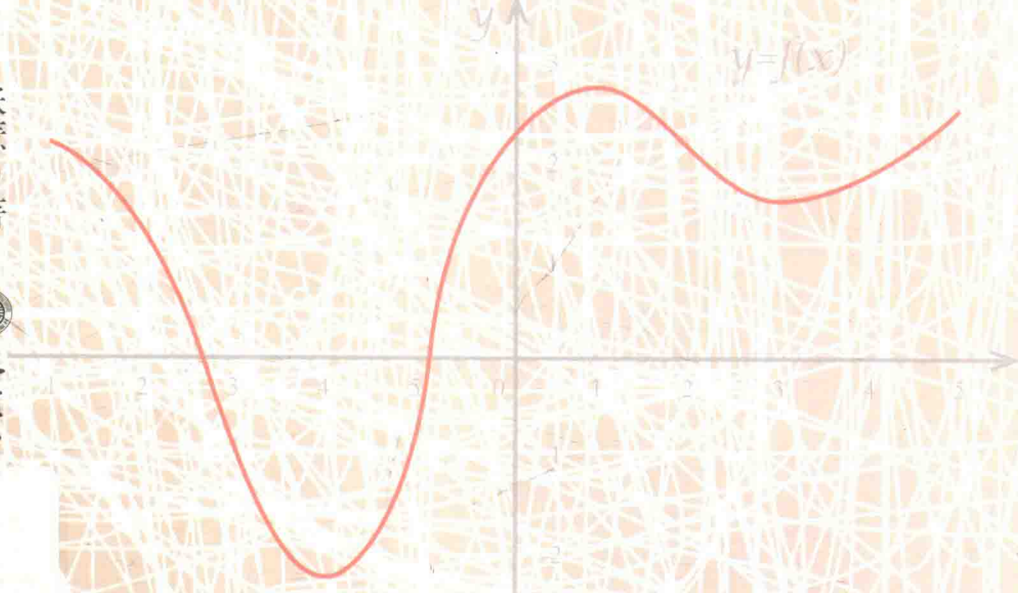
序伦财经文库

具有单调性的非参数函数估计方法

1928-2018



张薏 ⊙ 著

立信会计出版社

序伦财经文库

具有单调性的非参数函数估计方法

张薏⊙著

立信会计出版社

图书在版编目(CIP)数据

具有单调性的非参数函数估计方法/张薏著.
—上海：立信会计出版社，2018.4
(序伦财经文库)
ISBN 978-7-5429-5737-5

Ⅰ.①具… Ⅱ.①张… Ⅲ.①参数估计
Ⅳ.①O211.67

中国版本图书馆 CIP 数据核字(2018)第 048040 号

策划编辑 窦瀚修
责任编辑 王艳丽
封面设计 南房间

具有单调性的非参数函数估计方法

出版发行 立信会计出版社
地　　址 上海市中山西路 2230 号　　邮政编码 200235
电　　话 (021)64411389　　传　　真 (021)64411325
网　　址 www.lixinaph.com　　电子邮箱 lxaph@sh163.net
网上书店 www.shlx.net　　电　　话 (021)64411071
经　　销 各地新华书店

印　　刷 江苏凤凰数码印务有限公司
开　　本 710 毫米×1 000 毫米　1/16
印　　张 11.5　　插　　页 1
字　　数 151 千字
版　　次 2018 年 4 月第 1 版
印　　次 2018 年 4 月第 1 次
书　　号 ISBN 978-7-5429-5737-5/O
定　　价 39.00 元

前　　言

21 世纪以来，人类生活的每个领域都在不断积累和使用大量复杂多样的数据，从疾病诊断、流行病控制、生物医学研究、工农业生产、市场调查、宏观经济调控、天气预报、环境保护、公众交通管制、通讯导航、犯罪学到军事学，数据几乎无处不在. 大量不同类型的复杂数据的积累和应用给数据的储存、处理和分析提出了严峻的挑战. 借助于计算机技术的发展，统计学家开始利用各种统计方法为实际数据建立模型，并以此解释数据背后变量间的关系，诠释经济现象、判断并预测事物未来发展的方向.

