Міністерство освіти і науки України
Львівський національний університет імені Івана Франка

В.С.ГРИЦЕВИЧ

КОРЕЛЯЦІЙНИЙ ТА РЕГРЕСІЙНИЙ АНАЛІЗ В СУСПІЛЬНІЙ ГЕОГРАФІЇ. ТЕКСТИ ЛЕКЦІЙ

Львів 2016

Рецензенти:

докт. біолог. наук. З.Г.Гамкало
(Львівський національний університет імені Івана Франка)
канд. геогр. наук. Ю.М.Андрейчук
(Львівський національний університет імені Івана Франка)

Науковий редактор д.г.н. професор Шаблій О.І.

Рекомендовано до друку
Вченою радою географічного факультету
Львівського національного університету імені Івана Франка.
Протокол №5 від 15.06.2016

Грицевич В.С.
Кореляційний та регресійний аналіз в суспільній географії: тексти лекцій. –Львів: Малий видавничий центр. Лабораторія тематичного картографування географічного факультету, 2016. -24 с.

У тексті лекцій подані теоретичні та методичні основи обчислення коефіцієнтів кореляції, їхньої інтерпретації, побудови регресійних моделей при виконанні суспільно-географічних досліджень, зокрема при написанні курсових, дипломних і магістерських робіт.

Для бакалаврів заочної форми навчання спеціальності "географія".

КОРЕЛЯЦЙНИЙ АНАЛІЗ

План

1. Вимірювання лінійного зв’язку між ознаками. Коефіцієнт парної кореляції.
2. Коефіцієнт множинної кореляції.
3. Коефіцієнт часткової кореляції.
4. Вимірювання нелінійного зв’язку між ознаками.
5. Непараметричні коефіцієнти кореляції.

1. Вимірювання лінійного зв’язку між ознаками. Коефіцієнт парної кореляції

Об’єкти, які доводиться вивчати в суспільній географії, завжди характеризуються багатьма ознаками і ці ознаки дуже часто перебувають у певних зв’язках. Такі зв’язки у статистичній науці називають кореляцією ознак. Кореляція між ознаками виникає у двох випадках:

- якщо одна ознака є факторною по відношенню до другої (тобто описує причини мінливості другої ознаки);
- якщо існують спільні фактори, що визначають мінливість обох ознак.

Параметричний коефіцієнт парної кореляції є мірою лінійного статистичного зв’язку між двома кількісними ознаками. Він може набувати значення від -1 до +1 . Якщо коефіцієнт кореляції більший нуля, то при зростанні одної ознаки, друга ознака в середньому також зростає. Якщо він менший нуля, то при зростанні одної ознаки, друга ознака в середньому спадає. Якщо коефіцієнт кореляції близький до +1, то між ознаками є функціональний прямий лінійний зв'язок. Якщо він близький до -1, то між ознаками є функціональний обернений лінійний зв'язок. Якщо коефіцієнт кореляції близький до нуля, то це означає, що між ознаками нема лінійного зв’язку. У цьому випадку або зв’язку між ознаками нема взагалі, або він є, але нелінійний.

Коефіцієнт парної кореляції обчислюють для двох ознак на підставі масиву спостережень за ними. Нехай x, y - дві кількісні ознаки, а масив спостережень записаний у такій таблиці:

Номер спостереження	Значення першої ознаки	Значення другої ознаки
1	x_1	y_1
2	x_2	y_2
…	…	…
m	x_m	y_m

У літературі для обчислення коефіцієнта кореляції існує багато різних на вигляд, але математично еквівалентних формул. Одна з найпростіших і найменш чутлива до похибок має такий вигляд:

$$r_{xy} = \frac{S_{xy}}{\sqrt{S_{xx} \cdot S_{yy}}},$$

де $S_{xx} = \sum_{i=1}^{m} (x_i - \bar{x})^2$, $S_{xy} = \sum_{i=1}^{m} (x_i - \bar{x})(y_i - \bar{y})$, $S_{yy} = \sum_{i=1}^{m} (y_i - \bar{y})^2$.

Як завжди $\bar{x} = \frac{1}{m} \cdot \sum_{i=1}^{m} x_i$, $\bar{y} = \frac{1}{m} \cdot \sum_{i=1}^{m} y_i$.

