幸福医药黎用冠慈善基金资助

香港幸福医药黎用冠慈善基金培训课程系列

2013年度西部高校青年教师教育行政与政策研究

暑期研讨班

(2013.8.4-2013.8.12)

主办单位

西安交通大学高等教育研究所

NUFE

量化方法在教育研究中的应用

Application of quantitative method in education research

黄斌
南京财经大学 财税学院

内容安排

1. 交叉表格：量化分析的起点
2. OLS回归及应用
3. Stata软件入门及其数据管理初步
4. 有限类别因变量回归：logit回归基础
5. 多层线性模型化方法

1. 交叉表格：量化分析的起点

交叉表格有什么用？

- 表格作为量化研究最为基础的元素，常常被研究者所忽视
  - 在当前的量化研究中，表格多用于统计描述分析之中，其分析性功能被大大低估
  - 交叉表格（Cross-tabulation）是一种十分有效且"廉价"的量化工具，合理地运用交叉表格，同样可以实现对变量间关系的深度分析

1.1 案例讲解

- "Religion: opiate or inspiration of civil rights militancy among Negroes?（Gary Marx，1967）"
  - 探讨黑人宗教信仰与人权观之间的关系，依据以往研究文献与理论陈述，有以下三种待验证的假设
    - 假设1：因为宗教宣言忍耐，倡导对另一个世界精神升华，对当前世界世俗态度淡漠，因此宗教对黑人争取人权态度有负影响，对维系人权种族分离现状有帮助
    - 假设2：因为黑人教堂是黑人聚集的重要场所，宗教成为宣扬人权价值观的重要来源，因此宗教对黑人争取人权态度有正影响，对打破人权种族分离现状有帮助
    - 假设3：两者之间无相关关系

由研究问题而提出的假设（hypothesis）是任何一项量化研究的起点，如果假设进一步得到计量检验和合理解答后即可形成理论。所谓"理论"并不需要多么的宏大或抽象，任何一种有关"what cause what"，或者两种变量为什么与怎么相互联系的想法，都有机会可以成为一种理论

案例讲解：数据抽样

- 有了假设，就需要检验，而检验需要量化数据信息
  - 量化数据包括一手数据和二手数据
    - 对假设进行检验，最理想的数据莫过于一手的总体数据，但总体样本常常难以获得，因此，我们就需要运用一定的抽样方法和技术，以一定概率值从总体中抽取一定的子集，形成概率样本（probability sample）
      - 我们希望假设在概率样本中的检验结果，应当与在总体中的检验结果一致，或者说，能从概率样本的分析中推断出总体样本情况。这就需要我们所抽取的概率样本应具有对总体的代表性
      - 在Marx研究中，他抽取了1119位在都市居住的黑人作为样本，其中有492名来自北方，剩余的来自南方
      - 他的抽样是否具有代表性呢？

案例讲解：问卷设计

- 设定样本和被访对象后，我们需对我们感兴趣的数据信息进行调查
  - 通过开放（open-ended）或封闭的（close-ended）问项获得你所感兴趣的信息进行调查。当所测量的对象是一种较为抽象的概念，我们通常需要多个问项来进行度量
- 例如，被访者对人权运动时持何种态度（积极还是保守）应当如何判定呢？

案例讲解：问卷设计

- 单个问项测量
  - What would you say about the civil right demonstrations over the past few years-that they have helped Negroes a great deal, helped a little, hurt a little, or hurt a great deal?
    - helped a great deal 1
    - helped a little 2
    - hurt a little 3
    - hurt a great deal 4
    - Don’t know 5

如上问题可以通过赋值（code）输入形成数据，并形成一个变量。在设计回答项时，有一个问题是“don't know”，对于这个选项的赋值，可能有两种处理方法：一是将它看成是缺失值，二是将它看成是个人态度的一种表达，即把它看成是处于helped a little与hurt a little之间的一种态度。为分清这两种选择，有研究者在其问项中同时加入“don’t know”和“no answer”两个选择，将后者视为是缺失，而将前者看成是个人态度的一种表达

案例讲解：问卷设计

- 多个问项测量
  - Marx对被访者人权观的判定是运用了八个问题进行判断。他定义如对下述八个问题中给予了标准回复的话，那么这个被访者就是对人权持积极态度
    - In your opinion, is the government in Washington pushing integration too slow, too fast, or about right?(too slow)
    - Negros who want to work hard can get ahead just as easily as anyone else. (disagree)
    - Negros should spend more time praying and less time demonstrating. (disagree)
    - To tell the truth I would be afraid to take part in civil rights demonstrations. (disagree)
    - Would you like to see more demonstrations or less demonstrations? (More)
    - A restaurant owner should not have to serve Negroes if he doesn't want to. (disagree)
    - Before Negroes are given equal rights, they have to show that they deserve them. (disagree)
    - An owner of property should not have to sell to Negroes if he doesn't want to. (disagree)

1.2 交叉表格类型

- 频数表
  - 因为不同类别的总量个数不尽相同，所以频数表在分析变量因果与相关关系上没有太多意义，多用于统计描述

Religiosity	Militant	Nonmilitant	Total
Very religious	61	169	230
Somewhat religious	160	372	532
Not very religious	87	108	195
Not at all religious	25	11	36
Total	333	660	993

交叉表格类型

- 比例表
  - 根据此表，随宗教信仰程度的减弱，持有militant人权观点的比例在不断下降
    - 一定要加入百分数的行总和数，确定读数的方向
    - 应加入样本的个数，便于计算频数值
    - 应显示总样本的比例总和数
    - 两个变量谁应安排在列，谁应安排在行，并没有统一的说法
      - 从节省版面的角度看，一般是将类别多的一种变量放在行
    - 表格要分清谁是被解释变量，谁是解释变量。在本例中，宗教信仰程度是解释变量，个人人权观是被解释变量，我们想观察随信仰程度变化，持积极和保守人权观的人口比例会发生怎样变化，因此信仰程度在行，人权观在列
  - 似乎可以推测这宗教信仰对人权观存在负向影响，但一定是这样的吗？

Militancy	Very religious	Somewhat religious	Not very religious	Not at all religious	Total
Militant	27%	30%	45%	69%	33%
Nonmilitant	73%	70%	55%	31%	67%
Total	100%	100%	100%	100%	100%
N	(230)	(532)	(195)	(36)	(993)

1.3 伪相关相关（Spurious correlation）与控制变量

- 可能存在其他变量，该变量既对黑人宗教信仰有影响，同时对黑人的人权观有影响，若黑人宗教信仰与人权观之间的相关关系完全或部分被其他变量所解释，那么这两个变量之间的关系即为完全或部分伪相关关系
  - 例如，个人教育水平
    - 个人所受教育可能会对个人宗教信仰和人权观同时具有影响，受过良好教育的黑人可能会具有较强的人权观，且同时持有较弱的宗教信仰，也就是说，宗教信仰与人权观之间的相关关系可能只是通过个人教育水平这一变量才能产生

教育与宗教信仰、人权观

Religiosity	Education attainment		
	Grammar school	High school	College
Very religious	31%	19%	19%
Somewhat religious	57%	54%	45%
Not very religious	12%	24%	25%
Not at all religious	1%	4%	11%
Total	101%	101%	100%
N	(353)	(504)	(136)

Militant	Education attainment		
	Grammar school	High school	College
Militant	22%	36%	53%
Nonmilitant	78%	64%	47%
Total	101%	101%	100%
N	(353)	(504)	(136)

随教育水平变化，黑人的宗教信仰和人权观态度分别呈现下降与增加趋势

如何利用交叉表格控制变量

- 为检验伪相关关系是否存在，我们就需要在控制教育水平的条件下，考察宗教信仰和人权观是否还能呈现出之前所观察到的负相关关系
  - 由下表可见，在控制了教育水平变量后，持有积极人权观的人口比例依然会随着个人宗教信仰程度的变化而发生变化，宗教信仰对人权观的影响并不是完全伪相关关系

Militancy	Grammar school			High school			College		
	v	s	n	v	s	n	v	s	n
Militant	17%	22%	32%	34%	32%	47%	38%	48%	68%
Nonmilitant	83%	78%	68%	66%	68%	53%	62%	52%	32%
Total	100%	100%	100%	100%	100%	100%	100%	100%	100%
N	(108)	(201)	(44)	(96)	(270)	(138)	(26)	(61)	(49)

注：V=very religious; S=somewhat religious; N=not very religious。Not very religious和not at all religious被合并为一类

如何利用交叉表格控制变量

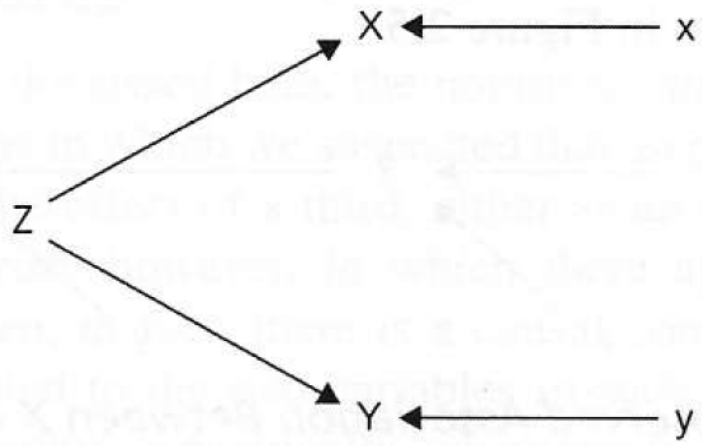
- 若被解释变量是两分类变量，那么上述表格可简化为下列形式：

Religiosity	Educational attainment		
	Grammar school	High school	college
Very religious	17% (108)	34% (96)	38% (26)
Somewhat religious	22% (201)	32% (270)	48% (61)
Not very or not at all religious	32% (44)	47% (138)	68% (49)

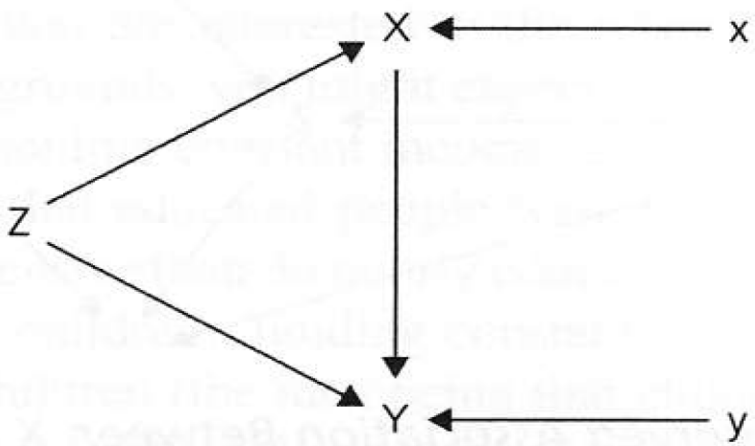
- 控制教育只是消除了一种可疑因素。事实上，可能还会有其他变量会同时对宗教信仰与人权观产生影响
  - 之后所考虑的控制变量，应尽量考虑哪些与教育水平相关度不大，且与宗教信仰与人权观同时具有影响的变量，例如年龄
  - 随着你需要控制变量的增加，表格这一工具的作用力度在减弱，一般用表格分析只适用于分析三个及以下变量的相关关系分析，因此，我们就需要运用回归分析

1.4 多种变量因果关系：伪相关

- 伪相关
  - 有其他变量对自变量和因变量同时具有影响



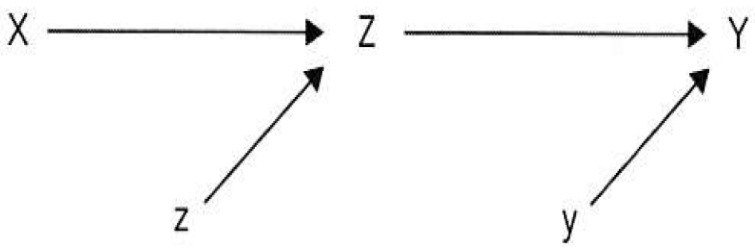
（完全伪相关）



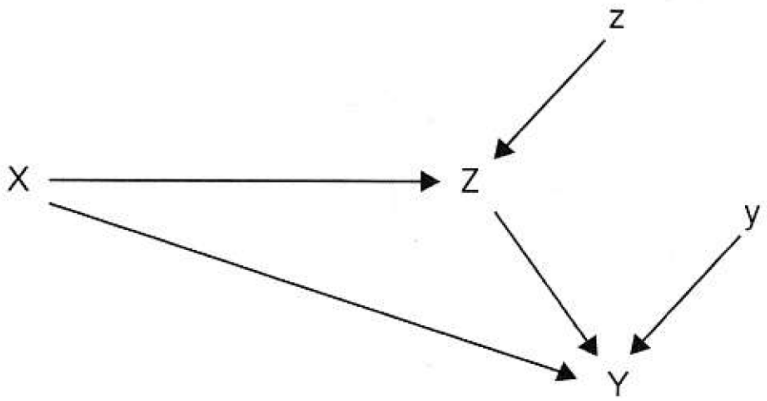
（部分伪相关）

多种变量因果关系：中介变量

- 中介变量（intervening/mediate variable）
  - X对Y有影响是因为X对Z有影响，Z对Y有影响。此Z变量即称为中介变量
    - 例如，父亲的职业地位对子女的教育水平有影响，而子女的教育水平又对他们的收入有影响，因此子女教育水平为父亲职业地位对子女收入水平影响的中介变量



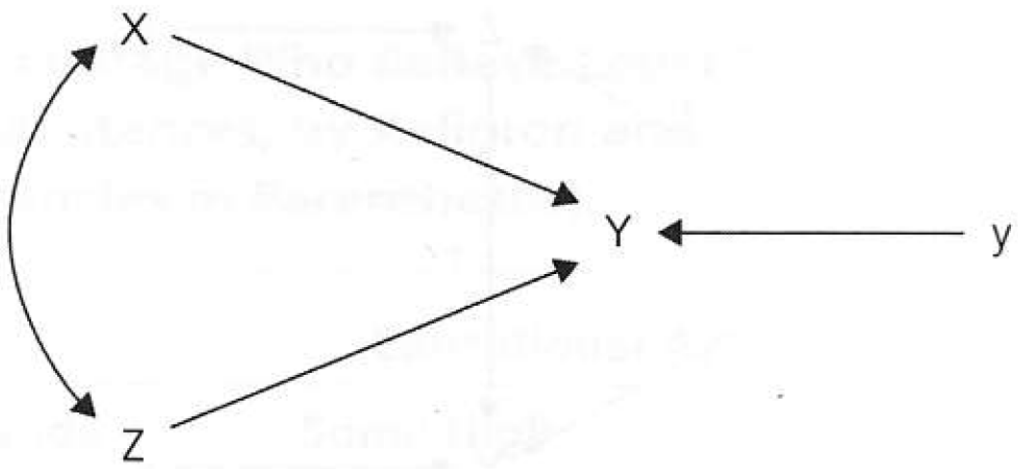
（完全中介效应）



（部分中介效应）

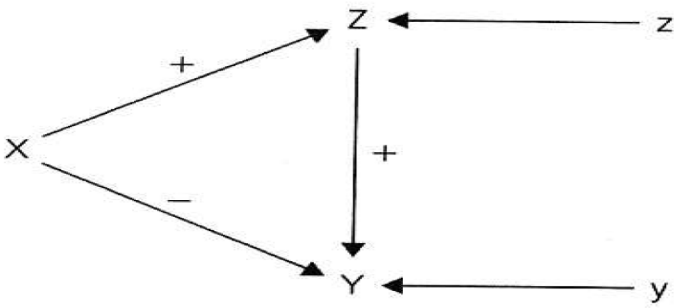
多种变量因果关系：中介变量

- 另一种中介变量的情况是x和z对y同时有影响，但x和z之间没有因果关系，只有相关关系



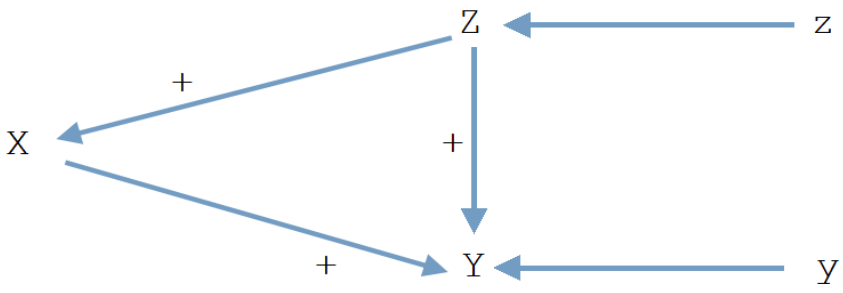
多种变量因果关系：压抑效应

- 压抑效应（suppressor effect）是指X对Y同时存在两种影响路径，一是对Y的直接影响，二是通过中介变量Z产生的间接影响，而且直接和间接影响的方向正好相反，相互抵消，由此导致在不控制Z的情况下，X对Y的影响很小，甚至为零，而实际上两者之间存在较大的因果关系
  - 如下图，考虑个人教育水平、收入和生育率之间的关系。一般认为，教育水平（X）越高，生育率越低（Y），同时教育水平越高，个人收入越高（Z），而个人收入越高，抚养孩子的能力越高，生育率也会越高
    - 在不控制Z变量个人收入变量的条件下，教育水平对生育率的影响是会被高估，还是被低估？
      - 低估！一正一负两种效应相互抵消我们可能会观测到教育水平对个人生育率只具有较小的影响，甚至无影响



多种变量因果关系：压抑效应

- 另一个例子
  - 个人教育水平会对收入有正向影响，个人天生能力对个人教育水平有正向影响，进而对个人收入产生正向影响，同时个人天生能力对个人收入水平有直接的正向影响
    - 若未控制个人天生能力，我们会高估还是低估个人教育水平对个人收入的影响？
      - 高估！



交互效应与加法效应

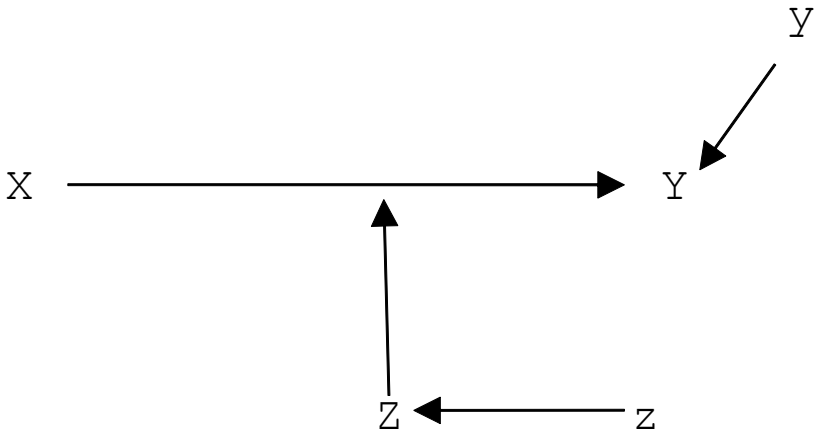
- 交互效应（interaction effect）
  - 是指X对Y的影响会受到另一变量Z的影响，或者X对Y的影响会随Z变化。在社会学研究中，此Z变量又常被称为调节变量（moderate variable）
    - 例如，我们想观察教育水平（X）对女性堕胎率（Y）的影响是否会受到个人信仰类型（Z）的影响（如下表）
      - 表格中各数字表示不同教育水平与宗教信仰下女性堕胎率的比例。由此表可知，教育对女性堕胎率的影响在不同的宗教信仰人群中表现有差别。在天主教信仰者中，这种影响关系并不明显，而在新教徒中，教育对堕胎率呈现明显的正影响，受教育水平越高，堕胎率的比例有明显提升。这表明，交互效应是存在的

Religion	Educational attainment			
	8th grade or less	Some high school	High school graduate	Some college or more
Catholic	31%(90)	33%(96)	33%(89)	31%(75)
Protestant	29%(287)	36%(250)	43%(256)	51%(225)

Source: Rossi(1966)

多种变量因果关系：调节变量

- 调节变量（moderate variable）
  - X对Y的影响会受到变量Z的影响，此时变量Z即为调节变量

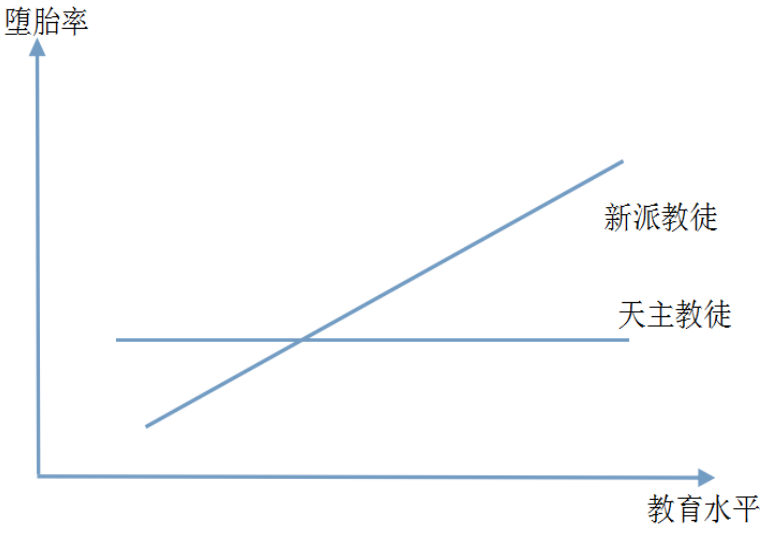


交互效应与加法效应

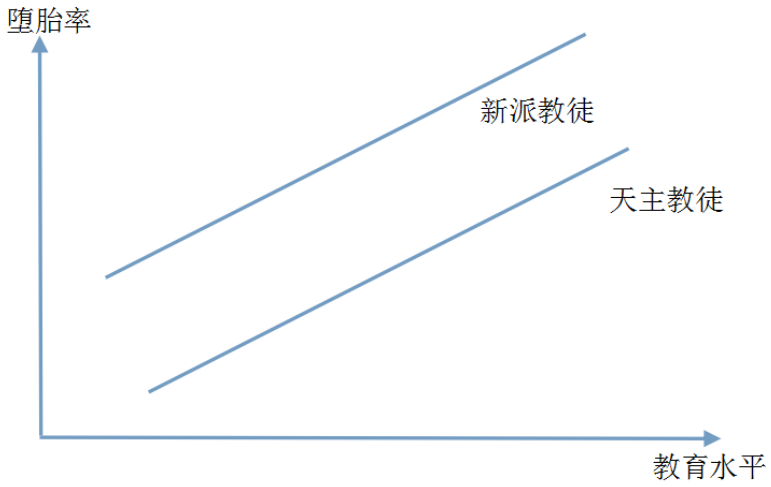
- 加法效应（additive effect）
  - 是指X对Y的影响不会受到另一变量Z的影响，但Z同时对Y存在影响（如下表）
    - 无论信仰哪种教派，教育水平对女性堕胎率都具有明显并且相同程度的正影响，这表明教育水平与女性堕胎率之间的关系不会受到宗教信仰的影响，即不存在交互效应
    - 无论是哪种教育水平，天主教信徒的堕胎率都要比新派教徒低10个百分点，这表明宗教信仰对女性堕胎率具有影响，即存在加法效应

