Statistical models with autogenerated units

PETER MCCULLAGH, *University of Chicago*,
e-mail: pmcc@galton.uchicago.edu

Keywords: parameter attenuation; quota sample; random effects; sampling bias; stratum distribution; unlabelled unit

AMS keywords:

In the formal theory of stochastic processes, which includes conventional regression models, the index set of potential experimental or observational units is fixed, and usually infinite. The response distribution, which is specified in a consistent manner by the regression model $p_{\mathbf{x}}(\mathbf{y})$ for each fixed finite sample of units, depends on the sample configuration $\mathbf{x} = (x(u_1), \ldots, x(u_n))$ of covariate values. Although the definition of a fixed sample is unambiguous mathematically, the meaning is not at all clear in the majority of applications. Random samples of units are hard to avoid in biological work because the population units are typically unlabelled. Sequential recruitment of units is standard practice in clinical work, ecological studies and market research: labels affixed to the sample units after recruitment tend to obscure the sampling scheme. It is by no means obvious that standard models with distributions determined for *fixed* samples are suitable for applications in which the units are unlabelled and samples are random. I propose an alternative process in which the formal concept of a fixed set of statistical units is absent. Instead, the process itself generates a stream of 'units' in time, each unit being identified with its (x, y, t) value. Samples are automatically random because the units themselves are random. The relation between the conditional distribution $p(\mathbf{y} \,|\, \mathbf{x})$ for a sequential samples, and the stratum distribution $p_{\mathbf{x}}(\mathbf{y})$ for fixed quota $\mathbf{x}$, will be discussed in the context of random-effects models. This analysis reveals that that the phenomenon of parameter attenuation in logistic models is a statistical illusion caused by sampling bias.