体育社会学研究中定类、定序变量的回归分析

杨 威[1], 杨 霆[2]
(1. 东北师范大学 体育学院,吉林 长春 130024; 2. 吉林大学 体育学院,吉林 长春 130012)

摘 要:回归分析是社会学研究领域中进行多因素分析最重要的研究方法之一,而将回归分析的方法应用到体育社会学实证研究中的较少。学习和掌握西方社会学这种定量分析的研究方法,对提高我国体育社会学研究方法的应用水平和研究水平具有重要意义。由于体育社会学研究中的变量主要为定类与定序变量,而体育统计教材对定类与定序变量的回归分析方法鲜有论及。从介绍定类与定序变量的回归分析方法入手,旨在为其在体育社会学研究中的应用提供方法指导。

关 键 词:回归分析; 定类变量; 定序变量; 体育社会学
中图分类号:G80-05 **文献标识码**:A **文章编号**:1006-7116(2004)05-0022-03

Nominal level variable and ordinal level variable of regression analysis in the research of physical sociology

YANG Wei[1], YANG Ting[2]
(1. School of Physical Education, Northeast Normal University, Changchun 130024, China;
2. School of Physical Education, Jilin University, Changchun 130012, China)

Abstract: regression analysis is the one of the most important methods which analyze many factors possessed causation in the sociology field. Few people apply the method to the practice of physical sociology. It is meaningful to study and master the method which can improve the level of application and research. This article will introduce the measurement of the social variables, the transform of nominal level variable and ordinal level variable and the regression analysis of nominal level variable and ordinal level variable. The purpose of the article is to guide the application of the regression analysis.

Key words: regression analysis; nominal level variable; ordinal level variable; physical sociology

近些年来,西方社会研究方法中有关实证研究的技术,特别是建立在概率论基础上的统计学方法受到了我国社会学界的普遍重视。学者们学习和借鉴西方社会学定量分析的研究方法,对我国的社会现象与社会问题进行了大量的实证研究。其中,许多学者应用回归分析的方法,对通过问卷或量表收集的经验事实进行多因素量化分析,进而对社会现象进行深入的、多层次的理论解释,这在当前的社会学界甚为流行。如《社会学研究》中发表的徐安琪的论文"择偶标准:五十年变迁及其原因分析",作者将择偶取向分为政治、经济和个性气质3方面建立多元回归模型,利用回归模型分析人口特征、个人资源、择偶方式的诸多指标对择偶标准的影响性质及影响力的大小;又如李树茁等的"中国农村子女的婚姻形式和个人因素对分家的影响研究",利用 Logistic 回归模型分析农村子女的家庭类型、婚姻形式和个人背景因素对与父母分家可能性的影响,这类论文还有李春玲的"文化水平如何影响人们的经济收入"、胡欣的"相对剥夺地位与阶层认知"等。

然而在我国的体育社会学界,还较少有人将回归分析的方法应用到体育社会学的实证研究中,其研究水平及定量研究方法的应用与母学科前沿差距较大。

体育领域中社会现象的发生与变迁,都可能是多种因素的影响所致,很少可以用一种现象去解释另一种现象的产生,往往需要用多种原因或多种因素进行解释。在这种多因素的分析与解释过程中,较为常用的是统计学中的回归分析方法。通过回归分析,可以分析多个自变量对某个因变量的共同影响,即纳入研究的多种因素对社会现象解释力的大小;还可以将因素的作用分解,以确定具有因果联系的自变量中哪些是稳定的、相关程度较高的因素,达到比较自变量对因变量主次作用的目的。如对城乡居民参加体育锻炼频度影响因素分析、对是否是体育人口影响因素的探讨和对学生体育态度影响因素的研究等都可以采用回归分析的方法。因此回归分析已成为深入研究体育社会现象,进行多因素定

收稿日期:2004-01-20
作者简介:杨 威(1950-),男,教授,研究方向:体育统计。

量分析,进而达到理论解释十分有利的方法。

回归分析中最基本的方法是多元线性回归,它除要求自变量间的关系是相互独立、线性可加等条件之外,还要求所涉及的变量均应达到较高的定距测量尺度。然而,在体育社会学研究中,经操作化后的社会变量大多数是低层次的定类与定序变量,如职业、家庭居住地、文化程度、生活满意度、体育活动项目、体育消费类型等,这无疑使多元线性回归分析方法的应用受到限制。因此学习和掌握定类、定序变量回归分析的方法,对提高我国体育社会学研究方法的应用水平和研究水平将具有极大的推动作用。

