第1章

引 言

回归分析的目的在于揭示因变量和自变量的关系。在实际应用中，自变量并不能精确地估计因变量。相反，与每个自变量的特定值相对应的因变量是一个随机变量。因此，我们常常使用集中趋势的测量方法，来概括自变量特定值域下的因变量变化情况，主要包括均值、中位数和众数。

传统的回归分析主要关注均值，即采用因变量条件均值的函数来描述自变量每一特定数值下的因变量均值，从而揭示自变量与因变量的关系。模型化和拟合条件均值函数(conditional-mean function)是回归模型法大族谱中的核心思想，具体包括常见的简易线性回归模型、多元回归、加权最小平方数下的异方差误差模型和非线性回归模型。

条件均值模型具有以下优点：在理想的条件下，它们可以为我们提供关于自变量和因变量分布关系的完整的和参数的描述。另外，采用条件均值模型可获得具有优越统计特性的估计量(最小二乘法和最大似然法)，它更容易计算，并且更容易解释。这种模型通过不同的方式被推广，从而适用于误差具有异方差性的情况，因此，对于特定的自变量，因变量条件均值和条件单位的模型化可以同时进行。

条件均值模型被广泛应用于社会科学中，尤其在过去的

半个多世纪里，使用最小二乘法及其衍化方法对连续型因变量和自变量的关系进行回归建模被认为是现代重要的统计工具。最近，分析二分因变量的 logistic 和 probit 模型、分析计数因变量的泊松回归模型在社会科学研究中的重要性不断提高。这些方法并没有超出条件均值模型的框架。当社会科学定量研究者已经应用更高级的分析方法来放宽条件均值框架下的一些建模假设时，这个框架本身却很少被质疑。

条件均值框架存在先天的局限性。首先，当归纳自变量特定数值下的因变量情况时，这个条件均值模型并不能轻易地扩展到非中心位置，而非中心位置往往正是社会科学研究的兴趣所在。例如，关于经济不平等和流动的研究对穷人（低尾）和富人（上尾）的情况有浓厚的兴趣。教育研究者会设法在既定的成绩水平下去理解和减少群体差异（如三层次参照标准：基础、熟练和高级）。这样，对中心位置的强调，长期阻碍了学者采用恰当的技术来研究有关因变量非中心位置的课题。而采用条件均值模型来分析以上问题是没有效率的，甚至会偏离研究重点。

其次，这些模型的假设在现实生活中并不总会得到满足。特别是方差齐性假设经常被违反。另外仅仅关注集中趋势会忽视关于因变量分布的有用信息。并且，社会现象中通常会出现重尾分布，从而导致离群值优势。正因为条件均值深受离群值的干扰，所以它对中心位置的测量是不恰当和具有误导性的。

最后，一直以来对中心位置的关注转移了学者对因变量整体分布性质的注意力。我们需要跳出预测变量的位置和数值范围对因变量的效应这一框架，进而探讨预测变量的变

化会如何影响因变量分布的基本形状。例如，许多社会科学研究关注社会分层和不平等，这一领域要求深入分析因变量的分布特征。对分布特征的描绘包括中心位置、数值范围、偏态和其他高阶特性，而不仅仅是中心位置。因此，采用条件均值模型来表述因变量分布与自变量的关系是具有先天性缺陷的。关于不平等主题的例子包括工资、收入和财富等经济不平等；在学业成绩上的教育不平等；在身高、体重、疾病发生概率、毒品上瘾、医疗和预期寿命上的健康不平等和由于社会政策而导致的生活质量的不平等。这些课题通常采用条件均值框架进行分析，从而忽略了其他更重要的分布特征。

条件均值模型的替代方法可以追溯到 18 世纪中期。这一方法被称为条件中位数模型，或简称中位数回归。它解决了一些上面提出的关于集中趋势测量方法的选择问题。这种方法用最小绝对距离估计代替最小二乘估计。最小二乘估计不需要大功率的计算机便可轻松实现，然而最小绝对距离估计必须借助强大的计算机力量。所以，直到 20 世纪 70 年代后期，当计算机技术融合了如线性优化等算法系统时，采用最小绝对距离估计的中位数回归模型才变得实用。

