Кореляційний аналіз

Кореляційний аналіз займається з'ясуванням питання є істотним (суттєвим, достатньо тісним) зв'язок між змінними, які досліджуються.

Основні етапи розв'язання задачі **кореляційного аналізу**:

- вибір числової характеристики статистичного зв'язку;
- визначення оцінки цієї характеристики статистичного зв'язку;
- на основі отриманого значення оцінки характеристики статистичного зв'язку прийняття рішення, чи є істотним зв'язок між змінними, які аналізуються.

І тільки після стверджувальної відповіді про істотність статистичного зв'язку між змінними, що розглядаються, має сенс переходити до наступного етапу – пошуку математичної моделі цього зв'язку засобами інших розділів аналізу та обробки інформації (даних).

Вимоги до характеристик статистичного зв'язку $K_{\eta\vec{\xi}}$ при їх конструюванні:

- якщо $K_{\eta\vec{\xi}} = 0$, то це відповідає ситуації, що зв'язок між η та $\vec{\xi}$ відсутній,
- зі збільшенням відхилення $K_{\eta\vec{\xi}}$ від нуля зростає суттєвість зв'язку між η та $\vec{\xi}$,

---

- нехай $\max\left|K_{\eta\vec{\xi}}\right| = K_{\max}, \left(K_{\max} < \infty\right)$; тоді якщо $\left|K_{\eta\vec{\xi}}\right| = K_{\max}$, то це відповідає ситуації, що зв'язок між η та $\vec{\xi}$ функціональний.

Нехай $\hat{K}_{\eta\vec{\xi}}$ є оцінкою характеристики $K_{\eta\vec{\xi}}$, тоді у загальному випадку спільна **процедура її використання** буде мати вигляд:

- якщо $\hat{K}_{\eta\vec{\xi}} = 0$, то зв'язок між η та $\vec{\xi}$ відсутній,
- якщо $\left|\hat{K}_{\eta\vec{\xi}}\right| = K_{\max}, \left(K_{\max} < \infty\right)$, то зв'язок між η та $\vec{\xi}$ функціональний,
- якщо $\left|\hat{K}_{\eta\vec{\xi}}\right| \in \left(0, K_{\max}\right)$, то потрібно перевірити гіпотезу про те, чи значимо відхиляється від нуля коефіцієнт $K_{\eta\vec{\xi}}$, тобто

Нехай аналізується зв'язок між ***залежною*** скалярною змінною η та вектором ***незалежних*** змінних $\vec{\xi}$ розмірності q деякого типу. Припустимо, що була обрана характеристика (коефіцієнт) статистичного зв'язку $K_{\eta\vec{\xi}}$.

Якщо $q=1$, то її називають ***парною характеристикою статистичного зв'язку***, інакше – ***множинною характеристикою статистичного зв'язку***.

- якщо $\left|\hat{K}_{\eta\vec{\xi}}\right| \in (0, K_{\max})$, то потрібно перевірити гіпотезу про те, чи значимо відхиляється від нуля коефіцієнт $K_{\eta\vec{\xi}}$, тобто здійснити ***перевірку*** $K_{\eta\vec{\xi}}$ ***на значимість***, а саме перевірити гіпотезу:

$$H_0 : K_{\eta\vec{\xi}} = 0,$$

  з деяким рівнем значущості $\alpha > 0$.

Приклади задач кореляційного аналізу:

1. чи суттєво впливає рівень безробіття на рівень злочинності у країні (+1% росту рівня безробіття дає +5% росту рівня злочинності),
2. вплив рівня алкоголю у крові водія на час його реакції,
3. вплив рівня допінгу у крові спортсмена на його спортивне досягнення,
4. вплив ряду основних економічних показників країни на рівень життя населення,
5. чи є суттєвим вплив на об’єми продажів товару/послуги має обсяг фінансування рекламної компанії товару/послуги,
6. чи істотно впливають на врожайність певної сільськогосподарські культури кількість опадів та сонячних годин протягом сезону,
7. чи суттєво залежить остаточний результат студента під час сесії від результатів його роботи протягом семестру.

