

基于逆概率加权的广义经验似然方法在平均处理效应估计中的应用

宁高佳, 侯文*

辽宁师范大学数学学院, 辽宁 大连

收稿日期: 2025年4月12日; 录用日期: 2025年5月6日; 发布日期: 2025年5月13日

摘要

传统的经验似然方法在处理复杂模型时存在计算效率低和灵活性不足的局限, 广义经验似然方法通过引入更一般的权重函数和优化目标, 提高了模型的适应性和计算效率。本文分析了名为NSW的数据集, 其中包含职业培训对收入影响的相关数据。基于线性模型和logit模型的得分函数, 构造矩条件, 通过广义经验似然方法得到了平均处理效应的估计结果。对比于单一的逆概率加权方法, 基于逆概率加权的广义经验似然方法的估计结果较为稳健。

关键词

广义经验似然, 逆概率加权, 平均处理效应, 倾向得分, 因果推断

Application of the Generalized Empirical Likelihood Method Based on Inverse Probability Weighting in the Estimation of Average Treatment Effects

Gaojia Ning, Wen Hou*

School of Mathematics, Liaoning Normal University, Dalian Liaoning

Received: Apr. 12th, 2025; accepted: May 6th, 2025; published: May 13th, 2025

Abstract

The traditional empirical likelihood method has limitations in terms of computational efficiency and

*通讯作者。

文章引用: 宁高佳, 侯文. 基于逆概率加权的广义经验似然方法在平均处理效应估计中的应用[J]. 应用数学进展, 2025, 14(5): 72-81. DOI: 10.12677/aam.2025.145234

flexibility when dealing with complex models. The generalized empirical likelihood method addresses these issues by introducing more general weight functions and optimization objectives, thereby enhancing the adaptability and computational efficiency of the model. This paper analyzes the NSW dataset, which examines the impact of vocational training on income. Based on the score functions of the linear model and the logit model, moment conditions are constructed, and the average treatment effect is estimated using the generalized empirical likelihood method. Compared to the single inverse probability weighting method, the estimation results of the generalized empirical likelihood method based on inverse probability weighting are more robust.

Keywords

Generalized Empirical Likelihood, Inverse Probability Weighting, Average Treatment Effect, Propensity Score, Causal Inference

Copyright © 2025 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

经验似然(Empirical Likelihood, EL)方法作为一种非参数或半参数的统计推断方法,自 Owen [1]提出以来,受到了广泛的关注和应用。经验似然方法通过最大化经验似然函数,能够在不需要假设数据分布的情况下,灵活地处理复杂的统计问题。其核心思想是利用观测数据的经验分布来构造似然函数,从而进行参数估计和假设检验。经验似然方法的一个重要优势在于其能够自然地结合矩条件,处理带有约束条件的统计模型。

然而,经验似然方法在处理复杂模型时,可能会面临计算效率低下或模型灵活性不足的问题。为了解决这些问题,广义经验似然方法(Generalized Empirical Likelihood, GEL)应运而生。广义经验似然方法是对传统经验似然方法的扩展,它不仅保留了经验似然的灵活性,还通过引入更一般的权重函数和优化目标,进一步提高了模型的适应性和计算效率。广义经验似然方法的核心思想是通过最大化一个广义的经验似然函数,同时满足一组矩条件,从而实现对模型参数的有效估计。在 1994 年, Qin 和 Lawless [2]将经验似然方法与一般估计方程相结合,提出了广义经验似然方法。Wooldridge [3]提出:使用条件选择概率的估计量,而非真实值,有助于提高逆概率加权方法(Inverse Probability Weighting, IPW)矩估计量的渐进效率。

