

因果推断下基于加速失效时间模型的网络营销策略研究

杨 慧

贵州大学数学与统计学院, 贵州 贵阳

收稿日期: 2025年9月4日; 录用日期: 2025年9月18日; 发布日期: 2025年10月11日

摘 要

本文构建了基于倾向性得分的加速失效时间模型, 探究电商平台促销活动与用户复购行为之间的因果关系。研究采用倾向性得分处理观测数据中的选择偏倚问题, 并利用Logistic回归模型对其进行估计。随后, 将倾向性得分作为协变量纳入加速失效时间模型, 对促销活动与用户复购时间间隔的关系进行建模, 并估计因果效应, 同时考虑了右删失数据的处理。实证结果显示, 在对一系列关键混杂因素进行调整后, 是否使用优惠券的促销行为, 对用户复购时间间隔产生统计学上的显著影响, 证明了模型的有效性。一定程度上, 促销活动可以加速用户复购速度, 帮助电商平台调整网络营销策略。

关键词

倾向性得分, 加速失效时间模型, 网络营销, 复购时间间隔

Research on Internet Marketing Strategy Based on Accelerated Failure Time Model under Causal Inference

Hui Yang

School of Mathematics and Statistics, Guizhou University, Guiyang Guizhou

Received: September 4, 2025; accepted: September 18, 2025; published: October 11, 2025

Abstract

This article constructs an accelerated failure time model based on propensity scores to explore the causal relationship between promotional activities on e-commerce platforms and user repurchase behavior. The study uses propensity scores to address the selection bias problem in observational data and estimates it using a logistic regression model. Subsequently, the propensity score is incorporated

into the accelerated failure time model as a covariate to model the relationship between promotional activities and the time interval between user repurchases, and the causal effect is estimated, while also considering the treatment of right-censored data. The empirical results show that after adjusting for a series of key confounding factors, the promotional behavior of using coupons has a statistically significant impact on the time interval between user repurchases, which proves the validity of the model. To a certain extent, promotional activities can accelerate the speed of user repurchases and help e-commerce platforms adjust their online marketing strategies.

Keywords

Propensity Score, Accelerated Failure Time Model, Online Marketing, Repurchase Interval

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

1.1. 研究背景

随着互联网和大数据等技术的发展,以及智能终端设备的普及,越来越多的消费者、商家和企业进入电子商务领域[1]。近来,该领域已成为全球零售业的核心驱动力,用户的很多行为数据都被记录下来。各大电商平台通过数据分析和用户反馈,能够实时调整营销策略,以适应市场变化和消费者需求,从而吸引更多用户进行购买。在大数据网络时代,电商平台投入大量资源于形式多样的网络营销活动[2]-[4],如发放优惠券、限时折扣、包邮等,以期刺激用户再次消费、缩短用户的购买决策周期。

日益激烈的市场竞争迫使电商平台从早期的流量扩张模式,转向以用户为中心的数据驱动精细化运营模式。在众多运营指标中,用户复购行为[5]不仅是衡量平台用户忠诚度和长期价值的关键,也是平台实现可持续增长、在竞争中脱颖而出的核心。然而,用户的复购行为是一个复杂的决策过程,它受到用户个人特征、历史消费习惯、商品属性及购物体验(如物流、评价)等多重因素的共同影响。因此,准确地从这些混杂因素中剥离出营销活动的真实因果效应,评估其投资回报率(ROI),是当前电商平台在制定营销策略时面临的关键挑战与核心议题[6]。

1.2. 研究目的

通过分析用户复购时间间隔与促销活动(如发放优惠券)之间的因果关系[7],希望能够帮助电商平台的管理者找到有购买意愿的用户,向他们发送促销活动的信息使得营销活动收益最大化。因此,促进用户复购是电商平台实现可持续增长的核心议题。在影响用户复购决策的众多因素中,促销活动(如折扣、优惠券等)被普遍认为是刺激短期消费、提升用户活跃度的重要手段。然而,用户复购决策是一个极为复杂的过程,受到诸多因素(协变量)的共同影响,如:历史购物体验(如物流速度、商品满意度)、个人消费习惯(如支付方式、消费金额)以及商品自身属性。简单地将复购行为的改变归因于促销,往往会忽略其他混杂因素的作用,从而导致对营销策略的误判。

因此,本研究目的:通过构建因果推断与生存分析相结合的模型[8][9],揭示促销活动、多维度用户交易特征与复购时间间隔三者之间的深层关系,识别影响用户复购决策的关键驱动因素。通过对促销活动的净效应进行量化分析,研究旨在为电商平台提供可靠的数据支持,帮助其制定更有效的市场策略,

