数量经济研究 The Journal of Quantitative Economics
第 4 卷 第 2 辑 VoL. 4. No.2
2013 年 9 月 September 2013

基于分位数回归的 S 市中低收入家庭可支配收入影响因素分析

张建同 王 萍

（同济大学,经济与管理学院，上海 200092）

摘要：本文通过调查获取 S 市民政局关于该市申请经济适用房家庭的相关数据，运用计量经济学中的分位数回归模型，分析家庭人口数、家庭所在区县、家庭主要收入者的年龄、受教育程度、就业状态以及所从事行业对该市中低收入家庭可支配收入的影响，细致和深入地研究了该市居民收入及其影响因素之间的回归关系。

关键词：经济适用房；家庭收入状况；分位数回归

The Analysis of the Disposable Income of City S Low-income Familiesan Based on Quantile Regression

Abstract: This paper obtains some important related data about the families who apply for affordable housing from the Check center in one Civil Affairs Bureau. Using econometrics quantile regression model, this paper analysis the factors impact the disposable income of low-income families in S city, including the family size, location, and age, educational attainment, employment status, and the industry of the family main income earner. This article gives a detailed and in-depth study on the regression relationship between incomes and its affecting factors.

Keywords: Affordable Housing; Family income; Quantile regression

引 言

住房问题关系国计民生，解决群众住房问题是建设社会主义和谐社会的重要方面之一。S 市经济适用房项目自 2009 年 12 月启动，实际核对工作在 2010 年开展并完成，共有 2415 户家庭接受委托进行核对；2011 年上半年和下半年共分两批开展，其中第一批即上半年开展的有 13 个区，第二批即 2011 年下半年全市 17 个区县开展。

目前的 S 市经济适用房是否真正实现了人们“居者有其屋”的愿望以及居民的家庭收入状况能否承担得起住房消费所带来的压力已成为许多专家、学者关注的焦点。随着经济发展和经济结构的调整，城镇居民的收入是动态变化的，而其家庭收入状况也随着家庭内部因素和外部宏观变量的影响而不断变化着。居民家庭收入状况不仅在学术界有广泛的研究意义，研究申请经济适用房家庭的收入状况也是政府制定相关政策制度的一个很重要参考指标。

如今关于家庭收入状况的研究一般集中于农民或牧民领域的笼统分析，较少有学者对发达城市居民的家庭收入进行研究，在实证研究方面的研究更是严重缺乏。本文通过调查获取到的 S 市申请经济适用房家庭的相关数据，运用计量经济学中的分位数回归模型，不仅定性更定量地研究分析居民经济状况的影响途径以及程度。

1 分位数回归国内外研究现状

在分位数回归理论的发展过程中，Koenker，Bassett，Powell 等都做出了非常重大的贡献。1978 年，Koenker 和 Bassett[1]首次提出了分位数回归的概念，他们是在 1818 年 Laplace 提出的中位数回归（最小绝对偏差估计）理论的基础上，提出了更为一般化的分位数回归模型，它弥补了最小二乘回归的不足，发展了线性分位数回归理论；1982 年，他们又研究了分位数回归的线性假设检验以及异方差的稳健性检验，为分位数回归的应用提供了保证[2][3]；1986 年，Bassett 等[4]研究了回归分位数的强相合性等性质；随着线性分位数回归的发展，同年 Powell[5]基于删失模型提出了非线性分位数回归；1987 年，Koenker 等[6]提出了线性模型的 L 估计法，同年 Koenker 等[7]提出了关于分位数回归的有效算法；Buchinsky[8][9][10]在 1995 年和 1998 年分别讨论了分位数回归模型渐近协方差矩阵的估计方法，以及分位数回归最新的一些发展，并应用它分析了美国女性薪水结构的变化情况；2000 年，Koenker 和 Zhijie Xiao[11]解决分位数回归过程中存在的特定推断问题；2001 年，Tasche[12]研究了分位数回归的无偏性；2002 年，Koenker 等[13]又讨论了线性异方差模型的 L 估计法。

