参赛密码 _____

（由组委会填写）

“华为杯”第十三届全国研究生数学建模竞赛

学　　校		浙江大学
参赛队号		10335004
队员姓名	1.	李克西
	2.	林苾湲
	3.	张苏锐

参赛密码 _____

（由组委会填写）

题 目 **具有遗传性疾病和性状的遗传位点分析**

摘 要：

本文根据全基因组关联分析的步骤，运用统计学和机器学习方法，对样本数据建立数学模型，主要完成了以下几方面的工作:

针对问题一：在对原始数据进行统计分析、清洗及质量控制检测后，比较了数值编码、哑元化以及 One-Hot 向量等几种对无序特征编码的方案，提出一种改进的数值型编码方式，既不增加样本向量空间维度，又考虑了碱基对排列组合方式间的距离问题。

针对问题二：利用卡方检验获取在显著水平为 0.01 的条件下，与疾病 A 相关性较高的 16 个位点。并进行进一步检验，确定位点 rs2273298 为疾病 A 的致病位点，位点 rs932372、位点 rs12036216 和位点 rs2807345 极有可能为疾病 A 的致病位点。Logistic 回归检验的准确率为 68.3%.

针对问题三：首先，在显著性水平为 0.05 的条件下，筛选出 167 个可能与疾病相关的位点。然后，根据位点信息对包含这些位点的 105 个基因进行重编码。使用卡方检验对这些基因与疾病 A 进行关联性分析；同时采用权重累加编码方式进行辅助检验，确定出与疾病 A 最相关的基因为 gene_55，其他较为相关的基因为 gene217、293、169。使用 Logistic 回归检验准确率为 79.9%.

针对问题四：采用典型相关分析法，将多维变量转为单一综合变量，具体使用 K-means 方法对相关性状数据进行聚类，根据多次实验的残差平方和确定类别数 k 为 2。接着使用卡方检验确定位点 rs10157835 为这些相关性状的致病位点，位点 rs12746773 和位点 rs3218121 极有可能为表现综合性状的致病位点。Logistic 回归检验准确率为 68.3%.

关键词：GWAS、卡方检验、Logistic 回归检验、K-means 聚类、典型相关分析

目 录

1. 问题重述

1.1 问题背景

人体的每条染色体携带一个 DNA 分子，人的遗传密码由人体中的 DNA 携带。DNA 是由分别带有 A,T,C,G 四种碱基的脱氧核苷酸链接组成的双螺旋长链分子。在这条双螺旋的长链中，共有约 30 亿个碱基对，而基因则是 DNA 长链中有遗传效应的一些片段。在组成 DNA 的数量浩瀚的碱基对（或对应的脱氧核苷酸）中，有一些特定位置的单个核苷酸经常发生变异引起 DNA 的多态性，我们称之为位点。

在 DNA 长链中，位点个数约为碱基对个数的 1/1000。由于位点在 DNA 长链中出现频繁，多态性丰富，近年来成为人们研究 DNA 遗传信息的重要载体，被称为人类研究遗传学的第三类遗传标记。

大量研究表明，人体的许多表型性状差异以及对药物和疾病的易感性等都可能与某些位点相关联，或和包含有多个位点的基因相关联。因此，定位与性状或疾病相关联的位点在染色体或基因中的位置，能帮助研究人员了解性状和一些疾病的遗传机理，也能使人们对致病位点加以干预，防止一些遗传病的发生。

近年来，研究人员大都采用全基因组的方法来确定致病位点或致病基因，具体做法是：招募大量志愿者（样本），包括具有某种遗传病的人和健康的人，通常用 1 表示病人，0 表示健康者。 对每个样本，采用碱基(A,T,C,G)的编码方式来获取每个位点的信息(因为染色体具有双螺旋结构，所以用两个碱基的组合表示一个位点的信息)。虽然碱基的组合不同，但每个位点只存在三种不同编码。研究人员可以通过对样本的健康状况和位点编码的对比分析来确定致病位点，从而发现遗传病或性状的遗传机理。

由于可以把基因理解为若干个位点组成的集合，遗传疾病与基因的关联性可以由基因中包含的位点的全集或其子集合表现出来。另外，人体的许多遗传疾病和性状是有关联的，如高血压，心脏病、脂肪肝和酒精依赖等。科研人员往往把相关的性状或疾病放在一起研究，这样能提高发现致病位点或基因的能力。在实际的研究中，科研人员往往把相关的性状或疾病看成一个整体，然后来探寻与它们相关的位点或基因。

1.2 问题提出

问题一：为了便于进行运算分析，首先需要选择合适的方法将样本数据中每个位点包含的碱基对信息转化为数值型编码。

问题二：给出 1000 个样本在某条有可能导致遗传病 A 的染色体片段上的 9445 个位

点的编码信息以及样本患有遗传病 A 的数据。设计或采用某种方法找出遗传病 A 最有可能的一个或几个致病位点，并给出合理的解释。

问题三：如果把基因理解为若干个位点组成的集合，那么遗传疾病与基因的关联性就可以由基因中包含的位点的全集或其子集合表现出来。针对问题二中给出的的数据，提供 300 个基因和其每个基因所包含的位点信息，要求从中找出遗传病 A 最有可能的一个或几个基因，并给出合理的解释。

问题四：给出 10 个相关联的性状信息，要求将它们作为一个整体，从问题二的 9445 个位点中找出与这 10 个性状有关联的位点，并给出合理的解释。

2. 问题假设与符号系统

2.1 问题假设

假设 1：位点中碱基的顺序对该位点是否致病没有影响。即

$$d(\mathrm{XY}) = d(\mathrm{YX}) \tag{2-1}$$

其中，d(XY)表示某位点上碱基对 XY 对疾病 d 的致病性，且

$$\begin{cases} d(\mathrm{XY}) = 1, \text{表示该位点碱基对XY致病} \\ d(\mathrm{XY}) = 0, \text{表示该位点碱基对XY不致病} \end{cases} \tag{2-2}$$

假设 2：假设样本为无相关个体。

假设 3：假设不考虑连锁不平衡的影响。

2.2 定义与符号系统

2.2.1 定义

1) 等位位点：指不同样本中同一基因同一位置的位点，可类比于等位基因。
2) 碱基对之间的距离：碱基对 A 变异为碱基对需要改变的碱基个数。
3) SNP: 单核苷酸的多态性，SNP 位点、SNP 和位点在本文中也可以指代题目中的位点。
4) 阳性 SNP 位点: 指按一定 P 值筛选出的可能致病 A 的 SNP 位点，用以表征基因对病 A 的关联性。

2.2.2 符号系统

X	样本输入矩阵
X’	最小化样本输入矩阵
Y	输出目标矩阵
W	权重矩阵
S	单个样本的位点编码向量
χ^2	卡方统计量
A	观察频数
E	期望频数
p_i	i 水平的期望频率
k	自由度

H_0	无效假设
G	基因编码总数
C_n	某位点的编码个数
W_{gene}	单个基因权重值
Chi_i	相关位点的卡方值
V	聚类中心向量
x_j	样本
μ_i	质心
K	质心个数
J	评价指标

3. 问题分析

3.1 相关概念

3.1.1 单核苷酸的多态性(SNP)

