基于典型相关性的城市居民健康影响因素研究

摘要

近年来，我国居民的慢性传染病患病率持续攀升，本文探究了患病率攀升的原因与生活习惯和饮食习惯等因素的关系，并依据《中国居民膳食指南》给出了居民降低患病率，保持身体健康的合理性建议。

针对问题一，题目要求根据附件分析居民饮食习惯的合理性，并说明存在的主要问题。本文首先对数据进行预处理，将缺失数据进行填充，并统一计量单位。在此基础上，对居民的用餐情况、食用食物的种类和各种食物所占比例进行了可视化分析，提出部分居民存在的饮食问题并给出了合理性建议。

针对问题二，题目要求分析居民的生活习惯和饮食习惯是否与年龄、性别、婚姻状况等因素相关。显而易见，这是一个两组变量间的相关性分析问题。因此，采用研究两组变量之间相关关系的多元统计方法——典型相关分析，使用 SPSS 软件求解，最后得出居民的生活习惯和饮食习惯与年龄等因素具有相关关系。

针对问题三，题目要求深入分析常见慢性病与吸烟、饮酒、饮食习惯、生活习惯等因素的关系以及相关程度。本文首先对数据进行预处理，将异常数据剔除，并对部分数据进行重现编码。在此基础上，分别建立糖尿病、高血压逻辑回归模型，采用 MATLAB 求解，得到了效果较好的模型，并得到了各自变量逻辑回归的系数和显著性，最后通过分析数据给出了预防和治疗糖尿病和高血压的合理性建议。

针对问题四，题目要求根据居民的具体情况对居民进行合理分类，并针对各类人群剔除有利于身体健康的合理建议。本文首先都数据进行预处理，将缺失数据剔除。在此基础上采用 k-means++聚类算法将居民进行分类，聚类后得到四个类别。其次，通过对四个类别进行特征分析，将四个类别定义为营养不良人群、肥胖人群、高血压人群和正常人群。最后，对四个人群分布提出合理性建议。

关键词：典型相关分析；逻辑回归；聚类算法

一、问题重述

1.1 问题背景

慢性非传染性疾病，如心脑血管疾病、糖尿病、恶性肿瘤以及慢性阻塞性肺病，已成为影响我国居民身体健康的重要问题。大量事实表明，身体健康状况受年龄、饮食、体育活动、职业等多方面因素的影响。因此，研究通过合理地安排膳食、适量的身体运动、践行健康的生活方式，可以达到促进身体健康的目的。

1.2 问题提出

问题一：参考附件 A3，分析附件 A2 中居民的饮食习惯的合理性，并说明存在的主要问题。

问题二：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。

问题三：根据附件 A2 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。

问题四：依据附件 A2 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、问题的分析

2.1 问题的整体分析

本题的所有问题都是建立在居民的生活习惯、饮食习惯和是否患有慢性病之间的关系之上。首先，我们需要分析居民饮食习惯的合理性。并在此基础上，探寻居民自身的各种因素是否与生活习惯和饮食习惯有关。其次，需要进一步深入分析常见慢性病与居民的生活习惯和饮食习惯的关系与关联程度。最后，需要对所有居民进行合理的分类，并针对各类人群剔除有利于身体健康的膳食、运动等方面的合理建议。

2.2 问题一的分析

问题一需要我们分析居民饮食习惯的合理性，并说明其中存在的主要问题。此问首先对附件 2 中的数据进行预处理，并分别对缺失值和计量单位进行填充和统一。此外，此问对居民的用餐情况、食用食物的种类和各种食物所占比例进行可视化分析，并提出针对居民饮食习惯的相关建议。

2.3 问题二的分析

问题二需要我们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。显而易见，这是一个两组变量间相关性分析问题。由于两组变量中含有多个指标，若采用简单的相关分析方法，只是考虑了单个 X 与单个 Y 之间

的相关性，而没有考虑 X、Y 变量组内部各变量间的相关性。因此，我们采用了研究两组变量之间相关关系的多元统计方法——典型相关分析，考虑两组变量的线性组合，并研究其相关系数。

2.4 问题三的分析

问题三需要我们分析高血压、糖尿病与居民各种自身因素的关系及相关程度。首先，此问对数据进行预处理操作，剔除了部分缺失数据并重新编码了部分数据。此外，建立逻辑回归模型来衡量关系与相关程度，并总结了各因素对糖尿病、高血压的影响程度。

2.5 问题四的分析

问题四需要我们对居民的健康状况进行分类，并针对各类人群提出相关的健康建议。首先我们以居民的生理指标作为判断健康状况的依据，剔除了少数缺失数据。其次，我们采用 k-means++ 聚类算法，根据居民的各项指标将其分为营养不良、肥胖、高血压、正常四类人群，并对各类人群提出相应的饮食、运动、生活习惯的建议。

三、模型的假设

1. 假设附件各指标数据都是可靠的，能够反映事实规律。
2. 假设居民的身体状况在短时间内不会发生突变。
3. 假设居民的生活规律在短时间内不会发生突变。
4. 假设中国居民膳食指南短时间内不会发生重大修订。

