Journal of the
Royal Statistical Society

SERIES B
Statistical
Methodology

J. R. Statist. Soc. B (2008)
70, *Part* 4, *pp.* 643–677

Sampling bias and logistic models

Peter McCullagh

University of Chicago, USA

[*Read before* The Royal Statistical Society *at a meeting organized by the* Research Section *on Wednesday*, *February 6th, 2008*, Professor I. L. Dryden *in the Chair*]

Summary. In a regression model, the joint distribution for each finite sample of units is determined by a function $p_{\mathbf{x}}(\mathbf{y})$ depending only on the list of covariate values $\mathbf{x} = (x(u_1),\ldots,x(u_n))$ on the sampled units. No random sampling of units is involved. In biological work, random sampling is frequently unavoidable, in which case the joint distribution $p(\mathbf{y}, \mathbf{x})$ depends on the sampling scheme. Regression models can be used for the study of dependence provided that the conditional distribution $p(\mathbf{y}|\mathbf{x})$ for random samples agrees with $p_{\mathbf{x}}(\mathbf{y})$ as determined by the regression model for a fixed sample having a non-random configuration $\mathbf{x}$. The paper develops a model that avoids the concept of a fixed population of units, thereby forcing the sampling plan to be incorporated into the sampling distribution. For a quota sample having a predetermined covariate configuration $\mathbf{x}$, the sampling distribution agrees with the standard logistic regression model with correlated components. For most natural sampling plans such as sequential or simple random sampling, the conditional distribution $p(\mathbf{y}|\mathbf{x})$ is not the same as the regression distribution unless $p_{\mathbf{x}}(\mathbf{y})$ has independent components. In this sense, most natural sampling schemes involving binary random-effects models are biased. The implications of this formulation for subject-specific and population-averaged procedures are explored.

Keywords: Autogenerated unit; Correlated binary data; Cox process; Estimating function; Interference; Marginal parameterization; Partition model; Permanent polynomial; Point process; Prognostic distribution; Quota sample; Random-effects model; Randomization; Self-selection; Size-biased sample; Stratum distribution

1. Introduction

Regression models are the primary statistical tool for studying the dependence of a response Y on covariates x in a population $\mathcal{U}$. For each finite sample of units or subjects $u_1, \ldots, u_n$, a regression model specifies the joint distribution $p_{\mathbf{x}}(\mathbf{y})$ of the response $\mathbf{y} = (Y(u_1), \ldots, Y(u_n))$ on the given units. Implicit in the notation is the exchangeability assumption, that two samples having the same list of covariate values have the same joint distribution $p_{\mathbf{x}}(\mathbf{y})$. All generalized linear models have this property, and many correlated Gaussian models have the same property, e.g.

$$p_{\mathbf{x}}(A) = N_n(X\beta, \sigma_0^2 I_n + \sigma_1^2 K[\mathbf{x}])(A), \qquad (1)$$

where $N_n(\mu, \Sigma)(A)$ is the probability assigned by the n-dimensional Gaussian distribution to the event $A \subset \mathcal{R}^n$. The mean $\mu = X\beta$ is determined by the covariate matrix X, and $K_{ij}[\mathbf{x}] = K\{x(u_i), x(u_j)\}$ is a covariance function evaluated at the points $\mathbf{x}$.

Depending on the area of application, it may happen that the target population is either unlabelled, or random in the sense that the units are generated by the process as it evolves. Consider, for example, the problem of estimating the distribution of fibre lengths from a specimen of woollen or cotton yarn, or the problem of estimating the distribution of speeds of highway

Address for correspondence: Peter McCullagh, Department of Statistics, University of Chicago, 5734 University Avenue, Chicago, IL 60657-1514, USA.
E-mail: pmcc@galton.uchicago.edu

1369–7412/08/70643

vehicles. Individual fibres are clearly unlabelled, so it is necessary to select a random sample, which might be size biased. Highway vehicles may be labelled by registration number, but the target population is weighted by frequency or intensity of highway use, so the units (travelling vehicles) are generated by the process itself. In many areas of application, the set of units evolves randomly in time, e.g. human or animal populations. The concept of a fixed subset makes little sense physically or mathematically, so random samples are inevitable. The sample might be obtained on the fly by sequential recruitment in a clinical trial, or by recording passing vehicles at a fixed point on the highway, or it might be obtained by simple random sampling, or by a more complicated ascertainment scheme in studies of genetic diseases. The observation from such a sample is a random variable, possibly bivariate, whose distribution depends on the sampling protocol. In the application of regression models, it is often assumed that the joint distribution $p(\mathbf{x},\mathbf{y})$ is such that the conditional distribution $p(\mathbf{y}|\mathbf{x})$ is the same as the distribution $p_{\mathbf{x}}(\mathbf{y})$ determined by the regression model for a sample having a predetermined covariate configuration. The main purpose of this paper is to reconsider this assumption in the context of binary and polytomous regression models that incorporate random effects or correlation between units.

2. Binary regression models

The conventional, most direct, and apparently most natural way to incorporate correlation into a binary response model is to include additive random effects on the logistic scale (Laird and Ware, 1982; McCullagh and Nelder, 1989; Breslow and Clayton, 1993; McCulloch, 1994, 1997; Lee *et al.*, 2006). The random effects in a hierarchical model need not be Gaussian, but a generalized linear mixed model of that type with a binary response Y and a real-valued covariate x suffices to illustrate the idea. The first step is to construct a Gaussian process η on $\mathcal{R}$ with zero mean and covariance function K. For example, we might have $K(x,x')=\sigma^2\exp(-|x-x'|/\tau)$, so that η is a continuous random function. Alternatively, K could be a block factor expressed as a Boolean matrix, so that η is constant on blocks or clusters, with block effects that are independent and identically distributed. Given η, the components of Y are independent and are such that

$$\text{logit}[\text{pr}\{Y(u)=1|\eta\}]=\alpha+\beta x(u)+\eta\{x(u)\}, \tag{2}$$

where $Y(u)$ is the response and $x(u)$ the covariate value on unit u. As a consequence, two units having the same or similar covariate values have identical or similar random contributions $\eta\{x(u)\}$ and $\eta\{x(u')\}$, and the responses $Y(u)$ and $Y(u')$ are positively correlated. Since η is a random variable, the joint density at $\mathbf{y}=(y_1,\ldots,y_n)$ for any fixed sample of n units having covariate values $\mathbf{x}=(x_1,\ldots,x_n)$ is

$$p_{\mathbf{x}}(\mathbf{y})=\int_{\mathcal{R}^n}\prod_{j=1}^{n}\frac{\exp\{(\alpha+\beta x_j+\eta_j)y_j\}}{1+\exp(\alpha+\beta x_j+\eta_j)}\phi(\eta;K)\,\mathrm{d}\eta. \tag{3}$$

The word model refers to these distributions, not to the random variable (2). In this instance we obtain a four-parameter regression model with parameters $(\alpha,\beta,\sigma,\tau)$.

The simplest polytomous version of model (3) requires k correlated processes, $\eta_0(x),\ldots,\eta_{k-1}(x)$, one for each class. The joint probability distribution

$$p_{\mathbf{x}}(\mathbf{y})=\int_{R^{nk}}\prod_{j=1}^{n}\frac{\exp\{\alpha_{y_j}+\beta_{y_j}x_j+\eta_{y_j}(x_j)\}}{\sum_0^{k-1}\exp\{\alpha_r+\beta_r x_j+\eta_r(x_j)\}}\phi(\eta;K)\,\mathrm{d}\eta \tag{4}$$

depends only on the distribution of differences $\eta_r(x) - \eta_0(x)$. Setting $\alpha_0 = \beta_0 = \eta_0(x) = 0$ introduces the asymmetry in model (3), but no loss of generality. In the econometrics literature, model (4) is known as a discrete choice model with random effects, and $\mathcal{C} = \{0, \ldots, k-1\}$ is the set of mutually exclusive choices or brand preferences, which is seldom exhaustive.

The term *regression model* used in connection with distributions (1), (3) and (4) does not imply independence of components, but it does imply *lack of interference* in the sense that the covariate value $x' = x(u')$ on one unit has no effect on the response distribution for other units (Cox (1958), section 2.4, and McCullagh (2005)). The mathematical definition for a binary model is

$$p_{\mathbf{x},x'}(\mathbf{y}, 0) + p_{\mathbf{x},x'}(\mathbf{y}, 1) = p_{\mathbf{x}}(\mathbf{y}), \tag{5}$$

which is satisfied by model (3) regardless of the distribution of η. Here $p_{\mathbf{x},x'}(\cdot)$ is the response distribution for a set of $n+1$ units, the first n of which have covariate vector $\mathbf{x}$. For further discussion, see Sections 6 and 8.1.

In any extension of a regression model to a bivariate process, two possible interpretations may be given to the functions $p_{\mathbf{x}}(\mathbf{y})$. Given $\mathbf{x} = (x_1, \ldots, x_n)$, the *stratum distribution* is the marginal distribution of $Y(u_1), \ldots, Y(u_n)$ for a random set of units selected so that $x(u_i) = x_i$. Ordinarily, this is different from the conditional distribution $p_n(\mathbf{y}|\mathbf{x})$ for a fixed set of n units having a random configuration $\mathbf{x}$. Stratum distributions automatically satisfy the no-interference property, so the most natural extension uses $p_{\mathbf{x}}(\mathbf{y})$ for stratum distributions, as in Section 3. In the conventional *hierarchical* extension, the two distributions are equal, and the regression model $p_{\mathbf{x}}(\mathbf{y})$ serves both purposes.

The distinction between conditional distribution and stratum distribution is critical in much of what follows. If the units were generated by a random process or selected by random sampling, then $\mathbf{x}$ would indeed be a random variable whose distribution depends on the sampling plan. In a marketing study, for example, it is usual to focus on the subset of consumers who actually purchase one of the study brands, in which case the sample units are generated by the process itself. Participants in a clinical trial are volunteers who satisfy the eligibility criteria and give informed consent. The study units are not predetermined but are generated by a random process. Such units, whether they be patients, highway vehicles or purchase events, are called *autogenerated*; the non-mathematical term *self-selected* is too anthropocentric for general use. Without careful verification, we should not expect the conditional distribution $p_n(\mathbf{y}|\mathbf{x})$ for autogenerated units to coincide with $p_{\mathbf{x}}(\mathbf{y})$ for a predetermined configuration $\mathbf{x}$. We could, of course, extend the regression model (4) to an exchangeable bivariate process by asserting that the components of $\mathbf{x}$ are independent and identically distributed with $p_{\mathbf{x}}(\mathbf{y})$ as the conditional distribution. This extension guarantees $p_n(\mathbf{y}|\mathbf{x}) = p_{\mathbf{x}}(\mathbf{y})$ by *fiat*, which is conventional but not necessarily natural. It does not address the critical modelling problem, that labels are usually affixed to the units *after* they have been generated by the process itself.

In principle, the parameters in models (3) or (4) can be estimated in the standard way by using the marginal likelihood function, either by maximization or by using a formal Bayesian model with a prior distribution on $(\alpha, \beta, \sigma, \tau)$. Alternatively, it may be possible for some purposes to avoid integration by using a Laplace approximation or penalized likelihood function along the lines of Schall (1991), Breslow and Clayton (1993), Wolfinger (1993), Green and Silverman (1994) or Lee and Nelder (1996).

The binary model (3) and the polytomous version (4) are satisfactory in many ways, but they suffer from at least four defects as follows.

(a) Parameter attenuation: suppose that $x = (z, x')$ has several components, one of which is the treatment status, and that βx is a linear combination. The odds of success are $p_{(1,x')}(1)/p_{(1,x')}(0)$ for a treated unit having baseline covariate value x', and $p_{(0,x')}(1)/p_{(0,x')}(0)$ for an untreated unit, and the treatment effect is the ratio of these numbers. In ordinary linear logistic models with independent components, the coefficient of treatment status is the treatment effect on the log-scale. However, the treatment effect in model (3) is a complicated function of all parameters. In itself, this is not a serious drawback, but it does complicate inferential statements about the principal target parameter if model (3) is taken seriously.

(b) Class aggregation: suppose that two response classes r and s in model (4) are such that $\alpha_r = \alpha_s$, $\beta_r = \beta_s$ and $(\eta_r, \eta_s) \sim (\eta_s, \eta_r)$ have the same distribution. Although these classes are homogeneous, the marginal distribution after aggregation of classes is not of the same form. In other words, the binary model (3) cannot be obtained from (4) by aggregation of homogeneous classes.

(c) Class restriction: suppose that the number of classes in model (4) is initially large, but we choose to focus on a subset, ignoring the remainder. In a study of causes of death, for example, we might focus on cancer deaths, ignoring deaths due to other causes. Patients dying of cancer constitute a random subset of all deaths, so the x-values and y-values are both random, with distribution determined implicitly by model (4). On this random subset, the conditional distribution of $\mathbf{y}$ given $\mathbf{x}$ does not have the form (4). In particular, the binary model (3) cannot be obtained from (4) by restriction of response classes.

(d) Sampling distributions: if the sampling procedure is such that the number of sampled units or configuration of x-values is random, the conditional distribution of the response on the sampled units $p_n(\mathbf{y}|\mathbf{x})$ may be different from model (3).

Parameter attenuation is not, in itself, a serious defect. The real defect lies in the fact that, for many natural sampling protocols, parameter attenuation is a statistical artefact stemming from inappropriate model assumptions. The illusion of attenuation is attributable to sampling bias, the fact that the sample units are not predetermined but are generated by a random process that the conventional hierarchical model is incapable of taking into account. The distinction that is frequently drawn between subject-specific effects and population-averaged effects (Zeger *et al.*, 1988; Galbraith, 1991) is a manifestation of the same phenomenon (Section 8.2).

3. An evolving population model

3.1. The process

Let $\mathcal{X}$ be the covariate space, and let ν be a measure in $\mathcal{X}$ such that ν is finite and positive on non-empty open sets. In other words $0 < \nu(\mathcal{X}) < \infty$, and $\tilde{\nu}(\mathrm{d}x) = \nu(\mathrm{d}x)/\nu(\mathcal{X})$ is a probability distribution on $\mathcal{X}$ with positive density at each point. In addition, $\mathcal{C} = \{0, \ldots, k-1\}$ is the set of response classes, and $\lambda(r, x)$ is the value at (r, x) of a random intensity function on $\mathcal{C} \times \mathcal{X}$, positive and bounded. For notational convenience we write $\lambda_r(x)$ for $\lambda(r, x)$ and $\lambda_{\bullet}(x) = \Sigma\, \lambda_r(x)$ for the total intensity at x. A Poisson process in $\mathcal{C} \times \mathcal{X} \times (0, \infty)$ evolves at a constant temporal rate $\lambda(r, x)\, \nu(\mathrm{d}x)\, \mathrm{d}t$. These events constitute the target population, which is random, infinite and unlabelled.

Let $\mathbf{Z}_t$ be the set of events occurring before time t, and $\mathbf{Z} = \mathbf{Z}_\infty$. Each point in $\mathbf{Z}$ is an ordered triple $z = (y, x, t)$ where $x(z)$ is the spatial co-ordinate, $t(z)$ is the temporal co-ordinate and $y(z)$ is the class. Given λ, the number of events in $\mathbf{Z}_t$ is Poisson with mean $t \int_{\mathcal{X}} \lambda_{\bullet}(x)\, \nu(\mathrm{d}x)$, proportional to t and finite. The number of events in $\mathbf{Z}$ is infinite, and the set of points $\{x(z) : z \in \mathbf{Z}\}$ is dense in $\mathcal{X}$.

The Cox process provides a complete description of the random subset $\mathbf{Z} \subset \mathcal{C} \times \mathcal{X} \times (0, \infty)$.

Since $\mathbf{Z}$ is a random set, there can be no concept of a fixed subset or sample in the conventional sense. Nonetheless, the distribution of $\mathbf{Z}$ is well defined, so it is possible to compute the distribution for the observation generated by a well-specified sampling plan. It is convenient for many purposes to take $\mathcal{U}=\{1,2,\ldots\}$ to be the set of natural numbers, and to order the elements of $\mathbf{Z}$ temporally, so that $t_j=t(j)$ is the time of occurrence of the jth event, $x(j)$ is the spatial co-ordinate and $y(j)$ is the type or class. The ordered event times $0\equiv t_0\leqslant t_1\leqslant t_2\leqslant\cdots$ are distinct with probability 1. With this convention, the sequence (x_j, y_j, t_j-t_{j-1}) is infinitely exchangeable. The components are conditionally independent given λ, and identically distributed with joint density

$$E\left[\Lambda_\bullet \exp\{-\Lambda_\bullet(t_j-t_{j-1})\}\,\mathrm{d}t_j \frac{\lambda_\bullet(x_j)\,\nu(\mathrm{d}x_j)}{\Lambda_\bullet}\frac{\lambda_{y_j}(x_j)}{\lambda_\bullet(x_j)}\right]$$

averaged over the intensity function λ. The total intensity $\Lambda_\bullet=\int\lambda_\bullet(x)\,\nu(\mathrm{d}x)$ is the rate of accrual, which is random but constant in time.

3.2. Sampling protocols

Six sampling protocols are considered, some being more natural than others because they can be implemented in finite time. The first is a quota sample with covariate configuration $\mathbf{x}$ as the target. The second is a sequential sample consisting of the first n events and the third is the set $\mathbf{Z}_t$ for fixed t. The fourth is a simple random sample from $\mathbf{Z}_t$ at some suitably large time. The final protocol is a retrospective or case–control sample in which the number of successes and failures is predetermined.

