The Korean Journal of Applied Statistics
2024, Vol. 37, No. 5, 675–690

DOI: https://doi.org/10.5351/KJAS.2024.37.5.675
Print ISSN1225-066X / Online ISSN 2383-5818

Analyzing the causal impact of streaming service usage on IPTV viewing

Dahai Jung[a], Yongho Yoon[b], Kwonsang Lee[1,b]

[a]Department of Statistics, Sungkyunkwan University; [b]Department of Statistics, Seoul National University

Abstract

In modern society, the rapid growth of streaming services has significantly changed the way people consume television. This study aims to analyze the causal impact of streaming service usage on IPTV viewing. To achieve this, we compared users who use streaming services with those who do not while controlling for many possible confounders. We employed causal inference matching methods, focusing particularly on several matching techniques to compare groups with similar characteristics. Additionally, we used regression methods using matching-driven weights to assess the statistical significance of the causal effect. The results indicate that streaming service usage has a significant impact on how IPTV is consumed. These findings provide important insights for content providers, broadcasters, and advertisers to understand viewer behavior patterns and make strategic decisions accordingly. This study offers new insights into the relationship between streaming services and traditional TV viewing and can serve as a foundation for future related research.

Keywords: causal inference, IPTV, matching, Netflix, panel data

1. 서론

최근 온라인동영상서비스(over the top; OTT)의 열풍은 빅데이터를 기반으로 시작되었다고 해도 과언이 아니다. 대표적인 예로, 유튜브 AI (artificial intelligence) 알고리즘은 시간 가는 줄 모르고 동영상을 끊임없이 시청하게 만들며, 넷플릭스는 개인화 추천 서비스(curation service)를 통해 시청을 유도하고 있다. 이처럼 사람들은 글로벌 OTT 플랫폼(유튜브, 넷플릭스, 디즈니+ 등)뿐만 아니라, 국내 OTT 플랫폼(티빙, 웨이브, 쿠팡플레이 등)에도 노출되고 있다. 특히, 이러한 OTT 시장의 성장을 촉진한 주요 원인 중 하나는 코로나19(COVID-19) 사태이다. 팬데믹 상황이 장기화되자, 격리된 공간에서 문화생활을 즐기기 위해 미디어 소비가 증가하였고, 월 이용료만 내면 콘텐츠를 무제한으로 시청할 수 있는 OTT 플랫폼으로 엄청난 유입이 발생했다. 그 중 가장 많이 성장한 사업자는 넷플릭스로, 월간 사용자수(monthly active users; MAU) 기준으로 코로나 이전보다 이용자가 약 50% 이상 증가했다. 이후 팬데믹이 해제되었지만, OTT 열풍은 쉽게 가라앉지 않고 있다. 영화관 관람료 인상에 따른 소비자 부담이 늘었을 뿐만 아니라, 핸드폰, 태블릿 등과 같이 휴대용 디바이스를 통해 어디에서나 시청이 가능하다는 장점이 크게 작용한 것으로 보인다.

팬데믹의 영향과 더불어 AI 기술의 발전이 OTT 플랫폼으로의 유입을 더욱 촉진시키는 것으로 나타난다. AI 기술을 활용한 개인화 추천 서비스는 사용자 경험을 크게 향상시키고 있다. AI알고리즘은 사용자의 과거

This work is supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (NRF-2021R1C1C1012750) and the New Faculty Startup Fund from Seoul National University.

[1]Corresponding author: Department of Statistics, Seoul National University, 1 Gwanak-ro, Gwanak-Gu, Seoul 08826, Korea. E-mail: kwonsanglee@snu.ac.kr

Published 31 October 2024 / journal homepage: http://kjas.or.kr

행동 데이터를 분석하여 개인 맞춤형 콘텐츠를 제공함으로써 사용자가 선호할 만한 동영상이나 프로그램을 예측하고 추천한다. 넷플릭스와 유튜브와 같은 OTT 플랫폼에서는 AI 기반의 추천 시스템을 통해 사용자의 시청 이력과 평가 데이터를 실시간으로 분석하여 관련성이 높은 콘텐츠를 제공하고 있다. 머신러닝, 자연어 처리, 실시간 분석 등의 AI 기술이 결합된 서비스는 사용자가 보다 쉽게 자신에게 맞는 콘텐츠를 발견하게 도와주며, 이를 통해 사용자 만족도를 높이고 플랫폼의 이용률을 증가시키고 있다.

이러한 변화로 인해 가장 큰 어려움을 맞게 된 사업이 바로 유료방송 사업이다. OTT 서비스가 유료방송의 '대체재' 역할을 하면서, 주요 미디어 소비 매체였던 유선방송을 해지하거나, 더 저렴한 상품으로 바꾸는 코드컷팅(cord-cutting)과 코드쉐이빙(cord-shaving)이라는 현상이 발생하였다. 한 유료방송 사업자는 이러한 현상을 적대시하는 것이 아니라 글로벌 OTT를 모두 품을 수 있는 오픈 플랫폼으로 진화하겠다고 선언하였고, 또다른 사업자는 넷플릭스 측에 망 이용대가를 요구하며 소송을 진행하기도 하였다.

본 논문에서는 이러한 AI 기술이 결합된 서비스를 제공하는 OTT 플랫폼의 컨텐츠 이용이 실제 유선방송을 포함한 IPTV (internet protocol television) 시청에 미치는 영향을 분석하고자 한다. 간접적으로는 실시간 TV시청과 같이 상대적으로 수동적으로 컨텐츠를 이용하던 소비자가 개인화 맞춤 서비스를 통해 능동적으로 컨텐츠를 접할 때 변화되는 양상을 이해할 수 있을 것으로 기대한다. 본 연구에서는 IPTV 시청의 패턴을 실시간 TV시청과 VoD시청으로 나누어 빅데이터 분석을 통해 어떠한 패턴으로 영향이 발생하는지 확인하고자 한다. 또한 패널데이터를 이용하여, 당장 나타나는 영향뿐만 아니라 1, 2개월 후의 영향도 같이 분석하고자 한다. OTT 시청여부가 미치는 영향을 단순한 상관관계에 대한 분석이 아니라 인과효과를 추정하는 것을 목표로 진행할 것이며, 이는 AI알고리즘의 성능향상을 위한 빅데이터 분석에서 인과추론의 역할이 중요하다는 것을 강조할 수 있을 것으로 기대한다.

본 연구를 위해 산학협력을 통해 수집된 패널 데이터를 활용한다. 해당 패널 데이터는 월 단위 자료로, 해당 사업자의 유료방송 가입자 중 무작위로 선별하여 가구별 넷플릭스 시청여부와 유료방송 내 실시간 시청시간, VoD (video on demand) 시청시간에 대해 반복 측정하였다. 특히, 해당 패널 데이터 특성에 따라 기존 패널 매칭(panel matching) 방법론과 비복원 패널 매칭(panel matching without replacement) 방법을 사용하여 OTT 구독에 따른 인과(처리)효과(treatment effect)를 비교하고자 한다. 또한, 해당 결과에 따라 유료방송 사업자는 해지방어 또는 이용증대 캠페인 등에 활용할 예정이다. 먼저 2장에서는 인과추론의 개요 및 매칭방법론를 소개하고 패널데이터에서의 적용에 대해 서술하며, 3장에서는 비복원추출을 이용한 최적매칭방법론을 소개한다. 4장에서는 데이터 구조를 소개하고, 5장에서 해당 데이터에 패널 매칭 방법론을 적용한 결과와 처리효과를 비교한다. 마지막으로 6장에서는 결론 및 향후 연구에 대해 논의한다.

2. 패널데이터에서의 인과추론

2.1. 인과추론의 개요

인과추론 방법론은 관심있는 두 변수 사이에 인과적 관계가 성립하는지 알아낼 수 있는 통계적 방법론이다. 그동안 인과관계를 확인할 수 있는 많은 인과추론 방법론이 개발되어 왔으며, 데이터에 맞는 다양한 가정을 통해 효과적으로 인과효과를 추정할 수 있는 방법론이 함께 개발되어 왔다. 본 논문에서 기본 데이터를 여러 시점에서 데이터가 반복적으로 축적되는 패널데이터로 가정하고, 패널데이터에서의 인과추론에 대한 내용을 서술한다.

패널데이터에서는 각 유닛과 시간을 아래첨자 it를 이용하여 표현한다. 여기서 i는 유닛을 뜻하며, t는 시간을 뜻한다. 각 it에 대하여, 처리변수 Z_{it}는 이진(binary)이며 처리를 받았을 때에는 1, 아닌 경우에는 0으로 정의한다. 예를 들어, 넷플릭스 데이터에서는 유닛 i가 시간t에서 넷플릭스를 구독하는 경우에는 $Z_{it} = 1$이 된다. 결과변수는 잠재적 반응 변수(potential outcome)로 가정하여, 두 $Y_{it}(1), Y_{it}(0)$를 정의한다. 모든 it에 대해 처리를 받았을 경우 $Y_{it}(1)$이 되며 받지 않았을 경우 $Y_{it}(0)$가 된다고 생각하여, 처리의 유무에 따라 둘 중 하나

만 관찰된다. 이 경우 각 it에 대한 인과효과는 $Y_{it}(1) - Y_{it}(0)$로 정의되지만 두 잠재변수 모두를 관찰할 수가 없기 때문에 이 인과효과는 관찰이 불가능하다 (Holland, 1986). 개별인과효과(individual causal effect) 대신 평균인과효과(average causal effect)는 몇 가지 가정하에서 확인(identification)이 가능하다.

