**МЕТОД ГОЛОВНИХ КОМПОНЕНТ: ОЦІНЮВАННЯ РИЗИКІВ ДЛЯ ПАЦІЄНТІВ ХВОРИХ НА СЕРЦЕВО-СУДИННІ ЗАХВОРЮВАННЯ**

**Рудзей Вікторія Олександрівна**

## Київський національний університет імені Тараса Шевченка

VI курс, спеціальність «Економічна аналітика та статистика»

На сьогодні серцево-судинні захворювання є основною причиною смерті в усьому світі. За даними Всесвітньої організації охорони здоров’я [1], приблизно 17,9 мільйонів людей щороку помирають від серцево-судинних захворювань, що становить 31% усіх смертей у світі.

Серцево-судинні захворювання також є основним економічним тягарем для систем охорони здоров’я. Прямі витрати на лікування серцево-судинних захворювань, включаючи госпіталізацію, медикаменти та операції, є досить значними.

Для зменшення рівня серцево-судинних захворювань можна застосовувати два основних підходи: профілактику та діагностику. Профілактика включає різні заходи, призначені для попередження розвитку захворювань у здорових людей. Надання допомоги людям, які мають ризик розвитку серцево-судинних захворювань, а також регулярний моніторинг стану здоров'я, дозволяють виявляти патологію на ранніх стадіях і раніше почати лікування. Також для досягнення позитивних результатів важливо забезпечити доступ до професійної медичної допомоги та заходів по підвищенню рівня освіти та свідомості населення щодо медичних питань. Статистичні дослідження допоможуть визначити потенційні ризики для розвитку серцево-судинних захворювань та виявити чинники, що впливають на їх розвиток. На основі цих досліджень можна розробляти програми профілактики і діагностики, що забезпечать захист здоров'я населення. Також вони дозволять оцінити ефективність вже існуючих програм профілактики та діагностики.

Статистична інформація щодо факторів ризику та виживаності пацієнтів включає досить велику кількість складових, що ускладнює виявлення основних причин захворюваності. Отже, дуже важливо попередньо обмежити кількість факторів і виявити найсуттєвіші та приховані закономірності. Саме це завдання і визначило вибір статистичного методу - методу головних компонент, який використовується для аналізу взаємозв’язків між великою кількістю змінних і пояснення цих взаємозв’язків з точки зору меншої кількості змінних (головних компонент), з мінімальною втратою інформації [2]. Використання цього методу дає змогу вивчити та пояснити взаємозв’язки між досліджуваними показниками, після чого можна визначити значущі групи факторів для подальшого аналізу.

Метод головних компонент є популярним методом обробки даних, який застосовується для скорочення розмірності вхідних даних з мінімальним втратами корисної інформації. Цей метод зводиться до пошуку додаткової системи прямих або площини у просторі змінних, що максимізують дисперсію вихідних значень. Такі прямі або площини називаються головними компонентами. Головною перевагою методу головних компонент є те, що він єдиний математично обґрунтований метод факторного аналізу [3]. К. Пірсон був першим, хто розробив цей метод, після чого він був розвинений, доопрацьований, описаний та обґрунтований в роботах таких вчених, як Г. Хотеллінг, Г. Харман, С. Рао, П. Андрукович, С. А. Айвазян та В.С. Мхитарян.

Розрахунки проводились в програмі SPSS (Statistical Package for the Social Sciences), яка використовується дослідниками в різних сферах для аналізу складних даних.

Статистична інформація для проведення дослідження була зібрана у госпіталізованих пацієнтів для прогнозування смертності [4]. Для цього набору даних було проведено факторний аналіз, а саме використано метод головних компонент задля визначення груп значущих факторів, які надалі будуть використані для аналізу виживаності. Щоб уникнути впливу розмірності та одиниць виміру всі показники були попередньо стандартизовані.

Всі подальші розрахунки було проведено за допомогою відповідної надбудови програми SPSS (Analyze > Dimension Reduction > Factor Analysis).

Важливим кроком у виконанні факторного аналізу є обчислення кореляційної матриці для визначення тісноти зв’язку між обраними 27 змінних за всією сукупністю пацієнтів та в розрізі пацієнтів, які вижили чи померли.

