多水平模型及其在人口科学研究中的应用

张风雨　王海东

【提要】　本文从分层整群随机抽样数据、层次数据的分析入手，引入多水平分析的一般线性模型和多水平二分类变量的Logit模型，并对其模型的假设条件予以阐述。对模型的产生背景和软件的研制情况进行了回顾，展望了多水平模型的方法学发展。最后指出了多水平分析方法在生育行为的社区因素分析、计划生育项目的评估、婴儿死亡率及人口迁移等人口学研究领域中的应用。该方法可在分析个体因素模型的基础上同时考虑宏观因素，并可考虑微观个体因素与宏观变量之间的交互作用，使数据分析的深度更进一步提高。

【作者】　张风雨　北京大学人口研究所，人口学专业博士生；王海东　国家计划生育委员会统计司工作。

1.引言

定量社会科学研究，往往要进行大规模的随机抽样调查。基于可操作性和成本效益考虑，调查多采用随机分层整群抽样，即从被研究的人群中，随机抽取一组个体进行调查。如全国性的生育节育调查，可先随机抽取县，然后在被抽中的县中随机抽取村，最后对所抽取村中的全部育龄妇女进行调查。然而不幸的是，一旦调查数据被收集后，分析者却很少考虑同一单位中各观测个体间的相关性，并且通常所使用的统计方法却假定每一观测个体之间是相互独立的。忽视这种“内在相关性”的后果将导致估计参数的标准误变小，会使研究者过高的估计研究结果的显著性（Hobcraft，etc.；1983，Guilkey，1992）。与此同时，在社会科学研究中，也常遇到一些层次结构的数据（hierachical structural data）。如教育领域学生学习成绩的评价研究中，由于学生被分成班级，除有描述个人或家庭特征的变量外，还有描述班级或教师的变量。班级又可存在于不同的学校，进而又可得到描述学校特征的信息。这些特征都可影响到学生的学习成绩，并且数据形式表现为层次结构。对层次数据的处理，常见的方法有两种：一是将高层次变量（如教师或班级）分解到个体水平（学生），同一班级内的学生，高层次变量的取值相同，然后以个体为单位进行分析，因此，违反了经典统计学所要求的观测之间独立这一基本假定；另一种方法是将个体变量汇总成高层变量，以高层单位进行分析，这样将会导致个体观测数据信息的损失，往往会高估计变量之间的关系(Bryk and Raudenbush，1992）。此外，社会科学的行为研究中，某些行为或现象的发生，除与个体特征有关外，还受周围环境的影响。如育龄妇女的节育生育行为，不但受个人或家庭经济的影响，同时还受社区环境特征的影响（宋瑞来，1993；田雪原，1993；Hermalin，1986），因此，在进行此类研究时，不但要分析个体变量还要分析社区变量的作用。

2.多水平一般线性模型

2.1 定义 假定两水平的观测数据，即个体自变量（微观）水平（水平-1）和环境（宏观）水平（水平-2），并且因变量个体水平的值取决于个体和周围的环境特征（以下将社区、环境和宏观水平，个体和微观水平视为同一术语）。为简化起见，还假设只有一个个体水平的因变量Y，一个个体水平的自变量I_1和一个环境自变量E_1，则两水平的线性模型可定义为：

水平-1：
$$Y_{ij}=\beta_{0j}+\beta_{1j}I_{1ij}+\varepsilon_{ij} \qquad (2.1)$$

水平-2：
$$\beta_{0j}=\eta_{00}+\eta_{01}E_{1j}+\alpha_{0j} \qquad (2.2)$$
$$\beta_{1j}=\eta_{10}+\eta_{11}E_{1j}+\alpha_{1j} \qquad (2.3)$$

这里，j=1，…J表示环境，i=1，……I表示第j环境中个体观测，并且$\eta_j\geqslant 2$，α_{0j}、α_{1j}为宏观水平上的误差项。

其基本思想是：同一环境中的个体变异较小，不同环境中的个体变异较大，因此不同环境中的系数β_{0j}和β_{1j}应有不同，即式（2.1）中的系数β_{0j}和β_{1j}受环境变量的影响。

