Statistics and Application 统计学与应用, 2017, 6(4), 386-395
Published Online October 2017 in Hans. http://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2017.64044

Hans汉斯

The Analysis on Influence Factors of the Transcripts of Mathematics and Advanced Mathematics Based on Hierarchical Linear Models

Chao Qian, Sen Lin, Tong Xue, Xiaoqiang Gao, Kai Zhang

China University of Mining & Technology, Beijing
Email: 891878406@qq.com

Received: Oct. 1st, 2017; accepted: Oct. 15th, 2017; published: Oct. 20th, 2017

Abstract

Transcript of Mathematics in college entrance examination represents an individual's learning situation in high school. At the same time, transcript of Advanced Mathematics represents an individual's learning situation in college. By using hierarchical linear models, this article identifies the relationship between the transcripts of Mathematics and Advanced Mathematics along with some factors that affects student's Advanced Mathematics' level greatly. Some suggestions that aimed at increasing college students' Mathematics level are also being provided. Students come from different provinces and situations, especially educational situations vary in provinces. Taking students' individual differences in consideration as well, it is obvious that the data has been studied nested structure, which indicates that traditional Hierarchical linear models are not capable of analyzing this category of hierarchical models. The hierarchical models adopted in this article have fully considered both provincial and individual differences in order to establish statistical models properly. The method has overcome the traditional linear models' limitation on analyzing nested or hierarchical structure. It establishes appropriate hypothetical relationships both inside and among the provinces, and estimate variables from different hierarchy. As a result, this model is more similar to basic phenomenon in reality and its model interpretation is more reasonable.

Keywords

Transcript of Mathematics in College Entrance Examination, Transcripts of Advanced Mathematics, Hierarchical Linear Models

高数成绩与高考成绩的分层回归模型影响分析

钱　超，林　森，薛　童，高小强，章　恺

文章引用: 钱超, 林森, 薛童, 高小强, 章恺. 高数成绩与高考成绩的分层回归模型影响分析[J]. 统计学与应用, 2017, 6(4): 386-395. DOI: 10.12677/sa.2017.64044

中国矿业大学，北京
Email: 891878406@qq.com

收稿日期：2017年10月1日；录用日期：2017年10月15日；发布日期：2017年10月20日

摘　要

高考数学成绩是学生高中学习状况的体现，高等数学成绩是大学生大学学习状况的体现，本文通过分层线性模型确定大学生高考数学成绩和高等数学成绩的关系及一些对高等数学成绩具有显著影响的因素，并给出了一些提高大学生数学水平的意见。由于学生来自不同省份，各省情况尤其是教育情况迥异，同时学生个体间存在差异，因此所研究的数据具有明显的嵌套结构，传统的线性回归模型无力分析此类分层模型，本文采用的分层线性模型同时考虑了省份间差异及个体间差异进行统计建模，突破了传统的线性模型在分析嵌套结构上的局限性，建立了各省内关系与各省间关系的假设，估计了各个层次上的变化量，故该模型更接近现实生活中的基本现象，模型解释更为合理。

关键词

高考数学成绩，高等数学成绩，分层线性模型

1. 引言

高等数学课程是高等院校理、工科各专业必修的基础课程，它为理工类专业后继课程提供基本的数学基础知识和数学方法，为培养学生的思维能力、分析和解决实际问题的能力打下良好的基础。成绩是学生学习能力的一种数值度量，高考数学成绩可一定程度上代表学生大学以前的数学能力，而高等数学考试成绩一定意义上代表学生大学期间的数学水平。一个人的数学学习能力不是与生俱来的，它来源于学生日常生活和学习的长期积累，因此高数成绩在一定程度上与高考数学成绩之间存在关联。

对于本文所要研究的问题，由于大学生来自不同的地区，而各个地区的教育投入、师资力量、人均GDP、人均可支配收入又不相同，同一个地区的这些变量之间具有一定的相关性，而不同省的学生高等数学成绩具有异质性，因此用普通线性回归模型是不合适的。该问题数据具有明显的嵌套结构，采用分层线性模型研究可以同时考虑省份间差异及个体间差异，突破了传统的线性模型在分析嵌套结构上的局限性[1]，使该模型更接近现实生活中的基本现象，故采用分层线性模型分析该问题是十分有必要的。

本文基于分层线性模型引入对于高数成绩有所影响的有关变量：学生的性别、年龄、生源地、高考成绩、高数成绩、不同地区的人均GDP、人均可支配收入、教育投入、师生比。通过建立分层线性模型进行定量分析可以数值化的确定各变量与高数成绩的关联，进而可以有针对性的通过一些措施，提高大学生高数成绩。

