武汉纺织大学学术著作出版基金资助出版

农业经济调查数据的缺失值处理：模型和方法

潘传快 著

Treatment of Missing Value in Agricultural Economic Research Data: Model and Method

汕頭大學出版社

纺织大学学术著作出版基金资助出版

农业经济调查数据的缺失值处理：模型和方法

潘传快 著

Treatment of Missing Value in Agricultural Economic Research Data: Model and Method

汕頭大學出版社

图书在版编目（CIP）数据

农业经济调查数据的缺失值处理：模型和方法 / 潘传快著 . —汕头：汕头大学出版社，2018.10
ISBN 978-7-5658-3658-9

Ⅰ . ①农… Ⅱ . ①潘… Ⅲ . ①农业经济—调查—统计数据—研究—中国 Ⅳ . ① F32

中国版本图书馆 CIP 数据核字（2018）第 129133 号

农业经济调查数据的缺失值处理：模型和方法
NONGYE JINGJI DIAOCHA SHUJU DE QUESHIZHI CHULI：MOXING HE FANGFA

著　　者：潘传快
责任编辑：李金龙
责任技编：黄东生
封面设计：汤　丽
出版发行：汕头大学出版社
广东省汕头市大学路 243 号汕头大学校园内　邮政编码：515063
电　　话：0754-82904613
印　　刷：北京市金星印务有限公司
开　　本：787mm × 1092mm　1/16
印　　张：14
字　　数：201 千字
版　　次：2019 年 1 月　第 1 版
印　　次：2019 年 1 月　第 1 次印刷
定　　价：50.00 元
ISBN 978-7-5658-3658-9

序　言

中国是农业大国，虽然农业产值占 GDP 的比重不到 10%，但农业就业人口仍占总就业人口的 28.3%，农业仍旧是国民经济和社会发展的基础。在科学研究中，农业经济以及农业管理的研究仍是必要和重要的，而这些农业经济管理研究很多都需要通过农业经济调查获取数据，然后在数据分析的基础上得出结论。

跟任何调查一样，农业经济调查会遇到一个几乎不可避免的问题：缺失值，农户不回答或者调查人员疏忽都会让农业经济调查数据产生缺失值。但跟其他调查（如市场调查、民意调查）不同的是，农业经济调查有很强的特殊性，比如调查仍使用古老的人员访问方法、调查问卷中存在大量的开放性问题、能获得较多的辅助信息、随机性不高但农户调查配合度较高等。

结合农业经济调查的特点和数据缺失的原因，本研究提出了特定的假设条件：调查数据来自一个正态总体；调查数据是随机获得的；变量类型以数量变量为主；数据的缺失模式是单一缺失和一般缺失；数据的缺失机制是完全随机缺失（MCAR）和随机缺失（MAR）。基于这些基本假设，本研究针对农业经济调查数据的缺失值处理构建了一套较系统的模型和方法。这套模型有一个完整的逻辑体系，但为了叙述方便，分成三部分：删

除模型、单一插补模型和多重插补模型，每一个模型又包括很多具体的方法。

本研究的基本逻辑是，根据假设和条件提出一个模型，在这个模型中提出基本方法，利用理论分析和模拟分析找出其缺陷，改进后提出新的方法；如果假设和条件改变，又使用新的模型，为新的模型寻求方法并分析改进。

删除是缺失值处理的最基础模型。大部分农业经济调查人员都利用该模型将缺失值当作无效数据删除，大部分的数据分析软件也默认删除缺失值。删除一般是指成列删除，也就是删除所有含缺失值的个案，留下完整数据。当数据的缺失比重很小时，删除缺失值倒也无所谓，但是当数据缺失比重较大或者变量很多时，就会导致大量个案被删除。本书用不同的缺失比例模拟随机产生农业经济调查缺失数据，分析发现当变量很少时，缺失比重略高一点删除比例也不会太高；但当变量稍微多一点，哪怕很小的缺失比重都将致使大量的数据被删除。

一个可供替代的删除方法是，如果我们不需要完整数据，只要使用可用的个案计算参数估计，就可以尽量减少数据删除，这就是成对删除。但成对删除会让估计量来自不同大小的样本，造成很多参数估计上的麻烦。此外，本书的模拟分析发现，其实成对删除在对缺失农业经济调查的相关关系估计上并没有显著超过成列删除。