单核苷酸的多态性(Single Nucleotide Polymorphisms, SNP)是指在基因组上单个核苷酸的变异，包括转换、颠换、缺失和插入。有些 SNP 位点并不直接导致疾病基因的表达，但由于它与某些疾病基因相邻，而成为重要的标记。理论上讲，SNP 既可能是二等位多态性，也可能是 3 个或 4 个等位多态性，但实际上，后两者非常少见，几乎可以忽略。因此，通常所说的 SNP 都是二等位多态性的。

3.1.2 全基因组关联性分析(GWAS)

全基因组关联分析(Genome-Wide Association Study, GWAS) 利用分布于全基因组范围内的分子标记，基于它们与分析性状的连锁不平衡关系，通过各种统计分析方法，以获得与这些性状关联的候选基因或基因组区域。采用 CASE-CONTROL 试验设计，比较全基因组范围内所有 SNP 位点的等位基因或者基因型频率在 CASE 与 CONTROL 组中的差异，如果某个 SNP 位点等位基因或基因型有病例组中出现的频率明显高或低于对照组，则认为该位点与疾病之间存在关联。

采用小样本数量进行第一阶段的全基因组范围 SNP 基因分型，统计分析过后一般能够筛选少量阳性 SNP 位点，之后的第二阶段再在更大数量的样本中对这些阳性 SNP 位点进行基因分型，最后整合两个阶段的结果进行分析。

3.2 问题一的分析

问题一需要根据位点的碱基编码方式来得到计算机易于理解地编码方式。由于 SNP 通常是二等位多态的，所以不同样本中等位位点的编码都是固定 2 种碱基的组合，至多有 3 种不同的编码方式。在后续数学分析的过程中，主要关注的是不同样本的等位位点或等位基因的差异与某一个或几个性状的关系，而与非等位位点之间碱基类型或非等位基因之间基因型的差别无关。首先需要对原始数据进行统计分析、清洗及质量控制检测，然后寻找合适的无序特征编码方案。

3.3 问题二的分析

本题给出了 1000 个样本的 9445 个位点，要求采用合理的方法找出某种疾病 A 最有可能的一个或几个致病位点。在求解过程中，为了便于从 9445 个位点中筛选出致病位点，暂时忽略位点之间由于位置相邻等因素的影响，只考虑位点与性状之间的双变量关

联性，从而使问题得以简化。由于 GWAS 的数据结构是频数型的，所以采用卡方检验的方法来分析双变量的关联性，在分析的过程中应当关注病患组与控制组之间等位位点差异性的比较。最后应当对关联分析的结果进行多重检验，排除假阳位点。

3.4 问题三的分析

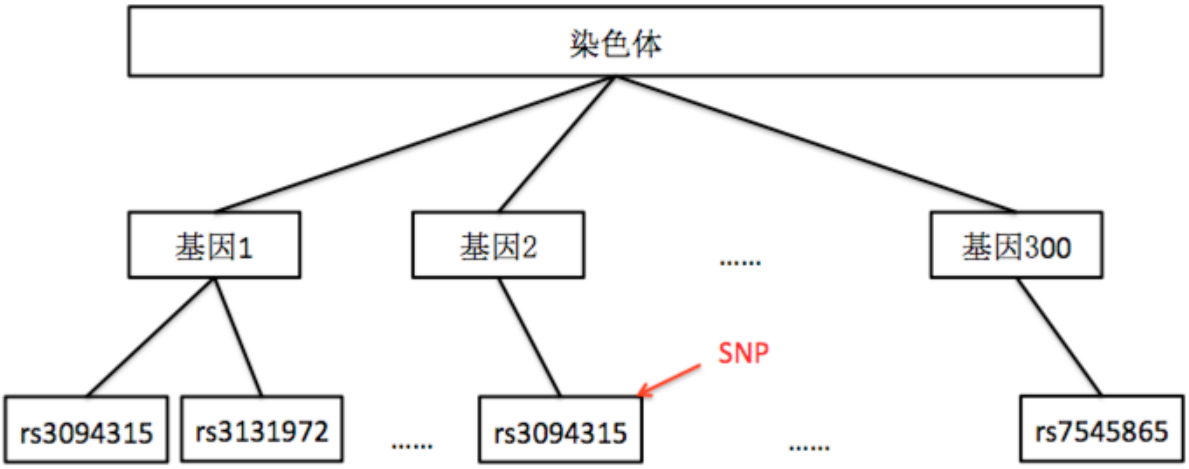

图 3-1 以阳性 SNP 表征基因

本题需要从样本中找出与疾病 A 表现性状相关的一个或几个致病基因。由于基因可以由若干 SNP 位点组成的集合表征，因此遗传疾病与基因的关联性可以由基因中包含的 SNP 位点的全集或其子集表现出来。所以本题的关键在于如何利用阳性 SNP 位点表征的基因量化与疾病 A 之间的相关性。

解决这一问题可以有两种思路：一种是对基因按阳性 SNP 位点集合的基因型进行重新编码，从而转化为问题二。此时同问题二一样，依然不考虑基因之间的相互影响，而只关心单个基因不同基因型频率与疾病 A 表现性状之间的双变量关联性。另一种思路则是对阳性 SNP 位点赋予合适的权值，按权重累加的结果指标评价基因与疾病 A 表现性状之间的关联关系。

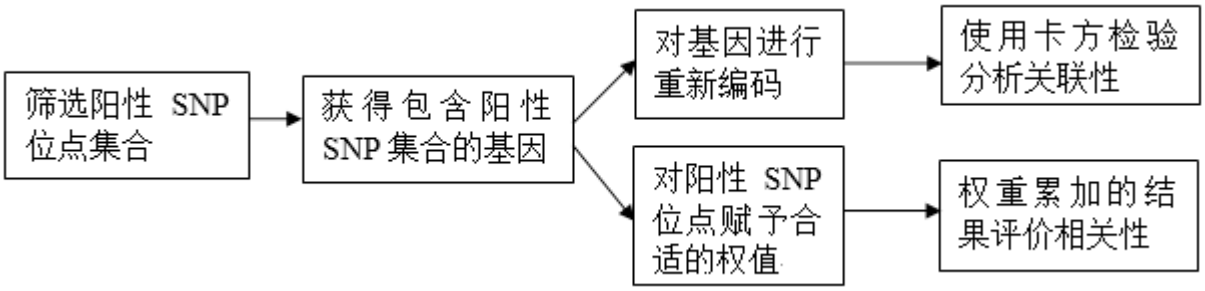

图 3-2 问题三流程图

3.5 问题四的分析

同样要求从 1000 个样本的 9445 个 SNP 位点中找出致病位点，与第二题不同的是，本题中给出了相互关联的 10 种性状。显然不能单独寻找各性状最可能的致病位点集合，而应将 10 个形状看作一个整体。采用典型相关分析等方法可以将若干性状转化为一个综合变量，然后考虑各 SNP 位点与该综合变量之间的关联关系，从而可以借鉴第二题的解决方法。

