微观计量分析中缺失数据的极大似然估计①

孙 凤

（清华大学人文学院）

【摘要】微观计量经济分析中常常遇到缺失数据，传统的处理方法是删除所要分析变量中的缺失数据，或用变量的均值替代缺失数据，这种方法经常造成样本有偏。极大似然估计方法可以有效地处理和估计缺失数据。本文首先介绍缺失数据的极大似然估计方法，然后对一实际调查数据中的缺失数据进行极大似然估计，并与传统处理方法的估计结果进行比较和评价。

关键词 极大似然估计 预期—最大算法 缺失数据

中图分类号 F224.0 **文献标识码** A

Maximum Likelihood Estimates for Missing Data of Micro - Econometric Analysis

Abstract: Study on the micro - econometric analysis frequently suffers from missing data. Traditional methods, such as listwise deletion, are inferiors. Maximum likelihood can be an effective and practical method for handling missing data. In this paper, I introdnce the basic principles of ML under missing data at first. Then, I present Expectation - Maximization Algorithm for missing data to obtain ML estimates. Finally, I use the EM to analyze a concrete survey data.

Key words: Maximum likelihood Estimates; Expectation—Maximization Algorithm; Missing Data

一、问题的提出

在微观计量经济分析中常常遇到缺失数据，如被访者拒绝回答问题、访问员疏忽遗漏了某些问题、或在 panel data 研究中，上次被访者在本次调查中可能由于搬迁或死亡而不能继续跟踪调查时，都会造成数据缺失。由于计量经济的分析方法是为完备的数据集设计的，缺失数据给分析带来许多不便，需要利用辅助信息对缺失数据进行补救。比较常用的补救方法

① 本文受国家自然基金资助（批准号：70173040），课题名称为“中国消费者行为研究”。

万方数据

如：①直接使用回答数据的均值替代缺失数据；②利用与缺失数据有关的辅助信息将总体单位分层，在分层基础上进行替代估计或加权调整估计（金勇进，1998）。

然而，缺失数据的存在意味着信息的丢失，如果直接使用回答数据的均值替代缺失数据，达不到我们所设计的估计效果。特别是，如果不回答问题的人群与回答问题的人群存在系统差异时，用现有的回答数据的均值替代缺失数据，分析结果将出现偏差。例如，美联储的消费者金融调查（SCF）采用两个样本框，一个是多阶段地区抽样框，另一个是纳税人信息抽样框，采用后一个抽样框的目的是使在全社会财富分布中处于右端尾部的人分配到更多的样本。Kennickell（1991）对 SCF 数据的分析发现：地区抽样框的响应率是 69%，而纳税人信息抽样框的响应率仅为 34%，其中的原因是富人更倾向于拒绝回答问题，响应率随财富的增加而递减。

比较而言，利用与缺失数据有关的辅助信息进行分层，在分层基础上进行替代估计或加权调整估计在克服回答与未回答问题人群的系统差异上不失为一种权宜之计，但当能够用于分层的辅助信息较多时，缺失数据的估计值将不是惟一的。考虑到这种情形，本文尝试运用极大似然估计法对缺失数据进行估计。

二、极大似然估计法与缺失数据的估计

极大似然估计法是计量经济分析中常用的参数估计方法，该方法通过建立似然函数将观测数据的概率密度表述为未知模型参数的函数，模型参数的最大似然估计是选择能够使这一函数值达到最大的参数估计值。其基本思想如下：

假定我们要估计一个参数 θ，如果 $f(y/\theta)$ 是在给定参数 θ 下的任意观察值 Y 的概率密度，n 个单位的似然函数为：

$$L(\theta) = \prod_{i=1}^{n} f(y_i/\theta) \tag{1}$$

其中 Π 为连乘符号，表示 n 个单位概率密度的联合分布。当然我们还需要明确 $f(y/\theta)$ 的实际分布。假如数据类型为二项分布，编码为 1，0。其似然函数可以表示为：

$$L(\theta) = \prod_{i=1}^{n} \theta^{y_i}(1-\theta)^{1-y_i} \tag{2}$$

极大似然法如何用于缺失数据的估计呢？假如有两个变量 X 和 Y，对于前 m 个单位，我们能够观察到 X 与 Y 值，对于剩余的 $n-m$ 个单位，X 是缺失数据，我们仅能观察到 Y 值。

对于前 m 个单位，由于能够观察到全部数据，我们可以建立概率密度函数 $f(x, y/\theta)$，其中 θ 是待估参数，决定着 X 和 Y 的分布。对于 X 变量中 $n-m$ 个缺失数据，每一单位的似然函数是 Y 的边际分布，以函数表示为：

$$g(y/\theta) = \sum_{x} f(x, y/\theta) \tag{3}$$

对于全部样本的似然函数可以表示为：

$$L(\theta) = \prod_{i=1}^{m} f(x_i, y_i/\theta) \prod_{i=m+1}^{n} g(y_i/\theta) \tag{4}$$

