广西科学院学报
Journal of Guangxi Academy of Sciences

2010,26(1):13~16
Vol. 26,No. 1 February 2010

基于 Logistic 回归模型的大学生视力影响分析*

Analysis on Myopia Influence Factors of College Students Based on Logistic Regression Model

黄恒振,邓春亮,王朋炎,尹长明
HUANG Heng-zhen,DENG Chun-liang,WANG Peng-yan,YIN Chang-ming

(广西大学数学与信息科学学院,广西南宁 530004)
(Department of Mathematics and Information Science,Guangxi University,Nanning,Guangxi,541004,China)

摘要:采取随机抽样调查的方法,抽取 1610 名广西南宁各高校在校大学生,对可能影响视力的每天的上网(使用电脑)时间、睡眠时间、学习时间、运动时间、读书时的视线距离、做眼保健操的情况、吃甜食情况、吃果蔬情况、营养状况和双亲近视情况等 10 个因素进行问卷调查,所收集的数据用 SPSS13.0 软件进行处理,通过建立 Logistic 回归模型对影响近视的因素进行分析。结果发现上网(使用电脑)时间,学习时间,视线距离,做眼保健操情况,运动时间,睡眠时间及营养状况是影响学生视力的 7 大因素,其中上网(使用电脑)时间和学习时间是最主要的因素。

关键词:Logistic 回归分析 近视 模型检验 参数检验
中图法分类号:O212.2 **文献标识码**:A **文章编号**:1002-7378(2010)01-0013-04

Abstract: By random sampling 1610 college students who from nanning city of guangxi are extracted. A questionnaire survey includes 10 factors (time of online, time of sleeping, time of study, time of exercises, view distance, the circumstance of doing the eye exercises, nourishment condition, the situation of eating sweet and fruit, etc.) which infulence myopia probably. The data was analyzed by SPSS13.0. Through established logistic regression model, it was found time of on line, time of study, view distance, the circumstance of doing the eye exercises, time of exercises, time of sleep and nourishment condition are the 7 factors mainly affecting student's sight, in which, time of online and time of study are the key factors.

Key words: logistic regression, myopia, model checking, parameter testing

许多学生在大学期间视力下降得很快,在大学校园里,无论走到哪里,都可以碰到不少戴眼镜的学生。视力的下降,不论对以后的工作还是生活,都会带来很多的不便。关于近视方面的研究文献有不少[1~4],但大多是从医学角度去研究。影响大学生视力下降的因素很多,但是其中最关键的因素是什么,这些因素对近视的影响程度到底有多大,本文从数学的角度出发,通过统计调查及建立 Logistic 回归模型[5]对影响近视的因素进行具体的分析。

1 资料来源与研究方法

1.1 资料来源

按照随机抽样调查的方法,采取无记名问卷调查的形式,在广西南宁各高校共抽取 1610 名在校大学生,对可能影响近视的平均每天的上网(使用电脑)时间、睡眠时间、学习时间、运动时间、读书时的视线距离、做眼保健操的情况、吃甜食情况、吃水果情况、营养状况和双亲近视情况等 10 个因素进行调查,调查时间为 1 周。调查时,问卷由学生现场亲自填写并当场收卷。由于这次问卷调查的数量比较大,范围较广,所有问卷是看着学生亲自填写并当

收稿日期:2009-12-08

作者简介:黄恒振(1984-),男,硕士研究生,主要从事广义线性模型研究。

*广西自然科学基金项目(0832108)资助。

场收卷，因此这次调查的结果具有较强的普遍性与真实性。

1.2 研究方法

在文献研究的基础上，以大学生平均每天的上网(使用电脑)时间，睡眠时间，学习时间及运动时间，视线距离，做眼保健操情况，吃甜食情况，吃果蔬情况，营养状况，双亲近视情况等10个因素为自变量(表1)，以大学生近视与否为因变量构建Logistic回归模型进行分析，在数学模型中，自变量与因变量都采用虚拟变量[6](0～1变量或是0,1,2,3)表示，采用最大似然估计法(MLE)进行估计，利用SPSS13.0软件进行拟合和分析变量之间存在的关系。在很多现实研究问题中各自变量间存在一定程度的线性依存关系，被称作多重共线性，这种多重共线性关系常常会增大估计参数的均方误差和标准误，有的甚至使回归系数的方向相反，导致方程极不稳定，从而引起回归模型拟合上的矛盾及不合理。因此为避免自变量之间多重共线性对模型估计带来的影响，我们首先用SPSS13.0软件对自变量进行共线性诊断，然后进行Logistic回归分析。

