Autorzy: Dzmitry Fiodarau

Ocena efektu edukacji na wynik testu słownictwa przy użyciu Propensity Score Matching

Warszawa 2025.

1.Wprowadzenie

W badaniach społeczno-ekonomicznych rzadko stosuje się eksperymenty losowe oparte na randomizacji ze względu na trudności logistyczne oraz etyczne. W ich miejsce coraz częściej wykorzystywane są badania obserwacyjne, które pozwalają na analizę rzeczywistych zachowań jednostek, jednak wiążą się z ryzykiem występowania obciążenia selekcyjnego. Obciążenie selekcyjne jest wynikiem systematycznych różnic między grupami analizowanymi, wynikających z nielosowego doboru jednostek. W takich przypadkach grupy poddane analizie mogą nie być porównywalne, co wpływa na jakość i wiarygodność wniosków. [1]

W celu rozwiązania tego problemu coraz częściej stosuje się metody quasi- eksperymentalne, takie jak model Neymana-Rubina oraz metody dopasowania, takie jak Propensity Score Matching. Model Neymana-Rubina umożliwia ocenę przyczynowych efektów działania poprzez porównanie efektów w grupach traktowania i kontrolnej, uwzględniając różnice w ich charakterystykach. Metoda PSM pozwala na redukcję obciążenia selekcyjnego poprzez dopasowanie grup na podstawie prawdopodobieństwa przypisania do traktowania, co zapewnia większą równoważność między grupami.

W niniejszej pracy wykorzystane zostaną dane z General Social Survey , które umożliwiają analizę wpływu różnych czynników, takich jak wykształcenie, wiek, płeć oraz miejsce urodzenia, na znajomość słownictwa w populacji badanej. Analiza, przeprowadzona z wykorzystaniem modelu Neymana-Rubina, odpowiada na pytania badawcze związane z przyczynowymi determinantami znajomości słownictwa w kontekście wybranych zmiennych.

[1] Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu, Taksonomia 24. Klasyfikacja i analiza danych – teoria i zastosowania, red. nauk. Krzysztof Jajuga, Marek Walesiak, Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu, Wrocław 2015, nr 384.

2.Opis problemu badawczego i pytań badawczych

W niniejszej pracy jest badane, jakie czynniki demograficzne i edukacyjne wpływają na poziom znajomości słownictwa w wybranej próbie populacji. Oceniony został wpływ edukacji na wyniki testu słownictwa. Natomiast wpływ wieku, płci oraz miejsca urodzenia na poziom wykształcenia.

Pytania badawcze:

1. Jak poziom wykształcenia wpływa na wynik testu słownictwa?
2. Czy płeć wpływa na poziom wykształcenia?
3. Czy wiek wpływa na poziom wykształcenia?
4. Czy miejsce urodzenia wpływa na poziom wykształcenia?

3.Dobór zbioru danych

Głównym źródłem danych jest General Social Survey (GSS)[2], jedno z najbardziej szanowanych badań społecznych w Stanach Zjednoczonych, prowadzonych przez National Opinion Research Center na Uniwersytecie Chicagowskim. Badanie GSS rozpoczęto w 1972 roku i od tego czasu regularnie zbierane są dane dotyczące wartości, demografii i różnych aspektów społeczno-ekonomicznych. Dane te stanowią cenne źródło do analiz społecznych i ekonomicznych oraz umożliwiają identyfikację trendów społecznych w Stanach Zjednoczonych.

Natomiast dane wykorzystane w niniejszej analizie zostały odnalezione na platformie Kaggle[3], gdzie były one już wstępnie oczyszczone i przekształcone na potrzeby dalszej analizy statystycznej. Oryginalne dane z GSS pochodzą z ankiet i zawierają odpowiedzi respondentów, które następnie zostały odpowiednio przekodowane.

Głównym powodem wyboru tego zbioru danych jest to, że ma on charakter obserwacyjny i obejmuje różne lata, co oznacza, że przypisanie poziomu edukacji(zmienna treatment) nie było randomizowane, a wynikało z naturalnych procesów edukacyjnych i społecznych.

