

NON-LINEAR REGRESSION MODELLING

J.C. WAKEFIELD

*Departments of Statistics and Biostatistics,
University of Washington,
Seattle,
WA 98195, United States
Email: jonno@u.washington.edu*

In this paper frequentist and Bayesian approaches to non-linear regression modelling are described, critiqued and contrasted. Estimating functions provide a unifying framework for frequentist inference, and sampling-based methods provide a flexible computational technique for carrying out Bayesian analyses. Special interest is focused upon the effects of model misspecification; in this regard the use of the (linear) exponential family is beneficial, and provides one advantage of using the generalized linear model class. A new application of the inverse polynomial models introduced in Nelder (1966) is presented: the analysis of data from a pharmacokinetic experiment.

1. Introduction

Over recent years, increases in computer power, accompanied by algorithmic development and the inclusion of such algorithms within statistical software, have unshackled the statistician in his/her ability to fit models of choice, as opposed to models imposed by mathematical and/or computational convenience. In this paper the analysis of data using non-linear models is considered.

In preparation for analysis the strategy that is stressed is:

- 1 To formulate an initial model class on the basis of the context.
- 2 To examine this class with respect to its statistical properties; specifically the behaviour of estimators and posterior distributions (in particular with respect to model misspecification) may be examined from, respectively, frequentist and Bayesian perspectives.
- 3 To examine computational aspects.

In either 2 or 3 the model may be altered to correct mathematical or computational shortcomings. The approach followed in this paper is rooted firmly

in the tradition of attempting to understand structure within data through parametric modelling of the mean, in contrast to the predictive view of statistical inference (see Breiman (2001) and the ensuing discussion).

There are a number of challenges associated with the ability to fit ever more complex models. First, the statistical properties of complex models are often not fully understood, in particular with respect to model misspecification. A second problem is the potential for loss of information on parameters of interest when the number of nuisance parameters is unnecessarily increased by expanding the model; further discussion of this issue is given in Section 3.2.

A third difficulty is that there now exists great potential for over-fitting in which models become too dataset-specific as they are refined on the basis of the examination of diagnostics. In practice, if refinement is carried out through the fitting of alternative models (e.g. transformation of covariates, choice of distribution for the responses), then interval estimates will often be too narrow since they are produced by conditioning on the final model, and hence do not reflect the mechanism by which the model was selected (see Chatfield (1995), and the accompanying discussion). From a frequentist standpoint estimators and test statistics should be examined via their long-run behaviour *given* the model-fitting process, including refinement. To be more explicit, let P denote the procedure by which a final model M is decided upon. Then suppose it is of interest to examine the bias of a statistic T ,

$$E[T|P] = E_{M|P}\{E[T|M]\}. \quad (1)$$

In general it will be incorrect to report $E[T | \hat{M}]$ where \hat{M} is the final model chosen, since this does not reflect the procedure by which \hat{M} was chosen, but rather acts as if the final model is the “truth”. From a Bayesian standpoint the same problem exists because the posterior distribution should reflect all sources of uncertainty and *a priori* all possible models that may be entertained should be explicitly stated, with prior distributions being placed upon different likelihoods and the parameters of these likelihoods; model averaging should then be carried out across the different possibilities, a process which is fraught with difficulties not least in placing “comparable” priors over what may be fundamentally different objects (see Section 6 for an approach to rectifying this problem).

One solution to this third difficulty is to never refine the model for a given data set. This approach is operationally pure but pragmatically dubious (unless one is in the context of a randomized experiment) since we

may obtain appropriate inference for a model that is a very poor description of the phenomenon under study. The philosophy suggested here is to think as carefully as possible about the initial model class before the analysis proceeds, but after fitting to carry out model checking and refine the model in the face of *clear* model misspecification, with refinement ideally being carried out within distinct *a priori* known classes^a. Inference then proceeds as if the final model were the one that were chosen initially. This is clearly a subjective procedure but can be informally justified via either philosophical approaches.

Under a frequentist approach inference follows from the behaviour of an estimator under repeated sampling from the true model, and if an initial model is clearly wrong on the basis of a residual plot (say), then it is very unlikely to be close to the “true” model and hence it is more appropriate to obtain properties of estimators under the assumed model. With reference to (1), if a model is chosen because it is clearly superior to the alternatives, then it may be reasonable to assume that $E[T | P] \approx E[T | \hat{M}]$, because \hat{M} would be consistently chosen in repeated sampling under these circumstances.

In a similar vein, under a Bayesian approach the above procedure is consistent with model-averaging but with the posterior model weight being concentrated upon the chosen model (since alternative models are only rejected on the basis of clear inadequacy). The aim is to provide probability statements, from either philosophical standpoints that are “honest” representations of uncertainty. The above approach is relevant to analyses that are more confirmatory in their outlook, as opposed to being used for prediction, or for more exploratory purposes (for example, to gain clues to models that may be appropriate for future data analyses).

The structure of this paper is as follows. The frequentist approach to the analysis of non-linear models is considered in Section 2, with an estimating functions approach being emphasized, and specific choices being suggested by likelihood and quasi-likelihood. The Bayesian approach is described in Section 3 with computation via direct sampling from the posterior being described. A critique and comparison of the frequentist and Bayesian approaches is carried out in Section 4; in particular, situations in which one may be preferred over the other are delineated. Specific non-linear model classes are considered in Section 5; with generalized linear models being

^aSo that, for example, examining quantile-quantile plots for different t distributions and picking the one that produces the straightest line would not be a good idea.

described in Section 5.1 and compartmental models in Section 5.2. The approach to modelling followed in the paper is illustrated with the analysis of a set of pharmacokinetic data in Section 6. The paper ends with a concluding discussion in Section 7.

2. Frequentist Inference

Under the frequentist approach to inference procedures are assessed with respect to their long-run properties under hypothetical repeated sampling. If estimation is the objective then the aim is to obtain an estimator whose distribution is “close” to the true value. A fundamental criteria is *consistency* which heuristically states that the estimator tends to the true value as the sample size increases. Another criteria by which we may compare two competing asymptotically unbiased estimators is via comparison of their asymptotic variances; an asymptotically *efficient* estimator is one that attains the lowest possible asymptotic variance.

Estimating functions have emerged as a unifying approach to much of frequentist inference and in the next section we review the basics before giving specific examples of estimating functions in the following two sections, specifically those arising from likelihood in Section 2.2 and quasi-likelihood in Section 2.3. In Section 2.4 sandwich estimation as a method of obtaining a consistent estimator of the variance of an estimator is described.

2.1. Estimating Functions

Let $\mathbf{Y} = (Y_1, \dots, Y_n)^\top$, represent n observations from a distribution indexed by a p -dimensional parameter $\boldsymbol{\theta}$, with $Y_i|\boldsymbol{\theta}$ (conditionally) independent. An *estimating function* is a function

$$\mathbf{G}(\boldsymbol{\theta}) = \sum_{i=1}^n \mathbf{G}(\boldsymbol{\theta}, Y_i) = \sum_{i=1}^n \mathbf{G}_i(\boldsymbol{\theta}) \quad (2)$$

of the same dimension as $\boldsymbol{\theta}$ for which

$$\mathbb{E}[\mathbf{G}(\boldsymbol{\theta})] = \mathbf{0}. \quad (3)$$

The estimating function $\mathbf{G}(\boldsymbol{\theta})$ is a random variable because it is a function of \mathbf{Y} . The corresponding *estimating equation* that defines the estimator $\hat{\boldsymbol{\theta}}$ has the form

$$\mathbf{G}(\hat{\boldsymbol{\theta}}) = \sum_{i=1}^n \mathbf{G}_i(\hat{\boldsymbol{\theta}}) = \mathbf{0}. \quad (4)$$

For inference, the frequency properties of the estimating function are derived and are then transferred to the resultant estimator. This is an ingenious approach because the estimating function may be constructed to be a simple function of the data, while the estimator of the parameter that solves (4) will often be unavailable in closed form. The estimating function (2) is a sum of random variables which provides the opportunity to evaluate its asymptotic properties via a central limit theorem. The *art* of constructing estimating functions is to make them dependent on distribution-free quantities, for example, the population moments of the data; in Section 5.1 we will see that estimators arising from exponential family models are particularly appealing. We now state a theorem that forms the basis for asymptotic inference.

Theorem: The estimator $\hat{\boldsymbol{\theta}}_n$ which is the solution to the estimating equation

$$\mathbf{G}(\hat{\boldsymbol{\theta}}_n) = \sum_{i=1}^n \mathbf{G}_i(\hat{\boldsymbol{\theta}}_n) = \mathbf{0},$$

has asymptotic distribution

$$\hat{\boldsymbol{\theta}}_n \rightsquigarrow N_p\left(\boldsymbol{\theta}, \mathbf{A}^{-1} \mathbf{B} \mathbf{A}^{\text{T}-1}\right),$$

where

$$\mathbf{A} = \mathbf{A}_n(\boldsymbol{\theta}) = \text{E} \left[\frac{\partial \mathbf{G}}{\partial \boldsymbol{\theta}} \right] = \sum_{i=1}^n \text{E} \left[\frac{\partial \mathbf{G}_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right],$$

and

$$\mathbf{B} = \mathbf{B}_n(\boldsymbol{\theta}) = \text{cov}(\mathbf{G}) = \sum_{i=1}^n \text{cov}\{\mathbf{G}_i(\boldsymbol{\theta})\}.$$

The form of the covariance of the estimator here, the covariance of the estimating function, flanked by the inverse of the Jacobean of the transformation from the estimating function to the parameter, is one that will appear again in Section 2.4 in the context of sandwich estimation.

In practice, $\mathbf{A} = \mathbf{A}_n(\boldsymbol{\theta})$ and $\mathbf{B} = \mathbf{B}_n(\boldsymbol{\theta})$ are replaced by $\hat{\mathbf{A}} = \mathbf{A}_n(\hat{\boldsymbol{\theta}}_n)$ and $\hat{\mathbf{B}} = \mathbf{B}_n(\hat{\boldsymbol{\theta}}_n)$, respectively. In this case, from a Slutsky Theorem (see for example Ferguson (1996), Chapter 6),

$$\hat{\boldsymbol{\theta}}_n \rightsquigarrow N_p\left(\boldsymbol{\theta}, \hat{\mathbf{A}}^{-1} \hat{\mathbf{B}} \hat{\mathbf{A}}^{\text{T}-1}\right), \quad (5)$$

since $\hat{\mathbf{A}} \rightarrow_p \mathbf{A}$ and $\hat{\mathbf{B}} \rightarrow_p \mathbf{B}$.

