文章编号:1000－2367(2015)02－0179－06 DOI:10.16366/j.cnki.1000－2367.2015.02.034

基于多元线性回归模型的我国世界冠军人数影响因素分析

冉艳恩

(郑州大学 体育学院,郑州 450044)

摘 要:运用多元线性回归模型分析了我国当年获得世界冠军人数与我国当年总人口数和当年国内生产总值两个外部环境,以及二级以上运动员人数和各级体育系统职工人数两个内部环境,这 4 个影响因素之间的关系.首先用灰色关联分析,得到作为子因素数列的 4 个影响因素和作为母因素数列的当年获得世界冠军人数之间的关联度.为了进一步知道每个因素和当年获得世界冠军人数的具体函数关系,利用多元线性回归分析,构建每年获得世界冠军的人数与 4 个影响因素之间的多元线性回归模型,且通过 F 检验证明模型与实际相符合,为我国运动员的训练和国家政策的制定提供理论依据.

关键词:世界冠军人数;总人口数;国内生产总值;回归模型;灰色关联分析

中图分类号:O212.4;G80－32 **文献标志码**:A

在经济和科技快速发展的今天,一个国家竞技体育水平的高低从侧面反映了综合国力的强弱[1],而国家每年获得世界冠军人数则反映了国家竞技体育水平的高低,因此研究对每年获得世界冠军人数的因素以及他们之间的关系就很必要了[2].每年获得世界冠军的人数不是一个孤立的单独的数据,是一个受很多方面影响的一个复杂结果.这些影响因素包括当年世界性体育赛事举办的多少,还有类似奥运周期年对运动员培养所产生的特定效果等[3].因此,我们在运用线性回归模型、灰色 GM(1, 1)模型或灰色线性回归模型等预测理论对每年获得世界冠军的人数这样较为复杂的观测量进行预测时,就必须要考虑一些限制性因素:优秀运动员的建设是一个不间断的整体训练过程,同时运动员的梯队建设也是影响这类观察量的主要因素;对运动员未来的发展要以长远的眼光来看待,不能仅仅把这作为我国体育事业健康发展的指向标[4－6].

当然,我国每年获得世界冠军的人数,在大环境上和国家当年的总人口数、国内的生产总值、国际上举行的赛事等有关,这里我们筛选出国家当年的总人口数和国内的生产总值作为外部驱动因素研究指标.从小环境来说也和国家对体育事业政策上的支持以及倾斜度有关[7].在上面所说的大环境总体上较固定、不易变动的影响因素下,本文认为我们自身的因素更重要,自身的因素体现在两点:运动员的优秀程度和优秀教练员的水平,这两个因素作为内部驱动因素可以用下面这两个数据来量化:二级以上运动员人数和各级体育系统职工人数[8].因此,我们分析每年获得世界冠军人数和我国当年总人口数、当年国内生产总值、二级以上运动员人数和各级体育系统职工人数之间的关系和内部联系,为我国运动员的训练和国家政策的制定提供理论基础.

1 研究方法

1.1 文献资料法

查询中国期刊网、国家体育总局官网等国内大型检索机构和官方网站,收集有关我国获世界冠军人数的研究资料和成果,为本次研究提供翔实的研究基础和理论支撑.

收稿日期:2014-08-27;**修回日期**:2014-11-25.

基金项目:河南省科技厅软科学项目(142400410173)

作者简介:冉艳恩(1979－),男,河南郑州人,郑州大学体育学院讲师,主要从事体育教育训练学、体育统计学等研究,E-mail:ranyanen@163.com.

1.2 灰色关联分析法

将世界冠军人数、国内生产总值、全国总人数、二级以上运动员人数、各级体育系统职工人数的数据进行无量纲化处理,计算他们的关联系数,在此基础之上得到他们的关联系数.

1.3 多元线性回归模型分析法

用多元线性回归模型来构建每年获得世界冠军的人数与每年国内生产总值、当年全国总人口数、当年二级以上运动员人数和当年各级体育系统职工人数之间具体的函数关系,用 Matlab 软件对回归结果进行 F 检验和残差分析,去除异常点后得到满意的回归模型[9].

2 研究结果与分析

2.1 灰色关联分析模型

灰色关联分析方法是一种系统分析方法,克服了数理统计中传统的回归相关分析、方差分析、主成分分析等因素分析方法的不足.我们假设"获得世界冠军人数"为主行为因子的时间序列为 $X_0(k)$,"国内生产总值"、"全国总人口数"、" 二级以上运动员人数"、"各级体育系统职工人数"为比较时间序列,分别为 $X_1(k)$, $X_2(k)$, $X_3(k)$, $X_4(k)$,具体数据见表 1.

