

Artur Pokropek
Instytut Filozofii i Socjologii PAN
Zespół Badawczy EWD

Analiza efektów kontekstowych – problemy związane z rzetelnością

Wprowadzenie

Uczniowie chodzący do różnych szkół, nawet gdy przydział do szkół był losowy, zaczęli się różnicować według pewnych wzorów. Jedni chodzą do lepszej szkoły – w pełnym tego słowa znaczeniu, drudzy do szkół kiepskich, jedni uczniowie mają genialnego matematyka i fizyka, a inni mają wspaniałą pracownię informatyczną i zdolnego informatyka. W jednych klasach uczniowie mają kolegów i koleżanki pozytywnie nastawione do szkoły i motywujących się nawzajem do nauki. W innych klasach może panować atmosfera antyszkolna, co wydatnie może utrudniać naukę i nauczanie. Cechy i umiejętności uczniów zaczynają się różnicować i stawać się podobne w zależności od różnych kontekstów środowiskowych. Badanie efektów kontekstowych ma duże znaczenie dla polityki edukacyjnej. Można z łatwością wymienić wiele z nich. Jakie cechy środowiskowe sprzyjają nauce, a które mogą ją utrudniać? Jakie warunki środowiskowe mogą sprzyjać występowaniu agresji w szkole, a jakie ją powstrzymują? Jaka atmosfera szkolna wpływa na równowagę psychiczną uczniów i proces socjalizacji?

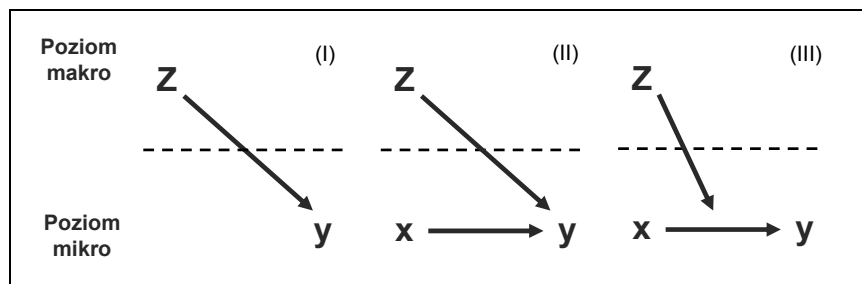
Od lat 80. dysponujemy narzędziami statystycznymi, które zaprojektowane zostały między innymi po to, by umożliwić poprawną analizę efektów kontekstowych – analizy wielopoziomowe. Klasyczne analizy wielopoziomowe zakładają jednak, iż używane w nich zmienne nie są obciążone błędem pomiaru. Sytuacja taka w edukacji zdarza się jednak bardzo rzadko. W badaniach edukacyjnych na ogół posługujemy się testami umiejętności lub kwestionariuszami badającymi postawy, pochodzenie społeczne, status społeczno-ekonomiczny etc., które charakteryzuje ograniczona rzetelność. Fakt ten bywa często pomijany. W artykule pokazane zostanie, jak nieuwzględnienie błędu pomiaru może odbić się na jakości uzyskiwanych wyników. Zaproponowane i przetestowane zostaną również metody, które uwzględnią błąd pomiaru podczas estymacji parametrów.

Analiza zmiennych kontekstowych

W zdecydowanej większości przypadków kontekst może wpływać na jednostki w trojaki sposób, co zostało przedstawione na rysunku 1 (za: Snijders i Bosker 1999)

Po pierwsze, mamy do czynienia z sytuacją, w której zmienna z niższego poziomu jest zależna od zmiennej z wyższego poziomu (I). Wyniki ucznia zależą od tego, do jakiej klasy trafił. Po drugie, można opisywać sytuację, gdy zmienna z niższego poziomu jest zależna zarówno od zmiennej z wyższego poziomu, jak i od zmiennej z niższego poziomu (II). Wyniki ucznia zależą jednocześnie od jego wcześniejszych osiągnięć i od tego, jaka atmosfera panuje w klasie. Po trzecie, gdy rela-

cje zmiennych z niższego poziomu są zależne od zmiennej z wyższego poziomu (III). Jest to sytuacja, w której analizuje się na przykład, jak w określonych kontekstach uczeń może wykorzystywać wcześniej zdobytą wiedzę lub inne zasoby, np. zasoby materialne.