广义经验似然方法的一个重要应用领域是因果推断,特别是在估计平均处理效应时。基于潜在结果模型的因果推断最早由 Rubin [4]在 1974 年提出,并将其扩展到了非实验数据和观察性研究。Zhao 和 Percival [5]提出,在不设定模型的情况下,熵平衡法相对于线性结果回归和逻辑回归的倾向评分具有双重稳健性,但该方法有许多限制。Zhang 等[6]针对高维时间序列数据,提出了基于平滑矩函数的惩罚 GEL。通过引入 Lasso 惩罚项,实现了变量选择与参数估计的统一。Hill [7]针对自回归模型中存在重尾误差和条件异方差的问题,提出了一种稳健的 GEL 估计方法。Ogata [8]针对金融数据的非正态性情况,将 GEL 扩展至多元稳定分布参数估计。牛翔宇等[9]在响应变量随机缺失时,提出经验似然距离、经验 Cook 距离等诊断统计量。通过随机模拟验证,使用 GEL 方法较传统正态逼近在置信区间覆盖率上提升了 10%~15%。Guggenberger 等[10]提出在弱工具变量场景下, GEL 估计量的渐近分布与检验统计量设计。Jin 等[11]将 GEL 扩展至高阶空间自回归模型,利用鞅结构构建估计方程。Li 和 Qin [12]提出动态面板数据的 GEL 估计方法,解决了固定效应与内生性问题。

在因果推断中, 广义经验似然方法在过度识别的情况下, 能够确保处理效应的估计是稳健的, 并可通过过度识别检验(如 LR 检验)来验证模型设定的合理性。基于逆概率加权的广义经验似然方法作为融合了逆概率加权和广义经验似然理论, 能够在复杂的数据结构和模型设定下, 实现对参数的有效估计和精确推断, 尤其在处理存在缺失数据、非正态分布数据等问题时展现出独特的优势。

本文探讨基于逆概率加权方法的广义经验似然方法在评估职业培训项目效果中的应用, 本文的结构如下: 第二部分介绍逆概率加权方法和广义经验似然方法的基本理论; 第三部分详细描述数据集; 第四部分根据第二部分理论计算分析结果; 第五部分总结全文。

2. 方法介绍

2.1. 广义经验似然方法

Owen (1988) [1]提出经验似然(EL)方法以观测的经验分布为基点, 对于 $k \times 1$ 维的参数向量 θ 以及独立观测样本 X_i ($i = 1, \dots, n$), 经验似然函数为:

$$L(\theta) = \prod_{i=1}^n \pi_i, \quad (2.1)$$

其中, π_i 是样本 X_i 对应的概率, 且满足约束条件 $\sum_{i=1}^n \pi_i \phi(X_i, \theta)$ 和 $\sum_{i=1}^n \pi_i = 1$ 。传统经验似然通过最大化 $L(\theta)$ 得到参数的估计值 $\hat{\theta}$ 。对数经验似然比统计量 $2 \ln [L(\hat{\theta})/L(\theta)]$ 渐近服从自由度为 k 的卡方分布 $\chi^2(k)$ 。

广义经验似然方法(GEL)是经验似然方法的拓展, 不需要经验分布的等权重假设, 通过引入距离函数(如指数倾斜、对数差异)调整权重, 形成更灵活的似然形式。核心思想是在满足矩的条件的前提下, 寻找与经验分布“最接近”的概率分布。对于样本 X_i , 满足 $r \times 1$ 维的矩条件 $E[\phi(X_i, \theta)] = 0$, 且 $r > k$ 。广义经验似然方法通过最大化广义经验似然函数求解,

$$\hat{\theta} = \arg \min_{\theta, \pi_i} \sum_{i=1}^n h_r(\pi_i), \quad (2.2)$$

约束条件为:

$$\sum_{i=1}^n \pi_i = 1, \quad (2.3)$$

$$\sum_{i=1}^n \pi_i \phi(X_i, \theta) = 0, \quad (2.4)$$

其中, $h_r(\pi_i) = (\tau^2 + \tau)^{-1} ((n\pi_i)^{\tau+1} - 1)/n$ 。引入函数 $h_r(\pi_i)$ 的目的是度量概率 π_i 与经验分布权重 $1/n$ 之间的差异, τ 是散度参数。当 $\tau = 0$, 即 $h_r(\pi) = -\ln \pi$ 时, 公式(2.2)退化为传统的经验似然估计量。