The accuracy of the asymptotic approximation to the sampling distribution of the estimator is dependent on the parameterization adopted. A rule of thumb is to obtain the confidence interval on a reparameterization which takes the parameter onto the real line (for example, the logistic transform for a probability, or the logarithmic transform for a dispersion parameter), and then to transform to the more interpretable scale; examples are presented in Section 6. Estimators for functions of interest, $\phi = g(\boldsymbol{\theta})$, may be obtained via $\hat{\phi} = g(\hat{\boldsymbol{\theta}})$, and the asymptotic distribution may be found using the delta method.

2.2. Likelihood

We begin by giving the definition of *likelihood* (as given by Fisher (1921), p. 24).

Definition: Viewing $p(\mathbf{y} | \boldsymbol{\theta})$ as a function of $\boldsymbol{\theta}$ gives the *likelihood function* which we denote by $L(\boldsymbol{\theta})$.

To follow a likelihood approach one must, therefore, specify the probability distribution of the observed data given the model parameters, that is $p(\mathbf{y} | \boldsymbol{\theta})$. In this paper we consider models that are appropriate when the data are independent and identically distributed conditional on $\boldsymbol{\theta}$, so that we have

$$p(\mathbf{y} | \boldsymbol{\theta}) = \prod_{i=1}^n p(y_i | \boldsymbol{\theta}).$$

The probability model for the full data, which determines the likelihood (and includes the independence assumption), is based upon the context and all relevant accumulated knowledge. The level of belief in this model will clearly be context-specific and in many situations there will be insufficient information available to confidently specify all components of the model. Depending on the confidence in the likelihood, which in turn depends on the sample size (since large n allows examination of the assumptions of the model), the likelihood may either be effectively viewed as “correct” in that inference proceeds as if the true model were known, or may instead be seen as an initial *working* model from which an estimating function is derived, the properties of the subsequent estimator then being determined in a more general setting. For example, in Section 2.4 we describe a method

for producing an estimator of the variance of the estimator that does not depend on the full probability model.

The value of $\boldsymbol{\theta}$ that maximizes $L(\boldsymbol{\theta})$, denoted $\hat{\boldsymbol{\theta}}$, is known as the *Maximum Likelihood Estimator* (MLE); the MLE is therefore that value of $\boldsymbol{\theta}$ that gives the highest probability to the observed data. We now define some functions of the likelihood which will aid in the development of the asymptotic distribution of the MLE.

For both computation and the evaluation of analytical properties, it is convenient to consider the *log likelihood* function which is given by

$$l(\boldsymbol{\theta}) = \log L(\boldsymbol{\theta}) = \sum_{i=1}^n \log p(Y_i | \boldsymbol{\theta}) = \sum_{i=1}^n l_i(\boldsymbol{\theta}),$$

and the *score* function

$$\begin{aligned} \mathbf{G}(\boldsymbol{\theta}) &= \frac{\partial l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \left[\frac{\partial l(\boldsymbol{\theta})}{\partial \theta_1}, \dots, \frac{\partial l(\boldsymbol{\theta})}{\partial \theta_p} \right]^T \\ &= [\mathbf{G}_1(\boldsymbol{\theta}), \dots, \mathbf{G}_p(\boldsymbol{\theta})]^T, \end{aligned}$$

which, as we show below, satisfies the requirements of an estimating function upon which inference may be based.

Definition: Fisher's *expected* information is given by

$$\mathbf{I}(\boldsymbol{\theta}) = E\{\mathbf{G}(\boldsymbol{\theta})\mathbf{G}(\boldsymbol{\theta})^T\},$$

a $p \times p$ matrix.

Result: Under regularity conditions:

$$E[\mathbf{G}(\boldsymbol{\theta})] = E\left[\frac{\partial l}{\partial \boldsymbol{\theta}}\right] = \mathbf{0}, \quad (6)$$

and

$$\mathbf{I}(\boldsymbol{\theta}) = E\{\mathbf{G}(\boldsymbol{\theta})\mathbf{G}(\boldsymbol{\theta})^T\} = -E\left[\frac{\partial \mathbf{G}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T}\right] = -E\left[\frac{\partial^2}{\partial \boldsymbol{\theta}^T \partial \boldsymbol{\theta}} l(\boldsymbol{\theta})\right]. \quad (7)$$

Since $E[\mathbf{G}(\boldsymbol{\theta})] = \mathbf{0}$ we have $\mathbf{I}(\boldsymbol{\theta}) = \text{cov}\{\mathbf{G}(\boldsymbol{\theta})\}$. From this result we have

$$-\mathbf{A}(\boldsymbol{\theta}) = \mathbf{B}(\boldsymbol{\theta}) = \mathbf{I}(\boldsymbol{\theta})$$

and so, from the theorem of Section 2.1:

$$\hat{\boldsymbol{\theta}}_n \sim N_p\{\boldsymbol{\theta}, \mathbf{I}(\boldsymbol{\theta})^{-1}\}.$$

It can be shown that MLEs are asymptotically efficient, *if* the model from which the score was derived is correct, see for example van der Vaart (1998), Chapter 8.

As an alternative to using the asymptotic distribution, resampling methods such as the bootstrap may be used to examine the sampling distribution, see Efron and Tibshirani (1993) and Davison and Hinkley (1997). We do not discuss the bootstrap further, but acknowledge that a large literature now exists on both its theoretical properties and its use in practice (though its use in small sample is not recommended). Likelihood ratio tests may be used to obtain confidence intervals (and are invariant to the parameterization adopted), and profile likelihood provides a method of examining the likelihood function for a parameter of interest alone.

In multiparameter situations *adjusted profile likelihood* may be used to create confidence intervals for a parameter of interest while making an attempt to account for the estimation of nuisance parameters. This approach can be computationally intensive and is not always reliable, and a number of modifications have been suggested, see for example Reid (1995).

If the model is misspecified then the MLE is that value of the parameter that brings the assumed model closest, in a Kullback-Leibler sense, to the true model (Huber (1967), White (1982)).

2.3. Quasi-Likelihood

In this section we describe an estimating function that is, at least on the surface, based upon the mean and variance of the data only. Specifically we assume

$$E[\mathbf{Y}|\boldsymbol{\beta}] = \boldsymbol{\mu}(\boldsymbol{\beta}),$$

$$\text{cov}(\mathbf{Y}|\boldsymbol{\beta}) = \phi \mathbf{V}\{\boldsymbol{\mu}(\boldsymbol{\beta})\},$$

where $\boldsymbol{\mu}(\boldsymbol{\beta}) = [\mu_1(\boldsymbol{\beta}), \dots, \mu_n(\boldsymbol{\beta})]^\top$ represents a regression function and \mathbf{V} is a diagonal matrix (so the observations are uncorrelated), with

$$\text{var}(Y_i|\boldsymbol{\beta}) = \phi V\{\mu_i(\boldsymbol{\beta})\},$$

and $\phi > 0$ is a scalar which is independent of $\boldsymbol{\beta}$. The aim is to obtain the asymptotic properties of an estimator of $\boldsymbol{\beta}$. The specification of the mean function in a parametric regression setting is unavoidable, and least squares would indicate that properties for an estimator may be obtained from the additional specification of the variance.

To motivate an estimating function we follow McCullagh (1991) (see also Firth (1994) for an exceptionally clear description of quasi-likelihood) and consider the sum of squares

$$(\mathbf{Y} - \boldsymbol{\mu})^\top \mathbf{V}^{-1}(\mathbf{Y} - \boldsymbol{\mu}), \quad (8)$$

where $\boldsymbol{\mu} = \boldsymbol{\mu}(\boldsymbol{\beta})$ and $\mathbf{V} = \mathbf{V}(\boldsymbol{\beta})$. To minimize this sum of squares there are two ways to proceed. Perhaps the more obvious route is to acknowledge that both $\boldsymbol{\mu}$ and \mathbf{V} are functions of $\boldsymbol{\beta}$ and differentiate with respect to $\boldsymbol{\beta}$ to give

$$-2\mathbf{D}^T\mathbf{V}^{-1}(\mathbf{Y} - \boldsymbol{\mu})/\phi + (\mathbf{Y} - \boldsymbol{\mu})^T \frac{\partial \mathbf{V}^{-1}}{\partial \boldsymbol{\beta}} (\mathbf{Y} - \boldsymbol{\mu})/\phi, \quad (9)$$

where \mathbf{D} is the $n \times p$ matrix of derivatives with elements $\partial\mu_i/\partial\beta_j$, $i = 1, \dots, n; j = 1, \dots, p$. Unfortunately, if we only assume that $E[Y_i] = \mu_i(\boldsymbol{\beta})$, the expectation of (9) is not necessarily zero, and so a consistent estimator of $\boldsymbol{\beta}$ will not generally result in this situation if it based on (9). However, if the true variance is equal to $\phi\mathbf{V}$, then there is an efficiency loss in ignoring the second term, since it contains information on $\boldsymbol{\beta}$. This illustrates the classic efficiency-robustness trade-off that must be addressed whenever a model (procedure) is chosen for inference.

Alternatively we may suppose that \mathbf{V} is not a function of $\boldsymbol{\beta}$ when we differentiate (8), and then solve

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

As shorthand we write this function as

$$\mathbf{U} = \mathbf{D}^T \mathbf{V}^{-1} \{\mathbf{Y} - \boldsymbol{\mu}\} / \phi. \quad (10)$$

This estimating function is linear in the data and so its properties are straightforward to evaluate. In particular:

- (1) $E[\mathbf{U}(\boldsymbol{\beta})] = \mathbf{0}$.
- (2) $\text{cov}\{\mathbf{U}(\boldsymbol{\beta})\} = \mathbf{D}^T \mathbf{V}^{-1} \mathbf{D} / \phi$.
- (3) $-E\left[\frac{\partial \mathbf{U}}{\partial \boldsymbol{\beta}}\right] = \text{cov}\{\mathbf{U}(\boldsymbol{\beta})\} = \mathbf{D}^T \mathbf{V}^{-1} \mathbf{D} / \phi$.

