

Analiza danych z badań międzynarodowych w Stata z użyciem pakietu `repest`

Michał Sitek, IBE PIB, m.sitek@ibe.edu.pl

Spis treści

1	Wprowadzenie	2
1.1	Dlaczego warto używać pakietu <code>repest</code> ?	3
1.1.1	Ograniczenia <code>repest</code>	3
1.1.2	Obsługiwane badania	3
1.2	Instalacja pakietu	4
2	Opis składni i opcje <code>repest</code>	4
2.1	Składnia komendy <code>repest</code>	4
2.2	Opcja <code>estimate()</code>	4
2.2.1	Wbudowane komendy <code>repest</code>	4
2.2.2	Komendy Stata	5
2.3	Kluczowe opcje <code>repest</code>	5
2.4	Dobre praktyki	6
2.5	Ograniczenia i alternatywy	6
2.6	Pobieranie danych	7
2.7	Dodatkowe informacje o danych	7
2.7.1	Struktura danych IEA	7
2.7.2	Struktura danych PISA	8
2.8	Łączenie danych w Stata	8
2.8.1	Dołączanie dodatkowych informacji (<code>merge</code>)	8
2.9	Zrozumienie struktury danych	9
2.9.1	Wartości prawdopodobne (<i>plausible values</i>)	9
2.10	Wagi analityczne	10
3	Analiza danych PISA 2022	11
3.1	Pobieranie danych	11
3.2	Wczytywanie danych w Stata	11

3.3	Analizy PISA krok po kroku - przykłady	13
3.3.1	Obliczanie średniego wyniku dla Polski i Czech	13
3.3.2	Różnice w wynikach chłopców i dziewcząt z testem istotności różnicy . .	14
3.3.3	Różnice w percentylach	14
3.3.4	Porównywanie średnich wyników między krajami	16
3.3.5	Odsetki uczniów według poziomów umiejętności	19
3.3.6	Prosta regresja liniowa	21
4	Analiza Danych TIMSS	24
4.1	Struktura plików	24
4.2	Przygotowanie danych TIMSS	24
4.2.1	Pobieranie i łączenie danych	24
4.2.2	Przygotowanie wag dla repest (Kluczowy etap!)	26
4.3	Analizy TIMSS krok po kroku - przykłady	27
4.3.1	Różnice między płciami w Polsce	28
4.3.2	Poziomy umiejętności	30
5	Analiza danych ICCS	32
5.1	Wczytywanie i przygotowanie danych	32
5.2	Analizy ICCS krok po kroku - przykłady	32
5.2.1	Odsetki kobiet i mężczyzn wśród nauczycieli w Polsce	32
5.2.2	Różnice w postrzeganiu problemów w szkole w podziale na płeć	33

1 Wprowadzenie

Pakiet **repest** w Stata to moduł stworzony m.in. dla analityków OECD (*Arvisati, F. and F. Keslair (2014), REPEST: Stata module to run estimations with weighted replicate samples and plausible values, Statistical Software Components S457918, Boston College Department of Economics.*), wspomagający analizy z wykorzystaniem wartości prawdopodobnych (*plausible values*, PVs). Jego głównym celem jest ułatwienie analizy danych z międzynarodowych badań umiejętności, takich jak **PISA**, **TIMSS**, **PIRLS** czy **PIAAC**.

Badania te charakteryzują się złożonymi schematami doboru próby i zaawansowanymi technikami skalowania wyników. Poprawna analiza wymaga uwzględnienia w modelu **wag statystycznych, replikacyjnych** oraz **wartości prawdopodobnych (PVs)**, aby skorygować niepewność pomiarową i schemat badania. W praktyce polega to na wielokrotnym powtórzeniu analiz i uśrednieniu wyników. **repest** automatyzuje ten proces.

1.1 Dlaczego warto używać pakietu `repest`?

Zalety pakietu `repest`

- **Automatyzacja:** Upraszcza przetwarzanie złożonych danych, automatycznie uwzględniając metodologię badań OECD i IEA, taką jak wagi replikacyjne czy wartości prawdopodobne (PV). Eliminuje to konieczność samodzielnego uwzględniania wag i replikacji w analizach.
- **Wygoda:** Pakiet oferuje łatwo dostępne funkcje do analiz podstawowych statystyk (średnich, częstości, korelacji), testów statystycznych oraz tworzenia i eksportowania zestawień.
- **Elastyczność:** Umożliwia analizy danych z wielu międzynarodowych badań edukacyjnych.

1.1.1 Ograniczenia `repest`

Ważne ograniczenia

Mimo swoich zalet, `repest` ma pewne ograniczenia. Pakiet jest ułatwieniem w obsłudze złożonych danych, ale nie jest wszechstronnym narzędziem do wszystkich typów analiz. W przypadku **bardziej zaawansowanych analiz** (np. modeli wielopoziomowych) wygodniejsze będzie wykorzystanie innych rozwiązań – wspominamy o nich pod koniec poradnika.

1.1.2 Obsługiwane badania

Badanie (Organizacja)	Obsługiwane
TIMSS (IEA)	
PIRLS (IEA)	
ICILS (IEA)	
ICCS (IEA)	
PIAAC (OECD)	
PISA (OECD)	
TALIS (OECD)	
SSSES (OECD)	

1.2 Instalacja pakietu

Aby rozpocząć pracę, należy jednorazowo zainstalować pakiet **repest**. W oknie komend Stata wpisz:

```
ssc install repest, replace
```

2 Opis składni i opcje repest

W tej części przyjrzymy się dokładniej składni **repest** i opcjom tej komendy. Część z nich jest zilustrowana w przykładach w dalszych częściach poradnika.

2.1 Składnia komendy repest

Podstawowa składnia komendy **repest** jest następująca:

```
repest svyname [if warunek] [in zakres] , estimate(cmd [,cmd_options]) [options]
```

- **svyname**: Musi być jedną z obsługiwanych przez **repest** nazw badań (np. PISA, TIMSS, PIRLS).
 - **estimate(cmd)**: Kluczowa opcja, w której określa się rodzaj analizy.
-

2.2 Opcja estimate()

2.2.1 Wbudowane komendy repest

- **means**: Oblicza średnie.

```
repest PISA, estimate(means pv@math) by(cnt) over(gender) display
```

- **summarize**: Oblicza statystyki opisowe. Wymaga opcji **stats()**.

```
repest PISA, estimate(summarize escs, stats(p5 p25 median p75 p95))
```

- **corr**: Oblicza macierz korelacji.

```
repest PISA, estimate(corr pv@math pv@read) by(cnt)
```

- **quantiletable**: Tworzy tabele kwantylowe, podobne do tych w raportach PISA.

```
repest PISA, estimate(quantiletable escs pv@math, nquantiles(5))
```

2.2.2 Komendy Stata

Można użyć dowolnej standardowej komendy Stata, która akceptuje wagi (**pweights** lub **awweights**), poprzedzając ją **stata**:

- **Regresja liniowa (reg)**:

```
repest PISA, estimate(stata: reg pv@math escs i.gender)
```

- **Regresja logistyczna (logit)**:

```
repest PIAAC, estimate(stata: logit zmienna_binarna pvnum@ age)
```

2.3 Kluczowe opcje repest

- **by(varname [, by_options])**: Wykonuje analizę oddzielnie dla każdej kategorii zmiennej **varname** (np. kraju).
 - **levels(string)**: Ogranicza analizę do wymienionych grup (np. **levels(POL FRA DEU)**).
 - **average(string)**: Oblicza uśrednione wyniki dla zdefiniowanych grup krajów (np. **average(OECD EU)**).
- **over(varlist [, test])**: Oblicza statystyki w podgrupach w ramach każdego kraju lub ogółem.
- **outfile(filename, replace)**: Zapisuje wyniki do pliku Stata (**.dta**).
- **display**: Wyświetla wyniki w oknie konsoli.

- **store(name)**: Zapisuje wyniki regresji do pamięci w celu późniejszego eksportu np. przez `esttab`.
- **results(results_options)**: Pozwala kontrolować, które wyniki są wyświetlane.
 - `add(addlist)`: Dodaje skalary (np. `r2`, `N`).
 - `combine(name: expression)`: Oblicza nowe statystyki na podstawie wyników.
- **fast**: Przyspiesza obliczenia dla dużych zbiorów danych (przydatne we wstępnych analizach).
- **flag**: Zastępuje wyniki oparte na małej liczbie obserwacji specjalnym kodem braku danych.
- **coverage**: Oblicza odsetek obserwacji uwzględniony w analizie.
- **svyparams(svy_options)**: Pozwala nadpisać domyślne parametry badania (wymagane dla `svyname = SVY`).

