J. R. Statist. Soc. B (1997)
59, *No.* 1, *pp.* 55–95

Inference for Non-random Samples

By J. B. COPAS† and H. G. LI

University of Warwick, Coventry, UK *The Open University, Milton Keynes, UK*

[*Read before* The Royal Statistical Society *at a meeting organized by the* Research Section *on Wednesday, April 17th, 1996*, Professor R. L. Smith *in the Chair*]

SUMMARY

Observational data are often analysed as if they had resulted from a controlled study, and yet the tacit assumption of randomness can be crucial for the validity of inference. We take some simple statistical models and supplement them by adding a parameter θ which reflects the degree of non-randomness in the sample. For a randomized study θ is known to be 0. We examine the profile log-likelihood for θ and the sensitivity of inference to small non-zero values of θ. Particular models cover the analysis of survey data with item non-response, the paired comparison t-test and two group comparisons using observational data with covariates. Some practical examples are discussed. Allowing for sampling bias increases the uncertainty of estimation and weakens the significance of treatment effects, sometimes substantially so.

Keywords: DARWIN'S DATA; HECKMAN'S TWO-STAGE METHOD; ITEM NON-RESPONSE; MISSING DATA; PROFILE LIKELIHOOD; RANDOMIZATION; SELECTIVITY BIAS

1. INTRODUCTION

One of the greatest contributions of Fisher was his insight into the importance of randomization, not only in the design of experiments and surveys but also as the logical underpinning of methods of analysis. In the third chapter of *The Design of Experiments*, Fisher (1966) (first edition 1935) discussed Darwin's famous data on the heights of cross- and self-fertilized plants and showed that a permutation analysis based on a simple model of randomization of treatment order to pairs gives almost exactly the same P-value as the t-test. The essence of Fisher's argument is that it is randomization, or an equivalent assumption of sampling from a population, which justifies the use of standard significance tests and other methods of normal inference. However, methods designed for analysing experimental data are also routinely applied to observational data, sometimes (often?) with little or no recognition of the fact that the absence of randomization has, in Fisher's sense, removed the grounds for the validity of these methods. Essentially, randomization becomes a *model* for the data rather than a factual statement of how the data were obtained.

Modern statistics places great emphasis on the testing of assumptions. But the argument that randomization underpins the standard model assumptions is not reversible — the empirical verification of these assumptions does not imply that the hidden assumption of randomization is necessarily justified so that standard inference statements can safely be made. Often, interesting features of observational data, such as a 'significant' difference between responses of subjects given different

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

0035–9246/97/59055

treatments, can just as well be explained by biases in the way that those subjects have been allocated to the treatments.

As an example to illustrate this, and to introduce the later discussions in this paper, we take the analysis of data on a new form of dialysis (ambulatory peritoneal dialysis) for patients with kidney disease, reported in Burton and Wells (1989). The response of interest here is the rate of hospitalization, measured as the number of days in hospital per annum. A feature of the data highlighted by Burton and Wells (1989) is that this rate appeared to decrease over time, as seen in Fig. 1 which plots log(rate) against date of commencement of therapy. The linear regression

$$y = 3.93 - 0.079x$$

gives a reasonable summary of the data, where y is log(rate) and x is the date in years measured from 1980.

Patients who were not assigned to the new treatment were simply given the standard therapy (haemodialysis), but the treatment allocations were not random. Data on patients on both treatments were available, so we can model the selection process by a probit analysis which gives

$$P(S|x) = \Phi(-0.514 + 0.17x) \tag{1}$$

where S is the event of allocation to the new treatment and Φ is the standard normal distribution function. Thus as time progressed an increasing proportion of patients were allocated to the new therapy. But, much more importantly, each allocation decision was made by the clinician involved and so would be influenced by the patient's clinical state, which could itself have a direct bearing on the outcome y. An alternative explanation of the trend in Fig. 1 is that the average of y in fact remains constant throughout but that it is the allocation process which has changed over time. Although in this context y is only defined for the patients on the new treatment, we could envisage extending its definition to other patients by letting y be the value that *would* have been observed *had* the patient been allocated to the new treatment. At least conceptually we can then extend equation (1) by allowing the probit to depend on both x and y, and, if we make some further assumptions to be set out in Section 2, then we can estimate this extended model by using the fact that the data in Fig. 1 come from the conditional distribution of y given both x and S. This leads to

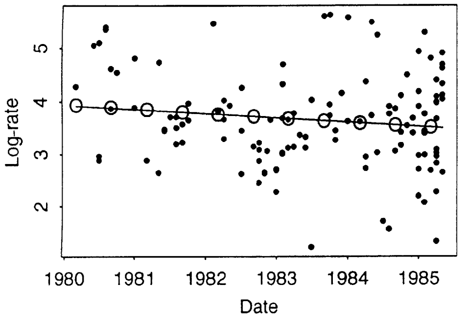

Fig. 1. Log(hospitalization rate) against start date: ●, date; ———, least squares; ○, selection

$$P(S|x, y) = \Phi(-7.79 + 0.34x + 1.85y). \tag{2}$$

For $x = 0$ (1980 admission), only a large value of y will make the right-hand side of equation (2) sufficiently large for that patient to have much chance of being selected, but as x increases smaller values of y are more likely to be among those selected. The model goes on to imply that

$$E(y|x, S) = 3.09 + 0.75\lambda(-0.514 + 0.17x) \tag{3}$$

where the function λ is the Mills ratio defined by $\lambda(u) = \phi(u)/\Phi(u)$, ϕ being the standard normal density function. Values of equation (3) are shown as the circles in Fig. 1 and give just as good an explanation of the trend as the linear regression. The question is, do the data in Fig. 1 mean that hospitalization rates have decreased during the study, or is the trend just an artefact of the selection process? If the trial had been randomized then of course this alternative model could not arise since the design would ensure that the coefficient of y in equation (2) would be known to be 0.

The approach of this paper is to extend some simple statistical models to include an additional parameter θ which models the degree of non-randomness in the mechanism generating the data. The special value $\theta = 0$ is the random hypothesis, that the data are *as if* they had resulted from a randomized experiment. We could regard θ as an unknown parameter along with those already in the model and proceed to parametric inference in the usual way, or we could proceed by testing the hypothesis that $\theta = 0$ and then take acceptance of this hypothesis as justification for the standard inference. But both approaches are fraught with difficulties, some of which will be discussed here. The more cautious approach which we adopt is to study inference conditional on a range of different values of θ, and to see how sensitively our conclusions depend on departures of θ from 0. We also study the (profile) log-likelihood for θ after maximizing out the other parameters, to see how much information the data give us about θ. Often this information will be very weak (profile log-likelihood very flat) and we have to entertain a range of inferences given by the range of values of θ which could be considered plausible in the light of the scientific context.

Section 2 discusses a generic model underlying the later sections of the paper and sets out some basic properties and methods including the reasoning behind equation (3) above. A direct application of this model is to item non-response, which is taken up in Section 3. Section 4 shows that the same model leads to an extension of the paired comparison t-test, in which the selection process now corresponds to the allocation of treatment order within pairs. A slight modification of the same model leads to the two-sample t-test in Section 5.1, generalized to two-group comparisons with covariates in Section 5.2. The paper concludes with some practical suggestions and pointers to further work.

There is a large but scattered literature on sample selection issues relating to the topics discussed in the paper, both in the statistical literature and, more prominently, in the econometrics literature. Much statistical literature is concerned with missing data, particularly in sample surveys. Basic aspects are covered in chapter 13 of Cochran (1953). A major reference is Little and Rubin (1987), who give an excellent survey of earlier papers and applications. The model of Section 3 of the paper is discussed in section 11.4 of Little and Rubin (1987). Several other models have also been studied, e.g. Little (1994) and Freedman (1986). Following Fay (1986), Little

(1985a) and Baker and Laird (1988), Chambers and Welsh (1993) set out a rather general theory for analysing categorical survey data with non-response. Wider aspects of selection mechanisms in sampling are discussed in Sugden and Smith (1984) and in many other papers. A central theme is that of ignorability, the conditions under which missing data can, and cannot, be ignored in inference. Explained in terms of equation (2), the selection method is 'informative' (non-ignorable) if the coefficient of y is non-zero. Data are 'missing completely at random' if the coefficients of both x and y are 0 — the presence of missing data can then safely be ignored. If the data are 'missing at random' then the coefficient of y but not x is zero — it is then possible to adjust the analysis for missing values or to impute the missing values by using covariates. These conditions correspond to special cases of the model in Section 3 as we shall explain. Similar issues arise in the analysis of longitudinal data, questions of ignorability now relating to 'drop-outs' during the study — see the recent paper and discussion Diggle and Kenward (1994). A broader setting in inference is the idea of 'coarsening' discussed in Heitjan and Rubin (1991) and Heitjan (1993, 1994).

Although based on a completely different model, the series of papers by Rosenbaum (1987, 1988) and also Rosenbaum and Rubin (1983) and Rosenbaum and Krieger (1990) is closely related to the sensitivity approach which we adopt later in the paper. Rosenbaum's approach is compared with our more model-dependent method in Section 4. Also see Rosenbaum (1995).

Questions of ignorability for missing data are placed in a wider context of comparative studies in the literature on causality. Cox (1992) has discussed some statistical aspects. Holland (1986) explains the basic issues involved and discusses the approaches found in the literature from various disciplines including statistics (much influenced by the pioneering work of Rubin), philosophy, economics and sociology. More recent contributions in the philosophy of science covering both formal and empirical views of causality can be found in Humphreys (1994). The 'fundamental problem of causality' arises when it is impossible to give two treatments simultaneously to the same individual. Given data in which different treatments are given to different individuals, we can only proceed to causal inference by making assumptions, assumptions which, as Holland explains, are in principle untestable. Some of these assumptions correspond to special cases of our model, as pointed out in Section 5.

An application which brings these problems into sharp focus is the evaluation of social programmes, e.g. estimating the effectiveness of an employment training scheme. This is an area where randomized trials would be difficult or impossible, and where reliable methods based on observational data would be particularly valuable. The question of whether the validity achievable by experimental studies (in which subjects are randomized in or out of the programme) can also be achieved by non-experimental studies (in which subjects are self-selected) is debated at length. Several non-experimental estimates of effectiveness have been proposed but, as widely noted, these can give quite different answers in practice. This has lead Barnow (1987), and others, to conclude that only experimental studies can be trusted. The alternative view is put forward in Heckman *et al.* (1987). Heckman and Hotz (1989) reviewed this literature and went on to list a menu of models and specification tests for the analysis of non-experimental data. They claimed that in applications their tests tend to reject the models which lead to estimates that are in conflict with experimental evidence. Our approach is closer to that of Holland's contribution to the discussion of Heckman and Hotz (1989), emphasizing sensitivity of inference to assumptions rather than attempting model choice leading to a definitive analysis.

Gronau (1974) seems to have been the first to formulate the model which we use here (Section 2). See also the commentary on Gronau's paper by Lewis (1974). This was followed by two important papers by Heckman (1976, 1979) and Nelson (1977) which led to the extensive literature on sample selectivity bias in econometrics. A useful introduction and discussion is in the text by Maddala (1983). Heckman's correction for selectivity bias has been very influential in economics, particularly labour economics, and its application was considered almost routine practice among empirical economists in the 1980s. Although the Heckman correction is consistent in the technical sense, it does not always give sensible answers and is now no longer regarded as the panacea for all data selection problems. Heckman's method has also been criticized on theoretical grounds in that it depends sensitively on strong model assumptions (Little, 1985b). We give a brief discussion of the Heckman method in Section 2, but see Idsen and Feaster (1990) for a more detailed review and an extended example. A topical example in the British context is Main and Shelly (1990), who used the method in their evaluation of the UK government's youth training scheme. Many variations of the Heckman approach have been proposed. Stelcner *et al.* (1989) used full maximum likelihood for a model similar to that used in Section 5.2 later, and other approaches include those by Duncan and Leigh (1980), Lee (1982, 1983) and Manski (1989). Closely related is the econometrics literature on tests for exogeneity; see for example Duncan and Leigh (1985) and references therein. Lee and Chesher (1986) and Melino (1982) discussed tests for sample selectivity, which we go on to discuss in Section 3.3.

2. BASIC MODEL

Let y be the response variable of interest, assumed linearly related to covariates x through the standard multiple regression

$$y = \beta^{\mathrm{T}} x + \sigma \epsilon_1. \tag{4}$$

Here, vector x has m components and $x_1 = 1$, so that the first component of vector β is the intercept term. Residual ϵ_1 is standard normal. This main model is supplemented by a 'selection equation'

$$z = \gamma^{\mathrm{T}} x + \epsilon_2. \tag{5}$$

We assume that ϵ_2 is also standard normal, and that (ϵ_1, ϵ_2) is standard bivariate normal with correlation coefficient ρ.

Our two main applications of this model are to missing data (y is only observed if $z > 0$; Section 3) and to comparative trials (a subject is allocated to treatment A if $z > 0$ and to treatment B if $z \leqslant 0$; Sections 4 and 5). We assume that covariates x are fixed and always observed. We never observe the actual value of z, but we always know whether it is positive or negative. In both applications we have observations on the conditional density $f(y|x, z > 0)$, which is

$$\sigma^{-1}\, \Phi^{-1}(\gamma^{\mathrm{T}} x)\, \phi\{\sigma^{-1}(y - \beta^{\mathrm{T}} x)\}\, \Phi\{(1 + \theta^2)^{1/2} \gamma^{\mathrm{T}} x + \theta \sigma^{-1}(y - \beta^{\mathrm{T}} x)\}, \tag{6}$$

where

$$\theta = \frac{\rho}{(1-\rho^2)^{1/2}}.$$

Equation (6) follows from the fact that

$$P(z > 0|x, y) = \Phi\left\{\frac{\gamma^T x + \rho\sigma^{-1}(y - \beta^T x)}{(1-\rho^2)^{1/2}}\right\}. \tag{7}$$

The parameter θ is a convenient reparameterization of ρ, which also has the advantage of improving some of the linear approximations to be developed later. If $\rho = 0$ then expression (6) is exactly the same as the marginal normal distribution of y in equation (4). Analogous equations follow when conditioning on $z \leqslant 0$.

The simplest special case of this model is $\beta = \gamma = 0$ and $\sigma = 1$, i.e. y and z are standard bivariate normal with correlation ρ. Then $f(y|z > 0)$ is $2\,\phi(y)\,\Phi(\theta y)$, the 'skew-normal distribution' studied by Azzalini (1985). This is shown in Fig. 2 for a few different values of ρ. This immediately shows the difficulty in trying to estimate the model from observations on the conditional distribution. Only when ρ is close to -1 or 1 is the conditional density appreciably skewed to the left or right (if $|\rho| = 1$ the conditional density is half-normal). Any attempt to estimate ρ from the shape of the conditional density will therefore depend sensitively on the normality assumptions.

It is easy to show that

$$E(y|x, z > 0) = \beta^T x + \sigma\rho\,\lambda(\gamma^T x) \tag{8}$$

where, as before, λ is the Mills ratio ϕ/Φ. This equation is the basis of Heckman's two-stage estimation procedure (Heckman, 1976, 1979). First estimate γ by noting which cases have $z > 0$. This is a standard probit analysis since

$$P(z > 0|x) = \Phi(\gamma^T x). \tag{9}$$

Use the resulting estimate of γ to form $\lambda(\gamma^T x)$ for each of the cases with $z > 0$, and then take this as an additional covariate in equation (8) and fit by least squares. The coefficient of the additional covariate then gives an estimate of $\rho\sigma$. The conditional variance is

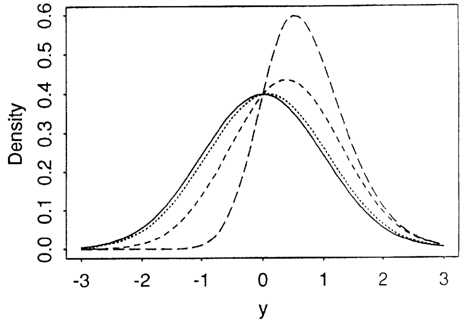

Fig. 2. Conditional distribution of Y given $Z > 0$: ——, $\rho = 0$; ········, $\rho = 0.1$; - - - -, $\rho = 0.5$; — — —, $\rho = 0.9$

$$\mathrm{var}(y|x,\, z>0)=\sigma^2[1-\rho^2\,\lambda(\gamma^{\mathrm{T}}x)\{\gamma^{\mathrm{T}}x+\lambda(\gamma^{\mathrm{T}}x)\}], \tag{10}$$

so, by equating the average value of the right-hand side of equation (10) to the observed residual variance of the second-stage regression, estimates of both ρ and σ can be obtained.

This was the method used in the example in Section 1. Here we identify selection to the new treatment with the event $z > 0$, take $x = (1, \text{date})$, and constrain β_2 to be 0. Equation (1) is the first-stage probit model, and equation (3) is the second-stage regression. Equation (2) is the implied selection mechanism found by substituting the estimated parameters into equation (7).

The Heckman method suffers from several deficiencies as pointed out by Little (1985b) and others. If all the non-intercept components of γ are 0, then $\gamma^{\mathrm{T}}x$ is a constant, duplicating the intercept term already included, and so the second-stage regression fails. Thus components of γ have to be sufficiently large for the non-linearity in $\lambda(\gamma^{\mathrm{T}}x)$ to safeguard the second-stage regression against multicollinearity between $\lambda(\gamma^{\mathrm{T}}x)$ and x. But the method then depends sensitively on the assumed linearity in equation (4). The requirements that the estimates satisfy $|\rho\sigma| \leqslant \sigma$ and $\sigma \geqslant 0$ can be violated. In practice Heckman's method is only useful if restrictions are placed on which covariates enter equations (4) and (5) to avoid the multicollinearity problem, as is done in many applications of the procedure in the economics literature. But again the results depend sensitively on which components of β and γ are constrained to be 0, and different choices could give sharply different estimates of ρ.