The similarity of these properties with those of the score function (equations (6) and (7)) is apparent and has lead to (10) being referred to as a *quasi-score* function. We can apply the theorem of Section 2.1 directly to obtain the asymptotic distribution of the maximum quasi-likelihood estimator (MQLE) as

$$\hat{\boldsymbol{\beta}} \sim N_p\{\boldsymbol{\beta}, (\mathbf{D}\mathbf{V}^{-1}\mathbf{D})^{-1}\phi\},$$

where we have so far assumed that ϕ is known. Note that $\hat{\boldsymbol{\beta}}$ does not depend on ϕ , a consequence of assuming that ϕ is a multiplier in the variance function. Since

$$E[(\mathbf{Y} - \boldsymbol{\mu})^T \mathbf{V}^{-1}(\boldsymbol{\mu})(\mathbf{Y} - \boldsymbol{\mu})] = n\phi,$$

an unbiased estimator of ϕ would be

$$(\mathbf{Y} - \boldsymbol{\mu})^T \mathbf{V}^{-1}(\boldsymbol{\mu})(\mathbf{Y} - \boldsymbol{\mu})/n.$$

A “degrees of freedom corrected” (but not in general, unbiased) estimate is therefore given by the Pearson statistic divided by its degrees of freedom

$$\hat{\phi} = \frac{1}{n-p} \sum_{i=1}^n \frac{\{Y_i - \hat{\mu}_i\}^2}{V(\hat{\mu}_i)}, \quad (11)$$

where $\hat{\mu}_i = \hat{\mu}_i(\hat{\boldsymbol{\beta}})$. The benefit of this approach, as opposed (say) to constructing an estimating function for ϕ from a likelihood, is again one of robustness, since (11) will in general be more appropriate under a broader range of circumstances (those in which the mean and variance-covariance models are correct) than a model-based alternative. The asymptotic distribution that is used in practice is given by

$$\hat{\boldsymbol{\beta}} \sim N_p\{\boldsymbol{\beta}, (\mathbf{D}\mathbf{V}^{-1}\mathbf{D})^{-1}\hat{\phi}\},$$

so that the uncertainty in $\hat{\phi}$ is not accommodated in the uncertainty for $\hat{\boldsymbol{\beta}}$ (see Section 3.2 for related discussion). McCullagh (1983) and Crowder (1986) give conditions under which this asymptotic result may be appealed to. Crowder (1987) gives counter-examples in which a linear estimating function such as (10) does not perform well; these examples are mostly of theoretical interest but do indicate that one should not assume that linear estimating functions always perform well.

Integration of the quasi-score (10) gives

$$l(\mu; y) = \int_y^\mu \frac{y-t}{\phi V(t)} dt,$$

which, if it exists, behaves like a log-likelihood, explaining the genesis of the label “quasi-likelihood”; Wedderburn (1974) was the first to consider this class. As an example, for the model $E[Y] = \mu$ and $\text{var}(Y) = \phi\mu$ we have

$$l(\mu; y) = \int_y^\mu \frac{y-t}{\phi t} dt = \frac{1}{\phi} [y \log \mu - \mu + c],$$

where $c = -y \log y - y$ and $y \log \mu - \mu$ is the log likelihood of a Poisson random variable. The word “quasi” refers to the fact that the score may or may not equate to a probability function. For example, the variance function $\mu^2(1-\mu)^2$ does not correspond to a probability distribution (but was shown by McCullagh and Nelder (1989), Example 9.2.4, to be useful in a particular application). If the estimating function (10) corresponds to

a score function then the subsequent estimator corresponds to the MLE. Hence, although the mean and variance only are specified in the estimating function, there may be an implicit model in the sense that the estimating function corresponds to a particular likelihood function. As a trivial example, the estimating function based on $E[Y] = \mu$, $\text{var}(Y) = \phi$ corresponds to the model $Y \sim_{ind} N(\mu, \phi)$.

The prediction of *observable* data Y is not possible with quasi-likelihood, since there is no probabilistic mechanism to appeal to.

2.4. Sandwich Estimation

A general method of avoiding stringent modelling conditions when the variance of an estimator is calculated is provided by *sandwich estimation*. The basic idea is to empirically estimate the variance of the data with minimum modelling assumptions, and to incorporate this in the estimation of the variance of an estimator. While the idea may be traced at least as far as Huber (1967), the paper of White (1980) implemented the technique for the linear model, and Royall (1986) provided a clear and simple account with many examples; Liang and Zeger (1986) and Zeger and Liang (1986) described the technique in the context of longitudinal data by using the replication across individuals to empirically estimate within-person correlations. Carroll et al. (1995), Appendix A.3 provide a good review.

We have seen that when the estimating function corresponds to a score equation, then *under the model*

$$\mathbf{I} = \mathbf{A} = -\mathbf{B}$$

so that

$$\text{var}(\hat{\boldsymbol{\theta}}) = \mathbf{A}(\boldsymbol{\theta})^{-1} \mathbf{B}(\boldsymbol{\theta}) \mathbf{A}(\boldsymbol{\theta})^{\text{T}-1} = \mathbf{I}(\boldsymbol{\theta})^{-1}.$$

If the model is not correct then this equality does not hold, and the variance estimator will be incorrect. An alternative is to *empirically* evaluate the variance via

$$\hat{\mathbf{A}} = \sum_{i=1}^n \frac{\partial}{\partial \boldsymbol{\theta}} \mathbf{G}(\hat{\boldsymbol{\theta}}, Y_i),$$

and

$$\hat{\mathbf{B}} = \sum_{i=1}^n \mathbf{G}(\hat{\boldsymbol{\theta}}, Y_i) \mathbf{G}(\hat{\boldsymbol{\theta}}, Y_i)^{\text{T}}.$$

This method is general and can be applied to any estimating function, not just those arising from a score equation.

Suppose we assume $E[\mathbf{Y}] = \boldsymbol{\mu}$ and $\text{var}(\mathbf{Y}) = \phi\mathbf{V}$ with $\text{var}(Y_i) = \phi V(\mu_i)$, and $\text{cov}(Y_i, Y_j) = 0$, $i, j = 1, \dots, n$, $i \neq j$, as a *working* covariance model. Under this specification it is natural to take (10) as an estimating function and

$$\text{var}_s(\hat{\boldsymbol{\beta}}) = (\mathbf{D}^T \mathbf{V}^{-1} \mathbf{D})^{-1} \mathbf{D}^T \mathbf{V}^{-1} \text{cov}(\mathbf{Y}) \mathbf{V}^{-1} \mathbf{D} (\mathbf{D}^T \mathbf{V}^{-1} \mathbf{D})^{-1},$$

and so the appropriate variance is obtained by substituting in the correct form for $\text{cov}(\mathbf{Y})$ which is, of course, unknown. However, a simple “sandwich” estimator of the variance is given by

$$\text{var}_s(\hat{\boldsymbol{\beta}}) = (\mathbf{D}^T \mathbf{V}^{-1} \mathbf{D})^{-1} \mathbf{D}^T \mathbf{V}^{-1} \mathbf{R} \mathbf{R}^T \mathbf{V}^{-1} \mathbf{D} (\mathbf{D}^T \mathbf{V}^{-1} \mathbf{D})^{-1},$$

where $\mathbf{R} = (R_1, \dots, R_n)^T$ is the $n \times 1$ vector with $R_i = Y_i - \mu_i(\hat{\boldsymbol{\beta}})$. This estimator is consistent for the variance of $\hat{\boldsymbol{\beta}}$, under correct specification of the mean, and with uncorrelated data. There is finite sample bias in R_i as an estimate of $Y_i - \mu_i(\boldsymbol{\beta})$ and versions that adjust for the estimation of the parameters $\boldsymbol{\beta}$ are also available, see for example Kauermann and Carroll (2001).

The great advantage of sandwich estimation is that it provides a consistent estimator of the variance in very broad situations. There are two things to bear in mind when one considers the use of this technique, however. The first is that for small sample sizes, the sandwich estimator may be highly unstable, and in terms of mean squared error model-based estimators may be preferable for small to medium sized n (for small samples one would anyway want to avoid the reliance on the asymptotic distribution). Hence *empirical* is a better description of the estimator than *robust*. The second consideration is that if the model is correct, then the model-based estimators are more efficient.

3. Bayesian Inference

3.1. Summarising the Posterior Distribution

In the Bayesian approach all unknown quantities which are contained in a probability model for the observed data (including, for example, missing or censored data) are considered to be random variables. This is in contrast to the frequentist view in which parameters are treated as *constants*.^b Let $\boldsymbol{\theta} =$

^bHere, strictly, *fixed* effects parameters are being referred to. So-called *random* effects are assumed to arise from a population distribution and are viewed as random.

$(\theta_1, \dots, \theta_p)^T$ denote all of the unknowns of the model, and $\mathbf{y} = (y_1, \dots, y_n)^T$ the vector of observed data. Also let \mathcal{I} represent all information relevant to the analysis that is currently available to the individual who is carrying out the analysis, in addition to \mathbf{y} .

Inference is made through the posterior probability distribution of $\boldsymbol{\theta}$, after observing \mathbf{y} :

$$p(\boldsymbol{\theta} | \mathbf{y}, \mathcal{I}) = \frac{p(\mathbf{y} | \boldsymbol{\theta}, \mathcal{I}) \times \pi(\boldsymbol{\theta} | \mathcal{I})}{p(\mathbf{y} | \mathcal{I})}, \quad (12)$$

where $p(\mathbf{y} | \boldsymbol{\theta}, \mathcal{I})$ is the likelihood, and $\pi(\boldsymbol{\theta} | \mathcal{I})$ the *prior* distribution representing the probability beliefs for $\boldsymbol{\theta}$ *before* observing the data \mathbf{y} , based on the current information \mathcal{I} . Different individuals will have different information \mathcal{I} and so in general priors, and for that matter likelihoods, may differ. The normalizing constant is given, for continuous $\boldsymbol{\theta}$, by

$$p(\mathbf{y} | \mathcal{I}) = \int_{\boldsymbol{\theta}} p(\mathbf{y} | \boldsymbol{\theta}, \mathcal{I}) \pi(\boldsymbol{\theta} | \mathcal{I}) d\boldsymbol{\theta} \quad (13)$$

and is the marginal distribution of the data, given the likelihood and prior. From this point onwards we suppress the dependence on \mathcal{I} , for notational convenience.

