Statistics and Application 统计学与应用, 2021, 10(1), 173-182
Published Online February 2021 in Hans. http://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2021.101017

Hans汉斯

基于嵌套结构的分层线性回归模型的统计推断

周梦雨[1]，田茂再[2,3*]

[1]兰州财经大学统计学院，甘肃 兰州
[2]中国人民大学应用统计科学研究中心，北京
[3]中国人民大学统计学院，北京
Email: *mztian@ruc.edu.cn

收稿日期：2021年1月25日；录用日期：2021年2月19日；发布日期：2021年2月26日

摘　要

通常在处理模型假设检验的问题时，统计推断是通过样本数据的观测信息来推断总体的主要方法，本文提出基于嵌套结构的分层线性回归模型的系数向量诊断方法，对于分层线性回归的第一层模型系数诊断主要利用传统的线性嵌套回归模型*F*检验进行统计推断。该论文的创新之处在于对分层线性回归模型的第二层系数进行统计诊断，利用嵌套多元线性回归模型推广到具有嵌套结构的分层线性回归模型中，主要构建分层线性回归模型似然函数比值来构造检验统计量。通过高校数学成绩分层数据进行分析，来验证该方法的有效性和可行性。

关键词

分层线性模型，嵌套模型，似然比，统计推断

Statistical Inference of Hierarchical Linear Regression Model Based on Nested Structure

Mengyu Zhou[1], Maozai Tian[2,3*]

[1]School of Statistics, Lanzhou University of Finance and Economics, Lanzhou Gansu
[2]Scientific Research Center of Applied Statistics, Renmin University of China, Beijing
[3]School of statistics, Renmin University of China, Beijing
Email: *mztian@ruc.edu.cn

Received: Jan. 25th, 2021; accepted: Feb. 19th, 2021; published: Feb. 26th, 2021

*通讯作者。

文章引用: 周梦雨, 田茂再. 基于嵌套结构的分层线性回归模型的统计推断[J]. 统计学与应用, 2021, 10(1): 173-182.
DOI: 10.12677/sa.2021.101017

Abstract

Generally, when dealing with the problem of model hypothesis testing, statistical inference is the main method to infer the population through the observation information of sample data. In this paper, the coefficient vector diagnosis method of Hierarchical Linear Regression Model based on nested structure is proposed. For the first level model coefficient diagnosis of hierarchical linear regression, the traditional F-test of linear nested regression model is used for statistical inference. The innovation of this paper lies in the statistical diagnosis of the second layer coefficient of Hierarchical Linear Regression Model. The nested multiple linear regression model is extended to the Hierarchical Linear Regression Model with nested structure. The likelihood function ratio of Hierarchical Linear Regression Model is mainly constructed to construct the test statistics. The effectiveness and feasibility of this method is verified by the hierarchical data analysis of college mathematics scores.

Keywords

Hierarchical Linear Model, Nested Model, Likelihood Ratio, Statistical Diagnosis

1. 引言

在实际观测数据中，数据往往具有分层结构，具有嵌套和分层结构的数据均可以使用多层线性模型进行分析，多层线性模型在处理缺失数据时不影响参数估计精度的特征，相比于多元线性回归模型，处理纵向数据有较大优势，同时对于重复测量的次数和时间跨度没有严格要求。对于分层线性回归模型的描述，不同的文献中名称是有差异的，在社会学研究中，通常被称为多层线性模型(Goldstein [1], 1997)，在生物计量学应用中，称为随机效应模型(Laird [2], 1982)，在统计学文献中，一般称为协方差成分模型(Dempster [3], 1981)。对分层线性回归模型未知参数一般分为对随机效应和固定效应两部分对应的回归系数和方差协方差部分进行估计，在本文中，主要考虑分层线性回归模型的回归系数诊断问题，对分层线性回归模型第一层回归系数进行回归诊断，即对应的为多元线性回归系数诊断问题，针对分层线性回归模型的第二层系数的回归诊断则有少量的参考文献，本文对分层线性第二层系数的回归诊断主要通过构造分层线性回归模型似然函数比的方法来进行统计推断。在多元线性回归模型中，吴喜之[4] (2016)用线性模型来近似因变量与自变量的线性关系，并给出线性模型中参数估计的方法和关于多自变量系数复合检验过程，谢宇[5] (2013)介绍了多元线性模型中嵌套模型的具体含义，并给出了嵌套模型多个回归系数的联合检验常用的 F 统计量，同时给出二分因变量嵌套模型的模型评价方法，但并没有给出分层线性嵌套模型的检验过程，Lindley & Smith [6] (1972)针对复杂的嵌套数据给出了分层线性模型的具体形式，并针对分层线性回归模型中的未知参数给出了具体的估计方法，田茂再[7] (2006)提出了基于 Guass-Seidel 迭代的条件分位分层线性回归模型的算法，解决了分层模型不能全面刻画高维情况下响应变量的条件分布问题，吴密霞[8] (2013)介绍了模型参数的似然比检验，给出了具体的假设过程，刘红云[9] (2005)介绍在追踪数据分析中，对于多层线性模型，可以利用 Wald 检验来对固定部分参数进行显著性检验。

