АНАЛІЗ ДАНИХ

Нікіта Скибицький

6 січня 2019 р.

У ваших руках конспект лекцій з нормативного курсу “Вступ до аналізу даних”, прочитаного доц., к.ф.-м.н. Слабоспицьким Олександром Сергійовичем на третьому курсі спеціальності прикладна математика факультету комп’ютерних наук та кібернетики Київського національного університету імені Тараса Шевченка восени 2018-го року.

Конспект у компактній формі відображає матеріал курсу, допомагає сформувати загальне уявлення про предмет вивчення, правильно зорієнтуватися в даній галузі знань. Конспект лекцій з названої дисципліни сприятиме більш успішному вивченню дисципліни, причому більшою мірою для студентів заочної форми, екстернату, дистанційного та індивідуального навчання.

Комп’ютерний набір та верстка – Скибицький Нікіта Максимович.

Зміст

1 Вступ

1.1 Етапи аналізу даних:

1. Отримання і збереження даних.
2. Обробка даних.
3. Аналіз отриманих результатів (найважливіший етап).

1.2 Основні розділи аналізу даних:

1. Попередня обробка (включаючи розвідувальний аналіз) даних.
2. Кореляційний аналіз – застосування наявності зв'язків.
3. Дисперсійний аналіз.
4. Регресійний аналіз.
5. Коваріаційний аналіз.
6. Кластерний аналіз.
7. Дискримінантний аналіз.
8. Аналіз часових рядів.

Пункти 3-5 – побудова математичних моделей зв'язків.

1.3 Класифікація змінних

ξ, η, ζ – змінні, які ми спостерігаємо.

$\{x_i\}_{i\in I}$, $\{y_j\}_{j\in J}$, $\{z_k\}_{k\in K}$ – спостереження за змінними.

Змінні: кількісні і якісні.

- Кількісні.
- Якісні:
  - ординальні (порядкові);
  - номінальні (класифікаційні).

Ординальні змінні – змінні, що приймають значення з деякої множини, елементи якої називаються градаціями, причому кожен елемент множини апріорі впорядкований відносно інших (задано чіткий порядок)

Приклад. Рівень освіти: бакалавр, спеціаліст, магістр – упорядковані змінні.

Приклад. Військові звання.

Номінальні змінні – змінні, що приймають своє значення з деякої множини, елементи (градації) якої не мають наперед заданого порядку (загальновідомого).

Категоризовані змінні – змінні, для яких апріорі відома множина їх значень (градацій) та алгоритм віднесення конкретного спостереження такими змінними до градації.

Некатегоризовані змінні – змінні, для яких апріорне задана або множина значень, або алгоритм віднесення спостереження до певної градації.

Приклад. Некатегоризовані змінні – назви юр. осіб на даний момент. Влаштувались на роботу, зарплату не заплатили, фірма зникла, на іншій вулиці з'явилась $\Rightarrow$ градація зникла.

Ще існує поділ на дискретні і неперервні змінні.

1.4 Групування даних

Проводиться при спостереженнях над неперервними змінними (кількість спостережень $n > 50$). У дискретному випадку звертають увагу на кількість змінних $m > 10$.

Ідея підходу: вся вибірка спостережень розбивається на підвибірки і кожен заміняється на типового представника і далі працюють з цими представниками.

Нехай є вибірка. По ній знаходимо min і max значення.

$$\{x_i\}_{i=1}^n : (x_{\min}, x_{\max}).$$

Цей інтервал розбиваємо на s підінтервалів. Зазвичай s вибирають так

$$5 \leq s \leq 30 \quad \text{і} \quad s = 1 + \lfloor \log_2(n) \rfloor.$$

Беруть підінтервали $(C_1^1, C_1^2]$, $(C_2^1, C_2^2]$, ..., (C_s^1, C_s^2). Потрібно, щоб в кожен інтервал потрапило більше 5 спостережень. Вибирають з кожного інтервалу єдиного представника.

Поставимо у відповідність $(C_i^1, C_i^2] \mapsto x_i^0$ (як правило середня точка середня точка), v_i – частота попадання.

Тобто переходимо від вибірки $\{x_i\}_{i=1}^n$ до вибірки $\{x_i^0, v_i\}_{i=1}^s$.

Зауваження: для випадкової величини ξ – $F_\xi(x)$ – *функція розподілу.*

$\widehat{F}_\xi(x)$ – *темпіричний розподіл*, n – *об'єм вибірки.*

$p_s(x)$, $\widehat{p}_s^{(n)}(x)$ – *неперервний випадок* (щільність).

Для дискретного випадку $\{y_i, p_i\}_{i=1}^m \mapsto \{y_i, \widehat{p}_i\}_{i=1}^m$ – *полігон частот.*

2 Розвідувальний аналіз

Займається розробкою методів попереднього експрес аналізу інформації шляхом представлення її у вигляді таблиць або різного роду графічних зображень.

2.1 Спостереження за однією змінною

Засоби спостереження:

1. пробіт-графік;
2. імовірнісний графік;
3. висячі гістобари;
4. підвішена коренеграма;
5. зображення "скринька з вусами";
6. зображення "стебло-листок".

У випадках 1-2 використовуємо інше зображення функцій розподілу, 3-4 – використання іншого розподілу емпіричних функцій щільності, 5-6 – сімейство розподілів зсув масштабу.

Сімейство розподілів $\mathcal{F}$ – *сімейство розподілів типу зсуву масштабу*, якщо існує функція розподілу

$$\exists F_0(\cdot) \in \mathcal{F} : \forall F(\cdot) \in \mathcal{F} : \exists a, b \in \mathbb{R}^1, (b > 0) : F(x) = F_0\left(\frac{x-a}{b}\right),$$

де a – параметр зсуву, b – параметр масштабу, F_0 – базова функція для сімейства розподілів $\mathcal{F}$.

Приклад. Нормальний розподіл $F(x) : N(m, \sigma^2)$. Базова функція $\Phi(x)$ з розподілу $N(0, 1)$. $a = m$, $b = \sigma$, $F(x) = \Phi\left(\frac{x-m}{\sigma}\right)$. Сімейство нормальних розподілів є сімейством зсуву масштабу.

Приклад. Експоненціальний розподіл з параметром $\lambda > 0$. $a = 0$, $b = \frac{1}{\lambda}$, $F(x) = \Phi_1(\lambda x)$, Φ_1 – базова функція експоненціального розподілу з параметром $\lambda = 1$.

2.1.1 Пробіт-графік

Будується наступним чином:

Маємо на вході вибірку $\{x_i\}_{i=1}^n$.

Обчислимо емпіричну функцію розподілу $\{x_i\}_{i=1}^n \mapsto \widehat{F}(x)$ (Сімейство розподілів $\mathcal{F}$ з базовою функцією F_0).

Пробіт-графік – графік функції $y = F_0^{-1}\left(\widehat{F}(x)\right)$.

Використовується для:

1. Перевірки гіпотези $H_0 : F_\xi(\cdot) \in \mathcal{F}$.

   У випадку, коли справедлива гіпотеза $H_0 : F_\xi(\cdot) \in \mathcal{F}$. пробіт-графік повинен уявляти собою майже пряму.
   Пояснення: маємо:

   $$y = F_0^{-1}(\widehat{F}_\xi(x)) \overset{H_0}{\approx} F_0^{-1}\left(F_0\left(\frac{x-a}{b}\right)\right) = \frac{x}{b} - \frac{a}{b}.$$

2. Виявлення наявності аномальних спостережень у вибірці.

2.1.2 Імовірнісний графік

Ідея та ж сама. Зі спотвореною віссю . Маємо множину $\{x \in \mathbb{R}, y \in [0,1]\}$, яку розтягують за правилом $(x,y) \mapsto \left(x, F_0^{-1}(y)\right)$, де $y = \widehat{F}_\xi(x)$.

Папір, де спотворюється масштаб називається *імовірнісним папером.*

Якщо в якості розподілу взяти нормальний розподіл, то такий папір називається *нормальним імовірнісним папером.*

Будуємо графік функції $y = F_\xi(x)$ – для спостереження величини ξ.

1. У випадку, коли $H_0 : F_\xi(\cdot) \in \mathcal{F}$, то отримаємо майже пряму.
2. Виявляємо наявність *аномальних спостережень*: якщо маємо точки, що лежать осторонь, то перевіряємо їх на аномальність.

2.1.3 Висячі гістобари

Використовується для перевірки нормальності вибірки. Нехай по вибірці ξ: x_1, …, x_n підраховано мат. сподівання $\bar{x}(n)$ та вибіркова дисперсія $s^2(n)$.

Найбільш узгодженим нормальним розподілом для спостережень за ξ будемо називати такий нормальний розподіл $N\left(x(n), s^2(n)\right)$.

Спочатку будуємо графік щільності з вибірки $\xi : x_1, \ldots, x_n$.

В центрах групування даних до графіка підвішуються прямокутні гістобари, довжина яких пропорційна відносній частоті потрапляння у відповідний інтервал групування. Якщо основа цих гістобар не суттєво відхиляється від осі – гіпотеза про нормальність вибірки приймається.

2.1.4 Підвішена коренеграма

Для кожного інтервалу групування даних визначають v_e – емпірична частота потрапляння в інтервал, а також теоретичне значення частоти v_T згідно гіпотези про найбільш узгоджений нормальний розподіл. Потім на графіку відкладають такі різниці: $\sqrt{v_e} - \sqrt{v_T}$. І якщо ці значення не значно відхиляються від нуля, то гіпотеза про нормальність вибірки приймається.

2.2 Спостереження за двома змінними

Використовуються:

1. Діаграма розсіювання.
2. Таблиця спряженості.

2.2.1 Діаграми розсіювання

Маємо дві вибірки $\xi : x_1, \ldots, x_n$ та $\eta : y_1, \ldots, y_n$. Використовують для з'ясування класу залежності між парою кількісних змінних, а також для з'ясування наявності аномальних спостережень у вибірці.

2.2.2 Таблиця спряженості

Використовуються для представлення спостережень над номінальними, ординальними, кількісними дискретними (скінченими), кількісними неперервними (згрупованими змінними).

Нехай є змінна ξ яка має r_1 градацій, та змінна ζ яка має r_2 градацій:

	1	2	...	r_1	$\sum$
1	n_{11}	n_{12}	...	n_{1r_1}	n_{1*}
2	n_{21}	n_{22}	...	n_{2r_1}	n_{2*}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
r_2	n_{r_21}	n_{r_22}	...	$n_{r_2r_1}$	n_{r_2*}
$\sum$	n_{*1}	n_{*2}	...	n_{*r_1}	n_{**}

Де n_{ij} – кількість таких спостережень $\{\xi = i, \zeta = j\}$, позначимо $n_{i*} = \sum_{j=1}^{r_1} n_{ij}$, $n_{*j} = \sum_{i=1}^{r_2} n_{ij}$.

3 Попередня обробка

До попередньої обробки відносять:

- розвідувальний аналіз;
- обчислення основних характеристик спостережуваних величин;
- видалення аномалій;

- перевірка основних гіпотез;
- перевірка на стохастичність вибірки.

3.1 Характеристики випадкових (скалярних) величин

3.1.1 Квантилі та процентні точки

Квантилем рівня $0 < q < 1$ для *неперервної* випадкової величини $\xi : F_\xi(\cdot)$ називається значення $u_q(F) : P\{\xi < u_q(F)\} = q$.

Квантилем рівня $0 < q < 1$ для *дискретної* випадкової величини $\xi : F_\xi(\cdot)$ називається будь-яке значення $U_q(F) \in (y_{i(q)}, y_{i(q)+1}]$, для границь якого виконується $P\{\xi < y_{i(q)}\} < q$ та $P\{\xi < y_{i(q)+1}\} \geq q$.

Вибіркові квантилі $\widehat{u}_q(F)$ визначаються як квантилі відповідних емпіричних розподілів.

Q-процентною точкою $(0 < Q < 100)$ для *неперервної* випадкової величини $\xi : F_\xi(\cdot)$ називається значення $\omega_Q(F) : P\{\xi \geq \omega_Q(F)\} = \frac{Q}{100}$.

Q-процентною точкою $(0 < Q < 100)$ для *дискретної* випадкової величини $\xi : F_\xi(\cdot)$ називається довільне значення $w_Q(F) \in (y_{i(Q)}, y_{i(Q)+1}]$, для границь якого виконується $P\{\xi \geq y_{i(Q)}\} > \frac{Q}{100}$ та $P\{\xi \geq y_{i(Q)+1}\} \leq \frac{Q}{100}$.

Квантиль та процентна точка пов'язані певним співвідношенням, а саме

$$\omega_Q(F) = u_{1-\frac{Q}{100}}(F) \quad \text{та} \quad u_q(F) = \omega_{(1-q)100}(F).$$

Введемо додаткові характеристики розподілу, похідні від перших двох.

Приклади квантилей:

1. *Медіаною* називається квантиль рівня 0.5: $u_{0.5}$.
2. $u_{0.75}, u_{0.25}$ – *верхній та нижній квартилі* відповідно.
3. Значення $\{u_{\frac{i}{10}}\}_{i=1}^{9}$ називаються *децилями*.
4. $\{u_{\frac{i}{100}}\}_{i=1}^{99}$ – *процентилі*.
5. *Інтерквантильною широтою рівня* $p : 0 < p < \frac{1}{2}$ називається величина $u_{1-p} - u_p$.

