# 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（从 A/B 中选择一项填写）： A

参赛报名队号为 A19B604

报名时所属学院（请填写完整的全名）： 计算机科学与技术学院

参赛队员姓名与学号 (打印，用二号字，并签名) ：

1. 陈凌灏 18130500143

2. 庞义人 17130130255

3. 林纯钢 18130500144

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 | | | | |

# 近视的预测及预警机制

## 摘要

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，已成为重大社会公共卫生问题。青少年近视比例较大已经成为社会上越来越严重的问题。“十三五”全国眼健康规划提出，要开展儿童青少年屈光不正的筛查和科学矫正，减少因未矫正屈光不正导致的视觉损伤。因此，建立一套近视的预测的预测及预警机制对我们整个社会来说尤为必要。

针对问题一，本文先确定影响视力的可能因素，并问卷的形式进行调查。然后，对各个变量进行KMO检验和Bartlett球形检验，发现Bartlett球形检验法的统计量的观测值为：498.87，相对应的概率值十分接近于0，小于指定的显著性水平$\alpha = 0.05$，应拒绝原假设。同时，KMO值满足KMO度量标准的要求，可以认为原有变量适合进行因子分析。对数据进行因子分析，原有变量被划分成为class1、class2、class3、class4、class5、class6六个影响因子，达到了降维的目的。根据SPSS软件得到的权重，本文对上述因子进行量化分析，计算出了视力下降的风险值。

针对问题二，本文以问题一的模型为基础，对相应影响因子进行显著性检验，保留显著性小于0.05的因子，最终确定使用class1、class2、class4三个因子作为近视演化机理模型的变量。同时，本文又将依据原有变量建立的模型与本模型进行比较，并结合和现代医学原理进行综合分析，验证了该模型的普适性和准确性。

针对问题三，本文利用因子分类的优点，采用了 *logistic* 回归分析、极大似然估计的方法，对所给影响因子指标进行分类，并计算出在所给影响因子条件下是否会作出人眼视力警告预警。在此基础上，本文通过建立线性判别模型来对警告值进行在再次预测，使模型更加准确。除此之外，本文通过建立实时监控机制，在用户收到警告的时候，准确判断出警告原因，并告诉孩子的父母、老师及时调整孩子的生活习惯。用户就能及时注意视力下降的风险，并对孩子的生活方式进行调整。

针对问题四，本文基于近视形成的机理模型、近视预警的机器学习模型对其所需数据进行分析，建立适合的训练集结构。并对问卷数据进行处理，进行相应模型训练。

针对问题五，本文基于近视形成的机理模型、近视预警的机器学习模型对其所需数据进行分析，建立适合的训练集结构。并根据相所需的的数据，探讨了可能的数据获取途径，例如，和学校进行合作、线下近视普查、开发近视预警小程序以及和眼科医院进行合作等方式。

综上所述，本文经过对用户用眼相关数据进行分析，建立了视力预警模型，直观地将孩子用眼数据反馈给父母，防止孩子发生近视。

**关键词：**KMO 检验、Bartlett 球形检验、独立性检验、*Pearson* 相关系数、*logistic* 回归分析、因子分析法、线性判别分析、机器学习

# 一、问题重述

## 1.1 背景资料与条件

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，已成为重大社会公共卫生问题。数据显示我国小学生近视比例为45.7%，初中生近视比例为74.4%，高中生近视比例为83.3%，大学生近视比例则高达87.7%。青少年近视比例较大已经成为社会上越来越严重的问题。“十三五”全国眼健康规划提出，要开展儿童青少年屈光不正的筛查和科学矫正，减少因未矫正屈光不正导致的视觉损伤。

近视问题同时也危及到国家安全。虽然中国人参军的热情丝毫没有减弱，但是由于这一门槛，很多人都将被拒之门外。因为视力不合格而被淘汰的占不合格人数已经接近半数。而且这已比例还有上升的趋势。从而加大了各地方政府的征兵难度。

因此，建立一套近视的预测及预警机制对我们整个社会来说尤为必要。

## 1.2 需要解决的问题

(1)确定影响视力的关键因素及其量化模型。

(2)建立一个眼睛视力的演化机理模型。

(3)建立一个基于真实数据，实时分析发出相应视力预警信息的模型。

(4)通过查阅资料和数据以及仿真等手段进行眼睛视力的机器学习模型。

(5)提出一套文中所述机理模型和学习模型的实现方案。

# 二、问题分析

## 2.1 问题一的分析

问题一要求基于一定的数据来源和调查，确定影响视力的关键因素并建立一个量化模型。

通过查阅相关文献，本文首先确定影响视力的可能因素，然后通过调查问卷的形式进行相关数据的收集并剔除坏数据。之后对数据进行预处理，将名义变量做归一化处理，尝试对变量进行 KMO 检验和 Bartlett 球形检验来确定是否可以使用因子分析法。如果可以，确定影响因子来对原有变量进行降维，对确定的影响因子建立量化模型，量化近视风险。如果不行则尝试对所有变量进行分析。

## 2.2 问题二的分析

问题二要求在问题一的基础之上，考虑模型的可执行性，建立眼睛视力的演化机理模型。基于问题一给出的因子，本文考虑将问题一给出的六种因子同近视人群的加深度数进行显著性分析，保留显著性小于 0.05 的因子。尝试将保留的因子进行线性回归拟合，确认相关因子对近视机理的具体演化模型。本文还将建立关于所有变量的模型，分析比较两个模型的差异，确定一个更优的模型作为最终的演化机理模型。为进一步确定模型的准确性，本文还将尝试借助现代医学的相关知识，对演化的机理进行分析。

## 2.3 问题三的分析

问题三要求我们在问题二的基础上，建立一个眼镜视力预警模型，并借助 AI 的手段完善预警机制，同时使得预警结果的呈现对用户比较友好。

基于问题二影响因子分类的优点，本文将采用 *logistic* 回归分析、极大似然估计的方法，对所给影响因子指标进行分类，并计算出在所给影响因子条件下是否会作出人眼视力警告预警，并将最后的分类结果反馈给用户。在此基础上，本文将通过建立线性判别模型来对警告值进行在再次预测，使模型更加准确。除此之外，本文将通过实时监控机制，在用户收到警告的时候，准确判断出警告原因，并及时调整生活习惯。用户就能及时注意视力下降的风险，并对生活方式进行调整。

### 2.4 问题四的分析

问题四要求我们建立可供学习的数据集，并进行机器学习模型的训练。我们将沿用影响因子分析的方法，将数据进行降维整合，按照问卷数据转化为及其可读取的数据，并进行代码实现。

### 2.5 问题五的分析

问题五要求我们通过采取切实可行的方法，已完成数据的实现。在问题五的求解过程中，本文将对形成近视的机理模型、机器学习的预警模型来对方法的实现进行分析。

对于确定最终的方法，本文将从和学校进行合作、开发相应近视风险预警微信小程序、线上线下进行免费近视普查、和眼科医院进行合作等切实可行的方式为模型实现确立最终的方法。

## 三、模型假设

（1）假设性别和饮食对视力不起影响；
（2）假设近视与非近视的人群分布具有随机性；
（3）假设问卷收集到的数据真实可靠；
（4）假设收集的人群没有进行视力矫正等手术。

## 四、符号说明

| 符号 | 说明 |
|---|---|
| $\alpha$ | 指定的显著性水平 |
| $class(i)$ | 降维后的第 $i$ 个影响因子 |
| $c_i$ | 降维后的第 $i$ 个影响因子的值 |
| $\omega$ | 降维后的第 $i$ 个影响因子的值的权重 |
| $D_i$ | 为近视人群视力加深的度数 |
| $T$ | 警告值 |
| $num$ | 数据集的数量 |

