



Projekt współfinansowany ze środków Unii Europejskiej w ramach Europejskiego Funduszu Społecznego.

TRZYLETNI WSKAŹNIK EGZAMINACYJNY

DOKUMENTACJA TECHNICZNA

Wersja 1.0

Styczeń, 2010



Artur Pokropek

Konsultacja merytoryczna:
Ewa Stożek, Aleksandra Jasińska

Korekta:

Magdalena Pokropek



Projekt współfinansowany ze środków Unii Europejskiej w ramach Europejskiego Funduszu Społecznego.

SPIS TREŚCI

Wstęp	3
Normalizacja wyników egzaminacyjnych za pomocą metody ekwicyntylowej	4
Szacowanie wyników końcowych za pomocą modeli wielopoziomowych	8
Szacowanie wskaźnika edukacyjnej wartości dodanej za pomocą modeli wielopoziomowych.	15
Estymacja parametrów w modelach wielopoziomowych.....	17
Przedstawianie wyników w dwuwymiarowej przestrzeni przy uwzględnieniu niepewności szacowania.	24
Jednowymiarowy rozkład normalny	25
Wielowymiarowy rozkład normalny	26
Powierzchnia ufności	31
Podsumowanie.....	34
Literatura.....	35

WSTĘP

Trzyletni wskaźnik egzaminacyjny publikowany przez Centralną Komisję Egzaminacyjną został opracowany przez grupę odpowiedzialną za metodologię szacowania wskaźnika Edukacyjnej Wartości Dodanej.

Wskaźnik ten został przychylnie odebrany przez osoby zainteresowane poszerzaniem swojej wiedzy o pracy konkretnych szkół. Forma graficzna, w jakiej przedstawiany jest wskaźnik, powoduje, iż nawet 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 precyzyjny. W dokumentacji technicznej przybliżone zostaną kulisy szacowania omawianego wskaźnika na tyle dokładnie, aby pozwolić czytelnikowi odtworzyć i zrozumieć kolejne kroki analiz statystycznych używanych podczas szacowania wskaźnika trzyletniego.

Warto podkreślić, iż publikacja wskaźnika jest elementem pilotażu i nie jest to wskaźnik ostateczny, to jest metodologia, jak i wartości publikowanych wyników dla szkół, mogą w przyszłości być dopracowywane po to, by uzyskać jeszcze większą precyzję oszacowania. W stosunku do poszczególnych szkół nie będą to zmiany rewolucyjne, wskaźnik już w tej chwili zapewnia zadowalającą dokładność.

W tej chwili szacowanie wskaźnika egzaminacyjnego można podzielić na cztery etapy:

1. Normalizacja wyników egzaminacyjnych za pomocą metody ekwicyntylowej.
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 kolejnych kroków. Przy czym rozszerzony zostanie opis logiki sposobów estymacji wykorzystywanych w szacowaniu wskaźników trzyletnich. Zaczniemy od zrównywania opartego na metodzie ekwicyntylowej, 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 istotę szacowanego wskaźnika. Skupimy się tu na modelowaniu, za pomocą którego uzyskujemy wyniki końcowe oraz wyniki wskaźnika edukacyjnej wartości dodanej. Jako odwołanie do punktu drugiego i trzeciego, poruszona zostanie również kwestia estymacji Bayesowskich, wykorzystywanych w szacowaniu trzyletnich wskaźników egzaminacyjnych. W punkcie czwartym omówiona zostanie koncepcja graficznego prezentowania

Projekt współfinansowany ze środków Unii Europejskiej w ramach Europejskiego Funduszu Społecznego.

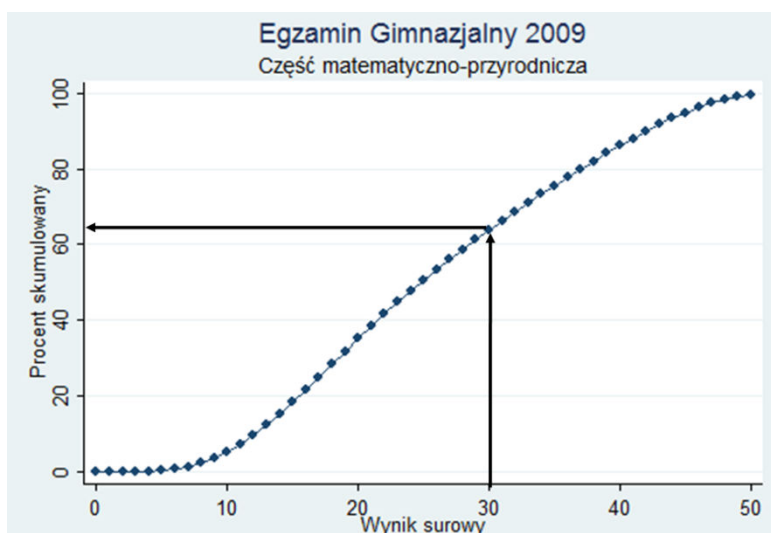
wskaźników ze szczególnym uwzględnieniem przesłanek statystycznych wiążących się z takim sposobem prezentacji wyników.

NORMALIZACJA WYNIKÓW EGZAMINACYJNYCH ZA POMOCĄ METODY EKWICENTYLOWEJ

Procedurę normalizacji wyników zastosowaną dla wskaźników trzyletnich można opisać w 3 krokach:

1. Rangowanie wyników
2. Ustalenie położenia rang w rozkładzie normalnym
3. Standaryzacja wyników do skali: średnia 100, odchylenie standardowe 15.

W pierwszym kroku rangowane są wyniki egzaminacyjne. Czyli każdemu wynikowi przypisana zostaje relatywna pozycja w rozkładzie wyników. W taki sposób, iż najwyższy wynik odpowiada 100% skumulowanych wyników, a poszczególne rangi określają relatywną pozycję danego wyniku w stosunku do wszystkich uzyskanych wyników. Pozycję tę można wyrazić za pomocą skumulowanego procentu wyników niższych i równych danemu wynikowi na skali punktowej. Graficzna reprezentacja tego przekształcenia przedstawiona została na rysunku 1. Na krzywej wyrażającej kształt funkcji rangowania kwadratami zaznaczono 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.

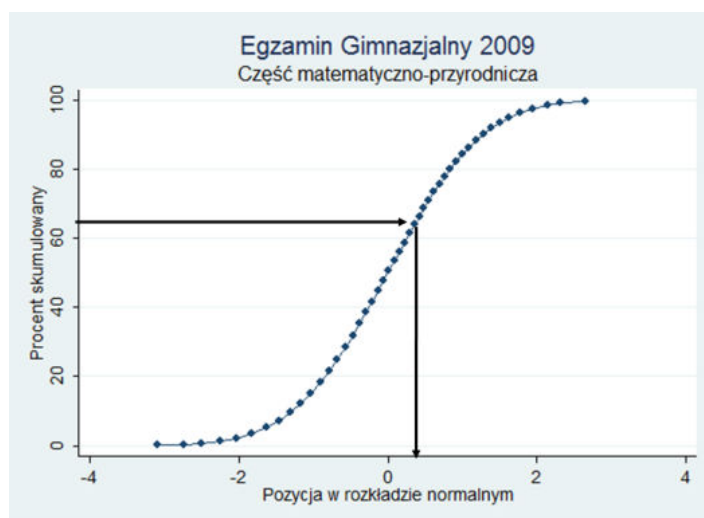


Rysunek 1. Funkcja wyznaczająca rangę pozycyjną dla egzaminu gimnazjalnego 2009, część matematyczno-przyrodnicza.

Projekt współfinansowany ze środków Unii Europejskiej w ramach Europejskiego Funduszu Społecznego.

Warto zauważyć, iż jeżeli podzielimy rozkład gęstości na 100 równych części (percentyle), wynik ten umieścimy na randze 64. Jeżeli jednak rozkład gęstości podzielimy na 1000 równych części, ranga będzie wynosiła 638. Wybór dokładności, z jaką przenosimy wyniki surowe na rangi, nie jest bez znaczenia. Im mniejsza ilość 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, 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.

W drugim kroku uzyskane rangi transponujemy na pozycje w rozkładzie normalnym. Kolejnym rangom przypisujemy takie pozycje, które globalnie gwarantowały będą wynikom rozkład najbliższy rozkładowi normalnemu o średniej 0 i odchyleniu standardowym 1. W ten sposób normalizujemy takie parametry rozkładu jak skośność i kurtoza. Graficzna reprezentacja takiego przekształcenia znajduje się na rysunku 2.



Rysunek 2. Funkcja normalizująca rangi pozycyjne dla egzaminu gimnazjalnego 2009, część matematyczno-przyrodnicza.

W trzecim kroku uzyskaną wcześniej pozycję w rozkładzie normalnym przedstawiamy na skali o średniej 100 i odchyleniu standardowym 15. Aby przekształcić uzyskane wyniki na taką skalę, musimy odwołać się do liniowej funkcji przekształcającej uzyskany wynik na skalę o zadanych

Projekt współfinansowany ze środków Unii Europejskiej w ramach Europejskiego Funduszu Społecznego.

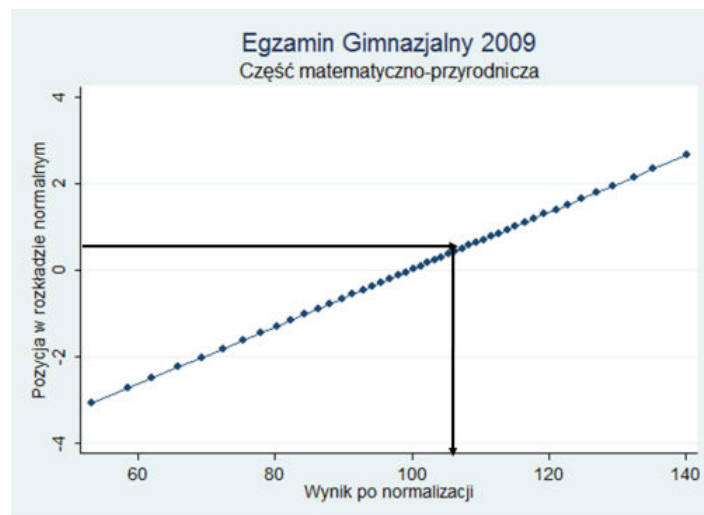
parametrach (czyli średniej 100 i odchyleniu standardowym 15. Możemy to zrobić, odwołując się do prostego przekształcenia liniowego, które możemy zapisać jako funkcję $L(p)$:

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

gdzie: σ_p to odchylenie standardowe pozycji w rozkładzie normalnym

μ_p to średnia wyniku wejściowego

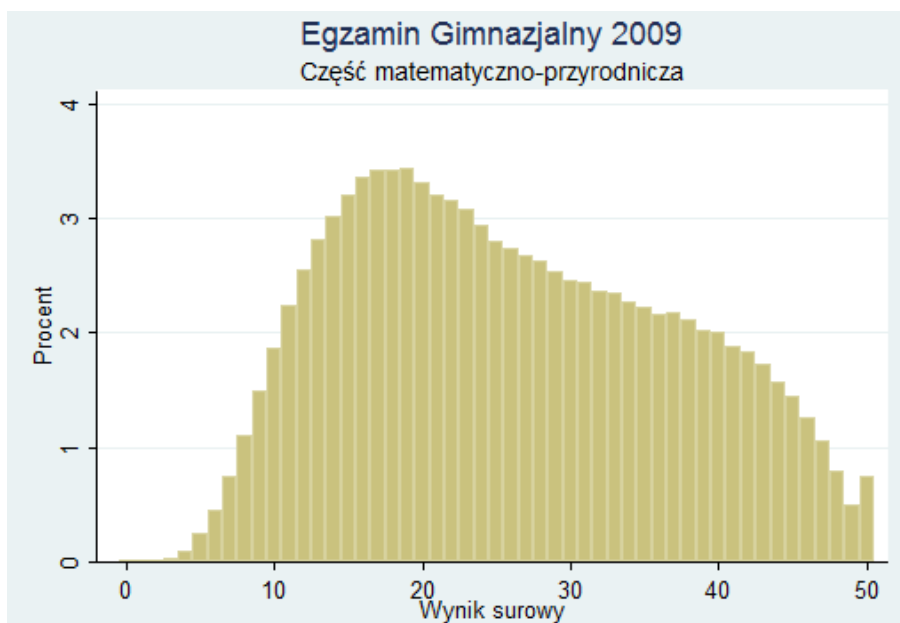
Graficzną reprezentację funkcji $L(p)$ przedstawia rysunek 3. Na tym rysunku można zaobserwować przekształcenie wcześniej omawianej rangi 638 po przypisaniu pozycji w rozkładzie normalnym na skalę o średniej 100 i odchyleniu standardowym 15, w której ranga ta ma wartość 105,32.



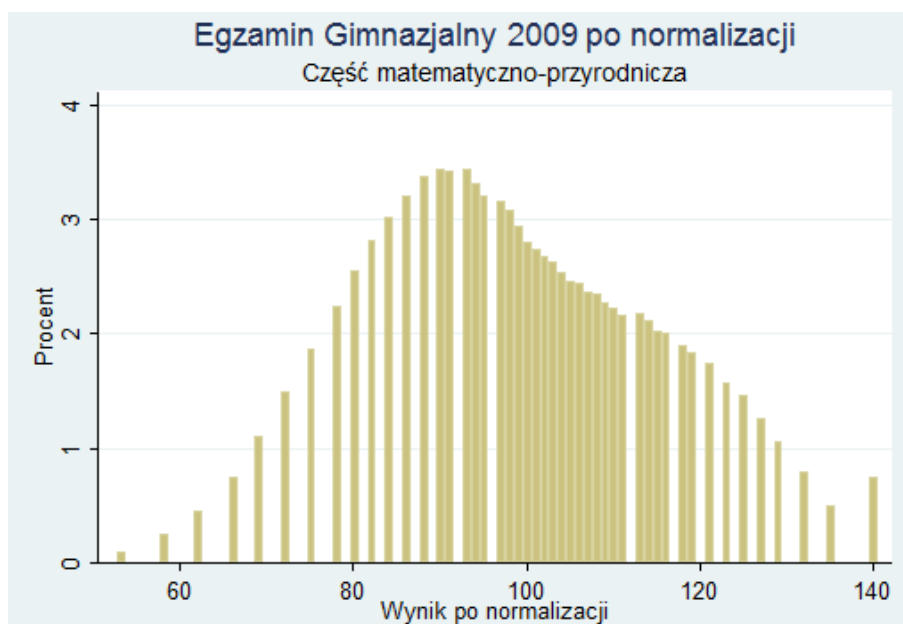
Rysunek 3. Funkcja przekształcenia liniowego wyznaczająca wynik po normalizacji

Poniżej zaprezentowane zostały wyniki normalizacji dla całego egzaminu. Rozkład wyników przed normalizacją (rysunek 4.), rozkład wyników po normalizacji (rysunek 5.) oraz tabela wyników surowych i po normalizacji (tabela 1.).

Projekt współfinansowany ze środków Unii Europejskiej w ramach Europejskiego Funduszu Społecznego.



Rysunek 4. Rozkład egzaminu gimnazjalnego 2009, część matematyczno-przyrodnicza przed normalizacją



Rysunek 5. Rozkład egzaminu gimnazjalnego 2009, część matematyczno-przyrodnicza po normalizacji

Projekt współfinansowany ze środków Unii Europejskiej w ramach Europejskiego Funduszu Społecznego.

Tabela 1. Wyniki surowe i po normalizacji egzaminu gimnazjalnego 2009, część 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								

SZACOWANIE WYNIKÓW KOŃCOWYCH ZA POMOCĄ MODELI WIELOPOZIOMOWYCH

Do szacowania wyników końcowych, 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 wykorzystać informacje o pogrupowaniu uczniów do klasycznego równania opisującego średni wynik uczniów w populacji z uwzględnieniem błędu pomiaru:

$$y_i = \beta_0 + e_i$$

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

wprowadzamy indeks j – oznaczający szkołę j .

Projekt współfinansowany ze środków Unii Europejskiej w ramach Europejskiego Funduszu Społecznego.

Wtedy równanie wygląda następująco:

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

gdzie:

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

gdzie: r_{ij} to wyraz błędu pierwszego poziomu,

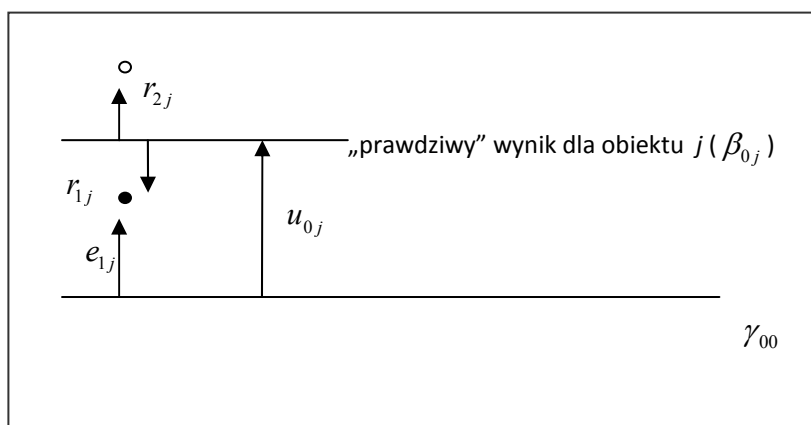
u_{0j} to wyraz błędu drugiego poziomu (utożsamiany też z efektem losowym)

γ_{00} to efekt stały równania wielopoziomowego.

Wprowadzenie informacji o pogrupowaniu powoduje, że analizy uwzględniają dwa poziomy: poziom jednostek (uczniów) i poziom grup, do których owe jednostki należą (poziom szkół). W równaniu, które uwzględnia oba poziomy, można wyodrębnić dwie części: tak zwaną część stałą i część losową (tutaj ujętą w nawias kwadratowy):

$$y_{ij} = \gamma_{00} + [u_{0j} + r_{ij}]$$

Model przedstawiony powyżej to prosty, dwupoziomowy model, który jest wykorzystywany do obliczania wskaźnika poziomu umiejętności 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}



Rysunek 6. Relacje między błędami, efektem stałym w populacji i wynikiem prawdziwym dla obiektu j

Czasami, szczególnie w teoriach testów, ten efekt stały 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

Projekt współfinansowany ze środków Unii Europejskiej w ramach Europejskiego Funduszu Społecznego.

przewidywania e_{ij} znany z wcześniejszego równania można teraz przedstawić jako sumę błędu na poziomie jednostkowym r_{ij} oraz grupowym u_{0j} tak, że:

$$e_{ij} = r_{ij} + u_{0j}$$

(Patrz rysunek 6.)

Zakładamy tutaj, że r_{ij} ma rozkład normalny w populacji pomiarów (uczniów w danej szkole). Ponadto u_{0j} , 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} .

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:

$$E(u_j r_{ij}) = 0$$

$$E(u_i u_j) = 0 \quad (i \neq j)$$

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

dla każdego i, j, t, s .

Pomiary y_{1j}, y_{2j} są niezależne od siebie na i -tym poziomie. Zależność ich widziana jest dopiero na wyższym poziomie, gdzie patrzymy na wszystkie wyniki zgromadzone w grupach, do których przynależą jednostki obserwacji, ich zależność odzwierciedlana jest przez błąd u_{0j} . Dlatego wariancję wartości y_{ij} można zapisać następująco (co również obrazowo przedstawia rysunek 6):

$$Var(y_{ij}) = Var(\gamma_{00} + u_{0j} + r_{ij}) = Var(u_{0j} + r_{ij}) = \tau_{00} + \sigma^2$$

Po co wprowadzać takie komplikacje, zamiast liczyć według średniej, w klasyczny sposób dla poszczególnych szkół? Odpowiedź jest prosta: chcemy minimalizować błędy oszacowań. Klasyczne szacowanie średniej i jej błędów standardowych nie jest jednak 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 predykcje¹ 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 (dowód można znaleźć w: Sorensen 2002). Mają też optymalne, czyli najniższe poprawne, błędy standardowe, dzięki którym wyznaczane są przedziały ufności (Rabe-Hesketh, Skrondal, 2008:82). 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 (β_{0j}).

¹ Konwencjonalnie o estymatorach Bayesowskich mówi się jako o predyktorach.

Projekt współfinansowany ze środków Unii Europejskiej w ramach Europejskiego Funduszu Społecznego.

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

Klasyczny estymator wartości oczekiwanej dla grupy j , a w języku modelowania wielopoziomowego – stała regresji wielopoziomowej ($\beta_j = \beta_{0j} + u_{0j}$), to oczywiście klasycznie liczona średnia²:

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

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

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

Gdzie n_j oznacza liczebność j -tej grupy, N to liczba badanych grup; M to całkowita liczebność populacji. Estymator Bayesowski jest średnią ważoną dwóch estymatorów: grupowego i populacyjnego tak, że (Snijders, Bosker, 2004: 56):

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

gdzie:

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

Jak widać zatem, na różnicę między estymatorem klasycznym i predyktorem Bayesowskim dla j -tej grupy wpływają 4 czynniki:

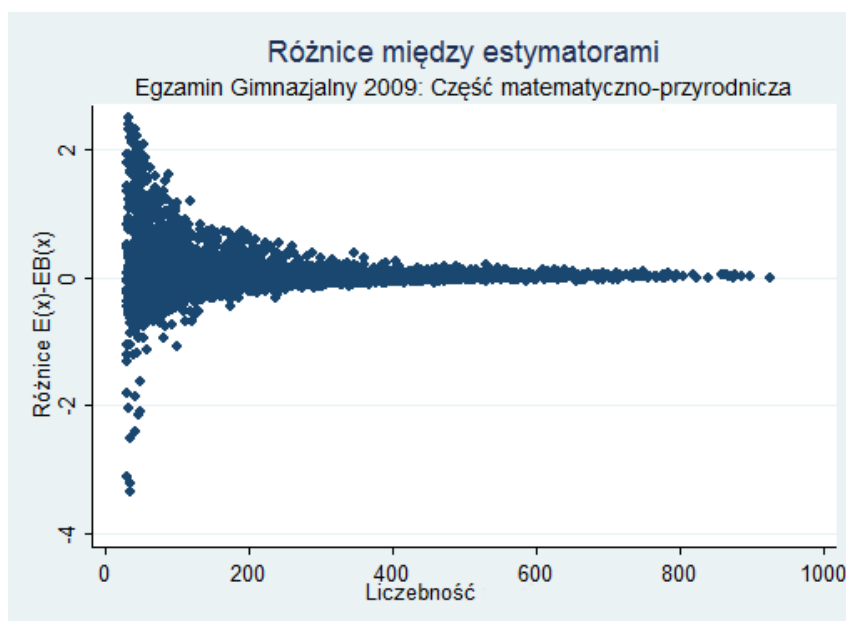
1. zróżnicowanie predyktorów grupowych w całej populacji charakteryzowane przez τ_{00}^2 ,
2. zróżnicowanie wewnątrzgrupowe: σ^2 ,
3. liczebność grupy: n_j ,
4. różnica między średnią w populacji a klasycznym estymatorem wyników końcowych w grupie.

