Poprzedni vs nowy model obliczania EWD

Podobieństwa i różnice

Bartosz Kondratek

everythingthatcounts@protonmail.com

Spis treści

Tabelaryczne podsumowanie różnic między poprzednim a nowym modelem EWD

Aspekt modelu	Poprzednie EWD	Nowe EWD	Konsekwencje zmian
Model IRT dla zadań ocenianych dwupunktowo (0-1)	2PLM (dwuparametryczny model logistyczny)	3PLM (trójparametryczny model logistyczny)	• Uwzględnienie efektu zgadywania odpowiedzi poprawnych w przypadku trudniejszych zadań zamkniętych. • Lepsze dopasowanie modelu pomiarowego do danych. • Bardziej adekwatne oszacowanie poziomu umiejętności na wejściu oraz na wyjściu.
Sposób szacowania poziomu umiejętności uczniów	EAP (punktowe oszacowanie *expected a posteriori*)	PV (warunkowane 'wartości możliwe', *plausible values*)	• Uwzględnienie niepewności pomiaru poziomu umiejętności ucznia przy wyznaczaniu wskaźnika EWD. • Mniejsze obciążenie wszystkich statystyk wyznaczanych z oszacowania umiejętności, zwłaszcza – wskaźnika EWD. • Efektywniejsze wykorzystanie całej zebranej informacji, zwłaszcza – istotnej korelacji między wynikiem na wejściu a wynikiem na wyjściu, poprzez warunkowanie w modelu adekwatnym dla pomiaru powtarzanego.
Postać modelu regresji będącego podstawą szacowania EWD	Dwupoziomowa regresja trzeciego stopnia bez średnich szkoły na wejściu	Dwupoziomowa regresja trzeciego stopnia uwzględniająca średnie szkoły na wejściu	• Brak korelacji między średnią szkoły na wejściu a oszacowaniami EWD na poziomie populacji. • Kontrola zmiennych egzogenicznych. • Mniejsza wariancja wskaźnika EWD na poziomie populacji (więcej wariancji poddanych kontroli statystycznej)

Model IRT dla zadań ocenianych dwupunktowo

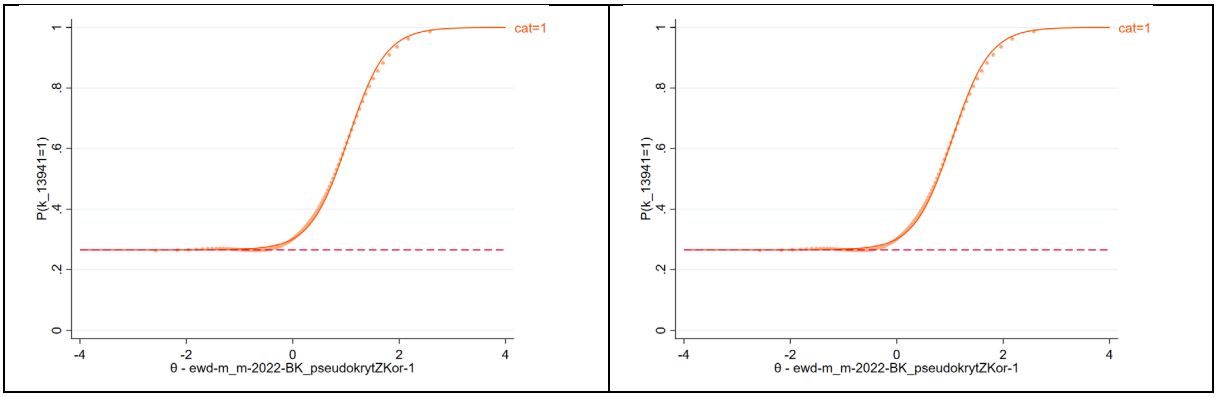
Zarówno poprzednie podejście do szacowania EWD, jak i nowe rozwiązanie – do wyznaczenia poziomu umiejętności ucznia na wejściu oraz na wyjściu, odwoływało się do modeli osadzonych w ramach IRT (*Item Response Theory*, teoria odpowiedzi na zadanie testowe).

Podstawowym „klockiem", z którego zbudowano każdy model IRT, jest funkcja określająca postać zależności między poziomem umiejętności ucznia (θ) a prawdopodobieństwem uzyskania określonego wyniku w zadaniu testowym.

W dotychczasowym modelu EWD wszystkie zadania, za które można było uzyskać 0–1 punktu, były modelowane dwuparametryczną krzywą logistyczną (*two-parameter logistic model*, 2PLM).