四、符号说明

符号	符号说明
D_i	附件 1 问卷对象的编号
Z_i^1	原始样本数据指标变量 X_i 标准化后的结果
Z_j^2	原始样本数据指标变量 Y_j 标准化后的结果
U_i	第一组变量中第 i 对标准化后的典型变量
V_i	第二组变量中第 i 对标准化后的典型变量
λ_i	第 i 对典型变量的系数
α_i, β_i	第 i 对典型相关变量系数矩阵的特征向量
u_{C_i}	样本 x_i 所在簇中心
$\|C_i\|$	簇 C_i 中样本数量
$\|u_i^j\|$	簇 C_i 中的第 j 个特征

五、模型的建立与求解

5.1 问题一

5.1.1 数据预处理

（1）结合附件 1 和附件 2，我们发现附件 2 表中含有一些异常缺失值，以下对各类异常缺失值进行判断与填充。

• 是否食用某食物

若附件 2 表中某位居民的“某食物的食用频率”和“某食物的平均每次食用量”存在数值，则相应的“是否食用某食物”的缺失值为 1，否则为 2。

• 某食物的食用频率

若附件 2 表中某位居民的“是否食用某食物”和“某食物的平均每次食用量”存在数值，则相应的“某食物的食用频率”的缺失值使用该食物所有居民的平均食用量进行填充，否则使用 0 进行填充。

• 某食物的平均每次食用量

若附件 2 表种某位居民的“是否食用某食物”和“某食物的食用频率”存在数值，则相应的“某食物的平均每次食用量”的缺失值使用该食物所有居民的平均每次食用量进行填充，否则使用 0 进行填充。

（2）为达到简化模型的目的，我们对“某食物的食用频率”进行了归一化处理，得到各居民某食物每天的食用频率，并根据某食物的平均每次食用量，得出了各居民某食物每天的食用量。

