MISSING OBSERVATIONS IN REGRESSION: A CONDITIONAL APPROACH

H. S. BATTEY AND D. R. COX†

SUMMARY. This note presents an alternative to multiple imputation and other approaches to regression analysis in the presence of missing covariate data. Our recommendation, based on factorial and fractional factorial arrangements, is more faithful to ancillarity considerations of regression analysis and involves assessing the sensitivity of inference on each regression parameter to missingness in each of the explanatory variables. The ideas are illustrated on a medical example concerned with success of hematopoietic stem cell transplantation in children, and on a sociological example concerned with socio-economic inequalities in educational attainment.

Key words: ancillarity, EM algorithm, fractional factorial, Hadamard matrix, missing data, regression

1. INTRODUCTION

Analysis of observational data is frequently hindered by observations on some explanatory variables being missing. There is a rich literature on this, stemming indirectly from the self-consistency property of maximum likelihood estimation (Fisher, 1925), which is perhaps best approached via Efron (1977; 1982, p. 351). This in particular forms the basis for the EM algorithm (Sundberg, 1974; Dempster *et al.*, 1977) for maximum likelihood estimation from incomplete data. In a regression context when the missing observations are on covariates, application of these ideas necessitates specification of a joint probability model for the vector of responses Y and the matrix of covariates X, which leads essentially to a probability model for the missing observations. Efficient algorithms are in fact rather similar to imputation methods, based on an assumption that observations are missing at random, not to be confused with missing completely at random, an even stronger assumption.

Modelling aspects of the distribution of X, or indeed the joint distribution of X and Y, is in contradiction with the practice of conditioning on X in a regression setting. At least when data are fully observed, the latter is appropriate based on ancillarity considerations, since the regression function is a property of the conditional distribution of Y given X, so that the distribution of X is irrelevant. In fact, when the d-dimensional vector of regression coefficients is a canonical parameter of an exponential family, the appropriate conditional calculation for the parameter of interest, the dth coefficient say, conditions on the realized values of $(x_{\cdot 1}^{\mathrm{T}}Y, \ldots, x_{\cdot (d-1)}^{\mathrm{T}}Y)$, where $x_{\cdot j}$ denotes the jth column of X and $x_{\cdot j}^{\mathrm{T}}$ is its transpose. This choice both eliminates the nuisance parameters and ensures that the precision attached to the conclusions is that actually achieved and not an average over hypothetical, so called recognizably distinct, situations that have not in fact occurred. A

Date: January 10, 2023.

† Note on the posthumous version: a partial draft of the work was completed in September 2021 and set aside while attending to other projects. D. R. Cox died on January 18, 2022. This version includes revisions and additions to that draft.

detailed discussion of these issues is not necessary for present purposes. See Fisher (1956, §IV.4) or Cox (1958a, 1958b, 1970, §4.2).

This short paper illustrates what we consider to be a more appropriate approach to presenting conclusions when covariate data are missing. This entails replacing unobserved entries by rather extreme assignments in such a way that the effect of missingness of each covariate can be assessed. One of several conclusions emerges: that the missing observations do not materially affect inference on the aspects of interest; that the missingness in some variables affects inference on some but not all of the regression coefficients; or that the missingness is so severe that reliable conclusions are not possible without strong uncheckable assumptions on the process through which the missingness arises. The focus is primarily on the main effects of missingness, estimated from a simple fractional factorial design for the missing entries of X, although in principle interactive effects can also be assessed, provided that the number of columns of X with missing entries is not impractically large. The idea appears partially, in less explicit form, in the example application of Battey, Cox and Jackson (2019).

2. Full and fractional factorial assessments of missingness

Let x_i for $i = 1, \ldots, n$ be d-dimensional vectors, the transposed rows of an $n \times d$ covariate matrix X and let y_i be the associated outcome variables. For a subset $\mathring{\mathcal{I}} \subset \{1, \ldots, n\}$ of observations, some entries of x_i are missing for reasons that are unknown. The set of observed vectors are denoted by $\mathring{x}_i$, for $i = 1, \ldots, n$. Note that $\mathring{x}_i = x_i$ for $i \notin \mathring{\mathcal{I}}$ and otherwise $\mathring{x}_i$ has at least one entry missing.

Suppose that the missing entries arise in $m \leq d$ of the columns of $\mathring{X}$, where $\mathring{X}$ is the $n \times d$ matrix with $\mathring{x}_i^{\mathrm{T}}$ as its rows. Consider initially replacing all missing entries from a single column of $\mathring{X}$ by the maximum or minimum of the observed entries in that column. This can in principle be done for all 2^m combinations of high and low values, leading to 2^m matrices, $\mathring{X}^{(c)}$ say, whose transposed rows are denoted by $\mathring{x}_i^{(c)}$ where $c = 1, \ldots, 2^m$ indexes a particular column-wise combination of substitutions for the missing entries of $\mathring{X}$. Associated with these are 2^m vectors of regression coefficient estimates, written $\hat{\beta}^{(1)}, \ldots, \hat{\beta}^{(2^m)}$, each in $\mathbb{R}^d$, and their estimated standard errors. For each entry of the unknown regression coefficient vector, the main effect of missingness of each variable and any low-order interactive effects can be estimated by the appropriate factorial contrast.

It is not critical that the maximum and minimum value of each observed column be used in the combinatorial replacements, only that relatively extreme assignments be made in order to assess the sensitivity, where "extreme" is calibrated against the data that have been observed. In Battey, Cox and Jackson (2019) the upper and lower quartiles of any continuous variables were used, giving a less conservative assessment. A caveat concerns the situation in which extremity is the reason for the missingness. For instance, if a measuring device is unable to record observations above or below a certain threshold. In that case, the analysis would not faithfully reflect the range of conclusions that could have been reached had the data been available in their entirety.

As elucidated in Appendix A for linear least squares, it cannot be guaranteed in general that the inaccessible estimate $\hat{\beta}$, based on the unobservable matrix X, is contained in the convex hull of $\hat{\beta}^{(1)}, \ldots, \hat{\beta}^{(2^m)}$. The sensitivity analysis to be outlined is a way of presenting the evidence in incisive form, while being transparent about the limitations of the data. By contrast, imputation procedures seek to present inferential statements as if the missing data had been observed. McCullagh (2023, Chapter 14) convincingly

articulates the matter and concludes that the practice can be rather misleading, a view that we share. In the present context it is also, at least from one point of view, a violation of conditionality as noted above.

