Methods of Imputation for Missing Data (Fifth Draft)

Melvyn Weeks
Faculty of Economics and Politics
and Department of Applied Economics,
University of Cambridge,
Sidgwick Avenue, Cambridge CB3 9DE

May 1999

Contents

1. General

The problem of estimation and inference within the confines of a non-experimental modelling framework has generated a wealth of statistical techniques, much of which is predicated upon the fundamental dichotomy between that which is observed by decision makers (i.e. consumers, firms) and that which is observed by the analyst. Thus, even if we take the highly unlikely case of a model in which the functional form is known with certainty up to a finite dimensional vector of parameters, we must utilise statistical methods to recover these unknowns together with an indication of their sampling variability.

As soon as we depart from full knowledge of both the process generating the data and a fully observed sample of observations, we are faced with the difficult task of estimating unknown parameters based upon an incomplete sample of data. We examine this problem by providing a general framework for analysing missing data together with a taxonomy of missing value processes. In this respect it is instructive to distinguish between *intentional* missing data such as individuals that are not part of a survey and unintentional missing data generated by an intermediate filter between the population and the sample. Examples of the latter include nonrandom non-response and censoring. To the extent that the intermediate filter is nonrandom, it is necessary to model both the data generating process (*dgp*) and the missing-data generating process (*mdgp*).[1] In this respect it is important to differentiate between simple imputation of missing values and the model-based procedures which combine modelling of the phenomenon of interest with a technique for filling-in missing values.

In the case of non iid data the existence of autoregressive processes suggests that techniques used to infer missing values should incorporate these properties. In this context we will examine the Kalman filter which represents a model-based framework for imputing missing values. Missing data in household or business surveys data may be manifest in the form of incomplete questionnaires. For example, in a survey of firms some respondents may decline to report their profits. Dependent upon the pattern of nonresponse across questions, the subset of individuals providing complete information may be relatively small, and may constitute a non-representative sample.

The importance of missing data for statistical inference is in general self evident. For example, even in the case of data which is missing at random where model parameters are, in general unbiased and consistent, there will be an efficiency loss thereby compromising the validity of model estimates through wider confidence bands. As a result we need to distinguish between imputation tech-

[1] Within a likelihood setting formulation that includes both a model of the *dgp* and *mdgp* is referred to as a *complete data* likelihood.

niques which facilitate parameter estimation *and* methods to assess the variability of these estimates alongside the impact upon confidence intervals for parameter estimates based upon a combination of observed and imputed data. The question why is data missing for a particular cross-section or time series is not explicitly covered in this report. However, the extent to which data is missing according to a non-random mechanism obviously requires an understanding of this process if reliable imputation is to occur.

Although the principal focus of this paper is to survey methods for imputing missing data, we also examine an alternative approach based upon *weighting* the observed data. The weighting of observed data by the probability of response has generated a wealth of techniques in applied econometrics under the rubric of sample selection and self-selection adjustments. Here the primary objective is to introduce an adjustment such that inferences from a non-randomly generated observed sample to the population are still valid. We compare the use of these two approaches in section 5.

In section 2 we introduce a general framework for analysing missing data and in particular the notion of an observational rule. We also distinguish between ignorable and non-ignorable missing data mechanisms. In section 3 we compare classical Bayesian approaches to the missing data problem. Section 4 introduces a number of example and 5 presents a taxonomy of methods for imputing missing values in the context of non model-based procedures. Section 6 examines two model-based imputation procedures. In 6.1 we introduce the EM algorithm which combines parameter estimation (within the confines of maximum likelihood) with a technique for filling-in missing values. In 6.2 we examine another model-based procedure, the Kalman Filter, which is particularly suitable for missing data in time series. In section 7 we consider the usefulness of entropy and cross-entropy techniques for imputing missing values based upon a limited set of assumptions. In section 8 we move away from standard imputation and provide a brief overview of methods which emphasise the impact of imputation upon the variability of parameter estimates where a subset of data points have been imputed. In section 9 we focus upon a number of issues related to implementation, including the availability of computer software. Finally in section 10 we examine a specific missing data problem and compare a number of possible solutions.

2. Observed Data and Missing Data Mechanisms

2.1. Missing Data and Sample Selection

Given that the overriding principle of statistical analysis is the use of sample information to make inferences to a larger population, then the existence of missing data is the foundation stone of estimation and inference in classical statistics. For example, the assumption of an i.i.d. sample allows us to make inference from the sample (of size n) to the population. In this context each observation has a uniform weight equal to $1/n$. The extent to which elements of n have a non-uniform probability of being selected, takes us immediately into the world of sample selection and attendant methodologies. A key decision at this juncture is whether or not the process which determines sample selection is known.

In both cross-section and time-series models a predominant construct is the notion of a data generating process which describes how the variable(s) of interest are a function of a set of covariates and a stochastic error term. However, it is frequently the case that we may identify an intervening process that can mask the true process. For example, in microeconometrics a whole class of models (i.e. binary response, censored regression and duration models) can be specified based upon the notion of an *observational rule* (see Weeks (1998)). In time-series models, observations may be missing due to aggregation to a lower frequency. In both instances the critical issue is the extent to which observations are missing randomly or whether the missing value process is in some way related to the variable of interest.

The predominant characteristic of these models is that observational rules are, in general, based upon non-random processes which are intrinsically related to the phenomenon of interest. As such, in the parlance of the statistics literature, the resulting observed data does not represent an *ignorable* random sample of the underlying data.

Below we examine alternative definitions of missing data processes before considering a number of examples by examining the relationship between the observed data, Y, and the underlying (latent) data, Y^*. The use of the observational rule facilitates the representation of the data generating process as either an 'incomplete data' or 'partial observability' process. In addition we can distinguish between random and non-random missing data mechanisms.

2.2. Missing at Random: A Closer Look

Following seminal work by Rubin (1976), the dominant theme in the statistical literature on missing data has been a focus on imputation techniques which rely on the assumption that the missing data are 'missing at random' or *mar.* As noted above, this emphasis is distinct from the econometrics literature, where progress has been made in correcting for various forms of self-selection, where the relationship between Y and Y^* is governed by a systematic observational rule. In this section we focus upon *ignorable* missing data mechanisms and take a closer look at *mar.*

We motivate the ensuing discussion by considering a data matrix $M = \{m_{ij}\}$ $i = 1, ...n$, $j = 1, ..., k$, where i indexes individual data points and j covariates. In the analysis that follows, missing data will refer to a situation where data is missing for one or more rows and columns of M. Obviously the applicability of any particular form of imputation will depend, inter alia, upon the dimensionality of M, and whether inference is unconditional or conditional. Related, the most appropriate form of imputation will differ depending upon whether univariate or multivariate analysis is to be performed.

We make a distinction between the columns of M based upon the endogenous/exogenous (or predetermined) dichotomy, and in doing so we let Y denote the first column of M, and M_{-1} represent the ($n \times k$-1) matrix of exogenous variables. In addition we partition the n observations into two mutually exclusive sets: M_ℓ denotes fully observed data and $M_{n-\ell}$ denotes missing observations. Further let $Q = \mathbf{1}(M \text{ is observed})$ be an $(n \times k)$ vector with typical element $q_{ij} = 1$ if M is observed and zero otherwise. We examine the extent to which data is both missing at random and observed at random by examining the *joint* distribution of M and Q. For example, we may obviously write the joint distribution of Y and Q as

$$f(M, Q \mid \theta, \beta) = f(M \mid \theta) f(Q|M, \beta), \tag{2.1}$$

where θ and β are vectors of parameters, and $f(Q \mid M, \beta)$ is the distribution of the missing data mechanism. The key issue here is obviously when estimation and inference can be based upon M_ℓ, thus ignoring the missing data mechanism. If this can be done then instead of (2.1) we can write

$$f(M_\ell, Q \mid \theta, \beta) = f(M_\ell \mid \theta) f(Q \mid M_\ell, \beta).$$

It therefore follows that if we can simplify $f(Q \mid M, \beta)$ such that

$$f(Q \mid M_\ell, M_{n-\ell}, \beta) = f(Q \mid M_\ell, \beta), \tag{2.2}$$

then data is missing at random or *mar.* Note that if in addition to *mar* the parameters determining the pattern of missing data (Q) are distinct from θ, then the missing-data mechanism is ignorable. Putting (2.2) into words we now define *mar.*

Definition 2.1. *mar*

The missing data are missing at random (mar) if the conditional probability of the observed pattern of missing data, given the missing data and the value of the observed data (*namely* $f(Q \mid M_\ell, M_{n-\ell}, \beta)$)*, is the same for all possible values of the missing* data such that $f(Q \mid M_\ell, M_{n-\ell}, \beta) = f(Q \mid M_\ell, \beta)$*. In more general terms, the probability of missing data depends upon the observed but not the unobserved data.*

As Little and Rubin (1987) note, if the pattern of missing data is *mar* then, for example, the likelihood that a particular element of Y is missing does not depend upon the value of Y. Perhaps a more intuitive perspective is to recast the problem in terms of the ability to predict Q. In this respect, for a missing value process that is *mar*, there is no predictive power in Y.

We note that in Definition (2.1) we focus on the conditional distribution of Q in and do not differentiate between Y and the effect of covariates in M_{-1}. However, since the workhorse of applied economic analysis, the linear regression model, involves estimating the *conditional* distribution of Y given M_{-1}, then we might wish to consider an alternate representation of a missing-data mechanism.

Definition 2.2. *oar*

The observed data are observed at random (oar) if for each value of the missing data the conditional probability of the observed pattern of missing data, given the missing data and observed data, is the same for all possible values of the observed data.

