МІНІСТЕРСТВО ВНУТРІШНІХ СПРАВ УКРАЇНИ
ХАРКІВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
ВНУТРІШНІХ СПРАВ

Факультет № 6
Кафедра соціології та психології

ТЕКСТ ЛЕКЦІЇ

з навчальної дисципліни
«**Комп'ютерні методи практичної психології**»
обов'язкових компонент освітньої програми
першого (бакалаврського) рівня вищої освіти

053 Психологія (практична психологія)

Тема №11. Аналіз психологічних даних в SPSS: дискримінантний аналіз, багатовимірне шкалювання, кластерний аналіз

Харків 2023

ЗАТВЕРДЖЕНО
Науково-методичною радою
Харківського національного
університету внутрішніх справ
Протокол від 30.08.2023 р. №7

СХВАЛЕНО
Вченою радою факультету №6
Протокол від 25.08.2023 р. №7

ПОГОДЖЕНО
Секцією Науково-методичної
ради ХНУВС з гуманітарних та
соціально-економічних дисциплін
Протокол від 29.08.2023 р. №7

Розглянуто на засіданні кафедри соціології та психології
Протокол від 15.08.2023 р. №8

Розробник:
Професор кафедри соціології та психології факультету №6
д-р соціол. н., професор Нечітайло Ірина Сергіївна

Рецензенти:
1. Керівник психологічної служби Харківського гуманітарного університету «Народна українська академія», доцент кафедри соціології та гуманітарних дисциплін, к. психол. н., Гога Н. П.;
2. Доцент кафедри соціології та психології факультету №6, к. психол. н., доцент Філоненко В. М.

ТЕМА №11. АНАЛІЗ ПСИХОЛОГІЧНИХ ДАНИХ В SPSS: ДИСКРИМІНАНТНИЙ АНАЛІЗ, БАГАТОВИМІРНЕ ШКАЛЮВАННЯ, КЛАСТЕРНИЙ АНАЛІЗ

План

11.1. Основні поняття та етапи дискримінантного аналізу.
11.2. Багатовимірне шкалювання.
11.3. Порівняння кластерного і факторного аналізів. Основні поняття та етапи кластерного аналізу.

Рекомендована література

Основна

1. Нечітайло І. С., Бірюкова М. В. Математичні методи в соціології : підручник для студентів ВНЗ / Нар. укр. акад., [каф. соціології]. Харків : Вид-во НУА, 2012. 243 с.

2. Татьянчиков А. О. Математичні методи в психології: навчально-методичні рекомендації (в допомогу до самостійної роботи для здобувачів вищої освіти ступеня бакалавра факультету психології, політології та соціології) ; кафедра психології НУ «Одеська юридична академія». Одеса : Фенікс, 2021. 48 с.

Допоміжна

3. Катаєв Є.С. Використання статистичних методів обробки даних у дослідженнях “я-концепції” особистості. Вісник Національного університету оборони України. 2012. №2 (27) /2012. С. 171-176.

4. Салюк М. А. Статистична обробка даних експериментального дослідження. Методичний посібник з курсу «Експериментальна психологія» / за ред. Е.Л. Носенко. Дніпропетровськ: Інновація, 2010. 26 с.

5. Старушенко Г. А. Статистична обробка даних в системі публічного управління : навч. посіб. Дніпро : ГРАНІ, 2018. 144 с.

6. Татьянчиков А. О. Методичні рекомендації до виконання лабораторних робіт з курсу «Методи психологічного дослідження: математичні методи в психології». Одеса : Вид-во Університету Ушинського, 2019. 38 с.

ТЕКСТ ЛЕКЦІЇ

11.1. Основні поняття та етапи дискримінантного аналізу

Дискримінантний аналіз дозволяє передбачити приналежність об’єктів до двох або більше груп, що перетинаються. Вихідними даними для дискримінантного аналізу є безліч об’єктів, розділених на групи так, що кожен об’єкт може бути віднесений тільки до однієї групи. Допускається при цьому,

що деякі об’єкти не належать до жодної групи (є «невідомими»). Для кожного з об’єктів є дані по ряду кількісних змінних. Такі змінні називаються дискримінантними змінними, або предикторами. Завданнями дискримінантного налізу є визначення:

- ✓ вирішальних правил, що дозволяють за значенням дискримінантних змінних (предикторів) віднести кожен об’єкт (в тому числі і «невідомий») до однієї з відомих груп;
- ✓ «ваги» кожної дискримінантної змінної для розподілу об’єктів на групи.

