Pure Mathematics 理论数学, 2026, 16(5), 54-65
Published Online May 2026 in Hans. https://www.hanspub.org/journal/pm
https://doi.org/10.12677/pm.2026.165130

Hans汉斯

基于变量选择的交通事故数据建模研究

张文静，董翠玲*

新疆师范大学数学科学学院，新疆 乌鲁木齐

收稿日期：2026年3月24日；录用日期：2026年5月13日；发布日期：2026年5月25日

摘　要

道路交通事故数据作为经典计数数据，其建模在复杂交通事故分析应用中具有重要意义。本研究基于英国交通部的道路交通事故统计系统(STATS19)中2022~2023年的道路交通事故记录数据，构建涵盖时间效应、自回归效应及交互项的Poisson回归模型，系统对比了逐步回归与五种正则化方法(Ridge回归，LASSO，Elastic Net，SCAD，Adaptive LASSO)的性能。实证表明，SCAD与Adaptive LASSO均显著优于传统逐步回归，其中SCAD在预测精度与模型简洁性之间取得了最优平衡，Adaptive LASSO展现出优异的变量选择一致性与稳定性。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

道路交通事故，Poisson回归，变量选择，SCAD，Adaptive LASSO

Research on Traffic Accident Data Modeling Based on Variable Selection

Wenjing Zhang, Cuiling Dong*

School of Mathematical Sciences, Xinjiang Normal University, Urumqi Xinjiang

Received: March 24, 2026; accepted: May 13, 2026; published: May 25, 2026

Abstract

Road traffic accident data, as a classic form of count data, plays a pivotal role in complex traffic accident analysis. This study utilizes road accident records from the UK Department for Transport's STATS19 system (2022~2023) to construct a Poisson regression model incorporating temporal effects, autoregressive terms, and interaction effects. We systematically compare the performance of stepwise regression against five regularization methods: Ridge, LASSO, Elastic Net, SCAD, and Adaptive

*通讯作者。

文章引用: 张文静, 董翠玲. 基于变量选择的交通事故数据建模研究[J]. 理论数学, 2026, 16(5): 54-65.
DOI: 10.12677/pm.2026.165130

LASSO. Empirical results demonstrate that both SCAD and Adaptive LASSO significantly outperform traditional stepwise regression. Specifically, SCAD achieves an optimal balance between predictive accuracy and model parsimony, while Adaptive LASSO exhibits superior consistency and stability in variable selection. This study provides robust modeling tools that combine precision and stability for practical applications involving count data.

Keywords

Road Traffic Accidents, Poisson Regression, Variable Selection, SCAD, Adaptive LASSO

1. 引言

计数数据广泛存在于公共卫生[1]、交通管理[2]、环境科学[3]等诸多研究领域，其观测值呈现非负整数特征，具备离散性与非负性，且数据分布普遍偏离正态假设。传统线性回归模型的构建基于连续型数据与正态分布假设，若直接将传统线性回归模型应用于计数数据建模，易导致回归系数估计偏差，掩盖变量间的真实关系，造成预测结果显著偏离实际观测值。Poisson 回归模型作为广义线性模型(Generalized Linear Model, GLM)的重要分支[4] [5]，通过引入对数链接函数，成功实现了线性预测空间向非负响应空间的映射，完美适配计数数据的分布特性，已成为计数响应变量建模的核心工具[6]。

道路交通事故的发生频次严格遵循非负整数特征，是典型的计数数据。国内外学者已围绕此类数据开展了大量研究。孟祥海等基于统计假设检验验证了交通事故数据的计数分布特性[7]；Miaou (1994)利用泊松回归量化了交通流量与道路几何特征(如弯道曲率、纵坡坡度)对卡车事故发生次数的影响[8]；陈昭明和徐文远引入随机参数负二项模型量化了道路几何条件(如平曲线曲率、纵坡坡度)对事故频次的影响[9]；王迎和周燕基于广义线性模型验证了 Poisson 回归在高速公路事故预测中的有效性[10]，Lord 和 Mannering 系统分析了交通事故频率数据方法的体系，评估了 Poisson 回归及其扩展形式(负二项、零膨胀、随机参数模型)的适用场景，为后续模型选择提供了重要的理论依据[11]。

然而，道路交通事故的发生受道路几何条件、交通运行状态、环境气象特征及时间周期性等多重因素的综合影响。在实际建模中，潜在候选变量众多且往往存在高度相关性，若盲目将所有变量纳入模型，不仅会增加模型复杂度，还可能因冗余变量引发多重共线性，导致参数估计有偏，削弱模型的预测精度与稳定性，降低模型可解释性，难以精准识别影响事故发生的核心因素。因此，如何在保证模型预测准确度的前提下，通过科学的变量选择方法提升模型简洁性和可解释性，构建兼具高预测精度与强可解释性的 Poisson 回归模型，是当前计数建模领域亟待解决的问题。

变量选择是当前兼顾模型精度与可解释性的关键策略，旨在从众多潜在影响变量中筛选出与响应变量密切相关的核心变量，剔除冗余及无关变量，能在简化模型结构的同时兼顾预测精度与可解释性[12]。目前变量选择方法已形成较为完善的体系，其中传统方法以逐步回归为代表，通过向前引入、向后剔除或双向筛选策略实现变量选择；惩罚类方法则通过引入正则化惩罚函数实现变量选择与系数收缩，如 Ridge 回归、最小绝对收缩与选择算子(Least Absolute Shrinkage and Selection Operator, LASSO)、弹性网(Elastic Net)、平滑削边绝对偏离法(Smoothly Clipped Absolute Deviation, SCAD)及自适应 LASSO (Adaptive LASSO)等，能有效缓解多重共线性并提升估计的稳健性，展现出更优越的统计性质[13]-[18] [19]。

