天津财经大学
Tianjin University of Finance and Economics

TUFE

硕士学位论文

三种线性回归多重插补法的模拟比较

一级学科：应用经济学

二级学科：统计学

论文作者：董世杰

指导教师：杨贵军　教授

二〇一七年五月

分类号：
密　级：

硕士学位论文

三种线性回归多重插补法的模拟比较

Simulation Comparison of Three Linear Regression Multiple Imputation

所属学院：理工学院

所在系别：统计系

年　　级：2014 级

学　　号：2014310192

论文作者：董世杰

天津财经大学学位论文原创性声明

本人郑重声明：所呈交的学位论文：

《三种线性回归多重插补法的模拟比较》，

是本人在导师指导下，在天津财经大学攻读学位期间进行研究所取得的成果。除文中已经注明引用的内容外，不包含任何他人已发表或撰写过的研究成果。对本论文研究工作做出贡献的个人和集体，均已在文中以明确方式标明。

本声明的法律责任完全由本人承担。

学位论文作者签名：董世杰

2017 年 6 月 20 日

天津财经大学学位论文版权使用授权书

本人完全了解天津财经大学关于收集、保存、使用学位论文的规定，即：按照学校要求提交学位论文的印刷版本和电子版本；同意学校保留论文的印刷版本和电子版本，允许论文被查阅和借阅。本人授权天津财经大学可以将本学位论文的全部或部分内容编入有关数据库进行检索；可以采用影印、缩印或其他复制手段保存或汇编论文；可以向有关机构或者国家部门送交论文的印刷本和电子版本；在不以赢利为目的的前提下，学校可以复制论文的部分或全部内容用于学术活动。

本学位论文属于：

（　）1. 经天津财经大学保密委员会审查核定的保密学位论文，于　　年　月　日解密，解密后适用上述授权。

（✓）2. 不保密，适用上述授权。

（请在以上相应括号内打“ √ ”或填上相应内容。保密学位论文应是已经天津财经大学保密委员会审定过的学位论文，未经天津财经大学大学保密委员会审定的学位论文均为公开学位论文。此声明栏不填写的，默认为公开学位论文，均适用上述授权。）

作者签名：董世杰　　日期：2017 年 6 月 20 日

导师签名：杨贵军　　日期：2017 年 6 月 20 日

内容摘要

在社会中众多领域的统计调查研究中，无回答现象时有发生。无回答会影响统计分析结果的真实性和精确程度，从而降低了调查数据的质量。因此，对无回答的处理在实际的统计调查研究中则显得十分重要。

处理无回答的办法主要分为两大类：一类是在实际调查之前采取预防措施，这类方法虽然可以减少无回答情况的发生频率，但其花费的成本较高，也不能完全消除无回答现象；另一类则是在调查之后采用不同的方法进行补救，此类方法在一定程度上降低了调查中的误差，增加了调查数据的可靠程度，并且节约了成本。

在处理含有无回答数据的众多方法当中，多重插补法是一类较为理想的办法。多重插补法则用多个插补值对无回答插补，得到的插补数据可信度较高，从而提高了调查结果的准确性。

在多重插补方法中，PMM、DA、EMB 多重插补法是目前使用较为普遍，且性能较好的插补方法，并且国内外很多学者对这些方法进行了相关细致深入的研究，但在对这些方法进行模拟研究时，得到的估计量的偏差、均方误差都相对较大。本文则采用普通线性回归多重插补法、贝叶斯线性回归多重插补法、贝叶斯自助线性回归多重插补法对无回答进行插补，在不同插补重数、不同无回答率、不同无回答机制下对线性回归模型系数的估计量做比较研究，并对估计量的偏差、均方误差的偏度以及峰度进行分析，考察其统计学性质。模拟结果显示，这三种多重插补法利用了无回答与解释变量之间的线性关系，显著改善了插补结果。当插补重数选择为 5 时，分别采用三种多重插补法都能给出较好的回归系数估计量，估计量的偏差绝对值和均方误差都显著小于采用 PMM 和 DA 插补法的情况；与采用 EMB 插补法的情况相比差异较小。在无回答率增加的情况下，系数估计量的偏差绝对值和均方误差呈较小递增趋势。在依赖无回答变量的非随机无回答机制下，截矩项估计量具有较大的偏差绝对值和均方误差，其他系数估计量的偏差绝对值和均方误差都相对较小。同时，各系数估计量偏差、均方误差相比于正态分布都有不同的偏斜；而其峰度值均为负，相比于正态分布更为平坦，数据更分散。

关键词：线性回归多重插补法；贝叶斯；无回答机制；无回答率；插补重数

Abstract

In the fields of social statistics research, non-response problem has occurred frequently. Non-response problem will affect the authenticity and reliability of statistical analysis results, thus affecting the quality of the survey data. Therefore, the handle of non-response in the actual investigation and study is very important.

There are two main ways to deal with non-response problem: one is to take precautions before the actual investigation. Although this method can reduce the frequency of non-response, but the cost of investigation is higher and it can not completely eliminate non-response. And the other is, after the investigation, using different methods to remedy the incomplete data set. These a variety of methods will reduce the error in survey and enhance the reliability of survey data, also decrease the cost.

Imputation method is one of the most desirable methods in dealing with non-response data. The multiple imputation methods use different imputation values for non-response data set, so the result of imputation data set is highly reliable, thus improving the accuracy of the survey results.

Under a variety of interpolation multiplicities, non-response rates and non-response mechanisms, this paper simulates coefficients estimators of linear model based on the Ordinary linear regression multiple imputation, Bayesian linear regression multiple imputation, Bayesian bootstrap linear regression under the condition of non-response. The simulation results show that the three multiple imputation methods significantly improve the imputation results, using the linear relationship between the non-response and explanatory variables. When the number of imputations are selected for 5, using three multiple imputation methods respectively can give better estimate of the regression coefficients, the absolute values of the bias and mean squared error of the coefficients estimators are significantly smaller than the ones using the PMM and DA imputation methods; the difference with the ones using the EMB imputation method is small. With the increase of non-response rate, the absolute values of the deviation and mean squared error of the coefficients estimators present a small increasing trend. Under non-response dependent variable of non-random at non-response mechanism, the absolute values of the bias and mean squared error of the intercept estimators are larger, the absolute value of the deviation and the mean square error of the other coefficients estimators are relatively small. Under the non-random non-response mechanism of the non-response variable, the intercept term has a

large deviation absolute value and mean square error, and the absolute value and mean square errors of the other coefficient estimators are relatively small. At the same time, the deviation of the coefficients and the mean square error are different from those of the normal distribution. The kurtosis value is negative, and the data are more dispersed than the normal distribution.

Key word: Linear Regression Multiple Imputation; Bayesian; Non-response Mechanism; Non-response Rate; Number of Imputation.

目录

第1章 导论

1.1 选题背景和意义

随着社会经济的发展，在许多领域中遇到的问题或者现象都需要用统计学方法加以解释和分析。尤其是在统计调查中收集到的数据往往决定着分析结果的优劣。由于无回答现象普遍存在于各类调查中，从而严重影响了调查结果的准确性。如何应对无回答带来的影响以及如何应用统计学方法处理无回答现象，这在诸多统计调查中是亟待解决的问题。

无回答现象会影响抽样调查的分析结果，带来的问题主要包括：无回答造成观测值的数量减少，导致数据信息损失增加，增大了统计估计量的偏差；在调查中得到的样本结果信息不能体现真实的情况，通常这样的统计分析结果很可能是无效的。无回答会使数据处理变得复杂，降低工作效率；而且会降低有效样本量，严重影响估计的精度，导致错误的分析结果。因此，深入探讨无回答的解决方法以及对无回答问题的处理在抽样调查工作中就十分重要。

处理无回答问题的主要方法包括提前采取预防措施和调查后的补救。事前预防是指通过采取精心设计调查方案，提高调查人员的政治素质和业务素质，明确统计调查对个人资料的保密原则，使被调查者免除后顾之忧等方法来降低问卷的无回答率，从而得到回答完整的数据信息。但在实际调查中会遇到许多复杂的情况，事前预防的措施尽管考虑较为细致，还是不能避免无回答的出现。因此目前较好的解决办法是做好事后补救。事后补救是通过对调查结果再抽样，然后对无回答进行插补，再对新的完整数据集进行参数估计的方法。对于实际调查的人员来说，如何采用适当的方法对无回答数据集进行研究，是数据处理过程存在的难题。目前解决此类问题较为广泛使用的方法是插补法。

插补法是解决无回答最主要的事后补救方法之一，其主要目的不是测算无回答值，而是对无回答的分布进行分析。此外，插补法能够融合数据收集者的知识，应用标准的完全数据分析方法，避免删除数据不完整所造成的信息丢失，某些情况下，插补法还能够降低无回答带来的偏差。插补法是解决无回答的最常用方法（Rubin,1987; Okafor and Lee,2000;

王璐和王飞，2006）。插补法可以通过计算得出无回答插补数值。依据插补值数量，插补法分为单重和多重插补方法。单重插补法仅给出无回答的一个插补值，单重插补法的优势是具有比较多的实践结果，能够应用完全的数据进行分析，其缺点主要是把缺失的数据当做已知的值，从而没有反应无回答的变异性。多重插补法是一种利用可能分布中两个或更多个值来替代缺失值的方法。多重插补法给出无回答的多个插补值。相比较而言，多重插补法能够给出总体参数估计量的方差估计，应用也更广泛（金勇进和邵军，2009）。

如今处理无回答的多重插补法已经有很多种，主要有EM、PMM、DA以及线性回归等多重插补法。本文主要研究普通线性回归多重插补法、贝叶斯线性回归多重插补法、贝叶斯自助线性回归多重插补法的统计性质。由于无回答的真值很难获得，本文采用随机模拟方法，研究这三种多重插补法在不同无回答机制、不同插补重数及不同无回答率下的统计性质，为三种方法的实际应用提供指导。

1.2 国内外研究综述

无回答在抽样调查中经常出现，从而对抽样调查的结果造成了一系列不利影响，它使得数据处理变得复杂，分析效率降低。这对统计分析来说，是数据结果处理常遇的难题。为了解决此类问题，国内外的统计学者研究出了许多解决无回答的方法。处理无回答的方法主要有事前预防和事后补救两大类，但由于事前预防受到各种因素的限制，随着调查研究的深入，这种方法并不能解决所遇到的某些问题。而事后补救的方法出现，很大程度上改善了无回答问题的处理结果，提高了统计分析的效率。目前主要的调查后补救方法主要有热卡、冷卡、均值、随机等单一插补法；同样也包括 PMM(Predictive Mean Matching)、DA(Data Augmentation)、MCMC(Markov Chain Monte Carlo)、EMB 等多重插补法。

1.2.1 无回答处理方法研究综述

国外学术界对无回答问题的研究较早，统计学家 Bowley 在 1915 年最早提出了无回答，该问题的提出使当时的许多研究者注意到造成抽样调查结果偏差的原因是无回答。在这个阶段，Bowley 阐述了调查结果的不确定性，以及导致结果产生误差的因素，并主要关注了怎样控制误差产生的问题。其他研究者也对无回答进行了一些基础的探究，同时强调了无回答在抽样调查结果中的重要性。随后，统计学家们对无回答问题进行了不同的研究，得出了回归、加权、均值、冷、热卡插补法等一系列处理无回答的单一插补方法。

20 世纪 30 年代，统计学家 Yates.F 没有直接删除分析结果中出现的无回答，同时对其进行填补，从而得到了较好的结果。Deming(1940)对抽样调查结果采用了分层多变量加权的插补方法来解决无回答问题；Nordbotten(1963)采用冷卡法进行了定期的实际调查，并探究了该方法的特点和作用；Chapman(1976)采用热卡插补法对涉及周期性的调查进行了研究，总结了此方法对调查结果产生的作用；Sande(1979,1982)通过热卡插补法，研究得出了距离函数匹配法，以此弥补了回归插补法存在的不足。

20 世纪 70 年代，国外学术界出现了一些新的处理无回答的方法，具有代表性的研究是 Laird、Dempster 和 Rubin(1977)首次提出了 EM 算法，并基于此种算法提出了用模拟的思路实现对缺失值的多重插补；Rubin(1978)又通过贝叶斯理论，提出了多重插补的思想；Herzog 和 Rubin(1983)提出，因为多重插补方法考虑了区间估计和显著性水平的变异性，这使得多重插补的结果比单一插补法得到的结果好；Rubin(1988)对多重和单重插补法做了较为详细的比较。尽管 Rubin 在 1977 至 1978 年间提出了多重插补的思想，但因为计算机的应用在当时不是很普遍，加上插补法的计算过程相当复杂，所以并没有得到广泛地应用。

20 世纪 90 年代，关于多重插补的研究主要有：Meng(1995)、Rubin(1996)和 Schafer(1997)对多重插补模型与分析模型的关系进行了探究；Maren K. Olsen(1998)提出了多元变量缺失值的多重插补法，并把该方法运用于实践当中；Ulrich Rendtel(2006)对有关无回答和多重插补方法的问题进行了总结，Rubin(2006)对其中的一些工作进行了深入的研究，阐述了多重插补法的有效区间估计等内容。随着统计软件的普遍应用，使多重插补法成为处理无回答数据的最重要的工具。在此段时期，主要的研究有：Mingxiu Hu，Sameena Salvucci，Stanley Weng(1996)对 Proc Impute、Schafer 等多重插补软件的特性进行了比较；Schafer(1997)设计了 NORM、CAT、MIX 和 PAN 等处理无回答的软件，并出版了相关著作；Yang，Yuan 和 Rockville(2000)探讨了多重插补在 SAS 软件上的实现。

对于缺失值的处理，除了对诸多方法研究之外，国外学术界也进行了广泛地应用。Thomas N. Herzog 和 Rubin 基于多重插补法对美国人口普查局的数据进行了研究，以此来衡量多重插补的效果；Rubin，N.Schenkeer(1991)对工业和职业普查中的无回答进行了多重插补；Barnard，J.和 Meng(1999)对艾滋病的调查数据使用了多重插补的方法；Niels Smits，Don Mellenbergh 和 Harrie Vorst(2002)应用多重插补法对学生课业成绩进行处理，并比较了 EM 和 DA 插补法的插补效果；Fiona M Shrive，Heather Stuart(2006)运用不同的插补法对 SDS(曾氏自评抑郁量表)的插补结果进行比较。国外学术界关于无回答的研究和应用已经很广泛，延伸到了各种不同的行业，形成了较为成熟的体系。

相比较而言，我国对无回答的研究较少，起步比较晚，相应的研究成果也不是很多，与国外的诸多研究相比，还有较大的差距。国内学术界对无回答处理的研究总体情况是在处理方法的理论研究方面很少有原创性的成果，主要是对国外提出的理论方法进行改进和总结，对于此类方法的应用也比较少。在经济迅速发展，国际往来越来越频繁的背景下，种类繁多的抽样调查以及经济普查工作在不断增加，传统的处理无回答的方法已经不能很好地解决更多的无回答问题，因此，对于无回答处理方法的深入研究有着举足轻重的意义。目前国内对无回答处理的研究主要有：金勇进(1998)提出了替代法、加权法等处理无回答的方法，比较了两种方法之间本质的差异和处理问题的效率，介绍了辅助信息对于减小估计量偏差的作用，并且将无回答造成的偏差表现出来；赵明德、谢邦昌(1999)在其著作中介绍了常用的处理无回答的多重插补法；王璐，王飞(2006)对无回答的处理进行了简要概述，介绍了随机抽样情况下的热卡插补方法，并采用蒙特卡洛方法对相关的方差问题进行研究，得出了此类问题的主要证据；庞新生(2005)对单一和多重插补方法进行了比较，深入分析了后者的特性以及处理无回答的基本思想；杨军、赵宇、丁文兴(2008)讨论了单重插补法对方差进行估计以及多重插补法进行的简要计算，并介绍了单重和多重插补一些重要的方法，对插补所面临的问题和发展方向进行了论述；谢邦昌、方匡南(2011)将聚类和关联规则用于无回答的处理，用聚类的方法对不完全数据集进行记录并归类，再采用调整的关联规则深入研究数据集中变量之间的相关性，然后以此来对无回答进行插补。经过验证，这种方法对先验信息缺失情况下的数据处理，具备良好的作用和效果。

1.2.2 线性回归多重插补法研究综述

目前，主要的多重插补法有PMM、DA、EMB、线性回归多重插补法等。其中，杨贵军、李静华(2014)通过PMM方法先找出无回答的渐近分布，抽取其随机数作为其插补值，对线性模型进行参数估计，对该方法的优良性做出评价和参考。杨贵军、骆新珍(2014)对DA多重插补法构造参数估计的迭代算法，将参数估计趋于稳定时的插补值用于估计无回答。杨贵军(2016)探究了EMB多重插补法的插补过程，该方法先采用 Bootstrap 法从样本中抽取数据，再用 EM 算法对不完整数据集进行插补，并给出未知参数的估计值等。这三种插补法的缺点是不考虑变量之间是否存在线性关系，插补值的随机性大。

线性回归模型主要用于研究不同变量的线性相关性，被广泛应用于很多复杂经济关系的研究。Rubin(1987)、Kalton G.(1983)论述了线性回归多重插补法利用具有无回答的变量

与其它变量之间的线性相关关系，得出插补值。依据插补值的残差随机性，线性回归多重插补法分为三类。第一类称为普通线性回归多重插补法，插补结果的残差服从均值为零的正态分布。线性回归模型较好估计了无回答的平均水平，残差随机性能够有效降低多个插补值之间的相关性。然而，普通线性回归多重插补法仅利用了总体信息与样本信息，并没有利用已有的先验信息。第二类称为贝叶斯线性回归多重插补法，以普通线性回归多重插补法为基础，利用先验信息改善回归模型参数估计和插补值残差的随机性（Little and Rubin,2002；Stef van Buuren,2012）。第三类称为贝叶斯自助线性回归多重插补法，在插补过程中采用自助法抽取样本，再估计回归模型参数（Efron,1979；Stef van Buuren,2012）。目前，普通线性回归多重插补法、贝叶斯线性回归多重插补法、贝叶斯自助线性回归多重插补法的应用较多，但它们的统计性质研究少。

