影响城市居民身体健康的因素分析研究

摘要

慢性非传染性疾病（心脑血管病、糖尿病等）严重影响了我国居民健康，患病率随生活方式变化上升。身体健康与多因素相关，践行健康生活方式是全社会关注的重点。

对于问题一, 需要评判居民饮食习惯的合理性。即根据《中国居民膳食指南》中所提出的八条准则，对所得数据进行处理，本文将居民的饮食习惯分为饮食内容与饮食规律。其中饮食规律为是否按时就餐，饮食内容除去糕点后，大致分为全谷物、鱼禽蛋瘦肉、奶制品等共 11 大类，对所有指标数据**归一化处理**后，通过**熵权法**计算得出各项指标的权重，在通过**稳定性检**验后，依据最终得分的高低，得出居民饮食习惯的合理性为：附件 2 中居民饮食习惯不合理，主要问题包括：居民日常摄入的新鲜蔬菜和水果量不足，以及烹调用油量普遍超标。同时，居民每日和每周的食物种类数也未达到《中国居民膳食指南》的标准。

对于问题二, 要求分析居民生活习惯与饮食习惯是否受年龄、性别等特征影响。本文在问题一模型的基础上，分析居民生活习惯和饮食习惯与年龄、性别等因素的相关性。选取差异性较大的饮食习惯作为变量，生活习惯中的吸烟状况、体育锻炼、家务劳动三个变量，由于所选取变量中既有分类变量，又有连续变量，本文将分类变量转化为**哑变量**，并运用**线性回归模型**分析其与生活习惯、饮食习惯的相关性。通过 **Pearson 相关系数**和 t **检验**分析连续变量与生活习惯、饮食习惯的相关性。最终发现，居民的生活习惯和饮食习惯与婚姻状况具有较强的相关性，与文化程度等因素并不相关，此外，居民的饮食习惯与其年龄相关，居民的生活习惯与其年龄并不相关。

对于问题三, 要求深入分析高血压、糖尿病等常见慢性病与吸烟、饮酒等因素的关系以及相关程度。本文在问题一、二所建立模型的基础上，对饮食习惯、生活习惯进行评级，从而简化问题。而后通过计算得出各种常见慢性病各自与吸烟、饮酒等因素的线性回归模型，从而综合分析得到常见慢性病（如高血压、糖尿病等）与吸烟、饮酒等因素之间的关系以及相关程度为：常见慢性病如高血压和糖尿病与多种生活方式因素相关。糖尿病与饮食、生活习惯和运动量有强线性关系，与较高的相关程度，但与吸烟和饮酒关系较小。高血压则与饮食、吸烟、饮酒和工作性质具有较强的线性关系与较高的相关程度，而与生活习惯的相关性较低.

对于问题四, 需要实现居民的合理分类，并为各分类群体量身定制建议。具体而言，本文参照附件 2 的信息，将居民依据其婚姻状况及是否罹患慢性病，细化为四大类别：患病有无配偶、健康有无配偶。基于此前构建的模型框架，我们进一步融合问题二与问题三的研究成果，深入优化模型，按年龄段对居民进行分类为青中老年三类人群，从而对每一类别的人群，提供具有针对性的指导，以促进整体健康水平的提升。

关键字： $0-1$ 标准化　线性回归模型　Pearson 相关系数　t 检验

一、 问题重述

1.1 问题背景

慢性非传染性疾病，主要包括心脑血管疾病、糖尿病、恶性肿瘤以及慢性阻塞性肺病等，已成为我国居民健康的重要威胁。近年来，随着生活方式的变迁，慢性病的发病率持续上升，对公众健康构成了严峻挑战。众所周知，个体的健康状况深受多种因素影响，包括但不限于年龄、饮食习惯、身体活动水平及职业特性等。因此，如何通过科学规划膳食结构、积极参与适量身体运动，并全面践行健康生活方式，以达成促进身体健康的目标，已成为社会各界广泛关注的焦点问题。

1.2 问题要求

问题 1 参考附件 3，分析附件 2 中居民的饮食习惯的合理性，说明存在的主要问题。

问题 2 分析居民的生活习惯和饮食习惯是否与年龄、性别、婚姻状况、文化程度、职业等因素相关。

问题 3 根据附件 2 中的数据，深入分析常见慢性病（如高血压、糖尿病等）与吸烟、饮酒、饮食习惯、生活习惯、工作性质、运动等因素的关系以及相关程度。

问题 4 依据附件 2 中居民的具体情况，对居民进行合理分类，并针对各类人群提出有利于身体健康的膳食、运动等方面的合理建议。

二、 问题分析

2.1 问题一分析

对于问题一，要求参考附件 3 去分析附件 2 中居民的饮食习惯的合理性，并对其映射出的问题进行剖析。本文首先依照附件 3 中的准则，将附件 2 中居民的饮食习惯分为饮食规律与饮食内容两大类，其中饮食规律指是否按时吃就餐，饮食内容按照《中国居民膳食指南》中所提出的八条准则，大致分为全谷物、鱼禽蛋瘦肉、奶制品、豆制品、新鲜蔬菜、新鲜水果、油炸面食、含糖饮料、酒精、食用油与食用盐共 11 大类。而后依据附件 3 中的数据，为提取的饮食习惯相关变量设置合理性范围，即评分指标，建立评分卡模型，将指标数据归一化处理后，通过熵权法计算各项指标的权重，导出最终得分并排序，从而得到附件 2 中各个居民饮食习惯的合理性，并分析存在的主要问题。

2.2 问题二分析

问题二需要分析居民的生活习惯、饮食习惯，两者与年龄、性别、婚姻状况、文化程度、职业等因素是否相关。本文在问题一模型的基础上，对于饮食习惯，选取差异性较大的部分变量，对于生活习惯，通过吸烟状况、体育锻炼和家务劳动三个变量体现。由于所选取变量中既有分类变量，又有连续变量，本文将职业、婚姻状况等分类变量转化为哑变量，通过线性回归模型，分析生活习惯和饮食习惯与其他变量的关系。将所得变量各自与年龄、性别、婚姻状况、文化程度、职业等因素进行相关性分析，对于年龄等连续变量，本文通过 Pearson 系数判断变量之间的相关性，结果通过 t 检验后，最终通过 Pearson 相关系数得出居民的生活习惯和饮食习惯是否与年龄、性别、婚姻状况、文化程度、职业等因素相关。

2.3 问题三分析

对于问题三，我们需要系统而深入地剖析高血压、糖尿病等普遍性慢性病症与一系列生活方式因素之间的关联程度，包括吸烟习惯、酒精摄入、饮食偏好、日常作息、职业特性以及体力活动水平等。基于问题一和问题二中构建的模型框架，本研究进一步细化了饮食习惯与生活习惯的评估体系，旨在将复杂问题化繁为简。随后，本文分别构建了针对高血压、糖尿病等常见慢性疾病的线性回归模型。这些模型旨在量化分析吸烟、饮酒、饮食模式、生活习惯、职业环境以及体育活动等因素对上述慢性病发病风险的独立影响及其相互间的关联性。通过上述综合分析方法，从而得到各因素与慢性病之间的潜在联系，并评估这些关联的具体程度。

2.4 问题四分析

问题四旨在实现居民的合理分类，并为各分类群体量身定制促进身体健康的膳食与运动建议。具体而言，本文参照附件 2 的信息，将居民依据其婚姻状况及是否罹患慢性病，细化为四大类别：患病有配偶、患病无配偶、健康有配偶、健康无配偶。基于此前构建的模型框架，我们深入融合问题二与问题三的研究成果，实施深度剖析与综合考量，同时进一步优化模型，按年龄段对居民进行分类为青年、中年、老年三类人群，旨在针对每一类别的人群，提供具有针对性的膳食优化方案与科学运动指导，以促进其整体健康水平的提升。

三、 模型假设

为简化问题，本文做出以下假设：

- 生活习惯与饮食习惯间相互独立，两者之间影响较小或互不影响。
- 常见慢性病间不存在并发症关系。
- 附件 2 中数据均属实，具有可信度。

四、 符号说明

符号	含义
k^*	离差标准化后数据
k	原始数据
E_j	第 j 个指标的熵值
p_{ij}	第 i 个样本在第 j 个指标的比重
C	常数
w_j	权重
t	统计量
Y	因变量
β	回归系数
x	自变量
L	损失函数
$\hat{Y}_i$	第 i 个观测值的预测值
N	样本数
σ_a	标准差
$cov(a,b)$	两变量间的协方差
$C_p(a,b)$	Pearson 相关系数

五、 问题一的模型的建立和求解

5.1 模型简介

对问 aa 题一，依据附件 3 分析附件 2 饮食习惯的合理性并剖析问题。本文按附件 3 准则，将饮食习惯分为饮食规律与内容两类，内容依据《中国居民膳食指南》分为 11 大类。依据附件 3 数据，设合理性范围，建立评分卡模型，归一化数据后计算权重，导出得分排序，评估合理性并分析主要问题，流程图如图 1：

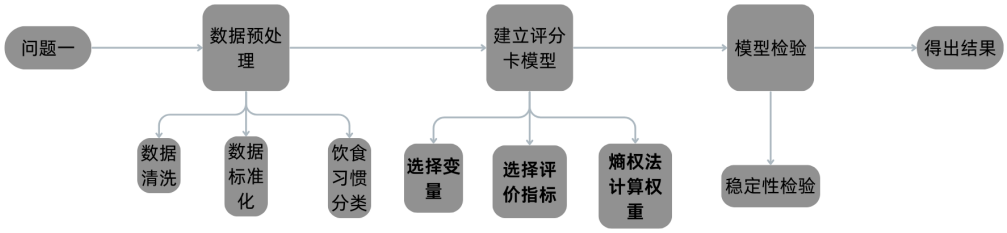


图 **1**　问题一求解流程图

5.2 模型建立

依据附件 3 中的数据，为提取的饮食习惯相关变量设置合理性范围，即评分指标，建立评分卡模型，即将指标数据归一化处理后，通过熵权法计算各项指标的权重，导出最终得分并排序，在通过稳定性检验后，可得到附件 2 中各个居民饮食习惯的合理性，并分析存在的主要问题。

5.2.1 数据预处理

对于附件 2，由于其中数据完整性和一致性较差，故需要进行数据预处理，以数据清洗、集成、归约、变换等方式提高数据质量。本文通过对附件 2 数据进行清洗，剔除异常值、缺失值以及与问题一无关的数据，同时为了方便数据处理, 本文将所有的 NAN 值转换为 0。最后运用 $0-1$ 标准化方法把原始数据进行线性变换，使结果落在 $[0,1]$ 区间内，设原始数据为 x 转换函数如式1：

$$k^* = \frac{k - \min k}{\max k - \min k} \tag{1}$$

式(1)中，k^* 表示离差标准化后的数据，$mink$ 为样本数据的最小值，$maxk$ 为样本数据的最大值。

5.2.2 数据可视化

对处理后的数据，利用 Matlab 绘制出附件 2 中居民饮食规律与饮食内容如下图所示：

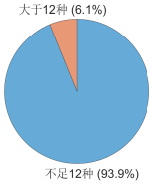


图 2　居民日摄入食物种类数达标比例

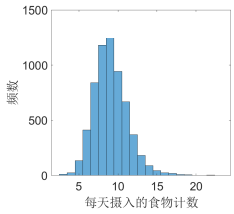


图 3　居民日摄入食物种类数分布

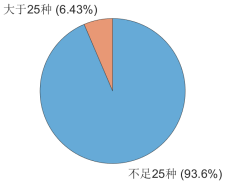


图 4　居民周摄入食物种类数达标比例

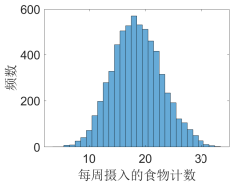


图 5　居民周摄入食物种类数分布

由上述 4 图可以得出，大部分居民每天、每周摄入的食物种类数都无法达到《中国居民膳食指南》中的要求，其中仅有 6.1% 居民每天摄入食物种类数达 12 种以上，仅 6.43% 居民每周摄入食物种类数达 25 种以上，较为不合理。

