The Role of the Propensity Score in Estimating Dose-Response Functions [Imbens, 1999]

Jai-Hua Kevin Yen

Department of Biostatistics
University at Buffalo

November 17, 2025

Table of Contents

Introduction

Introduction

- **Aim:** Extend propensity score approach for multi-valued treatments while in [Rosenbaum and Rubin, 1983] only can deal with two-valued treatments.
- **Proposed Method:** [Imbens, 1999] proposed General Propensity Score (GPS) approach for multi-valued treatments.
  - Weak Unconfoundness and liked it with the concept missing at random (MAR) [Rubin, 1976, Little and Rubin, 2019].
  - Provide a roadmap on how to use GPS (3-step approach) and gave a graphical diagonosis method to check overlaps.

Basic Setup

Notations

- $Y(t)$ is the potential outcome for treatment t.
- Y is the observed outcome.
- $\mathbf{X}$ is the vector of pre-treatment covariates.
- $D(t) = \mathbb{I}(T = t)$ is an indicator for receiving treatment t.

Comparison of Unconfoundness

Definition (Strong Unconfoundedness)

Assignment to treatment T is strongly unconfounded, given pre-treatment variables $\mathbf{X}$, if

$$T \perp\!\!\!\perp \{Y(t)\}_{t \in \mathcal{T}} \mid \mathbf{X}.$$

Definition (Weak Unconfoundedness)

Assignment to treatment T is weakly unconfounded, given pre-treatment variables $\mathbf{X}$, if

$$D(t) \perp\!\!\!\perp Y(t) \mid \mathbf{X}.$$

Notes on Weak Unconfoundness

- Under weak unconfoundness, the decision to take dose t $(D(t) = 1)$ is independent of the outcome at dose t.
- To estimate the average outcome for dose t, only need the assumption about the selection process for that dose. No need to consider the relationship between receiving dose t and potential outcome for receiving dose s $(t \neq s)$.
- **Link to MAR:** the probability of data being missing $(D(t) = 0)$ is independent of the value of the missing variable $(Y(t))$, conditional on observed covariates $(\mathbf{X})$. i.e. $D(t) \perp\!\!\!\perp Y(t) \mid \mathbf{X}$.

Propensity Score

Definition

The propensity score is the conditional probability of receiving the treatment (i.e., $T = 1$) given the pre-treatment variables :

$$e(x) \equiv \Pr(T = 1 \mid \mathbf{X} = x).$$

Definition

The Generalized Propensity Score (GPS) is the conditional probability of receiving a *particular level of the treatment* t given the pre-treatment variables :

$$r(t, x) \equiv \Pr(T = t \mid \mathbf{X} = x).$$

Balancing Properties

Lemma

The propensity score is a "balancing score": the pre-treatment variables $\mathbf{X}$ *are independent of treatment assignment* T*, conditional on the propensity score :*

$$T \perp\!\!\!\perp \mathbf{X} \mid e(\mathbf{X}).$$

Lemma

The GPS satisfies a "local" balancing property. For any treatment level $t \in \mathcal{T}$*, the indicator for receiving that treatment,* $D(t)$*, is independent of the covariates* $\mathbf{X}$ *conditional on the score for that level,* $r(t, \mathbf{X})$ *:*

$$D(t) \perp\!\!\!\perp \mathbf{X} \mid r(t, \mathbf{X}).$$

Estimating Potential Outcomes using Propensity Scores

Theorem

Suppose assignment to a binary treatment is weakly unconfounded. Then

- $\mu(t,e) = \mathbb{E}[Y(t)|e(X) = e] = \mathbb{E}[Y(t)|e(X) = e, T = t]$,
- $\mathbb{E}(Y(t)) = \mathbb{E}[\mathbb{E}[\mu(t,e)]]$,

for all $t \in \mathcal{T}$.

