具有遗传性疾病和性状的遗传位点统计分析

ReadMe：本文素材来自于研究生数模竞赛的题目，本人课余时间结合了几篇优秀论文的实现方法，在 R 语言中实现了四道题目。因为多元统计分析这门课的机缘巧合，将程序中的思想简单写成了下述论文，其中多元线性回归与典型相关分析是多元统计分析书本中重要内容。

本文建立了 bonferroni-卡方检验、多元线性回归、典型相关分析等数学模型，分析了全基因组关联中单/多性状-单/多位点的关系，并对模型进行了检验。

在问题一中，主要解决位点数据数值编码的问题。根据每个位点蕴含的碱基类型进行划分，划分得 A/T，A/G，A/C，C/G，C/T，G/T 六种基本位点类型。在每一个位点中，可构成的碱基对类型均为三种，如 A/T 类型可构成 AA，AT，TT 三种碱基对，我们可将其编码为 0、1、2。位点只与三种碱基对的比例相关，与其中碱基对的顺序，位置均无关，在上述编码方式下，我们可以继续进行相关的分析。

在问题二中，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。1000 个样本中分为患病组与对照组两组，如果某个位点与患病不相关，那么位点的编码在患病组与对照组分布几乎一致；如果某个位点与患病相关，那么两组编码应存在显著差异，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，检验过程利用 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。

在问题三中，主要分析基因（多位点）与性状之间的关系，为了简化问题的分析，假设基因与性状为简单的线性相关关系，且如果拥有患病基因，就会导致疾病，建立因变量为基因，自变量为位点的多元线性回归模型，利用 F 分布检验回归显著性，最终发现 Gene217，Gene245 具有较强的致病性。

在问题四中，首先对十种性状数据进行初步的统计检验，发现其存在较强的相关性，其次利用问题二中卡方检验的方法，利用较宽松的条件筛选与性状关联度较差的位点，减小后续问题分析的难度。进一步我们建立典型相关分析的模型来分析多形状-多位点之间的关系，求解其中的典型关系，利用显著性检验选取其中可以信赖的典型关系，并认定回归直线中系数较大的位点对性状具有较大的影响，最终，发现与 10 个性状相关性最大的位点为 rs12746773。

关键词：SNP，遗传统计学，卡方检验，多元线性回归，典型相关分析

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,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。染色体、基因和位点的结构关系见图 1.

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)；如表 1 中，不同样本在位点 rs100015 就有三种不同编码 *TT*,*TC* 和 *CC*。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。

表 1. 在对每个样本采集完全基因组信息后，一般有以下的数据信息（以 6 个样本为例，其中 3 个病人，3 个健康者):

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。

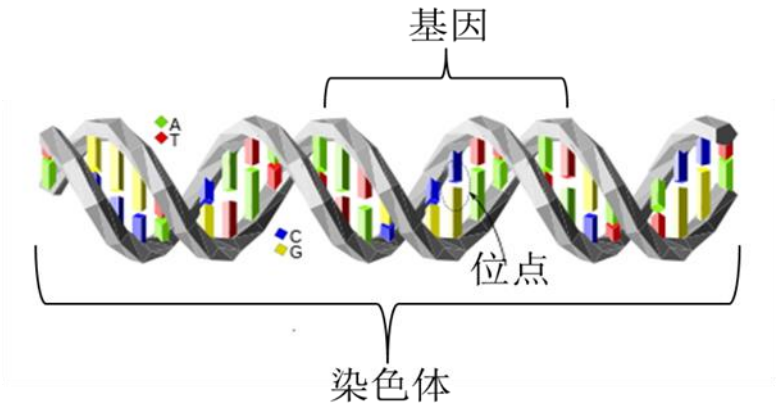

图 1. 染色体、基因和位点的结构关系.

本题目提供 1000 个样本的 11 种疾病（或性状）信息，以及其 9445 个位点编码信息，这些信息包含在附录中的 3 个文件(phenotype.txt，genotype.dat 和 multi_phenos.txt)和 1 个文件夹 gene_info(包含 300 个文件)中。

genotype.dat 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。该文件总共有 1001 行，9445 列。具体来说，第一行表示 9445 个位点的名称，都是以字母 rs 开头的；接下来，有 1000 行，每一行表示一个样本在该条染色体片段上所有位点(9445 个位点)的编码信息。例如，该文件中第 2 行，就表示 1 号样本在该条染色体片段上 9445 个位点的编码信息。如同上表中第三列到第六列的编码信息。

phenotype.txt 文件中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 0，500 个 1，表示我们现在共有 1000 个样本，其中 500 个 0 就是 500 个没患有疾病 A 的人，500 个 1 就是有 500 个患有遗传病 A 的人。如同表一中的第二列。

multi_phenos.txt 文件中包含了上述样本的 10 种相关性状(如肥胖、高血压、脂肪肝等)组成的信息，每一列表示一个性状。文件中的 0 和 1 信息同 phenotype.txt 文件。

文件夹 gene_info 中包含了 300 个 dat 文件，表示 300 个基因的信息；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，事实上，可以把基因理解为若干个位点组成的集合。注意到在 genotype.dat 文件中已包含所有位点的编码信息，所以我们可以得到每一个基因所包含位点的编码信息了。例如 gene_1.dat，表示基因 gene_1 包含了 rs3094315,..., rs4040617，共 7 个位点。

所有这些文件都可以利用 Notepad++软件打开。

§ 1.2 研究问题

问题一、请用适当的方法，把 genotype.dat 中每个位点的碱基(A,T,C,G) 编码方式转化成数值编码方式，便于进行数据分析。

问题二、现有一组 1000 个样本在某条有可能致病的染色体片段上 9445 个位点的编码信息(见 genotype.dat)。样本患有遗传疾病 A 的信息，见 phenotype.txt 文件。请设计或采用一个方法，找出某种疾病最有可能的一个或几个致病位点，并给出相关的理论依据。

