线性回归方法的相对有效性和估值漂移

葛永慧 著

科学出版社

线性回归方法的相对有效性和估值漂移

葛永慧 著

科学出版社

内 容 简 介

本书根据作者多年从事测量数据处理的教学与研究工作成果撰写而成。书中讨论和确定了常用稳健估计方法的相对有效性，以及总体最小二乘法与最小二乘法、稳健总体最小二乘法与稳健最小二乘法线性回归在不同误差模型影响下的相对有效性；提出了参数估计方法线性回归估值漂移的概念，讨论了最小二乘法和总体最小二乘法线性回归估值漂移的相关问题，建立了判定估值漂移的基本方法；讨论了一元线性回归自变量的优化、可线性化的一元非线性回归中直接观测值与间接观测值回归的差异和总体最小二乘法验后方差因子的实用性。

本书可供研究参数估计理论与方法的学者参考，也可作为高等院校测绘类相关专业研究生的参考用书，还可供参数估计领域的相关专业人员和工程技术人员参考。

图书在版编目（CIP）数据

线性回归方法的相对有效性和估值漂移 / 葛永慧著. —北京：科学出版社，2017.12

ISBN 978-7-03-056152-7

Ⅰ. ①线… Ⅱ. ①葛… Ⅲ. ①线性回归-应用-测量方法

Ⅳ. ①P204

中国版本图书馆 CIP 数据核字（2017）第 317905 号

责任编辑：裴 育 王 苏 / 责任校对：桂伟利

责任印制：张 伟 / 封面设计：蓝 正

科学出版社 出版

北京东黄城根北街 16 号

邮政编码：100717

http://www.sciencep.com

北京中石油彩色印刷有限责任公司 印刷

科学出版社发行 各地新华书店经销

*

2017 年 12 月第 一 版 开本：720×1000 B5

2017 年 12 月第一次印刷 印张：21 3/4

字数：426 000

定价：120.00 元

（如有印装质量问题，我社负责调换）

前　言

回归分析是建模和分析数据的重要工具。线性回归是利用数理统计中的回归分析确定两种或两种以上变量间相互依赖的定量关系的一种统计分析方法，运用十分广泛。按照自变量和因变量之间的关系类型，回归分析可分为线性回归分析和非线性回归分析。如果回归分析中只包括一个自变量和一个因变量且二者的关系可用一条直线近似表示，则称为一元线性回归分析。如果回归分析中包括两个或两个以上的自变量且因变量和自变量之间是线性关系，则称为多元线性回归分析。在进行非线性回归分析时，有些非线性回归方程可以通过适当的数学变换将非线性模型转化为线性模型进行求解，这类回归分析称为可线性化的非线性回归分析。线性回归求解的常用方法是最小二乘法、稳健最小二乘法、总体最小二乘法和稳健总体最小二乘法等。

本书讨论和确定了常用稳健估计方法的相对有效性。在粗差不可避免的情况下，用稳健估计方法能够有效地消除或减弱粗差对参数估计的影响。稳健估计方法的稳健特性取决于稳健估计方法自身、具体的参数估计问题及其观测值的数量等。对于不同观测值数量、不同粗差个数、不同母体随机误差的一元至五元线性回归，采用仿真实验的方法讨论和确定了常用稳健估计方法的相对有效性，以及当观测值中不包含粗差时常用稳健估计方法的精度损失。

本书讨论和确定了总体最小二乘法与最小二乘法、稳健总体最小二乘法与稳健最小二乘法线性回归在三种误差模型影响下的相对有效性。定义了三种误差模型：①因变量含有随机误差，自变量也含有随机误差；②因变量含有随机误差但自变量不含随机误差；③因变量不含随机误差但自变量含随机误差。对于不同观测值数量、不同母体随机误差的一元至五元线性回归，采用仿真实验的方法讨论和确定了总体最小二乘法与最小二乘法、稳健总体最小二乘法与稳健最小二乘法在不同误差模型影响下的相对有效性。对于最常用的一元线性回归，还讨论和确定了它们在不同斜率下的相对有效性。