Dane obejmują 28 867 obserwacji i pochodzą z okresu 1978-2016. W podczas dalszej analizy zostaną wykorzystane zmienne:

- rok przeprowadzenia badania
- wiek respondenta w latach
- Grupa wiekowa
- liczba lat edukacji
- kategoria poziomu wykształcenia
- płeć respondenta

[2] https://gss.norc.org/ (dostęp 21.01.2025)
[3] https://www.kaggle.com/datasets/utkarshx27/general-social-survey (dostęp 20.01.2025)

- informacja, czy respondent urodził się w USA
- wynik testu słownictwa, który mierzy liczbę poprawnych odpowiedzi na 10 pytań
- informacja, czy respondent posiada więcej niż 12 lat edukacji

W tabeli 3.1 przedstawione statystyki opisowe zmiennych wykorzystanych w badaniu. Większość z nich charakteryzuje się względnie symetrycznym rozkładem, co sugerują zbliżone wartości średniej i mediany.

Największą asymetrię wykazuje zmienna dotycząca miejsca urodzenia, której skośność wynosi -2,96, co oznacza, że większość wartości jest skupiona na poziomie 1. Co oznacza że większość badanych osób są urodzeni w USA. Wysoka kurtoza tej zmiennej (6,78) potwierdza dużą koncentrację danych wokół jednej wartości.

Największą zmienność obserwujemy w przypadku wieku, co wynika z szerokiego przedziału wiekowego respondentów. Natomiast zmienne liczba lat edukacji i wynik testu słownictwa mają wartości kurtozy bliskie 0, co oznacza rozkład zbliżony do normalnego.

Tabela 3.1 Statystyki opisowe zmiennych

Zmienna	**Średnia**	**Mediana**	**Odchylenie standardowe**	**Wariancja**	**Skośność**	**Kurtoza**
Płeć(1-mężczyźni)	0.433	0	0.49	0.24	0.27	-1.92
Wiek	45.74	43	17.38	302.25	0.45	-0.73
Liczba lat edukacji	13.24	13	3.01	9.04	-0.23	0.86
Wynik testu słownictwa	6	6	2.1	4.42	-0.23	-0.11
Czy respondent urodził się w USA(1-tak)	0.91	1	0.28	0.08	-2.96	6.78

Źródło: Opracowanie własne

W zbiorze danych 57% respondentów to kobiety, a 43% mężczyźni. 91% badanych urodziło się w Stanach Zjednoczonych. Wiek respondentów waha się od 18 do ponad 60 lat, a średni poziom edukacji wynosi 13 lat. Średni wynik w teście słownictwa to 6 na 10, przy czym najczęściej uzyskiwane wyniki to 5 i 6.

W tabelach 3.2 i 3.3 przedstawione rozkłady częstości według grupy wiekowej oraz poziomu edukacji, odpowiednio. Największą grupę stanowią osoby 60+ lat (23,7%) i 30–39 lat (22%), najmniejszą 50–59 lat (15%). Najwięcej respondentów ukończyło 12 lat edukacji (30,2%) lub 13–15 lat (25,4%), a najmniej >16 lat (11,3%). Różnice w strukturze badanej populacji mogą mieć istotny wpływ na dalszą analizę, zwłaszcza w kontekście porównywania wyników testu słownictwa między różnymi grupami wiekowymi i edukacyjnymi.

Tabela 3.2 Rozkłady częstości według grupy wiekowej

Grupa wiekowa	**% Respondentów**
60+	23.7
50-59	22
40-49	20.8
30-39	18.4
18-29	15