If Definition 2.2 holds, then we can further simplify (2.1) by writing

$$f(Q \mid M_\ell, M_{n-\ell}, \beta) = f(Q \mid \beta), \tag{2.3}$$

such that the missing-data mechanism is completely independent of M. Note that based upon (2.3) β represents the *unconditional* frequency of missing values, such that $Q \sim \beta in(\beta, n)$.

To elaborate upon the distinction between *mar* and *oar* we consider a *single* covariate X which is recorded for all observations, whereas Y contains missing values. Both X and Y are confined to the (positive) half real line. Following Little and Rubin (1987) we differentiate between three missing value processes, where the probability of response is determined by:

i) Y and X;

ii) X and not Y;

iii) is independent of X *and* Y.

These three cases are presented in table 1, where α and δ are threshold constants that restrict the sample space of, respectively, X and Y^*. In the case of i) we have neither a random sample from the conditional or the unconditional distribution of Y, since the observed data is only recorded when $Y^* > \delta$ and $X > \alpha^2$. In the case of iii) $Y = Y^*$ for all possible values of both Y^* and X and as such

Table 1: Missing Data Mechanisms

Two Continuous Variables: X and Y.
Y subject to Non-response

		mar	*oar*
(i)	$Y = \mathbf{1}(Y^* > \delta \cap X > \alpha)Y^*$	×	×
(ii)	$Y = \mathbf{1}(Y^* > 0 \cap X > \alpha)Y^*$	✓	×
(iii)	$Y = \mathbf{1}(Y^* > 0 \cap X > 0)Y^*$	✓	✓

there is no intervening process that restricts the observability of Y^*. Therefore we can say that the data is missing at random (mar) and observed at random (oar).[3] Note that in this instance $f(Q|Y,X) = f(Q|\beta)$, since the pattern of missing data (represented by Q) cannot be predicted with any sample information. For ii) the missing data are missing at random (mar) since the probability of observing (or not observing) Y does not depend on the value of Y. However, since Y only equals Y^* if $X > \alpha$ the observed data are not observed at random. In this instance, the observed values of Y constitute a random subsample from the unconditional distribution of Y, but do not represent a non-random sample *conditional* on the values of X. We may therefore write $f(Q|Y,X,\beta) = f(Q|X,\beta)$.

Another way to think about a *mar* process in terms of the relationship between Y and Y^*, is to consider another level of random sampling. For example, if n is a random sample from N but we observe a sample $n' \subset n$, then a random missing value process effectively acts like a single non-parametric bootstrap. As a result, it is instructive to write the observational rule as

$$Y = AY^*,$$

where A is an $(n \times n)$ selection matrix. Obviously if A is an identity matrix there is no missing data. Letting $Q = diag(A)$ represent the diagonal elements of A, we might think of q_i as the outcome of independent Bernoulli trials such that for $q_i = 1$ (0) the ith observation for Y^* is observed (missing). For $\Sigma q_i = n' < n$ we have missing data. As a result, we have a *mcar* process if for the *ith* element in the sub-sample n' we have

$$E(Y_i) = E(Y_i^*|q_i = 1) = E(Y_i^*).$$

The importance of the type of missing data mechanism will depend upon the modelling framework. For example, in a regression framework the focus is, in gen-

[2] This type of missing-data mechanism is generally referred to as nonignorable.

[3] This type of missing data process is often referred to as *mcar*, or missing completely at random.

eral, upon the *conditional* distribution of Y given X, and as a result consistency of the estimated regression function (and associated parameters) requires that the data are *mar.* Obviously if the focus is upon the marginal distribution of Y then biases will appear unless the data are *mcar.*

3. Bayesian versus Frequentist Approaches to Missing Data

Imputation procedures are designed, in general, to provide a complete dataset so that statistical inference on one or more unknown parameters (θ) can take place with the most information possible. However, the appropriate form of imputation will depend upon the underlying approaches to inferences. For example, within a classical frequentist paradigm ... Missing data are easily integrated within a Bayesian setting. In this context the imputation of missing data adds an additional component of uncertainty in the construction of the posterior distribution of θ, denoted $g(\theta|y_{obs})$. In the presence of missing data $g(\theta|y_{obs})$ is constructed by averaging over the posterior distribution of the missing data given the observed, such that we may write $g(\theta|y_{obs})$ as

$$g(\theta|y_{obs}) = \int h(\theta|y_{obs}, y_{mis}) f(y_{mis}|y_{obs}) dy_{mis}, \tag{3.1}$$

where $h(.)$ denotes the conditional density of θ given the complete data, and $f(.)$ is the predictive density of the unobserved data conditional on what is observed. The idea of sampling from $f(.)$ to generate multiple values of y_{mis} was first introduced by Rubin (1987) and is now generally referred to as multiple imputations. (3.1) is often referred to as the *augmentation identity* (see Wei and Tanner (1990)).[4]

The problem of estimating $g(\theta|y_{obs})$ has been partially solved by the use of simulation. By taking multiple imputations from $f(y_{mis}|y_{obs})$ and computing

$$\hat{g}(\theta|y_{obs}) = \frac{1}{m} \sum_{m=1}^{M} \hat{g}^{m}(\theta|y_{obs}), \tag{3.2}$$

where M indexes the imputations. Rubin and Schenker (1986) emphasise that a Bayesian perspective provides the most natural theoretical framework with which to consider multiple imputation....

At the core of the Bayesian approach to inference is the posterior distribution of the unknown quantities such as model parameters or unobserved data. Probability statements are made conditional on the value of y_{obs}. As Gelman et al (1998) note, this conditioning upon observed data differentiates the Bayesian from the

[4] Distinction between two estimands: θ and y_{mis} (see Gelman et al (...)

classical approach to inference. The distinction between a Bayesian and classical likelihood-based approach to missing data is revealed if we compare the data augmentation approach with the EM algorithm. Notice that in (3.1) we take multiple draws from $f(y_{mis}|y_{obs})$. In contrast the use of EM algorithm in incomplete data problems replaces these multiple draws by the expected value of the missing data, conditional on the observed data and current values of parameter estimates. As King, Honaker, Joseph, and Scheve (1998) note, the Bayesian approaches preserves the whole distribution of the two estimands - the imputed values and the parameters - whereas the EM approach delivers the single, maximum posterior values.

Efron (1994) consider a number of nonparametric bootstrap approaches to the problem of missing data which are rooted within frequentist framework ...

4. Some Examples

Example 4.1. *Truncated and Censored Observational Rules*

The observed data Y is related to the underlying data Y^* by the observational rule

$$Y = \mathbf{1}(Y^* = \alpha + X^{'}\beta + \varepsilon > c)Y^*$$

where $Y^* = \alpha + X'\beta + \varepsilon$ is the population regression function, c is a threshold constant, and $\mathbf{1}(\cdot)$ is the indicator function. If we only observe Y^* if its value exceeds c, and that this condition also affects the observability of X, then we have the truncated regression model.[5] *In this instance the selection rule is endogenous and in the context of missing data taxonomy, we may think of this process as unit nonresponse on Y^* and X with Y being neither* mar *or* oar. *If the set of covariates X are observed for the complete sample, then Y fails the* mar *condition and is characterised by unit nonresponse. The censored (or Tobit) regression model has been used to model this type of data.*[6]

Example 4.2. *Missing Regularly (Stock and Flow Data)*

In time-series analysis a typical missing data problem is as follows. Let two variables Y_q and X_q be observed for a particular frequency, say quarterly (q). However, although data for X is observed at a higher frequency, say monthly (m), this is not true for Y. The basic approach to imputation in this context following seminal work by Chow and Lin (1971), is to establish a relationship

[5] We note that this particular representation of truncated data i.e. using a latent variable formulation, is based upon econometric 'theory' and may be distinct from a statistician's perspective.

[6] There is a plethora of articles that deal with this problem. The classic reference is Heckman (1979), and an informative overview is provided by Green (1997).

between Y_q and X_q and use this to impute Y_m whilst respecting any adding-up constraints.

The relationship between Y_m and Y_q (here Y^* and Y respectively) may be represented using a matrix C where C is a $(n \times 3n)$ matrix that converts high frequency (monthly) observations into lower frequency (quarterly) observations. C has a different structure depending upon whether Y is a stock or flow.

In the case of a stock we may write

$$Y = C_s Y^*$$

and for flows

$$Y = C_f Y^*$$

where C_s and C_f are given in Appendix 1.

Example 4.3. *Missing Regularly (Aggregate and Disaggregate Data)*

Below we consider the problem of missing data which arises when aggregate data are observed but disaggregate constituents are not observed. Examples occur in regional economic modelling and input-output analysis. Here we focus on three aggregates $Y^A = \sum_{i=1}^{R} Y_i, Z^A = \sum_{i=1}^{R} Z_i$ and $Q^A = \sum_{i=1}^{A} Q_i$, where in each case the aggregates are additive in R components. In addition we impose additional structure by assuming that Y^A, Z^A and Q^A are related by the identity $Y^A \equiv Z^A + Q^A$. This situation would occur if, for example, Y^A denoted population and Z^A, Q^A represent, respectively, total labour force and total unemployment. Note that in this instance we have $3(R-1)$ unknowns and therefore the problem is ill-determined using classical techniques. However, by rewriting the unknowns in terms of the proprotions of the observed aggregates, namely $w_i = Z_i/S_A$, $v_i = Q_i/Q_A$ and $\alpha_i = Y_i/Y_A$,we add additional information given that w_i, v_i and α_i are bounded. We now may represent the relationship between each aggregate and its constituent, incorporating the identity as

$$\begin{bmatrix} 0 & 0 & 0 & w_1 & w_2 & w_3 & v_1 & v_2 & v_3 \\ \alpha_1 & \alpha_2 & \alpha_3 & 0 & 0 & 0 & -v_1 & -v_2 & -v_3 \\ \alpha_1 & \alpha_2 & \alpha_3 & -w_1 & -w_2 & -w_3 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} Y_A \\ Y_A \\ Y_A \\ Z_A \\ Z_A \\ Z_A \\ Q_A \\ Q_A \\ Q_A \end{bmatrix} = \begin{bmatrix} Y^A \\ Z^A \\ Q_A \end{bmatrix} \tag{4.1}$$

noting that the following adding-up and non-negativity constraints apply

$$\Sigma\alpha_i = \Sigma v_i = \Sigma w_i = 1 \tag{4.2}$$
$$\alpha_i, v_i, w_i \geq 0 \quad \forall_i = 1, \ldots, R$$

In section 10 we examine the usefulness of entropy-based techniques for finding estimates of the unknown parameters (w_i, v_i and α_i).