2.1. **Illustration for two columns with missing entries.** Suppose for illustrative simplicity that there are two columns with missing entries among those of the $n \times d$ matrix X. It is helpful to introduce in this illustration a more explicit notation coinciding with the usual one for factorial designs. Thus, associate with the incomplete columns of X the factors A, B and a general treatment combination $a^j b^k$ where j and k take values zero and one when the corresponding factor is at its high and low level respectively. Thus, when $j = k = 1$ the missing entries in both columns are replaced by the maximum of the column entries that have been observed. The treatment combination of all factors at their low level is written (1). The four treatment combinations are $X^{(1)}, X^{(a)}, X^{(b)}, X^{(ab)}$ and what would be called outcome variables in the usual experimental design terminology are the d-dimensional vectors $\hat\beta^{(1)}, \hat\beta^{(a)}, \hat\beta^{(b)}, \hat\beta^{(ab)}$ and their estimated standard errors. The effect of missingness in the column associated with A is then summarized by the usual factorial contrast for the main effect of A, namely $2\hat\tau^A$, where

$$\hat\tau^A = (\hat\beta^{(ab)} + \hat\beta^{(a)} - \hat\beta^{(b)} - \hat\beta^{(1)})/4 \tag{2.1}$$

and similarly

$$\begin{aligned} \hat\tau^B &= (\hat\beta^{(ab)} + \hat\beta^{(b)} - \hat\beta^{(a)} - \hat\beta^{(1)})/4 \\ \hat\tau^{AB} &= (\hat\beta^{(ab)} - \hat\beta^{(b)} - (\hat\beta^{(a)} - \hat\beta^{(1)}))/4. \end{aligned}$$

These are not to be interpreted as estimators of any population-level quantities as they are considered conditionally on the data. The factorial contrast $2\hat\tau^A$ is the effect of changing A from its low to its high level, averaged over the levels of B, and $2\hat\tau^{AB}$ is the difference between the effects of changing A when B is at its high level and when it is at its low level. Effectively, the contrast for variable j is a discrete approximation to the partial derivative of $\hat\beta$ with respect to the aspects of $x_{\cdot j}$ that have not been observed, holding all other such aspects fixed.

2.2. **Generalisation to higher-dimensional missingness.** If m is small, it is reasonable to report the $(m + \binom{m}{2}) \times d$ dimensional matrix of main effects of missingness and pairwise interactions. When the number m of covariates subject to missing observations exceeds 4 or 5, the 2^m sets of coefficients from the full factorial is unreasonably large for presentation, and wasteful for estimating the m main effects of missingness. These may instead be estimated from a $2^{m-\ell}$ fractional factorial arrangement with ℓ such that $2^{m-\ell} \geq m + 1$. For a comprehensive discussion of fractional factorial arrangements, conveniently studied using prime power commutative groups of order 2^m, see Cox and Reid (2000, §5.5.2) or Bailey (2008, §13). When interest lies in main effects only, construction is greatly simplified by using Hadamard matrices, with a Hadamard matrix of dimension 2^m being constructed from one of half the size as

$$H_{2^m} = \begin{pmatrix} H_{2^{m-1}} & H_{2^{m-1}} \\ H_{2^{m-1}} & -H_{2^{m-1}} \end{pmatrix}. \tag{2.2}$$

The 2^m rows of the sub-matrix

$$\begin{pmatrix} H_{2^{m-1}} \\ -H_{2^{m-1}} \end{pmatrix},$$

after discarding the first column, define the treatment combinations associated with the full factorial, and the 2^{m-1} rows of the sub-matrices $H_{2^{(m-1)}}$ and $-H_{2^{(m-1)}}$, after discarding the first column of each, define the treatment combinations associated with the two distinct half-replicates of the full factorial system. For instance, with $m = 3$,

$$H_{2^m} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 \\ 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 \\ 1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 \\ 1 & -1 & -1 & 1 & -1 & 1 & 1 & -1 \end{pmatrix}$$

and the assignments associated with the two half-replicates are $\{X^{(abc)}, X^{(b)}, X^{(a)}, X^{(c)}\}$ and $\{X^{(1)}, X^{(ac)}, X^{(bc)}, X^{(ab)}\}$. Orthogonality of the Hadamard sub-matrices ensures that each factor is present twice at each level with a different combination of levels for the other factors.

More generally, for ℓ such that $2^{m-\ell} \geq m+1$, there are ℓ fractions of size $2^{-\ell}$ of the full 2^m factorial design, and two distinct Hadamard matrices of dimension $2^{m-\ell}$ embedded within a given one of dimension 2^m. The last m columns specify $(2^{-\ell})$-replicates of the full factorial experiment. For example, with $m = 5$, a full factorial entails 32 treatments. The highest available degree of fractionation gives two quarter-replicates specified by the eight rows and last five columns of the two distinct Hadamard sub-matrices of H_{16} (or $-H_{16}$). In view of this, when m is large, the main effects of missingness can be constructed analogously to equation (2.1) by assigning high and low entries as determined by the last m columns of a Hadamard matrix of dimension M, say, where $M \geq m + 1$.

Let $\mathfrak{F} \subset \{1, \ldots, 2^m\}$ denote a treatment combination associated with, say, a half replicate of the full factorial, as detailed above. The main effects of missingness could be constructed either from $\mathfrak{F}$ or its complement $\overline{\mathfrak{F}} = \{1, \ldots, 2^m\} \backslash \mathfrak{F}$. The answers will be similar provided that appreciable interactive effects are not present. Since appreciable discrepancies can be checked for, fractional replication provides an internal mechanism for validating or refuting the reasonableness of the simplified analysis.

Two further comments are helpful. If the Hadamard matrix is not in the standard form (2.2), the column consisting only of 1 or -1 should be discarded. If $M > m + 1$, any of the redundant columns can also be used to estimate main effects, and the conclusions will be consistent, as these specify the same treatment combinations in different orders.

