承　诺　书

我们仔细阅读了大学生数学建模竞赛的竞赛规则.

我们完全明白，在竞赛开始后参赛队员不能以任何方式（包括电话、电子邮件、网上咨询等）与队外的任何人（包括指导教师）研究、讨论与赛题有关的问题。

我们知道，抄袭别人的成果是违反竞赛规则的，如果引用别人的成果或其他公开的资料（包括网上查到的资料），必须按照规定的参考文献的表述方式在正文引用处和参考文献中明确列出。

我们郑重承诺，严格遵守竞赛规则，以保证竞赛的公正、公平性。如有违反竞赛规则的行为，我们将受到严肃处理。

我们的题目是：B 题

参赛年级是（一年级，二年级以上，研究生）：二年级以上

所属学院（请填写完整的全名，可填多个）：格拉斯哥学院

参赛队员姓名学号：1.2021190503038

(打印并签名)　　2.2021190905014

3.

指导教师或指导教师组负责人　(有的话打印)：覃思义

是否愿意参加国内赛（是，否）：是

日期：2021 年 05 月 21 日

报名队号（请查阅《2023 校内赛报名队信息-0517》后填写）：H062

2023 电子科技大学数学建模竞赛

编 号 专 用 页

报名队号（请查阅《2023 校内赛报名队信息-0517》后填写）：

评阅记录：

评阅人				
评分				
备注				

基于岭回归，GMM 和贝叶斯层次决策的直肠癌化疗效果研究和手术辅助模型

摘要

根据患者的临床医学数据准确评估接受 nCRT 的局部晚期直肠癌患者淋巴结治疗反应情况对于精准判断是否需要进行 TME 以及在 TME 中是否需要切除淋巴结具有重要意义。本文根据实际接受 TME 手术的患者的临床医学数据，运用岭回归模型结合 K 交叉检验方法设计出根据患者临床医学指标预测 LRG 的模型。并在此基础上建立了基于分层贝叶斯模型和 GMM 聚类模型的决策方案，并使用断点回归设计探究 LRG 指标与 TRG 指标的关联。

在第一个问题中我们采用了基于岭回归和 K 交叉检验的方法利用临床数据建立了预测 LRG 指标的模型。岭回归是一种强健的多因变量预测算法，能够有效抑制过拟合的问题。岭回归算法在标准线性回归的基础上引入了正则化项，可以通过优化带有 L2 范数惩罚的目标函数来学习参数。经过 1000 次迭代，我们找到了最佳的正则化参数 α，并通过 K 交叉验证法确定了模型的最终参数，其中测试集的 R 方为 0.585，证明模型具有良好的预测效果。各项系数分析显示变量与 LRG 之间的线性关系不明显，但是预测的均方误差（MSE）接近于 0，这显示出模型的优秀性能。

问题二结合了 GMM 和分层贝叶斯模型为患者决策是否应该进行 TME 手术以及是否应该切除淋巴结。分层贝叶斯模型适合处理考虑了层次结构的数据，允许我们在数据分析中考虑不同层次的随机变量之间的关系。我们进一步引入了 GMM 聚类模型，其灵活性能处理复杂的数据分布并克服因临床测量数据不准确带来的误差。结果显示，使用 TME 决策模型和 GMM 聚类的策略，可以找出需要进行手术的人群与聚类后的人群有大规模的重合。总结来说，通过结合分层决策和 GMM 聚类，这种方法允许在筛选病人时更加精确和灵活，使手术决策过程更为科学且符合实际需求。

在问题三中，我们需要依据已有的数据集探讨 LRG 与肿瘤消退分级 TRG 的关系，也就是寻求两者间的因果联系。为解决此问题，我们引入断点回归设计（RDD）模型来进行数据分析。断点回归设计（RDD）模型是一种强大的因果推断工具，适用于分析一项干预（例如治疗或政策）是否基于某种阈值或“断点”而产生效果。研究结果显示，对 LRG 和 TRG 的数据进行断点回归设计分析后，我们得到的模型参数中 coef(4) 为 0.4389，这代表了阈值处指标 1 对指标 2 影响的变化程度。因其值较大且统计显著，我们认为 LRG 与 TRG 之间存在明显的因果关系。

最后本文对模型进行了分析总结，综合评价了模型的优势与不足。

关键字：岭回归，高斯混合模型，贝叶斯分层决策，断点回归设计

一、 问题重述

1.1 问题背景

新辅助放化疗（nCRT）和直肠全肠系膜切除术是局部晚期直肠癌（LARC）治疗的主要手段 [3]。淋巴结回归分级（LRG）是基于术后转移淋巴结病理的预后和术前 nCRT 反应的指标。因为 nCRT 后转移淋巴结会导致大量并发症，人们设计了 LRG 系统，主要关注残留癌细胞与退行性纤维化之间的关系，并估计残留癌细胞存在的淋巴结数量。在大多数情况下，nCRT 后的 LRG 与患者的预后相关。而现有计算 LRG 的模型不够成熟，以及医生并不只依赖 LRG 作判断 [2]。提出一个评估 LRG 的算法，大多数关注肿瘤消退的研究和应用都集中在原发肿瘤上，而 LRG 对肿瘤消退和预后的影响尚未得到充分的探讨。基于良好预测和评估回归的 nCRT 治疗有利于个体化临床决策。手术不是所有病人都需要做，因此，建立一个辅助医生判断是否需要做手术的模型是十分必要的，它能提供指导制定更准确的手术或治疗策略 [1]。最后，一个探究 TRG 和 LRG 关系的模型有助于医生以全局的眼光为病人服务 [4]。

1.2 问题分析

1.2.1 问题 1

问题一需要我们我们需要通过临床数据开发一个模型，以精确地预测直肠癌患者在接受新辅助化放疗（neo-adjuvant ChemoRadio Therapy, nCRT）后的淋巴结消退程度（LRG 指标）。这个问题是一个回归分析问题，也就是根据已知数据对预测量进行回归拟合。我们首先需要明确可以用于推测淋巴结消退分级（LRG）的关键因素。由于术前难以获取淋巴结组织标本，因此，我们需要探索其他可能的因素，如临床特征、生化指标等与 LRG 的关联。由于现有数据集的项目较多，为了简化模型，我们还需要对反映相同或相似情况的指标进行合并便于后续分析。接下来，我们需要构建一个 LRG 回归模型。这个模型的目标是预测接受 nCRT 后的淋巴结消退程度。我们可以通过对可以用于推测 LRG 的指标以及给出的 LRG 指标进行回归分析，基于分析结果构建预测模型，这其中模型的准确性和可靠性至关重要。

