Advances in Applied Mathematics 应用数学进展, 2026, 15(6), 219-229
Published Online June 2026 in Hans. https://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2026.156280

Hans汉斯

基于Poisson回归模型因果效应的目标最大似然估计

李思默，侯　文*

辽宁师范大学数学学院，辽宁　大连

收稿日期：2026年5月16日；录用日期：2026年6月7日；发布日期：2026年6月16日

摘　要

观察性研究中，估计二值处理的平均因果效应(ATE)时，传统回归方法易受模型误设影响。目标最大似然估计(TMLE)是一种半参数双稳健方法，仅需结果模型或倾向得分模型之一正确即可获得一致估计。本文以Poisson计数结果为背景，系统介绍TMLE的原理与算法，并通过蒙特卡罗模拟比较TMLE与其余方法在模型正确、倾向得分误设、结果模型误设三种场景下的表现。模拟结果表明，TMLE在所有场景下均保持低偏差和较小的均方根误差，表现出双稳健性。实例分析进一步验证了TMLE在真实计数数据中的实用性。TMLE是估计Poisson型ATE的可靠方法，建议作为观察性研究中计数结局因果推断的首选工具之一。

关键词

因果推断，目标最大似然估计，Poisson回归模型，双稳健估计

Targeted Maximum Likelihood Estimation for Causal Effects Based on Poisson Regression Models

Simo Li, Wen Hou*

School of Mathematics, Liaoning Normal University, Dalian Liaoning

Received: May 16, 2026; accepted: June 7, 2026; published: June 16, 2026

Abstract

In observational studies, conventional regression methods for estimating the average treatment

*通讯作者。

文章引用: 李思默, 侯文. 基于 Poisson 回归模型因果效应的目标最大似然估计[J]. 应用数学进展, 2026, 15(6): 219-229. DOI: 10.12677/aam.2026.156280

effect (ATE) of a binary treatment are vulnerable to model misspecification. Targeted maximum likelihood estimation (TMLE) is a semiparametric doubly robust method that requires only one of the outcome model or the propensity score model to be correctly specified to obtain a consistent estimate. Focusing on Poisson count outcomes, this paper systematically introduces the principles and algorithm of TMLE, and compares TMLE with other methods via Monte Carlo simulations under three scenarios: correct model specification, misspecified propensity score, and misspecified outcome model. Simulation results show that TMLE maintains low bias and small root mean squared error across all scenarios, demonstrating double robustness. An empirical example further validates its practical utility with real count data. TMLE is a reliable method for estimating Poisson-type ATE and is recommended as a preferred tool for causal inference with count outcomes in observational studies.

Keywords

Causal Inference, Targeted Maximum Likelihood Estimation, Poisson Regression Model, Doubly Robust Estimation

1. 引言

在观察性研究中，由于处理组与对照组在协变量分布上往往不可比，直接比较结果均值会产生混杂偏倚。为了从观察数据中估计处理对结果的因果效应，研究者常依赖潜在结果框架[1]并引入条件可交换性、正则性、一致性等假设，此外还需假设无干扰(即个体的潜在结果不受其他个体处理状态的影响)，该假设通常在研究设计部分以文字形式讨论其合理性[2]。在这些假设下，平均处理效应(ATE)可以被识别。

传统的协变量调整方法(如将处理变量作为协变量纳入回归模型)要求结果模型正确设定，且效应同质，这在实践中往往难以满足。倾向得分方法[3]通过平衡协变量分布来估计 ATE，但要求倾向得分模型正确设定。G-computation [4]直接标准化结果模型，同样依赖结果模型的正确性。

为降低模型误设带来的偏倚，双稳健方法应运而生。Bang 与 Robins [5]提出了双稳健估计量的理论框架，随后增强型逆概率加权(AIPTW)和目标最大似然估计(TMLE) [6]被发展出来。TMLE 具有替换估计量的优势，估计值始终位于参数空间内，且在大样本下达到半参数有效界[7]。

目前大多数 TMLE 教程聚焦于二元或连续结局[8]。然而，在许多医学和公共卫生研究中，结局变量常为计数数据(如住院次数、急诊就诊次数、肿瘤复发次数)。处理计数数据时，Poisson 回归是自然的起点[9]，但实际数据往往存在过度分散[10]。因此，将 TMLE 扩展至 Poisson 结果具有重要实用价值。为此，本文以 Poisson 结果为背景，提供 TMLE 的逐步实现教程，并通过蒙特卡罗模拟比较不同方法在模型正确与误设场景下的表现。

