Advances in Applied Mathematics 应用数学进展, 2021, 10(3), 790-800
Published Online March 2021 in Hans. http://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2021.103087

Hans汉斯

二元逻辑回归模型中的一阶近似刀切Liu估计

邹　媛

贵州民族大学，数据科学与信息工程学院，贵州　贵阳
Email: 1336718033@qq.com

收稿日期：2021年2月25日；录用日期：2021年3月23日；发布日期：2021年3月30日

摘　要

为了解决二元逻辑回归模型中的复共线性问题，我们结合一阶近似Liu估计和刀切法的优点提出了一个新的估计即一阶近似刀切Liu估计。研究得出了新估计偏差的优良性以及在均方误差矩阵、均方误差准则下优于一阶近似极大似然估计、一阶近似Liu估计和一阶近似刀切岭估计的充要或充分条件。更进一步使用了蒙特卡罗模拟和实证分析来探讨一阶近似刀切Liu估计偏差和在均方误差意义下的优良性。

关键词

二元逻辑回归模型，复共线性，一阶近似刀切Liu估计，偏差

A First-Order Approximated Jackknifed Liu Estimator in Binary Logistic Regression Model

Yuan Zou

School of Data Science and Information Engineering, Guizhou Minzu University, Guiyang Guizhou
Email: 1336718033@qq.com

Received: Feb. 25th, 2021; accepted: Mar. 23rd, 2021; published: Mar. 30th, 2021

Abstract

In order to solve the problem of multicollinearity in the binary logistic regression model, we combine the advantages of the first-order approximated Liu estimator and the jackknife procedure, and propose a new estimator, namely the first-order approximated jackknifed Liu estimator. The research obtained the sufficient and necessary or sufficient conditions for the new estimator to be

文章引用: 邹媛. 二元逻辑回归模型中的一阶近似刀切 Liu 估计[J]. 应用数学进展, 2021, 10(3): 790-800.
DOI: 10.12677/aam.2021.103087

superior to the first-order approximated maximum likelihood estimator, the first-order approximated Liu estimator and the first-order approximated jackknifed ridge estimatior under the bias, mean square error matrix or mean square error criterion. Furthermore, Monte Carlo simulation and empirical analysis are used to explore the first-order approximated jackknifed Liu estimator's performance in the sense of bias and mean square error.

Keywords

Binary Logistic Regression Model, Multicollinearity, First-Order Approximated Jackknifed Liu Estimator, Bias

1. 引言

逻辑回归模型是生物统计学和健康科学中常用的二元数据建模方法。这种模型有时也称为概率模型，因为给定一组协变量，事件发生的概率可以估计。二元逻辑回归模型的基本假设有：模型的响应变量 y 的分量 y_i 是相互独立且服从 Bernoulli (π_i)分布，其中：

$$\pi_i = \frac{\exp\left(x_i'\beta\right)}{1+\exp\left(x_i'\beta\right)},\quad i=1,\cdots,n \tag{1}$$

x_i' 是 $n\times p$ 样本资料矩阵 X 的第 i 行元素组成的向量。$\beta=\left(\beta_1,\cdots,\beta_p\right)'$ 为 $p\times 1$ 的系数向量。$\pi_i = P\left(y_i=1\middle|x_i\right)$ 是在 x_i 的条件下 $y_i=1$ 的概率。

在逻辑回归模型中，一般使用极大似然(ML)方法来估计回归参数 β，模型(1)的对数似然函数为：

$$L\left(\beta\right)=\sum_{i=1}^{n}\left[y_i x_i'\beta-\ln\left(1+\exp\left(x_i'\beta\right)\right)\right] \tag{2}$$

对等式(2)进行求导并令其等于 0 求 $L\left(\beta\right)$ 的极大值：

$$\frac{\partial L\left(\beta\right)}{\partial\beta}=X'\left(y-\pi\right)=0 \tag{3}$$

由于等式(3)不是线性的，因此 ML 估计是通过 Newton-Raphson 方法求解方程组(3)而得。用 Newton-Raphson 方法给出了 β 一个数值解：

$$\hat{\beta}^{(m)}=\hat{\beta}^{(m-1)}+\left(X'\hat{V}^{(m-1)}X\right)^{-1}X'\left(y-\hat{\pi}_i^{(m-1)}\right) \tag{4}$$

在上述迭代运算中，$\hat{\beta}^{(m-1)}$ 是 β 的第 $m-1$ 次迭代估计的向量，$\hat{\pi}_i^{(m-1)}$ 是 $\hat{\pi}^{(m-1)}$ 的第 i 个元素，$\hat{V}^{(m-1)}=diag\left(\hat{\pi}_i^{(m-1)}\right)\left(1-\hat{\pi}_i^{(m-1)}\right)$ 是权重矩阵。当 $\left|\hat{\beta}^{(m)}-\hat{\beta}^{(m-1)}\right|<\delta$ 时收敛，运算终止，其中 δ 为事先给定的计算精度，求得 $\hat{\beta}^{(m)}$ 为极大似然估计 $\hat{\beta}$ 的近似解 $\hat{\beta}_{MLE}$：

$$\hat{\beta}_{MLE}=\left(X'\hat{V}X\right)^{-1}X'\hat{V}\hat{z} \tag{5}$$

其中，$\hat{z}=X\hat{\beta}+\hat{V}^{-1}\left(y-\hat{\pi}\right)$，$\hat{V}$ 是一个对角矩阵且第 i 个对角元素为 $\hat{\pi}_i\left(1-\hat{\pi}_i\right)$，$\hat{\pi}$ 为收敛后的取值。

