社会学教材教参方法系列

SM

回归分析

（修订版）

Regression Analysis (Revised Edition)

谢　宇／著

- 回归分析无疑是社会科学领域中最基础同时也是最经典的定量分析方法
- 最基础，是因为新近发展出来的统计方法基本上都建立在回归分析之上
- 最经典，是因为回归分析尤其是多元回归分析集中体现了社会科学定量分析方法的基本出发点：通过统计控制来实现或部分实现组间的可比较性

社会科学文献出版社
SOCIAL SCIENCES ACADEMIC PRESS (CHINA)

社

SM

回归分析

（修订版）

Regression Analysis (Revised Edition)

谢　宇／著

社会科学文献出版社
SOCIAL SCIENCES ACADEMIC PRESS (CHINA)

图书在版编目(CIP)数据

回归分析/谢宇著. —修订本. —北京：社会科学文献出版社，2013.3
（社会学教材教参方法系列）
ISBN 978-7-5097-4289-1

Ⅰ.①回… Ⅱ.①谢… Ⅲ.①回归分析 Ⅳ.①O212.1

中国版本图书馆CIP数据核字（2013）第029670号

·社会学教材教参方法系列·
回归分析（修订版）

著　　者／谢　宇

出 版 人／谢寿光
出 版 者／社会科学文献出版社
地　　址／北京市西城区北三环中路甲29号院3号楼华龙大厦
邮政编码／100029

责任部门／社会政法分社（010）59367156　　责任编辑／杨桂凤
电子信箱／shekebu@ssap.cn　　责任校对／姜夕芬
项目统筹／童根兴　　责任印制／岳　阳
经　　销／社会科学文献出版社市场营销中心（010）59367081　59367089
读者服务／读者服务中心（010）59367028

印　　装／北京季蜂印刷有限公司
开　　本／787mm×1092mm　1/16　　印　　张／25
版　　次／2013年3月第2版　　字　　数／445千字
　　2013年3月第2次印刷
　　097-4289-1

序　言

“社会学不像物理学。唯独物理学才像物理学，因为一切近似于物理学家对世界的理解都将最终成为物理学的一部分。”

——奥迪斯·邓肯

我一直认为，社会科学与物理学存在本质上的差别。社会科学的分析单位是异质性的或彼此区别的，而物理学的分析单位则被假定为同质性的或可相互替换的。我将社会科学这一重要而普遍的属性称作“变异性原理”（Variability Principle）[①]。

由于变异性原理的存在，社会科学要发掘出“放之四海而皆准”的规律注定是困难的，甚至是不可能的，尤其在个体层次上更是如此。正因为这个原因，社会科学似乎是一门软性的、不严谨的科学。这也是许多学者一直对社会科学中的定量方法提出质疑而偏好定性方法的主要原因。

然而，那些主张定性方法的学者并没有意识到，使定量方法遭到质疑的特性——变异性——也同样使定性研究遭到质疑，甚至问题更为严重。例如，因为每一个分析单位都不同于另一个分析单位，建立在单一个案基础上的定性研究得出的结论很可能会因案例的选择而发生根本性的改变。

我曾说过，“尽管带有自身的缺陷、局限和不完善，定量方法依然是理解社

① 谢宇，2006，《社会学方法与定量研究》，北京：社会科学文献出版社，第15～16页。

会及其变迁的最佳途径。在黑格尔哲学的意义上，那些使定量社会学不可靠、成问题的特征恰恰同时使它成为研究社会现象的不可缺少的工具，即……变异性原则。变异是人类社会的本质。没有一种定量的方法，我们就无法表述这种变异性。其他可供选择的方法，比如思辨、内省、个人体验、观察和直觉，确实也能增进我们的理解。不过，我大胆地提出，它们能够起到补充作用，但不应取代定量方法成为当代社会学的核心”①。

本书所介绍的统计方法常用于描述社会现象的属性、规律性以及变异性，这些方法可被纳入回归分析这一广义范畴中。毋庸讳言，这些方法都有缺陷，因为它们都难以精确地反映复杂的社会现实，但这并不妨碍它们成为社会科学研究的有用工具。有的学生或许会有这样的错觉，即社会科学研究中存在某种完美的方法，或者某些方法本质上优于另一些方法。事实并非如此。没有一种完美的方案可以解决社会科学中所有方法论上的难题，也没有哪种方法能在一切情境中都必然地优于另一些方法。最好的方法就是最适用于既定研究情境的方法。

