半参数回归模型

柴根象　洪圣岩　编著

安徽教育出版社

半参数回归模型

柴根象　洪圣岩　编著

安徽教育出版社

（皖）新登字03号

半参数回归模型

安徽教育出版社出版发行

（合肥市金寨路381号）

新华书店经销　　六安新华印刷厂印刷

*

开本：850×1168　1/32　印张：9　字数：220,000

1995年12月第1版　　1995年12月第1次印刷

印数：1-2,000

ISBN 7-5336-1644-8/G·2031

定价：8.00元

内 容 简 介

半参数回归模型是80年代才发展起来的一种重要的统计模型。由于这种模型既有参数分量，又含有非参数分量，并可以描述许多实际问题，因而引起广泛的重视。本书将详细地介绍这一类模型的基本概念及目前研究的问题以及今后发展的趋势。

序　言

“半参数模型”这个提法出现在数理统计文献中，为时尚不算长，广义地说，这个提法可以概括一大批现时常见的统计模型，但恐怕多数统计学者都会同意：唯有在回归函数分解成参数的和非参数的这种结构中，“半参数”的含义体现得最为清楚，确切且最富现实意义，以此，半参数回归一直是半参数模型研究中的重点，也就可以理解了。

回顾回归分析研究的历史，大致上在本世纪七十年代以前，重点在于参数回归，尤其是线性回归，这个方面目前仍在向纵深发展。七十年代以来，非参数回归的研究日渐兴起，吸引了一批学者的注意，半参数回归介于二者之间——可以设想，在不少实际问题中，它可能是一个更接近于真实、更能充分利用数据中所提供的信息的提法。在理论上，处理这种模型的方法融合了参数回归中习用的方法与较近发展起来的非参数方法，但也并非这两类方法的简单叠加。总的看，可以认为其复杂性和难度，都超过了单一性质的回归模型。因此可以说，它实在是一个在实用上有重大意义且在理论上富有挑战性的领域。

检阅迄今为止本领域所取得的成果，我们有信心说：我国学者的贡献，其中包括本书二位作者的贡献，无论从质和量的角度看，都可在国际上列入先进水平。然而，本领域的有关文献，散见于国内外众多的期刊杂志中，欲窥全豹，谈何容易。有见于此，本书二位作者以自己的研究经验为本，广泛搜集资料，去芜取精，撰写了这本专著。笔者有幸拜读了其原稿，感到其取材精当，可

谓要而不繁、不支不蔓，概括了迄今本领域的主要成果。论理严谨，叙述清楚，可读性强，对有意进入本领域研究或对这个方面的动向和现状感兴趣的学者、教师和研究生，以至对半参数回归方法感兴趣的应用工作者，不失为一本有价值的读物。

半参数回归是一个历史不长，尚在发展中的领域，目前所取得的成果中，有一些还不能认为是最终的。本领域主要属大样本性质，凡是对大样本统计的研究有些经验的人都了解：对大样本统计的一些基本问题，如相合性和渐近正态性，作出初步结果较易，而要达到完满的结论，比如说必要充分条件，则十分艰难，拿最简单的线性回归说，虽则现今的研究已可算很有深度了，但有些基本问题，至今仍未达到完全和最后的解决。由于半参数回归模型的复杂结构，可以预料，要把工作在目前的基础上深入下去，还会有不少困难要克服。例如，在线性回归研究已达到的成果中，有哪些在半参数回归中也可以达到，如有差距，能达到的最好形式当如何？另外，从应用角度看，实用工作者多关心一个方法在合理样本大小下的表现，与传统方法的比较及计算上的方便等。这些方面，需要做的工作还很多，因此笔者希望，本书的出版，能对半参数回归的研究向纵深发展作出贡献。我想，这也是本书作者的希望。

陈希孺

1994年6月25日

后　记

一. 本书的初稿是1992年底完成，书中涉及的理论结果及引用的文献，均为截止1992年底以前的工作；

二、本书的主要结果多属作者的工作，其中也有一定的篇幅介绍了中国科大高集体同志的许多重要成果；此外，也引用了中国科大赵林城教授、系统科学研究所李国英教授及北大施沛德.系统所梁华等同志的工作，在此谨向他们致谢；

