Local model uncertainty and incomplete-data bias

John Copas
University of Warwick, Coventry, UK

and Shinto Eguchi
Institute of Statistical Mathematics, Tokyo, Japan

[Read before The Royal Statistical Society at a meeting organized by the Research Section on Wednesday, February 2nd, 2005, Professor J. T. Kent in the Chair]

Summary. Problems of the analysis of data with incomplete observations are all too familiar in statistics. They are doubly difficult if we are also uncertain about the choice of model. We propose a general formulation for the discussion of such problems and develop approximations to the resulting bias of maximum likelihood estimates on the assumption that model departures are small. Loss of efficiency in parameter estimation due to incompleteness in the data has a dual interpretation: the increase in variance when an assumed model is correct; the bias in estimation when the model is incorrect. Examples include non-ignorable missing data, hidden confounders in observational studies and publication bias in meta-analysis. Doubling variances before calculating confidence intervals or test statistics is suggested as a crude way of addressing the possibility of undetectably small departures from the model. The problem of assessing the risk of lung cancer from passive smoking is used as a motivating example.

Keywords: Coarsening; Hidden confounders; Ignorable data; Missing data; Missingness at random; Misspecified models; Publication bias; Selection bias

1. Introduction

Much of the theory and practice of statistics involves fitting parametric models to data. Most text-books, and much of the literature, assume that the model is known and that we obtain observations on the variables that are described by that model: we have model certainty and complete data. In reality this is almost never the case. Apart from simple sampling experiments like choosing balls from urns, there is no magical process telling us which model is correct. And, especially in large data sets, there are invariably complications in measurement, with observations missing, censored or corrupted in some way. In practice we have model uncertainty and incomplete data.

There is a long history of discussion of these problems at meetings of the Royal Statistical Society. Two of the most often quoted papers in the statistical literature were read to the Society in the 1970s and are about the analysis of incomplete data: Cox (1972) on censored data and Dempster et al. (1977) on the EM algorithm for computing maximum likelihood estimates (MLEs) in incomplete-data problems. Discussion papers on robustness, influence and model choice have addressed various aspects of model uncertainty. Draper (1995) addressed model uncertainty through Bayesian model averaging. Most recently, Greenland (2005) discusses problems of unidentifiable biases in epidemiological studies and raises many of the same issues that we go on to discuss here.

Address for correspondence: John Copas, Department of Statistics, University of Warwick, Coventry, CV4 7AL, UK.
E-mail: jbc@stats.warwick.ac.uk

© 2005 Royal Statistical Society

© 2005 Royal Statistical Society

1369–7412/05/67000

CRiSM Paper No. 05-1, www.warwick.ac.uk/go/crism
An example which illustrates some of these problems is the somewhat controversial assessment of the link between passive smoking and lung cancer. The report of the UK Government’s Scientific Committee on Tobacco and Health (Department of Health, 2001) concluded that environmental tobacco smoke was a cause of lung cancer and recommended that smoking in public places should be restricted on the grounds of public health. This advice has been widely heeded, although there remain many dissenting voices. The committee followed Hackshaw et al. (1997) in estimating the relative risk as 1.24 (95% confidence limits 1.13–1.36)—prolonged exposure of non-smokers to other people’s tobacco smoke increases the risk of lung cancer by 24%. This figure is based on a meta-analysis of 37 published epidemiological studies, each of which compares the risk of lung cancer in non-smokers according to whether the spouse of the subject does or does not smoke. But, as discussed by these and later researchers, this analysis is likely to be influenced by several sources of potential bias, suggesting that the figure of 24%, and the strength of the evidence that is claimed for a causal link, needs to be interpreted with considerable caution.

The problems with this analysis, as with many published meta analyses, include the following.

(a) **Publication bias**: the studies in the meta-analysis are those found in a literature review, not necessarily an unbiased sample of all the studies which have been done in the area. If, for example, studies showing a significant and positive effect are more likely to be published than those giving an inconclusive result, then an analysis that is based on the published studies alone will be biased upwards.

(b) **Confounding**: the risk of lung cancer among non-smokers may also depend on other factors which are themselves correlated with exposure to tobacco smoke. If, for example, a healthy (high fruit and vegetable) diet tends to protect against lung cancer, and diet in non-smoking households tends to be better than in households in which one or both partners smokes, then the apparent relative risk using smoking status alone will again be biased upwards.

(c) **Measurement error**: the response variable in a case–control study is the level of exposure, measured here by the reported smoking status of the subject’s spouse. But this is a very imperfect measure of actual exposure. If, for example, some of the spouses who claim to be non-smokers are in fact current or former smokers, then the relative risk based on observed exposure will be biased downwards.

Hackshaw et al. (1997) claimed that publication bias is not a problem in their analysis but gave extended discussions of the other two sources of potential bias. They concluded that, as far as it is possible to estimate from other sources, these biases will roughly cancel, leaving the original crude estimate of 1.24 as their best estimate of relative risk. Their arguments for ignoring these biases have since been vigorously challenged (Copas and Shi, 2000a, b; Nilsson, 2001).

Although these causes of bias are very different, they all come under our general theme of incomplete data and model uncertainty. We could, at least in principle, correct for these biases if we had further data available. For publication bias, we would need data on unpublished as well as published studies; for confounding bias we would need values of all possible confounders for each subject; for measurement error on exposure we would need an unbiased measure of actual exposure for each subject. The problem arises when these complete data are not available. To make progress with the incomplete data we must rely on a model which asserts that these biases are, in some sense, ignorable. For publication bias, standard methods of meta-analysis assume that the probability that a study is published may depend on ancillary quantities such as sample size, but not on the observed outcome of the trial. For confounders, we allow these to be correlated with the observed outcome but not with the main exposure or treatment variable (after
conditioning on available covariates). For measurement error, standard models condition on observed covariates and assume that the measurement errors are subsumed within the residual variation of the response.

Model uncertainty is important because these models make assumptions which cannot be tested with the available data. Ignorability assumptions are typically made as a matter of expediency rather than through any conviction that they are true. In general, it is difficult to make any useful inferences at all if models are grossly misspecified, but to study sensitivity to small departures from the model is a useful first step. Our aim is to suggest a rather general asymptotic setting for exploring the link between local model uncertainty, defined in an appropriate way, and the bias in likelihood inference. Our set-up includes the above, and other, special cases as discussed in detail in later sections.

Our general notation for complete data $z$ and incomplete data $y$ is set out in Section 2. This is similar to Heitjan and Rubin’s (1991) idea of ‘coarsened’ data, although we use a rather different and simpler formulation. A parametric model $f_Z = f_Z(z; \theta)$ specifies the distribution of $z$, but the observable likelihood is based on the derived distribution $f_Y = f_Y(y; \theta)$ of $y$. Some examples are discussed in Section 3.

Model misspecification is discussed in Section 4. We assume that inference is based on model $f_Z$ (with corresponding model $f_Y$), but that $z$ is in fact generated by a ‘nearby’ distribution $g_Z$ (with corresponding distribution $g_Y$). We follow Copas and Eguchi (2001) by expressing $g_Z$ in terms of $f_Z$ and their log-likelihood ratio, although our formulation here is both simpler and more general than in Copas and Eguchi (2001). Taking a geometric view, we see incomplete-data bias as the difference between the projection of true model $g_Y$ onto assumed model $f_Y$, and the corresponding projection which we would be able to make with complete data, in which $g_Z$ is projected onto assumed model $f_Z$. Some special cases are discussed in Section 5, generalizing the results of Copas and Eguchi (2001) for univariate missing data problems.

In ordinary statistical problems where we can observe data on $z$, we can make the useful distinction between $f_Z$ as a ‘true’ model, one that we are willing to assume is correct, and $f_Z$ as a ‘working’ model, one that we use because it gives a good description of the data. In our formulation, a true model means that $f_Z = g_Z$; a working model means that, roughly, the ‘distance’ between $f_Z$ and $g_Z$ is of the same order of magnitude as the sampling error in estimates of parameter $\theta$. Incomplete-data problems are much more difficult, because there may be assumptions within model $f_Z$ which cannot be assessed from data on $y$ alone. Thus $f_Z$ may be grossly misspecified (leading to a large bias) and yet appear to give a good fit to the available data. We argue that, in practice, the model uncertainty when we use $f_Y$ as a model for incomplete data $y$ should be at least as much as the uncertainty that we would have about $f_Z$ if we were to use it as a working model given the luxury of being able to observe the complete data $z$. This argument gives a lower bound to the size of the incomplete-data bias, as explained in Section 6. As the direction of this bias is unidentified, we treat it as a source of extra uncertainty in inference about $\theta$. We do this by widening the conventional confidence interval for $\theta$ by a factor which turns out to be a rather simple function of the amount of information that is lost in the transformation of $z$ into $y$. We show that this factor is bounded above by $\sqrt{2}$, i.e. ‘double the variance’.

In Section 7 we return to the problem of confounding in the passive smoking and lung cancer study, again using the data in Hackshaw et al. (1997) as an example. Estimates of relative risk are seen to be highly sensitive to assumptions about ignorability. We use this example to illustrate some of the underlying ideas of the paper, presenting this section in such a way that it can be read virtually independently of the more technical material in the earlier sections.

The paper concludes with some comments in Section 8, and a technical appendix giving the proof of the ‘double-the-variance’ result of Section 6.
2. Complete and incomplete data

Suppose that random variable $z$ has distribution $z \sim f_Z(z; \theta)$, indexed by unknown parameter $\theta$. Usually both $z$ and $\theta$ will be vectors. The presentation will assume that $z$ is continuous so $f_Z$ is a probability density function, but we take for granted the obvious changes in notation if some or all of the components of $z$ are discrete. We shall assume throughout that $f_Z$ satisfies the regularity conditions that are necessary for the asymptotic properties of maximum likelihood estimates to apply in the usual way.

Define $s_Z(z; \theta) = \partial \{ \log(f_Z) \}/\partial \theta$ and $I_Z = E[-\partial^2 \{ \log(f_Z) \}/\partial \theta \partial \theta^T]$ to be the score vector and information matrix of $f_Z$ respectively. Then, given a large sample $z_1, z_2, \ldots, z_n$, the MLE $\hat{\theta}_Z$ is given by $\sum_i s_Z(z_i; \theta) = 0$ and standard asymptotic theory gives

$$\hat{\theta}_Z \simeq \theta + \frac{1}{n} I_Z^{-1} \sum_i s_Z(z_i; \theta),$$

$$\text{var}(\hat{\theta}_Z) \simeq (n I_Z)^{-1}. \quad (1)$$

We shall refer to observations on $z$ as the complete data, the observations that we would ideally like to have for inference about $\theta$.

Sometimes we cannot observe $z$ directly but can observe only the derived random variable $y = h(z)$, for some given function $h$. Or, there may be other reasons for wishing to base inference on the marginal likelihood using $y$ rather than the full likelihood using $z$. If $h$ is a one-to-one smooth function then this is just a data transformation, but if $h$ is many to one then there will in general be a loss of information. We shall then refer to observations on $y$ as the incomplete data. Most forms of incomplete data can be written in this way by a suitable choice of $z$ and $h$. Note that $y$ is not necessarily a numerical scalar or vector in the usual sense, it may contain ranges or subsets in some or all of its components.

The nature of the function $h$ is given by its level sets $(y) = \{ z : h(z) = y \}$. Fig. 1 illustrates these level sets for six simple examples where $z$ has two components, $z = (z_1, z_2)$. Fig. 1(a) shows single points, so $h$ is 1:1 with $(y) = h^{-1}(y)$. Fig. 1(b) is for missing data, where $z_1$ is the response of interest and $z_2$ is the missing data indicator. If $z_2 = 1$ we observe $z_1$, but if $z_2 = 0$ then $z_1$ is missing so all that we can say about $z_1$ is that $-\infty < z_1 < \infty$. In Fig. 1(c), $z_1$ is the response of interest, but it can only be observed with measurement error. Here $z_2$ is the measurement error so what we observe is the sum $z_1 + z_2$.

Next, in Fig. 1(d), we have data which may be rounded. If $z_2 = 1$ we observe $z_1$ exactly, but if $z_2 = 0$ then $z_1$ is rounded to the nearest integer. Fig. 1(e) represents two competing risks, where we observe the minimum of two (potential) cause-specific lifetimes. If $z$ is in the upper octant, we observe $z_1$ but $z_2$ is censored so all we know is that $z_2 > z_1$, and the other way round if $z$ is in the lower octant. Finally, Fig. 1(f) illustrates the case of a hidden variable: we observe $z_1$ but not $z_2$.

These examples suggest that the components of vector $z$ will typically consist of the main response variables of interest plus subsidiary variables that are involved in the process of observation. For Fig. 1(b)–1(f), conventional models for $f_Z$ may assume that $z_1$ and $z_2$ are independent. Misspecifications of these models would then allow $(z_1, z_2)$ to be correlated, or the measurement process non-ignorable. Figs 1(b), 1(f) and 1(c) are simplified versions of the problems of publication bias, hidden confounders and measurement error which were mentioned in Section 1 in connection with the passive smoking study.

This set-up is quite similar to that of the ‘coarse data’ model of Heitjan and Rubin (1991), who envisaged a function that is analogous to $h$ above, but with two arguments, $z$ plus a stochastic ‘coarsening variable’. Equivalently we could think of $h$ as a stochastic observation equation
converting \( z \) into an observable \( y \). Here we subsume the coarsening variable of the observational process into the vector \( z \), so that \( h \) can be thought of as deterministic. Although this simplifies the relationship between \( y \) and \( z \), a disadvantage of our notation is that it can obscure the distinction between components of \( \theta \) of interest, and components of \( \theta \) which are simply nuisance parameters of the coarsening process.

The model \( f_Z(z; \theta) \) for \( z \) induces the corresponding model \( f_Y(y; \theta) \) for \( y \). In a somewhat informal notation,

\[
y \sim f_Y(y; \theta) = \int_{(y)} f_Z(z; \theta) \, dz,
\]

where \((y)\) on the integration sign means integration with respect to \( z \) over the level set that is given by \( y = h(z) \). If, for example, \((y)\) fixes one component of \( z \) but leaves the others as ranges, then the integral returns the value that is fixed, with the appropriate Jacobian for any transformation of that component, and integrates over the ranges of the other components. See Jacobsen and Keiding (1995) for a rigorous formulation of integrals of this kind.

The score function for \( f_Y \) is now

\[
s_Y(y; \theta) = \frac{\partial}{\partial \theta} \log \{ f_Y(y; \theta) \} = E \{ s_Z(z; \theta) | y \},
\]

where \( s_Z(z; \theta) \) is the score function of \( f_Z(z; \theta) \).
the conditional expectation of \( s_Z(z; \theta) \) over the level set \((y)\). If \( \hat{\theta}_Y \) is the MLE based on incomplete data \( y_1, y_2, \ldots, y_n \),

\[
\hat{\theta}_Y \approx \theta + \frac{1}{n} I_Y^{-1} \sum s(y; \theta),
\]

where \( I_Y \) is the corresponding information matrix for \( f_Y \).