Religion	Educational attainment			
	8th grade or less	Some high school	High school graduate	Some college or more
Catholic	30%	35%	45%	65%
Protestant	40%	45%	55%	75%

交互效应与加法效应的图形区分



交互效应图



加法效应图

如上图，所谓交互效应即表现为斜率差别，而加法效应则体现为斜率差别。
在线性回归中，交互效应通过交互项实现，而加法效应则通过虚拟变量实现

1.5 利用交叉表格实现协变量调整

- 协变量调整（covariate adjustment）有什么用？
  - 利用协变量调整，可以观测两个变量之间的关系在控制了另一个变量后会发生怎样的变化
    - 如之前的案例，我们已经发现宗教信仰对个人人权观的影响虽然能被教育水平部分解释，但两者之间并非完全伪相关关系。但接下来的问题是在控制了教育水平后，宗教信仰对人权观的剩余影响关系是多少

协变量调整的基本思路

- 我们为什么要控制教育水平？
  - 是因为教育水平对个人宗教信仰和人权观同时具有影响，但如果不同宗教信仰人口的教育水平分布完全相同，那么也不需要控制教育水平变量了
    - 协变量调整的基本思路是要通过一定方法（权重方法），将不同宗教信仰人群中各层次教育人口分布调整为相同分布

协变量调整步骤（1）

- 画出不同宗教信仰与教育层次下持不同人权观黑人人口比例表

Religiosity	Educational attainment		
	Grammar school	High school	college
Very religious	17% (108)	34% (96)	38% (26)
Somewhat religious	22% (201)	32% (270)	48% (61)
Not very or not at all religious	32% (44)	47% (138)	68% (49)

协变量调整步骤（2）

- 计算出总样本中各教育层次人口的分布比例
  - grammar school、high school和college教育水平人口占总样本数量的比例分别为0.356、0.508和0.137
  - 这些教育人口比例将作为权重进行协变量调整

协变量调整步骤（3）

- 将第二步中的教育人口比例作为权重，对各宗教信仰和教育层次下持有不同人权观的人口比例进行加权，计算得出不同宗教信仰人群中持不同人权观的人口比例分布

Religiosity	Educational attainment			
	Grammar school	High school	college	Weighted sum
Very religious	17%*0.356+	34%*0.508+	38%*0.137=	29%
Somewhat religious	22%*0.356+	32%*0.508+	48%*0.137=	31%
Not very or not at all religious	32%*0.356+	47%*0.508+	68%*0.137=	45%

协变量调整步骤（4）

- 比较调整前和调整后不同宗教信仰人群中持不同人权观的人口比例分布
  - 在协变量调整后，随宗教信仰程度加强，持积极人权观的黑人人口比例依然呈下降趋势。这表明教育水平不构成形成宗教信仰与人权观之间伪相关关系的一个原因
  - 在协变量调整前，随宗教信仰程度加大，黑人中持积极人权观的人口比例由27%变化至48%，跨度达到21%，而在协变量调整后，该跨度缩减为16%。也就是说，教育水平这一变量能解释24%（=（21%-16%）/21%）的宗教信仰与人权观关系

	Percent militant	Adjusted percent militant	N
Very religious	27%	29%	(230)
Somewhat religious	30%	31%	(532)
Not very or not at all religious	48%	45%	(231)
Percent spread	21%	16%	

1.7 因变量为连续变量的交叉表格分析

- 若因变量不是类别变量，而是连续变量，其表格分析应用方法与之前类别变量相似
  - 唯一不同在于是将之前表格中的比例换为均值、标准差、频数或其他描述性统计指标
  - 如下表

因变量为连续变量的交叉表格分析

Education	Male	Female
	Means	
Post-graduate	31864	14113
College graduate	27227	11789
Some college	19222	13003
High school graduate	16288	10324
Less than twelve years	15855	8399
Total	20415	11135
	Standard Deviation	
Post-graduate	17541	5019
College graduate	14618	6794
Some college	12912	9704
High school graduate	8935	7573
Less than twelve years	11488	6280
Total	13790	7750
	Frequencies	
Post-graduate	57	21
College graduate	46	35
Some college	68	52
High school graduate	131	105
Less than twelve years	78	33
Total	380	246

因变量为连续变量的交叉表格分析

- 利用均值和标准差（或方差），我们大致可以知道某一变量的分布状况。均值反映其分布水平，标准差反映其离散状况
  - 均值：男性平均收入不仅在总样本中要明显高于女性，而且在不同教育层次水平上，男性收入亦远高于女性。也就是说，在控制受教育水平后，性别对于个人收入依然存在明显的影响
  - 标准差：男性收入离散程度远高于女性。具有post-graduate学历的男性收入标准差是同具有post-graduate女性收入标准差的3倍多
    - 具有post-graduate学历女性的职业就业范围比较窄，这些职业的收入大多相近，且低于同为post-graduate学历男性的平均收入水平
      - 引申讨论：高学历女性大多都从事何种职业，这些职业具有哪些特征？高学历女性为什么大多被“锁定”在比较狭窄的就业范围中

1.8 差异度指标（index of dissimilarity）

- 在一些情况下，特别是在变量有大量分类且不同类别之间差别十分细微或不呈现明显规律的情况下，就无法运用交叉表格方法进行分析，此时我们可以运用差异度指标实现我们的研究目的
  - 案例
    - 我们观测劳动力职业地位是否存在性别与种族之间的分化
      - 不同性别人口的劳动力职业地位是否存在明显差别？
      - 不同种族人口的劳动力职业地位是否存在显著差别？
      - 职业地位的性别分化和种族分化，哪一种更严重？
    - 利用交叉表格，我们很难用肉眼观察到职业地位在性别和种族的分化程度（如下表）

差异度指标案例

Major Occupation Group	White Men	Black and Other Men	White Women	Black and Other Women
Professional and technical workers	15.6%	10.5%	16.4%	14.2%
Manager administrators	14.9%	6.9%	6.8%	3.4%
Sales workers	6.4%	2.5%	7.4%	3.1%
Clerical workers	6.0%	7.6%	35.9%	29.0%
Craft and kindred workers	22.0%	16.6%	1.9%	1.2%
Operative, except transport	11.2%	15.4%	10.2%	14.7%
Transport equipment operatives	5.6%	8.5%	0.8%	0.6%
Non-farm laborers	6.7%	12.7%	1.3%	1.6%
Private household workers	0.0%	0.1%	2.0%	6.8%
Other service workers	7.7%	15.7%	16.1%	24.6%
Farmers and farm managers	2.5%	0.6%	0.4%	0.1%
Farm laborers and supervisors	1.5%	2.9%	0.9%	0.7%
Total	100.0%	100.0%	100.0%	100.0%

差异度指标测量方法

- 差异度指标计算公式

$$\Delta = \frac{\sum |P_i - Q_i|}{2}$$

  - P_i表示第一种分布下的第i个分类的百分数值，Q_i表示第二种分布下的第i个分类的百分数值
  - 差异度指标值域在0-100之间，0表示完全相同，100表示完全不相同

差异度指标测量方法

- 首先，按性别分并控制种族，分为白人男性对白人女性和黑人男性对黑人女性
  - 将上表中的第二列百分数值减去第四列百分数值，和将第三列百分数值减去第五列百分位数值，并分别取绝对值加总除以2
    - 白人男性对白人女性的差异度指标：（|15.6-16.4|+|14.9-6.8|+...+|1.5-0.9|）/2=42.1
    - 黑人男性对黑人女性的差异度指标：(|10.5-14.2|+|6.9-3.4|+...+|2.9-0.7|)/2=41.3
- 其次，按种族并控制性别，分为男性白人对男性黑人和女性白人对女性黑人
  - 将上表中的第二列百分数值减去第三列百分数值，和将第四列百分数值减去第五列百分位数值，并分别取绝对值加总除以2
    - 男性白人对男性黑人的差异度指标：24.3
    - 女性白人对女性黑人的差异度指标：18.2
- 结论
  - 职业地位在性别间和种族间都存在一定的分化（差别）
  - 职业地位在性别间的差异要比种族差异大。性别分化具体表现为知需40%白人（或黑人）女性的职业发生变化，才能使得白人（或黑人）女性的职业分布与白人（或黑人）男性职业分布完全相同；相比之下，只要不超过25%的男性黑人职业发生变化，就可以使得男性黑人的职业分布与男性白人职业分布完全相同，而要使得女性黑人的职业分布与女性白人完全相同，女性黑人职业分布所需发生变化的比例就更少了

差异度指标的缺陷

- 差异度指标会随分类类别数量的增加而不断增大
  - 如本例，如果职业类型分类由12种增加到24种，那么根据性别和种族拆分出的类别数量势必会增加很多，此时差异度指标值便会增大。也就是说，差异度指标只适用于同一分类标准下的差异度比较
    - 例如，不同国家所采用的职业分类标准不同，那么根据差异度指标所计算出的结果就不可以进行国别之间职业阶层分化的比较研究
- 随控制变量数量和类别的增加，计算负担会不断加大，且其结果可能会变得难以解释
  - 如本例，如果种族分三类：白人、黑人和其他族裔，会有几种比较组合？
    - 会有9种比较组合
      - 按性别比较：白人男性对白人女性、黑人男性对黑人女性和其他族裔男性对其他族裔女性；
      - 按种族比较：男性白人对男性黑人、男性白人对男性其他族裔、男性黑人对男性其他族裔、女性白人对女性黑人、女性白人对女性其他族裔、女性黑人对女性其他族裔

1.6 重新讨论控制变量的逻辑

- 我们为什么在探讨X和Y两个变量之间关系时需控制其他变量？
  - 因为我们希望能采用控制手段剔除其他变量的影响，由此得到X对Y的“净”影响
  - 为实现这一目标，最科学的方法是标准实验

标准实验

- 标准实验试图构造出一种环境，在这种环境下
  - 研究对象被分为两组或若干组，有一组为控制组，其他为处理组，被试对象被随机分配到这些组中，如果我们观测到对处理组进行了某些特定的处理，其结果与未进行任何处理的控制组的结果存在差别，那么我们就可以确定，即所得到的结果差别是由处理手段导致的，该因果关系是真实存在的
  - 标准实验方法应用于社会行为研究存在很多困难
    - 成本巨大，不确定性较强，难以保证对象能被完全随机分配
    - 存在道德限制，尤其是在处理手段与被试对象的福利存在较大关联时，谁被选入控制组，谁被选为处理组，会存在道德争议
    - 研究者所关注的议题常常很难被实验运作
      - 例如，想研究在战争中集体凝聚力会更强，农业社会的社会分化与分层要比狩牧社会更严重

利用控制变量“模拟”实验环境

- 在无法营造标准实验环境的条件下，控制成为了一种替代手段
  - 通过控制其他变量，将它们保持不变，尽可能避免X变量对Y变量的影响作用不会受到其他变量的“干扰”，得到X变量对Y变量的“净”影响。
- 控制变量手段的缺陷
  - 影响变量间因果关系的变量很多，我们不可能将所有的相关变量控制住。因此，通过控制手段，是无法肯定我们通过控制得到的X对Y的净影响是否一定是真正的因果关系
    - 解决方法：立足充分的文献阅读，通过充足的理论探讨构造因果关系假设，设定计量模型，正确识别函数模型，合理确立主要解释变量与函数形式，并在此基础上，采用合宜的数据，对假设进行实证检验

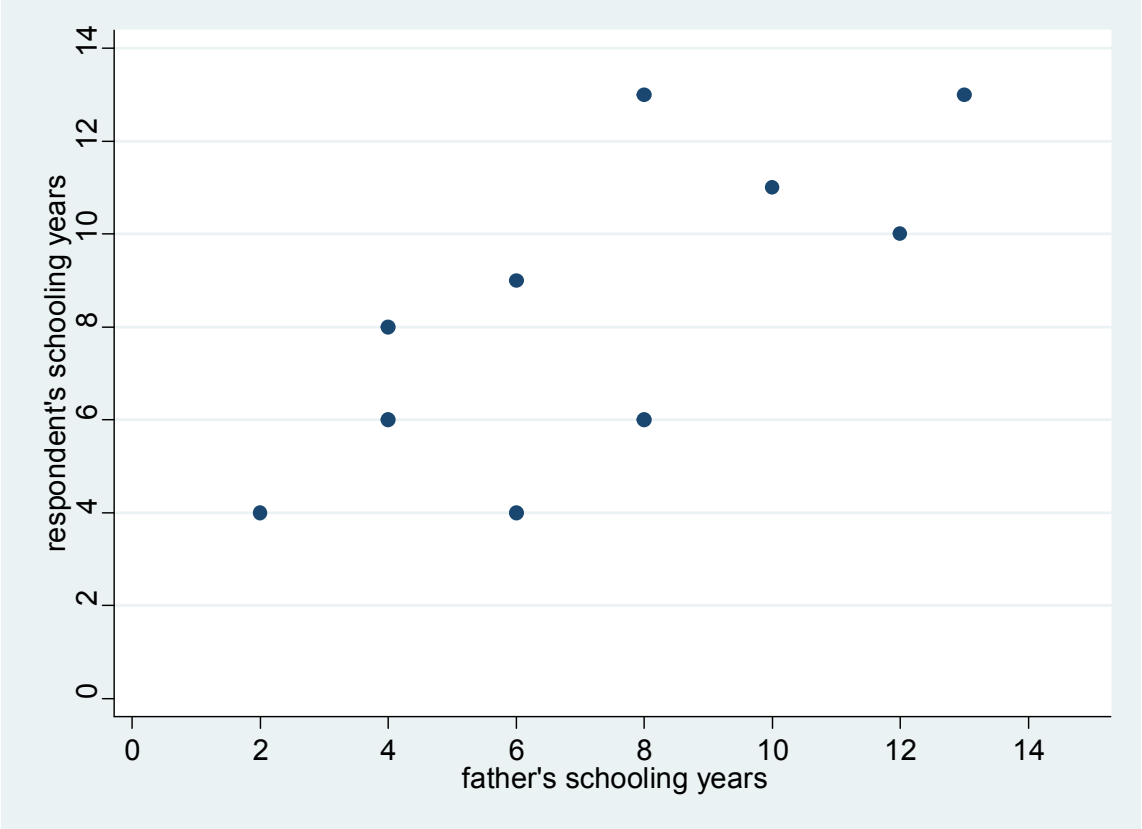
2. OLS回归及应用

2.1 一个案例

- 我们关注教育的代际流动问题，具体讨论父亲受教育水平是对于其子女受教育水平的影响作用
  - 是否具有影响，若存在，该影响具有何种特征？
  - 两者之间的相关关系到底有多强？

Father's schooling years	Respondent's schooling years
2	4
12	10
4	8
13	13
6	9
6	4
8	13
4	6
8	6
10	11

散点图



我们需找到一种直线能够“最优”刻画出以上散点变化关系

2.2 最小二乘法（OLS）原理

- 寻找一条拟合线，该直线能使得所有散点距离它的垂直距离的平方和达到最小
  - 符合这一条件的拟合线最能代表上图散点的变化趋势

2.3 OLS各项估计量及说明

- 估计方程：$y = b_1 + b_2 x + \varepsilon$
- 斜率估计量：$b_2 = \dfrac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$ （ $Var(b_2) = \dfrac{\sigma^2}{\sum (x_i - \bar{x})^2}$ ）
- 截距估计量：$b_1 = \bar{y} - b_2 \bar{x}$ （ $Var(b_1) = \sigma^2 [\dfrac{\sum x_i^2}{N \sum (x_i - \bar{x})^2}]$ ）
- 因变量预测值：$\hat{y} = b_1 + b_2 x$
- 残值估计值：$\varepsilon = y - \hat{y}$ （ $Var(\varepsilon) = \sigma^2$ ）

OLS各项估计量及说明

- 残值项方差σ^2越大，估计系数的方差越大，其估计系数的精确度便越低
  - 残值方差越大，因变量y的分布越广，OLS的估计系数就越不精确
- $\sum(x_i-\bar{x})^2$ 越大，说明自变量x的离散程度越大，估计系数的方差越小，其估计精度越高
- 样本容量N越大，估计系数的方差越小，其精度越大
- $\sum x_i^2$ 越小，截距估计系数的估计越精确。这是因为，即表示因变量x距离0值之间的离散程度，而截距表示当x＝0时，y的取值。所以当x离0越远，那么我们就更难估计出截距系数

拟合优度R^2

- 拟合优度指标是用于测评你构建的OLS模型所反映的各变量间关系优劣与强弱的一种统计量，其值域在0,1之间。拟合优度越高，说明因变量变异被你构建的模型解释的部分越多
  - 因变量的总变异（SST）=被自变量所解释的变量（SSR）+未被解释的变异（SSE）
  - 拟合优度R^2=SSE/SST=1-SSE/SST，即有：

$$r^2 = 1 - \frac{\sum (y_i - \hat{y})^2}{\sum (y_i - \bar{y})^2}$$

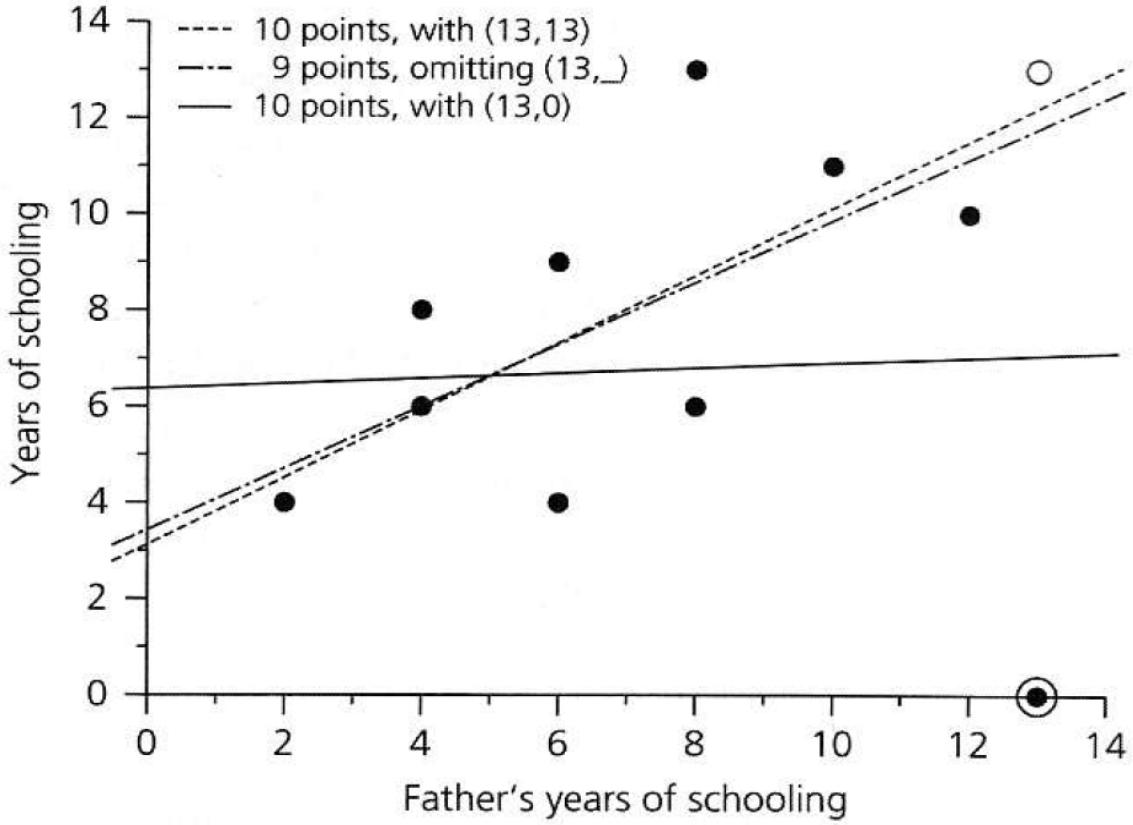
拟合优度R^2

- 拟合优度会随着模型所包含自变量个数的增加而不断增加
  - 调整后R^2=（SSE/(N-K)）/(SST/(N-1))
    - 其中，N为样本容量，K为模型中带估计的系数个数（一般情况下等于自变量个数+1）
    - 调整后拟合优度不再具有原先指标的解释能力了
- 当模型不包括截距项时，不应报告拟合优度
  - 只有当模型包括常数项时，SST=SSR+SSE等式才会成立
- 拟合优度等于自变量和因变量皮尔逊相关系数的平方
  - 皮尔逊相关系数值域在-1,1之间
- 不能仅依靠拟合优度来判断估计或预测的优劣性，一个拟合优度表现十分优异的模型，并不意味着它就一定具有显著的经济学意义。
  - 应该综合考虑回归系数估计值的符号和大小，回归结果所具有的统计学和经济学含义，估计的精确度，以及模型的预测能力

案例的估计结果及其解释

```
. reg r_edu fa_edu

      Source |       SS       df       MS              Number of obs =      10
-------------+------------------------------           F(  1,     8) =    9.23
       Model |  54.8496124     1  54.8496124           Prob > F      =  0.0161
    Residual |  47.5503876     8  5.94379845           R-squared     =  0.5356
-------------+------------------------------           Adj R-squared =  0.4776
       Total |       102.4     9  11.3777778           Root MSE      =   2.438

------------------------------------------------------------------------------
       r_edu |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      fa_edu |   .6873385   .2262642     3.04   0.016     .1655722    1.209105
       _cons |   3.382429   1.822797     1.86   0.101    -.8209482    7.585806
------------------------------------------------------------------------------
```