W praktyce dwie ostatnie wielkości mają największy wpływ na różnicę między oszacowaniem klasycznym a Bayesowskim: są to liczebność grupy j (w naszym wypadku szkoły) oraz odległość średniej dla grupy j od średniej w całej populacji odległość od średniej populacyjnej. Im szkoła mniejsza, a wynik jej bardziej oddalony od średniej populacyjnej, tym korekta związana z użyciem

² W tym wypadku średnia jest też tożsama z estymatorem MLE.

Projekt współfinansowany ze środków Unii Europejskiej w ramach Europejskiego Funduszu Społecznego.

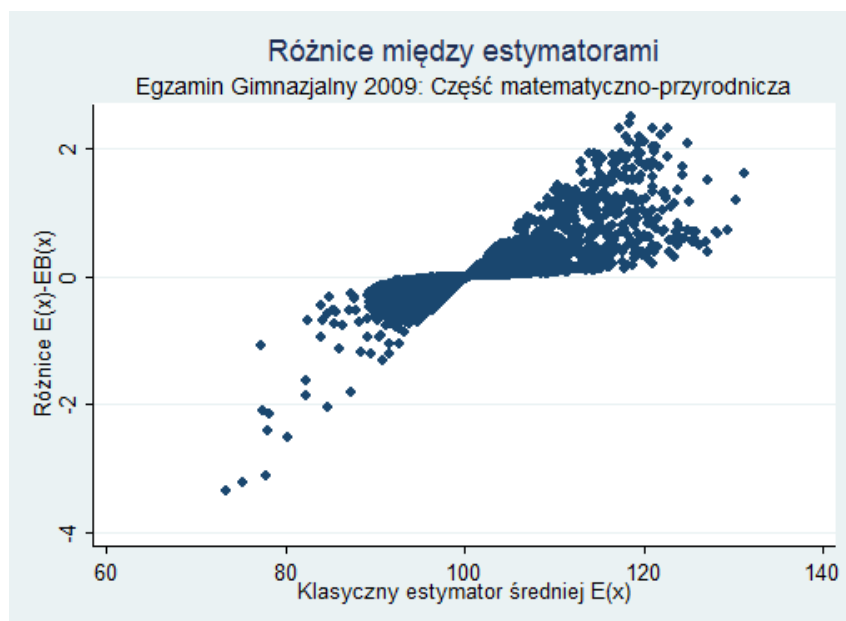
predyktora Bayesowskiego jest większa, czyli innymi słowy: w przypadku małych i oddalonych od średniej populacyjnej szkół predyktor Bayesowski będzie „ściągał” wyniki ku średniej. Przedstawione to zostało na rysunkach 7. i 8. Na pierwszym z tych rysunków widać, że największe różnice między klasycznym estymatorem a predyktorem Bayesowskim dotyczą szkół, w których w przeciągu 3 lat uczyło się mniej niż 100 uczniów. Poniżej tego progu różnice szybko maleją, a powyżej 400 jednostek obserwacji stają się marginalnymi i oscylują wokół punktu 0 (na skali średnia 100, odchylenie standardowe 15). Natomiast na rysunku 8 widać, iż różnice między klasycznym estymatorem a predyktorem Bayesowskim rosną wraz z oddalaniem się wyników szkół od średniej populacyjnej (100).



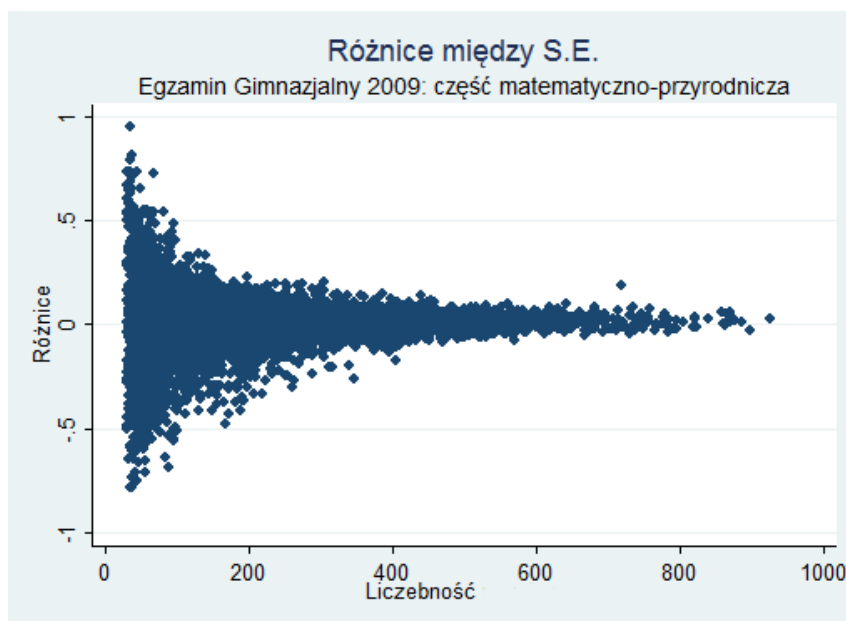
Rysunek 7. Różnice między estymatorem klasycznym i Bayesowskim w zależności od liczebności uczniów



Projekt współfinansowany ze środków Unii Europejskiej w ramach Europejskiego Funduszu Społecznego.

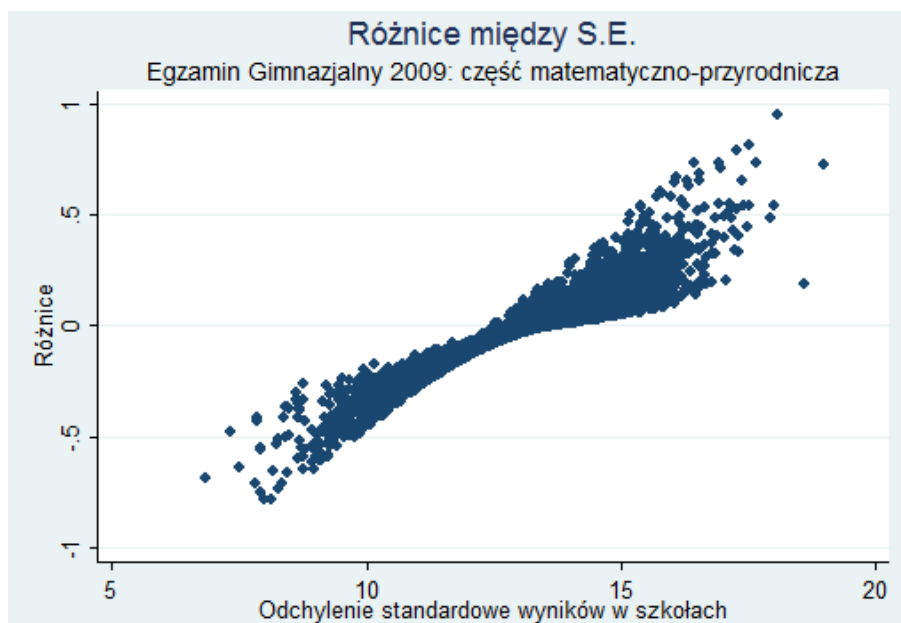


Rysunek 8. Różnice między estymatorem klasycznym i Bayesowskim dla różnych wyników estymatora klasycznego



Rysunek 9. Różnice między klasycznym błędem standardowym i błędem standardowym dla predyktora Bayesowskiego w zależności od liczebności uczniów

Projekt współfinansowany ze środków Unii Europejskiej w ramach Europejskiego Funduszu Społecznego.



Rysunek 10. Różnice między klasycznym błędem standardowym i błędem standardowym dla predyktora Bayesowskiego a zróżnicowanie wyników

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

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

Korekta błędów standardowych zależy przede wszystkim od dwóch czynników: wielkości szkoły i zróżnicowania wyników uzyskiwanych przez uczniów w szkołach. Pokazują to rysunki 9. i 10. Na rysunku 9. 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 10. pokazuje, jak korekty Bayesowskie zależą od zróżnicowania wyników wewnątrz szkoły. Gdy zróżnicowanie mierzone odchyleniem standardowym zbliża się do średniego zróżnicowania 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 procedura szacowania predyktorów Bayesowskich, szacowanie błędów oparte na tej metodzie jest optymalne, czyli w sposób optymalny minimalizuje popełnienie błędu podczas wnioskowania statystycznego.

SZACOWANIE WSKAŹNIKA EDUKACYJNEJ WARTOŚCI DODANEJ ZA POMOCĄ MODELI WIELOPOZIOMOWYCH

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):
$$y_{ij} = \beta_{0j} + \beta_{1j}x_{ij} + r_{ij}$$

gdzie:

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

$$\beta_{1j} = \gamma_{10}$$

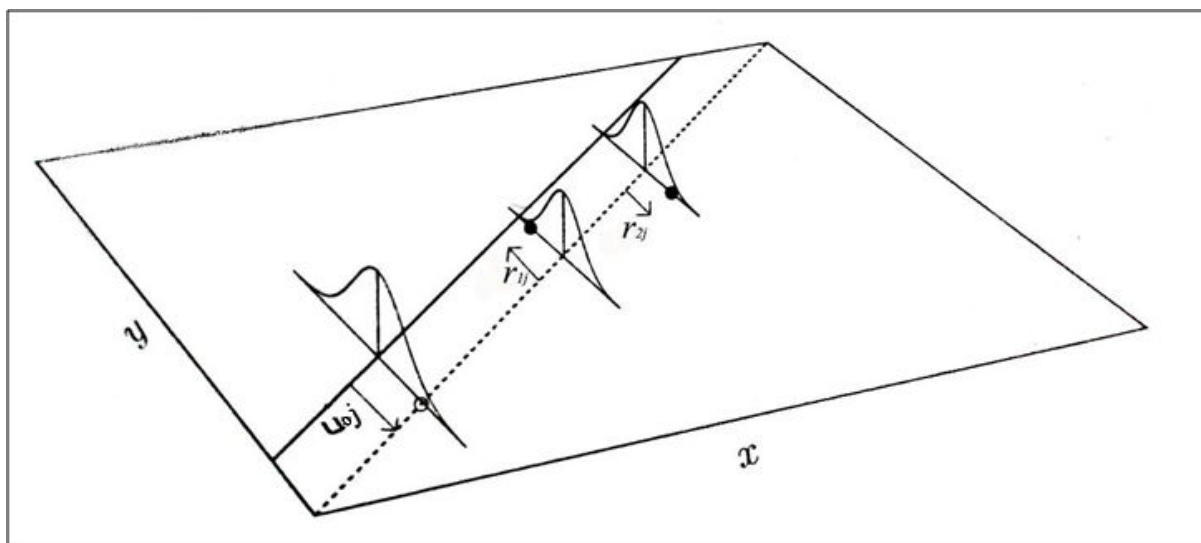
gdzie: γ_{10} to efekt stały - współczynnik nachylenia krzywej regresji.

