# Wybrane metody uzupełniania braków danych

Prezentacja na temat pracy magisterskiej

Marta Morawiec

# Spis treści



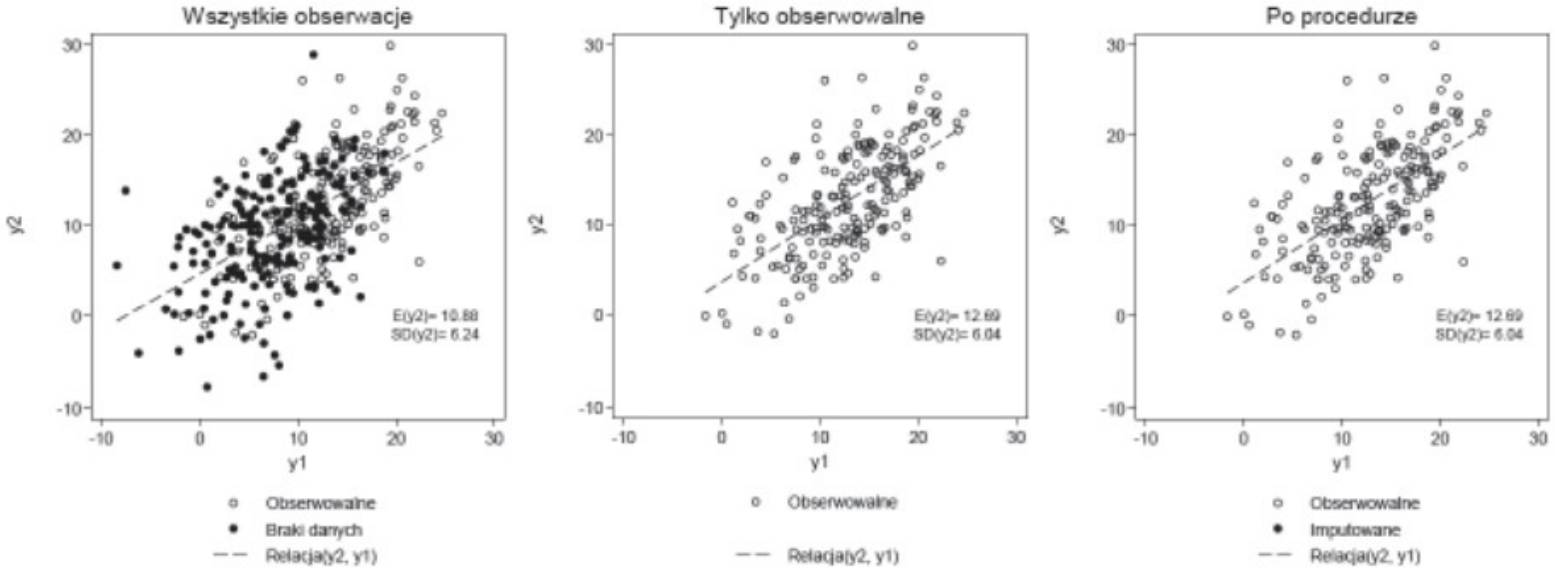
# Mechanizmy powstawania braków danych

- Mechanizm całkowicie losowy (**MCAR**: Missing completely at random);
- Mechanizm losowy (**MAR**: Missing at random);
- Mechanizm nielosowy (**MNAR** Missing not at random).

