分位回归

汉译经济学文库

Translated Economics Library

QUANTILE REGRESSION

[美]罗杰·康克（Roger Koenker） 著
马令杰 译

上海财经大学出版社

汉译经济学文库

分位回归 Quantile Regression

［美］ 罗杰·康克 (Roger Koenker) 著

马令杰 译

上海财经大学出版社

图书在版编目(CIP)数据

分位回归/(美)康克(Koenker,R.)著;马令杰译. —上海:上海财经大学出版社,2013.8

(汉译经济学文库)

书名原文:Quantile Regression

ISBN 978-7-5642-1595-8/F.1595

Ⅰ.①分… Ⅱ.①康… ②马… Ⅲ.①回归分析-统计模型 Ⅳ.①O212.1

中国版本图书馆 CIP 数据核字(2013)第 094143 号

□责任编辑 俞晓峰

□封面设计 钱宇辰

□责任校对 赵 伟 廖沛昕

FENWEI HUIGUI

分 位 回 归

[美] 罗杰·康克 (Roger Koenker) 著

马令杰 译

上海财经大学出版社出版发行

(上海市武东路 321 号乙 邮编 200434)

网址:http://www.sufep.com

电子邮箱:webmaster@sufep.com

全国新华书店经销

上海市印刷七厂印刷

上海远大印务发展有限公司装订

2013 年 8 月第 1 版 2013 年 8 月第 1 次印刷

787mm×1092mm 1/16 16.25 印张(插页:2) 354 千字

印数:0 001—3 000 定价:48.00

图字:09-2013-71 号

Quantile Regression

Roger Koenker

译者序

这本书得以以中文的形式与读者见面，我感到十分高兴，希望能对“分位回归”在中国的研究和应用有所裨益。

本书的原作者罗杰·康克(Roger Koenker)是我在读博士研究生时的导师。从我做博士论文时(2002年)开始着手翻译，到现在陆陆续续已经十年了，自己深感愧疚。刚开始翻译时用的是罗杰·康克的最原始稿件，而英文原著在2005年由英国剑桥大学出版社出版时，Roger已对许多章节作了相当大的改动。于是，我不得不对这些章节重新进行翻译，其间由于工作、家庭及自己的惰性而耽搁了好几年。今年，终于挤出时间将此书译完，算是对Roger和自己的一个交代。

在翻译过程中，我力求译文能够信、达、雅，并以信为先。对书中所涉及的绝大多数人名、地名和援引的著作，我都保留了原著中的英文，这样便于读者查找和阅读相关的文献。

我已注意到，“分位回归”在中国已经有了一定程度的引入和应用，希望此书的译本能给读者提供一个系统和深入的介绍。尤其是书中既有严谨的理论和分析，又有通俗易懂的例子，更有充满Roger风格的寓庄于谐的评论，所以我真诚地希望这本书能激发读者对“分位回归”研究和应用的兴趣。

在这里，我要感谢罗杰·康克对这项工作的无私帮助和全力支持，没有他的支持，就没有本书中文版的面世。我也十分感谢顾佳颖博士，她在本书的词句修正和图表文字说明方面做了大量的、十分出色的工作。Larry Pohlman给了我很多鼓励，在此深表谢意。同时，也衷心感谢我的妻子庄建萍对我的支持和对本书中文译稿的加工处理。最后，我要将本书的中文译本献给我的父母以及Ellen、Abby和Rachel。

马令杰

2012年4月27日于芝加哥

原著序言

当面对别人对统计学的贬低时，Francis Galton 有一段精彩的话用来阐述"统计学的魅力"，同时责备他的一些统计学同行们：

> "对那些仅满足于求得平均值而忽视探求细节的人来讲，他们的观点看起来相当滑稽：如果把瑞士的高山推倒在瑞士的大湖里，那么瑞士的地形就与英格兰的大片平原一样平坦。"(《自然遗传》第62页)