A consequence of equation (3) is that \( I_Z - I_Y \) is non-negative definite, so the eigenvalues of the matrix

\[
\Lambda = I_Y^{-1/2} I_Z^{-1} I_Y^{-1/2}
\]

are between 0 and 1. Matrix \( \Lambda \) reflects the proportion of information that is retained when we observe \( y \) instead of \( z \), and its eigenvalues \( \lambda_i \) can be thought of as relative efficiencies in estimating contrasts of \( \theta \) using \( \hat{\theta}_Y \) instead of \( \hat{\theta}_Z \). In particular, for estimating \( \phi = d^T \theta \),

\[
\frac{\lambda_{\min}}{\text{var}(\hat{\phi}_Z)} \leq \frac{\text{var}(\hat{\phi}_Y)}{\text{var}(\hat{\phi}_Y)} \leq \lambda_{\max}.
\]

3. Examples

3.1. Missing data

To start with the simplest case, suppose that we are trying to sample a scalar random variable \( t \), but that some observations are missing. Let \( r \) be the response indicator, which is equal to 1 if \( t \) is observed and equal to 0 if \( t \) is missing. Then \( z = (t, r) \), and

\[
y = h(t, r) = (t^{(r)}, r),
\]

where

\[
t^{(r)} = \begin{cases} t & \text{if } r = 1, \\ \infty & \text{if } r = 0. \end{cases}
\]

We use the symbol \( \infty \) here to mean that when \( r = 0 \) all we know is that \( t \) takes some value in \((-\infty, \infty)\). This is the set-up that was noted earlier in Fig. 1(b).

The simplest model here is the data missing completely at random (MCAR) model, which asserts that

\[
f_Z(z; \theta, \psi) = f_T(t; \theta) \psi^r (1 - \psi)^{1-r},
\]

where \( f_T \) is the marginal density of \( t \) and \( 1 - \psi \) is the probability that an observation is missing. We assume that parameters \( \theta \) and \( \psi \) are functionally independent. This model asserts that \( t \) and
Model Uncertainty and Incomplete-data Bias

7

r are independent, so the likelihood for \( \theta \) is the usual likelihood for the observed cases only, without the need to know the value of \( \psi \).

If \( s_T(t; \theta) \) is the score function for the main model \( f_T(t; \theta) \), the score functions for \( (\theta, \psi) \) are

\[
\begin{align*}
  s_Z &= \{s_T(t; \theta), s_R(r; \psi)\}, \\
  s_Y &= \{r s_T(t; \theta), s_R(r; \psi)\}
\end{align*}
\]

where \( s_R(r; \psi) = (r - \psi)/\psi(1 - \psi) \). The matrix \( \Lambda \) is diagonal, with diagonal entries equal to \( \psi \) for the \( \theta \)-components and 1 for the \( \psi \)-component. As expected, the relative efficiency of estimating \( \theta \) just reflects the reduction in the sample size that is caused by the missing data, and there is no loss of efficiency in estimating \( \psi \) because \( r \) is always observed.

More generally, let \( z = (t, r) \) where \( t = (t_1, t_2, \ldots, t_m) \) is a vector of \( m \) measurements and \( r \) is the corresponding vector of response indicators \( r = (r_1, r_2, \ldots, r_m) \), with \( r_i = 1 \) if \( t_i \) is observed and \( r_i = 0 \) if \( t_i \) is missing. Now \( y = h(t, r) = (t^{(1)}, r) \) with \( t^{(1)} = (t_1^{(1)}, t_2^{(1)}, \ldots, t_m^{(1)}) \), each component defined as in expression (6). Then the data missing at random (MAR) model is

\[
f_Z(t, r; \theta, \psi) = f_T(t; \theta) f_{R|T}(r, t; \psi)
\]

where \( f_{R|T}(r, t; \psi) \) is the conditional probability distribution of \( r \) given \( t \), which is assumed to depend on \( t \) only through the value of \( t^{(1)} \). This is the crucial data MAR assumption, that for any given \( r \) the value of \( f_{R|T}(r, t; \psi) \) depends on \( t \) only through those components \( t_i \) for which \( r_i = 1 \). Under data MAR, the missing data mechanism is not allowed to depend on the values of any unobserved components of \( t \). Lu and Copas (2004) showed that if \( f_T \) is a complete distribution family then data MAR is a necessary and sufficient condition for the missing data process to be ignorable, in the sense that the likelihood function for \( \theta \) can be constructed directly from the marginal distributions of the subsets of the \( t_i \)'s that are observed, without the need to know \( f_{R|T} \) explicitly.

Several special cases of \( f_{R|T}(r, t; \psi) \) are of interest. Data MAR are MCAR if \( f_{R|T}(r, t; \psi) \) depends on \( r \) but does not depend at all on \( t \). If \( t_1 \) are strata variables in a sample design, \( r_1 = 1 \) (always observed). Case non-response is when the \( r_i \)'s are all the same (1 for a responder; 0 for a refusal). The simplest set-up in meta-analysis has \( m = 2 \), \( t_1 \) the result that is reported in a typical study (e.g. estimated relative risk), \( t_2 \) the within-study estimated standard error of \( t_1 \), \( r_1 = r_2 = 1 \) if the study is published and \( r_1 = r_2 = 0 \) otherwise. Standard methods of meta-analysis ignore publication bias by assuming that \( f_{R|T} \) depends on \( t_2 \) but not \( t_1 \).

There is a very large literature on missing data problems, and extensive discussions of the data MAR assumption. Excellent texts include Little and Rubin (2002) and Schafer (1997), with many references therein. A good general introduction to statistical methods for meta-analysis is Sutton et al. (2000a).

3.2. Potential confounders

Here we envisage an observational study in which we wish to assess the dependence of a response \( t \) on a treatment or exposure variable \( x \). Suppose that \( t \) is also influenced by a hidden variable \( c \). Then, if \( c \) is independent of \( x, c \) is ignorable in the sense that it just contributes to the residual variation of \( t \) given \( x \). But, if \( c \) is associated with \( x \) as well as \( t \), then it is a potential confounder.

If we could observe \( z = (t, x, c) \) then, at least in principle, we could disentangle the influences of \( x \) and \( c \) on \( t \), but in practice we can observe only \( y = h(t, x, c) = (t, x) \).

Using Pearl's notation of the ‘do’ operator (Pearl, 2000), the causal effect of \( x \) on \( t \) can be described by
\[ p\{t|\text{do}(x)\} = \int p(t|x, c) \, p(c) \, dc. \]  

(10)

This is the distribution which we would obtain if the treatments \( x \) had been chosen at random, so the distribution of \( c \) is the same within each level of \( x \). However, the incomplete data \( y = (t, x) \) only give us information on

\[ p(t|x) = \int p(t|x, c) \, p(c|x) \, dc. \]  

(11)

Equations (10) and (11) are the same if and only if \( c \) and \( x \) are independent. This is our basic model, that we analyse observations on \( y \) as if they had arisen from a randomized experiment. The influence of \( c \) as a potential confounder will then be discussed as a misspecification from this model in Section 5.2.

We write our parametric model for \( y \) as

\[ f_Y(t, x; \theta, \psi) = f_{T|X}(t|x; \theta) \, f_X(x; \psi). \]  

(12)

Our modelling assumption is that \( f_Y \) is the marginal distribution of \( t|x \) from the distribution of \( (t, x, c) \) given by a complete-data model \( f_Z \) in which \( x \) and \( c \) are independent, namely

\[ f_Z(t, x, c; \psi, \beta, \gamma) = f_{T|XC}(t|x, c; \theta, \beta) \, f_X(x; \psi) \, f_C(c; \gamma). \]  

(13)

Here, \( f_Z \) may also involve other parameters, \((\beta, \gamma)\), but parameterized in such a way that \( \beta \) and \( \gamma \) disappear from equation (13) when we integrate out \( c \). The \( \theta \)-components of the score functions \( s_Y \) and \( s_Z \) come directly from the conditional distributions in equations (12) and (13).

In the special case of a linear model, suppose that \( t|x, c \sim N(\psi_1 + \theta^T x + \beta c, \psi_2^2) \) and \( c \sim N(0, \gamma^2) \).

(14)

Then the observable response distribution is just the ordinary regression model

\[ t|x \sim N(\psi_1 + \theta^T x, \psi_2^2). \]

Assume that the components of \( x \) are centred so that, under \( f_X \), \( x \) has mean 0 and covariance variance \( \Psi \). Under this set-up, the \( \theta \)-components of the score functions \( s_Z \) and \( s_Y \) are

\[(t - \psi_1 - \theta^T x - \beta c)x \]
\[
\frac{1}{\psi_2^2 - \beta^2 \gamma^2},
\]

\[ \frac{(t - \psi_1 - \theta^T x)x}{\psi_2^2} \]

respectively, and the \( \theta \)-submatrix of \( \Lambda \) is

\[
\left( \frac{1}{\psi_2^2} \right) = \{1 - \text{corr}(t, c|x)^2\} I.
\]

Under the ignorable data model, the proportion of information that is lost through not observing the hidden variable \( c \) is just the square of the partial correlation between \( t \) and \( c \) given \( x \).

3.3. Other examples

Fig. 1 illustrated some other examples of the general set-up, which we mention here but do not develop in detail.
Following Fig. 1(e), let \( t \) and \( v \) represent two competing risks, \( t \) a lifetime of interest and \( v \) the time of (potential) censoring. What we observe in censored survival analysis is the minimum of \( t \) and \( v \), and whether the event that is observed at this time was an actual failure or a censored observation. Thus \( z = (t, v) \), and \( y = h(t, v) \) is \( \{t, [t, \infty)\} \) if \( t \leq v \) and \( \{[v, \infty), v\} \) if \( t > v \). Standard models (ignorable censoring) assume that \( t \) and \( v \) are independent, so that \( f_Z = f_T(t; \theta) \times f_Y(v; \psi) \). If we define \( s_T \) to be the score vector for \( f_T \), the \( \theta \)-component of \( s_T \) is \( s_T(t; \theta) \) if \( t \leq v \) and \( \partial[\log\{S_T(v; \theta)\}] / \partial \theta \) if \( t > v \), where \( S_T(t; \theta) \) is the survival function for \( f_T(t; \theta) \).

The simplest example is when \( t \) and \( v \) are exponential survival distributions with rate parameters \( \theta \) and \( \psi \) respectively. We find that the \( \theta \)-submatrix of \( \Lambda \) is the scalar \( \theta / (\theta + \psi) \). In this special case (but not generally) the proportion of information that is lost through the censoring is just equal to the proportion of observations which are censored.

For an example with measurement error, following Fig. 1(c), suppose that we are interested in the linear regression of response \( y \) on covariates \( x \). But we can only observe \( t \) indirectly through the sum \( t + v \), where \( v \) is the measurement error. Here \( z = (t, v, x) \) and \( y = (t + v, x) \). Model \( f_Z \) is made up of the factors \( t| x \sim N(\theta^T x, \psi^2 - \beta^2), v \sim N(0, \beta^2) \) and \( x \sim f_X \). Model \( f_Y \) replaces the first two factors of \( f_Z \) by \( (t + v)| x \sim N(\theta^T x, \psi^2) \). In this model, the measurement error is pure random error and is ignorable in the sense that regression coefficient \( \theta \) can be estimated from the observable distribution \( f_Y \). The proportion of information about \( \theta \) that is retained in the reduction of \( z \) to \( y \) is just the squared correlation between \( t \) and \( t + v \), or \( 1 - \beta^2 / \psi^2 \). The non-ignorable case is when \( f_Z \) is misspecified so that there may be a dependence between \( v \) and \( t \) or \( x \) or both. In that case, the estimate of \( \theta \) from data on \( y \) may be biased.

4. Misspecified models

In practice we can never be sure that these or any other parametric statistical models are correct. Even if we could observe \( z \) there will be uncertainty about \( f_z \), and hence even more uncertainty about \( f_Y \) when we can only observe \( y \). Of particular interest will be uncertainty about the ignorability assumptions that are implied in models such as those of Section 3. In this section we develop a rather general asymptotic theory for local misspecifications of \( f_Z \) and \( f_Y \). By ‘local’, we mean model departures of a magnitude which could not easily be detected empirically from samples of the complete data \( z \).

To formulate distributions in a local neighbourhood of \( f_Z \), let \( u_Z(z; \theta) \) be any scalar function of \( z \) and \( \theta \), standardized to have mean 0 and variance 1 under the model \( f_Z \). Then for small values of \( \epsilon \)

\[
g_Z = g_Z(z; \theta, \epsilon, u_Z) = f_Z(z; \theta) \exp\{\epsilon u_Z(z; \theta)\} \tag{16}
\]

is non-negative and integrates to 1 up to and including first-order terms in \( \epsilon \), and so identifies a distribution in the neighbourhood of \( f_Z \). Essentially, any distribution in this close neighbourhood can be represented by equation (16) with an appropriate choice of \( \epsilon \) and \( u_Z \). Our assumption is that the actual distribution generating \( z \) is a member of this family for some small value of \( \epsilon \). The family of distributions (16) is a much more general version of the local misspecification family of Copas and Eguchi (2001).

If \( \epsilon = 0 \) then \( g_Z = f_Z \). Intuitively, \( \epsilon \) can be thought of as the ‘magnitude’ of misspecification and \( u_Z \) can be thought of as the ‘direction’ of misspecification. In fact the squared misspecification magnitude \( \epsilon^2 \) is, in its leading term, just half of the Kullback–Leibler divergence between \( f_Z \) and \( g_Z \). Geometrically, we can think of the model \( f_Z(z; \theta) \) as belonging to a curve in distribution space, different points on the curve corresponding to different values of \( \theta \). Then, if we fix \( \epsilon \) and
imagine $\theta$ and $u_Z$ ranging over all possibilities, $g_Z$ will cover all distributions within a ‘tubular neighbourhood’ of ‘radius’ $\varepsilon$ about this curve.

Using the same (informal) notation as in equation (2), the distribution of $y = h(z)$ that is induced by $g_Z$ is

$$
g_Y = g_Y(y; \theta, \varepsilon, u_Z) = \int f_z(z; \theta) \exp\{\varepsilon u_Z(z; \theta)\} \, dz
\simeq f_Y(y; \theta) \exp\{\varepsilon u_Y(y; \theta)\},
$$

where

$$
u_Y(y; \theta) = Ef\{u_Z(z; \theta)\}.
$$

These and later approximations are correct to first-order terms in $\varepsilon$.