本书提出了参数估计方法线性回归估值漂移的概念，讨论了最小二乘法和总体最小二乘法线性回归估值漂移的相关问题，建立了判定估值漂移的基本方法。用参数估计方法得到的参数估值显著偏离其真值的现象称为参数的估值漂移。线性回归通常采用相关系数或复相关系数和复判定系数来检验回归方程的拟合程度，相关系数或复相关系数和复判定系数越趋近于 1，说明因变量和自变量的线

性关系越密切，回归方程的有效性越好。然而，在实践中发现，最小二乘法或其他参数估计方法解算一元或多元线性回归系数时，即使相关系数或复相关系数和复判定系数都趋近于 1，也存在回归系数估值显著偏离其真值的现象。相对于仅用相关系数或复相关系数和复判定系数确定，增加回归系数估值漂移的确定，对线性回归参数特别是回归系数的有效性确定具有更高的可靠性。

本书讨论了一元线性回归自变量的优化、可线性化的一元非线性回归中直接观测值与间接观测值回归的差异和总体最小二乘法验后方差因子的实用性，介绍了各种线性回归方法的计算，提供了各种线性回归方法的具体算例。

本书根据作者多年从事测量数据处理的教学与研究工作成果撰写而成，基础理论是测量平差与稳健估计，而其应用不限于测绘类专业。本书既可作为高等院校测绘类相关专业研究生的参考用书，还可供参数估计领域的相关专业人员和广大工程技术人员参考。

在撰写本书过程中得到了许多同行和同事的热心帮助，他们提出了许多宝贵意见，高庚、董巧玲和刘清等硕士研究生对本书的插图和表格进行了整理与校对，在此深表感谢。同时，对本书参阅和引用的有关文献资料的作者表示真诚的感谢。

由于作者水平所限，书中难免存在不妥之处，敬请读者批评指正。

目　录

第1章 概 述

1.1 回归分析

“回归”这一概念是19世纪80年代由英国生物学家弗朗西斯·高尔顿(Francis Golton)在研究父代身高与子代身高时首先提出的。他发现子代身高有向族群平均身高“回归”的趋势。之后，高尔顿的学生卡尔·皮尔逊(Karl Pearson)把回归的概念同数学的方法联系起来，把代表现象之间一般数量关系的统计模型称为回归直线或回归曲线，从此诞生了统计学上著名的回归理论[1]。现代统计学的“回归”概念已不是原来生物学的特殊规律性，而是指变量之间的依存关系。

如今，回归分析已经成为社会科学定量研究方法中最基本、应用最广泛的一种数据分析技术。它既可以用于探索和检验自变量与因变量之间的因果关系，也可以基于自变量的取值变化来预测因变量的取值，还可以用于描述自变量和因变量之间的关系[2]。概括地说，回归分析是研究自然界变量之间存在的非确定性的相互依赖和制约关系，并把这种关系用数学表达式表达出来的一种方法。其目的是利用这些数学表达式以及对这些表达式的精确估计，对未知变量做出预测或检验其变化，为决策服务[3]。

较早的、比较成熟的回归模型是经典回归模型，它包括线性回归模型和非线性回归模型。其中，线性回归模型是最基本、最简单的回归形式。19世纪初，高斯首先提出了在线性关系下的回归方程的最小二乘法，这可以说是回归分析的起点，在实际应用中发挥了很重要的作用。然而，在实际应用中，严格符合线性回归模型规律的问题并不多见，虽然大多数问题可近似为线性回归模型，但在不少情况下，用非线性回归模型可能更加符合实际。从回归分析的发展来看，非线性回归模型是线性模型的自然推广，目前已发展成为近代回归分析的一个重要研究分支[4]。

从方法论的角度看，回归分析主要研究回归模型的参数估计、假设检验、模型选择等理论和有关计算方法。其一般步骤是，首先根据理论和对问题的分析判断，将变量分为自变量和因变量；其次设法找出合适的数学方程(即回归模型)描述变量间的关系；由于涉及的变量具有不确定性，接着还要对回归模型进行统计检验；统计检验通过后，最后是利用回归模型，根据自变量估计、预测因变量。