# Standardowe metody

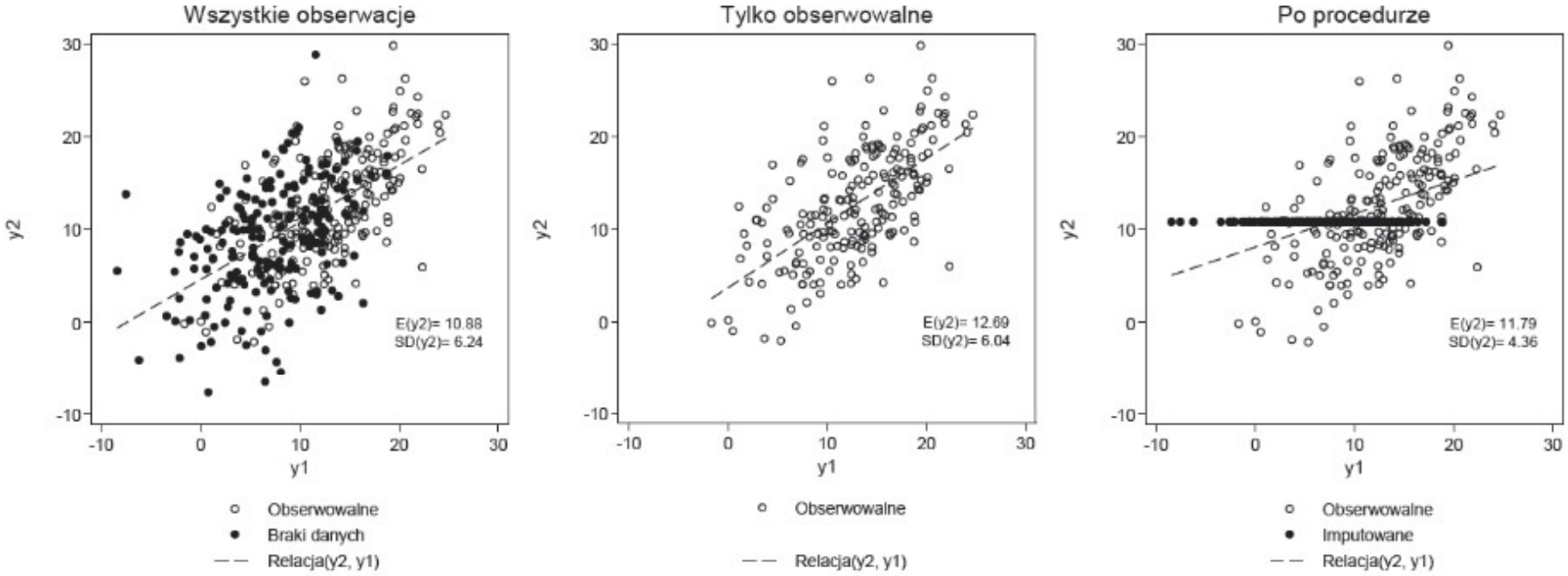
# Usuwanie braków danych

- Listwise deletion



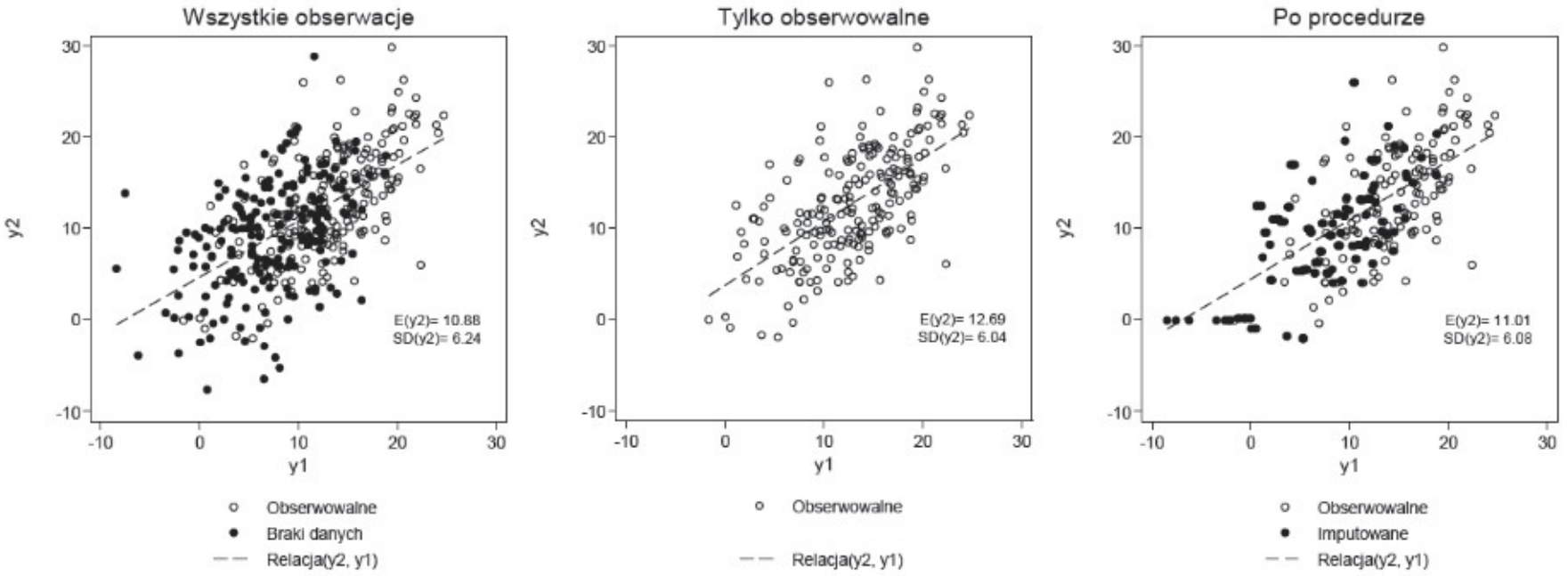

- Pairwise deletion – usuwanie rekordów parami rekord z brakiem i ten najbardziej z nim skorelowany

# Uzupełnianie średnią




Podejście do analiz