三、本书得以出版得力于本书的责任编辑杨晓源先生，为本书的编辑、出版作出了巨大努力。作者的老师陈希孺教授为本书写了序，这是对作者的宝贵支持和鼓励。在此一并致以衷心的谢意；

四、由于本书作者处于两地，全书的统稿及修改较为匆促，错漏及疏忽所在不少，敬请国内同行不吝指教。

作者

1994.7.5

目　录

第一章 引 论

半参数回归模型是八十年代发展起来的一种重要的统计模型．由于这种模型既含参数分量，又含非参数分量，可以概括和描述众多实际问题，因而引起广泛的重视．本章将介绍这一类模型的若干基本概念和所要研究的几个主要问题．

§1.1 模 型

一、古典回归模型

在许多实际问题中，往往考察对象（指标Y）同若干因素（指标为$Z_1,\cdots,Z_K$）有关，但给定$Z_1,\cdots,Z_K$尚不足以完全确定Y（只能确定Y的条件分布），实际上Y与$Z=(Z_1,\ \cdots,\ Z_K)'$服从如下关系

$$Y=\mu(Z)+\varepsilon, \tag{1.1.1}$$

其中$\mu(\cdot)$是一个未知的k元函数，ε为随机误差，称满足(1.1.1)的变量Y与Z有回归关系，$\mu(\cdot)$为Y关于Z的回归函数．为对$\mu(\cdot)$作统计推断，对(Z,Y)作n次观察得$(Z_i,\ Y_i)$，$i=1,\ \cdots,\ n$，则$\{(Z_i,\ Y_i)\}_{i=1}^{n}$满足

$$Y_i=\mu(Z_i)+\varepsilon_i,\ \ i=1,\ \cdots,\ n \tag{1.1.2}$$

通常假定误差序列$\{\varepsilon_i\}$为不相关，且

$$E\varepsilon_i=0,\ \ \sigma^2=E(\varepsilon_i)^2<\infty,\ \ i=1,\ \cdots,\ n \tag{1.1.3}$$

称观察$\{(Z_i,\ Y_i)\}_{i=1}^{n}$满足的(1.1.2)为回归模型．回归模型

(1.1.2)的基本问题是，基于$\{(Z_i, Y_i)\}_{i=1}^{n}$估计回归函数$\mu(\cdot)$，或对$\mu(\cdot)$的假设作统计检验，统称对回归函数的统计推断方法为回归分析．自然，推断方法是同对回归函数的基本假设密切相关(严格来说，还同误差分布的假定有关)，因此文献中，常常依对回归函数的基本假设，对回归模型进行分类．

一类最简单，且又十分有用的回归模型是线性回归模型，即假定回归函数有如下的线性形式：

$$\mu(Z)=\beta_0+\beta_1 Z_1+\cdots+\beta_k Z_k, \qquad (1.1.4)$$

此时在回归函数$\mu(\cdot)$中，只有$k+1$个参数β_0，β_1，…，β_k是未知的．

另一类常用回归模型是非线性回归，即存在一已知非线性函数$g(\cdot;\ \beta_0,\ \beta_1,\ \cdots,\ \beta_k)$使得

$$\mu(\cdot)=g(\cdot;\ \beta_0,\ \cdots,\ \beta_k), \qquad (1.1.5)$$

其中β_0，…，β_k为未知参数，例如g可以是

$$\beta_0 Z_1^{\beta_1},\ \beta_0+\beta_1 Z_1\exp(-\beta_2 Z_2^{\beta_3})$$

等等．

这两类回归模型有一个共同特点：在回归函数中除去有限个参数其余都是已知的．因而也称具有这种特点的回归模型为参数回归模型（此处并未对误差分布进一步设定．尽管实际上大多数常用参数回归模型都假定误差分布为正态，因而此处的“参数回归模型”是与“参数统计模型”的概念不相同）．由于参数回归模型容易处理，且对其研究已有相当长的历史，因而已形成一套成熟的理论和方法．

