Artur Pokropek[1]
Instytut Filozofii i Socjologii Polskiej Akademii Nauk

Metody statystyczne wykorzystywane w szacowaniu trzyletnich wskaźników egzaminacyjnych

Wstęp

Trzyletni wskaźnik egzaminacyjny publikowany przez Centralną Komisję Egzaminacyjną został opracowany przez grupę odpowiedzialną za metodologię szacowania wskaźnika Edukacyjnej Wartości Dodanej[2]. Wskaźnik ten został przychylnie odebrany przez osoby zainteresowane poszerzaniem swojej wiedzy o pracy szkół. Forma graficzna, w jakiej przedstawiany jest wskaźnik, powoduje, iż jego pobieżna interpretacja jest intuicyjna, nie sprawia większych problemów i nie wymaga specjalistycznej wiedzy statystycznej, a jedynie wyobraźni i zdrowego rozsądku. Pod pozorami tej prostoty kryją się jednak wyrafinowane procedury statystyczne, starannie wybrane i przetestowane po to, by przedstawiany wynik był jak najbardziej dokładny, trafny oraz rzetelny. W artykule tym przybliżone zostaną kulisy szacowania omawianego wskaźnika na tyle dokładnie, aby pozwolić czytelnikowi odtworzyć i zrozumieć kolejne kroki analiz statystycznych używanych podczas szacowaniu wskaźnika trzyletniego.

Szacowanie omawianego wskaźnika egzaminacyjnego można podzielić na cztery etapy:

1. Normalizacja wyników egzaminacyjnych za pomocą metody ekwicentylowej.
2. Szacowanie wyników końcowych za pomocą modeli wielopoziomowych.
3. Szacowanie wskaźnika edukacyjnej wartości dodanej za pomocą modeli wielopoziomowych.
4. Przedstawianie wyników w dwuwymiarowej przestrzeni przy uwzględnieniu niepewności szacowania.

W kolejnych częściach tego artykułu przedstawiony zostanie każdy z tych kroków. Zaczniemy od zrównywania opartego na metodzie ekwicentylowej, jednak z uwagi na to, iż metoda ta jest stosunkowo dobrze znana w polskich badaniach edukacyjnych, poświęcone jej zostanie stosunkowo mało miejsca. Kluczowymi punktami, które zostaną bardziej szczegółowo opisane, są punkty drugi i trzeci. Stanowią one bowiem rdzeń szacowanego wskaźnika. Skupimy się tu na modelowaniu, za pomocą którego uzyskujemy wyniki końcowe oraz wyniki wskaźnika edukacyjnej wartości dodanej. Jako odwołanie do punktu

[1] artur.pokropek@gmail.com

[2] Do której należy autor tego tekstu. Bliższe informacje patrz: www.ewd.edu.pl

drugiego i trzeciego, poruszona zostanie również kwestia estymacji Bayesowskich, wykorzystywanych w szacowaniu trzyletnich wskaźników egzaminacyjnych. W punkcie czwartym omówiona zostanie koncepcja graficznego prezentowania wskaźników ze szczególnym uwzględnieniem przesłanek statystycznych wiążących się z takim sposobem prezentacji wyników.

1.

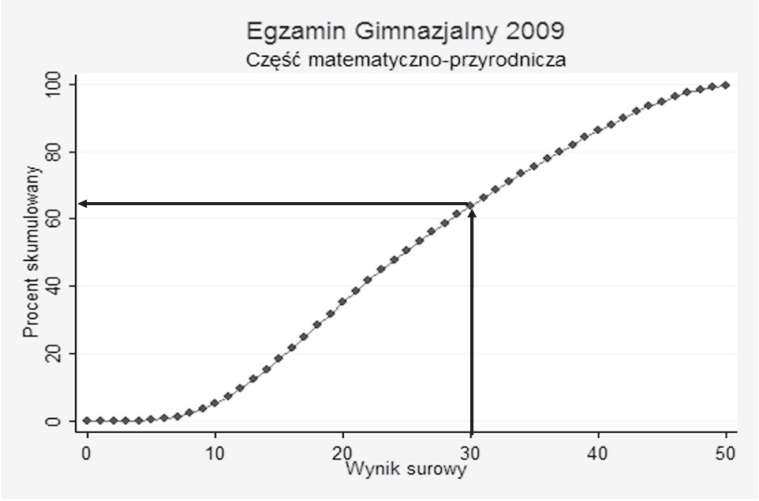
Przy normalizacji wyników za pomocą metody ekwicentylowej warto zdefiniować dwie funkcje. Niech *F(x)* będzie kumulatywną funkcją gęstości rozkładu zmiennej losowej *X* (reprezentującej wynik egzaminu), a funkcja *P(x)* funkcją wyznaczająca rangę pozycyjną[3] tak, że:

$$
\begin{aligned}
P(x) &= 100\{F(x^*-1)=[x-(x^*-0,5)][F(x^*)-F(x^*)-F(x^*-1)]\}; \\
&=0 \quad dla\ x < -0,5 \\
&=100 \quad dla\ x \geq max(x)+0,5
\end{aligned}
\qquad (1.1)
$$

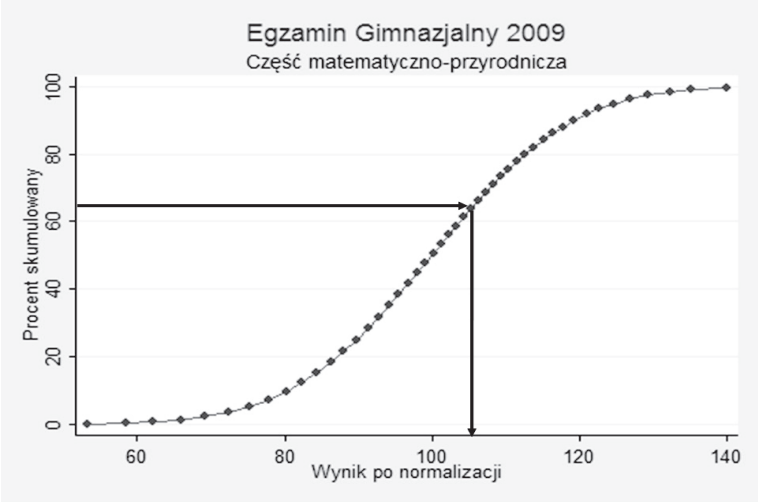
gdzie:
$-0,5 \leq x < \max(x) + 0,5$;
x^*: to najbliższa liczba całkowita wartości x tak, że:
$x^* - 0,5 <= x < x^* + 0,5$;
max(x): to maksymalna do uzyskania liczba punktów.

