第45卷 第3期
2026年9月

《新疆师范大学学报》(自然科学版)
Journal of Xinjiang Normal University
(Natural Sciences Edition)

Vol.45,No.3
Sept. 2026

Poisson回归模型的变量选择

张文静,董翠玲*
(新疆师范大学 数学科学学院,新疆 乌鲁木齐 830017)

摘 要: Poisson回归模型作为处理计数数据的核心工具,其变量选择精度直接影响模型的解释性与预测性能。文章应用R软件,通过数值模拟及共享单车租赁数据的实证分析,对比研究了LASSO、岭回归、Elastic Net、SCAD、Bridge和Adaptive Bridge六种变量选择方法在Poisson回归中的优劣。结果表明,SCAD方法在模型拟合优度和预测精度上表现最佳,适用于大样本和高相关性数据;Adaptive Bridge方法在模型复杂度与预测稳定性方面取得良好平衡。本研究为计数数据在实际中的应用提供了兼具精度与稳定性的建模工具。

关键词: 广义线性模型;Poisson回归模型;变量选择方法;SCAD;Adaptive Bridge

中图分类号: O212.1 **文献标识码:** A **文章编号:** 1008-9659(2026)03-0102-11

广义线性模型(Generalized Linear Model,GLM)在生物、医学、经济学和社会学中具有广泛应用,是线性模型的直接推广,它不仅能有效处理连续数据,也适用于离散型数据。Poisson回归模型是研究计数响应变量与其他自变量之间相互关系的一种广义线性模型,它假设响应变量服从Poisson分布,通过链接函数将条件均值与自变量的线性组合相关联,从而有效处理计数数据的回归分析问题。其响应变量常常为事件发生的次数,例如,疾病发作次数、事故次数等,此类计数数据通常具有非负整数值且方差与均值接近的特点。

随着现代化数据收集及存储能力的提高,变量的个数越来越多。在利用Poisson回归模型对计数数据进行建模时,尤其当解释变量中包含冗余变量或者变量间存在多重共线性时,可能导致模型过拟合,从而降低模型的解释能力和预测精度。因此,如何从众多变量中筛选出有效变量成为亟待解决的问题。

近年来,多种变量选择方法被应用于广义线性模型中,这些方法通过引入惩罚项实现变量选择和模型正则化,从而提高模型的预测性能和解释性。王倩等人[1]对比最小绝对收缩与选择算子(Least Absolute Shrinkage and Selection Operator,LASSO)、弹性网(Elastic Net)和带平滑削边绝对偏离法(Smoothly Clipped Absolute Deviation,SCAD)在Logistic回归模型的应用,发现SCAD方法能准确地将非重要变量的回归系数压缩为零。夏亚峰等人研究高维数据下广义线性模型的参数估计和变量选择问题,提出一种基于对数似然函数和自适应桥(Adaptive Bridge)的估计方法,验证了Adaptive Bridge估计量的相合性和Oracle性质[2]。对于带有测量误差Poisson回归模型的变量选择问题,刘芳提出了一种惩罚偏差校正方法,有效解决了变量选择和参数估计问题[3]。杨宜平等人研究了高维部分线性模型的变量选择问题[4]。卢改改采用稳健Poisson回归模型研究了中国31个主要城市空气质量的影响因素[5]。孙龙等人利用Poisson回归模型和负二项回归模型对大兴安岭地区林火发生次数进行预测[6]。卢颖研究了广义线性模型中基于Elastic Net的变量选择方法[7]。

本研究采用数值模拟的方法将LASSO、岭回归、Elastic Net、SCAD、Bridge和Adaptive Bridge六种变量选择方法应用于Poisson回归模型,从真阳性率(TPR)、假阳性率(FPR)、拟合优度(R^2)、均方误差(MSE)等方面进行对比分析,探究各方法的优势与局限,并将其应用于UCI数据库Capital共享单车租赁数据中,从14个自变量中选取影响共享单车租赁量的关键变量,并建立较精准的预测模型。

[收稿日期]2025-09-03 [修回日期]2025-09-29
[基金项目]新疆维吾尔自治区自然科学基金项目(2023D01A37)。
[作者简介]张文静(2000-),女,硕士研究生,主要从事计数数据的分析与应用方面研究,E-mail:1760810981@qq.com.
* [通讯作者]董翠玲(1978-),女,副教授,主要从事多变点的统计推断、复杂数据分析方面研究,E-mail:75894804@qq.com.

1 Poisson回归模型的定义

Poisson回归模型是一种广义线性模型，用于处理因变量为计数数据的情况。设$\{Y_i, X_i\}_{i=1}^n$为n个独立观测样本，其中响应变量Y_i为n维列向量，是计数数据，表示n个观测事件的发生次数，且服从Poisson分布，即

$$P(Y_i = y_i|X_i) = \frac{\lambda^{y_i}e^{-\mu}}{y_i!}, \quad y_i = 0, 1, 2, \cdots \tag{1}$$

自变量$X_i = (X_{i1}, X_{i2}, \cdots, X_{ip})^T$为$p$维向量$(i = 1, 2, \cdots, n)$.

模型假设[8]：

(1)独立性假设：设各个观测事件的发生次数Y_i相互独立；

(2)Poisson分布假设：设因变量Y为计数数据，且$Y_i \sim Poisson(\mu)$；

(3)对数线性假设：观测事件发生次数的对数条件期望与自变量呈线性关系，即$\ln\left(E(Y_i|X_i)\right) = \beta_0 + X_i^T\beta$.

