Advances in Applied Mathematics 应用数学进展, 2024, 13(8), 4021-4033
Published Online August 2024 in Hans. https://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2024.138383

Hans汉斯

基于面板数据的分组多折点回归模型估计

王　昊

西南大学数学与统计学院，重庆

收稿日期：2024年7月21日；录用日期：2024年8月13日；发布日期：2024年8月23日

摘　要

折点回归模型是指响应变量与某个协变量之间存在连续的分段线性关系，本文基于面板数据，研究了个体间具有群组结构的多折点回归模型。首先，建立一种基于贪心策略的坐标下降法用于预估折点位置，用较小的计算代价解决了折点估计量对初值敏感的问题，并使用信息准则选择合适的折点个数。然后，基于该折点预估算法的框架下，使用最大最小距离法选择初始聚类中心，用于K-means类型的算法去优化各组的模型参数，分组的个数由自动化手肘法确定。数值模拟和实证分析显示，该方法可得到良好的参数估计和群组结构估计，并且在真实的女性黄体酮数据中具有实际意义。

关键词

面板数据，多折点回归，群组结构，坐标下降法

Estimation of Grouped Multi-Kink Regression Model Based on Panel Data

Hao Wang

School of Mathematics and Statistics, Southwest University, Chongqing

Received: Jul. 21st, 2024; accepted: Aug. 13th, 2024; published: Aug. 23rd, 2024

Abstract

A kink regression model refers to a model where the response variable has a continuous piecewise linear relationship with a covariate. This paper studies a multi-kink regression model with grouped structure among individuals based on panel data. First, a coordinate descent method based on a greedy strategy is established to estimate the kink locations, addressing the issue of sensitivity to initial values in kink estimation with minimal computational cost. An information criterion is used to select the appropriate number of kinks. Then, within the framework of this kink estimation algorithm, the max-min distance method is used to select the initial clustering

文章引用: 王昊. 基于面板数据的分组多折点回归模型估计[J]. 应用数学进展, 2024, 13(8): 4021-4033.
DOI: 10.12677/aam.2024.138383

centers for a K-means type algorithm to optimize the model parameters for each group. The number of groups is determined using an automated elbow method. Numerical simulations and empirical analysis show that this method can achieve good parameter estimation and grouped structure estimation. Moreover, the grouped structure and within-group parameters have analytical value in the real-world data of female progesterone levels.

Keywords

Panel Data, Multi-Kink Regression, Grouped Structure, Coordinate Descent Method

Copyright © 2024 by author(s) and Hans Publishers Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
http://creativecommons.org/licenses/by/4.0/
Open Access

1. 引言

在回归模型中，线性回归模型作为一种基础且应用广泛的统计方法，被广大研究者与实践者所熟知和使用。然而，传统的线性回归模型在面对一些非线性的数据关系时，可能无法达到理想的拟合效果。而折点回归模型从一定程度上放松线性假定，假设变量间呈分段线性关系，在生物、经济等应用领域有着重要应用。

文献[1]首次提出单折点回归模型，用于研究响应变量与协变量之间连续的分段线性关系，与此同时文献[2]提出基于累积和的方法构造了关于单折点的假设检验方法。随后，折点回归模型在不同场景下都有所应用，比如文献[3]分析了哺乳动物的最大奔跑速度和身体质量之间的关系；文献[4]结合逻辑回归研究了病毒免疫测试数据的折点回归模型；文献[5]研究时间序列数据的单折点推断并用于分析增长与债务问题。也有一些文献研究单折点模型的稳健估计量，比如文献[6]和文献[7]分别基于分位数损失和秩的估计量。对于当折点的个数是未知时，文献[8]首次将单折点回归模型推广到多折点情形，并在分位数损失框架下研究了估计量的渐近性质和折点存在性检验。而文献[9]提出使用自助法为折点构建置信区间。

但由于折点的存在，这使得求解模型估计量时的优化问题变得比较棘手。一方面，导致损失函数在该处不可导，许多基于一阶导乃至二阶导的优化算法均变得不可行；另一方面，折点问题对应的损失函数都是非凸函数，一些局部极值点的存在使得优化结果十分依赖初值。现有的折点估计方法，基本上分为四种类型：第一种是网格搜索法，构造仅关于折点的条件损失函数，再通过网格搜索寻找条件损失函数的最值点，比如文献[5]；第二种是基于泰勒展开的局部线性近似法，拆开非线性项，通过迭代的方式更新寻找折点位置，比如文献[10]；第三种是基于马尔可夫链蒙特卡洛方法，比如文献[11]，但是该方法即使在简单的模型中也有很大的计算负担；第四种是使用平滑过渡核函数去替代不可导的函数，然后使用基于导数的优化方法，比如文献[12]。在文献[12]中系统地比较了单折点回归模型的这四种估计方法，除了马尔可夫链蒙特卡洛方法计算效果较差外，另外三种方法的估计效果都彼此接近。而当折点个数较多时，估计方法的选择将愈发困难。随着折点个数增加，局部线性近似法和平滑过渡核函数都面临优化结果对初值敏感的问题，网格搜索法和马尔可夫链蒙特卡洛方法都面临成倍的计算成本问题。

随着数据类型的复杂化，已有一些文献着力于研究基于面板数据的折点回归模型，比如文献[13] [14]，拓展了一般的折点问题，研究面板数据单折点和协变量具有相关性的模型。还有一些文献基于纵向数据背景下应用折点回归模型，比如文献[15]研究得到女性黄体酮数据的两个关键折点，以及文献[16]研究阿兹海默高风险中年人的认知与年龄关系。

而在面板数据背景下，个体间的同质性是一个十分具有意义的话题，利用这种同质性还能增加参数估计的效率，当数据的时间维度不大时，利用个体同质性能大大增加折点估计的精确程度。文献[17]采用二项分割算法，分别对几种不同类别的参数聚类，充分利用样本间的同质性信息去提高估计量的渐近效率，但是由于不同类别的参数之间的分组结构不同，因而缺乏一定的直观性。本文将从另一种聚类角度出发，从新的视角构建面板数据的分组多折点回归模型，即采用对个体聚类的角度，在保持模型解释力的基础上，拓展折点回归模型在面板数据应用中的灵活性。

本文剩余部分的结构安排如下：第二节主要介绍面板数据时单组和多组的多折点回归模型及其估计；第三节通过数值模拟实验展示了模型估计量的有限样本表现；第四节将该模型用于分析真实数据；第五节是总结部分。

2. 模型及估计

2.1. 面板数据的多折点回归模型的参数估计

基于面板数据，这一小节讨论的是将一部分个体视为同一组时给出相应的参数优化方法。当一些个体真实参数不相同时，如果依然将这些混杂的个体视为同一组，那么现有的折点优化算法均无法在低计算成本的同时快速收敛，这一小节给出一种特殊的坐标下降法解决这一问题。假设个体数量有 N 个，其中某 m 个属于同一组，现在需给出这一组的参数优化方法。

假设有界门限变量 x_{it} 与响应变量 y_{it} 有连续的分段线性关系，而 z_{it} 是一个 p 维协变量，其中 $i=1,\cdots,N$ 和 $t=1,\cdots,T_i$。出于记号上的简便，不妨假定数据是平衡面板数据，即对于任意 $i=1,\cdots,N$ 都有 $T_i=T$，对于非平衡面板数据，本文的模型及其算法依然可行。现假定数据来源于有个体固定效应的多折点回归模型，所有个体共享相同的系数参数和 K 个折点参数，即

$$y_{it}=\mu_i+\alpha_0 x_{it}+\sum_{k=1}^{K}\alpha_k\left(x_{it}-\delta_k\right)_+ +\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{z}_{it}+\varepsilon_{it}, \tag{1}$$