本文的结构如下：第 2 节介绍反事实因果的因果推断框架和 TMLE 方法；第 3 节报告模拟研究的设计与结果；第 4 节为实例分析；第 5 节讨论主要发现和实际应用建议，并给出结论。

2. 方法框架

2.1. 因果推断框架(反事实因果)

令 $A \in \{0,1\}$ 表示二值处理(例如新疗法 vs 标准疗法)，Y 表示结果变量(计数，如住院天数)，W 表示处

理前协变量向量。每个个体存在两个潜在结果：$Y(1)$表示接受处理时的结果，$Y(0)$表示未接受处理时的结果。观测结果由一致性假设给出：$Y = AY(1)+(1-A)Y(0)$。

目标参数为平均处理效应(ATE)：

$$\mathrm{ATE} = E\left[Y(1)-Y(0)\right]. \tag{1}$$

为了从观测数据中识别 ATE，需要以下三个可用数学公式表达的假设：

条件可交换性(Conditional Exchangeability)：

$$\left\{Y(1),Y(0)\right\} \perp A|\boldsymbol{W}. \tag{2}$$

这意味着在给定协变量$\boldsymbol{W}$的条件下，处理分配 A 与潜在结果$\left\{Y(1),Y(0)\right\}$相互独立。等价地，对于任意$a \in \{0,1\}$

$$\mathrm{E}\left[Y(a)|A=1,\boldsymbol{W}\right] = \mathrm{E}\left[Y(a)|A=0,\boldsymbol{W}\right] = \mathrm{E}\left[Y(a)|\boldsymbol{W}\right]. \tag{3}$$

该假设要求所有影响处理分配和结果的共同原因(混杂因子)均已测量并包含在$\boldsymbol{W}$中。

正则性(Positivity)：

$$0 < P(A=1|\boldsymbol{W}) < 1，\text{几乎必然成立.} \tag{4}$$

记倾向得分$e(\boldsymbol{W}) = P(A=1|\boldsymbol{W})$，则要求$0 < e(\boldsymbol{W}) < 1$对所有$\boldsymbol{W}$成立。该假设确保在每个协变量层内都存在处理组和对照组的个体，从而允许进行组间比较。

一致性(Consistency)：

$$Y = A\cdot Y(1)+(1-A)\cdot Y(0). \tag{5}$$

等价地，若$A=a$，则$Y=Y(a)$。该假设建立了观测数据与潜在结果之间的桥梁，使得可以用观测结果替代反事实结果。

在上述假设下，ATE 可通过以下公式识别：

$$\mathrm{ATE} = \mathrm{E}_{\boldsymbol{W}}\left[\mathrm{E}(Y|A=1,\boldsymbol{W}) - \mathrm{E}(Y|A=0,\boldsymbol{W})\right]. \tag{6}$$

2.1.1. 初始结果模型估计

假设结果变量Y在给定$A,\boldsymbol{W}$的条件下服从 Poisson 分布，即$Y|A,\boldsymbol{W} \sim \mathrm{Poisson}(\lambda)$其中$\lambda = \mathrm{E}(Y|A,\boldsymbol{W})$。对于计数结果变量$Y$，Poisson 回归模型采用对数链接函数(loglink)，将条件期望的对数表示为协变量的线性组合[9]：

$$\log \bar{Q}^0(A,\boldsymbol{W}) = \beta_0 + \beta_1 A + \sum_{j=1}^{p}\beta_{j+1}W_j, \tag{7}$$

其中$\bar{Q}^0(A,\boldsymbol{W}) = \mathrm{E}(Y|A,\boldsymbol{W})$为初始估计；$\beta_0,\beta_1,\cdots,\beta_{p+1}$为回归系数，通过最大似然估计得到；$W_j$为第 j 个协变量，p为协变量个数。上式等价于：

$$\bar{Q}^0(A,\boldsymbol{W}) = \exp\left(\beta_0 + \beta_1 A + \sum_{j=1}^{p}\beta_{j+1}W_j\right). \tag{8}$$

拟合后，对每个个体 i 预测：

$$\bar{Q}^0(0,\boldsymbol{W}_i) = \exp\left(\hat{\beta}_0 + \sum_{j=1}^{p}\hat{\beta}_{j+1}W_{ij}\right),\ \bar{Q}^0(1,\boldsymbol{W}_i) = \exp\left(\hat{\beta}_0 + \hat{\beta}_1 + \sum_{j=1}^{p}\hat{\beta}_{j+1}W_{ij}\right). \tag{9}$$

