西安电子科技大学 2019 年数学建模校内赛

承 诺 与 产 权 转 让 书

我们完全明白，在竞赛开始后参赛队员不能以任何方式（包括电话、电子邮件、网上咨询等）与队外的任何人研究、讨论与赛题有关的问题。

我们知道，抄袭别人的成果是违反竞赛规则的，如果引用别人的成果或其他公开的资料（包括网上查到的资料），必须按照规定的参考文献的表述方式在正文引用处和参考文献中明确列出。

我们郑重承诺，严格遵守竞赛规则，以保证竞赛的公正、公平性。如有违反竞赛规则的行为，我们将受到严肃处理。

我们同意将参赛论文以及支撑材料中的所建模型、算法以及程序产权归属西安电子科技大学以及合作单位共有。特别的，B 题参赛论文以及支撑材料中的相应产权西安电子科技大学拥有 50%，合作单位享有 50%。2019 年数学建模校内赛竞赛组委会，可将我们的论文以任何形式进行公开展示（包括进行网上公示，在书籍、期刊和其他媒体进行正式或非正式发表等）。

我们参赛选择的题号是：A

参赛报名队号为：19B100

报名时所属学院：电子工程学院

参赛队员姓名与学号：

1. 张校煜 17020130059

2. 刘祎敏 17020130084

3. 王显 18130500065

日期：2019 年 5 月 3 日

西安电子科技大学 2019 年大学生数学建模校内赛

评 阅 专 用 页

	评阅人 1	评阅人 2	评阅人 3	总评
成绩				

对近视的预测模型与预警机制

摘要

本文用统计回归模型模拟视力演化，用概率模型（非机器）、支持向量机（机器）预测视力变化，并在 AI 图像识别、数据获取技术的帮助下，设计出了一套完整的可实施的，面向家长、老师的近视预测预警机制。

问题一、二要求建立视力影响因素的量化模型和视力演变机理模型。通过文献调研及计算可能因素（例：双亲视力等）的相关系数，我们确定了主要影响视力的因素（学习、睡眠等 7 项），并建立了它们的量化模型。进一步，我们从统计回归模型入手，用最优逐步回归法，把以上 7 个变量逐个引入，检验评估其显著性，移出影响不显著的自变量。MATLAB 给出的最终视力演变机理模型中，“用眼距离”和“暗光看书”影响最显著，对应系数分别为-3.29200、3.51521。最后，我们从相关系数矩阵的角度检验，认为模型拟合度较好。

问题三要求建立基于人工智能的视力预警机制模型。该机制中，第一步由 AI 用眼数据采集模型通过 Adaboost 学习算法和霍夫变换算法实现瞳孔定位，采集学生“用眼距离”数据。第二步，我们采用了概率模型，计算数据箱线图的中位数，上、下四分位数和对应的视力值，并用基于最小二乘法的多项式拟合处理数据。第三步，在拟合基础上得出 4 组数据对应患近视（轻度、中度和重度）的概率。通过通知家长老师、远程管控电子设备等手段，实现对学生视力的预测预警。

问题四、五要求建立可供学习的信息系统和视力机器学习模型，并给出实现方案。我们采用了支持向量机作为机器学习策略，将数据库中的两千名青少年眼健康数据分为训练集和验证集。模型用交叉验证网格搜索的方法找到最佳损失函数乘法因子，再进行训练，预测。最终结果显示，准确性最高达 53.33%。为在实际中实现这些模型，我们从规范数据获取渠道和增加干预措施有效性的角度，给出了面向家长、老师的近视预防方案。

关键词：支持向量机 最小二乘法 逐步回归法 统计回归模型 近视预测

一、 问题重述

1.1 问题背景

紧张的学习生活节奏下，青少年近视问题日趋严重并且呈现低龄化的态势。先天因素、用眼习惯、用眼环境等影响着学生的视力。数据显示，2018年，我国小、中、高中学生近视比例分别为45.7%、74.4%、83.3%，大学生近视比例则高达87.7%。全国有近视眼的中小学生预估已超过1亿人。如何及早地预测、预警近视，使学校或家长有效地预防或矫正近视，这影响着军检和征兵，关系着万千中国家庭，是涉及民生的重大公共卫生问题和社会问题。

1.2 问题提出

我们需要从教育管理系统及其相关软件，硬件收集不同的数据，并借此通过数学模型分析预测学生的近视发生概率，从而及时预警，使学校或家长可以尽早进行干预，有效预防或矫正近视。为此，请讨论以下问题：

(1)提出影响视力的关键因素及其量化模型；
(2)提出具有可执行性的视力演化机理模型；
(3)提出基于人工智能的视力预警机制模型；
(4)提出可供学习的信息系统、进行眼睛视力的机器学习模型；
(5)就如何实现对应的机理模型和学习模型提出一份方案。

二、问题分析

2.1 问题一分析

为了提出影响视力的关键因素及其量化模型，我们可以调研文献，据此确定调查类目。通过随机整体抽样，可以获得所需数据并计算其相关系数，相关系数较大的可视为影响视力的关键因素。

2.2 问题二分析

为了提出具有可执行性的视力演化机理模型，我们从统计回归模型入手，采用最优逐步回归法，分析每种因素对视力的显著性，找到影响较为显著的几个因素。我们可以确定一个包含若干自变量的初始集合，在显著性水平标准下，对其引入、检验、移出变量，依此进行，直到不能引入、移出为止，从而得到反映眼睛视力演化机理的回归方程。

2.3 问题三分析

为了提出基于人工智能的视力预警机制模型，我们可以用 AI 算法收集实时的学生用眼数据，通过概率模型和统计回归模型对近视加重可能性进行预测，AI 判断是否向发出学生近视预警，再通过互联网终端将信息反馈给老师和家长，对学生活动采取一定的管控措施。

2.3 问题四分析

为了设计可供学习的信息系统的眼睛视力机器学习模型，我们认为随机森林、BP 神经网络等算法会出现过度拟合、复杂低效等缺点。我们可以采用 SVM 支持向量机进行机器学习。我们可以把数据分为训练集和测试集，在此基础上可以利用网格搜索法最佳参数，进而训练求解。