3.2.1. *Quota sample*

Let $\mathbf{x}=(x_1,\ldots,x_n)$ be a given ordered set of n points in $\mathcal{X}$, and let $\mathrm{d}x_j$ be an open interval containing x_j. For convenience of exposition, it is assumed that the points are distinct and the intervals disjoint. A sample from $\mathcal{U}$ is an ordered list of distinct elements $\varphi_1,\ldots,\varphi_n$, and the quota is satisfied if $x(\varphi_j)\in\mathrm{d}x_j$. The easiest way to select such a sample is to partition the population by covariate values $\mathbf{Z}_{\mathrm{d}x}=\{(x,y,t)\in\mathbf{Z}:x\in\mathrm{d}x\}$. Each stratum is infinite and temporally ordered. Define φ_j to be the index of the first event in $\mathbf{Z}_{\mathrm{d}x_j}$.

Distributions are computed for the limit in which each interval $\mathrm{d}x_j$ tends to a point, so the distribution of the spatial component is degenerate at $\mathbf{x}$. The temporal component $t(\varphi_j)$ is conditionally exponential with parameter $\Lambda_\bullet(\mathrm{d}x_j)$, so $t(\varphi_j)\to\infty$. The distribution of the class labels is

$$p_n(\mathbf{y}|\mathbf{x})=E\left\{\prod_{i=1}^{n}\frac{\lambda_{y_i}(x_i)}{\lambda_\bullet(x_i)}\right\}. \tag{6}$$

The conditional distribution is independent of ν and coincides with $p_{\mathbf{x}}(\mathbf{y})$ in model (4) when we set

$$\log\{\lambda_r(x)\}=\alpha_r+\beta_r x+\eta_r(x).$$

In other words, the Cox process is fully compatible with the standard logistic model (3) or (4), and quota sampling is unbiased in the sense that the conditional distribution $p_n(\mathbf{y}|\mathbf{x})$ coincides with $p_{\mathbf{x}}(\mathbf{y})$ in model (4).

3.2.2. *Sequential sample*

Let n be given, and let the sample φ consist of the first n events in temporal order. Given the intensity function, the temporal component of the events is a homogeneous Poisson process with rate $\Lambda_{\bullet} = \int \lambda_{\bullet}(x)\,\nu(\mathrm{d}x)$. The conditional joint density of the sampled time points is thus $\Lambda_{\bullet}^n \exp(-\Lambda_{\bullet} t_n)$ for $0 \leqslant t_1 \leqslant \cdots \leqslant t_n$. The components of $x\varphi$ are conditionally independent and identically distributed with density $\lambda_{\bullet}(x)\,\nu(\mathrm{d}x)/\Lambda_{\bullet}$, and the components of $y\varphi$ are conditionally independent given $x\varphi = \mathbf{x}$ with distribution

$$\mathrm{pr}\{y(\varphi_i) = r|\mathbf{x}\} = \lambda_r(x_i)/\lambda_{\bullet}(x_i).$$

Given λ, the joint density of the sample values at $(\mathbf{x}, \mathbf{y}, \mathbf{t})$ is

$$\begin{aligned} p_n(\mathbf{y}, \mathbf{x}, \mathbf{t}|\lambda)\,\mathrm{d}\mathbf{x}\,\mathrm{d}\mathbf{t} &= \Lambda_{\bullet}^n \exp(-\Lambda_{\bullet} t_n) \prod_{i=1}^n \frac{\lambda_{y_i}(x_i)}{\lambda_{\bullet}(x_i)} \frac{\lambda_{\bullet}(x_i)\,\nu(\mathrm{d}x_i)}{\Lambda_{\bullet}}\,\mathrm{d}t_i \\ &= \exp(-\Lambda_{\bullet} t_n) \prod \lambda_{y_i}(x_i)\,\nu(\mathrm{d}x_i)\,\mathrm{d}t_i, \end{aligned}$$

so the unconditional joint density is

$$p_n(\mathbf{y}, \mathbf{x}, \mathbf{t})\,\mathrm{d}\mathbf{x}\,\mathrm{d}\mathbf{t} = E\left\{ \exp(-\Lambda_{\bullet} t_n) \prod_{i=1}^n \lambda_{y_i}(x_i)\,\nu(\mathrm{d}x_i)\,\mathrm{d}t_i \right\}$$

averaged with respect to the distribution of λ.

The joint density of $(\mathbf{x}, \mathbf{t})$ is computed in the same way:

$$p_n(\mathbf{x}, \mathbf{t})\,\mathrm{d}\mathbf{x}\,\mathrm{d}\mathbf{t} = E\left\{ \exp(-\Lambda_{\bullet} t_n) \prod_{i=1}^n \lambda_{\bullet}(x_i)\,\nu(\mathrm{d}x_i)\,\mathrm{d}t_i \right\}$$

so the conditional distribution of $\mathbf{y}$ given $(\mathbf{x}, \mathbf{t})$ is

$$p_n(\mathbf{y}|\mathbf{x}, \mathbf{t}) = \frac{E\left\{ \exp(-\Lambda_{\bullet} t_n) \prod_{i=1}^n \lambda_{y_i}(x_i) \right\}}{E\left\{ \exp(-\Lambda_{\bullet} t_n) \prod_{i=1}^n \lambda_{\bullet}(x_j) \right\}}. \tag{7}$$

These calculations assume that event times are observed and recorded. Otherwise we need the conditional distribution of $\mathbf{y}$ given $\mathbf{x}$, which is

$$p_n(\mathbf{y}|\mathbf{x}) = \frac{E\left\{ \prod_{i=1}^n \lambda_{y_i}(x_i)/\Lambda_{\bullet} \right\}}{E\left\{ \prod_{i=1}^n \lambda_{\bullet}(x_i)/\Lambda_{\bullet} \right\}}. \tag{8}$$

In either case the conditional distribution is a ratio of expected values, whereas equation (6) is the expected value of a ratio.

If the intensity ratio processes $(\lambda_r(x)/\lambda_0(x))_{x\in\mathcal{X}}$ are jointly independent of the total intensity process $(\lambda_{\bullet}(x))_{x\in\mathcal{X}}$, the conditional distribution $p_n(\mathbf{y}|\mathbf{x})$ coincides with model (4). Otherwise, sequential sampling is biased in the sense that $p(\mathbf{y}|\mathbf{x}) \neq p_{\mathbf{x}}(\mathbf{y})$.

3.2.3. *Sequential sample for fixed time*

In this sampling plan the observation is the set $\mathbf{Z}_t$ for fixed t. Given λ, the number of sampled events $n = \#\mathbf{Z}_t$ is Poisson with parameter $t\Lambda_{\bullet}$. The probability of observing a specific sequence of events and class labels is

$$p_t(\mathbf{y}, \mathbf{x}, \mathbf{t})\,\mathrm{d}\mathbf{x}\,\mathrm{d}\mathbf{t} = E\left\{ \exp(-t\Lambda_{\bullet}) \prod_{i=1}^n \lambda_{y_i}(x_i)\,\nu(\mathrm{d}x_i) \right\} \mathrm{d}\mathbf{t}$$

for $n \geqslant 0$ and $0 \leqslant t_1 \leqslant \cdots \leqslant t_n \leqslant t$. Likewise, the marginal density of $(\mathbf{x}, \mathbf{t})$ is

$$p_t(\mathbf{x}, \mathbf{t})\,\mathrm{d}\mathbf{x}\,\mathrm{d}\mathbf{t} = E\left\{ \exp(-t\Lambda_.) \prod_{i=1}^{n} \lambda_.(x_i)\, \nu(\mathrm{d}x_i) \right\} \mathrm{d}\mathbf{t}$$

so the conditional distribution of $\mathbf{y}$ given $(\mathbf{x}, \mathbf{t})$ for this protocol is

$$p_t(\mathbf{y}|\mathbf{x}, \mathbf{t}) = \frac{E\left\{ \exp(-t\Lambda_.) \prod_{i=1}^{n} \lambda_{y_i}(x_i) \right\}}{E\left\{ \exp(-t\Lambda_.) \prod_{i=1}^{n} \lambda_.(x_i) \right\}}. \tag{9}$$

The conditional distribution depends on the observation period but is independent of the event times $\mathbf{t}$. Consequently $p_t(\mathbf{y}|\mathbf{x})$ coincides with equation (9).

3.2.4. *Simple random sample*

The aim of simple random sampling is to select a subset uniformly at random among subsets of a given size n. In the application at hand, the population is infinite, so simple random sampling is not well defined. However, a similar effect can be achieved by selecting $N \geqslant n$, and restricting attention to the finite subset $\mathbf{Z}_t \subset \mathbf{Z}$ where t is the first time that $\#\mathbf{Z}_t \geqslant N$. By exchangeability, the distribution of $(\mathbf{y}, \mathbf{x})$ on a simple random sample is the same as the distribution on the first n events in temporal order. Apart from the temporal component, the sampling distributions are the same as those in Section 3.2.2. Consequently, unless the intensity ratio processes are independent of $\lambda_.(\cdot)$, simple random sampling is biased.

3.2.5. *Weighted sample*

A weighted sample is one in which individual units are selected (thinned) with probability $w(y, x)$ depending on the response value. Examples with known weight functions arise in monetary unit sampling (Cox and Snell, 1979) and stereological sampling in mining applications (Baddeley and Jensen, 2005). Restriction to a subset of $\mathcal{C}$ is an extreme special case in which w is zero on certain classes and constant on the remainder. More generally, the self-selection of patients that occurs through informed consent in clinical trials may be modelled as an unknown weight function. One way to generate such a sample is to observe the units as they arise in temporal order, retaining units independently with probability $w(y, x)$. This amounts to replacing the intensity function $\lambda(y, x)$ with the weighted version $w(y, x)\,\lambda(y, x)$. Weighted sampling is clearly biased.

3.2.6. *Case–control sample*

Case–control sampling is essentially the same as weighted sampling, except that $k = 2$ and the quota sizes n_0 and n_1 are predetermined. The sample for a case–control study consists of the first n_0 events having $y = 0$ and the first n_1 events having $y = 1$. Event times are not observed. An observation consists of a list $\mathbf{x}$ of n points in $\mathcal{X}$ together with a parallel list $\mathbf{y}$ of labels or class types. The joint probability is

$$p_{n_1 n_2}(\mathbf{y}, \mathbf{x}) = E\left\{ \frac{\prod \lambda_{y_i}(x_i)\, \nu(\mathrm{d}x_i)}{\Lambda_0(\mathcal{X})^{n_0}\, \Lambda_1(\mathcal{X})^{n_1}} \right\},$$

and the conditional probability given $\mathbf{x}$ is proportional to

$$p_{n_1 n_2}(\mathbf{y}|\mathbf{x}) \propto E\left\{ \frac{\prod \lambda_{y_i}(x_i)}{\Lambda_0(\mathcal{X})^{n_0}\, \Lambda_1(\mathcal{X})^{n_1}} \right\}.$$

The approximation derived in Section 4 is equivalent to assuming that, for some measure ν, the random normalized function $\lambda_0(x)/\Lambda_0(\mathcal{X})$ is independent of the integral $\Lambda_0 = \int \lambda_0(x)\,\nu(\mathrm{d}x)$, and likewise for λ_1. Using this approximation, the conditional probability is proportional to

$$p_{n_1 n_2}(\mathbf{y}|\mathbf{x}) \propto E\left\{ \prod_{i=1}^{n} \lambda_{y_i}(x_i) \right\}.$$

In essence, this means that the observation from a case–control design can be analysed as if it were obtained from a prospective sequential sample or simple random sample as in Sections 3.2.2–3.2.4.

3.3. Exchangeable sequences and conditional distributions

The sequence $(y_1, x_1), (y_2, x_2), \ldots$ generated in temporal order by the evolving population model is exchangeable. In contexts such as this, two interpretations of conditional probability and conditional expectation are prevalent in applied work. The probabilistic interpretation is not so much an interpretation as a definition; $\mathrm{pr}(y_u = y|x_u = x) = p_1(y, x)/p_1(x)$ as computed in equation (8) for $n = 1$. Here, u is fixed, x_u is random and we select from the family of conditional distributions the one corresponding to the event $x_u = x$. The stratum interpretation refers to the *marginal* distribution of the random variable $y(u^*)$, where u^* is the first element for which $x_{u^*} = x$. Here, u^* is random, x_{u^*} is fixed and $p_x(y)$ is the marginal distribution of each component in stratum x as defined in equations (6) or (4) for $n = 1$.

In an exchangeable bivariate process, each finite dimensional joint distribution factors $p_n(\mathbf{y}, \mathbf{x}) = p_n(\mathbf{x})\, p_n(\mathbf{y}|\mathbf{x})$. If the conditional distributions satisfy the 'no-interference' condition (5), the stratum distributions coincide with the conditional distributions, the conditional distributions determine a regression model and the bivariate process is called *conventional* or *hierarchical*. Otherwise, if the conditional distributions do not determine a regression model, the stratum distributions are not the same as the conditional distributions. The risk in applied work is that the marginal mean $\mu(x) = \int y\, p_x(\mathrm{d}y)$ in stratum x might be mistaken for the conditional mean $\kappa(x) = \int y\, p(\mathrm{d}y|x)$.

The notation $E(y_u|x_u = x)$ is widely used and dangerously ambiguous. The preferred interpretation has the index u fixed and x_u random, so $E(y_7|x_7 = 3) = \kappa(3)$ is a legitimate expression. In biostatistical work on random-effects models, the stratum interpretation with fixed x and random u is predominant. This interpretation is not unreasonable if properly understood and consistently applied, but it would be less ambiguous if written in the form $\mu(x) = E(y_u|u\colon x_u = x)$. The longer version makes it clear that

$$E\{y_1 - \mu(x_1)|x_1 = 3\} = \kappa(3) - \mu(3) \neq 0,$$

with obvious implications for estimating equations (Section 7.1).

The evolving population model shows clearly that the response distribution for a set of units having a predetermined covariate configuration $\mathbf{x}$ is not necessarily the same as the conditional distribution for a simple random sample that happens to have the same covariate configuration. Thus, the sampling protocol cannot be ignored with impunity. For practical purposes, the plausible protocols are those that can be implemented in finite time, which implies sequential sampling, weighted sampling or case–control sampling.

3.4. Variants and extensions

Up to this point, no assumptions have been made about the distribution of λ. The evolving population model has two principal variants: one in which the k intensity functions $\lambda_0(\cdot), \ldots, \lambda_{k-1}(\cdot)$

are independent, and the conventional one in which the total intensity process $(\lambda_{\bullet}(x))_{x\in\mathcal{X}}$ is independent of the intensity ratios $(\lambda_r(x)/\lambda_0(x))_{x\in\mathcal{X}}$. The two types are not disjoint, but the intersection is small and relatively uninteresting. The characteristic property of the second variant is that the conditional sampling distribution for a sequential or simple random sample coincides with the distribution $p_{\mathbf{x}}(\mathbf{y})$ for predetermined $\mathbf{x}$. Ambiguities concerning the sampling distribution do not arise. Otherwise it is necessary to calculate the conditional distribution that is appropriate for the sampling protocol. Each version of the evolving population model has merit. Both are closed under aggregation of classes because this amounts to replacing k by $k-1$ and adding two of the intensity functions. Deletion or restriction of classes necessarily introduces a strong sampling bias. The total intensity is reduced so only the first variant is closed under this operation. The Gaussian submodel (4) is not closed under aggregation of classes; nor is the log-Gaussian process that is described in Section 5.2.

In the evolving population model, the response on each unit is a point $y(z)$ in the finite set $\mathcal{C}$. It is straightforward to modify this for a continuous response such as the speed of a vehicle passing a fixed point on the highway. Counting measure in $\mathcal{C}$ must be replaced by a suitable finite measure in the real line. To extend the model to a crossover design in which each unit is observed twice, it is necessary to replace $\mathcal{C}$ by $\mathcal{C}^2$, or by $(\mathcal{C}\times\{C,T\})^2$ for randomized treatment (Section 5.4). The random intensity function λ on $(\mathcal{C}\times\{C,T\})^2\times\mathcal{X}$ governs the joint distribution of the x-values and the response–treatment pair at both time points. In a longitudinal design, observations on the same unit over time are understood to be correlated, and there may also be correlations between distinct units. To extend the model in this way, it is necessary to replace $\mathcal{C}$ by a higher order product space, and to construct a suitable random intensity on this space. Such an extension is well beyond the scope of this paper.

4. Limit distributions

The conditional probability distribution

$$p_t(\mathbf{y}|\mathbf{x}) = \frac{E\left\{\exp(-\Lambda_{\bullet}t)\prod_{i=1}^{n}\lambda_{y_i}(x_i)\right\}}{E\left\{\exp(-\Lambda_{\bullet}t)\prod_{i=1}^{n}\lambda_{\bullet}(x_i)\right\}} = \frac{p_t(\mathbf{y},\mathbf{x})}{p_t(\mathbf{x})} \tag{10}$$

in equations (7) and (9) is the ratio of the joint density and the marginal density. Ideally, the conditional distribution should be independent of the baseline measure, but this is not so because ν enters the definition of $\Lambda_{\bullet}=\int\lambda_{\bullet}(x)\,\nu(\mathrm{d}x)$. However, this dependence is not very strong, so it is reasonable to proceed by selecting a baseline measure that is both plausible and convenient, rather than attempting to estimate ν. Plausible means that ν should be positive on open sets.

