南京理工大学
NANJING UNIVERSITY OF SCIENCE & TECHNOLOGY

毕业设计说明书

作　　者：陈泽宇　　学　号：916113700122

学　　院：理学院

专业(方向)：应用统计学

班　　级：9161137001

题　　目：医学中动脉硬化数据分析

指导者：赵慧秀 副教授

评阅者：陈萍 教授

2020 年 6 月

声　明

我声明，本毕业设计说明书及其研究工作和所取得的成果是本人在导师的指导下独立完成的。研究过程中利用的所有资料均已在参考文献中列出，其他人员或机构对本毕业设计工作做出的贡献也已在致谢部分说明。

本毕业设计说明书不涉及任何秘密，南京理工大学有权保存其电子和纸质文档，可以借阅或网上公布其部分或全部内容，可以向有关部门或机构送交并授权保存、借阅或网上公布其部分或全部内容。

学生签名：

年　月　日

指导教师签名：

年　月　日

毕业设计说明书中文摘要

牙齿健康状况被医学研究认为是与动脉硬化相关的因素之一。根据相关研究的数据集，本文采用基于正态分布，逆高斯分布和二项分布假定的广义估计方程模型验证了牙齿健康状况和动脉硬化统计学上的显著性，并对于回归方程中变量间的交互效应进行了现实意义的理解，发现二者的相关关系受性别，年龄和吸烟状态的不同而发生不同程度的改变。此外，文章借助基于结构方程模型改造的双胞胎基因环境模型探究了基因和环境因素导致牙齿健康状况指标和动脉硬化指标在人群中差异的比重，发现不同性别群体中基因和环境因素解释两个指标的比重存在明显区别，同时建立的模型为进一步探究基因在牙齿健康状况和动脉硬化的相关性中起到的作用是否关键提供了方向。

关键词　动脉硬化　牙齿健康状况　广义估计方程　双胞胎基因环境模型

毕业设计说明书外文摘要

Title Data Analysis of Arteriosclerosis in Medicine

Abstract

In medical research, dental health is considered to be one of the factors related to arteriosclerosis. According to the data set of related research, this paper uses the generalized estimation equation based on normal distribution, inverse Gaussian distribution and binomial distribution hypothesis to verify the statistical significance of dental health and arteriosclerosis. And the practical meaning of the significance in the interaction effects between variables in the model has been elaborated. It is found that the correlation between dental health and arteriosclerosis is affected by gender, age and smoking status to varying degrees. In addition, with ACE model, a transformation of structural equation model, this article explores in what proportions can the differences in dental health and arteriosclerosis among population be explained by genetic and environmental factors. The result shows that there are obvious different proportions in male and female. At the same time, this model provides directions for further exploring whether genes play a key role in the correlation between dental health and arteriosclerosis.

Keywords Arteriosclerosis Dental Health Generalized Estimating Equation ACE model in Twin Analysis

目 次

1 引言

动脉硬化是随年龄自然增长出现的一种血管疾病，根据调查，在我国动脉粥样硬化已严重危害了人们的健康[1]。引起动脉硬化的病因有很多，诸如高血压，高血脂，糖尿病等，同时不良的饮食习惯，如大量摄入油腻食物也会导致动脉硬化。动脉硬化是一种慢性疾病，其形成过程缓慢，往往至中老年时期才会加重发病，给疾病的治疗带来了困难。而由于动脉硬化是一种与人们的生活方式密切相关的疾病，尽早检测到相关的征兆并及时给予治疗对于改善动脉的健康状况具有重大意义。虽然目前动脉硬化已被明确为是一种慢性炎症性疾病[2]，但迄今为止，人们对其发病机制尚未有清晰的认知[3]。由于其关乎人类生命健康，故具有一定的研究价值。

心血管疾病是世界公认的死亡和残疾的主要原因之一，这些死亡中大多数可归因于由动脉粥样硬化导致的缺血性心脏病以及中风的血栓栓塞事件。美国心脏协会指出牙周炎和心血管疾病享有相同的风险因素[4]。近些年的流行病学研究已将牙周炎同心血管疾病相互关联了起来[5]，这种关联具有生物学和医学意义。研究表明，牙周袋的慢性感染可作为病原微生物，其毒素和降解的产物将增加系统的炎症负担，如果它们进入血液循环，将会导致进一步的局部和全身炎症反应，这些都促进了动脉粥样硬化的形成，也潜在地增加了心血管疾病的风险[6-7]。因此，围绕牙周炎和牙齿健康状况，各国学者均进行了试验设计和研究，采用统计方法对试验数据进行分析，来定量研究口腔健康和动脉硬化之间的相关关系和因果关系。

首先已有文献指出牙齿健康状况如牙齿脱落与动脉硬化等疾病间接相关，Kenji Wakai 等人基于日本牙医协会关于人们口腔健康状况的问卷调查数据，采用线性回归模型，表明了牙齿脱落数量与人们摄入的关键营养素如维生素、胡萝卜素等存在相关性[8]。虽然此文献仅仅从饮食习惯的角度进行了研究，没有更精细的生理医学数据的支持，但这一结论仍然提示我们和人们饮食习惯相关的动脉硬化问题也可能与口腔健康有关。K. Asai 等人收集了日本 8124 名 30 至 75 岁年龄段的人关于年龄，性别，体重指数，吸烟状况，血红蛋白以及降糖药史等数据，利用多元线性回归模型分析了日本成年人牙齿脱落与动脉僵硬程度之间的关系。研究发现，牙齿脱落与动脉僵硬度之间存在线性关系，并且该关联因性别而异，在男性中牙齿脱落和动脉僵硬的风险更高[9]。Xiao-Tao Zeng 等人运用 Meta 分析的方法，针对牙周疾病与颈动脉粥样硬化之间的关系进行了分析，发现牙周疾病与颈动脉粥样硬化有显著的相关关系，且吸烟是牙周疾病和颈动脉硬化的共同风险因子[10]。但此文

献指出目前难以推断出牙周疾病与动脉硬化之间的因果关系。

上述文献从人们的饮食习惯和生理的角度讨论了口腔健康和动脉硬化之间的关系，而 Yuko Kurushima 等人在此基础上考虑了基因这一影响因素。其通过日本大于 50 岁的 106 名双胞胎动脉硬化情况，牙齿健康状况，体重指数，是否吸烟等各项生理数据建立多种模型，深入研究了在基因相同的条件下，口腔健康与动脉硬化的关系[11]。针对先前研究未考虑基因差异的缺陷，其研究样本选取了双胞胎，针对双胞胎数据（又称孪生数据）的特殊结构，该文献提出利用考虑孪生数据组内组间差异的回归模型结合广义估计方程这一参数估计方法（Generalized Estimating Equations, GEE，下称广义估计方程为 GEE）进行分析。GEE 方法通常用于分析纵向数据和相关联响应数据，在流行病学研究中有着广泛使用。与一般的最小二乘估计相比，其考虑了孪生数据中的配对关系，具有一定的优越性[12]。虽然文献[11]主要探索了控制相同基因条件下牙周健康情况对动脉硬化的影响，但其没有讨论牙周治疗是否会对动脉硬化有所控制。而韩亚琨等人在孪生数据的基础上，继续讨论了这一点，根据 52 对同卵双生双胞胎动脉硬化患者牙周健康状况分为牙周病组和非牙周病组，通过对牙周病组治疗前后的动脉硬化情况数据的比较分析，得出了积极的牙周治疗可在一定程度上阻止动脉硬化的发展的结论[13]。

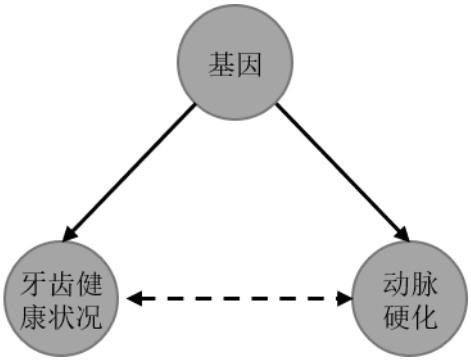

图 1-1 基因在动脉硬化和牙齿健康状况中可能的作用关系

当然，牙周炎和动脉硬化之间的相关关系并不是毫无争议的，首先是基因方面，许多研究报告发现，牙齿健康状况和动脉硬化均可在一定程度上归因于遗传因素[14-15]。也就是说可能遗传因素对动脉硬化的发展或者牙齿健康状况同时产生了影响，而并非牙齿健康状况本身直接影响了动脉硬化，示意图如图 1-1 所示。除此之外，还有研究质疑了这种相关关系的存在本身[16]。

本文使用参考文献[11]中的数据集来继续探究动脉硬化同口腔健康之间的关系。由于本文的数据集是双胞胎数据集，即我们取同性别的双胞胎的测量数据构成了数据集。考虑到一个双胞胎内数据的相关性，普通的线性模型因为假设观察值之间相互独立而不适用

于本数据结构，本文的第一部分延续文献[11]中的研究，采用广义估计方程的参数估计方法对建立的广义线性模型进行参数估计，这里的广义线性模型包括以 IMT 水平（颈动脉中膜层厚度）为因变量的线性模型和以 IMT 是否大于 1 形成二分类变量的 IMT01 为因变量的逻辑斯蒂模型。针对文献[11]中建立的模型的变量不显著的问题，我们引入变量间的交互效应，改进了原有的模型，使之更有说服力。同时，我们结合模型的拟合结果对交互效应项进行的具有实际医学意义的解读，更加深入地阐述了动脉硬化同牙齿健康状况间的联系。本文的第二部分着重考虑了文献[11]中未考虑的问题之一，同时也是之前指出的关于牙周炎和动脉硬化的相关性的争议之处——基因对于动脉硬化和牙齿健康状况的影响。针对这一情况，本文采用在双胞胎数据研究中常见的基于结构方程模型建立的基因—环境模型（又称 ACE 模型）来探究基因在动脉硬化和牙齿健康状况中扮演的角色。

2 模型方法概述

2.1 广义线性模型与广义估计方程

本节首先对广义线性模型进行简要概述，在引入广义线性模型的概念后，自然地导向处理广义线性模型中响应变量内部相关性的参数估计方法——广义估计方程的介绍。

2.1.1 广义线性模型

在一般的线性回归模型中，我们通常假设响应变量（因变量）服从正态分布，通过建立其均值和解释变量的线性组合的线性关系来研究响应变量和解释变量之间的相关性。然而在更加一般的假设中，响应变量并不一定服从正态分布，诸如当响应变量为分类变量或计数变量的情形。广义线性模型则拓展了普通线性模型，摆脱了普通线性模型中要求变量服从正态分布的限制，将其推广为更加一般的指数族分布。并且，其也允许处理响应变量的均值和被解释变量之间非线性的关系，通过连接函数建立响应变量的均值与解释变量的线性关系。

广义线性模型由线性预测部分、连接函数和随机成分三个部分组成。线性预测部分指代解释变量$\left(X_1, X_2, ..., X_p\right)$的线性组合$\beta_1 X_1 + ... + \beta_p X_p$；随机成分指代响应变量服从指数族概率分布；连接函数记为$\eta = g(\mu)$,其中$\mu = E(Y_i)$为响应变量的均值，连接函数表征了解释变量的线性组合和响应变量之间的关系。广义线性模型用数学形式表出：

- 假定 Y|X 服从指数族分布 $f(y;\theta,\phi) = exp(\frac{y\theta - b(\theta)}{\phi} + c(y,\phi))$，这里称$\theta$为自然参数，称$\phi$为尺度参数；
- 记 Y|X 的期望为$E(Y \mid X) = \mu$，一般来说，我们的目的是估计μ，而认为尺度参数已知；
- 解释变量 X 的线性组合记为η，即$\eta = \beta^T X = \beta_1 X_1 + ... + \beta_p X_p$，注意上述 Y|X 所服从的指数族分布中的参数$\theta$即为这里的$\eta$；
- 建立连接函数$g(E(Y \mid X)) = g(\mu) = \eta$。

对于广义线性模型，有以下几点模型假设：

- 响应变量$Y_1,...,Y_n$相互独立，模型估计误差相互独立；
- 解释变量可以是原始解释变量的非线性变化（如乘方）；
- 方差齐性不一定需要满足，模型误差需要相互独立，但不一定是正态分布；
- 由式推得$E(Y)=b^{'}(\theta)=\mu,Var(Y)=\phi b^{''}(\theta)=\phi V(\mu)$。

我们指出，普通线性模型，适用于响应变量为分类变量的逻辑斯蒂模型以及适用于响应变量为计数变量的泊松模型实际上都属于广义线性模型。

2.1.2 广义估计方程

广义估计方程是应用于广义线性模型的参数估计的一种方法，GEE 方法是基于面板数据和纵向数据等特殊的数据结构而提出的，在面板数据和纵向数据中，对于一个个体会有多次的重复测量值。在利用极大似然估计的广义线性模型中，我们知道，响应变量之间是相互独立的。然而在面板数据或纵向数据中，我们可以延续个体之间相互独立性这一假定，但一个个体的重复测量值之间可能是相关的，假使我们要将这种相关性考虑进参数估计中，原来的极大似然估计就不适用了。基于这个背景，Liang 和 Zeger 提出用广义估计方程衡量这种个体内的相关性[17]。这里我们先给出 GEE 模型的模型假设：

- 响应变量Y_i是聚类的或相关的，而不是相互独立的；
- 协变量可以是原始解释变量的非线性变化，可以包含交互效应；
- 方差齐性不需要满足，误差是相关的；
- 使用伪似然估计进行参数估计，而不是极大似然估计或最小二乘估计；
- 借助作业相关矩阵描述响应变量之间的相关性。