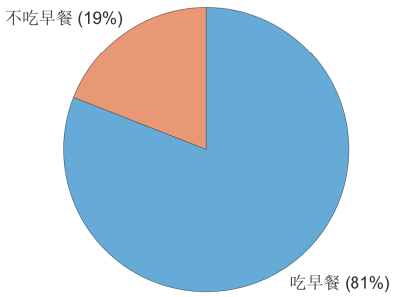


图 6　居民是否吃早餐比例

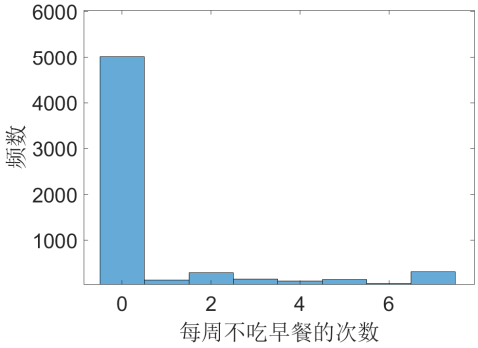


图 7　居民每周不吃早餐次数分布

由图 6与图 7不难看出，81% 居民有着吃早餐的习惯，且对于 19% 不吃早餐的居民，其每周不吃早餐次数以 2 次和 7 次居多，证明此类居民饮食不规律或有长期不吃早餐的习惯，即有着较差的饮食习惯。

由图 8、图 9可以看出，59.2% 的居民每天摄入食用盐的重量是符合标准的，且大部分超标的居民每天摄入食用盐的含量也在 10g 以内，因此，在此饮食习惯中，大部分居民较为合理。

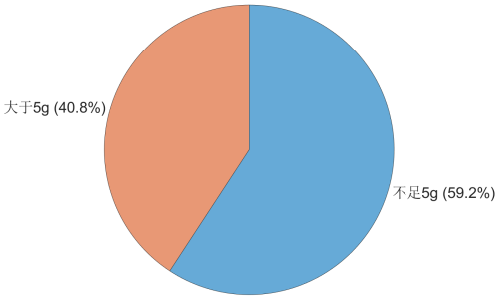


图 **8**　居民日摄入食用盐达标比例

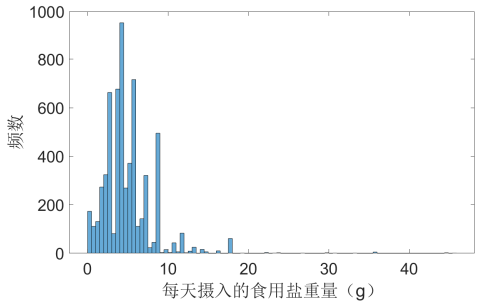


图 **9**　居民日摄入食用盐重量分布

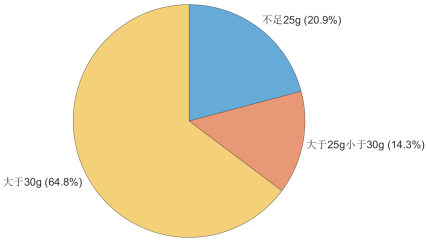


图 **10**　居民日摄入烹调油达标比例

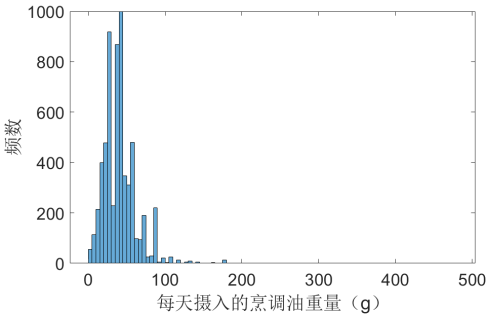


图 **11**　居民日摄入烹调油重量分布

由图 10、图 11可以看出，仅有 14.3% 的居民每天摄入烹调油的含量在《中国居民膳食指南》建议范围中，有 64.8% 居民每天摄入烹调油过量，这显然是不合理的。

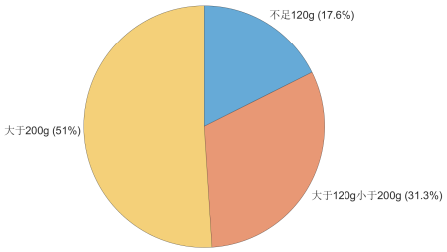


图 **12**　居民日摄入鱼肉蛋比例

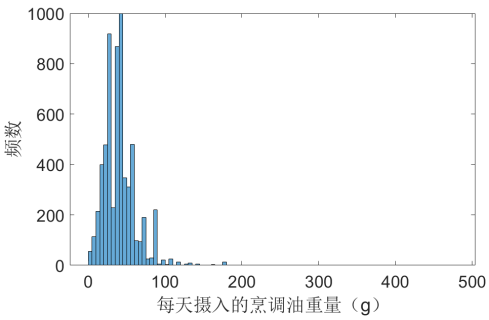


图 **13**　居民日摄入烹调油重量分布

由图 12、图 13不难看出，仅有 31.3% 的居民每天摄入鱼肉蛋符合《中国居民膳食指南》建议范围，其中 51% 居民超标，17.6% 居民未达标，但超标的居民中，近一半每天摄入的鱼肉蛋重量在 250 300g 间，超标重量较少。

由图 14、图 15不难得出，居民每天摄入新鲜水果量达标比例较低，为 26.5%，极大多数居民每天摄入新鲜水果重量在 100g 左右，且有小部分居民每天摄入新鲜水果量大于 350g，说明大部分居民在新鲜水果方面的饮食规律并不合理。

图 16、图 17直观反映了大部分居民每天摄入的新鲜蔬菜重量不足 300g，这一情况

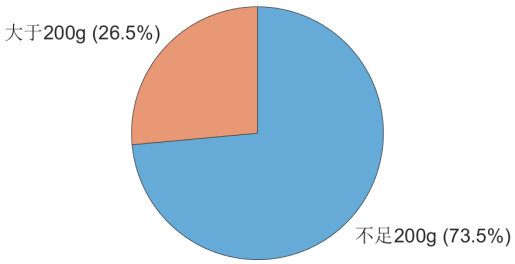


图 **14**　居民日摄入新鲜水果达标比例

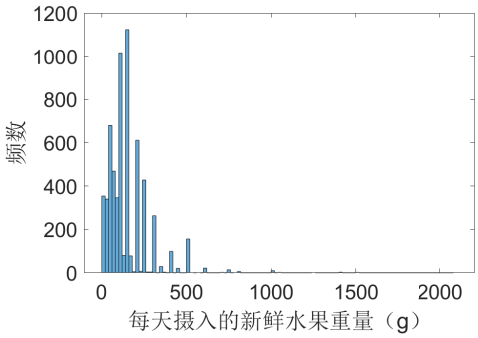


图 **15**　居民日摄入新鲜水果重量分布

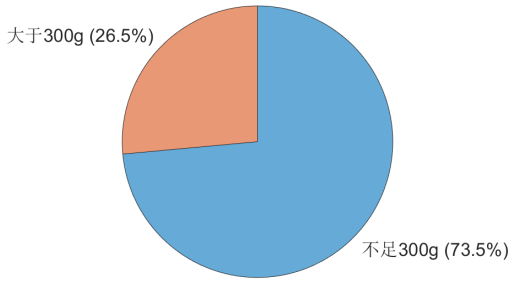


图 **16**　居民日摄入新鲜蔬菜达标比例

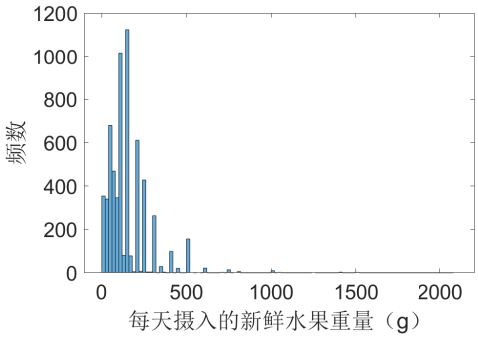


图 **17**　居民日摄入新鲜蔬菜重量分布

不符合《中国居民膳食指南》所设准则，并不合理。

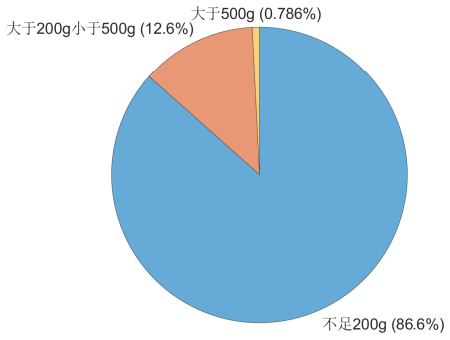


图 **18**　居民日摄入奶制品达标比例

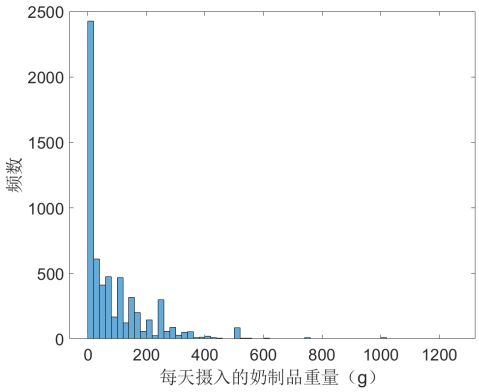


图 **19**　居民日摄入奶制品重量分布

《中国居民膳食指南》指出，奶制品的日摄入最高量为 500g，由图 18、图 19可以看出，绝大多数居民每天奶制品摄入量都符合这个标准，仅有 0.786% 的居民每天摄入奶制品超过 500g。这一情况说明绝大多数居民对此的饮食习惯较好。

5.2.3 建立评分卡模型

评分卡模型是广泛用于信用风险管理、营销决策和客户评级等领域的统计模型。它分析一组变量，评估特定对象的信用风险或其他相关特征，以辅助决策。

Step1：变量选择

选择饮食规律与全谷物、鱼禽蛋瘦肉、奶制品、豆制品、新鲜蔬菜、新鲜水果、油炸面食、含糖饮料、酒精、食用油与食用盐 11 大类饮食内容共 12 个变量。

Step2：选择评价指标

本文依据附件 3 中的健康饮食准则，列出了 10 个评价指标，分别为：指标 1 每天摄入 12 种以上食物，指标 2 每周摄入 25 种以上食物，指标 3 保证每天摄入不少于 300g 的新鲜蔬菜，指标 4 每天摄入 200 350g 的新鲜水果，果汁不能代替鲜果，指标 5 奶制品最高摄入量为 500g，指标 6: 平均每天摄入鱼禽、蛋类和瘦肉 120 200g，指标 7: 成人每天摄入烹调油 25 30g，指标 8: 食用盐＜ 5g，指标 9: 每周不吃早餐的次数，指标 10: 是否饮酒。其中，指标 3、4、6、7 属于区间型，指标 1 和 2 当做极大型处理，指标 8 需特殊处理。

Step3：熵权法计算指标熵值与权重

熵值可以反映指标的重要性与信息量，信息熵的计算公式为：

$$E_j = -k \sum_{i=1}^{m} p_{ij} \ln(p_{ij}) \tag{2}$$

其中，E_j 为第 j 个指标的熵值，p_{ij} 为第 i 个样本在第 j 个指标的比重，k 是常数。权重反映了指标的信息量，根据熵值计算权重的公式为：

$$w_j = \frac{1 - E_j}{\sum_{j=1}^{n} (1 - E_j)} \tag{3}$$

Step4：评分分级

根据计算出来的总得分，将居民的饮食习惯划分为不同的等级，以直观反映出居民饮食习惯的合理性。

5.2.4 基于 T 检验的稳定性检验

为了确保评分卡的有效性和稳定性，本文采用 t 检验对其进行检验和验证。t 检验是一种用于检验连续变量在不同时间段或样本中的均值是否存在显著差异的工具，即检验不同样本上重复评分中，模型的稳定性与一致性，具体步骤如下：

Step1：设置零假设 H_0

两个样本均值相等，没有差异

Step2：设置备择假设 H_1

两个样本均值不等，有差异

Step3：计算 t 统计量

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \tag{4}$$

式(4)中，$\bar{X}_1$ 和 $\bar{X}_2$ 是两个样本的均值，s_p 是两个样本的合并标准差，n_1 和 n_2 是两个样本的大小。

Step4：确定 t 分布的自由度和临界值

设定置信水平为 95%，则显著性水平 $\alpha = 1 - 95\% = 0.05$，根据样本量 n 和显著性水平 α 确定 t 分布的自由度和临界值，依据 t 分布的自由度和临界值选择拒绝域的边界。

Step5：检验 t 统计量

检验 t 统计量 t 值是否落在 t 分布的拒绝域内。如果 t 统计量落在拒绝域内，则拒绝零假设，认为两个变量之间存在显著的线性相关关系；否则接受零假设，认为两个变量之间不存在显著的线性相关关系。一般来说，如果 p 值小于显著性水平 α，就可以拒绝原假设。p 值是在给定原假设下观察到的统计量或者更极端统计量的概率。

