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

题 目 具有遗传性疾病和性状的遗传位点分析

摘 要：

本文采用全基因组关联性分析的方法来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，针对相应的数据结构建立多种数学模型：卡方检验（Chi-square test）、多元线性回归模型（multiple linear regression model）、典型相关分析（canonical correlation analysis）等等，并综合利用这几种方法完成了数据的处理。

问题一，每个位点由两种碱基组成，四种碱基共构成 6 种位点类型（A/T，A/G，A/C，G / T，C/T，C/G），每种位点类型共有三种编码方式。位点只与三种编码方式的构成比例和位点类型有关系，与位点内碱基对与顺序、位置无关，所以六种类型都可以由 A/B 表示，三种编码方式 AA、AB、BB 可以分别编码为 0、1、2。

问题二，认为每个位点相互独立，并且不考虑基因的存在，位点与患病之间具有直接关系。样本分为患病组和对照组两组，如果某个位点与患病相关，那么编码方式应该有明显的差异，利用卡方检验和显著性分析来衡量差异。将患病组和对照组的差异性作为衡量位点和患病关联强度的指标。患病组和对照组之间的差异性可以用卡方值来表示，卡方值越大，差异性越大，位点与患病之间关联强度越强。利用 Bonferroni 校正，得到一个比较保守的结论，与患病最相关的位点是 rs2273298。

问题三，为了简化基因与患病的关系，假设基因与患病之间是简单相关，即拥有患病基因就会导致患病，基因内的位点与患病进行多元线性回归分析和显著性检验，利用线性回归分析的残差和显著性强度综合衡量基因与患病之间

的关系。如果残差越小并且显著性水平高，基因与患病之间的关联越强，与患病关联最强的基因是 Gene102 和 Gene217。

问题四，首先需要分析多种性状进行初步统计，发现性状之间的具有很强的相关性，可以进一步降低性状的维度。然后筛选与性状无关的性状，降低位点的维度，得到候选名单，为下一步计算做准备。最后，问题转化为多个性状与候选名单内的位点之间的关系。由于多个性状是一个整体，不可分割，相当于求全局最优解，采用典型相关分析，求解其中的典型关系，认为回归直线的系数比较大的位点对性状具有比较大的影响，并运用显著性检验进行验证，确保假设成立。找到与 10 个性状相关性最大的位点 rs12746773。

关键词：遗传统计学，全基因组关联性分析(GWAS)，位点(SNPs)，卡方检验，多元线性回归，典型相关分析

目　录

一. 问题重述

1.1 研究背景

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、G、C 四种三项的脱氧核苷酸链接组成的双螺旋长链分子。如图 1.1 所示，碱基 A、T、G、C 即为位点，具有特定功能的 DNA 双螺旋片段就形成了基因，DNA 双螺旋与组蛋白缠绕形成核小体，进一步形成染色体。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，或者与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。

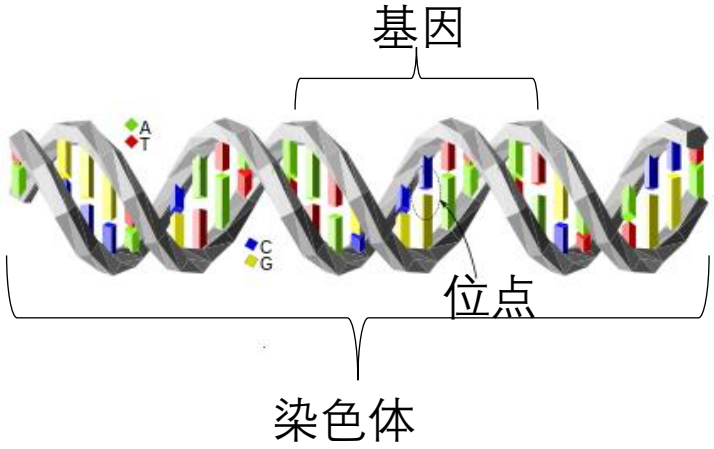

图1.1 染色体、基因和位点的结构关系

随着人类基因组计划和基因组单倍体图谱计划实施和飞速发展，一种新的性状/复杂疾病的遗传分析策略——全基因组关联分析（Genome Wide Association Study, GWAS）是目前研究复杂性状的主要方法。GWAS 是应用人类基因组中数以百万计的 SNPs 为标记进行病例对照关联分析，以期发现影响复杂性疾病发生的遗传特征的一种新策略[8-12]。2005 年 Science 杂志首次报道了年龄相关性视网膜黄斑变性 GWAS 结果，在医学界和遗传学界引起了极大的轰动，此后陆续出现了有关冠心病、肥胖、2 型糖尿病、甘油三酯、精神分裂症等的研究报道。截至 2013 年 12 月，已经报道了包括糖尿病等在内的 17 个复杂性性状的 GWAS 结果，见图 1.2，这些研究不但很好地重复发现了过去已证实的复杂性疾病或性状相关联信号，而且还产生了很多新的候选基因，这些研究成果极大地推进了性疾病或疾病的遗传学研究进展。