将式（2.2）、（2.3）代入式（2.1），有：

$$Y_{ij}=\eta_{00}+\eta_{01}E_{1j}+\eta_{10}I_{1ij}+\eta_{11}E_{1j}I_{1ij}+(\alpha_{0j}+\alpha_{1j}I_{1ij}+\varepsilon_{ij}) \qquad (2.4)$$

在统计学中，式（2.4）又称作为混合模型（mixed model）或协方差分解模型（covariance component model），它包含固定系数（η）和随机系数（α），随机系数随环境的不同而变化。同时也可看出，方程式（2.4）也较为直观，它由环境变量E_{1j}、个体变量I_{1ij}、两变量的交互项$E_{1j}I_{1ij}$和误差项（$\alpha_{0j}+\alpha_{1j}I_{1ij}+\varepsilon_{ij}$）四部分组成。

2.2 模型的基本假定 象一般线性模型一样，多水平线性模型假定误差项：

（1）ε_{ij}是正态分布误差项，服从于均数为0，方差为σ^2，I_n的正态分布。α_{0j}服从于N（0，γ_{00}）的分布，α_{1j}服从于N（0，γ_{11}）的分布。γ_{00}、γ_{11}分别为α_{0j}、α_{1j}的方差。

（2）误差项之间的协方差为：$Cov(\alpha_{0j},\alpha_{1j})=\gamma_{01}$，$Cov(\varepsilon_{ij},\alpha_{0j})=0$，$Cov(\varepsilon_{ij},\alpha_{1j})=0$。

（3）式（2.1）-（2.3）中所有误差项与自变量不相关。

可以看出，式（2.4）并不是普通最小二乘法假定的典型线性模型。因为基于最小二乘法的系数估计和准确的假设检验要求误差项独立、正态分布，并具有不变的方差。而这里的误差项（$\alpha_{0j}+\alpha_{1j}I_{1ij}+\varepsilon_{ij}$）较为复杂，误差项依赖于不同的环境，因为同一环境j内的α_{0j}，α_{1j}是不变的。同时，该误差项方差不齐。因为（$\alpha_{0j}+\alpha_{1j}I_{1ij}+\varepsilon_{ij}$）既依赖于$\alpha_{0j}$、$\alpha_{1j}$，随环境的变化而变化，也依赖于个体变量$I_{1ij}$而变化。尽管该模型不适合用最小二乘法回归分析，但可用迭代最大似然估计模型。

2.3 几种主要的简化线性模型 根据不同的条件，可将多水平线性模型进一步简化，常见的简化模型有：

（1）固定效应模型（fixed-effect model）。若α_{0j}，α_{1j}规定为0，则式（2.4）变成：

$$Y_{ij}=\eta_{00}+\eta_{01}E_{1j}+\eta_{10}I_{1ij}+\eta_{11}E_{1j}I_{1ij}+\varepsilon_{ij} \qquad (2.5)$$

这是具有交互项$E_{1j}I_{1ij}$的一个多重线性回归模型，则I_1对Y的效应为：

$$\partial y/\partial_1=\eta_{10}=\eta_{11}E_1 \qquad (2.6)$$

因此，当变量I_1和E_1之间具有交互作用时，I_1的效应依赖于环境变量E_1的取值。

（2）随机效应方差模型（random effects ANOVA）。这是式（2.4）的一个特殊情况，即除环境变量的常数项η_{00}外，所有的回归系数均规定为0。

$$Y_{ji}=\eta_{00}+\alpha_{0j}+\varepsilon_{ji} \qquad (2.7)$$

这里，η_{00}为总均数，α_{0j}是环境变量的随机变异，ε_{ji}是个体变量的随机变异。

（3）相加性方差分解模型（additive variance component model）。式（2.4）是假定模型具有相乘性（multiplicative model）。若对环境变量而言，不考虑环境变量对个体变量的交互作用，即模型不具备相乘性，则模型为：

$$Y_{ji}=\eta_{00}+\eta_{01}E_{1j}+\eta_{10}I_{1ji}+(\alpha_{0j}+\varepsilon_{ji}) \qquad (2.8)$$

