学校编码：10384　　分类号_____密级_____
学　　号：19020071152091　　UDC_____

厦门大学

硕　士　学　位　论　文

分层线性模型在住院费用影响因素分析中的应用研究

An applying research on HLM in the influential factors analysis of hospitalization expenses

王洁丹

指导教师姓名:　张志强　副教授
专　业　名　称:　概率论与数理统计
论文提交日期:　2010 年　5　月
论文答辩日期:　2010 年　6　月
学位授予日期:　2010 年　　月

答辩委员会主席：_____
评　　阅　　人：_____

2010 年 5 月

厦门大学学位论文原创性声明

本人呈交的学位论文是本人在导师指导下，独立完成的研究成果。本人在论文写作中参考其他个人或集体已经发表的研究成果，均在文中以适当方式明确标明，并符合法律规范和《厦门大学研究生学术活动规范（试行）》。

另外，该学位论文为（　　）课题（组）的研究成果，获得（　　）课题（组）经费或实验室的资助，在（　　）实验室完成。（请在以上括号内填写课题或课题组负责人或实验室名称，未有此项声明内容的，可以不作特别声明。）

声明人（签名）：

年　　月　　日

厦门大学学位论文著作权使用声明

本人同意厦门大学根据《中华人民共和国学位条例暂行实施办法》等规定保留和使用此学位论文，并向主管部门或其指定机构送交学位论文（包括纸质版和电子版），允许学位论文进入厦门大学图书馆及其数据库被查阅、借阅。本人同意厦门大学将学位论文加入全国博士、硕士学位论文共建单位数据库进行检索，将学位论文的标题和摘要汇编出版，采用影印、缩印或者其它方式合理复制学位论文。

本学位论文属于：

（　　）1.经厦门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厦门大学保密委员会审定过的学位论文，未经厦门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。）

声明人（签名）：

年　　月　　日

摘 要

住院费用影响因素的研究一直是医疗保险精算领域风险控制研究的主要内容。目前，用于住院费用影响因素分析的方法主要是传统的统计分析方法。然而，住院费用的产生往往不仅由病人层次的因素影响，同时也受到医院级别及地区因素的影响，需要分析具有层次结构的数据。而具有层次结构的资料不适宜用传统的统计方法，宜采用分层线性模型。

本文应用分层线性模型，参考已有文献资料，利用福建省莆田市荔城区某镇新型农村合作医疗保险所补偿的住院费用的半年记录，探讨了影响住院费用的个人层次因素与医院层次因素，并最终建立了住院费用对数值与肿瘤情况、住院天数、药品费对数值及医院级别的分层线性模型。本研究的结果表明：影响住院费用的个人层次因素有肿瘤情况、住院天数、药品费，医院层次因素有医院级别；医院级别越高平均住院费用越高；在药品费与住院天数一定的情况下肿瘤病人比非肿瘤病人的平均住院费用高；在药品费与肿瘤情况一定的情况下住院天数越多平均住院费用越高；在肿瘤情况以及住院天数一定的条件下，药品费的增加会导致住院费用的增加，但药品费对数值的增加在高级别医院所导致的住院费用对数值的增加比低级别医院要多。

关键词：分层线性模型、层次结构、住院费用、影响因素

Abstract

The study on factors affecting hospitalization expenses has always been the main contents of risk control study on medical actuarial field．Currently，the main analysis methods for hospitalization expenses are the traditional statistical analysis．However, the production of hospitalization expenses is influenced not only by the factors of patient-level, but also by the hospital-level and regional factors, which need to analyze the hierarchical data. Moreover, the traditional statistical methods are inappropriate to the information of hierarchical structure, the appropriate method is Hierarchical Linear Model (HLM).

In this article, the Hierarchical Linear Model is used to analyze the factors influencing hospitalization expenses of patient-level and hospital-level, basing on past documents and the records of the compensation costs by the new rural cooperative medical insurance in half a year, which is from a certain town in Licheng District, Putian City, Fujian Province. And a two-level HLM model is established, which is about ln(hospitalization expenses), tumor situation, hospitalization days, ln(drug costs) and hospital grade. The main results of this research reveal: the factors influencing hospitalization expenses in patient-level are tumor situation, hospitalization days, ln(drug costs) and hospital grade in hospital-level; the higher hospitalization grade, the higher average hospitalization expenses; when pharmaceutical costs and hospitalization days are under a certain circumstance, the average hospital cost of cancer patients is higher than non-cancer patients; when drug costs and cancer situation are under a certain circumstance, the more hospitalization days, the higher average hospital cost; when cancer situation and hospitalization days are under a certain circumstance, with an increase of drug costs the hospitalization costs increase, but the increase of ln(drug cost) will cause smaller increase of ln(hospitalization cost) in high grade hospitals than low grade.