2. 模型的介绍

从统计分析技术角度上讲，传统分析方法在分析多层数据时所遇到的问题可以通过多水平多层统计

分析模型得到解决，多水平多层模型以两水平为例的话，模型如下基本模型[2]：

$$y_{ij} = \beta_{0j} + \sum_{k=1}^{K} \alpha_k x_{kij} + \sum_{l=1}^{L} \beta_{lj} z_{lij} + e_{ij} \tag{2.1}$$

$$\beta_{0j} = \gamma_{00} + \sum_{h=1}^{H} \gamma_{0h} \omega_{hj} + u_{0j} \tag{2.2}$$

$$\beta_{1j} = \gamma_{10} + \sum_{h=1}^{H} \gamma_{1h} \omega_{hj} + u_{1j} \tag{2.3}$$

$$\cdots$$

$$\beta_{Lj} = \gamma_{L0} + \sum_{h=1}^{H} \gamma_{Lh} \omega_{hj} + u_{Lj} \tag{2.4}$$

式中 y_{ij} 表示在第 j 个水平 2 单位(如第 j 组)中的第 i 个个体的水平 1 结局测量；其中，$i = 1,2,\cdots,N$ (N 是总样本量)，$j = 1,2,\cdots,J$ (J 是水平 2 的单位数)，水平 1 截距 β_{0j} 是随机截距，K 个水平 1 解释变量(即变量 x_{kij})具有固定效应或有固定斜率，L 个水平 1 解释变量(即 z_{lij})具有随机效应或有随机斜率。每一个水平 1 随机回归系数($\beta_{0j}, \beta_{lj}, l = 1,2,\cdots,L$)被定义为 H 个水平 2 解释变量(即变量 ω_{mj})的线性函数，这样就建立了 $L+1$ 个宏观方程(即公式(2.2)~(2.4))。

将公式(2.2)~(2.4)代入(2.1)，得出以下组合模型：

$$y_{ij} = \gamma_{00} + \sum_{h=1}^{H} \gamma_{0h} \omega_{hj} + \sum_{k=1}^{K} \alpha_k x_{kij} + \sum_{l=1}^{L} \gamma_{l0} z_{lij} + \sum_{l=1}^{L} \sum_{h=1}^{H} \gamma_{lh} \omega_{hj} z_{lij} + \left(u_{0j} + \sum_{l=1}^{L} z_{lij} u_{lj} + e_{ij} \right) \tag{2.5}$$

与普通多元回归模型不同，多层回归模型有宏观和微观两种方程，每一种方程均有一个残差项，因而有一个总的复合残差结构。水平 1 模型的残差项(如公式 2.1 的 e_{ij})代表组内变异，水平 2 模型的残差项(如公式 2.2 和 2.3 的 u_{0j} 和 u_{1j})分别代表水平 1 随机回归系数 β_{0j} 和 β_{1j} 的跨组或组间变异。模型假设水平 1 残差符合正态分布，水平 2 残差符合多元正态分布，且水平 1 残差与水平 2 残差相独立。这些假设可表述为：

$$e_{ij} \sim N\left(0, \sigma^2\right) \tag{2.6}$$

$$\begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix} \begin{pmatrix} u_{u_0}^2 & u_{u_{01}}^2 \\ u_{u_{01}}^2 & u_{u_{01}}^2 \end{pmatrix} \right] \tag{2.7}$$

$$Cov\left(e_{ij}, u_{0j}\right) = 0, Cov\left(e_{ij}, u_{1j}\right) = 0 \tag{2.8}$$

$Cov\left(e_{ij}, u_{0j}\right) = 0$ 和 $Cov\left(e_{ij}, u_{1j}\right) = 0$

表示水平 1 残差与水平 2 残差之间无相关关系，但不同宏观方程中的水平 2 残差之间可以存在相关关系，如 $Cov\left(u_{0j}, u_{1j}\right) = \sigma_{u_{01}}^2$

3. 大学生高等数学成绩影响因素建模分析

本文研究数据取自某重点工科高校 2014 级不同专业不同班级大学生两个学期高等数学课程考试成绩和学生的一些基本资料。由于部分学生出现转专业、降级、退学、休学等情况，造成了数据的不完整，本文将相应数据进行了删除，通过变量之间的相关性分析和一些基本常识，本文初始选定学生性别 X_1、学生年龄 X_2、生源地 X_3、地区人均 GDP X_4、地区人均可支配收入 X_5、地区教育投入 X_6、地区师生比 X_7、将高考成绩依地区进行中心化后得到 X_8 作为自变量，因变量为学生高等数学成绩 Y。一般检验