Структура кореляційного аналізу:

- кореляційний аналіз кількісних змінних,
- кореляційний аналіз ординальних змінних,
- кореляційний аналіз номінальних змінних.

Кореляційний аналіз кількісних змінних

Ключове поняття у цьому розділі це функція регресії η щодо $\vec{\xi}$.

Означення. Нехай η і $\vec{\xi}$ – випадкові величина та вектор, відповідно, причому $M|\eta| < \infty$. Тоді ***функцією регресії*** η ***щодо*** $\vec{\xi}$ називається функція

$$f(\vec{x}) = M(\eta / \vec{\xi} = \vec{x}).$$

Тут $M(\eta / \vec{\xi} = \vec{x})$ - ***умовне математичне сподівання випадкової величини*** η ***відносно події*** $\left\{\vec{\xi} = \vec{x}\right\}$. Детальніше, відомо що:

$$M(\eta) = \int_{-\infty}^{\infty} y\, dF_{\eta}(y), \text{ де } F_{\eta}(y) = P\{\eta < y\}, \text{ тоді}$$

$$M(\eta / \vec{\xi} = \vec{x}) = \int_{-\infty}^{\infty} y\, dF_{\eta}\left(y / \vec{\xi} = \vec{x}\right), \text{ де } F_{\eta}\left(y / \vec{\xi} = \vec{x}\right) = P\left\{\eta < y / \vec{\xi} = \vec{x}\right\}.$$

Якщо $M\eta^2 < \infty$, тоді по аналогії буде існувати ***умовна дисперсія випадкової величини*** η ***відносно події*** $\left\{\vec{\xi} = \vec{x}\right\}$:

$$g(\vec{x}) = D(\eta / \vec{\xi} = \vec{x}).$$

Для $f\left(\vec{\xi}\right)$ та $g\left(\vec{\xi}\right)$ будемо використовувати такі нотації:

$$M\left(\eta / \vec{\xi}\right) = f\left(\vec{\xi}\right),$$
$$D\left(\eta / \vec{\xi}\right) = g\left(\vec{\xi}\right),$$

та називати відповідно ***умовним математичним сподіванням випадкової величини*** η ***відносно випадкового вектора*** $\vec{\xi}$ та ***умовною дисперсією випадкової величини*** η ***відносно випадкового вектора*** $\vec{\xi}$.

Самостійна робота №4. З навчального посібника «Слабоспицький О.С. Основи кореляційного аналізу даних, 2006». Пропрацювати матеріал наведений у **Додатку 1**:

Умовні ймовірності та математичні сподівання. Основні властивості.

Згадаємо деякі **властивості** для них:

1. $M\left\{M\left(\eta/\vec{\xi}\right)\right\} = M\eta,$
2. $M\left(\varphi\left(\vec{\xi}\right)\eta/\vec{\xi}\right) = \varphi\left(\vec{\xi}\right)M\left(\eta/\vec{\xi}\right), \quad \varphi(\cdot)\in\mathfrak{B}_q,$

   де $\mathfrak{B}_q$ – множина борелівських функцій на $\mathbb{R}^q$.

Наслідок.

1. $Mf\left(\vec{\xi}\right) = M\eta,$ (*)
2. $M\left[\left(\eta - f\left(\vec{\xi}\right)\right)/\vec{\xi}\right] = 0.$ (**)
3. $M\left(\varphi\left(\vec{\xi}\right)\left[\eta - f\left(\vec{\xi}\right)\right]\right) = 0, \quad \forall\varphi(\cdot)\in\mathfrak{B}_q.$ (***)

Доведення наслідку. 1. Очевидно з огляду на першу властивість.
2. Легко перевіряється, бо згідно другої властивості:

$$M\left[\left(\eta - f\left(\vec{\xi}\right)\right)/\vec{\xi}\right] = M\left(\eta/\vec{\xi}\right) - f\left(\vec{\xi}\right) = 0.$$