总的来说，对于无回答及处理等问题的探究，国外学术界的研究成果广泛、丰富，形成了一个较为完善的体系。国内学术界在无回答方面的研究成果也呈现出逐渐增加的趋势。计算机应用的普及，提高了无回答处理的效率，这使得处理方法的应用及改进有了很大的提升空间。无回答问题在社会各领域中尤为重要，对此类问题的妥当处理更得到国内外学术界的重视。

1.3 论文思路及结构安排

1.3.1 本文研究思路

本文采用普通线性回归多重插补法、贝叶斯线性回归多重插补法、自助线性回归多重插补法，基于不同的缺失率、插补重数等条件对无回答集进行模拟研究，得出三种方法特性的比较。由于在经济领域的研究中线性回归模型应用广泛，因此文中设定三种线性回归多重插补法主要的研究对象为线性回归模型，通过对回归模型变量前系数的估算，来研究三种方法插补的优良性。模拟中设定完全随机、随机、非随机等无回答机制，在每种机制下分别使用三种线性回归多重插补法，基于五种不同的无回答率和插补重数，对设定的回归模型进行插补，并分析模拟所得参数分析结果，讨论在不同无回答机制、无回答率和插补重数下估计量的偏差及均方误差，尝试寻找三种方法在经济实践问题中的应用价值。

在模拟时，主要采用三种无回答机制，无回答率产生响应变量的插补值。其中缺失机

制分别为完全随机无回答机制、依赖不同解释变量的随机无回答机制、非随机无回答机制。应用不同的无回答机制是为了探究含有缺失值的变量的插补效果，并计算参数的估计值来验证插补的效果。模拟中采取的无回答率分别为5%，10%，20%，30%，40%，使用这个范围的无回答率是为了较为全面地分析不同无回答率下所得插补值的效果。没有采用更高的缺失率是因为当无回答率超过了50%，数据则不具备较好的使用价值，从而导致插补方法失效。整个模拟过程中，插补重数也是影响分析结果的重要因素，本文采用的插补重数分别为5，10，20，30，40。Rubin(1987)在研究了多重插补的效率问题之后提出，插补重数为5次时能够得到较好的插补值以及参数估计结果，而插补重数的增加并没有明显改善插补值和参数估计值。没有选择更大的插补重数是参考Gramham(2012)对生物数据的研究，作者提出插补重数为20至40次时得到的结果是合适的。若取更大的插补重数，所得到的插补值的精度所能提高的空间很小，因此选择的插补重数的最大值为40。

本文应用三种线性回归多重插补法对模拟数据进行研究，选取偏差绝对值和均方误差作为插补结果的评价指标，基于以往的经验，将模拟过程重复200次，可以得到较为准确的插补结果，因此设定模拟的次数为200次。

1.3.2 本文结构安排

结合上文的研究思路，论文的结构安排如下：

第1章 导论介绍了无回答的选题背景和意义，详细概述了国内外学术界对无回答和处理无回答方法的研究情况，介绍了三种线性回归多重插补法的研究现状，并介绍了论文的研究思路和(创新之处)。

第2章 介绍了什么是无回答，从无回答的产生原因，缺失机制，处理无回答常用的方法等方面入手进行论述，并主要论述了本文所采用的普通线性回归多重插补法，贝叶斯线性回归多重插补法、自助线性回归多重插补法的基本理论。

第3章 采用三种线性回归多重插补法对文中设定的线性回归模型进行模拟研究，通过不同缺失机制、缺失率、插补重数等条件对模型的参数进行模拟，并计算其估计值的偏差和均方误差，并对这两个指标的偏度和峰度进行测度，得出其统计分布的性质，由此分析三种方法的插补结果。

第4章小结 综述本文的模拟分析结果，并给出相关建议。

1.4 本文的创新之处

本文主要采用普通线性回归多重插补法、贝叶斯线性回归多重插补法、贝叶斯自助线性回归多重插补法对线性回归模型产生的数据集进行模拟比较研究，通过测度回归系数估计量的偏差和均方误差以及它们的偏度和峰度来分析比较三种方法的特性。目前普遍使用的多重插补法主要有PMM、DA、EMB等多重插补法，国内外学术界对这些方法的研究较多。文章通过使用三种线性回归多重插补法对回归模型做研究分析之后，在每一小结部分与前文提到的三种多重插补法的性质进行了比较分析。而对于本文采用的三种线性回归多重插补法，在国内外的研究很少。本文的创新之处主要是：

在采用三种方法进行模拟研究时，将三种多重插补法设置在相同的R语言环境中同时运行，这样使得这三种方法的计算出的回归系数的偏差和均方误差有了很好的可比性，并同时与其他常用的多重插补法作比较分析，更能显示出本文所用三种线性回归多重插补法的特性。

第 2 章　无回答及其处理方法概述

2.1 无回答产生的原因

数据缺失使得已有的数据集失去了一些非常有利于分析结果的信息，比如在调查所收集的问卷中包含不能使用的信息，或者在一些有关个人及团体隐私的调查中，接受调查的人对问卷的一些问题，甚至对整个问卷不做回答等。在这样的情况下，只是根据已有的数据信息进行分析，则不能反映实际的情况。因此，无回答的处理成为了实际调查后续工作中较为主要的问题。

产生无回答的因素较多，涉及众多领域。许多与社会科学相关的试验，都可能出现意外的情况，而在统计调查中，无回答现象非常普遍，这种现象会给调查结果的分析造成比较严重的影响，因此国内外学术界普遍重视这类问题的。在实际调查中的无回答主要有：调查中不能够使用或包含无回答的数据。

统计调查工作中不能使用的数据信息是指类似于数据登记错误，数据录入时出现的增添或遗漏，调查数据结果严重不符合预期等错误的信息。这些信息通常会使调查分析结果受到严重的影响。为了减少这类信息，工作人员会在检查时将此类信息删除，从而也导致了无回答的发生。

实际调查中，无回答主要是指在调查时出现的接受调查的人没有对问卷进行完整的填写，甚至还有拒绝填写问卷的情况。在这种情况下，无回答主要包括两种：项目无回答和单元无回答。项目无回答是指接受调查的人没有填写完整调查中的某个或某些项，从而产生的无回答。造成项目无回答的主要原因有：问卷的某些问题涉及被调查者的隐私而拒绝作答，调查人员由于不细心遗漏了某些问题等；单元无回答是指工作人员没有能够和被调查者取得联系，或者被调查者不愿意接受调查，亦或是被调查者因为某些原因没能够接受调查所造成的无回答现象。

综上所述，无回答的来源主要有：实际调查中不可的用信息、数据收集结果中含有的无回答以及出现的错误信息等。在无回答的处理过程中，对这些空缺信息进行分类是很重要的。这关系到如何使用适当的方法来更好地处理缺失的数据，从而提高统计分析的效率，

改善分析结果。

2.2 无回答的产生机制和无回答模式

2.2.1 无回答的产生机制

无回答的产生机制由Rubin(1976)提出，他阐述了无回答机制的定义，将其主要分为三类，包括：完全随机无回答、随机无回答、非随机无回答。研究无回答机制的主要目的是尝试说明无回答的本质，无回答的产生机制主要指的是无回答值和变量值之间的关系。除此之外，无回答数据的无回答机制还包括：基于辅助变量的无回答、基于随机影响的无回答、基于前期数据的无回答。

1.完全随机无回答

完全随机无回答是指无回答值既和观测到的值无关，也和未观测到的值无关。用数学表达式描述，即对目标变量Y，有

$$P(R|X,Y)=P(R|X,Y_{obs},Y_{mis})=P(R) \tag{2.1}$$

成立，称目标变量Y为完全随机无回答。其中Y指调查者感兴趣的目标变量，Y_{obs}为目标变量中观测到的变量，Y_{mis}为目标变量中未观测到的变量，R表示目标变量是否缺失的指示变量。

式(2.1)主要说明了变量Y的数据与含有无回答值的数据相同的分布。完全随机无回答假定较强，其无回答信息与变量没有相关性，因此对无回答信息进行评价变得较为困难。此类无回答在实际调查中比较少见，比如在有关个人收入的调查中，回答者和无回答者之间收入的平均值差异很小。

2.随机无回答

随机无回答是指目标变量Y的无回答值与已知的Y_{obs}有关，与未知的Y_{mis}无关。用数学表达式，即：若

$$P(R|X,Y)=P(R|X,Y_{obs},Y_{mis})=P(R|X,Y_{obs}) \tag{2.2}$$

对全部Y_{mis}都成立，则称目标变量Y为随机无回答。

随机无回答机制假定的条件与完全随机无回答的相比较弱，此类无回答较为多见，比如在不同人群的身体健康检查中，如果某些检测指标超出正常范围，被检查者就必须入院

治疗，是否进入医院治疗是由已经观测到的数据决定的，而不会受到未观测到的数据影响。

3.非随机无回答

非随机无回答是指目标变量Y的无回答情况只与Y本身有关，不受Y_{obs}的影响。或者说，如果$P(R|X,Y)$与Y_{mis}有关，则称目标变量Y为非随机缺失。

与前两种无回答机制相比，非随机无回答机制的假定条件最弱，其只依赖于无回答，不能被忽略，因此，非随机无回答机制也称为不可忽略的无回答机制。例如，实际的个人收入调查中，收入较高的被调查者往往不愿意透露自己的收入状况，从而导致调查数据缺失，这样的数据类型属于非随机缺失，主要是因为此类无回答处理之后得到的分析结果的偏差会比较大，只有通过收集足够多的数据才可能减小非随机无回答所造成的偏差。

2.2.2 无回答缺失模式

在无回答处理过程中，除了需要考虑无回答机制外，无回答的缺失模式也是一个重要的指标。无回答的缺失模式主要研究的是在已有的数据中得出哪些是观测到的数据，哪些是未观测到的数据。明确地说，其注重的是指示变量R的分布。这能够让研究人员了解同一数据集各变量之间的关系，从而为无回答的处理给出有用的建议。

设目标变量Y的矩阵是由m个观测值、n个变量组成的，其维度是$m\times n$。对该矩阵进行分析，可以得到数据的缺失模式。

1.单变量缺失模式。在该模式下无回答只在单个变量中出现，最长见的是在种植业中存在的无回答。其中Y_i代表农作物产量，可以将温度、光照、降雨量等影响因子设定为试验中的变量，并分别用$Y_1,Y_2,\cdots Y_{i-1}$代替。在这些变量中，Y_i存在无回答，其余变量都可以完全观测到。

2.多变量缺失模式。多变量缺失模式是指在目标变量Y的矩阵中，从变量Y_j起始，之后的变量缺失的项目都和变量Y_j相同，而变量Y_j之前的所有变量都可以被完全观测。在实际调查中最常见的情况是对问卷全部回答，或者是对整个问卷不回答。

3.单调缺失模式。该模式下对目标变量Y的矩阵进行适当的调整，使其表现为一种缺失值逐层递增的模式，也就是说，矩阵Y中的某个变量Y_j出现缺失值时，对任意的变量$Y_p(p\geq j)$，其缺失值的数目比Y_j的要多。实际调查中常见的情况是，对于固定群体的调查，由于时间的延长，导致一些调查单元逐渐丢失，从而造成了单调缺失的模式。

4.一般缺失模式。该模式是指目标变量矩阵中，数据发生缺失比较偶然，一般不能分析得出该类数据的缺失规律。调查结果中最为多见的情形是接受调查的人没有填写问卷中的某些项目而造成的项目无回答。

2.3 处理无回答常用的方法

2.3.1 直接删除法

为了较好地处理无回答问题，国内外学者们研究出了很多方法，直接删除法就是其中之一。直接删除法主要是根据已经收集到的数据，忽略缺失值的影响，对删除后的数据进行分析。其主要包括：列表删除法和成对删除法。列表删除法是指在已有的数据集中，对含有缺失值的单元进行剔除，只分析含有完整数据信息的单元。该方法是一种较为简单的处理无回答的方法。其特点是在处理无回答时过程简洁，使用的都是真实信息。在完全随机缺失机制下，用此法处理无回答的结果是较好的。但这种方法的缺点是，在含有无回答的单元数目较多的情况下，能够被分析的完整数据单元就很少，如此会丢失很多有价值的信息，从而不能得到理想的分析结果。直接删除法主要应用于包含较少无回答结果的分析中。成对删除法是指利用已有数据集中已观测到的数据信息对其进行处理。这种方法也称作有效单位分析法，与之前的列表删除法相比，该方法不会对收集的数据信息造成浪费。其不足之处主要是在不同的缺失模式下，每个变量含有的样本信息是不断变化的，当数据集中重要变量缺失的信息较多，且无回答机制不是完全随机无回答机制时，该方法就不是很适用。上述两种方法虽然对无回答的处理有一定的效果，但其分析结果的偏差还是较大，需要进行调整。

2.3.2 加权调整法

加权调整法是指在处理含有无回答的单元时，把这些单元的权重分配到其他不含无回答的单元，也就是说，在分析中提高含有完整回答单元的权重，以此降低无回答造成的误差。加权调整法可以应用在含有无回答的单元，还能够应用到含有项目无回答的情况当中，相比较而言，前者应用居多。在实际的抽样中每个单元都有对应的权重，其作用主要是增加重要变量所占的比重，所有的权重通过抽样设计决定，权重之和即为总体的规模。因此

用该方法分析含有缺失信息的数据集时，无回答的数目是一个重要的因素。

采用加权调整法之后的权重主要包括基础权重和调整因子，前者符合抽样设计的要求，后者是回答概率的倒数。加权调整法中回答概率是其最重要的内容，这是因为回答概率的估计方法不同，所得到的加权调整方法也会不同。该方法所包含的分类较多，也有比较广泛的应用。加权调整法主要的特点是在无回答偏差不存在时，使用此方法则需要足够的无回答数据，从而得到的分析结果的偏差较小。但是在单元无回答和项目无回答同时存在的情况下，此方法不再适用。

2.3.3 插补法

插补法是处理缺失值常用的方法之一，其主要的原理是对数据集中每个缺失值给出替代值，从而得到新的完整数据集之后，再对其进行参数估计。一般情况下，虽然插补后的数据集与真实数据集有差异，但插补法对于无回答处理是必要的，是因为：首先，研究某个领域的总体需要完整的数据集。完整的数据集可以给调查者提供比较全面的信息，从而避免了无回答给分析结果带来的偏倚。用插补方法得到的数据集包含调查者收集数据采用的知识，这可以得出良好的插补结果，从而减少数据缺失带来的不便。其次，对无回答实施插补，能够让调查人员使用完全数据分析方法，简化了数据分析过程中的计算，提高了数据处理的效率。第三，使用插补法可以减小分析结果的偏差。在无回答较多的情况下，直接对数据集进行分析，可能会对分析结果造成较大的偏差，影响最终的决策。经过插补得出的分析结果，其偏差比插补前有了明显地改善，这对分析决策有很大的帮助。

对于插补法的研究，是国内外学者们一直重点关注的问题，研究成果也是层出不穷。插补法的发展历史起始于单一插补，到目前已经有了很多较为优良的多重插补方法，而且不断应用于实际的调查当中，其中的一些方法也是越来越符合无回答需要的处理模式，这给无回答的处理带来了很大的帮助。

1.均值插补

均值插补是指用数据集中已观测到数据的平均值来代替无回答，是一种比较简单方便的方法。主要分为以下两种方法：

（1）单一均值插补是指对数据当中包含的无回答，采用所有已知数据的平均值运算插补。在此种情况下，整个数据集只有一个用于替代的均值，因此称为单一均值插补。单一均值插补的过程较为单调，得到的插补值也过于集中，这可能会给分析结果带来不利的影响。如果有可使用的辅助信息，可以改善插补结果。

（2）分层均值插补是指先对所有的数据进行分层，使得每层数据单元的信息比较类似，之后用每层已观测到值的平均值对所在层的无回答运算插补。这个方法是对单一均值插补的精细化，改善了结果的准确性。但该方法的会低估参数的方差估计值，比较适用于简单描述性研究的情况。

2.演绎插补

演绎插补法是采用经济或人口普查数据、目前调查中的相关项目等已收集的辅助信息，对缺失的数据进行逻辑推理，进而得到插补值的一种方法。用数学公式可表述为 $W_i = f(X_i)$。其中，w_i 为无回答的插补值，X_i 为辅助变量，$f(*)$ 是含有运算关系的函数。该方法的插补过程较为简单，如果能得到很好的辅助信息，演绎插补就可以通过这些信息得出准确度较高的插补值。这种插补方法常用于数据信息的审核当中，比如，在处理时删除不合逻辑的数据，然后将插补值填入已删除数据所在的位置，进行插补后分析。

当插补值和实际值相等时，采用该方法计算所得估计量为

$$\hat{\bar{Y}} = (\sum_{i=1}^{n_1} y_i + \sum_{i=1}^{n_0} w_i)/n = \sum_{i=1}^{n} y_i/n$$

其中 n_1 为有回答项的数目，n_0 为未回答项的数目。由上式可以知道，当辅助信息准确的时，演绎插补得到的估计值是没有偏差的。

3.最近距离插补

最近距离插补是通过目标变量 Y 与辅助变量的距离是否接近而选择需要赋值的项，也就是说，通过辅助信息，设定不同预测单元的距离的函数，在与无回答项相邻的有回答项中，选取与辅助变量相对应的变量 Y 中已观测到的数据则作最终的插补值，这些插补值同样也满足设定的距离条件。

