多水平模型及其在经济领域中的应用

石 磊　向其凤　陈 飞　著

科学出版社

多水平模型及其在经济领域中的应用

石 磊 向其凤 陈 飞 著

科学出版社

北京

内 容 简 介

本书系统介绍多水平模型理论、方法以及在经济分析中的应用. 内容主要介绍多水平线性模型、多水平广义线性模型的理论和方法，包括模型定义、参数估计、模型检验等. 将多水平模型应用于宏观及微观经济数据的分析，提出多水平生长函数模型、多水平发展模型、多水平面板数据模型、多水平的因素分析模型等；结合实际经济问题，介绍如何使用多水平模型对微观经济、金融数据进行统计建模，为研究具有层次结构的经济数据提供了新的分析工具. 本书还介绍如何使用 MlwiN、SAS、Stata 软件计算多水平模型参数估计和检验统计量，并通过实例进行分析.

本书可作为统计学专业本科生、研究生的教材和参考书，也可作为经济学、管理学、社会学、生物医学等领域的研究人员和相关科技工作者的参考书.

图书在版编目(CIP)数据

多水平模型及其在经济领域中的应用/石磊，向其凤，陈飞著. —北京：科学出版社，2013.8

ISBN 978-7-03-038195-8

Ⅰ. ①多… Ⅱ. ①石… ②向… ③陈… Ⅲ. ①统计模型—研究 Ⅳ. ①C8

中国版本图书馆 CIP 数据核字 (2013) 第 168946 号

责任编辑：王丽平 / 责任校对：宣 慧
责任印制：钱玉芬 / 封面设计：陈 敬

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

北京彩虹伟业印刷厂 印刷
科学出版社发行 各地新华书店经销
*
2013 年 8 月第 一 版 开本：B5(720 × 1000)
2013 年 8 月第一次印刷 印张：16 1/4
字数：333 000

定价：79.00 元

(如有印装质量问题，我社负责调换)

前　言

在许多社会经济领域, 我们得到的数据都具有层次 (Multilevel or Hierarchical) 或聚类 (Aggregation or Clustering) 结构. 例如, 在教育学研究中, 学生嵌套于班级中, 班级又嵌套于学校中, 形成了一个具有三个层次 (学生 — 班级 — 学校) 的数据结构. 在分析地区经济发展的研究中, 城市嵌套于省, 省嵌套于国家或地区, 城市的经济发展与时间有关 (时间层), 也形成了一个多个层次的数据结构. 这种分层数据常见于人口普查、经济普查、抽样调查及跨地区跨文化的研究中. 对层次或聚类数据, 同一层次的数据具有较高的相似性, 不同层次的数据具有较强的差异性 (或异质性), 传统的最小二乘 (OLS) 估计的假设 (如同方差及独立性) 不再适合这类数据的要求, 使用传统的线性回归方法将会产生较大的估计误差, 并得出不正确的推断结果. 多水平模型 (Multilevel Model) 或多层次模型 (Hierarchical Model) 正是基于这样的结构数据发展而来的 (Longford, 1993; Goldstein, 1995; Randenbush, 1999; 石磊, 2008). 多层线性模型考虑了数据的层次结构和聚集结构特征, 能准确地反映变量间基于层次框架下的关系, 并给出不同层次数据的差异性估计及跨级相关估计. 多水平模型可以允许回归参数具有随机性, 能处理一些较为复杂的多层次结构数据, 进而减小估计的误差. 近年来, 多水平模型得到了广泛的应用和发展, 如提出了多水平非线性模型、多水平时间序列模型、多水平测量误差模型、多水平离散数据模型等 (Goldstein, 1995), 并研发了相应的计算软件 (Goldstein et al., 1998). 在教育学、社会学、卫生及心理学领域, 多水平模型的应用极为广泛 (张雷 等, 2002; 杨珉 等, 2007; 王济川 等, 2008).

在经济研究领域, 多水平模型的应用和发展还不多. 在国外, 利用多水平模型, Grieve 等 (2005) 研究了多国资源使用和成本的变化; Grieve 等 (2007) 研究了多国净收益增量的估计; Shu 等 (2007) 研究了全球化对中国城市劳动力市场性别差异的影响; Hansen (2007) 提出了经济计量模型 DD 估计 (Difference-in-Difference Estimation) 的多水平模型估计理论. 从这些文献可看出, 多水平模型的应用和研究已开始向经济学领域发展. 但在国内, 多水平模型在经济分析中的研究和应用未见相关文献报道.