6. *Інтерквартильною широтою* називається величина $u_{0.75} - u_{0.25}$ тобто $p = 0.25$

7. Половина інтерквантильної широти називається *імовірнісним відхиленням.*

3.1.2 Характеристики положення центру значень

1. *Математичне сподівання* $M\xi$, та його вибірковий аналог

$$\bar{x}(n) = \frac{1}{n}\sum_{i=1}^{n}\xi_i.$$

2. *Геометричне середнє* $G_\xi = e^{M\ln\xi}$ для $\xi : P\{\xi \le 0\} = 0$.

$$\widehat{G}_\xi(n) = \sqrt[n]{\prod_{i=1}^{n} x_i}.$$

3. *Середнє гармонічне* $H = M^{-1}\left(\frac{1}{\xi}\right)$ для $\xi : P\{\xi \le 0\} = 0$.

$$\widehat{H}_\xi = \left(\frac{1}{n}\sum_{i=1}^{n}\frac{1}{x_i}\right)^{-1}.$$

4. *Мода* $x_{\text{mod}} = \arg\max_a P\{\xi = a\}$ для дискретних випадкових величин (визначається за гістограмою) та $x_{\text{mod}} = \arg\max_x f(x)$ в неперервному випадку.

5. *Медіаною* називається $x_{\text{med}} = u_{0.5}$.

3.1.3 Характеристики розсіювання значень

Нехай маємо вибірку об'єму n спостережень x_1, x_2, …, x_n над випадковою величиною ξ.

1. *Дисперсія* $D\xi = M(\xi - M\xi)^2$. Вибіркове значення

$$S^2(n) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}(n))^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n}x_i^2 - n\bar{x}^2(n)\right).$$

2. *Стандартне (середньоквадратичне) відхилення* $\sqrt{D\xi}$. Вибіркове значення $S(n)$.

3. *Коефіцієнт варіацій* $V_\xi = \frac{\sqrt{D\xi}}{M\xi}100\%$, $M\xi \neq 0$. Вибіркове значення $\widehat{V}_\xi(n) = \frac{S(n)}{\bar{x}(n)}100\%$.

4. *Стохастичне розсіювання* (імовірнісне відхилення) – це половина інтерквартильної широти: $\frac{U_{0.75}-U_{0.25}}{2}$. Вибіркове значення $\frac{\widehat{U}_{0.75}-\widehat{U}_{0.25}}{2}$.

5. *Розмах (широта) вибірки*: $x_{\max} - x_{\min}$, де $x_{\max}, x_{\min}$ – найбільше та найменше значення у вибірці.

6. *Інтервал концентрації* $(M\xi - 3\sqrt{D\xi}, M\xi + 3\sqrt{D\xi})$. Вибіркове значення $(\bar{x}(n) - 3S(n), M\bar{x}(n) + 3S(n))$.

3.1.4 Характеристики скошеності та гостроверхості розподілу

Нехай є розподіл випадкової величини ξ і отримані спостереження x_1, x_2, …, x_n над нею.

1. *Коефіцієнт асиметрії* – характеристика скошеності розподілу (базується на третьому центральному моменті):

$$\beta_1 = \frac{M(\xi - M\xi)^3}{(M(\xi - M\xi)^2)^{3/2}}, \quad D\xi > 0.$$

   Вибіркове значення

$$\widehat{\beta}_1 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_k - \bar{x}(n))^3}{S^3(n)}.$$

   Дисперсія спостережуваної величини $D\xi > 0$.

   Якщо розподіл симетричний (наприклад нормальний) то $\beta_1 = 0$. Якщо $\beta_1 > 0$, то розподіл скошений вліво, якщо $\beta_1 < 0$, то вправо.

2. *Коефіцієнт ексцесу* – характеристика гостроверхості розподілу (базується на четвертому центральному моменті):

$$\beta_2 = \frac{M(\xi - M\xi)^4}{(M(\xi - M\xi)^2)^2} - 3, \quad D\xi > 0.$$

Вибіркове значення

$$\widehat{\beta}_2 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_k - \bar{x}(n))^4}{S^4(n)} - 3.$$

Для нормального розподілу коефіцієнт ексцесу дорівнює нулю. Якщо $\beta_2 > 0$, то розподіл більш гостроверхий ніж нормальний, якщо $\beta_2 < 0$ то відповідно менш гостроверхий.

3.2 Характеристики векторних величин

Аналіз q-вимірних векторних величин, отримано n спостережень над вектором $\vec{\xi} : x_1, x_2, \ldots, x_n$, $x_i \in \mathbb{R}^q$, $i = \overline{1,n}$.

3.2.1 Характеристики положення центру значень

1. *Математичне сподівання* (теоретичне середнє) $M\xi$. Вибіркове значення
$$\bar{x}(n) = \frac{1}{n}\sum_{i=1}^{n}\vec{x}_i.$$

2. *Мода* x_{mod}. У неперервному випадку – це точка максимуму функції щільності ξ. Для дискретного випадку – це значення, яке набуває ξ з найбільшою ймовірністю.

3.2.2 Характеристики розсіювання значень

1. *Коваріаційна матриця* $\sum = M(\xi - M\xi)(\xi - M\xi)^T$. Вибіркове значення
$$\widehat{\sum}(n) = \frac{1}{n-1}\sum_{k=1}^{n}(x_k - \bar{x}(n))(x_k - \bar{x}(n))^T.$$

2. *Узагальнена дисперсія* – визначник коваріаційної матриці: $\det\sum$. Вибіркове значення $\det\left(\widehat{\sum}\right)$.

3. *Слід коваріаційної матриці* $\operatorname{tr}\sum$. Вибіркове значення $\operatorname{tr}\left(\widehat{\sum}(n)\right)$.

3.3 Перевірка стохастичності вибірки

Перевіряємо, чи справді вибірка є випадковою, а не знаходиться під впливом деякого систематичного зміщення. Для цього запропоновано критерії:

- Критерій серій на базі медіани
- Критерій зростаючих та спадаючих серій
- Критерій квадратів послідовних різниць (критерій Аббе)

Нехай x_1, x_2, …, x_n – вибірка спостережень, яка досліджується.

Будемо перевіряти гіпотезу H_0: ця вибірка є стохастичною з рівнем значимості $\alpha(0 < \alpha < 1)$ (рівень значимості – ймовірність допустити помилку першого роду).

1. *Критерій серій на базі медіани.* Альтернативна гіпотеза H_1: наявність у вибірці систематичного монотонного зміщення середнього.

   Спочатку визначається вибіркове значення медіани $\widehat{x}_{\text{med}}$. Потім під кожним членом вибірки ставимо відповідно

   $$\begin{cases} +, & x_i > \widehat{x}_{\text{med}} \\ \text{нічого}, & x_i = \widehat{x}_{\text{med}} \\ -, & x_i < \widehat{x}_{\text{med}} \end{cases}.$$

   Отримаємо послідовність символів. *Серія* – послідовність підряд розташованих однакових символів + чи −. *Довжина серії* – це кількість членів у ній.

   Для отриманої послідовності обчислюємо дві статистики: загальну кількість серій в послідовності $v(n)$, довжину найдовшої серії $\tau(n)$. Запишемо область прийняття нашої гіпотези:

   $$\begin{cases} v(n) > v_\beta(n) \\ \tau(n) < \tau_{1-\beta}(n) \end{cases}$$

   де $v_\beta(n)$, $\tau_\beta(n)$ – квантилі рівня β статистик $v(n)$, $\tau(n)$ відповідно. При фіксованому значенні β рівень значимості α лежить у межах $\beta < \alpha < 2\beta - \beta^2$. Якщо порушується хоч одна з нерівностей, то гіпотеза відхиляється.

2. *Критерій зростаючих та спадаючих серій.* Альтернативна гіпотеза H_1: наявність у вибірці систематичного періодичного зміщення середнього. Спочатку у вибірці замінюємо підряд розташовані однакові виміри одним їх представником. В результаті отримаємо послідовність x'_1, x'_2, ..., x'_k. Під кожним членом послідовності ставимо відповідно

$$\begin{cases} +, & x'_i < x'_{i+1} \\ -, & x'_i > x'_{i+1} \end{cases}.$$

Далі для таким чином отриманої послідовності + та −, як і в попередньому випадку, обчислюємо дві статистики: загальну кількість серій в послідовності $v(n)$, довжину найдовшої серії $\tau(n)$. Запишемо область прийняття нашої гіпотези:

$$\begin{cases} v(n) > v_\beta(n) \\ \tau(n) < \tau_{1-\beta}(n) \end{cases}$$

де $v_\beta(n)$, $\tau_\beta(n)$ – квантилі рівня β статистик $v(n)$, $\tau(n)$ відповідно. При фіксованому значенні β рівень значимості α лежить у межах $\beta < \alpha < 2\beta - \beta^2$. Якщо порушується хоч одна з нерівностей, то гіпотеза відхиляється.

3. *Критерій квадратів послідовних різниць (критерій Аббе).* Він є найбільш потужним на класі усіх нормальних вибірок. Альтернативна гіпотеза H_1: наявність у вибірці систематичного зміщення середнього.

На основі вибірки підраховуємо наступну статистику:

$$\gamma(n) = \frac{\dfrac{1}{2(n-1)} \displaystyle\sum_{i=1}^{n-1} (x_{i+1} - x_i)^2}{\dfrac{1}{n-1} \left(\displaystyle\sum_{i=1}^{n} x_i^2 - n\bar{x}^2(n) \right)}.$$

Область прийняття гіпотези для цього критерію має вигляд $\gamma(n) > \gamma_\alpha(n)$, де $\gamma_\alpha(n)$ – квантиль рівня α статистики $\gamma(n)$, що при $n \leq$ 60визначається з таблиць, а протилежному випадку потрібно скористатися формулою

$$\gamma_\alpha(n) = 1 + \frac{u_\alpha}{\sqrt{n + 0.5(1 + u_\alpha^2)}},$$

де u_α – квантиль рівня α нормального розподілу з параметрами 0 та 1.

3.4 Рангові критерії однорідності

Розглянемо випадкові величини $\xi_1, \xi_2, \ldots, \xi_k$ з функціями розподілу $F_1(x)$, $F_2(x), \ldots, F_k(x)$. На їх основі сформуємо об'єднану вибірку $\nu_1, \nu_2, \ldots, \nu_n$, а для кожної змінної ξ_i отримаємо незалежні спостереження $x_1^{(i)}, x_2^{(i)}, \ldots, x_{n_i}^{(i)}$, $i = \overline{1,k}$. Тоді сформована вибірка буде об'ємом $n = \sum_{i=1}^{k} n_i$. Для спрощення вважаємо, що всі виміри ν_i, $i = \overline{1,n}$ різні. Розташувавши ці значення у порядку зростання, отримаємо варіаційний ряд $\nu_{(1)}, \nu_{(2)}, \ldots, \nu_{(n)}$. Члени варіаційного ряду називають порядковими статистиками.

Рангом спостереження ν_i $(i = \overline{1,n})$ називається його порядковий номер у побудованому варіаційному ряді, позначається $R_{i,n}$ – ранг спостереження ν_i $(i = \overline{1,n})$.

3.4.1 Статистика для лінійного рангового критерію

$$K_i = \sum_{j=N_i-n_i+1}^{N_i} \phi(R_{j,n}), \quad N_i = \sum_{j=1}^{i} n_j, \quad i = \overline{1,k}$$

K_i – статистика по спостереження над ξ_i, $\phi(R_{i,n})$ – мітка.

Потрібно перевірити гіпотезу $H_0 : F_1(x) = F_2(x) = \ldots = F_k(x)$, $\forall x$ з рівнем значимості α $(0 < \alpha < 1)$. Хочемо переконатись зо всі випадкові величини однаково розподілені.

3.4.2 Випадок двох вибірок

Гіпотеза $H_0 : F_1(x) = F_2(x)$, $\forall x$ з рівнем значимості α $(0 < \alpha < 1)$. Альтернативні гіпотези:

$$H_{11} : F_1(x) = F_2(x - \Delta), \forall x, (\Delta \neq 0)$$
$$H_{12} : F_1(x) = F_2(x - \Delta), \forall x, (\Delta > 0)$$
$$H_{13} : F_1(x) = F_2(x - \Delta), \forall x, (\Delta < 0)$$

Всі критерії розглядаються над першою змінною ξ_1.

Критерій нормальних міток (Фішера)

$C = \sum_{i=1}^{n_1} M(R_{i,n}, n)$, де $M(m, n)$ – математичне сподівання m-ої порядкової статистики вибірки довжини $n = n_1 + n_2$ нормально розподіленої величини з параметрами 0 та 1.

Статистика C має наступні характеристики при справедливості нульової гіпотези:

$$MC = 0, DC = \frac{n_1 n_2}{n(n-1)} \sum_{i=1}^{n} (M(i,n))^2$$

Критерій Ван дер Вардена

Статистика критерію має вигляд $V = \sum_{i=1}^{n_1} \Phi^{-1}\left(\frac{R_{i,n}}{n+1}\right)$, де $\Phi^{-1}(x)$ – функція обернена до функції розподілу з параметрами 0 та 1, причому коли справедлива нульова гіпотеза, то

$$MV = 0, DV = \frac{n_1 n_2}{n(n-1)} \sum_{i=1}^{n} \left(\Phi^{-1}\left(\frac{i}{n+1}\right)\right)^2$$

Критерій Вілкоксона

Статистика критерію має вигляд $S = \sum_{i=1}^{n_1} R_{i,n}$, причому коли справедлива нульова гіпотеза, то

$$MS = \frac{n_1(n+1)}{2}, \quad DS = \frac{n_1 n_2 n}{12}$$

Процедура використання статистик C, V, S для перевірки гіпотези H_0 однакова: позначимо через U деяку статистику (C, V, або S), $\bar{U} = \frac{U - MU}{\sqrt{DU}}$. В залежності від альтернативної гіпотези H_0 приймається якщо:

- $|\bar{U}| < u_{1-\alpha/2}$, якщо альтернатива H_{11},
- $\bar{U} < u_{1-\alpha}$, якщо альтернатива H_{12},
- $\bar{U} > u_{\alpha}$, якщо альтернатива H_{13},

де u_α – квантиль рівня α для нормального розподілу з параметрами 0 та 1.