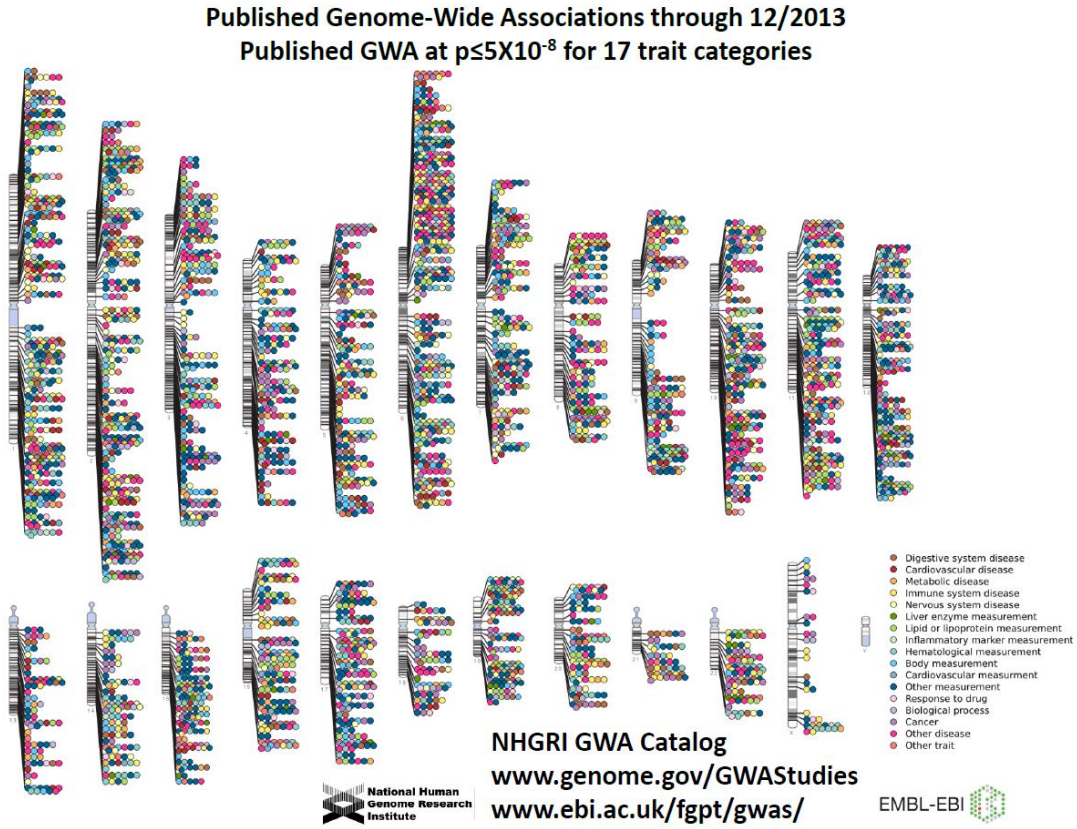

图1.2 全基因组关联分析研究结果(图中圆圈表示易感位点,不同颜色表示不同的性状/疾病)

1.2 研究问题

问题一：本题提供了“genotype.dat”文件，文件中包含了 1000 个样本在某条染色体片段上所有的位点信息，如表 1 所示第三列到第六列的编码信息。在“genotype.dat”文件中，共有 1001 行，9445 列，第一行为 9445 个位点的名称，下面的每一行表示一个样本在该条染色体片段上所有位点的编码信息。请用适当的方法，把“genotype.dat”文件中的每个位点的碱基编码方式转化成数值编码方式，以便进行数据分析。

表 1-1 染色体片段及位点信息

样本编号	样本健康状况	染色体片段位点名称和位点等位基因信息			
		rs100015	rs56341	...	rs21132
1	1	TT	CA	...	GT
2	0	TT	CC	...	GG
3	1	TC	CC	...	GG
4	1	TC	CA	...	GG
5	0	CC	CC	...	GG
6	0	TT	CC	...	GG

注：位点名称通常以 rs 开头。

问题二：大量的研究表明，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，因此定位与性状或疾病相关联的位点在染色体或者基因中的位置，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，从而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关联分析的方法来确定致病位点，对于每个样本，采用碱基

（A、T、G、C）的编码方式来获取每个位点的信息，如表 1 中的 rs100015 位置，不同样本的编码都是 T 和 C 的组合，有三种不同的编码方式 TT、TC 和 CC，类似地其他的位点虽然碱基的组合不同，但也只有三种不同编码。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。本题提供了“phenotype.txt”文件，里面包含了样本具有遗传疾病 A 的信息，即一列 0 和 1 组成的数据，其中共有 500 个 0，表示的是 500 个没患疾病 A 的人，500 个 1 表示的是 500 个患有遗传病 A 的人，如表 1 中的第 2 列所示。请设计或采用一种方法，找出遗传疾病 A 最有可能一个或几个致病位点，并给出相关的理论依据。

问题三：由于可以把基因理解为若干个位点组成的集合，所以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表现出来，根据问题二提供的“phenotype.txt”文件和本题提供的“gene_info”文件夹，找出与疾病最有可能相关的一个或几个基因，并说明理由。其中，“geno_info”文件夹中有 300 个“. dat”文件，每个“. dat”文件中包含了若干个位点的名称，表示该基因包含的位点信息。例如 gene_1.dat，表示基因 gene_1 包含了 rs3094315, rs3131972, ... , rs4040617，共 7 个位点。

问题四：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，研究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。附录中的“multi_phenos.txt”文件中提供了 1000 个样本的 10 种相关性状的信息，文件中的每一列表示一个性状，每一行对应一个样本。试根据文件给出的信息找出与 10 个性状有关联的位点。

二. 基本假设

1. 位点之间相互独立，位点与 XX 有关，与 YY 无关（见问题一的求解）；
2. 问题二中，假设有致病位点即患病；
3. 假设位点与基因是线性关系；
4. 问题三中，假设有致病基因即患病；
5. 问题四与疾病 A 无关。

三. 问题分析

3.1 问题一的分析

问题一要求将“genotype.dat”中的每个位点编码方式转化成数值编码方式，其中“genotype.dat”文件中的数据共有 1000 行，代表 1000 个样本，9445 列代表每个样本中含有 9445 个位点。四种碱基 A、T、C、G 共可组成 10 种类型，如图 3.1 所示。每个位点由两种碱基组成，四种碱基共构成 6 种位点类型（A/T，A/G，A/C，G / T，C/T，C/G），每种位点类型共有三种编码方式，如 A/T 类型有 AA 、AT、TT 三种编码方式。

在题 2、题 3 和题 4 中，位点只与三种编码方式的比例和位点类型有关系，与位点内碱基对与顺序、位置无关，所以六种类型都可以由 A/B 表示，三种编码方式 AA、AB、BB 可以分别编码为 0、1、2。

	A	T	C	G
A	AA			
T	TA	TT		
C	CA	CT	CC	
G	GA	GT	GC	GG

图 3.1 碱基组成类型

3.2 问题二的分析

问题二要求根据 1000 个样本的 9445 个位点信息和患病信息，找到与遗传疾病 A 相关的最可能的一个或者多个位点。生物学上，位点和患病之间具有非常复杂的关系，位点之间相互作用，位点还与基因相关等复杂因素。为了简化分析，认为每个位点相互独立，并且不考虑基因的存在，位点与患病之间具有直接关系。

如果将位点和患病当作一个函数映射关系，那么自变量是 9445 个相互独立的位点，因变量是患病。每个自变量与因变量的关联强度不同，关联比较大的位点可以认为是最可能的与患病相关的位点。1000 个样本分为患病组和对照组两组，如果某个位点与患病不相关，那么位点三种编码方式在患病组和对照组分布几乎一致；如果某个位点与患病相关，那么编码方式应该有明显的差异。将患病组和对照组的差异性作为衡量位点和患病关联强度的指标。

患病组和对照组之间的差异性可以用卡方值来表示，卡方值越大，差异性越大，位点与患病之间关联强度越强。设定一定的阈值找出相关性较大的位点。运用显著性检验来对总体分布进行假设，然后用样本信息来验证这个是否合理。