在经济领域, 层次结构数据是非常普遍的, 如在许多普查和抽样调查数据, 分层结构很明显. 在许多研究经济发展的结构变化中, 也会出现层次结构数据. 在许多情况下, 套用传统的 OLS 估计是不恰当的. 利用多水平模型, 不仅可以给出合理的预测模型和估计精度, 同时能有效地研究层次结构数据某些影响因素的影响特征. 对经济问题的研究, 不同的模型可能会产生截然不同的结论, 合理地选择和使

用模型是一项重要的工作. 因此, 本书的研究为具有层次结构的经济数据的统计建模提供了重要的研究工具, 对经济问题的计量方法研究具有广泛的应用价值.

本书共分 10 章. 第 1 章是预备知识, 介绍多水平模型研究中涉及的一些基础知识, 包括线性模型、混合线性模型、广义线性模型、广义线性混合模型的理论和方法. 第 2 章介绍多水平模型理论、方法和统计诊断. 第 3 章介绍多水平面板数据模型及其估计理论. 第 4 章利用多水平模型, 构建西部民族地区农民收入增长的多水平发展模型, 研究影响农户人均收入及其增长的影响因素. 第 5 章基于我国 210 个地级及其以上城市 1990 ～ 2007 年的经济发展数据, 采用多水平模型, 从两个不同的层次角度对我国区域经济增长收敛性特征进行分析研究, 并从理论及实证两方面说明对我国区域经济增长分析采用多水平模型分析的必要性. 系统讨论在层次结构数据分析中, 我国区域经济增长中各省区或城市发展水平等级区域内部收敛及区域之间的异质性问题. 第 6 章在 C-D 生产函数的基础上, 引入多水平线性模型, 通过对模型进行比较, 提出多水平线性模型在研究宏观经济收入问题的研究角度及方法, 并建立多水平 C-D 生产函数模型, 分别测算各要素对中国经济增长的贡献份额, 从而进行中国经济增长的源泉分析. 第 7 章介绍我国产业结构与经济增长和要素效率关系分析. 第 8 章基于层次结构的分析角度, 以地区经济增长与上市公司发展的规律性分析为例, 利用 1997 ～ 2007 年我国的各省份数据, 建立多水平模型与静态面板数据模型进行实证分析. 第 9 章从农村住户的调查数据出发, 以劳动力迁移理论为基础, 构建分层结构数据的多水平 Logistic 模型, 分析各种制约因素对西部民族地区劳动力转移的影响, 并针对西部民族地区的特殊情况提出相关政策建议. 这一章基于云南省红河哈尼族彝族自治州 13 个县 (市), 298 个行政村, 2985 户农户的微观数据集, 设计一个两水平农户收入函数模型, 实证分析地理因素对农户收入影响. 另外, 利用 1979 ～ 2007 年全国的经济数据, 引入多水平模型, 对我国产业结构与经济增长、要素效率分别进行实证分析. 第 10 章详细介绍基于 SAS、MLwiN、Stata 软件的使用方法, 同时给出了作者开发的基于 Matlab 软件平台的计算程序. 在第 3 章至第 9 章的实证研究中, 每一章都基于模型研究结果给出相关政策建议.

本书得到国家社科基金 (项目编号：08XTJ001), 国家自然科学基金 (项目编号：11161053), 国家自然科学基金数学天元基金 (项目编号：11126297) 的资助, 在此表示感谢. 本书由课题组成员共同完成, 课题负责人指导的研究生王焕英、程海生、张洵、王尚坤参与了课题的研究, 完成了本书的部分内容, 在此一并致谢. 由于作者水平有限, 疏漏和不足之处在所难免, 恳请专家和同行批评指正.

