КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
ІМЕНІ ТАРАСА ШЕВЧЕНКА
Кафедра прикладної статистики

Кваліфікаційна робота бакалавра
за спеціальністю 124 Системний аналіз
на тему:

МНОЖИННЕ ЗАПОВНЕНННЯ ПРОПУСКІВ ЯК МЕТОД БОРОТЬБИ З ПРОПУЩЕНИМИ ДАНИМИ.

Студента 4 курсу
Губара Артема Андрійовича
Науковий керівник:
доцент, доктор фіз.-мат. наук
Розора І.В.

Робота заслухана на засіданні кафедри прикладної статистики та рекомендована до захисту в ДЕК, протокол №11 від 06червня 2022 р.

Завідувач кафедри прикладної статистики
професор, доктор фіз.-мат. наук Розора І.В.

Київ – 2022

ЗМІСТ

РЕФЕРАТ

Обсяг роботи 46 сторінок, 11 ілюстрацій, 13 таблиці, 29 джерел посилань.

ІМПУТАЦІЯ ПРОПУЩЕНИХ ДАНИХ, ПОВНІСТЮ ВИПАДКОВІ ПРОПУСКИ, ПРАВИЛА РУБІНА, РЕГРЕСІЙНІ МОДЕЛІ, СИСТЕМАТИЧНІ ПРОПУСКИ, СТОХАСТИЧНА РЕГРЕСІЙНА ІМПУТАЦІЯ.

Об’єктом роботи є процес імпутації пропущених даних.

Предметом роботи є методи та засоби імпутації даних.

Мета роботи: оцінка ефективності трьох методів імпутації пропущених даних: регресійна модель, Байєсівська стохастична регресійна модель, множинна імпутація пропущених даних за методикою Рубіна.

У першому розділі розглянуто теоретичні підходи до методів заповнення (імпутації) пропущених даних, а також типи пропусків та проблеми при їх імпутації. Проаналізовано різні підходи до аналізу пропусків у датафреймах та аналіз їх якості. В роботі розглянуто 3 методи заповнення пропущених даних: регресійна модель, Байєсівська стохастична регресійна імпутація, множинна імпутація пропущених даних за методикою Рубіна. Досліджено адекватність застосування алгоритмів заповнення для пропусків різного походження: повністю випадкових пропусків (MCAR) та систематичних пропусків, розподіл яких залежить як від пропущених значень, так і від спостережуваних ознак (MNAR). У другому розділі реалізована генерація даних з різними типами пропусків та заповнення їх описаними методами. Генерація реалізована за допомогою програмної мови R пакетів для роботи з пропущеними даними: mice та naniar.

ВСТУП

Оцінка сучасного стану об’єкта розробки. З проблемою відсутніх значень даних стикаються дослідники у багатьох галузях та типах досліджень. Пропуски знижують якість набору даних, впливаючи на аналіз даних і можливість застосування початкового дизайну дослідження, призводять до втрати даних та спотворення результатів. Це вимагає проведення досліджень для розробки практичних підходів до відновлення відсутніх даних шляхом аналізу наявних взаємозв'язків і характеристик даних. Сьогодні таких методів розроблено досить багато, від найпростіших (наприклад, виключення неповних спостережень, заповнення пропусків середнім значенням) до складних, в основі яких лежать складні алгоритми підбору пропущених значень (включаючи множинну імпутацію) залежно від характеру пропусків і припущень дослідника. Суттєвий внесок у формулювання проблеми аналізу даних із пропущеними значеннями і класифікацію підходів до їх обробки зробили фундаментальні праці Д. Рубіна і Р. Літтла [1-3].

Дослідження щодо удосконалення методів роботи з відсутніми даними продовжуються, чому сприяє розвиток комп’ютерної техніки та інформаційних технологій. Тому актуальним є створення засобів для реалізації різних алгоритмів імпутації пропущених даних та оцінки їх ефективності для практичного використання у вибіркових дослідженнях.

Об’єктом дослідження в даній роботі є процес імпутації пропущених даних.

Предмет дослідження – методи та засоби імпутації даних.

Метою роботи є оцінка ефективності трьох методів імпутації пропущених даних: регресійна модель імпутації даних, Байєсівська стохастична регресійна модель, множинна імпутація пропущених даних за методикою Рубіна.

Практичній реалізації мети роботи передував аналіз літературних даних щодо поширених типів пропусків та методів їх заповнення (імпутації), а також

переваг та недоліків різних методів. Основну увагу приділено регресійним методам імпутації даних (множинна лінійна регресія та Байєсівська стохастична регресія) та відновленню даних за методикою Рубіна. Досліджено адекватність застосування алгоритмів заповнення для пропусків різного походження: повністю випадкових пропусків (MCAR) та систематичних пропусків, розподіл яких залежить як від пропущених значень, так і від спостережуваних ознак (MNAR). Генерація даних з різними типами пропусків (MCAR і MNAR) та їх відновлення описаними методами реалізовані за допомогою програмної мови R пакетів для роботи з пропущеними даними: mice та naniar.

РОЗДІЛ 1

МЕТОДИ МНОЖИННОГО ЗАПОВНЕННЯ ПРОПУСКІВ ДАНИХ

1.1 Проблема пропусків у даних різних досліджень

Відсутні дані створюють проблеми для аналізу реальних даних. В результатах багатьох досліджень (соціологічних, медичних, екологічних, соціально-економічних, розпізнаванні образів тощо) є пропуски в даних, що призводять до їх часткової втрати, зміщення результатів та обмеження у використанні різних методів обробки [4, 5, 6, 7].

Причини пропусків даних можуть бути різними, тому знання механізму, що призводить до відсутності значень, є ключовим при виборі методів аналізу та інтерпретації результатів. Основними причинами відсутності даних є неможливість отримання або обробки інформації, її викривлення або приховування, втрати частини даних під час їх передавання чи зберігання. Механізм породження пропусків дає розуміння ступеня важливості втраченої інформації, адже неповні дані несуть у собі нову інформацію, необхідну для дослідження, тому їх важливо включати в аналіз [6].

Існують методи, що дозволяють боротися з пропусками на етапі аналізу даних. На сьогоднішній день існує багато таких методів від найбільш простих – як то, виключення неповних спостережень, заміна пропущеного значення певним числом або середнім значенням, до складних, що базуються на підборі пропущених значень залежно від характеру пропусків даних та припущень дослідника [2, 5, 8].

Найбільш важливою для вибору методу відновлення пропущених даних є природа пропусків, що характеризується їх випадковістю або систематичністю, яка була вперше систематизована Rubin D. B. & Little R. J. [3] і описана в багатьох роботах [4, 9, 10, 11].

Рубін Д. (1976) [1] запропонував терміни, які класифікують зв'язок між відсутністю, яка є процесом, що призводить до відсутніх значень, і самими відсутніми та спостережуваними значеннями. Нехай $\boldsymbol{y}$ - вектор даних з одного суб'єкта, який містить як відсутні, так і спостережувані значення. Без втрати загальності вектор даних $\boldsymbol{y}$ можна розділити на $\boldsymbol{y_{mis}}$ (missing) та $\boldsymbol{y_{obs}}$ (observed), підвектори $\boldsymbol{y}$, які відсутні та спостерігаються відповідно. За $\boldsymbol{X}$ приймаємо будь-яку іншу повністю спостережувану змінну, що цікавить. Можливі три типи пропусків в залежності від міри випадковості їх виникнення:

• повністю випадкові пропуски або MCAR (missing completely at random): ймовірність того, що компоненти $\boldsymbol{y}$ відсутні не пов'язана з будь-якими іншими спостережуваними даними $\boldsymbol{y_{obs}}$, $\boldsymbol{X}$ або значенням неспостереженого $\boldsymbol{y_{mis}}$. Крім того, пропуски називають повністю випадковими, якщо їх виникнення зумовлене дизайном дослідження, тобто не залежить від самої одиниці спостереження. Для таких пропусків залежність між імовірністю пропуску, спостережуваними даними $\boldsymbol{y_{obs}}$ або можливими значеннями для пропущених даних $\boldsymbol{y_{mis}}$ відсутня, і уточнити прогноз про пропущені значення за допомогою наявної інформації не можна [12]. В цьому випадку спостереження з наявними значеннями утворять просту випадкову підвибірку, яка є незміщеною вибіркою з генеральної сукупності, до котрої можна застосовувати такі ж статистичні критерії, що і до оригінальної вибірки, однак їх потужність знижується через зменшення її обсягу;

• пропуск називають випадковим або MAR (missing at random), якщо імовірність того, що $\boldsymbol{y}$ відсутній, залежить лише від спостережуваних значень $\boldsymbol{y_{obs}}$ та $\boldsymbol{X}$, але не від значень самих відсутніх даних. Тобто, значення випадкових пропусків можна передбачити за допомогою інших змінних в масиві даних, для яких значення присутні [12]. У даному випадку не можна стверджувати, що спостереження без пропусків утворять випадкову підвибірку з реальної вибірки, але розподіл випадкової величини для таких пропусків має значення тільки в підвибірках, а не на всій генеральній сукупності [3]. Тому вилучення чи заміна

пропусків, як і у випадку MCAR, не призводить до істотного спотворення результатів. Пропуски виду MAR, так само як MCAR, називають ігнорованими;

• третій тип пропусків не є випадковим, оскільки розподіл пропусків залежить як від пропущених значень, так і від спостережуваних ознак. Тобто ймовірність того, що $\boldsymbol{y}$ відсутній, залежить від $\boldsymbol{y_{mis}}$, а також, можливо, від $\boldsymbol{y_{obs}}$ та $\boldsymbol{X}$. У цьому випадку пропуски називають систематичними або MNAR (missing not at random) і відносять до тих, що не ігноруються. Таких видів пропусків можна уникнути або усунути тільки на етапі збору інформації, але повторний збір може виявитися нездійсненним і здебільшого марним.