以提升用户满意度和忠诚度，最终实现可持续发展。期望能为电商行业的理论研究和实践应用提供有价值的参考，推动在线购物体验的不断改善。

本文使用基于倾向性得分的加速失效时间模型[10][11]对电商平台购物数据进行建模，模型中的倾向性得分采用 Logistic 回归模型进行估计，并对结果变量模型中的所有未知参数，尤其是因果效应参数利用秩的方法进行估计。最后，绘制出两种处理状态下的 Kaplan-Meier 曲线，探讨网络营销策略，从而验证模型的有效性。

2. 模型

2.1. 因果推断

潜在结果模型是因果推断的重要模型之一，其主要思想是比较同一个体在接受处理和不接受处理时的结果差异。潜在结果模型的提出为之后的理论发展奠定了重要基础，该模型是由统计学家 Rubin (1974) [12]提出，因此该模型又被称之为鲁宾因果模型(Rubin Causal Model)。Rubin 将随机化的思想推广到观测性数据的研究，潜在结果框架可以用来处理随机化实验和非随机化实验，并给出了因果效应的严格定义。

Rubin 认为在因果推断中，若没有干预(也叫处理，Treatment [13])就没有因果，这里的干预可以是商家的某项促销活动或者是一种新的治疗方案，本文主要考虑二值干预的因果效应。在干预状态实现前，有几种干预状态就有几种潜在结果，而干预状态实现之后，只有一个潜在结果可以被观测到，在实际应用中那些无法观测到的结果被称为反事实结果(Counterfactual Outcome)。许多因素可能会导致因果效应估计产生偏差，为了更准确地表达因果效应，需要引入潜在结果框架中常见的假设：

假设 1 稳定个体干预值假设：个体 i 的潜在结果不受其他个体干预的影响，并且对于每个个体和每一种干预只有一个潜在结果。

假设 2 强可忽略性假设：令 X 表示观测的协变量，如果满足：

$$(Y(1), Y(0)) \perp D | X, \quad 0 < P(D=1 | X) < 1$$

则称分配机制是可忽略的。其中， D 为一个二元处理变量，表示个体处理分配情况，即分配到控制组取值为 0，分配到处理组取值为 1； $Y(1)$ ， $Y(0)$ 表示个体分配到控制组或处理组时的潜在结果； $Y = DY(1) + (1-D)Y(0)$ 表示个体观察到的结果变量。

基于以上假设，下面给出总体平均因果效应(Average Causal Effect, ACE)，也叫平均处理效应(Average Treatment Effect, ATE)。对于一个感兴趣的总体，平均因果效应为：

$$\tau_{ATE} = E(Y(1) - Y(0)) = E(Y(1)) - E(Y(0)) \quad (1)$$

2.2. 加速失效时间模型

加速失效时间模型(Accelerated Failure Time Model, AFT)是一种常用于生存分析的统计模型，主要用于研究生存时间和其他因素之间的关系。AFT 模型是对数失效时间对协变量做回归，可以作为 Cox 比例风险模型的替代。此外，该模型可以直接根据两种不同处理方式的潜在失效时间来定义因果效应。AFT 模型的这些特点，使它们在生物医疗、经济以及在比较不同处理方式效果的背景下具有非常大的吸引力。

在实际应用中，AFT 模型更加稳健且易于解释。令 X_i 表示给定 d 维的协变量，则 AFT 模型可以表示为：

$$\log T_i = \beta^T X_i + \varepsilon_i \quad (2)$$

其中， $i=1, 2, \dots, n$ ； T_i 为个体失效时间； $\beta = (\beta_1, \dots, \beta_d)^T$ 为 d 维的参数向量； ε_i 为独立同分布的残差项。

2.3. 倾向性得分模型

研究人员常用观测性研究来估计平均因果效应。在此类研究中,由于缺乏随机处理分配,接受与未接受处理的个体存在差异,故需用统计方法消除或最小化两组观测或测量的基线协变量分布差异所造成的混杂影响。在1983年,Rosenbaum和Rubin [14]提出倾向性得分,用于从观测数据中估计因果效应。通过一维的倾向性得分可以消除或最小化观测协变量导致的偏差,以下是倾向性得分的定义:

倾向性得分(P propensity Score, PS)表示为给定协变量 X_i 的条件下,接受干预或处理 ($D_i = 1$) 的条件概率,即

$$\pi(X_i) = E[D_i | X_i] = P(D_i = 1 | X_i) \quad (3)$$

其中, $i = 1, 2, \dots, n$; $\pi(X_i)$ 为第 i 个个体的倾向性得分,且是最粗糙的平衡得分。

因此,可以用倾向性得分代替协变量,通过倾向性得分匹配法、加权法、分层法和回归调整法或它们的结合等估计因果效应,避免了高维协变量情况下的维数灾难问题。尽管这些方法很受欢迎且理论上也有吸引力,但实际应用的主要难点是需估计倾向性得分,而研究发现,倾向性得分模型的微小误差可能导致因果效应估计出现重大偏差。

