引文格式: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(6):44－47.

带有测量误差 Poisson 回归模型的变量选择*

刘　芳

(赣南师范大学 教务处,江西 赣州　341000)

摘　要:Poisson 回归模型是一种特殊的广义线性模型,被广泛用于计数数据的建模.然而在很多应用中,模型中的协变量很难被精确测量,在进行变量选择和参数估计时,这些测量误差通常难以处理,特别是当维数较高时.针对这一问题,本文提出了一种惩罚的偏差校正方法以确定带有测量误差的 Poisson 回归模型(MPR)的真实结构.该方法惩罚了专门设计用于处理测量误差的目标函数,并在一个有界的区域内利用复合梯度下降算法给出回归参数的估计.通过随机模拟验证了本文所提出方法的有限样本性质.

关键词:Poisson 回归模型;测量误差;变量选择;惩罚函数

中图分类号:O212.1　　**文献标志码**:A　　**文章编号**:2096－7659(2024)06－0044－04

0　引言

计数数据在实践中经常会遇到.例如,神经科学研究中的认知得分、传染病学研究中的死亡人数,以及电子商务平台上特定产品的点击次数,这些都是计数数据.在文献中,Poisson 回归是描述计数结果的最流行的模型,因为它自然地模拟了计数变量的偏态分布.另一方面,在大数据时代,与计数数据一起,常常会收集到大量的协变量.然而,由于不完善的数据采集和处理程序,这些协变量通常很难被精确观测,存在测量误差.忽略这些误差可能会产生有偏差的结果,最终可能导致对模型参数的误导性统计推断.本文的目的是对协变量带有测量误差的 Poisson 回归模型进行变量选择.

设协变量带有加性测量误差的 Poisson 回归模型如下:

$$\begin{cases} Y_i \mid X_i \sim Poisson\left(\exp(\beta_0^{\mathrm{T}} X_i)\right) \\ W_i = X_i + U_i \end{cases}, i = 1,2,\cdots,n, \tag{1}$$

其中 Y_i 是响应变量;β_0 是 p 维待估计的未知参数向量.此处 $X_i = (X_{i1},\cdots,X_{ip})^{\mathrm{T}}$ 是一个观测不到的 p 维协变量,取而代之的是 W_i 被直接观测.加性测量误差 U_i 与 (Y_i, X_i) 独立且服从均值为零的 p 维正态分布.与文献 Datta 和 Zou[1],Jiang 和 Ma[2]的假设相同,此处设 U_i 的协方差矩阵 Σ 已知.

本文研究高维协变量具有加性测量误差的 Poisson 回归模型(1)的变量选择问题.在协变量可以精确观测的情况下,已经有了一些方法来对高维 Poisson 回归模型进行变量选择,参见文献 Raginsky 等[3]、Li 和 Cevher[4]、Ivanoff 等[5]、Jia 等[6].然而,当协变量存在测量误差时,目前尚不清楚现有的变量选择方法是否适用,对高维 Poisson 测量误差回归模型的变量选择方法有待研究.事实上,构建模型(1)的变量选择方法主要有如下两个障碍:(I)Poisson 回归函数是非线性的,因此不能像线性测量误差回归模型(Loh 和 Wainwright[7],Datta 和 Zou[1],Li 等[8])一样直接构造一个合适的目标函数来近似无测量误差时的目标函数.(II)在 Poisson 回归模型中,响应变量的条件期望为 $e^{X_i^{\mathrm{T}}\beta}$,其比线性期望 $X_i^{\mathrm{T}}\beta$ 增长得快得多,即使 $X_i^{T}\beta$ 适度大小时,条件期望也可能爆炸,这使得构造一个非奇异的误差修正的 Hessian 矩阵非常困难.

针对上述两个障碍,本文提出了一种新的优化方法.该方法为研究高维非线性测量误差回归模型的统计和数值性质提供了一个新的思路.针对障碍(I),本文构造了一个目标函数,使其条件期望为 Jia 等[6]提出的

* **收稿日期**:2024－09－03　　DOI:10.13698/j.cnki.cn36－1346/c.2024.06.009

基金项目:江西省教育厅科技项目(GJJ211403)

作者简介:刘芳(1981－),女,福建南平人,赣南师范大学教务处讲师,研究方向:教育统计.