# 五、模型的建立与求解

## 5.1 问题一的模型

### 5.1.1 影响因素的确立

根据相关文献[1],[2],[3]的资料收集，本文先确定以下对视力存在影响的可能因素以及必要的个人信息，并据此制定调查问卷。影响因素如下：

个人信息：个人性别、所属省份、个人是否近视、父母近视情况；

生活环境：电子产品的屏幕种类、照明工具种类、作业负担、电子教具使用情况；

生活习惯：睡眠时间，电子产品使用时长，户外运动频率、阅读姿势；

护眼意识：确诊近视与配镜的时间间隔，确诊近视与配镜的时间间隔中度数的增加，视力复查的频率，眼镜的附加功能。

### 5.1.2 模型建立与求解

**1、近视影响因素的调查数据**

本次问卷共计发放 438 组，实际有效数据共计 385 组，数据有效率 88%。通过数据统计，获得的部分有效数据如下：

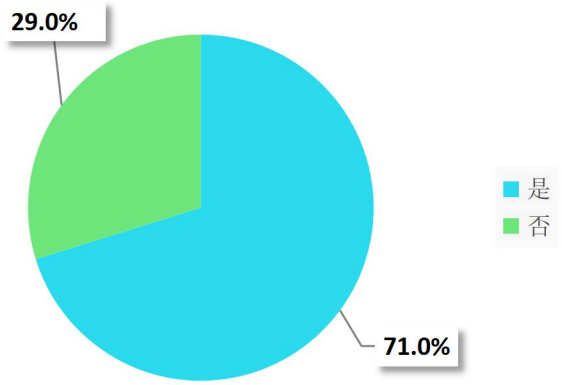


**图 5.1.2-1：调查人群的近视比例图**

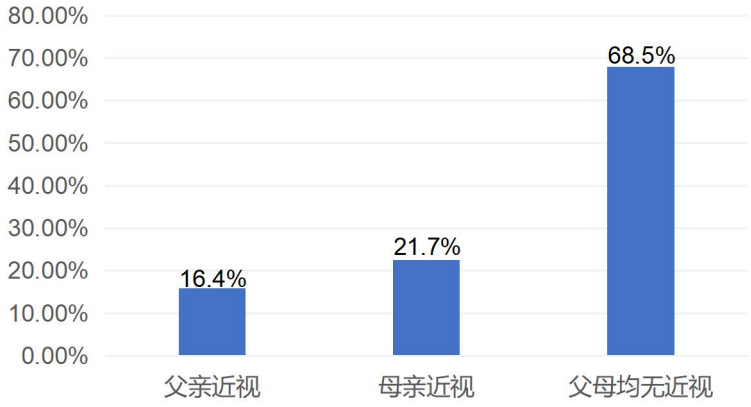


**图 5.1.2-2：调查人群亲属近视率**

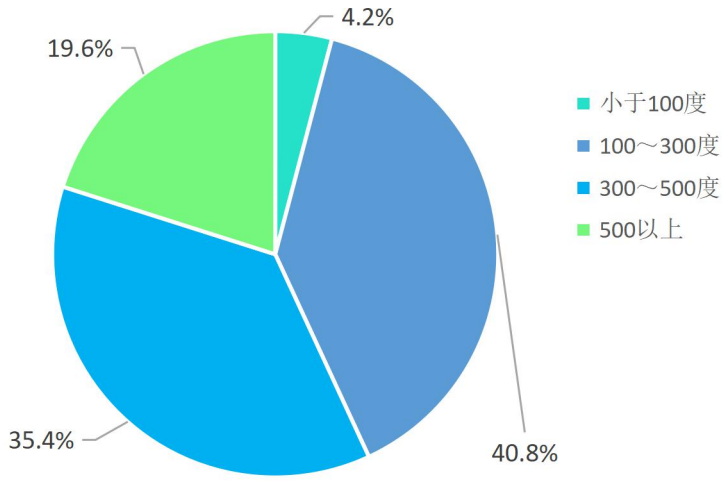


**图 5.1.2-3：所调查近视人群近视度数比例**

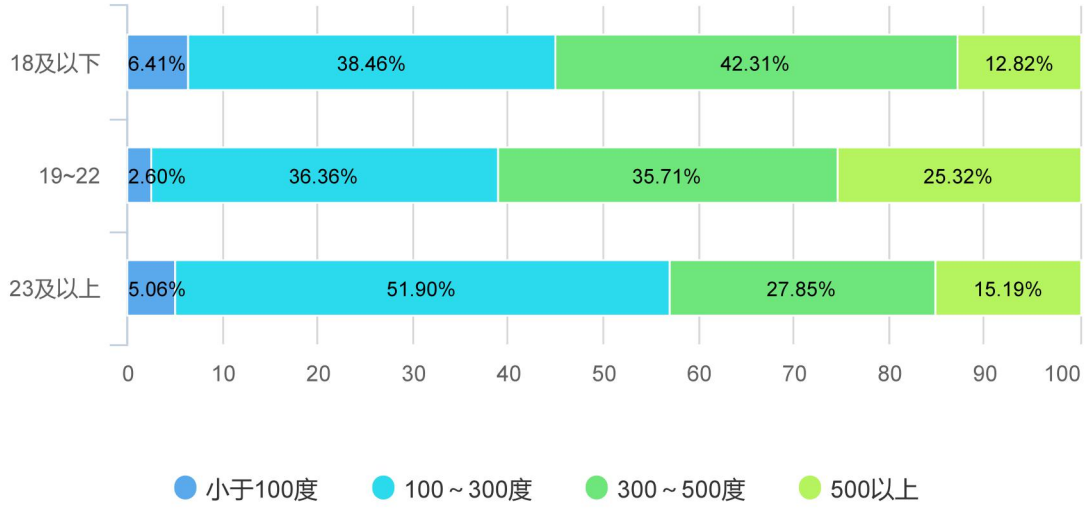


**图 5.1.2-4：所调查近视人群近视度数与年龄交叉性分析**

从代际近视率差异分析，调查人群中双亲均无近视的比例为 68.5%，，父母均近视的比例为仅为 6.6%，然而调查人群的无近视率仅为 29.0%。基于代际角度分析，可见人群的近视比例发生大幅度上升。

根据近视人群近视度数与年龄交叉性分析，年轻人群中高近视率的比例要高于相对年长者的比例，可见视力问题已经成为越来越严重的社会问题，且近视具有低龄化发展的态势。

## 2、因子分析

### (1) 影响因子确定与数值度量判别标准

本文将对近视的影响因素进行因子分析，对多组自变量使用 KMO 检验和 Bartlett 球形检验[4]确定影响视力的因子。

KMO 检验是使用于比较多个自变量间相关系数和偏相关系数的重要一个指标，并主要在多元统计的因子分析中广泛使用。在数值计算过程中，KMO 的取值在 0 和 1 之间。

当所有变量之间的简单相关系数平方和远大于偏相关系数平方和的时候，若 KMO 值接近于 1,则表明变量间的相关性较强，同时也表明原有变量适合作因子分析；反之，KMO 值越接近于 0,则表明变量间的相关性越弱，同时也表明原有变量不适合作因子分析。

Bartlett 球形检验法的统计量是由相关系数矩阵的行列式得到的。若该值较大,并且其对应的相伴概率值小于指定的显著性水平时(本文认为$\alpha = 0.05$),本文认为拒绝零假设,认为相关系数矩阵同单位阵有显著性差异；反之,零假设成立,原有变量之间不存在显著性。

**(2) 影响因子确定**

对于名义型变量，首先通过数据预处理，将其做归一化处理，转化为数值型变量(处理结果见附录 A)，借助 SPSS 软件，将有效数据进行 KMO 检验和 Bartlett 球形检验，经过数据处理得到以下数据：

**表 5.1.2-5：KMO 检验和 Bartlett 球形检验数据表**

| | |
|---|---|
| KMO 取样适切性量数 | 0.58 |
| 巴特利特球形度检验—近似卡方 | 498.87 |
| 自由度 | 55 |
| 显著性 | 0.00 |

由表 5.1.2-4 可知：Bartlett 球形检验法的统计量的观测值为：498.87，相对应的概率值十分接近于 0。由于概率值小于显著性水平$\alpha = 0.05$，则应该拒绝原假设，认为相关系数矩阵与单位矩阵存在显著性差异。在实验条件下，得到的 KMO 值为 0.58，根据 Kaiser 给出的 KMO 度量标准，原变量适合进行因子分析。

