格致方法·定量研究系列　吴晓刚　主编

回归诊断简介

[加] 约翰·福克斯（John Fox）著
於嘉译

- 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

5

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

回归诊断简介

[加] 约翰 · 福克斯(John Fox) 著
於嘉 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

回归诊断简介/(加)福克斯(Fox, J.)著;於嘉译. —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 978-7-5432-2117-8

Ⅰ. ①回… Ⅱ. ①福… ②於… Ⅲ. ①回归分析-研究 Ⅳ. ①O212.1

中国版本图书馆 CIP 数据核字(2012)第 127381 号

责任编辑　顾　悦

格致方法·定量研究系列
回归诊断简介
[加]约翰·福克斯 著
於嘉 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　4.75
字　数　91,000
版　次　2012 年 7 月第 1 版
印　次　2012 年 7 月第 1 次印刷
ISBN 978-7-5432-2117-8/C·69
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

此为试读，需要完整PDF请访问：www.ertongbook.com

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

在社会科学的数据分析中，回归可谓最常用的方法。通过计算机获得一个估计的回归方程就和数 1、2、3 一样简单，事实的确如此，因为利用任何一个软件程序，研究者都可以按如下步骤操作：(1)加载样本数据；(2)确定回归方程；(3)利用普通最小二乘法进行估计。这将获得一个类似下面这下等式的结果：

$$Y = 62 + 71.5X_1 + 5.4X_2 + e$$

但是，这个估计的结果如实反应了真实世界的状况吗？例如，在 X_2 保持不变的情况下，X_1 一个单位的变化是否将导致 Y 产生 71.5 的预期变化？我们往往可以非常自信地谈论总体估计的精确度。但是，我们对回归结果的信任程度取决于是否能够成功地处理以下常见问题：多元共线性、奇异值、非正态、异方差性以及非线性。

Fox 教授将“诊断”引申为发现上述问题。例如奇异观测值或更概括地讲，即强影响观测值产生的问题。除了那些可以展示某一极端值如何影响回归直线的常用图形外，他对

其他测量方法也进行了阐释：预测值、学生残差、Cook 距离以及偏回归散点图。这些测量方法大多可以通过常用的软件程序获得，例如 SAS 或 SPSS。

在对回归进行了诊断之后，Fox 专注寻找可能的解决办法。此类问题非常多，例如，如果具有高度的共线性，这个变量需要被剔除出回归方程吗？如果有奇异值出现，这个观测是否应该被舍弃？当误差的分布是偏斜的时候，是否应该对其进行一些变换？在异方差性存在的情况下，是否应该使用加权最小二乘法以解决这一问题？当非线性问题存在时，是否应该使用次方转换？在面对这些重要的问题时，应尽量避免使用机械的权宜方法。正如作者不断强调的，这些方法永远不能取代判别和思想。

为了使解释更加丰富，Fox 利用了许多数据作为例子：美国的人口普查、职业声望、人们报告的体重、加拿大公司中的董事会。这些例子使得本书中的诊断适用于广大的回归方法使用者。此外，有意愿受更高级训练的读者可以在附录中寻找答案（例如，对用于解决高度共线性的岭回归的评估）。每一个使用回归分析的人，理应进行一系列回归诊断。

迈克尔·S. 刘易斯-贝克

此为试读，需要完整PDF请访问：www.ertongbook.com

Regression Diagnostics
Copyright © 1991 by SAGE Publications, Inc.
All rights reserved. No part of this book may be reproduced or utilized in any form or by any means, electronic or mechanical, including photocopying, recording, or by any information storage and retrieval system, without permission in writing from the publisher.
This simplified Chinese edition for the People's Republic of China is published by arrangement with SAGE Publications, Inc. © SAGE Publications, Inc. & TRUTH & WISDOM PRESS 2012.

本书版权归 SAGE Publications 所有。由 SAGE Publications 授权翻译出版。
上海市版权局著作权合同登记号：图字 09-2009-547

目 录

第1章

概　论

在社会科学研究中，线性最小二乘回归分析可谓最常用的统计技术，并为许多其他的统计方法奠定了基础。但是，最小二乘回归往往面临许多困难，它对于数据结构有着较强且往往不切实际的假设。回归诊断是用于探索存在于回归分析中的问题及判断某些假设是否合理的一种技术。

回归诊断在当代的发展与计算机交互式的统计分析的实现是不可分割的，因此，回归诊断在很大程度上是近 20 年的产物。与回归诊断方法紧密相关的是用于纠正已发现问题的各种技术，其中许多方法都涉及对数据的转换。

作为一个初步的例子，我们首先考虑图 1.1 中来自 Anscombe(1973)的四幅散点图。统计分析的一个目的就在于为数据提供详尽的描述性归纳。Anscombe 的四个数据集已被设计得出相同的标准线性回归结果：斜率、截距、相关系数、回归标准误、系数标准误以及统计检验。但非常重要的是，它们不具有相同的残差。

在图 1.1(a)中，线性回归合理地描述了 y 随 x 的增长而增长这一趋势。在图 1.1(b)中，线性回归未能反映出数据具有的曲线形式，所以线性方程显然是错的。在图 1.1(c)中，某一点与其他点构成的直线偏离，这对拟合的回归直线产生

此为试读，需要完整PDF请访问：www.ertongbook.com