Corcoran [13]提出, 对公式(2.2)引入 $r \times 1$ 维的拉格朗日乘子向量 λ , 方程(2.3)和(2.4)的解等同于以下关于 θ 和 λ 的极值问题的解为:

$$\hat{\theta} = \arg \min_{\theta} \sup_{\lambda} \sum_{i=1}^n \rho(\lambda^T \phi(X_i, \theta)), \quad (2.5)$$

其中, $\rho(v) = -(1+v)^{1+1/\nu}/(1+\nu)$ 是一个关于 v 的函数, 在其定义域内是凹函数。当 $\rho(v) = \ln(1+v)$ 时, 退化为经验似然估计量; 当 $\rho(v) = -\exp(v)$, 为指数倾斜估计量(Empirical Transformation, ET); 当 $\rho(v) = -(1+v)^2$ 时, 为连续更新估计量(Continuously Updating Estimator, CUE)。基于满足公式(2.5)的估计量 $\hat{\theta}$ 和 $\hat{\lambda}$, 概率 π_i 的估计值为:

$$\hat{\pi}_i = \frac{\frac{\partial \rho}{\partial v}(\hat{\lambda}^\top \phi(X_i, \hat{\theta}))}{\sum_{j=1}^n \frac{\partial \rho}{\partial v}(\hat{\lambda}^\top \phi(X_j, \hat{\theta}))}, \quad (2.6)$$

重新代入公式(2.2)~(2.4), 最终通过迭代求解 $\hat{\theta}$ 。

在一定的正则条件下, 广义经验似然估计量

$$LR = 2 \sum_{i=1}^n [\rho(\lambda^\top \phi(X_i, \theta)) - \rho(0)], \quad (2.7)$$

渐近服从自由度为 $r-k$ 的卡方分布, 且效率达到 Cramér-Rao 下界。

广义经验似然方法无需假设数据分布, 仅依赖矩条件的存在性, 适用于异质性、非正态数据, 估计的偏差更小。

2.2. 基于逆概率加权的平均处理效应的估计

在观察性研究中, 处理组 and 对照组的个体特征可能存在差异, 一般用平均处理效应(Average Treatment Effect, ATE)度量。

设 Y_i 表示观测样本结果变量, 其对应的处理变量为 T_i , 协变量为 X_i 。当 $T_i=1$ 时, 表示样本接受处理, $T_i=0$ 时, 表示未接受处理。将样本分为处理组和对照组。

平均处理效应定义为:

$$\Delta_{ATE} = E(Y_i | T=1) - E(Y_i | T=0), \quad (2.8)$$

在观察性研究中, 处理组和对照组的个体特征 X_i 可能存在差异, 这些差异会干扰对平均处理效应的准确估计, 即存在混杂因素。

逆概率加权法(IPW)的本质是在观测数据中模拟随机对照试验, 通过“加权”消除个体选择处理(如治疗、受教育)带来的偏差, 从而达到无偏估计处理对结果的平均因果效应。常用的逆概率加权方法, 是通过影响处理的混杂因素的倾向得分计算权重, 调整每个样本的贡献, 使得加权后的分布类似随机试验。

Rosenbaum 和 Rubin [14] 在 1983 年首先提出倾向得分的概念, 即

$$e(X_i) = P(T_i=1|X_i), \quad (2.9)$$

表示在协变量 X_i 下个体接受处理的概率。多维协变量压缩为一维, 但保留所有混杂信息。

倾向得分是个体接受处理的条件概率, 通过它可以平衡处理组和对照组在协变量上的分布。当处理组和对照组在协变量上分布平衡后, 能减少样本选择偏差。估计倾向得分时常用的模型为 logit 模型, 将线性回归结果转换为介于 0~1 之间的概率值, 实现对目标变量的概率预测。logit 模型为:

$$e(X_i) = \frac{\exp(\gamma^\top X_i)}{1 + \exp(\gamma^\top X_i)}, \quad (2.10)$$