问题三、同上题中的样本患有遗传疾病 A 的信息（phenotype.txt 文件）。现有 300 个基因，请找出与疾病最有可能相关的一个或几个基因，并说明理由。（每个基因所包含的位点名称见文件夹 gene_info 中的 300 个 dat 文件，每个 dat 文件代表一个基因所包含的位点(位点信息见文件 genotype.dat)。）

问题四、在问题二中，已知 9445 个位点，其编码信息见 genotype.dat 文件。在实际的研究中，科研人员往往把相关的性状或疾病(如高血压，心脏病、脂肪肝和酒精依赖等)放在一起研究，这样能提高发现致病位点或基因的能力；即把这些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。现有 10 个相关的性状或疾病，其样本数据见 multi_phenos.txt 文件。试找出与 multi_phenos.txt 中 10 个性状都有关系的位点，并从理论上说明你所发现的致病位点的合理性。

二、条件假设

1、不对异常基因型值 DD,DI,II 进行考量

2、位点之间相互独立

3、问题二中，假设有致病位点即患病

4、问题三、四中，假设基因内的位点相互独立，基因与位点具有简单的线性关系

三、问题分析

§3.1 问题 1：数值编码

问题 1 要求采用适当的方法，将 genotype.dat 中每个位点的碱基(A,T,C,G)编码方式转换为数值编码方式，以保证后续数据分析开展的可靠性。

在 genotype.dat 文件中，共有 1000 个样本，每个样本含有 9445 个位点，每个位点的碱基对仅由两个固定的碱基构成，因此可根据其蕴含的碱基类型进行划分，划分得 A/T，A/G，A/C，C/G，C/T，G/T 六种基本位点类型。在每一个位点中，可构成的碱基对类型均为三种，如 A/T 类型可构成 AA，AT，TT 三种碱基对，我们可将其编码为 0、1、2。

在后三问中，位点只与三种碱基对（编码）的比例相关，与其中碱基对的顺序，位置均无关，在上述编码方式下，我们可以继续进行相关的研究。

$$\begin{array}{c} \\ A \\ T \\ C \\ G \end{array} \begin{array}{c} \begin{array}{cccc} A & T & C & G \end{array} \\ \begin{pmatrix} & AT & AC & AG \\ & & TC & TG \\ & & & CG \\ & & & \end{pmatrix} \end{array}$$

图 4.1 碱基组成类型

§3.2 问题 2：单独位点的致病基因分析

问题 2 要求在已知表现型（即个体是否患病，由 phenotype.dat 给出）的前提下，从 1000 个个体的位点信息中寻找最有可能的一个或多个致病位点。

在生物学上，基因型与表现型具有复杂的关系。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。

我们可以假设位点与患病之间存在函数映射关系，自变量是 9445 个位点，因变量为是否患病的零一变量（0 为健康，1 为患病）。为了找到最有可能的一个或多个致病位点，我们需要寻找自变量位点中与因变量函数映射关系最为显著的集合。1000 个样本中分为患病组与对照组两组，如果某个位点与患病不相关，那么位点的编码在患病组与对照组分布几乎一致；如果某个位点与患病相关，那么两组编码应存在显著差异，因而我们采用患病组与对照组的差异性作为衡量位点与患病关系显著性的指标。

个体\位点	**患病**	RS3094315	RS3131972	RS3131969	…（9445）
1	**0**	2	1	0	…
2	**0**	1	1	1	…
…	…	…	…	…	…
501	**1**	2	2	1	…
502	**1**	1	1	0	…
…（1000）	…	…	…	…	…

表 4.2 位点-患病关系概览

上述差异性可以利用卡方值来表示，卡方值越大，差异性越大，位点与患病之间的关系也就越显著。通过设定阈值，可以找出显著性较强的位点，运用显著性检验及 Bonferroni 校正，进一步验证其合理性

§3.3 问题 3：基因的致病性

问题三在上一问的基础上，提供了 300 个基因，基因是若干个位点的集合，此问旨在寻找与患病相关的一个或多个基因。

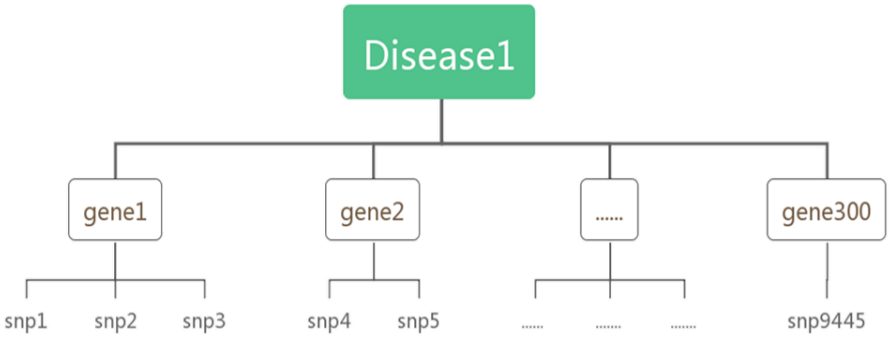

图 4.3 疾病-基因-位点关系图

生物学上，位点与基因之间、基因与患病之间具有复杂的关系。为了简化分析，假设基因内的位点相互独立，基因与位点具有简单的线性关系，假设基因 i 由 n 个位点组成，即

$$Gene_i = a_0 + a_1 SNP_1 + a_2 SNP_2 + \ldots + a_n SNP_n \tag{1}$$

对基因内的位点与患病进行多元线性回归分析，分别采取 F 检验判断回归方程的显著性，选取其中显著的基因，即可认为存在其疾病的致病性。

§3.4 问题 4：多性状位点分析

问题 4 要求找出与 multi_phenos.txt 文件中十个性状均存在相关性的位点。

个体\位点	患病 1	…	患病 10	RS3094315	RS3131972	RS3131969	…（9445）
1	0	…	0	2	1	0	…
2	1	…	1	1	1	1	…
…	…	…		…	…	…	…
501	1	…	1	2	2	1	…
502	1	…	1	1	1	0	…
…（1000）	…	…	…	…	…	…	…