2.4 Dobre praktyki

 Warto pamiętać o kilku rzeczach

- **Sprawdzenie danych**: Zawsze używaj komend `describe`, `summarize`, `tabulate`, `codebook`.
- **Dokumentacja skryptów**: Zawsze dokumentuj swoje skrypty w plikach `.do` lub `.qmd`, aby zapewnić powtarzalność analiz.
- **Weryfikacja wyników**: Zawsze porównuj uzyskane wyniki z oficjalnymi raportami międzynarodowymi.
- **Obsługa braków danych**: Pamiętaj, że `repest` domyślnie pomija obserwacje z brakami danych.

2.5 Ograniczenia i alternatywy

`Repest` świetnie się sprawdza do generowania zestawień statystyk opisowych z wielu krajów. Ma prostą i logiczną składnię i obsługuje wiele badań. Jest też wykorzystywany przez OECD i bieżąco utrzymywany przez analityków związanych z OECD.

- **pv w Stata**: Inny pakiet (`ssc install pv`). Działa wolniej od `repest`, ale jest bardziej elastyczny. Pozwala ręcznie definiować nazwy wag, co eliminuje potrzebę ich generowania dla TIMSS/PIRLS. W niektórych przypadkach lepiej współpracuje z innymi często wykorzystywanymi poleceniami (np. `estimates store`, `margins`, itp.).

- **mi i svyset**: Podejście dla ekspertów, dające pełną kontrolę, ale wymagające doskonałej znajomości Stata i metodologii badania. To bardziej zaawansowane rozwiązanie, którego główną zaletą jest możliwość wykorzystania szerszego zakresu poleceń Stata z zakresu analizy danych imputacyjnych.

2.6 Pobieranie danych

Dane z badań dostępne są na poniższych stronach:

- **PISA**: www.oecd.org/pisa/data
- **PIAAC**: <https://www.oecd.org/skills/piaac/data/>
- **TALIS**: <https://www.oecd.org/en/about/programmes/talis.html#data>
- **SSSES**: <https://www.oecd.org/en/about/programmes/oecd-survey-on-social-and-emotional-skills.html#data>
- **TIMSS, PIRLS, ICCS, ICILS**: <https://www.iea.nl/data-tools/repository>

W niektórych badaniach OECD przed pobraniem zbioru trzeba wypełnić krótką ankietę.

2.7 Dodatkowe informacje o danych

2.7.1 Struktura danych IEA

Zbiory IEA (TIMSS, PIRLS i ICILS) są zazwyczaj podzielone na dużą liczbę plików, pogrupowanych według kraju, poziomu klasy oraz zastosowanego narzędzia badawczego. Po ściągnięciu danych z badania TIMSS 2023 dla klasy 4 w folderze zobaczymy ponad 500 plików.

Nazwy plików

Przykład: Plik `asapolm8.sav` zawiera dane uczniów z Polski z badania TIMSS 2023 dla klasy 4 w formacie `.sav`

Gdzie: - **asa**: odpowiedzi na zadania i wyniki uczniów klasy 4 - **pol**: kod kraju (Polska)
- **m8**: cykl badania (TIMSS 2023)

Pierwsza litera w nazwie pliku wskazuje poziom klasy: - **a** – klasa 4 - **b** – klasa 8

Dalsze litery określają typ danych: - **asa/bsa** – wyniki uczniów oraz wartości prawdopodobne (PV) - **asp/bsp** – dane procesowe (np. czasy odpowiedzi) - **ash** – dane z kwestionariusza rodzica - **asg/bsg** – dane z kwestionariusza ucznia - **acg/bcg** – dane z kwestionariusza szkoły - **atg/btg** – dane z kwestionariusza nauczyciela

2.7.2 Struktura danych PISA

W przypadku badania PISA dane z różnych krajów są połączone w jeden zbiorczy plik (często bardzo duży), bez podziału na osobne pliki krajowe. Pliki są podzielone według cyklu badania oraz typu danych.

Nazwy plików PISA

Przykład: Plik `CY08MSP_STU_QQQ.sav` zawiera dane z kwestionariuszy uczniów z wszystkich krajów z badania realizowanego w roku 2022.

Gdzie: - `CY08` – cykl badania (tu: 2022) - `MSP` – Main Study (badanie główne) - `STU_QQQ` – dane uczniów (student)

Inne oznaczenia: - `SCH_QQQ` – kwestionariusze szkół - `TCH_QQQ` – kwestionariusze nauczycieli - `STU_COG` – wyniki testów kognitywnych uczniów (czytanie, matematyka, nauki przyrodnicze itp.) - `STU_FLT` – wyniki testów z edukacji finansowej - `STU_ICT` – kwestionariusz dotyczący technologii informacyjno-komunikacyjnych - `STU_WBQ` – kwestionariusz dobrostanu uczniów

2.8 Łączenie danych w Stata

Pakiet `repest` wymaga, aby wszystkie dane były w jednym, aktywnym zbiorze.

```
use "sciezka/do/pierwszego_kraju.dta", clear
append using "sciezka/do/kolejnego_kraju_A.dta"
append using "sciezka/do/kolejnego_kraju_B.dta"
```

2.8.1 Dołączanie dodatkowych informacji (`merge`)

Aby połączyć dane uczniów z danymi szkół (`m:1`) lub rodziców (`1:1`):

```
* Załaduj dane uczniów
use "sciezka/do/danych_uczniow.dta", clear

* Przyłącz dane szkół (wielu uczniów do jednej szkoły)
merge m:1 schoolid using "sciezka/do/danych_szkolnych.dta"
```

```
tab _merge
drop _merge

* Przyłącz dane rodziców (jeden uczeń do jednego rodzica)
merge 1:1 studid using "C:/sciezka/dane_rodzicow.dta"
tab _merge
drop _merge
```

2.9 Zrozumienie struktury danych

Aby `repest` działał poprawnie, zmienne w zbiorze danych muszą mieć określoną strukturę nazw, odpowiadającą konwencji w poszczególnych badaniach.

2.9.1 Wartości prawdopodobne (*plausible values*)

Wartości prawdopodobne to zestaw wielokrotnych imputacji niewidocznych cech (np. umiejętności), które pozwalają na uzyskanie nieobciążonych błędów pomiaru. Gdy `repest` napotka symbol @ w nazwie zmiennej (np. `pv@math`), zakłada, że dostępne są wielokrotne imputacje dla tej zmiennej (np. `pv1math`, ..., `pv10math`, w przypadku badania PISA 2022).

- **PISA:**
 - matematyka `pv1math`, `pv2math`, ..., `pv10math`;
 - rozumienie tekstu: `pv1read`, `pv2read`, ..., `pv10read`
 - rozumowanie w naukach przyrodniczych: `pv1scie`, `pv2scie`, ..., `pv10scie`
- **TIMSS**
 - matematyka: `BSMMAT01`, ..., `BSMMAT05` (wartości prawdopodobne wyników z matematyki; każda zmienna odpowiada jednej imputacji PV)
 - przyroda: `BSMSCI01`, ..., `BSMSCI05` (wartości prawdopodobne wyników z nauk przyrodniczych)
- **ICCS** wiedza i umiejętności obywatelskie
- **PIRLS**: `BSRRD01`, ..., `BSRRD05`
- **PIAAC:**
 - rozumienie tekstu: `PVLIT1`, ..., `PVLIT10`
 - umiejętności matematyczne: `PVNUM1`, ..., `PVNUM10`

– adaptatywne rozwiązywanie problemów (w II cyklu): PVAPS1, ..., PVAPS10

2.10 Wagi analityczne

Drugim ważnym rodzajem zmiennych są **wagi analityczne**. `repest` wymaga, by w aktywnym zbiorze danych były dostępne konkretne zmienne wag. Ich nazwy są automatycznie ustawiane przez `repest` po podaniu nazwy badania.

- **Wagi końcowe (Final weight)**: Np. `w_fstuwt` (dla uczniów PISA), `TOTWGTS` (dla ICCS/ICILS), `wGT` (dla PIRLS/TIMSS).
- **Wagi replikacyjne (Replicate weights)**: Np. `w_fstr1` do `w_fstr80` (PISA), `SRWGT1` do `SRWGT75` (ICCS/ICILS), `JR1` do `JR150` (PIRLS/TIMSS).