本研究将上述六种变量选择方法应用于 Poisson 回归模型，从均方误差(Mean Squared Error, MSE)、均方根误差(Root Mean Squared Error, RMSE)、平均绝对误差(Mean Absolute Error, MAE)、平均绝对百分比误差(Mean Absolute Percentage Error, MAPE)及决定系数(R^2)等方面进行对比分析，探究各方法的优势与局限，并将其应用于英国交通部的道路交通事故统计系统(STATS19)中 2022~2023 年的日度道路交通事故数据，整理并提取涵盖星期与假日效应、趋势与季节性、自回归效应、交互效应 4 个维度的 16 个自变量，应用变量选择方法选取影响日度道路交通事故发生频次的关键变量，建立更为精准的预测模型。

2. Poisson 回归模型的定义

2.1. 广义线性模型

传统线性回归模型假设响应变量服从正态分布且满足方差齐性，难以直接处理非负整数计数数据。为克服上述局限，Nelder 和 Wedderburn (1972)提出广义线性模型，通过放松正态性假设，将响应变量的分布扩展至指数族分布(Exponential Family)，为计数数据建模提供了统一的统计框架[4]-[6]。

广义线性模型由随机部分、系统部分和链接函数三部分构成[4] [5]，具体如下：

(1) 随机部分独立同分布且为指数族，其形式为：

$$f\left(y_i;\theta_i,\phi\right)=\exp\left\{\frac{y_i\theta_i-b\left(\theta_i\right)}{a\left(\phi\right)}+c\left(y_i,\phi\right)\right\}. \tag{1}$$

其中，y_i 为第 i 个观测值，θ_i 为与第 i 个样本相关的自然参数，ϕ 为公共散度参数(Dispersion Parameter)，$a(\cdot)$、$b(\cdot)$、$c(\cdot)$ 为已知函数。根据指数族性质，响应变量 Y_i 的期望与方差可表示为：$E\left(Y_i\mid x_i\right)=\mu_i=b'\left(\theta_i\right)$，$Var\left(Y_i\mid x_i\right)=a\left(\phi\right)b''\left(\theta_i\right)$。

广义线性模型通过方差函数 $Var\left(\mu_i\right)=b''\left(\theta_i\right)$ 刻画响应变量的离散特征，允许方差随均值变化，能够灵活适配包括正态分布(方差恒定)、Poisson 分布(方差等于均值)、Gamma 分布(方差为均值的二次函数)等多种数据类型。

(2) 系统部分用于描述自变量 x_i 与响应变量均值 $\mu_i=E\left(Y_i\mid x_i\right)$ 之间的线性关联，构建线性预测器 η_i，描述 p 个自变量 $x_i=\left(1,x_{i1},x_{i2},\cdots,x_{ip}\right)^{\mathrm{T}}$ 与响应变量的线性关系：

$$\eta_i=\boldsymbol{x}_i^{\mathrm{T}}\beta=\beta_0+\beta_1x_{i1}+\cdots+\beta_px_{ip}. \tag{2}$$

其中，$\beta=\left(\beta_0,\beta_1,\cdots,\beta_p\right)^{\mathrm{T}}$ 为回归系数向量。

(3) 链接函数用于建立响应变量均值 μ_i 与线性预测器 η_i 之间的单调可导映射 $g\left(\mu_i\right)=\eta_i$。特别的当 $g(\cdot)=\left(b'\right)^{-1}(\cdot)$，则称 $g(\cdot)$ 为正则链接函数(Canonical Link) [20]。

2.2. Poisson 回归模型

Poisson 回归模型是广义线性模型的重要分支，其响应变量服从 Poisson 分布且链接函数为对数链接函数的广义线性模型。

2.2.1. Poisson 回归模型的假设及构建[4] [5]

假设 1，独立性假设：各个观测事件的发生相互独立。

假设 2，等离散性假设：$Y_i\mid x_i\sim\mathrm{Poisson}\left(\mu_i\right)$，且 $E\left(Y_i\mid x_i\right)=Var\left(Y_i\mid x_i\right)=\mu_i$。

假设 3，对数线性性：响应变量均值的自然对数与自变量之间呈线性关系，即 $\ln\left(E\left(Y_i\mid x_i\right)\right)=x_i^{\mathrm{T}}\beta$。

基于上述模型假设，Poisson 回归模型严格遵循广义线性模型的结构：

(1) 随机部分的指数族形式为：

$$P(Y_i=y_i\mid\mu_i)=\frac{\mathrm{e}^{-\mu_i}\mu_i^{y_i}}{y_i!}=\exp\left\{\frac{y_i\theta_i-\mathrm{e}^{\theta_i}}{1}-\ln(y_i!)\right\}\quad y_i=0,1,2,\cdots \tag{3}$$

其中指数族函数中，$\theta_i=\ln(\mu_i)$，$a(\phi)=1$，$b(\theta_i)=\mathrm{e}^{\theta_i}=\mu_i$，$c(y_i,\phi)=-\ln(y_i!)$。由此可得均值与方差关系 $E(Y_i\mid x_i)=Var(Y_i\mid x_i)=\mu_i$。

(2) 系统部分中的线性预测器，为

$$\eta_i=\beta_0+\beta_1x_{i1}+\beta_2x_{i2}+\cdots+\beta_px_{ip}=x_i^{\mathrm{T}}\beta. \tag{4}$$