表 4.3 多患病-多位点数据概览表

首先单独从性状出发，求解十个形状的相关性，其结果应为相关性较强；其次探究位点与性状的关系，筛选出与性状无关的位点，利用问题二中的算法，我们可以将位点降维，得到一个含有患病位点的候选名单。

最后，问题转换为求解候选名单内位点与十个性状的关系。显然，问题二中求解一个性状与多个位点的关系，在这里进一步发展成为求解多个性状与多个位点的关系，从局部最优解问题演化成全局最优解问题。假设位点与性状之间存在线性关系，如果考虑如下多个多元线性回归模型的使用，并不能表现十个性状的相关性。

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \cdots \\ Y_i \end{bmatrix} = \begin{bmatrix} H_{1,1} & H_{1,2} & \ldots & H_{1,n} \\ H_{2,1} & H_{2,2} & \ldots & H_{2,n} \\ \cdots & \cdots & \cdots & \cdots \\ H_{10,1} & H_{10,2} & \ldots & H_{10,n} \end{bmatrix} \begin{bmatrix} SNP_1 \\ SNP_2 \\ \cdots \\ SNP_n \end{bmatrix} \tag{2}$$

于是，我们进一步考虑采用典型相关分析，最后利用典型相关系数进行模型的显著性检验

四、问题 1：数值编码

根据问题分析，在 genotype.dat 文件中，每一项位点由两个不同的碱基构成，如图 5.1 表示，四种碱基可以构成六种不同的位点类型（A/T，A/G，A/C，G/T，C/T，C/G）；每个位点类型共有三种编码类型，如 A/T 类型，有 AA，AT，TT 三种编码方式。

$$\begin{array}{c} \\ A \\ T \\ C \\ G \end{array} \begin{array}{c} \begin{array}{cccc} A & T & C & G \end{array} \\ \begin{pmatrix} & AT & AC & AG \\ & & TC & TG \\ & & & CG \\ & & & \end{pmatrix} \end{array}$$

图 5.1 碱基组成类型

在后续问题的分析中，位点只与三种编码方式的比例相关，与位点的碱基对的顺序、位置均无关，所以六种类型的位点均可视为 M/N 的形式，其中 MM，MN，NN 的编号分别为 0，1，2。以下给出各位点的编码映射表及操作流程图:

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

表 4.1 编码映射表（in alphabetic order）

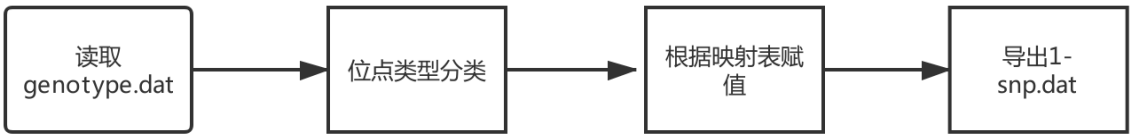

图 5.2 数值编码流程图

五、问题 2：单独位点的致病基因分析

§5.1 卡方检验模型建立

假设位点与患病之间存在函数映射关系，自变量是 9445 个位点，因变量为是否患病的零一变量

（0 为健康，1 为患病）。为了找到最有可能的一个或多个致病位点，我们需要寻找自变量位点中与因变量函数映射关系最为显著的集合。1000 个样本中分为患病组与对照组两组，如果某个位点与患病不相关，那么位点的编码在患病组与对照组分布几乎一致；如果某个位点与患病相关，那么两组编码应存在显著差异，因而我们采用患病组与对照组的差异性作为衡量位点与患病关系显著性的指标。

个体\位点	**患病**	RS3094315	RS3131972	RS3131969	…（9445）
1	**0**	2	1	0	…
2	**0**	1	1	1	…
…	…	…	…	…	…
501	**1**	2	2	1	…
502	**1**	1	1	0	…
…（1000）	…	…	…	…	…

表.6.1 位点-患病关系表

卡方检验是以χ^2分布为基础的一种常用假设检验方法，它的无效假设 H_0 是：观察频数与期望频数没有差别。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度太大，应当拒绝无效假设，表示比较资料之间有显著差异；否则就不能拒绝无效假设，尚不能认为样本所代表的实际情况和理论假设有差别。

在本问中，位点与患病相关意味着患病组与对照组差异显著，观察值与理论值偏差太大，应当拒绝无效假设，且此时的p 值以很小的值存在。对各位点的p 值计算使得我们可以进一步的寻找致病位点。

如下给出本题卡方检验的统计分类表（以 rs3094315 为例），并给出卡方检验的假设：

	AA（0）	AT（1）	TT（2）	合计
患病（1）	17	146	337	500
不患病（0）	26	147	327	500
合计	53	293	664	1000

表.6.2 卡方检验统计表

假设如下：

H_0：此位点不是最有可能的致病基因；　　H_1：此位点是有可能的致病基因

根据上述两行三列的统计表，给出其卡方检验对应的偏离度公式

$$\chi^2 = n\left(\sum \frac{A^2}{n_R n_C} - 1\right) \tag{3}$$

其中 n 为总例数，n_R为对应第R行的总频数，n_C为对应第C列的总频数，A为相应的观察频数。

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，χ^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，χ^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，χ^2值越大。