첫 번째 가정으로 SUTVA (stable unit treatment value assumption)가 있으며 각 유닛간에는 서로 영향을 주고 받지 않는다는 가정이 전제되어 있다. 이 가정에는 일관성(consistency) 가정을 포함하고 있다. 두 번째, 무시성(ignorability)이라는 가정으로 우리가 교란변수(confounder) X_{it}를 모두 찾아낼 수 있다면, 이를 통제하고 난 다음에는 처리변수와 결과변수는 독립이라고 본다. 즉,

$$(Y_{it}(1), Y_{it}(0)) \perp\!\!\!\perp Z_{it} \mid X_{it}. \tag{2.1}$$

세 번째, 처리를 받을 확률은 항상 0과 1사이에 있다는 중복(overlap) 가정이 필요하다. 이를 표현하자면 $0 < \Pr(Z_{it} = 1|X_{it}) < 1$이다. 서술한 세 가지 가정이 성립하는 경우 인과효과를 확인할 수 있다. 일반적으로 평균인과효과는 $\tau^{ATE} = \mathbb{E}[Y_{it}(Z_{it} = 1) - Y_{it}(Z_{it} = 0)]$으로 정의된다.

2.2. 패널데이터에서의 인과효과

본 논문에서의 인과효과에 대한 정의를 수정하여 살펴볼 예정이다. 패널데이터에서의 인과효과는 단순히 t 시점에 대한 효과를 넘어서 $t + F, F = 0, 1, \ldots$시점에 대한 인과효과도 고려한다. 일반적으로 결과변수는 t 시점 이후를 보는 경우가 많지만, 처리가 되는 시점($Z_{it} = 1$)과 결과가 측정되는 시점이 t에서 동시에 일어나는 경우도 생각하여 $F = 0$도 포함한다. $F = 0$은 처리 후 나타나는 동시효과를 의미하며, $F = 2$는 처리 후 두 기간 지난 뒤의 효과를 의미한다. F값의 크기에 따라 단기 또는 장기(누적) 효과를 살펴볼 수 있다 (Imai와 Kim, 2019). 평균인과효과는 다음과 같이 정의한다.

$$\begin{aligned}
\tau^{ATE}(F) &= \mathbb{E}\Big\{Y_{i,t+F}\left(Z_{it} = 1, Z_{i,t-1} = 0\right) - Y_{i,t+F}\left(Z_{it} = 0, Z_{i,t-1} = 0\right) \mid Z_{i,t-1} = 0\Big\} \\
\tau(F) &= \mathbb{E}\Big\{Y_{i,t+F}\left(Z_{it} = 1, Z_{i,t-1} = 0\right) - Y_{i,t+F}\left(Z_{it} = 0, Z_{i,t-1} = 0\right) \mid Z_{it} = 1, Z_{i,t-1} = 0\Big\}, \quad F = 0, 1, 2, \ldots.
\end{aligned} \tag{2.2}$$

시점 $t - 1$에는 처리를 받지 않는 집단($Z_{i,t-1} = 0$)을 대상으로 인과효과를 정의한다. 이러한 집단을 조건부로 두었을 때, 처리를 받기 전에는 모든 유닛이 처리를 받지 않는 동일한 상황을 설정하여 이전 처리에 대한 효과를 고려하지 않도록 설정할 수 있다. 또한 조건부확률에서 $Z_{it} = 1$이라는 조건을 추가로 설정한다면, 시점 t에서 처리를 받은 집단에 한정한 인과효과를 추정할 수 있으며 이를 처리군에 대한 평균인과효과(average treatment effect on treated; ATT)라고 정의한다. $\tau^{ATE}(F)$ 대신 $\tau(F)$ (ATT)를 추정하는 경우는 과거에 이미 발생한 처리의 효과에 대한 추정을 원할때 사용한다. 이를 후향적 효과라고 부르며 넷플릭스 사용 데이터에서는 과거의 스트리밍 서비스 사용이 IPTV 시청 기록에 미친 영향을 분석할 때 이러한 후향적 인과효과가 대상이 된다. 또한 인과추론 방법론에서 매칭방법론이 ATT에 대한 추정을 가능하게 한다.

식 (2.2)에서 새로 정의한 인과효과의 확인을 위해 2.1절에서 서술한 두 번째 무시성 가정을 수정해야 한다. 시점 $t - 1$에서 기록된 정보인 $X_{i,t-1}$을 조건부로 고려하는 것으로 조건부 독립가정이 성립한다고 생각한다.

$$Y_{i,t+F}\left(Z_{it} = 1, Z_{i,t-1} = 0\right), \quad Y_{i,t+F}\left(Z_{it} = 0, Z_{i,t-1} = 0\right) \perp\!\!\!\perp Z_{it} \mid Z_{i,t-1} = 0, Y_{i,t-1}, X_{i,t-1} \tag{2.3}$$

위의 가정은 음이 아닌 정수 L을 Lag 값으로 선택하여 시점 $t-1$부터 시점 $t-L$까지 과거의 처리변수 $\{Z_{i,t-l}\}_{l=1}^{L}$, 결과변수 $\{Y_{i,t-l}\}_{l=1}^{L}$ 그리고 교란변수 $\{X_{i,t-l}\}_{l=1}^{L}$ 모두를 조건부로 고려했을 때로 확장시킬 수 있다. 이는 순차적 무시성(sequential ignorability) 가정으로 볼 수 있다 (Robins 등, 2000). 이 때 $\{X_{i,t-l}\}_{l=1}^{L}$는 시간에 따라 변할 수 있는 시간가변(time-varying) 변수로 간주한다. 식 (2.3)는 다음의 일반화된 가정에서 $L = 1$인 경우로 생각할

수 있다.

$$Y_{i,t+F}\left(Z_{it}=1, Z_{i,t-1}=0, \{Z_{i,t-l}\}_{l=2}^{L}\right), Y_{i,t+F}(Z_{it}=0, Z_{i,t-1}=0, \{Z_{i,t-l}\}_{l=2}^{L}) \perp\!\!\!\perp Z_{it} \mid Z_{i,t-1}=0, \{Z_{i,t-l}\}_{l=2}^{L}, \{Y_{i,t-l}\}_{l=1}^{L}, \{X_{i,t-l}\}_{l=1}^{L} \tag{2.4}$$

식 (2.4)은 순차적 무시성에 대한 가정을 시점 $t-L$로 제한을 둔 경우라 생각한다. 이러한 가정을 바탕으로 $\tau(F)$를 수정하여 $t-L$시점까지의 과거 처리 기록을 고려한 $\tau(F,L)$을 정의할 수 있다. $\tau(F,L)$은 식 (2.4)을 바탕으로 확인할 수 있다 (Imai 등, 2023).

$$\tau(F,L)=\mathbb{E}\Big\{Y_{i,t+F}\left(Z_{it}=1, Z_{i,t-1}=0, \{Z_{i,t-l}\}_{l=2}^{L}\right) - Y_{i,t+F}\left(Z_{it}=0, Z_{i,t-1}=0, \{Z_{i,t-l}\}_{l=2}^{L}\right) \mid Z_{it}=1, Z_{i,t-1}=0\Big\}, \quad F=0,1,2,\ldots. \tag{2.5}$$

시점 $t-L$까지 기록을 고려하는 $\tau(F,L)$대신 주변화된(marginalized) $\tau(F)$를 고려하는 것이 실증적 연구에서는 결과에 대한 해석이 용이한 경우도 있다. 적절한 L의 선택은 이월(carryover) 효과를 고려하여 이루어져야 한다. 앞으로 소개할 매칭방법론과 연결하여 L의 선택에 관한 논의는 2절과 3절에서 자세히 다루도록 한다.

패널데이터를 사용하는 실증적 연구에서는 관찰되지 않은 교란 변수(unmeasured confounder)의 잠재적 존재가 우려요인이 될 수 있다. 이 때 식 (2.4)이 성립하지 않을 가능성이 존재하며, 따라서 $\tau(F,L)$에 대한 추정이 신뢰할 수 없을 가능성도 존재한다. 이를 해결하기 위해 식 (2.4)대신 DID (difference-in-differences) 디자인을 고려하고자 한다 (Abadie, 2005). 구체적으로, 처리, 결과 및 교란변수의 과거 기록을 조건으로 한 후 다음과 같은 평행추세(parallel trend) 가정을 한다.