Тому важливим є застосування методів, що дозволяють боротися з ігнорованими (MAR, MCAR) пропусками вже на етапі аналізу даних, коли інформація зібрана і повернутися до початкового етапу вже не можна.

Серед шляхів вирішення проблеми неповних даних розглядають кілька способів [3]. З них простим і досить поширеним способом є виключення некомплектних об'єктів із вихідної вибірки. Так, за даними аналізу 262 епідеміологічних досліджень з відсутніми даними в 81% досліджень проводився аналіз тільки повних випадків [13]. Даний метод легко реалізується, але необхідною умовою його застосування є тип пропусків MCAR та невелика кількість пропусків. Крім того, такий підхід не тільки зменшує обсяг вибірки, але за наявності великої кількості пропусків часто призводить до зміщення результатів і спотворення статистичних висновків, оскільки неповні дані несуть нову інформацію.

Другим шляхом роботи з неповними даними є застосування спеціально розроблених математичних методів їх аналізу, зокрема метод зважування або метод максимальної правдоподібності та EM-алгоритм [3]. Проте, при такому підході суттєво зростає складність проведеного аналізу.

Наступним шляхом боротьби з неповними даними, реалізованим у багатьох алгоритмах, є підхід із заповненням (імпутацією) пропущених значень. У цілому імпутацію можна описати як заміну відсутніх значень новими

значеннями за заданими критеріями. Існує багато способів імпутації пропусків: заповнення середнім або іншим певним значенням, розрахунок можливого значення за допомогою регресійної моделі тощо [3, 14, 15, 16, 17]. Особливістю цих алгоритмів є заповнення пропусків значеннями, які підбираються самим алгоритмом.

Слід відзначити, що використання будь-яких засобів заповнення пропусків може змінити структуру вибірки, яка буде отримана на основі існуючих неповних даних, що може викривити реальний розподіл даних у вибірці і зменшити значущість отриманих результатів. Обираючи конкретний алгоритм для заповнення пропусків, варто враховувати, що можливість його застосування істотно залежить від методу аналізу даних, який передбачається використати надалі [9].

Серед методів імпутації пропущених даних можна виділити найменш складні, зокрема методи із заповненням. При цьому підході пропущені значення вихідної вибірки заповнюються та отримані «повні» дані обробляються звичайними методами.

При заповненні пропусків середнім значенням обчислюється середнє арифметичне значення змінної, яка містить пропущені значення, за наявними даними. До недоліків методу можна віднести спотворення розподілу даних та зменшення дисперсії вихідних даних. Крім того, заповнення пропусків даних певними значеннями потребує обґрунтування методу формування підставлених значень. Зокрема, замість середнього значення більш адекватним може бути зважене середнє або медіана.

Підстановка з підбором усередині груп та підбір найближчого сусіда [18]. У першому випадку формуються групи, і пропуски в кожній групі заповнюються присутніми значеннями з неї. Заповнення з підбором поширене. Воно може включати складні схеми відбору об'єктів. Другий підхід заснований на введенні метрики d для вимірювання відстані між об'єктами, визначеної у просторі супутніх змінних, та виборі підстановки по об'єкту з присутнім

значенням, найближчим до об'єкта з пропуском. Такі схеми найближчого сусіда потребують значних обчислювальних витрат.

Наступна група методів з імпутації пропущених даних заснована на використанні регресії.

1.2 Регресійна модель імпутації даних

В основу цієї групи методів покладено відомі алгоритми регресійного аналізу [19]. При імпутації пропущених значень інформація інших змінних використовується для прогнозування відсутніх значень у змінній за допомогою регресійної моделі. Причому метод можна застосовувати лише за наявності вірогідної кореляції з іншими показниками масиву даних [20]. Зазвичай спочатку оцінюється регресійна модель у спостережуваних даних, а потім, використовуючи вагові коефіцієнти регресії, прогнозуються та замінюються відсутні значення. Іншими словами, доступна інформація для повних і неповних випадків використовується для прогнозування значення конкретної змінної.

У загальному вигляді регресійний метод імпутації пропущених значень у таблицях даних можна описати наступним чином. Позначимо дані як **Y**= (y_1,y_2,...,y_n) – вектор результуючих значень досліджуваної змінної, X= (x_1, x_2, … x_k) – вектор пояснювальних змінних (предикторів), пов’язаних з **Y**. Через **θ** позначаємо неспостережувані векторні величини або параметри досліджуваної сукупності, а через ˜y – невідомі, але потенційно спостережувані значення змінної.

Найпростішим і найбільш широко використовуваним варіантом цієї моделі є звичайна множинна лінійна модель, в який розподіл **Y** для заданого X є нормальним із середнім, яке є лінійною функцією X:

$$E(y_i \mid \beta, X) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik},$$

(2.1)

для i = 1,…,n.

В даному випадку **Y** представлений безперервними величинами, змінні **X** можуть бути дискретними або безперервними, а $\boldsymbol{\theta} = (\beta_0, \beta_1, \ldots, \beta_k)$.

Для визначення коефіцієнтів β використовують лише повні або комплектні (без пропусків) дані $\boldsymbol{y_{obs}}$, і метод найменших квадратів (МНК).

Отримані коефіцієнти надалі використовуються для відновлення пропусків.

Моделі лінійної регресії зазвичай використовуються для імпутації інтервальних та абсолютних (безперервних) змінних. При цьому розраховується конкретне значення змінної. Для дихотомічних змінних використовуються моделі бінарної логістичної регресії (2.2), для категоріальних змінних з кількістю категорій більше 2 – моделі політомічної (поліноміальної) логістичної регресії.

$$Pr(y_i = 1) = logit^{-1}(X_i\beta), \tag{2.2}$$

При цьому, якщо у випадку бінарної логістичної регресії незалежна змінна (**X**) може мати безперервну шкалу, то поліноміальна логістична регресія придатна лише для категоріальних незалежних змінних, причому має значення, чи належать вони до шкали найменувань чи до порядкової шкали. Крім того, на відміну від моделі множинної лінійної регресії для порядкових і номінальних значень з деякою ймовірністю передбачається категорія, до якої має бути віднесений об’єкт.

Регресійна модель передбачає найбільш ймовірне значення відсутніх даних, що в свою чергу залежить від успішного вибору взятої за основу регресійної моделі.

В цілому регресійні методи вимагають виконання ряду застережень та перевірок вхідних факторів на мультиколінеарність, гетероскедастичність, автокореляцію та застосування модифікованих версій МНК. Серед умов застосування даного методу можна виділити вимогу щодо типу пропусків MAR і вимоги до регресійного аналізу. Проблема полягає в тому, що розраховані дані не мають члена похибки, включеного в їхню оцінку, тому оцінки ідеально підходять уздовж лінії регресії без будь-якої залишкової дисперсії. Це призводить до зменшення варіації (дисперсії) значень, штучного посилення зв'язків (кореляції) між характеристиками і передбачає більшу точність замінених значень, ніж це виправдано [21].

1.3 Байєсівські підходи до імпутації пропущених даних

За наявності апріорної інформації щодо вірогідності певних варіантів відповідей (значень) можливо використання байесівського підходу для формування значень для заповнення пропусків [22].

Стохастична регресійна імпутація була розроблена для вирішення проблеми детермінованої регресії [21 Gelman]. Імпутація стохастичної регресії додає випадкову помилку до прогнозованого значення і, отже, може відтворити кореляцію X і Y більш належним чином. Байєсівські методи статистичного аналізу, враховують попередні знання про розподіли відсутніх даних та застосування їх до вибіркової сукупності. Застосування байєсівських методів для імпутації відсутніх даних може усунути деякі недоліки простіших методів імпутації для обробки даних, які не є MAR (Missing at Random).

Розглянемо формальне байєсовське обґрунтування умовного моделювання у контексті звичайної лінійної регресії [21 Gelman].

Числові «дані» в задачі регресії включають як X, так і Y. Таким чином, повна байєсівська модель включає розподіл для X, $p(X|\psi)$, індексований вектором параметрів ψ, і, таким чином, включає спільну ймовірність $p(X, y|\psi, \theta)$ разом із попереднім розподілом $p(\psi, \theta)$.

В методі звичайної лінійної регресії передбачається, що розподіл X не надає інформації про умовний розподіл Y за даними X; тобто ми припускаємо попередню незалежність параметрів θ, що визначають p(y|X, θ), і параметрів ψ, що визначають p(X|ψ).

Таким чином, з точки зору Байєса, визначальна характеристика «регресійної моделі» полягає в тому, що він ігнорує інформацію, надану X про (ψ, θ).

Припустимо, що ψ і θ незалежні у своєму попередньому розподілі; тобто

$$p(\psi, \theta \mid X, y) = p(\psi \mid X)p(\theta \mid X, y),$$

і ми можемо проаналізувати другий фактор самостійно (тобто як стандартну модель регресії), без втрати інформації:

$$p(\theta \mid X, y) \propto p(\theta)p(y \mid X, \theta)$$

Коли вибираються предикторні змінні X (наприклад, у спланованому експерименті), їх ймовірність p(X) відома, а параметри ψ відсутні.

Практична перевага використання такої регресійної моделі полягає в тому, що набагато легше вказати реалістичний умовний розподіл однієї змінної за k інших, ніж спільний розподіл для всіх k+1 змінних.

Ми визначаємо спочатку апостеріорний розподіл для коефіцієнтів регресії β, залежний від σ (середнього квадратичного відхилення), а потім граничний апостеріорний розподіл для σ^2 (дисперсії). Тобто ми розраховуємо спільний апостеріорний розподіл для β і σ^2 як $p(\beta, \sigma^2 \mid y) = p(\beta \mid \sigma^2, y)p(\sigma^2 \mid y)$.