z brakami danych – zastępowanie braków średnią wartością.
Mechanizm powstawania danych : **MAR**.

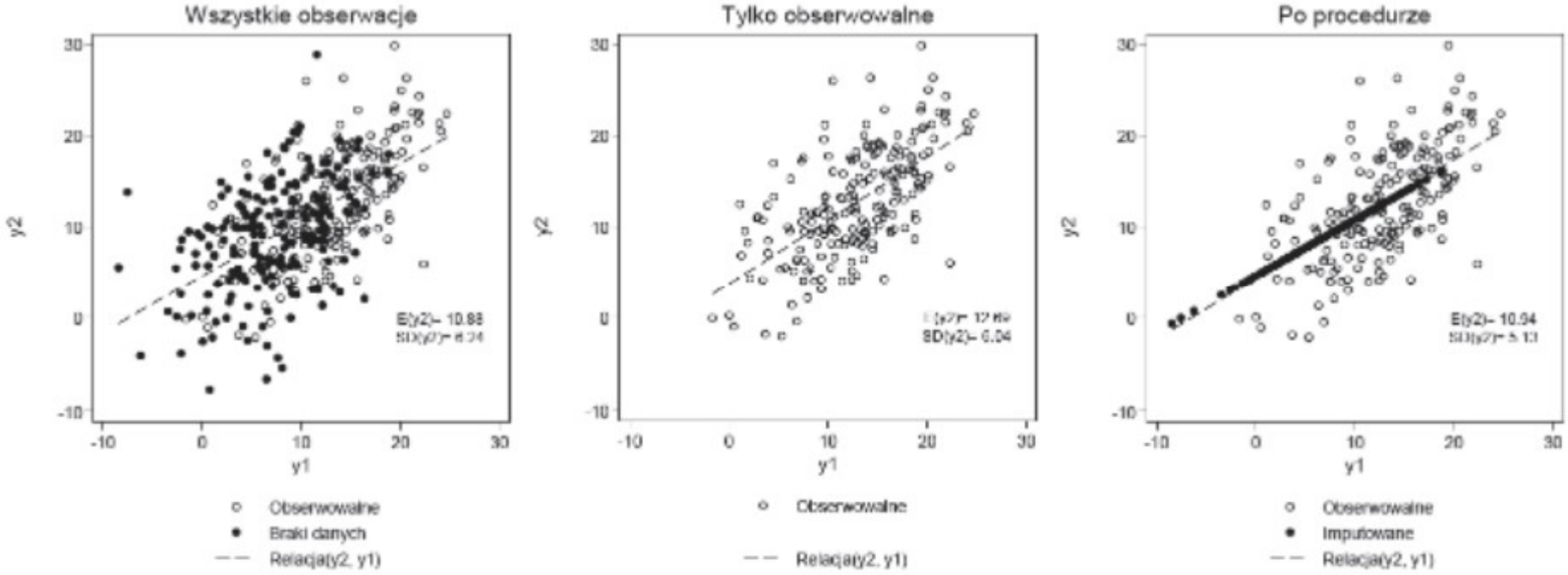
# Hot-deck imputation – imputacja nieparametryczna




Podejście do analiz z brakami danych – imputacja nieparametryczna
Mechanizm powstawania danych : **MAR**.