1.2　回归分析的分类

回归分析是一种最基础、最重要的统计分析方法，在建立实验模型和理论模型的检验系统中，回归分析起着不可或缺的作用。在统计学中，回归分析包括进行建模和分析几个变量的任何技术，其焦点在于一个因变量和一个或多个自变量之间的关系。更具体地说，回归分析有助于人们了解当任一自变量变化而其余自变量保持不变时，因变量典型值的变化情况[5]。

回归分析有不同的种类，按照自变量和因变量之间的关系类型，即回归模型的形式，回归分析可以分为线性回归分析和非线性回归分析；按照回归模型涉及的自变量数目，回归分析可以分为一元回归分析和多元回归分析。如果在回归分析中只包括一个自变量和一个因变量，且二者的关系可用一条直线近似表示，则称为一元线性回归分析；如果回归分析中包含两个或两个以上的自变量，且因变量和自变量之间是线性关系，则称为多元线性回归分析。这样，按自变量数目和回归模型的形式，回归分析的分类如表 1.1 所示[4]。

表 1.1　回归分析的分类

回归模型形式	变量的数目	回归类型
线性回归	一个因变量，一个自变量 一个因变量，多个自变量	一元线性回归 多元线性回归
非线性回归	一个因变量，一个自变量 一个因变量，多个自变量	一元非线性回归 多元非线性回归

实际分析时，应根据客观现象的性质、特点、研究目的和任务选取回归分析的方法。

1.3　两种参数估计方法的比较[6-9]

1.3.1　两种参数估计方法比较的指标

1. 绝对指标——残余真误差均方误差

定义 1-1　观测值的真误差与观测值的改正数之和，即观测值的估值与观测值的真值之差称为观测值估值的残余真误差，简称残余真误差(RTE)：

$$f_k = \Delta_k + V_k = \hat{L}_k - \tilde{L}_k, \quad k = 1,2,\cdots,n \tag{1-1}$$

式中，Δ_k 为观测值 L_k 的真误差；V_k 为通过参数估计方法得到的观测值 L_k 的改正数；$\tilde{L}_k$ 为观测值 L_k 的真值；$\hat{L}_k$ 为通过参数估计方法得到的观测值 L_k 的估值。$L_k = \tilde{L}_k + \Delta_k$，$\hat{L}_k = L_k + V_k$。$f_k$ 为观测值估值 $\hat{L}_k$ 相对于观测值真值 $\tilde{L}_k$ 的真误差，称为残余真误差。

定义 1-2　残余真误差均方误差(MSRTE)：

$$\hat{\sigma}_f = \sqrt{\frac{1}{n}\sum_{k=1}^{n} f_k^2} \tag{1-2}$$

式中，$\hat{\sigma}_f$ 称为残余真误差均方误差，它是观测值估值的标准差，能从实质上说明参数估计方法的有效性。为了从统计上说明参数估计方法的有效性，$\hat{\sigma}_f$ 通常取同一个参数估计方法对同一个参数估计问题的多次(如 1000 次)仿真实验的平均值，仍然称为残余真误差均方误差。

2. 相对指标——相对增益

定义 1-3　设有两种参数估计方法 A 和 B，它们对于同一个参数估计问题得到的残余真误差均方误差分别为 $\hat{\sigma}_{fa}$ 和 $\hat{\sigma}_{fb}$。B 方法相对于 A 方法的残余真误差均方误差比(简称均方误差比)为

$$\mathrm{RR} = \frac{\hat{\sigma}_{fb}}{\hat{\sigma}_{fa}} \tag{1-3}$$

当 RR>1 时，A 方法优于 B 方法；当 RR<1 时，B 方法优于 A 方法；当 RR=1(或接近于 1)时，A 和 B 两种方法等价。

定义 1-4　设有两种参数估计方法 A 和 B，它们对于同一个参数估计问题得到的残余真误差均方误差分别为 $\hat{\sigma}_{fa}$ 和 $\hat{\sigma}_{fb}$，B 方法相对于 A 方法的相对增益为

$$\mathrm{RG} = \frac{\hat{\sigma}_{fa} - \hat{\sigma}_{fb}}{\hat{\sigma}_{fa}} \times 100\% \tag{1-4}$$