Funkcja *P(x)* zdefiniowania w równaniu (1.1) jest receptą na przekształcenie wyniku surowego na rangę pozycyjną, czyli dzięki niej można umieścić wynik na kumulatywnej skali wyników. Oznacza to zrelatywizowanie danego wyniku do pozycji w rozkładzie skumulowanym. Graficzna reprezentacja tej funkcji przedstawiona została na rysunku 1.1. Na krzywej wyrażającej kształt funkcji kwadratami zaznaczono również 51 punktów odpowiadających 51 możliwym do uzyskania wynikom egzaminacyjnym (przykład odwołuje się do części matematyczno-przyrodniczej egzaminu gimnazjalnego 2009; skala tego testu wynosi od 0-50). Na rysunku zaznaczono również przykładowe przekształcenie wyniku surowego równego 30 punktom. Można odczytać, że 63,8% uczniów uzyskało wynik 30 punktów lub gorszy.

[3] Dalsze informacje patrz: M. J. Kolen i R. L. Brennan, *Test Equating, Scaling and Linking Methods and Practices Second Edition,* Springer, New York 2004.



Rysunek 1.1. Funkcja wyznaczająca rangę pozycyjną dla egzaminu gimnazjalnego 2009, część matematyczno-przyrodnicza



Rysunek 1.2. Funkcja przekształcenia liniowego wyznaczająca wynik po normalizacji

Jeżeli podzielimy rozkład gęstości na 100 równych części (percentyle), wynik ten będzie miał rangę 64. Jeżeli jednak rozkład gęstości podzielimy na 1000 równych części, ranga będzie wynosić 638. Wybór dokładności, z jaką przenosimy wyniki surowe na rangi, nie jest bez znaczenia. Im mniejsza liczba podziałów rozkładu kumulatywnego, czyli mniejsza ilość rang, tym rozkład wyników wyrażonych w rangach będzie bardziej przypominał rozkład normalny. Ceną tego jest jednak dokładność szacowania wyników. Przy podziale rozkładu skumulowanego na 10

części będziemy mieli tylko 10 rang w stosunku do 51 punktów. W przypadku podziału rozkładu skumulowanego na 100 części uzyskujemy (w zależności od rodzaju i edycji egzaminu gimnazjalnego) około 40 unikalnych rang. Dzieje się tak dlatego, że rzadkie wyniki, szczególnie na skrajach rozkładów, „sklejają się ze sobą". Nawet podział rozkładu skumulowanego na 1000 części nie gwarantuje tego, że kategorie się nie skleją. Można to prześledzić na rysunku 1.1, gdzie rangi zdefiniowane są na podstawie podziału rozkładu skumulowanego na 1000 części. Pięć najniższych wyników egzaminacyjnych uzyskuje identyczną, najniższą rangę. Jednak właśnie ten podział okazał się najbardziej optymalny przy zachowaniu dostatecznej równowagi między dokładnością pomiaru a normalizacją.

Tak uzyskana ranga przedstawiona zostaje na skali o średniej 100 i odchyleniu standardowym 15. Aby przekształcić uzyskane rangi na taką skalę (tak jak jest to pokazane na rysunku 1.2), musimy odwołać się do liniowej funkcji przekształcającej uzyskaną rangę $p = P(x)$ na skalę o zadanych parametrach (czyli średniej 100 i odchyleniu standardowym 15). Możemy to zrobić, odwołując się do prostego przekształcenia liniowego, które możemy zapisać jako funkcję $L(p)$:

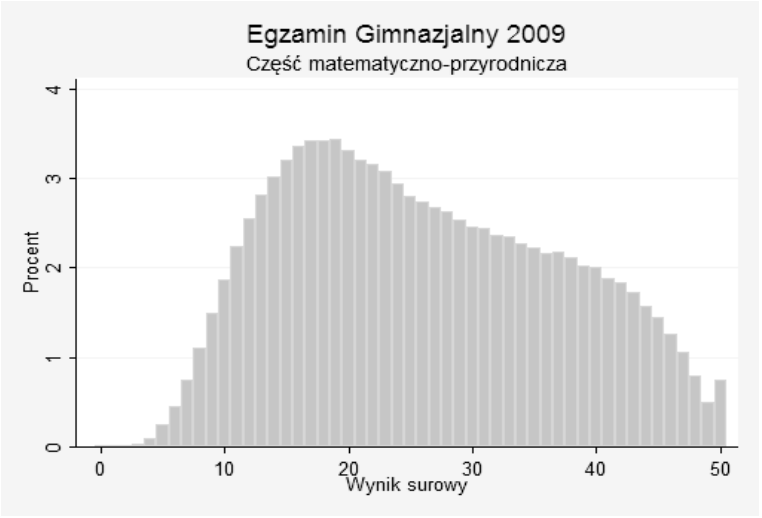
(1.2)