设目标变量中无回答值为 y_i，辅助变量为 x_i。则插补值的计算公式为 $w_i = f(x_i) = y_i$，其中 w_i 是第 i 个缺失单元的插补值，$f(*)$ 为赋值函数，该函数返回满足 $d(i,j) = \min d(i,i')$ 的 X_j 对应的值 y_i。$d(i,i')$ 为两个单元之间的距离函数。

在该方法中，根据辅助变量的数目的差异，距离函数的种类也不相同：

a.只采用一个辅助变量 X_1 设定距离函数，则数学表达式为：$d_1 = \left|x_{1i} - x_{1j}\right|$

b.采用多个辅助变量设定距离函数，其表达式为：$d_2 = Sup_k I_k \left|x_{ki} - x_{kj}\right|$，$I_k$ 表示第 k 个变量的相对重要程度。

c.马氏距离：$d_3 = (x_i - x_j)^T S_{xx}^{-1}(x_i - x_j)$，其中，$S_{xx}^{-1}$ 是 x_i 的协方差阵的估计矩阵。

不同的距离函数类型得到的插补值不同，具有伪随机性质，这增加了最近距离插补估计量的估计难度，也给该领域的研究带来了挑战(金勇进、邵军，2009)。

4.回归插补

回归插补法是在目标变量以及辅助变量之间设定函数模型，利用无回答对已观测单元的回归得出预测值，然后对无回答进行插补。其主要的思想是在响应变量和辅助变量之间建立回归模型，通过计算得出回归模型的系数估计。比较常见的情形是线性回归模型的插补。该方法的一般操作步骤是：首先选择与预测无回答有关的辅助变量，其次设定回归模型，之后再用无回答的条件期望进行插补。在响应变量和辅助变量之间的关系较为显著的情况下，采用该方法得到的插补结果和真实数值差异较小。

5.热卡插补

热卡插补是指在同一调查的有回答的数据中寻找与缺失值接近的值，并用该近似值对无回答进行插补。“热卡”一词来源于打孔记录计算机程序和数据的时代，在数据记录完成后，存储数据的卡片还是热的。关于热卡插补法，虽然目前还没有一致明确的定义，但国内外诸多文献中提到的方法，基本都是从同一个调查中采用某种方式为无回答确定插补值。因此，这一类方法都可以称为热卡插补法。

采用该方法从目标变量Y中按一定的概率抽取插补值，在这种情况下，总体均值估计可以表示为

$$\hat{\bar{Y}}=\frac{1}{n}\sum_{i=1}^{n}[a_i y_i+(1-a_i)\bar{y}^*]$$

其中，$\bar{y}^*$为插补值的均值。

热卡插补法可以分为随机、分层、序贯等热卡插补方法。随机热卡插补是指对目标变量Y的已观测单元进行有放回的简单随机抽样，从而获得插补值。与均值插补相比，该方法得到的插补值是随机的，不会使目标变量Y的样本分布产生较大偏差。该方法较为普遍地被应用于实际调查中。分层热卡插补是指采用调查中的辅助信息进行分层，得到具备较多性质相似的项，之后对每个层的数据运算插补。分层热卡插补利用辅助信息，使得缺失值和插补值更为接近，从而提高了插补的效率。序贯热卡插补首先确定数据插补的类型，并对数据进行分层，然后再对每层的数据单元进行排序，对于有缺失值的单元，将该层中最后一个被读取的数据当做插补值进行插补。热卡插补法在实践调查中应用广泛，与其他的单一插补法相比，该方法能够较好地保持变量的经验分布和利用辅助信息，插补的到的结果较好。

与热卡插补法相对应，另一种比较常用的方法是冷卡法。冷卡法是指先从以往的调查结果中获取插补值，然后对无回答集进行插补。冷卡插补主要分为冷卡替代和比率替代。冷卡替代的优点主要是在一定条件下经过替代插补得到的估计量的方差与目标变量的方差一致，这提高了插补的效率。比率替代则是较为充分地利用了辅助信息，其方差估计具有与冷卡替代的方差估计同样的性质。

6.双重稳健插补

双重稳健插补法与双重稳健加权法较为接近，两者都是通过充分利用辅助信息来构造估计量，从而得到稳定有效的插补值。该方法的主要思想是：首先在无回答的基础上，调整已观测值和缺失值的权数得到无偏或接近无偏的估计量；其次，通过调整权数改善估计结果，从而减少估计的偏差。

多重插补是对单一插补法的延伸，并以其为依据发展而来的。这个概念首先在1977年提出，该方法的主要思想是给每个无回答值构造多个插补值，如此就能产生多个完整数据集，采用相同的方法对完整数据集进行处理，并得到多个处理结果，最后对处理结果综合得出设定变量的估计。多重插补法很好的补充了单重插补法的缺点，其中包括：多重插补能够运用插补过程中得到的插补值的变异性反映出无回答值的不确定性；多重插补能够在运算过程中得出无回答值的分布，从而更好地保留了不同变量之间的相关性；多重插补能够反映很多与分析结果不确定性有关的信息。多重插补的主要问题是构造完全数据集时工作量较大。

多重插补法的提出，主要应用了贝叶斯理论。该理论的主要思想是某个变量的后验分布可以通过该变量的先验分布以及不同变量之间的联合分布得出。假设已观测到的数据集为 $Y=y(y_1,y_2,\cdots y_n)'$，这些数据集取自于参数的总体，参数 η 可以是向量，Y 的密度是 $p(Y,\eta)$。贝叶斯理论把 η 也当做是具有特定分布的变量，所以把 $p(Y,\eta)$ 看作已知 η 的条件概率密度。只要知道 η 的边缘概率密度 $\pi(\eta)$，就可以得出 Y 和 η 的联合概率密度，即

$$(\eta,Y)\sim\pi(\eta)p(Y,\eta)$$

因此 η 对于 Y 的条件概率密度为

$$h(\eta|Y)=\frac{\pi(\eta)p(Y,\eta)}{\int\pi(\eta)p(Y,\eta)d\eta}$$

其中 $\pi(\eta)$ 为先验分布，包含着人们在抽样前对 η 的认知程度，它具有与参数 η 相关的信息；$h(\eta|Y)$ 为后验分布，该分布反映了人们在抽样之后对 η 的认知程度，前者和后者的

主要区别在于后者可以被看作是用总体和样本信息对前者作了调整的结果，其整合了先验以及样本所包含的信息，所以分析结果都主要服从后验分布。根据这些理论，多重插补可以描述为：在每一次插补过程中，对未观测到的Y值，通过推导其后验分布，并重复m次之后构成缺失值Y的m次插补，每次重复对应于独立的参数和缺失值，这样可以得到m个完整数据集，之后对这些数据集进行分析，进而得出η的估计值和相应的方差估计，并将这些估计结果合并分析，最后得到对目标变量的弱偏估计。综上所述，多重插补法大致包含三步，其中包括：对目标参数的估计，通过插补值构造完整数据集，多重插补的统计推断等步骤(庞新生，2013)。

处理无回答的多重插补法在首次被提出时并没有受到广泛地关注，主要是因为其计算过程较为复杂。到了上世纪90年代，电脑和各类分析软件的广泛应用，才使得多重插补法渐渐成为解决无回答的主要方法。虽然目前多重插补方法在实际中解决了很多问题，但是某些实际的调查数据并不完全符合多重插补法的假定，所以，在复杂情况下还需要对处理无回答的多重插补方法做进一步的研究。

7.EM算法

EM算法的主要思想是通过计算参数极大似然估计或计算后验估计期望值的最大值来得到插补值。EM算法假定潜变量是存在的。这在很大程度上降低了似然函数的复杂性，并很好地处理了求解过程中遇到的问题。EM算法主要包含两个过程，首先是E步，此步骤是通过已观测到的值以及参数θ来预测无回答；其次是M步，该步骤则是通过E步计算出的完整数据集估计出一个新的参数θ'；重复以上的步骤，最终得到收敛的参数估计值，并将其预测值作为插补值。该方法采用无回答和模型参数来相互推导，如利用已经得到的无回答的插补值得出模型的参数估计，或者是用已知的参数估计值来推导无回答的估计值。因此，该方法主要是利用多次迭代的过程使估计值收敛，然后通过计算极大似然估计值得到插补值。这种方法的主要优点是其计算过程比较简单，而且较为稳健。而其不足主要有：在处理不同情况下的数据时，需要建立不同的模型以及编写不同的程序；当数据的无回答率比较高时，会影响到该算法的运算效果等。EM算法的问世，使得处理无回答的插补方法有了重要的发展。

8.MCMC方法

MCMC方法是一种通过贝叶斯统计理论来推导不完全数据集的后验分布的方法。该方法最初使用在观测分子无规则运动的试验中。实际应用中，MCMC方法通过DA方法计算数据的复杂分布来实现多重插补，并且由一系列方法组合而成。MCMC方法的基本思想是：

先通过计算得到Markov链的样本数据，之后对Markov链使用Monte Carlo积分得到插补结果，其主要的插补过程包括插补步和后验步。该方法的主要特点是在运算过程中提高了算法的收敛速度，其不足主要表现在运算过程复杂繁琐、需要多元正态的假定条件以及不能确定其是否收敛等方面。MCMC方法最初由Metropolis、Hastings等人对energy distribution的间接模拟算法的研究发展而来，此后German、Tanner和Wong等人对Gibbs抽样以及DA算法处理无回答数据的研究，使得MCMC方法有了更进一步的发展。随着SAS、R等分析软件的发展，MCMC方法也渐渐成为一种重要的处理无回答值的多重插补方法。

9.PMM方法

PMM多重插补法主要由Little在1988年提出，这种多重插补法适用于分析变量是连续型变量且单调缺失的情况，该方法依据Rubin的应用统计文件匹配理论为基础建立，它通过对不完全数据集的插补，计算出的插补值与预测值最为近似。这种插补法是以完整的数据集为基础，用相应变量进行回归，之后通过回归模型得到插补值，并对数据集进行插补。PMM多重插补法的优点是：在插补的过程中采用的插补值不是确定的数值，而是实际观测值，如此使得插补值与观测值更近似，从而具有较好的准确性。

综上所述，无论是加权调整或直接删除法还是插补法，处理无回答的方法相比较以前有了很多方面的改进。但现有的处理无回答的插补法中，还是有许多不足之处。其中，多数单一插补法的运算过程较为简单，得出的插补值相比于多重插补法过于集中，这会使分析结果可能存在较大的误差。而在上述的EM、MCMC、PMM等多重插补法中，计算得到的插补值比单重插补法算出的插补值更为可靠，但这些方法得出的插补值所计算的系数的偏差和均方误差比较大，从而增大了分析结果的误差。而本文采用的三种线性回归多重插补法，除了具备上述多重插补法的优良特性以外，该类方法所得到的系数的偏差及均方误差都要更小，在一定程度上降低了分析结果的误差。

2.4 线性回归多重插补法

线性回归多重插补法利用不含无回答的观测数据，建立线性回归模型，再利用拟合模型得到无回答的若干个插补值（Rubin,1977）。线性回归多重插补法给出的插补值与无回答的真值之间偏差相对较小，统计分析结论的可信度较高。

本文仅考虑响应变量无回答的情况。不失一般性，响应变量的 n 个观测值用

$y^T=(y_R^T,y_N^T)$表示，其中y_R为n_R个完整观测数据，y_N为n_N个无回答，y^T表示y的转置。$n_R+n_N=n$，解释变量的观测值也划分为两部分，X_R是相应于y_R的q个解释变量的观测数据矩阵，X_N是相应于y_N的q个解释变量的观测数据矩阵，X_R与X_N均不含无回答。

2.4.1 普通线性回归多重插补法

普通线性回归多重插补法的具体运算过程如下。

（1）建立响应变量y的拟合模型。线性回归模型为$y=X\beta+\varepsilon$，ε为回归模型的随机误差，其方差记为σ^2。利用观测数据y_R和X_R，采用普通最小二乘估计法，回归系数β的估计为$\hat{\beta}_R=(X_R^TX_R)^{-1}X_R^Ty_R$，随机误差的方差估计为$\hat{\sigma}_R^2=(y_R-X_R\hat{\beta}_R)^T(y_R-X_R\hat{\beta}_R)/(n_R-q)$。

（2）计算无回答的插补值。依据拟合模型$\hat{y}=X\hat{\beta}_R$，利用观测数据X_N估计无回答，得无回答y_N的拟合模型$y_N=X_N\hat{\beta}_R$。于是，无回答y_N的L个插补值分别为$\hat{y}_{N,i}=X_N\hat{\beta}_R+e_i$，$i=1,\cdots,L$，$L$为插补重数。其中，$e_i$是随机生成的残差项。普通线性回归多重插补法选择的$e_i$服从均值为零的正态分布（Stef van Buuren，2012），其方差为$\hat{\sigma}_R^2=(y_R-X_R\hat{\beta}_R)^T(y_R-X_R\hat{\beta}_R)/(n_R-q)$。

相对于PMM、DA和EMB多重插补法，普通线性回归多重插补法更好保持响应变量的分布规律，及其与因变量之间的线性相关性（Little and Rubin，2002）。

2.4.2 贝叶斯线性回归多重插补法

Rubin（1987）利用先验信息，提出贝叶斯线性回归多重插补法。该方法采用了正态线性模型，利用标准非信息的先验分布调整线性模型参数估计，并给出无回答的插补值（Stef van Buuren，2012）。贝叶斯线性回归多重插补法的主要步骤如下。

（1）建立响应变量y的拟合模型。首先，利用观测数据得到$S_R=X_R^TX_R$，记S_{DR}为对角矩阵，由矩阵S_R的对角元素构成。$V_R=(S_R+\kappa S_{DR})^{-1}$，$\kappa$是岭回归参数，一般选取接近于0的正数，以避免出现奇异矩阵的情况。记系数估计为$\hat{\beta}_R=V_RX_R^Ty_R$。其次，利用先验信息修正回归系数。记$\xi$服从自由度为$n_R-q$的$\chi^2$分布。随机误差的方差估计为$\hat{\sigma}_{BR}^2=(y_R-X_R\hat{\beta}_R)^T(y_R-X_R\hat{\beta}_R)/\xi$，回归系数估计为$\hat{\beta}_{BR}=\hat{\beta}_R+V_R^{1/2}\hat{\sigma}_{BR}\eta_1$，$\eta_1$为$q$维随机向

量，服从q维标准正态分布。$V_R^{1/2}$是由对矩阵V_R的Cholesky分解得到。

（2）计算无回答的插补值。利用观测数据X_N估计无回答，得无回答y_N的拟合模型$y_N = X_N\hat{\beta}_{BR}$。于是，无回答$y_N$的插补值分别为$\hat{y}_{N,i} = X_N\hat{\beta}_{BR} + \eta_i\hat{\sigma}_{BR}$，其中，$i = 1,\cdots,L$，$\eta_i$是服从$n_N$维标准正态分布的向量。

与普通线性回归多重插补法相比，贝叶斯线性回归多重插补法利用先验信息修正了回归模型参数估计，以改善插补值。

2.4.3 贝叶斯自助线性回归多重插补法

贝叶斯自助线性回归多重插补法兼具自助法与贝叶斯线性回归多重插补法的特点（Stef van Buuren，2012），先从已观测数据中抽取自助样本，采用正态线性模型，利用标准非信息的先验分布调整线性模型参数估计，并给出无回答的插补值。自助法是重复抽取样本的方法，参见Efron and Tibshirani（1993）。贝叶斯自助线性回归多重插补法的主要运算步骤如下：

（1）建立响应变量y的拟合模型。首先，从已观测数据(y_R, X_R)中有放回的抽取样本量为n_R的自助样本，记为(y_{RB}, X_{RB})。其次，经过计算得到矩阵$S_{RB} = X_{RB}^T X_{RB}$，$V_{RB} = (S_{RB} + \kappa S_{DRB})^{-1}$，其中$S_{DRB}$表示由矩阵$S_{RB}$的对角元素构成的对角矩阵，$\kappa$是岭回归参数。利用样本估计回归系数为$\hat{\beta}_{RB} = V_{RB}X_{RB}^T y_{RB}$。

（2）计算无回答的插补值。利用观测数据X_N估计无回答，得无回答y_N的拟合模型$y_N = X_N\hat{\beta}_{RB}$。于是，无回答$y_N$的插补值分别为$\hat{y} = X\hat{\beta}_{RB} + \eta_2\hat{\sigma}_{RB}$，其中，$\hat{\sigma}_{RB}^2 = (y_{RB} - X_{RB}\hat{\beta}_{RB})^T(y_{RB} - X_{RB}\hat{\beta}_{RB})/(n_R - q - 1)$，$\eta_2$是服从$n_N$维标准正态分布的向量。重复$L$次步骤（1）和（2），得到$L$个插补值。

贝叶斯线性回归多重插补法与贝叶斯自助线性回归多重插补法相比，两种方法都充分利用先验信息，调整回归系数的估计，保持了无回答变量与其解释变量的线性相关关系。前者直接使用了全部样本数据进行构造插补值，后者采用自助法从样本中抽取数据构造插补值，有利于插补值更好地保持随机性。后者避免了Cholesky分解，且不需要从χ^2分布中抽样，计算相对简化。