在随机化实验中,研究者可以通过实验设计控制不同个体接受干预的概率;而观测性研究中,倾向性得分是未知的,需依据观测样本提供的信息估计处理分配概率。最常见的方法是用广义线性模型(如 Logistic 回归模型或 Probit 模型)估计倾向性得分,其中带参数的 Logistic 回归模型为:

$$\pi(X) = P(D = 1 | X) = F(X^T \alpha) = \frac{\exp(X^T \alpha)}{1 + \exp(X^T \alpha)} \quad (4)$$

Probit 模型为:

$$\pi(X) = P(D = 1 | X) = \Phi(X^T \alpha) = \int_{-\infty}^{X^T \alpha} \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}} du \quad (5)$$

因果推断在医疗、流行病、经济学等领域应用广泛,其研究对象多为事件时间数据,由于随访时间的限制有时这些数据是不完整观测。例如,在研究用户复购的问题上,只有感兴趣的事件(复购发生)发生在观测期内,用户复购时间间隔(生存时间)才能完全观察到,否则研究者无法观察到真实的复购时间间隔,只能观察到观测期内用户的购物情况。因此,需要考虑右删失情况,设 T_i 表示第 i 个个体的失效时间,用 C_i 表示第 i 个个体的删失时间,则相应观察时间 $U_i = T_i \wedge C_i = \min(T_i, C_i)$ 。用 δ_i 表示第 i 个个体的删失指标,当 $T_i \leq C_i$ 时, δ_i 值为 1; 当 $T_i > C_i$ 时, δ_i 值为 0, 即 $\delta_i = I(T_i \leq C_i)$, $I(\cdot)$ 表示示性函数。因此,观测结果为 $Y = (1 - D)Y(0) + DY(1)$, $\delta = (1 - D)\delta(0) + D\delta(1)$ 。其中,潜在结果为 $Y(D) = \min(T(D), C(D))$, 潜在删失指标为 $\delta(D) = I(T(D) \leq C(D))$ 。

在潜在结果框架下,假设我们观测到的数据为 $\{(Y_i, \delta_i, D_i, X_i), i = 1, 2, \dots, n\}$, 且失效时间 T 和删失时间 C 独立。在满足假设 1 和假设 2 情况下,本文采用的倾向性得分模型为基于倾向性得分的协变量调整法,即考虑将倾向性得分作为一个协变量纳入加速失效时间模型,通过倾向性得分消除混杂因素从而分析数据的因果效应。因此,带倾向性得分的加速失效时间模型可以表示成:

$$\log T = \beta D + \eta \pi(X, \alpha) + \varepsilon \quad (6)$$

其中, β, η, α 均为未知回归参数; ε 为独立同分布的随机误差项。此模型中协变量通过倾向性得分影响失效时间,处理变量 D 的回归系数 β 即为因果效应的参数。

3. 估计

3.1. 倾向性得分估计

倾向性得分是在给定协变量 X 的情况下, 服从某一特定处理的条件概率。本文考虑用 Logistic 回归模型估计倾向性得分。公式如下:

$$\pi(X) = P(D=1|X, \alpha_0) = \frac{\exp(X^T \alpha_0)}{1 + \exp(X^T \alpha_0)} \quad (7)$$

令 $\pi(x, \alpha_0)$ 表示真实的倾向性得分, 其中 $\pi(x, \alpha) = P(D=1|X=x, \alpha)$ 。则令 $\hat{\alpha}$ 为 α_0 的最大似然估计, 即如下对数似然函数的估计:

$$\ln f(X_1, \dots, X_n, \alpha) = \sum_{i=1}^n D_i (X_i^T \alpha) - \sum_{i=1}^n \ln(1 + \exp(X_i^T \alpha))$$

或如下得分方程的解:

$$U_p(\alpha) = \sum_{i=1}^n X_i \left[D_i - \frac{\exp(X_i^T \alpha)}{1 + \exp(X_i^T \alpha)} \right] = 0$$

因此, 倾向性得分的估计为 $\pi(X, \hat{\alpha})$ 。

3.2. 因果效应估计

由(7)可得估计的倾向性得分, 则结果变量模型(6)进一步可得:

$$\log T = \beta D + \eta \pi(X, \hat{\alpha}) + \varepsilon \quad (8)$$

其中, β 表示因果效应的参数; η 为冗余参数; ε 为服从某一未知分布的随机误差项。

当满足强可忽略性假设以及结果变量和倾向性得分线性假设时, 有:

$$\begin{aligned} & E[\log(T)|D=1, \pi(X)] - E[\log(T)|D=0, \pi(X)] \\ &= \beta + \eta \pi(X) + E(\varepsilon|\pi(X)) - [\eta \pi(X) + E(\varepsilon|\pi(X))] \\ &= \beta \end{aligned}$$