$$L(p) = \frac{15}{\sigma_p} p + \left(100 - \frac{15}{\sigma_p} \mu_p \right)$$

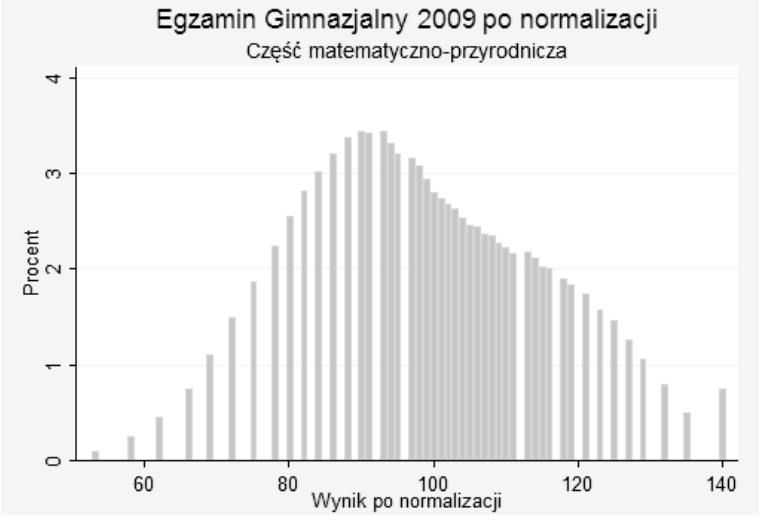
gdzie: σ_p to odchylenie standardowe wyniku wyrażonego w rangach p
μ_p to średnia wyniku wyrażonego w rangach p.

Graficzną reprezentację funkcji $L(p)$ można znaleźć na rysunku 1.2. Na tym rysunku można też zaobserwować przekształcenie wcześniej omawianej rangi 638 na skalę o średniej 100 i odchyleniu standardowym 15, w której ranga ta ma wartość 105,32.

Na rysunku 1.3 i 1.4 przedstawiono rozkłady egzaminu gimnazjalnego przed normalizacją i po niej. W tabeli 1.1 przedstawiono wartości punktowe przed normalizacją i po niej.



Rysunek 1.3. Rozkład egzaminu gimnazjalnego 2009, cześć matematyczno-przyrodnicza przed normalizacją



Rysunek 1.4. Rozkład egzaminu gimnazjalnego 2009, cześć matematyczno-przyrodnicza po normalizacji

Tabela 1.1 Wyniki surowe i po normalizacji egzaminu gimnazjalnego 2009, cześć matematyczno-przyrodnicza

WS	WN	WS	WN	WS	WN	WS	WN	WS	WN
0/1	53,32	**11**	77,92	**21**	95,53	**31**	106,34	**41**	117,81
2	53,32	**12**	80,28	**22**	96,78	**32**	107,35	**42**	119,34
3	53,32	**13**	82,39	**23**	97,97	**33**	108,34	**43**	121,00
4	53,32	**14**	84,40	**24**	99,12	**34**	109,38	**44**	122,73
5	58,49	**15**	86,34	**25**	100,21	**35**	110,46	**45**	124,68
6	62,05	**16**	88,07	**26**	101,28	**36**	111,55	**46**	126,97
7	65,90	**17**	89,75	**27**	102,30	**37**	112,64	**47**	129,33
8	69,28	**18**	91,36	**28**	103,31	**38**	113,87	**48**	132,37
9	72,43	**19**	92,84	**29**	104,32	**39**	115,13	**49**	135,12
10	75,29	**20**	94,25	**30**	105,32	**40**	116,44	**50**	140,04
WS	Wynik surowy								
WN	Wynik znormalizowany								

2.

Do szacowania wyników końcowych danego gimnazjum, czyli wyniku średniego z trzech kolejnych lat, wykorzystywane jest modelowanie wielopoziomowe. Stosuje się tutaj tak zwany „model pusty", w którym wykorzystywane są informacje o wynikach poszczególnych uczniów oraz informacja o pogrupowaniu ich w szkołach. Aby przywołać informacje o pogrupowaniu uczniów do klasycznego równania opisującego średni wynik uczniów w populacji z uwzględnieniem błędu pomiaru:

$$y_i = \beta_0 + e_1 \qquad (2.1)$$

gdzie: y_i to wynik i-tego ucznia
β_0 to średnia wyników w populacji
e_i wyraz błędu,

Wprowadzamy indeks *j* – oznaczający szkołę *j*.
Równanie wygląda wtedy następująco:

Poziom jednostki (1): $y_{ij} = \beta_{0j} + r_{1j}$ (2.1)
gdzie:

Poziom szkoły (2): $\beta_{0j} = \gamma_{00} + u_{oj}$

Gdy wprowadziliśmy informację o pogrupowaniu, analiza stała się analizą, w której bierzemy pod uwagę dwa poziomy rzeczywistości - poziom jednostek (uczniów) i poziom grup, do których owe jednostki należą (poziom szkół). Można to zapisać w jednym równaniu. Równanie to podzielmy na dwie części: tak zwaną część stałą i część losową, gdzie część losowa zamknięta została w nawiasie kwadratowym:

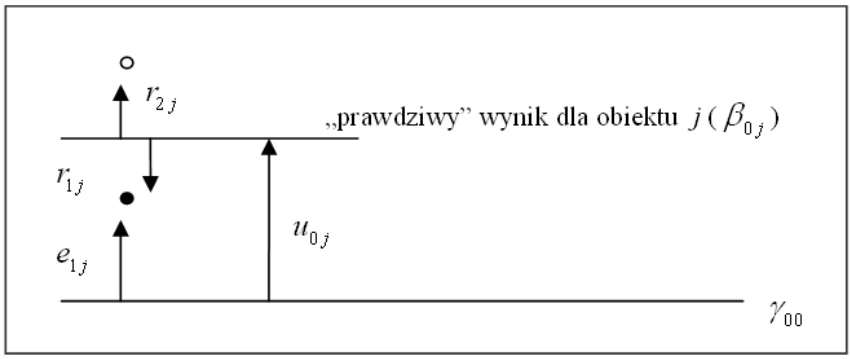
$$y_{ij} = \gamma_{00} + [u_{oj} + r_{1j}] \tag{2.2}$$