对于回归模型的统计推断推广中，马海强[10] (2008)给出了变系数模型的统计诊断问题，戴林送[11]

(2013)研究了广义泊松回归模型的统计诊断方法，Li & Lee [12] (2019)通过似然函数最大化拟合半参数零膨胀负二项分布回归模型，并通过似然比评价一个假设参数泛函形式的连续协变量效应的充分性，利用数据证明其方法的有效性。费宇[13] (2013)介绍了线性混合模型和广义线性混合模型的统计诊断方法。曾婕等[14] (2017)提出了结合残差、杠杆值和系数变化三者构造诊断统计量来诊断 logistic 回归模型中数据的异常点或强影响点问题。Chown & Ursula [15] (2019)介绍了一种多协变量非参数回归模型的异方差检验方法，利用局部多项式平滑构造残差，设计检测函数验证异方差。晏振等[16] (2016)利用杠杆值抽样后的大数据集来诊断异常点问题。梁晋雯[17] (2020)基于数据删失模型和均值漂移模型构建统计量进行异常点的诊断，研究体积抽样受异常点的影响。

本文基于线性回归模型的最小二乘方法得到参数的估计值，线性回归嵌套模型主要讨论增加一个变量的回归系数是否显著，与原本的回归系数是否有实质的改变，以此确定变量保留与否。基于线性回归模型中变量讨论的基础上，进一步讨论分层线性回归模型，这也是本文主要的创新点，利用具有嵌套结构的分层线性回归模型的似然函数的比值去判断分层线性嵌套模型假设的合理性。同时本文是基于同方差的基础上讨论的，对线性回归模型中异方差情况也有其他文章做过，这里不再详细说明。

在本文中，根据多元线性嵌套模型的含义，针对多元线性嵌套模型主要利用 F 统计量来检验限制性模型与非限制性模型的显著性问题，并通过波士顿房价来检验统计量的有效性，同时根据多元线性模型的嵌套结构，针对分层线性模型的嵌套结构进行合理的假设检验，主要通过构造具有嵌套结构的分层线性回归模型的似然函数的比值服从卡方分布，通过给定的拒绝域，来判断限制性分层回归模型和非限制性分层回归模型的显著性问题。

2. 线性嵌套模型的统计推断

若我们只考虑分层线性模型的第一层系数的回归诊断，那分层线性模型就可直接理解为多元线性回归模型的回归诊断。为了更好的进行第一层模型统计诊断，现在将第一层模型进行形式上的变换，其实际意义并无影响。

我们考虑数据存在随机线性模型的一般形式，线性模型意味着假定因变量 y 和自变量 x 之间的关系可以用线性关系来近似(吴喜之[4] 2016)：

$$Y_i = X_i^{\mathrm{T}}\beta + \varepsilon_i, i = 1, \cdots, n$$

其中，Y_i 为 $n\times 1$ 的观测向量，X_i 为 $n\times p$ 已知的向量矩阵，β 为待估计的未知参数，ε_i 是模型所无法描述的随机误差项。通常情况下，随机误差 ε_i 满足 3 个假设：1) $E(\varepsilon_i)=0$；2) $Var(\varepsilon_i)=\sigma^2$；3) $Cov(\varepsilon_i,\varepsilon_j)=0$，$i\neq j$。经常情况下，人们把未知的 σ^2 假设为相等，若这一假定不成立，则称线性回归模型存在异方差性，在存在异方差性情况下用传统的最小二乘法估计模型参数，得到的参数估计量不是有效估计量，这里不再具体介绍异方差情况下的回归诊断情况，下面线性回归诊断是基于同方差情况下介绍的。

对于线性回归模型的待估计参数 β 常用的估计参数的方法是普通最小二乘法，其目的是使得 $\varepsilon = y - x\beta$ 达到最小，即 $S(\beta)=(y-x\beta)^2$ 达到最小。即对未知参数 β 求偏导数，令函数 $S(\beta)$ 为零，可以得到：

$$\hat{\beta} = (X'X)^{-1}X'y$$

如果一个模型中的自变量为另一个模型的自变量的子集或者子集的线性组合，则称两个模型为嵌套模型(谢宇[5] 2013)。一个模型子集或子集的线性组合的模型称为限制性模型，对应的另一个模型称为非限制性模型，限制性模型嵌套在非限制性模型中。