故此时因果效应参数的估计是无偏的。

为简化推导过程中的表述, 现对相关参数进行重新定义: $\theta = (\beta, \eta)^T$ 和 $\theta_0 = (\beta_0, \eta_0)^T$; $r_i(\theta, \alpha) = \log T_i - \beta D_i - \eta \pi(X_i, \alpha)$, $K_i = (D_i, \pi(X_i, \alpha))^T$ 。定义计数过程 $N_i(t; \theta, \alpha) = I(r_i(\theta, \alpha) \leq t, \delta_i = 1)$ 和风险过程 $Y_i(t; \theta, \alpha) = I(r_i(\theta, \alpha) \geq t)$ 。估计加速失效时间模型的回归参数时, 常用秩统计量方法:

$$U(\theta, \hat{\alpha}) = \sum_{i=1}^n \int_{-\infty}^{\infty} \psi(t; \theta, \hat{\alpha}) [K_i(\hat{\alpha}) - \bar{K}(t; \theta, \hat{\alpha})] dN_i(t; \theta, \hat{\alpha}) = 0 \quad (9)$$

其中, $\psi(t; \theta, \alpha)$ 是一个权重函数; $\bar{K}(t; \theta, \alpha) = M^{(1)}(t; \theta, \alpha) / M^{(0)}(t; \theta, \alpha)$, 且

$M^{(c)}(t; \theta, \alpha) = n^{-1} \sum_{i=1}^n Y_i(t; \theta, \alpha) K_j^{(c)}(\alpha)$, $c=0, 1$ 。此时 $\psi(t; \theta, \alpha)$ 可以设为 $M^{(0)}(t; \theta, \alpha)$, 则对应的 Gehan 统计量可以表示为:

$$U_G(\theta, \hat{\alpha}) = n^{-1} \sum_{i=1}^n \sum_{j=1}^n \delta_i (K_i(\hat{\alpha}) - K_j(\hat{\alpha})) I(r_i(\theta, \hat{\alpha}) \leq r_j(\theta, \hat{\alpha})) = 0$$

其中, 令 $\tilde{\theta} = (\tilde{\beta}, \tilde{\eta})^T$ 表示该得分方程解的估计量。

由于 Gehan 得分方程的解存在不连续，可通过光滑化方法对该得分方程进行近似。此处采用分布函数来近似示性函数，因此，经光滑化的得分方程可表示为：

$$\tilde{U}_G(\theta, \hat{\alpha}) = n^{-1} \sum_{i=1}^n \sum_{j=1}^n \delta_i (K_i(\hat{\alpha}) - K_j(\hat{\alpha})) \Phi \left(\frac{r_j(\theta, \hat{\alpha}) - r_i(\theta, \hat{\alpha})}{k_{ij}(\hat{\alpha})} \right) = 0$$

其中， $k_{ij}(\alpha) = \sqrt{(K_i(\alpha) - K_j(\alpha))^T (K_i(\alpha) - K_j(\alpha)) / n}$ 。令 $\hat{\theta} = (\hat{\beta}, \hat{\eta})^T$ 表示新得分方程解的估计量。因此，此处 $\hat{\beta}$ 即为因果效应参数的估计量。

4. 实证分析

本文研究数据来源于 Kaggle Datasets 公布的巴西电商 Olist 的购物数据，该数据集包含由 96,096 名独立用户生成的 99,441 条订单记录。为了验证所提出方法的有效性，本节将基于所提出的倾向性得分协变量调整法，研究发放优惠券(促销活动)与用户复购时间间隔的因果关系。原始数据集包含主要变量的信息见表 1。

Table 1. Dataset variables
表 1. 数据集变量

	变量名	变量含义
结果变量	Repurchase Interval	用户复购时间间隔(生存时间)，天
处理变量	Promotion	促销活动，0 表示非促销活动、1 表示促销活动
	Cens	0 表示发生删失、1 表示未发生删失
删失示性变量	Gender	用户性别，分类为男性或女性
	Age	用户年龄，年
	Customer City (CC)	用户所在城市
	Product Category (PC)	商品品类，如家居用品、文具等
	Payment Installments (PI)	分期付款期数
	Payment Value (PV)	消费金额
协变量	Freight Value (FV)	商品运费
	Review Score (RS)	用户购物满意度
	Delivery Days (DD)	物流天数
	Order Item Count (OIC)	订单商品总数