5. A Taxonomy of Methods with Partially Missing Data

Below we provide a classification of missing data processes. The taxonomy covers both parametric and non-parametric models, and iid and non-iid data structures. Our principal focus will be on imputing data that is missing for a single endogenous variable Y. The n data points are, as above, partitioned into ℓ observed, $n - \ell$ missing and in most cases we assume that the matrix M_{-1} is fully observed. In each case we indicate the assumptions underlying the imputation (other than standard least squares), using the *mar/oar/mcar* distinction.

1. **Case Deletion**[7]

   Row i of M is deleted if at least one element of row i (m_{ij}) is missing. This method makes no use of the *observed* data for i.

   Information assumptions: *mcar.*

   Problem: as k (the number of columns in M) increases the probability of discarding any given observation increases.

2. **Imputation: Unconditional**

   Below we list two approaches based upon unconditional imputation. Both are based upon different degrees of smoothing.

   i) The most common form substitutes the mean (over observed data) for missing values.

   ii) An alternative procedure is to use information on one or more categorical auxiliary variables which may serve to improve the accuracy of the imputation. Consider the case of a single categorical variable v, which has k cells. Unconditional imputation using v simply replaces missing values with the average of the *observed* values in each k_{th} cell. In this respect we note that the form of the imputation utilises an ANOVA model.

Information assumptions: *mcar*

$$\begin{aligned} \text{i) } E(m_{ij}) &= \mu_j \ \forall_i \\ \text{ii) } E(m_{ij}) &= \mu_{jk} \ \forall_{ik}. \end{aligned}$$

[7] This procedure is also known as complete-case analysis.

3 Imputation: Conditional and Independent (y continuous)

Conditional imputation methods such as those first proposed by Buck (1960), are predicated upon a regression-based modelling strategy. That is, the focus is upon the conditional expectation (or regression function) of the random variable rather than unconditional.

Example: Assume that the underlying population regression function is given by

$$y = \alpha + M_{-1} \cdot \beta + \varepsilon \tag{5.1}$$

where α and β, are respectively an unknown scalar and a (k-1 × 1) vector of unknown parameters. The stochastic error term ε is iid $(0, \sigma^2)$. A common form of conditional imputation for the missing subset $n - \ell$[8] uses the expectation

$$E_\ell(y_{n-\ell} \mid M_{-1}) = \alpha + M_{-1,n-\ell}\beta, \tag{5.2}$$

and estimate

$$\widehat{E}_\ell(y_{n-\ell} \mid M_{-1}) = \widehat{\alpha} + M_{-1,n-\ell}\widehat{\beta}, \tag{5.3}$$

where $E_\ell(\cdot)$ denotes that the expectation is taken with respect to the information available - namely over the partition ℓ.

Information assumptions: *mar*

4 Random Regression-Based Imputation

Little (1988) has suggested that imputations should be drawn from the predictive *distribution* rather than simply substituting the conditional mean. Employing this method, Wang and Jinn (1992) refer to two variants of conditional imputation both based upon a modification to correct the systematic underestimation of the variance of the resulting $n \times 1$ vector containing a mixture of observed and imputed values.

i) Random residuals (e) are drawn from $N(0, \hat{\sigma}^2_\ell)$ where $\hat{\sigma}^2_\ell = \sum\limits_{\ell-1}^{\ell}(y_i - \hat{y}_i)^2/(\ell - k)$. These are then added to the conditional expectation (5.3). The $\ell_{th} + 1$ imputed value is therefore

$$\hat{y}_{\ell+1} = \hat{\alpha} + M_{\ell+1}\hat{\beta} + e_{\ell+1}. \tag{5.4}$$

[8] Since the observations are independent then the partition of observations into subsets ℓ and $n - \ell$ is possible.

ii) As above, random residuals are added to (5.3). However, a total of $(n-\ell)$ residuals are randomly sampled (with replacement) from the ℓ observed residuals $\hat{\varepsilon}_i = y_i - M_i\hat{\beta}$, $i = 1, ..., \ell$.

Information assumptions: *mar.*

Obviously the limitation of (i) is the assumed normality, whereas (ii) samples from the predicted residuals. A modification of this approach has been used by ? with the idea to utilise resampled residuals for which the predicted (not imputed!) value of y is close to that of the nonrespondents.

5 Imputation: Conditional and Independent (y is discrete)

We note that the regression-based imputation procedure above implicitly assumes that the dependent variable is continuously distributed. In many instances Y may be binary or polychotomous, and although this does not invalidate the logic of conditional imputation, the procedure used must be modified so that the imputed values have the same characteristics as the observed values. Below we consider a variant of (**??**) where Y is a binary random variable.

As above n data points are partitioned into ℓ observed, $n-\ell$ missing and the matrix M_{-1} is fully observed. In this instance the only modification to (5.2) is the introduction of a *link* function F, which maps the conditional expectation $E_\ell(\cdot)$ into the unit interval. For this particular data variant we may then rewrite (5.2) as $E_\ell(y_{n-\ell} \mid M_{-1}) = F(\alpha + M_{-1,n-\ell}\beta)$. Common forms for $F(\cdot)$ are the logistic and standard normal cumulative distribution function.

Information assumptions: *mar*

6 Imputation: Conditional and Non-independent.

Based upon the discussion in example (2.3) the relationship between Y and X for the lower frequency (quarterly) observation may be written

$$Y_q = CY_m = CX_m\beta + C\varepsilon = X_q\beta + \varepsilon_q. \tag{5.5}$$

Given an estimated form of (5.5), the interpolation (or distribution) of say p observations for Y_m is based upon the estimator

$$\hat{Y}_m = X_m\hat{\beta} + (V_m V_q^{-1})\hat{\varepsilon}$$

where $\hat{\beta}$ is the estimated regression coefficients from (5.5), $V_q = E(\varepsilon_q\varepsilon_q^{'})$ and $V_m = E(\varepsilon_m\varepsilon_m^{'})$. The principal problem with the above approach is that imputation is based upon the identification of a *static* relationship between the *levels* of the two series. As noted by Salazar, Smith, Weale, and Wright (1997), since the stationarity properties of the two series will determine whether estimation in levels or differences is appropriate and most regressions are estimated in logarithms, this approach is problematic. The procedures suggested by the authors are based upon estimating links between the interpoland and the higher frequency variable which accounts for lags and logarithmic terms[9].

Information assumptions: *mar*.

7 Imputation: Missing Covariates

To date we have introduced a number of imputation techniques to handle the problem of missing data in the endogenous variable Y with a fully observed matrix (M_{-1}) of covariates. Below we reverse the problem and assume that Y is fully observed and that one of the covariates, say $X_1 = M_{-1,1}$ has missing values. As noted above, the key issue in terms of appropriate imputation technique is the form of the missing data mechanism. For example, Little (1992) notes that the probability that X_1 is missing for one or more observations may be a) independent of data values, b) depend on X_1, c) depend on X_1, $M_{-1,2}, ..., M_{-1,k}$, or $d)$ depend on $M_{-1,2}, ...M_{-1,k}$ *and* Y.

A taxonomy of methods imputing missing covariates is provided by Little (1992) and covers a similar range of procedures as outlined above such as case deletion, unconditional and conditional imputation, alongside methods based upon combining models for the data and missing-data mechanism. Although the details are in principle the same as imputing missing data for an endogenous variable, below we examine conditional mean imputation.

Conditional Mean Imputation

We assume that the underlying population regression function is the same as (5.1) and instead of missing data on Y, we only observe ℓ data points for X. In this instance conditional mean imputation follows exactly the same logic as outlined in Section 3, except that we treat the missing values as random variables and utilise an auxiliary regression based upon the conditional expectation

$$E(X_1|M_{-1,2}, ..., M_{-1,k}).$$

Assuming that the missing data mechanism is *mar*, then the application of OLS using observed and imputed data produces consistent parameter

[9] See Harvey (1989) for further discussion.

estimates. However, following the work of Gourieroux and Montfort (1981), we note that the use of imputed data adds an additional error component thereby inflating the residual variance.

Information assumptions: *mar.*

8 **Imputation: Nonparametric**

The taxonomy of imputation procedures outlined in Section 3 included a number of methods based upon a *parametric* conditional expectation (or regression) function. An alternative procedure utilises a non-parametric approach[10] and thereby avoids the imposition of parametric assumptions on the conditional expectation function. The potential of non-parametric techniques for missing value imputation stems, in part, from a focus upon the "local" shape of the conditional mean function. For example, consider the case where a single endogenous variable Y has missing values. A parametric regression approach, imputes missing values based upon (5.3) and thereby attaches equal weights to components of M_{-1}. In contrast, a Kernel density estimator of the missing values (for a given bandwidth) is simply a *weighted* average of the covariates. We may also motivate nonparametric approaches to missing data by interpolating between adjacent points, as in time series.