在逻辑回归模型中，当 $X'\hat{V}X$ 的一些特征值很小时，极大似然估计(MLE)的方差会膨胀，可能导致符号与现实情况不符，统计推断可能出现错误。为了克服这个问题，学者们提出了很多有偏估计来改进 MLE。例如，Schaefer 等[1]提出了岭估计(RE)：

$$\hat{\beta}(k)=\left(X\hat{V}X+kI\right)^{-1}X\hat{V}X\hat{\beta}_{MLE},\ k>0 \tag{6}$$

k 为岭参数。

Månsson 等[2]提出的 Liu 估计(LE)，表达式如下：

$$\hat{\beta}(d)=\left(X'\hat{V}X+I\right)^{-1}\left(X'\hat{V}X+dI\right)\hat{\beta}_{MLE},\ 0<d<1 \tag{7}$$

为了减小复共线性的影响，基于 Newton-Raphson 方法，LeCessie 和 Van Houwelingen [3]提出了与 Schaefer 等人[1]提出的岭估计渐近等价的一阶近似岭估计(FAR)，表达式如下：

$$\hat{\beta}^{(1)}(k)=\left(X\hat{V}^{(0)}X+kI\right)^{-1}X\hat{V}^{(0)}X\hat{\beta}^{(1)}(ML),\ k>0 \tag{8}$$

其中 $\hat{V}^{(0)}$ 是真实参数值 β_0 估计的权重矩阵。$\hat{\beta}^{(1)}(ML)$ 是由等式(4)所得的一阶近似极大似然估计(FAE)，表达式为：

$$\hat{\beta}^{(1)}(ML)=\left(X\hat{V}^{(0)}X\right)^{-1}X\hat{V}^{(0)}\hat{z}^{(0)}$$

Özkale [4]提出了一阶近似 Liu 估计(FAL)，表达式如下：

$$\hat{\beta}^{(1)}(d)=\left(X'\hat{V}^{(0)}X+I\right)^{-1}\left(X'\hat{V}^{(0)}X+dI\right)\hat{\beta}^{(1)}(ML),\ 0<d<1 \tag{9}$$

为了减小估计的偏差，Quenouille [5]和 Tukey [6]提出了刀切法。它的基本思想是利用一种特殊的方法来处理实验数据，从而得到一个未知参数的统计估计量。即系统地从数据中去除每个观测值后重新计算估计量，再将这些估计量取平均值。在二元逻辑回归模型中，结合一阶近似岭估计和刀切法，Özkale 和 Arıcan [7]提出了一阶近似刀切岭估计(FAJR)。表达式如下：

$$\tilde{\beta}^{(1)}(k)=\left(I-k^2\left(X\hat{V}^{(0)}X+kI\right)^{-2}\right)\hat{\beta}^{(1)}(ML),\ k>0 \tag{10}$$

2. 提出的估计

在本文中，我们结合一阶近似 Liu 估计和刀切法，提出了一个新的估计即一阶近似刀切 Liu 估计。接下来我们应用刀切法来定义一阶近似刀切 Liu 估计。

当 X 和 y 的第 i 个观测值删除时一阶近似 Liu 估计的表达式为：

$$\hat{\beta}_{-i}^{(1)}(d)=\left(X'_{-i}\hat{V}_{-i}^{(0)}X_{-i}+I\right)^{-1}\left(X'_{-i}\hat{V}_{-i}^{(0)}\hat{z}_{-i}^{(0)}+d\hat{\beta}_{-i}^{(1)}(ML)\right) \tag{11}$$

其中，$X'_{-i}\hat{V}_{-i}^{(0)}X_{-i}=X'\hat{V}^{(0)}X-x_i\hat{v}_i^{(0)}x_i'$，$X'_{-i}\hat{V}_{-i}^{(0)}\hat{z}_{-i}^{(0)}=X'\hat{V}^{(0)}\hat{z}^{(0)}-x_i\hat{v}_i^{(0)}\hat{z}_i^{(0)}$。化简可得：

$$\begin{aligned}\hat{\beta}_{-i}^{(1)}(d)=&\hat{\beta}^{(1)}(d)-\frac{1}{1-h_{ii}}\left(X'\hat{V}^{(0)}X+I\right)^{-1}x_i\hat{v}_i^{(0)}\left(\hat{z}_i^{(0)}-x_i'\hat{\beta}^{(1)}(d)\right)\\&-d\frac{1}{(1-h_{ii})(1-h_i)}\left(X'\hat{V}^{(0)}X+I\right)^{-1}x_i\hat{v}_i^{(0)}x_i'\left(X'\hat{V}^{(0)}X+I\right)^{-1}\\&\cdot\left(X'\hat{V}^{(0)}X\right)^{-1}x_i\hat{v}_i^{(0)}\left(\hat{z}_i^{(0)}-x_i'\hat{\beta}^{(1)}(ML)\right)\end{aligned} \tag{12}$$

其中 $h_{ii}=\hat{v}_i^{(0)}x_i'\left(X'\hat{V}^{(0)}X+I\right)^{-1}x_i$，$h_i=\hat{v}_i^{(0)}x_i'\left(X'\hat{V}^{(0)}X\right)^{-1}x_i$。

根据 Hinkley [8]我们可以得出加权伪值：

$$Q_i = \hat{\beta}^{(1)}(d) + n(1-h_{ii})\left(\hat{\beta}^{(1)}(d) - \hat{\beta}_{-i}^{(1)}(d)\right) \tag{13}$$