其中函数 $(a)_+=\max\{a,0\}$，折点位置参数 δ_k，$k=1,\cdots,K$ 互不相同，μ_i 为个体固定效应，门限变量的系数参数满足 $\alpha_k\neq 0$，$k=1,\cdots,K$，协变量的系数参数 $\boldsymbol{\gamma}$ 是一个 p 维向量，ε_{it} 为扰动项。

本文使用平方损失函数估计模型(1)的参数。若记折点参数 $\boldsymbol{\delta}=(\delta_1,\cdots,\delta_K)^{\mathrm{T}}$ 和依赖于折点参数的变量 $\tilde{\mathbf{x}}_{it}^{\mathrm{T}}(\boldsymbol{\delta})=\left(x_{it},(x_{it}-\delta_1)_+,\cdots,(x_{it}-\delta_K)_+,\mathbf{z}_{it}^{\mathrm{T}}\right)$，并记相应的系数参数 $\boldsymbol{\theta}^{\mathrm{T}}=\left(\alpha_0,\alpha_1,\cdots,\alpha_K,\boldsymbol{\gamma}^{\mathrm{T}}\right)$，那么模型(1)的非线性最小二乘估计量 $\hat{\boldsymbol{\delta}}$，$\hat{\boldsymbol{\theta}}$ 和 $\hat{\boldsymbol{\mu}}$ 是使得损失函数

$$Q(\boldsymbol{\delta},\boldsymbol{\mu},\boldsymbol{\theta})=\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(y_{it}-\mu_i-\boldsymbol{\theta}^{\mathrm{T}}\tilde{x}_{it}(\boldsymbol{\delta})\right)^2 \tag{2}$$

达到最小值时相应的参数。关于 $Q(\boldsymbol{\delta},\boldsymbol{\mu},\boldsymbol{\theta})$ 的优化问题涉及两个关键点。其一，损失函数的优化是一个非线性最小二乘问题，事实上其参数 $\boldsymbol{\delta}$，$\boldsymbol{\mu}$ 和 $\boldsymbol{\theta}$ 的优化结果非常依赖初值的选择，如果先固定 $\boldsymbol{\delta}$，得到对应的估计量 $\hat{\boldsymbol{\mu}}(\boldsymbol{\delta})$ 和 $\hat{\boldsymbol{\theta}}(\boldsymbol{\delta})$，这时损失函数可以视为仅由 $\boldsymbol{\delta}$ 决定的函数，最后可通过搜索 $\boldsymbol{\delta}$ 得到最终的估计量 $\hat{\boldsymbol{\delta}}$，$\hat{\boldsymbol{\theta}}(\hat{\boldsymbol{\delta}})$ 和 $\hat{\boldsymbol{\mu}}(\hat{\boldsymbol{\delta}})$。当折点个数过多时即 K 过多时，难以平衡优化的速度和精度，本文使用一种特殊的坐标下降法结合网格搜索用于预估折点位置，在下一小节介绍聚类算法将个体分组之后，最后再来使用无导数优化算法提高 $\boldsymbol{\delta}$ 的估计精度。其二是折点参数 $\boldsymbol{\delta}$ 的维数 K 通常是未知的，所以 $\boldsymbol{\theta}$ 和 $\tilde{x}_{it}(\boldsymbol{\delta})$ 的维度是未知的，这里我们暂时假定折点个数 K 是已知的，在本小节结束时将使用信息准则选择最合适的 K。

针对问题的第一点，当给定折点个数 K 时，先给出损失函数(2)的优化方法。记

$$L(\boldsymbol{\delta})=\min_{\boldsymbol{\mu}\in\mathbb{R}^N,\boldsymbol{\theta}\in\mathbb{R}^{1+K+p}} Q(\boldsymbol{\delta},\boldsymbol{\mu},\boldsymbol{\theta}),$$

其中 $\mathbb{R}^a$ 表示 a 维向量空间，所以对于给定的 $\boldsymbol{\delta}$，可以先优化 $\boldsymbol{\mu}$ 和 $\boldsymbol{\theta}$ 得到对应损失函数。这一步的优化问题是常规的线性最小二乘问题，可以直接给出显式解

$$\left(\hat{\boldsymbol{\mu}}^{\mathrm{T}}(\boldsymbol{\delta}),\hat{\boldsymbol{\theta}}^{\mathrm{T}}(\boldsymbol{\delta})\right)^{\mathrm{T}}=\left(\boldsymbol{\chi}^{\mathrm{T}}(\boldsymbol{\delta})\boldsymbol{\chi}(\boldsymbol{\delta})\right)^{-1}\boldsymbol{\chi}(\boldsymbol{\delta})\boldsymbol{Y},$$

其中 $\boldsymbol{\chi}(\boldsymbol{\delta})=\left(\boldsymbol{D};\boldsymbol{X}(\boldsymbol{\delta})\right)$，$\boldsymbol{D}=\boldsymbol{I}_N\otimes\boldsymbol{e}_T$，$\boldsymbol{I}_N$ 表示 N 阶单位矩阵，$\boldsymbol{e}_T$ 为元素全为 1 长度为 T 的向量，而 $\boldsymbol{X}(\boldsymbol{\delta})$ 和 $\boldsymbol{Y}$ 分别是所有 $\tilde{\boldsymbol{x}}_{it}(\boldsymbol{\delta})$ 和 y_{it}，$1\le i\le N$，$1\le t\le T$ 按照 (i,t) 字典排序后竖向堆叠形成的矩阵和向量。由此外，使用如牛顿迭代等方法也能得到优化结果。

现在优化 $Q(\boldsymbol{\delta},\boldsymbol{\mu},\boldsymbol{\theta})$ 问题化为寻找

$$\hat{\boldsymbol{\delta}}=\arg\min_{\boldsymbol{\delta}\in\mathbb{D}^K}L(\boldsymbol{\delta}),\tag{3}$$

其中 $\mathbb{D}$ 是一个闭区间，其范围取决于有界门限变量 x_{it}，一般可假定区间下界和上界取为 x_{it} 观测值的 5% 和 95%分位数。由于目标函数 $L(\boldsymbol{\delta})$ 非凸且不可微，所以一些文献采用对 $(a)_+$ 函数局部线性近似的方法迭代得到优化结果，但由于该方法对于初值的选取比较敏感，导致其易于收敛到局部最优值甚至不收敛，当局部线性近似方法应用在分组多折点回归模型的优化时，由于不同个体的折点参数或系数不一样，收敛问题将会更加严重。这里使用一种基于坐标下降法的网格搜索方法，通过计算门限变量 x_{it} 观测值的一些分位点，得到候选折点向量 $A=(a_1,\cdots,a_s)^{\mathrm{T}}$，进而(3)的优化问题变成

$$\hat{\boldsymbol{\delta}}=\arg\min_{\boldsymbol{\delta}\in A^K}L(\boldsymbol{\delta}),\tag{4}$$