Key words: Hierarchical Linear Model; hierarchical structure; hospitalization expense; influencing factors

目 录

Contents

第一章 绪论

1.1 多层数据结构的普遍性

在社会研究中，很多都涉及多水平、多层的数据结构。比如在组织研究中就有工人从属于公司，公司从属于行业的现象。在该研究中，工人处于数据结构的第一层，公司处于数据结构的第二层，行业处于数据结构的第三层。对于第一层的工人数据，研究者可以提出一系列的研究问题。除此之外,也可以针对第二层的公司或行业提出又一系列的研究问题。当需要研究工作场所的特征，诸如决策的集中度如何影响工人的生产率时，工人和公司都是分析单位；变量是在两个层次进行测量的。

在教育研究中多层的数据结构尤为典型，这里学生镶嵌于班级，班级镶嵌于学校，这就是三层的数据结构。在多国研究中，当人口学家研究不同国家的经济发展如何与成年人教育程度互动并影响生育率时，研究中不仅要测量属于国家层次的经济指标，也要调查以住户为单位的教育与生育情况。这里住户与国家都是研究单位，其中住户就从属于国家，因而该数据结构是两层的[1]。

其它的多层数据可见于在不同的地理政治区域进行的大规模的评价和调查。比如,对大学生身体素质有影响的因素,在不同的地区可能会受到当地社会经济及气候特征的影响而起不同的作用,因而学生间的差异以及学生层变量间的相关的差异都可能受地区层的变量影响[2]。

类似的数据形式也往往存在于纵向研究或重复测量研究，其数据的收集往往是对同一组个体在不同时点所做的多次观察。这种重复测量包含了每一个体成长的轨迹信息。

例如，在发展心理学研究中，心理学家特别感兴趣的是一个人的个性及其所处的不同环境如何影响他的成长。研究者可以在一段时间内对儿童或其他被试进行多次观察。不同时间的观测数据形成了数据结构的第一层,而被试之间的个体差异就形成了第二层。当样本中每一个体均是按相同时点被不断进行观察时，通常将其视为一种不同个体与不同时点的交互研究设计。但是当对不同个体的观察在时点的数量和间隔上有所不同时，我们可以将重复观察视为嵌套于同一个体的

不同场合。

教育研究尤其具有挑战性，因为学生成长的研究总是涉及在结构上从属于个体的重复观察，并且每一个体又从属于某一组织结构。例如，研究大学生英语成绩的影响因素，要关注三个焦点：学生在一年(或其中的某一段)英语课程学习中的进步，教学方法对学生个人性格和学习收获上的效果，以及上述联系又如何受到班级设置和教师行为与特征的影响。因而，数据具有三个层次：第一层为各时点的重复观察，第二层为学生，第三层为班组或学校，其中各时点的重复观察从属于各学生，而每一学生又从属于班级或者学校[1]。

1.2 传统统计方法处理多层数据结构的局限性

1.2.1 传统线性回归模型的局限

传统的线性回归模型可分为一元回归模型和多元回归模型。一元回归模型只有一个自变量，因而也称为简单回归，可以表示为：

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

其中，β_0 和 β_1 称为回归参数，ε 为随机波动的误差项。

传统的一元回归模型通常有以下三个基本假设：

(1) Y 与 X 之间存在线性关系；

(2) X 是非随机变量；

(3)对于所有的取值 $X_i(i=1,\mathrm{K}\ ,N)$，误差项具有相同的方差，即 $Var(\varepsilon_i)=\sigma^2$。且 ε_i 是相互独立、服从正态分布 $N(0,\sigma^2)$ 的随机变量。

如果要建立住院费用的预测模型，考虑医院间(用 W 表示)的差异的影响，一般我们会建立如下传统的一般线性回归模型：

$$Y_{ij} = \beta_0 + \beta_1 X_{ij} + \beta_2 W_j + \varepsilon_{ij}$$

其中 Y_{ij} 代表第 j 家医院第 i 个住院病人的住院费用，β_0 是回归截距，β_1 是线性回归系数，β_2 表示第 j 家医院对住院费用的影响。ε_{ij} 表示第 j 家医院第 i 个病人住院费用的随机误差。在这种情况下，模型要求的线性假设以及误差项服从正态分

