2019 年（第六届）全国大学生统计建模大赛

基于 GWR 和路径分析的胎儿基因甲基化与孕妇吸烟的关系研究

参 赛 单位: 中 南 大 学

参赛者姓名: 王培锦、隋阳、宋玮琳

目 录

表格和插图清单

摘　要

目的：运用路径分析，对影响胎儿甲基化程度的直接影响因素和间接影响因素进行分析，并建立地理加权回归（GWR）模型，以排除实验误差对结果的影响，并探讨了孕妇吸烟情况及其它协变量的交互作用与胎儿甲基化程度的关系。

方法：首先，根据相关性检验结果进行路径分析，确定对胎儿甲基化产生影响的因素。随后，建立 GWR 模型对孕妇吸烟情况与胎儿甲基化程度的关系进行研究；在此基础上，对模型进行修正，引入孕妇的生育年龄、受教育程度、BMI 指标、种族以及胎儿的性别与孕妇吸烟情况的交互作用，对吸烟影响甲基化程度做更加深层次地研究。

结果：①在单因素方差分析中，只有母亲吸烟状况通过检验，说明其对甲基化水平影响显著，母亲分娩年龄、母亲 BMI 指标、母亲怀孕周数、胎儿性别、母亲受教育程度、种族等变量均未通过显著性检验。②在路径分析中，母亲受教育程度对母亲吸烟状况的总效应为-0.164，对胎儿甲基化水平的总效应为-0.004，吸烟情况对甲基化水平的总效应为 0.026。③在 GWR 模型中，回归方程的截距项和斜率项有明显的非平稳性，说明实验中的孔板及样品的位置对实验结果有比较严重的影响。④在修正的 GWR 模型中，母亲分娩年龄、母亲受教育程度、母亲 BMI 指标、种族、胎儿性别对于回归方程中的截距项的分布有较强的影响；母亲分娩年龄、母亲 BMI 指标、种族对于回归方程中的斜率项的分布有较强的影响。

结论：母亲吸烟状况对胎儿甲基化水平有显著的直接影响，而且吸烟通常会导致胎儿的甲基化程度偏大。由于受教育程度较高的女性存在晚孕倾向、BMI 指标普遍处于正常区间，这使得后代的甲基化程度越接近正常水平。年龄偏大、受教育程度较低、偏胖、非洲裔孕妇的男性胎儿更容易出现甲基化程度异常的现象。

建议：①科普孕期中吸烟对胎儿成长发育的不利影响，以此减少孕妇吸烟的人数以及吸烟次数；②孕妇应在孕产期中保持健康适中的体质，母亲 BMI 指数维持在合适的范围内，可以适当地增加体重；③建议较小年龄、偏胖、非洲裔的孕妇更加密切关注胎儿的甲基化程度，避免异常现象的出现。

关键词：甲基化 方差分析 路径分析 GWR

Abstract

OBJECTIVE: Use Path Analysis to analyze the direct and indirect factors affecting the degree of fetal methylation, and establish a Geographically Weighted Regression (GWR) Model to eliminate the influence of experimental error on the results, and explore the effect of pregnant women's smoking status and other covariates on the fetal methylation degree.

METHODS: First, Path Analysis was performed based on the results of the correlation test to determine the factors affecting fetal methylation. Subsequently, GWR Model was established to study the relationship between smoking status and fetal methylation status; on this basis, the model was modified by introducing the factors, such as maternal age, education level, BMI index, ethnicity, fetal gender and the interaction of them, to learn how smoking status has made a deeper effect on methylation.

RESULTS: ① In the one-way analysis of variance, only the F statistic of smoke passed the significance test. ② In the path analysis, the total effect of education level on smoke is -0.164, the total effect on methyl is -0.004, and the total effect of smoking on methylation level is 0.026. ③ In the GWR model, the intercept term and the slope term of the regression equation have obvious non-stationarity, indicating that the position of the orifice plate and the sample in the experiment has a serious impact on the experimental results. ④ In the modified GWR model, the variables age, education level, BMI, race, and gender have a strong influence on the distribution of the intercept term in the regression equation; the variables age, BMI, and race have strong influence on the distribution of the slope term in the regression equation.

CONCLUSIONS: Mother's smoking status has a significantly direct effect on fetal methylation levels, and smoking usually leads to a greater degree of methylation in the fetus. Because women with higher education tend to late pregnancy and BMI indicators are generally in the normal range, the degree of methylation of offspring is closer to

normal. Male fetuses, who are older, less educated and overweight, especially in African-American populations, are more prone to abnormal methylation.

RECOMMENDATIONS: ①The adverse effects of smoking during pregnancy on the growth and development of the fetus, in order to reduce the number and frequency of smoking in pregnant women; ②pregnant women should maintain a healthy and moderate condition during the maternity period, to keep BMI index in a suitable range, but appropriate increase of weight is also necessary; ③ It is recommended that pregnant women, who are young, overweight and African descent, pay more attention to the degree of methylation of the fetus and avoid abnormal phenomena.

KEYWORDS: Methylation; Analysis of Variance; Path Analysis; GWR

一、研究背景及意义

（一）研究背景

香烟的烟雾中包含超过 7000 种化学物质，其中有数百种是已知的有毒物质和至少 69 种致癌物质[1]。尽管人们已经熟知吸烟对健康的严重危害，但是吸烟仍然是世界最重要的可预测死亡原因之一。吸烟除了对吸烟者的身体产生有害的直接影响以外，吸烟还可能对发育中胎儿产生严重的间接影响。尽管吸烟是一种可控的危险行为，在美国，孕妇怀孕期间吸烟的概率约为 14%，根据疾病控制和预防中心 2012 年相关调查结果，在怀孕前 3 个月内吸烟的女性中，仅有 45%的孕妇会选择戒烟。所以，研究孕妇吸烟对胎儿甲基化产生的影响程度，对促进优生优育、降低胎儿出生后的患病概率都有着十分重要的意义。

DNA 甲基化是目前研究最深入的表观遗传调控机制之一，一般是指在 DNA 甲基转移酶的作用下，在基因组 CpG 二核苷酸的胞嘧啶 5 碳位共价键结合一个甲基基团。DNA 甲基化能引起染色质结构、DNA 构象、DNA 稳定性及 DNA 与蛋白质相互作用方式的改变，从而控制基因的表达。对于一般正常个体来说，甲基化水平应该接近 50%。基因甲基化异常容易导致单亲遗传病和肿瘤的出现，同时随着年龄的增长，基因组总体 DNA 甲基化水平逐渐降低[2]。

环境暴露因素会影响全基因组甲基化程度，也有相关研究证明有害环境污染物的暴露与受体表冠遗传状况之间存在关联关系，即环境中的污染物可以通过表冠遗传学的机制对人体的健康状况产生深远的影响[2]。基因甲基化在胎儿期就已经形成，DNA 的甲基化水平可以影响特定基因的表达，进而影响胎儿的生长发育。因此，可以认为孕妇在妊娠期吸烟将会导致胎儿甲基化异常。虽然孕期环境暴露对于基因不同的位点甲基化影响不尽相同，但依然能通过动物实验证明孕期暴露于香烟烟雾对一些特异基因甲基化会造成比较严重的影响[3]。

Markunas 等人在研究中也证实了孕期吸烟与胎儿 DNA 甲基化程度的改变有关，与此同时新发现了 FRMD4A、ATP9A、GALNT2 和 MEG3 这三个与尼古丁依赖、戒烟、胚胎发育的相关的基因，这些基因的变化还证实了直接和间接接触香烟烟雾会引起不同的表观遗传反应，这归因于受感染者年龄和对吸烟易感性的差异[1]。

（二）研究意义

虽然香烟对于胎儿甲基化程度的影响已经被诸多学者证实，但是胎儿甲基化异常应该是由许多影响因素共同作用的结果，这些影响因素除孕期的吸烟情况以外，还有分娩年龄、孕前身体状况、怀孕周数等等。根据相关研究，许多证据都

表明，胎儿在产前阶段处于一段非常敏感的时期，孕妇在妊娠期受到的不良影响将使得后代更容易感染一些潜伏期较长的疾病，例如 II 型糖尿病、心脏病、认知功能下降等等[4]。所以，尽可能地降低胎儿甲基化异常的概率，对于减少成年后患病概率和延长寿命都有着及其重要的意义。

多数学者在研究母亲孕期吸烟情况对胎儿甲基化程度的影响时，往往忽略了母亲的其它身体指标及孕期情况对胎儿甲基化程度的影响。所以本文为更加全面地分析胎儿甲基化与母亲怀孕时期的身体状况以及子宫环境的关系，选取了 608 份 DNA 样本，主要研究胎儿甲基化程度与母亲分娩年龄、身体质量指数、吸烟情况、胎儿性别、母亲的受教育程度、母亲的种族与胎儿甲基化水平之间的关系。目的在于分析影响胎儿甲基化程度的直接因素和间接因素，以及各种因素之间的交互作用对胎儿甲基化程度的影响；并为如何减少胎儿甲基化异常情况的出现提供相关建议。

二、研究思路

本文主要研究孕妇在怀孕期间的吸烟状况对胎儿甲基化程度的影响，在此基础上，还分析了孕妇的年龄、种族、受教育程度、身体状况、怀孕周数以及胎儿的性别等指标对胎儿甲基化程度的影响。共收集到 608 份孕妇及胎儿的样本数据，具体研究思路如下：

（1）由于数据存在缺失的现象，首先需要对数据进行预处理；

（2）对甲基化水平与各影响因素分别进行单因素方差分析，确定对婴儿甲基化水平有显著性影响的变量，并对变量之间的相关关系进行相关分析；

（3）基于相关关系研究结果，进行路径分析，具体刻画出变量之间的直接影响作用和间接影响作用。

（4）通过建立地理加权回归模型，消除实验过程中的样品所在仪器孔板的位置对甲基化测量结果的影响，并通过引入虚拟变量研究其它协变量与胎儿甲基化程度之间的关系。

三、数据收集与处理

（一）数据的收集

我们的数据来源于 Hoyo C.等人[5]在 2012 年关于印记胰岛素样生长因子 2（IGF 2）、血浆 IGF 2 和出生体重与脐血甲基化分数的相关性研究。实验中，该数据集来自一个多民族出生群体，旨在研究促进有关早期暴露表观遗传特征和