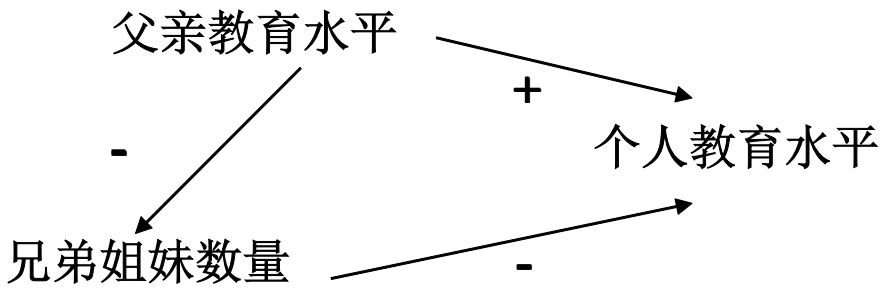
2.4 特异值（outlier）



OLS估计结果极易受到特异值的影响，一旦数据中出现特异值，即便只是个别数值，都可能导致OLS估计结果极大的变化，特别是当特异值位于散点中心位置较远的地方。此外，一般情况下，样本容量越大，特异值所可能造成的影响会越小

2.5 多变量OLS回归

- 在原先案例中加入一个自变量兄弟姐妹数量
  - 父亲受教育水平会对家庭兄弟姐妹数量产生负影响
  - 家庭兄弟姐妹数量会对个人受教育水平产生负影响
    - 兄弟姐妹数量作为中介变量，与其相关的两种效应负负得正，因此在控制该中介变量之后，父亲教育水平的估计系数会有所下降



    - 在未控制的情况下，兄弟姐妹数量变量的影响并未消失，而是被包含在残值项中，此时自变量父亲教育水平便会与残值项存在相关关系，遗漏变量会引发所谓“内生（endogeneity）问题，解决方案
      - 加入遗漏变量
      - 设计工具变量替代遗漏变量影响
        - 与父亲教育水平相关
        - 与残值项（或因变量）不相关

多变量回归结果

```
. reg r_edu fa_edu
      Source |       SS       df       MS              Number of obs =      10
-------------+------------------------------           F(  1,     8) =    9.23
       Model |  54.8496124     1  54.8496124           Prob > F      =  0.0161
    Residual |  47.5503876     8  5.94379845           R-squared     =  0.5356
-------------+------------------------------           Adj R-squared =  0.4776
       Total |       102.4     9  11.3777778           Root MSE      =   2.438

------------------------------------------------------------------------------
       r_edu |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      fa_edu |   .6873385   .2262642     3.04   0.016     .1655722    1.209105
       _cons |   3.382429   1.822797     1.86   0.101    -.8209482    7.585806
------------------------------------------------------------------------------
. reg r_edu fa_edu sibling
      Source |       SS       df       MS              Number of obs =      10
-------------+------------------------------           F(  2,     7) =    4.96
       Model |  60.0144184     2  30.0072092           Prob > F      =  0.0456
    Residual |  42.3855816     7  6.05508309           R-squared     =  0.5861
-------------+------------------------------           Adj R-squared =  0.4678
       Total |       102.4     9  11.3777778           Root MSE      =  2.4607

------------------------------------------------------------------------------
       r_edu |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      fa_edu |   .5644457   .2643104     2.14   0.070    -.0605491     1.18944
     sibling |  -.6398143   .6927667    -0.92   0.386    -2.277947    .9983186
       _cons |   6.262971   3.621132     1.73   0.127    -2.299645    14.82559
------------------------------------------------------------------------------
```

2.6 估计系数标准误与显著性检验

- 量化研究报告必须要报告估计系数标准误
  - 标准误本身能反映出系数的估计精确性
  - 标准误是进行系数显著性检验和计算系数置信区间的必要信息
- 利用估计系数标准误，我们可以快速进行显著性判定
  - 例如，某一自变量点估计系数为b，零假设为b=0，显著性水平选择α=0.05，（样本容量-自变量个数-1）≥30，此时若点估计系数达到其标准误的2倍以上（或者说，t值超过2），那么就可以明确拒绝原假设，即点估计系数显著不为0

$$t = \frac{b_k - c}{se(b_k)}$$

- 一个经常被采用的错误策略
  - 将不显著的自变量剔除出回归方程，只保留那些显著的自变量
    - 不显著的背后或许也有显著的经济学和统计学意义，是我们提出新观点和新理论的起点
      - 教育水平对农村劳动力外流影响同时具有正和负影响，相互抵消后，很可能表现为非显著
    - 是否应该纳入到回归方程不应以该变量最终的估计结果是否显著，而应该根据理论进行判定。如果理论既无说法，统计上又非显著，可以判定为无关变量，可以剔除
    - 控制应当控制的变量，才能得到自变量真正的“净”效应

2.7 如何比较不同自变量的估计系数大小

- 对于这一问题上，常有许多错误做法
  - 将两个单位不同的自变量估计结果直接进行大小比较
    - 如本例，fa edu估计系数是0.56，sibling的估计系数是-0.64，能否说兄弟姐妹数量对个人教育水平的影响超过父亲教育水平变量？
      - 兄弟姐妹数量与父亲教育水平的单位不同，因此需对自变量进行标准化转换（将因变量和所有自变量都减去均值后再除以其标准差，转换后所有变量的均值都为0，标准差都为1），再进行回归与估计系数。标准化后的回归系数表示当自变量值每变化一单位标准差，因变量会变化多少标准差
        - 一个好用的转换公式：$b^{|}=b*(s_x/s_y)$
  - 将两个单位相同的自变量的点估计结果直接进行比较
    - 即便你的样本估计结果是A变量的估计系数大于B变量，但在不同样本中该结果可能会有变化，因此根据点估计进行系数大小比较不具有统计学意义
    - 比较点估计大小应通过计量检验，F test
  - 就不同样本和不同回归方程的自变量估计结果进行比较
    - 不同样本因变量标准差不同，且不同方程所包含的自变量有差别，不能直接进行比较
    - Hausman检验+

同一回归方程的相同单位自变量的估计系数比较分析

- 分析义务教育阶段受教育年限与义务后教育阶段受教育年限对农村个人收入的影响是否存在差别

```
. reg lnincome edu_com edu_post
      Source |       SS       df       MS              Number of obs =    3545
-------------+------------------------------           F(  2,  3542) =  170.66
       Model |  443.248734     2  221.624367           Prob > F      =  0.0000
    Residual |  4599.63254  3542  1.29859755           R-squared     =  0.0879
-------------+------------------------------           Adj R-squared =  0.0874
       Total |  5042.88127  3544   1.4229349           Root MSE      =  1.1396

------------------------------------------------------------------------------
    lnincome |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     edu_com |   .0822851   .0069141    11.90   0.000     .0687291     .095841
    edu_post |   .1137816     .01154     9.86   0.000     .0911559    .1364072
       _cons |   5.510449   .0502071   109.75   0.000     5.412011    5.608886
------------------------------------------------------------------------------
. test edu_com=edu_post
 ( 1)  edu_com - edu_post = 0
       F(  1,  3542) =    4.33
            Prob > F =    0.0376
```

2.8 OLS模型整体的F检验

- 零假设
  - 所有自变量的回归系数为零
- 替代假设
  - 至少有一个自变量的回归系数不为零
- 检验逻辑
  - 先设立一个非限制模型，包含有你所设定的所有自变量，再设立一个只包含截距项的限制模型。若你所设定的模型中所有自变量对因变量都没有影响，估计系数都为0，因变量的总差异没有被任何自变量所解释，此时因变量总差异SST应与残值差异SSE相等，即接受原假设，若有显著变化，则拒绝原假设
- F统计量
  - F=((SST-SSE)/J)/(SSE/(N-K))，其中，J为自变量个数，N为样本容量，K为带估计系数个数
- 整体模型F检验的特点
  - 通过F检验，只能说明有部分自变量具有显著影响，但该假设并不能告诉我们究竟是哪一些变量应进入到模型之中，所以F检验在研究中的实际功效并不大，不一定需要文字报告

2.9 虚拟变量在OLS回归中的应用

- 一个案例
  - 教育收益率（明瑟方程）
    - $lnincome=b_0+b_1*edu+b_2*exp+b_3*exp2$
  - 研究问题
    - 男女性农村人口的平均收入是否存在差异？（加法效应）
    - 男女性农村人口的教育收益率是否存在差异？（交互效应）

虚拟变量在OLS回归中的应用

- 如何将类别变量拆解为虚拟变量
  - 若变量时n维类别变量，那么需将其拆分为n-1个虚拟变量，并将其中一组视为参照组（reference group）
    - x=(女性=1，男性=2)，拆分为：male（男性=1，其余=0），以female作为参照组
    - x=(东部=1，中部=2，西部=3)，拆分为：east（东部=1，其余=0），middle(中部=1，其余=0)，以west作为参照组
- 如何利用虚拟变量实现加法效应
  - $lnincome=b_0+b_1*edu+b_2*exp+b_3*exp2+b_4*male$
  - 男性：$lnincome_m=b_0+b_4+b_1*edu+b_2*exp+b_3*exp2$
  - 女性：$lnincome_w=b_0+b_1*edu+b_2*exp+b_3*exp2$
  - 男女性收入差别：$\Delta lnincome=b_4$
  - 观测b_4是否显著，其符号如何，以此来判断男女性平均收入是否有差异

虚拟变量在OLS回归中的应用

- 如何利用虚拟变量同事实现加法效应与交互效应
  - 构造交互项：edu_gender=edu*male
  - $lnincome=b_0+b_1*edu+b_2*exp+b_3*exp2+b_4*male+b_5*edu\_male$
  - 男性：$lnincome=b_0+b_4+(b_1+b_5)*edu+b_2*exp+b_3*exp2+b_4*male$
  - 女性：$lnincome=b_0+b_1*edu+b_2*exp+b_3*exp2$
  - 若b_4和b_5都显著，就表明男性与女性不仅平均收入有差别，而且其教育收益率亦存在差别，即同时存在加法效应与交互效应

加法效应与交互效应

```
. reg lnincome edu exp exp2 gender

      Source |       SS       df       MS              Number of obs =    3415
-------------+------------------------------           F(  4,  3410) =  166.53
       Model |  786.165331     4  196.541333           Prob > F      =  0.0000
    Residual |  4024.46199  3410  1.18019413           R-squared     =  0.1634
-------------+------------------------------           Adj R-squared =  0.1624
       Total |  4810.62732  3414  1.40908826           Root MSE      =  1.0864

------------------------------------------------------------------------------
    lnincome |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         edu |   .0460535   .0067282     6.84   0.000     .0328618    .0592452
         exp |   .0175667   .0050921     3.45   0.001     .0075828    .0275505
        exp2 |  -.0005268   .0000881    -5.98   0.000    -.0006996   -.0003539
      gender |   .6304397   .0381168    16.54   0.000     .5557057    .7051738
       _cons |   5.501039   .1014969    54.20   0.000     5.302038     5.70004
------------------------------------------------------------------------------
```

- male估计系数显著，表明男性收入比女性高

```
. reg lnincome edu exp exp2 gender edu_gender

      Source |       SS       df       MS              Number of obs =    3415
-------------+------------------------------           F(  5,  3409) =  136.84
       Model |  804.131816     5  160.826363           Prob > F      =  0.0000
    Residual |  4006.49551  3409  1.17527002           R-squared     =  0.1672
-------------+------------------------------           Adj R-squared =  0.1659
       Total |  4810.62732  3414  1.40908826           Root MSE      =  1.0841

------------------------------------------------------------------------------
    lnincome |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         edu |   .0650237   .0082837     7.85   0.000     .0487821    .0812653
         exp |   .0173413   .0050818     3.41   0.001     .0073776    .0273049
        exp2 |  -.0005171    .000088    -5.88   0.000    -.0006897   -.0003446
      gender |   .9377847   .0873267    10.74   0.000     .7665668    1.109003
  edu_gender |  -.0393316   .0100596    -3.91   0.000     -.059055   -.0196082
       _cons |    5.36166   .1073752    49.93   0.000     5.151134    5.572186
------------------------------------------------------------------------------
```

- 交互项估计系数显著，且为负值，表明女性教育收益率要高于男性；女性教育收益率为6.5%，男性教育收益率为2.6%（=6.5%-3.9%）

2.10 如何比较不同回归方程的优劣

- 使用F检验
  - 构建两个模型，R_1是镶套模型（如之前的加法效应模型）的拟合优度，R_2是被镶套模型（如之前的交互效应模型），前者镶套后者，后者镶套于前者
    - 构造F统计量：$F=((R_2-R_1)/m)/((1-R_2)/(N-k-1))$
      - 其中，m等于两个模型自变量数量差，为分子自由度；N为样本容量，k为镶套模型的自变量数量，N-k-1为分母自由度
      - 该统计量符合分子分母自由度分别为m和N-k-1的F分布
      - 原假设：镶套模型的拟合优度优于被镶套模型
    - 如本例
      - $F_{(1,3409)}=19$，$p<.000$，拒绝原假设
        - display F(1,3409,19)
        - .99998654
    - 注意：此种方法只适用于对同一样本的两个或多个镶套模型的优劣比较，不能用于不同样本，或非镶套模型的比较分析
      - 大样本的拟合度在相同的函数回归中一般要好于小样本，所以一个大样本回归的拟合度优于小样本，并不一定意味着模型本身设定得好
      - 如果模型不是镶套的，就无法构造出上述F统计量

如何比较不同回归方程的优劣

- 使用BIC（Bayesian information criterion）检验
  - BIC检验的优势
    - 不要求模型镶套，只要是对同一现象即可进行模型比较
    - BIC对大样本模型回归拟合优势进行了调整
    - BIC更偏好于简单模型，因此即便一个包含了更多自变量的复杂模型拟合度比简单模型高，但BIC值并不一定高
  - 统计量：$BIC=N(\ln(1-R^2))+k(\ln(N))$
    - 其中，R^2为模型拟合度，N为样本容量，k为模型自变量个数
  - BIC指标的判定标准
    - BIC是一个负值，其值越小，说明模型越优
    - 比较两个模型的BIC，可算其差值ΔBIC。“黄金法则”（Raftery，1995）
      - ΔBIC在0-2之间，能为后一种模型略优于前一种模型提供weak证据
      - ΔBIC在2-6之间，能为后一种模型较优于前一种模型提供positive证据
      - ΔBIC在6-10之间，能为后一种模型较优于前一种模型提供strong证据
      - ΔBIC在10以上，能为后一种模型较优于前一种模型提供very strong证据
    - 因为BIC虽经过调整，但依然会随样本增大而增大，所以以上法则只适用于小样本

BIC检验模型优劣的一个案例

- 如上例，我们想比较加法效应模型和交互效应模型孰优孰劣

```
. quietly reg lnincome edu exp exp2 male
. scalar bic1=e(N)*ln(1-e(r2))+e(df_m)*ln(e(N))
. quietly reg lnincome edu exp exp2 male edu_male
. scalar bic2=e(N)*ln(1-e(r2))+e(df_m)*ln(e(N))
. scalar D=bic2-bic1
. scalar list bic1 bic2 D
      bic1 =  -576.8161
      bic2 = -583.95995
         D =  -7.143851
```

2.11 缺失值（missing data）的处理

- 缺失值无处不在
  - 完全随机缺失（MCAR）
    - 缺失的发生与缺失的真值与模型所包含的其他变量无关
  - 随机缺失（MAR）
    - 缺失的发生与缺失的真值无关，但与模型所包含其他至少一个变量相关
  - 非随机缺失（MNAR or NI）
    - 缺失得发生与缺失得真值与模型所包含其他变量相关
  - 在教育研究中，个人受教育年限和收入水平是一个十分重要且通常都需要调查的关键变量，但这个变量的缺失问题与测量误差问题十分严重
    - 受教育年限变量缺失的发生概率不仅与个人真实的受教育水平存在关联，而且还与其他变量（例如年龄与地域）存在相关关系
    - 收入水平变量缺失的发生概率不仅与个人真实的收入水平相关，而且还与其他变量（例如教育水平和年龄）存在相关关系

2.12 如何处理缺失值

- 个案删除法（casewise deletion）
  - 一旦个案所包含的任一变量出现缺失，即将这个个案整条删除
    - 只适用于MCAR，否则会引起估计偏差
    - 即便是MCAR，若存在大量缺失，采用删除法会使得样本容量锐减
- 均值替代法
  - 样本均值替代法
    - 即某一变量存在缺失，以该变量的样本均值替代缺失值
  - 回归替代法
    - 即某一变量存在缺失，以回归的预测值替代该缺失值
  - 两种均值替代法也只适用于MCAR，否则会减少缺失变量的变异，进而导致（1）估计有偏，和（2）估计系数标准误变小，显著性检验无效
- 贝叶斯多重补值法（full Bayesian multiple imputation）
  - 过程
    - 任何一个具有缺失值的变量作为因变量，以其他自变量和因变量（如果与缺失变量存在相关关系的话）作为自变量进行回归，估计出该缺失变量的预测分布，并基于该预测分布，随机填补变量缺失值
    - 如果存在多个存在缺失值变量，就不断反复，直至形成第一个完整数据。如此反复的次数越多，填补计算量越大
    - 就上述过程再多次重复，得到若干个完整数据，对每个完整数据采用相同的量化方法进行分析，并对最终结果进行综合，得到最终的估计结果与统计推断
  - 优势
    - 能产生估计系数的无偏估计，以及标准误的无偏估计
      - Stata -lookup ice-
      - SPSS Missing data
  - 问题
    - 到底有哪些变量应该加入到预测缺失变量的自变量中，以及它们之间是一种怎样的关系（线性还是非线性？）
    - 适用于MAR，究竟能否用于解决完全非随机缺失问题，依然存在争议

2.13 OLS基本假设（1）

- 假设1：模型对参数而言是线性的
  - （注意）要求参数线性，而自变量可能是非线性
  - 也就是说，OLS可以用于估计非线性方程
- 一个案例

```
. reg lnincome edu edu2 exp exp2 gender

      Source |       SS       df       MS              Number of obs =    3415
-------------+------------------------------           F(  5,  3409) =  143.45
       Model |  836.216008     5  167.243202           Prob > F      =  0.0000
    Residual |  3974.41132  3409  1.16585841           R-squared     =  0.1738
-------------+------------------------------           Adj R-squared =  0.1726
       Total |  4810.62732  3414  1.40908826           Root MSE      =  1.0797

------------------------------------------------------------------------------
    lnincome |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         edu |  -.0586634   .0173248    -3.39   0.001    -.0926313   -.0246954
        edu2 |   .0068304   .0010425     6.55   0.000     .0047865    .0088743
         exp |   .0334594   .0056123     5.96   0.000     .0224555    .0444632
        exp2 |  -.0008327   .0000993    -8.39   0.000    -.0010273   -.0006381
      gender |   .6540217   .0380551    17.19   0.000     .5794086    .7286349
       _cons |   5.646952   .1033074    54.66   0.000     5.444401    5.849503
------------------------------------------------------------------------------
```

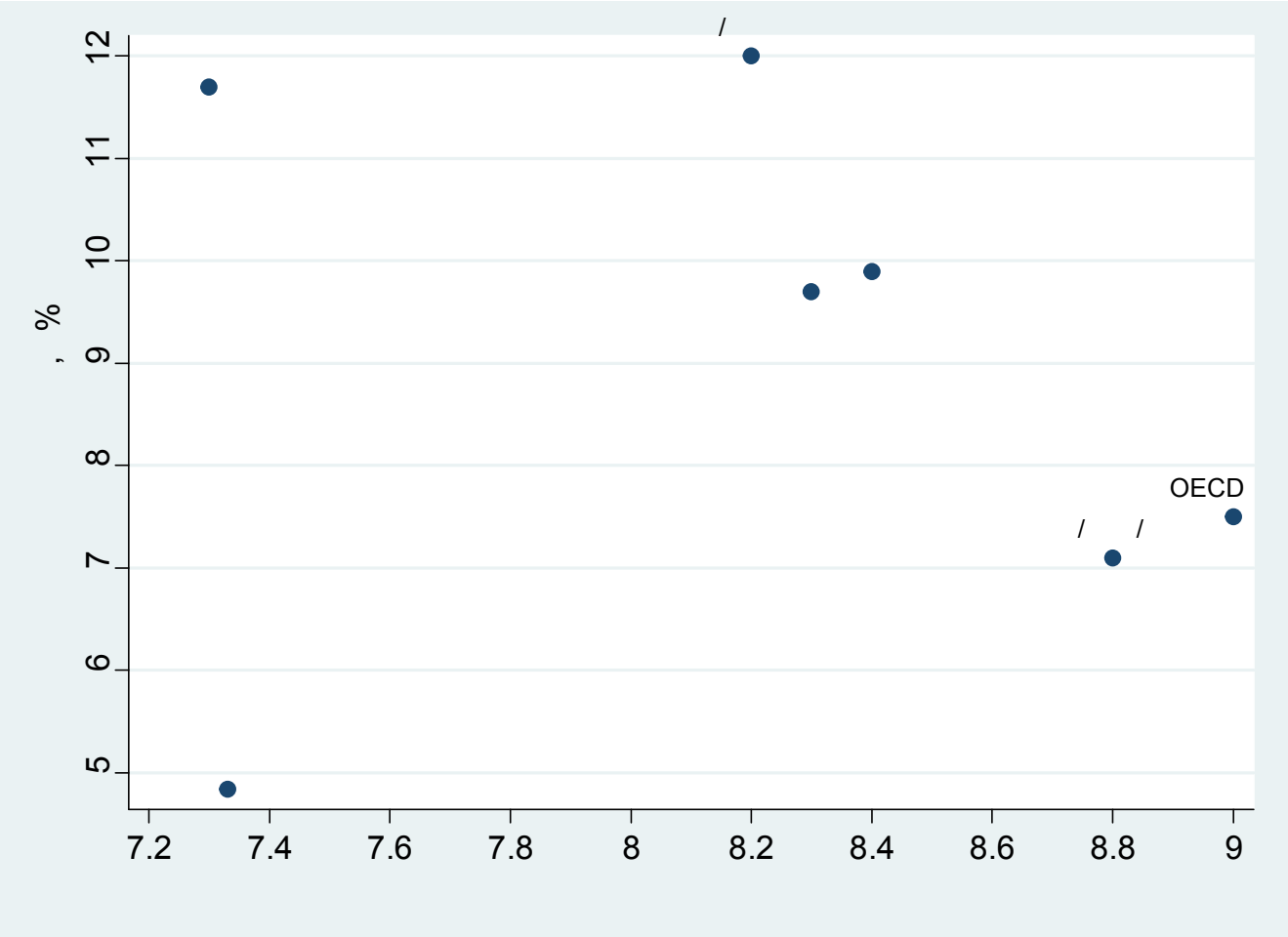
OLS基本假设（2）

- 假设2：残值项均值为零
  - 该假设包括三种含意
    - 计量模型没有遗漏任何重要自变量与其他模型识别错误（specification error）
    - 残值项与自变量之间不存在任何相关关系
      - 即若$E(e_i|x_i)=0$，必定有$cov(x_i,e_i)=0$
  - 若不满足，会导致内生问题，引起OLS有偏估计
    - 当自变量与残值项为正相关时，为高估
    - 当自变量与残值项为负相关时，为低估
  - 在无法改正模型识别错误的情况下，可使用工具变量法（两阶段回归）纠正内生性偏估问题

内生性的一个案例

- 测量教育收益率（明瑟方程）
  - $lnincome=b_0+b_1*edu+b_2*exp+b_3*exp2$
- 众多国际经验研究表明，教育之于个人和家庭收入具有极为重要的作用
  - 在农村地区
    - 能提高农民个人的劳动生产率与资源配置效率
    - 能增加农民个人的劳动迁移和非农就业的可能性
    - 能促进农民个人和家庭收入来源的多样化，以及分散家庭风险
- 以往针对中国的多数研究却发现
  - 中国的教育收益率无论是在城镇，还是在农村，都异乎寻常地低
    - 有学者曾将中国此种在经济高速发展的同时，教育收益率却常年处于较低水平的现象视为一种“特异（outlier）”现象

教育收益率的国别比较



两种解释

- 制度性解释
  - 认为是不完善的市场机制与制度性的劳动力市场分割压制了教育对于个人收入作用的进一步发挥，导致教育收益率低下
    - 然而，随多年来中国市场经济和劳动力市场环境的改善，中国教育收益率的经验结果却上升缓慢，收益率水平仍处于偏低水平
- 技术性解释
  - 认为是以往研究所采用的估计方法存在诸多计量问题，低估了中国的教育收益率

运用IV方法纠正教育收益率估计偏差

- 因遗漏个人天生能力变量与工作经验测量误差所导致的教育收益率内生偏估问题
  - 何为内生变量
    - 与残值项存在相关关系的变量称为内生变量，本例中为个人受教育水平与工作经验
  - 寻找有效的工具变量
    - 与内生变量（个人受教育水平变量相关），且对因变量个人收入无直接影响
      - 工具变量父母受教育水平、兄弟姐妹数量、家庭人口规模用于纠正因能力变量遗漏所引发的偏估问题
      - 工具变量年龄与年龄的平方用于纠正因工作经验测量误差所引发的偏估问题
  - 最终估计结果
    - 教育收益率OLS估计结果为6.5%
    - 教育收益率IV估计结果为13.1%

OLS基本假设（3）

- 假设3：残值（条件）方差同质（homoskedasticity）
  - 该条件若未得到满足，OLS的estimator仍是无偏的，但不是最优的，即不能使得估计系数的标准误达到最小。通常情况下，存在异方差时，OLS方法所估计出的系数标准误都偏大，使得原本应显著的系数检验表现为非显著

FIGURE 2.8

The simple regression model under homoskedasticity.

