

文章编号:2095-6134(2015)06-0728-07

现代变量选择方法在青少年近视研究中的应用*

海豹¹, 李仕明², 刘洛如³, 申立勇¹, 张三国^{1†}, 李偲圆², 李 翥³,
康梦田², 孙芸芸², 孟 博², 张庆昭¹

(1 中国科学院大学数学科学学院, 北京 101408; 2 首都医科大学附属北京同仁医院眼科中心, 北京 100730;

3 河南安阳市眼科医院, 河南 安阳 455000)

(2014 年 9 月 25 日收稿; 2015 年 3 月 16 日收修改稿)

Hai B, Li S M, Liu L R, et al. Juvenile myopia study using modern variable selection methods[J]. Journal of University of Chinese Academy of Sciences, 2015, 32(6): 728-734.

摘 要 通过分析一组医学数据挖掘出影响青少年近视的关键因素, 建立青少年近视患病概率预测模型. 数据集主要由两部分组成: 一是青少年眼睛的医学测量数据, 二是由生活学习习惯调查问卷得到的数据. 采用几种现代统计学方法, 并利用 ROC 曲线得到较优的患病概率模型. 结果表明, 性别、眼轴长度、角膜曲率、工作日睡眠时间、不戴眼镜远视力、远距离调节反应等因素对青少年近视有重要的影响作用, 并由此建立预测模型.

关键词 变量选择; logistic 回归; Lasso; MCP; ROC 曲线

中图分类号: 0212 文献标志码: A doi: 10. 7523/j. issn. 2095-6134. 2015. 06. 002

Juvenile myopia study using modern variable selection methods

HAI Bao¹, LI Shiming², LIU Luoru³, SHEN Liyong¹, ZHANG Sanguo¹, LI Siyuan²,
LI He³, KANG Mengtian², SUN Yunyun², MENG Bo², ZHANG Qingzhao¹

(1 School of Mathematical Sciences, University of Chinese Academy of Sciences,

Beijing 101408, China; 2 Beijing Tongren Eye Center, Beijing Tongren Hospital, Capital Medical University,

Beijing 100730, China; 3 Anyang Eye Hospital, Anyang 455000, Henan, China)

Abstract In this work we used some variable-selection techniques to find out the relevant factors that cause adolescent myopia, and established probabilistic models for myopia prediction. The research is based on a medical dataset consisting of two parts: medical measurement data of the youths and data on daily living habits obtained by questionnaire survey. We used some modern variable selection methods and the ROC curve to evaluate different modes. The results show that gender, axial length, corneal curvature, weekday sleeping time, distance vision without glasses, and remote adjustment reaction have important influences on adolescent myopia.

Key words variable selection; classical logistic regression; Lasso; MCP; ROC curve

* 国家 973 重点基础研究发展计划项目(2011CB504601)资助

† 通信作者, E-mail: sgzhang@ucas.ac.cn

据权威部门统计,目前中国共有4亿多近视眼患者,近视发病率达33.3%,其中青少年更是近视“重灾区”,小学生的近视率在40%左右,初中生近视率在60%左右,高中生近视率则高达70%以上.研究表明,过去几十年间很多地区面临近视患病率快速增长的情况^[1-2].因此,影响青少年视力健康的因素至关重要,因为这将决定应该如何有效预防近视或者防止近视程度的增加.

迄今为止,很多国内外学者一直在研究影响青少年近视的因素.Kathryn A. Rose等^[3]指出,较多户外活动(包括运动和休闲活动)的青少年具有较低的近视患病率,同时近距离工作较多并且户外活动少的青少年最容易近视;但是在一份纵向数据研究中^[4],近视和非近视的青少年之间的近距离工作量差距并不是很明显.此外,Amanda N. French等^[5]在其分组研究中指出在较小的年龄组中,较多的近距离工作量具有较高的近视患病率;然而在较高的年龄组中近距离工作量则不显著,并且父母双方近视的人数越多,则孩子的近视患病率越高.研究还表明不同种族对近视程度也有影响,相同种族的青少年在不同的地区近视患病率也不同^[2,6].另外,户外活动时间也是一个影响因素,基准线屈光度被认为是最重要的影响因素.同时,也有研究表明遗传因素对近视也有影响^[7-8].年龄、种族、母亲怀孕期间是否吸烟、脑瘫和唐氏综合症也是影响青少年视力的因素^[9].在国内,通过纵向数据获得青少年近视的患病和发病情况,并分析近视相关的影响因素^[10].目前为止,国内尚无基于本国数据进行视力影响因素变量选择的研究.