Умовний апостеріорний розподіл (векторного) параметру β, заданий σ, є експонентою квадратичної форми в β і, отже, є нормальним. Використовуємо позначення

$$\beta \mid \sigma, y \backsim N(\hat{\beta}, V_\beta \sigma^2),$$

$$\hat{\beta} = (X^T X)^{-1} X^T y,$$
$$V_\beta = (X^T X)^{-1}.$$

Граничний апостеріорний розподіл σ^2 може бути написаний як

$$p(\sigma^2 \mid y) = \frac{p(\beta, \sigma^2 \mid y)}{p(\beta \mid \sigma^2, y},$$

Як видно, він має масштабовану обернену χ^2 форму:

$$\sigma^2 \mid y \backsim Inv - x^2(n-k, s^2),$$

де

$$s^2 = \frac{1}{n-k}(y - X\hat{\beta})^T (y - X\hat{\beta}).$$

На практиці $\hat{\beta}$ і V_β можна обчислити за допомогою стандартного програмного забезпечення для лінійної регресії.

Формування Байєсівського висновку вимагає створення моделі розподілу за параметрами θ. Для цієї задачі, оцінка відсотка є середнє $\bar{y}$ для кінцевої популяції, яке можна виразити як

$$\bar{y} = \frac{n}{N}\bar{y}_{obs} + \frac{N-n}{N}\bar{y}_{mis}, \qquad (2.3)$$

де $\mathbf{y_{obs}}$ і $\mathbf{y_{mis}}$ – середні значення спостережуваних і відсутніх y_i.

Ми можемо визначити апостеріорний розподіл $\bar{y}$, використовуючи моделювання $\mathbf{y_{mis}}$ з його апостеріорним прогнозним розподілом. Почнемо з моделювання θ: θ_s, s = 1,...,S. Для кожного визначеного θ_s ми потім визначаємо вектор $\mathbf{y_{mis}}$ таким чином

$$p(y_{mis} \mid \theta^s, y_{obs}) = p(y_{mis} \mid \theta^s) = \prod_{i:i_i=0} p(y_i \mid \theta^s),$$

а потім усереднюємо значення змодельованого вектору, щоб отримати вибірку $\mathbf{y_{mis}}$ з його апостеріорного прогнозного розподілу. Оскільки $\mathbf{y_{obs}}$ відомий, ми можемо обчислити середнє ȳ для кінцевої популяції, з використанням (2.3) і вибірки $\mathbf{y_{mis}}$.

Хоча зазвичай оцінка розглядається як ȳ, у більш загальному вигляді вона може бути будь-якою функцією **y** (наприклад, медіана значень ȳi або середнє значення log ȳi).

Еквівалентність великої вибірки суперпопуляції та висновки про кінцеву популяцію. Якщо N − n є великим, то ми можемо використовувати центральну граничну теорему для апроксимації розподілу вибірки $\mathbf{y_{mis}}$:

$$p(\bar{y}_{mis} \mid \theta) \approx N(\bar{y}_{mis} \mid \mu, \frac{1}{N-n}\sigma^2),$$

де μ=μ(θ) = E(yi|θ) і σ^2 =σ^2(θ) = var(yi|θ). Якщо n також велике, то апостеріорні розподіли θ і будь-яких його компонентів, таких як μ і σ, є приблизно нормальними, отже, апостеріорний розподіл $\mathbf{y_{mis}}$ є приблизно сумішшю нормалей і отже сам по собі нормальний. Більш формально,

$$p(\bar{y}_{mis} \mid y_{obs} \approx \int p(\bar{y}_{mis} \mid \mu, \sigma) p(\mu, \sigma \mid y_{obs}) d\mu d\sigma;$$

оскільки і N, і n стають великими при фіксованому N/n, це приблизно нормально з

$$E(\bar{y}_{mis} \mid y_{obs}) \approx E(\mu \mid y_{obs}) \approx \bar{y}_{obs},$$

і

$$var(\bar{y}_{mis} \mid y_{obs}) \approx (\mid mu \mid y_{obs}) + E(\frac{1}{N-n}\sigma^2 \mid y_{obs})$$

$$\approx \frac{1}{n} s_{obs}^2 + \frac{1}{N-n} s_{obs}^2$$

$$= \frac{N}{n(N-n} s_{obs}^2,$$

де s^2_{obs} – вибіркова дисперсія спостережуваних значень y_{obs}. Об'єднавши це приближення апостеріорного розподілу з (2.3), отримаємо не тільки, що $p(y|y_{obs}) \rightarrow p(\mu|y_{obs}s)$, але, більше загалом, це

$$\bar{y} \,|\, y_{obs} \approx N(\bar{y}_{obs}, (\frac{1}{n} - \frac{1}{N}) s_{obs}^2) \,.$$

Це офіційне байєсівське обґрунтування висновку нормальної теорії для досліджень зі скінченною вибіркою.

Для нормального $p(yi|\theta)$ зі стандартним неінформативним попереднім розподілом точний результат є

$$\bar{y} \,|\, y_{obs} \backsim t_{n-1}(\bar{y}_{obs}, (\frac{1}{n} - \frac{1}{N}) s_{obs}^2) \,. \qquad (2.4)$$

Таким чином, при використанні байесівської стохастичної регресії невизначеність пропущених даних враховується не лише шляхом додавання дисперсії помилок до прогнозованих значень моделі лінійної регресії, а й шляхом урахування невизначеності в оцінці коефіцієнтів регресії моделі. Використовується байесівська ідея у тому, що немає одного (істинного) коефіцієнта регресії, але самі коефіцієнти регресії також підпорядковуються розподілу.

1.4 Множинна імпутація пропущених даних за методикою Рубіна

Одним із поширених способів коригування помилки, пов'язаної із заповненням пропущених значень, і врахування певної невизначеності підставлених значень є метод множинної (багаторазової) імпутації (Multiple Imputation (MI)) [3]. Ідея методу полягає в тому, щоб виконати процедуру

імпутації, яка містить випадковий компонент кілька разів, створюючи кілька повних наборів даних. Бажана оцінка аналізу, така як середнє значення або коефіцієнт регресії, потім розраховується для кожного набору даних окремо. Оцінки об'єднуються за допомогою простих правил комбінування, також відомих як правила Рубіна [2] і визначаються наступним чином.

Нехай δ — параметр, оцінку якого ми хочемо отримати з аналізу. Враховуючи M отриманих наборів даних, M оцінок δ : $(\delta\hat{}_1,\delta\hat{}_2,...,\delta\hat{}_m)$ генеруються та використовуються для обчислення наступних величин:

- Загальна оцінка – це середнє значення окремих точкових оцінок

$$\hat{Q} = \frac{1}{M}\sum_{m=1}^{M}\hat{\delta}_m \tag{2.5}$$

- Внутрішньогрупова дисперсія в межах імпутації для набору даних – це середнє значення індивідуальних дисперсій

$$U = \frac{1}{M}\sum_{m=1}^{M} Var(\hat{\delta}_m) \tag{2.6}$$

- Міжгрупова дисперсія між наборами даних – це дисперсія оцінок

$$B = Var(\hat{\delta}_1, \hat{\delta}_2, \ldots, \hat{\delta}_m) \tag{2.7}$$

- Загальна дисперсія комбінованої оцінки враховує як внутрішньогрупову, так й міжгрупову дисперсію

$$T = U + (1 + \frac{1}{M})B, \tag{2.8}$$

- 95% інтервали для параметрів обчислюються як

$$Q\hat{} \pm 1.96 * \sqrt{T} \tag{2.9}$$

Таким чином, Рубін [2] представив загальну структуру для створення множинних обчислень для відсутніх значень з урахуванням параметра моделі генерування даних, при якому на місце кожного пропуску підставляються кілька можливих значень, тобто генеруються кілька масивів даних. Багаторазова імпутація замінює кожне відсутнє значення кількома правдоподібними значеннями. На першому кроці набір даних із відсутніми значеннями (тобто неповний набір даних) копіюється кілька разів. Потім на наступному кроці пропущені значення замінюються розрахованими значеннями в кожній копії набору даних. При цьому в кожній копії ці значення будуть відрізнятися через випадкові варіації (обчислення проводиться з урахуванням генератора випадкових чисел). Це призводить до безлічі наборів даних. На третьому кроці отримані масиви даних окремо аналізуються з використанням стандартних методів, а результати об’єднуються для отримання неупереджених оцінок з використанням правила Рубіна. Отже, множинні методи імпутації використовують процеси вставки (заміни пропущених значень), аналізу та об'єднання [23].

Правдоподібність замінених значень можна встановити шляхом їх порівняння з реальними (спостережуваними) даними. Зокрема, графічно: якщо поставлені замість пропущених значення знаходяться в діапазоні даних, що спостерігаються, тобто немає великих відмінностей між розрахованими і спостережуваними значеннями, то можна зробити висновок, що ці значення правдоподібні. Дослідження ефективності множинної імпутації можна також провести шляхом аналізу точності введених значень зі зміною частки відсутніх даних в одному або кількох стовпцях або за допомогою регресійних моделей.

Слід відзначити, що проведення аналізу декілька разів на кожному масиві з наступним їх об'єднанням досить витратно. Навіть з урахуванням того, що певні статистичні пакети (SPSS, SAS, R тощо) автоматизують процедуру множинного заповнення пропусків. Зменшення обсягів роботи з методом множинного заповнення пропусків можна досягти врахуванням властивостей конкретного дослідження. Зокрема, такими властивостями можуть бути тип

шкали змінної, що вивчається з пропусками, частка пропусків у масиві і метод аналізу даних, який буде застосовуватися до досліджуваної змінної.