The third central moment of $f(y|x,\, z > 0)$, indicating distributional shape, is

$$\rho^3\sigma^3\,\lambda(\gamma^{\mathrm{T}}x)\{(\gamma^{\mathrm{T}}x)^2+3\gamma^{\mathrm{T}}x\,\lambda(\gamma^{\mathrm{T}}x)+2\,\lambda^2(\gamma^{\mathrm{T}}x)-1\}. \tag{11}$$

Comparing equations (8), (10) and (11) we see that the mean (and hence most inferences of interest) depends linearly on ρ, but the variance, and even more so the skewness, depends on ρ much less sensitively, as we have already seen in Fig. 2.

A further reparameterization of ρ is useful in comparing our approach with that of Rosenbaum (1987, 1988), and also as an aid to interpretation. If ρ is positive, values of y with $z > 0$ are likely to be larger than those with $z \leqslant 0$, and vice versa if ρ is negative. We can capture this in terms of a log-odds ratio by comparing the conditional probability that $z > 0$ at the upper quartile of the distribution of y with the probability at the lower quartile of y. From equation (7), this gives the *interquartile log-odds ratio* as

$$\eta=\log\left[\frac{\Phi\{(1+\theta^2)^{1/2}\gamma^{\mathrm{T}}x+0.674\theta\}\;\Phi\{-(1+\theta^2)^{1/2}\gamma^{\mathrm{T}}x+0.674\theta\}}{\Phi\{-(1+\theta^2)^{1/2}\gamma^{\mathrm{T}}x-0.674\theta\}\;\Phi\{(1+\theta^2)^{1/2}\gamma^{\mathrm{T}}x-0.674\theta\}}\right]. \tag{12}$$

For a single summary we suggest replacing $\gamma^{\mathrm{T}}x$ in equations (12) by $\Phi^{-1}(n/N)$, where n is the number of observations out of N with $z > 0$, on the grounds that the expected proportion of cases with $z > 0$ is the average of $\Phi(\gamma^{\mathrm{T}}x)$. Approximating equation (12) for small values of θ then gives

$$\eta\simeq\theta\frac{1.348N^2}{n(N-n)}\,\phi\left\{\Phi^{-1}\left(\frac{n}{N}\right)\right\}. \tag{13}$$

The linear approximation (13) is adequate for many practical purposes.

3. MISSING DATA

A direct application of the model in Section 2 is to item non-response in surveys. Consider a random sample of size N from a large population in which y is the response of interest, assumed to be related to covariates x by equation (4). Some of the y-values are missing, the response process being modelled by the sign of z in equation (5). A latent variable interpretation of the model is to suppose that the trio (x, y, z) exists for all members of the sample but they are not all recorded: vector x is always observed, only the sign of z is observed (since we know which of the ys are missing) and y is observed only if z is positive.

The crucial parameter in the model as far as ignorability is concerned is ρ. If $\rho = 0$ then the data are missing at random, and valid inference about the conditional distribution of y given x can be made from the data on the complete cases. Adjustments for missing data can then be made by using the covariates. If the non-intercept terms in γ, as well as ρ, are 0 then no adjustment is needed and the missing data are missing completely at random. If $\rho \neq 0$ then the missing data are informative or non-ignorable. Often in practice the assumption that $\rho = 0$ cannot be held with conviction, and so we shall be particularly concerned to study the sensitivity of inference to local departures of ρ from 0.

3.1. *Likelihood for Item Non-response*

Suppose that the data are arranged so that (y_i, x_i) are observed for $i = 1, 2, \ldots, n$ but only x_i is observed for $i = n+1, n+2, \ldots, N$. Then the log-likelihood function is

$$L(\beta, \sigma, \gamma, \theta) = -n \log \sigma - \frac{1}{2\sigma^2} \sum_1^n (y_i - \beta^{\mathrm{T}} x_i)^2 + \sum_1^n \log \Phi(u_i) + \sum_{n+1}^N \log \Phi(-\gamma^{\mathrm{T}} x_i) \tag{14}$$

where

$$u_i = (1+\theta^2)^{1/2} \gamma^{\mathrm{T}} x_i + \theta \sigma^{-1} (y_i - \beta^T x_i).$$

Note that, when $\theta = 0$, equation (14) is maximized by the usual estimates of β and σ based on the complete cases only and, for γ, by the vector of coefficients in the probit regression (9). Of these parameters, β is usually the principal object of inference.

It can be shown that for given θ, and when the other parameters are constrained to be solutions of $\partial L/\partial\beta = \partial L/\partial\sigma = \partial L/\partial\gamma = 0$, the Hessian matrix of L is always negative definite, and so the conditional maximum likelihood estimates $\hat{\beta}(\theta)$, $\hat{\sigma}(\theta)$ and $\hat{\gamma}(\theta)$ are uniquely defined for all θ. This enables us to define the profile log-likelihood

$$L^*(\theta) = \max_{\beta,\sigma,\gamma|\theta} \{L(\beta, \sigma, \gamma, \theta)\} = L\{\hat{\beta}(\theta), \hat{\sigma}(\theta), \hat{\gamma}(\theta), \theta\}.$$

It is helpful to separate the intercept terms in the model explicitly by writing $\gamma = (\gamma_1, \gamma_{-1}^{\mathrm{T}})^{\mathrm{T}}$ and $\beta = (\beta_1, \beta_{-1}^{\mathrm{T}})^{\mathrm{T}}$. Write $D^2 = \gamma_{-1}^{\mathrm{T}} \gamma_{-1}$, which describes the strength of the dependence of selection on the covariates.

Aspects of inference are relatively straightforward if $D = 0$. We then find that

$L^{*\prime}(0) = L^{*\prime\prime}(0) = 0$, and so the profile log-likelihood is very flat near $\theta = 0$, regardless of the data. Also

$$L^{*\prime\prime\prime}(0) = K_1 \sum_1^n \left\{ \frac{y_i - \hat{\beta}^{\mathrm{T}}(0) x_i}{\hat{\sigma}(0)} \right\}^3 \tag{15}$$

and

$$L^{*\prime\prime\prime\prime}(0) = K_2 \sum_1^n \left[\left\{ \frac{y_i - \hat{\beta}^{\mathrm{T}}(0) x_i)}{\hat{\sigma}(0)} \right\}^4 - 3 \right], \tag{16}$$

where K_1 and K_2 are simple polynomials in $\hat{\gamma}_1(0)$ and $\lambda\{\hat{\gamma}_1(0)\}$. Thus the local shape of the profile log-likelihood at $\theta = 0$ is determined by the skewness and kurtosis of the distribution of observed residuals. If the model is correct and $\theta = 0$, so that the residuals from the regression of observed y on x are normal, then the first *four* derivatives of L^* are all 0, or approximately so, at $\theta = 0$.

Formulae (15) and (16) suggest that inference for θ is very sensitive to distributional assumptions, and hence to the choice of any transformations before analysis. Later examples show that $L^*(\theta)$ is often quite flat near $\theta = 0$ even when $D \neq 0$, suggesting that the data give little information about local departures of θ from 0.

3.2. *Local Sensitivity to Selection*

Of particular interest is the dependence of $\hat{\beta}(\theta)$ on θ locally to $\theta = 0$. Finding $\hat{\beta}'(0)$ from equation (14) gives

$$\hat{\beta}(\theta) \simeq \left(\sum_1^n x_i x_i^{\mathrm{T}} \right)^{-1} \sum_1^n y_i x_i - \theta\, \hat{\sigma}(0) \left(\sum_1^n x_i x_i^{\mathrm{T}} \right)^{-1} \sum_1^n \lambda\{\hat{\gamma}^{\mathrm{T}}(0) x_i\} x_i. \tag{17}$$

This linear estimate is closely related to Heckman's procedure outlined in Section 2. In equation (8) replace σ by $\hat{\sigma}(0)$, ρ by θ (which are approximately equal when both are close to 0) and γ by $\hat{\gamma}(0)$, the probit estimate. Then equation (17) is just the value of β which minimizes the sum of squared differences between the values of the right-hand side of equation (8) and the observed values y_i.

Again everything simplifies if $D = 0$. For then $\hat{\gamma}_1(0) = \Phi^{-1}(n/N)$ and

$$\hat{\beta}_1'(0) = -\hat{\sigma}(0)\, \lambda\{\hat{\gamma}_1(0)\},$$

which tends to 0 as $n \to N$. The other components of $\hat{\beta}$ are insensitive to values of θ close to 0, as $\hat{\beta}'_{-1}(0) = 0$. We also find that $\hat{\sigma}'(0)$, $\hat{\gamma}_1'(0)$, $\hat{\gamma}_1''(0)$ and $\hat{\beta}''(0)$ are all 0, and so $\hat{\sigma}(\theta) = \hat{\sigma}(0) + O(\theta^2)$, $\hat{\gamma}_1(\theta) = \hat{\gamma}_1(0) + O(\theta^3)$ and the linear approximation (17) is correct to $O(\theta^3)$.

Some insight into what happens when $D \neq 0$ is given in the case when $x = (1, x_{-1}^{\mathrm{T}})^{\mathrm{T}}$ and x_{-1} is multivariate normal across the population. Asymptotically ($N \to \infty$), $\hat{\beta}'_{-1}(0)$ is a vector in the same direction as γ_{-1}. Hence all contrasts between non-intercept components of β in directions orthogonal to γ_{-1} are locally insensitive to departures from $\theta = 0$. Some algebraic manipulation leads to

$$\hat{\beta}'(0) = -\sigma\begin{pmatrix} \lambda(\gamma_1) + \frac{1}{2}D^2\,\lambda''(\gamma_1) \\ \gamma_{-1}[\lambda'(\gamma_1) + D^2\{\lambda(\gamma_1)\,\lambda''(\gamma_1) + \lambda'''(\gamma_1)\}] \end{pmatrix} + O(D^4).$$

If D is small the intercept $\hat{\beta}_1(\theta)$ depends on θ much more sensitively than $\hat{\beta}_{-1}(\theta)$ does.

Rather than attempting to estimate the whole vector β, we may only be interested in a linear combination of its components such as

$$\mu = \beta^{\mathrm{T}}\bar{x},$$

where $\bar{x} = N^{-1}\,\Sigma_1^N\,x_i$. Estimating μ would then provide an estimate of the population mean since we are assuming that the xs are sampled randomly.

Estimating μ by $\bar{y} = n^{-1}\,\Sigma_1^n\,y_i$, completely ignoring the missing data, is only unbiased if $\theta = 0$ and either β_{-1} or γ_{-1} or both are 0. A better estimate, for a given θ, is $\hat{\beta}^{\mathrm{T}}(\theta)\bar{x}$. Another approach is to impute the missing values by estimating their expectations

$$E(y|x,\,z<0) = \beta^{\mathrm{T}}x - \sigma\rho\,\lambda(-\gamma^{\mathrm{T}}x),$$

giving

$$\hat{\mu}(\theta) = N^{-1}\left(\sum_1^n y_i + \sum_{n+1}^N [\hat{\beta}^{\mathrm{T}}(\theta)x_i - \rho\,\hat{\sigma}(\theta)\,\lambda\{-\hat{\gamma}^{\mathrm{T}}(\theta)x_i\}]\right).$$

This estimate is particularly useful when, as is often the case in practice, a non-linear transformation is involved. If we model $y^* = f(y)$ instead of y then the estimate of the mean on the original scale is

$$N^{-1}\left(\sum_1^n y_i + \sum_{n+1}^N f^{-1}[\hat{\beta}^{\mathrm{T}}(\theta)x_i - \rho\,\hat{\sigma}(\theta)\,\lambda\{-\hat{\gamma}^{\mathrm{T}}(\theta)x_i\}]\right). \tag{18}$$

If the proportion of missing data is reasonably small, then the value of expression (18) will not depend sensitively on which particular transformation is used.

The sensitivity of $\hat{\mu}(\theta)$ with respect to θ near 0 is measured by

$$\hat{\mu}'(0) = N^{-1}\sum_{n+1}^N [\hat{\beta}'^{\mathrm{T}}(0)x_i - \hat{\sigma}(0)\,\lambda\{-\hat{\gamma}^{\mathrm{T}}(0)x_i\}].$$

If $D = 0$ this simplifies to

$$\hat{\mu}'(0) = -\hat{\sigma}(0)\,\frac{N}{n}\,\phi\left\{\Phi^{-1}\left(\frac{n}{N}\right)\right\}, \tag{19}$$

which tends to 0 rapidly as n/N tends to 1. Combining equation (19) with equation (13) gives

$$\hat{\mu}(\theta) - \hat{\mu}(0) \simeq -\frac{\hat{\sigma}(0)(N-n)}{1.348N}\,\eta. \tag{20}$$

It can be helpful to standardize differences in $\hat{\mu}$ with respect to the standard deviation of estimation. A crude estimate of this standard deviation is $s/\sqrt{n}$, where s

is the sample standard deviation of the observed ys. This suggests the standardized selectivity correction

$$\{\hat{\mu}(\theta) - \hat{\mu}(0)\}\surd n/s \tag{21}$$

which, for small D, simplifies to approximately

$$-(1 - R^2)^{1/2}\,\frac{(N-n)\surd n}{1.348N}\,\eta, \tag{22}$$

where R is the multiple correlation coefficient between y and x for the complete cases.

All these estimates are functions of θ (or ρ or η). We are very cautious about any proposal to estimate θ by a single numerical value. The Heckman method attempts to do this but is very unreliable, especially if D is not large. Maximum likelihood could be used, but unless the sample size is extremely large the likelihood function is not well behaved. We suggest setting up the calibration of θ in terms of the interquartile log-odds ratio η given in equation (13) and then examining plots of both $L^*(\theta)$ and $\hat{\mu}(\theta)$. If $L^*(\theta)$ attains a clear maximum far from 0, then this indicates that there is either strong selection bias or that the model is inadequate, for example that the population distribution of the residuals ϵ_1 is skewed. If the plot of $\hat{\mu}(\theta)$ shows widely differing estimates over a range of values of η which can be considered plausible in the context of the survey, then this may simply be a warning that the study is flawed and that little useful information about the population mean can be deduced. Plotting expression (21) instead of $\hat{\mu}(\theta)$ can suggest how the extra uncertainty in μ due to uncertainty in θ can be assessed relative to the inherent level of sampling error.

In the context of analysing survey data our model is of course very simple and often we shall be interested in the population distribution of several categorical variables rather than a single continuous variable y. Chambers and Welsh (1993) discussed a comprehensive model for categorical survey data. Their approach is to fit the conditional response probabilities (generalizing our $P(z > 0|x, y)$) by a variety of log-linear models of the kind familiar for ordered categorical data (Agresti, 1984). Although their setting is quite different from ours, and much more complicated, their conclusions are broadly similar — that the data themselves cannot identify any single model for non-ignorable non-response, and that the way forward in practice is to adopt a sensitivity approach involving a range of such models.

3.3. *Coventry Skills Audit*

To illustrate this section we consider some data from a local skills audit in Coventry, UK. Full details of this survey are given in Elias and Owen (1989) and not repeated here. One of the variables of interest is income (pounds per week), which is related to (among other things) sex and age. The $N = 1435$ cases chosen for analysis are all adults in the survey who were known to be in full-time employment, which we assume can be taken to be a random sample of all employed adults in Coventry. Data on sex and age were complete, but values of y were missing in 7.8% of cases, making $n = 1323$. Although the response rate is commendably high, the possibility of selection bias certainly cannot be ruled out: perhaps subjects with an unusually high income may be more likely to refuse to provide information on their income to the survey interviewer.

The fit of the model is illustrated for three different transformations on y: logarithmic, Box–Cox and square root. We take $m = 4$ and the components of x to be 1, sex (coded 0 and 1), age (in years) and $\{\text{age} - \text{mean(age)}\}^2$. Normal plots of the fitted residuals from the least squares regressions fitted to the observed cases suggest that if $\theta = 0$ then all three are reasonable. Estimates of γ_{-1} from equation (9) suggest that D^2 is small (not significantly different from 0), and so the approximations in Sections 3.1 and 3.2 for small D are relevant.

Fig. 3 shows the profile log-likelihoods L^* plotted against ρ—each likelihood has been scaled so that $L^*(0) = 0$. The curves are very flat near $\rho = 0$ as expected, but the point of maximum depends sensitively on the transformation used. By contrast, Fig. 4 displays values of $\hat{\mu}(\theta)$, after allowing for each transformation as in expression (18), showing that the estimates of the population mean for given ρ are much the same. For the Box–Cox transformation, the standardized selectivity correction (21) is plotted against η in Fig. 5, along with the simple linear approximation (22).

For a well-designed and well-executed survey such as this it is implausible that $|\eta|$ would be very large. With an overall rate of about 8%, a fairly extreme possibility might be that the probability of missing data at the lower quartile of y is 4% whereas at the upper quartile it is 12% (three times as large). This would give a log-odds ratio of 1.18. If values of η between say -1.2 and 1.2 were considered plausible, then from Fig. 5 the standardized selectivity correction lies between -2 and 2, directly comparable with the conventional standardized sampling error of ± 2. Roughly, the extra uncertainty could be thought of as doubling the standard error of estimation.