2.1.2. 倾向得分估计

使用 Logistic 回归：

$$\log \hat{g}(\boldsymbol{W}_i) = \log\left(\frac{\hat{g}(\boldsymbol{W}_i)}{1-\hat{g}(\boldsymbol{W}_i)}\right) = \alpha_0 + \sum_{j=1}^{p} \alpha_j W_{ij}, \tag{10}$$

其中 $\hat{g}(\boldsymbol{W}_i) = P(A_i = 1 | \boldsymbol{W}_i)$， $\alpha_0, \alpha_1, \cdots, \alpha_p$ 为 Logistic 回归系数。

2.1.3. 波动协变量与波动参数估计

设定波动协变量(fluctuation covariate)

$$H(1, \boldsymbol{W}_i) = \frac{A_i}{\hat{g}(\boldsymbol{W}_i)}, \; H(0, \boldsymbol{W}_i) = \frac{1-A_i}{1-\hat{g}(\boldsymbol{W}_i)}. \tag{11}$$

这两个协变量的作用是“目标化”初始结果模型，以吸收倾向得分的额外信息。

然后拟合如下 Poisson 回归模型(不包含截距项)：

$$\log \mathrm{E}(Y_i | A_i, \boldsymbol{W}_i)(\epsilon) = \log \bar{Q}^0(A_i, \boldsymbol{W}_i) + \epsilon_0 H(0, \boldsymbol{W}_i) + \epsilon_1 H(1, \boldsymbol{W}_i), \tag{12}$$

其中 $\log \bar{Q}^0(A_i, \boldsymbol{W}_i)$ 为已知项，其系数固定为 1 (即保持初始预测的对数值不变)； $\epsilon = (\epsilon_0, \epsilon_1)$ 为需要估计的波动参数。这一设定确保更新后的预测值 $\hat{Q}^* = \hat{Q}^0 \cdot \exp(\hat{\epsilon} H)$ 始终为正，且仅对初始模型进行最小幅度的调整。通过估计 ϵ，可以消除初始结果模型中可能存在的残差混杂，从而获得双稳健的 TMLE 估计量。ϵ 的估计通过最大似然法在 Poisson 回归中完成，得到 $\hat{\epsilon}_0$ 和 $\hat{\epsilon}_1$。

3. 模拟研究

符号说明：本节沿用上述符号定义，其中 A 表示二值处理变量，Y 表示结果变量，$\boldsymbol{W} = (W_1, W_2, W_3, W_4)$ 为协变量向量，$Y(0), Y(1)$ 为潜在结果，$g(\boldsymbol{W}) = P(A = 1 | \boldsymbol{W})$ 为倾向得分，$\bar{Q}(A, \boldsymbol{W}) = \mathrm{E}(Y | A, \boldsymbol{W})$ 为条件结果均值，ATE 为平均处理效应。

3.1. 模拟设计

我们通过蒙特卡罗模拟比较四种 ATE 估计方法：朴素均值差(Naive)、G-computation、增强型逆概率加权(AIPTW)和 TMLE。数据生成过程如下。

协变量独立生成：W_1~Bernoulli (0.5)，W_2~Bernoulli (0.6)，W_3~Uniform (0, 10)，W_4~Normal (0, 1)。

处理分配模型(倾向得分模型)：

$$\operatorname{logit} P(A = 1 | \boldsymbol{W}) = -1 + 0.5W_1 + 0.5W_2 + 0.1W_3 + 0.2W_4. \tag{13}$$

结果模型(Poisson)：

$$\log \mathrm{E}[Y(0) | \boldsymbol{W}] = 0.5 + 0.2W_1 + 0.2W_2 + 0.05W_3 + 0.2W_4, \tag{14}$$

$$\log \mathrm{E}[Y(1) | \boldsymbol{W}] = \log \mathrm{E}[Y(0) | \boldsymbol{W}] + 0.6. \tag{15}$$

观测结果由一致性确定： $Y = AY(1) + (1-A)Y(0)$。

真实 ATE 通过一个超大样本($n = 5 \times 10^6$)计算得到。在三种场景下分别为：

场景一(模型正确设定)：分析模型与数据生成模型一致(均只含主效应)。真实 ATE = 0.5983。

场景二(倾向得分模型误设)：数据生成时倾向得分模型包含交互项 W_1W_2 (系数 0.8)，分析时仅用主效应(误设)。结果模型正确。真实 ATE = 0.6120。