其中, γ 为模型 $q \times 1$ 维的参数向量。在实际应用中, 倾向得分往往是未知的, 需要用最大似然估计法来估计模型参数 γ 。模型(2.10)参数的得分向量为:

$$S_j(\gamma) = \sum_{i=1}^n [T_i - e(X_i)] X_{ij}, \quad (2.11)$$

其中, $j=1, 2, \dots, q$ 。

为了估计平均处理效应 Δ_{ATE} , 用倾向得分对样本进行加权调整, 对处理组赋予权重 $1/e(X_i)$, 对照组赋予权重 $1/[1-e(X_i)]$ 。

得到

$$E\left[\frac{T_i Y_i}{e(X_i)}\right] = E[Y_i(1)], E\left[\frac{(1-T_i) Y_i}{1-e(X_i)}\right] = E[Y_i(0)], \quad (2.12)$$

加权后, 根据公式(2.17)可得出基于逆概率加权估计的平均处理效应估计量为:

$$\hat{\Delta}_{ATE} = \frac{1}{n} \left(\sum_{i=1}^n \frac{T_i Y_i}{e(X_i)} - \sum_{i=1}^n \frac{(1-T_i) Y_i}{1-e(X_i)} \right). \quad (2.13)$$

2.3. 基于 GEL-IPW 平均处理效应的估计

不失一般性, 以常用线性回归模型为例, 考虑模型

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_{p-1} X_{p-1,i} + \varepsilon_i, i=1, \cdots, n,$$

其中, β 为 $p \times 1$ 维参数向量, ε_i 为随机误差项, 且满足 $E(\varepsilon_i) = 0$, $Var(\varepsilon_i) = \sigma^2$ 。由最小二乘法得到模型的正规方程, 设定矩条件为:

$$E\left[\psi(Y_i, X_i, \beta)\right] = E \begin{bmatrix} Y_i - \beta_0 - \beta_1 X_{1i} - \beta_2 X_{2i} - \cdots - \beta_{p-1} X_{p-1,i} \\ (Y_i - \beta_0 - \beta_1 X_{1i} - \beta_2 X_{2i} - \cdots - \beta_{p-1} X_{p-1,i}) X_{1i} \\ \vdots \\ (Y_i - \beta_0 - \beta_1 X_{1i} - \beta_2 X_{2i} - \cdots - \beta_{p-1} X_{p-1,i}) X_{p-1,i} \end{bmatrix} = 0, \quad (2.14)$$

对于有处理变量 T_i 的观测样本 Y_i , 将倾向得分作为调整的权重引入公式(2.12)。则在条件独立假设的设定下, 在 IPW 方法中满足的矩条件是:

$$E \left(\frac{\psi(Y_i, X_i, \beta) I(T_i = 1)}{e(X_i)} - \frac{\psi(Y_i, X_i, \beta) I(T_i = 0)}{1-e(X_i)} \right) = 0,$$

倾向得分向量满足的矩条件是:

$$E(S_i) = E[(T_i - e(X_i)) X_{ij}] = 0, \quad (2.15)$$

令 $\theta = (\beta_0, \beta_1, \cdots, \beta_{p-1}, \gamma_0, \gamma_1, \cdots, \gamma_{q-1})$, 将上述的条件联合, 基于约束条件(2.3)和(2.4), 得到 GEL-IPW 方法矩条件方程:

$$E[\phi(X_i, Y_i, \theta)] = E \begin{bmatrix} \frac{\psi(Y_i, X_i, \beta) I(T_i = 1)}{e(X_i)} \\ \frac{\psi(Y_i, X_i, \beta) I(T_i = 0)}{1-e(X_i)} \\ (T_i - e(X_i)) X_{i1} \\ \vdots \\ (T_i - e(X_i)) X_{iq} \end{bmatrix} = 0, \quad (2.16)$$