在以往近视因素的研究中^[3,5,9],大多利用边际相关性以及向前(向后)选择等传统变量选择方法.此种方法操作起来简单,但存在一些问题:首先,经常会发生重要变量与响应变量的边际相关系数很小,这样会导致选择出来的模型错误;其次,传统变量选择方法是不连续的,稳定性不强,数据的微小扰动可能导致选择的结果差异很大.相反,本文所使用的几种现代变量选择方法很好地克服了传统方法的缺陷,先用几种现代统计学方法建立预测模型,通过和传统的方法建立的模型相比较,得到最优模型.

1 数据来源

采用随机整群抽样方法.以学校为单位,在河

南安阳城区随机抽取4所初中^[11],对学生进行详细眼部检查和问卷调查.眼部检查过程中,采用1%环戊通和美多丽散瞳,电脑验光获得屈光度值并计算等效球镜度,近视定义为等效球镜 $\leq -0.5D$,非接触光学测量仪Lenstar LS900获取眼轴长度、角膜曲率和前房深度等.

经过数据初步处理之后,得到的完整的数据集共有1481个观测,46个变量,其中17个连续变量,4个0~1变量,25个多元属性变量.例如GENDER表示性别,RAXISLEG表示眼轴长度,st1工作日睡眠时间等等,更详细的变量标签见附录.考虑到各变量的数量级差距较大,在数据初步处理过程中已将连续变量标准化.

2 方法介绍

本文首先运用传统的Logistic回归的 P 值检验法(对照组);其次,由于数据集中有不少多元属性变量转化为哑变量,需要将它们分为一组,故可以用现代变量选择方法——Group Lasso和Group MCP选择变量;同时,若每组有多个变量,用上述方法不能实现组内挑选变量,这时采用Composite MCP方法建立模型.

这些方法有着各自不同的特点.传统Logistic回归方法简单易执行,原理简单;Group Lasso和Group MCP通过惩罚函数项,适当调节参数,可以组间挑选变量;composite MCP是Group MCP的推广,该方法可以进行双层变量选择——组内和组间,应用范围更为广泛.

2.1 变量选择方法介绍

鉴于数据集的响应变量为二值变量,可以用传统的Logistic回归模型.设响应变量为 $Y = (y_1, y_2, \dots, y_n)^T$, p 个自变量分别记为 x_1, x_2, \dots, x_p ,记 $\mathbf{x} = (1, x_1, x_2, \dots, x_p)^T$,在 p 个自变量的作用下出现成功的条件概率记为 $P\{Y = 1 | \mathbf{x}\}$,那么Logistic模型如下

$$P\{Y = 1 | \mathbf{x}\} = \frac{e^{\mathbf{x}^T \boldsymbol{\beta}}}{1 + e^{\mathbf{x}^T \boldsymbol{\beta}}} \in [0, 1],$$

其中, $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \dots, \beta_p)^T$, β_0 为常数项, $\beta_1, \beta_2, \dots, \beta_p$ 为Logistic回归系数.根据给定的观测,可以计算出该学生近视的概率,然后通过设定阈值(一般取0.5),来判定是否近视.

2.1.1 多元Logistic回归的 P 值判别法

该方法和医学上变量选择方法一致,通常根

据变量 P 值,逐步筛选变量(向后选择法).首先,让所有的变量都进入 Logistic 回归模型,然后找出 P 值最大的变量,然后将其删除,再重新建立 Logistic 回归模型,接着再删除 P 值最大的变量,依次类推,直至所有的变量都显著.一般给定显著性水平 α (通常取 0.05),当所有变量的 P 都满足 $P < \alpha$ 时,算法结束,最终得到的变量作为变量选择的结果,建立 Logistic 回归模型.

