Sampling bias and logistic models

Peter McCullagh *

November 2007

Abstract

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 $\mathbf{x}$. This 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.

Some key words: Auto-generated 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

*Support for this research was provided by NSF Grant DMS-0305009

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, for example

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

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 woolen or cotton yarn, or the problem of estimating the distribution of speeds of highway vehicles. Individual fibres are clearly unlabelled, so it is necessary to select a random sample, which might well 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, for example, 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, 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 pre-determined 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 among 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, Nelder and Pawitan, 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

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

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 $\eta(x(u)), \eta(x(u'))$, and the responses $Y(u), Y(u')$ 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

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

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

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

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 (3), but no loss of generality. In the econometrics literature, (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 (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, § 2.4; 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 (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 pre-determined, but are generated by a random process. Such units, whether they be patients, highway vehicles or purchase events, are called *auto-generated*; the non-mathematical term *self-selected* is too anthropomorphic for general use. Without careful verification, we should not expect the conditional distribution $p_n(\mathbf{y} \,|\, \mathbf{x})$ for auto-generated units to coincide with $p_{\mathbf{x}}(\mathbf{y})$ for a pre-determined 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 $p_n(\mathbf{y} \mid \mathbf{x}) = p_{\mathbf{x}}(\mathbf{y})$ 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 (3) or (4) can be estimated in the standard way 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.

1. 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 (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.

2. Class aggregation: Suppose that two response classes r, s in (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.

3. Class restriction: Suppose that the number of classes in (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 (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.

4. 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 (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 artifact 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 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 $\tilde{\nu}(dx) = \nu(dx)/\nu(\mathcal{X})$ 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 $\lambda_\bullet(x) = \sum \lambda_r(x)$ for the total intensity at x. A Poisson process in $\mathcal{C} \times \mathcal{X} \times (0, \infty)$ evolves at a constant temporal rate $\lambda(r, x)\nu(dx)\, dt$. These events constitute the target population, which is random, infinite and unlabelled.

Let $\mathbf{Z}_t$ be the set of events occurring prior to 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 coordinate, $t(z)$ is the temporal coordinate, and $y(z)$ is the class. Given λ, the number of events in $\mathbf{Z}_t$ is Poisson with mean $t \int_{\mathcal{X}} \lambda_\bullet(x)\, \nu(dx)$, proportional to t and finite. The number of events in $\mathbf{Z}$ is infinite, and the set of points $\{x(z) \colon z \in \mathbf{Z}\}$ 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 coordinate and $y(j)$ is the type or class. The ordered event times $0 \equiv t_0 \leq t_1 \leq t_2 \leq \cdots$ are distinct with probability one. 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

$$E\left(\Lambda_\bullet e^{-\Lambda_\bullet (t_j - t_{j-1})}\, dt_j \times \frac{\lambda_\bullet(x_j)\nu(dx_j)}{\Lambda_\bullet} \times \frac{\lambda_{y_j}(x_j)}{\lambda_\bullet(x_j)}\right)$$

averaged over the intensity function λ. The total intensity $\Lambda_\bullet = \int \lambda_\bullet(x)\, \nu(dx)$ 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 selected 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 pre-determined.

Quota sample. Let $\mathbf{x} = (x_1, \ldots, x_n)$ be a given ordered set of n points in $\mathcal{X}$, and let dx_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 $x(\varphi_j) \in dx_j$. The easiest way to select such a sample is to partition the population by covariate values $\mathbf{Z}_{dx} = \{(x, y, t) \in \mathbf{Z} : x \in dx\}$. Each stratum is infinite and temporally ordered. Define φ_j to be the index of the first event in $\mathbf{Z}_{dx_j}$.

Distributions are computed for the limit in which each interval dx_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

$\Lambda_\bullet(dx_j)$, so $t(\varphi_j)$ tends to infinity. The distribution of the class labels is

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

The conditional distribution is independent of ν, and coincides with $p_{\mathbf{x}}(\mathbf{y})$ in (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 $p_n(\mathbf{y} \,|\, \mathbf{x})$ in (6) coincides with $p_{\mathbf{x}}(\mathbf{y})$ in (4).

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 $\Lambda_\bullet = \int \lambda_\bullet(x)\nu(dx)$. The conditional joint density of the sampled time points is thus $\Lambda_\bullet^n e^{-\Lambda_\bullet t_n}$ for $0 \le t_1 \le \cdots \le t_n$. The components of $x\varphi$ are conditionally independent and identically distributed with density $\lambda_\bullet(x)\nu(dx)/\Lambda_\bullet$, 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

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

so the unconditional joint density is

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

averaged with respect to the distribution of λ.

