КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
ІМЕНІ ТАРАСА ШЕВЧЕНКА
ФАКУЛЬТЕТ КОМП'ЮТЕРНИХ НАУК ТА КІБЕРНЕТИКИ
КАФЕДРА ТЕОРЕТИЧНОЇ КІБЕРНЕТИКИ

Пашко А.О.

СТАТИСТИЧНИЙ АНАЛІЗ ДАНИХ

МЕТОДИЧНІ МАТЕРІАЛИ

до курсу "**ІНТЕЛЕКТУАЛЬНА ОБРОБКА ДАНИХ**"
для студентів

галузь знань	12 «Інформаційні технології»
спеціальність	122 «Комп'ютерні науки»
освітній рівень	бакалавр
освітня програма	«Інформатика»

Київ – 2019

УДК 519.2:681.3

СТАТИСТИЧНИЙ АНАЛІЗ ДАНИХ / Пашко А.О. : Електронне видання, -2019.-55 с.

Затверджено Вченою радою

факультету комп’ютерних наук та кібернетики

Протокол №1 від 16 вересня 2019 року

ЗМІСТ

ВСТУП

Мета дисципліни «Інтелектуальна обробка даних» – вивчення основних та найбільш перспективних напрямків аналізу даних: зберігання інформації, оперативний і інтелектуальний аналіз, а також методів та алгоритмів інтелектуального аналізу; знайомство з актуальними питаннями, що постають при розробці програмних продуктів, що обробляють великі обсяги даних.

Протягом вивчення студенти мають опанувати основні методи та моделі інтелектуального аналізу даних та засоби їх реалізації, навчитися аналізувати та уникати сучасних проблем, пов’язаних із збиранням та обробленням інформації.

Початком інтелектуального аналізу даних є попередній аналіз даних, що базується на методах та алгоритмах статистичного аналізу даних.

Методичні матеріали призначені для засвоєння студентами статистичного аналізу даних і вибору показників для подальшого аналізу і обробки.

В основу методичних матеріалів покладено навчальний посібник:

Бахрушин В.Є. Методи аналізу даних : навчальний посібник для студентів / В.Є. Бахрушин. – Запоріжжя : КПУ, 2011. – 268 с.

1. ДИСПЕРСІЙНИЙ АНАЛІЗ

Дисперсійний аналіз є сукупністю статистичних методів, призначених:

-для перевірки гіпотез про зв'язок між певною ознакою та досліджуваними факторами, які не мають кількісного опису,

- для встановлення ступеня впливу факторів та їх взаємодії.

У спеціальній літературі дисперсійний аналіз часто називають ANOVA (від англомовної назви Analysis of Variations).

Вперше цей метод було розроблено Р. Фішером в 1925 р.

Факторами називають контрольовані чинники, що впливають на кінцевий результат.

Рівнем фактора, або **способом обробки**, називають значення, що характеризують конкретний прояв цього фактора.

Ці значення подають у номінальній або порядковій шкалі вимірювань.

Значення вимірюваної ознаки називають **відгуком**.

Часто вихідні значення факторів вимірюють у кількісних або порядкових шкалах.

Тоді постає проблема групування вихідних даних у ряди спостережень, що відповідають приблизно однаковим значенням фактора.

Якщо кількість груп взяти дуже великою, то кількість спостережень у них може виявитися недостатньою для отримання надійних результатів.

Якщо її взяти малою, це може призвести до втрати суттєвих особливостей впливу досліджуваного фактора на систему.

Кількість і розміри інтервалів при однофакторному аналізі найчастіше визначають за принципом рівних інтервалів або за принципом рівних частот.

При багатофакторному аналізі застосовують три типи групування:

– групи з рівною кількістю спостережень;

– групи з різною кількістю спостережень;

– групи, кількості спостережень у яких відповідають певній пропорції.

При цьому існують певні особливості обробки даних, залежно від типу групування.

Однофакторний аналіз. Основною метою однофакторного аналізу є оцінка величини впливу конкретного фактора на досліджуваний відгук.

Іншою метою може бути порівняння двох або декількох факторів один з одним з метою визначення різниці їх впливу на відгук, яку часто називають **контрастом факторів**.

Попереднім етапом є перевірка нульової гіпотези про відсутність будь-якого впливу досліджуваного фактора (факторів), тобто гіпотези про те, що зміни значень ознаки в порівнюваних вибірках є випадковими, і всі дані належать до однієї генеральної сукупності.

Якщо нульову гіпотезу відкидають, то наступним етапом є кількісне оцінювання впливу досліджуваного фактора і побудова довірчих інтервалів для отриманих характеристик.

У випадку, коли нульова гіпотеза не може бути відкинутою, її приймають і роблять висновок про відсутність впливу.

Якщо є підстави вважати, що такий вплив має бути присутнім (наприклад, це може випливати з теоретичних уявлень про об’єкт дослідження), то необхідно перевірити наявність інших факторів, що можуть його маскувати.

При **однофакторному дисперсійному аналізі** вихідні дані подають у вигляді таблиць, у яких кількість стовпчиків дорівнює кількості рівнів фактора, а кількість значень у кожному стовпчику – кількості спостережень при відповідному рівні фактора (таблиця).

Для різних рівнів фактора кількість спостережень може бути різною.

При цьому виходять з припущення, що результати спостережень для різних рівнів є вибірками з нормально розподілених сукупностей, середні значення та дисперсії яких є однаковими і не залежать від рівнів.

Завданням аналізу є перевірка нульової гіпотези про рівність середніх значень сукупностей, що розглядаються.

Форма таблиці спостережень при проведенні однофакторного дисперсійного аналізу

Результати вимірювань	Рівні фактора			
	1	2	…	*k*
1	x_{11}	x_{12}	…	x_{1k}
2	x_{21}	x_{22}	…	x_{2k}
…	…	…	…	…
n_i	$x_{n_1 1}$	$x_{n_1 2}$	…	$x_{n_1 k}$

Метод базується на основній тотожності дисперсійного аналізу.

Сума квадратів відхилень спостережень від загального середнього **(загальна варіація**) дорівнює:

$$\sum_{j=1}^{k}\sum_{i=1}^{n_j}\left(x_{ij}-\bar{x}\right)^2=\sum_{j=1}^{k}n_j\left(\left\langle x_j\right\rangle-\bar{x}\right)^2+\sum_{j=1}^{k}\sum_{i=1}^{n_j}\left(x_{ij}-\left\langle x_j\right\rangle\right)^2,$$

$\bar{x}=\frac{1}{N}\sum_{j=1}^{k}\sum_{i=1}^{n_j}x_{ij}$ – загальне середнє;

$N=\sum_{j=1}^{k}n_j$ – загальна чисельність;

k – кількість вибірок;

nj (j = 1, 2, …, *k*) – кількість елементів у *j*-й вибірці;

$\left\langle x_j\right\rangle=\frac{1}{n_j}\sum_{i=1}^{n_j}x_{ij}\quad j=1,\ ...,\ k;$ – середнє значення *j*-ї вибірки.

Перший доданок (**факторна**, або **міжгрупова варіація**) є зваженою сумою квадратів відхилень групових середніх від загального середнього. Він характеризує коливання значень, зумовлені фактором, на основі якого здійснено групування даних.

Другий доданок (**залишкова**, або **внутрішньогрупова варіація**) є сумою квадратів відхилень спостережень від відповідних групових середніх.

Він характеризує коливання значень досліджуваної ознаки, зумовлені неврахованими факторами або випадковими чинниками.

Сутність методу полягає в тому, що за умови правильності нульової гіпотези величини є незміщеними оцінками дисперсії похибок спостережень σ^2 і мають бути приблизно рівними одна одній.

Перша з них є мірою варіації всередині вибірок і не пов'язана з припущенням про рівність середніх значень, тому σ^2 ≈ σ1^2 незалежно від справедливості нульової гіпотези.

Друга оцінка характеризує варіацію між вибірками. При справедливості нульової гіпотези σ2^2 ≈ σ^2 , а при її порушенні величина σ2^2 є тим більшою, чим більше відхилення від неї.

Значення критерію розраховують за формулою:

Ця величина має *F*-розподіл Фішера з параметрами $k-1$ та $N-k$.

Нульову гіпотезу відхиляють, якщо ймовірність $P(F \geq F^*)$ є достатньо малою, де F^* – значення, розраховане за емпіричними даними (σ2^2).

Непараметричним аналогом однофакторного дисперсійного аналізу є **ранговий однофакторний аналіз Краскела – Уолліса**.

Цей критерій призначено для перевірки нульової гіпотези про рівність ефектів впливу на досліджувані вибірки з невідомими, але рівними середніми. При цьому кількість вибірок має бути більшою ніж дві.

Нульова гіпотеза полягає в тому, що k вибірок обсягами $n1, n2, \ldots, nk$ отримані з однієї і тієї самої генеральної сукупності.

Критерій Краскела – Уолліса є узагальненням *U*-критерію Манна – Уїтні на випадок, коли кількість вибірок $k > 2$.

Розроблений американськими математиком Вільямом Краскелом та економістом Вільсоном Уоллісом в 1952 р.

Рангові методи, у тому числі й метод Краскела – Уолліса, не передбачають нормальності розподілу результатів спостережень і можуть застосовуватися як для кількісних даних з невідомим законом розподілу, так і для порядкових ознак.

У таблицю замість спостережень заносять їх ранги *rij*, отримані шляхом впорядкування за зростанням усієї сукупності спостережень *xij*.

Загальний вигляд вихідної таблиці рангового однофакторного аналізу

№ результату	№ вибірки			
	1	2	…	*k*
1	r_{11}	r_{12}	…	r_{1k}
2	r_{21}	r_{22}	…	r_{2k}
…	…	…	…	…
n_i	$r_{n_1 1}$	$r_{n_1 2}$	…	$r_{n_1 k}$

Для кожного рівня фактора, тобто для кожного стовпця, розраховують суму рангів

$$R_j = \sum_{i=1}^{n_j} r_{ij}, \qquad \langle R_j \rangle = \frac{1}{n_j} \sum_{i=1}^{n_j} r_{ij}. \qquad N = \sum_{i=1}^{k} n_i$$

Для контролю можна використовувати тотожність:

$$\sum_{i=1}^{k} R_i = \frac{N(N+1)}{2},$$

Якщо між стовпцями немає систематичної різниці, то середні ранги будуть близькими до середнього рангу, розрахованого за усією сукупністю, який дорівнює $(N+1)/2$.

Тому величини $Rj - (N+1)/2$ мають бути відносно малими, якщо нульова гіпотеза є правильною.

Обчислення критерію здійснюють за формулами:

$$\mathrm{H} = \frac{12}{N(N+1)} \sum_{j=1}^{k} \frac{R_j^2}{n_j} - 3(N+1),$$

або

$$\mathrm{H} = \frac{12}{N(N+1)} \sum_{j=1}^{k} n_j \left(\langle R_j \rangle - \frac{N+1}{2} \right)^2.$$

За $ni \geq 5$ й $k \geq 4$ статистика критерію асимптотично наближається до $\chi 2$-розподілу з кількістю степенів свободи $k - 1$.

Нульову гіпотезу відхиляють на рівні значущості α, якщо $H > \chi^2_{1-\alpha}$

де $\chi^2_{1-\alpha}$ – квантиль рівня $1-\alpha$ розподілу $\chi 2$ з $k - 1$ степенем свободи.

При $k = 2$ статистика Краскела – Уолліса стає еквівалентною статистиці W Уїлкоксона.

Критерій Джонкхієра (Джонкхієра – Терпстра) запропонований незалежно один від одного нідерландським математиком Т.Дж. Терпстрою в 1952 р. й британським психологом Е.Р. Джонкхієром в 1954 р.

Його застосовують тоді, коли заздалегідь відомо, що наявні групи результатів упорядковані за зростанням впливу досліджуваного фактора, який вимірюють у порядковій шкалі.