The numerator and denominator both have non-degenerate limits as either $t\rightarrow 0$ for fixed ν, or the scalar $\nu(\mathcal{X})\rightarrow 0$ for fixed t. The limiting low intensity conditional distribution

$$q_n(\mathbf{y}|\mathbf{x}) = \frac{E\left\{\prod_{i=1}^{n}\lambda_{y_i}(x_i)\right\}}{E\left\{\prod_{i=1}^{n}\lambda_{\bullet}(x_i)\right\}} \tag{11}$$

is convenient for practical work because it is independent of ν. In addition, the product densities in the numerator and denominator are fairly easy to compute for a range of processes such as log-Gaussian processes (Møller *et al.*, 1998) and certain gamma processes (Shirai and Takahashi, 2003; McCullagh and Møller, 2006). One can argue about the plausibility or relevance of the limit, but the fact that the limit distribution is independent of ν is a definite plus.

The same limit distribution is obtained by a different sort of argument as follows. Suppose that there is a measure ν such that the nk ratios $\lambda_0(x)/\Lambda_., \ldots, \lambda_{k-1}(x)/\Lambda_.$ for $x \in \mathbf{x}$ are jointly independent of $\Lambda_.$. Then the numerator in equation (10) can be expressed as a product of expectations

$$E\{\exp(-t\Lambda_.) \prod \lambda_{y_i}(x_i)\} = E\{\Lambda_.^n \exp(-t\Lambda_.)\}\, E\left\{\frac{\lambda_{y_1}(x_1)}{\Lambda_.} \cdots \frac{\lambda_{y_n}(x_n)}{\Lambda_.}\right\}$$
$$= \frac{E\{\Lambda_.^n \exp(-t\Lambda_.)\}}{E(\Lambda_.^n)} E\{\prod \lambda_{y_i}(x_i)\}.$$

The denominator can be factored in a similar way, so the ratio in equation (10) simplifies to equation (11). If this condition is satisfied by ν, it is satisfied by all positive scalar multiples of ν, and the conditional distribution is unaffected.

The condition here is one of existence of a measure ν satisfying the independence condition for the particular finite configuration $\mathbf{x}$. In other words, the measure may depend on $\mathbf{x}$, so the condition of existence is not especially demanding. Examples are given in McCullagh and Møller (2006) of intensity functions such that the ratios are independent of the integral with respect to Lebesgue measure on a bounded subset of $\mathcal{R}$ or $\mathcal{R}^d$. It is also possible to justify the independence condition by a heuristic argument as follows. Suppose that $\lambda(x) = T\,\hat{\lambda}(x)$ where $\hat{\lambda}$ is ergodic on $\mathcal{R}$ with unit mean, and $T > 0$ is distributed independently of $\hat{\lambda}$. Then, if $\nu(\mathrm{d}x) = \mathrm{d}x/2L$ for $-L \leqslant x \leqslant L$, we find that $\Lambda_. = T\hat{\Lambda}_. = T + o(1)$ for large L, and the ratios $\lambda(x)/\Lambda_. = \hat{\lambda}(x)/\hat{\Lambda}_.$ are jointly independent of T by assumption. The independence assumption is then satisfied in the limit as $L \to \infty$, and ν is, in effect, Lebesgue measure. For these reasons, the limit distribution (11) is used for certain calculations in the following section.

5. Two parametric models

5.1. Product densities and conditional distributions

All the models in this section are such that $\lambda_0, \ldots, \lambda_{k-1}$ are independent intensity functions. The conditional distribution (11) is a distribution on partitions of $\mathbf{x}$ into k labelled classes, some of which might be empty. Denote by $\mathbf{x}^{(r)}$ the subset of $\mathbf{x}$ for which $y = r$. The numerator in equation (11) is the product

$$\prod_{r=0}^{k-1} E\left\{\prod_{x \in \mathbf{x}^{(r)}} \lambda_0(x)\right\} = m_0(\mathbf{x}^{(0)}) \cdots m_{k-1}(\mathbf{x}^{(k-1)})$$

where m_r is the product density for λ_r, and $m_r(\emptyset) = 1$. The denominator is the product density at $\mathbf{x}$ for the superposition process with intensity $\lambda_.$. In other words, equation (11) is

$$q_n(\mathbf{y}|\mathbf{x}) = \frac{m_0(\mathbf{x}^{(0)}) \cdots m_{k-1}(\mathbf{x}^{(k-1)})}{m_.(\mathbf{x})}, \tag{12}$$

for partitions of $\mathbf{x}$ into k labelled classes. For prediction, the Papangelou conditional intensity is used. Suppose that $(\mathbf{y}, \mathbf{x})$ has been observed in a sequential sample up to time t, and that the next subsequent event occurs at x'. The prognostic distribution for the response y' is the conditional distribution given $\mathbf{y}$, $\mathbf{x}$ and the value x', which is

$$q_{n+1}\{y' = r | (\mathbf{y}, \mathbf{x}, x')\} \propto m_r(\mathbf{x}^{(r)} \cup \{x'\})/m_r(\mathbf{x}^{(r)}). \tag{13}$$

In general, the one-dimensional prognostic distribution is considerably easier to compute than the joint distribution.

The first task is to find product densities for specific parametric models.

5.2. Log-Gaussian model

Let $\log(\lambda)$ be a Gaussian process in $\mathcal{X}$ with mean μ and covariance function K. In other words

$$E[\log\{\lambda(x)\}]=\mu(x),$$

$$\operatorname{cov}[\log\{\lambda(x)\},\ \log\{\lambda(x')\}]=K(x,x').$$

Then the expected value of the product $m(\mathbf{x})=E\{\lambda(x_1)\cdots\lambda(x_n)\}$ is

$$\log\{m(\mathbf{x})\}=\sum_{x\in\mathbf{x}}\mu(x)+\tfrac{1}{2}\sum_{x,x'\in\mathbf{x}}K(x,x').$$

This expression enables us to simplify the numerator in equation (11). Unfortunately, the sum of log-Gaussian processes is not log-Gaussian, so the normalizing constant is not available in closed form. The logarithm of the conditional distribution (11) is given in an obvious notation by

$$\log\{q_n(\mathbf{y}|\mathbf{x})\}=\text{constant}+\sum\mu_{y_i}(x_i)+\tfrac{1}{2}\sum_{r=1}^{k}\sum_{x,x'\in\mathbf{x}^{(r)}}K_r(x,x').$$

The prognostic distribution for a subsequent event at x' is obtained from the product density ratios

$$q_{n+1}\{Y'=r|(\mathbf{y},\mathbf{x},x')\}\propto\exp\{\mu_r(x')+\tfrac{1}{2}K_r(x',x')+\sum_{x\in\mathbf{x}^{(r)}}K_r(x,x')\}.$$

Without loss of generality, we may set $\mu_0(x)=0$. If the covariance functions are equal, the prognostic log-odds are

$$\text{log-odds}(Y'=1|\ldots)=\mu_1(x')+\sum_{x\in\mathbf{x}^{(1)}}K(x,x')-\sum_{x\in\mathbf{x}^{(0)}}K(x,x')$$

for $k=2$. The conditional log-odds is a kernel function, an additive function of the sample values formally the same as Markov random-field models (Besag, 1974) except that the x-configuration is not predetermined or regular.

The log-Gaussian model with independent intensity functions is closed under restriction of classes. However, the sum of two independent log-Gaussian variables is not log-Gaussian, so the model is not closed under aggregation of homogeneous classes. This remark applies also to the model in Section 2. Failure of this property for homogeneous classes is a severe limitation for practical work.

5.3. Gamma models

Let $Z_1,\ldots,Z_d$ be independent zero-mean real Gaussian processes in $\mathcal{X}$ with covariance function $K/2$, and let $\lambda(x)=Z_1^2(x)+\cdots+Z_d^2(x)$. Denote by $K[\mathbf{x}]$ the matrix with entries $K(x_i,x_j)$. For any such matrix, the α-permanent is a weighted sum over permutations

$$\operatorname{per}_\alpha(K[\mathbf{x}])=\sum_\sigma\alpha^{\#\sigma}\prod_{i=1}^{n}K(x_i,x_{\sigma_i})$$

where $\#\sigma$ is the number of cycles. The product density of λ at $\mathbf{x}$ is

$$E\{\lambda(x_1)\cdots\lambda(x_n)\}=\operatorname{per}_{d/2}(K[\mathbf{x}]).$$

Under a certain positivity condition the process can be extended from the integers to all positive values of d. For a derivation of these results, see Shirai and Takahashi (2003) or McCullagh and Møller (2006).

In the homogeneous version of the gamma model, λ_r has product density $\mathrm{per}_{\alpha_r}(K[\mathbf{x}^{(r)}])$, and $\lambda_{\bullet}$ has product density $\mathrm{per}_{\alpha_{\bullet}}(K[\mathbf{x}])$. The conditional distribution (11) is

$$q_n(\mathbf{y}|\mathbf{x}) = \frac{\mathrm{per}_{\alpha_0}(K[\mathbf{x}^{(0)}]) \cdots \mathrm{per}_{\alpha_{k-1}}(K[\mathbf{x}^{(k-1)}])}{\mathrm{per}_{\alpha_{\bullet}}(K[\mathbf{x}])},$$

which is a generalization of the multinomial and Dirichlet–multinomial distributions. The prognostic distribution for a subsequent event at x' is

$$q_{n+1}(y' = r|\ldots) \propto \mathrm{per}_{\alpha_r}(K[\mathbf{x}^{(r)}, x'])/\mathrm{per}_{\alpha_r}(K[\mathbf{x}^{(r)}]).$$

By contrast with the log-Gaussian model, the prognostic log-odds is not a kernel function. For an application of this model to classification, see McCullagh and Yang (2006).

The homogeneous gamma model can be extended in various ways, e.g. by replacing $\lambda_r(x)$ with $\exp(\alpha_r + \beta_r x)\,\lambda_r(x)$. Then the product density at $\mathbf{x}^{(r)}$ becomes $\exp(n_r\alpha_r + \beta_r \mathbf{x}_{\bullet}^{(r)})\,\mathrm{per}_{\alpha_r}(K[\mathbf{x}^{(r)}])$, where $n_r = \#\mathbf{x}^{(r)}$ and $\mathbf{x}_{\bullet}^{(r)}$ is the sum of the components. Alternatively, if λ_r is replaced by $\tau_r\,\lambda_r(x)$, where $\tau_0, \ldots, \tau_{k-1}$ are independent scalars independent of λ, the product density is replaced by $h_r(n_r)\,\mathrm{per}_{\alpha_r}(K[\mathbf{x}^{(r)}])$ where $h_r(n)$ is the nth moment of τ_r. Finally, there is a non-trivial limit distribution as $k \to \infty$ and $\alpha_r = \alpha/k$ with α fixed (McCullagh and Yang, 2006).

5.4. Treatment effects and randomization

To incorporate a non-random treatment effect, we replace $\mathcal{C}$ by $\mathcal{C} \times \{C, T\}$, where $\{C, T\}$ are the two treatment levels. Consider the binary model with multiplicative intensity function

$$\lambda(y, v, x) = \lambda(y, x)\,\gamma(y, v, x) \tag{14}$$

in which γ is a fixed parameter, and v is treatment status. Given λ, the treatment effect as measured by the conditional odds ratio is

$$\tau(x) = \frac{\gamma(1, T, x)\,\gamma(0, C, x)}{\gamma(0, T, x)\,\gamma(1, C, x)},$$

which is a non-random function of x, possibly a constant. Given an event $z \in \mathbf{Z}$ with $x(z) = x$, the four possibilities for response and treatment status have probabilities proportional to

$$\mathrm{pr}\{y(z) = y, v(z) = v | z \in \mathbf{Z}\} \propto E\{\lambda(y, v, x)\} = m_y(x)\,\gamma(y, v, x).$$

Consequently the treatment effect as measured by the unconditional odds ratio is also $\tau(x)$ with no attenuation. The stratum distribution $p_x(\cdot)$ as defined in model (3) for fixed x gives a different definition of treatment effect, one that is seldom relevant in applications.

For simplicity we now assume that the treatment effect is constant in x. The conditional distribution (11) for a sequential sample reduces to

$$q_n(\mathbf{y}, \mathbf{v}|\mathbf{x}) \propto m_0(\mathbf{x}^{(0)})\,m_1(\mathbf{x}^{(1)}) \prod_{i=1}^{n} \gamma(y_i, v_i),$$

which is a bipartition model for response and treatment status. If m_0 and m_1 are known functions, this is the exponential family generated from (12) with canonical parameter $\log\{\gamma(r, s)\}$

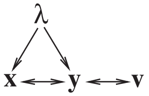
and canonical statistic the array of counts in which n_{rs} is the observed number of events for which $y=r$ and $v=s$.

Suppose that $(\mathbf{y}, \mathbf{v}, \mathbf{x})$ has been observed in a sequential sample. The prognosis for the response y' for a subsequent event at x' depends on whether $v'=T$ or $v'=C$ as follows:

$$\text{odds}(y'=1|\ldots, v')=\frac{m_1(\mathbf{x}^{(1)}, x')\,m_0(\mathbf{x}^{(0)})}{m_1(\mathbf{x}^{(1)})\,m_0(\mathbf{x}^{(0)}, x')}\,\frac{\gamma(1, v')}{\gamma(0, v')}.$$

However, the prognostic odds ratio is equal to the treatment effect, again without attenuation.

The preceding formulation of treatment effects is an attempt to incorporate into the sampling model the notion that treatment status is the outcome of a random process. The model with constant treatment effect can be written in multiplicative form $\lambda(y, x)\gamma(y, v)$ as an intensity on $\mathcal{C}\times\{C, T\}\times\mathcal{X}$. The graphical representation with one node for each random element



shows that treatment status is conditionally independent of $(\lambda, \mathbf{x})$ given $\mathbf{y}$. Although the concept of treatment *assignment* is missing, the multiplicative intensity model describes accurately what is achieved by randomization. The point process of events is such that treatment status $v(z)$ is conditionally independent of $x(z)$ given the response $y(z)$.

6. Interference

6.1. Definition

Let $p_{\mathbf{x}}(\cdot)$ or $p(\cdot|\mathbf{x})$ be a set of distributions defined for arbitrary finite configurations $\mathbf{x}$. Lack of interference is a mathematical property ensuring that the n-dimensional distribution $p_{\mathbf{x}}(\cdot)$ is the marginal distribution of the $(n+1)$-dimensional distribution $p_{\mathbf{x},x'}(\cdot)$ after integrating out the last component. In symbols $p_{\mathbf{x}}(A)=p_{\mathbf{x},x'}(A\times\mathcal{C})$ for $A\subset\mathcal{C}^n$, or

$$p(\mathbf{y}\in A|\mathbf{x})=p\{(\mathbf{y}, y')\in A\times\mathcal{C}|(\mathbf{x}, x')\}$$

if applied to conditional distributions. When this condition is satisfied, the distribution of the first n components is unaffected by the covariate value for subsequent components. This is also the Kolmogorov consistency condition for a $\mathcal{C}$-valued process in which the joint distribution of $Y(u_1), \ldots, Y(u_n)$ depends on the covariate values $x(u_1), \ldots, x(u_n)$ on those units. It is satisfied by regression models such as (1), (3) and (4).

In the statistical literature on design, interference is usually understood in the physical or biological sense, meaning carry-over effects from treatment applied to neighbouring plots. For details and examples, see Cox (1958) or Besag and Kempton (1986). The definition does not distinguish between physical interference and sampling interference, but this paper emphasizes the latter.

To understand how sampling interference might arise, consider the simplest evolving population model in which $\mathcal{X}=\{x\}$ is a set containing a single point denoted by x. Let $Y_1, \ldots$ be the class labels in temporal order. Given λ, the number m of events in unit time is Poisson distributed with parameter $\lambda_{\bullet}$, so m could be zero. However, given m, the values $Y_1, \ldots, Y_m$ are exchangeable with one- and two-dimensional distributions

$$p_1(Y_1=r|m\geqslant 1)=\frac{E[\{1-\exp(-\lambda_.)\}\lambda_r/\lambda_.]}{E\{1-\exp(-\lambda_.)\}}\simeq\frac{E(\lambda_r)}{E(\lambda_.)},$$

$$p_2(Y_1=r,Y_2=s|m\geqslant 2)=\frac{E[\{1-(1+\lambda_.)\exp(-\lambda_.)\}\lambda_r\lambda_s/\lambda_.^2]}{E\{1-(1+\lambda_.)\exp(-\lambda_.)\}}\simeq\frac{E(\lambda_r\lambda_s)}{E(\lambda_.^2)},$$

$$p_2(Y_1=r|m\geqslant 2)=\frac{E[\{1-(1+\lambda_.)\exp(-\lambda_.)\}\lambda_r/\lambda_.]}{E\{1-(1+\lambda_.)\exp(-\lambda_.)\}}\simeq\frac{E(\lambda_r\lambda_.)}{E(\lambda_.^2)}$$
$$\neq p_1(Y_1=r|m\geqslant 1).$$

In general, the probability assigned to the event $Y_1=r$ by the bivariate distribution p_2 is not the same as the probability assigned to the same event by the one-dimensional distribution p_1. The condition $m\geqslant 2$ in p_2 implies an additional event at x, which may change the probability distribution of Y_1.

