当代经济学系列丛书
Contemporary Economics Series
主编 陈昕

基本无害的计量经济学
实证研究者指南

当代经济学
教学参考书系

[美] 乔舒亚·安格里斯特
约恩-斯特芬·皮施克 著
郎金焕 李井奎 译

格致出版社
上海三联书店
上海人民出版社

基本无害的计量经济学

实证研究者指南

有趣而且不同寻常，这是本有着自身特点的计量经济学教科书。它为那些需要经常分析经济数据的人士提供了实实在在的答案和建议，以帮助他们解决自己面临的问题。

——Guido Imbens, Harvard University

这是本关于计量经济学实践的书籍，它内容出色、构思精妙。

——Orley Ashenfelter, Princeton University

无论是政治科学研究者，还是社会学家、历史学家、地理学家抑或是人类学家等等，对任何希望在社会科学领域构思研究并进行验证的科研人员而言，都应该将这本开创性的书籍列为必读书目。本书构思精巧、有趣，能够引导你穿越社会科学实证研究的迷雾。我真希望多年前就能读到这本书。

—— James Robinson, Harvard University

这是一本适合所有的实证研究者而非仅适合学生使用的书籍，任何一个实证计量工作者都能将它看作一份极佳的参考文献。

——Sandra Black, UCLA

ISBN 978-7-5432-2058-4
9 787543 220584
定价: 38.00元
易文网: www.ewen.cc
格致网: www.hibooks.cn

当代经济学系列丛书
Contemporary Economics Series
主编 陈昕

基本无害的计量经济学
实证研究者指南

[美] 乔舒亚·安格里斯特
约恩-斯特芬·皮施克 著
郎金焕 李井奎 译

当代经济学教学参考书系

格致出版社
上海三联书店
上海人民出版社

图书在版编目(CIP)数据

基本无害的计量经济学:实证研究者指南/(美)安格里斯特,(美)皮施克著;郎金焕,李井奎译. —上海:格致出版社:上海人民出版社,2012
(当代经济学系列丛书/陈昕主编. 当代经济学教学参考书系)
ISBN 978-7-5432-2058-4

Ⅰ. ①基… Ⅱ. ①安… ②皮… ③郎… ④李… Ⅲ. ①计量经济学-高等学校-教材 Ⅳ. ①F224.0

中国版本图书馆 CIP 数据核字(2012)第 013012 号

责任编辑 钱 敏
装帧设计 敬人设计工作室
吕敬人

[美]乔舒亚·安格里斯特
约恩-斯特芬·皮施克 著
基本无害的计量经济学:实证研究者指南 郎金焕 李井奎 译

格致出版社·上海三联书店·上海人民出版社
(200001 上海福建中路 193 号 24 层 www.ewen.cc)

编辑部热线 021-63914988
市场部热线 021-63914081
www.hibooks.cn

世纪出版集团发行中心发行
苏州望电印刷有限公司印刷
2012 年 4 月第 1 版
2012 年 4 月第 1 次印刷
开本:787×1092 1/16
印张:17 插页:5 字数:391,000

ISBN 978-7-5432-2058-4/F·510 定价:38.00 元

出版前言

为了全面地、系统地反映当代经济学的全貌及其进程，总结与挖掘当代经济学已有的和潜在的成果，展示当代经济学新的发展方向，我们决定出版“当代经济学系列丛书”。

“当代经济学系列丛书”是大型的、高层次的、综合性的经济学术理论丛书。它包括三个子系列：(1)当代经济学文库；(2)当代经济学译库；(3)当代经济学教学参考书系。该丛书在学科领域方面，不仅着眼于各传统经济学科的新成果，更注重经济学前沿学科、边缘学科和综合学科的新成就；在选题的采择上，广泛联系海内外学者，努力开掘学术功力深厚、思想新颖独到、作品水平拔尖的“高、新、尖”著作。“文库”力求达到中国经济学界当前的最高水平；“译库”翻译当代经济学的名人名著；“教学参考书系”则主要出版国外著名高等院校的通用教材。

本丛书致力于推动中国经济学的现代化和国际标准化，力图在一个不太长的时期内，从研究范围、研究内容、研究方法、分析技术等方面逐步完成中国经济学从传统向现代的转轨。我们渴望经济学家们支持我们的追求，向这套丛书提供高质量的标准经济学著作，进而为提高中国经济学的水平，使之立足于世界经济学之林而共同努力。

我们和经济学家一起瞻望着中国经济学的未来。

前　言

计量经济学的疆界正在不断扩张。作为这种扩张的结果，其方法和实践也有了长足发展，但即使是那些精于数据处理的个中老手，也会对如今如此繁多的计量方法感到困惑。幸运的是，并非所有方法都同样有用、同等重要。那些过于新奇的方法本来没必要如此复杂，而且还可能是有害的。从积极的方面讲，虽然对计量经济学基本工具的解释日趋精奥深微，但应用计量经济学（Applied Econometrics）的核心内容却保持着大体稳定。本书为实证研究者把握计量经济学的精义提供了一个向导，这些计量经济学的精义也就是我们所指的基本无害的计量经济学（Mostly Harmless Econometrics）。

在应用计量经济学家的工具箱中，最重要的几件工具可以列举如下：

（1）为了将可能掩盖因果关系的变量控制起来，而设计的回归模型（Regression Model）；

（2）用于分析真实实验以及自然实验的工具变量方法（Instrumental Variables Method）；

（3）在重复观察中用以处理不可观察的缺失变量的双重差分方法（Difference-in-Difference Strategies）。

对上面这些基本技巧的创造性使用要求读者对统计推断的作用机理有坚实的概念基础和良好的理解。应用计量经济学在这两方面的特点将会在本书中得到体现。

我们对计量经济学中那些内容重要的看法来自我们作为实证研究者的研究经验，而且特别来自我们的教学实践和指导经济学博士研究生的工作。正是

在与这些同学的思维交流中，我们完成了本书的写作。与此同时，我们还希望这本书能够吸引其他领域中正在苦苦探索如何选择计量方法、如何解释研究结果的研究者们。应用计量经济学所考虑的问题和其他社会科学或者流行病学所考虑的问题并无本质上的区别。任何希望运用数据指导公共政策或者推动公共卫生事业的人都要理解并使用统计结果。任何希望从数据中得到有用推断的人都可称为应用计量经济学家。

许多计量经济学方面的教科书都对研究方法提供一些指导，因此本书和其他广泛使用的教科书存在一些内容上的重叠。但这本书在多个方面有别于传统的计量经济学教科书。首先，我们认为使用数据回答特定因果关系的经验研究最有价值，这类似于在医学研究中经常出现的随机临床实验。我们研究所有问题的方法都体现这个观点。在缺乏真实实验时，我们寻找经过良好控制的对照组，或者说自然的“准实验(quasi-experiment)①”。当然，一些准实验研究设计要比其他一些方法更有说服力，但是在这些例子中计量经济学使用的方法几乎都很简单。因此，相比于其他教科书中对计量方法的处理，这本书对相应主题的讨论显得更短小更集中。我们主要对在自己的研究中读到和使用到的概念和简单的统计技巧进行强调，并与多个实证研究案例结合来解释这些观点和技巧。尽管我们对计量经济学中什么是重要的观点并未在应用经济学家中得到一致认同，但无可争议的事实是实验和准实验研究方法逐渐居于应用经济学中最具影响力的那些研究的核心。

我们要指出的第二个不同是本书在一定程度上忽略严格性。大多数计量经济学教科书都对计量模型进行严格处理。特别的，这些书对诸如线性性和同方差性等大家认为经典模型中普遍会被违背的假设进行大量讨论。虽然在行文中也会提及这些问题，但我们采取一种更加宽容和不那么迂腐的态度。能够支持我们以上述态度进行讨论的原因乃是：我们可以对计量经济学中得到广泛使用的估计值作出一个简单的解释，这一解释与模型本身并无太大关系。如果我们得到的估计值不是我们想要的那个，那么一定是做这项研究的计量经济学家错了，而不是计量经济学错了。一个典型的例子就是线性回归，它为我们提供了关于条件期望函数②的有用信息，而不论条件期望函数究竟是什么形状。同样的，工具变量方法可以估计出经过良好定义的总体的平均因果效应，即使这个工具变量无法影响所有个体。许多应用研究者往往从直觉上理解基本计量工具在概念上的严格性，因此隐藏在严格性背后的大部分理论将不会在本书中出现。本书在处理推断问题上也有所不同，我们并不过多地考虑渐进有效性，而是用大多数篇幅考虑实际中不易处理的有限样本问题。

---

① 准实验是指对控制组或者所研究因素几乎无法施加控制的实验。与一般的随机实验最大的区别在于准实验无法将个体随机分配，因此观察得到的结果可能和个体的某些不可观察因素有关，也就是说，我们不能保证得到结果是将观测个体混匀后的平均值。——译者注

② 这里的条件均值函数就是一般意义上所指的总体回归函数。在线性回归中，总体回归函数被假设为关于参数线性的。

本书的预修要求是掌握概率论和统计学的基本知识。我们特别希望读者熟悉统计推断的基本概念，比如 t-统计量和标准误(standard error)。对数学期望等概率论知识的熟悉也会有所帮助，但是之外的数学知识并不要求。虽然书中对部分重要结论进行证明，但是技术性的细节并不繁难。与很多计量经济学高级教材不同，本书仅仅少量地使用线性代数。因此，我们提供的这本指南应该比与之竞争的其他书籍更易阅读。最后，我们从 Douglas Adams 的系列轻松小说中持续获得灵感，在这种心境的引导下，我们的指南可能会偶尔地缺乏一点精确性，但是要比流行于市面上的多个百科全书式的大部头(Encyclopedia Galactica Econometrica)计量经济学教科书便宜。这里还要感谢普林斯顿大学出版社同意出版我们的这本指南。

致　谢

在这本书写作过程中，我们从很多朋友和同事的意见中受益良多。感谢 Alberto Abadie、Patrick Arni、David Autor、Amitabh Chandra、Monica Chen、Victor Chernozhukov、John DiNardo、Peter Dolton、Joe Doyle、Jerry Hausman、Andrea Ichino、Guido Imbens、Adriana Kugler、Rafael Lalive、Alan Manning、Whitney Newey、Derek Neal、Barbara Petrongolo、James Robinson、Gary Solon、Tavneet Suri、Jeff Wooldridge 以及 Jean-Philippe Wullrich，他们在本书构思和写作的不同阶段都给予了反馈，当然，作者文责自负。同样的感谢还要送给我们在伦敦经济学院和麻省理工学院的学生们，他们使用了本书的最初版本并帮助我们认清楚哪些部分是重要的。我们要特别感谢技巧高超的助教 Bruno Ferman、Brigham Frandsen、Cynthia Kinnan 以及 Chris Smith。我们感激绘图师 Karen Norberg 的无私奉献，他绘制了每章开始的那些图片并且在大大小小很多事情上给予反馈。我们还要感谢普林斯顿大学出版社的编辑 Tim Sullivan 和 Seth Ditchik、本书编辑 Marjorie Pannell 以及制作编辑 Leslie Grundfest 的热情帮助。最后，感谢我们的妻子给予的爱和支持。她们比任何人都了解做一个实证研究者伴侣的滋味。

本书结构

我们从两个作为引言的章节开始。第1章描述了对之后章节可能很有用的研究设计步骤。第2章讨论了在医学研究中用到的随机实验，这个实验为我们最感兴趣的问题提供了一个理想的基准。在引言章节之后，本书第二部分共有三章，分别讨论了回归、工具变量和双重差分法的核心内容。这三章内容既强调估计值的一般性质（比如回归总是可以近似条件期望函数等），也强调了对估计值赋予因果解释所需的假设（比如条件独立假设、工具变量“就像”随机分配、相似世界等）。在本书第三部分我们转入扩展。其中第6章考察对非连续实验的回归分析，我们既可将该部分内容看作回归—控制这种研究策略的变体，也可将其看作是一类工具变量估计法。在第7章我们讨论了用分位数回归来估计我们关心的变量对被解释变量分布的影响。最后一章则针对的是统计推断问题，我们在之前章节中对渐进性质进行考察时省略了这一部分内容。本书的一些章节里包含了更具技巧性或者专门性的小节，可以在不影响掌握本书主旨的前提下省略，这部分小节都用星号（*）标出。

目　录

第一部分　导　论

第二部分　核　心

第三部分 拓 展

第一部分　导　　论

▶1

关于“问题”的问题

电脑说：“我非常仔细地检查过了，它确实是我们需要的那个答案。但是，恕我直言，我认为问题的问题在于您还没有真正理解问题所在。”

Douglas Adams, *The Hitchhiker's Guide to the Galaxy*

本章简短讨论一个成功的研究项目所必备的基础。就像《圣经故事》里的《出埃及记》一样，一项研究计划可以围绕着四个问题展开。我们称它们为常见问题（frequently asked questions，简称 FAQs），因为它们的确需要被研究者不断质问才行。这些问题分别是——研究对象间的关系（relationship of interest）、理想条件下的实验（ideal experiment）、识别策略（identification strategy）以及推断模式（mode of inference）。

首先我们应该问的是：*我们关注的研究对象间的因果关系是什么？* 尽管纯粹描述性研究也起着重要作用，但是我们相信，社会科学中最有趣的研究是关于因与果的，例如第 2 章和第 6 章所讨论的班级规模对学生分数影响的研究。因果关系在预测情境变化和政策更迭的后果上也颇为有用，它告诉我们在各种可能发生但与现存状况不同的［或者称之为“反事实（counterfactual）”］的情况下将会发生什么。举例子来说，在一项调查人的生产能力——即劳动经济学家所谓的人力资本——的研究计划中，我们考察了教育水平对工资的因果效应［Card（1999）综述了这个领域里的研究］。在这里，教育水平对工资的因果效应是指个体接受更多教育所带来的工资增加量。一些研究表明，平均而言大学学历的因果效应是工资水平高出 40%，这是个相当大的回报。教育对工资的因果效应对预测上大学的成本发生变化或者加强义务入学法（compulsory attendance laws）所导致的收入变化十分有用。因为这一关系还可以从经济学模型中推导出来，所以它也是理论研究的兴趣所在。

作为劳动经济学家，我们非常喜欢将工人作为研究对象来考察因果关系，但是研究因果关系不必一定以个体劳动者为研究对象，企业和国家也可以拿来考虑。关于后者的一个例子就是 Acemoglu、Johnson 和 Robinson（2001）完成的殖民地制度对经济增长影响的研究。这一研究关心的是从殖民统治者那里继承了更多民主制度的国家后来是否享受到较高的经济增长。对此问题的回答对我们理解历史

以及思考当代发展政策的效果大有深意。比如说，今天我们可能对在伊拉克和阿富汗新创建的民主制度是否对其经济发展意义重大有所疑惑。而且，如今民主对经济增长的因果效应远非一目了然，某些东亚经济体就在没有享受完全政治自由化的同时获得了强劲增长，而众多的拉美国家虽然实现了民主化，却没有因此获得较高的经济增长。

常见问题中的第二个是考虑用于捕捉研究对象间因果效应的理想条件下的实验。比如在研究教育水平和工资间关系的例子中，我们可以想象这样一个政策——给那些有潜在辍学倾向的学生一笔奖励以鼓励他们完成学业，然后来研究相应的结果。实际上 Angrist 和 Lavy(2008)就完成了这样的一个实验。尽管这项研究着眼于诸如大学入学率等短期效果，但长期效果可以很好地指向工资①。在政治制度的例子里，我们可以来一个时空穿梭，回到过去，随机地将不同的政府结构赋予给那些独立之前的殖民地(这个实验很可能只会被拍成电影，而不是得到国家自然科学基金的支持)。

理想条件下的实验通常是假设出来的。即便如此，假想的实验仍然值得我们深思，因为它可以帮助我们挑出那些富有前景的研究主题。作为一名研究者，你可以想象没有预算约束、没有人权委员会因社会正当性而对你的研究进行规制的情况：类似于有丰厚基金支持的 Stanley Milgram，这位心理学家在 20 世纪 60 年代使用具有高度争议的实验设计，打破了关于权威服从性(the response to authority)的常规研究，在今天他的这一冒险行为却很可能会让他失去工作。

为了寻求对权威服从性的理解，Milgram(1963)表明他可以说服实验的被试者去对那些无辜的不断抗议的受害者执行痛苦的电击(这些电击是假的，受害人也是演员扮演的)。这种做法既聪明又极富争议性：有些心理学家声称对他人执行电击的被试者在心理上会受到实验的伤害。尽管如此，Milgram 的研究阐明了这样一点，即便有一些实验最好留在筹划阶段，但我们还是可以对很多实验进行思考②。如果你在具备理想实验条件的世界里都无法设计出一个实验来回答你的问题，那么在只有适度预算以及非实验调查数据的情况下，你能够得出有用结果的几率就相当渺茫了。对理想条件下的实验进行描述也可以帮助你准确地表达因果问题。在理想条件下进行实验的思考方法还可以凸显那些你希望操控的力量和那些你希望保持不变的因素。

对于那些不能被任何实验回答的研究问题，我们称之为根本无法识别的问题(Fundamentally Unidentified Questions，FUQ)。那什么是根本无法识别的问题

① 在 Angrist 和 Lavy(2008)中，对学生进行奖励来鼓励他们学习以通过大学入学考试，这一实验本身并不增加被实验者的人力资本，但是他们的研究发现被随机支付奖励的学生在 5 年后接受更高等教育的意愿更大，这会增加被实验者的人力资本，从而使他们在长期中获得较高的工资。——译者注

② 后来在一部电视专题节目中 William Shatner 扮演了 Milgram。至今，经济学家尚未获得过这种荣耀，但 Angrist(本书作者之一)对此仍抱有希望。

呢？第一印象告诉我们那些考虑种族和性别不同导致的因果效应问题应该比较接近于根本无法识别的问题，因为我们很难改变被试者的种族和性别，从而很难将这些因素与其他因素隔离分析（将其他因素隔离分析的最好办法是考虑同一个人在不同种族或性别下的情况，但这显然不现实，“想象一下你的染色体在出生的时候发生了改变”）。但是在另一方面，经济学家在种族、性别以及劳动市场歧视领域方面的研究集中于是否别人因为你是白人/黑人，男人/女人而对你区别对待（因此被试者究竟是白人/黑人，男人/女人并不重要，重要的是别人对此的看法）。而且，将男人看作女人，将女人看作男人的反事实世界已有很长的历史（比如罗莎琳德化装成古希腊神话中的美少年愚弄莎翁名剧《皆大欢喜》中的每一个人），因此我们对性别的因果效应进行研究并不需要亚当斯式的奇思异想。改变种族的想法显得不那么自然：在《人性污点》（The Human Stain）中，飞利浦·罗斯（Philip Roth，该剧作者）想象出一个以 Coleman Silk 为主角的世界，在那里 Coleman Silk 是一个在职业生涯中以白人身份示人的黑人文学教授。劳动经济学家也总是在设想此类场景。有时候我们甚至要构造这样的一些场景——比如在审计研究（audit studies）中伪造工作申请和简历①——以推动科学的进展。

将想象推进到研究设计需要走很长一段路，而且想象无法解决所有的问题。譬如说我们对晚一点入学的孩子是否在学校里表现更好这个问题感兴趣。因为七岁孩子的大脑可能比六岁孩子的大脑更适合学习的需要。这个问题来自于事实的政策含义在于有些学区为了提高考试成绩而推迟入学年龄（Deming and Dynarski，2008）。为了考察延迟入学对学习的影响，典型的做法就是我们随机挑选一些在六岁开始上学的孩子和在七岁开始上学的孩子。我们关心的问题是：是否如他们小学考试成绩显示的那样，晚一点入学的孩子可以学得更多。具体而言，让我们看一下一年级的考试成绩。

这个问题于是转化为入学年龄对一年级考试成绩的影响，但是问题的问题在于七岁入学的学生年龄本来就比较大。即使纯粹是发育的影响，年长的孩子也倾向于表现出好的成绩。为了避免这个因素对我们考虑问题的干扰，看上去我们应该控制年龄而不是控制学生所在的年级。于是假设我们考虑六岁上学的孩子到二年级时的成绩和七岁孩子在一年级时的成绩，在这个样本里所有孩子的成绩都是在其年龄为七岁时获得的。但是六岁上学的孩子在七岁时已经在学校有一年的时间，呆在学校里的任何有价值因素都可能引起学生成绩的提高（将这一现象称为在学影响）。因此，由于孩子们一直待在学校，所以无法将上面提到的发育影响和在学影响同时剔除以考察纯粹的入学年龄对成绩的影响。问题的核心乃是：对于学生而言，入学时间等于现有年龄减去在校时间（入学年龄＝现有年龄－在校时间，控制发育影响和控制在学影响相当于使现有年龄和在校时间都不变，那么这个恒

① 最近的一项研究是 Bertrand 和 Mullainathan（2004），他们考察了面对姓名首字母听上去像白人和听上去像黑人的简历，雇主们的反应有何区别。

等式告诉我们入学年龄也将不变，于是将无法观察不同年龄入学对学习成绩的影响)。上面这个恒等式在由成年人组成的样本中就不再成立了，因此我们可以考察纯粹的入学年龄对成年人的影响，这些影响包括收入水平或者高等教育完成水平(Black，Devereux and Salvanes，2008)。由于即使在随机实验中也无法获得入学年龄对小学生成绩的影响，因此这个问题就是根本无法识别的问题，即 FUQs。

第三个和第四个研究的常见问题(FAQ)主要涉及完成一项具体研究的细节。第三个问题是这样的：你的识别策略是怎样的？Angrist 和 Krueger(1999)使用识别策略这一术语来描述研究人员运用观察数据(也就是说，不是通过随机实验产生的数据)逼近真实实验的方式。我们再一次回到教育那个例子，Angrist 和 Krueger(1991)把美国学校的义务入学法和学生出生季节所成的交互项作为一项自然实验来估计高中学历对工资的影响(义务入学法要求学生在 16 岁生日以后才能辍学，但学生的出生季节影响学生的辍学率而与学生的能力无关)。本书从第 3 章到第 6 章的内容基本都是在考虑识别策略的概念性框架。

尽管对可靠的识别策略的关注可以作为现代经验研究的象征，但在计量经济学中将真实实验和自然实验相提并论还是经历了一个漫长的历史过程。下面是计量经济学先驱特里夫·哈尔维莫(Trygve Haavelmo，1944：14)呼吁对两种实验设计进行更为细致的讨论时所讲的一段话：

> 对于任何进行定量研究的理论而言，实验设计(物理学家称之为“决定性实验”)都是一项基本的附件。在我们构建理论时脑海中通常有一些这样的实验，虽然很不幸的是大多数经济学家无法精细地描述他们的实验设计。如果他们做到了这一点，就会发现其实脑海中的实验可以归为两类，也即(1)将所考虑问题从其他影响中分离出的实验，我们以此来考察特定的真实经济现象是否可以用来验证特定的假设；(2)自然从其巨大的实验室里不断川流而出的实验，对于此，我们主要是作为一个被动的观察者。无论哪种情况，理论的目的都是相同的：主，真实生活带来的幸福。

FAQ 的第四个问题借用了 Rubin(1991)的说法：你进行统计推断的模式是什么？为了回答这个问题，我们需要描述被研究的总体、所使用的样本以及构建标准误时所作的假设。有时候，推断是非常直观的，比如当你使用人口普查微观数据样本来研究美国人口的时候。不过，通常而言，推断则更为复杂，尤其是数据存在聚类问题或者被分组时。本书末章就涵盖了当你需要回答常见问题四时将会出现的实际问题。尽管推断问题往往有相当的技术性而且很难让人兴奋，但是一项经过精良构思而且从概念上来说激动人心的项目，其最终的成功也要依赖于这些统计推断的细节。这一有时令人大感沮丧的事实激发出了下面的一段计量经济学俳句，这是一位来自日本的计量经济学博士生 Keisuke Hirano 在完成自己的博士论文时妙手偶得的：

> t 统计言笑晏晏，如此迷人
> 试着将标准差集中——

显著性杳不可寻①

从上面的讨论中我们可以清楚地看到，FAQ 中涉及的四个研究问题是研究项目进行过程中的一部分。下列各章主要关注我们意欲回答这些问题时所要提出来的计量经济学疑难。换言之，一旦你的研究计划设定，这些问题都可能会在研究过程中遇到。不过，在转入经验研究的细节之前，我们对何以随机实验能够给我们以一个分析基准这一问题给出更为细致的解释。

① 试着将标准差集中指的是稳健的标准差(robust standard errors)，这句话是说也许 t 统计量很好，但是如果用稳健标准差，那么也许 t 统计量会变得不显著性。——译者注

▶2 理想的实验

重要而普遍的事实在于事情并非是我们看上去的那样。比如，在地球上，人类常常觉得自己的智商比海豚高，因为他们创造了诸如汽车、纽约和战争等东西，而所有的海豚所做的事情不过是潜入水下开心地生活。但是反过来想想，海豚也认为它们的智商比人类高——而且基于同样的原因。事实上这个星球上只有一种生物比海豚智商更高，他们花大量时间四处奔波，对各种行为进行研究并且对人类实施精妙之极的实验。由于造物主的计划，人类再次完全错误地解释了这种关系。

Douglas Adams, *The Hitchhiker's Guide to the Galaxy*

最可信和最有影响力的研究设计应该使用随机分配(random assignment)的方法。一个可以对此进行恰当解释的例子是 Perry 学前计划，它是一个在 1962 年实施的随机实验，意在估计对密歇根州伊斯兰堤(Ypsilanti, Michigan)的 123 个黑人学前儿童实施早期干预项目的效果。在 Perry 实验中，研究人员对处理组(treatment group)实施包括学前教育和家访在内的密集性干预措施。直到参与实验的被试者已经 27 岁的 1993 年，Perry 实验已经产生了相当的后续数据，该项实验的意义实在很难用言语描述。几十篇学术研究引用或者使用了 Perry 实验的发现(Barnett, 1992)。最重要的是，Perry 实验为大规模的入学前提前教育(Head Start Preschool Program)提供了理论基础，这项开始与 1964 年的入学前提前教育已经并将继续使百万计的美国儿童受益①。

2.1 选择性偏误

在正式讨论实验在揭示因果关系中所起的作用之前，我们先来考虑一个另外的问题。假设你对可以用“如果—那么”这种语句表述的某个因果关系感兴趣。具

① 特别是随着政策兴趣转向早期教育，Perry 实验的数据还在不断得到关注。由 Michael Anderson (2008)最近完成的分析确认了来自于原始的 Perry 实验的结论，但是 Anderson 还指出 Perry 实验从总体上反映出早期干预项目带来的正效应几乎完全来自于女孩，对男孩似乎没有起到什么作用。

体而言，让我们考虑一个简单的例子：医院能够让人变得更健康吗？就我们的目的而言，这个问题似乎有点隐喻在里面，但是它以令人惊讶的程度接近于健康经济学家所关心的因果关系。为了让这个问题更加贴近实际，想象我们正在研究一群贫穷的老年人，他们在医院的急诊室接受初级护理①。其中的一些人被医院接收。这种医疗方式比较昂贵并且拥挤的医院也使得相应的治疗不太有效(Grumbach, Keane and Bindman, 1993)。事实上，接触自身情况危险的重病患者对这些老年人的健康而言可能有负面影响。

因为在医院可能对健康状况有负面影响之外，被医院接收的那些人也能够得到很多有价值的服务，所以对医院是否能够让人变得更健康这一问题的回答似乎应该是“是”。但是数据支持这个说法么？对于一个倾向于进行经验研究的人而言，自然而然的方法就是比较去过医院和没去医院的人在健康状况上的差异。全国健康采访调研(National Health Interview Survey，简记为 NHIS)包含了进行这种比较的信息。具体而言，这个调研里包含这样一个问题，“在过去的 12 个月中，被访者是否曾因病在医院过夜?”，我们可以用这个问题来识别最近去过医院的人。全国健康采访调研还问，“总体而言，你觉得你的健康水平是极好、非常好、好、一般还是差?”下面的表格给出了最近去过医院和没有去过医院的人的平均健康状况(对健康状况最差的人赋值 1，对健康状况最好的人赋值 5，数据来自 2005 年的 NHIS)。

组　别	样本大小	平均健康水平	标准差
去过医院	7 774	3.21	0.014
没有去过医院	90 049	3.93	0.003

可见两者之间的平均差距是 0.72，没有去过医院的人健康状况更好，两者之差大且显著，其 t 统计值是 58.9。

从表面上看，这个结果意味着去医院会使人的健康状况变差。由于医院往往充满了可能会使我们受到感染的各类重病患者，危险的医疗仪器和化学药剂也可能伤害到我们，所以去医院会使人健康状况变差未必不是正确答案。但是，我们也很容易解释为什么这个结果不能只从表面上看：去医院的人可能本身健康水平就比较差。更进一步讲，平均而言，即使在医院接受过治疗，那些到医院寻求治疗的人的健康水平可能还是不如没有去医院的人，也就是说，对于去医院的那些人而言，不去医院只能让他们的健康状况变得更差，但是去医院也未必能让这些人的健康水平赶上不去医院的人。

① 根据书中提供的参考文献，在最近几十年，由于社区医疗机构和医疗器械的缺乏，美国人将原本用于对重伤或者有生命危险的患者进行治疗的急诊室用于初级护理。而且以这种方式进行初级护理的人往往是那些低收入、年老的人群，他们雇不起家庭医生，因此到医院进行初级护理。——译者注

为了更精确地描述这个问题，我们将接受医院治疗描述为一个二值随机变量，$D_i = \{0, 1\}$。我们所考虑的研究对象的结果——对健康水平的度量，记为 Y_i。我们的问题就是：Y_i 是否受医疗的影响。为了回答这个问题，我们想象去了医院的人如果没有去医院将会发生什么，没有去医院的人如果去了医院将会发生什么。因此，对于任何个体而言，他们的健康状况都有两种潜在结果：

$$\text{潜在结果} = \begin{cases} Y_{1i} & \text{if } D_i = 1 \\ Y_{0i} & \text{if } D_i = 0 \end{cases}$$

也就是说，假设一个人没有去医院，他的健康状况将是 Y_{0i}，而不论他事实上有没有去；假设一个人去医院接受了治疗，他的健康状况将是 Y_{1i}。我们想知道的是 Y_{1i} 和 Y_{0i} 之间的差距，这个差距就可以解释为第 i 个人在医院接受的治疗对其健康状况产生的影响。也就是我们一直希望研究的因果效应，这里的“因”是是否去医院接受治疗，“果”是两种选择下不同的健康状况，“因果效应”指的是两种健康状况之间的差别①。

观察到的结果 Y_i 可以用潜在结果的线性组合表示：

$$\begin{aligned} Y_i &= \begin{cases} Y_{1i} & \text{if } D_i = 1 \\ Y_{0i} & \text{if } D_i = 0 \end{cases} \\ &= Y_{0i} + (Y_{1i} - Y_{0i}) D_i \end{aligned} \tag{2.1.1}$$

可见在这个表达式中 $Y_{1i} - Y_{0i}$ 就是个体去医院接受治疗对其健康状况的影响。一般来说，Y_{1i} 和 Y_{0i} 在总体中都有相应的分布，因此对于不同的人，去医院接受治疗的因果效应是不一样的。但是由于我们不可能同时看到某个人的两种潜在的健康状况，所以我们必须比较同一类人去医院治疗和不去医院治疗对其健康状况的影响。

尽管在是否去医院接受治疗所带来的不同结果间进行简单比较并非我们想要的，但是这种肤浅的比较还是能告诉我们一些关于潜在结果的有益信息。下面这个公式就将去医院接受治疗与否带来的对平均健康水平的差异与我们感兴趣的平均意义上的因果效应(average casual effect)联系了起来：

$$\begin{aligned} \mathrm{E}[Y_i \mid D_i = 1] - \mathrm{E}[Y_i \mid D_i = 0] = & \underbrace{\mathrm{E}[Y_{1i} \mid D_i = 1] - \mathrm{E}[Y_{0i} \mid D_i = 1]}_{\text{处理的平均因果效应}} \\ & + \underbrace{\mathrm{E}[Y_{0i} \mid D_i = 1] - \mathrm{E}[Y_{0i} \mid D_i = 0]}_{\text{选择性偏误}} \end{aligned}$$

其中，

$$\mathrm{E}[Y_{1i} \mid D_i = 1] - \mathrm{E}[Y_{0i} \mid D_i = 1] = \mathrm{E}[Y_{1i} - Y_{0i} \mid D_i = 1]$$

① 这里使用的潜在结果的观点是目前对因果关系进行研究的基石。在发展这个概念过程中出现的重要参考文献包括 Rubin(1974, 1977)以及 Holland(1986)，其中 Holland(1986)将潜在结果中蕴含的因果框架称为 Rubin 的因果模型。

就是那些接受医院治疗的人因为在医院得到治疗而获得的平均因果效应。这里 $E[Y_{1i} \mid D_i = 1]$ 是接受住院治疗的人的平均健康水平，$E[Y_{0i} \mid D_i = 1]$ 是如果接受住院治疗的人本来没有得到治疗，他们的平均健康水平。我们能够观察到的健康状况的差异实际上由两部分组成，在我们关心的因果关系之外，剩下的那部分叫做选择性偏误(selection bias)。它是去医院接受治疗与不去医院接受治疗的人如果没有被治疗时健康状况的平均差别。由于患病者比健康人更加倾向于寻求治疗，所以那些接受住院治疗的人的初始健康水平 Y_{0i} 本身就比较低，从而使得选择性偏误是负的。在这个例子中，选择性偏误的绝对值可能会很大，当它大过我们想要寻找的因果效应时，就足以掩盖我们所要寻找的因果关系的符号，使得观察到的情况和真实情况相反。因此，经济学中大部分经验研究的目的就是剔除这种选择性偏误，从而阐释某个变量带来的效果，比如这里的变量 D_i[①]。

2.2 用随机分配解决选择性偏误

对 D_i 进行随机分配可以解决上文提到的选择性偏误，因为随机分配使得 D_i 独立于潜在的结果。为了理解这一点，让我们考虑：

$$\begin{aligned} E[Y_i \mid D_i = 1] - E[Y_i \mid D_i = 0] &= E[Y_{1i} \mid D_i = 1] - E[Y_{0i} \mid D_i = 0] \\ &= E[Y_{0i} \mid D_i = 1] - E[Y_{0i} \mid D_i = 0] \end{aligned}$$

其中，Y_{0i} 和 D_i 之间的独立性使得我们可以知道 $E[Y_{0i} \mid D_i = 1] = E[Y_{0i} \mid D_i = 0]$，从而可以将等式(2.1.1)中的第二行选择性偏误消去。事实上，给定随机分配下 D_i 的独立性，我们还可以对因果效应继续简化：

$$\begin{aligned} E[Y_{1i} \mid D_i = 1] - E[Y_{0i} \mid D_i = 1] &= E[Y_{1i} - Y_{0i} \mid D_i = 1] \\ &= E[Y_{1i} - Y_{0i}] \end{aligned}$$

也就是说对接受医院治疗的人考虑因果效应等同于随机分配患者进行治疗得到的因果效应。主要的发现就是，随机分配 D_i 消去了选择性偏误。这并不意味着随机分配本身不存在问题，但是总的来说它解决了在经验研究中遇到的最重要的问题。

上面我们讲到的医院治疗的故事有多切题呢？实验往往能够揭示一些基于简单比较所看不到的东西。来自医学的新近例子乃是对荷尔蒙替代疗法(hormone replacement therapy，简写为 HRT)治疗效果的研究。荷尔蒙替代疗法是推荐给中年女性用以减少更年期症状的医学干预疗法。来自护士健康研究(Nurses' Health Study)——一项大且颇有影响力的针对护士的非实验调研给出的证据指出 HRT 使用者拥有更高的健康水平。相比之下，最近的一项完全基于随机实验的结果指

① 这一小节标志着我们第一次使用条件期望算子(比如 $E[Y_i \mid D_i = 1]$ 和 $E[Y_i \mid D_i = 0]$)。我们在这里用这个算子来表示一个随机变量固定时另外一个随机变量在总体中的均值。更加正式的定义请见第 3 章。

出 HRT 几乎没有什么效果。更为糟糕的是，随机实验还揭示出多项副作用，这些副作用在非随机实验中很不明显的。[见女性健康行动计划，Women's Health Initiative[WHI]，Hsia 等(2006)]。

在我们自己的领域劳动经济学中，一项标志性的研究就是对政府补贴的培训计划的评估。这些培训计划为那些长期失业、瘾君子和刑满释放人员等在就业市场上处于不利地位的人提供一系列的培训项目，这些项目包括课堂内指导和在职培训。这个项目的目的在于提高这类人的就业率和收入。但矛盾的是基于非实验的研究方法对项目参与者和非参与者进行比较后发现，在接受培训后受培训者比相应的对照组赚得更少(Ashenfelter，1978；Ashenfelter and Card，1978；Lalonde 1995)。这里，由于受补贴的培训项目意在针对低收入人群，所以显然需要考虑其中是不是存在选择性偏误的问题。因此，对项目参与者和非参与者收入状况的简单比较往往可能显示出项目参与者会得到更低的收入。相比之下，对培训项目进行随机实验的研究发现这些培训项目大都具有正面效果(Lalonde，1986；Orr et al.，1996)。

就现在看来，相比于医学研究，随机实验在社会科学研究中的使用还不是那么广泛，但是它正逐渐变得流行。随机分配的重要性得到逐步提升的一个领域就是对教育的研究(Angrist，2004)。在 2002 年获得美国国会通过的教育科学改革法案(2002 Education Sciences Reform Act)正式规定，任何由联邦政府资助的教育学研究都必须使用严格的实验或者准实验的研究方法。因此我们可以预见在未来还会有更多的随机实验出现在教育学研究中。在教育学研究中开创性地使用随机化研究方法的是田纳西州的师生比例改进计划[Tennessee Student Teacher Achievement Ratio(STAR) experiment，之后用 STAR 表示这个项目]，该计划用以评估小学小班教学的效果。

劳动经济学家和其他的一些人一直都希望在课堂教学环境和学生学习成绩之间建立因果关系，我们将这一研究领域叫做“教育生产(education production)”。这个名称的含义来自于我们将花钱打造各种课堂环境看作对教育的投入，将学生成绩看作教育的产出。在对教育生产进行研究的过程中出现的一个关键问题就是给定成本下哪种投入可以使学生成绩最大化。在学校投入中最昂贵的一笔投入就是课堂规模，因为只有通过多雇用教师才能将课堂规模降下来。因此对昂贵的小班教学能否使学生获得好成绩的研究具有重要意义。STAR 实验就旨在回答这个问题。

很多对教育生产进行研究的论文都使用非实验数据，这些研究指出在课堂规模和学生成绩之间几乎没有联系。因此也许学校在不降低教学质量的情况下可以通过扩大课堂规模，减少雇用老师来降低成本。因为较羸弱的学生往往被有意编入规模较小的班级(也就是说运用非实验数据研究班级规模和学生成绩时，班级规模和学生的某些特点相联系，从而使得选择性偏误不为零)，所以我们不可以只通过简单比较可观察数据来看待课堂规模和学生成绩之间的关系。这时，随机实验

可以帮助我们跨越这个障碍，保证我们是在用苹果比较苹果，也就是说分配给不同规模的班级中的学生是可比的。来自 STAR 实验的结果指出小班教学有持久而强烈的回报[见 Finn 和 Achilles(1990)的最初研究以及 Krueger(1999)对 STAR 数据进行的计量分析]。

STAR 实验在宏大性和影响力上都值得我们在这里对它进行更加细致的讨论。这项研究花费 1 200 万美元，于 1985 年和 1986 年之间在一批幼儿园中开始实施，持续四年，调查了 11 600 位小孩子，直到最初还在幼儿园的孩子已经升入小学三年级为止。在 1985 年到 1986 年之间，田纳西州普通课堂的平均规模是 22.3。这项实验将学生分配至三个处理组(treatment group)：小班，课堂容量在 13—17 人之间；普通班级，课堂容量在 22—25 人之间并配备一位兼职助理教师(这是田纳西州对课堂规模的普遍设置)；普通/助理班级，课堂容量在 22—25 人之间，配备一名全职助理教师。每个年级有三个以上班级的学校都参加了这项实验。

对随机实验提出的第一个问题就是随机化是否成功地平衡了不同处理组间的各种特征。为了回答这个问题，往往需要比较处理之前的结果或者在组间比较不被处理过程影响但是影响被试者的变量。不幸的是，STAR 实验没有包括任何处理前的测试分数，因此可能的办法就是看看学生的个体特征，比如种族、年龄等。表 2.1 来自于 Krueger(1999)，可以帮助我们比较这些变量的平均值。表中代表学生个体特点的变量分别是免费午餐、学生种族和学生年龄。免费午餐是对学生家庭收入的良好度量，因为只有贫穷的家庭才满足享受免费午餐的标准。在三个类别中，上述变量之间的差距都很小，最后一列的 p 值显示出没有一个差距显著地不为零。这意味着随机分配如我们所希望的那样成立。

表 2.1 还为平均班级规模、损耗率、以百分比估计的测验成绩提供了信息。损耗率(没有出现在下一年样本中的学生)在幼儿园的小规模班级中更低。一般而言这将是一个潜在的问题①。班级规模在有意分配为小班的组中显著地缩小，这意味着实验成功获得了研究所希望的班级规模的不同变化。如果很多孩子的家长通过游说老师和校长，成功使得他们的孩子进入小规模班级，那么班级规模的差距将会大大缩小。

由于随机化实验可以去掉选择性偏误，所以各个处理组之间的不同就捕捉住了班级规模不同造成的平均因果效应(相对于拥有兼职助理教师的普通班级，也就是将普通班级作为实验的对照组)。在实际中，可以通过将测验成绩关于标志每个处理组的虚拟变量进行回归来得到处理组和对照组之间平均成绩的差异，这是我们在本章下一节将会展开的内容。用回归模型得到的对幼儿园中处理组—对照组之间差别的参数估计报告在表 2.2 中(来自于 Krueger，1999，表Ⅴ)，该表指出小班教学对教学效果大概有 5%的提高(该表中其他行显示的系数都是控制变量的

① Krueger(1999)花费相当篇幅讨论了损耗问题。不同组之间不同的损耗率可能导致较高年级的一部分学生不是随机分布的。对幼儿园的研究结果不受损耗率的影响，因此更加可信。

系数)。因果效应的影响范围大约是 0.2σ,其中 σ 是幼儿园百分位数成绩的标准差。小班教学的因果效应显著地不为零,但是普通/助理班级对学生成绩的影响则不显著。

表 2.1　在田纳西州的 STAR 实验中比较处理组和控制组的特点

变　量	课堂规模			关于组间等同性的 p 值
	小	普通	普通+助理教师	
免费午餐	0.47	0.48	0.50	0.09
白人还是亚裔	0.68	0.67	0.66	0.26
在 1985 年时的年龄	5.44	5.43	5.42	0.32
损耗率(Attrition rate)	0.49	0.52	0.53	0.02
幼儿园中的班级规模	15.10	22.40	22.80	0.00
幼儿园时期百分位数成绩	54.70	48.90	50.00	0.00

注:本表来自 Krueger(1999)的表Ⅰ。本表按照不同的班级规模列出了在幼儿园参加 STAR 实验学生在各个变量上的平均值。通过最后一列的 p 值可以知道三个组间变量均值的相等程度。免费午餐变量指的是学生中接受免费午餐的比例。百分位数成绩是基于三个斯坦福标准化试题的平均成绩得到的。损耗率是指在完成三年级之前在随后年份中从样本中损失掉的学生比例。

表 2.2　对班级规模对考试成绩影响的实验设计估计结果

解释变量	(1)	(2)	(3)	(4)
班级规模	4.82	5.37	5.36	5.37
	(2.19)	(1.26)	(1.21)	(1.19)
普通+助理教师	0.12	0.29	0.53	0.31
	(2.23)	(1.13)	(1.09)	(1.07)
白人还是亚裔	—	—	8.35	8.44
			(1.35)	(1.36)
女孩	—	—	4.48	4.39
			(0.63)	(0.63)
免费午餐	—	—	−13.15	−13.07
			(0.77)	(0.77)
白人教师	—	—	—	−0.57
				(2.10)
教师经验	—	—	—	0.26
				(0.10)
院长学历	—	—	—	−0.51
				(1.06)
学校固定效应	No	Yes	Yes	Yes
R^2	0.01	0.25	0.31	0.31

注:该表来自于 Krueger(1999)中的表Ⅴ。被解释变量是斯坦福标准化测试的百分位成绩。允许组内残差项之间相关的稳健的标准差见相应系数下的圆括号中。这个表对应的样本数是 5 681。

在社会科学史上属于随机实验范例的 STAR 研究同时也凸显了随机实验需要后勤保障、长持续时间和潜在的高成本所带来的困难。在很多情况下，进行这样的随机实验都是不现实的①。在其他一些例子中，我们希望尽早得到一个答案。因此我们做的很多研究尝试发掘更便宜、更好用的变异来源。我们希望找到自然实验或者准实验，通过改变感兴趣变量来模仿一个随机实验，同时将其他因素保持住。我们经常能够找到令人信服的自然实验吗？当然不能。我们将理论上的随机实验作为我们分析的基础。虽然不是所有的研究者都同意这个观点，至少很多是同意的。我们最早从我们的老师和毕业论文指导教师 Orley Ashenfelter 那里听到了这句话。Orley Ashenfelter 是一位在社会科学中最先支持使用实验和准实验研究方法的先驱，以下是 Ashenfelter(1991)对考察教育水平和收入之间关系的研究结论的可靠性给出的评价：

> 将教育水平和收入相联系的证据有多么可信？这里是我的答案：相当可信。如果我必须关于理想条件下的实验所可能得到的结论打赌的话，我赌这个实验将会指出教育水平越高的工人赚得越多。

对班级规模的准实验分析由 Angrist 和 Lavy(1999)完成，他们在这篇论文中展示了如何运用随机实验的思想来分析非实验数据。Angrist 和 Lavy 的研究依赖于这样的事实：在以色列，班级规模的上限是 40 人。因此，如果一个学生所在的五年级总共有 40 个学生，那么这些学生将被编排进入一个规模为 40 人的班级，但是如果一个学生所在的五年级总共有 41 名学生，那么这些学生所在的班级规模就是 41 的一半，也就是 20.5。由于一个年级的学生究竟是 40 个还是 41 个完全是随机的事情，所以不同规模的两个学生群体应该在诸如能力和家庭背景方面都比较相似，于是我们可以把年级规模为 40 和年级规模为 41 的两个学生群体间学习成绩的差异归因于班级规模，也就是说 40 个学生和 41 个学生之间学习成绩的差异可以看作是“就像随机分配出来的那样好”。

Angrist-Lavy 的研究以年级为单位，在没有真实实验的情况下，利用年级入学人数高于和低于官方班级规模上限的条件，估计出了班级规模锐减带来的因果效应，而且这个因果效应是在很好地控制了各方影响因素的条件下得到的。正如在 STAR 实验中得到的结果那样，Angrist 和 Lavy(1999)的结果指出班级规模和学生成绩之间有强烈的联系。这个结果和简单比较后得到的结果形成鲜明对比。简单比较指出的结论是小规模班级中的孩子在标准化测试中表现得更差。可见，我们刚才讲过的接受医院治疗对健康影响中发生的选择性偏误在这里也出现了②。

---

① 随机实验也不是完美的，即使 STAR 也不例外。重复出现或者跳级的小学生们被排除在实验之外。学生进入参加实验的学校上过一年课后会被加入实验从而随机地分配到一个班里。因此这个实验不幸的一面就是也许因为在普通班的学生家长的抗议，学生在幼儿园毕业后被重新分班，进入普通班或者有助理教师的普通班。而且在幼儿园毕业后，还有学生中途转班。但是 Krueger (1999)的分析指出，上面提到的这些操作性问题都不会影响该研究的主要结论。

② Angrist 和 Lavy(1999)的结果在第 6 章还会出现，将被作为准实验，非连续回归的研究设计的例子。

2.3 对实验的回归分析

无论使用的数据来自实验与否，回归都是研究因果关系的有用工具。假设(从现在起)因果效应对所有人都一样，也就是 $Y_{1i}-Y_{0i}=\rho$，是个常数。如果因果效应被假设为常数，那么我们可以将等式(2.1.1)写为：

$$Y_i=\underbrace{\alpha}_{E(Y_{0i})}+\underbrace{\rho}_{(Y_{1i}-Y_{0i})}D_i+\underbrace{\eta_i}_{Y_{0i}-E(Y_{0i})} \tag{2.3.1}$$

其中，η_i 是 Y_{0i} 的随机部分。根据处理状态(treatment status，指对被试者进行了何种处理)的有无，对上面这个等式求条件期望可得：

$$\begin{aligned}E[Y_i \mid D_i=1]&=\alpha+\rho+E[\eta_i \mid D_i=1]\\E[Y_i \mid D_i=0]&=\alpha+E[\eta_i \mid D_i=0]\end{aligned}$$

于是：

$$\begin{aligned}&E[Y_i \mid D_i=1]-E[Y_i \mid D_i=0]\\&=\underbrace{\rho}_{\text{处理效应}}+\underbrace{E[\eta_i \mid D_i=1]-E[\eta_i \mid D_i=0]}_{\text{选择性偏误}}\end{aligned}$$

因此选择性偏误意味着回归残差项 η_i 和回归元 D_i 之间的相关性。由于：

$$E[\eta_i \mid D_i=1]-E[\eta_i \mid D_i=0]=E[Y_{0i} \mid D_i=1]-E[Y_{0i} \mid D_i=0]$$

上面等式指的是得到处理和没有得到处理的人的潜在结果的差别。在医院治疗的那个故事里，得到医院治疗的人的健康状况要差于没有得到医院治疗的人，在Angrist和Lavy (1999)的研究中，在更小的班级中的学生本身的测验分数就比较低。

在STAR实验中，D_i 是随机分配的，所以选择性偏误项就消失了，对 Y_i 关于 D_i 的回归就估计出我们感兴趣的因果效应 ρ。表2.2给出了使用不同回归模型时估计出的不同参数，这些不同的模型有些包含了变量 D_i 之外的一些控制变量。在用回归模型分析实验数据时使用控制变量有两个用处。首先，在STAR实验中使用了条件随机分配方法。具体而言，在同一学校内，将学生分配至不同的班级是随机的，但是在学校间，这种分配不是随机的(有些学生必然会在某个学校)。在不同类型学校(比如城市里的学校和农村的学校)接受教育可能会影响学生被分配进入小规模班级的可能性。表2.2中第一列没有考虑这一问题，因此这一列估计出的参数可能会被不同类型学校对学生成绩的影响而干扰。为了进行调整，Krueger在一些回归方程中包含了学校固定效应，也就是对每个进入STAR数据的学校估计一个截距项。但事实上，对学校固定效应的调整并没有对结果带来大的改变，但是如果我们不这样做，就不会知道这个事实。我们在第5章将详细讨论包含固定效应的回归模型。

在Krueger模型中的其他控制变量描述了学生的个体特征，这些特点包括诸如种族、性别、是否参与免费午餐等。在之前我们就已知道这些个体特征在不同班

级类型之间已经得到平衡，也就是说这些特点已经系统性地与将学生分配至哪种类型的班级无关了。记这些控制变量为 X_i，它们与 D_i 不相关，因此也就不会影响对 ρ 的估计。换句话说，在长的回归方程：

$$Y_i = \alpha + \rho D_i + X_i'\gamma + \eta_i \tag{2.3.2}$$

里估计出的 ρ 与在等式(2.3.1)中短的回归中估计出的 ρ 将会很接近。这一点我们将在第 3 章详尽展开。

尽管在我们刚才考虑的这个例子中将变量 X_i' 包含进回归方程并无必要，但是一般来说这种做法可以为我们带来对因果关系的更加精确的估计。注意到在表 2.2 中，第三列里估计出的因果效应对应的标准误要小于第二列对应的标准误。虽然控制变量 X_i 与 D_i 无关，但是它们对被解释变量 Y_i 有相当的解释力度。因此将这些控制变量包含进回归可以减少残差的方差，从而降低回归的标准误。相同的，由于学校固定效应也能解释学生成绩变动的一部分，所以将其纳入回归也可以减少估计值 ρ 的标准误。表 2.2 的最后一行加入了教师特征。因为教师也是随机地分配给各种类型的班级的，所以教师特征对于数据里的学生成绩没有影响，这表现在加入教师特征后既没有改变估计值也没有改变标准误。

回归在经济学的实证研究中扮演着极为重要的角色。正如我们在这一章所见，用回归来分析实验数据是很恰当的。在某些例子中，回归还可以在缺少随机分配的情况下来近似实验。但是在开始讨论什么时候回归的结果可以被解释为因果关系之前，我们先来回顾几个关于回归的基本事实和性质。这些事实和性质对所有回归都成立，与我们进行回归的目的无关。

第二部分　核　　心

▶3

让回归变得有意义

"让我们思考那些原本无法考虑的事物，让我们做那些原本没法去做的事。
让我们准备好与不可言喻的事物作战，
看看是不是我们根本无法将其降服。"
Douglas Adams, *Dirk Gently's Holistic Detective Agency*

Angrist 讲述道：

我第一次做回归还是在 1979 年的夏季，当时我是奥伯林学院(Oberlin College)一名大一升大二的学生。作为一名研究助理，我为 Alan Meltzer 和 Scott Richard 工作，他们是在位于我家乡匹兹堡附近的卡内基—梅隆大学任教的老师。那个时候，我仍然对将来在特殊教育方面的职业生涯充满憧憬，因此计划像前一个暑期一样，作为一名勤杂人员到州立精神医院工作。但是经济学课程却使我陷入深思，也使我看到，在同样的工资水平上，一个研究助理的可支配时间和工作条件要比一个医院勤杂人员好得多。当时，我这个研究助理的职责主要包括数据收集和回归分析，虽然那个时候我压根不懂回归，甚至也不大了解统计知识。

那个夏天我所从事的论文(Meltzer and Richard, 1983)试图用 GDP 中政府支出比例作为对民主政府规模的衡量，将其与收入不平等之间建立某种联系。由于大部分收入分布都会显示出一个右偏的长尾，这意味着平均收入会大于收入的中位数。因此一旦不平等加剧，会有很多投票者发现他们低于平均收入水平。那些收入处在中位数和平均值之间的人们会因为自己收入水平相对于平均水平的下降而愤怒，于是就与收入低于中位数的人们联合起来投票支持劫富济贫的财政政策。于是政府规模扩大。

由于穷人的投票率比较低，所以我并不觉得 Meltzer 和 Richard 的论点总是成立，但当时还是被他们的基本理论所吸引。我还记得曾与 Alan Meltzer 就政府对教育的支出是否应该归类为公共品(那些对全社会所有人都有益并且对他们产生直接影响的商品)还是公共供给的私人品——一种类似于福利的社会再分配形式，进行过争论。你可能会说，这个研究项目标志着我对教育

的社会收益的最早兴趣，的确，后来在 Acemoglu 和 Angrist(2000)的论文中，我以更大的热情和更深刻的领会重新回到这个主题。

时至今日，我把 Meltzer 和 Richard 的研究理解为使用回归以发现和定量分析令人感兴趣的因果联系的一次尝试。不过在当时，我纯粹是一个“回归机器”。有时候我甚至发现研究助理的工作令人沮丧。在只与老板以及偶尔与那些不怎么会说英语的卡耐基—梅隆大学博士生的交流中，日子飘然而逝。这项工作最好的地方就是能和 Alan Meltzer 一起共进午餐，Alan 是一位杰出的学者，也是一位富有耐心和天性纯良的导师，在我们一起分享食物时他很乐于谈天(由于 Alan 吃的很少，而我又吃的太快，所以这种时间总是持续得很短)。我记得，曾经问 Alan，他是否对把时间花费在寻求回归结果上感到满意，因为那之后只不过是很多双面加宽绿色条的论文而已。他大笑，然后说这是他最愿意做的。

现在，我们也和我们在大学和研究院的老师和指导教师一样，整日快乐地追寻着回归结果。本章来解释个中缘由。

3.1 回归的基本原理

在前一章末尾，我们介绍了回归模型可以作为估计实验中处理组—控制组之间差异的计算工具，并分别在有无协变量的情况下进行了讨论。由于在 2.3 节讨论课堂规模影响的研究中我们关心的回归元(regressor)是随机分配的，所以可以对由此得到的估计量赋予一个因果解释。然而，在大多数研究中，回归模型中使用的是观察到的数据，而非随机实验产生的数据。在没有随机分配可以利用时，我们未必能对回归估计量赋予一个因果解释。在本章后半部分，我们会回来讨论一个核心问题——当回归满足什么条件时我们可以对其赋予因果解释。

这里我们暂时把相对抽象的因果问题搁置一旁，先来讨论一些回归估计的普遍性质。这些性质都是总体回归向量和相应的样本估计量的普适性质，与研究者如何解释他们的结果无关。这些性质主要分为两块，一是总体回归方程和条件期望函数之间的关系，二是回归结果的样本分布。

3.1.1 经济学中的关系和条件期望函数

在我们的研究领域劳动经济学中，经验研究尤其关心对个体经济状况的统计分析，特别是可用来解释人们财富水平差异的那些个体差异。人们所拥有的财富之间的差异是出了名地难以解释：一言以蔽之，它们都是随机发生的。不过，身为应用计量经济学家，我们相信能够以一种有效的方式来概括和解释这种随机性。关于“系统性随机性”(systematic randomness)的一个例子就是在导论中提到的教育水平和收入之间的联系。平均而言，教育水平更高的人赚得更多。尽管个体状

况千变万化，有时甚至会掩盖这一事实，但教育水平和收入之间的联系还是有相当的预测力。当然，教育水平更高的人赚得更多并不意味着教育水平的提高导致(cause)了收入提升。教育水平与收入之间是否具有因果性这一问题在我们的讨论中占据了极为重要的地位，我们还会多次回到这个问题上来。不过，即便无法解决因果性这一难题，我们也很显然地知道在一种狭义的统计学意义上，教育水平能够预测收入。这里我们使用条件期望函数(conditional expectation function，简写为 CEF)来概括和总结这种预测能力。

给定一个 $k\times 1$ 维的协变量 X_i(其中第 k 个元素记为 x_{ik})，被解释变量 Y_i 的条件期望函数就是给定 X_i 不变时 Y_i 的期望或者是总体均值，当 X_i 取遍所有的可能值后，得到的关于 Y_i 的各种期望值就成为关于 X_i 的函数。总体均值则可以看作一个无限大样本中的平均值，或者是可数有限总体(completely enumerated finite)的平均值。条件期望函数记为 $\mathrm{E}[Y_i \mid X_i]$，是关于 X_i 的函数。因为 X_i 是随机的，所以条件期望函数也是随机的，但有时我们会考虑条件期望函数的一个特定值，比如假设 42 是 X_i 的一个可能值，那么 $\mathrm{E}[Y_i \mid X_i = 42]$ 就是条件期望函数的一个特定值。在第 2 章，我们曾短暂地讨论过条件期望函数 $\mathrm{E}[Y_i \mid D_i]$，在那里 D_i 是只取 0 和 1 的变量。于是该条件期望函数取两个值，一个是 $\mathrm{E}[Y_i \mid D_i = 1]$，一个是 $\mathrm{E}[Y_i \mid D_i = 0]$。尽管这种特例很重要，但是我们更关注以多变量函数形式出现的条件期望函数，并且简单起见，将这些变量记为向量 X_i。对于向量 X_i 的某个特定值，比如 $X_i = x$，我们可以将条件期望函数记为 $\mathrm{E}[Y_i \mid X_i = x]$。对于条件密度函数为 $f_y(t \mid X_i = x)$ 的连续随机变量 Y_i 而言，其条件期望函数定义为：

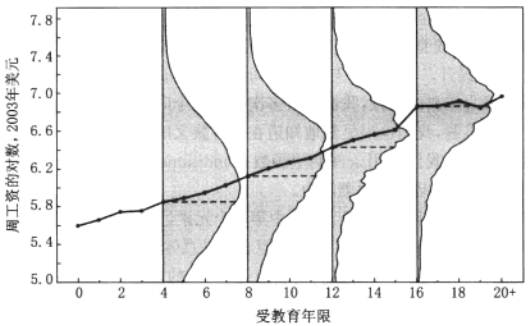
$$\mathrm{E}[Y_i \mid X_i = x] = \int t f_y(t \mid X_i = x)\mathrm{d}t$$

如果 Y_i 是离散的，$\mathrm{E}[Y_i \mid X_i = x]$ 等于 $\sum_t tP(Y_i = t \mid X_i = x)$，此处 $P(Y_i = t \mid X_i = x)$是给定 $X_i = x$ 时 Y_i 的条件概率质量函数[①](conditional probability mass function)。

期望是一个总体意义上的概念。在实际中，我们得到的数据通常来自总体所包含的一部分样本，而且很少有样本是由完整的总体构成的。因此我们要使用样本对总体作出推断。比如说，用样本的条件期望函数来了解总体的条件期望函数。通过样本对总体作出推断是必要且重要的，但我们先来考察总体，将通过样本推断总体的正式讨论推迟到 3.1.3 节。我们使用的这种“总体优先(population-first)”的教学方法来自于对以下事实的认识：在使用数据进行研究前，我们必须首先定义感兴趣的对象[②]。

① 概率质量函数与概率密度函数的不同之处在概率密度函数是对连续随机变量定义的，本身不是概率，概率质量函数是对可数离散随机变量定义的，本身就是概率。——译者注

② 使用“总体优先”这一教学方法的计量经济学教科书还有 Chamberlain(1984)，Goldberger(1991)和 Manski(1991)。



注：该样本由 1980 年公用微观样本数据集(IPUMS)中抽取的 5%的 40—49 岁之间的白人男性组成。

图 3.1　原始数据和给定受教育年限后对周工资对数平均值作出的条件期望函数

图 3.1 绘出了以 1980 年美国人口普查中抽取的中年白人男子作为样本，在教育水平给定后得到的对数化周工资曲线，显然这条曲线就是收入的条件期望函数。对于几个教育水平的关键值 4，8，12 和 16，图 3.1 还绘出了收入分布。在这幅图中，条件期望函数捕捉到的事实是：尽管个体状况差异很大，但是教育水平高的人通常赚得更多。与多接受一年教育相联系是平均收入提高十个百分点。

条件期望函数的一个重要性质就是它适用于迭代期望律(the law of iterated expectations)。该定律指出无条件期望(unconditional expectation)可被写作条件期望函数的无条件平均值(unconditional average)。换言之：

$$\mathrm{E}[Y_i] = \mathrm{E}\{\mathrm{E}[Y_i \mid X_i]\} \tag{3.1.1}$$

其中，处在外面一层的期望算子针对 X_i 的分布求期望。这里我们对具有联合密度函数 $f_{xy}(u, t)$ 的连续分布随机变量 (X_i, Y_i) 证明迭代期望律，其中 $f_y(t \mid X_i = u)$ 是给定 $X_i = u$ 时 Y_i 的条件密度函数，$g_y(t)$ 和 $g_x(u)$ 是边际密度函数：

$$\begin{aligned}
\mathrm{E}\{\mathrm{E}[Y_i \mid X_i]\} &= \int \mathrm{E}[Y_i \mid X_i = u] g_x(u)\,\mathrm{d}u \\
&= \int \left[\int t f_y(t \mid X_i = u)\,\mathrm{d}t\right] g_x(u)\,\mathrm{d}u \\
&= \iint t f_y(t \mid X_i = u) g_x(u)\,\mathrm{d}u\,\mathrm{d}t \\
&= \int t \left[\int f_y(t \mid X_i = u) g(u)\,\mathrm{d}u\right]\mathrm{d}t \\
&= \int t \left[\int f_{xy}(u, t)\,\mathrm{d}u\right]\mathrm{d}t \\
&= \int t g_y(t)\,\mathrm{d}t = \mathrm{E}[Y_i]
\end{aligned}$$

在这个表达式中积分遍历 X_i 和 Y_i 的所有取值(记为 u 和 t),所以我们看到的是二重积分。我们之所以在这里详细展开迭代期望律的证明,是因为条件期望函数及其性质居于本章剩余部分的核心①。

迭代期望律的威力在于它能将随机变量分成两部分,一部分是条件期望函数,一部分是有特殊性质的残差项。

定理 3.1.1:条件期望函数的分解性质(The CEF Decomposition Property)

$$Y_i = \mathrm{E}[Y_i \mid X_i] + \varepsilon_i$$

其中(i)ε_i 关于 X_i 均值独立,也就是说 $\mathrm{E}[\varepsilon_i \mid X_i] = 0$,因此有(ii)$\varepsilon_i$ 与关于 X_i 的任何函数都不相关。

证明:(i)$\mathrm{E}[\varepsilon_i \mid X_i] = \mathrm{E}[Y_i - \mathrm{E}[Y_i \mid X_i] \mid X_i] = \mathrm{E}[Y_i \mid X_i] - \mathrm{E}[Y_i \mid X_i] = 0$。(ii)令 $h(X_i)$是任意一个关于 X_i 的函数。由迭代期望律,$\mathrm{E}[h(X_i)\varepsilon_i] = \mathrm{E}\{\mathrm{E}[h(X_i)\varepsilon_i \mid X_i]\} = \mathrm{E}\{h(X_i)\mathrm{E}[\varepsilon_i \mid X_i]\}$,于是由 ε_i 关于 X_i 均值独立可知命题得证。

这个定理说明,任一随机变量 Y_i 都可以分解成"由 X_i 解释"的部分——也就是条件期望函数——以及正交于 X_i 的任何函数的部分(也就是说 ε_i 与 X_i 的任何函数不相关)。

有一系列的理由可以说明条件期望函数是对 Y_i 和 X_i 之间关系的良好概括。首先,我们往往将平均值看作是随机变量的代表值。更正式地说就是,在最小均方误的意义下,条件期望函数是对 Y_i 的最好预测。条件期望函数的这种预测性质乃是条件期望函数分解性质的一个推论:

定理 3.1.2:条件期望函数的预测性质(The CEF Prediction Property)

令 $m(X_i)$是关于 X_i 的任何函数。条件期望函数是下面问题的解:

$$\mathrm{E}[Y_i \mid X_i] = \arg\min_{m(X_i)} \mathrm{E}[(Y_i - m(X_i))^2]$$

因此条件期望函数就是给定 X_i 后对 Y_i 的最小均方误预测。

证明:记

$$\begin{aligned}(Y_i - m(X_i))^2 &= ((Y_i - \mathrm{E}[Y_i \mid X_i]) + (\mathrm{E}[Y_i \mid X_i] - m(X_i)))^2 \\ &= (Y_i - \mathrm{E}[Y_i \mid X_i])^2 + 2(\mathrm{E}[Y_i \mid X_i] - m(X_i)) \\ &\quad \times (Y_i - \mathrm{E}[Y_i \mid X_i]) + (\mathrm{E}(Y_i \mid X_i) - m(X_i))^2\end{aligned}$$

可见在上式第二个等号后的三个相加的表达式中,第一项不包含 $m(X_i)$,所以与我们要做的最小化问题无关。第二项可以写成 $h(X_i)\varepsilon_i$,其中 $h(X_i) = 2(\mathrm{E}[Y_i \mid X_i] - m(X_i))$,因此由条件期望函数分解定理,这一项的期望为零。当 $m(X_i) = \mathrm{E}[Y_i \mid X_i]$ 时,最后一项被最小化。

条件期望函数的最后一个性质与之前的条件期望函数分解性质和预测性质紧

① 一个简单的例子即可阐明迭代期望律的原理:在一个由男性和女性构成的总体中,平均收入等于男性收入平均值乘以其在总体中的比例加上女性收入平均值乘以女性在总体中的性别比例。

密相关，这就是方差分析(analysis of variance，简称为 ANOVA)定理：

定理 3.1.3：*方差分析定理*

$$V(Y_i) = V(\mathrm{E}[Y_i \mid X_i]) + \mathrm{E}[V(Y_i \mid X_i)]$$

其中，$V(\cdot)$表示方差，$V(Y_i \mid X_i)$表示给定 X_i 下 Y_i 的条件方差。

证明：由于 ε_i 与 $\mathrm{E}[Y_i \mid X_i]$ 不相关，所以条件期望函数分解性质意味着 Y_i 的方差等于条件期望函数的方差加上残差项 $\varepsilon_i \equiv Y_i - \mathrm{E}[Y_i \mid X_i]$ 的方差。由于 $\varepsilon_i \equiv Y_i - \mathrm{E}[Y_i \mid X_i]$，所以 $\mathrm{E}[\varepsilon_i^2 \mid X_i] = V[Y_i \mid X_i]$，残差项 ε_i 的方差是：

$$\mathrm{E}[\varepsilon_i^2] = \mathrm{E}[\mathrm{E}[\varepsilon_i^2 \mid X_i]] = \mathrm{E}[V[Y_i \mid X_i]]$$

这里提到的条件期望函数的两个性质以及方差分析定理听上去似乎很熟悉。比如也许你曾在回归的输出结果中见到过方差分析表。在对不平等的研究中方差分析也扮演了重要角色，劳动经济学家使用方差分析将收入分布的变化分解成两部分，一部分由工人个体特点的变化来解释，另一部分留作工人个体特点不能解释的变化(比如，Auto，Katz and Kearney，2005)。令我们颇感陌生的或许是条件期望函数的两个性质与方差分析定理不仅在总体中成立，在样本中也同样成立，而且无需将条件期望函数为线性作为这些性质和定理发挥作用的前提。事实上，线性回归作为经验研究的一种工具，其有效性也不依赖于线性性假设。

3.1.2 线性回归与条件期望函数

那么，你打算做的回归是什么？在我们工作的圈子里几乎每天都能听到这种问题或者与之类似的提问。因为回归与条件期望函数密切相关，所以回归估计几乎给所有的经验研究提供了一个有价值的基本点，而且条件期望函数还为经验研究中变量相互间的关系提供了一个天然的概括。我们至少可以从三个角度来阐述回归方程——通过最小化均方误(expected square errors)得到的最优拟合曲线——与条件期望函数之间的关系。为了准确地解释这些联系，我们首先应该精确地将心中所想的回归函数描述出来。本节主要关注的是总体回归方程的系数所组成的向量，我们将这个向量定义为总体最小二乘问题的解。在这里我们先不用担心因果问题。记 $k \times 1$ 维的回归系数向量 β 可以定义如下：

$$\beta = \arg\min_b \mathrm{E}[(Y_i - X_i'b)^2] \tag{3.1.2}$$

利用一阶条件后可得：

$$\mathrm{E}[X_i(Y_i - X_i'b)] = 0$$

于是最小二乘估计的解可以写作 $\beta = \mathrm{E}[X_iX_i']^{-1}\mathrm{E}[X_iY_i]$。注意到 $\mathrm{E}[X_i(Y_i - X_i'\beta)] = 0$。换句话说就是如果我们定义 $Y_i - X_i'\beta = e_i$ 为总体残差，那么这个残差与回归元 X_i 不相关。需要强调的是这个残差项本身并无价值。它的存在和价值

都是因为有了 β。我们在第 3.2 节讨论具有因果性质的回归时再来考虑这个问题。

在回归方程中只存在单一回归元 x_i 和一个常数的双变量回归模型中，斜率的系数就是 $\beta_1 = \frac{\text{cov}(Y_i, X_i)}{V(X_i)}$，截距项系数是 $\alpha = \text{E}[Y_i] - \beta_1 \text{E}[X_i]$。在多变量情形下，由于存在多个非常数的回归元，第 k 个回归元的斜率系数就是：

解构回归(regression anatomy)

$$\beta_k = \frac{\text{cov}(Y_i, \tilde{x}_{ki})}{V(\tilde{x}_{ki})} \tag{3.1.3}$$

其中，$\tilde{x}_{ki}$ 是将 x_{ki} 关于其他回归元回归后得到的残差项。

换个角度解释，$\text{E}[X_i X_i']^{-1}\text{E}[X_i Y_i]$ 是一个 $k \times 1$ 维的向量，其第 k 个元素是 $\frac{\text{cov}(Y_i, \tilde{x}_{ki})}{V(\tilde{x}_{ki})}$。由于这个重要公式揭示出的内容远比向量表达式 $\text{E}[X_i X_i']^{-1}\text{E}[X_i Y_i]$ 多，所以我们说等式(3.1.3)解构了多元回归系数。它告诉我们多元回归中每个回归元的系数都是该回归元在剔除其他回归元对自己的影响后与 Y_i 进行简单二元回归得到的斜率。

为了证明解构回归中的公式，将

$$Y_i = \alpha + \beta_1 x_{1i} + \cdots + \beta_k x_{ki} + \cdots + \beta_K x_{Ki} + e_i$$

代入等式(3.1.3)的分子中。因为 $\tilde{x}_{ki}$ 是所有回归元的线性组合，所以它与 e_i 不相关。而且，既然 $\tilde{x}_{ki}$ 是 x_{ki} 关于其他回归元回归后的残差，所以它也和其他回归元不相关。最后，因为同样的理由，$\tilde{x}_{ki}$ 和 x_{ki} 之间的协方差就是 $\tilde{x}_{ki}$ 自己的方差。我们于是有 $\text{cov}(Y_i, \tilde{x}_{ki}) = \beta_k V(\tilde{x}_{ki})$ ①。

解构回归中的系数公式可能与你在回归或者统计课上所学到的知识有些相似，但实际上还是有所不同：本节定义的回归系数不是估计值；确切地说，它们是解释变量和被解释变量的联合分布中的非随机特点。如果你有完整的总体(或者了解产生数据的随机过程)，那么你就可以知道这个联合分布。但你很可能没有这样的信息。不过，在为估计总体参数倍感忧心之前，思考一下总体参数到底意味着什么乃是经验研究的一个良好习惯。

---

① 解构回归的公式通常归功于 Frisch 和 Waugh(1933)。你也可以通过下面的方法得到该公式：

$$\beta_k = \frac{\text{cov}(\tilde{Y}_{ki}, \tilde{x}_{ki})}{V(\tilde{x}_{ki})}$$

其中，$\tilde{Y}_{ki}$ 是 Y_i 关于除 x_{ki} 之外的回归元进行回归后得到的残差项。因为从 $\tilde{Y}_{ki}$ 去掉的拟合值与 $\tilde{x}_{ki}$ 不相关，所以这样做总是可以的。通常描绘 $\tilde{Y}_{ki}$ 关于 $\tilde{x}_{ki}$ 的散点图也很有用，因为即使该散点图是二维的，在图中的最小二乘拟合曲线的斜率也还是 β_k。需要注意的是，只排除其他回归元对 Y_i 的影响，但是不排除这些回归元对 x_{ki} 的影响，结果可能是错误的。也即：

$$\frac{\text{cov}(\tilde{Y}_{ki}, x_{ki})}{V(x_{ki})} = \left[\frac{\text{cov}(\tilde{Y}_{ki}, \tilde{x}_{ki})}{V(\tilde{x}_{ki})}\right]\left[\frac{V(\tilde{x}_{ki})}{V(x_{ki})}\right] \neq \beta_k$$

除非 x_{ki} 与其他回归元无关，否则上面这个等式不会等于 β_k。

接下来我们讨论关注总体回归的参数向量的三个理由。这些理由可以总结成一句话：如果你对条件期望函数感兴趣，那么你就应该关注总体回归的参数向量。

定理 3.1.4：线性条件期望函数定理(linear CEF theorem)(回归的理由 1)。

假设条件期望函数是线性的，那么总体回归方程就应该是这个线性函数。

证明：假设 $\mathrm{E}[Y_i \mid X_i] = X_i'\beta^*$，其中 β^* 是某个 $k\times 1$ 维的向量。回忆条件期望函数分解性质可知 $\mathrm{E}[X_i(Y_i - \mathrm{E}[Y_i \mid X_i])] = 0$。将 $\mathrm{E}[Y_i \mid X_i] = X_i\beta^*$ 代入后可得 $\beta^* = \mathrm{E}[X_iX_i']^{-1}\mathrm{E}[X_iY_i] = \beta$。

线性条件期望函数定理带来这样一个问题：在什么条件下条件期望函数是线性的。经典的解决方案是加上联合正态分布的假设，也就是说假设向量 $(Y_i,\ X_i')'$ 是满足多元联合正态分布的。这是回归之父 Galton(1886)给出的条件。当时他对具有正态分布的一些特征——比如身高和智商——在不同世代之间的关联很感兴趣。由于回归元和被解释变量常常是离散的，而正态分布是连续分布，所以从经验研究角度来看这种正态性假设显然具有相当的局限性。另外一个线性化方案是使用饱和回归模型(saturated regression)。正如将在 3.1.4 节讲到的那样，在饱和回归模型中回归元向量的每一个取值都对应一个不同的参数。以存在两个虚拟变量作为协变量的饱和回归模型为例，这个模型包含两个协变量(它们的系数被称为主效应，main effect)和协变量的乘积(被称为交互项)。这个模型的总体回归方程一定是线性的，我们在 3.1.4 节再回头讨论这一点。

当线性条件期望函数定理不适用时，还存在两个原因使我们继续关注回归系数。

定理 3.1.5：最优线性估计量定理(The Best Linear Predictor Theorem)(回归的理由 Ⅱ)。

在最小均方误的意义下，给定 X_i，函数 $X_i'\beta$ 是对 Y_i 的最优线性估计量。

证明：$\beta = \mathrm{E}[X_iX_i']^{-1}\mathrm{E}[X_iY_i]$ 是总体的最小方差问题(3.1.2)的解。

换句话说，定理 3.1.2 告诉我们条件期望函数 $\mathrm{E}[Y_i \mid X_i]$ 是在给定 X_i 下，在所有关于 X_i 的函数中能够最好地(在最小均方误意义下)预测 Y_i 的函数；类似的，定理 3.1.5 告诉我们在所有的关于 X_i 的线性函数中，总体回归函数在最小均方误意义下能够最好地预测 Y_i。

定理 3.1.6：条件期望函数的回归定理(The Regression CEF Theorem)(回归的理由 Ⅲ)。

函数 $X_i'\beta$ 在最小均方误意义下为我们提供了对 $\mathrm{E}[Y_i \mid X_i]$ 的最优线性近似，也就是说：

$$\beta = \arg\min_b \mathrm{E}\{(\mathrm{E}[Y_i \mid X_i] - X_i'b)^2\} \tag{3.1.4}$$

证明：观察到 β 是等式(3.1.2)的解。记：

$$\begin{aligned}(Y_i - X_i'b)^2 &= \{(Y_i - \mathrm{E}[Y_i \mid X_i]) + (\mathrm{E}[Y_i \mid X_i] - X_i'b)\}^2\\ &= (Y_i - \mathrm{E}[Y_i \mid X_i])^2 + (\mathrm{E}[Y_i \mid X_i] - X_i'b)^2\\ &\quad + 2(Y_i - \mathrm{E}[Y_i \mid X_i])(\mathrm{E}[Y_i \mid X_i] - X_i'b)\end{aligned}$$

第一项不包括 b，根据条件期望函数分解性质中的(ii)，最后一项的期望是零。于是条件期望函数的线性近似问题和总体最小二乘问题(3.1.2)是相同的。

定理 3.1.5 和 3.1.6 为我们提供了两种额外的方式来看待回归。不存在约束条件下，条件期望函数是对被解释变量的最好估计，回归以相同的方式为我们提供了对被解释变量最好的线性估计。另一方面，如果我们考虑近似 $\mathrm{E}[Y_i \mid X_i]$ 而不是预测 Y_i，条件期望函数的回归定理告诉我们：即使条件期望函数不是线性的，回归也为我们提供了一个最好的线性近似。

我们最喜欢用条件期望函数的回归定理来说明使用回归的原因。用回归来近似条件期望函数的说法与我们对经验研究的看法相一致：我们将经验研究看作在无需精确计算变量间关系的同时捕捉到变量间统计关系之实质的一种努力。显然线性条件期望函数定理只适用于特例。虽然最优线性估计量定理则更加一般化，但是看上去这也鼓励了那种认为经验研究过于简陋的论调。我们并非真的对预测个体值 Y_i 感兴趣，我们真正关心的乃是 Y_i 的分布。

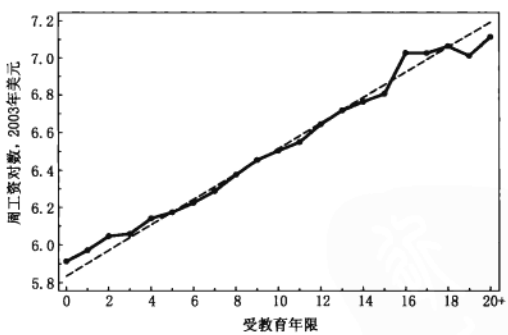


图 3.2　回归线贯穿了给定受教育年限下周平均工资的条件期望函数
(点＝条件期望函数；折线＝回归线)

图 3.2 以图 3.1 中绘出的那条条件期望函数曲线为例说明了条件期望函数的近似性质。回归曲线对一条非线性并且有时还不很平坦的条件期望函数进行了拟合，似乎我们是在估计模型 $\mathrm{E}[Y_i \mid X_i]$，而不是在估计 Y_i。事实上，事情确实是这样的。条件期望函数的回归定理告诉我们可以通过将 $\mathrm{E}[Y_i \mid X_i]$ 当作被解释变量进行回归来得到回归系数，而不需要将 Y_i 当作被解释变量。为了看清楚这一点，我们假定 X_i 是一个离散的随机变量，其概率质量函数为 $g_x(u)$。那么，

$$\mathrm{E}\{(\mathrm{E}[Y_i \mid X_i] - X_i'b)^2\} = \sum_u (\mathrm{E}[Y_i \mid X_i = u] - u'b)^2 g_x(u)$$

这意味着我们可以对 $\mathrm{E}[Y_i \mid X_i = u]$ 关于 u 进行加权最小二乘(WLS)来得到 β，其中 u 取遍 X_i 的可能取值。一个更为简单的方法是在关于 β 的公式中使用迭

代期望律：

$$\beta = \mathrm{E}[X_i X_i']^{-1}\mathrm{E}[X_i Y_i] = \mathrm{E}[X_i X_i']^{-1}\mathrm{E}[X_i \mathrm{E}(Y_i \mid X_i)] \qquad (3.1.5)$$

当我们从事的研究项目无法使用微观数据进行分析时，针对条件期望函数或者分组数据的回归公式就有了实际意义。例如，Angrist(1998)就使用分组数据研究了人们自愿参军对其后来收入的影响。该项目中使用的一个估计策略是用公民收入关于标志老兵身份的虚拟变量进行回归，回归中还包括了个人特征以及军队用以审查士兵的一些变量。他使用的收入数据来自美国社会保障系统，但是社会保障系统中个体收入的记录不能向公众开放，所以可能无从获得(也就是说无法得到 Y_i)。为此，Angrist 使用那些以种族、性别、考试成绩、教育以及老兵身份为协变量计算出平均收入(相当于计算 $\mathrm{E}[Y_i \mid X_i]$)，进而完成了这项研究。

为了阐明使用分组数据进行回归的方法，我们使用 21 个条件均值来估计工资方程中教育水平的系数，以及给定教育水平下来自样本的收入的条件期望函数。图 3.3 是用 Stata 产生的结果，它表明，以样本中各教育水平下的个体数目为权重，对该分组数据作回归得到的系数，与那些使用微观数据中成百上千观察值得到的

A - Individual-level data

```
. regress earnings school, robust

      Source |       SS       df       MS              Number of obs =  409435
-------------+------------------------------           F(  1,409433) =49118.25
       Model | 22631.4793      1  22631.4793            Prob > F      =  0.0000
    Residual |  188648.31 409433  .460755019           R-squared     =  0.1071
-------------+------------------------------           Adj R-squared =  0.1071
       Total | 211279.789 409434   .51602893           Root MSE      =  .67879

             |               Robust                        Old Fashioned
    earnings |      Coef.   Std. Err.      t                Std. Err.      t
-------------+----------------------------------------------------------------
      school |   .0674387   .0003447   195.63              .0003043   221.63
       const.|   5.835761   .0045507  1282.39              .0040043  1457.38
------------------------------------------------------------------------------
```

B - Means by years of schooling

```
. regress average_earnings school [aweight=count], robust
(sum of wgt is   4.0944e+05)

      Source |       SS       df       MS              Number of obs =      21
-------------+------------------------------           F(  1,     19) =  540.31
       Model | 1.16077332      1 1.16077332             Prob > F      =  0.0000
    Residual | .040818796     19 .002148358            R-squared     =  0.9660
-------------+------------------------------           Adj R-squared =  0.9642
       Total | 1.20159212     20 .060079606            Root MSE      =  .04635

             |               Robust                        Old Fashioned
average_earn |      Coef.   Std. Err.      t                Std. Err.      t
-------------+----------------------------------------------------------------
      school |   .0674387   .0040352    16.71             .0029013    23.24
       const.|   5.835761   .0399452   146.09             .0381792   152.85
------------------------------------------------------------------------------
```

资料来源：1980 年人口调查——公用微观样本数据集(IPUMS)，5%的样本。该样本由年龄在 40—49 岁之间的白人男性构成。稳健标准误是异方差一致的。表中 A 部分使用了来自个体层面的微观数据，B 部分使用了按照受教育年限进行平均的收入水平。

图 3.3　在教育回报的研究中分别使用微观数据和分组数据后得到的估计值(来自 Stata 的回归结果)

结果完全一致。不过，值得注意的是分组回归中得到的标准误没能度量使用重复抽样的微观数据得到的斜率估计值的渐进样本方差。为了估计这个值，你需要估计 $Y_i - X_i'\beta$ 的方差。这个方差的计算需要微观数据，特别是 $W_i \equiv [Y_i \quad X_i']'$ 的二阶矩，我们在下一节仔细讨论这个问题。

3.1.3 渐进最小二乘推断

在实际中，我们往往不知道条件期望函数是什么，也不知道总体回归向量。因此我们通过使用来自总体的某个样本来对这些数量进行推断。统计推断是大部分传统的计量经济学教科书所关注的内容。尽管在任何一本计量经济学教科书中都会涉及这些内容，但我们还是不愿意将推断部分完全省略。对基本的渐进理论的回顾使我们得以强调一个重要事实：统计推断与如何对一个特定回归结果进行解释是不同的。不论回归参数表示的是什么，它都有一个样本分布，我们可以很容易地描述这个分布并运用该分布进行统计推断①。

我们对重复抽样中总体回归向量

$$\beta = \mathrm{E}[X_i X_i']^{-1}\mathrm{E}[X_i Y_i]$$

的样本分布感兴趣。假设向量 $W_i \equiv (Y_i \quad X_i')'$ 在大小为 N 的样本中独立同分布。于是对总体一阶矩 $\mathrm{E}[W_i]$ 的很自然的估计便是 $\frac{1}{N}\sum_{i=1}^{N} W_i$。由大数定理，随着样本数量增大，这个样本矩所成的向量会无限接近于相应的总体矩所成的向量。我们也会以相同的方式考虑 W_i 的高阶矩，比如二阶矩所成的矩阵 $\mathrm{E}[W_i W_i']$ 的样本估计值就是 $\frac{1}{N}\sum_i^N W_i W_i'$。依照这个原则，对 β 进行矩估计就是将所有的期望算子换成连加符号。这个逻辑为我们带来了对 β 的最小二乘(ordinary least square，简称为 OLS)估计量：

$$\hat{\beta} = [\sum_i X_i X_i']^{-1}\sum_i X_i Y_i$$

虽然我们用矩估计得到了 $\hat{\beta}$，但之所以将其称为 β 的最小二乘估计量，乃是因为当我们在样本中(而不是总体中)考虑 3.1.2 节一开始所讨论的最小二乘问题时，$\hat{\beta}$ 正好是这个问题的解②。

$\hat{\beta}$ 的渐进样本分布完全依赖于我们对被估计量的定义(也就是说，我们试图估

① 在这部分对最小二乘推断的渐进性质的讨论大体上是对 Chamberlain(1984)中相应内容的一个缩写。渐进理论里的重要陷阱和问题会在最后一章提到。

② 因为矩阵可以很紧凑，所以计量经济学家喜欢使用这种记号。有些时候(但不是经常)我们也因此而使用它。假设 X 是矩阵，其第 i 行记为 X_i'，y 也是一个向量，其 i 位置上的元素是 Y_i，于是样本矩所成的矩阵 $\frac{1}{N}\sum_i X_i X_i'$ 用矩阵乘积表示就是 XX'/N，样本矩向量 $\frac{1}{N}X_i y_i$ 用向量来表示就是 $X'y/N$。于是我们可以记 $\hat{\beta} = (X'X)^{-1}X'y$，这是被广泛使用的矩阵记法。

计的 β 所具有的性质)和对数据来自于随机样本的假设。在得到这个分布之前，有必要先总结一下我们需要用到一般性的渐进分布定理。这些基本理论基本上可以用语言来表述。基于此，我们假设读者已经熟悉统计定理的核心术语和概念——矩、数学期望、概率极限和渐进分布。对这些术语和下面给出定理的正式的数学处理见 Knight(2000)。

大数定理(The Law of Large Numbers)

样本矩依概率收敛于相应的总体矩。也即是说，只要样本规模足够大，样本均值趋向于总体均值的概率可以足够得高。

中心极限定理(The Central Limit Theorem)

样本矩是渐进正态分布的(在减去总体矩，除以样本规模的平方根后)。渐进协方差矩阵由相应随机变量的方差给出。换句话说，样本规模足够大，经过合理标准化了的样本矩近似符合正态分布。

斯拉茨基定理(Slutsky's Theorem)

(1) 考虑两个随机变量之和，其中一个随机变量依分布收敛(也就是说这个随机变量有一个渐进分布)，另一个依概率收敛于一个常数。那么将依概率收敛的随机变量换成它收敛到的常数后得到的渐进分布与两个随机变量之和的渐进分布相同。正式地说就是，记 a_N 是具有渐进分布的一个统计量，记 b_N 是具有概率极限 b 的一个随机变量。那么 $a_N + b_N$ 和 $a_N + b$ 具有相同的渐进分布。

(2) 考虑两个随机变量的乘积，其中一个依分布收敛，另一个依概率收敛于一个常数；那么将依概率收敛的随机变量换成它收敛到的常数后得到的渐进分布与两个随机变量之积的渐进分布相同。正式地说就是，记 a_N 是具有渐进分布的一个统计量，记 b_N 是具有概率极限 b 的一个随机变量，那么 $a_N b_N$ 和 $a_N b$ 有相同的渐进分布。

连续映射定理(The Continuous Mapping Theorem)

概率极限算子可以穿过连续函数。比如，关于样本矩的任何连续函数之概率极限都等于在相应的总体矩处该连续函数的值。正式地说就是，函数 $h(b_N)$ 的概率极限是 $h(b)$，其中 $\operatorname{plim} b_N = b$，函数 $h(\cdot)$ 在 b 处连续。

德尔塔法(The Delta Method)

考虑一个渐进正态分布的向量值随机变量。这个随机变量的连续可微标量函数也是渐进正态分布的，其协方差矩阵由一个二次型[①]给出。这个二次型中，居于中间的那个矩阵是该随机变量的协方差矩阵，居于两边的两个向量是这个连续函数在该随机变量的概率极限处的梯度。正式地说就是，如果随机变量(这个随机变量可以是向量)b_N 是渐进正态分布的，其协方差矩阵是 Ω[②]，$\operatorname{plim} b_N = b$，$h(\cdot)$ 是

① 二次型是一个将矩阵作为权重的加权平方和。假设 v 是个 $N \times 1$ 的向量，M 是个 $N \times N$ 的矩阵。那么 v 的二次型就是 $v'Mv$。如果 M 是对角阵，其对角线上的元素为 m_i，那么 $v'Mv = \sum_i m_i v_i^2$。

② 运用斯拉茨基定理和连续映射定理得到德尔塔法的方法请见 Knight(2000:120—121)。我们所指的"$h(b_N)$ 的渐进分布"实际上指的是 $\sqrt{N}(h(b_N) - h(b))$ 的渐进分布。

在 b 处连续可微的函数，具有梯度 $\nabla h(b)$，那么 $h(b_N)$ 的渐进分布是正态的，其协方差矩阵为 $\nabla h(b)'\Omega\nabla h(b)$。

运用这些结论，我们可以通过两种方法得到 $\hat{\beta}$ 的渐进分布。从概念上看简明直接但是可能缺乏一些精致性的办法就是使用德尔塔法：$\hat{\beta}$ 是样本矩的函数，因此是渐进正态分布的。剩下的事就是从这个方程中求得渐进分布的协方差矩阵。(注意到从连续映射定理可以立刻得到 $\hat{\beta}$ 的一致性)①。一个更加简单并更富指导意义的方法是使用斯拉茨基定理和中心极限定理。首先注意到：

$$Y_i = X_i'\beta + [Y_i - X_i'\beta] \equiv X_i'\beta + e_i \tag{3.1.6}$$

如前，将其中的残差项 e_i 定义为被解释变量和总体回归方程之间的差。换句话说，由 $\beta = \mathrm{E}[X_iX_i']^{-1}$ 和 $e_i = Y_i - X_i'\beta$ 可以推导出 $\mathrm{E}[X_ie_i] = 0$，但 $\mathrm{E}[X_ie_i] = 0$ 并不是对变量间潜在的经济学关系做出的假设②。

用等式(3.1.6)替换 $\hat{\beta}$ 中的 Y_i 后我们有：

$$\hat{\beta} = \beta + [\sum X_iX_i']^{-1}\sum X_ie_i$$

于是 $\hat{\beta}$ 的渐进分布就是 $\sqrt{N}(\hat{\beta}-\beta) = N[\sum X_iX_i']^{-1}\frac{1}{\sqrt{N}}\sum X_ie_i$ 的渐进分布。由斯拉茨基定理可知，它与 $\mathrm{E}[X_iX_i']^{-1}\frac{1}{\sqrt{N}}\sum X_ie_i$ 有相同的渐进分布。由于 $\mathrm{E}[X_ie_i] = 0$，所以 $\frac{1}{\sqrt{N}}\sum X_ie_i$ 是由 $\sqrt{N}$ 调整过的 X_ie_i 的样本中心矩。由中心极限定理可知，它是渐进正态分布的，其均值为零，协方差矩阵是 $\mathrm{E}[X_iX_i'e_i^2]$。因此，$\hat{\beta}$ 是具有概率极限为 β，协方差矩阵为

$$\mathrm{E}[X_iX_i']^{-1}\mathrm{E}[X_iX_i'e_i^2]\mathrm{E}[X_iX_i']^{-1} \tag{3.1.7}$$

的渐进正态分布。

用来构造 t 统计量时所用的标准误就是等式(3.1.7)中对角元素的平方根。在实际中，用连加算子代替期望算子，用估计出的残差 $\hat{e}_i = Y_i - X_i'\hat{\beta}$ 来构造四阶矩阵 $\sum[X_iX_i'\hat{e}_i^2]/N$③。

通过这种方式计算出的渐进标准误被称为异方差修正标准误(heteroscedasticity-consistent standard errors)或者怀特(White，1980a)标准误，为了纪念艾克(Eicker，1967)对这个问题的开创性研究，该标准误也可被称为艾克—怀特(Eicker-White)标准误。在某些场合，该标准误还被称为稳健(robust)的标准误(比如在

① 说一个估计值是一致性，指的是这个估计值依概率收敛于目标参数。

② 以这种方式定义的残差项未必中值独立于 X_i，如果要求中值独立，那么我们需要线性条件期望函数。

③ 原书此处公式为 $\sum[X_iX_i\hat{e}_i^2]/N$，疑为作者笔误。——译者注

Stata 软件中）。之所以称这些标准误是稳健的，是因为当样本足够大时，它们在对数据和模型给定最少的假设下为我们提供了精确的假设检验和置信区间估计。特别的，除了保证类似于中心极限定理等基本的统计结论需要成立外，我们对分布没有做出任何限制性假设。但是，稳健的标准误并不是我们使用的计量经济学软件给出的默认的标准误。这些软件默认的标准误是在同方差假设下得到的，也就是 $E[e_i^2 \mid X_i] = \sigma^2$。给定这个假设，根据迭代期望律，我们有：

$$E[X_i X_i' e_i^2] = E(X_i X_i' E[e_i^2 \mid X_i]) = \sigma^2 E[X_i X_i']$$

于是 $\hat{\beta}$ 的渐进协方差矩阵就简化为：

$$\begin{aligned} & E[X_i X_i']^{-1} E[X_i X_i' e_i^2] E[X_i X_i']^{-1} \\ = & E[X_i X_i']^{-1} \sigma^2 E[X_i X_i'] E[X_i X_i']^{-1} \\ = & \sigma^2 E[X_i X_i']^{-1} \end{aligned} \tag{3.1.8}$$