当数据不是完全随机缺失（MCAR）时，无论成列删除还是成对删除都会产生有偏的估计。可以利用辅助信息将目标缺失变量分层，根据各层的完整观测数据计算各层均值，然后再将各层均值以缺失概率做权数加权平均，这样就可以在一定程度上弥补成列删除估计的有偏性，这就是加权调整的方法。本书通过模拟生成随机缺失（MAR）下的目标缺失变量和与之正相关的辅助变量，然后采用加权调整方法获得的均值估计非常接近真值，而成列删除的均值估计明显偏小。

用删除方法删掉的数据信息也许是有用的，再者，因为缺失值的存在而粗暴地删除农业经济调查数据，从心理上也是令人难以接受的，对数据的缺失值进行插补也许是一种更好的方法。插补分为单一插补和多重插补，前者指为缺失值提供单一插补值；后者是指每一个缺失值的插补值不止一个。插补的基本思想是根据数据的后验分布，用数据的观测部分为缺失部分提供合理的插补值。

简单均值插补是将目标缺失变量的观测部分的均值作为缺失值的插补，是最先能想到的单一插补方法。但简单均值插补的插补值完全集中于数据的中心位置，通过理论分析容易发现其显著低估了总体方差。一个解决方法是在其基础上加上随机误差项，这就是随机均值插补。本书还进一步做了一个模拟研究，那就是模拟产生变量正相关的农业经济调查缺失数据进行均值插补，最后发现其相关系数矩阵和协方差矩阵中的值明显小于真实相关系数矩阵和协方差矩阵的值。但无论简单均值插补还是随机均值插补，在数据非完全随机缺失（MCAR）的情况下，估计都是有偏的。分层均值插补可以修正这个问题，分层均值插补是指将目标缺失变量按照辅助信息分层，然后在各层中进行均值插补，这样其估计是无偏的。

分层均值插补虽然解决了一般均值插补的估计有偏问题，但插补值仍过于集中，回归插补可以解决这个问题。简单回归插补是指根据农业经济调查缺失数据的后验分布，利用数据的观测部分产生缺失部分的回归预测值，通过理论分析发现其对总体方差的估计仍偏小，可以加上随机残差项，这就是随机回归插补。对回归插补和均值插补进行对比模拟研究，结果显示回归插补是一个比均值插补更好的方法，尤其是随机回归插补有很好的插补效果，而简单均值插补是最不被推荐的。

如果农业经济调查缺失数据没有明显的后验分布，热平台插补方法会是更好的选择。热平台方法直接从数据的完整部分产生缺失部分的插补值，

其插补值一般比较稳健，不用担心像回归插补一样产生异常的插补值。一个简单的热平台插补是从完整观测数据中简单随机抽样产生插补值，这就是简单随机插补。如果数据是随机缺失（MAR）的，一个更好的方法是利用辅助信息将目标缺失变量分层，然后在各层的完整观测数据中随机产生该层的插补值，这就是分层随机插补。热平台插补还有一个很有效率的方法，就是利用辅助变量，找到缺失值最接近的观测值作为自己的插补值，这就是最近距离方法。本书中的一个针对热平台插补和均值插补、回归插补进行对比的模拟分析发现，在完全随机缺失（MCAR）的情况下，基于热平台的随机插补效果显著好于均值插补，但可能比回归插补略差。

根据单一插补后的数据进行估计检验时，其标准误差常常是被低估的，多重插补是解决这个问题的最有效的模型。多重插补的基本思想是，对同一缺失值产生多个插补值，这样就产生多个“完整”数据，然后对每一个“完整”数据进行估计检验，最后将其汇总成一个总的估计检验结果。

基于单一缺失的一元正态模型仍然利用回归插补产生插补值，但其从两个角度让缺失值的不同插补值差异加大，一是跟回归插补一样在插补值中加入残差项，二是让每一次插补的回归模型参数随机产生。回归模型参数的随机产生方法有两种，一是根据回归模型参数的后验分布随机产生模型参数，这就是贝叶斯方法；二是用数据的 Bootstrap 样本来产生模型参数，这就是 Bootstrap 方法。本书首先分析了这两种方法的假设和理论，然后为了比较这两种方法的应用效果，在完全随机缺失的假设下模拟产生缺失数据，然后分别用贝叶斯法和 Bootstrap 方法进行插补，并跟单一插补进行比较，结果发现无论贝叶斯法还是 Bootstrap 方法，都有很好的估计检验效果，其估计的准确性显著超过单一插补。