Щоб визначити кількість головних компонент та частку впливу кожного фактору було розраховано таблицю факторизації моделі для усієї вибірки пацієнтів (табл.1). Було сформовано лише 10 груп факторів, які пояснюють 71,4% загальної дисперсії.

Таблиця 1

Пояснення загальної дисперсії для усіх пацієнтів

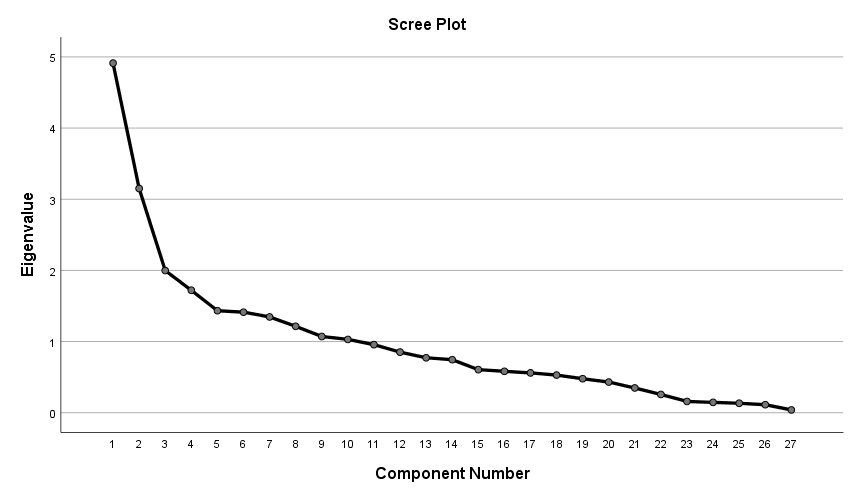
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Total Variance Explained** | | | | | | |
| Component | Initial Eigenvalues | | | Extraction Sums of Squared Loadings | | |
| Total | % of Variance | Cumulative % | Total | % of Variance | Cumulative % |
| 1 | 4.913 | 18.195 | 18.195 | 4.913 | 18.195 | 18.195 |
| 2 | 3.151 | 11.670 | 29.865 | 3.151 | 11.670 | 29.865 |
| 3 | 1.998 | 7.400 | 37.265 | 1.998 | 7.400 | 37.265 |
| 4 | 1.721 | 6.373 | 43.638 | 1.721 | 6.373 | 43.638 |
| 5 | 1.434 | 5.309 | 48.947 | 1.434 | 5.309 | 48.947 |
| 6 | 1.413 | 5.233 | 54.181 | 1.413 | 5.233 | 54.181 |
| 7 | 1.346 | 4.986 | 59.167 | 1.346 | 4.986 | 59.167 |
| 8 | 1.215 | 4.499 | 63.666 | 1.215 | 4.499 | 63.666 |
| 9 | 1.071 | 3.968 | 67.634 | 1.071 | 3.968 | 67.634 |
| 10 | 1.030 | 3.816 | 71.450 | 1.030 | 3.816 | 71.450 |
| Extraction Method: Principal Component Analysis. | | | | | | |

*Джерело: рохраховано автором на основі даних* [4]

Модель обмежує всю кількість досліджуваних змінних до десяти головних компонент, які пояснюють 71,4% загальної дисперсії. За допомогою графіку власних значення можна уточнити результати факторизації моделі і кількості компонентів, які варто розглядати (рис.1), в нашому випадку різкий кам’янистий осип починається після 3 компоненти, які загалом пояснюють лише 37,2% загальної дисперсії (табл.1).

Рисунок 1

Графік власних значень для усіх пацієнтів



Для визначення можливості використання факторного аналізу були використані два критерії: КМО (Кайзера-Мейєра-Олкіна) та критерій сферичності Бартлетта. Значення цих критеріїв підтвердили придатність факторного аналізу для подальшого використання [5]. Результат критерію КМО 0.709, що інтерпретується як прийнятна адекватність. Рівень значущості критерію сферичності Бартлетта (p<0,05) вказує на те, що дані цілком прийнятні для факторного аналізу (табл.2).