To summarise the posterior distribution marginal distributions for parameters of interest may be considered. For example the univariate marginal distribution for a component θ_i is given by

$$p(\theta_i | \mathbf{y}) = \int_{\boldsymbol{\theta}_{-i}} p(\boldsymbol{\theta} | \mathbf{y}) d\boldsymbol{\theta}_{-i}, \quad (14)$$

where $\boldsymbol{\theta}_{-i}$ is the vector of all parameters, $\boldsymbol{\theta}$, excluding θ_i . Posterior moments may be evaluated from the marginal distributions; for example the posterior mean is given by

$$E[\theta_i | \mathbf{y}] = \int_{\theta_i} \theta_i p(\theta_i | \mathbf{y}) d\theta_i. \quad (15)$$

Posterior means, in contrast to MLEs, are not invariant in the sense that $E[g(\boldsymbol{\theta}) | \mathbf{y}] \neq g(E[\boldsymbol{\theta} | \mathbf{y}])$, unless g is a linear function). Further summarisation may be carried out to yield the $100 \times q$ % quantile, $\theta_i(q)$ ($0 < q < 1$) by solving

$$\int_{-\infty}^{\theta_i(q)} p(\theta_i | \mathbf{y}) d\theta_i. \quad (16)$$

In particular, the posterior median, $\theta_i(0.5)$, will often provide an adequate summary of the location of the posterior margin, and a $100 \times p$ % equi-tailed *credible interval* ($0 < p < 1$) is provided by $[\theta_i\{(1-p)/2\}, \theta_i\{(1+p)/2\}]$

Another useful inferential quantity is the *predictive* distribution for future observations \mathbf{z} which is given, under conditional independence, by

$$p(\mathbf{z} | \mathbf{y}) = \int_{\boldsymbol{\theta}} p(\mathbf{z} | \boldsymbol{\theta}) p(\boldsymbol{\theta} | \mathbf{y}) d\boldsymbol{\theta}. \quad (17)$$

If we wish to compare models M_0 and M_1 then a natural measure is given by the *posterior odds*

$$\frac{\Pr(M_0 | \mathbf{y})}{\Pr(M_1 | \mathbf{y})} = \frac{p(\mathbf{y} | M_0)}{p(\mathbf{y} | M_1)} \times \frac{\Pr(M_0)}{\Pr(M_1)}, \quad (18)$$

where the *Bayes factor* $p(\mathbf{y} | M_0)/p(\mathbf{y} | M_1)$ is the ratio of the marginal distributions of the data under the two models, and $\Pr(M_0)/\Pr(M_1)$ is the *prior odds*. To calculate the former, integrals of the form (13) are required.

Bayesian inference is deceptively simple to describe probabilistically, but there have been two major obstacles to its routine use. The first is how to specify prior distributions and will be considered in Section 3.3. The second is how to evaluate the integrals required for inference, for example (13)–(17), given that for most models (and for all but the most trivial non-linear models always), these are analytically intractable. A general method for non-hierarchical non-linear models is described in Section 3.4.

If the likelihood is correctly specified then, under certain conditions, the posterior distribution is asymptotically normal with mean the true value, and variance-covariance matrix given by the inverse of the expected information, for non-technical derivations see for example O'Hagan (1994), p. 75 and Gelman et al. (1995), Appendix B. An important practical condition is that the prior does not exclude any part of the support of the parameter. More rigorous treatments can be found in, for example, Walker (1970) and Johnson (1970), where it is shown that under suitable regularity conditions the posterior distribution tends to a normal distribution with mean the MLE, and variance-covariance matrix given by the inverse of the observed information, evaluated at the MLE.

While the posterior distribution is asymptotically independent of the prior distribution, so that point and interval estimates are often robust to the prior choice with increasing n , Bayes factors are asymptotically sensitive to the prior, for further discussion see for example O'Hagan (1994), p. 195. Kass and Raftery (1995) give a review of Bayes factors, including discussion of computation and prior choice.

3.2. Model Misspecification

The behaviour of Bayesian estimators under misspecification of the likelihood has received less attention than frequentist estimators. As discussed above, under sensible prior distributions the posterior distribution mimics the sampling distribution of the MLE, and so properties of the latter, such as consistency, can be transferred to, for example, the posterior mean or median.

Rather than the effects of misspecification, the emphasis in the Bayesian literature has been on sensitivity analyses (O'Hagan (1994), Chapter 7, gives a review of approaches to address sensitivity to prior and likelihood choices), or on embedding a particular likelihood or prior choice within a larger class. If a discrete number of choices is considered then model averaging has been used (for a review see Draper (1995)), while others (e.g. Gelman and Meng (1995)) prefer to embed the model within a continuous class and then integrate over this class.

Box and Tiao (1964) argue that examining the behaviour of an estimator under model misspecification (which they term *criterion robustness*) is inadequate since as the model varies the criterion should change also. While this is certainly true in some situations, it is not true in general and so should not be used as a reason to reject the approach out of hand. Perhaps the reason that such an approach has not been followed is because it is more difficult to apply when no closed form estimator is available. The philosophy behind consideration of misspecification is therefore very different under frequentist and Bayesian approaches.

A major problem with considering model classes with large numbers of unknown parameters is that uncertainty on parameters of interest will be increased if a simple model is closer to the truth, so their will be an efficiency loss associated with considering models that are *too* large. In particular, model expansion may lead to a decrease in precision. The following discussion relates to likelihood inference as well as to Bayesian inference, but we include it here because the emergence of MCMC has encouraged the use of larger and larger models within a Bayesian approach.

We examine the form of the posterior variance. As n increases the prior effect is negligible and the posterior variance is given by the inverse of the observed information; for convenience we consider the expected information, which is asymptotically equivalent. Suppose that we have a $k \times 1$ vector of parameters, β , in an original model (and these include the parameters of interest), and $p - k$ additional parameters, γ , in an expanded

model. Then consider

$$\mathbf{I}(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \begin{bmatrix} \mathbf{I}_{11} & \mathbf{I}_{12} \\ \mathbf{I}_{21} & \mathbf{I}_{22} \end{bmatrix}, \quad (19)$$

where \mathbf{I}_{11} is a $k \times k$ matrix corresponding to the information for $\boldsymbol{\beta}$, and \mathbf{I}_{22} is the $(p - k) \times (p - k)$ information for $\boldsymbol{\gamma}$. In the simpler model the information on the parameters of interest is \mathbf{I}_{11} , while for the enlarged model it is

$$\mathbf{I}_{11} - \mathbf{I}_{12}\mathbf{I}_{22}^{-1}\mathbf{I}_{21},$$

which is never greater than \mathbf{I}_{11} .

To illustrate in a simple regression setting, consider an observational study in which the covariate of interest is not orthogonal to all other potential confounding variables. As a specific example, the model

$$Y_i = \beta_0^e + \beta_1^e(x_i - \bar{x}) + \gamma(z_i - \bar{z}) + \epsilon_i,$$

is an expansion of the model

$$Y_i = \beta_0^r + \beta_1^r(x_i - \bar{x}) + \epsilon_i,$$

$i = 1, \dots, n$, where $\epsilon_i \sim_{iid} N(0, \sigma^2)$, with σ^2 known. Here we have distinguished between β_0^e and β_1^e in the expanded model, and β_0^r and β_1^r in the reduced model, because the parameters have different interpretations, and we need to distinguish between them when the posterior variance of each is considered below. Letting \mathbf{x} denote the $n \times 3$ matrix with i -th row $[1 \ x_i \ z_i]$, and $\boldsymbol{\beta}^e = (\beta_0^e \ \beta_1^e)^\top$, we have

$$\mathbf{I}(\boldsymbol{\beta}^e, \boldsymbol{\gamma}) = \sigma^{-2}(\mathbf{x}^\top \mathbf{x}) = \sigma^{-2} \begin{bmatrix} n & 0 & 0 \\ 0 & S_{xx} & S_{xz} \\ 0 & S_{xz} & S_{zz} \end{bmatrix},$$

where $S_{xx} = \sum_i (x_i - \bar{x})^2$, $S_{xz} = \sum_i (x_i - \bar{x})(z_i - \bar{z})$, $S_{zz} = \sum_i (z_i - \bar{z})^2$. Hence

$$\mathbf{I}(\boldsymbol{\beta}, \boldsymbol{\gamma})^{-1} = \sigma^2 \begin{bmatrix} 1/n & 0 & 0 \\ 0 & S_{zz}/D & -S_{xz}/D \\ 0 & -S_{xz}/D & S_{xx}/D \end{bmatrix},$$

where $D = S_{xx}S_{zz} - S_{xz}^2$, giving

$$\text{var}(\beta_1^e | \mathbf{y}) = \frac{\sigma^2}{S_{xx} - S_{xz}^2/S_{zz}} \geq \frac{\sigma^2}{S_{xx}} = \text{var}(\beta_1^r | \mathbf{y}),$$

with equality if $S_{xz} = 0$, i.e. if x and z are orthogonal. Intuitively, the posterior variance is increased because when z is present in the model there

are competing explanations for the observed association between y and x . Of course, one of the reasons for including additional variables is to reduce bias; however, it is straightforward to phrase the above argument in terms of mean squared error and reach the same conclusion when the bias reduction due to the inclusion of z is not great.

With unknown σ^2 the situation is more complex since important covariates will reduce the size of the estimate of the variance, $s^2 = \text{RSS}/\text{DF}$ (where RSS is the residual sum of squares and DF the degrees of freedom), and (asymptotically, or with flat priors) $\text{var}(\beta_1^e | \mathbf{y}) = (\mathbf{x}^T \mathbf{x})^{-1} s^2$ and so the posterior variance will be reduced also. However, at some point this variance will also increase since s^2 increases as unimportant covariates are added to the model. So again the overall conclusion is the same: models should not be chosen to be as large as possible, because the variance of quantities of interest will be unnecessarily increased.

We now briefly discuss the general situation in which *estimated* parameters are used in the information matrix. In this situation the variance of quantities of interest is again increased (as just discussed in the normal case). The simplest example is in a generalized linear model with scale parameter ϕ . Assuming ϕ is known corresponds to one family, while ϕ unknown corresponds to another family, as examples the Poisson becomes the negative binomial, and the exponential becomes the gamma. If the data are truly from the simpler model then interval estimates will be unnecessarily widened if the larger model is assumed. This occurs even though the posterior distributions of β and ϕ are asymptotically independent (so that in (19) \mathbf{I}_{12} and \mathbf{I}_{21} are zero); the extra uncertainty is introduced when estimates are substituted into the information. As an aside, the Poisson and exponential scenarios are perhaps not the best illustrations since not allowing excess variation in these two models would be a very dangerous modelling strategy.