统计建模具体包括：明确问题、收集信息、模型假设、模型构建、模型求解及分析等步骤. 一般来说，对研究的具体问题如果非常清楚，研究目标就会非常明确. 如果是以变量间的因果关系为研究目标，通常将原因和结果分别对应为解释变量和被解释变量，例如在研究影响居民消费的因素时，将反映居民消费水平的指标设计为被解释变量(结果)，对其有影响作用的相关因素设计为解释变量(原因). 解释变量往往直接由研究目的所决定，相比之下，解释变量的确定，需要研究者对相关领域有充足的认识，或是与相关领域的专家合作. 一旦确定好解释变量，就要收集和整理这些变量的观测数据，依据相关领域的理论与结果，进行适当的假设，并建立相关的回

归模型.

最早的回归模型是根据先验信息建立的参数回归模型,其特点在于模型的结构事先已知,包含有若干未知的参数.最经典的参数回归模型当属线性回归模型,回归系数的估计方法有最小二乘估计、极大似然估计和贝叶斯估计等.参数回归模型的最大不足在于高度依赖先验信息,当对信息掌握不足时,极易出现模型选择错误的问题,据此所做的预测分析也是不准确、不适宜的.非参数回归模型不需要事先指定回归函数的形式,它是根据实际的样本数据对未知函数做出拟合的,但是受样本的限制,研究结论的外延推广存在困难.另外,当解释变量的维数不断增加时,空间几何测度的增加速度会变得更快,估计模型所需的数据量也相应地呈指数级增长,发生人们常说的"维数灾难".所以非参数回归模型不适于解决高维问题.半参数回归模型将参数结构恰到好处地融入到了非参数回归模型中.一方面参数结构降低了高维数据带来的复杂性,另一方面非参数函数确保模型不失灵活性.常见的半参数回归模型有:部分线性模型、变系数模型、单指数模型等,它们在对同质性总体数据建模时表现出优异特征.但是实际数据也有许多是来自于非均匀的总体.当我们有理由相信回归数据是来自两个或两个以上不同种类且无法观测其来自哪个种类时,使用有限混合回归模型进行建模是比较合适的.在计量经济学文献中,混合回归模型常被称为转换回归模型,由Goldfeld和Quandt(1976)首先提出.混合回归模型所反映的现象可以概括为两类:不同子类中,解释变量与被解释变量间存在不同的回归关系;不同的子类中,回归关系相同但误差项分布不同.根据模型中混合比例、回归函数以及误差项分布等是否具有非

参数结构,混合回归模型也被归纳为参数、非参数及半参数类型.

无论是在一般的回归模型中,还是在混合回归模型中,未知函数的光滑性假设都是非常重要的.非参数光滑的设计是出于建模和估计两方面的考虑.如果要估计完整的非参数函数,则不可能得到无偏估计,因为所估计的参数个数有无穷个,远远多于已知样本量.但如果函数是光滑的,那么在靠近待估点附近的样本点包含了很多关于待估点的重要信息.因此,可以利用局部平均的手段构造非参数函数在待估点处函数值的估计,甚至可以获得较小的偏差和方差.光滑化的技术手段包括:核光滑、最小邻域、样条光滑、局部多项式、B样条逼近以及神经网络等.

在微观经济理论中,除光滑性外,单调性、凹性等也是时常赋予系统性功能所具有的一定属性.例如当生产者最大化收益时,产品的产出是否会随生产的投入单调不减.收入的增加是否会导致消费的增加,又或者价格按一定比例增长时是否收益也会按类似比例增长.生物医学中关于毒性水平的发病率模型、工业生产中温度对钢体的强度影响等也属于单调回归问题.教育与心理测量领域极具影响力的项目反应理论中,项目特征曲线也被认为是单调的.另外,诸如随机变量的累积分布函数、生存分析中的累积危险率函数、计数过程的均值函数、剂量反应模型中的回归函数等都被认为具有单调性.在某些实际问题中,特别在经济社会现象的解释上,对估计的单调性的保证尤为重要.此外,单调化的设计还可以去除边界附近常见的小范围摆动,也是稳定函数估计的一种有效手段.随着保序回归技术和Box-Cox变换的应用,单调化在数据分析方面扮演着重要的角色.Wright和Wegman(1980)总结了有关单调性和光滑性相结

合的观点并得出了具有一定价值的结论.