作　者

2012 年 12 月 28 日

目　录

第1章 预备知识

1.1 线性模型理论

在现实问题的研究中, 所研究的对象之间的关系往往是我们感兴趣的问题. 具体而言, 通常希望了解哪些因素在影响着我们所研究的对象, 影响的方式又是怎样的. 如果用变量来表征研究对象和因素, 我们要了解的就是某个变量与其他某些变量之间的关系. 由于随机因素的存在, 变量之间的关系未必总是确定的函数关系, 而很可能仅表现为某种趋势. 故我们面对的问题就是怎样从数据出发, 获得关于变量间关系的推断. 如果先验理论或数据基本特征显示, 变量间关系的基本趋势呈线性, 那么线性假设往往可以作为有关推断的出发点. 以变量间的线性关系为基本假定和出发点的统计推断理论就是我们要介绍的线性模型理论.

1.1.1 最小二乘估计

1.1.1.1 线性模型简介

首先介绍线性模型的基本形式和相关假设. 变量间的关系如下:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \varepsilon \tag{1.1.1}$$

其中, y 称为因变量或被解释变量; $x_1, \cdots, x_k$ 称为自变量或解释变量; $\beta_0, \beta_1, \cdots, \beta_k$ 为参数; ε 为随机误差项. 设来自于该模型的样本为 $(x_{i1}, \cdots, x_{ik}, y_i), i = 1, \cdots, n$, 则模型的样本形式为

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i, \quad i = 1, \cdots, n \tag{1.1.2}$$

为简洁起见, 可将该模型表示为矩阵形式:

$$Y = X\beta + \varepsilon \tag{1.1.3}$$

其中, $Y = (y_1, \cdots, y_n)', X = (\mathbf{1}, X_1, \cdots, X_k), X_j = (x_{1j}, \cdots, x_{nj})', j = 1, \cdots, k, \mathbf{1} = (1, 1, \cdots, 1)', \beta = (\beta_0, \beta_1, \cdots, \beta_k)', \varepsilon = (\varepsilon_1, \cdots, \varepsilon_n)'$. 此外通常作如下假设:

(I) 自变量 $x_1, \cdots, x_k$ 均为非随机变量, 即 X 是非随机矩阵;

(II) 变量 $x_1, \cdots, x_k$ 相互之间线性无关, 即设计矩阵 X 是列满秩矩阵;

(III) $EY = X\beta$ 和 $\mathrm{Cov}(Y) = \sigma^2 I$, 其中 I 为单位矩阵. 这个假设称为高斯–马尔可夫 (Gauss-Markov) 条件.

1.1.1.2　最小二乘估计的定义及表达式

模型 (1.1.3) 中, β 为未知参数向量, 获取 β 的估计是该模型下统计推断的基本任务之一. 下面介绍最小二乘法. 称

$$Q(\beta)\triangleq\|Y-X\beta\|^2=(Y-X\beta)'(Y-X\beta) \tag{1.1.4}$$

的最小值点为线性模型 (1.1.3) 下 β 的最小二乘估计, 记作 $\hat{\beta}_{\mathrm{LS}}$.

注意到 $Q(\beta)=\|Y-X\beta\|^2$, 故有 $X\hat{\beta}_{\mathrm{LS}}$ 是 Y 在 X 的列空间 $L(X)$ 上的投影. 这意味着 $Y-X\hat{\beta}_{\mathrm{LS}}$ 应与 X 的所有列向量正交, 故有 $X'(Y-X\hat{\beta}_{\mathrm{LS}})=0$, 即

$$X'X\hat{\beta}_{\mathrm{LS}}=X'Y$$

上面方程称为正规方程组. 当 X 是列满秩时, 方程有唯一解, 即

$$\hat{\beta}_{\mathrm{LS}}=(X'X)^{-1}X'Y \tag{1.1.5}$$

上述对线性模型的拟合过程称为线性回归.

1.1.1.3　参数估计的性质

下面介绍最小二乘估计的性质. 若无特别申明, 下面所述性质均在基本假设 (I), (II), (III) 下获得.

(1) $\hat{\beta}_{\mathrm{LS}}$ 是 β 的线性无偏估计.