Źródło: Opracowanie własne

Tabela 3.3 Rozkłady częstości według poziomu edukacji

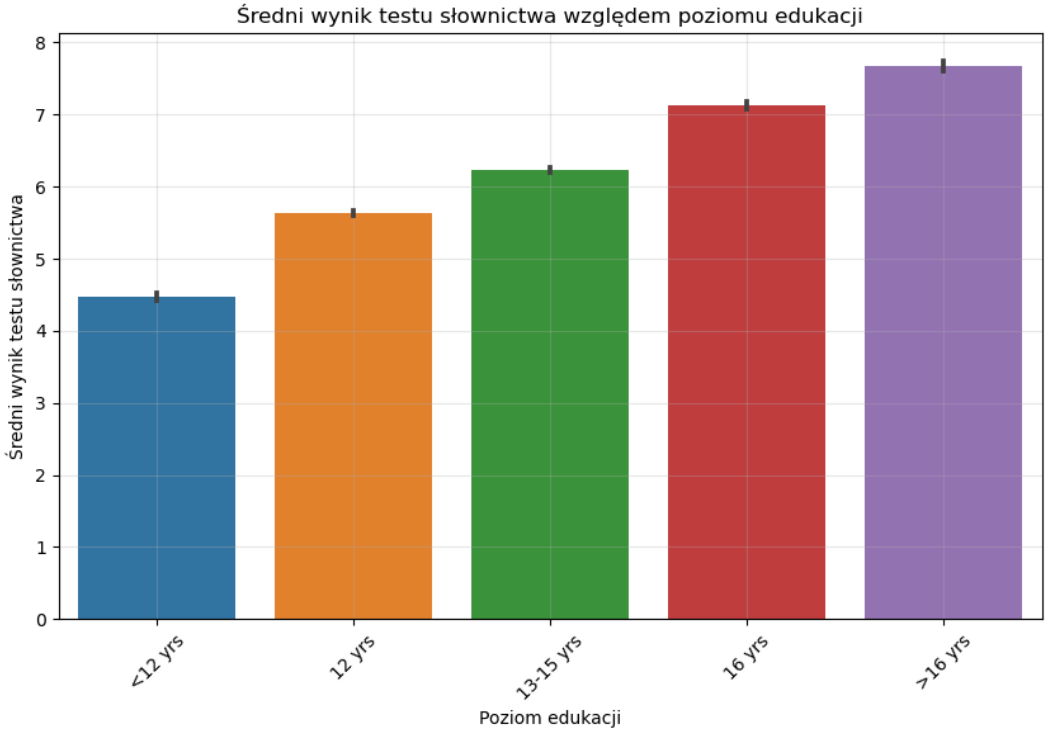
Poziom edukacji	**% Respondentów**
>16 lat	11.3
16 lat	13.9
13-15 lat	25.4
12 lat	30.2
<12 lat	19.2

Źródło: Opracowanie własne

Wstępna eksploracja danych wykazała, że różnice w wynikach testu słownictwa są wyraźnie widoczne w zależności od poziomu wykształcenia.

Rysunek 3.1 przedstawia średni wynik testu słownictwa w zależności od poziomu edukacji, pozując tendencję wzrostową- im wyższy poziom wykształcenia, tym lepszy wynik, co jest zgodnie z logiką.

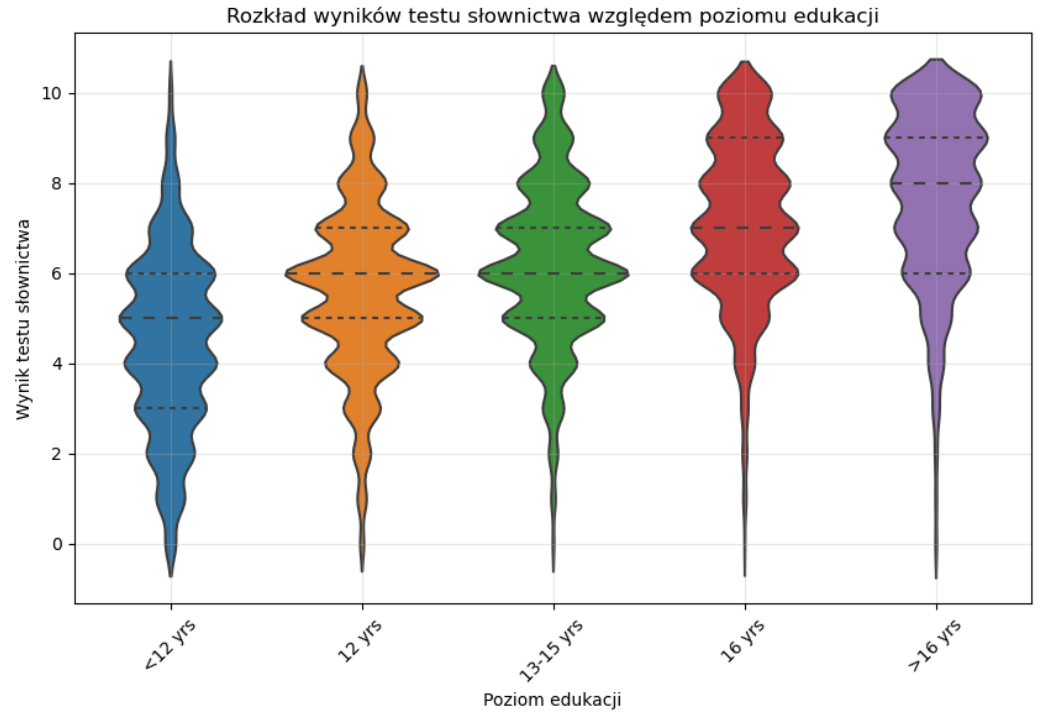
Rysunek 3.1 Średni wynik testu słownictwa



