医学论文中常用回归分析方法的审核要点及对策

王 曼
郑州大学学报编辑部,450001,郑州

摘 要 回归分析是医学研究中大量使用的统计学方法。本文介绍了医学论文中常用的回归分析方法,包括多重线性回归、logistic 回归及 Cox 风险比例回归模型,并对其统计学审核要点进行了总结,对稿件处理措施提出了建议。

关键词 医学论文;回归分析;统计学审核

Review and strategy of regression analysis methods commonly used in medical articles//WANG Man

Abstract Regression analysis is extensively applied in medical research. This paper introduces regression analysis methods commonly used in medical articles, including multiple linear regression, logistic regression and Cox risk ratio regression model, summarizes the review points of these methods, and proposes strategies for improvement.

Keywords medical article; regression analysis; statistical review

Author's address Editorial Department of Journal of Zhengzhou University, 450001, Zhengzhou, China

DOI:10.16811/j.cnki.1001-4314.2018.05.009

在医学临床研究中,分析某些因素对疾病发生发展、治疗效果、生存状况等的影响是经常遇到的问题。解决这类问题就需要采用多因素分析的方法,去伪存真,从众多影响因素中筛选出独立影响因素,这对指导临床意义重大。多因素分析常常应用各种回归分析,比如多重线性回归、非条件或条件 logistic 回归、广义线性模型、Cox 风险比例回归模型等等,其中多重线性回归、logistic 回归、Cox 回归是最常用的 3 种方法。笔者在日常稿件编辑过程和浏览相关文献中发现,医学论文回归分析处理中存在大量错误,一些不当的数据处理甚至可以得到与事实相反的结论[1-2],严重影响论文的科学性。减少学术论文中的统计学错误,对提升学术期刊质量至关重要。作为期刊编辑,不但要从思想上加以重视,还要掌握一些应对策略,才能把好发表的最后一关[3]。笔者结合编辑工作的特点,对医学论文回归分析中存在的问题进行总结,提出审核要点,以期为编辑同人提供可操作性的应对策略。

1 常用的回归分析方法

1)多重线性回归是研究一个因变量与多个解释变量之间线性的数量依存关系,要求解释变量是定性或定量资料,因变量服从正态分布。

2)logistic 回归是一种广义线性回归,与多重线性回归有很多相同之处,其中因变量要求是二分类或多分类资料,解释变量与因变量的 logistic 概率呈线性关系。在用 logistic 回归分析生存资料时,因变量只能是生存结局,无法对生存时间的影响进行分析。

3)Cox 回归是一种半参数生存分析方法,目前在医学研究尤其是临床研究中大量应用。该模型以生存结局和生存时间为因变量,可同时分析众多解释变量对生存期的影响,能分析带有截尾生存时间的资料,且不要求估计资料的生存分布类型,在随访研究资料分析中的应用比 logistic 回归模型具有明显的优势[4]。

2 审读要点

2.1 模型的选用 在审读论文中的统计学内容时,首要问题是判断数据类型,依据因变量和解释变量的数据类型和研究目的判断选用的分析方法是否合适。医学论文中回归分析常见问题就是模型选用不当。logistic回归常与多重线性回归误用,多表现为用多重线性回归处理因变量是分类变量的资料。在进行生存分析时,有些研究用 logistic 回归代替 Cox 回归,这样处理只能分析生存结局影响因素,但未考虑生存时间,数据信息利用不全,结论不合理。

2.2 解释变量的筛选和相互独立性 作者一般会根据自己的专业、经验或查阅文献列举出候选解释变量。在审读这部分内容时,首先要判断样本量和解释变量的数量能否满足分析要求。一般来说,阳性结局事件个数(Cox 回归)或样本量少的组别(logistic 回归)的例数是解释变量数量的 10 倍左右时,能够保证假设检验的精度。这是最简单的估算方法。比如疾病组有 60 人,对照组 40 人,那么回归分析时可以引入 4 个自变量(40/10=4)。其次,回归分析一般要求解释变量相互独立。但是,许多作者经常把候选解释变量不做处理或筛选,而是全部带入回归模型。如果解释变量间存在相关性,将会使得回归方程不稳定,筛选不出真实的影响因素,或者筛选出来的因素与实际意义不相符。因此,编辑应要求作者补充解释变量的共线性分析结果,判定其是否满足相互独立的条件;或者建议作者进行变量筛选,从有共线性的变量组中筛选出若干个变量来建立最优回归式。