国外关于分位数回归的技术研究趋于成熟，而国内关于分位数回归的理论研究总体相对偏少，并且大多停留在各领域应用上的研究。吴建南和马伟[14]把分位数回归技术与显著加权法通过估计极端行为参数的能力进行了比较。陈建宝等[15]利用分位数回归模型研究了中国居民收入与消费之间的关系；周耿[16]建立了网上商品热销的影响因素的分位数模型，通过分位数回归方法对不同热门程度产品的价格、信誉、保障标记以及口碑的需求弹性进行了实证分析。岳昌君，刘燕萍[17]建立了有关我国城镇居民收入状况的分位数回归模型，在研究中，他们发现低收入群体的收入不平等和受教育程度比较低是导致城镇居民收入差距拉大的重要原因。有关分位数回归理论和应用的研究正在逐步地完善与发展之中。

国外关于分位数回归的技术研究趋于成熟，也极大地促进了这项技术在各个领域的广泛应用。本文利用分位数回归模型，研究城市 S 的中低收入家庭的收入状况与其家庭内部因素的关系。

2 分位数回归模型以及 S 市中低收入家庭收入影响因素的实证研究

2.1 分位数回归模型的基本说明

根据 Koenker 和 Bassett 的设定，与最小二乘回归模型相对应的分为回归模型为：

$$y_i = \beta_0^{(p)} + \beta_1^{(p)} x_i + \varepsilon_i^{(p)} \tag{1}$$

其中$0 < p < 1$表示数值小于第 p 分位数的比例。

与最小二乘估计量不同，在分位数回归中，数据点到回归线距离的测量通过垂直距离的加权总和而求得[18]。估计第 p 个分位的参数 $\beta^{(p)}$ 即通过使表达式（2）完成最小化线性规划，以此最小化残差的总和，其中正向残差的权重为 p 而负向残差的权重为$(1-p)$。

$$\sum_{i=1}^{n} d_p(y_i, \hat{y}_i) = p \sum_{y_i \geq \beta_0^{(p)} + \beta_1^{(p)} x_i} \left| y_i - \beta_0^{(p)} - \beta_1^{(p)} x_i \right| + (1-p) \sum_{y_i < \beta_0^{(p)} + \beta_1^{(p)} x_i} \left| y_i - \beta_0^{(p)} - \beta_1^{(p)} x_i \right| \tag{2}$$

式（2）中，y_i 和 x_i 分别表示因变量和自变量，d_p 指的是绝对距离。 p 的基本含义是在回归线以上的数据占全体数据的百分比。若研究第 0.1 分位数回归线的系数时，位于回归线下方的观察值的

权重为 0.9，而上方的观察值的权重为 0.1。

分位数回归的目标函数使用的是绝对值偏差的加权求和，由于最小二乘法是使数据点距离回归平面的距离最小，它们提炼的是所有数据点的平均信息。而分位数回归本质上是通过分位数取 0 到 1 之间的任何值，调节回归平面的位置和转向，让自变量估计不同分位数的因变量。它也能在一定程度代表所有数据的信息，但更侧重于特定区域的数据，如极端位置的数据。

2.2 S 市中低收入家庭收入影响因素的实证研究

本文的数据来源于 S 市民政局，选取其中 31946 户申请经济适用房家庭的相关数据进行研究。

影响家庭包括外部宏观因素以及家庭内部因素，外部宏观因素包括地区经济发展水平、物价指数（CPI）、经济景气情况（PPI），年度 GDP 增长速度等，鉴于数据的局限性，数据的时间跨度较短，本研究只关注家庭内部因素。由于缺乏户主信息，在回归分析时选择家庭中可支配收入最高的成员为该家庭的主要收入者进行研究。本文选取家庭的地理位置、人口数、家庭主要收入者的年龄、性别、受教育程度、所从事的行业、就业状态为影响因素，研究它们与家庭人均年可支配收入的关系。

对于文化程度的量化方法是根据受教育的年限进行界定：0 年—其它（视为文盲）；3 年—学龄前儿童；6 年—小学；9 年—初中；12 年—高中或中专、职校；15 年—大专、高职；16 年—本科；19 年—硕士；22 年—博士。

