

Metody szacowania edukacyjnej wartości dodanej

1. Definicja i zastosowanie wartości dodanej.

1.1. Definicja wartości dodanej.

Wartość dodaną szkoły można zdefiniować jako przeciętną wzrostu umiejętności i wiedzy uczniów do niej uczęszczających w danym okresie czasu. O ile ma to być wskaźnik jakości pracy szkoły, to w określonych przypadkach wartość dodaną należałoby kalkulować po wyłączeniu czynników warunkujących przyrost wiedzy, na które szkoła nie może mieć wpływu, w tym różnej prędkości pozyskiwania wiedzy przez uczniów oraz obiektywnych warunków nauczania danej placówki. *Wartość dodaną nauczyciela* można zdefiniować jako miarę wpływu jego pracy na wzrost umiejętności i wiedzy podlegających mu uczniów po wyłączeniu czynników warunkujących przyrost ich wiedzy, na które szkoła i nauczyciel nie mogą mieć wpływu. Wyłączenie wszystkich tych czynników jest bardzo trudne. Kluczowy jest tu także problem rozdzielenia wpływu szkoły od wpływu nauczycieli. Nauczyciele stanowią bowiem o charakterze szkoły i ich praca ma wpływ nie tylko na uczniów, z którymi pracują, ale i może mieć wpływ na wszystkich uczniów w szkole.

W metodach przedstawionych poniżej wartość dodana szkoły obliczana jest na podstawie różnic między rzeczywistymi a oczekiwanymi wynikami uczniów do niej uczęszczających. Kluczowym problemem jest tu oszacowanie oczekiwanego wyniku ucznia i celem tej pracy jest dokonanie przeglądu metod, które mogą być w tym celu wykorzystane w kontekście egzaminów zewnętrznych i systemu szkolnictwa w Polsce. Warto też zauważyć, że choć w tekście skupiam się na liczeniu edukacyjnej wartości dodanej (dalej EWD) dla szkoły, to wszystkie metody mogą być także wykorzystane do wykonania podobnych obliczeń dla klasy lub innej, dowolnie zdefiniowanej grupy uczniów.

¹ Wydział Nauk Ekonomicznych, Uniwersytet Warszawski, email: mjakubowski@uw.edu.pl, tel. 0 602 748 740.

1.2. Zastosowanie wartości dodanej a metody jej szacowania.

Metody szacowania edukacyjnej wartości dodanej (dalej EWD) i rodzaj wymaganych danych muszą być podporządkowane temu, jaki charakter informacji mają tworzyć i kto będzie jej odbiorcą. Wartość dodana może przede wszystkim służyć jako informacja o:

- A. Efektach pracy szkoły skierowana do rodziców.
- B. Efektach pracy szkoły skierowana do organów prowadzących lub nadzoru pedagogicznego.
- C. Efektach pracy nauczycieli skierowana do organów prowadzących lub nadzoru pedagogicznego.

W przypadku A wartość dodana powinna być liczona dla całej placówki, bez wyłączenia wpływu zasobów szkolnych (zarówno materialnych, jak i np. składu społecznego szkoły), bowiem z punktu widzenia rodziców interesujący jest całościowy, potencjalny wpływ danej placówki na wzrost wiedzy ich dziecka. Ze względu na to, szacowanie wartości dodanej dla rodziców nakłada najmniejsze wymagania co do skomplikowania modeli i zakresu niezbędnych danych. Jednak sposób podawania informacji o wartości dodanej powinien uwzględniać różną jej wielkość w zależności od poziomu ucznia lub też określać w jakim przedziale umiejętności uczniów wartość dodana danej szkoły została wypracowana (por. Meyer, 1997).

Zupełnie inne wymagania nakłada szacowanie wartości dodanej celem oceny efektów pracy szkoły przez organy prowadzące lub nadzoru pedagogicznego. W tym przypadku celowe jest uwzględnienie zarówno zasobów materialnych dostępnych szkole, jak i środowiska w jakim ona pracuje. Jeśli szkoły znacząco różnią się np. bazą materialną i ma ona wpływ na osiągnięcia uczniów, to jej nieuwzględnienie w modelu powoduje, że placówki z lepszą bazą uzyskują dodatkowy bonus. Choć badania pokazują, że wpływ bazy materialnej jest znikomy (por. Hanushek, 2003), to już np. wpływ składu społecznego szkoły może być znaczący (por. Hoxby, 2000; Markman, 2003). Jednak wciąż nie wiemy dokładnie jakiego rodzaju jest to zależność, jakie cechy otoczenia społecznego ucznia wpływają na jego wyniki i czy dla wszystkich grup uczniów związek ten ma podobny charakter (por. Wilkinson i in., 2000; Kutnick i in., 2005).

Powyższa dyskusja wskazuje na dwa kluczowe problemy przy szacowaniu wartości dodanej celem oceny jakości pracy szkół:

- a) braku danych - zarówno ze względu na koszty ich pozyskania, jak i trudności pomiaru;
- b) braku teoretycznej i empirycznie potwierdzonej wiedzy na temat czynników wpływających na wzrost wiedzy ucznia (warunkujących proces nauczania).

Problem braku danych dotyczy wszystkich krajów. Nie zawsze jest on związany z niedojrzałością systemu danych oświatowych (jak np. w Polsce) lub wysokich kosztów jego utrzymania. Często dane niezbędne do precyzyjnego oszacowania wartości dodanej nie mogą być zbierane lub wykorzystywane ze względu na ich poufność (dane osobowe) lub też ze względów społecznych (np. zamożność) i politycznych (np. rasa, narodowość). Dodatkowo, ważne czynniki warunkujące osiągnięcia edukacyjne uczniów, takie jak ich talent, motywacja do pracy lub też zaangażowanie rodziców, są trudne lub wręcz niemożliwe do systematycznego mierzenia. W przypadku modeli szacujących wartość dodaną, podobnie jak w przypadku większości nieeksperymentalnych badań oświatowych, mamy zazwyczaj do czynienia z problemem endogeniczności zmiennych w modelu².

Równocześnie, szacowanie wartości dodanej utrudnia brak potwierdzonej empirycznie wiedzy dotyczącej czynników warunkujących osiągnięcia i wzrost wiedzy uczniów. Wieloletnie wysiłki badań oświatowych nie doprowadziły do zgody nawet co do siły wpływu czynników rodzinnych ucznia, nie mówiąc o wpływie zasobów edukacyjnych (por. Hanushek, 2003). Powoduje to, że nie do końca wiadomo, jakich zmiennych powinno się użyć w modelach wartości dodanej. Nie wiadomo też jakie dane powinno się zbierać, co przy znacznych kosztach i trudnościach w ich kolekcjonowaniu utrudnia rozwój systemu oceny szkół. Dodatkowo, badania empiryczne prowadzone są najczęściej dla danych z USA, co ze względu na odmienność kontekstu kulturowego oraz brak wiarygodnych badań ilościowych na danych krajowych rodzi oczywiste problemy w krajach takich jak Polska. Wszystko to powoduje, że wiele osób podważa zasadność wykorzystywania wartości dodanej jako narzędzia oceny jakości pracy szkoły powiązanego z systemem nagród i kar, tak jak to ma miejsce w niektórych stanach USA (por. Meyer 1997).