If we fit the model $f_Z(z; \theta)$ to a random sample of $n$ observations from $g_Z$, the limiting value of the MLE $\hat{\theta}_Z$ as $n \to \infty$ is

$$
\theta_{gZ} = \arg\theta\left[ Ef\{s_Z(z; \theta)\} = 0 \right]
\simeq \theta + \varepsilon I_Z^{-1} Ef\{u_Z(z; \theta) s_Z(z; \theta)\},
$$

in the sense of almost sure convergence. This follows from expression (16), noting that $Ef\{s_Z(z; \theta)\} = 0$. Similarly, if we are sampling from $g_Y$, the limiting value of $\hat{\theta}_Y$ is

$$
\theta_{gY} = \arg\theta\left[ Ef\{s_Y(y; \theta)\} = 0 \right]
\simeq \theta + \varepsilon I_Y^{-1} Ef\{u_Y(y; \theta) s_Y(y; \theta)\}.
$$

The important point to note is that when $\varepsilon \neq 0$ these are not the same. We define the first-order approximation to the difference $\theta_{gY} - \theta_{gZ}$ to be the incomplete-data bias $b_{\theta}$, which is given by

$$
\theta_{gY} - \theta_{gZ} \simeq b_{\theta} = \varepsilon \left[E_{f}\{u_Z(z; \theta)\} \{I_Y^{-1} s_Y(y; \theta) - I_Z^{-1} s_Z(z; \theta)\}\right].
$$

In Fig. 2 we illustrate geometrically what we are assuming about these distributions, and how the incomplete-data bias arises. For any given $\theta$ we can think of the misspecification quantity $\varepsilon u_Z(z; \theta)$ as the vector joining $g_Z(z; \theta)$ to $f_Z(z; \theta)$. This vector has ‘length’ $\varepsilon$ and ‘direction’ given

---

**Fig. 2.** Distributions of $z$ and $y$: (a) complete-data model; (b) incomplete data model
Model Uncertainty and Incomplete-data Bias

by the unit vector $u_Z(z; \theta)$. Fig. 2(a) is the orthogonal case when $E_f(s_Z u_Z) = 0$. This means that
$\theta$ in expression (16), and $\theta_{gZ}$ in expression (18), the value given by the projection from $g_Z$ onto
the line of the model, are the same. Fig. 2(b) is the corresponding diagram for the distributions
of $y$ rather than $z$. Now the vector $\varepsilon u_Y(y; \theta)$ that is defined in expression (17) is not orthogonal
to the line of the model, so the projection of $g_y$ onto $f_Y$ defines some other value of $\theta$. The
incomplete-data bias corresponds to the side of the right-angled triangle that is shown. We
have only attempted here a very superficial description of the geometrical aspects—concepts of
information geometry would be needed for deeper insights (Amari, 1985).

A natural standardized measure of the size of $b_\theta$ is the quadratic form that is defined with
respect to the incomplete-data information matrix $I_Y$. This gives

$$b_\theta^T I_Y b_\theta = \varepsilon^2 E_f \{u_Z(I_Y^{-1} s_Y - I_Z^{-1} s_Z)\}^T I_Y E_f \{u_Z(I_Y^{-1} s_Y - I_Z^{-1} s_Z)\}$$

$$\leq \varepsilon^2 E_f(u_Z^2) \| E_f \{I_Y^{-1/2} (I_Y^{-1} s_Y - I_Z^{-1} s_Z)(I_Y^{-1} s_Y - I_Z^{-1} s_Z)^T I_Y^{-1/2} \|$$

$$= \varepsilon^2 \| I - \Lambda \| = \varepsilon^2 (1 - \lambda_{\text{min}}),$$

where $\| \|$ denotes the largest eigenvalue. To derive this bound, inequality (22) follows from
equation (21) by using an elementary generalization of the Schwarz inequality in multivariate
analysis, that, for any scalar random variable $a$, any vector random variable $b$ with $V = E(bb^T)$,
and any conformable positive definite matrix $C$,

$$E(ab)^T C(E(ab)) \leq E(a^2) \| C^{1/2}VC^{1/2} \|.\]$$

The bound is attained when $a = \alpha^T b$, where $\alpha$ is the principal eigenvalue of the matrix $C^{1/2}VC^{1/2}$. Expression (23) follows from inequality (22) since

$$E_f(s_Y s_Z^T) = E_f \{s_Y E_f(s_Z^T)\} = E_f(s_Y s_Z^T) = I_Y. \quad (24)$$

The incomplete-data bias for estimating the scalar parameter $\phi = d^T \theta$ is $b_\phi = d^T b_\theta$. The corresponding inequality in the squared standardized bias is

$$\frac{b_\phi^2}{\text{var}_f(\phi_Y)} \leq \varepsilon^2 \left(1 - \frac{d^T I_Z^{-1} d}{d^T I_Y^{-1} d} \right)$$

$$\leq \varepsilon^2 (1 - \lambda_{\text{min}}). \quad (25)$$

Equality in expression (25) is attained when

$$u_Z(z; \theta) = \frac{d^T I_Y^{-1} s_Y \{h(z; \theta) - I_Z^{-1} s_Z(z; \theta)\}}{d^T (I_Y^{-1} - I_Z^{-1}) d^{1/2}}. \quad (27)$$

This misspecification direction $u_Z(z, \theta)$ is the ‘worst case’ as far as the incomplete-data bias in
$\phi_Y$ is concerned. Note that this $u_Z$ is orthogonal to the model, so $\theta_{gZ} = \theta$ to first order, which
is the case that we have illustrated in Fig. 2(a). This $u_Z$ is also the global worst case in the sense
of attaining the lower bounds in both inequality (22) and inequality (26) when $d$ is $I_Y^{-1/2}$ times the
eigenvector of $\Lambda$ with the smallest eigenvalue.

Inequalities (5) and (26) show how $\lambda_{\text{min}}$ plays a dual role in describing what happens when
we estimate a contrast from $y$ rather than from $z$. In inequality (5), $\lambda_{\text{min}}$ is the lower bound to
relative efficiency when the model is correct. In inequality (26), $1 - \lambda_{\text{min}}$ is the multiple of $\varepsilon^2$
which gives the upper bound to the standardized incomplete-data bias when the model is locally
misspecified.

By letting $u_Z$ be any standardized function of $z$, we are allowing for a very general class of
model misspecification. In most cases $f_Z$ will be made up of a number of submodels, with $\theta$
partitioned accordingly. It is then useful to represent \( u_Z \) as a corresponding sum of components. Through equation (20), this allows us to decompose the overall incomplete-data bias \( b_{\theta} \) into a sum of bias terms that are specific to misspecification of different parts of the model. Equivalently, additional constraints can be imposed on \( u_Z \) to focus attention on specific problems of interest, e.g. on bias caused by the failure of an ignorability assumption.

Using likelihood methods when sampling from a distribution which does not belong to the assumed model raises important questions which we have not discussed. We have assumed that \( \theta_Z \) and \( \theta_Y \) in expressions (18) and (19) are uniquely defined, and later we shall assume that the MLEs \( \hat{\theta}_Z \) and \( \hat{\theta}_Y \) are asymptotically normal. Using White’s (1982) more general discussion, Gustafson (2001) showed that these assumptions hold for local model misspecifications (in our notation, for sufficiently small \( \varepsilon \)) under quite weak regularity conditions (see Appendix A of Gustafson (2001)).

A more fundamental question is the interpretation of \( \theta \), since once we move outside a parametric model the concept of \( \theta \) as the ‘true value’ no longer has its usual meaning. Royall and Tsou (2003) distinguished between the ‘object of inference’, \( \theta_{\text{INF}} \) say, and the ‘object of interest’, \( \theta_{\text{INT}} \) say. The object of inference is the value of \( \theta \) for which the model is closest to the true distribution in the sense of Kullback–Liebler divergence. This corresponds to \( \theta_Z \) and \( \theta_Y \) defined above, for the complete- and incomplete-data models respectively. The object of interest, however, is a matter of the scientific objective of the study, and this may or may not be the same thing as the object of inference. For example, if \( \theta_{\text{INT}} \) is the mean of the population from which we are sampling, then \( \theta_{\text{INT}} = \theta_{\text{INF}} \) for the models \( N(\theta, \sigma^2) \) and Poisson(\( \theta \)), but not in general. Royall and Tsou (2003) argued that parametric inference about \( \theta \) is only meaningful when \( \theta_{\text{INF}} = \theta_{\text{INT}} \), and this is the assumption behind their idea of the robust adjusted likelihood function. With this assumption, \( b_{\theta} \) can be interpreted as the bias of \( \hat{\theta}_Y \) in the usual asymptotic sense, the difference between its expected value in large samples and the object of interest of the true distribution from which the complete data are sampled. In this setting, the difference between \( \theta_Z \) and \( \theta \) in expression (18) is just an artefact of the notation and is not a bias in any meaningful sense. We could, without loss of generality, reparameterize the model so that \( \theta = \theta_Z \) or, equivalently, assume that \( u_Z \) satisfies the orthogonality constraint \( E_f(uz) = 0 \) from the outset.

The misspecification function \( u_Z(z; \theta) \) needs to depend on \( \theta \) because of the constraint \( E_f(uz) = 0 \), which is the necessary condition for \( g_Z \) to integrate to 1 up to linear terms in \( \varepsilon \). However, the exact nature of this dependence is unimportant for the accuracy of the approximations that are studied here, as \( u_Z \) only enters our calculations through the first-order term \( \varepsilon u_Z \). To simplify the notation we can therefore write \( u_Z(z) \) instead of \( u_Z(z; \theta) \).

5. Examples continued

5.1. Missing data

To start with, return to the simple data MCAR model (7). The true model is now

\[
g_Z = f_T(t; \theta) \psi/(1 - \psi)^{-1} \exp\{\varepsilon u_Z(t, r)\}. \tag{28}
\]

If \( \varepsilon = 0 \) the model (28) is for data MCAR as before. If \( \varepsilon \neq 0 \), the function \( u_Z(t, r) \) allows the missing data process to be non-ignorable, in that \( P(r = 0|t) \) can now depend on the (unobserved) value of \( t \).

If \( I_T \) is the information matrix of the main model \( f_T(t; \theta) \), we find from expressions (8) and (20) that the \( \theta \)-component of the incomplete-data bias is

\[
\varepsilon(\psi I_T)^{-1} E_f \{(r - \psi)u_{ZST}\} = \varepsilon(1 - \psi) I_T^{-1} E_f \{u_D(t) s_T(t; \theta)\},
\]

CRiSM Paper No. 05-1, www.warwick.ac.uk/go/crism
where $u_D(t) = u_Z(t, 1) - u_Z(t, 0)$. The $\psi$-component of the incomplete-data bias is 0, as we have full information about $r$ under both $y$ and $z$.

The size of the quantity $\varepsilon u_D(t)$ indicates how much the actual missing data process differs from MCAR. It is easy to show that, for small $\varepsilon$,

$$s^2 \text{MCAR} = \text{var}_f \{ \varepsilon u_D(t) \}$$

This variance measures how much the probability of non-response varies across the population of different values of $t$. Under the assumption of data MCAR, the log-odds ratio in this expression is the same for all $t$ and so its variance is 0.

The standardized bias in equation (21) can now be written

$$\varepsilon^2 \psi (1 - \psi)^2 E_f (s_T u_D)^{-1} E_f (s_T u_D) \leq \varepsilon^2 \psi (1 - \psi)^2 \text{var}_f \{ u_D(t) \} \leq \psi (1 - \psi)^2 s^2 \text{MCAR}. \tag{29}$$

If $\theta$ is a scalar parameter, and we let $n_Y = n \psi$ be the expected actual sample size, and $\sigma_Y$ be the sample standard deviation of the incomplete-data MLE of $\theta$, then expression (29) reduces to

$$|b_0| \leq \sigma_Y (1 - \psi) n_Y^{1/2} \text{MCAR}. \tag{30}$$

This is equation (5) of Copas and Eguchi (2001). The maximum bias in inequality (30) is the product of four terms: $\sigma_Y$ (the first-order bias depends on the distribution of $t$ only through the standard error of the MLE), $1 - \psi$ (the proportion of observations that are missing), $\sqrt{n_Y}$ (bias becomes more important relative to the standard error the bigger is the actual sample size) and $s^2 \text{MCAR}$ (the standard deviation of the log-odds ratio measuring the contrast between data MCAR and the actual pattern of missing data). See Copas and Eguchi (2001) for further discussion of this formula and its generalizations.

For the more general data MAR model in equation (9), $u_Z$ is written as the sum of the two parts

$$u_T(t) = E_f \{ u_Z(t, r) \},$$

$$u_{R|T}(r, t) = u_Z(t, r) - u_T(t).$$

The misspecified model is then

$$g_Z = f_T(r; \theta) \exp \{ \varepsilon u_T(t) \} \ f_{R|T}(r, t; \psi) \ \exp \{ \varepsilon u_{R|T}(r, t) \}. \tag{20}$$

Note that $u_T(t)$, satisfying $E_f \{ u_T(t) \} = 0$, perturbs the complete-data distribution $f_T(t; \theta)$. The second part $u_{R|T}(r, t)$, satisfying $E_f \{ u_{R|T}(r, t) \} = 0$ for all $t$, is the perturbation on the conditional distribution of $r$ given $t$. The important point is that $f_{R|T}(r, t; \psi)$ satisfies the assumption of data MAR (for any given $r$, $f_{R|T}(r, t; \psi)$ depends on $t$ only through those $t_i$s with $r_i = 1$), whereas $u_{R|T}(r, t)$ can break the assumption of data MAR by allowing some dependence on the values of $t_i$ with $r_i = 0$.

Writing $u_Z = u_T + u_{R|T}$ in equation (20) shows that the incomplete-data bias also splits into two parts. If the main model $f_T$ is misspecified, then the fitted model involves a balance of discrepancies across the sample space, and this clearly changes if some some values of $t$ are more likely to be observed than others. This is the first component of bias. The second component of the bias is of particular interest in missing data problems since it describes the consequence of the missing data mechanism being non-ignorable. The $\psi$-elements of both components of bias are
0 (we observe \((r, t)\) under both complete and incomplete data) and so it is only the \(\theta\)-elements which are of interest. We focus on \(b^{\text{MAR}}\), the \(\theta\)-elements of this second bias component.