4. 问题求解

4.1 问题一的求解

4.1.1 数据预处理

首先对碱基对文件 genotype.dat 进行统计分析，去掉首行后利用 Spark 进行词频统计，得到文件中各碱基对的统计柱状图。

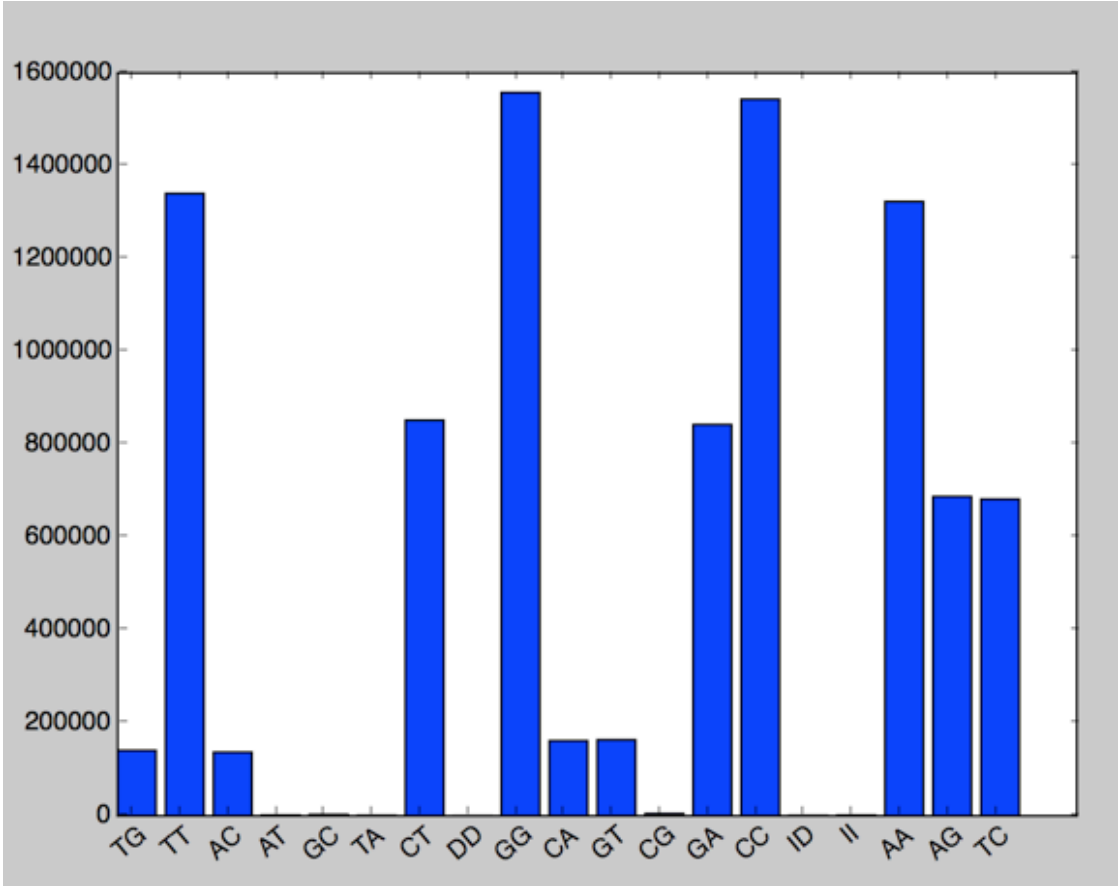

图 4-1 genotype.dat 文件各基因型频次统计柱状图

从统计结果发现文件中存在 I、D 两种异常碱基，根据要求 I、D 替换为 T、C 两种碱基。数据矫正后进行质量控制(QC)检验，发现不存在 MAF < 0.01 (或 0.05)的位点。最后确认碱基对的排列方式有 AA、AT、TT、AC、CC、AG、GG、TC、CG、TG 共 10 种情况。

4.1.2 无序特征的数值化编码方式

碱基对的编码方式呈现出一种离散的无序特征，无序特征常用的数值化编码方式包括直接数值化编码、哑元化编码和 One-Hot 向量化编码。

1) 直接数值化编码：

使用序号对无序特征进行编号，如采用 0、1、2 的数字编号。直接数字化编码的优

点在于简单易行，容易理解；缺点是编码后的向量大多情况下不能直接用作很多算法的输入。

2) 哑元化编码：

直接数值编会人为引入无关信息（如欧式空间距离），因此需要通过增加虚拟元来表示原来枚举值的序号。本题中 10 中碱基对组成需要增加 3 个虚拟元，原来的 1 个向量维度需要扩展成 4 维。

3) One-Hot 向量化编码：

即将原来枚举值表示为只有一个维度值为 1 的 0-1 向量，本问题中的 10 个无序特征，特征 1 及特征 8 的编码方式如下：

$$1 = [0,1,0,0,0,0,0,0,0,0]$$

$$8 = [0,0,0,0,0,0,0,0,1,0]$$

这种方式将原来一个样本的1×9445的向量扩展成了10×9445的矩阵，或者1×94450的向量。

哑元化和 One-Hot 向量两种编码方式都有一个共同的缺陷，会增加向量的空间维度，为后期处理带来极大的困难。针对本题每个位点 3 种碱基对排列方式，考虑采用一种改进的数值型编码方式。

4.1.3 碱基对数值化编码模型

设某个位点可能出现的碱基类型为 X 和 Y，在数值化的过程中，对同一位点不同碱基对 XY, XX, YY 的数值编码之间的距离有如下关系：

$$\begin{cases} |XX - XY| = |YY - XY| \\ |YY - XX| = 2\,|XX - XY| = 2\,|YY - XY| \end{cases} \tag{4-1}$$

可以从上文中 SNP 二态性的角度解释公式(4-1)的生物学意义：

在同一位点上，由碱基对 XY 转换为 XX，需经过该位点碱基 X 变异为 Y; 由编码 XY 转换为 YY, 需经过该位点碱基 Y 变异为 X; 由编码 XX 颠换为 YY，需经过该位点 2 个碱基 X 都变异为 Y; 其他碱基对的变异类似。

为了简化计算，对于任意一个位点，将 XY 型碱基对数值化编码为 1，XX、YY 型分别编码为 0 和 2，即令 XX=0, XY=1, YY=2，则公式(4-1) 成立。同样的，令 XX=2, XY=1, YY=0，则公式(4-1)也成立。

由此可以定义单个样本的位点向量为：

$$S = (s_1, s_2, ..., s_{9445}) \tag{4-2}$$

其中 $s_i \in \{0,1,2\}$.

4.1.4 模型的求解

```
DataToVector( file )
    loadFile( file )
    for 文件中每一行数据line
        按照空白符将line分割为数组arr
        初始化一个字符串构造器builder
        for arr中每个排列item
            code = getEncode( item )
            builder.append(code)
        将builder中的信息写入到输出文件
end

getEncode( item )
    if 符号表map中存在item
        return item对应编号
    else
        if item[0] == item[1]
            map.put( item, 1 ) return 1
        else
            if 符号表values含有0
                map.put( item, 2 ) return 2
            else
                map.put( item, 0 ) return 0
end
```

图 4-2 算法伪代码

按照图 4-2 中的算法，对 1000 个样本所给位点的数值编码结果见下表 4-1，详见附件.