当 RG>0 时，B 方法优于 A 方法；当 RG<0 时，A 方法优于 B 方法；当 RG=0(或接近于 0)时，A 和 B 两种方法等价。

RR 和 RG 能说明两种参数估计方法哪种更有效。为了从统计上说明两种参数估计方法哪种更有效，$\hat{\sigma}_{fa}$ 和 $\hat{\sigma}_{fb}$ 通常取 A 和 B 两种参数估计方法对同一个参数估计问题的多次(如 1000 次)仿真实验的平均值，仍然分别将 RR 和 RG 称为 B 方法相对于 A 方法的残余真误差均方误差比和相对增益。

1.3.2　仿真实验方法

设 $i=1,2,\cdots,S$，S 表示仿真实验的次(组)数；$j=1,2,\cdots,n$，n 表示观测值的数量。

用 $\tilde{L}_j$ 表示观测值的真值；用 δ_{ij} 表示服从正态分布 $N(0,\sigma_0^2)$ 的随机误差[10]，由随机误差模拟函数生成。

(1) 观测值中不包含粗差时：

$$\Delta_{ij}=\delta_{ij},\quad i=1,2,\cdots,S;\quad j=1,2,\cdots,n \tag{1-5}$$

(2) 观测值中包含粗差时：

$$\Delta_{ij}=\begin{cases}\varepsilon, & \theta_{ij}=1\\ \delta_{ij}, & \theta_{ij}=0\end{cases},\quad i=1,2,\cdots,S;\quad j=1,2,\cdots,n \tag{1-6}$$

式中，θ_{ij} 表示随机误差 δ_{ij} 是否被粗差 ε 代替，每一组 θ_{ij} ($j=1,2,\cdots,n$)由 g 个 1 和 $(n-g)$ 个 0 构成，由随机函数生成。对于每一组随机误差 δ_{ij} ($j=1,2,\cdots,n$)，当 $\theta_{ij}=1$ 时，随机误差 δ_{ij} 用粗差 ε 代替，生成 S 组同时包含 g 个粗差的随机误差 Δ_{ij}。

用观测值的真值 $\tilde{L}_j$ 加上对应的 S 组随机误差得到 S 组模拟观测值 L_{ij}：

$$L_{ij}=\tilde{L}_j+\Delta_{ij},\quad i=1,2,\cdots,S;\quad j=1,2,\cdots,n \tag{1-7}$$

对于 S 组模拟观测值中的每一组，用参数估计方法计算观测值的估值 $\hat{L}_{ij}$ 和改正数 V_{ij}，进而计算残余真误差均方误差。用 S 组残余真误差均方误差的平均值作为该参数估计方法在观测值中的残余真误差均方误差。用同样的方法计算不同参数估计方法的残余真误差均方误差。用不同参数估计方法得到的残余真误差均方误差可以计算它们相互之间的相对增益。

在仿真实验中，σ_0=1.0 或 σ_0=3.0，随机误差 $|\delta_{ij}|\leqslant 2.5\sigma_0$，仿真实验的次数 $S=1000$，粗差 ε 的取值为 $0.0\sigma_0$、$5.0\sigma_0$ 和 $10.0\sigma_0$。当需要迭代计算时，终止条件是相邻两次观测值改正数差值的绝对值均小于 0.1。