Важливим є питання оптимальної кількості наборів даних, що генеруються. Graham J. W. et al. рекомендували, щоб принаймні 20 введених наборів даних були необхідні, щоб обмежити втрату потужності під час тестування зв'язку між змінними [8]. Bodner T.E. (2008) [24] та White I. R. et al. (2011) [25] запропонували емпіричне правило, засноване на ідеї, що частка інформації, яку бракує, часто нижче, ніж відсоток пропущених значень. Тому кількість згенерованих наборів даних має дорівнювати принаймні відсотку відсутніх значень.

Van Buuren (2018) [26] стверджує, що кількість ітерацій може залежати від кореляції між змінними та відсотку відсутніх даних у змінних. Він припустив, що для досягнення доброї узгодженості даних достатньо 5-20 ітерацій. Це може бути скориговано, якщо відсоток відсутніх даних високий. Водночас, використання сучасних комп'ютерів дозволяє використовувати більшу кількість ітерацій.

Відомий алгоритм множинної імпутації пропущених даних за допомогою ланцюгових рівнянь (MICE) [26]. В алгоритмі MICE для отримання даних для заміни пропусків використовується ланцюг рівнянь регресії, тобто змінні з відсутніми даними обчислюються одна за одною. Регресійні моделі використовують інформацію з усіх інших змінних у моделі, тобто (умовних) моделей імпутації. Для того, щоб додати мінливість вибірки до обчислень, додається залишкова похибка для створення імпутованих значень. Ця залишкова похибка може бути додана безпосередньо до передбачуваних значень, що по суті подібне до повторюваної імпутації стохастичної регресії протягом кількох циклів імпутації. Залишкова дисперсія також може бути додана до оцінок параметрів регресійної моделі, тобто до коефіцієнтів регресії, що робить її байєсівським методом.

Багатоваріантна імпутація значень відсутніх даних за допомогою алгоритму MICE є найбільш підходящою формою для даних типу MAR та MNAR. Для імпутації за допомогою MICE існує майже 22 методи, які можна вибрати відповідно до наявного набору даних [27]. Виходячи з спостережень, для імпутації відсутніх даних безперервних змінних найкращим підходом буде використання лінійної регресії, для дихотомічних – логістичної регресії.

В цілому є багато модифікацій платформи MI. Зокрема, в роботі Reiter and Raghunathan [28] розглядаються ситуації, коли правила Рубіна діють і як їх слід виправляти за певних обставин. Raghunathan T.E. et. al. [29] стверджують, що найкращою або найбільш підходящою структурою для виконання багаторазової імпутації є повністю байєсівська модель.

РОЗДІЛ 2

ПРАКТИЧНА РЕАЛІЗАЦІЯ МЕТОДІВ МНОЖИННОЇ ІМПУТАЦІЇ ПРОПУЩЕНИХ ДАНИХ

У даному розділі наведені результати генерації кількісних даних з різними типами пропусків (MCAR і MNAR) та їх відновлення (імпутація) з використанням регресійних методів (множинна лінійна регресія, Байєсівська стохастична регресійна модель) та множинне відновлення пропущених даних за методикою Рубіна. Для реалізації були обрані пакети програмної мови R для роботи з пропущеними даними.

2.1 Генерація пропущених значень безперервних даних

Для початку імпортуємо усі необхідні пакети та обраний DataSet, який був сформований для аналізу та передбачення ймовірності того, що клієнт придбає певний продукт банку (ощадний рахунок, кредитну карту, інвестиції, тощо). DataSet описує демографічні та інші дані клієнтів певного банку.

Основою для практичної частини буде офіційній пакет R - “Mice” (Multivariate Imputation by Chained Equations), в якому реалізовано багато методів для заповнення пропущених даних. Також використовуються пакети для більш зручного зображення даних, такі як: Tidyverse, skimr, flextable.

На першому етапі імпортуємо усі необхідні пакети та бібліотеки:

```
library(tidyverse) # Набір пакетів для роботи з даними та якісного їх представлення
library(flextable) # Пакет для роботи з табличними даними
library(mice) # Бібліотека для множинних вставок пропущених даних
library(skimr) # Пакет для формування звіту
library(naniar) # Бібліотека для роботи та побудови графіків для пропущених
```

даних

```
library(knitr) # пакет, що поєднує обчислення та звітність
```

Зчитуємо дані DataSet з .csv файлу

```
df <- read_csv('Assignment-2_Data.csv')
df <- subset(df, select =-c(previous, y, pdays, contact, loan, campaign, default, poutcome, day, housing, Id))
df <- na.omit(df)
df<-df[!(df$age>=100 | df$age<=18),]
df<-df[!(df$balance<=1),]
df<-df[!(df$duration<=1),]
```

Генеруємо невеликий звіт та аналізуємо обраний DataSet.

#Генеруємо таблицю з перших 15-ти записів у DataSet

```
table <- head(df, 15)
```

#Конвертуємо таблицю у flextable

```
my_table <- flextable(table)
```

#Налаштовуємо зовнішній вигляд таблиці

```
my_table <- my_table %>%
```

#Коригуємо розмір таблиці

```
 autofit() %>%
```

#Друкуємо фрагмент таблиці даних (табл. 1)

```
my_table
```

Таблиця 1 – Фрагмент вихідних даних DataSet

age	job	marital	education	balance	month	duration
44	technician	single	secondary	29	may	151
33	entrepreneur	married	secondary	2	may	76
47	blue-collar	married	unknown	1,506	may	92
33	unknown	single	unknown	1	may	198
35	management	married	tertiary	231	may	139
28	management	single	tertiary	447	may	217
41	admin.	divorced	secondary	270	may	222
29	admin.	single	secondary	390	may	137
53	technician	married	secondary	6	may	517
58	technician	married	unknown	71	may	71
57	services	married	secondary	162	may	174
51	retired	married	primary	229	may	353
45	admin.	single	unknown	13	may	98
57	blue-collar	married	primary	52	may	38
60	retired	married	primary	60	may	219

```
#Аналіз DataSet
skim(df)
```

Вихідна таблиця даних DataSet містить 37707 записів та 7 колонок (змінних) (табл. 2). В тому числі 4 змінні – категоріальні (job, marital, education, month), 3 – кількісні (age, balance, duration). Характеристики категоріальних змінних представлені в таблиці 3.

Таблиця 2 – Підсумкові дані по DataSet

Name	df
Number of rows	37707
Number of columns	7
Column type frequency:	
character	4
numeric	3
Group variables	None

Таблиця 3 – Характеристика категоріальних змінних (Variable type: character)

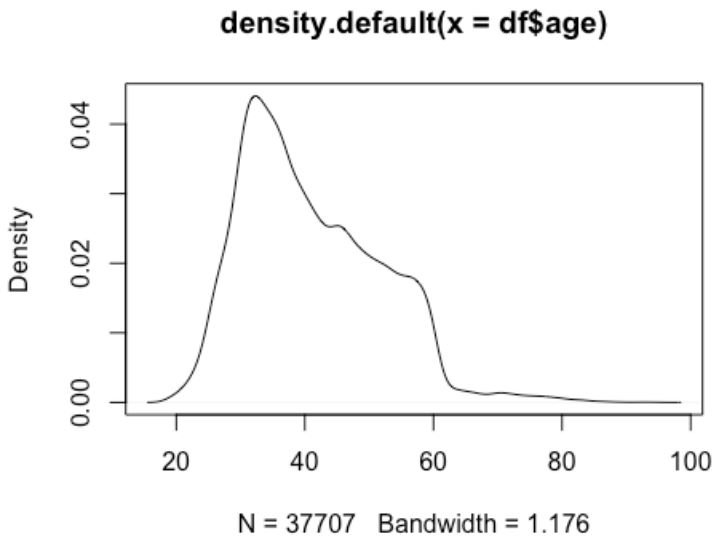
skim_variable	n_missing	complete_rate	empty	n_unique	whitespace
job	0	1	0	12	0
marital	0	1	0	3	0
education	0	1	0	4	0
month	0	1	0	12	0

Враховуючи, що реалізацію мети роботи (порівняльний аналіз обраних методів відновления пропущених значень) планувалось виконувати на кількісних змінних, перед початком генерації пропущених даних і аналізу фактичної DataSet було проаналізовано розподіли змінних.

```
d <- density(df$age)
plot(d)
d <- density(df$balance)
plot(d)
```

Відповідні графіки щільності розподілу змінних “age” і “balance” представлені на рис. 1 і 2. Із рисунків видно, що розподіл змінної “age” близький до нормального, в той час як розподіл змінної “balance” сильно відхиляється від нормального закону. Аналогічне відхилення від нормального закону отримали для змінної “duration”.

Рисунок 1 – графік щільності розподілу змінної “age”



на початку дослідження

Виходячи з цього значення змінних “balance” і “duration” у вихідній DataSet було прологарифмовано і увесь подальший аналіз проводився на перетворених даних:

```
df$balance <- log2(df$balance)
df$duration <- log2(df$balance)
```

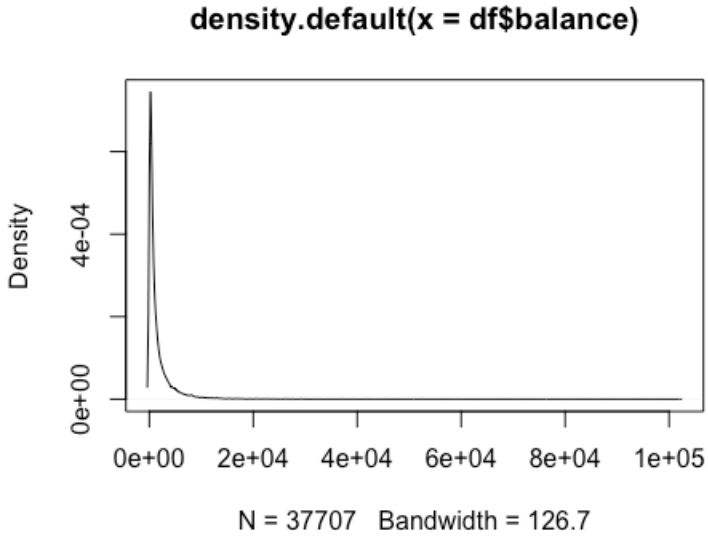


Рисунок 2 – графік щільності розподілу змінної “ balance ” на початку дослідження

Загальні і статистичні характеристики кількісних змінних представлені в таблиці 4.