布的假设较易得到满足，但方差齐性，尤其是个体间随机误差相互独立的假设却很难满足。因为这一假设意味着Y是从某个总体中随机抽样的，但是我们在对Y进行取样时，如果个体同属于同一个第二层单位，比如来自于同一家医院，那么Y就会受到相同的医院变量的影响，这样误差项就不能满足上述假设。即不同医院的个体间可以假设相互独立，但是来自于同一家医院的个体间由于受到相同的医院变量的影响，很难保证相互独立。在这种情况下，传统回归模型不能满足假设条件，所以据此建立的回归模型也是不准确的。

1.2.2 传统统计技术的局限

传统统计技术处理多层数据结构时，通常可能会采用两种并不令人完全信服的方法：

第一种方法，是把处于高层水平的变量分解到个体水平。医院的特征都被指定给个人，以便在病人层次上进行分析。这种方法的问题是，如果我们知道个人是来自于同一家医院的，那么他们在医院变量W_j上的取值就是相同的。但这样处理，不能满足其观察值独立性的假设，这是经典统计技术的基础。

第二种方法，是把处于个体水平的变量集中到高层水平，并且在高层水平上进行分析。于是我们就把个人的特征集中到了不同医院上，并对医院进行分析，这样就可能会忽略由于医院内看病人数的不同而产生不同的权重影响。这里的主要问题是我们丢失了组内信息，而它们在我们开始进行分析之前可能占到总变异的 80%以上。结果，集中变量之间的相关性往往都很高，并且它们与未集中变量之间的相关可能是极不相同的。这样我们就浪费了信息，并且，如果我们试图在个体水平来解释集中后的分析结果，那只会是一种曲解。

上述集中与分解的两种处理方式将很有可能得到不同的结果，那么在对结果的解释上也会很不一致。基于上述的讨论，这两种分析数据的方法有一个共同缺陷是，它们都没有考虑数据间分层的特点，有可能对数据结果做出不合理的甚至是错误的解释。这就是传统回归分析方法在分析分层数据时的局限性。

1.3 分层线性模型的简介

分层线性模型(Hierarchical Linear Models)是基于方差成分分析理论与多元

统计分析相结合的新的统计分析技术，主要用于研究具有层次结构或嵌套式结构的数据。该模型是针对经典统计技术在处理具有多层结构的数据时所存在的局限，以及可能产生的对分析结果的曲解而提出的，它适宜对广泛存在的多层数据结构进行恰当的、深入的分析和解释。

分层线性模型在不同领域的文献中有不同的称呼。在社会学研究中，它经常被称为多层线性模型(multilevel linear models)；在生物统计研究中更经常用的名字是混合效应模型(mixed-effects models)和随机效应模型(random-effects models)；计量经济学文献称之为随机系数回归模型(random-coefficient regression models)[1]。

1.4 分层线性模型的发展概况

将其称为分层线性模型(hierarchical linear models)，是因为它指出了即使在不同应用中，比如成长研究、组织效应、综合研究，其数据都存在一个相同的重要结构特征。这一称呼最早由 Lindley 和 Smith[4]在 1972 年提出，他们为具有复杂误差结构的嵌套数据研制了一个通用的研究框架。由于该模型的应用需要对非平衡数据进行方差协方差成分的估计，因而当时只能解决一些极简单的问题，不能提供通用的估计方法。直到 1977 年 Dempster、Laird 和 Rubin[5]在 EM 算法上取得巨大进展，才形成了切实可行并被广泛应用的方差协方差成分的估计方法。1986 年 Goldstein[6]提出了通过迭代再加权的一般最小二乘法的协方差成分的估计方法。1987 年 Longford[7]提出 Fisher 得分算法。随着计算机技术的发展与算法的程序化，目前已经有很多统计软件可以拟合分层线性模型，如 HLM、MIXOR、MLWIN、SAS 及 SPSS 等。

分层线性模型发展至今，已经产生了许多富有创造性的应用：第一，模型所采用的结果变量已由连续型分布推广至二分类结果变量、计数数据、序次分类变量以及多分类结果变量；第二，模型不仅可以包括纯粹的嵌套数据结构，而且可以与交互分类的数据结构相结合；第三，结果的多元影响问题已成为主流；第四，潜在变量被引入模型；第五，对分层模型的贝叶斯推断获得了更广泛的普及和应用。