W przypadku niektórych badań, np. **TIMSS** i **PIRLS**, konieczne jest przeliczenie wag. Opisujemy, jak to zrobić w części praktycznej.

W niektórych badaniach rozróżnia się schematy badań zależnie od rodzaju badanych, np. w badaniu **TALIS** mamy opcje: `TALISSCH`, `TALISTCH`, `TALISEC_STAFF`, `TALISEC_LEADER`, którą wprowadzamy zależnie od tego, jaką grupę chcemy analizować.

Aby analizować dane nauczycieli lub szkół w badaniach ICCS/ICILS, należy odpowiednio ustawić nazwę badania w poleceniu `repest`:

- **Dane nauczycieli**: użyj `ICCS_T` lub `ICILS_T` jako `svynome` w poleceniu `repest`. Dzięki temu zostaną automatycznie wykorzystane odpowiednie wagi nauczycielskie.
 - **Dane szkół**: użyj `ICCS_C` lub `ICILS_C` jako `svynome`, aby zastosować wagi szkolne.
-

Trzecią ważną informacją jest zmienna opisująca kraj, zazwyczaj odwołująca się do kodów ISO.

i Wielkość liter w Stata

W Stata **wielkość liter ma znaczenie**. Nazwy zmiennych dla PV powinny mieć wielkość zgodną z definicją w `repest`. Można je zmieniać komendą `rename`, np.:

```
* Zmienia nazwy wszystkich zmiennych pv*math na wielkie litery
rename pv*math, upper

* Zmienia nazwy wszystkich zmiennych na małe litery
rename *, lower
```

3 Analiza danych PISA 2022

Zacznijmy od badania PISA.

3.1 Pobieranie danych

- Wejdź na stronę OECD z danymi PISA 2022: <https://www.oecd.org/pisa/data/2022database/>
- Po wypełnieniu krótkiej ankiety zobaczysz listę plików w formacie SAS i SPSS. Są to dane z różnych modułów badania. Nas interesują dane z ankiety uczniów, w których są też ich wyniki. Stata ma wbudowaną możliwość importowania plików SAS (ale bez udostępnianego pliku z etykietami), możliwe jest też importowanie plików SPSS (.sav) przy pomocy komendy `import spss`, ale komenda ta słabo sobie radzi z bardzo dużymi zbiorami danych i dłuższymi etykietami zmiennych i wartości.

Utwórz na swoim komputerze folder na dane, np. `C:/pisa_data/`.

Pobierz międzynarodową bazę danych. Jest to spakowany (zip) plik zawierający dane w formacie SPSS. Rozpakowany plik nazywa się `CY08MSP_STU_QQQ.SAV`. Upewnij się, że plik jest zapisany w folderze. Do wczytywania pliku użyjemy komendy `import spss` (dostępnej od wersji Stata 16, we wcześniejszych wersjach można było korzystać z pakietu `usespss`).

3.2 Wczytywanie danych w Stata

```
* Ustawiamy ścieżkę do naszego folderu roboczego
* Pamiętaj, aby podać tu własną ścieżkę!
cd "C:/pisa_data/"

* Wczytujemy zbiór danych PISA 2022 w formacie SPSS.
import spss using CY08MSP_STU_QQQ.SAV, clear
```

Plik jest duży: zawiera 1278 zmiennych dla 613 744 uczniów.

! Problem z importem dużych plików SPSS w Stata

Niestety, Stata słabo sobie radzi z importem bardzo dużych plików SPSS (.sav), nawet jeśli mamy wystarczająco dużo pamięci RAM. Próba zaimportowania pliku może zakończyć się niepowodzeniem w niektórych systemach. Można zaimportować dane z pliku SAS (*.sas7bdat), jednak **stracimy etykiety wartości** (np. zmienna **gender** będzie zawierała wartości 1, 2, ale bez opisu etykiet wartości). Wprawdzie dysponujemy także plikiem *.SAS, który zawiera definicje etykiet, ale Stata go nie obsługuje.

W takiej sytuacji możemy użyć programu **R (pakiet haven)**, który pozwala zaimportować plik i zapisać go do formatu .dta z pełnymi metadanymi.

Jeśli interesuje nas tylko wybrany podzbiór danych, możemy najpierw ograniczyć zbiór, korzystając z narzędzia **IDB Analyzer**, który obsługuje pliki w różnych formatach, ale nie w formacie Stata. Wykorzystanie **IDB Analyzer** może być pomocne w zarządzaniu plikami – np. gdy chcemy stworzyć zbiór z kilku krajów czy połączyć dane z różnego rodzaju narzędzi (np. ankiety ucznia i nauczyciela).

Po załadowaniu pliku sprawdzimy, czy mamy najważniejsze zmienne, z których korzysta **repest**. Dla uproszczenia zapytamy o pierwsze 3 PV i 5 wag replikacyjnych.

```
. describe PV1MATH-PV3MATH W_FSTUWT W_FSTURWT1- W_FSTURWT5 CNT
```

Variable name	Storage type	Display format	Value label	Variable label
PV1MATH	double	%10.0g		Plausible Value 1 in Mathematics
PV2MATH	double	%10.0g		Plausible Value 2 in Mathematics
PV3MATH	double	%10.0g		Plausible Value 3 in Mathematics
W_FSTUWT	double	%10.0g		FINAL TRIMMED NONRESPONSE ADJUSTED STUDENT W
W_FSTURWT1	double	%10.0g		FINAL TRIMMED NONRESPONSE ADJUSTED STUDENT RI
W_FSTURWT2	double	%10.0g		FINAL TRIMMED NONRESPONSE ADJUSTED STUDENT RI
W_FSTURWT3	double	%10.0g		FINAL TRIMMED NONRESPONSE ADJUSTED STUDENT RI
W_FSTURWT4	double	%10.0g		FINAL TRIMMED NONRESPONSE ADJUSTED STUDENT RI
W_FSTURWT5	double	%10.0g		FINAL TRIMMED NONRESPONSE ADJUSTED STUDENT RI
CNT	str3	%7s		Country code 3-character

3.3 Analizy PISA krok po kroku - przykłady

3.3.1 Obliczanie średniego wyniku dla Polski i Czech

Ograniczmy nasz zbiór tylko do danych dla Polski i Czech. W Stata możemy do tego wykorzystać komendę `keep if` (zachowaj tylko) lub komendę `drop if` (usuń, jeśli). W naszym przypadku chcemy zachować tylko dane, dla których akronim kraju (zgodny z ISO) to POL (Polska) lub CZE (Czechy).

```
. keep if CNT=="POL" | CNT=="CZE"  
(599,273 observations deleted)
```

! Uwaga na wielkość liter

W Stata wielkość liter w nazwach zmiennych ma znaczenie. Często spotykanym problemem w `repest` są niespójności nazw zmiennych. W przypadku danych PISA, `repest` oczekuje np., że kluczowe zmienne są zapisane małymi literami. Jeżeli `repest` nie znajdzie określonej zmiennej, np. `cnt`, pokaże komunikat zawierający taką informację. W naszym przypadku konieczna jest zmiana wielkości liter kluczowych zmiennych definiujących badanie. W analizowanym przypadku identyfikatora kraju, wag oraz identyfikatora szkoły (`CNTSCHID`).

```
rename CNT W_FSTUWT W_FSTURWT* CNTSCHID, lower
```

Teraz możemy przeprowadzić analizy. Na początek spójrzmy na kilka statystyk opisowych wyników z matematyki dla Polski.

```
. repest PISA if cnt == "POL", estimate(means PV@MATH)
```

Survey parameters have been changed to PISA2015

```
_pooled.....  
: _pooled
```

	Coefficient	Std. err.	z	P> z	[95% conf. interval]	
PV_MATH_m	488.9601	2.27485	214.94	0.000	484.5014	493.4187

Wyniki tej komendy pokazują oszacowaną średnią, jej błąd standardowy i 95% przedział ufności.

Zauważmy, że pojawia się również komentarz `Survey parameters have been changed to PISA2015`. Informuje on o tym, że analizy są prowadzone dla 10 pv (od PISA 2015 w badaniu używanych jest 10 pv – we wcześniejszych edycjach było 5).

