格致方法 · 定量研究系列　吴晓刚　主编

虚拟变量回归

[美] 梅丽莎 · A.海蒂（Mellissa A.Hardy）著
贺光烨 译

- 革新研究理念
- 丰富研究工具
- 最权威、最前沿的定量研究方法指南

3

格致出版社　上海人民出版社

格致方法 · 定量研究系列　吴晓刚　主编

虚拟变量回归

[美] 梅丽莎 · A.海蒂(Mellissa A.Hardy) 著
贺光烨 译

SAGE Publications ,Inc.

格致出版社　上海人民出版社

图书在版编目(CIP)数据

虚拟变量回归/(美)海蒂(Hardy, M. A.)著;贺光烨译. —上海:格致出版社:上海人民出版社,2012
(格致方法·定量研究系列)
ISBN 978-7-5432-2107-9

Ⅰ. ①虚… Ⅱ. ①海… ②贺… Ⅲ. ①变量-回归分析 Ⅳ. ①O174

中国版本图书馆 CIP 数据核字(2012)第 122211 号

责任编辑 顾 悦

格致方法·定量研究系列
虚拟变量回归
[美]梅丽莎·A. 海蒂 著
贺光烨 译

出 版 世纪出版集团 www.ewen.cc
格致出版社 www.hibooks.cn
上海人民出版社
(200001 上海福建中路193号24层)
编辑部热线 021-63914988
市场部热线 021-63914081
发 行 世纪出版集团发行中心
印 刷 浙江临安曙光印务有限公司
开 本 920×1168 毫米 1/32
印 张 4.5
字 数 88,000
版 次 2012 年 7 月第 1 版
印 次 2012 年 7 月第 1 次印刷
ISBN 978-7-5432-2107-9/C·62
定 价 15.00 元

出版说明

由香港科技大学社会科学部吴晓刚教授主编的“格致方法·定量研究系列”丛书，精选了世界著名的SAGE出版社定量社会科学研究丛书中的35种，翻译成中文，集结成八册，于2011年出版。这八册书分别是：《线性回归分析基础》、《高级回归分析》、《广义线性模型》、《纵贯数据分析》、《因果关系模型》、《社会科学中的数理基础及应用》、《数据分析方法五种》和《列表数据分析》。这套丛书自出版以来，受到广大读者特别是年轻一代社会科学工作者的欢迎，他们针对丛书的内容和翻译都提出了很多中肯的建议。我们对此表示衷心的感谢。

基于读者的热烈反馈，同时也为了向广大读者提供更多的方便和选择，我们将该丛书以单行本的形式再次出版发行。在此过程中，主编和译者对已出版的书做了必要的修订和校正，还新增加了两个品种。此外，曾东林、许多多、范新光、李忠路协助主编参加了校订。今后我们将继续与SAGE出版社合作，陆续推出新的品种。我们希望本丛书单行本的出版能为推动国内社会科学定量研究的教学和研究作出一点贡献。

总序

往事如烟，光阴如梭。转眼间，出国已然十年有余。1996年赴美留学，最初选择的主攻方向是比较历史社会学，研究的兴趣是中国的制度变迁问题。以我以前在国内所受的学术训练，基本是看不上定量研究的。一方面，我们倾向于研究大问题，不喜欢纠缠于细枝末节。国内一位老师的话给我的印象很深，大致是说：如果你看到一堵墙就要倒了，还用得着纠缠于那堵墙的倾斜角度究竟是几度吗？所以，很多研究都是大而化之，只要说得通即可。另一方面，国内（十年前）的统计教学，总的来说与社会研究中的实际问题是相脱节的。结果是，很多原先对定量研究感兴趣的学生在学完统计之后，依旧无从下手，逐渐失去了对定量研究的兴趣。

我所就读的美国加州大学洛杉矶分校社会学系，在定量研究方面有着系统的博士训练课程。不论研究兴趣是定量还是定性的，所有的研究生第一年的头两个学期必须修两门中级统计课，最后一个学期的系列课程则是简单介绍线性回归以外的其他统计方法，是选修课。希望进一步学习定量研