所有社会科学中的统计方法都存在这样或那样的缺陷。因此，对我们而言，重要的是能够在将这些方法有效地运用到研究情境之前就知道它们的局限以及为什么会有这些局限。在本书中，我们特别关注了社会科学应用中各种统计方法的局限性以及在适用条件下改进这些方法的途径。权衡取舍在实践中普遍可见，因此，我希望学生们能够以灵活的思维来学习这些统计方法。通常，方法论上更大的解释力来自更多的信息——或是更丰富的数据，或是更强的理论基础。1996年，我在《美国社会学杂志》上评论 Charles Manski 发表于 1995 年讨论社会科学中识别问题的著作时，曾指出，“当观测数据不足时，我们只有通过强假定来获得清晰的结果。统计学中没有免费的信息。要么你收集它，要么你假定它”。

本书是根据我于 2007 年夏季在北京大学—密歇根大学学院举办的“调查方法与定量分析实验室项目”中教授回归分析课程时的讲义编写而成。我知道，目前中国国内有关回归分析的教材、专著和译著不胜枚举，这些著作都为中国学生与研究者了解和学习回归方法提供了有益的帮助。我认为，在社会科学领域，一本好的定量研究教材，既要涵盖量化研究与统计方法的重要理论，又要将方法原理与示范案例紧密相联，与此同时，对中文教材而言，最好还能结合中国的实际调查数据，以帮助读者对这些方法有更全面、更深入的了解。这本书是以

① 谢宇，2006，《社会学方法与定量研究》，北京：社会科学文献出版社，第 7～8 页。

CHIP88 数据作为主要的示例数据，之所以选用该数据，一方面是因为我在 1996 年与韩怡梅合作的文章中使用过这一数据，对其有较为详细的了解；另一方面是因为 CHIP88 数据也是许多其他学者做中国研究时常用的数据来源，因为该数据的全部原始个案和相关技术文档均可公开获得。我希望，借助对 CHIP88 原始数据所做的实例分析，读者既能将回归方法的基本原理和应用场合牢记于心，同时也能结合中国的实际研究数据来从事规范的社会科学定量研究。

这本书是许多人共同努力的成果。王广州教授在协调初稿写作阶段起了重要作用，我课堂上的六位学生——宋曦、刘慧国、王存同、李兰、傅强、巫锡炜，根据讲义编写了本书初稿中的部分章节。作为本身就有很强学术取向的学生和学者，这七人均是本书的合作者。我也从於嘉、赖庆、穆峥、周翔、黄国英、陶涛、任强、张春泥、程思薇在本书初稿读校的参与中获益良多。后记中细述了他们对本书所做的贡献。我对这些参与者的出色工作，还有历时三年的编写过程中他们同我的友谊以及对我的支持表示深深的感谢。对本书可能仍然存在的纰漏，我将独立承担责任。

本书的出版也得益于社会科学文献出版社的支持与鼓励。我在此感谢该社的谢寿光社长和杨桂凤编辑。正是他们致力于为中国社会科学界出版学术书籍的决心与付出鼓舞着我完成此书。

在此，还要感谢北京大学长江学者特聘讲座教授基金和密歇根大学 Fogarty 基金的资助。

最后，我还要感谢在我学术生涯中历经的无数老师与学生。他们让我知道，我对回归分析的理解仍旧有限。如果要论及此书的价值的话，它反映的是那些曾与我合作或共过事的人的集体智慧。我深知，与他们的合作和共事是我的幸运。

谢 宇

于安娜堡，2010 年 5 月 20 日

目录
CONTENTS

第1章

基本统计概念

1.1 统计思想对于社会科学研究的重要性

社会科学和自然科学存在本质的区别：自然科学以“发现”永恒的、抽象的、普遍的真理为最终目的，这是其精华所在；而社会科学则以“理解”暂时的、具体的、特定的社会现实为最终目的。历史上很多人曾希望在社会科学领域找到一种能够适用于各个方面的真理，并且为之做过许多尝试，但都没有成功。其实，定量研究方法并不可能使我们找到像自然科学那样的普遍真理。在社会科学研究中，我们的目的是理解现实社会（谢宇，2006）。

自然科学中真理的存在实质上反映了自然界中不同个体之间的同质性，即具体个体之间没有本质的差异。这一信念使自然科学家们认为，具体的、个体间的、看得见的差异只不过是表面的、人为的和微不足道的。然而，经验常识和从古到今的尝试表明，对于社会现象而言，异质性才是其突出的特性。由于具体个体间存在本质的差异，从而导致人们在社会科学研究中不能将所有个体等同对待。因此，社会科学中并不存在普遍真理，只存在一些原则和规律。对这些原则和相关逻辑进行探讨就是社会科学理论的任务。同时，受制于道德伦理和实际可行性，社会科学研究者基本上无法像自然科学家那样通过对实验室中的各种相关变量进行控制，从而寻找到社会现象的规律。因此，社会科学往往要依靠社会调查，通过样本来推断总体中的规律。这时，借用统计方法来完成研究工作便成为