Under the data MAR model, both \(I_Z\) and \(I_Y\) are block diagonal, with the \(\psi\)-submatrix the same in each case. Let

\[
\begin{align*}
    s_T^2(t, \theta) &= E_f \{ s_T(t; \theta) | t \}, \\
    I_T^2 &= \text{var}_f \{ s^*(t, \theta) \}, \\
    \bar{u}_{R|T}(t) &= E_f \{ u_{R|T}(r, t) | r, t \}.
\end{align*}
\]

Then, from equation (20),

\[
\begin{align*}
    b^{\text{MAR}} &= \varepsilon [I_T^{\psi - 1} E_f \{ u_{R|T}(r, t) s_T^* (t \psi) \} - I_T^{\psi - 1} E_f \{ u_{R|T}(r, t) s_T(t) \}] \\
    &= \varepsilon I_T^{\psi - 1} E_f \{ \bar{u}_{R|T}(t \psi) s_T^* (t \psi) \} \\
    &= -\varepsilon I_T^{\psi - 1} E_f \{ \{ u_{R|T}(r, t) - \bar{u}_{R|T}(t \psi) \} \{ s_T(t) - s_T^* (t \psi) \} \}.
\end{align*}
\]

Thus, by analogy with inequality (22),

\[
\begin{align*}
    (b^{\text{MAR}})^T I_T^2 b^{\text{MAR}} &\leq \varepsilon^2 \left[ E_f \{ s_T^* (t \psi) \} E_f \{ s_T(t) \} \right] I_T^{\psi - 1} \left[ E_f \{ u_{R|T} - \bar{u}_{R|T} \} \right]^2 \\
    &= \frac{1 - \lambda_{\min}}{\lambda_{\min}} s_{\text{MAR}}^2,
\end{align*}
\]

where

\[
s_{\text{MAR}}^2 = \varepsilon^2 E_f \{ \text{var}_f \{ u_{R|T}(r, t) | r, t \} \}.
\]

and \(\lambda_{\min}\) is the smallest eigenvalue of \(I_T^{\psi - 1} I_T^2\).

As before, \(s_{\text{MAR}}\) can be given a statistical interpretation, contrasting the actual non-ignorable missing data mechanism under \(g_Z\) with the closest matching data MAR model. These are given respectively by

\[
\begin{align*}
    p_{g_Z}(r | t) &= f_{R|T}(r, t) \exp \{ \varepsilon u_{R|T}(r, t) \}, \\
    p_{g_Z}(r | t \psi) &= f_{R|T}(r, t) \exp \{ \varepsilon \bar{u}_{R|T}(t \psi) \}.
\end{align*}
\]

Thus

\[
\begin{align*}
    s_{\text{MAR}}^2 &\simeq E_f \left[ \text{var}_f \left\{ \log \left( \frac{p_{g_Z}(r | t)}{p_{g_Z}(r | t \psi)} \right) \right\} | r, t \psi \right].
\end{align*}
\]

In this and similar expressions it is unimportant, as far as first-order accuracy in \(\varepsilon\) is concerned, whether the \(E\) and \(\text{var}\) operators in expression (32) are with respect to \(f\) or \(g\).

To interpret expression (32), we imagine that we can calculate, for each possible \(r\) and \(t\), the logarithm of the ratio of the actual probability of \(r\) given \(t\) to the probability of that \(r\) given only the values of the \(t\)s for which \(r_1 = 1\). Let this log-ratio be \(L_R\) say. If the data MAR assumption is true then \(L_R = 0\) for all possible values of the unobserved \(t\)s for which \(r_1 = 0\). The conditional variance in expression (32) is the variance of \(L_R\) over these potential unobserved values; the larger this is the more at fault is the assumption of data MAR. Taking the expected value of this variance over all possible \(t\)s with \(r_1 = 1\), and then over all possible missing data patterns \(r\), gives \(s_{\text{MAR}}^2\) as an overall measure of non-ignorability.

In the special case of \(m = 1\) and the data MCAR model \(f_{R|T}(r, t) = \psi^r (1 - \psi)^{1-r}\), we have \(s_T^2 = s_T\) if \(r = 1\) and \(s_T^2 = 0\) if \(r = 0\). Also \(u_{R|T}(1, t) = (1 - \psi) u_D(t)\). Hence equation (31) reduces to expression (29). Another special case of interest is \(m = 2\) with \(f_{R|T}(0, 1, t) = f_{R|T} \{ (0, 0), t \} = 0\) for all \(t = (t_1, t_2)\). This is a regression model with covariate \(t_1\), always observed, and response \(t_2\).
subject to missing observations. Let \( p(t_1) = f_{R_T \{1,1\}, t} \). Then in this case

\[
\sigma_{\text{MAR}}^2 = E_f \left( \rho^2(t_1) \left[ 1 - p(t_1) \right] \text{var}_f \left[ \log \left( \frac{P_g(r_2 = 1|t_1, t_2)}{P_g(r_2 = 0|t_1, t_2)} \right) \right] \right).
\]

There are many other special cases of the general model which are of interest in particular applications. Horton and Fitzmaurice (2002), for example, discussed the analysis of a childhood psychopathology study where one of the variables (\( t_1 \) say, always measured) identifies the reasons why particular outcomes were not observed. Some causes may be assumed ignorable; others non-ignorable. In the notation above, this could be modelled by assuming that \( u_{R_T, T, r} = 0 \) for some values of \( t_1 \) (the ignorable cases) but allowing \( u_{R_T, T, r} \neq 0 \) for other values of \( t_1 \) (the non-ignorable cases). Horton and Fitzmaurice (2002) presented a sensitivity analysis for assessing their assumptions on the non-zero part of this function.

### 5.2. Potential confounders

Our second example, continuing Section 3.2, concerns the association between response \( t \) and treatment \( x \), in the presence of a hidden variable \( c \). Under the ‘randomization’ model (13), \( c \) is an intermediate variable which is assumed to be independent of \( x \), giving the observable marginal model (12). But, under the perturbed model \( g_{Z, x, c} \) are allowed to be dependent, in which case \( c \) is a potential confounder.

Firstly, note that the information matrix \( I_f \) from equation (12) is block diagonal with respect to the parameter partition \( (\theta, \psi) \). Similarly, \( I_Z \) from equation (13) is block diagonal for the parameter partition \( \{ \theta, \beta, \gamma, \psi \} \). Let \( s_{\theta X} \) be the score function for \( \theta \) from the regression factor in equation (12), and \( s_{\theta XC} \) be the score for \( \theta \) from the regression factor in equation (13). Then, for any overall misspecification function \( u_Z \), the \( \theta \) component of the incomplete-data bias in equation (20) is

\[
b = \epsilon \left( I_{\theta X}^{-1} s_{\theta X} - \left[ I_{\theta XC}^{-1} s_{\theta XC} \right] \right) u_Z,
\]

where \( \cdot \) denotes the \( \theta \)-components of the relevant vector.

It is more informative, however, to decompose \( u_Z \) into additive terms affecting the different factors in equation (13). Let

\[
\begin{align*}
    u_X(x) &= E_f \{ u_Z(t, x, c) | x \}, \\
    u_C(c) &= E_f \{ u_Z(t, x, c) | c \},
\end{align*}
\]

and define

\[
\begin{align*}
    u_{XC}(x, c) &= E_f \{ u_Z(t, x, c) | x, c \} - u_X(x) - u_C(c), \\
    u_{T|XC}(t, x, c) &= u_Z(t, x, c) - u_{XC}(x, c).
\end{align*}
\]

Then the true distribution \( g_Z \) is

\[
g_Z = f_{T|XC}(t, x, c) \exp(\epsilon u_{T|XC}) f_X(x) \exp(\epsilon u_X) f_C(c) \exp(\epsilon u_C) \exp(\epsilon u_{XC}).
\]

Only the last term here brings in an association between \( x \) and \( c \); in fact evaluating equations (10) and (11) from equation (34) we find

\[
\log \left[ \frac{P_g(t \mid x)}{P_g(t \mid \text{do}(x))} \right] \asymp \epsilon E_f \{ u_{XC}(x, c) | t, x \}.
\]
This confirms that only the component $u_{Xc}$ of $u_Z$ affects the role of $c$ as a confounder. The corresponding component of the bias in equation (33) is

$$b_{\text{RAN}} = \varepsilon I_{T|X}^{-1} E_f(s_{T|X}u_{Xc}),$$

since $E_f(s_{T|X}|x,c) = 0$ for all $x$ and $c$. We use the notation $b_{\text{RAN}}$ (which is analogous to $b_{\text{MAR}}$ in the previous section) to emphasize that this is the consequence in terms of bias of the fact that the study is not randomized. Then we find

$$(b_{\text{RAN}})^T I_{T|X} b_{\text{RAN}} \leq (1 - \lambda_{\text{min}}) s^2_{\text{RAN}},$$

where

$$s^2_{\text{RAN}} = \text{var}_f \left[ \log \left( \frac{p_y(c|x)}{p_y(c)} \right) \right].$$

The variance $s^2_{\text{RAN}}$, which is analogous to $s^2_{\text{MAR}}$ in expression (32), measures the strength of the association between $c$ and $x$.

In the special case of the linear model (14), we find from expression (15) that the $\theta$-component of the incomplete-data bias is approximately

$$b = \varepsilon \beta \Psi^{-1} E_f(cu_{Xc}) = \varepsilon \beta \Psi^{-1} E_f(cu_{Xc}) = b_{\text{RAN}}.$$ 

The incomplete-data bias in $\theta$ is only affected by the non-ignorability component, $u_{Xc}$, of $u_Z$; the contributions to the bias from the other three components all reduce to 0 in this case. The size of the standardized bias is now

$$b^T(\psi^{-2})b = \left( \frac{\varepsilon \beta}{\psi_2^2} \right)^2 (E_f(cu_{Xc}))^T \Psi^{-1} E_f(cu_{Xc}),$$

(35)

which is maximized over $u_{Xc}$ when $u_{Xc} = cd^T x$ for some constant vector $d$. This means that for small $\varepsilon$ the conditional distribution of $c$ given $x$ is approximately

$$g(c|x) \sim N(\varepsilon \gamma^2 d^T x, \gamma^2).$$

(36)

This is the worst case as far as bias in the estimation of the treatment effect is concerned. This corresponds to our intuition, that the most troublesome confounder is one which is linearly correlated with treatment.

In many applications $x$ will just be a scalar, in which case the correlation between $c$ and $x$ that is implied by distribution (36) is $\varepsilon \gamma d \Psi^{1/2}$. The size of the squared standardized bias is bounded by

$$b^2 = \text{corr}(t,c|x)^2 \text{corr}(x,c)^2.$$ 

(37)

The right-hand side of inequality (37) is the value of equation (35) in the scalar case when $u_{Xc} = cdx$ and $\text{var}(u_{Xc}) = 1$. The first term on the right-hand side of inequality (37) is the proportion of the variance of $t$ that is accounted for by $c$ over the influence of $x$, and so measures how much we lose by not measuring $c$. The second term is the dependence between the treatment and confounder that is caused by the lack of randomization, and so is a measure of non-ignorability of the design.
5.3. The Heckman model

We now compare our approach with one of the earliest and probably best-known systematic approaches to modelling ignorability problems; the Heckman model for selection bias. Heckman's original formulation (Heckman, 1979) has led to a very large literature, mostly in econometrics. The main idea is that we have two linear models, one for the response variable(s) of interest, and the other for the mechanism by which these responses are observed or selected. The residuals of the models are correlated, with correlation \( \rho \) say. Then if \( \rho = 0 \) the selection mechanism is ignorable but, if \( \rho \neq 0 \), inference which ignores the selection mechanism will be biased. See Copas and Li (1997) for an extended discussion of the Heckman model and some statistical applications.

When \( \rho \) is small, Heckman-type models are special cases of our more general formulation. For example, suppose that

\[
t = \theta^T x + \sigma \delta_1, \tag{38}
\]

\[
v = \psi^T x + \delta_2, \tag{39}
\]

where \( \delta_1 \) and \( \delta_2 \) are standard normal residuals, jointly normal with correlation \( \rho \). We observe \( x \) and we observe the sign of \( v \) (but not its actual value), but we only observe \( t \) if \( v \geq 0 \). We are back in the set-up of Section 5.1 for missing data, with \( r = 1 \) if \( v \geq 0 \) and \( r = 0 \) if \( v < 0 \). The model is data MAR if and only if \( \rho = 0 \). Note that, since \( r \) and \( x \) are always observed, \( \psi \) can be consistently estimated by probit analysis as

\[
P(r = 1|x) = \Phi(\psi^T x), \tag{40}
\]

where \( \Phi \) is the standard normal distribution function.

Given a sample \((t_i, x_i, r_i)\), the least squares estimate of \( \theta \) based on the observed cases (the cases with \( r_i = 1 \)) is

\[
\hat{\theta}_Y = \left( \sum r_i x_i x_i^T \right)^{-1} \sum r_i t_i x_i.
\]

But, using elementary properties of the bivariate normal distribution,

\[
E(t|x, r = 1) = \theta^T x + \rho \sigma \lambda(\psi^T x), \tag{41}
\]

where \( \lambda \) is Mill's ratio \( \lambda = \phi/\Phi \) and \( \phi \) is the standard normal density function. Also, for any function \( a(x) \) of \( x \),

\[
E\{a(x)|r = 1\} = \frac{E\{a(x) \Phi(\psi^T x)\}}{E\{\Phi(\psi^T x)\}}.
\]

Hence the asymptotic bias of \( \hat{\theta}_Y \) is

\[
b = \rho \sigma [E\{\Phi(\psi^T x) xx^T\}]^{-1} E\{\phi(\psi^T x) x\}. \tag{42}
\]

Let \( f_X(x) \) be the distribution of \( x \). Then, in the notation of Section 5.1, the distribution \( g_Z \) of \( z = (t, x, r) \) can be written, for small values of \( \rho \), as

\[
g_Z \simeq \frac{1}{\sigma} \phi \left( \frac{t - \theta^T x}{\sigma} \right) \Phi(\psi^T x)^T \Phi(-\psi^T x)^{1-r} f_X(x) \exp\{\rho u^*(t, x, r)\}, \tag{43}
\]

where

\[
u^*(t, x, r) = \frac{t - \theta^T x}{\sigma} \lambda(\psi^T x)^T \{-\lambda(-\psi^T x)\}^{1-r}.
\]
The form of this function follows from the first-order approximation

\[
P(r = 1 | t, x) = \Phi \left\{ \frac{\psi^T x + \rho (t - \theta^T x) / \sigma}{(1 - \rho^2)^{1/2}} \right\}
\]

\[
\simeq \Phi(\psi^T x) \left\{ 1 + \rho \frac{t - \theta^T x}{\sigma} \lambda(\psi^T x) \right\}.
\]

The variance of \( u^* \) is

\[
\kappa = \text{var} \left\{ u^* (t, x, r) \right\} = E_f \left\{ \lambda(\psi^T x) \lambda(-\psi^T x) \right\},
\]

and so, if we set

\[
u_Z (t, x, r) = \kappa^{-1} u^* (t, x, r),
\]

\[
e = \kappa \rho,
\]

then expression (43) is of the form (16). From equation (45) we have an interpretation of the misspecification quantity \( \varepsilon \) in terms of the correlation coefficient \( \rho \).

