
Learning from Incomplete Data

Sameer Agarwal

sagarwal@cs.ucsd.edu

Department of Computer Science and Engineering
University of California, San Diego

Abstract

Survey non-response is an important problem in statistics, economics and social sciences. The paper reviews the missing data framework of Little & Rubin [Little and Rubin, 1986]. It presents a survey of techniques to deal with non-response in surveys using a likelihood based approach. The focuses on the case where the probability of a data missing depends on its value. The paper uses the two-step model introduced by Heckman to illustrate and analyze three popular techniques; maximum likelihood, expectation maximization and Heckman correction. We formulate solutions to heckman's model and discuss their advantages and disadvantages. The paper also compares and contrasts Heckman's procedure with the EM algorithm, pointing out why Heckman correction is not an EM algorithm.

1 Introduction

Surveys are a common method of data collection in economics and social sciences. Often surveys suffer from the problem of nonresponse. The reasons for nonresponse can be varied, ranging from the respondent not being present at the time of survey to a certain class of respondents not being part of the survey at all. Some of these factors affect the quality of the data. For example ignoring a certain portion of the populace might result in the data being biased. Care must be taken when dealing with such data as naive application of standard statistical methods of inference will produce biased or in some cases wrong conclusions. An often quoted example is the estimation of married women's labor supply from survey data collected from working women [Heckman, 1976]. Since a woman's participation in the labor force is based on what the job earns her y^* and her own valuation of her time y , we will only observe samples for which $y^* \geq y$. Hence the use of the sample average as an estimate of average reserve wage y for working women, below which she will stay at home, will be biased low.

Another example is the KDD-98 dataset [Georges and Milley,], based on responses from donors. A direct mailing strategy is to be designed so as to maximize the returns. This requires us to estimate the amount each potential patron will donate. However donation amount data is only available for those individual who donated in the last campaign, who form a very small part (5%) of the over all data set, and it is known that there is a negative correlation between the probability of

a patron donating and the amount he actually donates. Hence individuals who donate more frequently are inclined to donate smaller amounts as compared to donors who respond less frequently to donation mailers. Now if we were to estimate donation amounts only using the observed samples (which will contain a larger proportion of the more frequent low amount donors) our estimates will be biased low for individuals who have a low donation probability.

Survey nonresponse is an actively researched field of study [Little, 1985, Hall, , Heckman, 1976, Little, 1982, Little and Rubin, 1986, Maddala, 1983]. In [Little, 1985] the authors point out the two major approaches to the study of survey nonresponse, namely the *randomization* and the *model-based* approaches.

Randomization inference considers the population of all possible responses and a sampling distribution on top of it. The samples are selected using *probability sampling* which requires that the sampling probabilities be decided without knowledge of the data values and that each member of the over all population have a non-zero probability of being selected. This is a model free approach and various modes of sampling are possible ranging from the simple randomized sampling which utilizes no information about the data entries to sophisticated stratified sampling methods. However as we shall see later nonresponse cannot always be modeled by ignoring the data values and this approach is useful only when within recognizable subclasses of the overall population nonresponse can be approximated by probability sampling.

Model-based inference is based on considering each observation in the sample as a realization of some random variables distributed according to a an underlying model. This approach though more complicated gives better results and is more robust. In this paper we survey techniques for dealing with survey nonresponse using the latter approach.

The rest of the paper is organized as follows. Section 2 presents the missing data framework in which we will base all our analysis, Section 3 introduces the problem that we shall use for describing the various methods of dealing with missing data in surveys. Section 4 presents the maximum likelihood based approach to estimating the model parameters. Section 6 introduces the EM algorithm for dealing with missing data and how it can be used to estimate maximum likelihood estimates of our model. Section 7 presents Heckman's two step procedure for approximately correcting bias and discusses its strength and weaknesses. We also answer the question that started off this study, whether Heckman correction is one step EM algorithm, and if so, can it be generalized to an iterative algorithm which yields improved estimates? The final section concludes with a discussion and ideas for future work.

