Model Uncertainty and Incomplete Data

John Copas
University of Warwick, UK

Abstract

Some of the most challenging problems in statistics arise when data are incomplete and when we are uncertain about the model. The obvious example is missing data: we only observe part of the sample, and conventional models based on "missing at random" are almost certainly wrong (non-response may be correlated with the outcome of interest). Other equally important examples include: confounding in observational studies (lack of randomization leads to the possibility that treatment is correlated with unmeasured confounders), censoring (possible correlation between failure and censoring mechanisms); non-compliance (compliance may be correlated with outcome to treatment); publication bias (papers reporting significant results are more likely to be published); measurement error (errors in exposure may be correlated with response). A key difficulty in all of these areas is that mis-specification of an assumed model can bias the inference and yet be completely undetectable from the observed data.

These examples are all special cases of the general theory discussed in a recent paper read to the Royal Statistical Society (Copas and Eguchi, *Local Model Uncertainty and Incomplete Data Bias*, to appear in JRSSB 2005). We show how local asymptotic bias of maximum likelihood estimates depends on the "direction" as well as the "magnitude" of model mis-specification. We study "worst case" bias by maximizing over possible directions. Loss of efficiency in parameter estimation due to incompleteness in the data has a dual interpretation: the increase in variance when the assumed model is correct, the bias in estimation when the model is incorrect. Several special cases and examples are examined in detail: in some cases only a small degree of model mis-specification is needed to undermine the validity of standard methods of analysis.

These tutorial lectures will give an overview of the area, with particular reference to the Copas and Eguchi paper. Data from a meta-analysis of epidemiological studies into the health risks of passive smoking (inhaling other people's tobacco smoke) will be used to illustrate some of the ideas involved.

11.5.2005