Assumption 1. (*Parallel trend assumption*)

$$\mathbb{E}\Big[Y_{i,t+F}\left(X_{it}=0, X_{i,t-1}=0, \{X_{i,t-l}\}_{l=2}^{L}\right) - Y_{i,t-1} \mid X_{it}=1, X_{i,t-1}=0, \{X_{i,t-l}, Y_{i,t-l}\}_{l=2}^{L}, \{Z_{i,t-l}\}_{l=0}^{L}\Big] = \mathbb{E}\Big[Y_{i,t+F}\left(X_{it}=0, X_{i,t-1}=0, \{X_{i,t-l}\}_{l=2}^{L}\right) - Y_{i,t-1} \mid X_{it}=0, X_{i,t-1}=0, \{X_{i,t-l}, Y_{i,t-l}\}_{l=2}^{L}, \{Z_{i,t-l}\}_{l=0}^{L}\Big], \tag{2.6}$$

이 평행추세 가정이 관찰되지 않은 시간가변 교란변수들을 설명할 수 없다는 것이 알려져 있다. 처리를 받은 집단과 매칭된 통제 집단간에 결과에 대한 시간 추세가 실제로 평균적으로 평행한지 사전 처리 기간의 데이터를 사용하여 확인해야 한다.

2.3. 매칭방법론 리뷰

본 논문에서는 Assumption 1의 평행추세 가정을 통해 $\tau(F,L)$ (또는 $\tau(F)$)에 대한 추정을 진행하고자 한다. 추정 방법은 역확률 가중치를 이용하는 가중치(weighting)방법을 비롯하여 여러 방법들이 개발이 되었지만, 여기에서는 매칭(matching)방법론을 소개하고자 한다. 매칭을 사용하여 패널데이터를 분석하는 연구는 최근 Imai 등 (2023) 논문에서 소개되었다. 패널데이터에서의 매칭을 논의하기 전에 일반적인 매칭방법론에 대한 리뷰를 진행한다.

일반적으로 매칭은 두 단계로 진행된다. 첫 번째 단계는 설계(design)단계로 대조군에서 처리군의 공변량(covariate) 결합분포와 동일한 분포를 가지는 부분집합을 선택하는 과정을 가진다. 이때 선택하는 과정에서 처리군의 각각의 유닛은 비슷한 특성을 가지는 대조군과 짝지어지는 과정이 동반된다. 두 번째 단계는 분석(analysis)단계로 선택된 대조군의 부분집합 만을 이용하여 추정치를 계산한다. 설계 단계에서 주어지는 매칭 구조에 따라 분석하는 방식이 달라지며, 또한 같은 구조에 대해서도 회귀식을 이용한 방법과 무작위화 기반 방법과 같은 다양한 방법을 고려할 수 있다.

설계 단계에서 부분집합을 선택하고 짝짓는 과정을 수행할 수 있는 많은 매칭방법론이 연구되었다. 이 중에서 가장 단순한 방법인 1:1매칭을 고려할 수 있다. 먼저 두 집단 내의 유닛 간의 거리를 정의하고, 대조군의 크기가 처리군의 크기보다 큰 경우에는 각 처리군 유닛이 가장 가까운 대조군 유닛 하나를 찾는 방식으로 진행된다. 이렇게 진행되면 처리군의 크기와 같은 크기를 가지는 대조군의 부분집합이 선택된다. 만약 대조군의 크기가 처리군의 크기에 비해 많이 크다면, 1:1매칭을 이용하였을 때에는 선택된 부분집합의 크기가 기존 대조군의 크기에 비해 많이 작아 데이터를 충분히 활용하지 못한다는 단점이 있다. 이를 해결하기 위해 1:k매칭과 같은 방법을 생각할 수 있다. 적절한 k는 대조군의 크기/처리군의 크기를 고려하여 선택될 수 있다. 그러나 1보다 큰 k가 선택되었더라도 1:k매칭에서도 마찬가지로 모든 데이터를 활용하지 못하는 경우가 발생한다. 일부 연구자들은 매칭 과정에서 매칭되지 않은 유닛을 제외하는 것에 대해 회의적인 견해를 가지고 있다. 하지만 Rosenbaum (2010)에서도 언급되었듯이, 모든 데이터를 활용하지 못하는 점이 꼭 추정의 효율성(efficiency)을 감소시키지는 않는다. 오히려 비슷하지 않은 대조군의 유닛을 제외함으로써 효율성이 증가하는 현상이 발견되며, 이는 역확률 가중치 추정에서 몇몇 극단적인 가중치의 절단(truncation)이 효과적으로 추정의 효율을 증가시킨다는 연구와 연결되어 있다. 그럼에도 데이터를 버리지 않은 매칭에 대한 연구가 진행되었고 대표적으로 full matching (Hansen, 2004)이라고 부르는 매칭방법을 고려해볼 수 있다. 매칭과정을 더욱 세밀하게 나눠 부분집합을 선택하는 과정과 짝짓는 과정을 분리할 수도 있다. 이와 관련한 내용은 Zubizarreta 등 (2014)에서 논의되었다.

두 유닛간의 거리를 측정할 때는 두 유닛의 공변량 값이 얼마나 다른지에 초점을 맞출 수 있다. 공변량 값이 가까운 두 유닛이 선택되었을 때 공변량의 결합분포가 비슷해질 가능성이 커지기 때문이다. 가장 이상적인 방법은 같은 공변량을 가진 유닛들만 매칭할 수 있게 하는 exact matching방법을 고려할 수 있다. 이 경우에는 매칭 이후에는 처리군과 매칭된 대조군의 공변량의 분포가 정확하게 일치한다. 하지만 공변량의 차원이 증가함에 따라 exact matching의 사용은 어려워진다. 이렇게 exact matching이 불가능한 경우, 두 공변량 X_i와 X_j의 거리 $d(X_i, X_j)$는 마할라노비스 거리(Mahalanobis distance)를 이용하여 정의한다면 매칭을 적용할 수 있다. 마할라노비스 거리는 다음과 같이 정의된다:

$$d^M\left(X_i, X_j\right) = \sqrt{\left(X_i - X_j\right)^T \Sigma^{-1} \left(X_i - X_j\right)},$$

Σ는 X_i의 공분산행렬이다. 연구자에 따라 제곱근이 없는 거리를 사용하기도 한다. 마할라노비스를 이용한 거리는 X_i들이 정규분포를 따른다는 전제하에 정의하기 때문에 만약 X_i들이 정규분포와 다른 분포를 따른다면 $d^M(X_i, X_j)$가 제대로된 거리로서 역할을 하지 못할 수도 있다. 이를 해결하기 위해 순위 기반 마할라노비스 거리(rank-based Mahalanobis distance)가 사용된다:

$$d^{RM}\left(X_i, X_j\right) = \sqrt{\left(r\left(X_i\right) - r(X_j)\right)^T \tilde{\Sigma}^{-1} \left(r\left(X_i\right) - r(X_j)\right)},$$

$r(X_i)$는 각 요소별 순위(component-wise rank) 벡터이며, $\tilde{\Sigma}$는 $r(X_i)$의 공분산행렬이다. 순위 기반 마할라노비스 거리는 국지적(locally)으로 가까운 공변량을 선택하는 좋은 방법이라 볼 수 있다.

전역적(globally)으로 가까운 공변량을 선택할 수 있게 만드는 거리는 성향점수(propensity score)를 이용한 방법이다. 성향점수는 공변량이 주어졌을 때 처리를 받을 확률인 $e(X_i) = \Pr(Z_i = 1 \mid X_i)$을 말한다. 공변량이 다차원이라도 성향점수는 항상 일차원의 스칼라 값을 가진다. 성향점수는 중요한 성질 중 하나로 인과추론의 무시성 가정 (2.1)에서 조건부 부분을 X_{it}대신 $e(X_{it})$로 대체할 수 있는 점이 있다 (Rosenbaum과 Rubin, 1983). 나아가 소위 균형성질(balancing property)라고 부르는 조건 $X_i \perp\!\!\!\perp Z_i \mid e(X_i)$를 만족시키며, 따라서 매칭과정에서 공변량이 같을 필요가 없고 일차원 값이 성향점수만 같아도 됨을 의미한다. 성향점수를 이용한 거리 $d^{PS}(X_i, X_j)$는 다음과 같이 정의한다.

$$d^{PS}\left(X_i, X_j\right) = |e(X_i) - e(X_j)|.$$

그러나 실증적 연구에서는 $e(X_i)$는 알려지지 않아 추정이 필요하다. 거리의 계산도 성향점수에 대한 추정치를 바탕으로 정의된 버전 $d^{PS}(X_i, X_j) = |\hat{e}(X_i) - \hat{e}(X_j)|$을 사용한다. 성향점수거리를 이용한 매칭방법은 특별히 성향점수매칭(propensity score matching; PSM)으로 불리며 많은 응용연구에서 사용된다. 하지만 성향점수의 중요한 성질은 대표본(large sample)에서 성립하여 충분한 데이터가 있지 않을 때에는 성향점수를 이용한 방법은 효율적이지 않다. 국지적인 거리를 잘 나타내는 순위 기반 마할라노비스 거리와 결합하여 PSCM (propensity score caliper matching)을 이용하는 것이 추천된다. 성향점수 차이의 제한폭(caliper)을 설정하여, 두 유닛의 성향점수의 차이가 제한폭 바깥이면 매칭이 되지 않도록 하고 제한폭 안쪽이면 (순위 기반) 마할라노비스 거리를 이용하여 매칭한다. 제한폭 δ에 대해 거리 $d^{PSCM}(X_i, X_j)$는 다음과 같이 정의된다:

$$d^{PSCM}\left(X_i, X_j\right) = \begin{cases} d^{RM}\left(X_i, X_j\right), & \text{if } \; d^{PS}\left(X_i, X_j\right) \le \delta, \\ \infty, & \text{otherwise.} \end{cases}$$

다양한 거리의 정의와 이에 관련한 논의는 Stuart (2010)을 참고할 수 있다.