Cechą modelu 2PLM jest to, że wraz ze spadkiem poziomu umiejętności prawdopodobieństwo odpowiedzi na zadanie spada do zera. Stanowi to istotne ograniczenie modelu w sytuacji, gdy chcemy nim scharakteryzować odpowiedzi na zadanie zamknięte, zwłaszcza – na trudne zadanie zamknięte. Uczniowie o niskim poziomie umiejętności, rozwiązujący zadania tego typu, mają niezerową szansę na uzyskanie punktu: poprawną odpowiedź mogą odgadnąć. Rozwiązaniem tego problemu jest zastosowanie bardziej ogólnego, trójparametrycznego modelu logistycznego (3PLM), w którym dodatkowy parametr odpowiada za wyznaczenie z danych owej niezerowej dolnej asymptoty wykresu funkcji charakteryzującej zadanie.

Na poniższym wykresie zestawiono krzywe charakterystyczne dla jednego z zadań matury z matematyki na poziomie podstawowym z 2022 roku, uzyskane z wykorzystaniem 2PLM (z lewej) oraz 3PLM (z prawej). Linia ciągła opisuje krzywą charakterystyczną odpowiadającą parametrom IRT oszacowanym w ramach modeli, natomiast punkty naniesione na wykresy odpowiadają proporcjom poprawnych odpowiedzi w kolejnych 100 przedziałach umiejętności. Zgodność punktów z krzywą charakterystyczną pozwala na ocenę jakości dopasowania modelu do danych – jak widać, model 3PLM o wiele lepiej tłumaczy nasze dane: różnica jest widoczna zwłaszcza w odniesieniu do uczniów o niższym poziomie umiejętności.



Sposób szacowania poziomu umiejętności uczniów

Przejście z wyliczania EWD opartego na oszacowaniach punktowych EAP na wyliczanie EWD oparte na warunkowanym PV stanowiło bardzo istotną zmianę metodologiczną. Aby ją wyjaśnić, należy przyjrzeć się bliżej sposobowi, w jaki w IRT wnioskujemy o poziomie umiejętności ucznia.

Jeżeli dysponujemy:

1) wektorem odpowiedzi na zadania testu: $\boldsymbol{u} = (u_1, \dots, u_I)$,
2) wspomnianymi „klockami", z których zbudowano model – krzywymi określającymi prawdopodobieństwo zaobserwowania określonego wyniku u_i za każde zadanie i w zależności od poziomu umiejętności θ: $f_i(u_i, \theta)$,
3) rozkładem *a priori* umiejętności w populacji: $g(\theta)$,

możemy wyznaczyć tzw. rozkład *a posteriori* umiejętności ucznia:

$$\mathrm{P}(\theta|\boldsymbol{u}) = \frac{\prod_{i=1}^{I} f_i(u_i, \theta)\, g(\theta)}{\int \prod_{i=1}^{I} f_i(u_i, \theta)\, g(\theta) d\theta}$$

Taki rozkład *a posteriori* podsumowuje całą informację o poziomie umiejętności ucznia, a przy tym uwzględnia zaobserwowane dane (odpowiedzi na zadania: u_i), model pomiarowy (funkcje f_i) dopasowany do danych oraz informację o rozkładzie umiejętności w populacji ($g(\theta)$).

Estymator punktowy EAP, stosowany w poprzednim modelu EWD, uzyskuje się w wyniku pobrania wartości oczekiwanej z rozkładu *a posteriori* (stąd jego anglojęzyczna nazwa – *expected a posteriori*):

$$\hat{\theta} = E(\mathrm{P}(\theta|\boldsymbol{u}))$$

Gdy do dalszych analiz stosuje się oszacowanie punktowe EAP, zamiast całego rozkładu *a posteriori*, należy mieć na względzie, że:

1) Gubiona jest wszelka informacja o niepewności pomiaru. Dla niektórych uczniów pomiar jest wyznaczony z większą precyzją (rozkład *a posteriori* jest bardziej skoncentrowany wokół wartości oczekiwanej), a dla innych – z mniejszą (bardziej rozproszony rozkład *a posteriori* – ma to miejsce zwłaszcza dla najniższych oraz dla najwyższych wyników w teście).
2) Estymator jest obciążony w stronę średniej populacyjnej oraz wielkość tego obciążenia okazuje się tym większa, im większy staje się błąd pomiaru. Zatem obciążenie jest nieliniową funkcją prawdziwego poziomu umiejętności, a jego wartość zależy od unikalnych właściwości psychometrycznych danego testu, z generalną tendencją do przyjmowania wyższych wartości dla skrajnych poziomów umiejętności.