Таблиця 4 – Характеристика кількісних змінних (Variable type: numeric)

skim_variable	n_missing	complete_rate	mean	sd	p25	p50	p75
age	0	1	41.01	10.75	33.00	39.00	49.00
balance	0	1	9.17	2.37	7.85	9.35	10.80
duration	0	1	3.13	0.49	2.97	3.23	3.43

Як видно із представлених звітів (табл. 3, 4) вихідні дані не містять пропущених значень, тому для оцінки ефективності різних методів заповнення відсутніх даних останні будуть штучно згенеровано. В даній роботі генерація пропусків виконувалась лише для кількісних змінних (”Balance” та “Age”) і типів пропусків MNAR (**missing not at random**) і MCAR (**missing completely at random**). Частка пропусків у таблицях складає приблизно 20% від усіх записів.

Для формування DataSet з невипадковими пропусками в даних (MNAR) використана функція ampute() з пакету mice.

```
#Створюємо DataSet з невипадковими пропусками
MNAR_amp <- ampute(df, mech = "MNAR", prop =0.2 )
## Warning in lapply(X = X, FUN = FUN, ...): NAs introduced by coercion
## Warning in lapply(X = X, FUN = FUN, ...): NAs introduced by coercion
## Warning in lapply(X = X, FUN = FUN, ...): NAs introduced by coercion
## Warning in lapply(X = X, FUN = FUN, ...): NAs introduced by coercion
## Warning: Data is made numeric because the calculation of weights requires
## numeric data
df_spaces_MNAR <- MNAR_amp$amp
```

Аналогічно для DataSet, що буде містити пропуски, отримані **тільки випадковим чином (**MCAR), використовуємо колонки “Balance” та “Age”, а кількість пропусків у таблиці буде приблизно 20% від усіх записів.

```
#Створюємо DataSet з виключно випадковими пропусками
df_spaces_MCAR <- na.omit(df)
```

```
        #Функція для генерації пропусків в даних
miss_generator <- function(column_name){
  probs <- sample(seq(1,100), length(column_name), replace = T)
    for (i in seq(1, length(column_name))) {
    column_name[i] <- ifelse(probs[i] >= 80, NA, column_name[i])
    }
  return (column_name)
}
        #Генерація пропусків у колонці "Balance" для df_spaces_MCAR:
df_spaces_MCAR$balance <- miss_generator(df_spaces_MCAR$balance)
        #Генерація пропусків у колонці "Age" для df_spaces_MCAR:
df_spaces_MCAR$age <- miss_generator(df_spaces_MCAR$age)
```

Побудуємо графіки кількості пропущених даних у DataSet та подивимося на їх розподілення (рис. 3):

```
ggplot(df_spaces_MNAR,
aes(x = age, y = balance)) +  geom_miss_point()
```

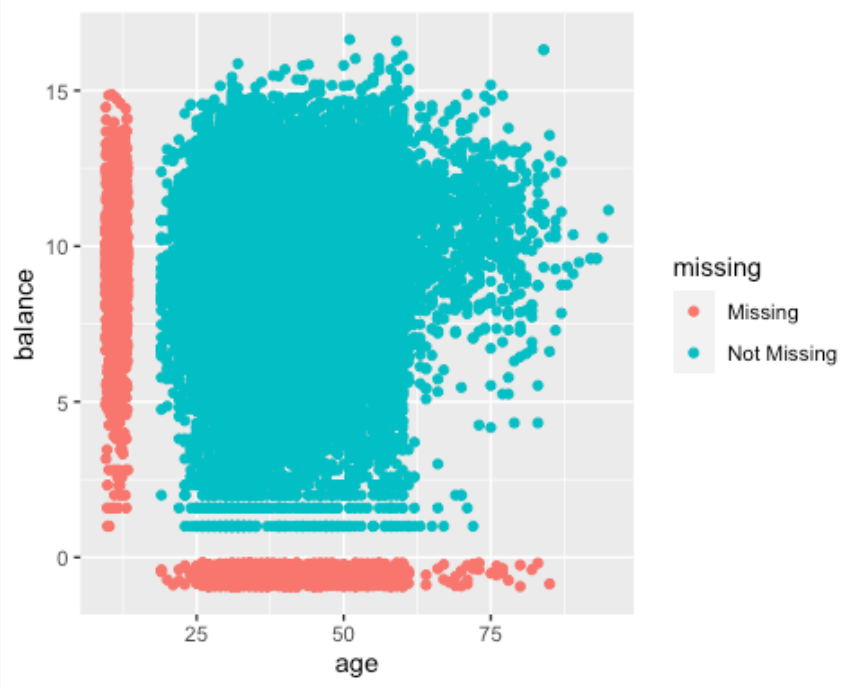


Рисунок 3 – Діаграма розсіювання згенерованих даних DataSet з невипадковим типом пропусків MNAR

Тепер, коли згенеровано DataSet, в яких достатньо пропущених даних, почнемо реалізовувати методи для їх заповнення. Імпутація даних у DataSet буде виконуватися за допомогою функції mice (), яка міститься у пакеті “mice”. На початку перевіримо чи існують у DataSet колінеарні данні, які будуть заважати роботі алгоритму для імпутації пропущених даних.

Колінеарні данні у DataSet MCAR — 0. Колінеарні данні у DataSet MNAR — 0. Отже колінеарних даних у сформованих DataSet немає, що дозволяє реалізовувати визначені у даній роботі алгоритми імпутації даних.

2.2 Імпутація пропусків методом лінійної регресії

Для проведення заповнення пропущених даних за регресійною моделлю у функціях mice() використовуємо метод “norm.nob”.

Спочатку створюємо нову змінну, куди буде записуватися результат виконання функції mice() для обох DataSet.

```
#Виконуємо метод імпутації за регресійною моделлю для типу пропусків MCAR
imp.regress_MCAR <- mice(df_spaces_MCAR, method = "norm.nob")
#Збираємо новий DataSet з імпутованими даними
completed_MCAR_Regression <- complete(imp.regress_MCAR)
#Виконуємо метод імпутації даних MNAR за регресійною моделлю
imp.regress_MNAR <- mice(df_spaces_MNAR, method = "norm.nob")
#Формуємо новий DataSet з імпутованими даними
completed_MNAR_Regression <- complete(imp.regress_MNAR)
```

Перевіряємо чи залишись пропущенні дані у новостворених DataSet.

```
any(is.na(completed_MCAR_Regression))
## [1] FALSE
any(is.na(completed_MNAR_Regression))
## [1] TRUE
```

Тепер проаналізуємо дані, отримані після імпутації пропущених значень для кількісних змінних методом лінійної регресії.

```
skim(completed_MCAR_Regression)
```

```
skim(completed_MNAR_Regression)
```

Основні статистичні характеристики результатів імпутації методом лінійної регресії представлені в табл. 5 і 6.

Таблиця 5 – Статистичні характеристики результатів імпутації пропущених даних типу MCAR методом лінійної регресії

skim_variable	n_missing	complete_rate	mean	sd	p25	p50	p75
age	0	1	41.00	10.73	33.00	39.28	48.39
balance	0	1	9.17	2.38	7.95	9.40	10.76
duration	0	1	3.13	0.49	2.97	3.23	3.43

Таблиця 6 – Статистичні характеристики результатів імпутації пропущених даних типу MNAR методом лінійної регресії

skim_variable	n_missing	complete_rate	mean	sd	p25	p50	p75
age	0	1	40.99	10.75	33.00	39.00	49.00
balance	0	1	9.17	2.38	7.86	9.36	10.80
duration	0	1	3.13	0.50	2.97	3.22	3.43

Як видно з табл. 5, 6, існує тісна залежність між відновленими даними різних типів пропусків.

Подібність розподілу вихідних (фактичних) і заповнених даних добре ілюструють графічні зображення. Для побудови графіків щільності розподілу даних використовуємо функцію densyplot() з пакету mice() (рис. 4, 5). На цих графіках синім кольором позначено вихідні (спостережувані) дані, а червоним – відновлені дані.

```
par(mfrow = c(1, 2))
densityplot(imp.regress_MCAR)
densityplot(imp.regress_MNAR)
```

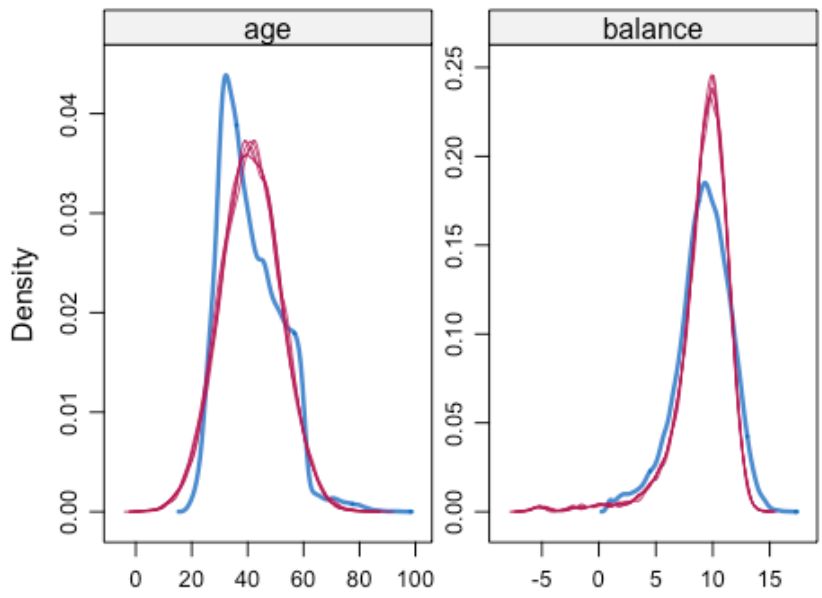


Рисунок 4 – Графік щільності розподілу вихідних (фактичних) і заповнених даних з типом пропусків MCAR методом лінійної регресії

На рис. 6 і 7 також добре видно подібність розсіювання спостережуваних і заповнених даних.