问题的核心是估计 θ 值使似然函数达到尽可能的大。如果两个变量 X，Y 中的 X 有缺失

万方数据

数据，$f(x, y)$ 的联合密度能够被写为 $h(x/y)\ g(y)$，其中 $g(y)$ 是 Y 的边际分布，而 $h(x/y)$ 是在给定 Y 条件下的 X 的分布。这样，我们可以重新构造似然函数为：

$$L(\lambda, \phi) = \prod_{i=1}^{m} h(x_i/y_i;\ \lambda) \prod_{i=1}^{n} g(y_i/\phi) \tag{5}$$

从（5）式可以看出：第一，（5）式是针对全部观察值，而不仅仅是在变量 X 上的缺失数据；第二，θ 参数被分为 λ 和 ϕ 两部分，其中 λ 表示在给定 Y 条件下的 X 的分布，ϕ 表示 Y 的边际分布。这种划分意味着我们能够将该似然函数分成两部分求最大值。

上述理论可以通过一个实例予以说明。假定运用简单随机抽样法抽取 200 人，描述两个二分变量 X，Y（取值为 1，2）的分布及关系。对于 X，Y 两个变量，二者都能被观察的单位是 150 个，如表 1 所示。

表 1　　能够被观察的变量 X、Y 的分布

	$Y=1$	$Y=2$
$X=1$	52	21
$X=2$	34	43

另外 50 个单位，X 是缺失数据，特别地，当 $Y=1$ 时，X 有 19 个缺失数据，当 $Y=2$ 时，X 有 31 个缺失数据。X 和 Y 的关系如表 2 所示。

表 2　　Y 已知情形下，缺失数据 X 的分布

	$Y=1$	$Y=2$
$X=1$	P_{11}	P_{12}
$X=2$	P_{21}	P_{22}

其中，p_{ij} 代表 $X=i$ 和 $Y=j$ 的概率。如果我们用 150 个单位的数据代表全部数据，其似然函数可以表示为：

$$L\ (p_{11})^{52}\ (p_{12})^{21}\ (p_{21})^{34}\ (p_{22})^{43} \tag{6}$$

以上公式的约束条件是四个概率之和为 1。四个概率的极大似然估计为：

$$\hat{p}_{ij} = \frac{n_{ij}}{n} \tag{7}$$

其中 n_{ij} 是落在相格（i，j）中单位的数量，因此我们能够计算得到：

$$\hat{p}_{11} = 0.346;\ \hat{p}_{21} = 0.227;\ \hat{p}_{12} = 0.140;\ \hat{p}_{22} = 0.287$$

然而，由于这些概率值没有用到变量 Y 的另外 50 个值的信息，在推断总体时，可能出现偏差，因此需要将在 X 为缺失数据下的 Y 变量的信息引入似然函数。

在 X 为缺失数据的情形下，$Y=1$ 的边际概率为 $p_{11}+p_{21}$；$Y=2$ 的边际概率为 $p_{12}+p_{22}$，所有样本的似然函数可表示为：

$$L = (p_{11})^{52}\ (p_{12})^{21}\ (p_{21})^{34}\ (p_{22})^{43}\ (p_{11}+p_{21})^{19}\ (p_{11}+p_{22})^{31} \tag{8}$$

如何能得到 p_{ij} 值，而使似然函数达到最大值呢？本例中，由于两个变量中只有一个变量具有缺失数据，我们可以分别估计在给定 Y 情形下 X 的条件分布和 Y 的边际分布。对于

万方数据

2×2的列联表，极大似然估计的一般形式为：

$$\hat{p}_{ij}=\hat{p}\ (X=i/Y=j)\ \hat{p}\ (Y=j) \tag{9}$$

等式右边的条件概率项用现有数据估计，这些值可以用 2×2 表中的次数与列合计相除得到。等式右边 Y 的边际概率项的估计则是将 Y 在每一列上的次数合计（包括在 X 上的缺失数据）除以样本规模得到。具体计算如下：

$$\hat{p}_{11}=\left(\frac{52}{86}\right)\left(\frac{86+19}{200}\right)=0.3174 \qquad \hat{p}_{21}=\left(\frac{34}{86}\right)\left(\frac{86+19}{200}\right)=0.2076$$

$$\hat{p}_{12}=\left(\frac{21}{64}\right)\left(\frac{64+31}{200}\right)=0.1559 \qquad \hat{p}_{22}=\left(\frac{43}{64}\right)\left(\frac{64+31}{200}\right)=0.3191$$

