2}\log\bigl(2\pi\sigma^2 V(y)\bigr),$$

which differs from Q^+ by an additive function of y. See, for example, Barndorff-Nielsen and Cox (1979), Nelder and Pregibon (1987), Efron (1986), Jørgensen (1987) or McCullagh (1987, Chapter 6). Note that saddlepoint approximations depend on the entire set of cumulants, and not just on the low-order cumulants.

More accurate approximations can be obtained for $E(D(Y;\mu))$ provided that information is available concerning higher-order cumulants of Y. To a certain extent, however, this requirement violates the spirit of least squares, which is based on first and second moment assumptions only. Using the representation (9.4) it can be shown that

$$E\bigl(D(Y;\mu)\bigr) \simeq \sigma^2 + \frac{1}{12V^2}\bigl\{6\sigma^4 VV'^2 - 3\sigma^4 V^2 V'' - 4V'\kappa_3\bigr\}. \tag{9.22}$$

If (9.21) can be justified up to order 4, this expression may be reduced to

$$\begin{aligned} E\bigl(D(Y;\mu)\bigr) &\simeq \sigma^2\bigl\{1 + (5\rho_3^2 - 3\rho_4)/12\bigr\}, \\ &= \sigma^2\bigl\{1 + \sigma^2(2V'^2/V - 3V'')/12\bigr\}, \end{aligned}$$

where the standardized cumulants $\rho_3^2 = \kappa_3^2/\kappa_2^3$ and $\rho_4 = \kappa_4/\kappa_2^2$ are both $O(\sigma^2)$. By the same argument we find

$$\begin{aligned} \operatorname{var}(D) &\simeq 2\kappa_2^2/V^2 = 2\sigma^4 \\ \operatorname{cov}(D, Y) &\simeq (\kappa_3 - \kappa_2\kappa_2')/V. \end{aligned}$$

The approximate covariance reduces to zero under the simplifying assumption (9.21) but not otherwise.

In what follows we assume that σ^2 is sufficiently small to justify the approximation $E(D) \simeq \sigma^2$. It follows then that the derivatives

$$\begin{aligned} \frac{\partial Q^+}{\partial \mu} &= \frac{Y - \mu}{\sigma^2 V(\mu)} \\ \frac{\partial Q^+}{\partial \sigma^2} &= \frac{D(Y; \mu)}{2\sigma^4} - \frac{1}{2\sigma^2} \end{aligned}$$

have zero mean and approximate covariance matrix

$$\begin{pmatrix} \dfrac{1}{\sigma^2 V(\mu)} & \dfrac{\kappa_3 - \kappa_2\kappa_2'}{2\sigma^6 V^2} \\ \dfrac{\kappa_3 - \kappa_2\kappa_2'}{2\sigma^6 V^2} & \dfrac{1}{2\sigma^4} \end{pmatrix}.$$

The expected value of the negative second derivative matrix is the same as the above except that the off-diagonal elements are zero. Note that if

$$\kappa_3 - \kappa_2\kappa_2' = O(\sigma^4)$$

for small σ, the correlation of the two derivatives is $O(\sigma)$, and hence negligible. Consequently, with this entirely reasonable condition, Q^+ has the properties of a quasi-likelihood with respect to both mean parameter and dispersion parameter. Further, the Fisher information matrix for (μ, σ^2) is diagonal, a property that simplifies some calculations.

The argument just given is a partial justification for the use of the extended quasi-likelihood for the joint parameter (μ, σ^2). The assumptions required are that σ^2 be small and that $\kappa_r(Y) = O(\sigma^{2(r-1)})$. Efron (1986) and Jørgensen (1987), using the stronger assumption (9.21), reach similar conclusions.

9.7 Bibliographic notes

The term quasi-likelihood seems first to have been used in this context by Wedderburn (1974) although calculations similar to those in section 9.2 appear also in unpublished work by Jarrett (1973).