在本文中，数值大的χ^2值表明患病组与健康组之间的差异大，即远离假设；数值小的χ^2值表明健康组与患病组并无过大差异，接近假设。同时，χ^2也是假设成立与否的度量指标，通过χ^2值我们可以查询卡方分布临界表，换算获得接受无效假设H_0的概率p。

§ 5.2 模型计算及 Bonferroni 校正

利用开源软件 R，我们得以更高效地对 9445 个位点进行卡方检验p值的计算，结果如下图：

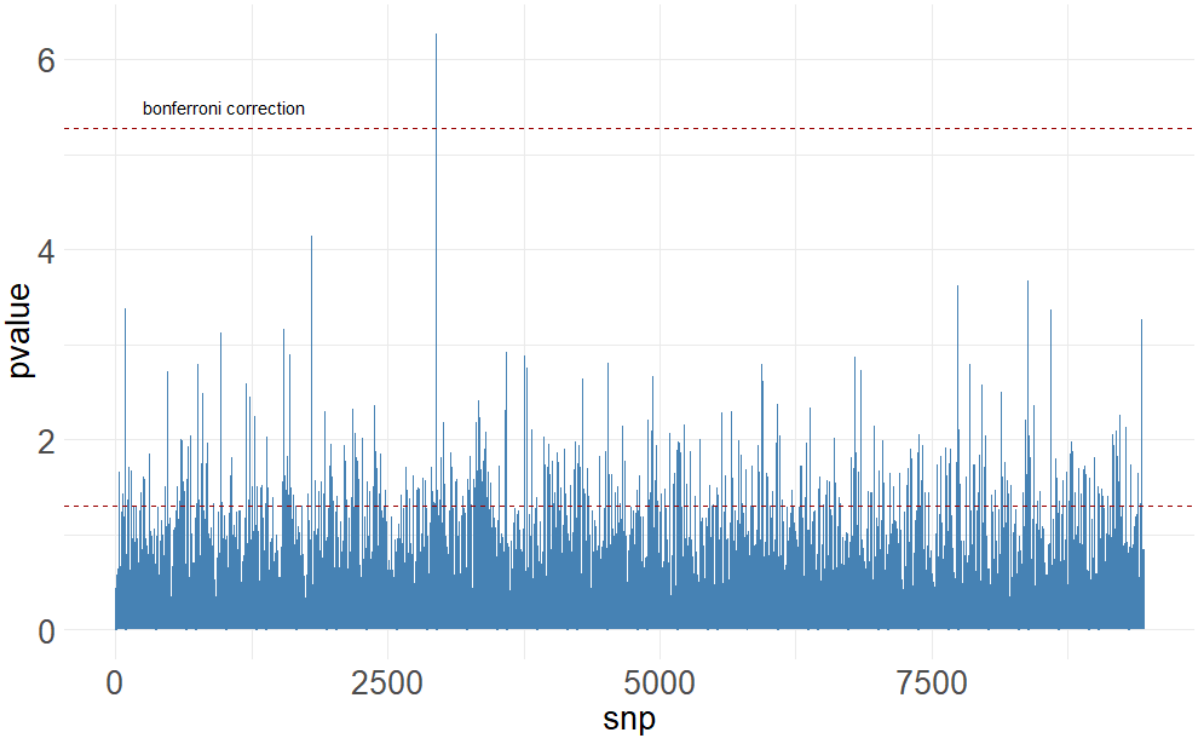

图.6.2 卡方检验各位点 p 值（$pvalue = -\lg p$）

在卡方检验中，我们一般选取设置显著性水平为 0.05，如图中下方虚线所示，在这种条件下我们会获得 447 个点。此时点数已到达了总数的一半，明显放宽了阈值判断。

因此，我们采用 Bonferroni 校正，在本问多重校验过程中，通过降低显著性水平来弥补同时进行检验次数太大造成的问题，其形式如下

$$P\left(SNP_i \text{ passes } | H_0\right) \le \frac{\alpha}{n} \tag{4}$$

此时p放宽到了 5.3e-6，如图 6.2 所示，在这个严格的阈值线上只有一个位点符合要求，即 rs2273298。

因此，位点与患病关联性最强的是位点 rs2273298。

六、问题 3：基因的致病性分析

§6.1 疾病-基因-位点多元线性回归模型

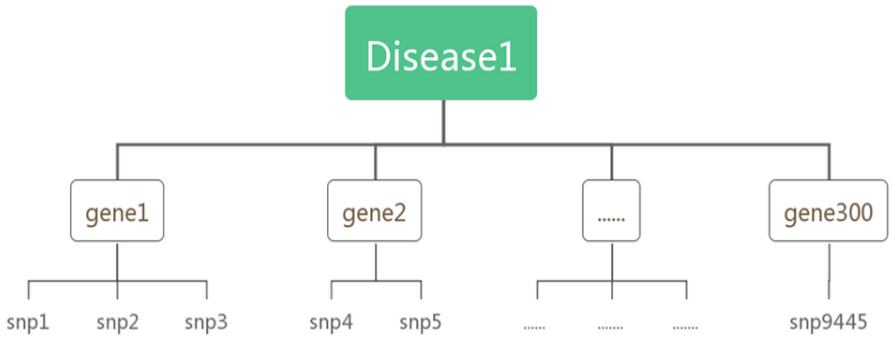

图 7.1 疾病-基因-位点关系图

由前文问题分析及上述疾病-基因-位点关系图可知，在不考虑生物学上各种复杂关系的前提下，假设基因与位点具有简单的线性关系，我们可以建立基因与位点之间的多元线性回归，类似其基本形式 $y = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_n x_n$，我们可以给出以下方程：

$$Gene_i = a_0 + a_1 SNP_1 + a_2 SNP_2 + \ldots + a_n SNP_n \tag{5}$$

以 gene_300.dat 中的数据为例，我们列出以下联表，以更清晰的来看待这个问题：

个体\位点	**患病**	RS6688664	RS6702129	RS7548805	…	RS7545865
1	**0**	0	1	2	…	1
2	**0**	0	1	1	…	2
…	…	…	…	…	…	…
501	**1**	1	1	2	…	1
502	**1**	1	0	1	…	2
…（1000）	…	…	…	…	…	…