There results an $m \times d$ matrix of main effects of missingness, where the (j, k)th entry is the effect of missing observations on variable j on the kth estimated regression coefficient. Whether a particular main effect of missingness is deemed large should be calibrated against the estimates themselves so as to make the assessment dimension free. This can be done for both the coefficient estimates and the estimates of standard errors. The ideas are best illustrated with an example; see §3.

3. Two examples

3.1. **A medical example.** The ideas are illustrated using publicly available data from Kalwak *et al.* (2010) on the success of hematopoietic stem cell transplatation in children. The measured covariates in this study included treatment modifications and numerous forms of mismatch between donors and recipients. This was a wide-ranging investigation

and we focus on one aspect of it. In particular, the binary outcome, coded as $\{0, 1\} = \{\text{unsuccessful, successful}\}$ classifies the treatment as unsuccessful if the patient died or relapsed within the follow-up period. The shortest such period for a surviving patient was 433 days. The covariate data are summarized in Table 1, where $x_{\cdot j}$ represents the jth column of the covariate matrix X. Key treatment variables are $x_{\cdot 1}$ to $x_{\cdot 4}$, which measure whether the stem cells were sourced from bone marrow or peripheral blood and the relative and absolute concentrations of $CD3^+$ and $CD34^+$ cell doses after infusion. Variables $x_{\cdot 5}$ to $x_{\cdot 8}$ measure the degree of compatibility between donor and recipient, with $x_{\cdot 5}$ and $x_{\cdot 6}$ specifying the number of antigens and alleles in which donor and recipient differ, $x_{\cdot 7}$ indicating whether their blood groups match, and $x_{\cdot 5}$ a measure of serological compatibility according to cytomegalovirus infection prior to transplantation (the higher the value, the lower the compatibility). Variables $x_{\cdot 9}$ to $x_{\cdot 18}$ are intrinsic features of the donor and recipient whose interpretations are mostly clear from Table 1. Exceptions are $x_{\cdot 9}$ and $x_{\cdot 10}$, indicating presence or absence of cytomegalovirus infection in, respectively, the recipient and donor of hematopoietic stem cells prior to transplantation, and $x_{\cdot 11}$, indicating presence of the Rh factor on the recipient's red blood cells.

Although these data appear to have been carefully collected, a small number of entries for 10 of the 18 variables are missing for reasons that are unknown. To quantify the effects of missingness on the estimates of the logistic regression coefficients, we constructed a 16-dimensional Hadamard matrix and discarded the first 6 columns. Note that 2^4 is the smallest power of 2 that is larger than m, and this specifies a 2^{-6}-replicate of the full factorial, the latter consisting of all 2^{10} treatments. Let H denote the resulting 16×10 matrix. Each row of H specifies a combination of values to be assigned to the missing entries of variables $x_{\cdot 2}$ and $x_{\cdot 3}$, and variables $x_{\cdot 5}$ to $x_{\cdot 12}$. For instance, since the top row of H consists only of ones, missing entries of $x_{\cdot j}$ are replaced by $\max\{x_{ij} : i \in \mathring{I}\}$. The second row of H is an alternating sequence of -1 and 1, specifying that the missing entries of $x_{\cdot 2}$ be replaced by $\min\{x_{i2} : i \in \mathring{I}\}$, those of $x_{\cdot 3}$ be replaced by $\max\{x_{i3} : i \in \mathring{I}\}$, and so on.

For each of the 16 combinations of missingness, estimates of the logistic constant parameter and 18 regression coefficients were stored, alongside their estimated standard errors. Let B and S denote the resulting 16×19 matrices. In direct analogy with (2.1), the main effect of missingness of the kth partially observed variable on coefficient ℓ is constructed orthogonally to the other effects as $(h_k^{\mathrm{T}} b_\ell)/8$, where h_k is the kth column of H and b_ℓ is the ℓth column of B. Let E_B denote the 10×19 matrix of such effects. The same analysis using S in place of B produces a matrix E_S of main effects of missingness on the standard error estimates. The entries of E_B and E_S are best calibrated against the entries of B and S respectively. For instance, dividing the absolute entries in the ℓth column of E_B by $\max_c |\hat{\beta}_\ell^{(c)}|$ produces a dimension-free number, comparable across rows and columns and between tables for coefficients and standard errors. These are presented in Tables 2 and 3. We do not put forward any definitive thresholds. As in other contexts, the appropriate exposition presents the evidence in incisive form, avoiding binary decisions to the extent feasible.

Missingness in variables $x_{\cdot 3}$ and $x_{\cdot 8}$ has an appreciable effect on several of the coefficient estimates. Some of the standard errors are also affected, particularly those on $\hat{\beta}_j$, $j \in \{8, 9, 10\}$. Other standard errors are relatively unaffected by the extreme reassignments of missing entries, typically varying by less than 5% and in many cases less than 1%. For the more stable coefficient estimates, namely $\hat{\beta}_0$, $\hat{\beta}_7$, $\hat{\beta}_{13}$, $\hat{\beta}_{14}$, $\hat{\beta}_{16}$, and their standard errors,

covariate	description	sample range	% missing
$x_{\cdot 1}$	stem cell source	$\{0 = \text{marrow}, 1 = \text{peripheral blood}\}$	0
$x_{\cdot 2}$	$CD3^+/CD34^+$	$[0.20 - 99.56]$	2.67
$x_{\cdot 3}$	$CD3^+$ per kg	$[0.040 - 20.02]$	2.67
$x_{\cdot 4}$	$CD34^+$ per kg	$[0.79 - 57.78]$	0
$x_{\cdot 5}$	antigen discrepancies	$\{0, 1, 2, 3\}$	0.53
$x_{\cdot 6}$	allele discrepancies	$\{0, 1, 2, 3, 4\}$	0.53
$x_{\cdot 7}$	blood group match	$\{0 = \text{mismatch}, 1 = \text{match}\}$	0.53
$x_{\cdot 8}$	CMV incompatibility score	$\{0, 1, 2, 3\}$	8.56
$x_{\cdot 9}$	recipient CMV	$\{0 = \text{absent}, 1 = \text{present}\}$	7.49
$x_{\cdot 10}$	donor CMV	$\{0 = \text{absent}, 1 = \text{present}\}$	1.07
$x_{\cdot 11}$	recipient Rh factor	$\{0 = \text{absent}, 1 = \text{present}\}$	1.07
$x_{\cdot 12}$	recipient body mass	$[6, 103]$	1.07
$x_{\cdot 13}$	previous relapse	$\{0 = \text{no}, 1 = \text{yes}\}$	0
$x_{\cdot 14}$	risk group	$\{0 = \text{standard}, 1 = \text{high}\}$	0
$x_{\cdot 15}$	disease type	$\{0 = \text{nonmalignant}, 1 = \text{malignant}\}$	0
$x_{\cdot 16}$	recipient sex	$\{0 = \text{female}, 1 = \text{male}\}$	0
$x_{\cdot 17}$	recipient age (years)	$[0.60, 20.2]$	0
$x_{\cdot 18}$	donor age (years)	$[18.65, 55.55]$	0

TABLE 1. summary of data.

it is reasonable to interpret the output from an arbitrary treatment assignment from the factorial combination. Such values are reported in Table 4.