对多元线性回归模型的系数提出假设如下：

$$H0: Y = X_1\beta_1 + X_2\beta_2 + \varepsilon \ \text{ vs } \ H1: Y = X_1\beta_1 + X_2\beta_2 + X_3\beta_3 + \varepsilon$$

对应的检验统计量：

$$F_j = \frac{\left(SSE_{H0} - SSE_{H1}\right)/\left(k-g\right)}{SSE_{H1}/\left(n-k-1\right)}$$

其中，SSE_{H0} 为原假设 $H0$ 对应的残差平方和，SSE_{H1} 为备择假设 $H1$ 对应的残差平方和，也称为限制性模型和非限制模型的残差平方和(谢宇[5] 2013)。这里 k 对应备择假设模型所包含的回归系数的数量，则 $n-k-1$ 对应备择假设残差平方和的自由度。其中 $k-g$ 这个自由度增量是备择假设与原假设对应模型之间回归系数个数的差值。

对于给定的显著水平 α，检验的拒绝域为 $F_j > F_\alpha\left(1, n-k-1\right)$。由于原假设去掉部分自变量，所以理论上原假设对应的残差平方和不小于备择假设的残差平方和。

由于这里原假设与备择假设对应模型之间只差一个参数，所以也可以使用 t 检验统计量 $t_{df}^2 = F_{1,df} \sim t\left(n-k-1\right)$，对于给定的显著性水平 α，检验的拒绝域为 $\left|t_j\right| > t_{\alpha/2}\left(n-k-1\right)$。其中 F 统计量的第一自由度为 1，这时既可以使用 F 统计量也可以使用 t 统计量。

对于上述嵌套结构的线性回归模型，同时也可以使用判定系数增量来解释回归模型拟合优度的问题，其具体为非限制性模型的判定系数减去限制性模型的判定系数(谢宇[5] 2013)，详细过程会从以影响波士顿房价因素的模拟中体现，通过构建多元限制性模型与非限制性模型的原假设与备择假设，来考虑加入每间住宅的平均房间数 x_1 这个自变量，通过方差分析来决定对这一变量是否保留的问题。

下面模拟数据来源于波士顿房价的部分数据，以自住房屋房价中位数为因变量 y，以每间住宅的平均房间数为 x_1，波士顿的五个就业中心加权距离为 x_2，城镇的学生与教师比例为 x_3，我们构造的线性嵌套模型检验，原假设为限制性模型，以 x_2，x_3 为自变量，备择假设为非限制性模型，加入每间住宅的平均房间数这一变量，以 x_1，x_2，x_3 为自变量，具体表达为：

$$H0: y = \beta_2 x_2 + \beta_3 x_3 + \varepsilon \quad \text{vs} \quad H1: y = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$$