5.3 模型求解

Step1：评分卡模型评分

选取 12 个变量，涵盖饮食规律及 11 大类饮食内容。依据健康饮食准则，列出 10 个评价指标，包括食物种类、摄入量等，其中指标 8 要求每天摄入食用盐 <5g，需要先对数据进行正向化处理后再进行标准化处理。本文对于指标 8 的正向化处理选择将对应的数值数据替换为它的相反。而后通过熵权法，利用熵值反映指标重要性，计算熵值与权重，公式分别为：

$$E_j = -k\sum_{i=1}^{m} p_{ij}\ln(p_{ij}) \tag{5}$$

$$w_j = \frac{1 - E_j}{\sum_{j=1}^{n}(1 - E_j)} \tag{6}$$

在确定指标权重后，计算得出居民饮食习惯合理性评分分布图 20：

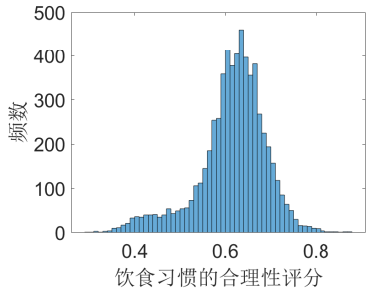


图 **20**　居民饮食习惯合理性评分分布图

图 20中大部分居民饮食习惯评分在 0.6 左右，此时需要一个评判标准，用于判断饮食习惯是否合理。

Step2：稳定性检验

得出评分结果后，本文运用 t 检验来检验模型的稳定性, 进行多次重复评分，将每次的的评分结果代入式(4)中得出 t 值，查找 t 分布表可知 t 值落在拒绝域内，故拒绝零假设，证明两次评分结果均值相等。多次重复评分卡模型，所得 t 值均落在拒绝域内，则所建立模型具有较好的稳定性。

Step3：评判合理性

经过稳定性检验后，仍需要寻找一个阈值去判断居民饮食习惯是否合理，本文选择通过设置各指标刚好满足时得到的饮食习惯评分作为评价的合理性的标准，计算得到具有合理性饮食习惯的最低评分为 0.7428。以此评分为分界点可以得到居民饮食习惯合理性比例如图 21：

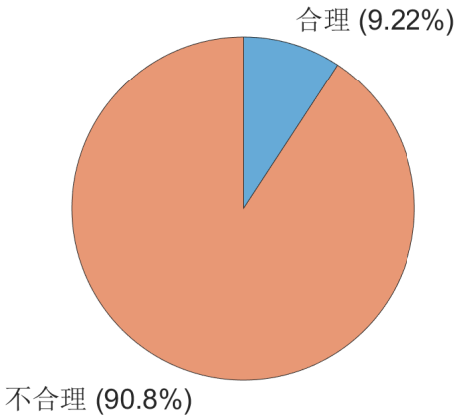


图 **21**　居民饮食习惯合理性比例

由图 21可以看出，仅有 9.22% 的居民饮食习惯是合理的，这一结果说明大部分居民的饮食习惯都有不合理处，使其综合评分降低。

Step4：总结

综合上述分析，可明确得出结论：附件 2 所展示的居民饮食习惯在整体上呈现出不合理性，其关键问题聚焦于两方面：一是广大居民每日新鲜蔬菜与新鲜水果的摄入量普遍未能达到充足水平；二是绝大多数居民每日烹调用油的摄入量明显超标。此外，还值得注意的是，绝大多数居民无论是每日还是每周食用的食物种类数，均未能达到《中国居民膳食指南》所规定的标准。

六、 问题二的模型的建立和求解

6.1 模型简介

针对问题二，题设探究居民生活习惯与饮食习惯是否受年龄、性别、婚姻、文化程度、职业影响。本文基于问题一所建立模型，选取饮食习惯中差异显著的变量，及以吸烟、体育锻炼代表生活习惯。采用 Pearson 系数分析变量间相关性，并经 t 检验，最终确定生活习惯、饮食习惯与上述因素的关系，流程如图 22：

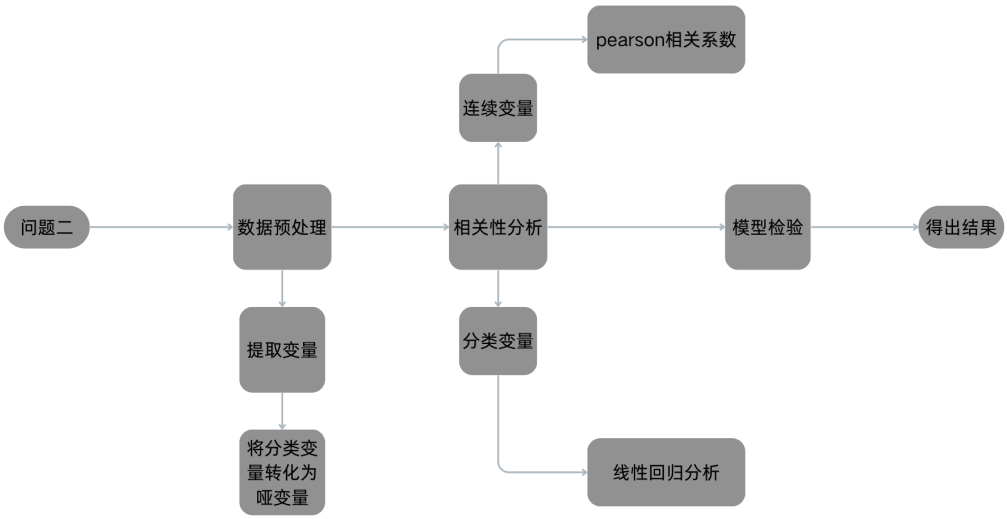


图 **22**　问题二求解流程图

6.2 模型建立

6.2.1 数据预处理

对于附件 2，在问题一数据清洗的基础上，选取居民生活习惯、饮食习惯、个人基本信息等与问题二相关的数据，重新进行归一化处理。本文运用 $0-1$ 标准化方法把原始数据进行线性变换，使结果落在 $[0,1]$ 区间内，设原始数据为 x 转换函数如下：

$$k^* = \frac{k - \min k}{\max k - \min k} \tag{7}$$

6.2.2 数据可视化

对处理后的数据，绘制生活习惯中所提取变量各自与居民个人信息对应的统计图如图 23。

由图 23易得，居民中吸烟人数占比为 19.4%，属于较少的人群，反映大部分居民在此方面的生活习惯较好。

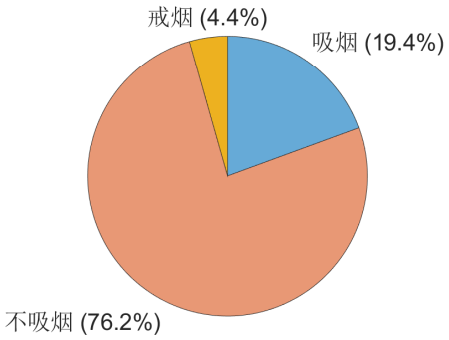


图 **23**　居民吸烟情况比例

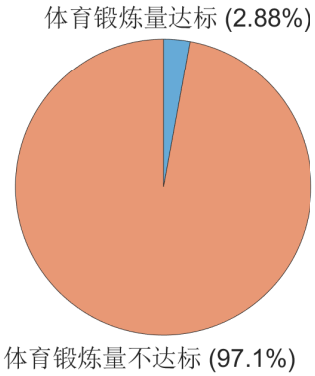


图 **24**　居民体育锻炼达标比例

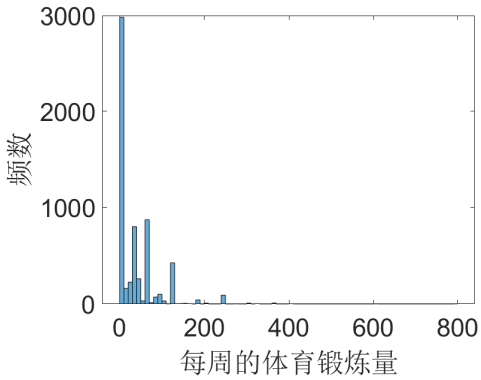


图 **25**　居民周体育锻炼情况分布图

由图 24与图 25可知，居民每周体育锻炼符合《中国居民膳食指南》所规定的标准的比例极少，仅有 2.88%，反映出附件 2 中居民在锻炼方面的懈怠，具有较大的不足，应改进自身在此方面的生活习惯，积极锻炼，养成良好的生活习惯，塑造健康的身体。

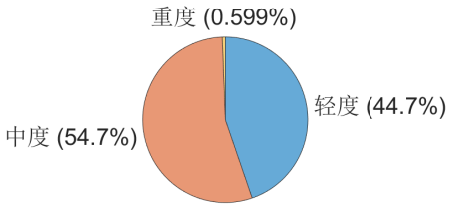


图 **26**　居民家务强度比例图

由图 26得，大部分居民家务强度处于居中状态，既可以做到一定量的身体活动，又不会带来太大负担，确保了身体活动量，故推荐居民适度地进行一些家务劳动。

6.2.3 建立线性回归模型

为分析生活习惯和饮食习惯与其他变量的关系，我们没需要建立线性回归模型。

线性回归的目标是找到一个最佳拟合线，通常表示为：

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \varepsilon \tag{8}$$

其中，Y 是因变量（目标变量）。β_0 是截距项（即当所有自变量为零时，因变量的预期值）。$\beta_1\beta_2...\beta_n$ 是自变量 $x_1x_2...x_n$ 的回归系数，表示各自变量对因变量的影响程度。ε 是误差项，表示模型未能解释的部分。

为寻找到最佳拟合曲线，需计算损失函数，用最小二乘法来估计回归系数。损失函数计算公式为：

$$L = \sum_{i=1}^{m} \left(Y_i - \hat{Y}_i\right)^2 \tag{9}$$

其中，m 是样本数量。Y_i 是第 i 个观测值的实际因变量值。$\hat{Y}_i$ 是第 i 个观测值的预测值。通过最小化损失函数，即使得误差平方和最小，就可以找到最佳的回归系数估计。最终，通过拟合曲线的 p 值，得到生活习惯和饮食习惯与其他变量的关系。

6.2.4 建立 Pearson 相关系数评价模型

Step1：计算变量的均值与标准差

均值是衡量变量集中趋势的统计量，其计算公式为：

$$\bar{X}_i = \frac{1}{N} \sum_{i=1}^{N} X_i \tag{10}$$

标准差是用来衡量数据集合中数据点的离散程度或波动性的统计量，以变量 a 为例，标准差公式如下：

$$\sigma_a = \sqrt{\frac{\sum_i^N (a_i - \bar{a})^2}{N}} \tag{11}$$

其中 N 表示变量的样本数目，即居民数，a_i 表示 a 的第 i 个取值，$\bar{a}$ 是 a 的均值。

Step2：计算两个变量间的协方差

变量间的协方差是用来度量两个变量间变化趋势是否一致，协方差公式如下：

$$cov(a,b) = \frac{1}{N} \sum_{i}^{N} (a_i - \bar{a})(b_i - \bar{b}) \tag{12}$$

Step3 计算变量间的 Pearson 相关系数

Pearson 相关系数可以用来度量两个变量 a 和 b 之间的线性相关关系，其值介于-1到 1 之间。这里定义两个变量 a 和 b 之间的 Pearson 相关系数 $C_p(a,b)$ 为：

$$C_p(a,b) = \frac{cov(a,b)}{\sigma_a \sigma_b} = \frac{\sum_i^N (a_i - \bar{a})(b_i - \bar{b})}{\sqrt{\sum_i^N (a_i - \bar{a})^2} \sqrt{\sum_i^N (b_i - \bar{b})^2}} \tag{13}$$

6.2.5 对于配对样本的 t 检验

对计算所得的相关系数进行假设检验，此处 t 检验是对于配对样本的 t 检验，与问题一检验过程相比，在计算 t 统计量时有所差异，其余步骤均相同，此处计算各组相关系数的 t 统计量为：

$$t = C_p(a,b)\sqrt{\frac{N-2}{1-[C_p(a,b)]^2}} \tag{14}$$

查找 t 分布表，确认 t 统计量 t 值是否落在 t 分布的拒绝域内。即检验 p 值是否小于显著性水平 α，如果 t 统计量落在拒绝域内，则拒绝零假设，认为两个变量之间存在显著的线性相关关系；否则接受零假设，认为两个变量之间不存在显著的线性相关关系。