Źródło: Opracowanie własne

Natomiast rysunek 3.2 przedstawia rozkład wyników testu słownictwa dla różnych poziomów edukacji, pokazując gęstość danych.

Rysunek 3.2 Rozkład wyników testu słownictwa



Źródło: Opracowanie własne

W tabeli 4 przedstawiono liczbę brakujących wartości w każdej zmiennej. Najwięcej braków dotyczy wyniku testu słownictwa , gdzie brakuje 1348 obserwacji.

Braki danych mogą wynikać z różnych przyczyn: niechęci respondentów do udzielenia odpowiedzi na pewne pytania, błędów w rejestracji informacji lub pominięcia niektórych zmiennych w ankiecie.

Ponieważ odsetek brakujących wartości nie przekracza 5%, zostały one usunięte ze zbioru danych. Należy podkreślić, że każda brakująca wartość niesie pewną informację, a jej usunięcie ze zbioru danych wiąże się z utratą informacji.

Tabela 3.4 Liczbę brakujących wartości

Zmienna	**Liczba brakujących wartości**
Rok przeprowadzenia badania	0
Wiek respondenta w latach	94
Grupa wiekowa	94
Liczba lat edukacji	81
Kategoria poziomu wykształcenia	81
Płeć respondenta	0
Informacja, czy respondent urodził się w USA	87
Wynik testu słownictwa	1348

Źródło: Opracowanie własne

4.Dobór metody analizy

Biorąc pod uwagę rodzaj posiadanych danych można rozważyć wykorzystanie metody Propensity Score Matching (PSM) w ramach modelu Neymana-Rubina, zakładającej redukcję obciążenia selekcyjnego przy estymacji efektu edukacji na wynik testu słownictwa. Została zdefiniowana grupa kontrolna oraz grupa traktowana. Grupa kontrolna w niniejszej pracy obejmuje respondentów z niższym poziomem edukacji (≤12 lat nauki), natomiast grupę traktowaną stanowią osoby, które zdobyły wyższy poziom wykształcenia (>12 lat nauki). Następnie dopasowanie grup przeprowadzono przy użyciu metody PSM, aby wyrównać różnice w cechach demograficznych i zapewnić porównywalność obu grup.

Założenia metody PSM w kontekście niniejszego opracowania są następujące[4]:

1. Warunkowa niezależność (Conditional Independence Assumption, CIA),zakłada, że po uwzględnieniu obserwowalnych cech X, przypisanie jednostki do grupy traktowanej lub kontrolnej nie zależy od jej potencjalnych wyników testu słownictwa. Założenie

[4] Abadie A., Imbens G.W., 2006, Large sample properties of matching estimators for average treatment effects, Econometrica, vol. 74(1), 235-267.

CIA nie w praktyce nie można sprawdzić, ponieważ nie mamy dostępu do cech nieobserwowalnych.[5]

2. Wspólny obszar określoności (Common Support Condition), oznacza, że dla każdej jednostki istnieje szansa znalezienia odpowiednika w drugiej grupie. Dane założenie można zweryfikować.
3. Stabilność efektu traktowania (Stable Unit Treatment Value Assumption, SUTVA) Zakłada, że wpływ edukacji na wynik testu słownictwa badanej jednostki powinien być niezależny od sytuacji innych respondentów. Założenia nie podlega bezpośredniej weryfikacji. W niniejszej analizie prawdopodobnie dochodzi do naruszenia tego założenia, ponieważ osoby z wyższym wykształceniem mogą wpływać na inne osoby.

