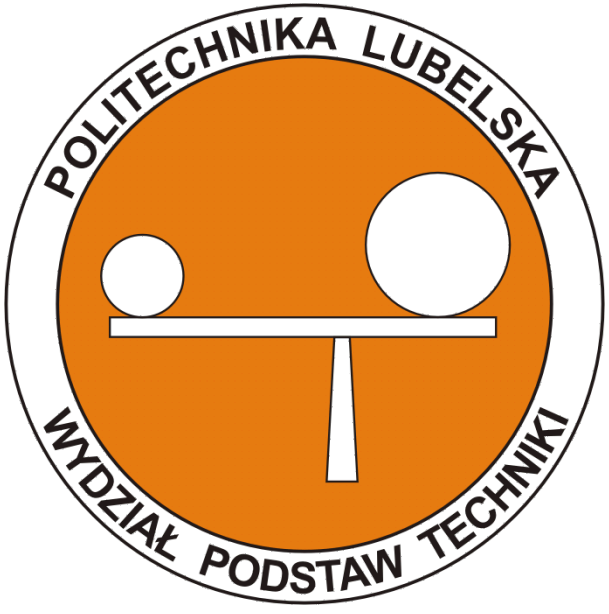
Zastosowanie techniki *Propensity Score Matching* w badaniach ewaluacyjnych

III Międzyregionalna Konferencja Ewaluacyjna



Dariusz Majerek

Katedra Matematyki Stosowanej
Wydział Podstaw Techniki
Politechniki Lubelskiej

Toruń
21-22 czerwca 2016

Model Neymana-Rubin'a[a]

- Y_i - wartość zmiennej wynikowej[b] dla i-tej jednostki;
- T_i będzie zmienną określającą czy dla i-tej jednostki został zastosowany bodziec[c];
  - $T_i = 1$ - grupa interwencji
  - $T_i = 0$ - grupa bez interwencji
- $\mathbf{X}_i$ - wektor kowariantów[d], których wartości nie zależą od tego, do której grupy zostanie zakwalifikowana jednostka.
- $Y_i(1)$ - wartość zmiennej wynikowej dla i-tej jednostki, gdy zastosowano wobec niej interwencję;
- $Y_i(0)$ - wartość zmiennej wynikowej dla i-tej jednostki, gdy nie zastosowano wobec niej interwencji.

[a] znany też jako *Rubin Casual Model*
[b] ang. *outcome*
[c] ang. *treatment*
[d] ang. *covariates*

Wielkość efektu

- Wielkość efektu dla pojedynczej jednostki

$$\tau_i = Y_i(1) - Y_i(0), \quad i = 1, \ldots, N. \tag{1}$$

- Przeciętny efekt oddziaływania bodźca (ATE[a])

$$\tau_{ATE} = \mathrm{E}Y(1) - \mathrm{E}Y(0). \tag{2}$$

- Przeciętny efekt oddziaływania bodźca dla jednostek dla których zastosowano interwencję (ATT[b])

$$\tau_{ATT} = \mathrm{E}[Y(1)|T = 1] - \mathrm{E}[Y(0)|T = 1]. \tag{3}$$

- W przypadku szacowania τ_{ATE} i τ_{ATT} mamy do czynienia albo z brakiem danych[c] albo ze stanem kontrfaktycznym[d].

---

[a] ang. *Average Treatment Effect*

[b] ang. *Average Treatment Effect on the Treated*

[c] żadna jednostka nie może się znaleźć w obu grupach jednocześnie

[d] $\mathrm{E}(Y(0)|T = 1)$

Metoda eksperymentalna

- W przypadku szacowania efektu interwencji najczęściej zaleca się przeprowadzenie badania eksperymentalnego (zwanego dalej eksperymentem).
- Polega ono na losowym doborze obiektów do badania (I zasada randomizacji) oraz losowego doboru obiektów, w stosunku do których zostanie zastosowany bodziec (II zasada randomizacji). Dla każdego obiektu zostanie przypisany kod zastosowanego bodźca[a] oraz wartość odpowiedzi na interwencję.
- Takie podejście pozwala na proste wyznaczenie nieobciążonego i zgodnego estymatora wielkości efektu występującego w populacji, w postaci różnicy średnich próbkowych (obiektów objętych interwencją i nieobjętych).
- W przypadku danych pochodzących z eksperymentu, rozkład cech (zarówno obserwowanych jak i nieobserwowanych), które charakteryzują obiekty jest jednakowy.