方法为似然比检验，但我们考虑到了检验在边界上的问题。最终在数据收集整理过程中得到的样本数据的样本容量为 1409，男女比例约为 3.9。生源地共计 28 个省及直辖市。表 1 为部分数据。

进行探索性分析，可以得到高考成绩与高数成绩的关系如图 1。

Table 1. The information on freshmen in science and engineering college come from different major and classes
表 1. 某重点工科高校 2014 级不同专业不同班级大学生个人信息情况

X_1	X_2	X_3	Y	X_4	X_5	X_6	X_7	X_8
男	22	新疆	21	40,648	15,097	599	12.8	−8
男	22	河北	62	39,984	16,647	1030	13.3	4
男	21	吉林	80	50,160	17,520	548	16.5	2
男	20	河北	77	39,984	16,647	1030	13.3	0
男	20	云南	41	27,264	13,772	901	15.6	4

注：资料来源：学校教务处。

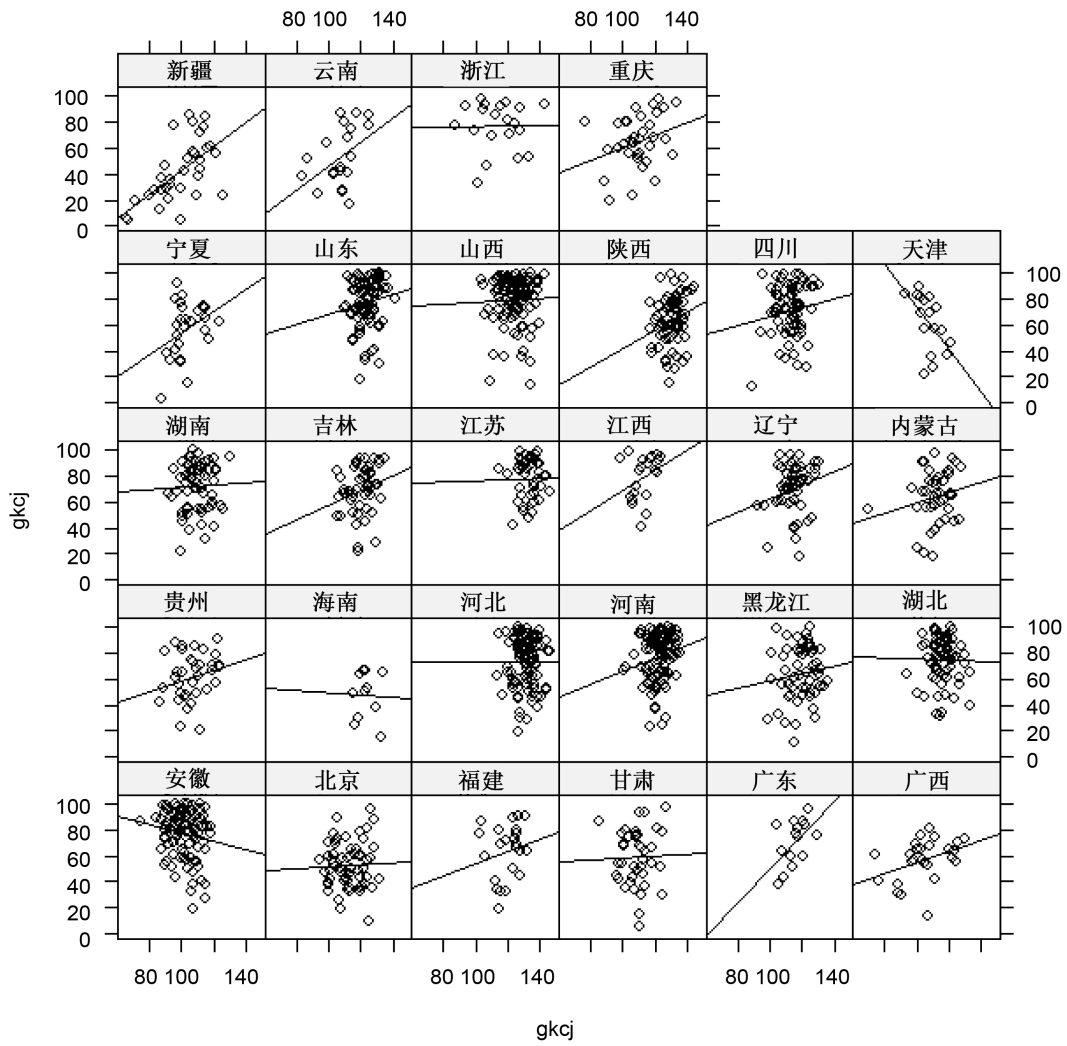

Figure 1. Diagram of regression analysis on the transcripts of Mathematics and Advanced Mathematics
图 1. 某重点工科高校 2014 级大学生高考数学成绩和高等数学成绩回归分析图