Jeszcze trudniejszym zadaniem, a w związku z tym rodzącym większe wątpliwości, jest wykorzystanie wartości dodanej dla oceny pracy nauczycieli. Przede wszystkim przy takim

² Endogeniczność w tym kontekście oznacza pozorny związek między wartością dodaną a zmiennymi kontrolnymi wynikający ze skorelowania tych zmiennych z cechami nie uwzględnionymi w modelu. Inaczej mówiąc zmienne uwzględnione związane są ze zmiennymi pominiętymi, które oddziałują na zmienną zależną, przez co oszacowania współczynników w równaniu regresji są błędne (por. klasyczny przykład badań nad wpływem wielkości klasy na wyniki: Akerhielm, 1995; Hoxby, 2000; Jakubowski, Sakowski, 2005). Więcej na ten temat można znaleźć w: Lee, 2005.

zastosowaniu niezbędne jest przypisanie uczniów do poszczególnych nauczycieli, co technicznie nie jest zadaniem łatwym, szczególnie jeśli mobilność uczniów i pedagogów jest wysoka. Rozstrzygnięcia wymagają kwestie zarówno przypisania uczniów zmieniających klasy i szkoły, jak i powiązania nauczycieli różnych przedmiotów z wynikami testów, które najczęściej mają charakter wielopredmiotowy. Ponadto, wpływ nauczycieli jest niezwykle trudny do oddzielenia od wpływu szkoły jako całości, o czym wspomniałem na wstępie. Jest to zadanie bardzo trudne, bowiem mamy tu do czynienia z mało wymiernymi czynnikami, np. atmosferą szkoły, która w pewnym stopniu stanowi cechę całej placówki, ale i zależy od starań poszczególnych osób z kadry pedagogicznej.

Kwestią niezwykle ważną jest też precyzja oszacowań, która w przypadku nauczycieli będzie zawsze mniejsza, niż w przypadku szkoły, ze względu na mniejszą liczbę obserwacji. W szczególności, problematyczne jest porównywanie wartości dodanej nauczycieli i szkół z niewieloma uczniami ze szkołami większymi. Pewnym rozwiązaniem jest tu szacowanie wpływu nauczycieli jako efektu losowego celem zmniejszenia ryzyka popełnienia błędu (przypisania krańcowo niskiej lub wysokiej wartości dodanej). Jednak takie podejście jest wymagające obliczeniowo i opiera się na czasem zbyt restrykcyjnych założeniach teoretycznych (por. McCaffrey, 2005).

Podsumowując tę dyskusję, należy zauważyć, że mimo wielu wad i problemów praktycznych związanych z szacowaniem wartości dodanej i jej praktycznym wykorzystywaniem dla oceny jakości pracy szkół i nauczycieli, metody tego typu dają znacznie większy zasób informacji niż surowe wyniki egzaminów. Ich świadome, ostrożne stosowanie, wraz z odpowiednim opisem i oparciem w innych wskaźnikach, może wprowadzać do systemu oświatowego niezwykle wartościową informację. Informację niezbędną dla rodziców decydujących o wyborze placówki, dla organu prowadzącego zarządzającego szkołami, a także dla nadzoru pedagogicznego, dając obiektywny wskaźnik jakości pracy nauczycieli i szkół. Wartość dodana ma też szerokie zastosowania w badaniach edukacyjnych i może być wykorzystywana jako cenne źródło oceny polityki oświatowej, zróżnicowania jakości nauczania, a także jako podstawa dla programów wspierających szkoły i regiony o szczególnych potrzebach.

1.3. System egzaminacyjny a możliwości oszacowania wartości dodanej.

Można wyróżnić dwa ogólne typy systemów egzaminacyjnych, które zasadniczo różnią się możliwościami szacowania wartości dodanej (por. McCaffrey i in., 2005):

- A. System, w którym egzaminy przeprowadzane są jednokrotnie na danym etapie kształcenia.
- B. System, w którym egzaminy przeprowadzane są kilka razy na danym etapie kształcenia, corocznie lub nawet częściej.

System typu A odpowiada egzaminom zewnętrznym w Polsce i wielu innych krajach europejskich. Na takich systemach skupiam się w tekście poniżej, jednak trzeba zauważyć, że z punktu widzenia szacowania wartości dodanej przynoszą one niewielki zasób informacji i wykluczają bardziej zaawansowane metody, możliwe do zastosowania w systemach typu B, rozwijanych od 20 lat w niektórych stanach USA. Zalety i ograniczenia systemów wieloletnich omawiam krótko w punkcie 2.6. Tutaj wskażę tylko, że tworzą one możliwość budowania wielopoziomowych modeli biorących pod uwagę „ścieżkę” rozwoju danego ucznia, co pozwala ocenić wzrost jego umiejętności w czasie i wpływ nakładów edukacyjnych na kilku poziomach kształcenia.

Inną, ważną cechą systemów jest stopień zróżnicowania egzaminów. W niektórych krajach kolejne egzaminy, przynajmniej na określonym etapie kształcenia, mierzą podobne umiejętności i przekładane są na jedną skalę pomiarową. W ten sposób różnice między wynikami w naturalny sposób mierzą wzrost wiedzy ucznia. Jednak w innych krajach egzaminy są na tyle różne, że ich przekładanie na jedną skalę jest trudne lub wręcz niemożliwe, a różnice w pomiarze stają się kluczowym problemem przy szacowaniu wartości dodanej.

1.4. Różnice indywidualne a wartość dodana.

Opisane poniżej metody nie powinny być wykorzystywane dla oceny wzrostu wiedzy pojedynczych uczniów. Należy stanowczo podkreślić, że w przypadku polskich egzaminów szacowanie indywidualnej wartości dodanej uczniów nie jest możliwe. Co prawda w trakcie obliczania wartości dodanej dla szkoły wykorzystywane są różnice między wynikami oczekiwanymi a uzyskanymi przez uczniów, to jednak ich wykorzystanie dla oceny przyrostu ich wiedzy i umiejętności jest nieuprawnione ze względu na zbyt duże znaczenie błędu pomiaru.

Zakładając, że błąd ten jest losowy oczekujemy, że nie będzie miał on wpływu na przeciętną różnicę między wartościami oczekiwanymi a rzeczywistymi. Inaczej mówiąc błędy pomiaru wiedzy uczniów „znoszą się” i dla odpowiednio licznej grupy nie mają znaczenia przy szacowaniu wartości dodanej. Jednak w przypadku pojedynczych uczniów stanowią one kluczowy problem i dla metod wykorzystujących jedynie wyniki z dwóch lat powodują, że ocena indywidualnych postępów nie jest możliwa.

1.5. Kryteria i ocena wyboru metody szacowania wartości dodanej.

Dokonanie wyboru właściwej metody szacowania wartości dodanej wymaga określenia kryterium ich oceny. Można przyjąć, że wartość dodana ma służyć ocenie jakości nauczania w szkole niezależnie od tego, jaki jest poziom wiedzy i umiejętności uczniów, którzy do niej uczęszczają. O ile zgodzimy się z takim założeniem, to narzucającym się i stosunkowo łatwym do zastosowania kryterium jest to, aby dla każdego poziomu wyników egzaminu na niższym poziomie przeciętna różnica między rzeczywistym a oczekiwanym wynikiem egzaminu na wyższym poziomie dążyła do zera. Inaczej mówiąc, najlepszą metodą będzie taka, która w całej populacji uczniów traktuje podobnie grupy o różnych wynikach uzyskanych na egzaminie na niższym poziomie. W kontekście sprawdzianu i egzaminu gimnazjalnego będzie to metoda, dla której przeciętna różnica między wynikiem oczekiwanym a uzyskanym na egzaminie będzie bliska zeru dla grup uczniów o różnych wynikach uzyskanych na sprawdzianie. O ile dana metoda spełnia to kryterium, to oszacowana nią wartość dodana szkoły nie będzie zależeć od poziomu uczniów do niej uczęszczających.

To podstawowe kryterium może być rozszerzone o kryteria odwołujące się do neutralności metody szacowania wartości dodanej względem lokalizacji szkoły (np. miasto a wieś), jej rozmiaru, czy też pochodzenia społecznego jej uczniów. Kryteria te będą jednak w dużym stopniu pokrywały się z opisanym powyżej. Innym, ważnym kryterium może być ocena trafności metody szacowania wartości dodanej jako metody oceny jakości pracy szkoły. Kryteria te powinny badać, czy wartość dodana oszacowana daną metodą jest skorelowana z czynnikami decydującymi o rzeczywistej jakości nauczania. W tym przypadku niezbędne jest jednak wykorzystanie danych, które zazwyczaj nie są kolekcjonowane w bazach danych systemów egzaminacyjnych. Ponadto, oczywistym problemem jest określenie czynników wpływających na jakość pracy placówek, a które można zmierzyć i analizować metodami ilościowymi.