Theorem

Suppose assignment to treatment t *is weakly unconfounded given pre-treatment variables* X. *Then*

- $\beta(t,r) = \mathbb{E}[Y(t)|r(t,X) = r] = \mathbb{E}[Y|T = t, r(T,X) = r]$,
- $\mathbb{E}[Y(t)] = \mathbb{E}(\beta(t, r(t,X)))$,

for all $t \in \mathcal{T}$.

Inverse Probability Weighting (IPW)

Theorem

Given weak unconfoundedness, the average potential outcomes can be estimated using weighting by the inverse of the probability of receiving the treatment actually received:

$$\mathbb{E}\left[\frac{D(1)\cdot Y}{e(\mathbf{X})}\right] = \mathbb{E}[Y(1)] \quad \textit{and} \quad \mathbb{E}\left[\frac{D(0)\cdot Y}{1-e(\mathbf{X})}\right] = \mathbb{E}[Y(0)].$$

Theorem

Suppose assignment is weakly unconfounded. Then for any $t \in \mathcal{T}$, the expected value of the observed outcome, weighted by the inverse of the GPS for the treatment actually received, is equal to the average potential outcome $\mathbb{E}[Y(t)]$:

$$\mathbb{E}\left[\frac{D(t)\cdot Y}{r(t,\mathbf{X})}\right] = \mathbb{E}[Y(t)].$$

Why GPS is Different?

Valid Causal Comparison in Binary Case

- The conditioning sets are identical (individuals in $\{x \mid r(1,x) = e\}$ and $\{x \mid r(0,x) = 1 - e\}$ are the same).
- $\mathbb{E}[Y(1)]$ and $\mathbb{E}[Y(0)]$ are estimated from the same group of units, then direct comparison is valid.
- For example,

$$\begin{aligned}\mathbb{E}[Y(1) \mid e(x) = 0.4] - \mathbb{E}[Y(0) \mid e(x) = 0.4] \\ = \mathbb{E}[Y(1) - Y(0) \mid e(x) = 0.4]\end{aligned}$$

The individuals with $e(x) = 0.4$ are the same, so the comparison is valid.

Invalid Causal Comparison in Multi-value Treatments

- For example, let $t =$ "high dose" and $s =$ "low dose". For people who have a **high probability** of receiving t is likely very **sick or old** than people who have a **high probability** of receive s. Comparing their outcome is **meaningless**.
- If we want to compute

$$E[Y(t)|r(t,X) = 0.8] - E[Y(s)|r(s,X) = 0.8],$$

then it is obvious the group of people in $\{x \mid r(t,x) = 0.8\}$ and $\{x \mid r(s,x) = 0.8\}$ are not the same.

- Even if we can compute that in using

$$\begin{aligned}\mathbb{E}[Y(t) \mid T = t, r(t,X), r(s,X)] - \mathbb{E}[Y(s) \mid T = t, r(t,X), r(s,X)] \\ = \mathbb{E}[Y(t) - Y(s) \mid T = t, r(t,X), r(s,X)],\end{aligned}$$

conditioning on additional variable is exactly what the propensity score approach attempts to avoid.

A Practical 3-Step Guide for Implementation

Step 1: Estimate the GPS, $r(t, x)$

The first step is to estimate $r(t, x) = \Pr(T = t \mid \mathbf{X} = x)$ for all treatment levels $t \in \mathcal{T}$. The choice of statistical model depends on the nature of the treatment T:

- **Qualitatively Distinct / Unordered Treatments:** If T represents distinct choices (e.g., surgery, drug treatment, no treatment), one can use standard discrete response models like a multinomial logit or nested logit model.
- **Ordered Treatments (Doses):** If T represents ordered levels (e.g., 0mg, 50mg, 100mg of a drug), an ordered logit/probit model could be used.

Step 2: Estimate the Conditional Expectation, $\beta(t, r)$

- The second step is to estimate the conditional expectation function $\beta(t, r) = \mathbb{E}[Y \mid T = t, r(t, \mathbf{X}) = r]$.
- $\beta(t, r)$ can be estimated by regressing the observed outcome Y on the observed treatment T and the *estimated GPS for the treatment received*, $\hat{r}(T, \mathbf{X})$.