По мірі спадання потужності критерії розташовуються так: критерій нормальних міток Фішера, критерій Ван дер Вардена, критерій Вілкоксона.

3.4.3 Загальний випадок

Гіпотеза H_0: $F_1(x) = F_2(x) = \ldots = F_k(x)$, $\forall x$ з рівнем значимості α $(0 < \alpha < 1)$.

1. Будуємо об'єднану вибірку $\nu_1, \nu_2, \ldots, \nu_n$ об'єму $n = \sum_{i=1}^{k} n_i$, а потім відповідний варіаційний ряд $\nu_{(1)}, \nu_{(2)}, \ldots, \nu_{(n)}$.

2. Для кожної ξ_i підрахуємо статистику $K_i = \sum_{j=N_i-n_I+1}^{N_i} \psi(R_{j,n})$.
   Для неї підійде будь-яка статистика з попередніх критеріїв:

$$C_i = \sum_{j=N_i-n_i+1}^{N_i} M(R_{j,n}, n),$$

$$V_i = \sum_{j=N_i-n_i+1}^{N_i} \Phi^{-1}\left(\frac{R_{i,n}}{n+1}\right),$$

$$S_i = \sum_{j=N_i-n_i+1}^{N_i} R_{j,n}.$$

3. Далі знаходимо їх стандартизовані значення

$$\bar{K}_i = \frac{K_i - MK_i}{\sqrt{DK_i}}.$$

4. Тепер рахуємо статистику

$$X^2 = \sum_{i=1}^{k} \bar{K}_i^2.$$

Нульова гіпотеза приймається якщо $X^2 < \chi^2_\alpha(k-1)$, де $\chi^2_\alpha(k)$ – $\alpha \cdot 100\%$ процентна точка χ^2-розподілу з k степенями свободи.

3.4.4 Перевірка симетрій розподілу ранговими критеріями

Маємо ряд незалежних спостережень $x_1, x_2, \ldots, x_n$ над випадковою величиною ξ з функцією розподілу $F(x)$. Перевіримо симетричність розподілу відносно точки x_0.

Гіпотеза: для дискретної випадкової величини $H_0 : F(x_0+x) = 1-F(x_0-x+0)$, $\forall x$, або ж для неперервної випадкової величини $H_0 : p(x_0+x) = p(x_0-x)$, $\forall x$.

Перевірка проводиться з деяким рівнем значимості α $(0 < \alpha < 1)$.

Побудуємо послідовність $z_1, z_2, \ldots, z_n$, де $z_i = |x_i - x_0|$, $i = \overline{1,n}$, а далі сформуємо варіаційний ряд $z_{(1)}, z_{(2)}, \ldots, z_{(n)}$.

Абсолютним рангом виміру x_i називається порядковий номер значення $x_i = |x_i - x_0|$ у варіаційному ряді $z_{(1)}, z_{(2)}, \ldots, z_{(n)}$, позначатимемо його як $R^+_{i,n}$ $(i = \overline{1,n})$.

Розіб'ємо вибірку $x_1, x_2, \ldots, x_n$ на дві вибірки, в першій всі виміри більше x_0. в іншій решта. Позначення для індексів з першої вибірки $I^+ = \{i : x_i > x_0, i = \overline{1,n}\}$. Тепер можна порівняти дві наші вибірки на однорідність.

Аналог критерію нормальних міток

Статистика критерію має вигляд $C^+ = \sum_{i \in I^+} M^+(R^+_{i,n}, n)$, при справедливості H_0:

$$MC^+ = \frac{n}{\sqrt{2\pi}}, \quad DC = \frac{1}{4}\sum_{i=1}^{n}(M^+(i,n))^2.$$

Аналог критерію Ван дер Вардена

Статистика критерію має вигляд $V^+ = \sum_{i \in I^+} \Phi^{-1}\left(\frac{1}{2} + \frac{R^+_{i,n}}{2(n+1)}\right)$, при справедливості H_0:

$$MV^+ = \frac{1}{2}\sum_{i=1}^{n}\Phi^{-1}\left(\frac{1}{2} + \frac{i}{2(n+1)}\right), \quad DV^+ = \frac{1}{4}\sum_{i=1}^{n}\left(\Phi^{-1}\left(\frac{1}{2} + \frac{i}{2(n+1)}\right)\right)^2$$

Аналог критерію Вілкоксона

Статистика критерію має вигляд $S^+ = \sum_{i \in I^+} R^+_{i,n}$, при справедливості H_0:

$$MS^+ = \frac{(n+1)n}{4}, \quad DS^+ = \frac{n(n+1)(2n+1)}{24}.$$

Нехай U^+ – одна з вищенаведених статистик. Стандартизуємо U^+:

$$\bar{U}^+ = \frac{U^+ - NU^+}{\sqrt{DU^+}}.$$

Область прийняття гіпотези H_0: $|\bar{U}^+| < u_{1-\alpha/2}$, де u_α – квантиль рівня α нормального розподілу з параметрами 0 та 1.

3.4.5 Визначення рангів у випадку наявності нерозрізнимих значень

Нехай $\nu_1, \nu_2, \ldots, \nu_n$ – об'єднана вибірка, побудована на основі спостережень над змінними, що досліджуються. Існує два варіанти однакових спостережень:

1. спостереження стосуються однієї змінних, тоді використовується метод випадкового рангу: ранг однакових елементів є довільним числом, яке припало на цю множину значень.

2. спостереження стосуються різних змінних, тоді використовують або вже відомий нам метод випадкового рангу, або метод середньої мітки: всім рівним спостереженням присвоюють середнє значення мітки підраховане за множиною рангів, яка відповідає цій групі нерозрізними вимірів.

Корекція алгоритмів рангових критеріїв:

$DC = \frac{n_1 n_2}{n(n-1)} \sum_{i=1}^{g} \tau_i \bar{M}_i^2$ для критерію нормальних міток Фішера, для критерію Ван дер Вардена $DV = \frac{n_1 n_2}{n(n-1)} \sum_{i=1}^{g} \tau_i \left(\bar{\Phi}_i^{-1}\right)^2$, де g – кількість груп нерозрізнимих спостережень, τ_i – кількість значень у i-ій групі.

3.5 Видалення аномальних спостережень

Критерії розглядаються для нормальних вибірок. Нехай маємо вибірку x_1, x_2, …, x_n. Гіпотеза H_0: найбільш підозрілий на аномальність вимір не є викидом із рівнем значущості α $(0 < \alpha < 1)$.

3.5.1 Випадок скалярних спостережень

1. *Критерій Граббса*

   Нагадаємо, що $\overline{x}(n) = \frac{1}{n} \sum_{i=1}^{n} x_i$, $s(n) = \sqrt{\frac{1}{n} \left(\sum_{i=1}^{n} x_i^2 - n\overline{x}^2(n)\right)}$. Побудуємо послідовність z_1, z_2, …, z_n, де $z_i = |x_i - \overline{x}(n)|$, $i = \overline{1,n}$, та її варіаційний ряд $z_{(1)}, z_{(2)}, \ldots, z_{(n)}$. Введемо допоміжне позначення $z_{(j)} = \left|x_{i(j)} - \overline{x}(n)\right|$, $j = \overline{1,n}$, тоді підозралим на аномальність є елемент вибірки що відповідає останньому елементу $z_{(n)}$ варіаційного ряду, тобто $x_{i(n)}$. Розглянемо статистику

$$T(n) = \frac{x_{i(n)} - \overline{x}(n)}{s(n)}.$$

Областю прийняття гіпотези H_0 буде $|T(n)| < T_{\alpha/2}(n)$, де $T_{\alpha/2}(n)$ – $100\frac{\alpha}{2}\%$ точка розподілу статистики

$$\frac{x_{i(n)} - \overline{x}(n)}{s(n)}.$$

Якщо спостереження аномальне, то його викидають і так продовжується допоки останній елемент нової вибірки перестає бути аномальним.

2. *Критерій Томпсона*

   Модифікація критерію Граббса із статистикою

$$t(n) = \frac{\sqrt{n-2} \cdot T(n)}{\sqrt{n-1-T^2(n)}}.$$

   Область прийняття гіпотези H_0: $|t(n)| < t_{\alpha/2}(n-2)$, де $t_{\alpha/2}(n-2)$ – $100\frac{\alpha}{2}\%$ точка розподілу Стьюдента з $(n-2)$ степенями свободи.

3. *Критерій Тітьєна-Мура*

   Дозволяє перевіряти кілька спостережень на аномальність одразу. На базі варіаційного ряду з критерію Граббса обчислюємо статистику:

$$E(n,k) = \frac{\sum_{i=1}^{n-k}(z_{(i)} - \overline{z}(n-k))^2}{\sum_{i=1}^{n}(z_{(i)} - \overline{z}(n))^2},$$

   де на аномальність перевіряються останні k членів варіаційного ряду, а $\overline{z}(m) = \frac{1}{m}\sum_{i=1}^{m} z_{(i)}$. Область прийняття гіпотези H_0: $E(n,k) \geq E_{1-\alpha}(n,k)$, де $E_{1-\alpha}$ – квантиль рівня α розподілу статистики $E(n,k)$.

3.5.2 Випадок векторних значень

Нехай маємо векторні величини $x_1, x_2, \ldots, x_n$, $x_i \in \mathbb{R}^q$, $i = \overline{1,n}$. Гіпотеза H_0: найбільш підозрілий на аномальність вектор не є викидом із рівнем значимості α $(0 < \alpha < 1)$.

Критерій на базі F-статистики

Введемо величини

$$\bar{x}_i = \frac{1}{n-1}\sum_{j\neq i} x_j, \quad \widehat{\sum}_i = \frac{1}{n-2}\sum_{j\neq i}(x_j - \bar{x}_i)(x_j - \bar{x}_i)^T, \quad i = \overline{1,n}.$$

Обчислимо вибіркові *відстані Махаланобіса*:

$$D_i^2 = (x_i - \bar{x}_i)^T \cdot \widehat{\sum}_i^{-1}(x_i - \bar{x}_i), \quad i = \overline{1,n}.$$

Визначимо наступну статистику:

$$F_i = \frac{(n-1)(n-1-q)}{n(n-2)q} D_i^2.$$

Знаходимо індекс i_0 найбільш підозрілого вектора як $i_0 = \arg\max_i F_i$. Гіпотеза H_0 приймається як $F_{i_0} < F_\alpha(q, n-1-q)$, де $F_\alpha(q, n-1-q)$ – $100\alpha\%$ точка F-розподілу з q та $n-1-q$ степенями свободи.

4 Кореляційний аналіз

З'ясовує наявність статистичною зв'яжу між змінними, що досліджуються.

Схема по які досліджується наявність статистичного зв'язку:

1. Вводиться характеристика статистичного зв'язку.
2. Обчислюється точкова чи інтервальна характеристика цієї оцінки.
3. Здійснюється перевірка на значимість характеристики статистичного зв'язку.

4.1 Випадок кількісних змінних

Нехай є змінні (скалярні) η, ξ (η – залежна, ξ – незалежна).

Треба з'ясувати по спостереженнях за η, ξ істотність зв'язку між ними. Зв'язок шукається у *вигляді функції регресії*:

$$f(x) = M(\eta/\xi = x), \quad g(x) = D(\eta/\xi = x)$$

– *умовні матсподівання і дисперсія*. $D\eta = Df(\xi) + Mg(\xi)$.

Індексом кореляції для змінних η та ξ називається

$$I_{\eta\xi} = \sqrt{\frac{Df(\xi)}{D\eta}} = \sqrt{1 - \frac{Mg(\xi)}{D\eta}}.$$

Властивості:

1. $0 \leq I_{\eta\xi} \leq 1$.
2. якщо $I_{\eta\xi} = 0$, то зв'язку між η та ξ *немає*.
3. якщо $I_{\eta\xi} = 1$, то є *функціональний* зв'язок між ними.

Коефіцієнт детермінації $I^2_{\eta\xi} = \frac{Df(\xi)}{D\eta}$ вказує яка частина варіації η визначаються варіацією функцій регресії в точці ξ.

4.2 Коефіцієнт кореляції

Розглянемо нормальний випадок. Є дві величини ξ та η.

$$\xi \sim N(m_\xi, \sigma^2_\xi), x_1, \ldots, x_n \qquad \eta \sim N(m_\eta, \sigma^2_\eta), y_1, \ldots, y_n \qquad .$$

$$r_{\eta\xi} = \frac{M(\xi - M\xi)(\eta - M\eta)}{\sqrt{D\xi D\eta}},$$

вибіркове значення:

$$\widehat{r}_{\eta\xi} = \frac{\sum_{i=1}^{n}(x_i - \bar{x}(n))(y_i - \bar{y}(n))}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x}(n))^2 \sum_{i=1}^{n}(y_i - \bar{y}(n))^2}}.$$

Можна довести, що $I_{\eta\xi} = |r_{\eta\xi}$.

Властивості:

1. $|r_{\eta\xi}| \leq 1$.
2. якщо $r_{\eta\xi} = 0$ то зв'язок між η і ξ відсутній.
3. Якщо $r_{\eta\xi} = \pm 1$ то зв'язок між η і ξ лінійний, причому *формула зв'язку*: $\eta = m_\eta + r_{\eta\xi}\sigma_\eta \frac{\xi - m_\xi}{\sigma_\xi}$.