Приклад. Самостійно обчислити коефіцієнт кореляції для такої таблиці даних:

Номер спостереження	Значення першої ознаки	Значення другої ознаки
1	1	5
2	11	17
3	6	9
4	16	13
5	21	21

(Відповідь: r=+0,9)

Для коефіцієнта парної кореляції визначають ступінь його вірогідності за критерієм Стьюдента. Визначають емпіричне $t_{емп}$ та теоретичне $t_{теор}$ значення критерію. Емпіричне значення обчислюють за формулою:

$$t_{емп} = \frac{\sqrt{m-3}}{2} \cdot \ln \frac{1+|r_{xy}|}{1-|r_{xy}|}.$$

Теоретичне значення визначають із спеціальних таблиць, входами яких служать кількість ступенів вільності $(m-2)$, і рівень значущості α (як правило 5%).

Якщо $t_{теор} \le t_{емп}$, то вважають, що вірогідність кореляції підтверджена. Якщо ж $t_{емп} < t_{теор}$, то нема підстав вважати кореляцію вірогідною.

З коефіцієнтом кореляції тісно пов'язаний його квадрат $d_{xy} = (r_{xy})^2$, який називається коефіцієнтом детермінації. Він показує, яка частина мінливості результуючої ознаки пояснюється лінійним впливом мінливості факторної ознаки. У зв'язку з цим коефіцієнт детермінації часто вимірюють у відсотках (тобто обчислюють $d_{xy} \cdot 100\%$). Наприклад, якщо $r_{xy} = 0{,}7$, то коефіцієнт детермінації близький до 50%, тобто майже половину мінливості ознаки y можна пояснити лінійним впливом мінливості ознаки x.

2. Коефіцієнт множинної кореляції

Коефіцієнт множинної кореляції застосовують тоді, коли є кілька факторних ознак і одна результуюча. Він є мірою лінійного зв'язку між сукупністю факторних ознак і результуючою ознакою. Коефіцієнт множинної кореляції може набувати значень від 0 до +1. Якщо він близький до нуля, то це означає, що результуюча ознака лінійно не залежить від сукупності факторних ознак. Навпаки, якщо коефіцієнт множинної кореляції близький до +1, то це свідчить про значний спільний лінійний вплив факторних ознак на результуючу.

Якщо факторних ознак є лише дві, то коефіцієнт множинної кореляції легко виражається через відповідні парні кореляції. Нехай x, y - дві факторних ознаки, а w - результуюча ознака. Спочатку, за відомою вже формулою, обчислюємо парні кореляції:

- r_{xy} - коефіцієнт кореляції між факторними ознаками,
- r_{xw} - коефіцієнт кореляції між першою факторною ознакою і результуючою,
- r_{yw} - коефіцієнт кореляції між другою факторною ознакою і результуючою

Тоді коефіцієнт множинної кореляції $r_{w(xy)}$ обчислюємо за такою формулою:

$$r_{w(xy)} = \sqrt{\frac{r_{xw}^2 + r_{yw}^2 - 2 \cdot r_{xy} \cdot r_{xw} \cdot r_{yw}}{1 - r_{xy}^2}} .$$

Оскільки кожна з факторних ознак доповнює вплив іншої, то їхній спільний вплив на результуючу ознаку завжди є сильнішим, ніж вплив кожної з них зокрема, тому мають місце нерівності:

$$r_{w(xy)} \geq \left| r_{xw} \right|, \quad r_{w(xy)} \geq \left| r_{yw} \right| .$$

Приклад. Нехай $r_{xy} = 0{,}939$, $r_{xw} = 0{,}610$, $r_{yw} = 0{,}794$. Самостійно обчислити $r_{w(xy)}$. *(Відповідь:* $r_{w(xy)} = 0{,}887$ *)*

Якщо факторних ознак є багато (більше двох), то нема простої формули для обчислення коефіцієнта множинної кореляції. В цьому випадку спочатку, на підставі спостережуваних значень w_i, $i = 1,...,m$ результуючої змінної, будують рівняння багатофакторної лінійної регресії і обчислюють регресійні значення $\tilde{w}_i$, $i = 1,...,m$. Після цього коефіцієнт множинної кореляції обчислюють за формулою:

$$r = \sqrt{1 - \frac{\sum_{i=1}^{m} (w_i - \tilde{w}_i)^2}{\sum_{i=1}^{m} (w_i - \overline{w})^2}} \ .$$

3. Коефіцієнт часткової кореляції

Коефіцієнт часткової кореляції застосовують тоді, коли є дві результуючі ознаки і деяка кількість факторних. Він є мірою лінійного зв'язку між результуючими ознаками після усунення лінійних впливів, викликаних факторними ознаками.

Коефіцієнт часткової кореляції може мати значення від -1 до +1, причому слід врахувати, що за величиною, а інколи навіть за знаком, він може відрізнятися від відповідного парного коефіцієнта кореляції між результуючими ознаками.