表 4-1 碱基对的数值型编码

	SNP1	SNP2	…	SNP9445
样本 1	1	1	…	1
样本 2	1	1	…	0
…	…	…	…	…
样本 1000	1	1	…	2

4.2 问题二的求解

4.2.1 位点与疾病 A 关联性分析模型

根据 GWAS 分析的原理，在经过质量控制的位点集合中，通过采用对各位点与疾病 A 表现性状之间关联性进行卡方检验的方法，假设某位点与患疾病 A 不相关，将病例组和控制组作比较，筛选与假设存在显著差异的位点集合。再经过 Logistic 回归等多重检验方法，排除其中的假阳性位点，最终从 9445 个 SNP 位点中得到疾病 A 最可能的致病位点集合。

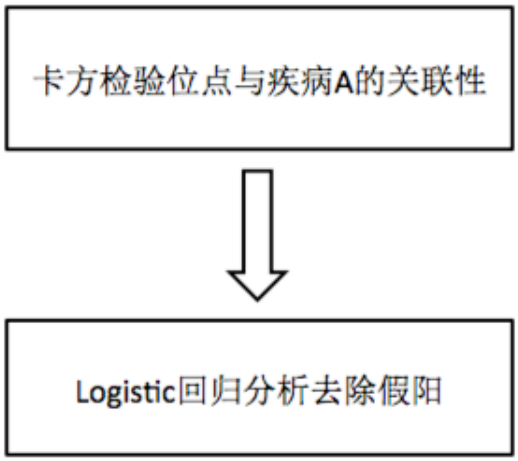

图 4-3 位点与疾病 A 关联性分析模型

卡方检验(Chi-square test / Chi-Square Goodness-of-Fit Test) 是一种用途很广的计数资料的假设检验方法。它属于非参数检验的范畴，主要是比较两个及两个以上样本率(构成比）以及两个分类变量的关联性分析。其根本思想就是在于比较理论频数和实际频数的吻合程度或拟合优度。

卡方统计量的公式如下:

$$\chi^2 = \sum \frac{(A-E)^2}{E} = \sum_{i=1}^{k} \frac{(A_i - E_i)^2}{E^2} = \sum_{i=1}^{k} \frac{(A_i - np_i)^2}{np_i} (i = 1, 2, ..., k) \tag{4-3}$$

其中，A_i 为 i 水平的观察频数，E_i 为 i 水平的期望频数，n 为总频数，p_i 为 i 水平的期望频率。i 水平的期望频数 E_i 等于总频数 n×i 水平的期望概率 p_i，k 为 C 行 R 列表资料的自由度其计算公式为:

$$k = (C-1)\ (R-1) \tag{4-4}$$

当 n 比较大时，卡方统计量近似服从 k-1(计算 E_i 时用到的参数个数)个自由度的卡

方分布。

此处的无效假设 H_0 是：患疾病 A 与某位点无关。该检验的基本思想是：首先假设 H_0 成立，基于此前提计算出卡方值，它表示观察值与理论值之间的偏离程度。根据卡方分布及自由度可以确定在 H_0 假设成立的情况下获得当前统计量及更极端情况的概率 P。如果 P 值很小，说明观察值与理论值偏离程度太大，应当拒绝无效假设，表示比较致病与某位点有关；否则就不能拒绝无效假设。

根据上述公式对每一个位点进行卡方检验，得到的卡方值进行排序，即可得到最有可能导致疾病 A 的位点集合。

4.2.2 位点与疾病 A 关联性的卡方检验

首先对每个位点各基因型进行频数统计。按照问题一中的编码方式(括号中为基因型)，以位点 rs2273298 为例分别统计病例组和控制组各基因型的频数，结果见表 4-1.

表 4-1 病例组与控制组 rs2273298 位点的基因型频数

位点 rs2273298	0(GG)	1(AG)	2(AA)
患病 A	60	218	222
未患病 A	34	161	305

卡方检验过程：

1) 零假设 H_0：患疾病 A 与位点 rs2273298 无关

2) 确定自由度为(3-1)×(2-1)=2，选择显著水平 α=0.01。

3) 对假设 H_0 进行卡方检验，获得的结果如下：

从表 5-2 卡方检验的结果中 rs2273298 对应数据可以看出：P 值远小于 0.01, 因此可以认为假设不成立，即位点 rs2273298 不同的基因型对是否患病 A 有显著性差别，即此位点很可能为疾病 A 的致病位点之一。

使用 Python 的机器学习工具包 scikit-learn 对所有位点是否导致疾病 A 进行卡方检验，按卡方值进行排序，得到如下可能致病位点的下标位置(下标从 0 开始)、卡方值和 P 值，如表 4-2：

表 4-2 疾病 A 可能的致病位点

位点	位置	卡方值	P 值
rs2273298	2937	20.95	0.000005
rs932372	7736	13.71	0.000213

rs12036216	79	13.37	0.000256
rs2807345	6793	10.94	0.000942
rs12133956	477	8.75	0.003095
rs11580218	6840	8.29	0.003995
rs12028945	7381	8.20	0.004185
rs7522344	1592	7.69	0.005548
rs9426306	8588	7.55	0.006001
rs731024	1923	7.44	0.006387
rs12036552	1598	7.34	0.006725
rs6429804	3937	7.15	0.007503
rs3818033	4919	7.13	0.007592
rs590368	3572	7.11	0.007661
rs3000851	8639	6.68	0.009761
rs6661776	3384	6.64	0.009961