The above is a very informal discussion, for a far deeper discussion of the choice between Student's t and normal errors see Hjort (1994). An interesting theoretical finding is that even if the errors are truly t , if the degrees of freedom are estimated, for small values of n , it will be more efficient to assume normal errors, because of the extra uncertainty involved in the estimation of the degrees of freedom.

3.3. *The Prior Distribution*

The specification of the prior distribution is clearly a crucial aspect of the Bayesian approach. We distinguish between two situations. In the first an analysis is required in which the prior distribution has minimal impact, so that the likelihood is concentrated upon. Such an analysis may be used as a comparison with other analyses in which more informative priors are specified, in order to determine the information being provided by the prior. Alternatively, the Bayesian formulation may be seen as a convenient way of carrying out computation for those with a likelihood bent. The second situation is one in which it is desired to incorporate more substantial prior information in the analysis.

For nonlinear models care must be taken to ensure that the posterior corresponding to a particular prior choice is proper. In particular the use of an improper uniform prior is not to be universally recommended. Such forms for fixed effects in a generalized linear model will usually lead to a proper posterior (Dellaportas and Smith (1993)) although not for some pathological cases; for example if a uniform prior is used on $\log\{p/(1-p)\}$ and $y = 0$ or $y = n$.

To illustrate the non propriety in more general non-linear models consider the model

$$Y_i \sim_{ind} N\{\exp(-\theta x_i), \sigma^2\},$$

$i = 1, \dots, n$, with $\theta > 0$ and σ^2 assumed known. With an improper uniform prior on θ we have the posterior

$$p(\theta | \mathbf{y}) \propto \exp \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - e^{-\theta x_i})^2 \right\}.$$

As $\theta \rightarrow \infty$, $p(\theta | \mathbf{y}) \rightarrow \exp \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^n y_i^2 \right\}$, a constant, so that the posterior is improper. Intuitively, the problem here is that as $\theta \rightarrow \infty$ the fitted values do not move increasingly away from the data but to the asymptote $y = 0$. There are no asymptotes in the linear model and so as the parameters increase/decrease to $\infty/-\infty$, the fitted line moves increasingly far from the data, and the likelihood tends to zero.

3.4. *Sampling-Based Inference*

Sampling-based methods have revolutionised the practical applicability of Bayesian methods. Such methods build on the duality between samples and densities (Smith and Gelfand (1992)); given a sample we can reconstruct

the density, and given an arbitrary density we can generate a sample, given the range of generic random variate generators available (see Devroye (1986)). With respect to the latter, the ability to obtain *direct* samples from a distribution decreases as the dimensionality of the parameter space increases and in this case Markov chain Monte Carlo (MCMC) methods may be used as an alternative, the disadvantage being that iteration is needed to produce samples that can be viewed as from the density of interest, and these samples are dependent. It is also not straightforward to calculate marginal densities such as (13) with MCMC, see DiCiccio et al. (1997) for a review.

For hierarchical models direct sampling is rarely possible (though feasible if the random effects may be integrated out, as in a linear hierarchical model), and MCMC needs to be considered. This paper concentrates on non-hierarchical models and in this case direct sampling is often feasible. We now describe a rejection algorithm that we will use to carry out Bayesian inference in Section 6.

Let $\boldsymbol{\theta}$ denote the unknown parameters and assume that we can evaluate the maximized likelihood $M = \sup_{\boldsymbol{\theta}} p(\mathbf{y} | \boldsymbol{\theta})$. The algorithm then proceeds as follows:

- (1) Generate $U \sim U(0, 1)$ and, independently, $\boldsymbol{\theta} \sim \pi(\boldsymbol{\theta})$.
- (2) Accept $\boldsymbol{\theta}$ if

$$U < \frac{p(\mathbf{y} | \boldsymbol{\theta})}{M},$$

otherwise return to 1.

The probability that a point is accepted is given by

$$p_a = \frac{\int p(\mathbf{y} | \boldsymbol{\theta}) \pi(\boldsymbol{\theta}) d\boldsymbol{\theta}}{M} = \frac{p(\mathbf{y})}{M},$$

(Wakefield (1996)). Hence the empirical rejection rate, \hat{p}_a , can be used to derive the prior predictive as

$$\tilde{p}(\mathbf{y}) = M \times \hat{p}_a. \quad (20)$$

An alternative importance sampling estimator that is more efficient (Evans and Swartz (1995), Pauler et al. (1999)) is given by

$$\hat{p}(\mathbf{y}) = \frac{1}{S} \sum_{s=1}^S p(\mathbf{y} | \boldsymbol{\theta}^{(s)}),$$

where $\boldsymbol{\theta}^{(s)} \sim_{iid} \pi(\boldsymbol{\theta})$. Hence it is straightforward to calculate Bayes factors using the rejection algorithm.

Clearly we need a proper prior distribution to implement the above algorithm, and the efficiency of the algorithm will depend on the correspondence between the likelihood and the prior, as measured through $p(\mathbf{y})$. For large n the algorithm will become less efficient (since M increases as n increases). As we will demonstrate in Section 6, it is straightforward to specify the prior distribution in one parameterization, and specify the likelihood in another. The latter is useful since we may be able to specify the prior in terms of a set of model-free parameters, and then compare different likelihoods with an “egalitarian” prior. Another potential advantage is that the above algorithm does not require the functional form of the prior. Wakefield (1996) used a predictive distribution from a Bayesian analysis of a set of data as the prior for the analysis of a separate data set; samples from the predictive distribution could be simply generated, even though no closed form was available for this distribution.

For a generalized linear model let $\boldsymbol{\theta} = (\boldsymbol{\beta}, \phi)$ where $\boldsymbol{\beta}$ is the vector of regression parameters and ϕ is a dispersion parameter. Standard software may be used to find the MLE $\hat{\boldsymbol{\theta}}$; however, some software (**R** for example), by default supplies a method of moments estimator, rather than the MLE, of ϕ . Since the MLE of $\boldsymbol{\theta}$ does not depend on ϕ in a GLM, one simply needs to find the MLE $\hat{\phi}$ (which is available in **R** for some families, and in particular for the normal and gamma likelihoods that are used in Section 6).

4. Comparison of Frequentist and Bayesian Methods

In this section we will describe situations in which frequentist and Bayesian methods are likely to agree, and when one is preferable over the other. We concentrate on estimation since point and interval estimation are directly comparable under the two paradigms. For model comparison the objectives of Bayes factors and hypothesis tests are fundamentally different, see for example Berger (2003), and so comparison is more difficult.

In terms of interpretation, the Bayesian approach is more straightforward since one can make probability statements, for example, credible intervals are probabilistic. In contrast, frequentist confidence intervals are not so simple to interpret.

Another appealing characteristic is that the Bayesian approach to inference may be derived via decision theory, see for example Bernardo and

Smith (1994). The likelihood principal, Berger and Wolpert (1988), also leads one towards a Bayesian approach since all frequentist criteria invalidate this principle, and a true likelihood approach, as followed by for example, Royall (1997), is difficult to calibrate. One may of course question the whole endeavor of establishing optimality, given that the subsequent use depends on the specification of likelihoods and priors, both of which are fraught with difficulties.

In contrast the frequentist approach has been justified within a frequentist set of guidelines. For example, there is a Gauss-Markov theorem for linear estimating functions (e.g. Godambe and Heyde (1987), McCullagh (1983)), while Crowder (1987) considers the optimality of quadratic estimating functions (which for implementation unfortunately require assumptions about the third and fourth moments). If one accepts that frequentist criteria are natural, then it would be desirable to find an estimator which minimises the mean squared error with respect to the sampling distribution. Unfortunately, this is not in general possible for finite n , and instead adjusted criteria such as minimum variance unbiased estimators becomes desirable.

A major problem with the frequentist approach is that, in contrast to the Bayesian approach, there is no rigid prescription for carrying out inference. Hence, for example, different types of likelihood (e.g. conditional, marginal, partial, profile, adjusted profile) exist as alternatives when conventional likelihood methods are inadequate (though in such cases the use of Bayesian methods usually requires careful prior specification). Some of these procedures are to deal with nuisance parameters, again the Bayesian approach is theoretically straightforward since posterior distributions for parameters of interest are obtained through marginalisation.

The greatest drawback of the Bayesian approach is the need to specify both a likelihood and a prior distribution. Sensitivity to each of these components can be carried out but the extent of such an approach is never clear, and one then is faced with the difficulty of how the results are reported. As we have discussed, assessing the behaviour of procedures under model misspecification is far more developed for frequentist methods than for Bayesian methods. For example, although a specific likelihood may be used to define the estimator, the properties of this estimator can be evaluated under more general models.

Bayesian methods are far more amenable to situations in which n is small. In this situation there is no way that the likelihood can be checked and inference will in general be sensitive to both likelihood and prior

choices. When the model is very complex then Bayesian methods are again advantageous since they allow a rigorous treatment of nuisance parameters; MCMC has allowed the consideration of more and more complicated hierarchical models. Spatial models, particularly those that exploit Markov random field second stages, provide a good example of models that are very naturally analysed using MCMC, where the conditional independencies may be exploited. Unfortunately assessments of the effects of model misspecification are rarely carried out for such complex models, instead sensitivity studies are again typically carried out. Bayesian methods are also a good idea in situations in which the maximum likelihood estimator provides a poor summary of the likelihood, for example in variance components problems where the likelihood may be highly skewed.

If n is sufficiently large for asymptotic normality of the sampling distribution to be accurate, then frequentist methods come into their own. In particular, sandwich estimation can be used to provide a consistent estimator of the variance-covariance matrix of the estimator. Hence if the estimator of a parameter of interest is consistent also, reliable confidence coverage will be guaranteed. We stress that n needs to be sufficiently large for the sandwich estimator to be stable. A typical Bayesian approach would be to increase model complexity, often through the introduction of random effects. The difficulty with this is that although this allows more flexibility, a specific form needs to be assumed for the mean-variance relationship, whereas sandwich estimation is consistent in more general situations (quasi-likelihood lies between the two, though there is usually an implicit model underlying the quasi-score function).

5. Non-Linear Regression Models

In this section we briefly review two classes of models, in anticipation of their use in Section 6.