显然, $\hat{\beta}_{\mathrm{LS}}$ 是 Y 的线性函数, 而 $\hat{\beta}_{\mathrm{LS}}$ 的期望

$$E(\hat{\beta}_{\mathrm{LS}})=(X'X)^{-1}X'EY=(X'X)^{-1}X'X\beta=\beta$$

(2) $\mathrm{Cov}(\hat{\beta}_{\mathrm{LS}},\hat{\beta}_{\mathrm{LS}})=\sigma^2(X'X)^{-1}$.

事实上, $\mathrm{Cov}(\hat{\beta}_{\mathrm{LS}},\hat{\beta}_{\mathrm{LS}})=(X'X)^{-1}X'\mathrm{Cov}(Y,Y)X(X'X)^{-1}=\sigma^2(X'X)^{-1}$.

(3) $\hat{\sigma}^2\triangleq SSE/(n-k-1)$ 是 σ^2 的无偏估计, 其中 $SSE\triangleq\left\|Y-X\hat{\beta}_{\mathrm{LS}}\right\|^2=(Y-X\hat{\beta}_{\mathrm{LS}})'(Y-X\hat{\beta_{\mathrm{LS}}})$ 称为残差平方和.

易证

$$\begin{aligned}E\hat{\sigma}^2=&\frac{1}{n-k-1}E\{Y'[I_n-X(X'X)^{-1}X'][I_n-X(X'X)^{-1}X']Y\}\\=&\frac{1}{n-k-1}\mathrm{tr}\{[I_n-X(X'X)^{-1}X']E(YY')\}\\=&\frac{\sigma^2}{n-k-1}\mathrm{tr}\{I_n-X(X'X)^{-1}X'\}\\=&\frac{\sigma^2}{n-k-1}\{n-\mathrm{tr}[(X'X)^{-1}X'X]\}\end{aligned}$$

$$=\sigma^2$$

(4) 对于任意的 $k+1$ 维实向量 $\alpha, \alpha'\hat{\beta}_{\mathrm{LS}}$ 是 $\alpha'\beta$ 的最小方差线性无偏估计, 即 $\alpha'\hat{\beta}_{\mathrm{LS}}$ 的方差小于或等于 $\alpha'\beta$ 的任一线性无偏估计的方差. 该性质即为高斯–马尔可夫定理, 其证明如下:

任给 $\alpha'\beta$ 的无偏估计 $\gamma'Y$, 则由无偏性可知 $\gamma'X=a'$, 从而有

$$\begin{aligned}\mathrm{Var}(\gamma'Y)-\mathrm{Var}(\alpha'\hat{\beta}_{\mathrm{LS}}) =&\sigma^2\gamma'\gamma-\sigma^2\alpha'(X'X)^{-1}\alpha\\ =&\sigma^2\gamma'\gamma-\sigma^2\gamma'X(X'X)^{-1}X'\gamma\\ =&\sigma^2\gamma[I-X(X'X)^{-1}X']\gamma\end{aligned}$$

注意到, $I-X(X'X)^{-1}X'$ 是正投影阵 (对称幂等阵), 故知其非负定性, 从而有 $\mathrm{Var}(\gamma'Y)\geqslant\mathrm{Var}(\alpha'\hat{\beta}_{\mathrm{LS}})$, 结论得证.

(5) 在 $Y\sim N(X\beta,\sigma^2I)$ 的假定下, 有如下结论:

(i) $\hat{\beta}_{\mathrm{LS}}\sim N(\beta,\sigma^2(X'X)^{-1})$;

(ii) $\dfrac{(n-k-1)\hat{\sigma}^2}{\sigma^2}\sim\chi^2(n-k-1)$;

(iii) $\hat{\beta}_{\mathrm{LS}}$ 与 $\hat{\sigma}^2$ 独立.

证明 (i) 由 $\hat{\beta}_{\mathrm{LS}}=(X'X)^{-1}X'Y$, Y 的正态性及性质 (I), (II) 知结论成立.

(ii) $\dfrac{(n-k-1)\hat{\sigma}^2}{\sigma^2}=\dfrac{1}{\sigma^2}(Y-X\hat{\beta}_{\mathrm{LS}})'(Y-X\hat{\beta}_{\mathrm{LS}})=\dfrac{1}{\sigma^2}Y'[I-X(X'X)^{-1}X']Y$.