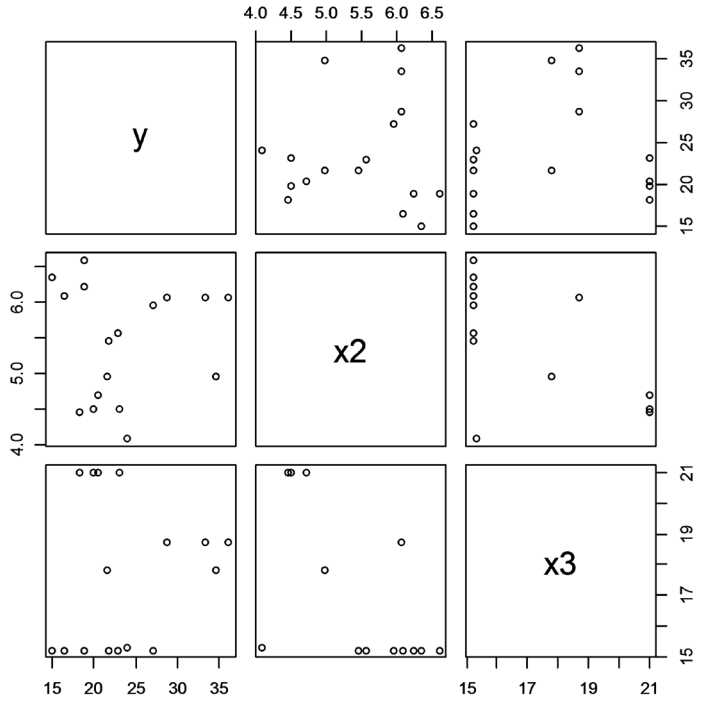

Figure 1. Scatter plot of restricted model
图 1. 限制性模型的散点图

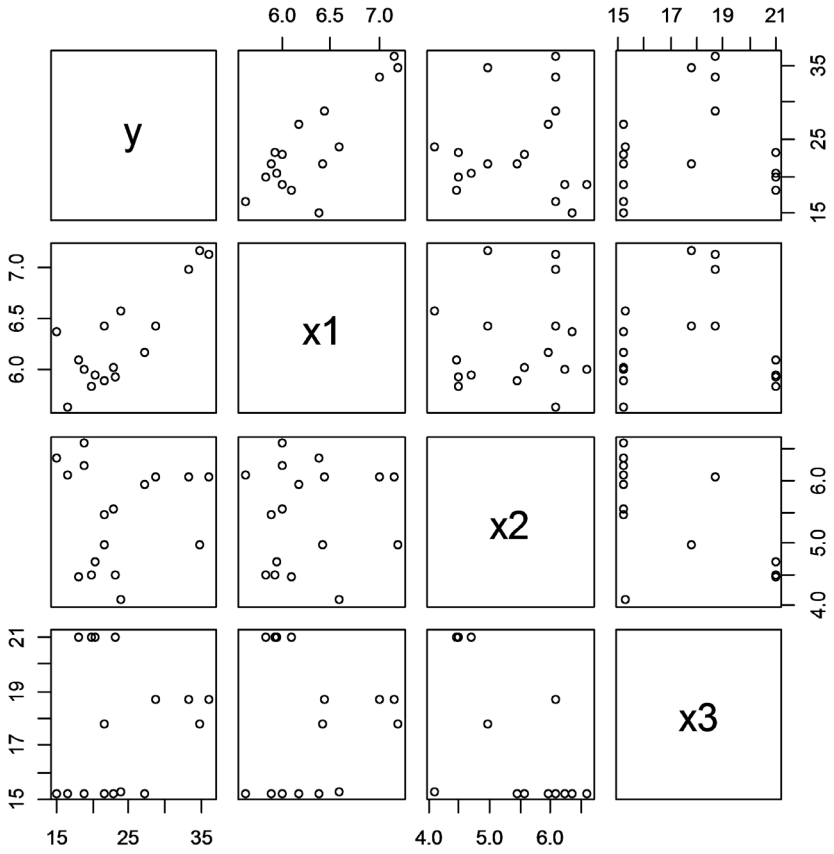

Figure 2. Scatter plot of unrestricted model
图 2. 非限制模型的散点图

从图 1 和图 2 可以看出在影响波士顿房价的因素中，每间住宅的平均房间数 x_1，波士顿的五个就业中心加权距离 x_2 以及城镇的学生与教师比例 x_3 与自住房屋房价中位数 y 有明显的线性关系。针对这种多元线性回归模型，我们构造限制性模型与非限制性模型，构造统计量的进行检验，具体可从下面的方差分析表中看出。

Table 1. Analysis of variance of restrictive model and unrestricted model
表 1. 限制性模型与非限制性模型的方差分析表

模型	残差平方和	残差平方和的自由度	嵌套模型 F 值	嵌套模型 F 值对应 P 值	嵌套模型 t 值	判定系数
$y \sim x_2, x_3$	603.18	14				0.067
			28.255	0.000	5.316	
$y \sim x_1, x_2, x_3$	190.09	13				0.706

根据表 1 方差分析表可以得到，嵌套模型对应的检验统计量计算得到 F 值为 28.255，从而根据嵌套模型 F 值对应的 P 值是明显小于显著性水平 0.05，所以拒绝原假设，接受备择假设，即非限制性模型通过了显著性检验。同时这里根据 F 值得到的 t 值为 5.316 是大于拒绝域 2.160。同时也可以根据嵌套模型计算判定系数的增量，可以看出，当限制性模型加入自变量 x_1，判定系数 R^2 增加，意味着更多的平方和被非限制性模型所解释。

在多元线性回归模型统计推断中，一般包括两个方面的内容：其一是对回归模型的整体检验，另一个是对回归系数的检验。多元线性回归与一元线性回归的方差分析大致相同，对于多元线性嵌套模型，我们常常利用构造 F 统计量来检测限制性模型的假设，若嵌套模型的原假设与备择假设的回归自变量只差一个回归系数，也可以使用 t 统计量来检验，但是对于两个不嵌套的模型是不能使用 F 统计量检验。同时对于嵌套模型，限制性模型自变量不仅可以是非限制性模型自变量的子集，而且非限制模型自变量也可以是限制性模型中自变量的线性组合。

3. 具有嵌套结构的分层线性模型统计推断

在统计数据过程中，数据往往存在分层结构，例如研究高校间不同学生的学习情况，或用于研究国家经济发展的差异如何与成人教育程度相互作用，或研究临床药物的治疗方法的差异等，这些情况中存在嵌套问题的研究，分层线性回归模型给出了良好的模型结构。分层线性模型是由 Lindley 和 Smith [6] (1972)提出的，作为对线性模型的贝叶斯估计的重要贡献，同时对复杂的嵌套结构数据给出了通用的分层线性模型的形式。