5.1.2 用餐情况分析

本文对附件 2 表中 D_1-D_3 中居民三餐的用餐情况进行分析，计算出居民三餐用餐与否的比例，如图 1 所示。

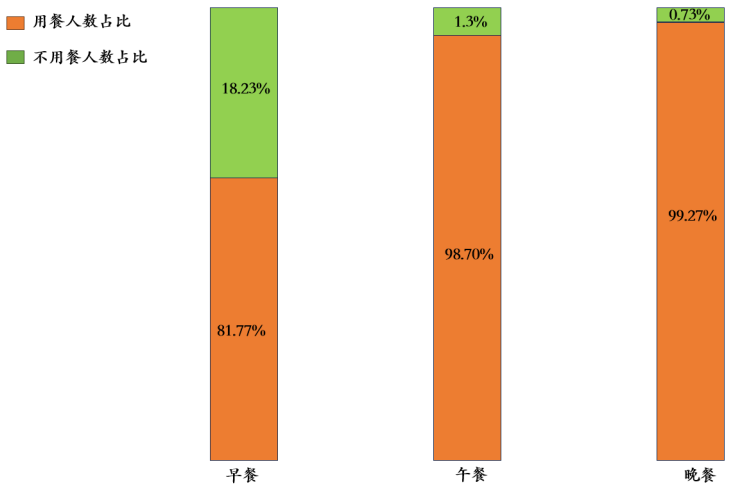

图 1　居民用餐情况分析

结合附件 3，做出以下分析：

考虑特殊情况，居民午餐和晚餐的用餐率符合膳食指南，而早餐的用餐率仅为81.77%，表明部分居民不重视早餐的食用，该现象不符合膳食指南的要求。

5.1.3 食用食物分析

对附件 2 表中 D_4-D_{30} 中居民是否食用该食物的情况进行分析。

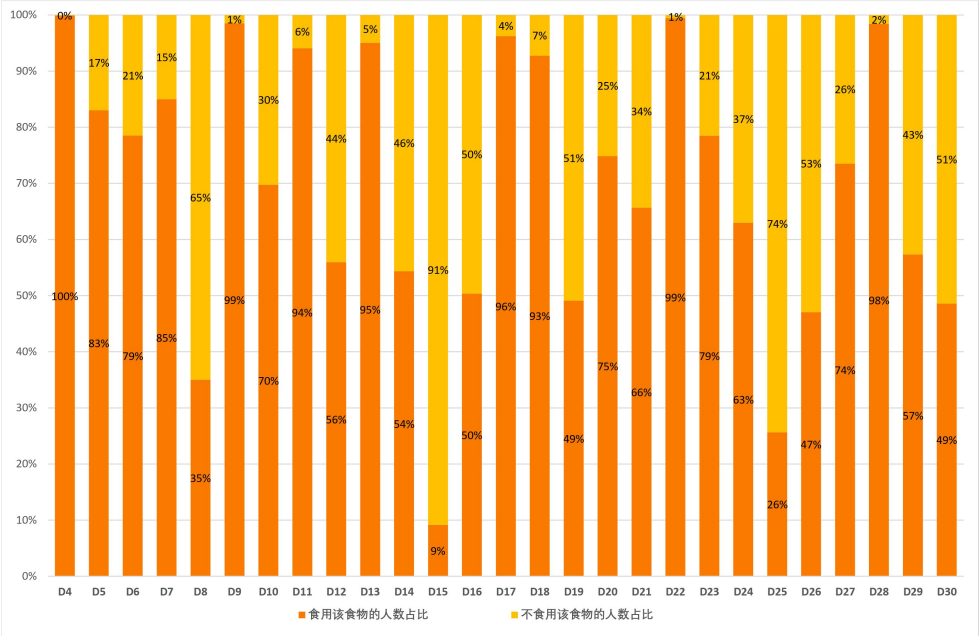

图 2　是否食用该食物分析

结合附件 3，做出以下分析：

（1）大米的食用率达到了 100%，小麦面粉的食用率达到了 83%，杂粮的食用率达到了 79%，满足以谷类为主、经常食用全谷物的膳食指南要求。

（2）各类奶制品的食用率均未超过 60%，表明居民未充分食用各类奶制品，不符合多食用各类奶制品的膳食指南的要求。

（3）果汁饮料和其他饮料的食用率均接近 50%，表明居民过度食用含糖饮料，不满足少食用各类含糖饮料得膳食指南的要求。

（4）豆类食物的食用率均低于 75%，不满足每天食用各类豆类食物的膳食指南得要求。

（5）水产品的食用率高于除猪肉外的肉质品，表明居民优先选择水产品食用，满足膳食指南的要求。

（6）新鲜蔬菜和水果的食用率分别为 99% 和 98%，满足多食用新鲜水果和蔬菜的膳食指南的要求。

（7）油炸食品的食用率较高，表明居民油的摄入含量过高，不满足膳食指南的要求。

（8）腌制食品的食用率较高，居民的盐摄入量过高，不满足膳食指南的要求。

5.1.4 各类食品具体食用情况分析

1. 为了结合附件 2 定量分析居民饮食习惯的合理性，从 D_4-D_{30} 提取了 7 类食物。

- 肉蛋类：猪肉、牛羊肉、鱼肉、禽肉、蛋肉和内脏；
- 奶制品：鲜奶、奶粉、酸奶；
- 豆类：豆腐、豆腐丝、千张、豆腐干、豆浆和干豆类；
- 蔬菜：新鲜蔬菜、海带和紫菜等海藻类；
- 腌制品：咸菜、泡菜和酸菜；
- 水果：新鲜水果；
- 饮料：果汁饮料和其他饮料；

对上述 7 类食物进行了描述性统计分析，如表 1 所示。

表 1　各类食物描述性统计分析

类别	最小值 (g)	最大值 (g)	均值 (g)	标准差
肉蛋类	0	577.143	174.111	89.019
奶制品	0	750.267	84.988	109.664
豆类	0	428.571	82.741	72.231
蔬菜	0	850.833	286.510	144.216
腌制品	0	75.000	7.977	10.696
水果	0	1250.000	144.552	123.624
饮料	0	1250.000	99.176	147.645

2. 我们从 D_4-D_{20} 提取了烹调油类和 5 种调味品类并对上述类别进行了描述性统计性分析，如表 2 所示。

表 2　烹调油和调味品描述性统计分析

类别	最小值 (g)	最大值 (g)	均值 (g)	标准差
盐	0	22.500	4.981	3.000
酱油	0	27.273	6.113	4.441
醋	0	13.636	2.097	2.526
酱类	0	6.364	0.526	1.050
味精	0	3.571	0.424	0.728
烹调油	0	129.629	40.227	19.820