2.3 问题五分析

为实现对应机理模型和学习模型，我们主要的问题是可靠的数据来源。问题二和问题三的模型是用数理统计的方法建立的，问题四中机器学习也需要大量的数据学习预测，数据的来源与普适性对我们的模型很重要。因此，我们可以精心设计一份调查问卷，合理可靠的数据对模型的实现有很大意义。

三、模型假设

(1)不考虑近视伴随的其他症状影响，包括斜视，散光，假性近视等;

(2)认为左右眼近视没有差异，即忽略用眼时角度因素;

(3)认为近视度数小于 75°时不近视，75°到 300°为轻度近视，325°到 600°为中度近视，600°以上为重度近视。

四、符号说明

符号	说明
t_1	日均学习时间
t_2	日均使用电子产品时间
t_3	日均户外运动时间
t_4	日均睡眠时间
t_5	日均最长连续用眼时间
d	平均用眼距离
l	暗光下用眼频率
q	父母平均屈光度
Q	学生预测视力

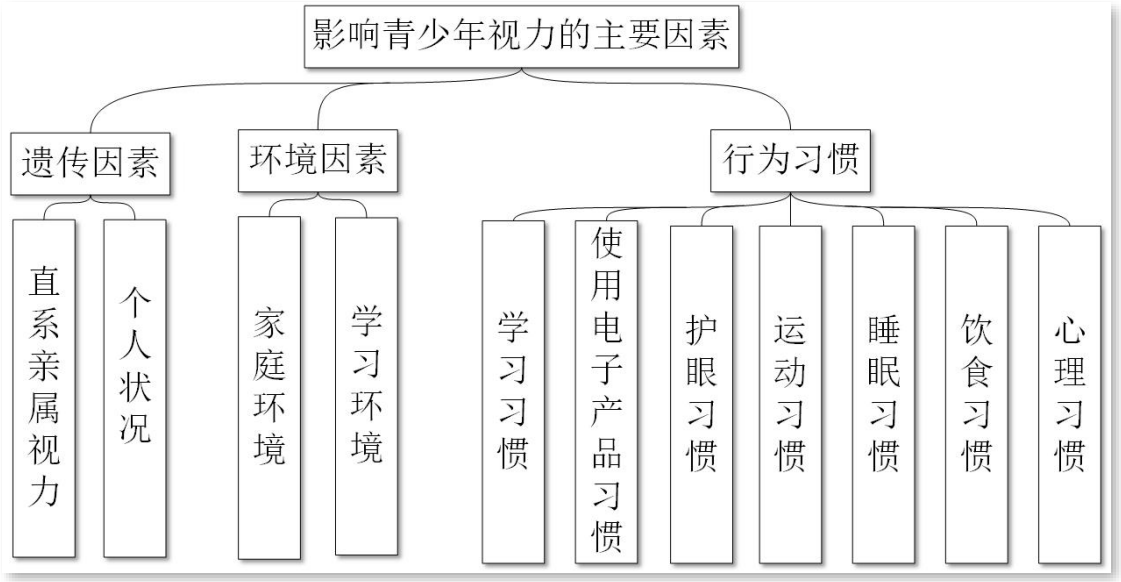
五、模型的建立与求解

5.1 问题一：影响视力的关键因素及其量化模型

5.1.1 模型准备

通过文献调研，我们找到了十个最影响视力的因素：直系亲属视力、个人状况（年龄和性别）、家庭环境、学习环境、学习习惯、使用电子产品习惯、护眼习惯、运动习惯、睡眠习惯、饮食习惯和心理习惯。

图 1 影响青少年视力的主要因素



就此，我们设计了《儿童青少年近视调查（初稿）》[1]，采取横断面研究，随机整体抽样，有效调查了 82 名小学生和 201 名大学生。它们与广州 2042 名青少年连续 5 年视力情况数据一起，为确立影响视力的重要因素提供了坚实的数据基础。

5.1.2 建立与求解

以上的工作中，我们收集了 82 名小学生和 201 名大学生以及广州 2024 名青少年的眼健康数据[5]。为了缩小研究对象，确定最影响视力的因素，我们将皮尔森相关系数应用到影响力大小的评估上。用 SPSS 软件可便捷地求出各个因素与近视的相关系数（保留四位有效数字）。

表 1 各因素与视力的皮尔森相关系数值（由高到低）

变量	r	变量	r
年龄	0.4316	照明条件	0.1019
长时间用眼	0.1812	户外运动	0.0994
睡眠时间	0.1783	学习时间	0.0803
父母视力	0.1601	饮食习惯	0.0689
用眼距离	0.1399	性别	0.0024
电子产品	0.1219	……	

皮尔森相关系数的绝对值越大，则相关性越强；反之，越弱。在 0.4 到 0.6 范围内的被认为为中等相关，而 0.2 以下的认为是极弱相关。

因此在这里，我们排除年龄、性别等不可改变的影响因素，取相关性较大的八个因素：日均学习时间t_1、日均使用电子产品时间t_2、日均户外运动时间t_3、日均睡眠时间t_4、日均最长连续用眼时间t_5、平均用眼距离d、暗光用眼频率l、直系亲属视力q。它们对视力的影响可以表示为：

$$Q = f(\varphi_1, \varphi_2, \varphi_3, \cdots, \varphi_n) \tag{1}$$

并且，建立该八个因素的量化关系如下：

表 2 主要影响视力因素的赋值详情

变量	0	1	2	3
因变量				
视力	<0.75D	≥ 0.75且< 3	≥ 3且< −6	≥ −6
自变量				
父母近视	不近视	一方近视	均近视	
暗光用眼	从不	偶尔	经常	
学习时用眼距离	<20cm	≥ 20且< 25	≥ 25且< 30	≥ 30
看电视用眼距离	<20cm	≥ 20且< 25	≥ 25且< 30	≥ 30
户外运动时间	<1h	≥ 1且< 2	≥ 2且< 3	≥ 3
使用电子产品时间	<2h	≥ 2且< 4	≥ 4且< 6	≥ 6
学习时间	<2h	≥ 2且< 4	≥ 4且< 6	≥ 6
睡眠时间	<6h	≥ 6且< 7	≥ 7且< 8	≥8
连续用眼时间	<1h	≥ 1且< 2	≥ 2且< 3	≥ 3