表1 Logistic回归模型中的变量及赋值

变量代码	变量意义	赋值
Y	近视与否	0:不近视(视力≤100度);1:近视(视力>100度)。
x_1	每天上网(使用电脑)时间	0:上网时间<5h/d;1:上网时间≥5h/d。
x_2	学习时间	0:学习时间<8h/d;1:学习时间≥8h/d。
x_3	做眼保健操情况	0:每小时做一次;1:偶尔做;2:从来没做。
x_4	视线距离	0:≥33cm;1:<33cm
x_5	睡眠时间	0:睡眠时间≥8h/d,睡眠正常;1:睡眠时间<8h/d;睡眠不足。
x_6	运动时间	0:运动时间≥1h/d;1:运动时间<1h/d。
x_7	甜食情况	0:偶尔吃;1:经常吃。
x_8	果蔬情况	0:经常吃;1:偶尔吃。
x_9	营养状况(身体健康的BMI指数=体重(kg)/身高^2(m))	0:身体健康指数>25,营养过剩;1:身体健康指数18～25,营养正常;2:身体健康指数<18,营养不良。
x_{10}	双亲近视情况	0:父母不近视;1:单亲近视;2:双亲近视。

2 结果与分析

2.1 调查结果

在所调查的1610名大学生中，近视的有1252名，总近视发生率为77.8%。其中53.4%的学生平均每天使用电脑时间超过5h，43.4%的学生平均每天学习时间超过8h；只有2.7%的学生在使用电脑或学习期间进行有规律地做眼保健操或小歇(每小时做一次)，71.6%的学生只是偶尔做，25.7%的学生根本没做；70.2%的学生没有与书本距离保持33cm，即坐姿不正确；44.8%的学生平均每天睡眠时间不足8h；68.1%的学生平均每天体育锻炼不到1h；19.6%的学生营养不良；3.4%的学生双亲近视，17.8%的学生单亲近视；52.0%的学生经常吃甜食，34.5%的学生不经常吃蔬菜水果。

2.2 回归诊断

为了建立较好的模型，我们对自变量进行共线性诊断[7]，结果方阵$X'X$(其中$X=(x_1,\cdots,x_{10})$)的10个特征根分别为$\lambda_1=2.083$,$\lambda_2=1.693$,$\lambda_3=1.210$,$\lambda_4=0.934$,$\lambda_5=0.908$,$\lambda_6=0.801$,$\lambda_7=0.686$,$\lambda_8=0.659$,$\lambda_9=0.561$,$\lambda_{10}=0.464$,条件指数为

$$k=\sqrt{\lambda_1/\lambda_{10}}=\sqrt{2.083/0.464}\approx 2.119<30。$$

因而自变量间的多重共线性程度很小。再对模型进行异常点、强影响点诊断，诊断结果发现模型中并不存在对回归结果影响较大的异常点、强影响点，因而可直接做Logistic回归分析[8]。

2.3 Logistic回归分析

Logistic回归分析结果见表2和表3，求得Logistic回归模型为

$$P=[\exp(-5.281+2.509x_1+1.782x_2+1.373x_3+1.735x_4+0.705x_5+1.219x_6+0.289x_7+0.281x_8+1.09x_9+0.081x_{10})]/[1+\exp(-5.281+2.509x_1+1.782x_2+1.373x_3+1.735x_4+0.705x_5+1.219x_6+0.289x_7+0.281x_8+1.090x_9+0.081x_{10})]。\quad(1)$$

模型(1)的似然比检验结果$\chi^2=776.612$,$P<0.001$，说明拟合的模型具有显著性意义。再从表3判别分析可以看出模型错判学生近视98例，错判学生不近视36例，总符合率(1610－98－36)/1610＝91.7%，也就是说通过该方程预测结果正确率可以达到91.7%。表2回归系数的P值除了x_7,x_8,x_{10}外，其它的都小于0.005，通过了Wald检验，说明学生每天的上网(使用电脑)时间、学习时间、睡眠时间及运动时间，有无做眼保健操，视线距离及营养状况对近视与否有着显著的相关关系。而吃甜食、水果及双亲近视的情况没有通过检验。

表 2 二值 Logistic 回归模型分析结果

变量	系数	标准误	Wald 值	P 值	OR 值
x_1	2.509	0.205	150.449	0.000	12.290
x_2	1.782	0.199	80.005	0.000	5.941
x_3	1.373	0.254	29.209	0.000	3.947
x_4	1.735	0.184	88.748	0.000	5.667
x_5	0.705	0.207	11.565	0.001	2.025
x_6	1.219	0.180	45.681	0.000	3.383
x_7	0.289	0.177	2.679	0.102	1.335
x_8	0.281	0.185	2.303	0.129	1.324
x_9	1.090	0.218	25.064	0.000	2.974
x_{10}	0.081	0.159	0.259	0.611	1.084
常数	−5.281	0.422	156.969	0.000	0.005