表 1 2000—2012 年获世界冠军人数、国内生产总值、全国总人数、二级以上运动员人数、各级体育系统职工人数的原始数值

k(年份)	获世界冠军人数 $X_0(k)$	国内生产总值 $X_1(k)$	全国总人数 $X_2(k)$	二级以上运动员人数 $X_3(k)$	各级体育系统职工人数 $X_4(k)$
k=1(2000)	109	99.216 4	126.743	21.993	153.599
k=2(2001)	138	109.655 2	127.627	24.752	153.091
k=3(2002)	123	120.332 7	128.453	31.469	147.778
k=4(2003)	94	135.822 8	129.227	22.309	97.062
k=5(2004)	175	159.878 3	129.988	28.836	143.665
k=6(2005)	159	184.937 4	130.756	21.521	141.849
k=7(2006)	169	216.314 4	131.448	23.148	126.910
k=8(2007)	217	265.810 3	132.129	24.422	147.929
k=9(2008)	151	314.045 4	132.802	22.798	150.575
k=10(2009)	223	340.902 8	133.450	22.753	153.398
k=11(2010)	180	401.512 8	134.091	46.341	155.527
k=12(2011)	198	473.104 0	134.735	38.380	157.333
k=13(2012)	140	518.942 1	135.404	46.412	159.762

注:数据来自于 2000—2012 年的《统计年鉴》.

其中,国内生产总值单位为千亿元,全国总人口数为千万,二级以上运动员人数和各级体育系统职工人数为千.

2.1.1 数据的选取及无量纲化处理

在表 1 中,国内生产总值(千亿元)、全国总人数(千万)、二级以上运动员人数(千)和各级体育系统职工人数(千),且数量级也不尽相同,有的是几百,有的是几十.这些不同的计算单位和不同的数量级导致这些数据不能直接进行综合分析,需要消除量纲的影响,对各指标的原值数据做无量纲化值的变换.均值化无量纲化方法在保留原始变量变异程度信息时,并不是仅取决于原始变量标准差,而是原始变量的变异系数,这就保证了变量变异程度信息的同时数据的可比性问题[10].本文采用均值法进行初始值转化,变换公式为:

$$x_i(k)=\frac{X_i^{(0)}(k)}{X_i}, \tag{1}$$

其中 $X_i^{(0)}(k)$, $i=0, 1, 2, \cdots, n$ 为最初原始数据,X_i 表示第 i 数据列的平均值,$x_i(k)$表示第 i 数据列的均值化数列,由公式(1)可得表 2.

表 2　2000—2012 年被无量纲化后的获世界冠军人数、国内生产总值、全国总人数、二级以上运动员人数、各级体育系统职工人数

k(年份)	无量纲化后的获世界冠军人数 $x_0(k)$	无量纲化后的国内生产总值 $x_1(k)$	无量纲化后的全国总人数 $x_2(k)$	无量纲化后的二级以上运动员人数 $x_3(k)$	无量纲化后的各级体育系统职工人数 $x_4(k)$
k=1(2000)	0.682 56	0.386 11	0.965 32	0.762 21	1.057 35
k=2(2001)	0.864 16	0.426 74	0.972 05	0.857 82	1.053 86
k=3(2002)	0.770 23	0.468 29	0.978 34	1.090 61	1.017 28
k=4(2003)	0.588 63	0.528 58	0.984 24	0.773 16	0.668 16
k=5(2004)	1.095 86	0.622 19	0.990 03	0.999 36	0.988 97
k=6(2005)	0.995 66	0.719 71	0.995 88	0.745 85	0.976 47
k=7(2006)	1.058 29	0.841 82	1.001 15	0.802 24	0.873 63
k=8(2007)	1.358 86	1.034 44	1.006 34	0.846 39	1.018 32
k=9(2008)	0.945 57	1.222 16	1.011 47	0.790 11	1.036 54
k=10(2009)	1.396 44	1.326 68	1.016 42	0.788 55	1.055 97
k=11(2010)	1.127 17	1.562 55	1.021 28	1.605 09	1.070 62
k=12(2011)	1.239 98	1.841 16	1.026 19	1.330 13	1.083 06
k=13(2012)	0.876 69	2.019 55	1.031 29	1.608 49	1.099 78

经过无量纲化后的数据，消除了获世界冠军人数、国内生产总值、全国总人数、二级以上运动员人数和各级体育系统职工人数这 5 个指标之间不同的计算单位和数量级的影响，解决了各指标数值不可综合性问题，就可以直接对这 4 个比较序列对主行为因子的影响进行评价和比较了[11].