4. Нехай $r_{\eta\xi} > 0$. Якщо $\xi \uparrow$, то і $\eta \uparrow$.

5. Нехай $r_{\eta\xi} < 0$. Якщо $\xi \uparrow$, то $\eta \downarrow$.

6. Якщо коефіцієнт кореляції прийняв проміжне значення, то перевіряємо гіпотезу H_0: $r_{\eta\xi} = 0$, $0 < \alpha < 1$. Для перевірки H_0 будемо розглядати статистику:

$$t(n-2) = \frac{\sqrt{n-2}r_{\eta\xi}}{\sqrt{1-r_{\eta\xi}^2}}.$$

Ця статистика має асимптотичний t-розподіл Стьюдента з $(n-2)$ степенями свободи. Тоді логічно вважати, що H_0 гіпотеза несправедлива, коли статистика приймає екстремальні значення. $|t(n-2)| < t_{\alpha/2}(n-2)$ – область прийняття гіпотези H_0, де $t_\alpha(n)$ – $100\alpha\%$ – точки t-розподілу Стьюдента з v степенями свободи.

4.3 Характеристика парного статистичного зв'язку в загальному випадку

Нехай спостерігаються ξ і η, з'ясуємо наявність зв'язку. Розглянемо 2 випадки:

- випадок групованих (за ξ) даних;
- випадок не згрупованих даних.

1. Спостереження над залежною змінною η: $y_{11}, \ldots, y_{1m_1}, \ldots, y_{s1}, \ldots, y_{sm_s}$, s інтервалів групування, в i-му інтервалі m_i спостережень.

   $\bar{y}_i$ – вибіркове середнє спостережень по групі i, $\bar{y}$ – загальне вибіркове середнє.

   S_y^2 – вибіркове значення дисперсії η, $S_{y(x)}^2$ – зважене вибіркове значення дисперсії вибіркових середніх $\bar{y}_i$.

   Запишемо оцінку для індексу кореляції (кореляційне відношення):

$$\widehat{p}_{\eta\xi} = \sqrt{\frac{S_{y(x)}^2}{S_y^2}}.$$

Властивості такі ж, як і в індексу кореляції. З'ясувалося, що

$$F = \frac{\widehat{p}^2_{\eta\xi}}{1-\widehat{p}^2_{\eta\xi}} \cdot \frac{n-s}{s-1}$$

має асимптотичний розподіл, який тотожньо рівний $F(s-1, n-s)$. Припускаємо, що спостереження нормальні.

Область прийняття гіпотези: $F < F_\alpha(s-1, n-s)$, де F_α – $100\alpha\%$-точка F-розподілу з параметрами $s-1$, $n-s$.

2. Функцію регресії f апроксимують на деякому класі параметричних функцій з точністю до вектор-параметру θ. $f(x,\theta), \theta \in \mathbb{R}^p$.

По спостереженням досліджуваних змінних: $\xi : x_1, \ldots, x_n, \eta : y_1, \ldots, y_n$.

Методом найменших квадратів визначаємо $\widehat{\theta}$, далі отримуємо деяку апроксимацію функції регресії $f(x,\theta)$.

Апроксимація індексу кореляції даних у вигляді:

$$\widehat{I}_{\eta\xi} = \sqrt{1 - \frac{\frac{1}{n-p}\sum_{i=1}^{n}(y_i - f(x_i,\widehat{\theta}))^2}{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y}(n))^2}}.$$

Приклад θ: $f(x,\theta) = \sum_{i=1}^{N} \theta_i f_i(x)$.

4.4 Частинний коефіцієнт кореляції

Частинним коефіцієнтом кореляції для змінних $x^{(i)}$, $x^{(j)}$ будемо називати величину:

$$r^*_{ij} = -\frac{R_{ij}}{\sqrt{R_{ii}R_{jj}}},$$

де R_{ij} – алгебраїчне доповнення для елемента (i, j) у звичайній кореляційній матриці:

$$R = \begin{pmatrix} 1 & r_{01} & \dots & r_{0q} \\ r_{10} & 1 & \dots & r_{1q} \\ \vdots & \vdots & \ddots & \vdots \\ r_{q0} & r_{q1} & \dots & 1 \end{pmatrix},$$

де r_{ij} – звичайний коефіцієнт кореляції.

Властивості частинного співпадають з властивостями звичайного коефіцієнта кореляції. Вибіркове значення коефіцієнта кореляції:

$$\widehat{r}^*_{ij} = -\frac{\widehat{R}_{ij}}{\sqrt{\widehat{R}_{ii}\widehat{R}_{jj}}},$$

$$\widehat{R} = \begin{pmatrix} 1 & \widehat{r}_{01} & \dots & \widehat{r}_{0q} \\ \widehat{r}_{10} & 1 & \dots & \widehat{r}_{1q} \\ \vdots & \vdots & \ddots & \vdots \\ \widehat{r}_{q0} & \widehat{r}_{q1} & \dots & 1 \end{pmatrix}.$$

При $r^*_{ij} = 0$ зв'язку не існує.

При $r^*_{ih} = \pm 1$ то зв'язок функціональний.

Якщо коефіцієнт прийняв проміжне значення, то перевіряється гіпотеза H_0: $r^*_{ij} = 0$, $\alpha > 0$. Використовуємо статистику:

$$t(n - m - 2) = \frac{\sqrt{n - m - 2} \cdot \widehat{r}^*_{ij}}{\sqrt{1 - (\widehat{r}^*_{ij})^2}},$$

де m – кількість третіх змінних зафіксованих на певному рівні.

Вона має t-розподіл Стьюдента з $n - m - 2$ степенями свободи. Критична область – обасть великих і малих значень. *Область прийняття* має вигляд:

$$|t(n - m - 2)| < t_{\alpha/2}(n - m - 2),$$

де $t_{\alpha/2}(n - m - 2)$ – $100\alpha/2\%$ точка t-розподілу Стьюдента з $n - m - 2$ степенями свободи.

4.5 Множинний коефіцієнт кореляції

Розглянемо залежну змінну η і незалежну змінну $\vec{\xi} \in \mathbb{R}^q$. Для з'ясування зв'язку використовується *множинний коефіцієнт кореляції*:

$$R_{\eta\xi} = \sqrt{D(f\xi)}D\eta = \sqrt{1 - \frac{M(g(\xi))}{D\eta}},$$

де умовні матсподівання і дисперсія визначаються так же як і раніше тільки для векторної ξ.

Множинний коефіцієнт детермінації: $R^2_{\eta\xi}$.

Властивості множинного коефіцієнта кореляції такі ж, як і звичайного коефіцієнта кореляції.

Вибіркове значення. Функцію регресії $f(\vec{x}, \theta)$ апроксимуємо на деякому класі параметричних функцій. $\vec{\xi} : \vec{x}_1, \ldots, \vec{x}_n$, $\eta : y_1, \ldots, y_n$.

По отриманим спостереженням методом найменших квадратів знаходимо оцінку $\widehat{\theta}$ і підставляємо в апроксимацію. Звідси оцінка нормальна.

$$\widehat{R}_{\eta\xi} = \sqrt{1 - \frac{\frac{1}{n-p}\sum_{i=1}^{n}(y_i - f(\vec{x}, \widehat{\theta}))^2}{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y}(n))^2}}.$$

4.5.1 Методика використання

Якщо $R_{\eta\xi} = 0$, то зв'язок неістотній.

Якщо $R_{\eta\xi} = 1$, то зв'язок функціональний.

Якщо $R_{\eta\xi}$ приймає проміжне значення, то перевіряється гіпотеза H_0.

Проаналізуємо наступну статистику:

$$F = \frac{\widehat{R}^2_{\eta\xi}}{1 - \widehat{R}^2_{\eta\xi}} \cdot \frac{n-p}{n-1}.$$

Вона має асимптотичний розподіл, який співпадає з F-розоділом з параметрами $(p-1, n-p)$. Тоді область прийняття – це область невеликих значень:

$$F < F_\alpha(p-1, n-p).$$

4.6 Кореляційний аналіз порядкових змінних

Нехай η – залежна порядкова змінна і $\vec{\xi} = (\xi_1, \ldots, \xi_q)^*$. Нехай $\xi^{(i)}$ – вектор спостережень над i-ою змінною, тобто $\xi_k^{(i)}$ – i-а змінна k-го предмету.

Розяглянемо ранжировку (перестановку чисел від 1 до n): $x^{(i)} = \left(x_1^{(i)}, \ldots, x_n^{(i)}\right)^*$, де $x_k^{(i)}$ – ранг k-го предмету по i-ій змінній.
Якщо всі прояви об'єктів різні, то маємо $x^{(0)}$, $x^{(1)}$, …, $x^{(q)}$ – спостереження, $x_*^{(0)}$, $x_*^{(1)}$, …, $x_*^{(q)}$ – ранжировка.

При наявності по деякій зміні групи об'єктів з однаковим проявом досліджуваної властивості, цим об'єктам присвоюють ранг, який дорівнює середньому арифметичному номерів тих місць, які припали на цю групу об'єктів з нерозрізненими рангами. Такий ранг називається зв'язаний (об'єднаний).

Будується таблиця рангів для доступу до об'єкта:

	$x^{(1)}$	$x^{(2)}$	$\ldots$	$x^{(q)}$
1	$x_1^{(1)}$	$x_1^{(2)}$	$\ldots$	$x_1^{(q)}$
2	$x_2^{(1)}$	$x_2^{(2)}$	$\ldots$	$x_2^{(q)}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
n	$x_n^{(1)}$	$x_n^{(2)}$	$\ldots$	$x_n^{(q)}$

рядки – об'єкти, стовпчики – змінні.

4.6.1 Характеристики парного статистичного зв'язку

Розглядаємо характеристики $x^{(i)}$, $x^{(j)}$.

В якості характеристики парного зв'язку між змінними $x^{(i)}$ та $x^{(j)}$ може-

мо використати *коефіцієнт Спірмана*, який визначається таким чином:

$$\widehat{\tau}_{ij}^{(s)} = 1 - \frac{\left\| x_*^{(i)} - x_*^{(j)} \right\|_2^2}{\frac{n^3 - n}{6}}.$$

Властивості рангу коефіцієнта Спірмана:

1. $-1 \leq \widehat{\tau}_{ij}^{(s)} \leq 1$;
2. якщо $\widehat{\tau}_{ij}^{(s)} = 0$, то зв'язок відсутній;
3. якщо $\widehat{\tau}_{ij}^{(s)} = 1$, то ранжировки по змінним співпадають, $x_*^{(i)} = x_*^{(j)}$;
4. якщо $\widehat{\tau}_{ij}^{(s)} = -1$, то ранжировки по змінним протилежні, $x_*^{(i)} = x_*^{(j)}$.

Розглянемо випадок наявності *нерозрізнених рангів*. В цьому випадку використовується *модифікований коефіцієнт*. Ранговий коефіцієнт Спірмана обчислюється за формулою:

$$\widehat{\widehat{\tau}}_{ij}^{(s)} = \frac{\frac{n^3 - n}{6} - \left\| x_*^{(i)} - x_*^{(j)} \right\|_2^2 - T^{(i)} - T^{(j)}}{\sqrt{\left(\frac{n^3 - n}{6} - 2T^{(i)} \right) \left(\frac{n^3 - n}{6} - 2T^{(j)} \right)}},$$

де

$$T^{(i)} = \frac{1}{12} \sum_{g=1}^{m^{(i)}} \left(n_g^{(i)} \right)^3 - n_g^{(i)}$$

– *корегуючий коефіцієнт*, $m^{(i)}$ – кількість груп об'єктів з нерозрізненими рангами по змінній $x^{(i)}$, $n_g^{(i)}$ – кількість членів у g-й групі нерозрізнимих рангів по i-й змінній.

Коли коефіцієнт приймає проміжне значення, то перевіряємо гіпотезу $H_0 : \widehat{\tau}_{ij}^{(s)} = 0$. Якщо об'єм вибірки невеликий, то перевіряємо по таблиці, при $n = 4..10$. Якщо ж $n > 10$, то розглядаємо статистику

$$\frac{\sqrt{n-2}\widehat{\tau}_{ij}^{(s)}}{\sqrt{1 - \left(\widehat{\tau}_{ij}^{(s)} \right)^2}},$$

що має t-розподіл Стьюдента з $(n-2)$ степенями свободи. Область прийняття гіпотези:

$$\left| \frac{\sqrt{n-2}\widehat{\tau}_{ij}^{(s)}}{\sqrt{1-\left(\widehat{\tau}_{ij}^{(s)}\right)^2}} \right| < t_{\alpha/2}(n-2).$$

Розглянемо іншу характеристику: *коефіцієнт Кендала. Ранговим коефіцієнтом Кендала* для змінних $x^{(i)}$ та $x^{(j)}$ називається величина

$$\widehat{\tau}_{ij}^{(k)} = \frac{4v\left(x_*^{(i)}, x_*^{(j)}\right)}{n(n-1)},$$

де $v\left(x_*^{(i)}, x_*^{(j)}\right)$ – кількість перестановок сусідніх елементів у ранжировці $x_*^{(i)}$, яка приводить її до ражировки $x_*^{(j)}$.

Властивості рангу коефіцієнта Кендала:

1. $-1 \leq \widehat{\tau}_{ij}^{(k)} \leq 1$;
2. якщо $\widehat{\tau}_{ij}^{(k)} = 0$, то зв'язок відсутній;
3. якщо $\widehat{\tau}_{ij}^{(k)} = 1$, то ранжировки по змінним співпадають, $x_*^{(i)} = x_*^{(j)}$;
4. якщо $\widehat{\tau}_{ij}^{(k)} = -1$, то ранжировки по змінним протилежні, $x_*^{(i)} = x_*^{(j)}$.