一种必要的手段。

社会现象的异质性是研究者在社会科学研究中面临的最大难题，它使社会科学的任何研究方法都具有局限性，统计方法也不例外。正因为如此，社会科学的任何结论，凡是利用统计方法得到的，都必然包含一定的假设条件。可以说，学习定量研究方法[①]的一个关键就是了解定量研究方法本身的缺陷、局限和不完善。而这些都根源于社会现象的异质性。

尽管定量研究得到的结论都建立在一定的假设条件上，也不一定具有普遍意义，但定量研究方法却是研究社会现象不可缺少的工具。这是因为，如果没有这种方法，我们就无法很好地捕捉和表述研究对象的变异性。其他可供选择的方法（比如思辨、内省、个人体验、观察和直觉等）确实也能增进我们对社会现象的理解，但这些方法都不能很好地反映社会现象的异质性。当然，它们能够起到一定的补充作用，但不应取代定量研究方法成为当代社会科学的核心。换言之，定量研究方法依然是理解社会及其变迁的最佳途径，它可以使我们避免一些因意识形态或先入之见而导致的偏见，确保研究活动的“价值中立”，从而得到更为客观和全面的认识。比如，它可以让我们知道从某一研究得出的结论在总体层面上是否有偏差或在多大范围内是有效的；它也使我们可以通过统计方法发现组间差异和组内个体差异。而关于组间差异和组内差异的统计信息就是我们想得到的有关社会现象的规律。

定量研究方法已成为现代西方社会科学研究的主要手段，但其在中国的发展仍处于初期阶段，在各种研究中的应用还很少见，这导致中国社会科学与国外主流社会科学之间的脱节和交流的匮乏。当前，中国正处在一个迅速变化的社会背景下，各种社会问题和矛盾不断涌现，这为社会科学研究提供了极好的契机。对研究者而言，学习并使用定量研究方法来研究、解决问题将是非常有价值的。

定量研究方法的核心内容之一是统计学。而统计学本身就是一门专业学科，具有自己的学科体系、逻辑推理和符号语言。对从事社会科学研究的人来说，我们需要掌握这一学科体系、逻辑推理和符号语言。但我们同时也应该知道统计学的知识只是社会科学的工具，它本身并不能取代对所研究社会现象的了解和社会科学研究所必需的研究设计。本书仅讨论社会科学研究中常见的与回归分析有关的统计学问题，而不讨论社会科学理论和社会科学的研究设计方法。所以，本书

① 本书会交替使用“统计方法”和“定量研究方法”两个术语，我们将其等同对待。

所讨论的主要内容与具体研究问题和理论取向无关。我们希望那些对定量研究持负面态度和批评意见的学者也能学习统计知识，因为只有在真正理解了统计学思想之后，一个人才能对定量研究方法进行评价。

1.2　本书的特点

本书主要针对已经修读过基础社会统计学课程或者具有一定统计学基础知识的学生或研究者，希望读者通过学习本书能够对社会科学中回归模型的理论和实际操作有更全面、更深入的了解。除了讲解统计理论外，本书还将结合具体问题，利用统计软件，指导读者如何利用这些方法解决实际研究问题。本书具有两大特点：第一，除了对经典的多元回归模型进行比较深入的讲解外，对一些重要的、非经典的回归模型也进行了扩展和补充；第二，不是仅仅停留在理论层面，同时更强调实际操作的重要性。在大部分章节中我们都会使用实际研究数据，通过实例分析和相应的 Stata 程序来讲解统计知识在研究中的应用以及对数据研究结果给出阐释。在数据使用上，我们选用了 1988 年和 1995 年两次中国居民收入调查（CHIP）数据，1990 年美国综合社会调查（GSS）数据，1998 年、2000 年、2002 年和 2005 年“中国老年人健康长寿影响因素调查”（CLHLS）项目数据，以及 1972 年美国高中毕业生有关职业选择问题的调查数据。其中，使用最多的是 1988 年中国居民收入调查（以下简称 CHIP88）数据中城市居民的部分。

CHIP88 数据来自 1988 年由中国社会科学院经济研究所主持的“中国居民收入分配”调查。它是中国改革早期较具规范性的社会调查数据，因此在中、英文文献中被广泛采用。CHIP88 包括两个部分：一个是针对城市居民的调查，另一个是针对农村居民的调查。此次调查采用分阶段抽样的方法：先从 30 个省级行政单位中抽选出 10 个省份，然后再从这 10 个省份的 434 个城市中抽选出 55 个城市作为代表。城市部分的调查在 1988 年 3 ~ 4 月进行，共调查 9009 户，调查问卷收集了每一户中所有家庭成员的资料，包括其基本情况、受教育情况和就业情况。在删除缺失数据和不完整观测个案之后，总共得到 15862 条居民个体的观测数据。