2.1.2　计算关联系数

无量纲化后的比较时间序列“国内生产总值”$x_1(k)$、“全国总人口数”$x_2(k)$、“二级以上运动员人数”$x_3(k)$、“各级体育系统职工人数”$x_4(k)$ 分别对无量纲化后的主行为因子时间序列“获得世界冠军人数”$x_0(k)$ 的影响大小可用关联系数 $\xi_i(k)(i=1,2,3,4)$ 来表示，$\xi_i(k)$ 的表达式为：

$$\xi_i(k)=\frac{\min\limits_i\min\limits_k|x_0(k)-x_i(k)|+\rho\max\limits_i\max\limits_k|x_0(k)-x_i(k)|}{|x_0(k)-x_i(k)|+\rho\max\limits_i\max\limits_k|x_0(k)-x_i(k)|},\tag{2}$$

ρ 为分辨系数，取值在(0,1) 之间，在这里取 $\rho=0.5$. 由公式(2)，可得关联系数 $\xi_i(k)(i=1,\cdots,4,k=1,2,\cdots,13)$，见表 3.

表 3　无量纲化后的国内生产总值、全国总人数、二级以上运动员人数、各级体育系统职工人数与无量纲化后的获世界冠军人数之间的关联系数

k(年份)	国内生产总值 $\xi_1(k)$	全国总人数 $\xi_2(k)$	二级以上运动员人数 $\xi_3(k)$	各级体育系统职工人数 $\xi_4(k)$
k = 1(2000)	0.658 67	0.669 23	0.878 00	0.604 14
k = 2(2001)	0.566 64	0.841 50	0.989 41	0.751 05
k = 3(2002)	0.654 53	0.733 32	0.641 00	0.698 43
k = 4(2003)	0.905 25	0.591 13	0.756 19	0.878 16
k = 5(2004)	0.546 98	0.844 06	0.855 85	0.842 74
k = 6(2005)	0.674 61	1.000 00	0.696 08	0.967 88
k = 7(2006)	0.725 54	0.909 45	0.690 83	0.756 06
k = 8(2007)	0.638 11	0.618 70	0.527 40	0.626 83
k = 9(2008)	0.674 10	0.896 95	0.786 43	0.863 00
k = 10(2009)	0.891 55	0.600 82	0.484 73	0.626 88
k = 11(2010)	0.567 78	0.843 99	0.544 77	0.910 30
k = 12(2011)	0.487 50	0.728 01	0.864 07	0.784 86
k = 13(2012)	0.333 46	0.787 36	0.438 64	0.719 49

关联系数 $\xi_i(k)$ 表示 1999＋k 年的第 i 个比较时间序列对应得数据对 1999＋k 年的主行为因子时间序列“获得世界冠军人数”$X_0(k)$ 的影响力大小[12]. 比如:$\xi_1(1)=0.658\ 67$ 表明2000年的国内生产总值对2000年的获世界冠军人数的影响力大小,并且从 $\xi_3(1)=0.878\ 00$ 可以看出 2000 年时 4 个比较时间序列中二级以上运动员人数对获世界冠军人数的影响力最大[13]. 但关联系数中的每一项 $\xi_i(k)(i=1,2,3,4,k=1,2,\cdots,13)$ 的数据都有 13 个,所以计算出来的关联系数也有 13 个,这样信息就过于分散. 如果把各点的关联系数取平均值,就可以把各点关联系数集中在一个值,称为关联度.

2.1.3 计算关联度

子因素数列 $\xi_i(k)(i=1,2,3,4,k=1,2,\cdots,13)$ 的关联度大小代表了与母因素数列 $\xi_0(k)(k=1,2,\cdots,13)$ 的关系密切情况,若大,则代表关系密切,从而该因素的影响就大,反之亦然. 关联度 γ_i 加权公式为:

$$\gamma_i=\frac{1}{13}\sum_{k=1}^{13}\xi_i(k), \tag{3}$$

由公式(3),得 $\gamma_1=0.640\ 36$, $\gamma_2=0.774\ 19$, $\gamma_3=0.704\ 11$, $\gamma_4=0.771\ 52$.