二、非参数回归模型

参数回归模型对回归函数提供了大量的额外信息（通常由经验和历史资料提供），因而当假设模型成立时，其推断有较高的

精度．例如熟知的线性模型的最小二乘估计，其方差有$O(\frac{1}{n})$的阶．但当参数假定与实际背离时，基于假设模型所作的推断其表现可以很差，这种情况（也包括参数回归可能作出的正态误差假定与实际背离）促使人们寻找别的出路，而非参数回归则是朝着这个方向的一种努力．非参数回归模型的特点是：回归函数的形式可以任意，(Z, Y)的分布也很少限制，因而有较大的适应性．详而言之，设(Z, Y)为$R^k \times R$值随机变量，(Z_i, Y_i)，$i=1, \cdots, n$为(Z, Y)的n次观察，满足

$$Y_i=\mu(Z_i)+e_i,\ i=1, \cdots, n,\ \mu\in\mathbf{H} \tag{1.1.6}$$

其中$\mathbf{H}$为R^k上某个函数空间，$e_1, \cdots, e_n$ iid·F，$F\in\mathscr{F}$，而$\mathscr{F}=\{F: F$为一维d·f，$\int xdF(x)=0\}$，在(1.1.6)中，诸$Z_1, \cdots, Z_n$可为随机或非随机，当$\{Z_i\}$为随机时，还假定

(i) $\{Z_i\}_{i=1}^n$ iid·v，$v\in\mathbf{V}$，

(ii) $\{Z_i\}$与$\{e_i\}$相互独立．

在实际应用中，常常视模型(1.1.6)中的$\{Z_i\}$为非随机的，这可作如下解释：依前述假定，$Z_1=z_1, \cdots, Z_n=z_n$，$Y_1, \cdots, Y_n$条件独立，且$Y_i|Z_i=z_i$有条件分布$F(\cdot-\mu(z_i))$，此时(1.1.6)中的$\mu(Z_i)$正好是此条件分布的期望(Z_i已是常量)．因而与参数回归模型不同，在非参数回归中，$\{Z_i\}$可为随机或非随机．

三、半参数回归模型

非参数回归自stone（1977）的一项著名工作发表后，其理论和方法已有重要进展．这种模型虽有前段所说的优点，但从实

际应用来说，尚有它的局限性。例如影响Y的因素（即解释变量）可分为两个部分，即x_1，…，x_p及t_1，…，$t_q(p+q=k)$，根据经验或历史资料可以认为因素x_1，…，x_p是主要的，而且Y同x_1，…，x_p是线性的；而t_1，…，t_q则是某种干扰因素（或者看作协变量），它同Y的关系是完全未知的，而且没有理由将其归入误差项。此时如用非参数回归加以处理，则会失去太多的信息，若采用线性回归一般拟合情况很差。比较自然地是采用两者的“混合”。Engle，Gronger等（1986）曾讨论气象条件对供电量的影响，就适合这种情况。

此外，在非参数回归模型中，各个解释变量对因变量作用的差别往往被忽略，这在实际问题对此未提供任何信息时，是不可避免的；但若有根据认为某些解释变量对Y的影响较显著时，而使用非参数回归会明显地降低模型的解释能力。

为在出现上述情况时，弥补非参数回归之不足，一个方向的努力是rice(1982)，Engle等（1986）提出的偏线性回归。这种模型的结构部分地受线性回归中协方差分析的启示，但与之不同的是此处主要解释变量可以取连续指标值，且Y同“协变量”之间的关系为“非线性”。详而言之，记$Z^{\tau}=(x_1, \cdots, x_p, t_1, \cdots, t_q)=(x^{\tau}, t^{\tau})$，设$\mu(\cdot)$有如下形式：

$$\begin{aligned}\mu(Z)&=\beta_1 x_1+\cdots+\beta_p x_p+g(t)\\&=x^{\tau}\beta+g(t)\end{aligned}\qquad(1.1.7)$$

其中$\beta^{\tau}=(\beta_1, \cdots, \beta_p)\in R^p$，$g\in W$。

今设$\{Z_i, Y_i\}_{i=1}^n$为(Z, Y)的n次独立观察，则

$$Y_i=x_i^{\tau}\beta+g(t_i)+e_i,\ i=1, \cdots, n\qquad(1.1.8)$$

通常假定

(i) $e_1, \cdots, e_n$ iid·F，$F\in\mathscr{F}$，$Ee_1=0$。

当$\{Z_i\}$为随机时，还假定

(ii) $Z_1, \cdots, Z_n$ iid, $\{Z_i\}$ 与 $\{e_i\}$ 相互独立.