图 1 所示横轴为高考数学成绩，纵轴为高等数学成绩，以广东省为例，随着横轴高考成绩的增加，纵轴高数成绩也在增加，整体表现为高考成绩越高的学生高等数学成绩越好；以天津省为例，随着横轴高考成绩的增加，纵轴高数成绩在减少，即高考成绩越高的学生高数成绩反而越低。故由图可知各省图示高考数学成绩与高等数学成绩的回归图像之间有较大差异，斜率和截距都显著不同，说明不同省份高考数学成绩对高数成绩的影响不同，初步可以得出分层是合理的。接下来进行数学建模，定量分析高考数学成绩与高等数学成绩的关系。

1) 模型 1：零模型

$$\begin{cases} Y_{ij} = \beta_{0j} + \varepsilon_{ij} & \varepsilon_{ij} \sim N\left(0, \sigma^2\right) \\ \beta_{0j} = \gamma_{00} + u_{0j} & u_{0j} \sim N\left(0, \sigma_{00}^2\right) \end{cases}$$
$$Y_{ij} = \gamma_{00} + u_{oj} + \varepsilon_{ij}$$

Y_{ij}：第 j 个地区的第 i 个学生高数成绩

u_{0j}：每个地区高数成绩均值与总均值 γ_{00} 之间的差异；组间差异

ε_{ij}：每个学生高数成绩与地区均值之间的差异；组内差异

通过模型的构建我们发现，通过检验在边界的方法，得到 $P < 0.05$，显著拒绝原假设[3]，即随机截距项显著不为零，说明截距项随省及直辖市的不用显著不同，所以使用分层线性模型构建模型是合理的，协方差参数估计部分报告了水平 1 随机截距方差($\hat{\sigma}_{u_0}^2 = 89.17$)和水平 1 残差方差估计($\hat{\sigma}^2 = 353.00$)。结果显示：① 各省的高等数学平均成绩有显著差异；② 尽管组间差异显著，组内差异仍然很大；水平 1 残差方差(即 $\hat{\sigma}^2$)约为组间方差($\hat{\sigma}_{u_0}^2$)的 3.96 倍。组内相关系数(ICC)得 0.2016，即结局测量中约 20.16%的总变异是由生源地的不同引起的。为了考虑省与省之间的差异是哪些因素造成的，我们建立模型 2。

2) 模型 2：用省级别变量解释省际之间变异

$$\begin{cases} Y_{ij} = \beta_{0j} + \varepsilon_{ij} & \varepsilon_{ij} \sim N\left(0, \sigma^2\right) \\ \beta_{0j} = \gamma_{00} + \gamma_{01}X_{4j} + \gamma_{02}X_{5j} + \gamma_{03}X_{6j} + \gamma_{04}X_{7j} + u_{0j} & u_{0j} \sim N\left(0, \sigma_{00}^2\right) \end{cases}$$
$$Y_{ij} = \gamma_{00} + \gamma_{01}X_{4j} + \gamma_{02}X_{5j} + \gamma_{03}X_{6j} + \gamma_{04}X_{7j} + u_{0j} + \varepsilon_{ij}$$

“X_4”人均 GDP“X_5”人均可支配收入

“X_6”教育投入“X_7”学生教师比

建模过程中最初考虑的省级别变量有地区人均 GDP X_4、地区人均可支配收入 X_5、地区教育投入 X_6、地区师生比 X_7，经过似然比检验，最终确定省级别变量地区教育投入 X_6 是显著的，纳入了我们的分层线性模型，从而得到最终用省级别变量解释省际差异的分层线性模型如下：

$$\begin{cases} Y_{ij} = \beta_{0j} + \varepsilon_{ij} \\ \beta_{0j} = \gamma_{00} + \gamma_{01}X_{6j} + u_{0j} \end{cases} \Rightarrow Y_{ij} = \gamma_{00} + \gamma_{01}X_{6j} + u_{0j} + \varepsilon_{ij} \quad \begin{cases} \varepsilon_{ij} \sim N\left(0, \sigma^2\right) \\ u_{0j} \sim N\left(0, \sigma_{00}^2\right) \end{cases}$$

由模型建模结果可以得到两个固定效应值：$\hat{\gamma}_{00} = 56.60$，即不存在地区教育投入差别时相应的总体均数结局的估计值。即是说，在不存在地区教育投入差别时，模型估计的平均高数成绩约为 56.60 分。

