·科研方法专题·

回归建模的基础与要领(Ⅰ)
——统计模型种类的划分方法

胡良平[1,2*]

(1. 军事科学院研究生院,北京 100850;
2. 世界中医药学会联合会临床科研统计学专业委员会,北京 100029
* 通信作者:胡良平,E-mail:lphu812@sina.com)

【摘要】 本文目的是介绍如何划分“统计模型的种类”。首先,介绍与“统计模型”有关的基本概念;其次,陈述从“不同角度划分统计模型”的构想;最后,分别基于“统计特性”“统计功能”和“预测结果”三个角度,给出更具有实际意义的“统计模型分类结果”。得到的结论是:基于“预测结果”划分统计模型,可以获得“最少种类”的划分结果;而且这种划分方法对实际工作者选择合适的统计模型具有更直接的指导作用。

【关键词】 一元与多元多重回归模型;线性与非线性回归模型;一水平与多水平回归模型;参数、半参数与非参数回归模型;非概率与概率回归模型

中图分类号:R195.1 文献标识码:A doi:10.11886/j.issn.1007-3256.2018.06.001

The basis and essential of the regression modeling (Ⅰ)
——the partition method of the variety of the statistical models

Hu Liangping[1,2*]

(1. *Graduate School, Academy of Military Sciences PLA China, Beijing* 100850, *China*;
2. *Specialty Committee of Clinical Scientific Research Statistics of World Federation of Chinese Medicine Societies, Beijing* 100029, *China*
* *Corresponding author: Hu Liangping, E-mail: lphu812@sina.com*)

【**Abstract**】 The purpose of this paper was to introduce how to divide the variety of the statistical models. Firstly, the basic concepts with respect to “statistical model” were presented. Secondly, the idea of dividing the statistical models from different angles was stated. Lastly, the more practical significance differentiate results of the statistical models were given by means of the following three aspects: ①statistical characters; ②statistical functions; ③predictive results. Based on the “predictive results” to divide the statistical models, the partition results with the least varieties were acquired. Furthermore, this partition method mentioned above was of more direct role of the guidance to select the rational statistical models for the practical workers.

【**Keywords**】 Univariate and multivariate multiple regression models; Linear and nonlinear regression models; One level and multilevel regression models; Parametric, semi-parametric and non-parametric regression models; Non-probability and probability regression models

1 统计模型概述

1.1 统计模型的概念

1.1.1 模型的概念

人们经常提及两类模型,即数学模型与统计模型。那么,首先要知道什么是“模型”。笼统地说,“模型”就是描述一个单一变量或向量如何随另一个变量或向量变化而变化的依赖关系的表达式或“函数”或“方程式”。当“模型”揭示的是“总体”中变量之间的关系时,称其为“模型”更恰当;而当“模型”揭示的是“样本”中变量之间的关系时,称其为“方程”更恰当。所谓“更恰当”是指:当表达式中带有“随机误差项”时,表达式呈现的是变量之间的“精确”数量关系;而当表达式中不带有“随机误差项”时,表达式呈现的是变量之间的“近似”数量关系。

1.1.2 数学模型的概念

“数学模型”是描述确定性事物或现象之间数量关系的表达式。换言之,它是一个“函数”,即给定自变量一个特定取值,因变量就有一个确定的值与其对应。事实上,可以这样认为:数学模型描述的是一般变量之间的数量依赖关系。

1.1.3 统计模型的概念

“统计模型”是描述随机变量随其他随机变量或随机过程或一般变量变化而变化的依赖关系的表

项目基金:国家高技术研究发展计划课题资助(2015AA020102)

达式或“方程式”或“模型”。事实上，可以这样认为：在统计模型中，因变量或是随机变量、或是随机变量的函数（被称为随机过程）；而自变量可以是一般变量、随机变量或随机过程。

在经典统计模型中，假定总体上的回归系数（含截距项）为常量，基于样本信息构建的样本回归系数（含截距项）被视为总体回归系数的估计值；而在贝叶斯统计模型中，假定总体上的回归系数（含截距项）是随机变量，通常，需要通过总体信息、样本信息和先验信息（有时还需借助随机模拟）来推断回归系数的估计值。

1.2 从不同角度划分统计模型[1-4]