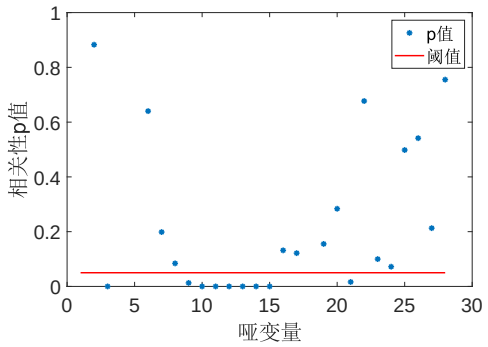
6.3 模型求解

Step1：将分类变量转化为哑变量

本文探讨的影响因素中，诸如性别、婚姻状况、文化程度、职业等因素都是分类变量，为寻找其与饮食习惯、生活习惯的相关性，需要将其转化为哑变量，本文通过标签编码，将分类变量中每个类别用一个整数代表，从而将分类变量转化为哑变量，结果见附录：

Step2：分析哑变量与生活、饮食习惯的相关性

本文通过线性回归模型，寻找最佳拟合曲线，以分析生活习惯和饮食习惯与其他变量的关系，这里计算得出哑变量与生活习惯评分、饮食习惯评分的最佳拟合曲线。最终得出哑变量与生活、饮食习惯评分的相关性 p 值分布：



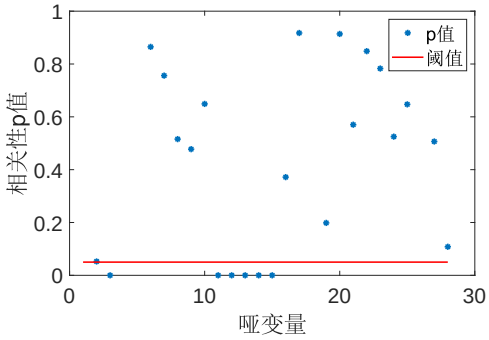


图 **27**　生活习惯评分与哑变量相关性 p 值分布　图 **28**　饮食习惯评分与哑变量相关性 p 值分布

图 27与图 28中，横坐标为哑变量，坐标数值代表不同的哑变量，p 值分布在阈值以下的哑变量大多在 10 15 之间，其展现的是婚姻状况这一分类变量，说明饮食习惯、生活习惯与婚姻状况这一因素具有明显的线性相关性。

Step 3：Pearson 相关系数分析年龄与生活、饮食习惯的相关性

计算连续变量与生活习惯评分、饮食习惯评分的 Pearson 相关系数，计算公式为：

$$C_p(a,b)=\frac{cov(a,b)}{\sigma_a\sigma_b}=\frac{\sum_i^N(a_i-\bar{a})(b_i-\bar{b})}{\sqrt{\sum_i^N(a_i-\bar{a})^2}\sqrt{\sum_i^N(b_i-\bar{b})^2}} \tag{15}$$

由于连续变量仅有年龄，得其与生活、饮食习惯评分的 Pearson 相关系数及 p 值如表1所示：

	相关系数	p 值
年龄与生活习惯	-0.125	0.3251
年龄与饮食习惯	0.461	0.00028429

表 **1**　年龄与生活习惯、饮食习惯的相关系数与 p 值

由表 1不难看出，年龄与生活习惯可能没有较强的相关性，但与饮食习惯具有一定的线性关系，且其 p 值小于显著性水平 0.05，即通过 t 检验。

Step4：总结

综上所述居民的生活习惯和饮食习惯与婚姻状况具有较强的相关性，与文化程度、职业、性别等因素并不相关。

此外，居民的饮食习惯与其年龄相关，居民的生活习惯与其年龄并不相关不相关，结果如表 2所示：

	年龄	性别	婚姻状况	文化程度	职业
生活习惯	不相关	不相关	相关	不相关	不相关
饮食习惯	相关	不相关	相关	不相关	不相关

表 **2**　求解结果

七、 问题三的模型的建立和求解

7.1 模型简介

本题旨在深入探讨高血压、糖尿病等慢性病与吸烟、饮酒、饮食习惯、生活习惯、工作性质和运动等因素之间的关系及其相关程度。本文基于先前建立的模型，对饮食和生活习惯进行评级以简化问题。而后通过计算 Pearson 相关系数并进行 t 检验，综合分析了这些慢性病与吸烟、饮酒、饮食偏好、生活习惯、职业特性及体育活动等因素之间的联系与相关性，流程如图 29：

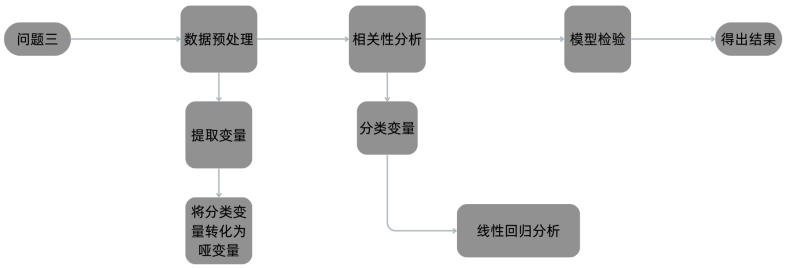


图 **29**　问题三求解流程图

7.2 模型建立

7.2.1 数据预处理

在问题一、二所用数据基础上，提取附件 2 中与常见慢性病相关数据，对其进行归一化处理，运用 $0-1$ 标准化方法转化原始数据，使结果落在 $[0,1]$ 区间内，设原始数据为 x 转换函数如 (16)：

$$k^{*}=\frac{k-\min k}{\max k-\min k} \tag{16}$$

7.2.2 建立 Pearson 相关系数评价模型

Step1：计算变量的均值与标准差

均值计算通式：

$$\bar{X}_i=\frac{1}{N}\sum_{i=1}^{N}X_i \tag{17}$$

标准差计算通式：

$$\sigma_a = \sqrt{\frac{\sum_i^N (a_i - \bar{a})^2}{N}} \tag{18}$$

Step2：计算两个变量间的协方差

协方差计算通式：

$$cov(a,b) = \frac{1}{N}\sum_i^N (a_i - \bar{a})(b_i - \bar{b}) \tag{19}$$

Step3：计算变量间的 Pearson 相关系数

两个变量 a 和 b 之间的 Pearson 相关系数 $C_p(a,b)$ 为：

$$C_p(a,b) = \frac{cov(a,b)}{\sigma_a \sigma_b} = \frac{\sum_i^N (a_i - \bar{a})(b_i - \bar{b})}{\sqrt{\sum_i^N (a_i - \bar{a})^2}\sqrt{\sum_i^N (b_i - \bar{b})^2}} \tag{20}$$

通过上述步骤计算得出高血压、糖尿病等常见慢性病与吸烟、饮酒、饮食习惯、生活习惯、工作性质、运动等因素的 Pearson 相关系数。

7.2.3 模型检验

对于得出的 Pearson 相关系数结果，需要进行假设检验以确定其可信度，本文在这里采用 t 检验，具体步骤如下:

Step1：设置零假设 H_0

相关性系数为 0，即两个变量间不存在线性相关关系。

Step2：设置备择假设 H_1

相关性系数不为 0，即两个变量间存在显著的线性相关关系。

Step3：计算 t 统计量

计算公式如下：

$$t = C_p(a,b)\sqrt{\frac{N-2}{1-[C_p(a,b)]^2}} \tag{21}$$

Step4：确定 t 分布的自由度和临界值

设定置信水平为 95%，则显著性水平 $\alpha = 1 - 95\% = 0.05$，根据样本量 n 和显著性水平 α 确定 t 分布的自由度和临界值，依据 t 分布的自由度和临界值选择拒绝域的边界。

Step5：检验 t 统计量

检验 t 统计量 t 值是否落在 t 分布的拒绝域内。如果 t 统计量落在拒绝域内，则拒绝零假设，认为两个变量之间存在显著的线性相关关系；否则接受零假设，认为两个变量之间不存在显著的线性相关关系。一般来说，如果 p 值小于显著性水平 α，就可以拒绝原假设。p 值是在给定原假设下观察到的统计量或者更极端统计量的概率。

7.3 模型求解

Step1：将分类变量转化为哑变量

问题三的影响因素中，皆为分类变量，为寻找其与饮食习惯、生活习惯的相关性，需要将其转化为哑变量，本文通过标签编码，将分类变量中每个类别用一个整数代表，从而将分类变量转化为哑变量，结果见附录：

Step2：线性回归分析哑变量与生活、饮食习惯的相关性

本文通过线性回归模型，寻找最佳拟合曲线，以分析生活习惯和饮食习惯与其他变量的关系，这里计算得出哑变量与生活习惯评分、饮食习惯评分的最佳拟合曲线为：

最终得出哑变量与生活、饮食习惯评分的相关性 p 值分布如下：

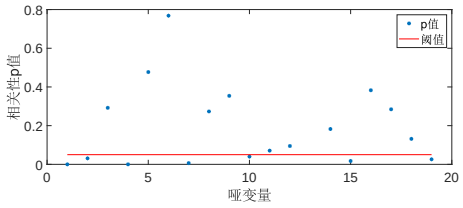


图 **30**　糖尿病与哑变量的相关性 p 值分布

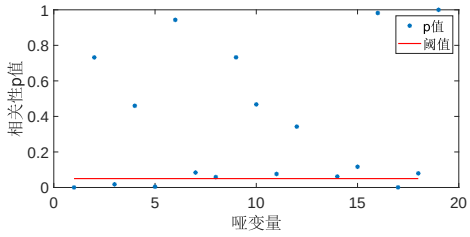


图 **31**　高血压与哑变量的相关性 p 值分布

图 30和图 31中，横坐标为哑变量，坐标数值代表不同的哑变量，反映出糖尿病与饮食习惯、生活习惯、运动量、工作性质等因素相关，高血压与饮食习惯、吸烟、饮酒、工作性质等因素相关。

Step3：总结

综上所述，可得模型求解结果如表 3所示：

		饮食习惯	生活习惯	吸烟	饮酒	运动量	工作性质
慢性病	高血压	相关程度高	相关程度低	相关程度高	相关程度高	相关程度低	相关程度高
	糖尿病	相关程度高	相关程度高	相关程度低	相关程度低	相关程度高	相关程度高

表 **3**　求解结果

7.4 求解结果

其中，常见慢性病（如高血压、糖尿病等）与吸烟、饮酒、饮食习惯、生活习惯、工作性质、运动等因素具有线性关系，其中，糖尿病与饮食习惯、生活习惯、运动量、工作性质有较强的线性关系，相关程度高，而糖尿病与吸烟、饮酒并无明显的线性关系，它们的相关程度较低，高血压与饮食习惯、吸烟、饮酒、工作性质具有较强的线性关系，它们具有较高的相关程度，而高血压与生活习惯可能并无较大的相关性。

八、 问题四的模型的建立和求解

8.1 模型简介

问题四要求对居民合理分类，并对各类人群提出有利于身体健康的膳食、运动等方面的合理建议，本文先将附件 2 中居民按婚姻状况及是否患有慢性病分为患病有配偶、患病无配偶、健康有配偶、健康无配偶。在前文所建立的模型基础上，根据问题 2 与问题 3 的结果，综合分析，针对各类人群提出有利于身体健康的的膳食、运动方面的合理建议。

8.2 综合分析

由前文三个问题所得结果可知，婚姻状况与生活、饮食习惯相关，常见慢性病与较多因素相关，故将附件 2 中居民按婚姻状况及是否患有慢性病分为患病有配偶、患病无配偶、健康有配偶、健康无配偶四类人群，具体分析如下：

患病有配偶：对于患慢性病有配偶人群，饮食方面要均衡膳食，增加膳食纤维摄入（如全谷物、蔬菜、水果），促进消化，选择低 GI（血糖生成指数）食物，减少精制糖及甜食摄入，同时遵循医生或营养师的建议，针对具体病症调整饮食，如糖尿病患者需严格控制碳水化合物。生活方面，建议每周至少 150 分钟的中等强度有氧运动，和伴侣一起参加健身活动，如散步、骑自行车或参加健身课程，增加家庭互动。如有吸烟、饮酒等情况，应适量减少或戒烟戒酒。

患病无配偶：对于患慢性病无配偶人群，需要个人定制膳食，尝试记录饮食，了解自己的饮食习惯，选择健康的饮食选项，同时选择易于携带和储存的健康零食（如坚果、酸奶等），避免快餐和高热量零食，在日常生活中，保持足够的水分摄入，限制含糖饮料，同时参与团体健身课程，结识志同道合的朋友，提高运动动力或者使用运动 App 或穿戴设备跟踪运动情况，设定目标，增强自我激励。如有吸烟、饮酒等情况，应减少吸烟、饮酒的频率或戒烟戒酒。

健康有配偶：对于身体健康且有配偶人群，要维持均衡营养，确保饮食中包含多样化的食品，包括新鲜水果、蔬菜、全谷物和瘦肉，同时鼓励家庭一起烹饪，尝试新的健康食谱，避免外出就餐。如若有饮酒情况保持适量，建议男性每日不超过两杯，女性不超过一杯。在生活习惯方面，应养成良好的锻炼习惯，夫妻间制定共同的运动计划，例如周末徒步旅行、骑行等，以确保每周的身体活动，同时也可以加入社区或社交团体，积极参加团体游戏或运动会。