3.3.2 Różnice w wynikach chłopców i dziewcząt z testem istotności różnicy

```
. repest PISA if cnt == "POL", estimate(means PV@MATH) over(ST004D01T, test) display
Survey parameters have been changed to PISA2015
over ST004D01T = 1 2
```

```
over var is ST004D01T
```

```
_pooled - ST004D01T = 1 .....
_pooled - ST004D01T = 2 .....
_pooled : _pooled
```

		Coefficient	Std. err.	z	P> z	[95% conf. interval]	
ST004D01T=1							
PV_MATH_m		486.1626	2.946798	164.98	0.000	480.387	491.9382
ST004D01T=2							
PV_MATH_m		491.6791	2.672776	183.96	0.000	486.4405	496.9176
ST004D01T=d							
PV_MATH_m		5.51646	3.326107	1.66	0.097	-1.002589	12.03551

3.3.3 Różnice w percentylach

W interpretacji wyników ważne są nie tylko średnie. Zawsze warto spojrzeć też na zróżnicowanie wyników. Przyjrzyjmy się wartościom wybranych percentyli w Polsce.

```
.
. repest PISA if cnt == "POL", estimate(summarize PV@MATH, stats(p5 p25 p50 p75 p95))
Survey parameters have been changed to PISA2015
```

```
_pooled.....
: _pooled
```

		Coefficient	Std. err.	z	P> z	[95% conf. interval]	
PV_MATH_p5		340.3515	3.870307	87.94	0.000	332.7658	347.9372
PV_MATH_p25		426.3688	3.188472	133.72	0.000	420.1195	432.6181
PV_MATH_p50		490.4081	2.941234	166.74	0.000	484.6434	496.1728
PV_MATH_p75		551.97	2.612002	211.32	0.000	546.8506	557.0894
PV_MATH_p95		634.7251	3.701408	171.48	0.000	627.4705	641.9797

```
. repest PISA if cnt == "POL", estimate(summarize PV@MATH, stats(p10 p25 p50 p75 p90)) over(
Survey parameters have been changed to PISA2015
over ST004D01T = 1 2
```

```
over var is ST004D01T
```

```
_pooled - ST004D01T = 1 .....
_pooled - ST004D01T = 2 .....
_pooled : _pooled
```

		Coefficient	Std. err.	z	P> z	[95% conf. interval]	
-----+							
ST004D01T=1							
PV_MATH_p10		374.9559	4.83756	77.51	0.000	365.4745	384.4373
PV_MATH_p25		428.5824	3.732784	114.82	0.000	421.2663	435.8985
PV_MATH_p50		487.2313	3.528813	138.07	0.000	480.315	494.1476
PV_MATH_p75		544.6233	3.83937	141.85	0.000	537.0983	552.1483
PV_MATH_p90		593.7016	4.100419	144.79	0.000	585.6649	601.7383
-----+							
ST004D01T=2							
PV_MATH_p10		365.8842	4.476712	81.73	0.000	357.11	374.6584
PV_MATH_p25		423.8901	4.072185	104.09	0.000	415.9088	431.8714
PV_MATH_p50		494.1357	3.850228	128.34	0.000	486.5894	501.682
PV_MATH_p75		559.0357	3.361573	166.30	0.000	552.4471	565.6243
PV_MATH_p90		613.7861	4.009809	153.07	0.000	605.927	621.6452

ST004D01T=d							
PV_MATH_p10		-9.0717	6.414851	-1.41	0.157	-21.64458	3.501176
PV_MATH_p25		-4.6923	4.729754	-0.99	0.321	-13.96245	4.577848
PV_MATH_p50		6.9044	4.563641	1.51	0.130	-2.040172	15.84897
PV_MATH_p75		14.4124	4.640495	3.11	0.002	5.317198	23.5076
PV_MATH_p90		20.0845	5.090743	3.95	0.000	10.10683	30.06217

W pierwszej tabeli widzimy średni wynik z matematyki (PV_MATH_m) osobno dla chłopców (ST004D01T=1: 486.16) i dziewcząt (ST004D01T=2: 491.68). Wiersz ST004D01T=d prezentuje **różnicę** między wynikami dziewcząt a chłopców (5.52 punktu). $P > |z|$ to p-wartość, która wskazuje na istotność statystyczną. W tym przypadku p-wartość (0.097) jest powyżej przyjętego progu 0.05, co oznacza, że różnica w średnich wynikach **nie jest statystycznie istotna**.

W drugiej tabeli, dla każdej grupy (chłopców i dziewcząt) obliczono i wyświetlono percentyle (np. PV_MATH_p10 dla 10. percentyla, PV_MATH_p50 dla mediany itd.). W wierszach ST004D01T=d ponownie znajdziemy różnice w percentylach między dziewczętami a chłopcami. Warto zwrócić uwagę na p-wartości: podczas gdy różnice na niższych percentylach (np. p10, p25) nie są istotne statystycznie, na wyższych percentylach (p75 i p90) różnice stają się **istotne statystycznie** (p-value odpowiednio 0.002 i 0.000). Oznacza to, że **repest** potrafi wykryć znaczące różnice między grupami na różnych poziomach rozkładu wyników.

Świetnie! Kontynuujmy formatowanie i poprawę czytelności. Oto kolejny fragment, z zastosowaniem wyróżnień i ustrukturyzowania treści.

3.3.4 Porównywanie średnich wyników między krajami

Poniżej przedstawiamy dodatkowe funkcje pakietu **repest**, które umożliwiają elastyczne zarządzanie wynikami analizy – ich wyświetlanie, zapisywanie i dalsze wykorzystanie:

- **display**: Domyślnie **repest** pokazuje wyniki bezpośrednio w oknie wyników Stata. Gdy używasz opcji **by()**, dla każdego poziomu (**by-level**) generowana jest osobna sekcja wyników.

- `outfile("nazwa_pliku.dta", [subopcje])`: Zapisuje wyniki do pliku `.dta`, co ułatwia ich dalsze przetwarzanie, zwłaszcza przy analizach porównawczych (np. między krajami). Możliwe jest dodanie subopcji, takich jak `pvalue` (p-wartości) czy `se` (błędy standardowe), aby uwzględnić więcej szczegółów statystycznych. Plik zawiera dane zbiorcze dla wszystkich poziomów określonych w `by()` lub `over()`.
- `store(nazwa_obiektu)`: Zapisuje wyniki jako obiekty w Stata, które można później przywołać. Dla każdego poziomu w `by()` tworzony jest osobny obiekt.

Aby porównywać wyniki między krajami, używamy opcji `by`.

```
* Analiza dla Polski i Czech z podziałem na kraje

repest PISA, estimate(means PV@MATH) by(cnt, levels(POL CZE))
```

Survey parameters have been changed to PISA2015

POL.....

cnt : POL

	Coefficient	Std. err.	z	P> z	[95% conf. interval]	
PV_MATH_m	488.9601	2.27485	214.94	0.000	484.5014	493.4187

CZE.....

cnt : CZE

	Coefficient	Std. err.	z	P> z	[95% conf. interval]	
PV_MATH_m	486.9992	2.093645	232.61	0.000	482.8957	491.1027

Czy średnia w Polsce była w 2022 r. wyższa niż w Czechach?

```
* Analiza dla Polski i Czech z podziałem na kraje

. repest PISA, estimate(means PV@MATH) over(cnt, test)
Survey parameters have been changed to PISA2015
option over() only allows numeric variables
```

```
over var is cnt
__00000U not found
r(111);
```

! Błąd option over() only allows numeric variables

W powyższej komendzie `repest` zakomunikował błąd. Zmienna użyta w `over` nie może być zmienną zawierającą znaki (`string`). Należy wtedy użyć drugiego identyfikatora dostępnego w zbiorze, zmiennej `CNTRYID`, czyli numerycznego kodu ISO krajów (Polska to 616, a Czechy to 203).

```
. repest PISA, estimate(means PV@MATH) over(CNTRYID , test)
Survey parameters have been changed to PISA2015
over CNTRYID = 203 616
```

```
over var is CNTRYID
```

```
_pooled - CNTRYID = 203 .....
_pooled - CNTRYID = 616 .....
_pooled : _pooled
```

		Coefficient	Std. err.	z	P> z	[95% conf. interval]	
=203							
	PV_MATH_m	486.9992	2.093645	232.61	0.000	482.8957	491.1027
=616							
	PV_MATH_m	488.9601	2.27485	214.94	0.000	484.5014	493.4187
=d							
	PV_MATH_m	1.960846	2.944474	0.67	0.505	-3.810218	7.731909

Średnie wyniki uczniów z Polski i Czech **nie różnią się istotnie statystycznie**. Przedziały ufności (Polska: 483–491, Czechy: 484–493) się pokrywają, a test różnicy (ostatni wiersz) pokazuje, że różnica 2 punktów nie jest istotna ($p = 0.505$). Ponieważ dane pochodzą z próby, uwzględniamy niepewność estymacji – tu brak podstaw, by uznać, że w populacji istnieje istotna różnica między krajami.