省级别变量地区教育投入 X_6 的主效应为 $\hat{\gamma}_{01} = 0.0108$ ($P = 0.002$)，表明在存在地区教育投入差别时，学生的高数成绩高于不存在地区教育差别时。模型估计结果表明，在存在 1 单位地区教育投入差别的情况下，学生的平均高数成绩估计约为 $\hat{\gamma}_{00} + \hat{\gamma}_{01} = 56.60 + 0.01 = 56.61$ 分。至于方差/协方差成分，现有模型中组内方差的估计值为 $\hat{\sigma} = 352.98$，与空模型中方差的估计基本相同($\hat{\sigma}^2 = 353.00$)。然而，组间方差的估

计值 $\sigma_{u_0}^2$ 却从 89.17 下降到 57.27，表明省级别解释变量地区教育投入 X_6 能解释组间变异，而不能解释组内变异。

在模型纳入协变量后，组内相关系数(*ICC*)被称为条件相关系数(conditional *ICC*)，其反映的是在控制省级别解释变量地区教育投入 X_6 后的组内同质性或组间异质性。现有模型的条件组内相关系数为

$$ICC=\frac{\hat{\sigma}_{u_0}^2}{\hat{\sigma}_{u_0}^2+\hat{\sigma}^2}=\frac{57.27}{57.27+352.98}\approx 0.1396$$

模型解释的结局测量组间变异可通过 Raudenbush & Bryk 方法或 Snijders & Bosker 方法进行估计。

采用 Raudenbush & Bryk 方法，现有模型的省级别解释方差的估计[4]为

$$\left(\text{RB}:\text{省级别可解释变异\%}\right)=1-\frac{57.27}{89.17}\approx 0.36$$

采用 Snijders & Bosker 方法，现有模型的省级别解释方差可以通过 $\hat{\sigma}_{\text{水平2-总}}^2$ 的比例缩减来估计[5]。首先，计算空模型和现有设定模型的省级别总变异；

$$\left(\text{空模型}\hat{\sigma}_{\text{水平2-总}}^2\right)=89.17+\frac{353.00}{48}\approx 96.52$$

$$\left(\text{设定模型}\hat{\sigma}_{\text{水平2-总}}^2\right)=57.27+\frac{352.98}{48}\approx 64.62$$

$$\left(\text{SB}:\text{省级别可解释变异\%}\right)=1-\frac{64.62}{96.52}\approx 0.33$$

与 Raudenbush & Bryk 方法计算的结果近似。也就是说，约有 34.50%的高数成绩跨生源地的变异可以由省级别解释变量地区教育投入 X_6 解释。

为了考虑是否有个体变量影响个人的高数成绩，我们建立模型 3。

3) 模型 3：个体解释变量纳入模型

$$\begin{cases}Y_{ij}=\beta_{0j}+\beta_{1j}X_1+\beta_{2j}X_2+\beta_{3j}X_3+\varepsilon_{ij} & \varepsilon_{ij}\sim N\left(0,\sigma^2\right)\\ \beta_{0j}=\gamma_{00}+\gamma_{01}X_{6j}+u_{0j} & u_{0j}\sim N\left(0,\sigma_{00}^2\right)\end{cases}$$

$$Y_{ij}=\gamma_{00}+\gamma_{01}X_{6j}+\beta_{1j}X_1+\beta_{2j}X_2+\beta_{3j}X_3+u_{0j}+\varepsilon_{ij}$$

“ X_1 ”性别“ X_2 ”年龄“ X_8 ”高考成绩依地区进行中心化

建模过程中最初考虑的个体级别变量有性别 X_1、年龄 X_2、高考成绩依地区进行中心化 X_8，经过似然比检验，最终确定个体级别变量性别 X_1、高考成绩依地区进行中心化 X_8 是显著的，纳入了我们的分层线性模型，从而得到最终用个体级别变量解释个体差异的分层线性模型如下：

$$\begin{cases}Y_{ij}=\beta_{0j}+\beta_{1j}X_2+\beta_{2j}X_8+\varepsilon_{ij}\\ \beta_{0j}=\gamma_{00}+\gamma_{01}X_{6j}+u_{0j}\end{cases}\Rightarrow Y_{ij}=\gamma_{00}+\gamma_{01}X_{6j}+\beta_{1j}X_2+\beta_{2j}X_8+u_{0j}+\varepsilon_{ij}\quad\begin{cases}\varepsilon_{ij}\sim N\left(0,\sigma^2\right)\\ u_{0j}\sim N\left(0,\sigma_{00}^2\right)\end{cases}$$

由模型建模结果可以得到，2 个个体级别的变量(性别、高考成绩依地区进行中心化)对结局变量有显著影响，在控制了其他个体级别变量和省级别变量(地区教育投入 X_6)后，女生的高数成绩较男生要更具规律性。另外，高考成绩对高数成绩更具有显著的正效应。年龄对结局测量则无统计显著效应。