***M*-критерій Бартлетта** запропонований британським статистиком Маурісом Стівенсоном Бартлеттом в 1937 р.

Його застосовують для перевірки нульової гіпотези про рівність дисперсій кількох нормальних генеральних сукупностей, з яких взяті досліджувані вибірки, що у загальному випадку мають різні обсяги (обсяг кожної вибірки має бути не менше чотирьох).

***G*-критерій Кокрена (Кочрена)** запропонований американським статистиком Вільмом Геммелом Кочреном в 1941 р.

Його використовують для перевірки нульової гіпотези про рівність дисперсій k ($k \geq 2$) нормальних генеральних сукупностей за незалежними вибірками рівного обсягу.

Непараметричний **критерій Левене**, запропонований американським математиком Ховардом Левене в 1960 р. є альтернативою критерію Бартлетта в умовах, коли немає впевненості у тому, що досліджувані вибірки підпорядковуються нормальному розподілу.

Критерії дають змогу встановити різницю дисперсій сукупностей, але не дають можливості дати кількісну оцінку впливу фактора на досліджувану ознаку, а також встановити, для яких саме сукупностей дисперсії є різними.

Критерій Кокрена

Дано k вибірок рівного об'єму: $x_1^n, \cdots, x_k^n$.

Позначимо s_i^2 - вибіркову оцінку дисперсії i-ї вибірки.

Гіпотеза H_0 - дисперсії всіх вибірок рівні: $\sigma_1 = \cdots = \sigma_n$.

Статистика критерія $$g = \frac{\max\limits_{1 \le i \le k} s_i^2}{\sum\limits_{i=1}^{k} s_i^2}$$.

Якщо $g > g_\alpha(k,n)$, то нульова гіпотеза відхиляється.

Квантилі розподілу знаходяться за таблицями F-розподілу

$$g_\alpha(k,n) = \frac{F_{\frac{k-1+\alpha}{k}}(n-1,(n-1)(k-1))}{k-1+F_{\frac{k-1+\alpha}{k}}(n-1,(n-1)(k-1))},$$

де $F_\gamma(f_1,f_2)$ - γ-квантиль F-розподілу з f_1 та f_2 степенями свободи.

Критерій Кохрена *G*

Рівень значимості $\alpha = 0,01$								
$k \backslash f$	1	3	6	10	16	36	144	∞
2	0,9999	0,9794	0,9172	0,8539	0,7949	0,7067	0,6062	0,5000
3	0,9933	0,8831	0,7606	0,6743	0,6059	0,5153	0,4230	0,3333
4	0,9676	0,7814	0,6410	0,5536	0,4884	0,4057	0,3251	0,2500
5	0,9279	0,6957	0,5531	0,4697	0,4094	0,3351	0,2644	0,2000
6	0,8828	0,6258	0,4866	0,4084	0,3529	0,2858	0,2229	0,1667
7	0,8376	0,5685	0,4347	0,3616	0,3105	0,2494	0,1929	0,1429
8	0,7945	0,5209	0,3932	0,3248	0,2779	0,2214	0,1700	0,1250
9	0,7544	0,4810	0,3592	0,2950	0,2514	0,1992	0,1521	0,1111
10	0,7175	0,4469	0,3308	0,2704	0,2297	0,1811	0,1376	0,1000
15	0,5747	0,3317	0,2386	0,1918	0,1612	0,1251	0,0934	0,0667
20	0,4799	0,2654	0,1877	0,1501	0,1248	0,0960	0,0709	0,0500
30	0,3632	0,1913	0,1327	0,1054	0,0867	0,0658	0,0480	0,0333
40	0,2940	0,1508	0,1033	0,0816	0,0668	0,0503	0,0363	0,0250
60	0,2151	0,1069	0,0722	0,0567	0,0461	0,0344	0,9245	0,0167
120	0,1225	0,0585	0,0387	0,0302	0,0242	0,0178	0,0125	0,0083
∞	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000

Для встановлення кількісного впливу досліджуваного фактора часто застосовують **адитивну модель**, яка передбачає, що значення відгуку є сумою впливу фактора і незалежної від нього випадкової величини:

$$x_{ij} = a_j + \varepsilon_{ij} \quad (j = 1, ..., k;\ i = 1, ..., n),$$

де aj – невипадкові невідомі величини, що визначаються значеннями рівнів фактора;

εij – незалежні випадкові величини, які мають однаковий розподіл і відображають внутрішню мінливість, що не пов'язана із значеннями рівнів фактора.

Модель можна записати у вигляді:

$$x_{ij} = \mu + \tau_j + \varepsilon_{ij} \quad (j = 1, ..., k;\ i = 1, ..., n),$$

де $\mu = \frac{1}{k}\sum_{j=1}^{k} a_j$ – середній рівень;

$\tau_j = a_j - \mu$ – відхилення від середнього рівня при *j*-му значенні рівня фактора.

У такій формі модель має на один невідомий параметр більше (середній рівень і *k* значень відхилень від нього), але кількість незалежних невідомих параметрів залишилася рівною *k*, оскільки відхилення пов'язані співвідношенням $\sum_{j=1}^{k} \tau_j = 0$.

2. КОРЕЛЯЦІЙНИЙ АНАЛІЗ

Кореляцією (кореляційним зв'язком) між випадковими величинами (ознаками) називають наявність статистичного або ймовірнісного зв'язку між ними. При цьому закономірна зміна певних ознак призводить до закономірної зміни середніх значень інших, пов'язаних з ними ознак.

Кореляційним аналізом називають сукупність методів виявлення кореляційного зв'язку.

Тому його можна застосовувати для формалізованого подання моделей зв'язків між окремими компонентами системи або між окремими процесами, що відбуваються в ній.

Наявність кореляційного зв'язку не означає існування причинно-наслідкового зв'язку між досліджуваними ознаками. Вона може бути зумовлена тим, що обидві ознаки мають причинно-наслідковий зв'язок з певним іншим фактором.

Наприклад, існує кореляція між цінами на нафту й на золото. Проте вона пояснюється тим, що обидві ціни виражаються у доларах США й

залежать від динаміки його індексу. Кореляція також може бути випадковою.

Сучасну класифікацію мір подібності запропонували австрійський та американський біостатистик та антрополог Роберт Сокал та британський таксономіст Пітер Сніс у 1963 р.

Згідно з нею виокремлюють такі типи мір подібності:

– міри асоціації, що відбивають різні співвідношення кількості ознак, що збігаються до загальної кількості ознак, а також близькі до них коефіцієнти спряженості (квантифіковані коефіцієнти зв'язку);

– вибіркові коефіцієнти зв'язку типу кореляції (нормовані косинусні міри);

– показники відстані у метричному просторі.

Перевірку зв'язку можна здійснювати лише для пов'язаних вибірок. Це означає, що між елементами обох досліджуваних вибірок існує взаємно однозначна відповідність, а кількість елементів у вибірках є однаковою.

Замість гіпотези про наявність кореляційного зв'язку часто розглядають протилежну гіпотезу про відсутність зв'язку між досліджуваними величинами. Нехай ознака A має r рівнів $A1, A2, \ldots Ar$, а ознака B – s рівнів: $B1, B2, \ldots Bs$.

Їх вважають незалежними, якщо події "ознака A набуває значення Ai" та

"ознака B набуває значення Bj" є незалежними для всіх можливих пар i, j, тобто:

$$P(Ai, Bj)=P(Ai)P(Bj).$$

Це можна сформулювати в інший спосіб: ознаки є незалежними, якщо значення ознаки A не впливає на ймовірності реалізації можливих значень ознаки B: $P(Bj / Ai) = P(Bj)$, $\forall(Ai, Bj)$.

Кореляційний аналіз здійснюють на початковому етапі вирішення всіх основних задач статистичного аналізу даних.

У задачах статистичного аналізу залежностей і побудови регресійних моделей він дає змогу встановити сам факт існування зв’язку між змінними та оцінити ступінь його прояву.

У задачах класифікації даних за допомогою кореляційного аналізу отримують вихідну інформацію у вигляді коваріаційних і кореляційних матриць та інших характеристик парних порівнянь. Це дає змогу визначити подібні один до одного або до певних еталонів об’єкти, сформувати класи подібних об’єктів і здійснити класифікацію.

У задачах зменшення розмірності досліджуваного простору ознак також за допомогою коваріаційних і кореляційних матриць визначають ознаки, що можуть бути без втрати суттєвої інформації подані через інші наявні дані.

Загальна методика перевірки гіпотези про існування зв’язку між ознаками передбачає етапи:

- визначення типу даних;

- перевірку гіпотези про відсутність зв’язку і, в разі її відхилення, оцінювання сили зв’язку.

Тип вихідних даних суттєво впливає на вибір методів і критеріїв, які можна застосовувати на наступних етапах аналізу.

Для визначення сили зв’язку використовують різноманітні показники. Їх прагнуть вибрати такими, щоб вони змінювалися від –1 до +1 або від 0 до 1.

Значення, що є близькими за модулем до одиниці, свідчать про наявність сильного зв’язку. Близькі до нуля значення вказують або на відсутність будь-якого зв’язку, або на відсутність зв’язку того типу (найчастіше лінійного), для якого розроблено відповідний коефіцієнт.

Знак коефіцієнта вказує на напрям зв’язку: прямий (для додатних значень) або зворотний (для від’ємних).

Кореляційний аналіз кількісних ознак. Універсальною характеристикою ступеня тісноти зв'язку між кількісними ознаками є коефіцієнт детермінації.

Вибірковий коефіцієнт детермінації певної ознаки *y* за вектором незалежних ознак X = (*x*(1) , *x*(2) ,..., *x*(*p*)) можна розрахувати як:

$$K_d(y;X) = 1 - \frac{s_\varepsilon^2}{s_y^2},\ s_y^2 = \frac{1}{n}\sum_{k=1}^{n}(y_k - \bar{y})^2,$$

n – кількість спостережень.

Вибіркове значення дисперсії нев'язок обчислюється за однією з формул:

$$s_\varepsilon^2 = \frac{1}{n}\sum_{k=1}^{n}(y_k - f(X_k))^2,$$

$$s_\varepsilon^2 = \frac{1}{m}\sum_{j=1}^{m}\frac{1}{\nu_j}\sum_{i=1}^{\nu_j}(y_{ij} - \bar{y}_j*)^2, \qquad \bar{y}_j* = \frac{1}{\nu_j}\sum_{i=1}^{\nu_j} y_{ji}$$

де $f(Xi)$ - є статистичною оцінкою значення функції регресії $f(X)$ у точці *Xi*,

νj – кількість даних, що потрапили до *j*-го інтервалу групування;

yji – значення *i*-го спостереження досліджуваної ознаки, що потрапило до *j*-го інтервалу;

yj* – її середнє значення за спостереженнями, які потрапили до *j*-го інтервалу;

m – кількість інтервалів.

Першу формулу застосовують у випадку, коли за результатами попереднього аналізу встановлено, що умовна дисперсія

$$D(\varepsilon|X) = \sigma_\varepsilon^2 = const,$$

тобто не залежить від X.

Другу формулу використовують, якщо ця умова не виконується, а також у всіх випадках, коли обчислення здійснюють за згрупованими даними.

У цьому випадку необхідно попередньо здійснити групування даних. Для цього їх впорядковують за зростанням значень однієї з ознак (ознаки X).

Потім задають кількість та межі інтервалів для цієї ознаки.

Підраховують кількості точок, що потрапили до кожного інтервалу ν_j, для змінної *Y* обчислюють загальне середнє та середні за інтервалами, далі розраховують значення коефіцієнта детермінації за формулами.

Величина коефіцієнта детермінації може змінюватися в межах від нуля до одиниці й відображає частку загальної дисперсії досліджуваної ознаки, яка зумовлена зміною функції регресії $f(X)$.