Two specific models are now considered: one exhibits interference; the other not. In the log-Gaussian model

$$\log(\lambda_0)\sim N(\mu_0,1),$$
$$\log(\lambda_1)\sim N(\mu_1,1)$$

are independent random intensities. For a low intensity value $(\mu_0,\mu_1)=(-5,-4)$, we find by numerical integration that $p_1(Y_1=0)=0.272$ whereas $p_2(Y_1=0)=0.201$, so the interference effect is substantial. The limiting low intensity approximations are 0.269 and 0.193 respectively. For $(\mu_0,\mu_1)=(-1,0)$ the mean intensities are $(\exp(-\frac{1}{2}),\exp(\frac{1}{2}))$, and the difference is less marked: $p_1(Y_1=0)=0.310$ *versus* $p_2(Y_1=0)=0.291$.

By contrast, consider the gamma model in which

$$\lambda_0\sim G(\alpha_0\theta,\alpha_0),$$
$$\lambda_1\sim G(\alpha_1\theta,\alpha_1)$$

are independent with mean $E(\lambda_r)=\alpha_r\theta$ and variance $\alpha_r\theta^2$. The total intensity is distributed as $\lambda_.\sim G(\alpha_.\theta,\alpha_.)$ independently of the ratio, which has the beta distribution $\lambda_0/\lambda_.\sim B(\alpha_0,\alpha_1)$. On account of independence, we find that $p_1(Y_1=0)=p_2(Y_1=0)=\alpha_0/\alpha_.$, so interference is absent.

6.2. Overdispersion

Suppose that events are grouped by covariate value, so that $T_.(x)$ is the observed number of events at x, and $T_r(x)$ is the number of those who belong to class r. Overdispersion means that the variance of $T_r(x)$ exceeds the binomial variance, and the covariance matrix of $T(x)$ exceeds the multinomial covariance. However, units having distinct x-values remain independent. This effect is achieved by a Cox process driven by a completely independent random intensity taking independent values at distinct points in $\mathcal{X}$. Since units having distinct x-values remain independent, it suffices to describe the one-dimensional marginal distributions of the class totals at each point in $\mathcal{X}$.

Under the gamma model, the class totals at x are independent negative binomial random variables with means $E(T_r)=\gamma\alpha_r$ and variances $\gamma\alpha_r(1+\gamma)$. Given the total number of events at x, the conditional distribution is Dirichlet–multinomial

$$p(T=t|T_.=m)=\frac{m!\,\Gamma(\alpha_.)}{\Gamma(m+\alpha_.)}\prod_{r\in\mathcal{C}}\frac{\Gamma(t_r+\alpha_r)}{t_r!\,\Gamma(\alpha_r)}$$

for non-negative t_r such that $t_{\bullet}=m$. The sequence of values $Y_1,\ldots,Y_m$ is exchangeable, and, since there is no interference, the marginal distributions are independent of m:

$$\mathrm{pr}(Y_1=r|m)=\alpha_r/\alpha_{\bullet}=\pi_r,$$

$$\mathrm{pr}(Y_1=r,Y_2=s|m)=\pi_r\pi_s+\begin{cases}\pi_r(1-\pi_r)/(\alpha_{\bullet}+1) & r=s,\\ -\pi_r\pi_s/(\alpha_{\bullet}+1) & \text{otherwise.}\end{cases}$$

Because of interference, no similar results exist for the log-Gaussian model.

7. Computation

7.1. Parameter estimation

Since **x** and **y** are both generated by a random process, the likelihood function is determined by the joint density. However, the joint distribution depends on the infinite dimensional nuisance parameter $\nu(\mathrm{d}x)$, which governs primarily the marginal distribution of **x**. It appears that the marginal distribution of **x** must contain very little information about intensity ratios, so it is natural to use the conditional distribution given **x** for inferential purposes. Likelihood calculations using the exact conditional distribution (7)–(9) or the limit distributions (11) and (12) are complicated, though perhaps not impossible. We focus instead on parameter estimation using unbiased estimating equations.

Let $m_r(x)=E\{\lambda_r(x)\}$ be the mean intensity function for class r, $m_{\bullet}(x)$ the expected total intensity at x, $\rho_r(x)=E\{\lambda_r(x)/\lambda_{\bullet}(x)\}$ the expected value of the intensity ratio at x and $\pi_r(x)=m_r(x)/m_{\bullet}(x)$ the ratio of expected intensities. Sampling bias is the key to the distinction between $\rho(x)$, the marginal distribution for fixed x, and $\pi(x)$, the conditional distribution for random x generated by a sequential sample from the process.

Consider a sequential sampling scheme in which the observation consists of the events $\mathbf{Z}_t$ for fixed t. The number of events $\#\mathbf{Z}_t$, the values $y(z)$, $x(z)$ and $\pi(x)=\pi\{x(z)\}$ for $z\in\mathbf{Z}_t$ are all random. It is best to regard $\mathbf{Z}_t$ as a random measure in $\mathcal{C}\times\mathcal{X}$ whose mean has density $tm_y(x)\,\nu(\mathrm{d}x)$ at (y,x). The expected number of events in the interval $\mathrm{d}x$ is $tm_{\bullet}(x)\,\nu(\mathrm{d}x)$, and the expected number of events of class r in the same interval is $tm_r(x)\,\nu(\mathrm{d}x)$. For a function $h\colon\mathcal{X}\to\mathcal{R}$, additive functionals have expectation

$$E\Big[\sum_{\mathbf{Z}_t}h\{x(z)\}\Big]=t\int_{\mathcal{X}}h(x)\,m_{\bullet}(x)\,\nu(\mathrm{d}x),$$

$$E\Big[\sum_{\mathbf{Z}_t}h\{x(z)\}y_r(z)\Big]=t\int_{\mathcal{X}}h(x)\,m_r(x)\,\nu(\mathrm{d}x),$$

where $y(z)$ is the indicator function for the class, i.e. $y_r(z)=1$ if the class is r. It follows that the sum $T_r=\Sigma_{\mathbf{Z}}\,h(x)\{y_r(z)-\pi_r(x)\}$ has exactly zero mean for each function h. This first-moment calculation involves only the first-order product densities. If it were necessary to calculate $E(T|\mathbf{x})$ given the configuration **x**, we should begin with the joint distribution or the conditional distribution (9). Because of interference, $E(T|\mathbf{x})$ is not zero; nor is $E(T|\#\mathbf{Z}_t)$. Consequently, the moment calculations in this section are fundamentally different from those of McCullagh (1983) or Zeger and Liang (1986).

The covariance of T_r and T_s is a sum of three terms: one associated with intrinsic Bernoulli variability, one with spatial correlation and one with interference. The expressions are simplified here by setting $h(x)=1$.

$$\begin{aligned}\operatorname{cov}(T_r, T_s) = t &\int_{\mathcal{X}} \{\pi_r(x)\delta_{rs} - \pi_r(x)\,\pi_s(x)\} m_{\bullet}(x)\,\nu(\mathrm{d}x)\\ &+ t^2 \int_{\mathcal{X}^2} \{\pi_{rs}(x,x') - \pi_{r\bullet}(x,x')\,\pi_{s\bullet}(x',x)\}\, m_{\bullet\bullet}(x,x')\,\nu(\mathrm{d}x)\,\nu(\mathrm{d}x')\\ &+ t^2 \int_{\mathcal{X}^2} \Delta_{r\bullet}(x,x')\,\Delta_{s\bullet}(x',x)\, m_{\bullet\bullet}(x,x')\,\nu(\mathrm{d}x)\,\nu(\mathrm{d}x'). \qquad (15)\end{aligned}$$

In these expressions, $m_{rs}(x,x') = E\{\lambda_r(x)\,\lambda_s(x')\}$ is the second-order product density, and $\pi_{rs}(x,x') = m_{rs}(x,x')/m_{\bullet\bullet}(x,x')$ is the bivariate distribution for ordered pairs of distinct events. Roughly speaking, $\pi_{r\bullet}(x,x')$ is the probability that the event at x is of class r given that another event occurs at x'. The difference $\Delta_{r\bullet}(x,x') = \pi_{r\bullet}(x,x') - \pi_r(x)$, which is a measure of second-order interference, is zero for conventional models. Both the gamma and the log-normal models exhibit interference, but the homogeneous gamma model has the special property of zero second-order interference.

The first integral in equation (15) can be consistently estimated by summation of $\pi_r(x)\delta_{rs} - \pi_r(x)\,\pi_s(x)$ over $\mathbf{x}$. The second and third integrals can be estimated in the same way by summation over distinct ordered pairs.

The marginal mean for an event in stratum x is $E\{y_r(x)\} = \rho_r(x)$ as determined by the logistic–normal integral, and the difference $y_r(x) - \rho_r(x)$ is the basis for estimating equations associated with hierarchical regression models (Zeger and Liang, 1986; Zeger *et al.*, 1988). If in fact the x-values are generated by the process itself, the estimating function $\Sigma_{\mathbf{z}}\, h(x)[y_r(z) - \rho_r\{x(z)\}]$ has expectation $\int_{\mathcal{X}} h(x)\{\pi(x) - \rho(x)\}\, m_{\bullet}(x)\,\nu(\mathrm{d}x)$, which is not zero and is of the same order as the sample size. Conventional estimating equations are biased for the marginal mean and give inconsistent parameter estimates. Similar remarks apply to likelihood-based estimates. The correct likelihood function (9) takes account of the sampling plan and gives consistent estimates; the incorrect likelihood (3) gives inconsistent estimates.

For the binary case $k = 2$, we write $\pi(x) = \pi_1(x)$ and revert to the usual notation with $y = 0$ or $y = 1$. If we use a linear logistic parameterization $\operatorname{logit}\{\pi(x)\} = \beta' x$, the parameters can be estimated consistently by using a generalized estimating equation of the form $X'\hat{W}(Y - \hat{\pi}) = 0$ with a suitable choice of weight matrix W depending on x. Recognizing that the target is $\pi(x)$ rather than $\rho(x)$, the general outline that was described by Liang and Zeger (1986) can be followed, but variance calculations need to be modified to account for interference as in equation (15).

The functional $y_r y_s' - \pi_{rs}(x,x')$ of degree two for ordered pairs of distinct events also has zero expectation for each r and s. The additive combination $\Sigma\, h(x,x')\{y_r y_s' - \pi_{rs}(x,x')\}$ can be used as a supplementary estimating function for variance and covariance components. However, variance calculations are much more complicated.

7.2. Classification and prognosis

By contrast with likelihood calculations, the prognostic distribution for a subsequent event at x' is relatively easy to compute. For the log-Gaussian model in Section 5.2, the prognostic distribution is a kernel function

$$\log[\operatorname{pr}\{Y(x') = r|\ldots\}] = \log\{m_r(x')\} + \sum_{x \in \mathbf{x}^{(r)}} K(x,x') + \text{constant}$$

for $r \in \mathcal{C}$. If necessary, unknown parameters can be estimated by cross-validation (Wahba, 1985). The theory in Section 3 requires K to be a proper covariance function defined pointwise, but the prognostic distribution is well defined for generalized covariance functions such as $-\gamma|x - x'|^2$ for $\gamma \geqslant 0$, provided that the functions $\log\{m_r(x)\}$ span the kernel of the process.

The gamma model presents more of a computational challenge because the prognostic distribution is a ratio of permanents,

$$\mathrm{pr}\{Y(x')=r|\ldots\} \propto \mathrm{per}_{\alpha_r}(K[\mathbf{x}^{(r)}, x'])/\mathrm{per}_{\alpha_r}(K[\mathbf{x}^{(r)}]),$$

which are notoriously difficult to compute. As it happens, the permanent ratio is easier to approximate than the permanent itself. For two configurations of 50 events, Fig. 1 shows the prognostic probability $\mathrm{pr}\{Y(x')=1\}$ computed for the homogeneous gamma model with $k=2$ and $\mathcal{X}=(0,4)$. Permanent ratios were approximated analytically by using a cycle expansion

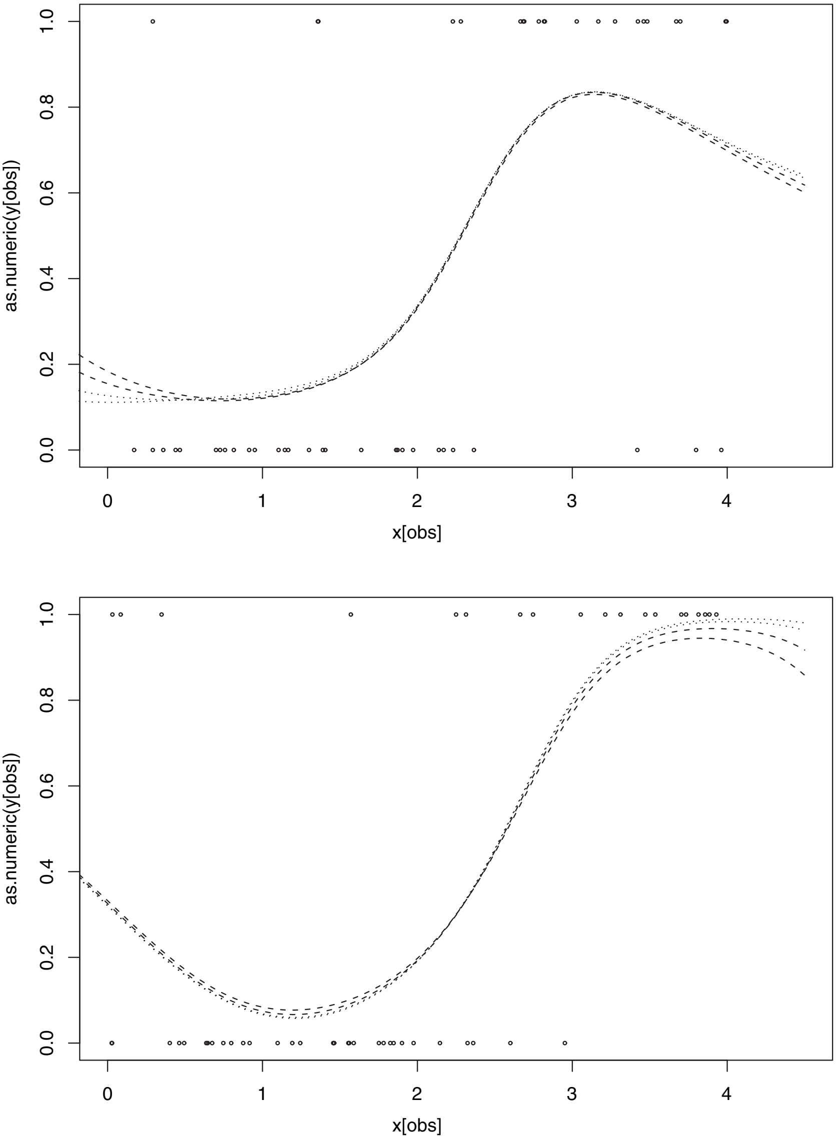


Fig. 1. Prognostic probability computed for the homogeneous gamma model with $K(x,x')=\exp\{-(x-x')^2\}$ and four values of α from 0.05 to 0.5: the two sample configurations of 50 events are indicated by dots at $y=0$ and $y=1$

truncated after cycles of length 4. In the homogeneous gamma model, the one-dimensional conditional probability $q_1(y=1|x)$ for a single event is $\frac{1}{2}$ for every x, so the prognostic probability graphs in Fig. 1 should not be confused with regression curves.

8. Summarizing remarks

8.1. Conventional random-effects models

The statistical model (3) has an observation space $\{0,1\}^n$ for each sample of size n, and a parameter space with four components $(\alpha,\beta,\sigma,\tau)$. Everything else is incidental. The random process η, used as an intermediate construct in the derivation of the distribution, is not a component of the observation space; nor is it a component of the parameter space. In principle, model (3) could have been derived directly without the intermediate step (2), so direct inference for η is impossible in model (3). On account of the consistency condition (5), we can compute the conditional probability of any event such as

$$E\{Y(u')|\mathbf{y}\}=E\left[\frac{\exp\{\alpha+\beta x'+\eta(x')\}}{1+\exp\{\alpha+\beta x'+\eta(x')\}}\middle|\mathbf{y}\right]=\frac{p_{\mathbf{x},x'}(\mathbf{y},1)}{p_{\mathbf{x}}(\mathbf{y})}, \tag{16}$$

using the model distribution with an additional unit having $x(u')=x'$. Sample space inferences of this sort are accessible directly from model (3), but inference for η is not. Similar remarks apply to the Gaussian model (1), in which case the conditional expected value (16) is a generalized smoothing spline in x'.

Since the likelihood function does not determine the observation space, we look to the likelihood function only for parameter estimation, not for inferences about the sample space or subsequent values of the process. This interpretation of model and likelihood is neutral in the Bayes–non-Bayes spectrum. It is consistent with expression (2) as a partially Bayesian model with parameters (α,β,η), in which the Gaussian process serves as the prior distribution for η. The Bayesian formulation enables us to compute a posterior distribution for $\eta(x')$ whether it is of interest or not. Despite the formal equivalence of expression (2) as a partially Bayesian model, and model (3) as a non-Bayesian model, the two formulations are different in a fundamental way. The treatment effect is ordinarily defined as the ratio of success odds for a treated individual to that of an untreated individual having the same baseline covariate values. Because of parameter attenuation, the value that is obtained for the partially Bayesian model (2) is not the same as that for the marginal model (3). Both calculations are unit specific, so this distinction is not a difference between subject-specific and population-averaged effects. This paper argues that neither definition is appropriate because neither model accounts properly for sampling biases.