3.3 问题三的分析

问题三在问题二的基础上，提供了 300 个基因，基因是由若干个位点组成的集合，患病与其中一个或者几个基因相关。如图 3.2 所示：

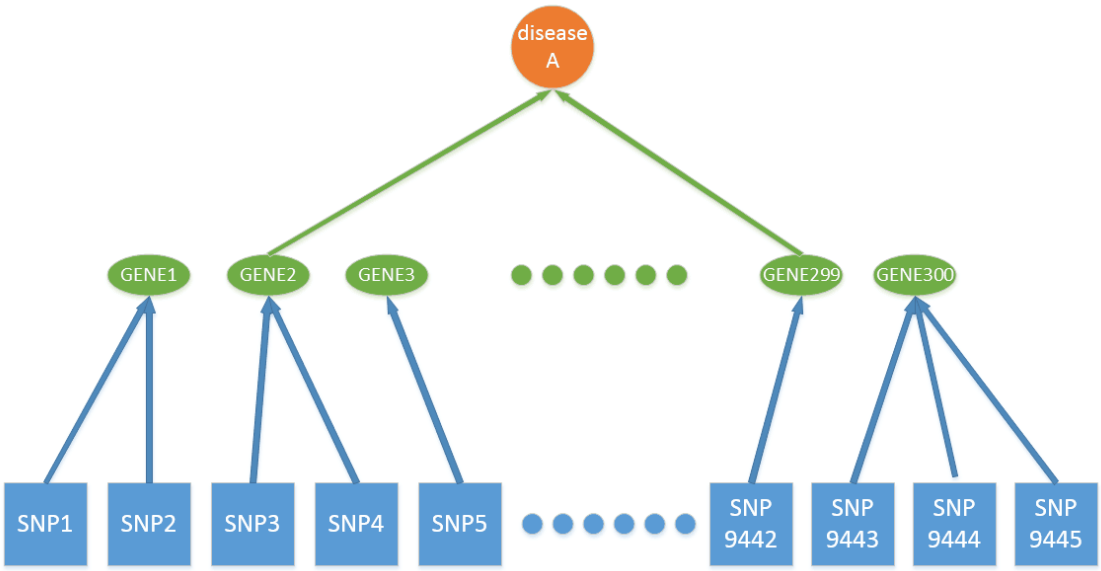

图 3.2 疾病 A、基因、位点关系图

生物学上，位点与基因之间具有相关性，但是相关关系复杂；基因与患病之间也具有复杂的相关性。为了简化位点和基因的关系，假设基因内的位点相互独立，基因和位点具有简单的线性关系，假设基因 i 是由 n 个位点构成，即

$$Gene_i = a_0 + a_1 SNP_1 + a_2 SNP_2 + ... + a_n SNP_n \qquad (3\text{-}1)$$

a_n 可以理解为不同位点对基因的贡献系数，贡献系数越大，说明位点与基因的相关越强，忽略贡献系数较小的位点，认为基因主要由贡献系数大的位点。

为了简化基因与患病的关系，假设基因与患病之间是简单相关，不存在基因之间的相互抑制等复杂关系。假设基因与患病之间存在相关，即拥有患病基因就会导致患病，基因内的位点与患病进行多元线性回归分析，将线性回归分析的残差作为基因与患病之间的关系。如果残差越小，说明拟合效果越好，基因与患病之间的关联越强，找到与患病关联最强的几个基因进行显著性检验。

3.4 问题四的分析

首先需要分析多种性状进行初步统计，并求解其中的相关性，性状之间的具有很强的相关性。

然后由于绝大部分位点与性状无关，需要进行筛选与性状无关的性状，降低位点的维度，为后续的计算做准备。利用题 2 的计算方法计算每个性状与位点之间的关系。如果位点与每个性状之间的关联性都很小，可以认为此位点一定不与多种性状相关，这是一种比较保守的筛选策略，可以筛选掉与所有性状都无关的位点，得到一个含有患病位点的候选名单，降低位点的维度。

最后，问题转化为多个性状与候选名单内的位点之间的关系。由于多个性状是一个整体，不可分割，相当于求全局最优解，题 2 可以认为是单个性状与多个位点之间的关系，相当于求局部最优解，局部最优解不一定是全部最优解，

全局最优解不是局部最优解的叠加。从函数的角度来看，候选名单内的位点是自变量，性状是因变量，假设位点与性状之间存在线性关系。对于 n 个位点，存在如下的关系式：

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \cdots \\ Y_i \end{bmatrix} = \begin{bmatrix} H_{1,1} & H_{1,2} & \cdots & H_{1,n} \\ H_{2,1} & H_{2,2} & \cdots & H_{2,n} \\ \cdots & \cdots & \cdots & \cdots \\ H_{10,1} & H_{10,2} & \cdots & H_{10,n} \end{bmatrix} \begin{bmatrix} SNP_1 \\ SNP_2 \\ \cdots \\ SNP_n \end{bmatrix} \quad (3\text{-}2)$$

采用典型相关分析，求解其中的典型关系，认为回归直线的系数比较大的位点对性状具有比较大的影响，并运用显著性检验进行验证，确保假设成立。

四. 问题一的模型建立与求解

4.1 解题思路概述

问题一要求将“genotype.dat”中的每个位点编码方式转化成数值编码方式，其中“genotype.dat”文件中的数据共有 1000 行，代表 1000 个样本，9445 列代表每个样本中含有 9445 个位点。四种碱基 A、T、C、G 共可组成 10 种类型，如图 4.1 所示。每个位点由两种碱基组成，四种碱基共构成 6 种位点类型（A/T，A/G，A/C，G / T，C/T，C/G），每种位点类型共有三种编码方式，如 A/T 类型有 AA 、AT、TT 三种编码方式。

在题 2、题 3 和题 4 中，位点只与三种编码方式的比例和位点类型有关系，与位点内碱基对与顺序、位置无关，所以六种类型都可以由 A/B 表示，三种编码方式 AA、AB、BB 可以分别编码为 0、1、2。

	A	T	C	G
A	AA			
T	TA	TT		
C	CA	CT	CC	
G	GA	GT	GC	GG

图 4.1 碱基组成类型

4.2 转化映射表

转化映射表如表 4-2 所示：

表 4-2 转化映射表

位点类型	编码方式		
A/T	AA(0)	AT(1)	TT(2)
A/G	AA(0)	AG(1)	GG(2)
A/C	AA(0)	AC(1)	CC(2)
C/G	CC(0)	CG(1)	GG(2)
C/T	CC(0)	CT(1)	TT(2)
G/T	GG(0)	GT(1)	TT(2)

4.3 流程图

编程流程图如图 4.2 所示：

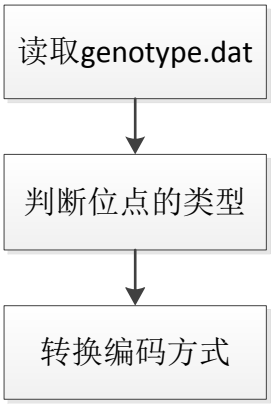

图 4.2 流程图

五. 问题二的模型建立与求解

5.1 解题思路概述

问题二要求根据 1000 个样本的 9445 个位点信息和患病信息，找到与遗传疾病 A 相关的最可能的一个或者多个位点。生物学上，位点和患病之间具有非常复杂的关系，位点之间相互作用，位点还与基因相关等复杂因素。为了简化分析，认为每个位点相互独立，并且不考虑基因的存在，位点与患病之间具有直接关系。