3. Скористаємося послідовно спочатку властивістю 1 справа наліво, а потім наслідком 2, в результаті отримаємо:

$$M\left(\varphi\left(\vec{\xi}\right)\left[\eta - f\left(\vec{\xi}\right)\right]\right) = M\left\{M\left(\varphi\left(\vec{\xi}\right)\left[\eta - f\left(\vec{\xi}\right)\right]/\vec{\xi}\right)\right\} =$$

$$= M\left\{\varphi\left(\vec{\xi}\right)M\left(\left[\eta - f\left(\vec{\xi}\right)\right]/\vec{\xi}\right)\right\} = 0. \qquad \blacksquare$$

Лема. Якщо η та $\vec{\xi}$ – випадкові величина та вектор відповідно, а $M\eta^2 < \infty$, тоді для них справедливо:

$$D\eta \overset{\mathrm{I}}{=} Df(\vec{\xi}) + Mg(\vec{\xi}),$$

або розгорнуто

$$D\eta = M\left\{\left(M\left(\eta/\vec{\xi}\right) - M\eta\right)^2\right\} + M\left\{M\left\{\left(\eta - M\left(\eta/\vec{\xi}\right)\right)^2/\vec{\xi}\right\}\right\}.$$

Доведення. Використаємо вищенаведені властивості:

$$D\eta = M(\eta - M\eta)^2 = M\left[\left(\eta - f(\vec{\xi})\right) + \left(f(\vec{\xi}) - M\eta\right)\right]^2 =$$

$$= M\left(\eta - f(\vec{\xi})\right)^2 + 2M\left[\left(\eta - f(\vec{\xi})\right)\left(f(\vec{\xi}) - M\eta\right)\right] + M\left(f(\vec{\xi}) - M\eta\right)^2 \overset{(***)}{=}$$

$$= M\left\{M\left[\left(\eta - f(\vec{\xi})\right)^2 / \vec{\xi}\right]\right\} + Df(\vec{\xi}) =$$

$$= M\left\{M\left[\left(\eta - M(\eta / \vec{\xi})\right)^2 / \vec{\xi}\right]\right\} + Df(\vec{\xi}) = M\left\{D(\eta / \vec{\xi})\right\} + Df(\vec{\xi}) =$$

$$= Mg(\vec{\xi}) + Df(\vec{\xi}).$$

■

Теорема(про фундаментальну властивість функції регресії). Нехай η і $\vec{\xi}$ – випадкові величина та вектор розмірності q, відповідно, причому $M\eta^2 < \infty$, $\mathfrak{B}_q$ – множина борелівських функцій на $\mathbb{R}^q$, тоді:

$$f(\cdot) = \arg \min_{\varphi(\cdot) \in \mathfrak{B}_q} M\left[\eta - \varphi(\vec{\xi})\right]^2,$$

де $f(\vec{x}) = M(\eta / \vec{\xi} = \vec{x})$.

$$f(\cdot) = \arg\min_{\varphi(\cdot) \in \mathfrak{B}_q} M\left(\eta - \varphi(\vec{\xi})\right)^2$$

$$\boxed{\eta = f(\vec{\xi}) + \varepsilon}$$

Доведення. Без втрати загальності будемо розглядати ті борелівські функції $\varphi(\cdot)$ на $\mathbb{R}^q$, для яких $M\varphi^2(\vec{\xi}) < \infty$. Скориставшись властивостями умовного математичного сподівання випадкової величини η відносно випадкового вектора $\vec{\xi}$ і врахувавши, що $M\left[f(\vec{\xi}) - \varphi(\vec{\xi})\right]^2 \geq 0$, легко бачити, що має місце такий ланцюжок перетворень:

$$M\left[\eta-\varphi\left(\vec{\xi}\right)\right]^2 = M\left[\left(\eta-f\left(\vec{\xi}\right)\right)+\left(f\left(\vec{\xi}\right)-\varphi\left(\vec{\xi}\right)\right)\right]^2 =$$
$$= M\left[\eta-f\left(\vec{\xi}\right)\right]^2 + 2M\left[\left(\eta-f\left(\vec{\xi}\right)\right)\left(f\left(\vec{\xi}\right)-\varphi\left(\vec{\xi}\right)\right)\right]+$$
$$+M\left[f\left(\vec{\xi}\right)-\varphi\left(\vec{\xi}\right)\right]^2 \overset{(***)}{\geq} M\left[\eta-f\left(\vec{\xi}\right)\right]^2.$$

У результаті, маємо

$$M\left[\eta-\varphi\left(\vec{\xi}\right)\right]^2 \geq M\left[\eta-f\left(\vec{\xi}\right)\right]^2,$$

тобто отримано нижню межу для нашого функціоналу $M\left[\eta-\varphi\left(\vec{\xi}\right)\right]^2$, яка досягається на функції регресії $f(\vec{x}) = M(\eta/\vec{\xi}=\vec{x})$. ■

Наведена теорема дозволяє стверджувати, що найкращою в середньоквадратичному розумінні апроксимацією η на класі борелівських функцій від $\vec{\xi}$ є функція $f\left(\vec{\xi}\right)$, тобто за математичну модель можна взяти таке співвідношення, яке будемо називати ***регресійною моделлю*** η ***щодо*** $\vec{\xi}$:

$$\eta = f\left(\vec{\xi}\right)+\varepsilon,$$

де ε – залишкова похибка апроксимації.

Для цієї моделі мають місце такі **властивості**:

1) $Mf\left(\vec{\xi}\right)=M\eta$, $M\varepsilon=0$;

2) $f\left(\vec{\xi}\right)$ та ε – некорельовані;

3) $D\eta = Df(\vec{\xi})+D\varepsilon$.

Зауваження. Перша властивість дозволяє запропонувати для останнього співвідношення ще одне представлення:

$$D\eta = Df(\vec{\xi})+M\varepsilon^2.$$

Доведення. Скористаємося властивостями умовного математичного сподівання випадкової величини η відносно випадкового вектора $\vec{\xi}$ та доведемо послідовно кожну з властивостей:

1) $Mf\left(\vec{\xi}\right) \overset{(*)}{=} M\eta$.

А це у свою чергу, дозволяє стверджувати, що

$$M\varepsilon = M\left\{\eta - f\left(\vec{\xi}\right)\right\} = M\eta - Mf\left(\vec{\xi}\right) = 0;$$

2) оскільки, згідно з попередньою властивістю $M\varepsilon = 0$, то для доведення некорельованості достатньо впевнитися, що $M\left[f\left(\vec{\xi}\right) - Mf\left(\vec{\xi}\right)\right]\varepsilon = 0$. Дійсно:

$$M\left[f\left(\vec{\xi}\right) - Mf\left(\vec{\xi}\right)\right]\varepsilon = M\left[f\left(\vec{\xi}\right) - M\eta\right]\left[\eta - f\left(\vec{\xi}\right)\right] \overset{(***)}{=} 0.$$

3) з доведеної у другій властивості некорельованості $f\left(\vec{\xi}\right)$ та ε випливає: $D\eta = D\left(f\left(\vec{\xi}\right) + \varepsilon\right) = Df\left(\vec{\xi}\right) + D\varepsilon.$ ■

Коефіцієнт детермінації та його властивості. Індекс кореляції.

Введемо універсальну характеристику статистичного зв'язку для змінних η та $\vec{\xi}$, $\left(\vec{\xi} \in \mathbb{R}^q\right)$. Будемо вважати, що $M\left|\eta\right| < \infty$, тоді існуватиме функція регресії η щодо $\vec{\xi}$, а саме: $f(\vec{x}) = M(\eta / \vec{\xi} = \vec{x})$. А для випадкової величини η можна використовувати таку математичну модель:

$$\eta = f\left(\vec{\xi}\right) + \varepsilon,$$

де ε – залишкова похибка апроксимації.