广义估计方程通过伪似然估计，引入作业相关矩阵来描述响应变量之间的相关性[18]。为说明 GEE 方法中的参数估计式，首先引入符号说明：设Y_{ij}代表第 i 个个体的第 j 次观测值，这里$i=1,2,...,N,\ j=1,2,...,n_i$，设有解释变量$x_{ij}$是$p\times1$的向量，则$Y_i=(Y_{i1},Y_{i2},...,Y_{in_i})^T$为$n_i\times1$维第 i 个个体的响应变量，$X_i=(x_{i1},...,x_{in_i})^T$为$n_i\times p$维第 i 个个体的解释变量。记$E(Y_{ij})=\mu_{ij}$。根据广义线性模型，有连接函数$g(\mu_{ij})=x_{ij}^T\beta$。我们借鉴普通线性回归中应用极大似然估计参数的式子估计广义线性模型的参数，有：

$$S(\beta)=\sum_{i=1}^{n}D_i^TV(Y_i)^{-1}(Y_i-\mu_i)=0 \tag{2.1}$$

这里D_i是$n_i \times p$维矩阵，其第 j 行元素为$\frac{\partial \mu_{ij}}{\partial \beta} = \dot{g}^{-1}(x_{ij}^T \beta)x_i$，这里$\dot{g}^{-1}(\cdot)$表示连接函数逆函数的一阶导数。我们称这种参数估计方法为伪似然估计。

由式 2.1 可知，我们要求参数β的估计，就要知道响应变量Y_i的协方差矩阵$V(Y_i)$。特别注意到根据广义估计方程的性质我们有$V(Y_{ij}) = \phi V(\mu_{ij})$。利用这一性质，当我们假定一个个体$Y_i$中的重复测量值间相互独立时，我们可以直接求得$V(Y_i) = \phi Diag(v(\mu_{i1},...,\mu_{in_i}))$。而当我们考虑这些重复观测值间的相关性时，我们将响应变量Y_i的协方差矩阵写为$V(Y_i) = A_i^{1/2} R_i(\alpha) A_i^{1/2}$，这里$A_i = Diag(v(\mu_{i1},...,\mu_{in_i}))$，我们称描述给定个体 i 的重复观测值间的相关系数矩阵$R_i(\alpha)$为Y_i的作业相关矩阵（Working Correlation Matrix）。

作业相关矩阵形式的选取以及参数α的估计即为 GEE 方法的精髓。表 2-1 总结了常见的作业相关矩阵形式。作业相关矩阵有结构性和非结构性两个类别，在结构性类别中，我们通常借助可交换相关，时间序列自相关，稳定相关来构造矩阵。其中可交换相关是相对直接，最常假设的结构，而对于由时间序列构成的纵向数据，我们可以考虑自相关，稳定相关是时间序列自相关的推广，将大于给定阶数的时间点上的观测值的相关性假定为 0，即认为一定时间间隔后，重复观测值之间不具有相关性。同时结构相关类中还有非稳定相关。除了结构性和非结构性的作业相关矩阵外，我们还可以从数据结构的本身出发，去构造合适的作业相关矩阵，以更合理地描述面板数据或纵向数据之间的相关关系。

表 2-1 常见作业相关矩阵结构

相关结构	$\mathrm{Cor}(Y_{iu}, Y_{iv})$	工作相关矩阵样例
可交换相关(Exchangeable)	$\begin{cases} 1 & u = v \\ 0 & u \neq v \end{cases}$	$\begin{pmatrix} 1 & \alpha & \alpha \\ \alpha & 1 & \alpha \\ \alpha & \alpha & 1 \end{pmatrix}$
自相关(AR(k))	$\alpha^{\lvert u-v \rvert}$	$\begin{pmatrix} 1 & \alpha & \alpha^2 \\ \alpha & 1 & \alpha \\ \alpha^2 & \alpha & 1 \end{pmatrix}$

一般认为，模型参数的估计值$\hat{\beta}$近似服从正态分布，从模型推导的角度上来看，我们有参数估计值的方差为$\sum_{i=1}^{n} D_i^T V_i^{-1} D_i$，这被称为基于模型的方差估计或者原始方差估计。

然而其并不能给出对参数β的标准差的准确估计，因为如果作业相关矩阵的假设形式不正确，那么这样的估计势必会带来误差，给我们判断参数估计的效果带来阻碍。但是，这一问题可以被我们称作鲁棒或三明治的方差估计解决：$\hat{V}_{LZ}=(\sum_{i=1}^{n}D_i^TV_i^{-1}D_i)^{-1}\hat{M}_{LZ}(\sum_{i=1}^{n}D_i^TV_i^{-1}D_i)^{-1}$，其中 $\hat{M}_{LZ}=\sum_{i=1}^{n}D_i^TV_i^{-1}Cov(Y_i)V_i^{-1}D_i$。注意到$Cov(Y_i)=(Y_i-\mu_i)(Y_i-\mu_i)^T$是对$V(Y_i)$的估计，即如果$V(Y_i)$正确给定了，那么$\hat{V}_{LZ}$将退化成$\sum_{i=1}^{n}D_i^TV_i^{-1}D_i$，即基于模型的方差估计。我们通常使用 Wald 检验来检验回归系数的显著性。

在模型的拟合优度的检验上，可以将广义线性模型中的 R^2 平移到的 GEE 模型中[19-20]。这里我们利用离差统计量给出伪 R^2 的定义。离差统计量$D(y;\mu)$等于全模型和已建立模型下对数似然函数的差的两倍，我们可以同一将伪 R^2 和针对多个解释变量调整后的伪 R^2 表示成如下形式：$R^2=1-\frac{D(y;\hat{\mu})}{D(y;\bar{\mu})}$ $R^2_{adjust}=1-\frac{(n-p-1)^{-1}D(y;\hat{\mu})}{(n-1)^{-1}D(y;\bar{\mu})}$。其中，根据被解释变量所假设的分布的不同，离差统计量$D(y;\mu)$分别具有不同的形式，如表 2-2 所示。

表 2-2 不同假设分布下的离差统计量

被解释变量的假设分布	离差统计量$D(y;\mu)$
正态分布	$\sum_i(y_i-\mu_i)^2$
逆高斯分布	$\sum_i(y_i-\mu_i)^2/y_i\mu_i^2$

同理，受限于似然函数的缺失，诸如 AIC 准则也不能用于 GEE 模型的模型选择中。针对这一问题，相关文献基于 AIC 准则提出了适用于 GEE 方法的 QIC 准则。QIC 准则的运用可以便于我们判断模型的好坏，从而找出关于响应变量的影响因素[21]。本文采用 QIC 和伪 R^2 来选择较优的模型。

2.2 结构方程模型与双胞胎基因环境模型

2.2.1 结构方程模型

如果说回归分析是一个探究变量间相关关系的常用手段，那么结构方程模型则是一个探究变量间因果关系的重要方法。在许多问题的研究中，我们不仅关心解释变量和被解释变量间的相关性，而可能更想知道二者间的因果联系，即具备一定的解释变量的条件能

否导致某一现象的发生。这种因果推断往往有别于一般的统计推断，在许多运用回归分析的医学统计研究中，研究者往往会强调他们建立的模型对于变量间因果关系的解释不足，正是由于普通的回归分析很难将变量间的相关关系转化为因果关系。当然，在计量经济学等学科的研究中，研究者往往通过工具变量的方法使得普通的回归模型具有因果关系的解释力。那么结构方程模型则是探究变量间因果关系的又一方法，它通过路径分析（Path Analysis）和潜变量加观测变量的结合为多元数据间的复杂结构的研究提供了一个灵活的框架，使得研究者可以通过他们构建的经验模型来判断这一因果解释模型的信服力[22]。简单说来，就是研究者首先通过路径分析的方法建立描述变量之间因果性和相关性的关系图，这种图被称为经验模型，那么我们结合结构方程的方法可以验证这种经验模型的适用性。在结构方程的构成中，路径分析是探究变量间直接或者间接因果关系的结构化的假设。相比于变量间简单的相关性的假设，路径分析的优点在于其可以解释变量间相关性的间接影响。在结构方程模型中，我们引入潜变量整合和解释观测变量中的信息以及观测变量内部的相关性，这一过程实际上就是验证性因子分析，我们称建立的模型为观测模型。同时，围绕不同的观测变量，我们可以建立一个或多个潜变量，多个潜变量之间我们也可以建立相关性或因果性。描述多个潜变量关系的模型称为结构模型。 除此以外，结构方程模型中定义了内生变量和外生变量的概念，内生变量是指受其他变量影响的变量，类似于回归分析中的因变量，当然，内生变量本身也可以影响其他变量，而外生变量是指起解释作用的变量，类似于回归分析中的解释变量，它们只影响别的变量，自身不受其余变量的影响。注意到内生变量和外生变量是一个相对性的概念，即在一个模型中的一部分来看，一个变量可能是内生的，但是在另一部分中，它扮演着外生的角色。一个简易的结构方程模型的路径分析图如图 2-1 所示，其中的x_1, x_2, x_3即为内生观测变量，它们之间可以存在相关性，而 x 是通过验证性因子分析提取的外生潜变量，这个潜变量实际上解释了我们的因变量 y，这是一个典型的观测的解释变量和因变量存在间接关系的模型，注意图中的箭头方向表示的是因果关系，这个模型可以用来进行因果推断。

结构方程模型常用极大似然估计的方法进行参数估计，在模型评价方面，我们可以采用衡量基于观察值的协方差矩阵和基于模型推导的协方差矩阵差别的卡方检验和衡量模型估计值和实际观察值的均方根误差等指标。得益于结构方程模型的直观解释和因果推断的能力，其在健康科学，生物信息，医学和传染病学中有着重要应用[23-25]。

2.2.2 双胞胎基因环境模型（ACE 模型）

在生物统计基因研究中，我们借助结构方程模型可以探究基因和环境的影响对变量和变量间相关关系的影响[25-27]。我们往往借助双胞胎数据对这一问题进行研究。在双胞胎数据中，研究者往往会分别围绕同卵双胞胎和异卵双胞胎采集实验数据，因为这样的数据结构可以帮助研究者解构基因和环境分别对一个特征的影响。这里我们可以把需要研究的携带特征的变量称为观测变量，而基因和环境可以作为潜变量，由此构建的结构方程模型称为双胞胎基因环境模型。

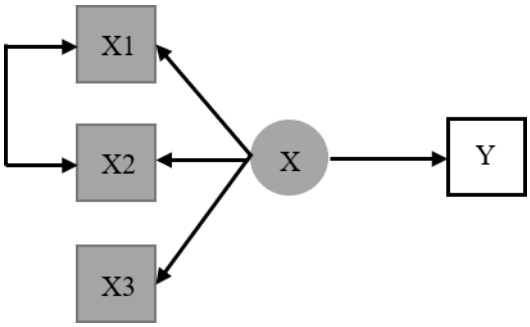

图 2-1 简单的结构方程模型

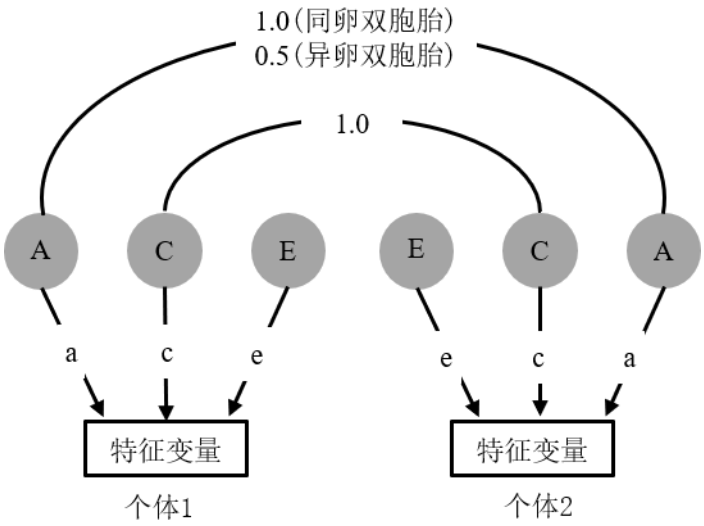

图 2-2 单变量 ACE 模型路径图

在基因环境模型中的一种典型模型是 ACE 模型，我们用 A 表示基因影响因子，用 C 表示双胞胎所处的共同的环境影响因子，用 E 表示双胞胎各异的环境影响因子[26-27]。从回归分析的角度上看我们有如下方程：$P=\beta_1 A+\beta_2 C+\beta_3 E$。实际上，ACE 模型是通过拆解我们要研究的特征变量的方差来实现的，我们将一对双胞胎内的单个个体的特征变量的方差记为Σ_p，特征变量的差异由基因和环境导致的差异共同组成，即有$\Sigma_p=\Sigma_A+\Sigma_C+\Sigma_E$。同卵双胞胎之间分享全部的基因和共同环境因素，而异卵双胞胎之间分享 50%的基因和全部的共同环境因素，因此一对同卵双胞胎中的两个个体之间的特征变量的协方差$Cov(P_{MZ1},P_{MZ2})=\Sigma_A+\Sigma_C$，而一对异卵双胞胎中的两个个体之间的特征变量的协方差等于$Cov(P_{DZ1},P_{DZ2})=0.5\Sigma_A+\Sigma_C$。综上，我们有同卵双胞胎和异卵双胞胎特征变量的协方差矩阵：$Cov_{MZ}=\begin{pmatrix}\Sigma_A+\Sigma_C+\Sigma_E & \Sigma_A+\Sigma_C\\ \Sigma_A+\Sigma_C & \Sigma_A+\Sigma_C+\Sigma_E\end{pmatrix}$

$Cov_{DZ}=\begin{pmatrix}\Sigma_A+\Sigma_C+\Sigma_E & 0.5\Sigma_A+\Sigma_C\\ 0.5\Sigma_A+\Sigma_C & \Sigma_A+\Sigma_C+\Sigma_E\end{pmatrix}$ 。为便于后续说明，我们简记以上两个矩阵为：

$$Cov_{MZ}=\begin{pmatrix}A+C+E & A+C\\ A+C & A+C+E\end{pmatrix},\ Cov_{DZ}=\begin{pmatrix}A+C+E & 0.5A+C\\ 0.5A+C & A+C+E\end{pmatrix} \tag{2.2}$$

