·继续教育园地·

多变量回归模型分析应用概述

于石成　亓晓　胡跃华　郑文静　王琦琦　么鸿雁
中国疾病预防控制中心流行病学办公室，北京 102206
通信作者：么鸿雁，Email：yaohy@chinacdc.cn，电话：010-58900522

【摘要】 多变量回归模型分析在医学研究中的应用非常广泛。本文从实际应用的角度出发，介绍常用的多变量回归分析方法：多重线性回归、logistic 回归、Poisson 回归和 Cox 比例风险回归模型，内容包括多变量回归模型的应用条件、分析步骤、自变量筛选策略、模型扩展讨论和应用注意事项。以期读者对多变量回归分析有所了解，在科研工作中能正确使用多变量回归模型分析，提高数据使用效率和统计分析水平。

【关键词】 回归分析；　模型，统计学；　概述
基金项目：国家重点研发计划“数据驱动的慢性病防控策略及应用研究”（2018YFC1315305）
DOI:10.3760/cma.j.issn.0253-9624.2019.03.020

Overview of multivariate regression model analysis and application

Yu Shicheng, Qi Xiao, Hu Yuehua, Zheng Wenjing, Wang Qiqi,Yao Hongyan
Office of Epidemiology, Chinese Center for Disease Control and Prevention, Beijing 102206, China
Corresponding author: Yao Hongyan, Email:yaohy@chinacdc.cn,Tel: 0086-10-58900522

【Abstract】 Analyses of the multivariate regression model are ued very widely in the medical research. Analytical methods of the mutivariate regression model including multiple linear regression, logistic regression, Poisson regression and Cox proportional hazard model were introduced in this article. The contents of the article covered the application conditions of regression models, analytical procedures, strategies of selecting independent variables, extended discussions of regression models and application notes. It is expected that authors could understand the principle of the mutivariate regression model, accurately use these analytical methods in their research, improve the efficiency of data utilization, and enhance the level of statistical analyses.

【Key words】 Regression analysis;　Models, statistical;　Overview
Fund program: National Key Research and Development Programme "The Data-driven Prevention and Control Strategy of Chronic Diseases and Applied Study(2018YFC1315305)
DOI:10.3760/cma.j.issn.0253-9624.2019.03.020

一、为什么使用多变量回归分析？

在医学研究中，经常要确定研究因素与疾病结果间的关联，关联有统计学关联和因果关联。统计学关联指因素与疾病结果间的关联不是由于随机误差或偶然性因素引起的，但不一定是因果关联，可能为生态学关联、间接关联或虚假关联。因果关联是指疾病结果由某一因素引起的，包括直接和间接因果关系。确定因素与疾病结果间因果关系强弱，最重要的是研究设计而非统计方法。因果关系确定是复杂的，对于慢性病目前普遍接受的病因理论是一果多因论。探讨致病因素的独立作用、联合作用以及他们的作用机制是病因研究的重要内容，其中，多变量回归分析起到了非常重要的作用。因为有混杂因素的存在，在分析因素与结果变量间的关联时，需要采用多变量分析技术，在控制多个混杂因素的影响后来阐述因素与疾病结果间的关联。

二、常用的多变量回归分析方法及选择

多变量回归分析是定量估计变量间关系的统计过程，研究多个自变量与一个因变量关联的统计规律。20 世纪30年代多变量回归分析的理论已发展起来，但由于计算复杂，实际应用不多。1970 年以后，由于计算机的发展和普及，多变量回归分析技术得到广泛的应用；进入 80 年代后期，计算分析的统计软件包迅速发展，使得多变量回归分析的应用更为普及。目前，常用的多变量回归分析包括多重线性回归、logistic 回归、Cox 比例风险回归和 Poisson 回归。“多重”是指多个自变量，“多元”是指多个因变量，在使用时注意区分这两个术语。