Якщо є лише одна факторна ознака, то коефіцієнт часткової кореляції легко виражається через відповідні парні кореляції. Нехай x - факторна ознака, а u, v - результуючі ознаки. Спочатку, за відомою вже формулою, обчислюємо парні кореляції:

- r_{uv} - парний коефіцієнт кореляції між результуючими ознаками,
- r_{xu} - коефіцієнт кореляції між факторною ознакою і першою результуючою,
- r_{xv} - коефіцієнт кореляції між факторною ознакою і другою результуючою.

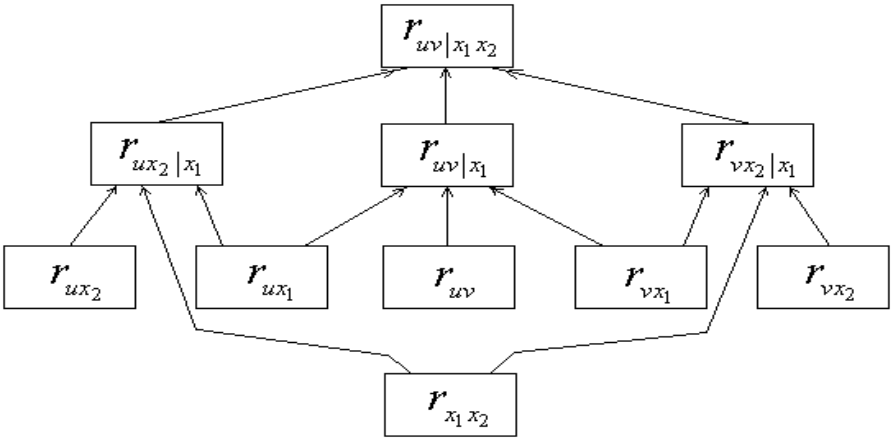
Тоді коефіцієнт часткової кореляції $r_{uv|x}$ обчислюємо за такою формулою:

$$r_{uv|x} = \frac{r_{uv} - r_{xu} \cdot r_{xv}}{\sqrt{(1 - r_{xu}^2) \cdot (1 - r_{xv}^2)}} \ .$$

Приклад. Нехай $r_{uv} = 0{,}948$, $r_{xu} = 0{,}610$, $r_{xv} = 0{,}654$.

Самостійно обчислити $r_{uv|x}$. *(Відповідь:* $r_{uv|x} = 0{,}917$ *)*

Якщо факторних ознак є багато (більше одної), то нема простої формули для обчислення коефіцієнта часткової кореляції. У цьому випадку послідовно перераховують коефіцієнти часткової кореляції вищого рангу через відповідні коефіцієнти нижчого рангу, причому найнижчий ранг мають відомі парні коефіцієнти кореляції. Наприклад, для двох факторних змінних x_1, x_2 схема обчислень коефіцієнта $r_{uv|x_1x_2}$ має чотири етапи перерахунку і виглядає так:



4. Вимірювання нелінійного зв'язку між ознаками

Розглянемо факторну ознаку x і результуючу ознаку y, над якими здійснено m спостережень (x_i, y_i), $i = 1,2,\ldots,m$. Нехай між ознаками встановлений нелінійний регресійний зв'язок $\hat{y} = f(x)$. Обчислимо регресійні значення результуючої ознаки: $\hat{y}_i = f(x_i)$, $i = 1,2,\ldots,m$. Тоді можна обчислити кореляційне відношення за формулою:

$$\eta = \sqrt{\frac{\sum_{i=1}^{m}\left(\hat{y}_i - \bar{y}\right)^2}{\sum_{i=1}^{m}\left(y_i - \bar{y}\right)^2}}.$$

Кореляційне відношення може приймати значення від 0 до 1. Якщо між ознаками x та y повністю відсутній будь-який зв'язок, то $\hat{y}$ буде константою, яка співпадає з $\bar{y}$, тому $\eta = 0$. Навпаки, якщо між ознаками x та y є функціональний зв'язок, то він описується формулою $y = f(x)$ і тому $\eta = 1$.

5. Непараметричні коефіцієнти кореляції

Описані вище коефіцієнти кореляції стосуються випадку, коли дослідник має в своєму розпорядженні числові значення кількісних ознак, причому бажано, щоб самі ознаки мали нормальний розподіл.

У практиці суспільно-географічних досліджень так буває не завжди. Нерідко ознаки мають довільний розподіл, крім цього можуть бути відомі не значення ознак, а лише їхні ранги, нарешті самі ознаки можуть бути не кількісними, а якісними. В цих ситуаціях також ставлять завдання виміряти зв'язок між ознаками, однак, в силу такої специфіки даних, для цього необхідно скористатись спеціальними непараметричними коефіцієнтами кореляції. В літературі описано багато різних непараметричних коефіцієнтів кореляції, тому зупинимось тут лише на найбільш часто вживаних.