거리를 정의하더라도 가까운 거리에 있는 처리군과 대조군의 유닛들을 선택하는 문제는 최적화 문제와 연결되어 있으며, 정수 프로그래밍(integer programming)을 이용하여 해결할 수 있다. 최적화 문제를 효과적으로 해결할 수 있는 최근 매칭 알고리즘에 관한 내용은 Rosenbaum (2020)에서 논의되었다. 또한, 데이터의 크기가 큰 경우에는 효율적인 계산을 위해 최적화 문제를 좀 더 단순하게 바꿀 필요가 있다. PSCM처럼 제한폭을 설정하여 매칭 가능한 유닛들을 선별하는 과정은 매칭에서 네트워크를 희소(sparse)하게 구성하는 방법 중 하나로 설명할 수 있다. 최적화 문제에서 희소한 네트워크는 많은 계산 부담을 덜어주어 효율적으로 문제를 해결할 수 있다. 이와 관련된 더욱 자세한 내용은 Yu 등 (2020)과 Yu와 Rosenbaum (2022)에서 논의되었다.

2.4. 패널데이터에서의 매칭

최근 패널데이터에서의 매칭방법론에 대한 연구가 많이 진행되고 있으며, 그중에서도 Imai 등 (2023)의 논문을 리뷰하고자 한다. 이 논문에서는 식(2.5)을 추정하기 위해서 가정(2.6)를 이용한다. 매칭과정에서는 최적화 문제의 해결을 피하고자 복원추출을 이용하여 대조군 유닛을 선택한다. 특징적으로 $\mathcal{M}_{it}$라는 집합이

$$\mathcal{M}_{it} = \{i' : i' \neq i, Z_{i',t} = 0, Z_{i',t-l} = Z_{i,t-l} \;\; \text{for all} \;\; l = 1, \ldots, L\} \tag{2.7}$$

로 정의되며 처리군 유닛 i의 시간 t 이전의 처리변수 기록과 같은 대조군에서의 부분집합을 선택하는 과정이라 볼 수 있다. 조건 $Z_{i',t-1} = 0$이 대조군에 포함됨을 의미하며, 조건 $Z_{i',t-l} = Z_{i,t-l}$는 모든 lag에서 같은 기록을 가지는 유닛을 선택함을 의미한다. 큰 값을 가지는 L을 선택했을 경우에는 $\mathcal{M}_{it}$가 존재하지 않을 수도 있어, 데이터를 통해 적절한 L을 선택하는 과정이 먼저 선행된다. 부분집합 $\mathcal{M}_{it}$ 선택이 이루어진 다음에는 1:k매칭처럼 가장 거리가 가까운 k개의 유닛을 선택한다. 이 때 처리군 유닛 i에 대해 대조군 유닛 i'에 대한 거리는 다음과 같이 정의한다.

$$S_{it}(i') = \frac{1}{L}\sum_{l=1}^{L} \sqrt{(X_{i,t-l} - X_{i',t-l})^{\top} \Sigma_{i,t-l}^{-1} (X_{i,t-l} - X_{i',t-l})}, \tag{2.8}$$

$X_{i,t-l}$은 시간 $t-l$에서 시간가변 변수들을 매칭을 통해서 통제하고자 하는 모든 변수를 포함한다. $\Sigma_{i,t-l}$은 $X_{i,t-l}$의 공분산 행렬이다. $S_{it}(i')$은 각각의 시간가변 변수들의 시간 $t-1$부터 시간 $t-L$까지의 거리를 평균낸 거리로 이해할 수 있다.

모든 유닛에 대해 거리 측정치 $S_{it}(i')$이 계산되면, 연구자가 지정한 칼리퍼 제약 C를 만족하는 가장 유사한 유닛을 최대 k개를 선택하여 매칭 집합을 정제하고, 나머지 매칭된 통제 유닛에는 가중치를 0으로 설정한다.

이렇게 함으로써, 정제된 부분집합 $\mathcal{M}_{it}^{*} \subset \mathcal{M}_{it}$을 선택한다.

$$\mathcal{M}_{it}^{*} = \left\{ i' : i' \in \mathcal{M}_{it}, S_{it}(i') < C, S_{it}(i') \leq S_{it}^{(k)} \right\}, \tag{2.9}$$

여기서 $S_{it}^{(k)}$는 $S_{it}(i')$ 중에서 k번째 순서통계량이고, C는 PSCM에서처럼 칼리퍼로 이용된다. 정제된 $\mathcal{M}_{it}^{*}$를 이용하여 유닛 i의 시간 t에서의 가중치를 계산할 수 있다. 즉,

$$w_{it}^{i'} = \begin{cases} \frac{1}{|\mathcal{M}_{it}^{*}|}, & \text{if } \ i' \in \mathcal{M}_{it}^{*}, \\ 0, & \text{otherwise.} \end{cases} \tag{2.10}$$

또한 이러한 가중치는 $\sum_{i' \in \mathcal{M}_{it}^{*}} w_{it}^{i'} = 1$을 만족시킨다.

정제된 집합 $\mathcal{M}_{it}^{*}$와 이를 기반으로 얻어진 가중치를 기준으로 식 (2.5)에서 정의된 처리군에 대한 평균인 과효과 ATT를 추정할 수 있다. 이를 위해, 가중평균을 이용하여 잠재적 반응 변수인 $Y_{i,t+F}(Z_{it} = 0, Z_{i,t-1} = 0, \{Z_{i,t-l}\}_{l=2}^{L})$을 먼저 추정한다. 그런 다음 각각에 대해 ATT의 DID추정치를 계산한다.

$$\hat{\tau}(F, L) = \frac{1}{\Sigma_{i=1}^{N} \Sigma_{t=L+1}^{T-F} D_{it}} \sum_{i=1}^{N} \sum_{L+1}^{T-F} D_{it} \left\{ (Y_{i,t+F} - Y_{i,t-1}) - \sum_{i' \in \mathcal{M}_{it}} w_{it}^{i'} (Y_{i',t+F} - Y_{i',t-1}) \right\}, \tag{2.11}$$

$D_{it} = Z_{it}(1 - Z_{i,t-1}) \cdot 1(|\mathcal{M}_{it}| > 0)$이고 $w_{it}^{i'}$은 식 (2.10)에서 얻은 가중치이다. $D_{it} = 1$은 $Z_{i,t-1} = 0, Z_{it} = 1$인 처리군에서 발생하며 매칭이 가능한 it에 대해서만 발생한다.

추정값은 회귀분석식을 통해서 쉽게 구할 수 있다. 특히 $\tilde{Y}_{i,t+F} = Y_{i,t+F} - Y_{i,t-1}$로 설정하여 마치 $\tilde{Y}_{i,t+F}$가 실제로 관측된 결과변수인 것처럼 생각하고 진행하면 된다. 가중치 $w_{it}^{i'}$을 이용하여 처리변수만 포함된 단순선형회귀모형에서 가중치가 적용된 가중회귀분석을 하게 된다면, 처리변수에 대한 계수 추정치와 함께 신뢰구간과 p값과 같은 통계적 추론 값들도 함께 얻을 수 있다. 매칭을 통해 처리군과 통제군 사이의 교란변수의 분포가 비슷하게 맞춰졌지만 완전히 동일하지는 않다. 이에 대한 영향을 제거하고자 교란변수를 이전의 단순선형회귀모형에 추가할 수 있다. 이때에도 마찬가지로 처리변수에 대응하는 계수에 대한 추론을 진행할 수 있다. 이때 다른 변수에 대응하는 계수에 대한 추론은 하지 않는다 (Ding, 2024).

소개한 매칭방법은 R-패키지 `PanelMatch`로 구현되었고, `panelmatch` 함수로 결과를 분석 할 수 있다. 분석결과는 가중치를 이용한 회귀모형을 적용한 결과이다.

3. 패널데이터에서의 최적매칭방법론

본 논문에서는 기존 Imai 등 (2023)에서 소개된 패널데이터 매칭을 비복원추출을 이용한 매칭으로 발전시키고자 한다. 부분집합 $\mathcal{M}_{it}$와 정제된 부분집합 $\mathcal{M}_{it}^{*}$의 선택과정에 있어서 Imai 등 (2023)에서는 기본적으로 복원추출을 상정한다. 예를 들어, 처리군에서 같은 처리변수 기록을 가지는 두 유닛 i_1과 i_2가 있다고 가정하자. 같은 처리변수 기록을 공유하는 두 유닛에 대한 부분집합 $\mathcal{M}_{it}$는 서로 같다. 즉, $\mathcal{M}_{i_1,t} = \mathcal{M}_{i_2,t}$이다. 하지만 처리군 유닛 i_1에 대응하는 $\mathcal{M}_{i_1,t}^{*}$의 선택과 다른 처리군 유닛 i_2에 대응하는 $\mathcal{M}_{i_2,t}^{*}$의 선택은 독립적으로 이루어진다. 따라서 $\mathcal{M}_{i_1,t}^{*} \cap \mathcal{M}_{i_2,t}^{*} \neq \emptyset$인 경우가 발생한다. 반면, 비복원추출의 경우에는 $\mathcal{M}_{i_1,t}^{*} \cap \mathcal{M}_{i_2,t}^{*} = \emptyset$이 되도록 설정한다. 대략적으로 말해서, 복원추출은 상대적으로 빠른 계산이 가능하다는 장점이 있지만, 한 유닛이 여러차례 매칭되어 결과값에 편향이 생기고 실제 랜덤실험 상황처럼 해석할 수 없다는 단점과 극단적인 가중치를 얻는 상황이 발생한다는 문제점이 있다. 비복원추출과 복원추출은 크기가 큰 매칭 문제에서는 비슷한 성능을 보이는 것으로 알려져 있지만, 크기가 작은 문제에서는 비복원추출이 선호된다. 특히 패널데이터 매칭에서는 L의 커짐에 따라 $\mathcal{M}_{it}$의 크기가 작아지고 결국 작은 매칭 문제로 바뀌기 때문에 복원추출의 문제가 발생할 수 있다.