这类数据的收集主要通过互联网问卷平台（例如：问卷星）完成。此外，我们所用的广州 2024 名青少年的眼健康数据来自于过去一项对眼科的研究论文，由八所广州眼科医院所收集。综上，我们可能的获取儿童青少年眼健康情况的途径有：互联网问卷、线下调查和医院病历。

5.2 问题二：具有可执行性的视力演化机理模型

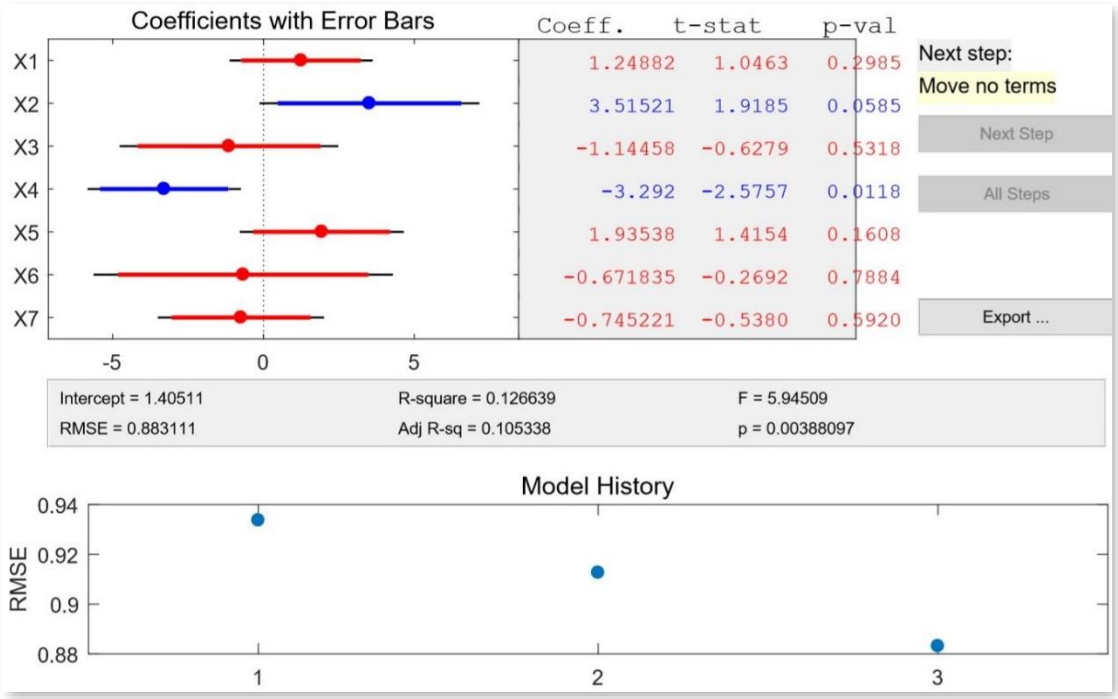
5.2.1 模型建立

统计回归模型对有效数据十分依赖。在第一问的调查数据和量化模型基础下，我们剔除了一些坏值，得有效的眼健康数据集。

我们先确定一个包含若干自变量的初始集合，在给定的显著性水平标准下，再对其引入、检验、移出变量，依此进行，直到不能引入、移出为止，从而得到反映眼睛视力演化机理的回归方程。

MATLAB 统计工具箱中的逐步回归命令是 stepwise，它提供人机交互式窗口，使用者可以在窗口内自由地引入和移除变量，进行统计分析。将附件二[错误! 未找到引用源。]的数据用 stepwise（x,y）命令（变量都没有进入初始模型）得到 Stepwise Regression 的初始界面窗口。当不能再引入、移出变量时，我们得到逐步回归的最终结果：

图 2 用 Stepwise Regression 逐步回归最终结果界面



在这个最终结果界面中，界面左上方是所有 7 个变量的回归系数的估计及误差界，其中彩色水平线表示置信度为 90%的置信区间，灰色为 95%。红色的水平线表示该变量未被选入到模型中，蓝色表示被选入到模型中。界面上方中间的表格显示每个变量的回归系数的估计值、检验的 t 统计量值以及 p 值，一般来说，每一步引入的变量因具有最小的 p 值或最大的 t 值。界面中间部分的表格给出了回归模型的所有计算结果，包括 *Intercept*（截距，即回归常数），决定系数 R^2，检验的 F 值（显著性检验），*RMSE*（剩余标准差）。界面最下面的部分是逐步回归中每一步所对应的模型的剩余标准差的点图。

5.2.2 求解与分析

在 MATLAB 帮助下，我们引入、移出变量直到不能引入和移除为止。在已有的工作基础下，我们得到了逐步回归的结果，其具体结果如下：

表 3 用 MATLAB 逐步回归具体解值

X_1	X_2	X_3	X_4	X_5	X_6	X_7
1.24882	3.51521	-1.14458	-3.292	1.93538	-0.67183	-0.74522

求解得出具有可执行性的视力演化机理模型，即将公式 (1)具体表示为：

$$Q = 3.51521 \times l - 3.292 \times d + 1.40511 \quad (2)$$

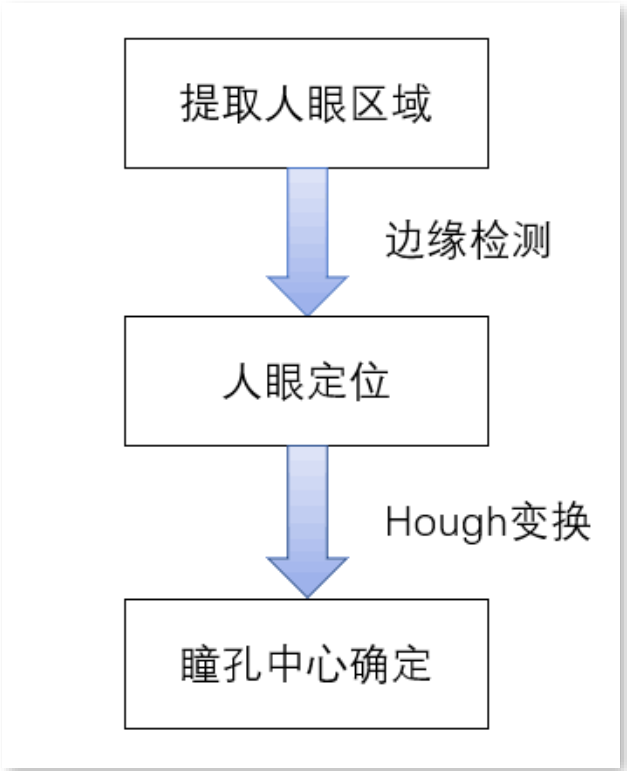
在最终的模型里回归变量只有 l 和 d 两个，二者对Q的影响程度比较接近，l 略重要一点，但二者回归系数相反，l 增加会加重近视，d 减少会加重近视。