和加权伪值相对应的加权刀切估计：

$$\tilde{\beta}^{(1)}(d) = n^{-1}\sum Q_i \tag{14}$$

根据等式(12)，(13)，(14)，$\sum_{i=1}^{n} x_i \hat{v}_i^{(0)} \hat{z}_i^{(0)} = X'\hat{V}^{(0)}\hat{z}^{(0)}$ 和 $\sum_{i=1}^{n} x_i \hat{v}_i^{(0)} x_i' = X'\hat{V}^{(0)}X$ 我们在逻辑回归模型中定义了一个新的估计即一阶近似刀切 Liu 估计(FAJL)，表达式为：

$$\begin{aligned}\tilde{\beta}^{(1)}(d) = &\left\{\left(I - \left(X'\hat{V}^{(0)}X + I\right)^{-1} X'\hat{V}^{(0)}X\right)\left(X'\hat{V}^{(0)}X + I\right)^{-1}\left(X'\hat{V}^{(0)}X + dI\right)\right.\\ &\left.+\left(X'\hat{V}^{(0)}X + I\right)^{-1} X'\hat{V}^{(0)}X\right\}\hat{\beta}^{(1)}(ML)\end{aligned} \tag{15}$$

3. 一阶近似刀切 Liu 估计的性质

为了方便讨论一阶近似刀切 Liu 估计的性质，我们对矩阵 $\Phi = X'V^{(0)}X$ 进行特征分解，可以表示为 $\Phi = X'\hat{V}^{(0)}X = T\Lambda T'$，这里 $\Lambda = T'\Phi T = Z'\hat{V}^{(0)}Z = diag(\lambda_j)$ 是由矩阵 $X'\hat{V}^{(0)}X$ 的特征值组成的对角矩阵，且 $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$。$T = (T_1 \cdots T_p)$ 是由矩阵 $X'\hat{V}^{(0)}X$ 的特征值所对应的标准化特征向量组成的 $p \times p$ 阶正交矩阵。$Z = XT$，$\alpha = T'\beta$。为了方便对所提出的新估计与其他估计进行比较，我们首先定义参数 β 的估计 $\hat{\beta}$ 的偏差和偏差的平方和分别为：

$$bias(\hat{\beta}) = E(\hat{\beta}) - \beta \tag{16}$$

$$\left\|Bias(\hat{\beta})\right\|^2 = bias(\hat{\beta})'\, bias(\hat{\beta}) \tag{17}$$

均方误差矩阵：

$$MSEM(\hat{\beta}) = E(\hat{\beta} - \beta)(\hat{\beta} - \beta)' \tag{18}$$

均方误差：

$$MSE(\hat{\beta}) = E(\hat{\beta} - \beta')(\hat{\beta} - \beta) \tag{19}$$

为了在均方误差矩阵准则下对一阶近似极大似然估计、一阶近似刀切岭估计、一阶近似 Liu 估计和一阶近似刀切 Liu 估计进行比较，我们使用到了如下引理：

引理 1 (Farebrother [9])设 A 是一个 $n \times n$ 阶的正定矩阵，c 是一个 $n \times 1$ 阶非零列向量，u 是正的标量。如果 $c'A^{-1}c < u$，则 $uA - cc'$ 是正定的。

定理 1. $\left\|Bias(\tilde{\beta}^{(1)}(d))\right\|^2 < \left\|Bias(\hat{\beta}^{(1)}(d))\right\|^2$。

证明：

$$bias(\hat{\beta}^{(1)}(d)) = (d-1)\left(X'\hat{V}^{(0)}X + I\right)^{-1} a^0 \tag{20}$$

$$bias(\tilde{\beta}^{(1)}(d)) = (d-1)\left(X'\hat{V}^{(0)}X + I\right)^{-2} a^0 \tag{21}$$

则

$$\left\|Bias\left(\hat{\beta}^{(1)}(d)\right)\right\|^2-\left\|Bias\left(\tilde{\beta}^{(1)}(d)\right)\right\|^2=\left(a^0\right)'G_1a^0$$

其中 $G_1=(d-1)^2\left(X\hat{V}^{(0)}X+I\right)^{-2}-(d-1)^2\left(X\hat{V}^{(0)}X+I\right)^{-4}$，$G_1$ 是对角元素为 $\frac{(d-1)^2\left((\lambda_i+1)^2-1\right)}{(\lambda_i+1)^4}$ 的对角矩阵，λ_i 是矩阵 $X\hat{V}^{(0)}X$ 的第 i 个特征值，所以 $\frac{(d-1)^2\left((\lambda_i+1)^2-1\right)}{(\lambda_i+1)^4}\geq 0$，即 $\left(a^0\right)'G_1a^0$ 是正定的。定理得证。

定理 2. 若 $k^4(\lambda_i+1)^4-(d-1)^2(\lambda_i+k)^4>0$ 则 $\left\|Bias\left(\tilde{\beta}^{(1)}(d)\right)\right\|^2<\left\|Bias\left(\tilde{\beta}^{(1)}(k)\right)\right\|^2$。

证明：由等式(16)得，一阶近似刀切岭估计的偏差为：

$$bias\left(\tilde{\beta}^{(1)}(k)\right)=-k^2\left(X\hat{V}^{(0)}X+kI\right)^{-2}a^0 \tag{22}$$

则

$$\left\|Bias\left(\tilde{\beta}^{(1)}(k)\right)\right\|^2-\left\|Bias\left(\tilde{\beta}^{(1)}(d)\right)\right\|^2=\left(a^0\right)'G_2a^0$$