如果将位点和患病当作一个函数映射关系，那么自变量是 9445 个相互独立的位点，因变量是患病。每个自变量与因变量的关联强度不同，关联比较大的位点可以认为是最可能的与患病相关的位点。1000 个样本分为患病组和对照组两组，如果某个位点与患病不相关，那么位点三种编码方式在患病组和对照组分布几乎一致；如果某个位点与患病相关，那么编码方式应该有明显的差异。将患病组和对照组的差异性作为衡量位点和患病关联强度的指标。

患病组和对照组之间的差异性可以用卡方值来表示，卡方值越大，差异性越大，位点与患病之间关联强度越强。设定一定的阈值找出相关性较大的位点。运用显著性检验来对总体分布进行假设，然后用样本信息来验证这个是否合理。

5.2 数据预处理

每个样本的位点数据对结果并不明显，需要将所有样本进行统计才能发现规律。位点的信息可以由其三种编码方式的 AA、AB 和 BB 的构成比例表达。样本可以分为患病组（cases）和对照组（controls），每组分别具有样本 500 例，需要分别统计患病组和对照组的每种位点的三种编码方式的分布情况，用三种编码方式的比例来表达位点信息。

5.3 利用卡方值来检验位点相似度

卡方检验是以 χ^2 分布为基础的一种常用假设检验方法，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 χ^2 值，它表示观察值与理论值之间的偏离程度。根据 χ^2 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序太大，应当拒绝假设 H_0，表示比较资料之间有显著差异，否则不能拒绝 H_0。

本题中我们假设：

H_0：位点 x 为最有可能的致病基因；

H_1：位点 x 不是最有可能的致病基因

A 代表某个类别的观察频数，可以做出如表 5-1 所示的统计结果表（以 A/T 基因型为例）。

表 5-1　A/T 基因型的统计结果表

	AA	AT	TT	合计
患病				
不患病				
合计				

计算这种偏离程度的公式如下所示：

$$\chi^2 = n\left(\sum \frac{A_{RC}}{n_R n_C} - 1\right) \qquad (5\text{-}1)$$

其中，R 为行数，C 为列数，A_{RC} 为相应的观察频数，n_R 为对应第 R 行的总频数，n_C 为对应第 C 列的总频数，。

由卡方的计算公式可知，当观察频数与期望频数完全一致时，x^2 值为 0；观察频数与期望频数越接近，两者之间的差异越小，x^2 值越小；反之，观察频数与期望频数差别越大，两者之间的差异越大，x^2 值越大。换言之，大的 x^2 值表明观察频数远离期望频数，即表明远离假设。小的 x^2 值表明观察频数接近期望频数，接近假设。因此，x^2 是观察频数与期望频数之间距离的一种度量指标，也是假设成立与否的度量指标。如果 x^2 值大，说明实际频数与理论频数相差太大，超出了误差允许的范围，从而怀疑 H_0 的正确性，继而拒绝 H_0，接受其对立假设 H_1，我们称这类错误为第一类错误；如果 x^2 值小，研究者就倾向于不拒绝 H_0，称为第二类错误。任何假设检验问题都存在 4 种可能的情况确定我们决策是正确还是错误的，表 5-2 概括了这 4 种情况。

表 5-2 统计假设检验的可能情况

	H_0 为真	H_0 不为真
不拒绝 H_0	正确决定	第二类错误
拒绝 H_0	第一类错误	正确决定

通过计算所有位点在患病组和对照组之间的卡方值，然后换算到 P 值，运用 P 值来进行显著性检测，P 值越大，则患病组和对照组的显著性越强，通过 P 值来衡量患病组和对照组的差异性，差异性越大，则位点越可能与患病相关。如图 5.1 所示，所有位点在患病组和对照组之间的 P 值：

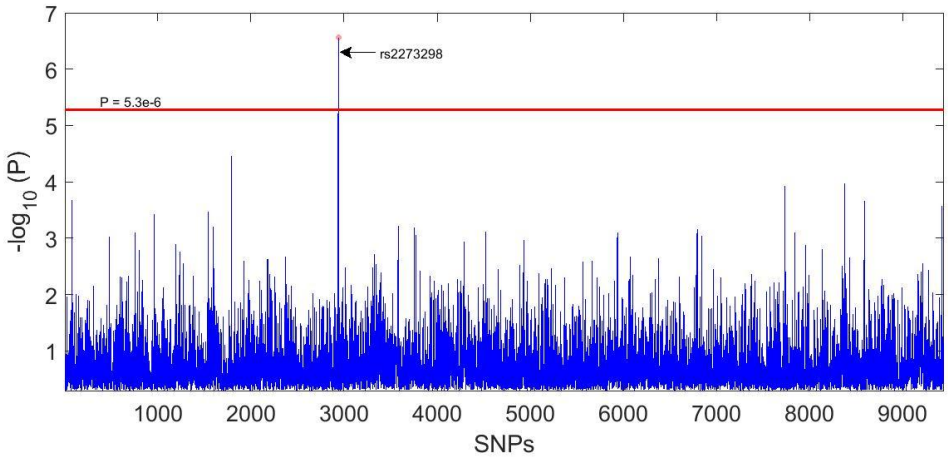

图 5.1 所有位点的 P 值

一般显著性检验设置显著水平 $\alpha = 0.05$ ，假设每个位点之间相互独立，为了保守起见，使用 Bonferroni 校正，P < 0.05 / 9445 = 5.3e-6。在这条比较严格阈值线上的位点只有一个，即 rs2273298。如果不加校正，P<0.05,则-log10(P)>1.3 都满足显著性检验，明显严重放宽了相关的判断阈值。

