Advances in Applied Mathematics 应用数学进展, 2026, 15(5), 1-9
Published Online May 2026 in Hans. https://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2026.155202

Hans汉斯

基于LsoCV准则的退出型缺失纵向数据模型平均估计

卢文暄，黄　彬*，沈皓明

北京化工大学数理学院，北京

收稿日期：2026年4月9日；录用日期：2026年5月2日；发布日期：2026年5月9日

摘　要

本文针对随机缺失假设下存在退出缺失的纵向数据，提出一种基于LsoCV准则的模型平均估计方法。该方法采用加权广义估计方程(WGEE)对各候选模型的参数进行估计，并通过去个体交叉验证(LsoCV)准则确定各模型的权重。模拟研究表明，所提出的模型平均方法展现出相较于其他替代方法更好的性能，且其优越性通过应用于PBC数据得到了进一步验证。

关键词

纵向数据，退出型缺失，加权广义估计方程(WGEE)，模型平均，LsoCV准则

LsoCV Criterion-Based Model Averaging for Longitudinal Data with Dropout

Wenxuan Lu, Bin Huang*, Haoming Shen

School of Mathematics and Physics, Beijing University of Chemical Technology, Beijing

Received: April 9, 2026; accepted: May 2, 2026; published: May 9, 2026

Abstract

In this paper, a model averaging estimation method based on the LsoCV criterion is proposed for longitudinal data with dropouts under the assumption of missing at random. This method adopts the weighted generalized estimating equations (WGEE) to estimate the parameters of each candidate model, and determines the weights of each model through the leave-subject-out cross-validation (LsoCV) criterion. Simulation studies reveal that the proposed model averaging method exhibits

*通讯作者。

文章引用: 卢文暄, 黄彬, 沈皓明. 基于 LsoCV 准则的退出型缺失纵向数据模型平均估计[J]. 应用数学进展, 2026, 15(5): 1-9. DOI: 10.12677/aam.2026.155202

much better performance compared with other competing methods, and its superiority is further verified by its application to the PBC data.

Keywords

Longitudinal Data, Dropout, Weighted Generalized Estimating Equations (WGEE), Model Averaging, LsoCV Criterion

1. 引言

纵向数据在临床试验或观察性研究中十分常见，它是指对同一组个体进行重复观测所得的多维数据，具有个体内相关、数据结构不平衡等特性。Liang 和 Zeger [1]提出的广义估计方程(GEE)方法可以有效地处理个体内相关问题，成为纵向数据分析中应用最广泛的边际模型之一。

在纵向数据分析过程中，为更好地解决模型不确定性问题、进一步提高模型预测的精准度，模型选择与模型平均是两类切实可行的解决路径。其中，模型选择的核心目标的是从众多候选模型中，筛选出最能揭示变量间内在关联的最优模型；而模型平均方法的核心思路则是全面考虑所有候选模型，通过对各模型的估计值进行加权整合，实现优势信息的综合利用。相较于模型选择，模型平均能够显著降低推断风险，进一步提升预测精度，有效弥补单一模型选择的局限性。

针对纵向数据，Pan [2]提出了独立模型下的拟似然(QIC)准则，该准则可同时用于 GEE 分析中的变量选择和相关结构选择。Gao 等[3]提出一种基于去个体交叉验证的频率模型平均方法，并在理论上证明了所得模型平均估计量 LsoMA 的渐近最优性。Zhao 和 Zuo [4]将 LsoMA 方法进一步拓展至高维纵向数据场景。针对变系数部分线性模型，Hu 等[5]提出了一种聚焦信息准则(FIC)和一个频率模型平均估计量。Li 等[6]研究了部分线性模型的模型平均问题，其权重选择准则为样本内期望平方误差损失的无偏估计量与一个常数之和。Jiang 等[7]研究了具有超高维协变量的稳健模型平均预测问题。值得一提的是，Yu 等[8]针对模型平均方法的通用性与有效性问题，提出了一种基于交叉验证且具有一般损失函数的统一最优模型平均方法，该方法可适用于离散响应纵向数据的场景。

上述方法大多仅聚焦于完整数据的情形，未考虑数据缺失等复杂场景。然而，在纵向数据分析过程中，响应变量的退出型缺失(dropout missingness)是一类常见问题，该情形往往由部分个体因死亡、研究依从性差等原因中途退出追踪观测导致。若直接采用常规统计方法进行参数估计，极易产生估计偏差。在退出型缺失机制为随机缺失(MAR)的假设下，Robins 等[9]提出了加权广义估计方程(WGEE)方法，该方法通过引入逆概率加权矩阵对 GEE 方法进行了拓展。文献研究表明，在存在退出型缺失且缺失机制为 MAR 的纵向数据应用场景中，QIC 准则的模型选择表现并不理想，Platt 等[10]与 Gosho [11]分别提出 QICWp 和 QICWr 准则对其进行修正。Shen 和 Chen [12]基于平方损失函数提出 MLIC 与 MLICC 准则，用于变量选择及工作相关矩阵的选择，进一步完善了 GEE 模型选择体系。此外，相较于模型选择领域的研究，模型平均方面针对此类数据的相关研究仍较为匮乏。

因此，本文拟针对退出型缺失纵向数据的模型平均问题展开研究，提出了一种可有效提升预测精度的模型平均方法。具体而言，我们将不同协变量与工作协方差矩阵纳入候选模型，采用 WGEE 对候选模