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

$$p_n(\mathbf{x}, \mathbf{t})\, d\mathbf{x}\, d\mathbf{t} = E\Big(\exp(-\Lambda_\bullet t_n) \prod_{i=1}^{n} \lambda_\bullet(x_i)\,\nu(dx_i)\, dt_i\Big)$$

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

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

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

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

In either case the conditional distribution is a ratio of expected values, whereas (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 (4). Otherwise, sequential sampling is biased in the sense that $p(\mathbf{y} \,|\, \mathbf{x}) \neq p_{\mathbf{x}}(\mathbf{y})$.

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

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

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

$$p_t(\mathbf{x}, \mathbf{t})\, d\mathbf{x}\, d\mathbf{t} = E\Big(\exp(-t\Lambda_\bullet) \prod_{i=1}^n \lambda_\bullet(x_i)\, \nu(dx_i)\Big)\, d\mathbf{t}$$

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

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

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 (9).

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 $N \geq n$, and restricting attention to the finite subset $\mathbf{Z}_t \subset \mathbf{Z}$ where t is the first time that $\#\mathbf{Z}_t \geq N$. 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 protocol II. Consequently, unless the intensity ratio processes are independent of $\lambda_\bullet(\cdot)$, simple random sampling is biased.

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 (Baddley 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.

Case-control sample. Case-control sampling is essentially the same as weighted sampling, except that $k = 2$ and the quota sizes n_0, 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

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

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

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

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 $\Lambda_0 = \int \lambda_0(x)\, \nu(dx)$, and likewise for λ_1. Using this approximation, the conditional probability is proportional to

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

In essence, this means that the observation from a case-control design can be analyzed as if it were obtained from a prospective sequential sample or simple random sample as in II, III or IV above.

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 (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 (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 $\mu(x) = \int y\, p_x(dy)$ in stratum x might be mistaken for the conditional mean $\kappa(x) = \int y\, p(dy \,|\, x)$.

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 sub-model (4) is not closed under aggregation of classes, nor is the log Gaussian process 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 made on the same unit over time are understood to be correlated, and there may also be correlations among 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

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

derived in (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 the case because ν enters into the definition of $\Lambda_\bullet = \int \lambda_\bullet(x)\,\nu(dx)$. 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 $t \to 0$ for fixed ν, or the scalar $\nu(\mathcal{X}) \to 0$ for fixed t. The limiting low-intensity conditional distribution

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

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 process (Møller, Syversveen and Waagpetersen 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 exists a measure ν such that the nk ratios $\lambda_0(x)/\Lambda_\bullet, \ldots, \lambda_{k-1}(x)/\Lambda_\bullet$ for $x \in \mathbf{x}$ are jointly independent of $\Lambda_\bullet$. Then the numerator in (10) can be expressed as a product of expectations

$$\begin{aligned} E\Big(\exp(-t\Lambda_\bullet)\prod \lambda_{y_i}(x_i)\Big) &= E\Big(\Lambda_\bullet^n e^{-t\Lambda_\bullet}\Big) E\Big(\frac{\lambda_{y_1}(x_1)}{\Lambda_\bullet} \cdots \frac{\lambda_{y_n}(x_n)}{\Lambda_\bullet}\Big) \\ &= \frac{E(\Lambda_\bullet^n e^{-t\Lambda_\bullet})}{E(\Lambda_\bullet^n)} E\Big(\prod \lambda_{y_i}(x_i)\Big). \end{aligned}$$

The denominator can be factored in a similar way, so the ratio in (10) simplifies to (11). Note that 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 $\nu(dx) = dx/(2L)$ for $-L \le x \le L$, we find that $\Lambda_\bullet = T\hat{\Lambda}_\bullet = T + o(1)$ for large L, while the ratios $\lambda(x)/\Lambda_\bullet = \hat{\lambda}(x)/\hat{\Lambda}_\bullet$ 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 of the models described 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 (11) is the product

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

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 $\lambda_\bullet$. In other words, (11) is

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

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

$$E(\log \lambda(x)) = \mu(x), \quad \operatorname{cov}(\log \lambda(x), \log \lambda(x')) = K(x, x').$$

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 (11). Unfortunately, the sum of log Gaussian processes is not log Gaussian, so the normalizing constant is not available in closed form. The log of the conditional distribution (11) is given in an obvious notation by