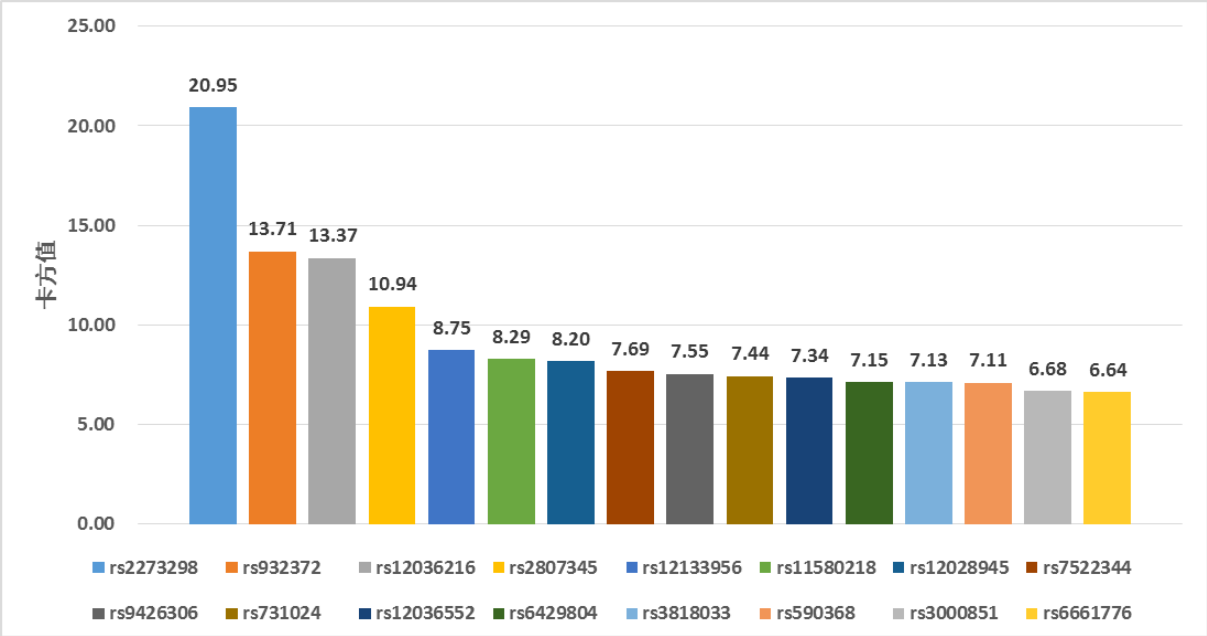

图 4-4 疾病 A 可能治病位点的卡方值

从图 4-4 中按卡方值对 P 值小于 0.01 的位点进行排序的结果。本题以卡方值表征位点与性状关联性的近似量化结果，结合以上图表得出结论：

位点 rs2273298 为疾病 A 的致病位点，位点 rs932372、位点 rs12036216 和位点

rs2807345 极有可能为疾病 A 的致病位点，图 4-4 中其余位点有可能为疾病 A 的致病位点。

4.2.3 Logistic 回归检验

本题也可以看做一个广义线性回归问题。给定输入编码后为一个1000×9445维度的矩阵 X，目标输出仅为 0 和 1（患病和非患病），即二分类问题。

$$X=\begin{bmatrix} S_1 \\ S_2 \\ ..... \\ S_{1000} \end{bmatrix} \tag{4-5}$$

其中

$$S_i=[s_1,s_2,.....s_{9445}],s_i\in\{0,1,2\}$$

若样本数据与致病性状存在线性相关性，那么目标函数为

$$Y=f(wX+b) \tag{4-6}$$

设 X’为 X 的一个子集，其中含有与 Y 相关的所有位点信息（列信息），那么有

$$Y=f(w'X'+b) \tag{4-7}$$

我们的目标是求得最小化的 X’，使得 Y 与目标输出 Y’的损失最小化。

Logistic 是一种最优化二分类器，是广义线性回归的一种推广，其核函数为 Sigmoid 函数，因此目标函数为

$$Y=f(wX+b)=sigmoid(wX+b)=\frac{1}{1+e^{wX+b}} \tag{4-8}$$

其中 w 为系数矩阵， b 为常量项，sigmoid 函数图像如下图 4-5：

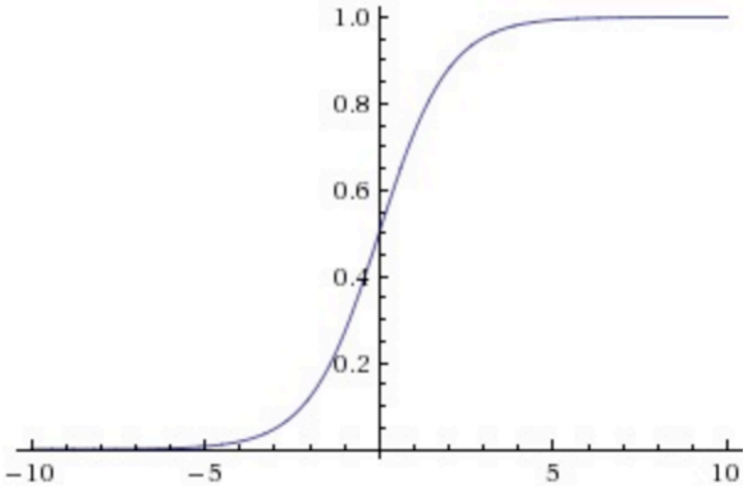

图 4-5 sigmoid 函数图

当 wX+b 大于阈值时，分类器输出预测为 1，小于输出预测为 0。

Logistic 回归同时是一种有监督学习的机器学习分类方法，需要通过带有目标类别的训练数据集训练模型，训练过程中采用随机梯度下降的方式调整各系数权重，使得损失函数最小化。

从经过卡方检验后的特征列中，采用逐步回归法选取组成样本输入矩阵训练 Logistic 分类器，并采用交叉验证方式进行评估，利用分类器指标检验特征选择的正确性。

使用 SPSS 进行 Logistic 回归分类，其结果如下，各变量系数见附件。

表 4-3 问题二 Logistic 回归检验

分类表[a]

已观测		已预测 label 0	已预测 label 1	百分比校正
步骤 1 label	0	344	156	68.8
	1	161	339	67.8
总计百分比				68.3

a. 切割值为 .500

由上图可得，分类的准确率为 68.3%，相比于单因素分析，准确率提高了 10%，故由疾病与位点关联性模型求解得到结果可信度较高，即位点 rs2273298 为疾病 A 的致病位点，位点 rs932372、位点 rs12036216 和位点 rs2807345 极有可能为疾病 A 的致病位点，图 5-1 中其余位点有可能为疾病 A 的致病位点。

4.3 问题三的求解

利用问题二的方法筛选合适的 SNP 集合作为阳性 SNP 位点，以阳性 SNP 位点表征可能的致病基因，并剔除不包含阳性 SNP 位点的基因。以阳性 SNP 位点表征基因有两种思路：一种是对基因按阳性 SNP 位点集合的基因型进行重新编码，从而转化为问题二。另一种思路则是对阳性 SNP 位点赋予合适的权值，按权重累加的结果指标评价基因与疾病 A 表现性状之间的关联关系。

4.3.1 基因重编码

对基因按问题二中方法筛选出阳性 SNP 位点集合的基因型进行重新编码，从而转化为问题二的思路，对基因与疾病 A 的关联性进行卡方检验，并以多重检验验证。

使用第二题的疾病与位点关联性模型，并将显著性水平设为 0.05，即 P 值小于 0.05

下，得到与疾病相关位点，将位点与基因对应，找到包含这些可能致病位点的基因，对基因的各位点的碱基对类型进行排列组合，则编码的种类至多为

$$G = \prod_1^n C_n \tag{4-9}$$

C_n 为该位点的编码的总数，编码以 0 起始，G-1 为结束，累加编码。

4.3.2 卡方检验

使用卡方检验来比较每个基因的基因频率在病例组和对照组的差异，若检验结果与假设 H_0：该基因与疾病 A 不存在关联，存在显著性差异，则可以表明该基因与疾病相关。

1) 设定无效假设 H_0：疾病 A 与某基因无关。
2) 计算自由度：

   其中，C 为某基因的编码类型种类数，R 为 2（代表健康与不健康）
3) 计算卡方统计量
4) 结论：若 P<0.01,则拒绝无效假设 H0, 即疾病 A 与某基因有关；若 P>=0.01,则接受无效假设 H0，即疾病 a 与某基因无关