2.1.2 Group Lasso 和 Group MCP

数据集中存在一些属性变量,在计算时需要引入哑变量.因此在进行变量选择时必须要保证某属性变量的一个哑变量进入模型,则该变量的其余哑变量也必须进入模型.因此将各属性变量的哑变量分别分为一组,以保证同进同出模型.分组之后,本文采用 2 种不同的方法来进行变量选择.

1) Group Lasso

1996 年, Tibshirani^[12] 提出 Lasso 方法,该方法的线性模型如下

$$\min_{\beta} \left\{ \frac{1}{2n} \|Y - X\beta\|_2^2 + \lambda \|\beta\|_1 \right\}, \quad (1)$$

设数据集有 n 个观测, p 个属性, X 为设计矩阵.

Lasso 方法是在回归系数的绝对值之和小于一个常数的约束条件下,使残差平方和最小化,从而使得某些自变量的回归系数为 0,得到解释力较强的模型.该方法优点是速度快、连续,缺点不是无偏估计.

上述方法不涉及分组问题,如果自变量分成 m 个不同的组,可以通过选择若干组达到变量选择的目的,这就是 Group Lasso^[13],其线性模型如下

$$\min_{\beta} \left\{ \frac{1}{2n} \|Y - \sum_{l=1}^m X^{(l)} \beta^{(l)}\|_2^2 + \lambda \sum_{l=1}^m \sqrt{p_l} \|\beta^{(l)}\|_2 \right\}, \quad (2)$$

其中, $\beta^{(l)}$ 表示第 l 组的回归系数, p_l 为 $\beta^{(l)}$ 的长度, $X^{(l)}$ 是 $\beta^{(l)}$ 对应的设计矩阵, λ 为调节参数.

该优化问题是通过最小化一个“损失”+“惩罚”的函数问题来解决.

线性模型对应的“损失函数”是残差平方和项,而我们的数据的响应变量是二元属性变量,把残差平方和项作为其“损失函数”已不再合适,于是引入 Logistic 模型的对数似然函数 $L(\beta)$ 作为

其“损失函数”^[14],即

$$\min_{\beta} \left\{ -\frac{1}{n} L(\beta) + \lambda \sum_{l=1}^m \sqrt{p_l} \|\beta^{(l)}\|_2 \right\},$$
$$L(\beta) = \sum_{i=1}^n \sum_{l=1}^m y_i X_i^{(l)} \beta^{(l)} - \sum_{i=1}^n \log(1 + e^{\sum_{l=1}^m X_i^{(l)} \beta^{(l)}}), \quad (3)$$

其中, $X_i^{(l)}$ 表示 $X^{(l)}$ 的第 i 行.

需要指出的是, Group Lasso 方法是 Lasso 方法的推广,应用范围更为广泛.如果 Group Lasso 中的每组只有一个变量, Group Lasso 则退化为 Lasso 方法.鉴于响应变量是二元属性变量,本文只用到各方法的 Logistic 模型形式,故后面的 group MCP 和 composite MCP 模型只介绍其 Logistic 模型,而线性模型则不再介绍.

2) Group MCP

MCP (minimax concave penalty) 方法^[15] 和 Lasso 类似,通过不同的惩罚函数,来进行变量选择. MCP 方法具有连续性,稀疏性及无偏性.

Group MCP^[16] 的 Logistic 模型如下:

$$\min_{\beta} \left\{ -\frac{1}{n} L(\beta) + \sum_{l=1}^m f_{\lambda, \gamma}(\sqrt{p_l} \|\beta^{(l)}\|_2) \right\},$$
$$f_{\lambda, \gamma}(t) = \lambda \int_0^t (1 - x/\gamma\lambda)_+ dx$$
$$= \begin{cases} \lambda t - t^2/2\gamma, & \text{若 } t \leq \gamma\lambda; \\ \gamma\lambda^2/2, & \text{若 } t > \gamma\lambda. \end{cases} \quad (4)$$

该惩罚函数具有连续导数

$$f'_{\lambda, \gamma}(t) = \begin{cases} \lambda - t/\gamma, & \text{若 } t \leq \gamma\lambda; \\ 0, & \text{若 } t > \gamma\lambda. \end{cases}$$

设数据共有 m 个组,第 l 组共有 K_l 个变量; $\lambda \geq 0$ 为调节参数, $L(\beta)$ 式同(3),参数 $\gamma > 0$,在变量标准化的情况下^[16],一般取 $\gamma = 3$.