[a] zmienna ze skali nominalnej lub porządkowej, najczęściej dychotomiczna

Dane obserwowane

- Badania metodą eksperymentalną mogą nie być wykonalne z kilu powodów: logistycznych, finansowych, czasowych i etycznych.
- Niestety w kontekście większości badań ex-post II zasada randomizacji nie jest spełniona. W tym przypadku wcześniej wspomniany estymator może charakteryzować się dużym obciążeniem[a] oraz brakiem zgodności (Guo et al., 2006).
- Różnice w dystrybucji cech mogą znacząco wpłynąć na wartość zmiennej Y w obu grupach, zatem oszacowanie efektu oddziaływania bodźca za pomocą różnicy średnich obu prób będzie obciążone sporym błędem.
- Wspomniane różnice skutkują tzw. obciążeniem selekcyjnym równym

$$\mathrm{E}[Y(0)|\,T=1]-\mathrm{E}[Y(0)|\,T=0]. \tag{4}$$

[a] ang. *bias*

- Pierwsze próby rozwiązania tego problemu opierały się o modele regresji zmiennej Y względem T oraz $\mathbf{X}$ (Cox, 1956; Robinson, 1973)

$$\hat{Y}_i = \mathrm{E}[Y_i | T_i, \mathbf{X}_i] = \beta_0 + \beta_\tau T_i + \beta_1 \mathbf{X}_i + \varepsilon.$$

  Wówczas β_τ jest oszacowaniem wielkości efektu.
- Istotnie taka metoda zmniejsza obciążenie selekcyjne ale metody dopasowania prób[a] wprowadzone później charakteryzują się jeszcze mniejszym obciążeniem selekcyjnym (Rosenbaum, 2002).
- Jedna z pierwszych form matchingu opierała się na dopasowaniu prób ze względu na wszystkie obserwowane cechy $\mathbf{X}$. Jednostki dobierane są wówczas w ten sposób, aby na poziomie każdej cechy wektora $\mathbf{X}$ próba eksperymentalna i kontrolna były identyczne.
- Istotnie obniża to błąd selekcyjny, natomiast trudno w praktyce zastosować tą metodę gdy dysponujemy niezbyt liczną próbą, a kowariantów w wektorze $\mathbf{X}$ jest wiele lub są mierzone na skali ilorazowej.

[a] ang. *matching*

- Próby dopasowane ze względu na cechy mogą posłużyć jako materiał do wyznaczenia wielkości efektu tylko wtedy, gdy spełnione są następujące założenia[a]:

$$Y(0), Y(1) \perp T|\mathbf{X}, \tag{5}$$

$$0 < \mathrm{P}(T = 1|\mathbf{X}) < 1. \tag{6}$$

- Istotą tych założeń jest zapewnienie, że rozkłady zmiennej zależnej zarówno w grupie eksperymentalnej, jak i kontrolnej są niezależne od bodźca jeśli znane są wartości kowariantów oraz niemożliwe jest aby na podstawie $\mathbf{X}$ jednoznacznie określić, do której grupy należy obiekt.

Propensity Score

Ze względu na praktyczne ograniczenia metody dopasowania prób wg cech, Rubin i Rosenbaum (1983) wprowadzili funkcję *propensity score*

$$e(\mathbf{x}) = \mathrm{P}[T = 1|\mathbf{X} = \mathbf{x}]. \tag{7}$$

[a] w angielskojęzycznych publikacjach założenia te są nazywane *Conditional Independence Assumption* lub *Strongly Ignorable Treatment Assigment*

Funkcja *propensity score* ma następujące własności:

- jest funkcją balansującą[a], czyli

$$T \perp \mathbf{X} \,|\, e(\mathbf{X}), \tag{8}$$

- jeśli spełnione jest założenie warunkowej niezależności (5 i 6) względem $\mathbf{X}$, to również względem $e(\mathbf{X})$, czyli

$$Y(0), Y(1) \perp T \,|\, e(\mathbf{X}) \tag{9}$$

$$0 < \mathrm{P}(T = 1 | e(\mathbf{X})) < 1. \tag{10}$$

Najczęstszą praktyką w wyznaczaniu wartości funkcji $e(\mathbf{x})$ jest zastosowanie regresji logistycznej lub probitowej. W przypadku gdy interwencja może przyjmować więcej niż dwa stany, do estymacji *propensity score* wykorzystamy wielomianową regresję logistyczną lub probitową (Caliendo and Kopeinig, 2008).

[a] ang. *balancing score*

Dobór zmiennych do modelu

- W poszukiwaniu właściwych kowariantów, które potencjalnie mogłyby wpłynąć na spełnienie założenia o niezależności warunkowej, należy zrobić dokładny *desk research*. Przegląd literatury przedmiotowej, wiedza teoretyczna oraz wyniki wcześniejszych badań mogą rzucić światło na kierunki poszukiwań.
- Z pewnością muszą to być zmienne, na które nie miał wpływu fakt przynależności do grupy eksperymentalnej czy kontrolnej. Unikamy również zmiennych jednoznacznie wpływających na otrzymanie lub nie otrzymanie interwencji.
- Heckman (1997), Dehejia i Wahba (1999) sugerują, żeby wziąć wszystkie te zmienne z wektora $\mathbf{X}$, które mają wpływ na zmienną Y i T.
- Bryson (2002), Augutrzky i Schmidt (2001) twierdzą, że włączanie do modelu *propensity score* wszystkich dostępnych zmiennych zwiększa wariancję estymatora efektu. Z drugiej strony Rubin i Thomas (1996) zaznaczają, iż tylko te zmienne, które nie mają związku ze zmienną wynikową lub nie są właściwym kowariantem można odrzucić.

Łączenie metodą najbliższego sąsiada[a]

- Polega na dobraniu do jednostki z grupy eksperymentalnej takiego elementu z puli kontrolnej, aby ich wartości *propensity score* były najbliższe.
- W zależności od wybranego algorytmu łączenia, metoda najbliższego sąsiada będzie prowadzić do różnego stopnia zbalansowania grup eksperymentalnej i kontrolnej. W przypadku metod 'chciwych'[b], czyli 1 : *n* i *bez powtórzeń*, otrzymamy estymator o większej precyzji (mniejsza wariancja) ale większym obciążeniu (Smith and Todd, 2005).
- Dodatkowo w metodzie *bez powtórzeń* zaleca się losowe ustawienie elementów w grupie eksperymentalnej przed wykonaniem łączenia, ponieważ jego jakość zależy od porządku elementów objętych interwencją (Smith, 1997).

[a] ang. Nearest Neighbor

[b] w literaturze obcojęzycznej funkcjonuje pod nazwą *greedy*

Algorytmy łączenia

Technika *Propensity Score Matching* obejmuje wiele algorytmów łączenia prób eksperymentalnej i kontrolnej w oparciu o wartość *propensity score*. Sposoby łączenia:

- Bez powtórzeń - każdy element z puli kontrolnej może być tylko raz użyty w grupie kontrolnej. Ten typ łączenia zwiększa obciążenie estymatora efektu ale zmniejsza jego wariancję.
- Z powtórzeniami - każdy element z puli kontrolnej może być wykorzystany kilka razy jako element w grupie kontrolnej. Ten typ łączenia zmniejsza obciążenie estymatora efektu ale zwiększa jego wariancję.
- 1 : 1 - czyli jednemu elementowi grupy eksperymentalnej jest przyporządkowany jeden ("najlepszy") element puli kontrolnej. Do dopasowanie sprawia, że jest mniejsze obciążenie estymatora efektu ale większa wariancja.
- 1 : n - czyli jednemu elementowi grupy eksperymentalnej jest przyporządkowane n ("najlepszych") elementów puli kontrolnej. Do dopasowanie sprawia, że jest większe obciążenie estymatora efektu ale mniejsza wariancja.