（4）随机系数模型（random coefficient regression）。若将环境变量E_1的系数规定为0（$\eta_{01}=\eta_{11}=0$），则方程式2.4可简化为：

$$Y_{ji}=\eta_{00}+\eta_{10}I_{1ji}+(\alpha_{0j}+\alpha_{1j}I_{1ji}+\varepsilon_{ji}) \qquad (2.9)$$

在无适当环境信息可利用时，这是一个比较有用的模型。事实上，当进行分层随机抽样时，没有收集相应的环境变量数据，并且用传统的统计技术分析个体观测数据时，测量个体间的独立性不满足时，即可用此模型来分析个体变量量间的关系，并估计未被观测的环境变量的效应。

关于多水平线性模型的参数估计与假设检验，可参阅有关专著（Bryk and Raudenbush，1992）。

3。二分类变量的Logit回归模型

社会科学中，常常要研究某些行为或事件的发生。由于测量的限制，只能对因变量进行定性测量，得到的数据多为二分类，即取值为0或1。因此二分类因变量的Logit模型得到了较为广泛地应用。

类似线性模型，假定有个体自变量I_1和环境自变量E_1，P_{1j}为第j环境中i个体的某事件发生的概率，依据多水平模型的定义有：

水平-1：
$$\ln\left(\frac{p_{ji}}{1-p_{ji}}\right)=\beta_{0j}+\beta_{1j}I_{1ji} \qquad (3.1)$$

水平-2：
$$\beta_{0j}=\eta_{00}+\eta_{01}E_{1j}+\alpha_{0j} \qquad (3.2)$$

$$\beta_{1j}=\eta_{10}+\eta_{11}E_{1j}+\alpha_{1j} \qquad (3.3)$$

于是，有：

$$\ln\left(\frac{p_{ji}}{1-p_{ji}}\right)=\eta_{00}+\eta_{01}E_{1j}+\eta_{10}I_{1ji}+\eta_{11}E_{1j}I_{1ji}+(\alpha_{0j}+\alpha_{1j}I_{1ji}) \qquad (3.4)$$

此称为随机效应的logit模型，因在水平-2模型中考虑了随机误差项。否则，$\alpha_{0j}=\alpha_{1j}=0$，式（3.4）就变为固定效应的logit模型（Mason，1986；Curtis，etc.，1993）：

$$\ln\left(\frac{p_{ji}}{1-p_{ji}}\right)=\eta_{00}+\eta_{01}E_{1j}+\eta_{10}I_{1ji}+\eta_{11}E_{1j}I_{1ji} \qquad (3.5)$$

若在随机模型中仅考虑未被观测的环境效应时，随机模型可变为：

$$\ln\left(\frac{p_{ji}}{1-p_{ji}}\right)=\eta_{00}+\eta_{01}E_{1j}+\eta_{10}I_{1ji}+\eta_{11}E_{1j}I_{1ji}+\mu_j \qquad (3.6)$$

在正态logistic模型中，假定μ_j服从于$N(0,\sigma^2)$正态分布，将μ_j标准化得到：

$$\ln\left(\frac{p_{ji}}{1-p_{ji}}\right)=\eta_{00}+\eta_{01}E_{1j}+\eta_{10}I_{1ji}+\eta_{11}E_{1j}I_{1ji}+\sigma v_j \qquad (3.7)$$

这里，v_j服从于标准正态分布，σ是μ_j的正态分布标准差。$\sigma=0$，表示不同社区环境之间无变异，即社区环境无效应，且同一社区内的不同个体间无相关关系。