None of the variables studied is statistically significant at typical thresholds, the strongest suggestion coming from $x_{\cdot 7}$ and $x_{\cdot 14}$. It is also worth noting that none of the omitted rows from Table 4 would be deemed statistically significant according to their notional p-values, which are all in excess of 0.39.

The analysis was repeated using a different 2^{-6} fraction obtained by multiplying the Hadamard matrix, or equivalently the reduced form H, by minus one. The most notable differences were on the relative effect of missingness of variable $x_{\cdot 3}$ on $\hat{\beta}_2$ and $\hat{\beta}_7$, which were 0.374 and 0.301 respectively in the second replicate, compared with 0.298 and 0.119 in the first replicate.

	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	$\hat{\beta}_6$	$\hat{\beta}_7$	$\hat{\beta}_8$	$\hat{\beta}_9$	$\hat{\beta}_{10}$	$\hat{\beta}_{11}$	$\hat{\beta}_{12}$	$\hat{\beta}_{13}$	$\hat{\beta}_{14}$	$\hat{\beta}_{15}$	$\hat{\beta}_{16}$	$\hat{\beta}_{17}$	$\hat{\beta}_{18}$
$x_{\cdot 2}$	0.1089	0.5169	0.3948	0.1455	0.0593	0.0502	0.1086	0.0285	0.0693	0.0473	0.0524	0.0697	0.0771	0.1277	0.0936	0.0845	0.0757	0.2560	0.0498
$x_{\cdot 3}$	0.0677	0.5609	0.2976	0.5676	0.4488	0.1012	0.0002	0.1194	0.4706	0.5574	0.3038	0.2773	0.0729	0.0269	0.0155	0.1321	0.2632	0.1075	0.4811
$x_{\cdot 5}$	0.0487	0.0367	0.0599	0.0949	0.1875	0.3416	0.2281	0.0995	0.0517	0.0491	0.0382	0.0628	0.0294	0.0752	0.0289	0.3551	0.0157	0.0817	0.0361
$x_{\cdot 6}$	0.0353	0.0514	0.1006	0.0034	0.1829	0.3288	0.5565	0.0394	0.0031	0.0069	0.0136	0.1156	0.1167	0.0875	0.0090	0.0059	0.0227	0.2064	0.0615
$x_{\cdot 7}$	0.0134	0.0143	0.0087	0.0004	0.0367	0.0127	0.0023	0.0779	0.0148	0.0195	0.0326	0.0174	0.0019	0.0435	0.0135	0.0767	0.0121	0.0015	0.0167
$x_{\cdot 8}$	0.0307	0.1179	0.0088	0.0254	0.0436	0.0022	0.0178	0.0389	0.9496	0.8474	0.5734	0.1171	0.0694	0.1603	0.0693	0.3356	0.1270	0.1300	0.0339
$x_{\cdot 9}$	0.1765	0.0987	0.0129	0.0220	0.1336	0.0440	0.0801	0.1110	0.5834	0.7935	0.3536	0.0493	0.0227	0.0906	0.0915	0.4983	0.0806	0.0382	0.1506
$x_{\cdot 10}$	0.0196	0.0342	0.0134	0.0010	0.1635	0.0130	0.0277	0.0157	0.0651	0.0519	0.0590	0.0230	0.0240	0.1004	0.0089	0.1985	0.0477	0.0441	0.0056
$x_{\cdot 11}$	0.1275	0.0249	0.0207	0.0585	0.0097	0.0574	0.0402	0.0150	0.0041	0.0003	0.0092	0.5274	0.0390	0.0014	0.0317	0.2230	0.0860	0.0904	0.0121
$x_{\cdot 12}$	0.0513	0.1529	0.0970	0.0314	0.1482	0.1870	0.1053	0.2996	0.0754	0.0588	0.0335	0.2392	0.3229	0.0589	0.0252	0.1013	0.1039	0.5435	0.0227

TABLE 2. Absolute main effects of missingness in variable $x_{\cdot j}$ on the estimated logistic regression coefficients $\hat{\beta}_k$ relative to the maximum absolute coefficient estimate for each column.