Poisson回归模型通过对数链接函数建立条件均值与自变量的线性关系，其结构为

$$\ln\left(E(Y_i|X_i)\right) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} = \beta_0 + X_i^T\beta \tag{2}$$

其中，$E(Y_i|X_i) = \mu$为条件期望，$\beta = (\beta_1, \beta_2, \cdots, \beta_p)^T$为回归系数向量，$\beta_0$为截距项。

由假设(1)观测值之间相互独立，则n个观测值的似然函数为

$$L(\beta_0, \beta) = \prod_{i=1}^{n}\frac{\lambda^{y_i}e^{-\mu}}{y_i!} \tag{3}$$

其对数似然函数为

$$l(\beta_0, \beta) = \sum_{i=1}^{n}[y_i\ln(\mu) - \mu - \ln(y_i!)] \tag{4}$$

由于$\ln(y_i!)$不包含待估参数β，式(4)可以进一步简化为

$$l(\beta_0, \beta) = \sum_i^n(y_i(\beta_0 + X_i^T\beta) - exp(\beta_0 + X_i^T\beta)) \tag{5}$$

2 变量选择方法

变量选择通过惩罚机制将冗余变量的系数压缩为零，实现模型稀疏化与预测性能的提升。文章对比了LASSO、岭回归、Elastic Net、SCAD、Bridge和Adaptive Bridge六种变量选择方法，在对数似然函数$l(\beta_0, \beta)$基础上引入惩罚项构建目标函数，其中正则化参数$\lambda \geqslant 0$控制惩罚强度。

Tibshirani于1996年针对线性回归模型提出了LASSO方法[9]。该方法是在对数似然函数的基础上加入L_1惩罚项$(\sum_{j=1}^{p}|\beta_j| < c, c$为常数$)$，将部分冗余变量的系数压缩为零，实现变量选择和模型的稀疏化。虽然计算高效，但是LASSO方法是有偏的，且不具备Oracle性质，可能会过度收缩非零系数[10]。回归系数β的估计形式为

$$\hat{\beta}_{Lasso} = arg\,min_{\beta_0,\beta}\left\{-l(\beta_0, \beta) + \lambda\sum_{j=1}^{p}|\beta_j|\right\} \tag{6}$$

Hoerl和Kennard于1970年提出了岭回归(Ridge Regression)[11]。该方法在对数似然函数的基础上加入L_2惩罚项$(\sum_{j=1}^{p}\beta_j^2 < c, c$为常数$)$，使回归系数的估计值向零收缩，但不会将系数精确压缩为零，即岭回归能压缩系数但无法实现变量选择，常用于预测，适用于变量之间存在多重共线性的问题，同样计算高效但有偏且不具备稀疏性和Oracle性质，模型解释性较差[10]。回归系数β的估计形式为

$$\hat{\beta}_{ridge} = arg\,min_{\beta_0,\beta}\left\{-2l(\beta_0, \beta) + \lambda\sum_{j=1}^{p}\beta_j^{\,2}\right\} \tag{7}$$

Zou等人于2005年提出了弹性网(Elastic Net)方法[12]。该方法结合了L_1和L_2惩罚项的优点,能同时实现变量选择和处理变量间的多重共线性问题。L_1惩罚项使模型具有变量选择的能力,类似于LASSO方法,能够将冗余变量的系数压缩为零,具备稀疏性;L_2惩罚项能有效处理变量之间的多重共线性,通过对系数的平方进行约束避免系数估计值的方差过大,提高模型的稳定性,但Elastic Net方法是有偏的,不具备Oracle性质,模型复杂度较高[10]。回归系数β的估计形式为

$$\hat{\beta}_{net} = arg\ min_{\beta_0,\beta}\left\{-l(\beta_0,\beta) + \lambda\left[\alpha\sum_{j=1}^{p}|\beta_j| + (1-\alpha)\sum_{j=1}^{p}\beta_j^2\right]\right\} \tag{8}$$

其中,$\alpha \in [0,1]$为混合参数,当$\alpha = 1$时,Elastic Net退化为LASSO,当$\alpha = 0$时,Elastic Net退化为岭回归。

Fan等人在2001年提出了SCAD方法[13]。该方法通过非凸惩罚函数,对较小的系数惩罚力度较大,加速冗余变量的系数向零收缩。对于较大的系数惩罚力度放缓,避免对重要变量系数过度收缩,确保在变量选择过程中既保证稀疏性,又解决LASSO和Elastic Net等方法的有偏性,具备Oracle性质,从而保证变量选择的稳定性和准确性。但是因为其惩罚函数的复杂性,导致计算复杂度增加[10]。回归系数β的估计形式为

$$\hat{\beta}_{SCAD} = arg\ min_{\beta_0,\beta}\left\{-l(\beta_0,\beta) + \sum_{j=1}^{p}\phi\left(|\beta_j|;\lambda,\gamma\right)\right\} \tag{9}$$

其中,$\phi\left(|\beta_j|;\lambda,\gamma\right)$为非凸惩罚函数,形式为

$$\phi\left(|\beta_j|;\lambda,\gamma\right) = \begin{cases} \lambda|\beta_j|, & |\beta_j| < \lambda \\ \dfrac{-|\beta_j|^2 + 2\gamma\lambda - \lambda^2}{2\gamma - 2}, & \lambda < |\beta_j| \leqslant \gamma\lambda \\ \dfrac{(\gamma+1)\lambda^2}{2}, & |\beta_j| > \gamma\lambda \end{cases}$$