多元正态模型是基于一般缺失模式的农业经济调查缺失数据的插补。多元正态模型，由于其缺失模式的复杂性，对缺失值的插补提出了更大的

挑战。本书研究了其中应用最为广泛的联合分布方法以及条件分布方法的假设和理论。本书更进一步模拟了一个多变量随机缺失的农业经济调查数据，然后运用这两个方法进行插补，结果显示两者都有很好的估计检验效果，而且两者之间差异并不大，都是很好的方法。

在理论和模拟分析的基础上，本书对一次实际农业经济调查缺失数据进行了应用分析并取得了较好的效果。通过实际应用分析可以得到一个基本的结论，那就是如果数据基本符合缺失值处理模型的假设，多重插补优于单一插补，而单一插补又优于删除；如果不符合假设，比如出现极端值，那么基于明确后验分布的缺失值插补的效果会大打折扣，而此时基于热平台的插补方法会得到更稳健的结果。

本研究基于研究结果为农业经济管理研究人员在缺失值处理前和缺失值处理中两个阶段分别给出了一定的具体建议。在缺失值处理前的建议：调查前通过良好的问卷设计减少缺失值产生；调查中通过与农户良好的沟通减少缺失值产生；及时处理无意义值，以免跟缺失值混淆；不要用不科学的方法消除缺失值。在缺失值处理中的建议：正视缺失值问题；尽量不要删除缺失值；善于利用分类变量处理缺失值；插补缺失值前对缺失数据进行描述考察；单一插补时选择回归插补；在数据一般缺失时使用多重插补。

本研究可能的创新有：

第一，本研究率先关注了农业经济调查数据的缺失值处理问题，并基本厘清了其学理。虽然在农业经济调查中缺失值无可避免，但绝大部分农业经济管理研究人员都将其忽略了，更鲜有人对其进行系统研究，使得该领域的研究，特别是国内研究基本空白，这也是笔者开始这项研究的重要原因。

第二，本研究专门针对中国农业经济调查的特点模拟缺失数据进行分析，具有一定的创新性和开创性。本书针对中国农业经济调查数据的缺失

值处理，提出了一整套具体而又可行的模型和方法体系，为了分析这些方法的可行性和使用条件，并对不同模型和方法的效果进行比较，采用了理论分析和模拟分析。而其中的很多模拟分析针对中国农业经济调查特点、缺失模式、缺失机制进行了专门的设计。

第三，本研究为农业经济调查数据中缺失值的实际处理和应用自编了一套具体的基于 R 软件的程序代码，并用于实际案例的应用分析，效果较好。该语言程序包括农业经济调查缺失数据的预分析、缺失值的处理和处理结果的定量分析。实际案例的处理结果显示，对于基本达到假设条件的农业经济调查缺失数据，本套语言程序能达到较好的缺失值处理效果。

目　录

1　导　论 …… 1

1.1　研究背景和研究意义 …… 1

1.1.1　研究背景 …… 1

1.1.2　研究意义 …… 4

1.2　主要概念界定 …… 5

1.2.1　农业经济调查 …… 5

1.2.2　缺失值 …… 6

1.3　问题的提出与研究目标 …… 7

1.3.1　问题的提出 …… 7

1.3.2　研究目标 …… 8

1.4　技术路线与结构框架 …… 8

1.4.1　技术路线 …… 8

1.4.2　结构框架 …… 9

1.5　研究方法与数据来源 …… 11

1.5.1　研究方法 …… 11

1.5.2 数据来源 …… 12
1.6 可能的创新与不足 …… 13
1.6.1 可能的创新 …… 13
1.6.2 不足之处与展望 …… 14

2 农业经济调查数据缺失值处理的文献综述 …… 15

2.1 缺失值处理的理论和方法的研究综述 …… 16
2.1.1 国外缺失值处理的理论和方法的研究综述 …… 16
2.1.2 国内缺失值处理的理论和方法的研究综述 …… 20
2.2 农业经济调查数据缺失值问题的研究综述 …… 21
2.2.1 国外农业经济调查数据缺失值问题的研究综述 …… 21
2.2.2 国内农业经济调查数据缺失值问题的研究综述 …… 23
2.3 结论和评价 …… 25