	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	$\hat{\beta}_6$	$\hat{\beta}_7$	$\hat{\beta}_8$	$\hat{\beta}_9$	$\hat{\beta}_{10}$	$\hat{\beta}_{11}$	$\hat{\beta}_{12}$	$\hat{\beta}_{13}$	$\hat{\beta}_{14}$	$\hat{\beta}_{15}$	$\hat{\beta}_{16}$	$\hat{\beta}_{17}$	$\hat{\beta}_{18}$
$x_{\cdot 2}$	0.0086	0.0249	0.3772	0.0701	0.0029	0.0001	0.0092	0.0007	0.0668	0.0570	0.0708	0.0067	0.0687	0.0015	0.0082	0.0060	0.0034	0.0595	0.0099
$x_{\cdot 3}$	0.0003	0.0129	0.0515	0.2387	0.0191	0.0083	0.0115	0.0177	0.4602	0.4557	0.2739	0.0094	0.0121	0.0114	0.0085	0.0075	0.0124	0.0169	0.0141
$x_{\cdot 5}$	0.0006	0.0069	0.0013	0.0126	0.0007	0.0043	0.0123	0.0040	0.0077	0.0076	0.0043	0.0014	0.0294	0.0062	0.0019	0.0048	0.0016	0.0119	0.0111
$x_{\cdot 6}$	0.0019	0.0066	0.0112	0.0114	0.0071	0.0022	0.0335	0.0008	0.0003	0.0005	0.0020	0.0001	0.0022	0.0035	0.0005	0.0011	0.0004	0.0086	0.0076
$x_{\cdot 7}$	0.0003	0.0021	0.0003	0.0016	0.0023	0.0016	0.0008	0.0006	0.0222	0.0195	0.0218	0.0003	0.0022	$< 10^{-4}$	0.0002	0.0012	$< 10^{-4}$	0.0022	0.0016
$x_{\cdot 8}$	0.0034	0.0044	0.0008	0.0066	0.0111	0.0012	0.0039	0.0057	0.0329	0.0419	0.0120	0.0049	0.0051	0.0098	0.0139	0.0024	0.0054	0.0048	0.0010
$x_{\cdot 9}$	0.0101	0.0012	0.0023	0.0027	0.0122	0.0019	0.0006	0.0024	0.0493	0.0725	0.0333	0.0000	0.0051	0.0100	0.0019	0.0002	0.0008	0.0045	0.0055
$x_{\cdot 10}$	0.0043	0.0002	0.0019	0.0032	0.0013	0.0019	0.0024	0.0032	0.0225	0.0218	0.0275	0.0009	0.0060	0.0016	0.0002	0.0039	0.0006	0.0054	0.0005
$x_{\cdot 11}$	0.0092	0.0008	0.0074	0.0025	0.0017	0.0008	0.0003	0.0032	0.0006	0.0007	0.0019	0.0218	0.0129	0.0010	0.0023	0.0025	0.0033	0.0006	0.0021
$x_{\cdot 12}$	0.0057	0.0092	0.0120	0.0031	0.0150	0.0123	0.0032	0.0196	0.0063	0.0063	0.0042	0.0049	0.0037	0.0020	0.0067	0.0030	0.0017	0.0169	0.0032

TABLE 3. Absolute main effects of missingness in variable $x_{\cdot j}$ on the standard errors of $\hat{\beta}_k$ relative to the maximum standard error for each column.

j	$\hat{\beta}_j$	standard error	p-value
0	-1.45	1.07	0.176
7	0.384	0.36	0.293
13	0.296	0.57	0.602
14	0.425	0.40	0.293
16	0.154	0.34	0.646

TABLE 4. estimates and estimated standard errors of logistic regression coefficients (additive on the logit scale).

3.2. **A sociological example.** The data to be analysed, from the US National Longitudinal Study of Youth (1979), were used by Battey, Cox and Jackson (2019) to illustrate different statistical issues to those in the present paper. The binary outcome, coded as $\{0, 1\}$, is enrolment on a four-year-degree-granting institution for at least one year. Five explanatory variables are: ability, measured as the respondent's score on the Armed Forces Qualifying Test, administered to all respondents in the 1981 wave of the survey; family income in childhood, measured as the log of total net family income in 1979; sex, as indicated by the respondent; race, recorded by interviewer observation; and whether respondents were living with at least one parent at the time of the first survey. The sample was restricted to those respondents who were classified as black or non-black and non-Hispanic. These data are summarized in Table 5.

covariate	description	sample range	% missing
$x_{\cdot 1}$	sex	$\{1 = \text{male}, 0 = \text{female}\}$	0
$x_{\cdot 2}$	AFQT score	percentage $(0 - 100)$	4.3
$x_{\cdot 3}$	log income	continuous $(3.00 - 11.23)$	51.2
$x_{\cdot 4}$	race	$\{1 = \text{black}, 0 = \text{non-black/non-Hispanic}\}$	0
$x_{\cdot 5}$	lives with parent	$\{1 = \text{yes}, 0 = \text{no}\}$	5.1

TABLE 5. summary of data.

With only three variables having missing observations, it is feasible to report the output from all 2^3 factorial combinations. However, to illustrate the previous ideas we calculate the main effects of missingness from the last 3 columns of the 8-dimensional standard Hadamard matrix. Information analogous to that in Tables 2 and 3 is given in Tables 6 and 7. The coefficient estimate $\hat{\beta}_3$ is highly unstable and to a lesser extent so is $\hat{\beta}_5$. Other coefficients are more secure and, for these, we report the output from an arbitrary treatment combination in Table 8.

	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$
$x_{\cdot 2}$	0.042	0.053	0.154	0.048	0.150	0.322
$x_{\cdot 3}$	0.293	0.012	0.0060	1.008	0.021	0.051
$x_{\cdot 5}$	0.024	0.018	0.0011	0.031	0.0074	0.468

TABLE 6. Main effects of missingness in variable $x_{\cdot j}$ on the estimated logistic regression coefficients $\hat{\beta}_k$ relative to the maximum absolute coefficient estimate for each column.

	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$
$x_{\cdot 2}$	0.021	0.032	0.097	0.019	0.031	0.033
$x_{\cdot 3}$	0.625	0.0015	0.0014	0.625	0.0060	0.017
$x_{\cdot 5}$	0.0042	0.0007	0.0007	0.0073	0.0024	0.064

TABLE 7. Main effects of missingness in variable $x_{\cdot j}$ on the estimated standard error of $\hat{\beta}_k$ relative to the maximum standard error for each column.

j	$\hat{\beta}_j$	standard error	p-value
0	-3.39	0.250	$< 10^{-9}$
1	-0.366	0.049	$< 10^{-9}$
2	0.0422	0.0010	$< 10^{-9}$
4	1.051	0.064	$< 10^{-9}$

TABLE 8. estimates and estimated standard errors of logistic regression coefficients (additive on the logit scale).

4. Discussion

When observations on explanatory variables are missing, a reasonable approach for general use is to assess the sensitivity of the inference to this missingness. Fragility would point to limitations of statistical inference on the data at hand. This is in contrast with the widely deployed strategy of multiple imputation (Rubin, 1987) in which missing entries are replaced by values drawn multiple times from a distribution, and the estimates and standard errors averaged. This produces a single answer without warning, regardless of how sensitive the conclusions may be to aspects that were not observed.

The approach here is semi-descriptive: an acknowledgement that statistics can only take us so far when the quality of the data is low. Any more formal statistical guarantees would entail strong assumptions on the process by which the missingness arises. While conceptually very different, it is possible that aspects of the proposal are operationally related to tests of the missing completely at random assumption (MCAR).