究方法的可以在第二年修读另外一个三学期的系列课程，其中头两门课叫“调查数据分析”，第三门叫“研究设计”。除此以外，还有如“定类数据分析”、“人口学方法与技术”、“事件史分析”、“多层线性模型”等专门课程供学生选修。该学校的统计系、心理系、教育系、经济系也有一批蜚声国际的学者，提供不同的、更加专业化的课程供学生选修。2001 年完成博士学业之后，我又受安德鲁·梅隆基金会资助，在世界定量社会科学研究的重镇密歇根大学从事两年的博士后研究，其间旁听谢宇教授为博士生讲授的统计课程，并参与该校社会研究院(Institute for Social Research)定量社会研究方法项目的一些讨论会，受益良多。

2003 年，我赴港工作，在香港科技大学社会科学部，教授研究生的两门核心定量方法课程。香港科技大学社会科学部自创建以来，非常重视社会科学研究方法论的训练。我开设的第一门课“社会科学里的统计学”(Statistics for Social Science)为所有研究型硕士生和博士生的必修课，而第二门课“社会科学中的定量分析”为博士生的必修课(事实上，大部分硕士生在修完第一门课后都会继续选修第二门课)。我在讲授这两门课的时候，根据社会科学研究生的数理基础比较薄弱的特点，尽量避免复杂的数学公式推导，而用具体的例子，结合语言和图形，帮助学生理解统计的基本概念和模型。课程的重点放在如何应用定量分析模型研究社会实际问题上，即社会研究者主要为定量统计方法的“消费者”而非“生产者”。作为“消费者”，学完这些课程后，我们一方面能够读懂、欣赏和评价别人在同行评议的刊物上发表的定量研究的文章；另一方面，也能在自己的研究中运用这些成熟的

方法论技术。

上述两门课的内容，尽管在线性回归模型的内容上有少量重复，但各有侧重。“社会科学里的统计学”(Statistics for Social Science)从介绍最基本的社会研究方法论和统计学原理开始，到多元线性回归模型结束，内容涵盖了描述性统计的基本方法、统计推论的原理、假设检验、列联表分析、方差和协方差分析、简单线性回归模型、多元线性回归模型，以及线性回归模型的假设和模型诊断。“社会科学中的定量分析”则介绍在经典线性回归模型的假设不成立的情况下的一些模型和方法，将重点放在因变量为定类数据的分析模型上，包括两分类的 logistic 回归模型、多分类 logistic 回归模型、定序 logistic 回归模型、条件 logistic 回归模型、多维列联表的对数线性和对数乘积模型、有关删节数据的模型、纵贯数据的分析模型，包括追踪研究和事件史的分析方法。这些模型在社会科学研究中有着更加广泛的应用。

修读过这些课程的香港科技大学的研究生，一直鼓励和支持我将两门课的讲稿结集出版，并帮助我将原来的英文课程讲稿译成了中文。但是，由于种种原因，这两本书拖了四年多还没有完成。世界著名的出版社 SAGE 的“定量社会科学研究”丛书闻名遐迩，每本书都写得通俗易懂。中山大学马骏教授向格致出版社何元龙社长推荐了这套书，当格致出版社向我提出从这套丛书中精选一批翻译，以飨中文读者时，我非常支持这个想法，因为这从某种程度上弥补了我的教科书未能出版的遗憾。

翻译是一件吃力不讨好的事。不但要有对中英文两种

语言的精准把握能力，还要有对实质内容有较深的理解能力，而这套丛书涵盖的又恰恰是社会科学中技术性非常强的内容，只有语言能力是远远不能胜任的。在短短的一年时间里，我们组织了来自中国内地及港台地区的二十几位研究生参与了这项工程，他们目前大部分是香港科技大学的硕士和博士研究生，受过严格的社会科学统计方法的训练，也有来自美国等地对定量研究感兴趣的博士研究生。他们是：

香港科技大学社会科学部博士研究生蒋勤、李骏、盛智明、叶华、张卓妮、郑冰岛，硕士研究生贺光烨、李兰、林毓玲、肖东亮、辛济云、於嘉、余珊珊，应用社会经济研究中心研究员李俊秀；香港大学教育学院博士研究生洪岩璧；北京大学社会学系博士研究生李丁、赵亮员；中国人民大学人口学系讲师巫锡炜；中国台湾“中央”研究院社会学所助理研究员林宗弘；南京师范大学心理学系副教授陈陈；美国北卡罗来纳大学教堂山分校社会学系博士候选人姜念涛；美国加州大学洛杉矶分校社会学系博士研究生宋曦。