W niniejszej analizie zastosowano cztery podejścia do dopasowania[6]:

1. Nearest Neighbor Matching polega na przypisaniu każdej jednostki z grupy traktowanej do najbliższej jednostki z grupy kontrolnej na podstawie wartości propensity score. Najbliższy sąsiad jest wybierany na podstawie minimalnej odległości w przestrzeni propensity score, mierzonej metryką euklidesową.

2. Metoda Mahalanobis Matching polega na dopasowaniu jednostek traktowanych do jednostek kontrolnych na podstawie odległości Mahalanobisa, która uwzględnia wzajemne korelacje między zmiennymi. Dopasowanie jest liczone w wielowymiarowej przestrzeni.
3. Inverse Probability Weighting (IPW) to metoda wykorzystuję wagi oparte na odwrotności propensity score. To pozwala skorygować różnice między grupą traktowaną a kontrolną. Obserwacje z grupy traktowanej i kontrolnej są ważone tak, aby jednostki były bardziej zbliżone do siebie. Jedna z głównych wad IPW jest podatność na duże wariancje.
4. Metoda Propensity Score Stratification polega na podziale jednostek na kwantyle na podstawie ich propensity score i porównaniu efektu przyczynowego wewnątrz każdej grupy. Wadą tej metody jest to, że wymaga dużej liczby obserwacji.

5.Interpretacja wyników

Przed zastosowaniem metody PSM przeprowadzony został test t-Studenta, polegający na porównaniu średnich dwóch grup, w celu sprawdzenia istotnych statystycznie różnic między grupą traktowaną a kontrolną w odniesieniu do wieku, płci oraz miejsca urodzenia respondentów. Wartości p = 0.000 dla wszystkich analizowanych zmiennych sugerują, że różnice między grupami nie są randomizowane, a odzwierciedlają rzeczywiste nierównomierności w próbie.

Warto podkreślić, że istotne różnice pomiędzy grupami, mogą prowadzić do błędnych wniosków podczas analizy.

W tabeli 5.1 przedstawione są wyniki oszacowania modelu regresji logistycznej. Wszystkie zmienne są istotne na poziomie 5%. Wiek, płeć oraz miejsce urodzenia ma statystycznie istotny wpływ na prawdopodobieństwo przynależności do grupy traktowanej. Interpretacja ilorazów szans są następująca:

[5] Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu, Taksonomia 24. Klasyfikacja i analiza danych – teoria i zastosowania, red. nauk. Krzysztof Jajuga, Marek Walesiak, Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu, Wrocław 2015, nr 384. Str. 61-65

[6] https://pages.uoregon.edu/waddell/metrics/matching.html (dostęp 29.01.2025)

Wraz ze wzrostem wieku o jeden rok, szansa na przynależność do grupy z wykształceniem powyżej 12 lat, maleje o około 1,3%.

Mężczyźni mają średnio o około 13,8% większą szansę na przynależność do grupy wykształceniem powyżej 12 lat.

Osoby urodzone w USA mają średnio o około 12,5% mniejszą szansę na przynależność do grupy z wykształceniem powyżej 12 lat. Dany wniosek ten jest nie do końca oczywisty.

Pseudo R^2 wynosi 0.0096, co sugeruje, że zmienne objaśniające mają stosunkowo niewielką zdolność do przewidywania zmiennej zależnej.

Tabela 5.1 Oszacowania modelu regresji logistycznej

Zmienna	**Oszacowania**	**Błąd standardowy**	**Z-value**	**p-value**	**Iloraz szans**
Wyraz wolny	0.6607	0.053	12.377	0.000	1,936
Wiek	-0.0125	0.001	-17.706	0.000	0.987
Płeć	0.1298	0.025	5.279	0.000	1,138
Urodzony w USA	-0.1337	0.044	-3.065	0.002	0,875
Pseudo R^2	0.0096				
AIC	37567,21				
BIC	37600,08				