借助 SPSS 软件进行因子分析，得到的数据如下：

**表 5.1.2-6：成分矩阵**

| | 成分 | | | | | |
|---|---|---|---|---|---|---|
| 数据名称 | class1 | class2 | class3 | class4 | class5 | class6 |
| 使用的照明灯具种类 | 0.041 | -0.03 | -0.269 | -0.424 | 0.003 | 0.794 |
| FACEID/虹膜识别 | -0.347 | 0.809 | -0.173 | -0.088 | -0.055 | 0.05 |
| 屏幕种类 | -0.643 | 0.625 | -0.039 | -0.021 | -0.12 | 0.014 |
| 是否使用电子教具(iPad 等) | -0.087 | 0.332 | 0.231 | 0.479 | 0.498 | 0.024 |
| 睡眠时间 | -0.412 | -0.086 | 0.069 | 0.013 | -0.128 | -0.08 |
| 每日手机使用频率 | 0.671 | 0.289 | -0.106 | -0.13 | 0.105 | -0.197 |

| | | | | | | |
|---|---|---|---|---|---|---|
| 每周户外运动时长 | 0.044 | 0.022 | 0.465 | -0.014 | 0.607 | 0.351 |
| 是否走路看手机 | 0.597 | 0.312 | 0.407 | -0.038 | -0.215 | -0.067 |
| 是否在床上使用电子设备 | 0.475 | 0.268 | -0.52 | 0.097 | 0.191 | -0.022 |
| 是否黑暗中使用电子设备 | 0.627 | 0.289 | 0.129 | 0.052 | -0.268 | 0.187 |
| 是否阳光下阅读 | -0.082 | 0.129 | 0.671 | -0.407 | -0.176 | -0.027 |
| 父母近视 | 0.026 | -0.037 | 0.119 | 0.705 | -0.433 | 0.41 |

【注】：提取方法：主成分分析法。提取了 6 个成分：class1、class2、class3、class4、class5、class6。

**表 5.1.2-6：旋转后的成分矩阵**

| | 成分 | | | | | |
|---|---|---|---|---|---|---|
| 数据名称 | class1 | class | class3 | class4 | class5 | class6 |
| 使用的照明灯具种类 | 0.013 | 0.034 | -0.051 | -0.012 | 0.004 | 0.939 |
| FACEID/虹膜识别 | 0.059 | 0.898 | -0.049 | 0.023 | -0.061 | 0.055 |
| 屏幕种类 | -0.255 | 0.858 | 0.128 | 0.011 | 0.038 | -0.037 |
| 是否使用电子教具(iPad 等) | 0.005 | 0.212 | -0.151 | 0.695 | 0.136 | -0.282 |
| 睡眠时间 | -0.365 | 0.118 | 0.184 | -0.098 | 0.06 | -0.1 |
| 每日手机使用频率 | 0.688 | -0.05 | -0.229 | -0.016 | -0.277 | -0.076 |
| 每周户外运动时长 | 0.034 | -0.13 | 0.196 | 0.778 | -0.09 | 0.199 |
| 是否走路看手机 | 0.753 | -0.03 | 0.279 | 0.022 | 0.081 | -0.134 |
| 是否在床上使用电子设备 | 0.416 | 0.062 | -0.646 | -0.014 | -0.131 | 0.057 |
| 是否黑暗中使用电子设备 | 0.726 | -0.02 | 0.027 | -0.046 | 0.247 | 0.106 |
| 是否阳光下阅读 | 0.149 | 0.09 | 0.789 | 0.054 | -0.127 | -0.001 |
| 父母近视 | 0.034 | -0.02 | -0.043 | 0.02 | 0.93 | -0.011 |

【注】：提取方法：主成分分析法。旋转方法：凯撒正态化最大方差法。旋转在 6 次迭代后已收敛。

由实验数据可知：可以将所给数据划分成为 class1、class2、class3、class4、class5、class6 六个影响视力的因子。数据名称所属因子对应表如下：

**表 5.1.2-7：数据名称所属因子对应表**

| 因子 | 数据名称 |
|---|---|
| class1 | 睡眠时间、是否走路看手机、是否黑暗中使用电子设备 |
| class2 | FACEID/虹膜识别、屏幕种类、每日手机使用频率 |
| class3 | 是否在床上使用电子设备、是否阳光下阅读 |
| class4 | 父母近视 |
| class5 | 是否使用电子教具(iPad 等)、每周户外运动时长 |
| class6 | 使用的照明灯具种类 |

建立因子-数据成分得分系数矩阵：

**表 5.1.2-8：因子-数据成分得分系数矩阵**

| | 成分 | | | | | |
|---|---|---|---|---|---|---|
| 数据名称 | class1 | class2 | class3 | class4 | class5 | class6 |
| 使用的照明灯具种类 | 0.005 | 0.042 | -0.011 | 0.027 | 0.041 | **0.895** |
| FACEID/虹膜识别 | 0.098 | **0.567** | -0.048 | -0.008 | -0.044 | 0.067 |
| 屏幕种类 | -0.056 | **0.514** | 0.067 | -0.026 | 0.038 | -0.013 |
| 是否使用电子教具(iPad 等) | -0.009 | 0.104 | -0.177 | 0.624 | **0.105** | -0.238 |
| 睡眠时间 | **-0.16** | 0.044 | 0.121 | -0.1 | 0.049 | -0.088 |
| 每日手机使用频率 | 0.337 | **0.02** | -0.131 | -0.015 | -0.249 | -0.092 |
| 每周户外运动时长 | -0.004 | -0.111 | 0.116 | 0.714 | **-0.095** | 0.22 |
| 是否走路看手机 | **0.414** | 0.03 | 0.269 | -0.025 | 0.079 | -0.118 |
| 是否在床上使用电子设备 | 0.176 | 0.082 | **-0.48** | 0.019 | -0.107 | 0.03 |
| 是否黑暗中使用电子设备 | **0.387** | 0.052 | 0.078 | -0.066 | 0.245 | 0.108 |
| 是否阳光下阅读 | 0.138 | 0.053 | **0.632** | -0.007 | -0.123 | 0.021 |
| 父母近视 | 0.026 | -0.005 | -0.041 | **0.001** | 0.865 | 0.024 |

【注】：基于表 5.1.2-7、表 5.1.2-8，所加粗的得分系数作为该因子的有效得分系数，若未加粗则表示该得分系数视为 0。

对于因子 $class(i)$ ,依据因子-数据成分得分系数矩阵，给出各因子取值的计算公式：

$$class(i)=\sum_{i=1}^{n}k_i x_i \ ,$$

其中 $x_i$ 表示数据名称所对应的的归一化值(见附录 A), $k_i$ 表示有效的得分系数。

为建立一套有效的影响视力的量化模型，本文给出一份视力影响打分因素，用于计算各项因子对视力的量化影响指标。定义近视风险值 $Value$ ，各因子在风险值下所占的绝对权重为 $\omega_{0i}$ ,则风险值的计算表达式为：

$$Value=\sum_{i=1}^{6}\omega_i \cdot class(i)\ ,$$

基于此公式，利用 SPSS 数据分析得到对应的绝对权重 $\omega_i$ 构成的向量值：

$$(\Omega^{'})^T=[17.90\%,12.65\%,10.89\%,9.26\%,8.62\%,8.42\%]\ ,$$

在 $(\Omega^{'})^T$ 中，由于各个绝对权重值较小，为得到更加有效且直观的风险值，本文定义相对权重：

$$\Omega=100\Omega^{'}\ ,$$

得到：

$$(\Omega)^T=[17.90,12.65,10.89,9.26,8.62,8.42]\ ,$$

并以所求得到的相对权重最为风险值的权重计算风险值(即风险预测得分)。

### 5.1.3 结果分析

本方法将已知的数据内容进行剔除坏值与归一化处理，体现了严谨、数字化的思想。其次借助 KMO 检验和 Bartlett 球形检验，对较多的因素进行数据处理，达到数据降维的目的，以此在保证模型准确性的同时有效地降低了模型的复杂度，使其更加易于求解。