При цьому нульове значення коефіцієнта детермінації відповідає відсутності будь-якого зв'язку, а його рівність одиниці – наявності строго функціонального зв'язку.

Коефіцієнт є універсальним показником зв'язку, він відбиває й такі зв'язки, що є немонотонними функціями.

Тому питання напряму зв'язку у цьому випадку не має сенсу.

Слід зазначити, що для обмеженого набору даних часто можна побудувати декілька різних адекватних регресійних моделей.

Групування даних також можна здійснювати різними способами.

Тому існує певна невизначеність коефіцієнтів детермінації: при застосуванні різних регресійних моделей або різних способів групування ми будемо отримувати дещо різні значення коефіцієнта детермінації.

Інші поширені характеристики ступеня тісноти зв'язку між ознаками можна розглядати як окремі випадки коефіцієнта детермінації, що отримані для конкретних математичних моделей зв'язку.

Розрізняють парні та частинні кореляційні характеристики.

Парні характеристики розраховують за результатами вимірювань тільки досліджуваної пари ознак. Тому вони не враховують опосередкованого або спільного впливу інших ознак.

Частинні характеристики є очищеними від впливу інших факторів, але для їх розрахунку необхідно мати вихідну інформацію не тільки про досліджувані ознаки, а й про всі інші, вплив яких необхідно усунути.

Для кількісних ознак найчастіше застосовують коефіцієнти кореляції Пірсона і Фехнера.

Коефіцієнт кореляції Пірсона (коефіцієнт кореляційного відношення Пірсона, парний коефіцієнт кореляції, вибірковий коефіцієнт кореляції, коефіцієнт Бравайса – Пірсона) вимірює ступінь лінійного кореляційного зв'язку між кількісними скалярними ознаками.

Коефіцієнт розраховують за формулою:

$$R_{xy} = \frac{\sum_{i=0}^{n-1}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=0}^{n-1}(x_i - \overline{x})^2 \sum_{i=0}^{n-1}(y_i - \overline{y})^2}}.$$

Він був запропонований К. Пірсоном у 1896 р. Часто, посилаючись на згадування К. Пірсона про ідеї математичного подання зв'язку, висловлені в 1846 р. відомим французьким фізиком та кристалографом Огюстом Браве, цей показник називають коефіцієнтом Бравайса – Пірсона (Бравайс – це викривлена транскрипція від французького Bravais, що закріпилася в літературі з кореляційного аналізу).

Коефіцієнт Пірсона можна виразити також через дисперсії σy і $\sigma \Delta y$, друга з яких характеризує розкид емпіричних точок стосовно рівняння лінійної регресії $y = ax + b$, де a та b – коефіцієнти, визначені за методом найменших квадратів: $r = \left(1 + \left(\frac{\sigma \Delta y}{\sigma y}\right)^2\right)^{-\frac{1}{2}}$.

За умови достатньо великого обсягу спостережень ($N \geq 30$) стандартне відхилення коефіцієнта кореляції Пірсона можна визначити за формулою:

$$\sigma_r = \frac{1 - r^2}{\sqrt{N}}.$$

На рівні значущості 0,01 гіпотезу про наявність кореляційного зв'язку приймають, якщо $\frac{|r|}{\sigma_r} \geq 2.6$.

Застосування коефіцієнта Пірсона як міри зв'язку є обґрунтованим лише за умови, що спільний розподіл пари ознак є нормальним. Тому перед його розрахунком слід перевірити виконання цієї гіпотези. Якщо вона справедлива, то квадрат коефіцієнта кореляції Пірсона дорівнює коефіцієнту детермінації.

Значення коефіцієнта кореляції може змінюватися від −1 до +1. Значення −1 та +1 відповідають чіткій лінійній функціональній залежності, яка в першому випадку є спадною, а у другому – зростаючою. Для функціональної залежності *y = const* коефіцієнт кореляції, як видно з наведеної формули, є невизначеним, оскільки в цьому випадку знаменник дорівнює нулю. Що ближчим є значення коефіцієнта кореляції до −1 або +1, то більш обґрунтованим є припущення про наявність лінійного зв'язку. Наближення його значення до нуля свідчить про відсутність лінійного зв'язку, але не є доказом відсутності статистичного зв'язку взагалі.

Для обох пар вибірок (рис. 1.) є очевидним існування статистичного зв'язку між параметрами *x* та *y*. Але коефіцієнти кореляції для них дорівнюють, відповідно, *r1* = 0,995 і *r2* = 0,006 .

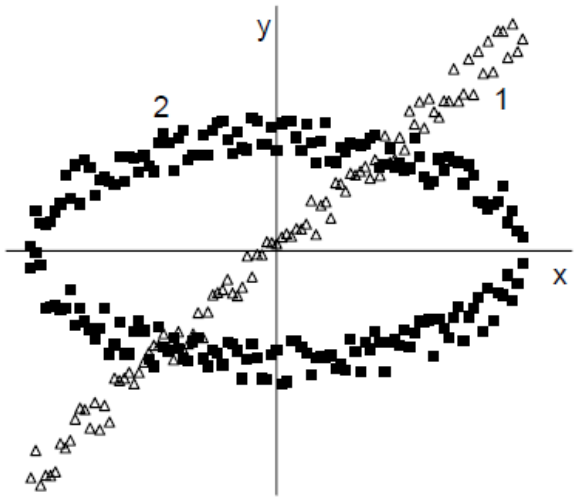


Рис. 1. Приклад різних пар вибірок.

Близькість коефіцієнта кореляції до нуля для другої пари вибірок пов’язана не з відсутністю зв’язку, а з його нелінійністю. Для порівняння, коефіцієнти детермінації для тих самих пар вибірок дорівнюють 0,98 та 1,00.

Приклад свідчить, що в багатьох випадках для попереднього аналізу припущення про наявність і тип зв’язку між певними ознаками доцільно нанести наявні дані на графік.

Як видно, близькість коефіцієнта кореляції Пірсона до нуля в загальному випадку не є доказом незалежності ознак. Але можна довести, що у випадку, коли сумісний розподіл випадкових величин (x, y) є нормальним, рівність $r = 0$ свідчить про статистичну незалежність x і y.

Коефіцієнт кореляції Пірсона часто розглядають як універсальну міру кореляційного зв’язку.

Але, як випливає з наведених вище даних, насправді сфера його обґрунтованого застосування є досить вузькою, оскільки лінійність залежності й нормальний розподіл даних навколо неї є скоріше винятком, ніж правилом.

При дослідженні багатовимірних сукупностей випадкових величин із коефіцієнтів кореляції, обчислених попарно між ними, можна побудувати

квадратну симетричну кореляційну матрицю з одиницями на головній діагоналі. Вона є основним елементом при побудові багатьох алгоритмів багатовимірної статистики, наприклад у факторному аналізі. Довірчий інтервал вибіркової оцінки коефіцієнта кореляції для двовимірної нормальної генеральної сукупності:

$$r \in \left[\tanh\left(z(r) - \frac{N\frac{1+p}{2}}{\sqrt{n-3}} \right); \tanh\left(z(r) + \frac{N\frac{1+p}{2}}{\sqrt{n-3}} \right) \right],$$

де n – обсяг вибірки;

$N\frac{1+p}{2}$ - квантиль нормального розподілу;

p – значення довірчого рівня;

z(r) – z-перетворення (перетворення Фішера) вибіркового коефіцієнта кореляції r.

Коефіцієнт кореляції Пірсона можна застосовувати для перевірки гіпотези про значущість зв'язку. Для нормально розподілених вихідних даних величину вибіркового коефіцієнта кореляції вважають значимо відмінною від нуля, якщо виконується нерівність:

$$r^2 > \left[1 + \frac{n-2}{t_\alpha^2} \right]^{-1},$$

де t_α - критичне значення *t*- розподілу з *(n-2)* степенями свободи.

У випадку, коли між двома наборами ознак існує нелінійний зв'язок, для оцінювання ступеня його тісноти часто використовують кореляційне відношення, яке було запропоновано К. Пірсоном.

Це можливо, якщо щільність розміщення емпіричних точок на координатній площині дає можливість їх групування за однією із змінних і підрахунку групових середніх значень другої змінної для кожного інтервалу.

Кореляційне відношення є квадратним коренем з відношення факторної варіації ознаки до її загальної варіації.

Тоді кореляційне відношення залежної змінної *y* за незалежною змінною *x* можна розрахувати за формулою:

$$\rho_{yx}^2 = s_{y(x)}^2 / s_y^2,$$

$$s_{y(x)}^2 = \frac{1}{n}\sum_{j=1}^{s} \nu_j \left(\overline{y}_{j*} - \overline{y}\right)^2;$$

$$s_y^2 = \frac{1}{n}\sum_{j=1}^{s}\sum_{i=1}^{\nu_j}\left(y_{ji} - \overline{y}\right)^2;$$

$$\overline{y} = \frac{1}{n}\sum_{j=1}^{s} \nu_j \overline{y}_{j*};$$

$$\overline{y}_{j*} = \left(\sum_{i=1}^{\nu_j} y_{ij}\right) / \nu_j,$$

де *n* – обсяг вибірки;

s – кількість інтервалів групування по вісі абсцис;

vj –кількість точок, що потрапили до *j*-го інтервалу.

Кореляційне відношення збігається з модулем коефіцієнта кореляції між тими самими змінними за наявності лінійного зв'язку, а також за відсутності зв'язку.

В інших випадках воно перевищує модуль коефіцієнта кореляції. Це дає можливість використовувати їх різницю як характеристику ступеня відхилення зв'язку від лінійності.

Для цього розраховують величину:

$$v^2 = \frac{(n-k)\left(\rho_{yx}^2 - r^2\right)}{(k-2)\left(1-\rho_{yx}^2\right)}$$

де *n* – кількість емпіричних точок;

k – кількість невідомих параметрів моделі.

Ця величина приблизно підпорядковується *F*-розподілу з параметрами (s – 2) та (n – s).

Якщо значення $\rho^2_{yx} = s^2_{y(x)} / s^2_y$, перевищує точку v^2_α розподілу *F* (*s* − 2, *n* − *s*) , то гіпотезу про лінійний зв'язок відхиляють на рівні значущості α.

Зауваження. У зв'язку з можливістю різних способів групування даних значення кореляційного відношення, як і значення коефіцієнта детермінації, у загальному випадку є дещо невизначеним.

Коефіцієнт кореляції Фехнера розраховують за формулою:

$$r_F = \frac{C - H}{C + H} = \frac{2C - n}{n} = \frac{2C}{n} - 1$$

де *C* – кількість збігів знаків відхилень варіант від відповідних середніх;
H – кількість знаків, що не збігаються.

Значення коефіцієнта Фехнера можуть змінюватися в межах від −1 до +1.

Як і коефіцієнт Пірсона, він показує наявність лінійного зв'язку: що ближчим до одиниці за модулем є значення коефіцієнта, то сильніший зв'язок.

Малі значення абсолютної величини коефіцієнта свідчать про відсутність лінійного зв'язку, але цього недостатньо для твердження про відсутність будь-якого зв'язку взагалі.

Цей показник було запропоновано німецьким психологом Густавом Фехнером у 1860 р.

Застосування для обчислення коефіцієнта лише кількості збігів або незбігів знаків відхилень від середніх значень можна розглядати як зведення первинної кількісної шкали до номінальної, що має призвести до втрати частини корисної інформації.

Тому цей критерій застосовують досить рідко, але у певних випадках, коли інформація про збіги й незбіги знаків відхилень потрібна й для інших цілей, він може виявитися зручнішим за критерій Пірсона.

Коваріацією називають змішаний момент другого порядку. Її розраховують за формулою:

$$Cov(X,Y)=\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i-\overline{x}\right)\left(y_i-\overline{y}\right)$$

На відміну від інших показників, що характеризують наявність статистичного зв'язку, вона не є безрозмірною величиною.