2 Models of missing data

We shall deal with the problem of survey-nonresponse in the missing data framework first introduced in [Little and Rubin, 1986] and reviewed in [Gaharamani and Jordan, 1993]. We begin by assuming that the data $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^N$ can be divided into two components, \mathbf{X}_{obs} and \mathbf{X}_{mis} being the observed and the missing components respectively. We further allow each data vector \mathbf{x}_i to have one or more components missing. We define a missing data matrix,

$$R_{ij} = \begin{cases} 1, & x_{ij} \text{ observed,} \\ 0, & x_{ij} \text{ missing.} \end{cases} \quad (1)$$

The R matrix is just an indicator matrix which makes explicit what data components are available and what are missing, it does not by itself contain any new

information that is not already contained in the data \mathbf{X} .

Considering the data generation and missing mechanisms to be stochastic we denote the joint probability density of the complete data by

$$f(\mathbf{X}, R|\theta, \psi) = g(\mathbf{X}|\theta)h(R|\mathbf{X}, \psi) \quad (2)$$

Here θ and ψ denote the parameters for the data generation and missing data mechanisms respectively. The specific form of the density h results in three different kind of missing data mechanisms.

1. **Missing Completely at Random(MCAR)** When the probability of missing data is independent of the values of the data vector :

$$h(R|\mathbf{X}, \psi) = h(R|\psi) \quad (3)$$

2. **Missing at Random(MAR)** This is the case when the probability of x_{ij} is independent of the missing data values, but may depend on the observed values :

$$h(R|\mathbf{X}_{obs}, \mathbf{X}_{mis}, \psi) = h(R|\mathbf{X}_{obs}, \psi) \quad (4)$$

3. **Not Missing at Random (NMAR)** This is the case when the probability of x_{ij} missing depends on the value of x_{ij} or some other missing data item, and no term in the density function expression for the missing data mechanism can be ignored. An example of this is a censored data model, where variable values above a certain threshold are not recorded, instead only the fact that they are greater than a certain threshold is known.

Let us now consider the following: Given that

$$p(\mathbf{X}_{obs}, R|\theta, \psi) = \int g(\mathbf{X}_{obs}, \mathbf{X}_{mis}|\theta)h(R|\mathbf{X}_{obs}, \mathbf{X}_{mis}, \psi)d\mathbf{X}_{mis} \quad (5)$$

for the case when

$$h(R|\mathbf{X}_{obs}, \mathbf{X}_{mis}, \psi) = h(R|\mathbf{X}_{obs}, \psi) \quad (6)$$

equation 5 becomes

$$p(\mathbf{X}_{obs}, R|\theta, \psi) = h(R|\mathbf{X}_{obs}, \psi) \int g(\mathbf{X}_{obs}, \mathbf{X}_{mis}|\theta)d\mathbf{X}_{mis} \quad (7)$$

$$= h(R|\mathbf{X}_{obs}, \psi)g(\mathbf{X}_{obs}|\theta) \quad (8)$$

This implies that if the data is MAR the likelihood can be factored into two pieces with only one of the pieces depending on θ . Hence if we were only estimating the parameters for the data generation mechanism using maximum likelihood, we could just maximize

$$L(\theta|\mathbf{X}_{obs}) \propto g(\mathbf{X}_{obs}|\theta) \quad (9)$$

as a function of θ and it will be equivalent to maximizing $L(\theta, \psi|\mathbf{X}_{obs}, R)$. This means that the missing data mechanism can be ignored, and we can estimate the parameters of the data generation mechanism by using the observed data only. Since MCAR data is just a special case of MAR data, the above analysis is valid for both the cases. However survey data is NMAR more often than not and the rest of the paper is devoted to methods of dealing with the same.

3 The Problem

We now present the problem which we will use to introduce and analyze the various techniques available for this task.