结果。IGF2 DMR 甲基化水平这一指标，是受试者（后代）等位基因在其 DNA 样本中甲基化的百分比的测试值。在 314 名受试者中，294 名受试者中该位点的甲基化被检测了两次，其余 20 名受试者只检测了一次，并将 608 份 DNA 样品排列在 22 个 96 孔板上。每个板有 8 行，记为“A”到“H”；12 列，记为 1 到 12；需要注意的是，在实验中每个板上只使用了一部分格子。

（二）数据的描述

数据集命名为 epigen 共有 11 个变量，各变量的具体信息如表 1 所示。

表 1. epigen 数据说明

变量名称	数据尺度	数据说明
母亲分娩年龄 *age*	定序尺度	0-分娩时小于 30 岁、1-分娩时为 30 至 39 岁、2-分娩时大于 40 岁
母亲身体质量指数 *BMI*	定序尺度	0-30 以下、1-大于或等于 30
母亲吸烟状况 *smoke*	定类尺度	0-不吸烟、1-怀孕早期吸烟
怀孕周数 *gestage*	定序尺度	1-小于 37 周、0-大于等于 37 周
胎儿的性别 *gender*	定类尺度	1-男、0-女
母亲的受教育程度 *edu*	定序尺度	0-低于高中水平、1-高中水平、2-至少是大学水平
母亲的种族 *race*	定类尺度	0-AA 非洲裔美国人、1-EA 高加索人种、2-其它
孩子甲基化水平 *methyl*	定比尺度	将百分比数据转化为 0～1 之间小数
孔板 *plate*	定类尺度	编号为 1～22 个孔板
孔板的行 *row*	定类尺度	每个孔板有 7 行，记为 A～H
孔板的列 *column*	定类尺度	每个孔板有 12 列，记为 1～12

进一步对母亲分娩年龄（*age*）、母亲身体质量指数（*BMI*）、母亲吸烟状况（*smoke*）、孕产期（*gestage*）、胎儿性别（*gender*）、母亲受教育程度（*edu*）、母亲种族（*race*）进行频数分析，频数表如表 2 所示。

表 2. 各定性变量的频数表

变量	频数	变量	频数
age		*edu*	
0(<30)	330	0(<高中)	225
1(30~39)	256	1(高中)	159
2(>40)	22	2(>大学)	224
BMI		*smoke*	
0(<30)	432	0(否)	487
1(≥30)	176	1(是)	121
gestage		*gender*	
0(<37)	533	0(女)	279

1(≥37)	75	1(男)	329
race			
0(非洲裔)	296		
1(高加索)	263		
2(其他)	49		

根据对数据集进行描述性统计分析，本文发现在数据集中，超过 50%的测试者的年龄少于 30 岁，其次是 30～39 岁，只有 22 位测试者的年龄大于 40 岁。对于大部分测试者来说，她们的孕前 BMI 指标处于正常水平，仅有 30%左右的测试者出现肥胖情况。绝大多数的测试者怀孕周数正常，仅有 75 位孕妇怀孕超过 37 周。测试者中大多为非洲裔与高加索裔，仅有 49 位来自其他种族的孕妇。测试者的受教育程度没有明显的集中现象，即均匀分布在各个受教育层次。测试者中有 121 位有在孕期吸烟的经历，超过 80%的孕妇不会选择在妊娠期吸烟。测试者所孕育的胎儿性别男女比例接近 1：1。

（三）数据的预处理

1. 缺失值处理

在 608 份测试样本中，有 20 个样本存在某些指标数据缺失的现象，缺失数据主要集中在指标 *BMI* 和 *gender* 中。因此，为了尽量保留更多真实合理的样本，采取神经网络算法对缺失值进行插补。

利用神经网络进行缺失值插补的基本思想是：利用完整的数据集来训练网络，把除待插补的变量以外的其它变量的数据作为神经网络模型的输入，将待插补变量的对应数据作为输出。在训练结束后，把缺失数据集除待插补的变量以外变量的数据作为模型的输入数据，得到结果即为数据集的预测值[6]。神经网络算法有传统方法所不具备的优点，即能够对数据的非线性关系进行映射。而且，对被建模对象的经验知识要求不高，可以通过网络本身的学习功能得到网络输入与输出的关系，适合应用于本数据集的插补。

2. logit 变换

根据基因甲基化的概念可知，甲基化水平数据应该位于[0,1]之间。如果用该指标直接进行建模，无法保证所有预测值位于该区间内，对超出范围的数据将无法进行实际意义的解释。因此，需要对甲基化指标进行 logit 变换：

$$y = \ln\left(\frac{methyl}{1-methyl}\right) \tag{1}$$

显然，通过变换得到的 y 的取值范围被扩展以 0 为对称中心的整个实数域，这使得在任何自变量取值下，模型的预测值均有实际意义，为模型结果的解释提供了便利。

（四）小结

本章主要介绍了本文的数据收集、处理过程，并对数据集进行描述性统计分析。在数据处理方面，主要对数据的缺失值进行填补，并且对其中取值范围为[0,1]的变量 *methyl* 进行 logit 变换。在描述性统计方面，本文主要从频数角度对各个指标的分布进行研究。

四、变量相关关系分析

（一）单因素方差分析

单因素方差分析是通过对数据变异情况的分析，来判断各组总体均值是否有差别的一种统计方法。该方法可以用于检验同一个影响因素的不同水平对因变量是否有显著影响，其数学定义为：

$$\mathrm{F}=\frac{\frac{SSA}{k-1}}{\frac{SSE}{n-k}}$$

其中，n 为样本总数，k 为样本组数，*SSA* 代表组间离差平方和，*SSE* 代表组内离差平方和，F 值应服从$F(k-1,n-k)$分布。当计算出的统计量 F 的值大于显著水平 0.05 时的临界值时，则可以认为该影响因素的不同水平对因变量有显著影响。反之，则影响不显著。

为了研究数据集中的各个指标对甲基化水平是否有显著影响，本文将 *methyl* 设为因变量，其他指标分别设为自变量，进行单因素方差分析，计算结果如表 3 所示。

表 3. 影响婴儿基因甲基化水平指标的单因素方差分析结果

指标	个案数	甲基化水平	*F*	*P*
age			0.653	0.521
0(<30)	330	0.476±0.007		
1(30~39)	256	0.481±0.008		
2(>40)	22	0.467±0.026		
BMI			0.548	0.460
0(<30)	432	0.479±0.006		
1(≥30)	176	0.475±0.010		
smoke			16.58	0.000
0(否)	487	0.472±0.005		

1(是)	121	0.499±0.015		
gestage			0.804	0.370
0(<37)	533	0.477±0.006		
1(≥37)	75	0.484±0.014		
gender			0.792	0.374
0(女)	279	0.480±0.008		
1(男)	329	0.475±0.007		
edu			0.641	0.527
0(<高中)	225	0.481±0.008		
1(高中)	159	0.478±0.010		
2(>大学)	224	0.474±0.009		
race			0.537	0.585
0(非洲裔)	296	0.477±0.008		
1(高加索)	263	0.477±0.007		
2(其他)	49	0.487±0.019		

从结果中可以看出，只有母亲的吸烟情况（*smoke*）对婴儿基因甲基化水平（*methyl*）有直接的显著性影响。在怀孕期间吸烟的母亲生下的后代基因的甲基化水平的平均值偏高，且组内标准差更大，说明后代的基因甲基化水平更不稳定，可能会存在一些健康问题的隐患。其他各指标的影响虽然不显著，但是也可以通过样本均值和组内标准差的区别观察到一些细微影响，具体的影响程度和影响方式还需进一步的研究。

（二）Pearson 相关性分析

Pearson 相关性分析可以用来判断各变量之间相互关系的密切程度。假设两个变量(X,Y)的 n 对样本为$(x_i,y_i), i = 1,2,..,n$，则 Pearson 相关系数的计算公式为：

$$\mathrm{r} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

其中，$\bar{x}$和$\bar{y}$分别表示变量X和Y对应的样本均值。

代入样本数据进行计算，得到的相关系数值$|r|$在 0.9～1.0 之间则判断为强高度相关，0.8～0.9 之间则判断为高度相关，0.5～0.8 之间则判断为中度相关，0.3～0.5 之间则判断为低度相关，0.0～0.3 判断为弱相关或者不相关。

因此，可以采用 Pearson 法来分析指标之间的相关关系，为进一步研究除 *smoke* 外其他指标是否会对甲基化水平产生间接影响提供依据。计算所得的 Pearson 相关系数结果如表 4 所示。

观察 *methyl* 指标和各个指标的相关系数可以发现，除了 *smoke* 指标的相关系数较大之外，其余的相关关系都比较弱，与单因素方差分析的结果相吻合。其中，*age*、*smoke*、*gestage*、*race* 指标与 *methyl* 指标的相关系数为正，说明生产年龄越高、怀孕周期越长的孕妇的后代基因甲基化水平更高。并且，如前文所述，孕妇的吸烟习惯会给后代基因甲基化水平带来显著提升。此外，除非洲裔以及高加索人群外的其他种族的后代甲基化水平也会略微偏高。

表 4. 各指标 Pearson 相关系数结果

	age	*BMI*	*smoke*	*gestage*	*gender*	*edu*	*race*	*methyl*
age	1.00	-0.02	-0.05	0.01	-0.08	0.33	0.06	0.01
BMI	-0.02	1.00	-0.03	-0.03	0.01	-0.12	-0.20	-0.08
smoke	-0.05	-0.03	1.00	0.01	0.05	-0.35	-0.06	0.16
gestage	0.01	-0.03	0.01	1.00	-0.01	-0.04	0.00	0.06
gender	-0.08	0.01	0.05	-0.01	1.00	-0.03	0.00	-0.04
edu	0.33	-0.12	-0.35	-0.04	-0.03	1.00	0.38	-0.05
race	0.03	-0.20	-0.06	0.00	0.00	0.38	1.00	0.03
methyl	0.01	-0.08	0.16	0.06	-0.04	-0.05	0.03	1.00