Questions of efficiency and optimality, and to a lesser extent robustness, are addressed by Godambe (1960), Bhapkar (1972), Cox and Hinkley (1968), Morton (1981), Cox (1983), McCullagh (1983, 1984), Firth (1987), Godambe and Heyde (1987) and Hill and Tsai (1988).

Estimates having increased efficiency can sometimes be obtained by considering non-linear estimating functions or by combining two or more estimating functions. This subject has been studied by Jarrett (1973) and subsequently by Crowder (1987), Firth (1987) and Heyde (1987).

The problem discussed in section 9.3.3 has previously been studied by Firth (1982).

9.8 Further results and exercises 9

9.1 Suppose, conditionally on $M = m$ that $Y \sim P(m)$, the Poisson distribution with parameter m, and that M in turn has the gamma distribution $M \sim G(\alpha\nu, \nu)$ with mean $\mu = E(M) = \alpha\nu$ and coefficient of variation $\nu^{-1/2}$. Show that the unconditional mean and variance are $E(Y) = \mu = \alpha\nu$, and

$$\text{var}(Y) = \alpha\nu + \alpha^2\nu.$$

Suppose now that $\mathbf{Y}$ has independent components generated in the above way with $\mu_i = E(Y_i)$ not all equal. Show that if $\nu_i = \nu$, a known constant, then the distribution of $\mathbf{Y}$ has the natural exponential-family form with variance function $V(\mu) = \mu + \mu^2/\nu$, which is quadratic in μ. On the other hand if $\alpha_i = \alpha$, a constant, show that the variance function has the standard over-dispersed Poisson form $V(\mu) = \phi\mu$ with $\phi = 1 + \alpha$, but that $\mathbf{Y}$ does not then have the linear exponential-family form.

More generally, if both α and ν vary according to the relations

$$\alpha_i = \theta + \psi\mu_i; \qquad \nu_i^{-1} = \psi + \theta\mu_i^{-1},$$

show that $V(\mu) = \mu + \theta\mu + \psi\mu^2$ and that the distribution of $\mathbf{Y}$ again does not have the linear exponential-family form. Compare the exact likelihood with the corresponding quasi-likelihood in the second and third cases.

9.2 Show that the likelihood function given at the end of section 9.3.3 can be written in the form

$$(1-\pi_1)^{m_1}\pi_2^{y}(1-\pi_2)^{m_2-y}P_0(\psi; m_1, m_2, y),$$

where $P_0(\psi)$ is defined in section 7.3.2. Hence show that the log-likelihood derivatives with respect to $\lambda_i = \text{logit}(\pi_i)$ are

$$\frac{\partial l}{\partial \lambda_1} = \sum_i \{\kappa_1(\psi) - m_1\pi_1\},$$
$$\frac{\partial l}{\partial \lambda_2} = \sum_i \{y - \kappa_1(\psi) - m_2\pi_2\},$$

where $\kappa_1(\psi)$ is the non-central hypergeometric mean, and summation runs over constituencies. Interpret these equations in terms of the EM algorithm.

9.3 Deduce that the maximum-likelihood estimates in the previous exercise satisfy $\sum y = \sum(m_1\hat{\pi}_1 + m_2\hat{\pi}_2)$. Show that the Fisher information matrix for $\boldsymbol{\theta} = (\pi_\cdot, \lambda_1 - \lambda_2)$ is a sum over constituencies of matrices of the form

$$\mathbf{I}_\theta = \begin{pmatrix} m_1V_1 + m_2V_2 & (m_1 - m_2)V_1V_2 \\ (m_1 - m_2)V_1V_2 & (m_1V_2 + m_2V_1)V_1V_2 - \kappa_2V_\cdot^2 \end{pmatrix} \frac{1}{V_\cdot^2},$$

where $V_1 = \pi_1(1-\pi_1)$, $V_2 = \pi_2(1-\pi_2)$, and κ_2 is the hypergeometric variance, which depends on ψ. Under what conditions are these parameters orthogonal? Deduce that the Fisher information matrix has rank 2 even if $m_{i1} = m_{i2}$ for each i, but that $\lambda_1 - \lambda_2$ is not consistently estimated unless $\mathbf{X}$ has rank 2.