Rysunek 1. Zależności między poziomem mikro i makro

W tym artykule analizowana jest sytuacja, w której dysponujemy zmiennymi agregowanymi określającymi średnie wyniki grupowe uczniów. Co można zapisać za pomocą prostego modelu:

$$y_{ij} = x_{ij} + \bar{x}_j + e_{ij} \quad (1.1)$$

Gdzie y_{ij} jest zmienną zależną dla i -tego ucznia z j -tego oddziału szkolnego, x_{ij} jest zmienną niezależną, a \bar{x}_j zmienną niezależną zagregowaną na poziomie klasy, natomiast e_{ij} to wyraz błędu. Mamy zatem zmienną, która opisuje zarówno indywidualne właściwości jednostki, jak i kontekst grupowy poprzez zagregowanie indywidualnych właściwości. Podobne analizy spotykamy stosunkowo często, gdy badamy zagadnienia związane ze szkołą, co wskazane było już we wstępie. Sztandarowym przykładem jest tu efekt rówieśnika, gdzie twierdzi się, że uczeń, który trafia do klasy, gdzie średnio rówieśnicy mają wysokie kompetencje mierzone wynikami egzaminacyjnymi (\bar{x}_j) z wcześniejszych egzaminów, uzyska większy wynik na egzaminie końcowym (y_{ij}) niż uczeń, który trafiłby do klasy o niskich średnich wynikach. W zdecydowanej większości przypadków wynik ten się potwierdza (Dolata 2009). Innym przykładem jest problem tak zwanej „dużej ryby w małym stawie”. W tym wypadku zmienną zależną jest wynik egzaminacyjny, zmienną niezależną na poziomie indywidualnym poczucie własnej skuteczności oraz zagregowana dla klasy wartość tego wskaźnika. Okazuje się, iż na poziomie indywidualnym, im wyższe poczucie własnej skuteczności, tym wyższe wyniki. Na poziomie grupowym dostrzegany jest natomiast negatywny związek. Im wyższe średnie poczucie własnej skuteczności w klasie, tym niższe wyniki indywidualne uczniów (przy kontroli indywidualnego poczucia skuteczności, Marsh 2003).

Empiryczne potwierdzenia tych efektów pozostawiają jednak pewne wątpliwości. Bardzo rzadko w tego typu analizach bierze się pod uwagę naturę wyko-

rzystywanych zmiennych. Precyzyjnie rzecz ujmując, nie uwzględnia się tego, że używane zmienne charakteryzują się błędem pomiaru i ograniczoną rzetelnością.

Błąd pomiaru i rzetelność

Podstawowym zadaniem pomiaru wiadomości i umiejętności jest określenie wiedzy ucznia w jakiejś dziedzinie. W edukacji często próbuje się określić poziom tych umiejętności na podstawie wyniku testu. Oprócz wiadomości i umiejętności ucznia na wynik uzyskany w danym teście mogą wpływać inne czynniki. Jednym z głównych czynników jest oczywiście dobór zadań (problem reprezentatywności). Uczeń może trafić na zadania reprezentujące wiedzę, którą powtarzał niedawno, ale może być też odwrotnie, może powtarzał dany wycinek materiału bardzo dawno temu albo trafił na zadania związane z materiałem, który był omawiany w szkole podczas jego choroby. Testy edukacyjne nie mogą mierzyć wszystkich wymaganych zagadnień jednocześnie. To, co robimy podczas pomiaru dokonywanego na podstawie testów, jest próbą określenia poziomu ogólnych umiejętności ucznia z danego przedmiotu na podstawie kilkunastu zadań wybranych według określonych zasad. Należy podkreślić, że żaden znany nam pomiar wiadomości i umiejętności (nawet ten, który zawierałby cały materiał) nie jest pomiarem w pełni precyzyjnym. Każdy pomiar obarczony jest błędem. Problematyka rzetelności zajmuje się analizą tych błędów, czyli próbuje odpowiedzieć na pytanie, w jakim stopniu dany pomiar jest dokładny. Rzetelność testu dotyczy precyzji, z jaką jesteśmy w stanie zmierzyć poziom wiedzy.

Aby mówić o rzetelności, należy wprowadzić pojęcie rozkładu wyników prawdziwych. Zwyczajowo oznacza się jego parametry: średnia równa zero i wariancja σ_t^2 ($N(0; \sigma_t^2)$). Średnia jest tu równa zero, co może dziwić – jest to konwencja wprowadzona przez statystyków, znacznie ułatwiająca obliczenia. Każdy rozkład zmiennej ciągłej łatwo można sprowadzić do średniej zero – odejmując od wszystkich wyników poszczególnych uczniów wynik średni testu. Wtedy uczniowie, którzy uzyskali wyniki poniżej średniej, charakteryzowani będą przez wyniki ujemne, a uczniowie powyżej średniej – przez wyniki dodatnie.