1.2.2 问题 2

在问题二中我们需要通过我们预测的 LRG 指标结合患者的其它医学数据为患者决策是否应该进行全直肠系膜筋膜切除术（Total Mesorectal Excision，TME）以及是否应该切除淋巴结。为此我们需要基于 LRG 和患者的其它未做手术时的临床医学指标来设

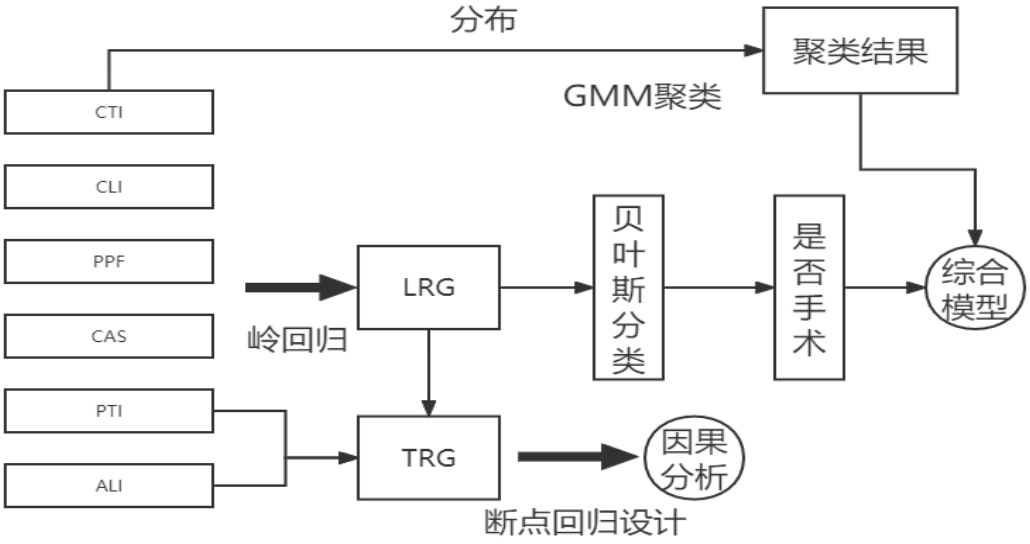

图 1　整体模型

计一个决策模型，这个决策模型应当考虑患者的身体状况是否适合手术，患者化疗效果，患者肿瘤病情是否需要手术以及患者淋巴结是否已经被肿瘤大规模侵略而需要切除。由于决策分为两层进行，也就是先决策是否进行 TME 手术，如果做手术再决策是否切除淋巴结，因此可以使用分层决策结构来处理。

1.2.3 问题 3

在问题三中，我们需要探讨 LRG 与肿瘤消退分级（Tumor Regression Grade，TRG）的关系。我们需要根据这些患者的临床医学数据，对这两个指标的数据进行因果性、相关性分析来推断二者之间是否存在联系。为此我们需要首先从已有的数据中整合反映 LRG 和 TRG 水平的指标，然后对代表两个指标水平的数据进行因果性分析，从而从数据层面探究两者之间是否存在因果联系。我们的整体模型如下

二、 模型假设

1. GMM 假设: 假设误差项服从正态分布，并且假设所有观测的误差项具有相同数量级的方差。
2. 断点回归设计假设：我们假设在阈值处存在一个“跳跃”，即阈值处的行为变化可以归因于阈值的存在。
3. 层次贝叶斯模型假设：我们假设数据具有层次结构，且上一层的参数会影响下一层的参数分布。

三、 符号说明

表 1　符号说明

符号	意义
CTI	术前患者肿瘤病情综合评估指数
CLI	术前患者淋巴病情综合评估指数
PPF	患者身体素质综合指数
CAS	化疗效果评估分数
PTI	术后患者肿瘤病情综合评估指数
ALI	实际淋巴病变指数
LRG	标准化后的 LRG 指标
Q	似然函数
$mathcal\{N\}$	高斯子分布

四、 模型建立

4.1 数据预处理

了便于模型分析，我们接下来将反映相同情况的指标合并并进行标准化：

1. **术前患者肿瘤病情综合评估指数 (CTI):** 该指数反映手术前对患者的肿瘤病情的评估结果，包含术前 CEA 和 cT 两个手术前检测的关于患者肿瘤病情的指标。归一化后该指标为 0-1 之间的数，越大表示术前认为肿瘤情况越严重，可以以下述公式算得：

$$CTI = \frac{CEA + CT}{5} \tag{1}$$

2. **术前患者淋巴病情综合评估指数（CLI）：** 该指数反映手术前对患者的淋巴病情的评估结果，包含 cN 和 cN+ 两个手术前检测的关于患者淋巴结病情的指标。归一化后该指标为 0-1 之间的数，越大表示术前认为淋巴病变情况越严重，可以以下述公式算得：（其中 cN 的 Y 处理成 4）

$$CLI = \frac{CN + CN^{+}}{5} \tag{2}$$

3. **患者身体素质综合指数（PPF）：** 该指数反映手术前对患者身体是否适合做手术的评估结果，包含 ECOG 和年龄两个关于患者身体素质的指标以及距肛距离这个关系到

患者情况是否适合做手术的指标（根据资料显示，患者肿瘤的距肛距离越大，手术操作越简单，术后并发症也会更小，保肛率也更高因此更适合执行手术）。归一化后该指标为 0-1 之间的数，越大表示患者身体越适合做手术，可以以下述公式算得：

$$PPF = [ECOG/10] + 0.5 * [1 - year/100] + 0.5 * [distance/20] \tag{3}$$

公式中 year 代表年龄，distance 代表距肛距离。

4. **化疗效果评估分数（CAS）：**该指数反映对患者的化疗（nCRT）效果的评估结果，包含 mrTRG 和术后 TRG 两个评估的关于患者化疗效果的指标。归一化后该指标为 0-1 之间的数，越大表示患者接受 nCRT 化疗的效果越好，可以以下述公式算得：

$$CAS = 0.5 * (\frac{(5 - mrTRG)}{5} + \frac{afterTRG}{2}) \tag{4}$$

afterTRG 代表术后 TRG

5. **术后患者肿瘤病情综合评估指数（PTI）：**该指数反映手术后对患者的肿瘤病情的评估结果，包含肿瘤分级，肿瘤沉积，微血管浸润，神经侵犯，CRM/R 和 pT 六个手术后检测的关于患者肿瘤病情的指标。归一化后该指标为 0-1 之间的数，越大表示肿瘤情况越严重，可以以下述公式算得：（原指标评级为 0-1 之类的情况取平均值 0.5）：