2. Wybrane metody szacowania wartości dodanej.

W świetle powyższych uwag i charakteru egzaminów w Polsce uzasadnione wydaje się skoncentrowanie na metodach szacowania wartości dodanej opierających się na wynikach dwóch egzaminów, uwzględniając, że są one przedstawiane na odmiennej skali pomiaru. We wszystkich metodach zmienną objaśnianą (zależną) jest wynik egzaminu na wyższym etapie, a zmienną objaśniającą (niezależną) jest wynik egzaminu na niższym etapie kształcenia. Pod koniec rozdziału w punktach 2.5 i 2.6 omawiam też nieco inaczej definiowane modele bezpośrednio wyjaśniające różnicę między wynikami dwóch egzaminów i modele wieloletnie.

Wszystkie metody opisane zostaną wzorami, w których przyjmuję wspólną notację:

- y_i - wynik i -tego ucznia na egzaminie na wyższym poziomie (np. gimnazjalnym)
- x_i - wynik i -tego ucznia na egzaminie na niższym poziomie (np. sprawdzian)
- d_i - różnica między wynikiem uzyskanym a oczekiwanym dla i -tego ucznia
- D_j - wartość dodana j -tej szkoły
- S_j - zbiór uczniów j -tej szkoły
- n_j - liczba uczniów j -tej szkoły
- $CT(z)$ - miara tendencji centralnej zmiennej z
- $CT(y | x = x_i)$ - miara tendencji centralnej wyników egzaminu na wyższym poziomie uzyskanych przez wszystkich uczniów o tym samym wyniku co i -ty uczeń na egzaminie na niższym poziomie.

Z miar tendencji centralnej jako $E(z)$ oznaczam średnią, a jako $Med(z)$ medianę zmiennej z .

2.1. Szacowanie wartości dodanej przez odniesienie wyniku ucznia do osiągnięć uczniów na podobnym poziomie.

Jest to najprostsza metoda, która różni się od kolejnych tym, że nie zakłada niczego o kształcie związku między wynikami egzaminów na poszczególnych etapach kształcenia. Punktem odniesienia w tej metodzie jest centralna tendencja, najczęściej mediana lub średnia, wyników egzaminu na wyższym etapie uzyskanych przez uczniów, którzy mieli ten sam wynik na egzaminie na etapie niższym. Tak więc na pierwszym etapie tej metody obliczamy dla każdego ucznia różnicę między uzyskanym przez niego rzeczywistym wynikiem a właściwym mu punktem odniesienia:

$$d_i = y_i - CT(y | x = x_i) \quad (1)$$

a w szczególności, kiedy za miary tendencji centralnej wybraliśmy średnią lub medianę:

$$d_i = y_i - E(y | x = x_i) \quad \text{lub} \quad d_i = y_i - Med(y | x = x_i) \quad (1a)$$

Na drugim etapie obliczamy centralną tendencję tych różnic dla wszystkich uczniów danej szkoły, które to stanowią jej wartość dodaną:

$$D_j = CT(d_i) \text{ dla wszystkich } i \in S_j \quad (2)$$

W szczególności centralną tendencję można określić przez średnią, medianę, średnią obciętą lub inne miary. Warto zauważyć, że miary tendencji centralnej we wzorach (1) i (2) mogą być różne. W (1) obliczane są one dla bardzo dużej próby (wszyscy uczniowie w kraju o tym samym wyniku na sprawdzianie), a w (2) zazwyczaj dla znacznie mniejszej liczby uczniów (wszyscy zdający egzamin gimnazjalny w danej szkole w danym roku).

Zaletą tej metody jest jej prostota. Jedynie obliczenie punktów odniesienia dla poszczególnych grup uczniów wymaga danych ogólnokrajowych. O ile zostałyby one opublikowane, to dalsze szacunki rodzice lub nauczyciele mogą bez problemu przeprowadzić samodzielnie³.

Metoda ta nie zakłada kształtu związku funkcyjnego między wynikami egzaminów, co może stanowić jej wadę w sytuacji, gdy średnie warunkowe opisane powyżej nie będą zawsze rosły

³ W ten sposób prezentowana jest wartość dodana w Anglii, gdzie jako punkt odniesienia stosuje się medianę wyników uczniów na tym samym poziomie, a wartość dodaną szkoły stanowi średnia indywidualnych różnic (por. Bartmańska, 2004). Jednak w Anglii poziom wejściowy i wyjściowy szacowany jest na podstawie wielu ocen a nie jednokrotnego egzaminu (więcej na: www.standards.dfes.gov.uk). Od niedawna wprowadzono tam też tzw. kontekstową wartość dodaną, opierającą się na metodzie regresji ze zmiennymi kontrolnymi, podobną do opisanej poniżej.

wraz z wynikami egzaminu na niższym poziomie. Innym zasadniczym jej brakiem jest trudność w uwzględnieniu zmiennych kontrolnych. Co prawda możliwe jest policzenie średnich warunkowych dla różnych kombinacji wyników egzaminu na etapie niższym i zmiennych kontrolnych, to jednak jedynie w przypadku, gdy te drugie są zmiennymi zerojedynkowymi lub jakościowymi. Wzrost liczby wartości zmiennych kontrolnych powoduje, że problem braku monotoniczności, wskazany powyżej, będzie coraz silniejszy ze względu na malejącą liczebność grup, dla których obliczane są wyniki stanowiące punkty odniesienia. Ponadto metoda ta będzie coraz bardziej pracochłonna. W przypadku ciągłych zmiennych kontrolnych jej zastosowanie nie ma uzasadnienia.

2.2. Szacowanie wartości dodanej przez reszty równania regresji.

Ta metoda szacowania wartości dodanej opiera się na równaniu regresji indywidualnych wyników egzaminu na niższym poziomie na wyniki egzaminu na wyższym poziomie. W metodzie tej konieczne jest przyjęcie założenia o kształcie związku między y i x , co nie było potrzebne w metodzie poprzedniej. Zaczniemy od najprostszego przypadku zależności liniowej. W takiej sytuacji na pierwszym etapie szacujemy metodą najmniejszych kwadratów równanie regresji na zbiorze wyników wszystkich uczniów:

$$\hat{y}_i = \hat{\alpha} + \hat{\beta}_1 x_i \quad (3)$$

gdzie znak „ $\hat{}$ ” przy współczynnikach równania oznacza ich ocenę dla danej próby, a przy y oznacza wynik oczekiwany egzaminu na wyższym poziomie dla określonej wartości wyniku z egzaminu na niższym poziomie⁴. Następnie obliczamy różnicę między rzeczywistym a oczekiwanym wynikiem dla każdego ucznia:

$$d_i = y_i - \hat{y}_i \quad (4)$$

Wartość dodaną szkoły obliczamy wg wzoru (2) jako wybraną miarę tendencji centralnej indywidualnych różnic.

Włączenie do analizy nawet szerokiego zestawu różnego typu zmiennych kontrolnych nie sprawia w tej metodzie technicznego problemu. W równaniu (3) możemy uwzględnić wektor

⁴ Opis metody najmniejszych kwadratów i analizy regresji liniowej można znaleźć w każdym podręczniku ekonometrii, np. Gujarati, 2002.

zmiennych \mathbf{R}_i zawierający charakterystyki rodziny oraz cech indywidualnych i -tego ucznia, a także wektor zmiennych \mathbf{P}_j określających warunki nauczania, nakłady oświatowe, środowisko społeczne w j -tej placówce:

$$\hat{y}_i = \hat{\alpha} + \hat{\beta}_1 x_i + \mathbf{R}'_i \hat{\beta}_2 + \mathbf{P}'_j \hat{\beta}_3 \quad (3a)$$

Współczynniki tego równania szacujemy podobnie jak poprzednio metodą najmniejszych kwadratów, a wartość dodaną szkoły obliczamy za pomocą wzorów (4) i (2).