Ze względu na powyższe ograniczenia estymatora EAP w kontekście wnioskowania statystycznego, w międzynarodowych badaniach edukacyjnych (PISA, TIMSS, PIRLS) wykorzystywane są – zamiast punktowych oszacowań umiejętności – tzw. wartości możliwe (*plausible values,* PV), których uzasadnienie wynika z metodologii wielokrotnych imputacji (*multiple imputations*).

Pojedyncza wartość możliwa jest po prostu realizacją wylosowaną z rozkładu *a posteriori*:

$$\mathrm{PV}_m \sim \mathrm{P}(\theta|\boldsymbol{u})$$

a wnioskowanie statystyczne z wykorzystaniem PV sprowadza się do przeprowadzenia analizy wielokrotnie na zbiorze niezależnie wylosowanych wartości możliwych ($\mathrm{PV}_1, \dots, PV_M$) i do odpowiedniego uśrednienia wyników oraz błędów standardowych.

Intuicja podpowiada, że wskutek takiego podejścia możliwe staje się uwzględnienie niepewności pomiaru (dla mniej rzetelnych oszacowań umiejętności PV będą bardziej zróżnicowane). Możemy też zauważyć, że – jeżeli skupić się na pojedynczym uczniu – uśrednienie wielu PV będzie zmierzało do wartości uzyskiwanej za pomocą estymatora EAP.

Poprawne zastosowanie metodologii opartej na PV wymaga jednak, aby przy ich generowaniu została odpowiednio uwzględniona informacja o wszystkich zmiennych niezależnych, które zostaną wykorzystane w dalszej analizie. Ten zabieg, nazywany warunkowaniem, sprowadza się do rozbudowania fragmentu modelu pomiarowego określającego rozkład *a priori* o odpowiednio skonstruowane równanie regresji latentnej. Zamiast ogólnego populacyjnego rozkładu *a priori*, przy warunkowaniu pojawia się rozkład – przewidywanego zgodnie z regresją latentną – poziomu umiejętności dla ucznia/uczniów o określonej konfiguracji zmiennych niezależnych.

W odniesieniu do modelu EWD to zadanie okazywało się nieoczywiste. Aby uchwycić wszystkie potencjalnie istotne zależności między zmiennymi wchodzącymi w skład modelu będącego podstawą wyliczania EWD, warunkowanie przeprowadzono z wykorzystaniem trójpoziomowej regresji latentnej, określonej globalnie, tj. przy jednoczesnym uwzględnieniu zarówno danych na wejściu, jak i danych na wyjściu:

$$\theta_{jkl} = \sum_{n=1}^{N} \gamma_n \cdot {}^{n}x_k + \sum_{n=1}^{N} \gamma_{N+n} \cdot j \cdot {}^{n}x_k + \varepsilon_l + j \cdot \varepsilon_l^* + \varepsilon_{kl} + \varepsilon_{jkl}$$

gdzie poziom umiejętności θ_{jkl} uzyskany na egzaminie j ($j \in \{0,1\}$), przez ucznia k, uczęszczającego na wyjściu do szkoły l, jest wyjaśniany poprzez uwzględnienie:

- efektów stałych, $\gamma_1, \dots, \gamma_{2N}$, zmiennych niezależnych ${}^{1}x, \dots, {}^{n}x$ (płeć, zaświadczenia o dysleksji) z możliwością interakcji każdej zmiennej z egzaminem;

- losowego efektu odchylenia średniej szkoły l od średniej globalnej, ε_l, oraz dodatkowego losowego efektu odchylenia średniej szkoły na egzaminie na wyjściu, ε_l^*; para tych zmiennych jest elementem modelu odzwierciedlającym istotną wariancję międzyszkolną w poziomie umiejętności uczniów oraz możliwość jej zróżnicowania na wejściu oraz na wyjściu – co okazuje się bardzo istotne w kontekście EWD; w strukturze modelu regresji dopuszczono możliwość skorelowania efektów losowych ε_l oraz ε_l^*;
- losowego odchylenia średniego poziomu umiejętności ucznia k w szkole l od średniej globalnej przy uwzględnieniu pozostałych zmiennych, ε_{kl}; ta zmienna pozwala na uchwycenie istotnej korelacji między wynikami tych samych uczniów na wejściu i na wyjściu;
- losowego składnika resztkowego, ε_{jkl}, odpowiadającego za niewyjaśnioną – z wykorzystaniem pozostałych zmiennych – część wariancji umiejętności ucznia k w szkole l na egzaminie j.