5.3 问题三：基于 AI 的视力预警机制模型；

5.3.1 基于 AI 的用眼数据采集模型及求解

我们需要对学生用眼的实时数据进行采集，AI 图像识别技术能帮助我们识别人脸、瞳孔等，获取用眼环境，用眼距离等数据。该技术经过了长时间的发展，已有模板匹配、子空间方法、基于机器学习的方法，（如神经网络、支持向量机等）。实际中，大多数使用 Adaboost (Adaptive Boosting)的学习算法和霍夫变换算法以实现瞳孔定位。我们采用后两种方法，操作流程如下：

图 3 Adaboost 人脸识别流程图



5.3.2 预测近视的概率模型及求解

由于第二问中的视力演化机理模型仅仅能反映眼睛近视程度并且不能用来求解屈光度未来的趋势，因此在这里，我们可以建立近视概率模型用以预测面向学生个体的“视力-时间”变化。

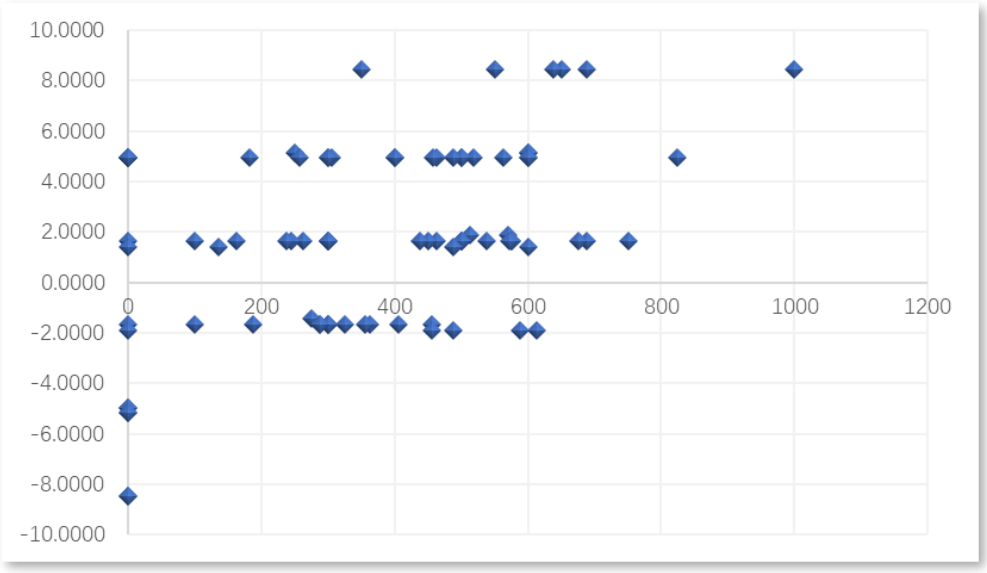
我们将视力屈光度Q作为衡量参数，代表不同的影响因素条件。在问题一量化模型的基础上，我们可以得到所有Q可能的值：

表 4 对应用眼距离和光线下Q的取值

Q的取值		用眼距离			
		0	1	2	3
用眼光线	0	1.40511	4.92032	8.43553	11.95074
	1	-1.88689	1.62832	5.14353	8.65874
	2	-5.17889	-1.66368	1.85153	5.36674
	3	-8.47089	-4.95568	-1.44047	2.07474

接着，我们计算了每个个体对应的Q值，并与每个个体的近视度数作比较。可视化数据如下：

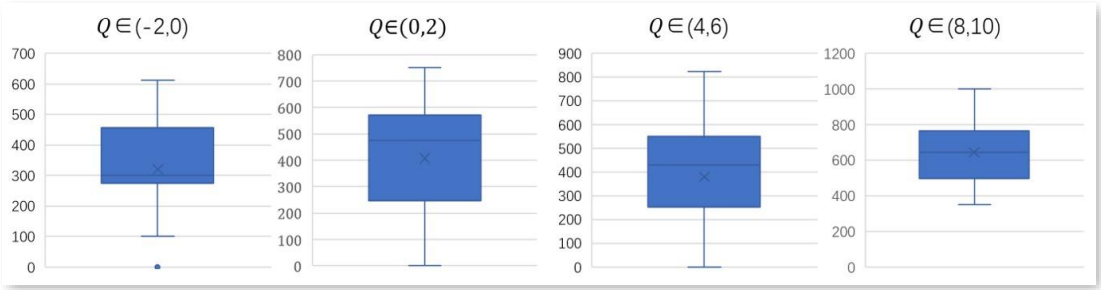
图 4 个体对应的近视屈光度Q值



观察可知，Q值与近视屈光度数具有相关性。为了进一步实现眼睛近视的预测，需要确立这一相关性，即研究Q为某一值的时候，近视度数为某一值的概率。为此我们要建立预测近视的概率模型。