其中,$\gamma > 2$是一个固定参数,用于确定惩罚函数的形状。

Frank等人于1993年提出Bridge回归[14]。该方法结合了L_1和L_2惩罚的思想,用$\sum_{j}^{p}|\beta_j|^{\gamma}$作为惩罚项对系数进行约束。$\gamma > 0$是调节稀疏程度的参数,决定了惩罚项对系数的压缩方式,γ越大对大系数惩罚越大;当$0 < \gamma < 1$时,随着γ的减小,模型的解会具有更高的稀疏性,甚至优于LASSO回归,并具有渐近Oracle性质;当$\gamma = 1$时,Bridge回归退化为LASSO回归,强调变量选择和模型的稀疏性;当$\gamma > 1$时,Bridge回归计算高效,估计值虽不具有稀疏性但具有渐近无偏性;当$\gamma = 2$时,Bridge回归退化为岭回归,更注重模型的稳定性和系数估计的准确性[10]。回归系数β的估计形式为

$$\hat{\beta}_{Bridge} = arg\ min_{\beta_0,\beta}\left\{-l(\beta_0,\beta) + \lambda\sum_{j}^{p}|\beta_j|^{\gamma}\right\}, \gamma > 0 \tag{10}$$

Zou和Hastie于2005年提出了Adaptive Bridge方法[15]。该方法对Bridge回归的系数引入自适应权重,使得惩罚项能够根据变量的重要性进行自适应调整。当$\gamma \leqslant 1$时,估计值的偏差和稀疏性都优于LASSO,并具备渐近Oracle性质,但其系数估计的准确性依赖初始值,且计算复杂度因需迭代更新权重而增加[10]。回归系数β的估计形式为

$$\hat{\beta}_{Ad\text{-}Bridge} = arg\ min_{\beta_0,\beta}\left\{-l(\beta_0,\beta) + \lambda\sum_{j=1}^{p}\omega_j|\beta_j|^{\gamma}\right\} \tag{11}$$

其中,$\gamma > 0$,$\sum_{j=1}^{p}\omega_j|\beta_j|^{\gamma}$是对Bridge回归系数$\beta_j$进行加权的约束惩罚项,$\beta_j$的初始值$\hat{\beta}_j$可通过Bridge回归获得,权重$\omega_j = 1/\left|\hat{\beta}_j\right|^{\delta}$,$\delta > 0$是常数,决定$\omega_j$对$\hat{\beta}_j$大小的敏感程度,$\delta$越大,小的$\hat{\beta}_j$对应的权重$\omega_j$越大,惩罚越大。

3 数值模拟

运用R软件通过数值模拟的方法,设置不同样本量及变量间的相关程度。将上述六种变量选择方法应

用于Poisson回归模型，在100次独立重复试验中，通过变量选择的真阳性率（TPR）、假阳性率（FPR）、模型的拟合优度（R^2）、均方误差（MSE）等指标来评价这六种变量选择方法的稳定性与准确性，为确保模拟结果的可复现性，设定种子为“123”.

3.1 模拟设计与调节参数的选择

假设响应变量Y服从Poisson回归

$$\ln(E(Y|X)) = \ln\mu = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p \tag{12}$$

其中，$X = (X_1, X_2, \cdots, X_p)$为解释变量，且$X \sim N_p(0, \Sigma)$，通过Toeplitz协方差矩阵$\Sigma_{ij} = P^{|i-j|}$来精确控制变量间的相关性，避免独立同分布假设的局限性。为了区分实际数据中的信号与噪声，相关系数$P = 0$、$P = 0.3$、$P = 0.5$、$P = 0.7$分别刻画变量间独立、低相关、中等相关和高相关的情形。用样本量n分别为200、500、800来模拟小样本、中样本和大样本场景。

设变量总数$P = 20$，真实回归系数$\beta = (\beta_1, \cdots, \beta_{20}) = (1, 0, 0, 0, 0, -1, 0, 0, 1, -1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0)$，即第1、6、9、10、18非零，其余15个系数为零；截距项$\beta_0 = 0$，在模型拟合阶段仍包含截距项以保证统计规范性。实际拟合时，模型保留式（12）中的截距项β_0（用于捕捉数据随机波动），且所有评估指标仅针对$\beta_1 \sim \beta_{20}$，截距项偏差不影响核心结论。

采用交叉验证法选择正则化参数λ，保证计算效率以及稳健性。对于LASSO和Elastic Net应用R软件中glmnet包的坐标下降算法，通过十折交叉验证从预设的λ参数序列中选择最优λ；对于Bridge及Adaptive Bridge应用optim函数的L-BFGS-B算法遍历λ的候选值并选出最优λ；对于SCAD，应用ncvreg包的局部二次逼近算法，借鉴文献[16]，设定参数γ=3.7.

3.2 评估指标

通过真阳性率（TPR）、假阳性率（FPR）、模型的拟合优度（R^2）、均方误差（MSE）等指标评估变量选择方法的性能。

真阳性率（True Positive Rate，TPR）是指真实非零系数中被正确选中的比例，其计算公式为

$$TPR = \frac{TP\_mean}{m}$$

其中，*TP_mean*为正确选中的有效变量个数的均值，m为总有效变量的个数（本试验中*TP_mean*为100次重复试验中正确选中有效变量个数的均值，m=5）。

假阳性率（False Positive Rate，FPR）是指错误选中冗余变量的比例，其计算公式为

$$FPR = \frac{FP\_mean}{p - m}$$

其中，*FP_mean*为错误选中的冗余变量数量的均值，m为总冗余变量个数（本试验中*FP_mean*为100次重复试验中错误选中冗余变量个数的均值，$p-m$=20−5=15）。