型进行估计，并通过 LsoCV 准则确定各候选模型的权重，最终通过模拟实验与实例分析进一步验证所提方法的优良性。

2. 模型平均估计

2.1. 模型假设

假设随机样本包含 n 个个体。对第 i 个个体，令 y_{it} 为第 it 个响应变量，$\boldsymbol{x}_{it}$ 为对应的协变量向量，$i=1,\cdots,n$，$t=1,\cdots,T_i$，观测总数 $T=\sum_{i=1}^{n}T_i$。记 $\boldsymbol{Y}_i=\left(y_{i1},\cdots,y_{iT_i}\right)'$，$\boldsymbol{X}_i=\left(\boldsymbol{x}_{i1},\cdots,\boldsymbol{x}_{iT_i}\right)'$，考虑纵向数据下的线性模型：

$$\boldsymbol{Y}_i=\boldsymbol{\mu}_i+\boldsymbol{\varepsilon}_i=\boldsymbol{X}_i\boldsymbol{\beta}+\boldsymbol{\varepsilon}_i,i=1,\cdots,n,$$

其中 $\mu_{it}=\boldsymbol{x}_{it}'\boldsymbol{\beta}$，$\boldsymbol{\mu}_i=\left(\mu_{i1},\cdots,\mu_{iT_i}\right)'$，$\boldsymbol{\beta}$ 为未知的 p 维回归系数，$\boldsymbol{\varepsilon}_i=\left(\varepsilon_{i1},\cdots,\varepsilon_{iT_i}\right)'$ 为随机误差，且满足 $E\left(\boldsymbol{\varepsilon}_i\mid\boldsymbol{X}_i\right)=\boldsymbol{0}$，$E\left(\boldsymbol{\varepsilon}_i\boldsymbol{\varepsilon}_i'\mid\boldsymbol{X}_i\right)=\boldsymbol{\Sigma}_i>0$。这里假定协变量 $\boldsymbol{x}_{it}$ 能被完全观测，但响应变量 y_{it} 可能在时间 $t>1$ 时出现退出型缺失。令 r_{it} 为缺失的指标变量，即当 y_{it} 缺失时 $r_{it}=0$，否则 $r_{it}=1$。退出型缺失会导致单调缺失数据模型，若 y_{it} 缺失意味着 $y_{i(t+1)},\cdots,y_{iT_i}$ 也缺失，即若 $r_{it}=0$ 则之后的 $r_{i(t+1)}=\cdots=r_{iT_i}=0$。另外假设所有个体的第一次试验都能被观测，即 $r_{it}=1$。

令 $\pi_{it}=P\left(r_{it}=1\mid\boldsymbol{X}_i,\boldsymbol{Y}_i\right)$，$\lambda_{it}=P\left(r_{it}=1\mid r_{i(t-1)}=1,\boldsymbol{X}_i,y_{i1},\cdots,y_{i(t-1)}\right)$，则有 $\lambda_{i1}=1$。在 MAR 机制下，$\pi_{it}=\lambda_{i1}\times\cdots\times\lambda_{it}$。在响应变量存在退出型缺失的情形下，利用如下的基于恒等连接函数的 WGEE 估计 $\boldsymbol{\beta}$，

$$U\left(\boldsymbol{\beta}\right)=\sum_{i=1}^{n}U_i\left(\boldsymbol{\beta}\right)=\sum_{i=1}^{n}\boldsymbol{X}_i'\boldsymbol{V}_i^{-1}\boldsymbol{W}_i\left(\boldsymbol{Y}_i-\boldsymbol{X}_i\boldsymbol{\beta}\right)=\boldsymbol{0},$$

其中 $\boldsymbol{W}_i=diag\left(r_{i1}/\pi_{i1},\cdots,r_{iT_i}/\pi_{iT_i}\right)$，$\boldsymbol{V}_i$ 为工作协方差矩阵。令 $\boldsymbol{Y}=\left(\boldsymbol{Y}_1',\cdots,\boldsymbol{Y}_n'\right)'$，$\boldsymbol{X}=\left(\boldsymbol{X}_1',\cdots,\boldsymbol{X}_n'\right)'$，$\boldsymbol{\mu}=\left(\boldsymbol{\mu}_1',\cdots,\boldsymbol{\mu}_n'\right)'$，$\boldsymbol{\varepsilon}=\left(\boldsymbol{\varepsilon}_1',\cdots,\boldsymbol{\varepsilon}_n'\right)'$，$\boldsymbol{V}=diag\left(\boldsymbol{V}_1,\cdots,\boldsymbol{V}_n\right)$，$\boldsymbol{W}=diag\left(\boldsymbol{W}_1,\cdots,\boldsymbol{W}_n\right)$。则 $\boldsymbol{\beta}$ 的估计可表示为

$$\breve{\boldsymbol{\beta}}=\left(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{W}\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{W}\boldsymbol{Y},$$

相应地，$\boldsymbol{\mu}$ 的估计可表示为

$$\breve{\boldsymbol{\mu}}=\boldsymbol{X}\breve{\boldsymbol{\beta}}=\boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{W}\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{W}\boldsymbol{Y}.$$

2.2. 模型平均准则

为实现对 $\boldsymbol{\mu}$ 的精准预测，本小节给出一种可行的权重选择准则，并基于该准则得到 $\boldsymbol{\mu}$ 的模型平均估计量。共考虑 M 个候选模型，其中第 m 个候选模型为

$$\boldsymbol{Y}_i=\boldsymbol{X}_i^{(m)}\boldsymbol{\beta}_{(m)}+\boldsymbol{\varepsilon}_i^{(m)},i=1,\cdots,n,$$