目标函数.虽然 Jia 等[6]的目标函数是凸函数,但是本文的目标函数由于测量误差的影响被证明是非凸的.因此,本文构建一个在 L_1 范数约束条件下的正则化稀疏模型,并利用复合梯度下降算法来求解.针对障碍(II),本文将未知参数的搜索空间限制在一个有界的 L_2 范数球内.这样,目标函数及其 Hessian 矩阵的特征值就不会在这个集合中爆炸.

本文第 1 节介绍了本文提出的方法,对高维协变量具有加性测量误差的 Poisson 回归模型进行变量选择和参数估计;第 2 节给出了基于复合梯度下降算法的估计的计算过程;第 3 节通过数值模拟验证所提方法的有效性;最后,第 4 节总结本文的工作内容.

1　主要方法

本文的目标是识别模型式(1)中的非零系数和零系数.如果协变量 X_i 可以被精确观测,Jia 等[6]通过惩罚加权得分函数方法(L_1 penalizedweighted score function method, LPWS)最小化目标函数 $\frac{1}{n}\sum_{i=1}^{n}(Y_i e^{-\frac{1}{2}X_i^{\mathrm{T}}\beta}+e^{\frac{1}{2}X_i^{\mathrm{T}}\beta})$ 来估计未知参数 β_0.当测量误差 U_i 服从正态分布 $N(0,\Sigma)$ 时,易证:

$$E\left\{\exp\left(-\frac{1}{2}W_i^{\mathrm{T}}\beta-\frac{\beta^{\mathrm{T}}\Sigma\beta}{8}\right)\,\middle|\,X_i\right\}=\exp\left(-\frac{1}{2}X_i^{\mathrm{T}}\beta\right),\ E\left\{\exp\left(\frac{1}{2}W_i^{\mathrm{T}}\beta-\frac{\beta^{\mathrm{T}}\Sigma\beta}{8}\right)\,\middle|\,X_i\right\}=\exp\left(\frac{1}{2}X_i^{\mathrm{T}}\beta\right).$$

当 X_i 给定时,Y_i 和 W_i 相互独立,则由上式可得

$$E\left\{Y_i\exp\left(-\frac{1}{2}W_i^{\mathrm{T}}\beta-\frac{\beta^{\mathrm{T}}\Sigma\beta}{8}\right)+\exp\left(\frac{1}{2}W_i^{\mathrm{T}}\beta-\frac{\beta^{\mathrm{T}}\Sigma\beta}{8}\right)\,\middle|\,X_i,Y_i\right\}=Y_i\exp\left(-\frac{1}{2}X_i^{\mathrm{T}}\beta\right)+\exp\left(\frac{1}{2}X_i^{\mathrm{T}}\beta\right).$$

定义偏差校正的目标函数 $\ell(\beta)$ 为

$$\ell(\beta)=\frac{1}{n}\sum_{i=1}^{n}\left[Y_i\exp\left(-\frac{1}{2}W_i^{\mathrm{T}}\beta-\frac{\beta^{\mathrm{T}}\Sigma\beta}{8}\right)+\exp\left(\frac{1}{2}W_i^{\mathrm{T}}\beta-\frac{\beta^{\mathrm{T}}\Sigma\beta}{8}\right)\right].$$

显然,当 X_i 被精确观测时,$\ell(\beta)$ 和 Jia 等[6]的目标函数有着一样的数学期望.因此,可以通过最小化 $\ell(\beta)$ 来估计 β_0.当 $n>p$ 时,利用梯度下降法最小化 $\ell(\beta)$ 可以直接得到 β_0 的估计量.但是,当 $n\leqslant p$ 时,不添加额外的正则化约束条件,最小化 $\ell(\beta)$ 将是一个不适定的数学问题.

考虑到模型式(1)的高维特性,给定正参数 R_1 和 R_2,本文通过求解如下的约束正则化稀疏模型来估计 β_0:

$$\hat{\beta}=\min_{\|\beta\|_1\leqslant R_1,\|\beta\|_2\leqslant R_2}\{\ell(\beta)+\rho_\lambda(\beta)\}\tag{2}$$

其中 $\rho_\lambda(\beta)$ 是一个惩罚函数,λ 为调整参数.