观察 *methyl* 指标和各个指标的相关系数可以发现，除了 *smoke* 指标的相关系数较大之外，其余的相关关系都比较弱，与单因素方差分析的结果相吻合。其中，*age*、*smoke*、*gestage*、*race* 指标与 *methyl* 指标的相关系数为正，说明生产年龄越高、怀孕周期越长的孕妇的后代基因甲基化水平更高。并且，如前文所述，孕妇的吸烟习惯会给后代基因甲基化水平带来显著提升。此外，除非洲裔以及高加索人群外的其他种族的后代甲基化水平也会略微偏高。

（三）路径分析模型

1. 路径分析法简介

路径分析模型是一种因果关系模型，相较于普通的多元回归模型，它可以容纳更多环节的因果结构。在建模过程中，可以通过路径图把这些因果关系清楚地展示出来，并且以此为基础进行更深层次的分析，可以计算各个指标之间的相对重要程度，计算各个指标之间的直接影响和间接影响。[7]

根据前文分析，除 *smoke* 外的各指标对甲基化水平的直接影响较小，所以采用路径分析的方法把指标对甲基化水平的影响分为直接影响和间接影响，并且将指标之间的影响关系通过路径图做可视化呈现。

2. 路径分析模型的估计

由于路径分析不能分析定类数据，所以此时不考虑后代性别和种族对甲基化水平可能带来的间接影响。根据上文的相关系数矩阵，选取母亲吸烟情况、年龄、受教育程度、BMI 指数、怀孕周期和婴儿甲基化水平建立模型。母亲年龄、受教育程度、BMI 指标、怀孕周期为相互影响关系。受教育程度与吸烟情况、吸烟情况与婴儿甲基化水平是因果关系。使用 AMOS 软件对模型得到标准化路径分析结果，如图 1 所示。

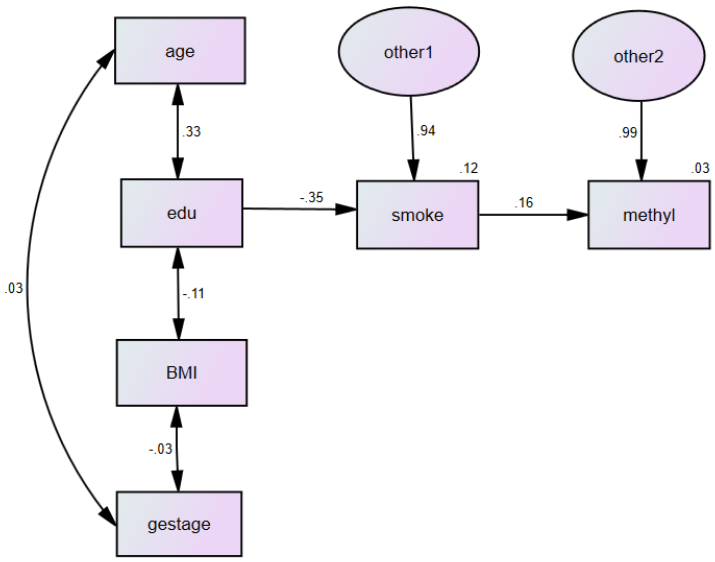

图 1 路径分析结果

根据路径分析模型结果，受教育程度对吸烟情况的总效应为-0.164，对甲基化水平的总效应为-0.004，吸烟情况对甲基化水平的总效应为 0.026。模型拟合优度指数$GFI = 0.993 > 0.9$，调整后拟合有度指数$AGFI = 0.978 > 0.9$，赤池信息$\mathrm{AIC} = 31.16$，大于独立模型的 AIC 值 198.376，由此可知，模型的拟合效果较好，建模时假设的因果关系和相互影响关系都有一定的说服力。

3. 路径分析模型结果分析

由路径分析模型结果可知，母亲吸烟对胎儿甲基化路径系数为正，说明二者之间有正效应，即吸烟习惯的母亲后代基因甲基化水平会有偏高的趋势。母亲受教育程度对吸烟情况的路径系数为负，说明二者之间有负效应，即受教育程度越高的母亲有吸烟习惯的可能越降低，从而可知受到高水平教育的母亲，产下的后代基因甲基化水平往往略微偏低。除此之外，母亲生育年龄与受教育程度有正向效应，BMI 指数与受教育程度呈现负向效应。可以得知，受教育程度高的女性往往生育年龄更高，更容易有晚育倾向。而该类人群的 BMI 值往往偏低，说明该类

人群更加自律以追求和保持健康的身材，并且会比较关注肥胖等因素带来的健康问题。这些指标都能给甲基化水平带来一定的弱效应。而怀孕周期和各个指标的相关关系都不强，对整体模型的影响程度也较小。

（四）小结

本章首先使用单因素方差分析和 Pearson 相关性分析，研究对变量之间的相关性，通过分析发现，在诸多变量中只有母亲的吸烟情况（*smoke*）对婴儿基因甲基化水平（*methyl*）的相关系数是显著的。接下来进行路径分析，确定了与胎儿甲基化异常有正效应的是母亲的吸烟情况，而年龄、受教育程度、BMI 指数、怀孕周期等指标作为间接影响因素对胎儿甲基化情况也存在影响。

四、模型的建立与求解

（一）地理加权回归 GWR 模型

在路径分析模型中，我们假设进行 DNA 样品实验测量甲基化程度的过程中不存在实验误差；但是在实际实验过程中会存在无法避免的实验误差，本节将考虑不同的孔板以及位于孔板的不同位置对实验结果产生的影响。在实验中一共用到了 20 个孔板上，每个孔板都是 8 行 12 列；并且假设这 608 份实验样品所处的孔板以及孔板上的具体位置是实验人员随机选择的。

绘制出全部实验结果的散点图，如图 2 所示，其中每个格子的深浅表示样品的甲基化程度。

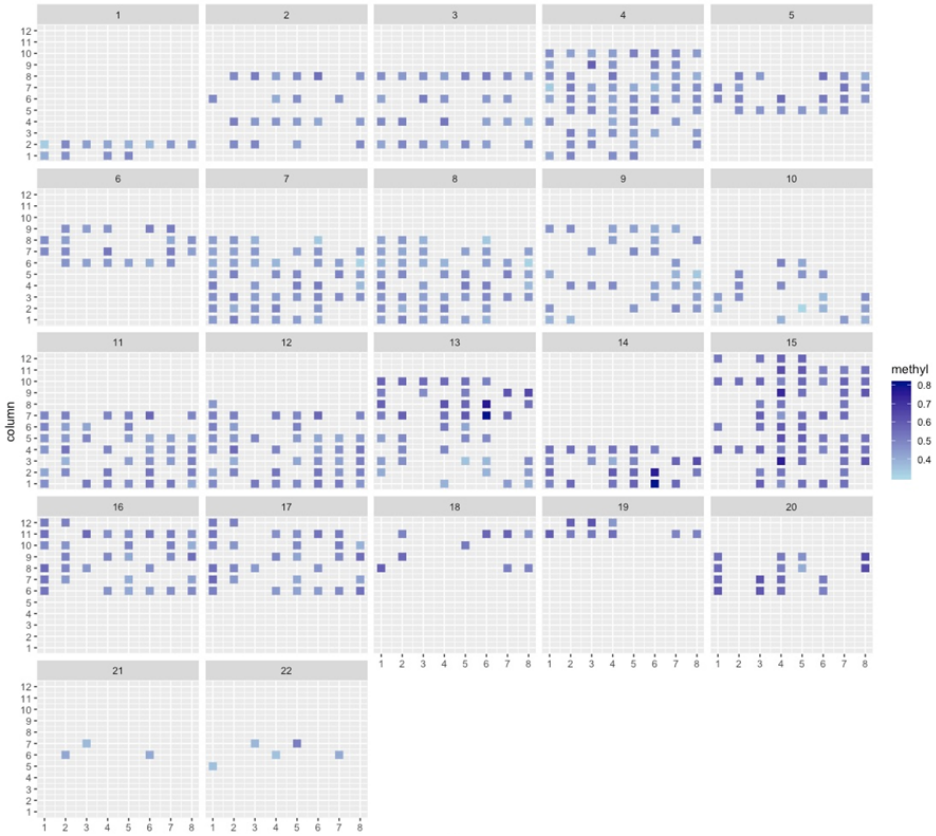

图 2 每个孔板甲基化水平

根据图 2 可以看出，孔板 1～10 上的样品甲基化程度较孔板 11-20 来说明显偏低。由于在实验过程中，实验员是随机选择的孔板，理论上每个孔板上样品的平均甲基化程度应该大致相同，但真实情况并非如此。这说明了在实验过程中孔板的不同会对实验结果产生一定程度的影响。

与此同时，对于样品数量较多的孔板进行分析，如孔板 15。我们发现位于孔板外围的样品的甲基化水平明显偏高。但是，由于实验数据数量较少，我们并不能完全确定孔板位置对实验结果是否产生影响。所以，本文将建立一维地理加权回归和二维地理加权回归模型，考察不同的孔板和同一孔板上的不同位置是与实验结果相关，通过对比两个模型预测结果的残差平方和以及其回归效果，选择最合适的模型。

1. 地理加权模型的建立

回归分析通常是指确定两种或两种以上变量相互依赖的定量关系的一种统计分析方法，但是因为它假定回归系数是全局一致的[8]，即没有考虑到空间非平稳性对研究的影响。因此，一般的线性回归不能研究孔板位置对测量结果的影响，所以需要建立 GWR（Geographically Weighted Regression）地理加权回归模型

对问题进行研究。GWR 模型就是在一般线性回归模型的基础上，将距离或坐标作为变量的权重建立修正后的线性回归模型。

GWR 模型中对每个样本点都设置了与其相对应的回归系数，其目的是反映出参数在区间内的非平稳变化。GWR 模型的基本形式是：

$$y_i = \beta_0(u_i, v_i) + \sum_{k=1}^{p} \beta_k(u_i, v_i) x_{ik} + \varepsilon_i \quad i = 1,2,\dots,n \tag{2}$$