这里以两层数据模型为例，给出两层分层线性模型的具体形式，假设有 (X,Y) 的一组独立同分布观测值 $\{(X_1,Y_1),(X_2,Y_2),\cdots,(X_n,Y_n)\}$，其中 Y_i 是实数被解释变量的值，X_i 是已知的 $1\times d$ 维第一层的解释变量，β_i 是未知的 $d\times 1$ 维系数向量，满足第一层模型(田茂再等[7] 2006)：

$$Y_i = X_i\beta_i + r_i,\quad r_i \sim N(0,\sigma^2)$$

其中，r_i 是 *i.i.d* 不可观测随机效应变量，假定与解释变量独立，并服从均值为 0，方差为 σ^2 的正态分布。

在第二层模型中，第一层模型中系数向量作为被解释变量，γ 为固定效应向量，w_i 为第二层已知的解释变量矩阵：

$$\beta_i = w_i\gamma + u_i,\quad u_i \sim N(0,T)$$

其中，u_i 是第二层 $d\times 1$ 维随机效应向量，假定与第二层解释变量和 r_i 独立，并服从均值向量为 0，协方差为 T 的多元分布。

将第二层模型带入第一层模型中，得到下列形式(Raudenbush 和 Bryk [18] 1992)：

$$Y_i = X_i w_i\gamma + X_i u_i + r_i,\quad i=1,\cdots,n$$

上述模型也称为线性混合模型，可以用于分析处理纵向数据和面板数据等各类重复测量数据，相比线性模型，对观测值的协方差矩阵可以有更灵活的设定，同时对于随机效应部分给出更方便和合理的假设。

关于对二分因变量进行嵌套模型分析的统计方法(谢宇[5] 2013)，其目的在于估计和预测成功或失败的概率是否受到协变量的影响。二分因变量解释为其取值只有两种可能，也通常称为 0-1 变量，常用于处理二分因变量的模型为 *logit* 模型。针对存在嵌套关系的二分因变量，常通过进行对数似然比检验来判断模型的拟合优度更佳问题，即具体为两个嵌套模型之间的对数似然比之差构造统计量，其统计量服从 χ^2 分布，相应为的统计量形式为：

$$\Delta G^2 = G_r^2 - G_u^2$$

其中，G_r^2 表示约束模型的对数似然比，G_u^2 为无约束模型的对数似然比，则二分因变量嵌套模型服从的 χ^2 分布对应的自由度为无约束模型的残差自由度与约束模型的残差自由度之差。

对于上述多元线性回归模型的假定检验和二分因变量的嵌套模型检验，扩展到分层线性回归模型下，

对于分层线性模型中未知参数估计主要是估计固定效应的回归系数和随机效应的方差协方差部分，具体可根据 Raudenbush [18] (1992)提出的利用完全数据充分统计量的条件期望代替期望步进行的迭代过程得到，这里不具体讨论参数估计过程，下面内容主要是得出分层线性模型的似然函数。

将第二层模型带入到第一层模型中，Y_i 具有线性混合模型形式，已知第一层随机误差 $r_i \sim N\left(0,\sigma^2\right)$ 和第二层随机误差 $u_i \sim N\left(0,T\right)$，所以线性混合模型 Y_i 服从 $y \sim N\left(xwr,V\right)$，相应的，$V = xTx^{\mathrm{T}} + \sigma^2 I_n$。从而给出分层线性模型的似然函数为：

$$\begin{aligned} L(y) &= \prod_{j=1}^{J}\left[\left(2\pi\right)^{-\frac{n}{2}}\left|V\right|^{\frac{1}{2}}\exp\left\{-\frac{1}{2}\left(Y-Xwr\right)^{\mathrm{T}}V^{-1}\left(Y-Xwr\right)\right\}\right] \\ &= \prod_{j=1}^{J}\left[\left(2\pi\right)^{-\frac{n}{2}}\left|x_j\tau x_j^{\mathrm{T}}+\sigma^2 I_n\right|^{\frac{1}{2}}\exp\left\{-\frac{1}{2}\left(Y-Xwr\right)^{\mathrm{T}}\left|x_j\tau x_j^{\mathrm{T}}+\sigma^2 I_n\right|^{-1}\left(Y-Xwr\right)\right\}\right] \\ &= \left(2\pi\right)^{-\frac{n^2}{2}}\left|x_j\tau x_j^{\mathrm{T}}+\sigma^2 I_n\right|^{\frac{n}{2}}\exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\left(Y-Xwr\right)^{\mathrm{T}}\left|x_j\tau x_j^{\mathrm{T}}+\sigma^2 I_n\right|^{-1}\left(Y-Xwr\right)\right\} \end{aligned}$$