Zakładając, iż błędy są losowe, niezależne od siebie i niezależne od wyniku prawdziwego, możemy powiedzieć, iż wariancja, czyli zróżnicowanie wyników, dla całego testu ma dwa źródła: wariancję błędu i wariancję wyniku prawdziwego. Jako że błędy nie są zależne od wyniku prawdziwego i od siebie nawzajem, możemy zapisać całą wartość zmienności wyników uzyskanych analogicznie do ich sumy:

$$\sigma_x^2 = \sigma_t^2 + \sigma_e^2 \quad (1.2)$$

gdzie:

σ_x^2 – wariancja wyniku uzyskanego

σ_t^2 – wariancja wyniku prawdziwego

σ_e^2 – wariancja błędu

Rzetelność mówi nam o jakości narzędzia, o tym, na ile precyzyjnie mierzy ono to, co ma mierzyć. Definiuje go się jako udział zróżnicowania wyniku prawdziwego w całym zróżnicowaniu. Inaczej mówiąc, jest to stosunek wariancji wyniku prawdziwego do wariancji wyniku uzyskanego na podstawie testowania (będącego sumą zróżnicowania wyniku prawdziwego oraz zróżnicowania błędu pomiaru):

$$\alpha = \frac{\sigma_t^2}{\sigma_t^2 + \sigma_e^2} \quad (1.3)$$

Zasadnicze pytanie postawione w tym artykule brzmi: jak fakt istnienia rzetelności wpływa na oszacowania estymatorów efektów kontekstowych. Po drugie, jakich metod używać, by uzyskać wyniki jak najmniej obciążone.

Rzetelność w modelowaniu efektów kontekstowych

Zaczynimy od prostego modelu z jedną zmienną wyjaśniającą:

$$y = \beta x + e \quad (1.4)$$

W sytuacji błędu pomiaru to, co obserwujemy to (obserwowalne wartości oznaczono wielkimi literami, wartości prawdziwe małymi):

$$Y = y + v \text{ oraz } X = x + u$$

gdzie v i u to błędy pomiaru, y i x to wyniki prawdziwe. Będziemy zakładać, że:

$$E(u) = E(v) = 0 \quad \text{var}(u) = \sigma_u^2 \quad \text{var}(v) = \sigma_v^2 \quad (1.5)$$

$$\text{cov}(u, x) = \text{cov}(u, y) = \text{cov}(v, x) = \text{cov}(v, y) = 0$$

Równanie z jedną zmienną wyjaśniającą dla zmiennych obserwowalnych można zatem zapisać jako:

$$Y - v = \beta(X - u) + e \quad (1.6)$$

w uproszczeniu:

$$Y = \beta X + w \bar{Y} \quad (1.7)$$

gdzie $w = e + v - \beta u$

Okazuje się, że wyniki klasycznego modelu regresyjnego będą szacowane błędnie, gdy nie weźmiemy pod uwagę błędu pomiaru, gdyż pogwałcony zostanie postulat: $\text{cov}(w, X) = 0$.

W naszym przypadku $\text{cov}(w, X) = \text{cov}(-\beta u, x + u) = -\beta \sigma_u^2$. Jeżeli szacujemy β w klasyczny sposób (metodą najmniejszych kwadratów), to estymator szukanego parametru przyjmuje następującą postać:

$$\hat{\beta} = \frac{\sum XY}{\sum X^2} = \frac{\sum (x+u)(y+v)}{\sum (x+u)^2} \quad (1.8)$$

Po przyjęciu założeń o nieskorelowaniu błędów i nieskorelowaniu błędów ze zmiennymi, można go wyrazić w następujący sposób:

$$\hat{\beta} = \frac{\text{cov}(xy)}{\text{var}(x) + \text{var}(u)} = \frac{\sigma_{xy}^2}{\sigma_x^2 + \sigma_u^2} \quad (1.9)$$

Okazuje się też, że błąd w zmiennej zależnej nie odgrywa roli podczas estymacji pożądanego parametru. Widać również, że estymator ten będzie nie doszacowywał wartości prawdziwej parametru z powodu dodatkowej wariancji błędu pomiaru w mianowniku.

Trzymając się poprzednich założeń, można jednak pokazać, że:

$$\hat{\beta} = \frac{\sigma_{xy}}{\sigma_x^2 + \sigma_u^2} = \frac{\sigma_{xy} / \sigma_x^2}{1 + \sigma_u^2 / \sigma_x^2} = \beta \frac{1}{1 + \sigma_u^2 / \sigma_x^2} = \beta \frac{\sigma_x^2}{\sigma_x^2 + \sigma_u^2} \quad (1.10)$$

Łatwo zauważyć, iż ostatni człon równania: $\frac{\sigma_x^2}{\sigma_x^2 + \sigma_u^2}$ to nic innego, jak rzetelność

dobrze znana z klasycznej teorii testu, czyli stosunek wyniku prawdziwego do wyniku obserwowalnego. Wynika stąd, iż znając rzetelność zmiennej niezależnej, możemy łatwo wprowadzić korektę, aby uzyskać nieobciążony estymator β , tak

że: $\beta = \frac{\hat{\beta}}{\alpha}$

Oczywiście nie znamy prawdziwej rzetelności zmiennej. Potrafimy ją jednak oszacować narzędziami z klasycznej teorii testu – Alfya Cronbacha. Obliczenie wskaźnika Alfya Cronbacha nie jest trudne i sprowadza się do użycia jednego wzoru:

$$\hat{\alpha} = \frac{k}{k-1} \left(1 - \frac{\sum_{i=1}^k \sigma_i^2}{\sigma_{suma}^2} \right) \quad (1.11)$$

gdzie:

k – ilość pytań w teście

σ_i^2 – wariancja dla i -tego pytania

σ_{suma}^2 – wariancja dla wyniku sumarycznego

Zatem wartość prawdziwą parametru można łatwo przybliżyć: $\beta \approx \frac{\hat{\beta}}{\hat{\alpha}}$