1.2.1 概述

统计模型不计其数，如何对其进行分类呢？事实上，从不同的角度来考量，就会有不同的分类结果。显然，这样分类的结果之间具有“交叉重叠”现象；然而，这或许是引导读者认识“统计模型”的最简易、最直接的思路或方法。

1.2.2 基于统计思想分类

基于统计思想可分为经典统计模型（可进一步划分为“参数统计模型”“半参数统计模型”和“非参数统计模型”）、贝叶斯统计模型、蒙特卡罗（随机模拟）统计模型和机器学习统计模型。

1.2.3 基于统计模型是否有解析式分类

基于是否有解析式可分为有解析式的统计模型（绝大部分统计模型都属于这一类）与无解析式的统计模型（机器学习和深度学习方法建模基本上属于这一类，还有所谓的“概率图模型”）。

1.2.4 基于统计功能分类

基于统计功能可分为广义差异性分析模型、相关与关联分析模型、回归分析模型、判别分析模型、聚类分析模型、综合评价模型和多元统计分析模型（包括通径分析模型、主成分分析模型、探索性与证实性因子分析模型、结构方程模型、典型相关分析模型、对应分析模型、多维尺度分析模型、结合分析模型等）。

1.2.5 基于模型的个数分类

基于模型个数可分为单一结局变量的统计模型（或称为一元统计模型）与多结局变量的统计模型（或称为联立方程组模型）。

1.2.6 基于模型的水平数分类

基于模型的水平数可分为单一水平统计模型（即通常的统计模型）与多水平统计模型（也称为随机系数统计模型）。

1.2.7 基于因变量与自变量之间的几何关系分类

基于因变量与自变量之间的几何关系可分为一般线性与非线性统计模型、广义线性与非线性统计模型。

1.2.8 基于回归系数的效应关系分类

基于回归系数的效应关系可分为固定效应统计模型、随机效应统计模型与混合效应统计模型。

1.2.9 基于时间变量分类

基于时间变量可分为时点统计模型（包括所有不以“时间”为自变量的统计模型或与“时间”无关的统计模型）与时序统计模型（包括各种线性与非线性时间序列统计模型、Cox 比例风险与非比例风险回归模型、生存资料的各种参数模型、纵向追踪或称为重复测量设计混合效应统计模型）。

1.2.10 基于因变量是否为“显变量”分类

在常规的“回归分析”中，在“经典统计思想和贝叶斯统计思想”框架下，人们所讨论的统计模型中的因变量基本上都是“显变量”或由“显变量变换所得到的结果”；在很多多元统计分析中，很少采用“统计模型”去描述所获得的最终结果，而是采用“典型变量”或“主成分变量”等去描述。本质上，它们就是以“隐变量”为因变量的“统计模型”。具体地说，在典型相关分析中，采用“显变量”来线性表达“典型变量（本质上就是隐变量）”。一个“典型变量对”就是一个“二元多重线性回归模型”或视为由两个“一元多重线性回归方程（注意：因变量为隐变量）”组成的回归方程组；假定在所研究的问题中，有 m 个“显变量（即定量结果变量）”，于是，在主成分分析中，用“显变量”的不同线性组合分别表达 m 个“主成分变量（本质上就是隐变量）”，实际上，全部 m 个主成分表达式就是由 m 个“一元多重线性回归方程（注意：因变量为隐变量）”组成的回归方程组；在探索性因子分析中，“因子得分模型”也是由 m 个“一元多重线性回归方程（注意：因变量为隐变量）”组成的回归方程组；同理，在定量资料对应分析（有公因子变量）、多维尺度分析（有公因子变量）和变量聚类分析（有类成分变量）中，都有“以隐变量为因变量”的统计模型。

1.2.11 基于统计模型中是否包含“未知参数”分类

一般来说，统计模型中会包含“未知参数”。然而，若按上述“基于因变量是否为‘显变量’来划

分”，“广义差异性检验”可被视为“基于概率分布”的“统计模型”，因为检验统计量，如 Z、t、F、χ^2 等，都可被视为“隐变量”，通过相应的“概率分布”把握其变化规律，而基于“样本信息”提取的是一般统计量，如样本均值、标准差、样本含量、观察频数与理论频数等，它们并不包含“未知参数”。由此可知，基于某种概率分布的“检验统计量”应属于“最简单的统计模型”，其他统计模型可被概括为反映“依赖关系的统计模型”。