结合附件 2 的膳食平衡要求，衡量出居民摄入部分食物、烹调油和调味品的情况，统计了居民在这些食品的缺乏率、合格率和超标率，如图 3 所示。

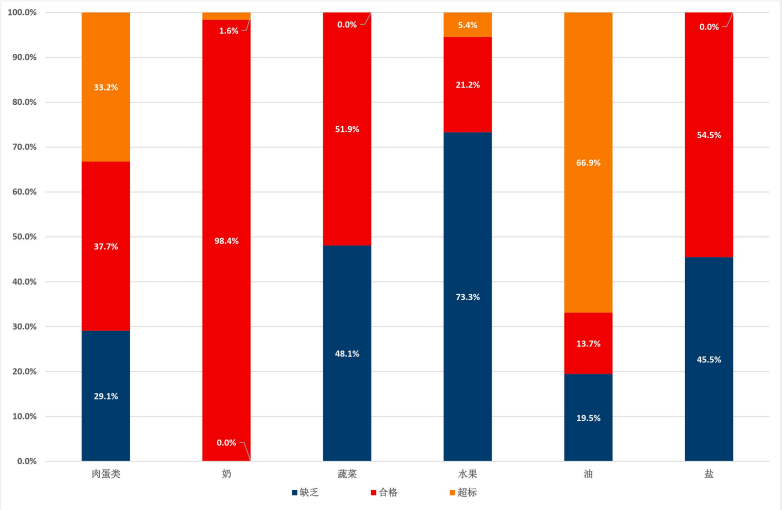

图 3　居民摄入部分食品情况

对此，我们做出以下分析：

（1）肉蛋类的合格率仅为 37.7%，缺乏率和超标率较高，表明居民肉蛋类的摄取不满足膳食平衡要求。

（2）奶制品的合格率接近 100%，表明居民奶制品的摄取量基本满足膳食平衡的要求。

（3）蔬菜的合格率接近 50%，表明居民的蔬菜摄取量不足，没有达到膳食平衡的要求。

（4）水果的合格率仅为 21.2%，表明居民水果摄取量没有达到膳食平衡的要求。

（5）烹调油的超标率达到了 66.9%，表明居民多喜欢食用多油食物，不满足膳食平衡的要求。

（6）盐的缺乏率接近 50%，表明居民盐的摄入量不足，不满足膳食平衡的要求。

5.1.5 对居民饮食习惯的建议

（1）居民应该按时就餐，不遗漏任何一餐，养成良好的三餐作息习惯。

（2）居民应该根据自身情况，增加或减少肉蛋类食物的摄入，满足膳食平衡的要求。

（3）居民应该增加摄入奶制品的种类，不能局限于几种奶制品。

（4）居民应该增加新鲜蔬菜和新鲜水果的食用，每日保证 300 克以上的摄入，以丰富膳食营养。

（5）居民应该减少多油食品的摄入，控制好摄入的油脂量。

（6）居民应该适当增加盐的摄入，满足膳食平衡的要求。

（7）居民应该减少含糖饮料的摄入，少喝或不喝糖含量高的饮料，不食用过量糖分的糕点。

5.2 问题二

问题二要求分析居民的生活习惯和饮食习惯与年龄、性别、婚姻状况、文化程度和职业的相关性。为了解决该问题，我们建立了典型相关分析模型[1]，并对各指标之间的关系进行分析。在此问题中，我们考虑休闲、家务活动以衡量居民的生活习惯。

5.2.1 典型相关分析模型的建立

典型相关分析模型的建立具体步骤如下：

令居民的个人信息年龄、性别、婚姻状况、文化程度和职业为第一组变量，居民的生活习惯和饮食习惯为第二组变量。

Step1. 建立原始矩阵；

根据附件 2，设居民的个人信息数据为 $X=(X_1,X_2,\cdots,Y_{7966})^T$，居民的生活习惯和饮食习惯数据为 $Y=(Y_1,Y_2,\cdots,X_{7966})^T$，Z 为总体观测矩阵。

$$Z=(X,Y)=\begin{bmatrix} X_{11} & X_{12} & \cdots & X_{16} & Y_{11} & Y_{12} & \cdots & Y_{1,14} \\ X_{21} & X_{22} & \cdots & X_{26} & Y_{21} & Y_{22} & \cdots & Y_{2,14} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ X_{7966,1} & X_{7966,2} & X_{7966,3} & \cdots & Y_{7966,1} & Y_{7966,2} & \cdots & Y_{7966,,14} \end{bmatrix} \tag{1}$$

Step2. 对原始数据进行标准化变换并计算相关系数矩阵；

对居民个人信息和居民生活习惯与饮食习惯的数据进行标准化处理，并计算样本相关系数矩阵 R，从而将 R 分为四个部分：

$$R=\begin{bmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{bmatrix} \tag{2}$$

其中，R_{11} 和 R_{12} 分别是居民个人信息指标间的相关系数和居民生活习惯与饮食习惯指标间的相关系数，R_{12} 和 R_{21} 是居民个人信息和居民生活习惯与饮食习惯指标间的相关系数矩阵。

Step3. 求典型相关系数及典型变量；

求 $A=R_{11}^{-1}R_{12}R_{22}^{-1}R_{21}$ 的特征根 λ_i^2 和特征向量 α_i；$B=R_{22}^{i}R_{21}R_{11}^{-1}R_{12}$ 的特征根 λ_i^2 和特征向量 β_i。

则居民个人信息 X 和居民生活习惯与饮食习惯之间的典型相关系数为 λ，典型变量为：

$$\begin{cases} U_i=\alpha_i^T X \\ V_i=\beta_i^T Y \end{cases} \tag{3}$$

Step4. 检验各典型相关系数的显著性。

最后对典型相关系数 λ_i 进行显著性检验。若 λ_i 不显著，则讨论该组典型变量的相关性无意义，λ_i 之后的典型相关变量也不显著。

5.2.2 典型相关分析模型的求解

采用 SPSS 软件进行典型相关分析：

表 3 典型相关系数表

	相关性	特征值	威尔克统计	F	分子自由度	分母自由度	p 值
1	0.442	0.243	0.659	39.133	84.000	42377.608	0.000
2	0.325	0.118	0.820	23.737	65.000	35935.320	0.000
3	0.264	0.075	0.917	13.906	48.000	29293.442	0.000
4	0.089	0.008	0.986	3.354	33.000	22406.433	0.000
5	0.066	0.004	0.993	2.510	20.000	15212.000	0.000
6	0.048	0.002	0.998	1.924	9.000	7607.000	0.044

如表 3 所示，P 值均小于 0.05，所以拒绝原假设，即认为两组变量之间存在相关性，且对典型变量的相关性均是显著的。

典型相关性的已解释方差比例如表 4 所示，根据方差解释比例，可以从定量角度的判断典型变量所包含的原始信息量，第一组变量和第二组变量的自身样本方差比例分布较均匀，故使用所有典型变量得到相应的典型相关性结果。

表 4 已解释的方差比例

典型变量	集合 1* 自身	集合 1* 集合 2	集合 2* 自身	集合 2* 集合 1
1	0.192	0.038	0.080	0.016
2	0.189	0.020	0.094	0.010
3	0.157	0.011	0.066	0.005
4	0.173	0.001	0.059	0.001
5	0.135	0.001	0.080	0.001
6	0.154	0.001	0.079	0.001

表 5 第一组变量标准化典型相关变量对应的线性组合系数

变量	1	2	3	4	5	6
年龄	-0.370	0.035	-0.979	0.108	-0.425	-0.159
性别	-0.918	-0.113	0.374	-0.130	-0.161	-0.202
民族	-0.058	-0.086	-0.027	-0.311	-0.361	0.875
文化程度	-0.160	-0.981	-0.375	-0.081	-0.040	-0.163
婚姻状况	-0.079	0.045	-0.016	-0.278	1.006	0.304
职业	-0.005	-0.194	0.009	0.887	0.214	0.435