从结果中可以看出,子因素数列国内生产总值、二级以上运动员人数、体育系统职工人数和全国总人口数这 4 个因素与母因素数列获世界冠军人数的密切度依次增加. 在外环境中,全国总人口数比国内生产总值对获世界冠军人数的影响更大;在内环境中,体育系统的科技人员和教练员比二级以上运动员对获世界冠军人数的影响更大.

关联度的计算只能得出这几个因素中说明哪个因素更重要,但每个因素和当年获得世界冠军人数的具体函数关系还不知道.

2.2 获得世界冠军的人数与 4 种影响因素的关系

我们就用多元线性回归模型来构建每年获得世界冠军的人数与每年国内生产总值、当年全国总人口数、当年二级以上运动员人数和当年各级体育系统职工人数之间的具体关系.

2.2.1 多元线性回归模型

设影响因变量 Y 的自变量为 x_1, x_2, x_3, x_4,如果满足下述关系

$$Y=\alpha_0+\alpha_1x_1+\alpha_2x_2+\alpha_3x_3+\alpha_4x_4+\varepsilon, \tag{4}$$

其中:$\varepsilon\sim N(0,\sigma^2)$ 是零均值的随机变量,Y, x_1, x_2, x_3, x_4 分别表示每年获得世界冠军的人数,每年国内生产总值,当年全国总人口数,当年二级以上运动员人数,当年各级体育系统职工人数[6]. α_0, α_1, α_2, α_3, α_4 为待定的回归系数.

2.2.2 回归系数计算

现有个独立观测数据 Y_i, x_{1i}, x_{2i}, x_{3i}, x_{4i}, $i=1,2,\cdots,n$,用最小二乘法求得未知参数的值,代入回归方程(4),得:

$$Y=\alpha X+\varepsilon, \tag{5}$$

其中 $Y=(Y_1,Y_2,\cdots,Y_n)=(109,138,\cdots,140)$, $\alpha=(\alpha_0,\alpha_1,\cdots,\alpha_4)$, $X=\begin{pmatrix}1&1&\cdots&1\\x_{11}&x_{12}&\cdots&x_{1n}\\x_{21}&x_{22}&\cdots&x_{2n}\\x_{31}&x_{32}&\cdots&x_{3n}\\x_{41}&x_{42}&\cdots&x_{4n}\end{pmatrix}=$

$$\begin{pmatrix}1&1&1&1&1&1&1&1&1&1&1&1&1\\99.2146&109.655\ 2&120.332\ 7&135.822\ 8&159.878\ 3&184.937\ 4&216.314\ 4&265.810\ 3&314.045\ 4&340.902\ 8&401.512\ 8&473.104\ 0&518.942\ 1\\126.74\ 3&127.627&128.453&129.227&129.988&130.756&131.448&132.129&132.802&133.450&134.091&134.735&135.404\\21.993&24.752&31.469&22.309&28.836&21.521&23.148&24.422&22.798&22.753&46.341&38.380&46.412\\153.599&153.091&147.778&97.062&143.665&141.849&126.910&147.929&150.575&153.398&155.527&157.333&159.762\end{pmatrix}.$$

由多元线性回归模型和 Matlab 软件,可得下面统计数据,见表 4. 其中可决系数 $R^2=0.725\ 6$,F 统计量为 5.3134,F 统计量对应的概率 $\rho<0.05$. 由表 4 的数据和模型(5),可得此时的模型(2) 为:

$$Y=-4147.5-0.5x_1+32.5x_2-x_3+1.4x_4, \tag{6}$$

相关系数的平方值为 $R^2=0.725\ 6$,说明模型拟合程度不高,且模型(6) 的残差杠杆图见图 1.

残差杠杆图 1 中的异常点是 2008 年的数据,去掉后重复上面的过程,再依次去掉异常点 2012 年和 2005 年的数据后,用 Matlab 再次作多元线性回归,得表 5.

