基于GMDH的多重共线性研究

鲁茂 著

天津大学出版社
TIANJIN UNIVERSITY PRESS

基于 GMDH 的多重共线性研究

鲁 茂 著

天津大学出版社
TIANJIN UNIVERSITY PRESS

图书在版编目(CIP)数据

基于 GMDH 的多重共线性研究/鲁茂著. —天津:天津大学出版社,2014.4

ISBN 978-7-5618-5021-3

Ⅰ.①基… Ⅱ.①鲁… Ⅲ.①共线性-研究 Ⅳ.①O212.1

中国版本图书馆 CIP 数据核字(2014)第 059711 号

出版发行 天津大学出版社
出 版 人 杨欢
地　　址 天津市卫津路 92 号天津大学内(邮编:300072)
电　　话 发行部:022-27403647
网　　址 publish.tju.edu.cn
印　　刷 天津大学出版社有限责任公司
经　　销 全国各地新华书店
开　　本 148mm×210mm
印　　张 6.75
字　　数 194 千
版　　次 2014 年 4 月第 1 版
印　　次 2014 年 4 月第 1 次
定　　价 48.00 元

前　言

计量经济学家古扎拉蒂在他的专著(*Basic Econometrics*, Mc Graw-Hill, New York, 1995)中举了一个说明性例子:消费支出与收入和财富的关系。其中以消费支出为因变量,收入和财富为自变量建立线性回归模型。得出了一个令人十分惊讶的违背常理的结果:财富项的系数为负值,意味着随着财富积累的增多消费支出反而变少!为何出现这种现象?原来收入和财富这两个自变量存在严重多重共线性!多重共线性是线性回归建模中一个棘手的问题,它可能使自变量设计矩阵$\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$呈病态,OLS方法估计的参数值可信度降低甚至失效,不仅会增大模型误差也会使模型丧失稳健性。因此,如何消除或降低多重共线性对建模有重要的理论意义和实际价值。

对于多重共线性人们已提出了多种补救措施,总体上可分为两大类,每一类又包括几种方法。一类措施从增加外部信息着手,包括增加样本数据和运用先验信息等途径。另外一类措施是从减少内部信息着手,包括剔除变量方法、有偏估计方法和多项式回归方法。减少内部信息是目前处理多重共线性的主要途径。通过剔除变量可以减少内部信息,但现有的剔除变量方法在选择变量时存在主观性,并且可能面临设定偏误的问题。用有偏估计方法来处理多重共线性问题时可以防止设定偏误的产生,它们的参数估计都有偏误,其中主成分回归方法还存在提取的主成分不能保证对因变量有最大解释能力的不足。用多项式回归处理多重共线性问题时,一般使用正交多项式,但它仅适用于单自变量的实验数据。鉴于

此,本书尝试通过 GMDH 方法建模途径,探索处理多重共线性问题的新的有效方法。

(1)提出处理多重共线性问题的剔除变量线性 GMDH 参数模型方法。在证明了选择最优变量子集与剔除变量是对偶关系的基础上,在本书中提出了用线性 GMDH 参数模型选择最优变量子集来处理多重共线性问题的方法。现有剔除变量方法在选择变量时存在主观性,并且可能面临设定偏误的问题。书中证明并用实验验证了在满足经典线性回归模型的假设条件下,线性 GMDH 参数模型的参数估计量具有线性无偏的性质。最后用算例对所有子集回归、逐步回归和线性 GMDH参数模型三种方法进行了比较验证。结果显示,线性 GMDH参数模型能通过客观地选择变量来处理多重共线性问题。

(2)提出处理多重共线性问题的有偏估计 C-GMDH 方法。书中提出的有偏估计 C-GMDH 方法,使选取的主成分尽量携带自变量集信息,并确保它们对因变量有最大的解释能力。在算例上对主成分回归、PLS 法和 C-GMDH 方法三种有偏估计方法进行了比较研究。通过算例比较证实,相对于 PLS 法和主成分回归,用 C-GMDH 方法建立的有偏估计模型能确保提取的主成分对因变量有最大的解释能力,有效降低多重共线性的影响,并改善参数估计的有偏性。

(3)提出处理多重共线性问题的多项式回归 B-GMDH 方法。现有处理多重共线性问题的多项式回归模型通常适用于单变量且样本是等距取值的数据。在本书中提出了适用于多变量情形的多项式回归 B-GMDH 方法。它通过迭代运用非线性 GMDH 参数模型和线性 GMDH 参数模型,消除多项式回归模型中的多重共线性问题,并简化模型。通过算例证实了