## 5.2 问题二的模型

### 5.2.1 模型准备

问题二的模型将以问题一的模型为基础，继续沿用问题一所得的影响因子，建立相关模型。

### 5.2.2 模型一的建立与求解

对于问题一所确定的相关因子，本文借助显著性检验，将显著性小于 0.05 的变量因子予以保留，确定了因子 class1、class2、class4 作为近视演化机理模型的自变量，通过显著性检验的数据得到的各因子显著性如下：

**表 5.2.2-1：保留因子的显著性表**

| 因子 | class1 | class2 | class4 |
|---|---|---|---|
| 显著性 | 0.018 | 0.002 | 0.038 |

基于上述显著性表，本文将通过 SPSS 软件对三个因子变量和近视人群的近视都市加深情况进行线性回归拟合，得到了如下的拟合演化机理模型：

$$L = 0.42 \times class1 + 1.03 \times class2 - 0.38 \times class4 + 0.79,$$

其中 $L$ 表示眼睛在三种影响因子的作用之下受到的损伤程度，并以此来刻画眼睛视力演化机理刻画模型。

### 5.2.3 模型二的建立与求解

为验证模型一的准确性，本文通过建立模型二与模型一进行比较。

首先，本文引入 *Pearson* 相关系数作为变量相关性的判断依据，记 $x_i$ 为影响视力自变量的值，$D_i$ 为近视人群视力加深的度数。*Pearson* 相关系数的计算模型如下：

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(D_i - \bar{D})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(D_i - \bar{D})^2}},$$

其中：

$$\bar{x} = \sum_{i=1}^{n} x_i \ ,$$

$$\bar{D} = \sum_{i=1}^{n} D_i \ ,$$

得到有显著相关性的变量及其 *Pearson* 相关系数如下表：

**表 5.2.2-2：有显著相关性的变量及其 *Pearson* 相关系数表**

| 因子 | 睡眠时间 | 是否走路看手机 | 是否阳光下阅读 |
|---|---|---|---|
| 显著性 | -0.873 | 0.817 | 0.935 |

将三个变量分别记作 $x_1, x_2, x_3$，并借助 SPSS 软件进行纤细你回归拟合，得到最终评估刻画眼睛视力演化机理刻画模型的函数：

$$D = -0.58x_1 + 0.91x_2 + 1.14x_3 + 2.486 ,$$

### 5.2.4 模型三的建立

根据相关文献[5][6]，我们从医学角度对近视的成因进行了简要的分析。我们得知，近视即屈光不正，主要分曲率性近视和轴性近视。前者主要是因为眼球的趋光性较强，而轴性近视的原因主要是眼球前后轴过长。另外近视也与家族性遗传存在一定联系。下图是近视成因的过程分析：

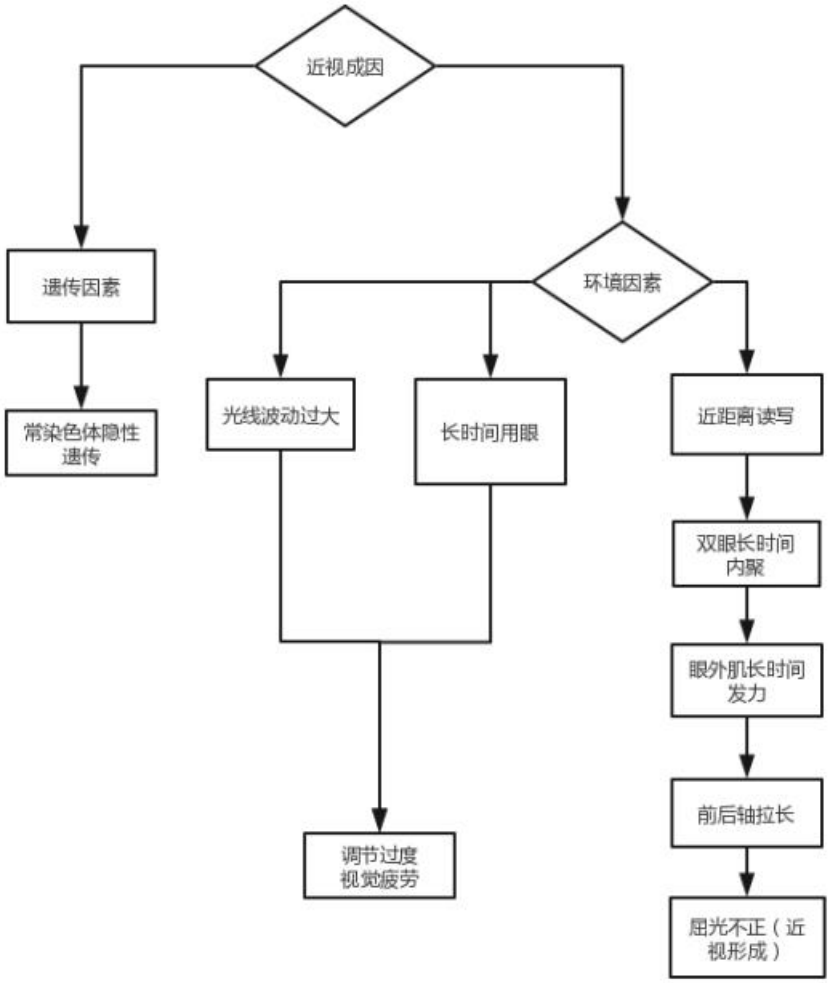


**图 5.2.2-3：近视成因分析图**

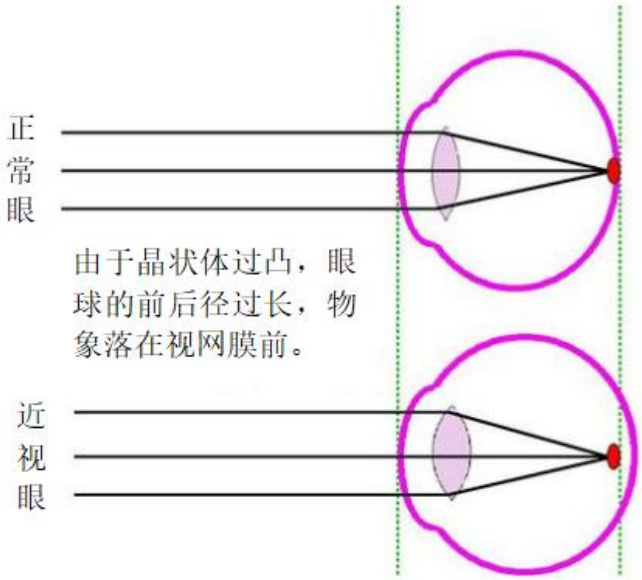


**图 5.2.2-4：近视对眼球形状影响的分析图**

### 5.2.4 模型的分析与准确性检验

根据模型三与图 5.2.2-3，本文认为近视与遗传存在一定的相关性，在 5.2.2 中模型一中确定的变量因子 class4(父母是否近视)与眼睛在三种影响因子的作用之下受到的损伤程度 $L$ 有高度的相关性。模型和医学分别从数理统计分析和眼睛的生理结构方面论证了近视与用眼习惯存在一定的相关性，在这一点上模型一有较好的准确性。除此之外，我们发现在受访人群中，父母的近视与孩子近视的关系不一定是正相关关系，因为近视的父母往往会更加关注下一代的视力问题，避免孩子和他们一样近视。由图 5.2.2-3、图 5.2.2-4，本文认为 class1、class2 中的变量：是否黑暗中使用电子设备、每日手机使用频率都是危害视力的重要因素，因此将其列为变量建立模型是非常有意义的，将会提高模型的准确率、客观性。在模型一与模型二的分析中，我们发现模型一的因子中已经对模型二的变量有一定的包含，而且在一定程度上由于模型一通过数据的整合降维，其优势更加明显。通过上述交叉分析，本文建议采用 5.2.2 中的模型一作为眼睛视力的演化机理模型。