5.1. Non-Ignorable Response

The principal problem with many of the approaches listed above is the assumption that the probability of nonresponse is independent of Y (i.e. *mar*) and as such ignorable (see Little and Rubin (1987)). For example, regression based (conditional) imputation (see Section 5, examples 3, 4, 5, and 6) are non-valid if, for example, the *dependent* variable is constrained to lie within a given interval. This was demonstrated in Example 2.1 in the case of truncated regression. In this instance $E(Y|X) \neq X\beta$ (even if β is unbiased) since the stochastic component will be correlated with X and will not have zero mean. As a result imputation based upon the $X\hat{\beta}$ will produce biased estimates of the missing values (see Greenlees, Reece, and Zieschang (1982) for further details). This is apparent if we consider a simple univariate case. If the observational rule is $Y = \mathbf{1}(Y^* > c)Y^*$ and we impute the missing values (i.e. observations for $Y^* < c$) using the set of observed values, imputed values will be systematically too large.

The problem of making inferences in models subject to a *nonignorable* nonresponse has been treated extensively within the econometrics literature (see Horowitz and Manski (1998)). A critical distinction can be made between methods which apply a set of observation-specific weights to the observed data, and

[10] See Hardle (1990) for an excellent overview of nonparametric regression.

techniques which utilise the observed data and information on the *mdgp* to impute missing values. We compare these two approaches by considering an example which has generated a voluminous literature - namely missing wage data for individuals not in the labour market. The generic form of this problem is known as stochastic censoring and involves the joint distribution of two variables, say y_1, and y_2, which we write as

$$\left(\begin{array}{c} y_1 \\ y_2 \end{array} \right) \sim BVN\left(\left[\begin{array}{c} x_1\beta_1 \\ x_2\beta_2 \end{array} \right], \left[\begin{array}{cc} \sigma^2 & \rho \\ \rho & 1 \end{array} \right] \right) \tag{5.6}$$

where $BVN(\mu, \Sigma)$ denotes the bivariate normal distribution with mean μ and covariance Σ. x_1, x_2 denote possibly overlapping sets of covariates, β_1 and β_2 are unknown parameter vectors, y_1 is a partially observed outcome variable and y_2 is an unobserved variable which controls the observability of y_1. In this example y_1 represents the observed wage for labour market participants and $y_2 = y_1 - W_R$, where W_R is the reservation wage. y_1 is observed *iff* $y_2 > 0$. Subsequently we may write the conditional regression function for the two groups as follows:

Labour Force Participants.

$$E(y_1|y_2 > 0) = x_1\beta_1 + \alpha\phi\underbrace{(\gamma)/(1 - \Phi}_{Q_1}(\gamma)) \tag{5.7}$$

Non Participants

$$E(y_1|y_2 < 0) = x_1\beta_1 - \alpha\underbrace{\phi(\gamma)/\Phi(\gamma)}_{Q2} \tag{5.8}$$

where $\gamma = x_2\beta_2$. ϕ (Φ) denotes the density (distribution function) of the standard normal distribution.

Note that if we base inference on sample observations for which $y_2 > 0$, then we use (5.7) where Q_1 (the inverse of the Mills ratio) represents an artificial regressors used to correct for the non-random selection of the data. Alternately, in other situations we may require estimates of wage rates for non-participants then following the work of Heckman (1976), the following two-stage procedure is now well known.

1. Run a probit regression of $z = \mathbf{1}(y_2 > 0)$ on x_2 and create an artificial variable $\hat{\gamma} = -x_2\widehat{\beta}_2$.

2. Estimate β_1, α and σ^2 by running a regression of y_1 on x_1 and Q_2.

The Heckman procedure is highly sensitive to model misspecification particularly with respect to bivariate normality and the division of the total set of covariates into sets x_1 and x_2. Olsen (1980) has noted that if x_1 and x_2 coincide, then the model is only identified by the nonlinear transformation on $\phi(\gamma)/\Phi(\gamma)$. In practical applications it is necessary for x_1 and x_2 to be distinct. However, prior knowledge as to the appropriate set of zero restrictions may be lacking.

Note that based upon (5.8) a test of whether the censored wage data may be considered *mar* or equivalently if there is no sampling selectivity - is a test of $\alpha = 0$. In addition if $x_1 \neq x_2$ then the missing-data mechanism is ignorable such that estimation of the parameters β_1 and σ^2 based upon an application of OLS on the participants will generate consistent and efficient parameter estimates.

In an analysis of imputation techniques applied to missing wage and salary data in the Current Population Survey (conducted by the Census Bureau), a number of studies have utilised a quasi-experimental framework. With access to a secondary data source from the Internal Revenue Service (IRS), the authors are able to compare imputed data with nonrespondents IRS wage data. Greenlees, Reece, and Zieschang (1982) find that an approach which utilises a stochastic censoring model represents an improvement over a standard regression approach which assumes that the missing data process is *mar.* Given access to a complete IRS wage series the authors were able to test hypothesis which in most circumstances are not verfiable. For example, a negative and highly significant coefficient on the wage variable in a probability of response model resulted in the rejection of non-ignorable nonresponse. In addition, despite finding approximate symmetry, a large kurtosis value on the residuals from the wage equation resulted in the rejection of the normality hypothesis.

David, Little, Samuhel, and Triest (1986) compare the CPS hot deck method for imputing wages with regression approaches. One of the distinctive features of this particular study is that. ....The hotdeck approach to imputation is based upon the use of a set of fully observed covariates (in this case age, sex and education) to allocate respondents and nonrespondents to groups which possess similar characteristics. Thereafter respondents act as donors, with nonrespondents assigned a particular respondents wage data[11]. Below we consider a variant of this approach based upon an adaptation of a bootstrap methodology which has been used for Bayesian imputation methods.

With reference to the stochastic censoring problem described above, let $\Phi(x_2\hat{\beta})$ represent an estimate of the probability that $y_2 > 0$. Using the quintiles of $\Phi(x_2\hat{\beta})$ we allocate the n data points into 5 equal parts. Within each quintile let Mq_j (Oq_j) denote the number of censored (observed) values on y_1. The total number

[11] See Lillard, Smith, and Welch (1982) and Little and Rubin (1987) for details.

of observed (censored) values is therefore

$$\sum_{j=1}^{5} Oq_j = n_1; \sum_{j=1}^{5} Mq_j = n_O \quad (n = n_1 + n_O). \tag{5.9}$$

A bootstrap variant of the hotdeck approach is based upon the following procedure. Draw a random sample with replacement of size Mq_j from Oq_j. Index these draws by m and repeat M times. An estimate of censored observation i in the jth quintile is given by

$$\hat{y}_{ij} = \frac{1}{M}\sum_{m=1}^{M} y_{Oj}^{m}, \tag{5.10}$$

where y_{0j} denotes an observed value in the jth quantile.

An advantage of the bootstrap procedure is that it does not impose an assumption of bivariate normality, nor does it impose a parametric model for the mean equation when imputing censored observations. i.e. it does not assume that imputed values are generated by (5.8). Instead differences between the two groups are controlled for by the specification of a missing data probability model. Thereafter the set of potential 'donor' values ($y_1 > 0$) are then stratified based upon the quintiles of $\Phi(\gamma)$. In this respect, the use of bootstrap approach is predicated on the specification of a probability model with high predictive power. The principal drawback of this methodology are analogous with the problems encountered with hotdeck imputation. Respondents and nonrespondents have the same wages distribution within the quintile defined by the missing-data probability model.[12] As noted above, the better the probability model, the better are we able to control for possible differences between the two groups.

To examine this procedure more closely we consider four types of individuals for whom $z = 1$. Let these four types be referred to as Group A $= \{a_i, b_i, c_i, d_i\}$. Based on observed wages (W) for Group A and $\Phi(x_{i2}\hat{\beta})$ we write the joint distribution of high/low W and high/low $\Phi(x_{i2}\hat{\beta})$ as

$\Phi(x_{i2}\hat{\beta})$

		low	high
W_i	low	a_i	b_i
	high	c_i	d_i
		e_i	f_i

[12] This implies that the 'nonresponse' mechanism is ignorable.

In the case of individuals for which wage data is unobserved (i.e. $2 = 0$) we define two types of individuals which we denote Group B $= \{e_i f_i\}$. a_i and c_i are potential 'donors' for e_i; b_i and d_i are potential 'donors' for f_i. By treating all censored observations within the jth quintile as iid, the bootstrap approach cannot account for the problem depicted in the diagram. Since the covariate sets x_1 and x_2 are not identical there is not a one-to-one mapping from a low $F(x\hat{\beta}_2)$ to the wage rate. i.e. $E(\hat{W}_{ij})$ is constant for each jth quintile. The obvious disadvantage of this assumption may be circumvented by increasing the number of percentiles on $F(x\hat{\beta}_2)$. Therefore we randomly sample from a_i, c_i to impute values for e_i. In contrast, the Heckman procedure adjusts the predicted conditional mean $x_i\hat{\beta}_1$ for selection bias - in this case $E(y_1|y_2 < 0) = E(\varepsilon_1|\varepsilon_1 > x_2\beta_2)$. Therefore we utilise both unconditional mean information and an adjustment based upon $P(y < 0)$.

Also

1. $\lim \frac{1}{M} \sum \hat{W}_{ij}^m = \bar{W}_{0j}$.
2. $W_{0j}^{\min} \leq \hat{W}_{ij}^m \leq W_{0j}^{\max}$

6. Model-Based Imputation Techniques

6.1. The EM Algorithm

The consistency of parameter estimates subsequent upon conditional and unconditional imputation techniques described above depends upon the assumptions that the missing data are missing at random (mar) and that the observed data are observed at random (oar). Namely, the missing data is missing completely at random ($mcar$). The model-based imputation technique using the EM algorithm only requires the weaker mar assumption.