$$PTI = 0.1 * (a + b + c + d + CRM/R + pT) \tag{5}$$

a 代表肿瘤分级,b 代表肿瘤沉积,c 代表微血管浸润,d 代表神经侵犯

6. **实际淋巴病变指数（ALI）：**该指数反映手术后对患者的淋巴的病变率的评估结果，包含 LNR，阳性淋巴结数，总淋巴结数三个手术后检测的关于患者淋巴结病情的指标。归一化后该指标为 0-1 之间的数，越大表示淋巴病变情况越严重，可以以下述公式算得：

$$ALI = \frac{LNR + PLO/TLO}{2} \tag{6}$$

PLO 代表阳性淋巴结数，TLO 代表总淋巴结数

7. **LRG:** 这个指标结合了 LRGsum 和 LRGmax, 其中 LRGsum 和 LRGmax 都已经标准化。使用下列公式根据 LRGsum 和 LRGmax 进行计算，反映 LRG 的综合水平。

$$LRG = \sqrt{LRGsum^2 + LRGmax^2} \tag{7}$$

最大最小和均值标准化：为了后续算法的效果，我们对以上的综合指标都进行了最大最小和均值标准化，首先，对每一综合指标，先进行最大最小标准化，即

$$index_i = \frac{index_i^j - \min index_i}{\max index_i - \min index_i} \tag{8}$$

$index_i^j$ 表示第 i 个综合指标的第 j 个元素，这个标准化让综合指标的每个元素按照比例映射到 0 到 1 上。然后我们进行了均值，这是为了消除最大最小标准化后出现的 0 对后续算法的影响，这里的均值是最大最小标准化后的均值。

4.2 问题一：基于岭回归的 LRG 预测模型

淋巴结消退分级 (Lymph node Regression Grade，LRG），是基于术后转移淋巴结病理的术后 nCRT 预后和反应的指标。在评估化疗效果方面有重要作用，医学界也常常使用它来辅助判断患者是否需要做手术。由于现有模型的不成熟，加上临床试验上的测量数据有无法避免的偏差，有必要通过其他辅助指标来预测 LRG 或者说通过其他指标建议 LRG 数据的正确性 [5]。

因为 LRG 的值已经被我们定义，这个问题也就转化成有标签的回归问题。在这个问题中，我们首先使用岭回归算法，利用各综合指标进行回归预测，并对结果进行了 R 方检验。岭回归算法是一种针对多因变量预测的有效的回归算法，它在线性回归的基础上引入了正则项，使之具有更强的适应性和鲁棒性。尤其可以缓解过多变量导致的过拟合问题。

首先需要介绍线性回归，线性回归的目标是学习函数

$$\hat{y} = \mathbf{x}^\top \beta + \beta_0 \tag{9}$$

x 是自变量，y 是因变量，β 和 β_0 是系数和截距，它通过完成目标

$$\min_{\beta,\beta_0} \sum_{i=1}^{n} \left(y_i - \mathbf{x}_i^\top \beta - \beta_0\right)^2 \tag{10}$$

来实现学习。它的解是

$$\beta = \left(X^T X\right)^{-1} X^T Y \tag{11}$$

岭回归在线性回归的基础上加入了正则化项，将目标修正为

$$\min_{\beta,\beta_0} \left\{ \sum_{i=1}^{n} \left(y_i - \mathbf{x}_i^\top \beta - \beta_0\right)^2 + \alpha \|\beta\|^2 \right\} \tag{12}$$

其中 $\alpha\|\beta\|^2$ 是正则化项，$\|\beta\|^2$ 是 β 的 L2 范数。我们使用岭回归，通过 6 个综合指标拟合 LRG。系数修正为

$$\beta = (X^T X + \alpha I)^{-1} X^T y \tag{13}$$

在岭回归的基础上，我们又使用了 K 交叉验证法以寻找最佳的 α。这两个方法可以有效缓解过拟合和多特征带来的问题。

K 交叉验证法，一种常用的模型评估方法。它通过将数据集划分为若干个互不重叠的子集，然后利用其中的一部分子集作为训练集进行模型拟合，利用其余子集作为测试集对模型进行验证和评估。可以更充分地利用数据集，也可以通过评估来寻找超参数 α。我们将数据集划分为 k 个大小相等的子集，并依次选取每个子集作为测试集，其余子集作为训练集。模型每次训练的误差可以定义为

$$E_{cv} = \frac{1}{k} \sum_{i=1}^{k} MSE_i \tag{14}$$

MSE 是均方差。

经过 1000 次迭代，我们找到的最佳 alpha 值为：0.5845845845845846，各综合指标的回归系数为: [-0.0308538 -0.06172541 -0.00471115 0.15498751 -0.05459138 1.05896919 -0.00129222],测试集 R 方为:0.5845757357461157,而测试集残差平方和为: 0.04965327371502274。对系数的分析显示变量与 LRG 之间线性关系并不明显，不过 MSE 在测试集上的值接近于 0，显示预测效果很不错。而且综合参数中的实际淋巴病变指数对 LRG 影响最大。

4.3 问题二：结合 GMM 和分层决策的 TME 决策模型

我们在前端使用 TME 决策模型，在后端使用 GMM 聚类，然后调整 GMM 的参数以得到相似的分布。结果显示聚类后的人群和需要做手术的人群有大规模的重合（反之亦然），这表明这个后续对病人的筛选可以先聚类，在使用决策模型进行精确判断。

4.3.1 分层决策模型

我们基于预测的 LRG 指标和患者术前可以得到的其它临床指标设计决策模型，通过这些指标对患者是否需要做手术以及如果要做手术是否需要切除淋巴结两个问题进行决策。本决策模型基于如下指标：（PLN 和 PTN 分别是术前对病变淋巴结数量和病灶区淋巴结总数的预测值，使用“淋巴结初始报告总计”指标的数据）

本决策模型基于如下原则：关于做不做手术，考虑患者肿瘤病情越严重，也就是 CTI 指标越高，则越倾向于做手术；但如果患者身体越差或化疗效果越好，也就是 PPF 或 mrTRG 指标越低，则越倾向于不做手术；关于切不切淋巴，则考虑患者病变淋巴结数量越多，比例越大，则越倾向于切除淋巴结。