## 5.3 问题三的模型

### 5.3.1 问题准备

问题二影响因子经过验证可以较为有效地对近视的风险进行预测，因此将 class1、class2、class4 三个重要影响因子作为自变量进行预测。

### 5.3.2 视力预警模型一

本文将引入 logistic 回归模型和似然极大函数对问题进行求解。变量 $c_1, c_2, c_3$ 分别表示已知 $class(1), class(2), class(3)$ 三个因子的值，$C = (c_1, c_2, c_3)^T$，$T \in (0,1)$ 表示警告值，警告值越接近 1 表示发出警告的可能性越大，若警告值越接近 0 则表示发出警告的可能性越低。

$$T = \frac{1}{1 + e^{-(w^T C + b)}},$$

所以：

$$\ln \frac{T}{1-T} = w^T C + b,$$

等式的右侧是一个线性函数。我们可以将 $T$ 视作样本 $C$ 作为正例的可能性，$1-T$ 视作样本 $C$ 作为反例的可能性，则有对数几率：

$$\ln \frac{T}{1-T},$$

所以：

$$\frac{\ln p(T=1 \mid C)}{\ln p(T=0 \mid C)} = w^T C + b,$$

为了简化讨论的复杂性，我们把 $w$ 和 $b$ 吸入向量形式 $\hat{w} = (w; b)$，同理 $\hat{C} = (C; 1)^T$，逻辑斯蒂回归模型即满足下列条件概率分布：

$$P(T=1 \mid C) = \frac{e^{(\hat{W}^T \hat{C})}}{1 + e^{(\hat{W}^T \hat{C})}},$$

$$P(T=0 \mid C) = \frac{1}{1 + e^{(\hat{W}^T \hat{C})}},$$

此时，若线性函数的值趋于正无穷大，概率 $P(T=1 \mid C)$ 就越趋近于 1；若线性函数的值越趋近负无穷，概率 $P(T=1 \mid C)$ 的值就越趋近 0。

获得完整的 logistic 回归模型就需要对参数进行进一步的确定，对于该模型给出有 *num* 组数据的训练集 $K$ 可以应用极大似然估计法对模型的参数进行估计，得到 $w$ 的估计值。设：

$$P(T=1 \mid C) = \tau(C), \qquad P(T=0 \mid C) = 1 - \tau(C),$$

则可以得到似然函数：

$$\prod_{i=1}^{num} [\tau(C_i)]^{T_i} [1 - \tau(C_i)]^{1-T_i},$$

对数似然函数为：

$$L(W)=\sum_{i=1}^{num}[T_i\ln\tau(\hat{C}_i)+(1-T_i)\ln(1-\tau(\hat{C}_i))]$$

$$=\sum_{i=1}^{num}[T_i\ln\frac{\tau(\hat{C}_i)}{1-\tau(\hat{C}_i)}+\ln(1-\tau(\hat{C}_i))]$$

$$=\sum_{i=1}^{num}[T_i(\hat{W}^T\cdot\hat{C}_i)-\ln(1+e^{\hat{W}^T\cdot\hat{C}_i})]\quad,$$

$L(\hat{W})$ 是关于 $\hat{W}$ 的高阶可导连续凸函数，根据凸优化理论[7]，logistic 回归学习最常用的方法是使用梯度下降法、牛顿法求得最优解，本文所需要求得的最优解为：

$$\hat{W}^*=\arg\min_{W}L(\hat{W})\ ,$$

利用牛顿法第 $p+1$ 次迭代解可以得到公式：

$$\hat{W}^{p+1}=\hat{W}^{p}-(\frac{\partial^2L(\hat{W})}{\partial\hat{W}\partial\hat{W}^T})^{-1}\frac{\partial L(\hat{W})}{\partial\hat{W}}\ ,$$

$$\frac{\partial L(\hat{W})}{\partial\hat{W}}=-\sum_{i=1}^{num}\hat{C}_i(T_i-\tau(C_i)),$$

$$\frac{\partial^2L(\hat{W})}{\partial\hat{W}\partial\hat{W}^T}=-\sum_{i=1}^{num}\hat{C}_i\hat{C}_i^T\tau(C_i)(1-\tau(C_i)),$$

将数据进行迭代即可求得最优解，对于最优解，只需将对应的近视影响因子代回模型，将会得到 $\tau(C)$ 的值，若 $\tau(C)$ 的值大于 0.5，则认为视力很可能正在受到损伤，应该对学生老师或家长发出预警；若 $\tau(C)$ 的值小于 0.5，则认为视力暂时处于安全状态，不发出预警。

### 5.3.3 视力预警模型二

对于视力的预警判别，本文将采用二分类线性判别分析(LDA)[8]的方式对研究对象的视力进行及时的预警。

线性判别分析是一种经典的线性学习方法，给定训练样例集，设法将样例投影到一条直线上，使得同类样例的投影点尽可能的近，异类样例的投影点尽可能原，在对新样本进行分类时，将其投影到同样的这条直线上，在根据投影点的位置来确定新样本的类别[9]。

定义变量 $c_1,c_2,c_3$ 分别表示已知 $class(1),class(2),class(4)$ 三个因子的值，$C=(c_1,c_2,c_3)^T$，$Y$ 表示是否发出近视警告。对于给定的样本集：$(c_{11},c_{12},c_{13}),(c_{21},c_{22},c_{23})\ .....(c_{num1},c_{num2},c_{num3})$,线性判别分析试图通过特征的

线性组合来特征化或区分它们，样本特征向量为$C=(c_1,c_2,c_3)^T$，那么线性判别分析的输出应该是：

$$Y=\omega^T C,$$

我们先求出投影前的均值向量好零件的均值向量：

$$(\frac{\sum_{i=1}^{num}x_{i1}}{num},\frac{\sum_{i=1}^{num}x_{i2}}{num},\frac{\sum_{i=1}^{num}x_{i3}}{num}),$$

投影后，各样本的均值：

$$\frac{\omega_1\sum_{i=1}^{num}c_{i1}+\omega_2\sum_{i=1}^{num}c_{i2}+\omega_3\sum_{i=1}^{num}c_{i3}}{num},$$

该式可以表示为：

$$\omega^T\overline{C},$$

所以类别“发出警告”$C'$和“不发出警告”$C''$的中心点之间的距离可以通过下列式子进行刻画：

$$Z=(\omega^T\overline{C}-\omega^T\overline{C}')^2$$

$$=\omega^T(\overline{C}-\overline{C}')(\overline{C}-\overline{C}')^T\omega,$$

对于发出警告的数据来说，投影后的方差：

$$\sum_{i=1}^{num}(\sum_{j=1}^{num}\omega_j C_{ij}-\omega^T\overline{C})^2$$

$$=\sum_{i=1}^{num}(\omega^T(C_i-\overline{C})(C_i-\overline{C})^T\omega)$$

$$=\omega^T(\sum_{i=1}^{num}(C_i-\overline{C})(C_i-\overline{C})^T)\omega,$$

同理，对于不发出警告的数据来说，投影后的方差：

$$\omega^T(\sum_{i=1}^{num}(C'_i-\overline{C}')(C'_i-\overline{C}')^T)\omega,$$

可以得到当二者之和：

$$\omega^T(\sum_{i=1}^{num}(C_i-\overline{C})(C_i-\overline{C})^T)\omega+\omega^T(\sum_{i=1}^{num}(C_i'-\overline{C}')(C_i'-\overline{C}')^T)\omega,$$

最小化时，二者的方差也达到最小化。

为便于分析，设：

$$S_O = (\overline{C} - \overline{C}')(\overline{C} - \overline{C}')^T ,$$

$$s_\omega = (\sum_{i=1}^{num}(C_i - \overline{C})(C_i - \overline{C})^T) + (\sum_{i=1}^{num}(C_i' - \overline{C}')(C_i' - \overline{C}')^T) ,$$