Różnice między `by` a `over`

Warto zwrócić także uwagę na różnice między `by` i `over`.

- Prefiks `by` wykonuje analizę osobno dla każdej grupy, traktując je jako niezależne zbiory danych. Nie umożliwia bezpośredniego testowania różnic między grupami – trzeba to robić oddzielnie.
- Natomiast opcja `over(*zmienna*, test)` w `repest` analizuje wszystkie grupy jednocześnie w ramach jednego modelu, automatycznie obliczając różnice między nimi i zapewniając, że błędy standardowe są poprawne.

3.3.5 Odsetki uczniów według poziomów umiejętności

W raportach z badania PISA często przedstawia się dane w postaci poziomów umiejętności – zazwyczaj w postaci odsetków uczniów osiągających kolejne progi, zdefiniowane w punktach PISA. W zbiorze PISA nie ma odpowiedniej zmiennej – musimy ją stworzyć.

O ile wcześniejsze analizy można było w zasadzie wykonywać, kopiując pojedyncze linie do okna poleceń i klikając Enter, to w praktyce wygodniej jest korzystać z **plików poleceń** (`do-file`). Jest to plik tekstowy (`.do`) zawierający sekwencję komend, działający jak skrypt. Jest to niezbędne narzędzie do zapewnienia powtarzalności, transparentności i efektywnej organizacji każdej analizy statystycznej. Umożliwia łatwe uruchamianie kodu, wprowadzanie zmian oraz dokumentowanie wszystkich kroków pracy z danymi.

Aby wyliczyć proporcje uczniów, musimy przekształcić PV (dla matematyki są to zmienne PV1MATH do PV10MATH) na kategoryjną zmienną odpowiadającą poziomom biegłości. Poniżej przekodowano wartości do poziomów zdefiniowanych w raporcie międzynarodowym PISA, gdzie wyróżniono poziomy 1a, 1b, 1c, poziom 2, 3, aż do poziomu 6, co łącznie daje 9 kategorii.

```
* Zdefiniuj etykiety wartości dla nowej zmiennej
label define math_level_labels 0 "Poniżej 1c" ///
                                1 "Poziom 1c" ///
                                2 "Poziom 1b" ///
                                3 "Poziom 1a" ///
                                4 "Poziom 2"  ///
                                5 "Poziom 3"  ///
                                6 "Poziom 4"  ///
                                7 "Poziom 5"  ///
```

8 "Poziom 6"

```
foreach i of numlist 1/10 {

* Zrekoduj pv dla matematyki na poziomy biegłości
recode PV`i'MATH    (min/233.169 = 0 "Poniżej 1c") ///
                    (233.17/295.469 = 1 "Poziom 1c") ///
                    (295.47/357.769 = 2 "Poziom 1b") ///
                    (357.77/420.069 = 3 "Poziom 1a") ///
                    (420.07/482.379 = 4 "Poziom 2") ///
                    (482.38/544.679 = 5 "Poziom 3") ///
                    (544.68/606.989 = 6 "Poziom 4") ///
                    (606.99/669.299 = 7 "Poziom 5") ///
                    (669.30/max = 8 "Poziom 6"), gen(math`i'level)

* Zastosuj etykiety wartości do nowej zmiennej

label values math`i'level math_level_labels

}
```

i Wyjaśnienie kodu do rekodowania PV na poziomy umiejętności

- **label define math_level_labels ...**: Definiuje opisowe etykiety (np. "Poziom 1c", "Poziom 6") dla wartości liczbowych, które będziemy przypisywać poziomom. Dzięki temu w wynikach zobaczymy czytelne nazwy, a nie tylko liczby.
- **foreach i of numlist 1/10 { ... }**: To pętla, która wykonuje ten sam zestaw komend dziesięć razy (dla każdej z dziesięciu PV, czyli PV1MATH, PV2MATH, ..., PV10MATH). Pozwala to uniknąć powtarzania kodu.
- **recode PViMATH (min/233.169 = 0 ...) ..., gen(mathilevel)**: Komenda **recode** przekształca każdy surowy wynik PISA (PV1MATH, PV2MATH itd.) na nową zmienną (**math1level**, **math2level** itd.). Odbywa się to poprzez przypisanie liczbowej wartości (0 do 8) na podstawie zdefiniowanych przedziałów punktowych (np. wynik od 233.17 do 295.469 staje się wartością 1). Opcja **gen()** tworzy nową zmienną, zachowując oryginalne dane.
- **label values mathilevel math_level_labels**: Na koniec, do nowo utworzonej zmiennej (**math1level**, **math2level** itd.) przypisywane są zdefiniowane etykiety, dzięki czemu wartości liczbowe (0, 1, 2...) są wyświetlane jako zrozumiałe opisy poziomów umiejętności (np. "Poniżej 1c", "Poziom 1c").

Powyżej zrekodowano wartości PV do wszystkich poziomów, ale możemy też uprościć

rekodowanie, np. wyróżnić kategorię uczniów, którzy są poniżej 2 poziomu (wartość 1) i na poziomie 2 i wyżej (wartość 0).

Mając te zmienne, możemy wyliczyć odsetki uczniów na poszczególnych poziomach. Wyliczmy to dla Polski.

```
. repeat PISA if cnt=="POL" , estimate(freq math@level)
Survey parameters have been changed to PISA2015

_pooled.....
: _pooled
```

	Coefficient	Std. err.	z	P> z	[95% conf. interval]	
math_level_0	.0740546	.0629373	1.18	0.239	-.0493001	.1974094
math_level_1	1.129739	.2436488	4.64	0.000	.6521962	1.607282
math_level_2	6.367772	.5078255	12.54	0.000	5.372453	7.363092
math_level_3	15.4304	.7576769	20.37	0.000	13.94538	16.91542
math_level_4	23.75994	.9219499	25.77	0.000	21.95295	25.56692
math_level_5	25.58255	.8678128	29.48	0.000	23.88167	27.28344
math_level_6	18.24058	.6846419	26.64	0.000	16.89871	19.58245
math_level_7	7.499555	.4674799	16.04	0.000	6.583311	8.415799
math_level_8	1.915414	.2991674	6.40	0.000	1.329057	2.501771

3.3.6 Prosta regresja liniowa

W tym przykładzie analizujemy wpływ statusu społeczno-ekonomicznego (ESCS) na wyniki z matematyki (PV0MATH).

Opcja **results(add(r2 N))**: Ta opcja pozwala na dodanie do tabeli wyników dodatkowych statystyk, które są domyślnie przechowywane przez Stata po estymacji. W tym przypadku dodajemy **R-kwadrat** (e_r2) – miarę dopasowania modelu, oraz **liczbę obserwacji** (e_N).

Analizy przeprowadzimy osobno dla Polski i dla Czech.

```
* Model regresji liniowej

* Zwykły, liniowy efekt
```

```
. repeat PISA, estimate(stata: reg PV@MATH ESCS) by(cnt) results(add(r2 N))
Survey parameters have been changed to PISA2015
`"CZE"' `"'POL"'
```

CZE.....

cnt : CZE

		Coefficient	Std. err.	z	P> z	[95% conf. interval]	
	ESCS	50.58072	1.800561	28.09	0.000	47.05168	54.10975
	_cons	492.6841	1.810705	272.10	0.000	489.1352	496.2331
	e_r2	.2204457	.0121808	18.10	0.000	.1965719	.2443196
	e_N	8301	.	.	.	.	.

POL.....

cnt : POL

		Coefficient	Std. err.	z	P> z	[95% conf. interval]	
	ESCS	40.34929	1.857624	21.72	0.000	36.70842	43.99017
	_cons	494.8693	1.851499	267.28	0.000	491.2404	498.4981
	e_r2	.1625078	.0129805	12.52	0.000	.1370666	.187949
	e_N	5875	.	.	.	.	.