Appendix A. Some geometric properties of least squares in the full factorial system

The recommended sensitivity analysis is conditional on the data, and it is not in keeping with the paper to introduce any mechanism through which the factorial contrasts could be treated as random. It is, however, of some interest to know how the set of coefficients arising from the factorial or fractional factorial treatment plan relate to those that would have been obtained had the data been available in their entirety.

When the number of covariates subject to missing observations is relatively small (less than four, say), it is feasible and reasonable to report the output from a full factorial arrangement. The present section provides a geometric characterization of the relationship between the notional regression coefficient estimate, $\hat{\beta}$ say, and those observable ones determined by substitution at each factorial combination of missingness. The discussion is restricted to linear least squares estimates so that $\hat{\beta} = (\sum x_i x_i^{\mathrm{T}})^{-1} \sum x_i y_i$ and $\hat{\beta}^{(1)}, \ldots, \hat{\beta}^{(2^m)}$ are analogously defined.

While one might expect the notional coefficient estimate $\hat{\beta}$ to reside in the convex hull of $\hat{\beta}^{(1)}, \ldots, \hat{\beta}^{(2^m)}$, this strong property has been neither proved nor refuted in the current work (see the discussion below Proposition A.1). Instead it has been demonstrated that $\hat{\beta}$ is in the convex hull of a closely related set which cannot be explicitly calculated from the observed data.

While worth reporting, the connection to §2 is slight, because the geometric result is much stronger than is required to justify the sensitivity analysis.

The proof of Proposition A.1 is in Appendix B.

Proposition A.1. *Provided that $w = \sum_i x_i x_i^{\mathrm{T}}$ is invertible and the missing entries in each column of $\mathring{X}$ are not one of the two extreme points from the corresponding column of X, the notional least squares estimate satisfies*

$$\hat{\beta} \in \mathrm{conv}\{w^{-1} w^{(c)} \hat{\beta}^{(c)} : c = 1, \ldots, 2^m\} \subset \mathbb{R}^d, \tag{A.1}$$

where $w^{(c)} = \sum_i \mathring{x}_i^{(c)} \mathring{x}_i^{(c)\mathrm{T}}$ and, for a finite set $\mathcal{A}$, $\mathrm{conv}(\mathcal{A})$ is the convex hull of $\mathcal{A}$.

Note that $\mathring{X}^{(c)} = X + P^{(c)}$, say, where $P^{(c)}$ is a matrix consisting primarily of zeros except in the positions corresponding to missing entries of X, where they take values $p_{ij}^{(c)} = \max\{x_{kj} : k \in \mathring{I}\} - x_{ij}$ or $p_{ij}^{(c)} = \min\{x_{kj} : k \in \mathring{I}\} - x_{ij}$ depending on the assignment prescribed by c. Thus,

$$w^{-1} w^{(c)} = I_n + (X^{\mathrm{T}} X)^{-1} \big[P^{(c)\mathrm{T}} X + X^{\mathrm{T}} P^{(c)} + P^{(c)\mathrm{T}} P^{(c)} \big] = I_n + A,$$

say, where I_n is the n-dimensional identity matrix. It follows that $w^{-1} w^{(c)} - I_n$ is positive semi-definite if A is. Whether this is satisfied is not immediately clear, as some configurations produce a constituent matrix $P^{(c)\mathrm{T}} X + X^{\mathrm{T}} P^{(c)}$ that is negative definite.

Since §2 suggests fractional factorial assessments of missingness, it is of some interest to consider whether a version of (A.1) holds with c varying over a restricted set. Again, the semi-descriptive assurances of §2 do not hinge on geometric results of this nature, which are much stronger.

Inspection of the proof of Proposition A.1 reveals that equation (A.1) holds for any subset $\mathfrak{K} \subset \{1, \ldots, 2^m\}$ of the full factorial set such that $x_i \in \mathrm{conv}\{\mathring{x}_i^{(c)} : c \in \mathfrak{K}\}$ for all i. Thus, if one considers $\mathfrak{K} = \mathfrak{F}$ for some fractional factorial combination, the above is in effect a restriction on the values of the missing entries of each x_i, since $\mathrm{conv}\{\mathring{x}_i^{(c)} : c \in \mathfrak{F}\} \subset \mathrm{conv}\{\mathring{x}_i^{(c)} : c = 1, \ldots, 2^m\}$. By Caratheodory's theorem (Lemma B.1),

$$\mathrm{conv}\{\mathring{x}_i^{(c)} : c = 1, \ldots, 2^m\} = \mathrm{conv}\{s_{i1}, \ldots, s_{i(d+1)}\} = \mathrm{conv}(\mathfrak{S}_i),$$

say, where s_{ij} are vectors in the finite set $\{\mathring{x}_i^{(c)} : c = 1, \ldots, 2^m\}$. Let $\mathfrak{C}_i$ denote the set of indices $c \in \{1, \ldots, 2^m\}$ corresponding to $\mathfrak{S}_i$. Then

$$\hat{\beta} \in \mathrm{conv}\{w^{-1} w^{(c)} \hat{\beta}^{(c)} : c \in \mathfrak{K}\} \tag{A.2}$$

would hold with $\mathfrak{K} = \cup_{i=1}^n \mathfrak{C}_i$.

To explore the strength of the condition $x_i \in \mathrm{conv}\{\mathring{x}_i^{(c)} : c \in \mathfrak{K}\}$ when $\mathfrak{K} = \mathfrak{F}$, a set defining a fractional factorial arrangement, consider $m = d$ so that the index i can be dropped in $\mathrm{conv}\{\mathring{x}_i^{(c)} : c \in \mathfrak{F}\}$, and $\mathring{x}^{(c)}$ represents a vertex in the design space. For instance, with $m = d = 3$, $\max\{x_{ij} : i \in \mathring{I}\} = 1$ and $\min\{x_{ij} : i \in \mathring{I}\} = 0$ for $j = 1, 2, 3$, the vertices corresponding to $\mathfrak{F}$ and $\overline{\mathfrak{F}}$ are $\{(1,1,1), (1,0,0), (0,1,0), (0,0,1)\}$ and $\{(0,0,0), (1,0,1), (0,1,1), (1,1,0)\}$ respectively. Figure 1 depicts $\mathrm{conv}\{\mathring{x}^{(c)} : c \in \mathfrak{F}\}$ and $\mathrm{conv}\{\mathring{x}^{(c)} : c \in \overline{\mathfrak{F}}\}$ for these two half-replicates of the 2^3 factorial.