The estimation-maximisation (EM) algorithm initially proposed by Dempster, Laird, and Rubin (1977) is a general framework for solving maximum likelihood problems when an observable model is derived from an underlying latent model. In this respect it is instructive to consider the observational rule introduced in Section 4, which facilitates comparison of an observed endogenous variable Y, and a latent variable Y^* via the mapping $Y = \lambda(Y^*)$. Thus, if Y^* were fully observed $\lambda(.)$ could be written as $\mathbf{1}(-\infty < Y^* < \infty)Y^*$, such that the maximum likelihood estimator of θ, (the vector of unknown parameters) would be the solution to the maximisation of log $\ell(y^* \mid x; \theta)$. In situations where the mapping from Y^* to Y results in either (non-random) missing observations or partial observability[13], the

[13] We point out that the EM algorithm is applicable in situations of both ignorable and non-

expectation step of the EM algorithm introduces the following criterion function

$$Q(\theta,\ \theta^t) = E_{\theta^t}(\log \ell(y^* \mid x;\ \theta) \mid Y = y)$$

which is the expectation of log $\ell(y^* \mid x,\ \theta)$ *given* the observed data y. θ^t denotes the value of θ at the t_{th} iteration of the optimisation routine.

The most transparent way to motivate this method is to present the following generic algorithm (from Little and Rubin (1987)) for handling missing data.

1. Replace missing values with *estimated* values

   In the EM algorithm this is the **E**xpectation step, where estimated values are based upon the expectation of the missing data, conditional upon the observed data and any estimated parameters.

2. Estimate model parameters

   This is the **M**aximisation step.

3. Re-estimate the missing values based upon 'updated' parameter values

4. Re-estimate model parameters

5. Iterate until prespecified convergence criteria is met.

Note that the iterative component of the EM algorithm follows from estimation by maximum likelihood. If there exists a closed form solution for the parameters of interest (as in ordinary least squares) then the algorithm would truncate at 2. Rubin and Little provide an extensive overview of both the theory and applications of the EM algorithm covering a broad range of multivariate statistical techniques.

6.2. The EM Algorithm for Incomplete Multivariate Normal Samples

Using the notation of Section 2 we assume that the k columns of M $(M_1, M_2, ..., M_k)$ $\sim MVN(\mu, \Sigma)$ where $\mu = (\mu_1, ..., \mu_k)$ and Σ is the $(k \times k)$ covariance matrix.

We write the observed data as

$$M_{obs} = (M_{obs,1}, M_{obs,2}, ..., M_{obs,n}),$$

where $M_{obs,i}$ represents the set of *variables* which are observed for observation i, $i = 1, ..., n$. The *expectation* step of the algorithm is written as

$$E(M_{ij} \mid M_{obs}, \theta^{(t)}) = M_{ij}^{(t)},$$

ignorable missing-data mechanisms. In the case of the latter the **E**xpectation step is modified to condition on the process determining the observed data.

where $M_{ij}^{(t)}$is written

$$M_{ij}^{(t)} = \mathbf{1}(M_{ij} \text{ is observed})\ M_{ij} + \mathbf{1}(M_{ij} \text{ is missing}) \cdot E(M_{ij} \mid M_{obs,i}, \theta^{(t)}). \quad (6.1)$$

In (6.1) we see that the EM algorithm partitions the data into *observed* and missing subsets, and at each iteration missing values are replaced by the conditional mean of M_{ij} given the data observed for that observation ($M_{obs,i}$), and current values of the parameter vector θ. The maximisation step, again assuming that θ is not available in closed form, simply updates $\theta^{(t)}$ by $\theta^{(t+1)}$ using the (estimated) complete data.[14]

6.3. Time Series

In section 4 we examined the case of imputing values of a time-series when observations are subject to contemporaneous aggregation, and there exists an 'indicator' variable which in a regression framework can be used to interpolate higher frequency observations. In instances where observations are simply missing for certain periods this is not possible. In this case a number of alternate methods are possible.

Below we examine a situation where for a time-series y_t $(t = 1, ..., m-1, m+1, ..., T)$ there is a missing observation at time $t = m$.

Example 6.1. $y_t = \alpha y_{t-1} + \varepsilon_t \quad (|\alpha\ | < 1)$
Using simple recursion, y_{m+1} *may be written*

$$y_{m+1} = \alpha(\alpha y_{m-1} + \varepsilon_m) + \varepsilon_{m+1} = \alpha^2 y_{m-1} + \varepsilon_{m+1} + \alpha\varepsilon_m.$$

Therefore, since $y_{m+1} = y_m$ *by definition, the best predicator of* y_m *is given by*

$$E(y_m \mid \Omega_{m-1}) = \alpha^2 y_{m-1},$$

which may then be substituted into the likelihood, along with the remaining $T-1$ *fully observed components.* Ω_{m-1}*denotes the information set at period* $m-1$.

6.3.1. The Kalman Filter

The Kalman filter, based upon the representation of a dynamic system in a state-space form, is an algorithm for sequentially updating a linear projection for the system. In this respect its potential for use within the context of missing values is immediately obvious. Below we briefly introduce the main components of both

[14] Details such as the importance of starting values are discussed at length in Rubin and Little. Convergence properties are discussed in Ruud (1991).

the state space model and the Kalman filter and link the exposition to a number of missing value problems.

Let y_t denote an $(n \times 1)$ vector of variables observed at date t. The state-space representation of the dynamics of y may be represented by the following system of equations:

$$\begin{aligned} \delta_{t+1} &= F\delta_t + v_{t+1} && (6.2) \\ y_t &= A'q_t + H'\delta_t + w_t && (6.3) \end{aligned}$$

where F, A' and H' are, respectively, matrices of parameter of dimension $(r \times r)$, $(n \times k)$ and $(n \times r)$. q_t is a $(k \times 1)$ vector of exogenous or predetermined variables. Equation (6.2) is the *state* equation and (6.3) is the *observation* equation. v_t and w_t are, respectively, $(r \times 1)$ and $(n \times 1)$ vectors of mean zero, iid observations.[15]

As demonstrated below, the formulation of the state-space is based upon writing any finite-ordered dynamic system as a first-order system, thereby simplifying the analysis (see (6.2) and (6.3)). The state vector δ_t carries all of the dynamics of the process, which itself is determined by the state (or transition) equation. Allowance for observational error is made using the observation (or measurement) equation[16]. It is of course true that all linear (and many non-linear) models in econometrics may be presented in a state-space form. However, in relatively simple models such as an AR(p) process, the state-space formulation is not such an advantage.

The following example will clarify notation.

Example 6.2. *AR(p) process*

$$y_{t+1} = \phi_1(y_t) + \phi_2(y_{t-1}) + ... + \phi_p(y_{t-p}) + \varepsilon_{t+1} \qquad (6.4)$$

where $E(\varepsilon_t \varepsilon_r) = \sigma^2 \ \forall_t = r$ *and zero otherwise. The state space form proceeds by writing the pth order difference equation (6.4) as a first-order difference equation in a vector* δ_t*, where the first element of* δ_t *is the value of* y *at time* t_1 *and so on.*

State (Transition) Equation $(r = p)$:

[15] For further details see Harvey (1989) and Hamilton (1994).

[16] Much of this discussion is taken from Diebold (1992).

$$\begin{bmatrix} \underbrace{\begin{matrix} y_{t+1}-\mu \\ y_t-\mu \\ \vdots \\ y_{t-p+2}-\mu \end{matrix}}_{\delta_{t+1}} \end{bmatrix} = \begin{bmatrix} \underbrace{\begin{matrix} \phi_1 & \phi_2 & \cdots & \phi_{p-1} & \phi_p \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{matrix}}_{F} \end{bmatrix} \begin{bmatrix} \underbrace{\begin{matrix} y_t-\mu \\ y_{t-1}-\mu \\ \vdots \\ y_{t-p+1}-\mu \end{matrix}}_{\delta_t} \end{bmatrix} + \begin{bmatrix} \underbrace{\begin{matrix} \varepsilon_{t+1} \\ 0 \\ \vdots \\ 0 \end{matrix}}_{v_{t+1}} \end{bmatrix} \tag{6.5}$$

Observation (Measurement) Equation $(n = 1)$:

$$y_t = \underbrace{\mu}_{A'q_t} + \underbrace{[1,\ 0 \ldots 0]}_{H'=2} \begin{bmatrix} \underbrace{\begin{matrix} y_t-\mu \\ y_{t-1}-\mu \\ \vdots \\ y_{t-p+1}-\mu \end{matrix}}_{\delta_{t+1}} \end{bmatrix} \tag{6.6}$$

Note that the first equation in the system is identical to (6.4), the second equation is simply the identity $y_{t-1} = y_{t-1}$ and so on[17]. In this way we see that first-order vector difference equation (6.2) - the state equation - is an alternative representation of the pth order *scalar* system (6.4), but has the advantage of being expressed as a first-order system.

6.3.2. Projection and Missing Values

In Section 3 we considered a number of regression-based imputation procedures, which essentially utilised conditional expectations. The Kalman filter may be viewed in a similar light, namely as an algorithm for calculating forecasts of the state vector based upon information available up to time t, (hereafter the information set, Ω_t). Using notation introduced in Section 4.3.1we may write this forecast as

$$\widehat{\delta}_{t+1|t} \equiv \widehat{E}(\delta_{t+1}|\Omega_t),$$

where $\Omega_t = (y_t, y_{t-1}, \ldots, y_1, x_t, x_{t-1}, \ldots x,)'$. $\widehat{E}(\cdot)$ is sometimes referred to as the linear projection of δ_{t+1} on Ω_t, but is, within the linear context, no different than the conditional expectation function.