其中 $G_2=k^4\left(X\hat{V}^{(0)}X+kI\right)^{-4}-(d-1)^2\left(X\hat{V}^{(0)}X+I\right)^{-4}$，$G_2$ 是对角元素为 $\frac{k^4(\lambda_i+1)^4-(d-1)^2(\lambda_i+k)^4}{(\lambda_i+1)^4(\lambda_i+k)^4}$ 的对角矩阵，所以当 $k^4(\lambda_i+1)^4-(d-1)^2(\lambda_i+k)^4>0$ 时 $\frac{k^4(\lambda_i+1)^4-(d-1)^2(\lambda_i+k)^4}{(\lambda_i+1)^4(\lambda_i+k)^4}>0$，即 $\left(a^0\right)'G_2a^0$ 是正定的。定理得证。

定理 3. 当 $0<d<1$ 时，一阶近似刀切 Liu 估计在 MSEM 准则下优于一阶近似极大似然估计当且仅当

$$(1-d)\left(a^0\right)'\left\{2\Lambda+4I+(d+1)\Lambda^{-1}\right\}^{-1}a^0<1。$$

证明：令 $M_1(d)=MSEM\left(\hat{\beta}^{(1)}(ML)\right)-MSEM\left(\tilde{\beta}^{(1)}(d)\right)$，由公式(18)可得：

$$M_1(d)=M_1-(d-1)^2(\Lambda+I)^{-2}a^0\left(a^0\right)'(\Lambda+I)^{-2}$$

其中

$$\begin{aligned}M_1=&\Lambda^{-1}-\left\{\left(I-(\Lambda+I)^{-1}\Lambda\right)(\Lambda+I)^{-1}(\Lambda+dI)+(\Lambda+I)^{-1}\Lambda\right\}\Lambda^{-1}\\&\cdot\left\{(\Lambda+dI)(\Lambda+I)^{-1}\left(I-(\Lambda+I)^{-1}\Lambda\right)+\Lambda(\Lambda+I)^{-1}\right\}\\=&(1-d)(\Lambda+I)^{-2}\left\{2\Lambda+4I+(d+1)\Lambda^{-1}\right\}(\Lambda+I)^{-2}\end{aligned}$$

因此 $M_1(d)$ 是正定的当且仅当 $M_1-(d-1)^2(\Lambda+I)^{-2}a^0\left(a^0\right)'(\Lambda+I)^{-2}$ 正定。由引理 1 可得，当 $0<d<1$，$(1-d)\left(a^0\right)'\left\{2\Lambda+4I+(d+1)\Lambda^{-1}\right\}^{-1}a^0<1$ 时，$M_1-(d-1)^2(\Lambda+I)^{-2}a^0\left(a^0\right)'(\Lambda+I)^{-2}$ 是正定的。定理得证。

定理 4. 一阶近似刀切 Liu 估计在 MSEM 准则下优于一阶近似 Liu 估计当且仅当

$$(d-1)\left(a^0\right)'\left\{2\Lambda^2+(d+3)\Lambda+2dI\right\}^{-1}a^0<1。$$

证明：令 $M_2(d)=MSEM\left(\hat{\beta}^{(1)}(d)\right)-MSEM\left(\tilde{\beta}^{(1)}(d)\right)$，由公式(18)可得：

$$M_2(d)=M_2+(d-1)^2(\Lambda+I)^{-1}a^0\left(a^0\right)'(\Lambda+I)^{-1}-(d-1)^2(\Lambda+I)^{-2}a^0\left(a^0\right)'(\Lambda+I)^{-2}$$

其中

$$\begin{aligned}M_2=&\left\{(\Lambda+I)^{-1}(\Lambda+dI)\right\}\Lambda^{-1}\left\{(\Lambda+dI)(\Lambda+I)^{-1}\right\}-\left\{\left(I-(\Lambda+I)^{-1}\Lambda\right)(\Lambda+I)^{-1}(\Lambda+dI)\right.\\&\left.+(\Lambda+I)^{-1}\Lambda\right\}\Lambda^{-1}\left\{(\Lambda+dI)(\Lambda+I)^{-1}\left(I-(\Lambda+I)^{-1}\Lambda\right)+\Lambda(\Lambda+I)^{-1}\right\}\\=&(d-1)(\Lambda+I)^{-2}\left\{2\Lambda^2+(d+3)\Lambda+2dI\right\}(\Lambda+I)^{-2}\end{aligned}$$

因为 $(d-1)^2(\Lambda+I)^{-1}a^0\left(a^0\right)'(\Lambda+I)^{-1}$ 是正定的，因此 $M_2(d)$ 是正定的当且仅当 $M_2-(d-1)^2(\Lambda+I)^{-2}a^0\left(a^0\right)'(\Lambda+I)^{-2}$ 正定。由引理 1 可得，当 $0<d<1$，$(d-1)\left(a^0\right)'\left\{2\Lambda^2+(d+3)\Lambda+2dI\right\}^{-1}a^0<1$ 时，$M_2-(d-1)^2(\Lambda+I)^{-2}a^0\left(a^0\right)'(\Lambda+I)^{-2}$ 是正定的。定理得证。

定理 5. 对于任意的 i，如果 $0<d\le\dfrac{\left(a_i^0\right)^2-(2\lambda_i+3)/(\lambda_i+2)}{\left(a_i^0\right)^2+1/\lambda_i}<1$，则 $MSE\left(\tilde{\beta}^{(1)}(d)\right)\le MSE\left(\hat{\beta}^{(1)}(d)\right)$。