注 1　设集合 $\{\beta:\|\beta\|_1\leqslant R_1,\|\beta\|_2\leqslant R_2\}$ 为可行集(Fletcher 和 Watson[9]),其中 R_1 和 R_2 分别为任意大于 $\|\beta_0\|_1$ 和 $\|\beta_0\|_2$ 的常数.条件 $\|\beta\|_1\leqslant R_1$ 保证式(2)中的目标函数满足 Loh 和 Wainwright[7]文中讨论的限制特征值条件;条件 $\|\beta\|_2\leqslant R_2$ 避免了函数项 $\exp\left(-\frac{W_i^{\mathrm{T}}\beta}{2}-\frac{\beta^{\mathrm{T}}\Sigma\beta}{8}\right)$ 和 $\exp\left(\frac{W_i^{\mathrm{T}}\beta}{2}-\frac{\beta^{\mathrm{T}}\Sigma\beta}{8}\right)$ 的爆炸.因此,根据 Loh 和 Wainwright[7]的证明,式(2)中的目标函数在可行集中是凸的.

注 2　惩罚函数的一般形式为 $\rho_\lambda(\beta)=\sum_{j=1}^{p}\rho_\lambda(|\beta_j|)$,其中 β_j 为向量 β 的第 j 个分量.本文考虑经典的惩罚函数:Lasso[10]罚 $\rho_\lambda(|t|)=\lambda|t|$.

2　计算

为了求解式(2)中的 $\hat{\beta}$,首先选取充分大的常数 R_1 和 R_2 分别满足 $\|\beta_0\|_1\leqslant R_1$ 和 $\|\beta_0\|_2\leqslant R_2$,然后利用复合梯度下降法求解 $\hat{\beta}$.具体来说,通过递归更新 β:

$$\beta^{t+1}=\operatorname*{argmin}_{\|\beta\|_1\leqslant R_1,\|\beta\|_2\leqslant R_2}\left\{\frac{\partial\ell(\beta^t)}{\partial\beta^{\mathrm{T}}}(\beta-\beta^t)+\frac{\eta}{2}\|\beta-\beta^t\|_2^2+\rho_\lambda(\beta)\right\}\tag{3}$$

其中 $\eta>0$ 为步长参数.

为了求解式(3),如果忽略对 β 的 L_1 和 L_2 范数约束,式(3)是一个典型的二次函数加上一个正则化惩罚函数.因此,考虑到这些约束,首先可以使用现有的算法来求解下式:

$$\tilde{\beta}^{t+1}=\operatorname*{argmin}_{\beta}\left\{\frac{\partial\ell(\beta^t)}{\partial\beta^{\mathrm{T}}}(\beta-\beta^t)+\frac{\eta}{2}\|\beta-\beta^t\|_2^2+\rho_\lambda(\beta)\right\}\tag{4}$$

然后，采用 Duchi 等[11]提出的单纯形投影法，将 $\tilde{\beta}^{t+1}$ 投影到半径为 R_1 的 L_1 范数球上得到 $\breve{\beta}^{t+1}$.最后，如果 $\| \breve{\beta}^{t+1} \|_2 > R_2$，则令 $\beta^{t+1} = \breve{\beta}^{t+1} R_2 / \| \breve{\beta}^{t+1} \|_2$，否则令 $\beta^{t+1} = \breve{\beta}^{t+1}$.

上述算法总结如下：

算法 1

输入：观测数据 (W_i, Y_i)，$i = 1, \cdots, n$，参数 $\Sigma, \eta, \lambda, N, tol$.

代入观测数据 (W_i, Y_i)，求解 Jia 等[6]的估计值 $\hat{\beta}_0$

令 $R_2 = 2 \| \hat{\beta}_0 \|_2$ 和 $R_1 = \sqrt{n/\log(p)} R_2$

循环：t 从 0 到 N

1.求解式(4)得到 $\tilde{\beta}^{t+1}$

2.将 $\tilde{\beta}^{t+1}$ 投影到半径为 R_1 的 L_1 球上得到 $\breve{\beta}^{t+1}$

3. $\beta^{t+1} = \breve{\beta}^{t+1} \min(R_2, \| \breve{\beta}^{t+1} \|_2) / \| \breve{\beta}^{t+1} \|_2$

如果 $\| \beta^{t+1} - \beta^t \| \leqslant tol$，停止.