其中 $x_i=(1,x_{i1},x_{i2},\cdots,x_{ip})^{\mathrm{T}}$ 为协变量向量，$x_{ij}(j=1,2,\cdots,p)$ 为第 i 个样本的第 j 个自变量观测值，$\beta=(\beta_0,\beta_1,\cdots,\beta_p)^{\mathrm{T}}$ 为回归系数向量。

(3) 链接函数为对数链接函数，即

$$g(\mu_i)=\ln(\mu_i)=\eta_i=\beta_0+\beta_1x_{i1}+\beta_2x_{i2}+\cdots+\beta_px_{ip}=x_i^{\mathrm{T}}\beta. \tag{5}$$

这表明响应变量的对数均值与自变量呈线性关系，且通过指数变换保证了预测均值 μ_i 恒为正。

2.2.2. Poisson 回归模型的极大似然估计

传统 Poisson 回归采用极大似然估计法(Maximum Likelihood Estimation, MLE)求解回归系数向量 β。该方法符合 Poisson 分布的指数族特性。且具备一致性、渐近有效性等性质。

基于观测值的独立性假设(1)，样本的似然函数为各观测概率密度函数之积，即：

$$L(\beta)=\prod_{i=1}^{n}\frac{\mu_i^{y_i}\mathrm{e}^{-\mu_i}}{y_i!}. \tag{6}$$

其中 $\mu_i=\exp(x_i^{\mathrm{T}}\beta)$。对式(6)取对数，由于 $\ln(\mu_i!)$ 不包含待估参数 β，得到对数似然函数为：

$$l(\beta)=\sum_{i=1}^{n}\left[y_i\ln(\mu_i)-\mu_i-\ln(\mu_i!)\right]=\sum_{i=1}^{n}\left[y_i(x_i^{\mathrm{T}}\beta)-\exp(x_i^{\mathrm{T}}\beta)\right]. \tag{7}$$

通过最大化式(7)得到 Poisson 回归模型中参数的极大似然估计为：$\hat{\beta}_{\mathrm{MLE}}=\arg\max_\beta l(\beta)$，该估计量具备相合性，渐近有效性等性质。

在实际应用中，传统极大似然估计面临多重共线性与冗余变量等挑战，导致估计方差膨胀、结果不稳定或模型过拟合。为克服这些局限，需引入变量选择方法，在参数估计的同时剔除无关变量，优化模型结构并提升估计的准确性[15]。

3. 变量选择方法及 Poisson 回归的参数估计

变量选择问题是损失函数与惩罚函数之和的极小化问题。在 Poisson 回归框架下，损失函数为负对数似然函数，设 $P_\lambda(\beta)$ 为依赖调节参数 $\lambda\geq0$ 的惩罚函数[15]，则 Poisson 回归模型式(5)中参数 β 的估计量为：

$$\hat{\beta}=\arg\max_\beta\left\{l(\beta)-P_\lambda(\beta)\right\}. \tag{8}$$

不同方法的差异主要体现在惩罚函数 $P_\lambda(\beta)$ 的构造及其统计性质(如稀疏性、无偏性等)。

3.1. 传统变量选择方法——逐步回归

逐步回归(Stepwise)是一种基于信息准则，在离散模型空间中搜索最优变量组合的方法[12]。设 p 为候选变量总数，$S\subseteq\{1,\cdots,p\}$ 为选入变量的子集，$|S|$ 为模型复杂度(变量个数)。则最优组合 $\hat{S}$ 为：

$$\hat{S}=\arg\max_S\left\{2l(\hat{\beta}_S)-\lambda|S|\right\}. \tag{9}$$

其中，$\hat{\beta}_S$ 为子集 S 的无约束最大似然估计。特别地，当 $\lambda=2$ 时，信息准则为 AIC 准则，当 $\lambda=\ln n$ 时信息准则为 BIC 准则[12]。Poisson 回归模型式(5)中参数 β 的估计量为：

$$\hat{\beta}_{Stepwise,j}=\begin{cases}\left(\hat{\beta}_S\right)_j, & \text{若 } j\in\hat{S}\\ 0, & \text{若 } j\notin\hat{S}\end{cases}$$

逐步回归计算简便，但易陷入局部最优，缺乏对系数的连续收缩能力，且在变量共线性较强时稳定性较差，因此现代统计学习更倾向于使用基于连续惩罚框架的正则化方法，实现更稳健的变量选择与参数估计。

3.2. 正则化变量选择方法

Ridge 回归[13]是 Hoerl 和 Kennard (1970)在损失函数的基础上，加入关于回归系数 β 的 L_2 惩罚 $P_\lambda(\beta)=\frac{\lambda}{2}\|\beta\|_2^2$，能有效缓解多重共线性且计算高效的方法，但无法将系数精确压缩为零实现变量选择，导致模型可解释性较弱，常用于预测。Poisson 回归模型式(5)中参数 β 的 Ridge 回归估计量为：

$$\hat{\beta}_{ridge}=\arg\max_{\beta}\left\{l(\beta)-\frac{\lambda}{2}\|\beta\|_2^2\right\}. \tag{10}$$

LASSO [14]是 Tibshirani (1996)在损失函数的基础上，加入关于回归系数 β 的 L_1 惩罚 $P_\lambda(\beta)=\lambda\|\beta\|_1$，能将冗余系数压缩为零，实现变量选择和模型稀疏化的方法。该方法计算高效，但对所有系数施加同等惩罚，导致大系数估计有偏，不具备 Oracle 性质。Poisson 回归模型式(5)中参数 β 的 LASSO 估计量为：

$$\hat{\beta}_{lasso}=\arg\max_{\beta}\left\{l(\beta)-\lambda\|\beta\|_1\right\}. \tag{11}$$