2.1.3 Composite MCP

按照 2.1.2 的分组方法,数据分的组数多且组内无法变量选择.从另一个角度考虑分组问题,根据定性分析,将变量分为 7 组,分别为遗传因素、个体参数、近距离工作量、户外活动量、用眼习惯、生活环境和饮食习惯.由于每组内变量的重要程度不尽相同,若按照 2.1.2 的方法,则会将一些不重要的变量纳入模型,故在选择分组后,最好能在挑选的组内进行二次筛选,剔除一些不重要的变量,从而达到组内变量选择的效果.

Composite MCP 方法^[17] 可以进行双重变量选择.

它的 Logistic 模型如下

$$\min_{\beta} \left\{ -\frac{1}{n}L(\boldsymbol{\beta}) + \sum_{l=1}^m f_{\lambda,a} \left(\sum_{k=1}^{K_l} f_{\lambda,\gamma}(|\beta_{lk}|) \right) \right\}, \tag{5}$$

设 β_{lk} 表示第 l 组的第 k 个变量的回归系数, $\lambda \geq 0, f_{\lambda,\gamma}(t), L(\boldsymbol{\beta}), K_l$ 同式(3)、式(4), $a = K_l \gamma \lambda / 2$ 为外部调节参数, 参数 $\gamma > 0$, 通常情况下取 $\gamma = 3$.

上面介绍了 4 种具体的建模方法, 响应变量都是二元变量, 因此可以用 ROC 曲线评价模型的好坏.

2.2 ROC 曲线

ROC 曲线^[18]即接受者操作特征曲线(receiver operating characteristic curve), 是一种坐标图式的判断一个二元分类器性能的分析工具.

ROC 曲线定义 Y 轴为灵敏度(sensitivity)或者真阳性率(TPR, 真阳性率指在所有实际为阳性的样本中, 被正确地判断为阳性之比率), 同时定义 X 轴为 $1 - \text{特异度}(1 - \text{specificity})$ 或者假阳性率(FPR, 假阳性率指在所有实际为阴性的样本中, 被错误地判断为阳性之比率).

对于一个二元分类器, 当给定一个阈值时, 将得到一个具体的 (FRP, TPR) 数据对, 将其和 ROC 曲线中的一个点对应起来; 当选取一组不同的阈值时, 可以得到 ROC 曲线中的一系列点, 并将这些点按顺序连接起来, 得到一条 ROC 曲线. 判断一个分类器好坏的标准是 ROC 曲线下的面积(AUC, the area under the curve). 当分类器趋于完美分类器时, 每个点 (FRP, TPR) 都会向 (0, 1) 靠近, 所以当 AUC 面积越大, 分类器越好.

本文将该数据量的 90% (1 332) 用来建立模型, 10% (149) 用来绘制对应模型的 ROC 曲线, 然后通过比较几种不同模型的 AUC, 从而选出较优的模型分类器.

3 实验结果及分析

3.1 传统 Logistic 回归模型的 P 值筛选法

该方法即向后选择变量法, 鉴于篇幅问题, 向前和逐步选择变量法未列出.

第一步, 计算运用 Logistic 回归模型, 计算所有自变量的 P 值, 其中属性变量 FAST 的 $P = 0.999\ 3$, 故先将该变量去掉; 第二步, 重新建立模

型, 得到具有最大 P 值 (0.977 6) 的变量 AMB, 将其删除; 以此类推, 直至所有留下来的变量的 P 值显著 ($P < 0.05$), 最终得到的变量如表 1.