均方误差（Mean Squared Error，MSE）是模拟预测值与真实值之差平方的平均值，用来量化预测值和真实值之间的平均偏差大小，值越小说明拟合效果越好。其计算公式为

$$MSE = \frac{1}{n}\sum_{i}^{n}(y_i - \hat{y}_i)^2$$

模型的拟合优度（R^2）是衡量回归模型对预测数据拟合程度的统计指标，用于量化因变量的总变异中可被自变量通过回归模型解释的比例。其计算公式为

$$R^2 = 1 - \frac{\sum_{i}^{n}(y_i - \hat{y}_i)^2}{\sum_{i}^{n}(y_i - \bar{y}_i)^2}$$

其中，$\hat{y}_i$为模型的预测值，y_i为观测值，$\bar{y}_i$为观测值的平均值。

绝对误差之和（Sum of Absolute Errors，SAE）是指估计值与真实值之间的绝对差异的总和，用于衡量模型估计值与真实值之间的偏离程度，其计算公式为

$$SAE = \sum_{j=1}^{p}\left|\beta_j - \hat{\beta}_j\right|$$

其中,β_j为真实系数,$\hat{\beta}_j$为模型系数的估计值。

3.3 模拟结果

在上述模拟设计与调节参数的选择下,表1至表4展示了样本量分别为$n = 200, n = 500, n = 800$以及变量间相关性分别为$P = 0, 0.3, P = 0.5, P = 0.7$时六种变量选择方法的TPR、FPR、MSE和$R^2$.

表1 $P = 0$时不同样本量下的变量选择方法性能对比

样本量	方法	TPR	FPR	MSE	R^2
$n = 200$	LASSO	1	0.40000	0.00251	0.95900
	岭回归	1	1.00000	0.02480	0.79100
	Elastic Net	1	0.73300	0.00248	0.96200
	SCAD	1	0.00000	0.01790	0.88400
	Bridge	1	0.33300	0.00331	0.94400
	Adaptive Bridge	1	0.60000	0.00221	0.96200
$n = 500$	LASSO	1	0.53300	0.00052	0.99000
	岭回归	1	1.00000	0.01850	0.86800
	Elastic Net	1	0.73300	0.00056	0.99000
	SCAD	1	0.00000	0.02050	0.92600
	Bridge	1	0.00000	0.01270	0.90700
	Adaptive Bridge	1	0.00000	0.01250	0.91400
$n = 800$	LASSO	1	0.46700	0.00020	0.97900
	岭回归	1	1.00000	0.00981	0.91700
	Elastic Net	1	0.53300	0.00021	0.97900
	SCAD	1	0.00000	0.02160	0.93100
	Bridge	1	0.06700	0.00170	0.96900
	Adaptive Bridge	1	0.00000	0.00192	0.96700

表2 $P = 0.3$时不同样本量下的变量选择方法性能对比

样本量	方法	TPR	FPR	MSE	R^2
$n = 200$	LASSO	1	0.26700	0.00068	0.96600
	岭回归	1	1.00000	0.03110	0.85500
	Elastic Net	1	0.40000	0.00073	0.96800
	SCAD	1	0.00000	0.08520	0.81700
	Bridge	1	0.46700	0.00069	0.96800
	Adaptive Bridge	1	0.06700	0.00051	0.96900
$n = 500$	LASSO	1	0.46700	0.00108	0.95700
	岭回归	1	1.00000	0.01520	0.85600
	Elastic Net	1	0.46700	0.00130	0.95500
	SCAD	1	0.00000	0.04010	0.72900
	Bridge	1	0.80000	0.00112	0.95700
	Adaptive Bridge	1	0.13300	0.00080	0.95800
$n = 800$	LASSO	1	0.66700	0.00054	0.95500
	岭回归	1	1.00000	0.00889	0.87300
	Elastic Net	1	1.00000	0.00064	0.95500
	SCAD	1	0.00000	0.06620	0.77300
	Bridge	1	0.93300	0.00050	0.95500
	Adaptive Bridge	1	0.06700	0.00052	0.95400

表3 $P = 0.5$时不同样本量下的变量选择方法性能对比

样本量	方法	TPR	FPR	MSE	R^2
$n = 200$	LASSO	1	0.53300	0.00280	0.93100
	岭回归	1	1.00000	0.02020	0.87100
	Elastic Net	1	0.46700	0.00431	0.92700
	SCAD	1	0.00000	0.00080	0.92700
	Bridge	1	0.26700	0.00425	0.92600
	Adaptive Bridge	1	0.06700	0.00126	0.92900
$n = 500$	LASSO	1	0.73300	0.00062	0.95900
	岭回归	1	1.00000	0.01860	0.87700
	Elastic Net	1	0.86700	0.00079	0.95900
	SCAD	1	0.00000	0.06150	0.78500
	Bridge	1	0.60000	0.00062	0.95900
	Adaptive Bridge	1	0.00000	0.00017	0.95800
$n = 800$	LASSO	1	0.60000	0.00043	0.98900
	岭回归	1	1.00000	0.03760	0.90700
	Elastic Net	1	0.73300	0.00064	0.98900
	SCAD	1	0.00000	0.02540	0.96600
	Bridge	1	0.13300	0.01020	0.94800
	Adaptive Bridge	1	0.00000	0.01780	0.93700