采用传统的估计方法(包括极大似然估计、最小二乘估计、贝叶斯估计、样条估计、核估计等)得到的函数估计都不具备单调性,但是通过构造组合、添加约束等手段可以得到单调性的估计.例如:Ramsay(1988)将单调回归样条的线性组合作为函数的估计,通过限制组合中各项系数非负来保证单调性.然而,组合系数非负只是具备单调性所需的充分而非必要条件.Mammen和Thomas-Agnan(1999)采取两步的方法估计单调回归函数.作者将单调性视为约束条件,根据光滑样条的方法得出无约束的估计,再将其投影约束集中,得到满足约束条件的估计.Hall和Huang(2001)提出了在核估计的基础上进行单调化估计的方法.主要是通过调整核估计中各项的概率分布使其在构成的核估计具有单调性的约束条件下与原概率分布的距离达到最小.Dette等(2006)利用严格单调的非参数函数$m(\cdot)$,构造随机变量$m(U)$,其中U为服从区间$[0, 1]$上均匀分布的随机变量.在非参数函数$m(\cdot)$的局部常数估计的基础上完成了$m(U)$的核密度估计以及累积分布函数的估计.根据累积分布函数的单调性以及反函数即为$m^{-1}(\cdot)$的特点,通过计算反函数得到了具有单调性的函数估计.

这些用来估计单调且光滑函数的方法目前只应用在一般的非参数回归模型中,尚未有文章对半参数回归模型、非参数混合回归模型中具有单调性的函数提出相应的估计方法.考虑到应用广泛的部分线性模型、变系数模型以及非参数的混合回归模型在近些年都有较为成熟的估计方法,本书以这些模型中具有单调性的非参数函数为研究对象,分别在这三个模型现有的局部常数估计、局部线性

估计以及局部似然估计的基础上，推广 Dette 等(2006)中对一元非参数回归模型所进行的单调化估计方法. 以下是各章节内容的简要介绍.

本书第 1 章从研究背景出发，根据总体的特征将回归模型分为两部分进行介绍. 其中一部分是针对同质性总体的一般回归模型，另一部分是针对异质性总体的混合回归模型. 除了各类模型的结构特征外，该章还重点介绍了①非参数函数常用的估计方法——局部光滑法以及样条估计法；②重要调节参数——窗宽的选择原理并列举了插入式和交叉验证等选择方法；③一般的非参数回归模型中单调函数的估计方法，混合模型、混合回归模型的估计方法以及 EM 算法等内容. 第 2 章讨论了部分线性模型产生的背景，国内外的研究水平. 对具有严格单调性的非参数函数，提出了保证单调性的估计方法. 从理论及应用角度对估计方法的合理性和有效性进行了验证. 第 3 章研究了应用更为广泛的变系数模型. 相比部分线性模型，变系数模型还能够反映解释变量之间的交互作用对回归变量的影响. 对比了模型常用的局部常数估计和局部线性估计，将适用于局部常数估计的单调化方法推广至局部线性估计上，所得估计量依然具备渐近正态分布的良好性质. 第 4 章的研究对象是一元非参数混合回归模型中具有单调性的非参数函数. 我们在回归函数的局部似然估计的基础上完成了单调化估计，并采用交叉确认准则的方法选择了局部似然估计时的最优窗宽. 在随机误差服从高斯分布的假设下，给出了单调化估计的大样本性质以及相关的正则条件. 第 5 章简要地概括了前面几章的研究内容，并指出了单调化估计方法的应用方向.

本书重点讲述的是对部分线性模型、变系数模型以及混合回归模型中具有单调性的一元回归函数的估计方法. 书中列举了非参数函数的传统估计方法,分析了估计原理及实现手段. 将具有单调性非参数函数的估计方法推广到了局部常数估计、局部线性估计及局部似然估计的基础上,促进了非参数技术的进一步发展,对具有形状约束的非参数函数的估计具有借鉴作用和指导意义.

本书在写作过程中得到了上海立信会计金融学院科研处和立信会计出版社的支持,上海财经大学统计与管理学院的王绍立老师也提出了不少中肯的意见. 本书的工作是在博士期间研究成果的基础上不断积累和完善的. 在此谨向对本书出版提供帮助的师长、同学和朋友表示衷心的感谢.

由于作者水平所限,书中难免有不足之处,尤其是在一些应用研究的体会性讨论中,恐有偏颇之处,恳切希望读者批评指正.