Існує безліч ситуацій, в яких було б дуже бажано обчислити ймовірність того чи іншого результату в залежності від сукупності вимірюваних змінних, наприклад, з’ясувати: чи підходить претендент на ту чи іншу посаду; чи страждає психічно хвора людина на шизофренію або психоз; чи повернеться ув’язнений до нормального життя після виходу на свободу; які фактори впливають на збільшення ризику пацієнта отримати серцевий напад тощо. У всіх перерахованих ситуаціях є дві спільні риси:

- ✓ по-перше, для багатьох суб’єктів є інформація про їх приналежність до тієї чи іншої групи;
- ✓ по-друге, про кожному суб’єкту є додаткова інформація для створення формули, яка дозволить спрогнозувати приналежність суб’єкта до тієї чи іншої групи.

Дискримінантний аналіз має певну схожість з кластерним аналізом, яка полягає в тому, що дослідник в обох випадках ставить перед собою мету розділити сукупність об’єктів (а не змінних) на декілька більш дрібних груп. Тим не менш процес класифікації у двох видах аналізу принципово різний. У кластерному аналізі об’єкти класифікуються на основі їх відмінності без будь-якої попередньої інформації про кількість і склад класів.

При дискримінантному аналізі кількість і склад класів задані від початку, і основна задача полягає у визначенні того, наскільки точно можна передбачити приналежність об’єктів до класів за допомогою даного набору дискримінантних змінних (предикторів).

Дискримінантний аналіз являє собою альтернативу множинного регресійного аналізу для випадку, коли залежна змінна являє собою не кількісну (номінальну, номінативну) змінну. При цьому дискримінантний аналіз вирішує, по суті, ті ж завдання, що і множинний регресійний аналіз: передбачення значень «залежної» змінної (в даному випадку категорій номінальної ознаки) і визначення того, які «незалежні» змінні найкраще підходять для такого передбачення. Дискримінантний аналіз заснований на складанні рівняння регресії, що використовує номінальну залежну змінну (зауважимо, що вона не є кількісною, як у випадку регресійного аналізу).

Рівняння регресії складається на основі тих об’єктів, про які відома групова належність, що дозволяє максимально точно підібрати його коефіцієнти. Після того як рівняння регресії отримано, його можна використовувати для угрупування об’єктів, які нас цікавлять, з метою прогнозування.

Команда дискримінантного аналізу вельми непроста і вимагає настройки безлічі параметрів, опис більшості з яких лежить за рамками теми даної лекції. При необхідності завжди можна звернутися за додатковою інформацією до керівництва користувача SPSS.

Як і для більшості складних статистичних операцій, параметри дискримінантного аналізу в основному визначаються особливостями даних, а також експериментаторськими прагненнями дослідника. Як завжди, ми розглянемо приклад (па цей раз єдиний) проведення дискримінантного аналізу в розділі покрокових процедур, а розділ «Представлення результатів» присвятимо інтерпретації даних, що виводяться.

Для демонстрації дискримінантного аналізу ми розглянемо приклад прогнозування успішності навчання на основі попереднього тестування. Наш робочий файл містить дані про 46 учнів (об'єкти з 1 по 46), які закінчили курс навчання, щодо яких відомі оцінки успішності навчання – для цього використовується змінна «оцінка» (1 – низька, 2 – висока). Крім того, до файлу включені дані попереднього тестування цих учнів до початку навчання (13 змінних):

- 11 показників тесту інтелекту;
- показник екстраверсії (за результатами тесту Р. Айзенка);
- показник нейротизму (за результатами тесту Р. Айзенка).

Для 10 претендентів на курс навчання відомі лише результати їх попереднього тестування (13 перерахованих змінних).