表4 $P = 0.7$时不同样本量下的变量选择方法性能对比

样本量	方法	TPR	FPR	MSE	R^2
$n = 200$	LASSO	1	0.53300	0.00285	0.91500
	岭回归	1	1.00000	0.02690	0.84000
	Elastic Net	1	0.66700	0.00315	0.91500
	SCAD	1	0.13300	0.00153	0.90900
	Bridge	1	0.53300	0.00291	0.91400
	Adaptive Bridge	1	0.26700	0.00168	0.91100
$n = 500$	LASSO	1	0.93300	0.00176	0.94900
	岭回归	1	1.00000	0.02480	0.89300
	Elastic Net	1	0.93300	0.00187	0.94900
	SCAD	1	0.20000	0.00125	0.94800
	Bridge	1	1.00000	0.00160	0.94800
	Adaptive Bridge	1	0.33300	0.00087	0.94900
$n = 800$	LASSO	1	0.33300	0.00187	0.91200
	岭回归	1	1.00000	0.02760	0.85700
	Elastic Net	1	0.53300	0.00215	0.91200
	SCAD	1	0.06700	0.00085	0.91400
	Bridge	1	0.40000	0.00183	0.91200
	Adaptive Bridge	1	0.33300	0.00073	0.91400

从表1~表4可以看出,六种方法的TPR均为1,说明这六种方法均能精准识别对模型有贡献的有效变量,提升模型的解释能力;随着样本量n的增大,FPR普遍降低,即在样本量增加的情况下,六种方法错误地将冗余变量纳入模型的可能性降低,有助于构建更为简洁准确的模型,其中SCAD在$n = 800$、$P = 0$时FPR降

至0,Adaptive Bridge在$n = 500$、$P = 0.3$时FPR仅为0.133,显著优于LASSO(0.467)和Elastic Net(0.467);随着样本量n的增加,所有方法的MSE均下降,R^2逐渐增大,即模型预测值与真实值之间的偏差逐渐减小,预测精度不断提高,反映出模型对数据的拟合效果越来越好。其中岭回归的TPR和FPR始终为1,无法变量选择,会将所有变量纳入模型,模型解释性最差。

随着变量间相关性P的增加,LASSO、岭回归、Elastic Net和Bridge方法的FPR显著上升,MSE增大,R^2减小,即在变量间存在强相关性时,这些方法难以剔除冗余变量,导致模型解释能力与预测精度下降。SCAD方法在所有相关性水平下均保持TPR = 1且FPR趋近于0,凸显了在复杂数据结构中变量选择的优势;Adaptive Bridge方法通过自适应权重机制,在高相关性场景中维持了较高的TPR与相对较低的FPR,体现了优异的稳健性和适应性;LASSO和Elastic Net具备变量选择能力,但在$P \geqslant 0.5$时,FPR显著上升(最高达0.933),导致模型冗余度增加。岭回归因保留全部变量(FPR = 1)而无法实现变量选择,虽能维持预测稳定性,但削弱了解释性;SCAD方法在精准变量选择方面表现卓越,适合处理高相关性数据;Adaptive Bridge方法则在模型复杂度与预测稳定性之间实现了更优平衡,在处理复杂相关性结构数据时具有优势。

以上为100次独立重复试验的平均效果。为了进一步对比六种变量选择方法对真实系数的估计效果,评估各方法在系数估计精度、变量选择能力和模型拟合效果方面的表现。表5展示了在样本量$n = 500$和高相关性$P = 0.7$情形下,在一次试验中,六种方法对模型(12)中回归系数的估计情况。

表5 真实系数与各个方法系数估计对比表($n = 500, P = 0.7$)

真实系数β	$\hat{\beta}_{\text{Lasso}}$	$\hat{\beta}_{\text{ridge}}$	$\hat{\beta}_{\text{net}}$	$\hat{\beta}_{\text{SCAD}}$	$\hat{\beta}_{\text{Bridge}}$	$\hat{\beta}_{\text{Ad-Bridge}}$
1	0.96500	0.85000	0.96400	0.98300	0.96400	0.96800
0	0.04570	0.08820	0.04700	0.00000	0.03720	0.03500
0	0.00000	0.06890	0.00000	0.00000	−0.000215	−0.000492
0	−0.02600	0.000839	−0.02580	0.00000	−0.01480	0.00000
0	−0.01380	−0.12700	−0.01590	0.00000	−0.01850	−0.02250
−1	−0.98100	−0.75800	−0.97700	−1.01000	−0.98000	−1.00000
0	−0.01320	−0.07970	−0.01500	0.00000	−0.01540	−0.02790
0	−0.07240	−0.03820	−0.07220	−0.08540	−0.06190	−0.03510
1	1.01000	0.71200	1.01000	1.02000	0.99200	1.01000
−1	−0.88700	−0.57900	−0.88300	−0.89600	−0.87300	−0.91700
0	−0.00774	−0.04780	−0.00886	0.00000	−0.01370	−0.02430
0	−0.03420	−0.01850	−0.03330	−0.06600	−0.03830	−0.00851
0	−0.05150	−0.04680	−0.05310	0.00000	−0.03380	−0.03020
0	0.00585	−0.12200	0.00484	0.00000	0.00469	0.00994
0	0.07670	0.11700	0.07890	0.03860	0.05760	0.05100
0	−0.02590	−0.05050	−0.02760	0.00000	−0.02150	−0.01470
0	−0.00835	0.03350	−0.00821	0.00000	−0.00615	−0.00000856
1	0.97000	0.76000	0.96700	0.99000	0.96800	0.97400
0	0.03160	0.11800	0.03330	0.00000	0.01780	0.01480
0	0.00445	0.10000	0.00625	0.00000	0.01170	0.00155
SAE	0.6229678	2.3988034	0.6449488	0.3496949	0.5761228	0.4313565