利用式(2.2)~(2.4), 采用迭代算法来求解上述联合矩估计方程中的参数估计值 $\hat{\theta}$, 检验统计量公式(2.7)服从卡方分布 $\chi^2(p)$ 。

3. 数据模拟

为确定 GEL-IPW 方法的稳健性, 本文进行一项模拟研究。

3.1. 模拟参数设定

设 $X = (X_1, X_2)$ 为协变量特征向量, Y 为结果变量, T 为处理变量, $T = 1$ 时表示接受处理, $T = 0$ 时表示未接受处理。处理组和对照组的结果变量 Y 分别由线性回归模型生成

$$\begin{aligned} Y_1 &= 2 + 3X_1 + 2X_2 + \varepsilon_1, \\ Y_0 &= 1 + 3X_1 + 2X_2 + \varepsilon_0. \end{aligned}$$

其中, 协变量 X_1 和 X_2 互独立, 随机误差项 ε_1 和 ε_0 互相独立, 并都服从标准正态分布。为模拟结果模型错误设置的情况, 设置错误结果估计模型为:

$$\begin{aligned} m(1, X) &= 1 + 3X_1 + 2X_2^2, \\ m(0, X) &= 3X_1^2 + 2X_2. \end{aligned}$$

模拟中设定真实的倾向得分为 logit 回归模型

$$e(X_i; \gamma) = \frac{\exp(\gamma_1 + \gamma_2 X_1 + \gamma_3 X_2)}{1 + \exp(\gamma_1 + \gamma_2 X_1 + \gamma_3 X_2)}.$$

为模拟倾向得分模型被错误设定的情况, 设定倾向得分被错误指定为:

$$e(X_i; \gamma) = \frac{\exp(\gamma_1 + \gamma_2 X_1 + \gamma_3 X_2 + X_1 X_2)}{1 + \exp(\gamma_1 + \gamma_2 X_1 + \gamma_3 X_2 + X_1 X_2)}.$$

为观察处理分配对估计量的影响, 选取参数 $\gamma = (\gamma_1, \gamma_2, \gamma_3)$ 的值分别为 $(\gamma_1, \gamma_2, \gamma_3) = (0.5, 1, 1), (0.5, 2, 2), (0.5, 3, 3)$ 。在这三种情况下, 正确的倾向得分均值分别为 59.98%、56.16%、53.3%, 错误的倾向得分均值为 60.42%、56.86%、54.18%, 设定样本量 n 为 500, 模拟重复次数 2000 次。

3.2. 模拟结果

当倾向得分模型被正确设定时, 各方法的估计结果见表 1。

Table 1. Estimation values when the propensity score model is set correctly
表 1. 倾向得分模型设定正确情况下的估计值

	$\gamma = (0.5, 1, 1)$			$\gamma = (0.5, 2, 2)$			$\gamma = (0.5, 3, 3)$		
	均值	标准差	偏差	均值	标准差	偏差	均值	标准差	偏差
结果模型设定正确									
$\hat{\Delta}_{ATE}$	1.8421	0.6299	0.8421	2.5507	1.6804	1.5507	3.0061	1.7234	2.0061
$\hat{\Delta}_{IPW}$	0.9922	0.4573	-0.0078	0.9965	0.4673	-0.0035	0.9771	0.4681	-0.0229
$\hat{\Delta}_{GEL-IPW}$	0.9719	0.2705	-0.0281	0.9872	0.3705	-0.0128	0.9822	0.5932	-0.0178
结果模型设定错误									
$\hat{\Delta}_{ATE}$	2.0839	0.5897	1.0839	2.9711	1.2463	1.9711	3.5531	1.4894	2.5531
$\hat{\Delta}_{IPW}$	0.9818	0.2851	-0.0182	1.7859	0.3838	0.7859	0.7983	0.6729	-0.2017
$\hat{\Delta}_{GEL-IPW}$	0.9829	0.2578	-0.0171	0.9399	0.5842	-0.0600	0.8890	0.6983	-0.1109

当倾向得分模型被错误设定时, 各方法的估计结果见表 2。

Table 2. Estimation values when the propensity score model is set wrongly
表 2. 倾向得分模型错误情况下的估计值