输出

注 3 在算法 1 中，R_2 限制了优化过程的搜索范围.本文将 W 作为 X，利用 Jia 等[6]提出的方法求得 β_0 的估计 $\hat{\beta}_0$，然后取 $R_2 = 2 \| \hat{\beta}_0 \|_2$.再根据 Loh 和 Wainwright[7]关于复合梯度下降算法的阐述，取 $R_1 = \sqrt{n/\log(p)} R_2$.

注 4 本文采用 He 等[12]的方法选取步长 η .类似文献 Friedman 等[13]，本文利用五折交叉验证来选择调整参数 λ .令 $I = \{1, 2, \cdots, n\}$ 是完整数据集的下标，I_k 和 I_k^C 分别为测试集和训练集的下标集，则交叉验证准则表示为 $CV(\lambda) = \frac{1}{5} \sum_{k=1}^{5} \sum_{i \in I_k} \{Y_i - \exp(\hat{\beta}_{(k)}^{\mathrm{T}} W_i - \hat{\beta}_{(k)}^{\mathrm{T}} \Sigma \hat{\beta}_{(k)} / 2)\}^2$.

3 数值模拟

本节通过数值模拟验证本文方法的有限样本性质.为此，首先生成一个设计矩阵 $X \in \Re^{n \times p}$，其中矩阵的每个元素 $X_{ij} \sim N(0,1)$，接着将 X 规范化，使得 $\sum_{i=1}^{n} X_{ij} = 0$ 且 $\frac{1}{n} \sum_{i=1}^{n} X_{ij}^2 = 1$.其次生成测量误差 $U_i \sim N(0, \Sigma)$，其中 Σ 的第 (k,l) 个元素为 $0.04 \times 0.5^{|k-l|}$.将 β_0 中的非零元素个数设置为 5，且每个非零元素从 $N(0,1)$ 中随机产生.最后生成响应变量 $Y_i \sim Poisson\left(\exp\left\{\sum_{j=1}^{5} X_{ij} \beta_{0j}\right\}\right)$.这里在 $n = 100$ 和 200，$p = 300$，500 和 1000 时分别执行模拟过程，且每种情况都进行 100 次重复模拟.在研究中，考虑 Lasso 惩罚函数，并将本文的偏差校正方法(MPR)与 Jia 等[6]的方法(LPWS)进行对比，其中 LPWS 方法的协变量用 W 替代.

表 1 中，“TP”给出了 β_0 中非零元素被正确选出的平均个数，“FP”表示 β_0 中零元素没有被正确选出的平均个数.表 1 括号中的值是对应的标准差.从表 1 可以看出，随着维数 p 的增加，MPR 和 LPWS 选出的参数个数差别不大，在变量选择上表现优异.但是，相比较 LPWS，MPR 的 TP 值更大，FP 值更小，说明 MPR 能有效的校正测量误差的影响.另外，随着样本量 n 的增加，MPR 的 FP 值显著变小，说明在大样本条件下 MPR 能选出更加准确的模型.

表 1 变量选择的结果

$n=100$				$n=200$			
p	方法	TP	FP	p	方法	TP	FP
300	LPWS	4.67(0.47)	14.87(3.53)	300	LPWS	4.02(0.14)	5.49(3.15)
	MPR	4.79(0.41)	9.37(3.34)		MPR	4.51(0.10)	1.71(1.77)
500	LPWS	4.88(0.84)	12.48(1.68)	500	LPWS	4.10(0.30)	6.23(3.54)
	MPR	4.94(0.81)	10.24(1.60)		MPR	4.55(0.22)	2.20(1.91)
1000	LPWS	4.87(0.67)	16.75(5.08)	100	LPWS	4.59(0.49)	10.51(4.30)
	MPR	4.97(0.17)	10.00(4.48)		MPR	4.67(0.50)	4.04(2.52)

表 2 给出了预测误差 $\| X(\hat{\beta} - \beta_0) \|_2$ 的平均值(PE)和 L_1 损失误差 $\| \hat{\beta} - \beta_0 \|_1$ 的平均值(LE).表 2 括号中的值是对应的中位数.表 2 的结果显示，随着维数 p 的显著增加，MPR 和 LPWS 的 PE 值和 LE 值都

只是略有增加,说明两个方法在估计的准确性上表现优异.但是,相比较 LPWS,MPR 的 PE 值和 LP 值更小,说明 MPR 能校正测量误差带来的影响,有效提高估计的准确性.另外,随着样本量 n 的增加,估计的准确性都会随之提高.