2.3. Efekty stałe.

Poprzednia metoda obliczania wartości dodanej daje obciążone rezultaty, jeśli grupy uczniów o różnym poziomie wyników egzaminacyjnych uczęszczają do szkół o różnej efektywności nauczania⁵. Jeśli np. słabsi uczniowie uczęszczają do gorszych szkół, a lepsi uczniowie do szkół o wyższej jakości nauczania, to opisana powyżej metoda będzie zawyżać oszacowaną wartość dodaną dla szkół z mniej zdolnymi uczniami i zaniżać dla szkół z bardziej zdolnymi. W takim przypadku lepszym rozwiązaniem jest włączenie do powyższych równań regresji z efektów stałych szkół (klas lub nauczycieli), które stanowiąc będą oszacowania ich wartości dodanej:

$$\hat{y}_i = \hat{\alpha}_0 + \sum_j \hat{\alpha}_j Z_j + \hat{\beta}_1 x_i \quad (5)$$

gdzie $Z_j=1$ dla uczniów uczęszczających do j -tej szkoły i $Z_j=0$ dla innych, a znak sumy oznacza sumowanie po wszystkich szkołach. W ten sposób $D_j = \hat{\alpha}_j$ stanowi oszacowanie wartości dodanej j -tej szkoły.

⁵ Obciążenie oznacza w tym przypadku, że oceny wartości dodanej systematycznie różnią się od ich rzeczywistego poziomu.

Oczywiście w metodzie tej także można uwzględnić zmienne kontrolne na poziomie ucznia i szkoły:

$$\hat{y}_i = \hat{\alpha}_0 + \sum \hat{\alpha}_j Z_j + \hat{\beta}_1 x_i + \mathbf{R}'_i \hat{\beta}_2 + \mathbf{P}'_j \hat{\beta}_3 \quad (5a)$$

gdzie przyjęto podobne definicje jak w równaniach (5) i (3a).

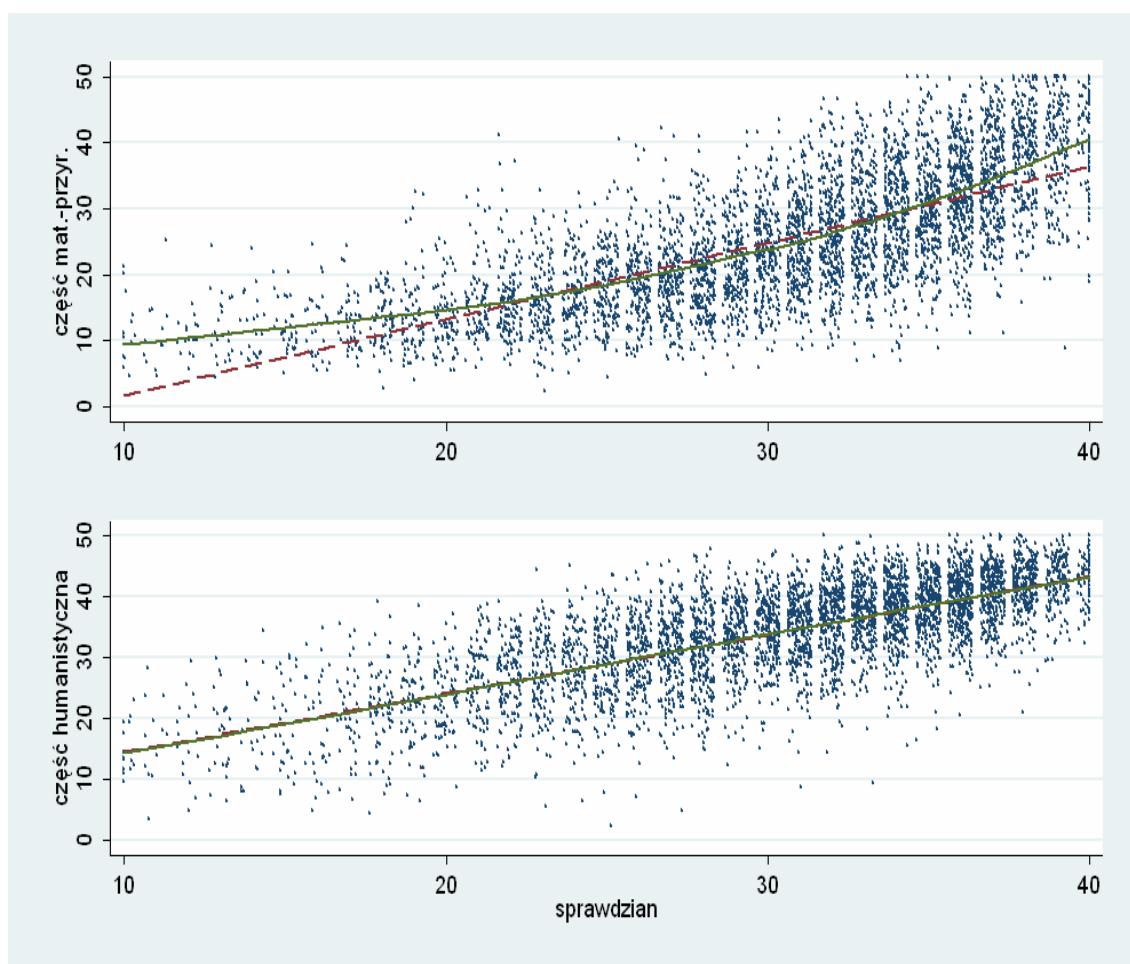
Należy zaznaczyć, że w modelu z efektami stałymi pakiety statystyczne najczęściej automatycznie wyłączają zmienną zerojedynkową dla pierwszej szkoły, aby uniknąć pełnej współliniowości. Za punkt odniesienia służy stała w równaniu dla tej placówki, a pozostałe efekty stałe szacowane są względem niej. Warto o tym pamiętać przy ich interpretacji.

2.4. Modelowanie nieliniowej zależności między wynikami egzaminów.

Zależność między wynikami egzaminów odzwierciedla nie tylko zmianę w poziomie umiejętności, ale i odmienne tempo przyrostu wiedzy dla różnych grup uczniów oraz rozbieżności w skalach pomiaru ich umiejętności na różnych egzaminach. Występowanie obu tych czynników powoduje, że zależność między wynikami egzaminów jest nieliniowa, przy czym warto podkreślić, że empirycznie rozróżnienie, który z nich ma decydujący wpływ na jej kształt jest często niemożliwe.

Problem odpowiedniej specyfikacji można ukazać na przykładzie relacji między wynikami uczniów na sprawdzianie w 2002 roku a wynikami w obu częściach egzaminu gimnazjalnego w 2005 roku. Zależności te ukazują wykresy poniżej, wykonane dla losowej próbki wyników uczniów jednego z województw. Widać, że dla części humanistycznej zależność liniowa całkiem dobrze opisuje relację między wynikami egzaminów. Dla części matematyczno-przyrodniczej związek między wynikami egzaminów nie jest liniowy. Dopasowana do danych prosta oznaczona czerwoną przerywaną linią znacząco odbiega od krzywej oznaczonej kolorem zielonym. Stąd niezbędna jest inna specyfikacja równania regresji uwzględniająca nieliniową zależność między sprawdzianem a częścią matematyczno-przyrodniczą egzaminu gimnazjalnego.

Wykres 1. Przykład liniowej i nieliniowej zależności między wynikami egzaminów.



Pominięcie nieliniowości związku między wynikami egzaminów może prowadzić do tego, że wartość dodana nie będzie neutralna względem poziomu umiejętności uczniów. Należy jednak zauważyć, że standardowa metoda regresji zakłada, że zmienna zależna, w tym przypadku wynik egzaminu na wyższym poziomie, będzie opisana liniową kombinacją zmiennych niezależnych lub ich funkcji, którymi w tym przypadku są wyniki egzaminu na niższym poziomie i zmienne kontrolne. Tak więc zastosowanie tej metody nie wymaga liniowego związku między wynikami egzaminów.