健康无配偶：对于身体健康但无配偶人群，应养成良好的生活习惯与饮食习惯，在饮食方面，学习基础的烹饪技巧，参与烹饪课程，减少不健康饮食，同时，选择以植物为基础的饮食，增加蔬菜和水果的摄入，保持低脂肪和高纤维，此外，可以尝试间歇性饮食，偶尔尝试简单的排毒饮食，以改善身体状态。在生活习惯方面，应多样化训练，

尝试多种形式的运动，如瑜伽、舞蹈、极限运动等，找到自己喜欢的项目，亦可以参加运动俱乐部或团体活动，通过社交互动增加锻炼乐趣。

总结：无论是患病还是健康人群，无论有无配偶，保持规律的锻炼、均衡的饮食和良好的生活习惯都是促进身体健康的关键，建议建立健康的生活模式，与家人和朋友保持良好的社交关系，从而提高整体生活质量。

8.3 模型改进

由于上述解决方案是基于问题二、三得到的粗糙分类，本文亦选取年龄作为分类标准，将居民分为青年、中年、老年三类人群，得到其生活、饮食习惯、吸烟、喝酒等个人情况的统计图，从而对客观数据进行分析，给出各类居民合理的膳食、运动等方面的建议。数据可视化结果如图 32。

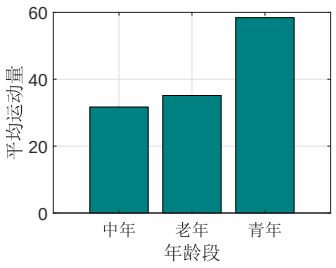
由图 32可知，三类人群的平均饮食习惯与生活习惯评分接近，说明每类人群都有各自不同的饮食、生活习惯的不足之处，需要分布进行细致分析，具体分析如下：

青年不吃早餐的占比在本类人群中较多，且青年具有饮酒、抽烟的现象，此外，青年的运动量最高，应注意保持良好的锻炼习惯，故而青年应注意合理膳食，培养健康的饮食规律，不能频繁不吃早餐，同时抽烟、饮酒都要少量，避免成瘾，伤害身体健康。值得注意的是，青年患高血压的比例与中年接近，这一趋势警戒青年要选择健康的饮食选项，同时可以选择易于携带和储存的健康零食，避免快餐和高热量零食在自身饮食中的占比。

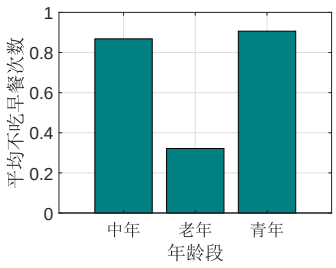
对于中年人群，中年人群亦有较高的不吃早餐频率，且中年的平均饮酒量与吸烟量都过高，需要有所减少，适量饮酒、抽烟，故而，中年人群应注意健康的饮食规律，控制油盐的摄入量。同时，中年人群的平均运动量最低，需要提高运动量，注意体育锻炼在个人生活中的重要性，建议每周至少 150 分钟的中等强度有氧运动，可以进行慢跑、骑自行车等活动，亦可参加健身课程。

对于老年人群，其主要问题集中在吸烟量极高且慢性病如高血压、糖尿病的患病比例较高，说明老年人群需要减少吸烟的频率，尽量戒烟，在日常饮食中需要主要控制减少油盐的摄入量，控制膳食的平衡与合理性，做到健康饮食，均衡膳食，增加膳食纤维摄入（如全谷物、蔬菜、水果），促进消化，选择低 GI（血糖生成指数）食物，减少精制糖及甜食摄入。

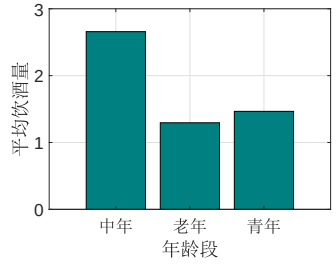
总结，青年、中年和老年三类人群均存在不良饮食和生活习惯问题。青年应避免不吃早餐、减少烟酒摄入，并保持运动；中年人需控制油盐摄入、增加运动量；老年人应戒烟、控制油盐摄入、均衡膳食，增加膳食纤维和低 GI 食物摄入，以预防慢性病。



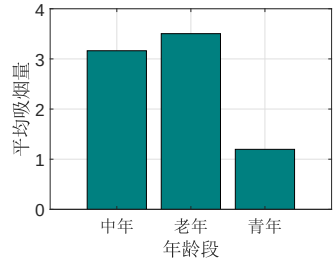
(a) 三类人群平均运动量分布



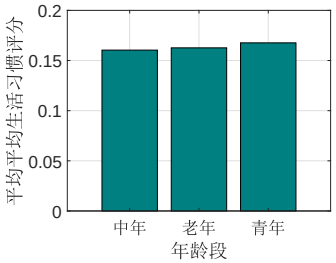
(b) 三类人群平均不吃早餐次数分布



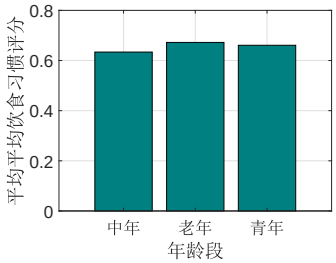
(c) 三类人群平均饮酒量分布



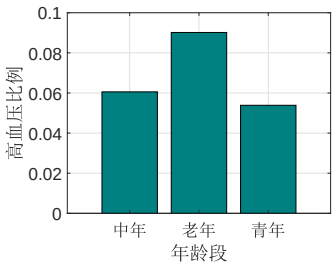
(d) 三类人群平均吸烟量分布



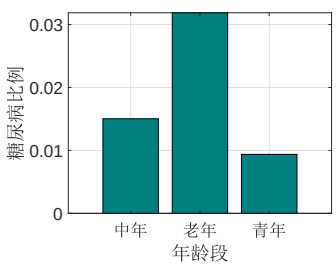
(e) 三类人群平均生活习惯评分分布



(f) 三类人群平均饮食习惯评分分布题



(g) 三类人群中高血压比例分布



(h) 三类人群中糖尿病比例分布

图 **32**　三类人群个人情况分布图组

九、 模型的评价

9.1 模型的优点

- Pearson 相关性分析简单易懂，表达直观，可解释性强。
- 线性回归模型易于理解和解释，计算速度较快，特别适合处理大型数据集，同时可以通过回归系数直观地了解各个自变量对因变量的影响程度。

• 直方图可以展示连续数据的分布特征，在处理大数据集和识别模式时表现较好，饼状图适用于展示分类数据中各部分的比例关系，适合简单的分类数据分析

9.2 模型的缺点

• Pearson 相关性分析只能检测到线性关系，对于非线性关系的相关性无法很好地反映，且要求变量符合正态分布，如果数据不符合正态分布，则相关性分析结果可能不准确。

• 线性回归要求变量之间存在线性关系，且假设误差项是独立同分布的，若这些假设不成立，可能导致误导性的结果，且对于具有非线性关系的数据，线性回归模型的预测能力较差，可能需要更复杂的模型。

• 饼状图适合少量类别，若类别过多，图形会显得拥挤且难以读取，直方图难以准确比较不同部分之间的比例，尤其是部分非常接近时，人的视觉感知会受限，导致较小的差异不易发现。

参考文献

[1] 司守奎, 孙玺菁. 数学建模算法与应用[M]. 北京: 国防工业出版社, 2011.

[2] 卓金武. MATLAB 在数学建模中的应用[M]. 北京: 北京航空航天大学出版社, 2011.

[3] COHEN I, HUANG Y, CHEN J, et al. Pearson correlation coefficient[J]. Noise reduction in speech processing, 2009:1-4.

[4] MISHRA P, SINGH U, PANDEY C M, et al. Application of student's t-test, analysis of variance, and covariance[J]. Annals of cardiac anaesthesia, 2019, 22(4):407-411.

[5] SHEUGH L, ALIZADEH S H. A note on pearson correlation coefficient as a metric of similarity in recommender system[C]//2015 AI & Robotics (IRANOPEN). [S.l.]: IEEE, 2015: 1-6.

[6] 郑英丽, 朴丽莎, 王丽珍. Pearson 相关系数下非对称相似度计算及其应用[J]. 云南民族大学学报 (自然科学版):1-11.

附录 A　文件列表

文件名	功能描述
q1.m	问题一程序代码
q2.m	问题二程序代码
q3.m	问题三程序代码
q4.m	问题四程序代码

附录 B　代码

pr1.m

```
1
2  %% 数据预处理
3  clc;clear
4  url = 'raw_data.xlsx';
5  data = readtable(url);
6  text = table2array(data);
7  [row,col] = size(text);
8  panduan = zeros(row,1);
9  %15 33 197
10 text = [text(:,1:14) panduan text(:,15:col)];
11 text = [text(:,1:32) panduan text(:,33:col+1)];
12 text = [text(:,1:196) panduan text(:,197:col+2)];
13 %数据预处理
14 [row,col] = size(text);
15 %方便数据处理,将所有的NAN值转换为0
16 text(find(isnan(text)==1)) = 0;
17 for i=1:row
18     %处理异常值
19     %吸烟
20     if text(i,9) == 3
21         if text(i,10) > 0 || text(i,11) > 0 || text(i,12) > 0
22             text(i,15) = -1;
23         end
24     end
```

```
    %饮酒
    if text(i,16) == 2
        if sum(text(i,17:32)) > 0
            text(i,33) = -1;
        end
    end
    %处理缺失值
    %吸烟
    if text(i,9) == 0
        if text(i,10) > 0 || text(i,11) > 0 || text(i,12) > 0
            text(i,1) = 1;
        else
            text(i,1) = 3;
        end
    end
    if text(i,16) == 0
        if sum(text(i,17:32)) > 0
            text(i,1) = 1;
        else
            text(i,1) = 2;
        end
    end
end
%饮食
%是否吃 天/次 周/次 月/次 平均每次的摄入量
[row,col] = size(text);
for i=1:row
    for j = 55:5:197-8
        %处理错误值
        if text(i,j) == 1
            if text(i,j+1)+text(i,j+2)+text(i,j+3) == 0 ||
   text(i,j+4) == 0
                text(i,197) = -1;
                break;
            end
```

```
        end
        %处理缺失值
        if text(i,j) == 0
            %存在摄入频率即为1
            if text(i,j+1)+text(i,j+2)+text(i,j+3)> 0
                text(i,j) = 1;
            else
                text(i,j) = 2;
            end
        end
    end
end

%去除掉无法计算家庭人数的数据
for i = 1:row
    if text(i,39)+text(i,40)+text(i,46)+text(i,47)+text(i,53)+
   text(i,54) == 0
        text(i,197) = -1;
    end
end

%去除吃的太多的数据
for i = 1:row
    if  text(i,146) > 4 || text(i,149) > 15 ||text(i,192) >105
    ...
        || text(i,179) >50 ||text(i,106) >4 ||text(i,111) >4
   || text(i,109) >100 || text(i,119) >100
        text(i,197) = -1;
    end
end


%去除年龄不合理的和年龄偏大的,其他指标不合理的
for i = 1:row
    if text(i,2) <=1943 || text(i,225) <=130 || text(i,226)
```

```
   <=35 || text(i,226)>=140
        text(i,197) = -1;
    end
    if find(text(i,227:col) == 0)
        text(i,197) = -1;
    end

end

t = 1;
for i = 1:1:row
    if(text(i,15) > -1 && text(i,33) > -1 && text(i,197) > -1)
        new1(t,:) = text(i,:);
        t = t + 1;
    end
end

%获得预处理后的数据new1 进行处理
[row,col] = size(new1);


%% 得出所有指标数据
%指标1:每天摄入12种以上食物计数
zhibiao1 = zeros(row,1);
for i=1:row%从第一个到最后一个居民
    count = 0;
    for j = 55:5:197-8
        if new1(i,j+1) ~= 0
            %判断天数
            count= count+1;
        elseif new1(i,j+2) >= 7
            %平均每天吃一次
            count=count+1;
        elseif new1(i,j+3)>=28
```

```
            %一个月看做是28天,平均每天一次
            count=count+1;
        end
    end
    for j=190:196
        if new1(i,j) > 0
            count=count+1;
        end
    end
    zhibiao1(i,1) = count;
end
%指标2 每周摄入25种以上
zhibiao2 = zeros(row,1);
for i=1:row%从第一个到最后一个居民
    count = 0;
    for j = 55:5:197-8
        if new1(i,j+1) ~= 0
            %判断天数
            count= count+1;
        elseif new1(i,j+2) ~= 0
            %每周都吃
            count=count+1;
        elseif new1(i,j+3)>=4
            %一个月看做是28天,至少平均每周一次
            count=count+1;
        end
    end
    for j=190:196
        if new1(i,j) > 0
            count=count+1;
        end
    end
    zhibiao2(i,1) = count;
end
%指标3 保证每天摄入不少于300g的新鲜蔬菜
```

```
zhibiao3 = meitianyinshi(145,row,new1,50);
%指标4 保证每天摄入200~350g的新鲜水果，果汁不能代替鲜果
zhibiao4 = meitianyinshi(175,row,new1,50);
%指标5 吃各种各样的奶制品，与第四版相比最高摄入量由原来的300g
   提高到500g
%奶粉 鲜奶 酸奶
zhibiao5 = meitianyinshi(105,row,new1,50)+meitianyinshi(110,
   row,new1,10)+meitianyinshi(115,row,new1,50);
%指标6:鱼禽、蛋类和瘦肉摄入要适量，平均每天120~200g
%80
zhibiao6 = meitianyinshi(80,row,new1,50)+meitianyinshi(85,row,
   new1,50)+meitianyinshi(90,row,new1,50)...
    +meitianyinshi(100,row,new1,50)+meitianyinshi(120,row,new1
   ,50);
%指标7:成人每天摄入烹调油25~30g
%将三餐就餐的平均人数四舍五入作为其家庭成员人数
zhibiao7 = jiatingyinshipanduan(190,row,new1,500)+
   jiatingyinshipanduan(191,row,new1,500);
% zhibiao7 = (植物油和动物油)
%指标8:食用盐＜5g
zhibiao8 = jiatingyinshipanduan(192,row,new1,50);

%指标9:饮食规律 不吃早餐+中餐+晚餐
zhibiao9 = new1(:,34);

%指标10:吸烟与否
% （可以删）
zhibiao10 = new1(:,9);


%指标11 :饮酒与否
zhibiao11 = new1(:,16);
zhibiao11((zhibiao11 == 3)|(zhibiao11 == 0)) = 2;