除非你要求它报告稳健的标准误，否则在统计软件 SAS 或者 Stata 中报告的将是等式(3.1.8)中的对角线元素的值。

我们将回归看作对条件期望函数的近似，这一观点使得异方差性变得很自然。如果条件期望函数不是线性的，而你用一个线性函数去近似它，那么拟合质量就会随着 X_i 的变化而变化。因此，平均而言，在 X_i 的某些值上，如果拟合效果不好，那么残差就会大。即使我们准备假设给定 X_i 时 Y_i 的条件方差是固定的，条件期望函数是非线性的这一事实也指出 $E[(Y_i - X_i'\beta)^2 \mid X_i]$ 将会随着 X_i 的变化而变化。为了看清这一点，注意到：

$$\begin{aligned} & E[(Y_i - X_i'\beta)^2 \mid X_i] \\ = & E\{[(Y_i - E[Y_i \mid X_i]) + (E[Y_i \mid X_i] - X_i'\beta)]^2 \mid X_i\} \\ = & V[Y_i \mid X_i] + (E[Y_i \mid X_i] - X_i'\beta)^2 \end{aligned} \tag{3.1.9}$$

因此，即使 $V(Y_i \mid X_i)$ 是常数，残差项的方差也会随着回归线和条件期望函数之间差距上升而增大，White(1980b)已经注意到了这个事实①。

同样的，值得注意的是条件期望函数为线性函数可以使同方差假设成立，但这并非同方差的充分条件。在这一知识点上我们最喜欢举出的例子是线性概率模型(LPM)。任何被解释变量只取 0—1 两个值的回归都可以叫做线性概率模型，比如标志劳动力是否参与工作的某个虚拟变量成为被解释变量时，该模型就可称为线性概率模型。假设回归模型是饱和模型，于是条件期望函数关于回归元是线性的。因为条件期望函数是线性的，所以残差项的方差就只有一项 $V[Y_i \mid X_i]$。但是由于被解释变量是一个伯努里(Bernoulli)实验结果，其条件方差是 $P[Y_i = 1 \mid X_i](1 - P[Y_i = 1 \mid X_i])$，仍然是 X_i 的函数。我们可以总结指出：除非回归元是个常数，否

① 由于 $Y_i - E[Y_i \mid X_i]$ 与 X_i 均值独立，所以等式(3.1.9)里平方展开项中交叉乘积的那部分的期望为零。

则线性回归模型的残差项一定是异方差的。

虽然上面提到的这些原理都是经验研究中会遇到的，但是异方差性实际上不会带来太大的影响。在图 3.3 中绘出的用微观数据对教育回报进行回归的例子中，稳健的标准误是0.000 347 7，而传统的标准误是 0.000 304 3，可见并未减少太多。在运用分组数据进行的回归中，如果分组样本的大小不同，那么我们得到的标准误必然是异方差的，所以这个值的变化会更大，于是得到的稳健的标准误是0.004，传统的标准误是 0.002 9。根据我们的经验，这些标准误之间的差别具有典型意义。如果异方差性变化很大，比如标准误上升了 30%或者有任何显著的下降，那么你应该注意是不是存在可能的程序错误或者是其他问题。如果稳健标准误低于传统意义上的标准误，可能标志着在稳健性计算中存在有限样本偏误。

最后，我们对你可能在其他教材中见过的讲解推断问题的方式做一个简要说明。在传统的计量经济学中，对推断问题的处理从更强的假设开始。传统的方式有时候被叫做经典正态回归模型，它假定回归元固定（非随机）、条件期望函数是线性的、残差符合正态分布以及同方差（比如，见 Goldberger，1991）。这些更强的假设为我们带来两个结果：(1)最小二乘估计量的无偏性，以及(2)最小二乘估计量的样本方差公式，这个公式在小样本和大样本中都成立。最小二乘估计量的无偏性意味着 $E[\hat{\beta}]=\beta$，这个性质在任何大小的样本中都成立，而且强于一致性（一致性只能告诉我们在大样本中 $\hat{\beta}$ 可以接近 β）。可以很容易地回答什么时候以及为什么我们可以得到无偏估计量。一般而言，

$$E[\hat{\beta}]=\beta+E\{[\sum X_i X_i']^{-1}\sum X_i e_i\}$$

如果回归元是非随机（在重复取样中固定）的，那么期望算子就可以进入大括号里，由于 $E[e_i]=0$，所以估计量是无偏的。另外，考虑回归元是随机的，如果 $E[e_i \mid X_i]=0$，那么用迭代期望律可得无偏性。但是只有当条件期望函数是线性时，这个结论才成立，在我们更一般化的“不做过多假设的回归”框架中，该结论将不再成立。

在经典假设下得到的方差公式与在同方差假定下的大样本方差公式相同，不过——如果强的经典假设都满足——那么该公式在任意大小的样本中都将成立。我们之所以选择从渐进的角度讨论推断，是因为现代经验研究工作主要建立在大样本理论之上，而这又是得到稳健的方差公式所需要的。这样做的好处就是在更弱的假设下得到有效的推断，特别重要的是，这样做使我们在一个有意义的框架中不那么学究式地考虑回归模型。当然，大样本的方法也不是没有其问题，我们会在第 8 章讨论推断时再回来说明，并在第 4 章讲解工具变量（instrument variable）时进行必要的讨论。

3.1.4 饱和模型、主效应和其他的有关回归的话

我们往往用饱和（saturated）和主效应（main effect）这样的术语来讨论回归模

型。这些术语都起源于实验者使用回归模型讨论离散型的处理变量(treatment-type variable)的传统。不过这个术语如今已在包括应用计量经济学在内的很多领域得到广泛应用。本节为那些还不甚熟悉这些术语的研究者提供一个简要回顾。

饱和回归模型指的是具有离散解释变量的回归模型，对解释变量的每一个可能取值，该模型都存在一个参数与之相对应。比如，我们只研究单一解释变量——一个用来表示工人是否大学毕业的解释变量，那么当模型包含一个用以表示是否大学毕业的虚拟变量和一个常数项时，我们说这个模型是饱和的。当回归元可取多个值时，我们同样可以将模型调整为饱和。比如考虑 $S_i = 0, 1, 2, \cdots, \tau$。于是针对 S_i 的饱和模型就是：

$$Y_i = \alpha + \beta_1 d_{1i} + \beta_2 d_{2i} + \cdots + \beta_\tau d_{\tau i} + \varepsilon_i$$

其中，$d_{ji} = 1[S_i = j]$ 是一个虚拟变量，表示个体是否接受了 j 年的教育，于是 β_j 便是受教育水平为 j 年级所带来的效应①。注意到：

$$\beta_j = \mathrm{E}[Y_i \mid S_i = j] - \mathrm{E}[Y_i \mid S_i = 0]$$

其中，$\alpha = \mathrm{E}[Y_i \mid S_i = 0]$。实际上我们可以将 S_i 的任何值取作参照组；一旦 $\mathrm{E}[Y_i \mid S_i = j]$ 中 j 的每一个可能取值都有一个参数与之对应时，就说这个模型是饱和模型。饱和回归模型能够完美地拟合条件期望函数，因为用来构造饱和模型的虚拟变量所构成的条件期望函数本来就是线性的。这是线性条件期望函数定理的一个重要特例。

如果模型中存在两个解释变量——比如一个是用来表示是否大学毕业的虚拟变量，另一个是用来表示年龄的虚拟变量——那么通过包含这两个虚拟变量以及它们的乘积和常数项，该模型达到饱和。虚拟变量的系数就叫做主效应(main effect)，两个虚拟变量的乘积叫做交互项。不过这并不是让参数饱和化的唯一方式；任何示性变量(虚拟变量)组成的集合，如果它们能识别出所有协变量的取值，那么它们就能构成一个饱和模型。比如，我们还可对刚才提到的以两个虚拟变量作为解释变量的模型构造另外一个饱和模型，在这个模型中包含标志男性大学毕业、男性非大学毕业、女性大学毕业和女性非大学毕业四个虚拟变量，但不包括截距项。

通过一些更具体的记号，我们可以将这个问题看得更加清楚。记 x_{1i} 表示大学毕业，x_{2i} 表示女性。给定 x_{1i} 和 x_{2i}，条件期望函数可以取四个值：

$$\mathrm{E}[Y_i \mid x_{1i} = 0, x_{2i} = 0]$$
$$\mathrm{E}[Y_i \mid x_{1i} = 1, x_{2i} = 0]$$
$$\mathrm{E}[Y_i \mid x_{1i} = 0, x_{2i} = 1]$$
$$\mathrm{E}[Y_i \mid x_{1i} = 1, x_{2i} = 1]$$

① 我们用记号 $1[S_i = j]$ 表示示性函数，在这个例子中示性函数就是当 $S_i = j$ 时第 j 个虚拟变量变为 1。

我们可以用下面的方式来记每个不同的值：

$$\begin{aligned}
&\mathrm{E}[Y_i \mid x_{1i}=0,\ x_{2i}=0]=\alpha\\
&\mathrm{E}[Y_i \mid x_{1i}=1,\ x_{2i}=0]=\alpha+\beta_1\\
&\mathrm{E}[Y_i \mid x_{1i}=0,\ x_{2i}=1]=\alpha+\gamma\\
&\mathrm{E}[Y_i \mid x_{1i}=1,\ x_{2i}=1]=\alpha+\beta_1+\gamma+\delta_1
\end{aligned}$$

由于这里存在四个希腊字母，而条件期望函数取四个值，所以上面的参数化过程并没有对条件期望函数加上什么限制。上面的希腊字母还可以表示为：

$$\mathrm{E}[Y_i \mid x_{1i},\ x_{2i}]=\alpha+\beta_1 x_{1i}+\gamma x_{2i}+\delta_1(x_{1i}x_{2i})$$

这就是有两个主效应和一个交互项的参数化过程①。这个饱和回归模型就变为：

$$Y_i=\alpha+\beta_1 x_{1i}+\gamma x_{2i}+\delta_1(x_{1i}x_{2i})+\epsilon_i$$

我们还可以将作为多值变量的教育变量和性别变量结合起来构造一个饱和模型，这个饱和模型有 τ 个教育的主效应，一个性别的主效应和 τ 于性别—教育的交互项：

$$Y_i=\alpha+\sum_{j=1}^{\tau}\beta_j d_{ji}+\gamma x_{2i}+\sum_{j=1}^{\tau}\delta_j(d_{ji}x_{2i})+\epsilon_i \tag{3.1.10}$$

交互项的系数 δ_j 告诉我们教育对不同性别的不同影响。在这个例子中条件期望函数取 $2(\tau+1)$ 个值，通过回归也可以得到同样多的参数。

注意到随着模型包含的变量越来越多，饱和模型对建模方法提出了相当的限制。由于饱和模型可以完美地逼近条件期望函数，所以我们很自然地从饱和模型开始讨论。但是从另一方面讲，饱和模型产生很多交互项，我们对其中的大部分可能都不感兴趣或者说很难对交互项的系数做出准确的估计。因此你可能会很明智地选择将部分或者全部的交互项省略。当不包含交互项时，方程(3.1.10)表现为只有主效应相加的模型。如果大学教育对男性和女性收入回报的影响大致相同，那么这个模型就可以很好地逼近条件期望函数。经过 3.3.1 节的讨论后我们会知道，在表现为只有主效应相加的所有例子中，教育水平前面的系数表达的都是两种性别的个体的教育回报的加权平均值。从另一个方面来讲，估计一个只包含交互项但是省略主效应的模型则显得十分古怪。在教育回报的这个例子中，类似于：

$$Y_i=\alpha+\gamma x_{2i}+\sum_{j=1}^{\tau}\delta_j(d_{ji}x_{2i})+\epsilon_i \tag{3.1.11}$$

这个模型要求教育只会影响女性的工资——一种和事实相差甚远的假设。因此，我们也很难对模型(3.1.11)的估计结果进行解释。

最后，我们还应该注意到饱和模型可以完美地拟合条件期望函数，这个性质与

① 如果在模型中加入第三个虚拟变量，比如 x_{3i}，那么饱和模型就包括三个主效应，三个二阶交互项 $\{x_{1i}x_{2i},\ x_{1i}x_{3i},\ x_{2i}x_{3i}\}$ 和一个三阶交互项 $x_{1i}x_{2i}x_{3i}$。

Y_i 的分布无关。当然，对于线性概率模型和其他的被解释变量受限模型（比如非负的 Y_i）来说，这个性质也是成立的，在本章最后我们再来详细讨论这个问题。

3.2 回归与因果关系

第 3.1.2 节告诉我们为什么回归是对条件期望函数的最佳线性逼近（在最小均方误的意义下）。但是，这个理解并没有回答我们提出的更深层次问题：何时可对回归赋予一个因果解释？何时可将回归系数看作对原本只在随机实验中出现的因果效应的近似？

3.2.1 条件独立假设

当条件期望函数可以近似因果关系时，我们称相应的回归也具有因果性。当然，这个回答并没有解决我们刚才提出的问题。由于回归可用来近似条件期望函数并具有条件期望函数所具有的性质，所以上面的这个回答只是将问题提到另外一个更高的层面上。因果性意味着不同的人会做出不同的决策，但是在很多学科领域中工作的研究人员都发现用第 2 章使用过的潜在结果来刻画因果关系显得相当有用，这种方式考虑的是去医院的人如果不去医院会发生什么。个体去医院和不去医院带来的结果的差别可被称为接受医院治疗的因果效应。如果对于给定的总体，条件期望函数刻画了平均潜在结果之间的不同，那么就说这个条件期望函数具有因果性。

由于在特定背景中很容易展开讨论看上去有点含糊的因果性条件期望函数的概念，所以我们这里还是以教育水平为例来讨论。教育和收入之间的因果关系可以定义为一个函数关系，这个函数描述的是如果给定个体接受不同的教育，他的收入会是多少。具体而言，我们可将人们做出接受不同水平教育的决策看作一段能够回头再过一遍的经历，在这段经历中人们可以做出这样或那样的决策。比如在高中时百无聊赖的 Angrist 郁闷地考虑着他的选择：是从高中辍学并得到一份工作，还是在高中上点容易的课，快速拿到一个不是很有价值的高中文凭，还是努力学习来上大学。虽然事先并不清楚这些选择带来的后果，但是不同的人生路径给每个特定个体带来不同结果这一点似乎是无可争议的。哲学家曾经争论这种个体潜在结果在科学的意义上是否足够精确，但是个体在做出决策时以这种方式来思考自己的生活和选择似乎是没有问题的（正如罗伯特·弗罗斯特在他《未选择的路》一诗中讲到：作为叙述者的旅行者曾经回忆到面对选择的那一刻。他相信沿着少有人迹的那条路“使得所有事情都变得不一样”，尽管他也明白走另一条路的结果可能不可知）。

在经验研究中，教育水平和收入之间的因果联系告诉我们如果人们在一个完美的受控实验中改变他的受教育水平，那么平均而言他们会赚到多少钱；或者

说如果人们随机选择受教育水平，从而使得他们在各方面都可比时，受教育水平的差异带来的收入水平的不同。正如在第 2 章中讨论的，实验保证了我们感兴趣变量与潜在的结果无关，从而使得被比较的组别之间是真正可比的。这里，我们将这一概念推广到因果变量取值超过两个并且有一系列控制变量需要给定的更复杂的情况，来使得因果推断得以成立。这就带来条件独立假设（conditional independence assumption，缩写为 CIA），这是能对回归赋予因果解释的核心假设。这个假设有时又被称为选择偏误来自可观察变量，因为这里假设：我们希望保持不变的那些协变量都是已知和可观察到的（Goldberger，1972；Barnow，Cain and Goldberger，1981）。因此现在最大的问题在于这些协变量是什么以及应该是什么。我们会在后面对此进行简短说明，就现在考虑的问题而言，我们还是使用计量经济学的方式，将这些变量称为协变量 X_i。随着对教育回报问题的深入讨论，我们会很自然地认为 X_i 是包含对能力和家庭背景进行度量的一个向量。

对于初学者来说，可将教育水平看作一个二值变量，比如可用它来表示 Angrist是不是去上大学。将此记为虚拟变量 C_i。于是可用第 2 章中描述实验中潜在结果的方法来刻画上大学和未来收入间的因果关系。方便讨论起见，我们想象有两个潜在的收入变量：

$$\text{潜在结果} = \begin{cases} Y_{1i} & \text{if } C_i = 1 \\ Y_{0i} & \text{if } C_i = 0 \end{cases}$$

这里 Y_{0i} 是第 i 个人不上大学的收入，Y_{1i} 是第 i 个人上大学的收入。我们想了解的是 Y_{1i} 和 Y_{0i} 之间的差别，它就是个体 i 上大学与不上大学带来的因果效应。如果我们能够回到过去让个体 i 选择不同的受教育水平，那么就可以度量这个因果效应。我们能够观察到的结果 Y_i 可表达为潜在结果的组合：

$$Y_i = Y_{0i} + (Y_{1i} - Y_{0i})C_i$$

因为我们只能看到 Y_{1i} 或者 Y_{0i}，但是不可能同时看到两者，所以我们希望估计 $Y_{1i} - Y_{0i}$ 的平均值，或对某些特定集合估计 $Y_{1i} - Y_{0i}$ 平均值，比如那些实际上大学的人接受大学教育的因果效应，这就是 $\mathrm{E}[Y_{1i} - Y_{01} \mid C_i = 1]$。

一般来说，接受过大学教育和没有接受过大学教育的个体间的收入差距很难度量接受大学教育带来的因果影响。沿着第 2 章的思路，我们有：

$$\underbrace{\mathrm{E}[Y_i \mid C_i = 1] - \mathrm{E}[Y_i \mid C_i = 0]}_{\text{观察到的收入差距}} = \underbrace{\mathrm{E}[Y_{1i} - Y_{0i} \mid C_i = 1]}_{\text{处理的平均因果效应}} + \underbrace{\mathrm{E}[Y_{0i} \mid C_i = 1] - \mathrm{E}[Y_{0i} \mid C_i = 0]}_{\text{选择性偏误}} \tag{3.2.1}$$

如果说上大学的那些人本来就可以赚得更多，那么这里出现的选择偏误就是正的，于是简单地比较 $\mathrm{E}[Y_i \mid C_i = 1] - \mathrm{E}[Y_i \mid C_i = 0]$ 可能夸大了接受大学教育带来的收益。

条件独立假设(CIA)指的是给定观察到的特点 X_i，选择性偏误消失。正式地说，也就是：

$$\{Y_{0i}, Y_{1i}\} \perp\!\!\!\perp C_i \mid X_i \tag{3.2.2}$$

其中，符号“$\perp\!\!\!\perp$”表示$\{Y_{0i}, Y_{1i}\}$与 C_i 之间相互独立的关系，黑色竖线右边的随机变量是协变量集合。给定条件独立假设，给定 X_i，可以对不同教育水平下平均工资差异赋予一个因果解释。换句话说，就是：

$$\mathrm{E}[Y_i \mid X_i, C_i = 1] - \mathrm{E}[Y_i \mid X_i, C_i = 0] = \mathrm{E}[Y_{1i} - Y_{0i} \mid X_i]$$

现在，我们将条件独立假设拓展到因果变量可以取多个值的情况，比如受教育年数 S_i 这类变量。由于受教育水平和收入之间的因果关系可能因人而异，所以我们用个体的收入函数：

$$Y_{si} \equiv f_i(s)$$

来描述这一因果关系。这个函数表示个体 i 接受 s 年教育后会获得的收入。如果 s 只取两个值 12 和 16，那么我们就回到了之前提到的接受/不接受大学教育的例子：

$$Y_{0i} = f_i(12); \; Y_{1i} = f_i(16)$$

更一般地，函数 $f_i(s)$告诉我们在任意的受教育水平 s 下个体 i 可能的收入。换句话说，$f_i(s)$回答了“如果……，就会……”这样的一个因果性问题。在考虑人力资本和收入之间关系的理论模型中，$f_i(s)$的具体形式可能由个体行为某个特点决定，可能被市场力量决定，或者二者兼而有之。

在更一般的条件下，条件独立假设(CIA)变为：

$$Y_{si} \perp\!\!\!\perp S_i \mid X_i, \; \forall s \tag{CIA}$$

在许多随机实验中，由于 S_i 是在给定 X_i 下随机分配的，所以条件独立假设自然成立(比如，在田纳西师生比例的实验中，在每个学校里，小班是被随机分配的)。在使用观察数据进行的研究中，条件独立假设意味着给定 X_i 下 S_i“就像被随机分配的那样好”。

给定 X_i，多接受一年教育带来的平均因果就是 $\mathrm{E}[f_i(s) - f_i(s-1) \mid X_i]$，多接受四年教育带来的平均因果就是 $\mathrm{E}[f_i(s) - f_i(s-4) \mid X_i]$。数据只能告诉我们 $Y_i = f_i(s_i)$，也就是当 $s = s_i$ 时的 $f_i(s)$。但是给定条件独立假设，给定 X_i，不同教育水平下平均收入的差异就可解释为教育的因果效应。换言之，

$$\begin{aligned} &\mathrm{E}[Y_i \mid X_i, S_i = s] - \mathrm{E}[Y_i \mid X_i, S_i = s-1] \\ &= \mathrm{E}[f_i(s) - f_i(s-1) \mid X_i] \end{aligned}$$

对任何的 s 都成立。比如，我们可以比较教育水平为 11 年和 12 年的个体间平均收入的差别，以此来了解高中毕业带来的平均因果效应：

$$\begin{aligned}&\mathrm{E}[Y_i \mid X_i, S_i=12]-\mathrm{E}[Y_i \mid X_i, S_i=11]\\&=\mathrm{E}[f_i(12) \mid X_i, S_i=12]-\mathrm{E}[f_i(11) \mid X_i, S_i=11]\end{aligned}$$

当条件独立假设成立时，就可以对上面的公式赋予一个因果解释：

$$\begin{aligned}&\mathrm{E}[f_i(12) \mid X_i, S_i=12]-\mathrm{E}[f_i(11) \mid X_i, S_i=11]\\&=\mathrm{E}[f_i(12)-f_i(11) \mid X_i, S_i=12]\end{aligned}$$

这里，选择性偏误来自两类收入之间的差别，一类是高中毕业生如果在 11 年级时辍学，他能获得的收入；另一类是在 11 年级时确实选择辍学的学生所获得的收入。但是，给定条件独立假设，给定 X_i，高中毕业与否和潜在收入没有关系，于是选择性偏误就消失了。在这个例子中我们还能注意到，对于高中毕业生而言，高中毕业的因果效应等于在每个 X_i 的取值处，高中毕业带来的平均效应：

$$\mathrm{E}[f_i(12)-f_i(11) \mid X_i, S_i=12]=\mathrm{E}[f_i(12)-f_i(11) \mid X_i]$$

这一点很重要，不过使用条件独立假设消除选择性偏误显得更为重要。

到现在为止，我们对 X_i 可取的每个值都构造了一个因果效应。这样做的结果是协变量 X_i 取多少值，就会有多少个因果效应，这些因果效应对我们来说太多了。经验研究者往往发现用一个综合指标来汇总一系列估计值会显得十分有用。由迭代期望律，高中毕业的无条件因果效应就是：

$$\begin{aligned}&\mathrm{E}\{\mathrm{E}[Y_i \mid X_i, S_i=12]-\mathrm{E}[Y_i \mid X_i, S_i=11]\} &(3.2.3)\\&=\mathrm{E}\{\mathrm{E}[f_i(12)-f_i(11) \mid X_i]\}\\&=\mathrm{E}[f_i(12)-f_i(11)] &(3.2.4)\end{aligned}$$

同理，我们还可能感兴趣于高中毕业对高中毕业生产生的因果效应：

$$\begin{aligned}&\mathrm{E}\{\mathrm{E}[Y_i \mid X_i, S_i=12]-\mathrm{E}[Y_i \mid X_i, S_i=11] \mid S_i=12\} &(3.2.5)\\&=\mathrm{E}\{\mathrm{E}[f_i(12)-f_i(11) \mid X_i] \mid S_i=12\}\\&=\mathrm{E}[f_i(12)-f_i(11) \mid S_i=12] &(3.2.6)\end{aligned}$$

这个系数告诉我们对于高中毕业生而言，仅仅由于高中毕业带来的收益。类似的，对大学毕业生而言，大学毕业的因果效应就是 $\mathrm{E}[f_i(16)-f_i(12) \mid S_i=16]$，大学毕业的无条件因果效应就是 $\mathrm{E}[f_i(16)-f_i(12)]$。

可以用 X_i 的边际分布做权重，通过对 X_i 每个可能值对应的因果效应进行加权平均来计算等式(3.2.3)所指的无条件因果效应，对于高中毕业生或者大学毕业生接受相应教育的因果效应而言，可以通过用 X_i 在相应集合中的分布函数为权重进行加权平均来计算该值。在这两个例子中，由此得到的都是匹配估计值：我们在具有相同协变量的那些个体之间比较不同教育水平下的平均收入差异，然后以某种方式进行加权平均。

在实际中，运用匹配策略进行的研究需要考虑很多的细节。我们在第 3.3.1 节对匹配的基本机理进行讨论时再涉及这些细节。这里要注意到匹配方法的一个

缺点是它不是自动执行的，而是需要两步走——匹配和平均，而且，如何从由此得到的残差中计算标准误也显得不那么明显。第三个需要我们考虑的问题是：本小节的核心是比较两种决策下的差别（高中毕业或者大学毕业的收入与辍学后的收入之间的比较），但是这种方法无法完全推广到我们真正考虑的问题上。因为 S_i 可取很多值，S_i 的每个可能取值带来的因果效应也许都不同，因此我们还需要想办法对不同的因果效应进行汇总①。这种考虑会使我们再次回到回归。

回归为我们提供了一个简单易用的经验研究策略，它可以自动地将条件独立假设转化为我们需要估计的因果效应。可以通过两种方式来完成这种转化：一种方式是假设 $f_i(s)$ 关于 s 是线性的，除了与其相加的误差项因人而异之外，其他部分对所有人都是相同的。在此假设下线性回归自然就是估计 $f_i(s)$ 的良好工具；另一种更为一般化同时理解起来也会复杂一点的方式则认为 $f_i(s)$ 未必关于 s 线性。即使允许 $f_i(s)$ 因人而异，允许 $f_i(s)$ 是非线性的，回归还是为我们提供了一个估计特定个体因果效应 $f_i(s)-f_i(s-1)$ 的加权平均值的工具。事实上，可以将回归值看作一类特殊的匹配估计结果，它捕捉到的是类似于等式(3.2.3)和等式(3.2.5)中那种平均的因果效应。

从现在起，我们集中讨论对回归赋予因果解释所需要的条件，暂时从对回归—匹配问题的讨论脱身。这里我们从上面提到的第一种方式出发，考虑一个线性的，因果效应为常数的模型。假设：

$$f_i(s)=\alpha+\rho s+\eta_i \tag{3.2.7}$$

除了模型是线性的之外，这个等式还假设我们关心的因果效应在不同个体之间是相同的。由于等式(3.2.7)告诉我们的是个体 i 在 s 的任意值下能够赚得的收入，而不是最终实现的值 s_i，所以这里还省略了 s 的下标 i。在这个例子中我们还同时假设在 $f_i(s)$ 中唯一因人而异的部分是误差项 η_i，其均值为零，用以捕捉决定潜在收入水平的其他不可观测因素。

将观察到的值 s_i 代入等式(3.2.7)可得：

$$Y_i=\alpha+\rho S_i+\eta_i \tag{3.2.8}$$

除了等式(3.2.7)将等式(3.2.8)中的参数解释为因果效应之外，等式(3.2.8)看上去像一个二值回归模型。重要的是，由于等式(3.2.7)是一个具有因果性的模型，所以 s_i 可能与潜在结果 $f_i(s)$ 相关，或者说在这个例子中与等式(3.2.8)中的残差项 η_i 相关。

现在考虑给定一系列可观察的协变量 X_i，条件独立假设成立。为了与等式

① 比如，我们可能会用 S_i 的分布来构造 S 的平均影响。换言之，对于每个 S，我们用匹配的方法估计 $\mathrm{E}[f_i(s)-f_i(s-1)]$，然后计算平均意义上的差别：

$$\sum \mathrm{E}[f_i(s)-f_i(s-1)]\mathrm{P}(s)$$

其中，$\mathrm{P}(s)$ 是 S_i 的概率权重函数。这是对导数的平均值 $\mathrm{E}[f_i'(S_i)]$ 的一种离散化近似。

(3.2.8)对潜在结果的函数形式的假设相一致，我们将潜在收入水平的随机项分解为可观察变量 X_i 和残差项 ν_i 的线性函数：

$$\eta_i = X_i'\gamma + \nu_i$$

其中，γ 是总体回归系数所成的向量，假设其满足 $\mathrm{E}[\eta_i \mid X_i] = X_i'\gamma$。由于我们将 γ 定义为对 η_i 关于 X_i 进行回归的结果，所以残差项 ν_i 和 X_i 不相关。更进一步，由条件独立假设，我们有：

$$\mathrm{E}[f_i(s) \mid X_i, S_i] = \mathrm{E}[f_i(s) \mid X_i] = \alpha + \rho s + \mathrm{E}[\eta_i \mid X] = \alpha + \rho s + X_i'\gamma$$

因此在线性模型

$$Y_i = \alpha + \rho s_i + X_i'\gamma + \nu_i \tag{3.2.9}$$

中的残差项与回归元 s_i 以及 X_i 都不相关，回归系数 ρ 就是我们感兴趣的因果效应。

这里需要再次强调的是我们做出的关键假设是：可观察的特点 X_i 是导致 η_i 和 s_i（等同于说 η_i 和 $f(s)$之间的相关性）相关的唯一原因。这就是在四分之一个世纪之前 Barnow，Cain 和 Goldberger(1981)针对回归讨论过的选择性偏误来自可观察变量的假设(selection-on-observable assumption)。它已成为经济学中绝大多数经验研究的基础。

3.2.2 遗漏变量偏误公式

除了我们感兴趣的变量 S_i 之外，我们还将一系列的控制变量 X_i 引入了回归模型。遗漏变量偏误(omitted variable bias)公式描述的是当回归包含不同的控制变量时，回归结果之间存在的关系。这个重要公式的出发点往往是：可以对类似于等式(3.2.9)的存在控制变量的回归方程赋予一个因果解释，但无法对不含有控制变量的回归方程赋予一个因果解释。因此在不含控制变量的较短的回归方程中得到的系数就被认为是有偏的(biased)。事实上，不论我们是否可以对含有控制变量的长回归方程赋予因果解释，遗漏变量偏误公式都提供了长回归方程和短回归方程估计系数之间的联系。方便起见，我们认为长回归方程中的系数和短回归方程中的系数由遗漏变量偏误公式决定。

为使讨论更加明确，假设在研究教育回报的回归方程中控制变量可以简化为家庭背景、智力和动机所组成的控制变量集合。将这些变量所组成的向量记为 A_i，并将此变量简记为“能力”。在控制了能力后，对工资关于教育水平进行回归的方程就可以写成：

$$Y_i = \alpha + \rho s_i + A_i'\gamma + e_i \tag{3.2.10}$$

其中，α，ρ 和 γ 是总体回归系数，e_i 是回归残差，由定义可知它和所有的回归元都无关。给定 A_i，如果条件独立假设成立，那么这里的 ρ 就等于等式(3.2.7)中的 ρ，在这里残差项 e_i 是控制了 A_i 后潜在收入水平的随机部分。

在实际中，能力是很难度量的。比如，美国当期人口调查（American Current Population Survey，简写为 CPS）——在应用微观经济学中大量使用的大型数据集（也是美国政府计算失业率的数据来源）——没有告诉我们任何关于被访者家庭背景、智商或者动机的信息。那么将能力排除在回归方程（3.2.10）之外的后果是什么？将能力排除在外的“短回归方程”中的参数与方程（3.2.10）中得到的参数之间的关系由下式给出：

遗漏变量偏误公式（Omitted Variables Bias Formula）

$$\frac{\text{cov}(Y_i,\ S_i)}{V(S_i)} = \rho + \gamma'\delta_A \tag{3.2.11}$$

其中，δ_A 是对 A_i 关于 S_i 回归得到的参数。解释一下，遗漏变量偏误说的是：

> 短回归参数等于长回归参数加上一个数，这个数等于遗漏变量效应乘以遗漏变量对被包含变量的回归系数。

将长回归方程代入短回归方程的系数公式 $\frac{\text{cov}(Y_i,\ S_i)}{V(S_i)}$ 就可很容易地求出这个公式。毫不令人惊讶，遗漏变量偏误与来自 3.1.2 节的解构回归公式（3.1.3）有紧密的关系。遗漏变量公式和解构回归公式都告诉我们当遗漏变量和纳入回归方程的变量不相关时，长回归和短回归得到的系数是一样的①。

我们可以使用遗漏变量偏误来感受一下在研究教育回报的例子中遗漏表征能力的那些变量会带来什么影响。表征能力的变量对工资有正的影响，这些变量也很可能与教育水平正相关。因此相对于我们想得到的那个系数，短回归方程的系数应该是偏大了。另一方面，正如经济理论所言，教育水平和能力之间的关系可能并不是很清晰。一些遗漏变量可能与教育水平是负相关的，在这些例子中短回归方程的系数又有可能偏小②。

表 3.1 用来自国家青年人纵向调研数据（简写为 NLSY）来阐述了上面提到的这些内容。在该表中前三列显示出当家庭背景——在这里是父母的受教育年限——加入回归并同时纳入一些表征人口基本特征的控制变量（年龄、种族和统计口径下的居住地区）后教育水平变量之前的系数由 0.132 下降至 0.114。为了进一步控制能力，我们用武装部队资格测验（Armed Forces Qualification Test，简称 AFQT，它是军方用于挑选士兵的测验）作为代理变量，这样将教育水平变量之前的系数降低至 0.087。遗漏变量偏误公式告诉我们回归系数的不断下降乃是与工

---

① 这里是多变量情况下对遗漏变量偏误公式的推广：记 β_1 为短回归方程中不存在其他变量时对 $k_1 \times 1$ 维向量 X_{1i} 回归得到的参数，记 β_1' 为包含了额外的 $k_2 \times 1$ 维向量 X_{2i} 时进行回归得到的参数，β_2' 是额外的向量在长回归中的系数。那么 $\beta_1 = \beta_1' + \text{E}[X_{1i}X_{1i}']^{-1}\text{E}[X_{1i}X_{2i}']\beta_2'$

② 作为接受过很高教育的人，我们倾向于认为能力和教育水平是正相关的。但这并不是一个可以先行预知的结论；可能因为对高能力的人而言，上学的机会成本太大了，所以 Mick Jagger（滚石乐队主唱）从伦敦经济学院退学，比尔·盖茨选择从哈佛退学（当然，他们也可能是一对非常幸运的大学辍学者）。

资和教育水平都正相关的控制变量不断加入的结果①。

表 3.1　在 NLSY 中估计男性的教育回报

控制变量	(1) 无	(2) 对年龄的 虚拟变量	(3) 第二列并加上额 外的控制变量*	(4) 第三列并加上 AFQT 分数	(5) 第四列加上职 业的虚拟变量
	0.132	0.131	0.114	0.087	0.066
	(0.007)	(0.007)	(0.007)	(0.009)	(0.010)

注：数据来自国家青年人纵向调研数据(他们在 1979 年出生，于 2002 年进行的调研)。上表报告了将对数化的工资关于受教育年龄进行回归后的系数以及相应的控制变量。括号里显示的是标准误。样本只考虑男性，并关于 NLSY 样本权重进行加权。样本规模是 2 434。

※ 额外的控制变量是父母的受教育年限，虚拟变量是种族和统计口径中的地区。

虽然看上去简单，但是遗漏变量偏误公式是我们理解回归时最重要的公式之一。遗留变量偏误公式的重要性在于下面的事实：如果你指出回归中不存在遗漏变量偏误，那等于说你得到的回归就是你想要的那个。而且，你想要的回归往往可以赋予一个因果解释。换句话说，你已经准备依靠条件独立假设来对长回归方程进行因果解释了。

在这时，需要考虑在什么情况下条件独立假设最有可能为经验研究提供一个可信的基础。最好的情况就是在某些实验(可能是自然实验)中给定 X_i 后对 S_i 进行随机分配。一个例子就是由 Black 等(2003)中对失业工人强制再培训项目所进行的研究。在该项研究中，作者关注强制再培训项目是否成功地提高了失业工人的工资。他们发掘的事实是：强制再培训项目的入选资格取决于基本的个体特征、过去的失业记录和工作历史。根据这些特征，工人被分入不同的组。虽然其中一些组别的工人不满足接受强制再培训项目的资格，但在其他组别中的工人是满足强制再培训项目的资格的，因此如果他们不工作，那么就必须参加培训。当某些强制接受培训的组别中工人数目大于受培训的限额数目时，接受培训的机会是以抽签的方式决定的。因此，给定导致工人被分配至不同组别的协变量，培训状况是随机分配的，于是用标志工人是否参加培训的虚拟变量以及表征个体特点、过去失业状况和就业历史的变量进行回归，就很可能得到对工人接受培训的因果效应的可靠估计②。

在对教育回报的研究中，人们往往不是通过抽签来决定某个人是继续上大学还是结束高中教育③。但我们仍可想象一下，将具有相似能力、来自相似的家庭背景的个体置于一个实验中，这个实验的目的是要鼓励他们继续上学。英国的教育维持计划(Education Maintenance Allowance)为英国特定地区的高中生支付学费，就是这样的一个政策实验(Dearden et al.，2003)。

① 大量经验研究文献讨论了在教育方程中遗漏能力变量所导致的后果。早期的重要参考文献包括 Griliches 和 Mason(1972)，Taubman(1976)，Griliches(1977)以及 Chamberlain(1978)。

② 这个项目看上去可以提高收入，主要因为接受培训的工人会更快地回到工作岗位上。

③ 曾经用抽签的方式来分配私立学校的学费补贴，见 Angrist 等(2002)。

能够使条件独立假设成立的第二个方式就是详细地考察决定 S_i 如何分配的制度知识。一个例子就是 Angrist(1998)关于志愿兵役服务对士兵退役后收入影响的研究。这项研究想要探寻的是从长期来看，参加美国军队志愿服役的男性其后来的收入是否得到了显著的提升。由于志愿兵役服役不是随机分配的，所以我们无法确切地知道这个问题的答案。因此 Angrist 用匹配和回归的技巧来控制在服役和没有服役的个体之间观察到的差异，这里不论是服役的个体还是没有服役的个体，他们在 1979 年到 1982 年间都申请了志愿服役。在这个例子中用来控制可观察特点的策略来自于如下事实：军方主要依据年龄、教育程度和考试分数来筛选士兵的申请。

要说明条件独立假设在 Angrist(1998)的研究中是成立的，等于要说明控制了所有可观察的因素后，服役和未服役的人之间是可比的。既然在 Angrist(1998)中给定 X_i 后服役状态的变化完全来自于以下事实：一些合格的申请者在最后一分钟没有被列入招募名单，所以在该项研究中使用的条件独立假设值得我们仔细把玩。当然，使合格的申请者没能进入征召名单的过程本身可能和被研究个体的潜在收入水平有联系，因此在这种情况下条件独立假设显然无法得到满足。

3.2.3 不合格的控制变量

我们已经指出：对协变量的控制可以提高使回归估计值获得因果解释的可能性，但并非控制变量越多越好。有些控制变量是不合格的控制变量，将其加入回归固然可以改变回归系数，但实际上却不该将其加入。由于我们总是可以将经验研究想象为一个实验，所以不合格的控制变量就是那些可以作为实验结果的变量。也就是说，不合格的控制变量本身可作为被解释变量。合格的控制变量是指当我们选定回归元后，它的取值已经固定给出的那些变量。

从本质上看，不合格的控制变量带来的问题仍然是选择偏误，但这个问题可能比第 2 章和第 3.2.1 节讨论过的那种选择偏误更加微妙。为了阐述清楚，假设我们感兴趣于大学教育对收入的影响，同时人们还可在白领和蓝领之间进行职业选择。由于接受大学教育无疑为我们开启了通往高收入白领阶层的大门，所以职业是不是我们在使用教育水平对收入进行回归时的一个遗漏变量呢？毕竟，职业选择和教育水平、收入都高度相关。也许最好的解决方法是在同一类职业中考察教育对收入的影响，比如说只在白领工人中考察该影响。上面这个建议的问题在于：一旦我们考虑到教育水平影响职业选择，那么即使教育水平是随机分配的，同一职业内不同教育水平下的工资差异也不再是可以相互比较的同类事物。

现在我们以刚才列举的大学教育/职业选择问题为背景，正式地考察不合格控制变量问题①。记 W_i 是表示个体 i 是否为白领工人的虚拟变量，该个体的收入水

① 在 3.4.2 小节详细讨论依正概率为条件进行比较(conditional-on-positive comparison)的例子中还会详细讨论这个问题。

平为 Y_i。每个个体接受或不接受大学教育都带来收入水平和职业选择的两种不同的潜在结果，分别记为$\{Y_{1i}, Y_{0i}\}$和$\{W_{1i}, W_{0i}\}$，于是收入水平 Y_i 和职业选择 W_i 的实现有赖于是否大学毕业和潜在可能结果的相互作用，于是我们有：

$$Y_i = C_i Y_{1i} + (1 - C_i) Y_{0i}$$
$$W_i = C_i W_{1i} + (1 - C_i) W_{0i}$$

其中，$C_i = 1$ 表示大学毕业水平，$C_i = 0$ 为其他。我们假设 C_i 是随机分配的，所以它独立于所有的潜在结果。由独立性可知，在估计 C_i 对 Y_i 和 W_i 的因果效应时不会遇到任何困难：

$$\mathrm{E}[Y_i \mid C_i = 1] - \mathrm{E}[Y_i \mid C_i = 0] = \mathrm{E}[Y_{1i} - Y_{0i}]$$
$$\mathrm{E}[W_i \mid C_i = 1] - \mathrm{E}[W_i \mid C_i = 0] = \mathrm{E}[W_{1i} - W_{0i}]$$

在实际操作中我们分别将 Y_i 和 W_i 关于 C_i 回归就可以得到平均因果效应。

不合格的控制变量意味着给定 W_i 后无法对收入水平的差异比较赋予一个因果解释。给定白领职业下，考虑大学毕业生和非大学毕业生的收入差距。我们可以在包含了 W_i 的回归模型中计算这个收入差距，也可以在 $W_i = 1$ 的所有个体中对 Y_i 关于 C_i 进行回归。后一个方法得到的估计值就是当 $W_i = 1$ 时 C_i 取值分别为 0 和 1 时显示出的平均收入上的差异：

$$\begin{aligned} &\mathrm{E}[Y_i \mid W_i = 1, C_i = 1] - \mathrm{E}[Y_i \mid W_i = 1, C_i = 0] \\ &= \mathrm{E}[Y_{1i} \mid W_{1i} = 1, C_i = 1] - \mathrm{E}[Y_{0i} \mid W_{0i} = 1, C_i = 0] \end{aligned} \tag{3.2.12}$$

由 $\{Y_{1i}, W_{1i}, Y_{0i}, W_{0i}\}$ 的联合分布与 C_i 相互独立可知：

$$\begin{aligned} &\mathrm{E}[Y_{1i} \mid W_{1i} = 1, C_i = 1] - \mathrm{E}[Y_{0i} \mid W_{0i} = 1, C_i = 0] \\ &= \mathrm{E}[Y_{1i} \mid W_{1i} = 1] - \mathrm{E}[Y_{0i} \mid W_{0i} = 1] \end{aligned}$$

这个等式指出了不合格控制变量带来的问题，相比较的不是同类事物：

$$\begin{aligned} &\mathrm{E}[Y_{1i} \mid W_{1i} = 1] - \mathrm{E}[Y_{0i} \mid W_{0i} = 1] \\ &= \underbrace{\mathrm{E}[Y_{1i} - Y_{0i} \mid W_{1i} = 1]}_{\text{因果效应}} + \underbrace{\{\mathrm{E}[Y_{0i} \mid W_{1i} = 1] - \mathrm{E}[Y_{0i} \mid W_{0i} = 1]\}}_{\text{选择性偏误}} \end{aligned}$$

换言之，给定所考虑的个体都是白领工人，是否拥有大学学历造成的工资差异等于大学文凭对 $W_{1i} = 1$（获得大学学历后会成为白领工人）的那些人带来的因果效应加上一个选择偏误项，这一选择偏误项反映出的是大学学历会改变白领工人组成这一事实。

在这个例子中，选择偏误项的符号不确定，它依赖于职业选择、是否上大学以及潜在收入水平之间的关系。我们想表达的主要观点是：即使 $Y_{1i} = Y_{0i}$ ——上大学对工资不具有因果效应，等式(3.2.12)也无法指出这一点（对 Y_i 关于 W_i 和 C_i 进行回归也会出现同样的问题）。与此同时，在给定职业选择下，我们也不能认为不同教育水平下的收入差异捕捉到了没能被职业解释的那部分因果效应。事实

上，在缺乏将教育、职业和收入联系起来的更加精致的模型下，给定职业选择结果后得到的不同教育水平的收入差异并未告诉我们更多的东西①。

在一个经验研究的例子中，我们可以看到将代表两位数职业代码的虚拟变量加入回归后确实降低了教育水平变量前的系数，在针对 NSLY 的模型中该系数从 0.087 下降到 0.066。但是我们很难解释是何种原因导致了这种下降。教育水平的系数变小可能仅仅是选择偏误的一种表现。因此我们最好还是用不由教育水平决定的那些变量作为控制变量。

当使用代理变量做控制变量时，也会出现不合格的控制变量问题，也就是说纳入回归方程的变量可能部分地控制遗漏变量，但是它本身被我们感兴趣的变量影响。举一个用代理变量做控制变量的简单例子，假设你对类似于(3.2.10)的长回归方程感兴趣：

$$Y_i = \alpha + \rho S_i + \gamma a_i + e_i \tag{3.2.13}$$

为了方便讨论，这里将控制变量向量 A_i 更换为表示能力的数值变量 a_i。可以将这个变量视为在八年级时学生的智力测验分数，这个分数可以度量学生的天生能力，同时八年级又先于学生的任何接受教育的决策(假设所有人都要完成八年的教育)。根据定义，这个方程中的误差项满足 $E[S_i e_i] = E[a_i e_i] = 0$。由于 a_i 是在个体做出 S_i 决策之前做出的，所以它是个好的控制变量。

等式(3.2.13)是我们感兴趣的回归方程，不幸的是关于 a_i 的数据不可得。但是你可能拥有另外一个度量能力的指标，这个指标是在个体接受完教育后得到的(比如说在求职申请中用到的测试分数)。不妨将这个变量叫做后天的能力，记为 a_{ki}。一般而言，在先天能力的基础上，接受过教育后个体提高了他们后天的能力。具体而言，就是：

$$a_{ki} = \pi_0 + \pi_1 S_i + \pi_2 a_i \tag{3.2.14}$$

也就是说先天能力以及教育水平都在提高后天的能力。在度量能力时几乎一定会存在某些随机性，但是用等式(3.2.14)这样的确定性的方程可以使我们的讨论变得简单。

当仅仅用 S_i 对 Y_i 进行回归时，你会担心遗漏变量引起的偏误，既然完美的控制变量 a_i 不可得，那么你希望用 S_i 和后天能力 a_{ki} 对 Y_i 进行回归。将等式(3.2.14)代入等式(3.2.13)后，关于 S_i 和 a_{ki} 的回归方程就变为：

$$Y_i = \left(\alpha - \gamma\frac{\pi_0}{\pi_2}\right) + \left(\rho - \gamma\frac{\pi_1}{\pi_2}\right)S_i + \frac{\gamma}{\pi_2}a_{ki} + e_i \tag{3.2.15}$$

在这个例子中，γ、π_1 和 π_2 都是正的，因此只有当 π_1 变得很小时，估计出的结果才能越接近因果效应。换言之，把随着 S_i 的提高而提高的代理控制变量加入回

① 在这个例子中，选择偏误很可能是负的，也就是说 $E[Y_{0i} \mid W_{1i} = 1] < E[Y_{0i} \mid W_{0i} = 1]$。任何大学毕业生都能获得白领的工作应该是比较合理的，因此 $E[Y_{0i} \mid W_{1i} = 1]$ 应该与 $E[Y_{0i}]$ 相去不远。但是那些没有大学学历(也即是 $W_{0i} = 1$)也能得到白领工作的人是比较特别的，所以对这些人而言，$E[Y_{0i} \mid W_{0i} = 1]$ 要比 Y_{0i} 的平均值高。

归，会使估计值小于我们想要的那个值。但重要的是我们还应该注意对参数 π_1 进行考察：如果用 a_{ii} 对 S_i 的回归结果为零，那么你可以相当自信地假设等式(3.2.14)中的 π_1 等于零。

在讨论代理控制变量问题时，我们遇到了一开始讨论不合格控制变量时没有遇到的模棱两可的问题。对被解释变量添加控制变量可能具有一定的误导性：如果你想在对教育回报的研究获得带有因果性的结果，那你就不会在回归方程中加入职业作为控制变量。但是在代理变量作为控制变量的问题中，你的出发点是对的。即使加入代理控制变量仍然没能得到我们感兴趣的回归参数，但相比于没有加入该变量，现在得到的结果改进了。回顾我们是根据等式(3.2.13)来使用代理控制变量的。遗漏变量偏误公式告诉我们当不存在控制变量时关于 S_i 做回归产生的系数是 $\rho+\gamma\delta_{ai}$，其中 δ_{ai} 是对 a_i 关于 S_i 进行回归得到的系数。相比之下，在等式(3.2.15)中得到的教育水平的系数要比没有控制变量情况下更接近于 ρ。更进一步，如果假设 δ_{ai} 为正，那么我们可以很确定地说我们感兴趣的因果效应处在无控制变量估计量和加入代理控制变量得到的估计量之间。

当我们开始思考使用何种变量做控制变量时，对不合格控制变量和代理性控制变量都适用的一个挑选准则是：考虑控制变量被决定的时间。一般来说在我们感兴趣的变量产生之前就被决定的变量都是好的控制变量。很明显，在我们感兴趣的变量产生前就被决定的变量不可能是我们考虑的因果关系的产物。但是有时我们必须面对控制变量被决定的时间不确定或未知的情况。在这种情况下，因果关系的准确考量需要我们做出哪个变量先被决定的假设，或者去说明没有任何一个控制变量是由我们感兴趣的变量所影响的。

3.3 异质性与非线性

正如前面几节所言，将线性因果模型与条件独立假设相结合，我们可以得到一个线性条件期望函数并对其赋予因果解释。虽说假设条件期望函数是线性的，那么总体回归方程就是它本身。但实际上，无需假设条件期望函数为线性，我们也可以对回归结果赋予一个因果解释。就像在3.1.2节讨论的那样，无论条件期望函数的具体形式是怎样的，我们都可以将用 X_i 和 S_i 对 Y_i 所进行的回归看作对相应的条件期望函数的最佳线性逼近。因此，如果条件期望函数具有因果性，那么回归可以逼近条件期望函数这一事实使得回归系数也具有了某种意义上的因果性。不过这种说法还是不够精确，我们还需要对回归和条件期望函数之间的联系作进一步阐发，这一阐发让我们理解到：回归是一种匹配估计量，并且由其提供的计算匹配估计量的方法具有良好性质。

3.3.1 回归与匹配

在过去的十年到二十年间，匹配(matching)开始成为一类经验研究工具并逐

渐引起了人们的兴趣。与前几节类似，匹配也是在条件独立假设上发展出的一种用于控制协变量的研究策略。比如 Angrist(1998)使用匹配法估计了志愿服兵役对之后收入的影响。给定军方在挑选士兵时所考虑的各种个体特征(比如年龄、教育水平和考试分数)，通过假设申请者能否成为士兵与其潜在收入水平无关，Angrist对用匹配法得到的估计值赋予了一个因果解释。匹配估计值的形成机制简单到令人心动：事实上，匹配法对由每个协变量的特定值所决定的个体计算处理组和控制组之间的平均差异，然后用加权平均的方式将这些平均因果效应汇总到一个总的因果效应中。

匹配策略的吸引人之处在于我们可以清楚地看到只有保证条件独立假设成立才能为匹配估计结果赋予一个因果解释。与此同时，我们已经看到对回归系数赋予一个因果解释也需要相同的假设。换言之，回归和匹配都是用来控制协变量的研究策略。既然在这两种研究策略下相应的因果推断都需要相同的核心假设，那么我们需要考虑在多大程度上回归和匹配之间是有区别的。我们的看法是可以将回归看做是一种特殊的匹配估计量，因此从经验研究的角度看，两者的区别并不重要。知道匹配和回归都是在得到因果解释中用于控制其他因素的一种策略。既然在得到因果推断中所使用的核心假设相同，那么弄清楚回归和匹配是否不同以及区别在哪里就很有必要了。我们的看法是，回归只是特定类型的一种加权后的匹配估计量，因此回归和匹配之间的区别看上去并不很重要。

为了充实这个观点并进一步考察匹配和回归估计值的数学结构——也就是考察在这些方法下得到的总体估计量——有助于我们的进一步理解。当然，对回归而言，估计量是总体回归参数所组成的向量。匹配法下得到的被估量则是一个加权平均值，其中每个被加权的值都是由协变量的特定取值所决定的那些个体中处理组和控制值之间的差异。由于在离散协变量下最容易看清匹配的作用机制，所以我们以志愿服兵役对之后收入的影响为例进行讨论，其中用虚拟变量 D_i 表示个体 i 是否参军。由于处理变量只取两个值，所以可以用 Y_{1i} 和 Y_{0i} 来表示潜在结果。这里我们主要感兴趣的变量就是处理组所表现出的平均处理效果 $\mathrm{E}[Y_{1i}-Y_{0i}\mid D_i=1]$。这个等式告诉我们士兵的平均收入 $\mathrm{E}[Y_i\mid D_i=1]$——这是可以观察到的数量，与这些士兵如果没有参军他们将会得到的收入 $\mathrm{E}[Y_{0i}\mid D_i=1]$ 之间的差别。除非标志是否参军的变量 D_i 与 Y_{0i} 相互独立，否则简单地比较自愿参军和未参军者之间的收入差距，得到的处理效果可能是有偏的。具体而言，就是：

$$\begin{aligned}&\mathrm{E}[Y_i\mid D_i=1]-\mathrm{E}[Y_i\mid D_i=0]\\&=\mathrm{E}[Y_{1i}-Y_{0i}\mid D_i=1]+\{\mathrm{E}[Y_{0i}\mid D_i=1]-\mathrm{E}[Y_{0i}\mid D_i=0]\}\end{aligned}$$

换言之，观察到的收入差别等于平均处理效应加上选择性偏误。这与第 2 章提到的选择性偏误是一样的。

在这里，条件独立假设是指：

$$\{Y_{0i},\ Y_{1i}\}\perp\!\!\!\perp D_i\mid X_i$$

如果条件独立假设成立，在给定 X_i 下，选择性偏误消失，因此可以通过对 X_i 使用迭代期望律计算被处理的平均处理效应。

$$\begin{aligned}\delta_{TOT} &\equiv \mathrm{E}[Y_{1i}-Y_{0i} \mid D_i=1] \\ &= \mathrm{E}\{\mathrm{E}[Y_{1i}-Y_{0i} \mid X_i, D_i=1] \mid D_i=1\} \\ &= \mathrm{E}\{\mathrm{E}[Y_{1i} \mid X_i, D_i=1]-\mathrm{E}[Y_{0i} \mid X_i, D_i=1] \mid D_i=1\}\end{aligned}$$

当然，$\mathrm{E}[Y_{0i} \mid X_i, D_i=1]$ 并非真实存在。但是，由条件独立假设：

$$\mathrm{E}[Y_{0i} \mid X_i, D_i=0]=\mathrm{E}[Y_{0i} \mid X_i, D_i=1]$$

因此，

$$\begin{aligned}\delta_{TOT} &= \mathrm{E}\{\mathrm{E}[Y_{1i} \mid X_i, D_i=1]-\mathrm{E}[Y_{0i} \mid X_i, D_i=0] \mid D_i=1\} \\ &= \mathrm{E}[\delta_X \mid D_i=1] \qquad (3.3.1)\end{aligned}$$

其中，

$$\delta_X \equiv \mathrm{E}[Y_i \mid X_i, D_i=1]-\mathrm{E}[Y_i \mid X_i, D_i=0]$$

是在 X_i 的每个特定值决定的个体中由于是否参军带来的收入的平均差别。对于 X_i 的每个特定值，比如 $X_i=x$，我们将相应的收入的平均差距记做 δ_x。

在 Angrist(1998)中使用了协变量 X_i 是离散的这一事实来构造等式(3.3.1)等号右端的样本估计值。在离散情况下，匹配估计量可以写为：

$$\mathrm{E}[Y_{1i}-Y_{0i} \mid D_i=1]=\sum_x \delta_x \mathrm{P}(X_i=x \mid D_i=1) \qquad (3.3.2)$$

其中，$\mathrm{P}(X_i=x \mid D_i=1)$ 是给定 $D_i=1$ 下 X_i 的概率质量函数①。在这个例子中，X_i 的取值是出生年份、考试分数、向军方递交申请所在年份以及申请年份中申请者受教育程度四个变量的所有可能组合。在这个例子中的考试分数来自于AFQT，它是军方用于区分申请者智力能力的测试(在 3.2.2 节中我们将这个变量作为控制变量加入回归)。在 Angrist(1998)中用由每个协变量可取值组合确定的个体中参军和非参军人士的平均收入差距作为 δ_x，然后用协变量在参军者中的分布来计算处理效应的加权平均值。

注意到我们还可以很方便地构造无条件平均处理效应：

$$\begin{aligned}\delta_{ATE} &= \mathrm{E}\{\mathrm{E}[Y_{1i} \mid X_i, D_i=1]-\mathrm{E}[Y_{0i} \mid X_i, D_i=0]\} \\ &= \sum_x \delta_x \mathrm{P}(X_i=x)=\mathrm{E}[Y_{1i}-Y_{0i}] \qquad (3.3.3)\end{aligned}$$

这是用 X_i 的边际分布得到的 δ_x 的加权期望值，而不是用被处理者中 X_i 的分布计算平均值。δ_{TOT} 告诉我们参军对代表性士兵带来的收入的增减量，而 δ_{ATE} 告诉我们代表性申请者收入的增减量(因为在 Angrist(1998)中，总体由所有的申请人构成)。

美国军方在挑选士兵时看上去相当挑剔，在冷战末期大规模裁军后这种情况更

① Rubin(1977) 也讨论了这个匹配估计量，Card 和 Sullivan(1988)运用它估计了受补贴的培训项目对个体就业状况的影响。

甚。基本上军方现在只招募那些考试分数分布在前一半的高中毕业生。因此如果只简单地比较参军者和非参军者之间的收入差距，那么军方挑选参军申请人的过程就会产生正的选择偏误。这一点可以从表 3.2 中看出，该表报告了在 1979 年到 1982 年间申请加入军队的男性由于在军队服役所造成的平均收入差距以及运用匹配和回归的方法计算出的这一差距，其中个人收入数据来源于 1988 年到 1991 年社保账户中纳税收入水平。匹配估计值的计算类似于等式(3.3.2)。虽然平均而言白人士兵在后来比非白人士兵多赚 1 233 美元，但是一旦将个体在协变量上的差异控制住，这种由参军造成的收入差别就变为负的。同理，虽然非白人士兵收入要比非白人为服役的那些人高 2 449 美元，但是控制协变量后这个差距下降到了 840 美元。

表 3.2 还报告了将构造匹配估计值时用到的协变量控制住后使用回归估计出的志愿服役对收入的影响，即下式中 δ_R 的估计值：

$$Y_i = \sum_x d_{ix}a_x + \delta_R D_i + e_i \tag{3.3.4}$$

其中，$d_{ix} = 1[X_i = x]$ 是一个虚拟变量，当 $X_i = x$ 时，该虚拟变量为 1，a_x 估计的是在回归方程中 $X_i = x$ 带来的影响，δ_R 是回归的待估参数。注意到该回归模型中每一组协变量都产生一个不同的参数，因此可以将该模型称为关于 X_i 饱和，当然这个模型不是完全饱和的，因为对 D_i 而言不存在交互项 $D_i \cdot X_i$。虽然在这个例子中匹配和回归都控制了相同的变量，但是表 3.2 中的结果显示出相比于匹配策略下得到的相应估计值，用回归对非白种人估计出的系数更大，对白种人估计出的系数则负得比较小。事实上，匹配和回归策略下得到的估计结果之间的差异在统计上是显著的。同时，用回归和匹配两种方法得到的服兵役对收入的影响又表现出大致相同的趋势。回归和匹配策略下得到的结果类似，这是因为可将回归看作一种匹配估计值：回归和匹配的差别只在于将处理效应 δ_X 加权平均到一个总体平均处理效应时使用的权重不同。具体而言，匹配策略中进行加权平均时使用的权重是处理组中协变量的分布，而回归结果使用的权重则是方差。

表 3.2　分别用无控制变量、匹配和回归三种方法得到的志愿赋予对收入产生的影响的估计

种　族	在 1988 年到 1991 年之间的平均收入 (1)	是否参军带来的平均收入的差距 (2)	匹配估计值 (3)	回归估计值 (4)	回归估计量减去匹配估计量 (5)
白　人	14 537	1 233.4	−197.2	−88.8	108.4
		(60.3)	(70.5)	(62.5)	(28.5)
非白人	11 664	2 449.1	839.7	1 074.4	234.7
		(47.4)	(62.7)	(50.7)	(32.5)

注：本表来自于 Angrist(1998)的表Ⅱ和表Ⅴ。标准误报告在括号中。这张表报告了在 1979 年到 1982 年间申请加入军队的男性由于在军队服役所造成的平均收入差距以及运用匹配和回归的方法计算出的这一差距，其中个人收入数据来源于 1988 年到 1991 年社保账户中纳税收入水平。匹配和回归估计值都是在控制了申请人的出生年份、申请时的受教育水平以及 AFQT 考分后得到的。样本中共有 128 968 位白人和 175 262 位非白人。

为了看清楚这一点，首先用解构回归的公式写出对 Y_i 关于 X_i 和 D_i 进行回归后得到的系数：

$$\delta_R = \frac{\text{cov}(Y_i, \tilde{D}_i)}{V(\tilde{D}_i)} \tag{3.3.5}$$

$$= \frac{\text{E}[(D_i - \text{E}[D_i \mid X_i])Y_i]}{\text{E}[(D_i - \text{E}[D_i \mid X_i])^2]}$$

$$= \frac{\text{E}\{(D_i - \text{E}[D_i \mid X_i])\text{E}[Y_i \mid D_i, X_i]\}}{\text{E}[(D_i - \text{E}[D_i \mid X_i])^2]}$$

$$= \frac{\text{E}\{(D_i - \text{E}[D_i \mid X_i])\text{E}[Y_i \mid D_i, X_i]\}}{\text{E}[(D_i - \text{E}[D_i \mid X_i])^2]} \tag{3.3.6}$$

第二个等号来自于如下的事实：模型关于 X_i 是饱和的，这意味着 $\text{E}[D_i \mid X_i]$ 是线性的。因此，根据对 $\tilde{D}_i$ 的定义，它就是 $D_i - \text{E}[D_i \mid X_i]$。第三个等号来自于如下事实：对 Y_i 关于 X_i 和 D_i 做回归等价于对 Y_i 关于 $\text{E}[Y_i \mid D_i, X_i]$ 做回归（这是我们在定理 3.1.6——条件期望函数的回归定理中得到的结论）。

为了更进一步简化，我们将条件期望函数 $\text{E}[Y_i \mid D_i, X_i]$ 展开，于是有：

$$\text{E}[Y_i \mid D_i, X_i] = \text{E}[Y_i \mid D_i = 0, X_i] + \delta_X D_i$$

将这个结果代入等式(3.3.6)的分子，于是有：

$$\begin{aligned}&\text{E}\{(D_i - \text{E}[D_i \mid X_i])\text{E}[Y_i \mid D_i, X_i]\}\\ =\ &\text{E}\{(D_i - \text{E}[D_i \mid X_i])\text{E}[Y_i \mid D_i = 0, X_i]\}\\ &+ \text{E}\{(D_i - \text{E}[D_i \mid X_i])D_i\delta_X\}\end{aligned}$$

因为 $\text{E}[Y_i \mid D_i = 0, X_i]$ 只是关于 X_i 的函数，而 $(D_i - \text{E}[D_i \mid X_i])$ 与 X_i 的任何函数都不相关，所以上式等号右边的第一项为零。类似的，上式第二项可以简化为：

$$\text{E}\{(D_i - \text{E}[D_i \mid X_i])D_i\delta_X\} = \text{E}\{(D_i - \text{E}[D_i \mid X_i])^2\delta_X\}$$

于是我们有：

$$\delta_R = \frac{\text{E}[(D_i - \text{E}[D_i \mid X_i])^2\delta_X]}{\text{E}[(D_i - \text{E}[D_i \mid X_i])^2]}$$

$$= \frac{\text{E}\{\text{E}[(D_i - \text{E}[D_i \mid X_i])^2 \mid X_i]\delta_X\}}{\text{E}\{\text{E}[(D_i - \text{E}[D_i \mid X_i])^2 \mid X_i]\}} = \frac{\text{E}[\sigma_D^2(X_i)\delta_X]}{\text{E}[\sigma_D^2(X_i)]} \tag{3.3.7}$$

其中，

$$\sigma_D^2(X_i) \equiv \text{E}[(D_i - \text{E}[D_i \mid X_i])^2 \mid X_i]$$

是给定 X_i 下 D_i 的条件方差。由此可知，回归模型(3.3.4)得到的参数乃是对匹配参数 δ_X 的一个加权平均，权数是给定 X_i 下 D_i 的条件方差。

由于我们感兴趣的回归元 D_i 是个虚拟变量，所以可以完成最后一步。在刚才提到的这个例子中，$\sigma_D^2(X_i) = \text{P}(D_i = 1 \mid X_i)(1 - \text{P}(D_i = 1 \mid X_i))$，所以：

$$\delta_R = \frac{\sum_x \delta_x [\mathrm{P}(D_i = 1 \mid X_i = x)(1 - \mathrm{P}(D_i = 1 \mid X_i = x))] \mathrm{P}(X_i = x)}{\sum_x [\mathrm{P}(D_i = 1 \mid X_i = x)(1 - \mathrm{P}(D_i = 1 \mid X_i = x))] \mathrm{P}(X_i = x)}$$

这说明回归估计值将每个特定协变量下的处理效应加权平均，权数是 $[\mathrm{P}(X_i = x \mid D_i = 1)(1 - \mathrm{P}(X_i = x \mid D_i = 1))]\mathrm{P}(X_i = x)$。相比之下，针对被处理者的处理效应计算的匹配估计值可以写作：

$$\begin{aligned} \mathrm{E}[Y_{1i} - Y_{0i} \mid D_i = 1] &= \sum_x \delta_x \mathrm{P}(X_i = x \mid D_i = 1) \\ &= \frac{\sum_x \delta_x \mathrm{P}(D_i = 1 \mid X_i = x)\mathrm{P}(X_i = x)}{\sum_x \mathrm{P}(D_i = 1 \mid X_i = x)\mathrm{P}(X_i = x)} \end{aligned}$$

上式用到如下结论：

$$\mathrm{P}(X_i = x \mid D_i = 1) = \frac{\mathrm{P}(D_i = 1 \mid X_i = x) \cdot \mathrm{P}(X_i = x)}{\mathrm{P}(D_i = 1)}$$

于是协变量取任何特定值时所对应的用来构造 $\mathrm{E}[Y_{1i} - Y_{0i} \mid D_i = 1]$ 的权重都与处理概率成正比。由此我们可知除非对个体进行的处理与协变量无关，否则回归结果和匹配结果总有差别。

通过考察匹配估计值对处理效应的加权方式，我们注意一个重要特点：使用匹配法对被处理者的处理效应进行估计时，对于由协变量的不同取值组合所决定的不同组别的个体而言，匹配法将最大权重赋予最可能被处理的那组个体的处理效应。相比之下，回归估计值将最大权重赋予条件方差最大的那组个体的处理效应。注意到，当 $\mathrm{P}(D_i = 1 \mid X_i = x) = \frac{1}{2}$ 时条件方差达到最大化，换言之，被处理者和被控制者数量相同的组别会被回归赋予最大权重。如果在不同组别间 δ_x 很少有变化（但是加权方式仍然影响估计值的统计有效性），那么使用哪种权重就显得无关紧要。在自愿参军对后来收入产生何种影响的研究中，看上去最愿意加入军队的那些男性从服役中获得的收益最小。这大概是因为最愿意服役的那些人最适合于服役，因而具有最高的潜在收入[①]。这个事实使得我们对服役带来的益处进行估计时，即使基于相同的控制变量，匹配估计值也要小于回归估计值[②]。

无论是回归估计值还是用依据协变量得到的匹配估计值，都没有对不包含处理者或者都是处理者的组别赋予权重，这一点同样重要。比如考虑 X_i 的一个特定值 x^*，由这个协变量确定的那组个体中不包含任何被处理者，或者该组个体中每

① 也就是说对这些潜在收入水平最高的人，服役不会对他们造成太大的影响。——译者注

② 由于对同方差、常处理效应模型，回归是有效的，所以回归将最大权重赋予 $\mathrm{P}(D_i = 1 \mid X_i = x) = \frac{1}{2}$ 的那组个体。我们应该希望将最大权重赋予估计得最精确的那些处理效应。当残差为同方差时，处理效应被估计得最精确的正是个体被处理概率为 $\frac{1}{2}$ 的组。

个人都被处理。那么 δ_x 就是个未定义值，回归中对应的权重 $[P(D_i = 1 \mid X_i = x^*)(1 - P(D_i = 1 \mid X_i = x^*))]$ 为零。在关于匹配的计量经济学文献中，当协变量达到饱和时，回归估计量和匹配估计量具有相同的支撑①，也即是说这两种方法只对既存在被处理者，又存在控制者的组别进行加权平均。

从估计量到估计值的这步转化显得有一些复杂。在实际中，都要对模型进行一定的假设才能得到回归和匹配估计量，这种假设乃是对由协变量所决定的子集进行一定的推断。比如，匹配估计量往往包含只有很少观察值的协变量。如果由这些协变量所决定的子集中的元素并非既有被处理的个体，也有作为控制的个体，那么具有共同支撑的假设就被违反。如果回归模型关于 X_i 不是饱和的，由协变量决定的子集所包含的个体并非既有被处理的，也有作为控制的，那么共同支撑假设还是无法保证，这些子集也就无法为基于推断的估计值作出贡献。尽管如此，这里还要指出的是我们看到了匹配和回归之间的一种对称性：总的来看，它们来自同样的统计方法，而且在实际操作中要求相同的建模假设②。

对回归和匹配的进一步讨论：有序处理和连续处理*

上一节将二元处理变量所在的回归解释为一种类似匹配的计量策略，那么对于有序处理和连续处理，这种解释是否也同样适用呢？对这个问题的详细回答具有很强的技术性，而且其中的一些内容我们并不想了解。因此简而言之的回答是：从某种程度是说，可以对有序处理和连续处理赋予同样的解释。

正如我们已经讨论过的，总体的最小二乘参数向量在最小均方误差意义下为我们提供了对条件期望函数的最优线性逼近。当然，这个定理不仅对二元处理变量适用，对有序处理变量和连续处理变量也同样适用。一个与之相关的性质就是可以将回归系数解释为“平均导数”。在多元回归模型中这种解释会变得很复杂，因为用最小二乘估出的斜率向量是对条件期望函数梯度的加权平均。除非特例，否则很难对使用矩阵进行加权的平均值作出解释(Chamberlain and Leamer，1976)。可以将平均导数性质用相对浅显的方式表达出来的一个特例就是协变量饱和、处理变量为有序处理或者连续处理的回归模型。为了避免啰嗦的求导过程，我们直接开始解释公式。求导过程附在本章附录，具体细节请参考 Angrist 和 Krueger (1999)中的附录。

① 随机变量的支撑是指能够以正概率出现的随机变量的集合。见 Heckman，Ichimura，Smith 和 Todd(1998)以及 Smith 和 Todd(2001)中对匹配策略中共同支撑的讨论。

② 对 X 变量具有良好分布的回归问题而言，解决之道是将协变量加总以得到相对粗糙的分组，或者比较那些虽不完全相同，但是很接近的协变量值所决定的子集中的个体。见 Cochran(1965)，Rubin(1973)或者 Rosenbaum(1995，第三章)中对这个问题的讨论。对于具有连续分布的协变量而言，匹配估计值将会有偏，因为无法做到完全的匹配。Abadie 和 Imbens(2008)最近已经指出在回归的基础上对偏误进行修正可以消除(在渐进的意义下)由无法完全匹配带来的偏误。

为了讨论的目的，假设标示处理水平的变量 S_i 是连续分布的随机变量，不一定非负。假设我们感兴趣的条件期望函数为 $h(t) \equiv \mathrm{E}[Y_i \mid S_i = t]$，其导数为 $h'(t)$。于是：

$$\frac{\mathrm{E}[Y_i(S_i - \mathrm{E}[S_i])]}{\mathrm{E}[S_i(S_i - \mathrm{E}[S_i])]} = \frac{\int h'(t)\mu_t \mathrm{d}t}{\int \mu_t \mathrm{d}t} \tag{3.3.8}$$

其中，

$$\mu_t \equiv \{\mathrm{E}[S_i \mid S_i \geqslant t] - \mathrm{E}[S_i \mid S_i < t]\}\{\mathrm{P}(S_i \geqslant t)[1 - \mathrm{P}(S_i \geqslant t)]\} \tag{3.3.9}$$

而且等式(3.3.8)中的积分元遍历 S_i 的所有可能取值。这个公式(由 Yizhaki (1996)得到)对 S_i 的每个可能取值都赋予一个权重，该权重与大于 S_i 和小于 S_i 的值的条件期望之差成正比。与 S_i 的中位数越接近，被赋予的权重也越大，因为在此邻域中 $\mathrm{P}(S_i \geqslant t) \cdot [1 - \mathrm{P}(S_i \geqslant t)]$ 被最大化。

如果加入协变量 X_i，那么等式(3.3.8)就发生变化，对每个特定的 X_i，都有等式(3.3.8)。于是经过协变量 X_i 调整的表示加权平均值的等式(3.3.8)就变为：

$$\frac{\mathrm{E}[Y_i(S_i - \mathrm{E}[S_i \mid X_i])]}{\mathrm{E}[S_i(S_i - \mathrm{E}[S_i \mid X_i])]} = \frac{\mathrm{E}\left[\int h'_X(t)\mu_{iX}\mathrm{d}t\right]}{E\left[\int \mu_{iX}\mathrm{d}t\right]} \tag{3.3.10}$$

其中，$h'_X(t) = \dfrac{\partial \mathrm{E}[Y_i \mid X_i,\ S_i = 1]}{\partial t}$ 并且：

$$\begin{aligned}\mu_{iX} \equiv &\{\mathrm{E}[S_i \mid X_i,\ S_i \geqslant t] - \mathrm{E}[S_i \mid X_i,\ S_i < t]\} \\ &\times \{\mathrm{P}(S_i \geqslant t \mid X_i)[1 - \mathrm{P}(S_i \geqslant t \mid X_i)]\}\end{aligned}$$

等式(3.3.10)反映出两类平均化的过程：一个积分是在固定协变量值后对非线性的条件期望函数进行平均化，另一个期望是在不同的协变量之间进行平均化。这里需要注意的是，对于那些使得 $\mathrm{P}(S_i \geqslant t \mid X_i)$ 等于 0 或 1 的协变量 X_i 所决定的子集而言，总体回归系数没有包含 S_i 对这些变量的影响。还值得注意的是，如果 S_i 是一个虚拟变量，那么我们可以从更一般化的模型(3.3.10)中得到等式(3.3.7)。

Angrist 和 Krueger(1999)用出生地和出生年份作为协变量为研究教育回报的回归模型构造了加权平均函数。虽然等式(3.3.8)和等式(3.3.10)看上去比较神秘，或者说至少不是那么显而易见，但是在我们举的这个例子中用于平均的权数 $\mathrm{E}[\mu_{iX}]$ 看上去是一个很好的具有平滑作用的关于 t 对称的函数，并以 S_i 的矩为中心。

对于具有给定回归元分布的模型，等式(3.3.8)和等式(3.3.10)的含义还可以进一步挖掘。比如假设 S_i 是正态分布的。令 $z_i = \dfrac{S_i - \mathrm{E}(S_i)}{\sigma_S}$，其中，$\sigma_S$ 是 S_i 的标准差，因此 z_i 是标准正态分布。于是：

$$\begin{aligned}\mathrm{E}[S_i \mid S_i \geqslant t] &= \mathrm{E}(S_i) + \sigma_S \mathrm{E}\left[z_i \mid z_i \geqslant \frac{t - \mathrm{E}(S_i)}{\sigma_S}\right] \\ &= \mathrm{E}(S_i) + \sigma_S \mathrm{E}[z_i \mid z_i \geqslant t^*]\end{aligned}$$

由截尾正态分布公式(Johnson and Kotz, 1970),我们知道:

$$\mathrm{E}[z_i \mid z_i > t^*] = \frac{\phi(t^*)}{[1-\Phi(t^*)]} \text{ 以及 } \mathrm{E}[z_i \mid z_i < t^*] = \frac{-\phi(t^*)}{\Phi(t^*)}$$

其中,$\phi(\cdot)$和$\Phi(\cdot)$都是标准正态分布的密度函数和分布函数。将这两个概率公式代入等式(3.3.9)可得:

$$\mu_t = \sigma_S\left\{\frac{\phi(t^*)}{[1-\Phi(t^*)]} - \frac{-\phi(t^*)}{\Phi(t^*)}\right\}[1-\Phi(t^*)]\Phi(t^*) = \sigma_S\phi(t^*)$$

因此我们有:

$$\frac{\mathrm{cov}(Y_i, S_i)}{V(S_i)} = \mathrm{E}[h'(S_i)]$$

换言之,当S_i是正态分布时,对Y_i关于S_i做回归得到的就是导数$h'(S_i)$的平均值。当然,这个特殊结果是一个特殊例子的产物①。当然,我们知道正态性的假设并非常常正确。而且,从我们做计量经济学的经验出发,从参数型非线性模型中构造出的平均导数(也被称为"边际效应")往往和线性回归结果的差别不大。我们在第3.4.2节对这一问题进行详细阐述。

3.3.2 用倾向评分控制协变量

回归理论中最重要的结论就是遗漏变量偏误公式(OVB),它告诉我们如果遗漏掉的变量与我们纳入回归模型的变量无关,那么是否加入这个遗漏变量不影响回归方程中已有变量的系数。通过Rosenbaum和Rubin(1983)发展出的倾向评分定理,我们可以将这个观点推广到以匹配法作为研究策略的更加一般化的情况,其中我们感兴趣的因果变量是表示处理与否的二值变量②。

倾向评分定理指的是:给定多元协变量构成的向量X_i,如果潜在结果与处理状态独立,那么给定协变量向量的某个值函数,潜在结果与处理状态仍然相互独立,这里协变量向量的值函数被称为倾向得分,定义为$p(X_i) \equiv \mathrm{E}[D_i \mid X_i] = \mathrm{P}[D_i = 1 \mid X_i]$。正式的定理是:

定理3.3.1:倾向评分定理(The Propensity Score Theorem)。

若条件独立假设成立,也就是$\{Y_{0i}, Y_{1i}\} \perp\!\!\!\perp D_i \mid X_i$。那么$\{Y_{0i}, Y_{1i}\} \perp\!\!\!\perp D_i \mid p(X_i)$。

① 同样思路下的其他特殊例子可见Yitzhaki(1996)和Ruud(1986),他们考虑了受限被解释变量不依赖于分布的估计量。

② 虽然仍需得到进一步理解,但我们还是可以将倾向得分方法推广到多值处理变量。这一方向上的努力可见Imbens(2000)。

证明：我们只要证明 $P[D_i = 1 \mid Y_{ji}, p(X_i)]$ 不依赖于 Y_{ji}，$j = 0, 1$ 即可：

$$\begin{aligned} P[D_i = 1 \mid Y_{ji}, p(X_i)] &= E[D_i \mid Y_{ji}, p(X_i)] \\ &= E\{E[D_i \mid Y_{ji}, p(X_i), X_i] \mid Y_{ji}, p(X_i)\} \\ &= E\{E[D_i \mid Y_{ji}, X_i] \mid Y_{ji}, p(X_i)\} \\ &= E\{E[D_i \mid X_i] \mid Y_{ji}, p(X_i)\} \end{aligned}$$

最后一个等号来自于条件期望独立假设。再由 $E\{E[D_i \mid X_i] \mid Y_{ji}, p(X_i)\} = E\{p(X_i) \mid Y_{ji}, p(X_i)\}$，这显然就是 $p(X_i)$。

类似于回归中的遗漏变量偏误公式，倾向评分定理指出我们只要将影响处理概率的协变量控制住就好。但实际上这个定理还能让我们走得更远：我们唯一需要控制的协变量就是处理概率本身。在实际操作中，往往需要分两步来使用倾向评分定理进行估计：首先，用类似于 logit 或 probit 等参数模型来估计 $p(X_i)$，然后运用匹配法对处理效应进行估计。在第二步的时候既可以依赖于第一步估计出的得分进行匹配，也可以使用下面提到的加权平均法（见 Imbens(2004)的一个综述）进行匹配。

对于直接用倾向评分进行匹配的方法和我们之前提到的用协变量进行匹配的方法类似，不同之处在于现在用倾向评分而不是直接用协变量来进行匹配。由倾向评分定理和条件期望假设可知：

$$\begin{aligned} &E[Y_{1i} - Y_{0i} \mid D_i = 1] \\ &= E\{E[Y_i \mid p(X_i), D_i = 1] - E[Y_i \mid p(X_i), D_i = 0] \mid D_i = 1\} \end{aligned}$$

于是可以通过两种方法估计被处理样本的处理效果，一种是估计出 $p(X_i)$，然后根据 $p(X_i)$ 对被处理样本进行分层，并用每一层的样本条件期望代替公式中（花括号内）的期望算子，另一种方法是在具有相似评分值的那些控制变量对应的个体中对被处理个体进行匹配（Dehejia 和 Wahba(1999)同时使用了这两种方法）。当然，基于模型或者使用非参数方法估计出的 $E[Y_i \mid p(X_i), D_i]$ 也可以代入条件均值函数，并将花括号外的那个期望算子变成连加符号就可以了（正如 Heckman, Ichimura 和 Todd(1998)的处理方式）。

还可以有更加巧妙的方法来对倾向评分估计值进行加权平均，这个方法避免了匹配过程。其原理在于条件独立假设意味着 $E\left[\frac{Y_i D_i}{p(X_i)}\right] = E[Y_{1i}]$ 和 $E\left[\frac{Y_i(1-D_i)}{(1-p(X_i))}\right] = E[Y_{0i}]$ ①。可知，给定在某种方法得到的对 $p(X_i)$ 的估计，我们可以运用下面的样本估计量来估计平均的处理效应：

---

① 为了看清楚这一点，我们对 X_i 求迭代期望：$E\left[\frac{Y_i D_i}{p(X_i)}\right] = E\left\{E\left[\frac{Y_i D_i}{p(X_i)} \middle| X_i\right]\right\}$；$E\left[\frac{Y_i D_i}{p(X_i)} \middle| X_i\right] = \frac{E[Y_i \mid D_i = 1, X_i] p(X_i)}{p(X_i)} = E[Y_{1i} \mid D_i = 1, X_i] = E[Y_{1i} \mid X_i]$。

$$\mathrm{E}[Y_{1i}-Y_{0i}]=\mathrm{E}\left[\frac{Y_iD_i}{p(X_i)}-\frac{Y_i(1-D_i)}{1-p(X_i)}\right]$$
$$=\mathrm{E}\left[\frac{(D_i-p(X_i))Y_i}{p(X_i)(1-p(X_i))}\right] \tag{3.3.11}$$

上式最后一个等号后的表达式正是由 Newey(1990)以及 Robins, Mark 和 Newey (1992)得到的结果。因此我们可以很简单地通过样本估计量

$$\mathrm{E}[Y_{1i}-Y_{0i}\mid D_i=1]=\mathrm{E}\left[\frac{(D_i-p(X_i))Y_i}{(1-p(X_i))\mathrm{P}(D_i=1)}\right] \tag{3.3.12}$$

计算被处理者的处理效应。用样本被选择的概率进行加权平均以调整非随机抽样带来的偏误的思想可以追溯到 Horvitz 和 Thompson(1952)。当然,为使该方法可用及被估计参数满足一致性,我们需要得到 $p(X_i)$的一致估计。

由于 Horvitz-Thompson 给出的倾向评分法无需繁琐的匹配,从本质上讲估计值是自动获得的,所以该方法具有相当的吸引力。Horvitz-Thompson 方法还凸显出运用倾向评分法进行匹配与回归之间的紧密联系,这正是我们在第 3.3.1 节讨论回归和匹配之间紧密关系时所关注的。再次考虑控制了以饱和模型的形式出现的协变量后,在总体回归方程中对 Y_i 关于 D_i 做回归后得到的估计值 δ_R,这个估计值可以写作:

$$\delta_R=\frac{\mathrm{E}[(D_i-p(X_i))Y_i]}{\mathrm{E}[p(X_i)(1-p(X_i))]} \tag{3.3.13}$$

与 Horvitz-Thompson 有关的两个匹配估计量(3.3.11)和(3.3.12)以及回归估计值(3.3.13)都属于由 Hirano, Imbens 和 Ridder(2003)给出的加权平均估计类中的一种,这个估计类可以写为:

$$\mathrm{E}\left\{g(X_i)\left[\frac{Y_iD_i}{p(X_i)}-\frac{Y_i(1-D_i)}{(1-p(X_i))}\right]\right\} \tag{3.3.14}$$

其中,$g(X_i)$是一个已知的权重函数。(如果我们需要从估计量得到估计值,那么用 $p(X_i)$的一致估计值替换 $p(X_i)$,用连加号代替期望算子。)如果想得到平均处理效果,可令 $g(X_i)=1$;如果想得到被处理者的平均处理效果,令 $g(X_i)=\frac{p(X_i)}{\mathrm{P}(D_i=1)}$;对回归而言,则可以令:

$$g(X_i)=\frac{p(X_i)(1-p(X_i))}{\mathrm{E}[p(X_i)(1-p(X_i))]}$$

这种相似性再次凸显出回归和匹配——包括倾向评分匹配——相互之间没有太多的差异,至少我们在对某个模型制定倾向评分函数的具体形式之前,它们都是相同的。

这里存在的大问题就是如何对 $p(X_i)$建模并进行估计,或者说指出在估计 $\mathrm{E}[Y_i\mid p(X_i),\ D_i]$时要平滑和分层到何种程度才可以,特别是当协变量为连续变

量时这个问题更加突出。对回归而言，这个问题转化为如何确定控制变量的参数形式（比如当协变量离散时，要用多项式还是用主效应加交互项）。这个问题的答案与我们面对的实际情况有关。在经验研究文献中不断发展的一部分文献指出在实际中用 logit 模型可对存在多项式的倾向评分进行很好的逼近，但这是规律，还无法上升到定理层面，而且无可避免地要使用一些实验来对这个方法进行验证（Dehejia and Wahba，1999）[①]。

一些还在不断完善的理论开始对使用倾向评分方法的有效性进行讨论，并得到了一些具有启发性的定理。首先，从渐进有效的角度来看，相比于使用协变量，使用倾向评分方法进行匹配总会带来一些有效性的损失。不论某个可以解释结果的协变量是否出现在倾向评分中，使用该协变量进行匹配得到的渐进标准误都会更低。我们从 Hahn(1998)的研究得到了这个结果，Hahn(1998)考察了在条件独立假设下，包含倾向评分对处理效果估计精确性的影响。比如在 Angrist(1998)中，尽管军队服役年限与出生年份无关，但是收入水平和出生年份有关，所以关于出生年份进行匹配使得估计有效性得到很大的提高。在回归中这个看法转化为：即使回归中不存在遗漏变量偏误，不论遗漏变量是否对结构有预测能力，长回归得到的对参数的估计都要比短回归更加精确（见第 3.1.3 节）。

Hahn(1998)的研究结果提出如下问题：我们为什么要不厌其烦地使用倾向评分法来估计参数。一个哲学角度的回答是：倾向评分法正确地将研究者的关注点聚焦在模型如何分配处理上，而不是将关注点聚焦在决定结果的更加复杂和神秘的过程上，在这个方面研究者拥有信息优势。当如何分配处理是人类制度或者政府规制的结果时，这个观点看上去更加令人信服，因为决定结果的实际过程可能更加神秘（比如类似于市场一类的东西）。比如在运用时间序列对货币政策进行的评估中，Angrist 和 Kuersteiner(2004)指出相比于决定 GDP 的过程，我们对美联储如何决定利率的过程要了解得更多。同样的道理，查验一个模型是如何分配处理的要比查验一个模型如何决定结果容易得多（这个说法的另一个版本请见 Rosenbaum 和 Rubin(1985)）。

用统计语言表达的更加精确地支持倾向评分法的讨论来自于 Angrist 和 Hahn(2004)。这篇论文指出即使基于倾向评分的估计值不是渐进有效的，在有限样本中使用这种方法也可以提高估计精度。既然所有来自真实世界的数据都是有限的，那么这种性质自然有其在经验研究上的作用。从直觉上讲，如果倾向评分中遗漏的协变量对结果的变化的解释力度很小（从纯粹统计意义出发），那么与其背上需要估计这种效应的重负，不如将其忽视。在使用诸如 NLSY 数据集进行的研究中，由于该数据集有上百个可预测结果的协变量，所以很容易在这里理解这种处理方法所蕴含的道理。在实际中，我们只关注一小部分协变量。对这些协变量的

① Andrea Ichino 和 Sascha Becker 已经在网络上公布了可以完成多种匹配任务的程序，见 Becker 和 Ichino(2002)。

选择往往基于我们想要预测什么样的处理。

最后，Hirano，Imbens 和 Ridder(2003)为在 Hahn(1998)中提出的“倾向评分悖论”提供了一个具有渐进性质的解。他们指出即便基于已知的倾向评分对处理效果的估计是无效的，对于具有连续协变量的模型而言，当使用非参数方式对评分进行估计并以此设定加权方式时，类似于 Horvitz-Thompson 式的加权估计值也是有效的。对 Hirano，Imbens 和 Ridder(2003)而言，倾向评分是被估计出来的，而且是使用非参数的方法估计出来的，这对他们的结论至关重要。

Hirano，Imbens 和 Ridder(2003)的结论是否解决了“倾向评分悖论”？就目前而言，我们更倾向于 Angrist 和 Hahn(2004)在有限样本下得到的结论。他们的研究强调了如下事实：正是因为研究者愿意对倾向评分加上一些限制，才使得基于倾向评分的推断有了概念上和统计上的威力。比如 Angrist(1998)就在高维离散协变量下运用了倾向评分法，对评分的无限制非参数估计值正好就是每个协变量确定的个体中被处理的概率。用这个非参数估计值代替 $p(X_i)$，很显然从代数的角度来看，等式(3.3.11)和等式(3.3.12)的样本估计值都等价于相应的居于所有协变量的匹配估计值。因此，既然基于完全协变量的匹配估计值是渐进有效的，那么基于倾向评分得到的估计值也应该是有效的。倾向评分法中的一个重要特点在于我们可以使用先验知识来降低协变量的维度。从统计学上来讲这样做的好处是对有限样本结果有所改进。如果你不准备平滑，限制或者用其他方法来降低匹配问题中协变量的维度(虽然在实际的经验研究中这个问题会产生相当的后果)，那么你应该使用基于所有协变量的匹配，或者使用饱和回归来控制协变量。

3.3.3 倾向评分模型与回归

倾向评分模型将我们的注意力从估计 $\mathrm{E}[Y_i \mid X_i, D_i]$ 转移到估计倾向评分 $p(X_i) \equiv \mathrm{E}[D_i \mid X_i]$。在实际运用中后者更加吸引人，因为它便于模型化。比如，Ashenfelter(1978)指出政府资助培训项目的参与者收入显著的偏低，之后的研究也多次发现这种情况。如果这种偏低的收入是项目参与者区别于他人的唯一因素，那么我们通过控制这些人过去收入的变化来估计该项培训对项目参与者收入的因果效应。在实际中，由于收入的历史数据是连续的、多维的，所以很难基于过去的收入进行匹配。Dehejia 和 Wahba(1999)在他们的论文中指出用倾向评分进行控制后得到的估计结果要好于用项目参与者的收入历史进行控制的结果。

在 Dehejia 和 Wahba 论文中，他们用随机实验得到的估计值作为比较的基准，然后发现用倾向评分方法得到的估计值显著地接近随机实验中得到的结果。不过我们仍然认为回归应该成为大部分经验研究项目的起点。毫无疑问，这个观点无法成为一个定理，因为在很多情况下，用倾向评分进行匹配为我们提供了一个更加可靠的对平均处理效果的估计。我们没有热情拥抱倾向评分法的第一个原因是基于对实际操作的考虑：在运用倾向评分方法进行匹配时有太多细节需要考虑，比如

如何模型化评分，如何进行推断，这些细节目前还没有形成一套标准。因此即使运用相同的数据和协变量，不同的研究者也可能得到相当不同的结果。更进一步，正如我们在 Horvitz-Thompson 估计量中看到的，基于倾向评分的加权平均和回归并无本质区别。如果回归模型的协变量相当灵活，比如接近于饱和模型，那么我们可以将回归看作某种使用倾向评分进行加权的方法，于是倾向评分和回归之间的区别在于如何执行这两种方法。在实际中，你所使用的回归模型可能远未饱和，但是如果使用的协变量恰当，那么问题就不是很大。

我们在这里使用与 Dehejia 和 Wahba(1999)[①]相同的数据集来阐述倾向评分和回归间的不同，这个数据集叫做国家支持的工作计划，简记为 NSW(National Supported Work)。NSW 是 20 世纪 70 年代中期实施的一个项目，该项目为劳动能力缺乏者提供工作经验。与之前不同的是对这个项目的评估使用的是随机实验的方法。Lalonde(1986)在其开创性分析中对两种估计结果进行比较，一种是来自 NSW 随机实验的估计值，另一种是来自 PSID 和 CPS 的对非随机控制组得到的计量结果。由于非实验方法产生的结果之间差别很大，而且与相应的随机实验结果相去甚远，所以他得到的结论比较悲观。而且他还指出不知道随机实验结果的客观的研究者可能无法选择最好的计量方法，也不知道如何选择基于观察得到的控制组。

在 Lalonde(1986)之后，Dehejia 和 Wahba(1999)发现用倾向评分法选择可被观察的控制组，基于此使用匹配法对 NSW 的处理组估计的处理效应很接近随机实验的结果，并用多个对照组来说明这一发现。沿着 Dehejia 和 Wahba(1999)的思路，我们再次考虑 CPS 中的两个对照组，其中一个是没有被选择的样本(记为 CPS-1)，另一个是从最近的失业样本中选择出来的进行比较的组别(记为 CPS-3)。

表 3.3(第一列到第四列是从 Dehejia 和 Wahba(1999)复制过来的)报告了 NSW 处理组、随机选择的 NSW 控制组和我们基于可观察控制变量选取的某个对照组的描述性统计。相比于来自于 CPS-1 样本的总体特征而言，NSW 处理组和随机选择的 NSW 控制组表出现其中的样本年龄更小、接受的教育更少、更可能是非白人并且收入也低于来自 CPS-1 总体的均值。CPS-3 样本显示出与 NSW 处理组更匹配，但还是表现出一些不同，特别是在种族和参与项目前的收入方面。

表 3.3.3 报告了对 NSW 中处理效应的估计值。被解释变量是 1978 年的年收入，这一年正好处在接受培训后的一年或两年之后。不同行表示的是使用不同控制变量后的估计值，分别是：不使用控制变量；使用表 3.3.2 中所有与人口统计学有关的变量做控制变量；将 1975 年的收入作为滞后变量进行控制；所有与人口统计学有关的变量加上滞后的收入变量；所有与人口统计学有关的变量加上 1974 年和 1975 年两年的收入作为滞后变量。所有的估计结果都是用表示是否接受培训

① 一个扩展了的比较倾向得分方法之间的不同的研究请见 Smith 和 Todd(2005)与 Dehejia(2005)之间的一次争论。

表 3.3 在 NSW 中协变量的均值和可观察控制组样本

变量	NSW		完全比较样本		用 p 得分值进行筛选后的样本	
	被处理的 (1)	控制组 (2)	CPS-1 (3)	CPS-3 (4)	CPS-1 (5)	CPS-3 (6)
年　龄	25.82	25.05	33.23	28.03	25.63	25.97
受教育年限	10.35	10.09	12.03	10.24	10.49	10.42
黑　人	0.84	0.83	0.07	0.20	0.96	0.52
西班牙语系	0.06	0.11	0.07	0.14	0.03	0.20
退　学	0.71	0.83	0.30	0.60	0.60	0.63
已　婚	0.19	0.15	0.71	0.51	0.26	0.29
1974 年的收入	2 096	2 107	14 017	5 619	2 821	2 969
1975 年的收入	1 532	1 267	13 651	2 466	1 950	1 859
观测值个数	185	260	15 992	429	352	157