B-GMDH 方法适用于多变量情形，建立的多项式模型剔除了导致多重共线性问题的变量项。B-GMDH 模型的复杂度远小于普通的多项式回归模型，它简化了多项式回归模型，有利于进行系统的结构分析和控制。

(4)为了研究共线性问题对建模带来的不良影响，在本书中定义了有害共线性概念，并给出有害共线性的判定准则以及消除或降低其对回归分析影响的方法。

本书作为一个基础性研究内容，研究过程中得到了四川师范大学科研创新团队项目“民营企业内部治理与创新发展研究”、四川省教育厅重点项目(13SA0136)、四川省教育厅重点项目(13SA0139)、国家自然科学基金项目(No. 71071101)等的资助和鼓励，特此致谢！

由于受研究水平、能力和时间所限，本课题研究难免存在不妥甚至错误之处，敬请同行、各位专家批评指正，恳请读者见谅。本书作者将继续从事相关问题的研究，随着实践的发展和认识水平的提高，必定会有更多的新认识，期待与读者分享。

鲁茂

2013 年 10 月 1 日

目　录

第1章　导论

1.1　问题提出

在自然科学、社会科学和工程科学等领域中，人们常常需要对研究对象用数学模型来（近似）描述，以达到对其进行预测、结构分析和控制的目的。比如在经济学领域，无论是宏观经济还是微观经济，都有大量的数学模型，这不仅是为了理论上的研究，更多的是为了更好地保持一个地区或国家经济的平稳、有效发展。又如一个大型的企业，从它的原材料的合理采购、产品的生产节奏、合格产品的生产控制，到产品的生存周期等，都有数学模型的身影。甚至可以说，管理科学的实现，离不开数学模型的支持。

数学模型从其表现特征可分为确定性模型和随机性模型[147]。显然，大多数实际问题是随机性的。线性回归模型是应用最为广泛的随机性模型之一。

参数估计是线性回归模型的重点，主要采用普通最小二乘方法（Ordinary Least Squares，OLS）来估计。自从高斯（C. F. Gauss）于1809年建立了OLS法，它就作为一个良好的估计方法而被广泛采用。

建立线性回归模型首先要做的工作是选择自变量。然而要选择合适的自变量并不是一件很容易的事情，不同的专家学者对同一个问题可能有不同的看法，加之对领域知识的缺乏，从而造成自变量的选取难以抉择。因此，建模者为避免遗漏重要的系统特征，倾向于更多地选取相关自变量，把能想到的自变量都罗列进来，唯恐丢掉与问题相关的自变量。这样一来，由于涉及的自变量过多，另外一个问题——多重共线性就容易产生了。

多重共线性可能会对建立的线性模型造成困难，使自变量设计矩阵 $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ 呈病态（奇异），参数估计值 β 可信度降低甚至失效，模型误差增大，也会使模型丧失稳健性[40,58,134]，并将严重限制回归模型在推断和预报中的作用[16]。

尽管对多重共线性问题的处理已经做了许多的研究，但到目前为止依然没有得到较好的处理。因此对多重共线性的研究和处理，一直是多元线性回归分析的一个重点和难点，人们一直在寻找处理多重共线性的更好方法。

1.2 国内外研究现状

1.2.1 多重共线性研究现状

多重共线性（Multicollinearity）一词最先由计量经济学奠基人 Ragnar Frisch 在1934年引入[32]。多重共线性对线性回归模型的影响与危害、产生原因、诊断方法和补救措施以及整个体系主要是在20世纪六七十年代构建完成。后来的工作主要是对以前的方法或规则做进一步的挖掘和完善。有关多重共线性问题的讨论，主要从两类文献渠道获得：回归分析和计量经济学。

尽管在新中国成立之初我国一些有识之士曾倡导学习、研究西方经济学中的定量分析方法，但直到20世纪80年代，相关研究在我国才得以全面展开。因此，对于多重共线性的理论研究成果基本上是来自于欧美国家。国内研究主要是对某种处理方法进行改进，更多的是实证研究。

1. 多重共线性定义

如果 n 个自变量中的 k 个自变量，满足下面的条件：

$$\lambda_1 X_1 + \lambda_2 X_2 + \cdots + \lambda_k X_k + \varepsilon = 0 \tag{1-1}$$