其中 $\boldsymbol{X}_i^{(m)}$ 为 $T_i\times k_m$ 阶协变量矩阵，$\boldsymbol{\beta}_{(m)},\boldsymbol{\varepsilon}_i^{(m)}$ 为对应的回归参数和随机误差，$m=1,\cdots,M$。在第 m 个候选模型下，$\boldsymbol{\mu}$ 的估计为

$$\breve{\boldsymbol{\mu}}^{(m)}=\boldsymbol{X}^{(m)}\left(\boldsymbol{X}^{(m)\prime}\boldsymbol{V}_{(m)}^{-1}\boldsymbol{W}_{(m)}\boldsymbol{X}^{(m)}\right)^{-1}\boldsymbol{X}^{(m)\prime}\boldsymbol{V}_{(m)}^{-1}\boldsymbol{W}_{(m)}\boldsymbol{Y}\triangleq\breve{\boldsymbol{P}}^{(m)}\boldsymbol{Y},$$

这里 $\breve{\boldsymbol{P}}^{(m)}=\boldsymbol{X}^{(m)}\left(\boldsymbol{X}^{(m)\prime}\boldsymbol{V}_{(m)}^{-1}\boldsymbol{W}_{(m)}\boldsymbol{X}^{(m)}\right)^{-1}\boldsymbol{X}^{(m)\prime}\boldsymbol{V}_{(m)}^{-1}\boldsymbol{W}_{(m)}$，$\boldsymbol{X}^{(m)}=\left(\boldsymbol{X}_1^{(m)\prime},\cdots,\boldsymbol{X}_n^{(m)\prime}\right)'$ 为 $T\times k_m$ 阶矩阵，$\boldsymbol{V}_{(m)}=diag\left(\boldsymbol{V}_{(m)1},\cdots,\boldsymbol{V}_{(m)n}\right)$ 为有可能错误指定的 $T\times T$ 阶协方差矩阵，$\boldsymbol{W}_{(m)}=diag\left(\boldsymbol{W}_{(m)1},\cdots,\boldsymbol{W}_{(m)n}\right)$，其中，

$\pi_{it}^{(m)}=P\left(r_{it}=1\mid \boldsymbol{X}_i^{(m)},\boldsymbol{Y}_i\right)=\lambda_{i1}^{(m)}\times\cdots\times\lambda_{it}^{(m)}$，$\lambda_{it}^{(m)}=P\left(r_{it}=1\mid r_{i(t-1)}=1,\boldsymbol{X}_i^{(m)},y_{i1},\cdots,y_{i(t-1)}\right)$，$\boldsymbol{W}_{(m)i}=diag\left(r_{i1}/\pi_{i1}^{(m)},\cdots,r_{iT_i}/\pi_{iT_i}^{(m)}\right)$。

令 $\boldsymbol{w}$ 为集合 $\mathcal{H}=\left\{\boldsymbol{w}\in[0,1]^M:\sum_{m=1}^{M}w_m=1\right\}$ 中的权重向量，可得 $\boldsymbol{\mu}$ 的一个模型平均估计

$$\breve{\boldsymbol{\mu}}(\boldsymbol{w})=\sum_{m=1}^{M}w_m\breve{\boldsymbol{\mu}}^{(m)}=\sum_{m=1}^{M}w_m\breve{\boldsymbol{P}}^{(m)}\boldsymbol{Y}\triangleq\breve{\boldsymbol{P}}(\boldsymbol{w})\boldsymbol{Y},\tag{1}$$

其中 $\breve{\boldsymbol{P}}(\boldsymbol{w})=\sum_{m=1}^{M}w_m\breve{\boldsymbol{P}}^{(m)}$。

值得注意的是，$\boldsymbol{W}_{(m)}$ 一般是未知的，这里我们利用逻辑(Logistic)回归模型估计 $\lambda_{it}^{(m)}$，$\text{logit}\left(\lambda_{it}^{(m)}\right)=\boldsymbol{z}_{it}^{(m)\prime}\boldsymbol{\gamma}_{(m)}$，其中 $\boldsymbol{\gamma}_{(m)}$ 为未知的 q_m 维系数，$\boldsymbol{z}_{it}^{(m)}$ 是某些协变量和历史响应变量组成的向量。记 $\hat{\boldsymbol{\gamma}}_{(m)}$ 为 $\boldsymbol{\gamma}_{(m)}$ 的极大似然估计，并将 $\hat{\boldsymbol{\gamma}}_{(m)}$ 代入 $\lambda_{it}^{(m)}$ 得到估计 $\hat{\lambda}_{it}^{(m)}$，从而得到 $\hat{\pi}_{it}^{(m)}=\hat{\lambda}_{i1}^{(m)}\times\cdots\times\hat{\lambda}_{it}^{(m)}$，$\hat{\boldsymbol{W}}_{(m)i}=diag\left(r_{i1}/\hat{\pi}_{i1}^{(m)},\cdots,r_{iT_i}/\hat{\pi}_{iT_i}^{(m)}\right)$，$\hat{\boldsymbol{W}}_{(m)}=diag\left(\hat{\boldsymbol{W}}_{(m)1},\cdots,\hat{\boldsymbol{W}}_{(m)n}\right)$。将 $\hat{\boldsymbol{W}}_{(m)}$ 代入(1)式，得到 $\boldsymbol{\mu}$ 的模型平均估计

$$\hat{\boldsymbol{\mu}}(\boldsymbol{w})=\sum_{m=1}^{M}w_m\boldsymbol{P}^{(m)}\boldsymbol{Y}\triangleq\boldsymbol{P}(\boldsymbol{w})\boldsymbol{Y},$$