该模型的截距估计值为 $\hat{\gamma}_{00}=55.03$，较先前模型的截距估计值小，这是因为它们分别代表了不同的结局均数估计。这里 $\hat{\gamma}_{00}$ 代表的是当所有解释变量均取 0 值(即： $X_1=0$ 、 $X_6=0$ 、 $X_8=0$)时模型的的结

局总均数计值。也就是说，这里的 $\hat{\gamma}_{00}$ 代表模型估计的在无教育投入地区($X_6=0$)具有平均高考成绩($X_8=0$)的男生($X_1=0$)的子样本平均高考成绩。在有教育投入地区的相应数字为 $55.03+0.0109\approx 55.04$ 。

紧接着，我们来看模型在多大程度上解释结局测量在个体和省级别的变异。为了计算个体和省级别的解释方差，我们需要将现有模型与不含任何可解释性变异的空模型进行比较。

用 Raudenbush & Bryk 方法估计的模型个体和省级别解释方差[2]分别为：

$$\left(\text{RB}:\text{个体可解释变异\%}\right)=1-\frac{\hat{\sigma}^2\left(\text{设定模型}\right)}{\hat{\sigma}^2\left(\text{空模型}\right)}=1-\frac{338.25}{353.00}\approx 0.0418$$

$$\left(\text{RB}:\text{省可解释变异\%}\right)=1-\frac{\hat{\sigma}^2\left(\text{设定模型}\right)}{\hat{\sigma}^2\left(\text{空模型}\right)}=1-\frac{59.22}{96.52}\approx 0.3864$$

用 Snijders & Bosker 方法计的模型个体和省级别解释方差分别为：

$$\left(\text{SB}:\text{个体可解释变异\%}\right)=1-\frac{\sigma^2_{\text{水平1-总}}\left(\text{设定模型}\right)}{\sigma^2_{\text{水平1-总}}\left(\text{空模型}\right)}=1-\frac{338.25+59.22}{353.00+96.52}\approx 0.1158$$

$$\left(\text{SB}:\text{省可解释变异\%}\right)=1-\frac{\sigma^2_{\text{水平2-总}}\left(\text{设定模型}\right)}{\sigma^2_{\text{水平2-总}}\left(\text{空模型}\right)}=1-\frac{59.22+338.25/48}{96.52+353.00/48}\approx 0.3620$$

使用不同方法，可解释个体变异的比例不同，用 Raudenbush & Bryk 方法，仅有低于 1%的个体变异可以被模型解释；而用 Snijders & Bosker 方法，则约有 11.58%的个体变异可以被模型解释。看来 Snijders & Bosker 方法为个体变异的解释提供了更为合理的结果。至于可解释性省变异的比例，两张方法的结果基本相同。省解释方差在现有模型中为 37.5%，而在仅带有省级别变量的模型(2)中仅为 34.50%。这表明了将个体变量纳入模型后，不仅可以解释个体差异(个体残差方差 σ^2 从 353.00 减少到 338.25)，而且可以解释省级别变异(省级别残差方差 $\sigma^2_{u_0}$ 从 96.52 减少到 59.22)。说明个体解释变量既可以在个体水平，也可以在省水平上影响结局变异，因为它们的值可能既有组内变异，也有组间变异。

4) 模型 4：个体解释变量斜率的随机性检验

$$\begin{cases} Y_{ij}=\beta_{0j}+\beta_{1j}X_8+\beta_{2j}X_1+\varepsilon_{ij} & \varepsilon_{ij}\sim N\left(0,\sigma^2\right) \\ \beta_{0j}=\gamma_{00}+\gamma_{01}X_{6j}+u_{0j} & u_{0j}\sim N\left(0,\sigma^2_{00}\right) \\ \beta_{1j}=\gamma_{11}+u_{1j} & u_{1j}\sim N\left(0,\sigma^2_{11}\right) \\ \beta_{2j}=\gamma_{21}+u_{2j} & u_{2j}\sim N\left(0,\sigma^2_{22}\right) \end{cases}$$

$$Y_{ij}=\gamma_{00}+\gamma_{01}X_{6j}+\gamma_{11}X_8+\gamma_{21}X_1+u_{0j}+u_{1j}X_8+u_{2j}X_1+\varepsilon_{ij}$$

“ X_1 ”性别“ X_8 ”高考成绩依地区进行中心化

考虑到检验在边界问题得 P 值如表 2 所示。

高考数学成绩依地区进行中心化(X_8)的 P 值显著小于 0.05，说明高考数学成绩依地区进行中心化(X_8)省与省之间有显著差异，即高考数学成绩依地区进行中心化(X_8)回归系数是随机系数，说明高考成绩对