作者

2018 年 1 月 16 日

目　　录

1 绪　　论

社会经济发展过程中表现出来的经济现象反映着客观经济规律.要认识和掌握这些规律就需要了解经济现象间不同经济变量的变化规律以及相互关系.因为在一定的社会环境、地理条件、政府决策影响下,一些因素会推动或制约另外一些与之相联系的因素发生变化.当一个或一些变量的变化可以完全决定另一个变量的变化时,我们将它们之间的关系理解为具体的函数关系.由于经济问题的复杂性,还有许多因素会因为我们的认识不够以及其他客观原因的影响而没有被考虑在内,或者由于试验误差、测量误差等其他偶然因素的影响,使得其他变量的取值带有一定的随机性.我们将变量间这种不能由某一个或某一些变量唯一确定另一变量的关系,称为统计关系,主要通过回归分析来研究.

回归分析是建立在对客观事物进行大量试验和观察的基础上,为寻找隐藏在那些看上去是不确定的现象中的统计规律性而使用的统计方法.在回归函数和误差项的某些假定下,根据变量的观测值估计回归函数,并利用统计推断的方法对所建立的回归关系的有效性以及由此关系所揭示的变量间的特征进行分析,进一步应用于预测、控制等实际工作.一般来说,回归分析包含三层含义:对观测到的数据进行整理分析、建立回归模型;探索模型的估计方法,得到具有良好性质的估计;对模型进行诊断检验.回归分析的结果不仅能够用来解释数据背后的变量关系,还能用于生产生活中的预报和控制,为人类社会创造更大的财富.我们观测到的数据通常可以将其归为两类,一类是直接能够观测或进行人为调整的

数据，我们称其为解释变量；另一类是我们比较感兴趣的数据，我们将其称为被解释变量. 被解释变量依赖于解释变量，通过内在的规律随着解释变量发生变化. 不同变量间的关系有的简单有的复杂. 例如美国 30～39 岁女性的体重与身高呈现二次函数关系（来源：The World Almanac and Book of Facts，1975）. 简单的回归模型可以表示为

$$Y = m(\boldsymbol{X}) + \varepsilon \tag{1.1}$$

其中 $\boldsymbol{X}$ 为解释变量，Y 为被解释变量，$\boldsymbol{X}$ 与 Y 的内在关系通过回归函数 $m(\cdot)$ 表现出来. ε 为不可观测的随机误差项（根据其与回归函数间的运算关系，也被称为加性误差项），用于反映除解释变量外其他各种微小因素对被解释变量的影响. 误差项是由测试过程中诸多因素随机作用而形成的，虽然具有抵偿性，但不可完全消除. 给定解释变量下的误差项 ε 通常被设计为零均值的随机变量，故 $m(\boldsymbol{X})$ 可以表示为给定 $\boldsymbol{X}$ 下，Y 的条件期望 $E[Y|\boldsymbol{X}]$.

同质性（homogeneity）和异质性（heterogeneity）是科学和统计学中常用的有关物质或有机体均匀性的概念. 如果总体是同质的，那么可以视群体之间无差异，用一个回归方程进行拟合. 如果总体是异质的，则认为总体是由多个不同的子总体混合而成的，需要利用统计建模识别出不同的子总体，并针对不同的子总体建立不同的回归方程. 回归模型可以根据总体是否具有异质性特征，分为一般的回归模型和混合回归模型.

1.1 一般的回归模型

一般的回归模型是对同质性总体所建立的统计模型. 根据回归函数的结构形式是否完全已知，可以分为参数回归模型、非参数回归模型以及半参数回归模型.

1.1.1 参数回归

1.1.1.1 模型的结构及估计方法

最早的回归函数是由人们根据先验信息直接给出的，其中含有某些未知参数，相应的回归模型可以记作

$$\begin{aligned} Y &= m(X_1, X_2, \cdots, X_p; \theta_1, \cdots, \theta_d) + \varepsilon \\ &= m(\boldsymbol{X}; \boldsymbol{\theta}) + \varepsilon \end{aligned} \tag{1.2}$$

其中 $m(\cdot)$为已知函数，$\boldsymbol{\theta}=(\theta_1, \cdots, \theta_d)^{\mathrm{T}}$为未知参数. 我们将这些事先知道回归结构及结构中未知参数个数的模型称为参数回归模型. 若回归函数为线性函数，则模型(1.2)被称为线性回归模型(Linear Regression Model)，形如

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon \tag{1.3}$$

其中$\beta_0, \beta_1, \beta_2, \cdots, \beta_p$为未知参数，也被称为回归系数. 线性回归模型也时常用矩阵形式表示，

$$Y = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta} + \varepsilon \tag{1.4}$$