Nieliniowość można uwzględnić wprowadzając do równania regresji wyrazy stanowiące przekształcenia wyników egzaminów. Często jako dodatkowy wyraz wprowadza się kwadrat zmiennej zależnej:

$$\hat{y}_i = \hat{\alpha} + \hat{\beta}_1 x_i + \hat{\beta}_2 x_i^2 \quad (6)$$

Związek między x a y można opisać dowolnym wielomianem x wprowadzając kolejne potęgi. O ile analizowana zależność jest monotoniczna, lub też powinna być, to często przyjmuje się, że opisuje ją funkcja wykładnicza:

$$\hat{y}_i = e^{\hat{\alpha} + \hat{\beta}_1 x_i} \quad (7a)$$

co odpowiada równaniu regresji liniowej, gdzie zmienną zależną jest logarytm naturalny y :

$$\ln(\hat{y}_i) = \hat{\alpha} + \hat{\beta}_1 x_i \quad (7b)$$

a wartości oczekiwane \hat{y}_i można uzyskać podstawiając $\ln(\hat{y}_i)$ jako potęgę liczby e .

Innym sposobem uwzględnienia nieliniowości związku między wynikami egzaminów jest przeprowadzenie regresji kawałkami (ang. „piecewise regression”). Metoda ta pozwala na osobne określenie postaci funkcyjnej równania regresji w różnych przedziałach zmiennej niezależnej. Można nie tylko zakładać, że w różnych przedziałach współczynniki w równaniu regresji mają inną wartość, ale i wprowadzać dodatkowe wyrazy ze zmiennymi niezależnymi. Dla przykładu, odnosząc się do kontekstu szacowania EWD, przypuśćmy, że egzamin z niższego poziomu w znacznie mniejszym stopniu różnicuje uczniów z niskimi osiągnięciami (ma niewielką wartość informacyjną dla tej grupy uczniów) w porównaniu z egzaminem na wyższym poziomie (o wysokiej wartości informacyjnej dla tej grupy). W takiej sytuacji dla uczniów o niskich wartościach x nie możemy przewidzieć wartości y lub też zależność między wynikami egzaminów jest dla tej grupy bardzo słaba. Jednak dla wyższych wyników zależność ta może być wyraźna. W tym przypadku celowa jest specyfikacja równania regresji zakładająca, że dla niższego przedziału wartości x równanie przyjmuje inną postać niż dla wyższego przedziału wartości x . Należy zatem oszacować regresję kawałkami, dla której w danym „kawałku” osobno szacujemy współczynniki równania regresji, które może też przyjmować różną postać w zależności od wartości x (np. w dolnym przedziale funkcja liniowa, a w górnym wielomian kwadratowy x).

Często, gdy znamy przyczynę zmiany charakteru zależności między zmiennymi, przyjmuje się, że w punkcie „załamania”, wyznaczającym granicę przedziałów x , mamy do czynienia ze skokiem wartości zmiennej zależnej y . Dla szacowania wartości dodanej należy jednak nałożyć wymóg, aby w punkcie „załamania” oszacowane w kawałkach równania regresji

przewidywały ten sam wynik egzaminu z wyższego poziomu⁶. Można to osiągnąć przez odpowiednią specyfikację równania regresji. Załóżmy, że związek między wynikami egzaminów można wyrazić prostą na płaszczyźnie, jednak dla $x < x^*$ ma ona inne nachylenie niż dla $x > x^*$. W takim przypadku oszacowane równanie regresji powinno mieć postać:

$$\hat{y}_i = \hat{\alpha} + \hat{\beta}_1 x_i + \hat{\beta}_2 (x_i - x^*) D_i \quad (8)$$

gdzie x^* oznacza punkt załamania dzielący dwa zbiory wartości x , D_i przyjmuje 0 dla $x_i < x^*$ oraz 1 w przeciwnym przypadku, a pozostałe zmienne zdefiniowano tak jak poprzednio. Nachylenie prostej w dolnym przedziale (dla $x_i < x^*$) określa współczynnik $\hat{\beta}_1$, a w górnym przedziale suma współczynników $\hat{\beta}_1 + \hat{\beta}_2$.

Do równania można też np. dodać wyraz $(x_i - x^*)^2 D_i$ zakładając, że w górnym przedziale zależność lepiej opisuje wielomian kwadratowy, podobnie dla dolnego przedziału dodając x_i^2 itp. W ten sposób dość dowolnie możemy modelować nieliniowość relacji między wynikami egzaminów, zmieniając postać funkcyjną w przedziałach, a także zmieniając wartości i liczbę punktów załamania. Wyboru punktów załamania można dokonać na podstawie wykresów rozrzutu wyników uczniów oraz analizując kształt dopasowanych krzywych nieparametrycznych (funkcja dostępna w większości pakietów statystycznych). Możliwe jest także zastosowanie bardziej zaawansowanych technik, tzw. regresji nieliniowej, celem oszacowania liczby i wartości punktów załamania. Wydaje się jednak, że w omawianym tu zastosowaniu regresja kawałkami powinna mieć jak najprostszą postać funkcyjną, a wybór przedziałów i punktów załamania powinien być dokonywany a priori w oparciu o znajomość różnic w charakterze egzaminów.

Podsumowując powyższą dyskusję w kontekście EWD trzeba dodać, że wybór specyfikacji postaci funkcyjnej równania regresji może być dokonany z uwagi na kryterium opisane w punkcie 1.5., a więc tak, aby przeciętna warunkowych różnic między wartością oczekiwaną a rzeczywistą dążyła do zera dla wszystkich grup uczniów wyznaczanych ze względu na wynik egzaminu na niższym poziomie. Można też analizować wpływ wyboru specyfikacji na uporządkowanie szkół względem EWD. Jeśli wpływ ten jest niewielki to należy wybrać taką

⁶ Często wielkość „skoku” przewidywanej wartości y w punkcie załamania jest celem analizy samym w sobie (np. gdy zmienną niezależną jest wiek, a zmienną zależną poziom spożycia używek wśród młodzieży pełno i niepełnoletniej). Jednak dla szacowania EWD kluczowym wymogiem jest to, aby krzywe wyników oczekiwanych oszacowane w regresji kawałkami były ciągłe. Inaczej, oczekiwania dla uczniów leżących przed i za punktem załamania mogłyby się za bardzo różnić.

specyfikację równania regresji, która jest najłatwiejsza w implementacji, a więc jest zrozumiała, prosta w interpretacji i stwarza niewielkie problemy obliczeniowe.

2.5. Wartość dodana jako różnica między wynikami egzaminów.

Przekładanie na jedną skalę i standaryzacja wyników egzaminów jest nie tylko zadaniem trudnym, ale i kontrowersyjnym. Powstało wiele metod, które można wykorzystać do tego celu, jednak większość z nich wymaga wieloletnich doświadczeń lub dodatkowych badań np. na losowej próbie uczniów, a przez to jest czasochłonna i kosztowna (por. Szaleniec, 2005). Wydaje się, że w przypadku Polski przekładanie egzaminów na jedną skalę celem szacowania wartości dodanej nie ma większego sensu, bowiem nie przynosi dodatkowych korzyści.