Також немає будь-яких обмежень на її значення. У загальному випадку за інших рівних умов вона збільшується (за модулем) із зростанням середніх значень досліджуваних показників.

Це робить коваріацію незручною для застосування як показника сили зв'язку. Але у багатьох алгоритмах її використовують як проміжний показник, що застосовують у подальших розрахунках.

При аналізі багатовимірних вибірок часто застосовують коваріаційні матриці:

$$C=\begin{pmatrix} c_{11} & c_{12} & \dots & c_{1n} \\ c_{21} & c_{22} & \dots & c_{2n} \\ \dots & \dots & \dots & \dots \\ c_{n1} & c_{n2} & \dots & c_{nn} \end{pmatrix}$$

де $c_{ij}=\operatorname{cov}\left(x_i,x_j\right)$.

Діагональні елементи матриці є дисперсіями $c_{ii}=\sigma^2\left(x_i\right)$ відповідних рядів спостережень.

Коваріаційна матриця є симетричною.

Кореляційний аналіз порядкових ознак. Під ранговою кореляцією розуміють статистичний зв'язок між порядковими ознаками. Вихідні дані подають у вигляді, де елемент *xik* є рангом *i*-го об'єкта за *k*-ю властивістю.

Таблиця вихідних даних для рангового кореляційного аналізу

Порядковий номер об'єкта	Порядковий номер досліджуваної ознаки						
	0	1	2	…	k	…	p
1	x_{10}	x_{11}	x_{12}	…	x_{1k}	…	x_{1p}
2	x_{20}	x_{21}	x_{22}	…	x_{2k}	…	x_{2p}
…	…	…	…	…	…	…	…
i	x_{i0}	x_{i1}	x_{i2}	…	x_{ik}	…	x_{ip}
…	…	…	…	…	…	…	…
n	x_{n0}	x_{n1}	x_{n2}	…	x_{nk}	…	x_{np}

Завданнями аналізу в цьому випадку можуть бути:

- вивчення структури досліджуваних об'єктів;

- перевірка сукупної узгодженості ознак та умовне ранжирування об'єктів за ступенем тісноти зв'язку кожної з них з іншими ознаками;

- побудова єдиного групового впорядкування об'єктів (задача регресії на порядкових змінних).

У першому випадку кожну послідовність впорядкованих за k-ю ознакою n об'єктів подають як точку

$$\mathbf{X}^{(k)} = \left(x_1^{(k)},\ x_2^{(k)},\ \ldots,\ x_n^{(k)} \right),\ k = 0,\ 1,\ \ldots,\ p$$

в n-вимірному просторі ознак.

Найхарактернішими типами структури є такі.

1. Аналізовані точки рівномірно розкидані по всій області їх можливих значень. Це означає відсутність будь-якого зв'язку між досліджуваними ознаками.

2. Частина точок утворює ядро (кластер) із точок, що розташовані близько одна до одної, а інші випадково розкидані навколо цього ядра. Це відповідає існуванню підмножини узгоджених ознак.

3. Аналізовані точки утворюють декілька кластерів, розташованих відносно далеко один від одного. Це відповідає наявності декількох таких підмножин ознак, що існує істотний статистичний зв'язок між ознаками, які належать до однієї і тієї самої підмножини, і не існує значущого зв'язку між ознаками, які належать до різних підмножин.

Прикладом завдання другого типу є визначення узгодженості думок групи експертів з наступним впорядкуванням їх за рівнем компетентності. Для цього розраховують коефіцієнти конкордації для різних сукупностей досліджуваних змінних.

Вирішення завдань третього типу зводиться до побудови такого впорядкування, яке б у певному значенні було б найближчим до кожного з наданих впорядкувань досліджуваних ознак.

Для цього часто застосовують середнє арифметичне або медіану наявних базових рангів.

Це можна розглядати як задачу найкращого у певному розумінні відновлення невідомого ранжирування за наявними емпіричними даними, що зумовлює можливість її розгляду як задачі регресії.

На практиці використовують:

φ-коефіцієнт Пірсона

коефіцієнт спряженості Крамера

поліхоричний коефіцієнт спряженості Чупрова

коефіцієнт (показник подібності) Жаккара

простий коефіцієнт зустрічальності (показник подібності Сокала й Міченера)

показник подібності Рассела і Рао

коефіцієнт спряженості Бравайса – Пірсона (показник подібності Чупрова)

коефіцієнт асоціації Юла

коефіцієнт колігації Юла

Хеммінгова відстань (метрика Хеммінга)

Множинна кореляція. Про множинну кореляцію мова йде в тому випадку, коли певна ознака може бути пов’язана не з однією, а із сукупністю декількох інших ознак.

У реальних дослідженнях можлива ситуація, коли на певну ознаку може впливати не одна, а декілька інших. В таких випадках парні показники кореляції будуть давати неправильну інформацію щодо наявності зв'язку між відповідними показниками, оскільки ці їх значення будуть викривлятися невраховуваними ознаками.

Для уникнення помилок використовують частинні показники кореляції, що усувають такий вплив.

Ідея введення таких показників вперше була висунута Г.У. Юлом у 1896 р., а пізніше розвинена ним та К. Пірсоном.

3. РЕГРЕСІЙНИЙ АНАЛІЗ

Завданням дослідження складних систем і процесів часто є перевірка наявності й встановлення типу зв'язку між незалежними змінними *xi* (**предикторами**, **факторами**), значення яких можуть змінюватися дослідником і мають певну заздалегідь задану похибку, та залежною змінною (**відгуком**) *z*.

Розв'язання таких завдань є предметом регресійного аналізу. Термін "Регресія" вперше був уведений Ф. Гальтоном наприкінці XIX ст.

На практиці завдання регресійного аналізу зазвичай формулюють так: необхідно підібрати достатньо просту функцію, що в певному розумінні найкращим чином описує наявну сукупність емпіричних даних.

Загальна характеристика методів і задач регресійного аналізу.

Класичний регресійний аналіз включає методи побудови математичних моделей досліджуваних систем, методи визначення параметрів цих моделей і перевірки їх адекватності.

Він припускає, що регресія є лінійною комбінацією лінійно незалежних базисних функцій від факторів з невідомими коефіцієнтами (параметрами). Фактори й параметри є детермінованими, а відгуки –

рівноточними (тобто мають однакові дисперсії) некорельованими випадковими величинами.

Передбачається також, що всі змінні вимірюють у неперервних числових шкалах.

Звичайна процедура класичного регресійного аналізу є такою. Спочатку обирають гіпотетичну модель, тобто формулюють гіпотези про фактори, які суттєво впливають на досліджувану характеристику системи, і тип залежності відгуку від факторів. Потім за наявними емпіричними даними про залежність відгуку від факторів оцінюють параметри обраної моделі. Далі за статистичними критеріями перевіряють її адекватність.

При побудові регресійних моделей реальних систем і процесів вказані вище припущення виконуються не завжди. У більшості випадків їх невиконання призводить до некоректності застосування процедур класичного регресійного аналізу і потребує застосування більш складних методів аналізу емпіричних даних.

Постулат про рівноточність і некорельованість відгуків не є обов’язковим.

У випадку його невиконання процедура побудови регресійної моделі певною мірою змінюється, але суттєво не ускладнюється.

Більш складною проблемою є вибір моделі та її незалежних змінних.

У класичному регресійному аналізі припускають, що набір факторів задається однозначно, всі суттєві змінні наявні в моделі й немає ніяких альтернативних способів обрання факторів.

На практиці це припущення не виконується. Тому виникає необхідність розробки формальних та неформальних процедур перетворення й порівняння моделей.

Для пошуку оптимальних формальних перетворень використовують методи факторного та дискримінантного аналізу.

На сьогодні розроблено комп’ютеризовані технології послідовної побудови регресійних моделей.

Фактори в класичному регресійному аналізі вважають детермінованими, тобто вважається, що дослідник має про них всю необхідну інформацію з абсолютною точністю.

На практиці це припущення часто не виконується. Відмова від детермінованості незалежних змінних зумовлює необхідність застосування моделей кореляційного аналізу.

В окремих випадках можна використовувати компромісні методи **конфлюентного аналізу**, які передбачають можливість нормально розподіленого та усіченого розкиду значень факторів.

Якщо ця умова виконується, побудову моделі можна звести до багаторазового розв’язування регресійної задачі.

Відмова від припущення про детермінованість параметрів моделей у регресійному аналізі призводить до суттєвих ускладнень, оскільки порушує його статистичні основи.

На практиці це припущення виконується не завжди. У деяких випадках можна вважати параметри випадковими величинами із заданими законами розподілу.

Як оцінки параметрів можна брати їх умовні математичні сподівання для відгуків, що спостерігалися.

Умовні розподіли та математичні сподівання розраховують за узагальненою формулою Байєса, тому відповідні методи називають **байєсівським регресійним аналізом**.

Регресійні моделі часто використовують для опису процесів, що залежать від часу.

У певних випадках це зумовлює необхідність переходу від випадкових значень відгуків до випадкових послідовностей, випадкових процесів або випадкових полів.

Однією з поширених і найпростіших моделей такого типу є **модель авторегресії**, згідно з якою відгук залежить не тільки від факторів, але також і від часу.

Якщо останню залежність можна виявити, то проблема зводиться до стандартної задачі побудови регресії для модифікованого відгуку.

В інших випадках необхідно використовувати більш складні прийоми.

Процедури класичного регресійного аналізу припускають, що закон розподілу відгуків є нормальним. Проте на практиці найчастішими є випадки, коли цей закон невідомий чи відомо, що він не є нормальним. Їх дослідження зумовило виникнення непараметричного регресійного аналізу, який не передбачає необхідності попереднього задання функції розподілу.

Важливою проблемою, яка виникає при оцінюванні параметрів регресійних моделей, є наявність грубих помилок серед набору аналізованих даних.

Ці помилки можуть виникати внаслідок неправильних дій дослідника, збоїв у роботі апаратури, неконтрольованих короткотривалих сильних зовнішніх впливів на досліджувану систему тощо.

У таких випадках використовують два підходи, що дають змогу зменшити вплив грубих помилок на результати аналізу.

У першому з них розробляють критерії та алгоритми пошуку помилкових даних. Потім ці дані відкидають.

У другому підході розробляють алгоритми аналізу, які є нечутливими до наявних помилкових даних (алгоритми робастного оцінювання параметрів).

Одним з основних постулатів класичного регресійного аналізу є припущення, що найкращі оцінки параметрів можна одержати, використовуючи метод найменших квадратів. На практиці оцінки,

одержані за допомогою цього методу, часто бувають недостатньо точними і містять великі похибки.

Причиною цього може бути структура регресійної моделі.

Якщо вона є лінійною комбінацією експонент або поліномом високого ступеня, то це призводить до поганої зумовленості матриці системи нормальних рівнянь і нестійкості оцінок параметрів.

Підвищення стійкості оцінок можна досягти шляхом відмови від вимоги щодо їх незміщеності.

Розвиток цього напряму досліджень призвів до виникнення гребеневого, або рідж-регресійного аналізу.

Найчастіше задачу побудови регресійної моделі формулюють так.

Необхідно знайти функцію заданого класу, для якої функціонал:

$$F(\alpha) = \sum_{i=1}^{n} \left(z_i(\alpha, X) - y_i \right)^2 \rightarrow \min .$$

У цьому виразі $z_i(\alpha, X)$ – значення функції, що апроксимує залежність, в *i*-ї точці, *yi* – відповідне значення емпіричної залежності,

α – вектор параметрів, які треба знайти,

X – вектор незалежних змінних.

Одержану функцію *z*(α, *X*) називають (**середньоквадратичною**) **регресійною моделлю**.

Метод її пошуку називають методом найменших квадратів.