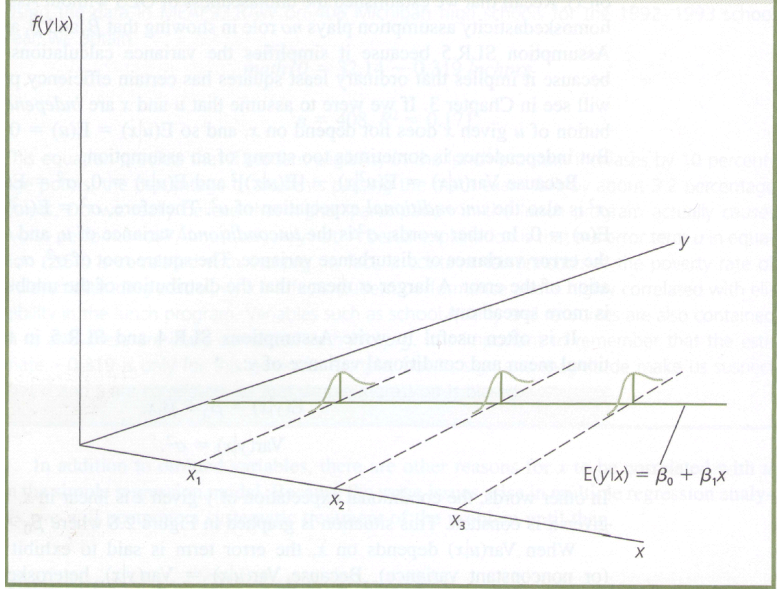
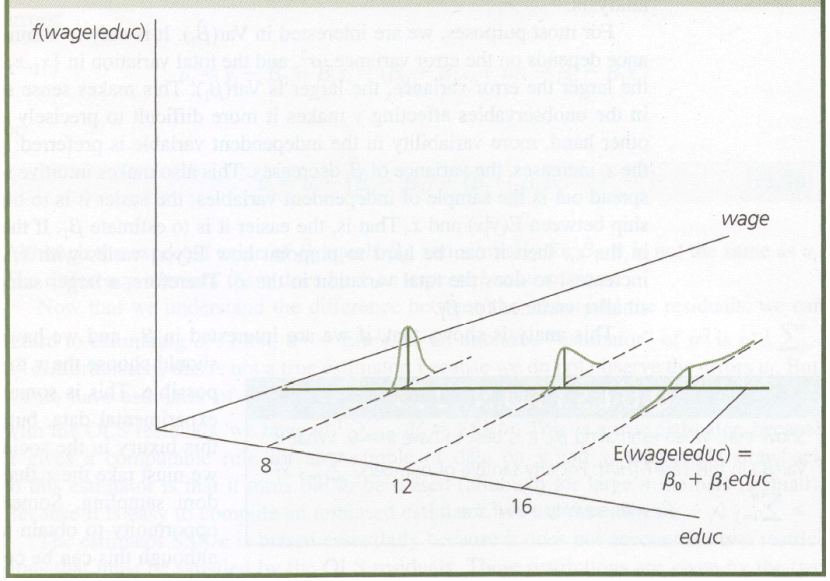


FIGURE 2.9

Var(*wage*|*educ*) increasing with *educ*.



异方差检验

- 如何检测异方差
  - 方法很多，包括观测法和检验法两种
    - 检验法的基本原理：若存在异方差，那么残值方差应当会与某些自变量相关，也就是说，残值方差会随着某些自变量的变化而发生变化。因此，我们就可以将模型所包含的所有自变量作为解释变量对残值方差进行回归，如果该回归方程中所有解释变量的估计系数都为零，那么说明不存在异方差，反之，则存在异方差
      - White test（原假设为方差同质）
        - quietly reg lnincome edu exp exp2 gender
        - White's test for Ho: homoskedasticity
        - against Ha: unrestricted heteroskedasticity
        - chi2(12)=36.37
        - Prob>chi2=0.0003
- 如何纠正异方差
  - 方法很多，包括一般最小二乘法、拆分样本回归、多层线性模型化方法、分位数回归和异方差稳健标准误，等等
    - 异方差稳健标准误
      - 可以不改变估计方法，依然采用OLS回归，并对因异方差导致的标准误错误进行纠正，但不是最优的估计量

异方差稳健标准误

```
reg lnincome edu exp exp2 gender
------------------------------------------------------------------------------
    lnincome |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         edu |   .0460535   .0067282     6.84   0.000     .0328618    .0592452
         exp |   .0175667   .0050921     3.45   0.001     .0075828    .0275505
        exp2 |  -.0005268   .0000881    -5.98   0.000    -.0006996   -.0003539
      gender |   .6304397   .0381168    16.54   0.000     .5557057    .7051738
       _cons |   5.501039   .1014969    54.20   0.000     5.302038     5.70004
------------------------------------------------------------------------------
```

```
. reg lnincome edu exp exp2 gender, vce(robust)
------------------------------------------------------------------------------
             |               Robust
    lnincome |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         edu |   .0460535   .0069224     6.65   0.000     .0324811    .0596259
         exp |   .0175667   .0054314     3.23   0.001     .0069175    .0282158
        exp2 |  -.0005268     .00009    -5.86   0.000    -.0007031   -.0003504
      gender |   .6304397   .0387719    16.26   0.000     .5544212    .7064583
       _cons |   5.501039   .1073872    51.23   0.000     5.290489    5.711589
------------------------------------------------------------------------------
```

- 比较后发现异方差问题并不是十分严重

OLS基本假设（4-5）

- 假设4：残值项之间不存在相关性
  - 若不满足，则存在自相关问题
  - 在自相关下，虽然OLS estimator仍是无偏的，但不再是最优的，它会使得估计系数标准误增大
- 假设5：自变量为固定的非随机变量，且至少取两个值以上（即需要存在一定的变异性）。此外，自变量之间不存在完全的线性关系
  - 自变量为非随机假设可以放松
  - 自变量之间必须存在一定的变异性，且越大越好
  - 自变量之间若存在完全线性关系，则估计系数的标准误将趋近于无穷大。为什么？

$$Var(\beta_1) = \frac{\sigma^2}{(1 - r_{x_1 x_2}{}^2)\sum (x_1 - \overline{x_1})^2}$$

  - 自变量之间若存在一定相关关系，但并非完全线性，此时，OLS 估计量仍然是BLUE，但估计系数标准误会增大，导致后果：
    - 一是自变量显著性难以通过检验
    - 二是估计系数的精度变差

OLS基本假设

- 假设6（optional）：残值项（或因变量）符合正态分布
  - 此假设可以放松
- GAUSS-MAEKOV Theorem
  - 如果以上1-5假设条件均得到满足，那么OLS估计量便是最优的线性无偏估计量（BLUE）
  - 注意：虽然BLUE的实现不需要假设6正态分布假设，但显著性检验的统计推断需要这一假设，因为只有残值项满足正态分布，那么OLS estimator在无数次随机抽样中才能符合正态分布，由此才能进行t检验。然而，中心极限定理又告诉我们：即便在残值不满足正态分布的条件，只要样本足够大， OLS estimator也会趋近于正态分布，其估计标准误亦会无限趋近于零。这即为大样本的一致性（consistent）特征

第一讲和第二讲参考文献

- Treiman D. J., Quantative data analysis: doing social research to test ideas
- Hill R. C & Griffith W. E., Principles of econometrics

3. Stata软件入门及其数据管理初步

内容

- 介绍Stata 11.0/12.0的基本功能、特点和窗口
- 介绍运用Stata 11.0/12.0如何实现数据管理

3.1 Stata的窗口

- 友好界面与自由界面的经典融合
  - Command Input Window
    - 用于输入命令
  - Output Window
    - 用于显示执行结果
  - Command review window
    - 用于回顾已实行过的命令，便于检查命令过程，以及重新执行或编辑命令，提高效率
  - Data management window
    - 用于检视数据库内各变量数据的基本情况，包括变量名、标签、类型与格式

3.2 初次体验Stata的便捷与快速

- 如何实现OLS回归
  - 汽车传递比对汽车价格的影响
    - sysuse auto, clear
    - reg price gear_ratio
    - sum price, detail
    - reg price gear_ratio in 1/50
    - reg price gear_ratio if foreign==1
    - reg price gear_ratio if foreign==0

Source	SS	df	MS
Model	62479817.3	1	62479817.3
Residual	572585579	72	7952577.48
Total	635065396	73	8699525.97

```
Number of obs =      74
F(  1,    72) =    7.86
Prob > F      =  0.0065
R-squared     =  0.0984
Adj R-squared =  0.0859
Root MSE      =    2820
```

price	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
gear_ratio	-2027.544	723.3597	-2.80	0.007	-3469.535	-585.5526
_cons	12278.03	2205.333	5.57	0.000	7881.776	16674.28

初次体验Stata的便捷与快速

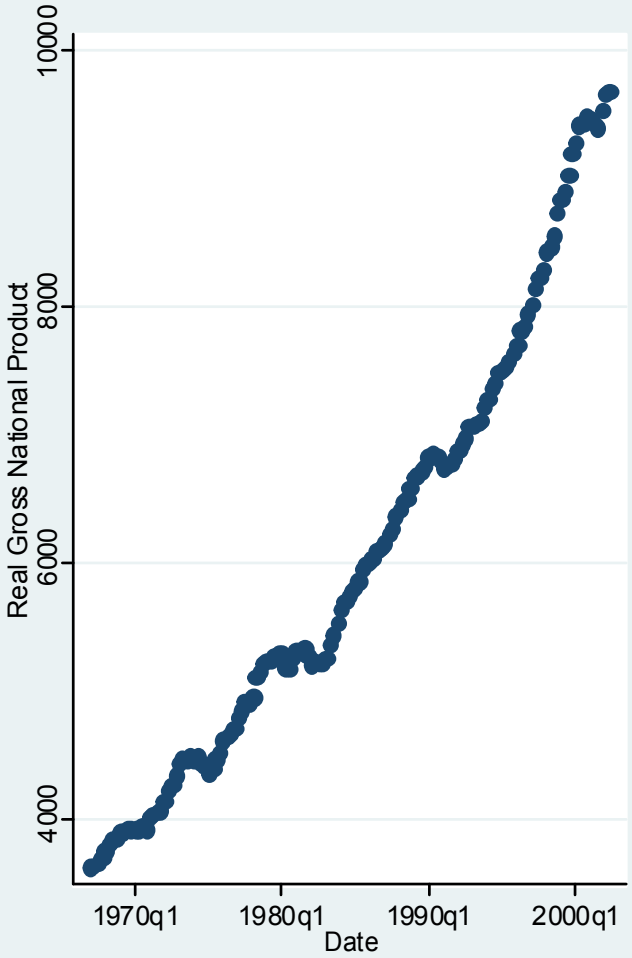
- 如何实现logit回归
  - 疾病治疗的效果评价
    - sysuse cancer, clear
    - sum died drug
    - tab drug, gen(drug)
    - logit died drug2 drug3
    - logit died drug2 drug3 if age>=60 （no convergence）

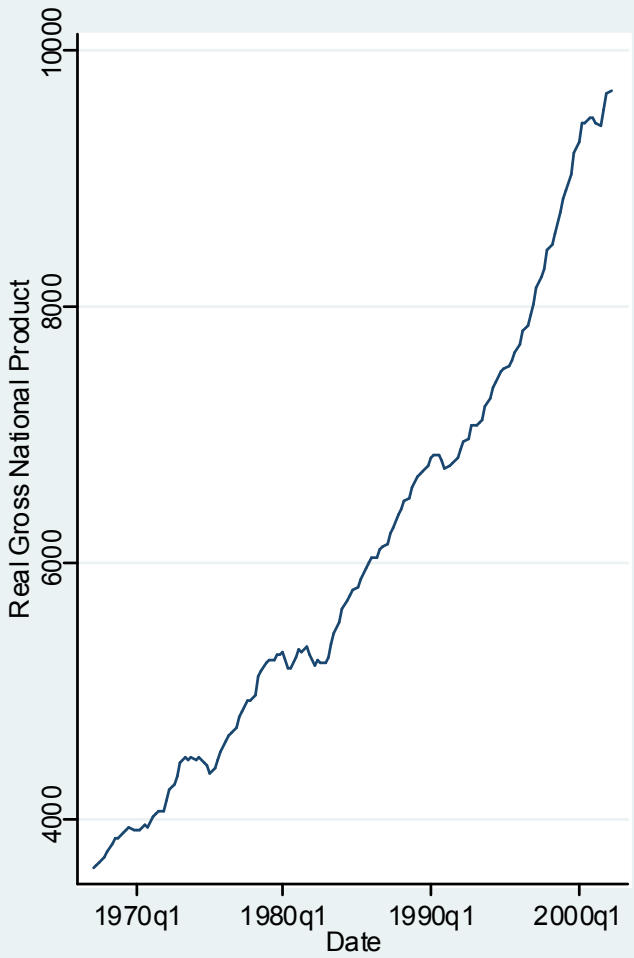
```
Logistic regression                               Number of obs   =         48
                                                  LR chi2(2)      =      16.22
                                                  Prob > chi2     =     0.0003
Log likelihood = -23.091732                       Pseudo R2       =     0.2599
```

died	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
drug2	-3.232121	1.159439	-2.79	0.005	-5.504579	-.9596628
drug3	-3.232121	1.159439	-2.79	0.005	-5.504579	-.9596628
_cons	2.944439	1.025978	2.87	0.004	.9335584	4.95532

初次体验Stata的便捷与快速

- 如何作图
  - sysuse gnp96, clear
  - twoway (scatter gnp96 date), name(a)
  - twoway (connected gnp96 date)
  - twoway (line gnp96 date), name(b)
  - graph combine a b, name(c)
  - sysuse auto, clear
  - graph bar (mean) price, over(foreign)


Real Gross National Product
4000
6000
8000
10000
1970q1
1980q1
1990q1
2000q1
Date


Real Gross National Product
4000
6000
8000
10000
1970q1
1980q1
1990q1
2000q1
Date

Stata的数据录入

- 如何手动输入数据，建立新的数据模板
  - 第一步，先输入第一个观测数据，以便于之后数据模板的制作
    - enter “data editor” mode
    - 横行输入，tab键
    - 纵行输入，enter键
  - 第二步，进入“variables management” mode
  - 第三步，定义变量名称
    - -rename-
  - 第四步，定义变量标签
    - -label variable-
  - 第五步，定义变量类型
    - stata为数值型变量提供五种数据精度
      - byte（字节），int（整数）,long（长整数），float（浮点），double（双精度）
      - 精度越低，数值运输因四舍五入后损失的信息就越多，结果越不精确
  - 第六步，定义变量格式
    - 数值型
    - 日期型
    - 字符型
  - 第七步，定义变量标签（类别变量）
    - 定义完标签后，记得要label
      - . label values race racelab

Stata的数据录入

- 如何将Excel、SPSS和SAS数据导入Stata
  - 手动导入，Ctrl＋C和Ctrl＋V
    - 适用于小规模样本数据的粘贴
  - 专用软件导入，Stattransfer
    - 适用于大型数据的转换

检查数据录入是否有误

- 昂贵的策略
  - Double data entry
- 常用的技巧
  - 利用一些常用的描述性命令寻找异常值
    - -tabulate-
    - -sum, detail-
    - -count-

4. 有限类别因变量回归分析：
Logit模型原理及应用

4.1 一个常被忽视的OLS基本假设

- 测量假设（measurement assumption）
  - 所有自变量是定距（interval）、比率（ratio）或类别（categorical）变量，并且因变量是连续的、无限制的（unbounded）和定距或比率变量。所有自变量不存在测量误差
- 如果模型自变量是有限的类别变量
  - 在实际分析中，我们通常面对的因变量实质上都是有限的
    - 例如，个人的收入多数只存在于一个有限的值域范围内，个人的年龄都限制在一定得寿命范围内，个人所受教育亦是有限的，但这些变量的取值毕竟是可以近似于连续，因此将这些变量看成是连续无限的变量也是合理的
    - 然而，有一些变量却明显不是连续的，属于类别变量（无序和有序）
      - 例如，个人性别、个人是否就学某一层级教育，个人是否纳税
        - 如果以这些变量作为因变量进行OLS回归，就会严重违背OLS有关测量的基本假设

4.2 线性概率模型（Linear Probability Model）

- 一个案例
  - 收入对于个人是否购买最新款Apple Ipad是否具有影响，
    - 自变量为个人收入，因变量为个人购买Ipad的决定，0表示不购买，1表示购买
    - 如果使用OLS，能形成一个线性概率回归方程：

$$E(y_i \mid x_i) = E(\alpha + \beta \cdot x_i + e_i) = \alpha + \beta \cdot x_i$$

$$\because E(y_i \mid x_i) = P(y_i = 1 \mid x_i)$$

$$\therefore P(y_i = 1 \mid x_i) = \alpha + \beta \cdot x_i$$

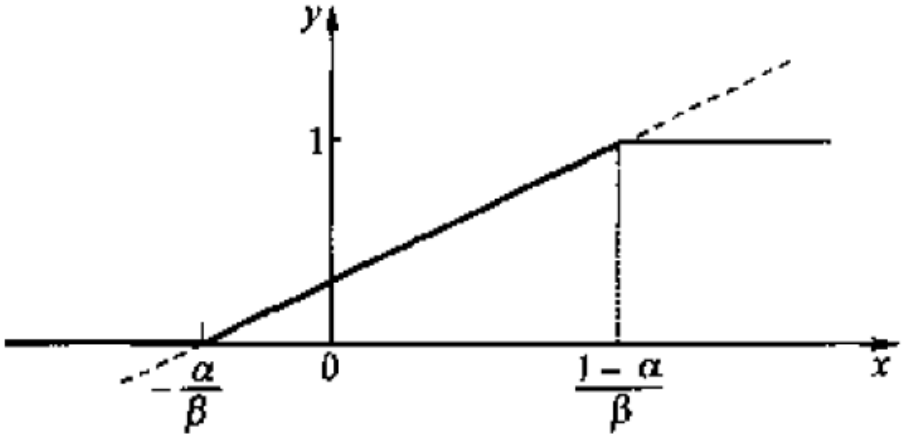
    - 因变量即为个人购买Ipad的条件概率，自变量的估计系数即表示当个人收入每增加一单位，个人购买Ipad的概率会上升多少
      - 如此做，好像也实现了我们的研究目的
      - 是否存在问题呢？

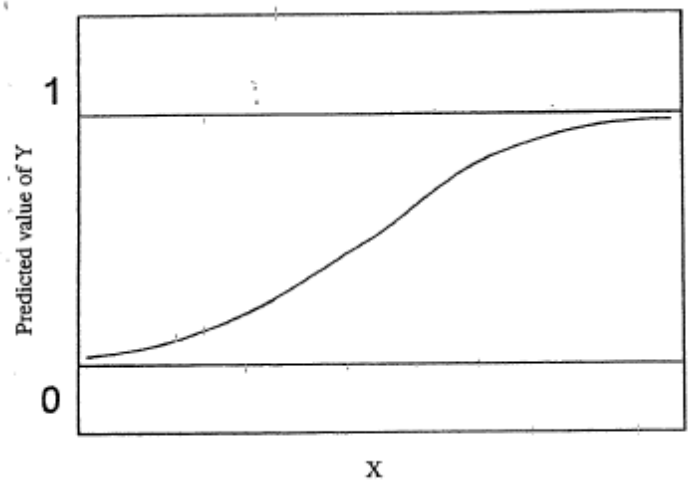
4.3 线性概率模型存在的问题

- 异方差问题
  - 若因变量只取值0和1值，那么线性概率模型的残值方差为：
    $Var(e_i) = P(y_i = 1 | x_i) \times P(y_i = 0 | x_i)$
  - 残值方差有赖于因变量的条件概率值，而该概率值又是以自变量取值为条件的，因此残值方差势必与自变量相关，为异方差
    - 估计系数标准误非最优，t显著性检验无效
- 预测值可能会超出因变量（0,1）的值域范围
  - 例如，运用线性概率模型OLS回归的结果如下：
    - y=-0.41+0.064*x，x的值域在（8,29）
      - 当x=8时，y=0.1，当x=29时，y=1.45（超出1）
- 残值不符合正态分布
  - 在小样本下，t显著性检验无效