Table 2. P-Value of boundary verification
表 2. 检验在边界 P 值表

变量名	P 值
高考数学成绩依地区进行中心化(X_8)	0.0057
性别(X_1)	0.5803

高等数学的影响随省与省的变化而变化，性别(X_1)不显著，无随机斜率，说明性别对高等数学的影响不随省与省的变化而变化。最终模型如下：

$$\begin{cases} Y_{ij} = \beta_{0j} + \beta_{1j}X_8 + \beta_{2j}X_1 + \varepsilon_{ij} & \varepsilon_{ij} \sim N\left(0,\sigma^2\right) \\ \beta_{0j} = \gamma_{00} + \gamma_{01}X_{6j} + u_{0j} & u_{0j} \sim N\left(0,\sigma_{00}^2\right) \\ \beta_{1j} = \gamma_{11} + u_{1j} & u_{1j} \sim N\left(0,\sigma_{11}^2\right) \end{cases}$$

$$Y_{ij} = \gamma_{00} + \gamma_{01}X_{6j} + \gamma_{11}X_8 + \beta_{2j}X_1 + u_{0j} + u_{1j}X_8 + \varepsilon_{ij}$$

5) 模型 5：跨层交互作用的显著性检验

由于模型中既有省级别变量，又有个体变量，为了考虑这两个不同级别的变量之间是否存在交互作用，引入跨出交互作用的显著性检验

$$\begin{cases} Y_{ij} = \beta_{0j} + \beta_{1j}X_8 + \beta_{2j}X_1 + \varepsilon_{ij} & \varepsilon_{ij} \sim N\left(0,\sigma^2\right) \\ \beta_{0j} = \gamma_{00} + \gamma_{01}X_{6j} + u_{0j} & u_{0j} \sim N\left(0,\sigma_{00}^2\right) \\ \beta_{1j} = \gamma_{11} + \gamma_{12}X_{6j} + u_{1j} & u_{1j} \sim N\left(0,\sigma_{11}^2\right) \end{cases}$$

$$Y_{ij} = \gamma_{00} + \gamma_{01}X_{6j} + \gamma_{11}X_8 + \gamma_{12}X_{6j}X_8 + \beta_{2j}X_1 + u_{0j} + u_{1j}X_8 + \varepsilon_{ij}$$

经过似然比检验，等到 P 值为 0.90，故地区教育投入 X_6 与高考数学成绩依地区进行中心化 X_8 之间的交互作用统计不显著，表明个体水平解释变量 X_8 的效应并不受省级别变量 X_6 的影响，也就是说男生和女生在高等数学成绩上的差异与他们属于哪个省或直辖市并不相关，最终模型如下：

$$\begin{cases} Y_{ij} = \beta_{0j} + \beta_{1j}X_8 + \beta_2X_1 + \varepsilon_{ij} & \varepsilon_{ij} \sim N\left(0,\sigma^2\right) \\ \beta_{0j} = \gamma_{00} + \gamma_{01}X_{6j} + u_{0j} & u_{0j} \sim N\left(0,\sigma_{00}^2\right) \\ \beta_{1j} = \gamma_{11} + u_{1j} & u_{1j} \sim N\left(0,\sigma_{11}^2\right) \end{cases}$$

$$Y_{ij} = \gamma_{00} + \gamma_{01}X_{6j} + \gamma_{11}X_8 + \beta_2X_1 + u_{0j} + u_{1j}X_8 + \varepsilon_{ij}$$

固定效应系数估计如表 3 所示。

由检验结果显示，截距项为 55.24，意味着男生在高考中数学成绩为该省的平均成绩，若他从未享受过该省的教育资源，则他高数成绩有 55 分；教育投入的回归系数为 0.01，意味着其他变量保持不变时，

Table 3. Estimation of fixed and random effects
表 3. 固定效应及随机效应估计表

固定效应			
变量名	估计值	标准差	P 值
截距项	55.24	3.36	<0.0001
教育投入	0.01	0.003	0.0017
高考数学	0.35	0.07	<0.0001
性别	6.79	1.24	<0.0001
随机效应			
截距项	——	7.99	——
高考数学	——	0.20	——
残差离差：18.31。AIC：12,279.35			