2.3 变量赋值 变量赋值直接影响模型的解读。通

过变量赋值才可以判定模型选用是否正确,也才能对求解出的模型进行解读,从而得出结论。但是,目前医学论文中常见问题有无变量赋值和赋值错误。在进行生存分析时,终点事件的定义一定要明确。对于无赋值的情况,编辑应要求作者补充。有赋值时,要考虑赋值是否正确。对于无序分类变量应进行哑变量化。有时候,将定量资料定性化处理后再赋值,以此变量构建出的模型更具有实际应用价值。比如将一些实验室指标、血压、年龄等按界值进行分段,然后赋值。近期,笔者发现,对于有些多分类变量进行赋值和哑变量化处理对结果影响很大。尤其是对于年龄这个因素,比如,当研究年龄对血压或血脂的影响时,尽管年龄分段后可视为有序变量,但有时候哑变量化处理比赋值更能体现试验的本质。

2.4 回归结果的审读 在作者补充、解决上述问题后编辑需要对回归结果进行核查。首先,编辑应核查统计报告是否完整。作者一般都会报告偏回归系数(β)及其显著性检验结果,OR、RR 或 HR 及其 95% CI。但对于一些研究,决定系数 R^2 至关重要,必须报告 R^2 是否有统计学意义及其大小。一般情况下,不能通过直接比较 β 的大小来判断某几个解释变量对因变量的影响程度,此时应要求作者报告标准回归系数。其次,核对数据,包括核对数值和方向。OR、RR 或 HR 的值等于 e^{β},显著性检验 $P>0.05$ 等价于 95% CI 包含 1。可以利用这些数据关系对统计报告中的数据进行相互印证。同时还要注意 β 的方向(正负号)与变量赋值的方向是否相符,是否能正确反映因变量与解释变量之间的关系。如果绘制有回归曲线,还要对曲线和数值进行比对分析。

3 案例分析

3.1 案例 1 某研究对 289 例接受肝癌根治切除术的 HBV 相关肝细胞癌患者进行 5 a 随访,随访率 89.8%,因肝癌死亡 155 例。分析术前凝血功能指标国际标准化比值(INR)对患者术后生存状况的影响。考虑到肝癌术后影响因素较多,同时收集了性别、年龄、术前分级、肝功能指标(ALT、AST)、总胆红素等指标。其中性别、年龄、术前分级为分类资料,肝功能指标、总胆红素为计量资料。

作者首先比较了不同性别、年龄段,不同分级患者 INR 的差异;对 INR 与肝功能指标、总胆红素进行了相关分析,发现 INR 与术前分级、总胆红素水平正相关。然后,作者以术后 5 a 是否存活为因变量,以性别、年龄、术前分级、INR、肝功能指标、总胆红素等指标为解释变量,进行 logistic 回归,回归结果见表 1。

表 1 案例 1 回归分析结果

变量	β	SE	OR(95% CI)	P
ALT	−0.411	0.290	0.663(0.376 ~ 1.170)	0.156
AST	0.587	0.234	2.508(1.138 ~ 2.844)	0.012
总胆红素	0.186	0.120	1.204(0.953 ~ 1.522)	0.120
INR	−0.553	0.218	1.738(1.134 ~ 2.665)	0.011

解析 1 1)这组资料属于生存分析,有随访数据,既有生存结局,也有生存时间,应选用 Cox 回归。2)该组资料阳性事件数(因肝癌死亡)228 例,引入回归模型的变量个数为 7,粗略估计样本量能够满足分析需要。3)作者的主要目的是探讨 INR 对术后生存的影响,故对 INR 和其他可能影响生存的因素进行了相关分析,并且结果显示术前分级、总胆红素水平与 INR 正相关。那么,在进行回归分析时,只是简单的将术前分级、总胆红素水平与 INR 一起作为解释变量带入模型的做法明显不合适,明显存在解释变量不独立的问题。因此,建议作者先进行变量筛选,再拟合最优回归方程。4)INR 偏回归系数为负数,值得怀疑。但因为没有变量赋值,无法判断筛选出来的 4 个变量的方向是否与实际相符。