表.7.1 位点-患病关系表

与表 6.1 不同的是，上表的位点缩小到了基因所拥有的位点，每一个位点的数值编码作为公式（5）中SNP_n的值，而$Gene_i$的值将取患病的零一变量进行计算，患病样本依然是前 500 位健康、后 500 位患病。

于是，我们相当于把疾病与基因的致病关系，转换为分析疾病与基因中相关位点的致病分析，

若基因中相应的位点能够很好的在多元线性回归模型中拟合患病值，那我们即可认可所分析的基因是最有可能致病的基因。

§6.2 回归计算

利用 R 语言，我们提取了 300 个基因进行多元线性回归，通过计算参数a的最小二乘估计:

$$\hat{a} = (Snp^T Snp)^{-1} Snp^T Gene \tag{6}$$

我们可以获得 300 对基因-位点的线性关系式，如下给出 gene_1 的回归系数示例:

(Intercept)	rs3094315	rs3131972	rs3131969
0.5093220271	0.0311338893	-0.0308687333	0.0184213079
rs1048488	**rs12562034**	**rs12124819**	**rs4040617**
-0.0006830496	-0.0254435477	-0.0066987832	-0.0021967523

表.7.2 gene_1 回归系数表

§6.3 F分布检验与拟合优度检验

在上述模型的环境下，F 检验能够检验因变量同多个自变量的整体线性关系是否显著，因为问题的核心在于寻找最有可能的致病基因而非位点，在这里我们不需要采用 t 分布去研究每一个位点的显著性。构造 F 检验如下

$$F = \frac{SSR / p}{SSE / (n-p-1)} \sim F(p, n-p-1) \tag{7}$$

上述统计量是用回归模型的可解释平方和、残差平方和以及各自的自由度建立起来的。我们假在无效假设中，回归模型的所有参数都是 0，模型没有存在的意义。这时，F 的分子和分母都服从卡方分布，而 F 服从 F 分布：$F\ k-1, N-k$ 。通过查表，我们就可以确定一个接受无效假设的概率，来确定哪一个基因具有最强的致病性。

通过计算，我们可以得到以下 F 分布 p 值检验的图表：

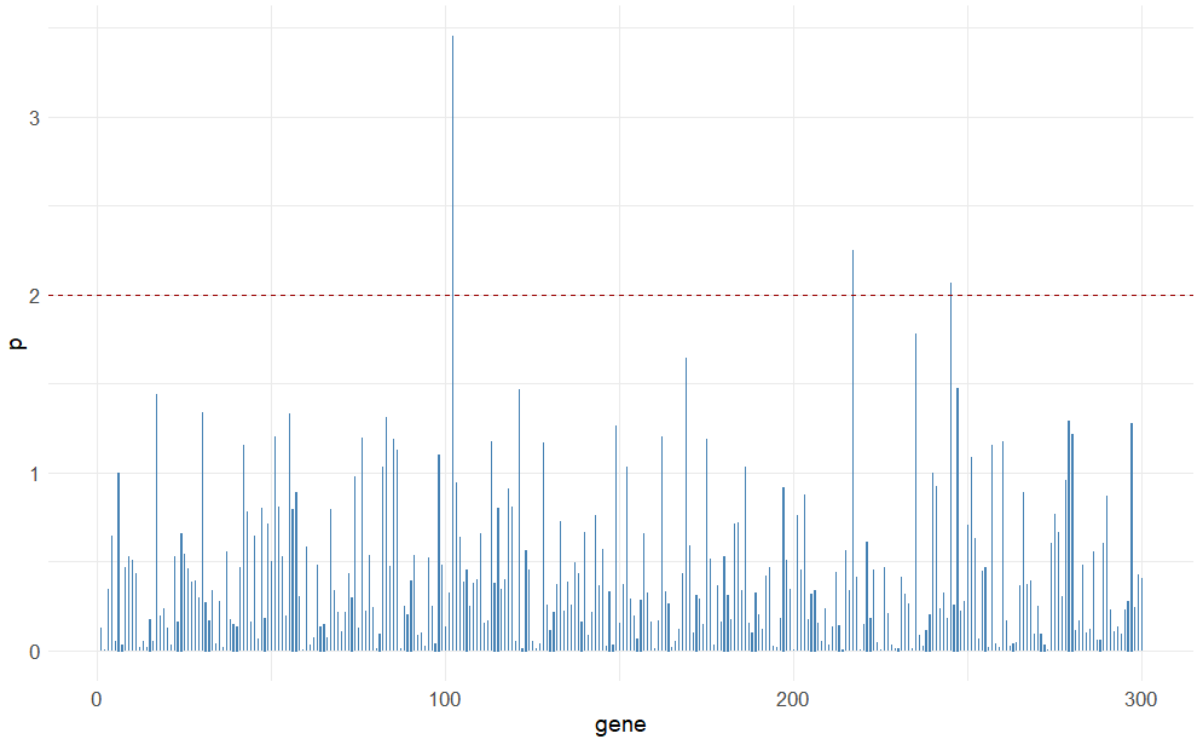

图 7.2 F 分布检验 p 值图

可以看出，Gene102 最符合显著性检验，Gene217 和 Gene245 次之。

F 分布的显著性检验让我们有一定的信心赋予模型现实的意义，进一步的，我们通过计算拟合优度来确定在多大程度上自变量决定了因变量的取值，也就是基因在多大程度上决定了相关疾病性状的出现。