Table 1: Summary statistics

Variables		Unmatched treated	Unmatched controls	Controls after 1:k matching	Controls after 1:1 matching
Netflix	N	1,167	14,041	-	-
purchase	구매	55%	46%	54%	54%
rltm_tot_hr.lag3	Mean	198.104	210.161	194.489	194.001
rltm_tot_hr.lag2	Mean	200.275	209.393	195.513	195.410
rltm_tot_hr.lag1	Mean	215.107	227.546	212.316	213.078
rltm_mv_hr.lag3	Mean	16.627	15.367	16.128	15.920
rltm_mv_hr.lag2	Mean	12.335	11.544	11.916	11.422
rltm_mv_hr.lag1	Mean	15.099	14.485	14.703	14.602
rltm_ent_hr.lag3	Mean	72.070	75.137	72.694	72.051
rltm_ent_hr.lag2	Mean	75.354	75.838	75.657	75.390
rltm_ent_hr.lag1	Mean	77.476	78.435	78.478	77.808
vod_tot_hr.lag3	Mean	15.141	13.566	14.346	14.535
vod_tot_hr.lag2	Mean	15.079	13.177	14.042	13.777
vod_tot_hr.lag1	Mean	19.088	15.577	17.278	17.626
vod_mv_hr.lag3	Mean	2.537	2.058	2.404	2.525
vod_mv_hr.lag2	Mean	2.601	1.968	2.445	2.630
vod_mv_hr.lag1	Mean	3.483	2.553	3.269	3.453
vod_drama_hr.lag3	Mean	4.390	3.700	4.133	3.840
vod_drama_hr.lag2	Mean	4.097	3.475	3.887	3.666
vod_drama_hr.lag1	Mean	5.954	3.952	5.546	5.612
vod_ent_hr.lag3	Mean	4.498	3.738	4.390	4.445
vod_ent_hr.lag2	Mean	4.655	3.790	4.493	4.358
vod_ent_hr.lag1	Mean	4.816	4.078	4.621	4.588
vod_tot_yn	시청	81%	78%	81%	81%

비복원추출을 기반으로 하는 매칭방법론 역시 같은 처리변수 기록을 공유하는 대조군의 부분집합 $\mathcal{M}_{it}$를 선별하는 작업이 선행되어야 한다. 같은 처리변수 기록을 공유하는 처리군의 유닛들은 모두 같은 $\mathcal{M}_{it}$를 가지게 된다. 따라서 비복원추출에서의 매칭은 처리변수 기록을 바탕으로 여러 개의 다른 기록을 가지는 하위집단을 설정하고 각 하위집단 내에서 매칭을 해결하는 문제로 바뀌게 된다. 이는 exact matching을 진행하는 과정과 동일하다. 하나의 큰 매칭 문제를 해결하는 대신에 여러 개의 작은 매칭 문제로 나누어지며 이는 크기가 큰 데이터에 대한 매칭을 적용하는데 유용하다. 예를 들어, 본 논문에서 사용하게 될 패널데이터에서는 $L = 3$으로 설정하여 4개의 하위집단을 구성하여 각 하위집단에서 매칭 문제를 해결한다. 일반적으로 m개의 같은 크기를 가지는 하위집단으로 분리되는 경우, 매칭 문제를 해결하는 시간은 $1/m^2$배가 된다. 각 하위집단 내에서의 매칭은 처리군과 대조군의 크기를 고려하여 1:1 매칭, 1:k 매칭 또는 full matching을 고려할 수 있다.

$\tau(F, L)$이 아닌 $\tau(F)$에 대한 추정을 원하는 경우에는 $\mathcal{M}_{it}$를 선별하는 작업이 필요하지 않으며 전체 처리군과 대조군을 매칭하는 하나의 큰 매칭문제를 풀면 된다. 하지만 크기가 큰 데이터의 경우에는 이러한 큰 매칭문제를 풀기위한 많은 계산량을 요구한다. 효율적으로 매칭문제를 해결하기 위해 $\mathcal{M}_{it}$와 같은 부분집합을 설정하는것이 도움이 된다. 가장 간단한 방법으로 성향점수를 이용하여 좀 더 희소한 네크워크를 설정하거나, 처리변수에 대한 기록에 대한 가중치를 설정하여 near-exact matching과 같은 방법으로 설정하는 방식이 사용될 수 있다.

비복원추출을 이용한 매칭방법론을 개발과 더불어 본 논문에서는 처리를 받는 기간에 대한 효과도 분석

Table 2: Classification by treatment history

		Description	Treatment history(t-3 ~ t+2)
T	T_1	Netflix subscription for 1 month at t time	**01**
	T_2	Netflix subscription for 2 continuous months since t time	**011*
	T_3	Netflix subscription for 3 continuous months since t time	**0111
C	C_1	Non-subscription to Netflix for 2 continuous months since $t-1$ time	**00**
	C_2	Non-subscription to Netflix for 3 continuous months since $t-1$ time	**000*
	C_3	Non-subscription to Netflix for 4 continuous months since $t-1$ time	**0000

하고자 한다. Imai 등 (2023)에서는 처리군과 대조군은 각각 $Z_{it} = 1, Z_{i,t-1} = 0$과 $Z_{it} = 0, Z_{i,t-1} = 0$을 만족하는 유닛으로 정의된다. 하지만 본 패널데이터에서의 처리는 넷플릭스의 시청여부와 연관되어 있어, 시간 t에서 넷플릭스를 시청하는 사람에 대한 특성을 고려해야 한다. 특징적으로 시간 t에서의 시청이 시간 $t+1$에서의 시청을 보장하지는 않는다. 따라서 더 오랜기간동안 넷플릭스를 시청하는 사람에 대한 인과효과가 다를 것이라고 생각하여 여러 다른 인과효과의 정의를 고려한다. 예를 들어, Table 2에서와 같이 시간 t에서부터 $t+2$까지 세 달 동안 넷플릭스를 시청한 처리군의 부분집단(T_3)을 고려할 예정이다. 이 때 처리군의 부분집단을 고려한 새로운 매칭방법론이 필요하며, 이는 더욱 세분화된 $\mathcal{M}_{it}$ 집합을 선택함으로써 해결할 수 있다. 관련 내용은 5절에서 더욱 자세히 논의할 예정이다.

4. 데이터 구조

4.1. 산학협력 자료

국내 유료방송 사업은 IPTV 및 위성방송 4개사, 종합유선방송사(5개사), 개별유선방송(9개사)로 총 18개의 사업자가 경쟁 중이다. 그 중에서도 IPTV는 인터넷을 통해 실시간 방송뿐만 아니라 VoD 및 정보검색, 쇼핑 등과 같은 인터넷 서비스까지 제공하기 때문에 빠르고 안정적인 인터넷이 기반이여야 하므로, 국내 IPTV 사업자는 망 사업자로 제한되어 있다. 일반적으로 소비자들은 인터넷과 IPTV 서비스를 묶음상품으로 이용하기 때문에, IPTV 사업자는 인터넷을 통한 넷플릭스 접속이력을 알 수 있다. 따라서, 해당 산학협력 자료는 국내 IPTV 사업자로부터 넷플릭스 접속이력과 IPTV 시청이력을 제공받아 진행되었다.

본 패널 데이터는 2020년 10월부터 2021년 3월까지 무작위로 추출된 유료방송 가입자의 회선별 자료로 구성되어있다. 변수는 기준년월, 고객 프로파일링 정보, 서비스 상품 정보, 해당 월 실시간 방송 전체 및 장르별(교양정보, 교육, 뉴스, 드라마, 만화, 스포츠, 시사다큐, 영화, 예능, 음악, 취미레저, 홈쇼핑) 시청시간(hour), 해당 월 VoD 전체 및 장르별(다큐, 드라마, 영화, 예능, 시사교양) 시청시간(hour), 최근 3개월동안의 VoD 총 구매금액(원), 해당 월 넷플릭스 구독여부(Y = 1/N = 0) 데이터로 총 29개이다. 단, 넷플릭스를 시청하는 세대 구성원의 수를 명확하게 알고, IPTV와 넷플릭스를 동시간에 시청하는 경우를 제외하기 위해 단일 IPTV를 사용하는 1인 가구로 한정하였다. 또한, 선형회귀모델(linear regression model), 랜덤 포레스트(random forest)를 통해 유의미한 변수를 선정하고, 유선방송 사업자의 관심 변수를 추가하므로써 최종 변수는 9개이고 총 관측치의 수는 15,208개이다. Table 1에는 매칭 전후에 따른 변수들의 요약 통계량을 나타낸다.