我们使用层次贝叶斯模型来设计这个决策模型。层次贝叶斯模型（Hierarchical Bayesian Model）用于建模具有层次结构的数据。它结合了贝叶斯统计学和层次建模的思想，允许我们在数据分析中考虑不同层次的随机变量之间的关系。在层次贝叶斯模型中，数据被组织成多个层次或嵌套的结构。每个层次都可以包含自己的参数和分布。通常，上一层的参数会影响下一层的参数分布，从而建立了不同层次之间的依赖关系。层次贝叶斯模型使用贝叶斯推断方法，通过结合先验知识和观测数据来估计模型参数的后验分布。这种贝叶斯方法允许我们在不同层次的参数估计中引入不确定性，并从数据中学习参数的分布，而不仅仅是获得点估计值。层次贝叶斯模型的能够处理复杂的数据结构和层次关系，适用于分析具有多层次变量的数据，且允许在参数估计中引入不确定性，提供更丰富的信息。层次贝叶斯模型还具备灵活性和可扩展性，可以根据数据的特性和问题的需求进行模型的设计和调整。

在此模型中，我们有两个主要决策：是否进行手术和是否切除淋巴结。因此，我们可以设定两个二元随机变量 S 和 L 表示这两个决策，其中 S=1 表示进行手术，S=0 表示不进行手术，L=1 表示切除淋巴结，L=0 表示不切除淋巴结。

CTI、CLI、PPF、mrTRG、PLN 和 PTN 指标是来自于观测到的数据，我们可以使用它们来更新我们的决策。我们假设这些数据来自某些已知的分布：假设 CTI、CLI、PPF、mrTRG 是从 Beta 分布中抽取的，PLN 和 PTN 是从 Poisson 分布中抽取的。我们的目标是根据观测到的数据来更新我们对 S 和 L 的信念。这可以通过贝叶斯更新完成。

我们将 CTI、PPF 指数结合起来，构成一个关于是否进行手术的决策参数：设定一个随机变量 P_surg，表示进行手术的概率。P_surg 由下列公式计算：

$$P_{surg} = 2 * CTI(i)/(1 + exp(PPF(i) -)) \tag{15}$$

其中 θ 是我们根据历史数据或专家意见设定的阈值。对于是否切除淋巴结的决策，我们同样定义一个随机变量 P_lymph，表示切除淋巴的概率，它受到指标 PLN、PTN 和 CLI 的影响，P_lymph 由下列公式计算：

$$P_{lymph} = (CLI(i) + 1) * ((PLN(i) + 1)/(1 + exp((PLN(i)/PTN(i)) -))) \tag{16}$$

两个计算公式基于逻辑斯蒂回归（Logistic Regression）模型。逻辑回归模型使用了逻辑函数（也被称为 sigmoid 函数），它的输出是一个在 0 到 1 之间的概率值。sigmoid 函数有一个” 饱和” 的特性，即当 PPF 或 PTN 远大于 θ 或 η 时，exp(PPF - θ) 或 exp(PTN - η) 将会非常大，导致 P_surg 或 P_lymph 接近 0。反之，如果 PPF 或 PTN 远小于 θ 或 η，P_surg 或 P_lymph 将接近 CTI 或 CLI*PLN。这样，我们就能在决策过程中充分考虑患者的身体素质和淋巴结的情况。对于每个患者，我们将他们的 P_surg 和 P_lymph 与从均匀分布 [0,1] 中抽取的一个随机数进行比较，如果 P_surg 或者 P_lymph 大于这个随机数，我们就决定进行手术或者切除淋巴结。

4.3.2 GMM 聚类模型

GMM（Gaussian Mixture Model）是一种概率模型，它假设数据是由多个高斯分布的混合生成的。每个高斯分布被称为一个组子分布，有自己的均值和协方差。GMM 是一种软聚类方法，因为它给出的是每个数据点属于每个子分布的概率。GMM 的参数包括每个子分布的均值、协方差和混合权重。混合权重表示每个组件生成数据点的概率。参数通常通过最大化数据的对数似然来估计，通过 EM（Expectation-Maximization）算法来实现。我们选择 GMM 是因为它的灵活性。由于每个子分布都可以有自己的均值和协方差，因此 GMM 可以拟合出非常复杂的数据分布。此外，通过调整混合权重，我们可以控制每个组件对模型的影响。这样可以很好的克服因为临床上测量数据的不准确带来的误差。

GMM 在数学上表示为：

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\mu_k, \Sigma_k) \tag{17}$$

其中，$(\mathbf{x})$ 是数据点，(K) 是组件的数量，π_k 是第 k 个组件的混合权重，$\mathcal{N}(\mathbf{x}|\mu_k, \Sigma_k)$ 是均值为 μ_k、协方差为 Σ_k 的高斯分布。

以下是 GMM 流程：

- 初始化每个组件的均值 μ、协方差 σ 和混合权重 α。
- E 步骤：计算每个数据点属于每个组件的后验概率，并以此计算权重。

$$\omega_{i,k}^t = \frac{\alpha_k^t N\left(x_i \mid \mu_k^t, \sigma_k^t\right)}{\sum_k \alpha_k^t N\left(x_i \mid \mu_k^t, \sigma_k^t\right)} \tag{18}$$

然后计算新的似然函数

$$Q\left(\Theta, \Theta^t\right) = \sum_i \sum_k \omega_{i,k}^t \left(\ln \alpha_k - \ln \omega_{i,k}^t - \ln \sqrt{2\pi\sigma_k^2} - \frac{\left(x_i - \mu_k\right)^2}{2\sigma_k^2} \right) \tag{19}$$

- M 步骤：更新每个组件的均值、协方差和混合权重。更新 σ,μ,α

$$\begin{aligned} &\frac{\partial \sum_i \sum_k \omega_{i,k}^t \left(\ln \alpha_k - \ln \omega_{i,k}^t - \ln \sqrt{2\pi\sigma_k^2} - \frac{(x_i-\mu_k)^2}{2\sigma_k^2} \right)}{\partial \mu_k} = 0 \\ &\sum_i \omega_{i,k}^t \frac{x_i - \mu_k}{\sigma_k^2} = 0 \Rightarrow \sum_i \omega_{i,k}^t \mu_k = \sum_i \omega_{i,k}^t x_i \\ &\Rightarrow \mu_k \sum_i \omega_{i,k}^t = \sum_i \omega_{i,k}^t x_i \end{aligned} \tag{20}$$