The \( \theta \)-components of the score and information matrices for the \( z \)- and \( y \)-versions of this model are given by

\[
s_T \mid X = \sigma^{-2}(t - \theta^T x)x,
\]

\[
s^* \mid X = \rho s_T \mid X,
\]

\[
I_T \mid X = \sigma^{-2} E(xx^T),
\]

\[
I^* \mid X = \sigma^{-2} E \left\{ xx^T \Phi(\psi^T x) \right\}.
\]

Hence

\[
E_f (u_Z s^*_T \mid X) = (\kappa \sigma)^{-1} E \left\{ x \phi(\psi^T x) \right\}
\]

and so the general asymptotic formula for the bias in equation (20) gives

\[
b = \varepsilon (I^*_T \mid X)^{-1} E_f (u_Z s^*_T \mid X)
\]

\[
= \rho \sigma \left\{ E \left\{ \Phi(\psi^T x)xx^T \right\} \right\}^{-1} E_f \left\{ \phi(\psi^T x)x \right\},
\]

which, to first order in \( \rho \), is the same as equation (42).

The standardized size of this bias, defined as in Section 4, is

\[
b^T I^*_T \mid X b = \kappa^{-2} \varepsilon^2 E_f \left\{ \phi(\psi^T x)x^T \right\} E_f \left\{ \Phi(\psi^T x)xx^T \right\}^{-1} E_f \left\{ \phi(\psi^T x)x \right\}.
\]

This can be compared with the maximum bias that is given by equation (23), which in this case is

\[
b^T I^*_T \mid X b \leq \varepsilon^2 \left[ 1 - \| E_f (xx^T) \|^{-1} E_f \left\{ \Phi(\psi^T x)xx^T \right\} \right].
\]

The simplest example of the Heckman model is when \( x \) is the scalar \( x = 1 \). Here, \( \theta \) is just the mean of \( t \sim N(\theta, \sigma^2) \) and \( \psi^T x \) is a constant. Then it is easy to check that expressions (46) and (47) are the same. Thus, in the problem of estimating the mean of a normal distribution with missing observations, the Heckman model is the worst case as far as bias is concerned. In general, however, the size of the bias that is given by the Heckman model is strictly less than the upper bound (47).
5.4. Publication bias

We mentioned publication bias in meta-analysis as an important example of incomplete-data analysis, which is particularly contentious in the case of the passive smoking example that was discussed in Section 1. The risk of bias arises because the studies that are available for analysis (the published studies) are not necessarily a random selection from the imagined population of all studies that have been done in the particular area of interest.

The simplest, and most common, setting is the meta-analysis of clinical trials in which a binary outcome is compared for patients given treatment and control. The standard approach is to assume a random-effects model based on the normal approximation of log-odds ratios from the resulting $2 \times 2$ tables (Sutton et al., 2000a). Each study gives an estimated log-odds ratio $t$, reported with (within-study) variance $\sigma^2_W$, and $\sigma^2_B$ is the between-study variance. It is usual to take these variances as known and to ignore the fact that in practice we use sample estimates. The meta-analysis assumes that, independently for each study,

$$t|x \sim N(\theta, \chi^2),$$

where $\chi^2 = \sigma^2_W + \sigma^2_B$ is the total variance and $\theta$ is the true value of the treatment effect that is to be estimated. Given observed pairs $(t, x)$ in the review, $\theta$ can be estimated by the sample weighted average of $t$ with weights inversely proportional to $x^2$.

Sutton et al. (2000b) reviewed the large literature on publication bias. One approach, going back to Lane and Dunlap (1978) and Hedges (1984), is to assume a weight function which we interpret as a selection probability

$$P(\text{select} | t, x) = w(t, x).$$

The estimate of $\theta$ is then found by maximizing the likelihood based on the conditional distribution of $(t, x)$ given that a study has been selected. The estimate clearly depends on the choice of weight function. For example, $w$ might be modelled to be an increasing function of the standardized size of effect $|t|/x$. Recognizing the arbitrary nature of such a choice, Greenhouse and Iyengar (1994) introduced an adjustable parameter into $w$ and reported a sensitivity analysis in which the bias is estimated for various values of this parameter.

If $f_X(x)$ is the distribution of $x$ across the population of all studies, the conditional joint distribution of $(t, x)$ given that the study is selected into the meta-analysis is, from expressions (48) and (49),

$$\frac{1}{p_X} \phi \left( \frac{t - \theta}{x} \right) f_X(x) w(t, x),$$

where $p = E\{w(t, x)\}$ is the marginal proportion of studies selected. The score function for $\theta$ from equation (48) is $(t - \theta)/\chi^2$, and so the bias $b$ in the estimation of $\theta$ from expression (50) is

$$b = \frac{E\{x^{-2}(t - \theta) w(t, x)\}}{E\{x^{-2} w(x)\}},$$

where $w(x) = E\{w(t, x) | x\}$ is the conditional probability of selection given $x$. Note that the size of $b$ depends on how strongly $w(t, x)$ depends on $t$. The bias is 0 if $w$ is a function of $x$ only, meaning that selection can depend on the 'size' of a study but not on its outcome.

If $w(t, x)$ depends only slightly on $t$, so that the publication bias $b$ is small, then this is just another special case of our general discussion. Define, for each study, the selection indicator $r$ to be 1 if the study is selected and to be 0 otherwise. Then the joint distribution of $(t, x, r)$ is
\[
\phi(t - \theta) f_X(x) w(t, x)^r \{1 - w(t, x)^{1-r}\}.
\]

Write
\[
w(t, x) = w(x) + \delta(t, x),
\]
so that \(E\{\delta(t, x)\mid x\} = 0\) for all \(x\). Then if \(\delta(t, x)\) is small (weak selection bias), expression (52) is approximately
\[
\frac{1}{x} \phi\left(\frac{t - \theta}{x}\right) f_X(x) w(x)^r \{1 - w(x)^{1-r}\} \exp\{u^*(t, x, r)\},
\]
where
\[
u^*(t, x, r) = \frac{\delta(t, x)\{r - w(x)\}}{w(x)\{1 - w(x)\}}.
\]

We have now written expression (52) in the general form (16), with \(z = (t, x, r)\). Note that the \(f_Z\)-model corresponding to expression (53) is not for data MAR, since \(P(r = 0 \mid x) = 1 - w(x)\) and \(x\) is not observed for unpublished studies.

In the notation analogous to that of Section 5.1, the incomplete-data score for \(\theta\) is
\[
s^*_T \mid X = r(t - \theta) / x^2
\]
and so the general formula for the bias in equation (20) gives
\[
I^*_T = E_f \left[ \frac{(t - \theta) \delta(t, x) r \{r - w(x)\}}{x^2 w(x) \{1 - w(x)\}} \right] \approx E_f \left\{ \frac{t - \theta}{x^2} w(t, x) \right\},
\]
but \(I^*_T = E_f \{w(x)/x^2\}\) and so expression (54) is exactly the same as equation (51). Since the distribution of \(x\) cannot be identified from the observed studies alone, a more useful form of expression (54) is
\[
b = E_{OBS}(x^{-2})^{-1} E_{OBS} \left[ \frac{E_f \{w(t, x) \mid x\}}{x^2 w(x)} \right],
\]
where the observable distribution \(p(x \mid r = 1) = f_{OBS}(x) = f_X(x) w(x)/p\) can now be identified with the empirical distribution of \(x\) over the selected studies.

Model uncertainty is a major issue here. Some researchers argue that so little is known about \(w(t, x)\) that a sensitivity analysis is the only sensible way forward. In a series of papers (Copas, 1999; Copas and Shi, 2000b, 2001; Shi and Copas, 2002) we used a Heckman model on the lines of Section 5.3, exploiting the analogy between \(w(t, x)\) and expression (44). Copas and Jackson (2004) avoided the approximation that \(\delta(t, x)\) is small by evaluating the upper bound for \(b\) over a wide class of possible weight functions. Others noted that, according to distribution (48), the conditional mean of \(t\) given \(x\) should not depend on \(x\), so any observed dependence and asymmetry in the scatterplot of \(t\) against \(x\) must be due to the multiplying factor \(w(t, x)\) in expression (50). This leads to the test for publication bias that was proposed by Egger et al. (1997) and the imputation method of Duval and Tweedie (2000).

6. Undetectable misspecification

6.1. Identifiability and tests of fit

Our local approximations have assumed that \(\varepsilon\) is small, but we have not discussed the size of misspecification that is needed for the bias approximations to be useful in the practical setting of inference from a sample of \(n\) observations. Standard asymptotic inference allows us to estimate \(\theta\) to within an accuracy of the order of magnitude \(O_p(n^{-1/2})\), and the local bias approximations
are of the order of magnitude \(O(\varepsilon)\). Thus our approximations allow us to combine these two sources of error in meaningful ways when \(\varepsilon = O(n^{-1/2})\). This size of \(\varepsilon\) means that the misspecification is ‘undetectable’ in the sense that empirical evidence for discriminating between \(f\) and \(g\) remains uncertain even when sample sizes are indefinitely large.

To see this, first consider the ideal situation in which we can observe a sample of the complete data \(z_1, z_2, \ldots, z_n\). If we knew \(\theta\) and the misspecification function \(u_Z\), then we could test \(g_Z\) in equation (16) against \(f_Z\), i.e. test the null hypothesis \(H_0: \varepsilon = 0\), with the uniformly most powerful standardized test statistic

\[
T_Z = n^{-1/2} \sum_{i=1}^{n} u_Z(z_i; \theta).
\]

(55)

Under the null hypothesis, \(T_Z\) is asymptotically standard normal. If, for significance level \(\alpha\), we reject hypothesis \(H_0\) when \(|T_Z| \geq d_\alpha = \Phi^{-1}(1 - \alpha/2)/\varepsilon\), the asymptotic power function is

\[
\Phi(-d_\alpha - n^{1/2} \varepsilon) + \Phi(-d_\alpha + n^{1/2} \varepsilon),
\]

(56)

since \(E_f(u_Z(z; \theta)) = \varepsilon + O(\varepsilon^2)\) from equation (16). With \(\varepsilon = O(n^{-1/2})\), the term \(n^{1/2} \varepsilon\) remains finite for large \(n\), and so the misspecification is undetectable in the sense that the power of the optimum test does not tend to 1 as \(n \to \infty\). Note that this argument is unaffected if we use \(\hat{\theta}_Z\) in place of \(\theta\) in equation (55), since

\[
n^{-1/2} \sum_{i=1}^{n} u_Z(z_i; \hat{\theta}_Z) \approx T_Z + \left(n^{1/2}(\hat{\theta}_Z - \theta)\right)^T E_f\left(\frac{\partial u_Z}{\partial \theta}\right).
\]

(57)

The last term vanishes, which is a consequence of the identity \(E_f(u_Z) = 0\) for all \(\theta\) which, on differentiating with respect to \(\theta\), yields \(E_f(\partial u_Z/\partial \theta) = -E_f(s_Z u_Z) = 0\), as discussed at the end of Section 4.

If we could sample \(z\), then we could use \(T_Z\) to test for misspecification in any given direction \(u_Z(z; \theta)\). Of particular interest would be the directions (27) which give maximum bias for estimating scalar contrasts of \(\theta\). However, the situation is quite different when we can only sample the incomplete data \(y_1, y_2, \ldots, y_n\), since there may be misspecifications in \(g_Z\) which cannot be detected from data on \(y\). If \(u_Z\) is such that \(u_\theta = d^T s_Y\) for some vector of constants \(d\), then

\[
g_Y \simeq f_Y(y; \theta)\left\{1 + \varepsilon d^T s_Y(y; \theta)\right\} \simeq f_Y(y; \theta + \varepsilon d)
\]

and so the misspecification is completely confounded with the unknown value of \(\theta\). An example of this happening is the simple pattern mixture model for missing data with \(z = (t, r)\) and \(t|r \sim N(\theta + r, 1)\). It is obvious that we have no information about the size of \(\varepsilon\) since we can only observe data on the conditional distribution of \(t\) given \(r = 1\). In this case we find \(d = 1\) since \(u_Y = s_Y = r(t - \theta)\).

To see that this can always happen, consider the analogue of \(T_Z\) for observations on \(y\), namely

\[
T_Y = n^{-1/2} \sum_{i=1}^{n} u_Y(y_i; \theta).
\]

When we make this test operational by replacing \(\theta\) by the incomplete-data estimate \(\hat{\theta}_Y\), the analogue of expression (57) is

\[
n^{-1/2} \sum_{i=1}^{n} u_Y(y_i; \hat{\theta}_Y) \approx n^{-1/2} \sum_{i=1}^{n} u_Y(y_i; \hat{\theta}) + \left\{n^{1/2}(\hat{\theta}_Y - \theta)\right\} E_f(u_Y s_Y) \\
\quad \approx n^{-1/2} \sum_{i=1}^{n} \left\{u_Y(y_i; \hat{\theta}) - (E_f(u_Y s_Y))\right\} I_Y^{-1} s_Y(y_i; \theta)\}.
\]

The term in braces is the sample residual when \(u_Y\) is projected onto the linear space that is spanned by the components of the score function \(s_Y\) and so is identically zero if \(u_Y = d^T s_Y\). But
this is exactly what happens in the worst case misspecification in equation (27), for then
\[ u_Y = E_f(u_Z|y) = \frac{d^T (I_y^{-1} - I_z^{-1}) s_y}{\{d^T (I_y^{-1} - I_z^{-1}) d\}^{1/2}}. \]

This is well defined provided that \( \lambda_{\text{max}} < 1 \). Referring to Fig. 2, the worst case for bias is when the triangle that is defined by the projection of \( g_T \) onto the model collapses onto a line along the model. The first-order approximation to \( \hat{\rho} \) conditional on the observed values of \( u_Z \) is unstable. If the test of fit that is based on \( \hat{\rho} \) fails because we are allowing misspecification to be in any direction, including the worst case function \( u_Z \) in equation (27). Here, in the notation of that section,
\[ \sum u_Y(t_i, x_i, r_i; \hat{\theta}_Y) = (\kappa \sigma)^{-1} \sum r_i(t_i - \hat{\theta}_Y^T x_i) \lambda(\psi^T x_i). \tag{58} \]

This is closely related to the standard method of fitting the Heckman model (Heckman, 1979), which is to estimate \( \psi \) from equation (40), to add the corresponding estimate of \( \lambda(\psi^T x) \) as an additional covariate to the linear regression of \( t \) on \( x \) and to refit by ordinary least squares. This is because, from equation (41), values of \( \hat{\rho} \) and \( x \) among those cases with \( r = 1 \) can be written
\[ t = \theta^T x + \rho \sigma \lambda(\psi^T x) + \delta^*, \tag{59} \]

where \( \delta^* \) is a random residual with mean 0. The (unweighted) least squares estimate of \( \sigma \rho \), the coefficient on the Mills ratio term in equation (59), gives
\[ \hat{\rho} = \frac{\sum r_i(t_i - \hat{\theta}_Y^T x_i) \lambda(\psi^T x_i)}{\sigma \sum r_i \{\lambda(\psi^T x_i) - \hat{\alpha}_Y^T x_i\}^2}, \tag{60} \]

where \( \hat{\alpha}_Y \) is the (unweighted) least squares coefficient in the observed regression of \( \lambda(\psi^T x) \) on \( x \). Thus equation (58) is proportional to the Heckman estimate \( \hat{\rho} \), the constant of proportionality depending only on the values of \( x \) in the observed case. Hence, if we test the hypothesis \( H_0: \rho = 0 \) conditional on the observed values of \( x \), the test based on equation (58) is equivalent to the regression test based on \( \hat{\rho} \).