第 2 章　一元线性回归

2.1　一元线性回归模型的建立

假设某一自变量 x 与某一因变量 y 之间呈线性相关关系，通过 n 组观测值得到一组数据为 (y_i, x_i)，其中 $i=1,2,\cdots,n$。假定一元线性回归模型结构为[11]

$$y_i = b_0 + b_1 x_i + \varepsilon_i$$

式中，b_0、b_1 为待定参数；$i=1,2,\cdots,n$ 表示观测值的个数；ε_i 为随机误差项。参数 b_0、b_1 一般是未知的，需根据 y_i 与 x_i 的观测值采用最小二乘法(least square method, LS 法)估计得到。设 β_0 和 β_1 分别为参数 b_0 和 b_1 的 LS 估值，可得一元线性回归模型为

$$\hat{y}_i = \beta_0 + \beta_1 x_i$$

式中，β_0 为常数；β_1 为回归系数；$i=1,2,\cdots,n$ 表示观测值的个数。

对回归模型进行回归分析时，通常有三个基本假定，结合一元线性回归模型进行具体说明[4]：

(1) 误差项 ε_i 是一个期望值为 0 的随机变量，即 $E(\varepsilon_i)$=0。这意味着回归模型中，b_0 和 b_1 都是常数，所以有 $E(b_0)=b_0$，$E(b_1)=b_1$。因此，对于一个给定的 x 值，y 的期望为 $E(y)=b_0+b_1x$。

(2) 对于所有的 x 值(即 x_i，$i=1,2,\cdots,n$)，误差项 ε_i 的方差 σ^2 都相同。

(3) 误差项 ε_i 是一个服从正态分布的随机变量，且相互独立，即 $\varepsilon_i \sim N(0,\sigma^2)$。独立性意味着对于一个特定的 x 值，它所对应的 y 值与其他 x 所对应的 y 值也不相关。

2.2　一元线性回归方程的通解

在线性回归中，通过将参数的线性组合作为因变量来确立模型。一元线性回归有一个自变量和两个回归系数(参数)。

设一元线性回归方程的一般形式[5]为

$$\hat{y} = \hat{a}z + \hat{b}x \tag{2-1}$$

数学模型是

$$y_i + v_i = z_i\hat{a} + x_i\hat{b}, \quad i = 1,2,\cdots,n \tag{2-2}$$

$$v_i = z_i\hat{a} + x_i\hat{b} - y_i, \quad i = 1,2,\cdots,n \tag{2-3}$$

$$\begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix} = \begin{bmatrix} z_1 & x_1 \\ z_2 & x_2 \\ \vdots & \vdots \\ z_n & x_n \end{bmatrix} \begin{bmatrix} \hat{a} \\ \hat{b} \end{bmatrix} - \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \tag{2-4}$$

式中，$\hat{a}$ 和 $\hat{b}$ 是回归系数；y_i 是观测值(因变量)；x_i 相当于自变量；z_i 是 $\hat{a}$ 的系数(通常为 1)；$v_i = \hat{y}_i - y_i$ 是观测值 y_i 的残差，$\hat{y}_i$ 是观测值 y_i 的估计值；n 表示观测值的数量。

由 LS 法得一元线性回归的法方程为

$$\begin{bmatrix} \sum(z_i z_i) & \sum(z_i x_i) \\ \sum(x_i z_i) & \sum(x_i x_i) \end{bmatrix} \begin{bmatrix} \hat{a} \\ \hat{b} \end{bmatrix} - \begin{bmatrix} \sum(z_i y_i) \\ \sum(x_i y_i) \end{bmatrix} = 0 \tag{2-5}$$

回归系数 $\hat{a}$ 和 $\hat{b}$ 的解为

$$\begin{bmatrix} \hat{a} \\ \hat{b} \end{bmatrix} = \begin{bmatrix} \sum(z_i z_i) & \sum(z_i x_i) \\ \sum(x_i z_i) & \sum(x_i x_i) \end{bmatrix}^{-1} \begin{bmatrix} \sum(z_i y_i) \\ \sum(x_i y_i) \end{bmatrix} \tag{2-6}$$

$$\begin{cases} \hat{a} = \dfrac{\sum x_i^2 \sum(y_i z_i) - \sum(x_i z_i)\sum(x_i y_i)}{\sum z_i^2 \sum x_i^2 - \sum(x_i z_i)\sum(x_i z_i)} \\ \hat{b} = \dfrac{\sum z_i^2 \sum(x_i y_i) - \sum(x_i z_i)\sum(y_i z_i)}{\sum z_i^2 \sum x_i^2 - \sum(x_i z_i)\sum(x_i z_i)} \end{cases} \tag{2-7}$$