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: $\beta_{0j} = \gamma_{00} + u_{0j}$. Cały model można przedstawić w jednym równaniu liniowym w następujący sposób:

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

Ilustracją dla tego modelu jest rysunek 11. 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_{0j} . Jak widać na rysunku, krzywa regresji dla grupy j (zaznaczona linią przerywaną) jest przesunięta w dół, względem krzywej dla całej populacji (linia ciągła).

Projekt współfinansowany ze środków Unii Europejskiej w ramach Europejskiego Funduszu Społecznego.



Rysunek 11. Ilustracja efektu losowego dla stałej regresji (za Rabe-Hesketh, Sophia i Anders Skrondal: s. 96)

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

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

Zastanówmy się teraz, co tutaj oznacza u_{0j} . Jest to 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. W naszym wypadku tym innym czynnikiem wyrażonym explicite jest wartość wyniku sprawdzianu po szkole podstawowej. Wartość u_{0j} jest zatem tym, co szkoła „dodała” (lub dodałaby, gdyby chodziło o 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.³ W tabeli 2 można znaleźć listę zmiennych niezależnych wykorzystanych w modelowaniu trzyletniego wskaźnika EWD.

³ 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.

Projekt współfinansowany ze środków Unii Europejskiej w ramach Europejskiego Funduszu Społecznego.

Tabela 2. Lista zmiennych niezależnych zawarta w wektorze X (przykład dla modelu 2009)

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

Metoda obliczania edukacyjnej wartości dodanej na podstawie modeli wielopoziomowych ma kilka zasadniczych zalet, przede wszystkim zalet statystycznych. 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 dla modelu pustego stosują się również do wskaźnika edukacyjnej wartości dodanej. Nie będziemy zatem powtarzać tych przykładów.

ESTYMACJA PARAMETRÓW W MODELACH WIELOPOZIOMOWYCH

Podczas estymacji parametrów i predyktorów Bayesowskich dla modeli wielopoziomowych napotykamy zasadniczą trudność. Dla modelu, za pomocą, którego szacujemy wyniki końcowe oraz EWD musimy znać τ_{00} , zdającego sprawę ze zróżnicowania populacyjnego (patrz rozdział: „Szacowanie wyników końcowych za pomocą modeli wielopoziomowych”). Aby poprawnie estymować τ_{00} należy odwołać się tutaj do metody największej wiarygodności. Estymacja efektów losowych za pomocą metody największej wiarygodności jest skomplikowanym procesem, który zostanie opisany w zarysie w dalszej części.

Szacowanie macierzy T odbywa się na podstawie metody największej wiarygodności. Aby ją pokrótce przedstawić, wprowadźmy kilka nowych oznaczeń.

Projekt współfinansowany ze środków Unii Europejskiej w ramach Europejskiego Funduszu Społecznego.

Dane są:

- $X = [X_1, \dots, X_N]^T$ – kolumnowy wektor zmiennych losowych
- $x = [x_1, \dots, x_N]^T$ – kolumnowy wektor obserwowanych wartości zmiennych losowych
- $f(x; \theta)$ – funkcje prawdopodobieństwa lub funkcje gęstości o nieznanym parametrze θ
- $\Theta = [\theta_1, \dots, \theta_p]^T$ – parametry wyznaczone są przez kolumnowy wektor, gdzie θ może być skalarem o $p = 1$; niech wartości parametru θ należą do przestrzeni prób Θ
- N – wielkość próby

Z definicji funkcja wiarygodności jest funkcją θ proporcjonalną do: $P(X | \theta)$, oznaczane jest jako $L(\theta; x)$ albo po prostu $L(\theta)$. Jest funkcją parametru dla określonych danych. Likelihood (funkcja wiarygodności) nie jest zatem funkcją prawdopodobieństwa i $L(\theta)$ nie może być interpretowana w kategoriach prawdopodobieństwa – co podkreślał sam Fisher (1926). Między innymi dlatego, że wartość prawdziwego parametru jest stała (*fixed*). Fisher miał zaproponować funkcję wiarygodności jako racjonalną miarę stopnia wiary (*rational measure of degree of belief*), bez właściwości probabilistycznych (Sorensen, Gianola: 121). $L(\theta)$ można interpretować jako pełny opis informacji o θ uzyskany przez aktualne dane. Z definicji zatem (Sorensen, Gianola 2002: 121):

$$L(\theta; x) = k(x)p(x; \theta) \propto p(x; \theta), \quad \theta \in \Theta$$

gdzie funkcja k nie zależy od parametru θ , lecz od danych i $k(x) > 0$.

W tym momencie warto również przypomnieć sobie podstawowe przekształcenia, które ułatwiają obliczenia. Szczególnie użyteczne okazują się logarytmiczne przekształcenia funkcji wiarygodności:

$$\log L(\theta; x) = l(\theta; x)$$

Zakładając, że elementy w X są stochastycznie niezależne, ich łączny rozkład można zapisać jako iloraz wiarygodności:

$$L(\theta) = \prod_{i=1}^n f(x_i; \theta)$$

Używając zatem logarytmu, wystrzegamy się niepotrzebnego mnożenia i zastępujemy je dodawaniem:

$$L(\theta; x) = \ln \left[\prod_{i=1}^N f(x_i; \theta) \right] = \sum_{i=1}^N \ln[f(x_i; \theta)]$$

Projekt współfinansowany ze środków Unii Europejskiej w ramach Europejskiego Funduszu Społecznego.

Co ma niebagatelne znaczenie w przypadku skomplikowanych obliczeń, gdzie wymnażając przez siebie dużo liczb, moglibyśmy napotkać problemy nawet przy użyciu nowoczesnych komputerów.

Estymacja metodą największej wiarygodności polega na tym, aby znaleźć taką wartość $\hat{\theta}$ (estymatora), która maksymalizuje wiarygodność zaobserwowanych danych. Innymi słowy, musimy znaleźć parametry rządzące rozkładem wyników doświadczenia losowego, przy których to, co się zdarzyło, będzie miało najwyższą wiarygodność.

Procedura estymacji największej wiarygodności polega na znalezieniu takiego θ^* , przy którym dla wszystkich innych θ

$$P(x | \theta^*) \geq P(x | \theta)$$

gdzie X jest wynikiem doświadczenia losowego.

Innymi słowy, z próby składającej się z obserwacji x dla losowego wektora \mathbf{X} trzeba znaleźć takie rozwiązanie θ , które maksymalizuje łączną wiarygodność wylosowania obserwowanych danych. Co oznacza maksymalizację funkcji $L(\theta)$. Formalnie możemy zapisać to w taki sposób::

$$L(\hat{\theta} | \mathbf{x}) = \max_{\theta \in \Theta} L(\theta | \mathbf{x})$$

Przywołajmy prosty przykład doświadczeń losowych, jakimi są rzuty monetą, gdzie rozkład wiarygodności można wyznaczyć na podstawie rozkładu Bernoulliego. Przywołując definicję funkcji wiarygodności, widzimy, że interesuje nas tylko ta część funkcji, która bezpośrednio odnosi się do szukanego parametru:

$$p^k (1-p)^{n-k}$$

Zakładając zatem, że $k(x) = 1$, funkcja największej wiarygodności po przekształceniu logarytmicznym wygląda następująco:

$$l(p | x) = \log[p^k (1-p)^{n-k}] = k \log(p) + (n-k) \log(1-p)$$

Znaleźć maksimum funkcji wiarygodności to tyle, co odnaleźć pierwszą pochodną tej funkcji (dla szukanego parametru) równą zero (płaska funkcja określająca maksimum):

$$\frac{\partial L(p)}{\partial p} = \frac{k}{p} - \frac{(n-k)}{(1-p)} = \frac{(k-np)}{p(1-p)}$$

$$\text{dla } \frac{\partial L(p)}{\partial p} = 0$$

$$\frac{(k-np)}{p(1-p)} = 0 \rightarrow \hat{p} = \frac{k}{n}$$

Projekt współfinansowany ze środków Unii Europejskiej w ramach Europejskiego Funduszu Społecznego.

Sprawdźmy, że wyznaczone ekstremum jest maksimum, tzn., że:

$$\frac{\partial^2 L(\hat{p})}{\partial p^2} < 0$$

$$\frac{\partial^2 L(\hat{p})}{\partial p^2} = \frac{k(2\hat{p} - 1) - n\hat{p}^2}{\hat{p}^2(\hat{p} - 1)^2} < 0$$

Co w tym wypadku zostało spełnione.

Estymacji metodą największej wiarygodności można używać, odwołując się do innych rozkładów, szukając różnych parametrów. Załóżmy dla przykładu, iż losujemy liczby z rozkładu normalnego. Załóżmy dalej, że chcemy znaleźć średnią, która maksymalizuje funkcję wiarygodności, czyli jest najbardziej wiarygodną średnią dla tej wylosowanej próbki. Dla uproszczenia złożmy też, że znamy wariancję rozkładu, z którego losowaliśmy naszą średnią. Tak, że: $X \propto N(\theta, \sigma^2)$. Wektor poszukiwanych parametrów, w tym wypadku, jest skalarą $\theta = [\mu]$, czyli $p(x; \theta) = p(x; \mu)$.

Zatem w przypadku rozkładu normalnego:

$$l(\mu) = \sum_{i=1}^N \ln \left[\frac{1}{\sqrt{2\pi\sigma^2}} \exp \left(\frac{-(x_i - \mu)^2}{\sigma^2} \right) \right] = -N \left(\ln \sqrt{2\pi\sigma^2} \right) - \frac{1}{2\sigma^2} \sum_{i=1}^N (x_i - \mu)^2$$

Z tego równania możemy opuścić stałe (znane wartości) i analizować jedynie niewielką część uzależnioną od danych i estymowanego parametru:

$$-\sum_{i=1}^N (y_i - \mu)^2$$

Posłużmy się prostym przykładem. Mamy wzrost 10 respondentów, który został przedstawiony w tabeli 3.

Projekt współfinansowany ze środków Unii Europejskiej w ramach Europejskiego Funduszu Społecznego.

Tabela 3. Wzrost 10 respondentów

Lp	Wzrost (x)
1	158
2	180
3	183
4	170
5	163
6	178
7	182
8	164
9	173
10	179

Estymacja metodą największej wiarygodności polega na poszukiwaniu najwyższej wartości funkcji wiarygodności, w tym przypadku danej jako: $-\sum_{i=1}^N (x_i - \mu)^2$.