证明：根据等式(19)可得出：

$$MSE\left(\tilde{\beta}^{(1)}(d)\right)=\sum_{i=1}^{p}\frac{\left(\lambda_i^2+2\lambda_i+d\right)^2+(d-1)^2\left(a_i^0\right)^2\lambda_i}{\lambda_i(\lambda_i+1)^4}\tag{23}$$

$$MSE\left(\hat{\beta}^{(1)}(d)\right)=\sum_{i=1}^{p}\frac{\left(\lambda_i+d\right)^2+(d-1)^2\left(a_i^0\right)^2\lambda_i}{\lambda_i(\lambda_i+1)^2}\tag{24}$$

它们的差：

$$\begin{aligned}&MSE\left(\tilde{\beta}^{(1)}(d)\right)-MSE\left(\hat{\beta}^{(1)}(d)\right)\\=&\sum_{i=1}^{p}\frac{\left(\lambda_i^2+2\lambda_i+d\right)^2+(d-1)^2\left(a_i^0\right)^2\lambda_i}{\lambda_i(\lambda_i+1)^4}-\sum_{i=1}^{p}\frac{\left(\lambda_i+d\right)^2+(d-1)^2\left(a_i^0\right)^2\lambda_i}{\lambda_i(\lambda_i+1)^2}\\=&\sum_{i=1}^{p}\left\{\frac{\left(\lambda_i^2+2\lambda_i+d\right)^2-(\lambda_i+1)^2+(d-1)^2(\lambda_i+d)^2}{\lambda_i(\lambda_i+1)^4}+\frac{(d-1)^2\left(a_i^0\right)^2}{(\lambda_i+1)^4}-\frac{(d-1)^2\left(a_i^0\right)^2}{(\lambda_i+1)^2}\right\}\\=&(1-d)\sum_{i=1}^{p}\frac{2\lambda_i^2+(d+3)\lambda_i+2d}{(\lambda_i+1)^4}+(d-1)^2\sum_{i=1}^{p}\frac{\left(a_i^0\right)^2}{(\lambda_i+1)^4}-(d-1)^2\sum_{i=1}^{p}\frac{\left(a_i^0\right)^2}{(\lambda_i+1)^2}\\=&(1-d)\sum_{i=1}^{p}\frac{1}{(\lambda_i+1)^4}f_i(d)\end{aligned}$$

其中 $f_i(d)=\lambda_i(2\lambda_i+3)-\alpha_i^0\lambda_i(\lambda_i+2)+d(\lambda_i+2)\left(1+\lambda_i\left(a_i^0\right)^2\right)$。

当 $1<\left(a_i^0\right)^2(\lambda_i+2)/(2\lambda_i+3)$，$d\le\dfrac{\left(a_i^0\right)^2-(2\lambda_i+3)/(\lambda_i+2)}{\left(a_i^0\right)^2+1/\lambda_i}$ 时 $f_i(d)\le0$。因为 $0<d<1$ 故 $MSE\left(\tilde{\beta}^{(1)}(d)\right)-MSE\left(\hat{\beta}^{(1)}(d)\right)\le0$。定理得证。

由定理 5，我们可以得出如下两个推论：

推论 1. 假设

$$0<d\le \min\left\{\frac{\left(a_i^0\right)^2-\left(2\lambda_i+3\right)/\left(\lambda_i+2\right)}{\left(a_i^0\right)^2+1/\lambda_i}\right\}<1$$

则 $MSE\left(\tilde{\beta}(d)\right)\le MSE\left(\hat{\beta}(d)\right)$。

推论 2. 假设

$$\max\left\{0,\frac{\left(a_i^0\right)^2-\left(2\lambda_i+3\right)/\left(\lambda_i+2\right)}{\left(a_i^0\right)^2+1/\lambda_i}\right\}<d<1$$

则 $MSE\left(\tilde{\beta}^{(1)}(d)\right)>MSE\left(\hat{\beta}^{(1)}(d)\right)$。

4. 蒙特卡罗模拟

为了进一步对理论成果进行说明，针对不同的复共线性程度及不同的自相关程度，本节我们用 Monte Carlo 模拟方法探讨上述各类估计在偏差和均方误差准则下的优良性。解释变量的数据产生采用与 McDonald 和 Galarneau [10]和 Kibria [11]相同的方法，即由以下方程生成：

$$x_{ij}=\left(1-\rho^2\right)^{1/2}z_{ij}+\rho z_{ip+1},\ i=1,\cdots,n;j=1,\cdots,p \tag{25}$$

其中，z_{ij} 是标准正态随机变量产生的随机数；ρ 是给定的常数；ρ^2 表示两个不同解释变量之间的相关性，因而 ρ^2 某种程度上体现了模型复共线性的程度。在模拟实验中，我们取协变量的数目 $p=4$ 和 $p=6$，样本数 n 考虑 100、150 和 200 三种情况，ρ 考虑 0.85、0.9、0.95 和 0.99 四种不同的情况。偏参数 d 我们考虑取 0.1、0.3、0.5、0.7 和 0.9 五种不同的取值。

响应变量对应的随机数来自伯努利分布 $Be\left(\pi_i\right)$，其中 $\pi_i=\dfrac{\exp\left(x_i'\beta\right)}{1+\exp\left(x_i'\beta\right)}$。对于系数向量 β，采用与 Kibria [11]相同的方法，对其做一定的限制，使其满足 $\sum_{j=1}^{p}\beta_j^2=1$。本次模拟重复 2000 次。估计的 MSE 可以通过以下式子得到：

$$MSE\left(\hat{\beta}\right)=\frac{1}{2000}\sum_{m=1}^{2000}tr\left(MSEM\left(\hat{\beta}_{(m)}\right)\right) \tag{26}$$