(iii) $g \in W$, W 为定义在 U ($\subset R^q$) 上的某个实值函数空间.

文献中称模型 (1.1.8) 为偏线性模型 (par tial linear model), 而 $x_i^T\beta$, $g(t_i)$ 分别称为该模型的参数分量与非参数分量, 习惯上也称 β 为回归参数. 我们的主要兴趣在于估计 β.

另外, 在模型 (1.1.8) 中往往假定误差方差 $Ee_1^2=\sigma^2>0$ 存在但未知, 此时对 σ^2 的估计也是一个重要的研究内容.

为了对该统计问题有进一步了解, 先引入如下记号:

记 $\Theta=\Theta_1\times\Theta_2$, $\Theta_2=W\times\mathscr{F}$, $\Theta_1=R^p$; Θ_2 中元用二元对 $\vartheta_2=(g, F)$ 表之, 又记对每一 $\theta=(\beta, \theta_2)\in\Theta$, (Z_i, Y_i) 有联合分布 $p_{\beta;\theta_2}^{\{i\}}$, $1\leqslant i\leqslant n$, 则 $\{(Z_iY_i)\}_{i=1}^n$ 的联合分布为

$$P_{\beta;\theta_2}=P_{\beta;\theta_2}^{\{1\}}*\cdots*P_{\beta;\theta_2}^{\{n\}},\ (\beta,\ \theta_2)\in\Theta \tag{1.1.9}$$

因此在偏线性回归模型 (1.1.8) 中估计 β, 等价于在具多余参数 θ_2 的模型 (1.1.9) 中估计 β. 显然回归参数是有限维, 而多余参数是无限维. 这是偏线性回归模型的一个重要特点, 因而文献中也常称模型 (1.1.8) 为半参数回归模型 (Semiparametric regression model).

半参数回归模型的优点是集中了主要部分(即参数分量部分)的信息, 因此有较强的解释能力, 深信随着此模型在理论和方法上的日益成熟, 必将有广阔的应用前景, 下面是这类模型的一个应用实例.

例1.1.1(截断回归模型)

设有线性模型

$$Y_i=x_i^T\beta+ei,\ i=1, \cdots, n$$

$\{e_i\}$ iid.F, $E|e_1|<\infty$, 但诸Y_i仅当$x_i^{\tau}\beta+e_i>0$时才能观察，记观察为Y_1^*，…，$Y_m^*(m\leqslant n)$，相应的e_i记为e_i^*，则e_i^*的分布为

$$F_i^*(W)=\{1-F(-x_i^{\tau}\beta)\}^{-1}\int_{-\infty}^{\infty}I\{y>-x_i^{\tau}\beta\}dF(y),$$

$i=1$，…，m

记

$$g(t)=\{1-F(-t^{\tau}\beta)\}^{-1}\int_{-t^{\tau}\beta}^{\infty}ydF(y),\quad t\in R^p$$

则易知Y_1^*，…，Y_m^*服从如下半参数回归模型

$$Y_i^*=x_i^{\tau}\beta+g(x_i)+\mu_i^*,\quad i=1,\ \cdots,\ m$$

这是模型（1.1.8）当$p=q$，$x_i=t_i$的特例.

至此我们已对本书要研究的模型作了简要的介绍. 这里要指出的是：以上提到的三种类型的模型各有其适应范围，离开各自的适应范围就无法评价这些模型的优劣，只能一般地说这些模型在其适用范围内都是十分有用的，其理论和方法都还在发展之中，即使如经典的线性模型，尽管对它的研究已有很长的历史，但无论从理论还是应用方法，仍在不断深化.

§1.2 估计方法

本节将在模型（1.1.8）的最为一般的假设下，介绍构造估计的主要思想及有关问题.