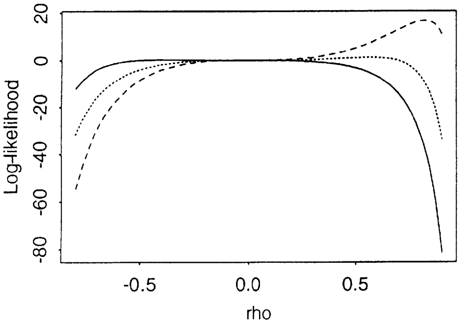

Fig. 3. Profile log-likelihood function for the skills data: ———, log-transformation; ·········, Box–Cox transformation; - - - -, square-root transformation

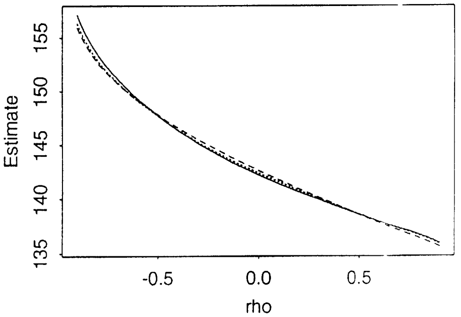

Fig. 4. Estimated average for the skills data: ———, log-transformation; ·········, Box–Cox transformation; - - - -, square-root transformation

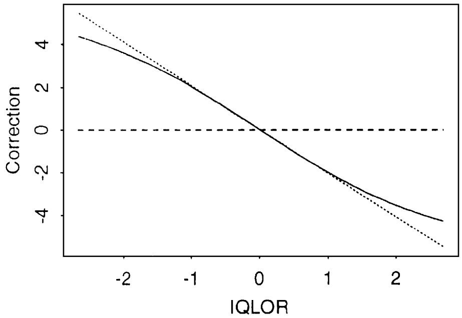

Fig. 5. Standardized selectivity correction

Over this range of values the linear approximation (22) is very adequate and could be used directly, avoiding the need to calculate any complicated estimates explicitly.

These data provide an opportunity to assess the effectiveness of standard asymptotic formulae about the estimation of θ. For the Box–Cox transformation we have $\hat{\theta} = 0.750$ from Fig. 3. The Hessian matrix of equation (14) gives the asymptotic standard error of $\hat{\theta}$ to be 0.178. This was tested by simulating 50 independent samples of the same size from the fitted model, and for each sample determining $\hat{\theta}$ by maximizing L^* numerically. These values averaged 0.515 with a standard deviation of 0.588, three times larger than the nominal asymptotic value. Similar results were obtained for the other two transformations, suggesting that $\hat{\theta}$ is substantially biased and is much more variable than we might expect from asymptotic theory.

As mentioned earlier, the second stage of the Heckman method of allowing for selection bias is to find the coefficient of $\lambda\{\hat{\gamma}^{\mathrm{T}}(0)x\}$ in the right-hand side of equation (8) by ordinary least squares. The usual significance test based on its 't-statistic' is then taken as a test of the null hypothesis of no selection bias. Melino (1982) showed that this regression significance test is exactly the same as the Lagrange multiplier test, essentially the score test of the hypothesis that $\rho = 0$ based on equation (14). This, to quote Melino (1982), 'allows us to deduce that the simple regression test also has desirable asymptotic properties'. Further asymptotic aspects are discussed by Lee and Chesher (1986). These limited bootstrap results, however, suggest that to assume that standard asymptotic theory applies even in relatively large samples is problematical and at worst grossly misleading.

4. PAIRED t-TEST

Consider the classical design in which two treatments A and B are to be applied in random order to N pairs of experimental units. We assume strict treatment additivity, in the sense that for a given experimental unit the response which would result if A were applied differs by a constant μ from the response which would result if B were applied. For a typical pair, let (r_1, r_2) be the responses if B were applied to both members. Then we describe the within-pair variability by supposing that $r_1 - r_2 \sim N(0, \sigma^2)$.

Now apply the treatments in random order, or equivalently let $z \sim N(0, 1)$ and allocate (A, B) if $z \geqslant 0$ and (B, A) if $z < 0$. Then if $z \geqslant 0$ the responses are $(r_1 + \mu, r_2)$

but are $(r_1, r_2+\mu)$ if $z<0$. Hence the treatment difference y (response for A minus response for B) is $\mu+\operatorname{sign}(z)(r_1-r_2)$. The complete model can therefore be written

$$y=\mu+\sigma\operatorname{sign}(z)\epsilon_1,$$
$$z=\epsilon_2.$$

If (ϵ_1, ϵ_2) is bivariate normal, then this is closely related to the general model of Section 2.

The correlation ρ between ϵ_1 and ϵ_2 measures the degree of selection bias in the allocation process. A positive ρ would indicate that the choice of the unit given A is biased in favour of the unit likely to give the higher response, so that the sample mean will tend to overestimate μ. If $\rho<0$ then A would favour the unit likely to give the lower response, giving an underestimate of μ. Properly conducted randomization of course would ensure that $\rho=0$. The size of this allocation bias can also be measured in terms of a log-odds ratio as in Section 2. Comparing the probability of (A, B) when r_1-r_2 is at its upper quartile with the probability of (A, B) when r_1-r_2 is at its lower quartile gives

$$\eta=2\log\left\{\frac{\Phi(0.674\theta)}{\Phi(-0.674\theta)}\right\}. \tag{23}$$

If θ is not too large then $\eta\simeq 2.15\theta$.

It is obvious that the conditional distribution of $\operatorname{sign}(\epsilon_2)\epsilon_1$ given $\epsilon_2\geqslant 0$ is the same as its conditional distribution given $\epsilon_2<0$ and hence must also equal its marginal distribution. Thus the marginal distribution of y here is exactly the same as the conditional distribution of y given $z\geqslant 0$ in the basic model in equations (4) and (5), where we set $m=1$, $\beta_1=\mu$ and $\gamma_1=0$. Properties of y in the paired experiment model can thus be deduced directly from Section 2. In particular

$$E(y)=\mu+\rho\sigma\sqrt{\left(\frac{2}{\pi}\right)},$$

$$\operatorname{var}(y)=\sigma^2\left(1-\frac{2\rho^2}{\pi}\right)$$

and hence if $\mu=0$ and N is large

$$t=\frac{\bar{y}}{\sqrt{\operatorname{var}(\bar{y})}}\sim N\left\{\rho\sqrt{\left(\frac{2N}{\pi-2\rho^2}\right)},1\right\}. \tag{24}$$

In large samples, where estimation error in the variance can be ignored, t is just the usual paired t-statistic for testing the hypothesis that $\mu=0$. The hypothesis is rejected at level α if

$$|t|>z_{\alpha/2}=\Phi^{-1}(1-\alpha/2),$$

and so the asymptotic type I error is

$$\Phi\left\{-z_{\alpha/2}-\rho\sqrt{\left(\frac{2N}{\pi-2\rho^2}\right)}\right\}+\Phi\left\{-z_{\alpha/2}+\rho\sqrt{\left(\frac{2N}{\pi-2\rho^2}\right)}\right\}. \tag{25}$$

Of course this equals α if $\rho=0$. Some values of expression (25) are shown in Fig. 6. This shows that the test is *very* sensitive to departures of ρ from 0—even a small amount of selection bias such as $\rho=0.1$ gives a type I error that is twice as large as its nominal level when $N=100$.

The log-likelihood function follows immediately from Section 3.1 and is

$$L=-N\log\sigma-\frac{1}{2\sigma^2}\sum_1^N(y_i-\mu)^2+\sum_1^N\log\Phi\{\theta\sigma^{-1}(y_i-\mu)\}.$$

As in that section, we find that $L^{*\prime}(0)=L^{*\prime\prime}(0)=0$, and that $L^{*\prime\prime\prime}(0)$ and $L^{*\prime\prime\prime\prime}(0)$ are proportional to the sample skewness and kurtosis of the ys respectively. Again the shape of L^* depends sensitively on the model specification.

If $\hat\mu(\theta)$ is the maximum likelihood estimate of the treatment effect then

$$\hat\mu(\theta)\simeq\bar y-\sqrt{\left(\frac{2}{\pi}\right)}\theta\,\hat\sigma(0). \tag{26}$$

Thus for the standardized deviate

$$t(\theta)=\frac{\hat\mu(\theta)}{\surd\mathrm{var}\{\hat\mu(\theta)\}}$$

we have the simple approximation

$$\begin{aligned}t(\theta)&\simeq t(0)-\sqrt{\left(\frac{2N}{\pi}\right)}\theta\\&\simeq t(0)-0.371\,N^{1/2}\eta.\end{aligned} \tag{27}$$

This allows us to see the sensitivity of the 'significance' of the data to local departures of η from 0.

As an example, take the famous data on heights of cross- and self-fertilized plants

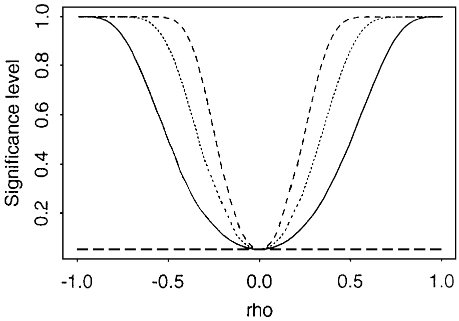

Fig. 6. Type I error of the t-test (nominal 5% level): ———, sample size 20; ·········, sample size 50; - - - -, sample size 100

(Darwin's data) analysed by Fisher (1966) and by many subsequent researchers. There are $N=15$ paired differences as follows (in eighths of an inch):

$$49,\ -67,\ 8,\ 16,\ 6,\ 23,\ 28,\ 41,\ 14,\ 29,\ 56,\ 24,\ 75,\ 60,\ -48.$$

These have $\bar{y}=20.93$ and $s=37.74$ giving $t=2.148$. In this experiment (Darwin, 1876), pairs of seedlings produced from cross- and self-fertilization were grown together so that members of each pair were reared under nearly identical conditions. Here, randomization could mean randomizing the growing positions of each plant within pairs, and allocation bias would be present if, for instance, the crossed seedlings tended to be placed in positions of more sunlight. In that case $\rho \neq 0$ and approximation (27) gives

$$t(\theta) \simeq 2.148 - 1.44\eta. \tag{28}$$

Even a small positive bias brings the value of $t(\theta)$ to below the conventional percentage point. Unless we can be sure that $\rho=0$, the evidence in these data for a real effect is surely considerably weakened.

A graph of the profile log-likelihood L^* is flat near 0 but has a marked cubic shape reaching a maximum near $\rho=-1$. The cubic shape results from the negative skewness of the data caused by the two outliers. Interestingly, a negative value of ρ, which is what the likelihood is indicating at face value, *increases* the size of the t-statistic.

If the t-test used in this analysis is replaced by the Wilcoxon signed rank test, the method of Rosenbaum (1987, 1988) gives an alternative approach to assessing sensitivity to selection bias. Following Rosenbaum (1988), and simplifying and adapting to the notation used here, we assume that the treatment orders are allocated by

$$\log\left\{\frac{P(A,\,B)}{P(B,\,A)}\right\} = \upsilon u, \tag{29}$$

where u is a latent covariate specific to each experimental pair and υ is a parameter, assumed non-negative. Rosenbaum fixes the arbitrary scale of u by supposing that u lies in the finite interval $[-1,\,1]$. (Our u equals the difference between the two unit-specific covariates defined in Rosenbaum (1988).)

The signed rank statistic is

$$T=\sum_1^N q_i I_i$$

where q_i is the rank of $|y_i|$ and I_i is 1 if $y_i>0$ and 0 otherwise. Under the null hypothesis $y=\pm\sigma\epsilon_1$, with the sign determined by the treatment order. Hence

$$P(y>0|\epsilon_1)=\frac{1}{2}\left\{1+\frac{\exp(\upsilon u)-1}{\exp(\upsilon u)+1}\operatorname{sign}(\epsilon_1)\right\}$$

which takes its maximum over u in $[-1,\,1]$ when $u=\operatorname{sign}(\epsilon_1)$ and its minimum when $u=-\operatorname{sign}(\epsilon_1)$. These values of u give the corresponding bounds to the P-value for testing the treatment effect. Equivalently, in large samples, the extremes of the corresponding standardized deviate are $t_{\mathrm{R}}(\pm\upsilon)$, where, for small values of υ,

$$t_{\mathrm{R}}(\upsilon) \simeq t_{\mathrm{R}}(0) - \upsilon\sqrt{\left\{\frac{3N(N+1)}{8(2N+1)}\right\}}.$$

For Darwin's data this is

$$t_{\mathrm{R}}(\upsilon) \simeq 2.045 - 1.70\upsilon. \tag{30}$$

Only a very small positive value of υ renders this 'insignificant'.

The parameter η measures selection bias by comparing the allocation probabilities for two cases (here pairs) at the upper and lower quartile of the variable governing the allocation process. From equation (29), if we compare pairs with covariates u_1 and u_2 the log-odds ratio is $\upsilon(u_1 - u_2)$. If we were to assume that the us are uniformly distributed in $[-1, 1]$ and set u_1 and u_2 at the quartiles $\pm\frac{1}{2}$ then this log-odds ratio is simply υ. Thus, in this sense, both η and υ can be interpreted as log-odds ratios measuring selection bias. Note that approximations (28) and (30) are rather similar. We can also study the distribution of T implied by our model rather than by equation (29), again with very similar results.

A crucial assumption in Rosenbaum's model is the finite range of u. In our model the selection probit (given y) is unbounded, but the selection process is more tightly specified by a fully parametric model. Rosenbaum's method can be applied to various generalizations of the signed rank statistic (Rosenbaum, 1988) and to permutation inference in two-sample problems (Rosenbaum and Krieger, 1990) but would be more complicated for other statistics (e.g. a t-test) and for more general comparisons in the presence of covariates.

5. TWO-SAMPLE COMPARISONS

5.1. *Two-sample t-test with Equal Groups*

A small change to the model of Section 4 leads to the comparison of two independent groups. The model is now

$$y = \mu + \operatorname{sign}(z)\delta + \sigma\epsilon_1,$$
$$z = \epsilon_2.$$

The interpretation is that we allocate treatment A if $z \geqslant 0$, in which case response y has mean $\mu + \delta$, but allocate treatment B if $z < 0$, in which case y has mean $\mu - \delta$. We suppose that, of N experimental units, n are allocated to A giving data $y_1, y_2, \ldots, y_n$, and the remaining $N - n$ are allocated to B with data $y_{n+1}, y_{n+2}, \ldots, y_N$. As n is ancillary, conditioning on n will not change the likelihood function.

As before, the parameter ρ measures the degree of selectivity bias in the treatment allocation. Imagine giving A to all N cases, and locate the lower and upper quartiles of the responses. Then the interquartile log-odds ratio η compares the probabilities that these two cases will be allocated treatment A rather than treatment B. The value of η is given by formula (23), exactly as before. A fully randomized design would of course ensure that $\rho = 0$. Conversely, $\rho = 0$ corresponds to the assumption of 'independence' in Holland's (1986) terminology.

In the obvious notation, the usual estimate of the treatment effect δ is

$$\hat{\delta} = \tfrac{1}{2}(\bar{y}_1 - \bar{y}_2),$$

and so

$$E(\hat{\delta}) = \delta + \rho\sigma\sqrt{\left(\frac{2}{\pi}\right)}$$

and

$$\operatorname{var}(\hat{\delta}) = \frac{\sigma^2}{N}\left(1 - \frac{2\rho^2}{\pi}\right) + O(N^{-2}).$$

Hence the asymptotic null distribution of the two-sample *t*-statistic is exactly the same as expression (24), and so the asymptotic type I error of the *t*-test is the same as in equation (25) and Fig. 6. The same conclusion follows, that the *t*-test is *very* sensitive to non-randomness in the treatment allocation.

The log-likelihood is

$$-N \log \sigma - \frac{1}{2\sigma^2}\sum_1^n (y_i - \mu - \delta)^2 + \sum_1^n \log \Phi\{\theta\sigma^{-1}(y_i - \mu - \delta)\}$$
$$- \frac{1}{2\sigma^2}\sum_{n+1}^N (y_i - \mu + \delta)^2 + \sum_{n+1}^N \log \Phi\{-\theta\sigma^{-1}(y_i - \mu + \delta)\}.$$

Trivial modifications to the earlier calculations show that $L^{*\prime}(0)$ and $L^{*\prime\prime}(0)$ are 0, but that $L^{*\prime\prime\prime}(0)$ is proportional to the *difference* between the skewness statistics for the two samples, suggesting that the local shape of L^* near $\theta = 0$ is less sensitive to data transformations than in the single-sample problem. From expression (11) we see that the skewness coefficients of the two samples are equal but of opposite sign. Hence, if we accept the assumption of additivity in the treatment effect, then the difference in skewness of the data between the two groups does provide some information about θ, and we would expect this to stay roughly the same over a modest range of different data transformations.

Evaluating the derivatives of the maximum likelihood estimate of δ with respect to θ shows that

$$\hat{\delta}(\theta) = \frac{1}{2}(\bar{y}_1 - \bar{y}_2) - \sqrt{\left(\frac{2}{\pi}\right)}\theta\,\hat{\sigma}(0) + O(\theta^3).$$

The local selectivity correction to $\hat{\delta}$ is exactly the same as in approximation (26) and so the '*t*-statistic' for testing the treatment difference also leads to approximation (27), just as in the paired case.

We illustrate these points by applying the model to some data from the national hearing survey, described in Davis (1995). One question of interest is the extent to which exposure to excessive noise at the workplace contributes to the impairment of hearing, after allowing for the confounding effect of age. For our first sample we take the data for male manual workers between the ages of 50 and 60 years who have been exposed to high levels of occupational noise, and we compare these with a control sample of similar age and occupation but who have been exposed to at most low levels of occupational noise. Not surprisingly those in the exposed sample show poorer levels of hearing than the controls, but the data are not randomized and non-zero values of ρ should at least be considered. (Perhaps workers with poor hearing are less concerned about noise and so are more likely to accept jobs in noisy factories.) We take y to be the average hearing threshold (volume of sound that can just be heard) for pure tones at frequencies 1, 2 and 3 kHz and, following Longford (1993), use the transformation $\log(y+20)$. We have $n = N - n = 78$. Statistical aspects of these data and other data in the survey are discussed in Bowater *et al.* (1996).

Normal plots of $\log(y+20)$ for the two groups show a clear difference in mean but with roughly the same variance. The plot of $L^*(\theta)$ is quite flat, corresponding to the fact that the normal plots just referred to show little difference in skewness. For these data $t(0) = 2.40$, indicating a nominally significant noise exposure effect. But from approximation (27)

$$t(\theta) \simeq 2.40 - 4.66\eta,$$

and so only a very small positive selection bias is sufficient to undermine this conclusion ($t(\theta) < 2$ if $\eta > 0.09$ or $\rho > 0.04$). Again, a rather marginal level of significance based on the usual analysis needs to be viewed with considerable caution.