注：来自于 Dehejia 和 Wahba(1999)中的表 1。表中前四列对应的样本已经在 Dehejia 和 Wahba(1999)中得到描述。最后两列对应的样本只比较了倾向评分在 0.1 和 0.9 下的两个组别。在估计倾向得分时使用了表中列出的所有协变量。

的虚拟变量加上一系列的控制变量对 1978 年收入所做的回归(不包含控制变量下的处理组—控制组之间的差别报告在第一行)后得到的。

由实验得到的控制组的估计结果报告在第一列，结果处在 1 600 美元到 1 800 美元之间。令人惊讶的是，虽然不同行表示对回归的不同设定，但是在随机实验下估计出的结果没有大的变化。相比之下，在第二列中报告的 NSW 参与者和 CPS-1 样本之间的收入差距超过了－8 500 美元，这意味着 CPS-1 中存在严重的选择偏误。将标志人口特征的控制变量和滞后的收入变量加入回归后在很大程度上减少了估计值与随机实验结果之间的差距；在最后一行中，估计出来的处理效果达到了正的 800 美元。在第三列中的结果表现得更好，该结果是使用更加细致的 CPS-3 比较组得到的。这个组别中的特点与 NSW 中参与者的特点更加接近，与之相吻合的是不加控制变量时人们的收入差别只有－635 美元。最后一行使用所有的控制变量后的估计结果接近 1 400 美元，与通过随机实验计算得到的处理效应的距离已经不远了。

我们从 CPS-1 出发去选择更加合理的 CPS-3 的过程存在着缺点，这个缺点就是在更加仔细地构造更小的 CPS-3 对照组时用到的规则过于特殊。挑选 CPS-3 的规则来自于 NSW 项目的选择规则，这一规则对低收入和劳动能力弱的人有特别关照，但是在实际中选择样本的方式可以有很多种。因此我们更希望有一个系统化进行筛选的方法。在最近的一篇论文中，Crump，Hotz，Imbens 和 Mitnik (2009)指出可以使用倾向评分作为一种系统性的样本选择方式，以此作为回归估计的前期准备。这项研究也同样印证了我们之前的讨论：倾向评分是估计值的基础。

表 3.4　在不同控制变量下对 NSW 中培训效果的回归估计值

设　定	对样本的完全比较			用 p 得分值筛选后的样本	
	NSW (1)	CPS-1 (2)	CPS-2 (3)	CPS-1 (4)	CPS-3 (5)
普通的比较	1 794	−8 498	−635		
	(633)	(712)	(657)		
用人口特征做控制变量	1 670	−3 437	771	−3 361	890
	(639)	(710)	(837)	(811)	(884)
				[139/497]	[154/154]
1975 年的收入	1 750	−78	−91	无观测值	166
	(632)	(537)	(641)	[0/0]	(644)
					[183/427]
用人口以及 1975 年的收入做控制变量	1 636	623	1 010	1 201	1 050
	(638)	(558)	(822)	(722)	(861)
				[149/357]	[157/162]
用人口以及 1974 和 1975 年的收入做控制变量	1 676	794	1 369	1 362	649
	(639)	(548)	(809)	(708)	(853)
				[151/352]	[147/157]

注：本表使用 Dehejia 和 Wahba(1999)中的数据，报告了在多种控制变量下对受培训带来的效果的回归估计。人口统计学方面的控制变量分别是年龄、受教育时间、标志是否为黑人、是否在西班牙语系、是否高中辍学以及是否已婚的虚拟变量。标准误报告在括号中。观察到的样本个数报告在方括号中[受处理/控制组]。在 CPS-1 数据中仅使用 1975 年收入数据作为协变量时，在倾向评分区间[0.1，0.9]中没有观察值。

我们使用 Crump 等(2009)的建议，首先在混同的 NSW 中选择处理组和机遇可观察变量构造对照组，并由此计算倾向评分，然后选择 $0.1 < p(X_i) < 0.9$ 的那些观察值。换句话说，我们对估计时使用的样本施加了限制，只有那些被处理概率在 10%—90%的观察值才会进入回归。这就保证了在协变量各种取值所决定的组别中，只把那些既有处理又有控制的组别纳入回归分析。因此用这些已被筛选过的样本进行的回归时我们无需再去考虑“共同支撑”问题——换言之，就是那些处理组和控制组之间的协变量不存在交叠的组别。依据倾向评筛选出的样本(对倾向评分的估计使用了表中所列的所有协变量)的描述性统计请见表 3.3 的最后两列。相比于未经过筛选的样本，在经过删选的 CPS-1 和 CPS-3 样本中协变量均值与第一列中表示的 NSW 样本中的均值已经十分接近。

随后我们还用另外一组协变量构造可以筛选共同支撑的方法，但是在筛选和使用迭代期望计算平均处理效应时仍然使用相同的协变量。由此得到的估计值在表 3.4 的最后两列给出。如果单独控制人口特征或者是滞后收入，由此得到的估计值在第二列和第三列给出，两者差别很小。相比于没有经过筛选的样本，同时控制人口特征和滞后收入后对经过筛选的样本 CPS-1 进行估计得到的结果与实验结果更加接近。将两个滞后收入变量都控制后针对经过筛选的样本 CPS-1 进行估计

得到的结果与实验结果也很接近。另外一方面，将滞后一期的收入加入控制并筛选具有共同支撑的样本后得到的估计值只在很小的程度上改进了估计结果，将滞后两期的收入加入控制后得到的结果反而变差了。

上面的这些讨论加强了我们使用回归的信心（当然使用回归的信心本来就已经很强了）。在 CPS-1 的例子中，在回归中使用正确的控制变量已经可以很好地剔除选择偏误了。使用我们对培训项目筛选标准的已有知识来对进入回归的样本进行限制后得到的回归结果更好，几乎可与 Dehejia 和 Wahba(1999)将两个滞后收入变量控制住后使用倾向评分匹配法得到的结果相媲美。对量大且粗糙的 CPS-1 样本使用系统性筛选以保证共同支撑的过程可以看作是回归估计的一个有用的助手。在经过筛选的 CPS-1 中得到的估计结果几乎和没有经过筛选的 CPS-3 估计结果一样好。但是我们注意到，没有对倾向评分法下得到的标准误进行调整，以反映我们在估计得分时的取样方差。正如通过 CPS-1 构造 CPS-3 的步骤所显示的，使用先验信息对样本进行筛选的优点在于我们无需再去构造 CPS-3。

3.4 回归的细节

3.4.1 加权回归

对于应用研究者而言，没有什么比确定样本权重更加令人困扰了。即使到现在，距我们拿到博士学位已有 20 年，我们还是对 Stata 手册中关于加权回归的相应章节感到不甚满意。在很多方面权重都有其用处，而且如何使用权重确实在很大程度上会影响你的结果。可惜的是，支持或反对使用加权平均的各种观点尚未达成一致，比如在如何计算权重这一问题上就存在种种争议。对加权平均正反两方面意见的详细讨论已经超越了本书的范围。对此可以参见 Pfefferman(1993)和 Deaton(1997)提供的两种思路。在这一小节中，我们为如何使用加权回归提供一些指导性的原则。

对回归进行加权平均的一个简单经验就是当加权可使你估计的数值更加接近总体的相应值时，你就应该使用加权回归。比如，我们的估计目标（或者说被估量）是总体回归方程，但是用来进行估计的样本是非随机的，那么其抽样权重就是 w_i，等于观察值 i 被抽样到的概率的倒数，那么这时用权重为 w_i 的加权最小二乘法就很有道理（你可以在 Stata 中使用 pweights 或者在 SAS 中使用 weight 命令来完成这个加权最小二乘计算）。使用观察值被抽取到的概率的倒数做权重可以为我们带来对总体回归方程的一致估计，即使你手头所使用的样本不是简单随机抽样得到的。

与加权回归相关的另一类问题是分组数据。假设你可能想了解总体回归向量 $\beta=\mathrm{E}[X_iX_i']^{-1}\mathrm{E}[X_iY_i]$，因此你在随机样本中对 Y_i 关于 X_i 做回归，但是如果你没有随机抽取的样本，有的只是按照 X_i 的各种取值分组的数据。也就是说你必须针对每个 x，运用来自随机抽样的数据去估计 $\mathrm{E}[Y_i \mid X_i = x]$。不妨记这个均值为

$\bar{y}_x$，而且你还知道 n_x，其中 n_x/N 就是在相应的随机样本中 x 出现的频率。正如我们在 3.1.2 节中指出的，用 n_x 做权重，对 $\bar{y}_x$ 关于 x 做加权回归得到的估计值与直接对随机抽样的微观数据做回归得到的结果是相同的。因此，如果你的目标是回到微观数据进行回归，那么用组规模作为权重就显得很有意义。不过我们注意到习惯于针对已发布数据（比如人均 GDP）进行分析的宏观经济学家往往会忽略相应的微观数据，可能会不同意在计算平均值时进行加权，或者从原则上同意加权平均的意义但在实际中还是坚持老一套的研究方法，这种研究方法更喜欢对加总数据进行无加权处理。

另一方面，类似很多教科书的处理方法，如果认为进行加权回归的唯一原因在于异方差性，那么我们（意指我们计量经济学家）可能比宏观经济学家还不愿意去使用加权平均。异方差的意义下进行加权的理由基本上可以这样叙述：假设我们感兴趣的是线性条件期望函数，$\mathrm{E}[Y_i \mid X_i] = X_i'\beta$。定义残差项为 $e_i \equiv Y_i - X_i'\beta$，它可能是异方差的。也就是说条件方差函数 $E[e_i^2 \mid X_i]$ 未必是个常数。在这个例子中，即使总体回归方程仍然是 $\mathrm{E}[X_iX_i']^{-1}\mathrm{E}[X_iY_i]$，其样本估计量也是无效的。对线性条件期望函数更加精确的估计就是使用加权最小二乘（WLS）——也即是说通过最小化以 $\mathrm{E}[e_i^2 \mid X_i]^{-1}$ 为权重的均方误后得到的估计结果。

正如在 3.1.3 节了解到的，线性概率模型（记为 LPM）——就是 Y_i 取虚拟变量的那种模型——必定是异方差的。假设条件期望函数实际上是线性的，如果模型是饱和的，那么该模型就是条件期望函数，于是有 $\mathrm{P}[Y_i = 1 \mid X_i] = X_i'\beta$ 和 $\mathrm{E}[e^2 \mid X_i] = X_i'\beta(1 - X_i'\beta)$，显然它们都是 X_i 的函数。这是模型本身即存在异方差性的例子。通过估计相应的回归方程，我们可以很容易地估计条件方差函数。线性概率模型的有效加权最小二乘估计——广义最小二乘估计（GLS）的特例——就是用 $[X_i'\beta(1-X_i'\beta)]^{-1}$ 做权重。因为假设条件期望函数是线性的，所以一旦最小二乘估计量得到，权重即刻得到。

有两个原因来说明为什么在这个例子中我们不会选择进行加权（虽然我们会用异方差一致性标准误）。首先，在实际中对 $\mathrm{E}[e_i^2 \mid X_i]$ 的估计可能不够理想。如果条件方差模型对条件方程函数的估计很差或者在估计过程中充满了噪音，那么加权最小二乘估计值的有限样本性质可能比未加权的情况还差。因此依赖于渐进性质作出的推断就可能很有问题，为了提高估计效率而使用加权最小二乘估计的想法便可能无法实现①。其次，如果条件期望函数不是线性的，对于最小二乘法估计不准确的变量，加权最小二乘法也爱莫能助。但是从另一个角度看，不加权的估计值得到的结论却易于解释：它在最小均方误的意义下是为我们提供了对总体条件期望函数的最好估计。

加权最小二乘估计值也还是为我们提供了某一类的近似，但是这种近似依赖于对权重的选择。至少，这种有赖于权重的估计值使得你的估计结果很难与其他

① Altonji 和 Segal（1996）在广义矩的背景下讨论了这一观点。

研究者的结果进行比较，并为如何设定权重打开了相互之间进行争议的空间。最后，我们需要记住一个古老的警告：如果东西还没坏，就别急着去修(if it ain't broke, don't fix it，言下之意就是不要破坏现状)。即使存在异方差性，我们对总体回归向量的解释也不受影响，因此为什么要担心异方差性呢？通过加权最小二乘法对估计效率做出的任何改进都是有限的，但是错误估计出的权重弊大于利。

3.4.2 有限被解释变量与边际效应

很多经验研究中使用的被解释变量都只取有限的一些值。Angrist 和 Evans (1998)关于生育孩子对妇女劳动力供给影响的讨论就是这样一个例子，这个例子在工具变量这一章还会出现。这项研究考虑的是生育子女对父母工作和收入的因果效应。因为生育小孩可能和潜在的收入水平相关，所以 Angrist 和 Evans 报告了基于兄妹性别组成和多胎生育的工具变量估计值，同时还报告了最小二乘估计结果。在这项研究中，几乎所有的被解释变量都是二元变量(比如表示雇佣状态的变量，只取 0 和 1，表示工作或失业)或者非负变量(比如表示工作时间，已工作周数和收入的变量)。被解释变量取值有限这件事情会影响经验研究实践吗？很多计量经济学教科书指出最小二乘适合于估计被解释变量是连续变量的情形，当我们感兴趣的被解释变量取值有限时(Limited dependent variable，简写为 LDV)，线性回归模型便显得不太适宜，这时用类似于 probit 和 Tobit 等非线性模型会更好。相比之下，我们认为有限被解释变量带来的问题并不严重，因为回归的合法性来自于它和条件期望函数之间的紧密关系。

如以往一样，随机实验可以作为我们讨论的基准，这时我们通过回归可以得到一个简单的处理组—控制组差别。比如在兰德健康保险实验(RAND Health Insurance Experiment，简写为 HIE；见 Manning 等(1987))中考虑对多个被解释变量进行的回归，其中回归元被随机分配，代表不同的处理组。在这项雄心勃勃的实验 ——也可能是美国社会科学领域中成本最高的研究中，兰德公司成立了一家不收保费的小型保险公司。在这项研究中将近有 6 000 参与者被随机分配了具有不同特点的保险计划。

任何保险计划都必不可少的一个条款是规定被保险者支付医疗费用的比例。兰德健康保险实验随机地为个体提供不同的保险计划。其中一个计划提供完全免费的医疗，其他的保险计划则包含各种规定，这些规定包括共同支付比例，费用上限以及免责条款，因此被分配了这类保险计划的人则要为他的健康成本支付一定额度。这个实验的主要目的在于了解人们对医疗的使用是否考虑成本，如果是的话，它对健康有什么影响。兰德健康保险实验的结果指出那些被提供了免费医疗或者低成本医疗的人确实会更多地享受医疗服务，但对于其中大部分人而言，并未因此而提高健康水平。这项研究成果为成本敏感型的健康保险计划以及受管理的医疗服务提供了理论基础。

在兰德健康保险实验中的大部分被解释变量都是有限被解释变量。这些变量包括标志被试个体在给定年份中是否有过医疗支出或是否去过医院的虚拟变量，以及标志和医生面对面交流的次数以及医疗总花费(包括自付和保险公司支付)等非负的变量。在样本中医疗花费为零的人大约占 20%。在兰德健康保险实验中两个处理组的结果在表 3.5 中给出，这个表来自 Manning 等(1987)表 2 报告的结果。表 3.5 报告了医疗费用全免和需要扣除一定费用的两个组的平均结果。在后一个需要扣除一定费用的组中，当被试个体接受门诊治疗时，他每年需要自己支付 150 美元，以家庭为单位的话每年需要支付 450 美元，之外的费用由保险公司支付(对住院治疗，保险公司不要求被试个体进行支付)。在这两个组中总的样本规模是 3 000 多一点。

为了简化对有限被解释变量的讨论，假设我们只对费用全免和需要扣除一定费用的两个组别之间的差别感兴趣，其中个体被提供何种保险计划是随机分配的①。令 $D_i = 1$ 表示个体被分配了需要扣除一定费用的健康保险计划。由随机分配的原理，$D_i = 1$ 和 $D_i = 0$ 之间的不同就给出了平均处理效应。这时我们对随机实验进行讨论得到的结论(见第 2 章)：

表 3.5　HIE 的两个处理组中的平均结果

计　划	面对面访问	门诊花费 (1 984 $)	受理 (%)	内科治疗概率 (%)	住院治疗概率 (%)	总花费 (1 984 $)
免　费	4.55	340	12.8	86.8	10.3	749
	(0.17)	(10.9)	(0.7)	(0.8)	(0.5)	(39)
扣　除	3.02	235	11.5	72.3	9.6	608
	(0.17)	(11.9)	(0.8)	(1.5)	(0.6)	(46)
扣除减去免费	−1.53	−105	−1.3	−14.5	−0.7	−141
	(0.24)	(16.1)	(1.0)	(1.7)	(0.7)	(60)

注：本表来自于 Manning(1987)的表 2。所有的标准误(标记在括号中)都在跨期和家庭内部得到了调整。数字用 1984 年 6 月的美元计价。去医院的形式包括和医生面对面交流以得到治疗；为了接受放射治疗而独自去医院；麻醉，但是没有包括病理学的治疗。去医院治疗和医疗花费没有包括牙科治疗以及门诊的心理辅导。

$$\begin{aligned} &\mathrm{E}[Y_i \mid D_i = 1] - \mathrm{E}[Y_i \mid D_i = 0] \\ &= \mathrm{E}[Y_{1i} \mid D_i = 1] - \mathrm{E}[Y_{0i} \mid D_i = 1] \\ &= \mathrm{E}[Y_{1i} - Y_{0i}] \end{aligned} \tag{3.4.1}$$

得到上式的原因是 D_i 与潜在结果无关。而且如前所见，$\mathrm{E}[Y_i \mid D_i = 1] - \mathrm{E}[Y_i \mid D_i = 0]$ 正是对 Y_i 关于 D_i 做回归后得到的系数。

等式(3.4.1)意味着不论 Y_i 是二元变量、非负变量还是连续变量，都没有影响

① HIE 考虑的问题远比我们这里讨论的复杂。它包含了 14 个不同的处理，包括预先支付的类似于健康维护的服务。在实验设计中也没有采用简单的随机抽样方法，而是采用了更加复杂的分层分配的方法，以更好地平衡各组之间的差别。

在随机实验中对因果效应的估计。尽管面对不同的解释变量时，我们对方程(3.4.1)右边部分的解释会有所改变，但是在计算平均因果效应时我们没有借助任何的特殊处理。比如，兰德健康保险实验的一个结果是用以标志是否有过医疗花费的虚拟变量。既然这时实验结果是一个伯努里实验，那么我们有：

$$\begin{aligned} &\mathrm{E}[Y_{1i}-Y_{0i}]=\mathrm{E}[Y_{1i}]-\mathrm{E}[Y_{0i}] \\ &=\mathrm{P}[Y_{1i}=1]-\mathrm{P}[Y_{0i}=1] \end{aligned} \tag{3.4.2}$$

这种记号和我们之前讨论所用的记号不同，但是相应的计算过程没有任何改变。比如在兰德健康保险实验中，在不同实验组之间进行类似于等式(3.4.1)的比较后发现，在给定的年份中享受免费医疗的组里有87%的人至少使用了医疗服务，而那些需要扣除一定费用的组中只有72%的人使用过医疗服务。因此那些需要扣除一定费用的组中因为使用医疗服务而被扣除的150美元是导致这种不同结果的关键。于是这两个组接受医疗服的概率之差就是−0.15，是对 $\mathrm{E}[Y_{1i}-Y_{0i}]$ 的估计，其中 Y_i 是标志是否发生过任何医疗费用的虚拟变量。因为这里的被解释变量是一个虚拟变量，所以平均因果效应也是不同成本导致的个体接受医疗治疗概率的因果效应。

意识到标志是否接受医疗治疗的变量实际上表示的是概率，那么我们现在可以考虑用probit模型来近似条件期望函数。试一试总是没错的！在使用probit模型时，往往假设个体是否参与的决策由一个潜变量(latent variable)Y_i^* 决定，这个潜变量满足：

$$Y_i^*=\beta_0^*+\beta_0^* D_i-v_i \tag{3.4.3}$$

其中，v_i 的分布是 $N(0,\ \sigma_v^2)$。注意到这个潜变量不可能是真实的医疗花费，因为真实的医疗花费不可能为负，因此不是正态分布，而正态分布的随机变量则是在实数轴上连续分布并可以取负值。给定潜在得分模型(latent index model)：

$$Y_i=1[Y_0^*>0]$$

条件期望函数可以记为：

$$\mathrm{E}[Y_i\mid D_i]=\Phi\left[\frac{\beta_0^*+\beta_1^* D_i}{\sigma_v}\right]$$

其中，$\Phi(\cdot)$是正态分布的累积分布函数。因此：

$$\mathrm{E}[Y_i\mid D_i]=\Phi\left[\frac{\beta_0^*}{\sigma_{\text{概}}}\right]+\left\{\Phi\left[\frac{\beta_0^*+\beta_1^* D_i}{\sigma_v}\right]-\Phi\left[\frac{\beta_0^*}{\sigma_v}\right]\right\}D_i$$

这是关于 D_i 的线性函数，因此在线性回归中用对 Y_i 关于 D_i 做回归得到的斜率系数正是用probit模型做拟合后的差：$\Phi\left[\frac{\beta_0^*+\beta_1^* D_i}{\sigma_v}\right]-\Phi\left[\frac{\beta_0^*}{\sigma_v}\right]$。但是probit模型的系数 $\frac{\beta_0^*}{\sigma_v}$ 和 $\frac{\beta_0^*}{\sigma_v}$ 本身并没有告诉我们 D_i 的因果效应的大小，只有将其代入正态累

积分布函数才能将因果效应大小看清楚(不过它们的符号是准确的)。相比之下，不论是否假设 probit 分布，回归都能直接计算出我们需要的东西。

在兰德健康保险实验中最重要的被解释变量之一就是加总的医疗支出，也就是医疗成本。那么需要支付一定医疗费用的个体会倾向于少消费一些医疗服务吗？这里我们用个体支付的医疗成本来度量对医疗服务的消费量。在兰德健康保险实验中，需要支付一定费用的个体在医疗上的花费要比享受全额保险的人少 141 美元，这大概是享受全额保险的人全年医疗费用的 19%。这个计算意味着让消费者支付一分部医疗费用可以在相当程度上减少医疗支出，尽管这个估计可能并非那么准确。

由于医疗花费是非负的随机变量，而且有时会等于零，所以它的期望可以记做：

$$\mathrm{E}[Y_i \mid D_i] = \mathrm{E}[Y_i \mid Y_i > 0,\ D_i]\mathrm{P}[Y_i > 0 \mid D_i]$$

在不同处理组之间医疗花费的差异就是：

$$\begin{aligned}
&\mathrm{E}[Y_i \mid D_i = 1] - \mathrm{E}[Y_i \mid D_i = 0] \\
&= \mathrm{E}[Y_i \mid Y_i > 0,\ D_i = 1]\mathrm{P}[Y_i > 0 \mid D_i = 1] \\
&\quad - \mathrm{E}[Y_i \mid Y_i > 0,\ D_i = 0]\mathrm{P}[Y_i > 0 \mid D_i = 0] \\
&= \underbrace{\{\mathrm{P}[Y_i > 0 \mid D_i = 1] - \mathrm{P}[Y_i > 0 \mid D_i = 0]\}}_{\text{参与效应}}\mathrm{E}[Y_i \mid Y_i > 0,\ D_i = 1] \\
&\quad + \underbrace{\{\mathrm{E}[Y_i \mid Y_i > 0,\ D_i = 1] - \mathrm{E}[Y_i \mid Y_i > 0,\ D_i = 0]\}}_{\text{COP效应}} \\
&\quad \times \mathrm{P}[Y_i > 0 \mid D_i = 0]
\end{aligned} \tag{3.4.4}$$

因此，可以将平均花费之间的总体差异分解为两部分：医疗消费为正的个体在不同处理组时接受医疗治疗的概率之差(将这一部分叫做参与效应)以及给定个体接受医疗治疗，不同处理组之间平均医疗花费之差，叫做选择偏误来自正数效应(conditional-on-positive effect，简记为 COP)。当然，对估计因果效应而言这个分解并无特殊含义；等式(3.4.1)包含的意思仍然成立：用对 Y_i 关于 D_i 做回归得到的是对医疗成本求出的无条件平均处理效应。

1. 好的正数效应，坏的正数效应：对正数效应的一些拓展

由于我们可以将成本这类非负随机变量的因果效应分解为两部分，所以一些应用研究者觉得他们应该对这两部分分而视之。事实上，很多人都使用了将因果效应分解为两部分的模型，其中第一部分是针对参与者的因果效应的估计，第二部分考虑正数效应(比如见 Duan 等(1983，1984)中将这个模型应用于兰德健康保险实验)。等式(3.4.4)中第一部分没什么特殊含义，因为如前面提到的，Y_i 是虚拟变量只意味着平均处理效应也反映出概率上的不同。就目前的这个两部分模型而言，问题在于即使在随机实验中我们也无法对正数效应赋予一个因果解释。这个问题和 3.2.3 节指出的由不合格的控制变量引入的选择偏误问题是相同的。

为了进一步分析正数效应，记：

$$
\begin{aligned}
&\mathrm{E}[Y_i \mid Y_i>0,\ D_i=1]-\mathrm{E}[Y_i \mid Y_i>0,\ D_i=0] \\
&=\mathrm{E}[Y_{1i} \mid Y_{1i}>0]-\mathrm{E}[Y_{0i} \mid Y_{0i}>0] \\
&=\underbrace{\mathrm{E}[Y_{1i}-Y_{0i} \mid Y_{1i}>0]}_{\text{因果效应}}+\underbrace{\mathrm{E}[Y_{0i} \mid Y_{1i}>0]-\mathrm{E}[Y_{0i} \mid Y_{0i}>0]}_{\text{选择性偏误}} \qquad (3.4.5)
\end{aligned}
$$

其中第二行使用了 D_i 是随机分配的这一假设。这个分解指出可将正数效应分解为两部分：一部分是需要支付费用的那些人的因果效应，另一部分是在需要支付医疗费用和无需支付医疗费用的个体之间 Y_{0i} 的不同。上式中的第二部分是选择偏误的另一种形式，尽管它要比我们在第 2 章中提到的选择偏误更加微妙。

这里之所以出现选择偏误，是因为随机实验改变了需要支付费用的那部分人的构成。那些 $Y_{0i}>0$ 的个体可能包括一些个体，当要对医疗支付费用时他们便选择不接受治疗。换言之，相对于 $Y_{1i}>0$ 的组别，$Y_{0i}>0$ 中个体花费的医疗成本会更低，包含的个体规模也会越大。由于选择偏误项会是正的，整个正数效应就可能接近于零，而不再是负的因果效应 $\mathrm{E}[Y_{1i}-Y_{0i} \mid Y_{1i}>0]$。这是 3.2.3 节不合格控制变量的另一个版本：除非实验处理对 Y_i 为正的这件事没有影响，否则在一个考虑因果效应的背景下，$Y_i>0$ 是表示结果的变量，不适合作为控制变量。

对正数效应带来的非因果性的一个解决办法依赖于类似 Tobit 模型的那种删失回归(censored regression)。这些模型对接受医疗服务的人假定一个潜在的成本(Hay and Olsen，1984)。传统 Tobit 模型规定观察到的 Y_i 是如下生成的：

$$
Y_i=1[Y_i^*>0]Y_i^*
$$

其中，Y_i^* 是满足正态分布的潜在成本变量，可以取负值。由于 Y_i^* 不是有限被解释变量，所以支持 Tobit 模型的人可能觉得用类似于等式(3.4.3)的传统线性回归模型将 D_i 和 Y_i^* 联系起来会比较合适。无论 Y_i 是正的还是其他情况，都可知 β_1^* 是 D_i 关于潜在变量 Y_i^* 的因果效应。如果我们愿意研究 D_i 对潜在变量 Y_i^* 的作用，那么这样做就避免了由正数效应引入的选择问题。

但是我们不会对 D_i 对 Y_i^* 产生的效应感到满意。第一个问题是“潜在医疗花费”的构造比较令人疑惑。对某些人而言，医疗成本确实为零，但这个变量不是统计上的人为误差，也不是某种删失数据。由于不存在潜在为负的 Y_i^*，所以我们很难将潜变量的概念以及潜在为负的 Y_i^* 运用在这个问题上。第二个问题乃是潜在变量模型中得到的参数 β_1^* 以及被观察到的 Y_i 的因果效应都依赖于对潜变量的正态性假设。为此我们考虑给定 D_i 时 Y_i 的期望(McDonald and Moffitt，1980)。这个表达式使用正态分布：

$$
\mathrm{E}[Y_i \mid D_i]=\Phi\left[\frac{\beta_0^*+\beta_1^*}{\sigma_v}\right][\beta_0^*+\beta_1^* D_i]+\sigma_v \phi\left[\frac{\beta_0^*+\beta_1^*}{\sigma_v}\right] \qquad (3.4.6)
$$

和 v_i 的同方差性得到的，并假设 Y_i 可以表示为 $1[Y_i^*>0]Y_i^*$。

Tobit 模型的条件期望函数为我们提供了基于观察到的成本来表达平均处理效应的方法，具体而言就是：

$$\begin{aligned}&\mathrm{E}[Y_i \mid D_i=1]-\mathrm{E}[Y_i \mid D_i=0]\\&=\left\{\Phi\left[\frac{\beta_0^*+\beta_1^*}{\sigma_v}\right][\beta_0^*+\beta_1^*]+\sigma\phi\left[\frac{\beta_0^*+\beta_1^*}{\sigma_v}\right]\right\}\\&\quad-\left\{\Phi\left[\frac{\beta_0}{\sigma_v}\right][\beta_0^*]+\sigma_v\phi\left[\frac{\beta_0^*}{\sigma_v}\right]\right\}\end{aligned}\tag{3.4.7}$$

这可是个相当让人畏惧的公式。但是既然唯一的回归元是虚拟变量 D_i，那么根本不需要用类似于上式一样的东西来估计 $\mathrm{E}[Y_i \mid D_i=1]-\mathrm{E}[Y_i \mid D_i=0]$。用 D_i 对 Y_i 做回归后得到的斜率就是等式(3.4.7)等号左边的条件期望函数之差，这与你是否采用 Tobit 模型来解释潜在结果可能的结构[①]无关。

研究者使用正数效应模型时往往基于以下的考虑：被解释变量围绕某个值分布，也就说这些值堆积在某个值周围，比如说零——或者这些被解释变量是重尾分布的，或者有些被解释变量具有这里提到的两种特点，但是这样做会使得对平均效应的分析缺少了一点东西，类似于特定值出现概率的变化或者对中位数的偏离情况等。但是为什么不直接考察处理对分布带来的影响呢？如果考虑分布的话，那么分布结果就是医疗成本超过 0 美元、100 美元、200 美元等的概率。换言之，令 $1[Y_i>c]$ 表示个体选择不同 c 的概率，将其放在我们感兴趣的回归方程的等号左边。从计量经济学的角度出发，这个结果的所有取值都落在等式(3.4.2)里。直接考察处理对分布的影响的想法是在 Angrist(2001)中得到阐述的，在这篇论文中，Angrist 分析了抚养小孩对工作时间的影响。当然，如果我们感兴趣的数量显示出存在一个焦点，那么我们可以使用分位数回归来处理这种问题。第 7 章会仔细讨论这种方法。

那么类似于 Tobit 的模型还有用么？是的，如果你使用的数据确实是删失的。真正的删失数据意味着潜变量与我们感兴趣的被解释变量之间在经验研究中是对应的。劳动经济学中的一个典型例子就是 CPS 收入数据，这个数据删去了收入非常高的那些值以保护被访者的隐私。但是我们感兴趣的是受教育对收入水平的因果效应，只要使用出现在受访者纳税回单上的数据就好了，无需使用这些被采访人相应的最高值被删去的 CPS 数据。Chamberlain(1994)指出在某些年中，CPS 中被删去的数据在很大程度上降低了我们估计出的教育回报，于是他指出基于类似 Tobit 模型的方式来调整这种删失数据带来的问题。我们还会在第 7 章提到用分位数回归来模型化删失数据的方法[②]。

---

① 更一般化的 Tobit 模型就是样本选择模型，其中决定样本参与概率的潜在变量与潜在成本变量不同，比如见 Maddala(1983)。与解释潜在变量带来的效应相关的问题也会出现在样本选择模型中。

② 我们应该注意到，即使使用我们喜欢的回归模型，如果对工资取自然对数后再进行回归，那么也会带来正概率效应的问题，因为将工资取对数会忽略那些收入为零的观察点。如果教育影响工人选择工作的概率，那么对工资取自然对数就会带来我们在正概率效应中提到的选择偏误问题。因此实际中，我们关注于处在成熟期的男性，因为这些人参与工作的概率很高，而且在不同教育水平的组别之间差别不大(比如表 3.1 中反映出的年龄在 40—49 岁的白人男性的工作概率)。

2. 协变量导致的非线性

像 CPS 这样的真正的删失数据是很少的，因此在实际应用中构造性地使用 Tobit 模型的地方还是不多的。但是从某些角度来看，我们可能要更少地选择使用 Tobit 模型。对实验的讨论之所以看上去很简洁，部分原因在于 $E[Y_i \mid D_i]$ 必须是 D_i 的线性函数，此时回归和条件期望函数是一回事。实际上，关于 Y_i 的任何函数，甚至包括表示分布的 $1[Y_i > c]$，都可以得到一个线性的条件期望函数。当然，在实际中我们感兴趣的解释变量不会经常是虚拟变量，条件期望函数中往往还存在一些协变量，比如在有限被解释变量（LDV）模型中 $E[Y_i \mid X_i, D_i]$ 就几乎一定是非线性的。从直觉上来看，随着估计出的均值越来越接近被解释变量的边界，有限被解释变量的条件期望函数的导数会变得越来越小（比如考虑正态累积概率分布函数在极限值附近几乎是平的）。

这样带来的结果就是在存在协变量的有限被解释变量模型中回归无需完美地拟合条件期望函数。因此，即使在条件独立假设成立下可以对相应的条件期望函数赋予一个因果解释，回归无法完美地啮合条件期望函数这一事实也是无法改变的。但是如果可以对条件期望函数赋予一个因果解释，那么对回归也可以赋予一个因果解释，因为在它为我们提供了条件独立假设下对条件期望函数的最优逼近。更进一步，如果模型关于协变量是饱和的，那么回归估计出的仍然是类似于等式(3.3.1)和等式(3.3.3)那样的加权平均处理效应。类似的，如果我们感兴趣的回归元是多值的或者说连续的，我们可以得到一个加权的平均导数，这正是我们对 3.3.1 节最后的几个公式进行讨论的结果。

到现在，我们可能还没有足够的数据来构造那种很有吸引力的关于协变量饱和的模型。因此回归可能会失掉一些条件期望函数的性质。这带来的一个问题就是它产生的拟合值可能是有限被解释变量无法取到的。这个事实让一些研究者很不满，于是产生了不用线性概率模型的压力。类似于 Tobit 和 probit 模型的好处在于它们给出的条件期望函数确实满足了有限被解释变量所需要的变量有界的性质。特别的，probit 模型拟合出的值都在 0 和 1 之间，Tobit 模型拟合出的值都是正的（这一点从等式(3.4.6)中不是很能看出来）。因此在简单的曲线拟合中，我们也会选择这些非线性模型。

但是从另一方面考虑，这里需要着重强调的是非线性模型产生的结果必须转化为边际效应才能有用。边际效应就是在非线性模型中条件期望函数的变化。如果没有边际效应，就很难去讨论对可观察的被解释变量的影响。如果我们继续假设感兴趣的回归元是虚拟变量 D_i，于是可以通过做减法来得到边际效应：

$$E\{E[Y_i \mid X_i, D_i = 1] - E[Y_i \mid X_i, D_i = 0]\}$$

或者通过微分，$E\left\{\frac{\partial E[Y_i \mid X_i, D_i]}{\partial D_i}\right\}$。当面对连续或者多值的回归元时，大部分都是用这种微分的形式。

最小二乘估计能够在多大程度上近似类似于 Tobit 模型或者 probit 模型所求

出的边际效应？我们首先求出边际效应，然后用一个经验研究来举例说明。一个带有协变量的 probit 模型的条件期望函数可以写为：

$$\mathrm{E}[Y_i \mid X_i, D_i] = \Phi\left[\frac{X_i'\beta_0^* + \beta_1^* D_i}{\sigma_v}\right]$$

于是平均而言，得到处理和没有得到处理的效应之间的差距为：

$$\mathrm{E}\left\{\Phi\left[\frac{X_i'\beta_0^*}{\sigma_v}\right] - \Phi\left[\frac{X_i'\beta_0^*}{\sigma_v}\right]\right\} \tag{3.4.8}$$

在实际中，可以通过求平均导数来近似等式(3.4.8)：

$$\mathrm{E}\left\{\phi\left[\frac{X_i'\beta_0^* + \beta_1^* D_i}{\sigma_v}\right]\right\} \cdot \left(\frac{\beta_1^*}{\sigma_v}\right)$$

(对于用虚拟变量做的回归元，Stata 可以将上面两种边际效应都计算出来，但是默认值是等式(3.4.8)。)

类似的，将等式(3.4.6)推广到有协变量的情形，于是对于非负的有限被解释变量，有：

$$\mathrm{E}[Y_i \mid X_i, D_i] = \Phi\left[\frac{X_i'\beta_0^* + \beta_1^* D_i}{\sigma_v}\right][X_i'\beta_0^* + \beta_1^* D_i] + \sigma_v\phi\left[\frac{X_i'\beta_0^* + \beta_1^* D_i}{\sigma_v}\right]$$

Tobit 模型的边际效应往往以平均导数的形式表现出，我们可以将其表示为一种极为简单的形式：

$$\mathrm{E}\left\{\Phi\left[\frac{X_i'\beta_0^* + \beta_1^* D_i}{\sigma_v}\right]\right\} \cdot \beta_1^* \tag{3.4.9}$$

(Wooldridge，2006)。由等式(3.4.9)立刻可知相对于 D_i 对 Y_i 的效应，Tobit 模型中求出的参数 β_1^* 总是显得比较大。直觉上看，这是因为给定关于潜在结果 Y_i^* 的线性模型，随着 D_i 的不同，潜在结果总是在发生变化。但是真实的 Y_i 却不会变：对很多人来说，这个值总是零或者别的什么值。

表 3.6 比较了最小二乘估计和非线性模型的边际效应，这两个回归模型都针对女性就业状况和工作时间而设定，被解释变量都是有限的。表中报告的估计值都使用来自 Angrist 和 Evans(1998)使用过的样本，这个样本来自于 1980 年美国人口普查，包括了年龄在 21—35 岁之间，至少有两个孩子的已婚女性。表示抚养孩子的变量由标志是否有两个以上孩子的虚拟变量或用孩子总数来表示。协变量包括母亲年龄、第一胎生育时的年龄、标志种族(黑人或西班牙裔人)的虚拟变量以及母亲的教育水平(用标志母亲是否高中毕业、是否接受过大学教育以及是否大学毕业的虚拟变量表示)。协变量显然不是饱和的；而且这里协变量之间都是可加的关系，没有交叉项，于是在这个例子中对应的条件期望函数确实是非线性的。

表 3.6　在有限被解释变量模型中，用各种估计方法估计抚养孩子对工作的影响以及结果之间的比较

被解释变量	均值	等号右手边包含的变量：多于两个孩子 OLS	Probit 平均效应，全样本	Probit 被处理者的平均效应	Tobit 平均效应，全样本	Tobit 被处理者的平均效应	孩子个数 OLS	Probit MFX 平均效应，全样本	Tobit MFX 平均效应，全样本	Tobit MFX 被处理者的平均效应
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
A　全样本										
就业率	0.528	−0.162	−1.63	−0.162	—	—	−0.113	−0.114	—	—
	(0.499)	(0.002)	(0.002)	(0.002)			(0.001)	(0.001)		
工作时间	16.7	−5.92	—	—	−6.56	−5.87	−4.07	—	−4.66	−4.23
	(18.3)	(0.074)			(0.081)	(0.073)	(0.047)		(0.054)	(0.049)
B　超过 30 岁的非白领大学生，在 20 岁后有第一个孩子										
就业率	0.832	−0.061	−0.064	−0.070	—	—	−0.054	−0.048	—	—
	(0.374)	(0.028)	(0.028)	(0.031)			(0.016)	(0.013)		
工作时间	30.8	−4.69	—	—	−4.97	−4.90	−2.83	—	−3.20	−3.15
	(16.0)	(1.18)			(1.33)	(1.31)	(0.645)		(0.670)	(0.659)

注：这张表报告了用最小二乘、平均处理效应以及边际效应估计出的抚养孩子对女性劳动力供给的影响。表中 A 部分使用了 254 654 个观测值，与 Angrist 和 Evans(1998)中所用的来自 1980 年人口普查中已婚女性的数据相同。协变量包括年龄，生第一胎的年龄，第一胎是否为男孩的虚拟变量以及第二胎是否为男孩的虚拟变量。表中 B 部分包含 746 个非白人女性，这些女性受过一些教育，年龄都已超过 30 岁并且在 20 岁之前就生了第一胎。第一列的括号中报告的是标准差，其他列的括号中报告的是标准误。第 4、6 和 10 列估计了平均处理效应，其中第 10 列报告的是对生育两个以上孩子的女性的估计值。

从表 3.6 中的第 2—4 列可以发现，对于是否有两个以上孩子的虚拟变量所产生的影响，最小二乘估计和 probit 模型中的边际效应几乎没有区别。其中第一行比较了在所有的 1980 年样本中各种估计方法的结果。对第三胎生育的最小二乘估计值是−0.162，相应的 probit 模型中得到的边际效应是−0.163 和−0.162。其中我们用等式(3.4.8)得到最小二乘估计值，用下面的等式

$$\mathrm{E}\left\{\Phi\left[\frac{X_i'\beta_0^* + \beta_1^*}{\sigma_v}\right] - \Phi\left[\frac{X_i'\beta_0^*}{\sigma_v}\right] \middle| D_i = 1\right\}$$

得到非线性模型的边际效应。

在 Tobit 模型中考虑生育对工作时间的影响，发现得到的边际效应与相应的最小二乘估计很接近，但并非没有差别。这个结果可以从第 5、第 6 两列中读出。比较而言，Tobit 估计值是−6.56 和−5.87，而在第二列中最小二乘估计值是−5.92。虽然有一个 Tobit 估计值比最小二乘估计高出了将近 10%，但是实质上这并不重要。表 3.6 剩下的列以标志养育小孩数目的有序数作为解释变量，比较了最小二

乘估计值和非线性模型中的边际效应。这些值的计算过程都使用了求导以计算边际效应(记为 MFX)。这里还可以发现，来自 probit 模型和 Tobit 模型这类非线性模型的边际效应和最小二乘估计值很相似。

有人说当被预测的概率接近 0.5 时 probit 模型产生的边际效应接近最小二乘估计值，这是因为在此时相应非线性条件期望函数最接近于线性。当概率预测值接近于 0 或者 1 时，我们可能会发现线性模型与非线性模型之间会有较大的差别。为了检验这种说法，我们在一个子样本中比较最小二乘估计值和非线性模型的边际效应。这个子样本中平均的就业率都比较高，样本由上过大学，在 20 岁之前生下第一胎并且现在年龄已超过 30 岁的女性。虽然在这个子样本中就业率超过了 83%，但是最小二乘估计值和边际效应再次变得十分相似。

于是这里讨论的结论就是：虽然在被解释变量有限的情况下使用非线性模型可以更好地近似条件期望函数，但是当我们考虑边际效应时，线性模型和非线性模型下结论的差别会变得很小。这个相当乐观的结论算不上是一个定理，但是正如这里的例子所显示的，这个结论看上去很稳健。

既然如此，那我们为什么还要不厌其烦地使用非线性模型以及它的边际效应呢？一个答案是计算边际效应已经相当容易了，在有些计量经济学软件包(比如 Stata)中这个值是自动给出的。但是对于非线性模型而言，还要处理很多问题(比如加权的方式，用微分还是算出一个差值)，但是在线性模型中，这些问题的处理都已经标准化了。而且，当我们在非线性模型中使用工具变量或者面板数据时，还会遇到更复杂的问题。最后，相同的复杂性还会出现在推断过程中，因为我们还需要边际效应的标准误。奥卡姆剃须刀(Occam's razor)原理指出，“如无必要，勿增实体”①。在这个精神下，引用我们以前的老师 Angus Deaton(1997)在琢磨由类似于 Tobit 模型产生的非线性回归函数时说的一段话：

> 不知道 F[误差的分布]，这个回归方程就不能识别 β[Tobit 模型中的参数]——见 Powell(1989)——但是更加本质的是，如果我们必须处理这种丑陋的、困难的而且不稳健的模型，我们得到了什么。

3.4.3 为什么取名为“回归”，回归对平均值意味着什么？

回归一词来自于 Francis Galton(1886)对身高的研究。在他书中第 26 页描述说当时他正在拜访自己的裁缝，希望研究一下几乎呈正态分布的父母和其子女的身高。他注意到给定自己父母的身高，孩子身高的条件期望函数是线性的，由此得到的参数就是二元回归中的斜率项和截距项。由于身高是静态的(即身高的分布不随时间的变化而改变)，所以二元回归中的斜率项也正是相关系数，这个值自然是处在 0 和 1 之间。

---

① 即“简单有效原理”。——译者注

在 Galton 的研究中，唯一的回归元 x_i 是父母的平均身高，被解释变量 Y_i 是成年子女的身高。回归斜率就是 $\beta_1 = \frac{\text{cov}(Y_i, x_i)}{V(x_i)}$，截距就是 $\alpha = \text{E}[Y_i] - \beta_1 \text{E}[X_i]$。但是因为高度在每一代之间并不改变，所以 Y_i 和 x_i 的均值和方差都是相同的。因此：

$$\beta_1 = \frac{\text{cov}(Y_i, x_i)}{V(x_i)} = \frac{\text{cov}(Y, x_i)}{\sqrt{V(x_i)}\sqrt{V(Y_i)}} = \rho_{xy}$$

$$\alpha = \text{E}[Y_i] - \beta_1 \text{E}[X_i] = \mu(1 - \rho_{xy})$$

其中，ρ_{xy} 是代际之间身高的相关系数，$\mu = \text{E}[Y_i] = \text{E}[X_i]$ 就是总体的平均身高。由此我们得到线性的条件期望函数：

$$\text{E}[Y_i \mid x_i] = \mu(1 - \rho_{xy}) + \rho_{xy} x_i$$

因此，给定父母的身高，后代的身高就是父母身高以及总体中平均身高的加权平均。因此平均而言，父母身高较高，那么其后代的身高就不会像其父那么高。对父母身高不高的那些孩子而言也是一样的，他们的身高会比父母高一些。具体而言，Pischke（本书的一个作者）有六英尺三英寸那么高，他知道自己的孩子应该也会比较高，但可能不会像他那么高。可喜的是，Angrist 只有六英尺高，所以他知道自己的孩子可能会比他高。Galton 将这种性质叫做“遗传的身高向平均水平回归”。在现在，我们将它叫做向均值回归。

Galton 还是达尔文（Charles Darwin）的表兄，并且创立了优生学会（Eugenics Society），他致力于培育更优良的人种。事实上，他对回归的的兴趣很大程度上来自于这种诉求。我们因此总结出的道理就是科学观点的价值不应掺杂其发明者的政治观点。

看上去 Galton 并没有表现出对多元回归的浓厚兴趣，而本章处理的却正是这个问题。在 Galton 的工作中，回归只不过是对静态随机变量分布性质的机械性描述；这种描述只对用父母身高对子女身高进行回归有用而且还无法得到因果性解释。Galton 可能自己也这么认为，因为他曾经拒绝了 Lamarck 的想法，即认为特定的特点是可以遗传的（后来斯大林在苏联推广了这个观点）。

用回归在统计意义上控制一些变量以寻找令人满意的因果关系来自于 George Udny Yule（1899）中对贫困率决定因素的回顾性研究。Yule 是一位统计学家，是 Karl Pearson（Pearson 是 Galton 的门生）的学生，他意识到只要求解一般性的最小方程问题就可以将 Galton 的回归参数推广到多元变量的情形，而这个问题早已被 Legendre 和 Gauss 解决掉。Yule（1899）的论文应该是第一篇有了多元回归估计结果的论文。他的论文将一个地区的贫困率的变化与英国当地贫困法案的变化相联系，控制了人口增长和该地区人口的年龄分布这两个因素。他特别感兴趣于为穷人提供收入帮助但是不要求他们搬进贫民院可能使得救助本身失去了

效力。这是一个经过很好定义的需求因果关系的问题，即使到今天，这类问题仍然是我们考虑的问题①。

最后，我们注意到 Steven Stigler(1986)对回归的历史做了详尽的回顾。Stigler是芝加哥大学非常有名的统计学家，但是相比于他的父亲，经济学家和诺贝尔经济学奖获得者 George Stigler，他的名气还是逊了一些。

3.5 附录：对加权平均导函数求导

我们从用 S_i 对 Y_i 做回归开始：

$$\frac{\text{cov}(Y_i, S_i)}{V(S_i)} = \frac{\text{E}[h(S_i)(S_i - \text{E}[S_i])]}{\text{E}[S_i(S_i - \text{E}[S_i])]}$$

令 $k_{-\infty} = \lim_{t \to -\infty} h(t)$，我们假设这个极限存在。由微积分基本定理可知：

$$h(S_i) = \kappa_{-\infty} + \int_{-\infty}^{S_i} h'(t)\text{d}t$$

代入 $h(S_i)$ 后，分子变为：

$$\text{E}[h(S_i)(S_i - \text{E}[S_i])] = \int_{-\infty}^{+\infty}\int_{-\infty}^{u} h'(t)(u - \text{E}[S_i])g(u)\text{d}t\,\text{d}u$$

其中，$g(u)$是在 u 处 S_i 的密度函数。调整积分顺序后我们得到：

$$\text{E}[h(S_i)(S_i - \text{E}[S_i])] = \int_{-\infty}^{+\infty} h'(t)\int_{t}^{+\infty}(u - \text{E}[S_i])g(u)\text{d}u\text{d}t$$

第一层积分内的项等于 $u_t \equiv \{\text{E}[S_i \mid S_i \geqslant t] - \text{E}[S_i \mid S_i < t]\}\{\text{P}(S_i \geqslant t)[1 - \text{P}(S_i \geqslant t)]\}$，在等式(3.3.9)中的加权函数显然是非负的。令 $S_i = Y_i$，分母也同样可以写成权重的积分。我们因此得到可以代表二元回归参数 $\frac{\text{cov}(S_i, Y_i)}{V(S_i)}$ 的一个加权导数值。对有协变量的回归方程也可以得到相同的结果，请见 Angrist 和 Krueger(1999)的附录部分。

① Yule 的第一篇关于贫困法的应用性论文发表于 1895 年的 *Economic Journal* 上，Pischke 曾经当过这份刊物的副主编，随着 Yule(1897)的发表，多元回归理论开始广为流传。

▶4

实践中的工具变量：得到你想要的

任何发生的事情，发生着。
任何事情，正在发生的，引起其他事情发生，
引起其他事情发生。
任何事情，正在发生的，
引起它自身再次发生、发生。
此时，时间已无法带来秩序。

Douglas Adams, *Mostly Harmless*

有两件事情将计量经济学从其姊妹学科——统计学——中区别开来。第一件事是它不再羞于讨论因果关系。因果推断往往是应用计量经济学研究的代名词。统计学家 Paul Holland(1986)曾提醒说"如果不对研究对象进行某种程度的控制，就无法推断因果关系"，这句类似于格言的话否定了使用非实验数据进行因果推断的可能。另外一些没有进行深入思考的学者则会回到"相关非因果"这种老生常谈的问题。和大多数通过与数据打交道来谋生的人类似，我们相信即使感兴趣的变量可能不是某个研究者或者实验者[①]通过控制相关因素后得到的，相关性还是为推断因果关系提供了良好证据。

将计量经济学家与大部分统计学家——或者说大部分社会科学家——相区别的第二件事是我们有一个统计学的武器库，这个武器库发轫于早期计量经济学对如何估计线性联立方程(linear simultaneous equations)的系数进行的研究。这个武器库中最有力的武器就是工具变量法(instrumental variables，简写为 IV)，这也正是本章的主题。随着本章的展开，我们会发现除了帮助我们在联立方程组中估计出具有一致性的参数外，工具变量法还可以做更多事情。

在 20 世纪 20 年代对农产品市场的研究中，由 Phillip Wright 和 Sewall

① 最近一些年我们看到统计学家逐渐开始在因果关系的框架中使用非实验数据，比如见 Freedman (2005)的回顾。

Wright 两父子组成的研究团队对下面这个相当有挑战性的因果推断问题感兴趣：当观察到的价格和数量信息是需求曲线和供给曲线相交的结果时，我们如何去估计需求曲线和供给曲线的斜率。换言之，均衡的价格和数量——我们唯一可以观察的值——乃是两个随机方程在同一时刻的解。因此，我们观察到的价格和数量形成的散点图会落在哪条曲线上呢？Phillip Wright 花了一点时间后理解了这样的事实：在联立方程组中，总体回归参数可能无法估计其中任何一个方程的斜率。最初由 Wright(1928)提出的工具变量法解决了联立方程组问题，他让只出现在一个方程中的变量变动，由此引起的该方程的平移会与其他方程相交并得到一个轨迹，这个轨迹就是另外一个方程。用来实现方程移动的这个变量就被称为工具变量(Reiersol，1941)。

在另外的研究中，人们指出还可以用工具变量法解决回归模型中由度量误差(measure error)带来的偏误[①]。在线性模型的统计理论中，最重要的结论是：对回归元的度量存在随机误差时，回归系数偏向于零(为了看清个中缘由，我们想象回归元只包含随机误差，于是该回归元和被解释变量不相关，因此用这个回归元对 Y_i 做回归得到的系数会是零)。工具变量法可以用来剔除这种偏误。

在计量经济学发展史中，联立方程组模型(simultaneous equations model，简记为 SEMs)占有极其重要的地位。但从现在来看，虽然用以讨论工具变量的语言仍然来自联立方程组模型，但最有影响力的那些应用计量经济学论文已经很少将正式的联立方程组模型作为分析框架。如今我们更愿意用工具变量法解决度量误差问题，而不是用其估计联立方程组中的系数。不过毫无疑问的是，人们对工具变量法的广泛使用是为了解决遗漏变量偏误。工具变量法就像随机实验那样，既避免了在回归中加入过多的控制变量，还解决了控制变量被遗漏或者存在未知控制变量时带来的问题[②]。

4.1 工具变量与因果关系

我们喜欢以递进的方式，分两步来讨论工具变量法。首先在因果效应为常数的框架下进行讨论，然后放松对因果效应的限制，在更一般的情况下考虑个体潜在结果存在异质性从而使因果效应不为常数的情况。对异质性因果效应的讨论没有改变工具变量法的基本统计原理(比如 2SLS 法，简写为 2SLS)，但丰富了我们对工具变量估计值的解释。先从因果效应为常数入手讨论，则使得我们以最不让人混乱的方式解释工具变量法的作用机制。

① 主要的历史参考文献来自于 Wald(1940)和 Durbin(1954)，这两篇论文我们在本章之后的内容中都会提及。

② 见 Angrist 和 Krueger(2001)对工具变量的使用和发展历史的回顾，Stock 和 Trebbi(2003)对工具变量法的诞生提供了详细的回顾，Morgan(1990)拓展了计量经济学思想发展的历史，其中包括对联立方程模型的研究。

为了在因果效应为常数的框架下讨论教育水平和收入之间的因果联系，类似于上一章中作出的假设，记潜在结果为：

$$Y_s = f_i(s)$$

和

$$f_i(s) = \alpha + \rho s + \eta_i \tag{4.1.1}$$

这与我们在 3.2 节讨论回归与因果关系时的假设相同。而且，正如之前的讨论，假设存在一个控制变量组成的向量 A_i，记为“能力”，于是我们可以将选择偏误由可观察变量引入(selection-on-observable)表达为：

$$\eta_i = A_i'\gamma + \nu_i$$

其中，γ 是总体回归系数组成的向量，由构造可知 A_i'和 ν_i 不相关。从现在起，我们假设 s_i 与 η_i 相关的唯一原因是 A_i'，因此：

$$\mathrm{E}[s_i\nu_i] = 0$$

换言之，如果 A_i 可以观察到，那么我们会很开心地将它包含在回归方程中，得到一个如下的长回归方程：

$$Y_i = \alpha + \rho s_i + A_i'\gamma + \nu_i \tag{4.1.2}$$

等式(4.1.2)是线性模型(3.2.9)的另一个版本。这个等式中的误差项 ν_i 是在潜在结果中控制了能力后剩下的随机部分。根据假设，此误差项与教育水平无关。如果这个假设是正确的，那么用 s_i 和 A_i 对 Y_i 做回归就得到等式(4.1.2)中的系数。

我们当初想解决的问题是当 A_i 无法观察时如何估计回归方程(4.1.2)中的系数 ρ。工具变量法可被用来解决这个问题，这种方法要求研究者得到一个变量(这个变量叫做工具变量，记为 Z_i)，它与我们感兴趣的 s_i 有关但是与决定被解释变量的其他因素都无关。这里，“与决定被解释变量的其他因素都无关”是指 $\mathrm{cov}(\eta_i, Z_i) = 0$，或者等价地说 Z_i 与 A_i、ν_i 都无关。由于我们没有在感兴趣的因果模型中放入变量 Z_i，所以这个假设被称为排他性约束(exclusion restriction)。

给定排他性约束，由方程(4.1.2)我们可以得到：

$$\rho = \frac{\mathrm{cov}(Y_i, Z_i)}{\mathrm{cov}(s_i, Z_i)} = \frac{\mathrm{cov}(Y_i, Z_i)/V(Z_i)}{\mathrm{cov}(s_i, Z_i)/V(Z_i)} \tag{4.1.3}$$

等式(4.1.3)中第二个等号表达的含义很有用，因为用回归系数来思考问题往往比用协方差思考来的容易。在这里，我们感兴趣的系数 ρ 就是 Y_i 关于 Z_i 的总体回归方程(称之为简约式)系数与 s_i 关于 Z_i 的总体回归方程(称之为第一阶段)系数的比值。工具变量估计值就是等式(4.1.3)的样本值。注意到工具变量估计值要求第一阶段估计值不能为零，不过这可以通过检查数据来验证。但是，如果第一

阶段估计值只是稍微地显著不为零，那么得到的工具变量估计值可能就不会很有意义，这一点我们之后还会回来继续展开。

这里需要重申一下保证等式(4.1.3)中协方差之比等于因果效应 ρ 所需要的假设。首先，工具变量应该对 s_i 有确定的影响。这是第一阶段。其次，只是由于第一阶段，Y_i 和 Z_i 之间才存在了联系。第二个假设被称为排他性约束，当在本章后面部分讨论异质性因果效应模型时，我们会发现第二个假设包含两层意思：第一层意思是说工具变量可以起到随机抽样的效果(也即在控制了协变量后，工具变量与潜在结果无关，类似于第3章中的条件期望独立假设)，第二层意思是说除了第一阶段的机制，工具变量不会通过其他机制影响被解释变量。

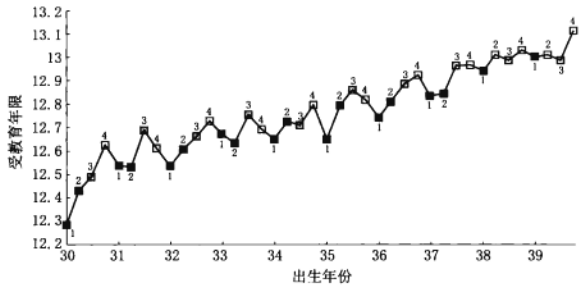
那么，我们在哪里找工具变量？好的工具变量来自于知识和想法的结合。一方面要对研究对象所处的制度背景有深刻理解，另一方面要对决定我们感兴趣变量的过程有好的看法。比如有关教育决策的经济学模型指出人们通过权衡成本和收益来作出接受多少教育的决策。因此，与个体能力或者未来收入潜力独立的贷款政策或者补贴政策，通过改变接受教育的成本改变了个体的教育决策，这类与受教育成本相联系的变量可以作为受教育水平的工具变量的一个来源。教育水平的工具变量的第二个来源可以考虑制度约束。与受教育水平相关的一系列制度约束都与义务教育法相关。Angrist 和 Krueger(1991)在他们的论文中发掘了由义务教育法带来的教育水平的变化，并代表性地使用“自然实验”纠正了遗漏变量引起的偏误。

在 Angrist 和 Krueger(1991)中，作者使用出生季度作为工具变量的出发点来自如下观察：美国的大部分州都要求学生在达到6岁的那个自然年度入学。因此入学年龄是出生日期的函数。具体而言，在每年下半年出生的孩子，其入学年龄会比较小。在那些将每年的12月31日作为截止日期的州里，前一年第四季度出生的孩子在还没有完全到六岁的时候就可以上学了，而那些出生在前一年一季度的学生，大约要到六岁半时才能入学。更进一步，由于义务教育法要求学生在16岁生日之前都必须待在学校里，所以这些出生在一年中不同季度的学生达到法定的可以离开学校的年龄时，他们会处在不同的年级。对学生入学年龄的要求与义务教育法的规定结合在一起，创造了一个自然实验，在这个实验里强制要求学生必须接受的教育水平和他们的出生日期有关。

Angrist 和 Krueger 运用美国人口普查数据考察了受教育水平与出生季度之间的关系。图4.1 A(来自 Angrist 和 Krueger(1991))使用1980年的普查数据绘出了在20世纪30年代出生的男性中受教育水平和出生季度之间的关系。这幅图清楚地显示出在一年中出生越早的人，其平均教育水平会越低。图4.1 A也是用图形表示的第一阶段回归。在运用工具变量法进行估计的一般化框架中，第一阶段是用协变量和工具变量对我们感兴趣的解释变量进行回归。由于按照年份和出生季度划分的平均受教育水平正好是用出生年份(协变量)虚拟变量和出生季度(工具变量)虚拟变量做回归后得到的拟合值，所以图4.1 A正好

就绘出了这个回归。

A. 按出生季度绘出的平均教育水平(第一阶段)



B. 按出生季度绘出的平均周工资(简约式)

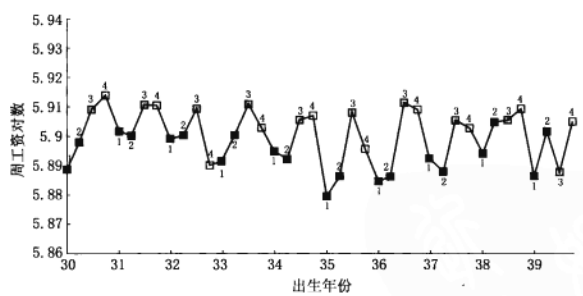


图 4.1　用出生季度作为工具变量来估计教育的经济回报,用图形绘出的第一阶段和简约式(Angrist and Krueger, 1991)

图 4.1 B 使用与图 A 相同的数据集绘出了按出生季度划分的平均收入。这幅图绘出了简约式下工具变量和被解释变量之间的关系。所谓简约式,就是用工具变量和模型中所有的协变量对被解释变量进行的回归。图 B 显示出年龄较大的人收入较高,这是因为收入会随着经验的增加而增加。该图同时还显示即使控制了出生年份和其他的一系列协变量(Angrist and Krueger, 1991),平均而言,出生的季度越早,收入会越少。重要的是,简约式显示出的规律和图 A 中出生季度与受教育年限之间的关系很相似,这意味着图 4.1 揭示出的两种规律是紧密联系在一起的。因为一个人的出生日期应该和他天生的能力、努力程度以及家庭背景无关,因此我们可以说造成收入随着出生季度变化而变化的唯一原因是教育水平随着出生季度的变化而变化。这就是我们把出生季度当作教育水平的工具变量时所作的

最重要假设。①

用数学来表示图 4.1 中描绘的第一阶段回归和简约式回归，这两个方程可以分别写为：

$$S_i = X_i'\pi_{10} + \pi_{11}Z_i + \xi_{1i} \tag{4.1.4a}$$

$$Y_i = X_i'\pi_{21} + \pi_{21}Z_i + \xi_{2i} \tag{4.1.4b}$$

方程(4.1.4a)中的 π_{11} 表示在控制了协变量 X_i' 后，在第一阶段工具变量 Z_i 对 s_i 的影响。方程(4.1.4b)中的 π_{21} 表示在控制了与第一阶段相同的协变量后，在简约式里 Z_i 对 Y_i 的影响。Angrist 和 Krueger(1991)中使用的工具变量是出生季度(也就是用来表示出生季度的虚拟变量)，协变量则是表示出生年份和出生地(以州为单位)的虚拟变量。用联立方程组模型的语言来讲，方程(4.1.4a)和(4.1.4b)中的两个被解释变量是内生变量(因为它们被这个系统的两个方程同时决定)，等号右边的那部分被称为外生变量(被系统之外的因素决定)。工具变量 Z_i 是外生变量的一个子集。不是工具变量的那些外生变量被称为外生的协变量。虽然在这个例子中，我们并未估计传统意义上的需求——供给系统，但是联立方程组模型中给予这些变量的名称已经在经验研究中得到广泛使用。

经过协方差调整的工具变量估计值就是联立方程组两个系数比值 $\frac{\pi_{21}}{\pi_{11}}$ 的样本值。为了看清这一点，注意到除以简约式中系数的那个参数和第一阶段估计值相等，因此这个比值就是

$$\rho = \frac{\pi_{21}}{\pi_{11}} = \frac{\mathrm{cov}(Y_i,\ \tilde{z}_i)}{\mathrm{cov}(S_i,\ \tilde{z}_i)} \tag{4.1.5}$$

其中，$\tilde{z}_i$ 是用外生的协变量 X_i 对 Z_i 做回归后得到的残差。因此等式(4.1.5)就用 $\tilde{z}_i$ 代替了原来在等式(4.1.3)右边出现的 Z_i。计量经济学家将等式(4.1.5)的样本值称为：在存在协变量的因果模型(4.1.6)中对 ρ 的间接最小二乘估计值，

$$Y_i = \alpha' X_i + \rho S_i + \eta_i \tag{4.1.6}$$

其中，η_i 代表复合残差项 $A_i'\gamma + \nu_i$。根据我们的构造，$\tilde{z}_i$ 与 X_i 不相关，同样根据假设，$\tilde{z}_i$ 与 η_i 无关，所以可以很容易地使用等式(4.1.6)来直接证明 $\mathrm{cov}(Y_i,\ \tilde{z}_i) = \rho\mathrm{cov}(S_i,\ \tilde{z}_i)$。

4.1.1 两阶段最小二乘回归

通过将方程(4.1.4a)代入因果关系(4.1.6)中，我们可以得到方程(4.1.4b)所

① 也可能存在另外的一些解释，最有可能的解释指出出生季度可能和影响收入的家庭背景存在关系(Bound, Jaeger and Baker, 1995)。对遗漏家庭背景对研究结果带来影响的回应来自于如下事实：在平均教育水平上表现出的出生季度特征在教育水平上表现得最明显，而教育水平主要受义务教育法的影响。

示的简约式，用联立方程组的语言讲，等式(4.1.6)被称为结构方程(structure equation)。通过这种变化我们得到：

$$\begin{aligned} Y_i &= \alpha' X_i + \rho[X_i'\pi_{10} + \pi_{11}Z_i + \xi_{1i}] + \eta_i \\ &= X_i'[\alpha + \rho\pi_{10}] + \rho\pi_{11}Z_i + [\rho\xi_{1i} + \eta_i] \\ &= X_i'\pi_{20} + \pi_{21}Z_i + \xi_{2i} \end{aligned} \tag{4.1.7}$$

其中，$\pi_{20} \equiv \alpha + \rho\pi_{10}$，$\pi_{21} \equiv \rho\pi_{11}$ 以及 $\xi_{2i} \equiv \rho\xi_{1i} + \eta_i$。等式(4.1.7)再次告诉我们 $\rho = \dfrac{\pi_{21}}{\pi_{11}}$。同时注意到对等式(4.1.7)稍加调整就有：

$$Y_i = \alpha' X_i + \rho[X_i'\pi_{10} + \pi_{11}Z_i] + \xi_{2i} \tag{4.1.8}$$

其中，$[X_i'\pi_{10} + \pi_{11}Z_i]$ 就是第一阶段用 X_i 和 Z_i 对 s_i 做回归得到的拟合值。因为简约式中的误差项 ξ_{2i} 与 X_i 和 Z_i 都不相关，所以用 X_i 和 $[X_i'\pi_{10} + \pi_{11}Z_i]$ 对 Y_i 做回归得到的 $[X_i'\pi_{10} + \pi_{11}Z_i]$ 前的系数就等于 ρ。

当然，在实际中我们往往要用样本数据来计算系数。给定一个随机抽样的样本，下面的等式可以让我们估计出具有一致性的第一阶段拟合值：

$$\hat{s}_i = X_i'\hat{\pi}_{10} + \hat{\pi}_{11}Z_i$$

其中，$\hat{\pi}_{10}$ 和 $\hat{\pi}_{11}$ 来自方程组(4.1.4)中第一阶段方程的最小二乘估计。用 $\hat{s}_i$ 和 X_i 对 Y_i 做回归后，$\hat{s}_i$ 前的系数被称为 ρ 的两阶段最小二乘回归(Two-stage least squares estimators，简写为 2SLS)估计值。换言之，通过对第二阶段方程进行最小二乘回归，我们可以构造出 2SLS 估计值。

$$Y_i = \alpha' X_i + \rho\hat{s}_i + [\eta_i + \rho(s_i - \hat{s}_i)] \tag{4.1.9}$$

将这个过程称为“两阶段最小二乘回归”是因为构造 2SLS 估计值需要分两步走，第一步运用方程组(4.1.4)中的第一阶段来估计 $\hat{s}_i$，第二步是估计等式(4.1.9)。因为协变量和第一阶段拟合值既与 η_i 不相关，也与 $(s_i - \hat{s}_i)$ 不相关，所以由此得到的估计值是 ρ 的一致估计。

虽然“两阶段最小二乘回归”里有个“两”字，但我们无需分两步去构造 2SLS 估计值。因为由此得到的标准误是错误的，这一点我们在本章后面部分会详细说明。特别的，我们用专门的统计软件程序(比如在 SAS 或 Stata 就有)进行相应的计算。这些软件可以帮助我们在得到正确的标准误的同时避免其他一些错误(见第 4.6.1 节)。不过，可以通过一系列最小二乘估计得到 2SLS 估计值这一事实有助于我们理解工具变量的作用机制。从直觉上来看，给定协变量后，2SLS 估计值只保留了由准实验——也即由工具变量 Z_i——带来的 s_i 的变化。

2SLS 是一件相当美好的事物。因为它是工具变量估计值：对等式(4.1.9)中 ρ 的 2SLS 估计值便是 $\dfrac{\text{cov}(Y_i, \hat{s}_i^*)}{\text{cov}(s_i, \hat{s}_i^*)}$ 的样本值，其中 $\hat{s}_i^*$ 是 $\hat{s}_i$ 关于协变量 X_i 进行回归后的残差。由多元回归中解构回归(见第 3.1.2 节)公式及 $\text{cov}(s_i, \hat{s}_i^*) = V(\hat{s}_i^*)$

可知 ρ 确实等于 $\frac{\text{cov}(Y_i,\ \hat{s}_i^*)}{\text{cov}(s_i,\ \hat{s}_i^*)}$。还可以很容易地知道在只有单个内生变量和单个工具变量的模型中，2SLS 估计值等于相应的间接最小二乘估计值①。

当出现多元工具变量时，2SLS 和工具变量之间的关系还需要进一步说明。假设每个工具变量都能捕捉到相同的因果效应（这是个很强的假设，我们会在之后放松），那我们希望将这些不同的工具变量估计值结合起来，构造出一个更为准确的工具变量估计值。当模型中存在多元工具变量时，2SLS 通过将多元工具变量转化为一元工具变量来达到这个目的。举个例子，假设我们有三个工具变量 Z_{1i}，Z_{2i} 和 Z_{3i}。在 Angrist 和 Krueger(1991)中，这三个工具变量都是虚拟变量，分别用来表示个体在第一、第二、第三季度出生。于是第一阶段回归变为：

$$s_i = X_i'\pi_{10} + \pi_{11}Z_{1i} + \pi_{12}Z_{2i} + \pi_{13}Z_{3i} + \xi_{1i} \qquad (4.1.10a)$$

第二阶段回归与等式(4.1.9)相同，只不过等式(4.1.9)中的拟合值 $\hat{s}_i$ 不再来自方程(4.1.4a)，而是来自等式(4.1.10a)。将这个 2SLS 估计值解释为工具变量估计值的方法和之前是一样的：用外生协变量对第一阶段拟合值进行回归，得到的残差就是工具变量。在这个例子中，排他性约束是指方程(4.1.10a)中标志出生季度的虚拟变量与方程(4.1.6)中的 η_i 无关。

用出生季度作为工具变量进行 2SLS 回归后求得的教育的经济回报请见表 4.1，该表分别报告了类似于 Angrist 和 Krueger(1991)中进行的最小二乘估计和 2SLS 估计的结果。在表中，每一列都包含了对 ρ 进行的最小二乘估计和 2SLS 估计，回归方程与(4.1.6)类似，只不过所用的工具变量和控制变量有所不同。第 1 列中的最小二乘估计值是用个体的工资对数进行回归后得到的，但没有包括任何控制变量。第 2 列中的最小二乘估计值是将表示出生年份和出生地（以州为单位）的虚拟变量作为控制变量加入回归后得到的估计值。在这两个例子中，估计出的教育的经济回报大概在 0.075 左右②。

第 3 列和第 4 列都报告了工具变量估计值，但是所使用的回归模型中没有包含外生的协变量。用于构造第 3 列中估计值的工具变量是表征出生于第一季度的单个虚拟变量，用于构造第 4 列的估计值的工具变量是表征出生于第一、第二、第三季度的三个虚拟变量。这些估计值从 0.10 到 0.11 不等。将出生年份和出生地所在州视作外生变量包括进模型后的估计值报告在第 5 列和第 6 列。毫不惊讶，这两种方式得到的结果几乎相同，因为出生季度与控制变量没有紧密联系。总体

---

① 注意到 $\hat{s}_i^* = \hat{z}_i\hat{\pi}_{11}$，其中 $\hat{z}_i$ 是 Z_i 关于 X_i 做回归后的残差，因此 2SLS 回归估计值就是 $\left[\frac{\text{cov}(Y_i,\ \hat{z}_i)}{V(\hat{z}_i)}\right](\hat{\pi}_{11})^{-1}$ 的样本值。但是 $\frac{\text{cov}(Y_i,\ \hat{z}_i)}{V(\hat{z}_i)}$ 的样本值正是简约式(4.1.4b)里 π_{21} 的最小二乘估计值，而 $\hat{\pi}_{11}$ 是对第一阶段系数 π_{11} 的最小二乘估计值。因此，当工具变量只有一个时，2SLS 估计值就是间接最小二乘估计值。这里间接最小二乘估计值是指简约式中工具变量的影响与第一阶段工具变量的影响之间的比值，这里假设第一阶段和简约式中都已包含协变量。

② 从表 4.1 中第 1 列和第 2 列看，估计出的教育回报都低于 0.075，但原文如此。——译者注

表 4.1　对接受教育的经济回报的 2SLS 估计

受教育年数	OLS		2SLS					
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	0.071	0.067	0.102	0.13	0.104	0.108	0.087	0.057
	(0.000 4)	(0.000 4)	(0.024)	(0.020)	(0.026)	(0.020)	(0.016)	(0.029)
外生协变量								
年龄(按季度)								✓
年龄(按季度)的平方								✓
9 个出生年份虚拟变量		✓			✓	✓	✓	✓
50 个出生州的虚拟变量		✓			✓	✓	✓	✓
工具变量								
表示 QOB=1 的虚拟变量			✓	✓	✓	✓	✓	✓
表示 QOB=2 的虚拟变量				✓		✓	✓	✓
表示 QOB=3 的虚拟变量				✓		✓	✓	✓
出生季度虚拟变量和出生年份虚拟变量间的交互项(共 30 个工具变量)							✓	✓

注:本表运用 Angrist 和 Krueger(1991)所使用的 1980 年人口普查数据,报告了运用最小二乘回归和 2SLS 回归方法估计出的教育的经济回报。研究所使用的样本包含的是在 1930—1939 年之间于美国本土出生、收入为正且关键变量都完备的男性白人。样本规模为 329 509 人。标准误报告在括号里。

而言,2SLS 估计值往往大于对应的最小二乘估计值。这意味着我们观察到的教育水平和收入之间的(正向)联系不是个体能力和家庭背景造成的,平均而言,教育水平可以提高收入。

表 4.1 的第 7 列报告了将交互项加入回归后的结果。特别的,这部分将 3 个标志出生季度的虚拟变量和 9 个标志出生年份的虚拟变量进行交互,从而得到了 30 个工具变量。于是第一阶段回归变为:

$$s_i = X_i'\pi_{10} + \pi_{12}Z_{2i} + \pi_{13}Z_{3i} + \sum_j (B_{ij}Z_{1i})\kappa_{1j} + \sum_j (B_{ij}Z_{2i})\kappa_{2j} + \sum_j (B_{ij}Z_{3i})\kappa_{3j} + \xi_{1i} \qquad (4.1.10b)$$

其中,B_{ij} 是虚拟变量,当个体 i 生于第 j 年时该虚拟变量等于 1,其中 j 取值在 1931—1939 年。系数 κ_{1j}、κ_{2j} 和 κ_{3j} 是相应的季度和年份交互项前面的系数。将这些交互项加入回归方程的目的是通过提高第一阶段的 R^2 来提高估计精度,因为对不同年份出生的人,表现在教育水平上的出生季度特点会不一样。在这个例子中,将额外的交互项加入工具变量只使得估计精度有了很小的提高;但是比较第 6 列和第 7 列却发现标准误从 0.019 下降到 0.016 。[①](图 4.1 中第一阶段和简约式的

① 这种估计精度上的微小提高可能是得不偿失的,因为过多地使用工具变量会提高估计偏误的可能,这一点我们在 4.6.4 节详细讨论。

图像就是从这个经过完全交互处理的模型中得到的。）

表 4.1 最后一列将个体在每个季度的年龄及其二次项作为外生协变量加入回归。也就是说，将 1930 年第一季度出生的人在普查日（4 月 1 日）的年龄记为 50 岁，将 1930 年第四季度出生的人在普查日的年龄记为 49.25 岁。因为年龄差别可能是某种遗漏变量的来源，会导致用出生季节做工具变量的识别策略存在疑点，所以将这种详细记录的年龄变量加入回归可以部分地控制这种遗漏变量偏误。如果年龄差别带来的影响是平滑的，那么将上面得到的这种年龄变量的二次型加入模型就可以很好地控制这种遗漏变量偏误。

表 4.1 的第 7 列和第 8 列阐述了识别（identification）方法和估计（estimation）方法如何相互作用。（在传统的联立方程组语言中，说一个参数可以被识别，是指我们可以从简约式中将这个参数解出来）。为使 2SLS 起作用，不论在模型中纳入何种外生协变量，第一阶段拟合值都不应是常数。如果第一阶段拟合值是协变量的线性组合，那么 2SLS 估计就不存在了。等式（4.1.9）就显示了完全的多重共线性（即 X_i' 和 $\hat{s}_i'$ 之间线性相关）带来的问题。用年龄的二次项作为控制变量的 2SLS 估计值是存在的，但是如果将个体在每个季度的年龄作为协变量纳入模型，那么由于个体在每个季度的年龄和表示出生季度的工具变量之间存在强相关性，所以这时第一阶段拟合值的变化就会很小。由于第一阶段拟合值的变化是标准误的主要来源，所以表 4.1 中的第 8 列估计值的精确性要比第 7 列差很多，但是它仍然接近于相应的最小二乘回归估计值。

1. 对工具变量和 2SLS 术语的简要概述

正如我们看到的，内生变量是被解释变量以及需要工具变量进行识别的解释变量；在联立方程组模型中，通过求解随机线性方程系统可以确定内生变量。将一个解释变量看作是内生变量，就是要对其使用工具变量，话句话说，就是在第二阶段用拟合值代替具有内生性的解释变量。在 Angrist 和 Krueger（1991）的研究中，内生的解释变量是教育水平。外生变量包括无需使用工具变量进行识别的外生协变量以及工具变量本身。在联立方程组模型中，外生变量被系统以外的因素决定。在 Angrist 和 Krueger（1991）中使用的外生协变量是表示出生年份和出生地的虚拟变量。我们则将外生的协变量看作控制变量。在那些痴迷于使用 2SLS 的研究者的世界里，他们对出现在 2SLS 中的随机变量赋予了相互区别的名称，以免引起混淆：在任何包含工具变量的经验研究中，被研究的随机变量包括被解释变量、存在内生性的解释变量、工具变量和外生协变量。有时我们将这些随机变量的名字简写为被解释变量、内生变量、工具变量和协变量（由于在联立方程组模型中被解释变量也是内生的，所以这里给出的术语和名称不同于传统的联立方程组模型）。

4.1.2 瓦尔德估计值

用单个虚拟变量做工具变量来估计只含一个内生变量且无协变量的回归模型，应该是我们所能见到的最简单的工具变量估计值。一个不含协变量的因果回

归模型写为：

$$Y_i = \alpha + \rho s_i + \eta_i \tag{4.1.11}$$

其中，η_i 和 s_i 可能是相关的。简单起见，假设工具变量是虚拟变量并以 p 的概率等于 1，那么有：

$$\text{cov}(Y_i, Z_i) = \{\text{E}[Y_i \mid Z_i = 1] - \text{E}[Y_i \mid Z_i = 0]\}p(1-p)$$

同样还能得到关于 $\text{cov}(s_i, Z_i)$ 的一个类似的公式。于是：

$$\rho = \frac{\text{E}[Y_i \mid Z_i = 1] - \text{E}[Y_i \mid Z_i = 0]}{\text{E}[s_i \mid Z_i = 1] - \text{E}[s_i \mid Z_i = 0]} \tag{4.1.12}$$

更加直接的得到上面这个估计值的方法是利用等式(4.1.11)和 $\text{E}[\eta_i \mid Z_i] = 0$ 两个条件可以推出：

$$\text{E}[Y_i \mid Z_i] = \alpha + \rho \text{E}[s_i \mid Z_i] \tag{4.1.13}$$

对方程(4.1.13)求解 ρ 亦能得到等式(4.1.12)。

等式(4.1.12)就是存在回归元误测[①]的二元回归中著名的瓦尔德估计值(Wald estimator)。在这里，瓦尔德估计值为我们提供了一种直观的方式来理解工具变量法是如何解决遗漏变量偏误的。用工具变量估计因果效应的核心观点在于：被解释变量和工具变量之间的唯一联系是工具变量通过影响我们感兴趣的解释变量进而影响被解释变量。当工具变量是虚拟变量时，很自然地应该用第一阶段的均值差除以简约式中的均值差。

在 Angrist 和 Krueger(1991)使用出生季度作为工具变量研究教育的经济回报中，他们就指出瓦尔德估计值确实在起作用。表 4.2 用 1980 年人口普查数据报告了用于构造瓦尔德估计值的两个均值差。出生在第一季度和出生在第四季度的男性之间的收入差距为 −0.013 5，而相应的受教育水平的差异为 −0.151。这两个差值之比就是教育回报的瓦尔德估计值，得到的结果是 0.089。毫不令人惊讶，这个估计值与表 4.1 中 2SLS 估计值没有太大差别。我们认为瓦尔德估计值和 2SLS 估计值应该相似的原因在于两者都是依赖相同的信息构造出来的：由于出生季度不同导致的收入差别。

Angrist(1990)研究了越战服役经历对退伍老兵收入的影响，这项研究也指出瓦尔德估计量很有用。在 20 世纪 60—70 年代，美国的年轻男性都有可能被招募入伍服役。出于对征兵政策公平性的考虑，美国政府在 1970 年建立了基于随机抽取的征兵制度，用来决定哪些人应该优先被征召入伍。由于这种参军资格完全由

① 正如在本章引言中指出的那样，回归元中存在度量误差时回归系数会趋于零。为了除去这种偏误，Wald(1940)指出可以将数据分类，分类方式应该与度量误差无关，那么类似于等式(4.1.12)，均值差的比值就是我们感兴趣想要估计的参数。Durbin(1954)指出 Wald 的方式实际上就是工具变量法，其中的工具变量就是 Wald 用以对数据分类的那个变量。Hausman(2001)回顾了处理度量误差的计量经济学方法。

基于出生日期的随机抽样决定，所以它可以是越战服役经历的一个良好的工具变量。具体而言，从 1970 年到 1972 年的三年中，每年都对年龄达到 19 岁的人的出生日期赋予一个随机数。如果该男性得到的随机数低于某个截断值，那么他就获得参军资格，如果某男性得到的随机数高于某个截断值，那么他就没有获得参军资格。在实际中，由于健康或者其他原因，很多获得参军资格的男性仍会被免于服兵役，同时没有获得参军资格的男性也可以志愿参军。因此，越战服役经历并不是完全由随机抽取的参军资格决定的，但是随机抽取的参军资格为我们提供了一个与越战服役经历高度相关的工具变量。

表 4.2　用出生季度做工具变量得到的教育回报的瓦尔德统计值

	(1) 在一年的第一个季度出生	(2) 在一年的第四个季度出生	(3) 差别 (标准误)
ln(周工资)	5.892	5.905	−0.013 5 (0.003 4)
受教育年限	12.688	12.839	−0.151 (0.016)
教育回报的瓦尔德估计值			0.089 (0.021)
教育回报的最小二乘估计值			0.07 (0.000 5)

注：来自于 Angrist 和 Imbens(1995)。样本由收入水平为正的出生于美国的白人男性组成，时间跨度为 1930—1939 年，数据来自 1980 年人口普查，样本规模为 162 515。

在 1970 年可能获得参军资格的那些白人男性中，参军资格与之后的低收入有显著联系。表 4.3 中第 2 列报告了随机抽取的参军资格对社保账户中可征税收入的影响。为了便于比较，第 1 列还报告了平均的年度收入。对于出生于 1950 年的男性而言，参军资格对 1971 年的收入水平有显著的副作用，此时他们刚开始在军队的服役，更令人吃惊的是这种副作用在十年后的 1981 年显得更大。相比之下，1969 年的收入水平表明参军资格对收入没有什么影响。注意到在这一年里政府只是对出生于 1950 年的人随机抽取了参军资格，但没有将他们征召入伍。

由于参军资格是随机分配的，所以用表 4.3 第 2 列的数字代表参军资格对收入的因果影响应该不会引起争议。但是从瓦德尔估计值的角度看，如果我们想通过参军资格对收入的影响计算越战服役经历对收入的影响，那还需要瓦尔德估计值中的分母，也就是拥有参军资格会如何影响一个人在越战服役的概率。这个信息报告在表 4.3 的第 4 列，它指出参军资格将越战服役概括提高了 16%。表中第 4 列[①]报告了用瓦尔德估计值计算出的越战服役经历对 1981 年收入造成的影响，

① 原文如此，疑为第 5 列。——译者注

表 4.3　出生在 1950 年的白人男性其参军经历对收入水平影响的瓦尔德估计值

	收　入		越战服役状态		
收入年	平均值 (1)	参军资格的影响 (2)	平均值 (3)	参军资格对服役概率的影响 (4)	服役对收入影响的瓦尔德估计值 (5)
1981	16 461	−435.8	0.267	0.159	−2 741
		(210.5)		(0.040)	(1.324)
1971	3 338	−325.9			−2 050
		(46.6)			(293)
1969	2 299	−2.0			
		(34.5)			

注：来自 Angrist(1990)的表 2 和 3。括号中的数字是标准误。收入数据来自社保账户的官方记录，都是名义值。参军经历的数据来自收入调查。样本包含 13 500 人。

大致导致平均收入下降 15%。1971 年参军对收入的影响(从百分比的角度看)要更大，这时研究样本中获得参军资格并入伍的士兵尚在服役之中。

瓦尔德估计值或者说工具变量估计值的一个重要特征是用它们识别因果关系时所作的假设既容易实现，也容易得到解释。令 D_i 表示是否有过越战服役经历，令 Z_i 表示是否获得参军资格。我们将瓦尔德估计值解释为 D_i 对收入的因果效应所基于的理由是：使 $E[Y_i \mid Z_i]$ 随着 Z_i 的变化而变化的唯一原因是 $E[D_i \mid Z_i]$ 在变化。有两种方式可以对这个理由进行检查。第一种方式可以考虑不受 D_i 影响的个体特征与 Z_i 之间的关系，这些个体特征包括种族、性别及其他特征。另一种方式是在 D_i 和 Z_i 之间不存在关系的样本中考察工具变量和最后结果之间的联系：如果参军资格影响收入的唯一原因是参军资格导致个体服役概率上升，那么在参军资格和入伍服役无关的样本中，这个影响应该是零。

Angrist(1990)在研究参军资格对收入产生的影响时就用第二种方式检查了工具变量可以发挥作用的前提条件。他考察了参军资格对 1969 年收入的影响，该估计值报告在表 4.3 最后一行。由于 1969 年的收入发生在 1970 年随机抽取参军资格之前，因此如果可以将瓦尔德估计值解释为越战服役经历对收入的影响，那么参军资格对 1969 年收入的影响应该为零，如果这个影响不为零，那么说明参军资格还通过另外的机制影响了收入。他得到的结果很令人满意，随机抽取的参军资格对 1969 年收入的影响确实为零。我们还可以变换一种检验方式，那就是考察 1953 年出生的人。虽然在 1972 年 2 月进行的随机分配将用于决定参军资格的随机数赋予了 1953 年出生的人，但是 1953 年出生的人没有一个被征召入伍(官方的征召行为在 1973 年 7 月结束)。因此，对于在 1953 年出生的男性，第一阶段回归中得到的参军资格和入伍状况(用 1952 年随机抽样的数字小于 95 为标准)之间的关系应该说明：获得参军资格和没有获得参军资格的人，他们服役的概率之间应该没有太大的差别。更进一步，对生于 1953 年的男性，其收入水平和参军资格确实没有显著的联系，这个结果支持我们的观点：参军资格影响收入的唯一原因是它影

响了个体在军队服役的可能性。

最后我们使用一个例子来结束对瓦尔德估计值的讨论，这个例子使用一组工具变量来考察家庭规模对母亲劳动力供给的影响。与对教育的经济回报或者越战服役经历对收入的影响所进行的研究一样，这个例子还会出现在本书的其他地方。一直以来，生育和劳动力供给之间的关系都是劳动经济学家研究的兴趣所在，但是这个问题中很明显地存在遗漏变量偏误：那些劳动参与水平较低或者潜在收入水平较低的女性可能更愿意多生孩子。这就使得我们很难解释观察到的家庭规模和工作参与之间的关系，因为家庭规模较大的母亲所能提供给市场的劳务本来就会少一些。Angrist 和 Evans(1998)用两个工具变量解决了这个遗漏变量问题，这两个工具变量估计值最后都采用了瓦尔德估计值的形式。

第一个瓦尔德估计值用双胞胎作为工具变量，这个方法最早由 Rosenzweig 和 Wolpin(1980)提出，用以研究家庭规模的影响。在 Angrist 和 Evans(1998)中，双胞胎工具变量是个虚拟变量，用以标志样本中至少有两个孩子的母亲曾在第二胎时生出了双胞胎。使用双胞胎做工具变量得到的第一阶段估计值是 0.625，报告在表 4.4 中第 3 列。这意味着有 37.5%的母亲在养育了两个孩子后还会选择生育第三胎①。双胞胎工具变量依赖于下面的想法：生育双胞胎往往是随机的，与家庭背景或者各种潜在结果无关。

表 4.4　家庭规模对劳动力供给影响的瓦尔德估计值

			工具变量估计值			
			双胞胎		性别组成	
被解释变量	平均值	最小二乘法	第一阶段	瓦尔德估计值	第一阶段	瓦尔德估计值
	(1)	(2)	(3)	(4)	(5)	(6)
就业率	0.528	−0.167	0.625	−0.083	0.067	−0.135
		(0.002)	(0.011)	(0.017)	(0.002)	(0.029)
工作周数	19.0	−8.05		−3.83		−6.23
		(0.09)		(0.76)		(1.29)
时间/周	16.7	−6.02		−3.39		−5.54
		(0.08)		(0.64)		(1.08)

注：本表报告了第三胎对劳动力供给影响的最小二乘估计和瓦尔德估计值，使用了双胞胎和性别组成作为工具变量。数据来自 Angrist 和 Evans(1998)，包含了 1980 年人口普查中 21—35 岁之间的至少有两个孩子的结婚女性。最小二乘模型包含了母亲年龄、生首胎时的年龄、表征首胎和第二胎性别的虚拟变量以及表征种族的虚拟变量。对所有的被解释变量而言，第一阶段是相同的。

① 记工具变量 Z_i 为虚拟变量，表示第二胎生双胞胎，记 D_i 为表示生育第三个孩子的虚拟变量，因此在用双胞胎做工具变量的模型中，第一阶段估计值为 0.625 意味着 $E[D_i \mid Z_i = 1] - E[D_i \mid Z_i = 0] = 0.625$，由于第二胎生双胞胎意味着该母亲必然生育了三个孩子，所以 $E[D_i \mid Z_i = 1] = 1$，由此得 $E[D_i \mid Z_i = 0] = 0.375$，这就说第二胎没有生双胞胎——两胎生育了两个孩子——的那些人选择生育第三个孩子的概率是 37.5%。——译者注

在表 4.4 中的第二个瓦尔德估计值使用性别组成作为工具变量，它基于以下的考虑：在美国家庭中，如果父母生育的前两个孩子性别相同，那么他们更愿意再生第三个孩子。这表现在表 4.4 的第 5 列，该数字指出前两个孩子都是相同性别的父母生育第三个孩子的概率会高 6.7%（当前两个孩子是不同性别时，父母愿意生第三胎的概率是 0.38）。用前两胎孩子性别相同作为工具变量基于的看法是：性别组成是随机的，只通过提高生育率来影响家庭的劳动力供给。

用双胞胎以及性别组成做工具变量得到的估计结果都指出生育第三胎对母亲的就业率、每年工作周数、每周工作时间数产生很大影响。使用双胞胎做工具变量得到的瓦尔德估计值指出精确估计得到的结果是：生育第三胎会导致就业率大约下降 0.08 左右，每年工作周数下降 3.8 周，周工作小时数下降 3.4 小时。这些结果报告在表 4.4 的第 4 列，它们都比第 2 列中用最小二乘估计得到的结果小。这意味着由于选择偏误，最小二乘估计结果被高估了。有趣的是，用表示性别组成的虚拟变量做工具变量构造出的瓦尔德估计值要比用双胞胎做工具变量估计出来的结果大（比如就业率下降了 0.135）。表 4.4 中在不同工具变量下得到的不同估计结果指出即使工具变量都是可行的，不同的工具变量也可能产生不同的结果。我们在第 4.4 节会对这一重要情况进行进一步阐述。不过就现在而言，我们还是继续沿着因果效应为常数的假设进行讨论。

4.1.3 分组数据和两阶段最小二乘

由于可以通过一组瓦尔德估计值构造出更加复杂的 2SLS 估计值，所以瓦尔德估计值是所有工具变量估计值的基础。而分组数据（grouped data）则将瓦尔德估计值和 2SLS 估计值联系在了一起：当用虚拟变量做工具变量时，2SLS 估计值实际上等于对一系列分组数据的组内均值做广义最小二乘估计（GLS）。因此，可以将广义最小二乘估计理解为瓦尔德估计值的线性组合，其中在每个分组数据里得到的组内均值都可以构造出一个进入该线性组合的瓦尔德估计值。瓦尔德估计值和 2SLS 估计值之间存在的这种一般意义上的关系看上去似乎有相当的局限性，因为它要求工具变量必须是虚拟变量。但是并非所有的工具变量都是虚拟变量，有些工具变量甚至都不是离散的，不过这种局限并不重要。首先，大量的工具变量都在定义类别，比如出生季度工具变量就是这一类的。更进一步，虽然有些工具变量看上去是连续的（比如决定参军资格时抽取的随机数取值从 1 到 365），但是我们可以在不损失太多信息的前提下将这些工具变量分组（比如可以用单个虚拟变量表征是否拥有参军资格，或者将随机抽取的数字分为 25 组，然后用虚拟变量表征不同组别）①。