Aby odnaleźć maksimum funkcji algebraicznie, musimy odwołać się do pierwszej pochodnej i znaleźć taki argument funkcji, dla którego pierwsza pochodna będzie równa 0, co będzie równoznaczne z odnalezieniem stycznej o nachyleniu 0.

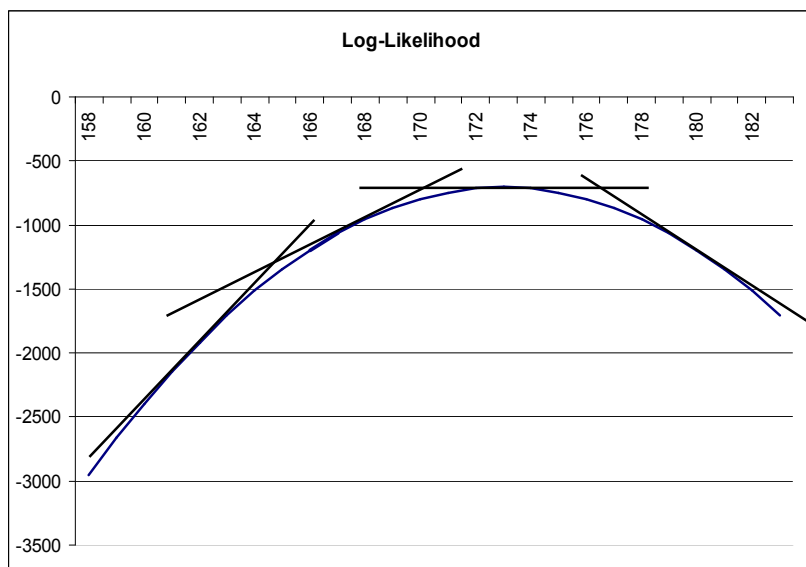
$$\frac{\partial}{\partial \mu} \left[-\sum_{i=1}^N (x_i - \mu)^2 \right] = 2 \sum_{i=1}^N x_i - 2N\mu$$

$$2 \sum_{i=1}^N x_i - 2N\mu = 0 \rightarrow \bar{\mu} = \frac{\sum_{i=1}^N x_i}{N}$$

W tym wypadku estymator największej wiarygodności, co nie może być wielkim zaskoczeniem, pokrywa się z klasycznie liczoną średnią.

Graficznie możemy sprawdzić, jaka wartość średniej byłaby najbardziej wiarygodna dla rozpatrywanych danych. Na rysunku 12. przedstawiona została funkcja wiarygodności z kilkoma przykładowymi stycznymi, w tym ze styczną reprezentującą pochodną funkcji wiarygodności o wartości 0. Jak łatwo dostrzec, wartość szacowanego parametru wynosi 173 (dla punktu, w którym styczna jest równoległa do osi X).

Projekt współfinansowany ze środków Unii Europejskiej w ramach Europejskiego Funduszu Społecznego.



Rysunek 12. Funkcja wiarygodności po przekształceniu logarytmicznym dla 10 respondentów – estymacja średniej wzrostu

Z formalnego punktu widzenia pozostało nam zweryfikowanie, czy to jest rzeczywiste maksimum. Nachylenie stycznej będzie równe 0 również dla minimum danej funkcji. Jakkolwiek w tym wypadku, jeśli posiadamy jeszcze wykres funkcji, wydaje się to niekonieczne, dla porządku przeprowadźmy taką weryfikację. Można ją przeprowadzić, odwołując się do drugiej pochodnej (pochodnej drugiego rzędu, która mówi nam o tym, czy funkcja jest wklęsła czy wypukła – nasza funkcja ma być wklęsła), jeżeli druga pochodna jest mniejsza niż 0, wtedy faktycznie naszym rozwiązaniem jest maksimum, co jest spełnione w naszym przypadku:

$$\frac{\partial^2}{\partial \mu \partial \mu} = \left[- \sum_{i=1}^N (x_i - \mu)^2 \right] = - \frac{\sigma^2}{N} < 0$$

Do wyznaczenia pozostała jeszcze jedna niezwykle ważna dla nas wartość, mianowicie wariancja estymatora. Może być ona wyznaczona przez funkcję drugiej pochodnej, a pisząc precyzyjniej (za: Eliason 1993) odwróconą ujemną wartość oczekiwaną drugiej pochodnej. W naszym przypadku:

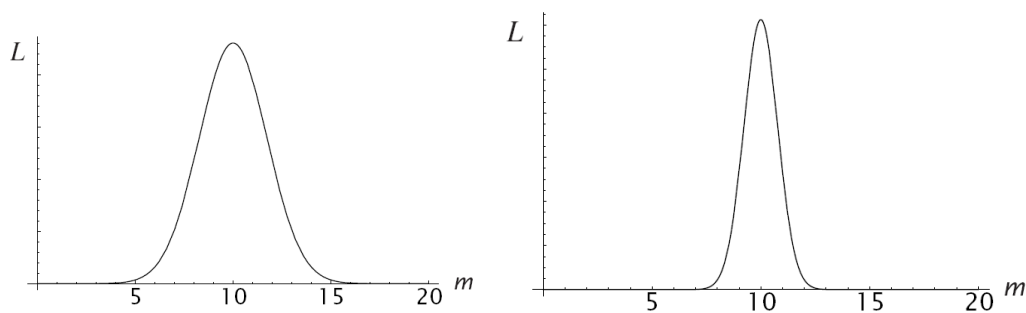
$$V(\hat{\mu}) = \left[-E \left(\frac{\partial^2 l(\mu)}{\partial \mu \partial \mu} \right) \right]^{-1} = \sigma / N$$

Takie sformułowanie wariancji estymatora ma stosunkowo łatwą interpretację geometryczną. Pierwsza pochodna reprezentuje nachylenie stycznej, czyli wskazuje na to, jak „stroma” jest rozpatrywana funkcja w danych punktach. Druga pochodna jest niczym innym jak pochodną funkcji reprezentującej pierwszą pochodną. Druga pochodna mówi nam zatem o tym, z jaką gwałtownością zmienia się „stromizna” badanej funkcji. Wariancja estymatora jest tym większa, im funkcja

Projekt współfinansowany ze środków Unii Europejskiej w ramach Europejskiego Funduszu Społecznego.

wiarygodności jest mniej „stroma”, czyli określenie jej maksimum jest „mniej pewne”. Gdy funkcja wiarygodności jest „stroma”, wtedy wariancja estymatora jest mniejsza, gdyż funkcja „szybciej” zbiega do swojego maksimum.

Na rysunku 13. przedstawiono dwie funkcje wiarygodności użyte przy szacowaniu parametru m . W przypadku pokazanym w lewej części rysunku użyto mniejszej próby niż dla przykładu z prawej strony. Funkcja wiarygodności z lewej strony jest mniej „stroma” niż funkcja wiarygodności z prawej strony. Druga pochodna funkcji wiarygodności z lewej strony rysunku będzie zatem przyjmowała niższe wartości niż druga pochodna funkcji wiarygodności z prawej strony rysunku. Jak pisaliśmy wcześniej, wartość funkcji wiarygodności dla danej wartości parametru możemy określać jako wiarygodność tego parametru względem posiadanych danych. Jak widać, zróżnicowanie wartości wiarygodności dla danych parametrów w lewej części rysunku jest większe niż w przykładzie z prawej strony rysunku, co związane jest z tym, iż szacowanie estymatora parametru m jest w tym przypadku mniej pewne, niż w przypadku funkcji przedstawionej z prawej strony rysunku. Zatem wariancja estymatora jest większa niż dla drugiej przedstawianej na rysunku 13. funkcji wiarygodności.



Rysunek 13. Dwie funkcje wiarygodności dla parametru m

Zależność między odwróconą ujemną wartością oczekiwaną drugiej pochodnej a wariancją estymatora można również udowodnić analitycznie, lecz dowód potwierdzający przywoływany związek wykracza poza ramy tej pracy (por. Eliason 1993).

Estymacja parametrów stałych i losowych w modelach wielopoziomowych jest analogiczna do prostej estymacji jednego parametru, tak jak zostało to przedstawione w poprzednim punkcie. W modelach wielopoziomowych nie estymujemy jednego parametru, lecz zadanie polega na oszacowaniu kilku parametrów, które można zapisać w formie wektora parametrów: θ . Szukamy wtedy takiego wektora parametrów, który byłby najbardziej wiarygodnym wektorem parametrów dla analizowanych danych (czyli poszukujemy najbardziej wiarygodnej konfiguracji parametrów dla danych). Zadanie obliczeniowe jest analogiczne do poprzednich przykładów, tutaj również szukamy pochodnej funkcji wiarygodności (najczęściej zlogarytmizowanej) z tym, że dla danego wektora parametrów (a nie jednego parametru):

$$\frac{\partial \log L(\theta | \mathbf{X})}{\partial \theta} = 0$$

Projekt współfinansowany ze środków Unii Europejskiej w ramach Europejskiego Funduszu Społecznego.

W przypadku funkcji wielu zmiennych do wyznaczenia ekstremum stosuje się iteracyjne metody numeryczne. To w zasadzie jedyna droga, gdy nie znamy dokładnie przebiegu badanej funkcji.

Wariancję estymatorów szacuje się natomiast na podstawie macierzy Hessego (hesjanu), czyli macierzy drugich pochodnych. Hesjan wyraża się jako (Eliason, 1993):

$$\mathbf{H}(\boldsymbol{\theta}) = \left(\frac{\partial^2 l(\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j} \right)$$

A wariancję estymatorów wyraża się jako:

$$V(\hat{\boldsymbol{\theta}}) = [-\mathbf{H}(\hat{\boldsymbol{\theta}})]^{-1}$$

Uważa się, że MNW jest przynajmniej teoretycznie najlepszym sposobem estymacji dla modeli wielopoziomowych. Estymacja MNW jest odporna (*robust*) oraz daje estymatory asymptotycznie efektywne i zgodne.

PRZEDSTAWIANIE WYNIKÓW W DWUWYMIAROWEJ PRZESTRZENI PRZY UWZGLĘDNIENIU NIEPEWNOŚCI SZACOWANIA

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, właściwszym rozwiązaniem jest tutaj posłużenie się technikami statystyki wielu zmiennych. Przypomnimy właściwości jedno- i wielowymiarowych rozkładów normalnych, a następnie pokażemy, dlaczego powierzchnie ufności, przy prezentacji wskaźnika trzyletniego, przybierają charakterystyczny kształt elips.

Projekt współfinansowany ze środków Unii Europejskiej w ramach Europejskiego Funduszu Społecznego.