5.2. *Two-group Comparisons with Covariance Adjustment*

Comparing two groups by using observational data usually involves an adjustment for relevant covariates x. For this the model of Section 5.1 extends to

$$y = \beta^{\mathrm{T}} x + \mathrm{sign}(z)\delta + \sigma\epsilon_1,$$
$$z = \gamma^{\mathrm{T}} x + \epsilon_2.$$

The log-likelihood is

$$-N\log\sigma - \frac{1}{2\sigma^2}\sum_1^n (y_i - \beta^{\mathrm{T}} x_i - \delta)^2 + \sum_1^n \log\Phi(u_i)$$
$$-\frac{1}{2\sigma^2}\sum_{n+1}^N (y_i - \beta^{\mathrm{T}} x_i + \delta)^2 + \sum_{n+1}^N \log\Phi(v_i)$$

where now

$$u_i = (1+\theta^2)^{1/2}\gamma^{\mathrm{T}} x_i + \theta\sigma^{-1}(y_i - \beta^{\mathrm{T}} x_i - \delta)$$

and

$$v_i = -(1+\theta^2)^{1/2}\gamma^{\mathrm{T}} x_i - \theta\sigma^{-1}(y_i - \beta^{\mathrm{T}} x_i + \delta).$$

This is very similar to the likelihood (14) and so most of the material in Section 3 adapts easily to cover the two-sample case. For example, if $\hat{\gamma}_{-1}(0) = 0$, we find

$$\hat{\delta}'(0) = -\frac{\hat{\sigma}(0)}{2}[\lambda\{\hat{\gamma}_1(0)\} + \lambda\{-\hat{\gamma}_1(0)\}], \tag{31}$$

$\hat{\beta}'_{-1}(0) = 0$ and $\hat{\beta}'_1(0)$ is the same as equation (31) but with the plus sign changed to minus. As in Section 3.2, we expect the major effect of selection bias to be in the intercept terms (here both δ and β_1) rather than the regression coefficients β_{-1}.

Approximating $\hat{\gamma}_1(0)$ in equation (31) by $\Phi^{-1}(n/N)$ as before, we have for small D

$$\hat{\delta}'(0) \simeq -\hat{\sigma}(0) \frac{N^2\, \phi\{\Phi^{-1}(n/N)\}}{2n(N-n)}.$$

Combining this with approximation (13) gives

$$\hat{\delta}(\theta) - \hat{\delta}(0) \simeq -\frac{\eta\, \hat{\sigma}(0)}{2.696}. \tag{32}$$

Thus if $t(\theta)$ is the usual asymptotic t-statistic $\hat{\delta}(\theta)/\surd \mathrm{var}\{\hat{\delta}(\theta)\}$ then for small θ and D we have

$$t(\theta) - t(0) \simeq -\frac{\eta s_{\mathrm{res}}}{2.696 s_\delta} \tag{33}$$

where $s_{\mathrm{res}} = \hat{\sigma}(0)$ is the least squares residual standard deviation and s_δ is the standard error of $\hat{\delta}(0)$. Even when θ and D are not particularly small, equation (33) provides a rough guide to how much the inference about treatment difference is affected by selection bias. No special calculations are needed since both standard deviations in approximation (33) are obtained by ordinary analysis of covariance.

Finally, we return to the medical example discussed in Section 1. Here there are two groups of patients with kidney disease, 123 patients on ambulatory peritoneal dialysis (treatment A) and a control group of 121 patients on haemodialysis (treatment B). The allocation to treatment was non-random, and as noted in Section 1 there was a steady increase in the proportion of patients assigned to treatment A as the study progressed. There was also a tendency for patients assigned to A to be older and less fit. We follow Burton and Wells (1989) by taking the following list of relevant covariates: date of commencement of therapy, age and powers of age up to a cubic term (the relationship between hospitalization rate and age is U shaped) and five binary covariates measuring the patient's initial clinical condition.

For these data, the slope of L^* is noticeably non-zero at $\theta = 0$, reflecting the fact that D is not small in this example (treatment assignments are strongly covariate dependent). However, the maximum value of $L^*(\theta)$ is less than one unit higher than $L^*(0)$. Here, $t(0) = 8.68$, and a plot of $t(\theta)$ shows that η would need to be as large as 1.3 ($\rho = 0.52$) before the treatment difference ceases to be significant at the nominal 5% level. This would mean an interquartile odds ratio of $\exp 1.3 = 3.7$, or, as there are roughly equal numbers for the two treatments, a probability of around $\frac{1}{3}$ that patients with y at the lower quartile are assigned to A, compared with a probability of around $\frac{2}{3}$ for patients at the upper quartile. This degree of selection bias is obviously possible but seems rather extreme. We would suspect that if there was this degree of bias in the allocation process then this would at least be commented on by the researchers involved, or else would be seized on by others to discredit the study. We would have arrived at essentially the same conclusion had we just used approximation (33) instead of calculating $t(\theta)$ exactly.

6. CONCLUSION

A recurring theme in the paper is that great caution is needed in making inferences about θ, the parameter which reflects the degree of non-randomness in the sample.

Often $L^*(\theta)$ is very flat, indicating that the data provide little information about sample selection. In fact if the likelihood suggests a clear inference, then this can be as much concerned with model misspecification as it is with selection bias. A second theme, echoing several of the papers cited earlier, is that if data are analysed by conventional methods which assume that $\theta = 0$, but in fact selection bias is present, then the conclusions can be grossly misleading. Even what may appear to be a small amount of bias in the sampling can have a substantial effect on inference. We believe that at least some kind of sensitivity analysis is essential when analysing observational data. Allowing for selection bias increases the uncertainty in estimation and increases type I errors of significance tests.

We suggest that the linear approximations in terms of the log-odds parameter η (related to θ by approximation (13)) which have been developed in this paper give a simple and useful way of assessing local sensitivity to selection bias. Thus, for any statistic T of interest, we calculate the sensitivity multiplier A such that

$$T(\theta) \simeq T(0) + A\eta.$$

For the cases considered, A is given in expressions (20), (22), (27), (32) and (33). In each case the calculation of A is trivial and there is no need to fit any selectivity models explicitly. If T is a standardized test statistic then calculating the size of $|\eta|$ needed to render $|T(\theta)|$ less than some conventional percentage point gives a rough guide to how much non-randomness would be needed to explain away the effect being tested. If T is a parameter estimate, and the context of the data suggests that $|\eta|$ could reasonably be assumed to be less than some η_0, then the limits $T(0) \pm A\eta_0$ could be thought of as adding extra uncertainty to the usual sampling confidence limits for $T(0)$. Determining η_0 in practice can be little more than an 'order of magnitude' guess—for the data in Section 3.3, for example, some value between 1 and 2 would seem to be sensible.

There are many unanswered questions, and in any case the material of Sections 3–5 only amounts to a few special cases of the model of Section 2. In Section 3, for example, only *item* non-response is considered, but in practice there will often also be *unit* non-response in which values of both y and x are missing. The model assumes that $P(z > 0|x, y)$ is monotonic in y, and this may need to be generalized. Data from survey recalls may provide extra information on both γ and ρ.

In Sections 4 and 5 we have made strong additivity assumptions. Allowing for unequal residual variances and treatment–covariate interactions is relatively straightforward. Analysis-of-covariance models for more than two treatment levels would be of interest. For ordered levels a vector of thresholds for z could be used. For categorical levels (no natural ordering), z could be a vector, perhaps with some simplified covariance structure. Another important generalization would be to binary or categorical responses.

At a more technical level, the profile likelihood for θ may be misleading if m is large relative to N. The above discussion of η_0 assumes that we have some informal prior information about ρ, but we have avoided taking it into account in any formal way. Clearly a full Bayesian analysis is possible, at least numerically. Sensitivity to prior assumptions would be an important issue and could provide new insights into the nature of these models.

Normality of residuals has been assumed throughout. A more general model is possible on the lines of the analysis of Darwin's data given in Box and Tiao (1962). This would involve first transforming ϵ_1 before assuming joint normality with ϵ_2.

Diagnostics for these models have not been discussed. For a fixed $\theta \neq 0$, a residual *Q–Q*-plot can be obtained by using an algorithm for the bivariate normal distribution function. Diagnostics for assumptions invariant to θ would be useful; for example in Section 5.1 the two distributions of y are mirror images of each other, up to an additive constant.

ACKNOWLEDGEMENTS

We are indebted to Dr Stan Siebert, Dr Peter Elias, Dr Adrian Davis and Dr Paul Burton for stimulating our interest in these problems and for giving us access to their data. One of us (HGL) was supported by a research grant (GR/H51590) under the Complex Stochastic Systems Programme of the former Science and Engineering Research Council (now the Engineering and Physical Sciences Research Council). We are grateful to the referees for many helpful comments.

REFERENCES

Agresti, A. (1984) *Analysis of Ordinal Categorical Data*. New York: Wiley.

Azzalini, A. (1985) A class of distributions which includes the normal ones. *Scand. J. Statist.*, **12**, 171–178.

Baker, S. G. and Laird, N. M. (1988) Regression analysis for categorical survey variables with outcome subject to nonignorable nonresponse. *J. Am. Statist. Ass.*, **83**, 62–69.

Barnow, B. (1987) The impact of CETA programs on earnings: a review of the literature. *J. Hum. Resour.*, **22**, 157–193.

Bowater, R. J., Copas, J. B., Machado, O. A. and Davis, A. C. (1996) Hearing impairment and the log-normal distribution. *Appl. Statist.*, **45**, 203–217.

Box, G. E. P. and Tiao, G. C. (1962) A further look at robustness via Bayes's theorem. *Biometrika*, **49**, 419–432.

Burton, P. R. and Wells, J. (1989) A selection adjusted comparison of hospitalization on continuous ambulatory peritoneal dialysis and haemodialysis. *J. Clin. Epidem.*, **42**, 531–539.

Chambers, R. L. and Welsh, A. H. (1993) Log-linear models for survey data with non-ignorable non-response. *J. R. Statist. Soc.* B, **55**, 157–170.

Cochran, W. G. (1953) *Sampling Techniques*, 2nd edn. New York: Wiley.

Cox, D. R. (1992) Causality: some statistical aspects. *J. R. Statist. Soc.* A, **155**, 291–301.

Darwin, C. (1876) *The Effect of Cross- and Self-fertilization in the Vegetable Kingdom*. London: Murray.

Davis, A. C. (1995) *Hearing in Adults*. London: Whurr.

Diggle, P. and Kenward, M. G. (1994) Informative drop-out in longitudinal data analysis (with discussion). *Appl. Statist.*, **43**, 49–93.

Duncan, G. M. and Leigh, D. E. (1980) Wage determination in the union and nonunion sectors: a sample selectivity approach. *Industrl Lab. Relatns Rev.*, **34**, 24–34.

——(1985) The endogeneity of union status: an empirical test. *J. Lab. Econ.*, **3**, 385–402.

Elias, P. and Owen, D. (1989) *People and Skills in Coventry*. Coventry: City of Coventry Press.

Fay, R. E. (1986) Causal models for patterns of non-response. *J. Am. Statist. Ass.*, **81**, 354–365.

Fisher, R. A. (1966) *Design of Experiments*, 8th edn. Edinburgh: Oliver and Boyd.

Freedman, D. A. (1986) A case study in nonresponse: plaintiff vs. California State Board of Equalization (with discussion). *J. Bus. Econ. Statist.*, **4**, 123–127.

Gronau, R. (1974) Wage comparisons: a selectivity bias. *J. Polit. Econ.*, **82**, 1119–1143.

Heckman, J. J. (1976) The common structure of statistical models of truncation, sample selection and limited dependent variables, and a simple estimator for such models. *Ann. Econ. Socl Measmnt*, **5**, 475–492.

——(1979) Sample selection bias as a specification error. *Econometrica*, **47**, 153–161.

Heckman, J. J. and Hotz, V. J. (1989) Choosing among alternative non-experimental methods for estimating the impact of social programs: the case of manpower training. *J. Am. Statist. Ass.*, **84**, 862–880.

Heckman, J. J., Hotz, V. J. and Dabos, M. (1987) Do we need experimental data to evaluate the impact of manpower training on earnings? *Eval. Rev.*, **11**, 395–427.

Heitjan, D. F. (1993) Ignorability and coarse data: some biomedical examples. *Biometrics*, **49**, 1099–1109.

——— (1994) Ignorability in general incomplete-data models. *Biometrika*, **81**, 701–708.

Heitjan, D. F. and Rubin, D. B. (1991) Ignorability and coarse data. *Ann. Statist.*, **19**, 2244–2253.

Holland, P. (1986) Statistics and causal inference (with discussion). *J. Am. Statist. Ass.*, **81**, 945–970.

Humphreys, P. (ed.) (1994) *Patrick Suppes: Scientific Philosopher*, vol. 1, *Probability and Probabilistic Causality*. Dordrecht: Kluwer.

Idson, T. L. and Feaster, D. J. (1990) A selectivity model of employer-size wage differentials. *J. Lab. Econ.*, **8**, 99–122.

Lee, L.-F. (1982) Some approaches to the correction of selectivity bias. *Rev. Econ. Stud.*, **44**, 355–372.

——— (1983) Generalized econometric models with selectivity. *Econometrica*, **51**, 507–512.

Lee, L.-F. and Chesher, A. (1986) Specification testing when score test statistics are identically zero. *J. Econometr.*, **31**, 121–149.

Lewis, H. G. (1974) Comments on selectivity biases in wage comparisons. *J. Polit. Econ.*, **82**, 1145–1155.

Little, R. J. A. (1985a) Nonresponse adjustments in longitudinal surveys: models for categorical data. *Bull. Int. Statist. Inst.*, **15**, no. 1, 1–15.

——— (1985b) A note about models for selectivity bias. *Econometrica*, **53**, 1469–1474.

——— (1994) A class of pattern-mixture models for normal incomplete data. *Biometrika*, **81**, 471–484.

Little, R. J. A. and Rubin, D. A. (1987) *Statistical Analysis with Missing Data*. New York: Wiley.

Longford, N. T. (1993) *Random Coefficient Models*. Oxford: Oxford University Press.

Maddala, G. S. (1983) *Limited-dependent and Qualitative Variables in Econometrics*. Cambridge: Cambridge University Press.

Main, B. G. M. and Shelly, M. A. (1990) The effectiveness of the Youth Training Scheme as a manpower policy. *Economica*, **57**, 495–514.

Manski, C. F. (1989) Anatomy of the selection problem. *J. Hum. Resour.*, **24**, 343–360.

Melino, A. (1982) Testing for selection bias. *Rev. Econ. Stud.*, **49**, 151–153.

Nelson, F. D. (1977) Censored regression models with unobserved stochastic censoring thresholds. *J. Econometr.*, **6**, 309–327.

Rosenbaum, P. R. (1987) Sensitivity analysis for certain permutation inferences in matched observational studies. *Biometrika*, **74**, 13–26.

——— (1988) Sensitivity analysis for matching with multiple controls. *Biometrika*, **75**, 577–581.

——— (1995) *Observational Studies*. New York: Springer.

Rosenbaum, P. R. and Krieger, A. M. (1990) Sensitivity of two-sample permutation inferences in observational studies. *J. Am. Statist. Ass.*, **85**, 493–498.

Rosenbaum, P. R. and Rubin, D. B. (1983) The central role of the propensity score in observational studies for causal effects. *Biometrika*, **70**, 41–56.

Stelcner, M., van der Gaag, J. and Vijverberg, W. (1989) A switching regression model of public–private sector wage differentials in Peru: 1985–86. *J. Hum. Resour.*, **23**, 545–559.

Sugden, R. A. and Smith, T. M. F. (1984) Ignorable and informative designs in survey sampling inference. *Biometrika*, **71**, 495–506.

DISCUSSION OF THE PAPER BY COPAS AND LI

C. J. Skinner (University of Southampton): The authors have tackled a particularly thorny problem, which has proved difficult to resolve. They are to be congratulated for setting out a clear way forward, providing a fairly general approach to modifying standard procedures to allow for selection and deriving some remarkably simple adjustments.

There is a wide range of observational settings where selection mechanisms are unknown. In these cases it is common to suppose that selection is non-informative given some choice of observed covariates x, i.e. that selection is conditionally independent of y given x. I prefer the term *non-informative* to *random* since unequal probability randomized sampling schemes are often informative with respect to arbitrary y and x.

Non-informative assumptions may be adequate if ‘appropriate’ covariate information is collected (Rubin *et al.*, 1995) and it is thus important to collect such information if possible. Nevertheless, there will often be reasons to doubt a non-informative assumption. For example, item non-response may be intrinsically informative by depending directly on the value taken by a respondent on an item, via reasons such as the sensitivity or difficulty of answering the question.

The basic problem with allowing for informative selection is that, in a certain sense, the data carry no information about the degree of informativeness. Thus, if we take the item non-response case, the informativeness of selection reflects the difference between the distributions of y given x among respondents and among non-respondents but the data carry no ‘direct’ information about the latter distribution.

The authors refer to econometric approaches which explicitly allow for informative selection, through modelling assumptions. In many applications prior knowledge may not be easily representable in this form and the authors propose instead to represent assumptions about the degree of informativeness more directly by specifying the parameter ρ or its transformed versions θ and η.

The econometric approach has been criticized for being sensitive to model misspecification but the nature of this sensitivity is I think subtle and the paper sheds some interesting light on this issue. In a somewhat paradoxical way the weakening of the model assumptions from restricted to unrestricted ρ appears to increase sensitivity (compare Figs 3 and 4). I find the evidence of insensitivity to Box–Cox transformation in Fig. 4 particularly interesting and would welcome the authors’ comments on any theoretical reasons or further empirical evidence for this finding. This has a bearing on whether the authors’ sensitivity analysis is sufficiently realistic.

