【观点综述】

统计数据预处理的理论与方法述评

程开明
(浙江工商大学 统计与数学学院,浙江 杭州 310018)

摘要:统计数据预处理是提升数据质量的重要阶段,包括数据审查、数据清理、数据转换和数据验证四大步骤。根据处理对象的特点及每一步骤的不同目标,统计数据预处理可采用的方法包括描述及探索性分析、缺失值处理、异常值处理、数据变换技术、信度与效度检验、宏观数据诊断等六大类。选用恰当的方法开展统计数据预处理,有利于保证数据分析结论真实、有效。

关键词:数据质量;数据预处理;缺失值;异常值;数据诊断
中图分类号:F 222 **文献标识码**:A **文章编号**:1007-3116(2007) 06-0098-06

统计学是对数据进行收集、整理和分析的方法论科学。学界普遍重视对数据收集和数据分析的研究,却相对忽视对数据收集之后、正式分析之前这一中间阶段的研究。该阶段一般被称为统计整理,内容包括统计分组及统计图表的绘制[1]。实际上,它包含着更为广泛的内容,称之为统计数据预处理更为合适。统计数据预处理直接决定着分析数据的质量,影响到统计产品的可信度及以此做出决策的科学性,对涉及的理论和方法进行研究十分必要。基于此,本文从提高数据质量的角度,结合数据的不同特点,开展统计数据预处理的过程、方法体系的综合性述评,以期对后续研究有所借鉴。

一、统计数据预处理的必要性

数据预处理本是数据挖掘中的一个概念,是指在数据挖掘前期,针对海量数据存在噪声数据、空缺数据和不一致数据等问题,所采取的一些步骤,包括数据清理、数据集成和变换、数据归约等几个方面[2]。对数据预处理的概念进行拓展,处理对象既包括微观企业数据、微观调查数据,又包括宏观统计数据,便形成统计数据预处理,其范围更广、内容更丰富,使用的方法也更多。

微观数据可分为统计调查数据和微观主体记录数据两部分。统计调查数据由于调查过程中的工作失误、被调查者不配合、抽样方法选取不当、问卷设计不合理等因素而存在误差。利用信息系统收集到的微观主体记录数据,由于数据录入、转换及数据库链接等过程中的失误,可能出现错误字段、记录重复或缺失等问题。政府统计机构生产的宏观统计数据,也会因人为干扰、体制缺陷等原因而存在数据质量问题。所以,在进行正式的数据分析之前,必须开展统计数据预处理,以便对数据质量进行诊断、评估及提升。

WTO 和数据公布通用系统(GDDS) 的加入,要求中国统计数据的发布程序、标准与国际接轨,对数据的可信度提出更高要求。市场体制下,政府往往根据宏观经济运行状况而采取相应的调控措施,为保证宏观调控的科学性、准确性,反映经济运行的统计数据务必准确、及时。信息化建设使企业的管理决策更加依赖于以企业信息系统为基础的决策支持系统,效果如何又依赖于反映交易状况数据的质量好坏。可见,宏观、微观决策的效果都依赖于数据质量,并且要求越来越高。但中国的宏观和微观数据质量却不容乐观。宏观上,经过统计界的努力,世界银行引用中国政府的宏观统计数据不再进行调整,但国际上对中国一些宏观数据依旧提出众多质疑。

收稿日期:2007-08-29
基金项目:2006 年浙江省教育厅科研计划项目"统计数据质量诊断的方法与应用研究"(20061101)
作者简介:程开明(1975-) ,男,湖北广水人,讲师,博士生,研究方向:统计方法与应用,城市与区域经济。

麦迪森提出中国官方1952～1995年的GDP增长速度高估了四分之一,罗斯基认为1998年中国经济增长速度存在严重高估,实际可能只有2%左右,甚至为负数[3]。国内,人们对政府统计数据也一直存有猜疑,最明显的例子就是质疑各省(区、市)GDP增速为何都明显高于国家统计局公布的全国GDP增速。微观方面,反映企业经营状况的数据质量也暴露出诸多问题,一些上市公司在财务数据上弄虚作假、发布虚假信息;一些商业性调查往往由于样本选择不规范、调查偷工减料、弄虚作假,甚至人为编造数据,数据质量让人不敢相信。