其中，(u_i, v_i)是第i个样本点的坐标，x_{ik}是第 i 个样本的第 k 个解释变量，回归系数β_k是(u_i, v_i)的函数。

GWR 模型的参数估计与一般线性回归模型的原理相同，使用最小二乘法，使得建立的目标函数满足每个样本点的残差平方和最小：

$$f(\hat{\beta}_{i0}, \hat{\beta}_{i1}, \dots, \hat{\beta}_{ip}) = min \sum_{i=1}^{n} w_{ij}(y_i - \hat{y}_i)^2$$

其中，w_{ij}是对应的距离函数。用矩阵表示第i个样本的回归参数$\hat{\beta}_i$的估计为：

$$\hat{\beta}_i = (X^T W_i X)^{-1} X^T W_{iy} \tag{3}$$

得到第i个点的估计值为：

$$\hat{y}_i = X_i \hat{\beta}_i = X_i (X^T W_i X)^{-1} X^T W_{iy} \tag{4}$$

最后，根据高斯函数建立空间权函数，权函数表示为：

$$w_{ij} = \exp\left(-\left(\frac{d_{ij}}{h}\right)^2\right)$$

其中，h表示带宽。在实际应用中，一般采用固定的带宽，它是在均值和方差之间平衡的一个参数，但是如果带宽过小，模型会趋于全局模型。为了避免上述情况的出现，本文选择 CV 法确定带宽，即利用交叉确认法来确定权重的带宽，得到带宽如下：

$$h = \frac{1}{n} \sum_{i=1}^{n} [y_i - \hat{y}_{\neq i}(b)]^2$$

其中，n是样本点的数量，y_i是因变量Y在观察点(u_i, v_i)的观测值，$\hat{y}_{\neq i}(b)$是在给定带宽下，去掉(u_i, v_i)处的观测值后，利用上述拟合方法求得因变量y在观察点(u_i, v_i)的拟合值[9]。通过将带宽函数带入空间权函数中，就可以得到每个样本点的估计值。

2. 地理加权模型的求解

（1）一维地理加权模型

首先，考虑不同的孔板对实验结果的影响，由于 GWR 模型要求观察点的坐标(u_i, v_i)是二维欧氏空间中的坐标，所以我们将所有观察点的坐标设置为(u_i, c)，其中u_i表示实验样品的孔板标号，c表示某一特定的常数。根据单因素方差分析的检验结果，只有 *smoke* 一个变量与甲基化水平存在相关关系，这说明在诸多变量中只有 *smoke* 对甲基化水平有直接影响。所以在这里，以为地理加权模型的自变量是 *smoke*，因变量是 *methyl*。

设 *smoke* 为 *x*，*methyl* 为 *y*。建立地理加权模型如公式（5）所示：

$$ln\frac{y_i}{1-y_i} = \beta_0(u_i, c) + \beta_1(u_i, c)x_{ik} + \varepsilon_i \ \ i = 1,2,\dots,n \tag{5}$$

其中，对y进行 logit 变换是为了限制预测结果在 0～1 之间，所以需要根据公式（1）进行 logit 变换；(u_i, c)是第i个样本点的坐标，表示样本点位于第u_i个孔板上。

接下来，根据公式（3）利用加权最小二乘法对地理加权模型中的参数进行求解：

$$\hat{\beta}_i(u_i, c) = (X^T W_i X)^{-1} X^T W_{iy} \tag{6}$$

最后，将公式（6）带入公式（5）中，得到第i个点的估计值为，如公式（7）所示。

$$ln\frac{\hat{y}_i}{1-\hat{y}_i} = X_i\hat{\beta}_i(u_i, c) = X_i(X^T W_i X)^{-1} X^T W_{iy} \tag{7}$$

（2）二维地理加权模型

与一维地理加权模型不同，二维地理加权模型考虑不同的孔板和孔板上的不同位置对实验结果产生的影响，所以我们将每个实验样本点的坐标记为(u_i, v_i)，其中u_i表示实验样品位于孔板的具体位置，v_i表示实验样品位于孔板的标号。对公式（5）进行调整，得到地理加权回归模型如下：

$$ln\frac{y_i}{1-y_i} = \beta_0(u_i, v_i) + \beta_1(u_i, v_i)x_{ik} + \varepsilon_i \ \ i = 1,2,\dots,n \tag{8}$$

其中，(u_i, v_i)是第i个样本点的坐标，表示样本点第v_i个孔板上的第u_i个孔。

接下来，利用加权最小二乘法对地理加权模型中的参数进行求解：

$$\hat{\beta}_i(u_i, v_i) = (X^T W_i X)^{-1} X^T W_{iy} \tag{9}$$

得到第i个点的估计值为：

$$ln\frac{y_i}{1-y_i} = X_i\hat{\beta}_i(u_i, v_i) = X_i(X^T W_i X)^{-1}X^T W_{iy} \tag{10}$$

3. 地理加权模型结果分析

使用 R 语言实现对一维地理加模型和二维地理加权模型的估计，本文选择使用较为广泛而且准确率较高的高斯法确定空间权函数，得到估计效果如表 5 和表 6 所示。根据表 5～6，显然二维 GWR 模型的残差平方和较小、AIC 值较小、R^2较大，这说明二维 GWR 模型（同时考虑实验样品放置的孔板及其在孔板上的位置）的模型较好。

表 5. 一维 GWR 模型估计效果

参数	Gauss 法
带宽	0.5640
残差平方和	24.4537
AIC 值	-116.4885
R^2	0.4421

表 6. 二维 GWR 模型估计效果

参数	Gauss 法
带宽	1.0822
残差平方和	9.5262
AIC 值	-556.2434
R^2	0.7826

判断数据是否适合 GWR 模型的主要依据之一是判断模型参数的估计结果是否具有空间非平稳性，所以本文通过对比 GWR 模型参数随坐标的分布，来判断 GWR 模型是否显著。

图 3 和图 4 显示的是一维 GWR 模型和二维 GWR 模型中参数β_0的估计值随坐标变换的分布情况，显然二维 GWR 模型中β_0具有明显的空间非平稳性，但是一维 GWR 模型中β_0却是空间平稳的。

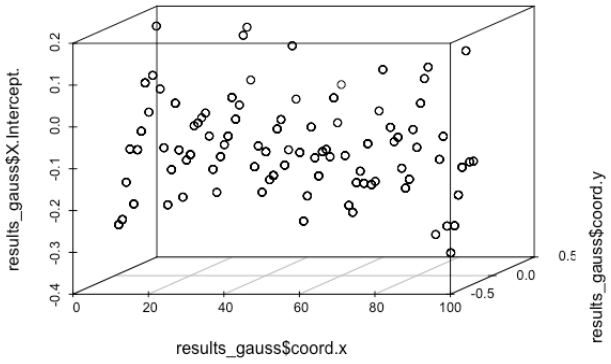

图 3 一维 GWR 模型中参数β_0估计值的分布图

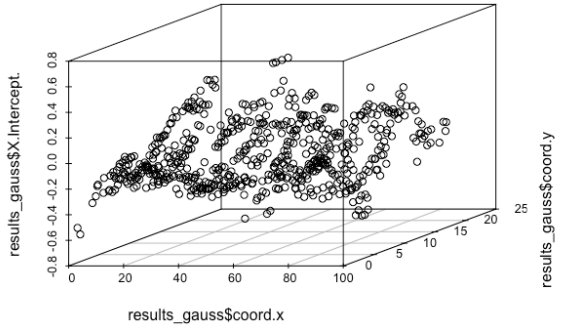

图 4 二维 GWR 模型中参数β_0估计值的分布图

根据图 5 和图 6 所显示的 GWR 模型中参数估计值β_1的分布情况，认为一维 GWR 模型中β_1具有微弱的空间非平稳性，二维 GWR 模型中β_1的空间非平稳性非常明显。根据图 2-6 和模型的参数估计结果，认为二维 GWR 模型更适合于分析本文所收集到的数据。

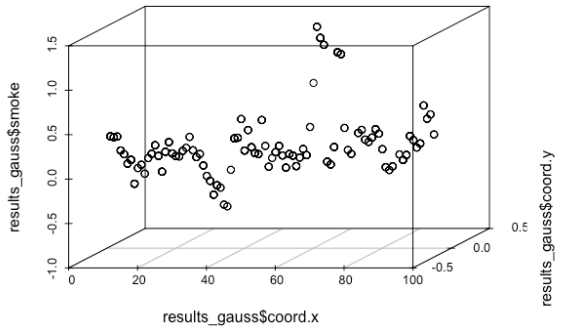

图 5 一维 GWR 模型中参数β_1估计值的分布图

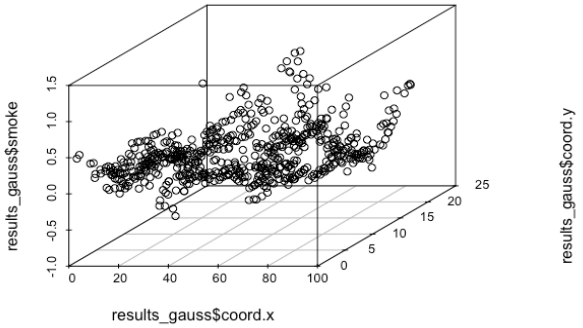

图 6 二维 GWR 模型中参数β_1估计值的分布图

通过上述分析证实了二维 GWR 模型更加适合解决本文所要研究的问题，随后对二维 GWR 模型的参数估计结果进行分析，模型的参数估计结果如表 7 所示。由于一维模型并不适用于本文，所以接下来所提及的 GWR 模型都表示二维 GWR 模型。

表 7. GWR 模型参数估计结果

参数	Min.	1st Qu.	Median	3rd Qu.	Max.	Global
$X.Intercept$	-0.5713	-0.2367	-0.1324	0.0041	0.4361	-0.1122
$\theta^{-1}(X.Intercept)$	0.3609	0.4411	0.4669	0.5010	0.6073	0.4719
$smoke$	-0.9664	-0.0502	0.0799	0.2092	1.3720	0.1135

根据表 7 的结果所示，由于截距项的特殊意义，所以对其进行 logit 逆变换。GWR 模型中截距项逆变换后的结果我们发现进行 logit 变换后的截距项约为 50%，这说明了对于大部分孕妇来说如果在妊娠期不吸烟，基本上会保证胎儿的甲基化程度处于正常水平。对于变量 *smoke* 来说，虽然回归系数的整体跨度较大，但是根据其中位数和平均数可以看出，吸烟在大多数情况下会导致胎儿的甲基化程度偏大.