4.1. Корелювання кількісних ознак з довільним розподілом

Нехай x, y - дві кількісні ознаки, за якими зроблено m спостережень. Спочатку, за відомими формулами обчислюють середні арифметичні значення спостережень за ознаками: $\bar{x}, \bar{y}$. Далі обчислюють всі різниці $(x_i - \bar{x})$ та $(y_i - \bar{y})$, $i = 1,...,m$. Ці різниці можуть бути як додатними так і від'ємними, тобто мати різні

знаки, тому підраховують кількість C випадків співпадінь знаків, і кількість H їх неспівпадінь. Тоді обчислюють коефіцієнт кореляції Фехнера за формулою:

$$r_{xy} = \frac{C - H}{C + H}.$$

Коефіцієнт кореляції Фехнера може приймати значення від -1 до +1.

4.2. Корелювання рангів

Розглянемо випадок, коли значення ознак неможливо встановити і відомими є лише їхні ранги. Нехай x, y - дві ознаки, за якими зроблено m спостережень. Позначимо через R_i^x ранг i-го спостереження за змінною x, а через R_i^y ранг i-го спостереження за змінною y, $i = 1,...,m$. Ранги – це числа від 1 до m. Якщо ранжування здійснюють за спаданням, то ранг 1 має найбільше за величиною спостереження, а ранг m - найменше за величиною. Ранжування за зростанням здійснюють навпаки. У будь-якому випадку ранговий коефіцієнт кореляції Спірмена обчислюють за формулою:

$$r_{xy} = 1 - \frac{6 \cdot \sum_{i=1}^{m} (R_i^x - R_i^y)^2}{m \cdot (m^2 - 1)}.$$

Коефіцієнт кореляції Спірмена також може приймати значення від -1 до +1.

4.3. Корелювання якісних ознак

Нехай X, Y - дві якісних ознаки, кожна з яких може приймати по два альтернативних якісних значення. Позначимо через A, B

альтернативні якісні значення ознаки X, а через C, D альтернативні якісні значення ознаки Y. На основі масиву спостережень за ознаками X, Y підрахуємо кількість випадків появи кожного поєднання значень ознак X, Y. В результаті отримаємо такі числа:

N_{AC} - кількість випадків, коли $X = A$ і $Y = C$ в масиві спостережень;

N_{AD} - кількість випадків, коли $X = A$ і $Y = D$ в масиві спостережень;

N_{BC} - кількість випадків, коли $X = B$ і $Y = C$ в масиві спостережень;

N_{BD} - кількість випадків, коли $X = B$ і $Y = D$ в масиві спостережень.

Цю інформацію можна розмістити в таблиці:

Y / X	C	D
A	N_{AC}	N_{AD}
B	N_{BC}	N_{BD}

На основі зроблених підрахунків обчислюємо міри звязку між ознаками X, Y.

- Коефіцієнт асоціації Юла:

$$Q = \frac{N_{AC} \cdot N_{BD} - N_{AD} \cdot N_{BC}}{N_{AC} \cdot N_{BD} + N_{AD} \cdot N_{BC}}.$$

- Коефіцієнт контингенції Пірсона:

$$K = \frac{N_{AC} \cdot N_{BD} - N_{AD} \cdot N_{BC}}{\sqrt{(N_{AC} + N_{BC}) \cdot (N_{AD} + N_{BD}) \cdot (N_{AC} + N_{AD}) \cdot (N_{BC} + N_{BD})}}$$

.

Кожен з цих коефіцієнтів вимірює рівень зв’язку між якісними ознаками за шкалою від -1 до +1.

Контрольні запитання

1. За якою ознакою статистичний зв’язок поділяють на прямий та обернений ?
2. За якою ознакою статистичний зв’язок поділяють на лінійний та нелінійний ?
3. За якою ознакою статистичний зв’язок поділяють на компонентний та балансовий ?
4. В якій формі залежності одному значенню аргумента відповідає одне значення функції ?
5. В яких межах лежить параметричний коефіцієнт лінійної кореляції ?
6. В яких межах лежить коефіцієнт множинної кореляції ?
7. В яких межах лежить коефіцієнт часткової кореляції ?
8. Який коефіцієнт кореляції вимірює зв'язок між рангами ознак ?
9. Який коефіцієнт вимірює зв’язок між якісними ознаками ?

