学校编码：10384　　分类号_____ 密级_____

学号：15420111151895　　UDC_____

厦门大学

硕 士 学 位 论 文

厦门大学数据挖掘研究中心

Data-Mining Reaearch Center, Xiamen University

姜 叶 飞

指导教师姓名：唐 三 藏 教授

专 业 名 称：西 天 取 经

论文提交日期：2012 年　　月

论文答辩时间：2012 年　　月

学位授予日期：2012 年　　月

答辩委员会主席：_____

评　　阅　　人：_____

2012 年　　月

厦门大学学位论文原创性声明

本人呈交的学位论文是本人在导师指导下，独立完成的研究成果。本人在论文写作中参考其他个人或集体已经发表的研究成果，均在文中以适当方式明确标明，并符合法律规范和《厦门大学研究生学术活动规范（试行）》。

另外，该学位论文为（　　）课题（组）的研究成果，获得（　　）课题（组）经费或实验室的资助，在（　　）实验室完成。（请在以上括号内填写课题或课题组负责人或实验室名称，未有此项声明内容的，可以不作特别声明。）

声明人（签名）：

年　　月　　日

厦门大学学位论文著作权使用声明

本人同意厦门大学根据《中华人民共和国学位条例暂行实施办法》等规定保留和使用此学位论文，并向主管部门或其指定机构送交学位论文（包括纸质版和电子版），允许学位论文进入厦门大学图书馆及其数据库被查阅、借阅。本人同意厦门大学将学位论文加入全国博士、硕士学位论文共建单位数据库进行检索，将学位论文的标题和摘要汇编出版，采用影印、缩印或者其它方式合理复制学位论文。

本学位论文属于：

（　　）1. 经厦门大学保密委员会审查核定的保密学位论文，于

年　　月　　日解密，解密后适用上述授权。

（　　）2. 不保密，适用上述授权。

（请在以上相应括号内打“√”或填上相应内容。保密学位论文应是已经厦门大学保密委员会审定过的学位论文，未经厦门大学保密委员会审定的学位论文均为公开学位论文。此声明栏不填写的，默认为公开学位论文，均适用上述授权。）

声明人（签名）：

年　　月　　日

摘 要

本文基于面板数据的非参数回归估计进行了实例研究以及 R 语言算法的实现，实例考察了居民生活标准的变化对经济发展的影响。首先，本文构建出相关指标，并使用因子分析降维成经济、社会和高校规模三个因子；然后分别采用固定效应和随机效应的非参数模型进行平均水平的估计以及逐点估计得其弹性影响。此外，本文还分别对各指标进行了上述的非参数逐点估计分析，从而更为完整的解释了居民生活标准的变化对经济增长的正向影响及其变化趋势。最后本文还对前面的估计作了灵敏度和优缺点分析以及未来研究的展望。

关键词： 面板数据, 非参数计量经济学, R 语言, 弹性影响

Abstract

This paper's analysis is based on the Nonparametric Regression Estimate theory of Panel Data, and also includes a Example and a achievement of R-Language algorithm. The Example Study examined how the changes of the living standard of residents affect the economic development. At first, we build a set of relevant indexes and reduce the dimensions to Economic Factor, Social Factor and University Scale Factor through Factor Analysis. Then we estimate the mean level with Nonparametric Fixed-Effects Model and Random-Effects Model respectively, and obtain the three factors' elasticity on the GDP's growth resulting from the pointwise estimation. Then this paper also adopts Nonparametric pointwise estimations to analysis all indexes, which more fully explains that the changes of the residents' living standards have positive effects on GDP's growth and its fluctuant tendency. Finally, we analysis the Robustness for previous estimations, and then summarize their Strengths and Weaknesses, while look forward to Extension and Further Work.