某省每增加教育经费 100 万元时，则该省大学生在高数考试中可多得 1 分；由于性别变量为 0~1 (男——0，女——1)虚拟变量，其回归系数为 6.7，意味着平均意义下，其他条件控制不变时，女生的高等数学成绩比男生平均意义下高约 6.79 分；高考数学成绩的回归系数为 0.35，意味着某学生高考数学每比该省数学平均分高 10 分，则其高数分数将多 3 分。同时，经过正态性检验，我们验证模型中 ε_{ij}、u_{0j}、u_{1j} 服从正态分布的假设是成立的。

4. 总结

两水平分层回归模型建模过程充分考虑到数据中存在的分层特性，不能用线性回归做简单处理，所以这一过程更具逻辑性。分层次逐层加入解释变量，不断利用向前法和向后法筛选变量，并且充分考虑个体变量回归系数的随机性以及不同层次间可能存在的交互作用，变量选择严谨，具有较好的拟合优度。

大学生高等数学课程成绩是学生在校学习情况的集中体现之一，对其影响因素的数值分析是十分有意义的。本文的模型拟合结果表明：① 平均意义下女生的高等数学成绩远高于男生，即女生的在校学习情况远优于男生。② 除性别这个定性变量之外，个体水平的另一个自变量为高考成绩依地区进行中心化，在其他变量保持不变的情况下，依地区进行中心化后的高考成绩对学生的高等数学成绩具有显著的正效应，符合了理、工科学校在录取选拔学生时，对高考数学成绩有基本要求并择优录取的原则。③ 除个体水平的变量对高等数学成绩有所影响外，省级别的变量也不容忽视，拟合结果表明不同省份学生的高等数学成绩差异显著，一方面是由学生所在省份对教育投入存在差异所决定的，教育投入高的地区，学生的高等数学成绩要高于教育投入低的地区的学生，另一方面是由学生依地区中心化后的高考成绩所决定的，是学生个人能力积累的体现，由模型检验结果知，个人能力的积累不仅能在个体水平对学生的高等数学成绩产生影响，也能一定程度的平衡由教育投入的差异所带来的学生高数成绩的差异。

高等数学分上、下两学期且知识点较多。作为工科后续学习课程的基础，一定要充分重视高等数学课程学习。对授课教师来讲，在保证授课课时的前提下，增加一些辅助学习环节，例如课下的答疑课、章节习题课、知识结构讲解课等；采取一些方法提高学生的学习兴趣，例如讲解一些与课程相关的数学史的知识，讲解一些与学生所学专业相关的课程知识等；转变一些教学方法，例如在期中考试过程中增加口试的过程，深入了解学生的课程掌握情况，学生对课程学习的疑惑等：例如平时作业不再从课后习题中指定而是任课教师自己出题等；同时在学习上可进行男女搭配模式的学习互助小组，发挥女同学的积极带头作用，对基础薄弱的同学起到帮扶作用；在授课过程中也应多关注一些男同学的听课注意力，多对男同学进行提问，加强对男生出勤率的考察等。对学生来讲，应更注重个人能力的积累，并以此去尽可能的降低由生源地教育投入差异所带来的对高等数学成绩的影响，做到在上课认真听讲下课自行完成作业的同时，还应尽可能抽时间看一些指定教材之外的教材，相同知识点采取不同的讲解方式更易于接受；应在学习章节知识点的同时自主去回顾构建相应课程的知识结构体系。对学校来说，在录取选拔学生时，应保持对高考数学成绩有基本要求并择优录取的原则，对录取的学生按生源地分班进行高等数学课程的学习而不是按照专业划分班级，尽可能的去平衡大多数学生的学习进度，使学生能在一个较为轻松地环境下学习，也方便老师集体对学生所学知识的查缺补漏。对教育部门来说，应加大对落后地区的教育投入，给落后地区的考生一个更为舒适，优秀的教育环境。

项目计划

本文受北京市大学生科学研究与创业行动计划项目(项目编号：k201607007)支持。

参考文献 (References)

[1] Stephen, W. Raudenbush, A., Bryk, S., 郭志刚, 等. 分层线性模型: 应用与数据分析方法[M]. 第 2 版. 北京: 社

会科学文献出版社, 2007.

[2] 王济川, 谢海义, 姜宝法. 多层统计分析模型: 方法与应用[M]. 第 2 版. 北京: 高等教育出版社, 2008.

[3] 茆诗松, 程依明, 濮晓龙. 概率论与数理统计教程[M]. 第 2 版. 北京: 高等教育出版社, 2011.

[4] Stephen, W. Raudenbush, A. and Bryk, S. (2002) Hierarchical Linear Models: Applications and Data Analysis Methods. 2nd Edition. New Delhi, International Educational and Professional Publisher.

[5] Snijders, T.A.B. and Bosker, R.J. (1994) Modeled Variance in Two-Level Models. *Sociological, Methods & Research*, 22, 342-363.