РЕГРЕСІЙНИЙ АНАЛІЗ

План

1. Поняття про регресійну залежність
2. Однофакторна лінійна регресія
3. Однофакторна нелінійна регресія
4. Двофакторна лінійна регресія
5. Багатофакторна лінійна регресія

1. Поняття про регресійну залежність

Ознаки, які вивчають у регресійному аналізі, можна поділити на факторні та результуючі. Факторною називають ознаку, мінливість якої є причиною мінливості результуючої ознаки. Результуючою називають ознаку, мінливість якої є наслідком мінливості факторної ознаки.

Між факторною та результуючою ознаками можуть спостерігатися функціональна та статистична залежності. Залежність називається функціональною, якщо одному значенню факторної ознаки відповідає єдине значення результуючої ознаки. Залежність називається статистичною, якщо одному значенню факторної ознаки може відповідати кілька значень результуючої ознаки.

Для виявлення статистичної залежності між кількісними ознаками використовують такі методи: аналітичне групування, побудова точкового графіка, кореляційний аналіз, дисперсійний аналіз та інші.

За напрямком статистичні залежності поділяють на прямі і обернені. У випадку прямої статистичної залежності, при зростанні факторної ознаки, результуюча ознака в середньому теж зростає. У випадку оберненої статистичної залежності, при зростанні факторної ознаки, результуюча ознака в середньому спадає.

За формою статистичні залежності бувають лінійними та нелінійними.

За змістом статистичні залежності бувають компонентні і балансові.

Для вивчення статистичних залежностей використовують рівняння регресії. Рівняння регресії – це така функціональна залежність, яка наближено, але максимально точно описує реальну статистичну залежність між ознаками.

Нехай x - факторна ознака, y - результуюча ознака.

Розглянемо загальний вигляд регресійних рівнянь, залежно від кількості факторних та результуючих ознак.

1. Одна факторна і одна результуюча ознака:

$$y = f(x).$$

2. n факторних ознак і одна результуюча ознака:

$$y = f(x_1, x_2, \ldots, x_n).$$

3. Одна факторна ознака і m результуючих ознак:

$$\begin{cases} y_1 = f_1(x) \\ y_2 = f_2(x) \\ \quad \ldots \\ y_m = f_m(x) \end{cases}.$$

4. n факторних ознак і m результуючих ознак:

$$\begin{cases} y_1 = f_1(x_1, x_2, \ldots, x_n) \\ y_2 = f_2(x_1, x_2, \ldots, x_n) \\ \quad \ldots \\ y_m = f_m(x_1, x_2, \ldots, x_n) \end{cases}.$$

Розглянемо головні форми однофакторної та двофакторної регресійних залежностей.

Головні форми однофакторної регресійної залежності (загалом їх безліч): лінійна, квадратична, гіперболічна, експонентна, степенева.

Головні форми двофакторної регресійної залежності (загалом їх безліч): лінійна, білінійна, квадратична.

2. Однофакторна лінійна регресія

Нехай x - факторна ознака, y - результуюча ознака.

Однофакторне рівняння лінійної регресії має вигляд:

$$y = a_0 + a_1 \cdot x .$$

Побудова цього рівняння полягає у визначенні коефіцієнтів a_0, a_1. Інформаційною базою для їх визначення служить масив спостережень за значеннями факторної та результуючої ознак. Цей масив можна записати у вигляді такої таблиці:

Номер спостереження	**Значення факторної ознаки**	**Значення результуючої ознаки**
1	x_1	y_1
2	x_2	y_2
⋮	⋮	⋮
m	x_m	y_m

Провідним методом визначення коефіцієнтів a_0, a_1 є метод найменших квадратів[1]. Після його застосування отримуються формули для визначення (ідентифікації) коефіцієнтів:

$$a_1 = \frac{S_{xy}}{S_{xx}}, \quad a_0 = \bar{y} - a_1 \cdot \bar{x},$$

де $$S_{xx} = \sum_{i=1}^{m} (x_i - \bar{x})^2, \quad S_{xy} = \sum_{i=1}^{m} (x_i - \bar{x})(y_i - \bar{y}),$$

$$\bar{x} = \frac{1}{m} \cdot \sum_{i=1}^{m} x_i, \quad \bar{y} = \frac{1}{m} \cdot \sum_{i=1}^{m} y_i .$$

[1] Суть цього методу полягає в тому, що коефіцієнти a_0, a_1 визначають з умови мінімуму суми квадратів різниць між регресійними значеннями результуючої ознаки та її спостережуваними значеннями.