Little (1985) pointed out, as have many others, that the estimate \( \hat{\rho} \) is unsatisfactory in practice because of its strong dependence on the correct specification of equations (38) and (39), and on the need for the range of variation of the propensity score \( \psi^T x \) to be sufficiently large for the non-linearity of \( \lambda(\psi^T x) \) to be evident in the observed cases. If the range of values of \( \psi^T x \) is small, the new regressor \( \lambda(\psi^T x) \) is highly collinear with the existing regressors \( x \), and so equation (60) is unstable. If the test of fit that is based on \( \hat{\rho} \) is unstable, then so is the test that is based on \( \hat{\rho} \). Copas and Li (1997) gave an example where two transformations of \( t \), apparently fitting the observed data equally well, lead to sharply different estimates of \( \rho \), and hence different estimates of \( \theta \).

In Section 5.3 we commented that, in the simpler problem of estimating the mean of a normal sample with missing observations, the Heckman model does attain maximum bias. Here \( \psi^T x \) is a constant, and so the added term \( \lambda(\psi^T x) \) in equation (59) is completely confounded with the main term in the regression. In this case, equation (58) is identically zero, as are both the numerator and the denominator of \( \hat{\rho} \) in equation (60). Again, this is a case where \( u_Y \) is a linear function of \( s_T \).

A rather similar situation arises in the literature on identifiability of competing risks. In a classic paper, Tsiatis (1972) showed that there is no available information on the dependence
between the potential lifetimes in the competing risks problem. However, Heckman and Honore (1989) showed that the problem is fully identified if we impose parametric models on the marginal life distributions. Crowder (1994) and others have since pointed out that the resulting estimates are highly dependent on the modelling assumptions that are made. See Crowder (2001) for a good review of this whole area.

There are also many other examples in the literature of identifiable models for the kind of problems that we are considering. For missing data, identifiable parametric models for the joint distribution of \( t \) and \( r \) of Section 5.1 have been proposed by Baker and Laird (1988), Chambers and Welsh (1993) and Park and Brown (1994), among many others. Identifiability may come through strong assumptions on other aspects of the model; for example Tang et al. (2003) assumed that the marginal distribution of covariates is known. Again, such models involve untestable assumptions or, in a Bayesian context, influential prior distributions.

6.2. Extra uncertainty

This discussion illustrates the central problem of incomplete-data analysis, that unless we make strong and unverifiable modelling assumptions we have little or no information about \( \varepsilon \), and hence little or no information about bias. Investing in a good model is always important, but particularly so here because of the lack of identifiability of important aspects of the model such as the ignorability assumptions that are implied in all our examples.

The strongest assumption is that \( \varepsilon = 0 \). This is modelling in the usual sense: we assume that \( f_Z \) (and hence \( f_Y \)) is the ‘true model’ in the sense that we are willing to rely on the inferences that are derived from it. In particular, if \( \phi = d^T \theta \), \( \hat{\phi}_Y = d^T \hat{\theta}_Y \) and \( \sigma^2_Y = (n_{I_Y})^{-1} d \), the asymptotic coverage probability of the confidence interval

\[
(\hat{\phi}_Y - d_\alpha \sigma_Y, \hat{\phi}_Y + d_\alpha \sigma_Y)
\]

is \( 1 - \alpha \). Equivalently, the pivotal quantity \( S = (\hat{\phi}_Y - \phi) / \sigma_Y \) is asymptotically standard normal.

In complete-data problems, where we can observe a sample of values of \( z \), a weaker interpretation of \( f_Z \) is as a ‘working model’: we do not assume that \( f_Z \) is necessarily true, but we use it for inference on the grounds that it gives an acceptable fit. We interpret this to mean that the actual distribution generating \( z \) is \( g_Z \), but that \( f_Z \) is accepted because \( |T_Z| \leq d_\alpha \) where \( T_Z \) is constructed for the worst case misspecification (27). Uncertainty of inference is now evaluated with respect to \( g_Z \) with \( \varepsilon \neq 0 \), but conditioning on the event \( |T_Z| \leq d_\alpha \). Since we are now allowing for misspecification, we expect this to increase uncertainty relative to the standard true model inference, but \( \varepsilon \) is unlikely to be too large because the null hypothesis that \( \varepsilon = 0 \) has been accepted by a goodness-of-fit test.

For this discussion to make sense, we need to ensure that the parameter \( \theta \) retains its meaning under both \( f_Z \) and \( g_Z \). In the terminology of Section 4, this means that we adopt Royall and Tsou’s (2003) assumption that \( \theta^{\text{NT}} = \theta^{\text{INF}} \), so that \( \theta = \theta_{gZ} \) and \( E_f(u_{Z|Z}) = 0 \). We now consider calculating the confidence interval (61) after accepting that \( f_Z \) is a ‘working model’. Our conjecture is that, to attain the same confidence coefficient (coverage), this interval needs to be widened to allow for the extra uncertainty through relaxing the status of \( f_Z \) from a true model to a working model. We shall find a factor \( k \geq 1 \) such that the coverage of \( \hat{\phi}_Y \pm k d_\alpha \sigma_Y \) in this broader sense remains at least \( 1 - \alpha \).

Similarly, when we can only observe incomplete data, we could describe \( f_Y \) as a working model if \( f_Y \) gives an adequate fit to the observed sample of values of \( y \). The difference now is that a good fit of \( f_Y \) no longer implies that \( \varepsilon \) is necessarily small. In the simple pattern mixture model that was mentioned in Section 6.1, for example, the observed values of \( t \) may give an
excellent fit to the normal distribution that is required by \( f_Y \), and yet \( \varepsilon \) may be large. If \( \varepsilon \) is large, then inference from \( \hat{\theta}_Y \) may be severely biased.

Since a good fit of \( f_Z \) to data on \( z \) necessarily implies a good fit of \( f_Y \) to the corresponding values of \( y \), we argue that the extra uncertainty that is implied by interpreting \( f_Y \) as a working model is as great as or greater than the extra uncertainty that is implied by interpreting \( f_Z \) as a working model. We therefore evaluate the factor \( k \) that was defined above and use this as a lower bound when basing inference on a working model on \( y \). If we are not willing to make the strong assumption that \( f_Y \) is the true model, and merely rely on its credentials as a working model, the actual error when we estimate \( \phi \) by \( \hat{\phi}_Y \) may be larger, and possibly substantially larger, than this calculation implies.

This is the key idea of this section. We can never know that our model is ‘correct’; the best that we can hope for is that it gives a good description of the data. The problem is that with data only on \( y \) we can never test \( f_Z \) fully, because of the identifiability problems that were discussed above. Instead we formulate our uncertainty on the assumption that \( f_Z \) gives a good fit to the (unobserved) data on \( z \) and use this as a lower bound to the actual uncertainty that we suffer when \( f_Y \) is used as a model for the data on \( y \). This argument leads to a confidence interval that is wider than expression (61), and hence less misleading than the native procedure which makes no allowance at all for model uncertainty.

To study \( k \), we need the joint distribution, under \( g_Z \), of the pivot \( S \) and the test statistic \( T_Z \) with \( u_Z \) in equation (27), which is

\[
T_Z = \frac{d^T}{\{nd^T(I^{-1}_Y - I^{-1}_Z)d\}^{1/2}} \sum (I^{-1}_Y s_Y(y; \theta) - I^{-1}_Z s_Z(z; \theta)) = \frac{\hat{\phi}_Y - \hat{\phi}_Z}{\sigma_Y(1 - \lambda)^{1/2}},
\]

where \( \lambda \) is the relative efficiency for estimating \( \phi = d^T \theta \) as defined in expression (5). As expected, if we can observe \( z \), we would test for the presence of incomplete-data bias by calculating \( \hat{\phi}_Z \) and \( \hat{\phi}_Y \) from the same set of data and testing the significance of the difference.

When \( \varepsilon = 0 \), both \( S \) and \( T_Z \) are asymptotically standard normal. Using equations (24), (1) and (4), the correlation between \( S \) and \( T_Z \) is \( (1 - \lambda)^{1/2} \). When \( \varepsilon \neq 0 \), \( \hat{\phi}_Y \) suffers the first-order approximate bias \( b_S \) from equation (20), and hence the corresponding approximate mean of \( S \) is \( b_S = \sigma_Y^{-1}b_S \). Note that if \( \varepsilon = O(n^{-1/2}) \) then \( b_S \) is \( O(1) \). Similarly, the expected value of \( T_Z \) is \( (1 - \lambda)^{-1/2}b_S \). Hence, if terms of size \( O(n^{-1/2}) \) and smaller are ignored, \( S \) and \( T_Z \) are jointly asymptotically normal with

\[
\begin{pmatrix}
S \\
T_Z
\end{pmatrix} \sim N\left\{ \begin{pmatrix}
1 \\
1
\end{pmatrix}, \begin{pmatrix}
1 & (1 - \lambda)^{1/2} \\
(1 - \lambda)^{1/2} & 1
\end{pmatrix} \right\}.
\]

Thus the conditional distribution of \( S \) given \( T_Z \) is approximately

\[
S|T_Z \sim N[b_S + (1 - \lambda)^{1/2}\{T_Z - (1 - \lambda)^{-1/2}b_S\}, 1 - (1 - \lambda)]
\]

\[
= N\{(1 - \lambda)^{1/2}T_Z, \lambda\}.
\] (62)

To this order of approximation, the conditional distribution of \( S \) given \( T_Z \) does not involve the misspecification bias \( b_S \). This argument applies for any misspecification function \( u_Z \) in \( g_Z \), not just the worst case function (27) that is used in \( T_Z \).

If we had actually observed \( T_Z \), we could use this conditional sampling distribution of the pivot \( S \) to construct a conditional confidence interval for \( \phi \). With the same significance level \( \alpha \) this would give

\[
\{\hat{\phi}_Y - (1 - \lambda)^{1/2}T_Z\sigma_Y - \lambda^{1/2}d_\alpha \sigma_Y, \hat{\phi}_Y - (1 - \lambda)^{1/2}T_Z\sigma_Y + \lambda^{1/2}d_\alpha \sigma_Y\}. \] (63)
If \( T_Z < d_\alpha \), the lower limit of interval (63) cannot be less than
\[
\hat{\phi}_Y - k_\lambda d_\alpha \sigma_Y,
\]
where
\[
k_\lambda = \lambda^{1/2} + (1 - \lambda)^{1/2}
= [1 + 2\{\lambda(1 - \lambda)\}^{1/2}]^{1/2}.
\]
Similarly, if \( T_Z > -d_\alpha \), the upper limit of expression (63) is at most expression (64) with the sign changed. Thus, if we assume that \( |T_Z| \leq d_\alpha \), then a conservative confidence interval for \( \phi \) is
\[
\{\hat{\phi}_Y - k_\lambda d_\alpha \sigma_Y, \hat{\phi}_Y + k_\lambda d_\alpha \sigma_Y\}.
\]
Since \( 0 < \lambda \leq 1 \) and so \( 0 \leq \lambda(1 - \lambda) \leq \frac{1}{4} \), the second equality in equation (65) shows that
\[
1 \leq k_\lambda \leq \sqrt{2}.
\]
Comparing expression (66) with expression (61) we see that relaxing the status of model \( f_Z \) from a true to a working model has led to a wider interval, by a factor which depends on the value of \( \lambda \), i.e. on the proportion of information that is retained in the incomplete data. The width of the interval, however, never increases by more than a factor of \( \sqrt{2} \), which we can think of as ‘doubling the variance’, recalculating the usual confidence interval with the variance \( \sigma_Y^2 \) doubled to \( 2\sigma_Y^2 \).

To see this in another way, if we could observe \( z \), then we could estimate \( \phi \) with variance (under \( f_Z \)) of \( \sigma_Z^2 = d^2 (n T_Z)^{-1} d \). With incomplete data, the variance increases to \( \sigma_Y^2 = \sigma_Z^2 / \lambda \). But, if \( f_Z \) is weakened to a working model, the expanded confidence interval (66) is the same as the conventional interval (61) but with the variance \( \sigma_Y^2 \) increased further to \( k_\lambda^2 \sigma_Z^2 \), which we could call the pseudovariance. From interval (65), the pseudovariance is
\[
\frac{k_\lambda^2}{\lambda} \sigma_Z^2 = \left\{ \frac{1}{\lambda} + 2\left[\frac{1 - \lambda}{\lambda}\right] \right\} \sigma_Z^2.
\]
The right-hand side of equation (67) splits the pseudovariance into the ordinary variance (assuming that \( f_Z \) is true), plus the effect of bias resulting from model uncertainty. Fig. 3 shows the

**Fig. 3.** Pseudovariance versus \( \lambda \)
total variance inflation factor in equation (67) in terms of \( \lambda \). The first term is the dotted line and the second the broken line, giving the total as the full line. All three curves are decreasing functions of \( \lambda \), as expected.

The rather informal argument leading to equation (65) is based on considering the upper and lower confidence limits separately. For a tighter bound, consider the (conditional) coverage probability of the two-sided confidence interval \( \hat{\phi}_Y \pm k_d \sigma_Y \) under the working model \( f_Z \). This is

\[
P_{gZ}(\hat{\phi}_Y - k_d \sigma_Y \leq \phi \leq \hat{\phi}_Y + k_d \sigma_Y \mid |T_Z| \leq d_\alpha)
\]

\[
= E_{gZ} \left[ \Phi \left( \frac{k_d \alpha - (1 - \lambda)^{1/2} T_Z}{\lambda^{1/2}} \right) - \Phi \left( -k_d \alpha - (1 - \lambda)^{1/2} T_Z \right) \right] \left| T_Z \leq d_\alpha \right]. \tag{68}
\]

Now define \( k^* = k^*(\lambda, \alpha) \) as the unique solution of

\[
\Phi \left( \frac{d_\alpha k^* - (1 - \lambda)^{1/2}}{\lambda^{1/2}} \right) - \Phi \left( -d_\alpha k^* + (1 - \lambda)^{1/2} \right) = 1 - \alpha.
\]

Then the coverage probability (68) is at least \( 1 - \alpha \) if \( k \geq k^* \). If \( k < k^* \) then the coverage falls below \( 1 - \alpha \) for at least some values of \( b_S \).