* Nieliniowy efekt ESCS

```
. repeat PISA, estimate(stata: reg PV@MATH c.ESCS##c.ESCS) by(cnt) results(add(r2 N))
Survey parameters have been changed to PISA2015
`"CZE"' `"'POL"'
```

CZE.....

cnt : CZE

		Coefficient	Std. err.	z	P> z	[95% conf. interval]	
	ESCS	49.8014	1.80252	27.63	0.000	46.26852	53.33427
c_ESCS	c_ESCS	-5.542008	1.618431	-3.42	0.001	-8.714074	-2.369941
	_cons	496.8342	2.284468	217.48	0.000	492.3567	501.3117
	e_r2	.2233969	.0118917	18.79	0.000	.2000896	.2467041
	e_N	8301	.	.	.	.	.

POL.....

cnt : POL

		Coefficient	Std. err.	z	P> z	[95% conf. interval]	
	ESCS	40.10634	1.979861	20.26	0.000	36.22588	43.98679
c_ESCS_c	ESCS	-1.108514	2.206676	-0.50	0.615	-5.433519	3.216491
	_cons	495.7264	2.62605	188.77	0.000	490.5794	500.8733
	e_r2	.1627424	.0128378	12.68	0.000	.1375808	.187904
	e_N	5875	.	.	.	.	.

i Interpretacja wyników regresji liniowej

Wyniki pierwszej regresji pokazują **liniowy związek** między ESCS a wynikami z matematyki w obu krajach.

- **Dla Czech:** Wzrost ESCS o jedną jednostkę wiąże się ze wzrostem wyniku z matematyki o około **50.58 punktu**. Model ten wyjaśnia około **22%** zmienności wyników ($e_r2 = 0.22$).
- **Dla Polski:** Ten sam wzrost ESCS przekłada się na wzrost wyniku o około **40.35 punktu**, a model wyjaśnia około **16%** zmienności ($e_r2 = 0.16$).

W obu przypadkach efekt ESCS jest statystycznie istotny (p-wartości bliskie 0.000). Widzimy, że siła związku między statusem społeczno-ekonomicznym a wynikami edukacyjnymi jest wyższa w Czechach niż w Polsce.

i Nieliniowy efekt ESCS (interakcja zmiennej ze sobą)

W Stata, gdy chcemy modelować **efekty nieliniowe** dla zmiennych ciągłych (takich jak ESCS), możemy dodać do modelu potęgę tej zmiennej. Najczęściej używa się kwadratu zmiennej. Stata ułatwia to za pomocą **notacji zmiennych czynnikowych** (**factor variables**), gdzie **c.** oznaczamy zmienne ciągłe, a **i.** zmienne kategoryjne (np. płeć). Fragment **c.zmienna##c.zmienna** tworzy interakcję zmiennej ciągłej ze sobą, co jest równoznaczne z dodaniem do modelu zmiennej kwadratowej (**zmienna²**). Stata automatycznie utworzy zarówno główny efekt zmiennej (ESCS), jak i jej kwadrat (**c.ESCS#c.ESCS** w tabeli, choć często wyświetlane jako **c.ESCS_c_ESCS**). Jest to wygodny sposób na sprawdzenie, czy związek jest ściśle liniowy.

W naszym przykładzie współczynnik kwadratowy jest **statystycznie istotny w Czechach** (p-wartość = 0.001), co wskazuje, że związek między ESCS a wynikami z

matematyki jest nieliniowy. Ujemny współczynnik kwadratowy sugeruje, że pozytywny wpływ ESCS na wyniki z matematyki jest silniejszy na niższych poziomach ESCS i stopniowo słabnie na wyższych poziomach statusu. R-kwadrat nieznacznie wzrósł do 0.2234, co sugeruje, że model z efektem nieliniowym jest lepiej dopasowany do danych.

4 Analiza Danych TIMSS

Przejdziemy teraz do badania TIMSS. Główna różnica, w porównaniu z badaniem PISA, wiąże się z koniecznością **samodzielnego przeliczenia wag replikacyjnych**.

4.1 Struktura plików

Przypomnijmy, że w badaniach IEA (TIMSS, PIRLS) zbiory są często podzielone na wiele plików. Typowa struktura nazwy pliku wygląda następująco: **asa pol m8 .sav**

Gdzie:

- **asa**: Typ danych, np. **asa** - dane uczniów Klasy 4 (dla Klasy 8: **bsa**).
- **pol**: Kod kraju (Polska).
- **m8**: Cykl badania (np. TIMSS 2019).

4.2 Przygotowanie danych TIMSS

4.2.1 Pobieranie i łączenie danych

Dane TIMSS są dostępne w repozytorium na stronie IEA: <https://www.iea.nl/data-tools/repository>. Po wejściu do zakładki TIMSS i zaakceptowaniu warunków korzystania możemy pobrać dane. Badanie TIMSS jest prowadzone w klasach 4 i 8. Nas interesuje klasa 4. Dane możemy pobrać w 3 formatach: SAS, SPSS i R. Wybieramy format **SPSS**.

Po pobraniu danych i skopiowaniu ich do naszego roboczego folderu możemy zacząć pracę. Pobrany plik zawiera nie tylko liczne pliki z danymi, ale także dokumentację umieszczoną w poszczególnych katalogach, co jest dużym ułatwieniem.

Spójrzmy, jak wygląda ten katalog z okna Stata:

```

cd C:\timss_data\TIMSS2023_IDB_SPSS_G4
C:\timss_data\TIMSS2023_IDB_SPSS_G4

. dir
<dir> 6/14/25 13:48 .
<dir> 6/14/25 13:48 ..
<dir> 6/14/25 13:48 1_User Guide
<dir> 6/14/25 13:48 2_Data Files
<dir> 6/14/25 13:48 3_Supplemental Material
<dir> 6/14/25 13:48 4_Reports and Encyclopedia
0.1k 6/14/25 13:48 timss2023.org.url

```

Nas interesują dane dla Polski i Czech. Znajdziemy je w folderze “2_Data Files”, w którym jest podfolder “SPSS Data”.

```

. cd "C:\timss_data\TIMSS2023_IDB_SPSS_G4\2_Data Files\SPSS Data\"
C:\timss_data\TIMSS2023_IDB_SPSS_G4\2_Data Files\SPSS Data

. dir *pol*
67.3k 6/14/25 13:48 acgpolm8.sav
8426.9k 6/14/25 13:48 asapolm8.sav
5005.8k 6/14/25 13:48 asgpolm8.sav
1049.6k 6/14/25 13:48 ashpolm8.sav
12.0M 6/14/25 13:48 asppolm8.sav
3152.6k 6/14/25 13:48 asrpolm8.sav
7076.1k 6/14/25 13:48 astpolm8.sav
283.6k 6/14/25 13:48 atgpolm8.sav

. dir *cze*
77.3k 6/14/25 13:48 acgczem8.sav
11.9M 6/14/25 13:48 asaczem8.sav
7284.3k 6/14/25 13:48 asgczem8.sav
1464.7k 6/14/25 13:48 ashczem8.sav
17.5M 6/14/25 13:48 aspczem8.sav
4020.3k 6/14/25 13:48 asrczem8.sav
6659.8k 6/14/25 13:48 astczem8.sav
252.3k 6/14/25 13:48 atgczem8.sav

```

Interesują nas zbiory zawierające **asa**, które wczytamy i zapiszemy w formacie Stata. W tym przykładzie używamy tylko dwóch krajów, ale przy pracy z wieloma krajami trzeba połączyć wiele plików. W takich sytuacjach pętle znacznie ułatwiają pracę.

```

local countries "pol cze"

foreach cnt of local countries {
  usespss asa`cnt`m8.sav
  save    asa`cnt`m8.dta, replace
}

```

i Zmienne lokalne w Stata

local w Stata to sposób na przechowywanie tekstu lub wartości pod określoną nazwą, którą można później wielokrotnie używać w kodzie. **local countries "pol cze"** tworzy zmienną lokalną o nazwie **countries**, która zawiera tekst "pol cze". W pętli możemy odwołać się do kolejnych elementów tej zmiennej. Dzięki temu wartość nowej zmiennej lokalnej **cnt** przyjmuje za pierwszym razem wartość **pol**, a za drugim **cze**.