- 重复步骤 2 和 3，直到参数收敛。EM 步骤至少保证了似然函数的值一直变大。

聚类结果如下图所示

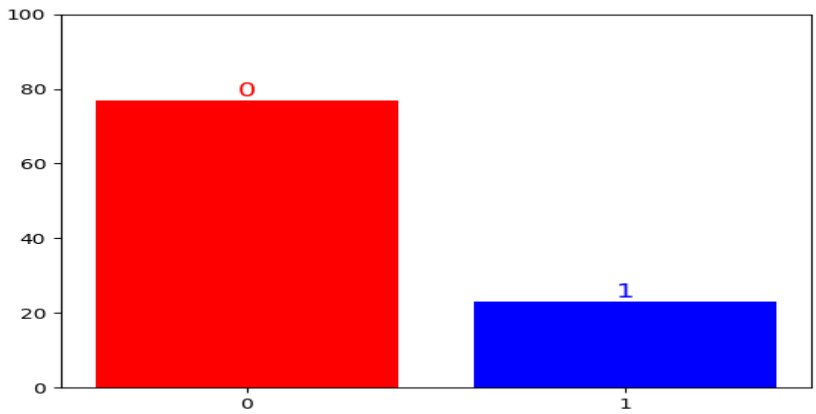

图 2　聚类结果

这里的 0 和 1 只代表不同的类别，不表示是否需要做手术。

4.4 问题三：基于断点回归设计的 LRG 和 TRG 关系模型

针对 LRG 与 TRG 两个指标的分析，我们选取化疗效果评估分数（CAS）作为 TRG 指标的数据，基于 Tanimoto 系数对相关指标的热力图分析证明了两者之间存在联系，而且是正相关。

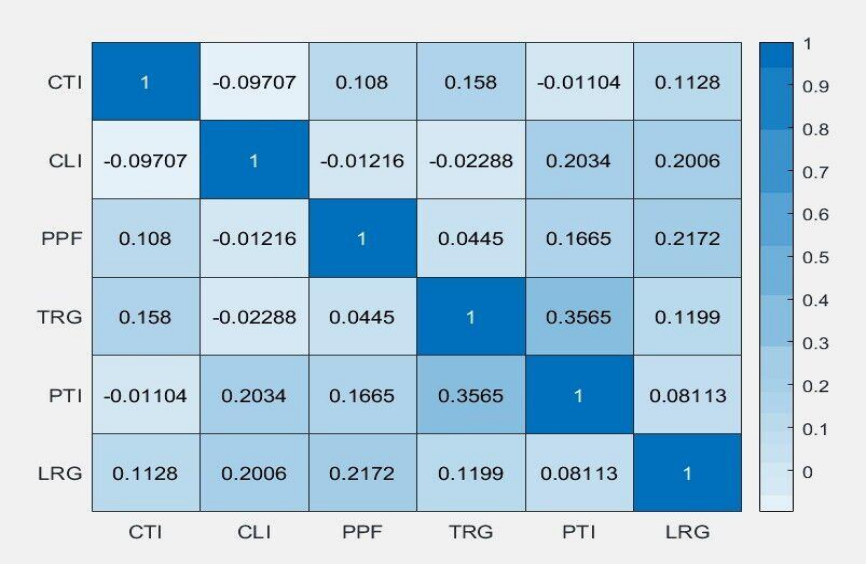

图 3　热力图

随后针对这两组数据我们使用断点回归设计进行因果性分析。

断点回归设计（Regression Discontinuity Design，RDD）是一种用于估计因果效应的统计策略。这种策略适用于当一个干预（例如一个治疗或政策）是基于某种阈值或“断点”决定的情况。RDD 的基本思想是，如果我们可以找到一个阈值或者“断点”，使得只有当一个度量超过这个阈值时，一个特定的干预才会被应用，那么我们就可以比较在阈值附近的个体，看看他们在结果上是否有显著的差异。因为在阈值附近的个体在其他很多特性上可能是非常相似的，所以任何在结果上的显著差异都可能被归因为干预的影响。一个典型的断点回归设计模型的形式可以表示为：

$$Y(i) = \alpha + \beta * Treatment(i) + f(X(i)) + \epsilon(i) \tag{21}$$

其中，Y(i) 是结果变量（指标 2），X(i) 是自变量（指标 1）Treatment(i) 是一个表征是否受到了干预的二值变量；X(i) 是分配到干预的决定变量，也就是那个“断点”；f(X(i)) 是一个关于 X(i) 的函数，表示结果对于 X(i) 的依赖关系，通常会假设在断点两侧的依赖关系可能是不同的，因此可能会分段设定；ϵ 是误差项。β 是我们感兴趣的参数，表示了在阈值处的处理效应。当合适的断点存在，以及阈值附近的个体可以被认为在其他方面是相似的时候，断点回归设计是一种强大的因果推断工具。

断点回归设计的核心思想是，在阈值处的行为变化可以归因于阈值的存在。因此，如果在阈值处观察到指标 2（因变量）的明显变化，那么这通常被解释为指标 1（自变量）对指标 2 有因果影响。进一步地，在 21 式的基础上，我们设计我们应用的断点回

归模型方程表示为：

$$indicator2 = coef(1) + coef(2) * indicator1 + coef(3) * D + coef(4) * indicator1 * D + error \tag{22}$$

这里，D 是一个虚拟变量（dummy variable），表示指标 1 是否大于阈值。indicator1 * D 是交互项，表示在阈值以上和以下，指标 1 对指标 2 的影响是否有差异。coef(1) 是截距项，表示在指标 1 等于 0，且 D 等于 0 时，指标 2 的期望值。coef(2) 是指标 1 的系数，表示在阈值以下，指标 1 每增加一个单位，指标 2 的期望变化量。coef(3) 是虚拟变量 D 的系数，表示在指标 1 等于 0，但 D 等于 1 时，指标 2 的期望值比 D 等于 0 时高出多少。coef(4) 是交互项的系数，表示在阈值以上，指标 1 每增加一个单位，指标 2 的期望变化量比阈值以下时多出多少。

我们关注的主要参数是 coef(4)，它代表了阈值处的" 跳跃" 大小，也就是在阈值处指标 1 对指标 2 影响的变化程度。如果这个数值较大（正或负）并且统计上显著（需要进行显著性检验），那么我们可以认为指标 1 对指标 2 有较强的因果影响。反之，如果这个数值较小或者不显著，那么我们就不能认为存在强烈的因果关系。