该优化结果即是折点参数的预估值。下面给出一种基于贪心策略的坐标下降法，用于优化(4)。因为 $\boldsymbol{\delta}$ 是 K 维向量，当使用信息准则选取最合适的 K 时，需要重复计算不同的 K 值，需要提供一种算法最好能在估计 K 维时的过程中顺便估计 $0,\cdots,K-1$ 维时的结果以节约计算时间。所以先假定 $\boldsymbol{\delta}$ 是 0 维的，每升一维时就立即找到当前维度的最佳估计。算法 1 给出了当 $\boldsymbol{\delta}$ 在升维时如何从候选折点 $a_1,\cdots,a_s$ 中找到目前最适合添加的折点的方法，这种算法便是直接找到使损失下降最快的候选折点 a_v，然后返回添加了该折点的新向量 $\check{\boldsymbol{\delta}}_v$。

算法 1：添加折点算法 $\mathrm{Add}(\boldsymbol{\delta},A)$

输入当前已有折点 $\boldsymbol{\delta}=(d_1,\cdots,d_r)^{\mathrm{T}}$，候选折点 $A=(a_1,\cdots,a_s)^{\mathrm{T}}$

for n in $[1,\cdots,s]$:

令 $\check{\boldsymbol{\delta}}_n\leftarrow(d_1,\cdots,d_r,a_n)^{\mathrm{T}}$

计算相应的损失大小 $L_n\leftarrow L(\check{\boldsymbol{\delta}}_n)$

找到 L_n 中最小值所在的索引 v，可知往 $\boldsymbol{\delta}$ 中添加 a_v 为最优选择

输出 $\check{\boldsymbol{\delta}}_v$

当 $\boldsymbol{\delta}$ 添加了新的一个折点之后，即升维之后，原来的折点需要更新，算法 2 给出替换折点算法，在所有折点中找到最需要被更新的折点，更新后返回替换某折点后的新向量 $\check{\boldsymbol{\delta}}_{u,v}$。

基于算法 1 和算法 2，算法 3 给出了损失函数(2)优化的整个过程。首先假定当前的折点为空向量，然后逐步升维，每次升维时调用算法 1，并在当前维度循环调用算法 2 替换折点直到折点向量不再变化，随后继续升维，依此类推，直到 $\boldsymbol{\delta}$ 升到 K 维并且不再发生变化后即可结束循环，最后输出每个维度下最合适的折点向量 $\boldsymbol{\delta}_0,\cdots,\boldsymbol{\delta}_K$。

算法 2：替换折点算法 $\mathrm{Rep}(\boldsymbol{\delta}, A)$

输入当前已有折点 $\boldsymbol{\delta}=(d_1,\cdots,d_r)^{\mathrm{T}}$，候选折点 $A=(a_1,\cdots,a_s)^{\mathrm{T}}$

for m in $[1,\cdots,r]$：

for n in $[1,\cdots,s]$：

令 $\breve{\boldsymbol{\delta}}_{m,n} \leftarrow (d_1,\cdots,d_{m-1},d_{m+1},\cdots,d_r,a_n)^{\mathrm{T}}$

计算损失大小 $L_{m,n} \leftarrow L\left(\breve{\boldsymbol{\delta}}_{m,n}\right)$

找到 $L_{m,n}$ 中最小值所在的索引 u,v，可知将 $\boldsymbol{\delta}$ 中的 b_u 替换为 a_v 为最优选择

输出 $\breve{\boldsymbol{\delta}}_{u,v}$

算法 3：折点预估算法 $\mathrm{Kink}(A,K)$

输入候选折点 $A=(a_1,\cdots,a_s)^{\mathrm{T}}$，指定折点个数 K

初始化当前的折点 $\boldsymbol{\delta}_0=\varnothing$

for k in $[1,\cdots,K]$：

令 $m \leftarrow 0$

计算 $\boldsymbol{\delta}_k^{(0)} \leftarrow \mathrm{Add}(\boldsymbol{\delta}_{k-1}, A)$，

#注释：其中 $\boldsymbol{\delta}$ 的下标表示当前维数，上标表示在当前维数时的迭代轮数

while True：

计算 $\boldsymbol{\delta}_{k+1}^{(m+1)} \leftarrow \mathrm{Rep}\left(\boldsymbol{\delta}_{k+1}^{(m)}, A\right)$

如果 $\boldsymbol{\delta}_{k+1}^{(m+1)}=\boldsymbol{\delta}_{k+1}^{(m)}$，那么记 $\boldsymbol{\delta}_{k+1} \leftarrow \boldsymbol{\delta}_{k+1}^{(m+1)}$ 并提前跳出当前的循环

否则令 $m \leftarrow m+1$

输出 $\boldsymbol{\delta}_0,\cdots,\boldsymbol{\delta}_K$

这个算法的本质是一种基于贪心策略的坐标下降算法，并且在坐标下降时使用的线搜索方法是网格搜索，但核心步骤是算法 2，一般的坐标下降会轮流更新每个参数，而算法 2 实际上只会优先更新能使损失值下降的最低的折点。至于精确寻找折点 $\boldsymbol{\delta}$，算法 3 得到的预估值作为初值，使用无梯度优化算法，比如 Nelder-Mead 算法，即可在给定精度下，迭代求得更精确折点位置，避免了初值敏感问题。

当给出指定 K 时模型(1)的最小二乘估计方法后，即可由贝叶斯信息准则确定组内的折点个数，通常假定折点的个数并不会很多，故设定最大折点个数为 5，并对于每个 $K=1,\cdots,5$ 计算相应的 BIC 值

$$\mathrm{BIC}(K)=\log\left(Q\left(\hat{\boldsymbol{\delta}},\hat{\boldsymbol{\mu}},\hat{\boldsymbol{\theta}}\right)\right)+P_K\frac{\log^2(N)}{2N},$$

其中 P_K 表示总共的未知参数个数，即 $\boldsymbol{\delta}$，$\boldsymbol{\mu}$ 与 $\boldsymbol{\theta}$ 的维数之和，最终选择使 BIC 值达到最小的 K 作为折点个数估计值 $\hat{K}$。

2.2. 分组多折点回归模型及其参数估计

现在考虑分组多折点回归模型，在上一小节讨论过面板数据的多折点回归模型的参数估计之后，其解决的问题即为每一组数据的参数估计。继续讨论面板数据 $\{y_{it},x_{it},\boldsymbol{z}_{it}\}$，$i=1,\cdots,N$，$t=1,\cdots,T$，为了记号上的简便，这里还是假定每个个体的观测次数一致，对于非平衡面板数据依然可行。现在引入群组结构，假设个体划分为 G 个不同的群组，对于任意个体 i，其所属的组别是 $g_i \in \{1,\cdots,G\}$，在同一组内的个

体共享除固定效应外的其它参数，即分组折点回归模型表示为

$$y_{it}=\mu_i+\alpha_{g_i0}x_{it}+\sum_{k=1}^{K_{g_i}}\alpha_{g_ik}\left(x_{it}-\delta_{g_ik}\right)_+ +\boldsymbol{\gamma}_{g_i}^{\mathrm{T}}\boldsymbol{z}_{it}+\varepsilon_{it}, \tag{5}$$

其中 g 组共享的参数包含折点个数 K_g，折点位置 $\boldsymbol{\delta}_g=\left(\delta_{g1},\cdots,\delta_{gK_g}\right)^{\mathrm{T}}$，门限变量系数 $\boldsymbol{\alpha}_g=\left(\alpha_{g0},\cdots,\alpha_{gK_g}\right)^{\mathrm{T}}$ 和协变量系数 $\boldsymbol{\gamma}_g$。