```

```
189
190 %% 数据可视化
191
192 %指标1:每天摄入12种以上食物计数
193 figure
194 % 绘制直方图
195 h = histogram(zhibiao1, 'Normalization', 'count');
196
197
198 % 为直方图添加标签（可选）
199 xlabel('每天摄入的食物计数');
200 ylabel('频数');
201 %title('直方图示例');
202
203 set(gca,'FontSize',20)
204
205
206 count_le12food = sum(zhibiao1 <= 12);  % 统计 <=12种食物 的数
    量
207 count_g12food = sum(zhibiao1 >12);      % 统计 >12种食物 的数
    量
208
209
210 % 绘制饼状图
211 figure;
212 piechart([count_le12food,count_g12food], {'不足12种', '大于12
    种'}); % 创建饼状图并标记图例
213 %title('每天摄入的食物不足12种和大于12种的比例'); % 添加标题
214 set(gca,'FontSize',20)
215
216 %%
217
218
219 %指标2:每周摄入25种以上
220 figure
```

```
% 绘制直方图
h = histogram(zhibiao2, 'Normalization', 'count');


% 为直方图添加标签（可选）
xlabel('每周摄入的食物种类计数');
ylabel('频数');
%title('直方图示例');

set(gca,'FontSize',20)



count_le25food = sum(zhibiao2 <= 25);  % 统计 <=12种食物 的数量
count_g25food = sum(zhibiao2 >25);      % 统计 >12种食物 的数量


% 绘制饼状图
figure;
piechart([count_le25food,count_g25food], {'不足25种', '大于25种'}); % 创建饼状图并标记图例
%title('每周摄入的食物不足25种和大于25种的比例'); % 添加标题
set(gca,'FontSize',20)

%%
%指标3:保证每天摄入不少于300g的新鲜蔬菜
figure
% 绘制直方图
h = histogram(zhibiao3, 'Normalization', 'count');


% 为直方图添加标签（可选）
xlabel('每天摄入的新鲜蔬菜重量（g）');
```

```
ylabel('频数');
%title('每天摄入的新鲜蔬菜重量（g）');

set(gca,'FontSize',20)


count_le300vega = sum(zhibiao3 <= 300);  % 统计 <=12种食物 的数量
count_g300vega = sum(zhibiao3 >300);      % 统计 >12种食物 的数量


% 绘制饼状图
figure;
piechart([count_le300vega,count_g300vega], {'不足300g', '大于300g'}); % 创建饼状图并标记图例
%title('每天摄入的新鲜蔬菜重量不足300g和大于300g的比例'); % 添加标题
set(gca,'FontSize',20)


%%
%指标4:保证每天摄入不少于200g的新鲜水果
figure
% 绘制直方图
h = histogram(zhibiao4, 'Normalization', 'count');


% 为直方图添加标签（可选）
xlabel('每天摄入的新鲜水果重量（g）');
ylabel('频数');
%title('每天摄入的新鲜蔬菜重量（g）');

set(gca,'FontSize',20)

```

```
284
285 count_le200fruit = sum(zhibiao4 <= 200);  % 统计 <=12种食物 的
    数量
286 count_g200fruit = sum(zhibiao4 >200);      % 统计 >12种食物 的
    数量
287
288
289 % 绘制饼状图
290 figure;
291 piechart([count_le200fruit,count_g200fruit], {'不足200g', '大
    于200g'}); % 创建饼状图并标记图例
292 %title('每天摄入的新鲜蔬菜重量不足300g和大于300g的比例'); % 添
    加标题
293 set(gca,'FontSize',20)
294
295 %%
296 %指标5: 保证奶制品200~500
297 figure
298 % 绘制直方图
299 h = histogram(zhibiao5, 'Normalization', 'count');
300
301
302 % 为直方图添加标签（可选）
303 xlabel('每天摄入的奶制品重量（g）');
304 ylabel('频数');
305 %title('每天摄入的新鲜蔬菜重量（g）');
306
307 set(gca,'FontSize',20)
308
309
310 count_le200milk = sum(zhibiao5 <= 200);  % 统计 <=200 的数量
311 count_g200le500milk = sum((zhibiao5 >200) &( zhibiao5 <500));
        % 统计 200~500 的数量
312 count_g500milk = sum( zhibiao5 >500);      % 统计 >500 的数量
313
```

```
% 绘制饼状图
figure;
piechart([count_le200milk,count_g200le500milk,count_g500milk],
    {'不足200g', '大于200g小于500g','大于500g'}); % 创建饼状图
   并标记图例
%title('每天摄入的新鲜蔬菜重量不足300g和大于300g的比例'); % 添
   加标题
set(gca,'FontSize',20)


%%
%指标6: 保证肉蛋奶120~200
figure
% 绘制直方图
h = histogram(zhibiao6, 'Normalization', 'count');


% 为直方图添加标签（可选）
xlabel('每天摄入的鱼蛋肉重量（g）');
ylabel('频数');
%title('每天摄入的新鲜蔬菜重量（g）');

set(gca,'FontSize',20)


count_le120meat = sum(zhibiao6 <= 120);  % 统计 <=200 的数量
count_g120le200meat = sum((zhibiao6 >120) &( zhibiao6 <200));
        % 统计 200~500 的数量
count_g200meat = sum( zhibiao6 >200);      % 统计 >500 的数量

% 绘制饼状图
figure;
piechart([count_le120meat,count_g120le200meat,count_g200meat],
    {'不足120g', '大于120g小于200g','大于200g'}); % 创建饼状图
   并标记图例
```

```
%title('每天摄入的新鲜蔬菜重量不足300g和大于300g的比例'); % 添加标题
set(gca,'FontSize',20)

%%
%指标7:成人每天摄入烹调油25~30g
%将三餐就餐的平均人数四舍五入作为其家庭成员人数

figure
% 绘制直方图
h = histogram(zhibiao7, 'Normalization', 'count');


% 为直方图添加标签（可选）
xlabel('每天摄入的烹调油重量（g）');
ylabel('频数');
%title('每天摄入的新鲜蔬菜重量（g）');

set(gca,'FontSize',20)


count_le25oil = sum(zhibiao7 <= 25);  % 统计 <=200 的数量
count_g25le30oil = sum((zhibiao7 >25) &( zhibiao7 <30));  % 统计 200~500 的数量
count_g30oil = sum( zhibiao7 >30);      % 统计 >500 的数量

% 绘制饼状图
figure;
piechart([count_le25oil,count_g25le30oil,count_g30oil], {'不足25g', '大于25g小于30g','大于30g'}); % 创建饼状图并标记图例
%title('每天摄入的新鲜蔬菜重量不足300g和大于300g的比例'); % 添加标题
set(gca,'FontSize',20)

% zhibiao7 = (植物油和动物油)
```

```
374
375 %%
376 %指标8:食用盐＜5g
377 figure
378 % 绘制直方图
379 h = histogram(zhibiao8, 'Normalization', 'count');
380
381
382 % 为直方图添加标签（可选）
383 xlabel('每天摄入的食用盐重量（g）');
384 ylabel('频数');
385 %title('每天摄入的新鲜蔬菜重量（g）');
386
387 set(gca,'FontSize',20)
388
389
390 count_le5salt = sum(zhibiao8 <= 5);  % 统计 <=200 的数量
391 %count_g120le200meat = sum((zhibiao8 >120) &( zhibiao6 <200));
        % 统计 200~500 的数量
392 count_g5salt = sum( zhibiao8 >5);      % 统计 >500 的数量
393
394 % 绘制饼状图
395 figure;
396 piechart([count_le5salt,count_g5salt], {'不足5g', '大于5g'});
   % 创建饼状图并标记图例
397 %title('每天摄入的新鲜蔬菜重量不足300g和大于300g的比例'); % 添
   加标题
398 set(gca,'FontSize',20)
399
400 %%
401 %指标9:饮食规律 吃早餐
402 figure
403 % 绘制直方图
404 h = histogram(zhibiao9, 'Normalization', 'count');
405
```

```
406
407 % 为直方图添加标签（可选）
408 xlabel('每周不吃早餐的次数');
409 ylabel('频数');
410 %title('每天摄入的新鲜蔬菜重量（g）');
411
412 set(gca,'FontSize',20)
413 count_breakfast = sum(zhibiao9 ==0);  % 统计 <=200 的数量
414 %count_g120le200meat = sum((zhibiao8 >120) &( zhibiao6 <200));
          % 统计 200~500 的数量
415 count_nobreakfast = sum( zhibiao9 >0);      % 统计 >500 的数量
416
417 % 绘制饼状图
418 figure;
419 piechart([count_breakfast,count_nobreakfast], {'吃早餐', '不吃
    早餐'}); % 创建饼状图并标记图例
420 %title('每天摄入的新鲜蔬菜重量不足300g和大于300g的比例'); % 添
    加标题
421 set(gca,'FontSize',20)
422
423
424 %%
425 %指标9:饮食规律 吃早餐
426 figure
427 % 绘制直方图
428 h = histogram(zhibiao9, 'Normalization', 'count');
429
430
431 % 为直方图添加标签（可选）
432 xlabel('每周不吃早餐的次数');
433 ylabel('频数');
434 %title('每天摄入的新鲜蔬菜重量（g）');
435
436 set(gca,'FontSize',20)
437 count_breakfast = sum(zhibiao9 ==0);  % 统计 <=200 的数量
```

```
%count_g120le200meat = sum((zhibiao8 >120) &( zhibiao6 <200));
        % 统计 200~500 的数量
count_nobreakfast = sum( zhibiao9 >0);      % 统计 >500 的数量

% 绘制饼状图
figure;
piechart([count_breakfast,count_nobreakfast], {'吃早餐', '不吃早餐'}); % 创建饼状图并标记图例
%title('每天摄入的新鲜蔬菜重量不足300g和大于300g的比例'); % 添加标题
set(gca,'FontSize',20)


%%
% 吸烟
count_smoke = sum(zhibiao10 == 1);  % 统计 吸烟 的数量
count_unsmoke = sum(zhibiao10 == 3);      % 统计 不吸烟 的数量
count_quitsmoke = sum(zhibiao10 == 2);      % 统计 不吸烟 的数量

% 绘制饼状图
figure;
piechart([count_smoke,count_unsmoke,count_quitsmoke], {'吸烟', '不吸烟','戒烟'}); % 创建饼状图并标记图例
%title('吸烟，不吸烟，戒烟所占比例'); % 添加标题
set(gca,'FontSize',20)

%%
% 指标11:喝酒
count_alcohol = sum(zhibiao11 == 1);  % 统计 喝酒 的数量
count_unalcohol = sum(zhibiao11 == 2);      % 统计 不喝酒 的数量

% 绘制饼状图
figure;
```

```
piechart([count_alcohol,count_unalcohol], {'喝酒', '不喝酒'});
    % 创建饼状图并标记图例
%title('喝酒和不喝酒所占比例'); % 添加标题
set(gca,'FontSize',20)


%%
%体育锻炼
figure
% 绘制直方图
h = histogram(data_pro4.exercise, 'Normalization', 'count');


% 为直方图添加标签（可选）
xlabel('每周的体育锻炼量');
ylabel('频数');
%title('每天摄入的新鲜蔬菜重量（g）');

set(gca,'FontSize',20)
count_exercise_meet = sum(data_pro4.exercise >150);  % 统计
   <=200 的数量
%count_g120le200meat = sum((zhibiao8 >120) &( zhibiao6 <200));
         % 统计 200~500 的数量
count_exercise_notmeet = sum( data_pro4.exercise <150);      %
    统计 >500 的数量

% 绘制饼状图
figure;
piechart([count_exercise_meet,count_exercise_notmeet], {'体育
   锻炼量达标', '体育锻炼量不达标'}); % 创建饼状图并标记图例
%title('每天摄入的新鲜蔬菜重量不足300g和大于300g的比例'); % 添
   加标题
set(gca,'FontSize',20)

%%
```

```
%家务
lifestyle =  readtable("raw_data.xlsx",'Sheet','生活习惯');
figure
% 绘制直方图
h = histogram(lifestyle.homework, 'Normalization', 'count');


% 为直方图添加标签（可选）
xlabel('家务类型');
ylabel('频数');
%title('每天摄入的新鲜蔬菜重量（g）');

set(gca,'FontSize',20)
count_home1 = sum(lifestyle.homework == 1);  % 统计 <=200 的数量

count_home2 = sum(lifestyle.homework == 2);  % 统计 <=200 的数量
count_home3 = sum(lifestyle.homework == 3);  % 统计 <=200 的数量
%count_g120le200meat = sum((zhibiao8 >120) &( zhibiao6 <200));
        % 统计 200~500 的数量
%count_exercise_notmeet = sum( data_pro4.exercise <150);
   % 统计 >500 的数量

% 绘制饼状图
figure;
piechart([count_home1 ,count_home2 ,count_home3 ], {'轻度', '中度','重度'}); % 创建饼状图并标记图例
%title('每天摄入的新鲜蔬菜重量不足300g和大于300g的比例'); % 添加标题
set(gca,'FontSize',20)

%% 1.对各项指标进行正向化处理
zhibiao_raw = [zhibiao1 zhibiao2 zhibiao3 zhibiao4 zhibiao5
```

```
   zhibiao6 zhibiao7 zhibiao8 zhibiao9  zhibiao11];
zhibiao_raw = [zhibiao_raw; 13 26 300 200 200 120 25 2.5 1 2];
row = row+1;
%指标3、4、6、7、9属于区间型
%
% zhibiao3e = Inter2Max(zhibiao3,300,500);
% zhibiao4e = Inter2Max(zhibiao4,200,350);
% zhibiao5e = Inter2Max(zhibiao5,300,500);
% zhibiao6e = Inter2Max(zhibiao6,120,200);
% zhibiao7e = Inter2Max(zhibiao7,25,30);

zhibiao3e = Inter2Max(zhibiao_raw(:,3),300,1000);
zhibiao4e = Inter2Max(zhibiao_raw(:,4),200,1000);
zhibiao5e = Inter2Max(zhibiao_raw(:,5),300,500);
zhibiao6e = Inter2Max(zhibiao_raw(:,6),120,200);
zhibiao7e = Inter2Max(zhibiao_raw(:,7),25,30);
%指标1 2当做极大型处理
zhibiao1e = normalize(zhibiao_raw(:,1), 'range');
zhibiao2e = normalize(zhibiao_raw(:,2), 'range');
zhibiao9e = normalize(-zhibiao_raw(:,9), 'range');
%zhibiao10e = normalize(zhibiao10, 'range');
zhibiao11e = normalize(zhibiao_raw(:,10), 'range');

%指标8特殊处理
zhibiao8e = teshuchuli(zhibiao_raw(:,8),5);
%2.归一化处理
zhibiao = [zhibiao1e zhibiao2e zhibiao3e zhibiao4e zhibiao5e
   zhibiao6e zhibiao7e zhibiao8e zhibiao9e  zhibiao11e];
%zhibiao = zhibiao ./ repmat(sum(zhibiao.*zhibiao) .^ 0.5, row
   , 1);
%3.利用熵权法计算各项指标的权重
W = Entropy_Method(zhibiao);
%4.计算最终得分
zuidajuli = sum([(zhibiao - repmat(max(zhibiao),row,1)) .^ 2 ]
    .* repmat(W,row,1) ,2) .^ 0.5;   % D+ 与最大值的距离向量
```

```
zuixiaojuli = sum([(zhibiao - repmat(min(zhibiao),row,1)) .^ 2 ] .* repmat(W,row,1) ,2) .^ 0.5;   % D- 与最小值的距离向量
score = zuixiaojuli ./ (zuidajuli+zuixiaojuli);    % 未归一化的得分
%%
figure
% 绘制直方图
h = histogram(score(2:end), 'Normalization', 'count');


% 为直方图添加标签（可选）
xlabel('饮食习惯的合理性评分');
ylabel('频数');
%title('饮食习惯的合理性评分');

set(gca,'FontSize',20)
count_reasonable = sum(score >= 0.7428);  % 统计 <=200 的数量
%count_g120le200meat = sum((zhibiao8 >120) &( zhibiao6 <200));        % 统计 200~500 的数量
count_unreasonable = sum(score < 0.7428);      % 统计 >500 的数量

% 绘制饼状图
figure;
piechart([count_reasonable,count_unreasonable], {'合理', '不合理'}); % 创建饼状图并标记图例
%title('每天摄入的新鲜蔬菜重量不足300g和大于300g的比例'); % 添加标题
set(gca,'FontSize',20)

%5.将分数按照从高到低的顺序进行排序
%[score,index] = sort(score,'descend');
%去除加入的辅助行
%new1 = [new1(:,1:14)  new1(:,16:32) new1(:,34:196) new1(:,198:col)];
```