[17] Note that in this case $A' = \mu$, $x_t = 1$ and $w_t = 0$.

The Kalman filter calculates one-step-ahead forecasts recursively, proceeding logically from unconditional projection at time $t = 1$ (i.e. $\widehat{E}(\delta_1) = \widehat{\delta}_{1|0}$), to conditional projection for $t = 2$ (i.e. $(\widehat{E}(\delta_2|\Omega_1 = (y_1, x_1)) = \widehat{\delta}_{2|1})$.[18] Generically the one-step-ahead forecast of the state variable is given by

$$\begin{aligned} \widehat{\delta}_{t+1|t} &= \widehat{E}(\delta_{t+1}|\Omega_t) \\ &= F \cdot \widehat{E}(\delta_t \mid \Omega_t) + \widehat{E}(v_{t+1}|\Omega_t) \\ &= F \cdot \widehat{\delta}_{t|t} + 0. \end{aligned} \tag{6.7}$$

Based upon the above analysis, the logical step to missing values is simple given that missing-values in time-series fall under the rubric of multi-step prediction. For example, given (6.2), (6.3), and (6.7), the formulae for the m-period ahead forecast, y_{t+m}, emerges from application of the Kalman filter

$$\hat{y}_{t+m} \equiv \hat{E}(y_{t+m} \mid \Omega_t) = A' q_{t+m} + H^{'} F^m \widehat{\delta}_{t|t}.$$

In this instance we may either think of information being available for $t = 1, ..., T$ and forecasting m periods ahead, or where a window of length m is missing within a particular series.

7. Maximum and Cross-Entropy Formulation

The major problem with the parametric approaches to imputing missing values as laid out in Section 5, is the need to specify some form of conditional mean regression model. Obviously the validity of such a procedure is completely dependent upon the parametric specification of the regression equation, the availability of a sufficiently large sample, and that covariates which enter the specified equation are observed for the sub-sample where the left-hand side variable is missing.

An alternative approach to this problem is to use either maximum or cross-entropy formulations. The fundamental distinction between this and the regression-based approach is the concept and valuation of information. In the classical regression framework the concept of information is inextricably linked to both the number of observations and covariates. Within the maximum entropy framework, adding information involves the inclusion of additional constraints which if consistent with the data will reduce the entropy value. For example, in the case of unknown regional wage data (and assuming a discrete probability distribution[19])

[18] Typically $\hat{\delta}_{1|0}$ is set at 0.

[19] We may formulate a continuous version of ME as the number of possible states → ∞.

if we reduce the number of unknowns, then the degrees of freedom will fall, and so will the number of possible ways of recovering the remaining unknowns.

To understand the concept of maximum entropy (ME) consider a set of unobservable frequencies $p = (p_1, p_2, ..., p_k)$ which represent a data generating process for a random variable with k possible outcomes. The ME principle seeks to maximise a function of p, such that the chosen p can be realised in the largest number of ways, *conditional* upon what we know. For example, based upon knowledge of a fair dice, the p that maximises entropy (or uncertainty in the system) is $p_1 = p_2 = ... = p_6 = 1/6$, with attendant mean 3.5. Thus, if we know nothing and the dice is fair then p must be uniformly distributed over the set of possible outcomes.[20]

As Golan, Judge, and Miller (1996) note, the maximum or cross entropy (CE) framework offers a solution to a range of problems which given limited data, are undetermined using conventional procedures.[21] For example, in many economic applications data may be available at the national level whereas at certain levels of regional disaggregation, data is either incomplete or non-existent. In essence, the use of the entropy principle facilitates data recovery using only the aggregate data, or if using cross-entropy, we use aggregate data in conjunction with an earlier set of disaggregate regional data. As such, the maximum level of entropy (uncertainty) follows from maximising entropy without any constraints which then yields a uniform distribution. The inclusion of additional, relevant information will therefore reduce uncertainty, and will result in a departure from the uniform.

Although entropy-based problems are usually cast in terms of probability distribution, Theil (1969) points out that the entropy measure, in its most general form, may be considered as the degree to which an aggregate is subdivided into its constituent parts. Obviously this problem may be immediately transformed into an examination of the characteristics of the implied probability distribution, but this need not define the original problem. For example, we might be interested in aggregates such as total population or total income of a nation state and the allocation across sub-regions. In Section 10 we demonstrate the use of entropy procedures in the context of a particular type of missing data, where aggregate wages are observed at the regional and industrial level but disaggregate data - wages by industry and region - are not observed.

[20] For an excellent discussion of these methods see Golan, Judge, and Miller (1996).

[21] The implementation of ME and CE procedures for data imputation is relatively straightforward. The software package Shazam (see chapter 17) is particularly useful in this context.

8. Accuracy of Estimation with Missing Data

To date we have focused upon the use of various imputation techniques as a tool to facilitate the use of as many data points as possible within the context of parameter estimation. As noted by Efron (1994), the drawback of most imputation techniques is that the process of replacing missing with fitted values ignores any residual variation insofar as missing values are subsequently treated as if they were known. As a result the unconditional variance (or conditional variance in regression) will be less than that of the original unobserved series, with similar implications for estimates of standard errors of model parameters.

Below we examine a number of techniques designed to examine the variability of parameter estimates where a subset of the data points have been imputed.

i) Stochastic regression imputation

Stochastic regression imputation (as introduced in Section 5, example 4) imputes missing values based upon the (estimated) conditional expectation of y, $\hat{E}(y \mid x)$, plus a residual component to reflect uncertainty in the predicted value. Herzog and Rubin (1983) outline the two-stage procedure for both normal and binary outcomes.

ii) Imputation: Confidence Bounds around Parameter Estimates

An approach suggested by Simon and Simonoff (1986), which does not assume that the missing data is *mar* (or *oar*), utilises a graphical technique to provide upper and lower limits for the parameter estimates. The authors demonstrate that the range is a function of the nonrandomness of the missing value process. For example, in the context of a non-random self-selection model, the procedure advocated by Simon and Simonoff (1986) could be utilised to compute an interval around the possibly biased set of estimated parameters induced by self-selection. This is obviously different from computing the inverse of the Mills ratio and correcting the bias.

iii) Multiple imputation

As noted above, regression-based imputation procedures will underestimate the variance in any subsequent parameter estimates. Based upon the work of Rubin (1987) and Rubin (1978), multiple imputation proceeds by nesting the estimation procedure within an outer loop which iterates over *multiple* random imputations of the missing data, rather than a single imputed value as in, for example, examples 3 and 4 of Section 5. In this way, we may construct a distribution of likely imputed values and thereby properly integrate this element of uncertainty in model estimation.

Although similar to bootstrap-based methods for parameter accuracy, multiple imputation is implemented (and best understood) using a Bayesian updating scheme. For example, one of the main tools of Bayesian analysis is the conditional distribution of the population parameters given the data. However, this perspective becomes particularly useful where a portion of the complete data has been imputed, and therefore there exists incertainty beyond parameter uncertainty. In this instance the additional uncertainty introduced by imputation can be examined by averaging the complete-data posterior distribution (i.e. the conditional distribution given both observed and missing data) over the posterior distribution of the missing data. [22]

9. Implementation

Below we focus on a number of issues related to the implementation of imputation both in terms of modelling strategies and the availability of computer software.

9.1. Parameter Estimation with Imputed Values

The distinction between exploratory data analysis and statistical modelling has an analogue in the missing data problem, insofar as the imputing of missing data may be distinct from the subsequent use of a complete dataset to estimate model parameters. Obviously this will be true if those in charge of imputing data are separate from the modeler. In addition, if there is complete separation of these two tasks, then the modeler may not have full information regarding the imputation techniques that were employed.

If we assume that these tasks are undertaken simultaneously, the study by Wang and Jinn (1992) presents an overview of a number of alternative strategies. These are:

i) apply case deletion and use only observed values

ii) impute values for missing data and treat observed and imputed values equally

iii) utilise imputation as in ii) but implement improved variance estimation (see Section 5, example 4)

iv) base parameter estimation on multiple imputation (see Section 8)

[22] See Efron (1994), Rubin and Schenker (1986), and Wei and Tanner (1990) for further details.

9.2. Computer Software

In standard econometric software packages such as SAS (1985) there are, in general, a limited number of in-built features for either imputing missing values or model estimation which adjusts parameter estimates given the additional variability introduced by imputation. Below we consider a number of exceptions.

i) STAMP 5.0 ((**S**tructural **T**ime Series **A**nalyser Modeler and **P**redictor) Koopman, Harvey, Doornik, and Shephard (1995)). STAMP is designed for modelling time series data using unobserved components. Version 5.0 runs on personal computers operating under DOS or Windows. Since STAMP utilises the Kalman filter to construct the likelihood it offers extensively powerful *model-based* procedures for imputing missing data in time series.[23]

ii) STATA (1997) (Release 5)

   STATA is a statistical package for managing, analysing and graphing data. Despite STATA's impressive array of sophisticated modelling tools, it does not have many features specifically designed to deal with missing data. However, the package does include an in-built *impute* function which fills in missing values using a parametric regression function as outlined in examples 3 and 4 in Section 5.

iii) GAUSS Third-Party Application: MISS

   MISS, a program for missing data, includes procedures for the computation of covariance matrices and means, and for the imputation of data sets with incomplete observations. Observations, mean vectors, and variance-covariance matrices are estimated by maximum likelihood. Regression-based imputation with or without variance adjustment is also provided.