Za wartość dodaną ucznia przyjmuje się różnicę między wynikami egzaminów tylko w sytuacjach, gdy z założenia w systemie szkolnictwa wyniki egzaminów podawane są na jednej skali, o wspólnym punkcie odniesienia i tej samej jednostce pomiaru w kolejnych latach, tzw. skali rozwoju (ang. „developmental scale”). Jest to często związane z samą koncepcją egzaminu, którego celem jest doroczny, a nawet częstszy, pomiar podobnych cech u uczniów. W takim przypadku naturalne jest modelowanie bezpośrednio wzrostu wiedzy uczniów. Jednak o ile egzaminy różnią się charakterem, to użycie tej metody jest dyskusyjne, nawet po, w praktyce często bardzo trudnym, przełożeniu wyników na jedną skalę. W takim przypadku modelując wartość dodaną ucznia jako różnicę między egzaminami trudno oddzielić efekty wzrostu od artefaktów tworzonych przez zestawienie wyników testów o innych cechach psychometrycznych i odmiennej skali pomiaru.

Schemat postępowania w przypadku, gdy operujemy wynikami egzaminów mierzonymi na tej samej skali, można opisać następująco. Przyjmując, że y i x mierzone są na tej samej skali, definiujemy:

$$\Delta_i = y_i - x_i \quad (8)$$

Wtedy $d_i = \Delta_i$, a wartość dodaną szkoły można obliczyć wg wzoru (2).

Jeśli jednak chcielibyśmy uwzględnić zmienne kontrolne, to szacujemy równanie regresji, gdzie zmienną objaśnianą jest różnica między wynikami egzaminów, a pozostałe zmienne zdefiniowano jak w równaniu (3):

$$\hat{\Delta}_i = \hat{\alpha} + \mathbf{R}'_i \hat{\beta}_2 + \mathbf{P}'_j \hat{\beta}_3 \quad (9)$$

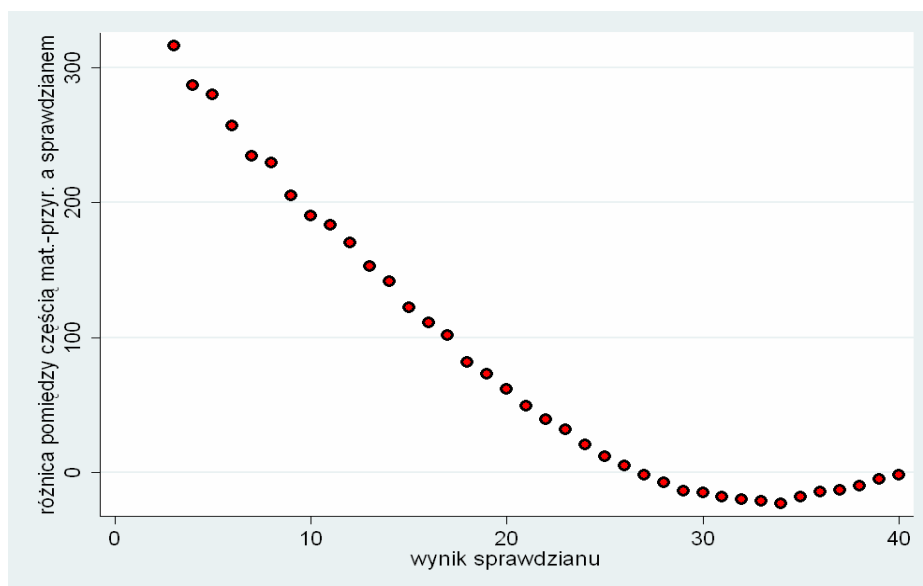
Wartość dodaną obliczamy wg wzoru (2) na podstawie różnic między rzeczywistą a oczekiwaną różnicą w wynikach egzaminów:

$$d_i = \Delta_i - \hat{\Delta}_i \quad (10)$$

Należy zauważyć, że równanie (9) jest równoważne równaniu (3a) przy założeniu, że $\beta_1 = 1$. Tak więc równanie (9) stanowi szczególną postać równania (3a), odpowiadający mu model jest więc mniej ogólny. Modele bezpośrednio wyjaśniające wzrost wiedzy, w rodzaju przedstawionego równaniem (9), opierają się na innych założeniach co do rozwoju ucznia, niż modele opisane równaniem (3a) i w większości przypadków prowadzą do uzyskania innych rezultatów (por. McCaffrey i in., 2005).

Zastosowanie tej metody do wyników egzaminów w Polsce, operujących różną skalą pomiaru, prowadzi do uzyskania oszacowań wartości dodanej, które nie są neutralne względem poziomu uczniów w szkołach. Wynika to z różnego rozkładu wyników sprawdzianu i egzaminu gimnazjalnego. Dla przykładu przekształciłem wyniki sprawdzianu z 2002 roku i części matematyczno-przyrodniczej egzaminu gimnazjalnego z 2005 roku w zmienne standaryzowane o średniej 500 i odchyleniu standardowym 100. Na wykresie poniżej przedstawiono średnie warunkowe różnic między tak zestandaryzowanymi wynikami egzaminów w odniesieniu do wyników sprawdzianu.

Wykres 2. Średnie warunkowe różnic między zestandaryzowanymi wynikami egzaminów.



Widać, że w tym przypadku zastosowanie różnic między wynikami egzaminów do szacowania wartości dodanej sztucznie podnosi wartość dodaną dla szkół, gdzie przeważają uczniowie, którzy uzyskali na sprawdzianie mniej niż 20 punktów. Tak więc ta prosta metoda, choć wydaje się intuicyjna, a przez wielu uznawana jest za naturalną i uzasadnioną, zdecydowanie nie jest odpowiednim sposobem szacowania wartości dodanej.

2.6. Modele dwuletnie a wieloletnie.

W przypadku szacowania EWD przez regresję wyniku danego ucznia z wcześniejszego egzaminu na wynik późniejszy, w rodzaju opisanej równaniami (3) lub (3a), zakłada się, że wcześniejszy wynik ucznia zawiera informacje na temat jego zdolności, przebiegu kształcenia, środowiska rodzinnego i innych zmiennych wpływających na poziom umiejętności, w tym nakładów szkolnych i jakości nauczania we wcześniejszych placówkach. Ponadto, ze względu na brak możliwości uwzględnienia charakterystyk niemierzalnych, takich jak zdolności ucznia i poziom jego motywacji, najczęściej przyjmuje się założenie, że w podobny sposób wpływają one na wyniki w kolejnych latach, a i ich efekty charakteryzuje takie samo tempo zaniku (por. Ladd, Walsh, 2002). Takie założenia są niezbędne, o ile dysponujemy jedynie dwoma wynikami egzaminów. Badania pokazują jednak, że są one nieprawdziwe, przynajmniej w przypadku możliwych do sprawdzenia zmiennych mierzalnych, i mogą silnie wpływać na wartość dodaną.

Założenia te można uchylić modelując wzrost wiedzy ucznia z wykorzystaniem wyników egzaminów z kilku lat. W ten sposób można uwzględnić wpływ nakładów edukacyjnych i nauczycieli w poszczególnych latach nauki, a także ich opóźnione efekty w kolejnych latach. Odpowiednio wyspecyfikowane modele wieloletnie dają bardziej precyzyjne i mniej obciążone oszacowania wartości dodanej. Są to jednak modele znacznie bardziej skomplikowane, wymagające zarówno pod względem zasobu danych, jak i metod obliczeniowych⁷.

Przewagę modeli wieloletnich nad dwuletnimi ukazuje prosty, ale nie tak daleki od polskich realiów przykład: założmy, że wszyscy uczniowie gimnazjum G1 uczęszczali do szkoły podstawowej SP1, a wszyscy uczniowie gimnazjum G2 uczęszczali do szkoły podstawowej

⁷ Por. McCaffrey i in., 2004, 2005. W tych pracach można znaleźć szczegółową analizę metod tego typu i bogate odniesienia do literatury

SP2. Przyjmijmy też, że grupa uczniów obydwu szkół ma identyczny rozkład istotnych dla wzrostu wiedzy cech indywidualnych i obydwie gimnazja charakteryzuje identyczna jakość nauczania, jednak nauczyciele SP1 wypracowali wyższą wartość dodaną, przez co ich uczniowie osiągnęli znacznie wyższe wyniki na sprawdzianie, niż uczniowie SP2. O ile efekty pracy nauczycieli z SP1 mają identyczny jak w przypadku sprawdzianu, ale opóźniony w czasie, wpływ na wynik egzaminu gimnazjalnego, to G1 i G2 uzyskają podobną wartość dodaną w modelu dwuletnim. O ile jednak efekty pracy nauczycieli ze szkoły podstawowej zanikają wraz z upływem czasu, to wartość dodana G1 szacowana poprzez model dwuletni będzie zaniżona. W skrajnym przypadku, gdy efekty pracy w szkole podstawowej nie mają żadnego wpływu na egzamin gimnazjalny, różnica między wartością dodaną G1 i G2 będzie zbliżona do wartości dodanej wypracowanej w SP1.