解析 2 论文退修后,作者进行了下列修改并采用 Cox 回归重新进行了分析。1)调整变量结构:以性别、年龄为调整因素。以术后存活状况为因变量,其中因肝癌死亡赋值为 1,否则为 0。2)补充变量赋值:ALT、AST、INR、总胆红素水平升高(大于正常参考值上限)赋值为 1,否则为 0;术前 B 级赋值为 1,否则为 0。3)用逐步回归方法从相关变量集中选出最优回归子集。4)最终结果显示 INR 是 HBV 相关肝细胞癌术后预后的独立危险因素($\beta=0.606$,RR 为 1.833,95% CI 为 1.182 ~ 2.842)。经数据验证,$e^{\beta}=e^{0.606}=1.833$,与 RR 值相等,方向与赋值相符,结果可信。

3.2 案例 2 某研究对河南省农村地区 1 万 3 851 对育龄夫妇进行问卷调查并对妊娠结局进行随访;以不良妊娠结局为因变量,以年龄(实测值)、文化程度(初中及以下、高中及以下、大专及以上分别赋值为 1、2、3)、职业(务农 =1,服务业 =2,工人 =3,经商 =4,教师公务员 =5,其他 =6)、既往不良妊娠史、被动吸烟、饮酒、孕前服药史、正确服用叶酸、有毒物质接触史(这几个指标“有”或“是”赋值为 1,“无”赋值为 0)等为解释变量,进行 logistic 回归分析,探索不良妊娠结局的风险因素,结果见表 2。

解析 1)该组资料因变量为二分类资料,可以用 logistic 回归分析。2)调查对象为 1 万 3 851 对夫妇,引入回归模型的解释变量有 9 个,样本量能够满足分析需要。3)虽有变量赋值,但问题较大。调查对象年

表 2　案例 2 回归分析结果

变量	β	SE	Wald χ^2	P	OR(95% CI)
年龄	0.012	0.002	4.946	0.026	1.012(1.008 ~ 1.016)
既往不良妊娠史	1.806	0.635	8.089	0.004	6.086(1.753 ~ 21.128)
叶酸服用	−0.524	0.239	4.807	0.028	0.592(0.371 ~ 0.946)
职业	0.799	0.321	3.991	0.044	2.223(1.184 ~ 4.179)

龄在 20 ~ 47 岁,其中 31 ~ 35 岁者占 4.1%(568/13 851),35 岁以上者占 1.0%(138/13 851)。考虑到年龄构成,该变量不宜用实测值,建议按是否高龄孕妇分组,是高龄孕妇则赋值为 1,否则为 0。职业是无序分类变量,应哑变量化,而不要赋值。表 2 结果也显示,职业和年龄为不良妊娠结局风险因素,但是年龄的 β 为 0.012,职业为 0.799,怀疑不符合实际情况,可能与解释变量赋值错误有关。

作者依建议重新进行了回归分析,结果见表 3。重新拟合的结果更符合实际情况。

表 3　案例 2 回归分析结果(修改后)

变量	β	SE	Wald χ^2	P	OR(95% CI)
年龄	1.281	0.576	4.946	0.026	3.600(1.164 ~ 11.134)
既往不良妊娠史	1.724	0.627	7.560	0.006	5.607(11.641 ~ 19.162)
叶酸服用	−0.514	0.228	5.082	0.024	0.598(0.383 ~ 0.935)