4.3.3 权重累加法

采用一种简单的权重累加法，即将基因中包含的所有相关位点的卡方统计量进行累加，结果作为基因的权重，按降序排列，从而得到与疾病 A 相关性递减的基因序列。单个基因权重值计算公式为：

$$W_{gene} = \sum_{i=1}^{m} Chi_i$$

通过该方法计算出所有含有相关性位点的基因的权重值后进行降序排列，从另一个角度得到相关性基因，对卡方检验的结果进行验证和微调。

4.3.4 模型求解

1) 基因重编码的求解

使用第二题对位点进行卡方检验的结果，将显著性水平的阈值设为 0.05，筛选出 167 个可能与疾病相关的位点，下表 4-4 为部分位点信息，全部位点信息见附件。

表 4-4 阳性 SNP

位点名称	显著性水平
rs12036216	0.000256

rs3813199	0.019115
rs2477777	0.036504
…	…
rs2803309	0.037514

根据文件夹 gene_info 中的 300 个 dat 文件，使用 Java 代码将筛选出来的位点与基因对应，找到包含这些致病位点的基因，代码流程图如下图 4-6：

```
matchGenos( fileList, indexList )
    获取位点名称-下标映射表map
    从map中选取indexList中包含的相关位点集合set
    初始化保存<基因, 相关位点列表>的hashMap
    for file in fileList
        读取基因信息文件file中包含的位点信息
        将set中含有的所有位点信息加入到一个新的list中
        if list所含元素不为空
        hashMap.put( 基因名, list )
    return hashMap
end
```

图 4-6 算法伪代码

运行得到包含与疾病相关的位点的基因有 105 个，表 4-5 列出部分基因以及包含的致病位点，全部基因及其包含的疾病相关位点信息见附件。

表 4-5 包含阳性 SNP 的部分基因

gene_169	rs16830759，rs473648，rs523919，rs648305，rs12097284
gene_113	rs2594289，rs761087，rs10779763
gene_115	rs12095517，rs12752700，rs6661776，rs1133398，rs10779765
gene_30	rs3765702，rs12117836，rs10910024，rs1181876
gene_55	rs7415936，rs10907214，rs9662668，rs12128558，rs6687987，rs7522344，rs12036552
gene_293	rs1201394，rs1924270，rs522294，rs1830705，rs7555715，rs1188399，rs1188403，rs3795438
gene_217	rs17356059，rs2473246，rs7513455，rs2473247，rs2473253，rs2505722，rs2807345
gene_114	rs12139487，rs9430624

根据上表，结合位点的编码信息，对基因进行编码，使用 Python 编程实现，伪代码如下：

```
GenoCode():
    1. 获取位点的编码 poinDataList 以及位点名称数据 pointNameList
    2. 获取包含相关位点的基因列表 genoNameList，以及相关位点 genoDataList
    3. 获取样本的疾病信息 sickData
    for genoData in genoDataList：
        得到该基因包含的所有位点的编码信息 point_geno_list
        编码 ID 设为 0
        for i in range(1000):
            获取每一个样本该基因位点组合编码信息
            判断该组合编码信息是否已经存在字典里
            如果不存在字典里：则添加该组合编码信息以及当前编码 ID 到字典
            编码 ID 加一并保存编码 ID
End
```

图 4-7 算法伪代码

结合 1000 个样本的的疾病信息和样本的基因编码信息，得到包含阳性 SNP 的基因的编码结果，表 4-6 为部分样本中包含阳性 SNP 的基因的编码结果，全部编码信息见附件。

表 4-6 部分包含阳性 SNP 的基因的编码结果

样本编号	基因编码信息				样本健康状况
	gene_55	gene_217	…	gene_218	
1	0	0	…	0	0
13	12	0	…	0	0
…	…	…	…	…	…
501	221	1	…	3	1

2) 疾病与基因关联性模型求解

使用 SPSS 对表 4-7 的信息进行卡方检验，可以得到 105 个基因的 P 值，如下图所

示：

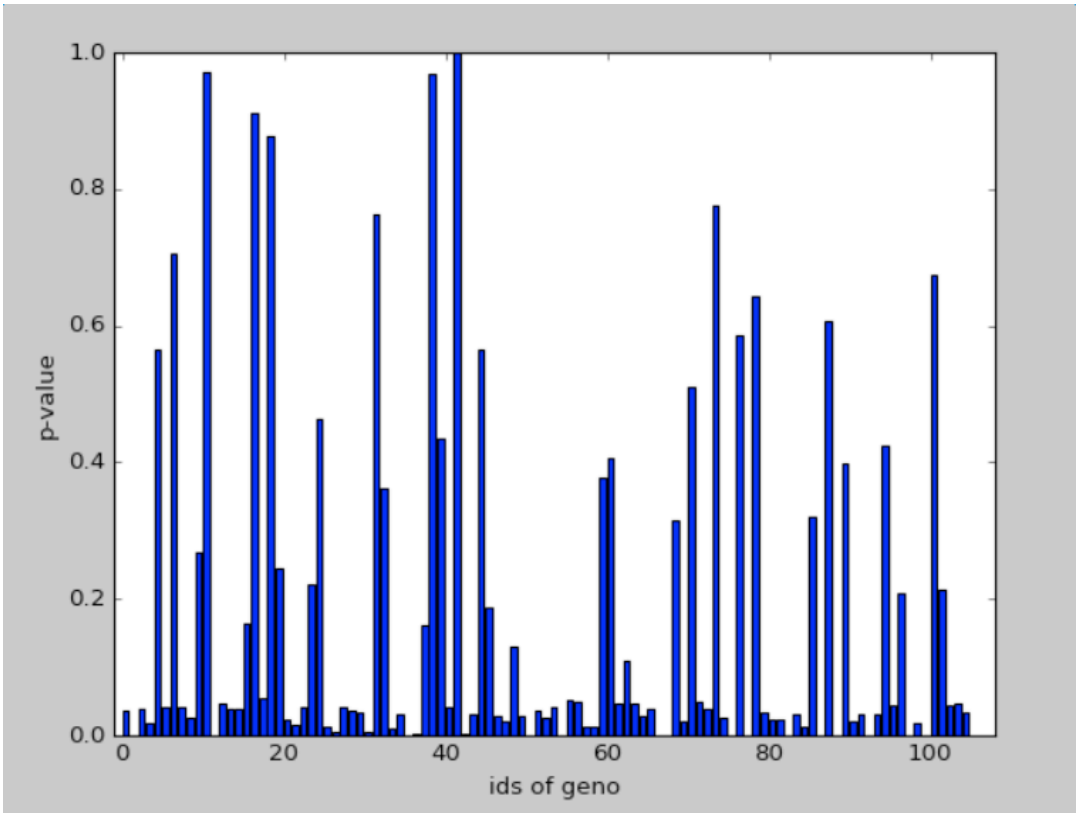

图 4-8 包含阳性位点的 105 个基因的 P 值

将显著水平设为 0.01，则 P 值小于 0.01 的基因有下表 4-7 所示：

表 4-7 疾病 A 的可能致病基因

基因名称	卡方值	P 值
gene_55	24016.26	0
gene_217	14720.63	0
gene_293	8584.559	0
gene_169	1548.839	0
gene_113	52.37188	4.59E-13
gene_115	41.37998	1.25E-10
gene_3	28.16968	1.11E-07
gene_67	27.92394	1.26E-07
gene_103	27.79321	1.35E-07
gene_114	22.12656	2.55E-06
gene_274	21.98909	2.74E-06

gene_102	20.95414	4.70E-06
gene_245	15.03672	0.00010544
gene_150	13.97119	0.000185634
gene_144	13.46283	0.000243337
gene_121	9.426932	0.002138217
gene_17	8.750487	0.003095193
gene_235	8.20155	0.004185461
gene_218	7.926829	0.004870708