表 6　第二组变量标准化典型相关变量对应的线性组合系数

变量	1	2	3	4	5	6
肉蛋类	0.490	0.039	0.051	-0.155	0.243	-0.015
豆类	-0.067	-0.257	-0.063	0.435	0.287	0.413
奶制品	-0.150	-0.496	0.388	0.009	-0.011	-0.305
蔬菜	-0.067	-0.050	-0.255	0.037	0.513	-0.105
腌制品	0.053	0.216	-0.104	0.263	-0.116	-0.179
水果	-0.230	-0.194	0.162	0.085	-0.375	0.024
饮料	0.424	-0.234	0.548	-0.179	0.063	0.025
盐	0.063	0.171	0.091	0.419	0.124	-0.014
酱油	-0.090	0.135	0.013	0.529	0.019	-0.075
醋	-0.108	-0.472	-0.505	-0.519	0.251	-0.224
酱类	0.024	-0.089	-0.066	0.086	-0.232	0.585
味精	0.039	-0.032	-0.088	-0.310	-0.314	0.483
烹调油	0.059	0.104	0.224	0.231	0.435	0.328
休闲、家务活动	-0.602	0.261	0.472	-0.352	0.217	0.124

表 5 和表 6 给出了两组变量标准化后对应的线性组合系数。此系数衡量了集合中各自指标的重要程度。各自以第一对变量为例，以下给出第一组变量和第二组变量的标准化典型变量，同理可得到剩余变量对的标准化典型变量。

$$U_1^* = -0.370Z_1^1 - 0.918Z_2^1 - 0.058Z_3^1 - 0.160Z_4^1 - 0.079Z_5^1 + 0.005Z_6^1 \tag{4}$$

$$\begin{aligned} V_1^* = {} & 0.490Z_1^2 - 0.067Z_2^2 - 0.150Z_3^2 - 0.067Z_4^2 + 0.053Z_5^2 - 0.230Z_6^2 + 0.424Z_7^2 \\ & + 0.063Z_8^2 - 0.090Z_9^2 - 0.108Z_{10}^2 + 0.024_{11}^2 + 0.039Z_{12}^2 + 0.059Z_{13}^2 - 0.602Z_{14}^2 \end{aligned} \tag{5}$$

表 7　第一组变量典型载荷

变量	1	2	3	4	5	6
年龄	-0.394	0.373	-0.834	0.097	-0.035	-0.005
性别	-0.922	-0.009	0.379	0.034	-0.042	-0.053
民族	-0.062	-0.115	-0.004	-0.357	-0.356	0.853
文化程度	0.122	-0.966	-0.097	-0.159	-0.045	-0.127
婚姻状况	-0.300	0.181	-0.301	-0.267	0.814	0.230
职业	-0.192	-0.116	0.053	0.898	0.133	0.351

表 8 第二组变量典型载荷

变量	1	2	3	4	5	6
肉蛋类	0.517	0.008	0.101	-0.103	0.318	0.012
豆类	-0.105	-0.456	-0.027	0.252	0.357	0.364
奶制品	-0.175	-0.664	0.359	0.020	0.046	-0.206
蔬菜	-0.099	-0.082	-0.212	0.019	0.521	-0.030
腌制品	0.123	0.139	-0.077	-0.303	-0.024	-0.057
水果	-0.251	-0.333	0.187	0.082	-0.180	0.000
饮料	0.496	-0.310	0.531	-0.191	0.053	0.058
盐	0.040	0.188	0.100	-0.339	0.282	-0.152
酱油	-0.074	0.122	0.011	0.290	0.182	0.052
醋	-0.170	-0.523	-0.429	-0.413	0.311	0.044
酱类	-0.001	-0.176	-0.124	-0.062	-0.165	0.627
味精	0.036	-0.112	-0.110	-0.361	-0.177	0.585
烹调油	0.083	0.116	0.155	0.082	0.505	0.366
休闲、家务活动	-0.662	0.245	0.407	-0.334	0.231	0.159

表 7 和表 8 给出了两组变量的典型载荷，以第一对典型变量为例，在第一组变量中性别的相关性最强，文化程度的相关性最弱。在第二组变量中，休闲、家务活动的相关性最强，酱类的相关性最弱。同理可得到剩下的典型变量。

5.3 问题三

问题三要求分析居民吸烟、饮酒、饮食习惯等因素与常见慢性病的关系以及关联程度。为解决此问题，我们建立了逻辑回归模型[2]。

5.3.1 数据预处理

（1）我们发现附件 2 中“是否被检查出糖尿病”、“是否被检查出高血压”和吸烟等指标具有部分缺失值，由于样本数据大，故将含有缺失值的数据直接去除。其中“是否被检查出糖尿病”指标剔除了 36 个缺失值，“是否被检查出高血压”指标剔除了 53 个缺失值，吸烟指标被剔除了 2 个缺失值，饮酒指标被剔除了 2 个缺失值，职业指标被剔除了 1 个缺失值，婚姻指标剔除了 1 个缺失值。

（2）我们对“是否被检查出糖尿病”和“是否被检查出高血压”指标进行了重新编码处理。其中，被检查出糖尿病记为 0，未被检查出糖尿病的记为 1。同理，被检查出高血压记为 0，未被检查出高血压的记为 1。