Consider the following equations:

$$y_{2i} = \mathbf{x}_{2i}\beta_2 + \nu_{2i} \quad (10)$$

$$y_{1i} = \mathbf{x}_{1i}\beta_1 + \nu_{1i} \quad \text{if } y_{2i} > 0 \quad (11)$$

$$y_{1i} = \text{not observed} \quad \text{if } y_{2i} \leq 0 \quad (12)$$

where Y_1 is the random variable of interest. (Upper case symbols refer to the random variables, lower case refers to a particular observation of that variable, hence y_{2i} is the i^{th} observation of the random variable Y_2 .) But it is not observed under all conditions, these conditions are specified by the dependent variable Y_2 . Y_1 is observed only when the corresponding value of Y_2 is greater than 0. X_1 and X_2 both refer to vectors of independent variables or covariates of Y_1 and Y_2 respectively. We will use the symbols X and Y when referring to X_1 , X_2 and Y_1 , Y_2 simultaneously. β_1 and β_2 are vectors of regression coefficients and ν_{1i} and ν_{2i} are the corresponding residuals. Without loss of generality we can assume that out of a total of N observations, the first N' are complete i.e. Y_1 values are known for them. Equation 10 is referred to as the response equation since it determines where a certain response is present in the survey or not. This model has a simple interpretation, in the case of the donor database the response equation states that the probability of a donor responding to a solicitation goes up with $-X_2\beta$. The particular shape of the distribution function for ν_2 decides how this probability varies with the magnitude of $-X_2\beta_2$. The magnitude of $-X_2\beta_2$ can perhaps represent the value that the donor associates to the act of donating.

Let us now look at the expectation for Y_1 given the independent variables X_1 and that $Y_2 > 0$

$$E[Y_1|X_1, Y_2 > 0] = X_1\beta_1 + E[\nu_1|\nu_2 > -X_2\beta_1] \quad (13)$$

As is evident from the above expression, simple regression estimates will be unbiased if ν_1 and ν_2 are independent. This corresponds to the Missing at Random (MAR) model introduced earlier. However we are more interested in the NMAR case where the two random variables are not independent. Further analysis requires us to assume a distribution on ν_1 and ν_2 .

Let us formalize the model within the framework presented above. The following model was proposed by Heckman [Heckman, 1976] The dependent variable Y_1 is incompletely observed, and Y_2 is never observed. The independent variable vector X is completely specified. The conditional density $f(Y|X, \theta)$ is given by

$$\begin{pmatrix} y_{1i} \\ y_{2i} \end{pmatrix} = N_2 \left[\begin{pmatrix} \mathbf{x}_i\beta_1 \\ \mathbf{x}_i\beta_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1 \\ \rho\sigma_1 & 1 \end{pmatrix} \right] \quad (14)$$

where $\mathbf{x}_i = (x_{i0}, x_{1i}, \dots, x_{ip})$. $x_{i0} \equiv 1$ is the constant term. $N_2(a, b)$ denotes the bivariate normal distribution with mean a and covariance matrix b . This is equivalent to saying that ν_1 and ν_2 are distributed as

$$\begin{pmatrix} \nu_1 \\ \nu_2 \end{pmatrix} = N_2 \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1 \\ \rho\sigma_1 & 1 \end{pmatrix} \right] \quad (15)$$

The covariance matrix as can be seen is constrained by setting the variance of the second variable to 1. This is an artifact of Heckman's estimation procedure and the reason for doing so will be explained later. It is sufficient to observe that setting the variance to 1 does not limit the generality of the model.

Further $f(R|X, Y, \psi)$ is given by

$$R_{1i} = \begin{cases} 1, & \text{if } y_{2i} > 0 \\ 0, & \text{if } y_{2i} \leq 0 \end{cases} \quad (16)$$

$$R_{2i} \equiv 0. \quad (17)$$

In most cases we are only interested in estimating θ , i.e. the parameters of the data generation mechanism, in some applications like cost-sensitive learning we are interested in estimating the parameters of the missing data mechanism also. The methods in the following sections perform a full model estimation and return both sets of parameters θ as well as ψ .

4 Maximum Likelihood Methods

The first method that one can use for estimating the parameters of the data generation mechanism is to construct the full data likelihood function and maximize it. Hall[Hall,] presents the following construction for the likelihood function.