模型的损失函数定义为

$$Q(\boldsymbol{\delta},\boldsymbol{\mu},\boldsymbol{\theta},\mathbb{G})=\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(y_{it}-\mu_i-\alpha_{g_i0}x_{it}-\sum_{k=1}^{K_{g_i}}\alpha_{g_ik}\left(x_{it}-\delta_{g_ik}\right)_+ -\gamma_{g_i}^{\mathrm{T}}z_{it}\right)^2, \tag{6}$$

其中折点位置参数 $\boldsymbol{\delta}^{\mathrm{T}}=\left(\boldsymbol{\delta}_1^{\mathrm{T}},\cdots,\boldsymbol{\delta}_G^{\mathrm{T}}\right)$，固定效应参数 $\boldsymbol{\mu}^{\mathrm{T}}=\left(\mu_1,\cdots,\mu_N\right)$，系数参数 $\boldsymbol{\theta}^{\mathrm{T}}=\left(\boldsymbol{\theta}_1^{\mathrm{T}},\cdots,\boldsymbol{\theta}_G^{\mathrm{T}}\right)$，这里 $\boldsymbol{\theta}_g^{\mathrm{T}}=\left(\boldsymbol{\alpha}_g^{\mathrm{T}},\boldsymbol{\gamma}_g^{\mathrm{T}}\right)$ 且 $\boldsymbol{\alpha}_g^{\mathrm{T}}=\left(\alpha_{g0},\cdots,\alpha_{gK_g}\right)$，群组结构参数 $\mathbb{G}=\left\{\mathcal{N}_1,\cdots,\mathcal{N}_G\right\}$，这里 $\mathcal{N}_g=\left\{i\middle|g_i=g,i=1,\cdots,N\right\}$ 表示第 g 组所含的全部个体集合。在优化(6)求解参数估计量时，先假定群组个数 G 已知，后续再来讨论 G 的选择方法。

将损失函数(6)改写为更加紧凑的形式

$$Q(\boldsymbol{\delta},\boldsymbol{\mu},\boldsymbol{\theta},\mathbb{G})=\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(y_{it}-\mu_i-\boldsymbol{\theta}_{g_i}^{\mathrm{T}}\tilde{\boldsymbol{x}}_{it}\left(\boldsymbol{\delta}_{g_i}\right)\right)^2,$$

其中 $\boldsymbol{\theta}_g^{\mathrm{T}}=\left(\boldsymbol{\alpha}_g^{\mathrm{T}},\boldsymbol{\gamma}_g^{\mathrm{T}}\right)$ 和 $\tilde{\boldsymbol{x}}_{it}\left(\boldsymbol{\delta}_g\right)=\left(\boldsymbol{x}_{it}^{\mathrm{T}}\left(\boldsymbol{\delta}_g\right),\boldsymbol{z}_{it}^{\mathrm{T}}\right)^{\mathrm{T}}$，这里 $\boldsymbol{x}_{it}^{\mathrm{T}}\left(\boldsymbol{\delta}_g\right)=\left(x_{it},\left(x_{it}-\delta_{g1}\right)_+,\cdots,\left(x_{it}-\delta_{gK_g}\right)_+\right)$，并可进一步写成可分离形式

$$Q(\boldsymbol{\delta},\boldsymbol{\mu},\boldsymbol{\theta},\mathbb{G})=\sum_{g=1}^{G}\left\{\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(y_{it}-\mu_i-\boldsymbol{\theta}_g^{\mathrm{T}}\tilde{\boldsymbol{x}}_{it}\left(\boldsymbol{\delta}_g\right)\right)^2 I\left(g_i=g\right)\right\},$$

其中如果表达式 a 成立则 $I(a)=1$，否则 $I(a)=0$。

考虑当给定 $\mathbb{G}=\left\{\mathcal{N}_1,\cdots,\mathcal{N}_G\right\}$ 时，估计每组参数。显然，对每一组数据 $\left\{y_{it},x_{it},\boldsymbol{z}_{it}\right\}$，$i\in\mathcal{N}_g$，$t=1,\cdots,T$ 都可以按照上一节讨论的方法计算相应的折点参数估计量 $\hat{\boldsymbol{\delta}}_g$，固定效应估计量 $\hat{\boldsymbol{\mu}}_g=\left(\hat{\mu}_i\right)_{i\in\mathcal{N}_g}^{\mathrm{T}}$ 和系数参数估计量 $\hat{\boldsymbol{\theta}}_g$。

考虑当已知每组参数 $\boldsymbol{\delta}_g$，$\boldsymbol{\mu}_g$ 和 $\boldsymbol{\theta}_g$ 时，估计 $\mathbb{G}$。我们需要先定义个体之间的距离和个体到群组中心的距离。因为个体固定效应参数 $\boldsymbol{\mu}_g$ 依赖于个体而不是群组，同一个个体分在不同的群组时，个体固定效应需要重新估计，所以定义 g 组的中心仅为参数 $\boldsymbol{c}_g=\left(\boldsymbol{\delta}_g^{\mathrm{T}},\boldsymbol{\theta}_g^{\mathrm{T}}\right)^{\mathrm{T}}$ 而不是 $\left(\boldsymbol{\mu}_g^{\mathrm{T}},\boldsymbol{\delta}_g^{\mathrm{T}},\boldsymbol{\theta}_g^{\mathrm{T}}\right)^{\mathrm{T}}$。基于此，我们定义个体 i 到 g 组的距离为

$$d_{i,g}=\min_{\mu\in\mathbb{R}}\left\{\frac{1}{T}\sum_{t=1}^{T}\left(y_{it}-\mu-\boldsymbol{\theta}_g^{\mathrm{T}}\tilde{\boldsymbol{x}}_{it}\left(\boldsymbol{\delta}_g\right)\right)\right\}.$$

以个体 i 自成一组得到参数估计量 $\tilde{\boldsymbol{\theta}}_i$ 和 $\tilde{\boldsymbol{\delta}}_i$，那么就可以定义个体 i 到个体 j 的距离为

$$\tilde{d}_{i,j}=\min_{\mu\in\mathbb{R}}\left\{\frac{1}{T}\sum_{t=1}^{T}\left(y_{it}-\mu-\tilde{\boldsymbol{\theta}}_j^{\mathrm{T}}\tilde{\boldsymbol{x}}_{it}\left(\tilde{\boldsymbol{\delta}}_j\right)\right)\right\},$$

其作用是找到与给定个体差异最大的个体，可用于选择初始中心。注意这里所定义的个体到个体的距离，事实上并不是度量空间中所定义的距离，比如不满足对称性等。

现在通过 K-means 类型的聚类算法即可完成分组。当完成分组之后，即可使用按照算法 3 得到的每组预估折点位置，然后将其作为初值使用 Nelder-Mead 算法得到精确的折点位置。因而将 g 组内的损失定义为

$$Q_g\left(\boldsymbol{\mu}_g,\boldsymbol{\theta}_g,\boldsymbol{\delta}_g\right)=\frac{1}{N_g T}\sum_{i\in\mathcal{N}_g}\sum_{t=1}^{T}\left(y_{it}-\mu_i-\boldsymbol{\theta}_g^{\mathrm{T}}\tilde{\boldsymbol{x}}_{it}\left(\boldsymbol{\delta}_g\right)\right)^2,\quad g=1,\cdots,G$$