8.2. Subject-specific and population-average effects

For all models considered in this paper, including (1)–(4), the probabilities are unit specific. That is to say, each regression model specifies the response distribution for every unit, and the joint distribution for each finite subset of units. Treatment effect is measured by the odds ratio, which may vary from unit to unit depending on the covariate value. The population-average effect, if it is to be used at all, must be computed after the fact by averaging the treatment effects over the distribution of x-values for the units in the target population. Note the distinction between unit u and subject $s(u)$: two distinct units u and u' in a crossover or longitudinal design correspond to the same patient or subject if $s(u)=s(u')$. This block structure is assumed to be encoded in x.

Numerous researchers have noted a close parallel between expression (2) and the one-dimensional marginal distributions that are associated with model (3). Specifically, if η_u is a zero-mean Gaussian variable,

$$\text{logit}[\text{pr}\{Y(u)=1|\eta;x\}]=\alpha+\beta x(u)+\eta_u \tag{17}$$

implies

$$\text{logit}[\text{pr}\{Y(u)=1;x\}]\simeq\alpha^*+\beta^* x(u) \tag{18}$$

by averaging over η_u for fixed u. Zeger *et al.* (1988) gave an accurate approximation for the attenuation ratio $\tau=\beta^*/\beta\leqslant 1$, which depends on the variance of η_u. Neuhaus *et al.* (1991) confirmed the accuracy of this approximation. They also gave a convincing demonstration of the magnitude of the attenuation effect by analysing a study of breast disease in two different ways. Maximum likelihood estimates $\hat{\alpha}$ and $\hat{\beta}$ were obtained by maximizing an approximation to the integral (3) by using the software package egret. The alternative, generalized estimating equations using expression (18) for expected values supplemented by an approximate covariance matrix, gives estimates $\hat{\alpha}^*$ and $\hat{\beta}^*$ of the attenuated parameters. The attenuation ratios β^*/β were found to be approximately 0.35, which is in good agreement with the Taylor approximation.

In biostatistical terminology, the regression parameters α and β in equation (17) are called subject specific or cluster specific, whereas the parameters in (18) are called population-averaged effects (Zeger *et al.*, 1988). The terms 'marginal parameterization' (Glonek and McCullagh, 1995), 'marginal model' (Heagerty, 1999) and even 'marginalized model' (Schildcrout and Heagerty, 2007) are also used in connection with (18). Certainly, it is important to distinguish one from the other because the parameter values are very different. Nonetheless, the population-average terminology is misleading because both (17) and (18) refer to a specific unit labelled u, and hence to a specific subject $s(u)$, not to a randomly selected unit or subject. The bivariate and multivariate version (3) is also specific to the particular set of units having covariate configuration $\mathbf{x}$. In other words, both of these are conventional regression models in which the concept of random sampling of units is absent.

Apart from minor differences introduced by approximating the one-dimensional integral by (18), and similar approximations for bivariate and higher order distributions, these are in fact the same model. They have different parameterizations, and they use different methods to estimate the parameters, but the distributions are the same. The distinction between the population-average approach and the cluster-specific approach is not a distinction between models, but a distinction between two parameterizations of essentially the same model, and two methods for parameter estimation.

Having established the point that there is only one regression model, it is necessary to focus on the parameterizations and to ask which parameterization is most natural, and for what purpose. Heagerty (1999) pointed out that individual components of β in the subject-specific parameterization are difficult to interpret unless the subject-specific effect η_u is known. Neuhaus *et al.* (1991), section 6, noted that, since each individual has her own latent risk, the model invites an unwarranted causal interpretation. Galbraith (1991) criticized the interventionist interpretation of parameters in equation (17) and pointed out correctly that additional assumptions are required to justify this interpretation in an observational study. If each pair of units having different treatment levels is necessarily a distinct pair of individuals or subjects, the treatment effect involves a comparison of distributions for two distinct subjects.

From this author's point of view, ephemeral unit-specific, subject-specific or cluster-specific effects such as $\{\eta_u\}$ or $\eta\{x(u)\}$ are best regarded as random variables rather than parameters, a distinction that is fundamental in statistical models. Given the parameters, the conventional model specifies the probability distribution for each unit and each set of units by integration. The intermediate step (17) shows a random variable arising in this calculation, leading to the joint distribution (3) whose one-dimensional distributions are well approximated by expression (18). Two units u and u' having the same baseline covariate values but different treatment levels have different response distributions. The treatment effect is the difference between these probabilities, which is usually measured on the log-odds-ratio scale. Although established terminology suggests otherwise, the treatment component of β^* in expression (18) is the treatment effect that is specific to this pair of units u and u'. If these units represent the same subject in a controlled crossover design, an interventionist interpretation is appropriate. Otherwise, if two units having different treatment levels necessarily represent distinct subjects, β^* is the difference of response probabilities for distinct subjects, so there can be no interventionist interpretation.

8.3. *Implications for applications*

Consider a market research study of consumer preferences for a set of products such as breakfast cereals. The relevant information is extracted from a database in which each purchase event is recorded together with the store information and consumer information. Breakfast cereal purchases are the relevant events. Following conventional notation, i denotes the purchase event, Y_i is the brand purchased and x_i is a vector of covariates, some store specific and some consumer specific. The aim is to study how the market share $\text{pr}(Y_i = r|x_i = x)$ depends on x, possibly by using a multinomial response model of the form (4). The random effects may be associated with store-specific variables such as geographic location, or consumer-specific variables such as age or ethnicity. The treatment effect may be connected with pricing, product placement or local advertising campaigns.

As I see it, the conventional paradigm of a stochastic process defined on a fixed set of units is indefensible in applications of this sort. Most purchase events are not purchases of breakfast cereals, so the relevant events (cereal purchases) are *defined* by selecting from the database those that are in the designated subset $\mathcal{C}$. An arbitrary choice must be made regarding the inclusion of dual use materials such as grits and porridge oats. Rationally, the model must be defined for general response sets, and we must then insist that the model for the subset $\mathcal{C}' \subset \mathcal{C}$ be consistent with the model for $\mathcal{C}$. Consistency means only that the two models are non-contradictory; they assign the same probability to equivalent events. The evolving population model with a fixed observation period is consistent under class restriction, but the conventional logistic model (4) with random effects is not.

The notation used above is conventional but ambiguous. The market share of brand r in stratum x is the limiting fraction of events in stratum x that are of class r, which is $\lambda_r(x)/\lambda_.(x)$ for both model (4) and the evolving population model. The expected market share is the stratum probability $\text{pr}(Y_i = r|i: x_i = x)$, which may be different from the conditional probability given x_i for fixed i. However, the central concept of a fixed unit i is clearly nonsense in this context, so the standard interpretation of $\text{pr}(Y_i = r|x_i = x)$ for fixed i is unsatisfactory.

The situation described above arises in numerous areas of application such as studies of animal behaviour, studies of crime patterns, studies of birth defects and the classification of bitmap images of handwritten decimal digits. The events are animal interactions, crimes, birth defects and bitmap images. The response is the type of event, so $\mathcal{C}$ is a set of behaviours, crime types, birth defects or the 10 decimal digits. This set is exhaustive only in the sense that events of other types are excluded: hence the need for consistency under class restriction.

In the biostatistical literature, which deals exclusively with hierarchical models, an expression such as $E(Y_i|X_i = x)$ is usually described as a conditional expectation but is often interpreted as the marginal mean response for those units i such that $X_i = x$. I do not mean to be unduly critical here because there can be no ambiguity if these averages are equal, as they are in a hierarchical model for an exchangeable process. For an autogenerated process, these averages are usually different. It is not easy to make sense of the literature in this broader context given that one symbol is used for two distinct purposes. To make the hierarchical formulation compatible with the broader context of the evolving population model, it is necessary to interpret models (3) and (4) as stratum distributions, not conditional distributions. Once the distinction has been made, it is immediately apparent that the stratum distribution does not determine the conditional probability given **x** for a sequential sample. Consequently, probability calculations using the stratum distribution, and efforts to estimate the parameters by using the wrong likelihood function (3), must be abandoned.

8.4. Sampling bias

The main thrust of this paper is that, when the units are unlabelled and sampling effects are properly taken into account by using the evolving population model as described in Sections 3, 5.4 and 7.1, there is no parameter attenuation. If the intensities are such that $\lambda_1(x)$ has the same mean as $\exp(\alpha+\beta x)\lambda_0(x)$, the correct version of expressions (17) and (18) for an autogenerated unit $u\in\mathbf{Z}_t$ is

$$\begin{aligned}\operatorname{logit}[\operatorname{pr}\{Y(u)=1|\lambda, u\in\mathbf{Z}_t\}]&=\log[\lambda_1\{x(u)\}]-\log[\lambda_0\{x(u)\}]\\&=\alpha+\beta x(u)+\eta\{x(u)\},\end{aligned}$$

$$\begin{aligned}\operatorname{logit}[\operatorname{pr}\{Y(u)=1|u\in\mathbf{Z}_t\}]&=\log[m_1\{x(u)\}]-\log[m_0\{x(u)\}]\\&=\alpha+\beta x(u),\end{aligned}$$

with no approximation and no attenuation. The distinction made in Section 8.2 between two parameterizations is simply incorrect for autogenerated units.

The subject-specific approach takes aim at the right target parameter in equation (17), but the conventional likelihood or hierarchical Bayesian calculation leads to inconsistency when sampling bias is ignored in the steps leading to model (3). Sample x-values occur preferentially at points where the total intensity $\lambda_.(\cdot)$ is high, which is not so for a predetermined **x**. As a result, parameter estimates from model (6) are inflated by the factor $1/\tau$ where τ is the apparent attenuation factor. The inflation factor reported by Neuhaus *et al.* (1991) is a little less than 3, so the bias in parameter estimates is far from negligible. The population-average procedure commits the same error twice, by first defining the stratum probability $\rho(x)$ as the target, and then failing to recognize that $E(Y|x)\neq\rho(x)$ for a random sample. But a fortuitous ambiguity of the conventional notation $E(Y|x)$ allows it to estimate the right parameter $\pi(x)$ consistently by estimating the wrong parameter $\rho(x)$ inconsistently.

For a sequential sample, the parameters α and β in equation (17) are exactly equal to the parameters α^* and β^* in the marginal distribution (18). The apparent attenuation arises not because of a real distinction between subject-specific and population-averaged effects, but because of failure to recognize and make allowance for sampling effects in the statistical model.

Acknowledgements

I am grateful to the referees for helpful comments on an earlier version, and to J. Yang for the R code that was used to produce Fig. 1.

Support for this research was provided by National Science Foundation grant DMS-0305009.

References

Baddeley, A. and Jensen, E. B. (2005) *Stereology for Statisticians*. Boca Raton: Chapman and Hall.

Besag, J. (1974) Spatial interaction and the statistical analysis of lattice systems (with discussion). *J. R. Statist. Soc.* B, **36**, 192–236.

Besag, J. and Kempton, R. (1986) Statistical analysis of field experiments using neighbouring plots. *Biometrics*, **42**, 231–251.

Breslow, N. E. and Clayton, D. G. (1993) Approximate inference in generalized linear mixed models. *J. Am. Statist. Ass.*, **88**, 9–25.

Cox, D. R. (1958) *Planning of Experiments*. New York: Wiley.

Cox, D. R. and Snell, E. J. (1979) On sampling and the estimation of rare errors. *Biometrika*, **66**, 125–132.

Galbraith, J. I. (1991) The interpretation of a regression coefficient. *Biometrics*, **47**, 1593–1596.

Glonek, G. F. V. and McCullagh, P. (1995) Multivariate logistic models. *J. R. Statist. Soc.* B, **57**, 533–546.

Green, P. and Silverman, B. (1994) *Nonparametric Regression and Generalized Linear Models*. London: Chapman and Hall.

Heagerty, P. J. (1999) Marginally specified logistic-normal models for longitudinal binary data. *Biometrics*, **55**, 688–698.

Laird, N. and Ware, J. (1982) Random effects models for longitudinal data. *Biometrics*, **38**, 963–974.

Lee, Y. and Nelder, J. A. (1996) Hierarchical generalized linear models (with discussion). *J. R. Statist. Soc.* B, **58**, 619–678.

Lee, Y., Nelder, J. A. and Pawitan, Y. (2006) *Generalized Linear Models with Random Effects*. London: Chapman and Hall.

Liang, K.-Y. and Zeger, S. L. (1986) Longitudinal data analysis using generalised linear models. *Biometrika*, **73**, 13–22.

McCullagh, P. (1983) Quasi-likelihood functions. *Ann. Statist.*, **11**, 59–67.

McCullagh, P. (2005) Exchangeability and regression models. In *Celebrating Statistics* (eds A. C. Davison, Y. Dodge and N. Wermuth), pp. 89–113. Oxford: Oxford University Press.

McCullagh, P. and Møller, J. (2006) The permanental process. *Adv. Appl. Probab.*, **38**, 873–888.

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

McCullagh, P. and Yang, J. (2006) Stochastic classification models. In *Proc. Int. Congr. Mathematicians*, vol. III (eds M. Sanz-Solé, J. Soria, J. L. Varona and J. Verdera), pp. 669–686. Zurich: European Mathematical Society Publishing House.

McCulloch, C. E. (1994) Maximum likelihood variance components estimation in binary data. *J. Am. Statist. Ass.*, **89**, 330–335.

McCulloch, C. E. (1997) Maximum-likelihood algorithms for generalized linear mixed models. *J. Am. Statist. Ass.*, **92**, 162–170.

Møller, J., Syversveen, A. R. and Waagepetersen, R. P. (1998) Log Gaussian Cox processes. *Scand. J. Statist.*, **25**, 451–482.

Neuhaus, J. M., Kalbfleisch, J. D. and Hauck, W. W. (1991) A comparison of cluster-specific and population-averaged approaches for analyzing correlated binary data. *Int. Statist. Rev.*, **59**, 25–35.

Schall, R. (1991) Estimation in generalized linear models with random effects. *Biometrika*, **78**, 719–727.

Schildcrout, J. S. and Heagerty, P. J. (2007) Marginalized models for moderate to long series of longitudinal binary response data. *Biometrics*, **63**, 322–331.

Shirai, T. and Takahashi, Y. (2003) Random point fields associated with certain Fredholm determinants, I: fermion, Poisson and boson point processes. *J. Functnl Anal.*, **205**, 414–463.

Wahba, G. (1985) A comparison of GCV and GML for choosing the smoothing parameter in the generalized spline smoothing problem. *Ann. Statist.*, **13**, 1378–1402.

Wolfinger, R. W. (1993) Laplace's approximation for nonlinear mixed models. *Biometrika*, **80**, 791–795.

Zeger, S. L. and Liang, K.-Y. (1986) Longitudinal data analysis for discrete and continuous outcomes. *Biometrics*, **42**, 121–130.

Zeger, S. L., Liang, K.-Y. and Albert, J. A. (1988) Models for longitudinal data: a generalized estimating equations approach. *Biometrics*, **44**, 1049–1060.

Discussion on the paper by McCullagh

N. T. Longford (*SNTL, Reading, and Universitat Pompeu Fabra, Barcelona*)

Linear regression of a single outcome variable is a key univariate method for studying what is essentially a multivariate setting of the outcome and one or several covariates. This paper finds a severe limitation to this univariateness in the context of logistic regression for correlated outcomes. It informs us of the non-ignorability of the process by which the values of the covariates are generated; by sampling (except for a special case), assignment or a mechanism of another kind.

The computational problem of fitting logistic regression to correlated outcomes has been resolved only

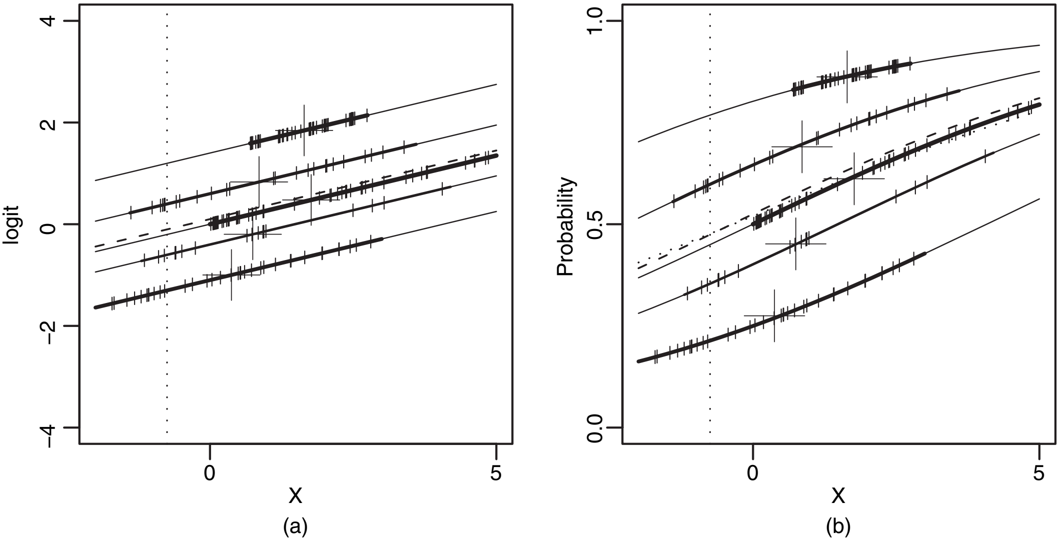


Fig. 2. Logistic regression with random coefficients—within-cluster sample sizes and distributions of *X* (⁚, point *x* where prediction is sought; +, within-cluster averages (before and after transformation)): (a) logit scale (———, conditional regression; – – –, marginal regression); (b) probability scale (– – –, transform of the average; · · · · · ·, average of the transform)

relatively recently (Stiratelli *et al.*, 1984; Schall, 1991), partly on the foundations of McCullagh and Nelder (1983). And now we hear of profound reservations about the application of these methods, which we would regard as perfectly well justified in a variety of settings. However depressing they may sound, these reservations are salient and far from overstated for a typical problem, when the values of the covariates X are not only outside our control, but also little or nothing is known about the function $X: u \rightarrow x$, which associates a unit with its value of X. Having to study the underlying multivariate (autogeneration) process erodes much of the attraction that regression has for the unsuspecting analyst.