1.2.12 基于统计模型是否为“最终模型”分类

若模型本身就是最终要求的模型，则该模型应被称为“目标模型”；若模型本身只是在计算过程中起一个“桥梁”作用，通过它来获得最终要求的模型中“未知参数”的估计值，则该模型可被称为“过程模型”。

事实上，所有以“检验统计量为别名的统计模型（它们在统计学教科书上被称为‘检验统计量’）”和反映“变量间依赖关系的统计模型”都是研究者希望构建的、具有解析式的统计模型，故它们都属于“目标模型”；而为了求解“目标模型”中的“未知参数”，需要先构造一个“目标函数”，再依据某种原则（如最小平方法或最大似然法）经由“目标函数”导出一个“正规方程组”或直接构建一个“广义估计方程组”，进而求出“目标模型”中的未知参数。为后续指代方便，不妨把“正规方程组”或“广义估计方程组”都统称为“过程模型”。

2 有解析式的广义统计模型的种类

2.1 概述

前面“从不同角度划分统计模型”给出了 11 种具有“交叉重叠”的分类结果，为读者了解和认识“统计模型”奠定了必要的基础。下面，再分别基于“统计特性”“统计功能”和“预测结果”三个角度，给出更具有实际意义的“统计模型分类结果”。其中，基于“统计特性”划分统计模型，其种类最多，而且，其数目会随着所找出的“统计特性”的数目增加而成倍增加；而基于“预测结果”划分统计模型，其种类最少，或许也是最有实用价值的分类方法。

2.2 基于“统计特性”对统计模型进行分类

根据同时考察模型是否具有下列 9 种“统计特性”（说明：事实上，可能还存在其他统计特性，此处归纳的仅是最常见的），可将统计模型归纳为 1 152 大类。9 种“统计特性”分别指“模型的水平数（2 种情况）、因变量的个数（2 种情况）、因变量的性质（3 种情况）、自变量的个数（2 种情况）、是否含隐变量（2 种情况）、是否考虑抽样权重（2 种情况）、因变量观测值是否独立（2 种情况）、因变量与自变量前回归系数是否为线性关系（2 种情况）以及是否基于‘参数’构建模型（3 种情形）”，于是，统计模型可被分解为以下 1 152 类，现概述如下：①模型的水平数（2 种情况）指“一水平模型”与“多水平模型”；②因变量的个数（2 种情况）指“一个因变量或称一元模型”与“多个因变量或称多元模型”；③因变量的性质（3 种情况）指“计量因变量”“计数因变量”和“定性因变量”；④自变量的个数（2 种情况）指“一个自变量或称一重模型”与“多个自变量或称多重模型”；⑤是否含隐变量（2 种情况）指“不含隐变量”与“含隐变量”；⑥是否考虑抽样权重（2 种情况）指“不考虑抽样权重”与“考虑抽样权重”；⑦因变量观测值是否独立（2 种情况）指“相互独立”与“相依（如‘时间序列资料’与‘具有重复测量的资料’）”；⑧因变量与自变量前回归系数是否为线性关系（2 种情况）指“线性”与“非线性”；⑨是否基于“参数”构建模型（3 种情况）指“参数法”“半参数法”和“非参数法”。

将上述 9 种“统计特性”全面组合起来构建统计模型，就有 $2^7 \times 3 \times 3 = 1\ 152$ 类。

2.3 基于“统计功能”分类

基于“统计功能”对统计模型进行分类，至少可以划分为以下 7 类：①差异性分析的线性模型；②相关分析模型；③关联分析模型；④回归分析模型；⑤判别分析模型；⑥聚类分析模型；⑦多元统计模型。

2.4 基于“预测结果”分类

2.4.1 概述

基于统计模型的“预测结果”划分统计模型的种类，可将统计模型划分为以下 4 类：①观测结果的预测值；②观测结果的概率值；③观测结果的综合值；④观测结果的统计量。

2.4.2 基于“观测结果的预测值”划分统计模型

何为“观测结果的预测值”？由模型计算的结果为观测结果 Y 的预测值，两者的属性和单位完全相同。例如：①计量资料线性与非线性回归分析模型；②时序资料线性与非线性时间序列分析模型；