其中向量 $\boldsymbol{\beta}=(\beta_0, \beta_1, \beta_2, \cdots, \beta_p)^{\mathrm{T}}$, $\boldsymbol{X}=(1, X_1, X_2, \cdots, X_p)^{\mathrm{T}}$. 若回归函数为指数函数等非线性函数，模型(1.2)被称为非线性回归模型(Nonlinear Regression Model). 比较具有代表性的非线性回归模型是20世纪70年代由 J. A. Nelder 等建立的广义线性模型(Generalized Linear Model, GLM)

$$Y = g(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}) + \varepsilon \tag{1.5}$$

其中 $g(\cdot)$为已知的连接函数. 虽然非线性回归模型的回归函数不是线性的，但是有些可以通过对解释变量或被解释变量进行函数变换转化为线性回归模型. 经济学中著名的柯布-道格拉斯生产函数模型就是一个典型的例子：

$$Y = AK^{\alpha}L^{\beta}\mu$$

模型中Y代表工业总产值，A代表综合技术水平，L代表投入的劳动力数量，K代表投入的资本量，参数α，β及$\mu(\mu\leqslant 1)$分别代表劳动力产出、资本产出的弹性系数和随机干扰项（不同于模型(1.1)中的误差项，该式中的误差项被称为乘性误差项）. 通过取对数可将模型变形为

$$\ln Y=\ln A+\alpha\ln K+\beta\ln L+\ln\mu$$

对参数回归模型的研究主要集中在根据样本的观测值估计未知参数及误差项的方差，对模型及模型中有关参数的假设进行检验，根据最后确定的模型对事物进行控制和预报. 如果已知误差项的分布，可以使用极大似然方法去估计未知参数. 具体步骤如下：假设$\{(\boldsymbol{x}_i, y_i), i=1, \cdots, n\}$为样本观测值，误差项$\varepsilon$的密度函数为$f(v)$，根据(1.2)式中的回归关系可以建立似然函数

$$L(\boldsymbol{\theta}; \boldsymbol{x}, \boldsymbol{y})=\prod_{i=1}^{n}f(y_i-m(\boldsymbol{x}_i; \boldsymbol{\theta}))$$

对似然函数取对数得到对数似然函数

$$l(\boldsymbol{\theta}; \boldsymbol{x}, \boldsymbol{y})=\sum_{i=1}^{n}\log\{f(y_i-m(\boldsymbol{x}_i; \boldsymbol{\theta}))\}$$

参数空间中使对数似然函数达到最大的估计量被称为极大似然估计，即

$$\hat{\boldsymbol{\theta}}_{\mathrm{MLE}}=\arg\max_{\theta\in\Theta} l(\boldsymbol{\theta}; \boldsymbol{x}, \boldsymbol{y})$$

最小二乘法也是估计参数回归模型时常用的方法，主要思想是通过样本观测值与模型的回归值构造离差平方和，并通过选择参数使得离差平方和达到最小，即

$$\hat{\boldsymbol{\theta}}=\arg\min_{\theta}\sum_{i=1}^{n}(Y_i-m(\boldsymbol{X}_i; \boldsymbol{\theta}))^2$$

将估计结果代回模型中，略去误差项就得到了经验回归方程$\hat{Y}=m(\boldsymbol{X};\hat{\boldsymbol{\theta}})$及拟合值$\hat{Y}$. 将被解释变量的观测值$Y_i$和拟合值$\hat{Y}_i$之差称

为残差,记为$\hat{\varepsilon}_i = Y_i - \hat{Y}_i$. 由残差向量$\hat{\boldsymbol{\varepsilon}} = (\hat{\varepsilon}_1, \hat{\varepsilon}_2, \cdots, \hat{\varepsilon}_n)^{\mathrm{T}}$所构造的残差平方和$\sum_{i=1}^{n}\left(\hat{\varepsilon}_i - \frac{1}{n}\sum_{i=1}^{n}\hat{\varepsilon}_i\right)^2$就是方差$\sigma^2$的一个无偏估计. 以线性回归为例,根据(1.4)式中回归函数的矩阵形式,可以将离差平方和记为

$$Q(\boldsymbol{\beta}) = (Y - \boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta})^{\mathrm{T}}(Y - \boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta})$$

通过令$Q(\boldsymbol{\beta})$关于参数$\boldsymbol{\beta}$的偏导数为零,可得参数的最小二乘估计$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}})^{-1}\boldsymbol{X}Y$. 在得到参数估计后,还需进行参数的区间估计、显著性检验、回归方程的显著性检验等回归诊断,以确认回归方程是否能真正描述被解释变量与解释变量之间的统计规律性. 经过回归诊断所确认的回归模型便可用于预测和控制了.