一方面社会各界对统计数据的需求越来越广泛,对数据质量的要求越来越高;另一方面数据质量的现状却不尽如人意。为解决这一矛盾,人们通常从完善统计制度、构建合理的指标体系、健全调查网络、选用恰当的调查分析方法、加强统计执法等方面进行探讨[4],却忽略了统计数据预处理这一过程,缺少对检测、诊断和提升数据质量关键性步骤的研究。

二、统计数据预处理的过程

概括起来,统计数据预处理的过程包括数据审查、数据清理、数据转换和数据验证四大步骤。

(一)数据审查

该步骤检查数据的数量(记录数)是否满足分析的最低要求,字段值的内容是否与调查要求一致,是否全面;还包括利用描述性统计分析,检查各个字段的字段类型、字段值的最大值、最小值、平均数、中位数等,记录个数、缺失值或空值个数等。

(二)数据清理

该步骤针对数据审查过程中发现的明显错误值、缺失值、异常值、可疑数据,选用适当的方法进行"清理",使"脏"数据变为"干净"数据,有利于后续的统计分析得出可靠的结论。当然,数据清理还包括对重复记录进行删除。

(三)数据转换

数据分析强调分析对象的可比性,但不同字段值由于计量单位等不同,往往造成数据不可比;对一些统计指标进行综合评价时,如果统计指标的性质、计量单位不同,也容易引起评价结果出现较大误差,再加上分析过程中的其他一些要求,需要在分析前对数据进行变换,包括无量纲化处理、线性变换、汇总和聚集、适度概化、规范化以及属性构造等。

(四)数据验证

该步骤的目的是初步评估和判断数据是否满足统计分析的需要,决定是否需要增加或减少数据量。利用简单的线性模型,以及散点图、直方图、折线图等图形进行探索性分析,利用相关分析、一致性检验等方法对数据的准确性进行验证,确保不把错误和偏差的数据带入到数据分析中去。

上述四个步骤是一个逐步深入、由表及里的过程。先是从表面上查找容易发现的问题(如数据记录个数、最大值、最小值、缺失值或空值个数等),接着对发现的问题进行处理,即数据清理,再就是提高数据的可比性,对数据进行一些变换,使数据形式上满足分析的需要;最后则是进一步检测数据内容是否满足分析需要,诊断数据的真实性及数据之间的协调性等,确保优质的数据进入分析阶段。

三、统计数据预处理的方法体系

对应于数据预处理的几个步骤,各有不同的处理方法。数据审查阶段主要是对调查数据进行信度、效度检验,利用描述及探索性分析手段对数据进行基本的统计考察,初步认识数据特征;数据清理阶段主要是利用多种插补方法对缺失值进行插补,采用平滑技术进行异常值纠正性平滑;数据转换阶段则根据不同的需要可供选择的方法较多,针对计量单位不同可采用无量纲化和归一化,针对数据层级不同可采用数据汇总、概化等方法,结合分析模型的要求可对数据进行线性或其他形式的变换、构造和添加新的属性,以及加权处理等;数据验证阶段包括确认上述数据准备操作的正确性与有效性,检查数据的逻辑转换是否对数据造成扭曲或偏差,并再次利用描述及探索性分析检查数据的基本特征,对数据之间的平衡关系及协调性进行检验。

(一)描述及探索性分析

描述性统计技术主要是对数据开展频数、描述统计量及列联表分析[1]。频数分析是利用非连续变量的频数表,报告出变量个数、记录数,以及缺失值数等;描述统计量分析主要是计算连续变量的均值、标准差、最小值、最大值、偏度、峰度等统计量,以便检查出超出范围的数据或极端值,譬如被调查者的学历分为1,2,…,6共六个等级,频数表中出现大于6的数据则说明存在错误数据。列联表主要起到交叉分类的作用,从中可轻易地发现逻辑上不一致的数据。譬如在一个列联表中有两个被调查者横行属性是"从未听说过"该产品,而纵栏的属性却是"经