Значення змінної «оцінка» для них, зрозуміло, невідомі, і у файлі даних їм відповідають порожні клітинки. В процесі дискримінантного аналізу ми, зокрема, спробуємо спрогнозувати успішність навчання цих 10 претендентів припускаючи, що результати тих, хто вже завершив навчання, і результати претендентів є ідентичними.

Етапи дискримінантного аналізу:

1. Вибір змінних-предикторів. Дослідник використовує свої теоретичні знання, практичний досвід, припущення тощо для того, щоб скласти список змінних, які можуть вплинути на результат групування (змінну-критерій). У цьому файлі крім змінної-критерію (оцінка) міститься 13 змінних, що характеризують кожного учня. Це дозволяє нам зробити всі 13 змінних предикторами і включити їх у рівняння регресії. Якби було велике число змінних (наприклад, кілька сотень), то було б неможливо застосувати дискримінантний аналіз до всіх змінним одночасно. Це обумовлено як концептуальними причинами (колінеарність змінних, втрата ступенів свободи тощо), так і практичними обмеженнями (недостатній обсяг оперативної пам'яті комп'ютера). Зазвичай на початковому етапі дискримінантного аналізу для предикторів формується кореляційна матриця. У даному контексті вона має особливий сенс, називається загальною внутрішньогруповою кореляційною матрицею і містить середні коефіцієнти кореляції для двох або більше кореляційних матриць (кожна для однієї групи). Крім загальної внутрішньогрупової кореляційної матриці можна також обчислити коваріаційні матриці для окремих груп, для всієї вибірки або загальну

внутрішньогрупову коваріаційну матрицю. Нерідко дослідники застосовують серію t-критеріїв між двома групами для кожної змінної або однофакторний дисперсійний аналіз, якщо число груп виявляється більше двох. Оскільки метою дискримінантного аналізу є складання найкращого рівняння регресії, додатковий аналіз вихідних даних ніколи не є зайвим. Так, в результаті застосування t-критеріїв для даних нашого прикладу були знайдені значущі відмінності між двома рівнями змінної «оцінка» (8 з 13 предикторів). Найбільш поширеним варіантом дискримінантного аналізу, є варіант, при якому програма автоматично виключає несуттєві для передбачення предиктори, але за критеріями, які встановлює сам користувач.

2. Вибір параметрів. За замовчуванням програма реалізує метод, який заснований на примусовому включенні в регресійне рівняння всіх предикторів, зазначених дослідником. В іншому варіанті використовується метод Вілкса (Wilks), що відноситься до категорії покрокових методів і заснований на мінімізації коефіцієнта Вілкса після включення в рівняння регресії кожного нового предиктора. Так само як і у випадку множинного регресійного аналізу, існує критерій для включення предикторів в рівняння регресії (за замовчуванням таким критерієм є F>3,84) і критерій для виключення предикторів з рівняння регресії (за замовчуванням – F<2,71). Коефіцієнт *X* являє собою відношення внутрішньогрупової суми квадратів до загальної суми квадратів і характеризує частку впливу предиктора на дисперсію критерію. Зі значенням *X* пов'язані величини F і *p*, які характеризують його значущість.

Як показує практика, найчастіше комп'ютер справляється зі складанням рівняння регресії краще, ніж дослідник, який задає список предикторів вручну. Однак зустрічаються ситуації, коли корисніше обмежити самостійність комп'ютера. Наприклад, якщо провести дискримінантний аналіз для наших даних з включенням всіх змінних, то неправильно класифіковані будуть 5 об'єктів з 46. Тієї ж точності прогнозу можна досягти всього з 7 предикторами, якщо вибрати покроковий метод з установками, що відрізняються від прийнятих за замовчуванням. У той же час, якщо використовувати покроковий метод з налаштуваннями за замовчуванням, лишаючи тільки 3 предиктори, кількість невірно згрупованих об'єктів збільшиться до 9. Крім розглянутих програма SPSS має й інші методи вибору предикторів, однак їх опис виходить за рамки теми цієї лекції, і при необхідності, ми рекомендуємо вам звернутися до керівництва користувача SPSS.