9.4 Show that the integral along the straight-line path $\boldsymbol{t}(s)$ from $\boldsymbol{t}(s_0) = \mathbf{b}$ to $\boldsymbol{t}(s_1) = \mathbf{c}$ of the tangential component of the vector $\mathbf{A}^T\boldsymbol{t}$ is given by

$$\int_b^c \boldsymbol{t}^T\mathbf{A}\,d\boldsymbol{t}(s) = \tfrac{1}{2}(\mathbf{c}+\mathbf{b})^T\mathbf{A}(\mathbf{c}-\mathbf{b}).$$

Find the value of the integral along the path from **b** to **c** to **d** and back to **b**. Hence deduce that if $\mathbf{A} = \mathbf{A}^T$ the integral around the loop is zero. Conversely, deduce that if the integral around every closed loop is zero then **A** must be symmetric.

9.5 Consider the model

$$Y_{i1} = \omega_1 + \rho R_i \cos\epsilon_i \cos\phi - \lambda R_i \sin\epsilon_i \sin\phi$$
$$Y_{i2} = \omega_2 + \rho R_i \cos\epsilon_i \sin\phi + \lambda R_i \sin\epsilon_i \cos\phi$$

for an ellipse centered at (ω_1, ω_2) with semi-axes of length ρ, λ inclined at an angle ϕ to the x-axis. Assume that R_i are independent and identically distributed with mean 1, and independently of the ϵs. Using the method described in section 9.4.3, construct an unbiased estimating function for the parameters $(\omega_1, \omega_2, \rho, \lambda, \phi)$.

Take as the elementary estimating functions

$$R_i - 1 = \left(\frac{X_{i1}^2}{\rho^2} + \frac{X_{i2}^2}{\lambda^2}\right)^{1/2} - 1,$$

where

$$X_{i1} = (Y_{i1} - \omega_1)\cos\phi + (Y_{i2} - \omega_2)\sin\phi = \rho R_i \cos\epsilon_i$$
$$X_{i2} = -(Y_{i1} - \omega_1)\sin\phi + (Y_{i2} - \omega_2)\cos\phi = \lambda R_i \sin\epsilon_i.$$

Show that the required coefficients (9.14) are

$$D_{i1} = \cos\epsilon_i \cos\phi/\rho - \sin\epsilon_i \sin\phi/\lambda,$$
$$D_{i2} = \cos\epsilon_i \sin\phi/\rho + \sin\epsilon_i \cos\phi/\lambda,$$
$$D_{i3} = \cos^2\epsilon_i/\rho,$$
$$D_{i4} = \sin^2\epsilon_i/\lambda,$$
$$D_{i5} = (\rho - \lambda)\cos\epsilon_i \sin\epsilon_i.$$

Hence compute the information matrix for the five parameters.

9.6 Suppose that the covariance matrix **V** can be written in the form

$$\mathbf{V} = \mathbf{DRD},$$

where **R** is independent of $\boldsymbol{\mu}$ and

$$\mathbf{D} = \text{diag}\{D_1(\mu_1), \ldots, D_n(\mu_n)\}.$$

Show that the quasi-likelihood function exists only if V_{ij} is independent of $\boldsymbol{\mu}$ for all $i \neq j$. In other words **R** must be diagonal or **D** must be independent of $\boldsymbol{\mu}$. [Section 9.3.2].

9.7 Let $\mathbf{A}$ and $\mathbf{B}$ be any two positive-definite matrices of the same order. Prove that

$$\mathbf{A} - \mathbf{B} \geq 0 \quad \text{implies} \quad \mathbf{A}^{-1} - \mathbf{B}^{-1} \leq 0,$$

where ≥ 0 means non-negative definite.