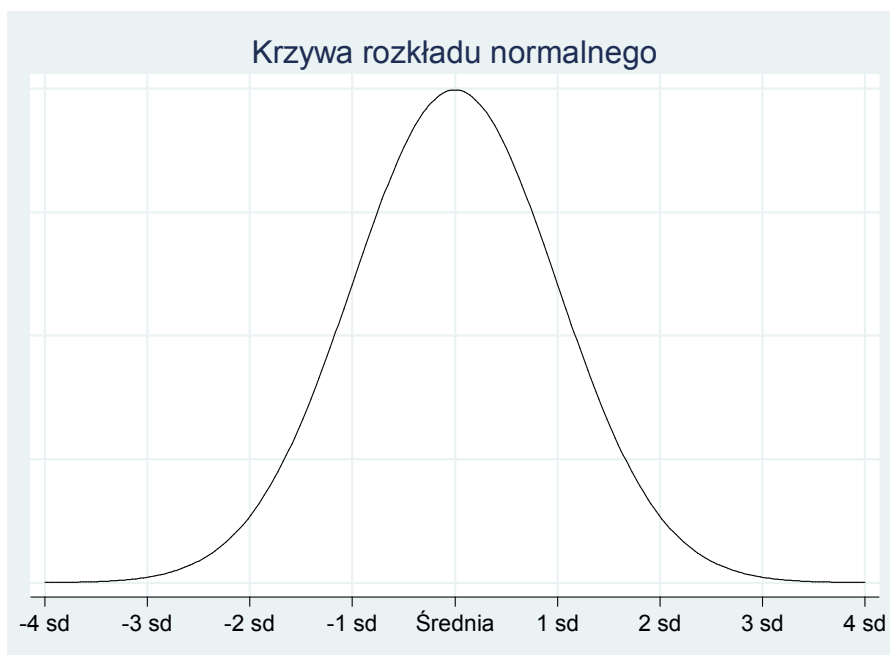
JEDNOWYMIAROWY ROZKŁAD NORMALNY

Jednowymiarowy rozkład normalny można zdefiniować, jako funkcję gęstości o wzorze:

$$f(y) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(y-\mu)^2/2\sigma^2}, \quad -\infty < y < \infty$$

gdzie y jest zmienną losową o średniej μ oraz o wariancji σ^2 , co w skrócie przyjęto zapisywać jako:

$N(\mu, \sigma^2)$. Funkcja ta graficznie da się przedstawić jako krzywa normalna, którą często nazywa się krzywą dzwonową, z uwagi na jej charakterystyczny kształt. Krzywa rozkładu normalnego została przedstawiona na rysunku 14.



Rysunek 14. Krzywa rozkładu normalnego

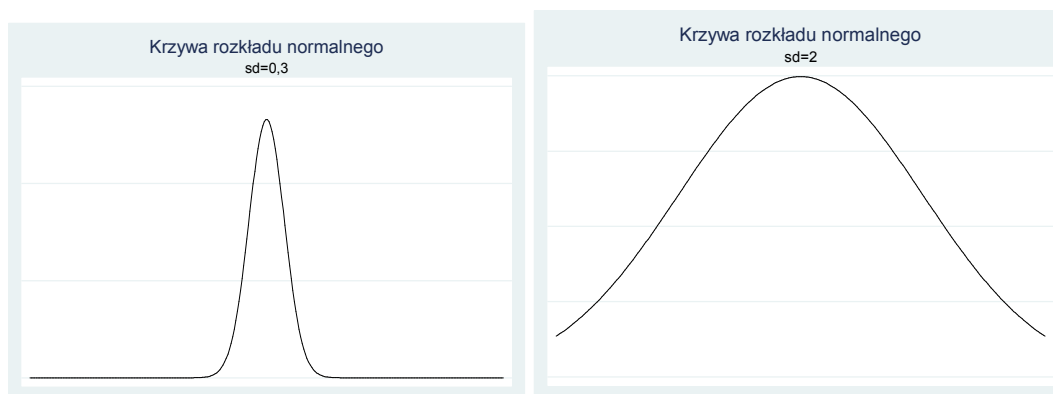
W rozkładzie normalnym najwięcej obserwacji ma wartości bliskie średniej, a liczba obserwacji maleje w miarę oddalania się od niej. Przy tym rozkład jest symetryczny, a więc jest tyle samo obserwacji powyżej, jak i poniżej średniej. Na rysunku 14. na osi poziomej zaznaczone zostały wartości odchyłeń standardowych (sd). Zakładając, że badamy zmienną o rozkładzie normalnym, w dowolnej populacji w przedziale (-1sd; 1sd) znajduje się około 68% jednostek obserwacji, około 95,5% w przedziale (-2sd; 2sd). Jeżeli chcielibyśmy dowiedzieć się, w jakim przedziale znajduje się 90% populacji, musimy przemnożyć odchylenie standardowe przez 1,64. Załóżmy, że rozkład badanej zmiennej ma średnią 100 i odchylenie standardowe 15, wtedy : $15 \times 1,64 = 24,6$. Oznacza to, że

Projekt współfinansowany ze środków Unii Europejskiej w ramach Europejskiego Funduszu Społecznego.

w przedziale $100 \pm 24,6$ będzie się zawierało 90% wyników (czyli między 75,4 a 124,6). Jeżeli chcemy uzyskać informacje o 95% wyników, odchylenie standardowe należy przemnożyć przez 1,96.

Za pomocą krzywej rozkładu normalnego możemy opisać jedną zmienną. Wskaźnik trzyletni jest natomiast złożeniem informacji o dwóch wskaźnikach (zmiennych) i aby poprawnie go opisać należy posłużyć się uogólnioną formą rozkładu normalnego, która pozwala analizować wiele zmiennych za pomocą jednej funkcji gęstości.

Warto zwrócić uwagę na to, iż funkcje gęstości mogą przybierać różne kształty w zależności od zróżnicowania w badanych populacjach. Jeżeli zróżnicowanie zmiennej w populacji jest małe (tak jak przedstawiono na lewym panelu rysunku 15.), krzywa ma kształt spiczasty. Jeżeli zróżnicowanie danej zmiennej w innej populacji jest duże (tak jak w przypadku prawej strony panelu rysunku 15.), krzywa rozkładu normalnego jest spłaszczona.



Rysunek 15. Przykłady rozkładów normalnych o różnym odchyleniu standardowym i tej samej średniej, na jednej skali wartości zmiennych

Parametrem za to odpowiadającym jest odchylenie standardowe σ (które oczywiście po podniesieniu do kwadratu interpretowane jest jako wariancja). Widać wyraźnie, iż gdy odchylenie standardowe jest tak małe, jak w lewym panelu ($sd = 0,3$), krzywa ma kształt spiczasty. Dla dużego odchylenia w prawym panelu ($sd = 2$) krzywa rozkładu normalnego jest „rozciągnięta wszerek”.

WIELOWYMIAROWY ROZKŁAD NORMALNY

Zanim przejdziemy do formalnego opisu rozkładu normalnego, musimy wprowadzić trzy pojęcia: (1) wektor zmiennych, (2) wektor średnich oraz (3) macierz kowariancji.

Wektor zmiennych

Można zdefiniować wektor zmiennych \mathbf{y} dla dowolnej liczby zmiennych p przy dowolnej liczebności próby n , składający się z wektorów obserwacji $\mathbf{y}_1 \dots \mathbf{y}_n$ tak, że:

Projekt współfinansowany ze środków Unii Europejskiej w ramach Europejskiego Funduszu Społecznego.

$$\mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$$

gdzie każdej jednostce obserwacji przypisany został wektor \mathbf{y}_i zawierający informacje o wartościach p zmiennych dla danej obserwacji tak, że:

$$\mathbf{y}_i = \begin{bmatrix} y_{i1} \\ \vdots \\ y_{ip} \end{bmatrix}$$

gdzie y_{ip} jest wartością zmiennej p dla obserwacji i .

Wektor średnich

Wektor średniej w ich w próbie $\bar{\mathbf{y}}$ można zdefiniować jako:

$$\bar{\mathbf{y}} = \frac{1}{n} \sum_{i=1}^n \mathbf{y}_i = \begin{bmatrix} \bar{y}_1 \\ \vdots \\ \bar{y}_p \end{bmatrix}$$

Mówiąc nie o próbie, lecz o populacji możemy posłużyć się analogicznym zapisem:

$$E(\mathbf{y}) = E \begin{bmatrix} y_1 \\ \vdots \\ y_p \end{bmatrix} = \begin{bmatrix} E(y_1) \\ \vdots \\ E(y_p) \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_p \end{bmatrix} = \boldsymbol{\mu}$$

Macierz kowariancji można natomiast wyrazić jako:

$$\boldsymbol{\Sigma} = E[(\mathbf{y} - \mathbf{u})(\mathbf{y} - \mathbf{u})']$$

lub, co jest równoważne:

$$\boldsymbol{\Sigma} = \text{cov}(\mathbf{y}) = \begin{bmatrix} \sigma_{11} & \cdots & \sigma_{1p} \\ \vdots & \ddots & \vdots \\ \sigma_{p1} & \cdots & \sigma_{pp} \end{bmatrix}$$

Dopiero w przypadku wektora zawierającego wiele zmiennych możemy mówić o wielowymiarowym rozkładzie normalnym. Wektor \mathbf{y} ma wielowymiarowy rozkład normalny o wektorze średnich $\boldsymbol{\mu}$ i macierzy kowariancji $\boldsymbol{\Sigma}$:

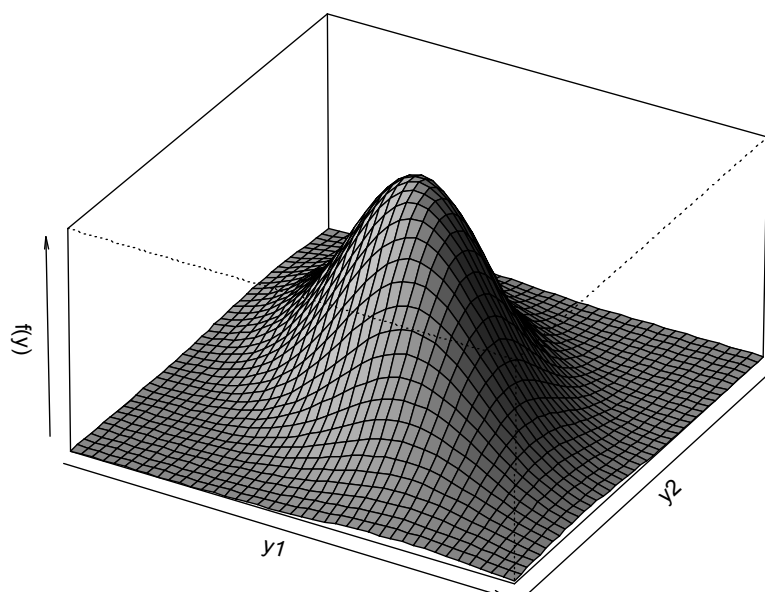
Projekt współfinansowany ze środków Unii Europejskiej w ramach Europejskiego Funduszu Społecznego.

$$g(\mathbf{y}) = \frac{1}{(\sqrt{2\pi})^p |\Sigma|^{1/2}} e^{-(\mathbf{y}-\mu)' \Sigma^{-1} (\mathbf{y}-\mu) / 2}$$

gdzie p wskazuje na ilość zmiennych. W skróconym zapisie wektor zmiennych o wielowymiarowym rozkładzie normalnym można zapisać jako $N(\mu, \Sigma)$.