表 4　第一次线性回归 F 检验结果

回归系数	回归系数估计值	回归系数置信区间
α_0	−4147.5	[−7 570.2, −7 249]
α_1	−0.5	[−1.1, 0.1]
α_2	32.5	[6.1, 58.8]
α_3	−1	[−4.0, 1.9]
α_4	1.4	[0.1, 2.6]
$R^2 = 0.725\ 6, F = 5.313\ 4, \rho < 0.021\ 8$		

表 5　第二次线性回归 F 检验结果

回归系数	回归系数估计值	回归系数置信区间
α_0	−4 216.7	[−5896.6, −2536.8]
α_1	−0.4	[−0.7, −0.1]
α_2	33	[20, 46]
α_3	−2	[−3.4, −0.6]
α_4	1.5	[1, 2.1]
$R^2 = 0.973\ 3, F = 45.552\ 8, \rho = 0.000\ 4$		

其中可决系数 $R^2 = 0.973\ 3$,F 统计量为 45.552 8,F 统计量对应的概率 $\rho < 0.000\ 5$. 由表 5 的数据和模型(5)得:

$$Y = -4\ 216.7 - 0.4x_1 + 33x_2 - 2x_3 + 1.5x_4, \tag{7}$$

对应的残差杠杆图为图 2,且图 2 中已经无异常点.

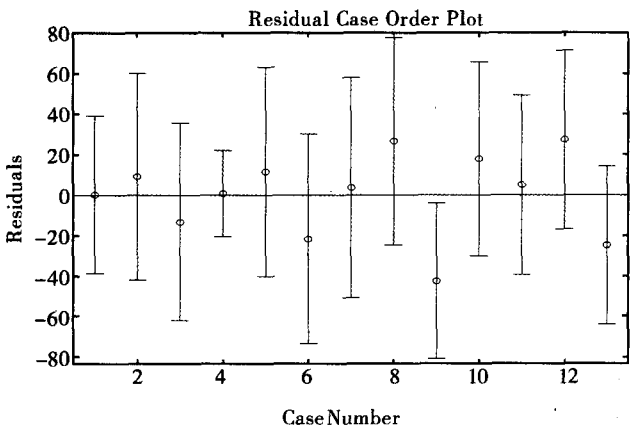

图1 模型(6)的残差杠杆图

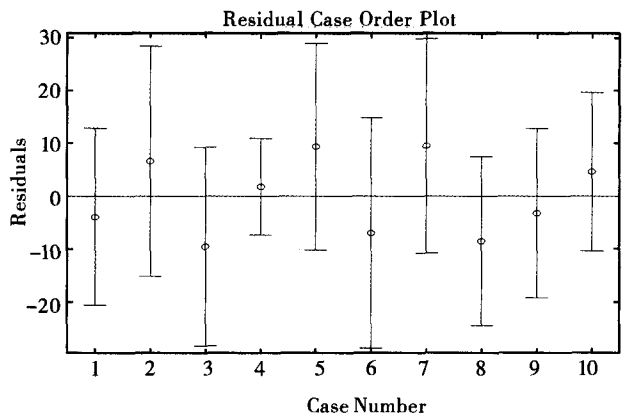

图2 模型(7)的残差杠杆图

在对多元线性回归模型进行 F 检验时,测定回归方程对各个观测点的拟合程度的可决系数 $R^2 \in [0,1]$ 的值越大(小)表明回归直线对各个观测点的拟合程度越高(低). 且由表 4 和表 5 的数据可得,去掉异常点后置信区间不包含零点,可决系数 R^2 从 0.726 5 增大到 0.973 3,F 值由 5.313 4 增大到 45.552 8,F 值对应的概率 $\rho < 0.000\ 5$. 这些数据表明我们的回归模型是成立的,回归方程是有意义的.

2.3　回归结果分析

从多元线性回归模型中可以看到,国内生产总值、全国总人数、二级以上运动员人数、各级体育系统职工人数对世界冠军人数的影响系数分别为:−0.4,33,−2,1.5. 从数据来看,外部因素中全国人口总数与世界冠军人数关联度最大,因为人数多便于人才的挑选和选拔,更有利于从大量的人数中选拔出优秀的运动员. 内部因素中各级体育系统职工人数对世界冠军人数影响较大,因为高素质的体育系统职工人数便于队员的训练、管理和营养等机制的良好的运行,为运动员的发展和成长提供良好的环境. 国内生产总值和二级以上运动员人数对世界冠军人数的影响较弱,这是因为国内生产总值的高低直接影响国家对体育竞赛、训练等的资助力度,间接地反映了国家对竞技体育和竞赛等的支持力度和资源保证情况. 二级以上运动员人数与世界冠军人数的关联度反而较弱,这是因为我国实行的是举国体制,造就了许多优秀的高水平运动员,促使我国竞技体育快速发展起来,使我国变成竞技体育强国,但是目前我国并不是体育强国,因为我国虽然靠举国体制使竞技体育快速发展起来,但是群众体育和群众基础并不扎实,体育的普及性不高,高水平、中高水平、低水平的运动员与优秀的高水平的运动人数差别不是太大,没有呈现金字塔形状的运动员梯形队伍,专业运动员培养可持续发展性不是很好,并且专业运动员退役的安置制度不是很完善,造成了很大的社会问题,这些原因导致了我国二级以上运动员人数与世界冠军人数的关联度不高.