表 2　估计的误差

$n=100$				$n=200$			
p	方法	PE	LE	p	方法	PE	LE
300	LPWS	10.77(10.63)	2.88(2.86)	300	LPWS	8.05(7.99)	1.61(1.58)
	MPR	8.93(8.81)	2.25(2.17)		MPR	7.05(7.10)	1.48(1.43)
500	LPWS	8.22(8.23)	2.32(2.33)	500	LPWS	8.88(8.82)	1.75(1.72)
	MPR	8.10(8.12)	2.30(2.29)		MPR	7.84(7.82)	1.59(1.56)
1000	LPWS	14.51(14.52)	4.38(4.21)	1000	LPWS	9.58(9.73)	1.90(1.91)
	MPR	12.19(11.88)	3.29(3.18)		MPR	8.34(8.28)	1.64(1.63)

4　总结

本文研究了高维协变量具有加性测量误差的 Poisson 回归模型的变量选择问题.在参数的稀疏性假设下,通过加性测量误差结构,构造了一个新的目标函数,并设计了一个计算算法.本文所提出的模型纠正了从无误差处理中得到的错误结果.此研究结果可以进一步扩展到其它回归模型和更复杂的测量误差结构.

参考文献:

[1] DATTA A, ZOU H. CoCoLasso for high-dimensional error-in-variables regression[J].Annals of Statistics, 2017,45(6):2400－2426.

[2] JIANG F, MA Y Y. Poisson regression with error corrupted high dimensional features[J].Statistica Sinica, 2022,32(4):2023－2046.

[3] RAGINSKY M, WILLET R M, HARMANY Z T, et al. Compressed sensing performance bounds under Poisson noise[J].IEEE Transactions on Signal Processing, 2010,58(8):3990－4002.

[4] LI Y H, CEVHER V. Consistency of ℓ_1-regularized maximum-likelihood for compressive Poisson regression[C]// IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2015:3606－3610.

[5] IVANOFF S, PICARD F, RIVOIRARD V. Adaptive Lasso and group-Lasso for functional Poisson regression[J].Journal of Machine Learning Research, 2016,17(55):1－46.

[6] JIA J Z, XIE F, XU L H. Sparse Poisson regression with penalized weighted score function[J].Electronic Journal of Statistics, 2019,13(2):2898－2920.

[7] LOH P L, WAINWRIGHT M J. High-dimensional regression with noisy and missing data: provable guarantees with non-convexity[J]. The Annals of Statistics, 2012,40(3):1637－1664.

[8] LI M Y, LI R Z, MA Y Y. Inference in high dimensional linear measurement error models[J].Journal of Multivariate Analysis, 2021,184:104759.

[9] FLETCHER R, WATSON G A. First and second order conditions for a class of nondifferentiable optimization problems[J].Mathematical Programming, 1980,18(1):291－307.

[10] TIBSHIRANI R. Regression shrinkage and selection via the lasso[J].Journal of the Royal Statistical Society: Series B(Methodological), 1996,58(1):267－288.

[11] DUCHI J, SHALEV-SHWARTZ S, SINGER Y, et al. Efficient projections onto the ℓ_1-ball for learning in high dimensions[C]//Proceedings of the 25th International Conference on Machine Learning (ICML 2008), 2008:272－279.

[12] HE K M, GKIOXARI G, DOLLÁR P, et al. Mask R-CNN[C]//Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), IEEE, 2017:2980－2988.

[13] FRIEDMAN J, HASTIE T, TIBSHIRANI R. Regularization paths for generalized linear models via coordinate descent[J].Journal of Statistical Software, 2010,33(1):1－22.

Variable Selection in Poisson Regression Model with Measurement Error

LIU Fang

(Registrar's Office, Gannan Normal University, Ganzhou 341000, China)

Abstract: Poisson regression is a special generalized linear model which is widely used to model count data. However, in many applications, covariates are often contaminated with errors. Errors in these covariates are usually difficult to handle, especially when the covariate dimension is high. This paper proposes a bias-corrected penalized method to determine the underlying structure of Poisson regression model with measurement errors (MPR). The procedure penalizes a target function that is specifically designed to handle measurement errors and provides the composite gradient descent algorithm within a bounded region. The numerical performance is demonstrated using simulation studies.

Keywords: Poisson regression model; measurement error; variable selection; penalty function