因为Q的取值是有限个，不同的近视度数在某一Q值附近聚集。利用箱线图辅助分析：

图 5 Q在不同值附近时近视度数箱线图



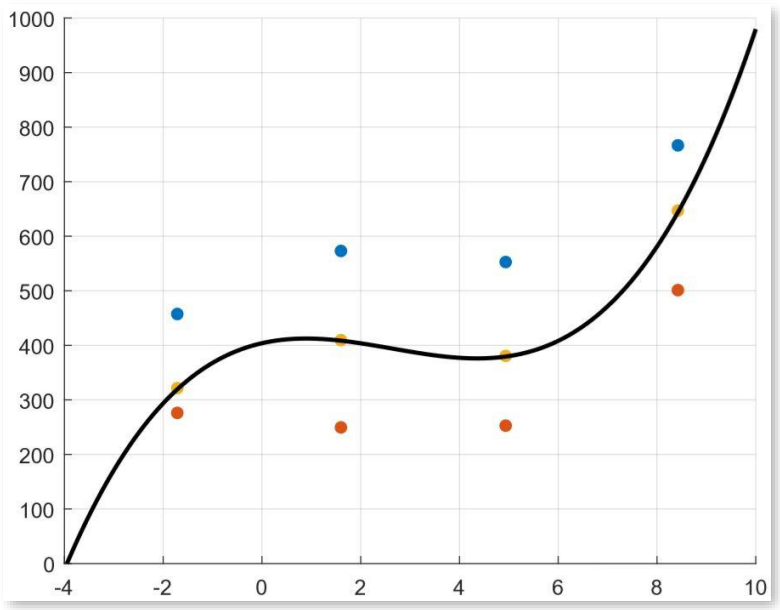
选取每张箱线图的中位数，上四分位数和下四分位数对应不同Q值，Q值则对应近视度数取平均值，可以得到下表：

表 5 箱线图中位数、Q值和上、下四分位数对应取值

中位数	上四分位数	下四分位数	Q
645.83	765.425	500	8.4355
379.479	551.56	251.5625	4.9482
408.3482	571.875	248.4375	1.6124
320.3927	456.25	275	-1.7017

基于最小二乘法的多项式拟合能够帮助我们对表格内数据进行处理。用 MATLAB 进行最小二乘法的多项式拟合：

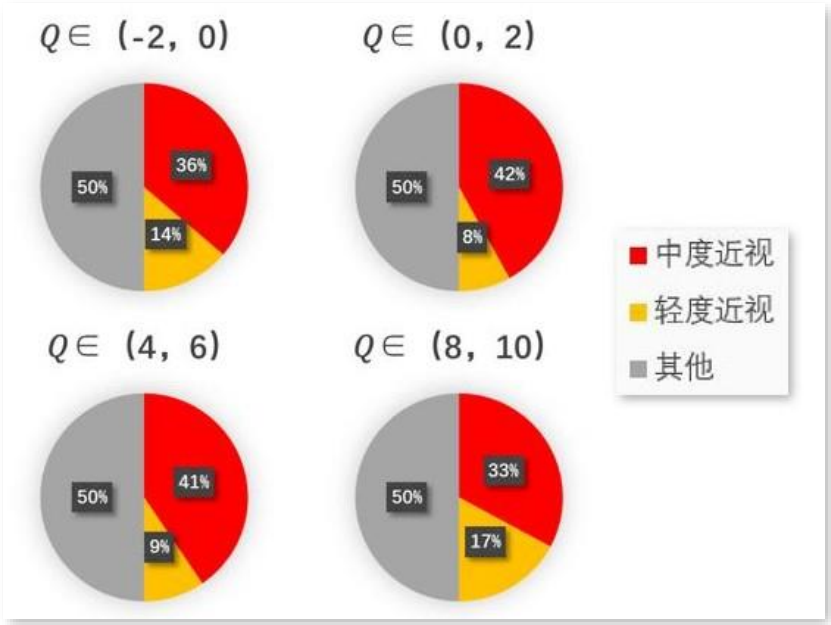
图 6 用 MATLAB 对Q的最小二乘法多项式拟合曲线



利用图中数据我们可以计算出不同Q值时近视度数的概率，假设上下四分位和中位数之间的数据分布是均匀的，并认为近视度数小于 75 度为不近视，75

到 300 度为轻度近视，300 到 600 度为中度近视，600 度以上为重度近视，那么可以得到：

图 7 不同值对应的近视程度概率



➢ Q在（-2，0）内时，学生有 36.23%的可能患中度近视，有 13.77%的可能患轻度近视；

➢ Q在（0，2）内时，学生有 41.94%的可能患中度近视，有 8.06%的可能患轻度近视；

➢ Q在（4，6）内时，学生有 40.53%的可能患中度近视，有 9.47%的可能患轻度近视；

➢ Q在（8，10）内时，学生有 32.86%的可能患中度近视，有 17.14%的可能患轻度近视；

值得注意的是，部份预测精度较低，可能患任意一种程度的近视，也可能不近视，我们这里将其统一归入“其他”，共占 50%。

5.3.3 预警与干涉

在上述的预测结果下，AI 可以辅助我们对学生未来近视的概率进行一个判断。当概率超过某一阈值，我们可以采取互联网等多种手段，通过通知家长老师、远程管控电子设备等手段，实现对学生视力的预测预警。例如，远程强制管控电脑使用时长等。

5.4 问题四：进行眼睛视力的机器学习模型

5.4.1 模型准备

5.4.1.1 归一化处理

为了消除指标之间的量纲影响，需要进行数据标准化处理，以使数据指标之间可比，使得预处理的数据被限定在一定的范围内，从而消除奇异样本数据导致的不良影响。

MATLAB 中 Mapminmax 函数可以按行地对数据进行标准化处理，将每一行数据分别标准化到区间$[y_{min}, y_{max}]$内，其计算公式是：

$$\mathrm{y} = (y_{max} - y_{min})\frac{x - x_{min}}{x_{max} - x_{min}} + y_{min}$$

5.4.1.2 函数引入

SVM 支持向量机涉及的关键函数

- Meshgrid 函数

Meshgrid 函数为 3-D 图生成 X 和 Y 阵列，其关系式表示为：

$$[X，Y] = meshgrid（x，y）$$

- Svmtrain 函数

Svmtrain 函数表示训练支持向量机，其关系式表示为：

$$model = svmtrain（train\_label，train\_matrix，'libsvm\_options'）$$

- Svmpredict 函数

Svmpredict 函数帮助我们使用支持向量机预测数据，表达式为：

$$[predict_{label}，accuracy] =$$

$$svmpredict（test\_label，test\_matrix，model）$$

5.4.2 模型建立

支持向量机的机器学习流程如下：

- 导入数据得到样本 N，进行归一化处理，
- 随机产生训练集和测试集，即随机选择 n 个样本数据用作训练（训练集），其余的（N-n）个用作测试（测试集），
- 对训练集进行训练（使用 svmtrain 函数），训练后再对测试集进行仿真预测（使用 svmtrain 函数）将仿真结果输出。