常使用”该产品，则说明数据存在矛盾[5]。

探索性分析利用图形直观地考察数据所具有的特征，反映数据的分布特征、发展趋势、集中和离散状况等，主要包括茎叶图、箱形图，散点图，直方图、折线图、条形图等。茎叶图把观测数据分为茎和叶两部分，使我们认识到数据接近对称的程度、是否有数据远离其它数据、数据是否集中、数据是否有间隙等特征。箱形图有助于直观地描述分布与离散状况，利用最大值、最小值、中位数、上四分位数和下四分位数等几个值反映出数据的实际分布。散点图用于直观地表现两个或多个变量之间有无相关关系，并反映数据的分布、集中、离散状况；直方图也是评估数据分布的常用图示法，P—P 图和Q—Q 图则可用于展示数据是否符合正态分布[6]，还有折线图、饼图、面积图、雷达图等，都从不同侧面直观地反映出数据的特征、趋势。

（二）缺失值处理

缺失数据的产生机制通过探讨缺失数据的出现与目标变量是否有关而界定，如果缺失数据是随机出现，就将缺失数据产生机制定义为可忽略的，如果缺失数据的产生与研究变量有关，则称之为不可忽略的[7]。对缺失数据的处理方法大体可以分为四类：

1. 忽略。若一条记录中有属性值缺失，则将该条记录被排除在数据分析之外。该方法简单易行，但是容易导致严重的偏差，仅适用于含有少量缺失数据的情况。

2. 插补（替代）。基本思想是利用辅助信息，为每个缺失值寻找替代值[8]。具体可采用以下几种策略：（1）使用一个固定的值代替缺失值：所有缺失值用一个常量代替，譬如用字母“N”代替缺失值。当某一属性的缺失值较多，使用此方法可能导致结果出现偏差，也只适合于缺失值不多的情况。（2）使用均值代替缺失值：对同一属性的所有缺失值都用其平均值代替。根据变量特征在简单及加权算术平均数、中位数、众数中选用合适的平均数，尽量使替代值更接近缺失值，减少误差。（3）使用同一类别的均值代替缺失值：对数据按某一标准分类，分别计算各个类别的均值来代替相应类别的缺失值，不同类别的均值可选用不同形式的平均数。（4）使用成数推导值代替缺失值：若同一属性的记录值只有少量几种，可计算各种记录值在该属性中所占比例，并对缺失值同比例赋值，该方法较适合缺失属性为是非标志的情况。（5）使用最可能的值代替缺失值：利用回归分析、决策树或贝叶斯方法等建立一个预测模型，利用模型的预测值代替缺失值。该方法相对复杂，但能够最大程度地利用现存数据所包含的信息。

3. 再抽样。包括以下三种情况：（1）多次访问。对无回答单位进行再次补充调查，尽可能多地获得调查数据。如果缺失数据是在不可忽略机制下产生，由于积极回答者和不积极回答者之间的数量特征有较大差异，多次访问很有必要，且差异越大，访问次数也需相应增加。（2）替换被调查单位。在出现无回答的情况下，为使样本量不低于原设计要求，补救方法之一是实行替换，用总体中最初未被选入样本的其他单位去替代那些经过努力后仍未获得回答的单位，替换时应尽可能保证替代者和被替代者的同质性。（3）对无回答进行子抽样。当后续访问的单位费用昂贵时，子抽样可作为减少访问次数的一种现成方法。

4. 加权调整。基本思想是利用调整因子来调整包含缺失数据所进行的总体推断，将调查设计中赋予缺失数据的权数分摊到已获取数据身上[9]。该方法的前提上缺失数据在可忽略机制下产生，即已获得数据与缺失数据之间没有显著差异，主要用于单位数据缺失情况下的调整。

（三）异常值处理

异常值又称为孤立点，异常值处理的首要任务是检测出孤立点。由于异常值可能是数据质量问题所致，也可能反映事物现象的真实发展变化，所以检测出异常值后必须判断其是否为真正的异常值。李金昌、徐雪琪把检测异常值的方法主要分为三类：统计学方法、基于距离的方法和基于偏离的方法[8]。