The calculation can be split into two parts. In the first part we calculate the likelihood for the case where we know that y_{1i} is observed and $y_{2i} > 0$:

$$\begin{aligned} Pr(y_{1i}, y_{2i} > 0|X) &= f(y_{1i})Pr(y_{2i} > 0|y_{1i}, X) \\ &= f(\nu_{1i})Pr(y_{2i} > 0|y_{1i}, X) \\ &= \frac{1}{\sigma_1} \phi\left(\frac{y_{1i} - X_i\beta_1}{\sigma_1}\right) \int_{-X_i\beta_2}^{\infty} f(\nu_{2i}|\nu_{1i})d\nu_{2i} \\ &= \frac{1}{\sigma_1} \phi\left(\frac{y_{1i} - X_i\beta_1}{\sigma_1}\right) \int_{-X_i\beta_2}^{\infty} \phi\left(\frac{\nu_{2i} - \frac{\rho}{\sigma_1}(y_{1i} - X_i\beta_1)}{\sqrt{1 - \rho^2}}\right) d\nu_{2i} \\ &= \frac{1}{\sigma_1} \phi\left(\frac{y_{1i} - X_i\beta_1}{\sigma_1}\right) \left[1 - \Phi\left(\frac{-X_i\beta_2 - \frac{\rho}{\sigma_1}(y_{1i} - X_i\beta_1)}{\sqrt{1 - \rho^2}}\right)\right] \\ &= \frac{1}{\sigma_1} \phi\left(\frac{y_{1i} - X_i\beta_1}{\sigma_1}\right) \Phi\left(\frac{X_i\beta_2 + \frac{\rho}{\sigma_1}(y_{1i} - X_i\beta_1)}{\sqrt{1 - \rho^2}}\right) \end{aligned} \quad (18)$$

In the second part we estimate the likelihood for when no information is available on y_{1i} except that $y_{2i} \leq 0$

$$Pr(y_{2i} \leq 0) = Pr(\nu_{2i} \leq -X_i\beta_2) = \Phi(-X_i\beta_2) = 1 - \Phi(X_i\beta_2) \quad (19)$$

Hence the total log likelihood is :

$$\begin{aligned} \log(\beta_1, \beta_2, \rho, \sigma|X_{obs}) &= \\ &= \sum_{i=1}^{N'} \left[\log \sigma_1 + -\log \sigma_1 + \log \phi\left(\frac{y_{1i} - X_i\beta_1}{\sigma_1}\right) + \log \Phi\left(\frac{X_i\beta_2 + \frac{\rho}{\sigma_1}(y_{1i} - X_i\beta_1)}{\sqrt{1 - \rho^2}}\right) \right] \\ &+ \sum_{i=N'+1}^N \log(1 - \Phi(X_i\beta_2)) \end{aligned} \quad (20)$$

We can now use any of a number of non-linear optimization techniques [Berndt et al., 1974] to minimize the above expression to yield values for β_1, β_2, σ and ρ .

4.1 Discussion

As Hall points out, the estimates so obtained will be consistent and asymptotically efficient [Hall,]. Consistency means that as the number of data points increase the difference between the estimates and the actual parameter values goes to zero with probability one. Asymptotic efficiency means that in the limit of a large sample, the estimator so constructed method will have the minimum possible variance amongst all the unbiased estimators possible. The problem however with MLE methods is that they are extremely sensitive to the correct specification of the underlying distribution and assumptions of homoskedasticity. The latter refers to the assumption that variance of the observed terms if independent from one observation to the other and can be represented as identically distributed independent samples from a fixed distribution.

Classically the maximum-likelihood estimators(MLE) have not been used except for very simple cases due to the amount of computation effort involved. Today however we do not face such problems and the use of MLE techniques is indeed feasible. There is however the added problem of finding the global optima of the log-likelihood expression. As the number of parameters to be estimated increases typically the number of local optima goes up combinatorially but it is only the global optima which corresponds to the ML estimates for the parameters. Even with the availability of high speed computers the calculations involved can be quite cumbersome involving the calculation of complex derivative matrices.

We can also formulate the log-likelihood estimator by considering the missing data as parameters too however this can introduce hundreds or thousands more parameters into the expression, making the task of finding the global optima harder still.