Ten prosty przykład pokazuje intuicję leżącą u podstaw rozwijania wieloletnich modeli wartości dodanej. Umożliwiają one uwzględnienie wpływu nauczycieli i nakładów z lat wcześniejszych, które jak pokazują badania mają duże znaczenie (por. Rivkin i in., 2005), oszacowanie stopnia deterioracji tego wpływu, a także modelowanie interakcji między zmiennymi w różnych latach. Trzeba jednak zaznaczyć, że modele wieloletnie wymagają testowania uczniów w kolejnych latach, a także zbierania danych indywidualnych i ich łączenia między latami. W przypadku Polski wprowadzenie modeli wieloletnich wymagałoby daleko idących zmian w systemie egzaminów zewnętrznych i systemie zbierania danych edukacyjnych.

3. Wiarygodność oszacowań wartości dodanej.

3.1. Mediana, średnia i inne miary tendencji centralnej.

W powyższych rozważaniach EWD liczona jest jako miara tendencji centralnej indywidualnych różnic między wynikami rzeczywistymi a oczekiwanymi uczniów. Najczęściej stosowaną miarą tego typu jest średnia, czasem wykorzystuje się też medianę. W przypadku prób o dużych liczebnościach i rozkładach zbliżonych do normalnego różnica między nimi dąży do zera. Jednak dla prób mniejszych, w sytuacji gdy często występują wartości odstające (ang. „outliers”), mediana okazuje się bardziej odporną statystyką (ang. „robust”). Mediana i średnia różnią się także, gdy mamy do czynienia z rozkładami skośnymi. Wybór jednej z tych statystyk zależy nie tylko od empirycznego charakteru analizowanych danych, ale i od założeń przyjętych co do tego, jakie cechy chcemy mierzyć je wykorzystując.

W kontekście EWD mediana będzie lepszą miarą, jeśli interesuje nas przede wszystkim poziom uczniów przeciętnych dla całej grupy (klasy lub szkoły). Średnia natomiast będzie brała pod uwagę zmiany w wynikach zarówno pojedynczych osób, jak i mniejszych grup, które to nie będą miały znaczenia dla poziomu mediany, o ile dotyczą uczniów o relatywnie niskich lub wysokich osiągnięciach w stosunku do całej grupy (szkoły lub klasy). Kluczowym pytaniem jest to, czy w przypadku większości szkół mamy do czynienia z występowaniem wartości skrajnych, czy też ze skośnością rozkładów, a także interpretacja przyczyn, dla których tak się dzieje. Wartości skrajne mogą tworzyć uczniowie, którzy niepoważnie potraktowali jeden z egzaminów lub też z przyczyn losowych osiągnęli wynik zupełnie nie odpowiadający ich rzeczywistym zdolnościom i wiedzy. Nie chodzi tu o losowy błąd pomiaru, który w tym samym stopniu dotyczy wszystkich uczniów, ale raczej o sytuacje, gdy uczeń w dramatyczny sposób zaniża swój wynik (np. ze względu na bardzo złe samopoczucie lub gdy z powodu braku chęci w ogóle nie wypełnia testu). W takich okolicznościach jego indywidualna różnica między rzeczywistym a oczekiwanym wynikiem może być bardzo duża, co dla mniejszych szkół może mieć kluczowe znaczenie dla ich wartości dodanej, o ile do jej szacowania wykorzystujemy średnią z indywidualnych różnic. Warto zauważyć, że indywidualne różnice o skrajnie dużych wartościach mogą być zarówno dodatnie, gdy uczeń nadspodziewanie źle napisał wcześniejszy egzamin, jak i ujemne, gdy nadspodziewanie źle napisał egzamin późniejszy.

Jeśli opisane powyżej sytuacje są dominującym problemem w precyzji szacowania wartości dodanej, to celowe jest wykorzystanie mediany lub innej z „odpornych” miar, o których za chwilę. Jednak możliwe jest także, że niektóre szkoły osiągają ponadprzeciętne rezultaty z całymi grupami uczniów słabszych lub lepszych. Jeśli wartość dodana szacowana jest przez medianę, to dla tych szkół pozostanie ona na tym samym poziomie, jak gdyby nie podejmowały one dodatkowego wysiłku. Te przykłady pokazują, że wybór pomiędzy średnią i medianą może być kluczowy i musi być uzasadniony celem stawianym przed metodami szacowania EWD. W badaniach używane są także inne miary, które łączą cechy mediany i średniej: średnia obcięta (ang. „trimmed-mean”), czyli średnia liczona z próby po wyłączeniu odpowiedniego procenta obserwacji skrajnych (np. dla poziomu 10% wyłącza się obserwacje z wartościami mniejszymi niż 1 decyl i większymi niż 9 decyl), czy też cała grupa znacznie bardziej skomplikowanych obliczeniowo tzw. „M-estimators” (por. Huber, 2003). Średnia obcięta stanowi niejako kompromis między medianą i średnią. Z jednej strony jest odporna na wartości skrajne, a z drugiej w większym stopniu niż mediana reaguje na zmiany w kształcie rozkładu. Jest też stosunkowo prosta obliczeniowo, co razem powoduje, że często wykorzystuje się ją w badaniach, gdzie ze względu na niewielkie próby i dużą liczbę obserwacji odstających średnia jest nieodpowiednią miarą tendencji centralnej. Warto rozważyć jej zastosowanie do szacowania EWD.

3.2. Ocena wiarygodności statystycznej wartości dodanej.

Wartość dodaną, obliczaną powyższymi metodami, można traktować jako statystykę z losowej próby uczniów określonej szkoły. Inaczej mówiąc, przyjmujemy, że uczniowie w danym roku stanowią losową próbkę jakości nauczania ich placówki. Takie podejście do wartości dodanej wydaje się uzasadnione, o ile w rzeczywistości interesują nas nie poszczególni uczniowie, lecz jakość pracy szkoły lub jej nauczycieli.

Konsekwencją takiego podejścia jest uznanie wartości dodanej za wartość losową, dla której niezbędne jest określenie precyzji oszacowania. Powszechnie stosowane jest podawanie przedziałów ufności, jakie z określonym prawdopodobieństwem zawierają prawdziwą wartość dodaną. W badaniach naukowych przyjmowanym poziomem jest 99%, 95% lub 90%. Oznacza to odpowiednio, że w 1%, 5%, 10% przypadków przedział ufności może nie zawierać prawdziwej wartości dodanej.

Przy standardowym podejściu obliczenie przedziałów ufności nie sprawia większego problemu i dokonane może być za pomocą wzorów zamieszczonych w każdym podręczniku statystyki. Precyzja oszacowania, a więc szerokość przedziałów ufności, zależy będzie nie tylko od przyjętych założeń (współczynnika ufności), ale przede wszystkim od liczby uczniów zdających egzamin i zróżnicowania ich wyników. Przy tym wyniki uczniów w kolejnych latach można traktować jako kolejną próbkę jakości pracy szkoły i włączając je do analizy zwiększać precyzję oszacowań. Także włączenie zmiennych kontrolnych zwiększających moc wyjaśniającą modelu powinno podnosić precyzję oszacowań. Dodatkowo zastosowanie odpornych miar tendencji centralnej w rodzaju średniej obciętej, zamiast średniej z całej próby, spowoduje wzrost precyzji oszacowań, o ile mamy do czynienia ze znaczną liczbą wartości odstających. Jeśli jednak mamy do czynienia z rozkładami bliskimi normalnemu, to średnia z całej próby będzie miarą najbardziej precyzyjną.