用路径分析的理论，我们作出单变量下的 ACE 模型的路径图如图 2-2 所示。根据路径系数的定义，我们可以把同卵双胞胎和异卵双胞胎的协方差矩阵写为路径系数的形式：

$$Cov_{MZ}=\begin{pmatrix}a^2+c^2+e^2 & a^2+c^2\\ a^2+c^2 & a^2+c^2+e^2\end{pmatrix},\ Cov_{DZ}=\begin{pmatrix}a^2+c^2+e^2 & 0.5a^2+c^2\\ 0.5a^2+c^2 & a^2+c^2+e^2\end{pmatrix}$$

另外，我们还可以定义基因和环境因素分别占总变量方差的百分比作为直观评价基因和环境对特征变量影响程度的指标。

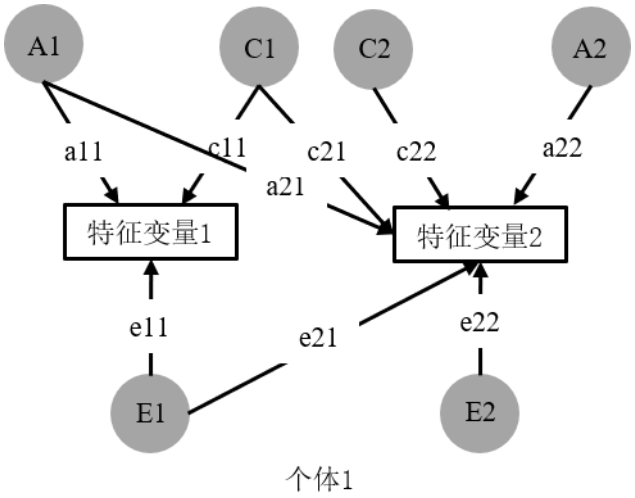

图 2-3 双变量 ACE 模型路径图

我们容易把单变量 ACE 模型推广到多变量的情形，下面以二变量作简要说明。同样的，我们依然延续之前用基因和环境因素拆解变量方差的思路，根据图 2-3 的双变量 ACE 模型的路径图，结合协方差矩阵正定性的要求，我们利用 Cholesky 分解的形式，定义由基因和环境分别解释的协方差成分为

$$\Sigma_A=aa^T,a=\begin{pmatrix}a_{11} & 0\\ a_{21} & a_{22}\end{pmatrix}\ \Sigma_C=cc^T,c=\begin{pmatrix}c_{11} & 0\\ c_{21} & c_{22}\end{pmatrix}\ \Sigma_E=ee^T,e=\begin{pmatrix}e_{11} & 0\\ e_{21} & e_{22}\end{pmatrix}$$

进一步地，我们可以得到同卵双胞胎和异卵双胞胎的特征变量协方差矩阵同式 2.2。

ACE 模型的参数估计也是通过极大似然估计得到的。在模型检验方面，我们可以利用似然比检验，将我们设计的模型同全模型进行比较，似然比检验的显著性在一定程度上可以说明模型的合理性。当然，更为直观的，我们可以把模型拟合的协方差矩阵同样本观察值计算的协方差矩阵进行直接比较，以说明模型的拟合效果。

3 牙齿健康状况与患动脉硬化风险相关性探究

在本节中我们首先对数据集进行探索性分析，着重探究在文献[11]的基础上引入交互项的意义，之后利用广义估计方程对分别以 IMT 水平（连续变量）和 ITM01（二分类变量）为被解释变量进行参数估计，接着围绕模型拟合与检验，结果解读等角度进行了详细讨论。

3.1 描述性统计与探索性分析

3.1.1 变量说明与数据选取

本文的数据集变量说明如表 3-1 所示。关于此数据集更加详细的变量采集过程参见文献[11]，这里不做赘述。我们用医学中的 IMT 水平即颈动脉中膜层厚度作为衡量动脉硬化的指标，一般地，颈动脉中层厚度最大值超过 1 被诊断为动脉硬化。同时，我们用牙齿数量，牙齿咀嚼能力得分和牙周袋平均深度作为衡量牙齿健康状况的三大指标。我们同时考虑将年龄，性别，体重指数和吸烟状态作为解释变量。我们一共有 290 个样本观察值，即 145 对相同性别的双胞胎数据，其中的 129 对为同卵双胞胎，16 对为异卵双胞胎。值得说明的是，为了后续研究的方便，我们将咀嚼能力得分同时划分为了三个档次分别为 0-3，4-6 和 7-9，变成一个新的分类变量。

表 3-1 变量说明

变量名		符号	说明
被解释变量	颈动脉中膜层厚度	IMT	取值>0，单位为 mm
		IMT01	当 IMT>1 时，IMT01=1；当 IMT<1 时，IMT01=0
解释变量	牙齿数量	totaltooth	记 5 个牙齿为一个单位的 NT
	牙齿咀嚼能力得分	scoreMP	得分取值为 0-9 的整数
	牙周袋平均深度	avePD	单位为 mm
控制变量	年龄	age	
	性别	sex	
	身体质量指数	BMI	单位为 kg/m^2
	吸烟状态	smoking	分为从不吸烟，从前吸烟和现在吸烟三类

为便于接下来的分析，我们进行缺失数据统计，在此数据集中有缺失值的变量为咀嚼能力得分和牙周袋平均深度，在后续分析中，我们分别删去存在缺失值的样本观察值。

在第一阶段中，我们考察动脉硬化同牙齿健康状况间的关系，为了避免基因的干扰，我们只选择同卵双胞胎的数据。这是由于，在基因解释的理论中，如果基因与动脉硬化或牙齿健康状况有关，那么基因的混杂会影响动脉硬化和牙齿健康状况之间相关性的强弱，同卵双胞胎由于比异卵双胞胎有更强的基因同源性，可能会造成二者之间相关性的增强。同时，我们考虑只采用年龄大于 50 岁的样本观察值。这是由于大于 50 岁前后会显著地呈现人体生理健康的变化。在图 3-2 中，我们展示了牙齿健康状况在两个年龄组中的分布情况，Wilcox 检验的显著性均表明两个年龄组的牙齿健康状况有显著不同且大于 50 岁年龄组的牙齿健康状况劣于小于等于 50 岁年龄组。在图 3-1 中，我们展现了 50 岁前后两组

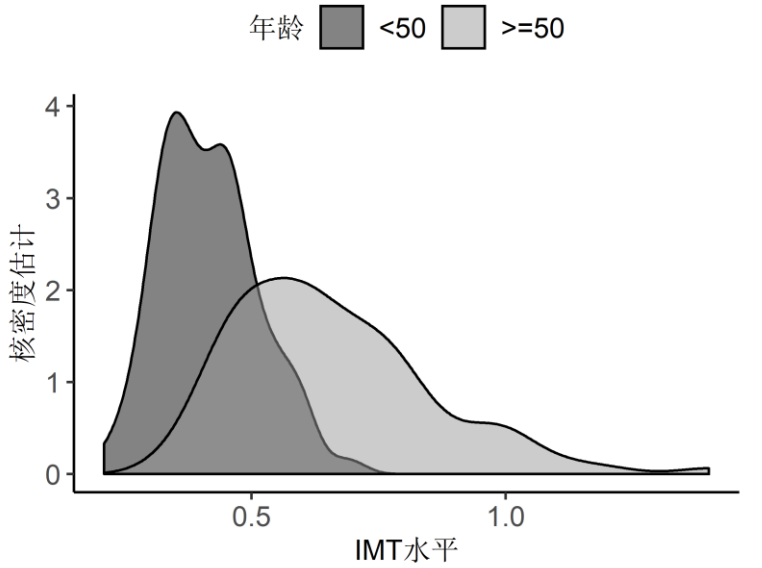

图 3-1 50 岁前后人群 IMT 水平差异

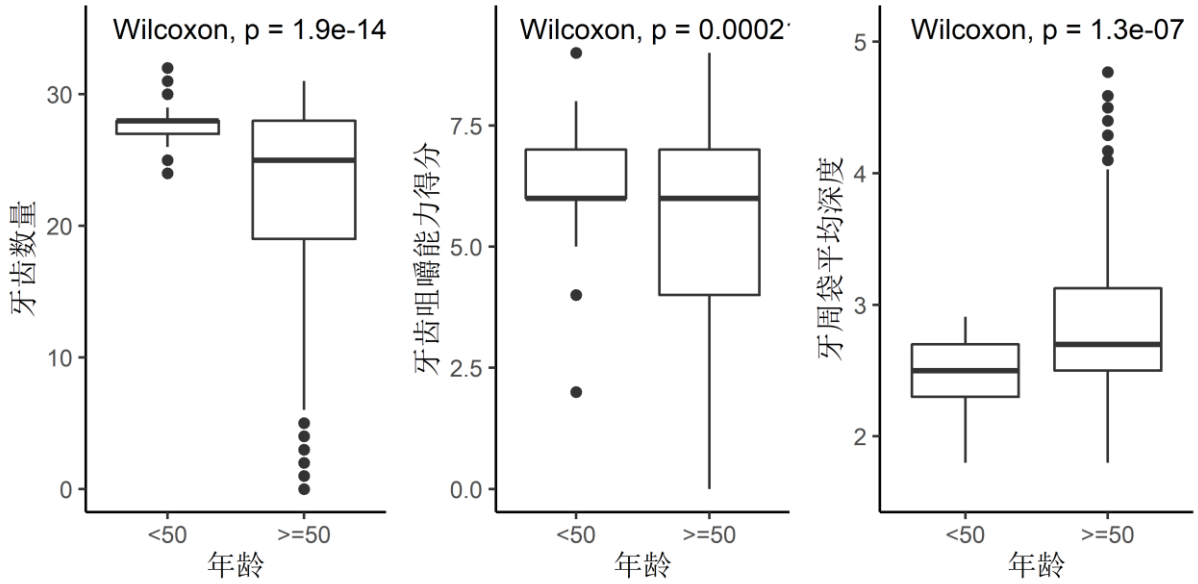

图 3-2 50 岁前后人群牙齿健康状况差异

样本的 IMT 水平分布图（采用核密度估计拟合），可以发现二者的位置参数和尺度参数有明显的不同，大于 50 岁组的 IMT 水平中位数高于小于 50 岁组，同时，大于 50 岁组的 IMT 水平的分布更加分散，预示着患动脉硬化的风险更高。因此，研究大于 50 岁组的样本更具有参考性，更加能揭示动脉硬化与牙齿健康状况之间的关系。

基于上述分析，我们选取大于 50 岁的个体 212 个，共计 106 对双胞胎。其中同卵双胞胎为 91 对，异卵双胞胎为 15 对，描述性统计分析见表 3-2。

表 3-2 大于 50 岁样本描述性统计

变量		全部（n = 212)		同卵双胞胎 （n = 182）		异卵双胞胎（n = 30）	
		均值	标准差	均值	标准差	均值	标准差
年龄		67.42	10.03	66.97	10.08	70.13	9.42
身体质量指数		22.49	3.14	22.44	3.14	22.77	3.17
牙齿数量		21.08	9.25	21.13	9.15	20.80	10.03
咀嚼能力得分		5.29	2.13	5.36	2.01	4.90	2.73
牙周袋平均深度		2.84	0.55	2.83	0.54	2.92	0.63
IMT 水平		0.67	0.24	0.66	0.24	0.71	0.24
		数量	比例	数量	比例	数量	比例
性别	男	84	39.6%	68	37.4%	16	53.3%
	女	128	60.4%	114	62.6%	14	46.7%
吸烟状态	从不吸烟	138	65.1%	120	65.9%	18	60.0%
	过去吸烟	44	20.8%	40	22.0%	4	13.3%
	当前吸烟	30	14.2%	22	12.1%	8	26.7%
咀嚼能力得分	0-3	44	20.8%	33	18.1%	11	36.7%
	4-6	101	47.6%	91	50.0%	10	33.3%
	7-9	67	31.6%	58	31.9%	9	30.0%

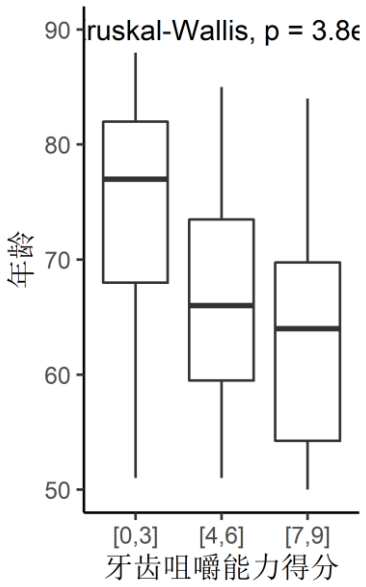

图 3-3 年龄和咀嚼能力得分的关系

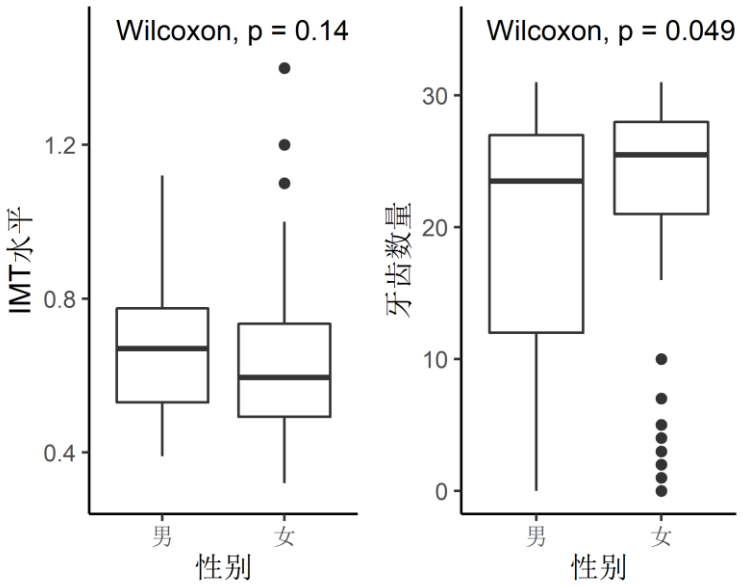

图 3-4 IMT 水平和牙齿数量的分性别差异

3.1.2 大于 50 岁样本的探索性分析

我们考察的重点是牙齿健康状况和 IMT 水平的关系，但是文献 1 中仅仅把所有变量纳入模型中的做法并不优，其系数的不显著性降低了模型的解释力。这是由于年龄，性别和吸烟状况会作为重要的调停变量参与二者相关性的影响中，例如，牙齿的健康状况受年龄的调节影响很大，高年龄造成牙齿健康状况的恶化，导致动脉硬化风险升高，而低年龄不造成牙齿健康状况发现显著变化，就可能埋没动脉硬化和牙齿健康状况间存在的潜在相关性。因此我们考虑在模型中引入交互效应，将调停变量与牙齿健康状况进行交互来改进原有的模型。 为了研究哪些调停变量和哪些牙齿健康状况指标可能具有交互，我们首先进行探索性分析。