鉴于就业状态过多，对其进行分类汇总如下：1—出租车司机；2—非正规就业；3—在职；4—无业；5—退休；6—纳保；7—征地养老； 8—支内回沪；9—其他。

居民所从事的行业包括：采矿业；电力燃气及水的生产供应业；房地产、公共管理和社会组织业；建筑业；交通运输、仓储和邮政业；教育业；金融业；居民服务和其他服务业；科学研究、技术服务和地质勘查业；农林牧渔业；批发和零售业；水利、环境和公共设施管理业；卫生、社会保障和社会福利业；文化、体育和娱乐业；信息传输、计算机服务和软件业；制造业；住宿和餐饮业；租赁和商务服务以及其他行业。

本研究所运用的回归模型设定如下：

$$y = a_0 + a_1 x_{age} + a_2 x_{num} + a_3 x_{edu} + b_i S_i + c_j Q_j + d_k H_k + e_m J_m + \varepsilon \qquad (3)$$

式（3）中，y 表示家庭人均可支配收入；x_{age} 表示年龄；x_{edu} 表示受教育年限，其定量化方法如前所述；对于定性变量，我们引入虚拟变量，其中 S_1 为性别的虚拟变量，Q_j 为 S 市区县的虚拟变量；H_k 为行业的虚拟变量；J_m 为就业状态的虚拟变量；ε 为随机误差项。

我们以年龄、受教育程度、性别、地区、所从事行业、就业状态为自变量，分别对经过处理得到的收入的 10%、25%、50%、75%和 90%的分位数进行回归，从而可以对申请经济适用房家庭的居民条件分布的不同位置进行分析。通过对不同分布点的差异做更详细的刻画，更深入地了解居民收入差异的真实原因所在。

利用 Stata11.2 软件计算，计算结果见表 1，由于区县、家庭主要收入者所从事行业及状态的变量较多，在下文中将选取以上变量中的一些特别情况进行研究，故在表 1 中只列出部分因素的回归结果。表 1 的第 2、3、4、5、6 行报告部分因素 10%、25%、50%、75%和 90%分位回归结果。

在虚拟变量中，选取某郊区县为地区变量的虚变量，选取女性为性别变量的基变量，行业变量的基变量被设定为农林牧渔业，将无业设定为就业状态的基变量。关于虚拟变量系数的解释都是相对于基变量而言。

表 1 影响收入因素分位数回归结果

自变量 / 分位	家庭人口数	年龄	教育	性别	拟 R^2
10%	-169.8756 （0.013）	-94.45225 （0.000）	351.1861 （0.000）	2073.845 （0.000）	0.1953
25%	-516.4972 (0.000)	-122.1583 (0.000)	464.5266 (0.000)	2011.07 (0.000)	0.1698
50%	-657.3629 (0.000)	-149.1623 (0.000)	535.0107 (0.000)	1625.417 (0.000)	0.1423
75%	-1135.741 (0.000)	-133.1585 (0.000)	656.812 (0.000)	1274.336 (0.000)	0.1179
90%	-2090.149 (0.000)	-173.7947 (0.000)	790.9636 (0.000)	1728.936 (0.000)	0.1285