Fig. 4 illustrates the values of \( k^* \) for \( \alpha = 0.05 \) and \( \alpha = 0.01 \). For each \( \lambda, k^* \) increases as \( \alpha \) becomes more extreme but is always less than the curve for \( k_\lambda \) in equation (65), which is also shown. In fact

\[
1 \leq k^*(\lambda, \alpha) < k_\lambda \leq \sqrt{2}. \tag{69}
\]

Of the three inequalities in expression (69), the first is attained when \( \lambda = 1 \) (no loss of information), the second is attained in the limit as \( \alpha \to 0 \) and the third is attained when \( \lambda = \frac{1}{2} \).
In summary, we have three asymptotic coverage statements. Firstly,

\[ P_f(\hat{\phi}_Y - d_\alpha \sigma_Y \leq \phi \leq \hat{\phi}_Y + d_\alpha \sigma_Y) = 1 - \alpha, \]

the conventional asymptotic confidence interval when \( f_Z \) is the true model. But if \( f_Z \) has the weaker status of a working model, defined by conditioning on the event \( |T_Z| \leq d_\alpha \) (so that \( \varepsilon = O(n^{-1/2}) \) for this event to happen with non-vanishing probability), then the conventional interval is no longer a confidence interval with this coverage, as

\[ P_g(\hat{\phi}_Y - d_\alpha \sigma_Y \leq \phi \leq \hat{\phi}_Y + d_\alpha \sigma_Y| |T_Z| \leq d_\alpha) < 1 - \alpha \]

for at least some possible misspecified distributions \( g_Z \). But

\[ P_g(\hat{\phi}_Y - k^* d_\alpha \sigma_Y \leq \phi \leq \hat{\phi}_Y + k^* d_\alpha \sigma_Y| |T_Z| \leq d_\alpha) > 1 - \alpha \]

for all possible distributions \( g_Z \) within the asymptotic set-up that is being discussed. Of course this is a hypothetical calculation since \( f_Z \) is unobserved, but the expanded confidence limits in inequality (71) involve only the \( y \). Our argument is that in practice, when we only accept \( f_Z \) as a working model, our uncertainty limits for \( \phi \) should be at least as wide as those in inequality (71).

The fact that, when \( \lambda = 1, k^* = k_\lambda = 1 \) emphasizes the importance of the assumption that \( \theta^{\text{INT}} = \theta^{\text{INF}} \). For complete data there is then no asymptotic penalty if we treat a working model as if it was a true model. But when \( \lambda < 1 \) the distinction is important, as seen in inequality (70).

The same argument also applies in multiparameter problems. Suppose that we want to find a confidence region for \( \theta \) itself, containing \( m \) components say. Let \( \Sigma_Y = \text{var}(\hat{\theta}_Y) \), and \( \Lambda = I_Y^{-1/2} T_Z^{-1} I_Y^{1/2} \) as before. Then the multivariate analogues of \( S \) and \( T_Z \) are

\[ S = \Sigma_Y^{-1/2} (\hat{\theta}_Y - \theta), \]
\[ T_Z = (I - \Lambda)^{1/2} \Sigma_Y^{-1/2} (\hat{\theta}_Y - \hat{\theta}_Z). \]

If \( \varepsilon = 0 \) then \( \text{var}(S) = \text{var}(T_Z) = I \). The distribution of \( S \) then allows us to write

\[ \theta = \hat{\theta}_Y + \Sigma_Y^{1/2} U, \]

(72)

where \( U \) is a random vector from \( N(0, I) \). The usual asymptotic confidence ellipsoid for \( \theta \) is the set of all values of equation (72) that are consistent with the inequality \( U^T U \leq d_\alpha \), where \( d_\alpha \) is now the \((1 - \alpha)\)-quantile of the \( \chi^2 \)-distribution on \( m \) degrees of freedom.

When \( \varepsilon = O(n^{-1/2}) \) we allow for first-order bias in the same way as before to give the generalization of distribution (62) as \( S|T_Z \sim N((I - \Lambda)^{1/2} T_Z, \Lambda) \). Now we can write, conditional on \( T_Z \),

\[ \theta = (\hat{\theta}_Y - \Sigma_Y^{1/2} (I - \Lambda)^{1/2} T_Z) - (\Sigma_Y^{1/2} \Lambda^{1/2} U), \]

(73)

The expanded confidence region is now the collection of all values of equation (73) that are consistent with the two inequalities \( T_Z^T T_Z \leq d_\alpha \) and \( U^T U \leq d_\alpha \).

To extend the univariate discussion, we now define \( k_{\min} \) to be the smallest value of \( k \) such that the region that is generated equation from (73) by \( T_Z^T T_Z \leq d_\alpha \) and \( U^T U \leq d_\alpha \) lies everywhere within the region that is generated from equation (72) by \( U^T U \leq k^2 d_\alpha \). The generalization of the ‘double-variance’ result is that \( k_{\min}^2 \leq 2 \). This is proved in Appendix A. The condition for the bound to be attained is that the terms \( \lambda_i - \frac{1}{2} \) are not all of the same sign, where the \( \lambda_i \)'s are the eigenvalues of \( \Lambda \). If \( \lambda_i \leq \lambda < -\frac{1}{2} \) for all \( i \), or \( \lambda_i \geq \lambda > \frac{1}{2} \) for all \( i \), then \( k_{\min} = k_\lambda \) in equation (65).

Fig. 5 shows two examples with \( m = 2, \alpha = 0.05 \) and
Fig. 5. Examples of confidence ellipsoid: (a) unattainable case; (b) double variance

\[ \hat{\theta}_Y = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \]
\[ \Sigma_Y = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}, \]
\[ \begin{pmatrix} \lambda_{\text{min}} \\ \lambda_{\text{max}} \end{pmatrix} = \begin{pmatrix} 0.1 \\ 0.2 \end{pmatrix} \]  
(Fig. 5(a))
or

\[
\begin{pmatrix}
0.05 \\
0.9
\end{pmatrix}
\]

(Fig. 5(b)).

In each of the two diagrams, the ellipses illustrate the two distinct components of the right-hand side of equation (73). The inner ellipse that is centred on the origin (drawn in bold) is the locus of the first term of equation (73) as \( T_Z \) varies over the circle \( T_Z^T T_Z = d_\alpha \). An ellipse corresponding to the values of the second term in equation (73) as \( U \) varies over the circle \( U^T U = d_\alpha \) is then centred on each point of the first ellipse, to give the collection of ellipses that are drawn with light lines. The outer envelope of these ellipses is the conservative confidence region for \( \theta \).

Of the two larger concentric ellipses that are drawn with bold lines in each graph of Fig. 5, the inner is the conventional confidence region for \( \theta \) given by \( S^T S = d_\alpha \). The outer is the region for \( \theta \) that is defined by \( S^T S = 2d_\alpha \). Note that the envelope allowing for all possible values of the bias that are consistent with the acceptance region \( T_Z^T T_Z \leq d_\alpha \) is contained within the conventional confidence ellipse but with the variances doubled.

The eigenvalues for these two examples show that \( k_{\min} = \sqrt{2} \) in the second case but not the first. This is confirmed in Fig. 5. In Fig. 5(a) we see that the envelope is everywhere inside the outer ellipse, but in Fig. 5(b) we see that the envelope touches the outer ellipse at exactly four points.

7. Example: passive smoking and lung cancer

We return to the study of passive smoking and lung cancer that was discussed in Section 1 and examine in more detail how sensitively the estimation of relative risk can depend on the influence of potential confounders. As mentioned, Hackshaw et al. (1997) was based on a meta-analysis of 37 separate epidemiological studies which compared the risk of lung cancer among non-smokers according to whether the spouse of the subject did or did not smoke. See Hackshaw et al. (1997) for full details.

Using the notation of Section 3.2, let \( x \) to be the binary exposure variable taking values 1 (exposed; spouse smokes) and 0 (unexposed; spouse a non-smoker), and let \( c \) be another (unmeasured) variable which may also affect the risk of cancer. A measure of quality of diet is just one possibility for \( c \). Suppose that an individual's risk of cancer is log-linear in \( x \) and \( c \) with coefficients \( \theta \) and \( \alpha \) respectively. Then, approximately, the estimate of log-relative-risk \( \hat{\theta} \) that is calculated from the \( j \)th study, \( \hat{\theta}_j \) say, will be biased by the amount \( \alpha d_j \), where \( d_j \) is the difference in the average values of \( c \) between the \( n_{1j} \) exposed cases and the \( n_{0j} \) unexposed cases.

If \( \sigma_c^2 \) is the variance of \( c \) within each level of \( x \), and \( \kappa_j = n_{0j}^{-1} + n_{1j}^{-1} \), then

\[
\hat{\theta}_j | d_j \sim N(\theta + \alpha d_j, \sigma^2 + \tau^2 - \alpha^2 \kappa_j / \sigma_c^2),
\]

where \( \sigma_j^2 \) is the within-study variance and \( \tau^2 \) is the between-study (heterogeneity) variance, defined as in the usual random-effects model for meta-analysis (Sutton et al., 2000a). We have formulated the variance in distribution (74) so that \( \hat{\theta}_j \) marginalizes over \( d_j \) to the standard random-effects model, as explained below.

Let

\[
\begin{align*}
T_Z &= \sum w_j d_j / \sigma_c \sigma_T \sum w_j, \\
\sigma^2_T &= \sum w_j^2 \kappa_j^{-1} / (\sum w_j)^2.
\end{align*}
\]

CRiSM Paper No. 05-1, www.warwick.ac.uk/go/crism
and define \( \hat{\theta} \) to be the pooled estimate of log-relative-risk by using the method of DerSimonian and Laird (1986). Then \( \hat{\theta} \), and its conditional distribution given \( T_Z \), are

\[
\hat{\theta} = \frac{\sum w_j \hat{\theta}_j}{\sum w_j},
\]

\[
\hat{\theta}|T_Z \sim N(\theta + \alpha \sigma_c \sigma_T T_Z, \sigma_\theta^2 - \alpha^2 \sigma_c^2 \sigma_T^2).
\]

where \( \sigma_\theta^2 = 1/\sum w_j \). For the pivot \( S = (\hat{\theta} - \theta)/\sigma_\theta \), we therefore have

\[
S|T_Z \sim N\left(\frac{\alpha \sigma_c \sigma_T T_Z}{\sigma_\theta}, 1 - \frac{\alpha^2 \sigma_c^2 \sigma_T^2}{\sigma_\theta^2}\right),
\]

(76)

By ignoring the problem of confounding, the published analysis tacitly assumes that \( x \) and \( c \) are conditionally independent within each study. If this is so then \( E(d_j) = 0 \) and so distributions (74) and (76) marginalize to the usual distributions \( \hat{\theta}_j \sim N(\theta, \sigma_j^2 + \tau^2) \) and \( S \sim N(0, 1) \). But \( x \) and \( c \) may be correlated—suppose that, for each individual in the \( j \)th study,

\[
c|x \sim N\{\psi_j + \epsilon \sigma_c (x - \frac{1}{2}), \sigma_c^2\}.
\]

(77)

This means that

\[
\begin{align*}
d_j &\sim N(\epsilon \sigma_c, \kappa_j \sigma_c^2), \\
T_Z &\sim N(\sigma_\theta^{-1} \epsilon, 1), \\
S &\sim N(\alpha \sigma_c \sigma_\theta^{-1} \epsilon, 1).
\end{align*}
\]

(78)

The size of \( \epsilon \) can be calibrated in terms of \( \rho = \text{corr}(x, c) \), by

\[
\epsilon = \frac{\rho}{\sigma_x \sqrt{1 - \rho^2}}.
\]

(79)

If \( \rho = 0 \) then \( \epsilon = 0 \) and vice versa.

If we let, for any given \( j, z = (\hat{\theta}_j, d_j) \) and \( y = \hat{\theta}_j \), then the above is just another special case of the general formulation of Section 2. Here we have the trivial extension of allowing for the different sample sizes at each value of \( j \). The model \( f_Z \) (ignorable confounding) is the product of the submodels (74) and \( d_j \sim N(0, \kappa_j \sigma_j^2) \). The model \( f_Y \) is \( \hat{\theta}_j \sim N(\theta, \sigma_j^2 + \tau^2) \). Comparing \( \text{var}(\hat{\theta}|T_Z) \) with \( \sigma_\theta^2 \), the usual meta-analysis variance, gives the loss of efficiency through ignoring \( c \) as

\[
1 - \lambda = \left(\frac{\alpha \sigma_c \sigma_T}{\sigma_\theta}\right)^2.
\]

(80)

In the notation of Section 4, the true model \( g_Z \) is the product of distribution (74) and the distribution of \( d_j \) in expression (78), and so for small \( \epsilon \)

\[
\log \left(\frac{g_Z}{f_Z}\right) \approx \frac{\epsilon d_j}{\sigma_c \kappa_j}.
\]

With the convention \( E_f(u_Z^2) = 1, \epsilon \) must be rescaled to \( \epsilon_z = \epsilon \kappa_j^{-1/2} \) for the general notation to apply to the \( j \)th study. Note that when \( \epsilon = 0 \) the confounding is ignorable and \( g_Z = f_Z \).

To see how this example illustrates Section 6, if we had observed the values of \( c \) we could use the test statistic \( T_Z \) in expression (75) to check the ignorability assumption in \( f_Z \) by testing the null hypothesis that \( \epsilon = 0 \). From expression (78), \( T_Z \) is standard normal if \( \epsilon = 0 \). Further, from equation (80) we see that distribution (76), the conditional distribution of the pivot \( S \) given \( T_Z \), agrees exactly with the general formula (62).

Following Section 6.2, we can now consider assessing the uncertainty in \( \hat{\theta} \) under three different scenarios, making decreasingly strong assumptions about \( \epsilon \):
(a) we assume that $\varepsilon = 0$;
(b) we are uncertain about $\varepsilon$ but assume that, if we had been able to measure $c$ for all the subjects in these studies, we would confirm that there is no significant correlation between $c$ and $x$;
(c) we are uncertain about $\varepsilon$ and cannot measure $c$.