为了提高机器学习的效率和准确度，我们需要寻找 svmtrain 函数、svmtrain 函数中的最佳参数。在这里，我们运用网格法在 meshgrid 函数生成的 3-D 列阵图来寻找最佳参数。调用时，它的参数是以一个参数对（参数名，参数值）的形式出现的。我们经过查找库的参数说明，得到如下常用参数信息：

表 6 支持向量机相关函数说明

t	核函数类型，核函数设置类型 RBF 函数 $\exp(-r\|u-v\|^2)$
v	n-fold 交互检验模式，n 为 fold 的个数，必须大于等于 2
g	核函数中的 gamma 函数设置
c	设置 C-SVC，e -SVR 和 v-SVR 的参数

目标经过网格搜索找到最佳的一个 *c* 参数和 *g* 参数，此时的 *t* 参数为 2，再遍历 *c*，*g* 的取值，当某一次迭代过程当中的性能 *c* 和 *g* 的值优于其目前的最优值时，将最优值替换，同时返回 *c*，*g* 的值。

如果在迭代过程中寻找误差且满足一定的精度，我们考虑到 *c*（参数）的重要意义（*c* 是损失函数乘法因子，它的大小代表了我们对一些异常样本的重视程度），以 *c*（参数）作为第一优先选择而不是以 *g*（参数）作为优先选择。最终我们将最好的 *c*，*g* 拼接起来作为我们 *svmtrain* 中 *option* 的一个参数项。

同时，我们在做网格搜索交叉验证的时候做了 2 层的嵌套循环，c，g 都是从-10 到 10，每个间隔都是 0.2.因此我们在做模拟验证训练输出的时候需要一定的时间。

接下来就是 *Svmpredict* 进行模型的预测，我们对其进行简单的处理后，将其结果打印出来并画图，使其更加直观。

图 8 样本 70 测试集 SVM 预测结果对比（RBF 核函数）

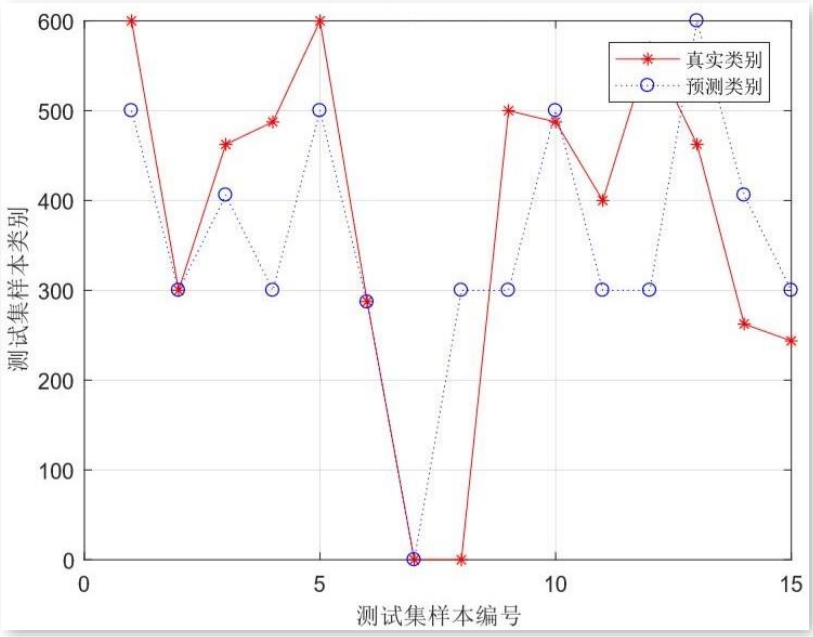
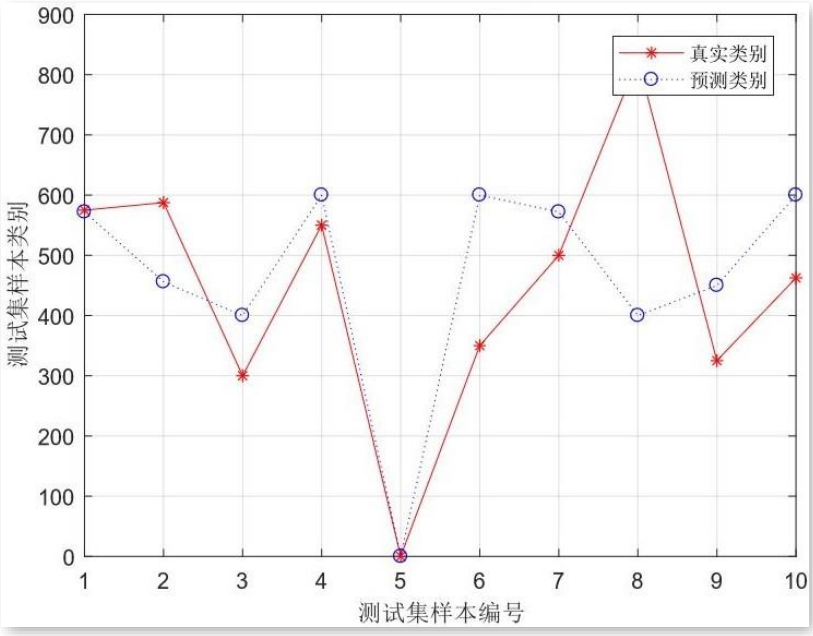