1. 统计学方法。首先对源数据假设一个分布或概率模型，然后根据模型采用相应的统计量做不一致性检验来确定异常值。常用的方法是用契比雪夫定理来检测异常值。该方法要求知道数据的分布参数，多数情况下这一条件难以满足，故具有一定的局限性。

2. 基于距离的方法。源数据中数据对象至少有 p 部分与数据对象O 的距离大于 d，则数据对象O 是一个带参数 p 和 d 的基于距离（DB）的异常值，即 DB（p，d），常用的距离是欧几里得距离。

3. 基于偏离的方法。通过检查一组数据对象的主要特征来确定异常值，与给出的描述相“偏离”的数据对象被认为是异常值。

检测出事实上的异常值，接下来还需对异常值进行处理。异常值的处理方法主要是采用数据平滑

技术,按数据分布特征修匀源数据[10]。具体方法包括分箱、聚类、回归等几种:

1.分箱。通过考察"邻居"来平滑异常数据的值,让其分布到一些"桶"或箱中,对于箱中的值可以按箱平均值、中值、或边界值。原理是参考相邻的值,进行局部平滑。

2.聚类。异常值可以被聚类检测,聚类将类似的值组织成群或类,将落在各类集合之外的异常值利用离其最近的类均值替代。

3.回归。通过让数据适合一个函数(譬如回归函数)平滑数据,找出适合数据的数学方程式,来帮助消除噪声。许多数据平滑方法还涉及离散化的数据归约问题。

(四)数据变换技术

数据变换是通过一定的方法将原始数据进行重新表达,以改变原始数据的某些特征,增进对数据的理解和分析。大致包括以下几类:

1.对原始数据重新分类、编码、定义变量和修改变量。对于以下两种情况,有必要将原始数据重新分类或重新编码。一是希望将数据分成更有意义的类别;二是希望将数据合并成更少的几大类别。譬如:调查时询问的是被调查者的具体年龄,实际分析时可将被调查者按年龄分为青少年、中年和老年等几大类。重新定义变量或修改现有变量也经常用到,有时变量间呈现出曲线关系,分析前可能需要利用现有变量定义新的变量,譬如令 $X'=X^2$、$X'=\sqrt{X}$ 或 $X'=\log(X)$ 等。重新规定变量的另一种情况是标准化,目的是为了使不同单位或不同量表的变量在分析中具有可比性。

2.数据的代数运算。当变量间的关系是非线性关系时,有时为了便于模型求解,对数据往往进行一些代数运算,譬如对数、指数、幂运算,当然也可能是多种运算的组合。

3.数据汇总和泛化。对数据进行汇总或合计操作,譬如对日销售额进行汇总可得到月销售额和年销售额;泛化处理则是利用更高层次的的概念取代低层次的数据[11],如街道属性可以泛化到城市、国家等更高层次的属性。

4.属性构造。根据给定的属性(字段),构造新的属性(字段),以更好地理解数据结构和更容易地发现变量间的关系。譬如可以根据"长"和"宽"添加属性"面积",根据"销售量"和"价格"得到"销售额"这样新的属性。

5.加权处理。有时对调查取得的数据,还需进行加权处理,以使样本更具有代表性或是强调某些被调查者群体的重要性[12]。

(五)信度和效度检验

问卷调查通过获取样本信息以推断总体特征,推断结果是否真实、可靠依赖于样本信息的准确性和代表性。如果样本不具有代表性,对总体的推断结果便会失真,因此,必须对样本数据所能达到的正确程度和水平高低作必要的检验,即信度和效度的检验[13]。信度是对调查对象而言的,它主要反映回答前后是否一致,即调查结果的可靠性问题;效度是针对调查统计所要研究的问题而言,主要回答调查工具是否合适,即调查结果的正确性问题。

信度是指调查统计结果的稳定性或一致性,也就是对同一调查对象多次重复进行调查或测量,所得结果的一致程度,可表示为 N 次调查中有多少次是正确的,或每次调查属于正确的概率是多少。信度的度量通常是以相关系数来表示的,又称信度系数,包括重测信度、复本信度、折半信度等,分别可以利用相关分析、计算 α 系数等方法来进行检验。