表 1 变量选择信息
Table 1 The information of variable selection

Parameter	Estimate	Standad error	Wald Chi-Squae	Pr > ChiSq
Intercept	2. 30	0. 72	10. 16	0. 001 4
GENDER	-0. 96	0. 30	10. 49	0. 001 2
DRNBASE	2. 58	0. 30	72. 17	<0. 000 1
NRNBASE	-0. 67	0. 34	3. 85	0. 049 6
RCORCU	1. 49	0. 23	41. 46	<0. 000 1
RAXISLEG	2. 69	0. 35	57. 60	<0. 000 1
RPR	-1. 11	0. 21	26. 70	<0. 000 1
YTR	-1. 73	0. 24	51. 36	<0. 000 1
st1	-0. 33	0. 14	5. 60	0. 0179
TUTOR1	0 1. 84	0. 71	6. 80	0. 009 1
TUTOR1	1 1. 50	0. 72	4. 38	0. 036 4

最终得到 9 个显著变量, 分别是 GENDER、DRNBASE、NRNBASE、RCORCU、RAXISLEG、RPR、YTR、st1 和 TUTOR1, 根据表 1 可以建立青少年近视概率 P 与上述变量的模型, 其中 TUTOR1 为属性变量, TUTOR1 = 0 表示没有参加户外类型辅导班, TUTOR1 = 1 表示参加户外类型辅导班, TUTOR1 = 2 表示不确定是否参加此类辅导班, 则有下列的预测模型:

$$f_1(x) = 2.30 - 0.96 \times \text{GENDER} + 2.58 \times \text{DRMBASE} - 0.67 \times \text{NRNBASE} + 1.49 \times \text{RCORCU} + 2.69 \times \text{RAXISLEG} - 1.11 \times \text{RPR} - 1.73 \times \text{YTR} - 0.33 \times \text{st1} + 1.84 \times \text{I}(\text{TUTOR1} = 0) + 1.50 \times \text{I}(\text{TUTOR1} = 1)$$

则 $P = e^{f_1(x)} / (1 + e^{f_1(x)})$, 其中 $\text{I}(\cdot)$ 为示性函数.

3.2 Group Lasso 和 Group MCP 模型

将属性变量转化的哑变量分为一组和连续变量共得到 46 组变量. 下述 2 种方法, 都是运用交叉验证的方法, 通过调整 λ , 找出使得交叉验证错误达到最小的 λ , 从而选出最优的模型.

3.2.1 Group Lasso

该方法共选择 20 组 (其中 8 组哑变量), 共 36 个变量, 分别为: GENDER、RANTECHA、RNCONPRE、IFTAI、BULB、DRNBASE、RCORCU、RAXISLEG、RPR、JTR、YTR、st1、TUTOR1、TUTOR2、READN1、JUICE、FAST、COLA、SUNP、DAI.

交叉验证选择过程如图 1, 横坐标表示调节

参数 λ 的对数值,纵坐标表示交叉验证错误. 每个 λ 对应一个交叉验证错误和变量选择的组数,例如当 $\lambda = 0.0056$ 时,交叉验证错误最小为 0.33,此时选择了 20 组变量.

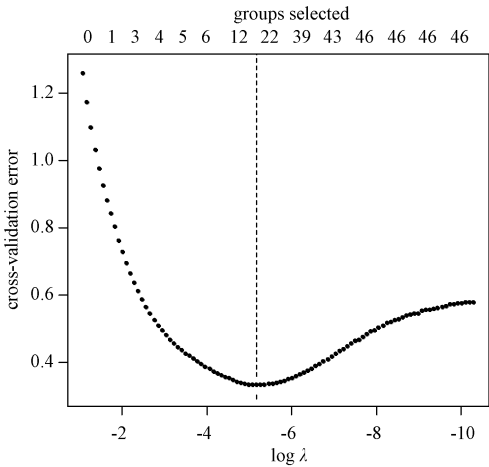


图 1 通过交叉验证错误率选择组数的过程
Fig.1 The process of group selection by cross-validation of error rates

3. 2. 2 Group MCP

与 Group Lasso 方法相比,Group MCP 采用不同的惩罚函数. 运用与 3. 2. 1 中相同的原则,根据交叉验证错误选择最佳模型.

最优模型在 $\lambda = 0.0108$ 处取得最小交叉验证错误 0.33,进入模型的变量有 8 组,分别为 GENDER、DRNBASE、RCORCU、RAXISLEG、RPR、YTR、st1 和 TUTOR1,具体模型如下

$$f_2(x) = 2.23 - 0.86 \times \text{GENDER} + 2.20 \times \text{DRNBASE} + 1.35 \times \text{RCORCU} + 2.42 \times \text{RAXISLEG} - 0.99 \times \text{RPR} - 1.70 \times \text{YTR} - 0.29 \times \text{st1} + 1.73 \times \text{I}(\text{TUTOR1} = 0) + 1.41 \times \text{I}(\text{TUTOR1} = 1),$$

则 $P = e^{f_2(x)} / (1 + e^{f_2(x)})$.