在本书中，我们统一使用 Stata 9.0 作为示例数据的统计分析软件。由于算法和默认设定上可能存在的差异，采用不同软件和同一软件的不同版本对复杂模型进行参数估计所得的结果可能会存在细微差异。

1.3 基本统计概念

本书假定读者已经对社会统计学有一定程度的了解，下面将简要回顾社会统计学中的一些基本概念以及它们的性质，对这些内容的理解将有助于我们更好地学习回归理论。

1.3.1 总体与样本

在社会科学定量研究中，我们首先需要建立区分总体（population）和样本（sample）的敏锐意识。本章开篇提到，异质性问题是在个体间普遍存在的，但如果不同的个体在分类上确实满足某种定义，那么我们就将它们组成的总和称为总体。需要注意的是，总体是一个封闭的系统，它具有时间上和空间上的清晰界限。例如，2005 年的所有中国人在定义上就是一个界定完好的总体。2005 年所有年龄在 20～35 周岁拥有北京户口的已婚妇女也是一个界定完好的总体。后一个例子可以看作是前一个例子对应总体的子总体。

样本是总体的一个子集。比如，我们关心 2005 年中国居民的受教育程度和收入之间的关系，那么这项研究的总体就应该是 2005 年的所有中国居民。但在实际研究过程中，由于研究技术和经费的限制，我们不可能对所有中国居民进行分析，这时我们就需要从总体中按一定方式抽取一部分个体（比如一万人）进行调查，那么这一万人就构成了该总体的一个样本。当然，从理论上讲，我们从同一总体中可以抽取出若干个不同的样本。

由于个体异质性的存在，来自总体的某一个体并不能代表总体中的另一个体，而个体之间也是不能相互比较的。因此，我们不能利用样本对总体中的个体进行任何推断。但是，概括性的总体特征是相对稳定的。总体的这种特征就被称为参数（parameter）。总体参数可以通过总体中的一个样本来进行估计。通过样本计算得到的样本特征叫做样本统计量（sample statistic）。[①] 当然，样本提供的信息是有限的。那么，接下来的问题就在于如何依据样本信息来认识所研究的总体。统计推断（statistical inference）在这里扮演着关键角色。所谓统计推断，就是通过样本统计量来推断未知的总体参数。统计学的主要任务就是关注这种被称

① 这里，我们应该建立另一种敏锐意识：参数与总体相联系，统计量与样本相联系。

作“统计推断”的工作。尽管可以通过不同的样本统计量对总体参数进行估计，但是为了方便起见，在本章中，我们主要讨论把原来适用于总体数据的计算式运用到样本数据，所得到的样本统计量被称为“样本模拟估计式”（sample analog estimator）。根据稍后将要讲到的大数定理，随着样本量的增加，样本逐渐趋于总体，而样本统计量（样本模拟估计式）和总体参数之间的差别也会逐渐消失。

1.3.2 随机变量

随机变量（random variable）是指由随机实验结果来决定其取值的变量。它具有两个关键属性：随机性和变异性。随机性也就是“不确定性”。在社会科学研究中，这种“不确定性”主要来自两个方面：一方面是由受访者个体行为或态度本身的不确定性造成的；另一方面来自群体中个体间的异质性，因随机取样而产生。

在实际研究中，作为随机变量的因变量的测量类型决定了研究者应该选择何种统计分析方法。[①] 丹尼尔·A. 鲍威斯和谢宇（Powers & Xie, 2008）在《分类数据分析的统计方法》一书中曾经根据三种标准将因变量划分为四种测量类型，如图1-1所示。

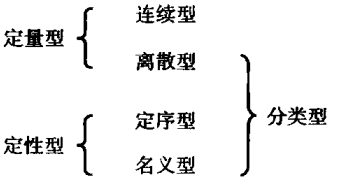

图1-1 随机变量的测量类型

首先，就定量和定性这一划分而言，在定量变量（quantitative variable）中，变量的数字取值具有实质性的意义；然而在定性变量（qualitative variable）中，变量的数字取值本身并没有什么实质意义，只是为了表明类别间的互斥性。例如，在贫困问题研究中，将贫困状况编码为“1=贫困”和“0=非贫困”，这里的数值1和0仅仅是划分研究对象是否处于贫困状态的标识而已，并没有表达贫困程度的含义。换句话说，定性变量的数字取值只是不同类别的代号。因此，定性变量都属于分类变量（categorical variable）。

① 更多的有关这部分的内容请参考Powers & Xie（2008）一书的前言。