3) 权重累加法的求解

使用 Java 代码实现，代码描述如下：

```
getGenoWeightSortList( k )
    构建 <位点名称-下标-卡方值>映射表lookUpTable
    读取 <基因-相关位点>映射文件，构造映射map
    初始化list
    for map中的每个基因
        通过lookUpTable查询对应位点的卡方值并累加
        将<基因，累加权重值>加入到list中
    对list按照累加权重值进行降序排列
    return 权重最高的k个基因信息
end
```

图 4-9 算法伪代码

排序后的最相关的 10 个基因及相关系数如下：

表 4-8 权重累加法结果

基因	相关系数
gene_217	42.48
gene_293	38.64
gene_55	37.52
gene_115	26.73
gene_169	26.00
gene_102	20.95
gene_245	19.01

gene_3	18.86
gene_30	18.24
gene_106	15.89

将卡方检验的结果与权重累加法结果进行对比，综合得出，与疾病 A 最相关的基因为 gene_55，其他较为相关的基因为 gene217、293、169。

4.3.5 问题三模型检验

使用 SPSS 进行 logistics 回归分类，其结果如下：

表 4-9 问题三的 Logistic 回归检验

分类表[a]

已观测			已预测		
			label		
			0	1	百分比校正
步骤 1	label	0	402	98	80.4
		1	103	397	79.4
	总计百分比				79.9

a. 切割值为 .500

由上图可得，分类的准确为 79.9%，分类效果较好，故由疾病与基因关联性模型求解得到结果可信度较高，即基 gene_55 与疾病 A 最相关，其他较为相关的基因为 gene217、293、169。

4.4 问题四的求解

4.4.1 位点与多性状关联性分析模型

为了将 10 个形状看作一个整体，采用典型相关分析法将 10 个相关联的性状转化为一个新的综合变量（或称为典型变量），具体采用的是 K-means 方法对 1000 个样本的 10 个性状的表现型进行聚类。然后利用卡方检验等方法分析 SNP 位点与该综合性状之间的关联关系，并进行多重检验的方法，最终选取与这 10 个性状都显著相关的致病位点。

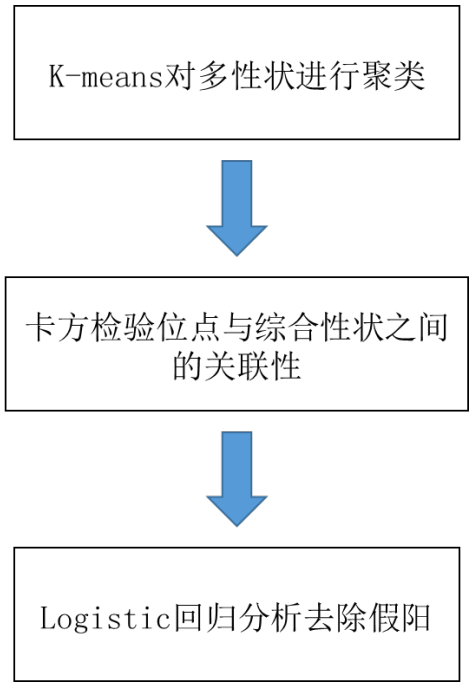

图 4-10　位点与多性状关联性分析模型

4.4.2 K-means 聚类

本文采用 K-means 聚类，K-means 算法是硬聚类算法，是典型的基于原型的目标函数聚类方法的代表，它是数据点到原型的某种距离作为优化的目标函数，利用函数求极值的方法得到迭代运算的调整规则。K-means 算法以欧式距离作为相似度测度，它是求对应某一初始聚类中心向量 V 最优分类，使得评价指标 J 最小。算法采用误差平方和准则函数作为聚类准则函数。

$$J=\sum_{i=1}^{k}\sum_{x_j\in S_i}\left(x_j-\mu_i\right)^2 \tag{4-10}$$

算法过程如下：

1）从 N 个样本随机选取 K 个样本作为质心

2）对剩余的每个文档测量其到每个质心的距离，并把它归到最近的质心的类

3）重新计算已经得到的各个类的质心

4）迭代 2～3 步直至新的质心与原质心相等或小于指定阈值，算法结束。

4.4.3 K 值的求解

K-means 方法的第一步是确定 K 值，首先做出 K 取 2-10 范围的残差平方和图 4-11，从图中可以得到，选取 K=2 时，聚类效果最好。

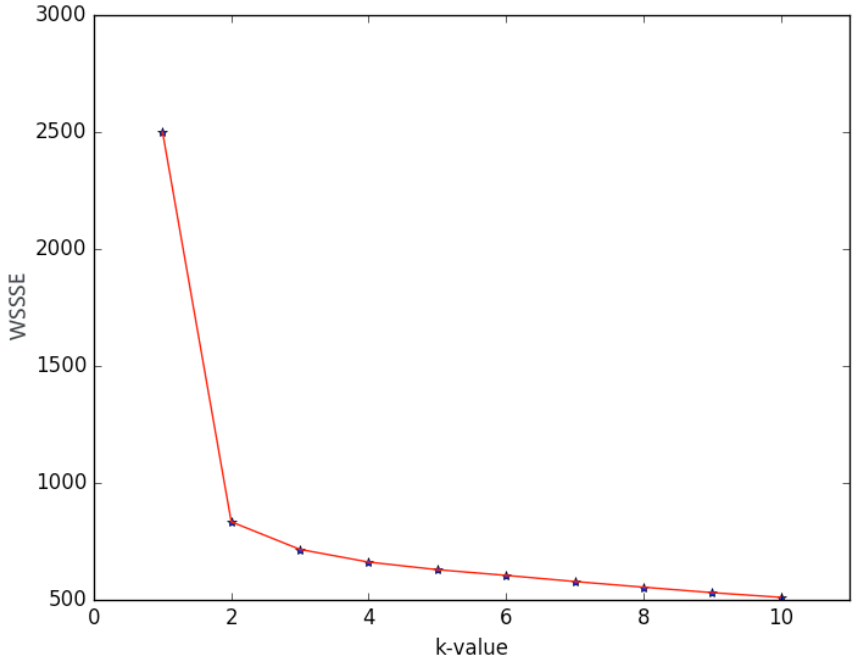

图 4-11 K 取 2-10 范围的残差和图

K 值取 2，利用 Python 的 scikit-learn 机器学习包对 1000 个样本的个 10 性状表现进行聚类，结果如下表 4-10

表 4-10 K-means 聚类部分结果

样本编号	10 种性状表现										聚类结果
1	0	0	0	0	0	0	0	0	0	0	0
2	1	1	1	1	1	0	0	0	1	1	1
3	0	0	1	0	1	1	1	1	0	0	0
…	…										…
1000	0	0	0	0	0	0	0	0	0	0	0

综合 1000 个样本的性状信息和样本的位点编码信息，得到表 4-11

表 4-11 位点与综合性状关系表

样本编号	位点名称及碱基对信息				综合性状
	rs3094315	rs3131972	...	rs7545865	
1	0	1	...	1	0
2	1	1	...	0	1
…	…				…