Model 2.2 to prosty, dwupoziomowy model, który został wykorzystany do obliczania wskaźnika wyniku egzaminacyjnego danej szkoły, gdzie *i*-te pomiary są obiektami pierwszego poziomu, a dane jednostki *j* są obiektami z drugiego poziomu (szkoła); γ_{00} oznacza tutaj wartość oczekiwaną dla populacji (efekt stały – niezmienny w kontekście całej populacji), natomiast β_{0j} jest parametrem składającym się z części stałej γ_{00} charakteryzującej wynik dla populacji, oraz części losowej - opisującej zróżnicowanie parametrów między grupami, wyrażanej za pomocą u_{0j}. Czasami, szczególnie w teoriach testów, parametr ten określa się jako wynik *prawdziwy* lub zmienną *ukrytą*, świadczącą o *prawdziwym* poziomie umiejętności, *prawdziwej* wartości inteligencji czy *prawdziwej* wartości, którą to próbujemy odkryć za pomocą zawodnych narzędzi. Całkowity błąd przewidywania e_i znany z równania (2.1) można teraz przedstawić jako sumę błędu na poziomie jednostkowym r_{ij} oraz grupowym u_j tak, że:

$$e_i = r_{ij} + u_j$$

(Zostało to przedstawione na rysunku 2.1)

Zakładamy tutaj, że r_{ij} ma rozkład normalny w populacji pomiarów (uczniów). Ponadto u_j, jako odchylenie od populacyjnej średniej, dla *j*-tych przedmiotów, ma rozkład normalny w całej populacji „wyników prawdziwych" o średniej zero i odchyleniu standardowym τ_{00} .



Rysunek 2.1. Relacje między błędami, efektem stałym w populacji i wynikiem prawdziwym dla obiektu *j*

Założenia są zatem analogicznie jak w klasycznej regresji liniowej z tym, że w rozbiciu na dwa poziomy. Ponadto zakłada się, że błędy są od siebie niezależne:

$$E(u_i r_{ij}) = 0 \tag{2.3}$$

$$E(u_i u_j) = 0 \quad (i \neq j) \tag{2.4}$$

$$E(r_{it} r_{is}) = E(r_{it} r_{jt}) = E(r_{it} r_{js}) = 0 \quad (i \neq j; t \neq s) \tag{2.5}$$

Pomiary y_{1j}, y_{2j} są niezależne od siebie w grupie pomiarów dla grupy. Zależność ich widziana jest dopiero na wyższym poziomie, gdzie patrzymy na wszystkie wyniki zagnieżdżone w grupach, do których przynależą jednostki obserwacji, ich zależność odzwierciedlana jest przez błąd u_i. Dlatego wariancję wartości y_{ij} można zapisać następująco, co również obrazowo przedstawia rysunek 2.1:

$$Var(y_{ij}) = Var(\gamma_{00} + u_j + r_{ij}) = Var(u_j + r_{ij}) = \tau_{00} + \sigma^2 \tag{2.6}$$

Po co wprowadzać takie komplikacje, zamiast liczyć średnią w klasyczny sposób dla poszczególnych szkół i traktować ją jako wartość oczekiwaną? Odpowiedź jest prosta: średnią traktujemy jako wartość oczekiwaną pracy szkoły i ważnym aspektem staje się dla nas estymowanie błędu szacowania. Klasyczne szacowanie wartości oczekiwanej (poprzez średnią) i błędów standardowych nie jest w tym przypadku optymalne. Klasyczne szacowanie wartości oczekiwanej jest optymalne tylko warunkowo, to znaczy wtedy, gdy odnosi się do jednej szkoły. Gdy estymujemy średnią standardową dla grupy szkół, najbardziej optymalne okazują się Bayesowskie predykcje[4] a posteriori. Predykcje te w przypadku oszacowań punktowych dają minimalną kwadratową funkcję błędu oszacowania i mówi się o nich, iż są BLUP, czyli *Best Linear Unbiased Predictors* (najlepsze liniowe nieobarczone błędem predykcje) dla całej populacji. Mają też optymalne, czyli najniższe poprawne, błędy standardowe, dzięki którym wyznaczane są przedziały ufności.[5] Wszystko to za cenę warunkowego obciążenia. Predykcje te wykazują pewien konserwatywny (czyli ukierunkowany w stronę średniej) błąd na poziomie szkół dla placówek o niewielkiej liczbie uczniów i/lub skrajnych wartościach średniej (b_{0j}).

Podstawową ideą, jaka kryje się za Bayesowskimi metodami wyniku końcowego, jest użycie dwóch typów informacji: (1) o jednostkach należących do *j-tej* grupy oraz (2) informacji o całym rozkładzie u_j (czy de facto o rozkładzie średniej: $\beta_{0j} = \beta_0 + u_j$) w całej populacji: średniej równej 0 i odchyleniu standardowym τ_{00}.

[4] Konwencjonalnie o estymatorach Bayesowskich mówi się jako o predyktorach.

[5] Rabe-Hesketh, Sophia i Anders Skrondal, *Multilevel and Longitudinal Modeling Using Stata*. College Station, Texas: Stata Press Publication – StataCorp LP. 2008; s. 82.

Klasyczny estymator wartości oczekiwanej dla grupy *j*, a w języku modelowania wielopoziomowego – stała regresji wielopoziomowej ($\beta_{0j} = \beta_0 + u_j$), to oczywiście klasycznie liczona średnia[6]:

$$\hat{\beta}_{0j} = \sum_{i=1}^{n} y_{ij} / n_j \tag{2.7}$$

Zaś estymator dla całej populacji, czyli średnia populacyjna, to oczywiście:

$$\hat{\gamma}_{00} = \sum_{i=1}^{N} \frac{n_j}{M} \hat{\beta}_{0j} \tag{2.8}$$

Gdzie n_j oznacza liczebność *j*-tej grupy, *N* to liczebność badanych grup; *M* to całkowita liczebność populacji. Estymator Bayesowski jest średnią ważoną dwóch estymatorów: grupowego i populacyjnego tak, że[7]:

$$\hat{\beta}_{0j}^{EB} = \lambda_j \hat{\beta}_{0j} + (1 - \lambda_j)\hat{\gamma}_{00} \tag{2.9}$$

gdzie:

$$\lambda_j = \frac{\tau_{00}^2}{\tau_{00}^2 + \sigma^2 / n_j} \tag{2.10}$$

Jak widać zatem z równań (2.9 i 2.10), na różnicę między estymatorem klasycznym i predyktorem Bayesowskim dla *j*-tej grupy wpływają 4 elementy:

1. Zróżnicowanie predyktorów grupowych w całej populacji charakteryzowane przez τ^2_{00}
2. Zróżnicowanie wewnątrzgrupowe:σ^2
3. Wielkość grupy: n_j
4. Różnica między średnią w populacji a klasycznym estymatorem wyników końcowych w grupie.