按照上一小节所讨论关于(2)式的优化方法去求解这个问题，即优化

$$\hat{\boldsymbol{\delta}}_g=\arg\min_{\delta_g\in\mathbb{D}^{K_g}}Q_g\left(\boldsymbol{\mu}_g\left(\boldsymbol{\delta}_g\right),\boldsymbol{\theta}_g\left(\boldsymbol{\delta}_g\right),\boldsymbol{\delta}_g\right),\quad g=1,\cdots,G$$

这里的优化是使用算法 3 预估 $\boldsymbol{\delta}_g$，并作为 Nelder-Mead 算法的迭代初值，进而得到精确估计结果。现在可以给出(6)式的估计算法如下：

算法 4：分组多折点回归参数估计算法 $\mathrm{GKinK}\left(\boldsymbol{\delta},G\right)$

输入折点位置 $\boldsymbol{\delta}$，群组个数 G
初始化聚类中心集合 $\boldsymbol{C}\leftarrow\varnothing$
随机选取一个个体 a，估计 $\tilde{\boldsymbol{\theta}}_a$ 和 $\tilde{\boldsymbol{\delta}}_a$ 作为首个聚类中心 $\boldsymbol{c}_1$
更新 $\boldsymbol{C}\leftarrow\{\boldsymbol{c}_1\}$
for g in $[2,\cdots,G]$：
    分别计算每个个体 i 到 $\boldsymbol{C}$ 中每个聚类中心的距离 $\tilde{d}_{i,s}$
    对于每个个体 i，都计算其到最近的中心的距离 $\tilde{d}_i=\min_s\left\{\tilde{d}_{i,s}\right\}$
    如果 $\tilde{d}_s=\max_i\left\{\tilde{d}_i\right\}$：
        计算个体 s 的参数估计 $\tilde{\boldsymbol{\theta}}_s$ 和 $\tilde{\boldsymbol{\delta}}_s$ 作为下一个聚类中心 $\boldsymbol{c}_g$
        更新 $\boldsymbol{C}\leftarrow\boldsymbol{C}\cup\left\{\boldsymbol{c}_g\right\}$
注释：现在已有 G 个聚类中心 $\boldsymbol{C}=\{\boldsymbol{c}_1,\cdots,\boldsymbol{c}_G\}$
while True：
    按照个体到中心的距离，将每个个体分配到最近的聚类中心得到群组估计 $\hat{\mathbb{G}}$
    分别计算每组的聚类中心，随后更新聚类中心集合 $\boldsymbol{C}\leftarrow\{\hat{\boldsymbol{c}}_1,\cdots,\hat{\boldsymbol{c}}_G\}$
    如果相邻两次群组估计不再发生改变，则跳出循环
注释：现在已得到群组估计 $\hat{\mathbb{G}}=\left\{\hat{\mathcal{N}}_1,\cdots,\hat{\mathcal{N}}_G\right\}$ 和聚类中心集合 $\boldsymbol{C}\leftarrow\{\hat{\boldsymbol{c}}_1,\cdots,\hat{\boldsymbol{c}}_G\}$
注释：聚类中心即为该组除固定效应外的参数估计，预估的折点位置
精确计算折点位置 $\hat{\boldsymbol{\delta}}=\left(\hat{\boldsymbol{\delta}}_1^{\mathrm{T}},\cdots,\hat{\boldsymbol{\delta}}_G^{\mathrm{T}}\right)^{\mathrm{T}}$，以及 $\hat{\boldsymbol{\mu}}\left(\hat{\boldsymbol{\delta}}\right)$ 和 $\boldsymbol{\theta}\left(\hat{\boldsymbol{\delta}}\right)$
输出分组标签 $\hat{\mathbb{G}}$，聚类中心集合 $\boldsymbol{C}$，$\hat{\boldsymbol{\mu}}$

关于确定群组个数 G 的方法，通常是手肘法。一般来说手肘法难以自动化，这里采用一种经验的方法，可用于自动提取 G 值。具体来说，先计算不同群组个数相应的均方误差 MSE_i，$i=1,\cdots,M$，其中 M 是指定的最大群组个数，然后依次计算相邻两线段 $\left(g-1,\mathrm{MSE}_{g-1}\right)-\left(g,\mathrm{MSE}_g\right)$ 和 $\left(g+1,\mathrm{MSE}_{g+1}\right)-\left(g,\mathrm{MSE}_g\right)$ 的夹角，最后将夹角最小的位置作为群组个数估计 $\hat{G}$。

3. 蒙特卡洛模拟

3.1. 数据生成

为了评价模型在参数估计方面的性能，我们准备几种不同的模型，个体数量 N 取 30 或 60，重复观测次数 T 取 30 或 60。

生成过程 1 (静态面板)：

$$y_{it}=\mu_i+\varepsilon_{it}+\begin{cases}\alpha_{10}x_{it}+\alpha_{11}\left(x_{it}-\delta_{11}\right)_+ +\gamma_1^{\mathrm{T}}\boldsymbol{z}_{it}, & i\in\mathcal{N}_1\\ \alpha_{20}x_{it}+\alpha_{21}\left(x_{it}-\delta_{21}\right)_+ +\gamma_2^{\mathrm{T}}\boldsymbol{z}_{it}, & i\in\mathcal{N}_2\end{cases}$$

其中两组的个体数量比为 $N_1:N_2=1:2$，两组折点位置分别为 $\delta_{11}=0$ 和 $\delta_{21}=1$。第一组系数 $\boldsymbol{\alpha}_1^{\mathrm{T}}=(\alpha_{10},\alpha_{11})=(1,0)$，$\boldsymbol{\gamma}_1^{\mathrm{T}}=(1,1)$ 而第二组 $\boldsymbol{\alpha}_2^{\mathrm{T}}=(\alpha_{20},\alpha_{21})=(1,-1)$，$\boldsymbol{\gamma}_2^{\mathrm{T}}=(0,1)$，$\mu_i$ 为取值于 $\{0,\cdots,9\}$ 上的离散均匀分布，ε_{it} 为相互独立服从标准正态分布，x_{it} 服从−5 到 5 区间上的均匀分布，$\boldsymbol{z}_{it}$ 服从一个期望为零且协方差矩阵为二阶单位矩阵的正态分布。

生成过程 2 (无协变量 $\boldsymbol{z}_{it}$)：去除协变量 $\boldsymbol{z}_{it}$，其它设定与生成过程 1 相同。

生成过程 3 (AR 误差项)：误差项满足 $\varepsilon_{it}=\rho\varepsilon_{i(t-1)}+\upsilon_{it}$，这里设定 $\rho=0.6$ 且 υ_{it} 服从标准正态分布，其它设定与生成过程 1 相同。

另外，为了评价群组个数识别能力，增加一个群组。模拟数据产生于模型

$$y_{it}=\mu_i+\varepsilon_{it}+\begin{cases}\alpha_{10}x_{it}+\alpha_{11}\left(x_{it}-\delta_{11}\right)_+ +\boldsymbol{\gamma}_1^{T}\boldsymbol{z}_{it}, & i\in\mathcal{N}_1\\ \alpha_{20}x_{it}+\alpha_{21}\left(x_{it}-\delta_{21}\right)_+ +\boldsymbol{\gamma}_2^{T}\boldsymbol{z}_{it}, & i\in\mathcal{N}_2\\ \alpha_{30}x_{it}+\alpha_{31}\left(x_{it}-\delta_{31}\right)_+ +\boldsymbol{\gamma}_3^{T}\boldsymbol{z}_{it}, & i\in\mathcal{N}_3\end{cases}$$