pro2.m

```
1
2  age = 2020*ones(length(new1),1) - new1(:,2);
3  new1(:,2) = age;
4  % 性别列和饮酒列
5  gender = new1(:,3);
6
7  %%
8  data = [gender zhibiao11];
9  % 计算男女饮酒和不饮酒人数
10 male_drink = sum(data(data(:, 1) == 1 & data(:, 2) == 1, 1));
11 male_no_drink = sum(data(data(:, 1) == 1 & data(:, 2) == 2, 1)
       );
12 female_drink = sum(data(data(:, 1) == 2 & data(:, 2) == 1, 1))
       ;
13 female_no_drink = sum(data(data(:, 1) == 2 & data(:, 2) == 2,
       1));
14
15 % 计算比例
16 male_total = male_drink + male_no_drink;
17 female_total = female_drink + female_no_drink;
18
19 male_ratio = [male_drink, male_no_drink] / male_total;
20 female_ratio = [female_drink, female_no_drink] / female_total;
21
22 % 绘制饼状图
23 figure;
24
25 subplot(1, 2, 1); % 创建1行2列的图，当前为第1个子图
26 piechart(male_ratio, {'饮酒', '不饮酒'});
27 title('男性饮酒比例');
28 set(gca,'FontSize',20)
29 subplot(1, 2, 2); % 当前为第2个子图
30 piechart(female_ratio, {'饮酒', '不饮酒'});
31 title('女性饮酒比例');
```

```
32
33  % 调整图形
34  %sgtitle('饮酒与性别的关系'); % 添加总标题
35  set(gca,'FontSize',20)
36
37  %%
38
39  %生活习惯和饮食习惯是否与年龄、性别、婚姻状况、文化程度、职业
        等因素相关
40  %clear;clc
41  %url = 'raw_data.xlsx';
42  text = readtable('raw_data.xlsx','Sheet','生活习惯');
43  text.x___ = -text.x___;
44  %对生活习惯数据进行评分
45  %1.对数据进行正向化
46  %其中1 3 5为极小型，2 4 6为极大型
47
48  %2.归一化处理
49  lifestyle = normalize(table2array(text), 'range');
50  row = length(lifestyle);
51  %3.利用熵权法计算各项指标的权重
52  W = Entropy_Method(lifestyle);
53  %4.计算最终得分
54  zuidajuli = sum([(lifestyle - repmat(max(lifestyle),row,1)) .^
        2 ] .* repmat(W,row,1) ,2) .^ 0.5;   % D+ 与最大值的距离向
        量
55  zuixiaojuli = sum([(lifestyle - repmat(min(lifestyle),row,1))
        .^ 2 ] .* repmat(W,row,1) ,2) .^ 0.5;   % D- 与最小值的距离
        向量
56  score_life = zuixiaojuli ./ (zuidajuli+zuixiaojuli);
57  %% 判断相关性
58  data = readtable('raw_data.xlsx','Sheet','pro2');  % 从CSV文件
        中读取数据
59
60  % 将分类变量转换为哑变量
```

```
data.gender = dummyvar(categorical(data.gender));
data.education = dummyvar(categorical(data.education));
data.marriage = dummyvar(categorical(data.marriage));
data.career = dummyvar(categorical(data.career));

% 提取连续变量和哑变量
X = [data.age, data.gender, data.education, data.marriage,
   data.career];
y_lifestyle = data.lifestyle;
y_dietary = data.dietary;


%%
% 计算 dietary 与哑变量的相关性
[correlations_dietary,p_dietary] = corr([y_dietary,X]);
disp('Correlation with Dietary:');
disp(correlations_dietary);

% 计算 lifestyle 与哑变量的相关性
[correlations_lifestyle,p_lifestyle] = corr([y_lifestyle,X]);
disp('Correlation with Lifestyle:');
disp(correlations_lifestyle);



%%
% 线性回归分析生活习惯
%生活习惯评分与个人因素的相关性分析
mdl_lifestyle = fitlm(X, y_lifestyle)
figure
plot(mdl_lifestyle.Coefficients.pValue,'*',LineWidth=1.5)
hold on
a = 0.05*ones(1,28);
plot(a,'r-',LineWidth=1.5)
legend('p值','阈值')
```

```
xlabel('哑变量')
ylabel('相关性p值')
set(gca,'FontSize',20)
% disp(mdl_lifestyle);

% 线性回归分析饮食习惯
mdl_dietary = fitlm(X, y_dietary);

figure
plot(mdl_dietary.Coefficients.pValue,'*',LineWidth=1.5)
hold on
a = 0.05*ones(1,28);
plot(a,'r-',LineWidth=1.5)
legend('p值','阈值')
xlabel('哑变量')
ylabel('相关性p值')
set(gca,'FontSize',20)

% 计算相关系数矩阵
[corr_lifestyle,p_lifestyle] = corr(data{:, {'age', 'lifestyle
    '}});
[corr_dietary,p_dietary] = corr(data{:, {'age', 'dietary'}});
disp('生活习惯评分和年龄的相关系数及p值：');
disp(corr_lifestyle(1,2));
disp(p_lifestyle(1,2));
disp('饮食习惯评分和年龄的相关系数及p值：');
disp(corr_dietary(1,2));
disp(p_dietary(1,2));

%%
%对数据进行斯皮尔曼相关系数
%将得出的数据保存到excel表格中并读取数据
text = xlsread(url,4);
text(:,1) = data_conversion(text(:,1));
text(:,2) = data_conversion(text(:,2));
```

```
[R,P]=corr(text, 'type' , 'Spearman');
%显著性水平判断
a = P < 0.01  % 标记3颗星的位置
b = (P < 0.05) .* (P > 0.01)  % 标记2颗星的位置
c = (P < 0.1) .* (P > 0.05) % % 标记1颗星的位置
```