在式(2-7)中，用 $\ln\hat{a}$ 替代 $\hat{a}$，得

$$\begin{cases} \hat{a} = \exp\left\{ \dfrac{\sum x_i^2 \sum(y_i z_i) - \sum(x_i z_i)\sum(x_i y_i)}{\sum z_i^2 \sum x_i^2 - \sum(x_i z_i)\sum(x_i z_i)} \right\} \\ \hat{b} = \dfrac{\sum z_i^2 \sum(x_i y_i) - \sum(x_i z_i)\sum(y_i z_i)}{\sum z_i^2 \sum x_i^2 - \sum(x_i z_i)\sum(x_i z_i)} \end{cases} \tag{2-8}$$

式(2-7)和式(2-8)是一元线性回归方程(2-1)的回归系数解的一般形式。根据一元线性回归方程的一般形式，可直接写出不同回归模型的一元线性回归方程和可转换成一元线性回归的非线性回归方程的解。

2.3　一元线性回归方程的拟合效果度量

回归方程在一定程度上描述了变量 Y 与 X 之间的内在规律。根据回归方程，

可由自变量 X 的取值来估计因变量 Y 的取值。但其估计的精度如何将取决于回归方程的拟合程度。分析一元线性回归方程的拟合程度时，最常用的指标是相关系数和判定系数。

2.3.1　相关系数

相关系数是测定变量之间关系密切程度的一个统计量，它能够通过定量的方式准确地描述变量之间的相关程度。相关系数有多种，对于不同类型的变量数据，应计算不同的相关系数。

皮尔逊简单相关系数(以下简称相关系数)是常用的相关系数之一，它主要是用来度量两个变量 x 与 y 之间的线性相关程度，如人均可支配收入与消费支出的相关程度、身高与体重之间的相关程度等。一般用 r 来表示。

设 $(x_i, y_i)(i=1,2,\cdots,n)$ 是 (x,y) 的 n 组观测值，相关系数的定义公式是

$$r=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})/n}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2/n}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2/n}}$$

上式可简化为

$$r=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2(y_i-\bar{y})^2}}=\frac{n\sum_{i=1}^{n}x_iy_i-\sum_{i=1}^{n}x_i\sum_{i=1}^{n}y_i}{\sqrt{n\sum_{i=1}^{n}x_i^2-(\sum_{i=1}^{n}x_i)^2}\sqrt{n\sum_{i=1}^{n}y_i^2-(\sum_{i=1}^{n}y_i)^2}}$$

式中，n 表示观测值数量；x 为自变量；y 为因变量。

相关系数的性质与具体含义理解如下[4]。

(1) r 的取值范围是 $[-1,+1]$，即 $-1\leqslant r\leqslant 1$。当 $|r|=1$ 时，表现为完全相关；当 $r=0$ 时，表现为无线性相关或完全不相关，但两个变量之间有可能存在非线性相关；当 $0<|r|<1$ 时，表现为不完全相关。

(2) $r>0$ 表明两个变量之间存在正线性相关关系；$r<0$ 表明两个变量之间存在负线性相关关系。

(3) r 具有对称性。x 与 y 之间的相关系数和 y 与 x 之间的相关系数相等。

(4) r 的数值与 x 和 y 的计量单位无关，改变 x 和 y 的计量单位，并不影响 r 的数值。

(5) r 是两个变量之间线性关系的度量指标，但无法反映两个变量之间的因果关系，即使 $|r|$ 接近于 1.0，也不一定意味着 x 与 y 之间一定存在着因果关系。

值得注意的是，相关系数是反映两个变量的线性相关程度，但它并不能够度