Key Words: Panel Data , Nonparametric Econometrics , R-Language , Elasticity Effect

目 录

第一章 学习

1.1 引言

1.1.1 研究背景

计量经济学 (Econometrics) 作为经济学科的一个分支学科，在 20 世纪 20 年代末由 R.Frish 创立，经过 40 多年的发展，其经典理论方法已经成熟。自 20 世纪 70 年代以来，随着经济活动的复杂性增强和计量经济学应用领域的扩展，计量经济学理论方法得到了很大的发展。除了 2000 年诺贝尔经济学奖获得者 J.J.Heckman 对选择性样本模型理论的发展和 D.L.Mcfadden 对离散选择模型理论的发展，以及以此开创的微观计量经济学之外，宏观领域中动态计量经济学模型理论方法的发展是这个阶段最重要的部分，该方法以非参数模型理论方法为基础，并形成了较为完整的内容体系，构成了现代计量经济学的主要部分[?]。然而到目前为止，它们仍旧处于新的领域，尚在研究和发展之中。因此，本文希望在前人研究的基础上，试图对它们的发展做出一点贡献。

1.1.2 R 语言概述

R 语言是主要用于统计分析、绘图的语言和操作环境。R 本来是由来自新西兰奥克兰大学的 Ross Ihaka 和 Robert Gentleman 开发。现在由“R 开发核心团队”负责开发。R 是基于 S 语言的一个 GNU 项目，所以也可以当作 S 语言的一种实现，通常用 S 语言编写的代码都可以不作修改的在 R 环境下运行。R 的语法也同样是来自于 Scheme[?]。

R 语言给我们提供了一套完整的数据处理、计算和制图软件系统。其功能包括：

- 有效的数据存储和处理功能
- 数组运算工具（其向量、矩阵运算方面功能尤其强大）
- 拥有完整体系的数据分析工具
- 为数据分析和显示提供的强大图形功能

- 一套（源自 S 语言）完善、简单、有效的编程语言（包括条件、循环、自定义函数、输入输出功能)。

R 的定位是一个完善、统一的系统，而非其他数据分析软件那样作为一个专门、不灵活的附属工具。R 很适合被用于发展中的新方法所进行的交互式数据分析。由于 R 是一个动态的环境，所以新发布的版本并不总是与之前发布的版本完全兼容。某些用户欢迎这些变化因为新技术和新方法的所带来的好处；有些则会担心旧的代码不再可用。简而言之，R 是 GUI 加上全球各地各种职业的个人和团体编写的包 (Packages) 组成的一个环境，使用者可以随时添加、删除以及更新这些包。从而在应用中，R 并不仅仅是局限于普通的统计软件。因此选用 R 作为本次模型的实现环境会有较好的通用性和自由性。

1.2 参数模型与非参数模型的比较分析

1.2.1 计量经济参数模型简析

在日常生活以及经济活动中，我们常常需要研究生产投入与产出之间的关系。例如，在宏观方面，以 $Y = A(t)L^{\alpha}K^{\beta}\mu$ 为基本形式的 Cobb-Douglas 生产函数通过分析两者的关系，分析并预测了各个国家、地区的工业系统或大企业的产出水平；在微观方面，商品消费支出和收入之间关系（或恩格尔曲线）的研究也对两者之间的关系进行分析和预测以商品消费支出和收入之间关系为例，Working(1943)[?]、Leser(1963)[?] 的实证研究表明，消费支出份额关于总支出的线性参数恩格尔曲线模型 (Workling-Leser 模型) 可以较好地拟合其实际样本数据，模型形式如下：

$$w_k = \xi_k + \beta_k \log M \tag{1.2-1}$$

其中，w_k 是商品类 k 的消费支出 m_k 占总消费支出 M 的比例，ξ_k 是与价格有关的常数，log 指以 e 为底的对数（下同）。这样表示的恩格尔曲线不但可以很好地拟合实际数据，而且还有一定的微观基础：它是几乎理想需求系统 (AIDS) 在价格给定的情况下的一个特例（详见 Deaton 与 Muellbauer(1980)[?] 的结论）。参照周先波与田凤平 (2008、2009)[? ?] 的研究结果可知，Working-Leser 参数模型设定的总消费支出的对数和各类商品消费份额之间的关系是线性的。较强线性约束使其不一定能很好地描绘所有的数据。与传统非参数核估计进行比较，使用非参数估计方法拟合的图形比参数估计拟合的图形更好地反映了两变量之间的变化趋势及特征。在下面的章节中我们将通过一组数据比较参数估计和非参数估计的拟合效果。