To consider further the sensitivity of the unrestricted ρ approach for the non-response case, suppose that $x = (x_1^T x_2^T)^T$, $\gamma = (\gamma_1^T \gamma_2^T)^T$and that the ‘model constraint’ $\gamma_2 = 0$ is applied. In large samples a fitting procedure searches for ρ which minimizes the ‘discrepancy’ between the ‘observed’ distribution $f_n(x_2|x_1)$ for non-respondents and the mixture

$$\int f_r(x_2|y,\, x_1)\, f_n(y|x_1;\, \rho)\, dy$$

of the observed distribution $f_r(x_2|y,\, x_1)$ for respondents with respect to the distribution $f_n(y|x_1;\, \rho)$ implied for non-respondents by the model. If the model is misspecified so that $f_n(x_2|x_1)$ cannot be obtained from $f_r(x_2|y,\, x_1)$ by mixing on y as above then the estimated distribution of y for non-respondents may be extreme and severe adjustment effects obtained. For analogous discrete variable models, Peter Smith and I have observed this empirically, finding that the estimated distribution of y for non-respondents may often be on the boundary, in the sense of Baker and Laird (1988). If ρ is fixed this effect is avoided.

The most innovative technical contribution of the paper is the use of small θ asymptotics which has some interesting parallels with the use of small measurement error asymptotics (e.g. Stefanski (1985)). It is a natural approach to producing first adjustments to standard procedures which ignore informative selection or measurement error and the authors produce some remarkably simple adjustments such as that in expression (32). The linear approximation provides simply the attractive inverse statements about how informative selection needs to be to make tests insignificant. I am intrigued by the idea of inflating standard errors to reflect non-response effects. This is analogous to the use of design effects to reflect complex sampling designs. Potential disadvantages of the approximation are that differentiation is needed for each new application and that the approximation may simply be unsatisfactory. Likelihood-based inference about β given θ does not suffer these problems and will often be computationally straightforward.

As the authors show, the choice of η is often crucial. Thus evidence on its magnitude should be sought whenever possible. Some analogies can be drawn with the measurement error problem, where validation studies and repeated measurements may provide evidence.

In conclusion, although the authors’ methods will not always provide comfortably precise inferences, they have considerable potential for clarifying what information can be extracted from observational data and what further information may be required from elsewhere. I believe that this paper makes an outstanding contribution to an important subject and it gives me great pleasure to propose the vote of thanks.

Andrew Chesher (University of Bristol): I enjoyed reading this paper. It raises issues that can be crucial, particularly when drawing inferences from data generated by thinking and reacting agents, as we do in much work with social, economic and medical data.

When there is 'non-random' sampling in the sense used in this paper, data are realizations from the density

$$\begin{aligned} f_+(y|x) &= f(y|x, Z > 0) \\ &= f(y|x)\, S(y, x), \end{aligned}$$

where

$$S(y, x) = \frac{P[Z > 0|x, y]}{P[Z > 0|x]},$$

rather than from the target density $f(y|x)$. We do have 'random sampling', but from $f_+(y|x)$ rather than from $f(y|x)$. Statistical procedures will reflect features of $f_+(y|x)$ rather than the target density and there is then the possibility of misspecification and misinterpretation.

Example 1

Regression functions under $f_+(y|x)$ and $f(y|x)$ are related by

$$E_+[Y|X = x] = E[Y\, S(Y, x)|X = x]$$

where E_+ and E denote expectation with respect to respectively f_+ and f. When $E[Y|X = x]$ is linear in x, $E_+[Y|X = x]$ will usually be non-linear, creating a potential for misspecification. However, a non-parametric regression estimator can make a good estimate of this regression function. There is a problem if we mistakenly interpret the estimate that it produces as an estimate of the regression function of Y on X for the target distribution or population.

The paper focuses on the *bivariate Gaussian* version of this model in which $S(y, x)$ is a ratio of normal distribution functions. In few applications involving non-aggregated data can a strong argument be made for joint normality so we might be concerned that the information about sensitivity conveyed by the techniques set out is itself sensitive to assumptions concerning distributional shape. It is interesting to ask what limits there are to the distortion induced by $S(y, x)$ under plausible assumptions about the target and selection processes, such as might flow naturally from the theory of the subject in which the application is set.

One approach is to ask what can be discovered about the target density under less detailed assumptions. The semiparametrics literature in econometrics surveyed by Powell (1994) has some useful results. For example Chamberlain (1986) showed that in the model used by Copas and Li, in the *absence* of the bivariate Gaussian assumption, identification of the coefficients in the regression of Y on X associated with $f(y|x)$ requires that a covariate figuring in the regression function of Z on X be *a priori* excluded from $E[Y|X = x]$. It follows that the 'small D' results (approximately *no* non-constant regressors in $E[Z|X = x]$) deals with a near semiparametrically unidentifiable case which we might expect to be challenging for conventional methods. Though $\theta = 0$ ($f_+ = f$) is an attractive point to expand around if we desire simple approximations it may be less interesting than other points and, because of the special properties of the likelihood function there, may give an atypical view of the sensitivity of inference to variations in θ.

First-order asymptotic approximations may be poor in the case considered here and the near lack of semiparametric identifiability may be a contributory cause, but much more evidence than is presented here is required before general conclusions can be drawn. The authors use the observed Hessian of the log-likelihood to produce estimated asymptotic standard errors. The choice of information matrix estimator can be important in determining the quality of approximation to the size of tests (Chesher and Spady, 1991) and the relatively inefficient observed Hessian is not obviously a leading contender. Where distributional assumptions may be in doubt a robust variance estimator such as the 'sandwich estimator' (Efron and Tibshirani, 1993) can produce a better result.

To obtain trustworthy estimates we must rely on more than distributional shape assumptions to achieve identification. Here statistics is not enough and information from the application is essential. Good examples arise in labour economics, the area that spawned Heckman's original work. Consider the problem of estimating models of labour force participation when wages are only available for those who choose and are able to obtain paid employment. Labour economic theory suggests that a person's choice to work is based on a comparison of the valuation of the marginal hour of non-work time that would be sacrificed if a work opportunity were to be taken and the wage that would then be received. So the model for selection into the labour force (and so for selection into a sample of wage earners) includes covariates characterizing the home environment of the potential worker (e.g. income from assets and other household earners) that are naturally excluded from the model for the wage received by potential labour force participants. Identification and estimation of interesting parameters is possible without reliance on assumptions about shapes of distributions. In contrast, in the pure missing data problem posed in the analysis of the Coventry skills audit data there is no theory (at least none advanced) to generate exclusion restrictions. The result is fragile estimates relying to a large extent on assumptions of symmetry in the underlying distributions and driven by observed departures from symmetry in the data.

It gives me great pleasure to second the vote of thanks to Professor Copas and Dr Li for their stimulating paper.

The vote of thanks was passed by acclamation.

David J. Hand (The Open University, Milton Keynes): It gave me great pleasure to read this paper, which deals with a crucial but underexplored topic. As the authors show, ignoring the selection issues can lead to mistaken conclusions and I would like to congratulate them on the significant progress that they have made in tackling these problems.

Situations of the kind explored by Professor Copas and Dr Li are ubiquitous. One such, with which I have been involved, is the following.

Statistical methods are used in consumer banking to construct classification rules to identify people who are at high risk of defaulting on their repayments. Ideally such rules would be based on a random sample from the application population, for each member of which the application details and the true good or bad class would be available. In practice, however, the design information usually includes application form details on all previous applicants, but true good–bad status on only those who were classified as good by an earlier classification rule. The question then arises whether the information which is available about those classified as bad by the earlier rule can help in the construction of an improved rule. Techniques of *reject inference* used in the credit industry purport to do this (Hand and Henley, 1993, 1994).

If the set of predictor variables used for the original rule includes only variables which are included as predictors for the proposed new rule then z is a deterministic function of x. In this case $z \leqslant 0$ corresponds to certain regions of the x-space so that for a given x the conditional distribution of y is either observed completely or is not observed at all.

Conversely, if the set of predictor variables used for the original rule is a superset of those used for the proposed new rule then the case of Copas and Li applies. However, in this case it is likely that the original rule, being based on a superset of variables, will have a superior performance to that of the proposed new rule. Moreover, in the credit scoring context, where much of the data are categorical, assumptions of normality for z would be risky.

In these credit scoring applications, extrapolation plays a key role. The accuracy of such extrapolation clearly depends on the validity of the assumed model for y. Could the authors say anything further about the effect of model misspecification on their conclusions?

Gillian M. Raab (Napier University, Edinburgh): I would like to congratulate the authors on this paper, which clarifies the properties and problems of a class of estimators based on the estimation of selectivity parameters. Their simple expressions for standardized selectivity corrections offer a safer route than that of using model-based estimators when assumptions cannot be checked.

A Bayesian analysis of these problems would now be straightforward using numerical methods. I would like to offer suggestions for this in the context of the example of missing data discussed in Section 3. Prior information is implicit in the paper, in terms of both the assumptions about the selectivity parameter and the parametric assumptions of the distributional form of the residuals for the regression model.

In discussing plausible values for η in relation to the Coventry skills audit data, it is suggested that for a *well-conducted* survey it is implausible that η could be very large. But, if other causes of non-response are removed by good fieldwork, it might be precisely the non-ignorable non-response that remains, and η may be substantial. It would be preferable to base the prior for η on follow-up studies of non-responders. It is also likely that η might vary with the X. A recent example would be data on the number of sexual partners in the National Survey of Sexual Attitudes and Lifestyles. Wadsworth *et al.* (1996) have presented a sensitivity analysis with differential response biases by gender. Extending the current model to cope with this would be straightforward.

Using parametric assumptions about the conditional distribution of Y as prior information is the least satisfactory aspect of the current approach. Close to the maximum, in an essentially flat region, the shape of the profile likelihood is sensitive to the distribution assumed for the residuals. A Bayesian solution to this would be to assign a fairly vague prior to the Box–Cox parameter. Integrating over this should flatten out the local bumps. As we move out to the extremes of this flat region, the corresponding variance parameter increases because the Y-values imputed for the missing observations are more extreme than those observed. The likelihood falls sharply when these imputed values would give residuals that are incompatible with the distribution assumed. A more satisfactory approach, that could easily be incorporated into a Bayesian analysis, would be to seek prior information on the extreme percentiles of the marginal distribution of Y.

R. L. Chambers (University of Southampton): The message that I receive from this paper is that a statistician must not only take responsibility for modelling the population from which sample data are obtained but also model the sampling process itself, *and integrate these two models in inference*.

Many 'mainstream' statisticians concentrate on the first of the above models, assuming implicitly that sample data are obtained 'at random'. What is required is systematic development of statistical methods for 'non-random' or informative samples, and I see this paper as a valuable step in that direction.

There are two mathematically equivalent, but conceptually distinct, ways of modelling data obtained via informative sampling. These are best described by considering the two ways that we can factorize the joint distribution $f(\mathbf{Y}, \mathbf{I})$ of $\mathbf{Y}$, the population vector of the response variable, and $\mathbf{I}$, the 0–1 population vector characterizing sample inclusion:

(a) the mixture model,

$$f(\mathbf{Y}, \mathbf{I}) = f(\mathbf{Y}|\mathbf{I})\, f(\mathbf{I});$$

(b) the selection model,

$$f(\mathbf{Y}, \mathbf{I}) = f(\mathbf{I}|\mathbf{Y})\, f(\mathbf{Y}).$$

In (a), sample inclusion or exclusion is determined first, and then the values in $\mathbf{Y}$ are determined. Once $\mathbf{I}$ has been determined, there is no need for any link between the sample and non-sample Y-values. In (b), the values in $\mathbf{Y}$ are determined first, and, depending on these, the values in $\mathbf{I}$ are determined. Here there is a link between the sample and non-sample Y-values which can be exploited for inference.

The authors focus on a rather specialized 'threshold' version of (b). I prefer to think instead of a population unit having a probability π of being included in the sample, where this probability can depend on $\mathbf{Y}$ as well as on other variables. In sample surveys we are often fortunate in having access to these probabilities, and there is a growing literature on how this information should be combined with the sample Y-values for inference. For example, a simple estimator of the population mean of Y is the Hájèk estimator

$$\sum_{\text{sample}} Y_i \pi_i^{-1} \left(\sum_{\text{sample}} \pi_i^{-1} \right)^{-1}.$$

For observational studies or non-response problems, the π-values are unknown. However, we know $\mathbf{I}$, and we often have access to population covariates which can explain a large part of the variability in $\mathbf{I}$. This suggests that we empirically model the π-values and then use these estimated inclusion probabilities in inference, e.g. in a version of the Hájèk estimator above where the π-values are replaced by estimates. This estimator will not be efficient, but it may be more robust to misspecification of the sample selection mechanism than the corresponding Mills ratio-based estimator implied by this paper.

Roger A. Sugden (Goldsmiths College, London): More generally, if the complete data are a vector $Y = (Y_1, \ldots, Y_N)'$ of responses and a matrix of covariates $X = (X_1, \ldots, X_N)'$, but selection occurs according to a vector of indicators $I = (I_1, \ldots, I_N)'$, then the *full likelihood* is just the joint distribution $f(Y, X, I)$ integrated or summed over unobserved data and the *face value likelihood*

(Dawid and Dickey, 1977) ignores selection by keeping I fixed. Both likelihoods can be conditioned as appropriate.

Sugden and Smith (1984) looked at partial design information under the assumption of *non-informative sampling* that I is conditionally independent of Y given X, which makes selection ignorable if both I and X are completely observed, and defines known non-stochastic functions $p(I|X)$, the selection (sampling) mechanism, and $\pi = \pi(X)$ where $\pi_i = E[I_i|X]$, $i = 1, \ldots, N$, are the selection (inclusion) probabilities of the units.

If the assumption of non-informative sampling is not made, as in this paper, then the joint distribution can be written either as

$$p(I|Y, X)\, f(Y|X)\, f(X),$$

the first term being the *selection model*, or

$$f(Y|X, I)\, p(I|X)\, f(X),$$

the *pattern mixture model*, leading to the *sample-based likelihood* of Skinner (1994), which attempts to treat the data as a random sample from a selection-modified population.

My interest is when the probabilities π are available (perhaps only for sampled units) and possibly little other design information. Rather than model π as a non-stochastic function of X and Y, we can regard it as a random variable in its own right and consider the joint distribution $f(Y, X, I, \pi)$. If the two assumptions

$$I \perp Y|X, \pi \quad \text{and} \quad I \perp X|\pi$$

(equivalent to $I \perp (Y, X)|\pi$) are satisfied, and this will be the case if π are the 'true' selection probabilities which fully specify the probabilistic rule determining the selected units $s \subset \{1, \ldots, N\}$, then the joint distribution can be written

$$f(Y|X, \pi)\, f(X|\pi)\, f(\pi)\, p(I|\pi),$$

and the first term of this reduces to the target conditional distribution $f(Y|X)$ under the non-informativeness condition $\pi \perp Y|X$.

Under informative sampling, we can either use the mixture approach with π as a further covariate or the selection approach, modelling the decomposition

$$f(\pi|Y, X)\, f(Y|X)\, f(X)\, p(I|\pi).$$

In the independent and identically distributed case with independent sampling of units and data (i, y_i, x_i, π_i), $i \in s$, we obtain the likelihood (further conditioning on I which is now typically an ancillary statistic)

$$\prod_{i \in s} f(y_i|x_i, \pi_i)\, f(x_i|\pi_i)\, \frac{f(\pi_i)\pi_i}{E[\pi_i]}.$$

This is in the form of the face value likelihood times a factor corresponding to size-biased sampling for inference on the parameters of the marginal distribution of the π_i. A similar derivation in this case was given by Smith (1988).

Ben Armstrong (London School of Hygiene and Tropical Medicine): I would like to make a comment at a somewhat different level from that of the other contributors to the discussion, more on the context than the technical content of the paper. As an applied statistician working with epidemiologists, I was delighted to see from the title of this paper that mainstream statisticians were addressing what I see as one of my major practical concerns, and I was even more delighted, on reading the paper, that the approach suggested was one that could be implemented without extensive software development.

However, I was concerned that in their introduction and in the context in which they presented some examples the authors, perhaps inadvertently, may have given the impression that scientists concerned with observational studies were unaware that inference from non-random samples must go beyond application of statistical methods based on the assumption of random samples, or that these scientists do not have approaches, albeit less formal, to doing so. It is my experience that epidemiologists, for example, are well aware of this problem — sometimes to obsession — and they have quite a wide range of approaches to address it. Formal models such as that presented here should have a role in making inference from non-random samples, but it is not easy to see such models incorporating more than a small part of the information brought to bear in the less formal approaches, even in contexts of no more than average complexity. Thus I suspect that less formal discussion of the implication of non-random samples will continue to predominate in the observational sciences even when this and similar approaches are widely known — and probably rightly so.

Garrett Fitzmaurice (Nuffield College, Oxford): I have some concerns regarding how sensitive the authors' proposed methods are to misspecification errors, in particular to departures from

(a) the assumption of normal residuals and
(b) the assumed linearity of the conditional expectation of ϵ_1 given ϵ_2.

With the assumption of normal residuals in equation (4), information about the selection process comes from the skewness in the *observed* residuals. However, if the population distribution of the residuals is skewed, but due to selection the observed residuals are symmetric, the proposed methods will mistakenly infer that there is no selection bias (Rubin, 1978). I also question how robust the proposed methods are to violations of (b).