5 Expectation Maximization

The EM algorithm has been proposed as a general framework for dealing with missing data in maximum likelihood estimation [Dempster et al., 1977]. Let us consider the general procedure for estimating dependencies using the EM algorithm. Assuming the same general framework as described in section 2, the algorithm steps are :

1. Find or assume initial estimates for the parameters $\theta^{(0)}$ and $\psi^{(0)}$.
2. At iteration t , calculate :

$$Q(\theta, \psi | \theta^{(t)}, \psi^{(t)}) = \int L(\theta, \psi | X_{obs}, X_{mis}, R) f(X_{mis} | X_{obs}, R, \theta^{(t)}, \psi^{(t)}) dX_{mis} \quad (21)$$

Here $L(\theta, \psi | X_{obs}, X_{mis}, R)$ refers to the complete data likelihood and $f(X_{mis} | X_{obs}, R, \theta^{(t)}, \psi^{(t)})$ is the conditional density of the missing data given the observed values, θ and ψ . This is the E step.

3. Find values of θ and ψ which maximizes Q ,

$$Q(\theta^{(t+1)}, \psi^{(t+1)} | \theta^{(t)}, \psi^{(t)}) \geq Q(\theta, \psi | \theta^{(t)}, \psi^{(t)}) \quad \forall \theta, \psi \quad (22)$$

call these values $\theta^{(t+1)}, \psi^{(t+1)}$ and replace $\theta^{(t)}, \psi^{(t)}$ in the next iteration. This is the M step.

4. Iterate steps 2 and 3 till convergence.

Dempster et al. [Dempster et al., 1977] proved the following theorems:

Theorem 1 *Let $L(\theta, \psi|X_{obs}, R)$ denote the likelihood function of the observed data, then every EM algorithm increases $L(\theta, \psi|X_{obs}, R)$ increases at each iteration, that is ,*

$$L(\theta^{(t+1)}, \psi^{(t+1)}|X_{obs}, R) = L(\theta^{(t)}, \psi^{(t)}|X_{obs}, R) \quad (23)$$

with equality if and only if

$$Q(\theta^{(t+1)}, \psi^{(t+1)}|\theta^{(t)}, \psi^{(t)}) = Q(\theta^{(t)}, \psi^{(t)}|\theta^{(t)}, \psi^{(t)}) \quad (24)$$

which has the following corollaries:

Corollary 1 *Let $M(\theta^t, \psi^t) = (\theta^{(t+1)}, \psi^{(t+1)})$ and suppose for some (θ^*, ψ^*) ,*

$$L(\theta^*, \psi^*|X_{obs}, R) \geq L(\theta, \psi|X_{obs}, R) \quad \forall \theta, \psi \quad (25)$$

Then for for every EM algorithm,

$$L(M(\theta^*, \psi^*)|X_{obs}, R) = L((\theta^*, \psi^*|X_{obs}, R)Q(M(\theta^*, \psi^*)|\theta^*, \psi^*) = Q(\theta^*, \psi^*|\theta^*, \psi^*)f(X_{mis}|X_{obs}, R, M(\theta^*, \psi^*))$$

almost everywhere.

and

Corollary 2 *Suppose for some (θ^*, ψ^*) ,*

$$L(\theta^*, \psi^*|X_{obs}, R) > L(\theta, \psi|X_{obs}, R) \quad \forall \theta, \psi \quad (26)$$

the for every EM algorithm

$$M(\theta^*, \psi^*) = (\theta^*, \psi^*) \quad (27)$$

The first theorem states that likelihood is nondecreasing with each iteration of the EM algorithm. The two corollaries imply that the ML estimate for (θ, ψ) is a fixed point of the EM algorithm.

6 EM for Heckman's Model

Let us now see a concrete realization of the EM algorithm for Heckman's model [Little and Rubin, 1986].

We begin by specifying the E step. Assuming we know the current estimates for the parameters. We start by calculating estimates for the missing data statistics.

$$E[y_{2i}|y_{2i} \leq 0] = \mu_{2i} - \lambda(-\mu_{2i}) \quad (28)$$

$$E[y_{2i}|y_{2i} > 0] = \mu_{2i} + \lambda(\mu_{2i}) \quad (29)$$

$$E[y_{1i}|y_{2i} \leq 0] = \mu_{1i} - \rho\sigma_1\lambda(-\mu_{2i}) \quad (30)$$

$$E[y_{2i}^2|y_{2i} \leq 0] = 1 + \mu_{2i}^2 - \mu_{2i}\lambda(-\mu_{2i}) \quad (31)$$

$$E[y_{2i}^2|y_{2i} > 0] = 1 + \mu_{2i}^2 + \mu_{2i}\lambda(\mu_{2i}) \quad (32)$$

$$E[y_{1i}^2|y_{2i}] = \mu_{1i}^2 + \sigma_1^2 - \rho\sigma_1\lambda(-\mu_{2i})(2\mu_{1i} - \rho\sigma_1\mu_{2i}) \quad (33)$$

$$E[y_{1i}y_{2i}|y_{2i} \leq 0] = \mu_{1i}(\mu_{1i} - \lambda(-\mu_{2i}))|\rho\sigma_1 \quad (34)$$