统计学家的基本任务就是设法从表面上看起来杂乱无章的科学观测数据中得出规律性的东西。通常这项工作可以由求得在限制变量条件下的一些参数平均值来实现，而统计学中流行的最小二乘法非常适合于求平均值。但是，正如 Galton 所指出的那样，这种一味地求平均值而忽略"多样性"的观点是很值得怀疑的。

作为美国地形最平坦的县之一的居民，我对瑞士的地形及其吸引人的多样性的认识与 Galton 所描述的只满足于"平均数"的人们是大不相同的。不仅仅是瑞士的地形，更有许多瑞士的著名统计学家也使我们认识到观测数据的多样性，以及局限于盲目求平均值的严重后果。

分位回归为我们提供了一个全面认识统计"地形"与随机变量之间关系的有效方法。用最优化来代替一个样本的分类和排序这样的一个简单思路，使我们联想到可把这个主意扩展到更广的统计模型中去。对于大多数模型来说，就像用最小二乘法来估计条件平均参数一样，我们可以通过最小化残差绝对值(非对称形式)法来求条件分位参数的估计值。对线性参数模型而言，由参数的线性规划方式来表达最优化问题大大简化了计算程序，而线性规划的标准对偶性为排序统计和以排序为基础的推论提供了一个新的研究途径。

我希望能通过这本书就分位回归方法给出一个全面的介绍，也希望它能激励别人在各自的研究领域中进一步探索和发展这些思路和方法。众所周知，任何一种统计方法的成功与否最终都在于实践的检验，我在本书中将尽可能地给出分位回归法的应用实例。虽然那种统计方法可以由数学演算而更清楚和有效地表达出来的观点在本书中被相应地淡化了，但整个第 4 章仍被用来探讨有关分位回归的

基本的渐近理论，并且其他各章中也列出了相关的数学证明细节。

分位回归的统计软件在许多应用广泛的统计软件包中都可以找到，如 S^+、SAS和 Stata。其中，S^+ 的应用者会发现本书中的计算语言和图形都是通过 S 来实现的。（读者可查询 http://www.econ.uiuc.edu。）

在此，我十分感谢多年来为这项工作提供支持的诸多同仁。Gib Bassett 的博士论文《最小绝对值 L_1 回归》，为这项工作打下了基础，而且他仍在继续提供有价值的观点和热情的支持。Jana Jureckova 在早期就参与了分位回归的研究工作，并且在工作中与 Cornelius Gutenbrunner 探讨了分位回归与排序统计学的关系。最近，我与 Zhijie Xiao，Ivan Mizera 分别在时间序列和多参数平滑问题上进行了合作研究。还有 David Ruppert，Ray Carroll，Alan Welsh，Tertius Dewet，Jim Powell，Gary Chamberlain，Xuming He，Keith Knight，Probal Chaudhuri，Hira Koul，Marc Hallin，Brian Cade，Moshe Buchinsky，Berndt Fitzenberger，Victor Chernozhukov 和 Andrew Chesher 等，都为分位回归法的发展作出了贡献。近些年来，我与我的博士研究生们进行了合作研究，在分位回归领域颇有建树，其中的佼佼者包括 Jose Machado，Pin Ng，Quanshui Zhao，Yannis Bilias，Beum-Jo Park，M. N. Hasan，Daniel Morillo，Ted Juhl，Olga Geling，Gregory Kordas，Lingjie Ma，Ying Wei，Roberto Perrelli 和 Carlos Lamarche。

本书中的一些内容来源于我为一些讲座所备的讲义，比如 1997 年巴西统计年会、2001 年南非统计年会、2003 年春季的伦敦大学学院讲座及同年秋季的格罗宁根 NAKE 讲座。在此，我对这些会议的组织者和参加者，尤其对他们的建议表示衷心的感谢。

另外，我要特别感谢 Steve Portnoy。我俩从 20 世纪 70 年代中期便开始合作，原打算一起写这本书，但由于 Steve 有其他研究在身，我就只有靠自己来完成它了。