其中，$\boldsymbol{P}(\boldsymbol{w})=\sum_{m=1}^{M}w_m\boldsymbol{P}^{(m)}$，$\boldsymbol{P}^{(m)}=\boldsymbol{X}^{(m)}\left(\boldsymbol{X}^{(m)\prime}\boldsymbol{V}_{(m)}^{-1}\hat{\boldsymbol{W}}_{(m)}\boldsymbol{X}^{(m)}\right)^{-1}\boldsymbol{X}^{(m)\prime}\boldsymbol{V}_{(m)}^{-1}\hat{\boldsymbol{W}}_{(m)}$。

此外，工作协方差矩阵 $\boldsymbol{V}_{(m)i}$ 通常基于随机误差 $\boldsymbol{\varepsilon}_i^{(m)}$ 的工作相关结构进行估计。在实际应用中，经常采用复合对称(EX)结构与自回归(AR)结构的工作相关结构。正如 Liang 和 Zeger [1]所指出的，即使相关结构被错误指定，与完全忽略组内相关性的方法相比，该方法仍有可能提升估计效率。作为协方差矩阵估计的一种替代方法，我们还可采用不同的工作协方差矩阵来构建候选模型集。我们所提出的模型平均法，可通过选取不同的协变量与工作协方差矩阵作为候选模型，直接用于降低模型不确定性。

模型平均估计 $\hat{\boldsymbol{\mu}}(\boldsymbol{w})$ 依赖于权重 $\boldsymbol{w}$ 的选取，我们采用去个体交叉验证(LsoCV [3])准则来选择 $\boldsymbol{w}$。首先，令 $\boldsymbol{Y}_{[-i]}=\left(\boldsymbol{Y}_1',\cdots,\boldsymbol{Y}_{i-1}',\boldsymbol{Y}_{i+1}',\cdots,\boldsymbol{Y}_n'\right)'$，$\boldsymbol{X}_{[-i]}^{(m)}=\left(\boldsymbol{X}_1^{(m)\prime},\cdots,\boldsymbol{X}_n^{(m)\prime}\right)'$，$\boldsymbol{V}_{(m)[-i]},\hat{\boldsymbol{W}}_{(m)[-i]}$ 分别表示删去 $\boldsymbol{V}_{(m)},\hat{\boldsymbol{W}}_{(m)}$ 第 i 个对角块后得到的矩阵，$i=1,\cdots,n$，采用去个体法(leave-subject-out)估计 $\boldsymbol{\mu}_i$ 为

$$\boldsymbol{\mu}_i^{*(m)}=\boldsymbol{X}_i^{(m)}\left(\boldsymbol{X}_{[-i]}^{(m)\prime}\boldsymbol{V}_{(m)[-i]}^{-1}\hat{\boldsymbol{W}}_{(m)[-i]}\boldsymbol{X}_{[-i]}^{(m)}\right)^{-1}\boldsymbol{X}_{[-i]}^{(m)\prime}\boldsymbol{V}_{(m)[-i]}^{-1}\hat{\boldsymbol{W}}_{(m)[-i]}\boldsymbol{Y}_{[-i]},$$

其中 $\boldsymbol{\mu}_i^{*(m)}=\left(\mu_{i1}^{*(m)},\cdots,\mu_{iT_i}^{*(m)}\right)'$，令 $\boldsymbol{\mu}^{*(m)}=\left(\boldsymbol{\mu}_1^{*(m)\prime},\cdots,\boldsymbol{\mu}_n^{*(m)\prime}\right)'$，$\boldsymbol{\mu}^{*}(\boldsymbol{w})=\sum_{m=1}^{M}w_m\boldsymbol{\mu}^{*(m)}$。在退出型缺失机制下，我们基于完全观测数据(complete case data)构造 LsoCV 准则。对第 i 个个体，假如 $y_{i1},\cdots,y_{iJ_i}$ 未缺失，且 $r_{i(J_i+1)}=0$ (若 $J_i<T_i$)，令 $\tilde{\boldsymbol{Y}}=\left(\tilde{\boldsymbol{Y}}_1',\cdots,\tilde{\boldsymbol{Y}}_n'\right)'$，$\tilde{\boldsymbol{Y}}_i=\left(y_{i1},\cdots,y_{iJ_i}\right)'$ 为 $\boldsymbol{Y}_i$ 中未缺失响应变量的子集，相应地，令 $\tilde{\boldsymbol{\mu}}_i^{(m)}=\left(\mu_{i1}^{*(m)},\cdots,\mu_{iJ_i}^{*(m)}\right)'$，$\tilde{\boldsymbol{\mu}}^{(m)}=\left(\tilde{\boldsymbol{\mu}}_1^{(m)\prime},\cdots,\tilde{\boldsymbol{\mu}}_n^{(m)\prime}\right)'$，$\tilde{\boldsymbol{\mu}}(\boldsymbol{w})=\sum_{m=1}^{M}w_m\tilde{\boldsymbol{\mu}}^{(m)}$。定义完全观测数据下的 LsoCV 准则为

$$\text{LsoCV}(\boldsymbol{w})=\left\|\tilde{\boldsymbol{Y}}-\tilde{\boldsymbol{\mu}}(\boldsymbol{w})\right\|^2,\tag{2}$$

通过极小化(2)式，得到 $\hat{\boldsymbol{w}}=\arg\min_{\boldsymbol{w}\in\mathcal{H}}\text{LsoCV}(\boldsymbol{w})$。此后，估计量 $\hat{\boldsymbol{\mu}}(\hat{\boldsymbol{w}})$ 被称为参数 $\boldsymbol{\mu}$ 的退出缺失机