Step 3: Average to Estimate the Dose-Response Function

- The final step is to estimate the $\mathbb{E}(Y(t))$ by applying

$$\mathbb{E}(Y(t)) = \mathbb{E}[\beta(t, r(t, \mathbf{X}))]$$

.

- The estimate for $\mathbb{E}(Y(t))$ is the average of these N predictions over the entire sample:

$$\hat{\mathbb{E}}(Y(t)) = \frac{1}{N} \sum_{i=1}^{N} \hat{\beta}(t, \hat{r}(t, \mathbf{X}_i))$$

Diagnosis of Overlap

Practical Diagnosis of Overlap

- The two distributions $f(e(X)|T=1)$ and $f(e(X)|T=0)$ should be sufficient overlap, If the histograms of these two distributions do not substantially overlap, it means there are no comparable units, and causal inference is unreliable.
- In the multi-valued case, this check is more complex. We must have sufficient overlap **for each treatment level** t that we wish to study.
- In applications it may be the case that there is sufficient overlap for pairs of values of the treatment for ranges of the pre-treatment variables, but not for others.

Practical Diagnosis of Overlap

This procedure must be repeated *for each treatment level* $t \in \mathcal{T}$. For a given level t:

1. **Create two groups of units:**
   - **Group 1:** Units that *received* treatment t (i.e., $T_i = t$).
   - **Group 2:** Units that *did not receive* treatment t (i.e., $T_i \neq t$).
2. **Calculate one score for all units:** For *every unit* in the sample (in both groups), calculate their predicted probability of receiving treatment t: $\hat{r}(t, \mathbf{X}_i)$.
3. **Compare the distributions:** Create two histograms or density plots and overlay them:
   - **Plot 1:** The distribution of $\hat{r}(t, \mathbf{X})$ for Group 1: $f(\hat{r}(t, \mathbf{X}) \mid T = t)$.
   - **Plot 2:** The distribution of $\hat{r}(t, \mathbf{X})$ for Group 2: $f(\hat{r}(t, \mathbf{X}) \mid T \neq t)$.

Summary

Binary Propensity Score vs. Generalized Propensity Score

Feature	Binary Propensity Score	Generalized Propensity Score
Definition	$e(x) = \Pr(T = 1 \mid \mathbf{X} = x)$	$r(t, x) = \Pr(T = t \mid \mathbf{X} = x)$
Structure	A **single scalar function** per unit.	A **family of functions**, one for each treatment level $t \in \mathcal{T}$.
Balancing	$T \perp\!\!\!\perp \mathbf{X} \mid e(\mathbf{X})$. Conditioning on $e(\mathbf{X})$ balances covariates for *both* groups ($T = 0, T = 1$).	$D(t) \perp\!\!\!\perp \mathbf{X} \mid r(t, \mathbf{X})$. Conditioning on $r(t, \mathbf{X})$ *only* balances for $T = t$ vs. $T \neq t$.
Causal Effects	**Valid.** Can be estimated directly by comparing means.	**Not Valid.** $\beta(t, r) - \beta(s, r)$ has no causal interpretation because the conditioning sets differ.
Primary Use	Stratification, matching, weighting, or adjustment. Can be used for within-strata comparisons.	Adjustment or weighting to estimate the *full population average*.

Summary

In the [Imbens, 1999], the author:

- Introduced a Weaker Assumption
- Proposed the Generalized Propensity Score (GPS)
- Provided a Critical Conceptual Comparison of PS and GPS
- Commented on a Practical Roadmap for Using GPS

References

References I


Imbens, G. W. (1999).
The role of the propensity score in estimating dose-response functions.
NBER Working Paper, (t0237).


Little, R. J. and Rubin, D. B. (2019).
Statistical analysis with missing data.
John Wiley & Sons.


Rosenbaum, P. R. and Rubin, D. B. (1983).
The central role of the propensity score in observational studies for causal effects.
Biometrika, 70(1):41–55.


Rubin, D. B. (1976).
Inference and missing data.
Biometrika, 63(3):581–592.