表 3 分类表

近视与否	调查值	预测值		正确分类比例(%)
		正确数	错判数	
近视	1252	1216	98	97.1
不近视	358	260	36	72.6
总计	1610	1476	134	91.7

2.4 修正模型的回归分析

由于模型(1)中,系数 x_7,x_8,x_{10} 的 P 值分别为 0.102,0.129,0.611,没有通过检验。为了建立更精确的模型,我们对模型(1)进行修正。用后向消去法,剔除对模型影响不显著的变量 x_7,x_8,x_{10},对剩下的自变量重新做回归分析,结果见表 4 和表 5。

表 4 剔除变量 x_7,x_8,x_{10} 后的模型分析结果

变量	系数	标准误	Wald 值	P 值	OR 值
x_1	2.462	0.202	149.065	0.000	11.730
x_2	1.741	0.197	77.988	0.000	5.704
x_3	1.344	0.253	28.127	0.000	3.834
x_4	1.703	0.178	91.242	0.000	5.492
x_5	0.648	0.202	10.257	0.001	1.912
x_6	1.217	0.174	48.669	0.000	3.377
x_9	1.167	0.210	30.895	0.000	3.213
常数	−4.997	0.380	172.484	0.000	0.007

表 5 剔除变量 x_7,x_8,x_{10} 后的分类表

近视与否	调查值	预测值		正确分类比例(%)
		正确数	错判数	
近视	1252	1216	96	97.1
不近视	358	262	36	73.2
总计	1610	1478	132	91.8

这时,我们有回归方程

$$P=[\exp(-4.997+2.462x_1+1.741x_2+1.344x_3+1.703x_4+0.648x_5+1.217x_6+1.167x_9)]/[1+\exp(-4.997+2.462x_1+1.741x_2+1.344x_3+1.703x_4+0.648x_5+1.217x_6+1.167x_9)]。\quad(2)$$

此时模型的似然比检验的结果 $\chi^2=771.940$,$P<0.001$,模型显著性明显。由判别分析,模型预测结果的正确率为 91.8%。这时表 4 回归系数的 P 值都小于 0.005,通过了 Wald 检验。因而模型(2)具有显著的统计学意义。再从 OR 值(相对危险度的一种估计值)来看,都超过 1。综上分析说明影响近视的主要因素为上网(使用电脑)时间、学习时间、视线距离、做眼保健操情况、运动时间、营养情况和睡眠时间等 7 个因素。

2.5 影响因素分析

2.5.1 上网(使用电脑)时间

大学生每天上网时间超过 5h 得近视的相对危险性是平均每天上网时间小于 5h 的 11.73 倍。这次调查中,1610 名学生中有 859 名平均每天上网时间超过 5 小时 h,这 859 名学生中就有 775 位近视,近视率高达 90.2%。可见上网(使用电脑)时间过长是大学生视力下降最主要的因素,而且使用电脑的时间越长对眼睛的危害就越大。原因是多数大学生拥有个人电脑,且大学学习环境相对比较自由,有较多的课余时间上网聊天,玩游戏,看电影等。这提示我们要合理使用电脑,尽量避免长时间上网或使用电脑,以保护我们的眼睛。

2.5.2 学习时间

学生每天学习时间超过 8h 得近视的危险性是每天上网时间小于 8h 的 5.704 倍,调查的结果显示平均每天学习时间超过 8h 的学生中近视率达 88.7%。可见学习时间过长对视力的影响,它是影响视力下降的第二大因素。原因是大学里学习仍是最主要的任务,很多勤奋好学的学生为努力学习,平均每天学习时间都不只 8h,久而久之影响了视力。学习永远是学生的天职,是学生无法避免的任务。因此建议在校大学生在勤奋学习期间学会让眼睛休息,避免眼睛过度疲劳。

2.5.3 视线距离

学生看书或写字时,视线距离不到 33cm 得近视的危险性是视线距离保持 33cm 的 5.492 倍,如一个长时间使用电脑和长时间学习及视线距离没有 33cm,而其他因素都正常的学生,即他的各指标为 $x_1=1,x_2=1,x_3=0,x_4=1,x_5=0,x_6=0,x_9=0$,根据模型(2),他得近视的概率 $P=0.889$。在调查的 1610 名学生中有 1131 名学生读书时视线距离不到 33cm,近视率高达 88%,原因是他们在看书学习时没有保持正确的坐姿。若上面的学生学习时视线能保持 33cm 的距离,即 $x_4=0$,则他得近视的概率 $P=0.592$,降低了 0.297,因此建议在看书写字时一定要注意保持正确的坐姿,以保证视线有 33cm