在结果的陈述和解释时，logit模型可按普通线性回归的形式来表述和解释回归系数，如某自变量变化一个单位，其因变量logit（p）变化一个相应的系数倍。然而这种解释的直观意义不明确。流行病学和公共卫生领域中，在研究某事件如疾病的发生或死亡时，往往以将系数转换为危险比（Odds Ratio$=e^{\beta}$）的方式来表述和解释，即某自变量变化一个单位，某事件（因变量）发生或不发生的可能性增加（或减少）e^{β}倍。这对大多数不熟悉危险比概念的社会科学研究者来说，不太容易理解。再者，社会科学研究中，尤其是与政策相关的研究，往往要根据分析结果了解在某自变量取值一定时所研究事件发生的概率有多大，因此通常需要根据分析结果计算出其事件发生或不发生的模拟概率水平（Diamond，1995）。然而，随机logit模型在危险比或模拟概率计算时，与传统的logit模型有所不同。

$$\hat{p}=\frac{\exp(\eta_{00}+\eta_{01}E_{1j}+\eta_{10}I_{1ji}+\eta_{11}E_{1j}I_{1ji}+\sigma_u\cdot s)}{1+\exp(\eta_{00}+\eta_{01}E_{1j}+\eta_{10}I_{1ji}+\eta_{11}E_{1j}I_{1ji}+\sigma_u\cdot s)} \qquad (3.8)$$

这里，s为均数上下标准差的个数，σ_u为所估计的未观测到的环境效应的标准差。根据这这个公式，可计算在其它变量不变的情况下，每个自变量取值的预测模拟概率（Curtis，etc.，1993）。

4.方法学发展及应用软件

多水平模型可追溯到在统计学中被称作随机效应的线性模型（Elston & Grizzle，1962）。它假定模型中的自变量是受其它因素影响的随机变量。70年代，统计学家开始研究这种模型的统计方法，并引入层次线性模型这一术语，同时提出这种模型的相应参数估计方法，并提出这种方法在分析层次结构数据的可能性。此后有人将此方法应用于增长的研究以及多水平结构的横断面数据分析(Steenio，ect.，1983；Mason，1983)。1985年Wong等人提出了多水平的Logistic回归模型及其经验Bayes估计的方法（Wong，etc.，1985）。1994年McDonald提出带潜变量（latent variable）的两水平路径分析模型，同时Muthen也提出了多水平的协方差结构模型(Covariance strructure analysis，也称作线性结构模型LISREL)（McDonald，1994；Muthen，1994）；此外，也有人指出可将多水平模型用于比例风险模型（Wang，1995）。

尽管可用常用统计软件去估计一些多水平模型，但已有许多专用统计软件包来帮助拟合多水平模型。较常见的有HLM、VARCL、GENMOD和ML3四种软件包（Kreft，etc.，1990），其中功能较强大的软件为ML3和GENMOD。最近，ML3已升级到MLn。在MLn软件中，除可拟合多水平一般线性模型和二分因变量的logit模型外，还可拟合多水平广义线性模型，并将模型拟合扩展到n水平（Rasbash and Woodhouse，1995）。