首先我们考虑性别，图 3-4 表明在不同性别中 IMT 水平的差异并不明显，而在牙齿健康方面，在牙齿数量和牙周袋平均深度上呈现显著的性别差异，而在咀嚼能力得分上的差异并不显著。在吸烟状态方面，女性大多数没有吸烟，而吸烟状态在男性中的分布较为分散。其次我们考虑年龄。我们发现牙齿数量和年龄间呈现明显的正相关，而牙周袋深度和年龄间呈现明显的负相关。虽然咀嚼能力得分和年龄的关系并不明显，可当我们将咀嚼能力得分分组化后可以发现，咀嚼能力得分和年龄也呈现一定程度上的负相关（如图 3-3）。接下来，我们考虑吸烟状况。图 3-5 表明不同吸烟状况和 IMT 的水平间的差异是显著的，不吸烟组的 IMT 水平的中位数低于吸烟组。在牙齿健康状态上，不同吸烟状态在牙齿数量上的差异是显著的，吸烟组中牙齿数量的波动显著高于不吸烟组；不同吸烟状态在牙周袋深度上的差异也是显著的，不吸烟组的牙周袋平均深度低于吸烟组。而在咀嚼能力得分上吸烟组和不吸烟组的差异并不显著。

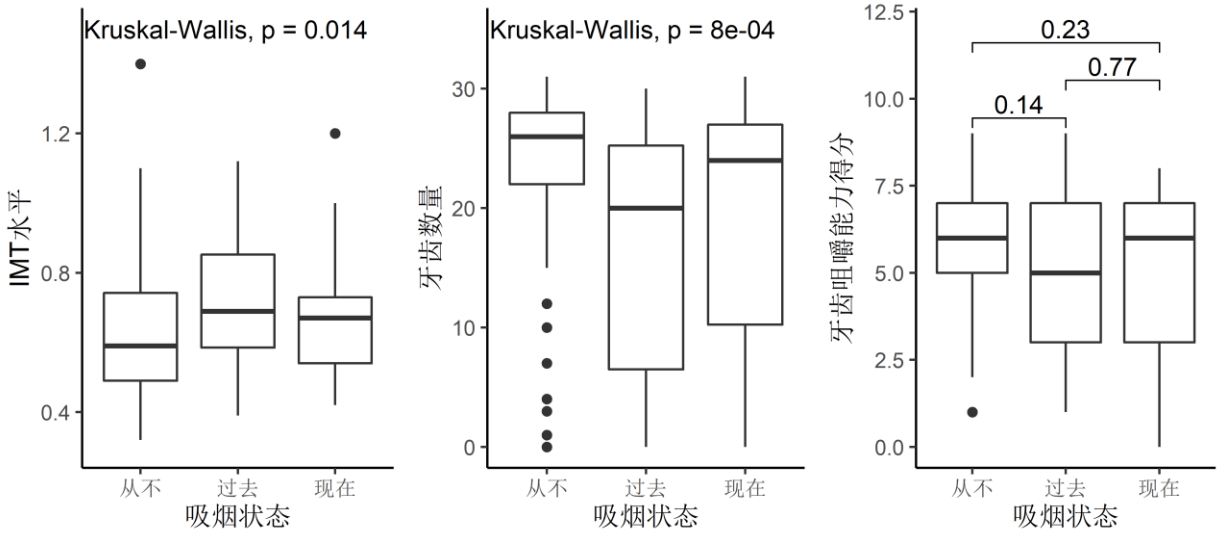

图 3-5 吸烟状态和 IMT 水平及牙齿健康状态的关系

除了上述考虑外，为了建立回归模型，我们还需要对大于 50 岁样本的 IMT 值的分布进行假定。当然，简单起见我们可以直接假定在解释变量和控制变量的影响下 IMT 水平服从正态分布。但是，从图 3-1 和 IMT 水平的定义上看，这是一个只取正值且明显有偏的连续型分布。为了更加具体的进行分布假定，我们考察被解释变量的均值和方差的关系。依据年龄，性别和吸烟状态这三个关键的控制变量，我们将 IMT 水平进行分组，计算组均值和组内方差。进一步，我们将组内均值和组内方差取对数后发现两者之间呈现线性关系，即有 $\ln V(\mu)=k\ln\mu$，因此有 $V(\mu)=\mu^k$，由此我们假定 IMT 水平的分布为逆高斯分布[28]。

3.2 含交互效应的广义估计方程模型

上述的探索性分析充分说明了我们在模型中引入交互效应从而考虑调停变量的影响的合理性。经过模型系数的显著性检验的选择，我们分别使用 GEE 方法，任选三种反映牙齿健康状态变量中的一种，和其他控制变量一起组成解释变量。被解释变量中我们分别选择 IMT 水平和 IMT01 建立模型，针对连续型分布的 IMT 水平我们分别采用了正态分布和逆高斯分布的假定，针对离散型分布的 IMT01 我们采用二项分布的假定。我们从变量的显著性和模型的拟合优度两个方面来进行模型选择，并针对拟合系数进行了医学意义上的解读。

3.2.1 以 IMT 水平为被解释变量的模型建立与选择

表 3-3 基于 IMT 水平正态分布假定的 GEE 估计结果

反映牙齿健康状况变量		牙齿数量		咀嚼能力得分		牙周袋平均深度	
解释变量		系数估计	P 值	系数估计	P 值	系数估计	P 值
（截距项）		0.672	0.175	1.292	0.059	-0.112	0.884
性别	男性						
	女性	-0.249	0.088	-0.258	0.079	0.033	0.524
年龄		0.005	0.137	-0.005	0.308	0.011	0.420
吸烟状态	从不吸烟						
	从前吸烟	-0.019	0.792	0.060	0.095	0.060	0.095
	现在吸烟	0.185	0.510	0.093	0.111	0.155	0.035
身体质量指数		-0.005	0.612	-0.005	0.634	-0.008	0.544
牙齿健康状况		-0.056	0.036	-0.213	0.008	0.147	0.698
女性×牙齿健康状况		0.059	0.021	0.052	0.018		
年龄×牙齿健康状况				0.003	0.004	-0.001	0.788
从前吸烟×牙齿健康状况		0.019	0.246				
现在吸烟×牙齿健康状况		-0.028	0.659				

在各个控制变量种类不改变的前提下，我们需要先确定三个反映牙齿健康状况的变量中哪一个作为解释变量最合适。在 IMT 水平为正态分布假定的模型中，我们可以看到，当我们将咀嚼能力得分作为衡量牙齿健康状况的指标时，显著的解释变量最多，在显著性水平为0.05时，咀嚼能力得分以及其与性别和年龄的交互都显著，在显著性水平为0.1时，性别项和吸烟状态项中的从前吸烟一项也显著。而当我们将牙齿数量作为衡量牙齿健康状况的指标时，在显著性水平为 0.1 的情形下也只有性别，牙齿数量和牙齿数量与性别的交互项是显著的。当我们将牙周袋平均深度作为衡量牙齿健康状况的指标时，这一指标并不显著。由此可见，这三个模型中，展现解释变量的显著性最好的是以咀嚼能力得分作为衡量牙齿健康状况的指标的模型。

下面我们在确定解释变量为牙齿咀嚼能力得分的条件下，评估 IMT 水平分布分别为正态分布假定和逆高斯分布假定下的回归效果。我们对 IMT 水平正态分布假定和逆高斯分布假定在不同的连接函数（包括相等连接和均值的负二次方连接）下进行了回归系数的 GEE 估计，结果如表 3-4 所示。在正态分布假定和采用相等连接的逆高斯分布的假定模型中，逆高斯分布假定下系数拟合的标准差明显较小，从伪 R^2 的角度上看，逆高斯分布假定的拟合效果也优于高斯分布。这里需要指出，从 QIC 的角度上看，采用均值的负二次方连接的逆高斯分布假定优于采用相等连接的逆高斯分布。但是，为了系数更加便于理解，我们选择采用相等连接的逆高斯分布假定。

表 3-4 以咀嚼能力得分为解释变量，在不同分布假定下的 GEE 模型

解释变量	正态分布（link=identity）		逆高斯分布（link=identity）		逆高斯分布 (link=1/mu^2)	
	系数估计	标准差	系数估计	标准差	系数估计	标准差
（截距项）	1.292	0.684	0.938	0.426	-2.049	2.973
性别 男性						
女性	-0.258	0.147	-0.158	0.122	1.402*	0.596
年龄	-0.005	0.005	-0.002	0.004	0.034	0.024
吸烟 从不吸烟						
状态 从前吸烟	0.060	0.036	0.049	0.036	-0.363	0.227
现在吸烟	0.093	0.058	0.047	0.042	-0.410	0.315
身体质量指数	-0.005	0.010	-0.002	0.006	0.049	0.054
咀嚼能力得分	-0.213*	0.081	-0.167*	0.049	1.418*	0.377
年龄×咀嚼能力得分	0.003*	0.001	0.002*	0.001	-0.018*	0.005
性别×咀嚼能力得分	0.052*	0.022	0.032*	0.017	-0.300*	0.106
QIC	194.16		-2106.48		-2430.66	
伪 R^2	0.246		0.314		0.313	
调整的伪 R^2	0.203		0.281		0.279	

3.2.2 以 IMT01 为被解释变量的模型建立与选择

表 3-5 基于 IMT01 的 GEE 估计结果

反映牙齿健康状况变量		牙齿数量		咀嚼能力得分		牙周袋平均深度	
解释变量		系数估计	P 值	系数估计	P 值	系数估计	P 值
（截距项）		-7.0223	0.3504	-12.5385	0.0486	-20.3831	0.095
性别	男性						
	女性	-1.7194	0.067	-2.32	0.0588	-16.7587	0.05
年龄		0.0676	0.2292	0.1101	0.0066	0.1668	0.084
吸烟状态	从不吸烟						
	从前吸烟	-0.6995	0.4929	1.1932	0.1692	5.4989	0.03
	现在吸烟	-0.0879	0.9469	2.2626	0.0281	8.1237	0.022
身体质量指数		0.0594	0.7104	0.0704	0.6292	0.1422	0.615
牙齿健康状况		-0.9864	0.0028	-0.2245	0.267	-1.1289	0.461
女性×牙齿健康状况		0.5982	0.0057	0.5202	0.0154	5.7861	0.051
从前吸烟×牙齿健康状况		0.5983	0.0412				
现在吸烟×牙齿健康状况		0.6379	0.0787				
QIC		81.9		75.3			
准确率		0.857		0.713			

在以 IMT01 为被解释变量并假定其为二项分布的 GEE 估计模型如表 3-5 中，可以看到，当我们将牙齿数量作为衡量牙齿健康状况的指标时，在显著性水平为 0.05 时，牙齿数量本身以及其与性别，年龄和吸烟状态的交互均显著；当我们将咀嚼能力得分作为衡量牙齿健康状况的指标时，在显著性水平为 0.05 时，年龄和吸烟状态中的现在吸烟都显著。同时，虽然咀嚼能力得分本身未体现显著性，但其与年龄的交互是显著的。当我们将牙周袋平均深度作为衡量牙齿健康状况的指标时，发现牙周袋深度本身以及其与其他变量的交互都未体现显著性。由此可见，以牙齿数量和咀嚼能力得分分别作为牙齿健康状况的指标因素的模型对变量显著性的展示较好。在模型的拟合优度上，我们计算了 QIC 值，同时我们得到模型预测结果的混淆矩阵，并进一步得到准确率，如表 3-5。我们发现二者的 QIC 值类似，而准确度上以牙齿数量作为衡量牙齿健康状况的指标的模型更好。综上所述，我们在以 IMT01 为被解释变量的 GEE 模型中选择以牙齿数量作为衡量牙齿健康状况的指标的模型。

3.2.3 模型结果理解

由 3.2.1 和 3.2.2 节，我们分别建立了以 IMT 水平为被解释变量，假定为逆高斯分布的 GEE 模型（简称 IMT 模型）和以 IMT01 为被解释变量，假定为二项分布的 GEE 模型

（简称 IMT01 模型）。表 3-6 展示了两个模型具体的回归系数和 95%的置信区间，值得注意的是，为了解读 IMT01 模型的回归系数的现实意义，表中显示的实际上是几率比（odds ratio），即等于以自然常数作底数，模型回归系数作指数计算的数，它表征了相较于对照组的事件发生的几率，实验组的时间发生的几率的大小。IMT 模型中的回归系数表明，在其他变量不变的情况下，牙齿咀嚼能力每减少 1 个得分，IMT 水平将升高 0.167。在 IMT01 模型中，在其他变量都相同的条件下，女性比男性患动脉硬化的几率降低 82.1%；在其他变量都相同的条件下，牙齿咀嚼能力每提高一个得分，患动脉硬化的几率将降低 62.7%。