Elastic Net [15]是 Zou 和 Hastie (2005)在损失函数的基础上，融合 L_1 范数与 L_2 范数，构建混合正则化惩罚项 $P_\lambda(\beta)=\lambda\left(\alpha\|\beta\|_1+\frac{1-\alpha}{2}\|\beta\|_2^2\right)$ 的方法，该方法兼具 LASSO 的稀疏性与 Ridge 回归的稳定性，但估计量有偏且不具备 Oracle 性质，模型复杂度较高。Poisson 回归模型式(5)中参数 β 的 ELastic Net 估计量为：

$$\hat{\beta}_{enet}=\arg\max_{\beta}\left\{l(\beta)-\lambda\left(\alpha\|\beta\|_1+\frac{1-\alpha}{2}\|\beta\|_2^2\right)\right\}. \tag{12}$$

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 Ridge 回归。

SCAD [16]是 Fan 和 Li (2001)在损失函数的基础上采用非凸平滑剪切惩罚的方法，其惩罚项的导数定义为 $p'_\lambda(|\beta|)=\lambda\left[I(|\beta|\le\lambda)+\frac{(a\lambda-|\beta|)_+}{(a-1)\lambda}I(|\beta|>\lambda)\right](a>2,(z)_+=\max(0,z))$，该方法通过非凸惩罚实现小系数强罚、大系数弱罚，避免过度收缩重要变量系数，估计量具备 Oracle 性质、稀疏性与无偏性，但计算复杂度高。Poisson 回归模型式(5)中参数 β 的 SCAD 估计量为：

$$\hat{\beta}_{SCAD}=\arg\max_{\beta}\left\{l(\beta)-\sum_{j=1}^{p}p_\lambda(|\beta_j|)\right\}. \tag{13}$$

其中 $a>2$ (常取 $a=3.7$)是一个固定参数，确保惩罚项平滑可微。

Adaptive LASSO [17]是 Zou (2006)在损失函数的基础上，加入关于回归系数 β 的自适应惩罚函数 $P_\lambda(\beta)=\lambda\sum_{j=1}^{p}\hat{w}_j|\beta_j|$，其中自适应权重 $\hat{w}_j=|\tilde{\beta}_j|^{-\gamma}$ ($\gamma>0$，通常取 $\gamma=1$，初始值 $\tilde{\beta}$ 由极大似然估计或 Ridge 回归获得)能实现对大系数弱罚、小系数强罚，克服了 LASSO 的有偏性，具备 Oracle 性质、无偏性与稀

疏性，且作为凸优化问题，其计算效率优于 SCAD。Poisson 回归模型式(5)中参数 β 的 Adaptive LASSO 估计量为：

$$\hat{\beta}_{Ad\text{-}LASSO} = \arg\max_{\beta}\left\{ l(\beta) - \lambda\sum_{j=1}^{p}\hat{\omega}_j|\beta_j| \right\}. \tag{14}$$

4. 实证分析

为比较在 Poisson 回归模型下，逐步回归、Ridge 回归、LASSO、Elastic Net、SCAD 及 Adaptive LASSO 六种方法的变量选择与参数估计的性能，对英国交通部的道路交通事故统计系统(STATS19)的数据进行分析，识别影响英国日度交通事故频次的关键因素。通过均方误差($\mathrm{MSE}=\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$)、均方根误差($\mathrm{RMSE}=\sqrt{\mathrm{MSE}}$)、平均绝对误差($\mathrm{MAE}=\frac{1}{n}\sum_{i=1}^{n}|y_i-\hat{y}_i|$)、平均绝对百分比误差($\mathrm{MAPE}=\frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{y_i-\hat{y}_i}{y_i}\right|$)及决定系数($\mathrm{R}^2=1-\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}$)等多个维度，比较不同变量选择方法在 Poisson 回归框架下的表现，系统评估各方法的优势与局限。

4.1. 数据来源与描述性分析

本研究数据来源于英国交通部的道路交通事故统计系统(STATS19) (https://www.gov.uk/government/statistical-data-sets/road-safety-open-data)，其原始数据记录了 2020~2024 年 5 年间英格兰、苏格兰及威尔士地区的每一起道路交通事故信息，涵盖事故编号、地理坐标、发生日期、星期等 41 项核心信息，数据权威、记录完整且时间连续。提取 2022~2023 年共 730 个观测日的事故数据作，样本记录完整且覆盖 2 个完整年度周期，为日度事故计数建模提供可靠的样本支撑，能有效支撑事故数的周期性与自回归效应分析。

对 2022~2023 年英国日度交通事故数据进行年度、星期、假日三个维度的分组统计，系统梳理事故数据的分布特征，为后续变量选择与模型构建提供数据基础，详见表 1。

从表 1 的年度分布来看，2022 年与 2023 年英国道路交通事故整体水平保持稳定，未出现显著的结构性波动，但具有星期效应和节假日效应，工作日事故量均值明显高于周末，公共假日事故量均值较非假日事故量少。

Table 1. Descriptive statistics of daily traffic accidents
表 1. 日度交通事故数描述统计

分组类型	组别	样本量	均值	标准差	中位数	最小值	最大值
年度	2022	365	290.4	53.5	289	132	461
	2023	365	286.4	52.0	288	115	428
星期	周一至周五	520	302.6	48.7	301	115	461
	周六至周日	210	252.4	44.4	249	132	369
假日	公共假日	19	201.0	46.8	201	115	283
	非假日	711	290.4	50.9	290	132	461
全样本	全样本	730	288.4	52.7	289	115	461