中位数回归模型可以实现与条件均值回归模型同样的目标：表述因变量的中心位置与一组协变量之间的关系。然而，当因变量的分布是高度偏态时，均值在解释的时候就会受到质疑，而中位数依然保有大量信息。因此，条件中位数模型具有更大的应用潜力。

中位数是一个特殊的分位数，它表示一种分布的中心位置。中位数回归是分位数回归的一种特殊情况，在这里，第 0.5 分位数被模型化为一个关于协变量的函数。一般地说，

其他分位数则可以用来描述一种分布的非中心位置。分位数概念可归纳为一些特定的名称，如四分位数、五分位数、十分位数和百分位数。第 p 个百分位数表示因变量的数值低于这一百分位数的个案数占总体的 $p\%$。因此，分位数可以指定分布中的任何一个位置。例如，有 2.5%的个案数值低于第 0.025 分位数。

凯恩克(Koenker) 和巴西特(Bassett)在 1978 年引入分位数回归，将条件分位数模型化为预测变量的函数。分位数回归模型是线性回归模型的自然扩展。随着协变量的变化，线性回归模型描述了因变量条件均值的变化，而分位数回归模型则强调条件分位数的变化。由于所有分位数都是可用的，所以对任何预先决定的分布位置进行建模都将是可能的。因而，研究人员可以选择适合他们特定研究议题的分位数进行分析。贫穷研究关心低收入人群，例如，在 2000 年 11.3%的社会底层生活在贫穷状态中(U. S. Census Bureau, 2001)。税收政策研究则关注富人，例如，最富有的 4%的人口(Shapiro & Friedman, 2001)。条件分位数模型为集中研究人口中的特定人群提供了灵活性，而条件均值模型则做不到。

由于多元分位数可被模型化，所以我们可以更加全面地理解因变量的分布是如何受到预测变量的影响的，包括形状变化等信息。一组间距相同的条件分位数(如总体中的每 5%或每 1%)可以描绘除中心位置外的条件分布的形状。这种模型化形状变化的能力是社会不平等研究领域在方法论上的一次飞跃。按照惯例，以往的不平等研究并不是建立在模型基础上的，这些方法包括洛伦兹曲线(Lorenz curve)、基尼系数(the Gini coefficient)、泰尔熵标准(Theil's measure of

entropy)、方差系数和对数转换分布的标准差等。

通过建立在线性优化基础上的算法系统，最小化关于距离的广义测量方法便可以轻松地建立分位数回归模型。因此，分位数回归目前是研究者的实用工具。社会科学家所熟悉的软件包则提供了简单易懂的命令来拟合分位数回归模型。

在凯恩克和巴西特首次引入分位数回归的15年后，有关分位数回归的实际应用开始迅速普及。实证研究者通过分位数回归来检验预测变量对因变量分布的影响。由经济学家(Buchinsky，1994；Chamberlain，1994)完成的两篇早期实证研究论文，为我们提供了如何将分位数回归应用到工资研究中的实际例子。借助分位数回归，他们全面分析了工资的条件分布，发现教育和工作经验的回报以及工会成员身份的效应在不同的工资分位点上是不同的。采用分位数回归分析工资的例子不断增加，并且扩展至另外一些话题，如工资分布的变化(Machado & Mata，2005；Melly，2005)、特定行业内的工资分布(Budd & McCall，2001)、白人与少数族裔(Chay & Honore，1998)、男性与女性(Fortin & Lemieux，1998)的工资差距、受教育水平和工资不平等(Lemieux，2006)以及收入的代际转移(Eide & Showalter，1999)。分位数回归同样应用于分析学校的教育质量(Bedi & Edwards，2002；Eide，Showalter & Sims，2002)以及人口特征对婴儿出生体重的影响(Abreveya，2001)。分位数回归还延伸至其他领域，特别是社会学(Hao，2005，2006a，2006b)、生态学和环境科学(Cade，Terrell & Schroeder，1999；Scharf，Juanes & Sutherland，1989)，还有医学和公共卫生等领域