5.1. *Generalized Linear Models*

Generalized linear models (GLMs) were introduced by Nelder and Wedderburn (1972), and the most comprehensive and interesting description is still McCullagh and Nelder (1989); an excellent review is also given by Firth (1994). A GLM is defined by two components:

- (1) The responses y_i follow an exponential family so that the distribu-

tion is of the form

$$p(y_i|\theta_i, \phi) = \exp(\{y_i\theta_i - b(\theta_i)\}/a(\phi) + c(y_i, \phi)),$$

where θ_i and ϕ are scalars. This is sometimes referred to as a *linear* or *natural* exponential family. It is straightforward to show (using the results of Section 2.2) that

$$E[Y_i|\theta_i, \phi] = \mu_i = b'(\theta_i)$$

and

$$\text{var}(Y_i|\theta_i, \phi) = b''(\theta_i)a(\phi),$$

$i = 1, \dots, n$, with $\text{cov}(Y_i, Y_j|\theta_i, \theta_j, \phi) = 0$ for $i \neq j$. This describes the stochastic part of the model.

- (2) We have $g(\mu_i) = \mathbf{x}_i\boldsymbol{\beta}$ where \mathbf{x}_i is $1 \times p$ and $\boldsymbol{\beta}$ is $p \times 1$ so that we have a linear predictor on a scale determined by the so-called link function $g(\cdot)$. This describes the deterministic part of the model.

While computational advances have unshackled the statistician from the need to use GLMs, they are still an extremely useful class of models. The use of the exponential family is advantageous because the score equation can be written

$$a(\phi)\frac{\partial l}{\partial \boldsymbol{\beta}} = \sum_{i=1}^n \frac{\partial l}{\partial \theta_i} \frac{\partial \theta_i}{\partial \boldsymbol{\beta}} = \sum_{i=1}^n \{y_i - \mu_i(\boldsymbol{\beta})\} \frac{\partial \theta_i}{\partial \boldsymbol{\beta}}$$

where $l = l(\boldsymbol{\theta}, \phi)$ is the log-likelihood, and so if the mean is specified correctly the MLE of $\boldsymbol{\beta}$ will be consistent (see the theorem of Section 2.1). Bayes estimators are consistent in this case also (so long as the priors do not exclude a part of the parameter space), due to the asymptotic equivalence between the sampling distribution of the MLE and the posterior distribution (Section 3.1). It is not necessary to have a linear predictor on any particular scale so, for example, the sums of exponentials models of the next section will share this consistency (so long as regularity conditions are satisfied), if the responses arise from the exponential family. So called canonical links in which $\theta_i = \mathbf{x}_i\boldsymbol{\beta}$ provide simplifications in terms of computation.

GLMs are also very useful pedagogically since they separate the deterministic and random components of the model; this aspect was emphasized by Nelder and Wedderburn (1972) who wrote in the abstract: “The implications of the approach in designing statistics courses are discussed”.

5.2. *Compartmental Models*

Pharmacokinetics is the study of the time course of a drug and its metabolites following introduction into this body. In this section we describe a class of models that have been extensively used in such studies to model individual drug concentrations, $y(x)$, as a function of time x . The drug may be introduced into the body via a variety of routes of administration including intravenously (directly into the bloodstream via either a bolus or an infusion), subcutaneously (beneath the skin), or orally. After introduction the drug undergoes the processes of absorption, distribution and elimination. These processes may be modelled by assuming the body consists of a series of homogenous pools or compartments, and then considering a set of differential equations that determine the rate of flow of drug between the different compartments, see Gibaldi and Perrier (1982) for a comprehensive account of pharmacokinetic models and principles, and Godfrey (1983) for an account of compartmental modelling in general.

As a simple example consider a model with a single compartment for the distribution and elimination, and an oral dose; we make use of this model in Section 6. We may think, nominally, of the compartment corresponding to the blood; in general pharmacokinetic modelling via a compartmental system is a convenient visualisation but the compartments often have no physiological meaning, rather physiological parameters such as the time to maximum concentration, maximum concentration, elimination half-life and clearance are of interest. These parameters are defined in Section 6.

Let $w_i(x)$ represent the amount of drug in compartment i , $i = 0, 1$, at time x , with compartment 0 representing the site from which absorption occurs. The differential equations describing the drug flow between the compartments may be assumed to be of the form

$$\frac{dw_0}{dx} = -k_a w_0 \quad (21)$$

$$\frac{dw_1}{dx} = k_a w_0 - k_e w_1 \quad (22)$$

where k_a is the absorption rate constant associated with flow from compartment 0 to compartment 1, and k_e is the elimination rate constant. Assuming that $w(0) = D$ is the dose at time zero and that the (apparent) volume of distribution (which converts total amount of drug into concentration) is V we may solve (22) to obtain the time course of the *concentration*,

$\mu(x)$, as

$$\mu(x) = \frac{Dk_a}{V(k_a - k_e)} \{\exp[-k_e x] - \exp[-k_a x]\}. \quad (23)$$

This model is sometimes known as the flip-flop model because there is a basic identifiability in that the same curve is achieved with the parameter sets (V, k_a, k_e) and $(Vk_e/k_a, k_e, k_a)$, and it is often assumed that $k_a > k_e$ in order to enforce identifiability.

We now consider the stochastic part of the model. In addition to measurement error in the assay technique, errors are introduced due to model misspecification (particularly at later phases of drug development which are carried out in a poorly controlled environment, and so the reported sampling times may be subject to error, for example). Assay precision is often found to increase with increasing true concentrations and models of the form

$$y(x) = \mu(x) + \delta(x),$$

where $\delta(x) \sim N\{0, \mu(x)^\gamma \sigma_\delta^2\}$ with $\gamma > 0$ have been used. The variance power γ is either fixed, with $\gamma = 2$ being a common choice (to produce a constant coefficient of variation), or estimated. A constant coefficient of estimation can also be approximately achieved by taking

$$\log y(x) = \log \mu(x) + \epsilon(x),$$

with $\epsilon(x) \sim N(0, \sigma^2)$.

Wakefield et al. (1999) provide a review of pharmacokinetic and pharmacodynamic modelling including more details on both biological and statistical aspects.

6. Pharmacokinetic Data Analysis

An oral dose of 1mg of Theophylline was administered to a new born baby, and concentration time data (x_i, y_i) were subsequently collected for $i = 1, \dots, 8$. These data were previously analyzed by Wakefield (1992), and are reproduced in Table 1.

Traditionally, the so-called one-compartment open model, as described in Section 5.2 would be fitted to these data. Under this model the concentration at time x is given by (23) which we reproduce here, along with an alternative form, in order to motivate a log-linear inverse polynomial

model:

$$\begin{aligned}\mu(x) &= \frac{Dk_a}{V(k_a - k_e)} \{\exp[-k_e x] - \exp[-k_a x]\} \\ &= D \exp(\beta_0 + \beta_1 x) \{1 - \exp[-(k_a - k_e)x]\},\end{aligned}\quad (24)$$

where $\beta_0 = \log\{k_a/(V(k_a - k_e))\}$ and $\beta_1 = -k_e$. Typically, interest does not focus upon (V, k_a, k_e) , but rather on the derived parameters:

- The elimination half-life, which is the time it takes for the drug concentration to drop by 50%, when elimination is the dominant process:

$$x_{1/2} = (\log 2)/k_e.$$

- The time to maximum

$$x_{\max} = \frac{1}{k_a - k_e} \log \left(\frac{k_a}{k_e} \right).$$

- The maximum concentration

$$\begin{aligned}\mu(x_{\max}) &= \frac{Dk_a}{V(k_a - k_e)} \{\exp(-k_e x_{\max}) - \exp(-k_a x_{\max})\} \\ &= \frac{D}{V} \left(\frac{k_a}{k_e} \right)^{k_a/(k_a - k_e)}.\end{aligned}$$

- The clearance, which is the amount of blood cleared of drug in unit time, is given by $Cl = D/\text{AUC}$ where AUC is the area under the concentration time curve so that

$$Cl = V \times k_e.$$

We assume that

$$\log y_i = \log \mu_i(x_i) + \epsilon_i,$$

where $\epsilon_i \sim_{iid} N(0, \sigma^2)$.

As an alternative to the above compartmental model, we here fit the log-linear fractional polynomial model, a GLM, given by

$$\mu(x) = D \exp(\beta_0 + \beta_1 x + \beta_2/x).$$

Comparison with (24) shows that β_2 is the parameter that determines the absorption; the model only makes sense if it produces an increasing absorption phase and a decreasing elimination phase which correspond, respectively, to $\beta_2 < 0$ and $\beta_1 < 0$. This model is very much in the spirit

of Nelder (1991) in which the inverse polynomial form was suggested as a model for enzyme-kinetic data.

Each of the derived parameters are functions of $\beta^T = (\beta_0, \beta_1, \beta_2)$. Specifically:

- The elimination half-life

$$x_{1/2} = -(\log 2)/\beta_1.$$

- The time to maximum

$$x_{\max} = (\beta_2/\beta_1)^{1/2}.$$

- The maximum concentration

$$\mu(x_{\max}) = D \exp\{\beta_0 - 2(\beta_1\beta_2)^{1/2}\}.$$

- From the definition as D/AUC the clearance is given by

$$Cl = \frac{\sqrt{\beta_1/\beta_2}}{2 \exp(\beta_0) \text{BesselK}\{2(\beta_1\beta_2)^{1/2}\}}, \quad (25)$$

where BesselK denotes a modified Bessel function of the third kind.

The data are assumed to be gamma distributed, specifically $Y_i \sim_{ind} \text{Ga}\{\phi^{-1}, (\mu_i\phi)^{-1}\}$, so that $\phi^{1/2}$ is the coefficient of variation. Lindsey et al. (2000) examine various distributional choices for pharmacokinetic data, and found the gamma assumption to be reasonable for their examples. A more extensive discussion of the application of this model to pharmacokinetic data may be found in Wakefield (2004). One disadvantage of this model compared to compartmental models is that if multiple doses are considered the mean function does not correspond to a GLM.

The lognormal compartmental and gamma log-linear models were fitted in \mathbb{R} , with maximum likelihood estimation of the fixed effects, and the moment estimator, (11), for the dispersion parameter. Confidence intervals based on the asymptotic distribution of the MLE were calculated for the parameters of interest using the delta method. These parameters are all positive and so the intervals were obtained for the log transforms, and then exponentiated (the delta method for the clearance under the log-linear model was not used because of the intractability of the calculations, the sampling-based Bayesian approach that we describe shortly is straightforward, however). The results are displayed in Table 2, and the fitted curves in Figure 1. Each of these summaries shows a remarkable level of agreement. The maximized log-likelihoods were -21.58 for the gamma model

and -20.89 for the lognormal model; these models are not nested and so a likelihood ratio statistic is not available, but the use of AIC is valid and suggests no significant differences between the models.