对数据集进行调整前，首先计算倾向性得分，图 1 为处理组和控制组数据的倾向性得分分布情况。其中，蓝色区域表示个体接受处理的倾向性得分分布，红色区域表示未接受处理的倾向性得分分布。这直观地表明，在原始状态下，处理组和控制组的混杂变量存在系统性差异，若直接比较这两组的结果，将会得出偏倚的结论。尽管两组有差异，但在倾向性得分的大部分区间内，重合区域较多，满足共同支撑域条件，便于后续研究。

接下来，验证倾向性得分对协变量的平衡效果，计算处理组与控制组在调整前后各个协变量的标准化均值差(SMD)。如图 2 所示，大部分协变量在经过倾向性得分调整后的 SMD 绝对值均小于 0.1，表明

倾向性得分有效地平衡了两组间的基线差异，满足了后续进行因果效应估计的前提条件。

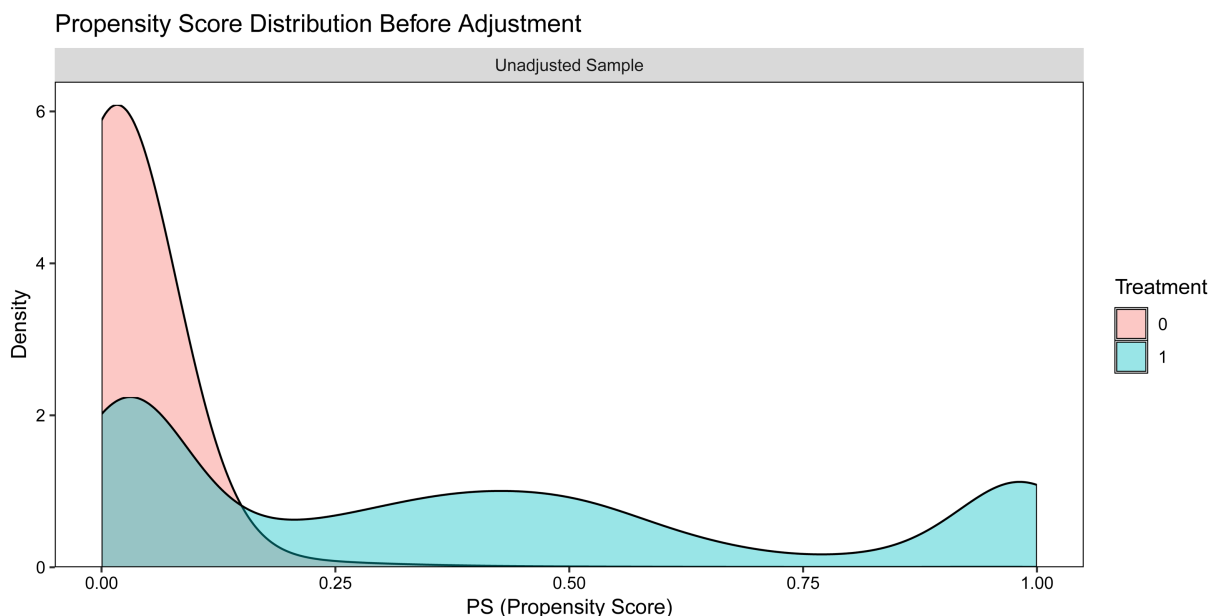


Figure 1. Propensity score distribution plot for the two treatment states

图 1. 两种处理状态的倾向性得分分布图

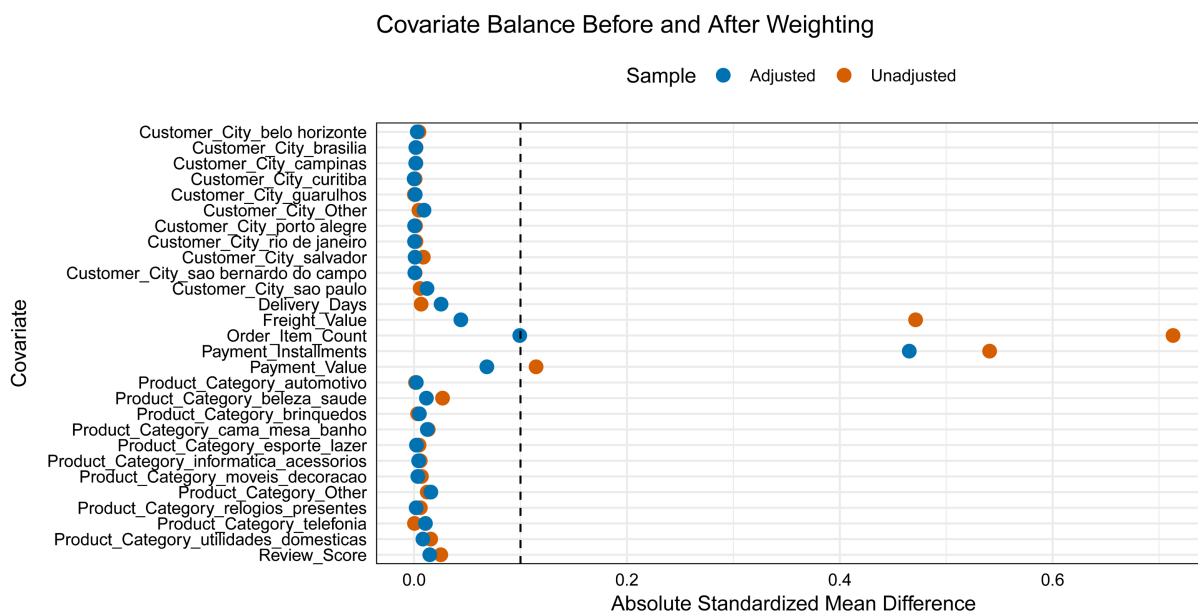


Figure 2. Partial covariate balance test

图 2. 部分协变量平衡性检验

为了估计因果效应，对此数据集搭建如下的处理分配模型：

$$PS = (Promotion = 1 | X) = \frac{\exp(\alpha_0 + \alpha_1 Gender + \alpha_2 Age + \alpha_3 CC + \alpha_4 PC + \alpha_5 PI + \alpha_6 PV + \alpha_7 FV + \alpha_8 RS + \alpha_9 DD + \alpha_{10} OIC)}{1 + \exp(\alpha_0 + \alpha_1 Gender + \alpha_2 Age + \alpha_3 CC + \alpha_4 PC + \alpha_5 PI + \alpha_6 PV + \alpha_7 FV + \alpha_8 RS + \alpha_9 DD + \alpha_{10} OIC)}$$