4.2. 변수 정의

해당 데이터를 통하여 넷플릭스 구독에 따른 유료방송 시청시간 변화를 확인하기 위해 처리군(treatment group)과 대조군(control group)을 구분하는 기준은 t시점에서의 넷플릭스 구독여부이다. 시점 $t-1$에는 넷플릭스를 시청하지 않았으나, t시점에서부터 넷플릭스 시청이 발생한 회선을 구독집단(treatment group), 시점 $t-1$부터 t시점까지 넷플릭스를 시청하지 않은 집단을 대조군(control group)으로 정의하였다. 이해의 편의를

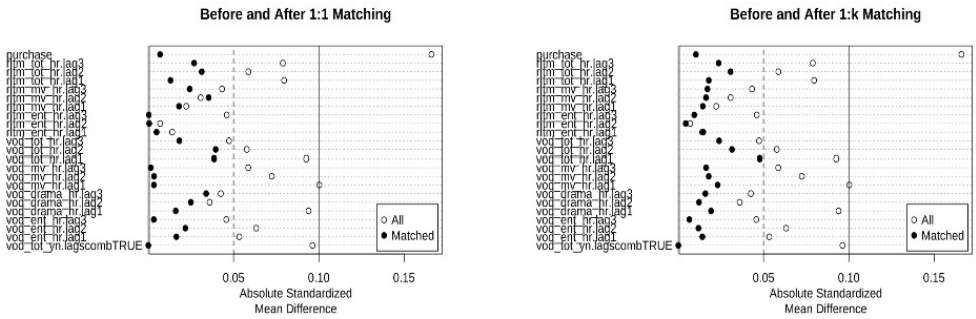

Figure 1: *Covariate balance before and after matching.*

위해 처리군 대신 구독집단이라는 용어를 사용한다. 더 나아가 넷플릭스를 꾸준히 이용하는 경우와 한 번만 이용하는 경우 즉, 넷플릭스 시청강도에 따라 유료방송 시청시간 변화에도 차이가 있는지를 살펴보고자 구독집단과 대조군을 상세히 분류하였다. 우선, lag(L)는 3, lead(F)는 2로 설정하여, 구독집단의 경우, t시점에만 확인하여 넷플릭스를 시청한 집단을 T_1, t시점부터 시점 $t+1$까지 연달아 시청한 집단을 T_2, 시점 $t+2$까지 세 달 연속으로 시청한 집단을 T_3 라고 하였다. $T_3 \subset T_2 \subset T_1$이다. 대조군은 시점 $t-1$부터 t시점까지 넷플릭스를 시청하지 않은 집단을 C_1, 시점 $t+1$까지 시청하지 않은 집단을 C_2, 시점 $t+2$까지 시청하지 않은 집단을 C_3 라고 하였다. 마찬가지로 $C_3 \subset C_2 \subset C_1$이다.

VoD 시청시간의 경우, 0의 값에 많이 치우쳐진 비대칭(skewed) 데이터 특성을 가지고 있어, VoD 시청여부(Y = 1/N = 0)라는 이진변수(binary variable)를 생성하여 매칭에 활용하였다. 최종적으로 변수 넷플릭스는 구독효과(treatment)에 해당하고, 실시간 전체 시청시간(rltm_tot_hr)과 VoD 전체 시청시간(vod_tot_hr), VoD 시청여부가 반응 변수에 해당한다.

5. 분석 및 결과

5.1. 패널데이터 매칭

3장에서 소개한 비복원추출을 이용한 최적매칭방법론을 4장의 산학협력 자료에 적용하고자 한다. 이때 매칭방밥론은 1:1 매칭과 1:k 매칭을 적용하여 분석한다. 구독집단은 $|T_1|$ =1,167명, $|T_3|$ = 928명이며, 대조군은 $|C_1|$ = 14,041명, $|C_3|$ = 12,668명이다. 넷플릭스 시청기간에 따른 인과효과를 구하기 위해서는 T_1은 C_1과 T_3은 C_3와 비교한다. 넷플릭스 구독집단이 t 시점 이후 넷플릭스를 시점 $t+1$ 또는 시점 $t+2$까지 연속으로 구독하는 경우, 이 기간의 IPTV 시청변화가 넷플릭스 구독 때문이라는 것을 설명하기 위해서는 동일 기간 동안 대조군의 넷플릭스 구독여부가 통제되어야 하기 때문이다. 이후 구독집단의 유닛 i의 t시점 이전의 처리변수 기록과 동일한 대조군 유닛들의 집단 $\mathcal{M}_{it}$을 선택한다. 마지막으로, 모든 변수들에 대해 매칭을 적용하며, 월별 VoD 시청여부 변수의 경우 시점 $t-3$부터 시점 $t-1$동안 VoD 시청한 적이 있는지 이진 변수로(Y = 1/N = 0)로 재정의하여 exact matching을 사용하였다.

매칭은 PSCM (propensity score caliper matching)의 제한폭(caliper, δ)을 0.2*표준편차(standard deviation; SD)로 설정하고, 순위 기반 마할라노비스 거리와 결합하여 사용하였다. 1:k 매칭에서는 $k = 3, 5, 8, 10$ 에 대해 진행하였고, 매칭 전후에 따른 변수 균형(covariate balance)이 가장 좋은 k = 5를 선택하였다. 이 때 균형을

Table 3: DID estimation of 1:1 matching

	Outcome	Result	Lead 0 (t)	Lead 1 ($t+1$)	Lead 2 ($t+2$)
T_1 & C_1	RLTM_TOT_HR	Estimation (SD)	−2.763 (2.498)	15.873 (3.015)	−2.148 (3.323)
		95% CI	(−7.662, 2.136)	(9.961, 21.785)	(−8.664, 4.367)
		p-value	0.269	0	0.518
	VOD_TOT_HR	Estimation (SD)	−3.451 (0.853)	−3.931 (0.881)	−4.196 (0.866)
		95% CI	(−5.123, −1.779)	(−5.659, −2.202)	(−5.894, −2.498)
		p-value	0	0	0
	VOD_WAT_YN	Estimation (SD)	−0.013 (0.015)	−0.067 (0.016)	−0.048 (0.017)
		95% CI	(−0.042, 0.016)	(−0.099, −0.035)	(−0.081, −0.015)
		p-value	0.379	0	0.004
T_3 & C_3	RLTM_TOT_HR	Estimation (SD)	−2.126 (2.814)	19.773 (3.419)	−1.815 (3.651)
		95% CI	(−7.646, 3.394)	(13.067, 26.480)	(−8.976, 5.346)
		p-value	0.45	0	0.619
	VOD_TOT_HR	Estimation (SD)	−2.863 (0.975)	−4.554 (1.092)	−4.613 (1.060)
		95% CI	(−4.775, −0.951)	(−6.696, −2.413)	(−6.691, −2.534)
		p-value	0.003	0	0
	VOD_WAT_YN	Estimation (SD)	−0.025 (0.016)	−0.084 (0.018)	−0.081 (0.018)
		95% CI	(−0.056, 0.006)	(−0.120, −0.048)	(−0.117, −0.045)
		p-value	0.118	0	0

맞추는 것이 중요한 이유는 균형이 맞지 않았을 때에는 두 집단 사이에 차이로 인한 편차가 생기게 되며, 이는 인과추론 문제에서는 가장 피해야할 문제이다. 더 많은 데이터를 사용하면 분산이 줄어들지만, 편차가 더 생기게 되어 흔히 말하는 편향-분산 트레이드오프(bias-variance trade-off) 현상이 발생한다. Figure 1에서 매칭 전과 비교하였을 때, 매칭 후 모든 변수의 표준화된 평균차의 절댓값(absolute standardized mean difference)이 0.05미만 값을 가져 일반적으로 생각하는 임계값(threshold) 0.1 보다 작다. 이 경우에는 만족할만한 매칭결과를 얻었다고 생각한고 분석단계로 넘어간다. 관련 논의는 Stuart (2010)을 참조한다. 특히, 최근 3개월동안의 VoD 구매이력(purchase)와 VoD 시청시간 변수들의 균형이 많이 개선되었다.

5.2. 인과효과분석

인과효과는 식 (2.11)을 이용하여 ATT의 DID추정치를 추정하였고, 1:1매칭과 1:k매칭에 따른 분석 결과는 각각 Table 3, 4에 정리하였다. Table 3에서 크게 T_1과 C_1의 비교와 T_3와 C_3의 비교 두 부분으로 이루어진다. 각각의 비교는 3가지 결과변수에 대해서 살펴보았고, 시간의 흐름에 따른 인과효과를 보기 위해 Lead 0, 1, 2 세 시점에서 결과를 분석하였다.