We now describe a Bayesian implementation of each of these models using the rejection algorithm described in Section 3.4. We place prior distributions on the half-life, $x_{1/2}$, time to maximum, x_{\max} , maximum concentration, $\mu(x_{\max})$ and coefficient of variation; this is more natural for each of the models (and in particular for the log-linear model within which β_2 is not straightforward to interpret). Another benefit of specifying the prior in terms of model-free parameters is that models may be compared using Bayes factors on an “even playing field”, in the sense that the prior input for each model is identical. For more discussion of this issue, see Pérez and Berger (2002). We choose independent lognormal priors for these four parameters. For a generic parameter, θ , denote the prior by $\theta \sim \text{Lognormal}(\mu, \sigma)$. To obtain the moments of these distributions we specified the prior median, θ_m , and the 95% point of the prior, θ_u , and then solved for the moments via:

$$\mu = \log(\theta_m), \quad \sigma = \{\log(\theta_u) - \mu\}/1.645.$$

The third line of Table 2 gives the illustrative prior choices; samples were simulated from the prior in order to empirically estimate the quantiles of the induced prior for Cl . This prior could be criticised for the assumption of independence; it would be straightforward to specify a multivariate lognormal, however, perhaps with the moments being based on a population pharmacokinetic analysis of a group of patients who are thought to be exchangeable with the specific subject being considered.

To implement the rejection algorithm we sample from the prior on the parameters of interest, and then back-solve for the parameters that describe the likelihood. For the compartmental model we transform back to the original parameters via

$$\begin{aligned} k_e &= (\log 2)/x_{1/2} \\ 0 &= x_{\max}(k_a - k_e) - \log\left(\frac{k_a}{k_e}\right) \\ V &= \frac{D}{\mu(x_{\max})} \left(\frac{k_a}{k_e}\right)^{k_a/(k_a - k_e)} \end{aligned} \quad (26)$$

so that k_a is not directly available, but must be obtained as the root of (26). For the log-linear model the transformation to β is via

$$\beta_1 = -\log 2/x_{1/2}, \quad \beta_2 = \beta_1 x_{\max}^2, \quad \beta_0 = \log \mu(x_{\max}) + 2(\beta_1 \beta_2)^{1/2}.$$

The rejection algorithm described in Section 3.4 was used, with the MLEs for the fixed effects obtained from the analyses reported earlier (and replacing the method of moments estimators with the MLEs for the dispersion parameters), and 500 samples being generated from the posterior distributions; the acceptance rates were 0.0030 and 0.0015 for the gamma and lognormal models, respectively. Table 2 summarizes inference for the parameters of interest with the interval estimates and medians being obtained as the sample quantiles. Note that inference for the clearance is straightforward since samples can be substituted directly in to the form (25). Figures 2 and 3 show the posteriors for the functions of interest under both models. These figures and Table 2 show that Bayesian inference under both of the models is very similar; frequentist and Bayesian methods are also in close agreement for this example. The posteriors are skewed for all functions of inference apart from the clearance parameter, indicating that the posterior medians are more representative than the MLEs. The clearance parameter is often found to be well-behaved, since it is a function of the area under the curve, which is very stably estimated.

We evaluated the normalizing constants using (20), and calculated the Bayes factor comparing the gamma and lognormal models (denoted M_G and M_L , respectively) on the \log_2 -base scale (which is suggested by Kass and Raftery (1995)), to obtain $\log_2 p(\mathbf{y} | M_G) - \log_2 p(\mathbf{y} | M_L) = -39.56 - (-39.52) = -0.04$, showing no significant differences between the models, in agreement with the AIC conclusion described earlier.

Residuals were examined to access the appropriateness of the mean function, the mean-variance relationship, and the distribution of the errors, and no clear inadequacy was seen, though with $n = 8$ data points, such checks are difficult to interpret.

7. Discussion

In this paper a review of parametric non-linear modelling has been presented, with both frequentist and Bayesian approaches to inference being described. It has been argued that models should first arise from the context, with mathematical and computational aspects being subsequently examined. The computational convenience of GLMs is a major benefit, and since their introduction in Nelder and Wedderburn (1972) GLMs have been widely used in an array of contexts, a testimony to their flexibility and their continued competitiveness with the increased array of models that are now computationally feasible for the practicing statistician. GLMs also have de-

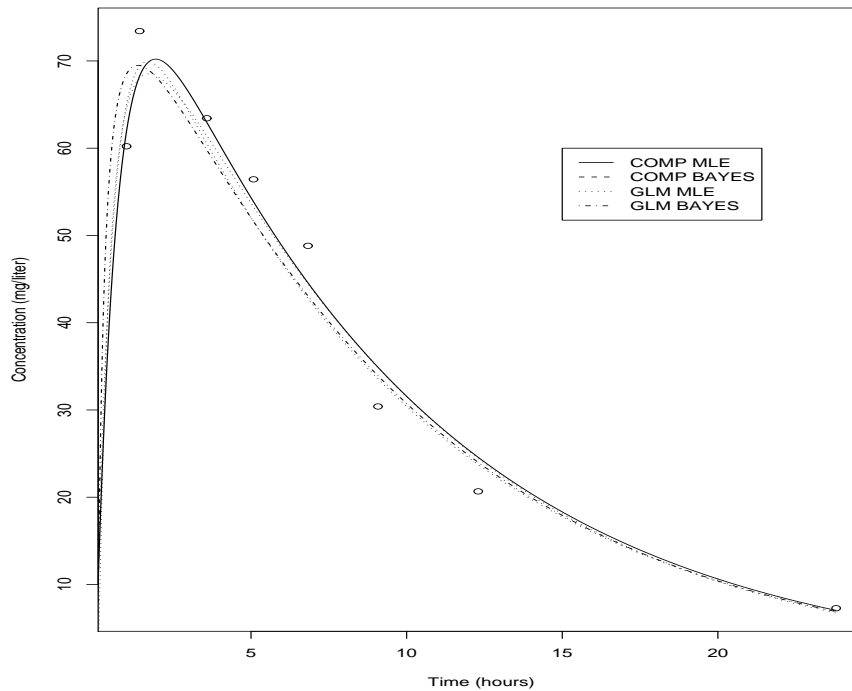


Figure 1. Fitted curves for Theophylline data.

sirable statistical properties; in particular the use of the linear exponential family yields consistent estimators from likelihood or Bayesian approaches, so long as the mean model is correctly specified.

We also described a simple rejection algorithm that may be used to produce independent samples from the posterior distribution and is very convenient in situations in which informative prior distributions are available, and the maximised likelihood can be simply calculated. The advantages of such sampling-based approaches have also been illustrated, in particular, inference for functions of interest is straightforward.

Acknowledgments

The author would like to thank John Nelder for discussions that greatly helped in the formulation of the models that were used for the pharmacokinetic data example.

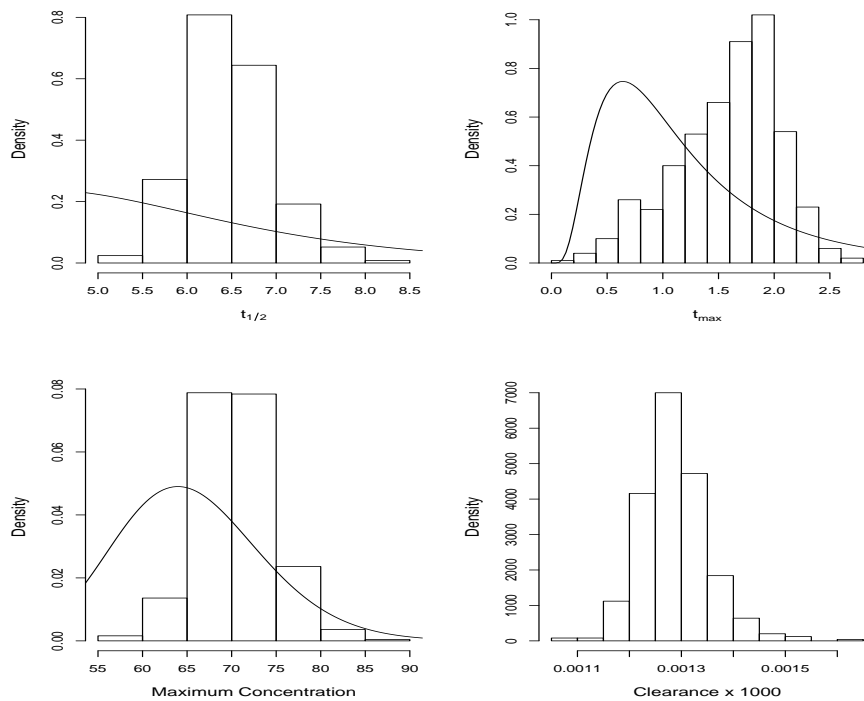


Figure 2. Histogram representations of posterior distributions for the compartmental model; solid curves denote the lognormal prior distributions.

References

- Berger, J.O. (2003). Could Fisher, Jeffreys and Neyman have agreed on testing? *Statistical Science* 18, 1–32.
- Berger, J.O. and Wolpert, R.L. (1988). *The Likelihood Principle: A Review, Generalizations, and Statistical Implications*. Hayward: Institute of Mathematical Statistics.
- Bernardo, J.M. and Smith, A.F.M. (1994). *Bayesian Theory*. John Wiley and Sons, New York.
- Box, G.E.P. and Tiao, G.C. (1964). A note on criterion robustness and inference robustness. *Journal of the Royal Statistical Society, Series B* 51, 169–173.
- Breiman, L. (2001). Statistical modeling: The two cultures (with discussion). *Statistical Science* 16, 199–231.