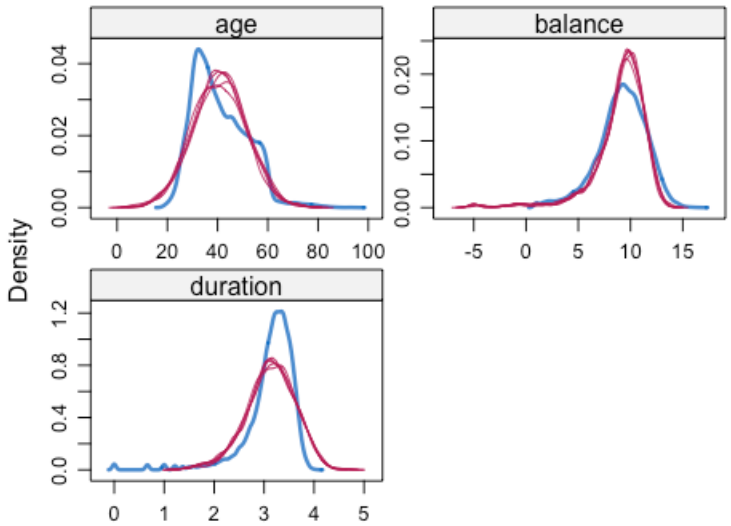


Рисунок 5 – Графік щільності розподілу вихідних (фактичних) і заповнених даних з типом пропусків MNAR методом лінійної регресії

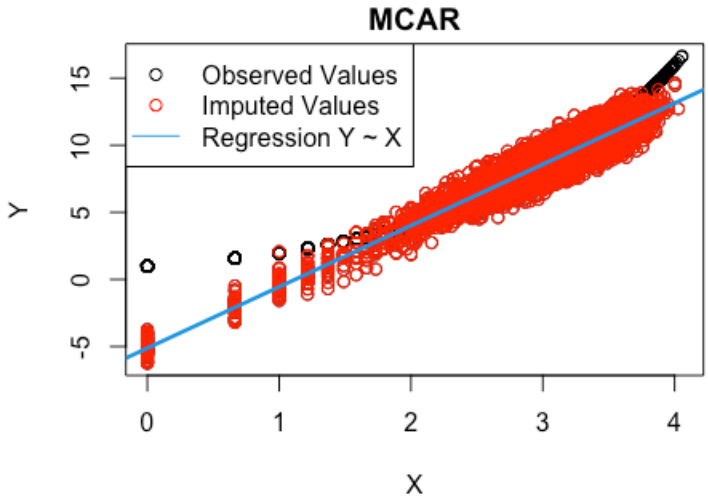


Рисунок 6 – Графік розсіювання спостережуваних і заповнених даних з типом пропусків MCAR методом лінійної регресії: вісь x – значення змінної “duration”, вісь y – значення змінної “balance”.

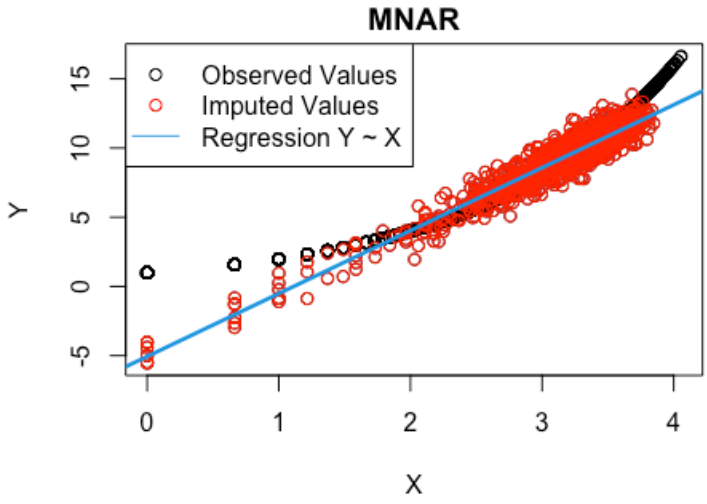


Рисунок 7 – Графік розсіювання спостережуваних і заповнених даних з типом пропусків MNAR методом лінійної регресії: вісь x – значення змінної “duration”, вісь y – значення змінної “balance”.

Далі проведемо такі ж самі операції, тільки с використання інших методів імпутації даних.

2.3 Імпутація пропусків за Баєсівською стохастичною регресією

Наступним буде дуже схожий метод - алгоритм імпутації за Баєсівською стохастичною регресією. У пакеті mice за цей алгоритм відповідає метод “norm”.

```
#Виконуємо Байєсівську стохастичну регресійну імпутацію для пропусків MCAR:
imp.regress_MCAR <- mice(df_spaces_MCAR, method = "norm")
#Збираємо новий DataSet з імпутованими даними
completed_MCAR_Bayesian <- complete(imp.regress_MCAR)
#Виконуємо Байєсівську регресійну імпутацію для пропусків MNAR:
imp.regress_MNAR <- mice(df_spaces_MCAR, method = "norm")
#Збираємо новий DataSet з імпутованими даними
completed_MNAR_Bayesian <- complete(imp.regress_MNAR)
```

Перевіряємо чи залишись пропущенні дані у DataSet.

```
## [1] FALSE
## [1] TRUE
```

Основні статистичні характеристики результатів імпутації методом Байєсівської стохастичної регресії представлені в табл. 7 і 8.

Таблиця 7 – Статистичні характеристики результатів імпутації пропущених даних типу MCAR методом Байєсівської стохастичної регресії

skim_variable	n_missing	complete_rate	mean	sd	p25	p50	p75
age	0	1	40.96	10.71	33.00	39.19	48.11
balance	0	1	9.17	2.38	7.95	9.40	10.76
duration	0	1	3.13	0.49	2.97	3.23	3.43

Таблиця 8 – Статистичні характеристики результатів імпутації пропущених даних типу MNAR методом Байєсівської стохастичної регресії

skim_variable	n_missing	complete_rate	mean	sd	p25	p50	p75
age	0	1	40.99	10.75	33.00	39.00	49.00

skim_variable	n_missing	complete_rate	mean	sd	p25	p50	p75
balance	0	1	9.17	2.37	7.87	9.36	10.80
duration	0	1	3.13	0.50	2.97	3.22	3.43

Як і в попередньому випадку відновлення пропущених даних методом лінійної регресії при застосуванні Байєсівського підходу досягнуто хороших результатів як між відновленими даними різних типів пропусків, так і між фактичними (спостережуваними) і обчисленими даними (рис. 8-11).

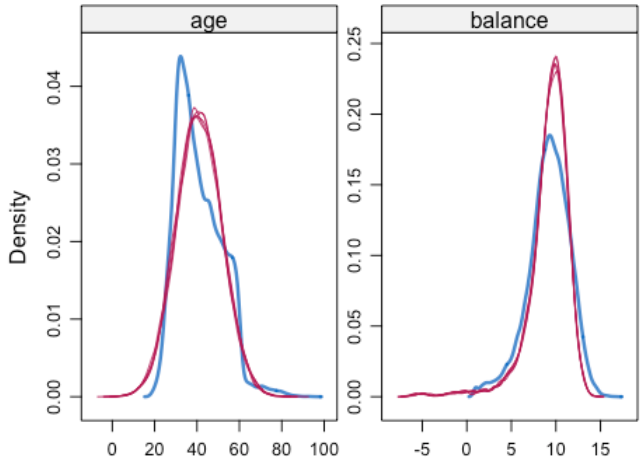


Рисунок 8 – Графік щільності розподілу фактичних і заповнених даних з типом пропусків MCAR методом стохастичної регресії Байєса

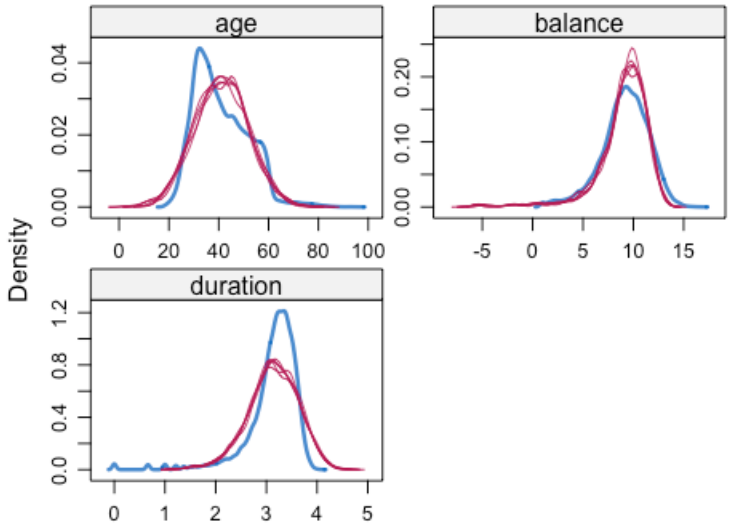


Рисунок 9 – Графік щільності розподілу фактичних і заповнених даних з типом пропусків MNAR методом стохастичної регресії Байєса