定义决定系数R^2为：

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} \tag{8}$$

其中，SST是因变量的方差，决定系数的取值范围是$[0,1]$，取值越大表示模型对因变量的解释越充分，越小表示模型对因变量的解释越不充分。进一步地，我们使用调整决定系数，获得更为准确简洁的答案：

$$\mathrm{adjusted}R^2 = 1 - \frac{(n-1)\left(1-R^2\right)}{n-p-1} \tag{9}$$

计算结果如下：

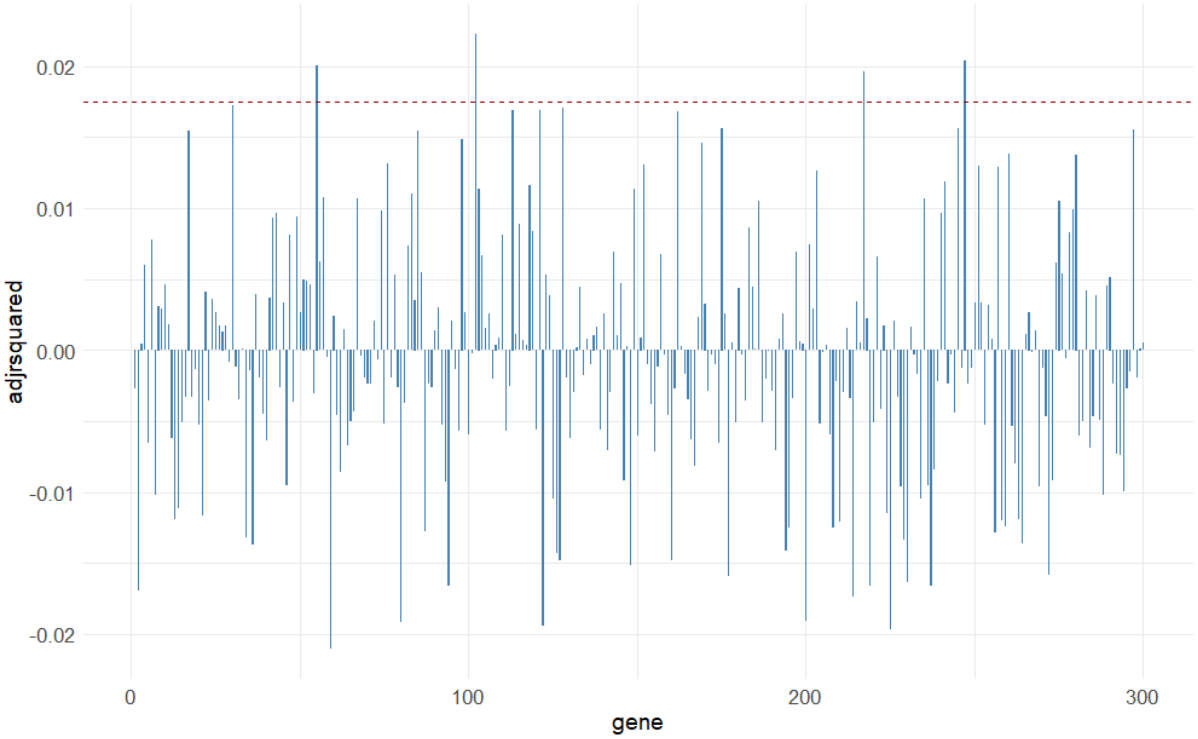

图 7.3 调整决定系数统计图

可以看出，Gene55、Gene102、Gene217、Gene247 的拟合优度较高，且 Gene102 的显著性水平与拟合优度均为最高，我们可以认定其基因与疾病之间有最强的相关性；Gene55、Gene247 显著性水平不高，不予考虑，Gene217、Gene245 有次优的显著性水平与拟合优度，因此我们认为 Gene217、Gene245 存在相对于 Gene102 较弱的致病性。

综上，Gene102 存在最强致病性，Gene217、Gene245 基因与致病存在一定的相关性。

七、问题 4：多性状位点分析

§7.1 性状相关性检验

为了寻找出与 multi_phenos.txt 中 10 个性状都有关系的位点，我们首先要初步统计这 10 个性状自身需要有很强的相关性。

其中相关系数公式为:

$$\rho_{XY} = \frac{Cov\ X,Y}{\sqrt{D\ X}\sqrt{D\ Y}}$$

计算结果如下表:

	1	2	3	4	5	6	7	8	9	10
1	1	0.716	0.740	0.708	0.684	0.680	0.740	0.712	0.744	0.728
2	0.716	1	0.712	0.688	0.680	0.692	0.716	0.732	0.736	0.748
3	0.740	0.712	1	0.700	0.724	0.692	0.724	0.728	0.712	0.748
4	0.708	0.688	0.700	1	0.684	0.676	0.720	0.732	0.720	0.712
5	0.684	0.680	0.724	0.684	1	0.748	0.680	0.708	0.668	0.692
6	0.680	0.692	0.692	0.676	0.748	1	0.684	0.708	0.676	0.668
7	0.740	0.716	0.724	0.720	0.680	0.684	1	0.736	0.720	0.760
8	0.712	0.732	0.728	0.732	0.708	0.708	0.736	1	0.728	0.760
9	0.744	0.736	0.712	0.720	0.668	0.676	0.720	0.728	1	0.744
10	0.728	0.748	0.748	0.712	0.692	0.668	0.760	0.760	0.744	1