W praktyce dwa ostatnie punkty mają największy wpływ na różnicę między oszacowaniem klasycznym a Bayesowskim. Czyli liczebność grupy *j* (w naszym wypadku szkoły) oraz jej odległość od średniej populacyjnej. Im szkoła mniejsza, a wynik jej bardziej oddalony od średniej populacyjnej, tym korekta związana

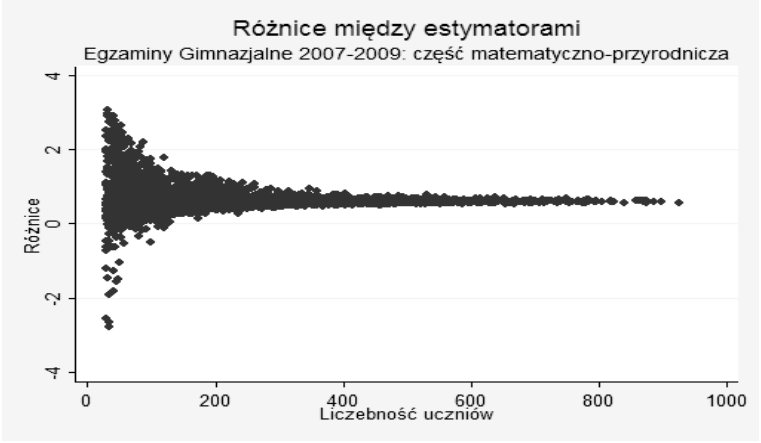
[6] W tym wypadku średnia jest też tożsama z estymatorem MLE.

[7] Snijders, Tom A.B. i Roel J. Bosker, *Multilevel Analysis*. Thousand Oaks – London – New Delhi: Sage 2004, s. 56-66.

z użyciem predyktora Bayesowskiego jest większa. Będąc bardziej dosłownym, można powiedzieć, iż w wypadku małych i oddalonych od średniej populacyjnej szkół predyktor Bayesowski będzie „ściągał" wyniki ku średniej. Przedstawione to zostało na rysunkach 2.2 i 2.3. Na pierwszym z tych rysunków widać, że największe różnice między klasycznym estymatorem a predyktorem Bayesowskim dotyczą szkół, w których w przeciągu 3 lat uczyło się mniej niż 100 uczniów. Poniżej tego progu różnice niezwykle maleją po to, by powyżej 400 jednostek obserwacji stać się marginalnymi i oscylować wokół 0,5 punktu[8] (na skali średnia 100, odchylenie standardowe 15). Natomiast na rysunku 2.3 widać, iż różnice między klasycznym estymatorem a predyktorem Bayesowskim rosną wraz z oddalaniem się wyników szkół od średniej populacyjnej (100).

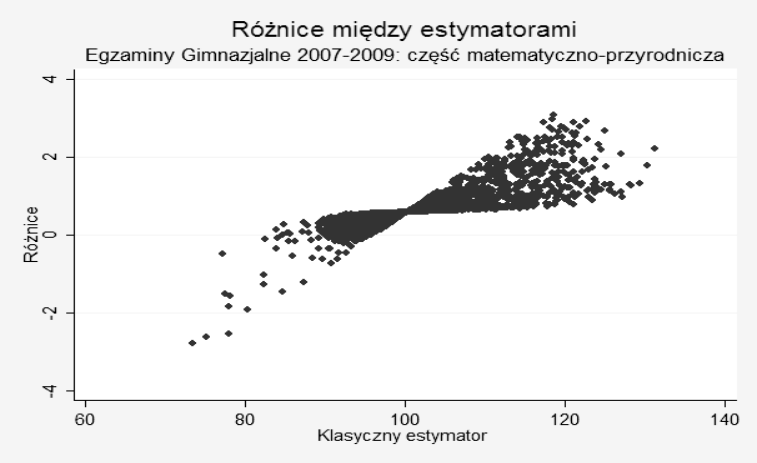
Największy zysk ze stosowania estymatorów Bayesowskich to korekcja błędów standardowych oszacowań. Błędy standardowe predyktorów Bayesowskich można zapisać w następujący sposób:

$$S.E(\hat{\beta}_{0j}^{EB}) = \frac{1}{\sqrt{\tau_{00}^{-2} + \sigma^{-2} n_j}} \tag{2.11}$$

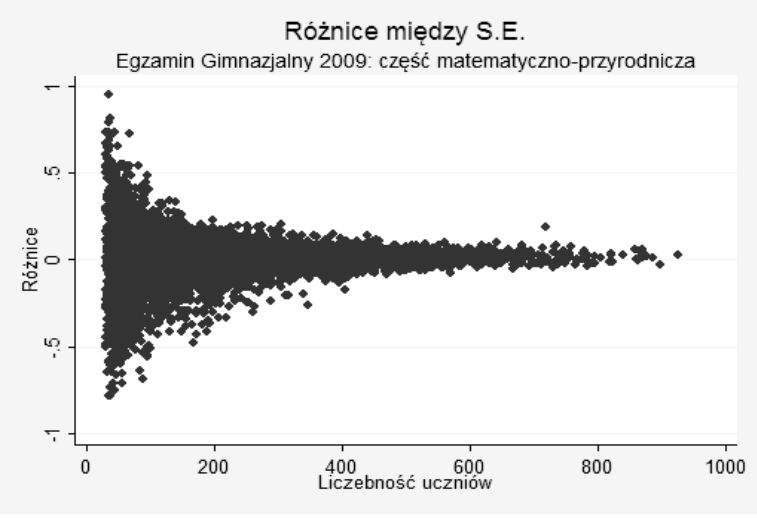


Rysunek 2.2. Różnice między estymatorem klasycznym i Bayesowskim a liczebnością grupy

[8] Różnice oscylują wokół punktu 0,5 a nie 0,0, ponieważ prezentowane poprawki odnoszą się do szkół o liczebności +30. Natomiast rozkład predyktorów grupowych w populacji, a zatem τ_{00} estymowana jest dla wszystkich szkół. Ponieważ wielkość szkoły jest skorelowana z wynikami, średnia różnic wyższa jest od zera.