制下去个体模型平均估计量(DM-LsoMA)。

3. 模拟计算

在本小节中，我们设计一个模拟实验来验证所提方法的性能，并与其他一些替代方法进行比较。假设数据生成模型如下：

$$y_{it} = \mu_{it} + \varepsilon_{it} = \beta_0 + x_{i1}\beta_1 + x_{it2}\beta_2 + x_{it3}\beta_3 + \varepsilon_{it},\ i = 1,\cdots,n;\ t = 1,\cdots,T_i,$$

其中 $x_{i1} \sim U(0,1)$，$x_{it2} = t-1$，$x_{it3} \sim N(0,1)$，$\boldsymbol{\beta} = (\beta_0,\beta_1,\beta_2,\beta_3)' = (-2,1,0.5,0.3)'$，$T_i = 3$，$n$ 取 50 或 100，随机误差项 $(\varepsilon_{i1},\cdots,\varepsilon_{iT_i})'$ 服从均值为 0，方差为 1，且相关矩阵为 $R(\rho)$ 的多元正态分布。我们考察真实相关矩阵 $R(\rho)$ 分别为可交换(EX)与一阶自回归(AR (1))结构的两种场景，其中 $\rho \in \{0.3,0.5\}$。考虑退出缺失模型如下：

$$\log\left(\frac{\lambda_{it}}{1-\lambda_{it}}\right) = \theta_0 + y_{i(t-1)}\theta_1 + h_{it}\theta_2,\ i = 1,\cdots,n;\ t = 2,\cdots,T,$$

其中 $h_{it} \sim U(-0.5,0.5)$，参数 $\boldsymbol{\theta} = (\theta_0,\theta_1,\theta_2)'$ 分别取值 $(-0.1,1.3,-1.2)'$ 和 $(-0.1,1.3,-0.7)'$ 将退出缺失率 mr 分别设定为 0.2 和 0.3 左右。

这里选取 5 种协变量：$\{\{x_{i1}\},\{x_{it3}\},\{x_{i1},x_{it2}\},\{x_{i1},x_{it3}\},\{x_{it2},x_{it3}\}\}$，并将单位矩阵、EX、AR (1)三种工作相关矩阵类型纳入候选模型。通过协变量与工作相关矩阵的完全交叉组合，共构建 15 个候选模型用于后续分析。

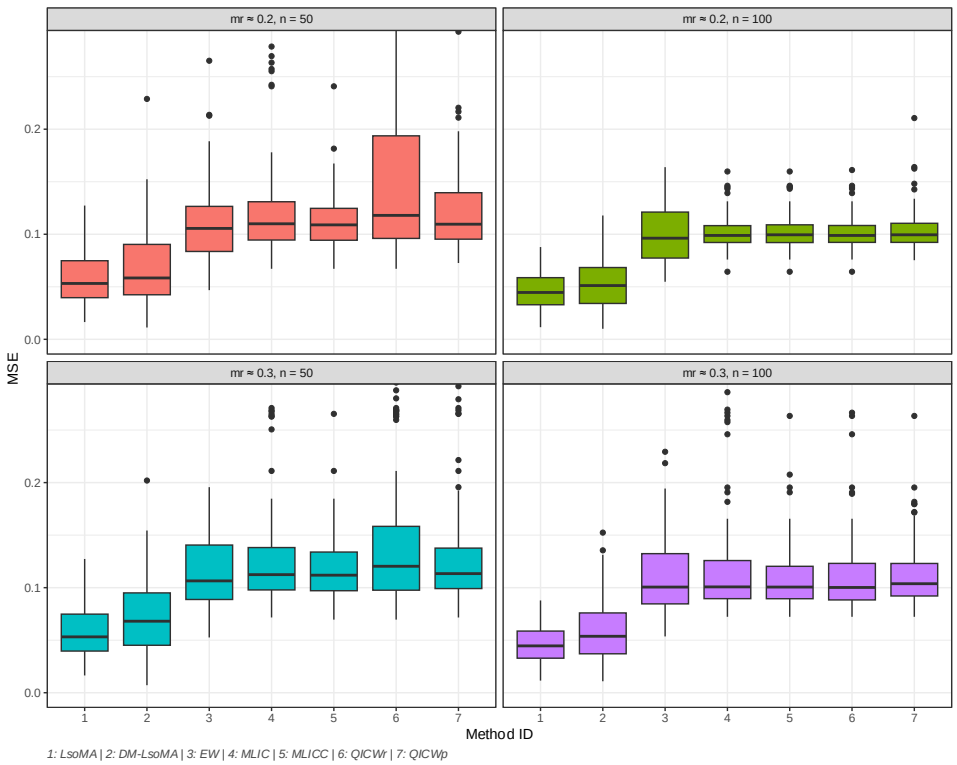

Figure 1. Boxplots of MSE under different methods (for $R(\rho)$ EX structure, $\rho = 0.3$)

图 1. 不同方法下 MSE 的箱线图($R(\rho)$ 为 EX 结构，$\rho = 0.3$)