表 8.1 不同性状之间的相关系数表

从上表可以看出，大多数不同性状之间的相关系数都约达到了 0.7，说明性状之间的相关性比较强，那么也很有可能存在与 10 个性状均相关的的一个或多个位点。

§7.2 位点筛选

由表 4.2 和表 4.3 可以看出，从第二问到第四问，问题从单患病-多位点关系演变成多患病-多位点关系，在第二问中，只有少数位点存在致病性，大多数位点与患病无关，因此，在这一问中，我们首先在 9445 个位点中，分别针对 10 种性状，筛选出相对宽松条件下有致病相关性的位点，从而到达简化问题分析的目的。

利用第二问中相同的方法，分布计算每种性状与位点之间的关联，通过p值来衡量性状与位点的相关性，p值越小，显著性越高，关联性越强。统计图如下：

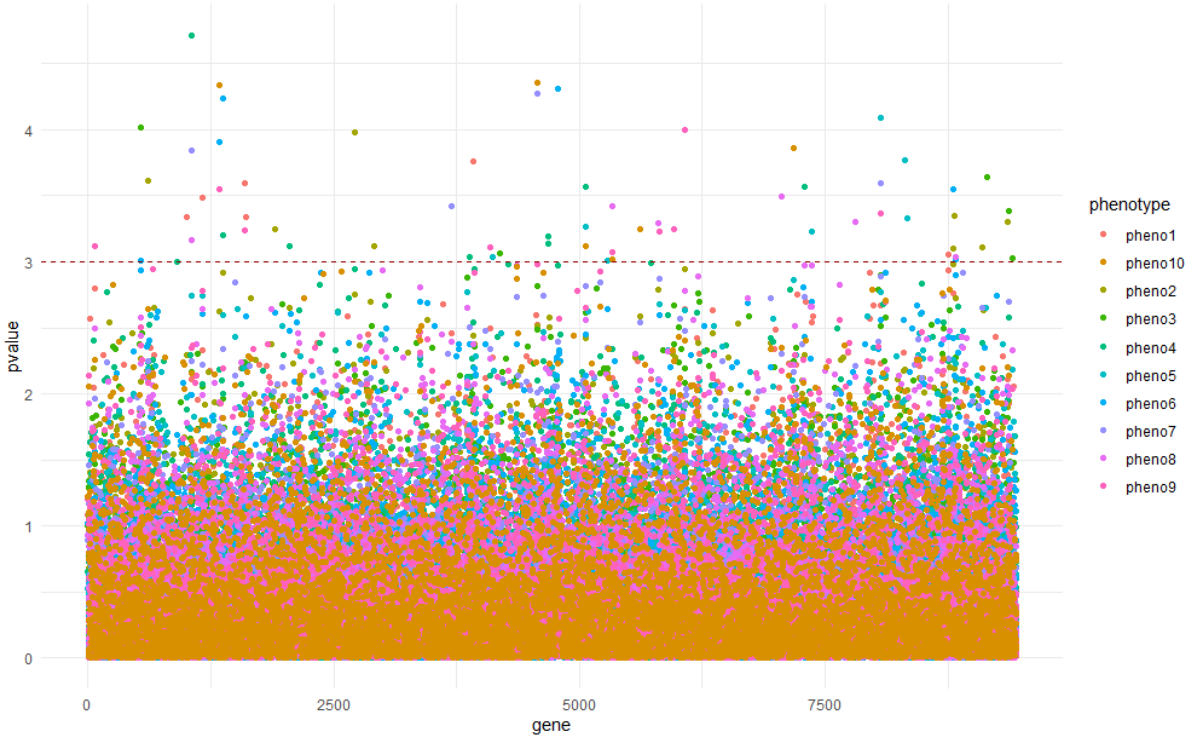

图 7.2 多性状-多位点 p 值统计图（$pvalue = -\lg p$）

从上图可以看出，绝大多数的位点与基因的关联性极低，我们需要设置一个相对于第二问更宽松保守的条件。因此，设置筛选条件为$-\lg p > 3$，若某个位点与 10 个性状的p值均满足宽松筛选条件，那么将此位点纳入候选名单，进入下一步的计算。

经过计算，共有 55 个位点进入候选名单。

§7.3 典型相关分析寻找典型位点

典型相关分析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：为了从总体上把握两组指标之间的相关关系，分别在两组变量中提取有代表性的两个综合变量 U1 和 V1（分别为两个变量组中各变量的线性组合），利用这两个综合变量之间的相关关系来反映两组指标之间的整体相关性。

在这里，综合变量对即为十个性状与已经进入候选名单的位点，通过典型相关分析，可以克服第二问中单性状-多位点分析方法中忽视了性状间关系的问题。其关系如下所示：

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \cdots \\ Y_{10} \end{bmatrix} = \begin{bmatrix} H_{1,1} & H_{1,2} & \cdots & H_{1,n} \\ H_{2,1} & H_{2,2} & \cdots & H_{2,n} \\ \cdots & \cdots & \cdots & \cdots \\ H_{10,1} & H_{10,2} & \cdots & H_{10,n} \end{bmatrix} \begin{bmatrix} SNP_1 \\ SNP_2 \\ \cdots \\ SNP_n \end{bmatrix} \tag{10}$$

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。从SNP_i中提取U，从Y_i中提取V，综合变量U,V也就分别是两类变量的线性组合:

$$U = a^T X = \sum_{i=1}^{p} a_i x_i, \mathrm{V} = b^T Y = \sum_{j=1}^{q} b_j y_j \tag{11}$$