线性概率模型存在的问题

- 函数识别错误
  - 自变量与因变量概率之间的关系应当是非线性还是线性的
    - 因变量概率的值域在（0，1）之间，因此随自变量值无限增大或减少，因变量概率值应当会无限接近1或0
    - 如下左图，对于两分类别因变量的分析应使用非线性函数
    - 如下右图，自变量x与因变量y概率之间呈S状变化，曲线上下两端无限趋近于0和1，此图形类似于一种随机变量的累积分布曲线





4.4 Logistic regression model

- 什么是Logistic regression model
  - 以Logistic分布来近似两分类别因变量概率的变化曲线
    - Logsitic分布的均值为0，方差等于 $\pi^2/3$ ，将该分布应用于两分因变量回归的最大好处在于，它能使得累积分布函数最终取得一个简单的函数公式

Logistic回归原理

- 假定有一个连续反应变量g表示某一事件发生的可能性，其值域无限制。y事件从不发生（0）到发生（1）的变化需g达到一个临界值（假设该临界值为0）
  - 当g>0时，y=1；当g≤0时，y=0
  - 假定g与自变量x之间存在线性关系，有：

$$g_i = \alpha + \beta \cdot x_i + e_i$$

$$P(y_{i=1} = 1 \mid x_i) = P(\alpha + \beta \cdot x_i + e_i > 0) = P(e_i > -\alpha - \beta \cdot x_i)$$

假定因变量y的分布为对称分布，于是有

$$P(y_{i=1} = 1 \mid x_i) = P(e_i \le \alpha + \beta \cdot x_i)$$

假定残值的累积分布函数为F，于是有：

$$P(y_{i=1} = 1 \mid x_i) = F(e_i = \alpha + \beta \cdot x_i)$$

此时我们需假设残值的分布具体是什么（应是对称的），可以假设它为Logistic分布，于是有：

$$P(y_{i=1} = 1 \mid x_i) = \frac{1}{1 + \exp(-e_i)}$$

当残值e_i趋近于无穷大时，$P(y_{i=1} = 1 \mid x_i)$无限接近于1，反之，无限接近于0

将$e_i = \alpha + \beta \cdot x_i$代入上式，可得：

$$P(y_{i=1} = 1 \mid x_i) = \frac{1}{1 + e^{-(\alpha + \beta \cdot x_i)}}$$，进一步转换，可得

$$\ln(\frac{p_i}{1 - p_i}) = \ln（\text{odds}）= \alpha + \beta \cdot x_i$$，这就是Logisitc回归模型式，亦称为Logit回归模型

Logistic回归的特点

- 因变量是因变量事件发生与不发生概率之比的对数值
  - 两个概率之比称为odds，有多种翻译，包括发生比、概率比、几率等，而log（odds）被称为对数发生比
    - 该值越大，说明事件发生的可能性越高
  - 该概率比对数值的值域在（0,1）之间

- 对数概率比对于参数是线性的，但自变量对事件发生的概率的影响是非线性的
  - 如果你想得到某一自变量对事件发生概率的边际影响，那么需要进行一定转换计算（如何用Stata绘图，之后会有介绍）

Logistic模型的估计

- 最大似然法
  - 对数似然函数（log-likelihood function）和迭代计算法
    - 即通过对估计系数α和β不断的取值，使得对数似然函数值实现最大
- 最大似然估计法对样本容量的要求
  - "It is risky to use ML with samples samller than 100, while samles over 500 seem adequate. These values should be rasised depending on characteristics of the model and the data. First, if there are many parameters, more observations are needed... A rule of at least 10 observations per parameters seems reasonable... This does not imply that a minimum of 100 is not needed if you have only two parameters. Second, if the data are ill-conditioned (e. g., independent variabless are highly collinear) of if there is little variation in the dependent variable (e.g., nearly all the outcomes are 1), a large sample is required. Third, some models seem to require more observations (such as the ordinal regression model or the zero-inflated count models)" (Long, 1997:54)

4.5 如何在Stata中实现Logistic回归分析

- 两分类别变量的逻辑回归
  - -logit-
  - 两分类别变量有两种结果，Stata进行logit回归默认以赋值小的那个结果作为参照结果（baseoutcome）进行回归
  - 例如，我们想探讨性别与受教育年限对农民外出从业可能性的影响，因变量是农民是否外出从业，0为在家从业，1为外出从业。此时逻辑回归的因变量即为：

$$\ln(p_{emigrant} / p_{\text{hom}e}) = \ln(p_{emigrant} /(1 - p_{emigrant}))$$

Logit回归估计结果

```
. logit emigrant gender edu

Iteration 0:   log likelihood = -2054.3767
Iteration 1:   log likelihood = -1983.1605
Iteration 2:   log likelihood = -1981.9849
Iteration 3:   log likelihood = -1981.9839
Iteration 4:   log likelihood = -1981.9839

Logistic regression                               Number of obs   =       3610
                                                  LR chi2(2)      =     144.79
                                                  Prob > chi2     =     0.0000
Log likelihood = -1981.9839                       Pseudo R2       =     0.0352
```

emigrant	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
gender	.4099229	.0800412	5.12	0.000	.2530451	.5668008
edu	.1078544	.0110317	9.78	0.000	.0862327	.129476
_cons	-2.179571	.1087227	-20.05	0.000	-2.392663	-1.966478

4.6 如何解释Logistic回归的估计结果

- 用对数概率比进行解释
  - 由于对数概率比没有实际意义，因此，推荐用趋势方法来解释估计结果
    - 由于因变量是对数概率比，这一数值本身没有太多现实含义，因此，我们一般不用“自变量每变化一单位，某一事件的对数概率比会变化多少”这一方法来解释，而是比较笼统地用“随自变量变大，某一事件发生的可能性会随之增大或减少”
  - 如本例，根据逻辑回归估计结果，我们可以说男性外出从业的可能性显著高于女性，受教育程度越高的农民外出从业的可能性越高，而不说，农民个人受教育年限每增加一年，农民外出从业的对数概率比会增加了0.11

如何解释Logistic回归的估计结果

- 用概率比进行解释

$$\ln(\frac{p_i}{1-p_i}) = \ln(odd_i) = \alpha + \beta \cdot x_i \Rightarrow odd_i = e^{(\alpha+\beta\cdot x_i)}$$

1、如果自变量是虚拟变 量，取值分别为0和1，那么

当$x=0$时，$odd_{x=0} = e^{(\alpha)}$

当$x=1$时，$odd_{x=1} = e^{(\alpha+\beta)}$

$odd_{x=1} / odd_{x=0} = e^{\beta}$

e^{β}表示随自变量由0变化为1，因变量取值为1相对于取值为0之间的概率比会变化e^{β}倍

2、如果自变量是连续变 量，那么

$\partial odd_i / \partial x_i = e^{(\alpha+\beta\cdot x_i)} \cdot x_i$

$(\partial odd_i / \partial x_i) / e^{(\alpha+\beta\cdot x_i)} = \beta$

$(\partial odd_i / \partial x_i) / odd_i = \beta$

β表示随自变量每增加一 单位，因变量取值为1相对于取值为0之间的概率比会变化β*100%

e^{β}

如何解释Logistic回归的估计结果

- 概率比解释方法。如本例，根据逻辑回归结果
  - 男性外出从业相对于在家从业的概率比是女性的1.51倍
  - 个人受教育年限每增加一年，农民外出从业相对于在家从业的概率比会显著增加10.8%
  - 在-logit-后面加上or选项，Stata会自动算出各回归系数的值

如何解释Logistic回归的估计结果

- 用概率进行解释
  - 虽然逻辑回归函数是线性函数，但只是有关概率比对数的线性函数，对于事件发生概率来说，该函数是非线性函数

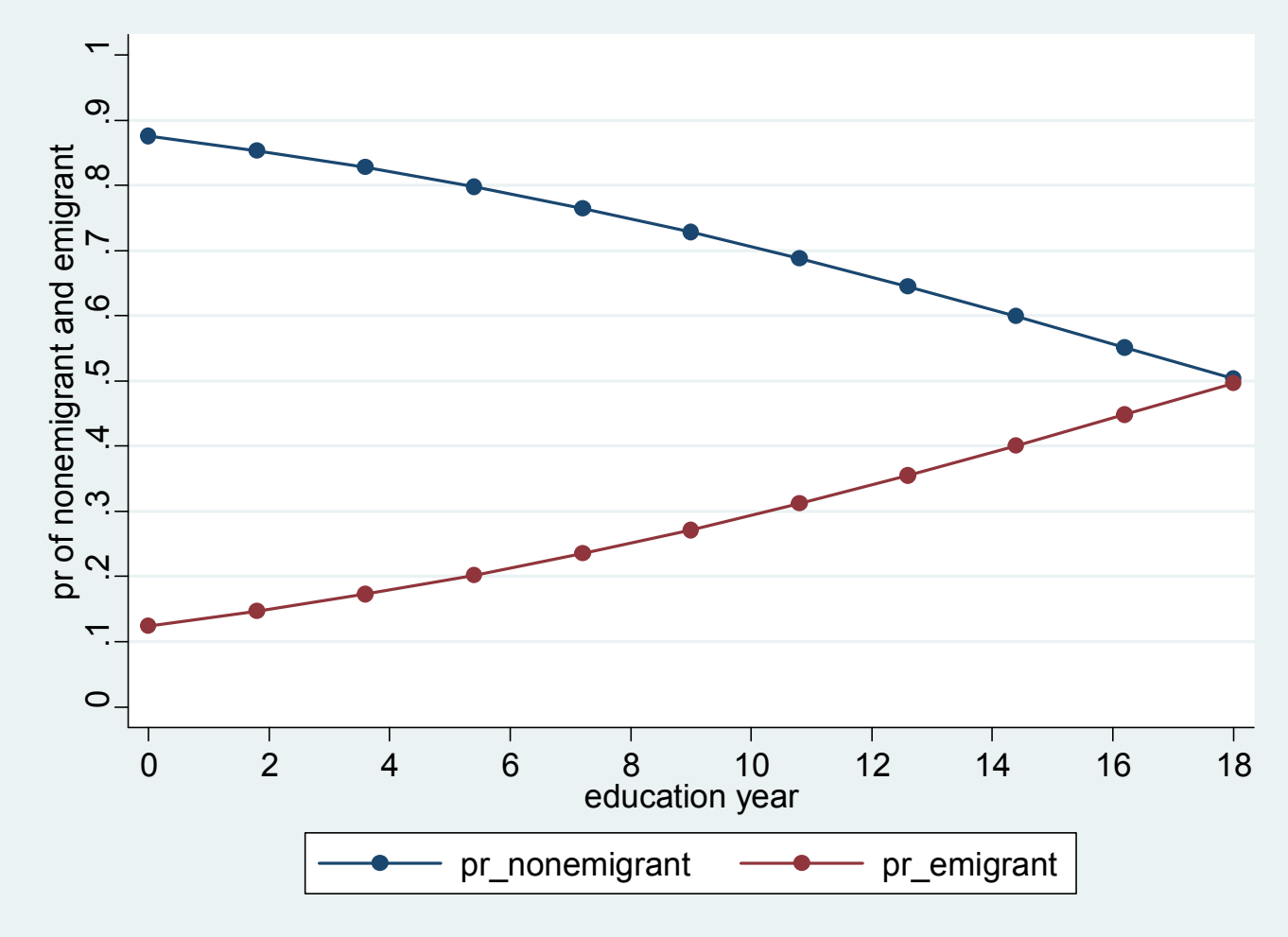
$$P(Y=1)=\frac{e^{(\alpha+\sum\beta\cdot x_i)}}{1+e^{(\alpha+\sum\beta\cdot x_i)}}$$

  - 在Logistic回归模型中，自变量对于事件发生概率的影响势必会随自变量的变动而发生变动，也就是说，自变量对事件发生概率的边际效应是变动的。如果自变量是连续变量，那么在我们对估计结果进行概率变动分析时
    - 坚持以自变量对事件发生概率的影响进行解释，会面临巨大的计算量，并且作用不大
    - 一个好的策略是
      - 借助Stata计算出在每一个自变量取值下因变量事件发生概率的预测值，并利用它绘制出一个随自变量变化，因变量事件发生概率的变动趋势图，这会给读者带来一些有价值的直观感受（next slide）

如何利用Stata作概率变化图

- quietly logit emigrant gender edu
- prgen edu, generate(all) rest(mean)
- label var allx "education year"
- label var allp0 "pr_nonemigrant"
- label var allp1 "pr_emigrant"
- graph twoway connected allp0 allp1 allx, ytitle("pr of nonemigrant and emigrant") ylabel(0(.1)1) xtitle("education year") xlabel(0(2)18)

如何利用Stata作概率变化图



4.7 Logistic回归模型的整体有效性检验

- 对模型整体是否有效的 χ^2 检验
  - 先计算出一个只包括截距，所有自变量回归系数强迫其为零的模型的log likelihood（对数似然）值，记为：
    - L（$M_{intercept}$）
  - 再计算出包括所有自变量的完全模型的log likelihood值，记为：
    - L（M_{full}）
  - 构造函数：LR=2L（M_{full}）-2L（$M_{intercept}$）
    - 该函数符合 $\chi^2(k)$ 分布，自由度K等于full model中的自变量个数
  - 卡方检验的零假设为：full model中所有的变量系数都为零，即$M_{intercept}$和M_{full}无差别
  - Stata实现命令
    - -fitstat-

4.8 Logistic回归模型的拟合优度（Fit of Goodness）

- Logsitic回归模型有多种拟合优度的计算指标
  - 皮尔逊 χ^2 检验及其修正Hosmer-Lemeshow拟合优度
  - 信息测量指标（information measures）
  - 类拟合优度指标（*Pseudo-*R^2）这些指标均基于对数似然函数进行计算，其原理与OLS拟合优度指标有本质不同，因此称为“类”拟合

皮尔逊卡方检验及其修正

- 皮尔逊卡方检验原理
  - 假定Logistic回归模型中的自变量都为类别变量
    - 例如我们想观察性别（男与女）与就读大学层次（重点与非重点）对个人是否能寻找到工作可能性的影响，此时模型自变量都为两分变量，个数为2，那么我们就可以得到4种人群组合，构建Logistic模型如下：

$$\ln\left(\frac{p}{1-p}\right)=\alpha+\beta_1\cdot gender+\beta_2\cdot univ$$

    - 通过如上模型回归，我们可以得到每一类人群找到工作的预测概率，将这一预测概率与每类人群人数相乘，就可以得到每类人群找到工作的频数，而皮尔逊卡方检验就是将模型的预测频数与实际观测频数相对比进行检验
    - 通过模型构造，可以得到一个符合卡方分布的统计量，可以进行检验，零假设是预测频数与观测频数无显著差别，如果皮尔逊卡方检验的卡方值越小，那么卡方检验势必不显著，这也就表明预测频数与观测频数有很好的吻合度，模型拟合度较高

协变量	n_i	观测结果		预测结果	
		是	否	是	否
重点大学					
男性	88	54	34	51.91	36.09
女性	44	9	35	11.09	32.91
非重点大学					
男性	462	205	257	207.09	254.91
女性	406	67	339	64.91	341.09

皮尔逊卡方检验及其修正

- 皮尔逊卡方检验的缺点
  - 如果模型中除类别自变量外，还有其他连续性自变量，人群分组数量就会变得非常多，而每一类别组内个体数量会变得很少，此时皮尔逊卡方检验就不再适用了。
    - Hosmer和Lemeshow就其进行了修正，HL检验的原理与皮尔逊相同，唯一不同在于检验统计量构造有差别，该方法克服了皮尔逊卡方检验严重依赖自变量为类别变量的缺点
      - HL检验的零假设依然为预测和观测频数无显著差异，检验的卡方值越大，表明预测和观测频数无显著差异
    - Stata的实现命令
      - -estat gof-

皮尔逊卡方检验及其修正

```
. do "C:\Users\Huang\AppData\Local\Temp\STD00000000.tmp"

. logit emigrant gender edu

Iteration 0:   log likelihood = -2054.3767
Iteration 1:   log likelihood = -1983.1605
Iteration 2:   log likelihood = -1981.9849
Iteration 3:   log likelihood = -1981.9839
Iteration 4:   log likelihood = -1981.9839

Logistic regression                               Number of obs   =       3610
                                                  LR chi2(2)      =     144.79
                                                  Prob > chi2     =     0.0000
Log likelihood = -1981.9839                       Pseudo R2       =     0.0352

------------------------------------------------------------------------------
    emigrant |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      gender |   .4099229   .0800412     5.12   0.000     .2530451    .5668008
         edu |   .1078544   .0110317     9.78   0.000     .0862327     .129476
       _cons |  -2.179571   .1087227   -20.05   0.000    -2.392663   -1.966478
------------------------------------------------------------------------------

. estat gof

Logistic model for emigrant, goodness-of-fit test

       number of observations =      3610
 number of covariate patterns =        38
             Pearson chi2(35) =        61.39
                  Prob > chi2 =         0.0038
```

信息测量指标：AIC和BIC

- 该类指标可以用于检验nested和nonnested模型之间的优劣
  - AIC（Akaike's Information Criterion）

$$AIC = \frac{\{-2\ln L(M_k) + 2P_k\}}{N}$$

    - $L(M_k)$是模型M_k的likelihood值，P_k是模型估计系数（包括截距和自变量估计系数）的个数
    - 如果有两个模型，一个是nonnested模型，一个是nested模型，在前者模型中，有部分自变量是后者模型所不包含的，并且除自变量有差别外，其他条件相同，那么我们可以分别计算出这两个模型的AIC值，AIC值越小，说明该模型的拟合度越高，是better-fitting model
    - AIC统计量的另一个变形是用样本容量N乘以AIC，即为N*AIC
  - BIC（Bayesian Information Crierion）

$$BIC = -2 \cdot \ln L(M_k) - df_k \cdot \ln N$$

    - df是$L(M_k)$的自由度，N为样本容量
    - BIC通常为负值，越小表明模型拟合度越高。同样假定有两个nested和nonnested模型，将它们的BIC分别算出并相减，如果结果小于0，说明nested模型的拟合度更好，如果小于0，说明nonnested模型拟合度更好
  - Stata实现命令
    - -fitstat-

类拟合优度

- McFadden's R^2
  - 两种模型构建之前的模型整体有效性检验

$$R^2_{MCF} = 1 - \frac{\ln L(M_{full})}{\ln L(M_{intercept})} \qquad (adjusted)\, R^2_{MCF} = 1 - \frac{\ln L(M_{full}) - k}{\ln L(M_{intercept})}$$

    - R^2_{MCF}值域在（0,1）之间。当full model和intercept model之间无任何差别时，两个模型的log likelihood完全相等，此时R^2_{MCF}等于0。R^2_{MCF}不可能完全等于1，只能无限接近于1
    - R^2_{MCF}的修正
      - 该指标由于会随着自变量个数的增加而不断增大，因此需进行调整
    - Stata的实现命令
      - -fitstat-

类拟合优度

- Cox & Snell's R^2
  - R^2_{MCF}又称为*Maximum likelihood* R^2，该指标原理与之前的McFadden's R^2类似，其性质基本相同

$$R^2_{MCF} = 1 - \left\{ \frac{L(M_{\text{intercept}})}{L(M_{full})} \right\}^{2/N}$$

- Cragg & Uhler's R^2
  - 由于Cox & Snell's R^2最大值不是1，而是$1 - L(M_{\text{intercept}})^{2/N}$
  - 因此，标准化后为：

$$R^2_{C\&U} = \frac{R^2_{MCF}}{\max R^2_{MCF}} = \frac{1 - \left\{ L(M_{\text{intercept}}) / L(M_{full}) \right\}^{2/N}}{1 - L(M_{\text{intercept}})^{2/N}}$$

- Stata实现命令
  - -fitstat-

Count & adjusted count R^2

- 原理
  - 假定个人行为有两种结果（例如，找到工作和没找到工作）根据Logistic回归结果，我们可以为每一个观测单位预测一个概率，假定若该预测概率大于0.5，该观测单位便会采取行动，而若该预测概率小于等于0.5，则不会采取行动。将这一预测的行为结果与实际发生的行为结果进行对比
  - count R^2就是将那些预测正确的case数除以样本总case数的一个比例值，假定n_{jj}表示对某一行为结果j预测正确的数量，那么，其公式为：

$$R^2_{count} = \frac{1}{N}\sum_j n_{jj}$$

  - 计数拟合指标常常会被视为是评价一个模型预测准确度的指标，但值得注意的是，如果我们对于模型到底应该包含哪些自变量没有太多知识判断的话，那么可能你随便加入一些自变量进行到Logistic回归模型中，也可能会获得很高的预测正确率，特别是在两分逻辑回归中，非此即彼，很容易得到较高的预测正确率，因此提出了一种修正后的计数拟合度指标（adjusted count R^2，该指标计算具体计算参见参考书Long & Freese（2006）：111）

拟合优度的测算

```
. quietly logit emigrant gender edu

. fitstat

Measures of Fit for logit of emigrant

Log-Lik Intercept Only:      -2054.377   Log-Lik Full Model:          -1981.984
D(3607):                      3963.968   LR(2):                         144.786
                                         Prob > LR:                       0.000
McFadden's R2:                   0.035   McFadden's Adj R2:               0.034
ML (Cox-Snell) R2:               0.039   Cragg-Uhler(Nagelkerke) R2:      0.058
McKelvey & Zavoina's R2:         0.067   Efron's R2:                      0.035
Variance of y*:                  3.526   Variance of error:               3.290
Count R2:                        0.742   Adj Count R2:                   -0.006
AIC:                             1.100   AIC*n:                        3969.968
BIC:                        -25582.639   BIC':                         -128.403
BIC used by Stata:            3988.542   AIC used by Stata:            3969.968
```

4.9 如何在Stata中实现多项类别逻辑回归

- 什么是多项类别变量的逻辑回归
  - 如果因变量不是两分类别变量，而是超过2个结果的多项类别变量，同样也可以运用Logistic模型进行回归分析
  - 如果因变量是n项类别变量，那么以其中一种结果作为参照结果，与其他结果两两对比，得到n-1个逻辑回归方程，分别进行回归。每一个方程的因变量是n-1个结果的概率与参照结果概率之比的对数
- Stata的实现方法
  - -mlogit-
  - mlogit employment gender edu, baseoutcome(1)
    - 若不输入选项outcome(#)，Stata默认是以频数最多的结果作为参照结果
    - option "rrr"将估计系数β转换成exp(β)，注意在-logit-命令中此选项命令为“or”
    - mlogit employment gender edu, baseoutcome(1) rrr