场景三(结果模型误设)：数据生成时结果模型包含交互项 W_1W_2（系数 0.8)，分析时仅用主效应(误设)。倾向得分模型正确。真实 ATE = 0.8509。

每个场景下，样本量分别取 $n=100,200,500$，每种设定重复 $R=1000$ 次。

3.2. 四种估计方法的计算

我们比较以下四种 ATE 估计方法：

(a) 朴素均值差(Naive)：直接比较处理组与对照组的结果样本均值，即

$$\widehat{\mathrm{ATE}}_{\mathrm{Naive}}=\frac{1}{n_1}\sum_{i:A_i=1}Y_i-\frac{1}{n_0}\sum_{i:A_i=0}Y_i, \tag{16}$$

其中 $n_1=\sum_i A_i$，$n_0=n-n_1$。该方法未调整混杂，通常存在偏倚。

(b) G-computation：基于第 2.1.1 节中拟合的 Poisson 结果模型(仅含主效应)，对每个个体预测 $\hat{Q}_i(1)=\hat{\mathrm{E}}(Y\mid A=1,\boldsymbol{W}_i)$ 和 $\hat{Q}_i(0)=\hat{\mathrm{E}}(Y\mid A=0,\boldsymbol{W}_i)$，然后取样本平均差：

$$\widehat{\mathrm{ATE}}_{\text{G-comp}}=\frac{1}{n}\sum_{i=1}^{n}\left[\hat{Q}_i(1)-\hat{Q}_i(0)\right]. \tag{17}$$

(c) 增强型逆概率加权(AIPTW)：利用第 2.1.2 节估计的倾向得分 $\hat{g}(\boldsymbol{W})$，结合结果模型的预测值，计算

$$\begin{aligned}\hat{\mathrm{E}}\left[Y(1)\right]&=\frac{1}{n}\sum_{i=1}^{n}\left(\frac{A_i}{\hat{g}(\boldsymbol{W}_i)}\left(Y_i-\hat{Q}_i(1)\right)+\hat{Q}_i(1)\right),\\ \hat{\mathrm{E}}\left[Y(0)\right]&=\frac{1}{n}\sum_{i=1}^{n}\left(\frac{1-A_i}{1-\hat{g}(\boldsymbol{W}_i)}\left(Y_i-\hat{Q}_i(0)\right)+\hat{Q}_i(0)\right).\end{aligned} \tag{18}$$

则

$$\widehat{\mathrm{ATE}}_{\mathrm{AIPTW}}=\hat{\mathrm{E}}\left[Y(1)\right]-\hat{\mathrm{E}}\left[Y(0)\right]. \tag{19}$$

(d) 目标最大似然估计(TMLE)：按照第 2.1.3 节描述的波动过程，最终

$$\widehat{\mathrm{ATE}}_{\mathrm{TMLE}}=\frac{1}{n}\sum_{i=1}^{n}\left[\hat{Q}^{*}(1,\boldsymbol{W}_i)-\hat{Q}^{*}(0,\boldsymbol{W}_i)\right]. \tag{20}$$

3.3. 评估指标

对于每种方法，基于 1000 次重复模拟计算以下指标：

(a) 绝对偏差(Absolute Bias)：

$$\mathrm{Abs\_Bias}=\left|\frac{1}{R}\sum_{r=1}^{R}\widehat{\mathrm{ATE}}_r-\mathrm{ATE}_{\mathrm{true}}\right|. \tag{21}$$

(b) 均方根误差(RMSE)：

$$\mathrm{RMSE}=\sqrt{\frac{1}{R}\sum_{r=1}^{R}\left(\widehat{\mathrm{ATE}}_r-\mathrm{ATE}_{\mathrm{true}}\right)^2}. \tag{22}$$

3.4. 模拟结果

表 1 汇总了三种场景下四种估计方法的绝对偏差和均方根误差(RMSE)。

在场景一(模型正确)下，由表 1 第 3、4 列所示，Naive 偏差大(约 0.39~0.40)，不随样本量减小；G-

computation、AIPTW 和 TMLE 偏差均很小(<0.015)，RMSE 随样本量增大下降。TMLE 表现与两者相当。

在场景二(倾向得分误设)下，由表 1 第 5、6 列所示，G-computation 因结果模型正确仍保持低偏差；AIPTW 和 TMLE 同样偏差很小，RMSE 随样本量增加而下降。TMLE 在倾向得分误设时依然稳健，RMSE 在 $n=500$ 时仅 0.1993，远优于 Naive。