则对应的结果变量模型为：

$$\log(\text{Repurchase Interval}) = \beta \text{Promotion} + \eta \widehat{PS} + \varepsilon \tag{10}$$

其中， β 为感兴趣的因果效应参数； η 为未知回归参数； ε 为误差项，假定其服从 Log-logistic 分布。

首先，验证所选误差项假设 Log-logistic 分布的 AFT 模型(结果模型)拟合优度，此处将模型拟合的生存曲线与原始数据的 Kaplan-Meier 非参数估计曲线进行了比较，如图 3 所示。从图中可以看出，两条曲线在整个观测期间都高度重合，这表明 Log-logistic 分布能很好地拟合本研究的生存时间数据，模型设定合理可靠。

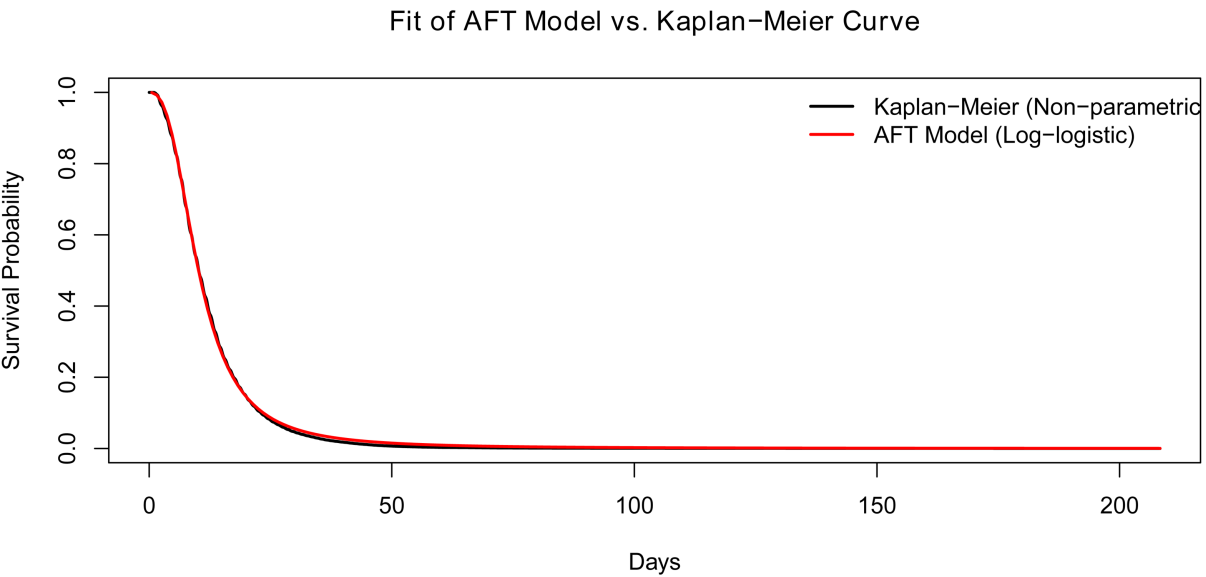


Figure 3. Comparison of AFT model fitting survival curve and Kaplan-Meier curve

图 3. AFT 模型拟合生存曲线与 Kaplan-Meier 曲线对比图

Table 2. Parameter estimation results

表 2. 参数估计结果

参数	估计值	标准误
β	-0.658	0.186
η	-0.917	0.201

其次，将数据依次代入处理分配模型与结果变量模型，通过 Logistic 回归模型估计出倾向性得分 PS 和秩的方法估计出因果效应参数 β ，估计结果见表 2。图 4 展示了两种处理状态下的 Kaplan-Meier 曲线。