表 3-6 模型选择后的回归结果

解释变量		IMT 模型（咀嚼能力得分）			IMT01 模型（牙齿数量）		
		系数	95%置信区间	P 值	exp(系数)	95%置信区间	P 值
（截距项）		0.938	(0.103,1.772)	0.028	0.001	(0.000,2242.000)	0.350
性别	男性						
	女性	-0.158	(-0.396,0.080)	0.194	0.179	(0.029,1.130)	0.067
年龄		-0.002	(-0.016,0.005)	0.292	1.070	(0.958,1.190)	0.229
吸烟状态	从不吸烟						
	从前吸烟	0.049	(-0.011,0.109)	0.107	0.497	(0.067,3.670)	0.493
	现在吸烟	0.047	(-0.036,0.130)	0.271	0.916	(0.069,12.200)	0.947
身体质量指数		-0.002	(-0.015,0.010)	0.729	1.060	(0.775,1.450)	0.710
牙齿健康状况		-0.167	(-0.263,-0.070)	0.001	0.373	(0.195,0.712)	0.003
女性×牙齿健康状况		0.032	(0.001,0.065)	0.054	1.820	(1.190,2.780)	0.006
年龄×牙齿健康状况		0.002	(0.001,0.003)	0.000			
从前吸烟×牙齿健康状况					1.820	(1.020,3.230)	0.041
现在吸烟×牙齿健康状况					1.890	(0.930,3.850)	0.079

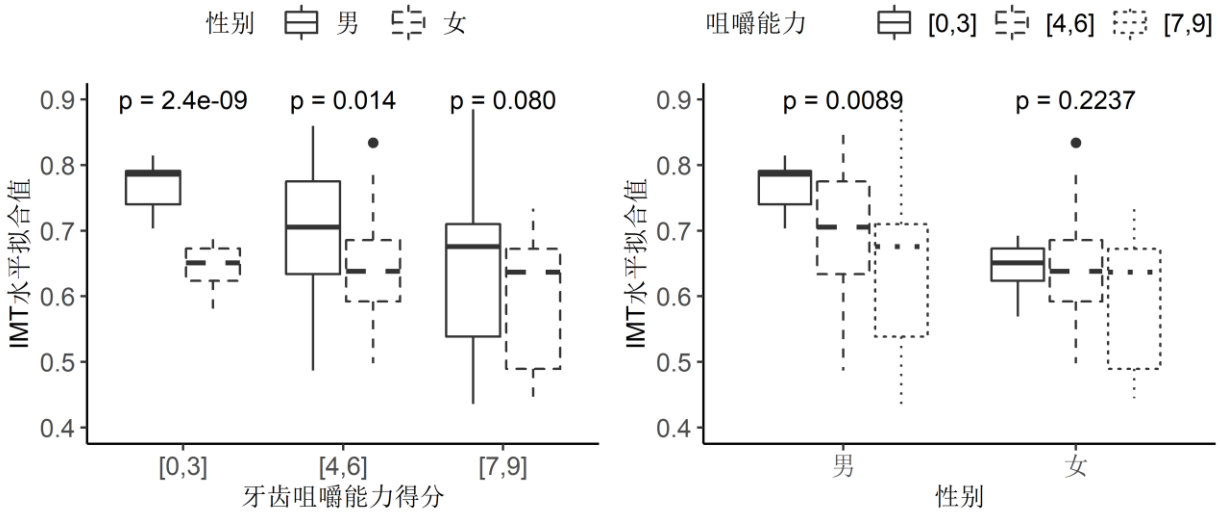

图 3-6 在性别的作用下咀嚼能力得分和 IMT 拟合值间的关系

下面我们利用交互图的方式对交互效应项进行解读。在咀嚼能力得分的 IMT 模型中，

我们可以把男女的咀嚼能力得分和模型的 IMT 水平拟合值作箱线图，如图 3-6 所示。可以看出在咀嚼能力得分为 0-3 和 4-6 的组中，男女的 IMT 水平存在显著差异，且咀嚼能力对于男性而言更加明显地影响了颈动脉膜层厚度。实际上这一点也可以在以 IMT01 为被解释变量，咀嚼能力得分为解释变量的模型中得到印证。进一步考察，我们将男性样本与女性样本分开，结合原有的模型分别回归，结果如表 3-7。从系数中可见，咀嚼能力得分都显著。对男性而言，在其他变量不变的情况下，咀嚼能力得分每升高一个单位，IMT 水平的平方的倒数上升 1.346；对女性而言，在其他变量不变的情况下，咀嚼能力得分每升高一个单位，IMT 水平的平方的倒数上升 0.864。由此可见咀嚼能力对男性动脉硬化的影响更大。

表 3-7 分性别，采用逆高斯分布假定的 IMT 模型（link=1/mu^2）

解释变量	男性		女性	
	系数估计值	P 值	系数估计值	P 值
（截距项）	-4.161	0.086	2.663	0.354
年龄	0.045	0.062	0.018	0.629
从前吸烟	-0.721	0.037	-0.181	0.554
现在吸烟	-0.314	0.434	-0.607	0.049
身体质量指数	0.118	0.007	-0.054	0.298
咀嚼能力得分	1.346	0.000	0.864	0.048
年龄×咀嚼年龄得分	-0.167	0.000	-0.014	0.047

同样地，我们可以把年龄分成一定的年龄段，通过交互图的方式研究年龄与牙齿咀嚼能力的交互作用。我们把大于 50 岁的样本以每十岁作为一个划分，得到 4 个年龄段，由此得到图 3-7。可以看到，在咀嚼能力和 IMT 水平之间，年龄起到了明显的调节作用，除

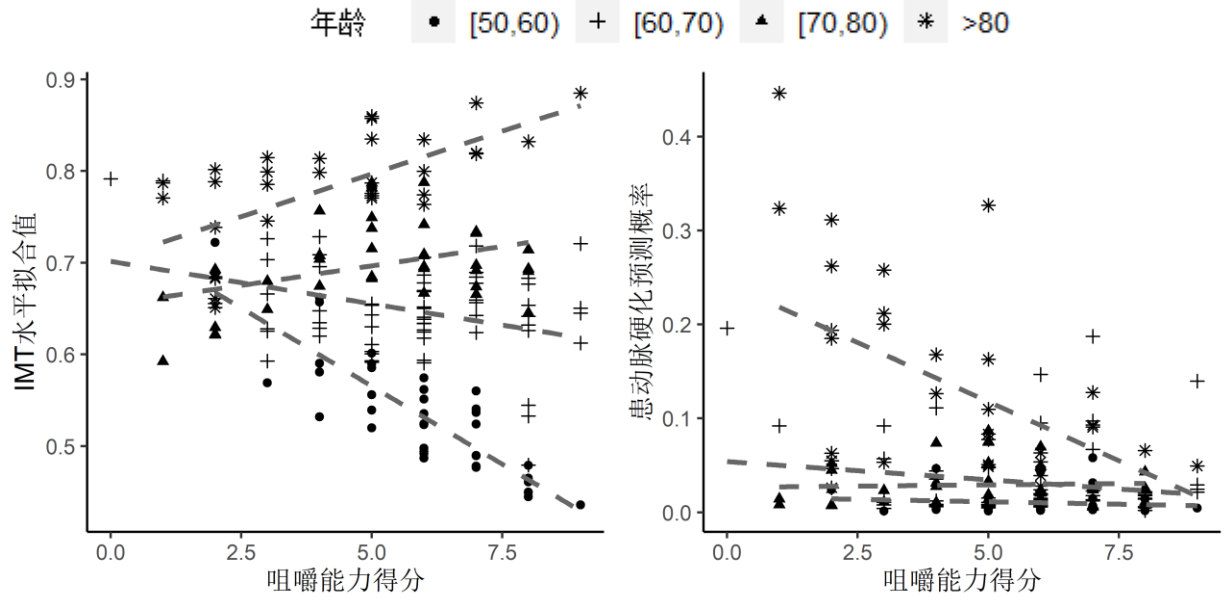

图 3-7 在年龄作用下咀嚼能力得分和 IMT 水平间关系

了 70-80 岁以及 80 岁以上的人群中这种相关性是相似的（因为两条拟合线较为平行），其他的年龄段与年龄段之间均有差异。另外，在 50 岁到 60 岁的人群中，咀嚼能力得分和 IMT 水平的负相关性是最明显的。值得注意的是，虽然在大于 70 岁的人群中拟合线呈现上升态势，我们依然不能得出这部分人群中咀嚼能力和 IMT 水平呈正比的趋势，因为这样的样本点并不多。实际上，在咀嚼能力的 IMT01 模型中，我们通过交互图 3-7 发现，大于 80 岁的人群中，随咀嚼能力的下降，患动脉硬化的概率的攀升是最明显的。在表 3-7 的两个回归模型中，我们剔除了性别的交互，年龄和 IMT 水平的交互依然显著。通过交互图 3-8，我们可以看到，在男性群体中，较低年龄中牙齿咀嚼能力和 IMT 的负相关性确实较为明显，较高年龄组中没有明显的关系。而在女性群体中，似乎呈现出咀嚼能力和 IMT 水平呈现正相关的性质，那么对于女性群体而言，牙齿的咀嚼能力可能就不是判断动脉硬化的一个较好的指标了。

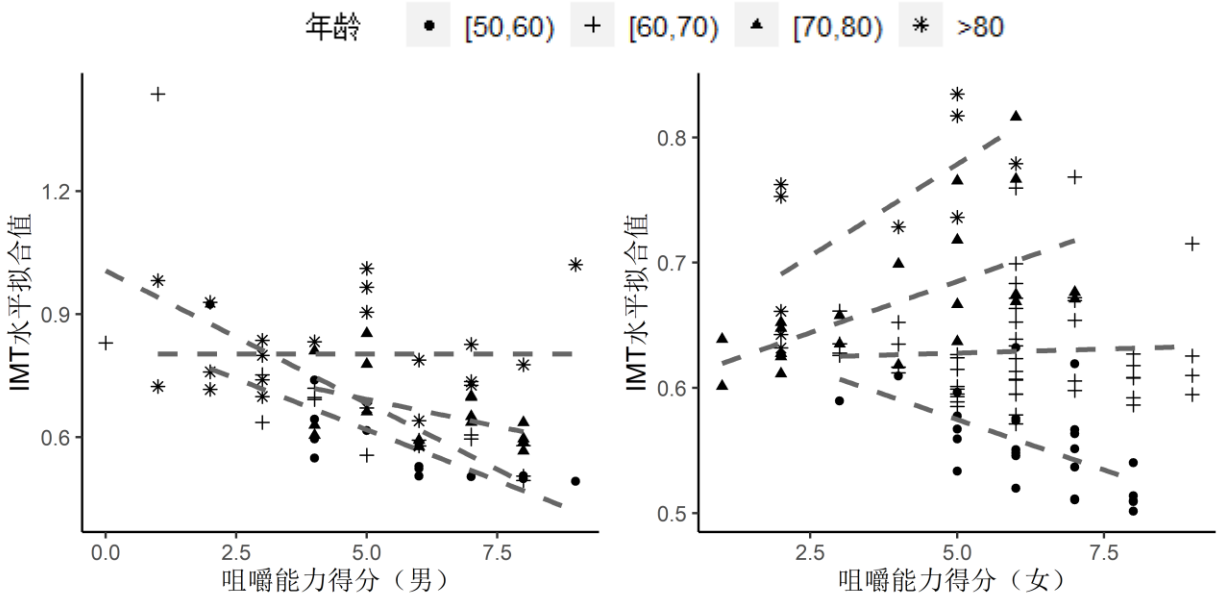

图 3-8 分性别下不同年龄段群体的牙齿咀嚼能力得分和 IMT 水平的关系

在牙齿数量的 IMT01 模型中，我们作牙齿数量和患动脉硬化预测概率的分性别交互图 3-9，可以发现随着牙齿数量的减少，动脉硬化的风险会升高，且男性群体的风险升高的比女性群体要快。另外，由于模型中吸烟状态和牙齿数量的交互显著，我们也可以考察这两个因素之间的交互。如图 3-9 所示，我们看到三种吸烟状态的调节下，牙齿数量和患动脉硬化的概率的相关性各有差异，其中在从前吸烟组中，二者的负相关性最强。当然，同时可以看出，吸烟会明显使得动脉硬化的风险升高。由于样本量的限制以及样本量中患动脉硬化的样本点（即 IMT>1）的数量太少，我们没有办法对 IMT01 模型的交互项作定量上更细致的解读，当然，这一缺点也降低了 IMT01 模型的可取性。因此，我们的模型

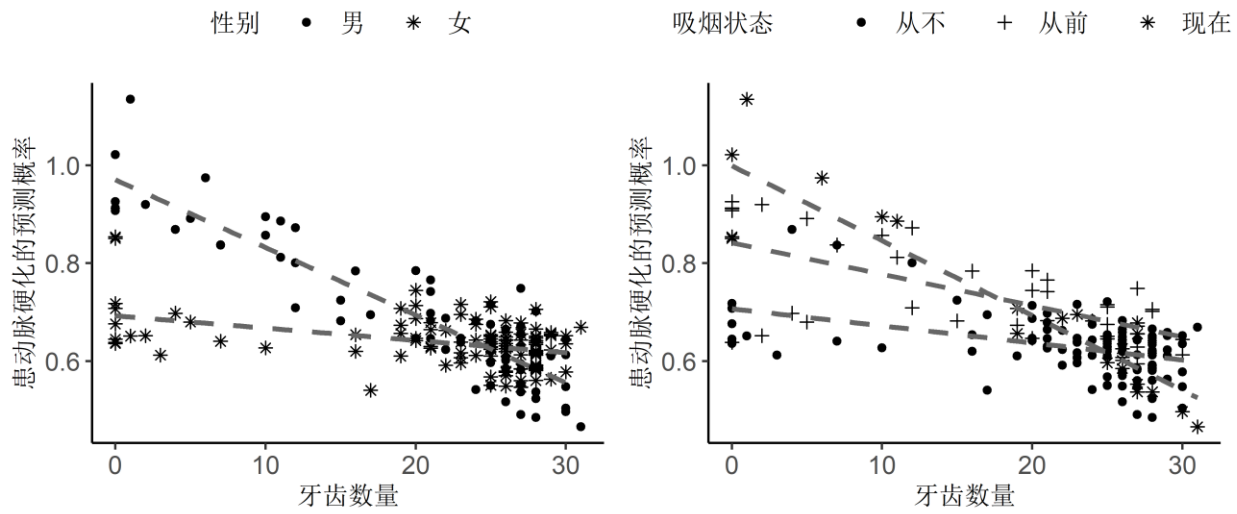

图 3-9 性别和吸烟状态分别作用下牙齿数量和患动脉硬化概率的关系

最终推荐以牙齿咀嚼能力作为衡量牙齿健康状况的主导因素，同时考虑性别，年龄，吸烟状态，身体质量指数以及它们和咀嚼能力的交互构成的 IMT 模型。

3.2.4 模型的进一步讨论

在 3.2.1 节，我们具体阐述了以 IMT 水平为被解释变量，假定逆高斯分布，采用相等连接的 GEE 估计模型，在 3.2.3 节中，我们对这个模型的系数进行了具体的解读。下面我们对这个模型进行进一步的讨论。在 GEE 模型中，我们将工作相关矩阵设置为了可交互结构。下面我们对工作相关矩阵进行讨论，由于数据结构是双胞胎数据，我们仅讨论工作相关矩阵是可交换结构和独立结构（即普通的 GLM 模型）。