3. Однофакторна нелінійна регресія

На загал, нелінійних регресійних рівнянь існує безліч, тому розглянемо тут деякі найчастіше вживані у наукових дослідженнях:

Нелінійне регресійне рівняння	**Вигляд рівняння**
квадратичне рівняння	$y = a_0 + a_1 \cdot x + a_2 \cdot x^2$
гіперболічне рівняння (три форми)	$y = a_0 + \frac{a_1}{x}$
	$y = \frac{1}{a_0 + a_1 \cdot x}$
	$y = \frac{x}{a_0 + a_1 \cdot x}$
експонентне рівняння	$y = a_0 \cdot e^{a_1 \cdot x}$
степеневе рівняння	$y = a_0 \cdot x^{a_1}$

Побудова кожного з цих рівнянь полягає у визначенні їхніх коефіцієнтів. Інформаційною базою для визначення коефіцієнтів служить масив спостережень за значеннями факторної та результуючої ознак. Провідним методом визначення коефіцієнтів є метод найменших квадратів.

Для ідентифікації трьох коефіцієнтів квадратичного рівняння існують спеціальні формули, однак вони достатньо громіздкі і тут розглядатися не будуть. Далі розглянемо ідентифікацію інших нелінійних рівнянь, які містять по два коефіцієнти.

Для визначення коефіцієнтів цих (а також багатьох інших) регресійних рівнянь використовують процедуру лінеаризації.

Лінеаризація – це перетворення нелінійного регресійного рівняння до лінійного вигляду з метою визначення (ідентифікації) його коефіцієнтів. Розглянемо спосіб лінеаризації наведених вище нелінійних рівнянь.

Нелінійне рівняння	**Перетворення лінеаризації**	**Результат лінеаризації**
$y = a_0 + \frac{a_1}{x}$	$u = \frac{1}{x}$	$y = a_0 + a_1 \cdot u$
$y = \frac{1}{a_0 + a_1 \cdot x}$	$w = \frac{1}{y}$	$w = a_0 + a_1 \cdot x$
$y = \frac{x}{a_0 + a_1 \cdot x}$	$u = \frac{1}{x}$ $w = \frac{1}{y}$	$w = a_1 + a_0 \cdot u$
$y = a_0 \cdot e^{a_1 \cdot x}$	$w = \ln(y)$	$w = \ln(a_0) + a_1 \cdot x$
$y = a_0 \cdot x^{a_1}$	$u = \ln(x)$ $w = \ln(y)$	$w = \ln(a_0) + a_1 \cdot u$

Бачимо, що в результаті кожної лінеаризації отримується знайоме лінійне регресійне рівняння, метод ідентифікації якого розглянутий у попередньому пункті лекції. Отже, є велика низка нелінійних рівнянь, які можна використовувати для моделювання залежностей між досліджуваними ознаками, причому ідентифікація коефіцієнтів нелінійних рівнянь зводиться до ідентифікації коефіцієнтів деякого лінійного рівняння.

4. Двофакторна лінійна регресія

Нехай x, y - факторні ознаки, z - результуюча ознака.

Двофакторне рівняння лінійної регресії має вигляд:

$$z = a_0 + a_1 \cdot x + a_2 \cdot y .$$

Побудова цього рівняння полягає у визначенні коефіцієнтів a_0, a_1, a_2. Інформаційною базою для їх визначення служить масив спостережень за значеннями двох факторних та одної результуючої ознаки. Цей масив можна записати у вигляді такої таблиці:

Номер спостереження	Перша факторна ознака	Друга факторна ознака	Результуюча ознака
1	x_1	y_1	z_1
2	x_2	y_2	z_2
⋮	⋮	⋮	⋮
m	x_m	y_m	z_m

Провідним методом визначення коефіцієнтів a_0, a_1, a_2 є метод найменших квадратів. Після його застосування отримуються формули для визначення (ідентифікації) цих коефіцієнтів:

$$\Delta = S_{xx} \cdot S_{yy} - S_{xy}^2 ,$$

$$a_1 = \frac{S_{xw} \cdot S_{yy} - S_{xy} \cdot S_{xw}}{\Delta}, \qquad a_2 = \frac{S_{xx} \cdot S_{yw} - S_{xy} \cdot S_{yw}}{\Delta},$$

$$a_0 = \overline{w} - a_1 \cdot \overline{x} - a_2 \cdot \overline{y} ,$$

$$S_{xx} = \sum_{i=1}^{m} (x_i - \overline{x})^2 , \quad S_{yy} = \sum_{i=1}^{m} (y_i - \overline{y})^2 , \quad S_{xy} = \sum_{i=1}^{m} (x_i - \overline{x})(y_i - \overline{y}) ,$$

$$S_{xw} = \sum_{i=1}^{m} (w_i - \overline{w})(x_i - \overline{x}) , \quad S_{yw} = \sum_{i=1}^{m} (w_i - \overline{w})(y_i - \overline{y}) .$$

5. **Багатофакторна лінійна регресія**

Нехай $x_1, x_2, \ldots, x_n$ - факторні ознаки, z - результуюча ознака.