从表5可看出,SCAD方法的SAE最小,表现出最精准的变量选择能力,5个真实非零系数的估计均接近真实值,且15个零系数中有12个被精确识别,显著优于其他方法。Adaptive Bridge方法对非零系数的估计虽然没有完全识别为0,但其估计接近0,且SAE = 0.431仅次于SCAD,体现了自适应权重机制在平衡估计精度与稳定性上的优势;LASSO和Elastic Net仅识别了1个零系数,偏差较大,Bridge回归未识别出零系数,但某些零系数的估计接近0;岭回归所有零系数均有较大非零估计,无变量选择能力且偏差最大。

4 实证分析

选取UCI数据库中Capital共享单车系统中2011—2012年每天的共享单车租赁数量数据（https://archive.ics.uci.edu/dataset/275/bike+sharing+dataset）作为研究对象，该数据集包含731天共享单车租赁数量（cnt）的数据，涉及季节（Season）、天气状况（Weather）、温度（Temp）等既有离散型又有连续型变量的14个自变量。考虑数据库中2011—2012年日期变量（dteday），月份变量（mnth）具有周期性，与变量年份（Year）、季节变量（Spring/Summer/Fall/Winter）信息重合，删去日期变量（dteday）、月份变量（mnth）；随机租赁数据（casual）与注册租赁数据（registered）之和为本研究对象实际租赁量（cnt），删去随机租赁数据（casual）与注册租赁数据（registered）。

考虑到周期性、季节性、节假日、工作日、天气等对共享单车租赁数据的影响，将年份（Year）转化为二分类变量，季节变量（Spring/Summer/Fall/Winter）以Fall为基准，拆分为三个二分类变量，将节假日（Holiday）、工作日（Workingday）、周末（Weekend）转化为二分类变量

$$yr_{2012}=\begin{cases}1, & year=2012\\0, & year=2011\end{cases},\quad Season{:}Spring=\begin{cases}1, & Spring\\0, & else\end{cases},\quad Season{:}Summer=\begin{cases}1, & Summer\\0, & else\end{cases}$$

$$Season{:}Winter=\begin{cases}1, & Winter\\0, & else\end{cases},\quad Holiday=\begin{cases}1, & holiday\\0, & else\end{cases},\quad Workingday=\begin{cases}1, & \text{工作日}\\0, & \text{休息日}\end{cases},$$

$$Weekend=\begin{cases}0, & \text{周一至周五}\\1, & \text{周六、周日}\end{cases}$$

对天气状况（Weather）、温度（Temp）、体感温度（Atemp）、湿度（Hum）、风速（Wind Speed）不作变换，采用原始数据。以下是经上述变换后12个变量的Pearson相关系数矩阵热力图。

由图1可以看出，工作日（Workingday）与周末（Weekend）具有强负相关性（-0.93），温度（Temp）与体感温度（Atemp）具有强正相关性（0.99），其他变量间存在弱相关性。

为确保模型评估具有代表性以及结果的可复现性，设定种子为“5432”，在数据集中随机选取80%（共584个数据）为训练集，其余20%（共147个数据）为测试集。应用Poisson回归模型，并使用LASSO、Elastic Net、岭回归、SCAD、Bridge和Adaptive Bridge六种变量选择方法进行参数估计（表6）。

表6 六种变量选择方法下的参数估计结果

Variable	$\hat{\beta}_{\text{Lasso}}$	$\hat{\beta}_{\text{ridge}}$	$\hat{\beta}_{\text{net}}$	$\hat{\beta}_{\text{SCAD}}$	$\hat{\beta}_{\text{Bridge}}$	$\hat{\beta}_{\text{Ad-Bridge}}$
截距项：β_0	7.964715	7.984871	7.9677480	7.962680	7.984871	7.987690
X_1：年份（2012年）	0.450164	0.440475	0.449267	0.453556	0.440475	0.438107
X_2：季节（春季）	−0.245173	−0.247343	−0.248408	−0.263888	−0.247343	−0.245444
X_3：季节（夏季）	0.103738	0.100658	0.100719	0.095234	0.100658	0.095701
X_4：季节（冬季）	0.199035	0.191500	0.194463	0.185619	0.191500	0.185619
X_5：节假日（是）	−0.112087	−0.120623	−0.109382	−0.117273	−0.120623	−0.117702
X_6：工作日（是）	0.023152	0.019644	0.022503	0.023763	0.019644	0.016748
X_7：周末（是）	0.000000	−0.005295	0.000000	0.000000	−0.005295	−0.006012
X_8：天气情况	−0.168457	−0.163915	−0.168153	−0.167837	−0.163915	−0.157673
X_9：气温	0.801455	0.638250	0.774729	0.000000	0.638250	0.637721
X_{10}：体感温度	0.421342	0.587950	0.440912	1.294377	0.587950	0.589101
X_{11}：湿度	−0.213229	−0.228557	−0.208377	−0.230835	−0.228557	−0.235469
X_{12}：风速	−0.611029	−0.616243	−0.603708	−0.597963	−0.616243	−0.621541
测试集R^2	0.8691868	0.8694270	0.8692802	0.8723627	0.8694270	0.8702391