5.3.2 糖尿病逻辑回归模型的建立与求解

1. 模型的建立

设糖尿病逻辑回归自变量向量 $X = (X_1, X_2, \cdots, X_{86})^T$，$\beta = (\beta_1, \beta_2, \cdots, \beta_{86})^T$ 为对应的回归向量系数。则设被检查患有糖尿病的概率函数为：

$$F(X,\beta) = \frac{exp(X\beta)}{1 + exp(X\beta)} = \frac{e^{X\beta}}{1 + e^{X\beta}} \tag{6}$$

构造极大似然函数：

$$lnL(\beta|y,X) = \sum_{i=1}^{86} y_i ln[F(X,\beta)] + \sum_{i=1}^{86}(1 - y_i) ln[1 - F(X,\beta)] \tag{7}$$

通过极大似然估计，估计出 $\hat{\beta}$，然后通过 $\hat{y}$ 进行预测，如果 $\hat{y_i} \geqslant 0.5$，认为其预测值 $y = 1$，否则 $y = 0$

2. 模型的求解

将训练集与测试集按照 8:2 的比例进行划分，经 MATLAB 模型求解，取得了较好的模型效果。其中患有糖尿病的正确预测率为 64.7%，未患有糖尿病的正确预测率为 97.1%，整体的预测正确率达到了 96.1%

除此之外，还求出了各自变量影响逻辑回归的系数和显著性，如表 9 所示。

表 9　自变量系数和显著性

变量名	系数	显著性
文化程度	0.017	0.807
奶制品食用量	0.001	0.740
酱类的食用量	-0.026	0.696
是否食用水果	0.023	0.671
是否吸烟	0.072	0.527
盐的食用量	0.027	0.321
性别	0.242	0.200
蔬菜的食用量	-0.01	0.189
食用豆类的频率	0.002	0.067
是否食用饮料	0.003	0.002
···	···	···

根据逻辑回归模型系数和显著性可以分析糖尿病与居民生活习惯有以下关系：

（1）性别是影响糖尿病的重要因素，其逻辑回归系数最大。

（2）文化程度与是否患有糖尿病的较大的关系。

（3）各类食物的食用量与是否患有糖尿病具有密切的关系。

（4）各类食物的食用频率与是否患有糖尿病具有较小的关系。

（5）民族、婚姻状况与是否患有糖尿病具有较小的关系。

（6）总体上看，面食、大米、谷物为主的饮食与糖尿病的关系十分紧密，水产、咸菜为主的饮食与糖尿病的关系较小。

综上所述，应当适当减少这些高危食物的摄入，同时多食用一些降低患病概率的食物，从而预防糖尿病的发生。

5.3.3 高血压逻辑回归模型的建立与求解

原理上文糖尿病逻辑回归模型的建立一致，在此不再赘述。经 MATLAB 模型求解，取得了较好的模型效果。其中患有高血压的正确预测率为 52.3%，未患有高血压的正确预测率为 96.4%，整体的预测正确率达到了 92.8%。

表 10　自变量系数和显著性

变量名	系数	显著性
食用豆类的频率	0.001	0.979
是否食用水果	0.005	0.704
是否吸烟	-0.059	0.409
盐的食用量	0.015	0.374
文化程度	-0.048	0.310
酱类的食用量	0.015	0.140
蔬菜的食用量	-0.001	0.021
是否食用饮料	0.001	0.008
奶制品食用量	0.002	0.002
性别	0.443	0.001
...	...	...

根据逻辑回归模型系数和显著性可以分析高血压与居民生活习惯有以下关系：

（1）性别是影响高血压的重要因素，其显著性最小且逻辑回归系数最大。

（2）是否吃水果和蔬菜对是否患有高血压具有较大关系。

（3）各类食物的食用量对是否患有高血压具有较大的关系。

（4）各类食物的食用频率对是否患有高血压具有较小的关系。

（5）总体上看，以小麦面食和油炸食物为主的饮食与高血压关联程度最高，应尽量减少摄入。

综上所示，应当通过调整饮食结构，减少高危食物的摄入，多吃一些有益的食物来预防和控制高血压。

5.3.4 影响慢性病因素的建议

（1）从上述结果来看，性别是影响是否患有慢性病的重要因素。通过数据分析，我们发现男性比女性更加容易患有慢性病，故男性应该更加注重自己的身体状况。

（2）是否吸烟是影响是否患有慢性病的重要因素，通过数据分析，我们发现吸烟者患有慢性病的概率是不吸烟者的 2 倍。

（3）是否锻炼和锻炼的多少是影响是否患有慢性病的重要因素，坚持锻炼的居民拥有更低的患病率。

（4）饮食习惯是影响是否患有慢性病的重要因素，经常使用高糖、高油的居民更容易患有慢性病。

（5）文化程度在一定程度上也会影响慢性病的发病率，文化程度高的居民拥有更高的患病率，这可能是由于文化程度高的居民大多工作压力大，生活不规律导致的。

5.4 问题四

问题四要求对居民进行合理分类，并针对各类人群提出有益建议。为解决该问题，我们建立了 k-means++ 聚类模型[3]，根据居民的身体指标进行聚类。并给出针对各类居民的合理性建议。