Для визначення параметрів регресійних моделей можна розв’язувати задачі мінімізації інших функціоналів, зокрема:

$$F(\alpha) = \sum_{i=1}^{n} \left| z_i(\alpha, X) - y_i \right| \rightarrow \min ;$$

$$F(\alpha) = \max \left| z_i(\alpha, X) - y_i \right| \rightarrow \min .$$

Одержувані при цьому регресійні моделі називають, відповідно, **середньоабсолютними** (**медіанними**) та **мінімаксними**.

Ці моделі найчастіше використовують при побудові робастних алгоритмів регресійного аналізу, але їх практичне застосування обмежується поганою збіжністю таких алгоритмів.

Апроксимуючу функцію у випадку однієї незалежної змінної (моделі простої регресії) часто шукають у вигляді полінома $z(x) = \sum_{j=0}^{M} \alpha_j x^j$,

оберненого полінома $z(x) = \frac{1}{\sum_{j=0}^{M} \alpha_j x^j}$,

експоненціальних або показникових функцій $z = \alpha e^x$, $z = \alpha b^x$;

степеневої функції $z = \alpha x^b$,

лінійно-логарифмічної функції $z = \alpha_1 + \alpha_2 x + \alpha_3 1nx,$

тригонометричного ряду Фур'є.

За наявності декількох незалежних змінних (моделі множинної регресії) найчастіше використовують функції, лінійні як за параметрами, так і за незалежними змінними

$$z = \alpha_0 + \sum_{i=1}^{p} \alpha_i x_i,$$

а також поліноміальні моделі, що є лінійними за параметрами, але нелінійними за незалежними змінними:

$$z = \alpha_0 + \sum_{i=1}^{p} \alpha_i x_i + \sum_{\substack{i,j=1 \\ i \geq j}}^{p} \alpha_{ij} x_i x_j + \sum_{\substack{i,j,k=1 \\ i \geq j \\ j \geq k}}^{p} \alpha_{ijk} x_i x_j x_k + \ldots .$$

Останні відповідають розкладу функції відгуку в ряд Тейлора. Проте можливе й використання для апроксимації інших видів залежностей.

Регресійні моделі називають **лінійними** або **нелінійними**, якщо вони є, відповідно, лінійними або нелінійними за параметрами. При цьому

визначення “лінійна” часто опускають. Значення найвищого степеня предиктора в поліноміальних моделях називають **порядком моделі**.

Наприклад: $z = \alpha_0 + \alpha_1 x + \alpha_2 x^2 + \alpha_3 x^3 + \varepsilon,$

де ε – похибка моделі, є лінійною моделлю третього порядку.

Вибір типу регресійної моделі є нетривіальним завданням. Для моделей, що містять одну незалежну змінну, рекомендують спочатку нанести наявні емпіричні дані на графік. Це дає можливість визначити наявність чи відсутність залежності між досліджуваними величинами, а також зробити певні припущення про тип залежності.

З наведених графіків видно, що ці моделі доцільно будувати у вигляді лінійної, квадратичної та експоненціальної функцій, відповідно.

Визначення типу моделі за графіком емпіричних даних є не настільки очевидним, тому доводиться перевіряти декілька варіантів моделі і вибирати кращий з них за певними критеріями (рис. 2.).

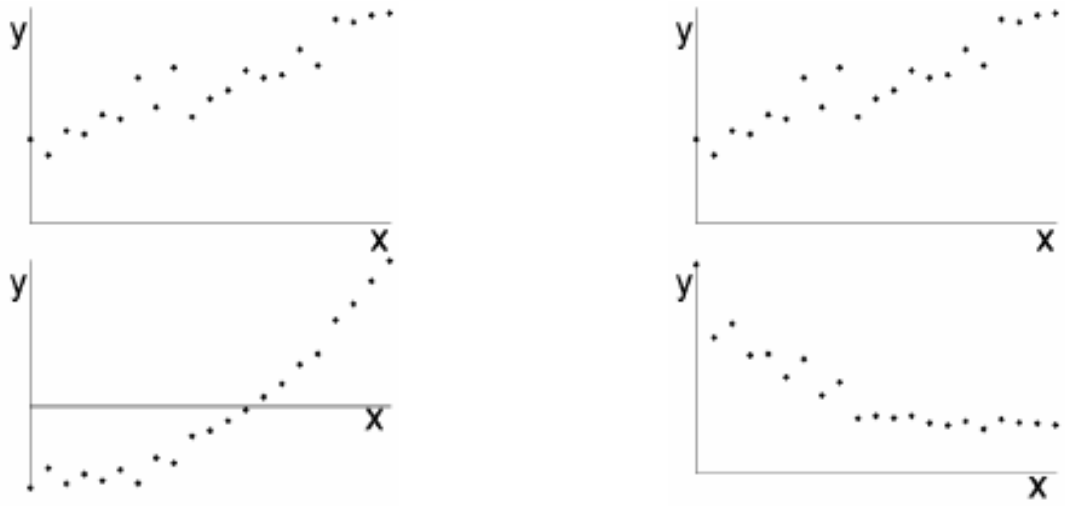


Рис. 2. Графіки емпіричних даних

Часто як попередній етап регресійного аналізу рекомендують за допомогою методів кореляційного аналізу перевіряти наявність значущого зв’язку між досліджуваними змінними.

Але при цьому слід ураховувати, що звичайні методи кореляційного аналізу дають змогу перевіряти лише гіпотезу про наявність лінійного зв'язку. Якщо зв'язок є, але він нелінійний, висновки, отримані за допомогою кореляційного аналізу, можуть бути помилковими.

Важливою особливістю регресійних моделей є те, що їх не можна застосовувати поза межами тієї області значень вихідних параметрів, для якої вони були побудовані.

При використанні регресійних моделей типу полінома, оберненого полінома, тригонометричного ряду та деяких інших слід враховувати, що, збільшуючи кількість членів ряду, можна одержати скільки завгодно близькі до нуля значення відповідних функціоналів.

Проте це не завжди свідчить про якість апроксимації, оскільки ці функціонали не дають інформації про ступінь наближення моделі до емпіричної залежності у проміжках між наявними точками.

Іншою проблемою може бути наявність декількох локальних екстремумів функціоналів.

У таких випадках необхідно враховувати, що більшість стандартних алгоритмів дає можливість знаходити лише локальні, а не глобальні екстремуми функціоналів, і результат мінімізації залежать від вибору початкових умов пошуку. Це часто зумовлює необхідність встановлення додаткових критеріїв вибору моделі, серед яких можуть бути як формальні критерії їх адекватності, так і неформальні критерії, що ґрунтуються на сукупності відомих даних про об'єкт дослідження.

Поліноміальні регресійні моделі, як правило, є формальними.

Їх використовують для опису систем і процесів, теорію яких розроблено недостатньо. При цьому спираються на відомі властивості ряду Тейлора для аналітичних функцій.

На практиці часто доводиться користуватися нелінійними за параметрами та багатовимірними моделями.

Під багатовимірними тут розуміють моделі, що розглядають декілька відгуків.

Задачам, що розв'язуються у межах відповідних напрямів регресійного аналізу, властиві й інші ускладнення.

Так у багатовимірних моделей окремі відгуки можуть бути пов'язані один з одним.

Сама регресійна модель часто задається у неявному вигляді та є неаналітичним розв'язком певної системи алгебраїчних або диференціальних рівнянь.

Нестійкість оцінок параметрів для нелінійних моделей різко зростає. Як правило, такі задачі мають декілька розв'язків або не мають розв'язків взагалі.

Лінійні однофакторні моделі. Найпростішим для аналізу і найбільш дослідженим є випадок лінійної кореляційної залежності між двома змінними *X* та *Y*. Наявність лінійного зв'язку можна перевірити, розрахувавши коефіцієнт парної кореляції Пірсона.

Задача зводиться до підбору параметрів лінійної моделі:

$$z(x) = \alpha_0 + \alpha_1 x + \varepsilon$$

за набором наявних емпіричних точок (*xi*, *yi*).

У **методі найменших квадратів** (МНК) виходять з припущення, що найкращими значеннями параметрів α0 і α1 будуть ті, для яких сума квадратів відхилень емпіричних значень *yi* від розрахункових значень *z*(*xi*) набуває мінімального можливого значення.

МНК оцінки мають такі властивості:

– вони є лінійними функціями результатів спостережень і незміщеними оцінками параметрів моделі;

– згідно з теоремою Гауса – Маркова, МНК оцінки мають найменші дисперсії серед усіх інших оцінок, що є лінійними функціями результатів спостережень;

– МНК оцінки збігаються з оцінками, які обчислюють методом найбільшої правдоподібності.

Для знаходження значень параметрів необхідно розв’язати систему:

$$\begin{cases} \dfrac{\partial}{\partial \alpha_0}\sum_{i=1}^{n}\left[z(x_i)-y_i\right]^2 = \dfrac{\partial}{\partial \alpha_0}\sum_{i=1}^{n}\left[\alpha_1 x_i+\alpha_0-y_i\right]^2 = 0\,, \\ \dfrac{\partial}{\partial \alpha_1}\sum_{i=1}^{n}\left[z(x_i)-y_i\right]^2 = \dfrac{\partial}{\partial \alpha_1}\sum_{i=1}^{n}\left[\alpha_1 x_i+\alpha_0-y_i\right]^2 = 0\,. \end{cases}$$

Вирази для оцінок *α0** і *α1** коефіцієнтів лінійної залежності:

$$\alpha_1^* = \frac{n\sum_{i=1}^{n}x_i y_i - \sum_{i=1}^{n}x_i\sum_{i=1}^{n}y_i}{n\sum_{i=1}^{n}x_i^2 - \left(\sum_{i=1}^{n}x_i\right)^2};$$

$$\alpha_0^* = \overline{Y} - \alpha_1\overline{X} = \frac{\sum_{i=1}^{n}y_i - \alpha_1\sum_{i=1}^{n}x_i}{n} = \frac{\sum_{i=1}^{n}x_i^2\sum_{i=1}^{n}y_i - \sum_{i=1}^{n}x_i\sum_{i=1}^{n}x_i y_i}{n\sum_{i=1}^{n}x_i^2 - \left(\sum_{i=1}^{n}x_i\right)^2}$$

Перше рівняння є відношенням коваріації

$$\operatorname{cov}(x,y) = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i-\overline{X}\right)\left(y_i-\overline{Y}\right)$$

до дисперсії

$$\sigma_X^2 = \frac{1}{(n-1)}\sum_{i=1}^{n}\left(x_i-\overline{X}\right)^2$$

тобто:

$$\alpha_1^* = \frac{\operatorname{cov}(x,y)}{\sigma_x^2} = \frac{\sum_{i=1}^{n}\left(x_i-\overline{X}\right)\left(y_i-\overline{Y}\right)}{\sum_{i=1}^{n}\left(x_i-\overline{X}\right)^2}.$$

У багатьох випадках, завдяки особливостям округлення чисел у ЕОМ, останній вираз дає змогу отримати точніші оцінки параметрів.

У випадку однофакторної лінійної моделі існує зв’язок між коефіцієнтом $a1$ моделі, коефіцієнтом кореляції предиктора і відгуку, а також їх дисперсіями:

$$r_{xy} = a_1 \frac{\sigma_x}{\sigma_y}.$$

Виключення вільного члена з моделі зазвичай є невиправданим. Використання моделі з $\alpha 0 = 0$ доцільно лише у випадках, коли з теорії відомо, що для нульових значень предикторів відгук має дорівнювати нулю.

Якщо це невідомо, але бажано одержати модель, що не містить вільного члена, більш доцільним є застосування центрування даних.

Звідси отримуємо **центровану модель**:

$$Y - \overline{Y} = \alpha_1^* \left(x - \overline{X} \right) + \varepsilon,$$

яка не містить вільного члена.

Незважаючи на те, що, реальні залежності відгуків від факторів є нелінійними, розглянутий випадок широко використовують у практиці побудови регресійних моделей.