其中 $2 \leqslant k \leqslant n$；$\lambda_1, \lambda_2, \cdots, \lambda_k$ 是不同时为零的常数；ε 为随机误差项。当 $\varepsilon = 0$ 时即为 Frisch 所指的多重共线性，现在则称为完全多重共线性；而当 $\varepsilon \neq 0$，且 $\varepsilon \to 0$ 时，则这 k 个自变量存在严重多重共线性[40]。

完全多重共线性在实际数据中是一种极端情况，其表现特征、后果及处理办法也是非常明确的。因此，对多重共线性问题的讨论，实际上

就是针对严重多重共线性这种情形。

2. 多重共线性的来源

多重共线性的来源被认为有多种原因。戈德伯格(Goldberger)认为多重共线性产生的原因主要是样本量的缺乏[38]。Gunst[42]，Mason等人[83]和 Montgomery 等人[88]认为多重共线性可能由以下四个因素导致：数据采集所用的方法、模型或从中取样的总体受到约束、模型设定和一个过度决定的模型。除以上原因外，陈希孺[134]、盛承懋[163]还认为自变量之间客观上就有相依性关系存在，这同样会导致多重共线性的出现。

以上这些观点，已经指出了产生多重共线性的各种原因。总体来说，缺乏足够的样本量被普遍认为是导致多重共线性产生的本质原因。

3. 多重共线性的诊断方法

为了侦察多重共线性，人们也提出了多种方法。这些方法包括：自变量间的简单相关、偏相关、t 检验、辅助回归、考察设计矩阵 $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$、容许度与方差膨胀因子、特征值与条件数等。在这些侦察方法中，最通用的方法是容许度和特征值两种方法。最简单但并不好的方法是简单相关分析。

尽管以上这些侦察方法能够诊断多重共线性是否存在以及其严重的程度，但人们也认识到它们对处理多重共线性并不能保证一定灵验[40]。

4. 多重共线性的补救措施

显然，采取何种措施来处理多重共线性对线性回归建模造成的困难，是研究多重共线性问题的最终目的。到目前为止，人们提出了多种补救措施。其中最主要的措施包括：追加样本信息、利用先验信息、变量转换、多项式回归、剔除一些不重要的变量、主成分回归、特征根估计、岭回归、PLS 法等。

这些措施总体上可分为两大类。一类从外部信息着手，包括追加样本信息和利用先验信息等方法。通常情况下追加样本信息既不现实也不一定有效[58]；而先验信息也未必满足相关的理论知识[40]。因此从外部信息着手处理多重共线性问题的方法在应用上受到极大的限

制。另外一类是从内部信息着手,包括剔除变量法、有偏估计法和多项式回归法。这些方法是处理多重共线性问题的主要方法。

处理多重共线性最直观的方法是剔除模型中引起多重共线性问题的变量,主要方法有岭回归和逐步回归。岭回归通过岭迹的判断来筛选自变量,这对建模者的主观判断能力有很高的要求。逐步回归选择的自变量与 F 统计量显著性水平 α 的主观取值有关[134],从而可能使重要的自变量被漏选。这两种方法在选择变量时都存在主观性的不足,另外剔除变量还可能面临设定偏误的问题。

有偏估计方法是目前研究应用最多、最普遍的一类处理多重共线性问题的方法,主要包括主成分回归、岭回归和 PLS 法。相对于剔除变量方法而言,有偏估计方法不会产生设定偏误,但参数的估计值都是有偏的。

主成分回归从自变量集中提取携带最大信息的主成分来建模。由于主成分之间相互正交,因此主成分之间不存在多重共线性问题;但选取的主成分并不一定对因变量有很好的解释能力,从而使建立的模型可靠性降低[143]。为了使提取的主成分与因变量有最大的相关性,Wold 提出用 PLS 法来处理多重共线性问题[167]。但 PLS 法在用交叉有效性 Q_h^2 确定主成分个数(h)的时候,Q_h^2 并不一定是减函数,从而存在 h 选取上的困难[166];另外 PLS 估计总体上压缩了 OLS 估计,但其中还会膨胀部分 β 的参数估计值,这是回归分析中首先要预防的[29]。

由于多重共线性的一个不良现象是设计矩阵 $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ 退化,岭回归的另一种处理多重共线性的策略是增加一个合适的 k 参数,使矩阵($\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}+k\boldsymbol{I}$)变得稳定从而防止多重共线性带来的危害;然而岭参数 k 值还没有一个公认的选取方法,只能由主观确定,限制了它的有效应用[134,59,137]。