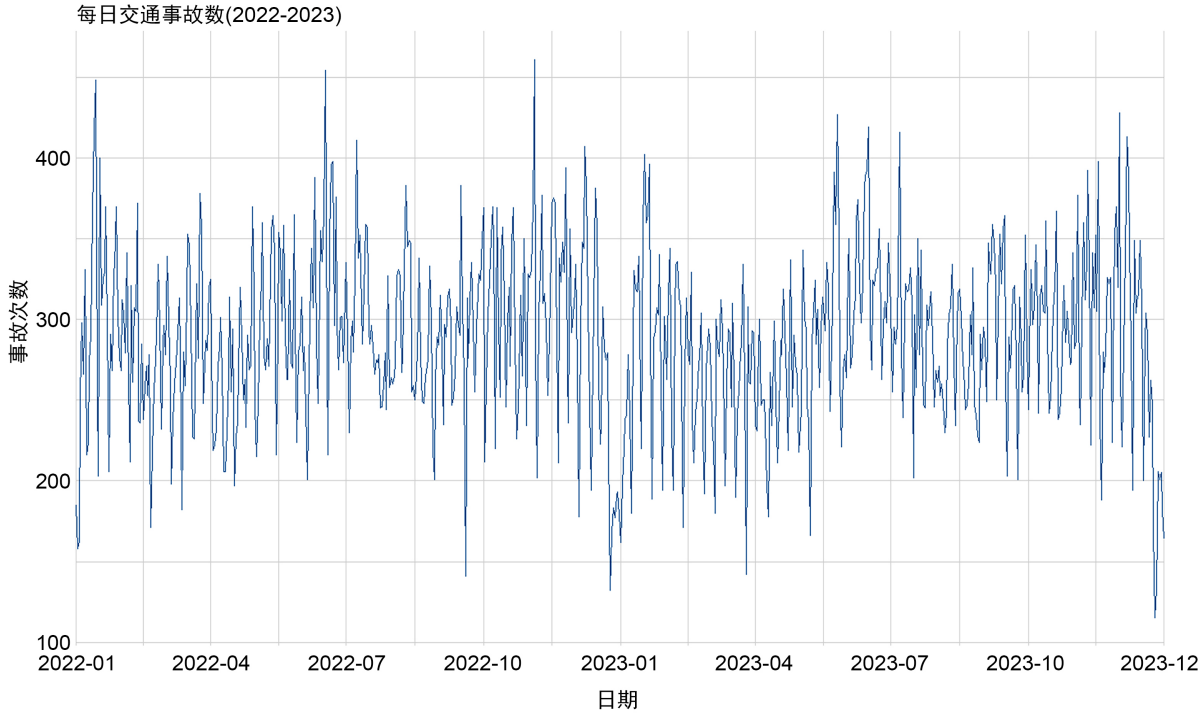

Figure 1. Time series of daily traffic accident counts (2022~2023)
图 1. 日度交通事故数时间序列(2022~2023)

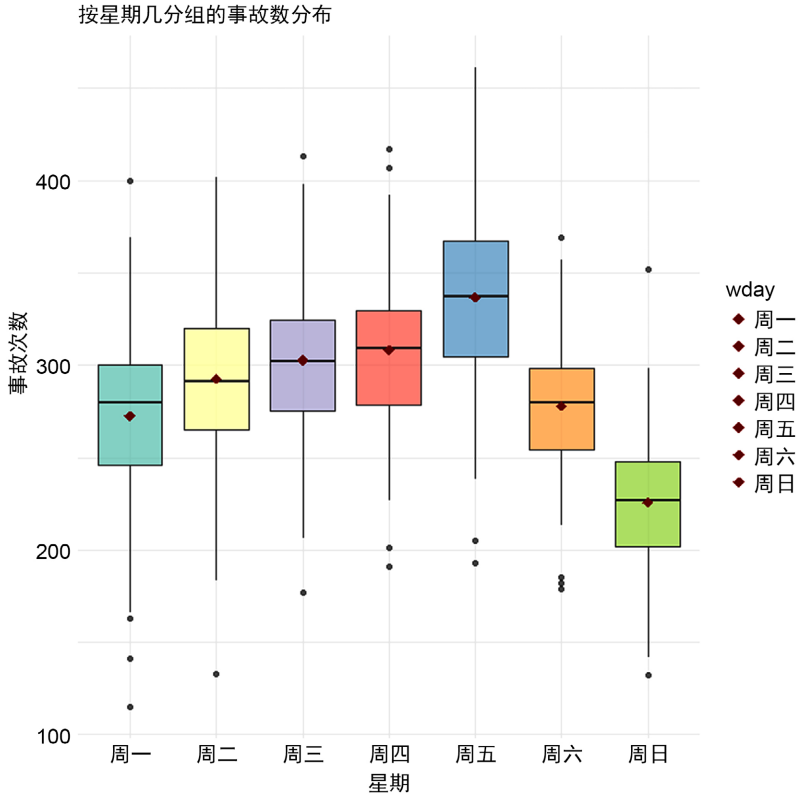

Figure 2. Box plot of daily accident count distribution grouped by weekday
图 2. 按星期分组的日度事故数分布箱线图

图 1 和图 2 分别为 2022~2023 年日度交通事故数时间序列图和按星期分组的每日交通事故数箱线图。图 1 呈现明显的周期性波动，峰值与谷值交替出现，周期长度约为 7 天，对应表 1 的星期效应。从图 2 中可以看出，工作日事故数量明显高于周末，与表 1 分析一致，箱线图分布近似对称、无明显偏态，可为后续 Poisson 回归建模提供依据。