먼저 T_1과 C_1의 비교를 해보면, Lead 1의 실시간 전체 시청시간(rltm_tot_hr)을 제외한 모든 결과에서 추정값(estimation)이 음수값을 가짐을 확인할 수 있다. 즉, 넷플릭스 구독집단의 실시간 시청시간, VoD 시청시간 그리고 Vod 시청여부 모두 줄어드는 현상이 발견된다. 결과변수별로 살펴보자면, 실시간 시청은 Lead 0과 2에서 통계적으로 유의미하지는 않지만 −2.7시간과 −2.1시간으로 나타난다. 특이한 패턴으로 Lead 1(2021년 2월)에서 실시간 시청시간이 약 15.9시간정도 구독집단에서 증가하였고 이는 표준편차 3시간에 비해 큰 값이다. 통계적으로도 유의미함을 확인할 수 있었다. 2021년 2월에 제공되는 컨텐츠를 살펴보았을 때, 미스트롯2(TV조선, 20.12.17–21.03.04)와 같이 실시간으로 시청하는 경향이 높은 프로그램의 후반부와 겹침을 확인하였다. 또한 해당 2021년 2월에는 설 연휴가 있어, 구독집단의 실시간 시청시간이 늘어났다는 추측도 가능하지만 본 데이터를 통해서는 확인이 어렵다.

Table 4: DID estimation of 1:5 matching

	Outcome	Result	Lead 0 (t)	Lead 1 ($t+1$)	Lead 2 ($t+2$)
T_1 & C_1	RLTM_TOT_HR	Estimation (SD)	−2.016 (1.99)	16.773 (2.636)	−2.884 (2.801)
		95% CI	(−5.918, 1.885)	(11.604, 21.941)	(−8.374, 2.606)
		p-value	0.311	0	0.303
	VOD_TOT_HR	Estimation (SD)	−3.374 (0.754)	−4.045 (0.8)	−4.574 (0.743)
		95% CI	(−4.852, −1.897)	(−5.612, −2.477)	(−6.03, −3.118)
		p-value	0	0	0
	VOD_WAT_YN	Estimation (SD)	−0.009 (0.011)	−0.061 (0.013)	−0.046 (0.013)
		95% CI	(−0.03, 0.013)	(−0.086, −0.036)	(−0.072, −0.02)
		p-value	0.439	0	0
T_3 & C_3	RLTM_TOT_HR	Estimation	−2.7 (2.305)	17.274 (2.981)	−3.882 (3.168)
		95% CI	(−7.218, 1.818)	(11.43, 23.119)	(−10.092, 2.328)
		p-value	0.241	0	0.22
	VOD_TOT_HR	Estimation (SD)	−2.829 (0.804)	−4.255 (0.891)	−5.072 (0.839)
		95% CI	(−4.405, −1.252)	(−6.002, −2.508)	(−6.716, −3.427)
		p-value	0	0	0
	VOD_WAT_YN	Estimation (SD)	−0.019 (0.013)	−0.077 (0.014)	−0.074 (0.015)
		95% CI	(−0.044, 0.006)	(−0.106, −0.049)	(−0.103, −0.045)
		p-value	0.129	0	0

다른 결과변수인 VoD 시청시간과 시청여부에서는 비슷한 패턴이 관측이 된다. VoD 시청시간에서는 Lead 0에서는 3.5시간 감소하였지만, 시간이 지남에 따라 감소폭은 조금씩 커져 Lead 1에서는 3.9시간, Lead 2에서는 4.2시간 감소하였다. 이러한 패턴은 넷플릭스 구독으로 인해 더 많은 VoD 시청시간을 빼앗김을 의미한다. VoD 시청여부를 보았을 때도 비슷한 패턴이 나타나지만 Lead 2에서는 Lead 1과 비교하여 감소폭이 줄어들었음을 확인할 수 있다. 이는 이용률에서는 시간이 지나면 안정화되는 패턴이 발생한다고 볼 수 있다.

T_1 vs. C_1에 비교하여, T_3 vs. C_3는 더 오랜기간동안 구독여부의 차이가 나기 때문에 구독집단과 대조군 사이에 더 큰 차이가 발생할 것으로 예측할 수 있다. Lead 1과 2에서는 전반적으로 예측한 결과에 부합한다. 예를 들어, VoD 시청시간은 Lead 2에서는 4.6시간 감소한 것으로 4.2시간과 비교하여 더 큰 감소를 보여준다. 그리고 Lead 1의 값과 Lead 2의 값은 크게 차이를 나타내지 않는다.

Table 4에서는 1:5매칭의 결과를 보여준다. Table 3에서 본 결과값과 크게 차이가 나지 않는 것을 확인할 수 있다. 차이점으로는 5배 많은 수의 대조군을 활용하였기 때문에 인과효과 추정치의 표준오차가 좀 더 작은 값을 가짐을 볼 수 있다. 하지만 인과추론 결과에 대한 질적인 차이는 없다.

Table 3, 4는 시점 $t-1$과 t시점에서의 처리이력을 exact matching한 결과이며, 식 (2.2)에서와 같이 $\tau(F)$를 추정하는 방식으로 얻은 결과이다. 본 논문에서는 $L=3$으로 선택하였기 때문에 $\tau(F, L=3)$에 대한 추정도 가능하다. Table 2에서 처리이력(treatment history)을 확인하면 $L=2,3$인 경우에 어떤 값을 가지냐에 따라 네 가지 다른 부분집합을 생각할 수 있다. 이때 매칭과정에서 같은 부분집합끼리 매칭이 되게 한다면, 시점 $t-3$부터 t 시점까지의 처리이력을 exact matching한 결과이다. Table 5에서는 1:1매칭을 이용한 결과를 보여준다. Table 3과 5 값을 비교해보면 전체 처리이력을 exact matching했을 때 표준편차 값이 전반적으로 줄어드는 것을 볼 수 있다.

위에서 설명한 결과는 Figure 2에 같이 표현되었다. 각 작은그림에서 6개의 추정값과 95% 신뢰구간을 확인할 수 있다.

Table 5: DID estimation of matching exactly on treatment history

	Outcome	Result	Lead 0 (t)	Lead 1 ($t+1$)	Lead 2 ($t+2$)
T_1 & C_1	RLTM_TOT_HR	Estimation	−2.694 (2.45)	15.14 (2.987)	−1.892 (3.266)
		95% CI	(−7.499, 2.11)	(9.282, 20.998)	(−8.296, 4.511)
		p-value	0.272	0	0.562
	VOD_TOT_HR	Estimation (SD)	−2.809 (0.851)	−3.28 (0.897)	−3.707 (0.826)
		95% CI	(−4.477, −1.141)	(−5.039, −1.521)	(−5.328, −2.087)
		p-value	0.001	0	0
	VOD_WAT_YN	Estimation (SD)	−0.005 (0.015)	−0.064 (0.017)	−0.058 (0.017)
		95% CI	(−0.035, 0.025)	(−0.096, −0.031)	(−0.091, −0.024)
		p-value	0.733	0	0.001
T_3 & C_3	RLTM_TOT_HR	Estimation	−1.777 (2.786)	19.192 (3.406)	−2.867 (3.619)
		95% CI	(−7.242, 3.688)	(12.512, 25.873)	(−9.966, 4.231)
		p-value	0.524	0	0.428
	VOD_TOT_HR	Estimation (SD)	−2.853 (0.949)	−4.551 (1.056)	−4.977 (1.026)
		95% CI	(−4.714, −0.992)	(−6.622, −2.48)	(−6.989, −2.965)
		p-value	0.003	0	0
	VOD_WAT_YN	Estimation (SD)	−0.014 (0.016)	−0.077 (0.018)	−0.088 (0.019)
		95% CI	(−0.046, 0.018)	(−0.112, −0.041)	(−0.125, −0.052)
		p-value	0.393	0	0

6. 결론 및 향후 연구

본 연구는 사람들의 미디어 소비 시간은 제한적이기 때문에 OTT를 시청하면 기존 유료방송 시청시간이 줄어들 것이라는 유료방송 사업자의 고민으로부터 시작되어, 실제 패널 데이터를 활용하여 인과효과를 추정하였다. 넷플릭스 구독하는 집단과 대조군을 이용하여 비교를 진행하였으며, 실시간 TV시청과 VoD시청 모두에서 구독집단의 시청시간이 줄어듬을 확인하였다. VoD시청여부 확인을 통해 약 5-8%정도 시청을 하지 않는 비율이 증가함을 확인하였고 이는 VoD시청시간에도 영향을 미쳤다. Lead 0-2까지 확인한 결과 구독집단에 대한 인과효과는 일시적이 아니며 적어도 2개월 후까지 효과가 지속되었음을 확인할 수 있다. 본 논문에서 찾아낸 특이한 패턴으로 Lead 1 시점에서 실시간 시청이 오히려 구독집단에서 증가함을 볼 수 있었다. 몇 가지 가설은 있지만, 이 현상을 이해하기 위해 새로운 연구가 필요할 것으로 보인다.