其中三组的个体数量比为 $N_1:N_2:N_3=1:1:1$。设定折点位置分别为 $\delta_{11}=0$，$\delta_{21}=1$ 和 $\delta_{31}=-1$，第一组系数 $\boldsymbol{\alpha}_1^{\mathrm{T}}=(\alpha_{10},\alpha_{11})=(-1,1)$，$\boldsymbol{\gamma}_1^{\mathrm{T}}=(1,1)$，第二组系数 $\boldsymbol{\alpha}_2^{\mathrm{T}}=(\alpha_{20},\alpha_{21})=(-1,2)$，$\boldsymbol{\gamma}_2^{\mathrm{T}}=(0,1)$ 和第三组系数 $\boldsymbol{\alpha}_3^{\mathrm{T}}=(\alpha_{30},\alpha_{31})=(-2,1)$，$\boldsymbol{\gamma}_3^{\mathrm{T}}=(1,2)$。至于随机变量 μ_i，ε_{it}，x_{it} 和 $\boldsymbol{z}_{it}$ 的模拟数据来自于与生成过程 1 同样的分布。相应的无协变量 $\boldsymbol{z}_{it}$ 和 AR 误差项的数据分别按照生成过程 2 和 3 中的定义。

3.2. 评价标准

考虑当真实群组个数 G 已知时，下面分别给出各参数估计量和群组估计量的评价标准。

为了评价分组多折点回归模型的个体折点识别能力，定义折点个数误判比率为

$$\mathrm{KNMR}=\frac{1}{N}\sum_{i=1}^{N}I\left(\hat{K}^{(i)}\neq K_0^{(i)}\right),$$

其中 $\hat{K}^{(i)}$ 和 $K_0^{(i)}$ 分别表示个体 i 的折点个数的估计值和真实值。为了评价系数参数 $\boldsymbol{\delta}$ 和 $\boldsymbol{\alpha}$ 的估计量的表现，考虑到估计的折点数量并不完全对应到真实的折点数量，从而导致参数 $\boldsymbol{\delta}$ 和 $\boldsymbol{\alpha}$ 估计量的长度与它们的真实长度不一定相同，在去除那些折点个数估计不一致的个体之后，才能较为直接评价参数估计效果，从而将这两个参数的均方误差定义为

$$\mathrm{MSE}\left(\hat{\boldsymbol{\delta}}\right)=\frac{1}{N^c}\sum_{i\in\mathcal{N}^c}\left\|\hat{\boldsymbol{\delta}}^{(i)}-\boldsymbol{\delta}_0^{(i)}\right\|_2^2,\ \mathrm{MSE}\left(\hat{\boldsymbol{\alpha}}\right)=\frac{1}{N^c}\sum_{i\in\mathcal{N}^c}\left\|\hat{\boldsymbol{\alpha}}^{(i)}-\boldsymbol{\alpha}_0^{(i)}\right\|_2^2,$$

其中 $\mathcal{N}^c=\left\{i\middle|\hat{K}^{(i)}=K_0^{(i)}\right\}$，$N^c$ 表示集合 $\mathcal{N}^c$ 内元素的个数，$\|a\|_2$ 定义为向量 a 的欧式范数，$\hat{\boldsymbol{\delta}}^{(i)}$ 和 $\hat{\boldsymbol{\alpha}}^{(i)}$ 是个体 i 的估计值，而 $\boldsymbol{\delta}_0^{(i)}$ 和 $\boldsymbol{\alpha}_0^{(i)}$ 是相应的真实值。后文的模拟显示模型的折点错误分类率足够低，只有极少数个体被排除在计算数据之外，因而上述评价标准整体上是有效的。按照一般定义，关于参数 $\boldsymbol{\gamma}$ 的均方误差计算方式为

$$\mathrm{MSE}\left(\hat{\gamma}\right)=\frac{1}{N}\sum_{i=1}^{N}\left\|\hat{\boldsymbol{\gamma}}^{(i)}-\boldsymbol{\gamma}_0^{(i)}\right\|_2^2,$$

其中 $\hat{\boldsymbol{\gamma}}^{(i)}$ 和 $\boldsymbol{\gamma}_0^{(i)}$ 分别是个体 i 的估计值和真实值。

标准化互信息是一种评价不同群组结构接近程度的准则，其取值范围为 $[0,1]$，其值越接近 1 表明两

种群组结构之间越彼此接近，在等于 1 时表明两个群组结构完全相同，具体地，记集类 $\mathbb{A}=\{\mathcal{A}_1,\cdots,\mathcal{A}_r\}$ 和集类 $\mathbb{B}=\{\mathcal{B}_1,\cdots,\mathcal{B}_s\}$，并用 $|\mathcal{A}_n|$ 表示集合 $\mathcal{A}_n$ 内元素的个数，那么当 $\sum_n|\mathcal{A}_n|=\sum_m|\mathcal{B}_m|=N$ 时，$\mathbb{A}$ 和 $\mathbb{B}$ 的互信息计算方式为

$$\mathrm{NMI}(\mathbb{A},\mathbb{B})=\frac{2\mathrm{I}(\mathbb{A},\mathbb{B})}{\mathrm{H}(\mathbb{A})+\mathrm{H}(\mathbb{B})},$$

其中 $\mathrm{I}(\mathbb{A},\mathbb{B})=\sum_{n,m}\left(|\mathcal{A}_n\cap\mathcal{B}_m|/N\right)\log\left(N|\mathcal{A}_n\cap\mathcal{B}_m|/\left(|\mathcal{A}_n||\mathcal{B}_m|\right)\right)$，而 $\mathrm{H}(\mathbb{A})=-\sum_n\left(|A_n|/N\right)\log\left(|A_n|/N\right)$ 表示集类 $\mathbb{A}$ 的信息熵。对于真实群组结构 $\mathbb{G}=\{\mathcal{N}_1,\cdots,\mathcal{N}_G\}$ 及其估计量 $\hat{\mathbb{G}}=\left\{\hat{\mathcal{N}}_1,\cdots,\hat{\mathcal{N}}_G\right\}$，这里使用 $\mathrm{NMI}\left(\hat{\mathbb{G}},\mathbb{G}\right)$ 作为群组结构估计的评价标准。

3.3. 模拟结果

模拟结果如表 1 所示，可以得出一些结论：无论对于什么数据生成过程，在重复观测次数 $T=60$ 时折点个数误判比率 KNMR 均为 0，而在重复观测次数 $T=30$ 时会轻微的错误判断折点个数；对于这三种数据生成过程，T 的增大能显著提升群组结构估计，但是 N 的提升对群组结构估计的效果相比之下并不大；从折点估计 $\hat{\boldsymbol{\delta}}$ 的均方误差数据来看，当固定 T 时，N 的增加有助于提高折点的估计精度，当固定 N 时，T 的增加对折点估计的精度提升效果更好；在 $NT=1800$ 时，由于 $(N,T)=(30,60)$ 这个组合比 $(N,T)=(60,30)$ 的各项指标都要更好，这是因为前者的个体固定效应的参数规模更小，所以样本数据平均在每个参数上含有的信息将会更多；由于协变量 $\boldsymbol{z}_{it}$ 的系数参数 $\boldsymbol{\gamma}$ 和折点参数 $\boldsymbol{\delta}$ 和系数 $\boldsymbol{\alpha}$ 共享同一群组结构，当无变量 $\boldsymbol{z}_{it}$ 时，群组结构的估计效果在一定程度上有所下降，这又反过来影响这些参数的估计效果；最后，受到群组结构估计的影响，AR 误差项时的均方误差比无变量 $\boldsymbol{z}_{it}$ 的均方误差更加稳定。