Rysunek 2.3. Różnice między estymatorem klasycznym i Bayesowskim a porównanie dla różnych wyników estymatora klasycznego



Rysunek 2.4. Różnice między klasycznym błędem standardowym i błędem standardowym dla predyktora Bayesowskiego a liczebność uczniów

Rysunek 2.5. Różnice między klasycznym błędem standardowym i błędem standardowym dla predyktora Bayesowskiego a zróżnicowanie wyników

W wypadku błędów standardowych, oprócz informacji o rozkładzie wyników we wszystkich szkołach, korekta błędów zależy przede wszystkim od dwóch czynników: wielkości szkoły i zróżnicowania wyników uzyskiwanych przez uczniów w szkołach. Pokazują to rysunki 2.4 i 2.5. Na rysunku 2.4 widać wyraźnie, iż największe korekty (w obie strony) dotyczą szkół, w których mamy stosunkowo niewiele obserwacji (poniżej 100). Im szkoły liczniejsze, tym korekty stają się coraz bardziej marginalne. Rysunek 2.5 pokazuje, jak korekty Bayesowskie zależą od zróżnicowania wyników wewnątrz szkoły. Gdy zróżnicowanie mierzone odchyleniem standardowym zbliża się do zróżnicowania średniego w populacji, korekty są niewielkie. Gdy zróżnicowanie w szkołach jest wyższe niż średnie zróżnicowanie w populacji, korekta jest dodatnia (czyli błąd standardowy predykcji Bayesowskich jest mniejszy niż klasyczny błąd standardowy), gdy zróżnicowanie w szkołach jest niższe niż średnia w całej populacji, korekta jest ujemna (czyli błąd standardowy predykcji Bayesowskich jest większy niż klasyczny błąd standardowy). Podobnie jak same predyktory, Bayesowskie błędy są optymalne, czyli w sposób optymalny minimalizują popełnienie błędu podczas wnioskowania statystycznego.

3.

Naturalnym rozszerzeniem dla modelu zaprezentowanego wcześniej, w którym szacowany był predyktor wyników końcowych, jest dodanie zmiennej niezależnej *x*' lub wektora zmiennych niezależnych, opisującego związek liniowy ze zmienną *Y* parametrem β_{1j} lub wektorem parametrów: β. Na model taki można patrzeć jako na rozszerzenie modelu pustego o zmienną wyjaśniającą lub rozszerzenie klasycznej regresji jednej zmiennej o informacje dotyczące

pogrupowania jednostek. Model z efektem losowym dla stałej regresji zapisujemy następująco (zapis dla jednej zmiennej niezależnej):

Poziom jednostki (1): $$y_{ij} = \beta_{0j} + \beta_{1j} x_{ij} + r_{ij} \quad (3.1)$$

gdzie:

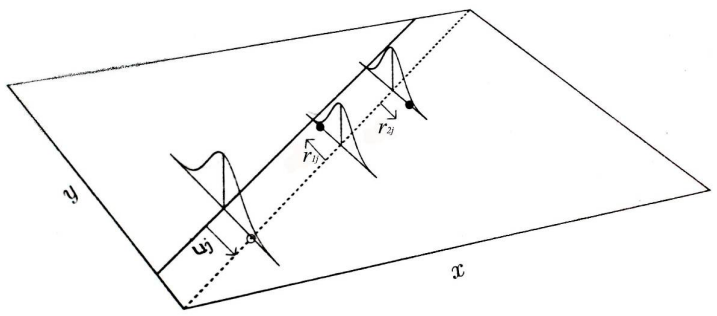
Poziom szkoły (2): $$\beta_{0j} = \gamma_{00} + u_{0j}$$

$$\beta_{1j} = \gamma_{00}$$

Jak widać, w tym modelu współczynnik nachylenia β_{1j} uznajemy za stałą dla wszystkich grup, tak jak w klasycznej regresji, i zapisujemy jako: $\beta_{1j} = \gamma_{10}$, 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: $\beta_{1j} = \gamma_{00} + u_{oj}$. Tak jak w modelu 2.2. Cały model można przedstawić w jednym równaniu liniowym w następujący sposób:

$$y_{ij} = \gamma_{00} + \gamma_{10} x_{ij} + u_{oj} + r_{ij} \quad (3.2)$$

Ilustracją dla tego modelu jest poniższy rysunek. 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. Mówiąc obrazowo, model ten pozwala na to, by wyraz wolny zmieniał się w zależności od danej grupy, przesuwając w ten sposób krzywą regresji o u_{oj}. Jak widać na rysunku, krzywa regresji dla grupy *j* (zaznaczona linią przerywaną) jest przesunięta w dół względem krzywej dla całej populacji (linia ciągła).