Another issue concerns how to determine a plausible range for η. The authors remark that this choice is often no more than an 'order of magnitude' guess. But surely determining a plausible range for η requires auxiliary information about the selection process. This leads to my final comment on the kinds of supplementary information that can alleviate problems of selection bias. Focus on the missing data case. When there are missing data, ideally we would like to have some external information about the selection process, e.g. through a follow-up survey of a sample of non-respondents. In practice, though, such information is rarely collected or available. However, information about the *marginal* distribution of Y is often readily available from secondary sources and can alleviate problems of selection bias. The following simple numerical example illustrates this idea. The data in Table 1 consist of 200 cases with complete data on both X and Y and 100 cases having data only on X. Let $p(y, x, r) = \mathrm{pr}(Y = y, X = x, R = r)$. The parameter of main interest concerns the association between Y and X, and it is natural to parameterize this association in terms of the log-odds ratio

$$\psi = \log\left\{\frac{p(1, 1, +)\,p(0, 0, +)}{p(1, 0, +)\,p(0, 1, +)}\right\},$$

where $p(1, 1, +) = p(1, 1, 0) + p(1, 1, 1)$, etc. Since ψ can be expressed as a non-linear function of $p(y, x, r)$, assessing its sensitivity to non-response can be regarded as a non-linear optimization problem, subject to the linear equality constraints imposed by the observed proportions in the contingency table (Fitzmaurice *et al.*, 1996). This yields a parameter window for ψ of $(-0.077, 3.245)$. Clearly, ψ is a parameter that is very sensitive to non-response. Because the range of non-identifiable values for ψ includes 0, the association between Y and X could be explained in terms of selection bias. However, suppose that additional information is available concerning $p(y, +, +)$, e.g. from a previous study it is known that $p(1, +, +) = p(0, +, +) = 0.5$. This imposes an additional linear equality constraint, i.e. $p(1, 0, 0) + p(1, 1, 0) + p(1, 0, 1) + p(1, 1, 1) = 0.5$, and the parameter window for ψ reduces to $(1.099, 2.398)$. With this supplementary information, selection bias can no longer explain the association between Y and X.

E. A. Molina (Universidad Simón Bolívar, Caracas): This is an important contribution to the study of the effect of treatment assignment and selection on statistical inference. The approach is model based. There are, however, many assumptions that should be explicit, since they limit the type of non-randomness that is permissible under the models to a very particular kind of *non-informativeness*. An example is the assumptions of independent and identical distribution behind expression (14), which imply other aspects of random allocation (or selection).

In general, observational studies encompass at least three activities: *selection* of the units, *allocation* of the treatments and *measurement* of the responses. The selection mechanism of sample units is absent

TABLE 1

2×2 contingency table with one partially classified margin†

		X	
		0	*1*
($R=1$) Y	0	100	30
	1	20	50
($R=0$) Missing		40	60

†Adapted from Little and Rubin (1987), chapter 11.

from the models. This leaves us with models whose relationship with the populations is neither explicit nor verifiable. The presentation in Section 1 may lead to the impression that models for treatment allocation (or selection) may be inspired by the observed data *solely*. Allocation (or selection) are activities that lead to the observed data and is that activity which the statistician should attempt to model if he wishes to extend the validity of its conclusions beyond the range of the observed data. It is at the heart of Fisher's argument that selection models should be probabilistic, which is granted under random selection.

It is impressive (and an important contribution from the authors) how, despite all the restrictions in the models, just a lack of non-informativeness leads to very misleading inferences. The paper also shows the need for a theory of selection, including the subject of the estimation of selection probabilities. This is particularly relevant to observational studies, where these probabilities are usually not considered, as opposed to the survey literature, where they are paramount.

Suppose that we have a finite population and let I_p and I_s respectively denote the vector of random indicators of selection and its observed value under a probabilistic mechanism of selection p. Let Λ_p and Λ_s denote the associated diagonal matrices and let Y denote the vector of responses under a model ξ. The observed data are $\Lambda_\mathrm{s} y$. If ξ is based on the observed data *only*, we have a model conditional on the observations, e.g. $E_\xi(\Lambda_\mathrm{p} Y|\Lambda_\mathrm{p} = \Lambda_\mathrm{s}) = \Lambda_\mathrm{s}\mu$. Then $E(\Lambda_\mathrm{p} Y) = E_\mathrm{p}\, E_\xi(\Lambda_\mathrm{p} Y|\Lambda_\mathrm{s}) = \Lambda_\pi \mu$, where Λ_π is the diagonal matrix of selection probabilities π_i. This says that the conclusions of the analysis should be weighted if we wish to extend them to the population.

A. J. Lawrance (University of Birmingham): This paper has presented an enlightening view of the often implicit assumption of randomness in some form in validating statistical analysis. The basic model is an elegantly simple perturbation of the ordinary regression model in an important respect. I enjoyed learning something new about the paired t-test! One way in which the basic model might be used is for a perturbation scheme in local influence diagnostics. In this respect we would be looking for cases of the data which seem to have responded in a way which has given them an unusual amount of influence on the estimates of the regression parameters. It is then for the investigator to deduce whether this is actually due to some selection or allocation effect. With the local influence approach to diagnostics (Cook (1986) initiated this), we have a different perturbation for each case, with a surface formed from them, and we work with the geometry of this surface. There is no estimation of perturbations *per se*. In the familiar case of perturbation to the constant variance assumption in the linear model, the local influence diagnostics point up the ordinary residuals. It would be nice to apply similar ideas to the selection perturbations and to find intuitively supported results, although this seems a tall order without doing some work.

Sander Greenland (University of California, Los Angeles): Copas and Li reach the conclusion that 'at least some kind of sensitivity analysis is essential when analysing observational data'. I applaud this — it is, after all, that reached by many others, including me (Greenland, 1990) and Rosenbaum (1995). None-the-less, I must admit discomfort with the approaches developed by Copas and Li and Rosenbaum. As their examples demonstrate, these approaches are capable of alerting us to sensitivity of inferences to randomization assumptions. But there is an aspect of the work that seems like an attempt to salvage a failing model — experimental statistics applied to observational data — by adding more unknowns to the model.

Both approaches begin by assuming that randomization has occurred conditionally on certain covariates, albeit by an unknown mechanism (the 'propensity score' in Rosenbaum's work; the 'selection equation' here) that is a very simple function of the covariates. Copas and Li assume that the covariates 'are always fixed and observed' (Section 2). In my field (epidemiology) there are usually many unobserved covariates that are associated with everything of interest; this renders ϵ_2 of model (5) correlated with x and with y given x, ϵ_1. Copas and Li (in Section 4) and Rosenbaum address this problem by allowing a 'hidden' covariate to affect allocation, but only through a very strict parametric model.

Both approaches produce a spectrum of p-values. I can only interpret these as representing the results of a spectrum of conditionally randomized experiments, one of which *might* have been performed. In much of epidemiology, however, I see no basis for assuming that *any* randomized experiment was performed. Even if I believed that some sort of randomized experiment was performed, I see no basis for assuming or testing that the experiment was in the spectra used by Copas and Li or Rosenbaum. These assumptions especially strain my credulity when (for example) I examine observational data concerning β-carotene and lung cancer, and I consider the enormous and unknown complexities that must characterize the relationships between health habits, diet, taste preferences, genetics, nutrient absorption, nutrient metabolism and cancer.

Copas and Li and Rosenbaum propose methods that are better than what is usually done (which, essentially, is to assume simple randomization conditional on observed covariates). I just fear that their methods will be misinterpreted as fully addressing the non-randomized nature of observational studies. Indeed, Greenland (1996) and Poole and Greenland (1997) argue that the presentations by Rosenbaum (1995) and Gastwirth *et al.* (1994) are misleading for precisely this reason. I commend Copas and Li for their more restrained interpretations, but how would they interpret a result like that of the medical example if the 'treatment' was (say) vegetarian diet and the outcome was mortality rate?

The following contributions were received in writing after the meeting.

Chris Chatfield (University of Bath): The theory of statistical inference generally assumes that data are random samples taken from some known population. In practice this is rarely the case, whether in sample surveys or in experiments. Thus I welcome this paper for starting to look at whether, and when, it is possible to make inferences from non-random samples. If we know from external contextual considerations that data are missing completely at random or are censored or, as in this paper, that it is reasonable to assume a linear selection equation as in the authors' equation (5), then it is possible to make some progress. However, equation (5) constitutes a fairly substantial assumption and it needs to be clearly understood that inference is not possible when the analyst has no idea how the probability of selection is related to the observed variables.

I have recently been brought firmly back down to earth by helping to organize a survey of households in my local area for our parish church. We achieved a 1 in 3 response rate which is very good for this type of survey. However, we will not try to pretend that we can say much, if anything, about the remaining two-thirds of households who did not respond. A descriptive, rather than inferential, analysis is indicated, and I suspect that this is more often the case in practice than the statistical literature implies.

At the other extreme to the above 'one-off' situation, scientific relationships are often formulated, not from a single sample, but by taking a series of samples under (slightly) different conditions to check for example that a result holds this month as well as last month, in the UK as well as in the USA, and so on. This may obviate the need to make potentially dubious assumptions about randomness in a single sample.

Peter J. Diggle (Lancaster University): The authors note a connection between their models and those of Diggle and Kenward (1994) for informative drop-out in longitudinal studies. A closer connection is with the models of Wu and Carroll (1988), who assumed that the response Y and the indicator variable $Z > 0$ are linked through an underlying random effects model. Specifically, the authors' equations (4) and (5) can be re-expressed as

$$Y = \beta^{\mathrm{T}} x + \sigma U + \nu \epsilon$$

and

$$P(Z > 0 | U) = \Phi(\gamma^{\mathrm{T}} x + \tau U)$$

where U and ϵ are mutually independent standard normal variates. Thus, the dependence of the indicator $Z > 0$ on Y is imparted indirectly, via the unobserved U, whereas the kind of model considered in Diggle and Kenward (1994) would correspond to assuming that

$$P(Z > 0 | Y) = \Phi(\gamma^{\mathrm{T}} x + \alpha Y).$$

At least in the longitudinal setting, it might be worthwhile to embed both of these kinds of model within a wider setting, say

$$P(Z > 0 | U, Y) = \Phi(\gamma^{\mathrm{T}} d + \tau U + \alpha Y), \tag{34}$$

and to note explicitly that different explanatory variables, x and d, might be used in the respective models for Y and Z conditional on Y.

I agree with the authors that, for models of this kind, the data typically give very little information about at least one parameter (the authors' θ); yet it can be grossly misleading to ignore the missing data mechanism (i.e. to assume that $\theta = 0$). See also Fitzmaurice *et al.* (1996). The authors' proposed sensitivity plots represent one very sensible way of dealing with this dilemma; another, admittedly not very original, is to let the practical context inform the choice of model. When dealing with drop-out in longitudinal studies, it might be reasonable to use contextual knowledge to decide whether drop-out of a subject should be modelled through explanatory variables, or as a direct consequence of their response history or as an indirect consequence of their latent characteristics, these being represented in equation (34) by the variables d, Y and U respectively. It would be rather optimistic to assume that the essence of the non-randomness in the underlying sampling can always be encapsulated in a single parameter.

A. S. C. Ehrenberg (South Bank University, London): As so often, I am totally baffled by what some of my statistical colleagues get up to, and why they do so.

Copas and Li start by noting that 'randomization . . . justifies the use of standard significance tests and other methods of normal inference'. But is it not the other way round — that a strict random sample *requires* inference to its population?

If the data are not a probabilistic sample, no inference to a population is required because there is no defined population. (The so-called 'sample' is simply a complete minipopulation.) Any attempt at *broader* inference (for instance to Fisher's undefined 'superpopulations') is usually dealt with quite differently, by checking out empirically whether the result generalizes to any *other*, different populations. There is no probabilistic short-cut to that, for arguing from one population to another, as far as I know.

The authors, however, suggest that we can allow statistically for any 'non-randomness' in our data — such as a 60% response rate in a survey — by estimating a 'non-randomness' parameter θ. But in their conclusion they warn that 'great caution is needed' in this. Less euphemistically, they are therefore saying that it does not work.

How could it? Surely one cannot make purely *probabilistic* inferences about the non-respondents in a survey. They might all be 'deceased', or 'have moved', or simply be 'not interested' or whatever.

When we empirically investigated 'non-co-operators' in on-going panel operations some years ago (Ehrenberg, 1960), they differed from the co-operators in virtually only one respect: they were not interested in co-operating. They did not differ in what they bought, or in what they said when they were probed in many other ways (including by psychologists). That does not mean that these non-co-operators were 'random' in any way, and hence that they either justified or required any kind of probabilistic inference. Instead, the data showed merely that they were not systematically biased.

D. A. Freedman (University of California, Berkeley): Selection bias is an endemic problem in social science research. Few samples are random, and data are often missing. One solution is to model the biases and to correct for them, as in Heckman (1979)—a paper that has been surprisingly influential. Heckman's model, however, does not flow from any real understanding of the mechanisms by which respondents are included in surveys, or excluded. The assumptions behind the model are therefore speculative, at least to some degree. If these assumptions remain in doubt, so do the merits of the corresponding procedures for estimating biases.

In their paper, Copas and Li make the following major points.

(a) Estimates based on models for selection bias are often remarkably sensitive to modelling assumptions.
(b) These assumptions can seldom be validated by examining the data.
(c) Even if the model is granted, conventional asymptotics take hold only with very large samples.

These points ring true, and the analysis is likely to help anyone wanting clarification of the selection bias literature.

In Section 1, Copas and Li allude to the somewhat problematic 'as if by randomization' assumptions underlying many statistical analyses. For a recent exchange on this topic, see Humphreys and Wojcicki (1995). There are, of course, other models for selection bias, non-compliance in clinical trials and similar problems, not covered in the present paper. Such models may be grist for a future essay by Copas and Li, and a general discussion postponed until then. Hierarchical logistic regression models for missing data, to mention one example, played some role in recent arguments over adjustments to the US census of 1990: see Breiman *et al.* (1994), pages 469, 489, 536. The US Supreme Court recently decided that the census would not be adjusted—without reaching the issue of imputation models.

Els Goetghebeur (Rijksuniversiteit Limburg, Maastricht) **and Krista Lapp** (University of Tartu): Our discussion draws on the related problem of estimating the effect of selective treatments when exposure to a treatment or placebo is randomized, but the level of exposure within the treatment group is generated from partial compliance which is possibly selective.

Using causal terminology, the effect of observed exposure on the gain from treatment is $E[Y_i^{\mathrm{T}} - Y_i^{\mathrm{P}} | C_i^{\mathrm{T}}]$, where Y_i^{P} and Y_i^{T} represent the possible response of subject i on the placebo or treatment arm respectively and C_i^{T} is the exposure resulting from partial compliance with the treatment.

A simple 'causal' model is

$$Y_i^{\mathrm{T}} = Y_i^{\mathrm{P}} + \beta C_i^{\mathrm{T}} + \epsilon_i^{\mathrm{T}} \qquad \text{with } E(\epsilon_i^{\mathrm{T}} | C_i^{\mathrm{T}}) = 0; \quad \mathrm{var}(\epsilon_i^{\mathrm{T}} | C_i^{\mathrm{T}}) = \sigma_{\mathrm{YT}}^2; \tag{35}$$

with selective C_i^{T}

$$Y_i^{\mathrm{P}} = \alpha + \gamma C_i^{\mathrm{T}} + \epsilon_i^{\mathrm{P}} \qquad \text{with } E(\epsilon_i^{\mathrm{P}} | C_i^{\mathrm{T}}) = 0; \quad \mathrm{var}(\epsilon_i^{\mathrm{P}} | C_i^{\mathrm{T}}) = \sigma_{\mathrm{YP}}^2. \tag{36}$$

$(C_i^{\mathrm{T}}, Y_i^{\mathrm{T}})$ and Y_i^{P} are observed in different arms only.

Exposure is random when $\gamma = 0$ and its effect β is then consistently estimated from a structural mean model (SMM) (Robins, 1994), from ordinary least squares (OLS) regression, or, given a parametric distribution for ϵ_i^{P}, ϵ_i^{T} and C_i^{T}, from maximum likelihood estimation (MLE). Under correct distributional assumptions, MLE is asymptotically the most efficient and SMM least.

When $\rho = \mathrm{corr}(C_i^{\mathrm{T}}, Y_i^{\mathrm{P}}) \neq 0$, SMM remains unbiased, but not OLS or MLE ignoring selection. Conditionally on C_i^{T}, we calculated exact mean-squared errors (MSEs) for the SMM and OLS approaches. For OLS, the bias depends linearly on ρ. With parameters mimicking a blood pressure reduction trial, the ratio of MSEs for SMM *versus* OLS is shown in Fig. 7.

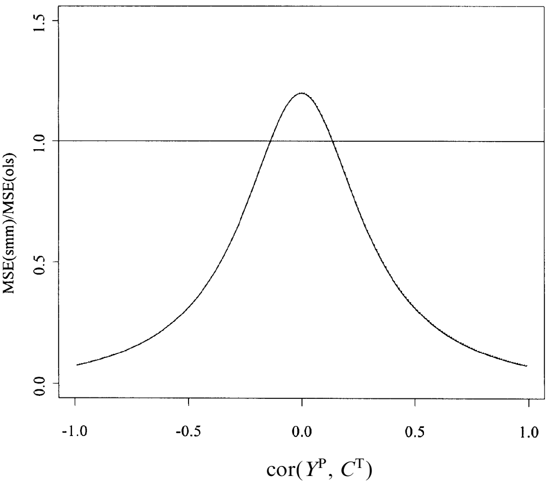

Fig. 7. Ratio of mean-squared errors: SMM/OLS

SMM wins when $|\rho|$ is large, but relies heavily on the linear form (35). A likelihood incorporating ρ (through $\gamma \neq 0$ in model (36)) is overparameterized when (Y_i^P, C_i^T, Y_i^T) are jointly modelled through normal (error) distributions and the structure described in models (35) and (36).

In practice, the analysis choice can combine prior beliefs in modelling assumptions (including (baseline) covariates) and the effect of their violation on results. Whereas Copas and Li studied violations of $\rho = 0$, the parametric model and normal errors were not tested. It would be worthwhile to investigate the relative importance of the three types of violation. Then, a weighted average of the different estimators could be chosen, which minimizes the MSE under prior beliefs. In practice, this involves much work. A prior investigation into extreme case scenarios can provide bounds on the parameters of interest (as in Balke and Pearl (1994)) and may indicate whether it is worth the trouble or whether we should rather seek external information on the selection mechanism.

