模型的魅力

2012全国统计建模大赛获奖论文选

全国统计建模大赛执行委员会
国家统计局统计教育培训中心 编

中国统计出版社
China Statistics Press

模型的魅力

2012全国统计建模大赛获奖论文选

全国统计建模大赛执行委员会
国家统计局统计教育培训中心 编

中国统计出版社
China Statistics Press

图书在版编目(CIP)数据

模型的魅力：2012全国统计建模大赛获奖论文选 / 全国统计建模大赛执行委员会，国家统计局统计教育培训中心编. — 北京：中国统计出版社，2013. 4

ISBN 978－7－5037－6797－5

Ⅰ. ①模… Ⅱ. ①全… ②国… Ⅲ. ①统计模型－文集 Ⅳ. ①C8－53

中国版本图书馆CIP数据核字（2013）第059218号

模型的魅力:2012全国统计建模大赛获奖论文选

作　　者/全国统计建模大赛执行委员会　国家统计局统计教育培训中心编
责任编辑/张　赏
特约编辑/孙　慧　李　锐
封面设计/李雪燕
出版发行/中国统计出版社
通信地址/北京市丰台区西三环南路甲6号　邮政编码/100073
电　　话/邮购(010)63376909　书店(010)68783171
网　　址/http://csp.stats.gov.cn/
印　　刷/河北天普润印刷厂
经　　销/新华书店
开　　本/710×1000mm　1/16
字　　数/600千字
印　　张/37.5
版　　别/2013年4月第1版
版　　次/2013年4月第1次印刷
定　　价/62.00元

国家统计局2012全国统计建模大赛组织委员会

主　任　马建堂

副主任　张为民　罗　兰　徐一帆　谢鸿光　许宪春
李　强

委　员　鲜祖德　郑京平　曾玉平　毛有丰　张仲梁
曹志刚　许剑毅　田鲁生　王立元　潘　璠
翟　艳　朱玉利　严建辉　叶植材

国家统计局2012全国统计建模大赛执行委员会

主　任　田鲁生

副主任　邱小聪　曹志刚　胡　帆　王立元　许亦频
万晓君　钟守洋　万东华

委　员　齐占林　汤魏巍　任全忠　孙　慧　王小舟
戴宏国

注：全国统计建模大赛执行委员会办公室设在统计教育培训中心。

序　言

历时数月，由全国统计系统66支代表队、200多位青年才俊参加的第三届全国统计建模大赛在2012年金秋九月的北京落下帷幕。

自2008年开始，国家统计局在全国统计系统每隔两年举行一次这项赛事，其目的就是要在统计青年中营造创新进取、钻研业务的氛围，逐步提高全系统利用模型分析处理数据、运用数据的能力。

本次大赛以“构建一个评估地区数据协调性的模型”为题，较好地结合了统计工作实际，既有一定难度，又相对开放，没有标准的方法和程式，有效地调动和发挥了参赛者的潜能。从参赛作品来看，整体水平较前两届有所提高。建模思路更显宽泛，建模方法更加多元，充分展现了统计系统注意追踪统计科技最新发展成果，不断学习的良好风尚，体现出参赛选手们很强的创新意识和团队合作精神，从一个侧面说明了举办统计建模大赛对促进统计队伍建设所起到的积极作用。

目前，人类已经进入产生海量电子化信息的“大数据”时代，统计工作面临着新的重大挑战和机遇。如何更好地利用这些“大数据”拓展充实完善统计数据收集渠道，如何通过“数据挖掘”技术，从这些海量电子化的数据中挖掘出有用的信息，是统计系统面临的紧迫难题。而统计建模正好从一个侧面提供了这方面的理论和实践。统计建模大赛则成为利用“大数据”的一个很好的历练平台。因此，希望大家能够以统计建模大赛为契机，不断深入思考和研究利用统计建模理论和方法解决实际问题的能力，为统计改革发展做出新的更大的贡献。

为了总结经验，使更多的人加入到利用模型分析处理、运用统计数据的行列，我们特将本次大赛的获奖论文选编集结成册，供大家参考。当然，由于理论水平和实践经验，以及时间所限，论文还比较粗浅，还存在着这样或那样的问题，欢迎读者批评指正。

郑京平

2013年2月

目录

一、全国统计建模大赛一等奖论文

二、全国统计建模大赛二等奖论文

三、全国统计建模大赛优秀专题论文

一、全国统计建模大赛一等奖论文

2012 全国统计建模大赛赛题

构建一个评估地区数据协调性的模型

统计数据在现代社会中的作用越来越大，是政府管理和人们决策的重要依据。高质量的统计数据可以帮助人们作出正确判断，而低劣的统计数据则会导致人们采取错误行动，从而付出惨痛代价。然而，统计数据又是现实的环境中产生的，参与主体互相博弈，如何确保博弈结果真实可靠一直是我国统计部门面临的难题，数据之间不协调、不匹配的现象时有发生。