Sprawa jest bardziej skomplikowana w przypadku modelu z dwoma zmiennymi, z których jedna jest agregatem z drugiego poziomu, tak jak w równaniu (1.1). W tym wypadku możemy założyć, iż zmienna z drugiego poziomu będzie miała bardzo niewielki, pomijalny błąd pomiaru, gdyż agregowana jest w grupie – poszczególne błędy znoszą się. W zależności od stopnia segregacji wewnątrzgrupowej, ze względu na indywidualną zmienną niezależną, będzie też skorelowana z wynikami indywidualnymi. W takiej sytuacji można pokazać (jednak pominiemy tutaj stosunkowo długi dowód, por. Maddala 1986), że estymator dla efektu indywidualnego będzie obciążony w następujący sposób:

$$\hat{\beta}_x = \beta_x \left(1 - \frac{1 - \alpha}{1 - \rho^2} \right) \quad (1.12)$$

Gdzie ρ to korelacja między zmienną kontekstową a indywidualną. Natomiast estymator efektu kontekstowego będzie obciążony w następujący sposób:

$$\hat{\beta}_x = \beta_x + \frac{\beta_x(1 - \alpha)\rho}{1 - \rho^2} \quad (1.13)$$

Oznacza to, iż efekt kontekstowy będzie przeszacowany o wyraz $\frac{\beta_x(1 - \alpha)\rho}{1 - \rho^2}$. Nawet jeżeli efektu kontekstowego w rzeczywistości nie ma, lecz rzetelność jest mała, a korelacja między zmienną kontekstową a indywidualną duża, to badacz używający klasycznych metod otrzyma wynik, który utwierdzi go w przekonaniu, iż efekt kontekstowy zaistniał. Podczas gdy w rzeczywistości będzie on tylko zwykłym artefaktem. W kolejnej sekcji pokazane zostanie, jak różne typy narzędzi statystycznych radzą sobie z przedstawioną analitycznie sytuacją.

Metody analizy zmiennych kontekstowych

W poprzedniej sekcji udało się pokazać, jaki wpływ może mieć nieuwzględnienie problematyki błędu pomiaru w analizach zmiennych kontekstowych. Nie mniej pozostają dwa znaczące problemy. Po pierwsze, założenia przyjęte w równaniu (1.5) są dosyć restrykcyjne i wydaje się, że w rzeczywistość rzadko są spełniane – szczególnie założenie mówiące o tym, że błąd pomiaru nie jest związany z poziomem umiejętności. Nie jest też łatwo udowodnić, jaki wpływ na estymacje ma fakt, iż rzetelność jest szacowana tylko w przybliżeniu. W takich sytuacjach

najlepiej odwołać się do symulacji. W symulacjach sprawdzone zostaną trzy narzędzia statystyczne: 1) regresja wielopoziomowa, która jest metodą bazową, 2) regresja wielopoziomowa z korektą na rzetelność (tak jak przedstawione zostało to w równaniach 1.12 i 1.13), 3) regresja wielopoziomowa z *plausible values* dla zmiennej niezależnej. We wszystkich¹ modelach zmienna zależna skalowana będzie za pomocą modelu dwuparametrycznego (2PL).

Model wielopoziomowy

Na model wielopoziomowy można patrzeć jako na rozszerzenie klasycznej regresji jednej zmiennej o informacje o pogrupowaniu jednostek. Model do badania zależności kontekstowych można przedstawić następująco:

$$\begin{aligned} \text{Poziom jednostki (1):} \quad & y_{ij} = \beta_{0j} + \beta_{1j}x_{ij} + \beta_{2j}\bar{x}_j + r_{ij} \\ & \text{gdzie:} \end{aligned} \tag{1.14}$$

$$\begin{aligned} \text{Poziom szkoły (2):} \quad & \beta_{0j} = \gamma_{00} + u_{0j} \\ & \beta_{1j} = \gamma_{10} \\ & \beta_{2j} = \gamma_{20} \end{aligned}$$

Jak widać, w tym modelu współczynniki nachylenia β_{1j} i β_{2j} uznajemy za stałe dla wszystkich grup, tak jak w klasycznej regresji, jednak pozostajemy przy założeniu z poprzedniego modelu związanym ze stałą regresji, którą traktujemy jako zmienną losową. Do wyrazu wolnego (stałej regresji) dołączony zostanie efekt losowy: $\beta_{0j} = \gamma_{00} + u_{0j}$. Cały model można przedstawić w jednym równaniu liniowym w następujący sposób:

$$y_{ij} = \gamma_{00} + \gamma_{10}x_{ij} + \gamma_{20}\bar{x}_j + u_{0j} + r_{ij} \tag{1.15}$$

Jak widać, zależność między Y i X jest stała dla całej populacji; tym, co różnicuje j -te grupy, jest efekt losowy u_{0j} , który wyraża to, iż w różnych grupach przewidujemy różne, warunkowe wartości oczekiwane. Model w symulacjach szacowany będzie metodą największej wiarygodności.