其中，$T = diag\left(\tau,\cdots,\tau\right)$，且 $j = 1,\cdots,J$。

根据上述内容，下面给出具有嵌套结构的分层线性模型的一般假定情况：

零假设情况下：

第一层：$Y_1 = X\beta_1 + \varepsilon_1$，$\varepsilon_1 \sim N\left(0,\sigma_1^2\right)$

第二层：$\beta_1 \sim N\left(w_1 r, T_1\right)$

备择假设情况下：

第一层：$Y_2 = X_1\beta_1 + X_2\beta_2 + \varepsilon_2$，$\varepsilon_2 \sim N\left(0,\sigma_2^2\right)$

第二层：$\begin{pmatrix}\beta_1\\ \beta_2\end{pmatrix} \sim N\left(\begin{pmatrix}w_1 r\\ w_2 r\end{pmatrix}, T_2\right)$

假设 $L_n\left(\theta \mid y\right)$ 是参数 θ 的似然函数，其中 $y = \left(y_1,\cdots,y_n\right)'$ 是一个样本容量为 n 的样本，参数 θ 的参数空间为 Ω，检验问题为 $H0:\theta \in \Omega_0$ vs $H1:\theta \notin \Omega_0$，则统计量定义似然比(吴密霞[8] 2013)为：

$$LR_n\left(y\right) = \frac{\sup_{\theta\in\Omega_0} L_n\left(\theta \mid y\right)}{\sup_{\theta\in\Omega} L_n\left(\theta \mid y\right)}$$

在多元分析过程中，似然比检验是常用的检验方法，统计量利用经典似然函数的比值构造为：$\Lambda\left(x_1,\cdots,x_n\right) = L\left(H0\right)/L\left(H1\right)$。这里我们构造分层嵌套检验的似然比，原假设为限制性分层线性回归模型，备择假设为非限制性分层线性回归模型，从而构造分层线性回归模型的似然比检验统计量。

零假设情况下似然函数：

$$L\left(H0\right) = \prod_{i=1}^{n}\left[\left(2\pi\right)^{-\frac{n}{2}}\left|V_1\right|^{\frac{1}{2}}\exp\left\{-\frac{1}{2}\left(Y_{1i}-X_{1i}w_1 r\right)^{\mathrm{T}}V_1^{-1}\left(Y_{1i}-X_{1i}w_1 r\right)\right\}\right]$$

其中，$V_1 = x_1 T_1 x_1^{\mathrm{T}} + \sigma_1^2 I_n$。

备择假设情况下似然函数：

$$L\left(H1\right) = \prod_{i=1}^{n}\left[\left(2\pi\right)^{-\frac{n}{2}}\left|V_2\right|^{\frac{1}{2}}\exp\left\{\frac{1}{2}\left(Y_{2i}-X_{1i}w_1 r-X_{2i}w_2 r\right)^{\mathrm{T}}V_2^{-1}\left(Y_{2i}-X_{1i}w_1 r-X_{2i}w_2 r\right)\right\}\right]$$

其中，$V_2 = x_1 T_1 x_1^{\mathrm{T}} + x_2 T_2 x_2^{\mathrm{T}} + \sigma_2^2 I_n$。

则构造统计量为：

$$\frac{L(H0)}{L(H1)}=\frac{\prod_{i=1}^{n}\left[(2\pi)^{-\frac{n}{2}}|V_1|^{\frac{1}{2}}\exp\left\{-\frac{1}{2}(Y_{1i}-X_{1i}w_1r)^{\mathrm{T}}V_1^{-1}(Y_{1i}-X_{1i}w_1r)\right\}\right]}{\prod_{i=1}^{n}\left[(2\pi)^{-\frac{n}{2}}|V_2|^{\frac{1}{2}}\exp\left\{-\frac{1}{2}(Y_{2i}-X_{1i}w_1r-X_{2i}w_2r)^{\mathrm{T}}V_2^{-1}(Y_{2i}-X_{1i}w_1r-X_{2i}w_2r)\right\}\right]}$$

$$=\frac{|V_1|^{\frac{1}{2}}\exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(Y_{1i}-X_{1i}w_1r)^{\mathrm{T}}V_1^{-1}(Y_{1i}-X_{1i}w_1r)\right\}}{|V_2|^{\frac{1}{2}}\exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(Y_{2i}-X_{1i}w_1r-X_{2i}w_2r)^{\mathrm{T}}V_2^{-1}(Y_{2i}-X_{1i}w_1r-X_{2i}w_2r)\right\}}$$

其中，分子表示原假设的似然函数最大值，分母表示备择假设下的似然函数最大值，如果统计量的值很大，说明原假设情况的可能性比备择假设情况下的可能性要小，于是，我们有理由认为原假设不成立。