	$\gamma = (0.5, 1, 1)$			$\gamma = (0.5, 2, 2)$			$\gamma = (0.5, 3, 3)$		
	均值	标准差	偏差	均值	标准差	偏差	均值	标准差	偏差
结果模型设定正确									
$\hat{\Delta}_{ATE}$	2.8421	0.8421	1.8421	3.5507	1.5507	2.5507	4.0061	2.0061	3.0061
$\hat{\Delta}_{IPW}$	1.0248	0.5391	0.0248	0.9643	0.5962	-0.0357	1.0885	0.7480	0.0885
$\hat{\Delta}_{GEL-IPW}$	1.0211	0.3327	0.0211	1.0267	0.4582	0.0267	1.0798	0.4893	0.0798
结果模型设定错误									
$\hat{\Delta}_{ATE}$	2.0005	0.6328	1.0005	2.8988	1.3880	1.8988	3.5657	1.5164	2.5657
$\hat{\Delta}_{IPW}$	0.0632	0.4210	-0.9368	2.3207	1.4694	1.3207	4.5269	1.5334	3.5269
$\hat{\Delta}_{GEL-IPW}$	1.0942	0.5538	0.0942	1.1011	0.5932	0.1011	1.1598	0.7638	0.1598

从模拟中可看出, GEL-IPW 方法在四种情况中都较为稳健。

4. 实例分析

4.1. 数据集描述

本节实例分析了职业培训对收入影响效应的案例。数据名称为 NSW 的数据集, 可以从 R 软件包 `causaldata` 得到。数据集记录的是美国联邦政府资助的一个示范项目, 时间是从 20 世纪 70 年代中期开始, 旨在为那些在加入项目之前面临经济和社会问题的个人提供 12 至 18 个月的职业培训, 评估项目参与者参加培训的效果。数据集包含 722 个观测值以及相关信息, 其信息都是通过初始调查和社会保障局的记录获得, 包括的信息有 1978 年的收入、是否接受职业培训、年龄、受教育年限、是否为黑人等等, 具体研究指标变量见表 3。

Table 3. Definition of variables and meaning
表 3. 变量定义与意义

变量	定义	意义
<i>re78</i>	1978 年的实际收入	结果变量(以千美元为单位)
<i>treat</i>	个体是否接受了职业培训	处理变量(1 = 接受培训, 0 = 未接受培训)
<i>black</i>	是否为黑人	协变量(1 = 是, 0 = 否)
<i>nodeg</i>	是否有学位	协变量(1 = 是, 0 = 否)
<i>re75</i>	1975 年的实际收入	协变量(以千美元为单位)

表 4 对个体相关情况进行统计性描述。

Table 4. Statistical description of variables
表 4. 变量描述性统计

变量	最小值	中位数	最大值	均值	标准差
<i>re78</i>	0	3.9521	60.3070	6.3492	58.9124

续表

<i>treat</i>	0	0	1	0.4157	0.2432
<i>black</i>	0	1	1	0.8427	0.1321
<i>nodeg</i>	0	1	1	0.7079	0.2068
<i>re75</i>	0	0.9363	38.5700	1.5674	12.8901

4.2. 计算结果与分析

为了评估项目实施的效果, 考虑 1978 年的收入(*re78*)与是否接受项目培训(*treat*)以及其他因素的影响, 对于此数据集, 建立如下线性回归模型:

$$re78 = \beta_0 + \beta_1 treat + \beta_2 black + \beta_3 nodeg + \beta_4 re75,$$
 (4.1)

回归方程的估计结果见表 5。

Table 5. Parameter estimation results of the linear regression model
表 5. 线性回归模型参数估计结果

变量	估计值	标准误	t 值	p 值
截距	6.4601	0.7289	8.863	0
<i>treat</i>	0.8008	0.4667	1.716	0.0866
<i>black</i>	-1.3954	0.5742	-2.430	0.0153
<i>nodeg</i>	0.1753	0.0454	3.859	0.0001
<i>re75</i>	-0.9635	0.5562	-1.732	0.0837