第 3 章　三种线性回归多重插补法的模拟比较研究

3.1 模拟研究设计

出于与其他线性回归多重插补法对比的角度，本文在模拟研究过程中借鉴了杨贵军和李静华（2014）的模拟步骤。在模拟过程中，用于生成数据集的线性模型设置为 $Y=\beta_0+\beta_1X_1+\beta_2X_2+\beta_3X_3+\varepsilon$，模型中的系数 $\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2,\hat{\beta}_3$ 的值分别为 1、1、2、3，随机误差项 $\varepsilon\sim N(0,1)$。对于自变量 X_1、X_2、X_3 以及随机误差项 ε，设定其分别服从正太分布 $N(8,4),N(4,8),N(4,4),N(0,1)$，并从中产生 100 个随机数作为自变量以及误差项的观测值，之后根据设定的线性模型计算出因变量 Y 的 100 个观测值 y。在不同的无回答率（5%,10%,20%,30%）和不同的无回答机制下，产生含有无回答的不完全数据集，记为 $y=(y_R,y_N)$，y_R 为观测数据，y_N 为无回答。在完全随机无回答、随机无回答、非随机无回答三种机制下，同时采用三种线性回归多重插补法对数据进行插补，计算出 L 组关于无回答的插补值，并可以计算出包含插补值的完全数据集，以此计算回归系数的估计量。将上面的步骤重复计算 200 次，则得到 200 组回归系数的估计值，计算回归系数估计值的偏差和均方误差作为三种多重插补方法的评价指标。对于第 $k=1,2,\cdots 200$ 次模拟，假定 $\hat{\beta}_{j,k}^{(i)}$ 代表采用数据集 $y_{N,k}^{(i)}$ 和 $y_{R,k}^{(i)}$ 线性模型系数 β_j 的估计值，其中 $y_{N,k}^{(i)}$ 代表采用回归多重插补法的第 i 重插补值，$i=1,2,\cdots m$ 代表第 i 重插补序号，$j=0,1,2,3$ 代表模型参数的下标。估计量 $\hat{\beta}_{j,k}^{(i)}$（$i=1,2,\cdots m$）的均值为 $\hat{\beta}_{j,k}=\sum_{i=1}^{m}\hat{\beta}_{j,k}^{(i)}/m$，相应的方差 $B_{j,k}=\sum_{i=1}^{m}(\hat{\beta}_{j,k}^{(i)}-\hat{\beta}_{j,k})^2/(m-1)$。$U_{j,k}^{(i)}$（$i=1,2,\cdots m$）的均值为 $\overline{U}_{j,k}=\sum_{i=1}^{m}U_{j,k}^{(i)}/m$。$\hat{\beta}_j$ 的偏差为 $B(\hat{\beta}_j)=\sum_{k=1}^{200}(\hat{\beta}_{j,k}-\beta_j)/200$，它是估计值和设定值的差值计算的平均数。$\hat{\beta}_j$ 的均方误差为 $MSE(\hat{\beta}_j)=\sum_{k=1}^{200}(\hat{\beta}_{j,k}-\beta_j)^2/200$，而均方误差是估计值与设定值差值平方的平均值。偏差及均方误差越小，插补法越优良。

此外，为了更好地研究三种线性回归多重插补法所得系数估计量的统计学性质，本文针对 $\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2,\hat{\beta}_3$ 的偏差和均方误差进行了与数据分布相关的偏度以及峰度的分析研究。其中，偏度是测度数据偏斜方向及偏斜程度的统计量，是对数据分布是否对称的数字特征。

3.2 完全随机无回答机制下的模拟结果

完全随机无回答机制是指无回答与响应变量和解释变量都无关的一种无回答机制。从 y 中采用不放回等概率地随机抽取 Y 的观测值，将这些已经取出的观测值定为无回答，取出的 Y 的值的数目占所有数据的比例则为无回答率。分别采用三种插补方法，对 Y 的数据集进行插补。

（1）偏差分析。在完全随机无回答机制下，表 3.1 给出三种多重插补法的回归系数估计的偏差。表 3.1 的第 2-5 列、6-9 列、10-13 列分别表示采用普通线性回归、贝叶斯线性回归、贝叶斯自助线性回归多重插补法得到的偏差。表 3.1 的列对应不同无回答率（5%,10%,20%,30%），行对应不同插补重数（5,10,20,30,40），每个格子里的四个数值自上而下依次为 $\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2,\hat{\beta}_3$ 的偏差。

表 3.1　完全随机无回答机制下 $\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2,\hat{\beta}_3$ 的偏差

重数	5%	10%	20%	30%	5%	10%	20%	30%	5%	10%	20%	30%
5	1.614	1.640	1.687	1.714	1.606	1.652	1.666	1.710	1.610	1.633	1.667	1.708
	0.003	0.001	-0.005	-0.002	0.004	0.000	-0.004	-0.002	0.003	0.002	-0.004	-0.002
	-0.004	-0.005	-0.003	-0.005	-0.004	-0.005	-0.003	-0.004	-0.004	-0.005	-0.004	-0.005
	0.001	0.002	0.002	-0.002	0.002	0.001	0.003	-0.002	0.002	0.001	0.002	-0.002
10	1.607	1.648	1.665	1.706	1.606	1.644	1.667	1.707	1.604	1.634	1.672	1.693
	0.003	0.001	-0.002	-0.001	0.004	0.002	-0.003	-0.002	0.004	0.001	-0.004	-0.001
	-0.004	-0.005	-0.004	-0.006	-0.004	-0.006	-0.004	-0.005	-0.004	-0.005	-0.004	-0.005
	0.003	0.001	0.001	-0.002	0.002	0.001	0.002	-0.002	0.002	0.002	0.002	-0.002
20	1.610	1.643	1.662	1.700	1.603	1.648	1.668	1.700	1.605	1.647	1.675	1.700
	0.003	0.001	-0.003	-0.001	0.004	0.001	-0.003	-0.001	0.003	0.001	-0.004	-0.001
	-0.004	-0.005	-0.004	-0.005	-0.004	-0.005	-0.003	-0.005	-0.004	-0.005	-0.004	-0.005
	0.002	0.001	0.003	-0.002	0.003	0.001	0.002	-0.001	0.002	0.001	0.002	-0.002
30	1.611	1.653	1.665	1.712	1.608	1.643	1.668	1.709	1.613	1.647	1.664	1.713
	0.003	0.001	-0.003	-0.002	0.004	0.001	-0.004	-0.002	0.003	0.001	-0.003	-0.002
	-0.004	-0.005	-0.004	-0.005	-0.004	-0.005	-0.004	-0.005	-0.004	-0.005	-0.003	-0.005
	0.002	0.001	0.003	-0.002	0.002	0.001	0.003	-0.002	0.002	0.001	0.003	-0.002
40	1.609	1.644	1.669	1.702	1.610	1.642	1.663	1.702	1.607	1.644	1.661	1.706
	0.003	0.001	-0.003	-0.002	0.003	0.001	-0.004	-0.002	0.003	0.001	-0.003	-0.002
	-0.004	-0.005	-0.004	-0.005	-0.004	-0.005	-0.003	-0.005	-0.004	-0.005	-0.004	-0.006
	0.002	0.001	0.002	-0.001	0.002	0.001	0.003	-0.001	0.002	0.001	0.003	-0.002

表 3.1 显示，在完全随机无回答机制下，对于相同无回答率和相同插补重数，采用普通线性回归、贝叶斯线性回归、贝叶斯自助线性回归，$\hat{\beta}_1,\hat{\beta}_2,\hat{\beta}_3$ 的偏差的绝对值都小，基本上都小于 0.006，$\hat{\beta}_0$ 的偏差的绝对值相对远大于其他系数的偏差的绝对值。在三种多重插补法下，当插补重数相同时，$\hat{\beta}_0$ 的偏差的绝对值随着无回答率的增加呈现递增趋势，$\hat{\beta}_1,\hat{\beta}_2,\hat{\beta}_3$ 则没有明显的变化趋势。当无回答率相同时，随着插补重数增加，$\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2,\hat{\beta}_3$ 的偏差绝对值没有递减趋势。

表 3.2　完全随机无回答机制下 $\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2,\hat{\beta}_3$ 的均方误差

重数	5%	10%	20%	30%	5%	10%	20%	30%	5%	10%	20%	30%
5	3.043	3.200	3.453	3.605	3.012	3.242	3.390	3.588	3.027	3.149	3.448	3.629
	0.003	0.003	0.004	0.005	0.003	0.003	0.004	0.004	0.003	0.003	0.004	0.004
	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
	0.003	0.004	0.004	0.004	0.003	0.003	0.004	0.004	0.003	0.003	0.004	0.004
10	3.032	3.205	3.398	3.564	3.015	3.213	3.406	3.553	3.008	3.172	3.410	3.531
	0.003	0.003	0.004	0.004	0.003	0.003	0.004	0.004	0.003	0.003	0.004	0.004
	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
	0.003	0.003	0.004	0.004	0.003	0.003	0.004	0.004	0.003	0.003	0.004	0.004
20	3.031	3.199	3.388	3.538	3.013	3.223	3.403	3.527	3.008	3.214	3.443	3.509
	0.003	0.003	0.004	0.004	0.003	0.003	0.004	0.004	0.003	0.003	0.004	0.004
	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
	0.003	0.003	0.004	0.004	0.003	0.003	0.004	0.004	0.003	0.003	0.004	0.004
30	3.033	3.235	3.380	3.569	3.016	3.198	3.406	3.561	3.039	3.215	3.392	3.576
	0.003	0.003	0.004	0.004	0.003	0.003	0.004	0.004	0.003	0.003	0.004	0.004
	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
	0.003	0.003	0.004	0.004	0.003	0.003	0.004	0.004	0.003	0.003	0.004	0.004
40	3.027	3.199	3.408	3.572	3.027	3.198	3.384	3.536	3.023	3.189	3.369	3.585
	0.003	0.003	0.004	0.004	0.003	0.003	0.004	0.004	0.003	0.003	0.004	0.004
	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
	0.003	0.003	0.004	0.004	0.003	0.003	0.004	0.004	0.003	0.003	0.004	0.004

（2）均方误差分析。表 3.2 给出在完全随机无回答机制下，分别采用三种多重插补法，$\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2,\hat{\beta}_3$ 的均方误差。表 3.2 的第 2-5 列、6-9 列、10-13 列分别表示采用普通线性回归、贝叶斯线性回归、贝叶斯自助线性回归多重插补法得到的均方误差。列对应不同无回答率（5%,10%,20%,40%），行对应不同插补重数（5,10,20,30,40），每个格子里的四个数值自上而下依次为 $\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2,\hat{\beta}_3$ 的均方误差。

表 3.2 显示，在完全随机无回答机制下，分别采用普通线性回归、贝叶斯线性回归、

贝叶斯自助线性回归多重插补法得到的 $\hat{\beta}_1,\hat{\beta}_2,\hat{\beta}_3$ 的均方误差均较小，均小于 0.005。$\hat{\beta}_0$ 的均方误差明显大于 $\hat{\beta}_1,\hat{\beta}_2,\hat{\beta}_3$ 的均方误差，均大于 3。在同时采用三种多重插补法的情况下，当插补重数相同时，随着无回答率增加，$\hat{\beta}_0$ 的均方误差呈现递增趋势，$\hat{\beta}_1,\hat{\beta}_2,\hat{\beta}_3$ 的均方误差则没有明显的变化。当无回答率相同时，随着插补重数增加，$\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2,\hat{\beta}_3$ 的均方误差没有明显减小。

（3）关于偏差和均方误差的偏度、峰度分析。表 3.3 给出在完全随机无回答机制下，分别采用三种多重插补法对 $\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2,\hat{\beta}_3$ 的偏差进行测度,从而得到偏差的偏度和峰度。表 3.3 中 3-7 行、8-12 行、13-17 行分别表示普通线性回归、贝叶斯线性回归、贝叶斯自助线性回归多重插补法在不同插补重数，不同无回答率下得到的 $\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2,\hat{\beta}_3$ 偏差的偏度和峰度。

表 3.3　完全随机无回答机制下系数偏差的峰度和偏度

方法	β_0		β_1		β_2		β_3	
	偏度	峰度	偏度	峰度	偏度	峰度	偏度	峰度
普通线性回归	0.490	-1.830	-0.104	-2.146	-0.546	-1.854	0.715	-1.711
	0.118	-2.302	-0.034	-1.878	-0.317	-1.986	0.197	-2.208
	-0.096	-1.962	-0.033	-2.092	-0.099	-2.292	-0.479	-1.884
	-0.143	-2.145	-0.013	-1.989	-0.669	-1.747	-0.373	-1.883
	0.194	-2.110	-0.015	-2.023	-0.683	-1.732	0.447	-1.820
贝叶斯线性回归	-0.123	-2.132	0.253	-2.134	-0.698	-1.726	0.479	-1.843
	0.064	-2.028	0.081	-2.245	-0.729	-1.702	0.420	-1.830
	0.080	-2.243	-0.051	-1.968	-0.562	-1.845	0.118	-1.883
	0.154	-2.066	0.073	-2.096	-0.640	-1.775	0.191	-2.026
	-0.005	-1.977	-0.044	-2.151	-0.611	-1.787	-0.324	-2.079
贝叶斯自助线性回归	0.444	-1.958	-0.262	-2.129	-0.706	-1.717	0.402	-1.827
	0.404	-1.941	0.104	-2.101	0.232	-1.863	0.744	-1.692
	0.196	-2.208	-0.051	-1.927	-0.594	-1.792	0.602	-1.798
	0.075	-2.288	-0.086	-1.968	-0.577	-1.828	0.382	-1.933
	0.103	-2.247	-0.200	-1.985	-0.651	-1.763	-0.168	-1.969

表 3.3 显示，在完全随机无回答机制下，在分别采用普通线性回归、贝叶斯线性回归、贝叶斯自助线性回归多重插补法得到的系数偏差的偏度和峰度值当中，$\hat{\beta}_1$ 偏差的偏度在不同的插补重数下较小，更接近正太分布；$\hat{\beta}_0,\hat{\beta}_2,\hat{\beta}_3$ 偏差的偏度相对较大，有不同程度的左偏和右偏。从各系数偏差的峰度来看，在使用三种多重插补法时各系数偏差的峰度值均小于 0，与正态分布相比为较平缓的平顶分布，平缓程度与正态分布的差异较大，且数据分

布不集中。

表 3.4　完全随机无回答机制下系数均方误差的峰度和偏度

方法	β_0		β_1		β_2		β_3	
	偏度	峰度	偏度	峰度	偏度	峰度	偏度	峰度
普通线性回归	-0.177	-2.157	0.493	-1.906	0.472	-1.857	0.219	-2.067
	-0.166	-1.936	0.377	-2.013	0.478	-1.861	0.282	-2.080
	-0.293	-2.067	0.410	-1.969	0.493	-1.852	0.137	-2.221
	-0.442	-1.936	0.332	-2.048	0.352	-1.937	0.275	-2.099
	-0.235	-2.039	0.374	-2.013	0.479	-1.834	0.075	-2.203
贝叶斯线性回归	-0.476	-1.923	0.422	-1.952	0.475	-1.863	0.263	-2.048
	-0.295	-1.957	0.382	-2.000	0.562	-1.805	-0.007	-2.269
	-0.273	-1.892	0.321	-2.068	0.515	-1.833	0.337	-1.980
	-0.219	-1.984	0.366	-2.027	0.445	-1.859	0.148	-2.131
	-0.292	-2.033	0.350	-2.036	0.383	-1.921	0.154	-2.174
贝叶斯自助线性回归	0.189	-1.988	0.261	-2.143	0.302	-1.979	0.177	-2.139
	-0.143	-2.052	0.385	-1.999	0.509	-1.813	0.334	-2.023
	0.097	-1.873	0.206	-2.171	0.298	-1.933	0.191	-2.093
	-0.060	-1.879	0.378	-2.002	0.511	-1.826	0.088	-2.253
	-0.271	-1.995	0.409	-1.988	0.447	-1.852	0.228	-2.143

表 3.4 显示，完全随机无回答机制下分别采用普通线性回归、贝叶斯线性回归、贝叶斯自助线性回归多重插补法计算的系数均方误差的偏度和峰度。其中，各系数均方误差的偏度的绝对值主要在 0.1-0.5 之间，与正态分布相比偏斜程度略大，$\hat{\beta}_0$ 在三种方法下均方误差的偏斜程度主要为左偏；其他系数均方误差的偏斜程度主要为右偏。从各系数均方误差的峰度来看，$\hat{\beta}_3$ 均方误差的峰度绝对值相对较大，其他系数均方误差的峰度的绝对值主要在 1.8-2.2 之间，说明各系数均方误差分布的平缓程度与正态分布差异较大，数据的值比较分散。

在这种无回答机制下，对应于插补重数的递增，本文分别采用三种插补法的系数估计的偏差和均方误差均没有出现递减趋势。在完全随机无回答机制下，与 PMM、DA、EMB 多重插补法相比，本文分别采用这三种多重插补方法，$\hat{\beta}_1,\hat{\beta}_2,\hat{\beta}_3$ 的偏差绝对值更小，都远小于采用 PMM 或 DA 多重插补法情况下的偏差绝对值，与采用 EMB 多重插补法情况下的偏差绝对值差异较小；$\hat{\beta}_1,\hat{\beta}_2,\hat{\beta}_3$ 的均方误差远远小于采用 PMM 或 DA 多重插补法情况下的均方误差，与采用 EMB 多重插补法的均方误差差异较小。

3.3 随机无回答机制下的模拟结果

在随机无回答机制下，无回答与解释变量有关。下面主要给出依赖变量X_1、X_2、X_3的随机无回答机制的模拟结果。根据不同的无回答的概率，确定变量X_1的分位数，将x_{1i}小于这个分位数的观测值$(y_i, x_{1i}, x_{2i}, x_{3i})$中的$y_i$设定为无回答。分别采用三种多重插补法，对$Y$的无回答进行插补，再估计模型系数。

3.3.1 依赖变量X_1的随机无回答机制下的模拟分析结果

（1）依赖变量X_1的偏差分析。表3.5给出在这种无回答机制下，分别采用普通线性回归、贝叶斯线性回归、贝叶斯自助线性回归多重插补法$\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3$的偏差。表3.5结构同表3.1。表3.5显示，分别采用三种多重插补法，$\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3$的偏差很小，小于0.011，$\hat{\beta}_0$的偏差最大。当插补重数相同时，对应于无回答率的递增，，$\hat{\beta}_0$的偏差绝对值有较小递增趋势，$\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3$则没有明显的变化趋势。当无回答率相同时，随着插补重数增加，$\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3$的偏差绝对值没有明显递减趋势。