Таблиця 2

Критерій КМО та критерій сферичності Бартлетта для усіх пацієнтів

|  |  |  |
| --- | --- | --- |
| **KMO and Bartlett's Test** | | |
| Kaiser-Meyer-Olkin Measure of Sampling Adequacy. | | .709 |
| Bartlett's Test of Sphericity | Approx. Chi-Square | 223395.893 |
| df | 351 |
| Sig. | .000 |

*Джерело: розраховано автором на основі даних*

Для отримання головних компонент з набору спостережень необхідно проаналізувати перевернуту матрицю компонентів. Метод головних компонент дає змогу отримати матрицю факторних навантажень, де коефіцієнти aij можуть бути використані для чисельного та формального пояснення коефіцієнтів кореляції. Це дозволяє зробити припущення про наявність фактору, який може бути причиною цих кореляцій. Щоб пояснити відібрані фактори потрібно в кожному рядку повернутої факторної матриці відмітити те факторне навантаження, яке має найбільше абсолютне значення. Факторні навантаження можна розглядати як коефіцієнти кореляції між змінними та факторами.

Відповідно до попередніх розрахунків для подальшого аналізу було обрано перші три, найбільш значущих компоненти. До першої компоненти увійшло 7 факторів, які відповідають за життєво-важливі ознаки та пояснюють 18,1% загальної дисперсії. Умовно ми її можемо назвати «основні причинно-наслідкові фактори серцево-судинних захворювань».

Перша компонента містить в собі такі змінні:

* Середній артеріальний тиск пацієнта протягом першої години перебування у відділенні (0,905)
* Середній артеріальний тиск пацієнта протягом перших 24 годин перебування у відділенні (0,901)
* Середній діастолічний артеріальний тиск пацієнта протягом першої години перебування в відділенні (0,833)
* Середній систолічний артеріальний тиск пацієнта протягом першої години перебування у відділенні (0,828)
* Середній діастолічний артеріальний тиск пацієнта протягом перших 24 годин перебування в відділенні (0,822)
* Середній систолічний артеріальний тиск пацієнта протягом перших 24 годин перебування у відділенні (0,818)
* Бал APACHE III для середнього артеріальний тиск, виміряного протягом перших 24 годин (0,652) .

Другу компоненту відповідно до факторів, які до неї увійшли, можемо назвати «показники пульсу». Вона пояснює 11,6% загальної дисперсії та включає такі змінні:

* Середній показник пульсу пацієнта протягом перших 24 годин перебування у відділенні (0,929)
* Середній показник пульсу пацієнта протягом першої години перебування у відділенні (0,902)
* Бал APACHE III для середньої частоти пульсу, виміряного протягом перших 24 годин (0,896).

Третя компонента - «показники частоти дихання», які пояснюють 7,4% загальної дисперсії та включає такі змінні:

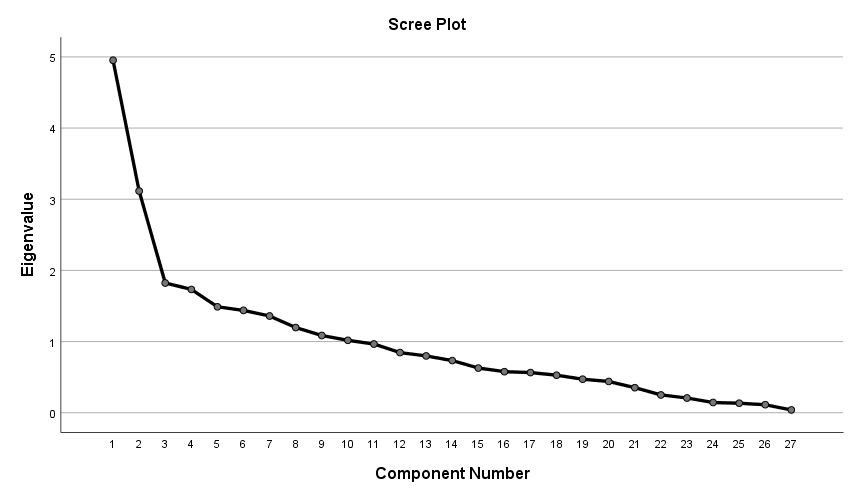
* Середня частота дихання пацієнта протягом перших 24 годин перебування у відділенні (0,875)
* Бал APACHE III для середньої частоти дихання, виміряного протягом перших 24 годин (0,832)
* Середня частота дихання пацієнта протягом першої години перебування у відділенні (0,771).