在多层线性模型中，对模型单个自变量参数估计值的统计诊断，可以通过极大似然估计得到固定部分参数估计结果已经对应的标准误，对于固定部分的显著性检验，可以用参数估计值除以标准误，即对应的$\gamma/se(\gamma)$进行(刘红云[9] 2005)。

对于多层嵌套线性回归模型，构造统计量为：

$$\Lambda(x_1,\cdots,x_n)=\frac{L(H0)}{L(H1)}$$

该统计量服从χ^2分布，其自由度等于备择假设参数的个数减去原假设中参数的个数，对于给出的分层线性回归零假设与备择假设情况，这里χ^2分布对应的自由度为1。

在显著性水平α下，其拒绝域为：$W=\{\chi^2<\chi_1^2=c\}$，如果落入拒绝域中，说明统计诊断不显著，则拒绝原假设，接受备择假设非限制性分层线性模型。

4. 数据分析

下面数据来自于160所学校7185名学生数学成绩，采用分层线性回归模型对数据进行分析，这里我们选取其中的部分数据进行嵌套分层线性回归模型分析，对于第一层水平，即学生层面，这里选取MATHACH (学生的数学成绩)作为因变量，即Y_{ij}，FEMALE(学生性别) (1表示女性，0表示男性)，SES (学生社会地位)由学生父母受教育程度、职业和收入合成作为自变量。对于第二层水平下，即学校层面，这里选取MEANSES (包含在水平1数据中，每个学校学生的平均社会地位)，DISCLIM (学科氛围)，SIZE (学校招生人数)作为第二层自变量。这里，构建限制性分层线性模型为：MATHACH作为第j所学校第i个学生因变量Y_{ij}，SES作为第一层水平下第j所学校第i个学生自变量X_{1i}，MEANSES和DISCLIM作为第二层水平下第j所学校的自变量，即w_0和w_1，非限制性分层线性模型构建为MATHACH作为第j所学校第i个学生因变量Y_{ij}，SES和FEMALE作为第一层水平下第j所学校第i个学生自变量X_{1i}和X_{2i}，MEANSES和DISCLIM和SIZE作为第二层水平下第j所学校的自变量，即w_0，w_1和w_2，具体模型构建形式如下：

原假设为：

第一层：$Y_{ij}=\beta_{0j}+\beta_{1j}*SES_{ij}+\varepsilon_{ij}$

第二层：$\beta_{0j}=\gamma_{00}+\gamma_{01}*MEANSES_j+u_{0j}$

$\beta_{1j}=\gamma_{10}+\gamma_{11}*DISCLIM_j$

将第二层模型带入第一层模型得到的混合效应模型为：

$$Y_{ij}=\gamma_{00}+\gamma_{01}*MEANSES_j+\gamma_{10}*SES_{ij}+\gamma_{11}*DISCLIM_j*SES_{ij}+u_{0j}+\varepsilon_{ij}$$

Table 2. Estimation of fixed effects in null hypothesis hierarchical linear model
表 2. 原假设分层线性模型中固定效应估计

固定效应		系数	标准误	*t-ratio*	*d.f.*	*p* 值
β_0	γ_{00}	12.746	0.149	85.705	158	<0.001
	γ_{01}	3.739	0.374	9.987	158	<0.001
β_1	γ_{10}	2.215	0.109	20.403	7023	<0.001
	γ_{11}	0.604	0.112	5.418	7023	<0.001

表 2 经过 6 次迭代对数似然函数变化值达到最小后停止，给出了在原假设情况下固定效应变量的系数估计值和对应的标准误，其中固定效应是通过最小二乘估计得到的，通过 p 值可以看出变量系数估计值都通过了检验，同时分层线性回归模型第一层对应的方差为 36.887。

备择假设为：

第一层：$Y_{ij} = \beta_{0j} + \beta_{1j} * FEMALE_{ij} + \beta_{2j} * SES_{ij} + \varepsilon_{ij}$

$\beta_{0j} = \gamma_{00} + \gamma_{01} * MEANSES_j + u_{0j}$

第二层：$\beta_{1j} = \gamma_{10} + \gamma_{11} * SIZE_j$

$\beta_{2j} = \gamma_{20} + \gamma_{21} * DISCLIM_j$

将第二层模型带入到第一层中得到备择假设的混合效应模型：

$$Y_{ij} = \gamma_{00} + \gamma_{01} * MEANSES_j + \gamma_{10} * FEMALE_{ij} + \gamma_{11} * SIZE_j * FEMALE_{ij} + \gamma_{20} * SES_{ij} + \gamma_{21} * DISCLIM_j * SES_{ij} + u_{0j} + \varepsilon_{ij}$$

Table 3. Estimation of fixed effects in alternative hypothesis hierarchical linear model
表 3. 备择假设分层线性模型中固定效应估计