1.1.1.2 重抽样方法

一般在获得参数估计后,还会通过计算估计量的偏差、方差来分析估计量的性质. 数值模拟过程中,可以根据已知的总体分布产生足够多的样本,但是在拟合实际数据时,往往只有一批样本,只能得到一个估计值. 因此,需要在原样本的基础上通过重抽样技术产生一批批样本,用于计算估计的偏差和方差.

常用的重抽样方法有 Jackknife 方法和 Bootstrap 方法. 其中,Jackknife 方法由 Quenodille(1949)提出,具体步骤是在原样本中每次删除一个或几个样本点,用剩下的样本和同样的估计量公式重新计算估计值,按照该方法进行逐个删除,便得到一系列估计值,利用一种类比关系即可求出偏差和方差的估计值. 关于 Jackknife 方法的具体内容可以参见 Miller(1974). Efron(1979)提出的 Bootstrap 方法在计算方式上与 Jackknife 类似,但是抽样方式不同.

假设相互独立的随机样本$X_1, \cdots, X_n$均来自于分布$F(x)$, $\theta = \theta(F)$为总体分布中的未知参数. 一方面,可以根据样本获得参数θ的估计量$\hat{\theta}_n$;另一方面,根据样本构造真实分布$F(x)$的估计——经验分布函数

$$F_n(x) = \frac{1}{n} \cdot \#\{X_i \mid X_i \leqslant x,\ i = 1, \cdots, n\}$$
$$= \frac{1}{n} \sum_{i=1}^{n} I\{X_i \leqslant x\} \tag{1.6}$$

其中 I 为示性函数. 可以证明该估计满足 $E(F_n(x)) = F(x)$, $\mathrm{Var}(F_n(x)) = \frac{1}{n}F(x)(1-F(x))$. 根据经验分布函数所反映的概率分布从原样本中抽取样本 $X_1^*, X_2^*, \cdots, X_n^*$ (Bootstrap 样本), 其实质为有放回的随机抽样. 利用 Bootstrap 样本重新计算参数 θ 的估计量 $\hat{\theta}_n^*$. 以总体均值为例, $\hat{\mu}_n = \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, $\hat{\mu}_n^* = \overline{X}^* = \frac{1}{n}\sum_{i=1}^{n} X_i^*$. 根据 $\hat{\theta}_n$ 及若干个 $\hat{\theta}_n^*$ 便可以计算估计量的偏差和方差. 通过估计量方差的估计还可以构造置信区间等. Bickel 和 Freedman(1981)证明了对于方差有限的总体来说, 样本均值的 Bootstrap 近似是渐近相合的. Singh(1981)进一步证明了在某些条件下 Bootstrap 近似比传统正态近似的收敛速度要快. 另外, Bootstrap 方法更适用于小样本数据.

对回归模型(1.4), 也可以利用 Bootstrap 方法对回归残差进行重抽样, 获得回归系数的近似分布. 具体步骤是:

(1) 根据样本估计回归系数 $\hat{\beta}$, 确定回归方程 $\hat{\boldsymbol{Y}} = \boldsymbol{X}^{\mathrm{T}}\hat{\boldsymbol{\beta}}$, 以及残差向量 $\hat{\boldsymbol{\varepsilon}} = (\hat{\varepsilon}_1, \hat{\varepsilon}_2, \cdots, \hat{\varepsilon}_n)^{\mathrm{T}}$;

(2) 考虑到误差项分布零均值的特征, 对残差进行中心化, 令

$$\tilde{\varepsilon}_i = \hat{\varepsilon}_i - \frac{1}{n}\sum_{i=1}^{n} \hat{\varepsilon}_i$$

(3) 对中心化的残差进行有放回的重抽样, 得到 $\boldsymbol{\varepsilon}^* = (\varepsilon_1^*, \varepsilon_2^*, \cdots, \varepsilon_n^*)^{\mathrm{T}}$;

(4) 将重抽样所得的残差向量与回归系数的估计值代入回归模型产生新的样本

$$Y_i^* = \boldsymbol{X}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}} + \varepsilon_i^*$$

(5) 根据新的样本 $\{(\boldsymbol{X}_i, Y_i^*): i = 1, \cdots, n\}$ 重新估计回归系数.