Rysunek 16. przedstawia przykładowy wielowymiarowy rozkład normalny dla $p = 2$. Precyzyjnie jest to dwuwymiarowy rozkład normalny. Jeżeli wcześniej mówiliśmy, że rozkład normalny ma kształt przekroju dzwonu, to o rozkładzie normalnym dwóch zmiennych można powiedzieć, iż swoim trójwymiarowym kształtem przypomina on bryłę dzwonu. Przy czym kształt rozkładu definiowany jest przez wartość wyznacznika $|\Sigma|$: im jest ona większa, tym wierzchołek rozkładu jest bardziej płaski; im mniejsza, tym wierzchołek bardziej spiczasty.

Dwuwymiarowy rozkład normalny



Rysunek 16. Przykład dwuwymiarowego rozkładu normalnego

Jednowymiarowy rozkład normalny jest szczególnym przypadkiem rozkładu wielowymiarowego, dlatego między dwoma wzorami istnieją daleko idące analogie. Tabela 4. wskazuje na analogie najważniejsze w prezentacji wskaźnika trzyletniego.

Projekt współfinansowany ze środków Unii Europejskiej w ramach Europejskiego Funduszu Społecznego.

Tabela 4. Porównanie jednowymiarowego i wielowymiarowego rozkładu normalnego

	Rozkład normalny	Wielowymiarowy rozkład normalny
Wzór	$\frac{1}{\sqrt{2\pi}\sqrt{\sigma^2}} e^{-(y-\mu)^2/2\sigma^2}$	$\frac{1}{(\sqrt{2\pi})^p \Sigma ^{1/2}} e^{-(\mathbf{y}-\boldsymbol{\mu})' \Sigma^{-1} (\mathbf{y}-\boldsymbol{\mu})/2}$
Odległość od średniej	$(y-\mu)^2/\sigma^2 = (y-\mu) (\sigma^2)^{-1} (y-\mu)$	$(\mathbf{y}-\boldsymbol{\mu})' \Sigma^{-1} (\mathbf{y}-\boldsymbol{\mu})$
Parametr zróżnicowania	$\sqrt{\sigma^2}$	$ \Sigma ^{1/2}$

Można zauważyć, iż wyrażenie $(y-\mu)^2/\sigma^2 = (y-\mu) (\sigma^2)^{-1} (y-\mu)$ jest miarą dystansu między y i μ wyrażoną w jednostkach odchylenia standardowego (σ). Podobnie wyrażenie $(\mathbf{y}-\boldsymbol{\mu})' \Sigma^{-1} (\mathbf{y}-\boldsymbol{\mu})$ jest niczym innym, jak odległością Mahalanobisa, czyli odległością punktu w przestrzeni wielowymiarowej od wektora średnich $\boldsymbol{\mu}$.

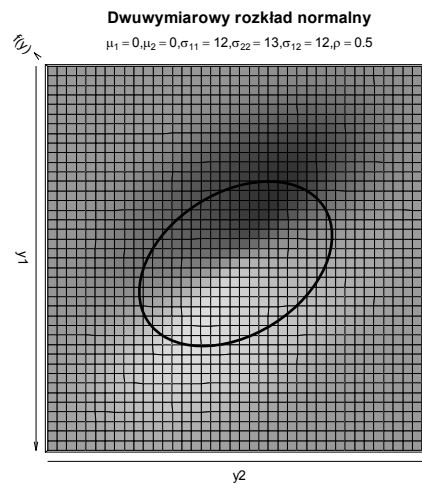
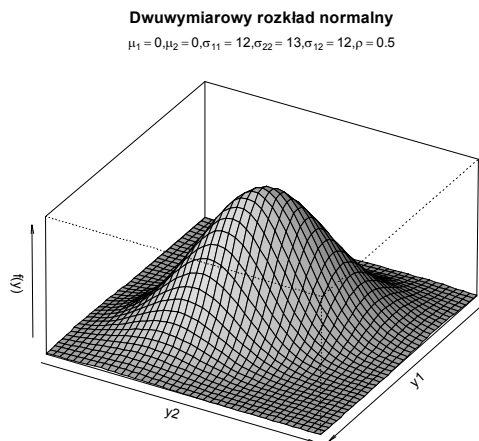
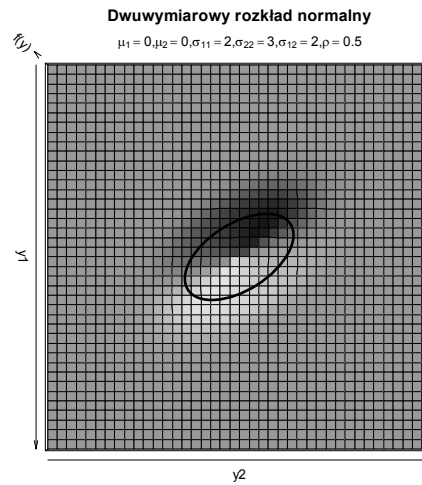
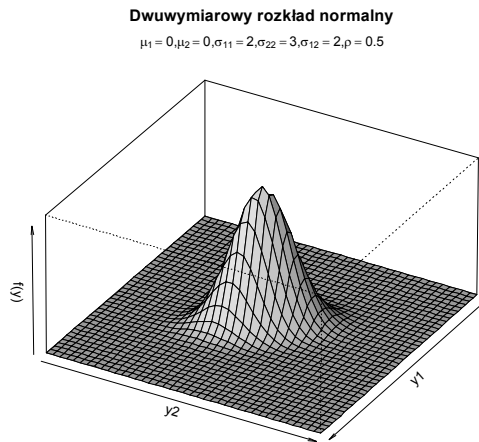
Natomiast wyrażenie $|\Sigma|^{1/2}$ będące pierwiastkiem wyznacznika macierzy jest tutaj analogiczne do $\sqrt{\sigma^2}$ we wzorze dotyczącym rozkładu normalnego. Jest podobne pod względem statystycznym. Oba wyrażenia mówią nam o zróżnicowaniu zmiennych (lub zmiennej w populacji): większa wartość wyznacznika Σ oznacza większe zróżnicowanie.

Ilustracją zależności kształtu od wyznacznika macierzy może być rysunek 17., gdzie dwa rozkłady zostały przedstawione w dwóch rzutach. Z lewej strony rysunku wielowymiarowy rozkład normalny przedstawiony został w perspektywie, a z prawej strony przedstawiony został rzut z góry. W dolnej części rysunku mamy do czynienia ze stosunkowo niewielkim $|\Sigma|$ – rozkład jest płaski, w górnej zaś ze stosunkowo dużym $|\Sigma|$ – kształt rozkładu jest wypiętrzony. Wierzchołkiem dwuwymiarowego rozkładu normalnego jest punkt wyznaczany przez wektor średnich:

$$E(\mathbf{y}) = E \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} E(y_1) \\ E(y_2) \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} = \boldsymbol{\mu}$$

Projekt współfinansowany ze środków Unii Europejskiej w ramach Europejskiego Funduszu Społecznego.

W prostej geometrycznej interpretacji $E(y)$ to punkt, w którym przecinają się wartości średnich: y_1 i y_2



Rysunek 17. Rozkłady normalne o różnym Σ

Natomiast dokładny kształt i nachylenie dwuwymiarowego rozkładu zależą od wartości zawartych w macierzy Σ . Gdy korelacja między zmiennymi wynosi 0, a wariancje są równe sobie, przekrój dwuwymiarowego rozkładu normalnego jest idealnym kołem. Szerokość rozkładu w wymiarze y_1 zależy od wielkości σ_{11} , w wymiarze y_2 od σ_{22} . Dlatego jeżeli σ_{22} będzie większe od σ_{11} , rozkład będzie rozciągnięty bardziej względem wymiaru y_2 . W odwrotnej sytuacji rozkład będzie rozciągnięty bardziej względem wymiaru y_1 .

Projekt współfinansowany ze środków Unii Europejskiej w ramach Europejskiego Funduszu Społecznego.

Gdy zmienne są ze sobą skorelowane, dochodzi również do „przekręcenia” rozkładu czy zmiany nachylenia elips, jakie mogą być tworzone na podstawie przekrojów poziomych rozkładu (elipsy można zauważyć przyglądając się prawej części rysunku 17.). Zależy to od różnicy między wariancjami dwóch zmiennych. Jeżeli σ_{22} jest większe od σ_{11} , rozkład przekręcony będzie w kierunku osi y_2 ; jeżeli natomiast σ_{22} będzie mniejsze od σ_{11} , rozkład przekręcony będzie w kierunku osi y_1 .

POWIERZCHNIA UFNOŚCI

Tak jak jednowymiarowy rozkład normalny jest szczególnym przypadkiem wielowymiarowego rozkładu normalnego, tak przedział ufności jest szczególnym przypadkiem powierzchni ufności. Tak jak w przypadku rozkładów normalnych, zaczniemy jednak od tego, co lepiej znane.

Aby obliczyć granice przedziału ufności dla jednej zmiennej, posługujemy się poniższym wzorem:

$$\mu \pm \sigma \sqrt{\chi^2(1 - \gamma; 1) / n}$$

Gdzie $\chi^2(1 - \gamma; 1)$ oznacza wartość testu Chi-kwadrat o jednym stopniu swobody dla obszaru krytycznego γ . I tak dla najpopularniejszego przykładu: jeżeli decydujemy się na 95% przedział ufności, czyli wartość krytyczną $\gamma = 0,05$

$$\mu + 1,96 \frac{\sigma}{\sqrt{n}}$$

W przypadku, gdy analizujemy dwie zmienne, operacje tę możemy wykonać dwa razy niezależnie, tak jak zostało to przedstawione na rysunku, gdzie mamy do czynienia ze zmienną charakteryzującą EWD (E) oraz zmienną charakteryzującą wyniki końcowe (W). Poprzez złożenie tych dwóch przedziałów ufności otrzymujemy prostokąt (na rysunku 18. zaznaczony przerywaną linią). Patrząc, czy punkt $E(\mathbf{y})$ różni się statystycznie od jakiegokolwiek innego punktu na tej zadanej płaszczyźnie, będziemy sprawdzać, czy punkt odniesienia będzie mieścił się w utworzonym prostokącie, czy nie. Jeżeli tak, nie będziemy mogli powiedzieć, iż jest on statystycznie różny od punktu $E(\mathbf{y})$ przy danym przedziale ufności. Jest to jednak podejście błędne. Nie zakłada ono, iż dwie zmienne są ze sobą skorelowane oraz tego, iż test statystyczny powinien dotyczyć dwóch parametrów jednocześnie. W takim wypadku musimy inaczej zdefiniować powierzchnię ufności.



Projekt współfinansowany ze środków Unii Europejskiej w ramach Europejskiego Funduszu Społecznego.