同时，我还要感谢许多研究所和大学为本书写作提供的良好环境，它们包括贝尔实验室、澳大利亚国立大学、查尔斯大学、伦敦大学学院，以及我的学术老家伊利诺伊大学。还有，我十分感谢（美国）国家自然科学基金多年来给予的支持。

最后，我将最诚挚的感谢献给我的妻子 Diane 和女儿 Hannah！

本书献给我已逝的女儿 Emma，以兹纪念。

罗杰·康克(Roger Koenker)

2004 年 7 月 2 日于厄班纳

目 录
CONTENTS

4 分位回归的渐近理论 / 89

1 引言

1.1 方法和结果

大多数应用统计学都可以看作对线性回归模型和与其相关的最小二乘法的论述。在介绍这些方法之前，先让我们引用一段很有影响的源于 Mosteller and Tukey(1977)的话：

> “传统回归曲线给出了关于集合 x 分布的一般平均值。我们可以走得更远些，即通过计算分布的几个百分位处的回归线而得出更完整的关于 x 的情况。但传统的回归分析还没有做到以上程度，因而给出的是一个不完整的有关 x 的状况。就如平均值只是给出了单一分布的不完整状况一样，传统回归曲线也只是给出了一组有关相应分布的不完整状况。”

我们的目标就是尽可能明白地描述怎样“走得更远些”，分位回归的目的就在于提供全面的回归信息。

为什么线性回归模型的最小二乘法在应用统计中如此流行？是什么原因使其如此成功？我想大概有以下三个原因：首先，我们不应该低估一个明显的事实，那就是线性估计在计算方法上的简便性，这是十分吸引人的。当然，这也是最小二乘法得以流行的最初动力。其次，如果观测到的误差是正态分布即 Gauss 分布的话，最小二乘法在很多方面具有最优性。但是，正如 Gauss 本人所指出的，这只是以一种过后的假设来代替实际的分布。再次，正如最近有人指出的那样，最小二乘法给出了一个估计条件平均参数的一般方法。

然而正如 Mosteller 和 Tukey 所指出的那样，即便是对单一样本的统计分析，平均值也是极少令人满意的。由此，有关分布的偏斜度、峰度、盒状图、矩形图和更复杂的密度估计便被人们用来获得对分布的更进一步的认识。对于回归来讲，我们是否可以进行类似的探索呢？一个自然而然的出发点就是，在应用最小二乘法获得条件平均值的同时，可以补充几个估计的条件分位值。在后续的章节里，我们将详细介绍这种方法。而这个方法的基本思想，则可以追溯到 Boscovich 在 18 世纪中叶和 Edgeworth 在 19 世纪末的工作。

1.2 第一个回归——历史的序曲

尽管最小二乘法如此流行，但具有讽刺意味的是，第一次回归尝试却是与分位回归紧密相连的。正如前面所述，我们现在的工作可以看作回到200年以前的统计学研究，那时还没有人提出最小二乘法。

如果最小二乘法可以追溯到1805年Legendre的工作的话，那么Boscovich对回归的最初工作则提前了半个多世纪。当时，令Boscovich感兴趣的是地球的椭圆性。牛顿和其他科学家提出，地球的转动可能使之成为两头平、中间凸出的形状——像一粒葡萄而非一枚柠檬。对于Boscovich的贡献和早期的回归历史，Stigler(1986)有着经典的论述。Smith(1987)详尽地描述了大地测量学的发展，并对表1.1列出的数据予以高度重视。

为估计地球的形状，Boscovich应用了5组测量数据（见表1.1）。这些测量值分别对应于地球上5个不同地点——从位于赤道的Quito到北纬66°19′的Lapland——的纬度1°的弧长。应该指出的是，取得这些测量值在当时的确是一项艰巨的任务。显然，弧长随着从赤道到极地的移动而增加，这恰巧符合牛顿的设想。但问题是，怎样把这5组测量值结合起来用以估计地球的椭圆率呢？答案在当时尚不清楚。