Це пов’язано з трьома основними причинами.

По-перше, він є найбільш простим і дослідженим. Зокрема, для нього достатньо повно розроблені процедури визначення статистичних характеристик одержуваних оцінок параметрів (дисперсії, довірчих інтервалів тощо) та перевірки адекватності моделей.

По-друге, у багатьох випадках складні залежності можна подати як набір лінійних (на малих відрізках змінювання факторів) залежностей.

По-третє, нелінійні залежності у деяких випадках можна перетворити до лінійного вигляду шляхом заміни змінних.

Приклади лінеаризації нелінійних залежностей

Вихідна залежність	Лінеаризована залежність	Нові змінні
$z = \alpha_0 exp\,(-\alpha_1 x)$	$\ln z = \ln \alpha_0 - \alpha_1 x$	x, $\ln z$
$z = \alpha_0[1 - exp(-\alpha_1 x)]$	$\ln\frac{\alpha_0}{\alpha_0 - z} = \alpha_1 x$	x, $\ln\frac{\alpha_0}{\alpha_0 - z}$
$z = \alpha_0 exp(-\alpha_1/x)$	$\ln z = \ln \alpha_0 - \alpha_1/x$	$1/x$, $\ln z$
$z = z = \alpha_0 x^{\alpha_1}$	$\ln z = \ln \alpha_0 + \alpha_1 \ln x$	$\ln x$, $\ln z$
$z = \alpha_0 x + \alpha_1 x^2$	$z/x = \alpha_0 + \alpha_1 x$	x, z/x
$z = \alpha_0 \sin(\alpha_1 x)$	$\arcsin(z/\alpha_0) = \alpha_1 x$	x, $\arcsin(z/\alpha_0)$

Перетворення нелінійних залежностей до лінійних є строго обґрунтованим, якщо вихідні дані є точними.

На практиці вони завжди вимірюються з деякою похибкою.

Розглянемо модель: $z = \alpha_0 x^{\alpha 1} + \varepsilon$, де ε – похибка вимірювань.

Її лінеаризована форма матиме вигляд: $\ln z = \ln \alpha_0 + \alpha_1 \ln x + \varepsilon'$, де ε’ є невідомою випадковою величиною.

Використання як лінеаризованої форми виразу:

$$\ln z = \ln \alpha_0 + \alpha_1 \ln x$$

є коректним лише у тому випадку, коли величина ε’ є малою порівняно з іншими доданками правої частини.

Точність оцінок. Розглянемо тотожність:

$$\left(y_i - \overline{Y}\right) = \left(y_i^* - \overline{Y}\right) + \left(y_i - y_i^*\right).$$

*yi** - є оцінкою значення величини *y* при *x* = *xi*.

Піднесемо обидві частини тотожності до квадрата та візьмемо суму від *i* = 1 до *n*, одержимо:

$$\sum_{i=1}^{n}\left(y_i - \overline{Y}\right)^2 = \sum_{i=1}^{n}\left(y_i^* - \overline{Y}\right)^2 + \sum_{i=1}^{n}\left(y_i - y_i^*\right)^2.$$

Ліва частина є сумою квадратів відхилень значень, що спостерігалися, стосовно загального середнього.

Перший доданок правої частини є сумою квадратів відхилень оцінок цих значень, зроблених на основі обраної моделі регресії, від загального середнього. Її часто називають сумою квадратів стосовно регресії.

Другий доданок правої частини є сумою квадратів відхилень значень, що спостерігалися, від їх оцінок, одержаних з використанням обраної моделі.

Цей доданок називають сумою квадратів, що зумовлена регресією.

Для того, щоб модель була придатною для прогнозування значень досліджуваної величини, необхідно, щоб він був малим порівняно із сумою квадратів стосовно регресії.

У граничному випадку він має дорівнювати нулю.

Будемо вважати сталими **дисперсію залишків** $\sigma^2_{\varepsilon_i}$,

дисперсію відгуків $\sigma^2_{Y_i}$

Дисперсія емпіричних точок **стосовно середнього** $\sigma^2_{Y_i}$ буде дорівнювати їх **дисперсії** $\sigma^2_{ХХ}$ **стосовно лінії регресії** у випадку, коли постульована модель є істинною.

У протилежному випадку $\sigma^2_{Y_i} > \sigma^2_{ХХ}$.

Оцінкою величини $\sigma^2_{ХХ}$ є відношення суми квадратів відхилень спостережень від середнього до кількості степенів вільності (рівна різниці між кількістю випробувань і кількістю констант, які визначаються незалежно одна від одної за їх результатами). У випадку, що розглядається, дорівнює $n-2$. Тобто: $\sigma^2_{ХХ} = \dfrac{\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2}{n-2}$.

Нехай *pi* є кількістю повторних вимірювань величини *Yi* при заданому значенні *xi*. Тоді квадратична форма, яку мінімізують в методі найменших квадратів:

$$Q = \sum_{i=1}^{n} \left(\overline{Y_i} - z(x_i) \right)^2 p_i = \sum_{i=1}^{n} p_i \left[\overline{Y_i} - \alpha_0 - \alpha_1 \left(x_i - \overline{x} \right) \right]^2 .$$

Розглядаючи її як функцію параметрів α0, α1, одержуємо оцінки параметрів:

$$\begin{cases} \alpha_0^* = \sum_{i=1}^{n} p_i \overline{Y_i} / \sum_{i=1}^{n} p_i = \overline{Y} ; \\ \alpha_1^* = \sum_{i=1}^{n} p_i \overline{Y_i} \left(x_i - \overline{x} \right) / \sum_{i=1}^{n} p_i \left(x_i - \overline{x} \right)^2 . \end{cases}$$

У припущенні, що умовний розподіл величини *Yi* при заданому *xi* є нормальним, оцінкою дисперсії буде величина:

$$\sigma_{Y_i}^{2\,*} = \frac{1}{n} \sum_{i=1}^{n} p_i \left(\overline{Y_i} - Y_i^* \right)^2 ,$$ де $Y_i^* = \alpha_0^* + \alpha_1^* (x_i - \overline{x_i}).$

Висновки, що одержувані на основі мінімізації дисперсії похибки, є правильними тільки тоді, коли постульована модель коректна. В інших випадках вони можуть виявитися помилковими. Перед прийняттям рішення стосовно моделі треба перевірити гіпотезу, що лінійна модель $z = \alpha_0 + \alpha_1 \left(x - \overline{x} \right)$ задовільно описує емпіричні дані із заданою точністю при заданому рівні значущості η. Для цього визначають міру похибки емпіричних даних:

$$S_a^2 = \frac{1}{n-2} \sum_{i=1}^{n} p_i \left(\overline{Y_i} - Y_i^* \right)^2 .$$

Ця величина є зміщеною оцінкою дисперсії $\sigma_{Y_i}^2$. Її називають **дисперсією неадекватності**.

Незміщеною оцінкою цієї дисперсії є величина:

$$S_e^2 = \frac{\sum_{i=1}^{n} \sum_{j=1}^{p_i} \left(Y_{ij} - \overline{Y_i} \right)^2}{\sum_{i=1}^{n} p_i - n},$$

де *Yij* – *j*-те одиничне вимірювання при *x* = *xi*.

Критерієм адекватності моделі при заданій надійності 1 – η є виконання нерівності: $S_a^2 / S_e^2 \le F_{1-\eta}$, де $F_{1-\eta}$ – відповідне значення функції розподілу Фішера, для кількостей степенів вільності $n1 = n2 = n - 1$.

Довірчі інтервали для параметрів α0, α1 можна знайти за допомогою коефіцієнтів *t*-розподілу Стьюдента з кількістю степенів вільності $\sum_{i=1}^{n} p_i - 2$

$$\begin{cases} \alpha^* - t_{1-\eta/2} S_{\alpha^*} \le \alpha \le \alpha^* - t_{1-\eta/2} S_{\alpha^*}; \\ S_{\alpha^*} = \sqrt{\dfrac{1}{n-2} \sum_{i=1}^{n} p_i \left(\overline{\alpha} - \alpha^*\right)^2}. \end{cases}$$

Важливим практичним завданням є перевірка гіпотези про збіг двох рівнянь регресії: $z_1(x) = \alpha_{01} + \alpha_{11} x;$ та $z_2(x) = \alpha_{02} + \alpha_{12} x.$

Воно передбачає перевірку трьох простих гіпотез. Спочатку перевіряють гіпотезу про рівність дисперсій неадекватності моделей: $H_0^{(1)} : \sigma_{a1}^2 = \sigma_{a2}^2.$

Для цього використовують дисперсійний критерій Фішера.

Якщо різниця дисперсій неадекватності є незначущою, то переходять до перевірки гіпотези про рівність кутових коефіцієнтів моделей:

$$H_0^{(2)} : \alpha_{11} = \alpha_{12}.$$

$$t_a = \frac{a_{11} - a_{12}}{\sigma \sqrt{\dfrac{1}{\sum_{i=1}^{k_1} n_{1i} \left(x_i^{(1)} - \overline{x}^{(1)}\right)^2} + \dfrac{1}{\sum_{i=1}^{k_2} n_{2i} \left(x_i^{(2)} - \overline{x}^{(2)}\right)^2}}},$$

Поліноміальні моделі. У багатьох випадках емпіричні залежності можна описати поліноміальними моделями вигляду: $z = \sum_{i=1}^{q} \alpha_i x^i$.

Оцінки параметрів таких моделей отримують шляхом розв'язування нормальних рівнянь виду:

$$\begin{pmatrix} n & \sum x_i & \sum x_i^2 & \dots & \sum x_i^q \\ \sum x_i & \sum x_i^2 & \sum x_i^3 & \dots & \sum x_i^{q+1} \\ \dots & \dots & \dots & \dots & \dots \\ \sum x_i^q & \sum x_i^{q+1} & \sum x_i^{q+2} & \dots & x_i^{2q} \end{pmatrix} \begin{pmatrix} \alpha_0 \\ \alpha_1 \\ \dots \\ \alpha_q \end{pmatrix} = \begin{pmatrix} \sum Y_i \\ \sum Y_i x_i \\ \dots \\ \sum Y_i x_i^q \end{pmatrix}.$$

Зазвичай стовпці, що утворюють матрицю *X*, не є ортогональними.

У зв'язку з цим у разі необхідності збільшення степеню полінома необхідно перераховувати оцінки всіх його коефіцієнтів. Тому для поліномів високих степенів більш раціональним методом побудови регресійної моделі є заміна вихідного рівняння іншим: $z = \sum_{i=1}^{q} \alpha_i' \zeta_i$, де $\zeta_i = \zeta_i(x)$ є поліномами *i*-го степеню за *x*, які задовольняють умови ортогональності:

$$\begin{cases} \sum_{j=1}^{n} \zeta_{ij} = 0, \quad i = 1, 2, ..., q; \\ \sum_{j=1}^{n} \zeta_{ij} \zeta_{i'j} = 0, \quad i \neq i' \end{cases},$$

ζ_{ij} є *i*-м поліномом для точки *xj*.

Однофакторні моделі інших типів. Апроксимацію емпіричних залежностей тригонометричними багаточленами називають **гармонічним аналізом**.

У цьому випадку модель має вигляд:

$$z(x) = \alpha_0 + \sum_{k=1}^{r} \alpha_k \cos\frac{2\pi kx}{T} + \sum_{k=1}^{r} \beta_k \sin\frac{2\pi kx}{T},$$

де T – період спостереження апроксимованої залежності;

r – кількість гармонік ($r < n/2$);

n – кількість частин, на які розділений період T.

Її параметри визначають за формулами:

$$\alpha_0 = \frac{1}{n}\sum_{k=0}^{n} y_k;$$

$$\alpha_m = \frac{2}{n}\sum_{k=0}^{n} y_k \cos\frac{2\pi km}{n}, \quad m = 1, 2, ..., r;$$

$$\beta_m = \frac{2}{n}\sum_{k=0}^{n} y_k \sin\frac{2\pi km}{n}, \quad m = 1, 2, ..., r,$$

де yk – значення апроксимованої функції у точках k $x_k = \frac{kT}{n}$.