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

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

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

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

$$\log \operatorname{odds}(Y' = 1 \mid \cdots) = \mu_1(x') + \sum_{x \in \mathbf{x}^{(1)}} K(x, x') - \sum_{x \in \mathbf{x}^{(0)}} K(x, x')$$

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, 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

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

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

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

If $K(x, x') \geq 0$ for all x, x', 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

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

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, for example by replacing $\lambda_r(x)$ with $e^{\alpha_r + \beta_r x}\lambda_r(x)$. Then the product density at $\mathbf{x}^{(r)}$ becomes $e^{n_r\alpha_r + \beta_r \mathbf{x}_\bullet^{(r)}} \mathrm{per}_{\alpha_r}(K[\mathbf{x}^{(r)}])$, 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 $\tau(x) = \gamma(1, T, x)\gamma(0, C, x)/(\gamma(0, T, x)\gamma(1, C, x))$, 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 (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 bi-partition model for response and treatment status. If m_0, 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 $v = s$.

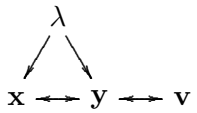
Suppose that $(\mathbf{y}, \mathbf{v}, \mathbf{x})$ has been observed in a sequential sample. The prognosis for the response y' for a subsequent event at x' depends on whether $v' = T$ or $v' = C$ as follows:

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

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 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

$$\begin{aligned} p_1(Y_1 = r \,|\, m \geq 1) &= \frac{E((1 - e^{-\lambda_\bullet})\lambda_r/\lambda_\bullet)}{E(1 - e^{-\lambda_\bullet})} \simeq \frac{E(\lambda_r)}{E(\lambda_\bullet)} \\ p_2(Y_1 = r, Y_2 = s \,|\, m \geq 2) &= \frac{E((1 - (1 + \lambda_\bullet)e^{-\lambda_\bullet})\lambda_r\lambda_s/\lambda_\bullet^2)}{E(1 - (1 + \lambda_\bullet)e^{-\lambda_\bullet})} \simeq \frac{E(\lambda_r\lambda_s)}{E(\lambda_\bullet^2)} \\ p_2(Y_1 = r \,|\, m \geq 2) &= \frac{E((1 - (1 + \lambda_\bullet)e^{-\lambda_\bullet})\lambda_r/\lambda_\bullet)}{E(1 - (1 + \lambda_\bullet)e^{-\lambda_\bullet})} \simeq \frac{E(\lambda_r\lambda_\bullet)}{E(\lambda_\bullet^2)} \\ &\neq p_1(Y_1 = r \,|\, m \geq 1). \end{aligned}$$

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 $m \geq 2$ 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 exhibiting interference, the other not. In the log Gaussian model

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

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 $(e^{-1/2}, e^{1/2})$, 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

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

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

6.2 Over-dispersion

Suppose that events are grouped by covariate value, so that $T_\bullet(x)$ is the observed number of events at x, and $T_r(x)$ is the number of those who belong

to class r. Over-dispersion 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

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

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:

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

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

7 Computation

7.1 Parameter estimation

Since $\mathbf{x}$ and $\mathbf{y}$ 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 $\nu(dx)$, which governs primarily the marginal distribution of $\mathbf{x}$. It appears that the marginal distribution of $\mathbf{x}$ must contain very little information about intensity ratios, so it is natural to use the conditional distribution given $\mathbf{x}$ 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 plan 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 $t\, m_y(x)\, \nu(dx)$ at (y, x). The expected number of events in the interval dx is $t\, m_\bullet(x)\nu(dx)$, and the expected number of events of class r in the same interval is $t\, m_r(x)\nu(dx)$. For a function $h\colon \mathcal{X} \to \mathcal{R}$, additive functionals have expectation

$$E\Big(\sum_{\mathbf{Z}_t} h(x(z))\Big) = t \int_{\mathcal{X}} h(x)\, m_\bullet(x)\, \nu(dx),$$
$$E\Big(\sum_{\mathbf{Z}_t} h(x(z))\, y_r(z)\Big) = t \int_{\mathcal{X}} h(x)\, m_r(x)\, \nu(dx),$$

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 $T_r = \sum_{\mathbf{Z}} h(x)\,(y_r(z) - \pi_r(x))$ 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 $\mathbf{x}$, 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 $h(x) = 1$.

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

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 log-normal models exhibit interference, but the homogeneous gamma model has the special property of zero second-order interference.