Rysunek 2.6. Ilustracja efektu losowego dla stałej regresji (za Rabe-Hesketh, Sophia i Anders Skrondal: s. 96)

Załóżmy, że zmienną wyjaśnianą jest wynik egzaminu gimnazjalnego, a zmienną wyjaśniającą sprawdzian po szkole podstawowej. W przypadku modelu wielopoziomowego możemy zapisać to następująco, lekko reorganizując równanie 3.2:

$$y_{ij} = (\gamma_{00} + u_{0j}) + \gamma_{10} x_{ij} + r_{ij} \quad (3.3)$$

Zastanówmy się teraz, co tutaj oznacza $u_{ij.}$ Jest ot efekt szkoły szacowany dla indywidualnego wyniku y_{ij}, a precyzyjniej relatywny efekt szkoły w stosunku do wartości przewidywanej dla całej populacji szkół γ_{00}. Ale to nie wszystko. Jest to efekt warunkowy, czyli przy założeniu, że wszystkie inne czynniki pozostają bez zmian, pozostaną zawieszone. W naszym wypadku tym innym czynnikiem wyrażonym explicite jest wartość wyniku sprawdzianu po szkole podstawowej. Wartość u_j jest zatem tym, co szkoła „dodała" (lub dodałaby, gdyby chodzili do niej inni uczniowie) bez względu na to, jakie wyniki otrzymali ze sprawdzianu po szkole podstawowej. To w edukacji nazywamy wartością dodaną. Dzięki modelom wielopoziomowym możliwe jest zatem estymowanie wartości dodanej dla szkół. Oczywiście modele szacowania EWD trzyletniego są bardziej skomplikowane, lecz skomplikowanie to wyraża się przede wszystkim w użyciu większej liczby zmiennych niezależnych zwiększających dopasowanie modelu do danych.[9] W tabeli 2.1 można znaleźć listę zmiennych niezależnych wykorzystanych modelowaniu trzyletniego wskaźnika EWD.

Tabela 2.1. Lista zmiennych niezależnych zawarta w wektorze *X*

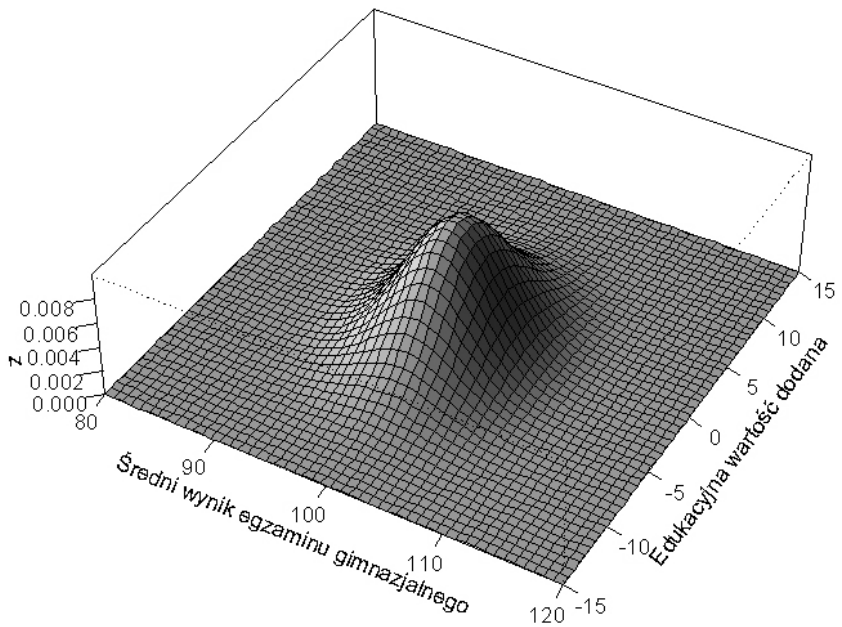
Oznaczenie	Opis	Uwagi
X_1	wynik sprawdzianu po szkole podstawowej	zmienna ciągła
X_2	płeć	zmienna 0-1
X_3	dysleksja na sprawdzianie po szkole podstawowej	zmienna 0-1
X_4	dysleksja na egzaminie gimnazjalnym	zmienna 0-1
X_3*X_4	dysleksja_spr*dysleksja_gim	interakcja (zmienna 0-1)
X_5	czy rok 2008?	zmienna 0-1
X_6	czy rok 2009?	zmienna 0-1
X_5*X_1	rok 2008*spr	interakcja
X_6*X_1	rok 2009*spr	interakcja
X_5*X_1^2	rok 2008*spr^2	interakcja
X_6*X_1^2	rok 2009*spr^2	interakcja
X_5*X_1^3	rok 2008*spr^3	interakcja
X_6*X_1^3	rok 2009*spr^3	interakcja

[9] Więcej informacji o zmiennych niezależnych w modelach edukacyjnej wartości dodanej można znaleźć w: R. Dolata (red.), *Edukacyjna wartość dodana jako metoda oceny efektywności nauczania na podstawie wyników egzaminów zewnętrznych,* CKE, Warszawa 2007.

Metoda obliczania edukacyjnej wartości dodanej na podstawie modeli wielopoziomowych ma kilka zasadniczych zalet. Niestety omówienie ich wykracza poza ramy tej publikacji. W skrócie można powiedzieć, iż podobnie jak w przypadku Bayesowskich predyktorów wyników końcowych, otrzymujemy optymalne oszacowania wskaźnika efektywności pracy szkół oraz optymalne oszacowania błędów standardowych. Warto podkreślić, iż wszystkie wzory i zależności pokazane w punkcie 2. stosują się również do wskaźnika edukacyjnej wartości dodanej. Nie będziemy zatem powtarzać tych przykładów.

4.

Ostatnim elementem niezbędnym do prezentacji wskaźnika trzyletniego jest połączenie informacji o wynikach końcowych i wskaźniku wartości dodanej. Jako że efektem zestawienia tych dwóch wskaźników ma być forma graficzna przedstawiona w dwuwymiarowym układzie współrzędnych, najwłaściwszym rozwiązaniem jest tutaj posłużenie się funkcją wiarygodności, a precyzyjniej złożeniem dwóch funkcji wiarygodności: wyniku egzaminu końcowego i wskaźnika edukacyjnej wartości dodanej dla danej szkoły. W obu wypadkach zakładamy, że funkcja wiarygodności w obydwu sytuacjach ma rozkład normalny i odchylenie standardowe proporcjonalne do błędu standardowego szacowanego wskaźnika; natomiast maksimum funkcji wiarygodności odpowiada punktowemu oszacowaniu wskaźnika. Poprzez złożenie takich dwóch funkcji otrzymujemy dwuwymiarową funkcję wiarygodności, tak jak zostało to przedstawione na rysunku 4.1.