3 农业经济调查数据缺失值处理的研究基础和假设 … 27

3.1 农业经济调查的特点和数据缺失原因 …… 28
3.1.1 农业经济调查的特点 …… 28
3.1.2 农业经济调查数据缺失的原因 …… 30
3.2 基本概念及符号表示 …… 31
3.2.1 基本概念及符号 …… 31
3.2.2 缺失数据及缺失信息的转换 …… 32
3.3 模拟方法介绍 …… 34
3.3.1 模拟方法的含义 …… 34
3.3.2 采取模拟方法的原因 …… 34
3.3.3 模拟方法的优势 …… 35

3.4 农业经济调查数据的缺失模式 …… 35
3.4.1 一般缺失模式 …… 35
3.4.2 单一缺失模式 …… 36
3.4.3 单调缺失模式 …… 37
3.5 农业经济调查数据的缺失机制 …… 37
3.5.1 农业经济调查数据缺失机制及模型 …… 37
3.5.2 农业经济调查数据缺失机制的模拟 …… 39
3.6 基本假设 …… 41
3.6.1 农业经济调查总体分布的假设 …… 41
3.6.2 农业经济调查样本随机性的假设 …… 42
3.6.3 农业经济调查的变量假设 …… 42
3.6.4 农业经济调查数据缺失模式的假设 …… 43
3.6.5 农业经济调查数据缺失机制的假设 …… 44
3.7 缺失值处理的统计软件 …… 44
3.7.1 分析软件 …… 44
3.7.2 本研究使用的软件 …… 45
3.7.3 本研究自编的 R 程序代码 …… 45

4 农业经济调查数据缺失值处理的删除及模拟分析 … 46

4.1 成列删除及其缺陷分析 …… 46
4.1.1 成列删除及其争议 …… 46
4.1.2 成列删除引致的估计错误分析 …… 48
4.1.3 成列删除引致数据损失和估计错误的模拟分析 …… 50
4.2 成对删除及比较分析 …… 54
4.2.1 成对删除及其争议 …… 54

4.2.2 成对删除的估计复杂性分析 …… 55
4.2.3 成对删除和成列删除在相关关系估计上的模拟比较分析 …… 57
4.3 随机缺失下成列删除有偏估计的加权调整分析 …… 60
4.3.1 加权调整的基本模型 …… 60
4.3.2 加权调整的方法 …… 61
4.3.3 加权调整效果的模拟分析 …… 63
4.4 结论和讨论 …… 65

5 农业经济调查数据缺失值处理的单一插补及模拟比较分析 …… 67

5.1 单一插补的模型和缺陷分析 …… 67
5.1.1 单一插补的基本思想 …… 67
5.1.2 单一插补的基本模型 …… 68
5.1.3 单一插补的缺陷分析 …… 69
5.2 均值插补及其改进分析 …… 70
5.2.1 关于均值插补的讨论 …… 70
5.2.2 简单均值插补对总体方差的低估分析 …… 71
5.2.3 均值插补离散性的改进分析 …… 73
5.2.4 随机缺失下均值插补估计偏差的修正 …… 75
5.2.5 简单均值插补对相关关系低估的模拟分析 …… 76
5.3 回归插补及其插补效果的比较分析 …… 78
5.3.1 关于回归插补的讨论 …… 78
5.3.2 简单回归插补及对总体方差的低估分析 …… 79
5.3.3 回归插补的改进分析 …… 81

5.3.4 回归插补效果的模拟比较分析 …… 83
5.4 基于模糊后验分布的热平台插补及比较分析 …… 86
5.4.1 关于热平台插补的讨论 …… 86
5.4.2 简单随机插补的稳健性分析 …… 87
5.4.3 随机缺失下随机插补的改进 …… 87
5.4.4 最近距离插补及其模型方法 …… 88
5.4.5 热平台插补效果的模拟比较分析 …… 89
5.5 结论和讨论 …… 90