3. Інтерпретація результатів. Метою дискримінантного аналізу є складання рівняння регресії з використанням вибірки, для якої відомі значення як предикторів, так і критерію. Це рівняння дозволяє за відомими значеннями предикторів визначити невідомі значення критерію для іншої вибірки. Зрозуміло, що точність розрахованих значень критерію для другої вибірки в загальному випадку не вище, ніж для вихідної. Так, у нашому прикладі регресійне рівняння забезпечило близько 90% коректних результатів для тієї вибірки, за допомогою якої воно було створено. Відповідно, точність

передбачення успішності навчання для 10 претендентів може досягати 90% лише в тому випадку, якщо вибірка претендентів абсолютно ідентична тим 46 учням, дані для яких послужили основою для прогнозу.

11.2. Багатовимірне шкалювання

Основне призначення багатовимірного шкалювання – представлення великих масивів даних про відмінність об'єктів в наочному, доступному для інтерпретації графічному вигляді. При багатовимірному шкалюванні матриця відмінностей між об'єктами представляється у вигляді одно-, двох- або тривимірного графічного зображення взаємного розташування цих об'єктів. Хоча доступно і більше трьох вимірювань, ця можливість рідко застосовується на практиці.

Основною перевагою багатовимірного шкалювання є можливість наочного візуального порівняння об'єктів аналізу. Якщо дві точки на зображенні віддалені одна від одної, то між відповідними об'єктами є значна розбіжність. Навпаки, близькість точок говорить про подібність об'єктів.

Багатовимірне шкалювання має багато спільних рис з факторним аналізом. Так само як і при факторному аналізі, створюється система координат простору, в якому визначається розташування точок. Так само як і при факторному аналізі, відбувається зниження розмірності і спрощення даних. Однак при факторному аналізі зазвичай використовуються коефіцієнти кореляції, а при багатовимірному шкалюванні – відмінності між об'єктами. Нарешті, при факторному аналізі найбільший інтерес викликають кути між точками, що представляють дані, а в багатовимірному шкалюванні ключовою величиною є відстань між цими точками.

Крім факторного аналізу багатовимірне шкалювання має кілька спільних рис з кластерним аналізом. В обох випадках аналізується відстань між об'єктами; однак при кластерному аналізі типовою є кількісна процедура об'єднання об'єктів в групи (кластери), а при багатовимірному шкалюванні якісний аналіз об'єктів проводиться візуально за допомогою діаграми.

Процедура багатовимірного шкалювання у SPSS, що має назву ALSCAL, являє собою набір невеликих процедур, кожна з яких відповідає своєму типу даних.

У якості прикладу опрацюємо гіпотетичну соціограму для групи учнів. У цьому прикладі їх кількісні оцінки відносин один до одного будуть перетворені у графічне зображення взаємного розташування учнів. У другому прикладі ми розглянемо результати тестування учнів за п'ятьма показниками і представимо графічно відмінності між учнями на плоскому зображенні. Нарешті, третій приклад буде представляти собою невелике дослідження сприйняття та розуміння студентами п'яти багатовимірних методів статистичного аналізу.

Найчастіше опис нового статистичного методу зручно проводити шляхом його порівняння з іншим методом. Саме таким чином ми і почнемо розгляд ієрархічного кластерного аналізу, порівнюючи його з факторним

аналізом. При численних загальних рисах між зазначеними статистичними методами існує чимало відмінностей.

Як звичайно, після теоретичної частини підуть приклади практичної реалізації статистичних методів засобами SPSS, оформлені у вигляді покрокових процедур.

11.3. Порівняння кластерного і факторного аналізу. Основні поняття та етапи кластерного аналізу

Головна схожість між кластерним і факторним аналізом полягає в тому, що і той, і інший призначені для переходу від вихідної сукупності безлічі змінних (або об'єктів) до істотно меншого числа факторів (кластерів). Тим не менш реалізація статистичних процедур і інтерпретація результатів для двох типів аналізу розрізняються. *Нижче наведемо основні відмінності*.

Метою факторного аналізу є заміна великої кількості вихідних змінних меншим числом факторів. Кластерний аналіз, як правило, застосовується для того, щоб зменшити кількість об'єктів шляхом їх групування. Іншими словами, у процедурі кластерного аналізу зазвичай змінні не групуються, а виступають в якості критеріїв для групування об'єктів.