由此，可以得到结论，与患病显著性相关的位点是 rs2273298，在这 9445 个位点中，此位点的与患病相关性最大，并且满足显著性检验。

六. 问题三的模型建立与求解

6.1 解题思路概述

问题三在问题二的基础上，提供了 300 个基因，基因是由若干个位点组成的集合，患病与其中一个或者几个基因相关。如图 6.1 所示：

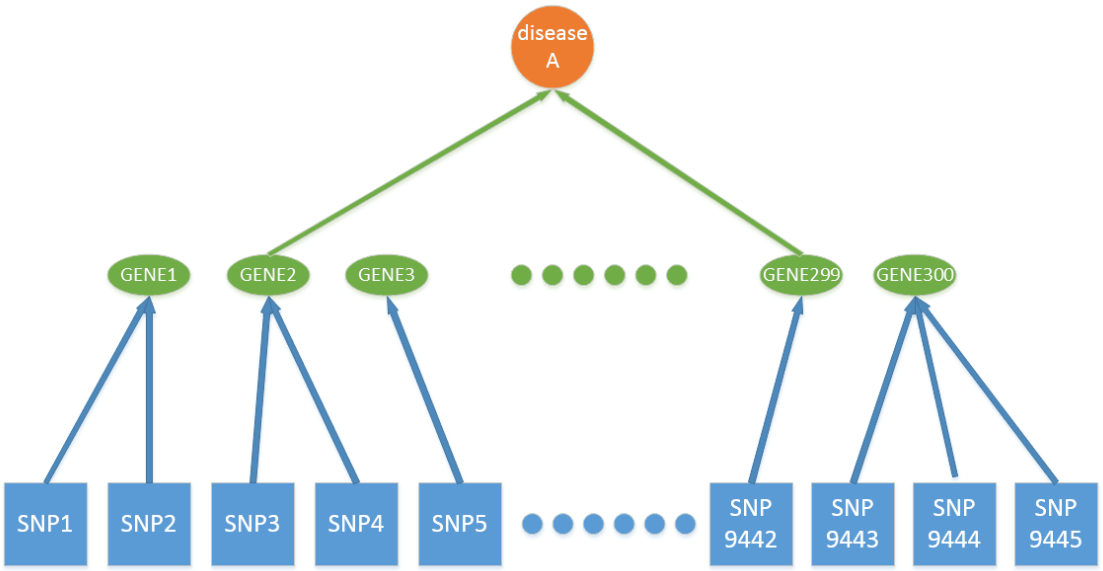

图 6.1 疾病 A、基因、位点关系图

生物学上，位点与基因之间具有相关性，但是相关关系复杂；基因与患病之间也具有复杂的相关性。为了简化位点和基因的关系，假设基因内的位点相互独立，基因和位点具有简单的线性关系，假设基因 i 是由 n 个位点构成，即

$$Gene_i = a_0 + a_1 SNP_1 + a_2 SNP_2 + ... + a_n SNP_n \qquad (6\text{-}1)$$

a_n 可以理解为不同位点对基因的贡献系数，贡献系数越大，说明位点与基因的相关越强，忽略贡献系数较小的位点，认为基因主要由贡献系数大的位点。

为了简化基因与患病的关系，假设基因与患病之间是简单相关，不存在基因之间的相互抑制等复杂关系。假设基因与患病之间存在相关，即拥有患病基因就会导致患病，基因内的位点与患病进行多元线性回归分析，将线性回归分析的残差作为基因与患病之间的关系。如果残差越小，说明拟合效果越好，基因与患病之间的关联越强，找到与患病关联最强的几个基因进行显著性检验。

6.2 多元线性回归模型

设 $Gene_i$ 为随机变量为 y，位点为非随机因素 x。为了简化位点和基因的关系，假设基因内的位点相互独立，基因和位点具有简单的线性关系，假设基因 i 是由 n 个位点构成，即

$$Gene_i = a_0 + a_1 SNP_1 + a_2 SNP_2 + ... + a_n SNP_n \qquad (6\text{-}2)$$

a_n 可以理解为不同位点对基因的贡献系数，贡献系数越大，说明位点与

基因的相关越强，忽略贡献系数较小的位点，认为基因主要由贡献系数大的位点。

在本问题中，我们建立了多元线性回归模型 $y=b_0+b_1x_1+b_2x_2+\cdots+b_nx_n$ 去拟合样本数据，从而得到参数 $b_0, b_1, b_2, \cdots b_n$ 的最小二乘估计。

利用显著性检验原理进行验证，假设 H_0：基因和位点有线性关系；H_1：基因和位点没有线性关系。首先假设 H_0 成立，根据 χ^2 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，应当拒绝假设 H_0，表示比较资料之间有显著差异，否则不能拒绝 H_0。

分布求解每个基因与其组成位点的多元线性模型，得到 P 值检验，可以看出 Gene102 最符合显著性检验，Gene217 和 Gene245 次之。计算结果如图 6.2 所示：

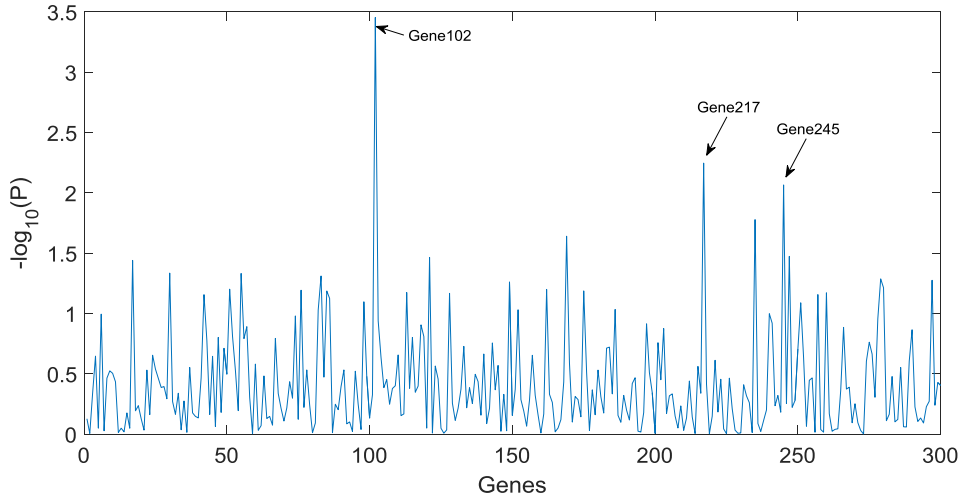

图 6.2　300 个基因的 P 值检验图

多元线性回归的拟合误差的方差可以衡量拟合的效果，方差越小，可以认为拟合效果越好，Gene102 拟合效果最好，Gene55、Gene217 和 Gene247 的拟合效果次之。计算结果如图 6.3 所示：