4.2. 自变量的构建

为分析影响事故频次发生的关键因素，结合描述性分析，同时考虑交通事故发生的时间惯性、周期性及交互效应，从星期与假日效应、趋势与季节性、自回归效应、交互效应四个维度构造 16 个自变量，详见表 2。

Table 2. Independent variable treatment table
表 2. 自变量处理表

特征维度	编号	变量名称	定义与构造方法	设计目的
星期与假日效应	x_1	is_monday	若日期为周一则取 1，否则 0	捕捉周一早高峰带来的事故风险上升
	x_2	is_friday	若日期为周五则取 1，否则 0	反映周末前出行模式变化对事故发生的影响
	x_3	is_weekend	若日期为周六或周日则取 1，否则 0	刻画周末交通量减少对事故发生的影响
	x_4	holiday	若日期为英国公共假日则取 1，否则 0	反映假日出行行为对事故发生的影响
趋势与季节性	x_5	days_since_start	从数据起始日到当前的天数	反映事故数随时间的变化
	x_6	month_sin	$\sin(2\pi\cdot\text{month}/12)$	以周期项建模分析季节性波动
	x_7	month_cos	$\cos(2\pi\cdot\text{month}/12)$	同上，与 month_sin 共同构成完整周期
自回归效应	x_8	lag_1	前 1 天的实际事故数	捕捉事故数的短期惯性(1 日记忆)
	x_9	lag_2	前 2 天的实际事故数	捕捉事故数的短期惯性(2 日记忆)
	x_{10}	lag_7	前 7 天的实际事故数	捕获以周为单位的周期记忆
	x_{11}	lag_14	前 14 天的实际事故数	捕获以两周为单位的周期记忆
	x_{12}	rolling_mean_7	过去 7 天的日均事故数	平滑近期波动，反映短期趋势水平
	x_{13}	rolling_mean_14	过去 14 天的日均事故数	平滑更长时间窗口，增强趋势稳健性
	x_{14}	diff_1_7	$\text{lag}_1-\text{lag}_7$	反映当日与一周前同期的变化
	x_{15}	pct_change_7	(lag_1-lag_7)/lag_7(当 lag_7>0)	反映事故数较上周同期的相对波动
交互效应	x_{16}	weekend_lag1	$\text{is}_{\text{weekend}}\times\text{lag}_1$	捕捉周末前一日数据对周末事故数的影响

图 3 给出各变量间的 Pearson 相关系数矩阵热力图，从图中可以看出，自回归变量间呈现高度相关性，如 x_{12} (rolling_mean_7)与 x_{13} (rolling_mean_14)高度相关(0.88)，x_{14} (diff_1_7)与 x_{15} (pct_change_7)近乎完全共线(0.97)，这两组变量可能导致估计方差膨胀；x_{10} (lag_7)与 x_{11} (lag_14)中度相关(0.52)；其余变量间相关性普遍较弱，因此在后续建模中引入正则化变量选择方法缓解多重共线性、提升模型稳健性尤为重要。

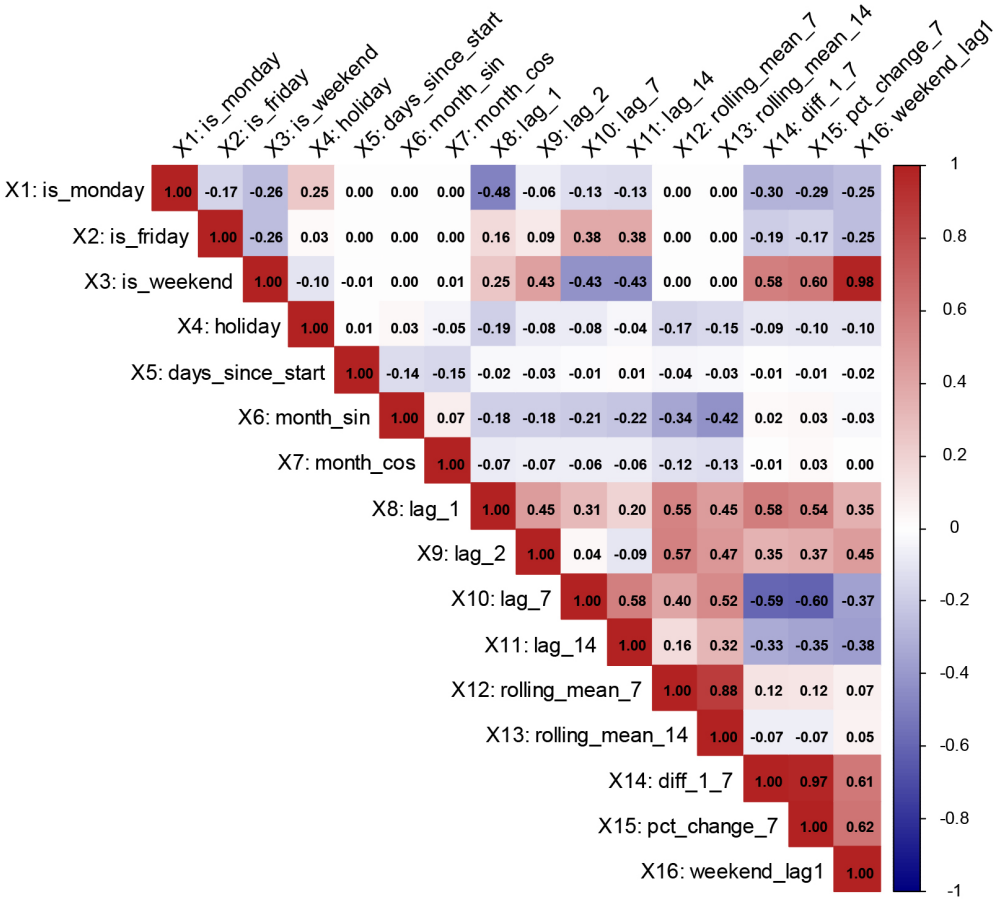

Figure 3. Heatmap of pearson correlation coefficient matrix
图 3. Pearson 相关系数矩阵热力图