6 农业经济调查数据缺失值处理的多重插补及比较应用分析 …… 93

6.1 多重插补的基本思想和基本模型 …… 94
6.1.1 多重插补的基本思想 …… 94
6.1.2 多重插补的基本模型 …… 94
6.1.3 关于多重插补的插补次数选择的讨论 …… 96
6.2 多重插补的参数估计和检验 …… 96
6.2.1 多重插补的点估计 …… 97
6.2.2 多重插补估计量的分布 …… 99
6.2.3 多重插补的参数估计和检验方法 …… 100
6.3 一元正态模型下的贝叶斯法多重插补及比较分析 …… 101
6.3.1 一元正态线性模型的假设 …… 101
6.3.2 贝叶斯多重插补方法的理论分析 …… 102
6.3.3 贝叶斯多重插补方法的参数估计 …… 103
6.3.4 贝叶斯多重插补方法的模拟比较分析 …… 104
6.4 一元正态模型下 Bootstrap 多重插补及比较分析 …… 108

6.4.1 Bootstrap 法和贝叶斯法在模型假设上的异同…… 108
6.4.2 Bootstrap 多重插补方法的理论分析…… 108
6.4.3 Bootstrap 多重插补方法的模拟比较分析…… 109
6.5 多元正态模型下联合分布多重插补及其应用分析 …… 113
6.5.1 农业经济调查的多变量缺失问题 …… 113
6.5.2 多元正态模型下联合分布多重插补方法的假设 …… 113
6.5.3 农业经济调查数据的具体缺失模式 …… 114
6.5.4 联合分布多重插补方法的理论分析 …… 115
6.5.5 联合分布多重插补方法的模拟应用分析 …… 118
6.6 多元正态模型下条件分布多重插补及模拟应用分析 …… 124
6.6.1 多元正态模型下条件分布多重插补方法的假设 …… 124
6.6.2 条件分布多重插补方法的模型 …… 125
6.6.3 条件分布多重插补方法的模拟应用分析 …… 126
6.7 结论与讨论 …… 127

7 农业经济调查数据缺失值处理的实例应用分析 ……130

7.1 数据缺失信息描述 …… 131
7.1.1 数据介绍 …… 131
7.1.2 数据整理 …… 132
7.1.3 缺失信息描述 …… 134
7.2 单一缺失数据处理 …… 136
7.2.1 目标变量和辅助变量选择 …… 136
7.2.2 单一插补分析 …… 138
7.2.3 多重插补分析 …… 139
7.3 一般缺失数据处理 …… 141

7.3.1 联合分布多重插补分析 …… 141
7.3.2 条件分布多重插补分析 …… 142
7.4 结论和讨论 …… 143

8 结论和建议 …… 145

8.1 结论 …… 145
8.1.1 关于农业经济调查特点的结论 …… 145
8.1.2 关于删除的结论 …… 146
8.1.3 关于单一插补的结论 …… 147
8.1.4 关于多重插补的结论 …… 148
8.1.5 关于实际应用分析的结论 …… 149
8.2 建议 …… 150
8.2.1 处理缺失数据前的建议 …… 150
8.2.2 处理缺失数据中的建议 …… 151

参考文献 …… 153
附录 1 符号表示 …… 172
附录 2 证明和说明 …… 173
附录 3 R 程序代码 …… 176
附录 4 原始数据（部分） …… 206
致 谢 …… 208

1 导 论

1.1 研究背景和研究意义

1.1.1 研究背景

农业作为第一产业，是国民经济和社会发展的基础。中国是农业大国，中华人民共和国国家统计局公布的数据显示，尽管农业产值仅占中国 GDP 总值的 9.0%，但农业就业人口却占到总就业人口的 28.3%。对农业发展、农业经济以及农业管理的研究仍然在中国的社会和经济研究中占有重要的地位。而这些研究很多都是基于数据的定量研究，这就需要开展农业经济调查以搜集数据。

虽然无法精确统计在农业经济管理研究中开展农业经济调查的数量，但可以通过文献搜索进行大致的估计。在中国知网（http://www.cnki.net/）中，以“调查”为关键词搜索所有 1990—2016 年的农业类文献，其中在篇名中含“调查”的有 52815 篇，摘要中含“调查”的有 159105 篇。由此可见，在中国的农业经济管理研究中开展的农业经济调查的数量是庞大的，图 1-1 进一步展示了这些文献在各年份的变化情况，可以看到在 2008 年后基于调查的农业经济管理研究以迅猛的速度在增长。