关于回归系数和标准误差的比较结果如表 3-8 所示，在两个模型中解释变量咀嚼能力得分和其与年龄，性别的交互这三个变量都是显著的，且可以看出两个模型估计的系数的标准差区别并不大，系数的估计值的差距也没有超过一个标准差的范围。虽然采用独立结构的 GEE 模型的 QIC 值更低，伪 R 方更高，但是在实际应用中我们并不能否认双胞胎数据一定不存在一个双胞胎内部个体的相关性，当然这样的相关性可能比较弱（例如本数据集中模型拟合的相关性为 0.48），使得在拟合效果上可以当成独立的个体来处理。

表 3-8 分别采用独立和可交换的工作相关矩阵的 GEE 估计结果对比

解释变量		独立结构		可交换相关结构	
		系数估计值.	标准误差	系数估计值.	标准误差
（截距项）		1.127	0.390	0.938	0.426
性别	男性				
	女性	-0.203	0.117	-0.158	0.122
年龄		-0.005	0.004	-0.002	0.004

吸烟状态	从不吸烟				
	从前吸烟	0.058	0.033	0.049	0.031
	现在吸烟	0.091	0.054	0.047	0.042
身体质量指数		-0.001	0.005	-0.002	0.006
牙齿咀嚼能力得分		-0.215	0.053	-0.167	0.049
年龄×咀嚼能力得分		0.003	0.001	0.002	0.001
性别×咀嚼能力得分		0.042	0.018	0.032	0.054
QIC		-2252.48		-2106.48	
伪 R^2		0.338		0.314	
调整的伪 R^2		0.306		0.281	

另外一个值得讨论的是我们所建立的模型对数据的拟合效果，具体说来就是模型拟合的个体 IMT 值和其实际值的差距。IMT 的实际值和拟合值的可视化如图 3-10 所示，由图知我们的预测模型效果并不完美，图中蓝线上分布的点为预测较为准确的样本点，而我们的样本点的分布较为分散。由此可见，实际上牙齿健康状况可能只是作为一个影响动脉硬化的因素，但是这种影响性不是绝对的，动脉硬化还受到很多其他因素的影响。

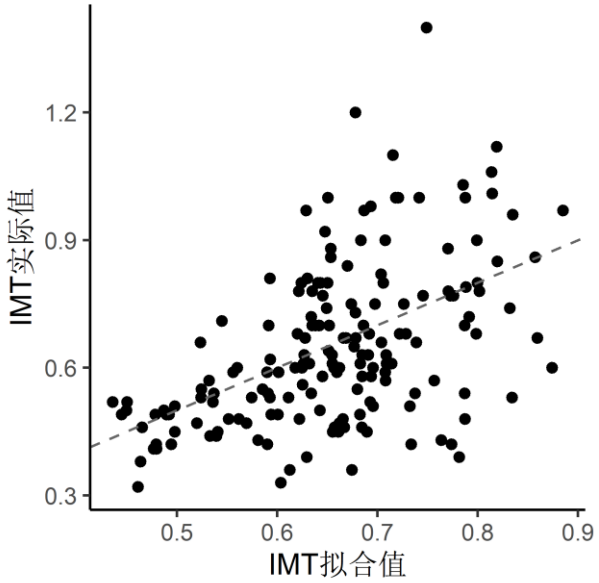

图 3-10 IMT 拟合值和实际值对比

4 基因环境因素与牙齿健康状况和动脉硬化的因果关系探究

本文的第二部分侧重研究基因在动脉硬化与牙齿的健康状态中扮演的角色。有文献指出基因对牙齿健康状态和动脉硬化有一定影响，这样的影响会混淆我们对于动脉硬化和牙齿健康状况之间相关性的判断，正如图 1-1 中所指出的那样，我们无法判断基因在这二者的相关性之间是否是一个调节变量。在第一部分中，我们利用基于广义估计方程的广义线性模型详细阐述了患动脉硬化风险和牙齿健康状况之间的相关关系，这一部分我们将利用双胞胎基因环境模型探究基因和环境因素对动脉硬化和牙齿健康状况的因果联系。

这一节中首先我们先利用单变量双胞胎 ACE 模型探究 IMT 水平和三种反应牙齿健康状况的指标（包括牙齿数量，咀嚼能力得分和牙周袋平均深度）中分别被基因和环境因素解释的比例。其次我们把 IMT 水平和三种反映牙齿健康状态的量联系起来，利用二元双胞胎 ACE 模型更加深入地解释了基因在 IMT 水平和牙齿健康状况的相关性中参与解释的部分。我们利用似然比检验的方法说明了模型拟合的优度，同时检验了基因因素的显著性情况。

4.1 考虑年龄差异的基因环境模型

在双胞胎 ACE 模型中，受制于数据结构的特点我们在建立模型时需要格外关注年龄和性别这样的调节变量对于我们探究基因环境影响的干扰，一般来说，我们希望在同一性别，年龄近似的双胞胎数据中建立基因环境模型[29]。由于本数据集的特点，样本的年龄分布较为分散，当我们把性别固定之后如果再考虑只选用一定年龄段的数据时，就会出现样本量太小的问题，降低模型的精度。因此本文在原有的 ACE 模型的基础上参考文献[29]中 Neale 的方法提出了加入性别这一调节因子的 ACE-age 模型，将年龄也作为解释样本数据中 IMT 水平或者牙齿健康状态有差异的原因之一，即有回归公式 $P=\beta_1 A+\beta_2 C+\beta_3 E+\beta_4 age$ 。相应的，我们的特征变量的方差就被分解为 $\Sigma_p=\Sigma_A+\Sigma_C+\Sigma_E+\Sigma_{age}$。在引入年龄因素后，我们使得原有的 ACE 模型更好地适应不同年龄的数据集。单变量 ACE-age 模型的路径图在 4.2.1 节中以具体特征变量的形式给出了。其中，我们假设年龄对牙齿健康状况变量或者 IMT 水平的影响有路径系数 s，则我们给出特征变量的模型拟合协方差矩阵结构：

$$Cov_{MZ}=\begin{pmatrix} a^2+c^2+e^2+s^2 & a^2+c^2+s^2 \\ a^2+c^2+s^2 & a^2+c^2+e^2+s^2 \end{pmatrix} Cov_{DZ}=\begin{pmatrix} a^2+c^2+e^2+s^2 & 0.5a^2+c^2+s^2 \\ 0.5a^2+c^2+s^2 & a^2+c^2+e^2+s^2 \end{pmatrix}$$

我们把 ACE-age 模型推广到了双变量基因环境模型的情形中。双变量 ACE-age 模型的路径图在 4.3.1 节中的图 4-2 给出，我们设年龄对 ITM 水平的影响为路径系数 s_1，对牙齿健康状况的影响为路径系数 s_2。为保证协方差均值的正定性，我们将含年龄影响的 ACE 模型下的同卵双胞胎和异卵双胞胎的，以 IMT 水平和牙齿健康状况作为两个特征变量的协方差矩阵设计为如下结构：

$$Cov_{MZ}=\begin{pmatrix} A+C+E+G & A+C+G \\ A+C+G & A+C+E+G \end{pmatrix} Cov_{DZ}=\begin{pmatrix} A+C+E+G & 0.5A+C+G \\ 0.5A+C+G & A+C+E+G \end{pmatrix}$$

其中表示年龄解释的方差部分为 Σ_{age}，简记为 G，我们约定 $G=ss^T, s=\begin{pmatrix} s_1 & 0 \\ 0 & s_2 \end{pmatrix}$。

4.2 基因环境分别对牙齿健康状况和动脉硬化差异的影响

4.2.1 单变量双胞胎基因环境模型的建立和选择

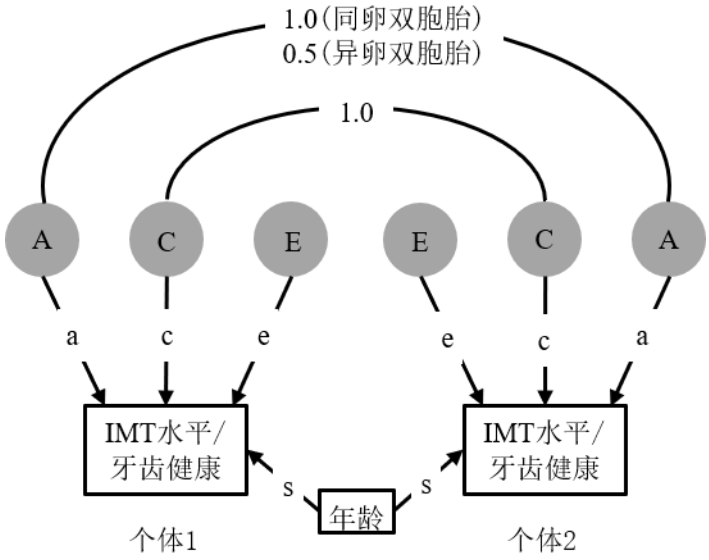

图 4-1 含年龄影响的单变量 ACE-age 模型路径图

我们建立的单变量 ACE 模型结构已在 2.2.2 节给出。单变量 ACE-age 模型如图 4-1 所示。我们建立了不同性别下以 IMT 水平和牙齿健康状况分别作为特征变量的 ACE 模型和 ACE-age 模型。

在模型选择上，我们首先检验了建立的 ACE 模型和 ACE-age 模型是否具有统计意义。双胞胎基因环境模型可以通过卡方拟合优度检验进行，我们将现有模型同饱和模型（即完美拟合的模型）比较，利用似然比检验。其中，完美拟合的模型是指所有的协方差都被当作自由参数进行估计，这种情形下，它们的极大似然估计实际上就是样本协方差，

那么一个不显著的卡方值（即 P>0.05）就意味着模型和数据反映的是一致的，而一个显著的卡方值则表示对数据的拟合效果很差。在似然比检验下，我们所有的单变量 ACE 和 ACE-age 模型相较于它们的饱和模型都是不显著的，这意味着我们的模型具有一定的可信度。由于模型较多，我们仅展示了部分的假设检验结果如表 4-1。

表 4-1 单变量 ACE 模型和 ACE-age 模型同饱和模型的似然比检验

特征变量	比较模型	男性			女性		
		-2×似然函数值	自由度	P 值	-2×似然函数值	自由度	P 值
IMT 水平	饱和模型	241.0	80	1.000	396.9	190	0.875
	ACE 模型	241.2	86		399.4	196	
	饱和模型（含年龄）	353.7	117	0.987	612.6	282	0.180
	ACE-age 模型	357.4	129		628.9	294	
牙齿数量	饱和模型	220.3	80	0.846	329.6	117	0.972
	ACE 模型	223.0	86		334.1	129	
	饱和模型（含年龄）	505.5	190	0.999	744.7	282	0.987
	ACE-age 模型	505.9	196		748.5	294	

当然，在假设检验中，不显著的 P 值只能让我们不拒绝原假设，而并不能接受原假设（即肯定地认为模型拟合与数据集之间具有一致性）。还有研究指出这样的卡方检验会显著地受到样本量的影响。由于我们的 ACE 模型是对原有数据集的特征变量的协方差矩阵的拟合，从直观上看我们可以直接比较模型拟合的协方差矩阵和数据集直接计算的协方差矩阵的差异来评判拟合的效果。当然，定量地，我们利用 Bollen 提出的函数来度量两个协方差矩阵之间的距离[30-31]：

$$D(\Sigma, S) = \ln|\Sigma| + tr(S\Sigma^{-1}) - \ln|S| - p \tag{4.1}$$

其中 Σ 表示模型估计的特征变量的协方差矩阵，S 表示数据集直接得到的特征变量协方差矩阵，p 表示特征变量的个数的两倍。

表 4-2 给出了反映两个模型在同卵双胞胎数据中的拟合效果的这种距离，从表中可见，实际上两个模型的效果都是比较好的，二者计算的基于模型的特征变量的协方差矩阵和数据集直接计算的协方差矩阵的差异都不明显。由于异卵双胞胎的数据较少，不具有代表性，这里对异卵双胞胎数据的拟合效果不做评价。在数据集拟合效果都比较一致的情况下，由于 ACE-age 模型引入的年龄的解释因素，能更加丰富地解释特征变量的协方差，我们选择 ACE-age 模型。

表 4-2 单变量 ACE 模型和 ACE-age 模型的拟合效果度量

特征变量	性别	ACE 模型	ACE-age 模型

IMT 水平	男性	0.00268	0.00268
	女性	0.00174	0.00174
牙齿数量	男性	0.01368	0.01368
	女性	0.00027	0.00035
咀嚼能力得分	男性	0.00413	0.00413
	女性	0.00037	0.00035
牙周袋平均深度	男性	0.00322	0.00322
	女性	0.00034	0.00037