图 9 样本 70 测试集 SVM 预测结果对比（RBF 核函数）



与预测结果对比图同时输出的还有一个参数对比表。

表 7 支持向量机参数结果

样本	Rho	Obj	nsv	Total nsv	Accuracy	Accuracy	Nu
70	0.233	-1	8	70	51.429%	53.33%	0.423
75	1.067	-1.699	5	75	52.857%	50.00%	0.759

obj 是对偶 SVM 问题的最佳目标值，rho 是决策函数的偏置项，nSV 和 nBSV 是支持向量和有界支持向量的个数

nu：选择的核函数类型的参数

obj：svm 文件转换成的二次规划求解得到的最小值

rho：决策函数的偏置项 b（决策函数 f(x)=w^T*x+b）

nSV：标准支持向量的个数（0<αi<c）

Total nSV：支持向量的总个数（二分类的话就等于 nSV，多分类的话就是多个分界面上 nSV 的和）

5.5 问题五：

问题二和问题三的模型是用数理统计的方法建立的，问题四中机器学习也需要大量的数据学习预测，因此，数据的来源与普适性对我们的模型有很大的影响。因此，在第五问中，我们需要精心设计一份调查问卷，并尽可能的是其合理可靠。

设计一份优秀的调查问卷需要考虑很多因素，尤其是在模型假设对数据依赖程度很高时。因此，我们需要对问卷的方方面面进行约束，查阅资料，问卷设计有四个基本原则——目的明确性原则，题项适当性原则，语句理解一致性原则，调查对象合适性原则。

结合四大基本原则，设计问卷时需要考虑以下问题：（问卷见附录 1[1]）

（1）目的

探究学生影响视力的关键因素对视力的影响程度，这要求问卷涉及的问题是视力和影响因素两个方面，同时还决定了调查对象是学生。

（2）调查对象

由调查目的可知调查对象是学生，但结合题中“近年来，我国儿童青少年的近视问题日益严重且低龄趋势明显”，我们应将调查对象重点放在小学和中学，并选取小学、初中、高中、大学及以上比例为 3：3：3：1，男女比例 1：1，为了保证数据来源的可靠性，问卷不应该在网上发布，而是联系教育部在各学校内抽样发布，保证调查对象的普适性。

（3）题项设计

题目的设计依赖于我们所考虑的因素，同时要保证题目对不同年龄对象的理解相同。选项的设计还要注意选项分布的合理性，不能只靠主观臆测，要有一定的理论依据。

（4）可信度检验

信度检验主要评价问卷的稳定性和一致性，常用的信度分析方法包括：同质信度、分半信度和重测信度。实际操作中，我们可以通过同一问题重复出现、类似问题变换选项、同一问卷再次测试等方法进行可信度检验。

（5）预测试

为了进一步提高调查问卷数据的可靠性，我们还可以在问卷发布前进行预测试，在小范围内发布问卷，得到问卷的初步反馈，对问卷中的问题进行改进，得到最终要发行的问卷。

此外，为了提高调查问卷的可适性，有可能的话，视力应由专业医护人员测量填入，小学生应在老师或家长的指导下完成问卷，涉及到家长的部分由家长完成。

六、模型检验与评价

6.1 模型优点

首先，本文从数学统计模型入手，在避免了复杂的近视机理动力学模型基础上，仍维持了较高的准确性。我们从近视成因上出发，列出了 7 种关键因素，但仍考虑到并不是每一种因素都对视力有很大影响。因此在建立眼睛视力演化机理模型时时，模型分析了每种因素的显著性，并通过逐步回归得到了眼睛视力在影响因素作用下的回归方程，对视力演化机理进行了高效准确的求解。

其次，我们考虑到了影响因素和眼睛视力之间的关系并不显著，不是一种一一对应的关系。在此基础上，我们利用箱线图分析，不同Q值对应的近视度数不再是一个值，而是一个区间，再利用 MATLAB 拟合得到不同Q值时近视发生的概率，从而提高预警的准确性。

6.2 模型缺点

首先，我们建立的统计回归模型对数据依赖性特别大。但现有数据库仅包含 82 名小学生和 201 名大学生以及广州 2024 名青少年的眼健康数据，并且数据大多数来自大学生，并不能代表全体学生。而且某些因素和问题的选项在问卷中设计得并不合理，并不具有很好的普适性。

其次，针对近视演化机理模型，我们利用 stepwise 逐步回归分析得到了眼睛近视的线性回归方程，但是根据其相关系数（见下表），眼睛近视度数与各影响因素之间的相关性并不强烈。

表 8 MATLAB 求得的相关系数

1.0000	0.0656	0.0598	0.0355	-0.0976	-0.1546	0.2013	0.1112
0.0656	1.0000	-0.1652	-0.1368	0.0626	-0.0299	0.0981	0.2366
0.0598	-0.1652	1.0000	0.4646	0.2228	-0.1199	-0.0174	-0.2149
0.0355	-0.1368	0.4646	1.0000	0.3763	-0.1445	-0.0093	-0.2957
-0.0976	0.0626	0.2228	0.3763	1.0000	0.1226	0.1740	0.0450
-0.1546	-0.0299	-0.1199	-0.1445	0.1226	1.0000	0.0370	0.0052
0.2013	0.0981	-0.0174	-0.0093	0.1740	0.0370	1.0000	-0.0334
0.1112	0.2366	-0.2149	-0.2957	0.0450	0.0052	-0.0334	1.0000

相关系数大于 0.85 时可认为线性关系较强，这里最大的相关系数在 0.3 附近。客观层面，近视是复杂的演化过程，与任何单个因素的相关关系都较弱；另一方面，不全有效的数据库和并不完美的模型也是主要原因。

再者，在 AI 的视力预警机制模型中，我们利用箱线图把离散的比较混乱的不一一对应的数据转换为值与区间的对应。在计算概率时，我们利用区间长度占比去计算的。事实上区间内分布是不均匀的，有密有疏。置信区间计算概率会收到更好的效果。

七、模型改进与推广

（1）改进：对于这种利用大量数据去回归分析各成分影响时，我们不能只用线性关系去拟合，要考虑可能存在的非线性关系，利用 SPSS 的非线性逐步回归分析以及 logistics 回归分析，可能会得到更精确的回归方程，更好的反映影响关系。

（2）推广：这种基于数理统计的逐步回归模型可以应用到各种多因素影响

下实际问题中，尤其是对那些影响机理十分模糊的情形，同时，我们设计的概率预警模型一定程度上解决了回归模型精确度不高无法有效预警的问题，在实际问题的预警上有较强的实用性。

八、参考文献

[1] 姜启源，谢金星，叶俊.数学模型（第四版）[M].北京：高等教育出版社，2011.