In the above equations, $\lambda(\cdot)$ refers to the inverse Mill's ratio and is defined as :

$$\lambda(x) = \frac{\phi(x)}{\Phi(x)} \quad (35)$$

and

$$\mu_{1i} = \mathbf{x}_{1i}\beta_1 \quad (36)$$

$$\mu_{2i} = \mathbf{x}_{2i}\beta_2 \quad (37)$$

The estimation is only carried out for those i for which the corresponding terms are missing.

The M step consists of the following sub-steps

1. Regress Y_2 on X_2 yielding coefficients $\hat{\beta}_2$.
2. Regress Y_1 on Y_2 and X_1 , yielding coefficients $\hat{\delta}$ for Y_2 and $\hat{\beta}_1^*$ for X_1 , and residual variance $\hat{\sigma}_{1.2}^2$.
3. $\hat{\beta}_1 = \hat{\beta}_1^* + \hat{\delta}\hat{\beta}_2$
4. $\hat{\sigma}_1^2 = \hat{\sigma}_{1.2}^2 + \hat{\delta}^2$
5. $\hat{\rho} = \hat{\delta}/\hat{\sigma}_1$.

6.1 Discussion

The EM algorithm is an iterative method for estimating the maximum likelihood estimates. Hence the estimates calculated with EM have the same properties as those calculated using the full ML estimation procedure. However the individual steps in the algorithm are simpler and do not involve any gradient calculations, which require further constraints like differentiability on the kind of density functions that can be used. Also EM does not directly require estimating the missing data, just the functions of missing data that appear in the likelihood function, which in many cases are much simpler to estimate than the entire missing dataset. It can be shown that the convergence rate of EM is proportional to the amount of information missing from the data [Little and Rubin, 1986]. However like any other non-linear optimization process EM suffers from the problem of getting stuck in local minima. It is also quite sensitive to the initial guess $\theta^{(0)}$ and $\psi^{(0)}$ that is used to start the algorithm.

7 The Heckman Correction

Heckman proposed a simple two step procedure for estimating the correction for the bias in the regression equation [Heckman, 1976]. The crux of his method is to estimate the conditional expectation of ν_1 given ν_2 and use this as an additional variable in the regression equation for Y_1 .

More specifically he begins by estimating the response model. i.e. finding the parameters to equation 10. Heckman chooses a bivariate normal distribution for the joint distribution for the two variables. Thus the marginals for ν_1 and ν_2 are normal. The choice of normal distribution for the response equation is known as choosing a probit model. Now since we never observe the random variable Y_2 ,

directly regressing over Y_2 is not possible, however we observe another variable y_{2i}^* which is defined as :

$$\begin{aligned} y_{2i}^* &= 1 \quad \text{if } y_{2i} > 0 \\ y^* &= 0 \quad \text{otherwise} \end{aligned} \quad (38)$$

We can now construct the likelihood estimator for Y_2^* and X_2 as :

$$Pr[y_{2i}^* = 1] = Prob[\nu_2 > -\mathbf{x}_{2i}\beta_2] = 1 - \Phi(-\mathbf{x}_{2i}\beta_2) \quad (39)$$

hence the likelihood function can be written as

$$L = \prod_{i=1}^{N'} (1 - \Phi(-\mathbf{x}_{2i}\beta_2)) \prod_{i=N'+1}^N \Phi(-\mathbf{x}_{2i}\beta_2) \quad (40)$$

A careful look at the expression for $\Phi(X_2\beta_2)$

$$\Phi(\mathbf{x}_{2i}\beta_2) = \int_{-\infty}^{\mathbf{x}_{2i}\beta_2/\sigma} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt \quad (41)$$

Shows that we will only be able to estimate β_2/σ and not β_2 and σ separately. Hence we might as well assume that $\sigma = 1$ to begin with. This does not affect the generality of the model since β_2 can be scaled to obey the variance restrictions without affecting the overall predictions of the model.