此线性判别分析的关键点为：投影后，不同类别的点尽可能远离，即 $Z$ 最大化；投影后，相同类别的点尽可能靠近，即 $S_\omega$ 最小化。本文的目标即求下列函数的最大值：

$$J = \frac{\omega^T S_O \omega}{\omega^T S_\omega \omega} ,$$

这里使用拉格朗日乘子法求解，我们将分母限制为长度为 1，则有：

$$R = \omega^T S_O \omega - \lambda(\omega^T S_O \omega - 1) ,$$

$$= S_O(\omega_1^2 + \omega_2^2 + \omega_3^2) - \lambda[S_\omega(\omega_1^2 + \omega_2^2 + \omega_3^2)] ,$$

对 $\omega$ 求偏导，令偏导等于 0，得：

$$S_O \omega = \lambda S_\omega \omega ,$$

即：

$$(\overline{C} - \overline{C}')(\overline{C} - \overline{C}')^T = \lambda S_\omega \omega ,$$

最终得到最优解：

$$\frac{\omega}{\| \omega \|} = S_\omega^{-1}(\overline{C} - \overline{C}') 。$$

对数据进行处理，得到实验结果如下：

$$\omega_o = \frac{\omega}{\| \omega \|} = (0.920, -0.231, -0.242) ,$$

组质心处对应的 $Y$ 值：

**表 5.3.3-1：组质心处对应的 $Y$ 值表**

| | |
|---|---|
| 近视 | 0.174 |
| 无近视 | -0.413 |

二者的均值为：-0.120

对于用户数据得到的影响因子 $C=(c_1,c_2,c_3)^T$，计算出：

$$Y=\omega C,$$

将计算出的 $Y$ 值与-0.120 进行比较，若 $Y$ 值较小，则不发出近视预警；若 $Y$ 值较大，则发出近视预警。

### 5.3.4 视力预警原因分析模型

基于上述模型，本文将对视力受损因素进行实时跟踪检测，以确定引起警报的原因，以便孩子的老师与家长及时地调整孩子的生活作息情况。

在下列流程机制下，本模型将建立在视力预警的情况下输出预警原因：

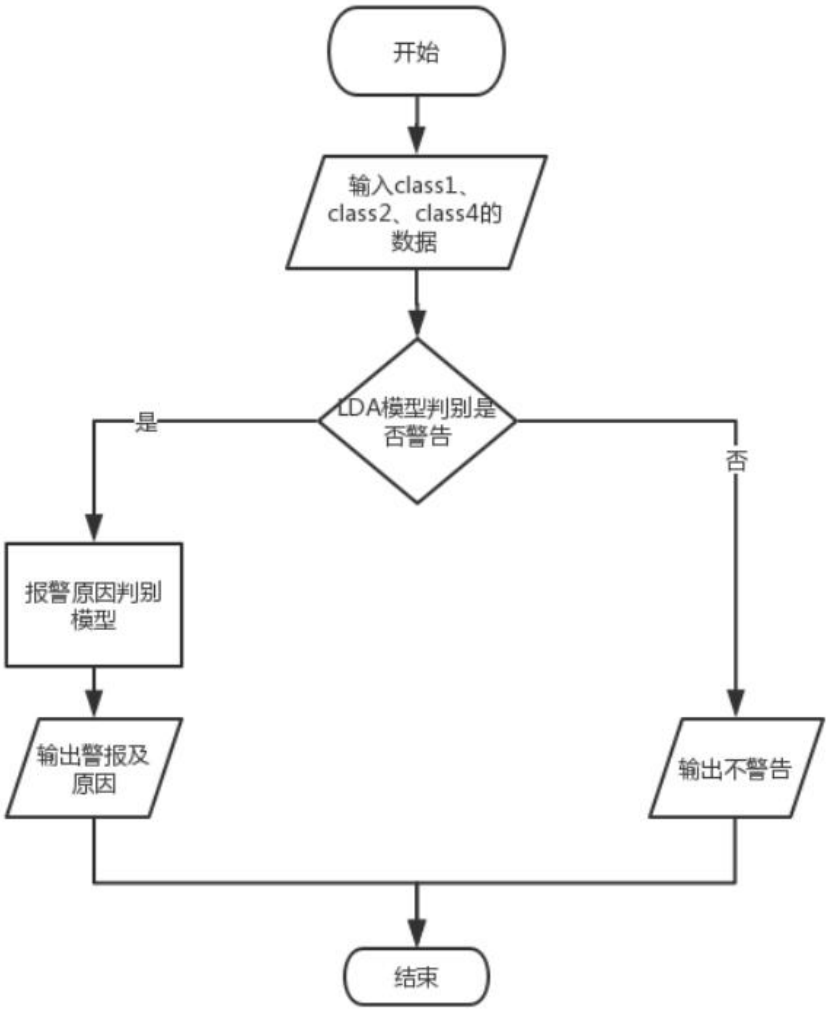


**图 5.3.3-2：预警原因输出流程图**

在问题二中，本文建立了视力演化风险的模型：

$$L = 0.42 \times class1 + 1.03 \times class2 - 0.38 \times class4 + 0.79,$$

在 $\Delta t$ 的时间内，本文将对视力进行实时监控，监控值为：

$$\Delta l_1 = 0.42 \times class1,$$

$$\Delta l_2 = 1.03 \times class2,$$

$$\Delta l_3 = -0.38 \times class4,$$

对 $\Delta l_1, \Delta l_2, \Delta l_3$ 三个变量因子进行比较，将三者的最大值作为报警原因进行输出，并对孩子的家长和父母进行直观化的反馈。

#### 5.3.4 模型的分析

本文建立了视力预警模型一、视力预警模型二。分别通过 logistic 回归分析和二分类线性判别分析(LDA)建立了视力预警模型。通过二者的比较，我们发现，“视力预警模型一”较为简洁，容易实现；而视力预警模型二的实验结果对用户更加友好，虽然用户只需要将相关因子输入模型即可从该警告系统中获得是否警告的反馈。在发出警告的时候，“视力预警原因分析模型”便能通过有效的分析对孩子的生活情况进行有效分析，以便家长和老师对孩子的生活作息进行调整。前两个模型相辅相成，在经过“视力预警原因分析模型”的原因分析，建立了模型之间的内在联系，最终实现了面向孩子家长及老师的用户反馈机制。

### 5.4 问题四的分析与求解

为了得到可供机器学习的数据集我们首先对问卷数据进行两步预处理(流程图见图 5.4-1)：

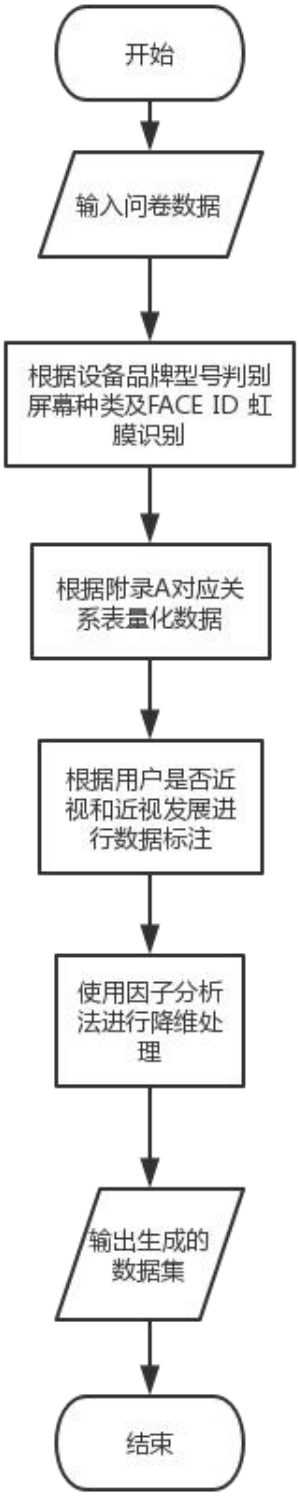


**图 5.4-1：数据处理流程图**

**数据初步处理：**

根据用户的设备品牌及型号得出用户使用的屏幕种类（OLED、LED）以及手机是否具有 FACE ID 和虹膜识别功能；

另外将问卷收集到的用户选项数据根据其对应性质量化成可以计算机可以处理的形式（具体对应关系见附录 A）；

同时将用户近视情况和近视发展情况进行处理，作为数据集的标注信息。

**数据的降维处理：**

因为影响视力的可能因素数量过多，不方便进行直接分析，所以我们首先通过问题 1 所述的因子分析法进行数据降维处理（具体细节见问题 3），最终得到包含 class1、class2、class4 三个维度的数据。

