

RSC115: Sampling bias and logistic models

By Peter McCullagh

### **Author's response to the discussants**

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 re-stating 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  $p_n(\mathbf{x}, \mathbf{y} \mid \mathbf{o} \equiv 1)$  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  $p_n(\mathbf{y} \mid \mathbf{x}, \mathbf{o} \equiv 1)$  and  $p_{\mathbf{o}=1}(\mathbf{y} \mid \mathbf{x})$  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 GLMM 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  $\lambda_{\cdot}(x)$  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 two or as low as one half, but I would not be surprised if it were in the range  $0.7 < \psi < 1.4$ . In principle, the decision to enroll 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 enrollment seems fundamentally different from the selection of fibres by left endpoint 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  $\psi$ , 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.

#### *Interpretation of conditional distributions*

I am grateful to Professors Lauritzen and Richardson for their interpretation and graphical representations of the various sampling schemes described in section 3. After a little effort the relation between graphs (b)–(d) and the various sampling plans becomes clearer, but the reason for the edge  $Y \rightarrow S$  in (a) remains obscure. An alternative suggestion is to include event times  $T \rightarrow X \rightarrow Y$ , with a single arrow  $T \rightarrow S$  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 (1)–(3) 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'm still not confident that I understand the subtleties of the distinction, but I'm 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 inter-unit 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 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 BLUP 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 enrollment 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  $\exp(\beta x_i)\lambda_i(t)\nu(dt)$ , 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

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

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. While 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 over-dispersion by a multiplicative adjustment as described by Ross is soundly established and entirely sensible.

**References**

- Bjørnstad, J.F. (1996) On the generalization of the likelihood function and the likelihood principle. *J. Am. Statist. Assoc.* **91**, 791–806.
- Efron, B. (2001) Selection criteria for scatterplot smoothers. *Ann. Statist.* **29**, 470–504.
- Henderson, C.R. (1975) Best linear unbiased estimation and prediction under a selection model. *Biometrics* **31**, 423–447.
- McCullagh, P. (2005) Exchangeability and regression models. In *Celebrating Statistics*. (Eds A.C. Davison, Y. Dodge and N. Wermuth), Oxford statistical Science Series, **3**, 89–113.
- 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*. SIAM Philadelphia.