

## Sampling bias in logistic models

*Peter McCullagh*

*University of Chicago*

This talk is concerned with regression models for the effect of covariates on correlated binary and correlated polytomous responses. In the conventional construction of generalized linear mixed models, correlations are induced by a random effect, additive on the logistic scale.

The joint distribution  $p_{bfx}(bfy)$  obtained by integration depends on the covariate values  $bfx$  on the sampled units. The thrust of this talk is that the conventional formulation may not be appropriate for natural sampling schemes in which the sampled units arise from a random process such as sequential recruitment of volunteers. The conventional analysis incorrectly predicts parameter attenuation due to the random effect, thereby giving a misleading impression of the magnitude of treatment effects.

The error in the GLMM analysis is a subtle consequence of selection bias that arises from random sampling of units. This talk will describe a non-standard but mathematically natural formulation in which auto-generated units are subsequently selected by an explicit sampling plan.

For a quota sample in which the  $bfx$ -configuration is pre-specified, the model distribution coincides with  $p_{bfx}(bfy)$  in the GLMM. However, if the covariate configuration is random, for example the values obtained by simple random sampling from the available population, the conditional distribution  $p(bfy \text{ given } bfx)$  is different from  $p_{bfx}(bfy)$ . By contrast with conventional models, conditioning on  $bfx$  is not equivalent to stratification by  $bfx$ . The implications for likelihood computations and estimating equations will be discussed.