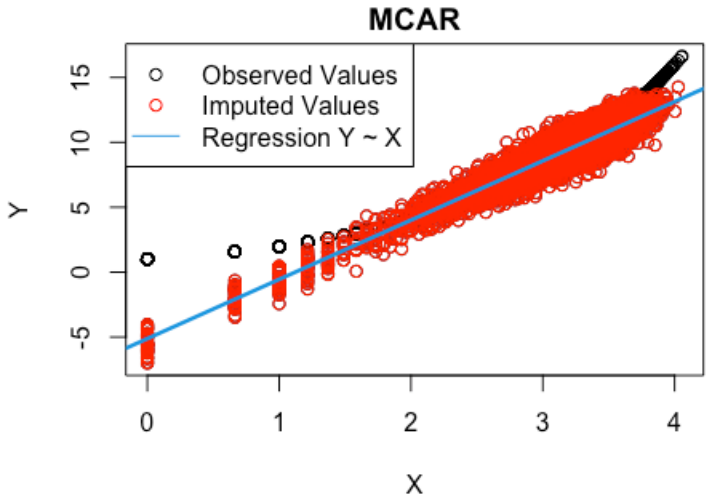


Рисунок 10 – Графік розсіювання спостережуваних і заповнених даних з типом пропусків MCAR методом стохастичної регресії Байєса: вісь x – значення змінної “duration”, вісь y – значення змінної “balance”.

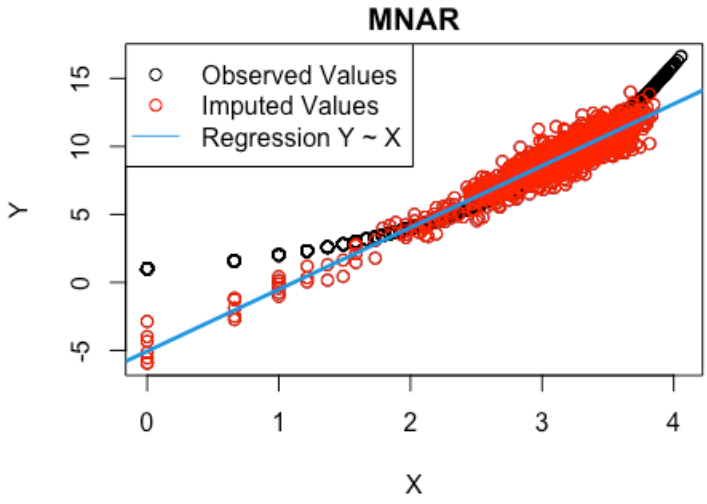


Рисунок 11 – Графік розсіювання спостережуваних і заповнених даних з типом пропусків MNAR методом стохастичної регресії Байєса

2.4 Імпутація пропусків за правилами Рубіна

Для множинних імпутацій за правилами Рубіна, нам потрібно буде декілька разів заповнити пропуски, а потім об’єднати їх за правилами Рубіна. Правила Рубіна призначені для об’єднання оцінок параметрів, таких як середні відмінності, коефіцієнти регресії, стандартні похибки, а також для виведення

довірчих інтервалів і p-значень. Для виконання правил Рубіна, буде використовуватись функція pool з пакету mice, з аргументом “rubin1987”.

```
imp_MCAR <- mice(df_spaces_MCAR, m=20, maxit=20)
completed_Rubin <- complete(imp_MCAR)
skim(completed_Rubin)
fit <- with(data = imp_MCAR, exp = lm(duration ~ balance+age))
fit$analyses
## [[1]]
##
## Call:
## lm(formula = duration ~ balance + age)
##
## Coefficients:
## (Intercept)      balance          age
##   1.3509249    0.1982255   -0.0009476
##
##
## [[2]]
##
## Call:
## lm(formula = duration ~ balance + age)
##
## Coefficients:
## (Intercept)      balance          age
##    1.363233     0.198234    -0.001263
##
##
## [[3]]
##
## Call:
## lm(formula = duration ~ balance + age)
##
## Coefficients:
## (Intercept)      balance          age
##   1.3487467    0.1979982   -0.0008496
##
##
## [[4]]
##
## Call:
## lm(formula = duration ~ balance + age)
##
```

```
## Coefficients:
## (Intercept)      balance          age
##   1.3484694    0.1981812   -0.0008803
##
##
## [[5]]
##
## Call:
## lm(formula = duration ~ balance + age)
##
## Coefficients:
## (Intercept)      balance          age
##    1.351703     0.198373    -0.000999
##
##
## [[6]]
##
## Call:
## lm(formula = duration ~ balance + age)
##
## Coefficients:
## (Intercept)      balance          age
##    1.358803     0.198325    -0.001162
##
##
## [[7]]
##
## Call:
## lm(formula = duration ~ balance + age)
##
## Coefficients:
## (Intercept)      balance          age
##   1.3493066    0.1981245   -0.0008929
##
##
## [[8]]
##
## Call:
## lm(formula = duration ~ balance + age)
##
## Coefficients:
## (Intercept)      balance          age
##    1.359496     0.198201    -0.001159
##
##
```

```
## [[9]]
##
## Call:
## lm(formula = duration ~ balance + age)
##
## Coefficients:
## (Intercept)      balance          age
##     1.35577      0.19854     -0.00113
##
##
## [[10]]
##
## Call:
## lm(formula = duration ~ balance + age)
##
## Coefficients:
## (Intercept)      balance          age
##    1.355520     0.198318    -0.001086
##
##
## [[11]]
##
## Call:
## lm(formula = duration ~ balance + age)
##
## Coefficients:
## (Intercept)      balance          age
##    1.352943     0.198368    -0.001025
##
##
## [[12]]
##
## Call:
## lm(formula = duration ~ balance + age)
##
## Coefficients:
## (Intercept)      balance          age
##    1.361203     0.198293    -0.001212
##
##
## [[13]]
##
## Call:
## lm(formula = duration ~ balance + age)
##
```

```
## Coefficients:
## (Intercept)      balance          age
##    1.365093     0.198466    -0.001349
##
##
## [[14]]
##
## Call:
## lm(formula = duration ~ balance + age)
##
## Coefficients:
## (Intercept)      balance          age
##    1.360469     0.198381    -0.001218
##
##
## [[15]]
##
## Call:
## lm(formula = duration ~ balance + age)
##
## Coefficients:
## (Intercept)      balance          age
##    1.356722     0.198234    -0.001096
##
##
## [[16]]
##
## Call:
## lm(formula = duration ~ balance + age)
##
## Coefficients:
## (Intercept)      balance          age
##    1.361683     0.198408    -0.001251
##
##
## [[17]]
##
## Call:
## lm(formula = duration ~ balance + age)
##
## Coefficients:
## (Intercept)      balance          age
##     1.35120      0.19848     -0.00101
##
##
```

```
## [[18]]
##
## Call:
## lm(formula = duration ~ balance + age)
##
## Coefficients:
## (Intercept)      balance          age
##    1.357619     0.198235    -0.001117
##
##
## [[19]]
##
## Call:
## lm(formula = duration ~ balance + age)
##
## Coefficients:
## (Intercept)      balance          age
##     1.35941      0.19824     -0.00116
##
##
## [[20]]
##
## Call:
## lm(formula = duration ~ balance + age)
##
## Coefficients:
## (Intercept)      balance          age
##    1.357141     0.198351    -0.001128
```

```
pool_MCAR <- pool(fit, rule = "rubin1987")
my_table <- flextable(summary(pool_MCAR))
```

Основні статистичні характеристики результатів імпутації пропущених даних за правилами Рубіна представлені в табл. 9-10.

Таблиця 9 – Статистичні характеристики результатів імпутації пропущених даних типу MCAR за правилами Рубіна

skim_variable	n_missing	complete_rate	mean	sd	p25	p50	p75
age	0	1	40.97	10.75	32.00	39.00	49.00
balance	0	1	9.17	2.37	7.86	9.35	10.80
duration	0	1	3.13	0.49	2.97	3.23	3.43

Таблиця 10 – Характеристики якості відновлення пропущених даних типу MCAR за правилами Рубіна

term	estimate	std.error	statistic	df	p.value
(Intercept)	1.3563	0.0067	203.93	52.64	0.0000000000
balance	0.1983	0.0004	534. 90	1,104.8	0.0000000000
age	-0.0011	0.0002	-6.90	31.92	0.0000000836

Теж саме виконуємо для DataSet за пропусками типу MNAR.

```
imp_MNAR <- mice(df_spaces_MNAR, m=20, maxit=20)
completed_Rubin <- complete(imp_MNAR)
skim(completed_Rubin)
        fit <- with(data = imp_MNAR, exp = lm(balance ~ age + duration))
pool_MNAR <- pool(fit, rule = "rubin1987")
my_table <- flextable(summary(pool_MNAR))
#Налаштовуємо зовнішній вигляд таблиці
my_table <- my_table %>%
        #Коригуємо розмір і виводимо таблицю результатів
  autofit() %>%
  my_table
```

Таблиця 11 – Статистичні характеристики результатів імпутації пропущених даних типу MNAR за правилами Рубіна

skim_variable	n_missing	complete_rate	mean	sd	p25	p50	p75
age	0	1	41.00	10.75	33.00	39.00	49.00
balance	0	1	9.17	2.37	7.86	9.35	10.80
duration	0	1	3.13	0.49	2.97	3.22	3.43