其中 $\hat{\beta}_{(m)}$ 是估计 $\hat{\beta}$ 的第 m 次所得的估计值。模拟结果见表 1~表 4。

观察表 1 和表 2 可以看到，在不同复共线性程度、样本量、协变量的数目和偏参数 d 的情况下，一阶近似刀切 Liu 估计的均方误差值小于极大似然估计和一阶近似极大似然估计的均方误差值，即一阶近似刀切 Liu 估计在均方误差准则下优于极大似然估计和一阶近似极大似然估计。同时由表 1 和表 2 可以看出，当偏参数 d 取 0.1 时一阶近似刀切 Liu 估计的均方误差值小于 d 取 0.3、0.5、0.7 和 0.9 时一阶近似刀切 Liu 估计的均方误差值。当固定给定的 d、n 和 p 值时，各估计的均方误差值随着复共线性程度 ρ 的增大而增大。当固定给定的 d、n 和 ρ 值时，各估计的均方误差值随着协变量的数目 p 的增大而增大。当固定给定的 d、p 和 ρ 值时，各估计的均方误差值随着样本量 n 的增大而减小。

Table 1. Estimated MSE values of the MLE, FAE and FAJL when $p = 4$
表 1. 当 $p = 4$ 时，估计 MLE、FAE 和 FAJL 的 MSE

d		MLE	FAE	FAJL				
				0.1	0.3	0.5	0.7	0.9
ρ	0.85							
	100	1.3927	1.2171	1.0357	1.0733	1.1125	1.1532	1.1954
n	150	0.8523	0.7907	0.7255	0.7395	0.7538	0.7684	0.7832
	200	0.6223	0.5851	0.5551	0.5616	0.5683	0.5751	0.5819
ρ	0.9							
	100	2.0610	1.8175	1.3578	1.4478	1.5448	1.6487	1.7595
n	150	1.2784	1.1843	0.9980	1.0366	1.0768	1.1186	1.1620
	200	0.9237	0.8726	0.7810	0.8005	0.8205	0.8410	0.8620
ρ	0.95							
	100	4.1806	3.6570	1.9873	2.2635	2.5939	2.9735	3.4173
n	150	2.5668	2.3768	1.5618	1.7121	1.8800	2.0655	2.2687
	200	1.8695	1.7673	1.2976	1.3890	1.4878	1.5940	1.7077
ρ	0.99							
	100	22.4064	19.5469	5.6987	6.5321	8.6477	12.0457	16.7260
n	150	13.6913	12.6788	4.0059	4.7671	6.1946	8.2885	11.0488
	200	9.9752	9.4003	3.2597	3.9126	4.9722	6.4384	8.3113

Table 2. Estimated MSE values of the MLE, FAE and FAJL when $p = 6$
表 2. 当 $p = 6$ 时，估计 MLE、FAE 和 FAJL 的 MSE

d		MLE	FAE	FAJL				
				0.1	0.3	0.5	0.7	0.9
ρ	0.85							
	100	2.6568	2.2046	1.7519	1.8419	1.9381	2.0403	2.1484
n	150	1.5989	1.4216	1.2474	1.2838	1.3215	1.3606	1.4009
	200	1.1273	1.0448	0.9611	0.9790	0.9973	1.0160	1.0351
ρ	0.9							
	100	4.1114	3.3823	2.2859	2.4872	2.7127	2.9624	3.2363
n	150	2.4261	2.1629	1.6946	1.7871	1.8862	1.9919	2.1042
	200	1.7291	1.5959	1.3494	1.4000	1.4530	1.5084	1.5661
ρ	0.95							
	100	8.5367	7.0108	3.3480	3.8997	4.6013	5.4527	6.4540
n	150	5.0352	4.5013	2.6422	2.9568	3.3277	3.7549	4.2384
	200	3.5709	3.3047	2.1915	2.3932	2.6210	2.8749	3.1549
ρ	0.99							
	100	46.5859	38.2250	9.5522	11.3917	15.8210	22.8402	32.4493
n	150	27.3406	24.4286	6.6708	8.2228	11.1429	15.4311	21.0874
	200	19.5496	18.0425	5.4978	6.7696	8.9077	11.9120	15.7824

Table 3. The sum of squares of the bias values of the MLE, FAE and FAJL when $p = 4$
表 3. 当 $p = 4$ 时，估计 FAL 和 FAJL 的偏差的平方和