3.3　案例 3　某项研究采用整群随机抽样的方法选择河南农村人群9 980人进行调查并测定血压、身体脂肪率(BFP),采用 logistic 回归分析 BFP 与高血压的关系。结果显示,在调整了年龄、性别、文化程度、婚姻状况、人均月收入、吸烟、饮酒、高脂饮食、较多蔬菜水果摄入、较多咸菜摄入、体力活动以及高血压家族史等因素后,BFP 与高血压患病存在关联,OR(95% CI)为 1.17(1.02 ~ 3.54)。

解析　1)该组资料因变量为二分类资料,可以选用 logistic 回归。2)调查对象有 9 980 人,样本量能够满足分析需要。3)将年龄、性别、文化程度、婚姻状况、人均月收入、吸烟、饮酒、高脂饮食、较多蔬菜水果摄入、较多咸菜摄入、体力活动以及高血压家族史等设定为调整因素,考虑较全面。4)因为 BFP 是个定量指标,所以作者直接将其作为连续型变量带入了模型,但这种考虑欠妥。根据经验,BFP 的单位变化对高血压患病风险的影响有限。因此,建议作者将 BFP 按四分位数分段,转换成有序变量。但是,考虑到 BFP 增加时高血压患病风险是增加还是减少并不确定,故并未对 BFP 按有序变量进行赋值,而是按无序变量进行了哑变量化处理,同时做了 $\chi^2_{趋势}$ 检验,结果见表 4。其中,Q2 的粗 OR(95% CI)为 1.16(1.02 ~ 1.04),95% CI 未能包括 1.16,怀疑作者笔误,经作者再次核对数据,Q2 的粗 OR(95% CI)为 1.16(1.02 ~ 1.24)。

表 4　案例 3 回归分析结果

BFP	n	高血压/例(占比/%)	粗 OR(95% CI)	调整后 OR(95% CI)
Q1	2 471	673(27.24)	1.00	1.00
Q2	2 501	755(30.19)	1.16(1.02 ~ 1.04/1.24)	1.71(1.48 ~ 1.97)
Q3	2 483	809(32.58)	1.29(1.14 ~ 1.46)	2.69(2.25 ~ 3.22)
Q4	2 525	1155(45.74)	2.25(2.00 ~ 2.54)	4.34(3.60 ~ 5.23)
$\chi^2_{趋势}$		187.88	185.57	140.76
P		<0.001	<0.001	<0.001

4　结束语

目前,各个医学期刊的外审主要负责学术上的把关,即使有统计学专家审稿,审稿人员也不可能像编辑一样在细节上、数据上一一核对,因此医学论文中的统计学问题仍较严重[5-5]。而编辑人员对论文的统计学审核往往由于专业的局限而浮于表面,严重影响论文的科学性[7]。回归分析是医学研究,尤其是临床研究中常用的统计方法,医学稿件中存在大量的回归分析不当甚至错误的问题,由错误的分析结果得出的结论也必然不可靠。因此,笔者对医学论文中回归分析内容的审核要点以及处理对策进行了部分总结。这些问题具有普遍性,提出的审核方法和处理对策也具有实际操作性,并不需要编辑人员学习具体的统计学软件。希望本文能对编辑同人在医学论文的统计学把关方面有所助益。

5　参考文献

[1] 周英智. 医学论文“统计学处理”常见问题分析及建议[J]. 中国科技期刊研究, 2016, 5(27): 480

[2] 王曼. 医学论文统计描述性数据审核的问题与方法[J]. 中国科技期刊研究, 2015, 26(4): 359

[3] 武建虎, 尤伟杰, 张楠, 等. 医学论文中关于“随机”的常见错误及编辑应对策略[J]. 编辑学报, 2016, 28(6): 558

[4] 孙振球. 医学统计学[M]. 4 版. 北京: 人民卫生出版社, 2014: 312

[5] 田云鹏, 陈丽. 医学论文中常见统计学错误例析[J]. 中华全科医学, 2017, 15(10): 1791

[6] 闵莹. 医学期刊文献中常见的统计学问题分析[J]. 天津科技, 2016, 43(1): 68

[7] 吴学军, 廖粤新. 科技期刊编辑应注重统计数据的审核[J]. 编辑学报, 2010, 22(5): 416

(2018-01-17 收稿;2018-07-15 修回)