   The two main components are:

   - EM - computes maximum likelihood estimates (using the EM method) of the covariance matrix and mean vector when data are missing and imputes the data as requested.
   - IMPUTE - imputes the data using either a mean or regression method substitution with an "equalisation" by either the random sample or the random variable method. Then the covariance matrix and mean vector may be calculated from the imputed data set.

iv) SOLAS (1997)

[23] See Harvey, Koopman, and Penzer (1997) for a number of examples.

SOLAS is a Windows-based statistical software package that incorporates a range of techniques for treating missing values. The package can handle both longitudinal and single observation survey data. Options for both single and multiple imputation are included. Within the context of multiputation it should be emphasised that SOLAS samples from the observed sample point after stratifying the data using the propensity score. Therefore, if this approach is to be viable it requires that within a given variable there are enough "donor" values available from which to sample. Subsequently there are certain missing data mechanisms that are not suitable for this approach. An obvious case is that given by example 4.3 (Section 4) where for any given time period only aggregate variables are observed and estimates of the disaggregate constituents are required.

v) NP-REG (**N**on-**P**arametric **REG**ression: See Duncan and Jones (1992)).

NP-REG is an interactive software package for Kernel density estimation and non-parametric regression. Written in GAUSS, it is menu-driven and incorporates a considerable number of non-parametric techniques. Although there is no specific missing data options, imputation using a non-parametric conditional mean function can be implemented.

vi) GAMS (**G**eneral **A**lgebraic **M**odelling **S**ystem: See Brooke, Kendrick, and Meeraus (1992))

GAMS provides a high-level language for the representation and solution of large linear and non-linear mathematical programming problems. GAMS is particulary useful for the solution of entropy and cross-entropy problems, which may be used in specific missing-data problems (see Secton 4, example 4.3).

10. An Example

In this section we will examine the problem of missing data for regional analysis. In particular, we focus upon situtaions where although data is available at, for example, the national level, data at the regional level is either incomplete or non-existent. Deutsch and Rodler (1990) highlight two key issues when the problem of missing data is one of aggregation:

i) there is an obvious information loss when aggregate observations are used in place of disaggregate ones;

ii) in terms of public policy, this loss of information may be critical when attempting to formulate policy rules when only aggregate information is available. It is also likely that the costs of using aggregate information will vary across different economic variables.

In a number of recent studies analysts have questioned the use of large macroeconomic models which focus on the modelling of income and output aggregates measured at the level of nation states. For example, the extent to which individual regions and industrial sectors exhibit heterogeneity in terms of the mechanism that determine growth, will affect the usefulness of aggregate studies. Quah (1994) notes that since the EC Cohesion Fund distributes resources at the level of NUTS-3 subdivisions (with over 800 regional units), the use of aggregate models is obviously misplaced. If one accepts that detailed disaggregate information is required, then in the face of limited information, aggregate indicators must be decomposed to reveal their constituent parts. As such, this problem is in general, underdetermined.

10.1. Missing Wages at the Regional Level

We will focus on the following problem. In an economy comprised of R regions we observe the average wage for sector k, denoted by w_k. However, w_{ki}, the wage in sector k, region i is unobserved. We let W^{RG} denote a $(R \times 1)$ vector denoting the total wage bill for each region W^{IND} is a $(n \times 1)$ vector denoting the total wage bill for each industry .

The missing data problem is the problem of determining additional information from a single aggregate mean. Obviously in the case of an economy based upon R identical regions, then it would be reasonable to set $w_{ki} = w_k \; \forall i = 1, ..., R$. Likewise, the more heterogenous the economy in terms of the processes that determine wages, the more unreasonable this assumption.

If the aggregate (national) mean wage is observed and data is missing for a *single* region, then a simple accounting constraint allows the missing data to be imputed without the use of a statistical model. However, with $R - 1$ degrees of freedom and two or more regional observations missing, the problem is underdetermined. This type of problem is common in econometrics and applied mathematics and may be solved by the use of prior or non-sample information. This may take a number of forms, including for example, knowledge of w_{ki} for the previous time period and an assumed growth rate. If we only observe aggregate data, then one possibility is to postulate an econometric model of wage determination at the national level,[24] and use this to impute the disaggregate regional wages.

[24] This obviously assumes we have time-series data for national wages.

Below we consider a number of alternative strategies based upon the various imputation techniques discussed and the extent of missing data.

10.1.1. Regression-Based: Only w_k observed

First we propose a model of wage determination at the national level. This might take the simple linear form

$$w_{tk} = \alpha + X_{tk}\beta_k + \varepsilon_{tk}, \tag{10.1}$$

where X is a $(n \times v)$ matrix of regressors, β is a $(v \times 1)$ parameter vector, $\varepsilon \sim iid(0, \sigma^2)$[25], and t indexes time. Based upon this aggregate model of wage determination, the imputed wages for region j, industry k at period t' would be

$$\hat{w}_{t'jk} = \hat{\alpha} + X_{t'jk}\hat{\beta}, \tag{10.2}$$

where $X_{t'jk}$ is the regressor matrix for region j, sector k. Obviously, the major problem with this approach is the assumption of an economy-wide constant set of parameters $\hat{\beta}$.[26] Note that since w_k is observed the imputation procedure must satisfy the constraint $w_k = \frac{1}{R}\sum_{i=1}^{R} \hat{w}_{ki}$. In addition (10.2) assumes that the missing-data mechanism is *mar*.

10.1.2. Regression-Based: w_k observed and a sub-sample of w_{ki}

Obviously if a subset of the w_{ki} regional data is missing at random then we may apply the same procedure as in (10.2). However, if there is some selection process which is endogenous, then the imputation based on (10.1) would not be appropriate since the information available in the selection process is not utilised. As an example, let us postulate the following non-random observational rule for regional wage data.

$$w_{ki} = \mathbf{1}(w_{ki} > \delta)w_{ki}, \tag{10.3}$$

such that regional data is only observed if wages exceed δ. For notational purposes let w_{ki}^B (w_{ki}^A) denote, respectively, the sub-sample of unobserved (observed) wages. The total information we observe comprises the national figures, estimates of α and β from (10.2) and w_{ki}^A.

[25] This iid assumption may be relaxed.

[26] Note that the problems of using imputed values in a subsequent modelling exercise, such as providing an accurate estimate of parameter uncertainty when elements of the regressand have been imputed, are discussed in section 8.

For the unobserved wage data, the population regression function conditional upon the observational rule may be written

$$w_{ki} \mid w_{ki} < \delta = \alpha^B + X_{ki}^B \beta^B + \varepsilon^B. \tag{10.4}$$

Since w_{ki} is unobserved we cannot recover parameter estimates α^B. We could utilise (10.1) and set $w_{ki}^B = \hat{\alpha} + X_{ki}^B \hat{\beta}$ but this is incorrect since this implicitly assumes that the data is both *mar* and *oar*, whereas in this instance $E(\varepsilon^B) < \delta - \alpha^B + x_{ki}^B \beta^B$. Therefore, an appropriate modification of (10.1) would be to use

$$\hat{w}_{ki}^B = \hat{\alpha} + X_{ki}^B \hat{\beta} + \frac{\phi(\gamma)}{\Phi(\gamma)}, \tag{10.5}$$

where $\gamma = \delta - \hat{\alpha} + X_{ki}^B \hat{\beta}$ and ϕ (Φ) are respectively, the standard normal probability density and cumulative distribution function. In (10.5) we note that we substitute $\hat{\alpha}$ and $\hat{\beta}$ for the unobserved $\hat{\alpha}^B$ and $\hat{\beta}.^B$

10.1.3. An Entropy-based Solution

We introduce a $(R \times n)$ coefficient matrix A, with typical element

$$a_{ij} = w_{ij}/w_i, \tag{10.6}$$

denoting the ratio of wages for industry j, region i to the total (aggregate) wage bill for industry j, $w_i = \sum_j w_{ij}$. We may write the accounting identity

$$W^{RG} = AW^{IND}, \tag{10.7}$$

with the first row of W^{RG} given by

$$W_1^{RG} = a_{11} W_1^{IND} + a_{12} W_2^{IND} + \cdots + a_{1n} W_n^{IND}, \tag{10.8}$$

where each a_{ij} coefficient allocates the total wage bill for the jth industry to region 1. Given (10.7), knowledge of the individual a_{ij} coefficients allows us to estimate the unobserved disaggregate wage data using $\hat{w}_{ij} = w_i \times a_{ij}$.

The objective is to find an estimate of the unknown A matrix given aggregate vectors W^{RG} and W^{IND}, which respects (10.7) and the following constraints and non-negativity restrictions:

$$\sum_j^n a_{ij} W_j^{IND} = W_i^{RG} \; \forall i = 1, \ldots, R; \tag{10.9}$$

$$\sum_{i}^{R} a_{ij} = 1 \; \forall j = 1, \ldots, n; \tag{10.10}$$

$$a_{ij} \geq 0 \; \forall \; i = 1, \ldots, R; j = 1, \ldots, n. \tag{10.11}$$

This type of problem has a considerable lineage in multisectoral regional equilibrium models (see Harrigan, McGilvray, and McNicoll (1980) and **?**). For example, in the context of updating input-output matrices the generic form of this particular missing data problem follows from the need to determine n_{t+1}^2 pieces of information from $2 \times n_{t+1}$ fully observed data points (for period $t+1$) and n_t^2 pieces of information in period t. In this context, many data recovery procedures such as RAS (see Schneider and Zenios (1990)) are based upon methods that rely on accounting constraints. Recent work by Golan, Judge, and Robinson (1994) have proposed an alternative method based upon the Maximum Entropy (ME) Principle, which in the case of updating a $(n \times n)$ interindustry flow matrix A, applies an optimisation criterion such that the estimate of A, say A^*, is that matrix which can be realised in the greatest number of ways, given existing knowledge.[27]

The maximum entropy solution to the above problem is to find the matrix A, given the observed aggregate data and the constraints (10.9)-(10.11), that may be realized in the largest number of ways. If certain w_{ij} are known this information may be incorporated via additional constraints, thereby reducing the amount of entropy in the system. Formally, we may state the problem as

$$Max \; H = -\sum_{j} \sum_{i} a_{ij} \log(a_{ij}) \tag{10.12}$$

subject to (10.9)-(10.11).