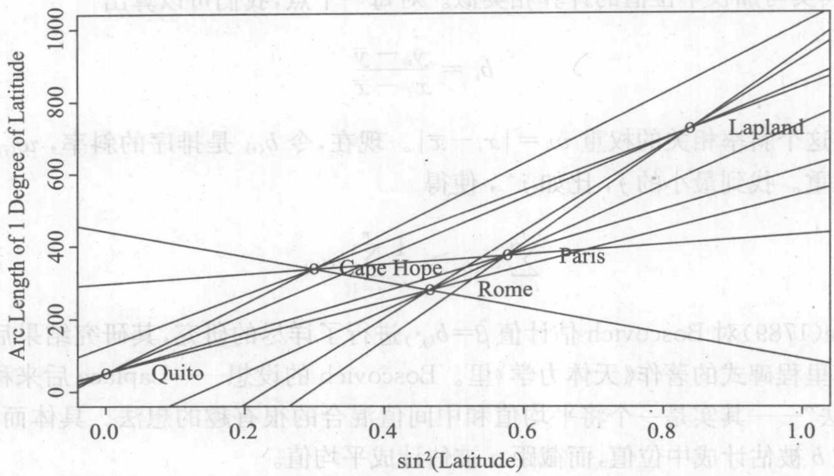
表1.1 **Boscovich椭圆率数据**

Location	Latitude	$\sin^2$(Latitude)	Arc Length
Quito	0°0′	0	56 751
Cape of Good Hope	33°18′	0.298 7	57 037
Rome	42°59′	0.464 8	56 979
Paris	49°23′	0.576 2	57 074
Lapland	66°19′	0.838 6	57 422

对于短弧而言，其弧长可以用下列方程来逼近：

$$y = a + b\sin^2\lambda \tag{1.1}$$

其中，y是弧长，λ是纬度，参数a可以看作赤道处的1°弧长，b可以看作极地与赤道各1°弧长之间的差值。椭圆率可以由下列公式来计算：$\eta=3a/b$，其中η为椭圆率的倒数。Boscovich注意到，任何一对观测值都可以用来计算a和b，从而得出η。于是，他采用了所有的10对观测值，这些回归线如图1.1所示。从图中我们可以看到，其中有几条线让人感到有些匪夷所思，比如经过罗马和好望角的向左下方倾斜的那条直线。Boscovich最终选用了两种估计方法：一种是对10个b的估计值取平均值，另一种是剔除两个最小的斜率估计值后取平均值。而a的估计值都是直接取自赤道处的弧长。两种估计方法给出了两个椭圆率：1/155和1/198。而现代的估计方法则是使用中位值即中间斜率来估计椭圆率，Theil(1950)所得的结果是1/255。



注：Boscovich 计算了全部 10 个斜率的估计值。最初，他报告了一个截尾均值(trimmed mean)作为对地球椭圆率的估计。基于对赤道和两极各 1°弧长之间差值的测算，可知赤道周长较两极圆周长多出 56 700 突阿斯(toise)。(此处 1 突阿斯≈1.95 米。)

图 1.1 Boscovich 椭圆率示例

有趣的是，(a,b) 的最小二乘法估计值可以表示为斜率估计值的加权平均值。令 h 代表 10 对数值，则

$$b(h)=X(h)^{-1}y(h) \tag{1.2}$$

其中，$h=(i,j)$，则

$$X(h)=\begin{bmatrix}1 & x_i\\ 1 & x_j\end{bmatrix},\qquad y(h)=\begin{bmatrix}y_i\\ y_j\end{bmatrix} \tag{1.3}$$

从而，我们可以将最小二乘法估计值写为

$$\hat{b}=\sum_h w(h)b(h) \tag{1.4}$$

其中，$w(h)=|X(h)|^2/\sum_h|X(h)|^2$。正如 Subrahmanyam(1972)和 Wu(1986)所指出的，这种最小二乘法估计值的典型形式可以推广到有 p 个参数的一般线性回归模型。在上述双变量的例子中，权重显然与多个测量点之间的距离成比例，这预示着最小二乘法对任何 x 或 y 的极端观测值的敏感性和脆弱性。