1000	1	1		2	0

至此，问题转化为类似问题二的形式，及只需要分析位点与综合性状之间的关联性即可。

4.4.4 位点与综合变量关联性的卡方检验

类似问题二，卡方检验过程：

1) 零假设 H_0：综合性状表现与位点 rs10157835 无关

2) 确定自由度为(3-1)×(2-1)=2，选择显著水平 α=0.005。

3) 对假设 H_0 进行卡方检验，获得的结果如下：

从表 5-2 卡方检验的结果中 rs10157835 对应数据可以看出：P 值远小于 0.005, 因此可以认为假设不成立，即位点 rs10157835 不同的基因型对综合性状表现型有显著性差别，即此位点很可能是综合性状表现为患病的致病位点之一。

使用 Python 的机器学习工具包 scikit-learn 对所有位点是否使综合性状表现为患病进行卡方检验，按卡方值高低排序，选择 P 值远小于 0.005 的部分，得到如下可能致病位点的下标位置(下标从 0 开始)，卡方值和 P 值如表 4-12：

表 4-12 疾病 A 可能的致病位点

位点	位置	卡方值	P 值
rs10157835	1338	14.44	0.0001
rs12746773	4568	11.63	0.0007
rs3218121	7180	10.89	0.0010
rs3128318	251	9.70	0.0018
rs2147905	460	8.78	0.0030
rs2526830	5067	8.70	0.0032
rs6698317	4785	8.62	0.0033
rs4654487	979	8.48	0.0036
rs351617	1049	8.29	0.0040
rs7515988	9373	8.08	0.0045
rs9659352	5296	8.07	0.0045

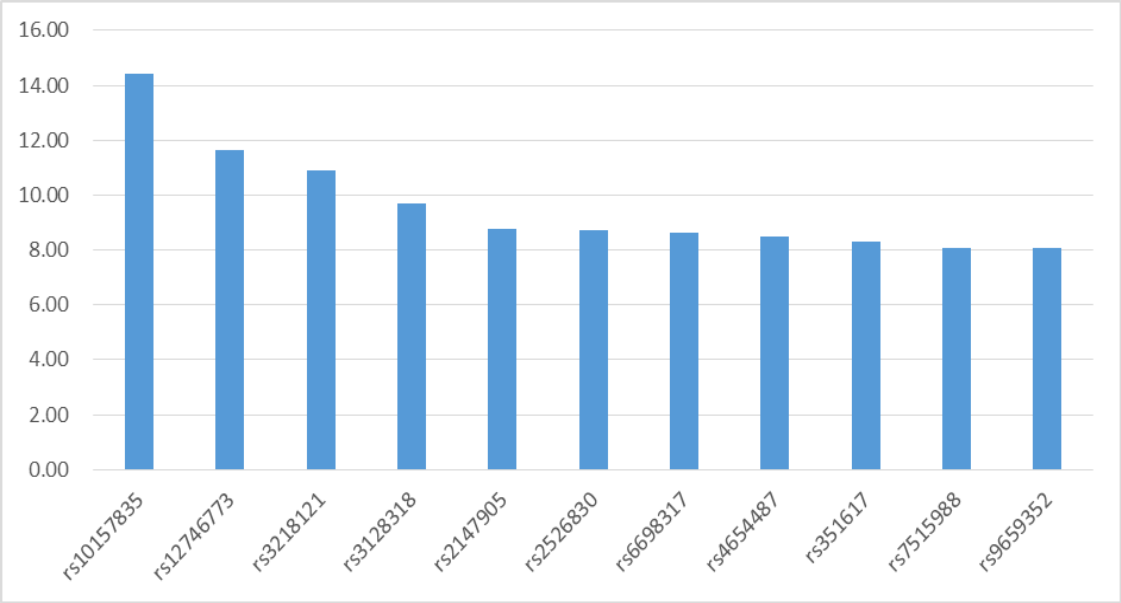

图 4-12 综合性状的可能治病位点的卡方值

从图 5-1 中按卡方值对 P 值小于 0.005 的位点进行排序的结果。本题依然以卡方值表征位点与综合性状关联性的近似量化结果，得出结论：

位点 rs10157835 为表现综合性状的致病位点，位点 rs12746773 和位点 rs3218121 极有可能为表现综合性状的致病位点，图 4-12 中其余位点有可能为表现综合性状的致病位点。

4.4.5 Logistic 回归检验

使用 SPSS 进行 logistics 回归分类，其结果如下

表 4-13 问题四的 Logistic 回归检验

分类表[a]

已观测			已预测 label 0	已预测 label 1	百分比校正
步骤 1	label	0	332	168	66.4
		1	149	351	70.2
	总计百分比				68.3

a. 切割值为 .500

由上表 4-13 可得，分类的准确为 68.1%，分类效果相比较于单因素分析，有较为明显的提升，故与 10 个性状相关的位点结果可信度较高，即位点 rs10157835 为表现综合性状的致病位点，位点 rs12746773 和位点 rs3218121 极有可能为表现综合性状的致病位

点。

5. 模型评价与推广

5.1 模型的评价

5.1.1 模型的优点

1) 对于位点编码，根据问题中每个位点只有 3 种组合方式的实际情况，选用考虑距离含义的数值编码方式，建立编码模型。后续实验证明，该编码模型具有良好的适应性，为后文研究奠定了良好的基础。
2) 在寻找与疾病 A 相关的位点时，首先采用了卡方检验模型，在显著性为 0.05 的条件下筛选出十余个位点作为候选位点，有效地减少了工作量，大幅降低了向量维度，给采用 Logistic 回归分析创造了条件。
3) 问题二划归为广义线性回归中的而分类模型，进而确定采用形式最相符合的 Logistic 回归进行结果的检验;
4) 本文在寻找基因与性状间的关系过程中，将基因看作位点的集合建立基因编码模型，给按照位点排列组合及频数给基因赋予了适当的权重。
5) 采用卡方检验为主，权重累加为辅的双重检验方法寻找与疾病 A 最相关的基因，两种方法结果高度相似，从而使结果更具说服力。
6) 在处理关联性状问题时，采用了 K-means 聚类算法将 1000 组性状分类两个类别，从而划归为与问题二类似的问题。

5.1.2 模型的缺点

研究过程中主要考虑单个位点或单个基因与性状之间的关联性，而忽略了位点之间或者基因之间的协同作用。

5.2 模型的推广

本文使用到的卡方检验、Logistic 回归、K-means 聚类等模型已得到较为广泛的应用。本文中的模型不仅适用于位点、基因与性状间的相关性分析,还可以应用在多种类型的层次关联分析问题上。

参考文献

[1] 全基因组关联分析. [E]. 维基百科. https://zh.wikipedia.org/wiki/全基因组关联分析

[2] 统计学习方法. 李航. [M]. 清华大学出版社. 北京. 2012

[3] 统计学：从数据到结论. 吴喜之. [M]. 中国统计出版社. 北京. 2006

[4] 多位点基因多态性分析与乙肝肝硬化风险的分子预测. 张焜和. [D]. 南昌:南昌大学医学院. 2009

[5] 全基因组关联研究中的多重校正方法比较. 黄杨岳，孔祥祯，甄宗雷，刘嘉. [J] 心理科学进展 2013, Vol. 21, No. 10, 1874–1882