Якщо є наявні нерозрізнені ранжировки, то використовують *модифікований коефіцієнт Кендала*:

$$\widehat{\widehat{\tau}}_{ij}^{(k)} = \frac{\widehat{\tau}_{ij}^{(k)} - \dfrac{u^{(i)} - u^{(j)}}{n(n-1)}}{\sqrt{\left(1 - \dfrac{U^{(i)}}{n(n-1)}\right)\left(1 - \dfrac{U^{(j)}}{n(n-1)}\right)}},$$

де

$$U^{(i)} = \sum_{g=1}^{m^{(i)}} n_g^{(i)}\left(n_g^{(i)} - 1\right),$$

$m^{(i)}$ – кількість груп об'єктів з нерозрізненними рангами по змінній $x^{(i)}$, $n_g^{(i)}$ – кількість членів у g-й групі нерозрізнимих рангів по i-й змінній.

Коли коефіцієнт приймає проміжне значення, то перевіряємо гіпотезу $H_0 : \widehat{\tau}_{ij}^{(s)} = 0$. Якщо об'єм вибірки невеликий, то перевіряємо по таблиці, при $n = 4..10$. Якщо ж $n > 10$, то використовуємо

$$\left|\widehat{\tau}_{ij}^{(k)}\right| \leq U_{\alpha/2}\sqrt{\frac{2(2n+5)}{9n(n-1)}}.$$

4.6.2 Характеристика множинних рангових статистичних зв'язків

Нехай аналізується m змінних $\zeta = \left(x^{(k_1)}, x^{(k_2)}, \ldots, x^{(k_m)}\right)^*$.

В якості характеристики використовується *коефіцієнт конкордації. Коефіцієнтом конкордації* для змінної $\zeta = \left(x^{(k_1)}, x^{(k_2)}, \ldots, x^{(k_m)}\right)^*$ називають величину

$$\widehat{w}_\zeta = \frac{12}{m^2(n^3-n)}\sum_{i=1}^{n}\left(\left(\sum_{j=1}^{m} x_i^{(j)}\right) - \frac{m(n+1)}{2}\right)^2.$$

Властивості:

1. $0 \leq \widehat{w}_\zeta \leq 1$;
2. якщо $\widehat{w}_\zeta = 1$, то ранжировки по змінним співпадають:
3. якщо $\widehat{w}_\zeta = 0$, то відсутній зв'язок між ранжировками.

У випадку двох нерозрізнених рангів використовуємо модифікований коефіцієнт $\widehat{\widehat{w}}_\zeta$:

$$\widehat{\widehat{w}}_\zeta = \frac{\left(\left(\sum_{j=1}^{m} x_i^{(k_j)}\right) - \frac{m(n+1)}{2}\right)^2}{\frac{m^2(n^3+n)}{2} - m\sum_{j=1}^{m} T^{(k_i)}},$$

де

$$T^{(k_j)} = \frac{1}{12}\sum_{g=1}^{m_j}\left(\left(n_g^{(k_j)}\right)^2 - n_g^{(k_j)}\right).$$

Якщо $\widehat{w}_\zeta$ приймає проміжне значення, то робимо перевірку на значимість $H_0 : \widehat{w}_\zeta = 0$, $0 < \alpha < 1$. Коли $n = 3..7$, $m = 2..20$, то за таблицею. Якщо $n > 7$, $m > 20$, то розглядаємо статистику $\widehat{w}_\zeta$:

$$\widehat{w}_\zeta < \frac{\chi^2_\alpha(n-1)}{m(n-1)},$$

має χ^2-розподіл з $(n-1)$ степенем свободи.

4.7 Кореляційний аналіз номінальних змінних

Нехай є змінна η яка має r_1 градацій, та змінна ξ яка має r_2 градацій:

	1	2	...	r_1	$\sum$
1	n_{11}	n_{12}	...	n_{1r_1}	n_{1*}
2	n_{21}	n_{22}	...	n_{2r_1}	n_{2*}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
r_2	$n_{r_2 1}$	$n_{r_2 2}$	...	$n_{r_2 r_1}$	$n_{r_2 *}$
$\sum$	n_{*1}	n_{*2}	...	n_{*r_1}	n_{**}

Де n_{ij} – кількість таких спостережень $\{\eta = i, \xi = j\}$, позначимо $n_{i*} = \sum_{j=1}^{r_1} n_{ij}$, $n_{*j} = \sum_{i=1}^{r_2} n_{ij}$.

Вводимо статистику яка називається *квадратичне спряження* і позначається

$$\chi^2_{\eta\xi} = \sum_{i=1}^{r_1}\sum_{j=1}^{r_2} \frac{\left(n_{ij} - \frac{n_{i*}n_{*j}}{n}\right)^2}{\frac{n_{i*}n_{*j}}{n}}.$$

Коефіцієнти:

1. $\phi_{\eta\xi} = \sqrt{\frac{\chi^2_{\eta\xi}}{n}}$ – *середнє значення квадратичної спряженості*;
2. $P_{\eta\xi} = \sqrt{\frac{\chi^2_{\eta\xi}}{n + \chi^2_{\eta\xi}}}$ – *коефіцієнт Пірсона*;
3. $T_{\eta\xi} = \sqrt{\frac{\chi^2_{\eta\xi}}{n\sqrt{(r_1-1)(r_2-1)}}}$ – *коефіцієнт Чупрова*;
4. $T_{\eta\xi} = \sqrt{\frac{\chi^2_{\eta\xi}}{n\min(r_1-1, r_2-1)}}$ – *коефіцієнт Крамера.*

Властивості коефіцієнтів

1. $k_{\eta\xi} \geq 0$, якщо коефіцієнт $p_{\eta\xi} \leq 1$;
2. $k_{\eta\xi} = 0$, тоді зв'язок відсутній.

Перевірку на значимість роблять шляхом перевірки гіпотези H_0: $\chi^2_{\eta\xi} = 0$, $0 < \alpha < 1$. З'ясувалося, що $\chi^2_{\eta\xi}$ має хі-квадрат розподіл з $(r_1 - 1)(r_2 - 1)$ степенями свободи.

Тоді критична область – область великих значень, та область прийняття гіпотези: $\chi^2_{\eta\xi} < \chi^2((r_1 - 1)(r_2 - 1))$.

4.8 Інформаційна міра зв'язку

Ентропією для змінної ξ називають величину

$$H_\xi = -\sum_i p(x_i) \ln(p(x_i)) = -M \ln(p(\xi)).$$

Позначимо ймовірність, з якою приймається пара значень (x_i, y_j) як $p(x_i, y_j)$, тоді *ентропією для пари* (ξ, η) називається величина

$$H_{\eta\xi} = -\sum_{i,j} p(x_i, y_j) \ln(p(x_i, y_j)).$$

Інформаційна міра зв'язку визначається як $I_{\eta\xi} = H_\xi + H_\eta - H\xi\eta$.

Властивості інформаційної міри зв'язку

1. $I_{\eta\xi} \geq 0$;
2. якщо $I_{\eta\xi} = 0$ то зв'язок між ξ та η відсутній.

Спробуємо визначити вибіркове значення. Спочатку визначимо вибіркове значення для ентропії η та ξ:

$$\widehat{H}_\eta = -\sum_i \frac{n_i}{n} \cdot \ln\left(\frac{n_i}{n}\right) \qquad \widehat{H}_\xi = -\sum_j \frac{n_j}{n} \cdot \ln\left(\frac{n_j}{n}\right),$$

де n_i, n_j – абсолютні частоти прийняття змінними η і ξ відповідно своїх значень. Далі, якщо n_{ij} – абсолютна частота прийняття змінними η і ξ пари відповідних значень, то

$$\widehat{H}_{\eta\xi} = -\sum_{i,j} \frac{n_{ij}}{n} \cdot \ln\left(\frac{n_{ij}}{n}\right),$$

звідки

$$\widehat{I}_{\eta\xi} = \frac{1}{n}\left(\sum_{i,j} n_{ij} \cdot \ln(n_{ij}) - \sum_i n_i \cdot \ln(n_i) - \sum_j n_j \cdot \ln(n_j) + n \ln n\right).$$

При перевірці характеристики на значимість $H_0 : \widehat{I}_{\eta\xi} = 0$, $0 < \alpha < 1$. $\widehat{\widehat{I}}_{\eta\xi} = 2n\widehat{I}_{\eta\xi} - n_0$, де n_0 – кількість нульових елементів у таблиці спряженості.

Виявилось, що така перетворена статистика: $\widehat{\widehat{I}}_{\eta\xi} < \chi^2((r_1 - 1)(r_2 - 1))$.

Оскільки ця статистика невід'ємна, то область прийняття гіпотези матиме такий вигляд $\widehat{\widehat{I}}_{\eta\xi} < \chi^2_\alpha((r_1 - 1)(r_2 - 1))$.

5 Дисперсійний аналіз

Нехай є деяка кількісна скалярна змінна η та є деякий вектор якісних змінних ξ.

Дисперсійний аналіз займається побудовою математичної моделі зв'язку між цими змінними, а також їх аналізом.

Приклад. З'ясувати вплив сорту зернових на врожай.

Залежна змінна – врожайність, якісна змінна – сорт зернових та тип міндобрив. η – врожайність, ξ_1 – сорт зернових, I_1 – всього сортів, ξ_2 – тип міндобрив, I_2 – всього міндобрив.

Позначимо y_{ijk} – врожайність i-го сорт зернових при застосуванні j-го тип міндобрив на k-му полі, α_i – вплив на залежну змінну, β_j – вплив на якісну змінну. Розглянемо тепер модель врожайності

$$y_{ijk} = \mu + \alpha_i + \beta_j + c_{ij} + e_{ijk},$$

де μ, α_i, β_j, c_{ij} – невідомі параметри, e_{ijk} – помилка моделі (наприклад вплив поля), c_{ij} – вплив взаємодії i-ї градації першої змінної та j-ї градації другої змінної на врожайність зернових.

Ця модель лінійна по всім параметрам, тому для її розв'язку напрошується метод найменших квадратів(МНК).

5.1 Перевірка лінійних гіпотез для регресійної моделі

Лінійну регресійну модель в матричному вигляді запишемо так: $\vec{y} = X\vec{\alpha} + \vec{e}$, де $\vec{y} \in \mathbb{R}^N$, $X \in \mathbb{R}^{N\times p}$, $\vec{\alpha} \in \mathbb{R}^p$, $\vec{e} \in \mathbb{R}^N$.

Нехай ранг X дорівнює p (тобто матриця має повний ранг по стовпчикам), а сама оцінка знаходилась з критерію

$$Q(\alpha) = \|y - X\alpha\|_2^2 \to \min$$

і має вигляд:

$$\widehat{\alpha} = (X^T X)^{-1} X^T y.$$

Розглянемо таку множину $L = \{\vec{\alpha} : A\vec{\alpha} = \vec{b}, \operatorname{rang} A = q\}$, де $A \in \mathbb{R}^{q\times p}$ (тобто матриця має повний ранг по рядкам).

Розглянемо оптимальну оцінку $\vec{\alpha}$ для лінійної моделі методом найменших квадратів при наявності лінійних обмежень L:

$$\widehat{\alpha}_L = \widehat{\alpha} + (X^T X)^{-1} A^T (A(X^T X)^{-1} A^T)^{-1} (b - A\widehat{\alpha}).$$

Теорема 1. *При справедливості гіпотези $H_0 : A\vec{\alpha} = b$, $\operatorname{rang} A = q$, $\gamma > 0$ наступна статистика*

$$F = \frac{\dfrac{Q(\widehat{\alpha}_L) - Q(\widehat{\alpha})}{q}}{\dfrac{Q(\widehat{\alpha})}{N-p}}$$

має асимптотичний F-розподіл $F(q, N-p)$, а відповідна область прийняття гіпотези: $F < F_\gamma(q, N-p)$, де $F_\gamma(q, N-p)$ це $100\gamma\%$ точка F-розподілу.

Зауваження. При справедливості гіпотези H_0: $Q(\widehat{\alpha}_L) - Q(\widehat{\alpha}) = \|X\widehat{\alpha}_L - X\widehat{\alpha}\|_2^2$.

Зауваження. Розглянемо наступну гіпотезу: $H : \alpha_i = 0$, $0 < \gamma < 1$, тобто перевіряємо, чи суттєво відхиляється від нуля відносний вплив i-ої градації. Тоді статистика, побудована по умовам теореми F_i називається *частинною статистикою по i-й змінній*, а відповідний критерій, побудований на цій статистиці, для перевірки гіпотези H, називається *частинним F-критерієм по i-й змінній.*

5.2 Однофакторний дисперсійний аналіз

Нехай η – деяка скалярна кількісна змінна, ξ – деяка якісна незалежна змінна, яка має I градацій. При фіксованій i-й градації вважаємо, що є J_i спостережень над залежною змінною, які позначимо через y_{ij}. Розглянемо модель

$$y_{ij} = \mu + \mu_i + e_{ij}, \quad i = \overline{1, I}, \quad j = \overline{1, J}.$$

Припустимо, що помилки моделі:

1. нормально розподілені $N(0, \sigma^2)$, $\sigma^2 > 0$;
2. незалежні.

Запишемо цю модель в матричному вигляді: $y = X\vec{\mu} + e$, де $X \in \mathbb{R}^{N \times (I+1)}$, $\vec{\mu} = (\mu, \mu_1, \ldots, \mu_I)^T$, де $N = \sum_{i=1}^{I} J_i$ – загальна кількість вимірів. Для існування оцінки $\widehat{\mu}$ потрібно $\operatorname{rang} X = I$, тобто

$$\exists w_i : \sum_{i=1}^{I} w_i \mu_i = 0, \quad \sum_{i=1}^{I} w_i = 1, \quad \forall i : w_i > 0.$$

Отримаємо оцінку $\vec{\mu}$ паралельно оцінивши суттєвість впливу однієї змінної на іншу. Математично це рівносильно перевірці гіпотези $H : \mu_1 = \mu_2 = \ldots = \mu_I = 0$, $0 < \gamma < 1$.