(Austin et al., 2005; Wei et al., 2006)。

本书旨在向那些对分布形状和位置的建模方法有着浓厚兴趣的社会科学家们介绍分位数回归模型。此外,本书同样适合那些关注线性回归模型易受偏态分布和离群值影响这一问题的读者们。该书的写作主要建立在凯恩克及其同事们的奠基性著作上(如 Koenker, 1994; Koenker, 2005; Koenker & Bassett, 1978; Koenker & d'Orey, 1987; Koenker & Hallock, 2001; Koenker & Machado, 1999)并作出了两大新贡献。在分位数回归估计值的基础上,我们发展了基于条件分位数上形状变化的测量方法。这些测量方法为我们提供了关于协变量如何影响因变量的分布形状这一研究问题的直接答案。另外,为了获取更好的模型拟合效应,不平等研究常常对右偏的因变量分布进行对数转换,并未考虑在这种情况下"不平等"指的是初始数值的分布。因此,我们发展出一套方法,从对数单位系数中计算协变量对条件分位数函数的位置和形状的绝对值效应。

从我们的研究经验中知道,这本书是为从事实证研究的学者而编写的。我们采用社会科学家熟悉的语言和步骤进行教学,具体包括定义清晰的术语、简化的方程式、插图、实证数据的图表和社会科学家熟悉的统计软件的计算编码。贯穿全书,我们从自己的家庭收入研究中提取实际例子进行讲述。为了更好地介绍分位数回归,我们使用简化的模型设定,在这里,不管是初始单位还是对数转换的因变量,其条件分位数函数对于协变量而言都是线性的和可加的。正如在线性回归中,我们介绍的方法可以轻松地应用于更加复杂的模型设定中,例如交互项和协变量的多项式或样条函数。

本书内容组织如下：第 2 章从两个方面定义分位数和分位数函数——运用分布函数和解决最小化问题。相对于分布矩阵（如均值、标准差），本章还提出测量分布位置和形状的分位差方法。第 3 章比较了线性回归模型和分位数回归模型（QRM）的基本原理，包括模型建立、估计量和特性。通过特定的分位数参数来建构多条分位数回归方程是分位数回归模型的独特性质。我们将展示如何运用最小距离原则来拟合分位数回归方程。QRM 假设分布具有单调同变性和稳健性等特性，这些特性可为我们提供灵活稳健的估计，此外，QRM 还具有其他线性回归模型所不具备的性质。在第 4 章里，我们讨论了分位数回归模型的推论方法。除了介绍分位数回归系数的渐近推论外，本章还强调自举法的实用性和可行性。另外，相对于线性回归模型，我们还简短地讨论了分位数回归模型的拟合优度。第 5 章提出了多种解释分位数回归估计值的方法。本章超越了协变量对特定条件分位数（如中位数或其他非中心分位数）效应的传统检验，主要关注对分布的理解。它阐述了对分位数回归估计值的图像化解释和从分位数回归估计值中对形状变化的定量测量，包括位置转移、尺度变化和偏态变化等。第 6 章讨论与单调转换因变量相关的话题。我们提出了两种方法，从对数单位系数中获得协变量对条件分位数函数的位置和形状的绝对值效应。第 7 章讲述了本书介绍并加以发展的技术的系统运用。在本章中，我们分析了美国在 1991 年至 2000 年持续并扩大的收入不平等的原因。最后，附录提供了第 2 章解决最小化问题的中位数和分位数的证明以及执行第 7 章所描述的分析任务的 Stata 命令。

第2章

分位数和分位数函数

描述并比较总体的分布特征，是社会科学的本质。描述分布最简单和最常见的方法，莫过于寻找代表中心位置的平均值和揭示离散程度的标准差。然而，将注意力仅仅局限于平均值和标准差，无疑会让我们忽视其他有助于深入挖掘分布特征的重要特性。对于许多学者而言，他们感兴趣的总体多为偏态分布，因此均值和标准差并不是测量位置和形状的最佳方法。为了描绘非对称分布的位置和形状特征，本章用累积分布函数(CDF)的方法介绍分位数、分位数函数及其特性，并且发展出测量分布位置和形状的分位差方法，最后将分位数重新定义为最小化问题的解决方法。