The second step is to estimate the conditional expectation of Y_1 given $Y_2 > 0$ and X :

$$E[Y_1|X_1, Y_2 > 0] = X_1\beta_1 + E[\nu_1|\nu > -X_1\beta_1] \quad (42)$$

$$= X_1\beta_1 + \rho\sigma_1 \frac{\phi(-X_2\beta_2)}{1 - \Phi(-X_2\beta_2)} \quad (43)$$

$$= X_1\beta_1 + \rho\sigma_1 \frac{\phi(X_2\beta_2)}{\Phi(X_2\beta_2)} \quad (44)$$

the term $\frac{\phi(X_2\beta_2)}{\Phi(X_2\beta_2)}$ is referred to as the inverse Mill's ratio and is denoted by $\lambda(X_2\beta)$. The above equation shows that the conditional expectation of y_1 given $y_2 > 0$ and X can be estimated by regressing the observed values of y_{1i} on X_1 and $\lambda(X_2\beta)$

7.1 Generalization to other distributions

In the previous section we described the Heckman correction when the conditional distribution of ν_2 given ν_1 was assumed to normal. This is referred to as a probit/normit response model. The procedure can be generalized for other distributions choices also. Olsen [Olsen, 1980] generalized it for linear and logit models of nonresponse.

Linear refers to the conditional distribution of ν_2 having a uniform distribution of the form :

$$f(u) = \begin{cases} 1/\sqrt{12}, & -\sqrt{3} < u < \sqrt{3} \\ 0, & \text{otherwise} \end{cases} \quad (45)$$

A logit response model corresponds to choosing a logistic density function for the conditional density :

$$f(u) = \frac{\pi e^{\frac{\pi u}{\sqrt{3}}}}{\sqrt{3}(1 + e^{\frac{\pi u}{\sqrt{3}}})^2} \quad (46)$$

This is also referred to as logistic regression.

We now look at a more general formulation. Let us assume :

1. ν_{1i} given ν_{2i} is normal with mean $\alpha\nu_{2i}$ and variance $\sigma_{1.2}^2$.
2. ν_{2i} has a known distribution with mean 0 and variance 1.

Let f and F denote the density and the distribution function describing ν_{2i} , and let $\gamma_i = -X_{2i}\beta_2$. Then :

$$Pr[y_{2i} > 0 | \mathbf{x}_{1i}, \mathbf{x}_{2i}] = 1 - F(\gamma_i) \quad (47)$$

$$E[y_{1i} | y_{2i} > 0, \mathbf{x}_{1i}, \mathbf{x}_{2i}] = \mathbf{x}_{1i}\beta_1 + \alpha\lambda_i \quad (48)$$

where

$$\lambda_i = E[\nu_{2i} | \nu_{2i} > \gamma_i] = \frac{g(\gamma_i)}{1 - F(\gamma_i)} \quad (49)$$

and

$$g(\gamma_i) = - \int_{-\infty}^{\gamma_i} u f(u) du \quad (50)$$

The overall procedure can again be described in two steps.

1. Estimate $\gamma_i = -\mathbf{x}_{2i}\beta_2$ by regressing the response indicator for y_{1i} on \mathbf{x}_{2i} based on the model assumed for the residuals in the response equation. The regression may be probit, linear or logistic depending whether ν_{2i} is assumed to be normal, uniform or logistic respectively.
2. Estimate β_1 and α by regressing the observed values of y_{1i} on \mathbf{x}_{1i} and λ_i respectively.

7.2 Discussion

Heckman in his original presentation required that the covariates for y_1 and y_2 should have atleast one variable which is not common. That is the variable vectors X_2 should have at-least one coordinate that it does not share with X_1 . This is referred to as the collinearity assumption.

Technically this assumption is not required when the response model is assumed to be probit or logit, since the dependence between X_{2i} and λ_i is nonlinear. The only case it is actually applicable is when the response model is assumed to be linear and ν_2 is distributed uniformly. The expression for λ_i becomes :

$$\lambda_i = 2(\sqrt{3} - X_2\beta_2) \quad (51)$$

which is linear in X_2 . This implies that if all the covariates of Y_2 are a subset of a the covariates of X_1 , the two step model will reduce to a single equation.