① 经典的度量误差模型是一个特例，在这个模型中无论是需要工具变量识别的变量还是工具变量本身都是连续的。这里，我们要记住工具变量法包含对遗漏变量偏误的处理。

为了更细致地解释分组数据中瓦尔德估计值与 2SLS 估计值之间的关系，我们仍以随机抽取参军资格的研究为例。在之前的讨论中我们注意到随机分配的参军资格是越战服役经历很好的工具变量。就抽取随机数的过程而言，对在 1950 年出生的男性，获得的随机数小于 195 的那些人得到参军资格；对 1951 年出生的男性，获得的随机数小于 125 的那些人得到参军资格；对 1952 年出生的男性，获得的随机数小于 95 的那些人得到参军资格。但是在实际中，用以确定参军资格的随机数（记为 R_i，是对 RSN 的随写）和实际发生的服役（记为 D_i，表征个体是否服役）之间存在更为复杂的关系。虽然当个体被赋予的随机数大于某个特定值后此人不会被征召，但这个数字并不能预先知道。因此随着获得的随机数的不同，一些人会存在志愿参军的动机，因为主动地志愿参军有助于他们获得到更好的服役条件并掌控服役时间。具体而言，那些随机数较低的人获得参军资格的概率高，因此入伍服役的概率也高。在入伍服役的压力驱使下这些人选择志愿服役的可能性就大。那些随机数较高的人获得参军资格的概率低，因此入伍服役以及在入伍服役压力驱使下志愿服役的可能性就小。因此，即使人们获得的随机数严格大于或者小于获得参军资格的截断值（如对于 1950 年出生的男性，这个数字是 195，对于 1951 年出生的男性，这个数字是 125，等等），$\mathrm{P}[D_i = 1 \mid R_i]$ 还是会有变化。比如生于 1950 年，获得的随机数在 200—225 之间的男性，在事前他们被抽中参军的概率会大于随机数在 226—250 的男性，即使从事后来看他们都没有获得参军资格。

用参军资格作为工具变量，对生于 1950 年的男性进行估计得到的瓦尔德估计值是在比较 $R_i < 195$ 和 $R_i > 195$ 这两组人的收入。但之前的讨论指出我们实际上还可以构造更多的组别进行比较，比如在满足 $R_i \leqslant 195$ 和 $R_i \in [26, 50]$ 的组别之间进行比较，在满足 $R_i \in [51, 75]$ 和 $R_i \in [76, 100]$ 的组别之间进行比较等等，直到对随机数划分出的 25 个组都列举进来。我们也可以更加仔细地划分组别，比如每隔五个数字甚至是每隔一个数字就划出一个区间进行比较。这样做的结果就是得到一组瓦尔德估计值。上面对不同分组的构造是完备的，因为我们完整地分割了工具变量所在的支撑。而且由于构造瓦尔德估计值时使用的分子之间是相互独立的，所以由此得到的一组瓦尔德估计值之间都是线性无关的。于是，一旦 R_i 与潜在结果无关但与服役状态有关（即瓦尔德估计值的分母非零），并假设待估计的每个因果效应都相同，那么每个瓦尔德估计值都一致地估计出了相同的因果效应。

可以对相同的因果效应构造多个瓦尔德估计值这一事实为我们提出了以下的问题：我们能对这些瓦尔德估计值做点什么。我们更希望得到一个单一的估计值，来将这一组瓦尔德估计值进行有效组合。随着下面讨论的展开，我们会发现：对用来构造瓦尔德估计值的组内均值拟合直线是一种很有效的方法，这个方法对一组完备的线性独立瓦尔德估计值进行了有效的线性组合。

用下面的方法可以直接估计对各个分组数据的组内均值拟合出的直线的斜率。类似于等式（4.1.11），我们使用二元常因果效应模型，表示为：

$$Y_i = \alpha + \rho D_i + \eta_i \tag{4.1.14}$$

其中，$\rho = Y_{1i} - Y_{0i}$ 是我们感兴趣的因果效应，$Y_{0i} = \alpha + \eta_i$。由于 R_i 是随机分配的，除了通过影响参军服役来影响个体收入之外，没有其他途径对个体的收入水平产生影响，所以 $\mathrm{E}[\eta_i \mid R_i] = 0$。再由 $\mathrm{P}[D_i = 1 \mid R_i] = \mathrm{E}[D_i \mid R_i]$，我们可得：

$$\mathrm{E}[Y_i \mid R_i] = \alpha + \rho \mathrm{P}[D_i = 1 \mid R_i] \tag{4.1.15}$$

换言之，给定随机数后，在平均收入和平均服役概率之间拟合出的直线的斜率就是服役对收入带来的影响 ρ。这也正好指出对 Y_i 关于 D_i 做回归——也就是求出是否服役对平均收入造成的影响——应该必然不同于 ρ，因为 Y_{0i} 和 D_i 是相关的。

等式(4.1.15)指出我们可以用 $\mathrm{E}[Y_i \mid R_i]$ 和 $\mathrm{P}[D_i = 1 \mid R_i]$ 的样本值来拟合曲线，从而估计出 ρ[①]。假设 R_i 取值为 $j = 1, \cdots, J$。在实际中，j 可以从 1 取到 365，但是在 Angrist(1990)中，随机抽取的数据按照每五个为一组进行分类，一共得到 69 个区间，再将 346—365 作为一个区间，一共得 70 个区间。因此我们可以认为有 70 个 R_i，i 从 1 取到 70。记 $\bar{y}_j$ 和 $\hat{\bar{p}}_j$ 分别表示对 $\mathrm{E}[Y_i \mid R_i = j]$ 和 $\mathrm{P}[D_i = 1 \mid R_i = j]$ 的估计，记 $\bar{\eta}_j$ 为等式(4.1.14)中的平均残差。由于样本矩收敛于总体矩，所以对等式(4.1.16)表示的分组方程

$$\bar{y}_j = \alpha + \rho \hat{\bar{p}}_j + \bar{\eta}_j \tag{4.1.16}$$

进行最小二乘回归估计出的参数关于 ρ 是一致的。但是在实际中人们更倾向于用广义最小二乘法进行计算，因为分组方程往往是方差结构已知的异方差模型。在线性常因果效应模型中可以使用加权最小二乘法计算参数，其中权重取为 $\bar{p}_j$ 的方差(Prais and Aitchison，1954；Wooldrige，2006)。假设分组数据内部的微观数据的残差项是同方差的，记为 σ_η^2，那么 $\bar{\eta}_j$ 的方差就是 $\frac{\sigma_\eta^2}{n_j}$，其中 n_j 是分组数据规模。正如第 3.4.1 节讨论的，这时对分组方程进行加权最小二乘的权重可以定为分组数据规模。

在方程(4.1.16)中用广义最小二乘(或者加权最小二乘)得到的 ρ 的估计值非常重要，原因有二。首先，运用广义最小二乘法从 J 个分组数据中构造出的斜率是任何 $J-1$ 个线性独立的瓦尔德估计值的渐进有效线性组合(Angrist，1991)。不用任何数学推导我们也可以看出这一点：首先，无论是广义最小二乘，还是瓦尔德估计值的其他线性组合，都是分组数据中被解释变量的线性组合。由于广义最小二乘估计值是分组数据的有效线性组合值，所以我们可以得到结论：(在渐进有效的意义下)瓦尔德估计值的最有效线性组合由广义最小二乘给出(再次说明一下，这里仍然假设 ρ 是常数，不随分组数据的变化而变化)。从一组完备的线性独立的瓦尔德估计值中构造广义最小二乘估计值的公式可以参阅 Angrist(1988)。

其次，由于每个瓦尔德估计值都是工具变量估计值，所以等式(4.1.16)中的广

① 原书中的符号是 e，可能是笔误。——译者注

义最小二乘估计仍然是 2SLS 估计值。这个例子中的工具变量就是标志被分组后的随机数的组别属性的一组完备的虚拟变量。为了看清楚这一点，定义虚拟工具变量集合 $Z_i \equiv \{r_{ji} = 1[R_i = j]: j = 1, \cdots, J-1\}$，其中 $1[\cdot]$ 表示用以构造虚拟变量的示性函数。现在，考虑对 D_i 关于 Z_i 做第一阶段回归。由于第一阶段回归是饱和的，所以拟合值应该就是样本的条件均值 $\hat{p}_j$，在等式(4.1.16)中每个 $\hat{p}_j$ 都会出现 n_j 次。

无论从概念上来看，还是从具体的经验研究实践来看，分组数据和 2SLS 之间的关系都很重要。从概念的角度讲，可以将任何使用一组虚拟变量做工具变量的 2SLS 估计值理解为相应瓦尔德估计值的线性组合，其中每个虚拟工具变量产生一个瓦尔德估计值。因此在本章后面部分讨论异质性潜在结果时，瓦尔德估计值为我们提供了一个简单的框架，可以用它来解释因果效应为异质性时工具变量估计值的含义。

虽然不是所有的工具变量都是离散的，也不是在所有情况下都可运用瓦德尔或者分组数据的方法来解释估计结果，但是很多情况下确实可以这么做，在上文中提到的决定参军资格的随机数，出生季度，双胞胎和性别组成等都是现成的例子(Bennedsen et al., 2007; Ananat and Michaels, 2008；这两项研究都使用第一胎为男性作为工具变量)。更进一步，表现为连续形式的工具变量在很多程度上都可以转化为离散变量。比如，Angrist, Graddy 和 Imbens(2000)将基于天气状况得到的连续型工具变量分解为三个虚拟变量：有风雨，晴朗，都可能，然后他们用这三个工具变量估计了鱼的需求函数。这种将连续变量变成虚拟变量的参数化过程看上去可以捕捉天气状况和鱼价①之间关系的主要部分。

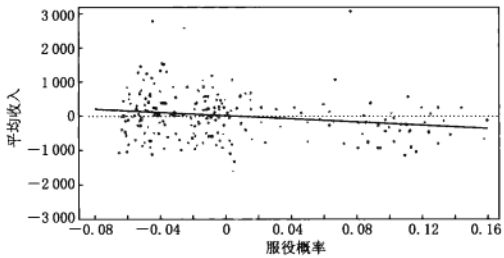
从经验研究的实践角度讲，分组数据与 2SLS 回归之间的等价性为我们解释和评价工具变量提供了简单方法。比如在随机抽取参军资格的例子中，分组模型体现出我们对工具变量所做的假设：随着决定参军资格的随机数的变化，平均收入水平发生变化的唯一原因是获得不同随机数的个体的参军概率不同。如果潜在的因果关系是线性的而且因果效应是常数，那么等式(4.1.16)可以很好地对组均值进行拟合，在下一节，我们还会把这个特点与正式的统计推断结合起来进行讨论。

有时候，劳动经济学家将在离散工具变量下绘制出的分组数据散点图称为可视化工具变量(visual instrumental variables, VIV)②。在 Angrist(1990)可以找到这样的例子，图 4.2 也对其进行了重新绘制。这幅图以每五个随机数(RSN)为一组，绘制出了在各个组之间平均收入和服役概率之间的关系。在这幅图中使用到的收入数据是生于 1950—1953 年的白人男性在 1981—1984 年之间的收入。穿过数据点绘出的直线的斜率就是参军带来的收入下降的工具变量估计值，在本例中，该斜率是 2 400 美元，与我们之前讨论过的瓦尔德估计值相去不远，但是它的

① 将连续型工具变量分解为一组虚拟变量，我们可得到用以估计第一阶段关系 $E[D_i \mid Z_i]$ 的一个简约型非参数模型。在同方差、常因果效应模型中，可以将 $E[D_i \mid Z_i]$ 视为渐进有效的工具变量(Newey, 1990)。

② 比如，见 Borjas(2005)的封面。

标准差更小(在本例子中是 800 美元)。



注:这是一个可视化工具变量的散点图,它以每五个数字为一个组,用 1981—1984 年的收入与参军的条件概率绘制散点图。样本包括了生于 1950—1953 年的白人男性。从这些点中得到的最小二回归估计值为−2 384,标准误为 778。

图 4.2 平均收入水平和服役概率之间的关系(来自 Angrist, 1990)

4.2 两阶段最小二乘的渐进推断

4.2.1 两阶段最小二乘回归系数的渐进分布

我们可以用类似于第 3.1.3 节中讨论最小二乘的方法来求得 2SLS 估计值的极限分布。令 $V_i \equiv [X_i', \hat{s}_i]'$ 表示等式(4.1.9)所表示的第二阶段回归中的回归元向量。于是 2SLS 估计值可以写为:

$$\hat{\Gamma}_{2SLS} \equiv \left[\sum_i V_i V_i'\right]^{-1} \sum_i V_i Y_i$$

其中,$\Gamma \equiv [\alpha' \rho']$ 是相应的系数向量。注意到:

$$\begin{aligned} \hat{\Gamma}_{2SLS} &= \Gamma + \left[\sum_i V_i V_i'\right]^{-1} \sum_i V_i [\eta_i + \rho(s_i - \hat{s}_i)] \\ &= \Gamma + \left[\sum_i V_i V_i'\right]^{-1} \sum_i V_i \eta_i \end{aligned} \tag{4.2.1}$$

其中,第二个等号来自于以下事实:在样本中,第一阶段的回归残差 $s_i - \hat{s}_i$ 和 V_i 是正交的。因此,两阶段最小二回归的系数向量的渐进分布就是 $\left[\sum_i V_i V_i'\right]^{-1} \sum_i V_i \eta_i$ 的渐进分布。相比于我们在讨论最小二乘回归系数向量的渐进性质时得到的统计量 $\left[\sum X_i X_i'\right]^{-1} \sum X_i e_i$,现在这个统计量更加难以处理,因为回归元中包含了来自第一阶段回归的拟合值 $\hat{s}_i$。不过斯拉茨基定理在这里可以派上用场,该定理指出将 $\left[\sum_i V_i V_i'\right]^{-1} \sum_i V_i \eta_i$ 中的样本拟合值替换为总体拟合值并不改变其极限分布(也就是用 $[X_i' \pi_{10} + \pi_{11} Z_i]$ 代替 $\hat{s}_i$)。由此可知 $\hat{\Gamma}_{2SLS}$ 是渐进正态分布的,概率极

限是 Γ，统计量 $\left[\sum_i V_i V_i'\right]^{-1}\left[\sum_i V_i V_i' \eta_i^2\right]\left[\sum_i V_i V_i'\right]^{-1}$ 则可一致地估计出协方差矩阵。与最小二乘估计的标准误类似，这个公式也是个三明治公式①(sandwich formula)(White，1982)。在给定协变量和工具变量下，如果 η_i 是条件同方差的，那么对协方差矩阵的一致估计就可化简为 $\left[\sum_i V_i V_i'\right]^{-1}\sigma_\eta^2$。

这里剩下的内容都没有什么新意，但有一点很微妙。通过估计等式(4.1.4)中第一阶段回归并将拟合值代入等式(4.1.9)，我们似乎可以很自然地构造 2SLS 估计值并在第二阶段用最小二乘法将所需要的参数估计出来。从估计参数的角度看，这样做无可厚非，但是在此过程中得到的标准误可能是错的。在使用软件进行回归时，传统的回归软件可能并不知道你是在构造一个 2SLS 估计值，因此软件会自动使用第二阶段最小二乘估计中得到的残差来计算标准误，也即对下面等式所指的残差计算方差：

$$Y_i - [\alpha' X_i + \rho \hat{s}_i] = [\eta_i + \rho(s_i - \hat{s}_i)]$$

式中，参数 α 和 ρ 由第二阶段估计值代替。但是，正确的残差方差估计值要使用最初的内生回归元来构造残差，而不是第一阶段拟合拟合值 $\hat{s}_i$ 来构造残差。换言之，我们想要的是用公式 $Y_i - [\alpha' X_i + \rho s_i] = \eta_i$ 估计出的残差，然后利用这个残差对 σ_η^2 作出一致性的估计，单不是用第二阶段回归中估计出的残差 $\eta_i + \rho(s_i - \hat{s}_i)$ 来对 σ_η^2 作出估计。虽然这个问题很容易解决(你可以用别的方式来构造合适的残差方差)，但是设计用来做 2SLS 的软件可以自动进行调整，同时帮助你避免一些大家在做 2SLS 中会遇到的共同问题。

4.2.2 过度识别与两阶段最小二乘的最小化元*

对常因果效应模型而言，如果使用的工具变量个数大于内生变量个数，那么我们说这个模型是过度识别的(工具变量数和内生变量数相同的模型被称为恰可识别(just-identified))。如果对感兴趣的参数进行识别时使用了过多的工具变量，那么我们可以用一种统计检验方法来评估过度识别模型的好坏。这种统计检验方法是模型设定检验的一种，它实际是在回答这样一个问题：在可视化工具变量中画出的那条直线是不是可以很好地拟合相应的条件均值(这里假设对条件均值的计算是精确的)。举个例子，在使用随机抽取的参军资格做工具变量的例子中，相应的条件均值就是针对每个分组数据计算出的平均收入和平均的参军概率。在本小节接下来的讨论中我们使用矩阵语言，因为这样做便于我们对过度识别检验的细节进行的讨论。

记 $Z_i \equiv [X_i' Z_{1i} \cdots Z_{Qi}]'$ 是由外生协变量和 Q 个工具变量构成的向量，记 $W_i \equiv [X_i' s_i]'$ 是由外生协变量和我们感兴趣的单个内生变量组成的向量。比如在用出

① 这是对该公式形状的一个形象的比喻。——译者注

生季度作为工具变量的研究中，相应的协变量就是表示出生年份和出生地所在州的虚拟变量，工具变量是表示出生季度的虚拟变量，内生变量则是受教育水平。为与之前使用的记号保持一致，记待估计系数向量为 $\Gamma \equiv [\alpha' \rho]'$。于是我们感兴趣的因果模型（第二阶段）中的残差可以定义为 Γ 的方程：

$$\eta_i(\Gamma) \equiv Y_i - \Gamma' W_i = Y_i - [\alpha' X_i + \rho s_i] \tag{4.2.1}$$

根据假设，这个残差与工具变量组成的向量 Z_i 不相关。换言之，η_i 满足下面的正交条件：

$$E[Z_i \eta_i(\Gamma)] = 0 \tag{4.2.2}$$

但是，在任何样本中这个方程都不会严格成立，因为等式(4.2.2)中的矩条件个数要大于 Γ 中待估计的参数的个数①。等式(4.2.2)的样本估计值就是关于 i 求和，也即：

$$\frac{1}{N}\sum Z_i \eta_i(\Gamma) \equiv m_N(\Gamma) \tag{4.2.3}$$

由此可见 2SLS 实际上是一种广义矩估计值(Generalized method of moments，简称为 GMM)，在这个意义下求解 2SLS 估计值就是在求解能使等式(4.2.3)所表示的样本矩最接近零的那个参数向量。

由中心极限定理，样本矩向量 $\sqrt{N} m_N(\Gamma)$ 的渐进协方差矩阵等于 $E[Z_i Z_i' \eta_i(\Gamma)^2]$，我们将这个矩阵称为 Λ。虽然粗看上去这个矩阵有点吓人，但是它和等式(3.1.7)中构造稳健标准误时得到的三明治型公式一样，也是一个四阶矩所构成的矩阵。正如 Hansen(1982)指出的，方程(4.2.2)的最优广义矩估计值应该可以最小化样本矩向量 $m_N(\hat{g})$ 的二次型。这里我们将 $m_N(\Gamma)$ 中的 Γ 替换为 $\hat{g}$，用它来表示 Γ 的估计值。在构造 $m_N(\hat{g})$ 的二次型时，最优权重矩阵就是 Λ^{-1}。当然在实际中 Λ 也是未知的，需要估计。解决这个问题的一个可行办法就是用 Λ 的一致估计来代替 Λ。既然在矩估计过程中使用的估计出来的加权矩阵与 Λ 有相同的渐进分布，那么我们现在先忽略两者之间的区别。于是我们要最小化的二次型可以写为：

$$J_N(\hat{g}) \equiv N m_N(\hat{g}) \Lambda^{-1} m_N(\hat{g}) \tag{4.2.4}$$

其中等式最前面出现的 N 来自于样本矩中用于正规化的两个 $\sqrt{N}$（也就是两个 $\sqrt{N} m_N(\Gamma)$ 相乘后得到的 N）。下面我们可立刻指出，当残差为条件同方差时，通过最小化 $J_N(\hat{g})$ 得到的估计值正好是 2SLS 估计值。当同方差性不成立时，通过最小化 $J_N(\hat{g})$ 得到的估计值就是 White(1982)得到的两阶段工具变量(two-stage IV，它是对 2SLS 的一般化)估计值，从这个角度出发，我们将 $J_N(\hat{g})$ 称为

① 假设存在 κ 个协变量，那么在内生变量只有一个但是工具变量多于一个的模型中，Γ 是 $[\kappa+1]\times 1$ 维的，但是 Z_i 是 $[\kappa+Q]\times 1$ 维的，其中 $Q>1$。由于矩条件个数多于待求解的参数个数，所以由此得到的线性系统存在多个解。如果工具变量之间存在某种线性相关性，使得一些变量是多余的，从而减少矩条件，那么这个线性系统可能存在唯一解。

2SLS 的最小化元。

当用广义矩估计来解释 2SLS 时，还需要考虑一些细节[①]。条件同方差性意味着：

$$\Lambda = \mathrm{E}[Z_i Z_i' \eta (\Gamma)^2] = \mathrm{E}[Z_i Z_i'] \sigma_\eta^2$$

用它替换 Λ^{-1} 并用 y、Z 和 W 表示相应向量和矩阵的样本值，于是需要最小化的二次型变为：

$$J_N(\hat{g}) = \frac{1}{N\sigma_\eta^2}(y - W\hat{g})' Z \mathrm{E}[Z_i Z_i']^{-1} Z'(y - W\hat{g}) \qquad (4.2.5)$$

最后，用样本的叉乘矩阵 $\left[\frac{Z'Z}{N}\right]$ 代替 $\mathrm{E}[Z_i Z_i']$，可得：

$$\hat{J}_N(\hat{g}) = \frac{1}{\sigma_\eta^2}(y - W\hat{g})' P_Z (y - W\hat{g})$$

其中，$P_Z = Z(Z'Z)^{-1}Z$。由此得到二次型的解是：

$$\hat{g} = \hat{\Gamma}_{2\mathrm{SLS}} = [W' P_Z W]^{-1} W' P_Z y$$

求出这个解的过程基于两个观察，一是映射元 P_Z 可以产生拟合值(比如 $P_Z W$ 就等于对 W 关于 Z 进行回归后的拟合值)，二是从等式(4.1.9)中对第二阶段进行最小二乘得到的估计值可知 P_Z 是幂等矩阵。更一般地，即使同方差假设不成立，我们还是可以通过最小化等式(4.2.4)得到类似于 2SLS 的有效估计值，还能用 $\mathrm{E}[Z_i Z_i' \eta_i (\Gamma)^2]$ 的一致估计来构造 $\hat{J}_N(\hat{g})$。典型的做法是构造经验四阶矩 $\sum_i Z_i Z_i' \hat{\eta}_i^2$，其中 $\hat{\eta}_i$ 是不考虑异方差问题时 2SLS 中的残差(见 White(1982)对具体的分布理论和细节的描述)。

通过求解 2SLS 的最小化元，我们还能得到检验过度识别问题的统计量。从直觉上看，由于假设要求 $\mathrm{E}[Z_i \eta_i] = 0$，所以这个统计量表示出样本矩向量 $m_N(\hat{g})$ 和 0 之间的距离是多少。而且，由于假设残差和工具变量不相关，那么最小化后的 $J_N(\hat{g})$ 满足 $\chi^2(Q-1)$ 分布。因此，通过比较 2SLS 最小化元的实际值和卡方分布表中的标准值，来对原假设 $\mathrm{H}_0: \mathrm{E}[Z_i \eta_i] = 0$ 做统计检验。

在本章上一节讨论瓦尔德估计值和分组数据之间的联系时，我们对一组互斥的虚拟变量做工具变量的情况表现出特殊的兴趣，这里我们对一组互斥的虚拟变量做工具变量时得到的 2SLS 最小化元也有特殊兴趣。在这个重要特例中，2SLS 估计值变为对分组数据所做的加权最小二乘估计值，2SLS 最小化元成为在最小二乘估计中我们希望最小化的那个平方和。为了看清这一点，假设一组互斥虚拟变量构成的工具变量可取 J 个值，产生 $N \times 1$ 维的拟合值向量，其中每

① 更多的细节可以参考 Newey(1985)，Newey 和 West(1987)，也可参考 Amemiya(1985)的高级教程以及 Hansen(1982)关于广义矩估计的原始论文。

个工具变量都表示一个分组数据，并对应于这个分组数据中拟合出的条件均值，那么易知在拟合值向量中每个分组数据的拟合值都会出现 n_j 次，并有 $\sum n_j = N$。在这个例子中工具变量的叉乘矩阵$[Z'Z]$是个 $J \times J$ 对角阵，对角线元素为 n_j。于是等式(4.2.6)可以简化为：

$$\hat{J}_N(\hat{g}) = \frac{1}{\sigma_\eta^2} \sum_j n_j (\bar{y}_j - \hat{g}' \bar{W}_j)^2 \tag{4.2.6}$$

这里 $\bar{W}_j$ 是在矩阵 W 中处于第 j 行的样本均值。由等式(4.2.6)可见，$\hat{J}_N(\hat{g})$ 是对 $\bar{y}_j$ 关于 $\bar{W}_j$ 进行广义最小二乘回归时得到的最小化元。对公式(4.2.6)再做一点小变化(这里我们省略一些细节)，就可以得到无需同方差假设时有效的两阶段工具变量过程的最小化元：

$$\hat{J}_N(\hat{g}) = \sum_j \left(\frac{n_j}{\sigma_j^2}\right)(\bar{y}_j - \hat{g}' \bar{W}_j)^2 \tag{4.2.7}$$

其中，σ_j^2 是第 j 组中 n_j 的方差。利用方程(4.2.7)进行的估计是可行的，因为当不考虑异方差问题时，2SLS 估计值虽然不是有效的，但却还是一致的，因此利用方程(4.2.7)进行估计时可以使用在第一阶段估计出的 σ_j^2。在 Angrist(1990，1991)中，作者为我们提供了满足有效性的两阶段工具变量估计值。

当 2SLS 中的工具变量由互斥的虚拟变量构成时，最小化元具有广义最小二乘的结构，这为我们解释过度识别统计量提供了新的角度：过度识别统计量实际上度量了将 $\bar{y}_j$ 和 $\bar{W}_j$ 连接起来的直线的拟合优度。换言之，这个统计量就是可视化工具变量回归散点图(图 4.2)中对回归直线的拟合优度所做的卡方检验。其中卡方分布的自由度等于工具变量的个数和待估参数①的个数之差。

当已经得到 2SLS 估计值时，我们可以用多种方式得到等式(4.2.7)表示的检验统计量。其中有两种方法值得大家了解。首先，无论工具变量是用来分组的虚拟变量还是其他情况，从工具变量的广义矩估计得到的最小化元等于检验过度识别问题的统计量，对后者的讨论可见以联立方程模型为主要内容的计量经济学参考书。比如，感兴趣的同学不妨查阅在计量经济学手册(Handbook of Econometrics)中由 Hausman(1983)写就的关于联立方程组问题的相关章节，那里就出现了我们在此提到的这个统计量。在该部分内容里，Hausman 提出了一个简单的计算过程：在同方差模型中，2SLS 最小化元等于样本大小乘以 R^2，其中 R^2 来自用 2SLS 的残差关于工具变量(以及外生的协变量)所进行的回归。具体的计算公式是 $N\left[\frac{\hat{\eta}' P_Z \hat{\eta}}{\hat{\eta}' \hat{\eta}}\right]$，其中，$\hat{\eta} = Y - W\hat{\Gamma}_{2SLS}$，它是 2SLS 的残差向量。

其次，我们还可以从“不同的方法，同一种结果”这一想法出发来诊断过度识别

① 比如，虚拟变量取三个值，其中之一是常数，模型包含一个常数和一个单一的内生变量，那么得到的检验统计量的自由度就是 1。

问题，这种方法也值得强调。当存在过度识别问题时，有不止一种工具变量可以对相同的因果关系进行识别，因此我们能构造多个恰好识别的工具变量并用这些工具变量对相同的因果关系进行识别，然后在不同估计结果之间进行比较。这种对估计结果进行的比较为我们提供了直接对过度识别问题进行检验的方法：如果每个恰好识别的估计值都是一致的，那么相对于样本方差，这些估计值之间的差别应该很小，而且随着样本规模的增大，估计的准确性会进一步提高，不同估计值之间的差距会进一步缩小。事实上，如果我们把“所有可能的恰好识别估计值都相等”看作原假设，那么对该原假设进行的检验实际上就是在构造瓦尔德检验①，而基于2SLS最小化元对该原假设进行的检验实际上是在构造拉格朗日乘子检验（Lagrange multiplier，LM），因为这个统计量与工具变量的极大似然估计中的得分向量有关。

在分组数据中使用工具变量估计法时，瓦尔德检验就是在检查所有线性独立的瓦尔德估计值是否相等。举个例子，我们将决定参军资格的随机数分为四组，分别是1—95，96—125，126—195和剩余的随机数，那么可以构造三个线性独立的瓦尔德估计值。相应的，对这四个分组数据中的条件均值做广义最小二乘回归，我们还可以构造出有效的分组数据估计值。将工具变量分为四组，意味着可以得到三个瓦尔德估计值，那么对这三个瓦尔德估计值都相等的假设实际上包含了两个等式约束，因此相应的瓦尔德统计量的自由度是2。从另一个角度讲，四个分组意味着可以用三个工具变量和一个常数来估计有两个参数的模型（比如在军队服役例子中，这两个参数就是常数项和表示因果效应的参数）。因此2SLS最小化元就可以产生一个自由度为$4-2=2$的过度识别检验统计量。而且，如果你估计加权矩阵时使用的方法与估计二次型中加权矩阵的方法相同，那么这两个统计量不仅在检验同一件事情，而且在数值上也相等。由于2SLS实际上是瓦尔德估计值②的线性组合，所以这种等价性是显然成立的。

最后，我们考虑在实际中使用过度识别检验需要注意的问题。由于$\hat{J}_N(\hat{g})$度量了经过方差调整的拟合优度，所以当估计出的参数不精确时，过度识别统计量倾向于偏低。由于工具变量估计值往往不准确，所以即使单个估计量显示出足够的精确性，我们也不能因为某个工具变量估计值和另外一个工具变量估计值相差不大而感到欣慰。从另一方面讲，当工具变量估计值相当精确时，过度识别检验拒绝

① 瓦尔德估计值和瓦尔德检验都是用来纪念Abraham Wald的，但是瓦尔德检验是在Wald(1943)中提出的。瓦尔德是计量经济学和数理统计学的巨匠，可惜的是他在48岁的时候死于一场飞机事故。

② 在线性模型中，对于相同的原假设，瓦尔德统计量和拉格朗日乘子检验的等价性由Newey和West(1987)建立。Angrist(1991)给出了一个更正式的论述。在这个背景下，Deaton(1985)提出了一个有趣的问题：当我们可以对数据分组时，分成多少组是最优的。分组数据和工具变量之间的相似性意味着更多的分组就有更多的工具变量，这种方式带来更高的渐进有效性，但却是以产生更多的估计偏误为代价的。Devereux(2007)针对分组数据存在多组的情况提出了一个对偏误进行修正的工具变量估计值。

原假设也并不意味着我们的识别策略错了。这可能是由于因果效应的异质性造成的，我们还会在本章后面部分对这个问题进行阐述。从概念的角度讲，对 2SLS 最小化元所具有的结构进行的理解是非常有价值的，因为它再次指出了分组数据和工具变量之间的联系。这一联系使得运用工具变量进行估计和假设检验的那种神秘性消失了，随之而来的是将注意力引向一般性的矩估计，因为这种估计为因果推断提供了基础。

4.3 双样本工具变量和剖分样本工具变量*

在上一节，我们用广义矩估计来解释 2SLS 估计值，在这种解释下我们发现无需使用微观数据(microdata)，只用样本矩就可以构造工具变量估计值了。回到样本矩条件(4.2.3)，重新整理后可以得到一个包含样本二阶矩，而且和回归方程很相似的等式：

$$\frac{Z'Y}{N}=\frac{Z'W}{N}\Gamma+\frac{Z'\eta}{N} \tag{4.3.1}$$

因为由矩条件可得 $E\left[\frac{Z'Y}{N}\right]=E\left[\frac{Z'W}{N}\right]\Gamma$，所以对等式(4.3.1)使用广义最小二乘估计后得到的系数是针对 Γ 的一致估计。

从等式(4.3.1)出发，我们还可以对 2SLS 最小化元作出新的阐释：对等式(4.3.1)做广义最小二乘，将需要被最小化的平方和乘以 $\sqrt{N}$，得到的就是 2SLS 最小化元。这里对平方和乘以 $\sqrt{N}$ 是为了保证相应的残差项不会因为样本规模的增大而消失。换言之，2SLS 的求解过程就是对一个二次型进行最小化，这个二次型是用等式(4.3.1)中的残差项和一个加权矩阵(该矩阵可能不是对角矩阵)构造出的。将 2SLS 问题转化成等式(4.3.1)的好处在于我们不需要用个体观察值来估计等式(4.3.1)。这个观点类似于我们在讨论最小二乘法时指出的结论——无需微观数据，运用来自样本的条件均值方程也能构造出最小二乘估计值。在这里，我们遇到的情况是——无需微观数据，用样本矩也能构造出工具变量估计值。在这一构造过程中，必须要有的样本矩是 $\frac{Z'Y}{N}$ 和 $\frac{Z'W}{N}$。被解释变量 $\frac{Z'Y}{N}$ 是一个 $[K+Q]\times 1$ 维的向量。回归元矩阵 $\frac{Z'\eta}{N}$ 是一个 $[K+Q]\times[K+1]$ 维的矩阵。除非 $Q=1$，否则工具变量的二阶矩方程有无数解，因此我们要尽可能地最小化残差项的二次型，以使二阶矩条件得到尽可能的满足。在这个目的下，构造残差项二次型的最有效加权矩阵就是 $\frac{Z'\eta}{N}$ 的渐进协方差矩阵。这样做的结果是得到的最小化元就是 2SLS 最小化元 $\hat{J}_N(\hat{g})$。

等式(4.3.1)左右两边都是样本矩的这一事实还启发我们得到了另一个结论：

假设等式(4.3.1)中等号左边和右边出现的矩阵来自于相同的总体，那么这两部分矩阵无需来自同一个数据集。基于这个看法，Angrist(1990)构造了双样本工具变量(two-sample instrumental variables，TSIV)估计值，之后在 Angrist 和 Krueger(1992)[①]中，两位作者正式提出了这种工具变量估计法。简便起见，用 Z_1 表示来自数据集 1 的工具变量/协变量构成的矩阵，用 Y_1 表示来自数据集 1 的被解释变量向量，数据集 1 的规模为 N_1；用 Z_2 表示来自数据集 2 的工具变量/协变量矩阵，用 W_2 表示来自数据集 2 的内生变量/协变量矩阵，数据集 2 的规模为 N_2。假设 $p\lim\left(\frac{Z_2'W_2}{N_2}\right)=p\lim\left(\frac{Z_1'W_1}{N_1}\right)$，对下式表示的双样本矩等式进行广义最小二乘估计得到的参数是 Γ 的一致估计。通过用 $\sqrt{N_1}$ 将上式正规化并假设 $p\lim\left(\frac{N_2}{N_1}\right)$ 是常数，我们还能得到这个估计值的渐进分布。

$$\frac{Z_1'Y_1}{N_1}=+\left\{\left[\frac{Z_1'W_1}{N_1}-\frac{Z_2'W_2}{N_2}\right]\Gamma+\frac{Z_1'\eta}{N_1}\right\}$$

双样本工具变量法的好处在于它拓宽了工具变量估计法的使用范围：即使很难在同一个数据集中找齐所有的解释变量，工具变量和感兴趣的内生变量，我们还是可以尝试运用工具变量法进行估计。比如我们可能会遇到下面的情况：某个数据集包含被解释变量和工具变量的信息，运用该数据集可以进行简约式估计，而另一个数据集则包含有关内生变量和工具变量的信息，可以对该数据集进行第一阶段回归。举个例子，在 Angrist(1990)中，来自社保管理系统(Social Security Administration，缩写为 SSA)的官方记录为我们提供了被解释变量(年收入)和工具变量(从出生日期中得到的随机抽取参军资格的数字以及诸如种族和出生年份等协变量)的信息。但是该记录不包含个体的服役情况，这个信息来自于军方的记录，而且军方的记录还包含了用以构造随机数的出生日期。Angrist(1990)使用军方的记录构造了 $\frac{Z_2'W_2}{N_2}$，这个矩阵表示第一阶段回归，也就是给定种族和出生年份后随机抽取的参军资格和服役经历之间的联系，而来自 SSA 的数据则用来构造 $\frac{Z_1'Y_1}{N_1}$。

下面讨论对双样本工具变量法的两个简化，从而使得该方法易于使用。第一，正如之前提到的，当工具变量由一系列互斥的虚拟变量构成时(类似于在 Angrist(1990)以及 Angrist 和 Krueger(1992)中出现的工具变量)，二阶矩方程(4.3.1)可被简化为条件均值模型。特别的，针对双样本问题的 2SLS 最小化元变为：

① 对双样本工具变量的应用包括 Bjorklund 和 Jantti(1997)，Jappelli，Pischke 和 Souleles(1998)，Currie 和 Yelowiz(2000)以及 Dee 和 Evans(2003)。在最近的论文中，Inoue 和 Solon(2009)比较了各种双样本工具变量估计值后提出了双样本工具变量的极大似然估计法(LIML)。他们同时纠正了 Angrist 和 Krueger(1995)在讨论分布理论时犯下的错误，这一点我们在后面还会提及。

$$\hat{J}_N(\hat{g}) = \sum_j \omega_j (\bar{y}_{1j} - \hat{g}' \bar{W}_{2j})^2 \tag{4.3.2}$$

其中，$\bar{y}_{1j}$ 是在第一个样本中当工具变量/协变量等于 j 时被解释变量的均值，$\bar{W}_{2j}$ 是在第二个样本中当工具变量/协变量为 j 时的内生变量和协变量的均值，ω_j 是相应的权重。等式(4.3.2)告诉我们：除了解释变量和被解释变量来自不同数据集之外，双样本工具变量估计值实际上就是可视化工具变量方程的加权最小二乘估计值。在 Angrist(1990)以及 Angrist 和 Krueger(1992)中，作者对这种方法进行了阐述。对于渐进有效的双样本工具变量法，等式(4.3.2)中的最优权重由 $\bar{y}_{1j} - \hat{g}' \bar{W}_{2j}$ 的方差给出。如果双样本工具变量法中的两个集合是相互独立的，那么这个方差便很容易计算。

第二，在 Angrist 和 Krueger(1995)中，两位作者引入了一个便于计算的双样本工具变量估计法，这种方法无需进行矩阵计算(matrix manipulation)，且在一般的回归软件中即可实现。在 Angrist 和 Krueger(1995)中，两位作者将得到的估计值称为剖分样本工具变量(split-sample IV，缩写为 SSIV)估计值，其工作原理如下[①]：在第一阶段估计拟合值时，先从数据集 2 的第一阶段估计中得到 $(Z_2'Z_2)Z_2'W_2$。然后将其拿到数据集 1 来构造跨样本拟合值 $\hat{W}_{12} \equiv Z_1(Z_2'Z_2)^{-1}Z_2'W_2$。在第二阶段时，用 $\hat{W}_{12}$ 对 Y_1 做回归。这个估计值的正确的渐进分布由 Inoue 和 Solon(2009)给出。他们指出要想得到 Angrist 和 Krueger(1992)中的分布，还需要加上额外的假设 $Z_1'Z_1 = Z_2'Z_2$（如果在重复抽样中工具变量和协方差的边际分布是固定的，那么这个假设是成立的）。但是值得注意的是，当内生变量前的系数为零时，SSIV 和 2SLS 的极限分布是相同的。在这个特例中对标准误的构造相当简单，而且估计出的标准误很可能为一般情况提供了一个很好的近似。

4.4 工具变量与异质性潜在结果

到现在为止，我们对工具变量的讨论尚局限在因果效应为常数的假设下。在用虚拟变量表征是否有服役经历的例子中，这个假设意味着对所有的 i，都有 $Y_{1i} - Y_{0i} = \rho_i$；在用可取多个值的变量来表征教育水平的例子中，这个假设意味着对所有的 s 和 i，都有 $Y_{si} - Y_{s-1,i} = \rho$。这两种假设都相当的理想化，特别是在我们关心的变量可以取多个值的情况下，因果效应为常数不仅假设了因果效应具有同质性，还假设了在我们关心的因果模型中因果效应是线性的。为了使异质性模型的讨论更加集中，我们每次只关注一个问题，先从因果变量只取 0—1 的最简单情形出发，此时因果变量是表征个体是否接受处理的虚拟变量。在这里，异质性体现在不同个

① Angrist 和 Krueger 将这个估计值叫做剖分样本工具变量估计值是因为在他们考虑的问题中故意将单个数据集剖分为两个数据集。正如在第 4.6.4 节中的讨论，由此得到的估计值可能比一般的 2SLS 估计值更无偏一些。Inoue 和 Solon(2009)将 Angrist 和 Krueger(1995)的估计值叫做双样本两阶段最小二乘归，简写为 2SLS 或者 TS2SLS。

体的因果效应不同，也即我们要用分布来描述这种异质性的因果效应。

为什么因果效应的异质性很重要？对这个问题的回答需要分清楚研究设计的两种不同类型的效度（validity）。一个是内部效度（internal validity），它考虑的是给定的研究设计是否成功地揭示了总体中令人感兴趣的因果效应。随机实验或者好的工具变量都能使研究的内部效度得到大幅提高。另一个是外部效度（external validity），它考虑的是当情况发生改变时研究结论的预测能力。比如说，如果随机实验中的研究总体特别愿意接受处理，因为这对他们有益，那么相应估计结果的外部效度便不会很好。比如，在利用随机抽取的参军资格作为工具变量的研究中我们估计出了越战服役对收入的影响，但是这个估计值无法度量志愿参军对收入的影响，因为志愿参军的那些人可能原本就在劳动力市场中处于劣势。利用针对异质性因果效应发展出的计量经济学框架，我们在运用工具变量时既能实现内部效度，还能兼顾外部效度。①

4.4.1 局部平均处理效应

在工具变量的框架中，使我们获得因果效应的原因在于工具变量 Z_i，但是我们感兴趣的变量却是 D_i。使用工具变量时表现出的这个特点使我们可以采取更为一般化的方式来描述潜在结果，这种方式将潜在结果表示为工具变量和处理状态的函数。记 $Y_i(d, z)$ 为个体 i 在处理结果为 $D_i = d$，工具变量取值为 $Z_i = z$ 时会选择的潜在结果。举个例子来说，$Y_i(d, z)$ 可以告诉我们给定个体 i 的处理状态和工具变量，他可能得到的工资是多少。当个体 i 已经知道自己是否获得了参军资格，那么他是否服役对其收入的影响就是 $Y_i(1, Z_i) - Y_i(0, Z_i)$；同样的，给定个体 i 的服役状态，他是否获得参军资格对其收入的影响就是 $Y_i(D_i, 1) - Y_i(D_i, 0)$。

我们可以认为工具变量的作用在于创造一个因果链条，在这个因果链条中工具变量 Z_i 影响我们感兴趣的变量 D_i，然后感兴趣的变量再去影响潜在结果 Y_i。更精确起见，这里引入一个记号来表达工具变量对 D_i 产生的因果效应。令 D_{1i} 表示工具变量 $Z_i = 1$ 时个体 i 的处理状态，令 D_{0i} 表示工具变量 $Z_i = 0$ 时个体 i 的处理状态。于是观察到的处理状态是：

$$D_i = D_{0i} + (D_{1i} - D_{0i})Z_i = \pi_0 + \pi_{1i}Z_i + \xi_i \tag{4.4.1}$$

如果用随机系数的记号来表示，那么有 $\pi_0 \equiv \mathrm{E}[D_{0i}]$ 和 $\pi_{1i} \equiv (D_{1i} - D_{0i})$，因此 π_{1i} 是工具变量带给 D_i 的异质性因果效应。因为我们现在讨论的是潜在结果，所以对任何一个人而言，都只可能观察到两个结果 D_{1i} 和 D_{0i} 中的一个。在随机抽取参军资格的例子中，D_{0i} 表示当个体 i 得到一个较大的随机数（即不太可能被征召入伍）

① 在社会科学中，针对内部效度和外部效度的争论已经持续了很长时间。比如 Shadish，Cook 和 Campell(2002)就用整个一章的篇幅来处理这个主题，它是对研究方法经典文献 Campbell 和 Stanley(1963)的传承。

时，他是否会在军队服役，D_{1i}表示当个体 i 得到一个较小的随机数（很有可能被征召入伍）时，他是否会在军队服役。根据 Z_i 的取值，我们只能看到潜在结果中的一种情况出现。Z_i 对 D_i 的因果效应的平均值就是 $\mathrm{E}[\pi_{1i}]$。

异质性因果效应框架下的第一个假设是工具变量应该起到随机分配的效果；也就是说，工具变量既与潜在结果无关，也与潜在处理状态无关。正式地讲就是：

$$[\{Y_i(d, z);\ \forall d, z\}, D_{1i}, D_{0i}] \perp\!\!\!\perp Z_i \tag{4.4.2}$$

这个独立性假设（independent assumption）足以保证我们对简约式赋予一个因果解释，这里的简约式是指对 Y_i 关于 Z_i 做回归。具体而言：

$$\begin{aligned}
&\mathrm{E}[Y_i \mid Z_i = 1] - \mathrm{E}[Y_i \mid Z_i = 0] \\
&= \mathrm{E}[Y_i(D_{1i}, 1) \mid Z_i = 1] - \mathrm{E}[Y_i(D_{0i}, 0) \mid Z_i = 0] \\
&= \mathrm{E}[Y_i(D_{1i}, 1) - Y_i(D_{0i}, 0)]
\end{aligned}$$

这个等式表示工具变量对 Y_i 的因果效应。独立性同时意味着：

$$\begin{aligned}
&\mathrm{E}[D_i \mid Z_i = 1] - \mathrm{E}[D_i \mid Z_i = 0] \\
&= \mathrm{E}[D_{1i} \mid Z_1 = 1] - \mathrm{E}[D_{0i} \mid Z_i = 0] \\
&= \mathrm{E}[D_{1i} - D_{0i}]
\end{aligned}$$

换言之，上式表示的 Z_i 对 D_i 的因果效应实际上就是在 2SLS 出现的第一阶段回归。

在异质性因果效应框架中的第二个关键假设是 $Y_i(d, z)$ 只是 d 的函数。具体而言，参军资格显然会影响到一个人是否服役，但是不论个体是否参军，他的收入都与参军资格无关。一般而言，这个假设要求工具变量通过唯一的已知途径对被解释变量产生影响，它就是排他性约束。正式地，排他性约束可以记为：

$$Y_i(d, 0) = Y_i(d, 1)，对于\ d = 0, 1$$

在因果效应为常数且在因果模型中是线性的时候，我们把工具变量排除在等式（4.1.14）表示的因果模型之外，并且宣称 $\mathrm{E}[Z_i\eta_i] = 0$，以此来实现排他性约束。但是值得一提的是在联立方程组模型中，传统的误差项无法让我们区分独立性假设和排他性约束。我们要求 Z_i 和 η_i 不相关，但是除非我们对独立性假设和排他性约束作出明确区分，否则要求 Z_i 和 η_i 不相关的原因不甚明朗。

在用随机抽取的参军资格作为工具变量的例子中，如果较低的随机数不仅提高了个体服役的概率，而且还从其他方面影响到这些人的潜在收入，那么排他性约束就无法满足。比如 Angrist 和 Krueger（1992）就考察了随机决定的参军资格与受教育水平之间的关系。他们的想法是这样的：以接受教育为理由可以逃避服役，因此对于那些获得的随机数较小的人而言，他们会选择在学校接受更长时间的教育。如果是这样，那么用于决定参军资格的随机数就通过至少两个途径影响了未来的潜在收入：第一个是它提高了个体在军队服役的概率；另一个途径是它提高了高等学校入学率。因此即使用以决定参军资格数字是随机分配的（因此满

足独立性假设），它也无法避免这两种机制同时影响潜在收入的问题。于是，“工具变量看上去就像随机分配的一样好”和“工具变量满足排他性约束”这两件事情之间就有了区别。排他性约束指的是工具变量只能通过唯一途径影响因果效应。①

使用排他性约束后，我们可以用不同的处理状态来定义潜在结果。具体而言就是：

$$\begin{aligned} Y_{1i} &\equiv Y_i(1, 1) = Y_i(1, 0) \\ Y_{0i} &\equiv Y_i(0, 1) = Y_i(0, 0) \end{aligned} \tag{4.4.3}$$

于是观察到的结果 Y_i 可以写为潜在结果的组合：

$$\begin{aligned} Y_i &= Y_i(0, Z_i) + [Y_i(1, Z_i) - Y_i(0, Z_i)]D_i \\ &= Y_{0i} + (Y_{1i} - Y_{0i})D_i \end{aligned} \tag{4.4.4}$$

用随机参数表示这个方程就是：

$$Y_i = \alpha_0 + \rho_i D_i + \eta_i$$

这是等式(4.4.4)的一种紧凑型的表示，其中 $\alpha_0 \equiv \mathrm{E}[Y_{0i}]$ 和 $\rho_i \equiv Y_{1i} - Y_{0i}$。

对异质性工具变量模型所做的最后一个假设是对所有的 i，要么 $\pi_{1i} \geqslant 0$，要么 $\pi_{1i} \leqslant 0$。这个单调性假设由 Imbens 和 Angrist(1994)提出，该假设允许一些人不被工具变量影响，但对于那些受工具变量影响的人，这种影响必须是以相同的方式发生的。换言之，对所有的 i，或者 $D_{1i} \geqslant D_{0i}$ 或者 $D_{1i} \leqslant D_{0i}$。在接下来的讨论中，假设 $D_{1i} \geqslant D_{0i}$。在用随机数决定参军资格的例子中，单调性假设意味着虽然对某些人而言参军资格不影响他入伍服役的概率，但是不存在获得参军资格而不允许其参军服役的事情。如果没有单调性假设，我们就无法保证工具变量估计值是对个体因果效应 $Y_{1i} - Y_{0i}$ 的加权平均值。

给定排他性约束，给定工具变量与潜在结果相互独立的独立性假设，如果第一阶段回归存在，而且工具变量对处理状态的影响满足单调性假设，瓦尔德估计值就可被解释为：对于受工具变量影响的那些人，由于在军队服役造成的收入的变化。这个参数就被称为局部平均处理效应(Local average treatment effect，LATE；见 Imbens 和 Angrist(1994))。下面的这个定理是对该结论的正式的叙述：

定理 4.4.1：局部平均处理效应定理(The LATE Theorem)。假设：

(假设 1，独立性) $[\{Y_i(D_{1i}, 1), Y_i(D_{0i}, 0), D_{1i}, D_{0i}\} \perp\!\!\!\perp Z_i$；

(假设 2，排他性) $Y_i(d, 0) = Y_i(d, 1) \equiv Y_{di}$，对于 $d = 0, 1$；

① 正如下面将要指出的，在 Angrist 和 Krueger(1992)的数据中，教育水平和随机数之间没有太多联系，这大概是因为在随机抽取数据的时候，通过延长教育时间来逃避服役的效果已经被平均化了。另外，在最近一篇论文中 Angrist 和 Chen(2007)指出由于老兵福利政策(比如 GI 法案)，越战老兵会获得更多的教育。但是由 GI 法案带来的越战老兵获得的更多教育并未违反排他性约束，因为享受老兵福利是军队服役的一个结果，本身并未影响人们是否服役的决策。

（假设 3，第一阶段）$\mathrm{E}[D_{1i}-D_{0i}]\neq 0$；

（假设 4，单调性）$D_{1i}-D_{0i}\geqslant 0\,\forall\, i$，当然这个假设中的不等号反过来也可；

于是：

$$\frac{\mathrm{E}[Y_i\mid Z_i=1]-\mathrm{E}[Y_i\mid Z_i=0]}{\mathrm{E}[D_i\mid Z_i=1]-\mathrm{E}[D_i\mid Z_i=0]}=\mathrm{E}[Y_{1i}-Y_{0i}\mid D_{1i}>D_{0i}]$$
$$=\mathrm{E}[\rho_i\mid \pi_i>0]$$

证明：由排他性约束可得 $\mathrm{E}[Y_i\mid Z_i=1]=\mathrm{E}[Y_{0i}+(Y_{1i}-Y_{0i})D_{1i}\mid Z_i=1]$，由独立性①，它应该等于 $\mathrm{E}[Y_{0i}+(Y_{1i}-Y_{0i})D_{1i}]$。同理可得 $\mathrm{E}[Y_i\mid Z_i=0]=\mathrm{E}[Y_{0i}+(Y_{1i}-Y_{0i})D_{0i}]$，因此瓦尔德估计值的分子就是 $\mathrm{E}[(Y_{1i}-Y_{0i})(D_{1i}-D_{0i})]$。单调性意味着 $D_{1i}-D_{0i}$ 等于 1 或者 0，因此：

$$\mathrm{E}[(Y_{1i}-Y_{0i})(D_{1i}-D_{0i})]=\mathrm{E}[Y_{1i}-Y_{0i}\mid D_{1i}>D_{0i}]\mathrm{P}[D_{1i}>D_{0i}]$$

相同的讨论指出：

$$\mathrm{E}[D_i\mid Z_i=1]-\mathrm{E}[D_i\mid Z_i=0]=\mathrm{E}[D_{1i}-D_{0i}]=\mathrm{P}[D_{1i}>D_{0i}]$$

这个定理指出如果工具变量能够像随机分配那样好，能够通过唯一的已知途径影响潜在结果，存在第一阶段，只通过一个方向影响因果性，那么这个工具变量就可以用来估计被其影响的群体的平均因果效应②。因此，在用随机抽取的参军资格做工具变量研究服役对收入影响的例子中，我们得到的瓦尔德估计值实际上捕捉到的是：对于那些仅仅因为在随机抽取中获得了参军资格而接受服役的人，服役对其收入的影响。这个估计值没有包含志愿参军对个人收入的影响，也没有在计算平均因果效应时考虑因病而免于服役的那些人。

局部平均处理效应的用处是什么？没有定理可以回答这个问题，但是这个问题值得讨论。我们感兴趣于越战服役对个人收入产生的影响，部分原因在于它可以回答这样一个问题：对退伍老兵（特别是对那些征召入伍的人）的补偿是否足以弥补他在为军队服务时做出的牺牲。我们可以使用具有内部效度的工具变量估计值来回答这个问题。而且，用随机抽取的参军资格作工具变量估计出越战征召入伍对收入的影响后，我们还可以用它来回答任何与未来征召政策有关的问题。从另一方面而言，虽然该估计值具有内部效度，但是就外部效度而言，也即时间和地点发生变化后该估计值的预测能力，在工具变量的框架中尚无法解决这个问题。在工具变量的框架中，无法解释为什么在越战中的服役会影响收入，对这个问题的

① 注意到在假设 1 中的独立性是对等式的简化，只涉及了满足 $Y_i(D_{Z_i}, Z_i)$ 的那些 $Y_i(d, z)$。

② 在本书中存在两个易混淆的名词，一是平均因果效应（average casual effect），另一个是平均处理效应（average treatment effect）。在同质性因果效应的框架中常用平均因果效应。在异质性因果效应的框架中，由于允许样本中存在不变工具变量影响的个体，因此估计出的因果效应只是来自受工具变量影响的那部分个体，精确起见，称其为平均处理效应。这两个词本质相同但在使用环境上存在微小区别。——译者注

解释需要一个理论。①

你可能在想为什么我们在局部平均处理效应定理中需要单调性，而在传统的常因果效应联立方程组模型中却没有这个假设。不满足单调性意味着工具变量使一些人的处理状态从 0 变成 1，却使得另外一个人的处理状态从 1 变成 0。在 Angrist，Imbens 和 Rubin(1996)中，作者将由于工具变量而使得处理状态从 1 变成 0 的那些人称为对抗者(defiers)。样本中存在的对抗者使局部平均处理效应和简约式之间的关系变得更为复杂。为了看清楚这一点，我们回到对局部平均处理效应定理的证明，在那里，简约式等于：

$$\mathrm{E}[Y_i \mid Z_i = 1] - \mathrm{E}[Y_i \mid Z_i = 0] = \mathrm{E}[(Y_{1i} - Y_{0i})(D_{1i} - D_{0i})]$$

如果不考虑单调性，上式等于：

$$\mathrm{E}[Y_{1i} - Y_{0i} \mid D_{1i} > D_{0i}]\mathrm{P}[D_{1i} > D_{0i}] - \mathrm{E}[Y_{1i} - Y_{0i} \mid D_{1i} < D_{0i}]\mathrm{P}[D_{1i} < D_{0i}]$$

由此我们发现即使所有人的因果效应都是正的，简约式估计值也还可能等于零，因为依从工具变量的个体的因果效应被对抗者的因果效应给抵消掉了。不过在常因果效应模型中，这个问题不会出现，因为无论第一阶段是否包括两种截然相反的行为②，简约式估计值始终是常因果效应与第一阶段的结果相乘。

如今在计量经济学领域流行着一种叫做潜在得分模型(latent index model)的方法，通过将局部平均处理效应与内生变量是虚拟变量(用该虚拟变量来表示处理状态)的潜在得分模型相联系，我们可以对局部平均因果效应有一个更深刻的理解。潜在得分模型描述的是个体通过比较效用或者成本来做决策的过程，但是这些效用或者成本部分可见，部分不可见(因此将这个模型叫做潜在得分模型，如见 Heckman(1978))。特别的，这些不可见的效用或者成本与最后的选择结果是有关系的，因此处理变量就具有了内生性(尽管从联立方程组模型的角度来看这个变量实际上并非内生的)。比如，我们可以将服役状态写为：

$$D_i = \begin{cases} 1 & \text{if } \gamma_0 + \gamma_1 Z_i > \nu_i \\ 0 & \text{其他} \end{cases}$$

其中，ν_i 是个随机元，用于表示在军队服役中看不到的成本和收益，假设这个随机

① Angrist(1990)用随机抽取的参军资格作为工具变量得到的估计值进行了解释，他指出由于参军导致的收入下降是一种惩罚，因为人们在军队服役期间失去了在劳动力市场中获得相关工作经验的机会。这意味着可以将该工具变量估计值推广到其他情形，从而使得这个估计值具有一定的外部效度，这个推测来自于 Angrist 和 Krueger(1994)对二战时期入伍士兵的研究结果。

② 如果是常因果效应 ρ，那么：

$$\begin{aligned} &\mathrm{E}[Y_{1i} - Y_{0i} \mid D_{1i} > D_{0i}]\mathrm{P}[D_{1i} > D_{0i}] - \mathrm{E}[Y_{1i} - Y_{0i} \mid D_{1i} > D_{0i}]\mathrm{P}[D_{1i} < D_{0i}] \\ &\quad = \rho\{\mathrm{P}[D_{1i} > D_{0i}] - \mathrm{P}[D_{1i} < D_{0i}]\} \\ &\quad = \rho\{\mathrm{E}[D_{1i} - D_{0i}]\} \end{aligned}$$

于是简约式中得到零意味着或者第一阶段估计结果为零，或者 $\rho = 0$。

元和工具变量 Z_i 相互独立。于是潜在得分模型用下面的方式刻画对处理状态的分配：

$$D_{0i}=1[\gamma_0>\nu_i] \text{ 以及 } D_{1i}=1[\gamma_0+\gamma_1>\nu_i]$$

注意到这个模型中单调性自动得到满足，因为 γ_1 是常数。假设 $\gamma_1>0$，于是可以将局部平均处理效应描述为：

$$\mathrm{E}[Y_{1i}-Y_{0i}\mid D_{1i}>D_{0i}]=\mathrm{E}[Y_{1i}-Y_{0i}\mid \gamma_0+\gamma_1>\nu_i>\gamma_0]$$

它是潜在的第一阶段参数 γ_0 和 γ_1 的函数，当然这个函数形式还与 $Y_{1i}-Y_{0i}$ 和 ν_i 的分布有关。一般而言，由此得到的局部平均处理效应与无条件的平均因果效应 $\mathrm{E}[Y_{1i}-Y_{0i}]$ 不同，也不等于被处理者的平均因果效应 $\mathrm{E}[Y_{1i}-Y_{0i}\mid D_i=1]$。在下一节我们考虑这两种不同的因果效应之间的区别。

4.4.2 依从工具变量的子集

在局部平均处理效应框架下，依照总体中个体对工具变量作出的反应，任何总体都可被分为三类子集：

定义 4.4.1：

依从工具变量者(compliers)：满足 $D_{1i}=1$ 和 $D_{0i}=0$ 的子集；

始终接受者(always-takers)：满足 $D_{1i}=1$ 和 $D_{0i}=1$ 的子集；

从不接受者(never takers)：满足 $D_{1i}=0$ 和 $D_{0i}=0$ 的子集。

局部平均处理效应便是依从工具变量者的平均因果效应。名词“依从工具变量者”来源于随机实验，在那里某些实验对象服从随机分配的处理方案(也即按照随机分配的方案服用药物)，但是有些人并不会这样做，而且被分配进入实验控制组的人也有可能进入实验处理组。如果随机实验要求某人服用药物，但他拒不服用，那么此人就被称为从不接受者，如果即使把某人放入实验控制组，他还是会服用药物，那么此人就被称为始终接受者。如果不对因果效应加入额外一些假设(比如常因果效应)，局部平均处理效应就无法估计从不接受者和始终接受者的因果效应。因为根据定义，工具变量并未使这两类人的处理状态发生改变。在依从工具变量者只占随机实验总体的一部分时，工具变量和随机实验之间的关系在于：工具变量解决了此类随机实验中的因果推断问题。这一点十分重要，值得我们用 4.4.3 整一个小节的篇幅进行讨论。

在转入重要特例之前，我们先来阐述些一般性观点。首先，依从工具变量者的平均因果效应往往不等于接受处理的人(也就是 $D_i=1$ 的那些人)的平均处理效应。根据 $D_i=D_{0i}+(D_{1i}-D_{0i})Z_i$，我们知道在总体中接受处理的个体由两个不相交的子集构成。根据单调性，我们无法同时得到 $D_{0i}=1$ 和 $D_{1i}-D_{0i}=1$，因为单调性要求 $D_{1i}\geqslant D_{0i}$，这意味着从 $D_{0i}=1$ 可以推出 $D_{1i}=1$，这与 $D_{1i}-D_{0i}=1$ 相矛盾。因此总体中接受处理的个体或者是 $D_{0i}=1$ 的人，或者是满足 $D_{1i}-D_{0i}=$

1 且 $Z_i = 1$ 的人，因此 $D_i = 1$ 可以写为两个互斥虚拟变量 D_{i0} 与 $(D_{1i} - D_{0i})Z_i$ 之和。换言之，总体中接受处理的那类人由两部分人组成，一部分是始终接受者，另一部分是当工具变量 $Z_i = 1$ 时愿意接受处理的人，这类人和工具变量 $Z_i = 0$ 时选择不接受处理的那部分人合起来构成依从工具变量者。既然工具变量能够像随机实验那样好，那么在工具变量 $Z_i = 1$ 时愿意接受处理的那部分依从工具变量者就能代表所有的依从工具变量者，由此我们有：

$$\begin{aligned}&\underbrace{\mathrm{E}[Y_{1i} - Y_{0i} \mid D_i = 1]}_{\text{接受处理的效应}}\\ &= \underbrace{\mathrm{E}[Y_{1i} - Y_{0i} \mid D_{0i} = 1]}_{\text{始终接受者的效应}}\mathrm{P}[D_{0i} = 1 \mid D_i = 1]\\ &\quad + \underbrace{\mathrm{E}[Y_{1i} - Y_{0i} \mid D_{1i} > D_{0i}]}_{\text{依从接受者的效应}}\mathrm{P}[D_{1i} > D_{0i},\ Z_i = 1 \mid D_i = 1] \qquad (4.4.5)\end{aligned}$$

由于 $\mathrm{P}[D_{0i} = 1 \mid D_i = 1]$ 和 $\mathrm{P}[D_{1i} > D_{0i},\ Z_i = 1 \mid D_i = 1]$ 加起来等于 1，所以接受处理的个体的因果效应乃是依从工具变量者和始终接受者的因果效应的加权平均值。

类似的，未接受处理的个体的平均因果效应是 $\mathrm{E}[Y_{1i} - Y_{0i} \mid D_i = 0]$，它也不等于局部平均处理效应。在使用随机抽取的参军资格做工具变量的例子中，未受处理的个体的平均因果效应等于没有选择服役的那些人如果当时选择了服役，其收入水平因此而受到的影响。具体而言，未受处理的个体的平均因果效应等于依从工具变量者和从不接受者的因果效应的加权平均。也就是：

$$\begin{aligned}&\underbrace{\mathrm{E}[Y_{1i} - Y_{0i} \mid D_i = 0]}_{\text{未受处理的效应}}\\ &= \underbrace{\mathrm{E}[Y_{1i} - Y_{0i} \mid D_{0i} = 0]}_{\text{从不接受者的效应}}\mathrm{P}[D_{1i} = 0 \mid D_i = 0]\\ &\quad + \underbrace{\mathrm{E}[Y_{1i} - Y_{0i} \mid D_{1i} > D_{0i}]}_{\text{依从接受者的效应}}\mathrm{P}[D_{1i} > D_{0i},\ Z_i = 1 \mid D_i = 0] \qquad (4.4.6)\end{aligned}$$

其中，我们注意到由于单调性，$D_{1i} = 0$ 的人一定是从不接受者。

最后，将等式(4.4.5)和等式(4.4.6)加权平均后可得：

$$\begin{aligned}\mathrm{E}[Y_{1i} - Y_{0i}] = {} & \mathrm{E}[Y_{1i} - Y_{0i} \mid D_i = 1]\mathrm{P}[D_i = 1]\\ & + \mathrm{E}[Y_{1i} - Y_{0i} \mid D_i = 0]\mathrm{P}[D_i = 0]\end{aligned}$$

上式指出无条件的平均因果效应乃是对依从工具变量者、始终接受者和从不接受者的因果效应的加权平均值。当然，给定单调性和定义(4.4.1)，我们也可以直接得到这个定理。

由于工具变量无法区分始终接受者和从不接受者，所以我们无法使用工具变量来估计所有受到处理的个体的因果效应，也无法估计未受处理的个体的因果效应。不过这个结论也有特例：如果总体中所有个体都依从工具变量，不存在始终接

受者和从不接受者，那么我们就可以用工具变量计算受到处理的个体的因果效应和未受处理的个体的因果效应。这种情况虽不常见，但它是重要的特例。这种特例的典型例子就是在 Rosenzweig 和 Wolpin(1980)，Bronars 和 Grogger(1994)，Angrist 和 Evans(1998)以及 Angrist，Lavy 和 Schlosser(2006)中都使用过的用双胞胎做工具变量研究生育的影响。另一个典型例子是 Oreopoulos(2006)使用义务教育法中发生的变化作为工具变量，对英国教育回报进行的估计。

为了在使用双胞胎做工具变量的这个例子中看清楚如何用工具变量估计未受处理的个体的因果效应，记 T_i 是表示第二胎出现多胞胎的虚拟变量。在 Angrist 和 Evans(1998)的研究中，他们在已有至少两个孩子的女性中估计了生育三个孩子对该女性收入造成的影响。在此例中，女性生育第三个孩子显得特别有趣，因为在 20 世纪 60 和 70 年代之间，美国家庭中女性生育数的下降主要是从生三个孩子向生两个孩子的转变。在这里，第二胎生育时出现双胞胎就为这种转变提供了一个准实验。令 Y_{0i} 表示女性生育两个孩子时的潜在收入，令 Y_{1i} 表示妇女生育三个孩子时的潜在收入，并将女性生育三个孩子这件事记为 D_i。假设 T_i“就像”随机分配的那样好，那么生育双胞胎意味着女性多生了一个孩子，由于生育双胞胎只通过影响生育数来影响个体收入，因此使用双胞胎工具变量 T_i 得到的局部平均处理效应就是 $E[Y_{1i}-Y_{0i} \mid D_i=0]$，它是生育了两个孩子的女性如果生育第三个孩子所带来的收入变化。该局部平均处理效应与等式(4.4.6)有所不同，原因在于第二胎生育双胞胎的所有女性最终都会有三个孩子，这个工具变量不存在从不接受者。

Oreopoulos(2006)也使用工具变量估计了未受处理者的平均因果效应。在这项研究中，他观察到英国义务教育法曾经将强制接受教育的年龄从 14 岁提高到 15 岁，利用这个变化，他估计了接受教育的经济回报。因为很多孩子可能在 14 岁之前就已经辍学了，所以不是所有人都严格遵守英国新的义务教育法。在这个例子中，作者感兴趣的因果效应是多接受一年高中教育带来的收入增长。我们可以将义务教育法修改后多增加的一年义务教育看做一种处理。由于在 Oreopoulos 的样本中，当义务教育法加强时所有人都接受了新增加的一年教育，所以他的样本中不存在从不接受者。因此，Oreopoulos 的工具变量估计策略就是要捕捉那些本来在 14 岁时就要离开学校的学生，由于加强了的义务教育法使他们多接受一年教育所带来的收入变化。Oreopoulos 的工具变量估计策略还要依赖于一个事实：英国的年轻人是著名的守法公民，但是这个事实在以色列则不成立，因此 Oreopoulos 的工具变量估计策略无法用在对以色列人教育的经济回报的研究中。在以色列，那里的学生对义务教育法的态度相对随意一些。因此，以色列的计量经济学家使用义务教育法的变化作为工具变量时就要注意到估计出来的值是局部平均处理效应而不是未受处理者的因果效应。

4.4.3 随机实验中的工具变量

在我们提出局部平均处理效应的分析框架时，使用的语言都来自于随机实验，

这是因为工具变量法和随机实验之间有很多相似之处。当然工具变量和随机实验之间的区别并不是泾渭分明的，因为很多工具变量本身就来自于随机实验。如果工具变量表示的是对处理状态的随机分配，那么局部平均处理效应就是依从工具变量但未受处理的人如果被处理，将会给他们带来的因果效应。这里，一个非常重要的例子就是工具变量产生于随机实验，而且在随机实验中只有一类不依从工具变量者①。在很多随机实验中都会出现这种情况：在随机分配了处理之后，那些选择接受处理的人往往是自愿的，同时在控制组中无人受到随机实验的处理。由于个体选择是否接受随机分配的处理本身就存在自选择(self-selected)问题，所以简单地比较处理组和控制组之间状况的差距并以此作为因果效应，可能存在较大的误导性。在这种情况下选择偏误几乎一定是正的：在随机实验中接受药物治疗的人本来就比较健康；通过接受培训项目提高自己收入能力的人本来的赚钱能力就比较强。

将随机分配的处理记做 Z_i，将个体接受的处理记为 D_i，用 Z_i 作为 D_i 的工具变量就可解决此类因为随机实验中存在不依从处理而带来的自选择问题。这时，局部平均处理效应就是针对接受处理者计算出的平均因果效应。假设工具变量 Z_i 为虚拟变量，表示随机分配来的处理，记 D_i 也是虚拟变量，表示个体是否接受处理。在实际中，由于存在不依从随机分配的处理的情况，所以 D_i 和 Z_i 之间可能不等。举个例子，在通过随机实验对 JTPA 培训项目进行评估的研究中，研究者随机派送出参加培训的资格，但是在那些获得该资格的人里只有 60%的人最终选择参加培训项目，同时控制组中也有 2%的人获得这种资格(见 Bloom 等(1997)；同样可以见 7.2.1 节)。可能是因为获得培训资格的人对这种培训不感兴趣，或者是因为实施这个项目的人没能成功说服人们参加培训项目，才导致 JTPA 培训项目中存在如此多不愿接受培训的人。由于不服从随机分配的现象主要发生在处理组而非控制组，所以用随机分配的 Z_i 作为实际接受的处理状态 D_i 的工具变量，由此得到的局部平均处理效应就是针对接受处理者计算出的平均因果效应。

表 4.5 就阐述了如何通过工具变量法来解决随机实验中存在的不依从分配问题，该表报告的是来自 JTPA 实验的结果。在 JTPA 实验中，我们主要关心的变量是接受随机处理后的个体在 30 个月内的总收入。表的第 1 列和第 2 列报告了接受培训和没有接受培训的人在收入方面的差异(第 2 列中结果来自于最小二乘估计，在该回归方程中我们加入了实验初期的一些个体特征作为协变量进行控制)。比较第 1 列和第 2 列的结果，我们发现接受培训使男性收入增加了将近 4 000 美元，使女性收入增加了将近 2 200 美元，无论对于男女，接受培训都将其平均收入提高了 20%，这是个相当大的影响。但是这个估计值可能存在误导性，因为这个结果是根据 D_i——实际得到的处理状态——进行比较的。由于在处理组中，拥有

① 也即在不依从工具变量的那些人中，随机实验只产生从不接受者或者始终接受者，这种只存在一类不依从工具变量者的情况，用英文表达为 one-sided noncompliance。——译者注

参加培训资格的那些人有权利选择不接受培训(有 40%的人确实这么做了),因此,除非个体做出的是否接受培训的决策与其潜在的收入水平无关,否则这种比较便有违随机分配的本意。但是看上去我们无法说明个体是否接受培训与其潜在的收入水平无关。

表 4.5 从 JTPA 实验中得到的结果:对培训项目的最小二乘估计值和工具变量估计值

	用接受培训状况进行比较(OLS)		用接受处理状况进行比较(ITT)		工具变量估计值(IV)	
	没有协变量	存在协变量	没有协变量	存在协变量	没有协变量	存在协变量
	(1)	(2)	(3)	(4)	(5)	(6)
A. 男性	3 970	3 754	1 117	970	1 825	1 593
	(555)	(536)	(569)	(546)	(928)	(895)
B. 女性	2 133	2 215	1 243	1 139	1 942	1 780
	(345)	(334)	(359)	(341)	(560)	(532)

注:本表来自于作者对 JTPA 研究数据的重新编制。该表表示 JTPA 实验中受补贴的培训项目对个体收入的影响,并分别报告了用最小二乘估计值(OLS)、ITT 估计值和工具变量估计得到的对该影响的估计值。第 1 列和第 2 列根据被处理状况进行比较;第 3 列和第 4 列使用随机分配的处理状况进行比较。第 5 列和第 6 列报告了使用随机分配的状态作为工具变量后进行的比较。在第 2、4、6 列中使用的协变量是高中或大学毕业、黑人、西班牙裔、已婚、在过去一年中工作时间少于 13 周、AFDC 以及表征 JTPA 服务策略、年龄分组和再次调研。括号中给出的是稳健标准误。样本中有 5 102 个男性和 6 102 个女性。

表 4.5 的第 3 列和第 4 列按照个体在随机分配中是否获得参加培训的资格进行了比较。换言之,这个比较是基于随机分配的变量 Z_i 作出的。在临床实验的语言中,第 3 列和第 4 列的比较被称为意向治疗(Intention to treat, ITT)效应。在表中我们看到意向治疗效应大致处在 1 200 美元的水平(当控制协变量后,这个值有所下降)。由于很多在随机分配中获得参加培训的资格的人后来没有参加培训,所以基于随机分配的 Z_i,我们可以对意向治疗效应赋予一个因果解释:它告诉我们个体接受培训所带来的因果效应。正是由于这个原因,相对于针对实际接受培训的人计算出的因果效应,意向治疗效应显得相当小。表 4.5 中的第 5 列和第 6 列将这两种情况结合在一起,为我们提供了最有趣的效应:用处理组和控制组之间参与率的差别(大约是 0.6)去除意向治疗效应,得到的数字大约是 1 800 美元,这个数字度量参加培训项目的人仅仅因为参加该项目而获得的收入提升。

我们如何知道将意向治疗效应除以参与程度就能得到参加培训项目的人的因果效应?我们可以将意向治疗效应看作随机分配处理后在简约式中得到的因果效应。参与率就是与工具变量相联系的第一阶段值,于是瓦尔德估计值就是简约式效应除以第一阶段估计值。一般而言,正是由于这个例子中(几乎)没有始终接受者,而且被处理的总体包含了所有的响应工具变量者,因此我们可以将这个值看作局部平均处理效应。故而,表 4.5 中第 5 列和第 6 列得到的工具变量估计值就一致性地估计出了被处理者的处理效应。

这个结论非常重要，它值得我们用另外一种方式进行严格讨论。囿于我们的见识所限，在随机实验中只存在一类不依从工具变量者的情况下，由 Howard Bloom(1986)最早提出了可以用工具变量法来估计被处理者的处理效应。这里是对 Bloom 结论的一个简单而又直接的证明。

定理 4.4.2：Bloom Result。*假设局部平均处理效应定理所要求的假设都成立，并且有* $E[D_i \mid Z_i = 0] = P[D_i = 1 \mid Z_i = 0] = 0$。*那么有*：

$$\frac{E[Y_i \mid Z_i = 1] - E[Y_i \mid Z_i = 0]}{P[D_i = 1 \mid Z_i = 1]} = E[Y_{1i} - Y_{0i} \mid D_i = 1]$$

证明：由于 $Z_i = 0$ 意味着 $D_i = 0$，故 $E[Y_i \mid Z_i = 1] = E[Y_{0i} \mid Z_i = 1] + E[(Y_{1i} - Y_{0i})D_i \mid Z_i = 1]$ 以及 $E[Y_i \mid Z_i = 0] = E[Y_{0i} \mid Z_i = 0]$。由独立性我们可知 $E[Y_{0i} \mid Z_i = 0] = E[Y_{0i} \mid Z_i = 1]$，于是有：

$$E[Y_i \mid Z_i = 1] - E[Y_i \mid Z_i = 0] = E[(Y_{1i} - Y_{0i})D_i \mid Z_i = 1]$$

但是由于 $D_i = 1$ 意味着 $Z_i = 1$，而且 $Z_i = 0$ 时没有人被处理，于是：

$$\begin{aligned} &E[(Y_{1i} - Y_{0i})D_i \mid Z_i = 1] \\ &= E[Y_{1i} - Y_{0i} \mid D_i = 1, Z_i = 1]P[D_i = 1 \mid Z_i = 1] \end{aligned}$$

因此，$E[Y_{1i} - Y_{0i} \mid D_i = 1, Z_i = 1] = E[Y_{1i} - Y_{0i} \mid D_i = 1]$。

局部平均处理效应除了可以帮助我们分析随机实验中存在不服从分配时的因果效应，这个分析框架还为由于客观原因或者道德原因无法让所有被处理的人都服从处理的情况提供了设计实验的指导。在犯罪学中具有相当创造性的一项研究乃是明尼阿波利斯市家庭暴力实验(Minneapolis Domestic Violence Experiment，MDVE)。这项实验的目的在于确定针对家庭暴力的最优政策(Sherman and Berk，1984)。一般而言，针对家庭暴力的政策包括一系列的策略，这些策略包括求助于社会辅导、分居令以及逮捕。由于研究者们注意到人们往往会放弃对家庭袭击的指控，因此对家庭暴力采取严厉措施——逮捕以及拘留——是不是有效一直存在着争议。

这场争议的一个结果就是明尼阿波利斯市当局进行了一项随机实验，在实验中，对某些家庭暴力的干预措施是随机分配的。在对这项随机实验进行设计时，研究者们先对彩色号码牌进行随机洗牌，然后进行随机抽取，抽到的纸牌上的号码对应于出警的某个警察，纸牌的颜色则告诉警察是将家庭暴力实施者进行逮捕，还是仅仅让两人分居或者寻求第三方进行辅导和干预。但是在实际操作中，警察实施干预时可以不遵循随机分配的结果。比如，对那些极端危险或者酗酒喝醉的侵犯者，无论抽到的纸牌指示警察采用何种措施，对该侵犯者都应该实施逮捕。因此，虽然随机抽取的干预措施和最终实施的干预措施之间存在高度相关，但有时两者存在不一致。

从使用明尼阿波利斯市家庭暴力实验数据进行研究并已得到发表的大部分论文来看，研究者们注意到了其中存在的问题，因此他们的研究都关注于对意向治疗

效应的估计，也就是说他们使用原始的随机分配进行估计，而不是针对被处理结果进行比较。但是本节的讨论指出还可以用明尼阿波利斯市家庭暴力实验数据来估计响应工具变量者的平均因果效应，也就是针对那些本该被逮捕，仅仅因为随机分配的指令漏网的人，计算如果这些人被逮捕，将会有何种结果。Angrist(2006)就沿着这个思路进行了研究。由于在明尼阿波利斯市家庭暴力实验中每个被随机赋予逮捕的人最终都被逮捕了，所以这部分人中没有不服从者，因此我们可以对Bloom定理进行一个有趣的转化：这里对每个人都有 $D_i = 1$。相应的，局部平均处理效应就是对未处理者的处理效应，也即：

$$E[Y_{1i} - Y_{0i} \mid D_{1i} > D_{0i}] = E[Y_{1i} - Y_{0i} \mid D_i = 0]$$

其中，D_i 表示是否被捕。使用明尼阿波利斯市家庭暴力实验数据并运用工具变量得到的估计值指出，逮捕家暴实施者的平均因果效应是：如果将那些在随机实验中本该逮捕却未逮捕的人抓起来，重新施暴的可能性会大大降低①。

4.4.4 计算并考察依从工具变量者所具有的特征

我们已经看到，除非出现特例，否则每个工具变量都对应于一个依从该工具变量者组成的集合，并识别出一个想要的参数。因此，至少从一般意义上（一个重要的特例就是存在一类完全依从工具变量的个体）来看，针对相同因果关系的不同工具变量估计出的参数之间会存在差异。虽然在因果效应同质性的假设下，2SLS可以对不同工具变量的估计值进行加权平均并产生一个平均因果效应，但是在完全异质性因果效应的框架下，我们在4.2.2节中讨论过的过度识别问题就不再会出现了，因为在那里我们根据不同工具变量是不是在估计同一件事情来进行过度识别，而在异质性框架下，不同工具变量确是在针对不同个体进行估计。

由于依从工具变量者不同，所以不同工具变量估计出的因果效应是不同的。因此我们希望尽可能地了解依从不同工具变量者之间的差别。更进一步，如果依从某个工具变量的个体所组成的集合与我们在别的研究中感兴趣的总体很类似，那么我们就可以很好地将该工具变量估计值的解释力推广到另外的研究中。在这个思路下，Acemoglu 和 Angrist(2000)指出，从本质上讲，出生季度和义务教育法（特别是对在出生所在州完成教育的最小年龄要求）两个工具变量通过相同的途径影响了相同的人群。因此可以认为无论用出生季度还是义务教育法做工具变量，估计出的因果效应应该是相同的。我们还应该想到，用出生季度做工具变量得到

① 在第2章提到的Krueger(1999)也用工具变量法分析了来自随机实验的数据。特别的，对于在一年级以及更高年级的学生，由于家长和老师会在实验开始后对学生所在的班级进行调动，所以真实的班级规模会与一开始随机分配的班级规模有差别。因此这项研究使用来自田纳西州STAR实验得到的数据，用随机分配的课堂规模作为工具变量估计真实课堂规模带来的因果效应。在Krueger(1999)中，作者也使用2SLS估计了变量的处理密度，这部分讨论见第4.5.3节。

的估计值可以用来预测加强义务教育法所带来的影响。

从另一方面讲，如果依从不同工具变量的个体之间差别很大，但是这些工具变量估计值却很相似，那么我们应该考虑这个问题中是不是存在着同质性的因果效应，并考虑是否运用同质性因果效应的框架进行估计。这样做的结果就是我们可重新利用过度识别问题来检验同质性因果效应假设是否成立，但与之前在同质性因果效应框架下的讨论不同，这时我们在某个工具变量下得到的估计值可以推广到依存其他工具变量者所成的总体上，从而使得研究结果具备一定的外部效度①。Angrist，Lavy 和 Schlosser(2006)对家庭规模如何影响儿童教育的研究就阐述了这种想法。他们的研究基于对现象的观察以及理论研究的需要。从现象上来看，在规模较大的家庭中，孩子们往往接受较少的教育；在规模较小的家庭中，孩子们接受的教育则较多。从理论研究的角度来看，对生育进行的研究中存在一个长期困扰人们的问题，即家庭和教育水平之间负相关的关系是不是具有因果性。随着研究的深入，Angrist，Lavy 和 Schlosser 使用了一系列的工具变量，每个工具变量都对应于一个依存工具变量者组成的集合，这些集合之间差别很大。但是，工具变量估计值却指出家庭规模没有对孩子的教育水平产生影响。Angrist，Lavy 和 Schlosser(2006)指出这个结论指向的内容是：在被研究的以色列人口中，家庭规模对孩子教育水平的影响为零②，这个结论具有普遍性。

我们已经看到，对依存工具变量者所成集合的规模进行度量是很容易的。给定单调性，它正好就是瓦尔德估计的第一阶段值，即：

$$\begin{aligned}\mathrm{P}[D_{1i} > D_{0i}] &= \mathrm{E}[D_{1i} - D_{0i}] \\ &= \mathrm{E}[D_{1i}] - \mathrm{E}[D_{0i}] \\ &= \mathrm{E}[D_{1i} \mid Z_i = 1] - \mathrm{E}[D_{0i} \mid Z_i = 0]\end{aligned}$$

我们还可以指出被处理组中有多少人是依从工具变量者。因为对于依从工具变量者而言，他们的处理状态完全由 Z_i 决定。从条件概率公式的定义出发，我们有：

$$\begin{aligned}\mathrm{P}[D_{1i} > D_{0i} \mid D_i = 1] &= \frac{\mathrm{P}[D_i = 1 \mid D_{1i} > D_{0i}]\mathrm{P}[D_{1i} > D_{0i}]}{\mathrm{P}[D_i = 1]} \\ &= \frac{\mathrm{P}[Z_i = 1](\mathrm{E}[D_i \mid Z_i = 1] - \mathrm{E}[D_i \mid Z_i = 0])}{\mathrm{P}[D_i = 1]} \qquad (4.4.7)\end{aligned}$$

等式(4.4.7)中的第二个等号来自于 $\mathrm{P}[D_i = 1 \mid D_{1i} > D_{0i}] = \mathrm{P}[Z_i = 1 \mid D_{1i} > D_{0i}]$。由独立性，我们还可以得到 $\mathrm{P}[Z_i = 1 \mid D_{1i} > D_{0i}] = \mathrm{P}[Z_i = 1]$。换言之，在被处理组中响应工具变量者的比例应该等于第一阶段估计值乘以工具变量 $Z_i = 1$ 的概率再除以样本中个体被处理的概率。

我们通过在有越战服役经历的人中计算依从工具变量者的规模，来阐述如何

① 事实上，如果在一个过度识别模型中所有工具变量都起作用，那么传统的过度识别检验就变为对因果效应是否同质的检验。

② 同时可见 Black，Devereux 和 Salvones(2005)对挪威的研究。

使用方程(4.4.7)。方程(4.4.7)中各部分的数据报告在表4.6中的第1行和第2行中。比如,对于出生在1950年的白人而言,第一阶段数值是0.159,获得参军资格的概率(也就是$P[Z_i = 1]$)是$\frac{195}{365}$,实际接受服役的边际概率是0.267。从这些统计数值出发,我们计算出实际接受服役的个体中依从工具变量者所占的比例是0.32。我们同时还对生于1950年的非白人计算出了这个值,结果在有越战服役经历的非白人中,依从工具变量者的比例下降至20%。这并不令人惊讶,因为对于非白人,第一阶段估计出的获得参军资格的概率相当小。表4.6第1行和第2行的最后一栏分别针对白人和非白人报告了:在没有越战服役经历的人中,获得参军资格就去参军的个体比例。对于非白人而言,这个数字是3%,对于白人而言这个数字是10%,该数字反映出的事实是:在没有越战服役经历的人中,大部分人因为豁免参军、未获得参军资格或者不符合服役的要求而没有入伍。

在Angrist(1990)的研究中,他关心的主要参数是强制服役对收入的影响,因此接受服役的人群中依从工具变量者只占少数这一事实并不是该研究的缺点。即使在越战时期,大部分士兵都是志愿参军的,在这一点我们要感谢越战老兵当时的觉悟。用局部平均处理效应去解释工具变量估计值还强调了如下事实:如果我们想估计志愿参军者的服役生涯对其后来收入的影响,那么需要其他的识别策略(Angrist(1998)在这个方向上作出了一些研究)。

表4.6中剩下的几行分别报告了用双胞胎和性别组成做工具变量时依从工具变量者的规模,以及用出生季度和义务教育法的改变做工具变量时依从工具变量者的规模。其中用双胞胎和性别组成做工具变量来自于Angrist和Evans(1998),他们研究了生育的影响;用出生季度和义务教育法的改变做工具变量来自于Angrist和Krueger(1991)以及Acemoglu和Angrist(2000),他们用这两个工具变量估计了教育的经济回报。在上面提到的每项研究中,依从工具变量者只占被处理集合中很小的一部分。比如,在所有的高中毕业生中,大约只有2%的人是因为出生季度或者义务教育法的改变而完成了高中教育。

由于依从工具变量者只占很小的一部分,因此我们是不是应该担心估计结果呢?对这个问题的回答依具体环境的不同而不同。在一些例子中,看上去我们可以放心地说:"你得到了你想要的。"举个例子,很多政策干预就属于这种情况,因为我们只对边际群体感兴趣,在McClellan, McNeil和Newhouse(1994)使用工具变量研究外科手术对心脏病人影响的奠基性论文中,他们也强调了这个观点。在他们的研究中,作者使用与心脏治疗机构的相对距离来构造工具变量,用来识别老年心脏病患者是不是会被施以外科手术。在他们的研究中,大部分病人得到的治疗都是相同的,但是对于某些病人,适宜的治疗方式(或者说我们已有的知识告诉我们该方法是合适的)是什么并不明确。在这种情况下,只有当装备良好的医疗机构比较近时,提供医疗服务的机构或者病人才会倾向于选择更加不保守的治疗方式。但是McClellan等人发现对这些处在边际上的群体而言,外科手术几乎没有用。相同的,义务教育法将辍学年龄提高到18岁显然与大部分美国高中生无关,但是

表 4.6　在工具变量研究中响应工具变量的概率

来源	内生变量	工具变量	样本	P[D=1]	第一阶段 P[$D_1 > D_0$]	P[Z=1]	依从概率 P[$D_1 > D_0$ | D=1]	依从概率 P[$D_1 > D_0$ | D=0]
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Angrist (1990)	服役状态	随机抽取的参军资格	生于 1950 年的白人男性	0.267	0.159	0.534	0.318	0.101
			生于 1950 年的非白人男性	0.163	0.060	0.534	0.197	0.033
Angrist 和 Evans(1998)	超过两个孩子	第二胎中的双胞胎	年龄在 21—35 岁的已婚女性，在 1980 年生育两个以上的孩子	0.381	0.603	0.008	0.013	0.966
		前两个孩子性别相同		0.381	0.060	0.506	0.080	0.048
Angrist 和 Krueger (1991)	高中毕业	在第三和第四季度出生	生于 1930 年到 1939 年的男性	0.770	0.016	0.509	0.011	0.034
Acemoglu 和 Angrist (2000)	高中毕业	州要求接受至少 11 年的学校教育	年龄在 40—49 岁的男性	0.617	0.037	0.300	0.018	0.068

注：本表针对一系列工具变量计算了依从工具变量者的相对规模和绝对规模。报告在第 6 列的第一阶段估计值告诉我们的是依从工具变量者的绝对规模。第 8 列和第 9 列则分别告诉我们相对于有越战服役经历的人和没有越战服役经历的人，依从工具变量者的相对规模。

会影响那些本来要选择辍学的人。工具变量估计指出对这些处在边际上的群体而言，接受教育的经济回报是相当可观的。

表 4.6 最后一行报告了在 4.4.2 节提到的双胞胎工具变量的特别之处。如前所述，令 $D_i = 0$ 表示在生育过两个孩子的女性组成的集合中只有两个孩子的那些女性，令 $D_i = 1$ 表示生育孩子多于两个的女性。由于在双胞胎生育中不存在从不接受者——所有在第二胎时生育双胞胎的女性最终至少会有三个孩子——因此在 $D_i = 0$ 的个体中依从工具变量的概率应该是 1(表中指出是 0.97)。因此在这个例子中局部平均处理效应就是 $E[Y_{1i} - Y_{0i} \mid D_i = 0]$，也就是对于没有生育第三个孩子的女性，如果她们选择生育第三个孩子，将会对收入带来的影响。

与计算依从工具变量者的规模不同，依从工具变量者的个体特征似乎是很难计算的。因为我们无法在每个个体上同时看到 $D_i = 0$ 和 $D_i = 1$，所以我们无法将满足 $D_{1i} > D_{0i}$ 的个体罗列出来并计算这些人的个体特征。不过，尽管我们无法把依从工具变量者一一列举，但是刻画这些人的个体特征的分布还是容易的。简单起见，我们关注了依从工具变量者的种族或者学位完成水平之类的个体特征，因为可以用虚拟变量来刻画这些特征。在这种情况下，我们通过对协变量做第一阶段回归，就可得到我们想要的一切。

令 x_{1i} 是一个遵循伯努里分布的个体特征，比如它是表征大学毕业的虚拟变量。我们的问题是：设现在考虑用性别组成做工具变量，对于那些依从工具变量的

女性而言，生育两个孩子的女性和生育两个以上孩子的女性，她们在教育水平方面会不会有不同？我们可以运用下面的计算方法来回答这个问题：

$$\frac{\mathrm{P}[x_{1i}=1 \mid D_{1i}>D_{0i}]}{\mathrm{P}[x_{1i}=1]}=\frac{\mathrm{P}[D_{1i}>D_{0i} \mid x_{1i}=1]}{\mathrm{P}[D_{1i}>D_{0i}]}$$
$$=\frac{\mathrm{E}[D_i \mid Z_i=1,\ x_{1i}=1]-\mathrm{E}[D_i \mid Z_i=0,\ x_{1i}=1]}{\mathrm{E}[D_i \mid Z_i=1]-\mathrm{E}[D_i \mid Z_i=0]} \tag{4.4.8}$$

换言之，依从工具变量者是大学生的相对可能性等于针对大学生群体做的第一阶段回归结果与总体的回归结果之比①。

在表 4.7 报告了以双胞胎和性别组成为工具变量时，针对第一胎生育年龄、非白人以及学位完成水平计算的依从工具变量者的个体特征。该表的构造基于 Angrist和 Evans(1998)，使用的样本是 1980 年按照 5%进行抽样的人口普查中年龄在 21—35 岁之间，有两个及以上孩子的已婚女性。在依从双胞胎工具变量的个体中，年龄超过 30 岁的女性比例远高于样本中女性的平均年龄，这意味着如果女性在较年轻的时候生育双胞胎，那么她们更倾向于再要一个孩子(虽然在 Angrist-Evans 的样本里，30 岁之后生育第一胎的女性十分少见)。依从双胞胎工具变量的

表 4.7　将双胞胎和性别组成作为工具变量时，依从该工具变量的个体所占的比例

		第二胎是双胞胎		前两个孩子性别相同	
变　量	$\mathrm{P}[x_{1i}=1]$ (1)	$\mathrm{P}[x_{1i}=1 \mid D_{1i}>D_{0i}]$ (2)	$\mathrm{P}[x_{1i}=1 \mid D_{1i}>D_{0i}]/\mathrm{P}[x_{1i}=1]$ (3)	$\mathrm{P}[x_{1i}=1 \mid D_{1i}>D_{0i}]$ (4)	$\mathrm{P}[x_{1i}=1 \mid D_{1i}>D_{0i}]/\mathrm{P}[x_{1i}=1]$ (5)
第一胎生育年龄大于 30 岁	0.002 9	0.004	1.39	0.002 3	0.995
黑人或者西班牙裔	0.125	0.103	0.822	0.102	0.814
高中毕业	0.822	0.861	1.048	0.815	0.998
大学毕业	0.132	0.151	1.14	0.090 4	0.704

注：本表报告了用双胞胎和性别组成做工具变量，对依从工具变量者的个体特征进行的分析。报告在第 3 列和第 5 列的比例给出的是拥有特定个体特征的依从工具变量者的相对概率，这里考虑的特定个体特征是表中最左边所显示的那些个体特征。这里用到的数据来自 1980 年按照 5%的比例进行抽样的人口普查数据，样本中包括了年龄在 21—35 岁之间的已婚母亲，Angrist 和 Evans(1998)中也是用了这个数据集。对所有列而言，使用的样本规模都是 254 654。

① 运用 Abadie(2003)构造的 Kappa 加权法，针对依从工具变量者个体特点的均值或者其他分布特征，我们有一个一般化的分析方法。比如：

$$\mathrm{E}[X_i \mid D_{1i}>D_{0i}]=\frac{\mathrm{E}[\kappa_i X_i]}{\kappa_i}$$

其中，

$$\kappa_i=1-\frac{D_i(1-Z_i)}{1-\mathrm{P}(Z_i=1 \mid X_i)}-\frac{(1-D_i)Z_i}{\mathrm{P}(Z_i=1 \mid X_i)}$$

上面的这个公式是有用的，因为加权函数 κ_i 的意义就在于它“找到了依从够工具变量者”，对它的讨论请见第 4.5.2 节。

个体的教育水平要高于平均水平，但是依从性别组成工具变量的个体接受的教育水平却不是很高。这有助于解释在使用双胞胎做工具变量时 2SLS 估计值（该估计值报告于表 4.4 中）较小的原因，因为在 Angrist 和 Evans（1998）中他们指出随着教育水平的提高，生育孩子对母亲劳动力供给的影响呈下降趋势。