还有一类处理多重共线性问题的方法是多项式回归,主要通过建立正交多项式回归来防止多重共线性的产生。但是正交多项式适用于单自变量的多项式回归,且自变量的样本要求等间隔取值[104]。这些条件使正交多项式的应用受到极大的限制。由于单变量非正交多项式的系数矩阵 $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ 病态的程度很严重[27],可以推知多变量的多项式回

归必然面临多重共线性的问题。作为一种建立多项式线性模型的非线性 GMDH 参数模型方法,其中的确存在多重共线性问题[110]。尽管人们已经提出了一些改进方法,但多重共线性问题依然存在。

针对目前处理多重共线性方法的不足,本书考虑从内部信息着手的措施,基于 GMDH 方法提出三种新的处理多重共线性的方法。在用 OLS 法建模存在多重共线性问题时,首先考虑用线性 GMDH 参数模型来剔除自变量集中导致多重共线性问题的变量,其基本条件是在难以确定模型应包含哪些自变量的情况下应用;如果能确定模型应包含哪些自变量,但仍然受到多重共线性问题的困扰,那么采用基于 GMDH 方法的有偏估计方法,以降低多重共线性带来的危害并保证所有重要的自变量都包含在模型中。对于多变量的复杂系统,如果直接建立线性回归模型难以达到建模目的,那么建立基于 GMDH 方法的多项式回归模型是值得考虑的方法,可以最大限度地建立一个较高拟合精度的线性回归模型。前面两种新方法有一定的联系,而它们与第三种方法联系不大。

1.2.2 GMDH 研究现状

GMDH(Group Method of Data Handling,数据分组处理)是由乌克兰科学院院士 A. G. Ivakhnenko 于 1967 年创立,用于对复杂系统建模[50,51,52]。它将黑箱思想、生物神经元方法、归纳法、概率论和 Godel 数理逻辑等方法有机地结合起来,实现了自动控制与模式识别理论的统一,极大地减少了人在建模过程中的参与,更具客观性与公正性[75]。建模者只需提供系统的样本数据,选择外准则并进行数据分组;计算机则选择模型结构和应该包含的变量。其最终结果为最优复杂度模型。

GMDH 方法在线性参数和非线性参数建模上都在发展,但线性参数模型仍然是主要的研究内容。从变量在模型中的变化形式看,GMDH方法建立的线性模型还可以细分为线性和非线性变量两类。这两种方法分别称为线性 GMDH 参数模型和非线性 GMDH 参数模型。

Ivakhnenko 提出的原始 GMDH 方法就是非线性 GMDH 参数模型。它是以 Kolmogorov-Gabor(K-G)多项式为模型形式来逼近未知非线性复杂模型。建模过程中的传递函数通常是二元二次多项式,经过逐层

迭代组合得到最终的最优复杂度模型。非线性 GMDH 参数模型建立的模型都是嵌套模型，如果将其展开，实际上就是多项式线性回归模型，可以发现其变量项极其庞大，少则几十项，多则成千上万项。因此非线性 GMDH 参数模型建立的模型中本身也存在多重共线性问题，也是亟待解决的问题之一。

如果建模过程中的传递函数是二元一次多项式，则得到的模型就是线性 GMDH 参数模型。尽管建模过程也要进行数据分组和模型逐渐复杂化，但线性 GMDH 参数模型的最终模型很容易化简。因此，最终模型与通常的线性回归模型无异，表达式为 $Y=\beta_0+\sum\beta_iX_i$。线性 GMDH 参数模型实际上是在自变量集合中选择最优变量子集，建立一个最优变量模型。多重共线性问题的一个处理办法是剔除变量，如用逐步回归法来进行变量的选择。那么是否可以用线性 GMDH 参数模型来处理多重共线性问题呢？这是本书要解决的一个问题。

1.3 研究内容、方法和创新

1.3.1 研究内容

本书介绍的研究内容拟解决的问题是目前处理多重共线性问题的补救措施存在哪些不足之处，哪些措施需要改进或者应用新的方法；如果这些措施对处理多重共线性问题效果不好，问题在哪里，又该如何解决。具体内容包括如下几点。

(1)通过对目前多重共线性研究的概述，指明其从定义、产生原因、诊断方法到补救措施等各个部分存在的不足。

(2)对剔除变量方法进行了探讨，进一步确定该方法在处理多重共线性问题上的可行性；在证明该方法的对偶问题是选择最优变量子集的基础上，确定了多重共线性问题可用变量选择方法来处理。由于现有剔除变量方法存在主观性，提出用线性 GMDH 参数模型方法来处理多重共线性问题；讨论并研究该方法的建模特点，包括变量选择的客观性、参数估计的无偏性等内容；与其他变量选择方法进行了比较研究。对剔除变量可能面临的设定偏误进行了研究。