Sampling bias, logistic regression and correlated outcomes conspire to distort the perceptions of good representation that are well founded under linearity and independence. Some of them are captured in Fig. 2 for clusters associated with varying logistic regressions that are parallel on the logit scale (Fig. 2(a)). The effects of the transformation to the probability scale are not dramatic, because the link is close to linearity, especially for probabilities that are close to 0.5. The lines or curves are drawn in the within-cluster support of X by thicknesses proportional to the population sizes, and the values of X realized in a replication are marked by vertical ticks.

There are several ways of averaging, taking into account the within-cluster distributions of X and the sample sizes, and these can vary across replications, when different clusters are selected. For a particular value of X, should we average over all clusters, or only clusters for which this value is plausible or apply weights that reflect the plausibility of this value of X?

The paper highlights an ambiguity much subtler than this; in essence, if we want to study within-stratum distributions $p_{\mathbf{x}}(y)$, we must stratify on the values $\mathbf{x}$ that are of interest. Setting the value of variable X and conditioning on a given value of X without stratifying on the values of X result in different distributions of outcomes. Even the established notation falls short because it has no means of indicating whether the condition in a probability statement is on a unit or only a value of X. The meaning of 'good representation' (by a sampling design) must be carefully qualified by the target(s) of inference. Some intrinsic honesty is established by the discovery that the answer to

'What happens to units that have $X = x$?'

can be obtained only by sampling from the stratum that is defined by x. Such a design is not necessary in linear models or with independent outcomes. Note the parallels with causal analysis (Holland, 1986; Rubin, 2005), in which the question

'What would happen if the unit had a different value of $\mathbf{x}$?'

can be inferred without any assumptions only when we can assign values of X to units (e.g. by randomization).

The results about the sampling protocols in Section 3 can be interpreted as failed attempts (not the author's failures!) at exchanging, mentally or analytically, two operations (ratio and expectation) which commute under linearity or certainty, but not otherwise. But in the mental processes these two operations are difficult to identify.

In most applications, quotas for the values of X cannot be enforced, and so we must study the process that generates the values of X. I think that the proposed models of evolving population are but models, i.e. they inject some realism into the analysis but do not make the bias disappear. In any case, the choice between the available options is not trivial for a secondary analyst who has modest resources to study a sampling process that may have involved some improvisation and the details of which are poorly recorded.

The paper is concerned solely with bias, which asymptotically dominates the mean-squared error in inferences with invalid models and assumptions. Parsimony is valuable with small samples, because it promotes smaller sampling variation. The inevitable uncertainty that will arise as we study and model the process that generates the values of x may bring about a variance inflation that is in excess of the squared bias. There is therefore some scope for compromise, to use models that are not valid but are associated with only limited bias. We should not pretend that such models are valid, but estimate the squared bias incurred or seek an upper bound for it. Some avenues for this are implied in the paper and I think that their elaboration would be useful.

I conclude by proposing the vote of thanks for this excellent paper.

Peter J. Diggle (*Lancaster University and Johns Hopkins University, Baltimore*)

At a very simple level, this paper is a plea for precision of notation. At a slightly deeper level, it warns us against indulging in statistical modelling without respecting the study design. At a considerably deeper level, it challenges conventional wisdom regarding the widespread use of unobserved random variables as fundamental building-blocks in what are variously called random effects, latent variable, multilevel or hierarchical models.

I agree with Peter McCullagh that confusion can easily arise through the use of the same everyday word 'given', and the same accompanying notation, to mean any one of three quite different things: fixing the value of a non-random variable; conditioning on the realized value of a random variable; indexing a family of distributions by the value of a parameter (although Bayesians may regard the second and third as being equivalent). I would agree with Peter McCullagh's use of $p_x(y)$ for the first of these, reserving $p(y|x)$ for the second. For the third, I favour $p(y;\theta)$ because we often use θ as the argument of a function, e.g. a likelihood.

The importance of good study design can never be overemphasized, and the same surely applies to the need to respect the study design when analysing the data. This is not an argument for accepting only inferences that are purely design based. But there is a difference between an unverifiable assumption and an assumption that is palpably false, and Peter McCullagh has shown us that, if we pretend that a stochastic sampling process is a prespecified study design, we can get ourselves into serious trouble.

I have lately taken an interest in this question in the specific setting of geostatistical modelling. There, the data in their simplest form consist of responses $Y_i : i=1,\ldots,n$ associated with locations $x_i \in A$, where A is a spatial region of interest. Typically, the Y_i are considered to be noisy measurements of the value at x_i of a spatial process $S(x) : x \in \mathbb{R}^2$. A widely used model is that $Y_i|S(\cdot) \sim N\{\beta + S(x_i), \sigma^2\}$, the Y_i are conditionally independent given $S(\cdot)$ and $S(\cdot)$ is a zero-mean Gaussian process. Standard geostatistical methods assume, if only implicitly, that x is either a non-random variable or is stochastically independent of $S(\cdot)$. But not infrequently this assumption is false. For example, in environmental monitoring networks the design prescription for the monitoring locations is rarely stated but in a rather vague sense will typically place monitors in areas that are suspected to be highly polluted. To investigate the implications of this, we extend our notation to include a point process of locations X, in addition to a measurement process Y, and an underlying spatial signal S, and adopt the notation $[\cdot]$ for 'the distribution of'. Then, the standard geostatistical model can be written as

$$[S, X, Y] = [S]\,[X|S]\,[Y|X, S] = [S]\,[X] \prod [Y_i|S(X_i)]$$

and conditioning on $[X]$ is innocuous for inference about S and/or Y. In contrast, for *preferentially sampled* geostatistical data, the model becomes

$$[S, X, Y] = [S]\,[X|S]\,[Y|X, S] = [S]\,[X|S] \prod [Y_i|S(X_i)].$$

The likelihood for the data X and Y is now

$$\int [S]\,[X|S] \prod [Y_i|S(X_i)]\,\mathrm{d}S$$

and conditioning on X is not the right thing to do. The consequences of this are explored in Diggle *et al.* (2008).

Similar issues arise in longitudinal studies, e.g. when study participants deviate from their assigned follow-up protocol for reasons unknown; see, for example, Lin *et al.* (2004), Lipsitz *et al.* (2002) or Sun *et al.* (2007).

Returning to Peter McCullagh's main focus of attention, the logistic regression model with a random effect added to the linear predictor, I accept that this model is overused, but to state that 'efforts to estimate the parameters by using the *wrong* likelihood (3) *must* be abandoned' (my italics) seems a little sweeping. A possible interpretation of model (3), when values of explanatory variables are not prespecified, is as follows. We assume an incomplete model

$$\operatorname{logit}\{P(Y=1|X,U)\}=\alpha+\beta X+\gamma U$$

for the random vector (Y,X,U) and wish to make inferences about β. We are unable to observe U but use independent random sampling from the population of interest to obtain data (y_i,x_i), $i=1,\ldots,n$. A complete-model specification takes the form

$$[Y,X,U]=[Y|X,U]\,[X|U]\,[U]. \tag{19}$$

We have no interest in the distributions of X or of U but are willing to assume a parametric form for the distribution of U induced by the random sampling design. If, additionally, we assume that X and U are independent, then in equation (19) we can condition on X and marginalize with respect to U, resulting in the likelihood

$$\int [Y|X,U]\,[U]\,\mathrm{d}U,$$

which is distribution (3). In most discussions of random-effects models the assumption that X and U are independent is left unstated and is probably unrealistic in many applications. But if this assumption is made explicit, and thereby open to criticism, I can see no logical objection to the use of model (3).

As always, Peter McCullagh has challenged us to think very carefully about matters of fundamental importance in the application of the statistical method to substantive problems. I am very pleased indeed to second the vote of thanks.

The vote of thanks was passed by acclamation.

John Nelder (*Imperial College London*)

Underneath equation (3), which defines the marginal likelihood after integrating out the random effects, the author states 'The word model refers to these distributions'. To me this seems to give a very restrictive class of random-effect models, in which the random effects are used only to produce overdispersion in the distribution of the responses. The author's definition does not allow the random effects to be estimated (a word which I prefer to 'predicted') and hence there is no way of checking whether their distribution satisfies what has been assumed in setting up the model. I regard such models as inherently defective from the model fitting point of view.

In seeking to deal with these problems, while keeping to a likelihood type of inference, Lee *et al.* (2006) were led to extend Fisher likelihood (of which equation (3) is an example) to produce an extended likelihood in which random effects could appear. We are well aware that the theory underlying inferences from this extended likelihood (or, more exactly, the *h*-likelihood that is derived from it) is as yet incomplete, but extensive simulation on a wide variety of models from the class of double hierarchical generalized linear models has shown consistently good results, as measured by bias and mean-squared error. The algorithm for fitting these models reduces to a set of interconnected generalized linear models, each using iterative weighted least squares. We allow random effects to appear in the linear predictors of either mean or dispersion or both, and to have distributions belonging to any exponential family conjugate distribution. Given enough data, all the assumptions about the parameters in the model can be checked by using standard model checking methods. The random effects may represent overdispersion effects (as in this paper) but also patterns in a field layout, or in a combined analysis of a set of clinical trials the effects of hospitals regarded as a sample from a distribution of a hypothetical population of effects. Although we have certainly not said the last word on our model class, we believe that development should start from the author's equation (2) rather than equation (3).

J. B. Copas (*University of Warwick, Coventry*)
I join others in welcoming Professor McCullagh back to the Society and thank him for his elegant and challenging paper: challenging to the reader (at least to this reader), but also a challenge to all of us that we need to think much more carefully about what we mean by regression. Elementary statistics is dominated by two fundamental concepts: the normal distribution and the linear model. The wonderful properties of the normal distribution, which is used so widely in approximations in applications, shields us from having to face up to deeper problems of inference. It is the same with the linear model: it applies so widely and gives such apparently neat answers that most of us have never had to face up to the issues that are raised in this paper.

As always, the key question is what is the population, that imagined structure behind our data that we want to find out about? In ordinary regression it is easy; we just fix the xs at the values we happen to observe, and there is no ambiguity. But, when random effects are in some sense correlated with the process generating the data, fixing the xs no longer captures all the features of interest, and this paper tells us that we must model the complete data process, as in the population model that is set out at the start of Section 3. But in practice this is a very big challenge. The Cox process here is a theoretical ideal; in practice there will be many complications along the way of arriving at our data which we would rather not have to know about. How convenient it would be if we can get away with a model which tells us that we do not need to know about them!

At another regression paper read to our Society over 20 years ago, one of the discussants commented that the then author's estimate of 100000 regressions *per day* was probably an underestimate by a factor of 10. He went on to wonder how many of these analyses were sensible! 20 years on, the number of applications of regression has probably grown by a further factor of 10. And users are no longer restricted to linear regression; we now have software to handle quite complicated non-linear random-effects models. How many of these applications I wonder fall foul of the author's comments in Section 8, that they are based on models which are 'indefensible' or using methods which 'must be abandoned'? How can we make the message of this important paper accessible to users of statistics?

Gavin J. S. Ross (*Rothamsted Research, Harpenden*)
In the analysis of overdispersion (Section 6.2) it is of interest to re-examine the classical approach adopted by R. A. Fisher in relation to bioassays with heterogeneous errors. Fisher advised that when the residual deviance is significantly large and there is no indication that the probit or logistic regression model is inappropriate then the estimated dispersion matrix of the parameters should be multiplied by a heterogeneity factor, the residual deviance divided by the degrees of freedom. See the introduction to Fisher and Yates (1974).

An alternative approach is to assume that the units are independent random samples from a beta–binomial distribution. If the sum of the indices is a specified constant for each sample it is possible to select the constant which will reduce the heterogeneity factor to 1. The log-likelihoods are easily specified in terms of log-gamma functions, and the maximum likelihood estimates of the parameters and their dispersion matrix can be computed.

Table 1. Logit fit to the data of Finney (1952)

Parameter	*Binomial errors with heterogeneity*	*Beta–binomial errors with sum of indices 5.4*
LD50	0.2384 (0.0224)	0.2422 (0.0231)
Slope	14.441 (3.921)	14.853 (3.773)
Deviance on 8 degrees of freedom	36.25	8.02

As an example of this analysis, the data that were used by Finney (1952) in his Table 10 may be fitted by using logits rather than probits, with the results that are given in Table 1 (with estimated standard errors in parentheses).

As the sum of beta-indices is progressively reduced the estimated variances increase. The likelihood contours in parameter space are slightly different for the two approaches, but Fisher's approach appears to be quite adequate for practical purposes.

Jesper Møller (*Aalborg University*)
I welcome this stimulating paper, with its use of point process models for discussing sampling bias in regression models. The central model is a Cox process model evolving in time t and with marks (y, x), where y is a response taking values in a finite set and x is a spatial location. It is driven by an intensity of the form $\lambda_y(x)\ \nu(\mathrm{d}x)\ \mathrm{d}t$, with a particular focus on log-Gaussian models or gamma models for the random functions $\lambda_y(x)$. Note that the marks (y_i, x_i) are independent of the times t_i. As discussed below, other point process models could be of relevance.

Could other types of Cox models be included? Shot noise Cox processes (e.g. Møller (2003)) are, in contrast with log-Gaussian models (Section 5.2), closed under aggregation, whereas moment expressions are less simple (e.g. Møller and Waagepetersen (2004)). Furthermore, McCullagh's random intensity may be extended to the time inhomogeneous case $\lambda_y(x)\ \nu(\mathrm{d}x)\ \kappa(\mathrm{d}t)$, say, so that, for example, quota sampling remains unbiased.

Another important class of point process models with marks is based on modelling the conditional intensity, i.e. when we at time t condition on the history of the evolving process (e.g. Daley and Vere-Jones (2003)). For example, expressions for the unconditional and conditional distributions considered under the various sampling protocols in Section 3.2 and the limit distribution in Section 4 will be of a similar form to that in McCullagh's paper. In particular, if the marks (y_i, x_i) are independent of the times t_i and distributed as in McCullagh's Cox model, we again obtain equation (6), and hence quota sampling is unbiased under the log-Gaussian model that is specified a few lines after equation (6), and we still obtain the limiting low intensity conditional distribution in equation (11), and so the theory in Section 5 applies.

For spatial point process modelling, the analogy with generalized linear models and random-effects models is to some extent discussed in Møller and Waagepetersen (2007).

The following contributions were received in writing after the meeting.

D. R. Cox (*Nuffield College, Oxford*)
Professor McCullagh's challenging paper on an important topic raises, implicitly or explicitly, many issues, in particular about generalized linear mixed models, especially those of logistic form.

It is natural to compare analyses of these logistic models with the corresponding analysis-of-covariance table (Wilsdon (1934), appendix A). Analysis, at least informally, of parallelism and of possible systematic departures from parallelism seems important for interpretation. Further the notion of two regressions, one within and the other between study individuals, may be relevant. These issues seem not much discussed in the more recent literature.

If the choice of individuals for study is not conditionally independent of outcome given the explanatory variables it is clear that careful analysis of the selection process is essential and the elegant discussion in the paper of the connection with point processes is very interesting. It would be valuable to see more explicit development in terms of one or two specific examples. Some of those sketched in the paper are not entirely convincing. For example, fibres in a traditional textile yarn are identified in principle by the position of their 'left' end, thus providing a sampling frame, and careful methods of sampling either respect that or introduce appropriate corrections. Selection of patients in a clinical trial seems broadly similar except for the issue of non-compliance. Non-compliance arising *during* a study has been quite extensively studied. So far as I know, although it is widely recognized that the population of individuals involved, those giving informed consent, deviates, possibly seriously, from the target population, there is little work on what, if anything, can be done about non-compliance before entry. Can the ideas in the paper make a contribution to this particular version of non-compliance?

Ruth Keogh (*University of Cambridge*)
I found Professor McCullagh's paper very interesting and challenging and am interested in his views on extensions to different types of study.

In Section 3 the author discusses studies of associations between covariates and event occurrence in an evolving population and considers several sampling protocols which may be used to select the study units. The functions $p_{\mathbf{x}}(\mathbf{y})$ and $p(\mathbf{y}|\mathbf{x})$ which arise under different protocols are compared. The case–control design is one of the protocols considered and I find it interesting to compare McCullagh's derivation of the conditional probability with the work of Prentice and Pyke (1979) and Farewell (1979), who established the validity of analysing data obtained from a case–control study as though they had arisen prospectively.