d		0.1		0.3		0.5		0.7		0.9	
		FAL	FAJL	FAL	FAJL	FAL	FAJL	FAL	FAJL	FAL	FAJL
ρ	0.85										
	100	0.0315	0.0040	0.0190	0.0024	0.0097	0.0012	0.0035	0.0004	0.0003	0.0000
n	150	0.0110	0.0007	0.0066	0.0004	0.0033	0.0002	0.0012	0.0000	0.0001	0.0000
	200	0.0049	0.0002	0.0030	0.0001	0.0015	0.0000	0.0005	0.0000	0.0000	0.0000
ρ	0.9										
	100	0.0849	0.0810	0.0513	0.0109	0.0262	0.0055	0.0094	0.0020	0.0010	0.0002
n	150	0.0313	0.0040	0.0189	0.0024	0.0096	0.0012	0.0034	0.0004	0.0038	0.0000
	200	0.0152	0.0013	0.0092	0.0007	0.0046	0.0004	0.0016	0.0001	0.0001	0.0000
ρ	0.95										
	100	0.3589	0.1426	0.2171	0.0863	0.1107	0.0440	0.0398	0.0158	0.0044	0.0017
n	150	0.1593	0.0464	0.0964	0.0281	0.0491	0.0143	0.0177	0.0051	0.0019	0.0005
	200	0.0889	0.0199	0.0538	0.0120	0.0274	0.0061	0.0098	0.0022	0.0010	0.0002
ρ	0.99										
	100	4.4535	3.4889	2.6941	2.1106	1.3745	1.0768	0.4948	0.3876	0.0549	0.0430
n	150	2.5537	1.8134	1.5448	1.0969	0.7881	0.5596	0.2837	0.2014	0.0315	0.0223
	200	1.7017	1.1097	1.0294	0.6713	0.5252	0.3425	0.1890	0.1233	0.0210	0.0137

Table 4. The sum of squares of the bias values of the MLE, FAE and FAJL when $p = 6$
表 4. 当 $p = 6$ 时，估计 FAL 和 FAJL 的偏差的平方和

d		0.1		0.3		0.5		0.7		0.9	
		FAL	FAJL	FAL	FAJL	FAL	FAJL	FAL	FAJL	FAL	FAJL
ρ	0.85										
	100	0.0683	0.0134	0.0413	0.0081	0.0210	0.0041	0.0075	0.0014	0.0008	0.0001
n	150	0.0246	0.0029	0.0148	0.0017	0.0075	0.0008	0.0027	0.0003	0.0003	0.0000
	200	0.0118	0.0009	0.0071	0.0005	0.0036	0.0002	0.0013	0.0001	0.0001	0.0000
ρ	0.9										
	100	0.1748	0.0541	0.1057	0.0327	0.0539	0.0167	0.0194	0.0060	0.0021	0.0006
n	150	0.0717	0.0152	0.0434	0.0092	0.0221	0.0047	0.0079	0.0016	0.0008	0.0001
	200	0.0343	0.0051	0.0208	0.0031	0.0106	0.0015	0.0038	0.0005	0.0004	0.0000
ρ	0.95										
	100	0.6484	0.3266	0.3922	0.1976	0.2001	0.1008	0.0720	0.0362	0.0080	0.0040
n	150	0.3162	0.1267	0.1913	0.0766	0.0976	0.0391	0.0351	0.0140	0.0039	0.0015
	200	0.1779	0.0582	0.1076	0.0352	0.0549	0.0179	0.0197	0.0064	0.0021	0.0007
ρ	0.99										
	100	7.2163	5.8941	4.3654	3.5656	2.2272	1.8191	0.8018	0.6549	0.0890	0.0727
n	150	4.0267	3.0332	2.4359	1.8349	1.2428	0.9361	0.4474	0.3370	0.0497	0.0374
	200	2.7304	1.9266	1.6517	1.1655	0.8427	0.5946	0.3033	0.2140	0.0337	0.0237

根据表 3 和表 4 可知，在不同复共线性程度、样本量、协变量的数目和偏参数的情况下，一阶近似刀切 Liu 估计的偏差的平方和始终小于一阶近似 Liu 估计的偏差的平方和。且当偏参数 d 取 0.9 时一阶近似 Liu 估计和一阶近似刀切 Liu 估计的偏差的平方和小于 d 取 0.1、0.3、0.5 和 0.7 时一阶近似 Liu 估计和一阶近似刀切 Liu 估计的偏差的平方和。当固定给定的 d、n 和 p 值时，各估计的偏差的平方和随着复共线性程度 ρ 的增大而增大。当固定给定的 d、n 和 ρ 值时，各估计的偏差的平方和随着协变量的数目 p 的增大而增大。当固定给定的 d、p 和 ρ 值时，各估计的偏差的平方和随着样本量 n 的增大而减小。

5. 实证分析

为了验证我们的理论结果，这部分我们考虑实例来分析所提出估计的优良性。我们所使用的数据来自 Agresti Alan [12]。数据涉及 Heinze 和 Schemper [13]所描述的一种关于子宫内膜癌的研究。分析了 79 个案例，因变量 y 为组织学分级(低分级 y 取 0，高分级时 $y=1$)，其中低分级的病人有 30 个，高分级患者有 49 个。涉及的三个风险因素：x_1 为新血管生成 (有：$x_1=1$，无：$x_1=0$)，x_2 为子宫动脉搏动指数(取值范围在 0 到 49 之间)，x_3 为子宫内膜高度(取值范围在 0.27 到 3.61 之间)。

迭代的计算精度 δ 我们取 10^{-6}，得到矩阵 $X\hat{V}X$ 的特征值 $\lambda_1=3068.0790$，$\lambda_2=7.1753$，$\lambda_3=0.3069$，$\lambda_4=1.2496\times10^{-7}$。条件数 $\kappa=\sqrt{\lambda_{\max}/\lambda_{\min}}=156687.8$，因此可以判断数据集存在严重的复共线性问题。

为了对我们所提的新估计一阶近似刀切 Liu 估计的优良性进行研究。我们得到极大似然估计、一阶近似极大似然估计和一阶近似刀切 Liu 估计的均方误差值，一阶近似 Liu 估计和一阶近似刀切 Liu 估计偏差的平方值和，如表 5 和表 6：