4.2.2 模型结果分析

表 4-3 单变量 ACE-age 模型路径系数估计

解释变量	男性				女性			
	基因因素(A)	双胞胎共同环境因素(C)	双胞胎个体差异环境因素(E)	年龄因素(age)	基因因素(A)	双胞胎共同环境因素(C)	双胞胎个体差异环境因素(E)	年龄因素(age)
IMT 水平	0.00 (-1.57,1.57)	0.55 (0.31,0.78)	0.73 (0.58,0.88)	0.36 (0.13,0.59)	0.00 (-0.41,0.41)	0.71 (0.60,0.82)	0.32 (0.27,0.36)	0.62 (0.45,0.78)
牙齿数量	0.00 (-0.68,0.68)	0.66 (0.47,0.85)	0.55 (0.44,0.66)	-0.47 (-0.71,-0.23)	0.63 (0.50,0.76)	0.00 (-1.29,1.29)	0.57 (0.49,0.65)	-0.51 (-0.67,-0.35)
咀嚼能力得分	0.00 (-1.47,1.47)	0.66 (0.43,0.89)	0.68 (0.54,0.82)	-0.24 (-0.49,0.01)	0.70 (0.56,0.84)	0.00 (-2.56,2.56)	0.59 (0.50,0.68)	-0.37 (-0.54,-0.20)
牙周袋平均深度	0.00 (-1.82,1.82)	0.63 (0.40,0.86)	0.66 (0.86,0.51)	0.35 (0.09,0.61)	0.67 (0.52,0.82)	0.00 (-1.40,1.40)	0.64 (0.55,0.73)	0.35 (0.18,0.52)

表 4-3 给出了考虑年龄变化影响的单变量 ACE-age 模型的路径系数估计及 95%的置信区间。特别注意的是，由于之前我们已经将数据标准化处理，我们得到的路径系数的平方大致就是基因或者环境影响导致的特征变量变化的比例。

表 4-4 单变量 ACE-age 模型对特征变量方差的基因环境因素分解

特征变量	性别	基因因素(A)	双胞胎共同环境因素(C)	双胞胎个体环境因素(E)	年龄因素(age)
IMT 水平	男性	0.00	0.30	0.53	0.13
	女性	0.00	0.50	0.10	0.38
牙齿数量	男性	0.00	0.43	0.30	0.22
	女性	0.39	0.00	0.32	0.26
咀嚼能力得分	男性	0.00	0.44	0.46	0.06
	女性	0.49	0.00	0.35	0.14
牙周袋平均深度	男性	0.00	0.40	0.43	0.12
	女性	0.45	0.00	0.41	0.12

从表 4-4 中看出，在 IMT 水平中，不论男女都没有表现出基因有明显的影响，而在

牙齿健康状况中男女出现了分化。男性中在牙齿数量，咀嚼能力得分和牙周袋平均深度上基因都没有显著影响，而在女性中基因对这三者的影响是显著不为 0 的，分别占到了 39%，49%和 45%，这说明基因会导致女性的牙齿健康状况的变化。另外，我们还注意到年龄因素对牙齿健康状况和 IMT 水平的影响几乎都是显著的，这正说明了我们建立的引入年龄的 ACE-age 模型的合理性。另外，除了观察系数的显著性水平，我们还可以通过似然比检验的方法对基因的影响进行检验，具体的做法是在 ACE 模型中，我们固定基因的影响为 0，这样拟合出来的模型我们称为 CE 模型，即假设双胞胎的特征变量的差异仅由环境因素导致。对比这两个模型，如果得出的结果是显著的，即表明 ACE 模型和 CE 模型是明显有差异的，那么就说明基因的影响是显著的，我们检验了上述所有的模型，具体结果如表 4-5。我们发现只在女性的牙齿数量上发现的基因的影响的几乎显著性，这进一步说明可能女性的牙齿数量的差异是受到基因一定程度的控制的。

表 4-5 通过 CE-age 模型和 ACE-age 模型的似然比检验基因影响的显著性

特征变量	性别	CE-age 模型		ACE-age 模型		P 值
		-2×对数似然函数值	自由度	-2×对数似然函数值	自由度	
IMT 水平	男性	357.4	129	357.4	130	1.000
	女性	628.9	29 4	628.9	295	1.000
牙齿数量	男性	334.1	129	334.1	130	1.000
	女性	748.5	294	750.9	295	0.119*
咀嚼能力得分	男性	343.1	123	343.1	124	1.000
	女性	744.2	282	745.9	283	0.188
牙周袋平均深度	男性	297.2	108	297.2	109	1.000
	女性	740.8	276	742.7	277	0.167

4.3 基因环境对牙齿健康状况和动脉硬化相关性的贡献

4.3.1 双变量双胞胎基因环境模型的建立和选择

类似的，我们通过似然比检验了二变量下的 ACE 模型和 ACE-age 模型的显著性，结果均说明我们的模型有一定的可信度。另外，我们也通过式 4.1 给出的模型估计的协方差矩阵和数据的协方差矩阵的度量函数计算了二者的差别。类型于单变量的情形，从拟合效果上看 ACE-age 模型和 ACE 模型没有明显的不同（如表 4-6），但前者的信息量显然更大。下面我们以 IMT 水平和牙齿数量的 ACE-age 模型为例，进行模型系数的解读。

表 4-6 双变量 ACE 模型和 ACE-age 模型拟合效果度量

特征变量	性别	ACE 模型	ACE-age 模型

牙齿数量	男性	2.028441	2.028441
	女性	2.125329	2.125842
咀嚼能力得分	男性	2.068706	2.068598
	女性	2.032815	2.032678
牙周袋平均深度	男性	2.090670	2.091306
	女性	2.011256	2.011872

4.3.2 模型结果分析

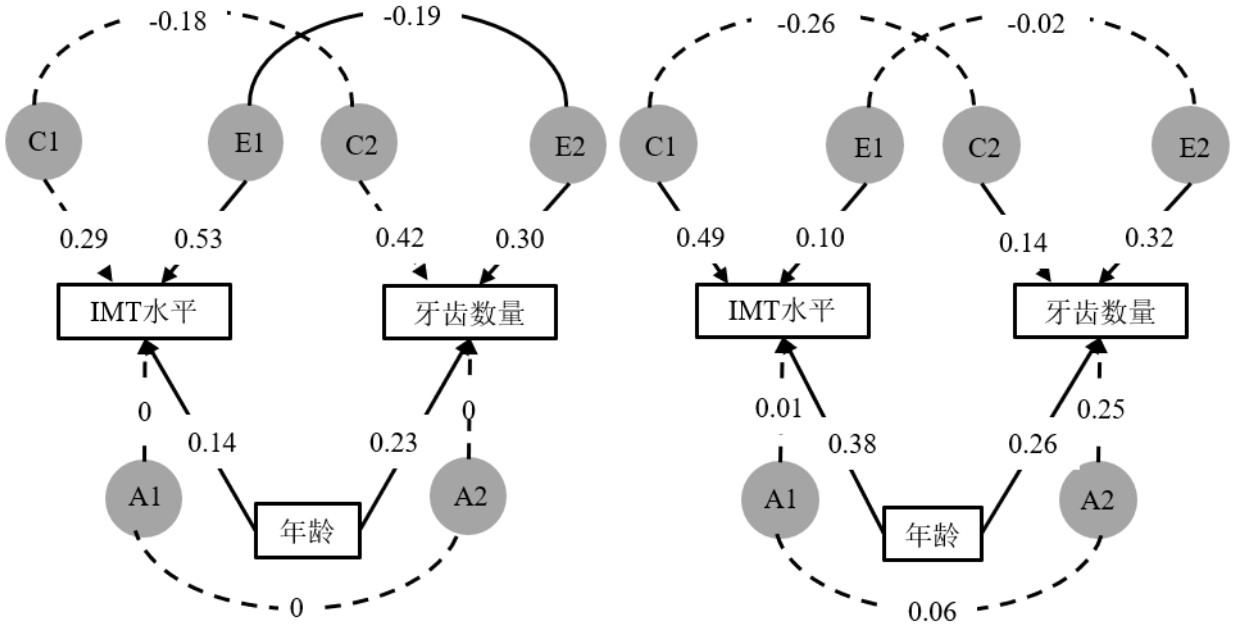

图 4-2 双变量 ACE-age 模型路径图（左边为男性，右边为女性）

我们以 IMT 水平和牙齿数量的关系为例，展示以男性和女性分别建立的 ACE-age 模型的路径图如图 4-2。从图中可见，图中的路径上的数字表示当前因素解释中心特征变量变化的百分比，而路径的虚实表示这一解释因素是否显著。从女性的模型可以看出虽然基因因素可以解释 IMT 水平的 25%，但这一系数是不显著的。这一结论是对模型单变量模型中得出的女性的牙齿数量受基因影响的结论的进一步修正。实际上我们指出，通过多变量模型的因素占比拟合结果与双变量模型有差距。这是因为多变量模型中我们还考虑了年龄的影响和跨双胞胎跨特征变量的基因因素的相关性和环境因素的相关性[27]。从男性的模型可以看出基因因素并不显著，这与我们之前通过单变量模型得出的结论是一致的。另外，我们发现在女性中年龄，双胞胎相同的环境和各异的环境因素对 IMT 水平和牙齿数量的影响显著。而对于男性而言我们发现牙齿数量和 IMT 水平主要受双胞胎各异的环境影响。在表 4-7，表 4-8 中，我们给出了基因和环境分别分解得到的关于特征变量牙齿数量和 IMT 水平的协方差矩阵和里面的元素的 95%置信区间。

表 4-7 男性 IMT 水平和牙齿数量协方差的基因环境因素分解

特征变量	基因因素		共同环境因素		个体环境因素	
	IMT 水平	牙齿数量	IMT 水平	牙齿数量	IMT 水平	牙齿数量
IMT 水平	0.00	0.00	0.29	-0.18	0.53*	-0.19
	(0.00,0.58)	(-0.39,0.14)	(0.00,0.63)	(-0.49,0.18)	(0.35,0.82)	(-0.36,-0.08)
牙齿数量	0.00	0.00	-0.18	0.42	-0.19*	0.30*
	(-0.39,0.14)	(0.00,0.47)	(-0.49,0.18)	(0.00,0.76)	(-0.36,-0.08)	(0.21,0.47)

表 4-8 女性 IMT 水平和牙齿数量协方差的基因环境因素分解

特征变量	基因因素		共同环境因素		个体环境因素	
	IMT 水平	牙齿数量	IMT 水平	牙齿数量	IMT 水平	牙齿数量
IMT 水平	0.01	0.06	0.49*	-0.26*	0.10*	-0.02
	(0.00,0.19)	(-0.12,0.19)	(0.28,0.71)	(-0.44,-0.05)	(0.08,0.13)	(-0.06,0.02)
牙齿数量	0.06	0.25	-0.26*	0.14	-0.02	0.32*
	(-0.12,0.19)	(0.00,0.52)	(-0.44,-0.05)	(0.00,0.46)	(-0.06,0.02)	(0.25,0.44)

关于年龄对 IMT 水平和牙齿数量的影响，模型拟合得到 IMT 水平和年龄呈正相关，在男性中路径系数为 0.37，女性中为 0.62。另外得到牙齿数量和年龄呈负相关，在男性中路径系数为-0.48，在女性中为-0.51。除此以外，我们特别指出我们这里对于年龄因素对特征变量的影响的假定是较为简单的，我们设计了两个路径系数分别反映年龄对 IMT 水平和牙齿健康状况的影响，这允许年龄对这两个因素的影响是可以存在不同的。但是我们没有控制年龄会影响 IMT 水平和牙齿健康数量二者的相关性，即我们没有考虑年龄在跨特征变量相关性中的影响。

实际上我们对三种牙齿健康状态的变量和 IMT 水平分别组成了 ACE-age 模型，出于篇幅考虑这里不对每一个模型的具体结果进行阐述，图 4-3 简要展示了不同性别中基因，双胞胎相同环境和各异环境因素解释特征变量协方差的情况。总的来说，在男性群体中，除了在牙周袋深度与 IMT 的相关性中发现了基因的贡献较明显外，基因贡献变量的方差或协方差的比例很小。即使前者的贡献是较明显的，其统计的假设检验也是不显著的。而在女性群体中，在牙齿健康状况中发现了基因有显著的贡献，当然这样的贡献在统计的假设检验中目前没有发现存在显著性。另外，两个群体中都说明，牙齿健康状况和 IMT 之间的相关性有一部分来自两者分享的环境因素，侧面说明了双胞胎数据中内部相关的特性。

综上，虽然我们借助双变量 ACE-age 模型在男性和女性群体中都没有发现基因在调节动脉硬化和牙齿健康状况的相关性中具有统计意义上的显著性，但这个模型为研究图 1-1 中三者的复杂关系提供了很好的渠道。我们期待更丰富的数据集来继续探究这一问题。