5.4.1 数据预处理

我们发现附加 2 中居民的身体指标部分数据存在少量缺失值，由于样本数据大，故将含有缺失值的数据直接去除。具体各指标剔除的缺失值如下表所示。

表 11　各指标缺失值剔除数量

指标名	缺失值数量
低密度脂蛋白	79
血糖	27
高密度脂蛋白	23
尿酸	20
甘油三酯	20
胆固醇	20
脉搏	6
臀围	6
腰围	6
舒张压	4
收缩压	4
体重	4
身高	4

5.4.2 k-means++ 聚类模型的建立与求解

1. 模型的建立

k-means++ 聚类模型的建立具体步骤如下：

表 12　各指标缺失值剔除数量

步骤	具体操作
Step1	从数据集 ψ 随机选取一个样本点作为第一个初始聚类中心 c_i
Step2	计算每个样本与当前已有聚类中心之间的最短距离，用 $D(x)$ 表示
Step3	最后选择最大概率值所对应的样本点作为下一个簇中心
Step4	重复第 2 步和第 3 步，直到选择出聚类中心

其中 k-means++ 的代价函数为误差平方和公式为：

$$S = \sum_{i=1}^{m} |x_i, u_{c_i}|^2 \tag{8}$$

簇中心的计算公式为：

$$u_i^j = \frac{1}{|C_i|} \sum_{x \epsilon C_i} x^j \tag{9}$$

2. 模型的求解

采用 SPSS 软件进行求解，将样本数据聚类为四类，四类样本数量如下表所示。

表 13　各聚类得分与数量

聚类	1	2	3	4
数量	2419	2776	609	1807

各类别各指标聚类中心如表 14 所示。

5.4.3 对不同人群的生活建议

结合附件 3，可以对表 14 四个聚类划分为以下四中人群。

- 聚类 1 人群体重系数不正常，身高和体重均低于正常水平，且脂蛋白量较低，为营养不良人群
- 聚类 2 体重系数不正常，体重超标，尿酸偏高为肥胖人群[5]。
- 聚类 3 腰围和臀围偏大，血压偏高，胆固醇偏高，尿酸偏高，为患有高血压人群[4]。
- 聚类 4 各类指标正常，为正常人群。

针对不同人群，以下给出合理的健康建议：

（1）针对营养不良人群，应该均衡饮食，摄取各类营养素。要注重高蛋白食物的摄取，如瘦肉、家禽、鱼类和豆类等。应该建立良好的运动习惯，适当运动，并保证充足睡眠，作息规律。

表 14 各类别各指标聚类中心

指标	1	2	3	4
身高	158.84	161.74	166.94	165.76
体重	55.38	60.62	71.09	66.74
腰围	75.51	79.98	88.11	84.44
臀围	91.86	93.99	98.03	95.84
收缩压	109.00	114.00	122.00	118.00
舒张压	71.00	75.00	81.00	78.00
脉搏	73.00	73.00	73.00	73.00
胆固醇	4.67	4.85	5.28	4.99
血糖	5.03	5.09	5.23	5.15
高密度脂蛋白	1.33	1.23	1.09	1.37
低密度脂蛋白	2.81	3.00	3.33	3.14
甘油三酯	1.13	1.47	2.66	1.90
尿酸	214.61	292.07	471.93	370.85

（2）针对肥胖人群，应该减少高能量、高脂肪、高蛋白的食物的摄入，多吃水果和蔬菜。应该增强有氧运动，逐步增加运动量，以达到降低体重和尿酸的目的。

（3）针对高血压人群，应该食用低盐、富含钾、高纤维的食物，适量摄取蛋白质。应该养成进行有氧运动的习惯，提高心肺能力，促进血压的降低。应该控制体重，控制体重有助于降低血压。定期测量血压，及时调整生活方式和药物治疗。

（4）针对健康人群，要保持良好的生活习惯，坚持适量的有氧运动，保持健康的体重，注意饮食结构。

六、模型的评价与改进

6.1 模型的优点

• 通过将数据描述性统计化可以方便直观地观察分析，帮助我们直观、快捷地寻找到数据间的关系，寻找普适规律，使模型建立的数据信息更加可靠，更加贴近实际。在定性的描述之后进行定量的计算，使结果更加可靠。

• 通过对模型的层层检验和比较，使模型更加可靠的同时，能适应更加复杂的实际情况，模型简洁实用，可移植性强。

6.2 模型的缺点

• 由于数据量过多，处理过程较为复杂，不同的模型所要求的数据不同，故我们未将建立的模型与其他模型进行对比。

• 问题三构建的逻辑回归模型，约束条件单一，忽略了一些因素。

- 模型求解时间过长，消耗计算资源过多。

6.3 模型的改进

- 丰富问题三和问题四的比较特征，从已有的数据中提取更多有用的信息，将结果进行横向对比，确定最优结果。
- 减少模型假设，以提高模型的普遍性。

参考文献

[1] 明道绪, 龙漫远. 典型相关分析及其应用 [J]. 四川农业大学学报,1987(04):269-274.

[2] 王正存, 肖中俊, 严志国. 逻辑回归分类识别优化研究 [J]. 齐鲁工业大学学报,2019,33(05):47-51.DOI:10.16442/j.cnki.qlgydxxb.2019.05.008.

[3] 郑建军, 甘仞初, 贺跃等. 一种基于 k-means 的聚类集成方法 [C]//中国电子学会工业工程分会, 中国优选法统筹法与经济数学研究会工业工程分会及管理学分会, 国际信息处理联合会中国委员会第五技术委员会, 北京理工大学. 全国第九届企业信息化与工业工程学术会议论文集.《电讯技术》杂志社,2005:5.