Łączenie z limitem lub promieniem[a]

Zasada działania jest zbliżona do poprzedniej, z tym wyjątkiem, że ustalony jest pewien limit różnicy pomiędzy wartościami *propensity score* obu elementów, którego nie można przekroczyć. Różnica pomiędzy metodą z limitem a metodą z promieniem jest taka, że w tej pierwszej należy ustalić wcześniej liczbę "sąsiadów" z grupy kontrolnej, których będziemy dopasowywać do jednostki z grupy eksperymentalnej. Natomiast w drugiej wszystkie jednostki, które spełniają warunek $|e_1(\mathbf{x}_i) - e_0(\mathbf{x}_j)| < r$ są włączane do puli kontrolnej, przy czym $e_1(\mathbf{x}_i)$ jest wartością *propensity score* dla ustalonej i-tej jednostki z grupy eksperymentalnej, a $e_0(\mathbf{x}_j)$ jest wartością *propensity score* dla dowolnej j-tej jednostki z puli kontrolnej. W przypadku kiedy różnica nie da się znaleźć w puli kontrolnej elementów spełniających powyższe warunki, to taką jednostkę eliminujemy z próby eksperymentalnej (Dehejia and Wahba, 2002). Niedogodności w metodzie z limitem jest konieczność arbitralnego ustalenia limitu.

[a] ang. *Caliper Matching* i *Radius Matching*

Warstwowanie ze względu na *propensity score*[a]

- Metoda ta polega na podziale wartości *propensity score* na rozłączne przedziały (warstwy). Najczęściej polecany jest podział kwantylowy. Rosenbaum i Rubin pokazali, że wówczas redukuje się w ten sposób 90% całkowitego obciążenia selekcyjnego (Rosenbaum and Rubin, 1984).
- Imbens (2004) twierdzi, że jeśli się weźmie warstwy o jednakowej liczebności, to do oszacowania efektu τ_{ATE} użyjemy średniej różnic średnich w poszczególnych warstwach. Natomiast do oszacowania τ_{ATT} zaleca się użycia średniej ważonej różnic średnich w poszczególnych warstwach, z wagami proporcjonalnymi do liczby elementów wystawionych na działanie bodźca w danej warstwie.

[a] ang. *Stratification on the Propensity Score*

Ważona metoda PSM[a]

Podejście wprowadzone przez Rosenbaum'a (1987) jest nieco inne od wcześniejszych, ponieważ nie dokonujemy w niej właściwego *matchingu*, natomiast wszystkie elementy w próbie włączamy do badania z odpowiednią wagą. Jest nią odwrotność prawdopodobieństwa otrzymania bodźca przez jednostkę

$$\omega_i = \frac{T_i}{e_i} + \frac{1 - T_i}{1 - e_i}, \tag{11}$$

gdzie e_i są wartościami *propensity score* dla i-tej jednostki.
Oznacza to, że w ważonej próbie rozkład kowariantów będzie niezależny od bodźca.

[a] znana jako ang. *Inverse Probability of Treatment Weighting*

Regresja z uwzględnieniem PS[a]

- Rozwiązanie to zakłada włączenie do modelu regresji wartości *propensity score*

$$\hat{Y}_i = \mathrm{E}[Y_i | T_i, \mathbf{X}_i] = \beta_0 + \beta_\tau T_i + \beta_1 e_i(\mathbf{X}_i) + \varepsilon. \tag{12}$$

- Metoda ta różni się znacząco od wcześniej wymienionych, podejściem do balansowania prób:
  - podział na planowanie doświadczenia i estymację efektu;
  - w modelach regresyjnych występuje tendencja do modyfikowania wyników w kierunku zakładanego efektu (Rubin, 2001);
  - podejście regresyjne jest dosyć wrażliwe na to czy wartość *propensity score* została właściwie oszacowana.