Крива Гомперця описується рівняннями $\hat{y}_t = ab^{c^t}$, $\hat{y}_t = ab^{c^{-t}}$, які логарифмуванням зводяться до узагальненої показникової функції першого або другого типу, відповідно.

Логістична функція $\hat{y}_t = \frac{1}{a + bc^t}$, $\hat{y}_t = \frac{1}{a + bc^{-t}}$ зводиться до модифікованої показникової перетворенням $y^* = 1/\hat{y}_t = a + bc^{\pm t}$.

Модифіковану показникову функцію використовують як модель у випадках, коли досліджуваній залежності властиве насичення, тобто при збільшенні значень незалежної змінної відгук поступово наближається до певного граничного значення, а його прирости наближуються до нуля.

У таких випадках існує певний обмежувальний фактор, вплив якого збільшується із зростанням досягнутого рівня.

Значення рівня насичення, як правило, можна задати, виходячи з наявних даних про об’єкт дослідження.

У такому разі інші параметри моделі можна визначити методом найменших квадратів після її лінеаризації.

Якщо вплив обмежувального фактора виявляється лише після досягнення певного рівня розвитку процесу, слід використовувати **моделі S-подібного зростання**, до яких належать крива Гомперця і логістична функція.

Вони описують процеси, в яких темп зростання поступово збільшується на початкових стадіях і поступово зменшується в кінці.

При цьому слід ураховувати, що крива Гомперця є асиметричною, а логістична крива симетрична стосовно точки перегину.

Процес побудови й дослідження логістичної моделі називають логістичним аналізом.

Логістичну криву часто називають законом зростання, оскільки вона описує залежність кількості популяції або її біомаси від часу.

Лінійні багатофакторні моделі. Лінійну як за параметрами, так і за незалежними змінними регресійну модель можна записати у вигляді:

$$Y = \alpha_0 + \sum_{i=1}^{p} \alpha_j x_j + \varepsilon = X\alpha + \varepsilon,$$

де Y – вектор-стовпчик відгуків, який має розмірність n ($n > p$);

$$X = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ \dots & \dots & \dots & \dots & \dots \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix}$$ – матриця значень p незалежних змінних

при n вимірюваннях;

α – вектор-стовпчик невідомих параметрів моделі, що має розмірність $p + 1$;

ε – вектор-стовпчик похибок моделі, який має розмірність n.

Оцінки параметрів моделі у методі найменших квадратів отримують мінімізацією скалярного добутку: $Q = (Y - X\alpha)^T (Y - X\alpha)$,

де символ “Т” позначає транспонування.

Для лінійної моделі є незміщеною оцінкою з найменшою дисперсією вектора α $\alpha = (X^T X)^{-1} X^T Y$.

Коваріаційною матрицею вектора α є: $\Sigma = \sigma^2 (X^T X)^{-1}$

де σ2 – дисперсія похибки.

Елементами головної діагоналі коваріаційної матриці є дисперсії компонентів вектора α, а позадіагональні компоненти є значеннями відповідних коефіцієнтів коваріації.

Для перевірки значущості регресії використовують F-критерій Фішера, розрахункове значення якого обчислюють за формулою:

$$F = \frac{Q_R / (p+1)}{Q / (n-p-1)},$$

де $Q_R = (X\alpha)^T (X\alpha)$ – сума квадратів відхилень, зумовлена регресією;

Q – сума квадратів відхилень спостережень від регресії.

За умови виконання нульової гіпотези $H_0: \alpha = 0$ $F < F_{кр}$,

де Fкр – критичне значення статистики Фішера для заданого рівня значущості й кількостей степенів свободи $(p+1)$ та $(n-p-1)$.

Значущість окремих коефіцієнтів регресії перевіряють за допомогою критерію: $t_j = \frac{\alpha_j}{\hat{s}\sqrt{(X^T X)^{-1}_{jj}}}$, де $\hat{s} = \sqrt{\frac{1}{n-p-1}Q}$ – незміщена оцінка стандартного відхилення залишків моделі.

За умови виконання нульової гіпотези $H0 : \alpha\, j = 0$ статистика критерію підпорядковується *t*-розподілу Стьюдента з кількістю степенів свободи $n - p - 1$.

Інтервальною оцінкою для коефіцієнта αj є:

$$\alpha_j \in \left[\widehat{\alpha_j} - t\hat{s}\sqrt{(X^T X)^{-1}_{jj}}; \widehat{\alpha_j} + t\hat{s}\sqrt{(X^T X)^{-1}_{jj}}\right].$$

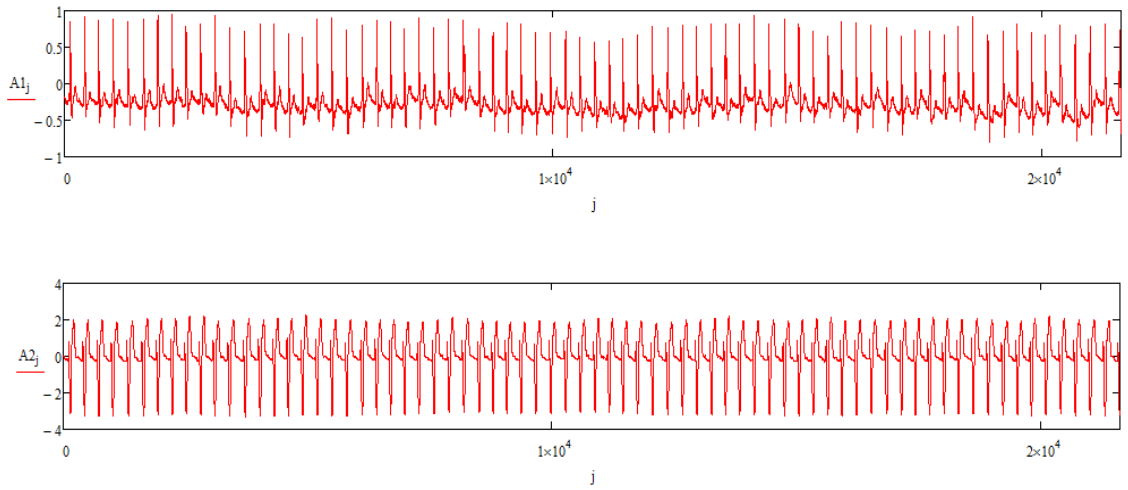
4. ЗАВДАННЯ ДЛЯ ПРОЕКТУВАННЯ

Завдання. У файлі data.txt міститься запис кардіограми людини на двох каналах. Час запису – 60 сек. Дискретність: 360 точок за 1 сек. Структура файлу – час в секундах, 1-й канал, 2-й канал (амплітуда у відносних одиницях).

Довжина запису N=21600, $\Delta t = 1/360 = 0.0028$

Алгоритм обробки

1. **Побудувати графік кардіограми.**



Обчислити середнє значення, дисперсію, нормувати початкові масиви.

Перевірити гіпотезу про нормальний закон розподілу.

2. Однофакторний дисперсійний аналіз.

Перевірити чи є результати вимірювання різними рівнями одного фактору.

Для кожного рівня знаходимо

$$S_i^2 = \frac{1}{n-1}\sum_{j=1}^{n}(x_{ij}-\bar{x}_i)^2 = \frac{1}{n-1}\left[\sum_{j=1}^{n}x_{ij}^2 - \frac{1}{n}\left(\sum_{j=1}^{n}x_{ij}\right)^2\right].$$

За припущенням дисперсійного аналізу - повинна мати місце рівність дисперсій. Перевірити рівність дисперсій за критерієм порівняння.

$$g = \frac{\max\limits_{1\leqslant i\leqslant k} s_i^2}{\sum\limits_{i=1}^{k} s_i^2}.$$

Критерій порівняння подується за формулою

При $g > g_\alpha(k,n)$ нульова гіпотеза про рівність дисперсій відхиляється.

Значення статистики $g_\alpha(k,n)$ для α=0,95

k	n					
	8	9	10	11	17	37
2	0,899	0,882	0,867	0,854	0,795	0,707
3	0,733	0,711	0,691	0,673	0,606	0,515
4	0,613	0,590	0,570	0,554	0,488	0,406
5	0,526	0,504	0,485	0,470	0,409	0,335
6	0,461	0,440	0,423	0,408	0,353	0,286
7	0,410	0,391	0,375	0,362	0,310	0,249
8	0,370	0,352	0,337	0,325	0,278	0,221
9	0,338	0,321	0,307	0,295	0,251	0,199
10	0,311	0,294	0,281	0,270	0,230	0,181
12	0,268	0,253	0,242	0,232	0,196	0,153
15	0,223	0,210	0,200	0,192	0,161	0,125
20	0,175	0,165	0,157	0,150	0,125	0,096
30	0,123	0,116	0,110	0,105	0,087	0,066

При виконанні припущення про рівність дисперсій, знаходимо оцінку дисперсії, що характеризує розсіювання поза впливом фактора,

$$S_0^2 = \frac{1}{k}\sum_{i=1}^{k}S_i^2\frac{1}{k(n-1)}\sum_{i=1}^{k}\sum_{j=1}^{n}(x_{ij}-\bar{x}_i)^2 = \frac{1}{k(n-1)}\left[\sum_{i=1}^{k}\sum_{j=1}^{n}x_{ij}^2 - \frac{1}{n}\sum_{i=1}^{k}\left(\sum_{j=1}^{n}x_{ij}\right)^2\right]$$

Знаходимо вибіркову дисперсію всіх спостережень

$$S^2 = \frac{1}{kn-1}\left[\sum_{i=1}^{k}\sum_{j=1}^{k}x_{ij}^2 - \frac{1}{kn}\left(\sum_{i=1}^{k}\sum_{j=1}^{n}x_{ij}\right)^2\right]$$

Знаходимо оцінку дисперсії, що характеризує зміни параметра, пов'язані з фактором

$$S_A^2 = \frac{n}{k-1}\sum_{i=1}^{k}(\bar{x}_i - \bar{\bar{x}})^2 . \qquad \bar{\bar{x}} = \frac{1}{k}\sum_{i=1}^{k}\bar{x}_i; \quad \bar{x}_i = \frac{1}{n}\sum_{j=1}^{n}x_{ij}.$$

Оцінка впливу фактора на зміни середнього значення визначається відношенням (вплив значущий з ймовірністю 1-α)

$$\frac{S_A^2}{S_0^2} > F_\alpha[k-1; k(n-1)]$$

де $F_\alpha(f_1,f_2)$ - α- квантиль F- розподілу з f_1 та f_2 степенями свободи.

3. Регресійний аналіз.

Знайти оцінки параметрів лінійної регресії $y = \alpha + \beta x$ методом найменших квадратів, де x - 2-й канал, y - 1-й канал.

Перевірити наявність викидів у регресії - якщо $R > R_\delta$, то y_i, що відповідає максимальному значенню відношення $\frac{e_i}{S_i}$ є викидом з достовірністю δ, де $\hat{y}_i = a + bx_i$

$$e_i = y_i - \hat{y}_i,$$

$$R = \max\left|\frac{e_i}{S_i}\right|$$

$$S_i^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-2}\left[1 - \frac{1}{n} - \frac{(x_i - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]$$

Значення $R_\delta \approx 4$.

Якщо є викиди, то їх видалити і провести оцінки спочатку.

Перевірити гіпотези про відповідність оцінок коефіцієнтів істинним значенням та адекватність моделі.