Багатофакторне рівняння лінійної регресії має вигляд:

$$z = a_0 + a_1 \cdot x_1 + a_2 \cdot x_2 + \ldots + a_n \cdot x_n = a_0 + \sum_{j=1}^{n} a_j \cdot x_j \ .$$

Побудова цього рівняння полягає у визначенні коефіцієнтів $a_0, a_1, \ldots, a_n$. Інформаційною базою для їх визначення служить масив спостережень за значеннями n факторних та одної результуючої ознаки.

Для ідентифікації коефіцієнтів багатофакторного регресійного рівняння перейдемо до векторно-матричного запису. Зробимо такі позначення:

$$\bar{a} = \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_n \end{pmatrix}, \qquad X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1n} \\ 1 & x_{21} & \cdots & x_{21} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{m1} & \cdots & x_{mn} \end{pmatrix}, \qquad \bar{z} = \begin{pmatrix} z_1 \\ z_2 \\ \vdots \\ z_m \end{pmatrix}.$$

Отже, $\bar{a}$ - вектор-стовбець коефіцієнтів регресійного рівняння. X - прямокутна матриця, яка має стільки рядків, скільки є спостережень і стільки стовпців, скільки є коефіцієнтів регресійного рівняння. Перший стовбець цієї матриці складається з одиниць, а наступні стовпці складаються із спостережень за значеннями факторних ознак. $\bar{z}$ - вектор-стовбець спостережень за результуючою ознакою.

У такому випадку коефіцієнти багатофакторного регресійного рівняння визначаються за формулою:

$$\bar{a} = \left(X^T X\right)^{-1} X^T \bar{z} \, .$$

Отже, процедура визначення коефіцієнтів $a_0, a_1, \ldots, a_n$ включає низку матричних операцій: транспонування матриці, множення матриць, обернення матриці, множення матриці на вектор-стовбець. При визначенні коефіцієнтів багатофакторного регресійного

рівняння може виникнути дві проблеми: мультиколінеарність та гетероскедатичність.

Мультиколінеарність – це висока кореляційна залежність між ознаками. Її негативними наслідками є зміщені оцінки та підвищена дисперсія коефіцієнтів регресійного рівняння. Для усунення мультиколінеарності використовують такі способи: об'єднання ознак, відкидання окремих ознак, збільшення кількості спостережень, застосування факторного аналізу і перехід від ознак до факторів.

Гетероскедатичність – це залежність внутрішньогрупової дисперсії спостережуваних ознак від групи спостереження. Її негативними наслідками є неможливість оцінити значущість обчислених коефіцієнтів та побудувати довірчі інтервали для регресії результуючої ознаки. Для усунення гетероскедатичності здійснюють функціональне перетворення ознак (як правило нелінійне), яке дає змогу вирівняти відповідні внутрішньогрупові дисперсії.

Контрольні запитання

1. Що є інформаційною базою для визначення коефіцієнтів регресійного рівняння ?
2. Який метод використовують для визначення коефіцієнтів регресійного рівняння ?
3. Скільки коефіцієнтів має однофакторне лінійне регресійне рівняння ?
4. Скільки коефіцієнтів має однофакторне квадратичне регресійне рівняння ?
5. Скільки коефіцієнтів має однофакторне експонентне регресійне рівняння ?
6. Скільки коефіцієнтів має однофакторне степеневе регресійне рівняння ?
7. Скільки коефіцієнтів має двофакторне лінійне регресійне рівняння ?
8. Скільки коефіцієнтів має двофакторне білінійне регресійне рівняння ?
9. Скільки коефіцієнтів має двофакторне квадратичне регресійне рівняння ?
10. Як називається метод перетворення нелінійного рівняння регресії до лінійного вигляду ?
11. Як називається висока залежність між змінними багатофакторної лінійної регресійної моделі ?

Рекомендована література

Базова

1. Вашків П.Г., Патер П.І., Сторожук В.П., Ткач Є.І. Теорія статистики. –К.: Либідь, 2001.
2. Захожай В.Б. Попов І.І., Коваленко О.В. Практикум з основ статистики. –К.: МАУП, 2001.
3. Беркита К.Ф. Економічна статистика. Курс лекцій. –К. 2004.
4. Крамченко Л.І. Статистика ринку товарів та послуг: Навчальний посібник. –Львів: Новий світ, 2006.
5. Матковський С.О., Марець О.Р. Теорія статистики: Навчальний посібник. –К.:Знання, 2009.