全国数据是根据地方数据计算出来的，因此只有地方数据搞准了，全国数据才有可靠的基础。为了提高地方数据的可靠性，对地方数据进行协调性和匹配性评估是重要一环。请构造一个评估地方统计数据协调性的模型，要求模型具有通用性，也就是适用于所有地区和较长的时期，评估对象可以是一个或几个重要的指标，评估时期可以是年度也可以是季度，评估可以从总量、结构、比例关系和增长率等角度进行。由于构建该类模型的难度很大，甚至很难找到理想的解决方案，因此该项研究侧重于探索性和实用性，而不是追求完美结果。

一、论文的具体要求

1. 数据的憐选和加工（占 10 分）。大赛组委会仅提供了部分相关数据，而不是全部，其他数据要靠参赛人员自己搜集。要从可比性和异常变动等角度对建模所用数据进行憐选，必须把不合要求的数据剔除或作出恰当的技术处理。

2. 模型的构建（占 40 分）。一是模型的构建要有理论依据，对模型的适用条件及存在的问题有必要说明；二是估计过程应遵循建模的基本规范，步骤完整清晰，技术难点和特殊问题有必要说明；三是对协调性评估模型的技术难题作了恰当处理，协调性评估的假设前提是数据可靠性有问题，而这正好与估计模型对数据的要求相悖。

3. 模型的应用（占 30 分）。以构建的模型为基础，对各地区数据进行评估，把不协调的数据找出来，并从更广泛的角度对评估的结果的合理性进行论证。根据评估结果出现的问题对构建的模型进行评价，分析引发问题的原因，进而提出改进的方向和途径。

4. 论文结构合理、表述清晰、逻辑严谨、文字精炼、格式规范（占 20 分）。

二、书写格式要求

论文正文以10000字为限(不包括图表及附录),摘要在1000字左右,使用四号仿宋字。全文由以下几个部分组成:

(1)论文题目

(2)作者,署"第X代表队",X为参赛队编码

(3)论文摘要

(4)正文

(5)附录

(6)参考文献

附录是可选的,其他几个部分不能缺。

三、基本数据

1. 全国及地区GDP总量和增长率
2. 全国及地区工业增加值增长率及比重
3. 全国及地区固定资产完成额
4. 全国及地区消费品零售额及总量
5. 全国及各地区能源消费总量及增长率
6. 全国及地区用电量、工业用电量
7. 全国及各地区建筑总产值及增长率
8. 全国及各地区的货运量和周转量、公路货运量及周转量
9. 全国及各地区城乡居民收入和消费支出
10. 全国及地区税收总额
11. 全国及地区增值税和营业税
12. 全国及各地区居民消费价格、工业生产者出厂价格、固定资产投资价格
13. 全国及各地区施工项目和新开工项目计划总投资

基于两种方法的部分省份宏观经济数据匹配性研究

陕西省统计局　　赵翔、程飞、潘英杰

摘　要

当前，国际国内对我国政府统计数据质量一直质疑不断。在质疑和反质疑的过程中，学界和统计部门发展出了两类评估数据协调性的方法，一类可概括为“标杆法”，另一类可概括为“统计诊断法”。这两类方法各有侧重，本文尝试采用这两种思路探索有效的数据评估方法。

此次竞赛的要求实质在于评估地方统计局所出数据的可靠性，本文在“标杆法”的思路上，假设外部数据（即非统计部门生产数据）具有客观性，探索利用稳健MM估计方法，分别以北京、天津、河南、陕西相关数据，构建了以地区财政收入、地区用电量、地区货运量等外部数据为自变量，地区GDP为因变量的评估模型，并通过拟合模型对四省市GDP数据进行预测，然后利用预测值与观测值（统计公布值）的对比，获得相对误差率，对超过5%的，认为不协调。

本文同时以“统计诊断法”思路为补充，尝试利用稳健距离自适应异常值检验方法（简称RDAOD方法），从多维数据异常值的角度来研究经济指标数据间是否存在不平衡性。考虑不再根据指标之间的经济内在联系构建模型，本文将北京、天津、河南和陕西等四省市的检验指标扩展到17个，时间跨度扩展到1978—2010，在此基础上用RDAOD方法对四省市的年度经济指标协调性进行了评估。