表3.5 依赖变量X_1随机无回答机制下$\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3$的偏差

重数	5%	10%	20%	30%	5%	10%	20%	30%	5%	10%	20%	30%
5	1.639	1.642	1.701	1.728	1.637	1.674	1.647	1.758	1.654	1.698	1.746	1.810
	0.005	0.003	-0.002	-0.004	0.004	0.000	0.003	-0.007	0.004	-0.001	-0.006	-0.011
	0.002	0.002	0.002	0.003	0.002	0.002	0.003	0.002	0.001	0.002	0.002	0.003
	-0.006	-0.005	-0.006	-0.006	-0.005	-0.005	-0.006	-0.005	-0.006	-0.006	-0.006	-0.007
10	1.648	1.654	1.691	1.737	1.649	1.684	1.683	1.721	1.655	1.680	1.645	1.735
	0.003	0.003	-0.002	-0.006	0.003	0.000	0.000	-0.004	0.002	0.000	0.002	-0.004
	0.002	0.002	0.002	0.003	0.002	0.002	0.003	0.003	0.002	0.002	0.003	0.003
	-0.006	-0.006	-0.004	-0.005	-0.005	-0.006	-0.007	-0.005	-0.005	-0.006	-0.003	-0.006
20	1.660	1.674	1.681	1.728	1.653	1.692	1.687	1.736	1.651	1.692	1.668	1.725
	0.002	0.001	-0.001	-0.005	0.003	0.000	0.000	-0.006	0.003	-0.001	0.001	-0.004
	0.002	0.002	0.003	0.002	0.002	0.002	0.002	0.003	0.002	0.002	0.003	0.002
	-0.005	-0.006	-0.005	-0.005	-0.006	-0.007	-0.006	-0.004	-0.006	-0.006	-0.005	-0.005

30	1.638	1.664	1.701	1.733	1.646	1.688	1.676	1.732	1.652	1.683	1.691	1.750
	0.004	0.002	-0.001	-0.005	0.003	0.000	0.000	-0.005	0.003	0.000	-0.001	-0.007
	0.002	0.002	0.002	0.003	0.002	0.002	0.003	0.002	0.002	0.002	0.003	0.003
	-0.005	-0.006	-0.006	-0.006	-0.005	-0.006	-0.006	-0.005	-0.006	-0.006	-0.006	-0.005
40	1.651	1.668	1.690	1.720	1.646	1.692	1.681	1.733	1.645	1.702	1.664	1.732
	0.003	0.002	-0.001	-0.004	0.003	-0.001	-0.001	-0.005	0.003	-0.001	0.001	-0.005
	0.002	0.002	0.003	0.003	0.002	0.002	0.003	0.002	0.002	0.002	0.003	0.003
	-0.005	-0.006	-0.005	-0.005	-0.005	-0.006	-0.005	-0.005	-0.005	-0.006	-0.005	-0.005

（2）依赖变量X_1的均方误差分析。表 3.6 给出在这种无回答机制下，分别采用普通线性回归、贝叶斯线性回归、贝叶斯自助线性回归多重插补法，$\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2,\hat{\beta}_3$的均方误差。表 3.6 结构同表 3.2。由表 3.6 可知，分别采用三种多重插补法，$\hat{\beta}_1,\hat{\beta}_2,\hat{\beta}_3$的均方误差很小，均小于 0.01。$\hat{\beta}_0$的均方误差明显大于$\hat{\beta}_1,\hat{\beta}_2,\hat{\beta}_3$的均方误差。当插补重数相同时，随着无回答率增加，$\hat{\beta}_0$的均方误差呈现递增趋势，$\hat{\beta}_1,\hat{\beta}_2,\hat{\beta}_3$则没有明显变化。当无回答率相同时，随着插补重数增加，$\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2,\hat{\beta}_3$的均方误差没有减小。

表 3.6　依赖变量X_1随机无回答机制下$\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2,\hat{\beta}_3$的均方误差

重数	5%	10%	20%	30%	5%	10%	20%	30%	5%	10%	20%	30%
5	3.366	3.441	3.939	4.330	3.358	3.430	3.847	4.115	3.421	3.477	4.008	4.702
	0.004	0.005	0.008	0.010	0.004	0.005	0.007	0.009	0.004	0.005	0.007	0.010
	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
	0.003	0.003	0.004	0.004	0.003	0.003	0.004	0.004	0.003	0.003	0.004	0.005
10	3.367	3.519	3.793	4.269	3.417	3.466	3.921	4.216	3.331	3.490	3.894	4.030
	0.004	0.005	0.007	0.009	0.004	0.005	0.007	0.009	0.004	0.005	0.007	0.009
	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
	0.003	0.003	0.004	0.004	0.003	0.003	0.004	0.004	0.003	0.003	0.004	0.004
20	3.388	3.490	3.849	4.261	3.347	3.477	3.908	4.265	3.381	3.477	3.916	4.199
	0.004	0.005	0.007	0.009	0.004	0.005	0.007	0.009	0.004	0.005	0.007	0.009
	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
	0.003	0.003	0.004	0.004	0.003	0.003	0.004	0.004	0.003	0.003	0.004	0.004
30	3.387	3.420	3.843	4.357	3.345	3.422	3.906	4.210	3.385	3.480	3.919	4.244
	0.004	0.005	0.007	0.010	0.004	0.005	0.007	0.009	0.004	0.005	0.007	0.009
	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
	0.003	0.003	0.004	0.004	0.003	0.003	0.004	0.004	0.003	0.003	0.004	0.004
40	3.366	3.456	3.886	4.164	3.363	3.453	3.935	4.190	3.375	3.432	3.978	4.090
	0.004	0.005	0.007	0.009	0.004	0.005	0.007	0.009	0.004	0.005	0.007	0.009
	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
	0.003	0.003	0.004	0.004	0.003	0.003	0.004	0.004	0.003	0.003	0.003	0.004

（3）依赖变量X_1的偏差和均方误差的偏度、峰度分析。表3.7给出在随机无回答机制下，分别采用三种多重插补法对$\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2,\hat{\beta}_3$的偏差进行测度。表中3-7行、8-12行、13-17行分别表示普通线性回归、贝叶斯线性回归、贝叶斯自助线性回归多重插补法在同一插补重数，不同无回答率下得到的$\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2,\hat{\beta}_3$偏差的偏度和峰度。

表3.7　依赖变量X_1随机无回答机制下系数偏差的偏度和峰度

方法	β_0		β_1		β_2		β_3	
	偏度	峰度	偏度	峰度	偏度	峰度	偏度	峰度
普通线性回归	0.703	-1.722	-0.711	-1.712	0.028	-2.011	0.076	-2.248
	0.677	-1.739	-0.733	-1.699	-0.330	-1.846	0.135	-1.951
	0.651	-1.766	-0.586	-1.775	-0.517	-1.805	-0.055	-2.123
	0.551	-1.781	-0.656	-1.748	-0.091	-1.917	0.132	-1.870
	0.645	-1.763	-0.653	-1.759	-0.101	-1.891	0.138	-2.109
贝叶斯线性回归	0.579	-1.798	-0.531	-1.874	-0.516	-1.830	-0.698	-1.719
	0.312	-1.951	-0.435	-1.960	-0.129	-1.936	-0.036	-2.132
	0.368	-2.035	-0.499	-1.899	-0.105	-1.876	0.283	-1.871
	0.343	-2.043	-0.398	-1.995	-0.526	-1.875	0.000	-2.437
	0.293	-2.090	-0.290	-2.114	-0.376	-1.836	-0.146	-2.017
贝叶斯自助线性回归	0.599	-1.810	-0.584	-1.825	0.352	-1.945	-0.436	-1.861
	0.530	-1.871	-0.366	-1.924	-0.120	-1.939	0.343	-2.016
	0.270	-2.120	-0.293	-2.099	-0.642	-1.754	0.062	-1.912
	0.574	-1.823	-0.592	-1.817	-0.342	-1.923	0.273	-2.000
	0.144	-2.213	-0.238	-2.171	0.019	-1.878	0.024	-2.388

表3.7显示，在随机无回答机制下，分别采用普通线性回归、贝叶斯线性回归、贝叶斯自助线性回归多重插补法得到的系数偏差的偏度和峰度。其中，$\hat{\beta}_1$和$\hat{\beta}_2$偏差的偏度多数为左偏、且$\hat{\beta}_1$偏差的偏斜程度略大一些；$\hat{\beta}_0$和$\hat{\beta}_3$偏差的偏度多为右偏，且$\hat{\beta}_0$偏差的偏度相对更大。从峰度上看，各系数偏差的峰度的绝对值主要介于1.7-2.3之间，因为是负值，所以相比于正态分布较为平缓，说明数据的分布比较分散。

表3.8　依赖变量X_1随机无回答机制下系数均方误差的偏度和峰度

方法	β_0		β_1		β_2		β_3	
	偏度	峰度	偏度	峰度	偏度	峰度	偏度	峰度
普通线性回归	0.501	-1.899	0.354	-1.999	0.669	-1.744	0.424	-1.939
	0.665	-1.748	0.575	-1.798	0.654	-1.762	0.385	-1.968
	0.617	-1.791	0.585	-1.805	0.627	-1.784	0.438	-1.957
	0.609	-1.802	0.555	-1.828	0.705	-1.719	0.474	-1.918

	0.540	-1.860	0.485	-1.888	0.694	-1.729	0.433	-1.936
贝叶斯线性回归	0.597	-1.811	0.563	-1.817	0.668	-1.752	0.463	-1.898
	0.553	-1.852	0.562	-1.823	0.689	-1.733	0.384	-1.993
	0.551	-1.846	0.553	-1.823	0.642	-1.761	0.434	-1.940
	0.539	-1.864	0.526	-1.858	0.654	-1.760	0.438	-1.947
	0.524	-1.876	0.547	-1.833	0.675	-1.743	0.377	-1.999
贝叶斯自助线性回归	0.583	-1.825	0.585	-1.802	0.661	-1.741	0.399	-1.969
	0.555	-1.838	0.551	-1.827	0.710	-1.717	0.385	-1.935
	0.589	-1.817	0.617	-1.780	0.667	-1.749	0.454	-1.930
	0.571	-1.832	0.571	-1.818	0.664	-1.750	0.440	-1.936
	0.473	-1.927	0.497	-1.880	0.705	-1.718	0.470	-1.908

表 3.8 显示，在随机无回答机制下，分别采用普通线性回归、贝叶斯线性回归、贝叶斯自助线性回归多重插补法得到的系数偏差的偏度和峰度。其中，各系数均方误差的偏值比较接近，而且偏斜程度比完全随机无回答机制下的大，$\hat{\beta}_2$ 均方误差的偏度最大。从峰度上看，在三种方法下各系数均方误差的峰度 $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 普遍比较接近、$\hat{\beta}_2$ 和 $\hat{\beta}_3$ 相差较多，且 $\hat{\beta}_3$ 均方误差的峰度绝对值最大，分布相对更为分散。

3.3.2 依赖变量 X_2 的随机无回答机制下的模拟分析结果

（1）依赖变量 X_2 的偏差分析。表 3.9 给出在这种无回答机制下，分别采用普通线性回归、贝叶斯线性回归、贝叶斯自助线性回归多重插补法，$\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2,\hat{\beta}_3$ 的偏差。表 3.9 结构同表 3.1。表 3.9 显示，分别采用三种多重插补法，$\hat{\beta}_1,\hat{\beta}_2,\hat{\beta}_3$ 的偏差较小，小于 0.008。当插补重数相同时，对应于无回答率的递增，$\hat{\beta}_0$ 的偏差最大，并有略微递增趋势，$\hat{\beta}_1,\hat{\beta}_2,\hat{\beta}_3$ 则没有明显的变化趋势。当无回答率相同时，随着插补重数增加，$\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2,\hat{\beta}_3$ 的偏差绝对值没有明显递减趋势。

表 3.9 依赖变量 X_2 随机无回答机制下 $\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2,\hat{\beta}_3$ 的偏差

重数	5%	10%	20%	30%	5%	10%	20%	30%	5%	10%	20%	30%
5	1.673	1.705	1.702	1.764	1.665	1.677	1.714	1.728	1.670	1.693	1.707	1.704
	0.001	0.000	0.002	0.006	0.001	0.002	0.004	0.006	0.001	0.001	0.003	0.004
	0.001	0.000	-0.003	-0.009	0.002	-0.001	-0.004	-0.006	0.001	-0.001	-0.004	-0.005
	-0.005	-0.006	-0.004	-0.009	-0.004	-0.004	-0.006	-0.007	-0.005	-0.005	-0.005	-0.005
10	1.670	1.684	1.744	1.769	1.671	1.680	1.708	1.723	1.675	1.681	1.711	1.743

	0.001	0.002	0.001	0.004	0.000	0.002	0.004	0.005	0.000	0.002	0.001	0.002
	0.001	0.001	-0.005	-0.008	0.001	0.000	-0.005	-0.006	0.000	-0.001	-0.003	-0.006
	-0.004	-0.006	-0.005	-0.008	-0.004	-0.005	-0.004	-0.006	-0.004	-0.005	-0.005	-0.007
20	1.662	1.682	1.728	1.733	1.661	1.671	1.719	1.739	1.672	1.676	1.716	1.734
	0.001	0.002	0.002	0.003	0.001	0.003	0.002	0.004	0.001	0.002	0.002	0.005
	0.001	0.000	-0.005	-0.005	0.001	0.000	-0.004	-0.006	0.000	0.000	-0.004	-0.007
	-0.004	-0.006	-0.005	-0.007	-0.004	-0.005	-0.005	-0.007	-0.004	-0.006	-0.005	-0.007
30	1.665	1.672	1.724	1.735	1.663	1.686	1.718	1.745	1.659	1.680	1.713	1.735
	0.001	0.002	0.003	0.004	0.001	0.002	0.003	0.004	0.001	0.002	0.003	0.003
	0.001	0.000	-0.004	-0.005	0.001	0.000	-0.004	-0.007	0.001	0.000	-0.004	-0.006
	-0.004	-0.005	-0.006	-0.008	-0.004	-0.006	-0.005	-0.007	-0.004	-0.006	-0.005	-0.006
40	1.669	1.674	1.720	1.735	1.663	1.683	1.721	1.731	1.664	1.685	1.704	1.738
	0.001	0.002	0.002	0.004	0.001	0.002	0.002	0.005	0.001	0.002	0.002	0.003
	0.001	0.000	-0.004	-0.006	0.001	0.000	-0.004	-0.006	0.001	0.000	-0.004	-0.006
	-0.004	-0.006	-0.005	-0.006	-0.004	-0.006	-0.005	-0.007	-0.004	-0.006	-0.004	-0.007

（2）依赖变量 X_2 的均方误差分析。表 3.10 给出在这种无回答机制下，分别采用普通线性回归、贝叶斯线性回归、贝叶斯自助线性回归多重插补法，$\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2,\hat{\beta}_3$ 的均方误差。表 3.10 结构同表 3.2。表 3.10 可知，分别采用三种多重插补法，$\hat{\beta}_1,\hat{\beta}_2,\hat{\beta}_3$ 的均方误差很小，均小于 0.006。$\hat{\beta}_0$ 的均方误差明显大于 $\hat{\beta}_1,\hat{\beta}_2,\hat{\beta}_3$ 的均方误差。当插补重数相同时，对应于无回答率的递增，$\hat{\beta}_0$ 的均方误差呈现递增趋势，$\hat{\beta}_1,\hat{\beta}_2,\hat{\beta}_3$ 则没有明显变化。当无回答率相同时，随着插补重数增加，$\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2,\hat{\beta}_3$ 的均方误差没有减小。

表 3.10　依赖变量 X_2 随机无回答机制下 $\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2,\hat{\beta}_3$ 的均方误差

重数	5%	10%	20%	30%	5%	10%	20%	30%	5%	10%	20%	30%
5	3.610	3.824	3.609	4.308	3.516	3.696	3.661	4.070	3.524	3.750	3.644	4.010
	0.003	0.004	0.004	0.005	0.003	0.003	0.004	0.005	0.004	0.004	0.004	0.005
	0.001	0.002	0.002	0.004	0.001	0.002	0.002	0.003	0.001	0.002	0.002	0.003
	0.004	0.004	0.004	0.006	0.004	0.004	0.005	0.006	0.004	0.004	0.005	0.005
10	3.548	3.950	3.783	4.296	3.573	3.755	3.624	3.973	3.615	3.851	3.629	4.050
	0.004	0.004	0.004	0.005	0.004	0.004	0.004	0.005	0.004	0.004	0.004	0.005
	0.001	0.002	0.002	0.003	0.001	0.001	0.002	0.003	0.001	0.002	0.002	0.003
	0.004	0.005	0.005	0.006	0.004	0.004	0.005	0.006	0.004	0.005	0.004	0.005
20	3.476	3.797	3.669	3.970	3.508	3.733	3.650	4.047	3.571	3.691	3.674	4.079
	0.004	0.004	0.004	0.005	0.004	0.004	0.004	0.005	0.003	0.004	0.004	0.005
	0.001	0.002	0.002	0.003	0.001	0.002	0.002	0.003	0.001	0.001	0.002	0.003
	0.003	0.005	0.004	0.005	0.003	0.004	0.004	0.006	0.004	0.004	0.004	0.006
30	3.525	3.709	3.666	4.057	3.529	3.785	3.649	4.032	3.492	3.692	3.628	4.027