4. 小结

综上所述，在测量 DNA 甲基化程度的实验过程中，不同的孔板以及孔板上不同的位置都会对实验结果产生比较严重的影响，所以为了避免实验误差影响模型的预测结果，本文选择建立 GWR 模型进行回归。通过 GWR 模型的参数可以看出，在如果测试者选择不吸烟，则胎儿的甲基化程度可以基本维持在正常水平，如果选择吸烟，则会导致胎儿的甲基化程度明显增大。

（二）修正的 GWR 模型

在上一小节中的 GWR 模型只考虑了 *smoke* 一个变量对胎儿甲基化程度的直接影响，但是通过路径分析发现，除了 *smoke* 以外，还有 *age*、*BMI*、*race*、*gender* 等变量对胎儿的甲基化程度有间接影响。为了更准确地分析影响胎儿甲基化程度的因素，本节将考虑一些协变量对甲基化程度的影响，对于一些变量属于定性变量，需要通过引入虚拟变量对其进行处理。因此，本节所建立的修正的 GWR 模型是指引入虚拟变量的二维 GWR 模型。

1. 修正的 GWR 模型的建立

对于数据集中的定性变量，如孕妇的种族以及胎儿的性别等等，如果想要把这些变量引入到 GWR 模型中，，所以就需要引入虚拟变量对其进行刻画。定性变量可以具有多个水平，当某种属性或状态不存在时，将虚拟变量定义为 0，反之定义为 1，表示具有某种属性[10]。

虚拟变量的个数设置规则是：如果定性变量有 *m* 个水平，则需要设置 *m-1* 个虚拟变量，如果引入 *m* 个就会存在多重共线性。如果用$A_i(i = 1,2,..,m)$表示定性变量的 *m* 个水平，则引入虚拟变量$D_j(j = 1,2,\dots,m-1)$:

$$D_j = \begin{cases} 1, x \in A_j \\ 0, x \notin A_j \end{cases} \tag{11}$$

接下来，将虚拟变量引入 GWR 模型中，即在公式（2）中引入虚拟变量，得到回归方程如公式（12）所示。

$$y_i = \beta_0(u_i, v_i) + \sum_{k=1}^{p} \beta_k(u_i, v_i) x_{ik} + \sum_{k=p+1}^{q} \beta_k(u_i, v_i) D_{il} + \varepsilon_i \ \ i = 1,2,\dots,n \tag{12}$$

仍然采用最小二乘法对模型中的参数进行估计，用矩阵表示第 *i* 个样本的回归参数$\hat{\beta}_i$的估计如公式（3）所示，第 *i* 个点的估计值如公式（4）所示。

2. 修正的 GWR 模型的求解

根据每个变量的水平，本文对 *race*、*gender* 两个定性指标引入相应的虚拟变量。

① *race*

变量 *race* 表示母亲的种族，一共有非洲裔美国人（AA）、高加索人种（EA）以及其他三个水平，所以根据公式（11）设置两个虚拟变量：

$$D_1=\begin{cases}1, x_{ij}\in AA\\ 0, x_{ij}\notin AA\end{cases}\quad D_2=\begin{cases}1, x_{ij}\in EA\\ 0, x_{ij}\notin EA\end{cases}$$

② *gender*

变量 *gender* 表示胎儿的性别，分为男孩（boy）和女孩（girl），所以根据公式（11）设置一个虚拟变量：

$$D_2=\begin{cases}1, x_{ij}\in boy\\ 0, x_{ij}\notin girl\end{cases}$$

根据相关性检验和路径分析的结果，除了变量 *smoke* 以外，其余变量 *age*、*edu*、*BMI*、*race*、*gender* 都是通过作用到变量 *smoke* 上进而对胎儿甲基化水平产生影响，所以在引入协变量时，需要考虑协变量和变量 *smoke* 的交互作用对甲基化水平的影响。但是为了清晰地分析每个协变量对胎儿甲基化程度的影响，本文选择每次只引入一个协变量对其进行分析，由公式（12）建立修正的 GWR 模型如下：

① *age* 与 *smoke* 交互作用：

$$ln\frac{y_i}{1-y_i}=\beta_0+\beta_1 smoke+\beta_2 age+\beta_3 age\times smoke \tag{13}$$

② *edu* 与 *smoke* 交互作用：

$$ln\frac{y_i}{1-y_i}=\beta_0+\beta_1 smoke+\beta_2 edu+\beta_3 edu\times smoke \tag{14}$$

③ *BMI* 与 *smoke* 交互作用：

$$ln\frac{y_i}{1-y_i}=\beta_0+\beta_1 smoke+\beta_2 BMI+\beta_3 BMI\times smoke \tag{15}$$

④ *race* 与 *smoke* 交互作用：

$$ln\frac{y_i}{1-y_i}=\beta_0+\beta_1 smoke+\beta_2 race+\beta_3 race\times smoke \tag{16}$$

⑤ *gender* 与 *smoke* 交互作用：

$$ln\frac{y_i}{1-y_i}=\beta_0+\beta_1 smoke+\beta_2 gender+\beta_3 gender\times smoke \tag{17}$$

同样使用加权最小二乘法对模型（13）～（17）进行求解。

3. 修正的 GWR 模型结果分析

用 R 语言对修正的 GWR 模型进行估计，在前文中已经证明过 GWR 模型的显著性，而且在引入虚拟变量后不会改变数据的空间非平稳性，所以在此就不再赘述与模型显著性相关的内容，而着重分析变量 *smoke* 与其它协变量的交互作用对胎儿甲基化产生的影响。

修正的 GWR 方程的截距项表示当孕妇不吸烟的情况下，胎儿甲基化程度受到某一协变量的影响程度；*smoke* 的系数表示胎儿甲基化程度对吸烟情况的敏感度受该协变量的影响程度。

① *age* 和 *smoke* 交互作用

根据图 7-A 所示，孕妇生育年龄小于 30 岁时，修正的 GWR 模型的截距偏离 0 最为明显，即胎儿的甲基化程度明显低于正常水平 50%，当孕妇的生育年龄大于等于 30 岁时，胎儿的甲基化程度接近正常水平 50%，这说明如果孕妇不吸烟，她的生育年龄不会使得胎儿的甲基化水平高于正常水平。根据图 7-B 所示，孕妇的生育年龄大于 39 都会导致胎儿甲基化程度对吸烟状况十分敏感，也就是说如果这个年龄区间段对孕妇吸烟将极大程度上导致胎儿甲基化程度偏大。

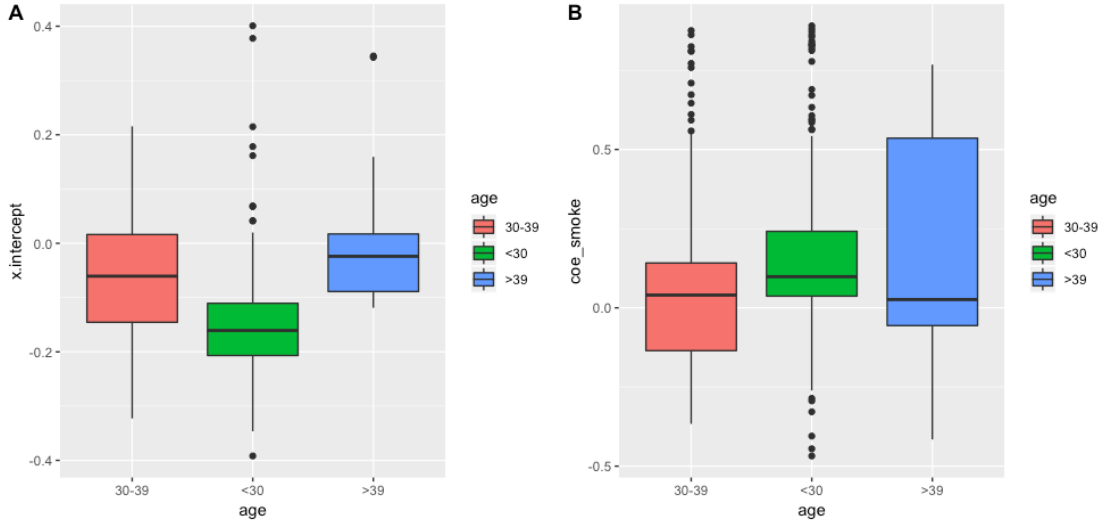

图 7 变量 *age* 对截距、斜率分布的影响

② *edu* 和 *smoke* 的交互作用：

根据图 8-A 所示，孕妇的受教育程度越高，胎儿的甲基化水平越接近正常水平，究其原因主要是受教育程度较高的孕妇更加注重自身的健康以及胎儿的健康，这将使得胎儿甲基化异常的概率大大下降。根据图 8-B 所示，胎儿的甲基化水平对吸烟的敏感度与母亲的受教育水平无关，这是因为如果孕妇在孕期吸烟，则她们子宫内的环境将和其受教育程度无关，自然也就不会影响到胎儿甲基化对吸烟的敏感度。