注意到 $I-X(X'X)^{-1}X'$ 是 $L(X)$ 上的正投影阵, 所以 $[I-X(X'X)^{-1}X']X=0$, 且存在正交阵 T, 使得 $I-X(X'X)^{-1}X'=T'\begin{pmatrix}I_{n-k-1}&0\\0&0\end{pmatrix}T$, 故而有

$$\frac{(n-k-1)\hat{\sigma}^2}{\sigma^2}=\left(\frac{Z}{\sigma}\right)'\begin{pmatrix}I_{n-k-1}&0\\0&0\end{pmatrix}\left(\frac{Z}{\sigma}\right),\quad \text{其中, } Z=T(Y-X\beta)$$

从而由 $Z/\sigma\sim N(0,I)$ 知, $\dfrac{(n-k-1)\hat{\sigma}^2}{\sigma^2}\sim\chi^2(n-k-1)$.

(iii) 由 $\begin{pmatrix}\hat{\beta}_{\mathrm{LS}}\\Y-X\hat{\beta}_{\mathrm{LS}}\end{pmatrix}=\begin{pmatrix}(X'X)^{-1}X'\\I-X(X'X)^{-1}X'\end{pmatrix}Y$, 知 $\begin{pmatrix}\hat{\beta}_{\mathrm{LS}}\\Y-X\hat{\beta}_{\mathrm{LS}}\end{pmatrix}$ 服从正态分布, 且

$$\begin{aligned}\mathrm{Cov}(\hat{\beta}_{\mathrm{LS}},Y-X\hat{\beta}_{\mathrm{LS}}) =&(X'X)^{-1}X'\mathrm{Cov}(Y,Y)[I-X(X'X)^{-1}X']\\ =&\sigma^2(X'X)^{-1}X'[I-X(X'X)^{-1}X']\\ =&0\end{aligned}$$

故而有 $\hat{\beta}_{\mathrm{LS}}$ 与 $Y-X\hat{\beta}_{\mathrm{LS}}$ 独立, 从而有 $\hat{\beta}_{\mathrm{LS}}$ 与 $\hat{\sigma}^2$ 独立.

1.1.1.4　中心化和标准化

从 1.1.1.2 节和 1.1.1.3 节的讨论知, 矩阵 $X'X$ 在线性模型下的参数推断中起着重要的作用. 因此, 可以考虑对模型作一些适当的变换以简化该矩阵. 将要讨论的中心化和标准化可以起到这个作用.

记 $\bar{x}_j = \dfrac{1}{n}\sum\limits_{i=1}^{n} x_{ij}, j = 1, \cdots, k, \bar{x} = (\bar{x}_1, \cdots, \bar{x}_k)'$, 则方程 $Y = X\beta + \varepsilon$ 可改写为

$$Y_i = \beta_0^* + \sum_{j=1}^{k} \beta_j (x_{ij} - \bar{x}_j) + \varepsilon_i, i = 1, \cdots, n,$$

其中, $\beta_0^* = \beta_0 + \beta_1 \bar{x}_1 + \cdots + \beta_k \bar{x}_k$. 这里, $x_{ij}^* \hat{=} x_{ij} - \bar{x}_j, i = 1, \cdots, n, j = 1, \cdots, k$ 称为中心化的解释变量样本. 上式显示, 对解释变量样本的中心化, 从参数角度看, 只是将原参数 $\beta_0, \beta_1, \cdots, \beta_k$ 作了一个线性变换, 且变换中 $\beta_1, \cdots, \beta_k$ 没有变化. 若将上式的矩阵形式记作 $Y = (\mathbf{1}, X^*)(\beta_0^*, \beta_1, \cdots, \beta_k)' + \varepsilon$, 则易证

$$(\mathbf{1}, X^*)'(\mathbf{1}, X^*) = \begin{pmatrix} n & 0 \\ 0 & X^{*\prime} X^* \end{pmatrix}$$

而 β_0^* 和 $(\beta_1, \cdots, \beta_k)'$ 的最小二乘估计为 $\hat{\beta}_0^* = \bar{Y}, (\hat{\beta}_1, \cdots, \hat{\beta}_k)' = (X^{*\prime} X^*)^{-1} X^{*\prime} Y$, 且 $(\hat{\beta}_0^*, \hat{\beta}_1, \cdots, \hat{\beta}_k)'$ 的协方差矩阵是

$$\sigma^2 \begin{pmatrix} 1/n & 0 \\ 0 & (X^{*\prime} X^*)^{-1} \end{pmatrix}$$

这意味着, 中心化使得回归系数 $\beta_1, \cdots, \beta_k$ 同新的常数项 β_0^* 可以被分开处理, 这在某些问题的分析中将会提供很大的便利.