注：表 1 中*表示系数， 括号内为 p 值。对于分位回归，t 统计量通过自举法（Bootstrapping）获得。

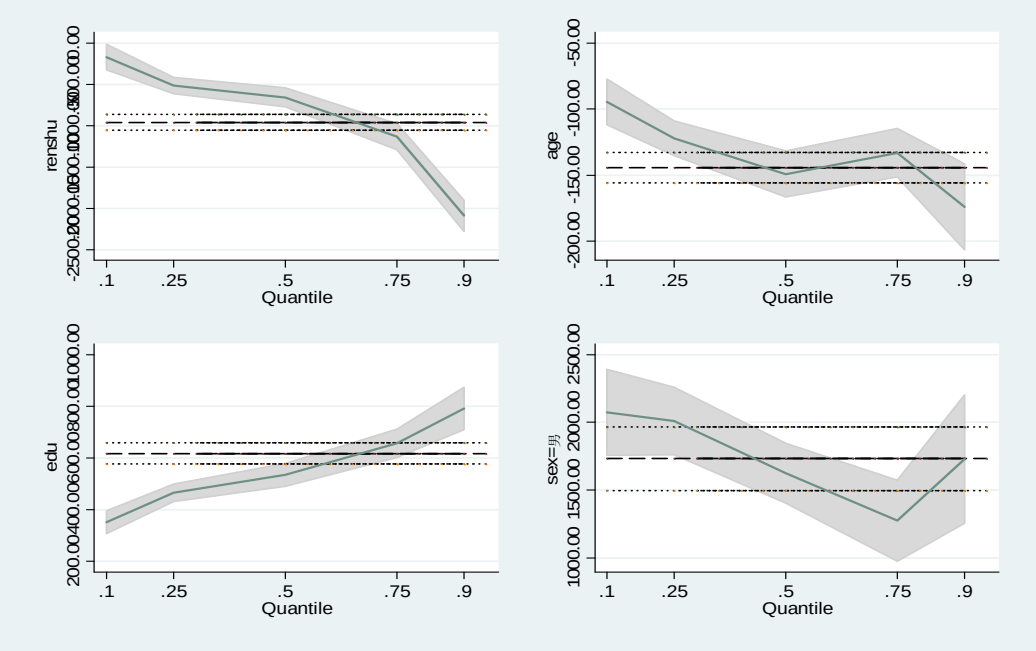

图 1　条件分位数系数图

注：renshu 为家庭人口数、age 为家庭主要收入者的年龄、edu 表示家庭主要收入者的受教育程度、sex=男表示家庭主要收入者的性别为男性。分位点的取值 0.1、0.25、0.5、0.75、0.9，图中的三条虚线为利用最小二乘法求的参数估计值及其 95%的置信区间上下限，灰色区域及其中的线段表示条件分位数模型系数的 95%Bootstrap 置信区间和点估计值。

通过分位回归，我们发现处于不同条件分布水平的中低收入居民，各变量的影响是不一样的。由表 1 和图 1 显示的结果可以看出：

在 0.1 的分位水平上，家庭人口数的回归系数为负值，说明家庭人口数对家庭人均收入有着负作用，并且这种负向的影响力随着分位数水平的上升而递增。表 1 的结果表明在 0.1 的分位水平上，家庭人口数的回归系数为-169.8756，在 1%的置信水平上并不显著，而在 0.9 的分位水平上，家庭人

口数的回归系数为-2090.149，并且高度显著，这个结果意味着对于处于条件分布（收入）底端的居民而言家庭人口数的负向影响作用不是很突出，原因在于这些居民文化程度相对较低，工作较早，大部分都能充当家庭的劳动力；而随着分位数的提高，家庭人口数的负影响作用越来越明显。

对于处于不同条件分布水平下的中低收入居民而言，家庭主要收入者的年龄的回归系数不仅普遍而且都很显著。如表 1 所示，年龄的回归系数在各分位水平上也都高度显著，家庭主要收入者的年龄的回归系数为负值，家庭主要收入者的年龄越大，家庭人均收入越低。随着分位数的上升，年龄的负影响越来越大。

对于处于不同条件水平的居民来说，受教育程度对家庭人均收入的影响都很显著。其系数随着10%、25%、50%、75%、90%的分位而上升，说明对于不同条件水平的居民来说，受教育程度是非常重要的一个影响因素。相对于其他因素，受教育程度是影响 S 市申请经济适用房家庭收入的最主要、最直接的因素之一，因此对教育的投入是提高中低收入家庭可支配收入的一个重要措施。

如表 1 所示，性别的回归系数为正，且在各分位水平上均高度显著，说明家庭主要收入者为男性的家庭人均可支配收入高于家庭主要收入者为女性的家庭人均可支配收入。如图 1 所示，性别的回归系数随着分位数的上升而减小，在 90%的分位回归中，其影响作用上升，这表明处在条件分布（收入）底端的居民，居民的工作能力存在着一定的性别的差距，随着平均收入向高水平位移，这种差距逐渐较弱。在 90%的分位回归，性别系数又有提高，说明进入高收入水平，相对而言男性的能力较强的事实又显现出来。