Запишемо цю гіпотезу як $H : A\mu = \vec{0}$, де $\operatorname{rang} A = I - 1$, тоді можна скористатися теоремою вище. Згідно теореми область прийняття гіпотези матиме вигляд:

$$F = \frac{\dfrac{Q(\widehat{\mu}_L) - Q(\widehat{\mu})}{I - 1}}{\dfrac{Q(\widehat{\mu})}{N - I}} < F_\gamma(I - 1, N - 1).$$

Зауваження. Якщо $I - 1$ параметр є нульовими, то і I-й параметр теж нульовий, згідно умов із w.

Знайдемо $Q(\widehat{\mu})$. Відомо, що оцінка методом найменших квадратів є розв'язком системи нормальних рівнянь $X^T X\widehat{\mu} = X^T y$. Перепишемо систему в розгорнутому матричному вигляді:

$$\begin{pmatrix} N & J_1 & J_2 & \ldots & J_I \\ J_1 & J_1 & 0 & \ldots & 0 \\ J_2 & 0 & J_2 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ J_1 & 0 & 0 & \ldots & J_I \end{pmatrix} \begin{pmatrix} \widehat{\mu} \\ \widehat{\mu}_1 \\ \widehat{\mu}_2 \\ \vdots \\ \widehat{\mu}_I \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{I} \sum_{j=1}^{J_i} y_{ij} \\ J_1 \bar{y}_1 \\ J_2 \bar{y}_2 \\ \vdots \\ J_I \bar{y}_I \end{pmatrix},$$

де $\bar{y}_i$ – вибіркове середнє i-ої групи.

Розглянемо окреме рівняння системи, отримаємо

$$\forall i : J_i\widehat{\mu} + J_i\widehat{\mu}_i = J_i\bar{y}_i \Rightarrow \bar{y}_i = \widehat{\mu} + \widehat{\mu}_i,$$

тобто оцінкою абсолютного впливу i-тої градації є $\bar{y}_i$. Звідси маємо $\widehat{\mu}_i = \bar{y}_i - \widehat{\mu}$.

З рівнянь на w знаходиться $\widehat{\mu} = \sum_{i=1}^{I} w_i\bar{y}_i$, тобто $\widehat{\mu}_i = \bar{y}_i - \sum_{i=1}^{I} w_i\bar{y}_i$.

Згідно викладеного вище отримаємо

$$Q(\widehat{\mu}) = \sum_{i=1}^{I}\sum_{j=1}^{J_i}(y_{ij} - \bar{y}_i)^2.$$

L – лінійне обмеження, еквівалентне умовам на w, тому $\widehat{\mu}_L : \widetilde{X}^T\widetilde{X}\widehat{\mu}_L = \widetilde{X}^T y$, де $\widetilde{X} = (1, \ldots, 1)^T$ – вектор стовпчик при обмеженні L, тому

$$\widehat{\mu}_L = \frac{1}{N}\sum_{i=1}^{I}\sum_{j=1}^{I_j} y_{ij} = \bar{y}$$

– загальне середнє по всім вимірам.

Зауважимо що тут ми часто користуємося μ як для позначення вектора $\vec{\mu}$ так і на позначення його першої компоненти, у марних сподіваннях на те що з контексту зрозуміло коли що використовується.

Наслідок/Зауваження: зі структури матриці X випливає, що

$$Q(\widehat{\mu}_L) - Q(\widehat{\mu}) = \|X\widehat{\mu} - X\widehat{\mu}_L\|_2^2 = \sum_{i=1}^{I}\sum_{j=1}^{J_i}(\bar{y}_i - \bar{y})^2 = \sum_{i=1}^{I} J_i(\bar{y}_i - \bar{y})^2.$$

Підставляючи це все в F знайдемо

$$F = \frac{\dfrac{\sum_{i=1}^{I} J_i(y_i - \bar{y})^2}{I-1}}{\dfrac{\sum_{i=1}^{I}\sum_{j=1}^{J_i}(y_{ij} - \bar{y}_i)^2}{N-I}} < F_\gamma(I-1, N-I).$$

5.3 Таблиця результатів однофакторного дисперсійного аналізу

Джерело варіацій	Сума квадратів	КСС	ССК	F	γ_*
Між градаціями	S_A	$I-1$	$\bar{S}_A = \dfrac{S_A}{I-1}$	$F = \dfrac{\bar{S}_A}{\bar{S}_E}$	γ_*
Всередині градацій	S_E	$N-I$	$\bar{S}_E = \dfrac{S_E}{N-I}$		

де

$$S_E = \sum_{i=1}^{I}\sum_{j=1}^{J_i}(y_{ij}-\bar{y}_i)^2, \qquad S_A = \sum_{i=1}^{I} J_i(\bar{y}_i-\bar{y})^2,$$

γ_* – максимальна ймовірність при котрій гіпотеза приймається (рівень значимості).

5.4 Перевірка контрастів

Якщо гіпотеза $\vec{\mu} = \vec{0}$ несправедлива, то нас цікавить питання: чи є серед градацій такі, що мають суттєві відхилення від нуля. Намагаємось виявити серед усіх градацій такі їх підмножини, що середні по ним несуттєво відхиляються від середніх сусідніх підмножин.

Абсолютний вплив i-ої градації, це $\theta_t = \mu - \mu_i$, та $\widehat{\theta} = \bar{y}_i$. *Контрастами* будемо називати статистики вище наведеного вигляду для коефіцієнтів яких справедлива умова. Нас цікавить перевірка гіпотези:

$$H_0 : \sum_i c_i\theta_i = 0, \quad \sum_i c_i = 0, \quad \theta_i > 0.$$

Алгоритм перевірки гіпотези

1. Будуємо *довірчий інтервал* з рівнем довіри $(1-\alpha)$.
2. *Якщо нуль належить* цьому інтервалу, то гіпотезу вважаємо справедливою, інакше її відхиляємо.

5.5 Довірчі інтервали для контрастів

1. Якщо c_i – *наперед задані*, то

$$\left|\sum_i c_i\bar{y}_i - \sum_i c_i\theta_i\right| < \bar{S}_E\sum_i \frac{c_i^2}{J_i}t_{\alpha/2}(N-I),$$

де

$$\left|\sum_i c_i\bar{y}_i - \sum_i c_i\theta_i\right|$$

– *усереднена залишкова сума квадратів.*

2. *S-метод Шофе*

$$\left|\sum_i c_i\bar{y}_i - \sum_i c_i\theta_i\right| < \sqrt{\bar{S}_E \sum_i \frac{c_i^2}{J_I}(I-1)F_\alpha(I-1, N-I)}.$$

3. *T-метод Кьюні* орієнтований на побудову довірчих інтервалів для контрастів статистики $\theta_i - \theta_j$, крім того припускаємо, що кількість вимірів при кожній градації однакова, тобто $\forall i : J_i = J$, тоді

$$|(\bar{y}_i - \bar{y}_j) - (\theta_i - \theta_j)| < \bar{S}_E q_\alpha(I, N-I),$$

де $q_\alpha(I, N-I)$ – $100\alpha\%$-на точка Стьюдентизованого розмаху.

Зауваження. Нехай є η_i, $i = \overline{1, I}$ – нормально розподілені величини з параметрами 0 та 1; статистика $\chi^2(k)$ має χ^2-розподіл; $\{\eta_i\}, \chi^2(k)$ – незалежні.

Тоді величина $\max_i \eta_i - \min_i \sqrt{\chi^2(k)}$ має розподіл Стьюдентизованого розмаху з параметрами i, k.

Нехай є η – залежна кількісна змінна, якісні змінні: T_A приймає значення з I-тої градації, T_B приймає значення з J-тої градації.

Якщо фактор приймає значення з i-тої градації, фактор з j-тої градації, то кількість вимірів $N = \sum_{i=1}^{I}\sum_{j=1}^{J} k_{ij}$.

В загальному випадку модель дисперсійного аналізу приймає вигляд:

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + e_{ijk},$$

де

- y_{ijk} – спостереження за нзалежною змінною η;
- μ – загальне середнє;

- α_i – кількісний вираз відносно впливу i-тої градації ξ_A на η (головний ефект i-того рівня фактору);
- β_j – кількісний вираз відносно впливу j-тої градації фактору ;
- γ_{ij} – кількісний вираз відносно впливу i-тої градації фактору , j-ї градації фактору на залежну змінну η;
- e_{ijk} – помилки моделі:
  1. $e_{ijk} \sim N(0, \sigma^2)$, $\sigma^2 > 0$;
  2. e_{ijk} – незалежні.

Нехай є така модель, відомі тільки e_{ijk} та які градації чому відповідають.

Припускаємо:

$$\forall v_i > 0 \exists w_j > 0 : \forall i, j : \sum_{i=1}^{I} v_i \alpha_i = \sum_{j=1}^{J} w_i \gamma_{ij} = \sum_{j=1}^{J} w_j p_j = \sum_{i=1}^{I} v_j \gamma_{ij} = 0$$

Наявність таких лінійних обмежень дозволяє стверджувати, що *методом МНК* ми зможемо знайти оцінки вектора невідомих параметрів з цієї моделі.

Зауваження. З метою спрощення виразів розглянемо випадок $\forall i, j$: $v_i = \frac{1}{I}, w_j = \frac{1}{J}, k_{ij} = k, N = I \cdot J \cdot K$. Перевіремо гіпотези $H_A : \alpha_1 = \alpha_2 = \ldots = \alpha_I$, $\gamma > 0$; $H_B : \beta_1 = \beta_2 = \ldots = \beta_J$, $\gamma > 0$; $H_{AB} : \forall i, j : \gamma_{ij} = 0$, $\gamma > 0$.

Оцінки МНК вищезгаданої моделі при наявності вищезгаданих лінійних обмежень обчислюються за формулами ($*$ замість індексу означає, що по цьому індексу береться усереднене):

$$\widehat{\mu} = \bar{y}, \quad \widehat{\alpha}_i = y_{i**} - \bar{y}, \quad \widehat{\beta} = y_{*j*} - \bar{y}, \quad \widehat{\gamma}_{ij} = y_{ij*} - y_{i**} - y_{*j*} + \bar{y}.$$

Введемо величини

$$S_E = \sum_{i,j,k} (y_{ijk} - \bar{y})^2, \quad S_A = JK \sum_{i} (y_{i**} - \bar{y})^2,$$
$$S_B = IK \sum_{j} (y_{*j*} - \bar{y})^2, \quad S_{AB} = K \sum_{i,j} (y_{ij*} - y_{i**} - y_{*j*} + \bar{y})^2.$$

Розглянемо *таблицю результатів*

Джерело варіації	СК	КСС
Головний ефект фактору A	S_A	$(I-1)$
Головний ефект фактору B	S_B	$(J-1)$
Головний ефект взаємодії	S_{AB}	$(I-1)(J-1)$
Загальна сума квадратів	S_E	$IJ(K-1)$

$$H_A : F_A = \frac{\frac{S_A}{I-1}}{\frac{S_E}{IJ(K-1)}} < F_\gamma(I-1, IJ(K-1)).$$

$$H_B : F_B = \frac{\frac{S_B}{J-1}}{\frac{S_E}{IJ(K-1)}} < F_\gamma(J-1, IJ(K-1)).$$

$$H_{AB} : F_{AB} = \frac{\frac{S_{AB}}{(I-1)(J-1)}}{\frac{S_E}{IJ(K-1)}} < F_\gamma((I-1)(J-1), IJ(K-1)).$$

Зауваження У випадку коли факторів більше двох, тоді в правій частині крім наявності головних ефектів будуть ще й ефекти по всім факторам, тобто $y_{i_1 i_2 \ldots} = \mu + \alpha_{i_1} + \ldots + w_{i_f}$.

6 Регресійний аналіз

Регресійний аналіз займається побудовою математичної моделі зв'язку з кількісними змінними.

Нехай маємо η – залежну кількісну скалярну змінну, $\xi \in \mathbb{R}^p$ – вектор незалежних змінних.

Зв'язок між змінними істотній. Ми хочемо побудувати математичну модель зв'язку між ними. В кореляційному аналізі його явне задання шукаємо у вигляді функції регресії (теоретично): $f(x) = M(\eta/\xi = x)$.

Лема 1. *Нехай* $M\eta^2 < \infty$. *Позначимо* $\mathcal{R} : \mathbb{R}^q \to \mathbb{R}^1$, *тоді*

$$f(\cdot) = \arg\min_{g(\cdot)\in\mathcal{R}} M((\eta - g(\xi))^2).$$

Доведення. Припустимо, що $Mg^2(\xi) < \infty$. Розглянемо

$$M(\eta - g(\xi))^2 = M(\eta - f(\xi) + f(\xi) - g(\xi))^2 =$$
$$= M(\eta - f(\xi))^2 + 2M(\eta - f(\xi))(f(\xi) - g(\xi)) + M(f(\xi) - g(\xi))^2 \geq$$
$$\geq M(\eta - f(\xi))^2 + 2M\left((M(\eta/\xi) - f(\xi))(f(\xi) - g(\xi))\right) = M(\eta - f(\xi)^2).$$

□

Нехай $\eta = f(\xi) + \epsilon$, позначимо спостереження над η як $y(i)$ і спостереження над ξ – $x(i)$, $i = \overline{1, N}$. $y(k) = f(x(k)) + e(k)$, $k = \overline{1, N}$.