图 4 绘制的是在不同处理状态下的复购生存概率(未复购概率)的 Kaplan-Meier 曲线，从中可以发现随着时间的推移，两组人员未复购的概率逐渐减小。其中，参与促销活动的处理组概率比未参与促销活动的控制组下降速度更快，说明收到促销活动信息的用户，其整体复购速度比未接受促销的用户要快一些，即复购时间间隔较短。为了进一步从统计学上验证该差异的显著性，进行了 Log-rank 检验，检验结果 P 值 = 0.061，在 $\alpha = 0.1$ 标准下是统计显著的，但在 $\alpha = 0.05$ 标准下是不显著的。在不考虑其他混杂因素时，参与促销活动的处理组与未参与的控制组在复购时间间隔的分布上存在一定的差异，但不明显。然而，通过倾向性得分控制了一系列关键混杂因素之后，促销活动对复购时间的真实效果才得以显现出来。

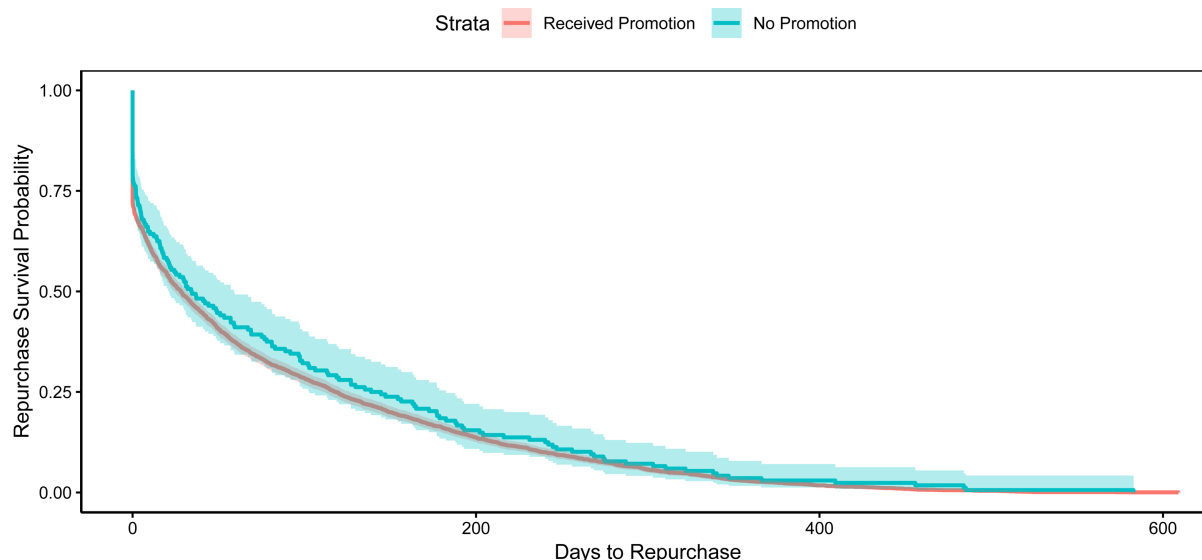


Figure 4. Kaplan-Meier curves for the two treatment states

图 4. 两种处理状态的 Kaplan-Meier 曲线

从表 2 可以看出, 因果效应参数 β 的估计量为 -0.658 , 其标准误 0.186 , P 值 $= 0.048 < 0.05$, 表明结果显著。其加速因子为 $\exp(-0.658) \approx 0.518$, 该结果的含义: 在控制了其他混杂因素后, 参与促销活动用户的预期复购时间间隔是未参与促销活动用户的 0.518 倍, 即促销活动能将用户的复购速度平均提升约 48.2% , 显著缩短了用户的复购周期。这表明用户复购时间间隔与促销活动(如发放优惠券)之间存在一定的因果关系, 且促销活动会“加速”复购(缩短复购时间间隔), 即收到促销活动信息的用户其未复购生存时间比未收到促销活动信息的用户短。倾向性得分估计系数 η 为 -0.917 , 这表明一个用户的倾向性得分越高, 他的复购时间间隔就越短。因此, 在一定程度上, 促销活动可以加速用户复购, 提高店铺的网络营销量, 与实际情况较相符。