	0.003	0.004	0.004	0.005	0.003	0.004	0.004	0.005	0.003	0.004	0.003	0.005
	0.001	0.002	0.002	0.003	0.001	0.002	0.002	0.003	0.001	0.002	0.002	0.003
	0.004	0.004	0.005	0.006	0.004	0.004	0.004	0.005	0.004	0.004	0.004	0.005
40	3.557	3.728	3.662	3.981	3.498	3.745	3.659	4.011	3.499	3.779	3.602	4.018
	0.003	0.004	0.003	0.005	0.003	0.004	0.003	0.005	0.003	0.004	0.003	0.005
	0.001	0.002	0.002	0.003	0.001	0.002	0.002	0.003	0.001	0.002	0.002	0.003
	0.004	0.005	0.005	0.005	0.004	0.004	0.004	0.006	0.004	0.005	0.004	0.006

（3）依赖变量 X_2 的偏差和均方误差的偏度、峰度分析。表 3.11 给出在这种无回答机制下，分别采用三种多重插补法对 $\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2,\hat{\beta}_3$ 的偏差进行测度。表中 3-7 行、8-12 行、13-17 行分别表示普通线性回归、贝叶斯线性回归、贝叶斯自助线性回归多重插补法在同一插补重数，不同无回答率下得到的 $\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2,\hat{\beta}_3$ 偏差的偏度和峰度。

表 3.11　依赖变量 X_2 随机无回答机制下系数偏差的偏度和峰度

方法	β_0		β_1		β_2		β_3	
	偏度	峰度	偏度	峰度	偏度	峰度	偏度	峰度
普通线性回归	0.372	-1.833	0.561	-1.846	-0.728	-1.701	-0.596	-1.766
	0.597	-1.785	0.344	-1.853	-0.743	-1.693	-0.082	-1.923
	0.225	-1.976	-0.053	-2.122	-0.665	-1.733	0.520	-1.792
	0.340	-2.035	0.292	-2.040	-0.658	-1.744	-0.283	-1.924
	0.348	-2.039	0.128	-2.067	-0.688	-1.732	0.489	-1.847
贝叶斯线性回归	0.237	-2.068	0.482	-1.890	-0.457	-1.841	-0.099	-2.215
	0.235	-2.092	0.456	-1.940	-0.688	-1.724	0.402	-1.955
	0.344	-2.015	0.148	-2.081	-0.626	-1.754	-0.183	-1.926
	0.304	-1.908	0.207	-1.887	-0.680	-1.727	0.123	-1.875
	0.235	-1.940	0.397	-1.935	-0.632	-1.749	0.387	-1.830
贝叶斯自助线性回归	-0.559	-1.838	0.630	-1.760	-0.463	-1.830	-0.620	-1.764
	0.498	-1.889	0.001	-2.330	-0.707	-1.714	0.022	-1.955
	0.321	-2.072	0.290	-1.956	-0.736	-1.698	0.166	-1.867
	0.222	-1.984	-0.130	-1.918	-0.676	-1.733	0.682	-1.736
	0.231	-1.970	0.016	-1.974	-0.692	-1.723	0.535	-1.806

表 3.11 显示，在随机无回答机制下，分别采用普通线性回归多重插补法、贝叶斯线性回归多重插补法、贝叶斯自助线性回归多重插补法得到的系数偏差的偏度和峰度。其中，$\hat{\beta}_1$ 和 $\hat{\beta}_1$ 偏差的偏度多数为右偏、且 $\hat{\beta}_2$ 偏差的偏斜程度略大一些，且多为左偏。从峰度上看，各系数偏差的峰度均小于 0，其绝对值主要介于 1.7-2.4 之间，所以系数的均方误差分布较为平缓，数据的分布比较分散。

表 3.12　依赖变量 X_2 随机无回答机制下系数均方误差的偏度和峰度

方法	β_0		β_1		β_2		β_3	
	偏度	峰度	偏度	峰度	偏度	峰度	偏度	峰度
普通线性回归	0.554	-1.852	0.127	-2.285	0.683	-1.730	0.485	-1.820
	0.285	-2.114	0.714	-1.711	0.440	-1.938	0.060	-2.355
	0.112	-2.160	0.476	-1.914	0.593	-1.808	0.209	-2.202
	0.530	-1.835	0.681	-1.739	0.611	-1.794	0.527	-1.869
	0.464	-1.915	0.627	-1.783	0.627	-1.775	0.137	-2.233
贝叶斯线性回归	0.559	-1.824	0.740	-1.695	0.564	-1.824	0.509	-1.888
	0.395	-1.988	0.357	-1.986	0.654	-1.760	0.571	-1.822
	0.435	-1.907	0.571	-1.798	0.538	-1.858	0.512	-1.892
	0.316	-2.030	0.597	-1.811	0.585	-1.808	0.273	-2.106
	0.352	-1.993	0.706	-1.720	0.566	-1.825	0.371	-2.030
贝叶斯自助线性回归	0.361	-1.947	0.624	-1.778	0.676	-1.743	0.356	-2.042
	0.252	-2.151	0.621	-1.792	0.445	-1.931	0.170	-2.173
	0.651	-1.759	0.602	-1.807	0.677	-1.737	0.556	-1.843
	0.482	-1.852	0.732	-1.700	0.628	-1.778	0.311	-2.073
	0.256	-2.039	0.605	-1.800	0.572	-1.828	0.282	-2.109

表 3.12 显示，在依赖变量 X_2 随机无回答机制下，分别采用普通线性回归多重插补法、贝叶斯线性回归多重插补法、贝叶斯自助线性回归多重插补法得到的系数均方误差的偏度和峰度。其中，各系数均方误差的偏度值均大于 0，偏斜程度为右偏，且偏度值大于完全随机无回答机制下的数值，$\hat{\beta}_1$、$\hat{\beta}_2$ 均方误差的偏度相对更大。从峰度上看，在三种方法下各系数均方误差的峰度 $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 普遍比较接近、$\hat{\beta}_2$ 和 $\hat{\beta}_3$ 相差较多，且 $\hat{\beta}_3$ 均方误差的峰度绝对值最大，数据分布相对较为分散。

3.3.3 依赖变量 X_3 的随机无回答机制下的模拟分析结果

（1）依赖变量 X_3 的偏差分析。表 3.13 给出在这种无回答机制下，分别采用普通线性回归、贝叶斯线性回归、贝叶斯自助线性回归多重插补法，$\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2,\hat{\beta}_3$ 的偏差。表 3.13 结构同表 3.1。表 3.13 显示，分别采用三种多重插补法，$\hat{\beta}_1,\hat{\beta}_2,\hat{\beta}_3$ 的偏差较小，小于 0.016。$\hat{\beta}_0$ 的偏差远大于其他系数的偏差。当插补重数相同时，对应于无回答率的递增，$\hat{\beta}_0$ 的偏差绝对值有略微递增的趋势，$\hat{\beta}_1,\hat{\beta}_2,\hat{\beta}_3$ 则没有明显的变化趋势。当无回答率相同时，随着

插补重数增加，$\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2,\hat{\beta}_3$的偏差绝对值没有明显递减趋势。

表 3.13　依赖变量X_3随机无回答机制下$\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2,\hat{\beta}_3$的偏差

重数	5%	10%	20%	30%	5%	10%	20%	30%	5%	10%	20%	30%
5	1.728	1.781	1.779	1.797	1.725	1.743	1.763	1.787	1.704	1.738	1.765	1.800
	0.001	-0.001	0.002	-0.003	0.000	-0.001	0.003	0.001	0.001	0.000	0.003	0.000
	-0.001	-0.002	-0.002	-0.002	-0.001	-0.001	-0.001	-0.002	0.000	-0.002	-0.001	-0.001
	-0.009	-0.013	-0.014	-0.012	-0.008	-0.009	-0.014	-0.015	-0.007	-0.010	-0.014	-0.015
10	1.717	1.760	1.732	1.807	1.717	1.749	1.760	1.797	1.731	1.766	1.755	1.776
	0.000	-0.001	0.003	-0.002	0.000	-0.001	0.002	-0.001	0.000	-0.001	0.001	0.000
	-0.001	-0.002	-0.002	-0.002	-0.001	-0.002	-0.002	-0.001	-0.001	-0.002	-0.001	-0.002
	-0.008	-0.011	-0.010	-0.014	-0.008	-0.010	-0.013	-0.014	-0.009	-0.011	-0.011	-0.013
20	1.716	1.745	1.744	1.763	1.727	1.747	1.746	1.796	1.724	1.762	1.729	1.780
	0.000	-0.001	0.002	-0.003	0.000	-0.001	0.003	0.000	0.000	-0.001	0.003	0.000
	-0.001	-0.002	-0.001	-0.001	-0.001	-0.002	-0.001	-0.002	-0.001	-0.002	-0.001	-0.001
	-0.008	-0.009	-0.011	-0.009	-0.009	-0.010	-0.012	-0.014	-0.008	-0.011	-0.011	-0.013
30	1.721	1.736	1.742	1.763	1.728	1.752	1.755	1.792	1.725	1.738	1.763	1.776
	0.000	0.000	0.002	-0.002	0.000	-0.001	0.002	0.000	0.000	0.000	0.002	0.000
	-0.001	-0.002	-0.001	-0.001	-0.001	-0.002	-0.001	-0.002	-0.001	-0.002	-0.001	-0.001
	-0.008	-0.009	-0.011	-0.010	-0.009	-0.010	-0.012	-0.014	-0.008	-0.009	-0.013	-0.013
40	1.720	1.753	1.742	1.785	1.723	1.743	1.756	1.795	1.720	1.753	1.750	1.784
	0.000	-0.001	0.002	-0.003	0.000	-0.001	0.002	0.000	0.000	-0.001	0.003	0.000
	-0.001	-0.002	-0.001	-0.001	-0.001	-0.002	-0.001	-0.002	-0.001	-0.002	-0.002	-0.001
	-0.008	-0.010	-0.011	-0.012	-0.008	-0.009	-0.013	-0.014	-0.008	-0.010	-0.012	-0.014

（2）依赖变量X_3的均方误差分析。表 3.14 给出在这种无回答机制下，分别采用普通线性回归、贝叶斯线性回归、贝叶斯自助线性回归多重插补法，$\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2,\hat{\beta}_3$的均方误差。表 3.14 结构同表 3.2。表 3.14 可知，分别采用三种多重插补法，$\hat{\beta}_1,\hat{\beta}_2,\hat{\beta}_3$的均方误差很小，均小于 0.02。$\hat{\beta}_0$的均方误差明显远大于$\hat{\beta}_1,\hat{\beta}_2,\hat{\beta}_3$的均方误差。当插补重数相同时，对应于无回答率的递增，$\hat{\beta}_0$的均方误差呈现递增趋势，$\hat{\beta}_1,\hat{\beta}_2,\hat{\beta}_3$则没有明显变化。当无回答率相同时，随着插补重数增加，$\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2,\hat{\beta}_3$的均方误差没有减小。

表 3.14 依赖变量 X_3 随机无回答机制下 $\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2,\hat{\beta}_3$ 的均方误差

重数	5%	10%	20%	30%	5%	10%	20%	30%	5%	10%	20%	30%
5	3.821	4.309	4.684	5.663	3.795	4.159	4.768	5.398	3.640	4.025	4.712	6.140
	0.004	0.004	0.004	0.005	0.004	0.004	0.004	0.005	0.003	0.004	0.004	0.005
	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
	0.005	0.007	0.011	0.018	0.004	0.006	0.011	0.016	0.004	0.006	0.011	0.019
10	3.813	4.137	5.007	5.610	3.813	4.070	4.700	5.446	3.906	4.211	4.688	5.377
	0.004	0.004	0.004	0.005	0.004	0.004	0.004	0.005	0.004	0.004	0.004	0.005
	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
	0.005	0.006	0.012	0.017	0.005	0.006	0.011	0.016	0.005	0.006	0.011	0.016
20	3.744	3.953	4.835	5.327	3.835	4.064	4.728	5.402	3.841	4.180	4.714	5.601
	0.004	0.004	0.004	0.005	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.005
	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
	0.004	0.005	0.012	0.016	0.005	0.006	0.011	0.016	0.005	0.006	0.011	0.017
30	3.828	4.002	4.717	5.376	3.830	4.107	4.710	5.451	3.876	4.016	4.712	5.442
	0.004	0.004	0.004	0.005	0.004	0.004	0.004	0.005	0.004	0.004	0.004	0.004
	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
	0.005	0.005	0.011	0.016	0.005	0.006	0.011	0.016	0.005	0.005	0.012	0.016
40	3.801	4.110	4.742	5.566	3.770	4.054	4.727	5.461	3.770	4.040	4.694	5.555
	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.005
	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
	0.005	0.006	0.011	0.017	0.004	0.005	0.011	0.016	0.004	0.005	0.011	0.017

（3）依赖变量 X_3 的偏差和均方误差的偏度、峰度分析。表 3.15 给出在此种无回答机制下，分别采用三种多重插补法对 $\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2,\hat{\beta}_3$ 的偏差进行测度。表中 3-7 行、8-12 行、13-17 行分别表示普通线性回归、贝叶斯线性回归、贝叶斯自助线性回归多重插补法在同一插补重数，不同无回答率下得到的 $\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2,\hat{\beta}_3$ 偏差的偏度和峰度。

表 3.15 依赖变量 X_3 随机无回答机制下系数偏差的偏度和峰度

方法	β_0		β_1		β_2		β_3	
	偏度	峰度	偏度	峰度	偏度	峰度	偏度	峰度
普通线性回归	-0.577	-1.772	-0.143	-2.028	0.690	-1.729	0.373	-1.848
	0.361	-1.979	0.345	-1.895	0.550	-1.815	-0.127	-1.901
	-0.269	-1.853	-0.089	-2.064	0.217	-2.006	-0.214	-1.894
	0.185	-1.911	0.039	-1.914	-0.518	-1.805	-0.214	-2.117
	0.190	-1.921	-0.041	-1.948	0.040	-1.875	0.482	-1.908

贝叶斯线性回归	-0.232	-2.147	0.274	-1.909	0.155	-1.908	-0.645	-1.751
	-0.405	-1.895	0.148	-1.934	0.012	-2.419	0.057	-1.934
	0.234	-1.862	0.290	-1.936	0.017	-1.901	0.069	-2.230
	0.103	-1.881	-0.068	-1.896	0.130	-1.913	0.105	-1.980
	0.250	-1.923	-0.048	-1.968	-0.343	-1.852	-0.062	-2.260
贝叶斯自助线性回归	0.152	-1.980	-0.362	-1.844	-0.389	-1.875	0.181	-2.232
	-0.551	-1.851	-0.399	-1.923	0.026	-2.408	-0.003	-2.425
	0.204	-2.193	0.207	-1.918	-0.214	-2.004	0.516	-1.814
	-0.033	-2.248	-0.323	-1.853	-0.661	-1.747	-0.581	-1.796
	-0.053	-1.881	0.036	-1.921	0.673	-1.734	0.161	-1.892

表 3.15 显示，在随机无回答机制下，分别采用普通线性回归、贝叶斯线性回归、贝叶斯自助线性回归多重插补法得到的系数偏差的偏度和峰度。其中，各系数偏差的偏度均有不同程度的左偏和右偏，与依赖变量 X_1 和 X_2 随机无回答机制下系数偏差的偏度相比，依赖变量 X_3 的系数偏差的偏度相对更接近正态分布。从偏差的峰度上看，各系数的峰度值都要比 0 小，数据分布较为平缓，且较为分散。

表 3.16　依赖变量 X_3 随机无回答机制下系数均方误差的偏度和峰度

方法	β_0		β_1		β_2		β_3	
	偏度	峰度	偏度	峰度	偏度	峰度	偏度	峰度
普通线性回归	0.607	-1.767	0.702	-1.715	0.491	-1.832	0.629	-1.759
	0.690	-1.732	0.435	-1.883	0.542	-1.791	0.695	-1.723
	0.709	-1.711	0.561	-1.810	0.609	-1.776	0.663	-1.753
	0.726	-1.702	0.580	-1.770	0.563	-1.780	0.681	-1.738
	0.697	-1.723	0.424	-1.857	0.587	-1.779	0.680	-1.734
贝叶斯线性回归	0.643	-1.747	0.221	-2.160	0.542	-1.796	0.622	-1.769
	0.701	-1.716	0.479	-1.809	0.520	-1.795	0.668	-1.745
	0.713	-1.712	0.492	-1.822	0.594	-1.765	0.683	-1.733
	0.698	-1.721	0.478	-1.822	0.491	-1.801	0.683	-1.731
	0.692	-1.721	0.179	-1.969	0.550	-1.787	0.652	-1.755
贝叶斯自助线性回归	0.654	-1.742	0.667	-1.752	0.383	-1.901	0.597	-1.801
	0.671	-1.737	0.632	-1.751	0.541	-1.807	0.647	-1.755
	0.684	-1.736	0.496	-1.839	0.554	-1.790	0.665	-1.744
	0.725	-1.703	0.369	-1.840	0.558	-1.778	0.660	-1.755
	0.701	-1.716	0.616	-1.757	0.599	-1.763	0.664	-1.749

表 3.16 显示，在随机无回答机制下，分别采用普通线性回归多重插补法、贝叶斯线性回归多重插补法、贝叶斯自助线性回归多重插补法得到的系数均方误差的偏度和峰度。其中，各系数均方误差的偏度均大于 0，偏斜程度均为右偏，且偏斜程度较大。各系数均