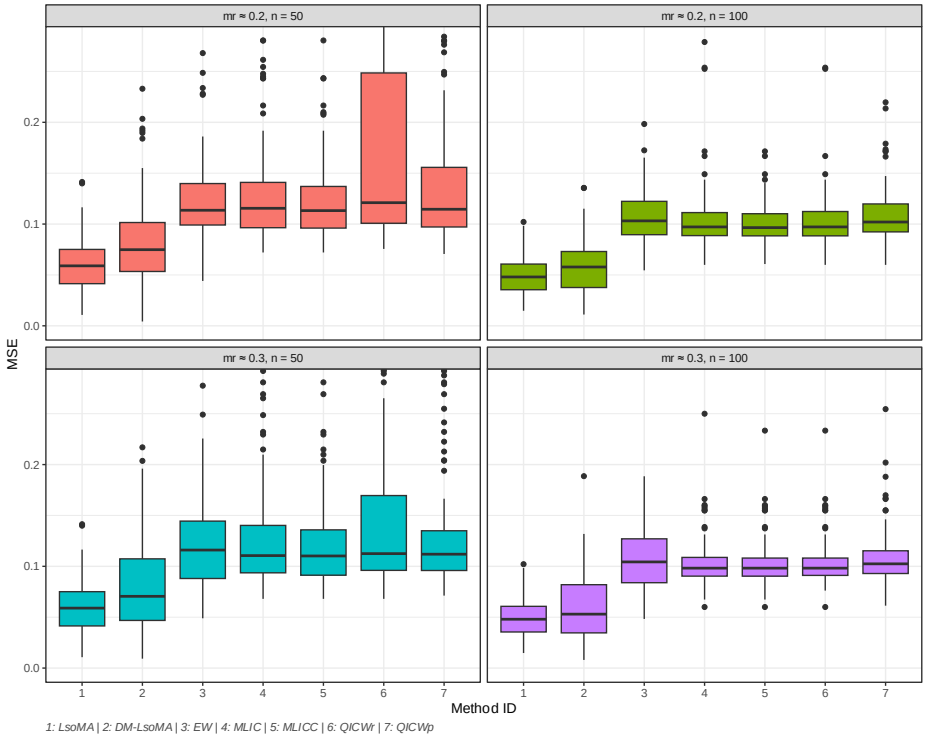

Figure 2. Boxplots of MSE under different methods (for $R(\rho)$ EX Structure, $\rho = 0.5$)

图 2. 不同方法下 MSE 的箱线图($R(\rho)$ 为 EX 结构， $\rho = 0.5$)

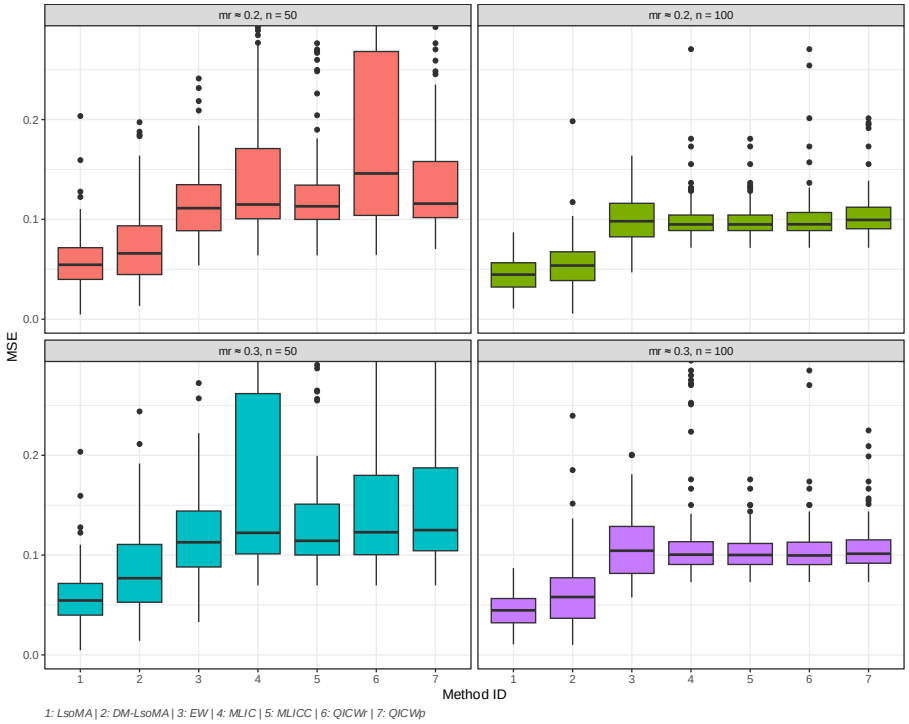

Figure 3. Boxplots of MSE under different methods (for $R(\rho)$ AR (1) Structure, $\rho = 0.3$)

图 3. 不同方法下 MSE 的箱线图($R(\rho)$ 为 AR (1)结构， $\rho = 0.3$)