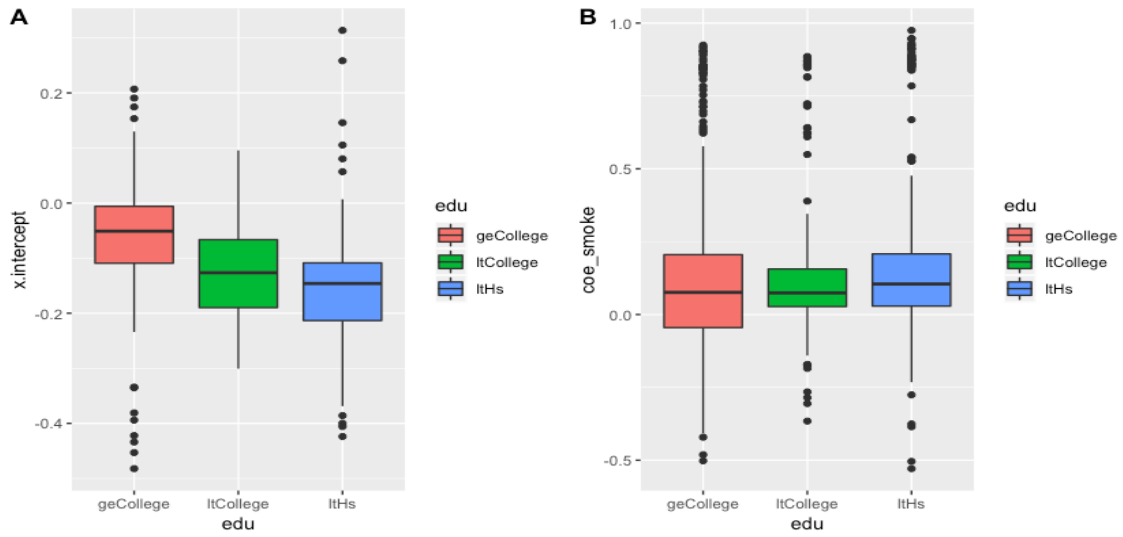

图 8 变量 *edu* 对截距、斜率分布的影响

③ *BMI* 和 *smoke* 的交互作用：

根据图 9-A 所示，孕妇的 BMI 指标偏低将使得胎儿甲基化程度偏低，而 BMI 指标稍高一些的孕妇的胎儿甲基化水平在正常值附近波动，这说明了孕妇在孕期的适当增重是有益于胎儿甲基化水平的，并不会导致胎儿的生长异常。根据图 9-B 所示，孕妇的 BMI 指标较大时，胎儿甲基化水平对吸烟情况较孕妇 BMI 指标偏小时敏感，也就是说偏胖的孕妇的子宫环境可以在一定程度上增加香烟烟雾对胎儿造成的不利影响。

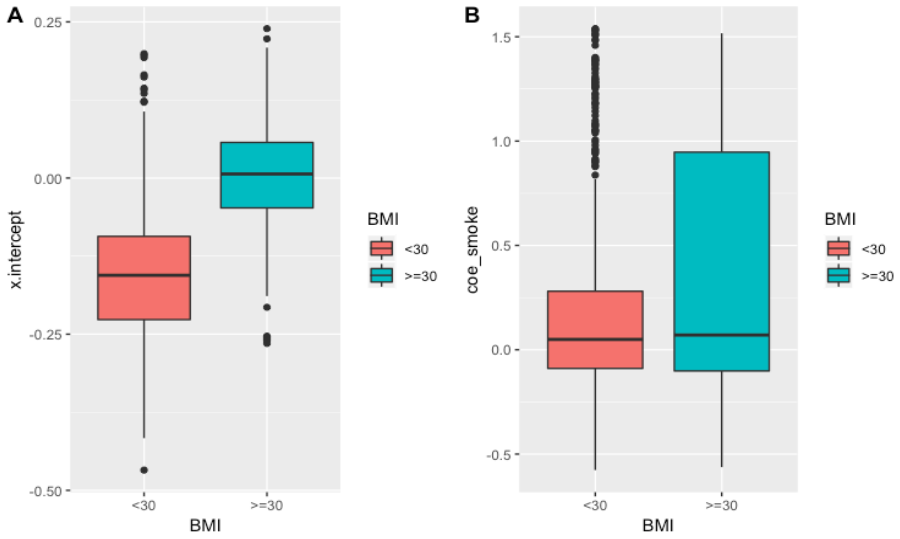

图 9 变量 *BMI* 对截距、斜率分布的影响

④ *race* 和 smoke 的交互作用：

根据图 10-A 所示，高加索人种的孕妇的胎儿甲基化程度较其他人种更接近正常值，非洲裔人种的孕妇的胎儿更容易因为人种的 DNA 问题，而出现甲基化程

度异常的现象。根据图 10-B 所示，非洲裔人种孕妇的子宫环境对吸烟状况是非常敏感的，吸烟将使得胎儿的甲基化程度明显高于正常值，其次是高加索人种的孕妇，她们子宫的环境虽然不像非洲裔人种孕妇的那么敏感，但仍然会导致胎儿的甲基化程度增加。

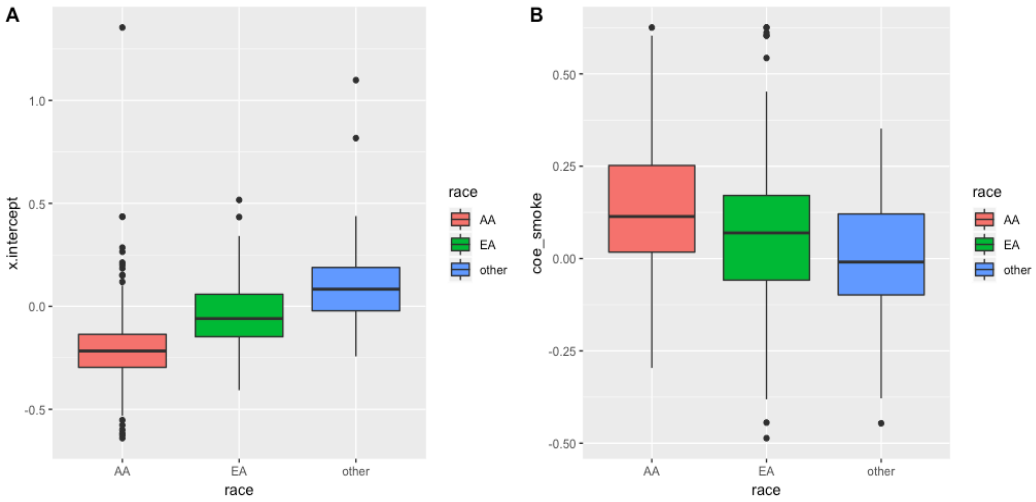

图 10 变量 *race* 对截距、斜率分布的影响

⑤ *gender* 和 *smoke* 的交互作用：

根据图 11-A 所示，胎儿性别为女性时，修正 GWR 方程的截距项跨度更大，但是通过比较中位数，女性胎儿的中位数更接近正常值，也就是说男性胎儿更容易出现甲基化程度偏低的情况。根据图 11-B 所示， 胎儿的性别与胎儿甲基化程度对吸烟的敏感度无关，也就是说吸烟对胎儿身体造成的影响与胎儿的性别无关。

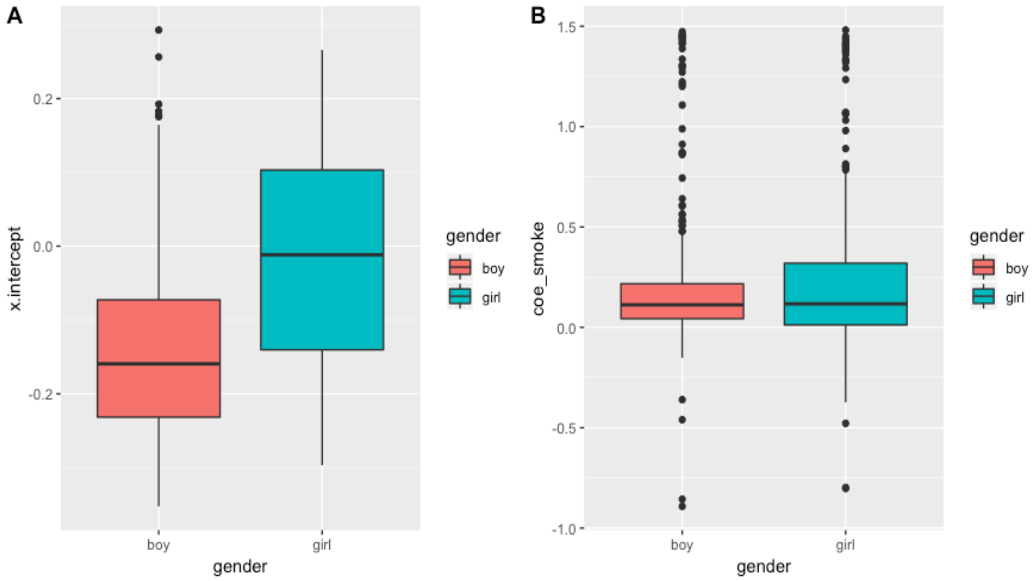

图 11 变量 *gender* 对截距、斜率分布的影响

4.小结

建立修正后的GWR模型，分析 *smoke* 和协变量 *age*、*edu*、*BMI*、*race*、*gender* 的交互作用得到如下结论：

（1）孕妇年龄偏小会导致胎儿的甲基化程度低于正常值，孕妇年龄偏大会使得胎儿的甲基化程度对吸烟状况十分敏感，也即对于年龄偏大的孕妇来说吸烟极易导致胎儿甲基化程度偏大；

（2）受教育水平较高的孕妇对自身和胎儿的健康重视程度更高，使得胎儿甲基化异常出现的概率较低；

（3）对于孕妇来说适当的增重不会导致胎儿甲基化异常现象的出现，但是偏胖孕妇的子宫环境会使得胎儿的甲基化程度因为孕 妇吸烟而急剧增加；

（4）非洲裔的孕妇的胎儿更容易出现甲基化异常现象的出现；

（5）男性胎儿的甲基化程度天生偏低，但是胎儿的性别与其甲基化水平对吸烟是否敏感无关。

（三）小结

本章通过建立地理加权模型排除了由于实验误差对分析结果产生的影响，提高了研究结论的准确率。根据地理加权模型的相关参数，发现孕妇吸烟会导致胎儿的甲基化程度偏大，而不吸烟的孕妇则不会出现类似情况。在上述模型的基础上，引入虚拟变量，研究 *smoke* 和协变量 *age*、*edu*、*BMI*、*race*、*gender* 的交互作用，通过分析模型的参数，得到了每个变量对胎儿甲基化的影响情况。