Jeżeli $y = N(\mu, \Sigma)$, zatem wektor średnich z próby można zapisać jako $m = N(\mu, \Sigma/n)$. Natomiast funkcja gęstości jest proporcjonalna do $e^{-(y-\mu)'\Sigma^{-1}(y-\mu)/2}$. Pozostawiając wszystkie stałe z równania wielowymiarowego rozkładu normalnego, możemy zapisać, że funkcja gęstości m jest proporcjonalna do

$$Q(m) = (y - \mu)'\Sigma^{-1}/n(y - \mu)$$

a funkcja ta ma rozkład Chi-kwadrat tak, że:

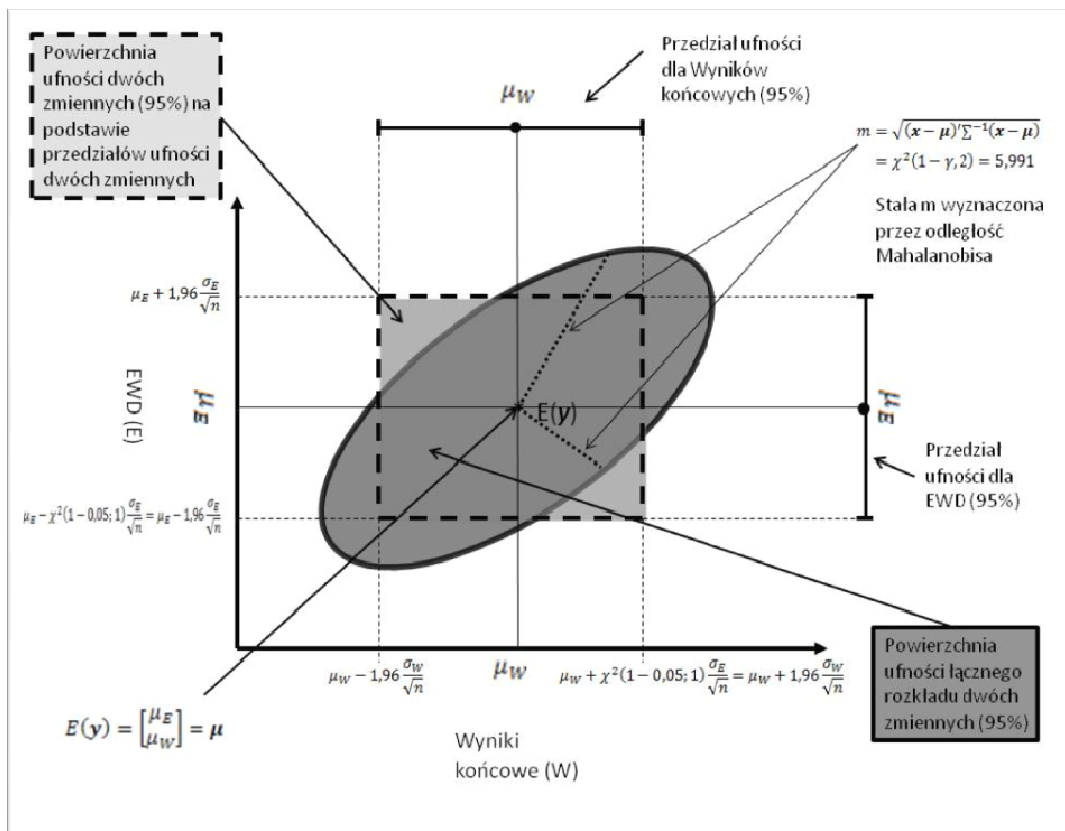
$$Q(m) = \chi^2(1 - \gamma, p)$$

W przypadku naszej dwuwymiarowej powierzchni ufności ($p = 2$) możemy zapisać, że:

$$n[m_1 - \mu_1 \quad m_2 - \mu_2] \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix}^{-1} \begin{bmatrix} m_1 - \mu_1 \\ m_2 - \mu_2 \end{bmatrix} = \chi^2(1 - \gamma, p)$$

Jeżeli przyjmiemy obszar krytyczny 0,05, to $\chi^2(1 - 0.05; 2) = 5,991$. Tak wyspecyfikowane równanie wyznacza zbiór punktów, które zamykają powierzchnię ufności rozkładu dwóch zmiennych (por. Chew 1966). Ów zbiór punktów przybiera kształt elipsy, co zostało zaprezentowane na rysunku 18.

Projekt współfinansowany ze środków Unii Europejskiej w ramach Europejskiego Funduszu Społecznego.

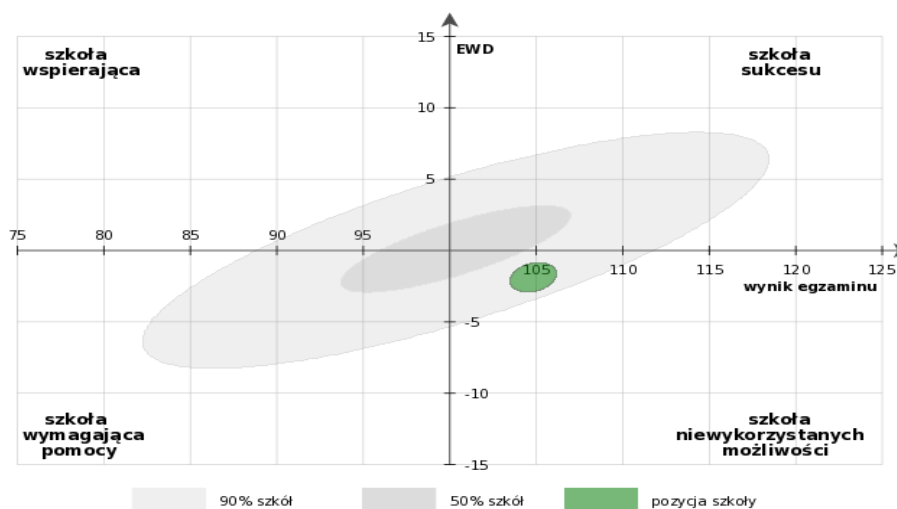


Rysunek 18. Elipsa jako powierzchnia ufności

Projekt współfinansowany ze środków Unii Europejskiej w ramach Europejskiego Funduszu Społecznego.

PODSUMOWANIE

Efektom końcowym procedur statystycznych omawianych w powyższych punktach jest graficzna prezentacja wskaźnika trzyletniego, tak jak zostało to przedstawione na rysunku 19. 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 pozioma 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 analogicznie do powierzchni ufności w punkcie 4 z tym, że powierzchni ufności odnoszącej się do populacji szkół, a nie uczniów jednej szkoły, oraz – co warto podkreślić – elipsy te szacowane są iteracyjnie. Szukana jest taka odległość Mahalanobisa, która wykreśli elipsę, która zawiera kolejno 50% populacji szkół oraz 90% populacji szkół.



Rysunek 19. Graficzne przedstawienie trzyletniego wskaźnika egzaminacyjnego: przykład

Projekt współfinansowany ze środków Unii Europejskiej w ramach Europejskiego Funduszu Społecznego.

LITERATURA

1. Agresti (2004) Alan, *Categorical Data Analysis*, John Wiley & Sons, Inc, Hoboken, New Jersey.
2. Agresti (2007) Alan, *An Introduction to Categorical Data Analysis*, John Wiley & Sons, Inc, Hoboken, New Jersey.
3. Blalock (1984) Hubert M., *Contextual-Effects Models: Theoretical and Methodological Issues*, "Annual Review of Sociology", 10: 353-372.
4. Chew (1966) Victor, *Confidence, Prediction, and Tolerance Regions for the Multivariate Normal Distribution*, "Journal of the American Statistical Association" Vol. 16 No. 315
5. DiPrete (1994) Thomas A. i Jerry D. Forristal, *Multilevel Models: Methods and Substance*, "Annual Review of Sociology", 20: 331-357.
6. Dolata (2007) Roman (red), *Edukacyjna wartość dodana jako metoda oceny efektywności nauczania na podstawie wyników egzaminów zewnętrznych*, CKE, Warszawa 2007.
7. Kolen (2004) M. J. i R. L. Brennan, *Test Equating, Scaling and Linking Methods and Practices Second Edition*, Springer, New York.
8. Goldstein (1999) Harvey, *Multilevel Statistical Models*, London: Institute of Education, Multilevel Models Projects, kwiecień.
[<http://www.arnoldpublishers.com/supportgoldstein.htm>].
9. Gujarati (2004) Damodar N., *Basic Econometrics, Fourth Edition*, The McGraw-Hill Companies.
10. Härdle (2003) Wolfgang i Simor Léopold, *Applied Multivariate Statistical Analysis*, Springer Verlag, Berlin.
11. Hox (1995) Joop, *Applied Multilevel Analysis*, TT-Publikaties, Amsterdam.
12. Hox (2002) Joop, *Multilevel Analysis Techniques and Applications*, Lawrence Erlbaum Associates, Publishers, Mahwah, New Jersey – Londyn.
13. Hox(2002b) Joop, *Multilevel Modeling: When and Why*, [w:] I Balderjahn, R. Mathar i M. Schader (red), *Classification, data analysis, and data highways*. New York: Springer Verlag.
14. Leeuw (2008), Jan de i Erik Meijer, *Handbook of Multilevel Analysis*, Springer, New York.
15. Luke (2004) Douglas A., *Multilevel Modeling*, Sage, Thousand Oaks – London – New Dheli.
16. Maddala (2006) G.S, *Ekonometria*, Wydawnictwo Naukowe PWN, Warszawa.
17. Mass (2000) J.M Cora, Hox Joop, *Sample Size for Multilevel Modeling*, Department of Methodology and Statistics, Faculty of Social Sciences, Utrecht University, [working paper].
18. Megiera (1994) Ryszard, *Modele i metody statystyki matematycznej*, Wydawnictwo GiS Gewert i Skoczylas, Wrocław.
19. Pawłowski (1981) Zbigniew, *Statystyka matematyczna*, PWN, Warszawa
20. Rabe-Hesketh (2008) Sophia i Anders Skrondal, *Multilevel and Longitudinal Modeling Using Stata*, A Stata Press Publication – StataCorp LP, College Station, Texas.
21. Raudenbush (2002) W. Stephen i Anthony S. Bryk, *Hierarchical Linear Models. Applications and Data Analysis (Second Edition)*, Sage, Thousand Oaks – London – New Dheli.
22. Silvey (1978) S.D., *Wnioskowanie statystyczne*, PWN, Warszawa.
23. Snijders (2004) Tom A.B. i Roel J. Bosker. *Multilevel Analysis*, Sage, Thousand Oaks – London – New Dheli.



Projekt współfinansowany ze środków Unii Europejskiej w ramach Europejskiego Funduszu Społecznego.

24. Snijders (2005), Tom A.B. *Fixed and Random Effects*. [w]: B.S. Everitt and D.C. Howell (red.), *Encyclopedia of Statistics in Behavioral Science*. Volume 2, 664-665, Wiley, Chichester.
25. Sorensen (2002) Daniel i Daniel Gianola *Likelihood, Bayesian, and MCMC Methods in Quantitative Genetics*, Springer-Verlag New York Berlin Heidelberg
26. Steenbergen (2002) Marco R. i Bradford S. Jones, *Modeling Multilevel Data Structures*, *American Journal of Political Science*, 46/1, 218-237.
27. Wooldridge (2006) Jeffery M., *Introductory Econometrics. A modern approach 2e*, South-Western College Publishing.