通过计算相关系数

$$r_{uv} = \frac{Cov\ a^T X, b^T Y}{\sqrt{D\ a^T X}\sqrt{D\ b^T Y}} = \frac{a^T \Sigma_{XY} b}{\sqrt{a^T \Sigma_{XX} x}\sqrt{b^T \Sigma_{YY} b}} = a^T \Sigma_{XY} b$$

使之达到值最大，且同时满足

$$\mathrm{var}(U) = \mathrm{var}\left(a' x\right) = a' \Sigma_{XX} a = 1 \quad , \quad \mathrm{var}(V) = \mathrm{var}\left(b' x\right) = b' \Sigma_{YY} b = 1$$

于是，我们可以获得关于a,b的，类似于回归的性状-位点表达式，同时获得了U,V第一对典型相关变量, 去除第一对典型关系后重复以上过程, 我们即可获得性状与位点之间的典型相关关系对。如下给出 10 个典型相关的相关系数:

1	2	3	4	5	6	7	8	9	10
0.5069	0.4045	0.3214	0.2926	0.2804	0.2631	0.2432	0.2206	0.2147	0.1665

表 8.2 典型相关关系系数表

同时进行显著性检验，数据如下

	STAT	APPROX	DF1	DF2	P.VALUE
1	1.10133221	1.8686604	550	9332	0.000000e+00
2	0.75534953	1.4535039	486	9352	7.852428e-10
3	0.55963912	1.2370136	424	9372	8.139179e-04
4	0.44437178	1.1465769	364	9392	3.057652e-02
5	0.35073513	1.0787971	306	9412	1.685790e-01
6	0.26537081	1.0011910	250	9432	4.833344e-01
7	0.19097541	0.9209692	196	9452	7.773590e-01
8	0.12806663	0.8423938	144	9472	9.140849e-01
9	0.07688927	0.7764180	94	9492	9.461735e-01
10	0.02853534	0.5900612	46	9512	9.877714e-01

表 8.2 典型相关检验系数表（Hotelling-Lawley Trace, using F-approximation）

由以上数据，取前两个典型相关，认为其具有意义，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 55 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：

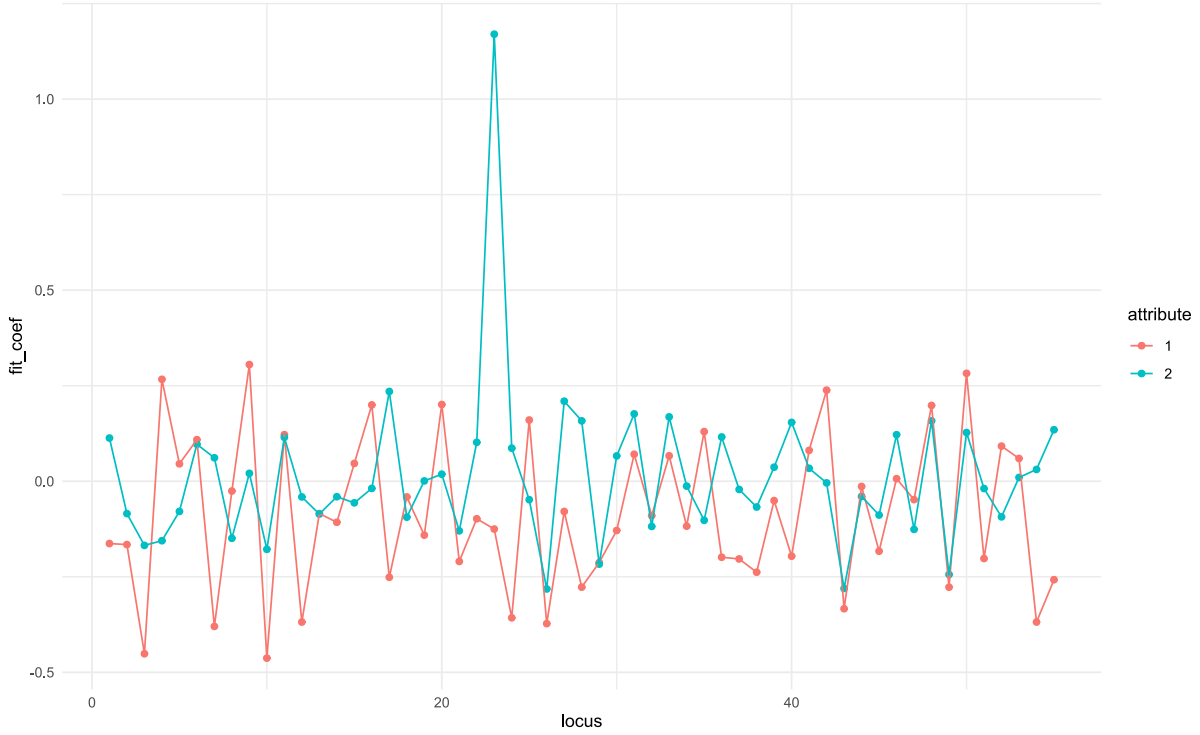

图 7.3 前两个典型相关关系系数结果图

考虑到显著性与相关系数两反面，rs12746773 取得最佳的成绩，因此，最终我们认定 rs12746773 是与 10 个性状都有关系的位点。

八、模型评判

1、 利用 0/1/2 的编码方式，为整个问题奠定了分析的基础，没有造成后续问题的不可分析

2、 第二问中通过卡方分布进行检验，能够很好的反映是否患病与三中位点类型的之间的深层关系，结果也证实了这个模型的合理性

3、 第四文中，从第二问中引伸，选取典型相关分析的方法，较好的解决了多位点-多性状的问题。

4、因中一些复杂的关系没有进一步进行考量