方误差的峰度均小于 0，其绝对值主要介于 1.7-1.9 之间，数据分布较为平坦。

与完全随机无回答机制下的模拟结果相比，在依赖三个自变量 X_1、X_2、X_3 的随机无回答机制下，分别采用三种多重插补法，$\hat{\beta}_1,\hat{\beta}_2,\hat{\beta}_3$ 的偏差绝对值和 $\hat{\beta}_1,\hat{\beta}_2,\hat{\beta}_3$ 的均方误差都没有明显差异，但 $\hat{\beta}_0$ 的均方误差明显变大。在依赖变量 X_1、X_2、X_3 的随机无回答机制下，与 PMM、DA、EMB 多重插补法相比，分别采用本文的三种多重插补法，$\hat{\beta}_1,\hat{\beta}_2,\hat{\beta}_3$ 的偏差绝对值更小，明显小于采用 PMM、DA 和 EMB 多重插补法情况下的偏差绝对值。$\hat{\beta}_1,\hat{\beta}_2,\hat{\beta}_3$ 的均方误差也小于采用 PMM 和 DA 多重插补法情况下的均方误差，与采用 EMB 多重插补法情况下的均方误差差异较小。随着插补重数增加，采用本文的三种多重插补法、PMM 和 EMB 多重插补法，$\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2,\hat{\beta}_3$ 的偏差绝对值和均方误差均没有出现递减趋势；采用 DA 多重插补法的 $\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2,\hat{\beta}_3$ 的偏差和均方误差出现递减趋势。

3.4 非随机无回答机制下的模拟结果

非随机无回答机制的无回答或与响应变量有关，或与解释变量有关。本节考虑依赖响应变量的非随机无回答机制和依赖自变量的非随机无回答机制。

3.4.1 依赖变量 Y 的非随机无回答机制下的模拟分析结果

在依赖变量 Y 的非随机无回答机制下，无回答或与响应变量有关。根据无回答率，计算变量 Y 的分位数，将小于该分位数的 y_i 设定为无回答。利用三种多重插补法对无回答进行插补，估计模型系数。

（1）偏差分析。表 17 给出在此种无回答机制下，分别采用普通线性回归、贝叶斯线性回归、贝叶斯自助线性回归多重插补法，$\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2,\hat{\beta}_3$ 的偏差。表 17 结构同表 1。由表 17 显示，分别采用这三种多重插补法，$\hat{\beta}_1,\hat{\beta}_2,\hat{\beta}_3$ 的偏差绝对值都小于 0.06，$\hat{\beta}_0$ 的偏差绝对值最大。当插补重数相同时，对应于无回答率的递增，$\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2,\hat{\beta}_3$ 的偏差绝对值均呈递增趋势，$\hat{\beta}_0$ 的偏差绝对值递增幅度相对更大。当无回答率相同时，随着插补重数增加，

$\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2,\hat{\beta}_3$的偏差绝对值没有明显递减趋势。

表 3.17　依赖 Y 的非随机无回答机制下 $\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2,\hat{\beta}_3$ 的偏差

重数	5%	10%	20%	30%	5%	10%	20%	30%	5%	10%	20%	30%
5	1.817	1.901	2.149	2.459	1.802	1.875	2.175	2.497	1.801	1.871	2.150	2.477
	-0.003	-0.003	-0.006	-0.009	-0.002	-0.003	-0.006	-0.009	-0.002	-0.003	-0.006	-0.008
	-0.005	-0.009	-0.016	-0.027	-0.005	-0.007	-0.016	-0.028	-0.005	-0.007	-0.016	-0.029
	-0.012	-0.018	-0.032	-0.048	-0.011	-0.017	-0.035	-0.052	-0.011	-0.016	-0.033	-0.049
10	1.797	1.899	2.135	2.542	1.799	1.885	2.156	2.526	1.808	1.898	2.140	2.497
	-0.002	-0.004	-0.006	-0.013	-0.003	-0.003	-0.007	-0.013	-0.003	-0.004	-0.006	-0.012
	-0.005	-0.008	-0.015	-0.029	-0.005	-0.008	-0.015	-0.029	-0.005	-0.008	-0.015	-0.028
	-0.011	-0.017	-0.032	-0.051	-0.011	-0.017	-0.033	-0.050	-0.012	-0.017	-0.033	-0.049
20	1.801	1.883	2.144	2.536	1.808	1.888	2.139	2.513	1.807	1.893	2.152	2.492
	-0.003	-0.004	-0.007	-0.013	-0.002	-0.004	-0.006	-0.012	-0.003	-0.003	-0.006	-0.011
	-0.005	-0.007	-0.015	-0.029	-0.005	-0.007	-0.016	-0.028	-0.005	-0.008	-0.016	-0.027
	-0.011	-0.016	-0.032	-0.051	-0.012	-0.017	-0.032	-0.050	-0.012	-0.017	-0.033	-0.050
30	1.808	1.875	2.135	2.523	1.808	1.885	2.151	2.513	1.806	1.879	2.156	2.510
	-0.003	-0.003	-0.006	-0.012	-0.002	-0.003	-0.006	-0.011	-0.002	-0.003	-0.006	-0.012
	-0.005	-0.007	-0.015	-0.028	-0.005	-0.008	-0.016	-0.028	-0.005	-0.007	-0.016	-0.029
	-0.012	-0.016	-0.032	-0.051	-0.012	-0.017	-0.033	-0.051	-0.012	-0.016	-0.034	-0.050
40	1.804	1.891	2.134	2.512	1.805	1.881	2.158	2.519	1.804	1.881	2.140	2.486
	-0.002	-0.003	-0.005	-0.012	-0.002	-0.003	-0.007	-0.012	-0.002	-0.003	-0.006	-0.012
	-0.005	-0.008	-0.015	-0.029	-0.005	-0.007	-0.016	-0.028	-0.005	-0.008	-0.015	-0.027
	-0.012	-0.017	-0.032	-0.050	-0.012	-0.016	-0.033	-0.051	-0.011	-0.017	-0.032	-0.049

（2）均方误差分析。表 3.18 给出在这种无回答机制下，分别普通线性回归、贝叶斯线性回归、贝叶斯自助线性回归多重插补法，$\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2,\hat{\beta}_3$的均方误差。表 3.18 结构同表 3.2。表 3.18 显示，分别采用三种多重插补法，$\hat{\beta}_1,\hat{\beta}_2,\hat{\beta}_3$的均方误差均小于 0.02。$\hat{\beta}_0$的均方误差明显大于$\hat{\beta}_1,\hat{\beta}_2,\hat{\beta}_3$的均方误差。当插补重数相同时，对应于无回答率的递增，$\hat{\beta}_0$的均方误差呈现递增趋势，$\hat{\beta}_1,\hat{\beta}_2,\hat{\beta}_3$则呈现略微递增的趋势。当无回答率相同时，随着插补重数增加，$\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2,\hat{\beta}_3$的均方误差没有明显减小趋势。

表 3.18　依赖 Y 的非随机无回答机制下 $\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2,\hat{\beta}_3$ 的均方误差

重数	5%	10%	20%	30%	5%	10%	20%	30%	5%	10%	20%	30%
5	4.153	4.739	6.317	9.723	4.068	4.621	6.788	10.186	4.108	4.557	6.978	9.841
	0.004	0.004	0.005	0.007	0.004	0.004	0.005	0.007	0.004	0.004	0.005	0.007

	0.001	0.002	0.003	0.005	0.001	0.001	0.003	0.005	0.001	0.001	0.003	0.005
	0.004	0.006	0.008	0.015	0.004	0.005	0.010	0.017	0.004	0.005	0.010	0.017
10	4.066	4.687	6.791	10.284	4.084	4.637	6.582	10.174	4.114	4.771	6.863	9.690
	0.004	0.004	0.005	0.007	0.004	0.004	0.005	0.007	0.004	0.004	0.005	0.006
	0.001	0.001	0.003	0.005	0.001	0.001	0.003	0.006	0.001	0.002	0.003	0.005
	0.004	0.005	0.010	0.015	0.004	0.005	0.009	0.015	0.004	0.005	0.011	0.015
20	4.072	4.529	6.696	10.410	4.121	4.653	6.570	10.118	4.099	4.683	6.768	9.381
	0.004	0.004	0.005	0.007	0.004	0.004	0.005	0.007	0.004	0.004	0.005	0.006
	0.001	0.001	0.003	0.005	0.001	0.001	0.003	0.005	0.001	0.001	0.003	0.005
	0.004	0.005	0.010	0.016	0.004	0.005	0.009	0.016	0.004	0.005	0.010	0.014
30	4.121	4.572	6.516	10.151	4.106	4.633	6.630	10.065	4.097	4.576	6.710	10.230
	0.004	0.004	0.005	0.007	0.004	0.004	0.005	0.007	0.004	0.004	0.005	0.007
	0.001	0.001	0.003	0.005	0.001	0.001	0.003	0.005	0.001	0.001	0.003	0.005
	0.004	0.005	0.009	0.015	0.004	0.005	0.009	0.016	0.004	0.005	0.010	0.015
40	4.102	4.673	6.497	10.282	4.078	4.612	6.706	10.203	4.097	4.574	6.364	9.971
	0.004	0.004	0.005	0.006	0.004	0.004	0.005	0.007	0.004	0.004	0.005	0.007
	0.001	0.001	0.002	0.006	0.001	0.001	0.003	0.005	0.001	0.001	0.002	0.005
	0.004	0.005	0.009	0.016	0.004	0.005	0.009	0.016	0.004	0.005	0.009	0.015

（3）依赖变量Y的非随机无回答机制下偏差和均方误差的偏度、峰度分析。表 3.19 给出在这种无回答机制下，分别采用三种多重插补法对$\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2,\hat{\beta}_3$的偏差进行测度。表中 3-7 行、8-12 行、13-17 行分别表示普通线性回归、贝叶斯线性回归、贝叶斯自助线性回归多重插补法在同一插补重数，不同无回答率下得到的$\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2,\hat{\beta}_3$偏差的偏度和峰度。

表 3.19　依赖变量Y非随机无回答机制下系数偏差的偏度和峰度

方法	β_0		β_1		β_2		β_3	
	偏度	峰度	偏度	峰度	偏度	峰度	偏度	峰度
普通线性回归	0.595	-1.807	-0.644	-1.772	-0.605	-1.796	-0.541	-1.847
	0.601	-1.799	-0.609	-1.783	-0.629	-1.776	-0.546	-1.845
	0.574	-1.826	-0.645	-1.767	-0.578	-1.823	-0.501	-1.886
	0.587	-1.817	-0.664	-1.754	-0.596	-1.808	-0.527	-1.869
	0.590	-1.811	-0.682	-1.735	-0.603	-1.801	-0.509	-1.877
贝叶斯线性回归	0.542	-1.857	-0.629	-1.782	-0.572	-1.833	-0.439	-1.945
	0.564	-1.834	-0.602	-1.808	-0.591	-1.811	-0.492	-1.890
	0.588	-1.814	-0.672	-1.743	-0.591	-1.813	-0.522	-1.868
	0.576	-1.825	-0.637	-1.776	-0.590	-1.813	-0.503	-1.887
	0.570	-1.831	-0.630	-1.780	-0.581	-1.822	-0.497	-1.894
贝叶斯自助线性回归	0.559	-1.841	-0.618	-1.792	-0.575	-1.827	-0.480	-1.911
	0.565	-1.830	-0.656	-1.756	-0.600	-1.800	-0.464	-1.915
	0.559	-1.837	-0.627	-1.781	-0.578	-1.820	-0.475	-1.908
	0.565	-1.835	-0.651	-1.764	-0.578	-1.824	-0.482	-1.907

	0.573	-1.828	-0.635	-1.780	-0.584	-1.816	-0.501	-1.886

表 3.19 显示，在此种无回答机制下，分别采用普通线性回归、贝叶斯线性回归、贝叶斯自助线性回归多重插补法得到的系数偏差的偏度和峰度。其中，$\hat{\beta}_0$ 偏差的偏度值在三种方法下均为正，其他系数偏差的偏度值均小于 0，整体上各系数的偏斜程度较大。从峰度上看，各系数偏差的峰度值比较接近，且分布较为平坦。

表 3.20 依赖变量 Y 非随机无回答机制下系数均方误差的偏度和峰度

方法	β_0		β_1		β_2		β_3	
	偏度	峰度	偏度	峰度	偏度	峰度	偏度	峰度
普通线性回归	0.687	-1.734	0.670	-1.747	0.693	-1.728	0.693	-1.728
	0.663	-1.753	0.682	-1.738	0.684	-1.737	0.620	-1.789
	0.647	-1.768	0.660	-1.757	0.666	-1.752	0.606	-1.804
	0.671	-1.748	0.682	-1.737	0.686	-1.735	0.657	-1.760
	0.669	-1.749	0.695	-1.728	0.682	-1.737	0.640	-1.772
贝叶斯线性回归	0.648	-1.766	0.663	-1.749	0.670	-1.746	0.630	-1.782
	0.661	-1.755	0.661	-1.751	0.686	-1.735	0.652	-1.763
	0.670	-1.748	0.677	-1.740	0.673	-1.745	0.662	-1.755
	0.660	-1.756	0.688	-1.732	0.672	-1.745	0.636	-1.776
	0.656	-1.759	0.679	-1.740	0.667	-1.751	0.638	-1.774
贝叶斯自助线性回归	0.631	-1.781	0.665	-1.750	0.667	-1.752	0.588	-1.820
	0.617	-1.790	0.657	-1.759	0.624	-1.782	0.555	-1.843
	0.609	-1.797	0.596	-1.810	0.615	-1.791	0.563	-1.838
	0.639	-1.774	0.682	-1.737	0.639	-1.774	0.597	-1.811
	0.663	-1.754	0.686	-1.734	0.679	-1.740	0.646	-1.768

表 3.20 显示，在依赖变量 Y 的非随机无回答机制下，分别采用普通线性回归、贝叶斯线性回归、贝叶斯自助线性回归多重插补法得到的系数均方误差的偏度和峰度。其中，各系数均方误差的偏度值均大于 0，其偏斜程度比较接近，且较大，均呈现右偏。各系数均方误差的峰度值均小于 0，其绝对值介于 1.7-1.85 之间，说明均方误差的分布较为平坦，且不集中。

与完全随机无回答机制和随机无回答机制的情况相比，在依赖变量 Y 的非随机无回答机制下，$\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2,\hat{\beta}_3$ 的偏差绝对值相对更大，估计量的均方误差也相对更大。在依赖变量 Y 的非随机无回答机制下，与采用 PMM 或 DA 多重插补法的模拟结果相比，分别采用本文三种多重插补法，$\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2,\hat{\beta}_3$ 的偏差绝对值明显变小，$\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2,\hat{\beta}_3$ 的均方误差也明显变小；与采用 EMB 多重插补法的偏差绝对值、均方误差差异较小。随着插补重数增加，

分别采用本文的三种多重插补法、PMM 和 EMB 多重插补法，$\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2,\hat{\beta}_3$ 的偏差和均方误差均没有出现递减趋势，DA 多重插补法的偏差和均方误差出现递减趋势。

3.4.2 依赖变量 X_3 的非随机无回答机制下的模拟分析结果

在依赖变量 X_3 的非随机无回答机制下，根据无回答率，计算变量 X_3 的分位数，将 x_{3i} 小于该分位数的观测 $(y_i,x_{1i},x_{2i},x_{3i})$ 中的 y_i 设定为无回答。在建立插补模型时，不考虑变量 X_3。分别采用三种多重插补法对无回答进行插补，再估计模型系数。

（1）偏差分析。表 3.21 给出在这种无回答机制下，分别采用普通线性回归、贝叶斯线性回归、贝叶斯自助线性回归多重插补法，$\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2$ 的偏差。表 3.21 的结构同表 3.1，每个格子里的三个数值依次为 $\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2$ 的偏差。表 3.21 显示，分别采用三种多重插补法，$\hat{\beta}_1,\hat{\beta}_2$ 的偏差绝对值小于 0.015，$\hat{\beta}_0$ 的偏差绝对值超过 25，远大于 $\hat{\beta}_1,\hat{\beta}_2$ 的偏差绝对值。随着插补重数增加，或对应于无回答率的递增，$\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2$ 的偏差绝对值无明显变化。

表 3.21　依赖 X_3 的非随机无回答机制下 $\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2$ 的偏差

重数	5%	10%	20%	30%	5%	10%	20%	30%	5%	10%	20%	30%
	25.767	25.753	25.747	25.702	25.773	25.753	25.757	25.697	25.759	25.755	25.756	25.727
5	-0.013	-0.011	-0.011	-0.006	-0.014	-0.011	-0.011	-0.007	-0.012	-0.011	-0.011	-0.008
	-0.007	-0.006	-0.006	-0.005	-0.006	-0.006	-0.006	-0.006	-0.006	-0.005	-0.006	-0.006
	25.771	25.754	25.746	25.712	25.771	25.752	25.750	25.712	25.769	25.763	25.739	25.688
10	-0.013	-0.011	-0.010	-0.007	-0.013	-0.011	-0.010	-0.007	-0.013	-0.012	-0.009	-0.005
	-0.007	-0.006	-0.006	-0.006	-0.007	-0.006	-0.006	-0.006	-0.007	-0.006	-0.006	-0.005
	25.766	25.756	25.741	25.712	25.764	25.760	25.752	25.710	25.765	25.753	25.736	25.707
20	-0.012	-0.011	-0.009	-0.008	-0.012	-0.012	-0.010	-0.007	-0.012	-0.011	-0.009	-0.007
	-0.007	-0.006	-0.006	-0.006	-0.007	-0.006	-0.006	-0.006	-0.007	-0.006	-0.006	-0.005
	25.769	25.753	25.743	25.712	25.764	25.757	25.749	25.706	25.767	25.760	25.750	25.707
30	-0.013	-0.011	-0.010	-0.007	-0.012	-0.012	-0.010	-0.007	-0.012	-0.012	-0.010	-0.007
	-0.007	-0.006	-0.006	-0.006	-0.007	-0.006	-0.006	-0.006	-0.007	-0.006	-0.006	-0.006
	25.769	25.760	25.743	25.716	25.764	25.756	25.746	25.711	25.766	25.755	25.748	25.710
40	-0.013	-0.012	-0.010	-0.007	-0.012	-0.012	-0.009	-0.007	-0.012	-0.012	-0.010	-0.007
	-0.007	-0.006	-0.006	-0.006	-0.006	-0.006	-0.006	-0.006	-0.007	-0.006	-0.006	-0.005