在场景三(结果模型误设)下，由表 1 第 7、8 列所示，G-computation 偏差明显增大($n=100$ 时 0.0561)，而 AIPTW 和 TMLE 因倾向得分正确，偏差保持很小(TMLE 在 $n=100$ 时 0.0470，$n=200$ 时降至 0.0048)。TMLE 的 RMSE 与 AIPTW 相当，显著低于 G-computation 和 Naive。

模拟结果表明，朴素均值差在所有场景下均存在严重偏倚，不推荐使用。G-computation 在结果模型正确时表现优异，但一旦结果模型误设则偏倚明显增大。AIPTW 和 TMLE 具有双稳健性，在任一模型正确时均能保持较小的偏倚。随着样本量增大，所有方法的 RMSE 均下降，符合大样本理论。综合来看，TMLE 在模型误设场景下表现稳健，是估计计数结果平均处理效应的可靠方法。

Table 1. Absolute bias and RMSE of four estimation methods under different scenarios
表 1. 四种估计方法在不同场景下的绝对偏差与 RMSE

		场景一(模型正确)		场景二(倾向得分误设)		场景三(结果模型误设)	
n	方法	绝对偏差	RMSE	绝对偏差	RMSE	绝对偏差	RMSE
100	Naive	0.4018	0.5745	0.5950	0.7417	1.2858	1.5575
	G-computation	0.0116	0.3768	0.0172	0.4105	0.0561	0.5550
	AIPTW	0.0145	0.3807	0.0204	0.4297	0.0462	0.5640
	TMLE	0.0142	0.3825	0.0164	0.4322	0.0470	0.5666
200	Naive	0.3970	0.4897	0.5870	0.6627	1.3227	1.4644
	G-computation	0.0038	0.2611	0.0124	0.3023	0.0184	0.3868
	AIPTW	0.0035	0.2617	0.0124	0.3118	0.0036	0.3910
	TMLE	0.0041	0.2626	0.0113	0.3140	0.0048	0.3918
500	Naive	0.3902	0.4301	0.5911	0.6272	1.3005	1.3543
	G-computation	0.0102	0.1634	0.0122	0.1948	0.0255	0.2386
	AIPTW	0.0103	0.1641	0.0094	0.1986	0.0121	0.2396
	TMLE	0.0102	0.1642	0.0094	0.1993	0.0121	0.2397

4. 实例分析

4.1. 数据来源与说明

本文研究数据取自 R 语言 MASS 包内置的 Insurance 数据集[11]。该数据记录了 1973 年第三季度某保险公司投保人的汽车保险索赔情况，包含 64 条聚合记录，每条代表一个协变量组合下的汇总信息。变量见表 2 (原始 Group 变量重组为处理变量 A)。

在计数数据分析中，观测结果 Y 往往依赖于某种“暴露量”或“基数”(如观测时间、面积、人口数)[9] [10]。例如，不同个体的保单持有者数量不同，直接比较原始索赔次数会因暴露量差异而产生偏误。此时，需要在 Poisson 回归模型中引入偏移量(offset)，即令：

$$\log E\left(Y\mid\cdot\right)=\log\left(\text{Exposure}\right)+\eta, \tag{23}$$

其中 η 为协变量的线性组合。等价地，$\log\left(E\left(Y\right)/\text{Exposure}\right)=\eta$，即模型实际预测的是单位暴露量的发生率。偏移项 log(Exposure)的系数被固定为 1，不参与参数估计，其作用是标准化不同个体的暴露水平。这一处理是 Poisson 回归中的标准步骤。为避免不同观测的保单持有者数量(Holders)对索赔次数的影响，本文后续所有 Poisson 模型(包括初始结果模型和 TMLE 的波动模型)均以 log(Holders)作为偏移量(offset)。此时模型实际建模的目标为 log(Claims/Holders)，即单位保单持有者的索赔发生率，从而在固定暴露量的基础上估计其他协变量(如排量组 *A*、地区 District、年龄 Age)对索赔风险的影响。

Table 2. Variable description
表 2. 变量说明

变量类型	变量名称	说明	取值
因变量	Claims	索赔次数(计数)	整数
自变量	A	汽车排量组(二值)	0：小于等于 1.5 升，1：大于 1.5 升
	District	居住地区	1，2，3，4 (4 表示主要城市)
	Age	驾驶员年龄组(有序)	1：小于 25 岁，2：25 至 29 岁，3：30 至 35 岁，4：大于 35 岁
	log(Holders)	保单持有人数的自然对数	连续数值(1.10 至 8.18)