1　定类变量与定序变量的量化处理

为了适应回归分析的需要,首先应对定类或定序变量进行量化处理。

1.1　定类变量的量化处理

一个变量被赋予 0 与 1 两个值变为定距变量,那么统计学上将这个变量称之为虚拟变量。用 k 个取值为 0 和 1 的虚拟变量分别代表各类的属性,当案例属于一个虚拟变量所代表的类时,这个虚拟变量就赋值为 1,否则为 0。对二分定类变量,可以直接以 0、1 表示两个类别即可。如性别可将男性赋予 0,女性赋予 1,那么就将性别转换为取值为 0 和 1 的虚拟变量。如果定类变量的类别或属性在 2 个以上时,可以用一组虚拟变量来表示。例如,在对单位职工参加体育活动状况影响因素的研究中,若单位所有制类型为一自变量,这一定类自变量的类别分别为:国家、集体、外资和民营,那么可将其转换为 4 个虚拟变量(见表 1)。定类变量转换为虚拟变量后可应用虚拟变量回归分析的方法。

表 1　单位所有制类型及其虚拟变量

原变量类别	虚拟变量			
	D_1(国)	D_3(集)	D_4(外)	D_{4+}(民)
集体	0	1	0	0
外资	0	0	1	0
国家	1	0	0	0
民营	0	0	0	1

1.2　定序变量的量化处理

对于抽象程度较低的定序变量,可以按照被测对象具体特征的高低或大小由低到高或由小到大依次赋予 1、2、3……等数值,将定序变量转化为定距变量。而对于抽象程度较高或综合性较强的概念,可以利用利克特量表进行测量。利克特量表的测量尺度是一种定序尺度,其测量结果属定序变量。此时可以依量表中陈述语句(变量)的正反方向将 5 级填答结果转换为 5、4、3、2、1 或 1、2、3、4、5,这样每一陈述句的得分或整个量表的总分都可视为定距变量。上述转换得到的变量虽然不是"真正"的定距变量,但都假定其具有定距变量的数学性质,因此可以进行有意义的多元线性回归分析。更为科学的量化处理是采用因子分析的方法,将量表中的多个观测变量简化为少数几个因子变量,然后计算各因子得分或加权综合得分,可用这些因子得分或综合分数代替原来的观测变量进行多元线性回归分析。

当然,也可将定序变量视为定类变量,再将其转换为虚拟变量,然后进行虚拟变量回归分析。

2　定类或定序自变量的回归分析

在体育社会学研究中,如果因变量达到定距测度,那么对于定类或定序自变量的回归分析,可先将其转换为虚拟变量,进行虚拟变量的回归分析。虚拟变量回归分析中需要特别注意的是,不能将由每个定类或定序变量转换的一组虚拟变量均引入回归方程,每组必须放弃其中的一个虚拟变量,否则将会使回归方程中的回归系数无解。被放弃的虚拟变量称为参照项,放弃哪一虚拟变量应根据研究需要而定。例如,欲研究单位所有制类型(分为国家、集体、外资和民营 4 类)、健康状况(分为良好、一般和较差 3 类)对单位职工参加体育活动情况的影响,拟以周参加体育活动时间作为因变量 y,单位所有制类型 x_1、健康状况 x_2 作为自变量进行回归分析。先将 2 个自变量转换为虚拟变量,然后做虚拟变量回归分析,得到如下虚拟变量回归方程:

$$\hat{y} = b_0 + b_1D_1 + b_2D_2 + b_3D_3 + b_4D_4 + b_5D_5$$

在上述虚拟变量回归方程中,虚拟变量 D_1 表示单位所有制类型中的国家,D_2 表示单位所有制类型中的集体,D_3 表示单位所有制类型中的合资;D_4 表示健康状况中的一般,D_5 表示健康状况中的较差。各类别虚拟变量的偏回归系数 b_j 则表示了该类别与其参照项均值之差,任意不同两类虚拟变量的偏回归系数也可以直接比较两类之间的差异。由于回归方程中没有引进单位所有制类型中的民营和健康状况中的良好 2 个虚拟变量(此两个虚拟变量均作为参照项),而 b_0 表示的则是其效果。