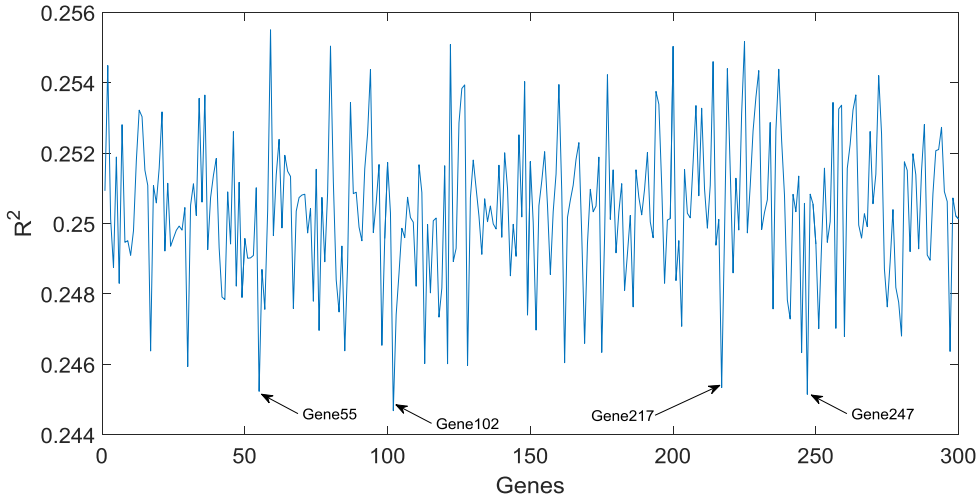

图 6.3　Gene102、55、217、247 拟合效果图

综合上面的两方面分析，Gene102 显著性水平最高，并且拟合方差最小，可以认为此基因与患病之间相关性最强，并且题 2 中与患病关系最强的位点 rs2273298 也在 Gene102，可以判断 Gene102 与患病之间相关性最强。此外，Gene217 的显著性水平次之，拟合方差较小，也可以认为 Gene217 与患病之间相关。由此，与患病相关的基因是 Gene102 和 Gene217。

七. 问题四的模型建立与求解

7.1 解题思路概述

首先需要分析多种性状进行初步统计，并求解其中的相关性，性状之间的具有很强的相关性。

然后由于绝大部分位点与性状无关，需要进行筛选与性状无关的性状，降低位点的维度，为后续的计算做准备。利用题 2 的计算方法计算每个性状与位点之间的关系。如果位点与每个性状之间的关联性都很小，可以认为此位点一定不与多种性状相关，这是一种比较保守的筛选策略，可以筛选掉与所有性状都无关的位点，得到一个含有患病位点的候选名单，降低位点的维度。

最后，问题转化为多个性状与候选名单内的位点之间的关系。由于多个性状是一个整体，不可分割，相当于求全局最优解，题 2 可以认为是单个性状与多个位点之间的关系，相当于求局部最优解，局部最优解不一定是全部最优解，全局最优解不是局部最优解的叠加。从函数的角度来看，候选名单内的位点是自变量，性状是因变量，假设位点与性状之间存在线性关系。对于 n 个位点，存在如下的关系式：

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \cdots \\ Y_i \end{bmatrix} = \begin{bmatrix} H_{1,1} & H_{1,2} & \cdots & H_{1,n} \\ H_{2,1} & H_{2,2} & \cdots & H_{2,n} \\ \cdots & \cdots & \cdots & \cdots \\ H_{10,1} & H_{10,2} & \cdots & H_{10,n} \end{bmatrix} \begin{bmatrix} SNP_1 \\ SNP_2 \\ \cdots \\ SNP_n \end{bmatrix} \quad (7\text{-}1)$$

采用典型相关分析，求解其中的典型关系，认为回归直线的系数比较大的位点对性状具有比较大的影响，并运用显著性检验进行验证，确保假设成立。

7.2 性状分析

经过统计，每个性状的患病组（cases）和对照组（controls）的数量各是 500 例。分别统计每个样本的含有与显性性状数量的个数，统计表如下：

表 7-1 样本性状统计表

显性性状数量	样本个数	显性性状数量	样本个数
0	300	6	27
1	76	7	36
2	48	8	50
3	32	9	80
4	27	10	294
5	30		

样本主要分布在全部隐形和全部显性上，分布以显性性状数量 5 为中心呈现对称分布。样本的分布具有一定的规律性，可能不同性状具有很强的相关性，计算不同性状之间的相关系数，得到相关系数如下：

	1	2	3	4	5	6	7	8	9	10
1	1	0.7160	0.7400	0.7080	0.6840	0.6800	0.7400	0.7120	0.7440	0.7280
2	0.7160	1	0.7120	0.6880	0.6800	0.6920	0.7160	0.7320	0.7360	0.7480
3	0.7400	0.7120	1	0.7000	0.7240	0.6920	0.7240	0.7280	0.7120	0.7480
4	0.7080	0.6880	0.7000	1	0.6840	0.6760	0.7200	0.7320	0.7200	0.7120
5	0.6840	0.6800	0.7240	0.6840	1	0.7480	0.6800	0.7080	0.6680	0.6920
6	0.6800	0.6920	0.6920	0.6760	0.7480	1	0.6840	0.7080	0.6760	0.6680
7	0.7400	0.7160	0.7240	0.7200	0.6800	0.6840	1	0.7360	0.7200	0.7600
8	0.7120	0.7320	0.7280	0.7320	0.7080	0.7080	0.7360	1	0.7280	0.7600
9	0.7440	0.7360	0.7120	0.7200	0.6680	0.6760	0.7200	0.7280	1	0.7440
10	0.7280	0.7480	0.7480	0.7120	0.6920	0.6680	0.7600	0.7600	0.7440	1

图 7.1　不同性状之间的相关系数表

通过图 7.1 可以看出，性状之间的相关系数都比较大，绝大部分都超过 0.7，说明性状之间的相关性比较强，很有可能所有性状都与一个或者几个位点具有一定的相关性。

7.3 位点筛选

由于绝大部分位点与性状无关联，需要对位点进行粗筛选。　　借鉴题 2 的算法，分布计算每种性状与所有位点之间的关联，以 P 值来衡量性状与位点之间的相关性，P 值越小，显著性越高，则关联性越强。每种性状与所有位点之间的关联图如下所示：