Подальше дослідження було продовжено аналізом групи пацієнтів, які вижили, використовуючи ті ж самі 27 змінних. Факторизація моделі у даній вибірці пацієнтів дала результати, аналогічні до всієї вибірки - 10 головних компонент, які пояснюють 71,1% загальної дисперсії. Результат КМО склав 0.710, що свідчить про прийнятну адекватність. Рівень значущості критерію сферичності Бартлетта становить p<0,05.

За допомогою графіку власних значень (рис.2) було прийняте рішення щодо кикористання, як і в попередньому випадку, трьох компонент, які пояснюють лише 36,6% загальної дисперсії.

Рисунок 2

Графік власних значень для пацієнтів, котрі вижили



Аналіз оберненої матриці компонентів для пацієнтів, котрі вижили, показав, що перші 3 компоненти містять в собі аналогічні змінні змінні, що і для всієї вибірки пацієнтів. Перша компонента пояснює 18,3% загальної дисперсії, друга - 11,5%, а третя - 6,7%.

Подальше дослідження було продовжено аналогічним аналізом для групи пацієнтів, які померли. Факторизація моделі у цій вибірці пацієнтів показала лише 9 факторів, які пояснюють 67,4% загальної дисперсії. Критерій КМО дав результат 0.689, що свідчить про прийнятну адекватність. Рівень значущості критерію сферичності Бартлетта склав p<0,05. Графік власних значень дозволив встановити, що потрібно враховувати 3 компоненти, які відображають лише 36,5% загальної дисперсії. (див.рис.3).

Рисунок 3

Графік власних значень для пацієнтів, котрі померли



Аналіз оберненої матриці компонентів для пацієнтів, які померли, показав ідентичність факторів, що були включені в перші три компоненти, що й для групи усіх пацієнтів та групи пацієнтів, котрі вижили.

Отже, проведене дослідження з використанням методу головних компонент для бази даних пацієнтів, які мають серцево-судинні захворювання, дозволило визначити взаємозв’язки між досліджуваними показниками та обрати 3 групи факторів (головні компоненти), а саме: «основні причинно-наслідкові фактори», «фактори показників пульсу», «фактори частоти дихання». Це дозволило скоротити кількість факторів та визначити найбільш значущі з усіх досліджених. Враховуючи те, що групи значущих факторів містять у собі однаковий набір змінних, можна зробити висновки, що поділ пацієнтів на різні групи в подальшому дослідженні не має сенсу і для подальшого аналізу виживаності варто розглядати всю сукупність пацієнтів загалом.

Визначення значущих факторів допоможе в подальшому аналізі виживаності пацієнті, і в результаті дозволить більш ефективно проводити лікування серцево-судинних захворювань або навіть запобігти їх появу на ранніх етапах.

**Література**

1. World Health Organization [Електронний ресурс] // World Health Organization: [сайт]. - Режим доступу: <https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1>
2. Чугаєвська, С. В., Ковтун, Н. В. Основи статистичного моделювання. Вид-во ЖДУ ім. І. Франка // Житомир – «Рута». - 2022
3. Т. С. Клебанова, Л. С. Гур’янова Бізнес-аналітика багатовимірних процесів // Мультимедійний навчальний посібник. - 2020 – Р. 8.4
4. Kaggle: Patient Survival Prediction [Електронний ресурс] // Patient Survival Prediction: [сайт]. - Режим доступу: [https://www.kaggle.com/datasets/mitishaagarwal/patient/](https://www.kaggle.com/datasets/mitishaagarwal/patient/versions/3?resource=download)
5. Громова Г.М. Інструменти вимірювання толерантності до невизначеності // Наукові студії із соціальної та політичної психології Вип. 47 (50)