有趣的是，仅仅在短短两年之后，Boscovich 就对椭圆问题发起了第二波“攻击”，而这次他使用的工具离我们现代的分位回归方法更接近了。他建议在误差之和等于零的前提下，通过取误差绝对值之和的最小值来估计式(1.1)中的 (a,b)。这个误差之和为零的前提要求回归曲线通过观测值的中心点即 $(\bar{x},\bar{y})$。Boscovich 提供了一个十分简单的几何算法来求估计值。在满足约束条件后，我们可将问题简化为：设想一条通过点 $(\bar{x},\bar{y})$ 的直线，绕这个点转动，直到误差绝对值之和最小时停止转动，即得到我们所要的回归线。正如后来 Laplace 所指出的，在代数方法

上，这其实与加权中位值的计算相类似。对每一个点，我们可以算出

$$b_i = \frac{y_i - \bar{y}}{x_i - \bar{x}} \tag{1.5}$$

以及与这个斜率相关的权重 $w_i = |x_i - \bar{x}|$。现在，令 $b_{(i)}$ 是排序的斜率，$w_{(i)}$ 为相关的权重。找到最小的 j，比如 j^*，使得

$$\sum_{i=1}^{j} w_{(i)} > \frac{1}{2}\sum_{i=1}^{n} w_{(i)} \tag{1.6}$$

Laplace(1789)对 Boscovich 估计值 $\hat{\beta} = b_{(j^*)}$ 进行了详尽的研究，其研究结果后来发表在其里程碑式的著作《天体力学》里。Boscovich 的设想——Laplace 后来称之为"情况法"——其实是一个将平均值和中间值混合的很有趣的想法。具体而言，斜率参数 b 被估计成中位值，而截距 a 被估计成平均值。

这些想法在被人们淡忘了近一个世纪后，由 Edgeworth 于 1888 年使其重焕光彩。Edgeworth 清楚地认识到了这些想法的价值，即使在其早期的对指数和加权平均值的研究中，他就已开始强调：一般所认为的作为位置估计值的样本平均值，其最优性能完全是建立在观测值服从正态分布的假设之上。但如果观测值偏离正态分布的话，中位数则要优于平均值。Edgeworth 仿佛已经"预测"到了 20 世纪 40 年代 Tukey 的研究工作，他将不同比例的正态分布相混合，从而得到相应的观测值，并比较了它们的平均值和中位数的渐近方差。他得出结论：对于相对比例大于 2.25 的等权重混合样本而言，中位数比平均值具有更小的渐近方差。

Edgeworth 对线性回归中间值估计法的研究，直接把我们带到了分位回归的领域。Edgeworth(1888)放弃了 Boscovich-Laplace 的"误差之和等于零"的限制条件，他提出了双重中位值法，即对截距和斜率都用最小绝对值之和法，并进一步将之推广到多重中位值法。他对双变量模型给出了一种几何算法并且讨论了最小绝对值法优于最小二乘法的条件。但是，这种几何算法相当艰涩，连 Edgeworth 本人后来也承认这种方法需要"数学家的关注和对超几何概念的把握(在有许多未知数的情况下)"。只是在很多年以后，线性规划的出现才为分位回归提供了一个概念上简单而计算上又行之有效的办法。

一旦有了一个中位回归估计值，一个很自然的问题就是"有类似的其他分位的回归吗?"我们将在下一节中对这个问题展开讨论。

1.3 分位、排序和最优化

任意实数随机变量可以由其(右连续)分布函数来描述：

$$F(x) = P(X \leqslant x) \tag{1.7}$$

同时，对任意 $0 < \tau < 1$，

$$F^{-1}(\tau) = \inf\{x: F(x) \geqslant \tau\} \tag{1.8}$$