Although the case–control design is discussed in the paper, the various cohort sampling designs that are commonly used in epidemiological studies do not appear to be considered. Under cohort sampling schemes cases and controls are sampled from an underlying cohort taking account of event times. The sampling may be performed during follow-up or retrospectively in a pre-existing cohort. I am interested to hear any comments from the author regarding how the issues raised in the paper may extend to studies using cohort sampling designs.

Under the cohort sampling designs the set of cases to be studied is commonly taken to be all individuals experiencing the event(s) of interest in a fixed time period. The identification of the cases thus seems essentially to follow the protocol in Section 3.2.3. Controls are selected according to one of several designs, including the nested case–control design using incidence density sampling (Goldstein and Langholz, 1992), the case–cohort design (Prentice, 1986), and more complex nested case–control designs using countermatching or quota sampling, for example (Borgan *et al.*, 1995). The analysis of data arising from epidemiological studies using cohort sampling designs differs from methods that are described in the paper; whereas the probabilities under the models in the paper are unit specific this is not quite so under the cohort sampling designs, in which the contribution to the likelihood from each case is conditional on covariate values in the comparison set. The control sampling procedure is incorporated in the analysis by using a weighted partial likelihood or pseudolikelihood. However, the way in which the cases arose does not appear to be considered. In light of the results in the paper, should we be concerned about the methods that are used to analyse data arising from cohort sampling designs?

Anthony Y. C. Kuk (*National University of Singapore*)

The author is to be congratulated for proposing an evolving population framework to factor in the effect of sampling plan and for pointing out that, for most sampling plans, the conditional distribution $p(\mathbf{y}|\mathbf{x})$ is not the same as the regression function $p_{\mathbf{x}}(\mathbf{y})$.

Analysis of clustered data with informative cluster size has received a fair amount of attention recently (Hoffman *et al.*, 2001; Williamson *et al.*, 2003; Benhin *et al.*, 2005). An often-quoted example is periodontal disease data where the dental health of a person will affect both the disease status of teeth as well as the number of teeth (cluster size) that a person has. I shall show how the simplest evolving population model discussed in Section 6.1 can be modified to model situations with non-ignorable cluster sizes. As in Section 6.1, we assume that, given λ, the number m of events in unit time is Poisson distributed with parameter λ. However, rather than letting the two components of λ, λ_0 and λ_1, be independent, we find it more appropriate for our purpose to model the joint distribution of (λ_0, λ_1) via the marginal distribution of their sum $\lambda = \lambda_0 + \lambda_1$ and the conditional distribution of λ_1/λ given λ. We can think of λ as a frailty term that reflects the health condition of a person. Given λ, a plausible way to model the effect of λ on λ_1 could be something like $E(\lambda_1/\lambda|\lambda) = \xi \exp(-\beta\lambda)$, $0 < \xi < 1$, $\beta \geqslant 0$, so that $\beta = 0$ corresponds to the case of ignorable cluster size and $\beta > 0$ will induce a negative correlation between cluster size and disease status. The marginal probability of $Y_1 = 1$ in the k-variate distribution of $(Y_1, \ldots, Y_k)$, $k \geqslant 1$, can be obtained as

$$p_k(Y_1 = 1|m \geqslant k) = E\left[\left\{1 - \sum_{i=0}^{k-1} \frac{\exp(-\lambda)\lambda^i}{i!}\right\} \frac{\lambda_1}{\lambda}\right] \bigg/ E\left\{1 - \sum_{i=0}^{k-1} \frac{\exp(-\lambda)\lambda^i}{i!}\right\}$$

$$= E\left[\left\{1 - \sum_{i=0}^{k-1} \frac{\exp(-\lambda)\lambda^i}{i!}\right\} \xi \exp(-\beta\lambda)\right] \bigg/ E\left\{1 - \sum_{i=0}^{k-1} \frac{\exp(-\lambda)\lambda^i}{i!}\right\}$$

which can be expressed in terms of quantities like $E\{\lambda^i \exp(-\gamma\lambda)\} = M_\lambda^{(i)}(-\gamma)$ involving derivatives of the moment-generating function of λ. A convenient case is gamma-distributed λ with easily differentiable $M_\lambda\{t\} = (1 - \theta t)^{-\alpha}$. As a demonstration, when $\alpha\theta = 10$ (the mean cluster size), $\alpha = 0.9372$ (which is chosen to make the probability of zero cluster size equal to 0.1), $\xi = 0.4$ and $\beta = 0.1$, the first five $p_k(1)$ are computed to be 0.184, 0.168, 0.153, 0.140 and 0.128, which decrease with cluster size, compared with $p_k(1) \equiv 0.4$ when $\beta = 0$. More generally, we can model λ by using the three-parameter family of distributions that was proposed by Hougaard *et al.* (1997) defined by the Laplace transform $L(s) = \exp[-\delta\{(\theta + s)^{\alpha} - \theta^{\alpha}\}/\alpha]$ which includes the gamma, inverse Gaussian and positive stable distributions as special cases, leading to examples where the distribution of cluster size can be different from negative binomial.

S. L. Lauritzen (*University of Oxford*) **and T. S. Richardson** (*University of Washington, Seattle*)

This thought-provoking paper seems to address questions which reach beyond those of simple bias in logistic regression models and point at fundamental issues of interpretation of conditional probabilities.

Recent work in causal inference (Spirtes *et al.*, 1993; Pearl, 1993, 1995, 2000) emphasizes the distinction between what we here shall term *conditioning by restriction* $P\{Y|\text{is}(X=x)\}$, which is often written as $P\{Y|X=x\}$, and *conditioning by intervention* $P\{Y|\text{do}(X=x)\}$. The latter, unlike the former, requires something in addition to a joint distribution, which may be specified by using potential responses, counterfactuals or graphical models, as here.

This paper highlights a distinction between conditioning by restriction and *conditioning by observation* which we shall denote by $P\{Y|\text{see}(X=x)\}$. To assess the latter, the protocol by which the information is revealed must also be modelled, a phenomenon which is illustrated by the *prisoner's dilemma* (Mosteller, 1965) and discussed extensively in different manifestations (Shafer, 1985; Grünwald and Halpern, 2003), also in the context of *coarsening at random* (Heitjan and Rubin, 1991; Jacobsen and Keiding, 1995; Gill *et al.*, 1997; Jaeger, 2005). In short, there might be information in the fact that a quantity is observed, over and above the observed value.

The issues involved can often be summarized by simple graphs which reflect the causal and observational processes. In Fig. 3(a) we have

$$P\{Y|\text{do}(X=x)\} = \int P(Y|X=x,\lambda)\,P(\mathrm{d}\lambda), \tag{20}$$

$$P\{Y|\text{is}(X=x)\} = \int P(Y|X=x,\lambda)\,P(\mathrm{d}\lambda|x), \tag{21}$$

$$P\{Y|\text{see}(X=x)\} = \int P(Y|X=x, S\in\mathbb{S}_x,\lambda)\,P(\mathrm{d}\lambda|x, S\in\mathbb{S}_x), \tag{22}$$

where $\mathbb{S}_x$ is the set of samples in which $X=x$ is observed.

If one of the links $\Lambda \to Y$ or $\Lambda \to X$ is missing equations (20) and (21) are equal since either $Y \perp\!\!\!\perp \Lambda|X$ or $X \perp\!\!\!\perp \Lambda$.

In the context of the paper, Λ, X and Y would represent corresponding elements of the stochastic processes in Section 3.1. The sampling process S indicates the selection of observations. The quota sample (Section 3.2.1) then corresponds to Fig. 3(b), which gives equality of equations (21) and (22) since $Y \perp\!\!\!\perp S|X$. Similarly, the sequential sample for fixed time (Section 3.2.3) corresponds to Fig. 3(d): inclusion in the sample does not depend on X or Y; hence we have $(X,Y) \perp\!\!\!\perp S$ and again equations (21) and (22) are equal. A case–control study (Section 3.2.6) has no link between X and S, as shown in Fig. 3(c), whereas the sequential sample of fixed size (Section 3.2.2) and simple random sample (Section 3.2.4) correspond to Fig. 3(a); in none of these cases do we have equality of equations (21) and (22).

Issues involving a treatment effect as in Section 5.4 would be represented by introducing another node T.

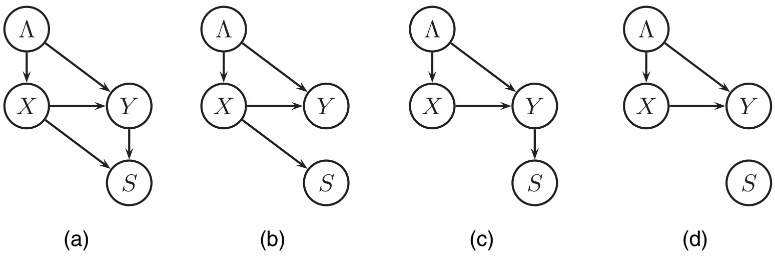


Fig. 3. Graphs corresponding to sampling schemes: (a) sequential sample of fixed size, simple random sample; (b) quota sample; (c) case–control sample; (d) sequential sample for fixed time

Youngjo Lee (*Seoul National University*)

I congratulate the author on finding that sampling procedures could lead to distributions of random effects without attenuation in binomial hierarchical generalized linear models. Such attenuation in the behaviour of parameters between two models, if it occurs, is based on a failure to compare like with like (Lee and Nelder, 2004). A question is whether the sampling model proposed is natural in biostatistics; for example, in a cohort study, the population base is defined at a moment in time.

It is often not clearly stated what we model, but in random-effect models we model $p(y|x,\eta)$, rather than $p(y|\eta)$ in model (2), leading to $p(y|x)$ via the integration (3). In my notation $p(y|x,\eta)=p_\theta(y|x,\eta)$ where the suffix θ is a parameter, which is suppressed for simplicity. When $\eta(\cdot)$ is a known monotone function such as $\eta(x)=\beta_0+x\beta_1$, we have $p(y|x,\eta)=p_{\beta_0,\beta_1}(y|\eta)=p_{\beta_0,\beta_1}(y|x)$, so that all $p(y|x)$, $p(y|\eta)$ and $p(y|x,\eta)$ give an equivalent model. However, if a functional form of $\eta(\cdot)$ is totally unknown it would be natural to model $p(y|x,\eta)$ for the random-effect model (2). Even if we could make a full joint model $p(y,x|\eta)=p(y|x,\eta)\ p(x|\eta)$, information would be hardly recoverable from $p(x|\eta)$ because of the infinite dimensional parameter space of η. Thus, we model $p(y|x,\eta)$.

Many interesting random-effect models have been proposed which involve modelling of unobservables such as random effects. We do not agree that the use of an integrated likelihood is non-Bayesian, whereas an unintegrated likelihood is Bayesian. The extended likelihood principle of Bjørnstad (1996) states that the evidence for unobservables is contained in the extended likelihood such as the *h*-likelihood, and Professor Nelder and I have shown how efficient inferences can be made from it, without resorting to estimating equations. The random-effect model is not just an intermediate for the marginal model. It can also be used for various purposes, such as robust estimation for generalized linear model classes, deriving new sandwich estimators, imputation, animal breeding, smoothing and image recovery... (Lee *et al.*, 2006).

Xiao-Li Meng (*Harvard University, Cambridge*)

It seems trite to reiterate the importance of identifying the data-generating mechanism for proper analysis, especially in a top statistical journal. Yet McCullagh's paper shows how easy it is to overlook such issues, even by many professional statisticians and for methods as well studied as logistic regression.

McCullagh is extremely careful with notation: $p(y|x)$ *versus* $p_x(y)$ and $E(y_u|x_u=x)$ *versus* $E(y_u|u,x_u=x)$. The importance of avoiding 'dangerously ambiguous' notation indeed cannot be overstated. McCullagh's formulation avoids the notion of units, but even without this advance a more explicit notation would clearly remind us of the importance of the data-generating mechanism. Let $\{(y_i,x_i), i=1,\ldots,n\}$ be the *observed* values of (Y,X) from a population. To model them, we typically write $P(y_i,x_i)$, forgetting that it should really be $P(y_i,x_i|O_i=1)$, where O_i is a Bernoulli variable indicating whether the ith subject's (Y,X) are *recorded and observed by the investigator*. With this notation, few statisticians would miss the distinction between $P(Y|X,O=1)$, similar to McCullagh's stratum distribution, and $P(Y|X)$, the distribution of interest. The two are the same if, and only if, Y and O are conditionally independent given X. This conditional independence is sometimes referred to as the ignorability of sampling mechanism in survey literature. In the context of missing data, a perfect example of 'autogenerated data', this conditional independence is key to ensuring an *ignorable missing data mechanism* (e.g. Rubin (1976)).

An on-going longitudinal study of young women's sexual behaviour and risks highlights well the critical importance of recognizing the conditioning on O. The study asks each subject to keep a diary to record her daily sexual activities and risks (if she had sexual intercourse, used a condom, whether the condom broke, etc.). Currently the shortest follow-up is about 3 years and the longest is about 7 years. However, as the study continues, the investigators are increasingly worried about the possibility that the very act of keeping a detailed diary could have changed a woman's sexual behaviour or risk exposure, since recording serves as a daily reminder, i.e. $P(Y|X,O=1)$ potentially can be very different from $P(Y|X,O=0)$. The problem was posed to me recently as a challenge: $P(Y|X,O=0)$ is logically impossible to estimate because if a woman did not keep a diary then the investigator would have no data! Although there is no space here to detail my brilliant solution (in the style of Pierre de Fermat), clearly carrying the 'O' notation at least helps to prevent a long 'Oops' when inferring from the study to the general population of young women.

Paul Rathouz (*University of Chicago*)

Professor McCullagh has provided an interesting perspective on sampling models for binary data in the presence of heterogeneity. His notion of 'autogenerated units' is applicable in epidemiologic studies of disease incidence. Here, I explore application of his ideas to case–control studies (Rothman (1986), page 62). I ask whether and how, in the presence of heterogeneity, data generated via case–control designs and subsequently analysed with logistic regression will lead to sampling bias in the estimation of disease incidence rates.

Consider a fixed population $\mathcal{U}$ of units $u_1, u_2, \ldots$ at risk for some disease event, and suppose that each unit is associated with a covariate vector $x_i \equiv x(u_i) \in \mathcal{X}$. Covariate $x_i = (w_i, z_i)$ may include vector $w_i \equiv w(u_i)$ containing exposure and adjustor variables; x_i may also include $z_i \equiv z(u_i)$ containing spatial or clustering information, thereby accommodating a model for population heterogeneity. Population size at $x \in \mathcal{X}$ is given by measure $\nu(x)\,\mathrm{d}x$.

Following McCullagh's 'evolving population model', disease incidence in $\mathcal{U}$ arises via random intensity function $\lambda_1(x)$ on $\mathcal{X}$. Let $E_\lambda\{\lambda_1(x)\} = m_1^*(x)$; in applications, it is useful to consider models wherein $m_1^*(x) = m_1(w)$. For any $\mathcal{A} \subset \mathcal{X}$, the *stratum disease incidence rate* is the expected number of disease events for the subpopulation with $x(u) \in \mathcal{A}$, divided by the size of that subpopulation, i.e.

$$h(\mathcal{A}) = E_\lambda\left\{\int_{\mathcal{A}} \lambda_1(x)\,\nu(x)\,\mathrm{d}x\right\} \bigg/ \int_{\mathcal{A}} \nu(x)\,\mathrm{d}x = \int_{\mathcal{A}} m_1^*(x)\,\nu(x)\,\mathrm{d}x/\nu(\mathcal{A}).$$

In applications, interest lies in stratum incidence rates for subpopulations with covariate values $w(u) = w_0$, i.e. $h(w_0) \equiv h(\mathcal{A}_{w0})$ for sets $\mathcal{A}_{w0} = \{(w, z) \in \mathcal{X} : w = w_0\}$, and in stratum incidence rate ratios $h(w_1)/h(w_0)$.

In a simple case–control design, controls are sampled independently of cases with intensity $\lambda_0(x)$. Control sampling is non-differential if $E_\lambda\{\lambda_0(x)\} = m_0$ does not depend on x. McCullagh's development can be applied to $\mathcal{X}$, $\lambda_1(x)$ and $\lambda_0(x)$ with case and control data 'sequentially sampled' under the protocols in Sections 3.2.2 and 3.2.3, yielding the *conditional probability* $p(Y_i = 1|x_i)$. It follows that $p\{Y = 1|x = (w, z)\} = \pi(w) \equiv E_\lambda\{\lambda_1(x)\}/E_\lambda\{\lambda_\bullet(x)\} = m_1(w)/\{m_1(w) + m_0\}$ can be consistently estimated by using the logistic regression estimating function methodology in Section 7.2.