第 1 节 | 分布函数、分位数和分位数函数

我们可通过分布函数描绘一个随机变量 Y 的分布。分布函数即在给定的函数 F_Y 中，对于每一个 y 值，当 $Y \leqslant y$ 时在总体中所占的比例。图 2.1 呈现了标准正态分布的分布函数。分布函数可用做计算 y 值在任意区间占总体的比例。由图 2.1 可知，$F_Y(0) = 0.5$ 和 $F_Y(1.28) = 0.9$。我们可以通过这一函数计算所有其他关于 Y 的概率。特别有 $P[Y > y] = 1 - F_y(y)$（例如，在图 2.1 中，$P[Y > 1.28] = 1 - F_y(1.28) = 1 - 0.9 = 0.1$）和 $P[a < Y \leqslant b] = F_Y(b) - F_Y(a)$）（例如，在图 2.1 中，$P[0 \leqslant Y \leqslant 1.28] = F_Y(1.28) -$

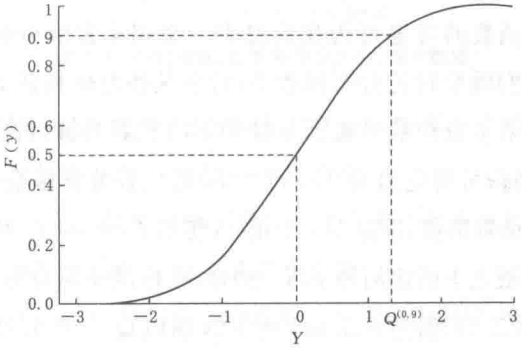


图 2.1　标准正态分布的累积密度函数

$F_Y(0)=0.40$)。分布函数最重要的两大特性是单调性(例如,当 $y_1 \leqslant y_2$ 时,$F(y_1) \leqslant F(y_2)$)和它的极限 $\lim_{y \to -\infty} F(y)=0$ 和 $\lim_{y \to +\infty} F(y)=1$。对于一个连续性随机变量 Y,我们同样可用概率密度函数来表示它的分布,对于所有 a 和 b 的取值,都有 $P[a \leqslant Y \leqslant b]=\int_{y=a}^{b} f_Y dy$。

让我们回到通过位置和离散并不足以充分描述一个分布的话题。假如我们知道白人家庭的平均收入(W)比黑人家庭的平均收入(B)多出 20500 美元。这可以简单描述成形状不变的分布图在位置上的移动(见图 2.2a 中对应的密度函数),因此这两种分布的关系可表示为 $F^B(y)=F^W(y-20500)$。但事实上,这两种分布的差异同时体现在位置和形状上(见图 2.2b 中对应的密度函数),所以对于常数 a 和 $c(a>0)$,两种分布之间的关系可归纳为 $F^B(y)=F^W(ay-c)$。这就是当 y 的均值和方差在总体 W 和 B 之间都不相同时出现的情况。对位置和尺度的测量方法,如均值和标准差,或者中位数和四分位距,有助于我们比较两种分布的 Y 属性。

分布越不对称时,需要越复杂的分析方法。对分位数和分位数函数的考虑可为我们提供一系列丰富的分析方法。下面我们继续讨论分布函数 F,对于某些总体特征而言,该分布的第 p 分位数可表示为 $Q^{(p)}(F)$(或者当被讨论的总体是已知时,可简化为 $Q^{(p)}$),$Q^{(p)}(F)$则代表分布函数在 p 点上的反函数的值,即存在一个值 y,使得 $F(y)=p$。所以,处于 $Q^{(p)}$ 值之下的比例为 $p\%$。例如,在标准正态分布的例子中(见图 2.1),因为 $F(1.28)=0.9$,所以 $Q^{(0.9)}=1.28$,那就是在值 1.28 之下的比例为 0.9 或 90%。