（2）均方误差分析。表 3.22 给出在此种无回答机制下，分别采用普通线性回归、贝叶斯线性回归、贝叶斯自助线性回归多重插补法，$\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2$ 的均方误差。表 3.22 结构同表

3.2，每个格子里的三个数值依次为 $\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2$ 的均方误差。表 3.22 显示，分别采用三种多重插补法，$\hat{\beta}_1,\hat{\beta}_2$ 的均方误差小于 0.12，$\hat{\beta}_0$ 的均方误差已经超过 650，远远大于 $\hat{\beta}_1,\hat{\beta}_2$ 的均方误差。随着插补重数增加，或对应于无回答率的递增、，$\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2$ 的均方误差无明显变化。

表 3.22　依赖 X_3 的非随机无回答机制下 $\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2$ 的均方误差

重数	5%	10%	20%	30%	5%	10%	20%	30%	5%	10%	20%	30%
	672.993	672.398	672.208	669.969	673.306	672.360	672.615	669.729	672.525	672.404	672.670	671.201
5	0.111	0.111	0.112	0.114	0.111	0.111	0.111	0.114	0.110	0.110	0.111	0.112
	0.023	0.023	0.022	0.023	0.023	0.023	0.022	0.023	0.022	0.022	0.022	0.023
	673.154	672.450	672.052	670.481	673.161	672.298	672.290	670.433	673.100	672.829	671.750	669.181
10	0.110	0.111	0.111	0.113	0.111	0.111	0.111	0.113	0.111	0.111	0.112	0.113
	0.023	0.022	0.022	0.023	0.022	0.023	0.022	0.023	0.023	0.022	0.022	0.023
	672.934	672.572	671.815	670.411	672.848	672.735	672.418	670.287	672.908	672.348	671.547	670.127
20	0.110	0.111	0.111	0.113	0.111	0.111	0.111	0.112	0.111	0.111	0.111	0.113
	0.023	0.023	0.022	0.023	0.023	0.022	0.022	0.023	0.023	0.023	0.022	0.023
	673.093	672.365	671.912	670.425	672.823	672.575	672.220	670.090	672.976	672.746	672.315	670.136
30	0.110	0.111	0.111	0.113	0.111	0.111	0.111	0.112	0.110	0.111	0.111	0.113
	0.023	0.023	0.022	0.023	0.023	0.023	0.022	0.023	0.023	0.023	0.022	0.023
	673.072	672.726	671.929	670.538	672.810	672.514	672.051	670.345	672.921	672.439	672.210	670.313
40	0.111	0.111	0.111	0.112	0.111	0.111	0.111	0.113	0.110	0.111	0.112	0.113
	0.023	0.022	0.022	0.023	0.023	0.022	0.022	0.023	0.023	0.023	0.022	0.023

（3）依赖变量 X_3 的非随机无回答机制下偏差和均方误差的偏度、峰度分析。表 3.23 给出在完全随机无回答机制下，分别采用三种多重插补法对 $\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2$ 的偏差进行测度。表中 3-7 行、8-12 行、13-17 行分别表示普通线性回归、贝叶斯线性回归、贝叶斯自助线性回归多重插补法在同一插补重数，不同无回答率下得到的 $\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2$ 偏差的偏度和峰度。

表 3.23　依赖变量 X_3 非随机无回答机制下系数偏差的偏度和峰度

方法	β_0		β_1		β_2	
	偏度	峰度	偏度	峰度	偏度	峰度
普通线性回归	-0.612	-1.761	-0.362	-1.839	0.460	-1.811
	-0.121	-1.951	0.572	-1.815	0.642	-1.760
	-0.311	-2.023	0.458	-1.811	0.444	-1.823
	-0.153	-1.954	0.486	-1.810	0.531	-1.795
	-0.375	-1.980	0.469	-1.859	0.627	-1.782
贝叶斯线	0.368	-1.906	0.544	-1.784	0.586	-1.768

性回归	-0.111	-1.871	0.661	-1.748	0.717	-1.711
	-0.689	-1.733	0.436	-1.965	0.629	-1.771
	-0.606	-1.787	0.323	-2.021	0.601	-1.763
	-0.448	-1.901	0.686	-1.723	0.446	-1.864
贝叶斯自助线性回归	-0.699	-1.720	0.559	-1.834	0.727	-1.701
	-0.424	-1.961	0.163	-2.188	0.667	-1.748
	0.061	-2.184	0.726	-1.705	0.687	-1.725
	-0.378	-1.963	-0.029	-2.130	0.640	-1.751
	-0.507	-1.814	0.369	-1.952	0.664	-1.738

表 3.23 显示，在这种无回答机制下，分别采用普通线性回归、贝叶斯线性回归、贝叶斯自助线性回归多重插补法得到的系数偏差的偏度和峰度。其中，$\hat{\beta}_0$ 偏差的偏度基本上都小于 0，呈现左偏分布，$\hat{\beta}_1$ 和 $\hat{\beta}_2$ 偏差的偏度值基本都大于 0，且偏斜程度较大。各系数偏差的峰度值均小于 0，分布较为平坦。

表 3.24　依赖变量 X_3 非随机无回答机制下系数偏差的偏度和峰度

方法	β_0		β_1		β_2	
	偏度	峰度	偏度	峰度	偏度	峰度
普通线性回归	-0.612	-1.761	-0.362	-1.839	0.460	-1.811
	-0.121	-1.951	0.572	-1.815	0.642	-1.760
	-0.311	-2.023	0.458	-1.811	0.444	-1.823
	-0.153	-1.954	0.486	-1.810	0.531	-1.795
	-0.375	-1.980	0.469	-1.859	0.627	-1.782
贝叶斯线性回归	0.368	-1.906	0.544	-1.784	0.586	-1.768
	-0.111	-1.871	0.661	-1.748	0.717	-1.711
	-0.689	-1.733	0.436	-1.965	0.629	-1.771
	-0.606	-1.787	0.323	-2.021	0.601	-1.763
	-0.448	-1.901	0.686	-1.723	0.446	-1.864
贝叶斯自助线性回归	-0.699	-1.720	0.559	-1.834	0.727	-1.701
	-0.424	-1.961	0.163	-2.188	0.667	-1.748
	0.061	-2.184	0.726	-1.705	0.687	-1.725
	-0.378	-1.963	-0.029	-2.130	0.640	-1.751
	-0.507	-1.814	0.369	-1.952	0.664	-1.738

表 3.24 显示，在此种无回答机制下，分别采用普通线性回归、贝叶斯线性回归、贝叶斯自助线性回归多重插补法得到的系数均方误差的偏度和峰度。其中，$\hat{\beta}_0$ 均方误差的偏度值多数小于 0，呈左偏分布，其他系数均方误差的偏度值基本上都大于 0，且三个系数偏差的偏斜程度都较大。从峰度上看，各系数均方误差的峰度值都小于 0，且峰度值也较为接近，均方误差的分布相比于正态分布更为平坦。

在依赖变量 X_3 的非随机无回答机制下，$\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2$ 的偏差绝对值和均方误差均大于在完全随机无回答机制下和在随机无回答机制下 $\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2$ 的偏差绝对值和均方误差。其中，$\hat{\beta}_1,\hat{\beta}_2$ 的偏差绝对值和均方误差的增加幅度相对小，$\hat{\beta}_0$ 的偏差绝对值和均方误差增加幅度过大。在依赖变量 X_3 的非随机无回答机制下，与采用 PMM 多重插补法的模拟结果相比，采用本文的三种多重插补法，$\hat{\beta}_1,\hat{\beta}_2$ 偏差绝对值更小。$\hat{\beta}_0$ 的均方误差要远大于采用 PMM 多重插补法所得到 $\hat{\beta}_0$ 的均方误差。$\hat{\beta}_1,\hat{\beta}_2$ 的均方误差则明显小于采用 PMM 多重插补法的情况。与采用 EMB 多重插补法的偏差绝对值和均方误差相比，差异不大。随着插补重数增加，采用本文三种多重插补法、PMM 多重插补法和 EMB 多重插补法，$\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2$ 的偏差和均方误差均没有出现递减趋势。

第 4 章 总结

本文在多种插补重数、多种无回答率和多种无回答机制下，分别采用普通线性回归、贝叶斯线性回归、贝叶斯自助线性回归多重插补法，计算无回答的插补值，并估计回归系数的偏差、均方误差，以及它们的偏度和峰度。讨论估计量的统计性质。模拟结果显示，在三种无回答机制下，回归系数估计量的偏差绝对值与均方误差都较小。在无回答率递增的情况下，系数估计量的偏差绝对值和均方误差呈较小递增趋势。同样当插补重数递增时，系数估计量的偏差绝对值和均方误差没有递减趋势。在相应的条件下，系数估计量的偏差绝对值和均方误差都明显小于采用 PMM 或 DA 插补法插补的情况；与采用 EMB 插补法的情况差异较小。在响应变量和自变量有线性关系的情况下，普通线性回归、贝叶斯线性回归、贝叶斯自助线性回归多重插补法都利用了这种线性关系，改善了插补结果。在响应变量和自变量有线性关系的情况下，建议选择普通线性回归、贝叶斯线性回归、贝叶斯自助线性回归多重插补法。其中当插补充数选择为 5 时，就可以给出较好的插补结果。贝叶斯线性回归多重插补法能够利用已知的先验信息，更好描述插补值的统计分布。贝叶斯自助线性回归多重插补法可以用于无先验信息的情况，适用范围更广。普通线性回归多重插补法相对这两种方法多重插补法，使用更简单。在依赖响应变量的非随机无回答机制下，截矩项估计量具有较大的偏差绝对值和均方误差，其他系数估计量的偏差绝对值和均方误差都相对较小。从系数的偏差、均方误差的偏度和峰度考量，在三种不同的无回答机制下，采用三种线性回归多重插补法得到的各系数指标的偏斜程度均比较小，与正太分布相比，左偏和右偏的情况都有发生；从各系数指标的峰度看，得到的数据的大致分布要比正态分布更为平缓，说明数据的集中程度不高。

综上所述，本文采用的三种线性回归多重插补法在三种无回答机制下，得到的结果分别有所不同。首先，在三种无回答机制下进行比较，完全随机无回答机制下采用三种线性回归多重插补法得到的系数的偏差和均方误差都要比其他两种无回答机制所得到的结果小，而在随机无回答机制下得到的系数的偏差和均方误差明显变大，在非随机无回答机制下得到的偏差和均方误差则最大。其次，与本文分析结果中所列举的 PMM、DA、EMB

等多重插补法相比，本文采用的三种线性多重插补法在三种无回答机制下，所得系数的偏差和均方误差在完全随机无回答机制、随机无回答机制下基本上小于 PMM、DA 等多重插补法得到的结果，而与 EMB 多重插补法得到的结果差异较小；在非随机无回答机制下，得到系数的偏差与均方误差要略大于 PMM 多重插补法，与 EMB 多重插补法得到的结果差异不大。对在三种不同无回答机制，三种不同的方法下，得到的系数指标的偏度的绝对值基本上都小于 1，并同时带有左偏或右偏的情况，说明数据在水平方向的分布要比正态分布略有偏斜；从峰度来看，在非随机无回答机制的情况下，各系数指标的峰度值都比较接近，而且数据的分布较为平缓，其他无回答机制下，峰度值不够稳定，分布不够集中。

参考文献

[1] 金勇进、邵军. 无回答的统计处理[M].中国统计出版社，2009

[2] 金勇进. 缺失数据的插补调整[J].数理统计与管理，2001（5）：47-53

[3] 金勇进，朱琳. 不同插补方法的比较[J].数理统计与管理，2000

[4] 刘燕.基于 Logistic 回归的邻近择优插补法[D].天津财经大学，2013

[5] 刘艳玲. 调查数据无回答的插补方法及模拟比较[D]. 天津财经大学，2012

[6] 庞新生. 多重插补方法与应用研究[M].经济科学出版社，2013

[7] 庞新生. 缺失数据插补处理方法的比较研究[J].统计与决策，2012（24）：18-22

[8] 庞新生，李萌.多重插补方法中插补模型的比较[J].统计与决策，2015（9）:82-84

[9] 乔丽华，傅德印.缺失数据的多重插补方法[J].统计教育，2006（12）：4-7

[10] 王璐、王飞. Hot deck 插补和插补后数据的方差模拟研究[J].数量经济技术经济研究，2006（2）：148-152

[11] 杨贵军、李静华. 基于 PMM 多重插补法的线性回归模型系数估计量的模拟研究[J].数量经济技术经济研究，2014（10）:139-150

[12] 杨贵军、骆新珍. 基于 DA 插补法的线性回归模型系数估计量的模拟研究[J].统计与信息论坛，2014（3）：3-8

[13] 杨贵军、孙玲莉、孟杰. 基于 EMB 多重插补法的线性回归模型系数估计量的模拟研究[J].数量经济技术经济研究，2016（10）：128-141

[14] 杨贵军，蔡娟，赵晓芸.高相关性辅助变量择优回归插补法[J].统计信息与论坛，2012，27（6）：8-11

[15] 张香云，张秀伟. 不同缺失率下 EM 算法的参数估计[J].数理统计与管理，2008，27（3）：428-431

[16] 赵晓云. 纵向数据缺失处理对分析模型影响的研究[D].天津财经大学，2013

[17] Allison,P.D. Missing data[M].SAGE publications,2001

[18] Barzi F, Woodward M. Imputations of missing values in practice:results from imputations of serum cholesterol in 28 cohort studies[J]. American Statistical Association,2008

[19] Efron B. Bootstrap methods: another look at the jackknife[J]. The Annals of Statistics, 1979,7(1) :1~26.

[20] Efron B. and Tibshirani R. An Introduction to the Bootstrap[M]. Chapman & Hall, London, 1993

[21] Graham, J.W. Missing Data:Analysis and Design[M].Springer,1997

[22] Hansen M H,Hurwitz W N. The problem of non-response in sample surveys[J]. Journal of the American Statistical Association,1946

[23] Horton N J, Lipsitz S R. Multiple imputation in practice:comparison of software packages for regression models with missing variables[J]. The American Statistician,2001

[24] Kalton G. Compensating for Missing Survey Data[J]. Research Report Series. Ann Arbor, MI:Instiute for Social Rsearch, University of Michigan, 1983

[25] Little R. and Rubin D. Statistical Analysis With Missing Data [M]. John Wiley & Sons, Inc, 2002

[26] Little, R.J.A., Missing Data in Large Survey(with Discussion)[J]. Journal of Business and Economic Statistics,1988

[27] Okafor F. and Lee H.Double Sampling for Ratio and Regression Estimation with Subsampling the Nonrespondents [J].Survey Methodology, 2000,26(2):183~188

[28] Rubin D. Multiple Imputation for Nonresponse in Survey[M]. New York John Wiley & Sons, Inc, 1987

[29] Stef van Buuren. Flexible Imputation of Missing Data [M]. Leiden, 2012

[30] Schafer J L. Analysis of incomplete multivariate data[M].Chapman and Hall,1997

[31] Tang G, Little R J A, Raghunathan T E. Analysis of multivariate missing data with non-ignorable non-response[J]. Biometrika,2003

后记

三年时光匆匆，转眼间已来到了自己的毕业季，回顾三年的学习生涯，虽充满着学习专业知识的艰辛，但也有获取知识的快乐相伴。

在三年前选择统计学专业作为自己的主修专业时，我还没有认识到在研究生阶段的学习中伴随的困难和挑战。此间我也曾迷茫过，但通过不懈地努力，我依然坚持自己的选择。三年来，我的导师杨贵军老师在学习和生活方面给予了我很大帮助，尤其在论文选题、讨论、分析、写作等方面，做了细致专业的指导，并给出很多好的建议，这让我受益匪浅。如今能顺利完成毕业论文，更是少不了杨老师的谆谆教导，在毕业之际，谨向杨老师表达内心诚挚的感谢。与此同时，还要感谢我的博士师兄孟杰，以及同门孙玲莉、张率、蔡凯月、邹文慧同学，正是有了他们在学习和生活中的鼎力相助，让我懂得如何在团队中发挥自己的作用，并使毕业论文得以完成。他们和杨老师给予我的帮助，是我在今后学习工作的重要财富。

三年的研究生学习生涯，使我具备了一定的走向社会，走向工作岗位的能力。在这里，我首先要特别感谢我的父母，没有他们的支持和理解，我也不可能经历这段难忘的求学历程；其次要感谢统计系的曹景林、李腊生、刘乐平等专业课教师，是他们在各项课程中的悉心指导，让我对统计学知识有了更深刻的理解和认识；还要感谢研究生院的陈楠书记，郭鑫、常晓春、张博超老师，是他们在平时给予了我许多展现、锻炼自己的机会，使我的为人处事能力有了较大提高；最终，要感谢和自己共处三年的七位室友，是他们给我的学习生活带来了许多快乐，同样也要感谢全班同学，感谢他们三年来对我工作的大力支持以及在生活中给予我的各种帮助。

从学校毕业，还有很长的路要走，在天财的岁月让我难忘。感谢母校给予我很不错的发展平台，让我结识了很多好的老师，还有很多不错的朋友，使我更好地成长。在此，衷心祝愿母校有更快更好的发展，更广阔的前景；也祝愿我的老师们身体健康，工作顺利；更祝愿我的同学和朋友们前程似锦。作为将要从母校毕业的学子，我更要在今后的工作中保持母校学思达信的优良传统，为社会贡献自己的一份绵薄之力。