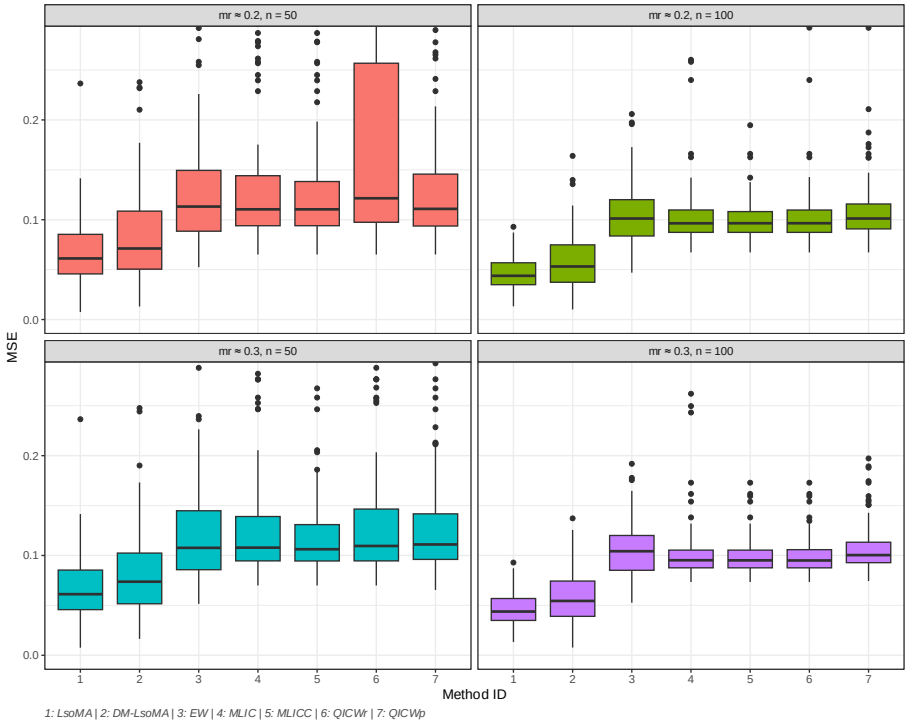

Figure 4. Boxplots of MSE under different methods (for $R(\rho)$ AR (1) Structure, $\rho = 0.5$)
图 4. 不同方法下 MSE 的箱线图($R(\rho)$ 为 AR (1)结构，$\rho = 0.5$)

我们将提出的 DM-LsoMA 方法与模型平均方法：不可行的、基于完全数据(complete data)的 LsoMA [3]，等权重(Equal Weight, EW)和模型选择方法：QICWp [10]，QICWr [11]，MLIC 及 MLICC [12]进行比较。采用均方误差 $\mathrm{MSE} = \frac{1}{BT}\sum_{b=1}^{B}\left\|\hat{\boldsymbol{\mu}}^{(b)} - \boldsymbol{\mu}\right\|^2$ 作为评判标准，其中 B (取 100)为实验重复次数，对第 b 次实验，$\hat{\boldsymbol{\mu}}^{(b)}$ 为 $\boldsymbol{\mu}$ 的估计量，$b = 1, \cdots, B$。

为对比分析不同方法的性能，我们绘制了相应的 MSE 箱线图，结果如图 1~4 所示。由图示结果可见，当真实相关矩阵分别为 EX 结构与 AR (1)结构时，模型平均方法始终表现出比模型选择方法更优的性能。所提出的 DM-LsoMA 方法相较于 EW 方法具有显著优势，且与完整数据下的 LsoMA 结果差异较小，尤其在缺失率较低时更为明显。此外，在 ρ 和缺失率保持不变的条件下，随着个体数 n 增加，所有方法的 MSE 均呈现下降趋势，并且表现更加稳定，尤为值得关注的是，DM-LsoMA 与 LsoMA 之间的 MSE 差距随之逐渐缩小，这充分表明 DM-LsoMA 的优势随个体数增大而愈发显著。综上所述，当数据存在退出型缺失时，DM-LsoMA 方法展现出优良的有限样本性能，可为相关预测问题提供有效工具。

另外，当样本量 n 较大时，本文所使用的 LsoCV 准则可能会产生较高的计算成本。为此，后续可借鉴 Zhang 和 Liu [13]的思路引入 K-折交叉验证准则，从而有效缩减计算量并提高运算效率。

4. 实证分析

本节我们将所提方法应用于原发性胆汁性肝硬化(PBC)数据集，该数据集可通过 R 语言包“joineRML”获取。该数据集涉及 312 例患者，研究初期记录了 age (登记时年龄)、gender (性别，男性与女性)、drug (药物类型，安慰剂与 D-青霉胺)等基线变量，并重复测量了多种生物标志物数据，如 serbili (血清胆红素)、alkaline (碱性磷酸酶)、prothrombin (凝血酶原)等。本研究剔除重复测量次数不足 8 次及存在数据缺失的

患者，最终纳入 89 例患者，共 952 条记录用于分析。

PBC 是一种慢性疾病，其特征为肝脏内小胆管的炎症性破坏，最终导致肝硬化并引发死亡。患者常表现为血液检测异常，如血清胆红素水平升高且呈渐进性上升趋势。因此，分析胆红素水平的时间变化规律具有重要医学意义。本研究以 log (serbili)为响应变量，参照 Bo 和 Zhang [14]的研究方法，选取 age、gender、drug、log (prothrombin)及 log (alkaline)共 5 个协变量构建候选模型。采用 EX 结构作为工作相关矩阵，通过对上述 5 个协变量进行组合，共构建 $2^5-1=31$ 个候选模型。

另外，考虑退出缺失模型如下：

$$\log\left(\frac{\lambda_{it}}{1-\lambda_{it}}\right)=\theta_0+age_i\theta_1+gender_i\theta_2+drug_i\theta_3+\log\left(prothrombin_{it}\right)\theta_4 +\log\left(alkaline_{it}\right)\theta_5+y_{i(t-1)}\theta_6+y_{i(t-2)}\theta_7+y_{i(t-3)}\theta_8,$$

其中 $i=1,\cdots,89$，$t=2,\cdots,T_i$，设定参数 $\boldsymbol{\theta}=(-0.1,0.9,1.2,1.3,0.5,-0.5,-0.1,0.2,0.1)'$ 使得退出缺失率约为 0.3。

将数据随机分为训练集和测试集，利用训练集进行参数估计，然后基于测试集评估不同方法的预测能力。假设训练集包含 $n_0=0.6n,0.7n$ 例患者的数据，测试集包含剩余 $n-n_0$ 例患者的数据，重复实验 $B=100$ 次，并计算均方预测误差(MSPE)：$\text{MSPE}=\frac{1}{B}\sum_{b=1}^{B}\frac{1}{N_1}\sum_{i=n_0+1}^{n}\sum_{t=1}^{T_i}\left(\hat{\mu}_{it}^{(b)}(\hat{\boldsymbol{\omega}})-y_{it}\right)^2$，其中 $N_1=\sum_{i=n_0+1}^{n}T_i$，$\hat{\mu}_{it}^{(b)}$ 为 μ_{it} 的估计。