一个多项类别逻辑回归的例子

- 考察性别与受教育年限对农民从业的影响
  - 农民从业类型分为纯务农从业、混合从业和纯非农从业三种，赋值分别为1、2和3
  - 自变量包括个人性别和受教育年限
  - 以农民纯务农为参照结果，考察农民选择混合从业（相对于纯务农）和选择纯非农从业（相对于纯务农）的可能性会受到个人性别和受教育年限怎样的影响
  - 估计结果
    - 男性选择混合从业和纯非农从业（相对于纯务农）的概率比分别是女性的1.9倍和2.5倍
    - 个人受教育水平每增加一年，农民选择混合从业（相对于纯务农）的概率比会显著增加2.8%（在P=0.1置信水平下），农村选择纯非农从业（相对于纯务农）的概率比会显著增加17.7%（在P=0.01置信水平下）

一个多项类别逻辑回归的例子

```
. mlogit employment gender edu, baseoutcome(1) rrr

Iteration 0:   log likelihood =  -3304.023
Iteration 1:   log likelihood = -3068.7528
Iteration 2:   log likelihood = -3064.8414
Iteration 3:   log likelihood = -3064.8402
Iteration 4:   log likelihood = -3064.8402

Multinomial logistic regression                   Number of obs   =       3610
                                                  LR chi2(4)      =     478.37
                                                  Prob > chi2     =     0.0000
Log likelihood = -3064.8402                       Pseudo R2       =     0.0724
```

employment	RRR	Std. Err.	z	P>|z|	[95% Conf.	Interval]
´¿ÎñÅ©´ÓÒµ	(base outcome)					
»ì°Ï´ÓÒµ						
gender	1.898668	.219968	5.53	0.000	1.512983	2.38267
edu	1.028294	.0154706	1.85	0.064	.9984152	1.059068
_cons	.3177133	.0380686	-9.57	0.000	.2512141	.4018156
´¿·ÇÅ©´ÓÒµ						
gender	2.513006	.2115632	10.95	0.000	2.130754	2.963835
edu	1.193608	.0137996	15.31	0.000	1.166866	1.220964
_cons	.4035717	.0394663	-9.28	0.000	.3331805	.4888346

```
. mlogit employment gender edu, baseoutcome(1)

Iteration 0:   log likelihood =  -3304.023
Iteration 1:   log likelihood = -3068.7528
Iteration 2:   log likelihood = -3064.8414
Iteration 3:   log likelihood = -3064.8402
Iteration 4:   log likelihood = -3064.8402

Multinomial logistic regression                   Number of obs   =       3610
                                                  LR chi2(4)      =     478.37
                                                  Prob > chi2     =     0.0000
Log likelihood = -3064.8402                       Pseudo R2       =     0.0724
```

employment	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
´¿ÎñÅ©´ÓÒµ	(base outcome)					
»ì°Ï´ÓÒµ						
gender	.6411525	.1158538	5.53	0.000	.4140831	.8682218
edu	.0279015	.0150449	1.85	0.064	-.001586	.057389
_cons	-1.146606	.1198205	-9.57	0.000	-1.38145	-.9117619
´¿·ÇÅ©´ÓÒµ						
gender	.9214798	.0841873	10.95	0.000	.7564758	1.086484
edu	.1769809	.0115612	15.31	0.000	.1543213	.1996404
_cons	-.907401	.0977926	-9.28	0.000	-1.099071	-.715731

第四讲参考文献

- Menard, S. W. (2002). Applied logistic regression analysis. Thousand Oaks: Sage（此书有中文版，题为《应用Logistical回归分析》）.
- Long, J. S., & Freese, J. (2001). Regression models for categorical dependent variables using Stata. TX: Stata press.
- O'Connell, A. A. (2005). Logistic regression models for ordinal response variables: Sage.

5. 多层线性模型化方法

Contents

1 什么是 **HLM?**

2 我们为什么需要运用 **HLM?**

3 **HLM**方法的基本原理

4 **HLM**方法在教育财政研究中的具体运用

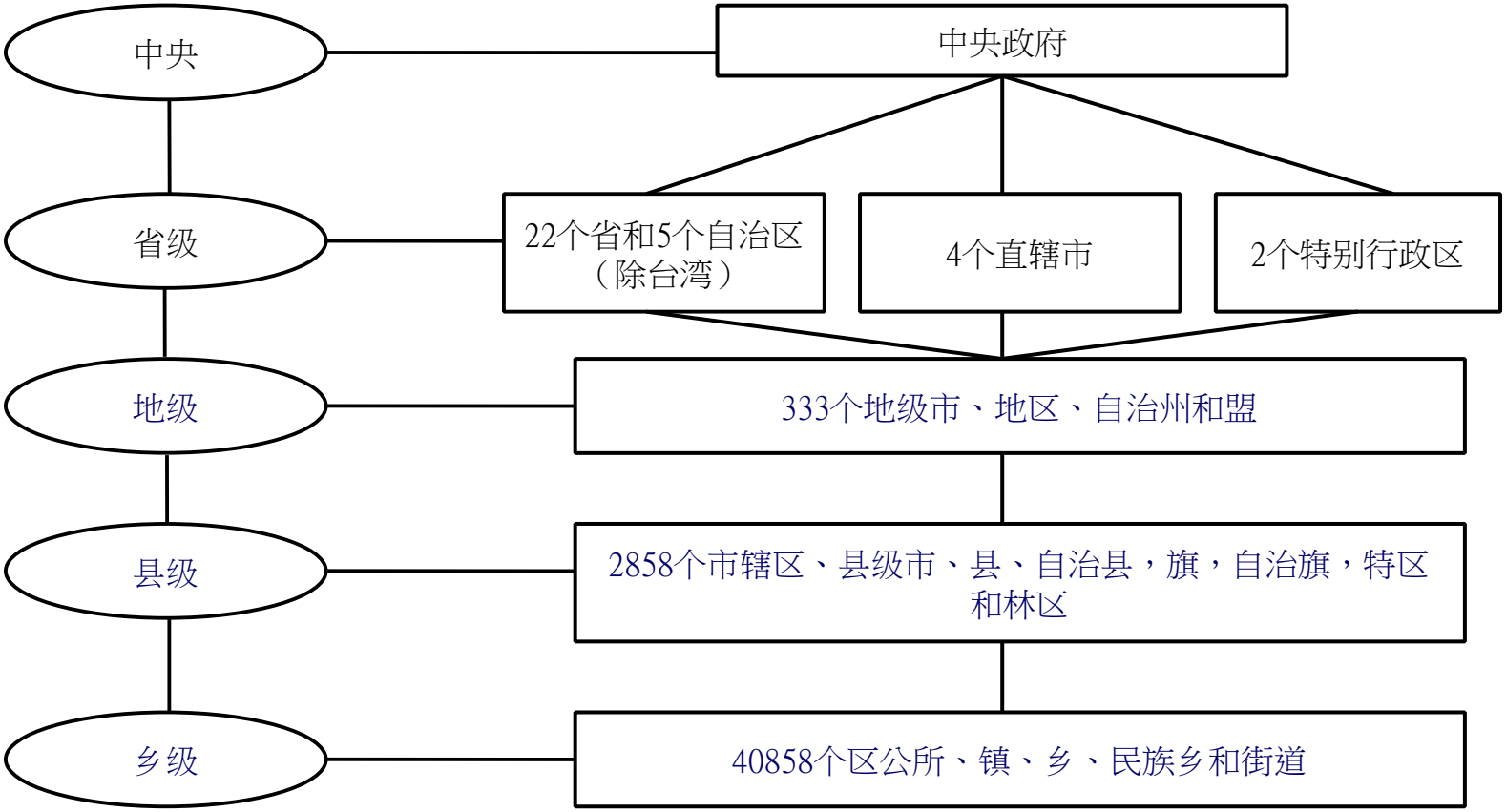
5.1 什么是多层线性模型化方法？

- 多层线性模型化方法（HLM）
  - 亦称为多层次模型化方法（Multilevel Modeling），它是应用于<u>多层数据结构</u>（hierarchical data structure）分析的一种统计方法

什么是多层数据结构？

- 多层数据结构有别于传统的数据结构，它是指较低层次数据嵌套于（nested）更高层次数据之内的一种层层镶嵌的数据结构
  - 此种数据结构是社会科学研究中十分常见的一种数据结构，在我们财政学研究尤为常见（next slide）
  - 与以往常用的单层数据相比，多层数据的优势
    - 能提供更为丰富的数据信息
      - 例如，在对政府财政支出的研究，传统的方法一般多采用全国数据或省级数据进行分析，此种分析只能告诉我们一个代表性的国家或省的财政支出作为一个整体呈现何种特征，而无法告诉我们其内部构成是否与整体呈现出相似的特征
    - 能实现更为深入的计量分析
      - 传统的方法一般仅能告诉截距项和斜率项的估计值，一般假定估计系数非随机，而HLM方法可以实现更深入的分析，允许估计系数随机变动，并分析这些估计系数会呈现出怎样变动特征

嵌套式的多层数据结构（2009年全国行政区划）



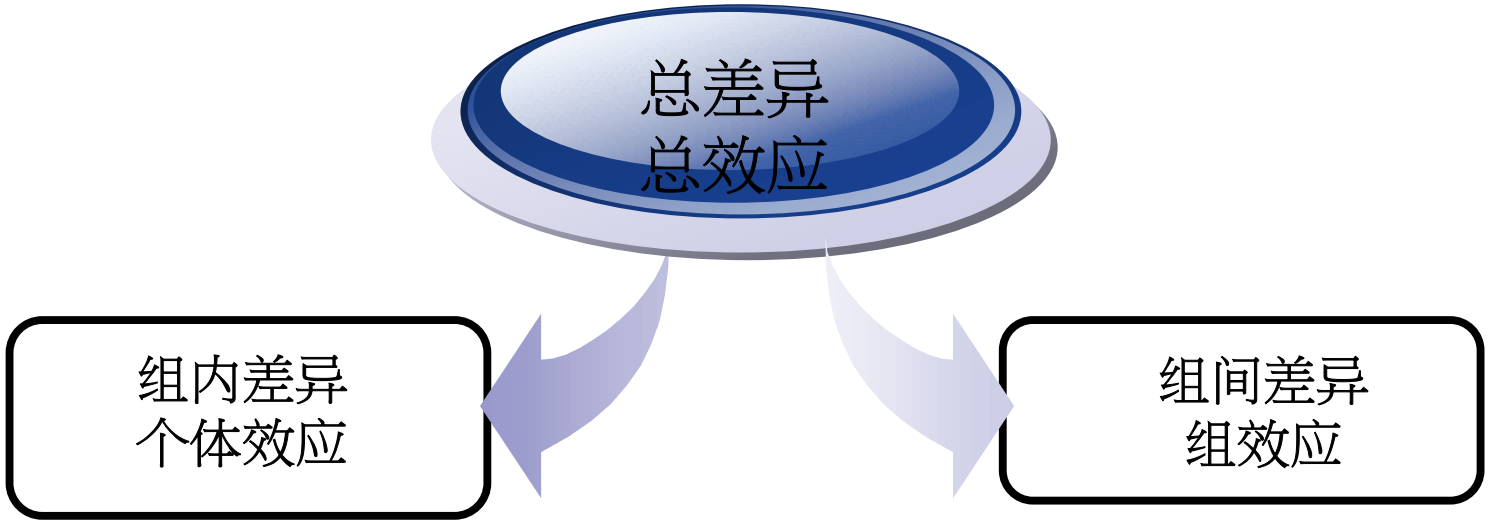
5.2 传统OLS方法应用于多层数据结构的局限性

- 分析需要局限
  - OLS无法处理和应对更为丰富和更为复杂的数据信息，并满足进行更深入计量分析的要求
- 技术性局限
  - 在多层数据结构下，OLS方法的基本假设无法得到满足

非技术性局限

- OLS无法同时处理和分析个体效应和组效应
  - 在实现中，个体行为及其后果既会受到个体特征的影响（即个体效应），又会受到其所身处环境的影响（即组效用）
    - 例如，
      - 个体消费习惯既会受到个体收入水平的影响，亦会受到个体所处收入阶层消费水平和习惯的影响
      - 各地方政府在财政支出水平上既会受到地方财政收入水平的影响，亦会受到省内财政体制安排的影响
      - 学生个体学业成就既会受到个体天生能力和努力水平的影响，又会受到班级学习风气的影响
  - 相对应地，个体行为及其后果的差异性亦可以分为组内差异和组间差异
    - 组内差异：同处于相似环境的个体因个体原本就存在天生能力与其他个体特征差异，其行为及其后果在同组内必定具有一定差异性
    - 组间差异：同处于相似环境中的个体也会因具有相似的背景而在行为上产生出一定的相似性，从而导致不同组个体行为与结果之间存在一定的变异性
    - 因此，个体行为活动及其后果差异既有组内差异，又有组间差异

个体效应与组效应



传统OLS方法应用于多层数据结构的局限性

- 传统的线性回归方法仅能将所有个案（case）视为是来源于同一单位，只就单层数据进行回归分析，因此无法分离个体效应与组效应。如果我们采用多层线性模型化方法，便可以在研究中同时考虑
  - 个体所具有的个性特征对于个体行为及其结构所具有的影响
  - 来自于同一环境之下个体所具有的共性特征对于个体行为及其结构所具有的影响
- 在OLS回归中，所有估计系数都是固定的，而实际上，因变量与某一自变量之间的关系可能是非固定而存在变异的
  - OLS: $Y_i = \beta_0 + \beta_1 X_i + r_i$
  - HLM:
    - Level 1: $Y_{ij} = \beta_{0j} + \beta_{1j} X_{ij} + r_{ij}$ → 个体效应
    - Level 2: $\beta_{0j} = \gamma_{00} + \mu_{0j}$, $\beta_{1j} = \gamma_{10} + \mu_{1j}$ } 回归的回归（regression of regression）

      → 组效应

在多层数据结构中，OLS的基本假设能否得到满足？

- 若为多层数据结构，且分析表明组内和组间差异同时存在，那么OLS的多种基本假设便无法得到满足
  - Level 1: $Y_{ij} = \beta_{0j} + \beta_{1j} X_{ij} + r_{ij}$
  - Level 2: $\beta_{0j} = \gamma_{00} + \mu_{0j}$

    $\beta_{1j} = \gamma_{10} + \mu_{1j}$
  - Reduced form:

$$Y_{ij} = (\gamma_{00} + \mu_{0j}) + (\gamma_{10} + \mu_{1j}) X_{ij} + r_{ij}$$

$$Y_{ij} = \gamma_{00} + \gamma_{10} X_{ij} + \mu_{0j} + \mu_{1j} X_{ij} + r_{ij}$$

  - 如果 μ_{0j}, μ_{1j} 都等于0，那么回归系数在第二层不存在变异，此时便等价于OLS模型
  - 如果 μ_{0j}, μ_{1j} 不等于0，那么便有以下后果
    - 残值均值不为零，残值与自变量相关，有偏且非一致
    - 异方差，非最优，标准误计算有误
      - 残值中，r_{ij} 为个体变异，而 μ_{0j}, μ_{1j} 为第二层单位之间变异，因此，在相同第二层单位内个体之间的相似性比第二层不同单位个体之间的相似性高，导致异方差现象
    - 残值之间可能会存在相关性，非最优，标准误计算有误

5.3 HLM的参数估计方法

- 采用收缩估计（shrinkage estimation）和广义最小二乘法（GLS）估计模型系数
- 采用首先的极大似然估计法（MLR）或贝叶斯方法（Beyesian method）估计系数的方差和协方差矩阵
- 详细统计推断参见：
  - Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical Linear Models: Applications and data analysis methods* (2th ed.). Thousand Oaks: Sage./（中译本）分层线性模型：应用于数据分析方法，郭志刚等译，社会科学文献出版社，2007年。

5.4 多层线性基本模型形式（两层模型）

- 零模型（Null Model）
  - Level 1: $Y_{ij} = \beta_{0j} + r_{ij}$
  - Level 2: $\beta_{0j} = \gamma_{00} + \mu_{0j}$
    - 下标*i*表示第一层的单位，例如县
    - 下标*j*表示第一层个体所隶属的第二层单位，例如省
    - 第一层模型中下标0表示截距，β_{0j} 便表示第一层模型的截距，它表示各个第二层单位组内的均值。若第一层因变量 Y_{ij} 表示第*j*省第*i*县的地方人均财政支出，那么第一层截距 β_{0j} 便表示第*j*省各县级地方人均财政支出均值，而其残值 r_{ij} 便表示该省各县级地方人均财政支出围绕省均值的一个变异。该残值方差为：$Var(r_{ij}) = \sigma^2$
    - 由于全国各省县级地方人均财政支出均值在省际之间存在变异，因此，我们便用第二层模型来解释这种变异。在第二层模型中，其截距 γ_{00} 表示各县级地方人均财政支出的全国样本均值，而其残值 μ_{0j} 表示各省支出均值围绕全国均值的一个变异。其残值方差为：$Var(\mu_{0j}) = \tau_{00}$

跨级相关

- 跨级相关（Intra-class correlation）：

$$\rho = \tau_{00} / (\tau_{00} + \sigma^2)$$

  - 该统计量表达何种含意？
    - 财政支出总体变异中有多少是由省际差异造成的
- 由于零模型主要是利用残值方差做变异检验，因此又称为方差成分分析模型（**VCA Model，Variance Component Analysis Model**）

多层线性基本模型形式（两层模型）

- 完整模型（Full Model）
  - Level 1: $Y_{ij} = \beta_{0j} + \beta_{1j} X_{1ij} + r_{ij}$
  - Level 2: $\beta_{0j} = \gamma_{00} + \gamma_{01} W_{1j} + \mu_{0j}$

    $\beta_{1j} = \gamma_{10} + \gamma_{11} W_{1j} + \mu_{1j}$
    - 第一层模型中，下标0依然表示截距，1表示模型第一个自变量x_1的回归系数，2表示第二个自变量的回归系数，以此类推。
    - 第二层模型中，下标第一个数字若为0，表示是进入到第一层模型截距系数的回归方程中，若为1表示是进入到第一层模型斜率系数的回归方程中；下标第二个数字若为1，表示其为第二层模型某一回归方程的截距系数，若为2表示其为第二层模型某一回归方程的斜率系数

多层线性基本模型形式（两层模型）

- 利用完整模型我们可以实现以下研究内容（仍以县级地方财政支出为例）
  - 在第一层模型中，零模型仅考虑各省县级地方人均财政支出在省内的均值情况及其变异程度，而完整模型不仅反映县级地方支出在省内的变异程度，而且还尝试引入一些各县级地方特征变量，对县级地方支出的省内变异进行解释。
    - 例如，我们可以考虑各县级地方自有财政收入对各县级地方财政支出的影响
      - 此时，地方自有财政收入的第一层回归系数 β_{1j} 有何经济学含意？
        - 表示的是县级地方自有财政收入对地方财政支出的影响，反映的是一种财政非公平
  - 在第二层模型中，零模型仅考虑县级地方支出的全国均值情况，及其在省际之间的变异程度，而完整模型不仅反映县级地方财政支出在省际之间的变异程度，而且还尝试引入一些省内共性特征变量（例如各省内财政分权化程度），
    - 对第一层截距系数的省际间变异进行解释，即对各省县级地方财政支出平均水平在省际之间的变异进行解释
    - 对第一层斜率系数的省际间变异进行解释，即对各省内财政非公平程度在省际之间的差异进行解释

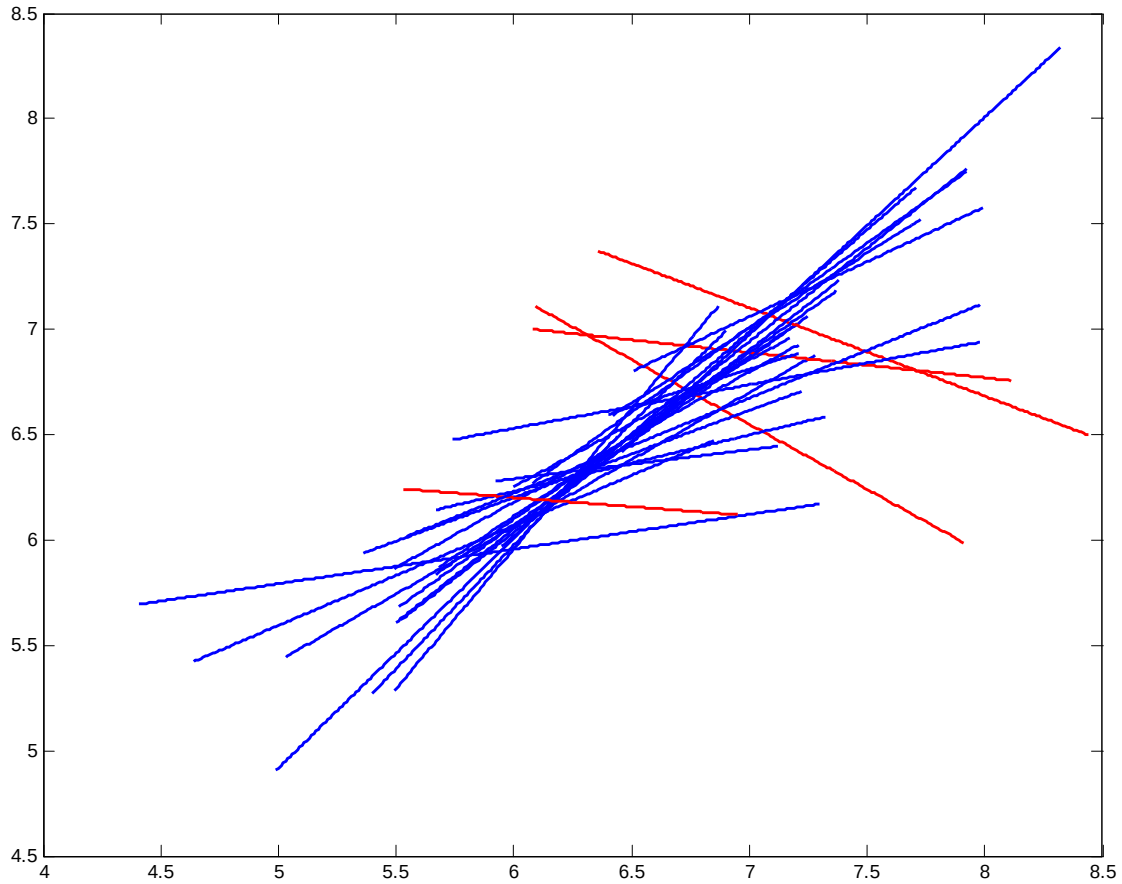
残值的方差分析

- 完整模型的残值分析
  - 第一层残值方差为 $Var(r_{ij}) = \sigma^2$ ，我们可以比较完整和零模型第一层残值方差的变化，来说明进入了一个个体特征变量后，该变量能解释省内支出变异多少
  - 第一层残值方差分别为 $Var(\mu_{1j}) = \tau_{11}$ 和 $Var(\mu_{0j}) = \tau_{00}$ ，我们同样可以观察和比较完整和零模型第二层残值方差变化，来说明第二层共性特征变量能解释省际支出差异多少？