Table 1. Given G = 2, the Kink number misidentification ratio (KNMR), mean squared error (MSE×10^2), and normalized mutual information (NMI) under different data generation processes, with the results shown as the averages from 500 simulation experiments

表 1. 给定 G = 2，不同数据生成过程时的折点个数误判比率 KNMR (×10^2)，均方误差 MSE (×10^2)和标准化互信息 NMI，所展示的结果为 500 次模拟实验的平均值

生成过程	N	T	KNMR	MSE$(\hat{\boldsymbol{\delta}})$	MSE$(\hat{\boldsymbol{\alpha}})$	MSE$(\hat{\boldsymbol{\gamma}})$	NMI
静态面板	30	30	0.000	3.221	1.236	0.651	0.990
		60	0.000	1.300	0.349	0.233	1.000
	60	30	0.003	1.457	0.502	0.382	0.991
		60	0.000	0.704	0.175	0.107	1.000
无变量 $\boldsymbol{z}_{it}$	30	30	0.000	4.469	3.093	-	0.863
		60	0.000	1.453	0.453	-	0.993
	60	30	0.007	3.212	1.915	-	0.903
		60	0.000	0.711	0.242	-	0.995
AR 误差	30	30	0.007	5.423	2.035	1.302	0.964
		60	0.000	2.283	0.549	0.356	0.999
	60	30	0.003	2.623	0.942	0.768	0.971
		60	0.000	0.944	0.272	0.186	0.999

使用自动化的手肘法选择群组个数，模拟结果如表 2 所示：在每一种情况下，都以高频率正确地选择了分为 3 个组，由于手肘法是基于损失函数值的方法，受到误差项的影响更大，所以数据生成过程是 AR 误差时，正确识别群组个数的频率有所降低。

Table 2. The frequency of the number of kinks selected by the automated elbow method under different conditions, with each condition repeated 100 times
表 2. 不同情况下自动化手肘法选择折点个数的频率，每种情况重复 100 次

生成过程	N	T	手肘法		
			2	**3**	4
静态面板	30	30	0.01	**0.94**	0.05
		60	0.00	**1.00**	0.00
	60	30	0.00	**0.91**	0.09
		60	0.01	**0.98**	0.01
无变量 $\boldsymbol{z}_{it}$	30	30	0.04	**0.87**	0.09
		60	0.02	**0.98**	0.00
	60	30	0.04	**0.83**	0.13
		60	0.00	**1.00**	0.00
AR 误差	30	30	0.05	**0.76**	0.19
		60	0.05	**0.93**	0.02
	60	30	0.12	**0.77**	0.11
		60	0.04	**0.92**	0.04

4. 实证分析

4.1. 数据集介绍

本文所提出的方法将用于分析文献[18]给出的纵向黄体酮激素数据。此数据集收集了 51 位女性在 1~5 个周期内的黄体酮激素含量值。一个完整的观测周期以排卵日为原点，包含其前 8 天和后 15 天在内共 24 日。数据集种共有 91 个观测周期和共 2004 个观测值。由于某些原因，某些观测周期内有缺失数据，所以是非平衡数据，实际数据如图 1 所示。

本文以观测周期为个体 i，所以 $i=1,\cdots,91$ 而 $t=1,\cdots,T_i$，以黄体酮激素含量的对数值作为响应变量 y_{it}，以观测日为门限变量 x_{it}，建立分组多折点回归模型应用于该数据。

4.2. 分组多折点回归模型拟合结果

接下来使用分组多折点回归模型拟合该数据集。设定可能的最大群组个数为 6，自动手肘法选择将个体分为 3 组，相应的模型参数估计结果为

$$y_{it}=\hat{\mu}_i+\hat{\varepsilon}_{it}+\begin{cases}0.07x_{it}, & i\in\mathcal{N}_1\\ 0.00x_{it}+0.40\left(x_{it}+1.21\right)_+-0.54\left(x_{it}-5.82\right)_+, & i\in\mathcal{N}_2\\ 0.02x_{it}+0.45\left(x_{it}+0.56\right)_+-0.36\left(x_{it}-4.60\right)_+, & i\in\mathcal{N}_3\end{cases}$$

这里 $\hat{\mu}_i$ 为个体固定效应，但并不是主要感兴趣的估计量，这里并未列出。从估计结果来看：在第一组

内，对数黄体酮与观测日成简单的线性关系，在排卵日前后共 24 天内，对数黄体酮只有略微的上升趋势，但并无折点；在第二组内，当排卵日未到来时，对数黄体酮激素含量无变化，约在排卵日前 1.21 天出现折点，对数黄体酮激素含量产生激增，直到排卵日后 5.82 天又快速发生回落；对于第三组，与第二组同为两个折点，但主要区别在于其对数黄体酮激素含量在第一个折点位置上升更快，在第二个折点回落的更慢。

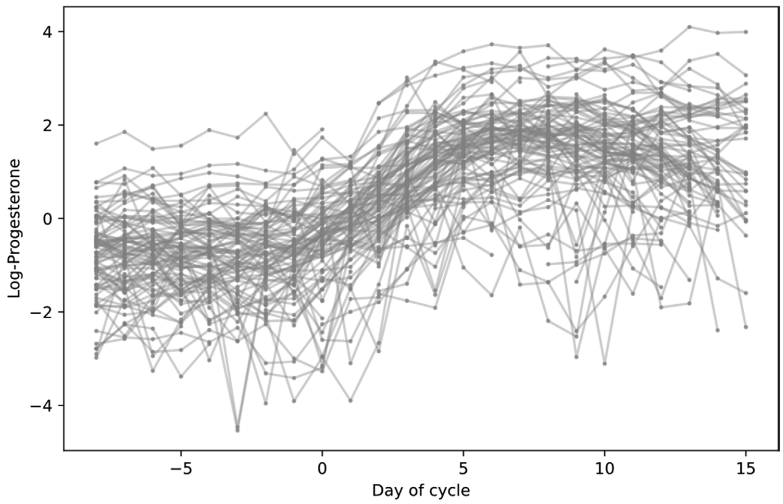

Figure 1. Logarithmic progesterone observation data, the horizontal axis is the date with ovulation day as the origin, and the vertical axis represents the logarithmic progesterone content
图 1. 对数黄体酮含量观测值，横轴为以排卵日为原点的日期，纵轴表示对数黄体酮含量

为便于展示各组对数黄体酮激素含量与观测日的关系，将个体数据减去其固定效应估计值，即 $y_{it}-\hat{\mu}_i$，得到原始数据上下平移后的新数据，并分别绘制各组的情况。结果如图 2 所示，直观地印证了前面的参数估计结果分析，其中第二组和第三组最显著的区别在于，第二组在在一个观测周期的最后几天已经开始回落，而第三组依然在上升。