4.3. 调节参数的选择

本研究所有模型的正则化参数 λ 均通过十折交叉验证确定，以兼顾计算效率与模型稳健性。Ridge 回归、LASSO、Elastic Net 及 Adaptive LASSO 基于 R 包 glmnet 的坐标下降算法求解；SCAD 基于 ncvreg 包的局部二次逼近(LQA)求解，并参考相关研究文献[16]设定 $a=3.7$。

5. 建模与结果分析

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“123”，在数据集中按照时间顺序将 730 天样本按 80%:20%划分为训练集(584 天)和测试集(146 天)。对连续型自变量进行标准化处理(Z-score)，确保惩罚项对各系数收缩力度的公平性。应用逐步回归、Ridge 回归、LASSO、Elastic Net、SCAD、Adaptive LASSO 进行变量选择与参数估计，得到的结果如表 3 所示。

Table 3. Regression coefficient estimation using six methods
表 3. 六种方法的回归系数估计

编号	变量	$\hat{\beta}_{Stepwis}$	$\hat{\beta}_{ridge}$	$\hat{\beta}_{lasso}$	$\hat{\beta}_{enet}$	$\hat{\beta}_{SCAD}$	$\hat{\beta}_{Ad\text{-}LASSO}$
β_0	(Intercept)	5.6517	5.6521	5.6518	5.6518	5.6519	5.6518
x_1	is_monday	0.0000	0.0002	−0.0021	−0.0022	0.0000	0.0000
x_2	is_friday	0.0320	0.0301	0.0311	0.0312	0.0282	0.0308
x_3	is_weekend	−0.1191	−0.0776	−0.1090	−0.1104	−0.0483	−0.1043
x_4	holiday	−0.0398	−0.0385	−0.0390	−0.0390	−0.0376	−0.0391
x_5	days_since_start	0.0000	−0.0012	−0.0007	−0.0007	0.0000	0.0000
x_6	month_sin	−0.0052	−0.0050	−0.0051	−0.0051	−0.0001	−0.0035
x_7	month_cos	−0.0058	−0.0063	−0.0059	−0.0059	−0.0020	−0.0049
x_8	lag_1	0.0778	0.0493	0.0600	0.0596	0.0705	0.0623
x_9	lag_2	−0.0197	−0.0191	−0.0191	−0.0191	−0.0206	−0.0190
x_{10}	lag_7	−0.0184	0.0165	0.0000	0.0000	0.0000	0.0000
x_{11}	lag_14	0.0181	0.0178	0.0179	0.0178	0.0209	0.0181
x_{12}	rolling_mean_7	0.1224	0.1128	0.1212	0.1213	0.1253	0.1211
x_{13}	rolling_mean_14	−0.0688	−0.0603	−0.0672	−0.0673	−0.0707	−0.0675
x_{14}	diff_1_7	0.0000	0.0269	0.0152	0.0163	0.0000	0.0127
x_{15}	pct_change_7	−0.0611	−0.0460	−0.0544	−0.0555	−0.0432	−0.0528
x_{16}	weekend_lag1	0.0712	0.0274	0.0599	0.0614	0.0000	0.0555

从表 3 中发现，不同变量选择方法在稀疏性与系数收缩上表现各异：逐步回归保留了 13 个变量；正则化方法中 Ridge 回归保留了 16 个变量，不具备变量选择能力；LASSO 与 Elastic Net 结果接近，均通过惩罚实现系数收缩与变量选择；SCAD 采用非凸惩罚，筛选最为严格，仅保留 11 个变量；Adaptive LASSO 保留 13 个变量，体现了其 Oracle 性质。所有方法均一致保留了 is_friday、holiday、lag_1 以及 rolling_mean_7，强有力地证实了周五高峰、假日效应及事故短期记忆是事故计数的核心影响因素。

表 4 列出了六种变量选择方法的 MSE、RMSE、MAE、MAPE 及 R^2五项指标对测试集、预测集的量化评估结果，具体如下。

表 4 结果显示所有惩罚类方法的 R^2 均高于传统逐步回归(0.7033)，表明引入正则化机制能够有效提升 Poisson 回归的预测精度。其中 SCAD 方法表现较好，以 11 个精简变量取得最低的误差和最高的拟合优度($R^2 = 0.7075$)，凸显其非凸惩罚机制的优势；LASSO 与 Elastic Net 性能相近($R^2 \approx 0.705$)，略优于 Ridge 回归(0.7046)，体现了稀疏性约束的积极作用；Adaptive LASSO 以 13 个变量获得次优性能($R^2 = 0.7051$)，在解释性与预测性间取得良好平衡；Ridge 虽能缓解多重共线性但缺乏变量选择能力。

基于上述分析，可采用 SCAD 方法建立日度道路交通事故频次的 Poisson 回归模型，其表达式为：

$$\ln(\hat{\mu}) = 5.6519 + 0.0282x_2 - 0.0483x_3 - 0.0376x_4 - 0.0001x_6 - 0.0020x_7 + 0.0705x_8 - 0.0206x_9 + 0.0209x_{11} + 0.1253x_{12} - 0.0707x_{13} - 0.0432x_{15}$$