Table 5. Estimated MSE values of the MLE, FAE and FAJL
表 5. 估计 MLE、FAE 和 FAJL 的 MSE

d	0.1	0.3	0.5	0.7	0.9
MLE	8002097	8002097	8002097	8002097	8002097
FAE	84.2875	84.2875	84.2875	84.2875	84.2875
FAJL	7.4009	13.1352	25.3562	44.0637	69.2580

Table 6. The sum of squares of the bias values of the FAL and FAJL when p = 6
表 6. 估计 FAL 和 FAJL 的偏差的平方和

d	0.1	0.3	0.5	0.7	0.9
FAL	5.4893	3.3207	1.6942	0.6099	0.0677
FAJL	4.6307	2.8013	1.4292	0.5145	0.0571

通过表 5 我们可以看出，对于给定的 d 值，新估计一阶近似刀切 Liu 估计的均方误差值小于极大似然估计和一阶近似极大似然估计的均方误差值，且当 $d=0.1$ 时一阶近似刀切 Liu 估计的均方误差值 $MSE\left(\tilde{\beta}^{(1)}(d)\right)=7.4009$ 最小。再由表 6 我们可以看到一阶近似刀切 Liu 估计偏差的平方和小于一阶近似 Liu 估计偏差的平方和，且当 $d=0.9$ 时一阶近似刀切 Liu 估计偏差的平方和取值最小 $\left\|Bias\left(\tilde{\beta}^{(1)}(d)\right)\right\|^2=0.0571$，同时对我们所得理论结果定理 2 进行了验证。

6. 结论

本文中，针对二元逻辑回归模型中的复共线性问题，我们在一阶近似 Liu 估计的基础上使用刀切法

的思想提出了一个新估计，即一阶近似刀切 Liu 估计。研究了一阶近似刀切 Liu 估计的偏差以及在均方误差矩阵和均方误差准则下的优良性。证明并得出了新估计的偏差平方和总是优于一阶近似 Liu 估计以及新估计优于一阶近似刀切岭估计的充分条件，得出了一阶近似刀切 Liu 估计在均方误差矩阵、均方误差准则下优于一阶近似极大似然估计、一阶近似 Liu 估计和一阶近似刀切岭估计的充要或者充分条件。此外，我们使用蒙特卡罗模拟，得到了一阶近似刀切 Liu 估计在均方误差准则下优于极大似然估计和一阶近似极大似然估计，各估计的均方误差值随着复共线性程度 ρ 的增大而增大，各估计的均方误差值随着协变量的数目 p 的增大而增大，各估计的均方误差值随着样本量 n 的增大而减小。然后利用实证分析探讨了一阶近似刀切 Liu 估计在实际应用中的实现问题，证明一阶近似刀切 Liu 估计能够有效地解决复共线性问题。

参考文献

[1] Schaefer, R.L., Roi, L.D. and Wolfe, R.A. (1984) A Ridge Logistic Estimator. *Communications in Statistics-Theory and Methods*, **13**, 99-113. https://doi.org/10.1080/03610928408828664

[2] Månsson, K., Golam Kibria, B.M. and Shukur, G. (2012) On Liu Estimators for the Logit Regression Model. *Economic Modelling*, **29**, 1483-1488. https://doi.org/10.1016/j.econmod.2011.11.015

[3] LeCessie, S. and VanHouwelingen, J.C. (1992) Ridge Estimators in Logistic Regression. *Journal of Applied Statistics*, **41**, 191-201. https://doi.org/10.2307/2347628

[4] Revan Özkale, M. (2016) Iterative Algorithms of Biased Estimation Methods in Binary Logistic Regression. *Statistical Papers*, **57**, 991-1016. https://doi.org/10.1007/s00362-016-0780-9

[5] Quenouille, M.H. (1956) Notes on Bias in Estimation. *Biometrika*, **43**, 353-360. https://doi.org/10.1093/biomet/43.3-4.353

[6] Tukey, J.W. (1958) Bias and Confidence in Not Quite Large Samples (Abstract). *Annals of Mathematical Statistics*, **29**, 614. https://doi.org/10.1214/aoms/1177706635

[7] Revan Özkale, M. and Arıcan, E. (2019) A First-Order Approximated Jackknifed Ridge Estimator in Binary Logistic Regression. *Computational Statistics*, **34**, 683-712. https://doi.org/10.1007/s00180-018-0851-6

[8] Hinkley, V. (1977) Jackknifing in Unbalanced Situations. *Technometrics*, **19**, 285-292. https://doi.org/10.1080/00401706.1977.10489550

[9] Farebrother, R.W. (1976) Further Results on the Mean Square Error of Ridge Regression. *Journal of the Royal Statistical Society B*, **28**, 248-250. https://doi.org/10.1111/j.2517-6161.1976.tb01588.x

[10] Gary, C., McDonald, D. and, Galarneau, I. (1975) A Monte Carlo Evaluation of Some Ridge-Type Estimators. *Journal of the American Statistical Association*, **70**, 407-416. https://doi.org/10.1080/01621459.1975.10479882

[11] Kibria, B.M.G. (2003) Performance of Some New Ridge Regression Estimators. *Communications in Statistics Simulation and Computation*, **32**, 419-435. https://doi.org/10.1081/SAC-120017499

[12] Agresti, A. (2015) Foundations of Linear and Generalized Linear Models. Wiley, Hoboken.

[13] Heinze, G. and Schemper, M. (2002) A Solution to the Problem of Separation in Logistic Regression. *Statistics in Medicine*, **21**, 2409-2419. https://doi.org/10.1002/sim.1047