根据结果变量的类型选取回归模型，结果变量为连续变量（血压、胆固醇、肺活量）选用多重线性回归，如在一现况调查中，要探讨性别、年龄、胆固醇、BMI、腰围与血压的关联，可使用多重线性回归，因为因变量血压为连续变量；结果变量为二分类变量（糖尿病：是/否；肝癌：是/否）选用 logistic 回归；结果变量为计数资料（跌倒性伤害发生次数、哮喘发作次数）选用 Poisson 回归，结果变量为 0、1、2、3 …，自变量为要研究的影响跌倒发生的因素；生存分析资料，结果变量为二分类，

但有明确的生存时间，此时使用Cox比例风险回归。

多变量回归分析的自变量可以为连续变量（年龄、收入、身高），分类变量（职业、血型、婚姻状况）和/或等级变量。根据变量类型和专业要求确定自变量以何种形式纳入模型，分类变量必须以哑变量的形式纳入模型。多个自变量如何引入模型遵循自变量纳入回归模型的策略，见后面的叙述。

三、多变量回归分析步骤及自变量筛选策略

多变量回归的拟合步骤包括分析方法的选择、确定自变量纳入模型的形式、模型应用条件检验、单因素分析、多因素分析、模型诊断及评价、模型修改、确定最终模型和得出统计结论。

（一）回归方法的选择

统计方法的确定是统计分析的关键，在课题设计或进行数据分析时，根据研究目的及收集数据的特点，选择统计方法。通常情况下的多变量回归分析，据因变量的类型确定分析方法，如前所述。

（二）确定自变量纳入模型的形式

自变量有三种类型：连续变量、分类变量和等级变量。在拟合模型前，对每个自变量进行统计描述。连续变量描述其集中趋势和离散趋势，进行正态性检验，如果资料符合正态分布，用均数、标准差、变异系数等指标描述其集中趋势和离散趋势；如资料为偏态，用中位数、四分位数间距和极差描述。分类变量描述其频数、频数百分比、累积频数和累计频数百分比等指标。根据资料的特点，等级资料可按定量或定性资料进行统计描述。

连续变量可以原变量纳入模型或以分组线性纳入模型，如年龄可直接纳入模型或将年龄分为几个年龄组，以分组线性的方式纳入模型，但这两种形式都要满足与因变量（变换的因变量）有线性关系的假设。分类变量必须以哑变量的方式纳入模型，等级变量可以分组线性或哑变量的方式纳入模型，如以分组线性纳入模型，其也要满足与因变量（变换的因变量）有线性关系的假设。

（三）模型应用条件检验

拟合多变量回归要对模型应用条件进行检验，多重线性回归要满足线性、独立、正态和方差齐性；logistic回归要满足独立和线性的要求：独立指因变量间独立，线性指连续自变量与$\ln[P/(1-P)]$为线性关系；Poisson回归要满足独立和线性的要求，线性指连续自变量与$\log_e(y)$成线性关系；Cox比例风险模型满足比例风险的假设。

（四）单因素分析

将每个自变量与因变量的关系进行单因素分析，可了解因变量的影响因素。通常情况下选取在单因素分析中有统计学意义的变量进行多因素分析。有些时候，如果认为某个或某些变量很重要，其致病的生物学意义清楚或以前的研究认为是危险因素，即使在单因素分析中无统计学意义，也可作为协变量纳入模型。模型纳入无统计学意义的变量可造成模型的拟合效果不好，因此模型纳入无统计学意义的自变量不能太多。

（五）多因素分析

单因素分析显著的变量以及需要分析的协变量如何纳入模型，即自变量的筛选策略。一般采取三种策略：

1.关注研究变量的策略：这种策略是有明确的要关注的研究因素，如某一新的危险因素，单因素分析之后，将关注的因素加上要控制的混杂因素一起纳入模型。这里感兴趣的是研究因素，检验研究因素的统计学意义，其他变量是调整变量，其统计学意义不是我们关注的。