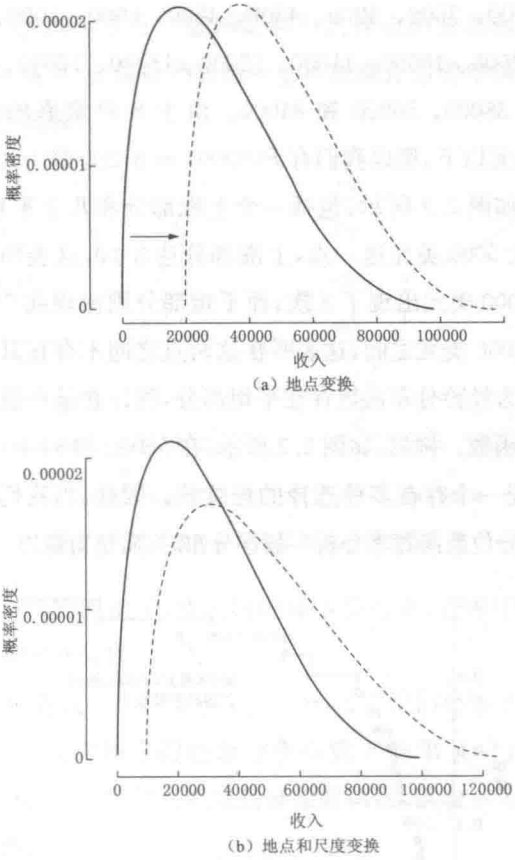


图 2.2　位置移动、位置和尺度变化(假设数据)

类似于总体的分布函数,我们考虑对应于一个样本的经验或样本分布函数。对于一个包含值 y_1, y_2, …, y_n 的样本,经验分布函数表示样本值小于或等于任意值 y 所占的比例。经验分布函数 $\hat{F}$ 的正式的定义为:

$\hat{F}(y)$ = 样本值小于或等于 y 值所占的比例

例如,考虑一个包含 20 户家庭收入情况的样本(单位:美

元)，3000、3500、4000、4300、4600、5000、5000、5000、8000、9000、10000、11000、12000、15000、17000、20000、32000、38000、56000 和 84000。由于 8 户家庭的收入在 5000 美元以下，所以我们有 $F(5000) = 8/20$。这一经验分布函数图如图 2.3 所示，包括一个上涨部分和几个平坦部分。例如，在 5000 美元这一点，上涨部分达 3/20，这表明在该样本中，5000 美元出现了 3 次；而平坦部分则出现在 56000 美元和 84000 美元之间，这表明在这两点之间不存在其他样本值。因为经验分布函数存在平坦部分，所以在某些值上存在多个反函数。例如，如图 2.3 所示，在 56000 和 84000 之间的 $Q^{(0.975)}$ 是一个存在多种选择的连续统。因此，当我们采用分位数和分位数函数来分析一般性分布时，需要留意以下定义。

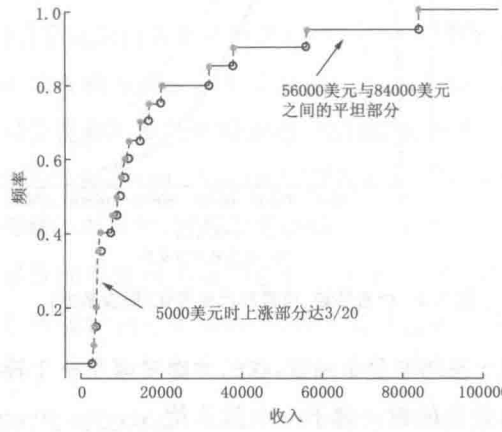


图 2.3 具有跳跃和水平部分的累积密度函数

定义：一条分布函数下的第 p 分位数 $Q^{(p)}$ 是一组 y 值中的最小值，从而使 $F(y) \geqslant p$。函数 $Q^{(p)}$（作为 p 的函数）正是 F 的分位数函数。