HLM在教育财政研究的应用

- 以全国县（县级市）2001年数据为例，讨论中国地方小学教育财政支出、地方自有财政收入与省内财政分权化之间的关系
- 研究问题
  - 中国县级地方小学教育财政支出的省内与省际差异表现如何？
  - 县级地方自有财政收入与地方小学预算内财政支出之间呈现何种关系，或者说，中国县级地方小学教育财政支出在财富或财政中立（Wealth/fiscal Neutrality）方面的公平性表现如何？
  - 省内财政分权化在影响省内县级地方小学教育财政支出水平，以及对省内小学教育财政支出的财富中立程度具有何种影响？

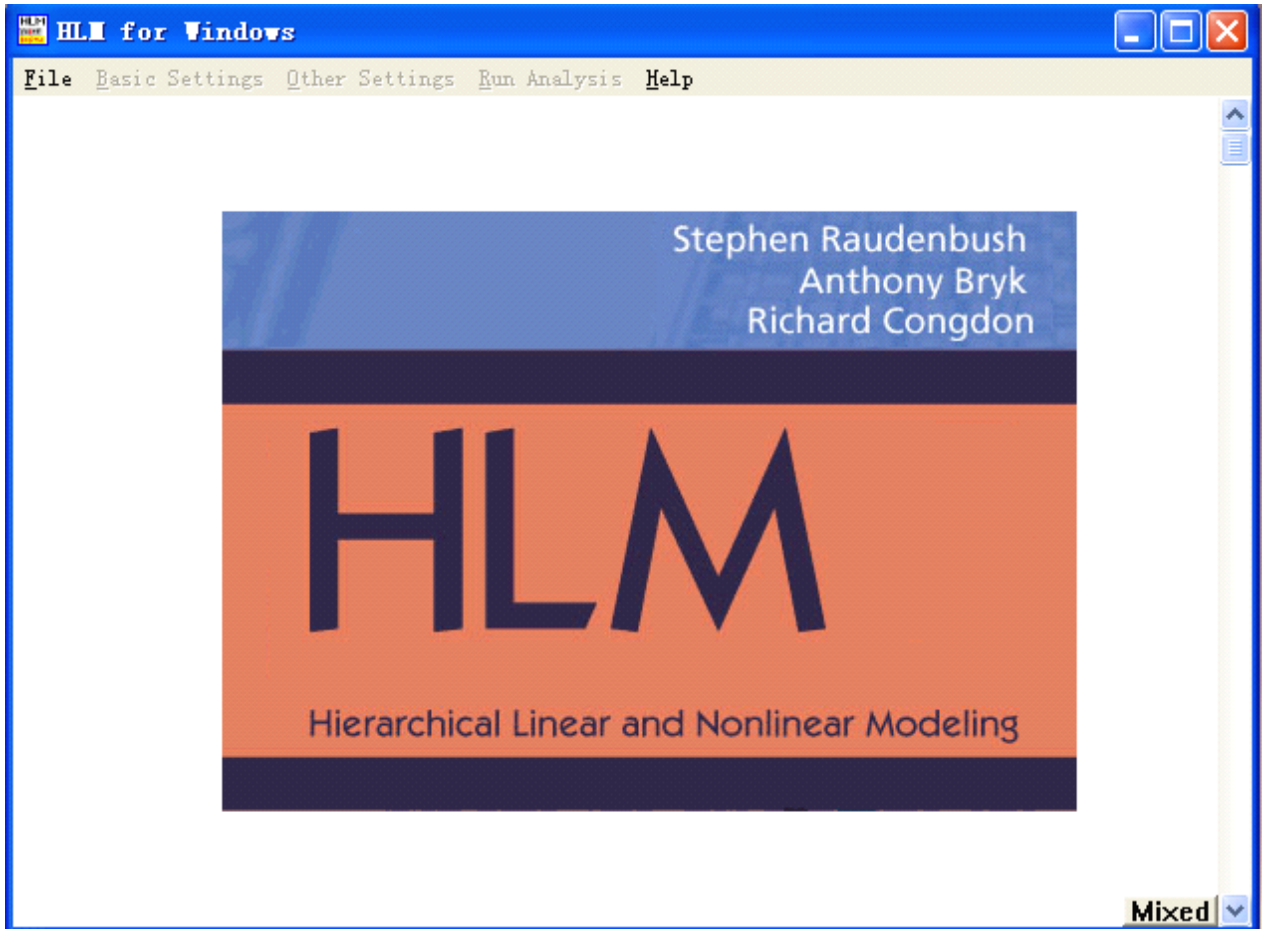
HLM在财政研究的应用



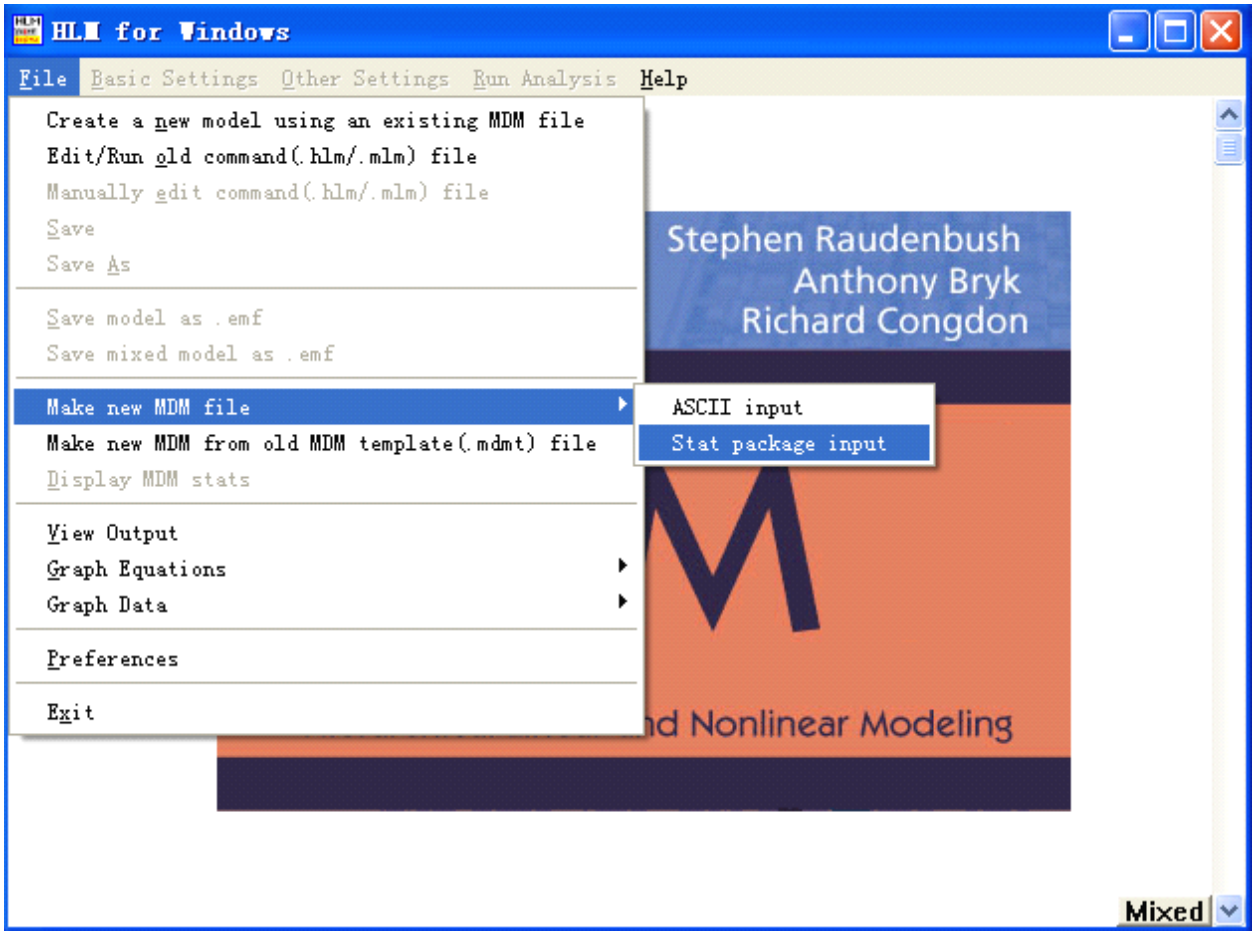
HLM在财政研究的应用

- 我们准备构建两层线性模型
  - 零模型 $$\ln EDU_{ij} = \beta_{0j} + r_{ij}$$ $$\beta_{0j} = \gamma_{00} + \mu_{0j}$$
  - 随机效应模型 $$\ln EDU_{ij} = \beta_{0j} + \beta_{1j} \ln POFR_{ij} + r_{ij}$$ $$\beta_{0j} = \gamma_{00} + \mu_{0j}$$ $$\beta_{1j} = \gamma_{10} + \mu_{1j}$$
  - 完整模型 $$\ln EDU_{ij} = \beta_{0j} + \beta_{1j} \ln POFR_{ij} + r_{ij}$$ $$\beta_{0j} = \gamma_{00} + \gamma_{01} FD_j + \mu_{0j}$$ $$\beta_{1j} = \gamma_{10} + \gamma_{11} FD_j + \mu_{1j}$$
- 设计好模型后，先通过SPSS进行数据输入与管理，并分别制作好第一层和第二层数据