4.5 对局部平均处理效应的推广

局部平均处理效应定理适用于无协变量，只用单个虚拟变量做工具变量去估计只存在单一处理（也就是 $D_i \in \{0, 1\}$）时的因果效应，它是进行因果推断的最基本模型。我们可以在三个方向上对该定理进行重要的推广：(1)存在多个工具变量的情况（比如用一组虚拟变量来表示出生季度）；(2)模型中存在协变量的情况（比如存在用以控制出生年份的控制变量）；(3)因果效应取多个值甚至是连续值的情况（比如接受一年教育和接受两年教育对收入的影响不同）。在所有这三类推广中，工具变量估计值都是对因果效应的加权平均，这里被加权平均的因果效应都是针对与特定工具变量相联系的依从工具变量者计算出的因果效应。在这三类推广中，用来计算的计量经济学工具都是 2SLS，对估计结果的解释也基本保持原样，只是有一些小的变化。具体而言，当存在多个工具变量时，每使用一个工具变量就能得到一个工具变量估计值，2SLS 估计值就是将这些工具变量估计值平均后得到的因果效应；当存在协变量时，每个协变量都和一个特定的局部平均处理效应相联系，2SLS 估计值则是这些局部平均处理效应的平均值；当因果效应取多个值甚至是连续值时，我们针对每个非线性的因果响应函数（这个函数的定义请见第 4.5.3 节）估计边际值，2SLS 就是对这些边际值的加权平均。在这三个方向上的推广使我们得到的结论更贴近于计量经济学实践所面临的情况，同时为 2SLS 提供了一个简单的因果解释。

4.5.1 多工具变量下的局部平均处理效应

局部平均处理效应定理在多工具变量下的推广是显而易见的。从本质上讲，这种推广与我们在分组数据中得到的结论相同。考虑一对以虚拟变量的身份出现的工具变量 Z_{1i} 和 Z_{2i}。不失一般性，假设这两个工具变量互斥（若否，我们可以构造一组互斥的工具变量 $Z_{1i}(1-Z_{2i})$、$Z_{2i}(1-Z_{1i})$ 和 $Z_{1i}Z_{2i}$）。那么用这两个虚拟变量可以构造出两个瓦尔德估计值。不失一般性，假设每个工具变量都满足单调性假设，得到的第一阶段估计值都是正的（若否，我们通过将虚拟变量取值互换，也即构造 $Z_{1i}^{*} = 1-Z_{1i}$、$Z_{2i}^{*} = 1-Z_{2i}$ 作为新的工具变量）。于是每个工具变量都估计出一个 $E[Y_{1i} - Y_{0i} \mid D_{1i} > D_{0i}]$，尽管满足 $D_{1i} > D_{0i}$ 的个体可能不同于在随机分配中获得处理的人，其中随机分配的处理状态由 Z_{1i} 和 Z_{2i} 表示。

除了使用瓦尔德估计值，我们还可以在 2SLS 中同时使用 Z_{1i} 和 Z_{2i}。既然这两

个工具变量和常数项包含了工具变量集合所有的信息，那么 2SLS 的估计过程就应该和分组数据估计过程相同，这里我们使用给定 Z_{1i} 和 Z_{2i} 后的条件均值来定义分组数据估计中用到的条件均值。正如 Angrist(1991)所指，由此得到的分组数据估计值是相应瓦尔德估计值的线性组合。换言之，这是与特定工具变量相联系的局部平均处理效应的线性组合，其中与特定工具变量相联系的局部平均处理效应是指使用某个工具变量后得到的那个局部平均处理效应(事实上，它就是在传统的线性常因果效应条件同方差模型中的有效线性组合)。

上面的这些讨论还不完整，因为我们尚未指明通过 2SLS 得到的局部平均处理效应的线性组合是一种加权平均值(比如，加权平均只要求权重非负且相加等于 1，这一点我们就没有说明)。Imbens 和 Angrist(1994)以及 Angrist 和 Imbens (1995)给出了我们想要求解的权重函数。在一般情况下的权重函数有点复杂，所以我们在这里以两个工具变量为例给出权重函数的一个简单形式。在这个例子中我们已经可以说明：同时使用工具变量 Z_{1i} 和 Z_{2i} 得到的 2SLS 估计值是对分别使用 Z_{1i} 和 Z_{2i} 得到的工具变量估计值的加权平均。令：

$$\rho_j = \frac{\operatorname{cov}(Y_i,\ Z_{ji})}{\operatorname{cov}(D_i,\ Z_{ji})},\ j = 1,\ 2$$

表示分别使用工具变量 Z_{1i} 和 Z_{2i} 后得到的两个工具变量估计值。

如果同时使用工具变量 Z_{1i} 和 Z_{2i}，我们在 2SLS 第一阶段得到的总体拟合值就是 $\hat{D}_i = \pi_{11}Z_{1i} + \pi_{12}Z_{2i}$，其中 π_{11} 和 π_{12} 都是正数。根据我们将 2SLS 估计值解释为工具变量估计值理由，2SLS 估计值就是：

$$\begin{aligned}
\rho_{2SLS} &= \frac{\operatorname{cov}(Y_i,\ \hat{D}_i)}{\operatorname{cov}(D_i,\ \hat{D}_i)} = \frac{\pi_{11}\operatorname{cov}(Y_i,\ Z_{1i})}{\operatorname{cov}(D_i,\ \hat{D}_i)} + \frac{\pi_{12}\operatorname{cov}(Y_i,\ Z_{2i})}{\operatorname{cov}(D_i,\ \hat{D}_i)} \\
&= \left[\frac{\pi_{11}\operatorname{cov}(D_i,\ Z_{1i})}{\operatorname{cov}(D_i,\ \hat{D}_i)}\right]\left[\frac{\operatorname{cov}(Y_i,\ Z_{1i})}{\operatorname{cov}(D_i,\ Z_{1i})}\right] \\
&\quad + \left[\frac{\pi_{12}\operatorname{cov}(D_i,\ Z_{2i})}{\operatorname{cov}(D_i,\ \hat{D}_i)}\right]\left[\frac{\operatorname{cov}(Y_i,\ Z_{2i})}{\operatorname{cov}(D_i,\ Z_{2i})}\right] \\
&= \psi\rho_1 + (1-\psi)\rho_2
\end{aligned}$$

其中，ρ_1[①]是使用工具变量 Z_{1i} 得到的局部平均处理效应，ρ_2 是使用工具变量 Z_{2i} 得到的局部平均处理效应，并且

$$\psi = \frac{\pi_{11}\operatorname{cov}(D_i,\ Z_{1i})}{\pi_{11}\operatorname{cov}(D_i,\ Z_{1i}) + \pi_{12}\operatorname{cov}(D_i,\ Z_{2i})}$$

是介于 0 和 1 之间的一个数字，它的大小依赖于第一阶段中每个工具变量的相对重要性。因此我们可以指出 2SLS 估计值是个加权平均值。每个工具变量估计值都等于它所对应的那个依从工具变量者的因果效应，然后 2SLS 对其进行加权平

① 原文在此处是 ψ_1 而不是 ρ，疑为作者笔误。——译者注

均。比如，设 Z_{1i} 表示双胞胎工具变量，Z_{2i} 表示性别组成工具变量，这两个工具变量都是针对家庭规模的，来自于 Angrist 和 Evans(1998)。在该项研究中，作者计算出第二胎生双胞胎可将女性生育第三个孩子的概率提高 0.6，而前两个孩子性别相同则将女性生育第三个孩子的概率提高 0.07。当同时使用这两个工具变量时，得到的 2SLS 估计值是单独使用两个工具变量得到的瓦尔度估计值的加权平均①。

4.5.2 存在协变量的异质性因果模型

行文至此，我们可能在想之前讨论过的协变量去哪里了？毕竟，在我们对回归和匹配的讨论中协变量扮演了主要角色，但是局部平均处理效应定理中却并未包含协变量。这是因为当我们将工具变量看作某种(自然导致或者人为的)随机实验时，协变量不会发挥作用——如果工具变量是随机分配的，那么它应该和协变量相独立。但是并非所有的工具变量都有这种性质。与我们在上一章对协变量的讨论类似，将其纳入使用工具变量进行因果分析的模型中，其主要原因为控制了协变量后条件独立性和排他性约束会更可能成立。即使像参军资格这样通过随机分配得到的工具变量，在控制了协变量后该工具变量依赖的假设会更有可能成立。在这个例子中，出生较早的人更可能获得参军资格，因为决定这些人能否获得参军资格的截断值更高。还因为不同出生年份(年龄)出生的人，其收入会有区别，所以控制了出生年份后工具变量才更有可能发挥作用。

正式地，存在协变量时工具变量估计需要的假设变为条件独立假设：

$$\{Y_{1i}, Y_{0i}, D_{1i}, D_{1i}\} \perp\!\!\!\perp Z_i \mid X_i \tag{4.5.1}$$

换言之，给定协变量 X_i，工具变量应该“像随机分配的那样好”(这里我们隐含地保留了排他性约束)。将协变量纳入模型的第二个理由是它可以降低被解释变量的变化。这可以带来更加精确的 2SLS 估计值。

存在协变量的常因果效应模型要求总体方程的函数形式为：

$$\mathrm{E}[Y_{0i} \mid X_i] = X_i'\alpha^*，\text{其中 } \alpha^* \text{ 是个 } k \times 1 \text{ 的参数向量}$$

$$Y_{1i} - Y_{0i} = \rho$$

将其与假设(4.5.1)结合起来，我们就可以得到在 4.1 节讨论过的 2SLS 估计方程(4.1.6)。

对常因果效应模型的直接推广允许我们将因果效应记为：

---

① 单独使用双胞胎工具变量得到的妇女生育第三胎对劳动力供给的影响是−0.084。单独使用性别组成工具变量得到的妇女生育第三胎对劳动力供给的影响是−0.138。同时使用两个工具变量得到的 2SLS 估计值是−0.098。在这个 2SLS 估计中，对双胞胎工具变量对应的估计结果赋予的权重是 0.74，对相同性别工具变量对应的估计结果赋予的权重是 0.26，这是因为在第一阶段中双胞胎工具变量表现得更强。

$$Y_{1i} - Y_{0i} = \rho(X_i)$$

其中，$\rho(X_i)$是关于协变量 X_i 的确定型函数。通过将 Z_i 和 X_i 的交互项加入第一阶段回归，将 D_i 和 X_i 的交互项加入第二阶段回归，我们就可以开始估计了。由于将 Z_i 和 X_i 的交互项加入第一阶段的结果是我们的内生变量增加了，所以第一阶段回归变为方程组可以写为：

$$D_i = X_i'\pi_{00} + \pi_{01}Z_i + Z_iX_i'\pi_{02} + \xi_{0i} \tag{4.5.2a}$$

$$D_iX_i = X_i'\pi_{10} + \pi_{11}Z_i + Z_iX_i'\pi_{12} + \xi_{1i} \tag{4.5.2b}$$

虽然等式(4.5.2b)中的 D_iX_i 看上去像是个常数，但是 D_iX_i 中的每个元素都应该对应于一个第一阶段方程。在本例中，第二阶段方程是：

$$Y_i = \alpha'X_i + \rho_0D_i + D_iX_i'\rho_1 + \eta_i$$

于是可得 $\rho(X_i) = \rho_0 + \rho_1'X_i$。可以选择在关于 X_i 饱和的子集中用 2SLS 对 $\rho(X_i)$ 进行非参数估计。

我们还可在条件独立假设(4.5.1)下求解异质性因果效应模型的局部平均处理效应定理，只不过这时对定理的解释会变得复杂一些。对 X_i 的每个特定值，我们定义与该协变量取值相对应的局部平均处理效应为：

$$\lambda(X_i) \equiv \mathrm{E}[Y_{1i} - Y_{0i} \mid X_i,\ D_{1i} > D_{0i}]$$

下面这个定理(来自于 Angrist 和 Imbens，1995)用“饱和加权”的方法估计出了存在协变量的局部平均处理效应 $\lambda(X_i)$。

定理 4.5.1：饱和加权定理(Saturate and Weight)。设给定 X_i 后局部平均处理效应定理所要求的假设都满足。也就是说：

(假设 1，独立性) $[\{Y_i(D_{1i}, 1), Y_i(D_{0i}, 0), D_{1i}, D_{0i}\} \perp\!\!\!\perp Z_i \mid X_i]$；

(假设 2，排他性) $\mathrm{P}[Y_i(d, 0) = Y_i(d, 1) \mid X_i] = 1$，对于 $d = 0, 1$；

(假设 3，第一阶段) $\mathrm{E}[D_{1i} - D_{0i} \mid X_i] \neq 0$；

如前，还是要假设单调性(假设 4)仍成立。我们基于第一阶段方程：

$$D_i = \pi_X + \pi_{1X}Z_i + \xi_{1i} \tag{4.5.3}$$

以及第二阶段方程：

$$Y_i = \alpha_X + \rho_cD_i + \eta_i$$

来估计 2SLS 估计值。

其中，π_X 和 α_X 代表协变量的饱和模型(表示 X_i 所有可能取值的虚拟变量组成的集合)，π_{1X} 表示给定 X_i 的每个取值，Z_i 在第一阶段的影响。然后 $\rho_c = \mathrm{E}[\omega(X_i)\lambda(X_i)]$，其中：

$$\omega(X_i) = \frac{V\{\mathrm{E}[D_i \mid X_i, Z_i] \mid X_i\}}{\mathrm{E}[V\{\mathrm{E}[D_i \mid X_i, Z_i] \mid X_i\}]} \tag{4.5.4}$$

并且

$$V\{\mathrm{E}[D_i \mid X_i, Z_i] \mid X_i\} = \mathrm{E}\{\mathrm{E}[D_i \mid X_i, Z_i](\mathrm{E}[D_i \mid X_i, Z_i] - \mathrm{E}[D_i \mid X_i]) \mid X_i\}$$

这个定理指出在第一阶段使用完全饱和模型，在第二阶段使用饱和模型得到的2SLS估计值就是对每个协变量取值下估计出的局部平均处理效应的加权平均。在 X_i 可以取到的每个值上，权重与第一阶段拟合值 $\mathrm{E}[D_i \mid X_i, Z_i]$ 的条件方差均值成正比[①]。得到这个定理基于如下事实：当方程(4.5.3)饱和时(也就是说第一阶段回归确实可以构造出条件期望函数)，那么第一阶段方程等于 $\mathrm{E}[D_i \mid X_i, Z_i]$。

在实际中，我们可能不想在第一阶段对所有 X_i 的可能取值都估计出一个系数。首先，这样做可能带来偏误，我们在本章最后再来讨论这一问题。其次，大量不精确的第一阶段拟合值看上去也不好。也许我们在模型中少加入一些参数，比如假设第一阶段中的 π_{1X} 为常数，也可以在平均意义上逼近局部平均处理效应。这个观点实际上是对的，但是对该观点进行的论证则比较复杂。Abadie(2003)指出一种2SLS可以为相应的因果关系给出一个最小均方误意义下的逼近。

在Abadie的研究中，他首先定义感兴趣的估计目标为 $\mathrm{E}[Y_i \mid D_i, X_i, D_{1i} > D_{0i}]$，这个等式表示给定处理状态和协变量下，关于依从工具变量者的 Y_i 的条件期望函数。该条件期望函数的重要特点在于如果给定 X_i 后局部平均处理效应定理所要求的假设成立，那么就可对该等式赋予一个因果解释。换言之，对于依从工具变量者，给定 X_i 后处理组和控制组之间的平均差别等于给定 X_i 后的局部平均处理效应：

$$\begin{aligned}
&\mathrm{E}[Y_i \mid D_i = 1, X_i, D_{1i} > D_{0i}] - \mathrm{E}[Y_i \mid D_i = 0, X_i, D_{1i} > D_{0i}] \\
&\quad = \mathrm{E}[Y_{1i} - Y_{0i} \mid X_i, D_{1i} > D_{0i}] \\
&\quad = \lambda(X_i)
\end{aligned}$$

得到这个结果的原因是：对于依从工具变量者有 $D_i = Z_i$，由假设(4.5.1)知给定 X_i 和 $D_{1i} > D_{0i}$，潜在结果与 Z_i 独立。这样做的关键意义是说明了：在依从工具变量的个体中对 Y_i 用 D_i 和 Z_i 做回归，我们可以对由此得到的结果做一个因果解释。虽然我们无法从这个回归中得到感兴趣的条件期望函数(除非这个条件期望函数是线性的，或者说模型是饱和的)，但是它始终可以为我们提供一个在最小均方误意义下的逼近。换言之，在依从工具变量的个体中对 Y_i 关于 D_i 和 Z_i 做回归可以近似 $\mathrm{E}[Y_i \mid D_i, X_i, D_{1i} > D_{0i}]$，这种处理类似于用最小二乘估计去近似 $\mathrm{E}[Y_i \mid D_i, X_i]$。可是在此过程中又遇到问题：我们不知道依从工具变量的个体是哪些，因此无法用样本值将其估计出来。按照下面的这个定理，我们可以解决该问题，从而找到依从工具变量的个体：

---

① 注意到给定 X_i 后 $\mathrm{E}[D_i \mid X_i, Z_i]$ 的变化完全来自于 Z_i 的变化。因此在工具变量变化带来拟合值更大变化的那些协变量上，加权方程赋予的权重会更大。

定理 4.5.2 阿坝蝶·卡帕定理(Abadie Kappa)。考虑给定 X_i 后局部平均因果效应定理所要求的假设都满足。记 $g(Y_i, D_i, X_i)$ 为定义在 (Y_i, D_i, X_i) 上的任意可测函数,其期望值有限。定义:

$$\kappa_i = 1 - \frac{D_i(1-Z_i)}{1-\mathrm{P}(Z_i=1 \mid X_i)} - \frac{(1-D_i)Z_i}{\mathrm{P}(Z_i=1 \mid X_i)}$$

那么:

$$\mathrm{E}[g(Y_i, D_i, X_i) \mid D_{1i} > D_{0i}] = \frac{\mathrm{E}[\kappa_i g(Y_i, D_i, X_i)]}{\mathrm{E}[\kappa_i]}$$

运用下述事实可对该定理进行直接的证明:设局部平均处理效应定理需满足的假设都满足,且我们考虑的任何均值都是对依从工具变量者、从不接受者和始终接受者所对应的均值的加权平均。由单调性,那些满足 $D_i(1-Z_i)=1$ 的人是始终接受者,因为他们有 $D_{0i}=1$,那些满足 $(1-D_i)Z_i=1$ 的人则是从不接受者,因为他们有 $D_{1i}=0$。因此,依从工具变量者就是除去始终接受者和从不接受者之后剩下的那部分人。

阿坝蝶·卡帕定理有很多重要的应用:比如它会出现在分位数因果效应中。在本节的讨论里,我们在线性回归中用它来逼近 $\mathrm{E}[Y_i \mid D_i, X_i, D_{1i} > D_{0i}]$。也就是说,令 α_c 和 β_c 是下面问题的解:

$$(\alpha_c, \beta_c) = \arg\min_{a,b} \mathrm{E}\{(\mathrm{E}[Y_i \mid D_i, X_i, D_{1i} > D_{0i}] - aD_i - X_i'b)^2 \mid D_{1i} > D_{0i}\}$$

换言之,$\alpha_c D_i + X_i'\beta_c$ 给出了对 $\mathrm{E}[Y_i \mid D_i, X_i, D_{1i} > D_{0i}]$ 的最小均方误近似,如果 $\mathrm{E}[Y_i \mid D_i, X_i, D_{1i} > D_{0i}]$ 是线性的,那么 $\alpha_c D_i + X_i'\beta_c$ 对 $\mathrm{E}[Y_i \mid D_i, X_i, D_{1i} > D_{0i}]$ 的近似是精确的。运用阿坝蝶·卡帕定理指出可以通过求解等式(4.5.5)来逼近函数 $\mathrm{E}[Y_i \mid D_i, X_i, D_{1i} > D_{0i}]$:

$$(\alpha_c, \beta_c) = \arg\min_{a,b} \mathrm{E}\{\kappa_i (Y_i - aD_i - X_i'b)^2\} \tag{4.5.5}$$

其中,$\mathrm{E}\{\kappa_i(Y_i - aD_i - X_i'b)^2\}$ 是相应最小二乘的最小化元[①],这里阿坝蝶·卡帕函数被用来对普通最小二乘中的最小化元 $(Y_i - aD_i - X_i'b)^2$ 进行加权。

Abadie 提出了对等式(4.5.5)进行估计的方法(以及相应的分布理论)。在第一步估计时可使用 $\mathrm{P}(Z_i=1 \mid X_i)$ 的参数或半参数估计结果来构造 κ_i,然后将第一步估计值代入等式(4.5.5)进行估计。不出意料,当唯一的协变量是常数时,Abadie 过程简化为瓦尔德估计值。但出人意料的是如果在构造 κ_i 时用线性模型来估计 $\mathrm{P}(Z_i=1 \mid X_i)$,那么最小化等式(4.5.5)的过程就是传统的 2SLS。换言之,如

① 用来逼近 $\mathrm{P}(Z_i=1 \mid X_i)$ 的函数无需是线性的。除了 $\alpha_c D_i + X_i'\beta_c$ 之外,我们还可使用指数函数(如果被解释变量是非负的)或者 probit 模型(如果被解释变量是 0 或者 1)等非线性函数,在本章最后一部分我们会再次回到这个问题。正如在 4.4.4 节提到的,可用 Kappa 加权函数来考察依从工具变量者的个体特征分布,这里个体特征是指用协变量表示的那些特征,同时我们还可用 Kappa 加权函数来估计被解释变量的分布。

果在构造对 κ_i 时使用 $P(Z_i = 1 \mid X_i) = X_i'\pi$，那么 Abadie 估计值就是 2SLS 估计值。因此我们可以总结：一旦可用线性模型对 $P(Z_i = 1 \mid X_i)$ 进行拟合或逼近，我们都可以将 2SLS 看作是对依从工具变量者的因果响应函数 $E[Y_i \mid D_i, X_i, D_{1i} > D_{0i}]$ 进行的逼近。但从另一方面讲，一般而言 α_c 不是 2SLS 估计值，β_c 也不是在 2SLS 中得到的协变量系数。不过 $P(Z_i = 1 \mid X_i)$ 为线性函数时 Abadie 过程和 2SLS 之间的等价性让我们知道：在绝大部分的应用中 Abadie 方法和 2SLS 会产生相似的结果。

举个例子，Angrist(2001)基于等式(4.5.5)对 Angrist 和 Evans(1998)中的结果进行了重新分析，发现与 2SLS 估计值没有太大差别。使用双胞胎工具变量估计生育第三胎对女性劳动力供给的影响，得到的 2SLS 估计结果是 −0.088，使用 Abadie 方法估计出的结果是 −0.089。类似的，当考虑生育第三胎对女性周工作时间的影响时，2SLS 和 Abadie 方法估计值相同，都是 −3.55。举这个例子不是想反对 Abadie 过程，而是想说明 2SLS 估计值确实可以近似我们感兴趣的因果关系①。

4.5.3 存在多种处理强度时的平均因果响应*

取值为{0，1}的虚拟变量带来的因果效应与取值为{0，1，2，…}的因果变量带来的因果效应之间有很大不同，这种不同在于对所有人而言，前者只产生一种因果效应，而后者却可能产生多种因果效应：比如因果变量取值从 0 变成 1 带来的因果效应，因果变量取值从 1 变到 2 带来的因果效应等。在研究教育带来的经济回报问题时我们注意到了这个问题，因此采取下面的记号，令：

$$Y_i \equiv f_i(s)$$

表示个体 i 在接受 s 年的教育后应该获得的收入。注意此函数中 f 带有下标 i，但是 s 却没有。函数 $f_i(s)$ 告诉我们在任意的教育水平 s 下个体 i 可能获得的收入，而不是指我们观察到 s 在现实中的真实取值 s_i 后个体 i 获得的收入。换言之，$f_i(s)$ 告诉我们的是“如果 s 取某个值，那么个体 i 会……”的这样一个因果问题。

假设 s_i 可以在集合{0，1，…，$\bar{s}$}中取值。那么就存在$\bar{s}$个因果效应，也就是说有$\bar{s}$个 $Y_{si} - Y_{s-1,i}$。线性因果模型假设对 s 的所有可能取值、对于所有人，$Y_{si} -$

① 在常见的线性或非线性回归软件中就可以计算 Abadie 估计值。但是巧妙之处在于构造一个权重都为正数的加权函数。这可以通过对等式(4.5.5)进行重复求期望而得到，于是我们要求解的 κ_i（对于从不接受者和始终接受者，这个值是负的）就可用下面的等式来计算：

$$E[\kappa_i \mid X_i, D_i, Y_i] = 1 - \frac{D_i(1 - E[Z_i \mid X_i, D_i, Y_i])}{1 - P(Z_i = 1 \mid X_i)} - \frac{(1 - D_i)E[Z_i \mid X_i, D_i, Y_i]}{P(Z_i = 1 \mid X_i)}$$

（也可在第 7.2.1 节的讨论中见到）。Abadie(2003)给出了计算标准误的公式，Alberto Abadie 将计算标准误和相应参数的程序公布。也可以用 Bootstrap 的方法求解 Abadie 估计值的标准误。

$Y_{s-1,i}$ 表示的因果效应都相同，但很显然这个假设的限制性很强。不过我们可以不用这么死板地对待该假设。因为 2SLS 提供了对单位因果效应(unit causal effect)计算加权平均的工具，而且我们还可以估计和研究相应的加权函数，以此了解特定工具变量所对应的行为来自哪里。对加权函数的研究可以告诉我们依从特定工具变量的个体在 s_i 的取值区间上是如何分布的。比如在用出生季度或义务教育法做工具变量对教育的经济回报进行的研究中，加权函数告诉我们估计出的教育回报来自：在所有高中学生中，能够完成高中教育的那些人的分布的移动。类似于 Card(1995)那样使用距离作为工具变量的研究则从另外的角度研究了不同教育水平在人群中的分布，因而捕捉到的是另一种教育的经济回报。

为了使这部分讨论更加丰满，假设有一个工具变量 Z_i，用以表征出生所在州是否有严格执行的义务教育法，并用它来估计教育的经济回报(正如 Acemoglu 和 Angrist(2000)的研究)。同时，记 s_{1i} 表示 $Z_i=1$ 时个体 i 接受的教育水平，记 s_{0i} 表示 $Z_i=0$ 时个体 i 接受的教育水平。下面这个定理来自于 Angrist 和 Imbens (1995)，它告诉我们在不同处理强度带来不同因果效应的情况下如何对瓦尔德估计值进行解释。注意到假设潜在结果 s 与工具变量独立，那么独立性假设和排他性约束都可以得到满足。

定理 4.5.3：平均因果响应定理(Average Causal Response)。假设：

(ACR1，独立性假设和排除性约束) $\{Y_{0i}, Y_{1i}, \cdots, Y_{si}; s_{0i}, s_{1i}\} \perp\!\!\!\perp Z_i$；

(ACR2，第一阶段) $\mathrm{E}[s_{1i}-s_{0i}] \neq 0$；

(ACR3，单调性) $s_{1i}-s_{0i} \geqslant 0 \ \forall i$，或者反之，

那么

$$\frac{\mathrm{E}[Y_i \mid Z_i=1]-\mathrm{E}[Y_i \mid Z_i=0]}{\mathrm{E}[s_i \mid Z_i=1]-\mathrm{E}[s_i \mid Z_i=0]} = \sum_{s=1}^{T} \omega_s \mathrm{E}[Y_{si}-Y_{s-1,i} \mid s_{1i} \geqslant s > s_{0i}]$$

其中

$$\omega_s = \frac{\mathrm{P}[s_{1i} \geqslant s > s_{0i}]}{\sum_{j=1}^{T} \mathrm{P}[s_{1i} \geqslant j > s_{0i}]}$$

这个权重 ω_s 是非负而且相加为 1 的。

平均因果响应(Average Causal Response，简称 ACR)定理告诉我们当不同处理强度带来不同因果效应时，瓦尔德估计值是对单位因果响应的加权平均。其中单位因果响应 $\mathrm{E}[Y_{si}-Y_{s-1,i} \mid s_{1i} \geqslant s > s_{0i}]$ 是指在点 s 处依从工具变量的个体在潜在结果上的平均差异。这里，在点 s 处依从工具变量的个体指的是在工具变量影响下个体的选择从小于 s 变到大于等于 s 的那些人。比如在 Angrist 和 Krueger (1991)使用出生季度作为工具变量时，样本中的一些个体本来只想读到 11 年级，由于该工具变量的作用改变了这些个体的决策，他们决定完成 12 年级甚至是更高

级别的教育，或者在样本中原本只想读完 10 年级的人由于工具变量的影响而改变决策，选择完成 11 年级或者更高级别的教育。使用出生季度作为工具变量的瓦尔德估计值可以将这些不同处理状态下的不同因果效应加权平均成一个平均因果响应。

点 s 处依从工具变量者所成的集合规模应该是 $\mathrm{P}[s_{1i} \geqslant s > s_{0i}]$。由单调性，这个数字应该是非负的，并由 s_i 的累积分布函数在 s 处的差值给出，为了看清楚这一点，注意到：

$$\begin{aligned}\mathrm{P}[s_{1i} \geqslant s > s_{0i}] &= \mathrm{P}[s_{1i} \geqslant s] - \mathrm{P}[s_{0i} \geqslant s] \\ &= \mathrm{P}[s_{0i} < s] - \mathrm{P}[s_{1i} < s]\end{aligned}$$

上面这个等式是非负的，因为单调性要求 $s_{1i} \geqslant s_{0i}$。更进一步，由独立性可得：

$$\mathrm{P}[s_{0i} < s] - \mathrm{P}[s_{1i} < s] = \mathrm{P}[s_i < s \mid Z_i = 0] - \mathrm{P}[s_i < s \mid Z_i = 1]$$

最后，注意到因为非负随机变量的均值等于对“1－累积分布函数”进行积分，于是我们有：

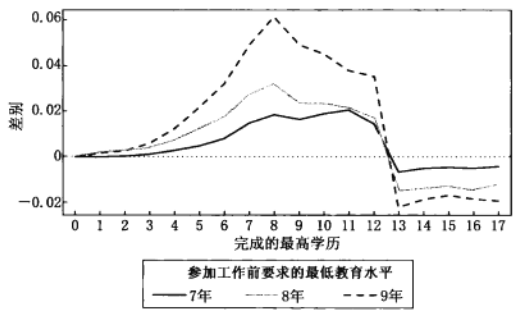
$$\begin{aligned}&\mathrm{E}[s_i \mid Z_i = 1] - \mathrm{E}[s_i \mid Z_i = 0] \\ &= \sum_{j=1}^{\bar{s}} (\mathrm{P}[s_i < j \mid Z_i = 0] - \mathrm{P}[s_i < j \mid Z_i = 1]) \\ &= \sum_{j=1}^{\bar{s}} \mathrm{P}[s_{1i} \geqslant j > s_{0i}]\end{aligned}$$

因此可见，利用工具变量等于 0 和等于 1 时内生变量的累积分布函数之间的差值，我们可以对加权函数进行具有一致性的估计。在第一阶段得到的加权函数是对估计出的加权函数进行正规化后得到的结果。

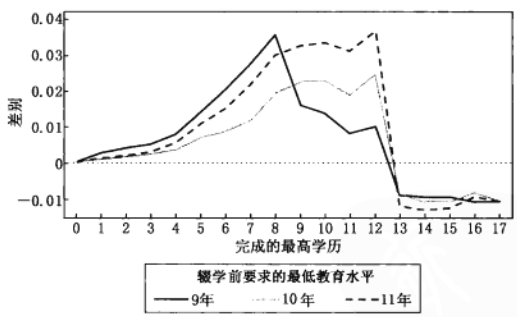
平均因果响应定理有助于我们进一步理解从 2SLS 中得到的估计值。比如用义务教育法和童工法做工具变量，我们可以在 6—12 年级的孩子中估计出受教育水平提高而带来的收入增加。但是这个估计值对我们考察小学教育对收入的影响则没有什么帮助。图 4.3 对这一因果效应进行了考察，该图来自 Angrist 和 Acemoglu(2000)。

图 4.3 中的 x 轴表示个体至少接受的教育水平，绘制出的曲线则表示由于工具变量的影响，x 轴上的数字所对应的个体在接受教育的可能性上的差别(也就是 1－累积概率函数)。在这里，作者对 1960 年、1970 年和 1980 年人口普查中年龄段在 40—49 岁之间的男性白人计算了接受教育水平的差异，这种差异是由童工法或者义务教育法的不同而造成的。作者使用了两个工具变量进行估计，一个是参加工作前要求的最低教育水平(A 图)，另一个是辍学前要求的最低教育水平(B 图)，并将法律最为宽松的那组个体的情况当做参照组。通过和参照组的情况作比较，每个工具变量(比如用以表示工作前是否要求必须接受至少 7 年义务教育的虚拟变量)都可用来构造一个瓦尔德估计值。

A.



B.



注：这个图显示出由工具变量所导致的接受教育的概率要比由 x 轴显示的实际教育水平高。用来参照的组别有两个，一个是图 A 中描绘的要求必须接受至少 6 年教育的情况，另一个是图 B，要求个体至少接受 8 年以上的教育。图 A 显示出的是由于童工法的加强带来的累积分布函数的不同。图 B 显示出的是由于加强义务教育法带来的累积分布函数的不同。

图 4.3　将义务教育作为工具变量得到的其对教育的影响

图 A 告诉我们面对最严格童工法的那些个体，接受 8—12 年级教育的概率会高 1—6 个百分点。这种变化的大小依赖于法律是否要求在工作之前接受 7 年、8 年或者 9 年的教育。但是在所有的例子中，在年级较低时，累积概率函数之间的差别会减少，并在 12 年级后急剧下降。图 B 显示了将义务教育法作为工具变量的结果，虽然产生的影响小一点而且变化发生在更高一点的年级，但是也出现了与图 A 显示的相同模式，出现这种情况是合理的，因为随着学生年级的提高，义务教育法产生的约束性要强于童工法的约束性。有趣的是，分别用童工法和义务教育法作工具变量得到的 2SLS 估计值很相似，这两个估计值大致在 0.08 到 0.10 之间。

本节讨论了如何对局部平均处理效应进行推广，在顺利结束这一节的讨论之前，我们需要注意的是这里讨论的内容往往都是结合起来使用的。比如，当存在多个工具变量且不同处理强度下的因果效应不同时，每个工具变量都可以得到一个平均因果效应(ACR)。类似的，我们也可以将饱和加权定理用于不同处理强度下因果效应不同的模型中(虽然我们并没有讨论存在这种情况时的 Abadie Kappa 加权函数)。最后一个重要的扩展考虑一下情形：我们感兴趣的因果变量是连续的，那么我们可以自然地将因果响应函数看作微分。

1. 再见，多谢你们的鱼[①]

还是像讨论教育水平时做的假设那样，我们用一个函数来描述不同因果变量带来的不同结果，假设感兴趣的因果变量可以取任何非负数且函数可微。比如需求曲线就满足上面的假设，需求量是价格的函数。具体而言，令 $q_i(p)$ 表示在市场 i 中价格为 p 时的需求量。类似于 $f_i(s)$，$q_i(p)$ 也是对潜在结果的表示，不同之处只在现在考虑的是一个时点或者地点上的需求曲线，不再是个体。比如，Angrist，Graddy 和 Imbens(2000)估计了纽约市富尔顿(Fulton)的鱼类批发市场上需求的价格弹性。这个需求曲线的斜率是 $q_i'(p)$；如果用自然对数来表示价格和数量，那么得到的那个斜率就是想要求解的弹性。

在 Angrist，Graddy 和 Imbens(2000)中，研究者们使用的工具变量来自长岛海岸线附近距大型商贸鱼类市场不远处的天气情况。当出现暴风雨天气时，渔民很难出海捕鱼，因此鱼类价格升高，供给减少。Angrist，Graddy 和 Imbens 使用虚拟变量 $stormy_i$ 来表示出现大风大浪的天气，以此作为虚拟变量来估计对鱼类的需求。其他的数据包括每天观察到的石首鱼批发价和数量，这种鱼是用来制作鱼饼和其他类似产品的一种较便宜的鱼类。

使用 $stormy_i$ 做工具变量得到的瓦尔德估计值可以用下面的方程来解释：

$$\frac{\mathrm{E}[q_i \mid stormy_i = 1] - \mathrm{E}[q_i \mid stormy_i = 0]}{\mathrm{E}[p_i \mid stormy_i = 1] - \mathrm{E}[p_i \mid stormy_i = 0]} = \frac{\int \mathrm{E}[q_i'(t) \mid P_{1i} \geqslant t > P_{0i}]\mathrm{P}[P_{1i} \geqslant t > P_{0i}]\mathrm{d}t}{\int \mathrm{P}[P_{1i} \geqslant t > P_{0i}]\mathrm{d}t} \tag{4.5.6}$$

其中，p_i 是在市场 i(或者说第 i 天)中的日观测价格，P_{1i} 和 P_{0i} 是用 $stormy_i$ 表示的潜在价格。可以用 $\mathrm{P}[P_{1i} \geqslant t > P_{0i}] = \mathrm{P}[p_i < t \mid stormy_i = 0] - \mathrm{P}[p_i < t \mid stormy_i = 1]$ 作为权重函数，来估计价格为 t 时需求量的导数的平均值。换言之，使用 $stormy_i$ 做工具变量可以估计出导数 $q_i'(t)$ 的平均值，用到的权重是由工具变量引

① 本小节的英文标题是：So Long，and Thanks for All the Fish。该名字取自 Douglas Adams 的同名小说《So Long，and Thanks for All the Fish》，是他著名科幻小说系列“银河系漫游指南”中的第四本书。科幻迷们往往用这种方式诙谐地与人说再见。由于本小节讨论的例子是计量鱼的需求函数，所以这里还有双关的意思。——译者注

起的价格分布的不同。这个权重与我们在讨论平均因果响应时用到的权重在本质上是一样的，只不过这里没有出现因单位处理不同带来的概率差，代之以一个导数。

运用独立性假设和微积分基本定理，等式(4.5.6)给出了连续情况下的平均因果响应公式，得到这个公式的原因来自于下面的事实：

$$\mathrm{E}[q_i \mid stormy_i = 1] - \mathrm{E}[q_i \mid stormy_i = 0] = \mathrm{E}\left[\int_{P_{0i}}^{P_{1i}} q_i'(t)\mathrm{d}t\right] \quad (4.5.7)$$

从等式(4.5.7)中可以清楚地得到两个有趣的特例。第一是因果响应函数为线性，也即 $q_i(p) = \alpha_{0i} + \alpha_{1i}p$，其中 α_{0i} 和 α_{1i} 是随机参数。于是我们就有：

$$\frac{\mathrm{E}[q_i \mid stormy_i = 1] - \mathrm{E}[q_i \mid stormy_i = 0]}{\mathrm{E}[p_i \mid stormy_i = 1] - \mathrm{E}[p_i \mid stormy_i = 0]} = \frac{\mathrm{E}[\alpha_{1i}(P_{1i} - P_{0i})]}{\mathrm{E}[P_{1i} - P_{0i}]} \quad (4.5.8)$$

这个公式正是对随机参数 α_{1i} 的加权平均。这里使用的权重与由市场 i 上因为天气变化引起的价格变化成正比。

第二个特例就是当我们可以将数量需求曲线写为：

$$q_i(p) = Q(p) + \eta_i \quad (4.5.9)$$

其中，$Q(p)$ 是个非随机函数，η_i 是可加的随机误差。由此我们实际上假设了在每天、在每个市场上都有 $q_i'(p) = Q'(p)$。于是平均因果响应函数变为：

$$\int Q'(t)\omega(t)\mathrm{d}t, \text{其中 } \omega(t) = \frac{\mathrm{P}[P_{1i} \geqslant t > P_{0i}]}{\int \mathrm{P}[P_{1i} \geqslant r > P_{0i}]\mathrm{d}r}$$

其中，r 是分母中的积分元。

这个特例凸显了平均因果响应定理中的两类加权平均的过程以及如何将该定理推广到等式(4.5.6)所示的连续情况。首先，在各个市场之间存在一个平均化过程，该过程中用到的权重与每个市场上天气对价格的影响成正比。那些对天气变化最敏感的市场得到的权重最大。其次，在每个给定的市场上，在因果响应函数对应的区间内，都存在对因果效应的加权平均。由于工具变量可以移动价格的累积分布函数，因此在造成累积分布函数移动最大的那个价格区间上，工具变量对需求的导数进行平均化。

4.6 工具变量的细节

4.6.1 两阶段最小二乘中常犯的错误

我们可以很容易地计算 2SLS 值，特别是有了 SAS 和 Stata 这类统计软件后，它们可以自动为我们计算该数值。但也许你想自己算算试一下，看看上面讲过的

理论是否真是那样。或者假设你被困于 Krikkit 星球，购买的软件都过期（在 Douglas Adams 的科幻小说里，Krikkit 星球被封装在一个时间被锁定的信封里，因此在这个星球上如果购买的软件过期，那么要花很长的时间来更新）。当遇到这样一些紧急状况时，手动计算 2SLS 估计值就显得很必要了。在手动计算 2SLS 估计值过程中，你首先计算第一阶段值（在任何状况下你都要关注这一阶段），然后将第一阶段拟合值代入第二阶段的方程进行最小二乘估计。回到本章一开始用联立方程系统对工具变量法的描述，第一阶段和第二阶段分别是：

$$s_i = X_i'\pi_{10} + \pi_{11}'Z_i + \xi_{1i}$$
$$Y_i = \alpha' X_i + \rho\hat{s}_i + [\eta_i + \rho(s_i - \hat{s}_i)]$$

其中，X_i 是一组协变量，Z_i 是一组满足排他性约束的工具变量，第一阶段估计值就是 $\hat{s}_i = X_i'\hat{\pi}_{10} + \hat{\pi}_{11}'Z_i$。

手动计算 2SLS 估计值消除了我们对软件封装处理过程的神秘感，但是也开启了犯错误之门。如前所述，手动计算时在第二阶段最小二乘里得到的标准误可能不正确（最小二乘估计的残差方差是 $\eta_i + \rho(s_i - \hat{s}_i)$，但是在 2SLS 估计中你希望获得的标准误却只是对 η_i 得到的残差）。当然，还存在一些类似的潜在风险。

1. 协变量不一致

假设协变量向量包含两类变量，一类记为 X_{0i}，我们对这类协变量感到满意，一类记为 X_{1i}，我们对是否将这类协变量纳入回归感到很矛盾。Griliches 和 Mason (1972)在为工资方程构造 2SLS 估计值时就遇到了这个问题。在那项研究中，研究者将 AFQT(军方能力测试得分)视为需要工具变量进行识别的内生变量。他们使用的工具变量则是早期教育水平（在参军之前完成的）、种族和家庭背景变量。然后他们按照下面给出的方程估计了一个联立方程组：

$$s_i = X_{0i}'\pi_{10} + \pi_{11}'Z_i + \xi_{1i}$$
$$Y_i = \alpha_0' X_{0i} + \alpha_1' X_{1i} + \rho\hat{s}_i + [\eta_i + \rho(s_i - \hat{s}_i)]$$

这个方程看上去和手动计算 2SLS 估计值的那个方程组很相似。

但是仔细一看后发现该方程组和我们之前讨论过的方程组有重要区别：纳入第一阶段和第二阶段的协变量是不一样的。比如，Griliches 和 Mason 将年龄纳入第二阶段回归，但是却没有将其纳入第一阶段回归，这个问题由 Cardell 和 Hopkins(1977)在 Griliches 和 Mason(1972)进行评论时发现。这是 Griliches 和 Mason所犯的错误。Griliches 和 Mason 第二阶段估计值与 2SLS 估计值不同。更严重的问题在于此时他们估计出的结果是不一致的，相比之下用 2SLS 估计出的结果则更好。为了看清个中缘由，由于最小二乘回归的残差与纳入回归的回归元不相关，所以我们注意到第一阶段残差 $s_i - \hat{s}_i$ 与 X_{0i} 不相关。但是由于第一阶段回归未包括 X_{1i}，那么 X_{1i} 就很可能和第一阶段残差 $s_i - \hat{s}_i$ 相关也就是说年龄很可能与 Griliches 和 Mason(1972)中第一阶段中 AFQT 的残差相关）。这种相关性使得第二阶段估计出的所有参数都不是对相应总体参数的一致估计。这个故事带给我

们的教益在于：第一阶段回归和第二阶段回归中使用的外生协变量应该相同。如果协变量在第二阶段效果良好，那么它在第一阶段也应该效果良好。

2. 禁止回归

禁止回归最早由麻省理工学院的 Jerry Hausman 教授在 1975 年提出，这个问题碰巧出现在一份待评审的论文中，不过由于技术所限，当时没有将该问题说清楚。当研究者将 2SLS 直接应用于非线性模型时，就会出现禁止回归的问题。一个普遍出现的情况是内生变量是虚拟变量的情况，比如假设我们感兴趣的因果模型是：

$$Y_i = \alpha' X_i + \rho D_i + \eta_i \tag{4.6.1}$$

其中，D_i 是表示参军状态的虚拟变量。于是普通 2SLS 的第一阶段就是：

$$D_i = \pi'_{10} X_i + \pi'_{11} Z_i + \xi_{1i} \tag{4.6.2}$$

它表示对 D_i 关于一系列协变量和工具变量 Z_i 进行的线性回归。

由于 D_i 是一个虚拟变量，所以与第一阶段相联系的条件期望函数 $\mathrm{E}[D_i \mid X_i, Z_i]$ 应该是非线性的。因此在第一阶段使用最小二乘只能得到对非线性条件期望函数的一个近似。因此，我们可能希望在第一阶段使用非线性的方法来对条件期望函数进行更好的近似。假设现在使用 probit 模型来逼近 $\mathrm{E}[D_i \mid X_i, Z_i]$。那么第一阶段的 probit 模型就是 $\Phi[\pi'_{p0} X_i + \pi'_{p1} Z_i]$，其中 π_{p0} 和 π_{p1} 是 probit 模型中的参数，拟合值为 $\hat{D}_{pi} = \Phi[\hat{\pi}'_{p0} X_i + \hat{\pi}'_{p1} Z_i]$。在此例中被禁止的回归就是用第一阶段拟合值 $\hat{D}_{pi}$ 代替 D_i 后的第二阶段回归：

$$Y_i = \alpha' X_i + \rho \hat{D}_{pi} + [\eta_i + \rho(D_i - \hat{D}_{pi})] \tag{4.6.3}$$

等式(4.6.3)中存在的问题在于：只有对等式(4.6.2)进行最小二乘回归，才能保证产生的第一阶段残差与拟合值、协变量都不相关。如果 $\mathrm{E}[D_i \mid X_i, Z_i] = \Phi[\pi'_{p0} X_i + \pi'_{p1} Z_i]$，那么来自第一阶段的残差与 X_i 和 $\hat{D}_{pi}$ 只是渐进不相关。但是谁能说第一阶段回归的条件期望函数就一定是 probit 的呢？相比之下，稍微改变一下 2SLS，我们就可以不必担心第一阶段的条件期望函数是不是线性的了①。

对被禁止进行回归的第二阶段(4.6.3)做一个小小的变化，就可以避免由第一阶段不正确地设定非线性模型而带来的问题了。我们不再将非线性回归的拟合值代入第二阶段，而是用非线性回归的拟合值做工具变量。换言之，在传统的 2SLS 中用拟合值 $\hat{D}_{pi}$ 做 D_i 的工具变量(当然，外生协变量 X_i 也应该包括进入工具变量中)。用拟合值作为工具变量，或者将第一阶段来自最小二乘的拟合值代入第二阶段，两者看似相同，实则不同。使用非线性拟合值做工具变量的好处在于：如果非

① 在传统的联立方程组模型中得到一致的 2SLS 估计值并不依赖于第一阶段条件期望函数的形式是否正确设定，这一洞见可以追溯到 Kelejian(1971)。在第一阶段用非线性模型不会带来太多问题——第一阶段使用 probit 模型可以很好地近似线性——但当不必这样做时，为什么要冒这个险呢？

线性模型可以更好地近似第一阶段的条件期望函数，那么用非线性拟合值做工具变量得到的 2SLS 会更加有效(Newey，1990)。

但是必须指出这样做也是有缺点的。用非线性拟合值做工具变量暗示着我们将第一阶段的非线性作为一种信息来源，用以识别参数。为了看清这一点，假设我们感兴趣的因果模型包含一组工具变量 Z_i：

$$Y_i = \alpha' X_i + \gamma' Z_i + \rho D_i + \eta_i \tag{4.6.4}$$

当我们使用第一阶段(4.6.2)的拟合值时，这个模型无法得到识别，等式(4.6.4)的 2SLS 估计值也不存在。事实上，等式(4.6.4)违反了排他性约束。但是使用 X_i 和 Z_i 并将 $\hat{D}_i$ 作为工具变量的 2SLS 估计值则是存在的，因为 $\hat{D}_i$ 是 X_i 和 Z_i 的非线性函数，而且 Z_i 被排除于第二阶段。你是否应该使用第一阶段的非线性性作为一种识别信息的来源呢？我们常常避免使用这种不正当的识别手段，因为我们不清楚将非线性性用作一种识别信息的时候，它所对应的随机实验是什么。

因此，单纯地将第一阶段的非线性拟合值放入第二阶段并不是个很好的想法。这会使由此得到的第二阶段回归模型和第一阶段一样成为非线性的。举个例子，假设你认为教育水平和收入之间的因果关系是二次型的，但是在其他方面仍然是同质的(类似于 Card(1995)的结构性模型)。换言之，我们感兴趣的模型是：

$$Y_i = \alpha' X_i + \rho_1 s_i + \rho_2 s_i^2 + \eta_i \tag{4.6.5}$$

假设存在两个工具变量，那么即使 s_i 和 s_i^2 都有内生性，我们也很容易估计方程(4.6.5)。在这个例子中存在两个第一阶段方程，一个是针对 s_i 的，另一个是针对 s_i^2 的。虽然我们至少需要两个工具变量来进行识别，但是很自然的处理方法是使用原来的工具变量及其二次方(除非唯一的工具变量是虚拟变量，如果这样的话我们需要更好的主意来解决这个问题)。

尽管如此，你可能还会尝试在第一阶段回归中使用类似等式(4.6.2)的方程并估计如下的第二阶段方程：

$$Y_i = \alpha' X_i + \rho_1 \hat{s}_i + \rho_2 \hat{s}_i^2 + [\eta_i + \rho_1 (s_1 - \hat{s}_i) + \rho_2 (s_i^2 - \hat{s}_i^2)]$$

这样做是错的，因为 $\hat{s}_i$ 可能与 $s_i^2 - \hat{s}_i^2$ 相关，而 $\hat{s}_i^2$ 可能和 $s_i - \hat{s}_i$ 与 $s_i^2 - \hat{s}_i^2$ 都相关。相比之下，一旦等式(4.6.5)中 X_i 和 Z_i 都与 η_i 不相关而且在向量 Z_i 中有足够的工具变量，那么很显然应该对方程(4.6.5)使用 2SLS。

4.6.2 同群效应

在社会科学中有大量文献在研究同群效应(peer effect)。大体上来说，同群效应是指群体特征对个体行为的因果效应。有时人们试图用回归分析揭开同群效应的面纱。在实际中，用回归模型揭示同群效应充满问题。尽管就问题本身而言同群效应不仅仅是用工具变量进行估计的问题，但是运用 2SLS 的语言和代数形式，

我们可以更好地理解为什么同群效应难以识别。

大体而言，存在两类同群效应。第一类同群效应考虑的是群体特征——某个州或城市的平均教育水平——对个体的影响，这里要用另外一个变量来描述对个体产生的影响。比如 Acemoglu 和 Angrist(2000)考虑了是否个体收入受其居住州的平均教育水平影响。人力资本外部性理论指出如果所在州的人均教育水平较高的话，那么这个州的居民的生产率会高一些，而且这种较高的生产率并非来自教育。这种外溢性就叫做教育的社会回报：不论个体是否接受教育，人力资本都会让每个人获益。

用来研究这种外部性的因果模型可以写为：

$$Y_{ijt} = \mu_j + \lambda_t + \gamma \bar{S}_{jt} + \rho s_i + \mu_{jt} + \eta_{ijt} \tag{4.6.6}$$

其中，Y_{ijt} 是个体 i 于第 t 年在州 j 获得的每周工资的对数值，μ_{jt} 是州—年份误差项，η_{ijt} 是个体的误差项。变量 μ_j 和 λ_t 用以控制居住所在地和观察值所在年份对应的固定效应。参数 ρ 是个体接受教育的经济回报，γ 用来捕捉在第 t 年州 j 的平均教育水平 $\bar{S}_{jt}$ 带来的效应。

除了要考虑如何识别 s_i 的因果性外，方程(4.6.6)中存在的最重要的识别问题在于误差项 μ_{jt} 中的州—年份效应会与州平均教育水平相关，因此带来遗漏变量偏误。比如在经济周期的上升期，公立大学系统可能存在周期性扩张，从而造成平均教育水平和平均收入水平的共同上升。在 Acemoglu 和 Angrist(2000)中，两位作者尝试用来自义务教育法的历史数据做工具变量来解决这个问题，这个工具变量与 $\bar{S}_{jt}$ 相关，但是与同期的 μ_{jt} 和 η_i 无关。

虽然州—年份效应中存在遗漏变量是 Acemoglu 和 Angrist(2000)进行工具变量估计的主要原因，但回归元 $\bar{S}_{jt}$ 是另一个回归元 s_i 的平均值这一事实令等式(4.6.6)的最小二乘估计值变得难以解释。为了看清楚这一点，将等式(4.6.6)简化到只存在截面维度。可以将总体回归方程写为：

$$Y_{ij} = \mu + \pi_0 s_i + \pi_1 \bar{s}_j + \nu_{ij} \tag{4.6.7}$$

其中，Y_{ij} 是个体 i 于在州 j 获得的周工资对数值，$\bar{S}_j$ 是该州的平均受教育水平。参数 π_0 和 π_1 的意义之前已经定义过，由构造可知 ν_{ij} 与回归方程(4.6.7)中的回归元都不相关。现在令 ρ_0 表示仅用 s_i 对 Y_{ij} 做回归得到的参数，令 ρ_1 表示仅用 $\bar{S}_j$ 对 Y_{ij} 做回归得到的参数。由本章前面对分组数据和 2SLS 之间关系的讨论可知，用标志所有城市的虚拟变量做工具变量后，ρ_1 就是 Y_{ij} 关于 s_i 的二元回归的 2SLS 估计值。在本章附录中我们使用这个事实来指出等式(4.6.7)中的参数可以表达为 ρ_0 和 ρ_1 的组合，即：

$$\begin{aligned} \pi_0 &= \rho_1 + \phi(\rho_0 - \rho_1) \\ \pi_1 &= \phi(\rho_1 - \rho_0) \end{aligned} \tag{4.6.8}$$

其中，$\phi = \dfrac{1}{1-R^2} > 1$，$R^2$ 为使用标志州的虚拟变量作 s_i 的工具变量时在第一阶

段回归的拟合优度。

等式(4.6.8)的要点在于，如果因为任何原因使得对工资用个人教育水平做回归得到的最小二乘估计值与使用州虚拟变量作为工具变量得到的 2SLS 估计值不同，那么等式(4.6.7)中平均教育水平前的系数都不会为零。例如，如果用州虚拟变量做工具变量来纠正由于 s_i 中存在度量误差而带来的微小偏误，我们可得 $\rho_1 > \rho_0$ 以及看似为正的教育的社会回报。相反，如果用州虚拟变量做工具变量消除了由 s_i 和不可观测的收入潜力之间存在正相关而带来的偏误，我们可得 $\rho_1 < \rho_0$ 以及教育的社会回报为负的结论①。因此，在实际中很难通过对方程(4.6.6)做最小二乘估计来分离教育的社会回报，不过同时将个体教育水平和群体平均教育水平当作内生变量，用更复杂的工具变量法进行估计的策略或许可行。

第二类同群效应是一类更难以去度量的同群效应，它考虑的是在同一个变量上群体均值如何影响个体值。当然，这个问题并不是真正意义上的工具变量问题；我们需要回到最基本的回归。为了看清楚这一点，假设 $\bar{S}_j$ 是学校 j 的高中毕业率，我们想知道的问题是：处在毕业率很高的高中是不是可以提高单个人高中毕业的可能性。为了寻找高中毕业率上存在的同群效应，我们可能要从下面这个回归方程入手：

$$s_{ij} = \mu + \pi_2 \bar{S}_j + \nu_{ij} \tag{4.6.9}$$

其中，s_{ij} 是个体 i 的高中毕业状态，$\bar{S}_j$ 是在学校 j 的平均毕业率，个体 i 的毕业状况也计入学校 j 的平均毕业率。

初看上去，等式(4.6.9)似乎是一个经过很好定义的因果模型，但实际上它没什么意义。如果我们意识到 $\bar{S}_j$ 就是在第一阶段用 s_{ij} 对关于所有标志学校的虚拟变量做回归得到的拟合值，那么就会明白用 $\bar{S}_j$ 对 s_{ij} 做回归得到的参数必将是 1②。因此，类似于等式(4.6.9)的方程是无法为因果效应带来有意义的信息的。对这样一个不甚良好的同群效应进行修改的方法就是将方程(4.6.9)修改为：

$$s_{ij} = \mu + \pi_3 \bar{S}_{(i)j} + \nu_{ij} \tag{4.6.10}$$

其中，$\bar{S}_{(i)j}$ 是在学校 j 中排除个体 i 后对 s_{ij} 计算出的平均值。这是正确方向上的第一步——π_3 不再自动等于 1——但是问题仍然存在，因为 s_{ij} 和 $\bar{S}_{(i)j}$ 都被学校层面

① 存在个体教育水平的方程中平均教育水平之前的系数可以解释为一种统计检验值，该统计检验是由 Hausman(1978)提出的，用以检验在研究教育的私人回报时最小二乘估计值和 2SLS 估计值之间的等价性，这里用表示州的虚拟变量做工具变量。Borjas(1992)讨论了类似的问题，那里部落背景会影响到估计结果。

② 对于用 s_{ij} 对 $\bar{S}_j$ 做回归得到的系数是 1 这件事实，这里提供一个简单的证明：

$$\frac{\sum_j \sum_i s_{ij}(\bar{S}_j - \bar{S})}{\sum_j n_j (\bar{S}_j - \bar{S})^2} = \frac{\sum_j (\bar{S}_j - \bar{S}) \sum_i s_{ij}}{\sum_j n_j (\bar{S}_j - \bar{S})^2}$$

$$= \frac{(\bar{S}_j - \bar{S})(n_j \bar{S}_j)}{\sum_j n_j (\bar{S}_j - \bar{S})^2} = 1$$

的随机干扰影响，但是我们只能将这种干扰置于 ν_{ij} 之中。在误差项中存在群体的随机效应成为统计推断中很重要的问题，我们将在第 8 章详细讨论这部分内容。但是在等式(4.6.10)中群体的随机效应更多地影响了标准误：任何对整体群体都相同的冲击都带来伪同群效应。例如，做事特别有效率的校长可能会在每一所他工作过的学校中提高每个学生的毕业可能性。但即使在群体均值和单个学生学习成绩之间没有因果联系，上面出现的这种情况意味着在 s_{ij} 和 $\bar{S}_{(i)j}$ 之间存在相关性，看上去很像一种同群效应。因此我们同样不选择使用类似于方程(4.6.10)的回归。

对同群效应中因果关系进行研究的最好着眼点是关注事前的群体特征，也即关注事情发生之前对群体质量的度量，因为这些变量不受随机冲击的影响。近期的一个例子来自于 Ammermueller 和 Pischke(2006)，他们在欧洲的小学中研究了同学的家庭背景和学生学习成绩之间的联系，这里用每个家庭中拥有的书籍数量来度量学生的家庭背景。Ammermueller 和 Pischke 的回归方程是：

$$s_{ij} = \mu + \pi_4 \bar{B}_{(i)j} + \nu_{ij}$$

其中，$\bar{B}_{(i)j}$ 是学生 i 的同学的家里藏书的平均量。这个回归方程看上去和(4.6.10)类似，但是有很大的不同。变量 $\bar{B}_{(i)j}$ 是对学生考试前就已经存在的家庭环境的度量，因此不会被学校层面的随机冲击所干扰。

Angrist 和 Lang(2004)为我们提供了另外一个例子，在该例中他们试图在学生成绩和其同学在事前的特点之间建立联系。Angrist 和 Lang 的研究考察了将学习成绩较差的学生转入学习成绩较好的班级后对该班级原有学生成绩的影响。在这个例子中他们感兴趣的回归乃是：

$$s_{ij} = \mu + \pi_5 \bar{m}_j + \nu_{ij} \tag{4.6.11}$$

其中，$\bar{m}_j$ 是转入学校 j 的学习较差的人数，s_{ij} 是学生 i 的测验分数。这里我们无需再为由共同的冲击带来的伪相关问题担心，原因有二：首先，$\bar{m}_j$ 是由回归(4.6.11)样本之外的学生决定的；其次，转入的低分学生的数量是由他们来自于何方决定的，不是被 s_{ij} 决定的。但是，由于学校层面的随机冲击带来的影响仍然是 ν_{ij} 的一部分，这对统计推断仍有很大影响，因为 $\bar{m}_j$ 是个群体层面的变量。

4.6.3 再论有限被解释变量

在 3.4.2 节，我们讨论了回归模型中存在有限解释变量的后果。当解释变量是二元值或者非负值时——比如表征就业状态的虚拟变量或者表征工作小时数的非负变量——条件期望函数会是非线性的。大部分非线性有限被解释变量模型都建立在将线性的潜在得分模型转化为非线性模型。这种例子包括 probit 模型、logit 模型和 Tobit 模型。这些模型可以捕捉到相应条件期望函数的特点(比如 probit 模型可以保证拟合值在 0 和 1 之间，Tobit 模型则可以保证拟合值是非负

的)。在那里我们讨论的结果是:因为存在有限被解释变量,我们研究了潜在得分模型,相比之下这类模型更加复杂,对计量结论也需要进一步解释,但这种增加了的复杂性似乎并不值得。

支持使用最小二乘估计的一个重要原因在于该方法在概念上是稳健的,这也正是结构模型所缺乏的。最小二乘法往往可以给出条件期望函数的最小均方误近似。而且,事实上我们还可以将最小二乘法看作一种用来计算边际效应的方法——一种简单、可自动执行且不同研究中使用该方法得到的结论之间可以相互比较的方法。而非线性的潜在得分模型则更像是广义最小二乘法:它使我们在估计效率上得到改进,但是却要求对函数形式和分布做出更强假设,但在这一点上我们往往不是很有信心①。支持使用最小二乘估计的第二个原因在于从潜在得分模型中估计出的系数和用最小二乘得到的因果效应之间存在不同,前者居于非线性模型的核心,而后者则是大部分研究项目所主要感兴趣的变量。

相比于受限被解释变量模型,我们更加支持最小二乘法,对存在内生变量的模型和 2SLS,上面举出的道理同样也成立。不论被解释变量取二元值、非负值还是连续值,工具变量法都可以捕捉到局部平均处理效应。当模型中存在协变量时,每个协变量都对应于一个局部平均处理效应,2SLS 估计值将这些局部平均处理效应进行加权平均。当模型中存在可变甚至是连续的处理时,2SLS 估计值告诉我们平均因果效应是多少,特别是当处理为连续时,2SLS 估计值求出的是因果效应函数的导数的平均值。尽管 Abadie(2003)指出,从一般意义上来看 2SLS 估计值无法提供依从工具变量者的因果效应函数的最小均方误逼近,但是在实际应用中 2SLS 估计值和严格的 Abadie 过程(当协变量饱和,2SLS 与 Abadie 过程相同)得到的估计值之间差别很小。而且,2SLS 直接估计出了局部平均处理效应,无需考虑计算边际效应的中间步骤。

2SLS 不是唯一的可行方法。另外一个更为复杂的方法试图通过仔细刻画产生有限被解释变量产生的过程来求解因果效应。二元变量的 probit 模型就是个很好的例子,这个方法应用于 Angrist 和 Evans(1998)的研究。假设一位女性通过比较生育第三胎的成本和收益做出是否生育第三胎的决策,其决策过程可以用收益

① 实际上,非线性有限被解释变量模型和广义最小二乘法之间具有很大的相似性。考虑具有非线性条件期望函数的 probit 模型 $E[Y_i \mid X_i] = \Phi\left[\frac{X_i'\beta^*}{\sigma}\right] \equiv r_i$。这个模型的极大似然估计的一阶条件为:

$$\sum_i \frac{(Y_i - r_i)X_i}{r_i(1-r_i)} = 0$$

对非线性回归模型的广义最小二乘考虑的是下面这个模型:

$$Y_i = \Phi\left[\frac{X_i'\beta^*}{\sigma}\right] + \xi_i$$

由于 Y_i 的条件方差是 $r_i(1-r_i)$,所以从渐进的角度讲,极大似然估计和广义最小二乘估计是一样的。probit 模型和广义最小二乘估计之间存在的唯一不同在于广义最小二乘法要通过两步完成。

函数或者潜在得分模型表示。假设用潜在得分模型表示，该模型关于协变量是线性的，不含工具变量，随机误差项记为 ν_i。那么这个二元变量的 probit 模型的第一阶段就可以写为：

$$D_i = 1[X_i'\gamma_0^* + \gamma_1^* Z_i > \nu_i] \tag{4.6.12}$$

其中，Z_i 是工具变量，给定协变量 X_i 后 Z_i 的变动会提高生育第三胎的收益。比如在美国家庭中，如果父母已经生育两个男孩或者两个女孩，那么他们会更愿意生育第三个孩子，可以将美国家庭的这种选择看作是在生育问题上的多样化，这种多样化提高了在前两个孩子性别相同时父母生育第三个孩子的收益。

在这个例子中，我们主要关心的变量是就业状况，这个状况可用伯努利随机变量表示，该变量的均值介于 0 和 1 之间。为了完成这个模型，假设就业状况是 Y_i，它是由下面的潜在得分过程决定的：

$$Y_i = 1[X_i'\beta_0^* + \beta_1^* D_i > \varepsilon_i] \tag{4.6.13}$$

其中，ε_i 是第二个随机部分或者说第二个误差项。可以将这个潜在得分模型看作是在工作的成本和收益之间进行权衡。

二元 probit 模型中存在的遗漏变量偏误与误差项 ν_i 和 ε_i 都有关系。换言之，决定生育但是我们无法度量的那些随机因素可能和决定就业状况但是我们无法度量的随机因素之间存在关系。通过假设这些随机因素符合正态分布且与工具变量 Z_i 相独立，我们可以对这个二元 probit 模型进行识别。给定正态性，可以通过极大似然估计得到方程(4.6.12)和(4.6.13)中的参数。相应的对数似然函数为：

$$\begin{aligned}\sum Y_i \ln \Phi_b\left(\frac{X_i'\beta_0^* + \beta_1^* D_i}{\sigma_\varepsilon}, \frac{X_i'\gamma_0^* + \gamma_1^* Z_i}{\sigma_\nu}; \rho_{\varepsilon\nu}\right) \\ + (1-Y_i)\ln\left[1 - \Phi_b\left(\frac{X_i'\beta_0^* + \beta_1^* D_i}{\sigma_\varepsilon}, \frac{X_i'\gamma_0^* + \gamma_1^* Z_i}{\sigma_\nu}; \rho_{\varepsilon\nu}\right)\right]\end{aligned} \tag{4.6.14}$$

其中，$\Phi_b(\cdot, \cdot; \rho_{\varepsilon\nu})$是二元正态分布函数，相关系数为 $\rho_{\varepsilon\nu}$。但是需要注意的是，用一个正数乘以倾向得分的系数和标准误(σ_ε, σ_ν)不改变这个似然函数，因此我们希望估计的参数倾向得分参数和标准误之比(也就是 $\beta_1^*/\sigma_\varepsilon$)。

于是由二元 probit 模型定义出的潜在结果就可记为：

$$Y_{0i} = 1[X_i'\beta_0^* > \varepsilon_i] \text{ 以及 } Y_{1i} = 1[X_i'\beta_0^* + \beta_1^* > \varepsilon_i]$$

在工具变量的影响下，对潜在处理状态的分配是：

$$D_{0i} = 1[X_i'\gamma_0^* > \nu_i] \text{ 以及 } D_{1i} = 1[X_i'\gamma_0^* + \gamma_1^* > \nu_i]$$

正如之前一直讨论的，对每个人，我们都只能观察到一个潜在结果和一个潜在的处理状态。上面两个表达式中还可以清楚地看出：ν_i 和 ε_i 之间的相关性和潜在结果与潜在处理状态之间的相关性是同一件事情。

就我们现在举出的这个例子而言，估计出的潜在得分系数并不能告诉我们因

果效应的大小，能告诉我们的只是因果效应的正负。为了看清楚这一点，注意到生育孩子的平均因果效应是：

$$\mathrm{E}[Y_{1i}-Y_{0i}]=\mathrm{E}\{1[X_i'\beta_0^*+\beta_1^*>\varepsilon_i]-1[X_i'\beta_0^*>\varepsilon_i]\}$$

其中，生育第三胎的女性的平均因果效应为：

$$\begin{aligned}&\mathrm{E}[Y_{1i}-Y_{0i}\mid D_i=1]\\&=\mathrm{E}\{1[X_i'\beta_0^*+\beta_1^*>\varepsilon_i]-1[X_i'\beta_0^*>\varepsilon_i]\mid X_i'\gamma_0^*+\gamma_1^*Z_i>\nu_i\}\end{aligned}$$

如果假设 ν_i 和 ε_i 的分布是另外一种分布，那么上面求出的这两个因果效应可以等于任何值。（如果误差项是异方差的，那么上面两个因果效应的符号也不确定了。）

如果我们假设了正态分布，那么由二元 probit 模型产生的平均因果效应就很容易计算。其平均因果效应是：

$$\begin{aligned}&\mathrm{E}\{1[X_i'\beta_0^*+\beta_1^*>\varepsilon_i]-1[X_i'\beta_0^*>\varepsilon_i]\}\\&=\mathrm{E}\left\{\Phi\left[\frac{X_i'\beta_0^*+\beta_1^*}{\sigma_\varepsilon}\right]-\Phi\left[\frac{X_i'\beta_0^*}{\sigma_\varepsilon}\right]\right\}\end{aligned}\tag{4.6.15}$$

其中，$\Phi[\cdot]$是正态分布的累积概率函数。生育第三胎的女性的因果效应计算起来会更加复杂一些，因为它包含了二元正态分布的累计概率函数：

$$\begin{aligned}&\mathrm{E}[Y_{1i}-Y_{0i}\mid D_i=1]\\&=\mathrm{E}\left\{\frac{\Phi_b\left(\frac{X_i'\beta_0^*+\beta_1^*}{\sigma_\varepsilon},\frac{X_i'\gamma_0^*+\gamma_1^*Z_i}{\sigma_\varepsilon};\rho_{\varepsilon\nu}\right)-\Phi_b\left(\frac{X_i'\beta_0^*}{\sigma_\varepsilon},\frac{X_i'\beta_0^*+\gamma_1^*Z_i}{\sigma_\nu};\rho_{\varepsilon\nu}\right)}{\Phi\left(\frac{X_i'\gamma_0^*+\gamma_1^*Z_i}{\sigma_\nu}\right)}\right\}\end{aligned}\tag{4.6.16}$$

很多软件包都有计算二元正态分布的标准函数模块，所以在实际中等式(4.6.15)和(4.6.16)表示的因果效应是很容易计算的。

应该说二元 probit 模型是满足在第 1 章我们对无害的计量经济学的要求的，因为它不是很复杂而且在常用的软件包中就可以计算。但是，这个模型还是存在我们在 3.4.2 节讨论的非线性潜在得分模型的缺点。首先，在非线性潜在得分模型中，花费精力后估计出的只是一个参数而非平均因果效应，这一点让很多研究者感到相当不满意。比如在计量经济学中有大量文献关注的是无需分布假设下得分系数的估计策略。对于那些只对因果效应感兴趣的研究者而言，这部分文献①确实可以忽略，因为无论怎样，该模型估计出来的都不是一个因果效应。

---

① 假设潜在误差项的分布未知，其累积概率函数为 $\Lambda[\cdot]$。这时平均因果效应就是：

$$\mathrm{E}\{\Lambda[X_i'\beta_0^*+\beta_1^*]-\Lambda[X_i'\beta_0^*]\}=\Lambda'[X_i'\beta_0^*+\tilde{\beta}_1]\beta_1^*$$

其中，（由中值定理）$\tilde{\beta}_1$ 是落在$[0,\beta_1^*]$中的一个数字。由于上面的等式往往有赖于 $\Lambda[\cdot]$的形状，所以我们无法单独知道参数的系数。

觉得非线性潜在得分模型还不错的第二种观点认为相比于 2SLS，该模型是有长处的。二元 probit 模型以及类似的其他模型可被用于估计无条件的平均因果效应以及被处理者的平均因果效应。相比之下，2SLS 只能为我们提供局部平均处理效应，而无法提供平均因果效应。但是从等式(4.6.15)中我们应该清楚地看到：之所以非线性潜在得分模型可以计算无条件平均因果效应，是因为我们假设潜在得分的随机部分是符合正态分布的。但是如果你无法做出这样的假设，那么你能做到的最好的结果就是计算依从工具变量者的平均因果效应。对于二元 probit 模型，可将局部平均处理效应记为：

$$\begin{aligned}&\mathrm{E}[Y_{1i}-Y_{0i}\mid D_{1i}>D_{0i}]\\=&\mathrm{E}\{1[X_i'\beta_0^*+\beta_1^*>\varepsilon_i]\\&-1[X_i'\beta_0^*]\mid X_i^*\gamma_0^*+\gamma_1^*>\nu_i>X_i'\gamma_0^*\}\end{aligned}$$

与方程(4.6.16)一样，我们也可以使用 ν_i 和 ε_i 的联合正态分布对这个方程进行估计。但是在估计 $[Y_{1i}-Y_{0i}\mid D_{1i}>D_{0i}]$ 时我们也可以不必使用正态分布的假设，因为对于协变量的每个取值，工具变量法可以估计出一个局部平均处理效应，然后用协变量的方差作为权重求得加权后的局部平均处理效应。或者就像在 4.5.3 节讨论过饱和加权定理那样，直接用 2SLS 对与协变量相对应的局部平均处理效应进行加权。

你可能想知道得到局部平均处理效应是否已经足够。可能你希望通过做出一些额外的假设来估计无条件平均因果效应或者被处理者的因果效应。那样做当然很好，但是我们的经验指出即使你敢于做出这种勇敢的假设，目标也很难达到。因为有关局部特征的信息完全包含在数据中，而且在实际操作中如果协变量的变化足够大，二元 probit 模型中得到的平均因果效应很可能接近于 2SLS 估计值。表 4.8 中的计量结果就传达了这样的看法，它报告了使用 Angrist-Evans(1998)中的性别组成工具变量，分别使用 2SLS 回归和二元 probit 模型得到的女性生育第三胎对劳动力供给的影响。其中被解释变量为上一年是否参加工作的虚拟变量，内生变量是表示是否生育第三胎的虚拟变量。在第一阶段由相同性别组成导致的生育第三胎的概率是 7%。

表 4.8 的 A 部分报告了不加入任何协变量的回归结果。第 1 列中的 2SLS 估计值是−0.138，与第 2 列中用 Abadie 方法估计到的结果几乎完全相同，当然理应如此。与 Abadie-Kappa 加权过程类似，当不存在协变量时 2SLS 估计出的斜率给我们提供了对依从工具变量者因果响应函数的最佳线性估计。如果我们针对 probit 模型的条件期望函数使用非线性最小二乘估计，而不是使用线性近似，那么得到的对边际效应的估计值几乎没有变化。估计边际效应的方法就是针对下式进行最小化：

$$\mathrm{E}\left\{\kappa_i\left(Y_i-\Phi\left[\frac{\beta_0^*+\beta_1^*D_i}{\sigma_\varepsilon}\right]\right)^2\right\}$$

表 4.8　生育第三胎对妇女劳动力供给的影响，分别用两阶段最小二乘回归、Abadie 方法以及二元 probit 模型

	Abadie 估计值		二元 Probit 估计值		
2SLS 法	线性	Probit	MFX	ATE	TOT
(1)	(2)	(3)	(4)	(5)	(6)
A. 无协变量					
−0.138	−0.138	−0.137	−0.138	−0.139	−0.139
(0.029)	(0.030)	(0.030)	(0.029)	(0.029)	(0.029)
B. 存在一些协变量(没有控制年龄)					
−0.132	−0.132	−0.131	−0.135	−0.135	−0.135
(0.029)	(0.029)	(0.028)	(0.028)	(0.028)	(0.028)
C. 协变量加上第一胎出生年龄					
−0.132	−0.139	−0.129	−0.133	−0.133	−0.133
(0.028)	(0.028)	(0.028)	(0.026)	(0.026)	(0.026)
D. 协变量加上第一胎出生年龄和是否在 30 岁前生育的虚拟变量					
−0.124	−0.125	−0.125	−0.131	−0.131	−0.131
(0.028)	(0.029)	(0.029)	(0.025)	(0.025)	(0.025)
E. 协变量加第一胎出生年龄和年龄					
−0.120	−0.121	−0.121	−0.171	−0.171	−0.171
(0.028)	(0.026)	(0.026)	(0.023)	(0.023)	(0.023)

注：这个表来自于 Angrist(2001)。表中考察妇女生育第三胎对其劳动力供给的影响，将 2SLS 估计值与非线性模型估计值进行了比较。所有的模型都是用性别组成作为工具变量。Abadie 方法下的标准误是使用大小为 20 000 的子样本进行 100 次的重复的 bootstrap 法得到的。MFX 表示边际效应；ATE 表示无条件的平均因果效应；TOT 是被处理者的平均处理效应。

这里我们估计出的是−0.137，报告在第 3 列中。这并不令人惊讶，因为当不存在协变量时我们对模型的函数形式没有假设。

可能更加令人吃惊的事情是我们运用公式(4.6.15)和(4.6.16)计算出的平均因果效应与 2SLS 估计值和 Abadie 方法下得到的估计值也相同。这些结果报告在第 4 至 6 列。用导数近似公式(4.6.15)中的差值后得到的边际效应是−0.138(在第 4 列，标记为 MFX)，而第 5 列和第 6 列的平均因果效应都是−0.139。从表 4.8 中的 B 部分可以看出，加入协变量对估计结果几乎没有影响。在 B 部分使用的协变量都是虚拟变量，用三个虚拟变量表征种族(黑人、西班牙裔和其他)，用两个虚拟变量表示第一胎和第二胎所生孩子为男孩(满足排他性约束的工具变量是这两个虚拟变量的交互项)。表 4.8 中的 C、D 部分指出加入生育第一胎时年龄的线性项以及加上标志母亲年龄的虚拟变量都不改变估计结果。

估计结果不随协变量的加入而改变看上去很合意：因为从本质上讲相同性别工具变量与协变量是相互独立的，因此加入协变量进行控制未必可以消除偏误，也不会对估计结果的精确性产生影响。当然，表 4.8 中的 E 部分显示用二元 probit 模型产生的边际效应估计值对一系列工具变量都很敏感。将标志母亲年龄超过三

十岁的工具变量加入模型，同时将年龄的线性项加入模型后，二元 probit 模型的估计值升至−0.171，而 2SLS 和 Abadie 方法下的估计值则也没有发生改变。这大概因为加入的年龄的线性项改变了原有协变量所对应的个体，使得估计结果向不太有数据的那些个体偏移。尽管在 E 部分报告二元 probit 效应本身是无害的，但是很难解释为什么不选择更加稳健的 2SLS 估计和 Abadie 估计值[①]。

4.6.4 两阶段最小二乘估计值的偏误*

很幸运，最小二乘估计不仅是一致的，而且是无偏的（在 3.1.3 节最后一部分我们提到了这个结论）。这意味着无论样本有多大，估计出的最小二乘系数向量都有一个以总体系数向量为中心的分布。[②]相比之下，2SLS 估计值是一致的，但是有偏的。这意味着只有在大样本 2SLS 估计值才会接近因果效应。在小样本中，2SLS 估计可能会系统性地偏离目标参数。

在过去很多年中，应用研究者知道 2SLS 估计值是有偏的，但却没有对这个事情给予过多关注。本书的两位作者当年读研究生时参加的计量经济学课程中也没有涉及 2SLS 估计值的有偏性。但是到了 90 年代初期，一系列论文开始改变人们对该问题的认识。这些研究指出在有关经验研究实践中，2SLS 估计值可能存在相当的偏误。[③]

当工具变量是“弱工具变量”时，2SLS 估计值很可能有偏，当存在过度识别时，2SLS 估计也很可能有偏。所谓弱工具变量，是指内生回归元和工具变量之间的相关性很低。当工具变量既存在过度识别问题又存在弱工具变量问题时，2SLS 估计值会依概率偏向相应的最小二乘估计值。在最差的情况下，当工具变量弱到总体中不存在第一阶段时，2SLS 估计值的样本分布就是以最小二乘估计值为中心的概率极限分布。这个结果背后的理论比较有技术性，但基本的想法还是易于理解的。2SLS 估计值的有偏性来自于求解第一阶段拟合值时引入的随机性。在实际中，由于第一阶段参数来自于内生变量关于工具变量的回归，所以第一阶段估计值反映出的是内生变量中包含的随机性。因此如果总体的第一阶段估计值为零，那么第一阶段中所有的随机项都是因为内生变量的缘故。因此内生变量与第二阶段的误差相联系（否则我们不会一开始就用工具变量了），所以当样本有限时，这种随机性

① Angrist(2001)用双胞胎工具变量也得到相同结论。该研究中作者考察了生育对个体劳动时间的影响，针对劳动时间进行 2SLS、Abadie 和非线性结构模型的估计后发现，估计值表现出与表 4.8 类似的模式。

② 更精确的陈述是：下列条件居其一时，最小二乘估计无偏：(1)条件期望函数线性；(2)回归元非随机，即重复抽样中回归元固定。不过在实际中这些要求并不起作用。$\hat{\beta} = [\sum_i X_i X_i']^{-1} \sum_i X_i Y_i$ 的样本分布以总体系数 $\beta = E[X_i X_i']^{-1} E[X_i Y_i]$ 为中心，因此与样本大小、条件期望函数是否线性及回归元是否随机无关。

③ 主要参考文献可见 Nelson 和 Startz(1990a, b)，Buse(1992)，Bekker(1994)以及 Bound，Jaeger 和 Baker(1995)。

可能导致第一阶段拟合值与第二阶段残差相关。

对 2SLS 估计有偏性的正式讨论可以如下进行。简单起见，我们使用矩阵和向量的语言，在常因果效应的框架中进行讨论(在异质性因果效应的框架下这个问题很难讨论，因为随着工具变量的改变，我们得到的目标系数实际上也在改变)。假设你想估计的单个内生回归元 x 对被解释变量 y 的影响，其中 x 和 y 都是向量，假设这里不存在协变量。于是我们关心的因果效应模型可以记为：

$$y = \beta x + \eta \tag{4.6.17}$$

工具变量 Z 所构成的是一个 $N \times Q$ 的矩阵，相应的第一阶段公式为：

$$x = Z\pi + \xi \tag{4.6.18}$$

等式(4.6.17)的最小二乘估计是有偏的，因为 η_i 和 ξ_i 相关。根据构造，我们知道 Z_i 与 ξ_i 不相关，根据对工具变量做出的排除性约束，Z_i 与 η_i 也不相关。

于是 2SLS 估计值就是：

$$\hat{\beta}_{2SLS} = (x'P_Z x)^{-1}x'P_Z y = \beta + (x'P_Z x)^{-1}x'P_Z \eta$$

其中，$P_Z = Z(Z'Z)^{-1}Z'$ 是产生第一阶段拟合值的映射矩阵。将 x 代入 $x'P_Z\eta$ 后我们有：

$$\begin{aligned}\hat{\beta}_{2SLS} - \beta &= (x'P_Z x)^{-1}(\pi'Z' + \xi')P_Z\eta \\ &= (x'P_Z x)^{-1}\pi'Z'\eta + (x'P_Z x)^{-1}\xi'P_Z\eta \end{aligned} \tag{4.6.19}$$

于是 2SLS 回归估计值的有偏性来自于等式(4.6.19)中等号右边的这部分的期望不为零。

由于矩阵求逆运算得到的是一个非线性函数，所以期望算子无法进入 $(x'P_Z x)^{-1}$，这就导致我们很难计算等式(4.6.19)的期望值。但我们还是可以指出等式(4.6.19)中等号右边的两个比值大致近似于其期望的比值，也即：

$$\mathrm{E}[\hat{\beta}_{2SLS} - \beta] \approx (\mathrm{E}[x'P_Z x])^{-1}\mathrm{E}[\pi'Z'\eta] + (\mathrm{E}[x'P_Z x])^{-1}\mathrm{E}[\xi'P_Z\eta] \tag{4.6.20}$$

这个近似比之前提到在大样本中成立的渐进估计更好，因此它带给我们关于 2SLS 估计值一个相当好的有限样本性质①。更进一步，因为 $\mathrm{E}[\pi'Z'\xi] = 0$ 以及 $\mathrm{E}[\pi'Z'\eta] = 0$，我们有：

$$\mathrm{E}[\hat{\beta}_{2SLS} - \beta] \approx [\mathrm{E}(\pi'Z'Z\pi) + \mathrm{E}(\xi'P_Z\xi)]^{-1}\mathrm{E}(\xi'P_Z\eta)$$

因此 2SLS 估计的渐进偏误来源于 $\mathrm{E}(\xi'P_Z\eta)$ 非零，当然，如果 η_i 和 ξ_i 不相关，那么 $\mathrm{E}(\xi'P_Z\eta)$ 为零。但正是 η_i 和 ξ_i 的相关性使我们一开始就使用 2SLS 估计。

① 见 Bekker(1994)以及 Angrist 和 Krueger(1995)。还可将此结论称为组渐进估计(group-asymptotic approximation)，因为我们可以从工具变量个数趋于无穷、观察值个数也趋于无穷但是每个工具变量对应的观察值不变的渐进序列中得到该结论。

对等式(4.6.20)的进一步转化产生下面这个等式，它特别有用：

$$\mathrm{E}[\hat{\beta}_{2\mathrm{SLS}}-\beta]\approx\frac{\sigma_{\eta\xi}}{\sigma_{\xi}^{2}}\left[\frac{\mathrm{E}(\pi'Z'Z\pi)/Q}{\sigma_{\xi}^{2}}+1\right]^{-1}$$

(见附录中对这个结论的证明)。$(1/\sigma_{\xi}^{2})\mathrm{E}(\pi'Z'Z\pi)/Q$ 就是对第一阶段回归中所有回归元都显著进行检验的 F 统计量①。记这个统计量为 F，所以我们可以将其写为：

$$\mathrm{E}[\hat{\beta}_{2\mathrm{SLS}}-\beta]\approx\frac{\sigma_{\eta\xi}}{\sigma_{\xi}^{2}}\frac{1}{F+1}\tag{4.6.21}$$

由此可知随着第一阶段 F 统计值变小，2SLS 估计值的偏误会趋向于 $\sigma_{\eta\xi}/\sigma_{\xi}^{2}$。最小二乘估计值的偏误是 $\sigma_{\eta\xi}/\sigma_{x}^{2}$，当 $\pi=0$ 时会有 $\sigma_{\eta\xi}/\sigma_{\xi}^{2}=\sigma_{\eta\xi}/\sigma_{x}^{2}$。于是我们已经指出，当第一阶段拟合值为零时，2SLS 估计与最小二乘估计的渐进分布具有相同的期望值。更进一步，我们可以说当第一阶段不起什么作用时，2SLS 估计会偏向最小二乘估计。从另一方面来讲，如果 F 统计量变得很大，那么 2SLS 估计的有偏性将会消失，这正是大样本估计中发生的情况。

当工具变量很弱时，随着工具变量个数的增加，F 统计值会变小。为了看清楚这一点，考虑将没用的工具变量加入 2SLS，也就是说加入第一阶段的这些工具变量不会提高第一阶段回归的 R^{2}。于是随着 Q 的增加，模型的均方误及残差的方差 σ_{ξ}^{2} 都保持不变，但 F 统计值会变小。由此我们知道在模型中加入更多的弱工具变量会增加 2SLS 估计值的偏误。

从直觉上来看，2SLS 估计值存在偏误基于以下的一个事实：第一阶段是估计出的而不是现实中的数据直接给出的。如果已知第一阶段参数，我们可以使用 $\hat{x}_{pop}=Z\pi$ 来代替第一阶段拟合值。这些拟合值与第二阶段回归的残差不相关。在实际中，我们使用的拟合值是 $\hat{x}=P_{Z}x=Z\pi+P_{Z}\xi$，它与 $\hat{x}_{pop}$ 之间的区别表现在 $P_{Z}\xi$ 上。2SLS 估计值的有偏性就是因为 $P_{Z}\xi$ 与 η 之间存在相关性，因此第一阶段残差和第二阶段残差之间的相关性就通过 $\hat{\pi}'$ 的样本值进入了 2SLS 估计值。这种相关性会渐进为零，但是真实世界往往不会表现出这样的渐进性。

公式(4.6.21)指出，其他条件不变，2SLS 估计值的偏误是工具变量数量的一个增函数，所以在恰好识别这种可以保持工具变量最少的情况下，2SLS 估计值的偏误最少。事实上，恰好识别的 2SLS 估计值(也就是瓦尔德估计量)是渐进无偏的。但这一点很难正式地讨论，因为恰好识别的两阶段最小二估计的矩不存在(也就是说样本分布是重尾的，以至于样本矩发散，无法收敛到一个数)。而且，即使所使用的工具变量是弱工具变量，恰好识别的 2SLS 估计值也是以总体

① 实际上真实的 F 统计值是 $(1/\tilde{\sigma}_{\xi}^{2})\mathrm{E}(\tilde{\pi}'Z'Z\tilde{\pi})/Q$，该等式中将 σ_{ξ} 和 π“戴上帽子”表示对该变量的样本估计值。因此有时我们将 $(1/\sigma_{\xi}^{2})\mathrm{E}(\pi'Z'Z\pi)/Q$ 称为总体的 F 统计量，因为它是我们在无限大样本中得到的。在实际中，总体和样本的 F 统计值之间的差别并不明显。有些计量经济学家倾向于将第一阶段的 F 统计值乘以工具变量个数来考察工具变量的能力。这个乘积被称为是“集中度参数(concentration parameter)”。

均值为中心符合渐进分布的。因此我们称恰好识别的2SLS估计值是中位数无偏(median-unbiased)的。但是这并不是说你可以开心地在恰好识别的模型中使用工具变量。在恰好识别的模型中，弱工具变量得到的结果可能太不精确以至于无法使用。

在过度识别的常因果效应模型中，有限信息极大似然(limited information maximum likelihood，LIML)估计值是渐进中位数无偏的，因此在每次只能使用一个工具变量的恰好识别模型(比如见 Davidson 和 MacKinnon(1993)以及 Mariano(2001))之外，它还为我们提供了另外的估计方法。有限信息极大似然法的优点在于它得到的渐进分布与2SLS法(常因果效应之下的2SLS法)相同，但是可以降低有限样本中2SLS估计值的偏误。还有其他的一些估计方法可以减少过度识别的2SLS中存在的偏误。但是在 Flores-Lagunes(2007)的研究中，作者用一系列蒙特卡洛方法指出：在大部分情况下，有限信息极大似然法能够做到和别的方法一样好(从有偏性、误差绝对值的均值及 t 统计值的拒绝率三个方面进行了评价)。有限信息极大似然法的另外一个优点是大部分统计软件都可以计算这个数字，但是其他的一些方法则可能要自己来编程序。①

我们使用一个小型的蒙特卡洛实验来解释上面提到的一些理论结论。其中得到的仿真数据来自于下面的模型：

$$y_i = \beta x_i + \eta_i$$

$$x_i = \sum_{j=1}^{Q} \pi_j z_{ij} + \xi_i$$

其中，$\beta = 1$，$\pi_1 = 0.1$，$\pi_j = 0$，对于所有的 $j = 2, \cdots, Q$ 并且有：

$$\begin{pmatrix} \eta_i \\ \xi_i \end{pmatrix} \Big| Z \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.8 \\ 0.8 & 1 \end{pmatrix} \right]$$

其中，z_{ij} 是均值为0，方差为1的相互独立且正态分布的随机变量。在这个仿真过程中，我们假设存在一个真正起作用的工具变量和 $Q-1$ 个不起作用的工具变量。样本规模为1 000。

① 在 SAS 和 Stata10 中都可以进行有限信息极大似然估计。当存在弱工具变量时，有限信息极大似然估计法的标准误可能不是很准确，但是 Bekker(1994)给出了一个简单的解决方法。为什么有线信息极大似然法是无偏的？等式(4.6.21)指出2SLS估计值的偏误与相应最小二乘估计值偏误成比例。由此我们可以想到2SLS估计值和最小二乘估计值的线性组合可能是无偏的。有限信息极大似然法正好提供了这样一种“线性组合后的估计值”。与2SLS估计值的有偏性一样，用 Bekker 的分组渐进序列可以近似有限信息极大似然法的渐进无偏估计值。需要指出的是如果模型存在一定程度的异方差问题，有限信息极大似然法估计值是有偏的。其细节见 Bekker 和 van der Ploeg(2005)以及 Hausman 等(2008)。与有限信息极大似然估计法不同的是，无论是否存在异方差性，Jackknife 工具变量估计值(简写为 JIVE；比如见 Angrist，Imbens 和 Krueger(1999))都是 Bekker 无偏的。Ackerberg 和 Devereux(2007)最近提出了一种改进过的 Jackknife 工具变量估计值，这种方法下的方差更小。

图 4.4 报告了最小二乘法估计值、恰好识别的工具变量估计值(也就是说 $Q=1$ 的 2SLS 法,记为 IV,其第一阶段 F 统计量是 11.1)、存在两个工具变量的 2SLS 估计值($Q=2$,记为 2SLS,第一阶段 F 统计值为 6.0)以及 $Q=2$ 的有限信息极大似然法估计值的结果。从图上可以看出最小二乘估计值是有偏的,其均值大致在 1.79。工具变量估计值的均值是 1,正是我们要求解的 β 的均值。存在一个弱工具变量和一个无效的工具变量时,2SLS 估计值轻微地偏向最小二乘估计值(中位数是 1.07)。当 $Q=2$ 时,即使使用了无效的工具变量,有限信息极大似然估计值的分布函数还是与恰好识别的工具变量估计值的分布函数很接近。

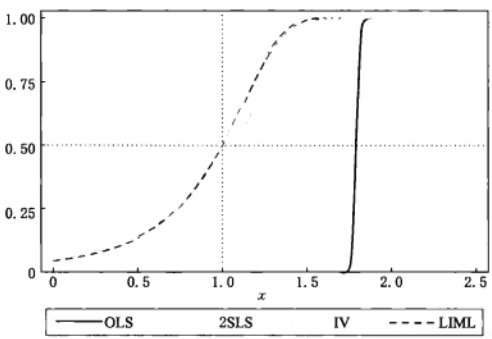


图 4.4 最小二乘、工具变量、2SLS 法和有限信息极大似然估计法得到的估计值的蒙特卡洛累积分布函数

图 4.5 报告了当 $Q=20$ 时的仿真结果。这里我们保留那个有用但弱的工具变量,同时加入 19 个无用的工具变量(第一阶段 F 统计值为 1.51)。该图同时给出了最小二乘、2SLS 以及有限信息极大似然法下估计值的分布。2SLS 估计值的有偏性变得更严重(中位数是 1.53,接近最小二乘估计值)。且 2SLS 估计值的样本分布要比 $Q=2$ 时更加收紧。有限信息极大似然估计值仍然表现良好且以 $\beta=1$ 为中心分布,但是比 $Q=2$ 时显得更加分散。

最后,图 4.6 报告了对无法识别模型所做的仿真结果。在该例子中,我们假设 $\pi_j=0;\ j=1,\cdots,20$(第一阶段的 $F=1.0$)。毫不惊讶,所有的样本分布都以最小二乘估计值的均值为中心。但是相比于有限信息极大似然估计值的分布,2SLS 估计值的分布显得更加紧凑。在这里我们可以将有限信息极大似然估计的这个性质看作是一种优点,因为估计值分布十分分散,正确地反映了:对我们感兴趣的变量而言这里的数据没有信息含量。

这些特点对我们在实际中进行的经验研究有什么意义呢?除了对第一阶段抱有一种不确定的担心之外,我们有以下建议:

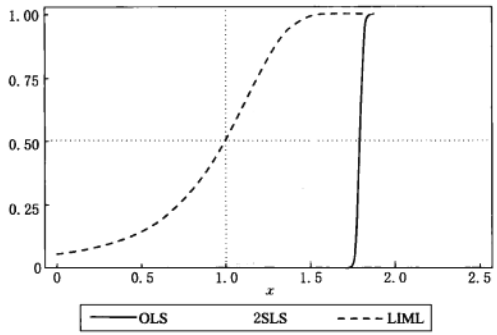


图 4.5　最小二乘、2SLS 法和有限信息极大似然估计法得到的估计值的蒙特卡洛累积分布函数(其中 $Q = 20$)

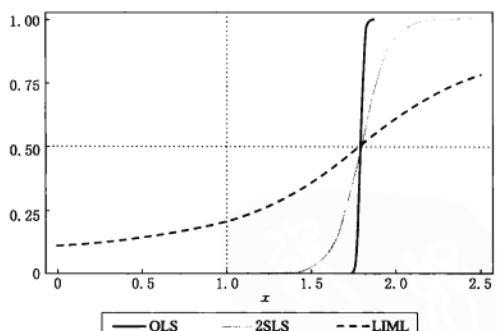


图 4.6　最小二乘、2SLS 法和有限信息极大似然估计法得到的估计值的蒙特卡洛累积分布函数(其中存在 $Q = 20$)

(1) 报告第一阶段回归结果并考虑相应估计值和拟合值是否有意义。估计值的大小和符号是不是符合你的预期，估计值是不是过大了，或者符号有问题？如果是这样，也许你假设的第一阶段并不存在，或者你只是撞大运了。

(2) 报告没有包括进回归的工具变量的 F 统计量。这个值越大越好。Stock，Wright 和 Yogo(2002)建议 F 统计量超过 10 的话你所做的回归就是安全的，显然，这只是一种经验结论，而非定理。

(3) 选择你觉得最好的那个工具变量并报告仅使用该工具变量得到的恰好识别估计值。恰好识别的工具变量总是中位数无偏的，因此不容易陷入弱工具变量的批评。

(4) 用有限信息极大似然估计法来检查过度识别的 2SLS 估计值。有限信息

极大似然估计值不如 2SLS 估计值精确，但相应的有偏性会小一些。如果两种方法得到的结果相似，那就应该高兴。如果不相似，那么要担心了，这时需要尝试找到更强的工具变量或者降低过度识别的问题。

(5) 在使用被解释变量关于工具变量做回归的简约式中考察工具变量的系数、t 统计值和 F 统计值。要记得，简约式中的估计结果是与我们感兴趣的因果效应成正比的，因为最小二乘估计是无偏的。正如 Angrist 和 Krueger(2001)注意到的，如果你在简约式中无法看到感兴趣的因果联系，那么很可能你错了。①

我们再次使用 Angrist 和 Krueger(1991)用出生季度做工具变量的研究中使用过的数据进行分析，以此说明上面的几条建议。Bound，Jaeger 和 Baker(1995)指出，即使样本大小超过 300 000(显然已经相对而言不是"小样本"了)，使用出生季度作为教育水平的工具变量时，最可能出现的问题是估计值有偏。在本章的前段，我们看到出生季度对教育水平的影响同样反映在简约式中，因此我们不该太过于鼓励估计值的有偏问题。而且，Bound，Jaeger 和 Baker(1995)还指出如果加入简约式中原本没有的控制变量，可能使得估计值偏误更加严重。表 4.9 再次给出了依据 Angrist 和 Krueger(1991)对回归方程的设定得到的估计值，同时也给出了依据 Bound，Jaeger 和 Baker(1995)对回归方程的设定得到的估计值。

表 4.9　用另外的工具变量法对教育的经济回报的估计

	(1)	(2)	(3)	(4)	(5)	(6)
2SLS	0.105	0.435	0.089	0.076	0.093	0.091
	(0.020)	(0.450)	(0.016)	(0.029)	(0.009)	(0.011)
LIML	0.106	0.539	0.093	0.081	0.106	0.110
	(0.020)	(0.627)	(0.018)	(0.041)	(0.012)	(0.015)
F-stata	32.27	0.42	4.91	1.61	2.58	1.97
控制变量						
出生年份	✓	✓	✓	✓	✓	✓
出生州					✓	✓
年龄和年龄的平方		✓		✓		✓
排他性工具变量						
出生季度虚拟变量	✓	✓				
出生季度 * 出生年份			✓	✓	✓	✓
出生季度 * 出生州					✓	✓
排他性工具变量个数	3	2	30	28	180	178

注：这个表比较了使用不同的工具变量和控制变量后得到的 2SLS 估计与有限信息极大似然估计的结果。年龄和年龄的平方项度量的是按季度计算的年龄。在第 1—4 列报告的相应模型的最小二乘估计结果是 0.071；在第 5 和 6 列报告的相应模型的最小二乘估计结果是 0.067。数据来自 Angrist 和 Krueger(1991)的研究中使用的 1980 年人口普查数据。样本规模是 329 509。标准误报告在相应参数下方的括号里。

① 最近由 Chernozhukov 和 Hansen(2008)完成的研究正式地研究了这个基本原则。

表 4.9 的第 1 列报告了使用三个出生季度虚拟变量作为工具变量进行的 2SLS 估计值和有限信息极大似然估计值，这两个估计值都用标志出生年份的虚拟变量做协变量。在这种模型设定下，最小二乘估计值为 0.071，2SLS 估计值高一点，为 0.105。第一阶段 F 统计量超过 32，很好地远离了危险域。很正常，这里的有限信息极大似然估计值和 2SLS 估计值很相似。

Angrist 和 Krueger(1991)尝试将年龄和年龄的平方项加入模型作为额外的控制变量，其中年龄用出生季度来衡量。加入这些控制变量的目的是去除因为遗漏年龄而带来的偏误以及对工具变量的干扰。将年龄和年龄的平方项加入模型使工具变量的个数减少到了两个，因为季度年龄、出生年份和出生季度之间是线性相关的。正如第 2 列所示，当将年龄和年龄的平方项加入模型后，第一阶段回归的 F 统计量降至 0.4，很明显，将这种控制变量加入回归是存在问题的。而且，2SLS 的标准误很高，以至于我们无法从估计值里得到任何有用的结论。有限信息极大似然估计值也很不精确。因此，我们没有识别出这个模型。

第 3 列和第 4 列报告了将出生季度和出生年份的交互项加入工具变量后得到的结果，因此这时已经有 30 个工具变量，如果把年龄和年龄的平方项作为控制变量加入模型，那么就是 28 个工具变量。这两个不同的模型设定下的第一阶段 F 统计量分别为 4.9 和 1.6。相比第一列的结果，2SLS 估计值要小一些，这意味着该估计值偏向于最小二乘的估计结果。但是有限信息极大似然估计值与 2SLS 估计值之间差距不大。尽管在第 4 列中我们发现有限信息极大似然估计下的标准误相当大，但并没有大到让我们一无所获的。总体而言，即使将年龄的二次项加入模型，我们也无需在具有 30 个工具变量的模型中担心弱工具变量的问题。

第 5 列和第 6 列对应的模型设定最不好。这两个模型在已有的出生季度和出生年份之间存在 30 个交互项的基础上，再将出生季度和出生州的 150 个交互项加入工具变量。将出生季度和出生州的交互项加入工具变量是为了捕捉不同州之间义务教育的不同实施水平。但是由于存在 180 个工具变量(或者 178 个)，其中大部分工具变量都是弱工具变量，所以这样做的结果是导致严重的过度识别问题。这两个模型的第一阶段回归的 F 统计量分别是 2.6 和 2.0。而且看上去有限信息极大似然估计值还是与 2SLS 估计值很相似。更进一步，有限信息极大似然估计的标准误没有比 2SLS 估计中的标准误大多少。这意味着你不能一直用一种很机械的方法(类似于“$F > 10$”)来判断工具变量是不是在发挥作用。在很多例子中，较低的 F 统计量并不是很致命。①

最后，值得注意的是当模型中存在多元内生变量时，传统的一阶段 F 统计量就不再可靠了。为了看清楚这一点，假设有两个工具变量可以来识别模型中的两个工具变量，其中一个工具变量强，可以很好地预测模型中的两个内生变量，但是另

① Cruz 和 Moreira(2005)也得到了相同的结论：即使在 Angrist 和 Krueger(1991)中包含了 180 个工具变量的回归得到的 F 统计量很低，但是其所执行的回归得到的参数几乎是无偏的。

一个工具变量则较弱。在这种情况下，两个内生变量对应的两个第一阶段的 F 统计值可能都比较高，但是由于有一个工具变量无法捕捉两个因果效应，所以模型的识别情况不好。在本章附录中我们给出在这种情况下正确的第一阶段 F 统计量。

4.7 附录

1. 等式(4.6.8)的推导

按照下面的方式重新写出等式(4.6.7)：

$$Y_{ij} = \mu + \pi_0 \tau_i + (\pi_0 + \pi_1)\bar{S}_j + \nu_{ij}$$

其中，$\tau_i \equiv s_i - \bar{S}_j$。既然根据构造 τ_j 和 $\bar{S}_j$ 不相关，那么我们有：

$$\rho_1 = \pi_0 + \pi_1$$

$$\pi_0 = \frac{\text{cov}(\tau_i, Y_{ij})}{V(\tau_i)}$$

展开第二个等式后有：

$$\begin{aligned}\pi_0 &= \frac{\text{cov}[(s_i - \bar{S}_j), Y_{ij}]}{[V(s_i) - v(\bar{S}_j)]} = \left[\frac{\text{cov}(S_i, Y_{ij})}{V(s_i)}\right]\left[\frac{V(s_i)}{V(s_i) - V(\bar{S}_j)}\right] \\ &\quad + \left[\frac{\text{cov}(\bar{S}_j, Y_{ij})}{V(\bar{S}_j)}\right]\left[\frac{-V(\bar{S}_j)}{V(s_i) - V(\bar{S}_j)}\right] \\ &= \rho_0\phi + \rho_1(1-\phi) = \rho_1 + \phi(\rho_0 - \rho_1)\end{aligned}$$

其中，$\phi \equiv \dfrac{V(s_j)}{V(s_i) - V(\bar{S}_j)}$ 是个正数。通过求解 π_1，我们还得到：

$$\pi_1 = \rho_1 - \pi_0 = \phi(\rho_1 - \rho_0)$$

2. 对两阶段最小二乘估计值的渐进偏误的推导

从等式(4.6.20)开始：

$$\text{E}[\hat{\beta}_{2SLS} - \beta] \approx [\text{E}(\pi' Z' Z \pi) + \text{E}(\xi' P_Z \xi)]^{-1} \text{E}(\xi' P_Z \eta)$$

用一些线性代数的知识，我们可以简化这个表达式：$\xi' P_Z \eta$ 是一个数，因此等于它的迹(trace)；迹函数是一个线性算子，可以穿过期望算子，并对交换运算是不变的；最后，P_Z 是一个幂等矩阵，因此它的迹等于它的秩 Q。使用这些事实，并对 Z 进行重复期望计算后，我们有：

$$\begin{aligned}\text{E}[(\xi' P_Z \xi \mid Z] &= \text{E}[\text{tr}(\xi' P_Z \eta) \mid Z] \\ &= \text{E}[\text{tr}(P_Z \eta \xi') \mid Z] \\ &= \text{tr}(P_Z \text{E}[\eta \xi' \mid Z]) \\ &= \text{tr}(P_Z \sigma_{\eta\xi} I) \\ &= \sigma_{\eta\xi} \text{tr}(P_Z) \\ &= \sigma_{\eta\xi} Q\end{aligned}$$

其中，我们假设 η_i 和 ξ_i 是同方差的。类似的，对 $\mathrm{E}[\xi' P_Z \xi]$ 使用上面提到幂等矩阵矩等于秩的技巧，我们立刻知道这一项等于 $\sigma_\xi^2 Q$。于是：

$$\begin{aligned}\mathrm{E}[\hat{\beta}_{2\mathrm{SLS}} - \beta] &\approx \sigma_{x\xi} Q[\mathrm{E}(\pi' Z' Z\pi) + \sigma_\xi^2 Q]^{-1} \\ &= \frac{\sigma_{x\xi}}{\sigma_\xi^2}\left[\frac{\mathrm{E}(\pi' Z' Z\pi)/Q}{\sigma_\xi^2} + 1\right]^{-1}\end{aligned}$$

3. 多元第一阶段 F 统计量

假设已经从工具变量中剔除了所有外生协变量带来的影响后，现在存在两个内生变量 x_1 和 x_2，其参数为 δ_1 和 δ_2。我们现在关心的是将 x_1 当做内生变量后用 2SLS 估计得到的 δ_2 的有偏程度是多少。易知第二阶段的等式是：

$$y = P_Z x_1 \delta_1 + P_Z x_2 \delta_2 + [\eta + (x_1 - P_Z x_1)\delta_1 + (x_2 - P_Z x_2)\delta_2] \quad (4.7.1)$$

其中，$P_Z x_1$ 和 $P_Z x_2$ 是来自于用 x_1 和 x_2 对 Z 做回归后得到的第一阶段拟合值。根据多元回归公式，方程(4.7.1)中的 δ_2 是对 y 关于来自回归 $P_Z x_1$ 和 $P_Z x_2$ 的残差进行回归得到的结果。这个残差是：

$$[I - P_Z x_1 (x_1' P_Z x_1)^{-1} x_1' P_Z] P_Z x_2 = M_{1x} P_Z x_2$$

其中，$M_{1x} = [I - P_Z x_1 (x_1' P_Z x_1)^{-1} x_1' P_Z]$ 是得到残差的矩阵。而且 $M_{1x} P_Z x_2 = P_Z [M_{1x} x_2]$。

由此我们可以总结出对 δ_2 进行的 2SLS 估计值是对 $P_Z[M_{1x}x_2]$ 进行最小二乘回归后得到的结果，换言之，就是用来自用 $M_{1x}x_2$ 对 Z 做回归后得到的拟合值进行最小二乘估计。这等同于在 2SLS 估计中使用 Z 做 $M_{1x}x_2$ 的工具变量。因此对 δ_2 的 2SLS 估计可以写为：

$$[x_2' M_{1x} P_Z M_{1x} x_2]^{-1} x_2' M_{1x} P_Z y = \delta_2 + [x_2' M_{1x} P_Z M_{1x} x_2]^{-1} x_2' M_{1x} P_Z \eta$$

决定针对 δ_2 进行的 2SLS 估计的有偏性是由第一阶段回归的平方和(F 统计量中的分母)决定的，也就是$[x_2' M_{1x} P_Z M_{1x} x_2]$的期望，到那时 2SLS 的有偏性来自于当 η 和 ξ 之间相关时期望 $\mathrm{E}[\xi' M_{1x} P_Z \eta]$是非零的。

这里是在实际操作中如何计算 F 统计量：(1)用另一个工具变量的第一阶段拟合值和所有的外生协变量对我们感兴趣的回归元 $P_Z x_2$ 进行回归。然后保存这一步得到的残差。(2)在第一阶段使用(1)中得到的残差进行回归并为工具变量构造 F 统计量。需要注意的是当你从用来自(1)的残差进行最小二乘估计并得到感兴趣的参数时，要用 Z 对来自(1)的残差进行工具，但是不要包含其他的协变量或者外生变量。用这一事实可以检查你的计算。

▶5

相似世界：固定效应、双重差分和面板数据

首要的事情是要认识到：平行的宇宙……并不平行。

Douglas Adams, *Mostly Harmless*

在第 3 章中，我们进行因果推断的关键是控制住干扰因果关系的可观察因素。如果对因果关系产生重大干扰的因素是观察不到的，那么我们可能要尝试使用第 4 章讨论的工具变量法。然而，良好的工具变量一般不易找到，所以需要发展一些别的手段来处理这类不可观察的干扰因素。本章探讨的主题是另一种使用控制变量来处理干扰因素的方法：考察数据在时间或者代际维度（cohort dimension）上的特点，用以控制不可观测但是固定的遗漏变量。这种估计策略总体上表现出的特点是：考虑个体在未受干预时表现出的趋势特征，然后在处理组和控制组中将这种趋势特征控制，最后比较两者的水平差异。在这一章我们还考察了两类方法，一类是用滞后被解释变量（lagged dependent variables）做控制变量，另一类是挖掘数据在时间维度上的不变因素。

5.1 个体固定效应

长久以来劳动经济学中就存在一个问题：工会成员的身份与工资之间存在何种联系。那些工资由集体议价确定的工人赚得比较多，但这种高工资仅仅是因为他们属于工会，还是由别的什么因素带来的？也许加入工会的工人本来就有更为丰富的经验和更为娴熟的技能。为了解决这一问题，令 Y_{it} 表示工人 i 在时间 t 所挣得的收入，令 D_{it} 表示其工会身份。于是可观察到的 Y_{it} 要么是 Y_{0it} 要么是 Y_{1it}，这里 Y_{1it} 和 Y_{0it} 分别指工会成员和非工会成员的工资。进一步假设：

$$\mathrm{E}[Y_{0it} \mid A_i, X_{it}, t, D_{it}] = \mathrm{E}[Y_{0it} \mid A_i, X_{it}, t]$$

其中，X_{it} 是由随时间变化的可观察的协变量构成的向量，A_i 是由不可观察但是固定的干扰因素构成的向量，我们将其称为能力。

换言之，上面这个等式表示的意思是：给定 A_i，给定类似于年龄、教育水平以及居住地这类可观察的协变量，工会成员身份“就像”随机分配的一样好。

固定效应估计法的关键假设在于：用线性模型表达 $E(Y_{0it}|A_i, X_{it}, t)$，$A_i$ 的下标中没有时间因素①：

$$E[Y_{0it} \mid A_i, X_{it}, t] = \alpha + \lambda_t + A_i'\rho + X_{it}\delta \tag{5.1.1}$$

我们同时假设工会成员身份带来的因果效应是不变且可加的：

$$E[Y_{1it} \mid A_i, X_{it}, t] = E[Y_{0it} \mid A_i, X_{it}, t] + \rho$$

将其与等式(5.1.1)结合起来，我们得到：

$$E[Y_{it} \mid A_i, X_{it}, t, D_{it}] = \alpha + \lambda_t + \rho D_{it} + A_i'\gamma + X_{it}'\beta \tag{5.1.2}$$

此处 ρ 是我们感兴趣的因果效应。相比于第 3 章讨论回归时施加的假设，等式(5.1.2)上附着的假设带来的限制更大；但是由于这一章中我们不再有工具变量可以使用，所以需要这种线性、可加的函数形式，以此讨论如何使用面板数据解决不可观察的干扰因素②。

方程(5.1.2)表明：

$$Y_{it} = \alpha_i + \lambda_t + \rho D_{it} + X_{it}'\beta + \epsilon_{it} \tag{5.1.3}$$

这里 $\epsilon_{it} \equiv Y_{0it} - E[Y_{0it} \mid A_i, X_{it}, t]$ 以及：

$$\alpha_i = \alpha + A_i'\gamma$$

这就是固定效应模型。给定面板数据(对个体进行重复观测后得到的数据)，通过将 α_i 看作一个需要估计的固定因素，我们通过回归就可以估计出工会身份对工资的因果效应。同样的，也可以将年份效应 λ_t 视为一个待估参数。通过设定表示不同个体的虚拟变量，通过设定表示不同年份的虚拟变量，不可观察的个体效应和年份效应就是相应虚拟变量前的系数。③

乍看起来，固定效应模型中的待估参数多到惊人。比如收入变化的面板调查(Panel Survey of Income Dynamics)是一个被广泛使用的面板数据集，该数据集中

---

① 言下之意是说 A_i 不随时间变化。——译者注

② 在有些情况下，我们可以模型中存在允许异质性的处理效应，也即：

$$E(Y_{1it} - Y_{0it} \mid A_i, X_{it}, t) = \rho_i$$

比如，Wooldridge(2005)就讨论了求解 ρ_i 平均值的估计量。

③ 与固定效应模型相对应的是随机效应(random effects)模型(例如可以参看 Wooldridge(2006))。随机效应模型假设 α_i 与回归元不相关。因为在随机效应模型中被遗漏的变量与包括进模型的回归元不相关，所以将这些变量遗漏不会导致估计值的偏误——实际上，这些被遗漏的变量成为模型残差(residual)的一部分。随机效应最为重要的结果是，对于给定的个人，他们在各期中的残差是相关的。我们在第 8 章会来讨论各期的残差相关对最小二乘估计带来的影响。如果随机效应模型的假设能够满足，那么用广义二乘法进行估计会更有效。不过正如在第 3 章中进行的讨论，相比于广义最小二乘法，我们更喜欢假设方差不变的最小二乘法。因为广义最小二乘法要求的假设比最小二乘法要强，但是用广义最小二乘法得到的估计效率的改进则比较微小，特别是其有限样本性质恐怕还不如最小二乘法。

大致包含 5 000 名处在工作年龄的工人近二十年的观察数据。因此使用这个数据集做固定效应回归，恐怕至少要估计 5 000 个固定效应。但是在实际中并不是这样的。从代数学的角度讲，将个体效应视为待估参数等同于估计个体对均值的偏离程度。换言之，我们首先估计个体均值：

$$\overline{Y}_i = \alpha_i + \overline{\lambda} + \rho\overline{D}_i + \overline{X}_i'\beta + \overline{\varepsilon}_i$$

然后在等式(5.1.3)的左右两端分别减去相应的均值，可得：

$$Y_{it} - \overline{Y}_i = \lambda_t - \overline{\lambda} + \rho(D_{it} - \overline{D}_i) + (X_{it} - \overline{X}_i)'\beta + (\varepsilon_{it} - \overline{\varepsilon}_i) \qquad (5.1.4)$$

可见，在用个体对均值的偏离程度来改写固定效应模型时，不可观察的个体效应消失了。①

与等式(5.1.4)相对应的另外一种方法是差分(differencing)，也即我们要估计的下面这个等式：

$$\Delta Y_{it} = \Delta\lambda_t + \rho\Delta D_{it} + \Delta X_{it}'\beta + \Delta\varepsilon_{it} \qquad (5.1.5)$$

这里符号 Δ 表示从一年到下一年的变化。例如，$\Delta Y_{it} = Y_{it} - Y_{it-1}$。如果模型考虑两期，那么从代数上看差分方程(5.1.5)和方程(5.1.4)是相同的，不过反过来看则不成立。这两种改写固定效应模型的方法都可拿来进行计算，但在同方差及 ε_{it} 不存在序列相关且考虑的时期大于两期时，方程(5.1.4)会更有效。当必须手动计算时，你会发现差分模型更为方便，需要注意的是由于差分模型的残差是序列相关的，所以在计算标准误时要进行调整。

有些回归软件包(regression packages)会自动给出方程(5.1.4)的估计值并计算出合适的标准误，因为它们在计算标准误时已经将估计 N 个个体均值时损失掉的自由度考虑进去。在残差同方差且序列不相关时，这么做就足以得到一个正确的标准误了。方程(5.1.4)的估计值有很多名字，包括“内估计量(within estimator)”和“协方差分析(analysis of covariance)”等。对方程(5.1.4)进行的估计过程则被叫做吸收固定效应。②

假设人们基于不可观测但是固定的个体特征选择是否加入工会，Freeman(1984)使用四个数据集估计了工会身份对工资的影响。表 5.1 报告了其研究中得到的一个估计值。对每个数据集，该表都报告了固定效应模型的估计值并同时报

① 为什么用个体对均值的偏离表示的固定效应模型等同于等式(5.1.3)所表示的最初的固定效应模型？因为根据解构回归公式(3.1.3)，我们可以分两步来计算任何多元回归的参数。为了得到一组多元变量的系数，首先用一个变量与模型中剩下的变量做回归，然后用被解释变量与第一步回归中得到的残差进行回归，就可得到该变量前的系数。在面板数据中，将表示所有个体和所有年份的虚拟变量放入回归得到的残差就是个体对均值的偏离程度。

② 在面板数据中，如果包含的时期数固定为 T 而截面上的个体数 $N \to \infty$，那么估计出的固定效应将不是一致的。这个问题叫做伴随参数问题(incidental parameters problem)，名字反映出了以下事实：随着样本规模的增长，待估参数的个数也在增加。尽管如此，对固定效应模型中的其他参数——我们关心的那些参数——的估计是一致的。

告了相应的截面估计值。相比于固定效应模型中的估计值(从 0.09 到 0.19)，截面估计值特别的高(从 0.14 到 0.28)。这可能意味着在截面模型中存在正的选择偏误，不过选择偏误不是固定效应模型估计值偏低的唯一原因。

表 5.1　估计出的工会身份对工资的影响

调　　查	截面估计值	固定效应估计值
May CPS，1974—1975	0.19	0.09
National Longitudinal Survey	0.28	0.19
Michigan PSID，1970—1979	0.23	0.14
QES，1973—1977	0.14	0.16

注：本表来自于 Freeman(1984)，报告了工会身份对工资影响的截面估计值和面板(固定效应)估计值。这些估计值都是用表最左边显示的调查数据得到的。在截面估计值中使用了人口特征和人力资本方面的控制变量。

尽管固定效应模型可以控制某一类遗漏变量，但我们还是强烈地怀疑固定效应估计值中存在由度量误差带来的微小偏误。从一方面来讲，诸如工会身份这样的经济变量一般倾向于长期稳定(一个工人今年是工会成员，下一年还是工会成员这是极为可能的)。从另一方面来讲，度量误差是每年都在变化着的(可能会在第一年误报工会身份或者对工会省份编码错误，但下一年报告正确)。因此，只要在任意年份中存在对工人工会身份的误报或者编码错误，那么我们观察到的工会身份在不同年份之间的变化就充满了噪音。换言之，相比于我们用水平值进行的回归，类似于方程(5.1.4)和方程(5.1.5)那样的差分方程中存在的度量误差会更加严重。这一事实可能造成固定效应估计值偏小。①

面板数据中的另外一类度量误差来自于如下事实：用来控制固定效应的差分法和偏离均值(方程(5.1.4))法将变量中好的变化和差的变化都处理掉了。换言之，这种转换一方面消除了一些遗漏变量偏误，但同时也丢掉了我们感兴趣变量的很多信息，这是一种“泼掉脏水同时扔掉孩子”的方法。用双胞胎估计教育水平对工资影响的研究就是这样一个例子。虽然这个问题没有考虑时间维度，但其基本思想和上面讨论过的工会身份对工资影响的例子相同：双胞胎拥有相似但基本上不可观察的家庭和基因背景。因此在一对双胞胎样本中放入家庭固定效应，我们就可以控制他们共同所有的家庭背景。

Ashenfelter 和 Krueger(1994)以及 Ashenfelter 和 Rouse(1998)用双胞胎做样本，在控制了家庭固定效应后估计了教育水平的经济回报。由于每个家庭都有两个双胞胎子女，所以该固定效应模型应该等同于用双胞胎之间教育水平差异去对双胞胎之间收入差距进行回归。但是令人惊讶的是在家庭内部估计出的教育的经济回报大于最小二乘估计值。但是，对于其他方面都类似的个体而言，教育水平如何会出现差异呢？Bound 和 Solon(1999)指出双胞胎之间存在一些小的差异，先

① Griliches 和 Hausman(1986)为面板数据中的度量误差提供了更为详尽的讨论。

出生的那个孩子一般有更高的体重和更高的 IQ 分数(这里出生时间是以分钟来度量的)。但是双胞胎之间的差别不是很大,他们在教育水平上的差别也不是很大。因此,双胞胎之间不可观测的能力差异可能造成了估计值中存在的相当一部分偏误。

在固定效应模型中,我们应该如何处理测量误差以及相关的问题呢?解决测量误差的一种可能的办法或许是使用工具变量。Ashenfelter 和 Krueger(1994)使用同一家庭兄弟姐妹报告的教育水平来构造针对双胞胎教育水平差异的工具变量。比如,双胞胎相互报告对方的教育水平,以此作为各自报告的教育水平的工具变量。解决测量误差的第二个方法是对度量误差使用外部信息,然后以此来调整估计值。在一项工会身份对工资影响的研究中,Card(1996)使用另外一个有效调查来调整面板数据估计值中存在的度量误差,这个度量误差是由人们报告工会身份时带来的。但是在 Ashenfelter 和 Rouse(1994)以及 Card(1996)中使用的多次报告和重复度量的数据不是很常见。因此,在解释固定效应估计值时我们要知道应该避免做出过于强的结论(在任何情况下,这都不是一个给应用计量学家的坏建议)。

5.2 双重差分:事前与事后,处理和控制

固定效应估计策略需要面板数据,也就是说需要对同样的个体(或者说公司以及任何可能的观察单位)进行重复观察。然而,经常出现的情况是我们关心的回归元只在更为加总水平上发生变动,比方说州或者某一代人这种范围的群体。例如,与怀孕工人有关的州一级层面上出台的政策可能会随着时间发生变化,但是对于州内的所有工人而言,该政策是一致和固定的。因此在考察这一类问题时,遗漏变量偏误可能主要是在州或者年份这个层面上出现的。在某些例子中,我们可以用群体层面的固定效应来解决这种遗漏变量偏误,这就引出了我们在这里需要讨论的有关双重差分(difference-in-difference,简称 DD)的识别策略。

或许双重差分的思想最早是由物理学家 John Snow(1855)提出的,当时他在研究 19 世纪中期伦敦市的霍乱传染问题。Snow 希望指出霍乱是由受污染的水传染而来的(这个理论与当时流行的"糟糕的空气"相左)。为了证明自己的观点,Snow 比较了由水厂 Southwark & Vauxhal 以及水厂 Lambeth 供水的地区的霍乱死亡率变化。在 1849 年,这两个水厂都从伦敦中部卫生较差的 Thames 地区汲取供给家庭的用水。但是到了 1852 年,水厂 Lambeth 将工厂迁往了上游较少受到下水道污染的地区来汲取水源。于是相比于由水厂 Southwark & Vauxhal 供水的地区,由水厂 Lambeth 供水的地区的霍乱死亡率剧降。

为了更具体地说明问题,我们现在回到经济学中的例子来。假设我们现在关心最低工资对就业的影响作用,这是劳动经济学的一个经典问题。在一个完全竞争的劳动市场上,最低工资的提高上移了向右下方倾斜的需求曲线。因此,最低工资越高,就越多地削减了就业量,可能还会伤害到政府实施最低工资政策原本打算救助的那些工人。Card 和 Krueger(1994)利用新泽西州最低工资上发生的显著变

化来辨别这一断言到底是不是正确的。①

在 1992 年 4 月 1 日，新泽西州将州最低工资从 4.15 美元每小时提高到 5.05 美元每小时。Card 和 Krueger 在 1992 年 2 月收集了新泽西州快餐店就业情况的数据，又在该年 11 月份再次收集了一次。这些快餐店(Burger King，Wendy's 等)都雇用了庞大的最低工资工人群体。然后他们穿过特拉华河，在该河对岸的宾夕法尼亚州东部同样类型的快餐店里收集数据。在同一时期，宾夕法尼亚的最低工资一直维持在 4.25 美元每小时。利用收集到的数据集，他们运用双重差分的方法计算了新泽西州最低工资提高后带来的效应。也就是说，他们比较了在新泽西州提高其最低工资这一段时间内，新泽西州就业量的变化和宾夕法尼亚州的就业量的变化。

双重差分法也是一种固定效应估计法，只不过这里使用的是加总的数据。为了看清楚这一方法，令 Y_{1ist} 表示如果面对的是较高的州最低工资，时期 t 中在州 s 的快餐店 i 里工人的就业人数，令 Y_{0ist} 表示如果面对的是较低的州最低工资，时期 t 中在州 s 的快餐店 i 里工人的就业人数。上面这两个变量都是潜在结果，在实际中我们只能看到其中的一个。比如，我们在新泽西州 1992 年 11 月份只能观察到 Y_{1ist}。双重差分法的核心在于假设没有受到处理的那个州的潜在结果可以写成两部分相加的形式。具体而言，假设：

$$\mathrm{E}(Y_{0ist} \mid s, t) = \gamma_s + \lambda_t \tag{5.2.1}$$

这里 s 代表州(新泽西州或宾夕法尼亚州)，t 表示时期(最低工资提高前的 2 月份，或最低工资提高后的 11 月份)。等式(5.2.1)是在说：如果没有最低工资变化这一事件，快餐店里的就业人数是由两部分因素决定的，一部分是不随时间变化的州效应，也即 γ_s，另一部分是对两个州都相同的年份效应 λ_t。这种以相加的方式出现的州效应就是我们在 5.1 节讨论过的不可观察的个体特征。

令 D_{st} 是虚拟变量，表示第 t 期实施较高的最低工资的那个州。假定 $\mathrm{E}(Y_{1ist} - Y_{0ist} \mid s, t)$ 是个常数，用 δ 表示，可观察到的就业人数 Y_{ist} 可以表示为：

$$Y_{ist} = \gamma_s + \lambda_t + \delta D_{st} + \varepsilon_{ist} \tag{5.2.2}$$

此处 $\mathrm{E}(\varepsilon_{ist} \mid s, t) = 0$，由此可得：

$$\begin{aligned}&\mathrm{E}[Y_{ist} \mid s = PA, t = Nov] - \mathrm{E}(Y_{ist} \mid s = PA, t = Feb)\\ &\quad = \lambda_{Nov} - \lambda_{Feb}\end{aligned}$$

以及：

$$\begin{aligned}&\mathrm{E}(Y_{ist} \mid s = NJ, t = Nov) - \mathrm{E}(Y_{ist} \mid s = NJ, t = Feb)\\ &\quad = \lambda_{Nov} - \lambda_{Feb} + \delta\end{aligned}$$

这里 PA 表示宾夕法尼亚州，NJ 表示新泽西州，Nov 表示 11 月份，Feb 表示 2 月份。总的差分就是：

---

① 最早用双重差分法的思想研究最低工资影响的研究是 Obenauer 和 von der Nienburg(1915)的研究，他们当时为联邦劳工部编写统计数据。

$$\{E(Y_{ist} \mid s = NJ, t = Nov) - E(Y_{ist} \mid s = NJ, t = Feb)\}$$
$$-\{E(Y_{ist} \mid s = PA, t = Nov) - E(Y_{ist} \mid s = PA, t = Feb)\} = \delta$$

就是我们所关心的因果效应。使用总体的样本值，我们可以轻松将此因果效应估计出来。

表 5.2　在新泽西州最低工资上升前后快餐店的平均雇员数

变　　量	宾夕法尼亚州(i)	新泽西州(ii)	差分，NJ－PA(iii)
1. 最低工资上升前的全职雇员，使用所有可用的观察值	23.33 (1.35)	20.44 (0.51)	−2.89 (1.44)
2. 最低工资上升后的全职雇员，使用所有可用的观察值	21.17 (0.94)	21.03 (0.52)	−0.14 (1.07)
3. 全职雇员平均数量的变化	−2.16 (1.25)	0.59 (0.54)	2.76 (1.36)

注：这个表来自于 Card 和 Krueger(1994)中的表 3。本表报告了在新泽西州最低工资上升前后，宾夕法尼亚州和新泽西州快餐店的平均雇员数。样本包括了在雇员数方面有数据记录的所有快餐店。将六个关闭的快餐店雇员数记为零，将四个暂时关闭的快餐店雇员数记为缺失。估计得到的标准误报告在括号里。

表 5.2(基于 Card 和 Krueger(1994)中表 3 而来)报告了在新泽西州最低工资变化前后，新泽西州和宾夕法尼亚州快餐店的平均就业人数。在头两行两列有四个方格，各自表示两期中两个州中快餐店的平均就业人数，剩下两个边上共有五组数据，其中一个边上的前两个数据分别表示每一期在两个州之间平均就业人数的差别，另一个边上的前两个数据表示每个州在最低工资上升前后平均就业人数的差别，最后一个数据则是双重差分估计值。在 1992 年 2 月份，宾夕法尼亚州快餐店就业人数高于新泽西州，但是在 11 月份却下降了。与此形成对照的是，新泽西州的就业人数略有上升。这两个变化产生了一个正的双重差分估计值，如果较高的最低工资推动企业上移其劳动需求曲线的话，我们应该期待的是相反的情况才对。

这一证据在多大程度上可以印证标准的劳动需求理论呢？这里可以进行验证的关键性假设是：如果新泽西州没有受到处理(即没有提高最低工资)，那么这两个州中的就业趋势应该是相同的。如图 5.1 所示，提高最低工资使新泽西州的就业趋势偏离了原有的趋势。虽然接受处理的那个州和作为控制的那个州本身就有不同，但是这种不同表现在相应州的固定效应中，这个固定效应发挥的作用和等式(5.1.3)①中不可观察的个体特征是一样的。

① 这种共同趋势假设可以应用到转换后的数据，比如：

$$E[\ln Y_{ist} \mid s, t] = \gamma_s + \lambda_t$$

然而需要注意的是，以对数形式表现的共同趋势排除了水平上的趋势(trends in levels)，反之亦然。Athey 和 Imbens(2006)引入了一种半参数的双重差分估计值，允许在未指明对被解释变量做何种函数变换的情况下存在共同趋势。Poterba，Venti 和 Wise(1995)与 Meyer，Viscusi 和 Durbin (1995)考察了可以应用于分位数的双重差分法。

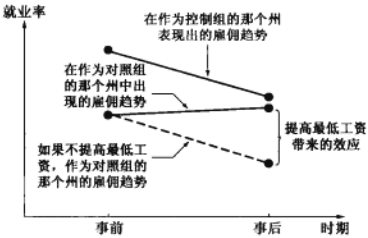
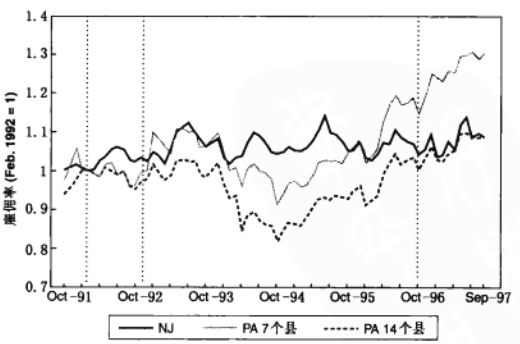


图 5.1　双重差分模型中的因果效应

可以利用多个时期的数据我们还能对共同趋势假设进行检验。作为最低工资的后续研究，Card 和 Krueger(2000)获得了来自官方的新泽西州和宾夕法尼亚州若干县的快餐店薪水数据。这些数据展示在图 5.2 中，该图与 Card 和 Krueger (2000)中的图 2 类似。第一条垂线表示 Card 和 Krueger 最初进行调查的时间，第三条垂线表示 1996 年 10 月联邦最低工资提升至 4.75 美元/每小时，这次最低工资变化影响了宾夕法尼亚州，但没有影响新泽西州。这些数据给我们提供了一个观察新的最低工资"实验"的机会。



注：横轴代表 Card 和 Krueger(1994)进行调研的时间以及 1996 年 10 月联邦最低工资法的调整。

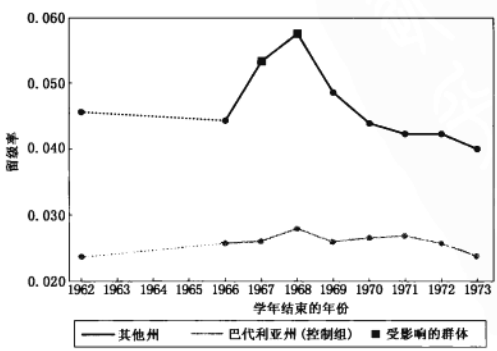
图 5.2　在新泽西州和宾夕法尼亚州快餐店雇员数的情况，1991 年 10 月到 1997 年 9 月

与 Card 和 Krueger 在原初调查中反映出的结论相同，这些官方数据反映出从 1992 年 2 月到 11 月，宾夕法尼亚州就业人数略有下降，同一时期新泽西州略有变化。同时这些数据还反映出在不同的年份两个州里的快餐店就业人数存在相当的变化，而且这些变化还很不一致。具体而言，在 1991 年年末，两个州快餐店就

业人数相似，在之后的三年中宾夕法尼亚州（特别是在由 14 个县组成的那一组中）快餐店的就业人数下降幅度超过了新泽西州，这些下降都发生在联邦提高最低工资之前。因此，宾夕法尼亚州快餐店就业水平不是度量反事实情况下新泽西州快餐店就业水平的良好指标，这里的反事实情况是指假设新泽西州未发生最低工资变化。

一个鼓舞人心的例子来自 Pischke(2007)，他利用德国在政策上的急剧变化所产生的差异来考察学校学期长度对学生成绩的影响。直到 20 世纪 60 年代，除了巴伐利亚州，德国各州的孩子们都是在春季开学。从 1966—1967 学年开始，春季开学被改为秋季开学。对于受到影响的学生而言，学校变成秋季开学意味着一年有两个短学年，学年长度不再是 37 周，而是变为 24 周。相比于其他未受影响的学生，同时也相比于一直从秋季开学的巴伐利亚州学生，受到影响的那些学生的在校时间被压缩了。

图 5.3 绘制了 1962—1973 年（在 1962—1965 年的数据缺失）巴伐利亚州和受影响地区的二年级学生的留级概率。从 1966 年开始，巴伐利亚州的学生留级率保持在比较合理的 2.5%左右。在改变学期长度的政策发生之前的 1962—1966 年，受到短学年影响的那些地区的学生留级率高一些，维持在 4%—4.5%左右。但是对于在巴伐利亚州之外的那些州受到影响的学生而言，留级率发生了跳跃，上升大约一个百分点，在回落到基本水平之前还是有点高。这幅图为我们提供了一个很强的直观印象：作为处理组的州和作为控制组的州，其潜在趋势是一样的，一个处理引起了处理组剧烈但是短暂的反应，该反应表现为对潜在趋势的偏离。对于受到政策变化影响的学生而言，短学期似乎确实提高了他们的留级比率。



注：该图来自 Pischke(2007)，数据时间跨度为巴伐利亚以外的学生改变学期长度前后。

图 5.3　在德国，分别处于处理组和对照组的那些学校中，二年级学生的平均留级率

5.2.1 双重差分回归

就像固定效应模型一样，我们也能使用回归去估计类似方程(5.2.2)表示的双重差分模型。令 NJ_s 为一个虚拟变量，用以表示处在新泽西州的餐馆，令 d_t 为表示时间的虚拟变量，当观察值来自于11月份(也就是在最低工资发生变化后)的那次调查时取值为1，代表一个时间哑变量(time-dummy)，于是：

$$Y_{ist} = \alpha + \gamma NJ_s + \lambda d_t + \delta(NJ_s \cdot d_t) + \epsilon_{ist} \tag{5.2.3}$$

与方程(5.2.2)是完全一样的，此处 $NJ_s \cdot d_t = D_{st}$。如果使用3.1.4节中讨论饱和模型时用到的语言，那么方程(5.2.3)包含表示地点(州)和表示时期(年)的两个主效应以及表示新泽西州11月份快餐店的一个交叉项。这是一个饱和模型，由于其条件期望函数 $E(Y_{ist}|s, t)$ 取四个值，所以有四个待估参数。回归方程(5.2.3)中的参数和双重差分模型中的条件期望函数之间的关系如下：

$$\begin{aligned}
\alpha &= E(Y_{ist} \mid s = PA, t = Feb) = \gamma_{PA} + \lambda_{Feb} \\
\gamma &= E(Y_{ist} \mid s = NJ, t = Feb) - E(Y_{ist} \mid s = PA, t = Feb) \\
&= \gamma_{NJ} - \gamma_{PA} \\
\lambda &= E(Y_{ist} \mid s = PA, t = Nov) - E(Y_{ist} \mid s = PA, t = Feb) \\
&= \lambda_{Nov} - \lambda_{Feb} \\
\delta &= \{E(Y_{ist} \mid s = NJ, t = Nov) - E(Y_{ist} \mid s = NJ, t = Feb)\} \\
&\quad - \{E(Y_{ist} \mid s = PA, t = Nov) - E(Y_{ist} \mid s = PA, t = Feb)\}
\end{aligned}$$

双重差分模型的回归公式为构造双重差分估计值和标准误提供了一种便利的方法。我们可以很容易地向该回归公式中添加更多的表示地点(州)和表示时期(年)的虚拟变量。比如，在研究新泽西州和宾夕法尼亚州最低工资的例子中，我们可以加入更多的州作为控制组，也可以收集最低工资提高前更多时期的情况加入样本。通过加入更多的表示州和表示时期的虚拟变量并保持其他部分不变，我们就能得到相应的更加一般化的方程(5.2.3)。

对双重差分模型进行回归的第二个好处在于：即使政策变化无法用虚拟变量来描述，我们也能对政策变化进行研究。比如，不同于对1992年新泽西州和宾夕法尼亚进行的那个研究，我们可能感兴趣于对美国所有州的最低工资进行研究。其中，可能有些州的最低工资只比联邦最低工资(联邦最低工资适用于所有人，不论他在哪个州生活)高一点，有一些州则高很多，还有一些州与联邦最低工资保持一致。因此，在不同州和不同时期里，表示最低工资的这个变量可能存在不同的处理强度。此外还要考虑到每个州最低工资在法律上的差异以及地区平均工资对最低工资产生的不同影响。举例来说，在20世纪90年代早期联邦最低工资为4.25美元/小时，可能就与康涅狄格州毫不相干，因为该州拥有更高的最低工资，但是对

于密西西比州来说该法案的影响大为不同。

Card(1992)探讨了联邦最低工资产生的影响中存在的地区差异。他的研究思路类似于下面这个方程：

$$Y_{ist} = \gamma_s + \lambda_t + \delta(FA_s \cdot d_t) + \varepsilon_{ist} \tag{5.2.4}$$

此处变量 FA_s 用以度量在每个州中最低工资提高可能影响到的青少年比例，d_t 是用来表示观察值来自 1990 年的虚拟变量，在那一年联邦最低工资从 3.35 美元/小时提高到 3.80 美元/小时。变量 FA_s 的主要用途是为每个州里每小时收入少于 3.80 美元的年轻劳动力比例设定一个基准。

就像对新泽西州和宾夕法尼亚进行的研究那样，Card(1992)仍然使用两期数据，在这里分别来自 1989 年和 1990 年，这两年分别处在联邦最低工资法变化之前和之后。但是不同之处在于该项研究使用了 51 个州(包括哥伦比亚特区)，总数为 102(州—年份)个观察值。由于等式(5.2.4)中不包含个体层面的协变量，所以该方程对应的估计与使用微观数据(假设群体层面的估计值是用个体规模加权平均过的)进行的估计相同。注意到 $FA_s \cdot d_t$ 是一个交互项，类似于方程(5.2.3)中的 $NJ_s \cdot d_t$，不过在每个州的观测数据中，这个交互项的取值都不同。最后，由于 Card(1992)分析的数据仅有两期，所以所报告的估计值来自一阶差分方程：

$$\Delta\bar{Y}_s = \lambda^* + \beta FA_s + \Delta\bar{\varepsilon}_s$$

此处 $\Delta\bar{Y}_s$ 是在 s 州里青少年平均就业量的变化量，$\Delta\bar{\varepsilon}_s$ 则是差分方程中的误差项。①

表 5.3 乃是基于 Card(1992)中的表 3 得来，它表明在那些最低工资的提高可能产生更大不良后果的州里(参看第 1 列中 0.15 的估计值)，工资提高更多。在 Card 的分析里这是很重要的一步——它验证了这样的看法：FA_s(受影响的青少年比例)的变化很好地预测了由联邦最低工资变化引起的工资变化。但是就业量似乎与 FA_s 基本无关，这一点可以从表 5.3 中的第 3 列看到。如此一来，Card(1992)中的结果就与对新泽西州和宾夕法尼亚州的研究结论相一致了。

表 5.3　最低工资对年轻人影响的双重差分回归估计值，1989—1990 年

解释变量	工作对数值的平均值的变化		年轻雇员在总人口中的比例	
	(1)	(2)	(3)	(4)
1. 受影响的青少年比例(FA_s)	0.15	0.14	0.02	−0.01
	(0.03)	(0.04)	(0.03)	(0.03)

① 另外的一些识别策略也沿袭了方程(5.2.4)的思路，但是它们用关于州和联邦最低工资的更为标准化的函数来代替 $FA_s \cdot d_t$。比如，可以参看 Neumark 和 Wascher(1992)，他们就使用了州与联邦最低工资之间的差别来研究，调整了最低工资覆盖条款，并且根据平均小时工资率进行了标准化。

（续表）

解释变量	工作对数值的平均值的变化		年轻雇员在总人口中的比例	
	(1)	(2)	(3)	(4)
2. 就业人数占总人口比重的变化	—	0.46 (0.60)	—	1.24 (0.60)
3. R^2	0.30	0.31	0.01	0.09

注：上表改编自 Card(1992)。此表报告了各州受最低工资影响的青少年就业量变化与各州最低工资变化之间的回归估计值。数据来源于 1989 年和 1992 年的 CPS。在回归中，用各州 CPS 样本规模进行了加权平均。

Card(1992)的分析进一步表明了用回归计算双重差分模型的好处：在这个框架中很容易加入更多的协变量。比如，成年人就业水平可能是在州这个层面上遗漏掉的某个趋势，我们可以将其加入回归模型进行控制。换言之，我们可以用下面这个方程来描述不存在最低工资变化的这种反事实的情况下各个州的就业状况：

$$E(Y_{0st} \mid s, t, X_{st}) = \gamma_s + \lambda_t + X'_{st}\beta$$

此处 X_{st} 是随着州与时间的变化而变化的协变量组成的向量，其中包括成年人就业量（如果成人就业量也会对最低工资变化作出反应，那么这样做可能不太合适，这时成年人就业量是个不合格的控制变量，参看 3.2.3 节）。正如表 5.3 中第 2 列和第 4 列所示，将成年人就业量加入控制变量几乎不影响 Card 的估计值。

值得强调的是 Card(1992)分析的是州这一层面上的平均值而非个体值。他应该也可以使用来自 CPS 的跨度为多个年份的混同(pooled)微观数据样本进行分析，去估计方程(5.2.5)：

$$Y_{ist} = \gamma_s + \lambda_t + \delta(FA_s \cdot d_t) + X'_{ist}\beta + \varepsilon_{ist} \qquad (5.2.5)$$

此处 X_{ist} 可以包括诸如种族这样个体层面的特征，也可以包含在州一级层面上度量出的随时间的变化而变化的变量。只有后者可能是一种遗漏变量偏误的来源，到那时个体层面的控制变量可以提高估计精度，这一点我们在 2.3 节业已提及。不过，在这样一个被解释变量来自于微观数据而回归元中包含群体层面数据的分析框架中，推断问题会显得略为复杂。在推断过程中最关键的问题是：如何最好地调整标准误，以应对群体层面可能存在的随机效应，这部分内容将在第 8 章进行讨论。

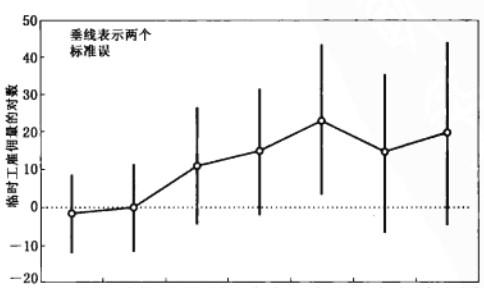
当样本中包含多年观测值时，借助于 Granger(1969)的思想，我们可以对双重差分的回归模型进行因果检验。Granger 的想法是：观察原因是否发生在结果之前，而不是相反（尽管在第 4 章开始部分的讽刺短诗里我们知道对于因果性推断来说这还不够）。假定我们感兴趣的政策变量是 D_{st}，该变量在不同州、不同时点上会发生变化。在这种情况下，Granger 因果性检验意味着对下面的问题做出检验：给定州和年份效应，是否过去的 D_{st} 可以预测 Y_{st} 而未来的 D_{st} 无法预测 Y_{st}。如果 D_{st} 引起了 Y_{st} 的变化而不是 Y_{st} 引起了 D_{st} 的变化，那么表征未来政策变化的虚拟变量的变化不应该在方程(5.2.6)中产生影响：

$$Y_{st} = \gamma_s + \lambda_t + \sum_{\tau=0}^{m} \delta_{-\tau} D_{s,t-\tau} + \sum_{\tau=1}^{q} \delta_{+\tau} D_{s,t+\tau} + X'_{st}\beta + \epsilon_{st} \qquad (5.2.6)$$

在等式右边的求和符号中允许存在 m 阶的滞后效应（δ_{-1}，δ_{-2}，…，δ_{-m}）以及 q 阶的提前效应或者说预想效应（δ_{+1}，δ_{+2}，…，δ_{+q}）。滞后效应表现出的特点往往也是我们的兴趣所在。比如，我们可能会觉得随着时间的推移，因果效应会减弱或者加强。

Autor(2003)在一项考察雇佣保护措施如何对企业使用临时工造成影响的研究中使用了 Granger 因果检验。在美国，雇佣保护措施是劳动法的一种，由州立法机关公布，或者更为典型的是通过州法院的普通法而设立。该法使得企业解雇工人变得更为困难。作为一项规则，美国劳动法允许随意雇佣，这意味着只要雇主乐意，他可以以一定的理由解雇工人，也可以无理由解雇工人。但是在某些州，法院允许不满足随意雇佣原则的例外存在，这就导致雇员可以用不正当解雇来起诉其雇主。Autor 关心的问题就是：是否因为企业害怕和雇员打官司，所以他们倾向于雇用更多的临时工，而临时工做的工作本来应该是正式雇员干的。因为临时工可以由企业之外的某些人雇用，只要完成既定的任务即可。所以如果企业让这些临时工走人，他们无法用不正当解雇的理由来起诉企业。

Autor 的经验研究策略将一个州里的临时工就业量作为被解释变量，用虚拟变量来表示该州法庭是否允许对随意雇佣的条例作出例外修改，然后在双重差分模型中考察两者之间的关系。类似于方程(5.2.6)，他的双重差分回归模型中既包含了该虚拟变量的滞后期，又包含了提前期。在估计中使用了虚拟变量的两期提前和四期滞后，估计得到的结果见图 5.4，该图复制自 Autor(2003)的图 3。这些估计值显示出在法庭允许对随意雇佣条例作出例外修改之前的两年，工具变量对临



注：其中被解释变量是 1979—1995 年之间州临时工就业量的对数值。在计算该估计值时，模型中既包含了州法院允许存在例外之前的效应，也包含了州法院正在允许存在例外和已存在例外的效应。

图 5.4　州法院允许对随意雇佣存在例外对临时工就业量的影响

时工就业量没有影响，但是在法庭允许对随意雇佣条例作出例外修改后的头几年里，工具变量对临时工就业量产生了剧烈的影响，之后这个影响趋于平坦但是临时工就业量始终维持在较高水平。这一模式看起来与 Autor 对其结果赋予的因果解释是一致的。

另外一个对双重差分识别策略进行检验的方法是在控制变量中加入与每个州相联系的时间趋势项。换言之，我们估计：

$$Y_{ist} = \gamma_{0s} + \gamma_{1s}t + \lambda_t + \delta D_{st} + X'_{ist}\beta + \epsilon_{ist} \tag{5.2.7}$$

与之前的设定相同，这里 γ_{0s} 是对每个州求出的截距项，γ_{1s} 是对每个州求出的时间趋势系数，在回归方程中这个系数是与时间趋势变量 t 相乘的。这样设定回归方程的好处是它允许处在处理组的州和处在控制组的州沿着不同的趋势发展，这种设定方法虽然有局限性但已经很吸引人了。如果将这些趋势控制后我们感兴趣的效应没有改变，那么这种回归就颇为令人兴奋，否则回归估计值就比较让人沮丧了。但是要注意到当模型中加入每个州的时间趋势后，我们至少需要三期数据来估计这些参数。但是在实际中三期往往无法让我们控制该趋势，也无法计算出处理效应。作为一种法则，当个体受到处理之前的数据可以清晰地反映出一种趋势，而且这种趋势可被外推到个体接受处理之后的时期中，那么存在州时间趋势项的双重差分模型估计值会更加稳健，更令人信服。

在对印第安纳州劳动管制对商业的影响的研究中，Besley 和 Burgess(2004)使用州趋势项做了一个稳健性检验。由于不同州在不同时间上改变了管制方法，这就自然地要求将双重差分法作为研究设计。正如 Card(1992)中一样，在 Besley 和 Burgess(2004)里观察的单位是每个州在每一年的平均值。表 5.4(基于他们论文中的表 IV)复制了那些关键结果。

表 5.4　在印度的各个州中估计出的劳动力管制对企业绩效的影响

	(1)	(2)	(3)	(4)
劳动力管制(滞后)	−0.186	−0.185	−0.104	0.000 2
	(0.064)	(0.051)	(0.039)	(0.020)
log(个人发展上的人均花费)		0.240	0.184	0.241
		(0.128)	(0.119)	(0.106)
log(人均已建成电力设备数)		0.089	0.082	0.023
		(0.061)	(0.054)	(0.033)
log(州人口)		0.720	0.310	−1.419
		(0.96)	(1.192)	(2.326)
国会中的多数席位			−0.000 9	0.020
			(0.01)	(0.010)
极左势力在国会中的席位			−0.050	−0.007
			(0.017)	(0.009)
Janata 政党在国会中的席位			0.008	−0.020
			(0.026)	(0.033)

（续表）

	(1)	(2)	(3)	(4)
地区代表在国会中的席位			0.006 (0.009)	0.026 (0.023)
州趋势项	无	无	无	无
R^2	0.93	0.93	0.94	0.95

注：此表改编自 Besley 和 Burgess(2004)中的表 Ⅳ。该表报告了用双重差分的回归模型估计出的劳动力管制对企业生产率产生的影响。被解释变量是企业人均产出的对数值。所有的模型都包含了州和年份的固定效应。括号中报告的是相应的稳健标准误。在推出劳动管制措施的整个期间，按照各州修改争端法时持有的态度，论文对采取支持工人态度的州赋值为 1，对支持雇主的州赋值为 −1，态度中立的州赋值为 0。log(人均已建成电力设备数)用人均千瓦时的对数值表示，log(个人发展上的人均花费)的取值等于每个州人均社会和经济服务支出的对数值。国会、极左、Janata 以及地区代表用这些政治势力在相应州占主要席位的时间来表示。样本中包含的数据来自于 1958—1992 年印度 16 个主要州，共计 552 个观测值。

第 1 列估计值对应的双重差分回归模型中没有包括州趋势项，得到的估计值指出劳动力管制降低了人均产出。用来计算第 2 列和第 3 列估计值的模型中加入了随时间变化的州趋势项作为协变量，这类协变量主要由人均政府支出和州人口组成。这种处理方法正是 Card(1992)中将州一级成年人就业率加入模型时所基于的考虑。从表中第 4 列可知，在模型中加入更多的控制变量几乎没有改变 Besley 和 Burgess 的估计结果。显然，产量下降的州里劳动管制水平得到了提高。因此将这种趋势控制后估计出来的效应降至零。

1. 挑选控制变量

我们业已在双重差分的框架中标识出来两个维度："州(states)"和"时间(time)"，因为这两个维度在应用计量经济学的双重差分方法里最具代表性。但是双重差分的思想实际上更加一般。即便不代表州，下标 s 也可以用来标识任何人口统计意义上的组；其中一些组受到政策影响，而另外一些组不受政策影响。比如 Kugler、Jimeno 和 Hernanz(2005)考察了西班牙与年龄有关的就业保护政策的影响。同样，即使 t 不代表时间，我们也可以用其代表基于出生年份或者不同个体特征进行分类的群体数据。Angrist 和 Evans(1999)就是这样的一个例子，他们研究了州堕胎法的变化对青少年怀孕的影响，使用了州和出生年份这样的变量。先不考虑对不同群体起什么名字，双重差分的实验设计经常都隐含地比较着处理组—控制组之间的差别。不过这种比较是不是一个好的比较则还有待于仔细的考量。

双重差分法可能存在一个缺点：随着处理结果的不同，控制组和比较组的组成人员可能发生着变化。我们回到基于地点和时间进行比较的实验设计，假定我们感兴趣的是公共救助对劳动力供给的影响。从历史上看，全美各州为贫穷的未婚妈妈提供的福利存在很大的差异。劳动经济学家们一直以来都对这类收入支持政策的效果很感兴趣：这些政策如何提高了她们的生活水平，是不是让这些未婚妈妈变得不愿工作了？(例如，可以参看 Meyer 和 Rosenbaum(2001)最近的一个研

究。)这里我们关心的问题——Moffitt(1992)在其有关福利研究的综述中也重点强调的问题，就是无论怎样都会成为就业市场上弱势群体的穷人会不会移居到福利更为优厚的州？在双重差分的实验设计中，这类由政策诱致的移民会使优厚福利对劳动力供给产生的效应变差。

如果我们知道某个人是从哪里迁移过来的，那么移民问题通常就可以迎刃而解。也就是说我们需要知道接受处理前的某个时期居民的居住州或者出生所在州。虽然某人出生所在州或者先前居住州不受处理的影响，但是却与该个体当前居住所在州高度相关。因此，如果用出生地或者先前居住州等维度进行比较，而不是用现在居住地进行比较，我们就可以解决移民问题。但是这又引出一个新的问题：那些确实选择移居的人可能选择了不正确的居住地。不过在实际中我们通过第 4 章讨论的工具变量法(可以用出生地所在州或者之前居住地来构造当前居住地的工具变量)就可以解决这个问题。

当控制组的情况很好时，我们可以对双重差分法进行变型，用更高阶的比较来推断因果效应。Yelowitz(1995)对全美医疗补助的覆盖范围扩大后带来的影响所做的研究就是这样一个例子。医疗补助是美国为穷人设置的一项大规模的医疗保险制度，AFDC 是美国实行的一项大的现金补助计划，政府曾经将符合医疗补助的资格与接受 AFDC 救助的资格联系在一起。但是，在 20 世纪 80 年代的不同时期，一些州将医疗补助资格扩大到了不符合 AFDC 救助资格的家庭中的孩子。Yelowitz 感兴趣的问题是：这种由政府提供的健康保险的受惠范围扩大至那些孩子后，对其母亲的劳动参与状态和收入产生了何种影响。

除了州与时间之外，由于不同年龄的孩子适用的医疗补助政策不同，所以孩子们的年龄为该项研究提供了第三个维度。Yelowitz 通过估计下式来探讨这一变化：

$$Y_{iast} = \gamma_{st} + \lambda_{at} + \theta_{as} + \delta D_{ast} + X'_{ast}\beta + \epsilon_{ast}$$

这里 s 表示州，t 表示时间，a 表示该家庭中最小孩子的年龄。这个模型为州时间趋势提供了完全的非参数控制，这种州时间趋势在不同年龄组(γ_{st})、随时间变化的年龄效应(λ_{at})以及州的年龄趋势(θ_{as})中产生的影响都是相同的。我们感兴趣的回归元 D_{ast} 表示在医疗救助覆盖的相应州和相应时期孩子处在接受医疗补助那个年龄段的家庭。相比于只考虑州和时间上存在的差别的双重差分模型，这里的三重差分模型产生的结果更加令人信服。

5.3 固定效应与滞后被解释变量

固定效应和双重差分估计值基于的假设是不随时间变化(或者在群体中不变)的遗漏变量。比如假设我们对参加受补贴的培训带来的效应很感兴趣，这个问题在 Dehejia 和 Wahba(1999)以及 Lalonde(1986)的研究中讨论过，也曾在本书 3.3.3 节出现过。在这个例子中，可以被验证的使固定效应估计可行的假设是：

$$E(Y_{0it} \mid \alpha_i, X_{it}, D_{it}) = E(Y_{0it} \mid \alpha_i, X_{it}) \tag{5.3.1}$$

此处 α_i 是一个不可观察的个人特征，该特征与协变量 X_{it} 一起决定了个体 i 是否接受培训。具体而言，α_i 可以是对职业技能的度量，但是与固定效应模型设定不相容的一个事实是：我们对不可观察变量的特点一无所知。在任何例子中，只要为 $E(Y_{0it} \mid \alpha_i, X_{it})$ 设定一个线性模型，那么假设(5.3.1)就可以为我们提供简单的估计策略，这些策略包括差分以及求与均值的偏离。

对于很多因果性问题而言，最为重要的遗漏变量不随时间变化这一看法似乎不是很有道理。对培训项目产生的效果的评估就是一个典型案例。看上去那些通过参与政府资助的培训项目以寻求改进自己在劳动力市场上的选择机会的人，他们可能在劳动力市场上遭受过某些挫折。很多培训项目明确地将目标人群定位为那些近期在劳动力市场上受到挫折的人，比如最近失去工作的人。与上面的这个看法相同，Ashenfelter(1978)以及 Ashenfelter 和 Card(1985)发现，查阅培训项目参与者的历史收入，往往发现参加项目前这些人的收入在下降。因此过去的收入不是时间不变的，因此不可以包含进入一个时间不变的遗漏变量 α_i。

受培训的个体在历史收入上表现出的差别激发了研究者使用一种新的估计策略，该策略直接控制过去的收入并省略固定效应。精确起见，与方程(5.3.1)不同，我们可能基于条件独立假设来作出因果推断：

$$E(Y_{0it} \mid Y_{it-h}, X_{it}, D_{it}) = E(Y_{0it} \mid Y_{it-h}, X_{it}) \tag{5.3.2}$$

方程(5.3.2)说明，使受培训者不同于别人的特征是他在 h 期前的收入。因此我们用面板数据来估计：

$$Y_{it} = \alpha + \theta Y_{it-h} + \lambda_t + \delta D_{it} + X'_{it}\beta + \varepsilon_{it} \tag{5.3.3}$$

此处接受培训的因果效应是 δ。为了让等式(5.3.3)更具一般性，Y_{it-h} 可被看作包含滞后多期的收入的一个向量。①

使用面板数据的应用研究者们经常会面对一个问题：选择固定效应模型还是选择滞后被解释变量模型。也就是说，要在方程(5.3.1)和方程(5.3.2)之间作出选择。面对这一困境，一个解决办法就是在模型中既放入滞后被解释变量，又放入不可观察的个体效应。换言之，基于下式进行识别：

$$E[Y_{0it} \mid \alpha_i, Y_{it-h}, X_{it}, D_{it}] = E[Y_{0it} \mid \alpha_i, Y_{it-h}, X_{it}] \tag{5.3.4}$$

等式(5.3.4)要求在给定 α_i 和 Y_{it-h} 的基础上进行识别。于是我们可以使用下面的这个函数形式来估计因果效应：

① Abadie, Diamond 和 Hainmueller(2007)为滞后被解释变量发展出了半参数估计法，相比于传统的回归模型，在这种方法下的模型设定可以更加灵活。比如在方程(5.3.2)中的关键假设就是给定滞后被解释变量，处理状态和潜在结果之间相互独立。Abadie, Diamond 和 Hainmueller 的方法可以用于微观数据，也可用于分组结构特征的数据。Dehejia 和 Wahba(1999)使用匹配策略时也使用了滞后的被解释变量。

$$Y_{it} = \alpha_i + \theta Y_{it-k} + \lambda_t + \delta D_{it} + X'_{it}\beta + \varepsilon_{it} \tag{5.3.5}$$

不幸的是，在方程(5.3.5)中得到 δ 的一致估计所要求的条件非常严格，远甚于固定效应模型或者滞后被解释变量模型所要求的假设。从一个简单的例子中我们就可以看到这一点，其中我们将滞后被解释变量表示为 Y_{it-1}。通过对等式(5.3.5)进行差分消除固定效应，我们得到下列结果：

$$\Delta Y_{it} = \theta \Delta Y_{it-1} + \Delta\lambda_t + \delta \Delta D_{it} + \Delta X'_{it}\beta + \Delta\varepsilon_{it} \tag{5.3.6}$$

这里的问题在于：由于 $\Delta\varepsilon_{it}$ 和 ΔY_{it-1} 可能是 ε_{it-1} 的函数，所以对残差进行差分后得到的 $\Delta\varepsilon_{it}$ 可能和滞后被解释变量的差分结果 ΔY_{it-1} 相关。因此，对方程(5.3.6)进行最小二乘估计得到的估计值可能与方程(5.3.5)中的参数不一致，这个问题最早是由 Nickell(1981)注意到的。这个问题可以得到解决，但是要求的假设比较强。最简单的解决方法是用 Y_{it-2} 作为等式(5.3.6)中 ΔY_{it-1} 的工具变量。[①]但是这又要求假设 Y_{it-2} 与差分后的残差 $\Delta\varepsilon_{it}$ 不相关，由于残差是控制了协变量后剩下的个体收入部分，因此这个假设看上去很难成立。对大部分人而言，他们在某一年的收入是和下一年收入高度相关的，因此过去的收入很可能与 $\Delta\varepsilon_{it}$ 相关。如果 ε_{it} 是序列相关的，那么等式(5.3.6)就不存在一致性的估计值。(还要注意，使用 Y_{it-2} 作为工具变量的策略需要至少三个时期，因此我们需要第 t，$(t-1)$，$(t-2)$时期的数据。)

看到估计(5.3.6)式时所遇到的诸种困难，会让我们不禁要问的是，固定效应和滞后被解释变量之间的区别是否真的在发挥作用。很不幸，答案是“yes”。固定效应模型和滞后被解释变量模型都不具有嵌套结构，因此我们无法在估计其中一个的同时将另一个当作特例。

那么，进行应用计量经济学研究的朋友们该怎么做呢？通常的答案是使用另外一种用以识别的假设来对你的研究结果做稳健性检验。这意味着你得使用另外一种可行的模型得到相似的结论。固定效应估计值和滞后被解释变量估计值都有一个括号性质(bracketing property)。本章后面的附录指出如果方程(5.3.2)是正确的，但是你错误地使用了固定效应模型，那么估计出的正的因果效应会偏大。从另一方面看，如果方程(5.3.1)是正确的，但是你却使用类似于方程(5.3.3)的滞后被解释变量模型进行估计，那么估计出的正的因果效应可能会偏小。因此你可以将固定效应估计值和滞后被解释变量估计值看作是我们感兴趣的因果效应(给定对选择偏误特性的一些假设)的极大值和极小值，因果效应会落在这两个值决定的区间里。Guryan(2004)研究了法院要求用公共汽车运送黑人学生对黑人学生高中毕业率的影响，在这个研究中作者运用了该想法。

① 对这种估计的具体细节和例子，请参见 Holtz-Eakin，Newey 和 Rosen(1988)，Arellana 和 Bond(1991)以及 Blundell 和 Bond(1998)。

5.4 附录：对固定效应模型和滞后被解释变量模型的进一步讨论

为了简洁起见，我们忽略协变量、截距项和年份效应，假设只有两期，对所有人而言第一期的处理都是0（这里表达的主要想法在更一般化的框架中仍然是一样的）。我们感兴趣的因果效应 δ 是正的。首先假设处理（是否接受培训）与不可观察的个体效应 α_i 相关，与滞后被解释变量的残差 ε_{it-1} 不相关，这时得到的方程可以写为：

$$Y_{it} = \alpha_i + \delta D_{it} + \varepsilon_{it} \tag{5.4.1}$$

这里 ε_{it} 是序列不相关的，而且也与 α_i 和 D_{it} 不相关。我们又有：

$$Y_{it-1} = \alpha_i + \varepsilon_{it-1}$$

这里 α_i 和 ε_{it-1} 是不相关的。你错误地在一个将滞后被解释变量 Y_{it-1} 当作控制变量但却忽略了固定效应的模型中估计 D_{it} 的效果。由此得到的估计值的概率极限是 $\dfrac{\text{cov}(Y_{it}, \tilde{D}_{it})}{V(\tilde{D}_{it})}$，其中 $\tilde{D}_{it} = D_{it} - \gamma Y_{it-1}$ 是用 D_{it} 对 Y_{it-1} 做回归得到的残差。

现在将 $\alpha_i = Y_{it-1} - \varepsilon_{it-1}$ 代入等式(5.4.1)，可得：

$$Y_{it} = Y_{it-1} + \delta D_{it} + \varepsilon_{it} - \varepsilon_{it-1}$$

由此，我们可以得到：

$$\frac{\text{cov}(Y_{it}, \tilde{D}_{it})}{V(\tilde{D}_{it})} = \delta - \frac{\text{cov}(\varepsilon_{it-1}, \tilde{D}_{it})}{V(\tilde{D}_{it})}$$

$$= \delta - \frac{\text{cov}(\varepsilon_{it-1}, D_{it} - \gamma Y_{it-1})}{V(\tilde{D}_{it})} = \delta + \frac{\gamma \sigma_\varepsilon^2}{V(\tilde{D}_{it})}$$

这里 σ_ε^2 是 ε_{it-1} 的方差。因为受训人的收入较低，所以 Y_{it-1} 较低，$\gamma < 0$，所以 δ 的估计值会偏小。

与上面的情况相反，假设处理由较低的 Y_{it-1} 决定。我们可以使用方程(5.3.3)的一个简化版来度量因果效应，即：

$$Y_{it} = \alpha + \theta Y_{it-1} + \delta D_{it} + \varepsilon_{it} \tag{5.4.2}$$

这里 ε_{it} 序列不相关，且与 D_{it} 不相关。这时假设你希望消除固定效应，但是却错误地估计了一个一阶差分方程。这时你忘记将滞后被解释变量加入回归方程。在这个简单的例子中，对所有人都有 $D_{it-1} = 0$，一阶差分估计值的概率极限就是：

$$\frac{\text{cov}(Y_{it} - Y_{it-1}, D_{it} - D_{it-1})}{V(D_{it} - D_{it-1})} = \frac{\text{cov}(Y_{it} - Y_{it-1}, D_{it})}{V(D_{it})} \tag{5.4.3}$$

在方程(5.4.2)的两端都消去 Y_{it-1}，我们有：

$$Y_{it}-Y_{it-1}=\alpha+(\theta-1)Y_{it-1}+\delta D_{it}+\varepsilon_{it}$$

将这个等式代入方程(5.4.3),不正确的差分方程产生:

$$\frac{\mathrm{cov}(Y_{it}-Y_{it-1},\ D_{it})}{V(D_{it})}=\delta+(\theta-1)\left[\frac{\mathrm{cov}(Y_{it-1},\ D_{it})}{V(D_{it})}\right]$$

一般来说,我们认为 θ 是一个小于 1 的正数,否则 Y_{it} 就是不平稳的(也就是说成为一个爆炸性的时间序列过程)。因此,由于受训人有较低的 Y_{it-1},在一阶差分中得到的 δ 的估计值就会过大。注意,在这个简单的模型里可以差分的原因在于方程(5.4.2)中的 θ 不太可能等于 1,不过在更一般化的情况下就不一定了。

第三部分　拓　　展

▶6 更进一步:断点回归设计

但是,一旦你运用那些规则,各种步骤都开始出现,你会发现有关人类的所有特质……这不过是思考问题的一种方式,它让问题以特定的形式表现出来。规则越多,规则划定的范围越小,规则越少,规则作用范围越广。

——Donglas Adams, *Mostly Harmless*(1995)

在使用断点回归(regression discontinuity,简称为 RD)进行研究设计时,要挖掘处理状态如何被决定的详细情况。断点回归式识别策略基于如下思想:在高度依赖规则而运行的世界中,有些规则的出现十分随意,这种随意性为我们提供了性质良好的实验。断点回归法可以分为两类,一类叫作模糊断点回归(fuzzy RD),另一类叫作清晰断点回归(sharp RD)。我们可将清晰断点回归设计看作一类选择偏误来自可观察变量(selection-on-observables)的经验研究方法。模糊断点回归则可被视为一种工具变量法。

6.1 清晰断点回归

当处理状态是协变量 x_i 的确定型、不连续函数时,我们可使用清晰断点回归法。举个例子,假设处理状态 D_i 可以写为如下函数:

$$D_i = \begin{cases} 1 & \text{if } x_i \geqslant x_0 \\ 0 & \text{if } x_i \leqslant x_0 \end{cases} \tag{6.1.1}$$

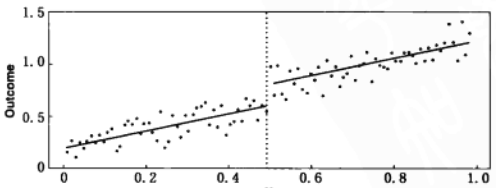
这里 x_0 是已知的阈值或临界值(threshold or cutoff)。由于我们一旦知道 x_i 的取值,就知道 D_i 的取值,所以上面给出的分配处理状态的机制是 x_i 的确定型函数。由于无论 x_i 有多么靠近 x_0,除非 $x_i = x_0$,否则处理状态不发生变化,所以上面给出的机制还是关于 x_i 的不连续函数。

这看起来似乎有点抽象,因此我们在这里给出一个例子。在美国,PSAT 考试是一项绝大多数即将报考大学的高中生都会参加的考试,那些将来准备参加 SAT 考试的高中生更是会去参加这项考试,依照该考试的成绩,美国的高中生

会被授予国家杰出奖学金。激发人们开始讨论断点回归法的一个问题便是：获得国家杰出奖学金的学生是否会因此而改变其职业规划或者学习计划。比如，获得国家杰出奖学金的高中生是不是会更愿意读研究生（Thistlewaithe and Campbell，1960；Campbell，1969）。清晰断点回归法通过比较 PSAT 分数刚好高于和低于国家杰出奖学金分数线的那些高中生的研究生入学率来回答这一问题。一般而言，我们可能认为学生在 PSAT 考试中得分越高，将来读研究生的概率会越大，但是通过回归来拟合研究生院入学率和 PSAT 分数之间的关系，我们可以控制这一趋势，或者说至少在 PSAT 分数线的邻域内，这一趋势可以得到很好的控制。在这个例子中，可将分数线附近 PSAT 成绩和大学入学率之间的关系中出现的跳跃视为存在处理效应的证据。断点回归这一名称正是源自回归线上的这一跳跃。①

Imbens 和 Lemieux(2008)在他们最近发表的一篇综述中强调了断点回归的一个有趣而且重要的特点：我们不可能在处理组和观察组中看到协变量 x_i 取值相同。在完全使用协变量进行匹配时，我们在给定协变量取值后比较处理组和控制组之间的不同，因此处理组和控制组中协变量的取值是相同的。与此不同的是，断点回归的有效性依赖于我们对协变量的外推，或者说至少在协变量不连续的那个邻域内的外推。这是我们为何将清晰断点回归与其他控制策略相区别的原因之一。正是基于同样的原因，在使用断点回归进行研究时，我们不能像第 3 章那样可以不去深究条件期望函数的具体函数形式。

图 6.1 展示了一个我们假设的断点回归情景，其中 $x_i \geqslant 0.5$ 的那些个体受到处理。在 A 图中，协变量 x_i 和结果之间的趋势关系是线性的，在图 B 中，这一关系是非线性的。在图 A 和图 B 两个例子中，我们观察到的条件期望函数 $E[Y_i \mid X_i]$ 在点 x_0 附近是不连续的，而 $E[Y_{0i} \mid x_i]$ 则是光滑的。



① 在好几个学科领域内，断点回归设计的基本结构差不多同时出现，只是在最近才在应用计量经济学中受到重视。Cook(2008)对这一方法的思想史进行了回顾。在一篇对研究结果进行比较的论文中(类似于 Lalonde(1986))，Cook 和 Wong(2008)发现断点回归可以很好地还原随机实验的结果。

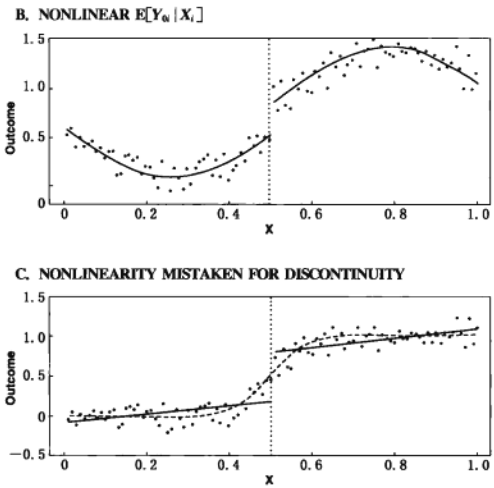


图 6.1 清晰回归断点设计

通过用一个简单的模型，我们可以清楚地表述断点回归的思想。除了用等式(6.1.1)表示的分配处理状态的机制，假设我们可以使用一个线性、常因果效应模型来描述潜在结果：

$$\mathrm{E}[Y_{0i} \mid x_i] = \alpha + \beta x_i$$
$$Y_{1i} = Y_{0i} + \rho$$

这意味着我们要做的回归是：

$$Y_i = \alpha + \beta x_i + \rho D_i + \eta_i \tag{6.1.2}$$

这里 ρ 是我们关心的因果效应。这个回归和其他我们曾用来估计处理效应的回归(例如在第 3 章使用的回归)之间的关键差异在于 D_i，我们感兴趣的这个回归元不仅与 x_i 相关，而且还是 x_i 的确定型函数。通过在关于 x_i 的光滑、线性函数中分离出非线性、不连续的函数 $1(x_i \geqslant x_0)$，断点回归便可捕捉到我们感兴趣的因果效应。

但是如果我们表示趋势关系的函数 $\mathrm{E}[Y_{0i} \mid x_i]$ 是非线性的，会发生什么情况呢？精确起见，假设 $\mathrm{E}[Y_{0i} \mid x_i] = f(x_i)$，基于易于处理的考虑，再假设 $f(x_i)$ 是个光滑函数。图 6.1 中的图 B 暗示，即使在这种更为一般性的假设下，也有可能使用断点回归设计。现在我们需要通过拟合等式(6.1.3)来构造断点回归估计值。

$$Y_i = f(x_i) + \rho D_i + \eta_i \tag{6.1.3}$$

这里 $D_i = 1(x_i \geqslant x_0)$ 仍然是 x_i 的函数，且在 x_0 处不连续。由于 $f(x_i)$ 在 x_0 的领域中是连续的，所以即使 $f(x_i)$ 的函数形式变化多样，我们也可以估计出等式(6.1.3)。比如，用 p 次多项式来模型化 $f(x_i)$，我们可以从下面的回归中构造断点回归估计值：

$$Y_i = \alpha + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_p x_i^p + \rho D_i + \eta_i \tag{6.1.4}$$

根据方程(6.1.4)我们可以进一步对断点回归法进行一般化，允许 $\mathrm{E}[Y_{0i} \mid x_i]$ 和 $\mathrm{E}[Y_{1i} \mid x_i]$ 是两个不同的趋势函数。通过分别对这两个条件期望函数进行 p 阶多项式展开，我们有：

$$\mathrm{E}[Y_{0i} \mid x_i] = f_0(x_i) = \alpha + \beta_{01}\tilde{x}_i + \beta_{02}\tilde{x}_i^2 + \cdots + \beta_{0p}\tilde{x}_i^p$$

$$\mathrm{E}[Y_{1i} \mid x_i] = f_1(x_i) = \alpha + \rho + \beta_{11}\tilde{x}_i + \beta_{12}\tilde{x}_i^2 + \cdots + \beta_{1p}\tilde{x}_i^p$$

这里 $\tilde{x}_i \equiv x_i - x_0$。将关于 x_i 的函数在 x_0 的邻域内展开是一种标准化的方式，这么处理的好处是当 $x_i = x_0$ 时，我们希望估计的处理效应正好等于将 D_i 当作交互项时所对应的系数。

为了求出一个可在本例中用来估计因果效应的回归模型，我们使用下列事实：D_i 是 x_i 的确定型函数，写作：

$$\mathrm{E}[Y_i \mid X_i] = \mathrm{E}[Y_{0i} \mid x_i] + (\mathrm{E}[Y_{1i} - Y_{0i} \mid x_i])D_i \tag{6.1.5}$$

将上面给出的两个条件期望函数的多项式展开代入等式(6.1.5)，我们可得：

$$Y_i = \alpha + \beta_{01}\tilde{x}_i + \beta_{02}\tilde{x}_i^2 + \cdots + \beta_{0p}\tilde{x}_i^p + \rho D_i + \beta_1^* D_i\tilde{x}_i + \beta_2^* D_i\tilde{x}_i^2 + \cdots + \beta_p^* D_i\tilde{x}_i^p + \eta_i \tag{6.1.6}$$

其中，$\beta_1^* = \beta_{11} - \beta_{01}$，$\beta_2^* = \beta_{12} - \beta_{02}$，$\beta_p^* = \beta_{1p} - \beta_{0p}$，$\eta_i$ 是误差项。

方程(6.1.4)是方程(6.1.6)的一个特例，其中 $\beta_1^* = \beta_2^* = \beta_p^* = 0$。在更为一般化的模型中，在 $x_i - x_0 = c > 0$ 处的处理效应是 $\rho + \beta_1^* c + \beta_2^* c^2 + \cdots + \beta_p^* c^p$，在 x_0 处的处理效应则为 ρ。在模型中加入交互项的好处在于它放松了我们对待估的条件期望函数所施加的限制。但是根据我们的经验，基于类似于等式(6.1.4)这类比较简单的模型对参数 ρ 进行断点回归估计，与基于等式(6.1.6)得到的估计值很接近。这种情况并不奇怪，因为上述两种方法都是在 x_0 的邻域中对 $\mathrm{E}[Y_i \mid X_i]$ 进行估计。

基于等式(6.1.4)或者等式(6.1.6)得到的断点回归估计值的有效性依赖于多项式模型能否足够精确地描述 $\mathrm{E}[Y_{0i} \mid x_i]$。如果不能，那么看上去由于个体被处理而发生的跳跃可能只不过是条件期望函数在某个点上的不连续，而我们并未提前预计到这种不连续性。我们在图 6.1 中的图 C 中展示了这种可能性，在这幅图中，我们可能将 $\mathrm{E}[Y_{0i} \mid x_i]$ 曲线的急剧转向错误理解为回归曲线发生的跳跃。为了降低出现这种错误的可能性，我们可以只去考察在不连续点的邻域中的数据，也就是考察区间 $[x_0 - \Delta, x_0 + \Delta]$，其中 Δ 是某个很小的正数。于是我们有：

$$\mathrm{E}[Y_i \mid x_0 - \Delta < x_i < x_0] \simeq \mathrm{E}[Y_{0i} \mid x_i = x_0]$$
$$\mathrm{E}[Y_i \mid x_0 \geqslant x_i \geqslant x_0 + \Delta] \simeq \mathrm{E}[Y_{1i} \mid x_i = x_0]$$

因此，

$$\lim_{\Delta \to 0} \mathrm{E}[Y_i \mid x_0 \leqslant x_i < x_0 + \Delta] - \mathrm{E}[Y_i \mid x_0 - \Delta < x_i < x_0] \quad (6.1.7)$$
$$\mathrm{E}[Y_{1i} - Y_{0i} \mid x_i = x_0]$$

换言之，在 x_0 左侧和右侧一个足够小的邻域内比较 Y_{1i} 和 Y_{0i} 的平均值之间的差别，就可估计出我们感兴趣的处理效应，而且这种方法与 $\mathrm{E}[Y_{0i} \mid x_i]$ 的具体设定形式无关。此外，这种非参数的估计方法还无需处理效应 $Y_{1i} - Y_{0i} = \rho$ 为常数的假设；在等式(6.1.7)中被估量 $\mathrm{E}[Y_{1i} - Y_{0i} \mid x_i = x_0]$ 表示对处理效应的平均化。

使用非参数方法对断点回归进行估计，需要分别对 x_0 左侧和右侧邻域中的 Y_i 的平均值作出精确估计。但是求得上述两个估计值具有相当难度。首先遇到的问题是如果我们在临界值的一个很小的邻域中进行估计，那么可用的数据不会多。而且，在有界邻域中对条件期望函数的估计也是有偏的(在这个例子中，有界邻域指的是 x_0 的某个邻域)。解决这个问题的方法就是使用非参数的局部线性回归(Hahn, Todd and van der Klaauw, 2001)，以及由 Porter(2003)发展出的部分线性、局部多项式回归(partial linear, local polynomial regression)。局部回归类似于对等式(6.1.6)使用加权最小二乘法，其中数据点越靠近临界值，该数据点上赋予的权重越大。

在经验研究实践中，针对断点回归的复杂非线性估计法并未得到广泛应用；大部分应用性的断点回归估计仍然是参数型的。但该方法中对接近临界值的数据点赋予更大权重的思想则暗示了一种非常有价值的稳健性检验。在 Angrist 和 Levy (1999)中，他们将这一思想称为“不连续样本”(discontinuity sample)。尽管随着不连续样本窗口的缩小，断点回归估计值会变得不精确，但是用来模型化函数 $f(x_i)$ 的多项式的阶数也会下降。乐观地看，当你以 x_0 为中心不断调整样本窗口大小时，控制变量会越来越少，但是针对 D_i 估计出的处理效应应该保持稳定。①第二个重要的检验则关注不连续点附近预处理变量的行为。由于预处理变量不受处理状态的影响，所以在 x_0 附近针对预处理变量估计出的条件期望函数不应该有跳跃。

① Hoxby(2000)也应用这一思想对班级规模带来的效应做了断点回归检验。完全的非参数估计法要求依照数据的状况来选择不连续样本窗口，这一窗口又被称为带宽(bandwidth)。为了保证对潜在的条件期望函数的一致估计，样本规模的带宽下降速度要足够慢。这方面的细节请参看 Imbens 和 Lemieux(2007)。我们倾向于使用类似等式(6.1.4)和等式(6.1.6)那样的纯参数模型来考虑回归问题：在任何给定的样本中，其他模型得到的估计值不会比你碰巧在用的模型得到的估计值好多少。那些如果你有更多数据，你该如何改变模型的建议都不会提高估计的精度。

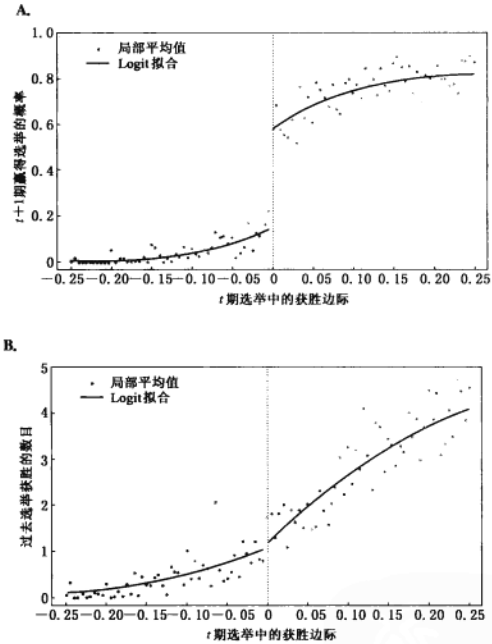
Lee(2008)研究了执政党地位对再次当选产生的影响，通过这项研究他阐明了清晰断点回归设计的使用方法。Lee 感兴趣的问题是如果民主党在上次竞选中获胜，那么它们是否会在本次竞选中具有优势。摆在大家面前的执政党常常连任的事实自然而然地向人们提出了这个问题：议员是否会利用他们的官方身份所带来的权利和资源为他们自己及其党派谋取利益？这一假说听起来非常合理，执政党的成功并不必然反映真正的选举优势。执政党——根据定义，就是那些业已显示出他们能够取得胜利的候选人和党派——可能只不过是在满足投票者或获取选票方面更胜一筹。

为了捕捉执政党地位带来的这一因果效应，Lee 将民主党候选人获胜视作前一次选举中相对得票份额的函数。具体而言，他发现一个事实：候选人能否获胜可由函数 $D_i = 1(x_i \geqslant 0)$ 决定，这里 x_i 是选举胜利者在边际上的得票份额（也就是说当民主党和共和党是最大的两个党时，两党的得票份额之差）。注意到由于 D_i 是 x_i 的确定性函数，所以在 x_i 之外并无其他变量带来干扰。这是断点回归设计的一个标志性特点。

图 6.2A 摘自 Lee(2008)，它显示出在该项研究中清晰断点回归设计确实发挥了作用。这幅图的纵轴和横轴分别绘出的是民主党人获胜的概率和在上一次选举中民主党和共和党得票份额之差。图中的点表示局部平均值（在相互无交叠的样本窗口中估计出的平均获胜概率，其中得票份额的边际变化是 0.005）；图中的线段是用一个在零点处不连续的参数模型进行拟合后得到的结果。[①]民主党获胜的概率是过去相对得票份额的增函数。但是这幅图中最重要的特点是在 0 点处获胜概率的大幅提高，在这个点处，民主党得多数票。基于这幅图中跳跃性的大小，执政党大约可以将再次当选的概率提高 40%。

图 6.2B 通过考察民主党在上一次获胜之前的获胜状况来检验清晰断点回归设计的假设。民主党在过去选举中的获胜率应该和上次选举中获胜的边际临界值无关，这个检验的结果很好，增强了我们对本例中使用断点回归研究方法的信心。Lee 对预处理个体获胜状况的考察实际上表达了一种思想：协变量应该被处理状态所平衡，因此给定处理状态，协变量应该像是在随机实验中赋予的。另外一个与之联系的检查通过计算接近 x_0 处的 x_i 的比例来考察不连续点附近 x_i 的分布密度。这么做的考虑是：在选举结果中存在私利的人可能会操控处在临界值附近的 x_i，从而使得临界值两边的状况不可比[McCrary(2008)对这个问题提出了一个正式的检验方法]。但是直到最近，我们还是可以说在 Lee 的研究中这种情况不太会出现。但是在 2000 年美国大选之后发生在佛罗里达的重新计票的情况意味着当美国选举中两党的票数接近时，可能存在着操纵选举的问题。

① 在这幅图里的拟合值来自于关于获胜概率的 logit 模型，该模型是关于下面一类变量的函数：表示临界点的 $D_i = 1(x_i \geqslant 0)$，关于 x_i 的四阶多项式以及多项式中各项和 D_i 组成的交互项。



注：本图来自 Lee(2008)(A)根据 t 期选举中的获胜边际(margin of victory)候选人在 $t+1$ 期选举获胜的概率；局部平均和参数拟合。(B)根据在 t 期选举中的获胜边际，候选人过去选举获胜的累积数目；局部平均和参数拟合。

图 6.2　过去赢得选举的概率以及未来投票份额

6.2　作为一种工具变量法的模糊断点回归

模糊断点回归设计(fuzzy RD)要挖掘的是给定某个协变量时，处理状态的概率或者期望值所发生的不连续变化。这样做的结果就是，在我们所用的研究设计中，不连续性成了针对处理状态的工具变量，不再和处理状态有确定性的联系。为了看清楚这个过程如何运行，和之前一样，令 D_i 表示处理状态，但是与清晰断点回归设计不同的是，这里 D_i 不再是临界值的确定型函数。不过，在 x_0 处个体被处理的概率还是应该有一个跳跃，也即：

$$\mathrm{P}(D_i = 1 \mid x_i) = \begin{cases} g_1(x_i) & \text{if } x_i \geqslant x_0 \\ g_0(x_i) & \text{if } x_i < x_0 \end{cases}, \text{其中 } g_1(x_0) \neq g_0(x_0)$$

函数 $g_0(x_i)$和 $g_1(x_i)$可以是任何函数，只要它们在 x_0 处不同即可（这种差别越大越好）。这里我们假设 $g_1(x_0) > g_0(x_0)$，也就是说当 $x_i \geqslant x_0$ 时，个体被处理的概率会增大。我们可以将被处理状态和 x_i 之间的关系记为：

$$\mathrm{E}[D_i \mid x_i] = \mathrm{P}[D_i = 1 \mid x_i] = g_0(x_0) + [g_1(x_0) - g_0(x_0)]T_i$$

其中，

$$T_i = 1(x_i \geqslant x_0)$$

虚拟变量 T_i 表示 $\mathrm{E}[D_i|x_i]$不连续的点。

模糊断点回归设计很自然地为我们带来了一个简单的工具变量估计策略。与我们在上一节处理 $f_0(x_i)$和 $f_1(x_i)$的方法一样，假设可以用 p 阶多项式来描述函数 $g_0(x_i)$和 $g_1(x_i)$，于是我们有：

$$\begin{aligned}\mathrm{E}[D_i \mid x_i] &= \gamma_{00} + \gamma_{01}x_i + \gamma_{02}x_i^2 + \cdots + \gamma_{0p}x_i^p \\ &\quad + [\pi + \gamma_1^* x_i + \gamma_2^* x_i^2 + \cdots + \gamma_p^* x_i^p]T_i \\ &= \gamma_{00} + \gamma_{01}x_i + \gamma_{02}x_i^2 + \cdots + \gamma_{0p}x_i^p \\ &\quad + \pi T_i + \gamma_1^* x_i T_i + \gamma_2^* x_i^2 T_i + \cdots + \gamma_p^* x_i^p T_i\end{aligned} \tag{6.2.1}$$

这里 γ^* 都是多项式和 T_i 所成的交互项前面的系数。

从这里的讨论我们可知，T_i 以及交互项$\{x_iT_i, x_i^2T_i, \cdots, x_i^pT_i\}$都可以作为方程(6.1.4)中 D_i 的工具变量来使用。①

最为简单的模糊断点回归估计值只用 T_i 做工具变量，不涉及它的交互项（如果将交互项纳入工具变量，那么在类似等式(6.1.6)的第二阶段也要加入交互项）。由此得到的恰好识别的工具变量估计值既直观又具有良好的有限样本性质。在这个例子中，第一阶段是：

$$D_i = \gamma_0 + \gamma_1 x_i + \gamma_2 x_i^2 + \cdots + \gamma_p x_i^p + \pi T_i + \xi_{1i} \tag{6.2.2}$$