通过对 LRG 和 TRG 的数据进行断点回归设计分析，我们得到了断点回归模型方程中的四个参量分别为：coef (1)= 0.7696，coef (2)= -0.0692，coef (3)= -0.2179，coef (4)=0.4389。

式中 coef(1) 是截距项，表示在指标 1 等于 0，且 D 等于 0 时，指标 2 的期望值。在这里，它的值为 0.7696。coef(2) 是指标 1 的系数，表示在阈值以下，指标 1 每增加一个单位，指标 2 的期望变化量。在这里，它的值为-0.0692。coef(3) 是虚拟变量 D 的系数，表示在指标 1 等于 0，但 D 等于 1 时，指标 2 的期望值比 D 等于 0 时高出多少。在这里它的值为-0.2179。coef(4) 则是交互项的系数，表示在超出阈值的时候，指标 1 每增加一个单位，指标 2 的期望变化量比阈值以下时多出多少。在这个里它的值为 0.438。

如果能在阈值处观察到指标 2（因变量）的明显变化，那么可以认为指标 1（自变量）对指标 2 有因果影响。因此我们主要关注的参数是 coef(4)，因为它代表了阈值处的" 跳跃" 大小，也就是在阈值处指标 1 对指标 2 影响的变化程度。在这里 coef(4) 为 0.4389，表示在阈值以上，指标 1 每增加一个单位，指标 2 的期望值将比阈值以下时多增加 0.4389 个单位。结合两个指标均为 0-1 之间的实数，我们认为 coef(4) 相比之下具有较大的数量级，因此我们认为 LRG 与 TRG 之间具有较为明显的相关性。

五、 结果分析

问题一中岭回归的回归系数为 [-0.0308538 -0.06172541 -0.00471115 0.15498751 -0.05459138 1.05896919 -0.00129222]，这显示实际淋巴病变指数（ALI）的大小对 LRG 的影响最大，印证其他文献中 ALI 和 LRG 关系紧密的结论 [6] 。岭回归在测试集上的

表现很好，第一原因是这些病人的 LRG 在理论上服从大致相同的分布，第二原因是约等于 0.5 的 α 使得岭回归具备了一定的泛化能力，而 α 是通过交叉验证法选择的。在问题二中，我们利用问题一的结论，认为所有人的分布大致一致，于是我们使用无标签的 GMM 来学习这种分布，并且构建了层次的贝叶斯决策模型。这个模型通过不同层次的指标分析，最终判断病人是否需要做手术。问题三中，断点回归设计分析的 coef(4) 为 0.4389，相对于指标的数量级来说比较大，可以认为 LRG 和 TRG 之间具有了紧密的联系，人们通过分析一个指标来分析另一个指标成为了可能。

六、 模型总结

6.1 模型优点

- 在解决淋巴结消退分级 (Lymph node Regression Grade，LRG）预测问题中，我们选择使用岭回归算法，利用各综合指标进行回归预测，并对结果进行了 K 交叉检验。岭回归算法是一种针对多因变量预测的有效的回归算法，它在线性回归的基础上引入了正则项，能够防止过拟合，提高模型的泛化能力，使模型具有更强的适应性和鲁棒性。岭回归能够处理多重共线性问题，即当预测变量之间存在高度相关性时，岭回归仍然可以提供稳定的解。它还能够处理特征数量大于样本数量的情况。尤其可以缓解过多变量导致的过拟合问题。我们还选用了 K 交叉验证法来选择模型参数，K 交叉验证法可以充分利用数据，它将数据集分成 K 份，每次使用 K-1 份数据作为训练集，剩下的一份数据作为验证集。这样，我们可以多次训练和验证模型，每份数据都有机会作为验证集被使用。这使得我们可以充分利用所有的数据，提高模型的性能。并且它可以减少过拟合的问题，避免模型在训练数据上过度拟合，提高模型的泛化能力。
- 在基于预测的 LRG 指标和患者术前可以得到的其它临床指标设计的决策模型中，我们使用层次贝叶斯模型来设计这个决策模型。层次贝叶斯模型能够处理具有层次结构的数据，考虑不同层次的随机变量之间的关系，适用于分析具有多层次变量的数据。该模型使用贝叶斯推断方法，通过结合先验知识和观测数据来估计模型参数的后验分布，允许在参数估计中引入不确定性，提供更丰富的信息。并且层次贝叶斯模型具备灵活性和可扩展性，可以根据数据的特性和问题的需求进行模型的设计和调整。其中的指标合并计算公式基于逻辑斯蒂回归（Logistic Regression）模型，使用了逻辑函数（sigmoid 函数），sigmoid 函数有一个” 饱和” 的特性，即在某变量向非定域区间变化时，原指标只会无限趋近于其它原始指标。这样，我们就能在决策过程中充分考虑患者的身体素质和淋巴结的情况。
- 针对 LRG 与 TRG 两个指标关系的分析中，我们使用断点回归设计进行因果性分析。断点回归设计（Regression Discontinuity Design，RDD）是一种用于估计因果效应的

统计策略。断点回归设计是一种强大的因果推断工具，能够在观察研究中估计因果效应。由于断点回归设计只关注那些在阈值附近的个体或群体，这些个体或群体可以被认为是随机选择的，因此它可以解决选择偏差的问题，即被研究的个体或群体并不是随机选择的，而是基于某些特征或条件被选择出来的而导致研究结果的偏差。

6.2 模型缺点

- 在进行 LRG 的预测中，岭回归引入的正则化项会使得模型不能进行特征选择，所有的特征都会在模型中得到保留，这可能会导致模型解释性降低。而且岭回归的正则化参数需要通过交叉验证等方式进行选择，我们就使用了 K 交叉验证来进行参数选择，这增加了模型的复杂性。
- 在决策模型中使用的层次贝叶斯模型的计算复杂性较高，需要引入较为复杂的算法进行参数估计。并且层次贝叶斯模型对先验分布的选择敏感，不合适的先验分布也可能导致模型效果较差。
- 另外，在分析 LRG 和 TRG 的关系时我们使用的断点回归设计的应用需要满足一定的假设，如连续性假设和可操作变量的随机性假设，因此如果数据集出现了变化，那么这些假设在应用中可能就会难以满足。而且该方法主要关注阈值附近的信息，对于阈值较远的数据利用不足，可能导致效率较低。并且断点的选择也可能影响模型的结果，需要根据实际情况和专业知识进行选择。
- 然而，GMM 也有一些缺点。首先，它需要预先设定组件的数量，这在实际应用中可能是一个挑战。其次，GMM 假设数据是由高斯分布生成的，这个假设在某些情况下可能不成立。最后，GMM 的计算复杂性较高，特别是当数据维度较高时。
- 放疗后正常淋巴细胞的变化尚不确定，大多数无反应，部分有纤维化，这使得更难区分正常的 LNs 和完全退化的 LNs，特别是在只有少量转移性肿瘤细胞存在的情况下。很难在手术前评估小纤维化组织病变是正常 LN 还是转移性 LN。区分无纤维化的 LNs 患者和残留的 ypN0 肿瘤患者为完全应答患者和无应答患者从逻辑上也非常困难。