基于表 5 分析表明: 有学位(*nodeg*)和种族为黑色人种(*black*)在 0.05 的显著性水平上是显著的, 接受培训(*treat*)和 1975 年收入(*re75*)在 0.1 显著性水平上显著。且有学位(*nodeg*)和接受培训(*treat*)有正向影响, 种族为黑色人种(*black*)和 1975 年收入(*re75*)有负向影响。

由于回归方程(4.1)是显著的, 所以以下矩条件成立是合理的:

$$E[\psi(\cdot)] = E \begin{pmatrix} re78 - \beta_0 - \beta_1 treat - \beta_2 black - \beta_3 no deg - \beta_4 re75 \\ (re78 - \beta_0 - \beta_1 treat - \beta_2 black - \beta_3 no deg - \beta_4 re75) treat \\ (re78 - \beta_0 - \beta_1 treat - \beta_2 black - \beta_3 no deg - \beta_4 re75) black \\ (re78 - \beta_0 - \beta_1 treat - \beta_2 black - \beta_3 no deg - \beta_4 re75) no deg \\ (re78 - \beta_0 - \beta_1 treat - \beta_2 black - \beta_3 no deg - \beta_4 re75) re75 \end{pmatrix} = 0,$$
 (4.2)

考虑到项目开始时参与者是否选择参与培训(*treat*)可能也受到如当时的收入(*re75*)等因素的影响, 接下来检验变量 *black*、*nodeg*、*re75* 对 *treat* 变量的影响, 建立 logit 模型:

$$e(\cdot) = \frac{\exp(\gamma_0 + \gamma_1 black + \gamma_2 nodeg + \gamma_3 re75)}{1 + \exp(\gamma_0 + \gamma_1 black + \gamma_2 nodeg + \gamma_3 re75)},$$
 (4.3)

模型参数估计结果见表 6。

Table 6. Parameter estimation results of the Logit model
表 6. Logit 模型参数估计结果

变量	估计值	标准误	z 值	p 值
截距	-0.0012	0.8599	0.505	0.6135

续表

<i>black</i>	0.0253	0.2635	-0.370	0.7116
<i>nodeg</i>	-0.4815	0.1814	-2.654	0.0079
<i>re75</i>	-0.0015	0.0151	-0.101	0.9198

由表 6 分析表明: 有学位(*nodeg*)在 0.01 水平上显著, 且与接受培训(*treat*)负相关, 说明 logit 回归模型成立, 倾向得分是合理的, 因此其对应矩条件为:

$$E[S(\cdot)] = E \begin{pmatrix} [treat - e(\cdot)]black \\ [treat - e(\cdot)]nodeg \\ [treat - e(\cdot)]re75 \end{pmatrix} = 0, \quad (4.4)$$

结合逆概率加权估计平均处理效应的定义, IPW-GEL 的矩条件为:

$$E[\phi(\cdot)] = E \begin{pmatrix} \frac{\psi(\cdot)I(treat=1)}{e(\cdot)} \\ \frac{\psi(\cdot)I(treat=0)}{1-e(\cdot)} \\ S(\cdot) \end{pmatrix} = 0, \quad (4.5)$$

将公式(4.5)中的 $\phi(\cdot)$ 代入式(2.1)~(2.4)中, 得到经验似然估计量, 检验统计量渐近服从 $\chi^2(5)$ 分布。

最后根据公式(2.13)利用 GEL-IPW 方法计算平均处理效应, 估计值结果见表 7。根据广义经验似然方法计算的结果分别在第 2 行、第 3 行, 为了和传统逆概率加权方法对比, 利用传统逆概率加权方法估计的平均处理效应结果列在第 1 行。