3. 3 Composite MCP 模型

该方法既考虑组间稀疏,又考虑组内稀疏. 这种方法是 3. 2. 2 的方法的拓展,具有更好的应用. 根据最小交叉验证错误原则,最优模型选择 2 个组,共 7 个变量,此时 $\lambda = 0.0144$,7 个变量分别为 GENDER、DRNBASE、RCORCU、RAXISLEG、RPR、YTR、st1,在模型中对应的回归系数分别为 2.23(常数项)、-0.86、2.20、1.35、2.42、-0.99、-1.70、-0.29.

3. 4 各模型 ROC 曲线比较

各模型的 ROC 曲线如图 2,图 3 是图 2 中方

框内曲线的放大.

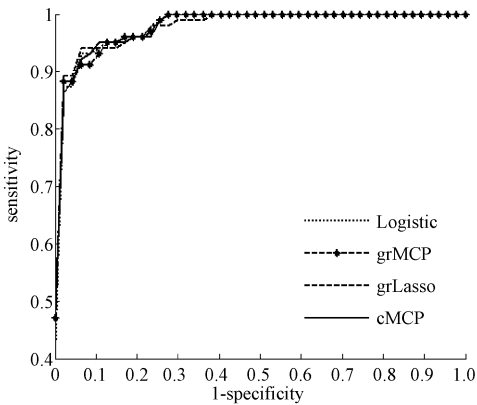


图 2 各模型的 ROC 曲线比较图
Fig.2 The ROC curve comparison among the models

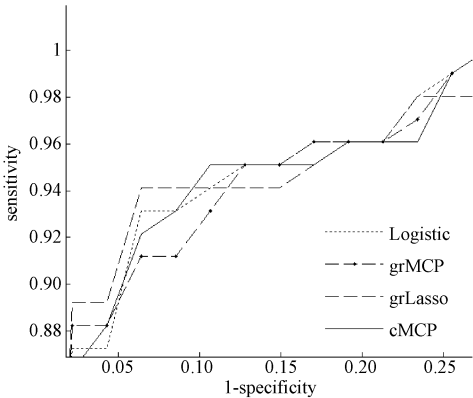


图 3 图 2 中方框内图形的放大
Fig.3 Enlarged view of the box in Fig.2

各模型 ROC 曲线的 AUC 信息如表 2.

表 2 各模型的 AUC 信息
Table 2 The AUC information of the models

模型名称	AUC	AUC 95% 置信区间
Logistic	0.973 1	[0.946 3,0.999 9]
grLasso	0.973 5	[0.948 4,0.998 6]
grMCP	0.973 1	[0.948 4,0.998 6]
cMCP	0.973 7	[0.948 8,0.998 6]

通过表 2 中 AUC 大小的比较,现代统计方法的结果都不比传统 Logistic 回归模型差. 其中 CMCP 模型的 AUC 最大,共选择了 7 个变量,所以应把该模型作为最佳模型,用来作为判断青少年近视患病率的标准.

4 结论和讨论

各个变量在 4 个模型中出现的频率直方图如图 4. 经比较发现,共有 7 个变量出现在 4 个

模型中,并且正好是 Composite MCP 方法选择的变量,可见这 7 个变量是影响青少年近视的重要因素,这些变量分别为性别(GENDER),眼轴长度(RAXISLEG),角膜曲率(RCORCU),工作日睡眠时间(stl),不戴眼镜远视力(DRNBASE),散瞳前电脑验光度数(RPR),远距离调节反应(YTR)。