1.2.2 截面数据非参数 Nadaraya-Watson 核估计概述

假设 Y 为被解释变量，$\boldsymbol{X} = (X_1, \cdots, X_q)$ 为解释变量向量，是影响 Y 的 q 个重要因素。给定样本观测值 $(Y_1, \boldsymbol{X_1})$,$(Y_2, \boldsymbol{X_2})$,$\cdots$,$(Y_n, \boldsymbol{X_n})$，假定 $(Y_i, \boldsymbol{X_i})$ 独立同分布 (下面简称 i.i.d)，建立非参数回归模型：

$$Y_i = m(\boldsymbol{X_i}) + u_i \tag{1.2-2}$$

其中，$m(\cdot)$ 是未知的函数，$m(\boldsymbol{X_i}) = E(Y_i \mid \boldsymbol{X_i})$，$\varepsilon_i$ 是均值为零、方差为 1 且与

$\boldsymbol{X_i}$ 独立的序列，随机误差项 $u_i = \sigma(\boldsymbol{X_i})\varepsilon_i$，且 $E(u_i) = 0$，$E(Y_i) = m(\boldsymbol{X_i})$。则此时回归函数 $m(x)$ 的核估计 $\hat{m}_n(x)$ 为最小化：

$$\sum_{i=1}^{n} (Y_i - m(x))^2 K\left(\frac{\boldsymbol{X_i} - x}{h_n}\right) \tag{1.2-3}$$

其中，h_n 为窗宽，$K(\cdot)$ 为满足 $K(u) \geq 0$，$\int K(u)\,du = 1$，$\int K(u)\,udu = 0$ 的核函数。于是，我们可以得到公式 (1.2-3) 的核估计的表达式为：

$$\hat{m}_n(x) = \frac{\sum_{i=1}^{n} K\left(\frac{\boldsymbol{X_i} - x}{h_n}\right) Y_i}{\sum_{i=1}^{n} K\left(\frac{\boldsymbol{X_i} - x}{h_n}\right)} \tag{1.2-4}$$

1.2.3 对比分析

从公式 (1.2-4) 中可以看出，核估计相当于作局部加权的最小二乘估计。那么，经典参数最小二乘估计与 Nadaraya-Watson 核估计在拟合效果上到底有多大差别呢？下面我们将通过使用一组数据来对其进行比较。

在 R 里面的 MASS 包中，我们可以找到一组名为 mcycle 的用来测试在事故发生时的摩托车加速度和时间的数据[?]。其中横轴为时间 X_i/ms，纵轴为加速度 Y_i。通过下列算法进行拟合。其中，线性回归选择使用 stats 包中的 lm() 函数，非参数回归选择使用 KernSmooth 包中的 dpill 函数确定窗宽，核函数选择 Gauss 核的 locpoly 函数进行拟合，具体算法见算法**??**。