问题三中我们使用的多元线性判别模型（LDA），所需的训练集主要包含三个部分：数据标示编码、降上述维处理后的的 class1、class2、class4 的数据以及是否近视的标注。

问题三中我们使用的 *logistic* 回归模型理论上需要的训练集是上述降维处理后的的 class1、class2、class4 的数据以及近视加深程度的标注，然而因为收集到的包含近视加深程度信息的样本数量非常有限，无法完成多分类模型训练，所以在论文中我们将此模型简化为使用是否近视标注的二分类模型，以此来简要展示我们的思想。

此外我们将问卷编号 386 号及以后的数据进行相同预处理后作为两种模型的测试集。

经过 Python 对数据进行数据的处理，得到的结果如下：

$$0.767c_1 - 0.277c_2 - 0.195c_3 + 0.097 = 0\ ,$$

并以此作为分类的标准的临界状态。

## 5.5 问题五的分析与求解

下面分机理模型和机器学习模型两方面进行叙述。

### 5.5.1 关于近视形成机理模型

对于机理模型，我们的模型需要获取的关键数据主要有四类：受调成员的基本信息、生活习惯信息、用眼环境信息以及否近视和成员近视发展情况信息。

在我们论文完成期间，因受到数据获取条件限制，我们主要采用问卷调查的方式获取相关信息，收集到的样本数量相对较少，在今后模型的实际应用过程中，可以从以下几方面获取更多、更详细的数据：

首先可以考虑和眼科医院进行合作，获得患者就诊信息的脱敏数据，根据就诊数据采用上述分析方法，探究得到准确模型。

针对青少年近视眼的调查，首先考虑和学校进行合作，对学生目前视力状况进行普查，在合规的前提下也可以利用学生的中小学生年度体检信息结合问卷进行处理得到数据。

此外，在准确度验证方面，我们也可以通过对患者的配镜数据进行追踪分析的方式获得更为详细的数据。

### 5.5.2 关于机器学习预警模型

对于机器学习模型，我们需要的数据主要包括四个方面：患者的个人信息、用眼环境信息、用眼习惯信息以及患者的近视情况信息。在论文完成期间我们使用的训练数据集主要依赖我们对收集到数百份问卷采用问题四所述方法进行整理得到，在模型的实际应用过程中我们设计如下微信小程序进行数据收集（简要思路如图 5.5-1 所示）：

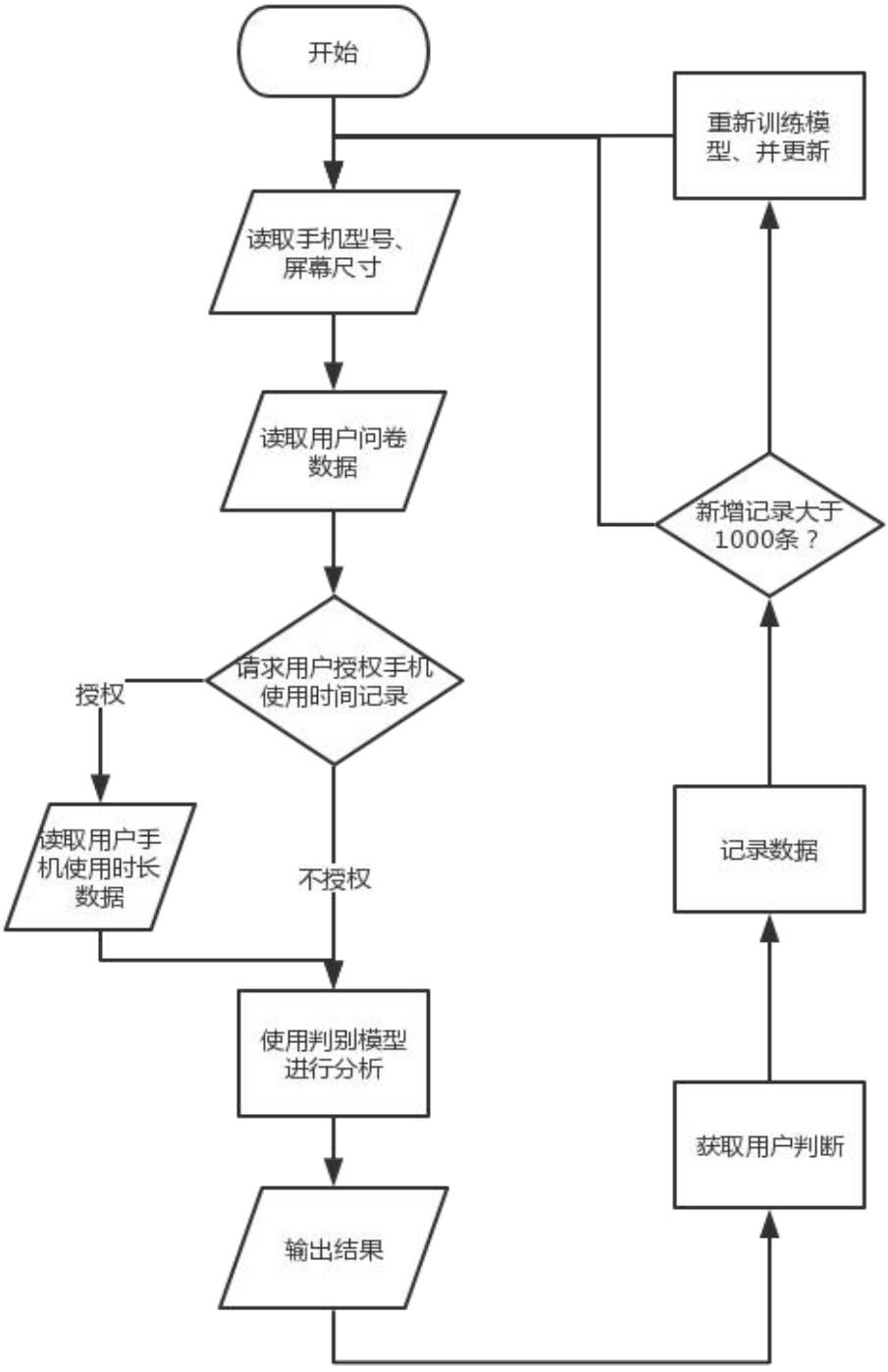

**图 5.5-1：数据收集流程图**

首先将我们问卷中选用的各项问题提出供用户选择，同时根据用户使用的设备型号，判断屏幕种类、屏幕尺寸等数据，请求用户授权读取手机使用时间信息（iPhone 和 android 均提供接口）；然后应用上面训练完成的模型作为初始模型进行是否近视和近三年近视增长的推断、给出健康提示，并询问用户我们的推断是否正确；之后将用户反馈的情况连同上述问卷作为新的样本数据存储起来，每当样本数量增加 1000 条的时候启动模型训练流程，更新我们的预测模型，并投放线上，如此循环逐步提升预测准确度。

我们可以考虑和学校进行合作，将这样的小程序作为学生近视增长风险的自我评估工具和良好用眼习惯的培养工具（健康提示）；

此外我们还可以和眼科医院进行合作，将我们的数据收集模型集成进医院的配镜眼光数据处理系统，帮助医院建立相应的近视发展数据库。

除此以外，我们还可以将小程序面向学生家长推广，抓住家长们爱子心切的内心，进行孩子相应生活习惯等数据的综合收集，并对孩子近视发展情况进行追踪调查，将调查范围覆盖到不同年龄段、不同地域进行大规模调查研究。