方程 $Y = X\beta + \varepsilon$ 还可进一步被改写为

$$y_j = \beta_0^* + \sum_{j=1}^{k} \tilde{\beta}_j \frac{x_{ij} - \bar{x}_j}{s_{jj}} + \varepsilon_i, \quad i = 1, \cdots, n$$

其中, $\tilde{\beta}_j = \beta_j \cdot s_{jj}$, $s_{jj}^2 = \sum\limits_{i=1}^{n} (x_{ij} - \bar{x}_j)^2$. 这里, $\tilde{x}_{ij} = \dfrac{x_{ij} - \bar{x}_j}{s_{jj}}$, $i = 1, \cdots, n, j = 1, \cdots, k$ 称为标准中心化的解释变量样本. 将上式的矩阵形式记作 $Y = (\mathbf{1}, \widetilde{X})(\beta_0^*, \tilde{\beta}_1, \cdots, \tilde{\beta}_k)' + \varepsilon$, 则

$$(\mathbf{1}, \widetilde{X})'(\mathbf{1}, \widetilde{X}) = \begin{pmatrix} n & 0 \\ 0 & \widetilde{X}' \widetilde{X} \end{pmatrix}$$

且 $\widetilde{X}' \widetilde{X}$ 恰为解释变量 $x_1, \cdots, x_k$ 的样本自相关矩阵. 标准化的另一个价值在于消除了自变量样本中的取值单位的影响, 避免了不同的解释变量样本值之间差异过大的问题.

1.1.1.5 极大似然估计

在 $Y \sim N(X\beta, \sigma^2 I)$ 的假设下, 可以获得 β 和 σ^2 的极大似然估计. 由假设易得似然函数

$$L(\beta, \sigma^2; Y) = p(Y; \beta, \sigma^2) = (2\pi)^{-\frac{n}{2}} (\sigma^2)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma^2}(Y - X\beta)'(Y - X\beta)\right\}$$

易见, $L(\beta, \sigma^2; Y)$ 达到极大值的必要条件是 $(Y - X\beta)'(Y - X\beta)$ 达到极小, 故有 β 的极大似然估计 $\hat{\beta}_{\mathrm{ML}} = \hat{\beta}_{\mathrm{LS}}$. 将 $\hat{\beta}_{\mathrm{ML}}$ 代入 $L(\beta, \sigma^2; Y)$ 得

$$L(\hat{\beta}_{\mathrm{ML}}, \sigma^2; Y) = (2\pi)^{-\frac{n}{2}} (\sigma^2)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma^2}(Y - X\hat{\beta}_{\mathrm{ML}})'(Y - X\hat{\beta}_{\mathrm{ML}})\right\}$$

极大化 $L(\hat{\beta}_{\mathrm{ML}}, \sigma^2; Y)$ 可得 σ^2 的极大似然估计:

$$\hat{\sigma}^2_{\mathrm{ML}} = \frac{1}{n}(Y - X\hat{\beta}_{\mathrm{ML}})'(Y - X\hat{\beta}_{\mathrm{ML}})$$

1.1.2 拟合优度和方程的显著性检验

在获得模型 $Y = X\beta + \varepsilon$ 中 β 的估计后, 一个值得关注的问题是, 得到的拟合方程 $Y = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k$ 与样本数据之间的接近程度如何. 这在一定程度上标志着拟合方程是否真正刻画了数据所遵从的规律. 注意到残差平方和 $SSE = \|Y - X\hat{\beta}_{\mathrm{LS}}\|^2$ 所刻画的就是各个数据点的因变量值与拟合值之间差异的总和, 所以 SSE 可以作为对上述问题进行推断的一个依据. 然而, SSE 仍存在不足, 就是对量纲的依赖. 在 SSE 的基础上, 考虑到样本数据因变量值本身的离散程度, 可以定义拟合优度(也称决定系数) 如下:

$$R^2 = 1 - \frac{SSE}{SST}$$

其中, $SST \hat{=} \|Y - \mathbf{1} \cdot \bar{y}\|^2 = \sum_{i=1}^{n} (y_i - \bar{y})^2$, 称为总平方和.