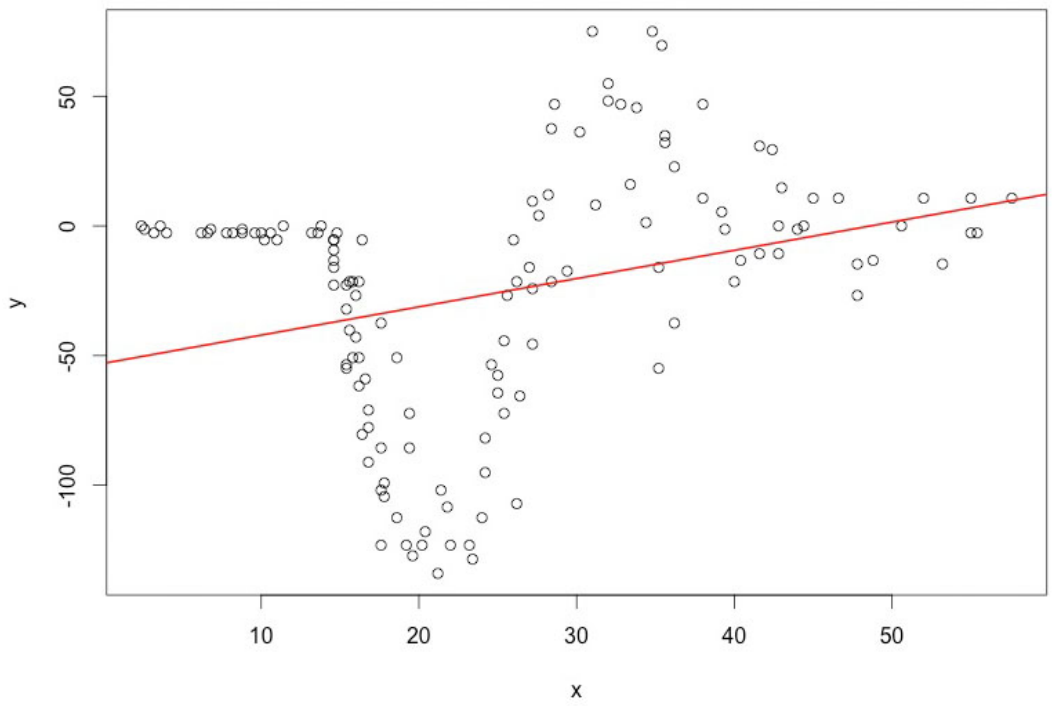

图 1.2-1: 摩托车数据及其线性拟合图

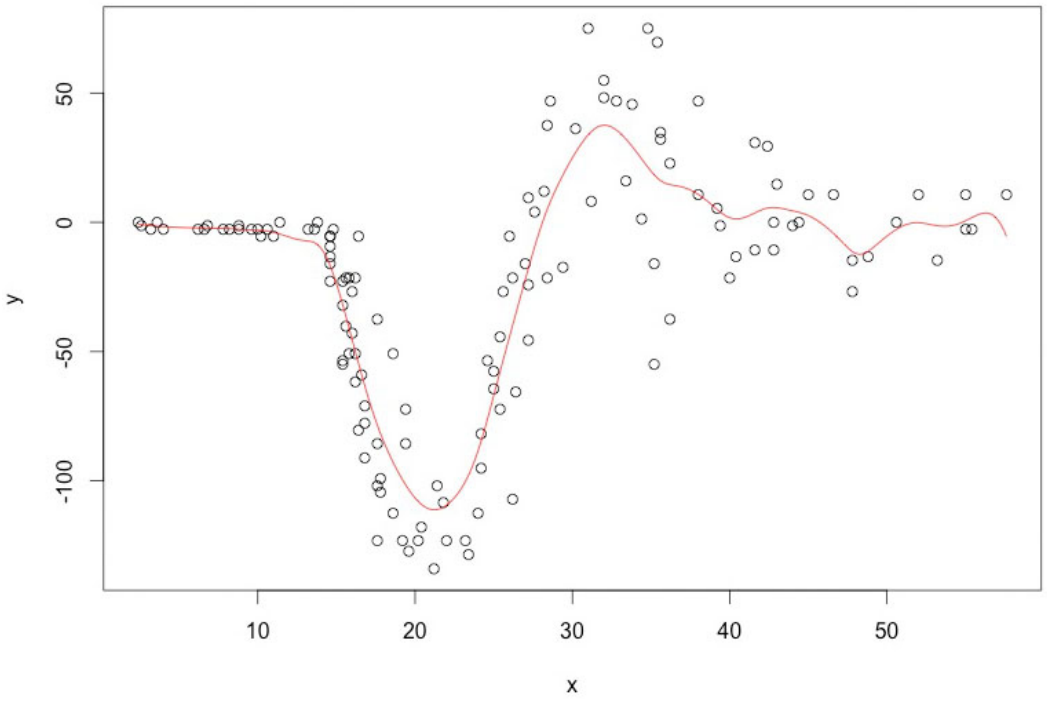

图 1.2-2: 摩托车数据的非参数回归拟合图

计算得到线性回归的拟合结果见图1.2-1，非参数回归见图1.2-2。从这些图中我

们可以清楚地看到，Y_i 与 X_i 的关系是非线性关系，给定已知的线性函数，拟合效果很差；即使使用其他已知的非线性函数，Y_i 与 X_i 的拟合效果也显然不能令人满意。而在图1.2-2中的非参数拟合图中我们可以看到非参数回归有很好的拟合效果，可以真实反映 Y_i 与 X_i 的实际关系（在不同的核函数和窗宽的选择下，拟合效果也有一定的差异)。