2.逐步回归选择法：逐步回归多用于变量的筛选，开始时模型中无任何自变量，然后按自变量对因变量的贡献大小依次将其引入方程。每引入一个变量，对已在模型里的变量进行逐个检验，如果无统计学意义时，将其剔除。每一次引入或剔除都要进行统计学检验，以保证之前模型中所有变量都有统计学意义。反复进行这个过程，直到没有统计学意义的变量引入，模型中也没有不具有统计学意义的变量为止。

3.最佳模型组合筛选法：单因素分析确定有8个因素要进行多变量分析，先以关注变量与其他7个变量分别组成有2个自变量的7个回归模型，按模型的拟合优度判断出一个最佳模型；以有2个自变量的最佳回归模型与其他6个变量分别组成有3个自变量的6个回归模型，按模型的拟合优度判断出一个最佳模型；依次进行下去，直到没有统计学意义的变量进入模型为止。这里需要指出的是，不但考虑拟合优度判断模型的拟合好坏，更重要的是在进行模型组合时，考虑变量的专业意义。最后可以将各种最佳模型组合结果都呈现出来，详细观察研究因素是如何受协变量的影响。

（六）模型诊断与评价

按上述步骤建立起来的模型为初步模型，还不清楚这个模型是否较好地揭示了自变量与因变量之间的关系，以及是否符合实际情况，因此须对模型进行诊断与评价。模型诊断与评价包括统计学和专业评价。根据拟合模型提供的统计学指标评价模型的拟合优度，如模型拟合优度不佳，需要考虑数据质量是否存在问题，或数据是否存在异常点、多重共线性等问题。如果自变量不足以解释因变量的变异时，还需要考虑选用的模型是否合适，是否存在交互作用等问题。专业评价指拟合的模型要符合实际和专业知识，统计中允许多个模型的存在，因为很多疾病的机制不清，可能存在多种解释。能够被专业知识合理解释的模型才是一个好的模型。

（七）修改模型和最终确定模型

得到统计学有意义和专业上能解释的模型，实为修改模型的过程。在这个过程中，可以加入交互作用项，以增加模型的拟合优度。

四、多变量回归分析的扩展

（一）多重线性回归的扩展

多重线性回归强调连续自变量与因变量要呈线性关系，可以横坐标为自变量，纵坐标为因变量图示，检验线性关系是否成立。如果呈现非线性，在有一个拐点的情况下，可在模型中加入二次项；有两个拐点的情况下，可在模型中加入三次项，此为多项式回归。如年龄（age）与血压呈先升高再下降，有一个拐点，可在模型中加入二次项，即将年龄取平方（age*age）后的变量纳入模型，同时注意也应将age

纳入模型。当不同自变量或自变量组合所对应的残差随着变量值增加而变大或减少时，即方差不齐，可考虑加权最小二乘回归。自变量存在共线性时，采用偏最小二乘回归处理。稳健回归是一类方法总称，主要是针对异常值的处理。检测异常点并在有异常点的情况下给出模型的稳健估计。当模型残差不满足正态性时，常规的最小二乘回归容易导致结果的偏倚，此时可以考虑采用分位数回归，分别描述不同分位数下自变量对因变量的影响情况。

（二）logistic 回归的扩展

一般 logistic 回归指非条件 logistic 回归，队列研究、现况研究和成组病例-对照研究资料用非条件 logistic 回归分析。配对或配比的病例-对照研究，使用条件 logistic 回归分析资料。当因变量为分类或等级变量时，可分别采用多分类和有序结果 logistic 回归。

1. 多分类结果 logistic 回归：有时因变量是多分类的，如肿瘤的 TNM 分期，病例-对照研究中的一个对照有两个或多个病例组，这些因变量都是多分类结果变量，如用前述二分类结果 logistic 回归模型，可能会增加犯Ⅰ类错误的概率，要用多分类结果 logistic 回归模型处理。如因变量 y 为体重，按肥胖的严重程度分三类：y=2（肥胖）、y=1（超重）和 y=0（正常体重）。设 y=0 为参考组，则 y=1 与 y=0 相比，y=2 与 y=0 相比的三分类结果 logistic 回归。多分类 logistic 回归可回答：与正常体重组相比，研究因素对超重和肥胖组的效应是否相同？综合效应如何？这是上述分别拟合两个二分类 logistic 回归不能回答的问题。