Model wielopoziomowy z poprawką na błąd pomiaru

Ten model jest analogiczny jak poprzedni, z tym że do estymacji wprowadzona zostanie poprawka przedstawiona w równaniach 1.12 i 1.13.

¹ Dla modelu z *plausible values* skalowanie będzie bardziej skomplikowane, choć nadal z użyciem modelu 2PL.

$$\beta_{1jMLE} = \left(1 - \frac{1-\alpha}{1-\rho^2}\right) / \hat{\beta}_{OLS} \quad (1.16)$$

$$\beta_{2jMLE} = \hat{\beta}_{2jOLS} - \frac{\beta_{1jMLE}(1-\alpha)\rho}{1-\rho^2} \quad (1.17)$$

Estymatory modelu wielopoziomowego oparte o metodę największej wiarygodności (MLE) będą korzystały z estymatora regresji najmniejszych kwadratów (OLS), gdy wprowadzimy korektę na rzetelność i skorelowanie zmiennych (1.16-1.17).

Model wielopoziomowy z plausible values

Klasyczną metodą szacowania umiejętności na podstawie modeli IRT jest estymacja wyników metodą największej wiarygodności. Inną metodą jest traktowanie umiejętności uczniów jako braków danych, które muszą być oszacowane na podstawie obserwowalnych odpowiedzi na pytania. Takiej metodologii używa PISA, TIMS czy NAEP – zwykle nazywana jest ona *plausible values* lub w skrócie PV.

PV są losowymi próbami z warunkowego rozkładu a posteriori cechy ukrytej każdego badanego ucznia (por. Mislevy 1991, Mislevy i inni 1992). Niech y oznacza wartość zmiennej zależnej, θ oznacza wartość cechy ukrytej – czyli wartość badanej cechy bez błędu pomiaru. Jeżeli θ byłaby znana dla każdego ucznia, możliwe byłoby obliczenie statystyki $t(\theta, y)$ takiej jak na przykład średnia warunkowa ze względu na płeć uczniów. Wtedy za pomocą klasycznych metod statystycznych można obliczyć wartość populacyjną dla tej statystyki T .

Jednak θ charakteryzuje cechę ukrytą uczniów, która nie jest bezpośrednio obserwowalna. Aby poradzić sobie z tym problemem, można przyjąć rozwiązanie Rubin (1987) i potraktować θ jako braki danych. Wtedy przybliżeniem dla $t(\theta, y)$ jest wartością oczekiwaną $t^*(x, y)$, gdzie x jest wektorem odpowiedzi na pytania testowe. Rubin pokazał, że:

$$t^*(x, y) = E|t(\theta, y) | x, y| = \int t(\theta, y) p(\theta | x, y) d\theta \quad (1.18)$$

Uzyskanie nieobciążonego estymatora statystyki t jest możliwe dzięki losowaniom z warunkowego rozkładu umiejętności przy danych odpowiedziach ucznia (x) na pytania testu i parametrach pytań oraz dodatkowych zmiennych (y). Rubin (1987) wskazuje na to, iż należy powtórzyć kilkakrotnie proces losowania tak, żeby uzyskać kilka wartości statystyki T . Średnia tych wylosowanych statystyk będzie przybliżeniem statystyki T . Precyzyjniej odnosząc metodę PV do naszych rozważań – generujemy pięć (ilość zbiorów może być większa, ale zwyczajowo stosuje się pięć) zbiorów na podstawie równania (1.18), w każdym z nich losując inną wartość θ dla zmiennej zależnej. Dalej zostanie estymowanych 5 osobnych modeli wielopoziomowych, następnie wyniki parametrów z pięciu analiz zostaną

uśrednione wedle wzoru (1.8), takie uśrednienie będzie przybliżeniem prawdziwych poszukiwanych parametrów.