与普通回归结果相比，分位数回归不但反映了各影响因素与家庭收入状况的基本关系，而且显示了不同收入水平条件下，各影响因素对其的影响趋势。

表 2 部分区县变量的分位回归分析结果

分位 变量	Quant10	Quant25	Quant50	Quant75	Quant90
A1 区	1645.727*	3139.381	4039.317	4867.489	6327.793
	（0.134）	(0.001)	（0.003）	(0.000)	(0.004)
A2 区	749.923	2664.935	3309.353	4471.293	7198.228
	（0.521）	(0.004)	(0.020)	(0.003)	(0.001)
A3 区	1362.189	3564.303	3927.323	4721.201	6046.898
	（0.213）	(0.001)	(0.006)	(0.001)	(0.006)
A4 区	1927.61	3369.564	3721.562	4567.825	6018.476
	(0.077)	(0.000)	(0.007)	(0.001)	(0.007)
A5 区	2039.579	3785.697	4395.9	5895.695	7339.821
	(0.059)	(0.0000	(0.002)	(0.000)	(0.002)

注：表 2 中*表示系数， 括号内为 p 值。对于分位回归，t 统计量通过自举法（Bootstrapping）获得。

根据回归的结果，在 10%的分位水平上，各区县的回归结果均不显著，这个现象说明处在条件分布（收入）底端的居民，各区县与该郊区县的差异不是很显著。对申请经济适用房家庭的数据进行分位回归，发现一个现象：所在位置为 A1 区、A2 区、A3 区、A4 区以及 A5 区的家庭的人均收入普遍高于其他地区，由表 2 可知，这五个区县的回归结果从 25%分位回归到 90%分位回归都是显著的。

如表 3 所示在行业变量中，申请经济适用房的家庭主要收入者所从事行业为：电力燃气及水的生产供应业(H2)、教育业(H7)、金融业(H8)以及卫生、社会保障和社会福利业（H15）的家庭总收入普遍高于其他行业，并且在各个水平上均显著，说明家庭主要收入者所从事行业为以上一个行业的，其家庭的人均收入相对较高。此外，信息传输、计算机服务和软件业的回归系数分别为 5817.127、4555.581、1751.901、468.6729 和 2543.235，其影响作用明显减弱，仅在 10%和 25%的水平上显著，在 50%的分位回归中其影响作用不再显著。

表 3 部分行业变量的分位回归分析结果

分位 变量	Quant10	Quant25	Quant50	Quant75	Quant90
H2	8066.102*	7567.521	6335.772	4266.42	4664.899
	（0.001）	（0.000）	(0.000)	(0.061)	(0.016)
H7	8831.182	8302.149	6204.667	4580.641	8745.434
	（0.000）	（0.000）	(0.000)	(0.019)	(0.000)
H8	8939.644	6830.976	4357.691	4614.321	8061.895
	（0.000）	（0.000）	(0.001)	(0.021)	(0.000)
H15	6797.499	6474.402	5093.565	5111.725	9421.35
	（0.003）	（0.0000	(0.000)	(0.022)	(0.000)
H18	5817.127	4555.581	1751.901	468.6729	2543.235
	（0.015）	(0.007)	(0.194)	(0.814)	(0.102)

注：表 3 中*表示系数， 括号内为 p 值。对于分位回归，t 统计量通过自举法（Bootstrapping）获得。

如表 4 所示在就业状态变量中，申请经济适用房的家庭主要收入者的就业状态为：退休(J5)、在职(J7)和支内回沪（J9）的家庭人均收入普遍高于其他状态，并且在各个水平上均高度显著。在某种程度上说明 S 市对于退休的居民以及支内回沪的居民的保障力度较好。

表 4 就业状态变量的分位回归分析结果

分位 变量	Quant10	Quant25	Quant50	Quant75	Quant90
J5	14619.78*	15780.34	16034.13	11563.55	11097.62
	（0.000）	(0.000)	(0.000)	(0.000)	(0.000)
J7	10164.23	10228.18	9773.833	5840.145	6310.414
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
J9	14378.36	15769.17	16373.73	12145.93	11377.6
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)