为了研究 R^2 的含义和性质, 先介绍一个有用的结论: $SST = SSE + SSR$, 其中, $SSR = \left\|X\hat{\beta}_{\mathrm{LS}} - \mathbf{1} \cdot \bar{y}\right\|^2 = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$, 称为回归平方和. 事实上,

$$SST = \|Y - \mathbf{1} \cdot \bar{y}\|^2 = \left\|Y - X\hat{\beta}_{\mathrm{LS}}\right\|^2 + \left\|X\hat{\beta}_{\mathrm{LS}} - \mathbf{1} \cdot \bar{y}\right\|^2 + 2(Y - X\hat{\beta}_{\mathrm{LS}})'(X\hat{\beta}_{\mathrm{LS}} - \mathbf{1} \cdot \bar{y})$$

而 $(Y - X\hat{\beta}_{\mathrm{LS}})'(X\hat{\beta}_{\mathrm{LS}} - \mathbf{1} \cdot \bar{y}) = Y'[I - X(X'X)^{-1}X'][X(X'X)^{-1}X' - \frac{1}{n}\mathbf{1} \cdot \mathbf{1}']Y = 0$(因为$\mathbf{1}$是 X 的第 1 列, 而 $[I - X(X'X)^{-1}X']X = 0$). 综上所述, 上述平方和分解式成立.

由平方和分解式可知, $0 \leqslant R^2 \leqslant 1$. 拟合优度 R^2 刻画的是在拟合方程下, 因变量的取值变化中可以由自变量的取值变化解释的部分所占比例, 因而 R^2 可视作拟合好坏的评价标准. R^2 的一个缺点是, 在方程中添加自变量会导致 R^2 的减小. 这意味着, 在 R^2 的标准下, 方程中自变量越多越好, 这是不合理的. 因此, 人们往往采用调整后的拟合优度:

$$R^2_{\text{adj}} = 1 - \frac{SSE/(n-k-1)}{SST/(n-1)}$$

R^2_{adj} 中加入了对自变量个数的惩罚, 弥补了上述 R^2 的缺陷.

为了准确地推断所使用的线性模型是否真正合适, 也就是, 在目前的线性模型下, 自变量是否确实对因变量产生着影响, 我们需要确定一个 R^2(或 R^2_{adj}) 的临界值, 这就是方程的显著性检验问题. 作为 R^2 的单调函数, 统计量 $F \doteq \dfrac{(SST-SSE)/k}{SSE/(n-k-1)}$ 服从第一、二自由度分别为 k 和 $n-k-1$ 的 F- 分布 (记作 $F(k, n-k-1)$). 由此可得检验的拒绝域:

$$\frac{(SST-SSE)/k}{SSE/(n-k-1)} > F_\alpha(k, n-k-1)$$

其中, $F_\alpha(k, n-k-1)$ 是 $F(k, n-k-1)$ 的上 α 分位点. 事实上, 该检验是 1.1.3 节中将要介绍的一般线性假设检验的特例. 有关理论细节, 如统计量分布的证明等, 见 1.1.3 节.

1.1.3　一般线性假设的检验及参数的置信区域

1.1.3.1　一般线性假设的检验

考虑一般线性假设如下

$$H_0 : H\beta = \gamma_0 \tag{1.1.6}$$

其中, H 为 $s \times (k+1)$ 阶行满秩矩阵, 方程 $H\beta = \gamma_0$ 有解. 这个假设涵盖了许多常见的待检假设 (只需将 H 取为相应的矩阵). 下面列举两个重要假设.

(1) 取 $\gamma_0 = 0, H' = (0, \cdots, 0, 1, 0, \cdots, 0)'$, 即取 H' 为一个列向量, 其第 i 个元素为 1, 其余元素为 0, 则假设为

$$H_0 : \beta_{i-1} = 0 \tag{1.1.7}$$

对该假设的检验就是第 $i-1$ 个回归系数的显著性检验. 该检验的结果代表着第 $i-1$ 个自变量是否对因变量产生影响.