9.8 Suppose that the random variables $Y_1, \ldots, Y_n$ are independent with variance $\text{var}(Y_i) = \sigma^2 \mu_i^2$, where the coefficient of variation, σ, is unknown. Suppose that inference is required for β_1, where

$$\log(\mu_i) = \beta_0 + \beta_1(x_i - \bar{x}_i).$$

Show that the quasi-likelihood estimates of β_0, β_1 are uncorrelated with asymptotic variances

$$\text{var}(\hat{\beta}_0) = \sigma^2/n, \quad \text{and} \quad \text{var}(\hat{\beta}_1) = \sigma^2 / \sum (x_i - \bar{x}_i)^2.$$

9.9 Suppose, for the problem described above, that a Normal-theory likelihood is used even though the data may not be Normal. Show that the 'maximum likelihood' estimates $\tilde{\beta}_0, \tilde{\beta}_1$ thus obtained are consistent under the assumptions stated. Show also that the true asymptotic variance of $\tilde{\beta}_1$, as opposed to the apparent value given by the Normal-theory log likelihood, is

$$\text{var}(\tilde{\beta}_1) = \frac{\sigma^2\{1 + 2\sigma\rho_3 + \sigma^2(\rho_4 + 2)\}(1 + 2\sigma^2)^{-2}}{\sum (x_i - \bar{x})^2},$$

where $\rho_3 = \kappa_3/\kappa_2^{3/2}$ and $\rho_4 = \kappa_4/\kappa_2^2$ are the standardized third and fourth cumulants, assumed constant over i. For a range of plausible values of σ^2, ρ_3, ρ_4, compare the efficiencies of the two methods of estimation.

Derive the asymptotic relative efficiency of $\tilde{\beta}_1$ to $\hat{\beta}_1$ under the assumption that the data are in fact Normal. [McCullagh, 1984b].

9.10 Suppose, conditionally on $Y_{i\cdot} = m_i$, $Y_{\cdot j} = s_j$, that $\mathbf{Y} = (Y_{11}, Y_{12}, Y_{21}, Y_{22})$ has the non-central hypergeometric distribution (7.9) with odds ratio ψ. Deduce that

$$g(\mathbf{Y}; \psi) = Y_{11}Y_{22} - \psi Y_{12}Y_{21}$$

is an unbiased estimating function for ψ. Show also that if $\psi = 1$

$$\operatorname{var}\big(g(\mathbf{Y};1)\big) = \frac{m_1 m_2 s_1 s_2}{m_\cdot - 1}.$$

Hence deduce that for n independent 2×2 tables $\mathbf{Y}^{(i)}$ with common odds-ratio ψ

$$\sum_i \frac{Y_{11}^{(i)} Y_{22}^{(i)} - \psi Y_{12}^{(i)} Y_{21}^{(i)}}{m_\cdot^{(i)}}$$

is an unbiased estimating equation for ψ, but is not optimal unless $\psi = 1$.

The estimator produced by this function

$$\hat{\psi}_{\text{MH}} = \frac{\sum Y_{11}^{(i)} Y_{22}^{(i)} / m_\cdot^{(i)}}{\sum Y_{12}^{(i)} Y_{21}^{(i)} / m_\cdot^{(i)}}$$

is known as the Mantel-Haenszel estimator. Find an expression for the asymptotic variance as $n \to \infty$ of $\hat{\psi}_{\text{MH}}$ when $\psi = 1$. [Mantel and Haenszel (1959); Mantel and Hankey (1975); Breslow and Day (1980, p.240); Breslow (1981); Breslow and Liang (1982)].

9.11 Use the above estimating equation to estimate the common odds-ratio for the data in Table 7.2. Compare your estimate and its estimated variance with the values obtained in section 7.4.3.