Wyznaczenie wartości możliwych zgodnie z założonym modelem, czyli z modelem określonym jednocześnie dla danych z egzaminu na wejściu i z egzaminu na wyjściu, technicznie przeprowadzono przez rozbicie problemu na dwa kroki:

1) W pierwszym etapie, niezależnie dla egzaminu na wejściu i dla egzaminu na wyjściu, dopasowano wielogrupowe modele IRT, podobnie jak rozwiązywano to w poprzednim podejściu do EWD – z jedyną różnicą, że zamiast modelu 2PLM zastosowano model 3PLM.
2) W drugim etapie dla połączonych danych z wejścia i z wyjścia wygenerowano wartości możliwe zgodnie z przyjętym trójpoziomowym poziomem regresji przy zakotwiczeniu parametrów modelu IRT na wartościach wyznaczonych w kroku pierwszym. Losowanie wartości możliwych dla znacznych zbiorów danych i dla skomplikowanej funkcji warunkującej wymagało stworzenia specjalnego oprogramowania, efektywniej replikującego metodę generowania wartości możliwych, zaimplementowaną w oprogramowaniu do analiz IRT: UIRT.

Postać modelu regresji będącego podstawą szacowania EWD

Ostatnią istotną zmianą w podejściu do wyznaczania EWD stało się włączenie – do dwupoziomowej regresji będącej podstawą szacowania efektu EWD – efektów dla średniej szkoły na wejściu.

W poprzednim podejściu EWD wyznaczano z równania regresji dwupoziomowej – co można opisać następująco:

$$\hat{\theta}_{1kl} = w^z(\hat{\theta}_{0kl}) + \sum_{n=1}^{N} \gamma_n \cdot {}^n x_k + \varepsilon_l + \varepsilon_{kl}$$

Gdzie $\hat{\theta}_{1kl}$, czyli punktowe oszacowanie umiejętności na wyjściu uzyskane przez ucznia k, uczęszczającego na wyjściu do szkoły l, jest wyjaśniane poprzez uwzględnienie:

- efektów stałych punktowego oszacowania umiejętności tego ucznia na wejściu, $\hat{\theta}_{0kl}$, w potęgach do stopnia z (zazwyczaj $z = 3$),
- efektów stałych, $\gamma_1, \dots, \gamma_N$, zmiennych niezależnych ${}^1x, \dots, {}^n x$,
- oraz efektu losowego na poziomie szkoły, ε_l, utożsamianego z EWD; estymator BLUP efektu ε_l był w poprzednim podejściu oszacowaniem EWD.

Natomiast analogiczny model regresji dla nowego podejścia do obliczania EWD ma następującą postać dla każdej ${}^m\mathrm{PV}$:

$${}^m\mathrm{PV}_{1\mathrm{kl}} = w^z({}^m\mathrm{PV}_{0\mathrm{kl}}) + w^{*z}(\overline{{}^m\mathrm{PV}_{0\mathrm{l}}}) + \sum_{n=1}^{N} \gamma_n \cdot {}^n x_k + \varepsilon_l + \varepsilon_{kl}$$

EWD jest również szacowane przez jego estymator BLUP dla efektu ε_l (dla każdej PV niezależnie, a potem – uśredniane).

Najistotniejsza różnica polega natomiast na włączeniu – do równania regresji –efektów stałych dla dodatkowej zmiennej niezależnej: średniego wyniku na wejściu w szkole l, tj. zmiennej $\overline{{}^m\mathrm{PV}_{0\mathrm{l}}}$.

Gdy uwzględniamy – jako jedną ze zmiennych wyjaśniających – średnią szkoły na wejściu, uzyskujemy w modelu dwupoziomowym pełniejszą kontrolę wyników na wyjściu. Pozwala to na uniknięcie tzw. problemu zmiennych egzogenicznych w modelach hierarchicznych, którym było obciążone poprzednie podejście do wyznaczania EWD – co objawiało się istotną korelacją między średnią szkoły na wejściu a oszacowaniami EWD.

W nowym podejściu, dzięki włączeniu – do równania regresji – efektów dla średniej szkoły na wejściu, uzyskujemy w zasadzie zerową korelację między tymi średnimi a oszacowaniami EWD.

Jednak oznacza to również, że chociaż nowe oszacowania EWD kontrolują większą część wariancji międzyszkolnej wyników na wyjściu, to mają mniejszą wariancję niż w poprzednim podejściu. Wartości oszacowanych wskaźników EWD są średnio bliższe zeru oraz rzadziej przekraczają próg istotności statystycznej.