Now let $q_{L_1}(\varepsilon)$ and $q_{U_1}(\varepsilon)$ be the lower and upper 2.5-percentiles of the marginal distribution of $S$, and $q_{L_2}(\varepsilon)$ and $q_{U_2}(\varepsilon)$ be the same percentiles but for the conditional distribution of $S$ given $|T_Z| < 2$. Then the corresponding confidence intervals for $\theta$ are, for $k = 1, 2$,

$$C_k(\varepsilon) = \{\hat{\theta} - q_{U_k}(\varepsilon)\sigma_\theta, \hat{\theta} - q_{L_k}(\varepsilon)\sigma_\theta\}.$$

Note that $q_{U_1}(0) = -q_{L_1}(0) \approx 2$.

Under the first scenario, $C_1(0)$ is the ordinary 95% confidence interval for $\theta$ as calculated by Hackshaw et al. (1997). But, if $\varepsilon \neq 0$, $C_1(\varepsilon)$ reflects the confounding bias which we risk under scenario (c). Interval $C_2(\varepsilon)$ also reflects this bias, but conditional on a check that $\varepsilon = 0$ would seem sensible in the light of measurements on $c$. According to Section 6.2, $C_2(\varepsilon)$ is everywhere within the interval $C^*$, which is defined to be the interval $C_1(0)$ widened by the factor $\sqrt{2}$. As $\varepsilon$ is unknown, $C^*$ seems a safe inference under scenario (b). Our state of ignorance about $\varepsilon$ is greater under scenario (c) than under scenario (b), and so our argument is that in practice, when we know little or nothing about $\varepsilon$, it is less misleading to report our inference as $C^*$ than the usual confidence interval $C_1(0)$.

Fig. 6 shows these confidence intervals for the meta-analysis of Hackshaw et al. (1997), plotted against $\rho$. All values are shown on the original relative risk scale. We have calculated $\hat{\theta}_j$ and $\kappa_j$, and estimated the parameters $\sigma_j$ and $\tau$ directly from Table 1 of Hackshaw et al. (1997). We assume that $\alpha$ and $\sigma_c$ are such that $\lambda = \frac{1}{2}$: this means that, under scenario (a), if we had been
able to design these studies properly by controlling on levels of $c$, then only half the sample size would have been needed to give the same accuracy in assessing the effect of exposure. To calibrate $\varepsilon$ in terms of $\rho$ we have used equation (79) and taken $\sigma_x$ to be the standard deviation of the binary variable given by the relative numbers of exposed and non-exposed subjects in these studies. The percentiles that are needed for $C_1(\varepsilon)$ and $C_2(\varepsilon)$ are estimated by simulation using 10000 replications of the joint distribution of $S$ and $T_Z$ in distributions (76) and (78).

The graph illustrates how sensitively inference can depend on assumptions about ignorability. The confidence interval of Hackshaw et al. (1997) for a relative risk of 1.13–1.36 is shown as the inner horizontal dotted lines. This agrees with $C_1$ (full lines) when $\rho = 0$. However, the correlation $\rho$ has to rise to only 0.03 before the conclusion is compromised: if $\rho > 0.03$, $C_1$ includes values for the relative risk of less than 1, so the assertion of a causal link between passive smoking and lung cancer is no longer significant. This accords with the view that is taken by some epidemiologists that any observed relative risk of less than about 2 should be regarded with considerable caution. Interval $C_2$ (the chain curves) controls the risk by staying within the expanded interval $C_\ast$, which is shown as the outer horizontal dotted lines. Unless we can be confident that the correlation is less than about 0.01, $C_1$ contains values that are outside both $C_1(0)$ and $C_\ast$, but $C_\ast$ is closer. In this sense it seems safer to give the inference as $C_\ast$ rather than $C_1(0)$. For these data, $C_\ast$ gives the relative risk from 1.08 to 1.41, still suggesting a causal effect but with a considerably wider margin of error.

8. Comments

Our formulation of $y$ as a deterministic function of $z$ means that the components of $z$ must be a mixture of responses of interest, and any subsidiary stochastic variables which determine the incompleteness of the data. Heitjan and Rubin (1991) avoided this by modelling the measurement process separately. Their formulation is more complicated algebraically, but assumptions such as ‘coarsening at random’ are more transparent. Here, details of the measurement process, including any nuisance parameters that are involved, are buried within the overall model $f_Z$. Covariates are also included in $z$, so the single model $f_Z$ implies a fully random model rather than conditioning on observed covariates as would be usual in regression models. If $f_Z$ makes the covariates ancillary for $\theta$, as in equation (13) for example, then this distinction is unimportant as far as asymptotic maximum likelihood estimation is concerned.

The approximations in this paper have been relatively simple because we have only retained linear terms in $\varepsilon$, and we have assumed standard asymptotics of maximum likelihood. The major simplifications include that the maximum bias in equation (23) depends on the model only through the information matrices $I_Z$ and $I_Y$, and the effect of misspecification on the variances can be ignored. Higher order approximations are much more complicated, although some progress is possible in particular cases. Copas and Li (1997), for example, showed that, for the Heckman model (Section 5.3), linear approximations to bias are in fact accurate up to and including second-order terms in $\varepsilon$.

By taking $\varepsilon = O(n^{-1/2})$ we are working with local asymptotics and not asymptotics in the more usual statistical sense of keeping the model fixed but letting the sample size $n \to \infty$. If $\varepsilon = O(1)$ then the bias will dominate the variance if $n$ is sufficiently large, and the power function in expression (56) tends to 1. Our approximations only provide a sensitivity analysis for undetectably small misspecifications of $f_Z$. Our interest in this is closely analogous to the theory of locally optimal statistical tests, which is concerned with evaluating power functions not against global alternative hypotheses but against alternatives that are increasingly close to the null. Sen (1985), page 96, wrote
'for a meaningful study of the asymptotic power of tests, one would naturally confine oneself to the locality of the null hypothesis for which the power functions may not converge to 1... we remark that a locally optimal test may not perform that well for nonlocal alternatives, particularly when the sample size is not large'.

Here we have the added difficulty that tests of $f_z$ against alternatives $g_Z$ need the data on $z$ and not just $y$, and hence our discussion in Section 6.

Our discussion of the consequences of misspecification is strongly parametric in the sense that we are interested only in the bias of the parameters that are defined by $f_z$. Even if $\theta = \theta_{gZ}$, and $\epsilon$ is sufficiently small for the misspecification to be undetectable, there may be other aspects of the fitted distribution which differ sharply from the corresponding true values for $g_Z$. For example, in fitting a normal distribution we always have consistent estimates for the mean and variance, but this does not mean that the tails of the distribution are correctly estimated. Lawless (2003), pages 250–252, gave some striking examples of this in the context of estimating life distributions. Gustafson (2001) suggested a more general setting for this discussion, in which, for data $z$ generated by true distribution $g(z)$, the object of interest is the functional $\gamma_g = T\{g(\cdot)\}$. For parametric model $f(z; \theta)$, with score function $s(z; \theta)$, let $\gamma(\theta) = T\{f(\cdot; \theta)\}$. Then by concentrating on bias in estimates of $\theta$ we are tacitly making the strong assumption that the functional satisfies the identity $E_g s(z; \gamma^{-1}(\gamma_g)) = 0$ for all $g$. For other functionals $T$, $\gamma_g$ may be grossly misspecified, as the examples in Lawless (2003) show.

Our paper raises the fundamental question of how to combine two kinds of uncertainty: uncertainty in the model and uncertainty arising from sampling variability in the usual sense. The Bayesian paradigm provides a complete solution, at least in principle, since the model, parameters and data are then all thought of as randomly sampled from some appropriately defined superpopulation. In our notation, the typical approach is to choose a parametric model for $uZ$ and to assume a joint prior distribution for the parameters of this model along with the parameter $\theta$ of the main model $f_z$. Many Bayesian papers take this approach for particular incomplete-data problems, e.g. Forster and Smith (1998) on missing data and Givens et al. (1997) on publication bias in meta-analysis. Problems of identifiability of model misspecification which we have discussed re-emerge in the form of strong dependence on the prior distribution or strong dependence on the particular form that is assumed for $uZ$, or both. Sometimes there may be substantive prior information; for example Rubin (1977) discussed how the effects of missing data in a survey can be assessed by using subjective notions about the similarity between respondents and non-respondents, and Scharfstein et al. (2003) argued that the doctors who are involved in a longitudinal trial of human immunodeficiency virus drugs are likely to have clear opinions about which patients are likely to drop out before the end of the trial. However, it is difficult to see that this would be the case in most applications, and so the difficulty of combining the model and sampling uncertainty remains. Greenland (2005), section 4.5, discusses the same issue, emphasizing the inadequacy of methods which claim to ‘let the data speak for themselves’, pointing out that ‘without external inputs, observational data say nothing at all about causal effects’. We have proposed a rather contrived approach to one aspect of the problem in Section 6.2. Whether there is a fully satisfactory solution in the frequency domain remains an open question.

Another fundamental question is raised by our use of the term ‘working model’. This term is rather misleading, since it refers to both the model and the data and is not a property of the model as such. The aim is to capture normal practice in exploratory data analysis: we search possible models and use only one for inference after verifying that it gives an acceptable fit to the data. The model chosen then depends on the data and so it is misleading to use (unconditional) sampling distributions as if the model was fixed. A general formulation of this seems difficult.
Our approach in Section 6.2 is to condition on the value of a goodness-of-fit test statistic, and then to take bounds over the range of this statistic which would lead us to accept this particular model. In this way, by conditioning out the first-order dependence on \( e \), we can work within both the single model \( f_Z \) and its close neighbourhood \( g_Z \). If the model search had been sufficiently structured, we could work within a nested sequence of models \( f_Z \) and define model choice in terms of some appropriate penalty function. Whether there is a satisfactory formulation of the problem which avoids making strong assumptions about such a nested sequence again seems an open question.

We have had space to discuss only a few special cases of incomplete data where model uncertainty is problematic. There are many other examples, which are equally important (and difficult), each with its own large literature. We could mention non-compliance in clinical trials (White and Pockock, 1996; Goetghebeur and Lapp, 1997), drop-out in longitudinal trials (Diggle and Kenward, 1994; Little, 1995; Scharfstein et al., 1999), non-ignorable censoring in survival analysis (Moeschberger and Klein, 1995; Scharfstein and Robins, 2002; Siannis et al., 2005) and errors in variables (Gustafson, 2002). Some of these problems, particularly those of Section 5.2, are closely related to the discussion of estimating causal effects (Rubin, 1974; Holland, 1986; Pearl, 2000). The design and analysis of observational studies raises many of these same issues: see Rosenbaum (1995) for a route into this large literature. Rosenbaum (2004) is related to Section 5.2 above and gives a useful review of the author’s pioneering papers on sensitivity analysis, which are also reviewed in Copas and Li (1997). Greenland (2003) has given an accessible account of confounding and measurement problems in the context of a topical case-study (power lines and childhood leukaemia), with a broader discussion of the same application in Greenland (2005).

**Appendix A**

We now verify the statement that is made in Section 6.2, that the double-the-variance property extends to confidence regions for the vector parameter \( \theta \).

First we simplify the notation by transforming the parameter space of \( \theta \) by multiplying by \( I_1^2 \) after the shift to \( \hat{\theta}_Y \). From equation (73), the confidence region now consists of the values of the new parameter \( \omega \) satisfying

\[
\omega = I_Y^{1/2} (\hat{\theta}_Y - \hat{\theta}_Y) = - (I - \Lambda)^{1/2} T_Z - \Lambda^{1/2} U. \tag{81}
\]

Our problem reduces to whether the region of \( \omega \) satisfying equation (81) with constraints \( T_Z^T T_Z \leq d_\alpha \) and \( U^T U \leq d_\alpha \) is included in \( \{ \omega \mid \omega^T \omega \leq 2d_\alpha \} \).

Let \( B_r (c) \) be an \( m \)-dimensional ball with centre \( c \) and radius \( r \). Then the assertion to be proved reduces to

\[
\bigcup_{\vec{c} \in \{ I - \Lambda \}^{1/2} B_r (0)} \Lambda^{1/2} B_r (c) \subseteq B_{2d_\alpha} (0).
\]

Thus it suffices to show the following lemma.

**Lemma 1.**

\[
\max \left\{ \sum_{i=1}^m b_i^2 : \sum_{i=1}^m a_i^2 \frac{1}{1 - \lambda_i} = 1, \sum_{i=1}^m (a_i - b_i)^2 \frac{1}{\lambda_i} = 1, \ 0 < \lambda_i < 1 \right\} \leq 2.
\]

**Proof.** Consider the Lagrangian function

\[
f(a, b, \lambda, \nu, \eta) = \sum_{i=1}^m b_i^2 - \nu \left( \sum_{i=1}^m \frac{a_i^2}{1 - \lambda_i} - 1 \right) - \eta \left( \sum_{i=1}^m \frac{(a_i - b_i)^2}{\lambda_i} - 1 \right),
\]

where
where $\nu$ and $\eta$ are Lagrangian multipliers. If $f$ has an equilibrium point at $(a, b, \lambda, \nu, \eta)$, then
\[
\frac{\partial f}{\partial a_i} = -2\nu \frac{a_i}{1 - \lambda_i} - 2\eta \frac{a_i - b_i}{\lambda_i} = 0,
\]
\[
\frac{\partial f}{\partial b_i} = 2\eta \frac{a_i - b_i}{\lambda_i} = 0,
\]
\[
\frac{\partial f}{\partial \lambda_i} = -\nu \frac{a_i^2}{(1 - \lambda_i)^2} + \eta \frac{(a_i - b_i)^2}{\lambda_i^2} = 0.
\]

We can, without loss of generality, assume that $0 \leq a_i \leq b_i$ for $i = 1, 2, \ldots, m$, these inequalities being strict for at least some values of $i$. Then $\eta$ must be positive from equation (83), and so is $\nu$ from equation (82). From equation (84) we obtain
\[
\sqrt{\nu} \frac{a_i}{1 - \lambda_i} + \sqrt{\eta} \frac{a_i - b_i}{\lambda_i} = 0,
\]
which, comparing with equation (82), gives $\eta = \nu$. Hence
\[
a_i = (1 - \lambda_i)b_i.
\]

For this to be compatible with equation (83) we must also have $\eta = 1$.

Finally, we substitute equation (85) into the two constraint equations to give
\[
1 = \sum_i a_i^2 = \sum_i (1 - \lambda_i)b_i^2,
\]
\[
1 = \sum_i \frac{(b_i - a_i)^2}{\lambda_i} = \sum_i \lambda_i b_i^2.
\]

Adding these two equations gives $\Sigma b_i^2 = 2$. Thus all equilibrium points of $f$ satisfy $f = 2$, and so this must be the global maximum (since $\Sigma b_i^2$ is clearly bounded).

References

J. Copas and S. Eguchi