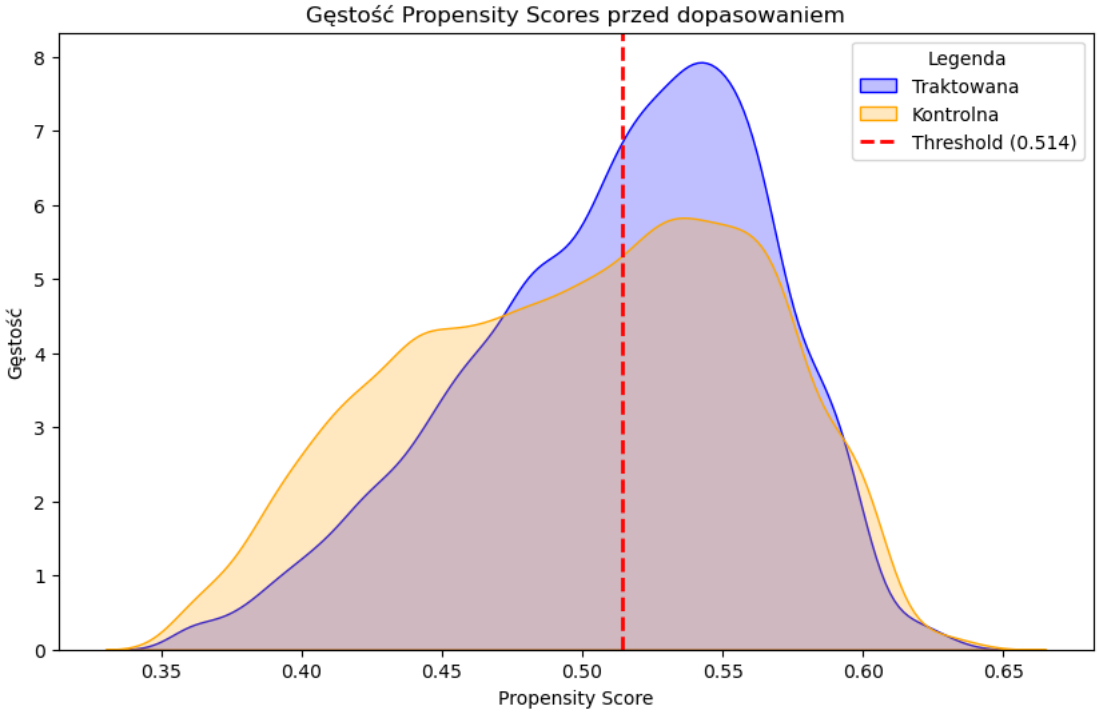
Źródło: Opracowanie własne

Następnie na podstawie oszacowanego modelu regresji logistycznej dokonano predykcji prawdopodobieństwa przynależności danej jednostki do grupy osób z wykształceniem powyżej 12 lat. To prawdopodobieństwo, nazywane również propensity score.W kolejnym etapie przeprowadzono dopasowanie jednostek, wykorzystując metodę Nearest Neighbor Matching oraz Radius Matching.

Dla metody Nearest Neighbor Matching wybierany był jeden najbliższy sąsiad z grupy kontrolnej. Średni wynik testu słownictwa w grupie z wykształceniem powyżej 12 lat wyniósł 6,8. W przypadku osób z wykształceniem poniżej 12 lat średni wynik testu słownictwa wyniósł 5,42. Szacowany efekt przyczynowy stanowi 1,38, co oznacza, że średni wynik testu słownictwa w grupie osób z wykształceniem powyżej 12 lat jest o około 1,38 punktu wyższy niż w grupie kontrolnej po dopasowaniu.