The first integral in (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, Liang and Albert 1988). If in fact the x-values are generated by the process itself, the estimating function $\sum_{\mathbf{Z}} h(x)(y_r(z) - \rho_r(x(z)))$ has expectation $\int_{\mathcal{X}} h(x)(\pi(x) - \rho(x))m_\bullet(x)\,\nu(dx)$, 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 $\text{logit}(\pi(x)) = \beta' x$, the parameters can be estimated consistently 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 described by Liang and Zeger (1986) can be followed, but variance calculations need to be modified to account for interference as in (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, s. The additive combination $\sum h(x, x')(y_r y_s' - \pi_{rs}(x, x'))$ 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

$$\log \text{pr}(Y(x') = r \mid \cdots) = \log m_r(x') + \sum_{x \in \mathbf{x}^{(r)}} K(x, x') + \text{const}$$

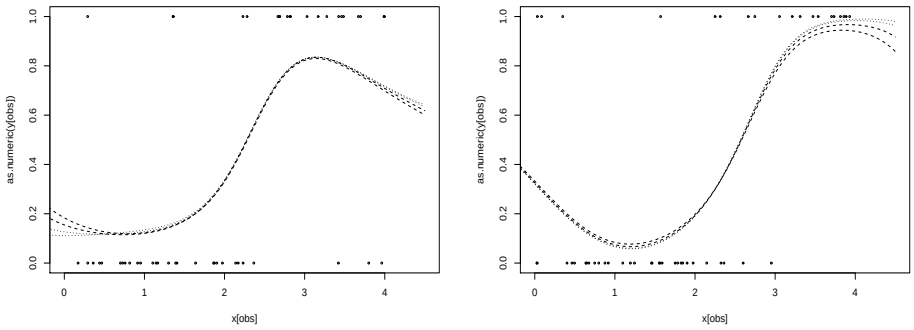


Figure 1: Prognostic probability computed for the homogeneous gamma model with $K(x, x') = \exp(-(x-x')^2)$ and four values of α from 0.05 to 0.5. The two sample configurations of 50 events are indicated by dots at $y = 0$ and $y = 1$.

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 $\gamma \geq 0$, 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,

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

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

8 Summary

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 (3). On account of the consistency condition (5), we can compute the conditional probability of any event such as

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

using the model distribution with an additional unit having $x(u') = x'$. Sample-space inferences of this sort are accessible directly from (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 (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 (2) as a partially Bayesian model, and (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 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, 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 authors have noted a close parallel between (2) and the one-dimensional marginal distributions associated with (3). Specifically, if η_u is a zero-mean Gaussian variable,

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

implies

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

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

In biostatistical terminology, the regression parameters α, β in (17) are called subject-specific or cluster-specific, while 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 expressions (17), (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 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) points 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) note that since each individual has her own latent risk, the model invites an unwarranted causal interpretation. Galbraith (1991) criticizes the interventionist interpretation of parameters in (17), and points 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 η_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 (18). Two units u, u' having the same baseline covariate values but different treatment levels have different response distributions. The treatment effect is the difference between these probabilities, usually measured on the log odds ratio scale. Although established terminology suggests otherwise, the treatment component of β^* in (18) is the treatment effect specific to this pair of units u, 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 $\mathrm{pr}(Y_i = r \,|\, x_i = x)$ depends on x, possibly 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 $\lambda_r(x)/\lambda_\bullet(x)$ for both (4) and the evolving population model. The expected market share is the stratum probability $\mathrm{pr}(Y_i = r \,|\, i\colon x_i = x)$, 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 $\mathrm{pr}(Y_i = r \,|\, x_i = x)$ 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 list of behaviours, crime types, birth defects or the ten decimal digits. This list 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 $E(Y_i \mid X_i = x)$ 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 don't 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 auto-generated 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. In order to make the hierarchical formulation compatible with the broader context of the evolving population model, it is necessary to interpret (3) and (4) as stratum distributions, not conditional distributions. Once the distinction is made, it is immediately apparent that the stratum distribution does not determine the conditional probability given $\mathbf{x}$ 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 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 $e^{\alpha+\beta x}\lambda_0(x)$, the correct version of (17) and (18) for an auto-generated unit $u \in \mathbf{Z}_t$ is

$$\begin{aligned}
\operatorname{logit} \operatorname{pr}(Y(u) = 1 \mid \lambda, u \in \mathbf{Z}_t) &= \log \lambda_1(x(u)) - \log \lambda_0(x(u)) \\
&= \alpha + \beta x(u) + \eta(x(u)), \\
\operatorname{logit} \operatorname{pr}(Y(u) = 1 \mid u \in \mathbf{Z}_t) &= \log m_1(x(u)) - \log m_0(x(u)) \\
&= \alpha + \beta x(u),
\end{aligned}$$

with no approximation and no attenuation. The distinction described in section (8.2) between two parameterizations is simply incorrect for auto-generated units.