关于每一位译者的学术背景，书中相关部分都有简单的介绍。尽管每本书因本身内容和译者的行文风格有所差异，校对也未免挂一漏万，术语的标准译法方面还有很大的改进空间，但所有的参与者都做了最大的努力，在繁忙的学习和研究之余，在不到一年的时间内，完成了三十五本书、超过百万字的翻译任务。李骏、叶华、张卓妮、贺光烨、宋曦、於嘉、郑冰岛和林宗弘除了承担自己的翻译任务之外，还在初稿校对方面付出了大量的劳动。香港科技大学霍英东南沙研究院的工作人员曾东林，协助我通读了全稿，在此

我也致以诚挚的谢意。有些作者，如香港科技大学黄善国教授、美国约翰·霍普金斯大学郝令昕教授，也参与了审校工作。

我们希望本丛书的出版，能为建设国内社会科学定量研究的扎实学风作出一点贡献。

吴晓刚

于香港九龙清水湾

序

第一次听到“虚拟变量”这个词的时候，许多定量研究方法的学生都会觉得有趣，但很快他们就会意识到，这个听上去“虚拟”的方法，在定量研究中却起着至关重要的作用。我们知道，在回归分析中，用定序或者名义变量作为自变量来进行回归分析，既不能有效地反映因变量与自变量之间的实际关系，而且又容易出现拟合不足的情况。然而，引入了“虚拟变量”的概念后，我们就可以在不违反测量相关假设的情况下，运用最小二乘法进行回归分析。

那到底什么是“虚拟变量”呢？简单地说，虚拟变量是由原先的定性变量构建出来的二分变量。对于二分法，通常需要 $G-1$ 个数字来涵盖所有信息，其中 G 为原先类别的个数。例如，在民意调查中，如果我们希望表达公民的政治兴趣（其中包括 3 个类别——非常同意、有点同意、不同意），研究者必须构建两个二分变量。假设它们分别为 X_1（编码 1 表示非常同意，0 表示除非常同意外的类别）和 X_2（编码 1 表示有点同意，0 表示除有点同意外的类别），如果 X_1、X_2 两个变量的编码都为 0，那么暗示了受访者所属类别为不同意。在这里，

“不同意”这个类别被设置成了底线，或者说是一个参照组，从而 X_1 和 X_2 的回归系数都是在其他组与该组比较后估计得到的。

但是为什么选择“不同意”作为参照组而不选其他类别，如“有点同意”呢？曾经使用过虚拟变量的研究者基本都遇到过这样的问题。在这里，Hardy 教授给出了明确的答案。在本书中，一个有关收入的、精心设计的例子贯穿全文，从一个简单模型（含有一个虚拟变量的回归模型，我们常常将其简化到均值差异的检验）到一系列复杂模型（含有多个虚拟变量、多个定量变量及多个交互项的回归模型）。所幸的是，通过严谨的语言叙述，这种复杂性可以用不同条件下所得的回归系数来表达。

对虚拟变量回归有了基本了解后，Hardy 教授还提出了有关虚拟变量回归的一些特殊问题。除此以外，她还对如何处理异方差性，在因变量取对数或者 logit 后，如何对回归系数进行诠释，如何在显著性检验下进行多重比较，如何进行效果编码和对比编码以及如何检验曲线性和如何进行分段线性回归作出了解释。

总之，本书以通俗易懂的语言，从不同角度对虚拟变量的用法进行了详述。在有关统计方法的书籍中，没有任何一个作者可以如此全面地诠释一个问题。可以说，这本书无疑是一部有关虚拟变量回归的重要著作。

迈克尔·S. 刘易斯-贝克

目 录

第1章

简　介

回归分析是定量分析中运用最灵活、最广泛的一种方法。一个典型的回归模型试图将因变量 Y_i 映射到一系列特定的自变量 X_i 上，并通过相应的线性函数来解释因变量 Y_i 的变异。利用最小二乘估计，我们可以得到一个预测方程，用来估计自变量的条件均值，即特定自变量组合下的 Y 的期望值，从而得到因变量的条件均值。当自变量像定量变量那样可测量时，我们可以假设其为一系列任意的相对零点且间隔大致相等的定量变量，此时，所有可能的 Y 的期望值都是无限的。此外，当因变量和自变量都是定量变量时，其相应的关系可用几何图形表示。