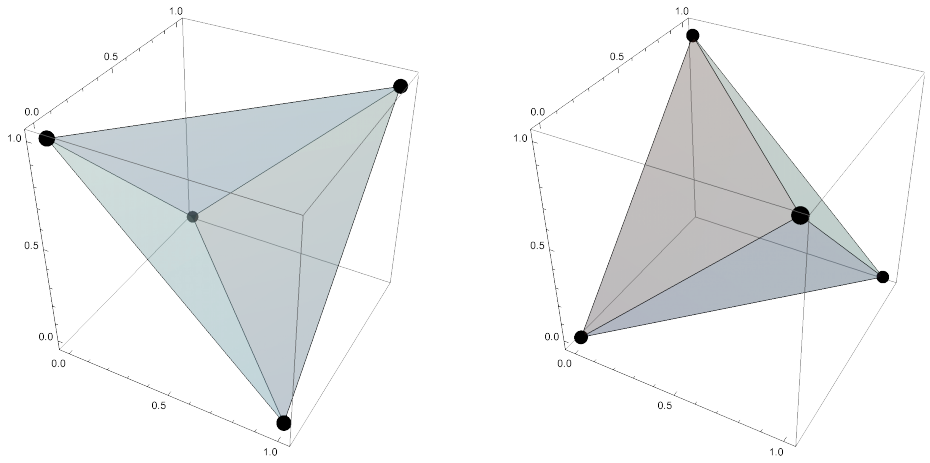

FIGURE 1. Convex hull of the four treatment combinations specified by the two half-replicates of the 2^3 factorial.

It is visually clear that the intersection $\mathcal{B}$, say, of $\text{conv}\{\mathring{x}^{(c)} : c \in \mathfrak{F}\}$ and $\text{conv}\{\mathring{x}^{(c)} : c \in \overline{\mathfrak{F}}\}$ is rather small, so that $x_i \in \mathcal{B}$ for all i is a strong requirement, especially for large m. The conclusion is only amplified when higher levels of fractionation are used. It follows that equation (A.2) will rarely be satisfied simultaneously when $\mathfrak{K}$ is taken as any of the relevant fractional factorial sets.

While worth pointing out, this conclusion is immaterial for assessing the sensitivity to missingness as detailed in §2.

APPENDIX B. PROOFS

B.1. **Preliminary lemmata.** Proofs of the two important results in convex analysis, Lemma B.1 and Lemma B.2 can be found in the stated references. These lemmata will be used in the proof of Proposition A.1, together with Lemma B.3, proved here.

Lemma B.1 (Caratheodory's convex hull theorem (e.g. Rockafellar, 1970, Theorem 17.1))**.** *Let $\mathcal{A}$ be any set of points and directions in $\mathbb{R}^d$, and let $\mathcal{C} = \text{conv}(\mathcal{A})$. Then $z \in \mathcal{C}$ if and only if z can be expressed as a convex combination of $d+1$ of the points and directions in $\mathcal{A}$ (not necessarily distinct).*

Lemma B.2 (Krein and Šmulian (1940, Theorem 3))**.** *For sets $\mathcal{A}$ and $\mathcal{B}$, denote their Minskowski sum by $\mathcal{A} + \mathcal{B} = \{a + b \mid a \in \mathcal{A}, b \in \mathcal{B}\}$. Minskowski summation commutes with the formation of convex hulls, that is:* $\text{conv}(\mathcal{A} + \mathcal{B}) = \text{conv}(\mathcal{A}) + \text{conv}(\mathcal{B})$.

Lemma B.3. *Let $\mathcal{A}$ and $\mathcal{B}$ be finite subsets of $\mathbb{R}^d$ and $\mathcal{A} \times \mathcal{B}$ their Cartesian product. Then* $\text{conv}(\mathcal{A} \times \mathcal{B}) = \text{conv}(\mathcal{A}) \times \text{conv}(\mathcal{B})$.

Proof. One inclusion, $\text{conv}(\mathcal{A} \times \mathcal{B}) \subseteq \text{conv}(\mathcal{A}) \times \text{conv}(\mathcal{B})$, is evident. To prove the converse, let $(a, b) \in \text{conv}(\mathcal{A}) \times \text{conv}(\mathcal{B})$. By Lemma B.1, there exists $a_1, \ldots, a_{d+1} \in \mathcal{A}$ and $b_1, \ldots, b_{d+1} \in \mathcal{B}$ such that $a = \sum_j \lambda_j a_j$ and $b = \sum_j \gamma_j b_j$, where $\lambda_j, \gamma_j \geq 0$ for all j and $\sum_j \lambda_j = \sum_j \gamma_j = 1$. Write

$$(a, b) = \left(\sum_{j=1}^{d+1} \lambda_j a_j, \sum_{j=1}^{d+1} \gamma_j b_j\right) = \left\{\sum_{k=1}^{d+1} \gamma_k \left(\sum_{j=1}^{d+1} \lambda_j a_j\right), \sum_{k=1}^{d+1} \lambda_j \left(\sum_{j=1}^{d+1} \gamma_j b_j\right)\right\} = \sum_{j,k} \lambda_j \gamma_k (a_j, b_k).$$

Since $\sum_{j,k} \lambda_j \gamma_k = 1$, $(a, b) \in \text{conv}(\mathcal{A} \times \mathcal{B})$. Thus $\text{conv}(\mathcal{A}) \times \text{conv}(\mathcal{B}) \subseteq \text{conv}(\mathcal{A} \times \mathcal{B})$. □

B.2. **Proof of Proposition A.1.**

Proof. Let $\mathcal{X}_i = \{\mathring{x}_i^{(c)} : c = 1, \ldots, 2^m\}$ and $\mathcal{M}_i = \{\mathring{x}_i^{(c)} \mathring{x}_i^{(c)\mathrm{T}} : c = 1, \ldots, 2^m\}$. By the constraint on the missing entries of $\mathring{X}$ relative to the two extremes of the corresponding column of X, $x_i \in \text{conv}(\mathcal{X}_i)$ and $(x_i, x_i) \in \text{conv}(\mathcal{X}_i \times \mathcal{X}_i) = \text{conv}(\mathcal{X}_i) \times \text{conv}(\mathcal{X}_i)$ by Lemma B.3. Thus, since the tensor product is a surjective bilinear map from $\mathbb{R}^d \times \mathbb{R}^d$ to $\mathbb{R}^d \otimes \mathbb{R}^d$, $x_i x_i^{\mathrm{T}} \in \text{conv}(\mathcal{M}_i)$. By the definition of the Minkowski sum, $\sum_i x_i x_i^{\mathrm{T}} \in \sum_i \text{conv}(\mathcal{M}_i)$ which is equal to $\text{conv}(\sum_i \mathcal{M}_i)$ by Lemma B.2.