사용된 패널데이터는 2020년 10월부터 2021년 3월까지의 데이터이며, 이를 이용하여 시점 t = 2021년 1월에 대한 결과를 분석하였다. 2021년 1–3월까지 넷플릭스 구독의 영향을 분석할 수 있었지만, 넷플릭스 구독은 시간 트렌드에 크게 영향을 받는다는 사실이 같이 고려되지 않았다. 예를 들어, Lead 1이었던 2021년 2월에 실시간 시청에서 큰 인과효과가 나타나는 것은 구독집단의 전반적인 Lead 1에 대한 해석이 아니며 2021년 1월 구독집단에서만 나타나는 효과로 제한된 해석만 가능하다. 어느 특정한 달에 사람들의 관심을 많이 받는 작품이 출시되는 경우에 다른 달과 비교하여 더 많은 사람들이 구독을 진행할 것이며 이는 해당하는 달의 구독집단이 다른 달의 구독집단과 달라짐을 의미한다. 또한 Lead에 대한 결과 해석은 바뀐 처리군에 대한 결과이기 때문에 해석을 더욱 어렵게 한다. 이를 해결하기 위해서는 시간 트렌드를 확인할 수 있도록 더 많은 시점 t를 설정하는 여러 개의 패널데이터를 사용하는 방법이 고려될 수 있다.

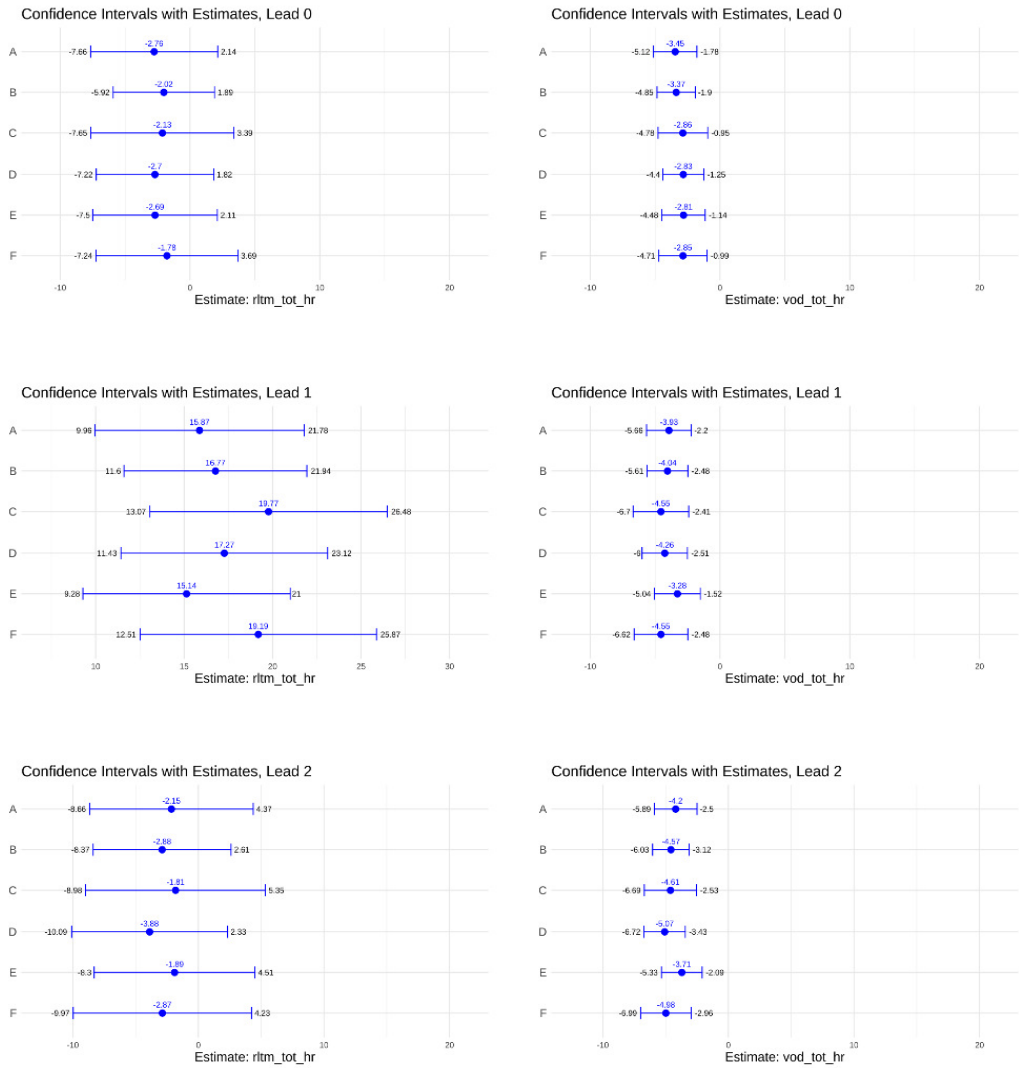

Figure 2: *95% confidence intervals for the estimates. (A) T_1 vs C_1 using 1:1 matching, (B) T_1 vs C_1 using 1:k matching, (C) T_3 vs C_3 using 1:1 matching, (D) T_3 vs C_3 using 1:k matching, (E) T_1 vs C_1 using exact matching on treatment history, and (F) T_3 vs C_3 using exact matching on treatment history.*

References

Abadie A (2005). Semiparametric difference-in-differences estimators, *The Review of Economic Studies*, **72**, 1–19.

Ding P (2024). *A First Course in Causal Inference*, CRC Press, Boca Raton, FL.

Hansen BB (2004). Full matching in an observational study of coaching for the SAT, *Journal of the American Statistical Association*, **99**, 609–618.

Holland PW (1986). Statistics and causal inference, *Journal of the American Statistical Association*, **81**, 945–960.

Imai K and Kim IS (2019). When should we use unit fixed effects regression models for causal inference with longitudinal data?, *American Journal of Political Science*, **63**, 467–490.

Imai K, Kim IS, and Wang EH (2023). Matching methods for causal inference with time-series cross-sectional data, *American Journal of Political Science*, **67**, 587–605.

Robins JM, Hernan MA, and Brumback B (2000). Marginal structural models and causal inference in epidemiology, *Epidemiology*, **11**, 550–560.

Rosenbaum PR and Rubin DB (1983). The central role of the propensity score in observational studies for causal effects, *Biometrika*, **70**, 41–55.

Rosenbaum PR (2010). *Design of Observational Studies*, Springer, New York.

Rosenbaum PR (2020). Modern algorithms for matching in observational studies, *Annual Review of Statistics and Its Application*, **7**, 143–176.

Rosenbaum PR and Zubizarreta JR (2023). Optimization techniques in multivariate matching. In *Handbook of Matching and Weighting Adjustments for Causal Inference* (1st ed), (pp. 63–86), Chapman and Hall/CRC, United Kingdom.

Stuart EA (2010). Matching methods for causal inference: A review and a look forward, *Statistical Science*, **25**, 1–21.

Yu R and Rosenbaum PR (2022). Graded matching for large observational studies, *Journal of Computational and Graphical Statistics*, **31**, 1406–1415.

Yu R, Silber JH, and Rosenbaum PR (2020). Matching methods for observational studies derived from large administrative databases, *Statistical Science*, **35**, 338–355.

Zubizarreta JR, Paredes RD, and Rosenbaum PR (2014). Matching for balance, pairing for heterogeneity in an observational study of the effectiveness of for-profit and not-for-profit high schools in Chile, *The Annals of Applied Statistics*, **8**, 204–231.

Received August 01, 2024; Revised August 14, 2024; Accepted August 15, 2024

스트리밍 서비스 사용이 IPTV 시청에 미치는 인과적 영향 분석

정다해[a], 윤용호[b], 이권상[1,b]

[a]성균관대학교 통계학과; [b]서울대학교 통계학과

요 약

현대 사회에서는 스트리밍 서비스의 급속한 성장이 전통적인 TV 시청 방식에 큰 변화를 가져왔다. 본 연구는 스트리밍 서비스 사용이 IPTV 시청에 미치는 인과적 영향을 분석하는 것을 목적으로 한다. 이를 위해, 스트리밍 서비스를 사용하는 사용자와 사용하지 않는 사용자를 비교하였으며, 다양한 잠재적 교란변수를 통제하였다. 특히 유사한 특성을 가진 그룹을 비교하기 위해 여러 매칭 기법을 활용한 인과추론 매칭 방법을 적용하였다. 또한, 인과적 효과의 통계적 유의성을 평가하기 위해 매칭으로부터 얻어진 가중치를 활용하는 회귀 분석 방법을 사용하였다. 연구 결과, 스트리밍 서비스 사용이 IPTV 시청 방식에 유의미한 영향을 미치는 것으로 나타났다. 이러한 결과는 콘텐츠 제공자, 방송사, 광고주들이 시청자의 행동 패턴을 이해하고, 이에 맞춘 전략적 결정을 내리는 데 중요한 정보를 제공한다. 본 연구는 스트리밍 서비스와 전통적인 TV 시청 간의 관계에 대한 새로운 통찰을 제공하며, 향후 관련 연구의 기초 자료로 활용될 수 있을 것이다.

주요용어: 인과추론, IPTV, 매칭, 넷플릭스, 패널데이터

본 연구는 한국연구재단의 지원 (NRF-2021R1C1C1012750)과 서울대학교 신임교수 연구정착금으로 지원되는 연구비에 의하여 수행되었음.

[1]교신저자: (08826), 서울특별시 관악구 관악로 1, 서울대학교 통계학과. E-mail: kwonsanglee@snu.ac.kr