Допоміжна

1. Козаченко І.В. Статистика. –К.: Вища школа, 1992.
2. Громыко Г.Л. Статистика: учебник для студ. ун-тов, обучающихся по спец. "география". –М.: МГУ, 1981.
3. Грицевич В.С. Збірник практичних робіт з курсу "Статистичні методи в соціально-економічній географії". -Львів: Видавничий центр ЛНУ імені Івана Франка, 2006.
4. Грицевич В.С. Статистичні ознаки та характеристики їхньої центральної тенденції: Тексти лекцій. -Львів: Видавничий центр ЛНУ імені Івана Франка, 2008.
5. Грицевич В.С., Котик Л.І. Завдання та методичні рекомендації до виконання практичних робіт з курсу "Статистичні методи в соціально-економічній географії" для студентів географічного факультету. -Львів: Видавничий центр ЛНУ імені Івана Франка, 2011. –96 с.
6. Грицевич В.С., Котик Л.І. Статистичні методи в суспільній географії: навчальний посібник-практикум для самостійної роботи студентів / В.С. Грицевич, Л.І.Котик. − Львів: Видавничий центр ЛНУ імені Івана Франка, 2015. − 64с.
7. Толбатов Ю.А. Загальна теорія статистики засобами EXCEL. –К.: Четверта хвиля, 1999.
8. Ковтун Н.В., Столяров Г.С. Загальна теорія статистики: Курс лекцій. –К.: Четверта хвиля, 1996.
9. Гетало В.П., Борух В.О., Алямкін Р.В. Економічна статистика. Навчальний посібник. –Полтава, 2002.

Публікації автора

1. Грицевич В.С. Геостатистичний аналіз розселення міського населення Тернопільської області : матеріали третьої звітної наук.-практ. конф. викл. та студ. Тернопільського держ. пед. ін-ту за 1992 рік (Тернопіль, 1993 р.). – Тернопіль, 1993. – С. 77 – 80.
2. Грицевич В.С. Методичні вказівки до застосування математичних методів при виконанні курсових і дипломних робіт для студентів 4-5 курсів географічного факультету спеціальності "Економічна і соціальна географія".(Множинна і часткова кореляція) / В.С. Грицевич . - Львів : ЛДУ, 1994. – 16 с.
3. Грицевич В.С. Геостатистика в географії : матеріали IV звітної наук. – практ. конф. виклад. та студ. географ. ф-ту (Тернопіль, 1994 р.). – Тернопіль, 1994. – С. 26-29.
4. Грицевич В.С. Центрографічний аналіз географічних полів : матеріали IV звітної наук. – практ. конф. виклад. та студ. географ. ф-ту за 1993 рік (Тернопіль, 1994 р.). – Тернопіль, 1994. – С.30-31.
5. Грицевич В.С. Методичні вказівки до застосування математичних методів при виконанні курсових і дипломних робіт для студентів 4-5 курсів географічного факультету спеціальності "Економічна і соціальна географія".(Одновимірний регресійний аналіз/ В.С. Грицевич . - Львів : ЛДУ, 1995 . – 20 с.
6. Грицевич В.С. Геостатистичний моніторинг соціально-економічних явищ : зб. наук. праць за матеріалами міжнарод. наук.-практ. конф. - Тернопіль, 1997. -С.154-155.
7. Грицевич В.С. Геостатистичний аналіз топонімії населених пунктів українсько-польського прикордоння : матеріали наук. - практ. конф. [“Економіко-, соціально- і еколого-географічні проблеми західноукраїнського прикордоння”] (Львів, 1997 р.). – Львів : Ред.-вид. відд. ЛДУ ім. І.Франка, 1997. - С.100-105.
8. Грицевич В.С. Геостатистика у регіональних соціально-економічних дослідженнях : матеріали наук. – практ. семінару [“Науково-практичний семінар з проблем розвитку регіональної статистики в Україні”] (Львів - Київ, 1997 р.). - Львів, Київ : ІРД НАН України, 1997. - С.77-78.
9. Грицевич В.С. Територіальні індекси як засіб регіональних економіко-статистичних досліджень /В.С. Грицевич // Формування нової парадигми економічної освіти в Україні. - Львів, 2000. - С.156-158.

ЗМІСТ

Навчально-методичне видання

Грицевич Володимир Степанович

Математичні методи в суспільній географії:
тексти лекцій для студентів
напряму підготовки 6.040104 − географія

Підготовлено до друку 19.04.2016 формат 60*84/16
Умовн. друк.арк.
Наклад 50 прим.

Малий видавничий центр
Лабораторія тематичного картографування географічного факультету,
Львівський національний університет імені Івана Франка
Україна, 79000, Львів, вул. П.Дорошенка, 41