# 六、模型的评价

## 6.1 模型的优点

(1)该模型通过问卷形式以考虑到的所有可能的因素为影响视力的因素作为问题来收集数据，不同于收集已有数据，该方法比较合理并且具有与时俱进的性质，具有创新意义。

(2)鉴于收集到的诸多变量之间通常会存在或多或少的相关性，该模型采取“主成分分析法（PCA）”进行降维操作，消除了变量之间的相关影响，减少了指标选择的工作量，不同于传统的采取全部变量，在变量较多的情况下，在保留绝大部分信息的情况下用少数几个综合指标代替原指标进行分析。

(3)在第二问中，该模型在现有医学的经验上进行分析改进，具有强有力的说服力。

(4)在第三问中，选取了同一个人在不同时间段的数据进行实时监控，给孩子的父母和老师进行直观化的反馈。

(5)在视力的预警中重新进行降维，通过“用二分类线性判别分析(LDA) ”在原始变量的基础上进行降维，同时保持区分类别的信息。 不仅降低了分类任务的计算量，而且减小了参数估计的误差，从而避免过拟合。

## 6.2 模型的缺点

(1)由于真实数据难以搜集全面，数据缺乏的问题使我们不得不在数据缺少的情况下进行数据模拟，并采取最后获得的问卷的变量数据作为测试级。

(2)进行“主成分分析”后主成分的解释和含义一般带有模糊性，不像原始变量那么清楚、准确。

# 七、模型的改进与推广

## 7.1 模型的改进

本文因为数据收集限制并没有考虑饮食因素，在接下来的模型改进过程中可以考虑加入饮食因素，进一步探究。

## 7.2 模型的推广

本文所述模型对于分类问题具有普适性，可以推广到其他类似的分类问题，可以进行相应研究推广。

# 八、参考文献

[1]付少雄,林艳青.手机使用对用户健康的负面影响研究——以大学生为调查对象,图书情报知识,2019；
[2]李倩楠,吴楠,戚海琴.学龄前儿童视力下降的因素及相关预防措施,科技风,2017；
[3]王赟,谌丁艳,熊华威,张浩,周丽,我国青少年近视影响因素与防治措施研究,实用预防医学,2016；
[4]金丕焕,主编,医用统计方法,上海,上海医科大学出版社,2000；
[5]张保身,抢救视力,北京:电子工业出版社,2007；
[6]张寅,关注视力健康.西安:西安电子科技大学出版社,2013；
[7]Dimitri P. Bertsekas,凸优化理论,北京：清华大学出版社,2015；
[8]张玲,陈收,张昕,基于多元判别分析和神经网络技术的公司财务困境预警,系统工程,第23卷11期,49-56,2005；
[9] 周志华, 机器学习, 北京: 清华大学出版社, 2016。

# 附录

A、数据归一化转化

| 调查数据 | 名义变量 | 转义成数字变量 |
|---|---|---|
| 性别 | 男 | 1 |
| | 女 | 2 |
| 年龄 | 小于 18 岁 | 1 |
| | 19 ~ 22 岁 | 2 |
| | 23 岁以上 | 3 |
| 父母近视情况 | 无人近视 | 0 |
| | 一人近视 | 1 |
| | 两人近视 | 2 |
| 本人是否近视 | 无近视 | 0 |
| | 近视 | 1 |
| 近视度数 | 100 度以下 | 1 |
| | 100 ~ 300 度 | 2 |
| | 300 ~ 500 度 | 3 |
| | 500 度以上 | 4 |
| 灯具 | LED | 1 |
| | 日光灯管 | 2 |
| | 灯泡 | 3 |
| | 其它 | 4 |

| | | |
|---|---|---|
| FACEID/虹膜识别 | 无 | 0 |
| | 有 | 1 |
| 手机屏幕 | LCD | 0 |
| | OLED | 1 |
| 电子教具使用 | 不使用 | 0 |
| | 课上使用 | 1 |
| | 课下使用 | 2 |
| 睡眠时间 | 6 小时以下 | 1 |
| | 6 ~ 8 小时 | 2 |
| | 8 小时以上 | 3 |
| 手机使用频率 | 几乎不用 | 1 |
| | 第二档 | 2 |
| | 第三档 | 3 |
| | 第四档 | 4 |
| | 非常频繁 | 5 |
| 每周户外活动时间 | 2 小时以下 | 1 |
| | 2 ~ 6 小时 | 2 |
| | 6 小时以上 | 3 |
| 发现近视后多久配镜 | 立刻配镜 | 0 |
| | 2 个月以内 | 1 |
| | 2 ~ 6 个月 | 2 |

| | | |
|---|---|---|
| | 6 个月以上 | 3 |
| 等待期间度数增长 | 几乎无增长 | 1 |
| | 50 ~ 100 度 | 2 |
| | 100 ~ 200 度 | 3 |
| | 200 度以上 | 4 |
| 视力检查频率 | 半年多次 | 1 |
| | 半年一次 | 2 |
| | 一年一次 | 3 |
| | 少于一年一次 | 4 |
| 近半年度数增长 | 100 度以下 | 1 |
| | 100 ~ 200 度 | 2 |
| | 200 ~ 300 度 | 3 |
| | 300 ~ 400 度 | 4 |
| | 400 ~ 500 度 | 5 |
| | 500 度以上 | 6 |
| 各项不良用眼习惯 | 无 | 0 |
| | 有 | 1 |
| 各项眼镜附加功能 | 无 | 0 |
| | 有 | 1 |

B、*logistic* 回归分析代码

```
import pandas as pd                    # 用于读取数据文件 import tensorflow
as tfimport matplotlib.pyplot as plt    # 用于画图 import numpy as np
```

```
df = pd.read_csv("input.csv", header=None)train_data = df.values
print(train_data)
train_X = train_data[:, :-1]train_y = train_data[:, -1:]feature_num =
len(train_X[0])sample_num = len(train_X)print("Size of train_X:
{}x{}".format(sample_num, feature_num))print("Size of train_y:
{}x{}".format(len(train_y), len(train_y[0])))
X = tf.placeholder(tf.float32)y = tf.placeholder(tf.float32)
W = tf.Variable(tf.zeros([feature_num, 1]))b = tf.Variable([-.9])
db = tf.matmul(X, tf.reshape(W, [-1, 1])) + bhyp = tf.sigmoid(db)
cost0 = y * tf.log(hyp)cost1 = (1 - y) * tf.log(1 - hyp)cost = (cost0 +
cost1) / -sample_numloss = tf.reduce_sum(cost)
optimizer = tf.train.GradientDescentOptimizer(0.001)train =
optimizer.minimize(loss)
init = tf.global_variables_initializer()sess =
tf.Session()sess.run(init)
feed_dict = {X: train_X, y: train_y}
for step in range(1000000):
    sess.run(train, {X: train_X, y: train_y})
    if step % 10000 == 0:
        print(step, sess.run(W).flatten(), sess.run(b).flatten())
# 绘图
w = [0.7672361 , -0.276697 , -0.19542742]b = 0.09650069
from mpl_toolkits.mplot3d import Axes3D
x1 = train_data[:, 0]x2 = train_data[:, 1]x3 = train_data[:, 2]y =
train_data[:, -1:]
fig=plt.figure()ax=Axes3D(fig)

for x1p, x2p, x3p, yp in zip(x1, x2, x3, y):
    if yp == 0:
        ax.scatter(x1p, x2p, x3p, c='r')
    else:
        ax.scatter(x1p, x2p, x3p, c='g')
        ax.set_zlabel('Z')  # 坐标轴
ax.set_ylabel('Y')ax.set_xlabel('X')

a = 0.7672361b = -0.276697 c = -0.19542742d = 0.09650069
x1 = np.linspace(-1,1,10)y1 = np.linspace(-1,1,10)
 X,Y = np.meshgrid(x1,y1)Z = (d - a*X - b*Y) / c
 fig = plt.figure()ax = fig.gca(projection='3d')
 surf = ax.plot_surface(X, Y, Z)
```