Testowanie modeli – symulacje

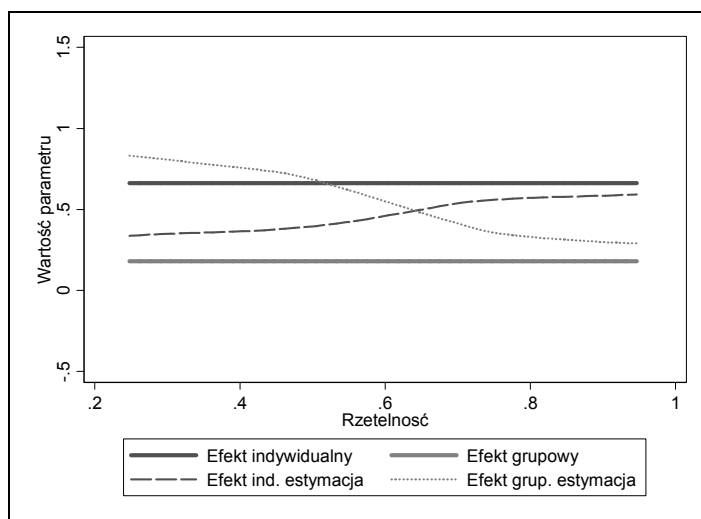
Analizy symulacyjne często używane są do diagnostyki modeli statystycznych. Strategia symulacyjna polega na stworzeniu sztucznych zbiorów danych o zadanych parametrach. Na tak skonstruowanych zbiorach przeprowadza się analizy, patrząc w jakim stopniu narzędzia statystyczne pozwalają odtworzyć zadane parametry. W metodach symulacyjnych wykorzystuje się od kilkudziesięciu do kilkuset zbiorów danych, które charakteryzują się takimi samymi wartościami szukanych parametrów, lecz różnymi wartościami zmiennych indywidualnych.

W analizie wykorzystano 400 zbiorów danych. W każdym z nich znalazło się 6000 obserwacji (można przyjąć, iż są to uczniowie). Każdemu z uczniów wylosowano (z rozkładu normalnego) wartości prawdziwe zmiennej zależnej (można przyjąć, iż jest to wynik prawdziwy na wyjściu) i niezależnej (wynik prawdziwy na wejściu), tak aby w populacji korelacja między zmiennymi wynosiła 0,7. Następnie uczniom przypisano do równolicznych trzydziestoosobowych klas w taki sposób, żeby korelacja wewnątrzspołowa ze względu na zmienną niezależną wynosiła 0,3 – co odpowiada sytuacji w Polsce na poziomie szkół gimnazjalnych, gdy zmienną niezależną są wyniki ze sprawdzianu po szkole podstawowej. Dla każdej klasy obliczono średnią grupową dla zmiennej niezależnej. Następnie wprowadzono efekt kontekstowy tak, iż każdemu uczniowi w klasie do zmiennej zależnej dodano 0,2 średniej klasowej (utworzonej ze zmiennej niezależnej) plus błąd o średniej 0 i odchyleniu standardowym 1,35. Oznacza to, że średnio rzecz biorąc, uczniowie klas o wyższych wynikach na wejściu będą otrzymywali wyniki wyższe niż uczniowie klas z niskimi średnimi wynikami na wejściu (przy znanej relacji 0,2 razy średnia klasowa na wejściu). W tak przygotowanych zbiorach znane są zatem wszystkie parametry. Efekt kontekstowy (czyli na ile średnia grupowa na wejściu wpływa na wyniki indywidualne) wynosi dokładnie 0,2. Efekt indywidualny zmniejsza się nieznacznie, ponieważ efekt grupowy generowany został poprzez dodanie wyrazu z błędem losowym i wynosi 0,65. Wygenerowany zbiór zawiera jedynie parametry prawdziwe. W dalszym kroku generowane są sztuczne testy o różnych charakterystykach pytań. Dla każdego zbioru danych wygenerowano 10 różnych testów. Każdy z testów ma 30 pytań. Pytania w każdym teście mają średnią 0 przy odchyleniu standardowym 1. Dziesięć testów różni się od siebie dyskryminacją pytań. W pierwszym średnia dyskryminacja wynosi 0,25, dalej: 0,3, 0,5, 0,7, 0,9, 1, 1,25, 1,5, 1,75 i 2,5, przy odchyleniu standardowym 0,2. Im wyższa średnia dyskryminacja pytań, tym większa precyzja pomiaru i większa rzetelność testu. Dla przyjętych wartości dyskryminacji Alfa Cronbacha przyjmuje wartości 0,33 do 0,94. Mając parametry pytań dla każdego testu oraz wartości cechy ukrytej uczniów, można przewidywać, jakie wyniki z poszczególnych testów uzyskają uczniowie o wygenerowanych wcześniej wartościach prawdziwych. Wyniki uzyskane na podstawie parametrów pytań oraz wartości prawdziwych umiejętności obarczone będą oczywiście błędem pomiaru, wielkość tego błędu uzależniona będzie od średniej dyskryminacji pytań, co bezpośrednio przekłada się na rzetelność. Dla każdego ucznia uzyskujemy w ten sposób 10