的距离。

2.5.4 眼保健操

从不做眼保健操的学生得近视的危险性是经常有规律做眼保健操的3.834倍，尤其是那些长时间使用电脑和长时间学习的学生，不做眼保健操对眼睛伤害更大，调查中的1610名学生中有414名学生从来没做，近视率达94.2%。如一个长时间使用电脑和长时间学习从没做眼保健操，而其他因素都正常的学生，即他的各指标为 $x_1=1, x_2=1, x_3=2, x_4=0, x_5=0, x_6=0, x_9=1$，根据模型(2)，他得近视的概率 $P=0.955$。这次调查中从不做眼保健操及偶尔做眼保健操的学生共占97.3%，近视率达79.8%，其主要原因是大部分学生没养成做眼保健操的习惯。若上面的学生进行有规律的眼保健操，即 $x_3=0$，那么他得近视的概率为 $P=0.592$，降低了0.363，因此建议大学生培养做眼保健操的习惯，在学习或使用电脑期间至少每小时做一次眼保健操或是休息几分钟，对眼睛可以起到很好的保护作用。

2.5.5 运动时间

平均每天运动不到1h的学生得近视的危险性是平均每天运动1h以上的3.377倍，在所调查的学生中平均每天运动时间不到1h的占68.1%，近视率达88.2%。原因是一部分学生学习紧张，而更多的学生因为沉迷于电脑游戏、电影等不愿出外活动。为此建议在校大学生要多进行体育锻炼，至少每天保持1h的运动时间，这样可以对眼睛起到休息保护作用。

2.5.6 营养状况

营养不良的学生得近视的危险性是正常良好的3.213倍，原因可能是部分学生生活比较贫困，因节俭造成营养不良，还有很多学生偏食、挑食导致营养不均衡。为此建议广大学生不要偏食、挑食，注意保持营养良好平衡。

2.5.7 睡眠时间

学生睡眠不足得近视的危险性是睡眠充足的1.912倍，调查数据显示有44.8%的学生睡眠不足，其中平均每天上网时间超过5h及学习时间超过8h的学生中睡眠不足的达76.9%。原因可能是大学生科目多，学习负担重，需要较多的时间学习，或是沉迷于电脑游戏、看电影亦或是参加其他活动，导致睡眠时间不足。睡眠不足不但会引起注意力不集中，长此以往对眼睛带来的影响也是不容小视的，因此建议大学生除了学习和工作外尽量避免参加无意义或是没必要的活动以保证充足的睡眠。

至于吃甜食和果蔬，及父母亲的近视情况，在模型(2)中不显著，可能是由于所取样本受地域影响及调查数量有限，这有待进一步研究。

3 结束语

在这个科技信息迅速发展、竞争激烈的时代，大学生近视已是一个相当普遍的情况，学生的高近视率必须引起社会的普遍关注和重视。预防近视必须从娃娃抓起，因此提醒广大家长对小孩就本文提到的7个因素进行科学用眼教育。提醒已近视或没有近视的学生及其他社会人士要注意改善长时间上网或学习，睡眠不足，视线距离不够，不注意做眼保健操，营养不良，缺乏体育锻炼的状况，提倡劳逸结合，科学用眼，避免用眼过度。以保护视力及身心健康。

Logistic模型是一种广义线性模型，对变量的要求要宽松很多，不一定要求变量连续或服从正态分布，其分析结果中的OR值对危险因子的解释简洁明了，能避免其他非研究因素对模型的干扰和影响，可以快速的从众多的危险因素中筛选出与近视密切的因素。但是值得注意的是Logistic模型也如同线性模型一样需考虑回归诊断问题，很多文献都把这个问题忽略了。而本文模型恰好不存在共线性，所以可直接进行Logistic回归。Logistic模型其实也是一种概率模型，根据模型中自变量的取值能较准确的预测得近视的概率，因此该模型具有较强的适应性。

参考文献

[1] 贾艳合，陈会云，陈静，等. 1933名中学生近视发生率及影响因素现况调查[J]. 现代预防医学，2007，34(20)：3825-3827.

[2] 张雪飞，王平，曾巍，等. 武汉市武昌城区学生近视状况及影响因素分析[J]. 中国公共卫生，2007，23(6)：383-384.

[3] 汪芳润. 近视眼研究的现状与存在问题[J]. 中华眼科杂志，2003，39(6)：381-386.

[4] 中国学生体质与健康研究组. 2000年中国学生体质与健康调研报告[M]. 北京：高等教育出版社，2002：8-15.

[5] 王济川，郭志刚. Logistic回归模型——方法与应用[M]. 北京：高等教育出版社，2001：91-120.

[6] 张尧庭. 定性资料的统计分析[M]. 桂林：广西师范大学出版社，1991：114-144.

[7] 陈希孺. 广义线性模型(六)[J]. 数理统计与管理，2003，22(4)：58-63.

[8] 范金城，梅长林. 数据分析[M]. 北京：科学出版社，2002：124-135.

(责任编辑：韦廷宗)