2. 有序结果 logistic 回归：多分类 logistic 回归的结果变量是分类变量，如血型、职业或病例对照组，有些结果变量是有序资料，如跌倒性伤害严重程度（轻度、重度、重度）或血尿严重程度（-、+、++、+++、++++）。这类结果变量可用多分类 logistic 回归，把结果变量按分类资料来处理，还可以用有序结果 logistic 回归处理，包括有序结果的累积优势比 logistic 回归和有序结果的相邻优势比 logistic 回归两种模型。

（三）Poisson 回归的扩展

Poisson 回归的因变量为计数资料，自变量可为连续变量、分类变量和等级变量。连续变量或等级变量要检验其与 $\log_e(y)$ 的线性关系，分类变量要用哑变量纳入模型。当采用 Poisson 回归模型时，Deviance/Pearson 卡方值与自由度的比值>1，为过度离散，其后果是标准误估计不准确。原因为：数据间的变异很大，Poisson 回归不能很好地描述计数资料；也意味着观测可能不独立；实验条件未能很好地控制，未知参数的变化不仅随测量的协变量变化也随未控制的协变量和潜变量变化。此时可使用过离散参数校正，也可改用负二项回归，负二项回归模型可处理资料不独立造成的过度离散，如具有传染性、地方性和家庭聚集性的疾病。

（四）Cox 比例风险回归模型的扩展

Cox 比例风险回归模型假定所有预后因素的作用强度在所有时间上都保持一致，即具有某预后因素的病人的死亡风险和不具有该因素的病人的死亡风险在所有的时间上都保持一个恒定的比例。在应用时要对这一条件进行检验，如果变量不满足比例风险的假设，可采用下列方法处理非等比例的情况：分层：将非等比例的变量分层，但作为分层的变量将无法估计其效应，因此一般只用于协变量；分段估计：从交叉点处划分成多个区间，在每个区间内是等比例的，分别对每个区间估计；在模型中加入非等比例变量与时间的交互项；改用时依协变量 Cox 模型。

五、小结

多变量回归模型分析在应用时需注意以下几点：在分析时注意离群点、高杠杆点和强影响点对模型的影响，要进行回归诊断，将其对结果的影响降到最小；在用模型进行预测时注意变量的取值范围，取值范围一定要在拟合模型的自变量取值范围之内；有些回归要求连续自变量与因变量（转换的因变量）呈线性关系，如果不满足线性关系，可改用哑变量拟合模型或采用数据转换满足线性条件；有些多变量分析单位不是个体而是由个体组成的二水平单位，如班级、医院或村庄，在二水平单位上用平均值进行回归分析，这种分析可产生生态学谬误和降低检验效能，应谨慎对待。

利益冲突　所有作者均声明无利益冲突

【选择题】（单选）

1. 常用的多变量回归分析方法不包括：
   A. logistic 回归
   B. F 检验
   C. Poisson 回归
   D. Cox 比例风险模型
2. logistic 回归的因变量不包括：
   A. 二分类变量
   B. 多分类变量
   C. 等级变量
   D. 连续变量
3. 多变量回归分析的自变量可为：
   A. 连续变量
   B. 等级变量
   C. 分类变量
   D. 以上都可
4. 在使用 Cox 比例风险模型时，下列哪项不是处理非等比例的方法：
   A. 将非等比例变量分层
   B. 使用多水平模型
   C. 在模型中加入非等比例变量与时间的交互项
   D. 改用时变协变量 Cox 模型
5. Poisson 回归模型过度离散的校正可用：
   A. 负二项分布回归
   B. 高斯回归
   C. 二项分布
   D. 正态分布

（收稿日期：2019-01-17）
（本文编辑：张振伟）