本文得出以下结论：(1)我国各省的经济发展阶段不同，经济运行规律差距很大，试图寻找适用全国各省的评估标杆，实践中可能较不现实。(2)北京、陕西、天津、河南等省市地区GDP数据和地区财政收入、地区用电量、地区货运量的匹配程度较好。

关键词：协调性　稳健估计　异常值检验

一、相关研究综述

改革开放以来，我国经济飞速发展，GDP增速一直居于全球前列，经济总量不

断攀升，目前已居全球第二位，成就举世瞩目。但是，国际、国内对我国政府统计数据质量一直存有质疑。在质疑和反质疑的过程中，学界和政府统计部门均纷纷加大了对统计数据质量评估方法的探索力度。从现有的文献看，当前的评估方法主要分为两大类。

(1)第一类方法，可概括为“标杆法”。其核心思想是，选取一定的指标，根据这些指标数据的特征，通过指数法、相关分析法、生产函数法等，拟合出合适的模型，并通过拟合模型对目标指标数据进行预测，确立一个“标杆”。然后利用预测值与实际值(统计公布值)的对比，获得相对误差率，对选定目标数量进行评估分析。如Rawski(2001)分析认为中国官方数据经济增长率数据与能源消耗数据之间不一致、生产数据之间以及生产数据与投资数据之间不一致、消费数据之间以及消费数据与收入数据之间不一致等，由此认定中国官方统计数据有问题。孟连、王小鲁(2000)根据货物运输业增长、电力和能源消费量增长各自与工业增长之间的相关分析估计工业增长速度统计误差，认为1991—1998年我国工业增加值年均增长率的统计误差约为4.5个百分点。Klein、Ozmucur (2002)选取了包括能源、交通、通讯、劳动力、农业、贸易、公共部门、工资、通货膨胀等15个具有代表性的、来源相对独立的经济变量的变动率，对中国GDP的增长进行了解释，结论显示中国官方估计的GDP增长的相关关系是完全符合经济规律的。阙里、钟笑寒(2005)利用Klein、Ozmucur (2002)的方法，对我国各地区GDP增长统计的真实性进行了检验，结果显示我国各地区的若干基础经济变量(包括能源消费量)各自相对于GDP的变化趋势是符合基本经济规律的，没有发现GDP统计数据质量存在系统的、长期的错误的证据。孟祥兰(2011)对居民消费价格指数(CPI)及其影响因素进行建模分析，并根据模型预测结果来评价数据的有效性。柴士改(2012)利用全要素生产率(TFP)对各省GDP数据进行评估等等。李庭辉(2011)、李庭辉、许涤龙(2012)、李庭辉、薛丽娜(2012)把GDP、工业增加值等作为一个投入产出系统，利用该系统的结构稳定性，探索用用电量等指标对GDP、工业增加值数据质量进行了评估。

国家统计局在实践中也主要利用这种思路对各地主要统计数据的协调性进行评估。如利用全社会用电量增速、税收总额增速评估生产总值增速的协调性；利用综合能耗、货运量、工业用电量、增值税等增速评估工业增加值增速的协调性；利用城镇居民人均可支配收入、城镇居民人均消费性支出增速评估社会消费品零售总额增速的协调性等。

(2)第二类方法，可概括为“统计诊断法”。其核心思想是从系统的观点出发，把经济数据视为经济系统这个多维空间中的点，根据这些数据自身所具有的统计特征或在特定模型(如生产函数)下反映出来的特征，探寻利用统计的方法来衡量统计数据的质量。如李盼(2012)尝试利用奔福德定律分析了政府统计数据质量

的可信度。刘洪、黄燕(2009)利用最小二乘法估计得到生产函数,并通过残差、Cook 的 D 统计量、DIFFITS 统计量等诊断统计量,对某地区的 GDP 数据中存在的异常点进行了诊断。卢二坡、黄炳艺(2012)探索了基于稳健 MM 估计的统计数据质量评估方法,利用基于 MM 估计的稳健回归方法及异常值诊断原理,对我国改革开放以来的 GDP 数据的可靠性进行了评估。

以上两种思路大大丰富了统计数据质量评估方法。但在实践中很难确定哪类方法的优劣。第一类方法总体上要受到所选指标数据准确性的影响,如果用于评估的多个指标数据或者某个指标的多年数据都有质量问题,那么必将产生严重后果。第二类方法在运用中如果需要依靠特定模型,也存在类似问题。