Dane możemy połączyć, używając komendy **append**. Dokleja ona kolejne obserwacje z kolejnego pliku. Gdy mamy większą liczbę krajów, tu również możemy wykorzystać pętlę, która po kolei dodaje do otwartego zbioru kolejne dane.

W naszym przypadku wystarczy:

```

* Wczytaj dane uczniów z Polski (asapolm8.dta)
use      asapolm8.dta
append using asaczem8.dta // Zmiana z .sav na .dta, gdyż wcześniej je zapisaliśmy

save timss_pol_cze_2023.dta, replace

```

4.2.2 Przygotowanie wag dla **repest** (Kluczowy etap!)

Niedogodnością pracy z danymi PIRLS i TIMSS (oraz innych badań wykorzystujących schemat ważenia **jackknife**) jest konieczność przeliczenia wag. Przydatny do tego będzie omówiony wcześniej skrypt poleceń (**do-file**). Poniższy kod należy przekleić do skryptu **do-file**. Po wczytaniu zbioru danych wystarczy uruchomić **tylko raz** po załadowaniu danych TIMSS (lub PIRLS), aby wygenerować wagi replikacyjne, których **repest** potrzebuje.

```

* repest dla TIMSS/PIRLS oczekuje zmiennej wagowej o nazwie WGT
gen WGT = TOTWGT

* Pętla tworząca 150 wag replikacyjnych (JR1 do JR150)
* na podstawie 75 stref losowania Jackknife (JKZONE)

```

```

forval i = 1/75 {
    local j = 75 + `i'

    * Generuj pierwszą wagę replikacyjną dla danej strefy
    gen JR`i' = WGT
    replace JR`i' = 2*WGT if JKZONE == `i' & JKREP == 1
    replace JR`i' = 0 if JKZONE == `i' & JKREP == 0

    * Generuj drugą wagę replikacyjną dla tej samej strefy
    gen JR`j' = WGT
    replace JR`j' = 2*WGT if JKZONE == `i' & JKREP == 0
    replace JR`j' = 0 if JKZONE == `i' & JKREP == 1
}

```

Ten kod tworzy 150 wag replikacyjnych (JR1 do JR150) metodą **Jackknife Repeated Replication (JRR)**. Symuluje ona wielokrotne losowanie próby, co pozwala na precyzyjną estymację wariancji, biorąc pod uwagę złożony schemat losowania w badaniach IEA.

4.3 Analizy TIMSS krok po kroku - przykłady

Podstawowe polecenia są bardzo podobne do tych, które pokazywaliśmy dla analiz danych PISA. Różnice sprowadzają się do różnic w nazwach zmiennych.

Policzmy najpierw średnie wyniki. W aktywnym zbiorze danych mamy dane dla dwóch krajów. Ale podobnie będzie to wyglądać, gdy będziemy mieli więcej krajów w zbiorze.

```

. repeat TIMSS , estimate(means ASMMATO@ ) by(CTY)
`"CZE" ' ` "POL" '

CZE.....
CTY : CZE

```

	Coefficient	Std. err.	z	P> z	[95% conf. interval]	
ASMMATO__m	530.4311	2.046724	259.16	0.000	526.4196	534.4426

```

POL.....
CTY : POL

```

	Coefficient	Std. err.	z	P> z	[95% conf. interval]	
ASMMATO__m	546.0228	1.953554	279.50	0.000	542.1939	549.8517

Czy średni wynik uczniów w Polsce był wyższy niż w Czechach? Tak, ale jaka jest różnica w wynikach uczniów? Możemy odjąć punktowe oszacowania, co da nam różnicę, ale nam potrzebne są też oszacowania błędu standardowego tej różnicy (i jej przedział ufności).

W zbiorze mamy dwa kraje, więc możemy to łatwo zrobić, korzystając z opcji `over`.

```
repest TIMSS , estimate(means ASMMATO@ ) over(IDCNTY , test)
over IDCNTY = 203 616
```

```
over var is IDCNTY
```

```
_pooled - IDCNTY = 203 .....
```

```
_pooled - IDCNTY = 616 .....
```

```
_pooled : _pooled
```

	Coefficient	Std. err.	z	P> z	[95% conf. interval]	
=203						
ASMMATO__m	530.4311	2.046724	259.16	0.000	526.4196	534.4426
=616						
ASMMATO__m	546.0228	1.953554	279.50	0.000	542.1939	549.8517
=d						
ASMMATO__m	15.59173	2.61209	5.97	0.000	10.47213	20.71133

4.3.1 Różnice między płciami w Polsce

Policzmy teraz różnice między chłopcami i dziewczętami w Polsce. Policzenie różnic w średnich jest proste i nie różni się od przykładów opisanych powyżej. Użyjemy opcji `estimate(means ASMMATO@ ) over(ITSEX , test)`, czyli podmienimy zmienną IDCNTY na zmienną, w której jest informacja o płci, czyli ITSEX.

Poniżej przyjrzymy się oszacowaniom wybranych percentyli i różnicom tych oszacowań dla chłopców i dziewcząt.

```
repest TIMSS if CTY == "POL", estimate(summarize ASMMAT0@, stats(p10 p25 p50 p75 p90)) over
over ITSEX = 1 2
over var is ITSEX
```

```
_pooled - ITSEX = 1 .....
_pooled - ITSEX = 2 .....
_pooled : _pooled
```

		Coefficient	Std. err.	z	P> z	[95% conf. interval]	
-----+-----							
ITSEX=1							
ASMMATO__p10		442.7281	3.916419	113.04	0.000	435.052	450.4041
ASMMATO__p25		493.9585	3.463758	142.61	0.000	487.1697	500.7474
ASMMATO__p50		545.2986	3.436177	158.69	0.000	538.5638	552.0334
ASMMATO__p75		591.5121	3.020694	195.82	0.000	585.5917	597.4326
ASMMATO__p90		632.209	5.727096	110.39	0.000	620.9841	643.4339
-----+-----							
ITSEX=2							
ASMMATO__p10		447.9956	6.46978	69.24	0.000	435.3151	460.6762
ASMMATO__p25		503.0338	3.975474	126.53	0.000	495.242	510.8256
ASMMATO__p50		557.1689	4.440397	125.48	0.000	548.4659	565.8719
ASMMATO__p75		604.0593	3.653707	165.33	0.000	596.8981	611.2204
ASMMATO__p90		646.2404	4.288714	150.68	0.000	637.8347	654.6461
-----+-----							
ITSEX=d							
ASMMATO__p10		5.267572	7.444711	0.71	0.479	-9.323793	19.85894
ASMMATO__p25		9.075292	5.397231	1.68	0.093	-1.503087	19.65367
ASMMATO__p50		11.87034	5.377029	2.21	0.027	1.331557	22.40912
ASMMATO__p75		12.54714	4.419928	2.84	0.005	3.884238	21.21004
ASMMATO__p90		14.03136	6.94566	2.02	0.043	.4181176	27.64461
-----+-----							

Wyniki pokazują **percentyle wyników z matematyki dla dziewcząt (ITSEX=1) i chłopców (ITSEX=2) w Polsce oraz różnice między nimi (ITSEX=d)**. Różnice są wyraźniejsze w górnej części rozkładu – np. mediana (p50) chłopców jest wyższa o około 12 punktów, a różnica ta jest statystycznie istotna. W dolnej części rozkładu (p10, p25) różnice są mniejsze i często nieistotne.

To może sugerować, że przewaga chłopców dotyczy głównie uczniów z lepszymi wynikami.