Let $\mathcal{S}_i = \{\mathring{x}_i^{(c)} y_i : c = 1, \ldots, 2^m\}$. By Lemma B.1 there exists vectors $s_{i1}, \ldots, s_{i(d+1)} \in \mathcal{X}_i$ such that $x_i = \sum_j \lambda_{ij} s_{ij}$ where $\sum_j \lambda_{ij} = 1$ and $\lambda_{ij} \geq 0$ for all $i \in \{1, \ldots, n\}$ and all $j \in \{1, \ldots, d+1\}$. It follows immediately that $x_i y_i \in \text{conv}(\mathcal{S}_i)$, and from Lemma B.2 that $\sum_i x_i y_i \in \text{conv}(\sum_i \mathcal{S}_i)$.

Let $w = \sum_i x_i x_i^T$ and $w^{(c)} = \sum_i \mathring{x}_i^{(c)} \mathring{x}_i^{(c)\mathrm{T}}$. By the definition of the least squares estimator, $w\hat{\beta} = \sum_{i=1}^n x_i y_i$, the right hand side of which belongs to $\text{conv}(\sum_i \mathcal{S}_i)$. Since for all $c \in \{1, \ldots, 2^m\}$, $w^{(c)} \hat{\beta}^{(c)} = \sum_{i=1}^n \mathring{x}_i^{(c)} y_i$, it follows that

$$w\hat{\beta} \in \text{conv}(\mathcal{Q}), \tag{B.1}$$

where

$$\mathcal{Q} = \{w^{(c)} \hat{\beta}^{(c)} : c = 1, \ldots, 2^m\}.$$

For any $q \in \text{conv}(\mathcal{Q})$ there exists $q_1, \ldots, q_{(d+1)} \in \mathcal{Q}$ such that $q = \sum_j \alpha_j q_j$ by Lemma B.1, where $\alpha_j \geq 0$ for all $j \in \{1, \ldots, d+1\}$ and $\sum_j \alpha = 1$. Thus $w^{-1} q = \sum_j \alpha_j w^{-1} q_j$, showing that any $w^{-1} q$ for $q \in \mathcal{Q}$ belongs to

$$\text{conv}\{w^{-1} w^{(c)} \hat{\beta}^{(c)} : c = 1, \ldots, 2^m\}.$$

The conclusion follows by setting $q = w\hat{\beta}$. □

Acknowledgement: I am grateful to the two referees and the associate editor for their careful reading and very constructive suggestions.

Data accessibility statement: the bone marrow data and source code for reproducing the analysis is available from Dryad at the following URL:
https://datadryad.org/stash/share/p_d3Boi6ErHmrrjDLyxLbN1nlqHUMtOmGlq-i1lDgtc.

Funding statement: the work was partially supported by a UK Engineering and Physical Sciences Research Fellowship (EP/T01864X/1).

References

1. Bailey, R. A. (2008). *The Design of Comparative Experiments*. Cambridge University Press.
2. Battey, H. S., Cox, D. R. and Jackson, M. V. (2019). On the linear in probability model for binary data. *Roy. Soc. Open Science*, 6, 190067.
3. Cox, D. R. (1958a). Some problems connected with statistical inference. *Ann. Math. Statist.*, 29, 357–372.
4. Cox, D. R. (1958b). The regression analysis of binary sequences (with discussion). *J. Roy. Statist. Soc., B*, 20, 215–242.
5. Cox, D. R. (1970). *The Analysis of Binary Data*. Methuen, London.
6. Cox, D. R. and Reid, N. (2000). *The Theory of the Design of Experiments*. Chapman and Hall, London.

7. Davison, A. C., Koch, E. and Koh, J. (2019). Comment: models are approximations! *Statist. Sci.*, 34, 584–950.
8. Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc., B*, 39, 1–38.
9. Efron, B. E. (1977). Discussion of "Maximum likelihood from incomplete data via the EM algorithm" by Dempster, Liard and Rubin. *J. Roy. Statist. Soc., B*, 39, p.29.
10. Efron, B. E. (1982). Maximum likelihood and decision theory. *Ann. Statist.*, 10, 340–356.
11. Fisher, R. A. (1925). Theory of statistical estimation. *Proc. Camb. Phil. Soc.*, 22, 700–725.
12. Fisher, R. A. (1956). *Statistical Methods and Scientific Inference*. Oliver and Boyd, Edinburgh.
13. Kalwak, K. *et al.* (2010). Higher CD34$^+$ and CD3$^+$ cell doses in the graft promote long-term survival, and have no impact on the incidence of severe acute or chronic graft-versus-host disease after in vivo T cell-depleted unrelated donor hematopoietic stem cell transplantation in children. *Biol. Blood Marrow Transplant*, 16, 1388–1401.
14. Krein, M. and Šmulian, V. (1940). On regularly convex sets in the space conjugate to a Banach space. *Ann. Math.*, 41, 556–583.
15. McCullagh, P. (2023). *Ten Projects in Applied Statistics*. Springer, in press.
16. Rockafellar, R. T. (1970). *Convex Analysis*, Princeton University Press, Princeton, New Jersey.
17. Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley, New York.
18. Sundberg, R. (1974). Maximum likelihood theory for incomplete data from an exponential family. *Scand. J. Statist.*, 1, 49–58.

Department of Mathematics, Imperial College London, UK
Email address: h.battey@imperial.ac.uk

Nuffield College, University of Oxford, UK