6.1 Основні етапи розв'язку задачі регресійного аналізу

1. Вибір класу апроксимуючих функцій $g(x, \alpha) \in \mathcal{J}$.
2. Отримання точеної чи множинної оцінки для вектора α невідомих параметрів, а також її характеристики розсіяності. $M(\widehat{\alpha} - \alpha)(\widehat{\alpha} - \alpha)^T$, де α – точне значення, $\widehat{\alpha}$ – оцінка для α.
3. Перевірка на значимість відхилення параметрів моделі від нуля: $H_0 : \alpha_i = 0$ з рівнем значимості $0 < \gamma < 1$.
4. Перевірка на адекватність отриманої моделі:

$$M((\eta - g(\xi, \alpha))^2), \quad \frac{1}{N}\sum_{k=1}^{N}(y(k) - g(x(k), \alpha))^2.$$

6.2 Класичний регресійний аналіз

Постановка: в якості апроксимації береться функція, лінійна по параметрах:

$$y(k) = \sum_{i=1}^{p} \phi_i(x'(k))\alpha_i + e(k), \quad k = \overline{1, N}.$$
$$x(k) = (\phi_1(x'(k)), \phi_2(x'(k)), \ldots, \phi_p(x'(k)))^T, \quad \alpha = (\alpha_1, \alpha_2, \ldots, \alpha_p)^T.$$
$$y(k) = \sum_{i=1}^{p} x_i(k)\alpha_i + e(k) = x^T(k)\alpha + e(k), \quad k = \overline{1, N}.$$

Функції $x_i(k)$ називаються *регресорами*, це просто деякі функції від векторів незалежних змінних. Введемо ще трохи позначень:

$y = (y(1), \ldots, y(N))^T$, $e = (e(1), \ldots, e(N))^T$, $X = (x^T(1), \ldots, x^T(N))^T$.

Тоді модель можна переписати у вигляді $y = X\alpha + e$.

6.3 Основні припущення класичного регресійного аналізу

1. Помилки моделі вважати нормально розподіленими $e(k) \sim N(0, \sigma^2)$.
2. Вони незалежні.
3. X – відома, і має повний ранг по стовпчикам $rankX = p$.
4. Немає ніяких обмежень на вектор невідомих параметрів α.

Будемо шукати оцінку мінімізуючи функціонал:

$$\sum_{k=1}^{N} e^2(k) = \|y - X\alpha\|_2^2 = \sum_{k=1}^{N} (y(k) - X^T(k)\alpha)^2 \to \min_{\alpha}.$$

Точкою мінімуму буде $\widehat{\alpha} = (X^T X)^{-1} X^T y$.

Доведемо це, враховуючи $\nabla_\alpha(\alpha^T\beta) = \beta$, $\nabla_\alpha(\alpha^T A\alpha) = (A + A^T)\alpha$.

Розпишемо функціонал: $\|y - X\alpha\|_2^2 = \alpha^T X^T X\alpha - 2\alpha^T X^T y + \|y\|_2^2$.

$$(\nabla_\alpha \|y - X\alpha\|_2^2)|_{\alpha=\widehat{\alpha}} = (2X^T X\alpha - 2X^T y)|_{\alpha=\widehat{\alpha}}.$$

Отже, в системі $X^T X\widehat{\alpha} = X^T Y$, матриця $X^T X$ – не вироджена, тому

$$\widehat{\alpha} = (X^T X)^{-1} X^T y.$$

Таким чином

$$\widehat{y}(k) = X^T(k)\widehat{\alpha}, \quad \widehat{y} = X\widehat{\alpha}, \quad \widehat{\sigma}^2 = \frac{\|y - X\widehat{\alpha}\|_2}{N - p}$$

– *незміщена оцінка максимальної правдоподібності.*

6.4 Властивості оцінок

1. $\widehat{\alpha} \sim N(\alpha, \sigma^2(X^T X)^{-1})$;
2. стастистика $\dfrac{(\widehat{\alpha} - \alpha)X^T X(\widehat{\alpha} - \alpha)}{\sigma^2} \sim \chi^2(p)$;
3. $\widehat{y} \sim N(X\alpha, \sigma^2 X(X^T X)^{-1} X^T)$;
4. $M\widehat{\sigma}^2 = \sigma^2$ – *незміщена* оцінка;

5. $\widehat{\alpha}$, $\widehat{\sigma}^2$ – незалежні оцінки (випадкові величини);

6. $\widehat{\alpha}$, $\widehat{y}$, $\widehat{\sigma}^2$ – *ефективні на класі незміщених оцінок.*

Доведемо деякі властивості:

1.

$$\widehat{\alpha} = (X^TX)^{-1}X^Ty = (X^TX)^{-1}X^T(X\alpha + e) = \\ = (X^TX)^{-1}X^TX\alpha + (X^TX)^{-1}X^Te = \alpha + (X^TX)^{-1}X^Te.$$

Тобто $M\widehat{\alpha} = \alpha$.

$$M(\widehat{\alpha} - \alpha)(\widehat{\alpha} - \alpha)^T = M((X^TX)^{-1}X^Tee^TX(X^TX)^{-1}) = \\ \sigma^2(X^TX)^{-1}(X^TX)(X^TX)^{-1} = \sigma^2(X^TX)^{-1}.$$

2. Доведемо спершу лему:

Лема 2. *Нехай* $\xi \in \mathbb{R}^q$, $\xi \sim N(m, R)$, $R > 0$, *тоді виконується*

$$(\xi - m)^TR^{-1}(\xi - m) \sim \chi^2(q).$$

Доведення. $M(\xi - m) = \vec{0}$. $R^{1/2}(\xi - m) \sim N(\vec{0}, E_q)$.

$$R = T\begin{pmatrix} \mu_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \mu_n \end{pmatrix}T^T$$

$\|R^{1/2}(\xi - m)\|_2^2 \sim \chi^2(q)$ співпадає з квадратичною формою. □

Сама властивість є просто наслідком застосування леми до $\widehat{\alpha}$.

3. $\widehat{y} = X\widehat{\alpha} \sim N(X\alpha, \sigma^2X(X^TX)^{-1}X^T)$.

6.5 Довірчі області та інтервали для невідомих параметрів моделі

1. Довірча область для α з рівнем значимості $(1 - \gamma)$.

З властивості 2 маємо

$$\chi^2(p) = \frac{(\widehat{\alpha}-\alpha)X^TX(\widehat{\alpha}-\alpha)}{\sigma^2} =$$
$$= \frac{e^TX(X^TX)^{-1}(X^TX)(X^TX)^{-1}X^Te}{N-p} = \frac{e^TX(X^TX)^{-1}X^Te}{\sigma^2}$$

$\chi^2(N-p) \sim \dfrac{(N-p)\widehat{\sigma}^2}{\sigma^2}$, де $\widehat{\sigma}^2 = \dfrac{\|y-X\alpha\|_2^2}{N-p}$. Маємо

$$\frac{\|y-X(X^TX)^{-1}X^Ty\|_2^2}{N-p} = \frac{\|(E-X(X^TX)^{-1}X^T)(X\alpha+e)\|_2^2}{N-p} =$$
$$\frac{\|PX\alpha+Pe\|_2^2}{N-p} = \frac{e^TP^TPe}{N-p} = \frac{e^TP^2e}{N-p} = \frac{e^TPe}{N-p},$$

де $P = E - X(X^TX)^{-1}X^T$, $PX = O$, $P^2 = P$. Тобто, маємо $\widehat{\sigma}^2 = \frac{e^TPe}{N-p}$, звідки

$$\frac{(N-p)\widehat{\sigma}^2}{\sigma^2} = \frac{(N-p)e^TPe}{(N-p)\sigma^2} = \frac{e^TPe}{\sigma^2}.$$

Лема 3. *Нехай $\xi \in \mathbb{R}^n$, $\xi \sim N(m, \sigma^2E)$, A, B – матриці розмірності $N \times N$, тоді квадратичні форми $\xi^TA\xi$, $\xi^TB\xi$ будуть незалежні тоді і тільки тоді, коли $AB = O$.*

$$\frac{(\widehat{\alpha}-\alpha)^TX^TX(\widehat{\alpha}-\alpha)}{\sigma^2P} < F_\gamma(p, N-p)$$

– *довірча область.*

2. Довірчий інтервал для α_i отримуємо аналогічно:

$$\left|\frac{\widehat{\alpha}_i-\alpha}{\widehat{\sigma}\sqrt{d_i}}\right| < t_{\gamma/2}(N-p).$$

3. Інтервали Бонферроні.

Якщо для кожної компоненти побудувати довірчий інтервал з рівнями довіри $1-\gamma/(\dim\alpha)$ і A_i – ймовірність того, що i-та компонента $\in$ своєму довірчому інтервалу, який побудований за попереднім пунктом. Тоді $\bigcap\limits_{i=1}^{p} A_i$ – всі компоненти $\in$ своїм довірчим інтервалам. Зрозуміло що

$$P\left\{\bigcap_{i=1}^{p} A_i\right\} \geq 1-\gamma.$$

6.6 Перевірка на значимість параметрів моделі

1. Перевіряємо гіпотезу: $H_0 : \alpha = 0$ – вектор параметрів, γ – рівень довіри.

   Якщо H_0 справедлива, то наступна статистика $\dfrac{\widehat{\alpha}^T X^T X \widehat{\alpha}}{p\widehat{\sigma}^2} \sim F(p, N-p)$.

   Критична область – область великих значень. Область прийняття:

   $$\frac{\|X\widehat{\alpha}\|_2^2}{p\widehat{\sigma}^2} < F_\gamma(p, N-p)$$

   – $100\gamma\%$ точка F-розподілу з параметрами $(p, N-p)$.

2. Перевіряємо гіпотезу: $H_0 : \alpha_i = 0$, $\gamma > 0$. $\frac{\widehat{\alpha}_i - \alpha_i}{\widehat{\sigma}\sqrt{d_i}} \sim t(N-p)$. При справедливості H_0, статистика: $t_i = \frac{\widehat{\alpha}_i}{\widehat{\sigma}\sqrt{d_i}} \sim t(N-p)$, де d_i – i-й діагональний елемент матриці $(X^T X)^{-1}$.

   Критична область – область дуже малих і дуже великих значень.

   Область прийняття:

   $$\frac{|\widehat{\alpha}_i|}{\widehat{\sigma}\sqrt{d_i}} < t_{\gamma/2}(N-p).$$

3. Нехай в моделі перший регресор $x_1(K) \equiv 1$. Тоді

   $$y(k) = \alpha_1 + \sum_{i=2}^{p} \alpha_i x_i(k) + e(k), \quad k = \overline{1, N}.$$

   Потрібно перевірити на значимість всі α_i, $i = \overline{2, p}$. $H_0 : \alpha_2 = \alpha_3 = \ldots = \alpha_p = 0$, $\gamma > 0$. Якщо H_0 справедлива, то достатньо обмежитись тим, що $y(k) = \alpha_1 + e'(k)$, $k = \overline{1, N}$.

   Або H_0 можна переписати у вигляді: Множина коефіцієнту кореляції залежної змінної, множина незалежної змінної суттєво відхиляються від 0.

Тобто $\widehat{\widehat{\alpha}} = \widehat{\widehat{y}}(K) = \bar{y} = \dfrac{1}{N}\sum\limits_{k=1}^{N} y(k)$.

$$H_0 : \begin{pmatrix} 0 & 1 & \cdots & 0 \\ 0 & 0 & \ddots & 0 \\ \cdots & \cdots & \cdots & 1 \\ 0 & 0 & \cdots & 0 \end{pmatrix}, \quad \alpha = (0, \ldots, 0)^T, \quad \gamma > 0, \quad \operatorname{rang} A = p{-}1.$$

Тоді за теоремою про перевірку лінійної гіпотези:

$$F = \frac{Q(\widehat{\alpha}_L) - Q(\widehat{\alpha})}{(p-1)\widehat{\sigma}^2} = \frac{\|X\widehat{\alpha} - X\widehat{\alpha}_L\|_2^2}{(p-1)\widehat{\sigma}^2} = \\ = \frac{\sum\limits_{i=1}^{N}(X^T(k)\widehat{\alpha} - \bar{y})^2}{(p-1)\widehat{\sigma}^2} = F < F_\gamma(p-1, N-p).$$

6.7 Довірчі інтервали та області для функції регресії

Нас цікавить довірчий інтервал і область для величин $x^T(k)\alpha$ та для всього $X\alpha$.

1. *Довірча область для вектора значень функції регресії* $X\alpha$. Враховуючи довірчу область знайдену для $\widehat{\alpha}$ у попередньому пункті, можемо записати:

$$\frac{(\widehat{\alpha} - \alpha)^T X^T X(\widehat{\alpha} - \alpha)}{p\widehat{\sigma}^2} = \frac{\|X\widehat{\alpha} - X\alpha\|_2^2}{p\widehat{\sigma}^2} < F_\gamma(p, N-p).$$

2. *Довірча область для* $x^T(k)\alpha$. За властивістю 3:

$$\widehat{y}(k) \sim N(x^T(k)\alpha, \sigma^2 x^T(k)(X^T X)^{-1}x(k)),$$

тому наступна статистика

$$\frac{\widehat{y}(k) - x^T(k)\alpha}{\sigma\sqrt{x^T(k)(X^T X)^{-1}x(k)}} \sim N(0,1).$$

За властивістю 4:

$$\frac{(N-p)\widehat{\sigma}^2}{\sigma^2} \sim \chi^2(N-p),$$

тому

$$\frac{\widehat{y}(k) - x^T(k)\alpha}{\frac{\sigma\sqrt{x^T(k)(X^TX)^{-1}x(k)}}{\sqrt{\frac{(N-p)\widehat{\sigma}^2}{(N-p)\sigma^2}}}} = \frac{\widehat{y}(k) - x^T(k)\alpha}{\widehat{\sigma}\sqrt{x^T(k)(X^TX)^{-1}x(k)}} \sim t(N-p),$$

тому довірча область має вигляд

$$\left| \frac{\widehat{y}(k) - x^T(k)\alpha}{\widehat{\sigma}\sqrt{x^T(k)(X^TX)^{-1}x(k)}} \right| < t_{\gamma/2}(N-p).$$

6.8 Перевірка на адекватність

Перевірка на адекватність здійснюється шляхом перевірки спільномірності оцінки $\widehat{\sigma}$, отриманої на базі основної вибірки, з оцінкою σ^2 на базі спостережень з додаткової вибірки вимірів у фіксованій точці фазового простору.