基于这种现状,本文尝试同时从这两个思路入手,各探索出一种评估方法,并同对特定地区的相关统计数据的可靠性进行评估,以此探寻相对合适的评估方法。

二、基于"标杆法"的模型构建及运用

2.1 建模思路

考虑本次竞赛主要目的是构造一个评估地方统计数据协调性的模型,其实质在于评估地方统计局所出数据的可靠性,主要包括:各地区 GDP 总量及增速,GDP 中工业增加值总量,各地区规模以上工业总产值、增加值及增加值增长率,各地区固定资产投资统计数据,各地区消费品零售额,各地区城乡居民收入和消费支出等等。

因此,本文建模的思路为:选取一个待评估目标,这个指标有比较多的相关指标,并且这些相关指标主要都是外部数据(即非统计部门生产的数据)。在相关经济理论的基础上,构建一个待评估指标与这些相关指标的关系模型,并通过拟合模型对目标指标数据进行预测,确立一个"标杆"。然后利用预测值与观测值(各地区统计局公布值)的对比,获得相对误差率,对选定目标数量进行评估分析。

经过认真遴选,并统筹考虑相关指标数据的完整性和可补充性,我们发现,各地区 GDP、财政收入、各地区用电量、各地区的货运量之间有显著相关关系,可以用于建模。

2.2 理论基础及基本假设

2.2.1 经济理论基础

GDP 是宏观经济中最受关注的经济统计数字。从经济系统发展来看,GDP 与经济结构相关变量之间的关系具有稳定性。在投入产出环节,与用电量指标相关;在后续流通或消费环节与货运量指标相关;在终期形态上,与财政收入指标相关。而且这种相关具有渐进的稳定性和结构的稳定性。这些相关指标具体如下:

(1)货运量。货运量是反映经济景气程度的一个重要内容,它将生产与消费相连,是经济发展的晴雨表,在经济的转折期间表现得尤其明显。货运量反应出

来的物流情况，是经济景气与否的一个重要指标，两者是正相关关系。虽然从货运量看不出多少服务业情况，但一国或地区经济主要还是需要运输来完成的经济，特别是在服务业不太发达的地区，货运量与经济发展水平更加密切，因此本研究将其作为检测GDP数据质量的一个重要指标。

(2)全社会用电量。经济发展与能源的消耗密切相关。能源中主要三大能源包括原油、煤和电，这三个观测指标对全国的整体数据而言是很有意义的，但具体到省区的数据，则会因为自然禀赋的问题，在很多省份中原油与煤的产量与消费量都很低而且不同地区之间的差异很大，这就很难反映出各个省区能源方面的状况，因此选用了最为广泛使用的电力消费量这一指标，这一指标的地区数据体现了各个地区能源的消费情况。

(3)财政收入。财政收入既是维持国家有效运转的经济基础，又是国家调节经济的有效手段。在宏观上，财政收入是社会总产出的重要组成部分。在微观上，财政收入中的税收收入伴随着企业生产产品、提供服务、进行交易和发生其他应税行为而产生。同时，在一定的政策下，地区财政收入的多少与产出之间存在高度的相关关系。

2.2.2　基本假设

根据建模目的，本文假设：外部门数据(即非统计系统生产的数据)，可视作一手数据，具有客观属性，可作为评价统计数据协调性的标准。

2.3　数据的遴选与加工

(1)聚类分组。确立以上思路后，我们根据2010年我国各地区的GDP、人口、财政收入，居民收入等数据，用系统聚类法进行分析，简化结果后，将全国各地区分成三类。

表1　全国各省市聚类分组表

地区	类别	地区	类别
北京市	2	湖北省	3
天津市	3	湖南省	3
河北省	3	广东省	1
山西省	3	广西壮族自治区	3
内蒙古自治区	3	海南省	3
辽宁省	2	重庆市	3
吉林省	3	四川省	2
黑龙江省	3	贵州省	3
上海市	2	云南省	3

续表

地区	类别	地区	类别
江苏省	1	西藏自治区	3
浙江省	2	陕西省	3
安徽省	3	甘肃省	3
福建省	3	青海省	3
江西省	3	宁夏回族自治区	3
山东省	2	新疆维吾尔自治区	3
河南省	2		

第一类包括:江苏省、广东省;

第二类包括:北京市、上海市、辽宁省、浙江省、山东省、河南省、四川省;

第三类:其余省市地区。

(2)探索试建模型。分类后,我们曾尝试利用各组平均值建模,但结果显示模型只对平均值异常有反应,与各地区观测值差距较大。我们分析,虽然我们进行了聚类分组,但组与组之间、各组各地区之间的数据大小不一,不适宜用平均值的方式建模。这次尝试也启示了我们,在"标杆法"的思路上,想建立一个通用的(即使是相近规模的省份通用)的评估模型都不太切合实际,在实践中可能需要以省为单位来建立评估模型。