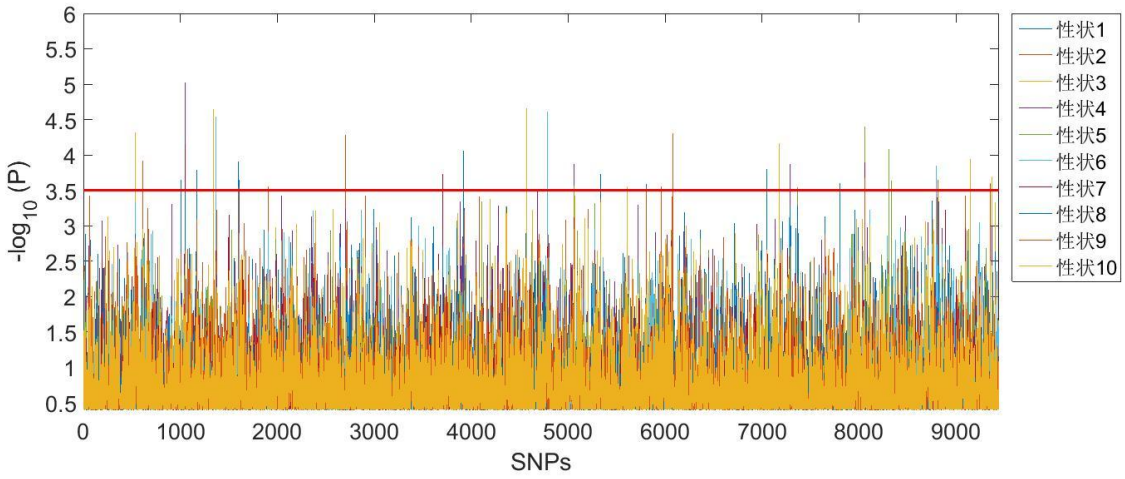

图 7.2　性状与所有位点之间的关联图

从图中可以看出，不同性状的相关位点不同，绝大部分位点与 10 个性状无相关性。需要设立一个筛选条件进行位点筛选，筛选的条件不能过于苛刻以防相关位点被筛选出去，设立一个相对宽松和保守的筛选条件来筛选无关的位点。设置筛选条件为-log10(P)<3.5，即如果某个位点的与 10 个性状的 P 值都满足筛选条件，则可以判定，该位点与 10 个性状整体无相关性。经过筛选可以得到 36 个位点，这 36 个位点进入候选列表，准备进行下一步的检验。

7.4 典型相关分析模型

典型相关分析就是利用综合变量对之间的相关关系来反映两组指标之间的整体相关性的多元统计分析方法，它的基本原理是：在所有的在所有的线性组合中，找一对相关系数最大的线性组合，用这个组合的单相关系数来表示两组变量的相关性，叫做两组变量的典型相关系数，而这两个线性组合叫做一对典型变量。在两组多变量的情形下，需要用若干对典型变量才能完全反映出它们之间的相关性。下一步，再在两组变量的与 u_1、v_1 不相关的线性组合中，找一

对相关系数最大的线性组合，它就是第二对典型变量，而且 $p(u_2，v_2)$就是第二个典型相关系数。这样下去，可以得到若干对典型变量，从而提取出两组变量间的全部信息。

典型相关分析的实质就是在两组随机变量中选取若干个有代表性的综合指标，用这些指标的相关关系来表示原来的两组变量的相关关系。综合变量 U 和 V 分别是两类变量的线性组合：

$$\begin{cases} U = a_1X_1 + a_2X_2 + \ldots + a_pX_p = a'X \\ V = b_1Y_1 + b_2Y_2 + \ldots + b_qY_q = b'Y \end{cases} \quad (7.2)$$

U 和 V 被称为典型相关变量，它们之间的相关系数称为典型相关系数，则有：

$$G_U = Cov(X,U) = Cov(X,a'X) = E(XU') = E(XX'a) = E(XX')a = \Sigma_{11}a \quad (7.3)$$

$$G_V = Cov(Y,V) = Cov(Y,b'X) = \Sigma_{22}b \quad (7.4)$$

G_U、G_V 是衡量原始变量与典型变量相关性的尺度，例如 X_i 与第一典型变量 U 的相关系数 G_U 最大，则表明变量 X_i 与第一典型变量 U 的关系密切，反之则不甚密切。

而待估计的系数 a=（a_1，a_2，…a_p）’和 b=（b_1，b_2，…b_p）’叫做典型系数。当典型相关系数足够大时，可以像回归分析那样，由一组变量的数值预测另一组变量的线性组合的数值。

在本题中，我们寻找的是 36 维的特征向量 X 与 10 个性状 Y 之间的线性关系，其中 $X \in R^n$，$Y \in R^n$，那么可以建立等式 Y=AX 如下：

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix} = \begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1n} \\ w_{21} & w_{22} & \cdots & w_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ w_{m1} & w_{m2} & \cdots & w_{mn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \quad (7.5)$$

其中 $y_i=w_i^Tx$，形式和线性回归一样。然后使用 Pearson 相关系数

$$\rho_{X,Y} = corr(X,Y) = \frac{\mathrm{cov}(X,Y)}{\sigma_X\sigma_Y} = \frac{E\left[\left(X-\mu_X\right)\left(Y-\mu_Y\right)\right]}{\sigma_X\sigma_Y} \quad (7.6)$$

来度量 U 和 V 的关系，我们期望寻求一组最优的解 a、b，使得 Corr(u，v)最大，这样得到的 a 和 b 就是使得 U 和 V 就有最大关联的权重。

此外在典型相关分析中，常常把典型变量对原始变量总方差解释比例的分析以及典型变量对另外一组原始变量总方差交叉解释比例的分析统称冗余分析。在统计上，如果一个变量中的部分方差可以由另外一个变量的方差来解释或预测，就说这个方差部分与另一变量方差相冗余。典型相关分析中的冗余分析是对分组原始变量总变化的方差分析。

使用 SPSS 软件对 36 个位点和 10 个性状之间求解典型相关分析，计算 10 个典型相关的相关系数，相关性从大到小排列，计算结果如下图所示：

Canonical Correlations	
1	.436
2	.387
3	.264
4	.245
5	.217
6	.213
7	.185
8	.167
9	.161
10	.133

图 7.3　36 个位点与 10 个性状之间的相关系数表

显著性检验结果如下图所示，只有前两个典型相关满足显著性检验，因此认为前两个典型相关具有意义，计算结果如下图所示：

Test that remaining correlations are zero:

	Wilk's	Chi-SQ	DF	Sig.
1	.492	691.936	360.000	.000
2	.608	485.748	315.000	.000
3	.715	327.853	272.000	.011
4	.768	257.454	231.000	.112
5	.817	196.883	192.000	.389
6	.858	149.819	155.000	.602
7	.899	104.336	120.000	.845
8	.930	70.334	87.000	.904
9	.957	42.849	56.000	.902
10	.982	17.312	27.000	.923

图 7.4　显著性检验结果图

取前两个典型相关关系，画出其中的典型相关关系的系数，其中系数比较大的位点是 36 个位点和 10 个性状之间相关性最大的。计算结果如下图所示：

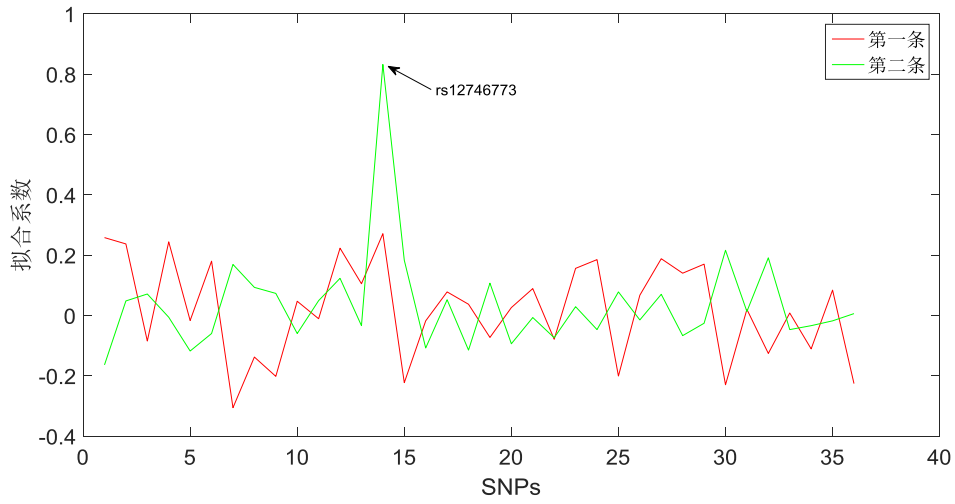

图 7.5　前两个典型相关关系系数结果图

采用冗余分析进行方差解释，可以看出第一条典型关系可以解释 68.2%的

方差，第二条可以解释 4.6%的方差，前两条方差解释相比其他典型关系数值较高，计算结果如下图所示：

Proportion of Variance of Set-2 Explained by Its Own Can. Var.