七、 非技术文章

癌症无疑是当今世界公共卫生的最大威胁之一。它不仅带来了病痛与死亡的恐惧，而且其治疗过程常常伴随着身体和精神的极大困扰，影响了全球人类的健康和生活。根据全球肿瘤疾病负担统计，2020 年全球新发癌症病例 1929 万例，其中中国新发癌症 457 万人，因癌症死亡 300 万人。这些令人震惊的数据揭示了癌症对全球健康的严重威胁。

直肠癌作为常见的消化系统恶性肿瘤，在所有癌症中的发病率和死亡率都居于前列。在中国，直肠癌已经成为继肺癌之后的第二大癌症。其患病人群广泛，既包括老年

人，也包括一部分年轻人。直肠癌的发病率与死亡率的提高，反映出癌症防控工作的紧迫性和重要性。在直肠癌的治疗中，尤其是对局部晚期的直肠癌，术前新辅助放化疗（neo-adjuvant ChemoRadio Therapy, nCRT）和全直肠系膜筋膜切除术（TME）的结合已经被证明是有效的治疗方法。随着医学技术的进步，nCRT 后接受 TME 已经成为局部晚期直肠癌的标准治疗模式。nCRT 能够缩小肿瘤的大小，提高手术成功率，而 TME 能够彻底切除肿瘤，降低局部复发率。nCRT 的引入使得治疗方案更加全面和个性化，它的应用不仅极大地改善了局部晚期直肠癌患者的预后，显著降低了局部复发率，同时也提高了低位直肠癌的保肛率。一部分局部晚期直肠癌患者接受 nCRT 后，肿瘤可达到完全消退状态，使非手术治疗成为可能。

然而，尽管我们在癌症治疗方面取得了显著的进步，但仍面临许多挑战。其中一个重要的问题是如何评估患者对 nCRT 的反应，这个问题的解决对于制定个体化治疗方案以及提高治疗效果具有重要意义。目前临床医生在 nCRT 后的 6-8 周左右用肿瘤消退分级（Tumor Regression Grade，TRG）指标来评估原发肿瘤对治疗的反应来判断患者是否需要继续接受手术治疗或选择何种手术方式。但是实际上，盆腔转移淋巴结等指标对患者的疗效与预后的评估也同样具有重要的临床意义。在我们的日常生活中，淋巴系统作为身体的防御系统，负责清除体内的异物和废物，维持体液的平衡。然而，当肿瘤细胞侵入淋巴系统，淋巴结就可能成为癌症转移的通道。这也是为什么，对淋巴结的检查和治疗，在癌症的治疗中占有重要的地位。因此，盆腔转移淋巴结的病理状态直接影响直肠癌患者的治疗策略和预后。接受 nCRT 后，如果淋巴结中全为正常细胞或已全部纤维化（无残留肿瘤细胞）则无需手术切除，如果淋巴结中还含有较高比例的肿瘤细胞，则有较高的转移风险，需要切除。这使得淋巴结的治疗反应评估在治疗决策中占有重要的地位。也就是说，TRG 指标并不能完全反映患者的治疗反应，尤其是对于淋巴结的反应。因为在淋巴结中发现的肿瘤细胞可能是癌症的早期迹象，也可能是癌症转移的标志。然而，由于术前难以获取淋巴结组织标本，目前对于 nCRT 后淋巴结治疗反应的判断是诊疗的难点。

为解决这个问题，我们需要一个更精准的模型来评估淋巴结的治疗反应，即淋巴结消退分级（Lymph node Regression Grade，LRG）。用于评估接受 nCRT 的局部晚期直肠癌患者淋巴结治疗反应情况。LRG 指标的引入对于精准判断是否需要进行 TME，或在 TME 中是否需要切除淋巴结具有重要意义。然而，我们对于 TRG 和 LRG 的理解还远远不够。这两个指标的判断，现在还主要依赖于病理检查，这意味着需要在手术后才能得到结果。如果我们能够在术前就能准确地预测 TRG 和 LRG，那么我们就能更早、更准确地做出治疗决策，更好地利用我们的医疗资源，更有效地改善患者的生存质量和生存期。

总的来说，癌症的防治是一个重要的公共卫生问题，而直肠癌作为癌症中的一员，更是严重威胁着人类的健康和生活。对于直肠癌的治疗，nCRT 和 TME 的结合已经取

得了显著的效果。然而，如何更准确地评估患者对治疗的反应，以及如何根据这个反应进行更个性化的治疗决策，仍然是我们面临的挑战。我们进行的研究就是希望建立一个能够准确预测 LRG 的模型，并在大规模的临床数据中验证这个模型的准确性和可靠性，希望通过建立 LRG 模型和分析 LRG 和 TRG 的关系，为解决这个问题提供新的思路和方法。

参考文献

[1] 赵小立. 直肠癌根治性切除术前新辅助化疗的疗效观察 [J]. 中国实用医药,2014(22):52-53.

[2] 申朔豪. 直肠癌新辅助放化疗后病理分期对预后的影响及应用研究 [D]. 中国医学科学院,2021.

[3] 袁贵前. 直肠癌根治性前切除术后常见并发症与术前营养状况的探讨 [D]. 广西: 广西医科大学,2018.

[4] He L, Xiao J, Zheng P, Zhong L, Peng Q. Lymph node regression grading of locally advanced rectal cancer treated with neoadjuvant chemoradiotherapy. World J Gastrointest Oncol. 2022 Aug 15;14(8):1429-1445.

[5] 王燕. 应用时间序列分析 [M]. 北京：中国人民大学出版社 2005.

[6] Lee HG, Kim SJ, Park IJ, Hong SM, Lim SB, Lee JB, Yu CS, Kim JC. Effect of Responsiveness of Lymph Nodes to Preoperative Chemoradiotherapy in Patients With Rectal Cancer on Prognosis After Radical Resection. Clin Colorectal Cancer. 2019 Jun;18(2):e191-e199.