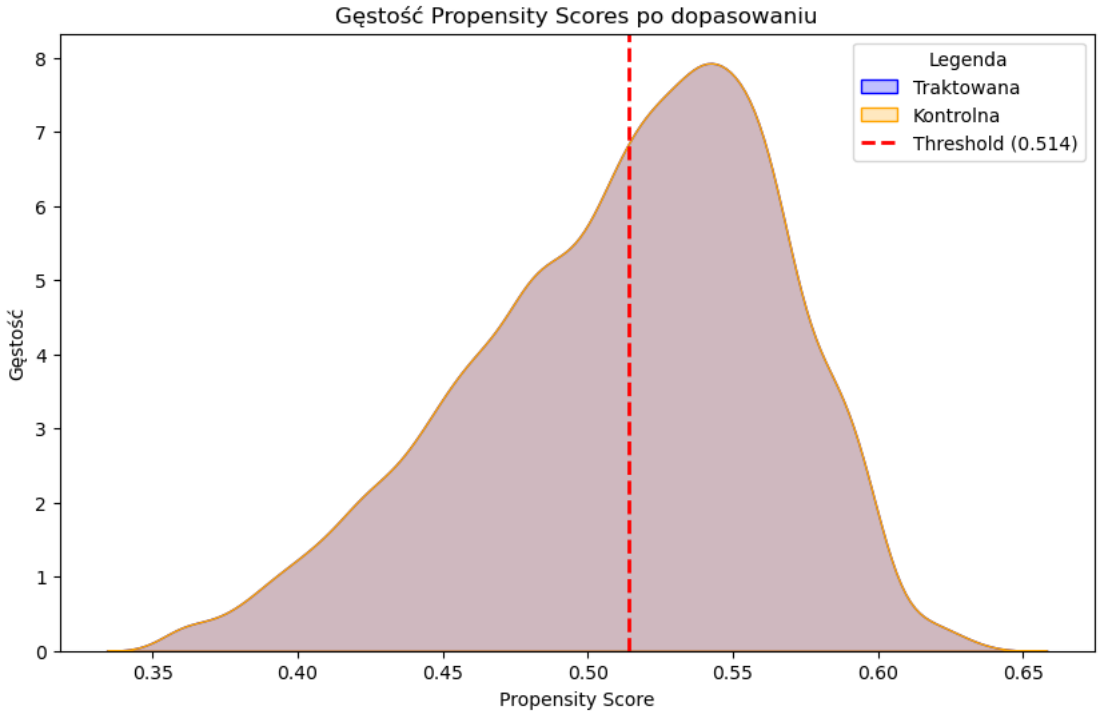
Na rysunku 5.1 i 5.2 przedstawione rozkłady propensity score przed i po dopasowaniu dla metody Nearest Neighbor Matching. Czerwona linia oznacza threshold, ustalony na poziomie proporcji jednostek w grupie traktowanej. Możemy zauważyć, że grupa traktowana ma wyraźnie wyższe wartości propensity score przed dopasowaniem, wskazując na to, że grupy nie były dobrze zbilansowane. Natomiast po dopasowaniu rozkłady grup kontrolnej i traktowanej niemal całkowicie się pokrywają, wyrównując różnice między grupami.

Rysunek 5.1 Rozkład propensity score przed dopasowaniem



Źródło: Opracowanie własne

Rysunek 5.2 Rozkład propensity score po dopasowaniu



Źródło: Opracowanie własne

Dla metody Propensity Score Stratification, zmienna propensity score została podzielona na 5 kwantyli. Szacowany efekt przyczynowy wyniósł 1.65, co oznacza, że osoby z wykształceniem powyżej 12 lat uzyskały średnio o 1.65 punktu wyższy wynik w teście słownictwa w porównaniu do grupy osób z wykształceniem poniżej 12 lat . Wynik jest nieco wyższy niż w przypadku Nearest Neighbor Matching.

W przypadku metody Mahalanobis Matching wybierany był jeden najbliższy sąsiad z grupy kontrolnej. Szacowany efekt przyczynowy wyniósł 1,24, co oznacza, że osoby z wykształceniem powyżej 12 lat uzyskały średnio o 1,24 punktu wyższy wynik w teście słownictwa w porównaniu do grupy osób z wykształceniem poniżej 12 lat . Wynik jest niższy niż w przypadku powyższych metod, co wskazuję na lepsze dopasowanie grup.

Metoda Inverse Probability Weighting(IPW) oszacowała efekt przyczynowy na poziomie 3,04, co oznacza, że osoby z wykształceniem powyżej 12 lat uzyskały średnio o 3,04 punktu wyższy wynik w teście słownictwa w porównaniu do grupy osób z wykształceniem poniżej 12 lat . Oszacowana wartość jest wyraźnie wyższa niż w pozostałych metodach, możemy mieć do czynienia z przeszacowywaniem efektu przyczynowego.

W celu weryfikacji wyników dopasowania przez cztery metody skorzystano z metryki pozwalającej na ocenę jakości dopasowania. Standaryzowana różnica średnich to miara równowagi między grupą traktowaną i kontrolną przed oraz po dopasowaniu. Wartość standaryzowanej różnicy średnich po dopasowaniu powinna być jak najbliższa zeru.