(3)对处理多重共线性问题的有偏估计方法进行概述，从现有主

成分回归方法存在的不足，提出一种新的有偏估计 C-GMDH 方法，弥补主成分回归提取的主成分不能保证对因变量有最大解释能力的缺点。

(4)对多项式回归面临的多重共线性问题做了概述，从现有方法存在的不足，提出了适用于多变量情形的多项式回归 B-GMDH 方法，消除多项式回归模型中的多重共线性问题，并简化了模型。

在以上的建模思想及理论研究基础上，以一定量的实验和算例对部分主要研究内容进行了验证和说明，并得出一些有意义的结论。

(5)为了确定哪些共线性对建模会带来不良影响，本书定义了有害共线性概念，并给出它的判定准则以及消除或降低其对回归分析影响的方法。

1.3.2　研究方法

本书的研究主体就是多重共线性问题，处理手段主要是基于GMDH方法展开的。主要探讨目前多重共线性问题的补救措施存在的不足，如何改进或提出一些新的方法，同时也涉及一些相关的问题。

本研究拟采用规范研究与实证研究相结合、定性与定量相结合的集成方法。在理论上，以回归分析、计量经济学、运筹学、数学建模等为理论基础，重点探讨多重共线性的各种补救措施存在的不足，如何基于 GMDH 方法提出新的补救措施。在理论研究的过程中，在必要的地方增加相应的实际算例，对部分研究内容进行实证说明。

本课题研究的基本思路如图 1-1 所示。

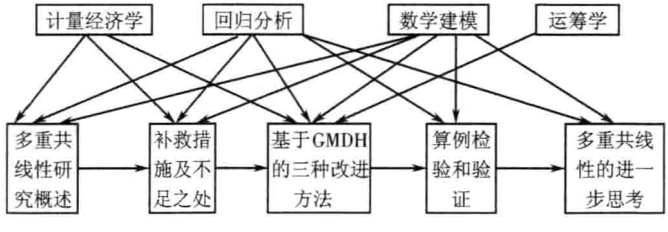

图 1-1　本课题研究基本思路

1.3.3 研究创新

在处理多重共线性问题方面有很多补救措施，其中最主要的是剔除变量方法、使用有偏估计方法和多项式回归建模。本书基于 GMDH 理论与方法，在这些方面提出了一系列新的途径。

(1)提出处理多重共线性问题的剔除变量线性 GMDH 参数模型方法。

在证明了选择最优变量子集与剔除变量是对偶关系的基础上，本书提出了用线性 GMDH 参数模型选择最优变量子集来处理多重共线性问题的方法。现有剔除变量方法在选择变量时存在主观性，并且可能面临设定偏误的问题。本书证明并用实验验证了在满足经典线性回归模型的假设条件下，线性 GMDH 参数模型的参数估计量具有线性无偏的性质。最后用算例对所有子集回归、逐步回归和线性 GMDH 参数模型三种方法进行了比较验证。结果显示，线性 GMDH 参数模型能通过客观地选择变量来处理多重共线性问题。

(2)提出处理多重共线性问题的有偏估计 C-GMDH 方法。

针对主成分回归提取的主成分不能保证对因变量有最大解释能力的不足，本文提出的有偏估计 C-GMDH 方法，使选取的主成分在尽量携带自变量集信息的同时，确保它们对因变量有最大的解释能力。在算例上对主成分回归、PLS 法和 C-GMDH 方法三种有偏估计方法进行了比较研究。通过算例比较证实，用 C-GMDH 方法建立的有偏估计模型能降低多重共线性的影响，能确保提取的主成分对因变量有最大的解释能力；同时发现，相对于 PLS 法和主成分回归，用 C-GMDH 方法建立的有偏估计模型能改善参数估计的有偏性。

(3)提出处理多重共线性问题的多项式回归 B-GMDH 方法。

现有处理多重共线性问题的正交多项式回归建模法，通常适用于单变量且样本是等距取值的数据。本文提出了适用于多变量情形的多项式回归 B-GMDH 方法。它通过迭代运用非线性 GMDH 参数模型和线性 GMDH 参数模型，处理多项式回归模型中的多重共线性问题，并简化模型。通过算例证实了 B-GMDH 方法适用于多变量情形，建立的多项式模型在处理多重共线性问题后，模型的拟合精度依然较高。B-GMDH 模型的复杂度远小于普通的多项式回归模型，它简化了多项式