五、 结论与建议

（一）结论

文章主要考察了母亲分娩年龄、母亲孕前BMI指标、孕期吸烟状况、孕产期、胎儿性别、母亲的受教育程度、母亲的种族及胎儿的甲基化程度等七个指标，目的在于研究胎儿甲基化程度与母亲孕期是否吸烟状况之间的关系，同时也研究了其他指标对吸烟与胎儿甲基化程度之间关系的影响。

首先，进行变量相关关系分析，研究影响胎儿甲基化的主要因素；根据相关性检验结果进行路径分析，研究胎儿甲基化除了受到母亲吸烟状况影响以外，还受到哪些协变量的影响；接下来，考虑检测样本在不同的孔板以及不同的位置上对实验结果带来的影响，首先分别建立只考虑孔板的一维GWR模型和考虑孔板以

及孔板的位置的二维 GWR 模型进行比较，确认拟合效果更好的模型。在此基础上引入各虚拟变量研究吸烟状况对甲基化程度的影响是否与其他的指标相关。

在变量相关关系分析中，本文主要进行了单因素方差分析和相关性检验。通过单因素方差分析，发现仅有母亲的吸烟状况对胎儿甲基化水平有显著的直接影响，而且吸烟通常会导致胎儿的甲基化程度偏大。在相关性检验中，*methyl* 指标和除 *smoke* 之外其余的指标相关关系都比较弱，与单因素方差分析的结果相吻合。其中，*age*、*smoke*、*gestage*、*race* 指标与 *methyl* 指标的相关系数为正，说明生产年龄越高、怀孕周期越长的孕妇的后代基因甲基化水平更高，同时孕妇的吸烟习惯会给后代基因甲基化水平带来显著提升。此外，除非洲裔以及高加索人群外的其他种族的后代甲基化水平也会略微偏高。

根据相关关系分析的结果建立路径分析模型，研究吸烟状况对甲基化成的的直接影响和分娩年龄、教育程度、BMI 水平对甲基化成的的间接影响。通过路径分析，本文得到如下结论：①母亲在妊娠期吸烟会导致胎儿甲基化程度偏大；②由于受教育程度较高的女性存在晚孕倾向，而且 *BMI* 指标正常，这使得后代的甲基化程度越接近正常水平；③母亲生育年龄与受教育程度有正向效应，*BMI* 指数与受教育程度呈现负向效应。

在建立 GWR 模型中，首先分别建立只考虑孔板的一维 GWR 模型和考虑孔板和位置的二维 GWR 模型，通过比较残差平方和以及 AIC 统计量判断二维 GWR 模型拟合效果更好。在此基础上分析得到 *smoke* 变量的回归系数具有空间非平稳性，进一步证实了样本在检测过程中存在由位置带来的结果偏差。随后，在 GWR 模型中引入母亲分娩年龄、母亲孕前 *BMI* 指数、母亲受教育程度、种族、胎儿性别变量对应的虚拟变量，建立修正 GWR 模型。通过分析 GWR 方程的截距项和斜率项在不同指标上的不同分布，得到如下结论：①孕妇年龄偏小会导致胎儿的甲基化程度低于正常值，孕妇年龄偏大会使得胎儿的甲基化程度对吸烟状况十分敏感，也即对于年龄偏大的孕妇来说吸烟极易导致胎儿甲基化程度偏大；②受教育水平较高的孕妇对自身和胎儿的健康重视程度更高，使得胎儿甲基化异常出现的概率较低；③对于孕妇来说适当的增重不会导致胎儿甲基化异常现象的出现，但是偏胖孕妇的子宫环境会使得胎儿的甲基化程度因为孕妇吸烟而急剧增加；④非洲裔的孕妇的胎儿更容易出现甲基化异常现象的出现；⑤男性胎儿的甲基化程度天生偏低，但是胎儿的性别与其甲基化水平对吸烟是否敏感无关。

（二）建议

基于对上述修正的 GWR 模型进行分析，本文为孕妇在怀孕事情的相关事项提出如下建议:

①科普孕期中吸烟对胎儿成长发育的不利影响，以此减少孕妇吸烟的人数以及频率；必要时，也可以出台公共场所控制吸烟条例，尤其是非洲裔群体聚集的地区，以减少二手烟对孕妇及胎儿产生的健康影响。

②提倡适龄生育理念，女性最佳生育年龄在 25-29 岁之间。在适龄生育利于女性孕育和抚育婴儿，可避免发育不良，妊娠合并症及其他潜在疾病的发生。

③普及科学的备孕方法，促使孕妇在孕产期中保持健康适中的体质，合理饮食、适度锻炼、保持心情愉悦，将 BMI 指数维持在合适的范围内；

④加强孕期检查的重要性的宣传力度，提倡孕妇在怀孕期间，定期到医院进行产期检查，以便及时了解胎儿的发育情况，并及时采取对应的措施。尤其注意性别为男的胎儿的甲基化水平，减少健康隐患。

参考文献

[1]Markunas C A , Xu Z , Harlid S , et al. Identification of DNA Methylation Changes in Newborns Related to Maternal Smoking during Pregnancy[J]. Environmental Health Perspectives, 2014, 122(10):1147---1153.

[2]黄镇铭, 肖靖雨, 付玉, 赵伟, 张云辉. DNA 甲基化在肺癌与环境暴露间关联的研究进展[J]. 肿瘤, 2016, 36(11):1272-1279.

[3]温世宝. 新生儿出生特征及 H19 DNA 甲基化与孕期环境暴露关系[D]. 郑州大学, 2012.

[4]Rolle-Kampczyk, Ulrike. (2016). Environment-induced epigenetic reprogramming in genomic regulatory elements in smoking mothers and their children. 10.13140/RG.2.1.3637.0323.

[5]Hoyo, C., Fortner, K., Murtha, A. P., Schildkraut, J. M., Soubry, A., Demark-Wahnefried, W., … & Huang, Z. (2012). Association of cord blood methylation fractions at imprinted insulin-like growth factor 2 (IGF2), plasma IGF2, and birth weight. Cancer Causes & Control, 23(4), 635-645.

[6]廖祥超. 九种常用缺失值插补方法的比较[D]. 云南师范大学, 2017.

[7]何晓群. 多元统计分析. 第 3 版[M]. 中国人民大学出版社, 2012

[8]张金亭, 赵瑞. 基于地理加权回归的环渤海城市群房价影响因子研究[J]. 国土与自然资源研究, 2019(01):87-93.

[9]王梦晗. 基于时空地理加权回归模型的北京市房价影响因素研究[D]. 山东农业大学, 2018.

[10]戴金辉. 虚拟变量回归及其应用[J]. 统计与决策, 2019, 35(05):77-80.

附录

附录 1 原始数据

在此仅摘选前 38 行原始数据

age	*BMI*	*smoke*	*gestage*	*gender*	*edu*	*race*	*methyl1*	*plate1*	*row1*	*column1*	*well1*
30to39	0	0	0	0	geCollege	EA	38.36	I	A	1	A1
lt30	0	0	0	0	geCollege	EA	37.85	I	B	1	B1
30to39	1	0	0	1	geCollege	EA	38.57	L	D	1	D1
30to39	1	0	0	0	geCollege	EA	39.75	A	A	1	A1
30to39	0	0	0	1	geCollege	EA	43.83	L	G	1	G1
30to39	0	0	0	0	ltHS	AA	39.08	L	H	1	H1
lt30	0	0	0	1	geCollege	Other	39.12	L	A	2	A2
30to39	1	0	0	0	geCollege	AA	30.15	L	E	2	E2
30to39	1	0	0	0	geCollege	EA	38.04	L	F	2	F2
lt30	0	0	0	1	ltHS	EA	55.95	W	A	8	A8
30to39	1	0	0	1	ltCollege	AA	46.28	A	B	1	B1
lt30	1	0	0	1	ltHS	AA	46.67	A	D	1	D1
lt30	0	0	0	1	geCollege	EA	43.12	L	H	2	H2
30to39	0	0	0	1	geCollege	EA	48.72	A	E	1	E1
lt30	0	0	0	0	ltCollege	EA	43.48	L	A	3	A3
lt30	0	1	0	0	ltHS	EA	46.43	L	B	3	B3
lt30	1	0	0	1	ltHS	AA	45.32	I	A	2	A2
lt30	1	1	1	1	ltHS	AA	48.02	B	B	2	B2
lt30	0	1	0	1	ltHS	EA	48.85	B	C	2	C2
lt30	0	0	1	1	ltHS	EA	34.84	O	E	3	E3
30to39	0	0	0	1	geCollege	EA	36.56	L	F	3	F3
30to39	0	0	0	0	geCollege	EA	42.62	B	E	2	E2
lt30	0	0	0	1	geCollege	EA	47.58	L	H	3	H3
30to39	0	0	0	0	geCollege	EA	48.14	L	B	4	B4
30to39	1	0	0	0	geCollege	EA	48.28	L	D	4	D4
lt30	0	1	0	1	ltHS	EA	44.12	I	C	7	C7
lt30	0	1	0	0	ltCollege	EA	47.69	B	H	2	H2
30to39	1	0	0	1	ltHS	AA	50.97	B	B	4	B4
30to39	0	0	0	0	geCollege	EA	47.93	I	E	7	E7
30to39	1	0	0	1	geCollege	AA	45.55	I	F	7	F7
30to39	0	0	0	1	geCollege	EA	44.85	B	C	4	C4
30to39	0	1	0	0	ltHS	AA	45.59	B	D	4	D4
30to39	0	0	0	0	ltCollege	AA	43.9	B	E	4	E4
lt30	1	0	0	1	ltCollege	EA	41.19	B	F	4	F4
lt30	1	0	0	1	ltHS	AA	44.05	B	H	4	H4
lt30	0	0	0	1	ltHS	EA	48.16	B	A	6	A6