应用回归模型进行多因素分析时,在大多数情况下引入回归方程中的多个自变量可能既有定类变量、定序变量,还有不同计量单位的定距变量(如在上述自变量中增加经济收入 x_3),此时可先将定性变量量化后与其它定量变量一起建立含虚拟变量的回归方程:

$$\hat{y} = b_0 + b_1D_1 + b_2D_2 + b_3D_3 + b_4D_4 + b_5D_5 + b_6x_3$$

回归方程中偏回归系数的符号反映了自变量对因变量影响的方向,通过对偏回归系数 b_j 的检验,可以判定单位所有制各种类型、各种健康状况和职工经济收入对周参加体育活动时间影响显著性的大小;也可以计算标准回归系数,通过对标准回归系数的比较,反映各自变量对因变量影响力的大小;同时还可以计算回归方程的确定系数 R^2,以此表示引入回归方程的全部自变量共同解释因变量解释力的大小,R^2 数值越大,建立的回归模型的解释力越强。

虚拟变量回归分析的特点是不要求自变量具有较高的测量层次,不必假设自变量之间是线性关系,因此,对体育领域中的社会现象、社会问题进行多因素分析时都可以采用含

虚拟变量回归分析的方法，故其在体育社会学研究中具有较为广泛的适用性。但在进行分析时，定类或定序变量的类别越多，分析就越为麻烦。因此，做虚拟变量回归，应尽量减少分类的数目。

3 定类或定序因变量的回归分析

在体育社会学研究中也会遇到定类因变量的回归分析问题。如居民是否是体育人口，往往受其年龄、健康状况、受教育程度、余暇时间和体育认知及情感等因素的影响；又如居民健身时选择的体育活动项目常常与其性别、年龄、身体状况和群体的影响等因素有关。此类研究的因变量“是否是体育人口”和“体育活动项目”均为定类变量，多元回归分析在这种条件下不宜再使用。随着社会统计方法的发展，一些适用于定类因变量的回归分析方法相继应运而生。

在 20 世纪 80 年代以前，对于二分的定类因变量是将其视为虚拟变量，进行线性概率回归分析。但这种回归分析方法具有许多局限性。

目前多采用 Logistic 回归对定类因变量进行回归分析。Logistic 回归分析根据因变量取值类别的不同又可分为 Binary Logistic 回归和 Multinomial Logistic 回归。

Binary Logistic 回归适用于二分的定类因变量。若某类事件发生则令 $y=1$，不发生令 $y=0$。事件发生的概率 $p=p(y=1)$，不发生的概率 $1-p=p(y=0)$，其发生比为 $p/(1-p)$。发生比是不对称的，需取对数。以对数发生比作为因变量，计算其与自变量 $x_1, x_2, \cdots\cdots x_k$（如自变量中有定类或定序变量须经量化处理）的 Logistic 回归方程，方程式如下：

$$\mathrm{Ln}\left(\frac{p}{1-p}\right) = b_0 + b_1x_1 + b_2x_2 + \cdots\cdots + b_kx_k$$

Binary Logistic 回归方程的右半部形式与一般多元线性回归方程在形式上相同，原则上对回归系数的检验与解释也相同。但是回归系数 b_i 反映的是自变量 x_i 的变化对对数发生比的作用，由于对数发生比没有直观意义，需将相应的 b 转换为 e^b，这样 b_i 就有了直观的解释。

因变量也可由多种分类所构成，在这种情况下需要做较为复杂的 Multinomial Logistic 回归分析。Multinomial Logistic 回归是对 Binary Logistic 回归分析的扩展，是由一组 Logistic 回归方程所构成。其对回归系数的解释和检验与 Binary Logistic 回归相同。

当因变量为定序变量时，可将定序变量视为定类变量，采用上述 Logistic 回归分析的方法。

回归分析是社会学研究领域中进行多因素分析最重要的研究方法之一，且运用 SPSS 或 SAS 统计软件均可进行虚拟变量回归分析与 Logistic 回归分析，非常迅速、便捷。学习和掌握这些方法，并将其运用到体育社会学研究实践中去，必将使我国体育社会学研究从定性走向定量，由思辨走向实证，从而结出丰硕的研究成果。

参考文献：

[1] 于秀林．多元统计分析[M]．北京：中国统计出版社，2003.

[2] 余建英．数据统计分析与 SPSS 应用[M]．北京：人民邮电出版社，2003.

[3] 斯迪虎，王胜利．体育社会学研究中的因果关系本质——数理统计与逻辑实证关系的辩证思考[J]．体育学刊，2002，9(4)：127－129.

[4] 风笑天．社会学方法二十年：应用与研究[J]．社会学研究，2000(1)：31－33.

[5] 徐安琪．择偶标准：五十年变迁及其原因分析[J]．社会学研究，2000(6)：23－25.

[编辑：周成]