[a] ang. *Covariate Adjustment with Propensity Score*

Testowanie dopasowania

- Najczęstszym sposobem testowania jakości dopasowania próby kontrolnej do eksperymentalnej jest ocena zbalansowania kowariantów za pomocą standaryzowanych różnic.
- Różnice standaryzowane

$$SB = \frac{\bar{X}(1) - \bar{X}(0)}{\sqrt{0.5 \cdot (S^2_{X(1)} + S^2_{X(0)})}} \cdot 100\% \tag{13}$$

- Ocena zbalansowania dokonywana jest przez porównanie różnic standaryzowanych przed i po *matchingu*.
- Zdarza się, że wspomniana różnica jest też testowana testem istotności różnic. O ile przed dopasowaniem próby różnica może być istotna, to po dopasowaniu już nie powinna.

Przykład

- Dane częściowo pochodzą z badania przeprowadzonego w USA w latach 1975-1980 i dotyczą programu *National Support Work Demonstration*, którego zadaniem było sprawdzenie jaki jest poziom zarobków, po odbyciu stażu w zakładzie pracy chronionej. Badaniem były objęte osoby z problemami: przestępcy, narkomani, kobiety, które dłuższy czas przebywały na zasiłku socjalnym oraz osoby, które porzuciły naukę w wieku 17-20 lat.
- Badanie zostało przeprowadzone metodą eksperymentalną w wynik oszacowanego efektu dla mężczyzn wyniósł 1794$.
- W latach późniejszych dane te zostały uzupełnione przez LaLonde'a (1986) o wyniki dwóch spisów powszechnych *Panel Study of Income Dynamics* i *Current Population Survey*.
- Dane przygotowane przez LaLonde'a zostały wykorzystane do weryfikacji jakości techniki PSM przeprowadzonej w (Dehejia and Wahba, 1999) i (Dehejia and Wahba, 2002).

Model Dehejia i Wahba

- Grupa eksperymentalna składała z 185 mężczyzn objętych interwencją, a pula kontrolna zawierała dane ze spisu powszechnego PSID na temat 2490 mężczyzn w wieku poniżej 55 lat, pracujących.
- Dla obu grup obserwowane były zmienne: wiek (AGE), lata edukacji (EDUC), czy mężczyzna przerwał edukację przed 12 rokiem życia (NODEGREE), czy jest czarnoskóry (BLACK), czy jest latynosem (HISP), czy jest żonaty (MARR), czy był bezrobotny w roku 1974 (U74), czy był bezrobotny w roku 1975 (U75), poziom zarobków w 1974 (RE74), poziom zarobków w 1975 (RE75) , poziom zarobków w 1978 (RE78). Ostatnia zmienna była traktowana jako wynikowa.

Rozpatrywane modele

- regresja wieloraka

$$RE78_i = \beta_0 + \beta_\tau T_i + \beta_1 \mathbf{X}_i, \tag{14}$$

gdzie $\mathbf{X} = AGE, AGE^2, EDUC, NODEGREE, BLACK, HISP, RE74, RE75$

- propensity score

$$e(\mathbf{X}_i) = \frac{e^{\beta_1 \mathbf{X}_i}}{1 + e^{\beta_1 \mathbf{X}_i}}, \tag{15}$$

gdzie $\mathbf{X} = AGE, AGE^2, EDUC, EDUC^2, NODEGREE, BLACK, HISP, MARR, RE74, RE75, RE74^2, RE75^2, U74 * BLACK$

Wyniki

- Oszacowanie efektu interwencji przeprowadzono w oparciu o kilka technik:
  - różnica średnich bez dopasowania;
  - regresję wieloraką;
  - dopasowaniem po cechach;
  - PSM metodą najbliższego sąsiada (1:1) z powtórzeniami;
  - PSM metodą z promieniem (r = 0.001, 0.0001,0.00001);
  - PSM metodą warstwową (6 warstw);
  - regresja wieloraka z PS.
- Zbalansowania grup znacząco się różniły w zależności od zastosowanej metody. Najlepsze zbalansowanie uzyskano przy metodzie łączenia wg cech. Natomiast nieco gorsze zbalansowanie wykazywały techniki PSM.