32 REFERENCES

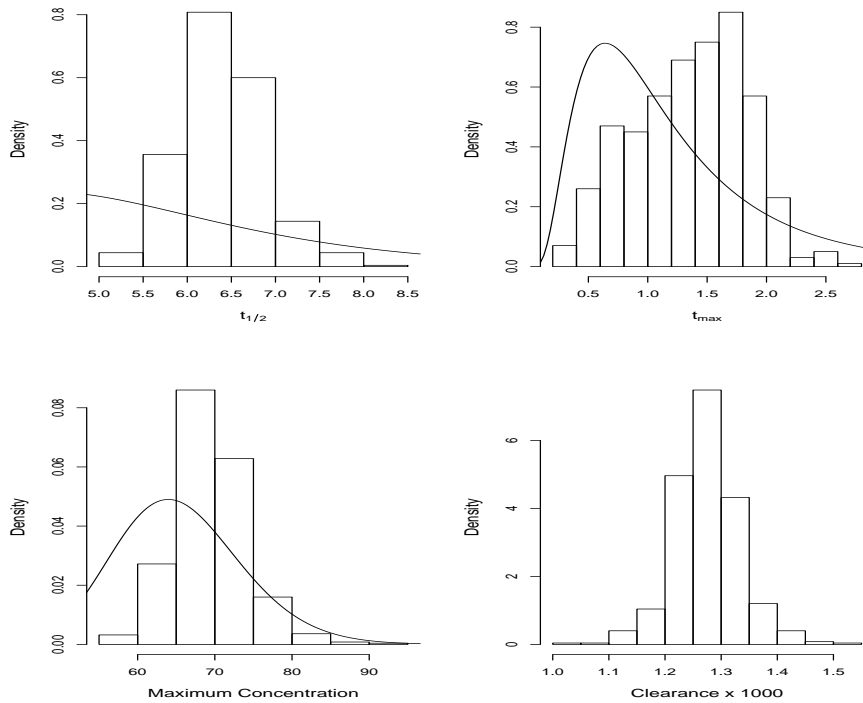


Figure 3. Histogram representations of posterior distributions for the log-linear inverse polynomial model; solid curves denote the lognormal prior distributions.

- Carroll, R.J., Ruppert, D., and Stefanski, L.A. (1995). *Measurement Error in Nonlinear Models*. Chapman and Hall/CRC, Boca Raton.
- Chatfield, C. (1995). Model uncertainty, data mining and statistical inference (with discussion). *Journal of the Royal Statistical Society, Series A 158*, 419–466.
- Crowder, M. (1986). On consistency and inconsistency of estimating equations. *Econometric Theory* 2, 305–330.
- Crowder, M. (1987). On linear and quadratic estimating functions. *Biometrika* 74, 591–597.
- Davison, A.C. and Hinkley, D.V. (1997). *Bootstrap Methods and their Application*. Cambridge University Press.
- Dellaportas, P. and Smith, A.F.M. (1993). Bayesian inference for generalized linear and proportional hazards models via Gibbs sampling. *Applied*

REFERENCES 33

- Statistics* 42, 443–459.
- Devroye, L. (1986). *Non-Uniform Random Variate Generation*. Springer-Verlag, New York.
- DiCiccio, T.J., Kass, R.E., Raftery, A., and Wasserman, L. (1997). Computing Bayes Factors by Combining Simulation and Asymptotic Approximations. *Journal of the American Statistical Association* 92, 903–915.
- Draper, D. (1995). Assessment and propagation of model uncertainty (with discussion). *Journal of the Royal Statistical Society, Series B* 57, 45–97.
- Efron, B. and Tibshirani, R.J. (1993). *An Introduction to the Bootstrap*. Chapman and Hall/CRC, Boca Raton.
- Evans, M. and Swartz, T. (1995). Rejoinder to: Methods for approximating integrals in statistics with special emphasis on Bayesian integration problems. *Statistical Science* 11, 54–64.
- Ferguson, T.S. (1996). *A Course in Large Sample Theory*. Chapman and Hall/CRC.
- Firth, D. (1994). Recent developments in quasi-likelihood methods. In *Proceedings of the ISI 49th Session*, pp. 341–358.
- Fisher, R.A. (1921). On the "probable error" of a coefficient of correlation deduced from a small sample. *Metron* 1, 3–32.
- Gelman, A., Carlin, J.B., Stern, H.S., and Rubin, D.B. (1995). *Bayesian Data Analysis*. Chapman and Hall, London.
- Gelman, A. and Meng, X.L. (1995). Discussion of: Assessment and propagation of model uncertainty, by D. Draper. *Journal of the Royal Statistical Society, Series B* 57, 83.
- Gibaldi, M. and Perrier, D. (1982). *Drugs and the Pharmaceutical Sciences, Volume 15: Pharmacokinetics, Second Edition*. Marcel Dekker.
- Godambe, V.P. and Heyde, C.C. (1987). Quasi-likelihood and optimal estimation. *International Statistical Review* 55, 231–244.
- Godfrey, K.R. (1983). *Compartmental Models and their Applications*. Academic Press, London.
- Hjort, N.L. (1994). The exact amount of t -ness that the normal model can tolerate. *Journal of the American Statistical Association* 89, 665–675.
- Huber, P.J. (1967). The behavior of maximum likelihood estimators under non-standard conditions. In L.M. LeCam and J. Neyman (Eds.), *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, pp. 221–233. University of California Press.
- Johnson, R.A. (1970). Asymptotic expansions associated with posterior distributions. *The Annals of Mathematical Statistics* 41, 851–864.
- Kass, R.E. and Raftery, A.E. (1995). Bayes factors. *Journal of the American*

34 REFERENCES

- Statistical Association* 90, 773–795.
- Kauermann, G. and Carroll, R.J. (2001). A note on the efficiency of sandwich covariance matrix estimation. *Journal of the American Statistical Association* 96, 1387–1396.
- Liang, K.Y. and Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* 73, 13–22.
- Lindsey, J.K., Byrom, W.D., Wang, J., Jarvis, P., and Jones, B. (2000). Generalized nonlinear models for pharmacokinetic data. *Biometrics* 56, 81–88.
- McCullagh, P. (1983). Quasi-likelihood functions. *The Annals of Statistics* 11, 59–67.
- McCullagh, P. (1991). Quasi-likelihood and estimating functions. In D.V. Hinkley, N. Reid, and E.J. Snell (Eds.), *Statistical Theory and Modelling*, pp. 265–286. Chapman and Hall.
- McCullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models, Second Edition*. Chapman and Hall/CRC, Boca Raton.
- Nelder, J.A. (1966). Inverse polynomials, a useful group of multi-factor response functions. *Biometrics* 22, 128–141.
- Nelder, J.A. (1991). Generalized linear models for enzyme-kinetic data. *Biometrics* 47, 1605–1615.
- Nelder, J.A. and Wedderburn, R.W.M. (1972). Generalized Linear Models. *Journal of the Royal Statistical Society, Series A* 135, 370–384.
- O’Hagan, A. (1994). *Kendall’s Advanced Theory of Statistics, Volume 2B: Bayesian Inference*. Arnold, London.
- Pauler, D.K., Wakefield, J.C., and Kass, R.E. (1999). Bayes factors for variance component models. *Journal of the American Statistical Association* 94, 1242–1253.
- Pérez, J.M. and Berger, J.O. (2002). Expected-posterior prior distributions for model selection. *Biometrika* 89, 491–512.
- Reid, N. (1995). The roles of conditioning in inference. *Statistical Science* 10, 138–199.
- Royall, R. (1986). Model robust confidence intervals using maximum likelihood estimators. *International Statistical Review* 54, 221–226.
- Royall, R. (1997). *Statistical Evidence – A Likelihood Paradigm*. Chapman and Hall/CRC, Boca Raton.
- Smith, A.F.M. and Gelfand, A.E. (1992). Bayesian statistics without tears: a resampling approach. *The American Statistician* 46, 84–88.
- van der Vaart, A.W. (1998). *Asymptotic Statistics*. Cambridge University Press, Cambridge.

REFERENCES 35

- Wakefield, J.C. (1992). *The Bayesian Analysis of Pharmacokinetic Models*. Ph. D. thesis, Nottingham University.
- Wakefield, J.C. (1996). Bayesian individualization via sampling-based methods. *Journal of Pharmacokinetics and Biopharmaceutics* 24, 103–131.
- Wakefield, J.C. (2004). Gamma generalized linear models for pharmacokinetic data. Technical report, Department of Biostatistics.
- Wakefield, J.C., Aarons, L., A., and Racine-Poon (1999). The Bayesian approach to population pharmacokinetic/pharmacodynamic modelling. In Gatsonis C., Kass R.E., Carlin B.P., Carriquiry A.L., Gelman A., Verdinelli I., and West M. (Eds.), *Case Studies in Bayesian Statistics, Volume IV*, pp. 205–265. Springer-Verlag, New York.
- Walker, A.M. (1970). On the asymptotic behaviour of posterior distributions. *Journal of the Royal Statistical Society, Series B* 31, 80–88.
- Wedderburn, R.W.M. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika* 61, 439–447.
- White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica* 48, 1721–746.
- White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica* 50, 1–26.
- Zeger, S.L. and Liang, K.Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics* 42, 121–130.

36 REFERENCES

Table 1. Concentration (y) as a function of time (x), obtained from a new-born baby following the administration of a 1mg dose of Theophylline.

i	x_i (hours)	y_i (mg/liter)
1	1.00	60.22
2	1.42	73.41
3	3.58	63.43
4	5.08	56.43
5	6.83	48.81
6	9.08	30.40
7	12.3	20.67
8	23.8	7.28

Table 2. Point and 90% interval estimates for the data of Table 1, under various models and estimation techniques; Cl denotes the clearance and CV the coefficient of variation, the latter expressed as a percentage. The Bayesian point estimates correspond to the posterior medians.

Model	$x_{1/2}$	x_{max}	$\mu(x_{max})$	Cl ($\times 10^3$)	CV ($\times 10^2$)
Comp MLE	6.27 (5.66,6.95)	1.87 (1.39,2.53)	70.5 (56.3,88.3)	1.28 (1.19,1.36)	11.3 (7.46,17.0)
GLM MLE	6.12 (4.46,8.39)	1.72 (1.36,2.17)	68.5 (53.0,88.5)	1.27 (-,-)	9.68 (7.89,11.9)
Prior	5.00 (2.78,9.00)	1.00 (0.333,3.00)	65.0 (52.8,80.0)	1.50 (3.29,13.9)	10.0 (2.50,40.0)
Comp Posterior	6.44 (5.74,7.28)	1.69 (0.700,2.23)	70.4 (64.4,78.1)	1.28 (1.19,1.40)	12.3 (7.82,20.3)
GLM Posterior	6.37 (5.69,7.18)	1.40 (0.556,2.06)	69.2 (62.6,77.2)	1.27 (1.16,1.37)	12.5 (7.95,21.2)