3 结 论

1)从灰色关联分析结果可以看出,外部因素中全国人口总数与世界冠军人数关联度最大,而内部因素中各级体育系统职工人数对世界冠军人数影响较大.国内生产总值和二级以上运动员人数对世界冠军人数影响相对较弱.

2)从获得世界冠军的人数与4种影响因素的多元线性回归模型可以看出,全国人口数有利于运动员的选拔;体育系统职工人数有利于运动员的训练、管理等机制的运行;国内生产总值间接地反映了国家对体育事业的支持力度;二级以上运动员人数是世界冠军人数的基础,因为我国是举国体制,虽然在短时间内训练出大量的世界冠军,但基础较弱,所以两者的关联度较低.

3)从长远来看,我国首先要大力发展经济,提高国内生产总值,优化人口数量和素质;其次,大力发展群众体育,提高我国后备人才力量,解决退役运动员的安置问题,提高运动员素质,使我国的竞技体育得到可持续发展;最后要发展我们的优势项目,大力开发潜优势项目,使我国的体育出现"遍地开花"的繁华景象.

参 考 文 献

[1] 姜桂萍. 杰出女性体育竞技人才成长特点及其时空分布特征研究——以我国历届夏奥会女性冠军为例[J]. 成都体育学院学报,2013(2):59-65.

[2] 赵光娟,肖枝洪. 历年中国运动员获世界冠军数的时序分析[J]. 湖北师范学院学报:自然科学版,2009,29(3):95-98.

[3] 张玉华. 奥运会奖牌数与5种影响因素的模型构建与定量分析[J]. 山东体育科技,2013,35(3):43-47.

[4] 杨延村,赵炳新. 基于残差分析的GM(1,1)模型有效性研究[J]. 控制与决策,2010,25(9):1413-1419.

[5] 鲍一丹,吴燕萍,何 勇. 基于GM(1,1)模型和线性回归的组合预测新方法[J]. 系统工程理论与实践,2004(3):95-98.

[6] 刘晓叙. 灰色预测与一元线性回归预测的比较[J]. 四川理工学院学报:自然科学版,2009,22(1):107-109.

[7] 刘 慧. 2001-2010年我国获世界冠军分析[J]. 体育文化导刊,2012(5):42-45.

[8] 汤 攀,许大庆. 中国乒乓球队囊括世界冠军分析[J]. 体育文化导刊,2010(2):36-38.

[9] 张玉华. 基于线性回归动态模型的中国第31届奥运会奖牌数预测[J]. 河南师范大学学报:自然科学版,2013,41(2):24-27.

[10] 刘国仕,何亮云,薛建华,等. 灰色线性回归组合模型在沉降监测中的应用[J]. 长沙理工大学学报:自然科学版,2012,9(4):32-36.

[11] 穆 瑞,张家泰. 基于灰色关联分析的层次综合评价[J]. 系统工程理论与实践,2008(10):125-130.

[12] 赵 聂. 时间序列模型在我国年度世界冠军预测中的应用[J]. 成都体育学院学报,2008,34(2):68-71.

[13] 冯守平,石 泽,邹 瑾. 一元线性回归模型中参数估计的几种方法比较[J]. 统计与决策,2008,24:152-153.

Analysis of Influencing Factors in Number of Chinese Population the World Champion Based on Multiple Linear Regression Model

RAN Yanen

(College of Physical Education, Zhengzhou University, Zhengzhou 450044, China)

Abstract: In this paper, using multiple linear regression, we analyzed the relationship between the number of China's world champion and four factors, i. e. China's population and GDP as external environment, and the number of more than second level athletes and the number of all levels of sports workers as internal environment. Firstly, by the gray realationsl analysis, we obtain the correlation degree between four influence factors as sub factors sequence and the number of world champions as maternal factors sequence. In order to know the function relationship further between each factor and the number of world champions, by multiple linear regression analysis, we construct the multiple linear regression model of the number of annual world champion and four influence factors, and prove that the model is in accord with the fact through the F test, providing a theoretical basis for the training of our athletes and national policy.

Keywords: the world champion number; the total number of population; gross domestic product; regression model; grey relational analysis