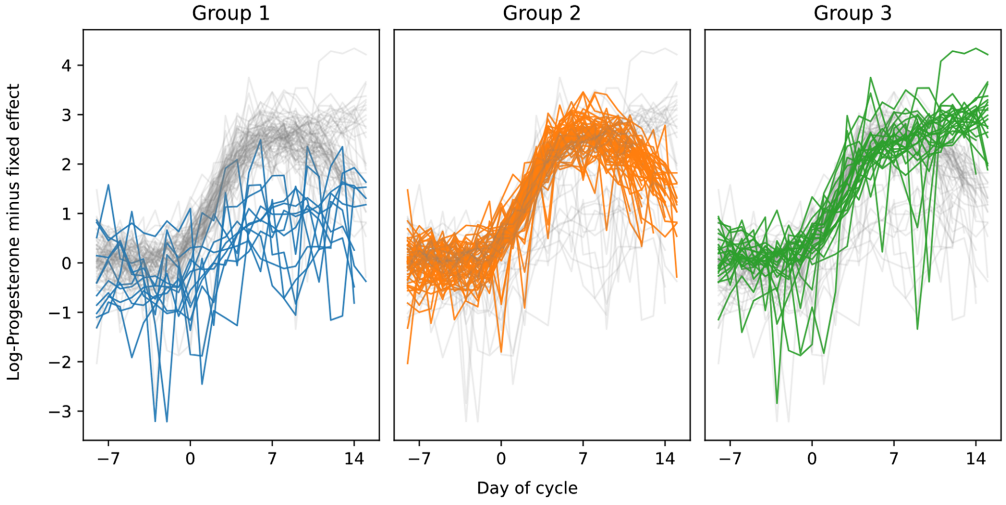

Figure 2. The data were divided into 3 groups and the logarithmic progesterone observations were translated up and down according to the fixed effect
图 2. 将数据分为 3 组并且按照固定效应上下平移后的对数黄体酮观测数据

使用未分组的多折点回归模型拟合，即只有一组时，得到模型估计如下

$$\hat{y}_{it} = \hat{\mu}_i + 0.00x_{it} + 0.40\left(x_{it} + 0.71\right)_+ - 0.45\left(x_{it} - 5.32\right)_+ .$$

相较于群组个数为 3 的分组折点回归模型，这个模型只能从整体层面得到对数黄体酮和观测日的关系，只能观察到一些到被掩盖和混合的规律，即第一组无明显折点，第二组和第三组在观测末期有一些相反的趋势，整体拟合后得出走势平缓的关系。

5. 总结

本文基于多折点回归模型和 K-means 类型的聚类方法，提出了分组多折点回归模型和响应的估计方法。相对于以往的折点估计方法来说，该方法能将观测个体划分成不同的群组，所以能够更加灵活地捕捉数据的内在信息。数值模拟实验和实际数据分析表明其有良好的估计性能和实际可行性。

参考文献

[1] Lerman, P.M. (1980) Fitting Segmented Regression Models by Grid Search. *Journal of the Royal Statistical Society. Series C*, **29**, 77-84. https://doi.org/10.2307/2346413

[2] Hinkley, D., Chapman, P. and Runger, G. (1980) Change-Point Problems. Institute of Mathematical Statistics.

[3] Chappell, R. (1989) Fitting Bent Lines to Data, with Applications to Allometry. *Journal of Theoretical Biology*, **138**, 235-256. https://doi.org/10.1016/s0022-5193(89)80141-9

[4] Fong, Y., Di, C., Huang, Y. and Gilbert, P.B. (2016) Model-Robust Inference for Continuous Threshold Regression Models. *Biometrics*, **73**, 452-462. https://doi.org/10.1111/biom.12623

[5] Hansen, B.E. (2017) Regression Kink with an Unknown Threshold. *Journal of Business & Economic Statistics*, **35**, 228-240. https://doi.org/10.1080/07350015.2015.1073595

[6] Li, C., Wei, Y., Chappell, R. and He, X. (2010) Bent Line Quantile Regression with Application to an Allometric Study of Land Mammals' Speed and Mass. *Biometrics*, **67**, 242-249. https://doi.org/10.1111/j.1541-0420.2010.01436.x

[7] Zhang, F. and Li, Q. (2017) Robust Bent Line Regression. *Journal of Statistical Planning and Inference*, **185**, 41-55. https://doi.org/10.1016/j.jspi.2017.01.001

[8] Zhong, W., Wan, C. and Zhang, W. (2021) Estimation and Inference for Multi-Kink Quantile Regression. *Journal of Business & Economic Statistics*, **40**, 1123-1139. https://doi.org/10.1080/07350015.2021.1901720

[9] Fong, Y. (2019) Fast Bootstrap Confidence Intervals for Continuous Threshold Linear Regression. *Journal of Computational and Graphical Statistics*, **28**, 466-470. https://doi.org/10.1080/10618600.2018.1537927

[10] Muggeo, V.M.R. (2003) Estimating Regression Models with Unknown Break-Points. *Statistics in Medicine*, **22**, 3055-3071. https://doi.org/10.1002/sim.1545

[11] Gössl, C. and Küchenhoff, H. (2001) Bayesian Analysis of Logistic Regression with an Unknown Change Point and Covariate Measurement Error. *Statistics in Medicine*, **20**, 3109-3121. https://doi.org/10.1002/sim.928

[12] Li, Y., Hu, Z., Liu, J. and Deng, J. (2021) A Note on Regression Kink Model. *Communications in Statistics—Theory and Methods*, **51**, 8246-8263. https://doi.org/10.1080/03610926.2021.1890780

[13] Yang, L., Zhang, C., Lee, C. and Chen, I. (2020) Panel Kink Threshold Regression Model with a Covariate-Dependent Threshold. *The Econometrics Journal*, **24**, 462-481. https://doi.org/10.1093/ectj/utaa035

[14] Zhou, M., Ye, F., Li, Y., Liu, F. and Wan, C. (2024) A Note on the Covariate-Dependent Kink Threshold Regression Model for Panel Data. *Communications in Statistics—Theory and Methods*. https://doi.org/10.1080/03610926.2024.2324985

[15] Wan, C., Zhong, W., Zhang, W. and Zou, C. (2022) Multikink Quantile Regression for Longitudinal Data with Application to Progesterone Data Analysis. *Biometrics*, **79**, 747-760. https://doi.org/10.1111/biom.13667

[16] Du, L., Koscik, R. L., Betthauser, T. J., Johnson, S. C., Larget, B. and Chappell, R. (2022) Bayesian Bent-Line Regression Model for Longitudinal Data with an Application to the Study of Cognitive Performance Trajectories in Wisconsin Registry for Alzheimer's Prevention. arXiv: 2211.09915. https://doi.org/10.48550/arXiv.2211.09915

[17] Sun, Y., Wan, C., Zhang, W. and Zhong, W. (2024) A Multi-Kink Quantile Regression Model with Common Structure for Panel Data Analysis. *Journal of Econometrics*, **239**, Article ID: 105304. https://doi.org/10.1016/j.jeconom.2022.04.012

[18] Munro, C.J., Stabenfeldt, G.H., Cragun, J.R., Addiego, L.A., Overstreet, J.W. and Lasley, B.L. (1991) Relationship of

Serum Estradiol and Progesterone Concentrations to the Excretion Profiles of Their Major Urinary Metabolites as Measured by Enzyme Immunoassay and Radioimmunoassay. *Clinical Chemistry*, **37**, 838-844.
https://doi.org/10.1093/clinchem/37.6.838