Rysunek 4.1. Dwuwymiarowa funkcja wiarygodności

Kształt tej dwuwymiarowej funkcji gęstości wiarygodności da się opisać za pomocą macierzy kowariancji *K* tak, że:

$$K=\begin{bmatrix} \upsilon_{00} & \upsilon_{01} \\ \upsilon_{10} & \upsilon_{11} \end{bmatrix} \quad (4.1)$$

gdzie:

υ_{01} jest wariancją pierwszej funkcji wiarygodności proporcjonalną do błędu standardowego wyznaczanego wskaźnika wyników końcowych

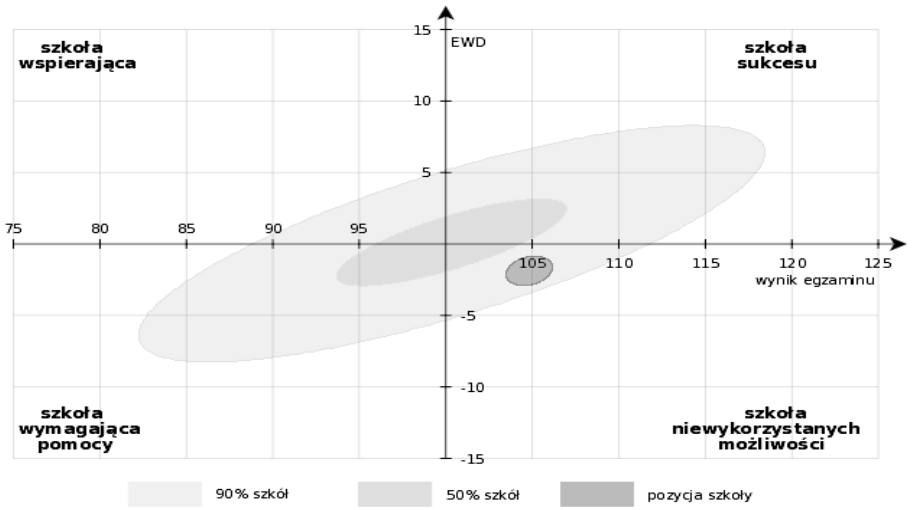
υ_{10} jest wariancją drugiej funkcji wiarygodności proporcjonalną do błędu standardowego wyznaczanego wskaźnika edukacyjnej wartości dodanej

$\upsilon_{00} = \upsilon_{11}$ to kowariancja między funkcjami wiarygodności, proporcjonalna do korelacji wskaźnika końcowego i indywidualnych reszt regresji wielopoziomowej.

Wierzchołek tej dwuwymiarowej funkcji wiarygodności wyznacza nam najbardziej wiarygodny łączny wynik końcowy i wynik edukacyjnej wartości dodanej dla dwuwymiarowej przestrzeni. Zaś kształt funkcji wiarygodności opisany macierzą kowariancji wykorzystany zostaje do określenia 95-procentowego przedziału ufności. Mówiąc obrazowo, dwuwymiarowa funkcja wiarygodności zostaje „odcięta" w taki sposób, by pozostało tylko 5% gęstości tej funkcji. Można sobie wyobrazić, że to, co zostaje po „odcięciu" owych 95%, rozkładu gęstości to przekrój „pienia" w kształcie elipsy o wymiarach modelowanych przez macierzy kowariancji *K*. Eliptyczny kształt przekroju „pienia" swym polem wyznacza 95% pole ufności, w którym znajduje się wynik prawdziwy, w centrum tego pola zaś wynik najbardziej wiarygodny.

Podsumowanie

Efektem końcowym procedur statystycznych przedstawianych w powyższych punktach jest graficzne przedstawienie wskaźnika trzyletniego, tak jak zostało przedstawione to na rysunku 5.1. Wszystkie wyniki podawane są w punktach znormalizowanych za pomocą procedury przedstawionej w punkcie 1. Oś pozioma reprezentuje wyniki końcowe szacowane tak, jak opisano to w punkcie 2. Oś pionowa reprezentuje wskaźniki edukacyjnej wartości dodanej szacowane tak, jak zostało to przedstawione w punkcie 3. Pozycje szkoły reprezentuje eliptyczne pole ufności, opisane w punkcie 4. Szare pola informują o zawierającym się w nich procencie szkół. Pola te zostały wyznaczone za pomocą odległości Mahalanobisa i można je traktować analogiczne do powierzchni ufności w punkcie 4. z tym, że powierzchni ufności odnoszącej się do populacji szkół, a nie uczniów jednej szkoły.



Rysunek 5.1. Graficzne przedstawienie trzyletniego wskaźnika egzaminacyjnego: przykład

Bibliografia:

1. R. Cole (red.), *Relative difficulty of examinations in different subjects,* CEM Centre, Durham University 2008.
2. R. Dolata (red.), *Edukacyjna wartość dodana jako metoda oceny efektywności nauczania na podstawie wyników egzaminów zewnętrznych,* CKE, Warszawa 2007.
3. M. J. Kolen i R. L. Brennan, *Test Equating, Scaling and Linking Methods and Practices Second Edition,* Springer, New York 2004.
4. S. W. Raudenbush i A.S. Bryk, *Hierarchical Linear Models,* Sage, Thousand Oaks-London-New Delhi 2002.
5. A. Skrondal i S. Rabe-Hesketh, *Generalized Latent Variable Modeling*, A CRC Press Company, Boca Raton-London-New York-Washington, D. C. 2004.
6. A. Skrondal i S. Rabe-Hesketh, *Multilevel and Longitudinal Modeling Using Stata*. College Station, Texas: Stata Press Publication – StataCorp LP. 2008.
7. Snijders, Tom A. B. i Roel J. Bosker, *Multilevel Analysis*. Thousand Oaks – London – New Delhi: Sage 2004.