Так, в прикладі факторного аналізу 11 субтестів інтелекту (змінних) були зведені до трьох чинників, кожен з яких об'єднав кілька споріднених вихідних змінних. Кластерний аналіз застосовується зазвичай для виділення груп об'єктів, виходячи з їх подібності за виміряними ознаками. Відповідно до прикладу з 11 субтестами інтелекту типовою задачею кластерного аналізу була б класифікація учнів (об'єктів) таким чином, щоб вимірювані за 11 показниками всередині кожної групи об'єкти були схожі один на одного, а не на об'єкти з інших груп. Групи об'єктів, виділені в результаті кластерного аналізу на основі заданої міри подібності між об'єктами, називаються *кластерами*.

Заявлені в попередньому пункті відмінності між кластерним і факторним варіантами аналізу з усією повнотою категоричності можуть бути віднесені лише до попередніх версій SPSS. Починаючи з версії SPSS 10.0, програма дозволяє з однаковим успіхом проводити кластерний аналіз не лише об'єктів, а й змінних. В останньому випадку кластерний аналіз може виступати як більш простий і нерідко більш ефективний аналог факторного аналізу. У розділі покрокових процедур продемонструємо обидва варіанти кластерного аналізу.

Дії, що виконуються в ході статистичних операцій в кожному з варіантів аналізу, принципово різняться. У факторному аналізі на кожному етапі вилучення фактору для кожної змінної підраховується частка дисперсії, яка обумовлена впливом цього чинника. При кластерному аналізі обчислюється відстань між поточним об'єктом і всіма іншими об'єктами, і утворює кластер та пара, для якої відстань виявилася найменшою. Подібним чином кожен об'єкт або групується з іншим об'єктом, або включається до складу існуючого кластера. Процес кластеризації триває до тих пір, поки всі об'єкти не будуть

об’єднані в один кластер. Зрозуміло, подібний результат в загальному випадку не має сенсу, і дослідник повинен самостійно визначити, в який момент кластеризація повинна бути припинена.

У контексті кластерного аналізу особливе місце займає один із його видів, званий ієрархічним кластерним аналізом. У SPSS він реалізується за допомогою команди Hierarchical Cluster (Ієрархічна кластеризація). Цей вид кластерного аналізу частіше використовується в економіці, соціології, політології, ніж у психології. Психологи зазвичай аналізують змінні з метою знайти статистичні зв’язки між ними. Ці зв’язки, як правило, вказують на схожість між тими чи іншими досліджуваними факторами.

Розподіл вибірки на групи при психологічному аналізі рідко представляє інтерес. У випадках, коли це виявляється необхідним, психологи віддають перевагу дискримінантному, а не кластернрму аналізу. Оскільки кластеризація змінних виявляється вельми доступною операцією, було б цікаво порівняти її результати з результатами більш складного факторного аналізу. Як і у випадку факторного аналізу, виконання кластерного аналізу та його результати залежать від ряду параметрів: способу обчислення відстані між об’єктами, кластеризації індивідуальних об’єктів тощо.

Етапи кластерного аналізу. Кластерний аналіз виконується за кілька етапів, що призводять до кінцевого результату. Спочатку розглянемо приклад, створений спеціально для демонстрації суті кластерного аналізу. Зазначимо, що кластерний аналіз непридатний до файлів даних, що використовувалися раніше, оскільки при їх складанні основну увагу було приділено змісту і зв’язкам між змінними, а вміст об’єктів (тобто інформація, що стосується суб’єктів) практично не грав ролі. Для демонстрації кластерного аналізу нами був підготовлений спеціальний файл cars.sav, що містить гіпотетичні дані про 15 автомобілів різних марок, виставлених на продаж. Файл має структуру, відповідну для наочної ілюстрації кластерного аналізу.