在二元回归中，我们预测 Y 为唯一自变量的函数，则两个变量之间的关系可由回归线直接表示。线上所有的点代表 Y 的条件均值。当有第二个自变量包含到函数中时，一维回归线扩展成二维，一个由南北方向和东西方向的线组成的平面生成了，此时代表 Y 的条件均值的是所有处于该平面上的点。由此可见，当自变量的数量增加时，这些原则是保持不变的，尽管其几何形态可能变得难以描述。

但如果所有用来预测的自变量都用间隔尺度来衡量，那么回归模型的有效性将会受到严重制约。我们研究的问题

经常涉及组差异，如社会学家感兴趣的对民族/种族差异、性别差异，或行为、态度及社会经济特征的区域差异的解释。又如，市场调研人员希望从人口统计数据中了解消费者偏好。研究人员常常想知道对于所有组别，自变量的影响是否一样，或者在同一关系的强度或方向上，组差异是否依然存在。由此可知，我们大多数的研究问题是为了区别各级因变量下的组差异以及不同自变量影响下的因变量的组差异。

当感兴趣的自变量为定性变量时（即“只在名义水平上测量”），我们需要一种方法，它既能定量地代表这种信息，又能防止将不切实际的测量假设强加于分类变量。例如，我们可以将职业分类按 1 到 12 进行编码（该分类用于人口普查中的单数代码），但我们不可以简单地说，职业的范围是从低值 1 到高值 12，因为这种描述是建立在假定的间隔相等的基本衡量标准上的。定义一系列虚拟变量可以使我们捕捉到分类方案里的分级信息，然后把此信息用到标准回归估计中。事实上，回归方程中的自变量可以是任意定性和定量预测因子的组合。

例如，“社会资源是通过收入进行分配的”，这个现象既是那些对不平等感兴趣的学者所关注的焦点，也是那些努力为维持生活水平而奋斗的人民群众所关心的问题。我们关于社会公正的信念往往建立在对资源分布的认识上，以及是否有某些特定团体在分配过程中处于优势或劣势。我们知道，对于研究劳动收入分配中的歧视，有一种常见的方法，即首先确定一个组差异，比如男人和女人的差异或者黑人和白人的区别，以这个组差异作为在劣势群体的总效应，然后探讨加入其他决定性因素后，这个总差异如何变化，它是不是仍然维持不变？通过此方法，那些形成于社会进程中的、可

察觉的不平等从而可被识别。

为了之后讨论统计方法时的连贯性，我会引用一个例子，即预测收入是个体特征的函数，并用定性或定量变量描述相应的个体特征。我所用的数据来自全国老年男性纵向调查。通过第一次入户结果[1]可知，在最初的样本中，我们的研究对象大约为美国1500万45岁至59岁且未收容到专门机构（如监狱、精神病院）的男性。在该例中，我们比较感兴趣的变量包括种族、职业（美国人口普查分类）、教育（受教育年限）和工作任期（在同一个雇主下的工作年限）。尽管其他变量，例如劳动力的供给、工作技能、健康等也可以被假设为（通过薪酬得到的）年收入的预测因子，但是对于此例，我们不予考虑，而用只含有四个预测因子的函数提供一个定性定量相结合的估测。通过讨论逐步复杂化的模型来阐述虚拟变量回归的方法，我会尽量解释清楚有关任意特定的虚拟变量的系数是如何随模型整体而变的问题。同时，我还希望通过这些努力，减少读者在不适用的情况下，对此方法进行演绎的可能性。

本书以讨论我们最初关注的问题——黑人和白人之间的收入差异（用“美元/年”衡量）开始。之后，我们会不断加入新的假设并逐步建立复杂的模型进行检验。我们所要估计的是，当控制了更多的自变量（包括定性的或定量的）后，黑人和白人之间的平均收入差异是否仍然存在。还有，各个自变量的净效应在黑人和白人中是否一样。最后，我们将使用虚拟变量回归的形式来估计种族对回归模型所有参数的具体影响。有关这个逐步深入的过程，我们将在第4章具体描述。尽管未必所有读者都对收入分配这个话题感兴趣，但是由于其中所涉及的方法比较直接简单，所以适合各个学科