5. 结束语

本文旨在探究电商平台中促销活动对用户复购时间间隔的因果效应, 为此构建了一个基于倾向性得分的加速失效时间模型, 并利用 Logistic 回归模型及秩的方法进行了统计推断。具体而言, 为解决观测数据中存在的混杂偏倚问题, 首先通过 Logistic 回归模型, 利用用户地理位置、历史交易特征、商品偏好及购物满意度等 8 个维度的协变量, 为每个购买行为估计了其接受促销的倾向性得分。随后, 将倾向性得分作为一个协变量, 纳入加速失效时间生存模型中, 对促销活动的处理效应进行了估计, 并考虑了右删失数据以保证结果的稳健性。实证结果显示用户复购时间间隔与促销活动(如发放优惠券)之间存在一定的因果关系, 这也与网络营销的实际情况相符合, 说明了模型的可行性和有效性。

但是在实际情况中, 用户的复购决策是一个极为复杂的过程, 不仅受到文中所述协变量的影响, 还受到诸多因素的共同影响, 如: 优惠券的面额(折扣力度)、品牌信誉、支付方式等。如果忽略这些因素, 可能导致研究结论受到潜在的内生性问题的挑战。倾向性得分法主要处理的是基于可观测变量的选择偏误, 而对于由不可观测的混杂因素所导致的内生性则无能为力, 如: 用户的内在购物意愿与价格敏感度、平台忠诚度、外部营销活动的干扰等不可观测混杂因素, 这些不可观测因素会导致本研究对促销活动的因果效应产生高估。若倾向性得分模型未能完全控制所有重要的混杂因素, 则会导致结果变量模型的准确性下降。

因此,针对本文存在的潜在内生性问题等局限性,未来的研究可以从以下几个方面展开,以获得更稳健的因果效应估计,如:寻找工具变量、应用断点回归设计、构建面板数据并使用双重差分模型等方法。针对混杂因素的选取,未来的研究可以从以下几个重要方向进行深化:一是尝试获取包含更丰富用户画像的数据,以构建更精确的倾向性得分模型;二是对促销活动进行更细致的划分,并探索异质性处理效应,即分析不同力度的促销对不同用户群体(如新/老用户、高/低消费用户)是否存在差异化影响;三是可以将研究结果扩展到其他商业指标,如客单价(AOV)或用户生命周期价值(LTV),以更全面地评估促销活动的商业价值。

参考文献

- [1] 黎春艳. 电子商务环境下的网络营销策略[J]. 中国商界, 2025(3): 154-155.
- [2] 耿冠宇. 网络经济背景下企业的市场营销举措分析[J]. 商展经济, 2023(19): 105-108.
- [3] 肖晖. 大数据时代电子商务网络营销探讨[J]. 现代商贸工业, 2015, 36(14): 50-51.
- [4] Mohammed, R., Fisher, R.J., Jaworski, B.J., *et al.* (2004) Internet Marketing. Auflage.
- [5] 石静. 基于 Stacking 与 Cox 比例风险模型对用户购买行为的预测及影响因素分析[D]: [硕士学位论文]. 上海: 上海师范大学, 2025.
- [6] 王艺瑾. 拼多多电商平台营销策略优化研究[D]: [硕士学位论文]. 兰州: 兰州理工大学, 2024.
- [7] Waisman, C., Nair, H.S. and Carrion, C. (2025) Online Causal Inference for Advertising in Real-Time Bidding Auctions. *Marketing Science*, **44**, 176-195. <https://doi.org/10.1287/mksc.2022.0406>
- [8] Rossi, P.E. (2018) Causal Inference in Marketing Applications. In: *Handbook of Marketing Analytics: Methods and Applications in Marketing Management, Public Policy, and Litigation Support*, Edward Elgar Publishing, 135-152. <https://doi.org/10.4337/9781784716752.00015>
- [9] Majeed, A.F. (2020) Accelerated Failure Time Models: An Application in Insurance Attrition. *The Journal of Risk Management and Insurance*, **24**, 12-35.
- [10] 周玲米, 汤杰, 莫靖林, 等. 桂林市 1996-2019 年 HIV/AIDS 患者生存时间影响因素分析[J]. 中国皮肤性病杂志, 2024, 38(6): 663-670.
- [11] 王淑影, 张亚男, 程云飞. 基于模型平均的右删失数据下加速失效时间模型的统计推断研究[J]. 统计与信息论坛, 2022, 37(12): 3-9.
- [12] 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>
- [13] Herman, M. and Robins, J.M. (2020) Causal Inference: What If. Chapman & Hall/CRC.
- [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>