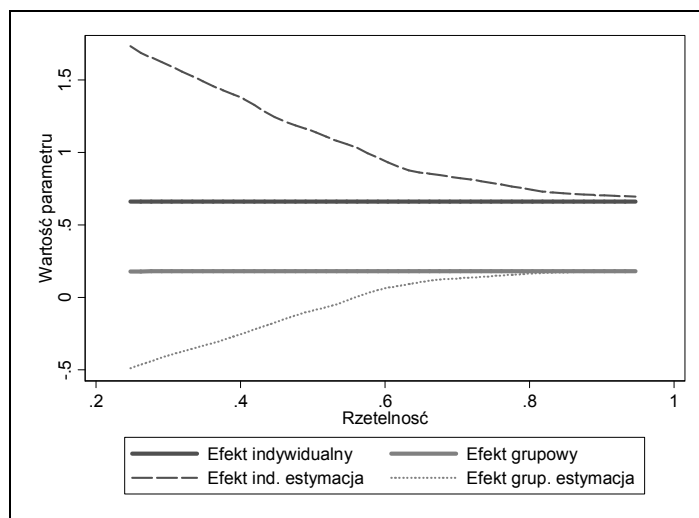
wyników obarczonych różnym błędem pomiaru (dodatkowo znamy wyniki prawdziwe). Dysponując takimi danymi, możemy sprawdzić, jak dobrze poszczególne metody estymują wynik prawdziwy w oparciu o wyniki testowe obciążone błędem pomiaru. Proste porównania wartości estymowanych parametrów z wartościami założonymi parametrów mówić będą nam o jakości metody. Im różnica między wartością parametru prawdziwego a wartością parametru estymowanego większa, tym metoda jest gorsza.

Na rysunkach 2-4 przedstawione zostały wyniki symulacji. Grube linie przedstawiają wartości efektów prawdziwych. Cienkie przerywane linie – estymowane efekty. Na rysunku 2. przedstawiono wyniki modelu wielopoziomowego bez uwzględnienia jakichkolwiek korekt. Jak widać, model ten znacznie przeszacowuje efekt kontekstowy i jednocześnie niedoszacowuje efektu indywidualnego. Te niedoszacowania i przeszacowania są na tyle duże, iż w pewnym momencie się przecinają. Do wartości rzetelności około 0,65 efekt kontekstowy wedle modelu wielopoziomowego szacowany jest jako większy niż indywidualny. Mimo iż w rzeczywistości dane zostały wygenerowane w ten sposób, by efekt indywidualny był około trzykrotnie wyższy.

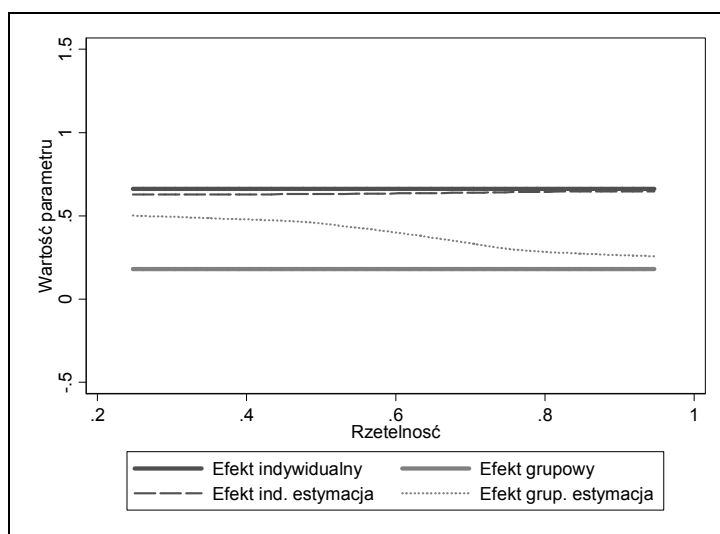
W przypadku regresji wielopoziomowej z korektą na rzetelność (rys. 3) mamy do czynienia z inną sytuacją. Tutaj efekt indywidualny jest przeszacowywany, a efekt kontekstowy niedoszacowywany. Do wartości rzetelności 0,6 obciążenie estymatorów jest niezwykle duże. Lecz należy zauważyć, iż od rzetelności około 0,8 metoda ta jest stosunkowo dokładna, szczególnie w przypadku zmiennych kontekstowych. Jak widać, zastosowana korekta nie sprawdza się w warunkach testowych. Jest to spowodowane niespełnieniem założeń (1.5) oraz niedokładnym szacowaniem rzetelności. Trzecim powodem może być błąd w estymacji korelacji między zmiennymi. Korelacja ta jest również liczona na podstawie wyników obarczonych błędem. W sytuacji niskiej rzetelności musi być szacowana ze znaczącym błędem.



Rysunek 2. Wyniki symulacji dla modelu wielopoziomowego



Rysunek 3. Wyniki symulacji dla modelu wielopoziomowego z korektą na rzetelność



Rysunek 4. Wyniki symulacji dla modelu wielopoziomowego z plausible values

Model z *plausible values*, który nie ma silnych założeń, tak jak model z analityczną korektą, zachowuje się znacznie lepiej w przypadku estymowania efektu indywidualnego. Jednak estymowany efekt kontekstowy jest znacząco przeszacowany dla rzetelności mniejszej niż 0,7. Dla wyższych wartości rzetelności przeszacowanie to spada, lecz nadal zostaje istotne i może wprowadzać badacza w błąd.