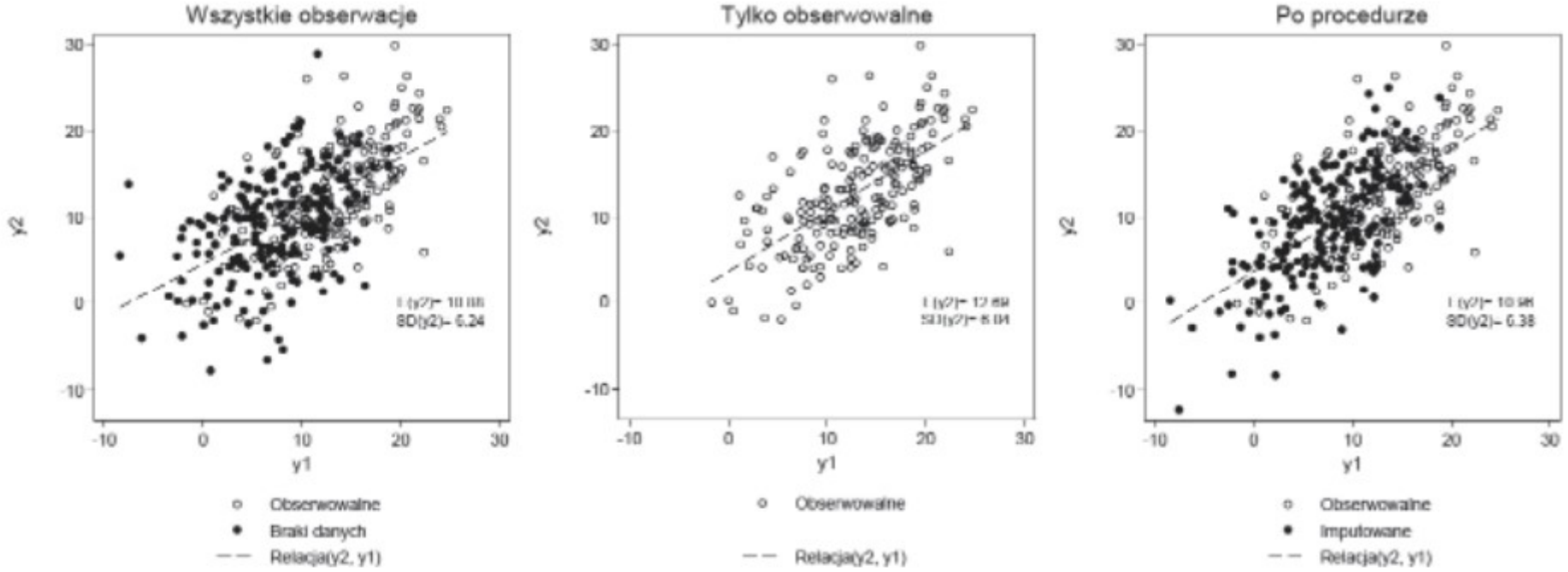
# Imputacja regresyjna

$$y_2 = \beta_0 + \beta_1 y_1 + e$$




Podejście do analiz z brakami danych –
imputacja regresyjna
Mechanizm powstawania danych : **MAR**.

# Stochastyczna imputacja regresyjna




Podejście do analiz z brakami danych – stochastyczna imputacja regresyjna
Mechanizm powstawania danych : **MAR**.

# Nowoczesne metody

# Metoda największej wiarygodności

- Początek lat 20. XXw Fisher
- **Główna zaleta:**

  Pozwala na uwzględnienie obserwacji z brakami danych w procesie estymacji wraz z kompletnymi danymi
- Najczęściej problem szacunków parametrów funkcji wiarygodności rozwiązuje się algorytmem EM.

# Wielokrotne imputacje

- Składa się z dwóch etapów:
  1. **Etap „I”** - imputacyjny
  2. **Etap „P”** – posterior
- Zazwyczaj liczba wielokrotnych imputacji nie przekracza kilkunastu.
- Można je zastosować do każdego typu danych.
- Jedyną ich wadą jest wrażliwość na błędną lub niekompletną specyfikację modelu.

# PODSUMOWANIE

- Poza stochastyczną metodą regresyjną klasycznych metod powinno się unikać.
- Rekomendowane metody to MNW i wielokrotne imputacje.
- W przypadku MCAR i małej liczbie braków usunięcie obserwacji z brakami nie będzie bardzo szkodliwe.
- Jeżeli odsetek danych jest znaczący i/lub prawdopodobne jest, iż nie jest to MAR, należy przeprowadzić dodatkowe analizy.

# Dziękuję za uwagę!

Wykresy i przykłady: https://pfp.ukw.edu.pl/archive/article-full/374/prokopek_wybrane_statystyczne_metody/