表 1 给出了 7 种方法对应的 MSPE 的均值。从表 1 可以看出，在各类情形下，本文所提 DM-LsoMA 方法的预测精度均优于其他对照方法，仅略低于基于完全数据的 LsoMA 方法，这一结果与模拟分析结论一致。说明在存在退出型数据缺失时，DM-LsoMA 能够有效利用缺失数据信息，其预测性能显著优于 EW、MLIC、MLICC、QICWr 及 QICWp 等方法。

Table 1. Mean of MSPE under different methods
表 1. 不同方法下 MSPE 的均值

	LsoMA	DM-LsoMA	EW	MLIC	MLICC	QICWr	QICWp
$n_0=0.6n$	0.8460	0.8738	0.9075	0.8989	0.9057	0.9122	0.8818
$n_0=0.7n$	0.8719	0.8886	0.9166	0.9268	0.9285	0.9271	0.8919

5. 总结

本文研究了基于 LsoCV 准则的退出型缺失纵向数据模型平均估计，经模拟实验与实例分析验证了所提方法在预测精度与适用性上的优良表现，表明该方法可以有效利用缺失数据信息，并为相关预测问题提供有效工具。未来还可进一步考虑高维纵向数据的情形(Zhao 和 Zuo [4])，通过边际相关性实现协变量分组与候选模型构建，以此降低计算成本。此外，本文方法亦可拓展至广义线性模型，借鉴 Yu 等[8]的思路，以 KL 散度度量损失，并通过二阶近似去个体交叉验证(SEAL)准则选择权重，从而提高计算效率，并从理论上分析其渐近最优性。

参考文献

[1] Liang, K. and Zeger, S.L. (1986) Longitudinal Data Analysis Using Generalized Linear Models. *Biometrika*, **73**, 13-22. https://doi.org/10.1093/biomet/73.1.13

[2] Pan, W. (2001) Akaike's Information Criterion in Generalized Estimating Equations. *Biometrics*, **57**, 120-125. https://doi.org/10.1111/j.0006-341x.2001.00120.x

[3] Gao, Y., Zhang, X., Wang, S. and Zou, G. (2016) Model Averaging Based on Leave-Subject-Out Cross-Validation. *Journal of Econometrics*, **192**, 139-151. https://doi.org/10.1016/j.jeconom.2015.07.006

[4] Zhao, Z. and Zou, G. (2020) Average Estimation of Semiparametric Models for High-Dimensional Longitudinal Data. *Journal of Systems Science and Complexity*, **33**, 2013-2047. https://doi.org/10.1007/s11424-020-9343-1

[5] Hu, G., Cheng, W. and Zeng, J. (2019) Focused Information Criterion and Model Averaging for Varying-Coefficient Partially Linear Models with Longitudinal Data. *Communications in Statistics—Simulation and Computation*, **50**, 2399-2417. https://doi.org/10.1080/03610918.2019.1609029

[6] Li, N., Fei, Y. and Zhang, X. (2024) Partial Linear Model Averaging Prediction for Longitudinal Data. *Journal of Systems Science and Complexity*, **37**, 863-885. https://doi.org/10.1007/s11424-024-2187-3

[7] Jiang, B., Lv, J., Li, J. and Cheng, M. (2024) Robust Model Averaging Prediction of Longitudinal Response with Ultra-high-Dimensional Covariates. *Journal of the Royal Statistical Society Series B*: *Statistical Methodology*, **87**, 337-361. https://doi.org/10.1093/jrsssb/qkae094

[8] Yu, D., Zhang, X. and Liang, H. (2025) Unified Optimal Model Averaging with a General Loss Function Based on Cross-Validation. *Journal of the American Statistical Association*, **120**, 2697-2708. https://doi.org/10.1080/01621459.2025.2487215

[9] Robins, J.M., Rotnitzky, A. and Zhao, L.P. (1995) Analysis of Semiparametric Regression Models for Repeated Outcomes in the Presence of Missing Data. *Journal of the American Statistical Association*, **90**, 106-121. https://doi.org/10.1080/01621459.1995.10476493

[10] Platt, R.W., Brookhart, M.A., Cole, S.R., Westreich, D. and Schisterman, E.F. (2013) An Information Criterion for Marginal Structural Models. *Statistics in Medicine*, **32**, 1383-1393. https://doi.org/10.1002/sim.5599

[11] Gosho, M. (2016) Model Selection in the Weighted Generalized Estimating Equations for Longitudinal Data with Dropout. *Biometrical Journal*, **58**, 570-587. https://doi.org/10.1002/bimj.201400045

[12] Shen, C.W. and Chen, Y.H. (2012) Model Selection for Generalized Estimating Equations Accommodating Dropout Missingness. *Biometrics*, **68**, 1046-1054. https://doi.org/10.1111/j.1541-0420.2012.01758.x

[13] Zhang, X. and Liu, C. (2023) Model Averaging Prediction by k-Fold Cross-Validation. *Journal of Econometrics*, **235**, 280-301. https://doi.org/10.1016/j.jeconom.2022.04.007

[14] Bo, X. and Zhang, W. (2023) Subgroup Analysis for Longitudinal Data via Semiparametric Additive Mixed Effects Model. *Journal of Systems Science and Complexity*, **36**, 2155-2185. https://doi.org/10.1007/s11424-023-2011-5