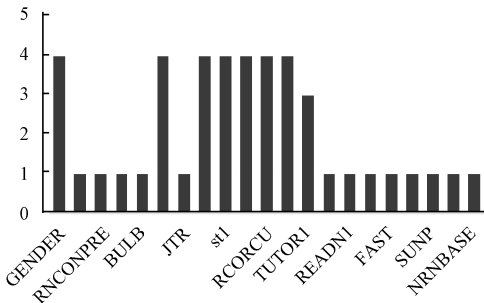


图 4 各变量选择方法的频率统计

Fig. 4 Frequency statistics of variable selection methods

从各模型的 AUC 来看,其 AUC 都比较大,接近于 1,这说明各模型所选的变量都有一定的参考价值。从上面的变量可以看出,青少年应该注意平时的睡眠时间,尽量不要熬夜,这样有利于近视的预防及减缓。从图 4 可以看出,是否参加户外类辅导班(TUTOR1)出现在 3 个模型中,该变量也可以参考,可以认为该变量是影响近视的一般因素,在平时也应该注意;同时,虽然有些变量只进入一个模型,例如父母戴眼镜的人数(DAI),在一些研究中^[5],该变量是一个显著的影响因素,我们可以作为参考;还有一些变量例如是否使用台灯(IFTAI),使用何种灯泡(BULB),虽然最终没有选择这些变量,但是这些变量和青少年的学习生活息息相关,也应给予关注;在饮食方面,对于果汁类,快餐类,碳酸饮料类食品可能也是影响青少年视力的次要因素。

综上所述,影响近视的因素是多方面的,要想有效的预防近视或者减缓近视的发展,青少年要养成良好的学习生活习惯,例如合理安排作息时

neighbouring Malaysia and Singapore[J]. British journal of ophthalmology, 2006, 90(10): 1 230-1 235.

[3] Rose K A, Morgan I G, Ip J, et al. Outdoor activity reduces the prevalence of myopia in children[J]. Ophthalmology, 2008, 115(8): 1 279-1 285.

[4] Jones-Jordan L A, Mitchell G L, Cotter S A, et al. CLEERE Study Group. Visual activity before and after the onset of juvenile myopia[J]. Invest Ophthalmol Vis Sci, 2011, 52: 1 841-1 850.

[5] French A N, Morgan I G, Mitchell P, et al. Risk factors for incident myopia in Australian schoolchildren: the Sydney adolescent vascular and eye study[J]. Ophthalmology, 2013, 120(10): 2 100-2 108.

[6] Rose K A, Morgan I G, Smith W, et al. Myopia, lifestyle, and schooling in students of Chinese ethnicity in Singapore and Sydney[J]. Archives of ophthalmology, 2008, 126(4): 527-530.

[7] Mutti D O, Mitchell G L, Moeschberger M L, et al. Parental myopia, near work, school achievement, and children's refractive error [J]. Investigative ophthalmology & visual science, 2002, 43(12): 3 633-3 640.

[8] Ip J M, Huynh S C, Robaei D, et al. Ethnic differences in refraction and ocular biometry in a population-based sample of 11-15-year-old Australian children[J]. Eye, 2008, 22(5): 649-656.

[9] Borchert M S, Varma R, Cotter S A, et al. Risk factors for hyperopia and myopia in preschool children: the multi-ethnic pediatric eye disease and Baltimore pediatric eye disease studies[J]. Ophthalmology, 2011, 118(10): 1 966-1 973.

[10] Li S M, Liu L R, Li S Y, et al. Design, methodology and baseline data of a school-based cohort study in central China: the Anyang childhood eye study [J]. Ophthalmic Epidemiology, 2013, 20:348-359.

[11] 李嵩,李仕明,刘洛如,等.河南安阳初中学生眼屈光度及生物学参数分布[J].中华医学杂志,2014,94(17):1 284-1 288.

[12] Tibshirani R. Regression shrinkage and selection via the lasso [J]. Journal of the Royal Statistical Society. Series B (Methodological), 1996: 267-288.

[13] Yuan M, Lin Y. Model selection and estimation in regression with grouped variables [J]. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2006, 68(1): 49-67.

[14] Meier L, Van De Geer S, Bühlmann P. The group lasso for logistic regression [J]. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2008, 70(1): 53-71.

[15] Zhang C H. Penalized linear unbiased selection [J]. Department of Statistics, Rutgers University, 2007: 2 007-2 003.