Nicholas T. Longford (De Montfort University, Leicester): I subscribe to the view that the value of θ, which, for a two-groups comparison, characterizes the informativeness of the allocation to the groups, is fundamentally unknowable in the sense that the data (X, y) contain no information about θ. Certainly, the assumptions of the Heckman (or a similar) model are contrived and cannot be substantiated in the analysis of the Coventry skills audit or the national hearing survey (or in similar settings). I would be concerned more about global sensitivity, as emphasized by Rosenbaum (1995), than about local sensitivity. Intuition suggests that the addition of salient x-variables leads to a reduction of $|\theta|$. But when do we have *all* the xs?

Published analyses of observational studies deal with this issue by ubiquitous statements calling for 'caution in interpreting the results' (including Lockheed and Longford (1991)). The authors' approach is a much more satisfactory way of addressing this issue, not presenting a miracle solution, but formulating the issue in a way that promotes thinking about the uncertainty additional to that emanating from the experimental-design-based analysis.

Geert Molenberghs (Limburgs Universitair Centrum, Diepenbeek): The paper unifies the well-known attemps of econometricians to entertain selectivity bias and the related but relatively recent quest of statisticians for models that accommodate non-ignorable missingness. They rightly argue that in general inference is complicated by for instance very flat likelihoods and go on to conclude that a sensitivity analysis should be preferred over identifying a single model or carrying out a single hypothesis test. For this a sensitivity parameter is included in their model.

However, a major concern remains the sensitivity to *model specification*. This was noted by many discussants to Diggle and Kenward (1994). One can suspect that the sensitivity analysis might be driven by the particular choice of a selection model, whether it is a logistic regression as in Diggle and Kenward

(1994) or a joint normal model for y and z as proposed by the authors. This is the price to pay for apparent identifiability in selection models, as opposed to pattern mixture models (Glynn *et al.*, 1986; Little, 1994), where a sharp distinction exists between identifiable and unidentifiable parameters. Arguably, the choice of a correlation coefficient to describe the degree of informativeness might determine the conclusions (see equations (15) and (16)). Therefore, it is necessary to broaden the sensitivity analysis from a mere sensitivity parameter to a range of plausible parametric representations.

Clearly, goodness-of-fit tools cannot solve this problem since, as noted by Baker and Laird (1988), a selection model for incomplete data not only spends earned degrees of freedom but also 'predicts' data pertaining to missing data degrees of freedom. Whereas two models might agree completely on the former aspect, they can totally disagree on the latter, while this shows hardly or not at all in an assessment of model fit. The best that we can hope for with this fundamental untestability is that the inference is fairly stable for a range of plausible parametric approaches.

James M. Robins (Harvard School of Public Health, Boston): The authors' stimulating paper raised some provocative questions.

Question 1

If a model places no restriction on the law $f(Y|X)$ of Y given $X = (1, X'_{-1})'$, can we distinguish selection bias ($\theta \neq 0$) from its absence ($\theta = 0$) in the probit selection model $\Pr[R = 1|X, Y] = \Phi(\alpha X + \theta Y) \equiv \Phi(\alpha_1 + \alpha_{-1} X_{-1} + \theta Y)$ with $R \equiv I(Z > 0)$? When $\theta \neq 0$, it is necessary and sufficient for non-distinguishability of $\theta \neq 0$ from $\theta = 0$ based on data (X, RY, R) that the observed law $f[Y, R = 1|X] = f(Y|X)\, \Phi(\alpha X + \theta Y)$ can be written as $f^*(Y|X)\, \Phi(\alpha^* X)$ with $\int f^*(Y|X)\, \mathrm{d}Y = 1$, i.e.

$$\alpha^* X = \Phi^{-1}\left\{ \int f(Y|X)\, \Phi(\alpha X + \theta Y)\, \mathrm{d}Y \right\},$$

a sufficient (and essentially necessary) condition for which is that both $\alpha_{-1} = 0$ and Y and X are independent, giving $\alpha^*_{-1} = 0$ and $\alpha^*_1 = \Phi^{-1}\{E[\Phi(\alpha_1 + \theta Y)]\}$. This same condition is also sufficient for non-distinguishability when $\theta = 0$.

Question 2

In the setting of question 1, can θ be estimated at the usual $\surd n$-rate whenever $\theta = 0$ is distinguishable from $\theta \neq 0$? Rotnitzky and Robins (1996a) give necessary conditions for $\surd n$-estimation in a class of semiparametric non-ignorable missing data models. It follows from their propositions A1.6 and A2.6 that a necessary condition for $\surd n$-estimation of θ is that $h(X)$ be linearly independent of ($g(X)$, $X_{-1}\, g(X)$) where $h(X) = E[Y\, \lambda(\alpha X + \theta Y)|X]$, $g(X) = E[\lambda(\alpha X + \theta Y)|X]$ and $\lambda = \phi/\Phi$ is the Mills ratio. If $\theta = 0$ and $E(Y|X) = \eta_1 + \eta_{-1} X_{-1}$ is linear, then $h(X) = \{\eta_1 + \eta_{-1} X_{-1}\}\, g(X)$ and θ is not $\surd n$ estimable; the optimal achievable rate when $\theta = 0$ is identified (i.e. when Y and X are dependent and/or α_{-1} is non-zero) remains open.

Question 3

Are β and the selection law $\pi(Y, X) \equiv \Pr[R = 1|Y, X]$ identified if we are given that

(a) Y follows the model $Y = \beta X + \epsilon$ with $\epsilon \sim N(0, 1)$,
(b) $\pi(Y, X)$ is left totally unrestricted and
(c) $1 > c_1 > \pi(Y, X) > c_0 > 0$ with probability 1 for constants (c_0, c_1)?

In this model, β and $\pi(Y, X)$ are not identified if and only if, for some $\beta^* \neq \beta$, the observed law $f[Y, R = 1|X] = \phi(Y - \beta X)\, \pi(Y, X)$ equals $\phi(Y - \beta^* X)\, \pi^*(Y, X)$ with

$$0 \leqslant \pi^*(Y, X) = \pi(X, Y)\{\phi(Y - \beta X)/\phi(Y - \beta^* X)\} \leqslant 1 \qquad \text{for all } (Y, X).$$

Since this is never true, β and $\pi(Y, X)$ are identified. (However, by a similar argument, in the larger model with $\epsilon \sim N(0, \sigma^2)$ and σ^2 unknown, β and $\pi(Y, X)$ may not be identified.)

Question 4

In the setting of question 3, can β be estimated at a $\surd n$-rate? Rotnitzky and Robins (1996b) show that no parameter β of any model for $f(Y|X)$ can be estimated at a $\surd n$-rate when $\pi(Y, X)$ is left unrestricted and $\pi(Y, X) < c_1 < 1$ with probability 1. Specifically, they show that the Fisher information for β is 0 in the parametric submodel

$$\operatorname{logit}\{\pi(Y, X; \alpha)\} = \operatorname{logit}\{\pi(Y, X)\} - \alpha[S^{\mathrm{F}}_{\beta}(Y, X)/\{\pi(Y, X) - 1\}].$$

$S^{\mathrm{F}}_{\beta} \equiv S^{\mathrm{F}}_{\beta}(Y, X) = \partial\{\log f(Y|X; \beta)\}/\partial\beta$ and $S^{\mathrm{F}}_{\alpha} = -(R - \pi)S^{\mathrm{F}}_{\beta}/\{\pi(Y, X) - 1\}$ are the scores at the truth for β and α based on the full data (X, Y, R) and $S_{\beta} = E[S^{\mathrm{F}}_{\beta}|X, R, YR]$ and $S_{\alpha} = E[S^{\mathrm{F}}_{\alpha}|X, R, YR]$ are observed data scores. Since

$$S_{\beta} = S_{\alpha} = RS^{\mathrm{F}}_{\beta} - (1 - R)\, E[S^{\mathrm{F}}_{\beta}\pi(Y, X)|X]/\{1 - E[\pi(Y, X)|X]\},$$

the Fisher information for β is 0.

Paul Rosenbaum (University of Pennsylvania, Philadelphia): Copas and Li argue that econometric selection models are formally identified, but barely so (Tukey, 1986), and are not robust to changes in the model (Little, 1985), concluding that, if the models are used at all, they should be the basis for sensitivity analysis rather than joint estimation of parameters. These findings are, I believe, correct and important, and they support the concern that excessive hopes for selection models may diminish efforts to collect data of high quality (Silber and Rosenbaum, 1996).

This comment restates the point abstractly. In Fig. 8, the curved segment S is a ridge of constant maximized likelihood and the closed curve C is a set estimate of the parameter $\boldsymbol{\theta} = (\theta_1, \theta_2)$. As the sample size increases, S barely changes but C shrinks to wrap S closely.

It is often said that inferences require identified models (e.g. Basu (1983)). Though not identified, yielding no consistent estimate of $\boldsymbol{\theta}$, Fig. 8 contains information about $\boldsymbol{\theta}$. It is clear that $\theta_2 > 1$ and unlikely that $\theta_2 > 2$ if $\theta_1 > 2$. A sensitivity analysis extracts this information and may or may not settle practical questions. For instance, Cornfield *et al.* (1959) found that, if smoking caused no increase in lung cancer ($\theta_2 = 1$), the association being due to an unobserved binary attribute, then the attribute must predict lung cancer almost perfectly and be $\theta_2 = 9$ times more prevalent among smokers.

Nonparametric sensitivity analyses are an alternative approach that do not require assumptions about the shapes of unobserved distributions (Rosenbaum, 1995).

Andrea Rotnitzky (Harvard School of Public Health, Boston): I wish to thank the authors for an insightful paper. They reference Little (1985) and Lee and Chesher (1986) regarding the poor asymptotic behaviour of tests of $\rho = 0$ when $\gamma_{-1} = 0$. In fact, when $\gamma_{-1} = 0$ the information matrix is singular and the standard $\sqrt{n}$ asymptotic theory for maximum likelihood estimation does not hold. The asymptotic distribution of the maximum likelihood estimate (MLE) of ρ (and of corrected Wald tests) is a corollary to the following theorem proved in Rotnitzky and Robins (1996c).

Theorem. Suppose that a density $f(Y; \boldsymbol{\psi})$ indexed by a $p \times 1$ parameter $\boldsymbol{\psi}$ with true value $\boldsymbol{\psi}^*$ has information matrix of rank $p - 1$ at $\boldsymbol{\psi}^*$ so there exists a $(p - 1) \times 1$ constant vector K_1 satisfying for a component of $\boldsymbol{\psi}$, say ψ_1, $\{\partial\{\log f(Y; \boldsymbol{\psi})\}/\partial\psi_1\}|_{\boldsymbol{\psi}^*} = K_1^{\mathrm{T}}\Gamma$ where Γ is the score vector for $\psi_2, \psi_3, \ldots, \psi_p$ at $\boldsymbol{\psi}^*$ and $E(\Gamma\Gamma^{\mathrm{T}})$ is non-singular. Suppose that there exist an integer s and $(p - 1) \times 1$ (possibly null) vectors $K_2, \ldots, K_{s-1}$ defined iteratively by

$$\left.\frac{\partial^r\left[\log f\left\{Y; \boldsymbol{\psi} - \sum_{j=0}^{r-1} A_j(\psi_1 - \psi_1^*)^j\right\}\right]}{\partial\psi_1^r}\right|_{\boldsymbol{\psi}^*} = K_r^{\mathrm{T}}\Gamma$$

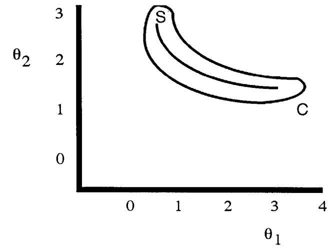

Fig. 8. Likelihood ridge and set estimate

where $1 \leqslant r \leqslant s-1$, $K_0 = \mathbf{0}$, $A_j = (\mathbf{0}, K_j^{\mathrm{T}})^{\mathrm{T}}$ for $0 \leqslant j \leqslant s-1$, such that

$$\Omega \equiv \left. \frac{\partial^s \left[\log f \left\{ Y; \boldsymbol{\psi} - \sum_{j=1}^{s-1} A_j (\psi_1 - \psi_1^*)^j \right\} \right]}{\partial \psi_1^s} \right|_{\boldsymbol{\psi}^*}$$

is neither 0 nor a linear combination of the elements of Γ. Then, under the smoothness conditions of Rotnitzky and Robins (1996c),

$$n^{1/2} \begin{pmatrix} (\hat{\psi}_1 - \psi_1^*)^s \\ \hat{\boldsymbol{\psi}}_{-1} - \boldsymbol{\psi}^*_{-1} + \sum_{j=0}^{s-1} K_j (\hat{\psi}_1 - \psi_1^*)^j \end{pmatrix} \xrightarrow{\mathcal{L}} \begin{cases} \mathbf{Z}\, I(Z_1 > 0) + \mathbf{U}\, I(Z_1 < 0) & \text{if } s \text{ is even}, \\ \mathbf{Z} & \text{if } s \text{ is odd}, \end{cases}$$

where $\boldsymbol{\psi} = (\psi_1, \boldsymbol{\psi}_{-1}^{\mathrm{T}})^{\mathrm{T}}$, $\mathbf{Z} = (Z_1, \ldots, Z_p)^{\mathrm{T}} \sim N(0, C^{-1})$, $\mathbf{U} = (0, \mathbf{W}^{\mathrm{T}})^{\mathrm{T}}$, $\mathbf{W} = (W_2, \ldots, W_p)^{\mathrm{T}} \sim N(0, \Sigma^{-1})$ with $C = E\{(\Omega, \Gamma)\, (\Omega, \Gamma)^{\mathrm{T}}\}/(s!)^2$ and $\Sigma = E(\Gamma\Gamma^{\mathrm{T}})/(s!)^2$.

To illustrate how these results can be applied to test for selectivity bias, consider, for simplicity, estimating $(\beta, \sigma, \alpha_1, \alpha_2)$ in the model $Y = \beta + \epsilon$, $\epsilon \sim N(0, \sigma^2)$, when Y is observed only if an always observed binary indicator R is equal to 1 and $P(R = 1|Y) = \pi\{\alpha_1 + \alpha_2(Y - \beta)\}$. The case $\pi(u) = \Phi(u)$ corresponds to the model of the authors' equations (4) and (7) with $x \equiv x_1 \equiv 1$, $\epsilon \equiv \sigma\epsilon_1$, $\alpha_1 = \gamma_1/(1-\rho^2)^{1/2}$ and $\alpha_2 \equiv \theta\sigma^{-1}$. To simplify the calculations, consider instead a logistic selection model $\pi(u) = \mathrm{e}^u/(1 + \mathrm{e}^u)$. Similar results hold for a probit selection model. Using the theorem, one can verify that when $\alpha_2 = 0$ and $\sigma = 1$

$$n^{1/2} \begin{pmatrix} (\hat{\beta} - \beta)^3 \\ \hat{\sigma} - 1 - \pi(\alpha_1)\{1 - \pi(\alpha_1)\}^{-1}(\hat{\beta} - \beta)^2 \\ \hat{\alpha}_1 - \alpha_1 + \{1 - 2\pi(\alpha_1)\}\{1 - \pi(\alpha_1)\}^{-2}(\hat{\beta} - \beta)^2 \\ \hat{\alpha}_2 + \{1 - \pi(\alpha_1)\}(\hat{\beta} - \beta) \end{pmatrix} \xrightarrow{\mathcal{L}} \mathbf{N}(0, C^{-1})$$

for a non-singular matrix C that I do not present here for lack of space. Thus, both $\hat{\beta}$ and $\hat{\alpha}_2$ converge at rate $O_p(n^{-1/6})$ and both $\hat{\sigma}$ and $\hat{\alpha}_1$ converge at rate $O_p(n^{-1/3})$. D. R. Cox (personal communication) earlier derived, from first principles, the asymptotic distribution of the MLE of $(\beta, \alpha_1, \alpha_2)$ when σ^2 is known. A corrected Wald test for the null hypothesis $\alpha_2 = 0$ of no selection bias can be derived from the asymptotic distribution of $\hat{\alpha}_2$. However, this test will be sensitive to local (Pitman) departures from the null hypothesis of order $O(n^{-1/6})$ and not of order $O(n^{-1/2})$. Thus extremely large samples will be required for the detection of small departures from the null hypothesis.

T. M. F. Smith (University of Southampton): I would like to congratulate the authors on a very interesting paper. Of particular importance are the results on the sensitivity of standard procedures, such as the paired *t*-test, to quite small selection effects, even when the underlying distributions are normal. In non-normal cases the combined effects of model misspecification and selection could be even more dramatic.

My main criticism of the paper is the title. I had expected to read a paper in which the process of selecting units would feature and the samples of units would be selected by using non-random sampling schemes. Sampling, in the sample survey sense, does not feature at all. The analysis is entirely model based and the sampling scheme implicit in likelihoods such as equation (14) is simple random sampling with replacement (independent and identically distributed). The paper tackles item non-response, which is part of the measurement process, not the sample selection process, and the non-random allocation of treatments to units in an observational study, not the selection of the units.

The paper also seems to imply that random sampling (random selection of units) and random treatment allocation are identical in some sense. This is not true. In random sampling there is a well-defined real finite population, samples are selected according to a specified sampling rule and inferences can be made about finite population parameters by using the randomization distribution. In randomization inference for experiments the set of experimental units comprises the population. There is no selection of units from a larger finite population. The randomization distribution refers to the hypothetical distribution corresponding to the allocation of different treatments to the same fixed unit; it is not a real population. As demonstrated by Rubin (1978) the validity of a causal inference about treatment effects depends not only on an ignorable allocation scheme such as randomization but also on an untestable assumption about unit treatment additivity or the stable unit treatment value assumption. Random sampling inference does not require this additional assumption.