调查统计资料的效度就是指调查结果反映客体的准确程度,反映出调查问卷本身设计的问题。如果问题设计的科学、合理,能够对调查对象进行很好地测量,那么效度就高,反之,则低。效度具体包括内容、准则和建构三个方面,分别对应内容效度、准则效度和构建效度,可以利用相关分析和因子分析等方法进行检验。

(六)宏观统计数据诊断

数据质量是计量经济模型赖以建立和成功应用的基础条件,确保进行计量分析的宏观数据质量,必须对数据进行严格的诊断。所谓数据诊断是通过适当的理论方法,发现对研究结果的可靠性产生显著不良影响的数据[14]。对于横截面数据的质量诊断主要基于计量模型通过各种诊断统计量来进行,而对于时间序列数据则通过时间序列分析来进行。方法包括:

1.分量指标对总量指标的支撑度判断。选取与总量指标密切相关的分量指标进行多元回归分析,建立相应的模型,测算出分量指标数据所能支撑的总量数值,再将支撑数据与现实数据进行比较。

2.宏观统计数据的因果性分析。如果某个变量的统计数据存在异常,利用与其存在因果关系的变量进行推论,可以得到该变量的真实数据,以对其进行修正。周建指出因果关系检测的方法大致有五种[14]:一是Haugh 和Pierce 提出的相关分析法;二

是Granger 和Sargent 提出的单侧分布滞后的方法；三是Sims 提出的双侧滞后的方法；四是Hsiao 提出的最终预测误差(FPF) 检验因果关系法；五是Hafida 提出的多元自回归移动平均模型方法。

3.各专业数据之间的匹配关系判断。国民经济各指标间存在着一定比例关系，把握主要经济指标的合理数量界限，界定其趋势范围，是检验这些数据质量的关键。利用主要经济指标间的比例关系，能够检测出未来短期内的数据置信程度。譬如利用GDP 增长与通货膨胀的关系、社会总供给与社会总需求的关系、经济增长与能源消耗之间的关系，经济增长与交通运输量之间的关系等，可检验GDP 增长数据的可信度[3]。

4.时间序列的预测值与实际值的比较。以经济指标的现有数据为基础，利用各个经济变量自身发展情况的走势进行最优化模拟，建立相应的时间序列模型，对相应指标进行预测，可得到该指标在理论上应该达到的数值，然后与实际数据相对比，以此评价实际数据与理论值的接近程度。

5.其他手段。包括全面调查与抽样调查的结果相验证，投入产出调查与国民经济核算资料相验证，利用统计执法检查的结果对数据进行调整等。

四、结论及需注意的问题

统计数据预处理是数据收集之后、数据分析之前进行数据质量评估、诊断和提升的重要步骤。从统计数据预处理过程来看，无论是微观数据，还是宏观数据，一般都可以进行描述及探索性分析、异常值和缺失值的处理、数据转换等。当然，随着数据本身质量好坏及数据分析要求的不同，方法的使用各有侧重。针对微观调查数据，还需进行调查数据的信度和效度检验，对宏观统计数据一般利用数据诊断技术进行平衡关系和协调性的检验。

以上各项统计数据预处理方法具有坚实的理论支撑，也有现实可操作性，可利用SAS、SPSS 等统计分析软件来具体实施。实际操作过程中，以下几个问题需加以注意：1. 统计数据预处理必须以目的为导向。即以数据分析的要求为出发点，预处理的目的是提高进入分析阶段的数据质量，保证分析结果客观、有效。2. 依据数据特点选用恰当的预处理方法，且应重点突出。并非每一次统计数据预处理都要对所有步骤进行操作，而应根据研究目的、内容及数据特点，选用合适的方法和步骤。3. 统计数据预处理必须与数据收集、数据分析的方法相结合。统计数据预处理只能对已有数据进行一定程度的检测、诊断和提升，尽量不让坏数据进入到分析阶段。要从根本上提高数据质量，还须数据收集阶段的方法得当，加强质量监控等；合理地选取数据分析方法，也是保证分析结论真实、有效的必要条件。

参考文献：

[1] 贾俊平.描述统计[M].北京：中国人民大学出版社，2003：115－122.