Table 4. Summary of predictive performance of six variable selection methods
表 4. 六种变量选择方法的预测性能汇总表

方法	选择的变量数量	MSE	RMSE	MAE	MAPE (%)	R^2
Stepwise	13	868.1847	29.4650	23.4698	8.6241	0.7033
Ridge	16	864.3702	29.4002	23.5613	8.6268	0.7046
LASSO	15	863.8598	29.3915	23.4171	8.5931	0.7048
Elastic Net	15	864.2431	29.3980	23.4142	8.5926	0.7046
SCAD	**11**	**855.9890**	**29.2573**	**23.3729**	**8.5746**	**0.7075**
Adaptive LASSO	13	862.9968	29.3768	23.4346	8.6036	0.7051

6. 结论

本研究基于英国交通部的道路交通事故统计系统(STATS19) 2022~2023 年日度事故数据，应用逐步回归、Ridge 回归、LASSO、Elastic Net、SCAD 及 Adaptive LASSO 六种变量选择方法，构建 Poisson 回归模型，结果显示星期效应(is_friday)、节假日效应(holiday)、自回归效应(lag_1, rolling_mean_7)为影响英国道路事故发生的重要因素，证实了正则化方法的有效性，为交通事故频次的建模及短期预测提供了较为精准的模型，也为国内复杂交通数据分析提供了参考。但仍存在一定局限性，未纳入实时气象数据(气温，能见度等)，后续研究可纳入气象数据以提升预测效果。

基金项目

新疆维吾尔自治区自然科学基金项目(2023D01A37)。

参考文献

[1] Okorie, I.E., Afuecheta, E., Nadarajah, S., Bright, A. and Akpanta, A.C. (2024) A Poisson Regression Approach for Assessing Morbidity Risk and Determinants among under Five Children in Nigeria. *Scientific Reports*, **14**, Article No. 21580. https://doi.org/10.1038/s41598-024-72373-4

[2] Ma, J. and Kockelman, K.M. (2006) Bayesian Multivariate Poisson Regression for Models of Injury Count, by Severity. *Transportation Research Record*: *Journal of the Transportation Research Board*, **1950**, 24-34. https://doi.org/10.1177/0361198106195000104

[3] Joseph, J.F., Furl, C., Sharif, H.O., Sunil, T. and Macias, C.G. (2021) Towards Improving Transparency of Count Data Regression Models for Health Impacts of Air Pollution. *Applied Sciences*, **11**, Article 3375. https://doi.org/10.3390/app11083375

[4] Nelder, J.A. and Wedderburn, R.W.M. (1972) Generalized Linear Models. *Journal of the Royal Statistical Society. Series A* (*General*), **135**, 370-384. https://doi.org/10.2307/2344614

[5] McCullagh, P. (2019) Generalized Linear Models. Routledge. https://doi.org/10.1201/9780203753736

[6] Cameron, A.C. and Trivedi, P.K. (2013) Regression Analysis of Count Data. 2nd Edition, Cambridge University Press. https://doi.org/10.1017/cbo9781139013567

[7] 孟祥海, 覃薇, 霍晓艳. 基于统计与假设检验的高速公路交通事故数据分布特性[J]. 交通运输工程学报, 2018, 18(1): 139-149.

[8] Miaou, S. (1994) The Relationship between Truck Accidents and Geometric Design of Road Sections: Poisson versus Negative Binomial Regressions. *Accident Analysis & Prevention*, **26**, 471-482. https://doi.org/10.1016/0001-4575(94)90038-8

[9] 陈昭明, 徐文远. 基于负二项分布的高速公路交通事故影响因素分析[J]. 交通信息与安全, 2022, 40(1): 28-35.

[10] 王迎, 周燕. 基于广义线性模型的高速公路交通事故预测[J]. 公路工程, 2015, 40(5): 115-119.

[11] Lord, D. and Mannering, F. (2010) The Statistical Analysis of Crash-Frequency Data: A Review and Assessment of

Methodological Alternatives. *Transportation Research Part A*: *Policy and Practice*, **44**, 291-305. https://doi.org/10.1016/j.tra.2010.02.001

[12] Akaike, H. (1974) A New Look at the Statistical Model Identification. *IEEE Transactions on Automatic Control*, **19**, 716-723. https://doi.org/10.1109/tac.1974.1100705

[13] Hoerl, A.E. and Kennard, R.W. (1970) Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, **12**, 55-67. https://doi.org/10.1080/00401706.1970.10488634

[14] Tibshirani, R. (1996) Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society Series B*: *Statistical Methodology*, **58**, 267-288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

[15] Zou, H. and Hastie, T. (2005) Regularization and Variable Selection via the Elastic Net. *Journal of the Royal Statistical Society Series B*: *Statistical Methodology*, **67**, 301-320. https://doi.org/10.1111/j.1467-9868.2005.00503.x

[16] Fan, J. and Li, R. (2001) Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties. *Journal of the American Statistical Association*, **96**, 1348-1360. https://doi.org/10.1198/016214501753382273

[17] Zou, H. (2006) The Adaptive Lasso and Its Oracle Properties. *Journal of the American Statistical Association*, **101**, 1418-1429. https://doi.org/10.1198/016214506000000735

[18] Hastie, T., Tibshirani, R. and Friedman, J.H. (2009) The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer. https://doi.org/10.1007/978-0-387-84858-7

[19] 张文静, 董翠玲. Poisson 回归模型的变量选择[J]. 新疆师范大学学报(自然科学版), 2026, 45(3): 102-112.

[20] Agresti, A. (2013) Categorical Data Analysis. John Wiley & Sons.