Zamiast zakończenia: efekt rówieśnika

W tej części zamiast podsumowania pokazana zostanie prosta analiza efektu rówieśnika. Szeroko przyjęta hipoteza mówi o tym, iż im wyższe wyniki rówieśników w klasie mierzone wynikiem na wejściu, tym uczniowie uzyskują wyższe wyniki. Zakłada się tutaj, iż w klasie „dobrej” ze względu na wysokie wyniki egzaminacyjne łatwiej jest nauczać, a atmosfera sprzyja nauce, co odbija się na indywidualnych wynikach (por. Dolata 2009). Hipoteza ta zostanie sprawdzona na polskich danych egzaminacyjnych z 2010 roku². Zmienną zależną będzie wynik egzaminu gimnazjalnego z obydwu części, a zmiennymi niezależnymi sprawdzian po szkole podstawowej (efekt indywidualny) i średnia sprawdzianu w klasie (efekt kontekstowy). Dane te zostały przeanalizowane trzema zaprezentowanymi modelami. W tabeli 1. Można odnaleźć wyniki dla trzech analiz odzienie dla części humanistycznej (hum) i matematyczno-przyrodniczej (mat). W tabeli podane zostały współczynnik regresji, wszystkie efekty są istotne statystycznie na poziomie $p \leq 0,01$. W regresji z korektą na rzetelność przyjęta została wartość 0,8³.

Tabela 1. Wyniki modelowania efektu rówieśnika dla trzech metod

Efekt	Regresja bez korekty		Regresja z korektą		Regresja z PV	
	hum	mat	hum	mat	hum	mat
Indywidualny	0,764	0,709	0,995	0,9230	0,753	0,683
Kontekstowy	0,191	0,146	-0,039	-0,067	0,185	0,189

Jak widać regresja bez korekty i regresja z PV wskazują na umiarkowanie silny efekt kontekstowy wahający się od 0,15 do 0,19 w zależności od zastosowanej metody i egzaminu gimnazjalnego. Natomiast regresja z korektą na rzetelność wskazuje na niewielki, ale ujemny efekt kontekstowy. W przypadku efektu indywidualnego regresja bez korekty i z PV spójnie wskazują na wartość w okolicach 0,7. Natomiast regresja z korektą na rzetelność estymuje znacznie większy efekt.

Jeżeli wyniki te porówna się z analizami symulacyjnymi okazuje się iż dają one spójny obraz. Regresja bez korekty i regresja z PV zawyżają efekt kontekstowy, podczas gdy regresja z korektą na rzetelność zawyża efekt indywidualny. Wyniki symulacji wskazują na to iż chcąc oszacować efekt kontekstowy najmniej obciążone oszacowanie zapewnia regresja z korektą na rzetelność. W badanym przypadku efekt rówieśników okazał się być ujemny. Jednak co widać na rysunku 4. przy rzetelności 0,8 regresja z korektą na rzetelność może lekko niedoszacowywał prawdziwego efektu. Najbezpieczniejszym wnioskiem płynącym z tych przedstawionych analiz jest następujący: efekt rówieśnika w polskiej szkole jest bliski zera, czyli nie wystę-

² Dane dla całej populacji z wyłączeniem klas mniej licznych niż 8 uczniów i klas liczniejszych niż 40 uczniów.

³ Centralna Komisja Egzaminacyjna raportuje znacznie wyższe wartości rzetelności sprawdzianu po szkole podstawowej – lecz szacuje je błędnie, traktując pytania w jednej wiązce zadań jako osobne zadania.

puje. W klasach o wyższych wynikach na wejściu uczniowie nie zyskują znacząco więcej niż ich koledzy w klasach słabszych.

Bibliografia

1. Dolata, R. (2009). *Szkola - segregacje - nierówności*. Warszawa: Wydawnictwo Uniwersytetu Warszawskiego.
2. Little, R. J. A., Rubin, D. B. (1987). Statistical analysis with missing data.
3. Maddala, G. S. (1986). *Limited-dependent and qualitative variables in econometrics*. Cambridge University Press.
4. Marsh, H. W., Hau, K. T. (2003). Big-Fish-Little-Pond Effect on Academic Self-Concept: A Cross-Cultural (26-Country) Test of the Negative Effects of Academically Selective Schools. *American Psychologist*, 58, 364-376.
5. Mislevy, R. J., Johnson, E. G., Muraki, E. (1992). Scaling procedures in NAEP. *Journal of Educational and Behavioral Statistics*, 17, 131.
6. Mislevy, R. J., Beaton, A. E., Kaplan, B., Sheehan, K. M. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement*, 29, 133-161.
7. Mislevy, R. J., Gitomer, D. H. (1995). The role of probability-based inference in an intelligent tutoring system. *User Modeling and User-Adapted Interaction*, 5, 253-282.
8. Mislevy, R. J., Almond, R. G., Yan, D., Steinberg, L. S. (1999). Bayes nets in educational assessment: Where do the numbers come from. (ss. 437-446).
9. Snijders, T. A. B. Bosker, R. J. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. SAGE publications Ltd.