从两种计算方法的比较看，不考虑缺失数据和考虑缺失数据的两种计算结果存在差异，后者较前者对现有信息的利用更多。

三、缺失数据极大似然估计的实际应用

上例是在两个属性变量，且一个变量有缺失数据情形下的极大似然估计方法的应用，当变量是连续变量，且每个变量都存在缺失数据时，该如何运用极大似然法估计缺失数据呢？

Dempster et（1977）、McLachlan 和 Krishnan（1997）的研究结果发现，如果在数据存在缺失情形下，求极大似然估计值的最好方法是 EM 算法。该方法之所以被称作 EM，是由于它的计算过程由两步构成：一步为预期过程（expectation step）；另一步为最大似然过程（maximization step）。这两步通过不断地迭代，最终收敛的结果就是 ML 估计值。

EM 算法的基本思想是：首先确定未知参数均值和方差向量的初始值，初始值是以剔除缺失数据后所计算的均值和协方差确定，基于该初始值，再计算任意一个 X 对其他 X 的回归系数。例如数据集中包括四个变量，X_1、X_2、X_3、X_4，假定变量 X_1 和 X_2 是能够被全部观察到的数据，而 X_3 和 X_4 存在缺失数据。我们用初始值的协方差向量计算 X_3 对 X_1 和 X_2 的回归，以及 X_4 对 X_1 和 X_2 的回归，然后再用该回归模型得到 X_3 和 X_4 的缺失数据的估计值。

当得到所有缺失数据的估计值后，接下来就是通过最大似然过程对非缺失数据计算新的均值和协方差向量。当得到新的均值和协方差的估计量，我们再开始重复 E 过程，即用新的估计量对缺失数据计算新的估计量。不断地轮番使用预期步骤和最大化步骤，直到下一步的估计量与本步的估计量相同，我们认为该估计量达到收敛①。

研究目的是通过对居民就诊满意度指标及数据的分析，了解缺失数据的分布，运用 EM 方法估计缺失数据，将估计的缺失数据填充到数据集中，计算均值、方差、建立回归模型，并将 EM 的计算结果与其他常用的缺失数据处理方法进行比较，以判断各种方法的优劣。

本次调查的有效样本是 620 个，调查的时段为自此调查前 10 个月内是否到过医院。在此调查时段内曾经到过医院的样本是 402 个，本文仅对到过医院的样本数据做一分析。表 3

① EM 方法的迭代过程需要借助统计软件来实现，比较通用的软件如 SPSS 12.0，AMOS4.0，以及 SAS。本文运用实际调查数据对缺失数据的 EM 处理方法做一考察，数据来源于中国人民大学 2002 年北京市居民医疗消费行为意愿调查（调查问卷见易丹辉《北京市居民医疗消费行为及意愿研究》，中国人民大学出版社，2004），分析软件是 SPSS 12.0。

万方数据

是对居民就诊满意度的 9 个指标的基本描述①，数据显示每个变量都有不同程度的缺失数据，最高占全部数据的 47.8%，最低占 21.6%。

表 3　缺失数据的分布

	样本数	非缺失	缺　失	
			频　数	频　率
tt4	402	261	141	35.1
tt5	402	292	110	27.4
tt6	402	315	87	21.6
tt7	402	257	145	36.1
tt8	402	250	152	37.8
tt9	402	210	192	47.8
tt10	402	214	188	46.8
tt11	402	259	143	35.6
tt12	402	234	168	41.8

表 4 和表 5 反映的是对缺失数据的不同处理方法，产生不同的均值和标准差。其中包括剔除缺失值（listwise），预期最大似然法（EM），以及回归法（Regression）。表下方的 Little's MCAR 检验结果为拒绝原假设，认为本次调查数据的缺失不是完全随机缺失，用剔除法或均值替代，将使结果有偏。从各种均值与标准差的比较看，剔除法与其两类方法存在较大差异。比较看，EM 法与回归法的计算结果比较接近，而与剔除法则有较大出入。

表 6 和表 7 是将 EM 模型的估计值与回归法的估计值分别替代缺失值，然后再分别建立回归方程而得到的回归系数以及检验结果。比较看，以 EM 方法估计参数的 t 检验结果优于回归法的 t 检验结果。从本研究各项指标的比较看，在对缺失数据的处理上，EM 方法最优，回归方法次之，而剔除缺失数据的方法由于损失了信息，与前两种方法有较大的出入。