附录 A　代码

```
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.mixture import GaussianMixture
from sklearn.cluster import DBSCAN

#读取csv文件前7列
data = pd.read_csv('C:\\Users\\zzy\\Desktop\\testdata7.csv', header=None,
    usecols=[0,1,2,3,4,5],skiprows=1)

#使用GMM进行聚类
gmm =
    GaussianMixture(n_components=2,tol=1e-6,max_iter=1000,init_params='kmeans',random_state=1)
```

```
labels_gmm = gmm.fit_predict(data)

#使用DBSCAN进行聚类
dbscan = DBSCAN(eps=0.5, min_samples=5)
labels_dbscan = dbscan.fit_predict(data)

#将标签写入csv文件，并保留表头
header_names = ['col1', 'col2', 'col3', 'col4', 'col5', 'col6']
data.columns = header_names
data['Label_GMM'] = labels_gmm
data['Label_DBSCAN'] = labels_dbscan
data.to_csv('C:\\Users\\zzy\\Desktop\\testdata7.csv', index=False, header=True)
```

```
import pandas as pd
import numpy as np

# 读取数据
data = pd.read_csv('C:\\Users\\zzy\\Desktop\\testdata7.csv')
A = data.iloc[:, 1]
B = data.iloc[:, 2]
C = data.iloc[:, 3]
D = data.iloc[:, 4]
E = data.iloc[:, 5]
F = data.iloc[:, 6]

# 初始化决策变量
S = np.zeros(len(A))
L = np.zeros(len(A))

# 决策阈值（可以调整）
theta = 0.2 # Adjust based on your domain knowledge
eta = 0.1 # Adjust based on your domain knowledge

# 遍历每个患者，做出决策
for i in range(len(A)):
    # 计算决策参数
    P_surg = 2 * A[i] / (1 + np.exp(C[i] - theta))
    P_lymph = (B[i] + 1) * ((E[i] + 1) / (1 + np.exp((E[i] / F[i]) - eta)))

    # 根据决策参数和随机数进行决策
    if P_surg > np.random.rand():
        S[i] = 1
    if S[i] == 1 and P_lymph > np.random.rand():
        L[i] = 1
```

```
# Save decision variables S and L into a new DataFrame
decision_df = pd.DataFrame({'S': S, 'L': L})

# Concatenate the original data with the decision variables
result_df = pd.concat([data, decision_df], axis=1)

# Save the result to a new CSV file
result_df.to_csv('C:\\Users\\zzy\\Desktop\\result.csv', index=False)



import pandas as pd
from sklearn.model_selection import cross_val_score, train_test_split
from sklearn.linear_model import Ridge
from sklearn.metrics import r2_score, mean_squared_error
import numpy as np
import matplotlib.pyplot as plt

# 读取csv文件，前6列作为自变量，第11列作为因变量
df = pd.read_csv("C:\\Users\\zzy\\Desktop\\testdata8.csv", usecols=[0,1,2,3,4,5,6,9],
    skiprows=0)

# 将自变量和因变量分离
X = df.iloc[:, :-1]
y = df.iloc[:, -1]

# 划分训练集和测试集
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)

# 创建一个岭回归模型对象
model = Ridge()

# 设置一组备选的alpha值
alphas = np.linspace(0, 1, 1000)

# 使用交叉验证选择最佳的alpha值
scores = []
for alpha in alphas:
    model.set_params(alpha=alpha)
    this_scores = cross_val_score(model, X_train, y_train, cv=10)
    scores.append(this_scores.mean())

# 找到最佳alpha值
best_alpha = alphas[scores.index(max(scores))]
print("最佳alpha值为：", best_alpha)

# 使用最佳alpha值对数据进行拟合
model.set_params(alpha=best_alpha)
```

```
model.fit(X_train, y_train)



import pandas as pd
from sklearn.preprocessing import MinMaxScaler

#读取CSV文件
df = pd.read_csv('C:\\Users\\zzy\\Desktop\\testdata7.csv')

#所有列数据
data = df.iloc[:, :]

#最大最小标准化处理
scaler = MinMaxScaler()
data_norm = scaler.fit_transform(data)

#获取每一列的均值
means = data.mean()

#将标准化后的结果加上自身标准化之前的均值
data_norm = data_norm + means

#将结果写回到CSV文件的所有列
df.iloc[:, :] = data_norm

#保存修改后的CSV文件
df.to_csv('C:\\Users\\zzy\\Desktop\\testdata7.csv', index=False)
```

```
    data = csvread('D:\Desktop\决策数据备份.csv',1,1); % Assume the first row is header
A = data(:,2);
B = data(:,3);
C = data(:,4);
D = data(:,5);
E = data(:,6);
F = data(:,7);

% 初始化决策变量
S = zeros(length(A),1);
L = zeros(length(A),1);

% 决策阈值（可以调整）
theta = 0.2; % Adjust based on your domain knowledge
eta = 0.1; % Adjust based on your domain knowledge

% 遍历每个患者，做出决策
for i = 1:length(A)
```

```
    % 计算决策参数
    P_surg = 2*A(i) / (1 + exp(C(i) - theta));
    P_lymph = (B(i) +1)* ((E(i)+1) / (1 + exp((E(i)/F(i)) - eta)));

    % 根据决策参数和随机数进行决策
    if P_surg > rand()
        S(i) = 1;
    end
    if S(i) == 1 && P_lymph > rand()
        L(i) = 1;
    end
end




data = csvread('D:\Desktop\LRG_VS_TRG_DATA.csv', 0, 0); % 假设数据从第二行开始，第一行是表头

% 指标1和指标2
indicator1 = data(:,1);
indicator2 = data(:,2);

% 假设阈值在指标1的中位数
threshold = median(indicator1);

% 创建一个指示变量，表示是否超过阈值
over_threshold = indicator1 > threshold;

% 为了拟合断点回归模型，我们使用多元线性回归函数，输入变量包括指标1、指示变量和他们的交互项
X = [indicator1, over_threshold, indicator1.*over_threshold];
Y = indicator2;

% 拟合模型
mdl = fitlm(X, Y);

% 输出模型系数
coef = mdl.Coefficients.Estimate

%
    其中，coef(1)是截距，coef(2)是指标1对指标2的平均影响（在阈值以下），coef(3)是阈值的效应，coef(4)是阈值以上时，
```