Trzeba zaznaczyć, że precyzja z jaką oszacowana zostanie wartość dodana może być niewielka. Analizując badania amerykańskie wielu autorów podkreśla, że nawet wykorzystanie złożonych zbiorów danych, z wynikami egzaminów z wielu lat i szerokim zestawem zmiennych kontrolnych, prowadzi do oszacowań, które nie pozwalają rozróżnić efektów nauczania większości szkół. Może to oznaczać, że np. dla $\frac{3}{4}$ szkół na podstawie uzyskanych rezultatów nie można stwierdzić, które z nich charakteryzuje wyższa lub niższa jakość nauczania. Tym bardziej, jeśli stosowane są „bezpieczne” kryteria statystyczne zmniejszające ryzyko popełnienia błędu (np. wysoki współczynnik ufności lub szacowanie wartości dodanej jako efektu losowego, a nie stałego, powodujące tzw. „shrinking”). Stąd, niektórzy autorzy powątpiewają w sens wykorzystania wartości dodanej do tworzenia rankingu szkół (por. Meyer, 1997), a inni wskazują, że główną z niej korzyścią jest możliwość określenia szkół lub nauczycieli, którzy zdecydowanie odstają od przeciętnej (por. McCaffrey i in., 2005).

Dla innych niż średnia miar tendencji centralnej, w przypadku, gdy empiryczne rozkłady różnią między przewidywanymi a rzeczywistymi wynikami uczniów będą dla wielu szkół odbiegać od normalnych, celowe może być szacowanie przedziałów ufności przez tzw. „bootstrapping”. Jest to metoda wykorzystująca znaczne możliwości obliczeniowe współczesnych komputerów polegająca na obliczaniu statystyk z próbek tworzonych przez wielokrotne losowanie ze zwracaniem z analizowanej próby. W ten sposób można oszacować rozkład statystyki, np. średniej obciętej, bez przyjmowania założeń o rozkładzie zmiennej, dla

której tę statystykę obliczamy. O ile rozkład ten odbiega od normalnego, to oszacowane tą metodą przedziały ufności dla statystyk innych niż średnia będą się różnić od uzyskanych standardowymi metodami. W szczególności, nie muszą być one symetryczne. Obecnie wiele pakietów statystycznych umożliwia szacowanie przedziałów ufności dowolnych statystyk przez „bootstrapping”.

Kwestia wyboru metody oceny wiarygodności statystycznej i jej prezentacji może być rozważana jedynie w odniesieniu do rzeczywistych danych oraz przyjętych w systemie rozwiązań. Przy tym kluczowym problemem jest nie tyle określenie przedziałów ufności, co znalezienie sposobu na ich odpowiednie przedstawienie rodzicom i szkołom. Trzeba jednak podkreślić, że pominięcie wiarygodności statystycznej wartości dodanej daje nieprawdziwy obraz jakości nauczania w szkołach i podważa sens wprowadzania takiej informacji do systemu. Faktem jest, że dwie szkoły o różnej wartości dodanej, których przedziały ufności zawierają wspólny odcinek, mogą mieć w rzeczywistości takie same osiągnięcia. Rozwijanie świadomości tych zagadnień oraz opracowanie metod ich odpowiedniej publikacji jest równie ważne jak doskonalenie metod szacowania i oceny precyzji wartości dodanej. Pomijanie tych aspektów może prowadzić do utraty zaufania wobec całego systemu egzaminów zewnętrznych.

4. Podsumowanie.

Wymienione w powyższym tekście metody szacowania edukacyjnej wartości dodanej stanowią propozycje, które należy przetestować w badaniach na rzeczywistych danych z egzaminów zewnętrznych w Polsce. Do kluczowych kwestii, które należy rozważyć w odniesieniu do danych empirycznych, należą tu:

- porównanie rezultatów różnych metod szacowania EWD
- odpowiednia specyfikacja równań regresji (dobór zmiennych i określenie zależności funkcyjnych),
- wybór statystyk do oceny tendencji centralnej jako miary EWD,
- ocena precyzji EWD,
- opracowanie sposobów prezentacji uzyskanych wyników wraz z informacją o ich statystycznej wiarygodności.

Zarysowane problemy mogą być rozstrzygnięte jedynie w oparciu o dyskusję nad wynikami badań empirycznych w szerokim gronie obejmujących zarówno ekspertów, jak i osoby pragnące wykorzystać uzyskane w ten sposób informacje dla rozwijania systemu oświaty w Polsce.

Bibliografia

- Akerhielm Karen. 1995. Does Class Size Matter? *Economics of Education Review*, XIV, 229-241.
- Bartmańska A. 2004. Wartość dodana wyniku kształcenia a interpretacja wyników egzaminu zewnętrznego. *Biuletyn badawczy 3/2004*, Centralna Komisja Egzaminacyjna.
- Gujarati Damodar. 2002. *Basic Econometrics*. McGraw-Hill.
- Hanushek Eric A. 2003. The Failure of Input-based Schooling Policies. *Economic Journal* 113, February, s. F64-F98.
- Hoxby Caroline Minter. 2000. Peer Effects in The Classroom: Learning From Gender and Race Variation. NBER Working Paper 7867.
- Huber Peter J. 2003. *Robust Statistics*. Wiley-IEEE.
- Jakubowski Maciej, Sakowski Paweł. 2005. Quasi-Experimental Estimates of Class Size Effect in Primary Schools in Poland. Wersja robocza dostępna na: www.wne.uw.edu.pl/mjakubowski w zakładce publikacje.
- Kutnick Peter, Sebba Judy, Blatchford Peter, Galton Maurice, Thorp Jo. 2005. The Effects of Pupil Grouping: Literature Review. Department for Education and Skills, Research Report RR688 (dostępne na: www.dfes.go.uk/research).
- Ladd H., Walsh R. 2002. Implementing value-added measures of school effectiveness: getting the incentives right. *Economics of Education Review* 21 (2002) 1–17
- Lee M. 2005. *Micro-Econometrics for Policy, Program and Treatment Effects*, Series: Advanced Texts in Econometrics, Oxford University Press
- Markman Jacob M., Hanushek Eric A., Kain John F., Rivkin Steven G. 2003. Does peer ability affect student achievement? *Journal of Applied Econometrics*, vol. 18(5), s. 527-544.
- McCaffrey D., Lockwood J., Koretz M., Hamilton L. 2005. *Evaluating Value-Added Models for Teacher Accountability*. Rand Corporation MG-158.
- McCaffrey D., Lockwood J., Koretz M., Louis Thomas A., Hamilton L. 2004. Models for Value-Added Modeling of Teacher Effects. *Journal of Educational and Behavioral Statistics*, pp. 67-101, Volume 29, Number 1, Spring 2004.
- Meyer Robert. 1997. Value-Added Indicators of School Performance: A Primer. *Economics of Education Review*, Vol. 16, No.3, s. 283-301.
- Rivkin S., Hanushek E., Kain J. 2005. Teachers, Schools, and Academic Achievement. *Econometrica*, March 2005.
- Szaleniec Henryk. 2005. Wykorzystanie probabilistycznych modeli zadania testowego do zrównywania wyników sprawdzianu 2003–2005 i budowania banku zadań. Materiały z XII Krajowej Konferencji Diagnostyki Edukacyjnej, dostępne na www.oke.krakow.pl.
- Wilkinson Ian A. G., Hattie John A., Parr Judy M., Townsend Michael A. R., Fung Irene, Ussher Charlotte, Thrupp Martin, Lauder Hugh, Robinson Tony. 2000. *Influence of Peer Effects on Learning Outcomes: A Review of the Literature*. Auckland UniServices Limited, The University of Auckland.