	Prop Var
CV2-1	.682
CV2-2	.046
CV2-3	.049
CV2-4	.028
CV2-5	.034
CV2-6	.028
CV2-7	.029
CV2-8	.029
CV2-9	.029
CV2-10	.046

图 7.6　前两个典型相关方差解释表

综合上面的分析可以看出，第一条典型关系显著并且可信度最高，第二条显著但是可信度较低，其他典型关系不显著并且可信度低。综合来看，第一条典型关系最可信，相关的位点是 rs12746773。

八. 总结与展望

根据题设要求以及所提供的原始数据，我们完成了所有四个问题的建模、解算及讨论，针对各问题给出了相应的解答和注释，简要总结如下：

(1) 将每个位点的三种编码方式分别编码为 0、1、2，为后续数据处理提供了便利；

(2) 将位点与患病建立一个函数映射关系，即 9445 个相互独立的位点是自变量，因变量为患病，通过计算卡方值，根据自变量与因变量的关联强度的不同，筛选出最可能与患病相关的位点，最后能过显著性检验来对总体分布进行假设，用样本信息证实了这个模型是合理的；

(3) 针对这一问题，为了简化位点和基因的关系，我们假设基因内的位点相互独立，基因和位点具有简单的线性关系，建立多元线性回归模型，用残差来表示基因与患病之间的关系，筛选出了与患病关联最强的几个基因。最后用这几个基因进行显著性检验，证实了这个模型是合理的；

(4) 由于性状之间有很强的关联性，所以将 10 个性状看为一个整体，求全局最优解，我们假设位点与性状之间存在线性关系，采用典型相关分析，选择回归系数比较大的位点对性状具有比较大的影响，并运用显著性检验进行验证，证实该模型是合理的。

全基因组关联分析(GWAS)是一种在人类或动植物全基因组中寻找变异序列的方法，是应用基因组中数以百万计的单核苷酸多态性为分子遗传标记，进行全基因组水平上的对照分析或相关性分析，通过比较发现影响复杂性状的基因变异的一种新策略。

GWAS 分析方法的原理是，借助于 SNP 分子遗传标记，进行总体关联分析，在全基因组范围内选择遗传变异进行基因分型，比较异常和对照组之间每个遗传变异及其频率的差异，统计分析每个变异与目标性状之间的关联性大小，选出最相关的遗传变异进行验证，并根据验证结果最终确认其与目标性状之间的相关性。

GWAS 在寻找致病基因方面有很大的作用，因为 GWAS 可以在全基因组范围内进行高通量的大规模筛选，可以发现用候选基因策略很难发现的遗传变异，这种方法的引入，使对遗传流行病的发病预测不再停留在传统的年龄、家族史等“环境性”因素分析，而是通过对人体全基因组的分析，找出可能导致今后发病的基因，并结合“环境性”因素，得出包括癌症在内的多种流行病的发病率。

然而 GWAS 也存在局限性。首先，GWAS 对疾病的选择有所限制。遗传关联性好的疾病使用该方法比较容易筛选易感位点，而低遗传度的疾病遗传学关联研究的把握度低，因此筛选易感位点需要非常大的样本量。其次，通过 GWAS 发现的新位点局限在 DNA 的序列水平上，只能用于预测和风险评估，临床应用还需要更深入的研究。而且，受种群差异、生活环境等因素的影响，不同人群同种疾病的易感基因可能不一致。

在 GWAS 研究后要确定一个基因型-表型因果关系还有许多困难，由于连锁不平衡的原因，相邻的 SNP 之间会有连锁现象发生。同样，在测序时同样存在连锁不平衡现象，而且即使测序的费用降到非常低的水平，要想如 GWAS 研究一般地获得大量样本的基因组数据还是非常困难的。

但是，随着基因组研究和基因芯片技术的不断发展和完善，必将迎来GWAS的广泛应用。

九. 参考文献

[1] Robert J. Klein, et al. Complement Factor H Polymorphism in Age-Related Macular Degeneration. Science, 2005(4), 308: 385-389.

[2] Anna Helgadottir, et al. A Common Variant on Chromosome 9p21 Affects the Risk of Myocardial Infarction. Science, 2007(7), 316:1491-1493.

[3] Ruth McPherson, et al. A Common Allele on Chromosome 9 Associated with Coronary Heart Disease. Science, 2007(7), 316:1488-1491.

[4] 赵肖肖，朱宁，黄云腾. Logistic 回归和 T 检验在基因特征提取中的应用. 桂林电子科技大学学报, 2012, 32(1):69-81.

[5] 张雁明，邢国芳，刘美桃等. 全基因组关联分析：基因组学研究的机遇与挑战. 生物技术通报, 2013(6):1-6.

[6] 涂欣，石立松，汪樊等. 全基因组关联分析的进展与反思. 生理科学进展, 2014，42(2):87-94.

[7] 罗旭红，刘志芳，董长征. 基因水平的关联分析方法. 遗传, 2013，35(9):1065-1071.

[8] Hirschhorn，J. N. and M. J.Daly. Genome-wide association studies for common diseases and complex traits. Nature Reviews Genetics, 2005，6(2):95-108.

[9] Easton，D. F.，Pooley，A. M. Dunning，et al. Genome-wide association study identifies novel breast cancer susceptibility loci. Nature, 2007 ，447(7148):1087-1093.

[10] Wang，W. Y. S.，B. J. Barratt，D. G. Clayton，et al. Genome-wide association studies: theoretical and practical concerns. Nature Reviews Genetics, 2005，6(2):109-118.

[11] Hunter，D. J.，P. Kraft，K. B. Jacobs，et al. A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. Nature genetics, 2007，39(7):870-874.

[12] Yeager，M.，N. Orr，R. B. Hayes，et al. Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. Nature genetics, 2007，39(5):645-649.