[2] 瞿佳.视光学理论和方法[M].北京：人民卫生出版社，2004.

[3] 龚焱宏.青少年近视预测模型与干预策略研究[D].北京：北京师范大学，2011.

[4] 蔡培.网络学习环境对中小学视力的影响研究[D].武汉：华中师范大学，2017.

[5] Lin H, Long E, Ding X, Diao H, Chen Z, Liu R, et al. Prediction of myopia development among Chinese school-aged children using refraction data from electronic medical records: A retrospective, multicentre machine learning study[J]. PLOS Med. 2018; 15(11): e1002674.

附录

[1] 附录一：儿童青少年近视调查表

儿童青少年近视调查表				
性别	男 □	女 □	年龄	
学历	小学	初中	高中	本科及以上
本人	□	□	□	□
父亲	□	□	□	□
母亲	□	□	□	□
是否近视	否	□	是否戴眼镜	□
	是	□		
	左眼	____度	右眼	____度
父亲是否近视	否	□		____度
母亲是否近视	否	□	是	____度
您经常在光线较暗的环境下用眼吗？				
总是	经常	偶尔	从不	
□	□	□	□	
您经常吃烧烤炸鸡等辛辣油腻的食物吗？				
总是	经常	偶尔	从不	
□	□	□	□	
您经常感到心理压力很大吗？				
总是	经常	偶尔	从不	
□	□	□	□	
您经常做眼保健操吗？				
总是	经常	偶尔	从不	
□	□	□	□	
您会主动地从工作娱乐中抽身以放松眼眼睛吗？				
总是	经常	偶尔	从不	
□	□	□	□	
您看书离书的距离为？				
<20cm	20cm 至 25cm	25cm 至 30cm	30cm 至 40cm	≥40cm
□	□	□	□	□
您用电脑时前胸离书桌的距离为？				
<30cm	30cm 至 40cm	40cm 至 60cm	60cm 至 80cm	≥80cm
□	□	□	□	□
您使用电子设备时眼睛到屏幕的距离为？				
<20cm	20cm 至 25cm	25cm 至 30cm	30cm 至 40cm	≥40cm
□	□	□	□	□

您的平均每日学习时长是？				
<1 小时	1 至 2 小时	2 至 3 小时	3 至 4 小时	≥4 小时
□	□	□	□	□
您的平均每日使用电子设备时长是？				
<1 小时	1 至 2 小时	2 至 3 小时	3 至 4 小时	≥4 小时
□	□	□	□	□
您的平均每日户外运动时长是？				
<1 小时	1 至 2 小时	2 至 3 小时	3 至 4 小时	≥4 小时
□	□	□	□	□
您每天一次性连续用眼时长最大是？				
<10 分钟	10 至 30 分钟	30 至 60 分钟	1 至 2 小时	≥2 小时
□	□	□	□	□
您的平均每日睡眠时间是？				
<5 小时	5 至 7 小时	7 至 9 小时	9 至 11 小时	≥11 小时
□	□	□	□	□

[2] 附件二：代码

main.m

```
%% I. 清空环境变量
clear all
clc

%% II. 导入数据
load data.mat

%%
% 1. 随机产生训练集和测试集
n = randperm(size(matrix,1));

%%
% 2. 训练集——80 个样本
train_matrix = matrix(n(1:75),:);
train_label = label(n(1:75),:);

%%
% 3. 测试集——26 个样本
test_matrix = matrix(n(76:end),:);
test_label = label(n(76:end),:);

%% III. 数据归一化
[Train_matrix,PS] = mapminmax(train_matrix');
```

```
Train_matrix = Train_matrix';
Test_matrix = mapminmax('apply',test_matrix',PS);
Test_matrix = Test_matrix';

%% IV. SVM 创建/训练(RBF 核函数)
%%
% 1. 寻找最佳 c/g 参数——交叉验证方法
[c,g] = meshgrid(-10:0.2:10,-10:0.2:10);
[m,n] = size(c);
cg = zeros(m,n);
eps = 10^(-4);
v = 5;
bestc = 1;
bestg = 0.1;
bestacc = 0;
for i = 1:m
    for j = 1:n
cmd = [' ',num2str(v),' -t 2',' -c ',num2str(2^c(i,j)),' -g ',num2str(2^g(i,j))];
        cg(i,j) = svmtrain(train_label,Train_matrix,cmd);
        if cg(i,j) > bestacc
            bestacc = cg(i,j);
            bestc = 2^c(i,j);
            bestg = 2^g(i,j);
        end
        if abs( cg(i,j)-bestacc )<=eps && bestc > 2^c(i,j)
            bestacc = cg(i,j);
            bestc = 2^c(i,j);
            bestg = 2^g(i,j);
        end
    end
end
cmd = [' -t 2',' -c ',num2str(bestc),' -g ',num2str(bestg)];

%%
% 2. 创建/训练 SVM 模型
model = svmtrain(train_label,Train_matrix,cmd);

%% V. SVM 仿真测试
[predict_label_1,accuracy_1,decision_values1] = svmpredict(train_label,Train_matrix,model);
[predict_label_2,accuracy_2,decision_values2] =
```

```
svmpredict(test_label,Test_matrix,model);
result_1 = [train_label predict_label_1];
result_2 = [test_label predict_label_2];

%% VI. 绘图
figure
plot(1:length(test_label),test_label,'r-*')
hold on
plot(1:length(test_label),predict_label_2,'b:o')
grid on
legend('真实类别','预测类别')
xlabel('测试集样本编号')
ylabel('测试集样本类别')
string = {'测试集 SVM 预测结果对比(RBF 核函数)';
    ['accuracy = ' num2str(accuracy_2(1)) '%']};
title(string)
```

ployfit.m

```
clear all%清空环境变量
clc

M=[765.425;551.56;571.875;456.25]';%近视度数的上四分位
m=[500;251.5625;248.4375;275]';%近视度数的中位
d=[645.83;379.479;408.3482;320.3927]';%近视度数的下四分卫
Q=[8.4355;4.9482;1.6124;-1.7017]';%
x=-4:0.01:10;
P=polyfit(Q,d,3);
y=polyval(P,x);
```