[4] 段秀芳, 吴锡桂, 顾东风等. 中国人群血压分类与高血压患病率研究-1991 年血压抽样调查资料的进一步分析 [J]. 高血压杂志,2002(03):78-80.DOI:10.16439/j.cnki.1673-7245.2002.03.024.

[5] 李禄伟. 影响因素交互作用对超重及肥胖人群高尿酸血症患病的相关研究 [D]. 桂林医学院,2022.DOI:10.27806/d.cnki.gglyx.2022.000139.

附 录

附录 1：代码

```
%数据处理代码 chuli_you.m
clc;clear;
load data_all
% 遍历每一列
for col = 1:size(data_all, 2)
    % 获取当前列，并忽略NaN值
    currentColumn = data_all(:, col);
    validValues = currentColumn(~isnan(currentColumn));
    % 计算当前列的均值和标准差，忽略NaN
    meanValue = mean(validValues, 'omitnan');
    stdValue = std(validValues, 'omitnan');
    % 找到大于3倍标准差的异常值索引，同时确保不是NaN
    outlierIndex = abs(currentColumn - meanValue) > 3 * stdValue & ~isnan(
        currentColumn);
    % 用均值替换异常值
    data_all(outlierIndex, col) = meanValue;
end

%问题1代码 Question_1.m
clc;clear;
load mydata2
%将频率归一为天
for i=23:5:153
    %先找出频率为天的空缺值
    index_1=isnan(data2(:,i));
    %找出频率为周的非空缺值
    index_2=~isnan(data2(:,i+1));
    %找出频率为月的非空缺值
    index_3=~isnan(data2(:,i+2));
    data2(logical(index_1.*index_2),i)=data2(logical(index_1.*index_2),i+1)./7;
    data2(logical(index_1.*index_3),i)=data2(logical(index_1.*index_3),i+2)./30;
end
%对标记为吃或者平均每次食用量不为0，但缺失食用频率的进行均值填充
for i=23:5:153
    %找出归一后频率(天)的空缺值的下标
    index_1=isnan(data2(:,i));
    mean_temp=nanmean(data2(:,i));
    %找出标记为吃或者平均每次食用量不为0的下标
    index_2=~isnan(data2(:,i-1)+data2(:,i+3));
    data2(logical(index_1.*index_2),i)=mean_temp;
```

```
end
%对是否吃的空值进行修正
for i=22:5:152
    index_1=isnan(data2(:,i));
    index_2=~isnan(data2(:,i+1)+data2(:,i+4));
    %填补为1的下标
    index_3=logical(index_1.*index_2);
    %进行错误值纠正，以及缺失值填补
    data2(index_3,i)=1;
    data2(isnan(data2(:,i)),i)=2;
end
%对平均每次食用量的空值进行填充
for i=26:5:156
    %找出食用量的空缺值的下标
    index_1=isnan(data2(:,i));
    mean_temp=nanmean(data2(:,i));
    %找出标记为吃或者平均每次食用量不为0的下标
    index_2=~isnan(data2(:,i-4)+data2(:,i-3));
    data2(logical(index_1.*index_2),i)=mean_temp;
end


%问题四代码 question_4_kmean.m

clc;clear;close all;
data=xlsread("D:\huashubei\第四问_kmean++.xlsx");
%使用kmean++确定初始聚类中心
%聚类个数
cluster_num=4;
%产生一个初始聚类中心
init_1=randi([1,size(data,1)]);
cluster_center=zeros(cluster_num,2);
cluster_center(1,:)=data(init_1,:);
%初始化距离矩阵
distance=zeros(size(data,1),cluster_num)+999999;
%计算每个样本与当前已有聚类中心的最短距离
for i=1:cluster_num-1
    for j=1:i
        %将第j个中心复制size(data,1)份便于计算
        temp=repmat(cluster_center(j,1),size(data,1),1);
        %计算每个点到第j个中心的距离
        distance(:,j)=abs(temp(1)-data(:,1),2);
    end
    %选取第i+1个聚类中心
```

```
    min_distance=min(distance');
    rate=min_distance/sum(min_distance);
    cum_rate=cumsum(rate);
    %产生一个随机数
    rand_1=rand();
    newin=1;
    fitin=1;
    %使用轮赌法找到聚类中心
    while newin<=1
        if(rand_1)<cum_rate(fitin)
            cluster_center(i+1)=data(fitin);
            newin=newin+1;
        else
            fitin=fitin+1;
        end
    end
end

%进行聚类
[m,n]=size(data);
%定义聚类参数
iter_max=1000;
deta=0.001;
iter_num=0;
while(iter_num<iter_max)
    iter_num=iter_num+1;
    for i=1:cluster_num%计算每个数据点到每个聚类中心的距离
        distance=(data-repmat(cluster(i,1),m,1));
        new_distance(:,i)=abs(sum(distance'));
    end
    %找到当前点最近的聚类中心，并把他分配给当前的聚类中心
    [~,index_cluster]=min(new_distance');
    %更新聚类中心
    for j=1:cluster_num
       new_cluster(j,1)=mean(data(find(index_cluster==j),1));
    end
    %判断聚类中心是否变化
    if (sqrt(sum((new_cluster-cluster).^2))>deta)
        cluster=cluster_new;
    else
        break;
    end
end
```