4.3.2 Poziomy umiejętności

W zbiorze TIMSS mamy zmienną opisującą poziomy umiejętności, a właściwie 5 osobnych zmiennych wyliczonych dla każdej z `pv`. Korzystając z komendy `fre` (`ssc install fre`) przyjrzymy się pierwszej z tej grupy zmiennych:

```
. fre ASMIBMO1
```

```
ASMIBMO1 -- INTERN. MATH BENCH REACHED WITH 1ST PV
```

			Freq.	Percent	Valid	Cum
Valid	1 Below 400		170	3.64	3.64	3
	2 At or above 400 but below 475	622	13.33	13.33	16.97	
	3 At or above 475 but below 550	1509	32.34	32.34	49.31	
	4 At or above 550 but below 625	1695	36.33	36.33	85.64	
	5 At or above 625	670	14.36	14.36	100.00	
	Total	4666	100.00	100.00		

Ta zmienna to przekodowana zmienna `pv`, według zdefiniowanych w badaniu granic poszczególnych poziomów. Powyższy rozkład jest nieważony. Aby uzyskać rozkład populacyjny, musimy użyć wag i uśrednić je dla wszystkich wartości `pv`. Tu z pomocą przychodzi `repest`:

```
. repest TIMSS if CTY=="POL" , estimate(freq ASMIBMO@)
```

```
_pooled.....
```

```
: _pooled
```

	Coefficient	Std. err.	z	P> z	[95% conf. interval]
ASMIBMO__1	3.818529	.4509266	8.47	0.000	2.934729 4.702329
ASMIBMO__2	13.21859	.7408295	17.84	0.000	11.76659 14.67059
ASMIBMO__3	32.43265	.9411117	34.46	0.000	30.5881 34.27719
ASMIBMO__4	36.15421	1.041231	34.72	0.000	34.11343 38.19498
ASMIBMO__5	14.37602	.8306454	17.31	0.000	12.74799 16.00406

W kolumnie **Coefficient** mamy wyliczenia odsetków uczniów na kolejnych poziomach, np. pierwszy wiersz pokazuje oszacowanie odsetka uczniów, których wynik jest niższy niż 400 pkt na skali TIMSS. W Polsce jest to około **4% uczniów** (przedział ufności CI95 to 2.9–4.7% uczniów).

Możemy też wyliczyć bardziej rozbudowane porównania. Wróćmy do przykładu z różnicami ze względu na płeć:

```
repest TIMSS if CTY=="POL" , estimate(freq ASMIBMO@) over(ITSEX, test)
over ITSEX = 1 2
over var is ITSEX
```

```
_pooled - ITSEX = 1 .....
_pooled - ITSEX = 2 .....
_pooled : _pooled
```

		Coefficient	Std. err.	z	P> z	[95% conf. interval]	
-----+							
ITSEX=1							
ASMIBMO__1		3.807798	.607817	6.26	0.000	2.616498	4.999097
ASMIBMO__2		14.41528	.9618309	14.99	0.000	12.53012	16.30043
ASMIBMO__3		34.56772	1.214754	28.46	0.000	32.18684	36.94859
ASMIBMO__4		35.13991	1.556351	22.58	0.000	32.08952	38.19031
ASMIBMO__5		12.06929	1.339601	9.01	0.000	9.443721	14.69486
-----+							
ITSEX=2							
ASMIBMO__1		3.829155	.5566068	6.88	0.000	2.738225	4.920084
ASMIBMO__2		12.0337	1.021051	11.79	0.000	10.03247	14.03492
ASMIBMO__3		30.31861	1.271462	23.85	0.000	27.82659	32.81063
ASMIBMO__4		37.15851	1.628476	22.82	0.000	33.96676	40.35026
ASMIBMO__5		16.66003	1.143771	14.57	0.000	14.41828	18.90178
-----+							
ITSEX=d							
ASMIBMO__1		.0213571	.7379917	0.03	0.977	-1.42508	1.467794
ASMIBMO__2		-2.381582	1.318514	-1.81	0.071	-4.965822	.2026571
ASMIBMO__3		-4.24911	1.625268	-2.61	0.009	-7.434576	-1.063644
ASMIBMO__4		2.018596	2.410272	0.84	0.402	-2.705451	6.742643
ASMIBMO__5		4.590739	1.85521	2.47	0.013	.9545949	8.226883

Porównanie odsetków uczniów osiągających różne poziomy umiejętności pokazuje, że **chłopcy częściej niż dziewczęta osiągają najwyższy poziom (625+), a rzadziej poziomy**

niższe. Na przykład 16.7% chłopców osiąga najwyższy poziom wobec 12.1% dziewcząt. Różnica ta jest istotna statystycznie.

Największe różnice występują na poziomach 3 i 5 — chłopcy rzadziej znajdują się na poziomie 3 (30.3% vs 34.6%), a częściej na poziomie najwyższym (16.7% vs 12.1%). Może to sugerować, że wśród chłopców jest więcej uczniów o bardzo wysokich osiągnięciach, mimo że różnice w średnich wynikach były umiarkowane.

Kontynuujemy
z
formatowaniem
i
ulepszaniem
czytelności
Twojego
tekstu.
Oto
kolejna
część, z
zastosowaniem
odpowiednich
wyróżnień
i
struktury.

5 Analiza danych ICCS

5.1 Wczytywanie i przygotowanie danych

Dane do Międzynarodowego Badania Edukacji Obywatelskiej (ICCS) znajdziemy na stronie IEA:

<https://www.iea.nl/data-tools/repository/iccs>

Pobieramy plik w formacie **SPSS**. Zawiera on, podobnie jak w przypadku PISA i TIMSS, pliki z danymi oraz przydatną dokumentację.

5.2 Analizy ICCS krok po kroku - przykłady

5.2.1 Odsetki kobiet i mężczyzn wśród nauczycieli w Polsce

Na początku przyjrzyjmy się proporcjom kobiet i mężczyzn w Polsce.

	Coefficient	Std. err.	z	P> z	[95% conf. interval]	
TD_GENDER_0	23.07338	.8382269	27.53	0.000	21.43048	24.71627
TD_GENDER_1	76.92662	.8382269	91.77	0.000	75.28373	78.56952

5.2.2 Różnice w postrzeganiu problemów w szkole w podziale na płeć

Sprawdźmy teraz, czy postrzeganie problemów społecznych w szkole (zmienna T_PROBSC) różni się ze względu na płeć nauczyciela. Najpierw przyjrzymy się tej zmiennej. Widzimy, że w Polsce przyjmuje ona wartości od 27 do około 75, dla 11 nauczycieli brakuje wartości.

```
codebook T_PROBSC
```

```
T_PROBSC Teachers' perceptions of social problems at school
```

```

Type: Numeric (double)
Label: labels202, but 94 nonmissing values are not labeled

Range: [27.16187,75.7119]      Units: .00001
Unique values: 94              Missing .: 11/2,259

Examples: 40.48312
           49.00636
           52.76074
           55.99976

```

Użyjemy `repest` z opcją `over` do porównania średnich i przetestowania istotności różnicy. Tak, jak w poprzednich przykładach korzystamy z *d* na oznaczenie różnicy między grupami (*difference*)

```

repest ICCS_T, estimate(mean T_PROBSC) over(TD_GENDER, test)
  over TD_GENDER = 0 1
over var is TD_GENDER

_pooled - TD_GENDER = 0 .
_pooled - TD_GENDER = 1 .

```

```
_pooled : _pooled
```

		Coefficient	Std. err.	z	P> z	[95% conf. interval]	
TD_GENDER=0							
T_PROBSC		45.72595	.5795467	78.90	0.000	44.59006	46.86184
TD_GENDER=1							
T_PROBSC		46.93191	.3567352	131.56	0.000	46.23272	47.6311
TD_GENDER=d							
T_PROBSC		1.205963	.5854271	2.06	0.039	.0585475	2.353379

Różnica między kobietami a mężczyznami jest statystycznie istotna ($p = 0.039$).

To samo możemy wyliczyć, korzystając z modelu regresji liniowej. Wynik dla predyktora `_1_TD_GENDER` jest identyczny z różnicą (`TD_GENDER=d`) z poprzedniej komendy.

```
repest ICCS_T, estimate(reg T_PROBSC i.TD_GENDER) results(add(r2 N))
```

```
_pooled.
: _pooled
```

		Coefficient	Std. err.	z	P> z	[95% conf. interval]	
_Ob_TD_GENDER		0 (omitted)					
_1_TD_GENDER		1.205963	.5854271	2.06	0.039	.0585475	2.353379
_cons		45.72595	.5795467	78.90	0.000	44.59006	46.86184
e_r2		.0027756	.0026804	1.04	0.300	-.0024778	.008029
e_N		2248	137.5354	16.34	0.000	1978.435	2517.565

Do komendy regresji dodaliśmy opcję, która zapisuje dodatkowe zmienne. Są to liczba obserwacji (`e_N`) oraz współczynnik determinacji (`e_r2`). W `repest` są one wyświetlane jako kolejne wiersze na dole tabeli wyników, co pozwala łatwo ocenić dopasowanie modelu oraz wielkość analizowanej próby. W naszym przypadku obserwujemy nieznaczne różnice między płciami, ale wyjaśniają one bardzo niewielką część — około 0.3% zróżnicowania wartości wskaźnika postrzegania problemów w szkole.