4.2. 描述性分析

在进行因果推断之前，我们先对关键变量进行描述性分析，以初步了解数据特征。表 3 展示了因变量索赔次数(Claims)的详细统计量。

Table 3. Descriptive statistics of variables
表 3. 变量描述性统计

变量	样本量	最小值	第一四分位数	中位数	均值	第三四分位数	最大值
Claims	64	0	9.5	22.0	49.2	35.5	400

表 4 展示了分类变量的频数分布，每个分类变量的各个水平均为均衡分布，这一均衡性有助于后续的因果推断。

Table 4. Frequency distribution of categorical variables
表 4. 分类变量频数分布

变量	水平	频数	百分比(%)
A	0	32	50.0
	1	32	50.0
District	1	16	25.0
	2	16	25.0
	3	16	25.0
	4	16	25.0
Age	1	16	25.0
	2	16	25.0
	3	16	25.0
	4	16	25.0

4.3. 因果推断关键假设的合理性论证

本节基于 Insurance 数据集，系统评估条件可交换性、正则性和一致性三大假设的合理性。

(1) 条件可交换性：该假设要求给定协变量 $\boldsymbol{W}$ 后，处理分配 A 与潜在结果 $\{Y(1),Y(0)\}$ 独立。本数据中协变量包括地区(District)、年龄(Age)和保单持有数(Holders)，它们同时影响排量选择与索赔次数。例如，年轻驾驶员更倾向大排量汽车且事故率更高；城市地区(District = 4)交通密集且大排量车比例高。因此，调整这些变量可在一定程度上阻断混杂路径。**局限性**：数据未提供驾驶行为(如年均里程、违章记录)等潜在强混杂因素，若这些因素与排量和索赔风险均相关，则条件可交换性可能被违背。建议未来研究纳入此类变量或进行敏感性分析(如 E-value 计算)。

(2) 正则性：倾向得分范围在 0.15 至 0.90 之间，未出现极端接近 0 或 1 的值，且各协变量分层内均存在处理组与对照组个体(见表 4)，假设满足。

(3) 一致性：处理定义明确(排量 >1.5 升 vs ≤1.5 升)，不同保单持有者的索赔行为可视为相互独立(无干扰)，假设合理。

4.4. TMLE 估计结果

本节按照目标最大似然估计(TMLE)的标准流程，依次估计初始结果模型、倾向得分模型，并通过波动协变量对初始预测进行目标化更新，最终得到平均处理效应(ATE)的 TMLE 估计量。

4.4.1. 初始结果模型

首先拟合一个不含交互项的 Poisson 回归模型作为初始结果模型 $\bar{Q}^0(A,\boldsymbol{W})=\mathrm{E}(Y\mid A,\boldsymbol{W})$，协变量包括处理变量 A、地区(District)、年龄数值编码(Age_num)，并以保单持有人数的对数 log(Holders)作为偏移量(offset)。模型形式为：

$$\log \mathrm{E}(Y\mid A,\boldsymbol{W})=\log(\mathrm{Holders})+\beta_0+\beta_1 A+\beta_2\mathrm{District}_2+\beta_3\mathrm{District}_3+\beta_4\mathrm{District}_4+\beta_5\mathrm{Age\_num}. \quad (24)$$

其中年龄按 1，2，3，4 编码。回归结果如表 5 所示。处理变量 A 的系数显著为正(0.331, $p<0.001$)，表明大排量汽车平均索赔率更高；年龄系数显著为负(−0.190, $p<0.001$)，说明索赔率随年龄增长呈下降趋势；District4 相比 District1 有更高的索赔率(0.265, $p=0.002$)。

Table 5. Poisson initial outcome model
表 5. Poisson 初始结果模型

变量	系数	标准误	z 值	p 值
(Intercept)	−1.607	0.154	−10.467	<0.001
A	0.331	0.046	7.207	<0.001
District2	0.037	0.046	0.810	0.418
District3	0.059	0.060	0.981	0.327
District4	0.265	0.085	3.124	0.002
Age_num	−0.190	0.038	−4.942	<0.001
log(Holders)	1.016	0.034	29.504	<0.001