However in practice it is observed that for the estimation procedure to work even in the case of probit and logit model, the two sets of variables should indeed be distinct and there should be at-least one variable that predicts response strongly.

Practically this a problematic requirement, as identifying variables that affect Y_1 and are not directly responsible for Y_2 is hard.

Little [Little, 1985] in his analysis of Heckman's procedure observes that the estimated selection bias ($E[\nu_1 | \nu_2 > -X_2\beta_2]$) has a very high variance Heckman's procedure observes that the estimated bias has a very high variance, making the estimates suspect.

One of the questions that prompted this study was the observed similarity between the EM algorithm for dealing with missing data and Heckman's procedure [Elkan,]. The analogy was based on the treatment of the λ_i estimates as missing variables, which is how Heckman originally described the solution. Hence it was conjectured that perhaps Heckman's procedure can be iterated just like EM to yield better estimates.

The similarity however is superficial and the Heckman procedure as it stands is not a one-step version of the EM algorithm. There are two principal reasons for this :

1. The dependence in the model is one way. The estimation of the response model does not depend on the actual observed values of y_1 and y_2 , only on whether y_1 was observed or not. In other words the response model can be estimated using X_1 and the R matrix. However since after performing an entire step of the Heckman procedure the R matrix and X_1 remains the same, there is no change in the probability estimates.
2. The collinearity requirement for the two variable sets is artificial and is not necessary for the full EM approach to work.

There is however the possibility that we use the estimated values of y_{1i} as an additional variable in the response equation and estimate the response model again and iterate this process. We are not certain whether the process will even converge. We are in the process of doing some experiments with synthetic data to experimentally verify this.

If one argues that we could use the estimated values for y_{1i} as an additional variable in the equation describing y_{2i} , we make the following two observations :

1. The variables that figure in the set X_1 become a superset of the variables appearing in X_2 , violating the collinearity assumption. As mentioned earlier, even though technically the method will still work, the performance of the methods goes down considerably.
2. The collinearity requirement for the two variable sets is artificial and is not necessary for the full EM approach to work.

8 Conclusions and future work

In this paper we have surveyed the various techniques that are based on model estimation using likelihood methods to deal with survey nonresponse. We looked at full maximum likelihood, expectation maximization and Heckman's method. We also showed why Heckman's method does not correspond to EM.

It is proposed that the regressed value of y_{1i} be used as a additional variable in doing the probit model iterating the two step process. We conjecture that the process will not work very well. We hope to have results soon to demonstrate this claim.

References

- [Berndt et al., 1974] Berndt, E. B., Hall, B., Hall, R., and Hausman, J. A. (1974). Estimation and inference in nonlinear structural models. *Annal of Economic Social Measurements*.
- [Dempster et al., 1977] Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*.
- [Elkan,] Elkan, C. Cost-sensitive learning and decision-making when costs are unknown.
- [Gaharamani and Jordan, 1993] Gaharamani, Z. and Jordan, M. I. (1993). Learning from incomplete data. Technical report, Artificial Intelligence Laboratory, MIT.
- [Georges and Milley,] Georges, J. and Milley, A. H. KDD'99 competition: Knowledge discovery context.
- [Hall,] Hall, B. H. Notes on sample selection models.
- [Heckman, 1976] Heckman, J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables, and a simple estimator for such models. *Annal of Economic Social Measurements*.
- [Little, 1982] Little, R. J. A. (1982). Models for nonresponse in sample surveys. *Journal of American Statistical Association*.
- [Little, 1985] Little, R. J. A. (1985). A note about models for selectivity bias. *Econometrica*.
- [Little and Rubin, 1986] Little, R. J. A. and Rubin, D. B. (1986). *Statistical Analysis with Missing Data*. John Wiley & Sons.
- [Maddala, 1983] Maddala, G. S. (1983). *Limited-dependent and qualitative variables in econometrics*. Cambridge University Press.
- [Olsen, 1980] Olsen, R. J. (1980). A least squares correction for selectivity bias. *Econometrica*.