这里 π 是 T_i 在第一阶段的效应。

通过将等式(6.2.2)代入方程(6.1.4)，我们可以得到模糊断点回归的简约式：

$$Y_i = \mu + \kappa_1 x_i + \kappa_2 x_i^2 + \cdots + \kappa_p x_i^p + \rho\pi T_i + \xi_{2i} \tag{6.2.3}$$

其中，$\mu = \alpha + \rho\gamma_0$ 以及 $\kappa_j = \beta_j + \rho\gamma_j$，$j = 1, \cdots, p$。与清晰的断点回归类似，模糊断点回归的识别能力也依赖于从第一和第二阶段用多项式做控制变量的过程中将

① 将被处理概率上出现的跳跃当作识别因果效应的信息来源的思想最初出现在 Trochim(1984)中，不过将这一思想表达为工具变量则出现得较晚。并非所有的人都同意模糊断点回归设计是一种工具变量法，但这种观点十分吸引人。在最近的一篇对断点回归设计的历史进行回归的论文中，Cook(2008)针对模糊断点回归写道："在很多情况下，临界值都可以达到工具变量的作用并带来对因果效应的无偏估计……与早先的看法相比，如今人们认为对分配处理状态的过程的模糊认识不再是一个很严重的问题。"

Y_i 和不连续函数 $T_i = 1(x_i \geqslant x_0)$之间的联系分离出来的能力。在应用计量经济学中首次使用断点回归设计的研究中，van der Klaauw(2002)使用模糊断点回归设计来估计对大学生进行资助的奖学金会怎样影响大学入学率。在 van der Klaauw 的研究中，D_i 是进行资助的奖学金的规模，T_i 是个虚拟变量，表示能力得分高于预设的获得奖学金的临界值的人，在模糊断点回归设计中，他用该得分的多项式做控制。①

当处理效应是随着 x_i 的变化而变化的 x_i 的函数时，可以使用具有处理—控制交互项的两阶段最小二乘估计来构造模糊断点回归估计值。具有交互项时的第二阶段模型与等式(6.1.6)相同，第一阶段与等式(6.2.1)类似，唯一的不同在于为了能够和第二阶段的参数化过程保持一致，我们在点 x_0 构造多项式。在这个例子中，没有在第一阶段包含进来的工具变量是 $\{T_i,\ \tilde{x}_i T_i,\ \tilde{x}_i^2 T_i,\ \cdots,\ \tilde{x}_i^p T_i\}$，而变量 $\{D_i,\ \tilde{x}_i D_i,\ D_i\tilde{x}_i^2,\ \cdots,\ D_i\tilde{x}_i^p\}$ 则是作为内生变量来处理的。于是 D_i 的第一阶段就变成了：

$$\begin{aligned} D_i = {} & \gamma_{00} + \gamma_{01}\tilde{x}_i + \gamma_{02}\tilde{x}_i^2 + \cdots + \gamma_{0p}\tilde{x}_i^p \\ & + \pi T_i + \gamma_1^* \tilde{x}_i T_i + \gamma_2^* \tilde{x}_i^2 T_i + \cdots + \gamma_p^* \tilde{x}_i^p T_i + \xi_{1i} \end{aligned} \tag{6.2.4}$$

对集合 $\{\tilde{x}_i D_i,\ D_i\tilde{x}_i^2,\ \cdots,\ D_i\tilde{x}_i^p\}$ 中的每个多项式而言，也要构造出一个类似的第一阶段模型。

使用非参数方法对模糊断点回归进行估计时，要在不连续点的领域里进行工具变量估计。在点 x_0 附近，Y_i 的条件期望函数的简约型是：

$$\mathrm{E}[Y_i \mid x_0 \leqslant x_i < x_0 + \Delta] - \mathrm{E}[Y_i \mid x_0 - \Delta < x_i < x_0] \simeq \rho\pi$$

同样，对于第一阶段的 D_i，我们有：

$$\mathrm{E}[D_i \mid x_0 \leqslant x_i < x_0 + \Delta] - \mathrm{E}[D_i \mid x_0 - \Delta < x_i < x_0] \simeq \pi$$

因此，

$$\lim_{\Delta \to 0} \frac{\mathrm{E}[Y_i \mid x_0 \leqslant x_i < x_0 + \Delta] - \mathrm{E}[Y_i \mid x_0 - \Delta < x_i < x_0]}{\mathrm{E}[D_i \mid x_0 \leqslant x_i < x_0 + \Delta] - \mathrm{E}[D_i \mid x_0 - \Delta < x_i < x_0]} = \rho \tag{6.2.5}$$

等式(6.2.5)所对应的样本值就是在第 4.1.2 节讨论过的瓦尔德估计值，这个估计值是在 x_0 的 Δ 邻域中使用 T_i 做 D_i 的工具变量而得到的。②由于这里使用虚拟变量做工具变量，所以得到的结果是一个局部平均处理效应。具体而言，模糊断点回归中得到的瓦尔德估计值捕捉到的是响应工具变量的个体表现出的因果效

① Van der Klaauw 最初的工作论文在 1997 年就开始流传。需要注意的是等式(6.2.2)是可加的模型，因此它只是对 $\mathrm{E}[D_i \mid x_i]$的一个近似。不过这并不重要，第二阶段估计值仍然是一致的。

② 为了允许在临界值的两边都可以出现斜率的变化，Imbens 和 Lemieux(2008)建议在计算等式(6.2.5)时，应该在临界值的一个很小的邻域内对 D_i 使用工具变量 T_i 进行两阶段最小二乘回归，同时将交互项$\{\tilde{x}_i T_i,\ \tilde{x}_i^2 T_i,\ \cdots,\ \tilde{x}_i^p T_i\}$作为内生控制变量包括在内。

应，也就是说当我们令 x_i 的取值从恰好小于 x_0 变到恰好大于 x_0 时，被处理状态发生变化的那些个体表现出的因果效应。对模糊断点回归的上述解释由 Hahn、Todd 和 van der Klaauw(2001)给出。但是从另外一个角度也可以指出这里得到的局部平均处理效应确实是局部的：估计值是针对 x_0 的邻域中的 x_i 得到的，具有非参数的清晰断点回归的特征。

最后，就像非参数方法下清晰断点回归估计值一样，等式(6.2.5)的估计值的有限样本性质不是很好。Hahn，Todd 和 van der Klaauw(2001)发展出一套非参数的工具变量估计过程，该过程使用局部线性回归来估计瓦尔德估计值的上下限，以此来减少估计偏误。这种方式实际上是用线性和多项式做控制变量的两阶段最小二乘模型，只不过我们使用不连续模型进行估计并且依照数据来选择带宽大小。使用不连续模型进行回归的想法也适用于如下情景：首先基于等式(6.1.4)使用全样本进行参数式的两阶段最小二乘估计。然后将样本限制在靠近不连续点的附近，去掉大部分或者所有的多项式控制变量。在完美情况下，只用很少的控制变量在不连续样本中进行两阶段最小二乘估计得到的估计值，与使用大样本得到的更精确的估计值相去不远。

Angrist 和 Lavy(1999)使用模糊断点回归研究设计来估计班级规模对学生考试成绩的影响，这个问题在我们第 2 章讨论过的 STAR 实验中已经得到解决。在 Angrist 和 Lavy 的研究中，他们强调模糊断点回归是特别强大且灵活多变的一种研究设计方法，并且从上面提到的两种方式入手对模糊断点回归进行了一般化。首先，我们感兴趣的因果变量是班级规模，它可以取多个值(正如在第 4 章讨论平均因果响应时看到的那样)。因此第一阶段考察的是班级平均规模上表现出的跳跃，而不是概率的跳跃。第二，Angrist 和 Lavy(1999)的研究设计使用了多个不连续点。

Angrist 和 Lavy 的研究从下述观察开始：在以色列的学校中，班级规模是 40 人封顶。如果学校中某个年级的学生数少于 40 人，那么这些学生将会被编入一个人数小于 40 人的班级，如果某个年级有 41 个学生，那么这个年级将被分为两个班级，有 81 个学生的年级将被分为三个班级，以此类推。Angrist 和 Lavy 将此称为“迈蒙尼德(Maimonides)”法则，因为中世纪犹太经典《塔木德》学者迈蒙尼德最早提出了班级规模最多为 40 人的看法。为了正式地讨论迈蒙尼德法则，我们令 m_{sc} 表示使用迈蒙尼德法则，我们在学校 s 中某个给定年级计算出的分配给班级 c 的学生人数，其中该年级总的人学人数记为 e_s。假定各个年级分成同等规模的小班，应用迈蒙尼德法则后预计班级规模为：

$$m_{sc} = \frac{e_s}{\text{int}\left[\frac{(e_s - 1)}{40}\right] + 1}$$

这里 int(a)表示实数 a 的整数部分。在图 6.3 中，虚线表示针对四年级和五年级的班级绘出的迈蒙尼德法则，可见这一函数呈锯齿状，在 40 的整数倍处出现不连

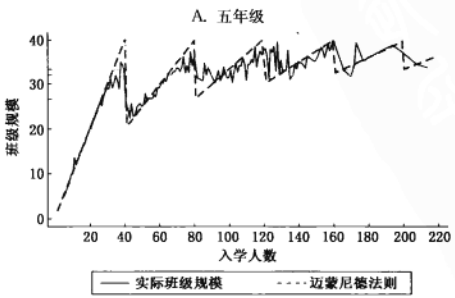
续(在这个例子中,不连续性来自于用迈蒙尼德法则预测出的班级规模的急剧下降)。与此同时,m_{ic} 显然是入学人数 e_s 的增函数,这使得入学人数是一个重要的控制变量。

Angrist 和 Lavy 通过对下面的方程构造两阶段最小二乘估计来利用迈蒙尼德法则中存在的不连续性:

$$Y_{ic} = \alpha_0 + \alpha_1 d_s + \beta_1 e_s + \beta_2 e_s^2 + \cdots + \beta_p e_s^p + \rho n_{sc} + \eta_{ic} \qquad (6.2.6)$$

这里 Y_{ic} 是学校 s 中 c 班学生 i 的测试成绩,n_{sc} 是这个班的班级规模,e_s 是入学人数。在这个模糊断点回归中,m_{ic} 扮演 T_i 的角色,e_i 扮演 x_i 的角色,而班级规模 n_{ic} 则扮演 D_i 的角色。这里 Angrist 和 Lavy 还将一个与入学人数无关的协变量 d_s 加入回归,用以控制学校中具有残障或者贫困等不利背景的学生比例。对于断点回归而言,这种做法是没有必要的,因为在断点回归模型中唯一的遗漏变量偏误来源于 e_s,不过这样做的好处在于我们可以拿如此设定出的模型与构造相应最小二乘回归估计值的模型进行比较。①

图 6.3 分别针对四年级和五年级学生绘出了实际的班级规模和用迈蒙尼德法则计算的班级规模。迈蒙尼德法则并没有完美地预测出班级规模,大部分是因为对于很多学校而言,即使年级人数没有超过 40,他们也会将学生分班。正是这个原因,使得我们要从模糊断点回归出发来设计研究。不过,我们仍然可以看到在入学人数为 40、80 和 120 处发生的班级规模的剧降。还要注意到的是,工具变量 m_{ic} 暗含着不连续性和斜率的不连续性两者的交互项,类似于在单变量模型中等式(6.2.4)中的 $\tilde{x}_i T_i$(在每个节点上,m_{ic} 都变成了 e_i 的一个更小的函数)。这种十分紧凑的参数化过程来源于我们对决定以色列学校班级规模的制度和规则的详细了解。



① Angrist 和 Lavy(1999)的研究与这里的叙述存在一点微小的区别,这里作者用班级规模的平均值来估计等式(6.2.6)。由于协变量都是定义在班级或者学校层面上的,所以,使用学生层面数据得到的结果和使用班级层面数据得到的结果之间的唯一差别在于对学生层面的数据进行平均时使用的加权方式。

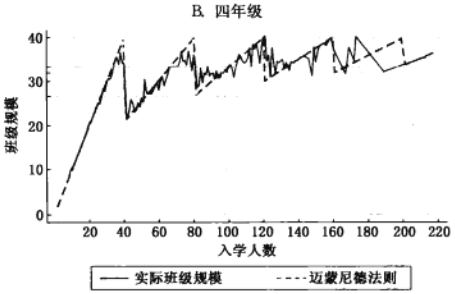


图 6.3　在使用模糊断点回归研究班级规模对学生成绩产生的影响中，第一阶段进行的不连续回归（来自 Angrist 和 Lavy，1999）

针对五年级学生的数学成绩，我们对等式（6.2.6）进行回归的结果报告在表 6.1 中，该表首先报告了最小二乘回归的结果。当不存在控制变量时，班级规模和考试成绩之间存在强烈的正相关关系。当把学校中具有残障或者贫困等不利背景的学生比例当作控制变量纳入回归后，大部分正相关关系都消失了。当把入学率作为控制变量加入回归后，班级规模和学生成绩之间的正相关关系变得不再显著，我们可以从第 3 列中看到这个结果。虽然在田纳西州的 STAR 随机实验中我们发现班级规模越小越好，但是在最小二乘回归得到的结果中，我们并未证实这一结论。

表 6.1　分别用最小二乘和模糊断点回归估计的班级规模对五年级学生数学成绩的影响

	最小二乘估计			两阶段最小二乘估计				
				全样本		不连续样本		
						±5		±3
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
平均得分		67.3			67.3		67.0	67.0
（标准误）		(9.6)			(9.6)		(10.2)	(10.6)
回归元								
班级规模	0.322	0.076	0.019	−0.230	−0.261	−0.185	−0.443	−0.270
	(0.039)	(0.036)	(0.044)	(0.092)	(0.113)	(0.151)	(0.236)	(0.281)
残疾人百分比		−0.340	−0.332	−0.350	−0.350	−0.459	−0.435	
		(0.018)	(0.018)	(0.019)	(0.019)	(0.049)	(0.049)	
入学			0.017	0.041	0.062		0.079	
			(0.009)	(0.012)	(0.037)		(0.036)	
入学的平方/100					−0.010			
					(0.016)			

（续表）

	最小二乘估计			两阶段最小二乘估计				
				全样本		不连续样本		
						±5		±3
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
第一组（入学在 38—43 之间）								−12.6 (3.80)
第二组（入学在 78—83 之间）								−2.89 (2.41)
R^2	0.048	0.249	0.252					
班级数量		2 018			2 018		471	302

注：改编自 Angrist 和 Lavy(1999)。该表报告了文中使用班级平均值的方程(6.2.6)的估计结果。标准误在圆括号内报告，是经过校内关联修正过的。

相比于第 3 列中的最小二乘估计值，使用 m_{ic} 作为 n_{ic} 的工具变量进行的估计强烈地指出较小的班级规模可以提高考试成绩。第 4 列报告的结果将入学率的线性项纳入回归，第 5 列报告的结果将入学率的二次项纳入回归，得到的结论分别是 −0.23 和 −0.26，标准误在 0.1 左右。这个结果意味着如果班级规模下降 7 人（正如田纳西州 STAR 实验结果那样），那么数学成绩将提高 1.75 个百分点，或者说是 0.18σ，这里 σ 是指班级平均得分的标准误。这个数值与田纳西州随机实验得到的结果相去不远。

重要的是，用入学人数做控制变量时其函数形式并不重要（表 6.1 中未报告不含控制变量的结果，不过该结果更小且更不显著）。第 6、7 列进一步使用 ±5 的不连续样本来验证主要结论的稳健性。毫不惊讶，其精确性不如第 4、5 列中的结果，因为它们只用了全样本数据的四分之一来构造估计值。但是，这两个结果都在 −0.25 之间。最后，第 8 列使用了更小的不连续样本，只考虑入学人数与临界值 40、80 和 120 相差不超过 3 人的样本（用与这些不连续性相关的虚拟变量做控制变量）。按照 Hahn、Todd 和 van der Klaauw(2001)的看法，这个估计值是等式(6.2.5)所表示的瓦尔德估计值；用来构造估计值的工具变量是个虚拟变量，该工具变量用以表示所在学校的入学人数是否恰好大于临界值。这个估计结果是 −0.270，不是很精确，但仍然与该表中其他估计值很接近。这个估计值显示出当我们将样本局限在不连续点附近时，所要付出的代价是牺牲精确度。不过，值得欣慰的是，表 6.1 所给出的结果仍然相当清晰明了。

▶7 分位数回归

这是篇为你准备的祈祷文。拿支铅笔来？……“保佑我不要知道那些我不必要知道的事。保佑我甚至不要知道有些事情需要知道我不知道。保佑我不要知道我决定不要知道的那些我决定不要知道的事情。阿门。”下面是和它一起的另外一篇祈祷文。“主啊，主啊，主。保佑我不要承受上面祷告之苦。”

——Douglas Adams, *Mostly Harmless* (1995)

且不同是对是错，95%的应用计量经济学考虑的都是平均值。举个例子，如果一个培训项目提高了平均收入，使其足够补偿成本，我们会对此感到满意。对平均值的关注部分因为很难对平均因果效应进行很好的估计。而且如果因变量是诸如就业这类的虚拟变量的话，均值就能描述其整个分布。但是，很多像收入和测验分数这类变量其分布是连续的。这时只考察平均值就无法揭示出整个分布的变化；比如，分布可能会变得更加离散，也可能变得更加紧凑。在了解平均值之外，应用计量经济学家一直都想知道整个分布是如何发生变化的，想知道分布发生变化时谁的相对状况变差，谁的相对状况变好。

政策制定者和劳动经济学家一直对工资分布的变化十分感兴趣。比如，我们知道在过去 25 年中没有发生较大变化的平均真实工资只是劳动力市场上发生的很小一部分事情。在整个社会的收入水平中，收入最高的那部分人所占的百分比在不断上升，收入最低的那部分人所占的百分比在不断下降。换言之，富人变得更富，穷人变得更穷。最近，社会不平等出现了非对称的变化；比如，在大学毕业生中，富人变得更富，但是收入水平处在较低水平上的那些人的收入则没有发生变化。工资分布发生的变化的全貌是很相当复杂的，而且看上去也很难总结和描述。

百分数回归是一类十分有用的工具，即使所考虑的问题很复杂并且有多个维度，我们也可以运用这个方法以使对分布进行模型化的任务变得简单。通过使用百分数回归，我们可以考察参加某个培训项目或者成为工会会员是否会在影响平均收入的同时影响到工资的不平等水平。在这种方法下我们也可以使用交互项，从而方便地考察教育水平和不平等之间的关系是否以及如何随着时间的变化而变化。百分数回归和传统的回归很相似：通过在模型中纳入协变量，我们可以解决存

在干扰因素的问题；交互项的作用也类似。当不太可能出现选择偏误来自可观察变量时，我们还能用工具变量法来估计在百分比上发生的因果效应。

7.1 分位数回归模型

分位数回归的起点是条件分位数函数(conditional quantile function，简写为CQF)。假定我们对于连续分布随机变量 Y_i 的分布感兴趣，其密度函数性质良好(无缝隙或者尖角)。那么给定回归元所在的向量 X_i，在分位数 τ 处的条件分位数函数可定义为：

$$Q_\tau(Y_i \mid X_i) = F_y^{-1}(\tau \mid X_i)$$

此处 $F_y(y \mid X_i)$ 是在给定 X_i 时 Y_i 在 y 处的分布函数。比如，当 $\tau = 0.10$ 时，$Q_\tau(Y_i \mid X_i)$ 描述的是给定 X_i 下 Y_i 的第一个十分位数，而 $\tau = 0.5$ 时该函数告诉我们的是条件中位数①。通过将收入的条件分位数函数看成教育水平和时间的函数，我们可以了解收入的分布是否会随着教育水平而上下变动。将收入的条件分位数函数看作教育水平和时间的函数，我们还可以了解教育水平和不平等之间的关系是否随时间变化而变化。

条件分位数函数实际上是一类特殊的条件期望函数。回想一下，通过求解在均方误意义下的预测问题，我们可以求得条件期望函数：

$$\mathrm{E}[Y_i \mid X_i] = \arg\min_{m(X_i)} \mathrm{E}[(Y_i - m(X_i))^2]$$

同理，条件分位数函数也是下面这个最小化问题的解：

$$Q_\tau(Y_i \mid X_i) = \arg\min_{q(X)} \mathrm{E}[\rho_\tau(Y_i - q(X_i))] \tag{7.1.1}$$

这里 $\rho_\tau(u) = (\tau - 1(u \leqslant 0))u$ 被称为“校验函数(check function)”，因为当你在图中绘出这个函数时，会发现该函数的形状类似于一个校验符号②。如果 $\tau = 0.5$，等式(7.1.1)变为离差的绝对值的最小值，因为 $\rho_{0.5}(u) = \frac{1}{2}(\mathrm{sign}\, u)u = \frac{1}{2} \mid u \mid$。在此情况下，由于条件中位数是使得离差的绝对值最小的数字，所以 $Q_\tau(Y_i \mid X_i)$ 是条件中位数。在其他情况下，校验函数分别对正数和负数赋予一个权重，但是这两个权重不同：

$$\rho_\tau(u) = 1(u > 0) \cdot \tau \mid u \mid + 1(u \leqslant 0) \cdot (1 - \tau) \mid u \mid$$

① 更一般地，当随机变量为离散的或者随机变量的密度函数性质不好时，我们可将条件分位数函数定义为：

$$Q_\tau(Y_i \mid X_i) = \inf\{y : F_y(y \mid X_i) \geqslant \tau\}$$

② 也就是我们所说的对号(√)。——译者注

这种不对称加权可以产生一个最小化元，该最小化元能将条件分位数挑选出来（我们没法立刻看出这一事实，但是稍微花些工夫即可予以证明；参看 Koenker (2005)）。

当 X_i 是连续的或者维度过高时，条件分位数函数具有和条件期望函数一样的缺点：很难对条件分位数函数进行回归和概括。基于此，我们希望简化该函数中的未知参数，使得 X_i 中的每个元素对应于一个参数。通过用一个线性模型取代等式 (7.1.1) 中的 $q(X_i)$，我们可以达到这个目的，从而得到：

$$\beta_\tau = \arg\min_b \mathrm{E}[\rho_\tau(Y_i - X_i'b)] \tag{7.1.2}$$

分位数回归估计量 $\hat{\beta}_\tau$ 是等式 (7.1.2) 的样本估计值。在这种处理下，我们将最小化问题变成了一个线性规划的问题，从而变得相当容易（对计算机而言）求解。

正如最小二乘估计通过最小化均方误来拟合 Y_i 的线性模型一样，分位数回归通过使用不对称的损失函数（loss function）$\rho_\tau(u)$ 来拟合 Y_i 的线性模型。如果 $Q_\tau(Y_i|X_i)$ 事实上就是线性的，那么分位数回归的最小化元就可以找到这个函数（就像当条件期望函数为线性时，最小二乘回归也可以找到这个函数一样）。在 Koenker 和 Bassett(1978) 引入对分位数回归的讨论时，最初的分位数回归就是假设条件分位数函数是线性的。不过，随着研究的发展，对条件分位数函数的线性假设是不必要的：不论你是否相信，分位数回归总是有用的。

在转入对分位数回归进行更为一般的理论讨论之前，我们以对工资分布的研究为例来展示对这一计量工具的使用。使用分位数回归来考察工资分布的动机来源于劳动经济学家感兴趣的问题：给定类似于教育水平和工作经验等协变量后，工资不平等是如何变化的（Buchinsky，1994）。在 20 世纪 80 和 90 年代，在不同教育水平组成的群体之间的总体收入差距有显著提升（比如接受高等教育的工资溢价）。但是我们并不清楚在具有相同教育水平和工作经验的群体内部，工资水平是如何变化的。很多劳动经济学家认为，群体内部工资不平等的扩大为劳动力市场发生的根本性变化提供了强有力的经验证据，而这种现象不是诸如从属工会的工人百分比这类制度特征的变化所能解释的。

表 7.1 报告了使用 1980 年、1990 年和 2000 年人口普查数据进行分位数回归后估计出的教育水平的参数。用来构造这些估计值的模型控制了种族以及潜在的劳动力市场经验（定义为年龄减去教育年限再减去六）的二次项。0.5 分位数回归系数——也就是条件中位数——接近于表格最右边的最小二乘回归系数。比如，针对 1980 年人口普查数据得到的最小二乘估计值是 0.072，与使用相同数据得到的分位数回归估计值 0.068 相去不远。如果给定协方差后工资对数的分布是对称的，也就是说条件中位数等于条件均值，那我们应该预期上面提到的两个估计值相等。还值得注意的是，在 1980 年的人口普查数据中，不同分位数水平下得到的系数大致相同。多接受一年教育可以让工资中位数增加 6.8%，当条件工资分布的分位数水平变得更高或者更低时，多接受一年教育带来的该分位数水平上工资的

增加稍微高一些，分别是 0.074 和 0.070。虽然在 1980 年和 1990 年之间，教育水平对工资分位数的影响快速上升（比如对工资中位数的影响达到 0.106，相应的最小二乘估计值为 0.114），但是 1990 年人口普查中收入分位数表现出的模式还是相当稳定的。教育水平带来最大影响的分位数是上十分位数，其系数值为 0.137，而其他的分位数回归系数则处于 0.11 附近。

表 7.1　对 1980 年、1990 年、2000 年人口普查中教育程度进行分位数回归得到的系数

普查年份	观察值	Desc. Stats.		分位数回归					最小二乘估计	
		均值	标准误	0.1	0.25	0.5	0.75	0.9	系数	Root MSE
1980	65 023	6.4	0.67	0.074 (0.002)	0.068 (0.001)	0.070 (0.001)	0.079 (0.001)	0.072 (0.001)	0.072 (0.001)	0.63
1990	86 785	6.5	0.69	0.112 (0.003)	0.110 (0.001)	0.106 (0.001)	0.111 (0.001)	0.137 (0.001)	0.114 (0.001)	0.64
2000	97 397	6.5	0.75	0.092 (0.002)	0.105 (0.001)	0.111 (0.001)	0.120 (0.001)	0.157 (0.004)	0.114 (0.001)	0.69

注：本表改编自 Angrist、Chernozhukov 和 Fernandez-Val(2006)。该表报告了在对数工资模型中教育水平回报的分位数回归估计值，同时在最右端给出了最小二乘估计以便于对照。样本中包含的个体由在美国本土出生的年龄在 40—49 岁的白人和黑人男性。样本规模、工资对数的均值和标准差报告在本表的左侧。标准误报告在括号中。所有的模型都控制了种族和潜在的工作经验。在针对 2000 年人口普查数据的估计值中使用了抽样加权。

如果教育水平对工资的影响类似于一种“位移”，那么我们应该会看到在不同分位数下的系数是相同的。这里，“位移”指的是教育水平提高收入水平，工资分布的其他部分则水平移动（也即是说在具有相同教育水平的群体内部，工资差异不会发生改变）。比如，假设我们可以用经典线性回归模型来描述对数工资：

$$Y_i \sim N(X_i'\beta,\ \sigma_\varepsilon^2) \tag{7.1.3}$$

这里 $E[Y_i \mid X_i] = X_i'\beta$ 而且 $Y_i - X_i'\beta \equiv \varepsilon_i$ 是一个具有恒定方差 σ_ε^2 的正态分布变量。同方差性意味着不论大学毕业还是高中毕业，对数工资的条件分布的离散程度都不会发生变化。大学毕业生的与高中毕业生的分布展开方式是一样的。下式可以告诉我们线性同方差性模型对分位数来说意味着什么：

$$P[Y_i - X_i'\beta < \sigma_\varepsilon \Phi^{-1}(\tau) \mid X_i] = \tau$$

这里 $\Phi^{-1}(\tau)$ 是标准正态分布累积密度函数的反函数。由此得 $Q_\tau(Y_i \mid X_i) = X_i'\beta + \Phi^{-1}(\tau)$。换言之，如果不考虑变化的截距项 $\sigma_\varepsilon \Phi^{-1}(\tau)$，在每个分位数水平上，分位数回归系数都应该相同。表 7.1 中针对 1980 年和 1990 年人口普查数据得到的结果与这一特征性事实相去不远。

与 1980 年和 1990 年人口普查数据得到的简单模式相对比，在不同的分位数水平上，2000 年人口普查数据得到的分位数回归估计值显得大不一样，特别是处在分布右边的那部分收入水平。对下十分位数而言，多接受一年教育带来收入水

平增加 9.2%，对于中位数而言，该效应为 11.1%，对于上十分位数而言，该效应为 15.7%。因此，除了在 1980 年和 1990 年人口普查数据中显示的总体不平等水平在上升（从简单的描述性统计中也可以看出这个现象）之外，到 2000 年，不平等还随着教育水平的提高而提高（随着分位数水平的提高，教育对收入的影响也在增加，这意味着随着教育水平的提高，工资分布的离差也越大）。这一发展趋势成为劳动经济学家经常讨论的主题，该趋势是否表明劳动力市场出现了根本性的变化或者制度变迁（Autor，Katz and Kearney，2005；Lemieux，2008）。

用一个参数化的例子，我们可以看清分位数回归系数和条件方差之间的联系。具体而言，通过将异方差性加入经典正态分布回归模型（7.1.3）中，我们可以得到递增的分位数回归系数。假设：

$$Y_i \sim N(X_i'\beta,\ \sigma^2(X_i))$$

这里 $\sigma^2(X_i)=(\lambda'X_i)^2$，$\lambda$ 是个参数为正且满足 $\lambda'X_i>0$ 的向量（也能与 β 成比例，这样条件方差随着条件均值的增加而增加）①。那么：

$$\mathrm{P}[Y_i - X_i'\beta < (\lambda'X_i)\Phi^{-1}(\tau) \mid X_i] = \tau$$

其含义如下：

$$Q_\tau(Y_i \mid X_i) = X_i'\beta + (\lambda'X_i)\Phi^{-1}(\tau) = X_i'[\beta + \lambda\Phi^{-1}(\tau)] \qquad (7.1.4)$$

因此在各个分位数水平上，$\beta_\tau = \beta + \lambda\Phi^{-1}(\tau)$ 分位数回归系数都提高了。

将本节讨论的各部分放在一起，表 7.1 很简洁地总结出两个关于组内不平等水平发生变化的故事。首先，从 2000 年人口普查中得到的结果指出随着教育水平的提高，不平等水平急剧上升。但是，这种不平等的上升是不对称的，而且在工资分布的右半部分表现得更加显著。其次，这种不平等的发展是一种新趋势。在 1980 年和 1990 年，教育水平以一种近似于简单“位移”的方式对工资分布产生影响。②

7.1.1 删失分位数回归

分位数回归允许我们在 Y_i 的分布中的一段被隐藏时仍然可以去考察 Y_i 的条

① 参看 Card 和 Lemieux(1996)给出的一个经验研究例子，在这个例子中，回归模型具有此类异方差的特点。Koenker 和 Portnoy(1996)将其称为线性成比例位移模型(linear location-scale model)。

② 在渐近分位数回归标准误的公式中，人们假设了一个线性的条件分位数函数，该公式是这样的：

$$\tau(1-\tau)\mathrm{E}[f_{u_\tau}(0 \mid X_i)X_iX_i']^{-1}\mathrm{E}[X_iX_i']\mathrm{E}[f_{u_\tau}(0 \mid X_i)X_iX_i']^{-1}$$

这里 $f_{u_\tau}(0|X_i)$是分位数回归残差在零处的条件密度函数。如果残差是同方差的，那么这个函数可以简化为 $\frac{\tau(1-\tau)}{f^2_{u_\tau}(0)}\mathrm{E}[X_iX_i']^{-1}$，这里 $f^2_{u_\tau}(0)$是非条件残差的密度函数的平方。Angrist、Chernozhukov 和 Fernandez-Val(2006)给出了在更一般情况下条件分位数函数为非线性时的公式。

件分布的特征。假如你所得到的数据形式如下：

$$Y_{i,obs} = Y_i \cdot 1[Y_i < c] + c \cdot 1[Y_i \geqslant c] \tag{7.1.5}$$

这里 $Y_{i,obs}$ 是你看到的观察值，而 Y_i 是你本来应该看到的观察值。变量 $Y_{i,obs}$ 是 Y_i 的删失(censored)数据——出于保密的原因或者收集某些数据过于困难、耗时，$Y_{i,obs}$ 中有关 Y_i 的信息受到限制。比如在 CPS 中，为了保护高收入者的隐私，他们收入的前几位数被重新编码。这意味着该收入水平上的收入数据都不是真实的。持续性数据(duration data)也是删失的：在一项研究失业保险对雇佣期限产生的影响的研究中，我们只能跟踪失业时间在 40 周以内的受访者。任何失业超过 40 周的人，其失业期限都是删失的，只有 40 周。注意到类似于我们在 3.4.2 节讨论过的工作时间或医疗费用等有限被解释变量都不是删失数据；它们因为自己的特有性质而取值为零，就像表示就业状态的虚拟变量只取 0 和 1 两个值一样。

在处理删失被解释变量(censored dependent variables)时，可以使用分位数回归针对低于某个删失点(假设删失是从上而下开始的)的数据估计协变量对条件分位数的影响。这意味着处在删失点之上的收入数据，比如说处于中位数之上的数据，对中位数无影响。因此，如果美国当期人口调查中的编码加密行为只影响到很少的人(通常这个情况是真实的)，那么出现的删失状况不影响对条件中位数的估计，甚至不影响 $\tau=0.75$ 时的 β_τ。类似的，如果给定 X_i 的所有取值，小于 10%的数据是删失的，那么直到 $\tau=0.90$ 时的 β_τ 都是可以被我们估计的。从另一个角度出发，你可以对样本做出限制，只考察 $Q_\tau(Y_i \mid X_i)$ 低于(如果删失现象是自下而上的，那么 $Y_{i,obs} = Y_i \cdot 1[Y_i > c] + c \cdot 1[Y_i \leqslant c]$)$c$ 的那些样本的协变量 X_i 的取值。

Powell(1986)正式地阐明了对删失数据做分位数回归的思想。由于我们可能不知道哪个条件分位数处于删失点之下，所以(比如继续以高收入者收入被加密的故事为例)，Powell 建议我们考虑下式：

$$Q_\tau(Y_i \mid X_i) = \min(c,\ X_i'\beta_\tau^c)$$

参数向量 β_τ^c 为下式的解：

$$\beta_\tau^c \equiv \arg\min_b \ E\{1[X_i'b < c] \cdot \rho_\tau(Y_i - X_i'b)\} \tag{7.1.6}$$

换言之，我们针对 $X_i'\beta_\tau^c < c$ 的 X_i 求解分位数回归中的最小化问题。(在实际中，我们最小化等式(7.1.6)的样本值。)只要有足够多的未删失数据，我们求得的估计值就给出了原本在不存在删失数据时才能求得的分位数回归函数(假设条件分位数函数是线性的)。而且，如果你要回归的条件分位数都处于删失点之下，那么你得到的就是常规分位数回归。

等式(7.1.6)的样本值将不再是一个线性规划问题，但是 Buchinsky(1994)给出了一个简单的迭代线性规划算法(iterated linear programming algorithm)，看上去该算法是可以用来进行求解的。该迭代算法如下运行：首先在忽略删失问题的情况下估计 β_τ^c；之后找到 $X_i'\beta_\tau^c < c$ 的数据集；然后用这些数据集进行分位数回归，

如此往复。这个算法不一定收敛，但是在实际中却往往可以收敛。可以通过自助法来求得标准误。Buchinsky(1994)使用这种方法估计了教育水平对具有很高经验的工人所产生的影响，而这些工人的收入可能已经超过美国当期人口调查中收入被加密的那个收入水平。对删失问题进行的调整倾向于提高此类个体中教育水平对收入的影响。①

7.1.2 分位数回归的近似性质*

给定教育水平，对数工资水平的条件分位数函数不一定是线性的，这时我们在一开始对分位数回归进行讨论时设定的假设就不再成立了。不过幸运的是，我们可以将分位数回归理解为一种最小均方误意义下的对条件分位数函数的线性近似，只不过相比于回归的条件期望函数定理，在这个例子中的近似过程显得有点复杂，求解过程也比较难。对任何的分位数指标 $\tau\in(0,1)$ 而言，定义分位数回归的设定偏误为：

$$\Delta_\tau(X_i,\beta_\tau)\equiv X_i'\beta-Q_\tau(Y_i\mid X_i)$$

正如 Angrist、Chernozhukov 和 Fernandez-Val(2006)在下面的定理中指出的，通过最小化模型设定偏误 $\Delta_\tau^2(X_i,\beta)$ 的加权平均值，我们可以找到总体分位数回归向量：

定理 7.1.1：分位数回归的近似性质(Quantile Regression Approximation)。

假设(i)条件密度 $f_y(y|X_i)$ 确定存在，(ii)$\mathrm{E}[Y_i]$、$\mathrm{E}[Q_\tau(Y_i|X_i)]$和 $\mathrm{E}\|X_i\|$ 是有限的，而且(iii)β_τ 是等式(7.1.2)的唯一解。则：

$$\beta_\tau=\arg\min_b \mathrm{E}[w_\tau(X_i,b)\cdot\Delta_\tau^2(X_i,b)] \tag{7.1.7}$$

其中，

$$\begin{aligned}w_\tau(X_i,b)&=\int_0^1(1-u)\cdot f_{\varepsilon(\tau)}(u\Delta_\tau(X_i,b)\mid X_i)\mathrm{d}u\\&=\int_0^1(1-u)\cdot f_y(u\cdot X_i'b+(1-u)\cdot Q_\tau(Y_i\mid X_i)\mid X_i)\mathrm{d}u\\&\geqslant 0\end{aligned}$$

$\varepsilon_i(\tau)$是分位数回归残差(quantile-specific residual)，

$$\varepsilon_i(\tau)\equiv Y_i-Q_\tau(Y_i\mid X_i)$$

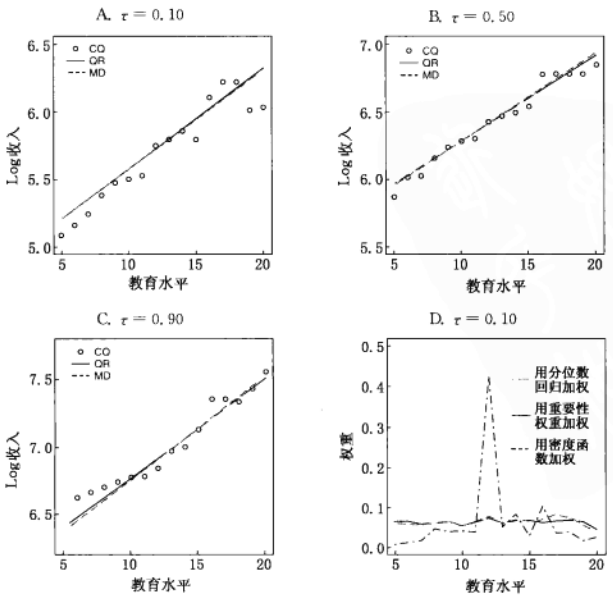
在 $\varepsilon_i(\tau)=e$ 处其条件密度为 $f_{\varepsilon(\tau)}(e|X_i)$。此外，当 Y_i 是平滑的条件密度函数时，对于在 β_τ 邻域内的 β，我们有：

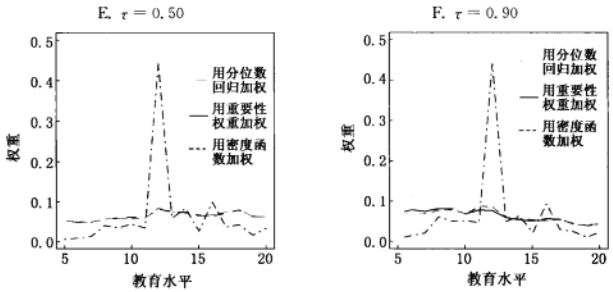
$$w_\tau(X_i,\beta)\approx 1/2\cdot f_y(Q_\tau(Y_i\mid X_i)\mid X_i) \tag{7.1.8}$$

① 参看 Buchinsky 和 Hahn(1998)以及 Chernozhukov 和 Hong(2002)为了得到更好的理论性质而给出的更加复杂的估计值。

分位数回归的近似定理看上去很复杂，但是其基本意思还是很容易理解的。类似于最小二乘回归是对 $E[Y_i \mid X_i]$ 的近似，我们可以将分位数回归看作是对 $Q_\tau(Y_i \mid X_i)$ 的近似。在最小二乘回归中的加权函数是 X_i 出现的概率，记为 $P(X_i)$。在分位数回归中，加权函数是 $w_\tau(X_i, \beta_\tau) \cdot P(X_i)$，要比 $P(X_i)$ 更为复杂(由于等式(7.1.7)中的期望值是在 X_i 的分布函数上求得的，所以 X_i 出现的概率是分位数回归中加权函数的一部分)。$w_\tau(X_i, \beta_\tau)$ 中包含了分位数回归向量 β_τ，但是可以将分位数回归向量从其中分离出来，使之成为只是 X_i 的函数(具体细节参看 Angrist，Chernozhukov 和 Fernandez-Val(2006))。在任何情况下，在条件分位数函数的邻域中，分位数回归权重都近似地与 Y_i 的密度函数成比例。

可用图 7.1 来阐明分位数回归的近似性质，在使用 1980 年人口普查数据，给定最高完成的学位水平下，该图绘出了对数工资的条件分位数函数。这里我们利用教育水平的离散分布和人口普查的大样本数据这两个特征，通过计算每个教育水平下工资水平的分位数，来估计非参数的条件分位数函数。图 A—C 分别在 0.10、0.50 和 0.90 三个分位数水平下绘出了用非参数方法得到的对 $Q_\tau(Y_i \mid X_i)$ 的线性分位数回归结果，这里协变量 X_i 只包括受教育水平和一个常数。图中的圆圈则表示针对不同教育水平的个体分别估计出的条件分位数函数，而分位数回归曲线则是实线。这幅图表现了线性分位数回归可以如何近似条件分位数函数。





注：本图来自 Angrist，Chernozhukov 和 Fernandez-Val(2006)。这里使用 1980 年人口普查数据，在给定最高学历已完成的情况下，用另外的方法估计了工资对数的条件分位数函数，同时说明了在估计中用到的加权函数。其中图 A—C 分别报告了非参数估计、分位数回归以及最短距离估计法下求得的估计值，对应的 τ 分别是 0.1、0.5、0.9。用 D—F 给出了分位数回归的加权函数，对这些加权函数的解释请见正文。

图 7.1　分位数回归近似性质

将分位数回归与使用 $P(X_i)$ 做权重拟合得到的结果进行比较也是十分有趣的。这里用 X_i 出现的概率做权重，对条件分位数函数进行拟合的方法类似于我们对条件期望函数做最小二乘回归。使用 X_i 出现的概率做权重的方法由 Chamberlain(1994)提出。Chamberlain 通过最小化距离函数(minimum distance)得到的估计值是下面这个向量 $\hat{\beta}_\tau$ 的样本值，它是通过求解下式得到的：

$$\begin{aligned}\beta_\tau &= \arg\min_b \mathrm{E}[(Q_\tau(Y_i \mid X_i) - X_i'b)^2] \\ &= \arg\min_b \mathrm{E}[\Delta_\tau^2(X_i, b)]\end{aligned}$$

换言之，$\hat{\beta}_\tau$ 是对 $Q_\tau(Y_i \mid X_i)$ 线性回归后在 X_i 处斜率的加权值，这里权重取 X_i 出现的概率。相比之下，分位数回归结果只对数据进行一步处理，最小化距离函数的方法则依赖于首先对 $Q_\tau(Y_i \mid X_i)$ 进行非参数的一致估计。

图 7.1 用虚线绘出了最小化距离下得到的拟合值。分位数回归和最小化距离下得到的拟合曲线十分接近，但是由于在分位数回归的拟合中使用了 $w_\tau(X_i, \beta_\tau)$ 做加权函数，所以两者之间并不会完全等同。使用 $w_\tau(X_i, \beta_\tau)$ 做加权函数的方法强调了当 Y_i 在条件分位数函数的邻域中密集分布时，在 X_i 的取值处进行回归的质量。在图 7.1 中，D—F 针对 X_i 绘出了分位数回归的总体权重函数 $w_\tau(X_i, \beta_\tau) \cdot P(X_i)$。这三幅图还绘出了 $w_\tau(X_i, \beta_\tau)$ 的估计值，记为“重要权重”以及它们的密度函数近似值 $1/2 \cdot f_Y(Q_\tau(Y_i \mid X_i) \mid X_i)$。重要权重和密度函数权重之间很相似，而且也相当平坦。总体的加权函数看上去很像教育水平的出现概率，因此对教育水平为 12 年和 16 年的那些个体赋予了最大权重。

7.1.3 微妙之处

条件分位数的语言是非常微妙的。有时候我们讨论的是关于“在中位数上的分位数回归系数”，或者关于“在下十分位数上产生的影响”。但是，重要的是要记住分位数回归的系数告诉我们的是对分布的影响，而不是对个体的影响。比如，如果我们发现一个培训项目可以提高工资分布的下十分位数，这并不意味着那些原本贫穷的人(如果那些在下十分位数处没有受到培训的人)变得不那么贫穷了。它只是意味着在下十分位数处，接受培训的人变得不那么贫穷了。

使一部分穷人变得更加富裕以及改变个体在穷人中的相对富裕程度，两者之间的差别是很微妙的。这个差别，也与我们是否认为一个干扰可以保持个体在收入分布中的排序有关。如果一个扰动不影响排序，那么下十分位数的提高意味着所有穷人都变富裕了，因为扰动不影响排序意味着人们收入水平之间的相对差距是不变的。否则，我们只能说穷人——定义为收入分布中处在最后百分之十的人，无论他是谁——其状况都变好了。我们在 7.2 节对此进行简短的讨论。

另外一个微妙之处在于从条件分位数到边际分位数的转换。在条件分位数和边际分位数之间存在的联系，允许我们考察分位数回归系数的改变对总体不平等水平的影响。比如假设教育水平对分位数系数的影响变得更加离散，超过了我们对 2000 年人口普查数据作出的结论。那么这对上十分位数和下十分位数处的个体的工资之比有什么影响？或者换个角度，我们可以问：在总体不平等(比如总上十分位数和下十分位数的收入比来度量)的上升中，有多少可以被组内不平等的上升所解释？这里用分位数回归系数的离散程度来度量组内不平等水平。对这类问题的回答显然极为棘手。困难就在于要处理这一事实：要对所有的条件分位数找到一个特定的边际分位数(Machado and Mata, 2005)。具体而言，$Q_\tau(Y_i \mid X_i) = X_i'\beta_\tau$ 并不意味着 $Q_\tau(Y_i) = Q_\tau(X_i)'\beta_\tau$。相比之下期望算子显得易于处理，如果 $\mathrm{E}(Y_i \mid X_i) = X_i'\beta$，那么迭代期望，我们有 $\mathrm{E}(Y_i) = \mathrm{E}(X_i)'\beta$。

1. 求解边际分位数*

为了更为正式地表明条件分位数和边际分位数两者之间的联系，我们假设条件分位数函数是线性的，这样就会有 $Q_\tau(Y_i \mid X_i) = X_i'\beta_\tau$。令 $F_y(y \mid X_i) \equiv \mathrm{P}[Y_i < y \mid X_i]$ 为给定 X_i 下 Y_i 的条件累积密度函数，其边际分布为 $F_y(y) = \mathrm{P}[Y_i < y]$。累积密度函数及其反函数之下的关系由下式给出：

$$\int_0^1 1[F_y^{-1}(\tau \mid X_i) < y]\mathrm{d}\tau = F_y(y \mid X_i) \tag{7.1.9}$$

其中，$F_Y^{-1}(\tau \mid X_i)$也是条件分位数函数 $Q_\tau(Y_i \mid X_i) = X_i'\beta_\tau$。

换句话说，在给定 X_i 下小于 y 的那些总体的比例等于小于 y 的条件分位数的比例。[①]用线性模型替换积分号里面的条件分位数函数，可得：

① 比如说，如果 y 是条件中位数，那么 $F_y(y \mid X_i) = 0.5$，有一般的条件分位数小于 y。通过改变变量所在的公式，我们可以证明关系(7.1.9)。

$$F_y(y \mid X_i) = \int_0^1 1[X_i'\beta_\tau < y]\mathrm{d}\tau$$

接下来，我们使用迭代期望律来求解边际分布函数 $F_y(y)$：

$$F_y(y) = \mathrm{E}[\int_0^1 1[X_i'\beta_\tau < y]\mathrm{d}\tau] \tag{7.1.10}$$

最后，通过对函数 $F_y(y)$进行转置，我们可以求得边际分位数，也就是 $Q_\tau(Y_i)$，其中，$\tau \in (0, 1)$：

$$Q_\tau(Y_i) = \inf\{y: F_y(y) \geqslant \tau\}$$

用加号代替等式(7.1.10)中的积分和期望，我们就能得到边际分布的估计值。这里的加号针对来自分位数回归的估计值进行求和，分位数水平可以是 0.01。如果样本规模是 N，那么这个等式就变为：

$$\hat{F}_y(y) = N^{-1}\sum_i (1/100)\sum_{\tau=0}^{\tau=1} 1[X_i'\hat{\beta}_\tau < y]$$

相应的边际分位数估计值就是 $\hat{F}_y(y)$的反函数。

在实际操作中这种方法存在很多困难。一方面，你不得不进行大量的分位数回归。另一方面，这里的渐进分布理论十分复杂(尽管这不是不可解的；比如参看 Chernozhukov, Fernandez 和 Melly(2005))。如何将条件分位数简化为边际分位数是一个非常活跃的研究领域。Gosling、Machin 和 Meghir(2000)以及 Machado 和 Mata(2005)是第一批考察如何从条件分位数求得边际分位数的经验研究。当我们在分位数回归模型中主要关心的变量是类似于表征处理状态的虚拟变量，而其他回归元为控制变量时，用倾向得分模型中的加权方法，我们可以估计出对边际分布的影响。Firpo(2007)给出了一个具有外生性的例子，而 Frölich 和 Melly (2007)提供了能够在内生处理效应模型中使用的边际化方法，我们在下一节讨论如何求解内生处理效应模型。

7.2 对分位数处理效应的工具变量估计

在之前讨论的那个有关 42 000 美元的例子中，我们考虑的问题是是否能为任何一个回归估计值赋予一个因果解释。但是对于分位数回归而言，这将不再成立。假设我们感兴趣的是估计培训项目对收入的影响。最小二乘回归估计值度量的是该培训项目对平均收入的影响，但是分位数回归估计则可以用来度量该项目对收入中位数的影响。在这两个例子中，我们必须考虑估计出的效应的效果是不是会受到遗漏变量偏误的影响。

在这里，我们还是要用工具变量法来解决遗漏变量偏误的问题，但是针对分位数回归的工具变量法目前尚在发展之中，而且也不如传统的两阶段最小二乘回归那样富于变化。在这里，我们讨论的方法主要针对使用二元工具变量捕捉二元变

量对分位数的因果效应(也就是处理效应)。Abadie、Angrist 和 Imbens(2002)给出了使用工具变量估计分位数处理效应(quantile treatment effect,简写为 QTE)的方法,该方法所基于的假设与局部平均处理效应框架下使用的假设相同。其估计结果是一个加权估计值,被加权的则是依从工具变量者表现出的分位数处理效应。①

我们对分位数处理效应估计值的讨论建立在一个条件分位数的可加模型上,所以在这个带有协变量的模型中,我们只估计单个处理效应。由此得到的估计值简化了 Koenker 和 Bassett(1978)在不存在工具变量时得到的线性分位数回归。因此,分位数处理效应和分位数回归之间的关系就类似于传统的两阶段最小二乘回归和最小二乘回归之间的关系,只不过这里我们感兴趣的回归元是个虚拟变量。

我们感兴趣的参数定义如下。对于 $\tau \in (0, 1)$,我们假设存在 α_τ 和 β_τ,满足:

$$Q_\tau(Y_i \mid X_i, D_i, D_{1i} > D_{0i}) = \alpha_\tau D_i + X_i'\beta_\tau \tag{7.2.1}$$

这里 $Q_\tau(Y_i \mid X_i, D_i, D_{1i} > D_{0i})$ 表示给定依从工具变量者的 X_i 和 D_i,Y_i 的 τ 分位数。这样一来,α_τ 和 β_τ 就是依从工具变量者的分位数回归系数。

回忆一下,给定 X_i 和 $D_{1i} > D_{0i}$,D_i 独立于潜在结果,这和我们在方程(4.5.2)中讨论的一样。因此该模型中的参数 α_τ 给出了给定 X_i 后依从工具变量者在 Y_{1i} 和 Y_{0i} 上的分位数之差。换言之:

$$Q_\tau(Y_{1i} \mid X_i, D_{1i} > D_{0i}) - Q_\tau(Y_{0i} \mid X_i, D_{1i} > D_{0i}) = \alpha_\tau \tag{7.2.2}$$

举个例子,上式可以告诉我们依从工具变量者收入的中位数或者下十分位数。注意到参数 α_τ 并未告诉我们处理是否会改变 Y_{1i} 和 Y_{0i} 的分位数的无条件分布。为此,我们要使用类似于在第 7.1.3 节中讨论的办法,对分位数回归结果进行积分。

同样值得强调的是,α_τ 并不是个体处理效应($Y_{1i} - Y_{0i}$)的条件分位数。比如说,你可能想知道中位数处理效应是否是正的。不幸的是,如果不作出诸如不变排序这样很强的假设,我们很难回答此类问题②。即使在完美依从工具变量的随机实验中,我们可也能无法求得($Y_{1i} - Y_{0i}$)的分布。虽然均值的差等于差的均值,但是由于我们不可能在同一个人身上同时看到 Y_{1i} 和 Y_{0i} 两个值,所以 $Y_{1i} - Y_{0i}$ 的分布函数中的其他特征则无法观察到。不过,由于对社会福利的比较往往只要求比较 Y_{1i} 和 Y_{0i} 的分布函数之间的差别,而不要求去比较两者差别的分布(例如,可参看 Atkinson(1970)),所以分布函数之间的差别往往要比处理效应之差的分布更为重要,这对应用计量经济学家而言是个好消息。因此,不借助分位数,我们也可以达到社会福利比较所要求的研究目的。当评估一个雇佣计划时,我们更希望看到该项目提高了平均的就业率。换言之,如果 Y_{1i} 的平均值大于 Y_{0i} 的平均值,那我们

① 其他的方法可以从 Chenozhukov 和 Hansen(2005)中找到,该文允许任何类型的回归元(即不仅针对虚拟变量),但是他们却引入了一个不变排序(rank-invariance)的假设,在分位数处理效应的框架中,这个假设是没有必要的。

② 在这种情况下,不变排序(rank-invariance)意味着 Y_{1i} 和 Y_{0i} 是通过一个不可逆的函数联系在一起的。比如可以参看 Heckman、Smith 和 Clements(1997)。

会很高兴地认为该项目的目的达到了。由于一个好的项目必然会让更多的人受益，所以对谁得到了工作(即 $Y_{1i}-Y_{0i}=1$)和谁失去了工作(即 $Y_{1i}-Y_{0i}=-1$)的讨论则处于第二感兴趣的位置。

7.2.1 分位数处理效应估计值

启发我们得到分位数估计量的想法来自于如下观察：通过对依从工具变量者总体进行分位数回归，我们可以估计出依从工具变量者的分位数回归系数。在某个给定的数据集中，我们无法列出依从工具变量的个体是哪些，但是正如我们在第4.5.2节中看到的，可以使用 Abadie Kappa 定理去发现它们。具体而言：

$$\begin{aligned}(\alpha_\tau,\ \beta_\tau) &= \arg\min_{a,\,b} \mathrm{E}\{\rho_\tau(Y_i-aD_i-X_i'b)\mid D_{1i}>D_{0i}\} \\ &= \arg\min_{a,\,b} \mathrm{E}\{\kappa_i\rho_\tau(Y_i-aD_i-X_i'b)\}\end{aligned} \tag{7.2.3}$$

其中，

$$\kappa_i = 1-\frac{D_i(1-Z_i)}{1-\mathrm{P}(Z_i=1\mid X_i)}-\frac{(1-D_i)Z_i}{\mathrm{P}(Z_i=1\mid X_i)}$$

和在第4.5.2节中讨论的一样。而分位数处理效应估计值正是等式(7.2.3)对应的样本值。

在实际构造非位数处理效应时，有一系列的现实问题需要我们考虑。首先，我们要对 κ_i 进行估计，依照相应的渐进分布定理，我们还应该从第一步估计中得到抽样方差。通过使用非参数方法估计出 κ_i，Abadie、Angrist 和 Imbens(2002)导出了等式(7.2.3)的有限样本分布。但是在实际中，通过自助法(bootstrap)而不是渐进分布的公式，我们可以更容易地完成上述整个过程。

其次，当 $D_i\neq Z_i$ 时，κ_i 是负的。用 Kappa 函数加权的分位数回归最小化元因而成为非凸的，而且常规的分位数回归估计值不同，这时最小化元不存在线性规划解。通过最小化下式，这个问题可以得到解决：

$$\mathrm{E}\{\mathrm{E}[\kappa_i\mid Y_i,\ D_i,\ X_i]\rho_\tau(Y_i-aD_i-X_i'b)\} \tag{7.2.4}$$

通过对等式(7.2.3)求重复期望，我们可以得到这一最小化元。在实际中，等式(7.2.3)和等式(7.2.4)之间的区别在于下面这一项：

$$\mathrm{E}[\kappa_i\mid Y_i,\ D_i,\ X_i]=\mathrm{P}[D_{1i}>D_{0i}\mid Y_i,\ D_i,\ X_i]$$

这一项是一个概率，因此介于0和1之间。[①]更进一步的简化来自下面这个事实：

$$\begin{aligned}\mathrm{E}[\kappa_i\mid Y_i,\ D_i,\ X_i] = 1&-\frac{D_i(1-\mathrm{E}[Z_i\mid Y_i,\ D_i=1,\ X_i])}{1-\mathrm{P}(Z_i=1\mid X_i)} \\ &-\frac{(1-D_i)\mathrm{E}[Z_i\mid Y_i,\ D_i=0,\ X_i]}{\mathrm{P}(Z_i=1\mid X_i)}\end{aligned} \tag{7.2.5}$$

① 因为 κ_i 的意义在于找到依从工具变量者，所以 κ_i 的期望是一个概率。Abadie、Angrist 和 Imbens(2002)的引理3.2中正式给出了这一结论。

依上述程序，Angrist(2001)分别在 $D_i=0$ 和 $D_i=1$ 的子样本中对 probit 模型的 $E[Z_i|Y_i, D_i, X_i]$进行了分位数处理效应估计，先使用等式(7.2.5)构造 $E[\kappa_i|Y_i, D_i, X_i]$，然后将任何处在单位区间之外的对 $E[\kappa_i|Y_i, D_i, X_i]$的估计值整理出来。使用 Stata 中的 qreg 函数，我们可以将得到的对 $E[\kappa_i|Y_i, D_i, X_i]$的第一步非负估计作为权重代入方程来构造第二部中的加权分位数回归估计值。①

1. 估计培训为受训者收入分位数的影响

工作培训合伙法案(Job Training Partnership Act，简写为 JTPA)是一个大规模的联邦项目，它为 20 世纪 80 年代那些遭受失业痛苦的美国工人们提供资助性培训。联邦政府在 649 个城市中提供了 JTPA 服务，这些城市也被称为服务提供区(Service Delivery Areas，即 SDAS)，它们遍布全国。最初对 JTPA 服务如何影响劳动力市场的研究基于一个包含男性和女性的样本，该样本的数据包括了在随机分配给 JTPA 服务后至少 30 个月的收入状况，该收入状况来自于州立失业保险中的收入记录，或者是两个随后进行的调研。②在这个样本中，有 30 个月的收入信息的成年男性共有 5 102 人。

在我们使用的符号中，Y_i 表示 30 个月的收入，D_i 表示是否参与了 JTPA 服务，而 Z_i 则表示对 JTPA 服务进行的随机分配结果。在大部分社会科学实验中，同时在大量对药品和治疗方法进行的随机实验中，都存在一个共有的关键性特征：项目参与人可能不愿意接受实验提供的干预。在 JTPA 中，并不强迫人们接受随机分配的服务。因此，虽然对由财政补贴的服务是随机分配的，但是只有 60%的人最终接受了 JTPA 服务。因此，对处理的接受中存在自选择的问题，而且这种自选择问题可能和潜在的结果有关系。从另一个方面看，正如我们在 4.4.3 节中的讨论，随机实验为接受实验的那部分人提供了一个很好的工具变量，因为随机分配的处理状态和最后接受的处理状态之间显然有着密切的联系，而且对处理状态的分配显然是独立于潜在结果的。更进一步，由于在控制组中几乎没有人获得JTPA 服务，所以我们可以将针对依从工具变量的个体计算出的效应解释为被处理的效应(类似于我们在第 4.4.3 节进行的讨论，局部平均处理效应等于依从工具变量者和从不接受工具变量干预的个体上表现出的处理效应)。

既然在全国性的 JTPA 研究中，对培训的提供是随机分配的，那么我们不需要对协变量(X_i)对服从工具变量者造成的影响进行估计。但即使在类似于 JTPA 的随机实验中，习惯上也要控制协变量以降低处理状况和个体特点之间的联系，提高

① 按照下面的方法，一步步进行：
(1) 分别在 $D_i=0$ 和 $=1$ 的子样本中对 Z_i 做关于 Y_i 和 X_i 的 probit 回归，保留拟合值。
(2) 在全部样本上对 Z_i 做关于 X_i 的 probit 回归，保留拟合值。
(3) 将上面得到的两组拟合值代入等式(7.2.5)中来构造 $E[\kappa_i|Y_i, D_i, X_i]$。将任何小于 0 的数设为 0，将任何大于 1 的数设为 1。
(4) 在分位数回归中将由此得到的 kappa 函数当作权重使用。
(5) 用自助法模拟整个程序，以构造标准误。

② 参看 Bloom 等(1997)和 Orr 等(1996)。

估计精度(见第 2 章)。在这样的研究中使用的协变量来自于在 JTPA 项目执行过程进行的基本度量：表征黑人和西班牙裔的虚拟变量，表征是否高中毕业(将大学毕业也包含在内)的虚拟变量，表征婚姻状态的虚拟变量，将年龄按照每五年一组进行分组后得到的虚拟变量以及表征在该随机分配项目开始前的 13 周内申请人是否工作的虚拟变量。在该研究项目中还包括了表征 JTPA 服务特点的虚拟变量(课堂培训、在职培训、求职帮助等)以及一个表征收入数据是否来自第二次后续调研的虚拟变量。由于这些协变量大部分都度量的是被试个体的地域和社会经济状况，所以我们认为分位数回归分析告诉我们的是在由地域和社会经济特征决定的组群中 JTPA 实验如何影响了收入的分布状态。

作为一个分析基准，在表 7.1 中的第 1 列报告了最小二乘回归和通常的工具变量(两阶段最小二乘回归)回归估计出的培训项目对成年男性的影响。最小二乘回归的估计值是精确的 3 754 美元。该系数是在用 D_i 和 X_i 对 Y_i 做回归后得到的 D_i 的系数。这些估计值忽略了受训者存在的自选择问题。在表 7.1 中的两阶段最小二乘估计用对处理状态的随机分配 Z_i 作为 D_i 的工具变量。两阶段最小二乘估计值是 1 593 美元，标准误是 895 美元，比最小二乘估计中得到的参数小了一半还多。

分位数回归估计表明，相对于高于中位数的受训者，低于中位数的那些受训者在不同分位数水平下的状态之间差别很大。从表 7.1 中最右边的那几列就可看出来，这些列报告了分别在 0.15、0.25、0.5 和 0.75 分位数水平下的分位数回归结果。具体而言，0.85 分位数上的受训者的收入大约比相应的未受培训者的收入高出 13 个百分点，而 0.15 分位数则高出 136 个百分点。和表中的 OLS 估计值一样，这些分位数回归系数并不必然具有因果性解释。而且，它们给出了受培训者和未受培训者收入分布之间的描述性比较。

表 7.1　来自 JTPA 实验的分位数回归估计值和分位数处理效应

A. 最小二乘与分位数回归估计值

变　量	最小二乘	分位数				
		0.15	0.25	0.50	0.75	0.85
培训效应	3 754	1 187	2 510	4 420	4 678	4 806
	(536)	(205)	(356)	(651)	(937)	(1 055)
受培训影响百分比	21.2	135.6	75.2	34.5	17.2	13.4
高中或大学毕业	4 015	339	1 280	3 665	6 045	6 224
	(571)	(186)	(305)	(618)	(1 029)	(1 170)
黑人	−2 354	−134	−500	−2 084	−3 576	−3 609
	(626)	(194)	(324)	(684)	(1 087)	(1 331)
西班牙裔	251	91	278	925	−877	−85
	(883)	(315)	(512)	(1 066)	(1 769)	(2 047)
已婚	6 546	587	1 964	7 113	10 073	11 062
	(629)	(222)	(427)	(839)	(1 046)	(1 093)

（续表）

A. 最小二乘与分位数回归估计值

变量	最小二乘	分位数				
		0.15	0.25	0.50	0.75	0.85
每年工作时间小于13周	−6 582	−1 090	−3 097	−7 610	−9 834	−9 951
	(566)	(190)	(339)	(665)	(1 000)	(1 099)
常数项	9 811	−216	365	6 110	14 874	21 527
	(1 541)	(468)	(765)	(1 403)	(2 134)	(3 896)

B. 两阶段最小二乘与QTE估计值

变量	两阶段最小二乘	分位数				
		0.15	0.25	0.50	0.75	0.85
培训效应	1 593	121	702	1 544	3 131	3 378
	(895)	(475)	(670)	(1 073)	(1 376)	(1 811)
受培训影响百分比	8.55	5.19	12.0	9.64	10.7	9.02
高中或大学毕业	4 075	714	1 752	4 024	5 392	5 954
	(573)	(429)	(644)	(940)	(1 441)	(1 783)
黑人	−2 349	−171	−377	−2 656	−4 182	−3 523
	(625)	(439)	(626)	(1 136)	(1 587)	(1 867)
西班牙裔	335	328	1 476	1 499	379	1 023
	(888)	(757)	(1 128)	(1 390)	(2 294)	(2 427)
已婚	6 647	1 564	3 190	7 683	9 509	10 185
	(627)	(596)	(865)	(1 202)	(1 430)	(1 525)
每年工作时间小于13周	−6 575	−1 932	−4 195	−7 009	−9 289	−9 078
	(567)	(442)	(664)	(1 040)	(1 420)	(1 596)
常数项	10 641	−134	1 049	7 689	14 901	22 412
	(1 569)	(1 116)	(1 655)	(2 361)	(3 292)	(7 655)

注：该表报告了培训对收入的影响效应的最小二乘回归、分位数回归、两阶段最小二乘回归和分位数处理效应估计值（改编自 Abadie、Angrist 和 Imbens(2002)）。该样本包括 5 102 个成年男性。分配状况作为表 B 中培训情况的工具变量来使用。除了在表中所显示的协变量之外，所有模型都包括推荐服务计划和年龄群体的虚拟变量，以及用以表明收入数据来自第二次调研的虚拟变量。显著标准误在括号中报告。

对中位数收入水平而言，分位数处理效应估计出的接受培训带来的效应与两阶段最小二乘估计得到的结果在数量级上类似，但是不如两阶段最小二乘估计结果精确。另一方面，分位数处理效应估计值显示出一种与分位数回归估计值截然不同的模式。在低分位数上的估计值的确比相应的分位数回归估计值要小，而且在绝对值上也是小的。比如说，在 0.15 分位数水平上，分位数处理效应的估计值为 121 美元，而相应的分位数回归估计值是 1 187 美元。相似的是，在 0.25 分位数水平上，分位数处理效应的估计值为 702 美元，而相应的分位数回归估计值是 2 510 美元。然而，和低分位数上的结果不同，在高于中位数的收入水平上，分位数

处理效应的估计值要更大，而且统计上也更显著(尽管仍然比相应的分位数回归估计值小)。

JTPA 成年男性培训并没有提高收入分布处在较低分位数上人的境况，这是从这一分析中得到的最有趣的发现。这表明在表 7.1 的上半部分的分位数回归估计值受到了正的选择偏误的干扰。对这一发现的一个回应可能是这样的：极少有 JTPA 的申请者是非常幸运的，所以人们主要关心该项目从整体上看确实帮助了很多项目申请人，但是人们很少关心在按照地域或者社会经济状况确定的群体内部，培训项目对分配的影响。不过，在收入的顶端分位数处，参与全国性 JTPA 项目对成年男性的影响较高也是合理的。虽然提高高收入群体的收入分布并不是政策制定者的一个优先目标。

▶8

非标准的标准误问题

我们是正常的。我再重复一遍，我们是正常的。
因此，任何你无法解决的问题都是你自己的问题。

Douglas Adams, *The Hitchhiker's Guide to the Galaxy*

如今，软件包已经可以帮助我们计算对抽样过程和模型施加较弱假设下的渐进标准误。比如，你在 Stata 软件的回归过程中使用选项 robust，就可得到用等式(3.1.7)计算出的标准误。稳健标准误(robust standard errors)比传统标准误的改进之处在于：当残差是异方差时，基于稳健标准误的推断是渐进有效的。如果我们希望用回归近似的条件期望函数是非线性的，那么残差几乎一定是异方差的，但传统的标准误则是在同方差假设下得到的。这里想要讨论的问题是：当我们对待估参数得到的渐进近似不是很好时，相应的稳健标准误可能是错的。本章第一部分考察用稳健标准误进行推断时可能存在的推断失效问题以及一些简单的补救措施。

在传统的截面数据推断中——也就是我们在第 3.1.3 节中的讨论——所施加的假设的数据是独立的。我们将每个观察值都视为对同一个总体进行随机抽样得到的。无论何时，这些观察值之间都不相关。现在我们知道对抽样模型施加的这种假设过于理想化，甚至可以说有点蛮干。与宏观经济学中经常出现的时间序列分析类似，对截面数据的分析也必须关注观察值之间的相关性。数据间存在相关性的最重要表现形式就是组群结构(group structure)——比如我们在一个班级或者一个学校中观察到的学生考试成绩就属于这种数据类型。因为同一班级或同一学校中的学生往往受到相同环境或者家庭背景的影响。我们将数据中存在的这一类相关问题称为聚类问题(clustering problem)，或者叫做 Moulton 问题，用以纪念 Moulton(1986)对这个问题作出的贡献。与聚类问题紧密相连的另一类问题是数据在前后期存在相关性，我们曾在第 5 章用双重差分法研究此类数据。比如针对州这一级行政单位研究其最低工资对就业的影响时，我们必须面对的事实是不同时期的州平均就业水平可能是相关的。我们将这类问题视为序列相关问题，以区别于 Moulton 问题。

被数据中存在的聚类问题和序列相关问题所困扰的研究者们也必须面对这一事实：用软件 Stata 中的 cluster 选项这类最简单的方法来解决这些问题，效果往往不好。对分析聚类数据和序列相关数据的分析经常要依赖大规模聚类或者大量时间序列观察值。但是大量的聚类和长时间序列的数据也无法保证得到正确的结论。虽然解决这类问题的最好办法是获得更多的数据，不过相应的推断难题并非无法克服。我们在本章第二部分考察了用计量经济学方法解决上面两类问题的办法。本章的内容离开矩阵代数的语言将很难描述，因此我们使用这些记号。

8.1 在估计稳健标准误时存在的偏误*

使用矩阵语言，我们可知：

$$\hat{\beta} = [\sum_i X_i X_i']^{-1} \sum_i X_i Y_i = (X'X)^{-1} X'y$$

其中，X 是个 $N \times K$ 的矩阵，每一行的元素是 X_i'，y 是 $N \times 1$ 维的向量，其元素为 Y_i。在 3.1.3 节我们已经看到 $\hat{\beta}$ 是渐进正态分布的。我们可以将其写为：

$$\sqrt{N}(\hat{\beta} - \beta) \sim N(0, \Omega)$$

其中，Ω 是渐进协方差矩阵，$\beta = \mathrm{E}[X_i X_i']^{-1} \mathrm{E}[X_i Y_i]$。再次回到等式(3.1.7)，这里 Ω 就等于：

$$\Omega_r = \mathrm{E}[X_i X_i']^{-1} \mathrm{E}[X_i X_i' e_i^2] \mathrm{E}[X_i X_i']^{-1} \tag{8.1.1}$$

其中，$e_i = Y_i - X_i'\beta$。当残差同方差时，协方差矩阵可以简化为 $\Omega = \sigma^2 \mathrm{E}[X_i X_i']^{-1}$。

这里我们担心的问题是在独立抽样(也就是说不存在聚类或者序列相关问题的样本)得到的样本中计算出的稳健标准误可能存在偏误。为了简化对该偏误的求解，假设可将回归元向量看作是固定的，类似于对 X_i 进行分层抽样得到的结果。非随机回归元提供了一个可以作为基准的抽样模型，我们往往可以用其考察有限样本的分布性质。这种假设不会影响我们讨论的问题的理论意义，但在很大程度上简化了求解过程。

当回归元固定时，我们有：

$$\Omega_r = \left(\frac{X'X}{N}\right)^{-1} \left(\frac{X'\Psi X}{N}\right) \left(\frac{X'X}{N}\right)^{-1} \tag{8.1.2}$$

其中，

$$\Psi = \mathrm{E}[ee'] = \mathrm{diag}(\psi_i)$$

是残差的协方差矩阵。在同方差的假设下，对所有的 i 都有 $\psi_i = \sigma^2$，由此我们可得：

$$\Omega_c = \sigma^2\left(\frac{X'X}{N}\right)^{-1}$$

对矩阵 Ω_r 和 Ω_c 对角线元素除以 N 后求解平方根就可得相应的渐进标准误。

在实际中，用样本矩来估计渐进协方差矩阵中的元素。传统的协方差矩阵估计值是：

$$\hat{\Omega}_c = (X'X)^{-1}\hat{\sigma}^2 = (X'X)^{-1}\left(\sum \frac{\hat{e}_i^2}{N}\right)$$

其中，$\hat{e}_i = Y_i - X_i'\hat{\beta}$ 是估计出的回归残差，并且

$$\hat{\sigma}^2 = \sum \frac{\hat{e}_i^2}{N}$$

估计出的是残差方差。于是相应的稳健协方差矩阵估计值就是：

$$\hat{\Omega}_r = N(X'X)^{-1}\left(\sum \frac{X_iX_i'\hat{e}_i^2}{N}\right)(X'X)^{-1} \tag{8.1.3}$$

我们可以将等式(8.1.3)中大括号里的那部分看做是对 $\sum \frac{X_iX_i'\hat{\psi}_i}{N}$ 的估计，这里用 $\hat{\psi}_i = \hat{e}_i^2$ 来估计 ψ_i。

由大数定理和斯拉茨基定理可知，$N\hat{\Omega}_c$ 依概率收敛于 Ω_c，同时 $N\hat{\Omega}_r$ 依概率收敛于 Ω_r。但是在有限样本中上述两个估计值都是有偏的。在经典的最小二乘理论中，$\hat{\Omega}_c$ 的有偏性已经为大家所熟知。但是如果残差是同方差的，那么稳健标准误比普通标准误偏误更厉害的这一事实却少有人知晓。由此我们可以总结道：当异方差性不是很严重时，稳健标准误带来的偏误远大于普通标准误。我们还根据传统标准误的最大值和稳健标准误来提出一个法则，来避免对估计值精确性的大量误判。

我们的分析从讨论 $\hat{\Omega}_c$ 的偏误开始。当回归元非随机时，我们有：

$$\mathrm{E}[\hat{\Omega}_c] = (X'X)^{-1}\hat{\sigma}^2 = (X'X)^{-1}\left(\sum \frac{\mathrm{E}(\hat{e}_i^2)}{N}\right)$$

为了分析 $\mathrm{E}[\hat{e}_i^2]$，我们先来展开表达式 $\hat{e} = y - X\hat{\beta}$：

$$\hat{e} = y - X(X'X)^{-1}X'y = [I_N - X(X'X)^{-1}X'](X\beta + e) = Me$$

其中，e 是总体残差组成的向量，令 $M = I_N - X(X'X)^{-1}X'$，它是一个非随机矩阵，当与 e 相乘后就可得到残差 $\hat{e}$。这个矩阵中的第 i 行是 m_i'，I_N 是 $N \times N$ 的单位阵。于是 $\hat{e}_i = m_i'e$，并且有：

$$\begin{aligned}\mathrm{E}(\hat{e}_i^2) &= \mathrm{E}(m_i'ee'm_i)\\ &= m_i'\Psi m_i\end{aligned}$$

为了更进一步地简化，记 $m_i = l_i - h_i$，其中 l_i 是矩阵 I_N 的第 i 列，$h_i = X(X'X)^{-1}X_i$，是映射矩阵 $H = X(X'X)^{-1}X'$ 的第 i 列。然后：

$$\begin{aligned}\mathrm{E}(\hat{e}_i^2) &= (l_i - h_i)'\Psi(l_i - h_i) \\ &= \psi_i - 2\psi_i h_{ii} + h_i'\Psi h_i\end{aligned} \tag{8.1.4}$$

其中，h_{ii} 是矩阵 H 的第 i 个对角线元素，满足：

$$h_{ii} = h_i'h_i = X_i'(X'X)^{-1}X_i \tag{8.1.5}$$

解释一下，h_{ii} 被称为第 i 个观察值的杠杆。注意到第 i 个拟合值（也即是 Hy 的第 i 个元素）由等式(8.1.6)可得，所以这个杠杆告诉我们 X_i 的特定值对回归曲线的作用大小：

$$\hat{Y}_i = h_i'y = h_{ii}Y_i + \sum_{j\neq i} h_{ij}Y_j \tag{8.1.6}$$

如果 h_{ii} 的值比较大，那么第 i 个观测值对拟合值的作用就比较大。在只有一个回归元 x_i 的二元回归中：

$$h_{ii} = \frac{1}{N} + \frac{(x_i - \bar{x})^2}{\sum(x_j - \bar{x})^2} \tag{8.1.7}$$

这个等式指出 x_i 与其平均值的距离越远，杠杆的作用就越大。等式(8.1.6)还告诉我们 h_{ii} 是个取值在[0, 1]之间的数字，而且 $\sum_{i=1}^{N} h_{ii} = K$，也即回归元的个数（比如见 Hoaglin 和 Welsch(1978)）。①

假设残差是同方差的，因此 $\psi_i = \sigma^2$。那么等式(8.1.4)可以简化为：

$$\mathrm{E}(\hat{e}_i^2) = \sigma^2[1 - 2h_{ii} + h_i'h_i] = \sigma^2(1 - h_{ii}) < \sigma^2$$

因此 Ω_c 的样本估计值 $\hat{\Omega}_c$ 会显得偏小。使用 h_{ii} 的性质，我们进一步得到：

$$\sum \frac{\mathrm{E}(\hat{e}_i^2)}{N} = \sigma^2 \sum \frac{1 - h_{ii}}{N} = \sigma^2\left(\frac{N-K}{N}\right)$$

因此，通过修正自由度，我们可以很简单地纠正 $\hat{\Omega}_c$ 的偏误：在计算 $\hat{\sigma}^2$ 时除以 $N-K$ 而不是 N。在大多数的回归软件中都已经自动地进行了这种纠正。

现在，我们指出在同方差假设下 $\hat{\Omega}_r$ 的偏误要比 $\hat{\Omega}_c$ 的偏误严重。我们想要的稳健协方差矩阵的估计值是：

$$\mathrm{E}[\hat{\Omega}_r] = N(X'X)^{-1}\left(\sum \frac{X_iX_i'\mathrm{E}(\hat{e}_i^2)}{N}\right)(X'X)^{-1} \tag{8.1.8}$$

---

① 这里，$\sum_{i=1}^{N} h_{ii} = K$ 来自于如下事实：矩阵 H 是幂等阵，因此它的迹等于秩。我们也可以使用等式(8.1.7)来验证二元回归中 $\sum_{i=1}^{N} h_{ii} = 2$。

其中，$E(\hat{e}_i^2)$由等式(8.1.4)给出。在同方差假设下 $\varphi_i=\sigma^2$ 并且在矩阵 $\hat{\Omega}_c$ 中我们有 $E(\hat{e}_i^2)=\sigma^2(1-h_{ii})$。很清楚，估计值 $\hat{e}_i^2$ 中存在的偏误倾向于让稳健标准误偏低。但是更加一般化的等式(8.1.8)很难求解。Chesher 和 Jewitt(1987)指出如果不存在太多的异方差问题，那么基于 $\hat{\Omega}_r$ 求得的文件标准误会向下有偏。①

我们如何知道 $\hat{\Omega}_r$ 的有偏程度大于 $\hat{\Omega}_c$ 呢？这个结果部分来自于蒙特卡罗方法(比如可见 MacKinnon 和 White(1985)以及我们在后面讨论的这个小型研究)。这里我们也在二元回归中证明该结论。假设唯一的回归元 $\tilde{x}_i$ 表示回归元与其平均值之间的差异，因此我们只有一个待估参数。在这个例子中，我们感兴趣的参数是 $\beta_1=\dfrac{\sum\tilde{x}_i^2Y_i}{\sum\tilde{x}_i^2}$，杠杆 $h_{ii}=\dfrac{\tilde{x}_i^2}{\sum\tilde{x}_i^2}$(通过对常数项做除法，我们可以将等式(8.1.7)中的$\frac{1}{N}$消去)。令 $s_x^2=\dfrac{\sum\tilde{x}_i^2}{N}$。对于普通的协方差估计值，我们有：

$$E[\hat{\Omega}_c]=\frac{\sigma^2}{Ns_x^2}\left[\frac{\sum(1-h_{ii})}{N}\right]=\frac{\sigma^2}{Ns_x^2}\left[1-\frac{1}{N}\right]$$

因此这里的偏误更小。用(8.1.8)做一个小小的计算，我们就可知道在同方差的假设下稳健标准误的期望值是：

$$\begin{aligned}E[\hat{\Omega}_r]&=\frac{\sigma^2}{Ns_x^2}\sum\frac{1-h_{ii}}{N}\left(\frac{\hat{x}_i^2}{s_x^2}\right)\\&=\frac{\sigma^2}{Ns_x^2}\sum(1-h_{ii})h_{ii}=\frac{\sigma^2}{Ns_x^2}\left[1-\sum h_{ii}^2\right]\end{aligned}$$

因此如果 $\sum h_{ii}^2>\frac{1}{N}$，那么 $\hat{\Omega}_r$ 的偏误要高于 $\hat{\Omega}_c$ 的偏误；由 Jensen 不等式也可以知道，除非从回归元中计算得到的杠杆为常数$\left(\text{即对所有的 } i \text{ 有 } h_{ii}=\frac{1}{N}\right)$，否则 $\hat{\Omega}_r$ 的偏误要高于 $\hat{\Omega}_c$ 的偏误。②

我们可以通过更准确地计算 φ_i 的估计值 $\hat{\varphi}_i$ 来降低 $\hat{\Omega}_r$ 中存在的偏误。White(1980a)指出在计算 $\hat{\Omega}_r$ 时可以令 $\hat{\varphi}_i=\hat{e}_i^2$，这个估计值也是我们在这一节进行讨

① 特别的，如果 φ_i 中的最大值和最小值之比超过 2，稳健标准误会向下有偏。

② 假设 h_{ii} 是样本中符合均匀分布的随机变量。那么：

$$E[h_{ii}]=\frac{\sum h_{ii}}{N}=\frac{1}{N}$$

由 Jensen 不等式，除非 h_{ii} 是常数，否则就有：

$$E[h_{ii}^2]=\frac{\sum h_{ii}^2}{N}>(E[h_{ii}])^2=\left(\frac{1}{N}\right)^2$$

因此 $\sum h_{ii}^2>\frac{1}{N}$。当$(\tilde{x}_i^2)$为常数时杠杆成为常数。

论的出发点。在 MacKinnon 和 White(1985)中讨论的残差方差估计值在用 $\hat{\psi}_i = \hat{e}_i^2$ 对 $\hat{\psi}_i$ 进行估计之外还包含了其他三种方法：

$$HC_0: \hat{\psi}_i = \hat{e}_i^2$$

$$HC_1: \hat{\psi}_i = \frac{N}{N-K}\hat{e}_i^2$$

$$HC_2: \hat{\psi}_i = \frac{1}{1-h_{ii}}\hat{e}_i^2$$

$$HC_3: \hat{\psi}_i = \frac{1}{(1-h_{ii})^2}\hat{e}_i^2$$

HC_1 是在求解 $\hat{\Omega}_c$ 时简单地使用自由度进行修正得到的估计值。当残差是同方差时，HC_2 使用杠杆得到了第 i 个的残差方差的无偏估计值，HC_3 是对 jackknife 估计值[①]的近似。在应用中我们可以看到，从 HC_0 到 HC_3，估计出的标准误不断变大，不过这个规律还不是一个定理。

1. 谈一点自助法

自助法(bootstrap)是一种重复抽样的方法，在基于渐进性质的基础上为我们提供了另外一种推断方法。自助法下得到的样本是从我们已有的数据中再次抽样得到的。换言之，如果有一个大小为 N 的样本，我们将此样本当做总体然后不断从中抽样(存在替换的情况下)。自助法下的抽样分布就是在如此这般的抽样中得到的某个估计值的分布。从直觉上来看，我们希望从数据中抽样得到的分布可以为我们希望求解的抽样分布提供一个很好的近似。

有很多方法可以实现在自助法下进行的回归估计。最简单的方法就是在数据中抽取一对值$\{Y_i, X_i\}$，有时我们将这种方法称为“成对自助法(pairs bootstrap)”或者“非参数自助法(nonparametric boostrap)”。另外，我们也可以保持 X_i 的取值固定，从残差($\hat{e}_i$)的分布中进行抽样，然后基于从每个观察中抽取的预测值和残差构造一个新的被解释变量。这个过程是一种参数化的自助法，模仿的是从非随机回归元中抽取的样本并且保证了 X_i 和回归残差之间是相互独立的。从另一方面讲，如果我们感兴趣的是在异方差假设下的标准误，那么我们并不想要这种相互独立性。另一种利用残差进行的自助法被称为 wild 自助法(wild bootstrap)，这种方法分别用 0.5 的概率抽取 $X_i'\hat{\beta} + \hat{e}_i$ 和 $X_i'\hat{\beta} - \hat{e}_i$(例如，见 Mammen(1993)和 Horowitz(1997))。这种方法保留了我们在原始样本中观察到的残差方差和 X_i 之间的关系，并要求残差和回归元之间是均值独立的，如果这个假设成立，那么基于

① 通过每次忽略一个观察值，我们可以从数据中得到一系列的经验分布，jackknife 方差估计值就是从这些经验分布中计算抽样方差。Stata 软件可以计算 HC_1、HC_2 和 HC_3。用 Meser 和 White(1984)提供的一个小窍门，你还可以如下计算：对模型进行变型——用 $\sqrt{\hat{\psi}_i}$ 去除 Y_i 和 X_i，然后用 $X_i/\sqrt{\hat{\psi}_i}$ 做工具变量来求解变型后的模型，并最后计算出你感兴趣的 $\hat{\psi}_i$。

自助法的统计推断结果可以得到改进。

作为一种需要大量计算但是原理比较浅显的方法，自助法对我们求解渐进标准误有帮助。特别是当一个估计值的渐进难以计算或者计算过程包含多个步骤（比如在第7章讨论分位数回归和分位数处理效应的渐进分布时要求对密度函数进行估计）时，自助法中得到的估计值显得特别有用。但是值得强调的是在最小二乘估计求解标准误的渐进性质是不存在困难的。

在本章的这些内容中，使用自助法的主要原因在于它可以帮助我们改进推断的效果。统计推断的效果得以改进，一般表现为两种形式：(1)在具有一致性的估计值中降低有限样本偏误（比如在估计稳健标准误时存在的偏误）；(2)在推断过程中使用如下事实：相比于渐进近似，自助法中检验统计量的抽样分布可能很接近于我们感兴趣的有限样本分布。这两个性质被称为渐进改善（asymptotic refinement，比如可以参考 Horowitz(2001)）。

这里我们最为感兴趣的就是使用自助法进行渐进改善。回归估计值的渐进分布是很好计算的，但我们担心的是传统的稳健协方差估计值（HC_0）是有偏的。自助法可以用来估计这种偏误并且通过一个简单的转换来构造具有最小偏误的标准误估计值。但是至少从现在来看，人们还很少在经验研究实践中使用自助法来纠正对回归标准误估计的偏误，因为软件还无法自动完成对偏误的计算，也可能因为用自助法纠正偏误的过程中又会引入另外的变异。不过，像回归参数这样的简单估计值，是可以在软件中使用诸如 HC_2 和 HC_3 来对偏误进行修正的（比如在Stata中）。

基于渐进枢轴量（asymptotically pivotal statistics）[①]的假设检验（以及置信区间）都可以为我们带来渐进改进。渐进枢轴量是渐进分布不依赖于任何未知参数的统计量。比如 t 统计量就是这样的一个统计量：它是渐进标准正态分布的。但是回归参数则不是渐进枢轴量，因为它们的渐进分布确实依赖于未知残差方差。为了改进我们对回归参数的推断，你可以在自助法中的每个样本里估计 t 统计值，由此得到一个来自于自助法的 t 统计值分布，再用来自原始数据的 t 统计值进行比较。如果来自原始数据的 t 统计值的绝对值过高，那么我们就可以拒绝 t 统计值等于某个值的假设检验，这种情况相当于是原始数据的 t 统计值的绝对值大于自助法中 t 统计值分布的95%。

虽然从理论上看渐进枢轴量的性质很好，但是作为应用研究者，我们非常不喜欢在自助法中使用枢轴量。部分原因在于我们不只感兴趣于进行正式的假设检验：我们喜欢在回归参数下的括号里看到标准误。这个量为我们提供了对估计精度的一种度量，可以用来构造置信区间，可以在估计值之间进行比较而且可以检验任何我们感兴趣的假设。因此在我们看来，在实践中担心稳健标准误有限样本偏

① 枢轴量适合于构造统计检验，因为它的性质允许我们控制统计检验中的第一类错误，而这种错误和未知参数无关，也和数据来自何方无关。——译者注

误的研究者们应该关注于使用 HC_2 或者 HC_3 来纠正这些偏误。正如我们在下面将要看到的，至少在异方差性不是很严重的情况下，使用传统标准误或者纠正过估计偏误的标准误以及可以为我们带来很好的结果：在不失精确性的前提下降低估计偏误。

2. 一个例子

为了更进一步考察不同的稳健协方差估计值之间的区别，我们分析一个简单且重要的例子，这个例子曾经出现在本书之前的章节中。假设你希望在下面这个模型中估计 β_1：

$$Y_i = \beta_0 + \beta_1 D_i + \varepsilon_i \tag{8.1.9}$$

其中，D_i 是个虚拟变量。那么最小二乘估计值就是 D_i 等于 1 和等于 0 时被解释变量在均值上的差别。将样本中 D_i 等于 1 和等于 0 的子集分别用下标 1 和 0 标记，于是我们有：

$$\hat{\beta}_1 = \overline{Y}_1 - \overline{Y}_0$$

在这种记号下，我们认为 D_i 是非随机的，因此 $\sum D_i = N_1$ 且 $\sum (1-D_i) = N_0$ 也固定的。令 $r = N_1/N_0$。

有统计理论，我们已经知道了 $\hat{\beta}_1$ 的一些有限样本性质。假设 Y_i 是正态分布的，方差未知但是在 $D_i = 1$ 和 $D_i = 0$ 的子集中 Y_i 的方差相等，那么 $\hat{\beta}_1$ 的 t 统计量具有 t 分布。这就是经典的双样本 t 检验。在这个例子中，异方差性意味着 $D_i = 1$ 和 $D_i = 0$ 的子集中 Y_i 的方差不相等。这时在小样本中的检验问题变得异常困难：即使在这个例子中精确的小样本分布也是未知的。[①]在方差不等的情况下，稳健的方差估计值 HC_0—HC_3 为我们提供了有限样本中分布未知情况下对相应方差的渐进近似。

HC_0、HC_1、HC_2 和 HC_3 之间的不同在于我们如何处理由 D_i 定义的两个子集的方差。定义 $S_j^2 = \sum D_{i-j}(Y_i - \overline{Y}_j)^2$，其中 $j = 0, 1$。在这里例子中杠杆等于：

$$h_{ii} = \begin{cases} \dfrac{1}{N_0} & \text{if} \quad D_i = 0 \\ \dfrac{1}{N_1} & \text{if} \quad D_i = 1 \end{cases}$$

用上面这个等式，我们可以直接计算出正在讨论的五个方差估计值分别是：

传统意义上的方差：$\dfrac{N}{N_0 N_1}\left(\dfrac{S_0^2 + S_1^2}{N-2}\right) = \dfrac{1}{Nr(1-r)}\left(\dfrac{S_0^2 + S_1^2}{N+2}\right)$

① 这个问题被称为 Behrens-Fisher 问题（比如见 DeGroot 和 Schervish(2001)的第八章）。

$$HC_0:\ \frac{S_0^2}{N_0^2}+\frac{S_1^2}{N_1^2}$$

$$HC_1:\ \frac{N}{N-2}\left(\frac{S_0^2}{N_0^2}+\frac{S_1^2}{N_1^2}\right)$$

$$HC_2:\ \frac{S_0^2}{N_0(N_0-1)}+\frac{S_1^2}{N_1(N_1-1)}$$

$$HC_3:\ \frac{S_0^2}{(N_0-1)^2}+\frac{S_1^2}{(N_1-1)^2}$$

传统意义上的方差估计值将两个子样本中的方差混在一起：当这两个方差相同时该方差估计值是有效的。White(1980a)中的方差估计值是 HC_0，用一致但是有偏的方差估计值$\frac{S_j^2}{N_j^2}$分别估计了来自两个子样本中的方差的均值并将其相加。HC_2 分别对每个子样本中的方差使用了无偏估计，因为它运用正确的自由度调整了这两个估计值。HC_1 则是在加总后的方差估计值上进行了自由度的修正，这个修正的过程只能有助于但仍无法得到正确的估计值。由于我们知道 HC_2 是在同方差假设下得到的抽样方差的无偏估计值，所以估计值 HC_3 应该是偏大了。①注意到当 $r=0.5$ 时我们称回归设计是平衡的，那么传统的方差估计值等于 HC_1 估计值，所有的五个方差估计值之间的差别都不太大。

基于方程(8.1.9)进行的蒙特卡洛研究阐述了上述几个估计的长处与短处，并且讨论了在什么情况下一开始建立的那个标准可以用来纠正如 HC 估计值中存在的偏误。我们选择 $N=30$ 来体现所研究的是小样本，令 $r=0.10$（样本中有 10% 的个体受到处理），这意味着当 $D_i=1$ 时 $h_{ii}=\frac{1}{3}$，当 $D_i=1$ 时 $h_{ii}=\frac{1}{27}$。这样就给出了一个轻微不平衡的回归设计。我们从下面这个分布中抽取残差：

$$\varepsilon_i \sim \begin{cases} N(0,\ \sigma^2) & \text{if} \quad D_i=0 \\ N(0,\ 1) & \text{if} \quad D_i=1 \end{cases}$$

并报告了三种情况下的估计结果。第一种情况总存在大量的异方差问题，$\sigma=0.5$，而在第二个例子中的异方差性相对少一点，其 $\sigma=0.85$，第三个例子中无异方差性，是作为基准的。

表 8.1 报告估计结果。第 1 列和第 2 列报告了在 25 000 次抽样实验中估计出的对各种标准误估计出的均值和标准离差。$\hat{\beta}_1$ 的标准离差就是我们试图估计的样本方差。如表中 A 部分所示，当存在大量异方差性时，传统的标准误估计存在很大的偏误，平均而言，估计值只有蒙特卡洛抽样方差的一半。但是除了 HC_3 显

① 在这个简单的例子中，不论残差是否同方差，HC_2 都是无偏估计值。

得过小[①]之外，其他的稳健标准误则表现良好。

标准误本身也是个估计值，因此也存在抽样变动。特别值得一提的是相比于传统的标准误，稳健标准误的变动程度更大，我们可以在第 2 列看到这个结果。[②]当我们通过对残差除以 $1-h_{ii}$ (HC_2)或者 $(1-H_{ii})^2$ (HC_3)以降低估计偏误时，估计出的标准误的抽样变动开始增加。情况最差的是 HC_3 其估计值的标准误比 HC_0(White，1980a 中得到的标准误)要高 50%。

表 8.1 的最后两列报告了在 5%水平上拒绝原假设 $\hat{\beta}_1 = 0$ 的次数，这里 $\hat{\beta}_1 = 0$ 是本例中来自总体的参数。由此得到的 t 统计量与正态分布估计值相比较，同时也和自由度为 $N-2$ 的 t 统计值进行比较。对于所有的统计量，拒绝原假设的次数都很高，即使对于 HC_3 也是如此。使用 t 统计量而不是正态分布只在边际上改进了结果。