5. 在人口科学研究中的应用

自60年代Becker用微观经济学和新家庭学的理论研究生育水平的决定因素以后，微观生

育行为分析模式曾一度盛行。而大量的实证分析，特别是对发展中国家的生育行为问题的实证分析结果却另人失望。社会人口学家开始从社区特征的角度研究生育率的变化，并强调以现实地区差异为前提。然而，这些研究都是采用微观研究或宏观研究分离的方法。于是人们开始在生育率的研究中将宏观与微观研究相联系，并在首次世界生育率的研究中，于1982年对7个国家进行了社区特征与生育率关系的抽样调查。1983年在英国伦敦召开的“世界生育率调查中的社区数据收集与分析”研讨会上，收到了许多用多层次模型将社区和个体特征结合起来分析生育率决定因素的文章（Caterline，1985）。这次会议不仅引起了人口学界对社区与生育率关系的重视，同时也扩大了多水平模型的应用（Entwisle and Hermalin，1984）。1985年Wong和Mason提出了层次Logistic回归模型，并用经验Bayes估计的方法分析了世界生育率调查数据（将个体视为微观观测，国家视为宏观观测）（Wong and Mason，1985）。1986年Hermalin和Mason分别将多水平分析方法的理论思想和一个logit分析的例子写入联合国的出版物中（Hermlin，1986；Mason，1986），大大推动了多水平模型在生育节育研究中的应用（Wang，1987）。1990年美国国际发展署资助北卡罗莱纳大学开展了一项长达5年的人口项目评估的研究，其主要目的在于改善目前的计划生育项目评估方法，包括评估的设计、测量、数据收集以及分析方法（Tsui and Hermalin，1993）。该项目组提出将多水平模型作为人口项目评估的主要方法之一，用于评价计划生育项目对发展中国家的生育水平下降的影响（Bucker，etc.，1995），并于1994和1995年资助美国夏威夷东西方中心举办了“计划生育和卫生干预项目评估中的多水平分析方法”研讨班。进一步推广该方法在人口学研究中的应用（Beegle，1994；Oliver，1995）。

在婴幼儿死亡率的研究中，有学者用多水平模型分析了个体和社区效应对婴儿死亡率的效应（Rosenzweig & Schultz，1982）。而后，许多学者也注意到婴幼儿的死亡具有家庭聚集性，并提出了相应的处理方法（Guo，1993）。Curtis and Diamond用水平logit模型分析了巴西出生间隔对婴幼儿死亡率的影响。在分析中将家庭效应视为第二水平，估计出未被观测的家庭效应对婴幼儿死亡率的影响。在分析中将家庭效应视为第二水平，估计出未被观测的家庭效应对婴幼儿死亡具有显著影响（Curtis and Diamond，1993）。

迁移作为一种个人行为，不但受个体特征的影响，同时还受社区因素的影响。Hugo首先提出迁移研究的社区影响（Hugo，1985）。祝俊明用多水平logit模型研究了中国广东农村迁移的影响因素，他提出个人迁移不但受社区环境的影响，还受家庭因素的影响。此外他还分析了家庭迁移的社区影响（Zhu，1995）。

人口事件作为一种行为现象，除受个人特征的影响外，还受家庭、社区等特征的影响，如生育节育行为、婴儿死亡、迁移等。此外，在妇女地位与生育率的研究中，除可分析表示个体妇女地位因素外，还可用多水平分析研究一些反映社区妇女地位的因素。因此，多水平分析模型应用必将会大大提高定量人口学的研究质量。此外，多水平分析还可用于研究不同观察单位内的增长变化和Meta—analysis等其它情况（Bryk and Raudenbush，1992）。随着定量人口学研究的深入，多水平模型在人口学研究领域的应用将更为广泛。

（致谢：本文的初始意念来源于北京大学人口研究所人口与生物统计学教授涂平博士，并感谢他提供有关资料。此外，本人还感谢1995年夏季夏威夷东西方中心“多水平分析研讨班”的教授们，他们是：Amy Ong Tsui、David Guilkey、Ian Diamond和Minjia Kim Choe博士）。

参 考 文 献

1 Beegle, Kathleen (1994), The quality and evaluation of family planning services and contraceptive use in Tanzania. Working paper at Michigan State University.

2 Bertrand, Jane T., etc (1994), Handbook of Indicators for Family Planning Program Evaluation; The EVALUATION Project, University of North Carolina at Chapel Hill.

3 Bryk, A.S.& S.W. Raudenbush (1992), Hierarchical Linear Models; Application and Data Anlysis Methods; Sage Publications.1-3.

4 Bucker, B. C. and Amy O. Tsui, etc. (1995), A Guide to Methods of Family Planning Program Evaluation, 1965-1990, The EVALUATION Project, University of North Carolina at Chapel Hill.

5 Casterline, J.ed. (1985), The Collection and Analysis of Community Data, Proceedings of the WFS Seminar on the Collection & Analysis of Data on Community and Institutional Factors, London, 1983.