(2) 取 $\gamma_0 = 0, H = (0, I_k)$, 则假设为

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0 \tag{1.1.8}$$

直观上看, 若该假设无法被拒绝, 则说明模型中自变量对因变量均无显著的线性形式的影响, 这就意味着方程是不显著的. 事实上, 下面的讨论将会说明假设 (1.1.8) 的检验统计量与 1.1.2 节中的线性模型显著性检验统计量是相等的, 而对假设 (1.1.8) 的检验就是线性模型的显著性检验.

下面介绍假设 (1.1.6) 的检验方法. 使用似然比检验方法导出检验统计量. 易见似然比为

$$\lambda = \frac{\sup\limits_{\beta\sigma^2} L(\beta,\sigma^2)}{\sup\limits_{H\beta=\gamma_0,\sigma^2} L(\beta,\sigma^2)} = \left(\frac{SSE_1}{SSE}\right)^{\frac{n}{2}}$$

其中, $SSE \hat{=} \min\limits_{\beta} \|Y - X\beta\|^2 = \left\|Y - X\hat{\beta}_{\mathrm{LS}}\right\|^2$, $SSE_1 \hat{=} \min\limits_{H\beta=\gamma_0} \|Y - X\beta\|^2$. 不难证明

$$SSE_1 - SSE = (H\hat{\beta}_{\mathrm{LS}} - \gamma_0)'[H(X'X)^{-1}H']^{-1}(H\hat{\beta}_{\mathrm{LS}} - \gamma_0) \tag{1.1.9}$$

证明的细节可以参见有关文献 (卞国瑞等, 1979).

由 1.1.1.3 节性质 (5) 知: ①$SSE/\sigma^2 \sim \chi^2(n-k-1)$; ②在 H_0 下, $H\hat{\beta}_{\mathrm{LS}} - \gamma_0 \sim N(0, \sigma^2 H(X'X)^{-1}H')$, 进而有 $(SSE_1 - SSE)/\sigma^2 \sim \chi^2(s)$; ③$\hat{\beta}_{\mathrm{LS}}$ 与 SSE 独立, 从而有 $SSE_1 - SSE$ 与 SSE 独立. 由上述事实, 得

$$\frac{(SEE_1 - SSE)/s}{SSE/(n-k-1)} \sim F(s, n-k-1) \tag{1.1.10}$$

综上所述, 假设 (1.1.6) 的检验水平为 α 的拒绝域为

$$\frac{SSE_1 - SSE}{SSE/(n-k-1)} > F_\alpha(s, n-k-1) \tag{1.1.11}$$

特别地, 注意到 t- 分布与 F- 分布之间的关系, 则易知, 对于第 $i-1$ 个回归系数 β_{i-1} 的显著性检验 (对假设 (1.1.7) 的检验), 实际上也可表达为一个 t- 检验的形式. 此外, 对假设 (1.1.8) 的检验是在假设 (1.1.6) 中取 $\gamma_0 = 0, H = (0, I_k)$, 此时

$$SSE_1 = \min_{\beta_1=\cdots=\beta_k=0} \|Y - X\beta\|^2 = SST, \quad s = k$$

所以, 该检验的检验统计量就是 1.1.2 节中所述线性模型的显著性检验的统计量. 由此可见, 对假设 (1.1.8) 的检验就是线性模型的显著性检验.

1.1.3.2 置信区域

本节讨论 $\theta \hat{=} H\beta$ 的置信区域. 这里, H 为 $s \times (k+1)$ 阶行满秩矩阵. 特别地, 取 $H' = (0, \cdots, 0, 1, 0, \cdots, 0)'$ 时, $H\beta = \beta_{i-1}$, 故所求就是 β_{i-1} 的置信区间; 取 $H' = (1, x_0)'$ 时, $H\beta$ 的置信区域即为在自变量 $x = x_0$ 处, 变量 Y 的均值 $E(Y_0) = \beta_0 + x_0(\beta_1, \cdots, \beta_k)'$ 的置信区间.