Отже, виділяють кілька етапів кластерного аналізу:

1. Вибір показників-критеріїв для кластеризації. У нашому прикладі кластеризація буде здійснюватися за наступними змінними: «ціна» (вартість); «т_стан» (експертна оцінка технічного стану за 10-бальною шкалою); «вік» (кількість років експлуатації); «пробіг» (пройдений кілометраж з початку експлуатації).

2. Вибір способу вимірювання відстані між об’єктами або кластерами (спочатку вважається, що кожен об’єкт відповідає одному кластеру). За замовчуванням використовується квадрат Евклідової відстані, згідно з яким відстань між об’єктами дорівнює сумі квадратів різниць між значеннями однойменних змінних об’єктів. Припустимо, що марка автомобіля «А» має показники технічного стану і віку 5 і 6, а марка «В», відповідно, 7 і 4. У цьому випадку відстань між марками обчислюється наступним чином: $(5-7)^2 + (6-4)^2 = 8$. При виконанні аналізу сума квадратів різниць обчислюється для всіх змінних. Одержувані відстані використовуються програмою при формуванні кластерів. Крім Евклідова існують і інші види відстаней, обчислювані за іншими формулами, проте в рамках даного навчального курсу

немає гострої необхідності на них зупинятися.

Щодо обчислення відстані може виникнути наступне запитання: чи буде адекватним результат кластерного аналізу у тому випадку, якщо змінні мають різні шкали вимірювання? Так, всі змінні файлу car.sav мають різні шкали. Для вирішення проблеми шкалювання в SPSS використовується стандартизація, зокрема, її простий метод – нормалізація змінних, що приводить всі змінні до стандартної z-шкали (середнє дорівнює 0, стандартне відхилення – 1). Крім однакової шкали нормалізовані змінні також мають рівні ваги. У разі якщо всі вихідні дані мають одну й ту саму шкалу вимірювання або ваги змінних за змістом повинні бути різними, стандартизацію змінних проводити непотрібно.

3. Формування кластерів. Існує два основних методи формування кластерів: метод злиття і метод дроблення. В першому випадку вихідні кластери збільшуються шляхом об'єднання до тих пір, поки не буде сформований єдиний кластер, що містить всі дані. Метод дроблення заснований на зворотній операції: спочатку всі дані об'єднуються в один кластер, який потім ділиться на частини до тих пір, поки не буде досягнутий бажаний результат. За замовчуванням програмою SPSS використовується метод злиття.

У методі злиття передбачено кілька способів об'єднання об'єктів. Спосіб, використовуваний за замовчанням, називається міжгруповим зв'язуванням, або зв'язуванням середніх всередині груп. SPSS обчислює найменше середнє значення відстані між усіма парами груп і об'єднує дві групи, які видалися найбільш близькими. На першому кроці, коли всі кластери являють собою поодинокі об'єкти, дана операція зводиться до звичайного попарного порівняння відстаней між об'єктами.

Термін «середнє значення» набуває сенсу лише на другому етапі, коли сформовані кластери, які містять більше одного об'єкта. Так, в окремому прикладі на початковому етапі є 15 кластерів (об'єктів); спочатку в кластер об'єднуються два об'єкти з найменшою відстанню один від одного. Потім підрахунок відстаней повторюється, і в кластер об'єднується ще одна пара змінних. На другому етапі отримуємо або 13 вільних об'єктів і 1 кластер, який об'єднує 2 об'єкти, або 11 вільних об'єктів і 2 кластера по 2 об'єкти в кожному. В кінцевому рахунку всі об'єкти опиняться в одному великому кластері. Існують й інші методи об'єднання об'єктів, проте в рамках даного курсу їх детальний розгляд немає сенсу.

4. Інтерпретація результатів. Як і у випадку факторного аналізу, бажане число кластерів і оцінка результатів аналізу залежать від цілей дослідника. Для розглянутого прикладу нам представляється найкращим число кластерів, рівне 3. Як показує аналіз, всі марки автомобілів можна розділити на 3 групи: перша група має високу вартість (середнє значення – 15 230), недовгий термін експлуатації (4 роки) і середній пробіг (85 400 км); друга група має середню вартість, невеликий пробіг, максимальний вік, але гарний технічний стан; третя група містить недорогі моделі з великим пробігом і невисоким рейтингом технічного стану.