Philip Young (University of York): I would like to comment on the dependence of the method proposed by the authors on the assumed model. In epidemiology a common technique is to match subjects with certain similar characteristics. However, this is clearly of limited value, unless we have a vast sample of data, since with only a few factors matched on we shall soon have trouble finding suitable matches. Therefore, as the authors have done, we are forced to fit a model and to use this, in effect, to 'match' subjects. In essence this then becomes a problem of calibration. Of course, even if we detect no treatment allocation bias, using the covariates that we have, we cannot know that there are no differences for some other factors that we have not measured.

Instead of the situation considered by the authors, consider the following discriminant problem, where subjects belong to one of two populations. One aspect that has received much attention in the calibration literature is the way in which we consider the errors within any model. In this particular situation we need to consider whether the errors lie in our final allocation to a population or in the covariates of interest. For example, if the populations are surviving and dead patients after a course of treatment then we would sensibly attribute the errors to the covariates, and we might assume that subjects in the two populations are distributed normally with identical covariance structures but different mean vectors; hence we use linear discriminant analysis. Yet if we were testing the efficiency of a new method of diagnosis we might assume that the errors lie in the technique and not the covariates, and one way in which to model the data would be to use logistic discriminant analysis. I suspect that the linear discriminant model is more appropriate since the very idea of a design matrix seems inappropriate for many of the situations that the authors discuss.

Now, assume that we have two treatments A and B. This can then be incorporated into the linear discriminant analysis by using the so-called location model. Some preliminary simulations that I have done suggest that if we assume the location model, when the logistic model is correct, then we are much more likely to conclude non-randomness in the treatment allocation when there is none. This must be a feature of the misspecification of the errors, because the way in which the covariates are handled is identical.

The **authors** replied later, in writing, as follows.

Firstly we thank the discussants for their perceptive and constructive comments about our paper.

Several discussants note the slightly ironical nature of the title. If by 'random' we mean the outcome of some stochastic mechanism, then of course the whole paper is a contradiction — our methods amount to analysing 'non-random' data by assuming that they *are* generated by a (albeit more complicated) stochastic model. If by random we refer to selection from a population then, as Professor Smith points out, the later parts of the paper are all about allocation and not selection. We hope that, after reading the paper, our rather more superficial use of 'random' is sufficiently clear.

A major theme echoed by many discussants is the nature of the model in Section 2, the technical basis of the whole paper. Professor Greenland writes '... no basis for assuming that *any* randomized experiment was performed ...'. Surely one could make the same comment about any non-experimental application of statistical inference — a model is a model and not necessarily a statement of reality. We are sorry if Professor Ehrenberg is totally baffled by our paper; I think that we are equally baffled by his opening remarks. We have *not* said that one can always allow statistically for *any* non-randomness — we have attempted throughout to be very constrained in our interpretations and have only advocated a local approach (only small deviations from the standard randomness assumption). Regarding his second paragraph, we are just quoting Fisher (1966), p. 21. Our main point is, if you are tempted to use a conventional model (tacitly assuming $\theta = 0$), then pause and ask what happens to your conclusion if $\theta \neq 0$. This can warn us if inference rests on essentially untestable assumptions. The paper stands or falls, not on whether the model in Section 2 is realistic, but on whether it is a simple and useful way of embedding the usual model to do sensitivity analysis.

Of course many other models are possible, as pointed out. Linking z to y via an independent unobserved random U, as suggested by Professor Diggle (compare the Rosenbaum approach), is a useful way of thinking about our model, and, as Diggle points out, can be generalized in various ways. This will be useful if we do have a richer data source such as in a longitudinal study where we have extraneous information on reasons for drop-out. In particular, the suggested generalization allows the covariates to be modelled in more detail at the different model stages — there is much discussion of such issues in the econometrics literature (separate roles of x in the two Heckman stages). We agree with Professor Raab that a general form of heteroscedasticity could be modelled without conceptual difficulty. We surmise, however, that if a small non-zero θ in our model makes a big change to inference then it will also tend to do so in other families in which we embed the standard model.

In emphasizing the local approach we concur with Dr Chatfield when he says that if the conventional model is completely implausible then there is unlikely to be any useful stochastic model for the data. So, as he says, we can only attempt a descriptive analysis. But then any inference on a population, or statement of causality, must rest on other considerations and not just on the data. It is difficult to hold the line on this — whether we like it or not the media will always report something like 'Dr Blogg's results show that . . .' rather than 'In Dr Blogg's data, . . ., but this cannot be generalized outside the particular cases he used'. (In our presentation of the paper we quoted some recent press headlines to emphasize this point.)

Some discussants comment on the population, or lack of it. Although it is stated at the start of Section 3 that the (complete) cases arise from simple random sampling, this is not really assumed in the development — we are estimating $\beta^{\mathrm{T}}\bar{x}$ and not $\beta^{\mathrm{T}} E(x)$ where the expectation here is over the population of possible xs. Following Professor Smith's comments, identity of random sampling and treatment allocation is established in a technical sense within our model. Essentially, if we assume unit–treatment additivity, then the treatment comparison is like two missing data problems in one — data on A are like the observed cases, and data about the responses the other units would give if given A are like the unobserved cases, and vice versa for the cases on B. The mathematics of the selection models in Sections 4 and 5, based on additivity, turn out to be exactly the same as that for unit non-response in Section 3. Again our argument is very local — a minimal approach to trying to understand the extra uncertainty caused by the (hopefully small number of) missing cases, or by the possibility that A and B cannot *both* be given to our particular set of experimental units. Strictly, generalizing to a wider population is beyond the scope of the paper.

Several discussants reinforce, and much more elegantly than we have done, the suggestion in the paper that inference about θ is difficult. The results mentioned by Professor Chesher, and extended in the theoretical results of Dr Robins and Dr Rotnitzky, show that without special assumptions the model is unidentifiable. Equivalently, inference about θ depends sensitively on such assumptions, as we have seen in equations (15) and (16) which show that the shape of the likelihood function depends on the shape of the sample, and hence on the assumption of normality. This also explains Professor Skinner's 'paradox', that weakening the assumption of $\theta = 0$ increases sensitivity, such as in Fig. 3. When $\theta = 0$ we have the simple sufficient statistics for normality, but when $\theta \neq 0$ these statistics are no longer sufficient. We thank Professor Lawrance for his suggestion — if we are trying to estimate θ then model assumptions are crucially important, and we need good diagnostics; for example the shape of $L^*(\theta)$ given in equations (4) and (5) is strongly affected by sample outliers. The idea of using local influence diagnostics rather than the Q–Q-plots mentioned in Section 6 is likely to be a good one.

In reacting to many of the comments we repeat that in our examples only a local sensitivity approach is used, for which it seems that these assumptions, such as assumption (a) in Dr Fitzmaurice's comment, are much less critical (compare Figs 3 and 4). The sensitivity approach (the multiplier A in Section 6) *does*, however, depend on the model connecting z with x, or equivalently on Dr Fitzmaurice's assumption (b). We mentioned in Section 6 the assumed monotonicity of the dependence of $P(z > 0|y, x)$ on y. If this was U shaped, people with unusually high or low incomes both being more likely to refuse to respond in the survey, then the sensitivity calculation would be quite different.

We agree with Professor Molenberghs that an advantage of the pattern mixture approach is that, by having separate models for $f(y|z > 0)$ and $f(y|z < 0)$ rather than deriving both of them from a single model as we have done, it is made explicit that the selection mechanism is unidentifiable. However, we find the selection modelling approach more natural for the kinds of example that we are considering — one starts with the income and then determines whether or not to respond, rather than the other way round.

Several discussants refer to the importance of outside information on selectivity, and we agree. Dr Armstrong makes a very fair comment, but we only advance our approach as *one* way of looking at local sensitivity. Of course there are other less formal considerations — any information on the reasons for non-response or the factors behind self-selection will be more important than any formal analysis ignoring such information. Professor Raab mentions a formal Bayes approach, but we comment in Section 6 on the problem of sensitivity to prior assumptions—you always obtain an answer even if the data tell you nothing at all about the parameter of interest! How to use data on survey recalls to provide information about θ is an interesting problem for research. The example of follow-up information discussed by Professor Ehrenberg is interesting—suggesting in this case that θ is not too large.

Some of the examples used in the discussion are about discrete data (Professor Rosenbaum and Dr Fitzmaurice) and so are quite different from the models in the paper. Suppose that y is known always to lie in (0, 1), and that in a sample of 100 cases we have 50 missing observations. Suppose that the 50 observed ys add to 25. Then we know for certain that the sum of all 100 ys cannot be less than 25 or more than 75. With normally distributed data no such bounds are possible, and ranges of uncertainty have to depend on the model rather than on the geometry of the relevant spaces. This is related to Professor Rosenbaum's approach with the bounds on the hidden covariate u, but, as he has pointed out in a separate communication, this requirement is relaxed somewhat in Section 4 of Rosenbaum (1987) by putting a bound on the number of cases with values of u outside a finite interval.

Dr Sugden and Professor Chambers discuss non-response in the more usual survey sampling framework. Surely the main point in Dr Sugden's notation is that the Is are *not* usually ancillary, and that is the problem. In Professor Chambers's notation, stage I of Heckman's method is essentially modelling the πs. If D is small Heckman's stage II fails, warning us that the model is almost unidentified. But, if there is little evidence of variation in the πs (as appears to be the case in Section 3.3), then the estimated πs will be almost constant and so the Hájèk estimate will be close to the sample mean, i.e. assumes that θ is close to 0. Surely this is not sensible. We would prefer a model relating the πs to the other variables including y, and seeing how sensitive inference is to varying the coefficients in this model. This is essentially what we have done, but within a completely different framework.

The role of the covariates in selection is emphasized by Professor Hand and Dr Longford. In Professor Hand's application (consumer credit) the screening is done by a deterministic z—essentially this is the case of large γ (so that ϵ_2 is relatively unimportant). Surely these other variables used in making the first screening *would* be known, so ignorability is not an issue in our technical sense. But, more interestingly, the problem of ignorability will come in at the next stage, in that applicants may be offered a loan but not actually take it up. On his question of model misspecification, small θ asymptotics seem robust to normality but depend on linearity of the structure linking z to y, as we have already remarked. Identification of θ is non-robust, but we are not trying to do that in the examples in the paper. In response to Dr Longford's comment, adding more xs does not necessarily 'reduce $|\theta|$', but increasing D (the size of γ) does reduce the dependence of equation (7) on y. We say little about global sensitivity, but Fig. 4, for example, suggests that the linear local approximation is useful for quite large ρ.

We ended our presentation at the meeting by remarking that we make no great technical claims for the paper but hope that we have contributed to awareness of the practical importance of these selection issues. The discussion has shown the diversity of models and interpretations that are possible in this area, but surely all are agreed on one thing—that this *is* a very important practical problem for statisticians.

REFERENCES IN THE DISCUSSION

Baker, S. G. and Laird, N. M. (1988) Regression analysis for categorical survey variables with outcome subject to nonignorable nonresponse. *J. Am. Statist. Ass.*, **83**, 62–69.

Balke, A. and Pearl, J. (1994) Non-parametric bounds on causal effects from partial compliance data. *Technical Report R-199-J.* Computer Science Department, University of California, Los Angeles.

Basu, A. P. (1983) Identifiability. In *Encyclopedia of Statistical Sciences* (eds S. Kotz and N. Johnson), vol. 4, pp. 2–6. New York: Wiley.

Breiman, L., Freedman, D., Wachter, K., Belin, T. R. and Rolph, J. E. (1994) Three papers on the census adjustment. *Statist. Sci.*, **9**, 458–537.

Chamberlain, G. (1986) Asymptotic efficiency in semi-parametric models with censoring. *J. Econometr.*, **32**, 189–218.

Chesher, A. D. and Spady, R. H. (1991) Asymptotic expansions of the information matrix test statistic. *Econometrica*, **59**, 787–816.

Cook, R. D. (1986) Assessment of local influence (with discussion). *J. R. Statist. Soc.* B, **48**, 133–160.

Cornfield, J., Haenszel, W., Hammond, E., Lilienfeld, A., Shimkin, M. and Wynder, E. (1959) Smoking and lung cancer: recent evidence and a discussion of some questions. *J. Natn. Cancer Inst.*, **22**, 173–203.

Dawid, A. P. and Dickey, J. M. (1977) Likelihood and Bayesian inference from selectively reported data. *J. Am. Statist. Ass.*, **72**, 845–850.

Diggle, P. and Kenward, M. G. (1994) Informative drop-out in longitudinal data analysis (with discussion). *Appl. Statist.*, **43**, 49–93.

Efron, B. and Tibshirani, R. J. (1993) *An Introduction to the Bootstrap*. London: Chapman and Hall.

Ehrenberg, A. S. C. (1960) A study of some potential biases in the operation of a consumer panel. *Appl. Statist.*, **9**, 20–27.

Fisher, R. A. (1966) *Design of Experiments*, 8th edn. Edinburgh: Oliver and Boyd.

Fitzmaurice, G. M., Heath, A. F. and Clifford, P. (1996) Logistic regression models for binary panel data with attrition. *J. R. Statist. Soc.* A, **159**, 249–263.

Gastwirth, J., Krieger, A. and Rosenbaum, P. (1994) How a court accepted an impossible explanation. *Am. Statistn*, **48**, 313–315.

Glynn, R. J., Laird, N. M. and Rubin, D. B. (1986) Selection modeling versus mixture modeling with non-ignorable nonresponse. In *Drawing Inferences from Self-selected Samples* (ed. H. Wainer), pp. 115–142. New York: Springer.

Greenland, S. (1990) Randomization, statistics, and causal inference. *Epidemiology*, **1**, 421–429.

——— (1996) Review of "Observational Studies" by P. Rosenbaum. *Statist. Med.*, **15**, in the press.

Hand, D. J. and Henley, W. E. (1993) Can reject inference ever work? *IMA J. Math. Appl. Bus. Indstry*, **5**, 45–55.

——— (1994) Inference about rejected cases in discriminant analysis. In *New Approaches in Classification and Data Analysis* (eds E. Diday, Y. Lechevallier, M. Schader, P. Bertrand and B. Burtschy), pp. 292–299. Berlin: Springer.

Heckman, J. (1979) Sample selection bias as a specification error. *Econometrica*, **47**, 153–161.

Humphreys, P. and Wojcicki, R. (1995) *Found. Sci.*, **1**, 19–83.

Lee, L.-F. and Chesher, A. (1986) Specification testing when score test statistics are identically zero. *J. Econometr.*, **31**, 121–149.

Little, R. J. A. (1985) A note about models for selectivity bias. *Econometrica*, **53**, 1469–1474.

——— (1994) A class of pattern-mixture models for normal incomplete data. *Biometrika*, **81**, 471–484.

Lockheed, M. E. and Longford, N. T. (1991) School effects on mathematics achievement gain in Thailand. In *Schools, Classrooms and Pupils: International Studies of Schooling from a Multilevel Perspective* (eds S. W. Raudenbush and J. D. Willms). London: Academic Press.

Poole, C. and Greenland, S. (1997) How a court accepted a possible explanation. *Am. Statistn*, **50**, in the press.

Powell, J. L. (1994) Estimation of semiparametric models. In *Handbook of Econometrics* (eds R. F. Engle and D. L. McFadden), vol. IV, ch. 41. Amsterdam: Elsevier Science.

Robins, J. M. (1994) Correcting for noncompliance in randomized trials using structural nested mean models. *Communs Statist. Theory Meth.*, **23**, 2379–2412.

Rosenbaum, P. R. (1987) Sensitivity analysis for certain permutation inferences in matched observational studies. *Biometrika*, **74**, 13–26.

——— (1995) Quantiles in nonrandom samples and observational studies. *J. Am. Statist. Ass.*, **90**, 1424–1431.

Rotnitzky, A. and Robins, J. M. (1996a) Analysis of semiparametric regression models with non-ignorable non-response. *Statist. Med.*, to be published.

——— (1996b) Semiparametric regression for repeated outcomes with non-ignorable non-response. To be published.

——— (1996c) The asymptotic distribution of the maximum likelihood estimator when the information matrix is singular. *Technical Report*. Department of Biostatistics, Harvard School of Public Health, Boston.

Rubin, D. B. (1978) Multiple imputations in sample surveys: a phenomenological Bayesian approach to nonresponse. *Proc. Surv. Res. Meth. Sect. Am. Statist. Ass.*, 20–34.

Rubin, D. B., Stern, H. S. and Vehovar, V. (1995) Handling 'don't know' survey responses: the case of the Slovenian plebiscite. *J. Am. Statist. Ass.*, **90**, 822–828.

Silber, J. and Rosenbaum, P. R. (1995) Letter: Measuring the quality of hospital care. *J. Am. Med. Ass.*, **273**, 21.

Skinner, C. J. (1994) Sample models and weights. *Proc. Surv. Res. Meth. Sect. Am. Statist. Ass.*

Smith, T. M. F. (1988) To weight or not to weight that is the question. In *Bayesian Statistics 3* (eds J. M. Bernardo, M. DeGroot, D. V. Lindley and A. F. M. Smith), pp. 437–451. Oxford: Oxford University Press.

Stefanski, L. A. (1985) The effects of measurement error on parameter estimation. *Biometrika*, **72**, 583–592.

Sugden, R. A. and Smith, T. M. F. (1984) Ignorable and informative designs in survey sampling inference. *Biometrika*, **71**, 495–506.

Tukey, J. (1986) Comments. In *Drawing Inferences from Self-selected Samples* (ed. H. Wainer), pp. 58–62, 108–110. New York: Springer.

Wadsworth, J., Johnson, A. M., Wellings, K. and Field, J. (1996) What's in a mean?—an examination of the inconsistency between men and women in reporting sexual partnerships. *J. R. Statist. Soc.* A, **159**, 111–123.

Wu, M. C. and Carroll, R. J. (1988) Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process. *Biometrics*, **44**, 175–188.