lt30	0	1	0	1	ltHS	AA	40.37	B	D	6	D6

数据说明

共有 314 行，每行对应一个样本，每列对应如下指标：

age：母亲年龄编码为“IT30”（分娩时小于 30 岁）、“30 至 39”（分娩时为 30 至 39 岁）和“GE40”（分娩时大于 40 岁）；

BMI：母亲的身体质量指数，孕前体重（单位：kg）除以身高（单位：米）的平方和。编码为“0”（=30 以下）或“1”（大于或等于 30）；

smoke：如果不吸烟，则编码为“0”；如果在怀孕早期吸烟，则编码为“1”；

gestage：孕产期，“1”表示小于 37 周、“0”表示大于等于 37 周；

gender：胎儿的性别（“1”=男，“0”=女）；

edu：母亲的受教育程度（‘ltHs’表示低于高中水平，‘ltCollege’表示高中/GED，‘geCollege’表示至少是大学）

race：母亲的种族/民族（‘AA’表示非洲裔美国人，‘EA’表示高加索人种或其它）

methyl1：第一次测量受试者孩子的甲基化水平；

methyl2：第二次测量受试者孩子的甲基化水平，20 个样本的第二次测量数据是缺失的；

plate1(and ‘plate2’)：受试者第一次和第二次测量所在的板子，‘plate2’有 20 个只进行一次实验的数据是缺失的；

row1（and ‘row2’）：将受试者的第一次和第二次重复测量值分别放在 ID 分别出现在“plate1”和“plate2”中的平板上的行，20 名受试者缺少“row2”；

column1（and column2）：将受试者的第一次和第二次重复测量值分别放在 ID 分别出现在“plate1”和“plate2”中的平板上的列。

20 名受试者缺少“row1”

well1（and well2）：在这个实验中，受试者的第一次和第二次重复测量被放在 ID 显示在“plate1”和“plate2”的平板上。

附录 2 BP 神经网络处理缺失值 R 语言代码

```
#BP-net
set.seed(2019)
```

```
epigen=read.csv("epigen1.csv")[,-c(5,9:12)]
original=epigen

library(lattice)
library(mice)
md.pattern(epigen)

#BMI_fill
sub=which(!complete.cases(epigen)==T)
missingdata=epigen[sub,]
completedata=epigen[-sub,]

library(nnet)
BPnet=nnet(BMI/max(BMI)~.,data=completedata,size=11,maxit=1000,decay=
0.01,trace=F)
missingdata[,2]=predict(BPnet,missingdata)*max(completedata$BMI)

nepigen=epigen
nepigen[c(sub),2]=predict(BPnet,epigen[c(sub),])*max(completedata$BMI)
anyNA(nepigen)

for(i in 1:608){
  if(nepigen$BMI[i]<=0.5)
    nepigen$BMI[i]=0
  else nepigen$BMI[i]=1
}

write.csv(nepigen,"epigen1_BMI.csv")

#gender_fill
epigen=read.csv("epigen1.csv")[,-c(2,9:12)]
original=epigen

#bp-net
sub=which(!complete.cases(epigen)==T)
```

```
missingdata=epigen[sub,]
completedata=epigen[-sub,]

library(nnet)
BPnet=nnet(gender/max(gender)~.,data=completedata,size=11,maxit=1000,
decay=0.01,trace=F)
missingdata[,4]=predict(BPnet,missingdata)*max(completedata$gender)

nepigen=epigen
nepigen[c(sub),4]=predict(BPnet,epigen[c(sub),])*max(completedata$gen
der)
anyNA(nepigen)



for(i in 1:608){
  if(nepigen$gender[i]<=0.5)
    nepigen$gender[i]=0
  else nepigen$gender[i]=1
}
write.csv(nepigen,"epigen1_gender.csv")
```

附录 3 实验样本孔板分布图 R 语言代码

```
epigen=read.csv('epigen4.csv')

library(ggplot2)
ggplot(epigen,aes(x=row,y=column,colour=methyl))+
  geom_point(shape=15,size=3)+
  scale_color_gradient(low="lightblue",high="darkblue")+
  facet_wrap(~plate,nrow=5)+
  scale_y_continuous(breaks = seq(0,12,1))+
  scale_x_continuous(breaks = seq(0,8,1))
```

附录 4 GWR 模型语言代码

```
library(spgwr)
```

```
epigen=read.csv("epigen_new.csv")[,-1]

##lm regression
#well and plate
epi.bw=gwr.sel(y~smoke,data=epigen,
               coords=cbind(epigen$well,epigen$plate))
epi.gauss=gwr(y~smoke,data=epigen,
              coords=cbind(epigen$well,epigen$plate),bandwidth      =
epi.bw,hatmatrix = T)
epi.gauss
results_gauss=as.data.frame(epi.gauss$SDF)



#only plate
epi.bw=gwr.sel(y~smoke,data=epigen,
               coords=cbind(epigen$plate,rep(0,608)))
epi.gauss=gwr(y~smoke,data=epigen,
              coords=cbind(epigen$well,rep(0,608)),bandwidth        =
epi.bw,hatmatrix = T)
epi.gauss
results_gauss=as.data.frame(epi.gauss$SDF)

#GWR 模型散点图
library(scatterplot3d)
scatterplot3d(results_gauss$coord.x,results_gauss$coord.y,results_gau
ss$smoke)
scatterplot3d(results_gauss$coord.x,results_gauss$coord.y,results_gau
ss$X.Intercept.)
```

附录 5 修正 GWR 回归 R 语言代码

```
##smoke:age
epigen=read.csv("epigen_new_dx.csv")[,-1]

epi.bw=gwr.sel(y~smoke*age,data=epigen,
               coords=cbind(epigen$well,epigen$plate))
```

```
epi.gauss=gwr(y~smoke*age,data=epigen,
              coords=cbind(epigen$well,epigen$plate),bandwidth        =
epi.bw,hatmatrix = T)
results_gauss=as.data.frame(epi.gauss$SDF)
A=data.frame(results_gauss$X.Intercept.,results_gauss$smoke+results_g
auss$smoke.age,epigen$age)

#boxplot
library(ggplot2)
library(magrittr)
library(ggpubr)

A=data.frame(x.intercept=results_gauss$X.Intercept.,coe_smoke=results
_gauss$smoke+results_gauss$smoke.age,age=epigen$age)
A[order(A$age),]
A$age=c(rep('<30',330),rep('30-39',256),rep('>39',22))

p1=ggplot(data=A,aes(x=age,y=x.intercept))+geom_boxplot(aes(fill=age))
p2=ggplot(data=A,aes(x=age,y=coe_smoke))+geom_boxplot(aes(fill=age))
ggarrange(p1,p2,ncol=2,nrow=1,labels = c("A","B"))

##boxplot_edu
epi.bw=gwr.sel(y~smoke*edu,data=epigen,
               coords=cbind(epigen$well,epigen$plate))
epi.gauss=gwr(y~smoke*edu,data=epigen,
              coords=cbind(epigen$well,epigen$plate),bandwidth       =
epi.bw,hatmatrix = T)
results_gauss=as.data.frame(epi.gauss$SDF)

A=data.frame(x.intercept=results_gauss$X.Intercept.,
coe_smoke=results_gauss$smoke+results_gauss$smoke.edu,edu=epigen$edu)
A[order(A$edu),]
A$edu=c(rep('ltHs',225),rep('ltCollege',159),rep('geCollege',224))

p1=ggplot(data=A,aes(x=edu,y=x.intercept))+geom_boxplot(aes(fill=edu))
```

```
p2=ggplot(data=A,aes(x=edu,y=coe_smoke))+geom_boxplot(aes(fill=edu))
ggarrange(p1,p2,ncol=2,nrow=1,labels = c("A","B"))

##boxplot_race
epi.bw=gwr.sel(y~smoke*race,data=epigen,
               coords=cbind(epigen$well,epigen$plate))
epi.gauss=gwr(y~smoke*race,data=epigen,
              coords=cbind(epigen$well,epigen$plate),bandwidth      =
epi.bw,hatmatrix = T)
results_gauss=as.data.frame(epi.gauss$SDF)

A=data.frame(x.intercept=results_gauss$X.Intercept.,

coe_smoke=results_gauss$smoke+results_gauss$smoke.race,race=epigen$ra
ce)
A[order(A$race),]
A$race=c(rep('AA',296),rep('EA',263),rep('other',49))

p1=ggplot(data=A,aes(x=race,y=x.intercept))+geom_boxplot(aes(fill=rac
e))
p2=ggplot(data=A,aes(x=race,y=coe_smoke))+geom_boxplot(aes(fill=race))
ggarrange(p1,p2,ncol=2,nrow=1,labels = c("A","B"))

##boxplot_gender
epi.bw=gwr.sel(y~smoke*gender,data=epigen,
               coords=cbind(epigen$well,epigen$plate))
epi.gauss=gwr(y~smoke*gender,data=epigen,
              coords=cbind(epigen$well,epigen$plate),bandwidth      =
epi.bw,hatmatrix = T)
results_gauss=as.data.frame(epi.gauss$SDF)


A=data.frame(x.intercept=results_gauss$X.Intercept.,
coe_smoke=results_gauss$smoke+results_gauss$smoke.gender,gender=epige
n$gender)
```

```
A[order(A$gender),]
A$gender=c(rep('boy',279),rep('girl',329))

p1=ggplot(data=A,aes(x=gender,y=x.intercept))+geom_boxplot(aes(fill=gender))
p2=ggplot(data=A,aes(x=gender,y=coe_smoke))+geom_boxplot(aes(fill=gender))

ggarrange(p1,p2,ncol=2,nrow=1,labels = c("A","B"))
```

致谢

在此我们要特别感谢指导老师王老师对我们的指导、支持和鼓励。从最开始的定题，到数据资料的收集，到最终写作定稿，她耐心地帮助我们分析面临的问题，给予了我们无私的帮助。虽然研究的过程是艰难的且充满了挑战，但是她的指导让我们受益匪浅，再次衷心感谢指导老师王老师。