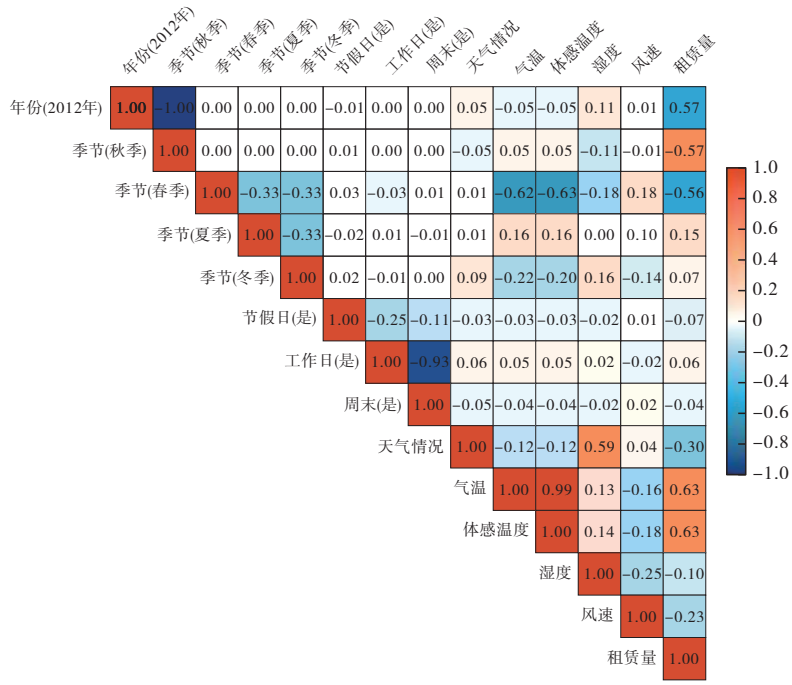

图1 Pearson相关系数矩阵热力图

表6展示了六种变量选择方法在共享单车租赁数据上的参数估计结果。SCAD方法将两个变量的系数压缩为0,且拟合优度值最大,表明SCAD方法能够有效剔除冗余变量,解释因变量的差异性,尤其在处理季节等分类变量时,能更精准地捕捉其对租赁量的影响,特别是在处理共线性变量(如Temp与Atemp)时表现突出;Adaptive Bridge方法拟合优度仅次于SCAD方法,而且对冗余变量(如workingdayWorking、is_weekendWeekend)的系数收缩明显,但没有压缩为0;LASSO和Elastic Net方法也具备变量选择能力,能将部分冗余变量压缩为0,模型解释性略低于SCAD和Adaptive Bridge方法;岭回归虽然在处理多重共线性问题上具有优势,但无法实现变量选择。

图2绘制了基于表6中不同Poisson回归模型在测试集上的预测值与真实租赁量的偏离程度$|y_i - \hat{y}_i|$,$i = 1, 2, \cdots 147$的箱线图(绝对误差$|y_i - \hat{y}_i|$越小,预测精度越高),各方法的绝对误差均值约在570. 其中SCAD方法的绝对误差均值为564.4,Adaptive Bridge方法的绝对误差均值为570.8,二者都相对较小,箱线图长度相对较短,说明SCAD方法和Adaptive Bridge方法得到的预测值相对较为精准。

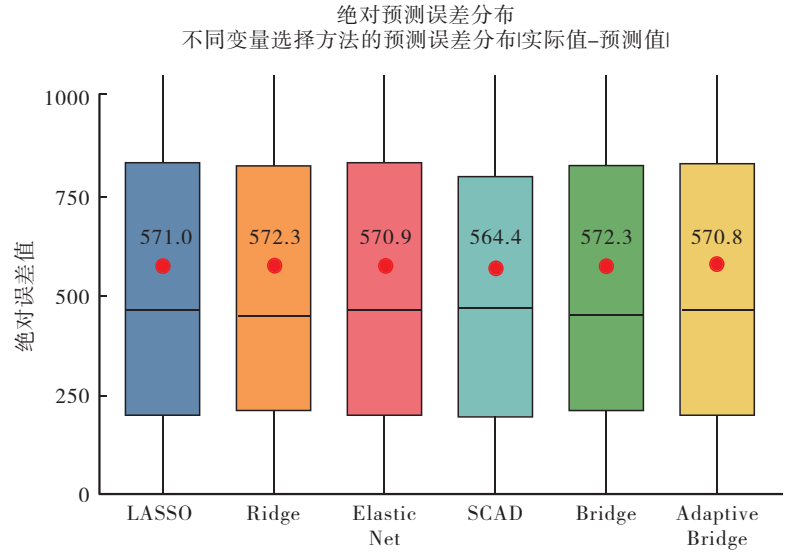

图2 六种方法在测试集上的绝对误差箱线图

基于上述分析，可采用SCAD方法建立共享单车租赁数据的泊松回归模型，其表达式为

$$\log(E(\mathrm{cnt})) = 7.9627 + 0.4536X_1 - 0.2639X_2 + 0.0952X_3 + 0.1856X_4 - 0.1173X_5 + 0.0238X_6 - 0.1678X_8 + 1.2944X_{10} - 0.2308X_{11} - 0.5980X_{12}$$

该模型在预测集和测试集上均有较好的表现。

5 结论

文章应用数值模拟和共享单车租赁量的实际数据，对比研究了LASSO、岭回归、Elastic Net、SCAD、Bridge和Adaptive Bridge六种变量选择方法在Poisson回归模型中的表现。模拟结果表明，SCAD方法处理大样本和高相关性数据时，变量选择和预测性能方面均表现卓越，能够精准识别重要变量并有效剔除冗余变量，展现出较高的真阳性率和较低的假阳性率；Adaptive Bridge方法通过自适应权重机制，在预测精度和模型稳定性之间取得了良好的平衡，其假阳性率显著低于传统方法，同时保持了与SCAD方法相近的预测精度；LASSO和Elastic Net方法倾向于保留更多的变量，导致模型复杂度增加，模型解释性较弱；岭回归由于保留所有变量而无法实现变量选择，模型解释性较差。在UCI数据库中Capital共享单车租赁数据上的实证分析进一步验证了SCAD方法在识别关键变量方面的优越性。因此，在面对计数数据建模需求时，若数据具有大样本量和高相关性特点，SCAD方法可以有效识别显著变量；当需要在模型简洁性与预测稳定性之间寻求平衡时，Adaptive Bridge方法也是理想的选择。本研究为计数数据在实际中的应用提供了兼具精度与稳定性的建模工具。

参考文献：

[1] 王倩，李风军. Logistic回归模型的变量选择[J]. 统计与决策，2021，37(16)：48-51.