In general, $\pi(w) \neq h(w)$, so sampling is biased. Consider, however, the special case where $\log\{\lambda_1(x)\}$ and $\log\{\lambda_0(x)\}$ are Gaussian processes with respective means $\beta_0 + w'\beta$ and α_0 and variances $\sigma_1^2\,K_1(z, z')$ and $\sigma_0^2\,K_0(z, z')$, with $K_j(z, z) = 1$. Then $h(w) = \exp(\beta_0 + \frac{1}{2}\sigma_1^2 + w'\beta)$ and

$$\pi(w) = \frac{\exp(\beta_0^* + w'\beta)}{1 + \exp(\beta_0^* + w'\beta)},$$

where $\beta_0^* = \beta_0 + \frac{1}{2}\sigma_1^2 - \alpha_0 - \frac{1}{2}\sigma_0^2$. It follows that logistic regression will yield consistent estimation of the log-hazard ratio β.

Bias under different case–control sampling schemes and under non-log-Gaussian random intensity functions warrants further study.

Tyler J. VanderWeele (*University of Chicago*)

I thank the author for an interesting and stimulating paper concerning which I offer two comments. First, it would be useful to consider further the types of applications in which the form of sampling bias the author describes is likely to be an issue. Although this form of sampling bias may be common in problems of market research or in assessing vehicle speed, I am less convinced that the bias is likely to be present in clinical trials with volunteers enrolling over time. The author notes that the sampling bias he describes will not be present if the intensity ratios $\lambda_r(x)/\lambda_\bullet(x)$ are independent of the overall intensity $\lambda_\bullet(x)$; this would include all settings in which the intensities $\lambda_r(x)$ or the ratios $\lambda_r(x)/\lambda_\bullet(x)$ are deterministic rather than random. Consider also a clinical trial, evaluating a drug that is thought to delay the onset of Alzheimer's disease in high risk individuals. Suppose that the total enrolment period is of length T and that the outcome of interest is Alzheimer's disease at 5 years after a subject's enrolment. If the set of covariates X is sufficiently rich it seems unlikely that within strata of X the decision to enrol will depend on an unknown outcome to take place in 5 years' time. If, within strata of X, the decision to enrol does not depend on the future outcome then for each x the possibly random proportion $p(x)$ who develop Alzheimer's disease after 5 years will be independent of the possibly random number $e(x)$ of those who will enrol. Consequently,

$$\left(\frac{\lambda_0(x)}{\lambda_\bullet(x)}, \frac{\lambda_1(x)}{\lambda_\bullet(x)}\right) = \left(\frac{\{1 - p(x)\}e/T}{e/T}, \frac{p(x)e/T}{e/T}\right) = (1 - p(x), p(x))$$

will be independent of $\lambda_\bullet(x) = e/T$ and thus this particular form of sampling bias will not be present.

Second, even if $\lambda_r(x)$ are deterministic rather than random functions of x, a different form of sampling bias might still be present. Consider, for example, the causal diagram given in Fig. 4 (see Pearl (1995)).

Suppose in this example that U_1 and U_2 are unmeasured and that the intensity functions $\lambda_r(x)$ are deterministic but vary with x_2. Let $X = \{X_1, X_2, V\}$ and $X' = \{X_1, V\}$. Without controlling for X_2, the

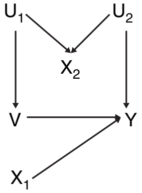


Fig. 4. Causal diagram demonstrating collider stratification bias

effect of V on Y is unconfounded but because the sampling process is such that $\lambda_r(x)$ varies with x_2 it can be shown that estimates of the effect of V on Y will be biased because of 'collider stratification' (Pearl, 1995; Greenland, 2003; Hernán *et al.*, 2004). The issue here is not that the distributions $p_x(y)$ and $p(y|x)$ differ but rather that the distribution $p(y|x)$ does not reflect the outcome distribution $p(y_v|x)$ of the sample under a particular intervention to set V to v. Note that $p(y|x')$ and $p(y_v|x')$ will also in general differ.

The **author** replied later, in writing, as follows.

I thank the discussants for their thoughtful and stimulating remarks. So far as possible, my responses are arranged by topic.

Biased sampling

It is worth restating the point made in Section 3.3, using Meng's notation in which $O_i = 1$ indicates that unit i will volunteer if asked. We distinguish between the conditional distribution $p_n(\mathbf{x}, \mathbf{y}|\mathbf{o} \equiv 1)$ given that a fixed sample of n individuals happens to have no refusers, and the distribution $p_{\mathbf{o}=1}(\mathbf{x}, \mathbf{y})$ for the first n volunteers. In a random-effects model where the responses for distinct units are correlated, these distributions are usually different. Longford will be disappointed to learn that the conditional distributions $p_n(\mathbf{y}|\mathbf{x}, \mathbf{o} \equiv 1)$ and $p_{\mathbf{o}=1}(\mathbf{y}|\mathbf{x})$ are seldom equal either. Whether they are the same or different, it is the stratum distribution $p_{\mathbf{o}=1}(\mathbf{x}, \mathbf{y})$ that is relevant for volunteer samples.

In defence of the generalized linear mixed model

Copas, Cox, Diggle, Meng and VanderWeele point to the critical assumption in the generalized linear mixed model (3), namely that, for given x, the selection of units must be independent of the outcome. The generation of units in the point process model is governed by the total intensity, and the intensity ratios determine the response distribution, so the corresponding assumption is that $\lambda_{\bullet}(x)$ be independent of the ratios $\lambda_r(x)/\lambda_s(x)$. The paper recognizes this possibility, but does not insist on it because the sampling scheme also matters. The independence assumption is guaranteed in certain settings such as agricultural field trials where the sample of plots is fixed, or laboratory experiments where the units do not participate in selection. It is indefensible in settings such as the classification of bitmap images of handwritten decimal digits, where the set of classes is not exhaustive because images of non-decimal characters are excluded.

For volunteer samples, the independence assumption is both fragile and unconvincing. Consider two patients having equal covariate values, one a volunteer the other a refuser. VanderWeele argues that the risks for Alzheimer's disease must be equal simply because the outcome will not be known for several years. My immediate reaction is quite the reverse. I would be surprised if the odds ratio were as high as 2 or as low as $\frac{1}{2}$, but I would not be surprised if it were in the range $0.7 < \psi < 1.4$. In principle, the decision to enrol could be associated with a personality trait that is a more effective predictor of the response than any other baseline measurement. Meng's example is more complicated but has many of the same characteristics. In this regard, patient enrolment seems fundamentally different from the selection of fibres by left end point in a traditional textile yarn. Regardless of the details specific to Alzheimer's disease, it seems unwise to use a statistical model founded on an assumption that is unnecessary in a randomized trial for the estimation of treatment effects, and unverifiable from data collected solely on participants.

There can be no logical objection to the assumption that the ratio $\lambda_0(x)/\lambda_1(x)$ be independent of the sum. But there is a logical objection to the assumption that the ratio be independent of $\lambda_0(x) + \psi\,\lambda_1(x)$ for more than one value of ψ, as required in the Alzheimer's example. Diggle's remark that *the assumption [of independence] is probably unrealistic in many applications* is absolutely correct, but also a substantial understatement.

Interpertation of conditional distributions
I am grateful to Professor Lauritzen and Professor Richardson for their interpretation and graphical representations of the various sampling schemes described in Section 3. After a little effort the relation between graphs 3(b)–3(d) and the various sampling plans becomes clearer, but the reason for the edge $Y \to S$ in 3(a) remains obscure. An alternative suggestion is to include event times $T \to X \to Y$, with a single arrow $T \to S$ representing samples of fixed size.

The graphs represent the entire process, but probabilities of interest refer to the observation $X[S]$, $Y[S]$, and presumably (20)–(22) refer to these. I had tried to understand the connection between conditioning by stratification and Pearl's concept of conditioning by intervention. For a while, I suspected that they might be the same, but I was unable to understand the latter sufficiently well to be sure. I am still not confident that I understand the subtleties of the distinction, but I am willing to believe that they are different.

Models, likelihoods and estimates
In my lengthy collaboration with John Nelder on generalized linear models, I recall no differences of opinion regarding the notion of a statistical model or the definition of likelihood. It is puzzling that a little correlation should lead to such divergent views. The accommodation of interunit correlation is a major task in *model construction*, but in my view correlation plays no role in *model definition*. A regression model is a parametric family of distributions such that, for each sample of units having covariate configuration $\mathbf{x}$, the response distributions $p_{\mathbf{x}}(\mathbf{y};\theta)$ satisfy the no-interference condition. As always, the likelihood ratio is $p_{\mathbf{x}}(\mathbf{y};\theta)/p_{\mathbf{x}}(\mathbf{y};\theta')$. I respect Nelder and Lee's firmly held opposing view based on Bjørnstad (1996), but I do not understand it.

Estimation and inference include parameter estimation and parametric inference, activities that take place within the parameter space under the auspices of the likelihood function and likelihood principle. But much inferential activity takes place in the observation space where the likelihood function is unknown and the likelihood principle irrelevant. Examples include the prediction of future values in the sense of the conditional distribution given the data, and the estimation of random variables such as infinite stratum averages in a random-effects model. The body of computational techniques associated with h-likelihood looks superficially similar to penalized likelihood (Wahba, 1985, 1990; Efron, 2001), which computes the conditional expected value for subsequent units by a smoothing algorithm (McCullagh, 2005). There is also a close affinity with Henderson's (1975) scheme for computing the so-called best linear unbiased predictor estimate. Each of these operations has a clear interpretation as the conditional expectation of a certain random variable in a Gaussian process given the observed data. My guess is that h-likelihood estimates have a similar approximate interpretation for many non-Gaussian random-effects processes.

Case–control and cohort designs
The evolving population model has an artificial temporal component introduced solely to label the events in a definite order, and not to be confused with time measured from enrolment in a cohort study. In the limited space available, it is not possible adequately to address the implications for cohort studies or case–control designs (Rathouz and Keogh). However, consider a cohort survival study in which the hazard of failure at time t for individual i is $\exp(\beta x_i)\ \lambda_i(t)\ \nu(\mathrm{d}t)$, the individual frailties $\lambda_i(\cdot) \sim \lambda_j(\cdot)$ being exchangeable but otherwise arbitrary. Given that a failure occurs among those in the risk set at time t, the probability that it is i who fails is the ratio of expected hazards

$$\frac{E\{\lambda_i(t)\exp(\beta x_i)\}}{\sum E\{\lambda_j(t)\exp(\beta x_j)\}} = \frac{\exp(\beta x_i)}{\sum \exp(\beta x_j)},$$

not the expected value of the ratio. As a result, the apparently weaker assumption of exchangeability of frailties is not distinguishable from equality of frailties.

The *sample* in a cohort study may be selected at one moment in time, but the *population* is quite a different matter. A cohort study cannot be worth the cost and effort unless the conclusions are judged to have implications for subsequent generations. Although not necessarily infinite, the population must include multiple cohorts: it is not fixed at a single moment in time as Professor Lee suggests.

Brief remarks
In practice, more complicated versions of the Cox process may be needed to model temporal dependence as described by Møller, or cluster size dependence as described by Kuk. The accommodation of overdispersion by a multiplicative adjustment as described by Ross is soundly established and entirely sensible.

References in the discussion

Bjørnstad, J. F. (1996) On the generalization of the likelihood function and the likelihood principle. *J. Am. Statist. Ass.*, **91**, 791–806.

Borgan, O., Goldstein, L. and Langholz, B. (1995) Methods for the analysis of sampled cohort data in the Cox Proportional Hazards model. *Ann. Statist.*, **23**, 1749–1778.

Daley, D. J. and Vere-Jones, D. (2003) *An Introduction to the Theory of Point Processes*, vol. I, *Elementary Theory and Methods*, 2nd edn. New York: Springer.

Diggle, P. J., Menezes, R. and Su, T.-L. (2008) Geostatistical inference under preferential sampling. *Working Paper 162*. Department of Biostatistics, Johns Hopkins University, Baltimore.

Efron, B. (2001) Selection criteria for scatterplot smoothers. *Ann. Statist.*, **29**, 470–504.

Farewell, V. T. (1979) Some results on the estimation of logistic models based on retrospective data. *Biometrika*, **66**, 27–32.

Finney, D. J. (1952) *Probit Analysis*, 2nd edn. Cambridge: Cambridge University Press.

Fisher, R. A. and Yates, F. (1974) *Statistical Tables for Biological, Agricultural and Medical Research*, 6th edn. London: Longman.

Gill, R. D., van der Laan, M. J. and Robins, J. M. (1997) Coarsening at random, characterizations, conjectures and counter-examples. *Lect. Notes Statist.*, **123**, 255–294.

Goldstein, L. and Langholz, B. (1992) Asymptotic theory for nested case-control sampling in the Cox regression model. *Ann. Statist.*, **20**, 1903–1928.

Greenland, S. (2003) Quantifying biases in causal models: classical confounding versus collider-stratification bias. *Epidemiology*, **14**, 300–306.

Grünwald, P. and Halpern, J. Y. (2003) Updating probabilities. *J. Artif. Intell. Res.*, **19**, 243–278.

Heitjan, D. F. and Rubin, D. B. (1991) Ignorability and coarse data. *Ann. Statist.*, **19**, 2244–2253.

Henderson, C. R. (1975) Best linear unbiased estimation and prediction under a selection model. *Biometrics*, **31**, 423–447.

Hernán, M. A., Hernández-Diaz, S. and Robins, J. M. (2004) A structural approach to selection bias. *Epidemiology*, **15**, 615–625.

Holland, P. W. (1986) Statistics and causal inference. *J. Am. Statist. Ass.*, **81**, 945–970.

Jacobsen, M. and Keiding, N. (1995) Coarsening at random in general sample spaces and random censoring in continuous time. *Ann. Statist.*, **23**, 774–786.

Jaeger, M. (2005) Ignorability of categorical data. *Ann. Statist.*, **33**, 1964–1981.

Lee, Y. and Nelder, J. A. (2004) Conditional and marginal models: another view (with discussion). *Statist. Sci.*, **19**, 219–238.

Lee, Y. and Nelder, J. A. (2006) Double hierarchical generalized linear models (with discussion). *Appl. Statist.*, **55**, 139–185.

Lee, Y., Nelder, J. A. and Pawitan, Y. (2006) *Generalized Linear Models with Random Effects*. London: Chapman and Hall.

Lin, H., Scharfstein, D. O. and Rosenheck, R. A. (2004) Analysis of longitudinal data with irregular, outcome-dependent follow-up. *J. R. Statist. Soc.* B, **66**, 791–813.

Lipsitz, S. R., Fitzmaurice, G. M., Ibrahim, J. G., Gelber, R. and Lipshultz, S. (2002) Parameter estimation in longitudinal studies with outcome-dependent follow-up. *Biometrics*, **58**, 621–630.

McCullagh, P. (2005) Exchangeability and regression models. In *Celebrating Statistics* (eds A. C. Davison, Y. Dodge and N. Wermuth), pp. 89–113. Oxford: Oxford University Press.

McCullagh, P. and Nelder, J. A. (1983) *Generalized Linear Models*. London: Chapman and Hall.

Møller, J. (2003) Shot noise Cox processes. *Adv. Appl. Probab.*, **35**, 614–640.

Møller, J. and Waagepetersen, R. P. (2004) *Statistical Inference and Simulation for Spatial Point Processes*. Boca Raton: Chapman and Hall–CRC.

Møller, J. and Waagepetersen, R. P. (2007) Modern spatial point process modelling and inference (with discussion). *Scand. J. Statist.*, **34**, 643–711.

Mosteller, F. (1965) *Fifty Challenging Problems in Probability with Solutions*. Reading: Addison-Wesley.

Pearl, J. (1993) Graphical models, causality and intervention. *Statist. Sci.*, **8**, 266–269.

Pearl, J. (1995) Causal diagrams for empirical research (with discussion). *Biometrika*, **82**, 669–710.

Pearl, J. (2000) *Causality: Models, Reasoning, and Inference*. Cambridge: Cambridge University Press.

Prentice, R. L. (1986) A case-cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrika*, **73**, 1–11.

Prentice, R. L. and Pyke, R. (1979) Logistic disease incidence models and case-control studies. *Biometrika*, **66**, 403–411.

Rothman, K. J. (1986) *Modern Epidemiology*. Boston: Little, Brown.

Rubin, D. B. (1976) Inference and missing data. *Biometrika*, **63**, 581–592.

Rubin, D. B. (2005) Causal inference using potential outcomes: design, modeling, decisions. *J. Am. Statist. Ass.*, **100**, 322–331.

Schall, R. (1991) Estimation in generalized linear models with random coefficients. *Biometrika*, **78**, 719–727.

Shafer, G. (1985) Conditional probability. *Int. Statist. Rev.*, **53**, 261–277.

Spirtes, P., Glymour, C. and Scheines, R. (1993) *Causation, Prediction and Search*. New York: Springer.

Stiratelli, R., Laird, N. M. and Ware, J. (1984) Random-effects statistical model for serial observations with binary response. *Biometrics*, **40**, 961–971.

Sun, J., Sun, L. and Liu, D. (2007) Regression analysis of longitudinal data in the presence of informative observation and censoring times. *J. Am. Statist. Ass.*, **102**, 1397–1406.

Wahba, G. (1985) A comparison of GCV and GML for choosing the smoothing parameter in the generalized spline smoothing problem. *Ann. Statist.*, **13**, 1378–1402.

Wahba, G. (1990) *Spline Models for Observational Data*. Philadelphia: Society for Industrial and Applied Mathematics.

Wilsdon, B. H. (1934) Discrimination by specification statistically considered and illustrated by the standard specification for Portland cement (with discussion). *J. R. Statist. Soc.*, suppl., **1**, 152–206.