注：表 4 中*表示系数， 括号内为为 p 值。对于分位回归，t 统计量通过自举法（Bootstrapping）获得。

3 总　结

如前文分析，提取 S 市申请经济适用房家庭主要收入者的年龄、受教育程度、所从事行业、就业状态以及家庭的人口数、家庭的地理位置等因素来分析其对 S 市中低收入家庭收入状况的影响。采用分位数回归来研究各因素对 S 市申请经济适用房家庭的收入的影响程度。主要得到以下结论：

居民的受教育程度对不同收入水平、不同地区、不同行业的家庭的年可支配收入影响程度均高

度显著，受教育水平是影响收入的最主要、最直接的因素之一。相对而言受教育程度较低的居民收入较低，高收入的行业专业性强，要求的学历高，低学历的居民难以进入高收入的行业（例如：金融业、科学研究等）。

家庭人口数以及主要收入者的年龄对收入的负影响也都十分显著，在保持其他因素不变的情况下，家庭人口数越多、主要收入者的年龄越大家庭的总收入越少。

从分位数回归的结果可以看出，对于处在条件分布（收入）底端的居民，居民的工作能力存在着一定的性别差距，随着平均收入向高水平位移，性别差距逐渐较弱。

某些行业，如金融业、教育业、电力燃气及水的生产供应业以及卫生、社会保障和社会福利业的家庭总收入普遍高于其他行业，并且在各个水平上均高度显著。这些行业基本上都是高学历要求行业，或者是紧缺资源行业，从数据也可以看出 S 市中低收入家庭从事这些行业的人数相对其他行业而言较少。

在家庭主要收入者的就业状态中，各个就业状态均显著区别于无业状态，在保持其他变量不变的情况下，家庭主要收入者的就业状态对收入的影响有着较大的影响。

参考文献:

陈建宝，杜小敏，董海龙. 2007．基于分位数回归的中国居民收入和消费的实证分析.统计与信息论坛，(24)：44～50

吴建南，马伟. 2006．估计极端行为模型：分位数回归方法及其实现与应用.数理统计与管理，(25)：536～542

肖东亮. 2012．分位数回归模型.上海：格致出版社，40～52

岳昌君，刘燕萍. 2006．教育对不同群体收入的影响.北京大学教育评论，(2)：86～92

周耿. 2011．网上商品热销的影响因素分析-基于分位数回归的实证研究.财经论丛，(5)：100～104

Bassett, G.a, Koenker, R.．1986．Strong Consistency of Regression Quantiles and Related Empirical Processes. Econometric Theory，2:191～201

Buchinsky, M.．1998．Recent Advances in quantile Regression Models. The Journal of Human Resources，1：88～126

Buchinsky, M.．1998．The Dynamics of Changes in the Female Wage Distribution in the USA: A Quantile Regression Approach. Journal of Applied Econometrics,13：1～30

Buchinsky，M. 1995．Estimating the Asymptotic Govariance Matrix for quantile Regression Models; A Monte Carlo Study. Journal of Econometrics, 68(2)：303～338.

Kim T H, Muller C. 2004. Two-stage quantile regression when the first stage is based on quantile regression[J]. The Econometrics Journal, 7(1): 218~231.

Koenker R, Zhao Q. 1994. L-estimatton for linear heteroscedastic models. Journaltitle of Nonparametric Statistics, 3(3-4): 223-235.

Koenker R W, d'Orey V. 1987. Computing regression quantiles. Applied Statistics, 36(3): 383-393.

Koenker, R., Bassett,G. 1978．Regression Quantiles.Econometrical, 46 (1) ： 33～50

Koenker, R., Z.Xiao.2002．Inference on the Quantile Regression Process. Econometrica.,70:1583～1612

Koenker, R.and Bassett, G.．1982b． Tests of Linear Hypotheses and L1 Estimation.Econometrica, 50:1577～1584

Koenker,R, S.Portnoy. 1987. L-Estimation for Linear Models. Journal of the American Statistical Association., 82:851～857

Koenker,R., Bassett,G.1982．Robust Tests for Heteroscedasticity Based on Regression Quantiles.Econometrica, 50:43～61

Powell, J.L．1986．Censored Regression Quantiles.Journal Econometrics.,32:143～155