[16] Breheny P, Huang J. Penalized methods for bi-level variable

参考文献

[1] Morgan I G, Ohno-Matsui K, Saw S M. Myopia[J]. The Lancet, 2012, 379(9827): 1 739-1 748.

[2] Saw S M, Goh P P, Cheng A, et al. Ethnicity-specific prevalences of refractive errors vary in Asian children in

selection[J]. Statistics and its Interface, 2009, 2(3): 369-380.

[17] Huang J, Breheny P, Ma S. A selective review of group selection in high dimensional models[J]. Statistical Science, 2012, 27: 481-499.

[18] Zhou X H, McClish D K, Obuchowski N A. 诊断医学统计学[M]. 北京:人民卫生出版社,2005:13.

附录

各变量对应的标签:

变量	标签
RA	是否近视
GENDER	性别
RAXISLEG	右眼眼轴
RANTECHA	右眼前房深度
RNCONPRE	右眼眼压均值
HEIGHT	身高
WEIGHT	体重
YOU3	独立阅读年龄
ETEST	是否定期做眼科检查
AMB	是否有弱视
DIST	读书时脸到书本的距离
COSTM	是否有歪头看字的习惯
FTP	写字时手指尖距离笔尖的距离
CTT	看电视时离电视机的距离
REST	连续读书多久停下来休息一会
TUTOR1	是否参加户外运动类辅导班
TUTOR2	是否参加室内学习类辅导班
READN1	每周阅读多少页
IFTAI	读书时,是否采用台灯
BULB	读书时,用哪种照明灯泡
TWORK	连续不断地近距离工作多长时间才停下来休息一会
TUTOR	是否有家教、音乐课、美术课、辅导班
EYE	是否做眼保健操
FEEL	按摩时是否有酸胀感觉
JUICE	过去 4 周内,喝 100% 水果汁频率
REDM	过去 4 周内,吃红肉频率
BEAN	过去 4 周内,吃豆类食品频率
FRIES	过去 4 周内,吃膨化食品频率
FAST	过去 4 周内,吃快餐频率
SUGAR	过去 4 周内,吃糖果频率

ICECRM	过去 4 周内,吃甜食频率
COLA	过去 4 周内,喝碳酸饮料频率
DRINK	过去 4 周内,喝运动饮料频率
DRTEA	过去 4 周内,喝茶频率
SUNP	暑假,使用防晒霜频率
DRNBASE	不戴眼镜远视力
NRNBASE	不戴眼镜近视力
RPR	散瞳前电脑验光
JTR	近距离调节反应
YTR	远距离调节反应
RCORCU	角膜曲率
DAI	父母带眼镜的人数
ACI	上学期间近距离工作量
BCI	上学期间远距离活动量
CCI	假期间近距离工作量
DCI	假期远距离活动量
st1	工作日睡眠时间

附录中共有 46 个变量,二元属性变量有 RA, GENDER,IFTAI,BULB,TUTOR;三元属性变量有 ETEST, AMB, COSTM, TUTOR1, TUTOR2, READN1,EYE,DAI;四元属性变量有 FEEL;五元属性变量有 SUNP,JUICE,REDM,BEAN,FRIES, FAST,SUGAR,ICECRM,COLA,DRINK,DRTEA, DIST,FTP,CTT;七元属性变量有 REST, TWORK. 连续变量共 17 个,RAXISLEG 单位 mm,取值范围[18.98,28.09];RANTECHA 单位 mm,取值范围[2.06,4.49];RNCONPRE 单位 mmHG,取值范围[6,28];HEIGHT 单位 cm,取值范围[115,179];WEIGHT 单位 kg,取值范围[20,85];YOU3 单位 a,取值范围[3,12];DRNBASE 取值范围为[−0.22,1];NRNBASE 取值范围[−0.30,1];RPR 的取值范围为[−8.875,6.125];JTR 的取值范围为[−8.93,3.57];YTR 取值范围[−7.955,5.225];RCORCU 单位 mm,取值范围[38.615,48.99];ACI 单位 h,取值范围[20.75,137.67];BCI 单位 h,取值范围[8.5,225];CCI 单位 h,取值范围[24.67,177.38];DCI 单位 h,取值范围[16.25,225.5];st1 单位 h,取值范围[4,18.5].