图 2.4 展示了分位数函数和与之相应的分布函数。由此可以观察到：分位数函数是一条从底端开始的单调非递减的连续性函数。

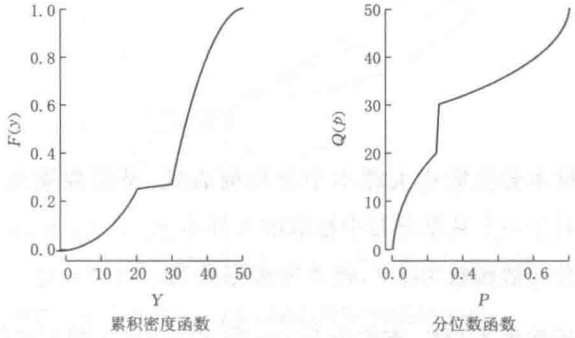


图 2.4　累积密度函数及其相应的分位数函数

作为特殊的例子，我们讨论样本分位数，它可用做估计抽样分布的分位数。

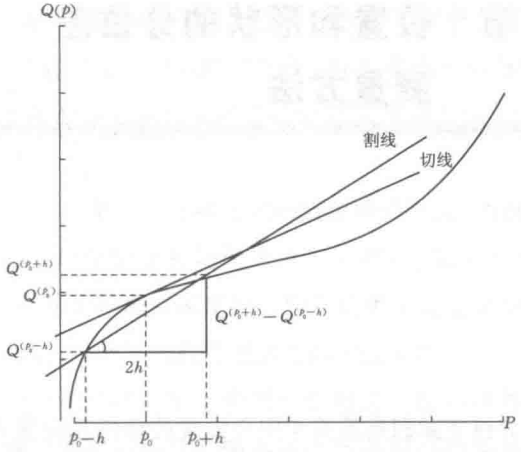
定义：给定一个样本 y_1，y_2，…，y_n，我们将第 p 样本分位数 $\hat{Q}^{(p)}$ 定义为相应的经验分布函数 $\hat{F}$ 的第 p 分位数，即 $\hat{Q}^{(p)} = \hat{Q}^{(p)}(F)$。与之相应的分位数函数，就表示为样本分位数函数。

样本分位数与顺序统计量密切相关。假定样本 y_1，y_2，…，y_n，我们按其大小从低到高排列，则表示为 $y_{(1)}$，…，$y_{(n)}$，有 $y_{(1)} \leqslant y_{(2)} \leqslant \cdots \leqslant y_{(n)}$。有些数值如果出现多次，它们会被重复。顺序统计量和样本分位数的关系可简单描述如下：对于大小为 n 的一个样本，第 (k/n) 样本分位数由 $y_{(k)}$ 确定。例如，在以上的 20 户家庭收入数据中，第 (4/20) 的样本分位数，即第 20 百分位数，等于 $\hat{Q}^{(0.2)} = y_{(4)} = 4300$。

第 2 节 | 样本分位数的抽样分布

样本分位数在大样本中会如何表现，是需要重点关注的。对于一个从某分布中抽取的大样本 $y_{(1)}$，…，$y_{(n)}$，该分布的分位数函数为$\hat{Q}^{(p)}$，概率密度函数为 $f = F$，$\hat{Q}^{(p)}$ 的分布接近均值为 $Q^{(p)}$、方差为 $\frac{p(1-p)}{n} \cdot \frac{1}{f(Q^{(p)})^2}$ 的正态分布。特别的，这一样本分布的方差完全由在分位点上估计而来的概率密度决定。这种对分位点上的密度的依赖有简单直观的解释：如果分位数附近有较多数据点（更高的密度），那么样本分位数更稳定；相反的，如果分位数附近有较少数据点（更低的密度），那么样本分位数较不稳定。

为了估计分位数抽样的变异性，我们可以利用以上的方差近似值，但这需要事先估计未知的概率密度函数。图 2.5 给我们展示了一种标准的估计方法，函数$\hat{Q}^{(p)}$在点 p 的切线斜率是分位数函数在 p 点上的导数，同样的，有密度函数的倒数：$\frac{d}{dp}Q^{(p)} = 1/f(Q^{(p)})$。这一项式接近点（$p-h$，$\hat{Q}^{(p-h)}$）和（$p+h$，$\hat{Q}^{(p+h)}$）的割线斜率 $\frac{1}{2h}(\hat{Q}^{(p+h)} - \hat{Q}^{(p-h)})$，尤其当 h 为极小值时。



注：函数在 p_0 点的导数（切线的斜率）约等于割线的斜率。

图 2.5　如何估计分位数函数的斜率的图示