4.4.2. 倾向得分模型

采用 Logistic 回归将处理变量 A 对地区(District)、年龄数值(Age_num)和 log(Holders)进行回归，以平衡处理组与对照组之间的协变量分布并构造后续的波动协变量。模型为：

$$\log\frac{P(A=1|\boldsymbol{W})}{1-P(A=1|\boldsymbol{W})}=\alpha_0+\alpha_1\text{District}_2+\alpha_2\text{District}_3+\alpha_3\text{District}_4+\alpha_4\text{Age\_num}+\alpha_5\log(\text{Holders}). \quad (25)$$

估计结果如表 6 所示。District3、District4、Age_num 以及 log(Holders)均与处理分配显著相关($p < 0.05$)，说明这些协变量对排量选择有较强的预测能力，存在混杂。倾向得分范围为 0.15 至 0.90，均值为 0.50，未出现极端接近 0 或 1 的值，满足正则性假设。

Table 6. Logistic regression of treatment A on covariates
表 6. Logistic 回归：处理变量 A 对协变量的回归结果

变量	系数	标准误	z 值	p 值
(Intercept)	7.748	2.344	3.305	<0.001
District2	−1.167	0.956	−1.220	0.222
District3	−2.529	1.128	−2.243	0.025
District4	−3.894	1.365	−2.853	0.004
Age_num	1.878	0.539	3.481	<0.001
log(Holders)	−2.127	0.549	−3.873	<0.001

4.4.3. 倾向得分平衡诊断(SMD)

为评估倾向得分模型的平衡能力，计算各协变量在加权前后的标准化均差(SMD)。采用倾向得分逆概率权重(IPW)进行加权。一般要求 SMD < 0.1 表明平衡良好。结果如表 7 所示。加权前，District 和 Age 的 SMD 为 0 (因数据本身完全平衡设计)，logHolders 的 SMD 为−0.673；加权后所有协变量的 SMD 绝对值均小于 0.25，且大部分小于 0.1，说明倾向得分模型有效平衡了协变量分布。

Table 7. Standardized mean differences (SMD) of covariates before and after IPW weighting
表 7. 倾向得分加权前后各协变量的标准化均差(SMD)

协变量	未加权 SMD	加权后 SMD
District_1	0.000	−0.010
District_2	0.000	0.002
District_3	0.000	0.016
District_4	0.000	−0.008
Age_num	0.000	−0.041
logHolders	−0.673	−0.217

4.4.4. Poisson 模型过度分散检验

计数数据常存在过度分散(overdispersion)，即方差大于均值。本文采用两种方法检验：

1) 拟合优度卡方检验：Pearson $\chi^2 = 65.996$，残差自由度 df = 58，过度分散因子 = 1.138 ($p = 0.220$)。2) 负二项模型比较：拟合负二项回归模型(与 Poisson 模型相同的协变量及 offset)，似然比检验统计量 LRT = 0.432 ($p = 0.511$)，负二项模型的过度分散参数 $\theta = 572.06$。综上，数据不存在显著过度分散，Poisson 模型适用。

4.4.5. TMLE 估计的 ATE

在获得初始结果模型和倾向得分模型后，构造波动协变量：

$$H(1,\boldsymbol{W}_i)=\frac{A_i}{\hat{g}(\boldsymbol{W}_i)},\ H(0,\boldsymbol{W}_i)=\frac{1-A_i}{1-\hat{g}(\boldsymbol{W}_i)}, \tag{26}$$

其作用是“目标化”初始结果模型，以吸收倾向得分的额外信息。然后拟合一个不含截距项的 Poisson 回归模型(仍以 log(Holders)为偏移量)：

$$\log \mathrm{E}(Y_i \mid A_i,\boldsymbol{W}_i)(\epsilon)=\log \hat{Q}^0(A_i,\boldsymbol{W}_i)+\epsilon_0 H(0,\boldsymbol{W}_i)+\epsilon_1 H(1,\boldsymbol{W}_i), \tag{27}$$

其中 $\log \hat{Q}^0(A_i,\boldsymbol{W}_i)$ 作为偏移量项(系数固定为 1)，$\epsilon=(\epsilon_0,\epsilon_1)$ 是需要估计的波动参数。通过最大似然法估计得到 $\hat{\epsilon}_0$ 和 $\hat{\epsilon}_1$，进而更新结果模型预测值：

$$\hat{Q}^*(1,\boldsymbol{W}_i)=\hat{Q}^0(1,\boldsymbol{W}_i)\cdot\exp\left(\hat{\epsilon}_1 H(1,\boldsymbol{W}_i)\right),\ \hat{Q}^*(0,\boldsymbol{W}_i)=\hat{Q}^0(0,\boldsymbol{W}_i)\cdot\exp\left(\hat{\epsilon}_0 H(0,\boldsymbol{W}_i)\right). \tag{28}$$