③通径分析或路径分析模型。

其中，“计量资料线性与非线性回归分析模型”包括一般线性与非线性回归分析模型、主成分回归分析模型、岭回归分析模型、基于正交化方法的回归分析模型、稳健回归分析模型、反应曲面回归分析模型、分位数回归分析模型、加性与广义加性回归分析模型、局部模型回归分析和有限混合模型回归分析模型等。

2.4.3 基于“观测结果的概率值”划分统计模型

何为“观测结果的概率值”？由模型计算的结果为观测结果 Y 取某特定值（对离散型随机变量而言）或某个小的取值区间内的值（对连续型随机变量而言）的概率，两者的属性和单位完全不同。例如：①生存资料回归分析；②计数资料回归分析；③定性资料回归分析。

2.4.4 基于“观测结果的综合值”划分统计模型

何为“观测结果的综合值”？由模型计算的结果为观测结果 $Y_1 - Y_k$ 的综合值，前者为隐变量、后者为显变量。例如：①主成分分析模型；②因子分析模型；③结构方程模型；④对应分析模型；⑤多维尺度分析模型；⑥典型相关分析模型；⑦结合分析模型；⑧判别分析模型；⑨经典综合评价模型。

其中，“经典综合评价模型”包括三十多种方法，主要有如下几种，即熵值法、Topsis 法、秩和比法、基于标准化变换的求和法、投影寻踪法、模糊综合评价法和层次分析法等[5]。

2.4.5 基于“观测结果的统计量”划分统计模型

何为“观测结果的统计量”？由模型计算的结果为“检验统计量”的值，它是由观测结果 Y 的一般统计量构造出来的检验统计量。例如：①Z、t、F、χ^2、W 等；②T^2、$Wilks'\ \lambda$ 等。

值得一提的是：对于最后一种分类结果，人们通常并不认为它们是“统计模型”，而认为它们只是假设检验的“检验统计量”。事实上，在统计学上，可以认为：一般线性模型包含了“假设检验”，或者说，假设检验属于“统计模型”的“特例”。

参考文献

[1] Ramon C. Littell, Rudolf J. Freund, Philip C. Spector. SAS System for Linear Models [M]. 3rd ed. Cary, NC: SAS Institute Inc, 1991: 1 - 292.

[2] Ramon C. Littell, George A. Milliken, Walter W. Stroup, et al. SAS System for Mixed Models [M]. Cary, NC: SAS Institute Inc, 1996: 1 - 490.

[3] David A. Freedman. 统计模型理论和实践 [M]. 吴喜之，译. 北京：机械工业出版社，2010：42 - 147.

[4] 杨珉，李晓松. 医学和公共卫生研究常用多水平统计模型 [M]. 北京：北京大学医学出版社，2007：6 - 186.

[5] 郭春雪，沈宁，胡良平. 基于标准化变换的求和法：一种新的样品聚类分析方法 [J]. 四川精神卫生，2017，30(3)：211 - 216.

（收稿日期：2018 - 12 - 12）

（本文编辑：唐雪莉）

科研方法专题策划人——胡良平教授简介

胡良平，男，1955 年 8 月出生，教授，博士生导师，曾任军事医学科学院研究生部医学统计学教研室主任和生物医学统计学咨询中心主任、国际一般系统论研究会中国分会概率统计系统专业理事会常务理事和北京大学口腔医学院客座教授；现任世界中医药学会联合会临床科研统计学专业委员会会长、中国生物医学统计学会副会长，《中华医学杂志》等 10 余种杂志编委和国家食品药品监督管理局评审专家。主编统计学专著 48 部，参编统计学专著 10 部；发表第一作者学术论文 260 余篇，发表合作论文 130 余篇，获军队科技成果和省部级科技成果多项；参加并完成三项国家标准的撰写工作；参加三项国家科技重大专项课题研究工作。在从事统计学工作的 30 年中，为几千名研究生、医学科研人员、临床医生和杂志编辑讲授生物医学统计学，在全国各地作统计学学术报告 100 余场，举办数十期全国统计学培训班，培养多名统计学专业硕士和博士研究生。近几年来，参加国家级新药和医疗器械项目评审数十项、参加 100 多项全军重大重点课题的统计学检查工作。归纳并提炼出有利于透过现象看本质的“八性”和“八思维”的统计学思想，独创了逆向统计学教学法和三型理论。擅长于科研课题的研究设计、复杂科研资料的统计分析与 SAS 实现、各种层次的统计学教学培训和咨询工作。