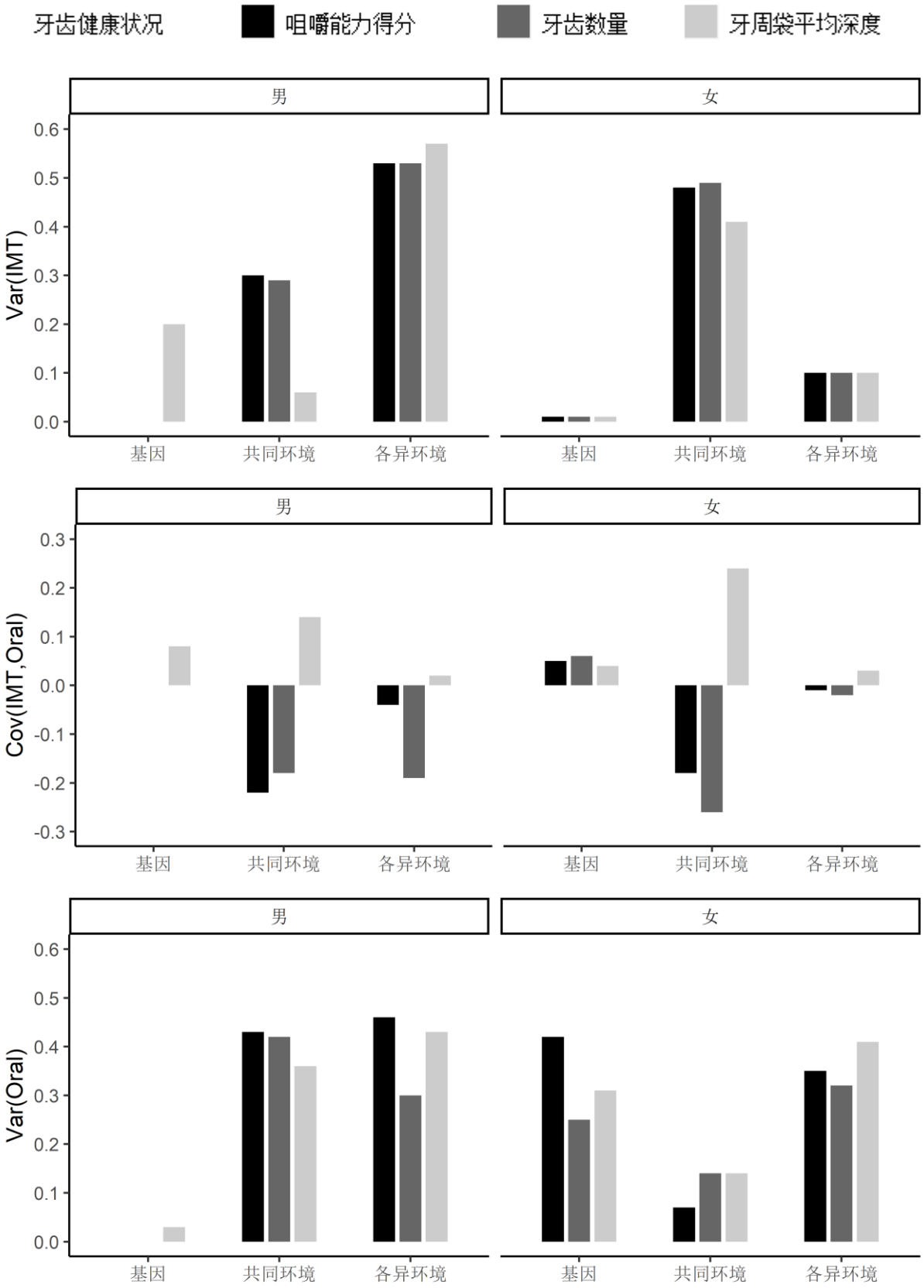

图 4-3 IMT 值和牙齿健康状况的方差和协方差结构的基因环境因素分解

结 论

本文对牙齿健康状况和动脉硬化之间的关系进行了研究，利用 GEE 估计方法，通过 IMT 模型和 IMT01 模型，论述了二者相关性的显著性，借助于牙齿数量，咀嚼能力得分二个反映牙齿健康状况的指标，发现了在不同年龄段，不同性别，不同吸烟状态的群体中，牙齿健康状况和动脉硬化间的相关性呈现强弱差异。例如，吸烟群体，高龄和男性都会显著增加患动脉硬化的风险。相较于女性，咀嚼能力的降低使男性患动脉硬化有更大的风险。且在 50-60 岁人群中的男性比 80 岁以上的男性，咀嚼能力和动脉硬化风险间的负相关性更强。同时，本文利用双胞胎基因环境模型（ACE 模型和 ACE-age 模型），探究了基因和环境因素在群体中动脉硬化，牙齿健康状况的差异上，以及动脉硬化与牙齿健康状况的相关性中扮演的角色，发现基因因素对女性群体的牙齿健康状况的差异可能有一定影响，而男性群体的牙齿健康状况主要受环境影响。

当然，本文受制于有限的样本数据集，暂未发现基因与动脉硬化间显著的因果联系，也没有发现基因对动脉硬化和牙齿健康状况之间的相关性有显著的贡献。此外，本文侧重于探究人们的牙齿健康状况和患动脉硬化的风险间的相关性。而许多文献指出，通过回归方法建立的模型可以用于探索变量间的相关性，但在因果关系上却难以具有说服力。我们期待有更加深刻的因果推断来研究牙齿健康状况和动脉硬化之间的因果关系。

虽然在动脉硬化众多的影响因素里，牙周可能只是其中一个因素，然而研究牙周状况却对实际中及时判断和预测动脉硬化情况提供了一个可能的途径，从而为人们提前防范动脉硬化提供了科学的依据。在这一方面，未来我们还需要借助更多的牙周炎和动脉硬化患者的临床数据以及有效的模型进行进一步的研究。

致 谢

从 2019 年年末毕业论文选题工作启动至今，经过了 6，7 个月左右。从选题策划到论文后续的方向进展的过程中，我首先要向本文的指导老师赵慧秀老师表示衷心的感谢。从去年启动国外大学申请工作后，我萌生了申请生物统计方向的硕士项目的想法，在与赵老师的交流中确定了以动脉硬化相关数据为方向的题目。在老师的引导下，我开始接触生物医学数据的分析处理方法，在这一过程中，我逐渐意识到生物统计研究的实际价值，对这一方向有了更深入的了解和更浓厚的兴趣。感谢赵老师在论文写作过程中给出的宝贵意见和对我悉心的鼓励！

岁月不居，时节如流，大学四年时光匆匆。值此毕业之际，感谢南京理工大学各位专业老师四年时间对我的培养，感谢陈培鑫，张正军等理学院老师对我科研训练，出国申请等工作的大力支持与帮助。正是各位老师将我引入统计学的大门，让我逐步对数据分析产生兴趣，也为我奠定了理解数据，研究数据所需要的扎实的数学基础。同时也感谢陪伴我四年的舍友们，同学们，那些与你们的朝夕相处，共同探讨学术话题，参与活动竞赛的过程构成了我大学四年美好难忘的回忆。总而言之，向在四年时间中帮助过我的前辈们，老师们，同学们，朋友们表示诚挚的谢意！

自 2020 年年初新冠疫情爆发以来，全球公共卫生秩序面临极大挑战。这一挑战更坚定了我攻读生物统计方向硕士学位的决心。感谢布朗大学的 Stavroula Chrysanthopoulou 教授就毕业论文内容和生物统计学前景与我进行的友好交流。希望在未来生物统计领域学习和研究的过程中，我能为公共卫生领域贡献自己的一份力量。

当然，我还要向我的父母，家人们表示特别的感谢，正是他们给我提供的资金和精神支持帮助我攻读学士学位，也让我平稳地度过这大学四年。同时也非常感谢我的父亲继续支持我申请国外大学，攻读硕士学位。

最后，向参加毕业论文评阅工作的各位老师和专家表示感谢！

前路漫漫，愿只争朝夕，不负韶华！

参 考 文 献

[1]. 臧伟进, 吴立玲.《心血管系统》[M], 北京：人民卫生出版社，2015.
[2]. Rossr. Atherosclerosis-an inflammatory disease[J]. N Engl J Med, 1999, 340(2): 115-126.
[3]. 刘俊田.动脉粥样硬化发病的炎症机制的研究进展[J].西安交通大学学报(医学版), 2015, 36(02): 141-152.
[4]. P.B. Lockhart, A.F. Bolger, P.N. Papapanou, O. Osinbowale, M. Trevisan, M.E. Levison, et al., Periodontal disease and atherosclerotic vascular disease: does the evidence support an independent association? a scientific statement from the American Heart Association[J], Circulation 125 (2012) 2520–2544.
[5]. Papapanou, Panos, N,Trevisan, Maurizio. Periodontitis and atherosclerotic vascular disease: what we know and why it is important[J]. Journal of the American Dental Association (1939), 2012, 143(8).
[6]. Gheorghita, Dorottya, Eördegh, Gabriella,Nagy, Ferenc, Antal, Márk. [Periodontal disease, a risk factor for atherosclerotic cardiovascular disease][J]. Orvosi hetilap, 2019, 160(11).
[7]. Thomas T. Nguyen, Kevin Y. Wu, Maude Leclerc, Hieu M. Pham, Simon D. Tran. Cardiovascular diseases and periodontal disease[J]. Current Oral Health Reports, 2018, 5(1) 13-18
[8]. Kenji Wakai, Mariko Naito, Toru Naito, Masaaki Kojima, Haruo Nakagaki, Osami Umemura, Makoto Yokota, Nobuhiro Hanada, Takashi Kawamura. Tooth loss and intakes of nutrients and foods: a nationwide survey of Japanese dentists[J]. Community Dentistry and Oral Epidemiology, 2010, 38(1).
[9]. Asai. K, Yamori. M, Yamazaki. T, Yamaguchi. A, Takahashi. K, Sekine. A, Kosugi. S, Matsuda. F, Nakayama. T, Bessho. K. Tooth loss and atherosclerosis: the Nagahama Study.[J]. Journal of Dental Research, 2015, 94(3 Suppl).
[10]. Xiao-Tao Zeng, Wei-Dong Leng, Yat-Yin Lam, Bryan P. Yan, Xue-Mei Wei, Hong Weng, Joey S.W.Kwong. Periodontal disease and carotid atherosclerosis: A meta-analysis of 17330 participants[J]. International Journal of Cardiology, 2016, 203 1044-1051.
[11]. Kurushima, Yuko, Ikebe Kazunori, Matsuda Ken-Ichi, Enoki Kaori, Ogata Soshiro, Yamashita Motozo, Murakami Shinya, Maeda Yoshinobu. Examination of the Relationship between Oral Health and Arterial Sclerosis without Genetic Confounding through the Study of Older Japanese Twins[J]. PloS one, 2015, 10(5).
[12]. John B Carlin, Lyle C Gurrin, Jonathan AC Sterne, Ruth Morley, Terry Dwyer. Regression models for twin studies: a critical review[J]. International Journal of Epidemiology, 2005, 34(5) 1089-1099.
[13]. 韩亚琨,林晓萍.相同基因条件下牙周治疗与动脉硬化的相关性[J].上海口腔医

学,2017,26(02):209-212.

[14]. C. Song, Z. Chang, P. K. E. Magnusson, E. Ingelsson, N.L. Pedersen. Genetic factors may play a prominent role in the development of coronary heart disease dependent on important environmental factors[J]. Journal of Internal Medicine, 2013, 275(6) 631–639.

[15]. Beck J D, Elter J R, Heiss G, Couper D, Mauriello S M, Offenbacher S. Relationship of periodontal disease to carotid artery intima-media wall thickness: the atherosclerosis risk in communities (ARIC) study[J]. Arteriosclerosis, Thrombosis, and Vascular Biology, 2001, 21(11).

[16]. Hayashi Shuji, Yamada Hirotsugu, Fukui Makoto, Ito Hiro-O, Sata Masataka. Correlation Between Arteriosclerosis and Periodontal Condition Assessed by Lactoferrin and α1-Antitrypsin Levels in Gingival Crevicular Fluid[J]. International Heart Journal, 2015, 56(6).

[17]. Liang K Y, Zeger S L. Longitudinal data analysis using generalized linear models[J]. Biometrika, 1986, 73(1): 13-22.

[18]. Wang M. Generalized estimating equations in longitudinal data analysis: a review and recent developments[J]. Advances in Statistics, 2014, 2014.

[19]. Zheng B. Summarizing the goodness of fit of generalized linear models for longitudinal data[J]. Statistics in medicine, 2000, 19(10): 1265-1275.

[20]. Heinzl, H., Mittlböck, M. Adjusted R^2 Measures for the Inverse Gaussian Regression Model. Computational Statistics 17**,** 525–544 (2002).

[21]. Cui J, Qian G. Selection of working correlation structure and best model in GEE analyses of longitudinal data[J]. Communications in Statistics—Simulation and Computation®, 2007, 36(5): 987-996.

[22]. Fan Y, Chen J, Shirkey G, et al. Applications of structural equation modeling (SEM) in ecological studies: an updated review[J]. Ecological Processes, 2016, 5(1): 19.

[23]. Dahly D L, Adair L S, Bollen K A. A structural equation model of the developmental origins of blood pressure[J]. International journal of epidemiology, 2009, 38(2): 538-548.

[24]. Beran T N, Violato C. Structural equation modeling in medical research: a primer[J]. BMC research notes, 2010, 3(1): 267.

[25]. Pahlen S, Hamdi N R, Aslan A K D, et al. Age-moderation of genetic and environmental contributions to cognitive functioning in mid-and late-life for specific cognitive abilities[J]. Intelligence, 2018, 68: 70-81.

[26]. Rijsdijk F V, Sham P C. Analytic approaches to twin data using structural equation models[J]. Briefings in bioinformatics, 2002, 3(2): 119-133.

[27]. Voronin I, Ismatullina V, Zakharov I, et al. Structural equation modeling in the genetically informative study of the covariation of intelligence, working memory and planning[C]//ITM Web of Conferences. EDP Sciences, 2016, 6: 02010.

[28]. Dunn P.K., Smyth G.K. (2018) Chapter 11: Positive Continuous Data: Gamma and Inverse Gaussian GLMs. In: Generalized Linear Models With Examples in R. Springer Texts in Statistics. Springer, New York, NY.

[29]. Neale M, Cardon L R. Methodology for genetic studies of twins and families[M]. Springer Science & Business Media, 2013.

[30]. Bollen,K.A. Structural equations with latent variables. New York:Wiley, 1989.

[31]. 宋铭,刘宇洁.结构方程模型参数 ML 估计的具体案例[J].课程教育研究,2014(02):231-232.

[32]. Hardin J W, Hardin J W, Hilbe J M, et al. Generalized linear models and extensions[M]. Stata press, 2007.

[33]. Ballinger G A. Using generalized estimating equations for longitudinal data analysis[J]. Organizational research methods, 2004, 7(2): 127-150.

[34]. Halekoh U, Højsgaard S, Yan J. The R package geepack for generalized estimating equations[J]. Journal of Statistical Software, 2006, 15(2): 1-11.