Причому $D\eta = Df(\vec{\xi}) + D\varepsilon$.

Після цих міркувань, враховуючи унікальну властивість функції регресії, доведену в теоремі, цілком природним здається використання, як характеристики статистичного зв'язку для кількісних змінних η та $\vec{\xi}$, нижчевизначеного коефіцієнта.

Означення. Нехай η і $\vec{\xi}$ – випадкові величина та вектор розмірності q, відповідно, причому $0 < D\eta < \infty$. Тоді ***коефіцієнтом детермінації*** η ***щодо*** $\vec{\xi}$ називається величина

$$I_{\eta\vec{\xi}}^2 = \frac{Df(\vec{\xi})}{D\eta} = 1 - \frac{D\varepsilon}{D\eta} = 1 - \frac{M\varepsilon^2}{D\eta}.$$

Зауваження 1. Коефіцієнт детермінації $I_{\eta\vec{\xi}}^2$ приваблює ще тим, що *має прозору інтерпретацію*, а саме: він вказує, яка частина дисперсії змінної η визначається варіацією (дисперсією) функції регресії $f(\vec{\xi})$.

Зауваження 2. Коефіцієнт детермінації $I_{\eta\vec{\xi}}^2$, де $\vec{\xi} \in \mathbb{R}^q$, у випадку $q = 1$ ще називають ***парним коефіцієнтом детермінації*** η ***щодо*** $\vec{\xi}$, а коли $q > 1$ його ще називають ***множинним коефіцієнтом детермінації*** η ***щодо*** $\vec{\xi}$.

Властивості коефіцієнта детермінації:

1) $0 \le I_{\eta\vec{\xi}}^2 \le 1$;
2) якщо $I_{\eta\vec{\xi}}^2 = 0$, то відсутній вплив $\vec{\xi}$ на η;
3) якщо $I_{\eta\vec{\xi}}^2 = 1$, то існує функціональний зв'язок між η та $\vec{\xi}$, а саме, з ймовірністю 1 справедливо $\eta = f\left(\vec{\xi}\right)$.

Зауваження. Тут і далі для випадкових величин/векторів рівності/нерівності вважаються справедливими з ймовірністю 1, якщо не наголошується на іншому.

Доведення. Скористаємось відомими властивостями умовного математичного сподівання та умовної дисперсії:

1) так як $D\varepsilon \geq 0$, то отримуємо:

$$0 \leq I^2_{\eta\vec{\xi}} = 1 - \frac{D\varepsilon}{D\eta} \leq 1;$$

2) припустимо, що $I^2_{\eta\vec{\xi}} = 0$. Тоді

$$Df(\vec{\xi}) = 0 \Rightarrow f(\vec{\xi}) = Mf(\vec{\xi}) = \text{const}.$$

Звідси випливає, що функція регресії, за допомогою якої апроксимується η, не залежить від значень свого аргументу – $\vec{\xi}$;

3) нехай $I^2_{\eta\vec{\xi}} = 1$, тоді

$$1 = I^2_{\eta\vec{\xi}} = 1 - \frac{D\varepsilon}{D\eta} \Rightarrow D\varepsilon = 0.$$

Згідно першої властивості регресійної моделі $M\varepsilon = 0$, тоді

$$D\varepsilon = 0 \Rightarrow \varepsilon = M\varepsilon = 0.$$

Але $\varepsilon = \eta - f\left(\vec{\xi}\right)$, а це, у свою чергу, дозволяє стверджувати, що з ймовірністю 1 має місце така функціональна залежність між η та $\vec{\xi}$:

$$\eta = f\left(\vec{\xi}\right). \qquad \blacksquare$$