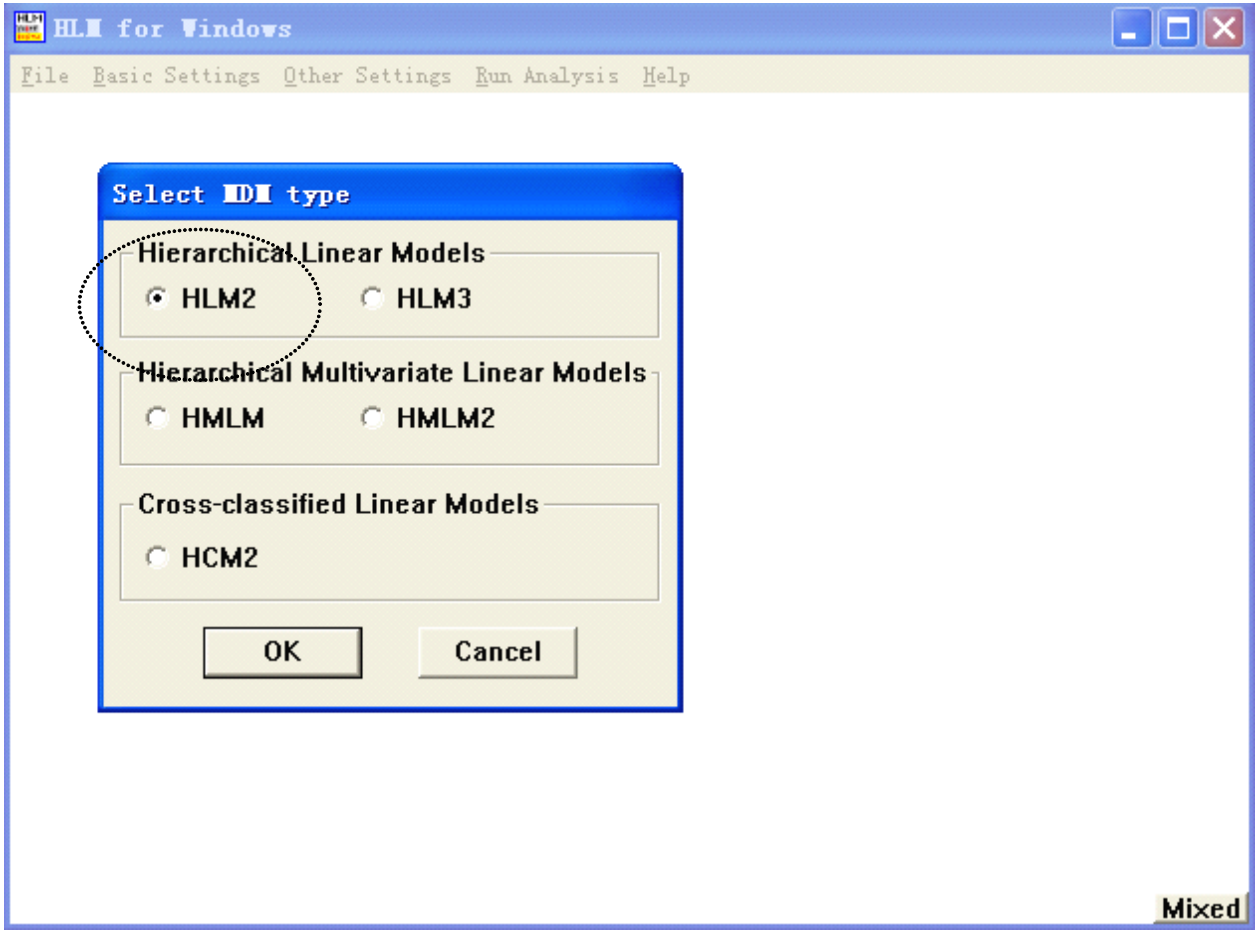
HLM在财政研究的应用



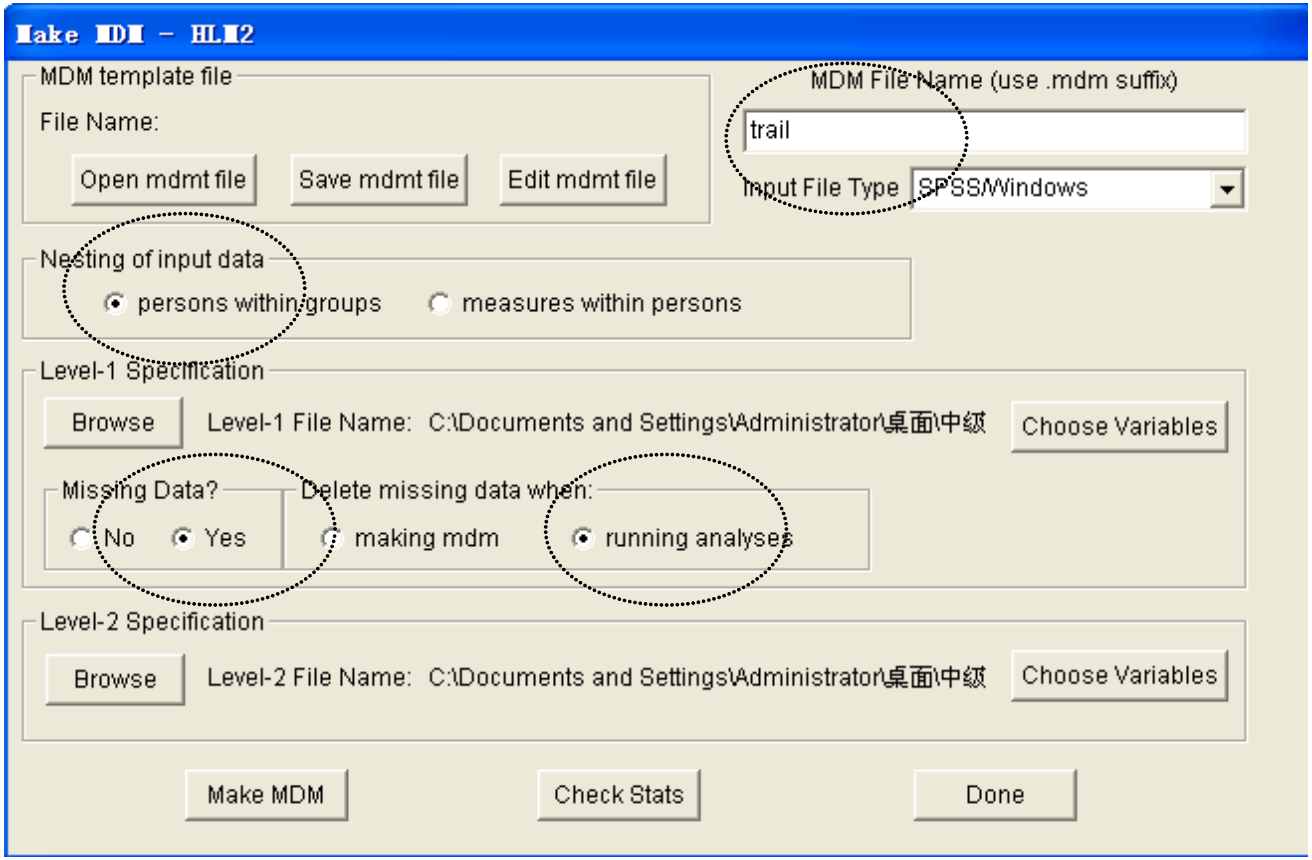
HLM在财政研究的应用



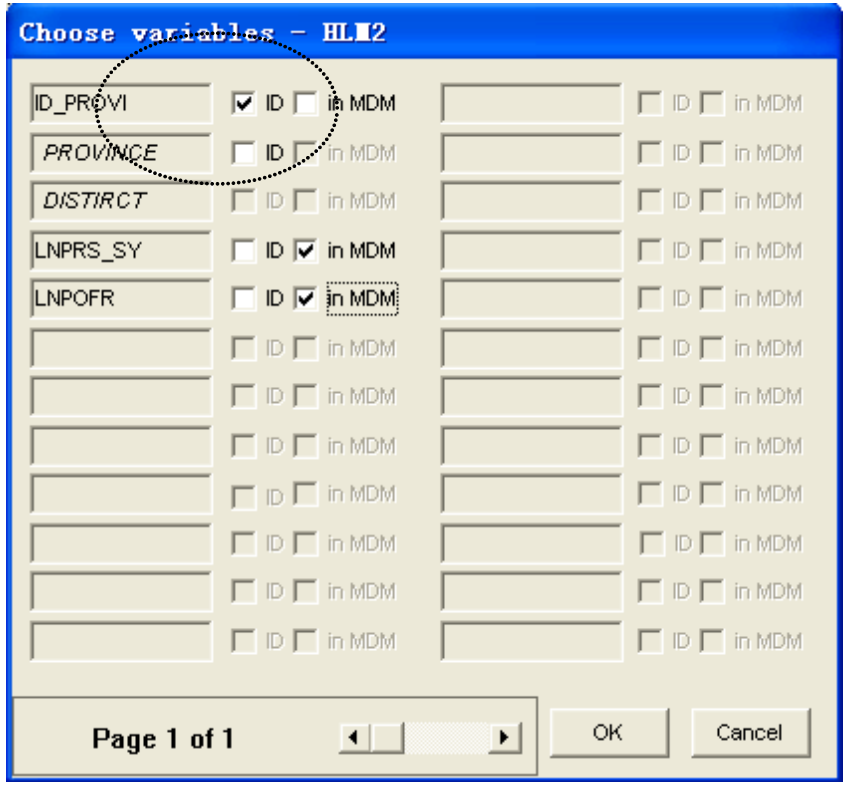
HLM在财政研究的应用

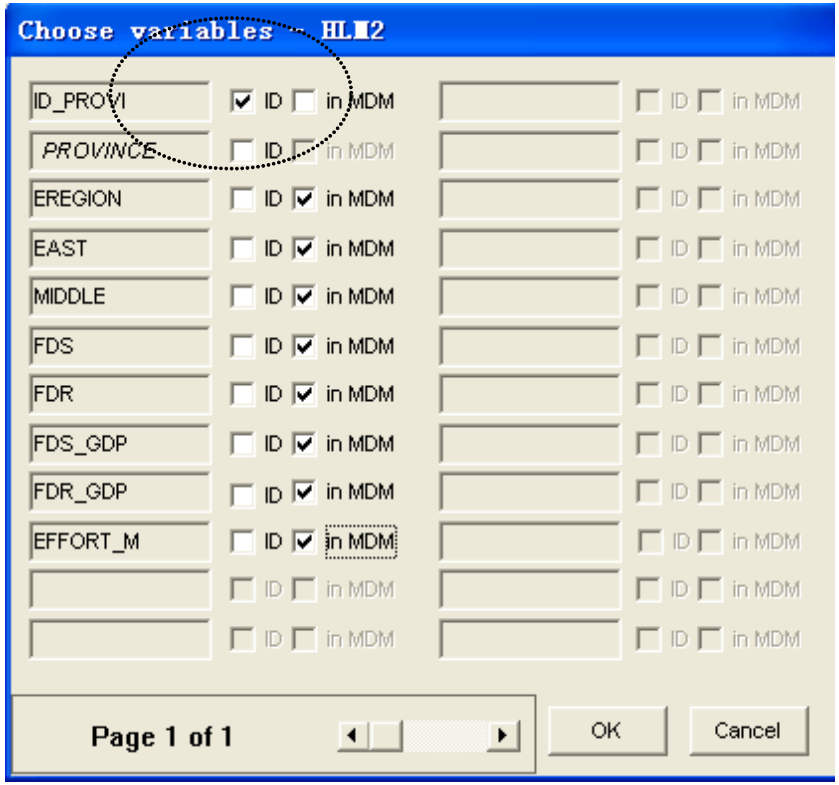


HLM在财政研究的应用

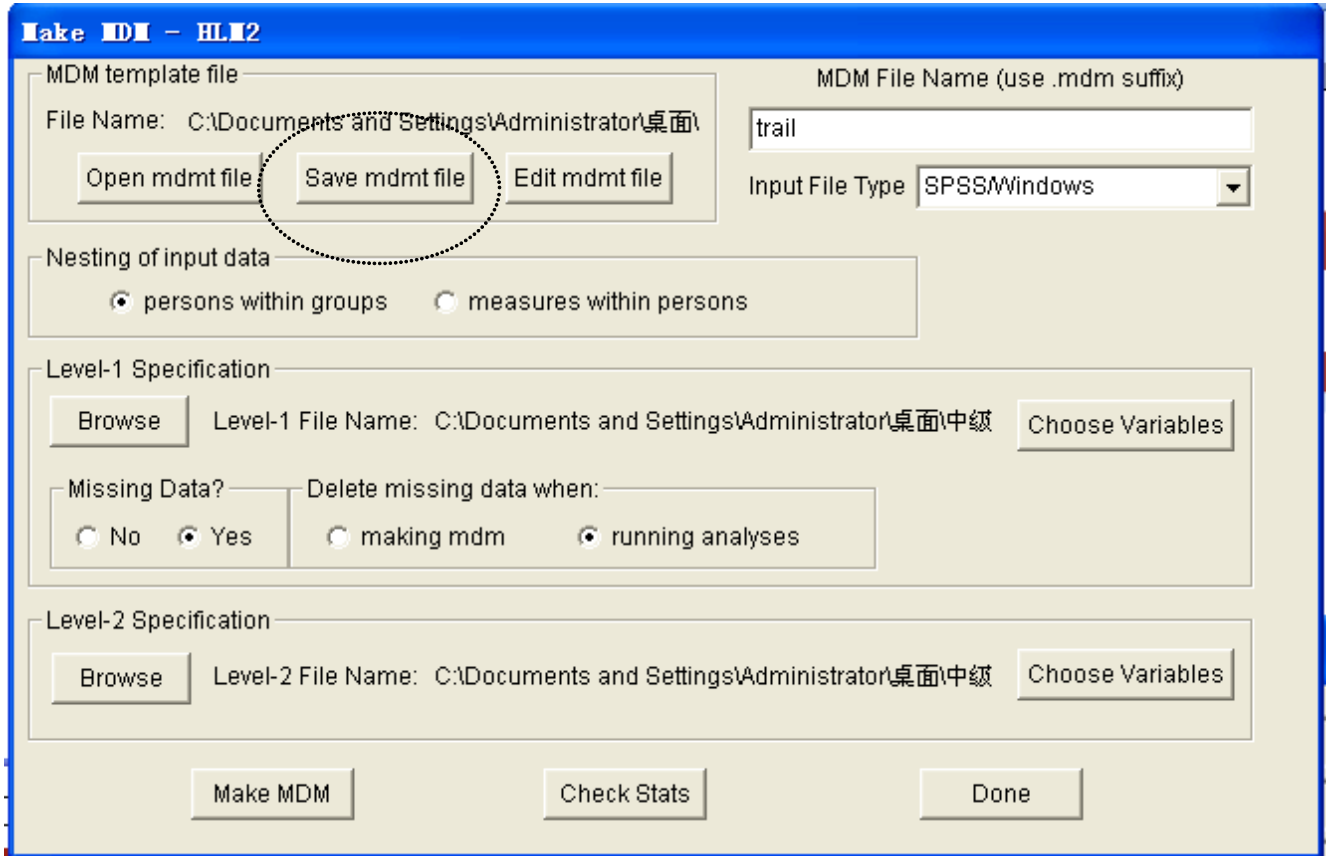


HLM在财政研究的应用

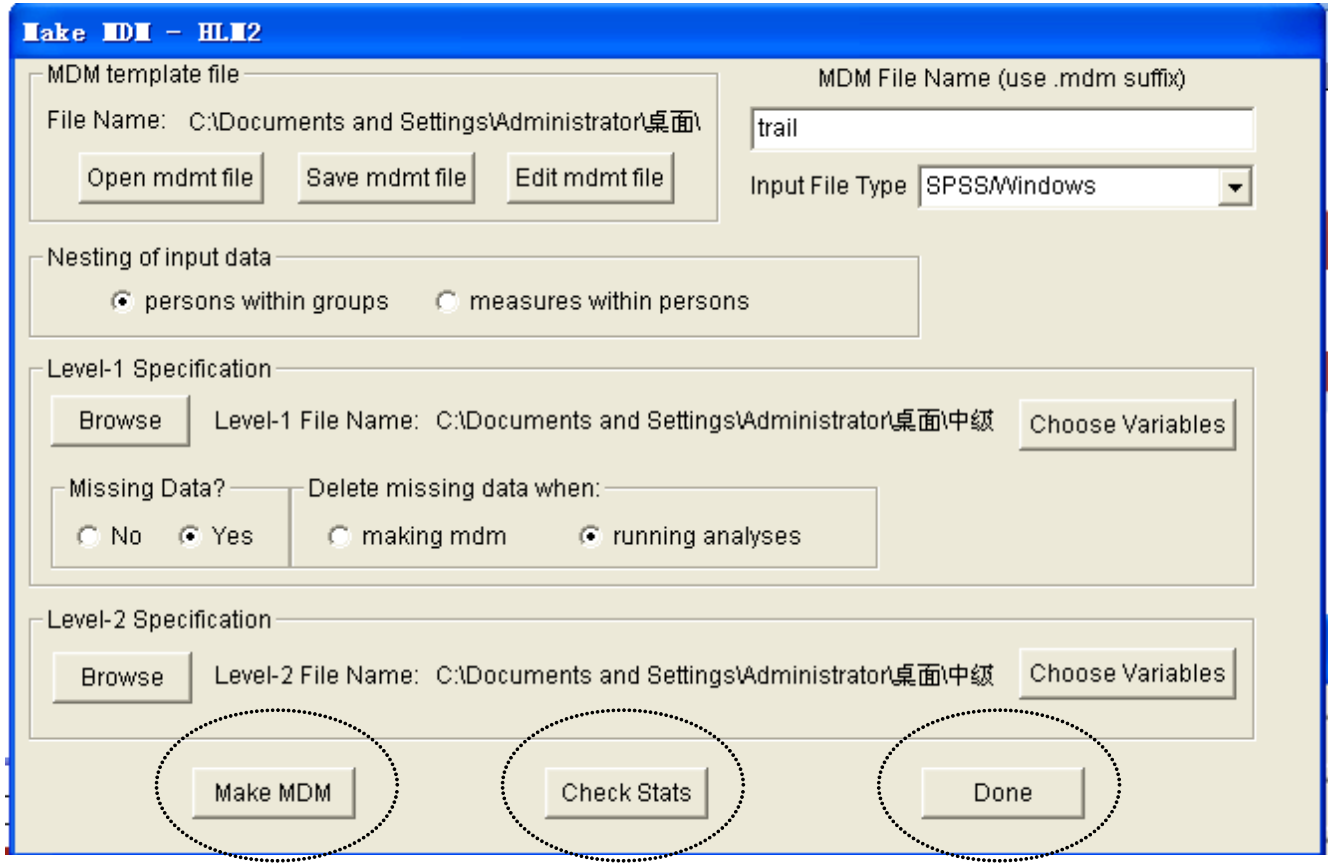




HLM在财政研究的应用

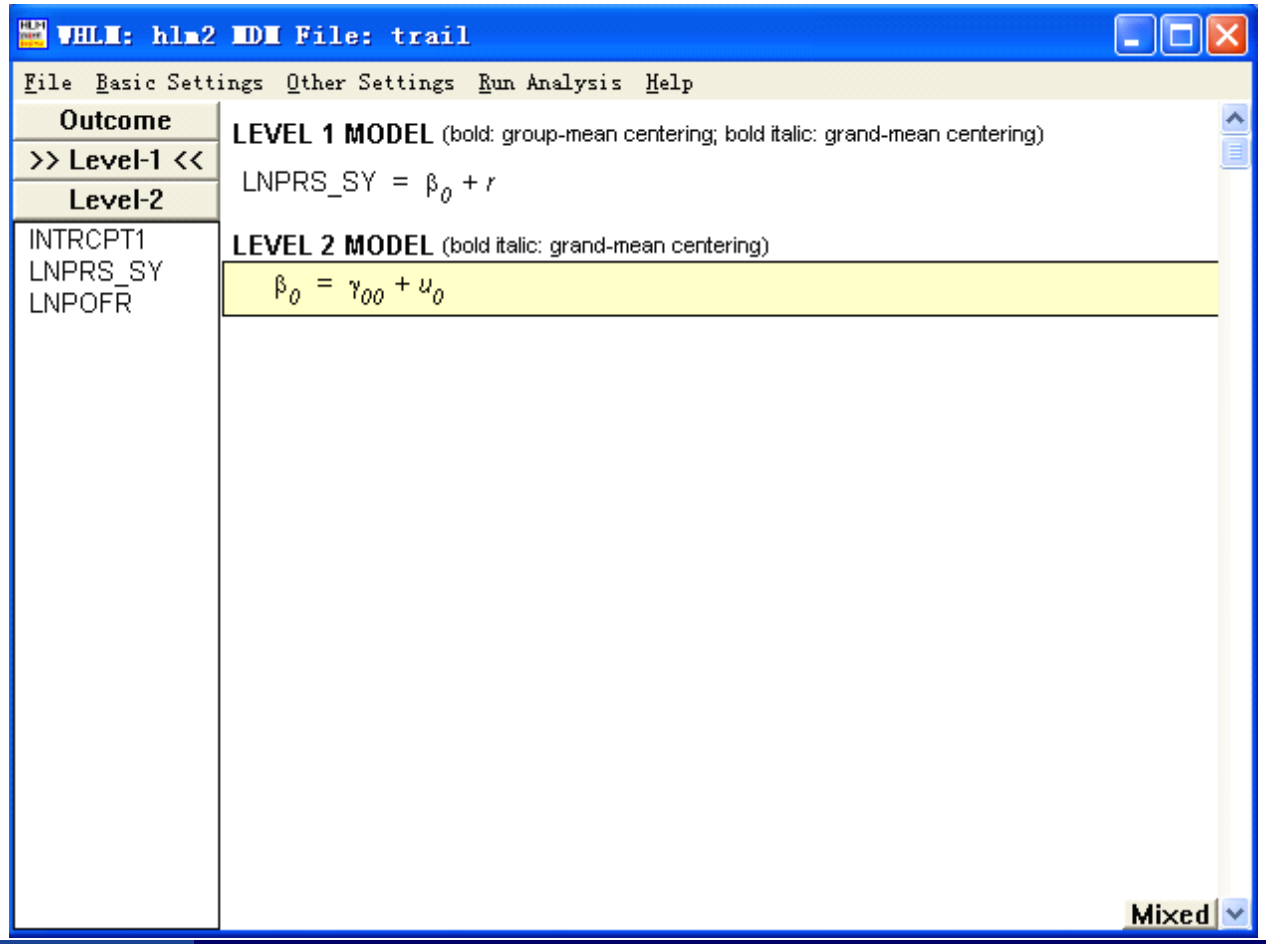


HLM在财政研究的应用



HLM在财政研究的应用

■ 零模型的构建



HLM在财政研究的应用

- 零模型
  - 那个方差成分代表县级地方小学教育财政支出的省内差异，那个代表省际差异？
  - 县级地方小学教育财政支出在省际间是否存在显著的差异？
    - 存在，看*Chi2*检验结果，我们需加入第二层变量对县级地方小学支出在省际间的变异进行解释
  - 县级地方小学教育财政支出的省内差异占总差异多少？
    - 省内＝0.15/(0.18+0.15)

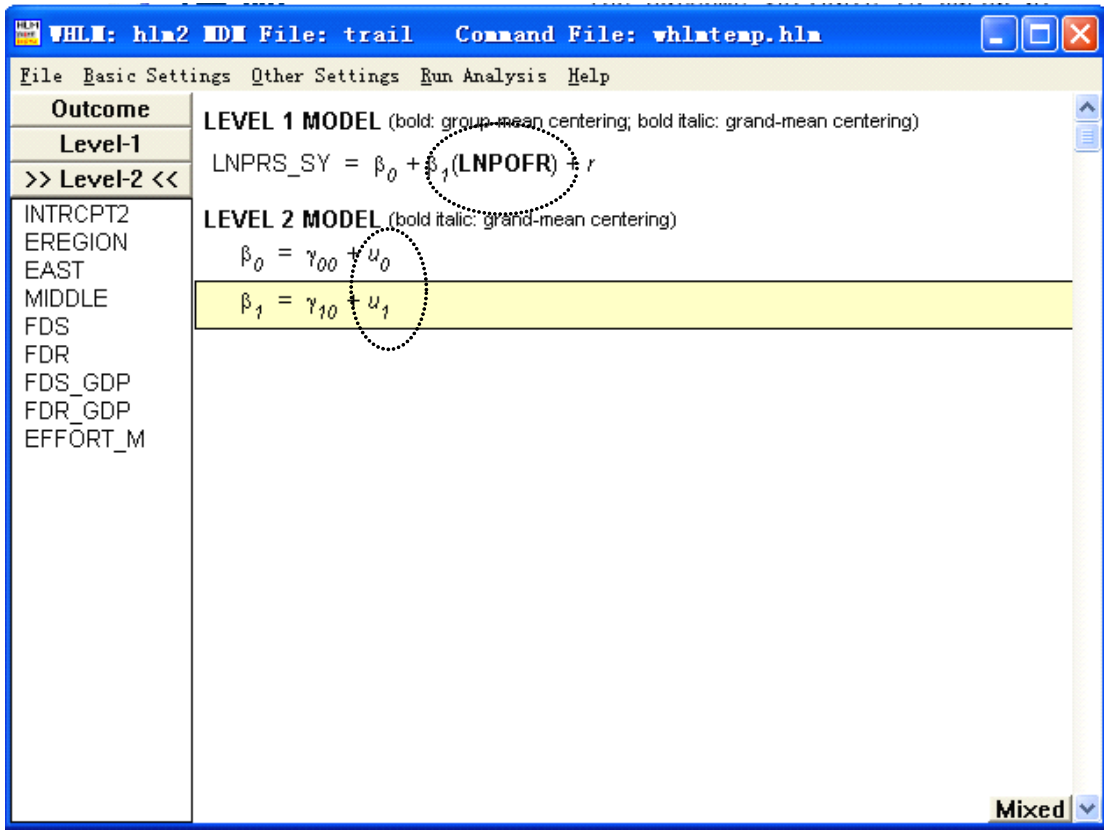
```
Final estimation of fixed effects:
----------------------------------------------------------------------------
                                          Standard            Approx.
   Fixed Effect          Coefficient      Error      T-ratio  d.f.     P-value
----------------------------------------------------------------------------
For       INTRCPT1, B0
   INTRCPT2, G00            6.580439      0.077591    84.810       30    0.000
----------------------------------------------------------------------------

The outcome variable is LNPRS_SY

Final estimation of fixed effects
(with robust standard errors)
----------------------------------------------------------------------------
                                          Standard            Approx.
   Fixed Effect          Coefficient      Error      T-ratio  d.f.     P-value
----------------------------------------------------------------------------
For       INTRCPT1, B0
   INTRCPT2, G00            6.580439      0.076326    86.215       30    0.000
----------------------------------------------------------------------------


Final estimation of variance components:
----------------------------------------------------------------------------
Random Effect            Standard       Variance     df    Chi-square  P-value
                         Deviation      Component
----------------------------------------------------------------------------
INTRCPT1,       U0         0.42535        0.18093    30    1979.54255    0.000
 level-1,       R          0.38931        0.15156
----------------------------------------------------------------------------


Statistics for current covariance components model
--------------------------------------------------
Deviance                        = 2096.980103
Number of estimated parameters = 2
```

组内差异

组间差异

HLM在财政研究的应用

- 随机效应模型的构建
  - 需要注意第一层自变量的group-mean centering的处理方法及其含义
    - 在本例中，即将LNPOFR转化为LNPOFR减去其组（省）的mean



HLM在财政研究的应用

- 随机效应模型（Random effect model）
  - 加入地方自有财政收入一层变量后，能解释多少省内差异？
    - （0.15-0.12）/0.15＝20%
  - 县级地方自有财政收入与小学教育财政支出之间的关系是否存在显著的省际差异（财富非中立）？
    - 存在，*lnPOFR*的回归系数为0.25，显著
  - 方差分析显示，各省的生均支出水平与财富非中立水平在省际间还存在显著的变异，需引入第二层省级变量予以解释，因此我们尝试加入各省财政分权化这一变量

The outcome variable is LNPRS_SY

Final estimation of fixed effects
(with robust standard errors)

Fixed Effect	Coefficient	Standard Error	T-ratio	Approx. d.f.	P-value
For INTRCPT1, B0					
INTRCPT2, G00	6.582383	0.076765	85.747	30	0.000
For LNPOFR slope, B1					
INTRCPT2, G10	0.248135	0.033914	7.317	30	0.000

Final estimation of variance components:

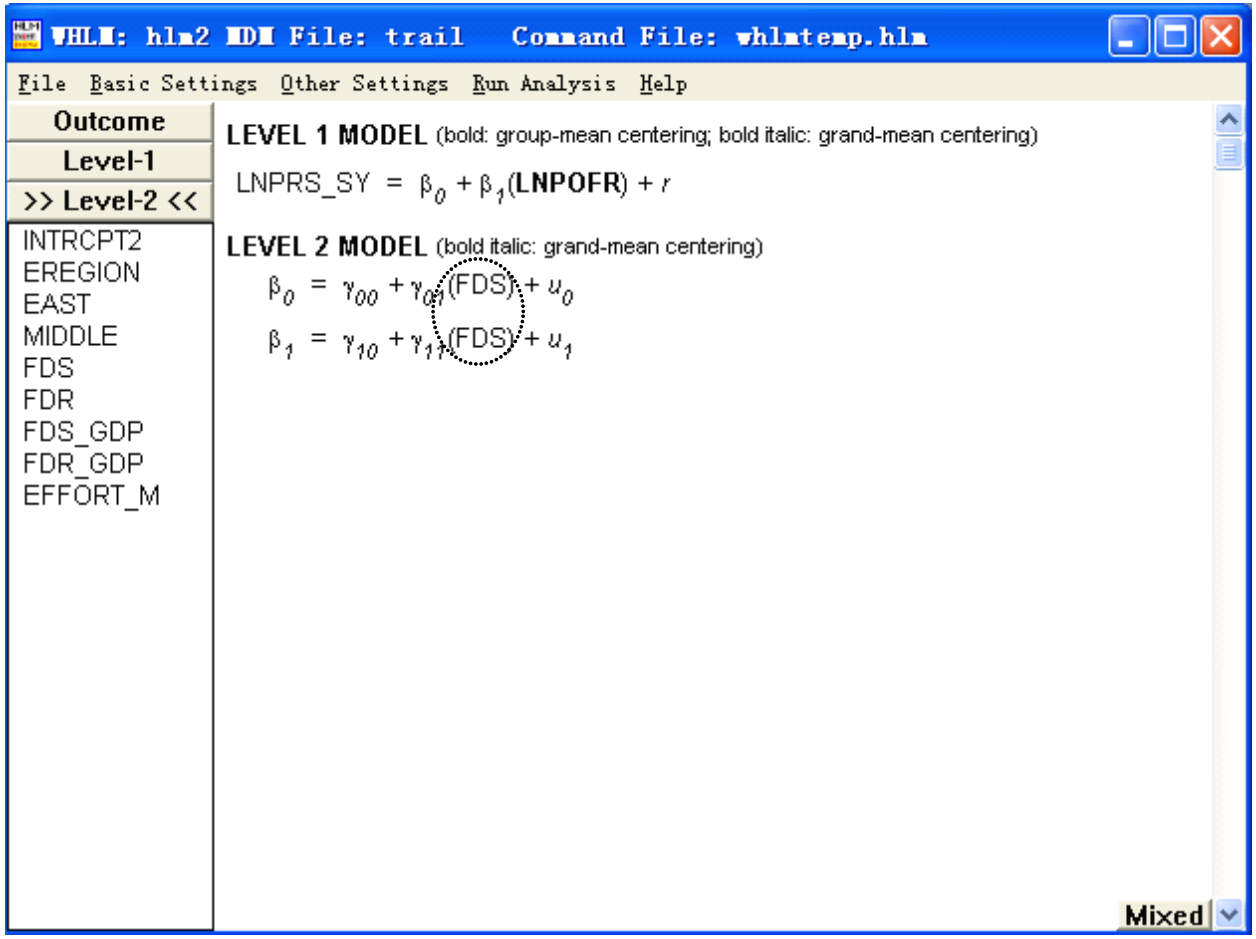
Random Effect		Standard Deviation	Variance Component	df	Chi-square	P-value
INTRCPT1,	U0	0.42896	0.18401	30	2412.05053	0.000
LNPOFR slope,	U1	0.16200	0.02624	30	159.43304	0.000
level-1,	R	0.35311	0.12469			

Statistics for current covariance components model

Deviance = 1752.050476
Number of estimated parameters = 4

HLM在财政研究的应用

- 完全模型的构建



HLM在财政研究的应用

- 完全模型
  - 省内财政分权化对各省县级地方教育财政支出水平具何种影响，为什么？
    - 为显著负影响
    - 财政越分权意味着地方政府拥有更高的支出量裁权。如果地方不偏好教育财政支出，分权度越大，教育支出水平便会越低
  - 省内财政分权化能解释县级地方小学教育财政支出省内差异多少？
    - （0.184-0.158）/0.184=14.13%
  - 省内财政分权化对于省内小学教育财政财富非中立是起到改善作用，还是恶化作用？
    - 为恶化作用，分权化会使得地方教育支出更加依赖地方自有的财政收入水平（虽非显著，但在第一层考虑更多解释变量后，变为显著，略去）
  - **财政分权化在对地方小学教育财政支出的影响实际上是一种多层的结构性影响**

The outcome variable is LNPRS_SY

Final estimation of fixed effects
(with robust standard errors)

Fixed Effect	Coefficient	Standard Error	T-ratio	Approx. d.f.	P-value
For INTRCPT1, B0					
INTRCPT2, G00	7.597111	0.409032	18.573	29	0.000
FDS, G01	-1.534104	0.599361	-2.560	29	0.016
For LNPOFR slope, B1					
INTRCPT2, G10	0.004087	0.155320	0.026	29	0.979
FDS, G11	0.364309	0.244134	1.492	29	0.146

Final estimation of variance components:

Random Effect		Standard Deviation	Variance Component	df	Chi-square	P-value
INTRCPT1,	U0	0.39749	0.15800	29	2185.51506	0.000
LNPOFR slope,	U1	0.15854	0.02514	29	151.51651	0.000
level-1,	R	0.35312	0.12469			

Statistics for current covariance components model

Deviance = 1745.201440
Number of estimated parameters = 4

第五讲参考文献

- Raudenbush, S. W., & Bryk, A. S. (2002). Hierarchical Linear Models: Applications and data analysis methods (2th ed.). Thousand Oaks: Sage（此书有中译本，题为《分层线性模型:应用与数据分析方法》）.
- 张雷、雷雳、郭伯良，多层线性模型应用，北京：教育科学出版社，2003.