Table 7. Estimated value of the average treatment effect
表 7. 平均处理效应估计值

方法	估计值	标准误	t 值	p 值
$\hat{\Delta}_{IPW}$	0.7982	0.4803	1.6618	0.0969
$\hat{\Delta}_{GEL-ET}$	0.7902	0.3996	1.9776	0.0734
$\hat{\Delta}_{GEL-CUE}$	0.7932	0.4088	1.9402	0.0768
$\hat{\Delta}_{GEL-IPW}$	0.7923	0.3951	2.0051	0.0709

表 5 分析表明: 在接受职业培训的情况下, 1978 年个体的平均收入比未接受培训的个体平均高出约 0.8 千美元。

注意到由于上述分析 1978 年收入(*re78*)不服从正态分布, 使用线性回归模型可能会产生偏差, 广义经验似然方法属于非参数方法, 对分布没有假设要求。

5. 结语

本文研究了基于逆概率加权的广义经验似然方法在平均处理效应估计中的应用。通过实证研究, 验证了该方法在解决观察性研究中对于处理不符合正态分布数据的平均处理效应的有效性。GEL 方法通过优化调整权重, 避免了传统 IPW 方法中可能出现的数据偏态问题, 将二者结合的 GEL-IPW 方法估计的

结果较为精准稳健。

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] Owen, A.B. (1988) Empirical Likelihood Ratio Confidence Intervals for a Single Functional. *Biometrika*, **75**, 237-249. <https://doi.org/10.1093/biomet/75.2.237>
- [2] Qin, J. and Lawless, J. (1994) Empirical Likelihood and General Estimating Equations. *The Annals of Statistics*, **22**, 300-325. <https://doi.org/10.1214/aos/1176325370>
- [3] Wooldridge, J.M. (2002) Inverse Probability Weighted M-Estimators for Sample Selection, Attrition, and Stratification. *Portuguese Economic Journal*, **1**, 117-139. <https://doi.org/10.1007/s10258-002-0008-x>
- [4] 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>
- [5] Zhao, Q. and Percival, D. (2016) Entropy Balancing Is Doubly Robust. *Journal of Causal Inference*, **5**, 1-19. <https://doi.org/10.1515/jci-2016-0010>
- [6] Zhang, J., Shi, H., Tian, L. and Xiao, F. (2019) Penalized Generalized Empirical Likelihood in High-Dimensional Weakly Dependent Data. *Journal of Multivariate Analysis*, **171**, 270-283. <https://doi.org/10.1016/j.jmva.2018.12.010>
- [7] Hill, J.B. (2015) Robust Generalized Empirical Likelihood for Heavy Tailed Autoregressions with Conditionally Heteroscedastic Errors. *Journal of Multivariate Analysis*, **135**, 131-152. <https://doi.org/10.1016/j.jmva.2014.12.008>
- [8] Ogata, H. (2013) Estimation for Multivariate Stable Distributions with Generalized Empirical Likelihood. *Journal of Econometrics*, **172**, 248-254. <https://doi.org/10.1016/j.jeconom.2012.08.017>
- [9] 牛翔宇, 冯予. 缺失数据下广义非线性回归的经验似然及诊断[J]. 重庆工商大学学报, 2016, 33(6): 15-21.
- [10] Guggenberger, P. and Smith, R.J. (2005) Generalized Empirical Likelihood Estimators and Tests Under Partial, Weak, and Strong Identification. *Econometric Theory*, **21**, 667-709.
- [11] Jin, F. and Lee, L. (2019) GEL Estimation and Tests of Spatial Autoregressive Models. *Journal of Econometrics*, **208**, 585-612. <https://doi.org/10.1016/j.jeconom.2018.07.007>
- [12] Li, Y. and Qin, Y. (2022) Empirical Likelihood for Spatial Dynamic Panel Data Models. *Journal of the Korean Statistical Society*, **51**, 500-525. <https://doi.org/10.1007/s42952-021-00150-4>
- [13] Corcoran, S. (1998) Bartlett Adjustment of Empirical Discrepancy Statistics. *Biometrika*, **85**, 967-972. <https://doi.org/10.1093/biomet/85.4.967>
- [14] 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>