If w_{ij} information is available from a previous period, say w_{ij}^{t-1}, then this may be incorporated into the problem[28]. In this instance we note that the appropriate reference distribution is not one which reflects a complete lack of prior information as to the distribution of the total wage bill over industrial sectors and regions, but rather we use the distribution of wages in a previous period as the reference distribution. The resulting cross entropy problem[29] may be written as

$$Min \; CE = \sum_{i} \sum_{j} a_{ij} \log(a_{ij/} a_{ij}^{t-1}) \tag{10.13}$$

[27] The earliest reference is Shannon (1948). More recent discussion and applications are to be found in Wilson (1970) and Zellner (1990).

[28] A similar problem is described in Theil (1969), who uses entropy methods in the analysis of financial statements.

[29] This is also referred to as the Kullback-Leibler information criterion.

subject to the same data and adding up constraints.

In Golan, Judge, and Miller (1996) the use of ME and CE is demonstrated using artificial data for a small inter-industry flow matrix. The authors show that the use of CE to impute missing disaggregate data represents an improvement over the traditional RAS approach. We could just as easily use the same data to represent the missing wages problem. For example, in table 1 the rows represent regions and columns industries, such that $w_{11} = 45$ signifies that 45 units of wages are earned in industry 1, which is located in region 1; the total wage bill for region 1 is 140. Similarly, since $w_{23} = 0$ the data indicates that are no firms from industry 3 in region 2. If we normalise the wage data by total industry wages, the we can generate the coefficient matrix A, with elements a_{ij}.

Table 1: Wage Data and Coefficient Matrix

	wage data				w_i	A			
	45	0	15	80	140	0.726	0.000	0.165	0.301
	10	15	0	120	145	0.161	0.268	0.000	0.451
	7	38	65	0	110	0.113	0.678	0.714	0.000
	0	3	11	66	80	0.000	0.054	0.121	0.248
$w_{\cdot j}$	62	56	91	266					

References

Brooke, A., D. Kendrick, and A. Meeraus (1992): *GAMS: A User's Guide, Release 2.25.* Scientific Press, South San Francisco, CA, USA.

Buck, S. (1960): "A Method of Estimation of Missing Values in Multivariate Data Suitable for Use with An Electronic Computer," *Journal of the Royal Statistical Society, Ser.B, 22,*, pp. 302–306.

Chow, G., and A. Lin (1971): "Best linear unbiased interpolation, distribution and extrapolation of time series by related series," *Review of Economics and Statistics*, 53, 372–75.

David, M., R. Little, M. Samuhel, and R. Triest (1986): "Alternative Methods for CPS Income Imputation," *Journal of American Statistical Association*, 81(393).

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977): "Maximum Likelihood from Incomplete Data via the EM Algorithm (with discussion)," *Journal of the Royal Statistical Society*, pp. 1–38.

Deutsch, E., and K. Rodler (1990): "Aggregation Problems in a Model of Wage Formation and Employment Demand," in *Disaggregation in Econometric Modelling*, ed. by T. Barker, and M. H. Pesaran, chap. 5. Routledge.

Diebold, F. (1992): "Advanced Econometrics," Lecture Notes, University of Pennsylvania.

Duncan, A., and A. Jones (1992): "NP-REG: An Interactive Package for Kernel Density Estimation and Nonparametric Regression," Discussion Paper W92/07, Institute for Fiscal Studies.

Efron, B. (1994): "Missing Data, Imputation, and the Bootstrap," *Journal of the American Statistical Association*, 89(426), 463–475.

Golan, A., G. Judge, and D. Miller (1996): *Maximum Entropy Econometrics: Robust Estimation with Limited Data.* John Wiley and Sons.

Golan, A., G. Judge, and S. Robinson (1994): "Recovering Information from Incomplete or Partial Multisectoral Economic Data," *The Review of Economics and Statistics*, LXXVI(3), 541–550.

Gourieroux, C., and A. Montfort (1981): "On the Problem of Missing Data in Linear Models," *Review of Economic Studies*, XLVIII, 579–586.

GREEN, W. H. (1997): *Econometric Analysis*. Prentice Hall, third edn.

GREENLEES, J., W. REECE, AND K. ZIESCHANG (1982): "Imputation of Missing Values When the Probability of Response Depends on the Variable Being Imputed," *Journal of the American Statistical Association*, 77(378).

HAMILTON, J. (1994): *Time Series Analysis*. Princeton University Press.

HARDLE, W. (1990): *Applied nonparametric regression*. Cambridge University Press.

HARRIGAN, F., W. MCGILVRAY, AND I. MCNICOLL (1980): "Simulating the Structure of a Regional Economy," *Environment and Planning*, A(12), 927–936.

HARVEY, A. (1989): *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge University Press.

HARVEY, A., S. KOOPMAN, AND J. PENZER (1997): "Messy Time Series: A Unified Approach," *Advances in Econometrics*, 13.

HECKMAN, J. (1979): "Sample Selection Bias as a Specification Error," *Econometrica*, 47, 153–161.

HERZOG, T., AND D. RUBIN (1983): "Using multiple imputations to handle nonresponse in sample surveys," in *Incomplete Data in Sample Surveys (Vol. 2 - Theory and Bibliographies)*, ed. by W. Madow, I. Olkin, and D. Rubin, pp. 209–245. New York: Academic Press.

HOROWITZ, J., AND C. MANSKI (1998): "Censoring of outcomes and regressors due to survey nonresponse: Identification and estimation using weights and imputations," *Journal of Econometrics*, 84, 37–58.

KING, G., J. HONAKER, A. JOSEPH, AND K. SCHEVE (1998): "Listwise Deletion is Evil: What to Do About Missing Data in Political Science," Dept. of Government, Harvard University.

KOOPMAN, S., A. HARVEY, J. DOORNIK, AND N. SHEPHARD (1995): *Stamp 5.0: Structural Time Series Analyser, Modeller and Predictor*. Chapman and Hall.

LILLARD, L., J. P. SMITH, AND F. WELCH (1982): "What Do We Really Know About Wages: The Importance of Non-Reporting and Census Imputation," Discussion paper, Rand Corporation, 1700 Main Street, Santa Monica, CA 90406.

Little, R. (1992): "Regression with Missing X's: A Review," *Journal of the American Statistical Association*, 87(420), 1227–1237.

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

Quah, D. (1994): "One Business Cycle and One Trend from (Many,) Many Disaggregates," *European Economic Review*, 38, 605–613.

Rubin, D. (1976): "Inference and missing data," *Biometrika*, 63(3), 581–92.

——— (1978): "Multiple Imputations in Sample Surveys - A Phenomenological Bayesian Approach to Nonresponse," *Proceedings of the Survey Research Methods Section, American Statistical Association*, pp. 20–34.

——— (1987): *Multiple Imputation for Nonresponse in Surveys.* John Wiley, New York.

Rubin, D., and N. Schenker (1986): "Multiple Imputation for Interval Estimation From Simple Random Samples With Ignorable Nonresponse," *Journal of the American Statistical Association*, 81(394), 366–.

Ruud, P. (1991): "Extensions of Estimation Methods using the EM Algorithm," *Journal of Econometrics*, 49(3), 305–342.

Salazar, E., R. Smith, M. Weale, and S. Wright (1997): "A Monthly Indicator of GDP," *National Institute Economic Review*, 161, 84–90.

SAS (1985): *SAS User's Guide: Basics, Version 5 Edition.*

Schneider, M., and S. Zenios (1990): "A Comparative Study of Algorithms for Matrix Balancing," *Operations Research*, (38), 439–455.

Shannon, C. (1948): "A Mathematical Theory of Communication," *Bell System Technical Journal*, (27), 379–423.

Simon, G., and J. Simonoff (1986): "Diagnostic Plots for Missing Data in Least Squares Regression," *Journal of the American Statistical Association*, 81(394), 501–509.

SOLAS (1997): *SOLAS for Missing Data Analysis 1.0.*Statistical Solutions Ltd., 8 South Bank, Crosse's Green, Cork, Ireland.

STATA (1997): *Stata Statistical Software: Release 5.0*Stata Press, College Station, TX: Stata Corporation.

Theil, H. (1969): "On the Use of Information Theory Concepts in the Analysis of Financial Statements," *Management Science*, 15, 459–480.

Wang, J., and J. Jinn (1992): "Secondary Data Analysis When There are Missing Observations," *Journal of the American Statistical Association*, 87(420), 952–961.

Weeks, M. (1998): "Methods of Imputation for Missing Data," Report for Eurostat relating to the project Model Based Regional Indicators.

Wei, G. C., and M. Tanner (1990): "A Monte Carlo Implementation of the EM Algorithm and the Poor Man's Data Augmentation Algorithms," *Journal of the American Statistical Association*, 85(411), 699–704.

Wilson, A. (1970): *Entropy in Urban and Regional Modeling.* London: Pion Ltd.

Zellner, A. (1990): "Bayesian Methods and Entropy in Economics and Econometrics," *W.T. Grandy and L.H. Shick (eds.) Maximum Entropy and Bayesian Methods*, pp. 17–31, Dordrecht: Kluwer Academic Publishers.