(3)遴选选定目标省份。基于以上情况,考虑数据整理的时间和工作量关系,本文从第二类中选取北京和河南作为目标省份,从第三类中选取陕西和天津作为目标省份。这四个省区基本涵盖了东、中、西三个区域。

(4)数据处理。考虑地区 GDP、财政收入、用电量、货运量等 4 个指标既有价值量,又有实物量,本文对价值量的指标如 GDP、财政收入作了不变价处理。四省(区)缩减后指标数据详见附录 1 。

2.4 模型的构建——基于稳健 MM 估计方法

2.4.1 模型的构建及估计方法

通过绘制散点图及分析相关系数发现,地区的 GDP 与地区税收收入、地区用电量、地区货运量之间存在较强的线性关系,建立如下回归模型:

$$Y_t = \alpha_0 + \alpha_1 X_{1t} + \alpha_2 X_{2t} + \alpha_3 X_{3t} + \varepsilon_t \tag{1}$$

其中 Y_t 是地区 GDP 收入,X_{1t} 是财政收入,X_{2t} 表示用电量,X_{3t} 表示货运量,ε_t 为随机误差项。

在对上述模型的参数估计中,传统的统计回归方法都是不稳健的,很容易受到数据集中少数异常值的影响。当数据集中包含有异常值时,使用普通最小二乘

回归方法会出现两种后果：一是多变量估计得不到正确的结果；二是根据拟合得到模型残差不能检测出所有的异常值。

本文所选用的指标属于小样本数据，考虑这种情况，本文拟采用稳健的统计方法。稳健估计方法比其他估计方法效果要好一些，稳健统计方法可以有效的克服传统统计方法的不足，不仅可以产生较少受到异常值影响的估计结果，而且拟合的残差可以更少偏倚、更突出地给出关于异常值的信息，能更好地识别异常值。

稳健回归方法的目的是使求出的回归估计不受异常值的强烈影响，并且通过稳健回归能更好的识别异常点。回归方法的稳健性特征可以用破坏点(Breakdown point)来评价。破坏点指的是强烈影响估计偏离其“实际”情况的异常值数与估计所包含的点数的比值。某方法的破坏点越高，其稳健性越好，可容忍的异常点数目越多，并且对于稳健性回归而言，应同时不受X与Y两方面异常的影响。OLS估计的破坏点是$1/n$，即只要n个观测值中有一个异常点，它就可以破坏最小二乘估计。稳健回归的另一个重要概念是效率。稳健回归估计的效率指的是稳健方法的误差均方根除以OLS估计的误差均方根，此比值越接近于1，稳健回归方法的效率越高。

在实践中，由Yaohai(1987)提出的稳健MM估计组合了高效率但不稳健的M估计以及高破坏点而低效率的S估计，同时具有高破坏点(50%)和高的效率(在高斯－马尔可夫假设成立的条件下，相当于普通最小二乘估计95%的效率)，成为当今最受欢迎的稳健估计方法之一。因此，本文也使用稳健的MM估计作为异常值诊断的方法，对统计数据可靠性进行评估。

稳健MM估计的基本原理是，首先基于迭代的S估计方法得出稳健的初始估计，然后再由M估计导出回归系数。假设一个因变量Y可由p个独立的自变量$X_1, X_2, \cdots, X_p$的线性组合解释，则对于所有的观测值$(x_{i1}, x_{i2}, \cdots, x_{ip}, y_i)$可由如下多元线性回归模型表示：

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta x_{ip} + \varepsilon_i \qquad i = 1, \cdots, n \tag{2}$$

式(2)中，ε_i是独立同分布的误差项，令$x_i = (x_{i1}, \cdots, x_{ip})$，$\theta = (\beta_0, \beta_1, \cdots, \beta_p)$，使用回归技术可产生$p+1$个回归系数$\hat{\theta} = (\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_p)$，第$i$个观测值的残差定义为$r_i(\hat{\theta}) = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_p x_{ip})$。

系数向量的MM估计$\hat{\theta}_{MM}$定义如下：

$$\hat{\theta}_{MM} = \arg\min_{\theta} \sum_{i=1}^{n} \rho\left(\frac{r_i(\theta)}{\hat{\sigma}^S}\right) \tag{3}$$

式(3)中，$\rho(\cdot)$是满足一定条件的损失函数，$\hat{\sigma}^S$是回归残差r_i的散布度量。求解式(3)可以采用Salibian－Barrera和Yohai(2006)提出的重复加权迭代最小