[2] 夏亚峰，何佳. 高维数据下广义线性模型自适应桥惩罚估计的变量选择[J]. 甘肃科学学报，2022，34(01)：7-15.

[3] 刘芳. 带有测量误差Poisson回归模型的变量选择[J]. 赣南师范大学学报，2024，45(06)：44-47.

[4] 杨宜平，薛留根，王学娟. 高维部分线性模型中的变量选择[J]. 北京工业大学学报，2011，37(02)：291-295.

[5] 卢改改. 基于稳健Poisson回归的中国城市空气质量影响因素研究[D]. 南昌：江西财经大学，2023.

[6] 孙龙，尚喆超，胡海清. Poisson回归模型和负二项回归模型在林火预测领域的应用[J]. 林业科学，2012，48(05)：126-129.

[7] 卢颖. 广义线性模型基于Elastic Net的变量选择方法研究[D]. 北京：北京交通大学，2011.

[8] NELDER J A，WEDDERBURN R W M. Generalized Linear Models [J]. Journal of the Royal Statistical Society：Series A (General)，1972，135(03)：370-384.

[9] TIBSHIRANI R. Regression Shrinkage and Selection via the Lasso [J]. Journal of the Royal Statistical Society：Series B (Methodological)，1996，58(01)：267-288.

[10] 李高荣. 统计学习(R语言版)[M]. 北京：北京师范大学出版社，2025.

[11] HOERL A E，KENNARD R W. Ridge Regression：Biased Estimation for Nonorthogonal Problems[J]. Technometrics，1970，12(01)：55-67.

[12] ZOU H，HASTIE T. Regularization and Variable Selection via the Elastic Net[J]. Journal of the Royal Statistical Society：Series B (Statistical Methodology)，2005，67(02)：301-320.

[13] FAN J，LI R. Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties [J]. Journal of the American Statistical Association，2001，96(456)：1348-1360.

[14] FRANK I E，FRIEDMAN J H. A Statistical View of Some Chemometric Regression Tools [J]. Technometrics，1993，35(02)：109-148.

[15] ZOU H，HASTIE T. Adaptive Bridge Estimation[J]. Journal of the Royal Statistical Society：Series B (Statistical Methodology)，2005，67(05)：717-743.

[16] JIN B，WU Y，SHI X. Consistent Two-stage Multiple Change-point Detection in Linear Models [J]. The Canadian Journal of Statistics，2016，44(02)：161-179.

[17] 安子祯，董翠玲. 线性回归模型中基于GMD算法的两阶段组Lasso多变点估计[J]. 新疆师范大学学报(自然科学版)，2025，4(04)：1-9.

Variable Selection of Poisson Regression Model

ZHANG Wen-jing, DONG Cui-ling*

(*School of Mathematical Sciences, Xinjiang Normal University, Urumqi, Xinjiang, 830017, China*)

Abstract: As a core tool for counting data, the accuracy of variable selection in the Poisson regression model directly affects the interpretability and prediction performance of the model. In this paper, R software is used to study the advantages and disadvantages of six variable selection methods in Poisson regression, including LASSO, Ridge regression, Elastic Net, SCAD, Bridge and Adaptive Bridge, through numerical simulation and empirical analysis of shared bicycle rental data. The results show that the SCAD method performs best in terms of model goodness of fit and prediction accuracy, and it is suitable for large samples and highly correlated data; The Adaptive Bridge method achieves a good balance between model complexity and prediction stability. This study provides a modeling tool with both accuracy and stability for the practical application of counting data.

Keywords: Generalized linear model; Poisson regression model; Variable selection methods; SCAD; Adaptive Bridge

(上接第101页)

A Delay-diffusion Malware Propagation Model with Limited Treatment Resources based on PD Control

GUO Jia[1], LI Ting[1]*, WANG Huai-zhu[2]

(*1.School of Mathematics and Statistics, Ningxia University, Yinchuan, Ningxia, 750021, China*;
2.School of Advanced Interdisciplinary Studies, Ningxia University, Zhongwei, Ningxia, 755000, China)

Abstract: In order to investigate the spatial-temporal dynamic propagation characteristics of malware in the Internet of Things system, this paper establishes a delay-diffusion malware propagation model with limited treatment resources based on the theory of differential equations. By analyzing the relevant characteristic equations, the conditions for Turing instability and Hopf bifurcation are derived. To enhance the system stability, Proportional-Derivative (PD) control is introduced into the model, and the stability of the controlled model is analyzed. Numerical simulations confirm the correctness of the theoretical derivations and demonstrate that the occurrence of Turing instability and Hopf bifurcation can be controlled by choosing appropriate parameters of the PD controller.

Keywords: Internet of Things; Malware; Delay-diffusion; Limited treatment resources; PD control