最终，TMLE 对平均处理效应的估计量为：

$$\widehat{\mathrm{ATE}}_{\mathrm{TMLE}}=\frac{1}{n}\sum_{i=1}^{n}\left[\hat{Q}^*(1,\boldsymbol{W}_i)-\hat{Q}^*(0,\boldsymbol{W}_i)\right]. \tag{29}$$

在本实例中，估计得到 ATE = 13.41 ($p < 0.001$)。这意味着在调整了地区、年龄及风险暴露后，大排量汽车比小排量汽车平均多索赔约 13.4 次。

本实例分析表明，倾向得分模型有效识别了影响排量选择的混杂因素(地区、年龄、保单持有量)，且平衡性良好。描述性分析显示索赔次数分布右偏，大排量组的索赔次数整体高于小排量组。Poisson 回归(含偏移量)显示数据不存在显著过度分散，适合采用 Poisson 模型；年龄对索赔次数的影响呈线性下降趋势。采用 TMLE 估计得到 ATE 为 13.41，说明大排量汽车显著增加了索赔次数，与保险实务中“大排量车辆风险更高”的认知一致。

5. 讨论与结论

本文针对 Poisson 计数结果，介绍了目标最大似然估计(TMLE)的原理与实现，并通过蒙特卡罗模拟和实例分析验证了其性能。模拟结果表明，TMLE 在模型正确、倾向得分误设、结果模型误设三种场景下均保持低偏差和较小的均方根误差，表现出双稳健性。实例分析中，我们系统讨论了因果推断假设的合理性，展示了倾向得分平衡诊断(SMD 表)和过度分散检验结果，进一步支持了 TMLE 在实际计数数据中的可用性。

应用 TMLE 时需注意：验证正则性(倾向得分避免极端值)，对计数数据应进行条件过度分散检验，小样本下谨慎解释推断结果。本研究模拟仅涉及简单交互项，未来可探索 TMLE 与机器学习算法的结合，以及纵向计数数据或零膨胀情形的扩展。

综上，TMLE 是估计 Poisson 型平均处理效应的可靠方法，建议在观察性研究中作为计数结局因果推断的常规工具。

基金项目

本研究由 2022 年度辽宁省研究生教育教学改革研究项目(2022-180-39510165)资助。

参考文献

[1] Rubin, D.B. (1974) Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies. *Journal of*

Educational Psychology, **66**, 688-701. https://doi.org/10.1037/h0037350

[2] Hernán, M.A. and Robins, J.M. (2020) Causal Inference: What If. Chapman & Hall/CRC.

[3] Rosenbaum, P.R. and Rubin, D.B. (1983) The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika*, **70**, 41-55. https://doi.org/10.1093/biomet/70.1.41

[4] Robins, J. (1986) A New Approach to Causal Inference in Mortality Studies with a Sustained Exposure Period—Application to Control of the Healthy Worker Survivor Effect. *Mathematical Modelling*, **7**, 1393-1512. https://doi.org/10.1016/0270-0255(86)90088-6

[5] Bang, H. and Robins, J.M. (2005) Doubly Robust Estimation in Missing Data and Causal Inference Models. *Biometrics*, **61**, 962-973. https://doi.org/10.1111/j.1541-0420.2005.00377.x

[6] van der Laan, M.J. and Rubin, D. (2006) Targeted Maximum Likelihood Learning. *The International Journal of Biostatistics*, **2**, Article No. 213. https://doi.org/10.2202/1557-4679.1043

[7] van der Laan, M.J. and Rose, S. (2011) Targeted Learning: Causal Inference for Observational and Experimental Data. Springer.

[8] Luque-Fernandez, M.A., Schomaker, M., Rachet, B. and Schnitzer, M.E. (2018) Targeted Maximum Likelihood Estimation for a Binary Treatment: A Tutorial. *Statistics in Medicine*, **37**, 2530-2546. https://doi.org/10.1002/sim.7628

[9] McCullagh, P. and Nelder, J.A. (1989) Generalized Linear Models. 2nd Edition, Chapman and Hall.

[10] Venables, W.N. and Ripley, B.D. (2002) Modern Applied Statistics with S. 4th Edition, Springer.

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