固定效应		系数	标准误	*t-ratio*	*d.f.*	*p* 值
β_0	γ_{00}	13.344	0.168	79.313	158	<0.001
	γ_{01}	3.627	0.367	9.885	158	<0.001
β_1	γ_{10}	-0.524	0.298	-1.760	7021	0.078
	γ_{11}	-0.001	0.000	-2.603	7021	0.009
β_2	γ_{20}	2.178	0.108	20.111	7021	<0.001
	γ_{21}	0.602	0.111	5.426	7021	<0.001

从表 3 经过 6 次迭代似然函数变化值达到最小，可以看出对备择假设情况下分层线性回归模型中固定效应相对应的系数估计值，且显著性检验大多在显著性水平为 0.05 通过了检验，系数 γ_{10} 在显著性水平 0.01 下没有通过显著性检验，同时第一层水平对应的方差为 36.628。

通过计算可以得到嵌套模型分层线性的似然比结果为 1.001，对应的，在显著性水平 $\alpha = 0.01$ 下，χ_1^2 的值为 0.0002，根据给定拒绝域，所以不能拒绝原假设，故接受原假设，即接受限制性分层线性模型，说明从学校层面上引入学校招收人数变量，从学生层面引入性别这一变量，对高校学生数学成绩没有显著影响。

5. 小结

本文通过多元线性嵌套模型的假设检验过程，提出具有嵌套结构的分层线性回归模型的假设检验，通过分层线性模型的似然函数比值构造检验统计量，来判断分层线性模型对于引入新的变量能否保留给出对应的理论依据，同时通过高校数学成绩数据分析分层线性模型假设检验的可行性和构造的具有嵌套结构似然比统计量的实用性。

参考文献

[1] Courgeau, D. and Goldstein, H. (1997) Multilevel Statistical Models. *Population*, **52**, 1043-1046. https://doi.org/10.2307/1534624

[2] Laird, N.M. (1982) Random Effects Model for Longitudinal Data. *Biometrics*, **38**, 963-974. https://doi.org/10.2307/2529876

[3] Dempster, A.P. and Tsutakawa, D.B.R.K. (1981) Estimation in Covariance Components Models. *Journal of the American Statistical Association*, **76**, 341-353. https://doi.org/10.1080/01621459.1981.10477653

[4] 吴喜之. 应用回归及分类: 基于 R [M]. 北京: 中国人民大学出版社, 2016.

[5] 谢宇. 回归分析[M]. 第 2 版. 北京: 社会科学文献出版社, 2013.

[6] Smith, D.V.L.F.M. (1972) Bayes Estimates for the Linear Model. *Journal of the Royal Statistical Society*, **34**, 1-41. https://doi.org/10.1111/j.2517-6161.1972.tb00885.x

[7] 田茂再, 陈歌迈. 条件分位中的分层线性回归模型[J]. 中国科学 A 辑, 2006, 36(10): 1103-1118.

[8] 吴密霞. 线性混合效应模型引论[M]. 北京: 科学出版社, 2013.

[9] 刘红云. 追踪数据分析方法及其应用[M]. 北京: 教育科学出版社, 2005.

[10] 马海强. 变系数模型的统计诊断和影响分析[D]: [硕士学位论文]. 长沙: 中南大学, 2008.

[11] 戴林送, 林金官. 广义泊松回归模型的统计诊断[J]. 统计与决策, 2013(21): 29-33.

[12] Li, C.S., Lee, S.M. and Yeh, M.S. (2019) A Test for Lack-of-Fit of Zero-Inflated Negative Binomial Models. *Journal of Statal Computation and Simulation*, **89**, 1301-1321. https://doi.org/10.1080/00949655.2019.1577856

[13] 费宇. 线性和广义线性混合模型及其统计诊断[M]. 北京: 科学出版社, 2013.

[14] 曾婕, 胡国治. Logistic 回归模型的统计诊断[J]. 数理统计与管理, 2017, 36(4): 620-631.

[15] Chown, J. and Müller, U.U. (2019) Corrigendum: Detecting Heteroscedasticity in Non-Parametric Regression Using Weighted Empirical Processes. *Journal of the Royal Statistical Society Series B* (*Statistical Methodology*), **81**, 805-806. https://doi.org/10.1111/rssb.12324

[16] 晏振, 戴晓文, 田茂再. 基于杠杆值大数据集抽样的异常点诊断[J]. 数理统计与管理, 2016, 35(5): 794-802.

[17] 梁晋雯, 田茂再. 大数据下基于体积抽样的异常点诊断及估计问题[J]. 数理统计与管理, 2020, 39(2): 223-235.

[18] Bryk, A.S. and Raudenbush, S.W. (1992) Hierarchical Linear Models: Applications and Data Analysis Methods. *Journal of the American Statistical Association*, **98**, 436-450.