Випадки неадекватності:

1. або більше параметрів;
2. або менше параметрів.

1. Нехай модель істинна: $My = X\alpha + X_1\alpha_1$, вибрана модель: $My = X\alpha$.

   Не всі регресори включені. $\widehat{\alpha} = (X^TX)^{-1}X^Ty$.

   Знаходимо оцінки $\widehat{\alpha}$, $\widehat{\sigma}$: $\widehat{\sigma}^2 = \dfrac{\|y - X\widehat{\alpha}\|_2^2}{N-p}$.

   Оцінка $\widehat{\alpha}$ зміщена, тобто

   $$M\widehat{\alpha} = \alpha + \Delta\alpha = \alpha + (X^TX)^{-1}X^TX_1\alpha_1.$$

   $\widehat{\alpha}$ – неслушна оцінка, $\widehat{\sigma}$ – зміщена оцінка, тобто:

   $$M\widehat{\sigma}^2 = \sigma^2 + \Delta\sigma^2.$$

2. Нехай в істинній моделі менше параметрів, ніж у вибраній, у якій є зайві. Тобто: $My = X\alpha$ істинна, а $My = X\alpha + X_1\alpha_1$ – вибрана.

$$\bar{X} = (X \vdots X_1), \quad \bar{\alpha} = (\alpha, \alpha_1)^T.$$

В цьому випадку: $\widehat{\bar{\alpha}}$ – незміщена, і $M\widehat{\bar{\alpha}} = (\alpha, O)^T$. $\widehat{\bar{\alpha}}$ – слушна оцінка.

Оцінка $\widehat{\sigma}^2$ – незміщена, $M\widehat{\sigma}^2 = \sigma^2$.

Ми втратили точності оцінки у вигляді:

$$M(\widehat{\bar{\alpha}} - M\widehat{\bar{\alpha}})(\widehat{\bar{\alpha}} - M\widehat{\bar{\alpha}})^T = \sigma^2(X^TX)^{-1} + \Delta u, \quad \Delta u \geq 0.$$

Точність оцінки може збільшуватись.

6.9 Не класичний регресійний аналіз

Розглянемо тепер випадки, коли деякі з припущень моделі, яку ми розглядали не виконуються.

6.9.1 Випадок корельованих збурень

Для нашої моделі $y = X\alpha + e$, де в класичному регресійному аналізі $e \sim N(\vec{0}, \sigma^2 E)$.

Нехай тепер це не так, а $e \sim N(\vec{0}, R)$, де $R > 0$ – кореляційна матриця. Спробуємо і для цього випадку отримати оцінку з тими ж хорошими властивостями, як і в класичному випадку.

$$R^{-1/2}y = R^{-1/2}X\alpha + R^{-1/2}e,$$

тоді, згідно властивості лінійного перетворення нормально розподілених випадкових величин $R^{-1/2}e \sim N(\vec{0}, E)$, тобто тут помилки уже будуть незалежними.

Ввівши позначення отримаємо нову модель $\widetilde{y} = \widetilde{X}\alpha + \widetilde{e}$ причому $\widetilde{e} \sim N(\vec{0}, R)$ і тоді будуємо оцінку класичним методом найменших квадратів :

$$\widehat{\alpha} = (\widetilde{X}^*\widetilde{X})^{-1}\widetilde{X}^T\widetilde{y} = (X^TR^{-1/2}R^{-1/2}X)^{-1}X^TR^{-1/2}R^{-1/2}y = \\ = (X^TR^{-1}X)^{-1}X^TR^{-1}y.$$

Цю оцінку називають *Марківською*.

На неї розповсюджуються всі раніше сформульовані властивості, все прекрасно, але радіти рано: треба знати матрицю R, а на практиці вона зазвичай невідома. Часто можна знайти лише деяке наближення до неї, а тоді наближене значення R^{-1} може бути дуже далеко від реального значення, а потім ще й до $X^T R^{-1} X$ треба шукати обернену...

Марківська оцінка має наступні властивості:

1. $\widehat{\alpha} \sim N(\alpha, (X^T R^{-1} X)^{-1})$.
2. $\widehat{\alpha}$ ефективна на класі всіх незміщених оцінок.
3. Якщо кількість вимірів зростає (тобто залежить від об'єму), тоді

$$\widehat{\alpha}(N) = (X_N^T R_N^{-1} X_N)^{-1} X_N^T R_N^{-1} y_N$$

є сильно слушною тоді і тільки тоді, коли

$$(X_N^T R_N^{-1} X_N)^{-1} \xrightarrow[N\to\infty]{} O.$$

Марківську оцінку можна розглядати також як *розв'язок наступної оптимізаційної задачі*:

$$\widehat{\alpha} = \arg\min_{\alpha} \|e\|^2_{R^{-1}}.$$

Ця оцінка – це фактично один із прикладів, коли теорія – теорією, а практика – практикою.

6.9.2 Оцінки параметрів в умовах мультиколінеарності

Для $y = X\alpha + e$ в класичних припущеннях ми вважали, що $\operatorname{rang} X = p$, тобто повний по стовпчиках ($p = \dim \alpha$). Знімемо це припущення.

Нехай $\operatorname{rang} X = p - q$, $q \geq 1$. Тоді, як результат, отримаємо, що класична оцінка просто не існує, оскільки матриця XX^T буде виродженою. Фактично це означає, що оцінка методом найменших квадратів буде не єдиною, і можна вибрати ту, яка найкраща для нас.

Нехай $\exists a_i : Xa_i = \vec{0}$.

В цьому випадку кажуть, що ми знаходимось в умовах строгої мультиколінеарності. Якщо ж ця умова виконується тільки приблизно, то кажуть,

що ми знаходимось в умовах мультиколінеарності.

Проаналізуємо ці випадки:

1. Строга мультиколінеарність: оцінка не єдина, множина оцінок є розв'язком системи
$$X^T X\widehat{\alpha} = X^T y.$$
В принципі тут можна використати псевдообернення, але це виходить за рамки нашого курсу.

2. Мультиколінеарність: класична оцінка існує і вона єдина, але це теорія, а на практиці матриця $X^T X$ виходить погано обумовленою (існують власні числа близькі нуля), тому оцінка *нестійка* та *малоефективна* ($\sigma(X^T X)^{-1}$ може бути дуже великою).

Є різні підходи до розв'язання цієї проблеми. Викладемо такий:

6.10 Гребеневі оцінки (ridge-оцінки)

Їх запропонував Херл в 1962р. **Ідея проста**: вводимо параметр θ і збурюємо $X^T X$:
$$\widehat{\alpha}(\theta) = (X^T X + \theta E)^{-1} X^T y,$$
де $\theta > 0$, але досить мале.

Введення цього зміщення відсуне і результат, але матриця $(X^T X + \theta E)$ вже не буде погано обумовленою. Виявилось також, що при деякому θ ця оцінка буде мати навіть кращі властивості, ніж класична, а саме, – меншу середньоквадратичну помилку.

Теорема 2. *Для гребеневої оцінки виконуються наступні властивості:*

1. $\widehat{\alpha}(\theta) = (E+\theta(X^T X)^{-1})^{-1}\widehat{\alpha}$, *де* $\widehat{\alpha}$ – *класична оцінка,* $\widehat{\alpha} = (X^T X)^{-1} X^T y$.
2. $\widehat{\alpha}(\theta) = (E - \theta(X^T X + \theta E)^{-1})^{-1}\widehat{\alpha}$.
3. $M\widehat{\alpha}(\theta) = \alpha + \Delta\alpha$, *де зміщення* $\Delta\alpha = -\theta B\alpha$, *а* $B = (X^T X + \theta E)^{-1}$.
4. $M(\theta) = M\|\widehat{\alpha}(\theta) - \alpha\|_2^2 = M\|\widehat{\alpha}(\theta) - M\widehat{\alpha}(\theta)\|_2^2 + \|\Delta\alpha\|_2^2$ *(перша рівність – просто позначення).*

6.11 Нелінійний регресійний аналіз

Моделі $y(k) = f(x(k), \alpha) + e(k)$, $k = \overline{1, N}$, де $e(k)$ – помилка моделі, поділяються на два підкласи:

1. Внутрішньо лінійні;
2. Внутрішньо нелінійні.

Внутрішньо лінійні – це моделі які шляхом перетворень зводяться до розв'язання деякої лінійної задачі.
Внутрішньо нелінійні – це моделі для яких не існує шляхів зведення до лінійної моделі.

Оцінка $\widehat{\alpha}$ шукається як розв'язок нелінійної (бажано квадратичної) задачі

$$\widehat{\alpha} = \arg\min_{\alpha} \sum_{k} e^z(k) = \arg\min_{\alpha} I(\alpha).$$

7 Коваріаційний аналіз

Треба побудувати модель залежності кількісної змінної від як від якісної так і від кількісної. Специфіка постановки задачі $\vec{\xi_g} \in \mathbb{R}^q$ – вектор незалежних якісних змінних, $\vec{\xi} \in \mathbb{R}^p$ – вектор кількісних змінних, η – залежна скалярна кількісна змінна.

На характеристику впливають як якісні так і кількісні змінні. Нехай $y(k)$ – спостереження над η. Тоді модель класичного коваріаційного аналізу має вигляд:

$$y(k) = \sum_{i=1}^{q} x_g^{(i)}(k)\alpha_g^{(i)} + \sum_{i=1}^{p} x^{(i)}(k)\alpha^{(i)} + e(k) = x_g^T(k)\alpha_g + X^T(k)\alpha + e(k), \quad k = \overline{1, N}.$$

Перепишемо модель матричному вигляді: $y = X_g\alpha_g + X\alpha + e$.

Для знаходження оцінок параметрів моделі використовується покроковий метод найменших квадратів.

Основні припущення:

1. $e \sim N(\vec{0}, \sigma^2 E)$
2. стовпчики матриці X не залежать від умов експерименту.

3. вважаємо, що лінійні обмеження враховані, тобто $\operatorname{rang} X_g = q$, $\operatorname{rang} X = p$.

4. Немає обмежень на α_g та α.

Для розв'язку задачі коваріаційного аналізу, враховуючи структуру моделі використовують покроковий метод найменших квадратів:

7.1 Двокроковий метод найменших квадратів

1. Нехай $\alpha = \theta$, тоді маємо задачу дисперсійного аналізу: знаходимо оцінку $\widehat{\alpha}_g$ методом найменших квадратів: $\widehat{\alpha}_g(\theta) = (X_g^T X_g)^{-1} X_g^T y$ також залишкову суму квадратів $\text{СК}_e(\theta) = y^T Q y$, де $Q = E - X_g(X_g^T X_g)^{-1} X_g^T$.

2. Заміняємо в y на $y - X\alpha$: $\text{СК}_e(\alpha) = (y - X\alpha)^T Q (y - X\alpha)$. Оцінка
$$\widehat{\alpha}_g = (X^T Q X)^{-1} X^T Q y, \quad \text{СК}_q(\widehat{\alpha}) = (y - X\widehat{\alpha})^T Q (y - X\widehat{\alpha})$$

3. Замінюємо y на $y - X\widehat{\alpha}$:
$$\widehat{\alpha}_g(\widehat{\alpha}) = (X_g^T)^{-1} X_g^T (y - X\widehat{\alpha}).$$

8 Аналіз часових рядів

Нехай η_t – випадковий процес з дискретним часом, де $t \in I = \{1, 2, 3, \ldots\}$, або $= \{\ldots, -2, -1, 0, 1, 2 \ldots\}$, y_t – спостереження над цим процесом $t \in I$, тобто конкретна реалізація.

Така послідовність спостережень в часі називається *часовим рядом* (інколи η_t називається часовим рядом).

В якості математичної моделі для часового ряду будемо використовувати наступне:
$$\eta_t = f(t) + \xi_t, \quad t \in I,$$
де $f(t)$ – детермінована систематична складова часового ряду ξ_t – стохастична складова часового ряду $f(t)$ ще називається *трендом*.

Якщо тренд можна апроксимувати на деякому класі параметричних функцій, то використовуються параметричні методи.

Якщо є можливість апроксимувати поліномом високого степеня у околі кожної точки, то можна використати метод ковзного середнього.

8.1 Часові ряди з поліноміальними трендами

Нехай відомо, що для часового ряду $\eta_t = f(t) + \xi_t$, $t \in I$ функцію тренда можна апроксимувати поліномом високого степеня.

Відносно ξ_t справедливо

1. $\forall t : \exists M\xi_t$;
2. $\forall t \neq s : \exists M\xi_t\xi_s$.

Задача полягає у:

1. знаходженні оцінок параметрів апроксимуючого полінома;
2. перевірка коефіцієнтів апроксимуючого полінома на значимість;
3. визначити порядок апроксимуючого полінома.