1. $H_0: \beta = b$. Значення коефіцієнта є значимим з достовірністю δ, якщо $|b| > t_{\frac{1+\delta}{2}} S_\beta$,

де $t_{\frac{1+\delta}{2}}$ – коефіцієнт розподілу Стюдента з (*n-2*) степенями свободи,

$$S_\beta = \frac{S}{S_x\sqrt{n-1}}; \quad S^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - a - bx_i)^2;$$

$$S_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2; \quad \bar{x} = \frac{1}{n}\sum_{i=1}^{n}x_i;$$

2. $H_0: \alpha = a$. Значення коефіцієнта є значимим з достовірністю δ, якщо $|a| > t_{\frac{1+\delta}{2}} S_\alpha$,

де $t_{\frac{1+\delta}{2}}$ – коефіцієнт розподілу Стюдента з (*n-2*) степенями свободи,

$$S_\alpha = S\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{(n-1)S_x^2}},$$

3. H_0: **модель адекватна.** Модель адекватна з достовірністю δ, якщо $\frac{S^2}{S_y^2} < F_\delta$,

де F_δ – квантиль розподілу Фішера з (*n-2*) та (*n-1*) степенями свободи.

Побудувати регресійні моделі більш високих порядків.

Вибрати найкращу за коефіцієнтом детермінації

$$R^2 = 1 - \frac{RSS}{TSS},$$

$$RSS = \sum_{i=0}^{n-1}(y_i - \hat{y}_i)^2, \qquad RSS = \sum_{i=0}^{n-1}(y_i - \bar{y}_i)^2,$$

$$\bar{y} = \frac{1}{n}\sum_{i=0}^{n-1} y_i$$

$\hat{y}$ - значення, що обчислене за відповідною моделлю.

4. Кореляційний аналіз

Визначити коефіцієнт кореляції між x та y.

$$R_{xy} = \frac{\sum_{i=0}^{n-1}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=0}^{n-1}(x_i - \bar{x})^2 \sum_{i=0}^{n-1}(y_i - \bar{y})^2}}$$

Оцінити кореляційну функцію параметрів x та y.

$$R_x(j)=\frac{\sum_{i=0}^{n-j-1}\left(x_i-\overline{x}\right)\left(x_{i+j}-\overline{x}\right)}{(n-j-1)\sum_{i=0}^{n-1}\left(x_i-\overline{x}\right)^2},\ j=0,1...n-2$$

Якщо сигнал періодичний – обчислити його характеристики, виділити періодичний тренд.

5. Перетворення Фур'є.

Виконати перетворення Фур'є за формулами

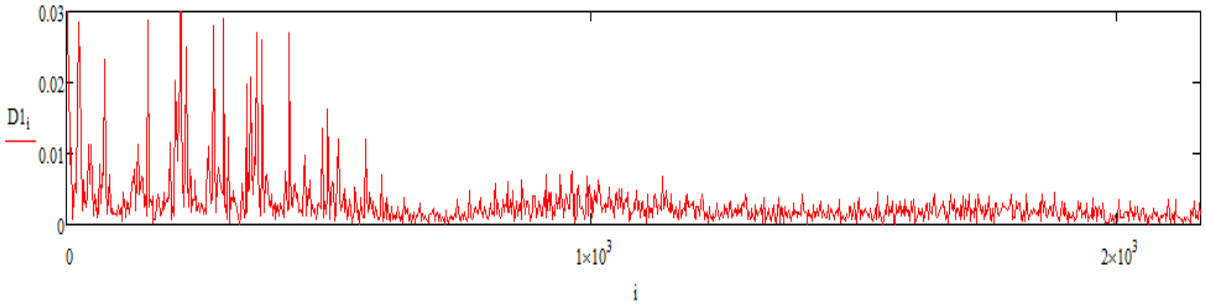
$$A_0 := \frac{1}{N}\cdot\sum_{i=0}^{N-1}\left(s1_i\cdot\cos\left(\frac{2\pi\cdot i\cdot 0}{N}\right)\right) \qquad A_{\frac{N}{2}} := \frac{1}{N}\cdot\sum_{i=0}^{N-1}\left(s1_i\cdot\cos\left(\frac{\pi\cdot i}{1}\right)\right)$$

$$A_l := \frac{2}{N}\cdot\sum_{i=0}^{N-1}\left(s1_i\cdot\cos\left(\frac{2\pi\cdot i\cdot l}{N}\right)\right) \qquad l := 1,2..\frac{N}{2}-1$$

$$B_j := \frac{2}{N}\cdot\sum_{i=0}^{N-1}\left(s1_i\cdot\sin\left(\frac{2\pi\cdot i\cdot j}{N}\right)\right) \qquad j := 0,1..\frac{N}{2}$$

$$C_j := \sqrt{\left(A_j\right)^2+\left(B_j\right)^2} \qquad j := 0,1..\frac{N}{2}$$

Знайти спектр .



Побудувати графік. Обчислити частоту першої синусоїди (або крок по частоті).

Порівняти отримані результати.

Виконати обернене перетворення Фур'є

$$d1_i := \sum_{j=0}^{\frac{N}{2}} \left(A_j \cdot \cos\left(\frac{2\pi j \cdot i}{N}\right)\right) + \sum_{j=0}^{\frac{N}{2}} \left(B_j \cdot \sin\left(\frac{2\pi j \cdot i}{N}\right)\right)$$

.

Порівняти початковий і результуючий масиви. Співпали?

6. Вейвлет аналіз.

Виконати пряме і обернене вейвлет - перетворення для кожного базису.

M1=N,

$$f2(j,k,x) := 2^{\frac{j}{2}} \cdot g1\left(2^j \cdot x - k\right)$$

, $g1(x)$ - батьківський вейвлет,

$$W(1,j) := \sum_{i=0}^{N-1} \left(s1_i \cdot f2(1,j,i)\right)$$

,

В якості батьківського вейвлета розглянути вейвлет Хаара, гауссів вейвлет, похідні від гауссівського вейвлета.

Обернене вейвлет-перетворення.

$$d_i := \sum_{l=0}^{M} \sum_{j=0}^{M1} \left(W(1,j) \frac{f2(1,j,i)}{2^{2l}}\right)$$

.

Порівняти початковий і відновлений масиви.

7. **Підготувати звіт.** У звіті відобразити функції Python, які використовувались для обробки інформації

ВИСНОВКИ

Для отримання додаткової інформації з питань теоретичних основ, програмного забезпечення й практики застосування сучасних методів аналізу даних можна використовувати інформацію, що рекомендована в навчальному підручнику [3]:

1. http://datan.ucoz.ru;
2. http://www.basegroup.ru;
3. http://www.statsoft.ru/home/textbook/default.htm;
4. http://orlovs.pp.ru;
5. http://www.aup.ru/books/m163/;
6. http://www.ami.nstu.ru/~headrd;
7. http://uk.wikipedia.org, http://www.wikipedia;
8. http://www.biometrica.tomsk.ru;
9. http://dvo.sut.ru/libr/opds/i130hodo_part1/index.htm;
10. http://www.dvo.sut.ru/libr/opds/i130hod2/index.htm;
11. http://www.gmdh.net/gmdh.htm;
12. http://www.machinelearning.ru;
13. http://riskcontrol.ru;
14. http://attestatsoft.narod.ru/index.htm;
15. http://www.medstatistica.com.

ЛІТЕРАТУРА

1. Кобзарь А.И. Прикладная математическая статистика. Для инженеров и научных работников / А.И. Кобзарь. – М.:ФИЗМАТЛИТ. - 2006. - 816 с.
2. Закс Л. Статистическое оценивание / Л. Закс. –М.: СТАТИСТИКА. – 1976. – 598 с.
3. Бахрушин В.Є. Методи аналізу даних : навчальний посібник для студентів / В.Є. Бахрушин. – Запоріжжя : КПУ, 2011. – 268 с.
4. Кремер Н.Ш. Теория вероятностей и математическая статистика / Н.Ш. Кремер. - М.: ЮНИТИ-ДАНА, 2010. - 551 с.
5. Гирко В.Л. Многомерный статистический анализ / В.Л. Гирко. –

К. : Высшая школа, 1988. – 320 с.

6. Енюков И.С. Методы, алгоритмы, программы многомерного статистического анализа / И.С. Енюков. – М. : Финансы и статистика, 1986. – 232 с.
7. Іващенко П.О. Багатовимірний статистичний аналіз / П.О. Іващенко, І.В. Семеняк, В.В. Іванов. – Х. : Основа, 1992. – 144 с.
8. Айвазян С. А. Прикладная статистика: Исследование зависимостей: Справ. изд. / С. А. Айвазян, И. С. Енюков, Л. Д. Мешалкин. — М.: Финансы и статистика. - 1985. — 487 с.
9. Айвазян С. А. Прикладная статистика: Классификация и снижение размерности: Справ. изд. / С. А. Айвазян, В.М. Бухштабер, И. С. Енюков, Л. Д. Мешалкин. — М.: Финансы и статистика. - 1989. — 607 с.
10. Джессен Р.Д. Методы статистических обследований / Р.Д. Джессен. - М.: Финансы и статистика. - 1985.

11. Добеши И. Десять лекций по вейвлетам / И. Добеши. – Москва-Ижевск: НИЦ «Регулярная и хаотическая динамика». – 2004. – 464 с.

12. Бендат Дж. Прикладной анализ случайных данных / Дж. Бендат, А. Пирсол. М.: Мир. - 1989. – 540 с.

13. Джонсон Н. Статистика и планирование эксперимента в технике и науке. Методы обработки данных / Н. Джонсон, Ф. Аннон. – М.: Мир. – 1980. - 610 с.

14. Малла С. Вейвлеты в обработке сигналов / С. Малла. – М.: Мир.- 2005. – 672 с.

15. Майборода Р.Є. Регресія: Лінійні моделі: Навчальний посібник / Р.Є. Майборода. – К.:ВПЦ «Київський університет». - 2007. – 296 с.

16. Гнеденко Б.В. Курс теории вероятностей / Б.В. Гнеденко. - М.: Наука. - 1988.

17. Закс Л. Статистическое оценивание / Л. Закс. –М.: СТАТИСТИКА. – 1976. – 598 с.

18. Шеффе Г. Дисперсионный анализ / Г. Шеффе. – М.: Наука. – 1980. – 512 с.

19. Закс Ш. Теория статистических выводов / Ш. Закс. – М.: Мир. – 1975. – 776 с.

20. Кокрен У. Методы выборочного исследования / У. Кокрен. – М.: Финансы и статистика. – 1976.

21. Блаттер К. Вейвлет – анализ. Основы теории / К. Блаттер. – М.: Техносфера. – 2004. 276 с.

22. Бахтин В.И. Введение в прикладную статистику / В.И. Бахтин. – Минск: БГУ. – 2011. – 91 с.

23. Шитиков В.К., Классификация, регрессия и другие алгоритмы Data Mining с использованием R / В.К. Шитиков, С.Э. Мастицкий. −

Электронная книга, адрес доступа: https://github.com/ranalytics/data-mining. -2017. – 351 с.

24. Спиридонов А.А. Планирование эксперимента при исследовании технологических процессов / А.А. Спиридонов. – М.: Машиностроение. – 1981. – 184 с.

25. Столниц Э. Вейвлеты в компьютерной графике. Теория и приложения / Э. Столниц, Т. ДеРоуз, Д. Салезин. – Ижевск: НИЦ «Регулярная и хаотическая динамика». – 2002. – 272 с.

26. Чуи К. Введение в вэйвлеты / К. Чуи. – М.:Мир. – 2001. – 412 с.

27. Новиков Л.В. Основы вейвлет-анализа сигналов / Л.В. Новиков. – С-Пб.: ООО»МОДУС». – 1999. – 152 с.

28. Лайонс Р. Цифровая обработка сигналов. / Р. Лайонс. - М.: ООО «Бином-Пресс». - 2006. - 656с.

29. Сергиенко А.Б. Цифровая обработка сигналов / А.Б. Сергиенко. - С-Пб.: ООО «ПитерПринт». - 2002. - 605с.