[2] TOM Soukup IAN Davidson.可视化数据挖掘——数据可视化和数据挖掘的技术与工具[M].朱建秋，等，译.北京：电子工业出版社，2004：59－115.

[3] 岳希明，张曙光，等.中国经济增长速度研究与争论[M].北京：中信出版社，2002：3－51.

[4] 李金昌.论什么是统计数据质量[J].统计与决策，1998(9)：6－8.

[5] 柯惠新，丁立宏.市场调查与分析[M].北京：中国统计出版社，2000：196－201.

[6] SOIBELMAN L M，HYUNJOO Kim.Data Preparation Process for Construction Knowledge Generation through Knowledge Discovery in Databases[J].Journal of Computing in Civil Engineering，2002(1)：39－48.

[7] RODERICH J A Little，RUBIN Donald B.缺失数据统计分析(中文版)[M].孙山泽，译.北京：中国统计出版社，2004：3－16.

[8] 李金昌，徐雪琪.数据挖掘质量问题探讨[J].统计研究，2004(7)：49－52.

[9] CARRIERE K C.Methods for Repeated Measures Data Analysis with Missing Values[J].Journal of Statistical Planning and Inference，1999(7)：221－236.

[10] SMOLINSKI，WALCZAK，EINAX J W.Exploratory Analysis of Data Sets with Missing Elements and Outliers[J].Chemosphere，2002(49)：233－245.

[11] PYLE Dorian.Data Preparation for Data Mining[M].Paperback，Bk &CD edition，1999：89－190.

[12] ZAFFALON Marco.Exact Creedal Treatment of Missing Data[J].Journal of Statistical Planning and Inference，2002(105)：

105—122.
[13] 隗斌贤,程利仲.论统计调查质量评价的信度与效度[J].浙江统计,1997(8):18—20.
[14] 周建.宏观经济统计数据诊断理论、方法及其应用[M].北京:清华大学出版社,2005:13—172.

(责任编辑:马 慧)

The Theory and Methods of Data Preparation: An Overview

CHENG Kai-ming
(School of Statistics and Mathematics, Zhejiang Gongshang University, Hangzhou 310018, China)

Abstract: In order to improve the quality of data for analyzing, data must be prepared. Data preparation can be decomposed to four steps such as data examination, data cleaning, data transformation and data validation. The methods of data preparation include descriptive and exploratory analysis, missing data analysis, outlier processing, transformation techniques, reliability and validity analysis, and national economic data diagnosis. Data preparation can be operated by software and some possible problems must be noticed.

Key words: data quality; data preparation; missing data; outlier; data diagnosis

(上接第81页)

[10] KOOPMANS Tjalling C. On the Concept of Optimal Economic Growth[J]. In The Economic Approach to Development Planning. Amsterdam: Elsevier, 1966:324—55.
[11] BARRO Robert J, SALA-I-MARTIN Xavier. Economic Growth[M]. The MIT Press, Cambridge, Massachusetts, London, England, 1995:155—234.
[12] WOOLDRIDGE Jeffrey M. Econometric Analysis of Cross Section and Panel Data[M]. The MIT Press, Cambridge, Massachusetts, London, England, 2002:435—475.
[13] WOOLDRIDGE Jeffrey M. Introductory Econometrics: A modern Approach (Second Edition)[M]. Tsinghua University, 2004:425—481.

(责任编辑:张治国)

Constant or Increase Returns to Scale: An Empirical Study on China and Five OECD Countries

WANG Jun-hui, HOU Fang-yu
(Research Institute for Fiscal Science, Ministry of Finance, Beijing 100036, China)

Abstract: The constant or increase returns to scale is the assumption of many economic theoretical models. This paper analyzes the issue of returns to scale under the normal economic growth theory. Returns to scale is also studied on economy of China from the year of 1978 to 2004. Panel Data model is taken to study on it of five OECD countries. The result indicates that the assumption of constant returns is not suitable for the modern economy. This is one of the reasons that many theoretical models are encountered difficulties.

Key words: constant returns to scale; increase return to scale; empirical study