Таблиця 11 – Характеристики якості відновлення пропущених даних типу MNAR за правилами Рубіна

term	estimate	std.error	statistic	df	p.value
(Intercept)	-5.2622	0.0279	-188.58	29,78	0.0000000000
age	0.0062	0.0004	16.72	6,401.6	0.0000000000
duration	4.5296	0.0079	572.78	37,29	0.0000000000

Побудуємо таблиці з помилками наближення відновлених даних до спостережуваних при об’єднанні за правилами Рубіна (табл. 12, 13):

```
miceMCError <- function(pooledRes) {
 monteCarloSE <- sqrt(pooledRes$pooled$b/pooledRes$m)
 ciLower <- pooledRes$pooled$estimate -
qt(0.975,df=pooledRes$m-1)*monteCarloSE
 ciUpper <- pooledRes$pooled$estimate +
qt(0.975,df=pooledRes$m-1)*monteCarloSE
 mcTable <- cbind(pooledRes$pooled$estimate, monteCarloSE, ciLower, ciUpper)
 colnames(mcTable) <- c("Estimate", "Monte Carlo SE", "95% CI lower limit",
"95% CI upper limit")
 return(mcTable)
 print("Warning: 95% CI only quantifies Monte-Carlo uncertainty!")
}
table_errors_MNAR <-miceMCError(pool_MNAR)
```

Таблиця 12 – Характеристики якості відновлення пропущених даних типу MNAR при об’єднанні масивів даних за правилами Рубіна

Term	Estimate	Monte Carlo SE	95% CI lower limit	95% CI upper limit
(Intercept)	-5.26221603	6.48e-04	-5.2635166	-5.2608039
age	0.0062333	1.81e-05	0.0061954	0.0062711
duration	4.5295954	7.93e-05	4.5294293	4.5297615

```
table_errors_MCAR <-miceMCError(pool_MNAR)
```

Таблиця 13 – Характеристики якості відновлення пропущених даних типу MCAR при об’єднанні масивів даних за правилами Рубіна

Term	Estimate	Monte Carlo SE	95% CI lower limit	95% CI upper limit
(Intercept)	1.3604218	0.0009017	1.3585345	1.3623092
age	0.1984089	0.0000295	0.1983472	0.1984705
duration	-0.0012239	0.0000240	-0.0012742	-0.0011736

Аналіз помилок наближення відновлених пропущених даних за правилами Рубіна до спостережуваних даних, представлений в табл. 10-13, свідчить про відмінну якість імпутації за всіма змінними, що вивчались.

ВИСНОВКИ

У проведеному дослідженні виконано порівняння ефективності 3 підходів до заповнення пропусків у кількісних змінних: регресійна модель імпутації даних, Байєсівська стохастична регресійна імпутація, множинна імпутація пропущених даних за методикою Рубіна. Практична реалізація методів відновлення пропусків здійснена на масиві даних клієнтів певного банку, який містив 37707 записів, з генерацією 20% пропущених значень типів MCAR і MNAR у кількісних змінних за допомогою програмної мови R пакетів для роботи з пропущеними даними: mice та naniar.

Встановлено, що усі застосовані методи імпутації показують якісне відновлення пропущених значень у кількісних змінних ($p<0,001$) за умов нормального розподілу спостережуваних даних, відсутності мультиколінеарності, застосування моделей множинної лінійної регресії. Крім того, отриманий відмінний ефект, на нашу думку, можливо був пов’язаний з великою вибіркою даних.

З урахуванням даних наукової літератури та власного дослідження про переваги і недоліки у застосуванні кожного методу імпутації вважаємо, що для відновлення відсутніх значень кількісних (безперервних) змінних типу MNAR найкращим підходом буде використання байєсівської стохастичної регресійної моделі.

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ

1. Rubin D. B. Inference and Missing Data // Biometrika. – 1976. – N 63. – P. 581–592.

2. Rubin D.B. Multiple Imputation for Nonresponse in Surveys. – John Wiley & Sons Inc., New York, 1987. – 258 p.

3. Little R. J. A., Rubin D. B. Statistical Analysis with Missing Data, – Wiley, 2nd ed., 2002. – 389 p.

4. Pigott T.D. A Review of Methods for Missing Data // Educational Research and Evaluation. – 2001. – Vol. 7, N 4. – P. 353–383.

5. Камінський Р. М. Відновлення пропусків у результатах тестування та ідентифікації операторського персоналу / Камінський Р. М., Кунанець Н. Е., Пасічник В. В., Худий А. М. // Вісник Національного університету «Львівська політехніка» «Інформаційні системи та мережі», 2018, Випуск 887. – С. 92 – 104.

6. Міщук О. С. Методи оброблення та заповнення пропущених параметрів у даних екологічного моніторингу / О. С. Міщук, Р. О. Ткаченко // Науковий вісник НЛТУ України. - 2019. - Т. 29, № 6. - С. 119-122.

7. Newman D.. Missing Data: Five Practical Guidelines // Organizational Research Methods. – 2014. – Vol.17 (4). – P. 372–411.

8. Graham J. W. How Many Imputations Are Really Needed? Some Practical Clarifications of Multiple Imputation Theory / Graham J. W., A. E. Olchowski, T. D. Gilreath // Preventive Science. – 2007. – Vol. 8 (3). – P. 206–213.

9. Guan N.C., Yusoff M.S.B. Missing values in data analysis: Ignore or Impute? // Education in Medicine Journal. – 2011. – Vol. 3 (1). P. e6-e11.

10. Carpenter J., Kenward M. Multiple imputation: current perspectives // Statistical Methods in Medical Research. – 2007. - Vol. 16 (3). – P. 199-218.

11. Zhang P. Multiple imputation: theory and method // International Statistical Review. – 2003. – Vol. 71 (3). – P. 581-592.

12. Honaker J. Analyzing Incomplete Political Science Data: An Alternative Algorithm for Multiple Imputation / Honaker J., Joseph A., King G., Scheve K. // The American Political Science Review. – 2001. – Vol. 95 (1). – P. 49-69.

13. Eekhout I. Missing Data: a Systematic Review of How They Are Reported and Handled / Eekhout I., de Boer R. Michiel, Twisk, Jos W. R., de Vet Henrica C. W., Heymans Martijn W. //Epidemiology: September 2012. – Vol. 23, Issue 5. – P. 729-732.

14. Schlomer G. L. Best practices for missing data management in counseling psychology / Gabriel L. Schlomer, Sheri Bauman, Noel A. Card // Journal of Counseling Psychology. – 2010. – Vol. 57 (1). – P. 1–10.

15. Soley-Bori M. Dealing with missing data: key assumptions and methods for applied analysis // Technical Report. – 2013. – N 4. – P. 1–20.

16. Антоненко С. В. Методи поповнення пропусків даних гідрологічного моніторингу / Антоненко С. В., Земляний О. Д., Ізмайлова М. К. //Актуальні проблеми автоматизації та інформаційних технологій.- 2020. –Т. 24. – С. 3-16.

17. Мацуга О.М., Шеремет В.С. Кластеризація даних з пропусками методом К-середніх //Актуальні проблеми автоматизації та інформаційних технологій.- 2019. – Т. 23. – С. 69-78.

18. Злоба Е., Яцкив И. Статистические методы восстановления пропущенных данных // Computer Modelling & New Technologies. − 2002. − Vol. 6. − № 1. − P. 51-61.

19. Бідюк П.І. Моделі і методи прикладної статистики / П.І. Бідюк, Л.О. Коршевнюк, Н.В. Кузнєцова . – К.: НУТУ «КПІ», 2014. – 722 с.

20. Sefidian A. M., Daneshpour N. Estimating missing data using novel correlation maximization based methods // Applied Soft Computing. 2020. – Vol. 91. DOI: 10.1016/j.asoc.2020.106249.

21. Gelman A. Bayesian Data Analysis / Gelman Andrew, Carlin John B, Stern Hal S., Dunson David B., Vehtari Aki, Rubin Donald B. – 2013. - 675p.

22. Shi F. Missing Value Estimation for Microarray Data by Bayesian Principal Component Analysis and Iterative Local Least Squares / F. Shi, D. Zhang, J. Chen, H.R. Karimi // Mathematical Problems in Engineering. Article ID 162938. – 2013. – P. 17.

23. Huque M.H. A Comparison of Multiple Imputation Methods for Missing data in Longitudinal Studies / Huque, M.H., Carlin, J.B., Simpson, J.A. and Lee, K.J. //BMC Medical Research Methodology, 2018. – N18.– P. 1-16.

24. Bodner T.E. What Improves with Increased Missing Data Imputations? // Structural Equation Modeling, 2008. – Vol. 15 (4). – P. 651–675.

25. White I. R. Multiple imputation using chained equations: Issues and guidance for practice / White I. R., P. Royston, A. M. Wood // Stat Med. – 2011. – Vol. 30 (4). – P. 377–399.

26. Van Buuren S. Flexible Imputation of Missing Data. Second Ed. Boca Raton, FL: Chapman & Hall/CRC, 2018. – 444 p.

27. Alruhaymi Abdullah Z., Kim Charles J. Why Can Multiple Imputations and How (MICE) Algorithm Work? // Open Journal of Statistics. – 2021. – Vol.11 (5). – P. 759-777. DOI: 10.4236/ojs.2021.115045.

28. Reiter J. P., Raghunathan T. E. The Multiple Adaptations of Multiple Imputation // Journal of the American Statistical Association. – 2007. – Vol. 102. – P. 1462–1470.

29. Raghunathan T. E. A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models // Raghunathan, T. E., Lepkowski, J. M., Hoewyk, J. V. //Survey Methodology. – 2001. – Vol. 27. – P. 85–95.