6 Curtis,etc. (1993), Birth interval and family effects on postneonatal mortaliy in Brazil, Demography, Vol.30, No. 1; 33-44.

7 Diamond, I. (1995), Multilevel data. Handout at the Workshop on Multilevel Analysis on the Evaluation of Family Planning Program and Health Intervention, East-West Center, Hawaii, June, 1995.

8 Elston & Grizzle (1962), Estimation of Time Response Curve and their Confidence Bands, Biometrics, 18, 148-159.

9 Entwisle, B., etc. (1984), A multilevel model model of family planning availability and contraceptive use in rural Thailand, Demography, Vol. 21, No. 4; 559-574.

10 Guilkey, D.K. (1992), Community effects in demographic and health survey. Carolina Population Center, University of North Carolina.

11 Guo, Chuang (1993). Use of sibling data to estimate family mortality effects in Guatemala, Demography, Vol.30, No.1; 15-32.

12 Hermalin, A.I. (1986), The multilevel approach; theory and concepts.见; U.N.Addendum to Manual IX, 15—24.

13 Hobcraft, J. N. etc. (1983), Child-spacing Effects on Infant and Early Child Mortality, Population Index 49; 585-618.

14 Hugo (1985), The nature of community effects in the study of migration an their empirical investigation, In John Casterline, ed., The Collection and Analysis of Community Data, Proceedings of the world Fertility Survey Seminar on the Collection and Analysis of Data on Commuity and Institutional Factors, London, June, 1983.

15 Kreft, I.G., etc. (1990), Comparing four different statistical packages for hierarchical linear regression; Genmod, HLM, ML2, and VARCL (Statistics Series No. 50), Univ.of California at Los Angeles.

16 Mason, W.M. (1986), The multilevel approach; illustrative example.见; U.N. Addendum to Manual IX, 24-31.

17 McDonald, R.P. (1994), The Bilevel reticular action model for path analysis with

latent variables, Sociological Methods and Research,22: 399-413.
18 Muthen,B.O. (1994),Multi-level covariance analysis, Sociological Methods and Research, 22: 376-398.
19 Oliver, Raylynn (1995), Contraceptive Use in Ghana: The role of services availability, quality, and price, Living Standards Measurement Study,Working Paper No. 111, World Bank.
20 Rasbash, Jon & G.Woodhouse (1995), MLn Command Reference, V1.0. Multilevel Models Project,Institute of Education, Univ.of London.
21 Rosenzweig, M.and T.P.Schultz (1982),Child morality and fertility in Colombia: individual and community effects, Health Policy and Education 2: 305-348.
22 宋瑞来(1993).社区发展与生育率转变：世界生育率调查有关研究的评述.中国人口科学,第2期：14-19.
23 Strenio, J.L., etc. (1983), Empirical Bayes estimation of individual growth curve parameters and their relationship to covariates, Biomesrics, 39: 71-86.
24 田雪原(1991)."中观"人口控制与社区综合发展. 中国人口科学, 1: 1-6.
25 Tsui, A.O.& A.Hermalin (1993), Improving the effectiveness of family planning programs by improving evaluation capabilities, Paper Presented at the 1993 IUSSP General Conference, Montreal, Canada.
26 Wang, Feng (1987), China's Reproductive Revolution: Individual and Community Determinants of Fertility Variation in Hebei, China (Dissertation), The University of Michigan.
27 Wang, Jichuan (1995), Multilevel Analysis, The handout at the methodogical work shop of NIDA's cooperative agreement project, Vienna, VA, May 5, 1995.
28 Wong, G. Y.& W.M.Mason (1985), The Hierarchichical logistic regression model for multilevel analysis, J.A.S.A.Vol.80, No.391: 513-524.
29 Zhu, Junming (1995), Multilevel analysis of rural migration in Guangdong, China. Paper prepared for the 1995 PAA Annual Meeting,April 6-8,San Francisco, U.S.A.

（本文责任编辑 徐 莉）