国内的研究多数仅限于对国外理论与模型的翻译与消化吸收上，有郭志刚等

译的《分层线性模型：应用与数据分析方法》[1]，王济川等著的《多层统计分析模型》[3]，孟庆茂、侯杰泰老师整理编写的《协方差结构模型与多层线性模型》等。

在应用研究方面1992年荷兰的J.jerweel对影响学生的学习因素进行了研究。同时考虑了来自五个层次的影响因素。研究结果发现，具有中等学习能力的学生，其学习成绩更容易受教学与学习环境的影响。国内陈柏熹、王文中对香港中学生科学成绩的影响因素用三层次模型进行了分析，分析结果显示，科学成绩的变异量中有三分之二来自学生层次，另外的三分之一来自学校与教师层次[11]。

1.5 住院费用影响因素的研究概况

对医疗费用风险因素进行分析，既可用作医疗保险精算时进行风险分类的依据，又是医疗保险管理机构进行风险控制的数量基础。确定哪些是影响医疗服务（门诊或住院）费用的主要风险因素，并对其影响强度和作用机制进行定量描述，是整个风险控制方法研究的主要内容。

目前国内的对医疗费用风险因素的研究多采用传统的统计方法，比如单因素分析方法和多元线性回归模型等。沈华亮、徐德法和王龙星等[8]在对郑州市公费医疗费用的风险因素进行分析时就采用单因素分析的方法对次均门诊费用、就诊者人均半年门诊费用、次均住院天数和住院费用四个指标的均数进行了比较和假设检验。程晓明、赵永明[9]在对上海市徐汇、虹口和长宁三区和上海、崇明两县的少年儿童住院医疗保险资料进行分析时，发现影响住院医疗费用的主要因素除被保险人的年龄外，还包括所在地区、医院级别、家长的医疗保健制度以及是否有手术等。而在多元线性回归模型中，很多研究都表示被保险人每次门诊和住院的医疗花费是由被保险人的病因和病情，医疗机构的级别和治疗、检查方式等众多风险因素综合作用的结果。因而，应用最广的是多元线性回归模型[10]。

然而，个人的病情、医疗机构的级别和地区因素分属于不同层次上的影响因素。个人病情以及治疗方式属于个人层次，而医疗机构级别属于医院层次，地区因素又属于地区层次。个人从属于医院，医院从属于地区。这是一个典型的三层数据结构。用传统的统计方法来处理这种的具有层次的数据，必然会导致部分信息丢失，参数估计失效，从而使得解释出现错误。因而，有必要利用分层线性模

型来研究住院费用的影响因素。

第二章 分层线性模型的原理

2.1 分层线性模型的设定

2.1.1 一般的分层线性模型

以两层线性模型为例，假设数据由来自层-1 和层-2 的两部分变量构成，其中层-1 嵌套于层-2。假设Y为模型应变量，$X_1, X_2, ..., X_Q$表示层-1 的特征，为层-1 自变量，$W_1, W_2, ..., W_S$表示层-2 的特征，为层-2 自变量。那么模型由两部分组成，第一部分拟合结果变量Y随层-1 各因素$X_1, X_2, ..., X_Q$变化的状况，与传统回归模型不同的是，截距与斜率不再是一个常数而是一个随机变量。模型的第二部分描述了层-2 每个单位回归方程的截距与斜率依据组水平因素$W_1, W_2, ..., W_S$变化的状况，这样就将来自不同水平因素引起的变化同时纳入到回归方程中进行了分析。那么层-1 的一般模型为：

$$Y_{ij} = \beta_{0j} + \sum_{q=1}^{Q} \beta_{qj} X_{qij} + r_{ij} \tag{2.1}$$

层-2 的一般模型为：

$$\begin{cases} \beta_{0j} = \gamma_{00} + \sum_{s=1}^{S} \gamma_{0s} W_{sj} + u_{0j} \\ \beta_{1j} = \gamma_{10} + \sum_{s=1}^{S} \gamma_{1s} W_{sj} + u_{1j} \\ \text{M} \\ \beta_{Qj} = \gamma_{Q0} + \sum_{s=1}^{S} \gamma_{Qs} W_{sj} + u_{Qj} \end{cases} \tag{2.2}$$

其中$\gamma_{k0}, \gamma_{k1,}\text{L}\ , \gamma_{ks} (k = 0,\text{K}\ ,Q)$是层-2 的系数且被称为固定效应，$r_{ij}$是层-1 的随机效应，而$u_{0j}, u_{1j}, u_{2j}, ..., u_{Qj}$是层-2 的随机效应，并假设

(1) $E(r_{ij}) = 0, Var(r_{ij}) = \sigma^2$