pro3.m

```
drink = new1(:,19).*new1(:,20) + 0.5*new1(:,22).*new1(:,23)
   ...
    + 0.3*new1(:,25).*new1(:,26) +0.1*(new1(:,28).*new1(:,29)
   + new1(:,31).*new1(:,32));
%%
data_pro3 = readtable('raw_data.xlsx','Sheet','pro3');

rows_to_delete = find((data_pro3.diabetes == 0)|(data_pro3.
   hypertension == 0));

% 删除这些行
data_pro3(rows_to_delete, :) = [];


data_pro3.career = dummyvar(categorical(data_pro3.career));
%%
% 将hypertension列转换为逻辑变量（1: 有, 0: 没）
data_pro3.hypertension = data_pro3.hypertension - 1;  % 将2变
   为1，将1变为0
data_pro3.diabetes = data_pro3.diabetes - 1;  % 将2变为1，将1
   变为0

%%
X_pro3 = [data_pro3.dietary,data_pro3.lifestyle,data_pro3.
   smoke,data_pro3.exercise,data_pro3.drink,data_pro3.career];
% 逻辑回归模型
%mdl = fitglm(data_pro3, 'hypertension ~ dietary + lifestyle +
    smoke + exercise + drink + career', 'Distribution', '
```

```
   binomial');
mdl_h = fitglm(X_pro3,data_pro3.hypertension, 'Distribution',
   'binomial');
% 显示模型分析结果
disp(mdl);
figure
plot(mdl_h.Coefficients.pValue,'*',LineWidth=1.5)
hold on
a = 0.05*ones(1,18);
plot(a,'r-',LineWidth=1.5)
legend('p值','阈值')
xlabel('哑变量')
ylabel('相关性p值')
set(gca,'FontSize',20)
%%
mdl_d = fitglm(X_pro3,data_pro3.diabetes, 'Distribution', '
   binomial');
% 显示模型分析结果
disp(mdl);
figure
plot(mdl_d.Coefficients.pValue,'*',LineWidth=1.5)
hold on
a = 0.05*ones(1,19);
plot(a,'r-',LineWidth=1.5)
legend('p值','阈值')
xlabel('哑变量')
ylabel('相关性p值')
set(gca,'FontSize',20)


%%

%%
clear;clc
%导入饮酒数据新型处理
```

```
% data_title = ["是否饮酒","饮酒年数","高度每周白酒饮用量","低度每周白酒饮用量","啤酒每周饮用量","黄酒、糯米酒每周饮用量","葡萄每周酒引用量"];
data_title = ["是否饮酒","饮酒年数","每周饮酒量"]
url = 'D:\Desktop\饮酒.xlsx';
data = xlsread(url);
[row,col] = size(data);
Drinking = zeros(row,1);
for i=1:row
    count = 1;
    sum = 0;
    if data(i,2) == 99
        data(i,2) = mean(data(:,2));
    end
    for j=4:3:col
        sum = sum + data(i,j)*data(i,j+1)*50;
        count=count+1;
    end
    Drinking(i) = sum;
end
a = [data(:,1) data(:,2) Drinking];
xlswrite('D:\Desktop\饮酒处理后数据.xlsx',[data_title;[data(:,1) data(:,2) Drinking]])
clear;clc
[row,col] = size(data);
score = zeros(row,1);
for i = 1:row
    if data(i,1) == 1
        score(i) = 0;
    elseif data(i,1) == 2
        score(i) = 1.5*data(i,2);
    elseif data(i,1) ==3
        score(i) = 4*data(i,2);
    elseif data(i,1) == 4
        score(i) = 6*data(i,2);
```

```
    end
end
```

pro4.m

```
%对居民进行合理分类，并针对各类人群提出有利于身体健康的膳食、运动等方面的合理建议。

% 创建数据表
data_pro4 = readtable("raw_data.xlsx",'Sheet','pro4');

% 显示原始数据

% 定义年龄段
ageGroups = cell(size(data_pro4.age)); % 创建一个单元格数组以存储年龄段

for i = 1:height(data_pro4)
    if data_pro4.age(i) < 35
        ageGroups{i} = '青年';
    elseif data_pro4.age(i) < 60
        ageGroups{i} = '中年';
    else
        ageGroups{i} = '老年';
    end
end

% 将年龄段添加到数据表中
data_pro4.ageGroup = categorical(ageGroups);

%%
% 计算各年龄段的平均吸烟量
averageSmoke_hyper = groupsummary(data_pro4, 'hypertension', 'mean', 'smoke');

% 显示平均吸烟量
```

```
disp('高血压患者和正常人的平均吸烟量：');
disp(averageSmoke_hyper );

% 绘制条形图
figure;
bar(categorical(averageSmoke_hyper.hypertension),
   averageSmoke_hyper.mean_smoke, 'FaceColor', [0 0 1]);
xlabel('高血压患者和正常人');
ylabel('平均吸烟量');
%title('各年龄段的平均吸烟量');
grid on;
set(gca,'FontSize',20)


%%
% 计算各年龄段的平均吸烟量
averageExercise_hyper = groupsummary(data_pro4, 'hypertension'
   , 'mean', 'exercise');

% 显示平均吸烟量
disp('高血压患者和正常人的平均运动量：');
disp(averageExercise_hyper );

% 绘制条形图
figure;
bar(categorical(averageExercise_hyper.hypertension),
   averageExercise_hyper.mean_exercise, 'FaceColor', [0 0 1]);
xlabel('高血压患者和正常人');
ylabel('平均运动量');
%title('各年龄段的平均吸烟量');
grid on;
set(gca,'FontSize',20)


%%
```

```
% 计算各年龄段的平均吸烟量
averageSmoke = groupsummary(data_pro4, 'ageGroup', 'mean', '
   smoke');

% 显示平均吸烟量
disp('各年龄段的平均吸烟量：');
disp(averageSmoke);

% 绘制条形图
figure;
bar(categorical(averageSmoke.ageGroup), averageSmoke.
   mean_smoke, 'FaceColor', [0 0.5 0.5]);
xlabel('年龄段');
ylabel('平均吸烟量');
%title('各年龄段的平均吸烟量');
grid on;
set(gca,'FontSize',20)
%%
% 计算各年龄段的平均吸烟量
averageExercise = groupsummary(data_pro4, 'ageGroup', 'mean',
   'exercise');

% 显示平均吸烟量
disp('各年龄段的平均运动量：');
disp(averageExercise);

% 绘制条形图
figure;
bar(categorical(averageExercise.ageGroup), averageExercise.
   mean_exercise, 'FaceColor', [0 0.5 0.5]);
xlabel('年龄段');
ylabel('平均运动量');
%title('各年龄段的平均吸烟量');
grid on;
set(gca,'FontSize',20)
```

```
92
93  %%
94  % 计算各年龄段的平均吸烟量
95  averageDrink = groupsummary(data_pro4, 'ageGroup', 'mean', '
       drink');
96
97  % 显示平均吸烟量
98  disp('各年龄段的平均饮酒量：');
99  disp(averageDrink);
100
101 % 绘制条形图
102 figure;
103 bar(categorical(averageDrink.ageGroup), averageDrink.
       mean_drink, 'FaceColor', [0 0.5 0.5]);
104 xlabel('年龄段');
105 ylabel('平均饮酒量');
106 %title('各年龄段的平均吸烟量');
107 grid on;
108 set(gca,'FontSize',20)
109
110 %%
111 % 计算各年龄段的平均吸烟量
112 averageBreakfast = groupsummary(data_pro4, 'ageGroup', 'mean',
        'breakfast');
113
114 % 显示平均吸烟量
115 disp('各年龄段的平均不吃早餐次数：');
116 disp(averageBreakfast);
117
118 % 绘制条形图
119 figure;
120 bar(categorical(averageBreakfast.ageGroup), averageBreakfast.
       mean_breakfast, 'FaceColor', [0 0.5 0.5]);
121 xlabel('年龄段');
122 ylabel('平均不吃早餐次数');
```

```
%title('各年龄段的平均吸烟量');
grid on;
set(gca,'FontSize',20)
%%
% 计算各年龄段的平均吸烟量
averageDietary = groupsummary(data_pro4, 'ageGroup', 'mean', '
    dietary');

% 显示平均吸烟量
disp('各年龄段的平均饮食习惯评分：');
disp(averageDietary);

% 绘制条形图
figure;
bar(categorical(averageDietary.ageGroup), averageDietary.
    mean_dietary, 'FaceColor', [0 0.5 0.5]);
xlabel('年龄段');
ylabel('平均平均饮食习惯评分');
%title('各年龄段的平均吸烟量');
grid on;
set(gca,'FontSize',20)

%%

% 计算各年龄段的平均吸烟量
averageLifestyle = groupsummary(data_pro4, 'ageGroup', 'mean',
     'lifestyle');

% 显示平均吸烟量
disp('各年龄段的平均生活习惯评分：');
disp(averageLifestyle);

% 绘制条形图
figure;
bar(categorical(averageLifestyle.ageGroup), averageLifestyle.
```

```
   mean_lifestyle, 'FaceColor', [0 0.5 0.5]);
xlabel('年龄段');
ylabel('平均平均生活习惯评分');
%title('各年龄段的平均吸烟量');
grid on;
set(gca,'FontSize',20)
%%
% 计算各年龄段的平均吸烟量
averageHypertension = groupsummary(data_pro4, 'ageGroup', '
   mean', 'hypertension');

% 显示平均吸烟量
disp('各年龄段的高血压比例: ');
disp(averageHypertension);

% 绘制条形图
figure;
bar(categorical(averageHypertension.ageGroup), (1-
   averageHypertension.mean_hypertension), 'FaceColor', [0 0.5
   0.5]);
xlabel('年龄段');
ylabel('高血压比例');
%title('各年龄段的平均吸烟量');
grid on;
set(gca,'FontSize',20)

%%
% 计算各年龄段的平均吸烟量
averageDiabetes = groupsummary(data_pro4, 'ageGroup', 'mean',
   'diabetes');

% 显示平均吸烟量
disp('各年龄段的糖尿病比例: ');
disp(averageDiabetes);

```

```
% 绘制条形图
figure;
bar(categorical(averageDiabetes.ageGroup), (1-averageDiabetes.
    mean_diabetes), 'FaceColor', [0 0.5 0.5]);
xlabel('年龄段');
ylabel('糖尿病比例');
%title('各年龄段的平均吸烟量');
grid on;
set(gca,'FontSize',20)
%%
writetable(data_pro3,'data_pro3.xlsx' );
writetable(data,'data_pro2.xlsx' );
%%

%读入数据
url = "D:\Desktop\聚类分析.xlsx"
Youth_data = xlsread(url,4);
Middle_aged_data = xlsread(url,6);
Elderly_data = xlsread(url,8);
%对青年数据进行分析,青年存在运动性水平低和生活习惯水平偏低的问
    题
y_p = size(Youth_data,1);
m_p = size(Middle_aged_data,1);
e_p = size(Elderly_data,1);
%运动型水平
Athletic_level_avg = (sum(Middle_aged_data(:,5))+sum(
    Elderly_data(:,5)))/(m_p+e_p)
%生活习惯
Lifestyle_habits = (sum(Middle_aged_data(:,6))+sum(
    Elderly_data(:,6)))/(m_p+e_p)
%判断各种情况人群人数
[level1,level2,level3,level4]=Type_judgment(Youth_data,
    Athletic_level_avg,Lifestyle_habits,5,6)
%运动型水平和生活习惯水平正常
all_normal = level1/y_p;
```

```
%运动型水平低
abnormal1 = level2/y_p;
%生活习惯水平低
abnormal2 = level3/y_p;
%二者均低
all_abnormal = level4/y_p;
y_c = [level1,level2,level3,level4];
y_pp = [all_normal abnormal1 abnormal2 all_abnormal];

%对中年数据进行分析,中年人存在吸烟水平和饮酒水平水平偏高的问题
%饮酒水平
Drinking_level_avg = (sum(sum(Youth_data(:,1:2)))+sum(sum(
   Elderly_data(:,1:2))))/(y_p+e_p)
%吸烟水平
Smoking_level_avg = (sum(sum(Youth_data(:,3:4)))+sum(sum(
   Elderly_data(:,3:4))))/(y_p+e_p)
%判断各种情况人群人数
[level1,level2,level3,level4]=Type_judgment1(Middle_aged_data,
   Drinking_level_avg,Smoking_level_avg,1,2,3,4)
%吸烟和喝酒正常
all_normal = level1/m_p;
%喝酒偏高
abnormal1 = level2/m_p;
%吸烟偏高
abnormal2 = level3/m_p;
%吸烟和喝酒均偏高
allm_abnormal = level4/m_p;
m_c = [level1,level2,level3,level4];
m_pp = [all_normal abnormal1 abnormal2 all_abnormal];
%老年人数据分析,老年人血糖和血压偏
%获取 140 90对应的值
%data = get_data(130,80);
data = 0.80977;
[level1,level2,level3,level4]=Type_judgment2(Elderly_data,data
   ,7,15,12);
```

```
all_normal = level1/e_p;
%高血压
abnormal1 = level2/e_p;
%糖尿病
abnormal2 = level3/e_p;
%均患有
allm_abnormal = level4/e_p;
o_c = [level1,level2,level3,level4];
o_pp = [all_normal abnormal1 abnormal2 allm_abnormal];
```

附录 C　哑变量