一、概述

在文献中，关于半参数回归的早期工作大多沿着这样的思路：对函数空间W施加一定的限制，这主要是指光滑性、由于W是无穷维的，通常由光滑性可使用合理的逼近形式，使得W中的元参

数化．例如W中的元在选定的基 $\{e_i\}$ 下有线性表示 $g(t)\approx\sum_{i=1}^{\infty}\theta_i e_i(\tau)$．若W中元有某种光滑性，使此级数一致收敛，则可用有限和 $g_\lambda(t)=\sum_{i=1}^{\lambda}\theta_i e_i(t)$ 逼近，此处 λ 是一个待选定的光滑参数，这样W中每一元 g，在选定的基 $\{e_i(\cdot)\}$ 及 λ 下，对应一个有限维参数 $\theta=(\theta_1,\cdots,\theta_\lambda)'$，因而将估计 g 的问题转化为估计有限维参数，然后使用最小二乘法或其它类似方法同时估计 β 及与 g 对应的参数 θ．由于这种估计以非参数分量 g 的参数化为特征，故文献中大多估计量以使用的参数化的方法命名，例如偏光滑样条估计，偏分块多项式估计，偏Fourier级数估计等等．

另外一种途径则是两阶段估计，其典型例子是核估计．第一步先设 β 已知，使用标准非参数回归方法，基于 $\{t_i, Y_i-x_i'\beta\}_{i=1}^n$ 估计 g，记估计量为 $\hat{g}_\lambda(\cdot,\beta)$；第二步再以 $\hat{g}_\lambda$ 代 g，使用最小二乘法找 β 的下述极小问题的解：

$$\sum_{i=1}^{n}[Y_i-x_i'\beta-\hat{g}_\lambda(t_i;\beta)]^2=\min$$

而 g 的最后估计为 $\hat{g}_\lambda(\cdot;\hat{\beta})$，此处 λ 是非参数回归估计中的光滑参数，λ 的取值可是实数、向量也可以是集值，例如在核估计时，则 λ 就是窗宽 h．

以上两种途径的一个共同特点是使用最小二乘法或其类似，而且在多数情况下得到的估计是线性的，这对理论分析带来不少方便。但众所周知，由最小二乘法得到的估计缺乏稳健性，新近，一些学者使用由Huber（1964）提出的稳健估计思想，考虑半参数M估计，即引进一个定义在 R 上的凸函数 $p(\cdot)$，找 $\hat{\beta}$ 及 $\hat{g}$ 使

$$\sum_{i=1}^{n}p(Y-x_i'\beta-g(t_i))=\min.$$

在此，上面提到的使 g 参数化及二阶段估计的思想同样适用。

顺便指出的是，非参数分量g的估计，一般都带有光滑参数λ．在实际使用时有个λ的选择问题，一个自然的途径是由数据本身来挑选，文献中已提出了一些方法，如交叉核实（CD）法、广义交叉核实（GCD）法．这些方法在非参数回归估计中，已证实是有效的，但对半参数情形尚少深入的研究．

二、函数空间的设定

为避免在叙述及记号上的复杂性，暂假定$q=1$，$\overline{U}$为R中有限区间$[a, b]$．对W的一个最为简单的设定是$W=L_2[a, b]$，即$[a, b]$上平方可积函数，这个空间包含了足够多的函数，但明显不足的是：依L_2相等的两个函数可在一个Lebesgue零集上不同，因而此空间中的函数并不是在$[a, b]$上处处确定的，这当然不适合于描述一个待估函数的目的．为此，W中的元还必须有某种光滑性．一些可能的选择是：取W为

$C[a, b]=\{g: g$连续$\}$

$C^m[a, b]=\{g: g^{(j)}$连续，$j=0, 1, \cdots, m\}$

$W_2^m[a, b]=\{g: g^{(i)}$连续，$i=0, 1, \cdots, m-1$，$g^{(m)}\in L_2[a, b]\}$

$P_m[a, b]=\{g: g$为次数$\leqslant m$的多项式$\}$

等等，今以$W_2^m[a, b]$说明前段的方法．设$g\in W_2^m[a, b]$由Taylor展式知存在$\theta_1, \cdots, \theta_m$使

$$g(t)=\sum_{i=1}^{m}\theta_i t^{i-1}+[(m-1)!]^{-1}\times\int_a^b g^{(m)}(u)(t-a)_+^{m-1}du$$

$$\hat{=}\sum_{i=1}^{m}\theta_i t^{i-1}+R_c(t) \qquad (1.2.1)$$

如余项$R_c(t)$关于$t\in[a, b]$一致地小，则可用多项式$\sum_{i=1}^{m}\theta_i$