表 4　三种估计方法的均值比较

	tt4	tt5	tt6	tt7	tt8	tt9	tt10	tt11	tt12
剔除法	5.2857	4.7143	3.8571	3.4286	4.0000	4.8571	5.2857	4.4286	5.0000
预期最大似然法	5.8978	5.1676	4.7061	5.0225	4.8982	5.7845	5.9125	5.1485	5.3560
回归法	5.9127	5.1422	4.6226	5.0051	4.7697	5.8180	5.9254	5.1229	5.4516

Little 的完全缺失检验（Little's MCAR test）：卡方（Chi - Square）＝1715.089，自由度（DF）＝321，显著性检验（Sig）.＝0.000。

表 5　三种估计方法的标准差比较

	tt4	tt5	tt6	tt7	tt8	tt9	tt10	tt11	tt12
剔除法	1.40184	1.92906	1.57064	1.86290	2.23061	1.37169	1.77084	1.41668	1.79430
预期最大似然法	1.21936	1.64377	1.75574	2.02807	1.86620	1.35530	1.38853	1.56133	1.61446
回归法	1.27885	1.63125	1.69656	1.99851	1.75970	1.35406	1.33830	1.52205	1.59747

Little 的完全缺失检验（Little's MCAR test）：卡方（Chi - Square）＝1715.089，自由度（DF）＝321，显著性检验（Sig）.＝0.000。

① 本次调查反映居民就诊满意度的指标有 9 个，分别是：tt4 总体满意度；tt5 总体收费合理性；tt6 药品费用合理性；tt7 治疗费用合理性；tt8 就诊手续方便度；tt9 环境卫生满意度；tt10 硬件设施满意度；tt11 医务人员专业水平；tt12 医务人员服务态度。该问卷通过设计 10 级量表，判断满意程度，其中“1”为特别不满意，“10”为特别满意。

万方数据

表 6　EM 法填充缺失值后的回归系数

模　型		非标准化系数		标准化系数	t 统计量	显著性检验值
		估计值	标准差	估计值		
1	常数项	5.010	0.129		38.743	0.000
	tt5	−0.472	0.018	−0.658	−25.792	0.000
	tt6	0.307	0.016	0.468	18.684	0.000
	tt7	0.502	0.014	0.853	34.643	0.000
	tt8	−0.293	0.014	−0.444	−20.491	0.000
	tt9	−0.934	0.034	−0.999	−27.675	0.000
	tt10	0.624	0.030	0.670	20.833	0.000
	tt11	0.775	0.017	0.994	46.228	0.000
	tt12	−0.276	0.017	−0.333	−16.607	0.000

a　因变量：tt4。

表 7　回归法填充缺失值后的回归系数

模　型		非标准化系数		标准化系数	t 统计量	显著性检验值
		估计值	标准差	估计值		
1	常数项	3.365	0.351		9.595	0.000
	tt5	−0.061	0.042	−0.077	−1.452	0.147
	tt6	0.147	0.038	0.201	3.824	0.000
	tt7	0.199	0.032	0.321	6.194	0.000
	tt8	−0.178	0.035	−0.251	−5.104	0.000
	tt9	−0.054	0.052	−0.059	−1.050	0.294
	tt10	−0.007	0.051	−0.007	−0.129	0.898
	tt11	0.425	0.041	0.510	10.483	0.000
	tt12	0.043	0.037	0.053	1.180	0.239

a　因变量：tt4。

以上实证研究表明，缺失数据的极大似然估计，不仅可以从未缺失数据的分布情况中推算出缺失数据的估计值，从而有效地使用所有数据进行分析，提高统计结果的可信度。而且，用估计值替代缺失值也可以消除数据中隐藏的偏向，将所有层级间的差异组在分析结果中表现出来，从而得出更加精确的结论。

微观计量经济学分析中的数据来自抽样调查，缺失数据常常发生。因此在建立计量经济模型、进行参数估计和拟合优度检验前，首先需要对缺失数据进行有效处理，只有这样，才能使分析建立在坚实的基础之上，所得出的结论和政策建议才有实际价值。

参 考 文 献

[1] Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), *Maximum likelihood estimation from incomplete data via the EM algorithm*, Journal of the Royal Statistical Society, series B, 39, 1～38.

[2] Little, R. J. A. (1988), *Missing data in large surveys (with discussion)*, Journal of Business and Economic Statistics, 6, 287～301.

[3] McLachlab, G. J., and Krishnan, T. (1997), *The EM algorithm and extensions*, New York: Wiley. Paul D. Allison, (2001), *Missing data*. Sage publication, Inc.

[4] 金勇进：《处理缺失数据中辅助信息的利用》，《统计研究》1998 年第 1 期。

[5] 易丹辉：《北京市居民医疗消费行为及意愿研究》，中国人民大学出版社，2004。

（责任编辑：朱长虹）

万方数据