Metryka jest obliczana następująco:

Równanie 5.1 Standaryzowana różnica średnich

$$\text{SRS} = \frac{|\overline{X_1} - \overline{X_0}|}{\sqrt{\frac{S_1^2 + S_0^2}{2}}}$$

gdzie:

$$(\overline{X_1}) - \textit{średnia wartość zmiennej w grupie traktowanej}$$

$$(\overline{X_0}) - \textit{średnia wartość zmiennej w grupie kontrolnej}$$

$$(S_1) - \textit{odchylenie standardowe zmiennej w grupie traktowanej}$$

$$(S_0) - \textit{odchylenie standardowe zmiennej w grupie kontrolnej}$$

Tabela 5.2 przedstawia wartości standaryzowanej różnicy średnich obliczonej przed i po dopasowaniu:

Tabela 5.2 Wartości standaryzowanej różnicy średnich

Zmienna	Przed dopasowaniem	Po dopasowaniu (Nearest Neighbor Matching)	Po dopasowaniu (Mahalanobis Matching)	Po dopasowaniu (Propensity Score Stratification)	Po dopasowaniu (Inverse Probability Weighting)
Wiek	0,22	0,00028	0	0,134	0,22
Płeć	0,072	0	0	0,0703	0,072
Czy urodzony w USA	0,044	0,0015	0	0.0476	0,044

Źródło: Opracowanie własne

Możemy wnioskować, że metody: Nearest Neighbor Matching oraz Mahalanobis Matching skutecznie zredukowali różnice między grupami dla wszystkich zmiennych, osiągając prawie idealne dopasowanie. Metoda Propensity Score Stratification gorzej wyrównało grupy niż dwie poprzednie. Metoda Inverse Probability Weighting nie zmieniła wartości w stosunku do wartości przed dopasowaniem, nie poprawiając równowagi między grupami.

6.Zakończenie

W niniejszej pracy przeprowadzona została analiza wpływu czynników demograficznych i edukacyjnych na poziom znajomości słownictwa. Badano rolę wykształcenia w kształtowaniu wyników testu słownictwa, a istotność wieku, płci oraz miejsca urodzenia na poziom edukacji. W pracy zastosowane również metody dopasowania jednostek: Nearest Neighbor Matching, Mahalanobis Matching, Propensity Score Stratification oraz Inverse Probability Weighting,

Na podstawie oszacowanego modelu regresji logistycznej możemy stwierdzić, że wiek, płeć oraz miejsce urodzenia istotnie wpływają na prawdopodobieństwo osiągnięcia wyższego poziomu wykształcenia w badanej próbie. Wyniki szacowanego efektu przyczynowego różniły się w zależności od metody dopasowania, natomiast każda metoda potwierdziła , że osoby z wykształceniem powyżej 12 lat osiągały średnio wyższy wynik testu słownictwa w porównaniu do pozostałych.

Również standaryzowana różnica średnich pomogła wyznaczyć jaka metoda najskuteczniej dopasowuje jednostki z obu grup. Najlepsze dopasowanie jednostek zapewniła metoda Mahalanobis Matching, natomiast Inverse Probability Weighting była najmniej skuteczna.

Podsumowując, wyniki badania potwierdziły, że poziom wykształcenia ma kluczowy wpływ na znajomość słownictwa. Wybór odpowiedniej metody dopasowania jest istotnym elementem analizy. Dalsze badania mogłyby rozszerzyć analizę o inne czynniki.

7. Bibliografia

1.Card, D., & Krueger, A. B. (1994). Minimum wages and employment: A case study of the fast-food industry in New Jersey and Pennsylvania. American Economic Review, 84(4), 772–793.
2.Abadie A., Imbens G.W., 2006, Large sample properties of matching estimators for average treatment effects, Econometrica, vol. 74(1), 235-267.
3.Strawiński P., 2008, Quasi-eksperymentalne metody ewaluacji, [w:] Środowisko i warsztat ewaluacji, red. A. Haber, RARP, Warszawa, s. 1-220.
4.https://gss.norc.org/ (dostęp 21.01.2025)
5.https://www.kaggle.com/datasets/utkarshx27/general-social-survey(dostęp 20.01.2025)
6.https://pages.uoregon.edu/waddell/metrics/matching.html (dostęp 29.01.2025)
7.Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu, Taksonomia 24. Klasyfikacja i analiza danych – teoria i zastosowania, red. nauk. Krzysztof Jajuga, Marek Walesiak, Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu, Wrocław 2015, nr 384. Str. 1-74