The subject-specific approach takes aim at the right target parameter in (17), but the conventional likelihood or hierarchical Bayesian calculation leads to inconsistency when sampling bias is ignored in the steps leading to (3). Sample x-values occur preferentially at points where the total intensity $\lambda_{\bullet}(\cdot)$ is high, which is not the case for a predetermined $\mathbf{x}$. As a result, parameter estimates from (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 $E(Y \mid x) \neq \rho(x)$ for a random sample. But a fortuitous ambiguity of the conventional notation $E(Y \mid x)$ 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, β in (17) are exactly equal to the parameters α^*, 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.

9 Acknowledgement

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

References

[1] Baddley, A. and Jensen, E.B (2005) *Stereology for Statisticians.* Boca Raton, Chapman and Hall.

[2] Besag, J. (1974) Spatial interaction and the statistical analysis of lattice systems (with discussion). *J. Roy. Statist. Soc.* B 36, 192-236.

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

[4] Breslow, N.E. and Clayton, D.G. (1993) Approximate inference in generalized linear mixed models. J. Amer. Statist. Assoc., 88, 9-25.

[5] Cox, D.R. (1958) *Planning of Experiments* Wiley, New York.

[6] Cox, D.R. (1972) Regression models and life tables. J. Roy. Statist. Soc. B 34, 187-220.

[7] Cox, D.R. and Snell, E.J. (1979) On sampling and the estimation of rare errors. *biometrika* 66, 125-132.

[8] Galbraith, J.I. (1991) The interpretation of a regression coefficient. *Biometrics* 47, 1593-1596.

[9] Glonek, G.F.V. and McCullagh, P. (1995) Multivariate logistic models. *J. Roy. Statist. Soc.* B 57, 533-546.

[10] Green, P. and Silverman, B. (1994) Nonparametric regression and generalized linear models. Chapman and Hall, London.

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

[12] Laird, N. and Ware, J. (1982) Random effects models for longitudinal data. Biometrics 38, 963-974.

[13] Lee, Y. and Nelder, J.A. (1996) Hierarchical generalised linear models (with discussion). J. Roy. Statist. Soc. B, 58, 619-656.

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

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

[16] McCullagh, P. (1983) Quasi-likelihood functions. *Annals of Statistics* 11, 59-67.

[17] McCullagh, P. (2005) Exchangeability and regression models. In *Celebrating Statistics*, A.C. Davison, Y. Dodge and N. Wermuth, editors. 89-113.

[18] McCullagh, P. and Nelder, J.A. (1989) *Generalized Linear Models.* Chapman and Hall, London.

[19] McCullagh, P. and Yang, J. (2006) Stochastic classification models. Proc. International Congress of Mathematicians, 2006, vol. III, 669-686.

[20] McCullagh, P. and Møller, J. (2006) The permanental process. *Adv. Appl. Prob.* 38, 873-888.

[21] McCulloch, C.E. (1994) Maximum likelihood variance components estimation in binary data. J. Amer. Statist. Assoc., 89, 330-335.

[22] McCulloch, C.E. (1997) Maximum-likelihood algorithms for generalized linear mixed models. J. Amer. Statist. Assoc., 92, 162-170.

[23] Møller, J., Syversveen, A.R. and Waagpetersen, R.P. (1998) Log Gaussian Cox processes. *Scandinavian Journal of Statistics* 25, 451-482.

[24] 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.

[25] Schall, R. (1991) Estimation in generalized linear models with random effects. Biometrika, 78, 719-727.

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

[27] Shirai, T. and Takahashi, Y. (2003) Random point fields associated with certain Fredholm determinants I: Fermion, Poisson and boson point processes. *J. Functional Analysis* 205, 414-463.

[28] Wahba, G. (1985) A comparison of GCV and GML for choosing the smoothing parameter in the generalized spline smoothing problem. *Annals of Statistics* 13, 1378-1402.

[29] Wolfinger, R.W. (1993) Laplace's approximation for nonlinear mixed models. Biometrika, 80, 791-795.

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

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