Wyniki

Model	N_1	N_0	ATT	SE[a]
Surowe średnie	2490	2490	-15205	657
Regresja wieloraka	2490	2490	218	768
Łączenie po cechach	185	185	2037	1731
PSM NN	185	57	1890	1202
PSM r = 0.001	2021	583	1824	1187
PSM r = 0.0001	337	76	1973	1191
PSM r = 0.0001	193	13	1893	1198
PSM warstwowe	1086[b]	1146	1452	1141
Regresja z PS	2490	185	1149	2000

[a] niektóre błędy standardowe były liczone za pomocą techniki bootstrap

[b] Korekta ze wzledu na *common support*

Wady PSM

- W ostatnich latach coraz częściej pojawia się krytyka techniki PSM (King et al., 2011), (King and Nielsen, Working Paper).
- Wskazuje się przykłady zastosowań PSM, w których po jej użyciu balans pomiędzy grupami się nawet pogarsza.
- Zbalansowanie odbywa się tylko po obserwowanych zmiennych, zmienne nieobserwowane nie są balansowane.
- Alternatywą mogą być:
  - *Mahalanobis Distance Matching*;
  - *Coarsened Exact Matching*.

Bibliografia (1)

Donald B. Rubin. Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics*, 6(1):34–58, 1978.

Shenyang Guo, Richard P. Barth, and Claire Gibbons. Propensity score matching strategies for evaluating substance abuse services for child welfare clients. *Children and Youth Services Review*, 28(4):357 – 383, 2006.

David R. Cox. A note on weighted randomization. *The Annals of Mathematical Statistics*, 27(4):1144–1151, 1956.

John Robinson. The analysis of covariance under a randomization model. *Journal of the Royal Statistical Society. Series B (Methodological)*, 35(2): 368–376, 1973.

Paul R. Rosenbaum. Covariance adjustment in randomized experiments and observational studies. *Statistical Science*, 17(3):286–304, 2002.

Bibliografia (2)

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

Marco Caliendo and Sabine Kopeinig. Some practical guidance for the implementation of propensity score matching. *Journal of Economic Surveys*, 22(1):31–72, 2008.

Petra E. Todd James J. Heckman, Hidehiko Ichimura. Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *The Review of Economic Studies*, 64(4):605–654, 1997.

Rajeev H. Dehejia and Sadek Wahba. Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American Statistical Association*, 94(448):1053–1062, 1999.

Alex Bryson. The union membership wage premium: an analysis using propensity score matching. *Discussion Paper*, 2002.

Bibliografia (3)

Boris Augutrzky and Christoph M. Schmidt. The propensity score: a means to an end. *IZA Discussion Paper*, 2001.

Neal Thomas Donald B. Rubin. Matching using estimated propensity scores: Relating theory to practice. *Biometrics*, 52(1):249–264, 1996.

Jeffrey A. Smith and Petra E. Todd. Does matching overcome lalonde's critique of nonexperimental estimators? *Journal of Econometrics*, 125 (1–2):305 – 353, 2005. Experimental and non-experimental evaluation of economic policy and models.

Herbert L. Smith. Matching with multiple controls to estimate treatment effects in observational studies. *Sociological Methodology*, 27(1):325–353, 1997.

Rajeev H. Dehejia and Sadek Wahba. Propensity score matching methods for non-experimental causal studies. *Review of Economics and Statistics*, 84:151–161, 2002.

Bibliografia (4)

Paul R. Rosenbaum and Donald B. Rubin. Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*, 79(387):516–524, 1984.

Guido Imbens. Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics*, 2004.

Paul R. Rosenbaum. Model-based direct adjustment. *Journal of the American Statistical Association*, 82(398):387–394, 1987.

Donald B. Rubin. Using propensity scores to help design observational studies: Application to the tobacco litigation. *Health Services and Outcomes Research Methodology*, 2(3):169–188, 2001.

Robert J. LaLonde. Evaluating the econometric evaluations of training programs with experimental data. *The American Economic Review*, 76(4): 604–620, 1986.

Bibliografia (5)

Gary King, Richard Nielsen, Carter Coberley, James E. Pope, and Aaron Wells. Comparative effectiveness of matching methods for causal inference. 2011.

Gary King and Richard Nielsen. Why propensity scores should not be used for matching. Working Paper.