表 8.1　稳健标准误的蒙特卡罗估计结果

			5%的拒绝率	
	平均值	标准误	正态分布	t 值
参 数 估 计	(1)	(2)	(3)	(4)
A. 存在大量异方差问题				
$\hat{\beta}_1$	−0.001	0.586		
标准误				
传统的	0.331	0.052	0.278	0.257
HC_0	0.417	0.203	0.247	0.231
HC_1	0.447	0.218	0.223	0.208
HC_2	0.523	0.260	0.177	0.164
HC_3	0.636	0.321	0.130	0.120
max{HC_0，传统的}	0.448	0.172	0.188	0.171
max{HC_1，传统的}	0.473	0.190	0.173	0.157
max{HC_2，传统的}	0.542	0.238	0.141	0.128
max{HC_3，传统的}	0.649	0.305	0.107	0.097
B. 存在很少的异方差问题				
$\hat{\beta}_1$	0.004	0.600		
标准误				
传统的	0.520	0.070	0.098	0.084
HC_0	0.441	0.193	0.217	0.202

① 虽然 HC_2 是抽样方差的无偏估计值，在整个抽样实验中 HC_2 标准误的均值(0.52)都低于 $\hat{\beta}_1$ 的标准误(0.59)。这是因为：标准误是抽样方差的平方根，但抽样方差本身就是个估计值，它也存在变化，而求平方根的函数却是个凹函数。

② Chesher 和 Austin(1991)最先注意到稳健标准误估计值的抽样方差较大。Kauermann 和 Carroll (2001)提出了对置信区间进行调整以纠正这种偏误的方法。

（续表）

			5%的拒绝率	
	平均值	标准误	正态分布	t 值
参数估计	(1)	(2)	(3)	(4)
HC_1	0.473	0.207	0.194	0.179
HC_2	0.546	0.250	0.156	0.143
HC_3	0.657	0.312	0.114	0.104
max{HC_0，传统的}	0.562	0.121	0.083	0.070
max{HC_1，传统的}	0.578	0.138	0.078	0.067
max{HC_2，传统的}	0.627	0.186	0.067	0.057
max{HC_3，传统的}	0.713	0.259	0.053	0.045
B. 存在很少的异方差问题				
$\hat{\beta}_1$	−0.003	0.611		
标准误				
传统的	0.604	0.081	0.061	0.050
HC_0	0.453	0.190	0.209	0.193
HC_1	0.486	0.203	0.185	0.171
HC_2	0.557	0.247	0.150	0.136
HC_3	0.667	0.309	0.110	0.100
max{HC_0，传统的}	0.629	0.109	0.055	0.045
max{HC_1，传统的}	0.640	0.122	0.053	0.044
max{HC_2，传统的}	0.679	0.166	0.047	0.039
max{HC_3，传统的}	0.754	0.237	0.039	0.031

注：本表报告了用 25 000 个复制值作出的抽样实验结果。第 1 列和第 2 列报告的是估计出的标准误的均值和离差，但是在 A、B、C 三个部分的第一列都报告了 $\hat{\beta}_1$ 的均值和标准误。用于计算本表的模型表示在模型(8.1.9)中，其中 $\beta_1 = 0$，$r = 0.1$，$N = 30$，各种异方差性都表示在 A、B、C 三部分的表头中。

在异方差性很小的情况下得到的估计结果报告在表 8.1 的 B 部分，显示出传统的标准误估计值还是显得过小，不过这种偏误已经减小到了 15%。HC_0 和 HC_1 对应的估计值也偏小，而且从比例上看和 B 部分的情况类似，但是相对于普通的标准误，这两个估计值变得更差了。平均而言，HC_2 和 HC_3 下的标准误估计值也还是比普通标准误估计值大，实际中的拒绝原假设的次数也比普通标准误要大。这意味着稳健标准误有时会“碰巧”很小，当这种“碰巧”很小的事情发生多次后，就足以使得拒绝率变大以至于超过传统标准误下拒绝原假设的比例。

从上面的结果和比较中我们得到的教训是：稳健标准误并非灵丹妙药。在两个原因的作用下，它可能比普通标准误更小：我们已经讨论过的小样本偏误以及过大的抽样方差。因此，当稳健标准误小于传统标准误时，我们认为它是经验研究结论可能存在问题的危险信号。很可能是因为存在某种偏误或者没有考虑到的情况，才会出现稳健标准误小于传统标准误。在这个看法下，取传统标准误和稳健标

准误的最大值可能可以更好地度量估计的精确性。这种经验法则在两个方面有所帮助：它剔除了稳健标准误估计中可能得到的较低值，降低了偏误并且降低了波动。表 8.1 给出了使用 $\max(HC_j$，传统标准误）后得到的拒绝概率。其 B 部分的估计结果指出这种经验法则看起来相当好。即使在存在大量异方差性的 A 部分，该经验法则下的结果也比单独使用稳健标准误得到的结果要好。①

由于没有付出便没有收获，使用经验法则 $\max(HC_j$，传统标准误）也一定存在着某些方面的损失。这个损失就是在不存在异方差性时最好的标准误是传统的标准误。这个结论报告在表 8.1 的 C 部分。在同方差假设下使用经验法则使得标准误被不必要地夸大，从而压低了拒绝原假设的比例。而且，C 部分的结果还指出拒绝原假设的比例不会下降。我们认为对估计精度的低估要好于对估计精度的高估。当低估估计精度时，我们会考虑数据中的信息量是不是不足，是不是要收集更多的数据或者改进研究设计，但是当高估估计精度时，我们可能会错误地作出一些重要的结论。

对这个蒙特卡洛方法的最后一点意见与样本规模有关。像我们这样的劳动经济学家往往要和成千上万的观察值打交道。但是有些时候你遇到的情况可能不是这样。在研究乘坐公共汽车上学对公立学校学生的影响中，Angrist 和 Lang (2004)中的观察值来自 56 个学校的 3 000 名学生。在这项研究中我们感兴趣的回归元只是在年级这个层面上的变量，因此其中的一些研究使用来自 56 个学校的均值。因此很正常的，Angrist 和 Lang(2004)在面对学校层面的数据时得到的 HC_1 型标准误估计值要比传统的最小二乘标准误小。因此，即使你是从来自个体的微观数据出发的，当我们感兴趣的回归元是在更高层面——比如学校、州或者其他组别和聚类——的加总结果时，有效的样本规模将更接近于聚类的数量，而不是微观数据的数量。下一节我们对聚类数据的推断问题进行详细讨论。

8.2 面板数据中的聚类问题和序列相关问题

8.2.1 聚类与 Moulton 因子

在进行统计推断时，异方差性不会导致结果的巨大变化。特别是在大样本数据里对标准误估计的偏误不大会成为很大的问题，比如我们会发现 HC_1 估计值只比传统的标准误高 25%。但相比之下，聚类问题带来的偏误可能会很大。

用一个在回归中使用的数据具有群结构的简单的二元回归就可以描述聚类问题。假设我们感兴趣于估计下面的这个二元回归：

$$Y_{ig} = \beta_0 + \beta_1 x_g + e_{ig} \tag{8.2.1}$$

① Yang、Hsu 和 Zhao(2005)正式地给出了用一组具有不同效率和稳健特性的检验统计量中的最大值进行检验的过程。

其中，Y_{ig}是在聚类(或者说群结构)g中第i个个体的被解释变量，这里假设有G个聚类。重要的是，我们感兴趣的回归元只在群结构这个层面发生变化。比如在Krueger(1991)分析过的STAR实验数据中，Y_{ig}是处在班级g中的第i个学生的考试成绩，x_g是班级g的规模。

虽然在STAR实验中学生被随机分配进某个班级，但是STAR数据的不同观察值之间应该不是相互独立的。同一个班级中学生的考试分数应该是相关的，因为在同一个班级的学生可能有着相同的背景、个体特征并且面对相同的老师和教学环境。因此我们应小心地假设在同一班级g中的学生i和j之间应该有：

$$\mathrm{E}[e_{ig}e_{jg}] = \rho_e\sigma_e^2 > 0 \tag{8.2.2}$$

其中，ρ_e在同一班级中学生成绩残差之间的相关系数，σ_e^2是残差方差。

我们往往可以用随机项相加的方式来表现组内部存在的相关性。具体而言，我们假设残差e_{ig}具有下面的这个群结构：

$$e_{ig} = \nu_g + \eta_{ig} \tag{8.2.3}$$

其中，ν_g是群g的随机项，该随机项对群g中的所有个体都是一样的，η_{ig}是剩下的在学生个体层面的随机部分，假设其均值为0。我们在这里只关心残差中存在的相关性问题，因此假设这里的残差都是同方差的。由于我们假设群结构中个体之间的相关性完全由群随机项造成，因此诸η_{ig}之间是不相关的。①

当我们感兴趣的回归元只在群这一层面发生变动时，等式(8.2.3)中的那种残差结构可能会使得标准误急速上升。这个不幸的事实并非新鲜事——Kloek(1981)以及Moulton(1986)都指出了这一点——但是可以很公平地讲在15年前聚类问题尚未进入计量经济学的范畴中。

给定等式(8.2.3)中的残差结构，班级内部的相关系数就成为：

$$\rho_e = \frac{\sigma_\nu^2}{\sigma_\nu^2 + \sigma_\eta^2}$$

其中，σ_ν^2是ν_g的方差，σ_η^2是η_{ig}的方差。这里对用到的术语说明一下：我们可以将ρ_e称为组内相关系数，不论我们所考察的问题是班级规模还是别的什么问题。

令$V_c(\hat{\beta}_1)$是传统的回归斜率的最小二乘方差公式(在上一节中它就是矩阵$\hat{\Omega}_c$在对角线上的元素)，这里我们用$V(\hat{\beta}_1)$来表示在给定残差结构(8.2.3)后修正的抽样方差。当非随机的回归元在群这一层面固定并假设群规模相等，都是n，那

① 残差项的这种相关结构也是分层抽样的结果(比如见Wooldridge(2003))。我们这里考察的样本中的绝大部分都可以看作是随机的，因此我们只需控制在群结构层面的残差相关性，而无需观察有抽样本身带来的聚类问题。注意到当残差结构是(8.2.3)时，我们无法用广义最小二乘法来估计方程(8.2.1)，因为我们关心的回归元是群层面上的，因此在每个群结构中该回归元都是固定的。不论在这里还是在其他的例子中，我们都选择"固定标准误"的方法，而不使用广义最小二乘估计。

么我们用：

$$\frac{V(\hat{\beta}_1)}{V_c(\hat{\beta}_1)} = 1 + (n-1)\rho_e \tag{8.2.4}$$

在本章附录部分我们会具体求解这一等式。我们将这个比值的平方根称为 Moulton 因子，用以纪念 Moulton(1986)的这一有广泛影响力的研究。方程(8.2.4)告诉我们如果忽略组内相关性，我们会在多大程度上高估精确性。随着 n 和 ρ_e 的变大，传统标准误的"准头"会越来越差。比如假设 $\rho_e = 1$。在这个例子中群结构内部的所有个体之间的残差都是相同的，因此在这个群结构的个体中 Y_{ig} 也是相等的。那么通过对一个较小的样本复制 n 次而得到的一个大样本，然后对这个样本进行分析得到的结论也许毫无意义。因为这样得到的 $V_c(\hat{\beta}_1)$ 也许是原来 $V_c(\hat{\beta}_1)$ 的 n 倍。随着群规模的增大，Moulton 因子也变大的原因在于如果我们给定总样本数，群规模的扩大意味着聚数目的减少，这时总样本中能够提供独立信息的来源就变少了(因为在聚类之间，数据是相互独立的，但是在聚类内部则不是)。①

即使组内相关系数很小，也可能产生一个很大的 Moulton 因子。比如在 Angrist和 Lavy(2004)的研究中有来自 40 个学校的 4 000 个学生，所以平均群规模就是 $n = 100$。我们感兴趣的回归元是学校层面上的处理状态：在所有受到处理的学校中，如果他们能够通过预科考试，那么就给予现金奖励。在这项研究中，组内相关性大致在 0.1。使用公式(8.2.4)，Moulton 因子超过了 3，因此用传统方法得到的标准误只是其正确值的三分之一。

等式(8.2.4)包含了一个特别重要的特例，这个特例就是在群内回归元固定，同时群规模固定。在一般性的公式中是允许回归元 x_{ig} 变化的，同时也允许群规模 n_g 发生变化。在这些个例子中，Moulton 因子就是等式(8.2.5)的平方根：

$$\frac{V(\hat{\beta}_1)}{V_c(\hat{\beta}_1)} = \left[\frac{V(n_g)}{\bar{n}} + \bar{n} - 1\right]\rho_x\rho_e \tag{8.2.5}$$

其中，$\bar{n}$ 是平均群规模，ρ_x 是 x_{ig} 的组内相关系数：

$$\rho_x = \frac{\sum_g \sum_j \sum_{i \neq j} (x_{ig} - \bar{x})(x_{jg} - \bar{x})}{V(x_{ig}) \sum_g n_g (n_g - 1)}$$

注意到 ρ_x 中并没有包含类似于等式(8.2.3)那样的残差结构；这里 ρ_x 是对群内回归元相关性的一个一般性度量。一般意义上的 Moulton 因子告诉我们当 ρ_x 较大而且群规模可变时聚类问题对标准误产生了很大的影响。当 $\rho_e = 0$ 时这种影响消失。换言之，如果在群内诸 x_{ig} 之间都是不相关的，那么残差的群结构就不影响对

① 当回归元是非随机的而且残差为同方差时，Moulton 因子是个有限样本结果。调查统计学家(survey statisticians)将 Moulton 因子称为设计效应(design effects)，因为这个因子告诉我们在简单随机抽样的分层样本中应该如何去调整标准误(Kish，1965)。

标准误的估计。这就是当回归元固定时我们在回归中担心聚类问题的原因。

我们用田纳西州 STAR 实验中得到的数据来阐述等式(8.2.5)。用班级规模对幼儿园学生成绩的百分位数做回归后得到的结果是−0.62，稳健(HC_1 型)标准误是 0.09。由于对班级内部的每个人而言，班级规模都是固定的，所以在这个例子中 $\rho_x = 1$，而且由于班级规模会发生变动，所以 $V(n_g)$ 是正数(在这个例子中 $V(n_g) = 17.1$)。残差的组内相关系数为 0.31，平均的班级规模是 19.4。将这些数字代入等式(8.2.5)后我们得到 $\frac{V(\hat{\beta}_1)}{V_c(\hat{\beta}_1)}$ 的取值大致为 7，因此我们应该对传统的标准误乘以 2.65 得到 $\sqrt{7}$。因此被纠正过的标准误应该是 0.24。

Moulton 因子的工作原理和 2SLS 估计的工作原理是一样的。特别的，用 $\rho_{\hat{z}}$ 来代替 ρ_x，这里 $\rho_{\hat{z}}$ 是第一阶段拟合残差的组内相关系数，取 ρ_r 为第二阶段拟合残差的组内相关系数(Shore-Sheppard(1996))，我们就可以使用等式(8.2.5)。为了理解这个机理，回忆我们对 2SLS 的讨论就知道：用第一阶段拟合值的方差除以第二阶段方差中的残差方差就可得到 2SLS 估计的传统标准误。这个标准误与最小二乘估计的间接方差公式相同，只不过在这里第一阶段拟合值扮演的是回归元的角色。

在此我们进行总结，列出并比较 Moulton 问题的求解方式，先从参数估计法开始：

(1) 参数化方法：用等式(8.2.5)来计算传统的标准误。用一些统计软件的描述性统计就可方便地计算组内相关系数 ρ_r 和 ρ_x。①

(2) 聚类标准误：Liang 和 Zegar(1986)对 Wihte(1980)的稳健协方差矩阵进行了推广，使得这个公式中既可以允许出现聚类问题，也可以允许异方差问题出现。聚类的协方差矩阵就是：

$$\hat{\Omega}_d = (X'X)^{-1}\left(\sum_g X_g \hat{\Psi}_g X_g\right)(X'X)^{-1} \tag{8.2.6}$$

其中，

$$\hat{\Psi}_g = a\hat{e}_g\hat{e}_g' = a\begin{bmatrix} \hat{e}_{1g}^2 & \hat{e}_{1g}\hat{e}_{2g} & \cdots & \hat{e}_{1g}\hat{e}_{n_g g} \\ \hat{e}_{1g}\hat{e}_{2g} & \hat{e}_{2g}^2 & \cdots & \vdots \\ \vdots & \vdots & & \hat{e}_{n_g-1,g}\hat{e}_{n_g g} \\ \hat{e}_{1g}\hat{e}_{n_g g} & \cdots & \hat{e}_{n_g-1,g}\hat{e}_{n_g g} & \hat{e}_{n_g g}^2 \end{bmatrix}$$

这里，H_g 是群 g 中回归元所在的矩阵，a 是对自由度进行调整的因子，类似于在 HC_1 中调整自由度的那个部分。不论群中的残差结果是不是等式(8.2.3)所描述的那样，这个聚类估计值都是一致的，随着样本规模的扩大，会收敛到它的真实值。

① 例如用软件 Stata 中的命令 loneway 就可完成这种计算。

但是即便群规模趋于无穷，只要群数是固定的，那么 Ω_{cl} 便不会是一致估计。一致性是由大数定理给出的，这个定理指出样本矩会收敛于总体矩（见第 3.1.3 节）。但是这里我们可以看到是在对群层面上的数据进行加总的，而不是对个体层面的数据进行加总。因此当聚类数较少时，估计出的聚类标准误可能就不是那么牢靠，这一点在后面我们还会回头进行讨论。

（3）用群均值而不是微观数据：令 $\overline{Y}_g$ 是群 g 中 Y_{ig} 的平均值。用将群规模作为权数的加权最小二乘法来估计下面的等式：

$$\overline{Y}_g = \beta_0 + \beta_1 x_g + \bar{e}_g$$

这与使用微观数据的最小二乘法是等价的，只不过分组方程的标准误反映的是等式(8.2.3)所表达的残差结构。[①]同样的，这里的渐进性质也是基于群个数而非群规模。但是当群规模差不多合适时，群均值接近于正态分布，这时进行回归的方程具有正态分布的残差，而且还由良好的有限样本性质。因此相比于在只有不多的聚类的样本中的得到的聚类标准误，从分组估计中得到的标准误可靠得多。

用两步估计的办法，我们可以将分组数据估计推广到存在微观协变量（micro-covariates）的情况。假设我们感兴趣的方程是：

$$Y_{ig} = \beta_0 + \beta_1 x_g + \beta_2 W_{ig} + e_{ig} \tag{8.2.7}$$

其中，W_{ig} 是在不同群中发生变化的协方差。在第一步构造经过协方差调整的群效应 μ_g，用的是下面这个方程：

$$Y_{ig} = \mu_g + \beta_2 W_{ig} + \eta_{ig}$$

被称为群效应的参数 μ_g 是表示各个群的完备虚拟变量前的系数。这里还根据个体协变量 W_{ig} 的不同对 μ_g 的估计值 $\hat{\mu}_g$ 进行了调整。注意到根据等式(8.2.7)和等式(8.2.3)，$\mu_g = \beta_0 + \beta_1 x_g + \nu_g$。因此，在第二步时我们用群这一层面上的变量对估计出的群效应进行回归：

$$\hat{\mu}_g = \beta_0 + \beta_1 x_g + \{(\hat{\mu}_g + \mu_g)\} \tag{8.2.8}$$

对方程(8.2.8)进行有效估计的广义二乘法就是加权最小二乘法，其权重取群这一层面上残差 $(\nu_g + (\hat{\mu}_g - \mu_g))$ 的方差的倒数。但是这里存在一个问题，当群数太小时 ν_g 的方差无法得到很好的估计。因此在实际中我们使用估计出的群效应的方差、群规模做权重，甚至也可不用权重。[②]为了更好地近似相应的有限样本分布，Donald 和 Lang(2007)指出对类似于方程(8.2.8)那样的分组方程进行的推断可以基于自由度为 $G-K$ 的 t 分布。

注意到当 x_{ig} 在群内存在变化时，分组的方法就不能用了。将 x_{ig} 平均化为 $\bar{x}_g$

① 除非各个群的规模相等，否则分组残差就会是异方差的，但是相比于来自微观数据的残差存在群结构，异方差性实际上是没那么重要的。

② 可将 Angrist 和 Lavy(2008)看作一个例子，在那篇论文中作者用后两种加权方法进行了计算。

就是一种工具变量估计法，这部分内容我们在第4章讨论过。因此，当我们感兴趣的回归元在微观层面上存在变动时，分组估计法得到的参数与我们想要在模型(8.2.7)中估计的参数是不一样的。

(4) Block自助法：一般而言，自助法借助于重新抽样得到的经验分布进行推断。但是在这个例子中简单的随机抽样是无效的。处理聚类数据的技巧在于在我们希望估计的总体中保持相互依赖的结构。我们可以通过Block自助法来实现这个目标，在群 g 定义好的群中抽取数据。以田纳西州STAR实验为例，我们用block自助法时重新抽样的是各个群，而不是个体。

(5) 在一些例子中，你可能会用广义最小二乘法或者极大似然估计法来估计具有残差结构(8.2.3)的方程(8.2.1)。这种办法解决了聚类问题，但是如果条件期望函数是非线性的，那么这种回归就改变了相应条件期望函数中我们感兴趣的变量。这一点在(3.4.1)讨论有限被解释变量时已经提到过。因此我们要选择其他的方法。

表8.2以田纳西州STAR项目为例报告了调整后的标准误：该表报告了六个估计值，分别是：传统的稳健标准误(使用 HC_1)；两个使用Moulton公式(8.2.5)得到的修正标准误，其中第一个是根据Moulton给出的组内相关性公式求出的，第二个根据Stata中命令loneway得到的估计值求出的；聚类标准误；block-自助法标准误以及在群这一层面上使用加权估计得到的标准误。在这个例子中求解出的参数是−0.62。所有经过聚类调整的标准误都得到了相同的结果，这些结果指出标准误大致在0.23左右。得到这样好的结果的大部分原因在于我们研究的问题中有318个班级，因此我们得到的聚类数足以使得群层面上的渐进估计一致。如果聚类的数量偏少，那么结果就很不确定了，在本章最后一部分我们再来讨论这个问题。

表8.2　在入学前提前教育实验中得到的课堂规模的标准误

方差估计值	标准误
稳健的(HC_1)	0.090
参数化的Moulton修正(用Moulton组内相关系数)	0.222
参数化的Moulton修正(用Stata软件中的组内相关系数)	0.230
聚类的	0.232
Block-自助法	0.231
用群均值的估计(用班级规模做权重)	0.226

注：本表报告了用班级规模对幼儿园学生测验分数百分位分布做回归得到的各种标准误。其中数据来自田纳西州STAR项目公开的数据。班级规模前的系数是−0.62。聚类过程中将每个班级当做一个群。观测值有5 743个。自助法估计值使用了1 000次重复抽样。

8.2.2　面板数据中的序列相关问题及双重差分

序列相关——一个观察与之前观察值之间存在相关性——曾被人们看作是别

人的问题(somebody else's problem)，有这种想法的人是如此的不幸，因为他们生活在没有时间序列数据的世界里(比如说，宏观经济学家)。因此，应用微观计量经济学家长期以来对这个问题都是忽视的。[①]但是我们所使用的数据往往具有时间上的维度，特别是在双重差分模型中。这一事实与聚类问题相结合，可能对统计推断产生重要的影响。

像在5.2节那样，我们对州最低工资带来的效果感兴趣。在这个例子中对双重差分模型的回归分析包含着州效应和时间效应，这两个效应都是以加的方式进入总体回归方程的。因此我们可以得到类似方程(5.2.2)的方程，重新写出来就是：

$$Y_{ist} = \gamma_s + \lambda_t + \delta D_{st} + \epsilon_{ist} \tag{8.2.9}$$

和之前的讨论类似，Y_{ist}是在年份t处于州s的个体i的某个结果，D_{st}是个虚拟变量，用以表示在处理期过后的处理状态。

在方程(8.2.9)中的残差项反映出的是在不同年份处在不同州的不同个体的潜在结果会存在变动。对于同一年在同一州的那些人而言，他们的结果变量中可能存在一个共同的变动，比如区域性的商业周期带来的影响。我们可以想象这种共有的变化ϵ_{ist}由州—年份冲击ν_{st}和异质性的个体变动η_{ist}两部分组成。因此我们有：

$$Y_{ist} = \gamma_s + \lambda_t + \delta D_{st} + \nu_{st} + \eta_{ist} \tag{8.2.10}$$

我们假设在州和年份之间进行随机抽样时有$E[\nu_{st}] = 0$，同时由定义可知$E[\eta_{ist} \mid s, t] = 0$。

对双重差分模型而言，州—年份冲击是个坏消息。正如我们在讨论Moulton问题时指出的，州和年份的随机效应产生了一个聚类问题，这个问题影响到统计推断。但是在这个例子中它至少是我们关心的一个问题。为了看清楚为什么，类似Card和Krueger(1994)对新泽西州和宾夕法尼亚州进行的研究，假设只有两个时期和两个州。在那里我们求出的双重差分估计值是：

$$\hat{\delta}_{OLS} = (\hat{Y}_{s=NJ,\ t=Nov} - \hat{Y}_{s=NJ,\ t=Feb}) - (\hat{Y}_{s=PA,\ t=Nov} - \hat{Y}_{s=PA,\ t=Feb})$$

由于$E[\nu_{st}] = E[\eta_{ist}] = 0$，所以上面这个估计值是无偏的。从另一方面而言，假设所选择的时期和州不变，我们考虑的是群规模变大后估计值的概率极限，那么州—年份冲击使得估计值$\hat{\delta}_{OLS}$不一致：

$$plim\,\hat{\delta}_{OLS} = \delta + \{(\nu_{s=NJ,\ t=Nov} - \nu_{s=NJ,\ t=Feb}) - (\nu_{s=PA,\ t=Nov} - \nu_{s=PA,\ t=Feb})\}$$

我们可以不断扩大在新泽西州和宾夕法尼亚州的两个时期中获得的样本规模，但

① “别人的问题”是指人们往往忽视对自己不重要但对别人很重要的事情。Douglas Adams在其科幻小说《生命、世界及宇宙》中提到了这种自然现象，维基百科对该词条的解释是“一种人为生成的能量场，它会影响到人们的认知。任何处在这个能量场中的事物都会被别人视为外部事物，这个外部事物就是别人的问题，除非观察者在需找特定的事物，否者处在这个能量场中的所有事物都被有效屏蔽。”请见 http://en.wikipedia.org/wiki/Somebody_Else%27s_Problem。——译者注

是在这个越来越大的样本中进行平均化的过程却未能将特定地区和时期的冲击消除。当只有两个州和两个时期时，我们无法将下面两件事情相区别：一是由政策变化引起的双重差分，二是由特定冲击引起的双重差分。比如在 1992 年，宾夕法尼亚州经历下行的商业周期而新泽西州未经历商业周期带来的双重差分效应。因此，ν_{st} 的存在使得我们在 5.2 节对中讨论的共同趋势假设失效。

解决由双重差分模型中存在的随机冲击带来的不一致性的方法就是分析包含多个时期多个州（或者两者兼有）的样本。比如，Card(1992)使用 51 个州分析了最低工资变化带来的影响，Card 和 Krueger(2000)用更长期的月度工资支付数据分析了新泽西州—宾夕法尼亚州的实验。当使用的数据存在多个州和时期时，我们可以希望 ν_{st} 平均化为零。正如在本章第一部分讨论 Moulton 问题那样，推断框架建立在基于很多组的渐进理论上，而不是建立在组规模上（或者说至少不是组规模单独决定的）。推断中最重要的问题变为考察 ν_{st} 的变化行为。特别的，如果我们想假设在不同州和不同时期发生的冲击是随机的——也即它们是序列不相关的——那我们就回到了 8.2.1 节中基本的 Moulton 问题，在那里由州×年份引起的聚类标准误可能使得推断失效。但是在大部分例子中，我们很难说明 ν_{st} 是序列不相关的。比如我们可以很确定地说地区性冲击是序列相关的：如果在某一个月中宾夕法尼亚州在某个指标上表现不好，那么在下一个月这个指标还可能不是很好。

Bertrand，Duflo 和 Mullainathan(2004)以及 Kézdi(2004)强调了聚类面板数据中的序列相关问题所带来的后果。任何研究设计，只要存在群结构从而使得群均值之间是相关的，那么都可以说这个研究设计存在序列相关问题。目前对存在群结构的数据中序列相关问题进行研究的结论是：正如我们要调整标准误以应对存在 ν_{st} 时带来的组间相关问题，我们必须进一步对 ν_{st} 本身进行调整以解决序列相关问题。有一系列方法可以解决这个问题，不过这些方法不是同样有效的。如何解决序列相关问题目前仍在研究之中，而且尚未形成共识。

最简单且被广泛运用的方法就是把聚类问题留给更高一个层面。在州—年份的例子中，我们可以报告在州这一个层面上的聚类标准误(Liang 和 Zeger(1986)，比如可以在 Stata 中使用 Cluster 命令）而不是报告州—年份层面上的聚类标准误。乍一看似乎觉得这种方法很奇怪，因为模型本来就是控制了州效应的。在方程(8.2.10)中的州效应 γ_s 捕捉到的是 ν_{st} 的均值，我们将其记为 $\bar{\nu}_s$。不过 $\nu_{st}-\bar{\nu}_s$ 似乎还是序列相关的。计算州这一层面上的聚类标准误则可以解决这个问题，因为在更高一个层面上的聚类协方差估计值允许聚类之间的残差存在任何的相关性，当然也允许在 $\nu_{st}-\bar{\nu}_s$ 中存在时间序列相关。这种解决方法快速且简单。[①]不过这里的问题是将聚类问题放在更高一个层面上考虑减少了聚类数。任何的渐进推断都假设我们有大量的聚类，因为需要很多州和时期，才能很好地计算 $\nu_{st}-\bar{\nu}_s$ 与 $\nu_{st-1}-\bar{\nu}_s$ 之间的相关性。缺少聚类会带来有偏的标准误以及错误的推断。

① Arellano(1987)第一次指出可以在面板数据中用更高一层次的聚类来解决该问题。

8.2.3 小于 42 个聚类

由于聚类数目过少而导致的估计偏误是 Moulton 问题和序列相关问题中共同存在的风险因为在这两个例子中推断都是基于聚类的。当聚类过少时，我们可能会低估由随机冲击 ν_{it} 带来的序列相关问题，也可能会在 Moulton 问题中低估组内相关系数 ρ_e。在 Moulton 问题中，聚类数目就等于群数量 G。在双重差分模型中，由于我们希望在州或者其他的截面维度上计算聚类，所以相应的聚类数目就是州或者横截面群数。因此，类似于 Douglas Adam 在其科幻小说《生命、世界及宇宙》给出的最后答案是 42，我们也可以在本小节提出一个相应的问题：当使用公式(8.2.6)来调整聚类标准误时，多少数量的聚类可以允许我们做出可靠的推断？

如果 42 个聚类就足够我们调整聚类标准误并得到可靠的结论，但是过少就不够了，那么当聚类数偏低时你该怎么办呢？最好的解决办法是收集更多的数据，扩大聚类数目。但是有时我们懒得去那样做，或者说聚类数是被自然所给定的，这时就需要下面的一些方法了。不过在讲述这些例子的时候需要说明的是下面的各种处理方法在解决 Moulton 问题和序列相关问题上不是同等有效的。

(1) 对聚类标准误偏误的纠正：由于 $\mathrm{E}(\hat{e}_g\hat{e}_g')\neq\mathrm{E}(e_ge_g')=\Psi_g$，所以聚类标准是有偏的，这也正好类似于 8.1 节中的残差方差的矩阵。$\mathrm{E}(\hat{e}_g\hat{e}_g')$往往较小。一个解决办法就是通过增加残差来减少偏误。Bell 和 McCaffrey(2002)建议可以用下面的方法(减小偏误的线性化)进行调整：

$$\hat{\Psi}_g=a\tilde{e}_g\tilde{e}_g'$$
$$\tilde{e}_g=A_g\hat{e}_g$$

其中，A_g 是下面方程的解：

$$A_gA_g'=(1-H_g)^{-1}$$
$$H_g=X_g(X'X)^{-1}X_g'$$

a 是对自由度进行的调整。

这是存在聚类情况下另外一个版本的 HC_2。减小偏误的线性化(bias-reduced linearization)方法本可运用于聚类数量较少的 Moulton 问题，但是由于技术的原因，这个方法无法运用于双重差分模型中存在序列相关问题的情况。①

① 矩阵 A_g 不是唯一的，这里存在多种分解方法。Bell 和 McCaffrey(2002)使用$(1-H_g)^{-1}$的对称平方根，也即：

$$A_g=R\Lambda^{1/2}$$

其中，R 是矩阵 $(1-H_g)^{-1}$ 的特征值。Bell 和 McCaffrey(2002)中存在的一个问题是$(1-H_g)$可能不是满秩的，因此其逆矩阵并非一定存在。比如，如果回归元是个虚拟变量，而且对某个特定的聚类这个虚拟变量都等于 1，对其他，这个虚拟变量是 0，这时逆矩阵就不存在了。Bertrand 等(2004)在讨论面板数据的双重差分模型时就遇到的这个问题，该模型中存在一组完备的虚拟变量用以表征所有的州，而聚类正好是根据州划分的。

(2) 认识到基本观测单位是聚类而非出于聚类中的个体，Bell 和 McCaffrey (2002)以及 Donald 和 Lang(2007)指出可以基于自由度为 $G-K$ 的 t 分布来进行推断，而不用标准的正态分布。当 G 较小时，使用 t 分布和使用标准的正态分布之间就存在不同；使用 t 分布时，置信区间可能会更宽，因此会避免一些错误。Cameron，Gelbach 和 Miller(2008)使用来自蒙特卡洛方法的例子，报告了使用减小偏误的线性化方法和 t 分布表的结果。

(3) Donald 和 Lang(2007)指出当推断基于自由度为 $G-K$ 的 t 分布时，即使 G 较小，使用群均值进行的回归也可以很好。但是正如我们在 8.2.1 节进行的讨论，对于分组估计而言在每个组内的回归元应该是固定的。应该将加总进行到你希望聚类的那个层次，比如在 Angrist 和 Lavy(2008)中就是学校。对于序列相关问题，聚类可以是在州这一级进行，但是如果模型中已经有州效应，那么就不能再在州之间进行平均化以得到更高一级的聚类。而且，由于州之间的状态会不同，所以平均化到州这一级也是使得我们感兴趣的回归元的本身。到平均化，以我们不喜欢的方式改变了问题本身(使用组虚拟变量做工具变量使得估计值变成了工具变量估计值)。因此组均值的方法就不再是序列相关问题了。注意到如果组残差是异方差的，那么你可以使用稳健的标准误，并考虑第 8.1 节讨论过的问题会不会出现。在一些例子中，我们可以用组规模来解决这种分组残差中存在的异方差性。但是当条件期望函数为非线性时，加权的过程改变了估计量，因此加权的例子不是一目了然的(Angrist 和 Lavy(1999)选择不去对学校层面的平均值进行加权，因为他们研究中需要考虑的变异主要来自于小学校)。是否进行加权？在组一级层面进行平均化时的一个保守的办法就是我们在 8.1 节提出的经验法则：在传统标准误和稳健标准误中选择更大的那个，以此作为你度量精确性的基础。

(4) Cameron，Gelbach 和 Miller(2008)指出当组数目较小时，一类 block 自助法得到的结果很好，特别是这个结果超越了在 Stata 中计算出的聚类标准误。这个结论在 Moulton 问题和序列相关问题中都成立。不过 Cameron，Gelbach 和Miller (2008)关注于使用(枢轴量)得到的检验统计量，而我们则主要关注于标准误。

(5) 参数化纠正：对于 Moulton 问题而言，这个方法是指运用 Moulton 因子。当存在序列相关问题时，这个方法意味着自组这一层面为序列相关问题纠正其标准误。基于对 Moulton 问题做出的抽样实验以及对一系列文献的阅读，我们发现参数化方法可能是很好用的，要比 8.2.6 节中讨论的非参数聚类估计值更好，特别是如果参数模型不是很难获得的时候(比如见 Hensen(2007a)，在这篇论文中作者也提出了一种在序列相关模型中修正度量的参数偏误的方法)。但是不幸的是蒙特卡洛方法只是被人所创造出来的一个受控研究，我们并不知道参数化的假设是不是成立的。

哎呀，这里的底线还不是很明确，聚类数达到多少时会对推断问题产生致命影响这类基本问题的答案我们也尚不清楚。在这里，偏误的严重性与你研究的问题

有关，特别的，在 Moulton 问题和序列相关问题中偏误的严重性就不同。正如 Donald 和 Lang(2007)指出的：看起来在回归元固定且没有太多异方差存在的 Moulton 问题中，将数据加总到组层面的效果很好。至少，我们可以指出从分析组均值中得到的推断应该与你的结论是一致的，因为这种方法比较保守而且明显。Angrist 和 Lavy(2008)使用减小偏误的线性化方法得到标准误在学校这一层面调整聚类，并指出得到的结论和使用协方差调整的组均值相同。

一旦序列相关问题不存在了，大量的证据指出当你很幸运地在美国这样一个在州一级层面上有 51 个聚类的国家进行研究时，简单地使用 Stata 中的命令 cluster 就好了。但是也许你得研究加拿大，在那里州一级层面上的聚类只有 10 个，比 42 小了很多。Hansen(2007b)发现 Liang 和 Zeger(1986)(Stata 中的聚类)中的标准误在修正面板数据中的序列相关问题时表现良好，即使对加拿大也是如此。Hansen 同时建议使用自由度为 $G-K$ 的 t 分布作为查找临界值的来源。

聚类问题使得应用微观计量经济学家必须得忍气吞声。出于对大样本数据结论的满意，我们想要嘲笑只用一点点时间序列样本的宏观经济学家。但是笑到最后的人才是笑得最好的人：如果我们感兴趣的回归元只在组这一层面变化，比如随着时间或者州或者国家的不同而变化，那么是宏观经济学家拥有最为现实的推断方式。

8.3 附录：对简单 Moulton 因子的计算

记：

$$y_g=\begin{bmatrix}Y_{1g}\\Y_{2g}\\\vdots\\Y_{ng}\end{bmatrix}\qquad e_g=\begin{bmatrix}e_{1g}\\e_{2g}\\\vdots\\e_{ng}\end{bmatrix}$$

以及

$$y=\begin{bmatrix}y_1\\y_2\\\vdots\\y_G\end{bmatrix}\qquad x=\begin{bmatrix}l_1x_1\\l_2x_2\\\vdots\\l_Gx_G\end{bmatrix}\qquad e=\begin{bmatrix}e_1\\e_2\\\vdots\\e_G\end{bmatrix}$$

其中，l_g 是矩阵 n_g 的列向量，G 是聚类(群结构)的数目。

$$\mathrm{E}(ee')=\Psi=\begin{bmatrix}\Psi_1&0&\cdots&0\\0&\Psi_2&&\vdots\\\vdots&&\ddots&0\\0&\cdots&0&\Psi_G\end{bmatrix}$$

$$\Psi_g = \sigma_\varepsilon^2 \begin{bmatrix} 1 & \rho_\varepsilon & \cdots & \rho_\varepsilon \\ \rho_\varepsilon & 1 & & \vdots \\ \vdots & & \ddots & \rho_\varepsilon \\ \rho & \cdots & \rho_\varepsilon & 1 \end{bmatrix} = \sigma_\varepsilon^2 [(1-\rho_\varepsilon) I + \rho_\varepsilon l_g l_g']$$

其中，$\rho_\varepsilon = \dfrac{\sigma_\nu^2}{\sigma_\nu^2 + \sigma_\eta^2}$。

现在，

$$X'X = \sum_g n_g x_g x_g'$$

$$X'\Psi X = \sum_g x_g l_g' \Psi_g l_g x_g'$$

但是

$$x_g l_g' \Psi_g l_g x_g' = \sigma_\varepsilon^2 x_g l_g' \begin{bmatrix} 1+(n_g-1)\rho_\varepsilon \\ 1+(n_g-1)\rho_\varepsilon \\ \cdots \\ 1+(n_g-1)\rho_\varepsilon \end{bmatrix} x_g'$$

$$= \sigma_\varepsilon^2 n_g [1+(n_g-1)\rho_\varepsilon] x_g x_g'$$

令 $\tau_g = 1+(n_g-1)\rho_\varepsilon$，然后我们有：

$$x_g l_g' \Psi_g l_g x_g' = \sigma_\varepsilon^2 n_g \tau_g x_g x_g'$$

$$X'\Psi X = \sigma_\varepsilon^2 \sum_g n_g \tau_g x_g x_g'$$

有了这个结果以后，我们可以记：

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

$$= \sigma_\varepsilon^2 \left(\sum_g n_g x_g x_g' g\right)^{-1} \sum_g n_g \tau_g x_g x_g' g \left(\sum_g n_g x_g x_g'\right)^{-1}$$

我们要做的正是在上式得到的协方差估计值和最上二乘协方差估计值之间进行比较：

$$V_c(\hat{\beta}) = \sigma_\varepsilon^2 \left(\sum_g n_g x_g x_g'\right)^{-1}$$

如果群规模都相同，也即 $n_g = n$ 且 $\tau_g = \tau = 1+(n-1)\rho_\varepsilon$，那么：

$$V_c(\hat{\beta}) = \sigma_\varepsilon^2 \tau \left(\sum_g n x_g x_g'\right)^{-1} \sum_g n x_g x_g' \left(\sum_g n x_g x_g'\right)^{-1}$$

$$= \sigma_\varepsilon^2 \tau \left(\sum_g n x_g x_g'\right)^{-1}$$

$$= \tau V_c(\hat{\beta})$$

由此可以推出等式(8.2.4)。

最后的几句话

如果应用计量经济学很容易，那么理论家也会去做了。但是应用计量经济学也不像大部头的《计量经济学期刊》(*Econometrica*)所传达出的那种厚重和困难。提出合乎逻辑的因果问题，回归和 2SLS 几乎总是在起作用的。你的标准误也许不是很正确，但它们在大部分时间都是正确的。做最好的那个质疑者，避免困难，最重要的是：**别怕**！

术语表及名词缩写

技术性名词

2SLS 两阶段最小二乘回归，是一种工具变量估计值。

ACR 平均因果响应，对一个经过排序的处理得到的加权平均因果响应。

ANOVA 方差分析，将总方差分解为条件期望函数(CEF)的方差与条件方差的均值。

BRL 有偏的简约线性估计值，针对聚类数据得到的修正偏误的协方差矩阵估计值。

CDF 累积分布函数，随机变量的取值小于或等于某个给定值的概率。

CEF 条件期望函数，给定 X_i 后 Y_i 的期望值。

CIA 条件独立假设，是能够对回归或匹配估计值赋予一个因果解释的关键性假设。

COP 给定正的处理效应，对一个非负随机变量只考虑正值时处理组和控制组之间的差。

CQF 条件分位数函数，给定 X_i，该函数在分位数 τ 处的值可定义为 Y_i 的 τ 分位数。

DD 双重差分估计值。在最简单的形式下，它等于处理组和控制组之间随时间变化而表现出的不同。

GLS 广义最小二乘估计值，当模型出现异方差或者序列相关时使用的一种回归估计量。当条件期望函数为线性时，广义最小二乘估计可以改进估计的效率。

GMM 广义矩估计法，这是一种计量经济学的估计框架，在这个框架中，要求估计值是样本值矩和总体矩之差的加权平方和的最小化元。

HC_0-HC_3 在 MacKinnon 和 White(1985)中得到讨论的异方差一致协方差矩阵估计值。

ILS 间接最小二乘估计值，在工具变量估计中，简约式中得到的系数与第一阶段得到的系数之比。

ITT 意向处理效应，被提供处理后出现的效应。

IV 工具变量估计值或工具变量估计法

JIVE Jackknife 工具变量估计值。

LATE 局部平均处理效应，依从工具变量者表现出的因果效应。

LDVs 受限被解释变量，比如处于回归模型或其他统计模型等号左边的虚拟变量，序数或者非负的随机变量。

LIML 信息受限下的极大似然估计值，这是一种偏误小于两阶段最小二乘回

归的方法。

LM 拉格朗日乘数检验，是一种考察有估计值施加的限制是否成立的统计检验。

LPM 线性概率模型，被解释变量是虚拟变量的一种线性回归模型。

MFX 边际效应。在非线性模型中，对条件期望函数关于回归元求导，得到的就是边际效应。

MMSE 最小均方误，使预测误差的平方达到最小的值，或者使估计值和某个目标值之间距离的平方最小的值。

OLS 普通最小二乘估计值，是总体回归向量的样本值。

OVB 遗漏变量偏误，当控制变量不同时，回归估计值之间的关系。

QTE 分位数处理效应，对依从工具变量者，处理对条件分位数造成的因果效应。

RD 不连续回归设计，这是一种识别策略，其中处理、或者处理的概率、或者平均处理密度，是协变量的一个已知的不连续函数。

SEM 联立方程组模型，是一个计量经济学框架，其中变量间的因果关系由一组方程来刻画。

SSIV 分样本工具变量估计值，这是双样本工具变量估计值的一种形式。

TSIV 双样本工具变量估计值，这是一种工具变量估计法，当两个数据集包含的信息都不充足时，可以使用该方法利用两个数据集来构造工具变量估计值。

VIV 可视化工具变量估计值，当工具变量为虚拟变量时，在简约式估计值与第一阶段拟合值之间绘制直线。

WLS 加权最小二乘法，这是一种广义最小二乘法，其中加权均值是对角阵。

数据集和变量名

AFDC 对需要抚养孩子的家庭提供的一种援助，这是一个在美国已经取消了的福利计划。

AFQT 人伍资格测试，由美国军方使用，目的在于考察被征召者的文化和认知能力。

CPS 当期人口调研，这是一项大型的针对美国家庭的月度调研数据，是美国就业率数据的来源。

CED 普通教育发展证书，用以替代传统的高中毕业证，通过一项测试即可获得该证书。

IPUMS 经过整合的可公开使用的微观数据集，根据美国和其他国家的普查数据，统一整合而成。

NHIS 国民健康访问调查，在美国进行的一项大的调查，其中很多问题都与健康有关。

NLSY 美国青年面板调查数据，是一项从 1979 年的高中生开始的长期面板

调查数据。

PSAT SAT 预考，在这项考试中达到一定成绩的美国高中生可以获得美国优秀学生奖学金。

PSID 收入动态变化的面板数据研究，这是一项从 1968 年开始的针对美国家庭的面板调查。

QOB 出生季节。

RSN 随机数序列，将随机抽取的数据分配给出生日期，这项随机分配实验发生在 1970—1973 年越战时期。

SDA 服务派送区，是 649 个派送点之一，这些派送点可以提供在职培训。

SSA 社保管理部门，是美国的一个政府部门。

研究项目名称

HIE 健康保险实验，由兰德公司实施，是一个随机实验，其中参与者面临不同类型的保险项目。

JTPA 职业培训伙伴关系法，这是一项大规模的由联邦政府资助的培训项目，其中包括对项目的随机评估。

MDVE 明尼阿波利斯家暴实验，这是一个随机实验，实验中警察对家暴采取的干预措施部分取决于随机实验。

NSW 国家劳动培训示范项目，是一个在 1970 年中期完成的实验性质的培训计划，旨在向劳动力水平不高的人提供工作经验。

STAR 田纳西州小班教学实验，这是一个随机实验，旨在研究小学班级规模的影响。

WHI 妇女健康启动计划，由一系列随机实验组成，包括对荷尔蒙替代疗法的评估。

参考文献

Abadie, Alberto (2003): "Semiparametric Instrumental Variable Estimation of Treatment Response Models." *Journal of Econometrics* 113, 231–63.

Abadie, Alberto, Joshua D. Angrist, and Guido Imbens (2002): "Instrumental Variables Estimates of the Effect of Subsidized Training on the Quantiles of Trainee Earnings." *Econometrica* 70, 91–117.

Abadie, Alberto, Alexis Diamond, and Jens Hainmueller (2007): "Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program." Working Paper No. 12831. National Bureau of Economic Research, Cambridge, Mass.

Abadie, Alberto, and Guido Imbens (2006): "Large Sample Properties of Matching Estimators for Average Treatment Effects." *Econometrica* 74, 235–67.

——— (2008): "Bias-Corrected Matching Estimators for Average Treatment Effects." Mimeo. Department of Economics, Harvard University, Cambridge, Mass.

Acemoglu, Daron, and Joshua Angrist (2000): "How Large Are the Social Returns to Education? Evidence from Compulsory Schooling Laws," in *National Bureau of Economics Macroeconomics Annual 2000*, ed. Ben S. Bernanke and Kenneth S. Rogoff, pp. 9–58. MIT Press, Cambridge, Mass.

Acemoglu, Daron, Simon Johnson, and James A. Robinson (2001): "The Colonial Origins of Comparative Development: An Empirical Investigation." *The American Economic Review* 91, 1369–401.

Ackerberg, Daniel A. and Paul J. Devereux (2008): "Improved JIVE Estimators for Overidentified Linear Models With and Without Heteroskedasticity." *The Review of Economics and Statistics*, forthcoming.

ADAMS, DOUGLAS (1979): *The Hitchhiker's Guide to the Galaxy*. Pocket Books, New York.

——— (1990): *Dirk Gently's Holistic Detective Agency*. Simon & Schuster, New York.

——— (1995): *Mostly Harmless*. Harmony Books, New York.

ALTONJI, JOSEPH G., AND LEWIS M. SEGAL (1996): "Small-Sample Bias in GMM Estimation of Covariance Structures." *Journal of Business and Economic Statistics* 14, 353–66.

AMEMIYA, TAKESHI (1985): *Advanced Econometrics*. Harvard University Press, Cambridge, Mass.

AMMERMUELLER, ANDREAS, AND JÖRN-STEFFEN PISCHKE (2006): "Peer Effects in European Primary Schools: Evidence from PIRLS." Discussion Paper No. 2077. Institute for the Study of Labor (IZA), Bonn, Germany.

ANANAT, ELIZABETH, AND GUY MICHAELS (2008): "The Effect of Marital Breakup on the Income Distribution of Women with Children." *Journal of Human Resources*, forthcoming.

ANDERSON, MICHAEL (2008): "Multiple Inference and Gender Differences in the Effect of Early Intervention: A Reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects." *Journal of the American Statistical Association*, forthcoming.

ANGRIST, JOSHUA D. (1988): "Grouped Data Estimation and Testing in Simple Labor Supply Models." Working Paper No. 234. Princeton University, Industrial Relations Section, Princeton, N.J.

——— (1990): "Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records." *American Economic Review* 80, 313–35.

——— (1991): "Grouped Data Estimation and Testing in Simple Labor Supply Models." *Journal of Econometrics* 47, 243–66.

——— (1998): "Estimating the Labor Market Impact on Voluntary Military Service Using Social Security Data on Military Applicants." *Econometrica* 66, 249–88.

——— (2001): "Estimations of Limited Dependent Variable Models with Dummy Endogenous Regressors: Simple Strategies for Empirical Practice." *Journal of Business and Economic Statistics* 19, 2–16.

——— (2004): "American Education Research Changes Track." *Oxford Review of Economic Policy* 20, 198–212.

——— (2006): "Instrumental Variables Methods in Experimental Criminological Research: What, Why and How." *Journal of Experimental Criminology* 2, 22–44.

ANGRIST, JOSHUA, ERIC BETTINGER, ERIK BLOOM, ELIZABETH KING, AND MICHAEL KREMER (2002): "Vouchers for Private Schooling in Colombia: Evidence from a Randomized Natural Experiment." *The American Economic Review* 92, 1535–58.

ANGRIST, JOSHUA D., AND STACEY H. CHEN (2007): "Long-Term Consequences of Vietnam-Era Conscription: Schooling, Experience, and Earnings." Working Paper No. 13411. National Bureau of Economic Research, Cambridge, Mass.

ANGRIST, JOSHUA D., VICTOR CHERNOZHUKOV, AND IVAN FERNANDEZ-VAL (2006): "Quantile Regression Under Misspecification, with an Application to the U.S. Wage Structure." *Econometrica* 74, 539–63.

ANGRIST, JOSHUA D., AND WILLIAM N. EVANS (1998): "Children and Their Parents' Labor Supply: Evidence from Exogenous Variation in Family Size." *American Economic Review* 88, 450–477.

——— (1999): "Schooling and Labor Market Consequences of the 1970 State Abortion Reforms," in *Research in Labor Economics*, ed. Solomon W. Polachek, vol. 18, pp. 75–113. Elsevier Science, Amsterdam.

ANGRIST, JOSHUA D., KATHRYN GRADDY, AND GUIDO W. IMBENS (2000): "The Interpretation of Instrumental Variables Estimators in Simultaneous Equations Models with an Application to the Demand for Fish." *Review of Economic Studies* 67, 499–527.

ANGRIST, JOSHUA D., AND JINYONG HAHN (2004): "When to Control for Covariates? Panel Asymptotics for Estimates of Treatment Effects." *Review of Economics and Statistics* 86, 58–72.

ANGRIST, JOSHUA D., AND GUIDO W. IMBENS (1995): "Two-Stage Least Squares Estimation of Average Causal Effects in Models with Variable Treatment Intensity." *Journal of the American Statistical Association* 90, 430–42.

ANGRIST, JOSHUA D., GUIDO IMBENS, AND ALAN B. KRUEGER (1999): "Jackknife Instrumental Variables Estimation." *Journal*

of Applied Econometrics 14, 57–67.

ANGRIST, JOSHUA D., GUIDO IMBENS, AND DONALD B. RUBIN (1996): "Identification of Causal Effects Using Instrumental Variables." *Journal of the American Statistical Association* 91, 444–72.

ANGRIST, JOSHUA D., AND ALAN B. KRUEGER (1991): "Does Compulsory Schooling Attendance Affect Schooling and Earnings?" *The Quarterly Journal of Economics* 106, 976–1014.

——— (1992): "The Effect of Age at School Entry on Educational Attainment: An Application of Instrumental Variables with Moments from Two Samples." *Journal of the American Statistical Association* 418, 328–36.

——— (1994): "Why Do World War II Veterans Earn More than Nonveterans?" *Journal of Labor Economics* 12, 74–97.

——— (1995): "Split-Sample Instrumental Variables Estimates of the Return to Schooling." *Journal of Business and Economic Statistics* 13, 225–35.

——— (1999): "Empirical Strategies in Labor Economics," in *Handbook of Labor Economics*, ed. Orley C. Ashenfelter and David Card, vol. 3. North Holland, Amsterdam.

——— (2001): "Instrumental Variables and the Search for Identification: From Supply and Demand to Natural Experiments." *Journal of Economic Perspectives* 15(4), 69–85.

ANGRIST, JOSHUA D., AND GUIDO KUERSTEINER (2004): "Semiparametric Causality Tests Using the Policy Propensity Score." Working Paper No. 10975. National Bureau of Economic Research, Cambridge, Mass.

ANGRIST, JOSHUA D., AND KEVIN LANG (2004): "Does School Integration Generate Peer Effects? Evidence from Boston's Metco Program." *The American Economic Review* 94, 1613–34.

ANGRIST, JOSHUA D., AND VICTOR LAVY (1999): "Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement." *The Quarterly Journal of Economics* 114, 533–75.

——— (2008): "The Effects of High Stakes High School Achievement Awards: Evidence from a Group-Randomized Trial." *The American Economic Review*, forthcoming.

Angrist, Joshua D., Victor Lavy, and Analia Schlosser (2006): "Multiple Experiments for the Causal Link Between the Quantity and Quality of Children." Working Paper No. 06–26. Department of Economics, Massachusetts Institute of Technology, Cambridge, Mass.

Arellano, Manuel (1987): "Computing Robust Standard Errors for Within-groups Estimators." *Oxford Bulletin of Economics and Statistics* 49, 431–34.

Arellano, Manuel, and Stephen Bond (1991): "Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations." *The Review of Economic Studies* 58, 277–97.

Ashenfelter, Orley A. (1978): "Estimating the Effect of Training Programs on Earnings." *Review of Economics and Statistics* 60, 47–57.

——— (1991): "How Convincing Is the Evidence Linking Education and Income?" Working Paper No. 292. Princeton University, Industrial Relations Section, Princeton, N.J.

Ashenfelter, Orley A., and David Card (1985): "Using the Longitudinal Structure of Earnings to Estimate the Effect of Training Programs." *The Review of Economics and Statistics* 67, 648–60.

Ashenfelter, Orley A., and Alan B. Krueger (1994): "Estimates of the Economic Return to Schooling from a New Sample of Twins." *American Economic Review* 84, 1157–73.

Ashenfelter, Orley A., and Cecilia Rouse (1998): "Income, Schooling, and Ability: Evidence from a New Sample of Identical Twins." *The Quarterly Journal of Economics* 113, 253–84.

Athey, Susan, and Guido Imbens (2006): "Identification and Inference in Nonlinear Difference-in-Difference Models." *Econometrica* 74, 431–97.

Atkinson, Anthony B. (1970): "On the Measurement of Inequality." *Journal of Economic Theory* 2, 244–63.

Autor, David (2003): "Outsourcing at Will: The Contribution of Unjust Dismissal Doctrine to the Growth of Employment Outsourcing." *Journal of Labor Economics* 21, 1–42.

Autor, David, Lawrence F. Katz, and Melissa S. Kearney

(2005): "Rising Wage Inequality: The Role of Composition and Prices." Working Paper No. 11628. National Bureau of Economic Research, Cambridge, Mass.

Barnett, Steven W. (1992): "Benefits of Compensatory Preschool Education." *Journal of Human Resources* 27, 279–312.

Barnow, Burt S., Glen G. Cain, and Arthur Goldberger (1981): "Selection on Observables." *Evaluation Studies Review Annual* 5, 43–59.

Becker, Sascha O., and Andrea Ichino (2002): "Estimation of Average Treatment Effects Based on Propensity Scores." *The Stata Journal* 2, 358–77.

Bekker, Paul A. (1994): "Alternative Approximations to the Distributions of Instrumental Variable Estimators." *Econometrica* 62, 657–81.

Bekker, Paul A. and J. van der Ploeg (2005): "Instrumental Variable Estimation Based on Grouped Data." *Statistica Neerlandica* 59, 239–267.

Bell, Robert M., and Daniel F. McCaffrey (2002): "Bias Reduction in Standard Errors for Linear Regression with Multistage Samples." *Survey Methodology* 28, 169–81.

Bennedsen, Morten, Kasper M. Nielsen, Francisco Pérez-González, and Daniel Wolfenzon (2007): "Inside the Family Firm: The Role of Families in Succession Decisions and Performance." *The Quarterly Journal of Economics* 122, 647–92.

Bertrand, Marianne, Esther Duflo, and Sendhil Mullainathan (2004): "How Much Should We Trust Differences-in-Differences Estimates?" *The Quarterly Journal of Economics* 119, 249–75.

Bertrand, Marianne, and Sendhil Mullainathan (2004): "Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination." *The American Economic Review* 94, 991–1013.

Besley, Timothy, and Robin Burgess (2004): "Can Labour Market Regulation Hinder Economic Performance? Evidence from India." *The Quarterly Journal of Economics* 113, 91–134.

Bjorklund, Anders, and Markus Jantti (1997): "Intergener-

ational Income Mobility in Sweden Compared to the United States." *The American Economic Review* 87, 1009–18.

Black, Dan A., Jeffrey A. Smith, Mark C. Berger, and Brett J. Noel (2003): "Is the Threat of Reemployment Services More Effective than the Services Themselves? Evidence from Random Assignment in the UI System." *The American Economic Review* 93, 1313–27.

Black, Sandra E., Paul J. Devereux, and Kjell G. Salvanes (2005): "The More the Merrier? The Effect of Family Size and Birth Order on Children's Education." *The Quarterly Journal of Economics* 120, 669–700.

——— (2008): "Too Young to Leave the Nest: The Effect of School Starting Age." Working Paper No. 13969. National Bureau of Economic Research, Cambridge, Mass.

Bloom, Howard S. (1984): "Accounting for No-shows in Experimental Evaluation Designs." *Evaluation Review* 8, 225–246.

Bloom, Howard S., Larry L. Orr, Stephen H. Bell, George Cave, Fred Doolittle, Winston Lin, and Johannes M. Bos (1997): "The Benefits and Costs of JTPA Title II-A Programs: Key Findings from the National Job Training Partnership Act Study." *The Journal of Human Resources* 32, 549–76.

Blundell, Richard, and Stephen Bond (1998): "Initial Conditions and Moment Restrictions in Dynamic Panel Data Models." *Journal of Econometrics* 87, 115–43.

Borjas, George (1992): "Ethnic Capital and Intergenerational Mobility." *Quarterly Journal of Economics* 107, 123–50.

——— (2005): *Labor Economics*, 3rd ed. McGraw-Hill/Irwin, New York.

Bound, John, David Jaeger, and Regina Baker (1995): "Problems with Instrumental Variables Estimation when the Correlation between the Instruments and the Endogenous Variables Is Weak." *Journal of the American Statistical Association* 90, 443–50.

Bound, John, and Gary Solon (1999): "Double Trouble: On the Value of Twins-Based Estimation of the Returns of Schooling." *Economics of Education Review* 18, 169–82.

Bronars, Stephen G., and Jeff Grogger (1994): "The Economic Consequences of Unwed Motherhood: Using Twin Births as

a Natural Experiment." *The American Economic Review* 84, 1141–56.

Buchinsky, Moshe (1994): "Changes in the U.S. Wage Structure 1963–1987: Application of Quantile Regression." *Econometrica* 62, 405–58.

Buchinsky, Moshe, and Jinyong Hahn (1998): "An Alternative Estimator for the Censored Quantile Regression Model." *Econometrica* 66, 653–71.

Buse, A. (1992): "The Bias of Instrumental Variable Estimators." *Econometrica* 60, 173–80.

Cameron, Colin, Jonah Gelbach, and Douglas L. Miller (2008): "Bootstrap-Based Improvements for Inference with Clustered Errors." *The Review of Economics and Statistics* 90, 414–27.

Campbell, Donald Thomas (1969): "Reforms as Experiments." *American Psychologist* 24, 409–29.

Campbell, Donald Thomas, and Julian C. Stanley (1963): *Experimental and Quasi-experimental Designs for Research.* Rand McNally, Chicago.

Card, David (1992): "Using Regional Variation to Measure the Effect of the Federal Minimum Wage." *Industrial and Labor Relations Review* 46, 22–37.

——— (1995): "Earnings, Schooling and Ability Revisited," in *Research in Labor Economics*, ed. Solomon W. Polachek, vol. 14, pp. 23–48. JAI Press, Greenwich, Conn.

——— (1996): "The Effect of Unions on the Structure of Wages: A Longitudinal Analysis." *Econometrica* 64, 957–79.

——— (1999): "The Causal Effect of Education on Earnings," in *Handbook of Labor Economics*, ed. Orley C. Ashenfelter and David Card, vol. 3. North Holland, Amsterdam.

Card, David, and Alan Krueger (1994): "Minimum Wages and Employment: A Case Study of the Fast Food Industry in New Jersey and Pennsylvania." *The American Economic Review* 84, 772–84.

——— (2000): "Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania: Reply."

The American Economic Review 90, 1397–420.

CARD, DAVID, AND THOMAS LEMIEUX (1996): "Wage Dispersion, Returns to Skill, and Black-White Differentials." *Journal of Econometrics* 74, 316–61.

CARD, DAVID E., AND DANIEL SULLIVAN (1988): "Measuring the Effect of Subsidized Training on Movements in and out of Employment." *Econometrica* 56, 497–530.

CARDELL, NICHOLAS SCOTT, AND MARK MYRON HOPKINS (1977): "Education, Income, and Ability: A Comment." *Journal of Political Economy* 85, 211–15.

CHAMBERLAIN, GARY (1977): "Education, Income, and Ability Revisited." *Journal of Econometrics* 5, 241–57.

——— (1978): "Omitted Variables Bias in Panel Data: Estimating the Returns to Schooling." *Annales De L'INSEE* 30–31, 49–82.

——— (1984): "Panel Data," in *Handbook of Econometrics*, ed. Zvi Griliches, and Michael D. Intriligator, vol. 2, pp. 1247–318. North Holland, Amsterdam.

——— (1994): "Quantile Regression, Censoring and the Structure of Wages," in *Proceedings of the Sixth World Congress of the Econometrics Society, Barcelona, Spain*, ed. Christopher A. Sims, and Jean-Jacques Laffont, pp. 179–209. Cambridge University Press, New York.

CHAMBERLAIN, GARY, AND EDWARD E. LEAMER (1976): "Matrix Weighted Averages and Posterior Bounds." *Journal of the Royal Statistical Society, Series B* 38, 73–84.

CHERNOZHUKOV, VICTOR, AND CHRISTIAN HANSEN (2005): "An IV Model of Quantile Treatment Effects." *Econometrica* 73, 245–61.

——— (2008): "The Reduced Form: A Simple Approach to Inference with Weak Instruments." *Economics Letters* 100, 68–71.

CHERNOZHUKOV, VICTOR, AND H. HONG (2002): "Three-step Censored Quantile Regression and Extramarital Affairs." *Journal of the America Statistical Assoc.* 92, 872–82.

CHERNOZHUKOV, VICTOR, IVAN FERNANDEZ-VAL, AND BLAISE MELLY (2008): "Inference on Counterfactual Distributions." Working Paper No. 08–16. MIT Department of Economics, Cambridge, Mass.

Chesher, Andrew, and Gerald Austin (1991): "The Finite-Sample Distributions of Heteroskedasticity Robust Wald Statistics." *Journal of Econometrics* 47, 153–73.

Chesher, Andrew, and Ian Jewitt (1987): "The Bias of the Heteroskedasticity Consistent Covariance Estimator." *Econometrica* 55, 1217–22.

Cochran, William G. (1965): "The Planning of Observational Studies of Human Populations." *Journal of the Royal Statistical Society, Series A* 128, 234–65.

Cook, Thomas D. (2008): "Waiting for Life to Arrive: A History of the Regression-Discontinuity Design in Psychology, Statistics, and Economics." *Journal of Econometrics* 142, 636–54.

Cook, Thomas D., and Vivian C. Wong (2008): "Empirical Tests of the Validity of the Regression-Discontinuity Design." *Annales d'Economie et de Statistique*, forthcoming.

Crump, Richard K., V. Joseph Hotz, Guido W. Imbens, and Oscar A. Mitnik (2009): "Dealing with Limited Overlap in the Estimation of Average Treatment Effects." *Biometrica*, forthcoming.

Cruz, Luiz M., and Marcelo J. Moreira (2005): "On the Validity of Econometric Techniques with Weak Instruments: Inference on Returns to Education Using Compulsory School Attendance Laws." *Journal of Human Resources* 40, 393–410.

Currie, Janet, and Aaron Yelowitz (2000): "Are Public Housing Projects Good for Kids?" *Journal of Public Economics* 75, 99–124.

Davidon, Russell, and James G. MacKinnon (1993): *Estimation and Inference in Econometrics*. Oxford University Press, New York.

Dearden, Lorraine, Sue Middleton, Sue Maguire, Karl Ashworth, Kate Legge, Tracey Allen, Kim Perrin, Erich Battistin, Carl Emmerson, Emla Fitzsimons, and Costas Meghir (2003): "The Evaluation of Education Maintenance Allowance Pilots: Three Years' Evidence. A Quantitative Evaluation." Research Report No. 499. Department for Education and Skills, DFES Publications, Nottingham, UK.

Deaton, Angus (1985): "Panel Data from a Time Series of Cross-sections." *Journal of Econometrics* 30, 109–126.

——— (1997): *The Analysis of Household Surveys: A Microeconometric Approach to Development Policy*. Johns Hopkins University Press for the World Bank, Baltimore, Md.

Dee, Thomas S., and William N. Evans (2003): "Teen Drinking and Educational Attainment: Evidence from Two-Sample Instrumental Variables Estimates." *Journal of Labor Economics* 21, 178–209.

DeGroot, Morris H., and Mark J. Schervish (2001): *Probability and Statistics*, 3rd ed. Addison-Wesley, Boston.

Dehejia, Rajeev H. (2005): "Practical Propensity Score Matching: A Reply to Smith and Todd." *Journal of Econometrics* 125, 355–364.

Dehejia, Rajeev H., and Sadek Wahba (1999): "Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs." *Journal of the American Statistical Association* 94, 1053–62.

Deming, David, and Susan Dynarski (2008): "The Lengthening of Childhood." *The Journal of Economic Perspectives* 22(3), 71–92.

Devereux, Paul J. (2007): "Improved Errors-in-variables Estimators for Grouped Data." *The Journal of Business and Economic Statistics* 27, 278–287.

Donald, Stephen G., and Kevin Lang (2007): "Inference with Difference-in-Differences and Other Panel Data." *Review of Economics and Statistics* 89, 221–33.

Duan, Naihua, Willard D. Manning, Jr., Carl N. Morris, and Joseph P. Newhouse (1983): "A Comparison of Alternative Models for the Models for the Demand for Medical Care." *Journal of Business & Economic Statistics* 1, 115–26.

——— (1984): "Choosing Between the Sample-Selection Model and the Multi-Part Model." *Journal of Business & Economic Statistics* 2, 283–289.

Durbin, James (1954): "Errors in Variables." *Review of the International Statistical Institute* 22, 23–32.

Eicker, Friedhelm (1967): "Limit Theorems for Regressions with Unequal and Dependent Errors," in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, pp. 59–82. University of California Press, Berkeley and Los Angeles.

FINN, JEREMY D., AND CHARLES M. ACHILLES (1990): "Answers and Questions About Class Size: A Statewide Experiment." *American Educational Research Journal* 28, 557–77.

FIRPO, SERGIO (2007): "Efficient Semiparametric Estimation of Quantile Treatment Effects." *Econometrica* 75, 259–76.

FLORES-LAGUNES, ALFONSO (2007): "Finite Sample Evidence of IV Estimators under Weak Instruments." *Journal of Applied Econometrics* 22, 677–94.

FREEDMAN, DAVID (2005): "Linear Statistical Models for Causation: A Critical Review," in *The Wiley Encyclopedia of Statistics in Behavioral Science*, ed. B. Everitt, and D. Howell. John Wiley, Chichester, UK.

FREEMAN, RICHARD (1984): "Longitudinal Analyses of the Effect of Trade Unions." *Journal of Labor Economics* 3, 1–26.

FRISCH, RAGNAR, AND FREDERICK V. WAUGH (1933): "Partial Time Regression as Compared with Individual Trends." *Econometrica* 1, 387–401.

FRÖLICH, MARKUS, AND BLAISE MELLY (2007): "Unconditional Quantile Treatment Effects Under Endogeneity." Working Paper No. CWP32/07. Centre for Microdata Methods and Practice.

FRYER, ROLAND G., AND STEVEN D. LEVITT (2004): "The Causes and Consequences of Distinctively Black Names." *The Quarterly Journal of Economics* 119, 767–805.

GALTON, FRANCIS (1886): "Regression Towards Mediocrity in Hereditary Stature." *Journal of the Anthropological Institute* 15, 246–63.

GOLDBERGER, ARTHUR S. (1972): "Selection Bias in Evaluating Treatment Effects: Some Formal Illustrations." Working paper. Department of Economics, University of Wisconsin, Madison.

——— (1991): *A Course in Econometrics*. Harvard University Press, Cambridge, Mass.

GOSLING, AMANDA, STEPHEN MACHIN, AND COSTAS MEGHIR (2000): "The Changing Distribution of Male Wages in the U.K." *Review of Economic Studies* 67, 635–66.

GRANGER, CLIVE W. J. (1969): "Investigating Causal Relations by

Econometric Models and Cross-spectral Methods." *Econometrica* 37, 424–38.

GRILICHES, ZVI (1977): "Estimating the Returns to Schooling: Some Econometric Problems." *Econometrica* 45, 1–22.

GRILICHES, ZVI, AND JERRY A. HAUSMAN (1986): "Errors in Variables in Panel Data." *Journal of Econometrics* 31, 93–118.

GRILICHES, ZVI, AND WILLIAM M. MASON (1972): "Education, Income, and Ability." *Journal of Political Economy* 80, S74–103.

GRUMBACH, KEVIN, DENNIS KEANE, AND ANDREW BINDMAN (1993): "Primary Care and Public Emergency Department Overcrowding." *American Journal of Public Health* 83, 372–78.

GURYAN, JONATHAN (2004): "Desegregation and Black Dropout Rates." *American Economic Review* 94, 919–43.

HAAVELMO, TRYGVE (1944): "The Probability Approach in Econometrics." *Econometrica* 12, S1–115.

HAHN, JINYONG (1998): "On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects." *Econometrica* 66, 315–31.

HAHN, JINYONG, PETRA TODD, AND WILBUR VAN DER KLAAUW (2001): "Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design." *Econometrica* 69, 201–9.

HANSEN, CHRISTIAN B. (2007a): "Asymptotic Properties of a Robust Variance Matrix Estimator for Panel Data When T Is Large." *Journal of Econometrics* 141, 597–620.

——— (2007b): "Generalized Least Squares Inference in Panel and Multilevel Models with Serial Correlation and Fixed Effects." *Journal of Econometrics* 140, 670–94.

HANSEN, LARS PETER (1982): "Large Sample Properties of Generalized Method of Moments Estimators." *Econometrica* 50, 1029–54.

HAUSMAN, JERRY (1978): "Specification Tests in Econometrics." *Econometrica* 46, 1251–71.

——— (1983): "Specification and Estimation of Simultaneous Equation Models," in *Handbook of Econometrics*, ed. Zvi Griliches, and Michael Intriligator, vol. 1, pp. 391–448. North Holland, Amsterdam.

——— (2001): "Mismeasured Variables in Econometric Analysis: Problems from the Right and Problems from the Left." *Journal of Econometric Perspectives* 15(4), 57–67.

HAUSMAN, JERRY, WHITNEY NEWEY, TIEMEN WOUTERSEN, JOHN CHAO, AND NORMAN SWANSON (2008): "Instrumental Variable Estimation with Heteroskedasticity and Many Instruments." Unpublished manuscript. Department of Economics, Massachusetts Institute of Technology, Cambridge, Mass.

HAY, JOEL W., AND RANDALL J. OLSEN (1984): "Let Them Eat Cake: A Note on Comparing Alternative Models of the Demand for Medical Care." *Journal of Business & Economic Statistics* 2, 279–82.

HECKMAN, JAMES J. (1978): "Dummy Endogenous Variables in a Simultaneous Equations System." *Econometrica* 46, 695–712.

HECKMAN, JAMES J., HIDEHIKO ICHIMURA, AND PETRA E. TODD (1998): "Matching as as Econometric Evaluation Estimator." *Review of Economic Studies* 62, 261–94.

HECKMAN, JAMES J., JEFFREY SMITH, AND NANCY CLEMENTS (1997): "Making the Most out of Programme Evaluations and Social Experiments: Accounting for Heterogeneity in Programme Impacts." *The Review of Economic Studies* 64, 487–535.

HIRANO, KEISUKE, GUIDO W. IMBENS, AND GEERT RIDDER (2003): "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score." *Econometrica* 71, 1161–89.

HOAGLIN, DAVID C., AND ROY E. WELSCH (1978): "The Hat Matrix in Regression and ANOVA." *The American Statistician* 32, 17–22.

HOLLAND, PAUL W. (1986): "Statistics and Causal Inference." *Journal of the American Statistical Association* 81, 945–70.

HOLTZ-EAKIN, DOUGLAS, WHITNEY NEWEY, AND HARVEY S. ROSEN (1988): "Estimating Vector Autoregressions with Panel Data." *Econometrica* 56, 1371–1395.

HOROWITZ, JOEL L. (1997): "Bootstrap Methods in Econometrics: Theory and Numerical Performance," in *Advances in Economics and Econometrics: Theory and Applications*, ed. David M. Kreps and Kenneth F. Wallis, vol. 3, pp. 188–222. Cambridge University Press, Cambridge, UK.

——— (2001): "The Bootstrap," in *Handbook of Econometrics*, ed.

James J. Heckman and Edward E. Leamer, vol. 5, pp. 3159–228. Elsevier Science, Amsterdam.

Horvitz, Daniel G., and Donovan J. Thompson (1952): "A Generalization of Sampling Without Replacement from a Finite Population." *Journal of the American Statistical Association* 47, 663–85.

Hoxby, Caroline (2000): "The Effects of Class Size on Student Achievement: New Evidence from Population Variation." *The Quarterly Journal of Economics* 115, 1239–85.

Hsia, Judith, Robert D. Langer, JoAnn E. Manson, Lewis Kuller, Karen C. Johnson, Susan L. Hendrix, Mary Pettinger, Susan R. Heckbert, Nancy Greep, Sybil Crawford, Charles B. Eaton, John B. Kostis, Pat Caralis, Ross Prentice, for the Women's Health Initiative Investigators (2006): "Conjugated Equine Estrogens and Coronary Heart Disease: The Women's Health Initiative." *Archives of Internal Medicine* 166, 357–65.

Imbens, Guido (2000): "The Role of the Propensity Score in Estimating Dose-Response Functions." *Biometrika* 87, 706–10.

——— (2004): "Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review." *The Review of Economics and Statistics* 86, 4–29.

Imbens, Guido, and Joshua Angrist (1994): "Identification and Estimation of Local Average Treatment Effects." *Econometrica*, 62, 467–76.

Imbens, Guido, and Thomas Lemieux (2008): "Regression Discontinuity Designs: A Guide to Practice." *Journal of Econometrics* 142, 615–35.

Inoue, Atsushi, and Gary Solon (2009): "Two-Sample Instrumental Variables Estimators." *The Review of Economics and Statistics*, forthcoming.

Jappelli, Tullio, Jörn-Steffen Pischke, and Nicholas S. Souleles (1998): "Testing for Liquidity Constraints in Euler Equations with Complementary Data Sources." *The Review of Economics and Statistics* 80, 251–62.

Johnson, Norman L., and Samuel Kotz (1970): *Distributions*

in Statistics: Continuous Distributions, vol. 2. John Wiley, New York.

KAUERMANN, GORAN, AND RAYMOND J. CARROLL (2001): "A Note on the Efficiency of Sandwich Covariance Estimation." *Journal of the American Statistical Association* 96, 1387–96.

KELEJIAN, HARRY H. (1971): "Two Stage Least Squares and Econometric Systems Linear in Parameters but Nonlinear in the Endogenous Variables." *Journal of the American Statistical Association* 66, 373–74.

KENNAN, JOHN (1995): "The Elusive Effects of Minimum Wages." *Journal of Economic Literature* 33, 1950–65.

KÉZDI, GÁBOR (2004): "Robust Standard Error Estimation in Fixed-Effects Panel Models." *Hungarian Statistical Review* (*Special English Volume*) 9, 95–116.

KISH, LESLIE (1965): "Sampling Organizations and Groups of Unequal Sizes." *American Sociological Review* 30, 564–72.

KLOEK, TEUN (1981): "OLS Estimation in a Model Where a Microvariable Is Explained by Aggregates and Contemporaneous Disturbances Are Equicorrelated." *Econometrica* 49, 205–7.

KNIGHT, KEITH (2000): *Mathematical Statistics*. Chapman & Hall/CRC, Boca Raton, Fla.

KOENKER, ROGER (2005): *Quantile Regression*. Cambridge University Press, Cambridge, UK.

KOENKER, ROGER, AND GILBERT BASSETT (1978): "Regression Quantiles." *Econometrica* 46, 33–50.

KOENKER, ROGER, AND STEPHEN PORTNOY (1996): "Quantile Regression." Working Paper No. 97-0100. College of Commerce and Business Administration, Office of Research, University of Illinois at Urbana-Champaign.

KRUEGER, ALAN B. (1999): "Experimental Estimates of Education Production Functions." *The Quarterly Journal of Economics* 114, 497–532.

KUGLER, ADRIANA, JUAN F. JIMENO, AND VIRGINIA HERNANZ (2005): "Employment Consequences of Restrictive Permanent Contracts: Evidence from Spanish Labor Market Reforms." FEDEA Working Paper No. 2003–14. FEDEA: Foundation for Applied Economic Research, Madrid, Spain.

LaLonde, Robert J. (1986): "Evaluating the Econometric Evaluations of Training Programs Using Experimental Data." *The American Economic Review* 76, 602–20.

——— (1995): "The Promise of Public Sector-Sponsored Training Programs." *Journal of Economic Perspectives* 93(2), 149–68.

Lee, David S. (2008): "Randomized Experiments from Non-random Selection in U.S. House Elections." *Journal of Econometrics* 142, 675–97.

Lemieux, Thomas (2008): "The Changing Nature of Wage Inequality." *Journal of Population Economics* 21, 21–48.

Liang, Kung-Yee, and Scott L. Zeger (1986): "Longitudinal Data Analysis Using Generalized Linear Models." *Biometrika* 73, 13–22.

Machado, Jose, and Jose Mata (2005): "Counterfactual Decompositions of Changes in Wage Distributions Using Quantile Regression." *Journal of Applied Econometrics* 20, 445–65.

MacKinnon, James G., and Halbert White (1985): "Some Heteroskedasticity Consistent Covariance Matrix Estimators with Improved Finite Sample Properties." *Journal of Econometrics* 29, 305–25.

Maddala, Gangadharrao Soundalyarao (1983): "Methods of Estimation for Models of Markets with Bounded Price Variation." *International Economic Review* 24, 361–78.

Mammen, Enno (1993): "Bootstrap and Wild Bootstrap for High Dimensional Linear Models." *Annals of Statistics* 21, 255–85.

Manning, Willard G., Joseph P. Newhouse, Naihua Duan, Emmett B. Keeler, Arleen Leibowitz, and Susan M. Marquis (1987): "Health Insurance and the Demand for Medical Care: Evidence from a Randomized Experiment." *American Economic Review* 77, 251–77.

Manski, Charles F. (1991): "Regression." *Journal of Economic Literature* 29, 34–50.

Mariano, Roberto S. (2001): "Simultaneous Equation Model Estimators: Statistical Properties," in *A Companion to Theoretical Econometrics*, ed. B. Baltagi. Blackwell, Oxford, UK.

McClellan, Mark B., Barbara J. McNeil, and Joseph P. New-

HOUSE (1994): "Does More Intensive Treatment of Acute Myocardial Infarction Reduce Mortality? Analysis Using Instrumental Variables." *Journal of the American Medical Association* 272, 859–66.

MCCRARY, JUSTIN (2008): "Manipulation of the Running Variable in the Regression Discontinuity Design: A Density Test." *Journal of Econometrics* 142, 698–714.

MCDONALD, JOHN F., AND ROBERT A. MOFFITT (1980): "The Uses of Tobit Analysis." *The Review of Economics and Statistics* 62, 318–21.

MELTZER, ALLAN H., AND SCOTT F. RICHARD (1983): "Tests of a Rational Theory of the Size of Government." *Public Choice* 41, 403–18.

MESSER, KAREN, AND HALBERT WHITE (1984): "A Note on Computing the Heteroskedasticity Consistent Covariance Matrix Using Instrumental Variables Techniques." *Oxford Bulletin of Economics and Statistics* 46, 181–84.

MEYER, BRUCE D., W. KIP VISCUSI, AND DAVID L. DURBIN (1995): "Workers' Compensation and Injury Duration: Evidence from a Natural Experiment." *The American Economic Review* 85, 322–40.

MEYER, BRUCE D., AND DAN T. ROSENBAUM (2001): "Welfare, the Earned Income Tax Credit, and the Labor Supply of Single Mothers." *The Quarterly Journal of Economics* 116, 1063–114.

MILGRAM, STANLEY (1963): "Behavioral Study of Obedience." *Journal of Abnormal and Social Psychology* 67, 371–78.

MOFFITT, ROBERT (1992): "Incentive Effects of the U.S. Welfare System: A Review." *Journal of Economic Literature* 30, 1–61.

MORGAN, MARY S. (1990): *The History of Econometric Ideas*. Cambridge University Press, Cambridge, UK.

MOULTON, BRENT (1986): "Random Group Effects and the Precision of Regression Estimates." *Journal of Econometrics* 32, 385–97.

NELSON, CHARLES R., AND RICHARD STARTZ (1990a): "The Distribution of the Instrumental Variables Estimator and Its *t*-Ratio when the Instrument Is a Poor One." *Journal of Business* 63, 125–40.

——— (1990b): "Some Further Results on the Exact Small-Sample

Properties of the Instrumental Variable Estimator." *Econometrica* 58, 967–76.

NEUMARK, DAVID, AND WILLIAM WASCHER (1992): "Employment Effects of Minimum and Subminimum Wages: Panel Data on State Minimum Wage Laws." *Industrial and Labor Relations Review* 46, 55–81.

NEWEY, WHITNEY K. (1985): "Generalized Method of Moments Specification Testing." *Journal of Econometrics* 29, 299–56.

——— (1990): "Semiparametric Efficiency Bounds." *Journal of Applied Econometrics* 5, 99–135.

NEWEY, WHITNEY K., AND KENNETH D. WEST (1987): "Hypothesis Testing with Efficient Method of Moments Estimation." *International Economic Review* 28, 777–87.

NICKELL, STEPHEN (1981): "Biases in Dynamic Models with Fixed Effects." *Econometrica* 49, 1417–26.

OBENAUER, MARIE, AND BERTHA VON DER NIENBURG (1915): "Effect of Minimum Wage Determinations in Oregon." Bulletin of the U.S. Bureau of Labor Statistics, No. 176. Washington, D.C., U.S. Government Printing Office.

OREOPOULOS, PHILIP (2006): "Estimating Average and Local Average Treatment Effects of Education When Compulsory Schooling Laws Really Matter." *American Economic Review* 96, 152–75.

ORR, LARRY L., HOWARD S. BLOOM, STEPHEN H. BELL, FRED DOOLITTLE, AND WINSTON LIN (1996): *Does Training for the Disadvantaged Work? Evidence from the National JTPA Study.* Urban Institute Press, Washington, D.C.

PFEFFERMAN, DANIEL (1993): "The Role of Sampling Weights When Modeling Survey Data." *International Statistical Review* 61, 317–37.

PISCHKE, JÖRN-STEFFEN (2007): "The Impact of Length of the School Year on Student Performance and Earnings: Evidence from the German Short School Years." *Economic Journal* 117, 1216–42.

PORTER, JACK (2003): "Estimation in the Regression Discontinuity Model." Unpublished manuscript. Department of Economics, University of Wisconsin, Madison, Wis.

POTERBA, JAMES, STEVEN VENTI, AND DAVID WISE (1995): "Do 401K

Contributions Crowd Out Other Personal Savings." *Journal of Public Economics* 58, 1–32.

POWELL, JAMES L. (1986): "Censored Regression Quantiles." *Journal of Econometrics* 32, 143–55.

——— (1989): "Semiparametric Estimation of Censored Selection Models." Unpublished manuscript. Department of Economics, University of Wisconsin, Madison.

PRAIS, SIG J., AND JOHN AITCHISON (1954): "The Grouping of Observations in Regression Analysis." *Revue de l'Institut International de Statistique* (*Review of the International Statistical Institute*) 22, 1–22.

REIERSOL, OLAV (1941): "Confluence Analysis by Means of Lag Moments and Other Methods of Confluence Analysis." *Econometrica* 9, 1–24.

ROBINS, JAMES M., STEVEN D. MARK, AND WHITNEY K. NEWEY (1992): "Estimating Exposure Effects by Modeling the Expectation of Exposure Conditional on Confounders." *Biometrics* 48, 479–95.

ROSENBAUM, PAUL R. (1984): "The Consequences of Adjustment for a Concomitant Variable That Has Been Affected by the Treatment." *Journal of the Royal Statistical Society, Series A* 147, 656–66.

——— (1995): *Observational Studies*. Springer-Verlag, New York.

ROSENBAUM, PAUL R., AND DONALD B. RUBIN (1983): "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika* 70, 41–55.

——— (1985): "The Bias Due to Incomplete Matching." *Biometrics* 41, 106–16.

ROSENZWEIG, MARK R., AND KENNETH I. WOLPIN (1980): "Testing the Quantity-Quality Fertility Model: The Use of Twins as a Natural Experiment." *Econometrica* 48, 227–240.

RUBIN, DONALD B. (1973): "Matching to Remove Bias in Observational Studies." *Biometrics* 29, 159–83.

——— (1974): "Estimating the Causal Effects of Treatments in Randomized and Non-Randomized Studies." *Journal of Educational Psychology* 66, 688–701.

——— (1977): "Assignment to a Treatment Group on the Basis of a Covariate." *Journal of Educational Statistics* 2, 1–26.

——— (1991): "Practical Implications of Modes of Statistical Inference for Causal Effects and the Critical Role of the Assignment Mechanism." *Biometrics* 47, 1213–34.

RUUD, PAUL A. (1986): "Consistent Estimation of Limited Dependent Variable Models Despite Misspecification of Distribution." *Journal of Econometrics* 32, 157–87.

SHADISH, WILLIAM R., THOMAS D. COOK, AND DONALD T. CAMPBELL (2002): *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Houghton-Mifflin, Boston.

SHERMAN, LAWRENCE W., AND RICHARD A. BERK (1984): "The Specific Deterrent Effects of Arrest for Domestic Assault." *American Sociological Review* 49, 261–72.

SHORE-SHEPPARD, LARA (1996): "The Precision of Instrumental Variables Estimates with Grouped Data." Working Paper No. 374. Princeton University, Industrial Relations Section, Princeton, N.J.

SMITH, JEFFREY A., AND PETRA E. TODD (2001): "Reconciling Conflicting Evidence on the Performance of Propensity-Score Matching Methods." *American Economic Review* 91, 112–18.

——— (2005): "Does Matching Overcome LaLonde's Critique of Nonexperimental Estimators?" *Journal of Econometrics* 125, 305–53.

SNOW, JOHN (1855): *On the Mode of Communication of Cholera*, 2nd ed. John Churchill, London.

STIGLER, STEPHEN M. (1986): *The History of Statistics: The Measurement of Uncertainty Before 1900*. The Belknap Press of Harvard University Press, Cambridge, Mass.

STOCK, JAMES H., AND FRANCESCO TREBBI (2003): "Who Invented Instrumental Variables Regression?" *The Journal of Economic Perspectives* 17(3), 177–94.

STOCK, JAMES H., JONATHAN H. WRIGHT, AND MOTOHIRO YOGO (2002): "A Survey of Weak Instruments and Weak Identification in Generalized Method of Moments." *Journal of Business & Economic Statistics* 20, 518–29.

TAUBMAN, PAUL (1976): "The Determinants of Earnings: Genetics,

Family and Other Environments: A Study of White Male Twins." *American Economic Review* 66, 858–70.

THISTLEWAITE, DONALD L., AND DONALD T. CAMPBELL (1960): "Regression-Discontinuity Analysis: An Alternative to the Ex Post Facto Experiment." *Journal of Educational Psychology* 51, 309–17.

TROCHIM, WILLIAM (1984): *Research Designs for Program Evaluation: The Regression Discontinuity Design*. Sage Publications, Beverly Hills, Calif.

VAN DER KLAAUW, WILBERT (2002): "Estimating the Effect of Financial Aid Offers on College Enrollment: A Regression-Discontinuity Approach." *International Economic Review* 43, 1249–1287.

WALD, ABRAHAM (1940): "The Fitting of Straight Lines if Both Variables Are Subject to Error." *Annals of Mathematical Statistics* 11, 284–300.

——— (1943): "Tests of Statistical Hypotheses Concerning Several Parameters When the Number of Observations Is Large." *Transactions of the American Mathematical Society* 54, 426–82.

WHITE, HALBERT (1980a): "A Heteroskedasticity-Consistent Covariance Matric Estimator and a Direct Test for Heteroskedasticity." *Econometrica* 48, 817–38.

——— (1980b): "Using Least Squares to Approximate Unknown Regression Functions." *International Economic Review* 21, 149–70.

——— (1982): "Instrumental Variables Regression with Independent Observations." *Econometrica* 50, 483–99.

——— (1984): *Asymptotic Theory for Econometricians*. Academic Press, Orlanda, Fla.

WOOLDRIDGE, JEFFREY (2003): "Cluster-Sample Methods in Applied Econometrics." *American Economic Review* 93, 133.

——— (2005): "Fixed-Effects and Related Estimators for Correlated Random-Coefficient and Treatment-Effect Panel Data Models." *The Review of Economics and Statistics* 87, 385–90.

——— (2006): *Introductory Econometrics: A Modern Approach*. Thomson/South-Western, Mason, Oh.

WRIGHT, PHILLIP G. (1928): *The Tariff on Animal and Vegetable Oils*. Macmillan, New York.

YANG, SONG, LI HSU, AND LUEPING ZHAO (2005): "Combining Asymptotically Normal Tests: Case Studies in Comparison of Two Groups." *Journal of Statistical Planning and Inference* 133, 139–58.

YELOWITZ, AARON (1995): "The Medicaid Notch, Labor Supply and Welfare Participation: Evidence from Eligibility Expansions." *The Quarterly Journal of Economics* 110, 909–39.

YITZHAKI, SHLOMO (1996): "On Using Linear Regression in Welfare Economics." *Journal of Business and Economic Statistics* 14, 478–86.

YULE, GEORGE UDNY (1895): "On the Correlation of Total Pauperism with Proportion of Out-Relief." *The Economic Journal* 5, 603–11.

——— (1897): "On the Theory of Correlation." *Journal of the Royal Statistical Society* 60, 812–54.

——— (1899): "An Investigation into the Causes of Changes in Pauperism in England, Chiefly During the Last Two Intercensal Decades (Part I)." *Journal of the Royal Statistical Society* 62, 249–95.

译 后 记

这本著名的应用计量经济学教材能够有幸由我们所翻译，实在有些偶然，也很感到荣幸。

2009年，我们在网络上下载到了本书的英文电子初稿，当时我们正在为如何一起学习应用计量经济学而苦恼，这本书给我们带来了无尽的惊喜。后来，复旦大学的挚友陆铭教授来浙大讲授应用计量经济学，也大力推荐此书。我们在之后的翻译过程中，得到了他不少的鼓励和支持，在此特别向他表示感谢。

正是由于这些原因，我们突发奇想，这么好的书，为什么不把它介绍给更多的国内经济学同行和同学们呢?! 于是，我们就找到了格致出版社的朋友来问询这本书的版权情况。没想到，很快就收到了格致出版社李娜编辑的热情回信。在回信中得知，李娜编辑自己已经开始翻译了这本书的第1章，愿意与我们共同合作，很巧的是，我们已经翻译好了本书的前面两章，并发给了李娜编辑看。在这里我们特别感佩李娜编辑的敬业精神，她立刻同意这本书由我们二人来翻译完成，虽然她已经付出了不少心血，却毅然废稿。我们两人甚为李娜编辑的认真负责和敬业态度所敬服，因此对这一翻译工作更是诚惶诚恐，不敢懈怠。

在翻译这本书的过程中，我们发现首先要有一定的计量经济学基础才能顺利地阅读本书。特别是，我们甚至在中文文献中找不到书中出现的一些计量经济学方法所对应的中文名称。因此，在整个翻译过程中，我们动了很多脑筋来思考如何“信、达、雅”地表现出这些方法的实质。当然，面于两位译者在计量经济学、英文理解和中文表达上的造诣不精，其中存在不少可改进的空间。其次，从本书的名字上看，“基本无害”的名称显示出两位原作者的强大信心。回顾全文，本书只讲了回归、工具变量、双重差分、断点回归、分位数回归以及时间序列数据中出现的非标准误的处理，这六部分内容是两位作者认为大家都该学，学了没坏处的计量经济学方法，“基本无害”的名字也正是源于此。如此自信甚至有些狂妄的姿态自然引来多方关注和评论，因此，在2010年美国经济学会出版的期刊 *Journal of Economic Perspective* 春季卷中(Vol. 24, No. 2, Spring 2010)，本书两位原作者和来自计量经济学多个领域的人士就计量经济学中什么是重要的、其方法的改进应该向何处去等关键问题进行了一次辩论。我们认为，任何理论、书籍都有其局限性，因此想进一步深入了解本书内涵以及局限性的经济学同行和同学们不妨从这次争论开始。而且幸运的是，这一组文章在美国经济学会的官方网站上可以免费下载。

在本书的翻译过程中，格致出版社的麻俊生副总编、钱敏编辑多次表达了对本书翻译进展的关怀和关心，并对我们的拖延表现了十足的耐心，特向他们二位表示感激！两位译者的博士生导师浙江大学经济学院的史晋川教授对于我们的工作多

有鼓励；浙江财经学院的副校长卢新波教授，经贸学院谢作诗院长在本书翻译过程中都给予了不少帮助和支持；浙江财经学院的优秀毕业研究生，将赴复旦大学攻读博士学位的刘志阔同学，还有浙大经济学院的王昊博士，在翻译过程中提出了不少建议，在此一并致谢！

本书是由浙江财经学院李井奎老师和浙江大学郎金焕博士生通力合作完成的，郎金焕同时还承担了本书的统校工作，他是浙大数学系出身，校对十分认真，希望我们的工作能够嘉惠学林，有助于中国计量经济学事业的发展。毕竟我们的水平有限，难免有不少错误，谨就此书就教于列位方家，不吝赐教！