АНАЛІЗ ДАНИХ

@Лекція 1

Етапи аналізу даних:

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

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

3-5 – Побудова математичних моделей зв'язків.

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 => => градація зникла.

Ще існує поділ на дискретні і неперервні змінні.

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

$5 \le s \le 30$

$s \approx 1 + [\log_2 n]$

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

Поставимо у відповідність $(C_i^1, C_i^2] \to x_i^0$ (як правило середня точка середня точка), ν_i - частота попадання

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

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

$\hat{F}_\xi^{(n)}(x)$ - емпіричний розподіл, n – об'єм вибірки.

$p_s(x), \hat{p}_s^{(n)}(x)$ - неперервний випадок (щільність).

Для дискретного випадку $\{y_i, p_i\}_{i=1}^m \to \{y_i, \hat{p}_i\}_{i=1}^m$ - полігон частот.

Розвідувальний аналіз.

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

I. 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 F – *сімейство розподілів типу зсуву масштабу*, якщо існує функція розподілу $\exists F_0(\cdot) \in F : \forall F(\cdot) \in F \quad \exists a, b \in R^1 (b > 0): \quad F(x) = F_0\left(\dfrac{x-a}{b}\right)$

a - параметр зсуву

b - параметр масштабу

F_0 - базова функція для сімейства розподілів F

Приклад.

Нормальний розподіл $F(x)$: $N(m,\sigma^2)$

Базова функція Φ(x) з розподілу $N(0,1)$

$$a = m \qquad F(x) = \Phi\left(\frac{x-m}{\sigma}\right)$$
$$b = \sigma$$

Сімейство нормальних розподілів є сімейством зсуву масштабу.

Приклад.

Експоненціальний розподіл з параметром λ

$F(x)$ з $\lambda > 0$

$$a = 0$$
$$b = \frac{1}{\lambda} \qquad F(x) = \Phi_1(\lambda x)$$

Φ_1 - базова функція експоненціального розподілу з параметром $\lambda = 1$

1.Пробіт-графік.

Будується наступним чином.:

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

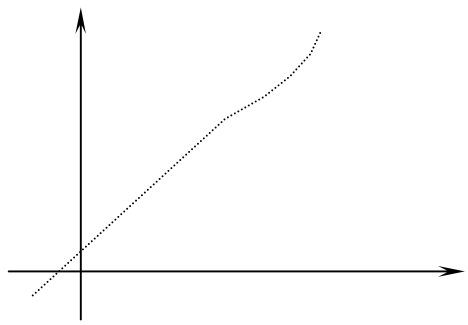
Обчислимо емпіричну функцію розподілу $\{x_i\}_{i=1}^n \to \hat{F}(x)$ (Сімейство розподілів F з базовою функцією F_0)

Пробіт-графік – графік функції $y = F_0^{-1}(\hat{F}(x))$

Використовується для:

1) Перевірки гіпотези $H_0 : F_\xi(\cdot) \in F$

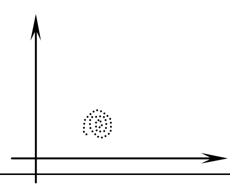
У випадку, коли справедлива гіпотеза $H_0 : F_\xi(x) \in F$ пробіт-графік повинен уявляти собою майже пряму.



Пояснення: маємо:

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

2) Виявлення наявності аномальних спостережень у вибірці.



@Лекція 2

2. Імовірнісний графік

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

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 $H_0 : F_\xi(\circ) \in f$, то отримаємо майже пряму. Якщо маємо точки, що лежать осторонь, то перевіряємо їх на аномальність.
2.Виявляємо наявність аномальних спостережень

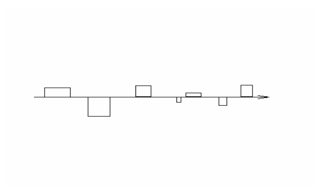
3.Висячі гістобари

Використовується для перевірки нормальності вибірки. Нехай по вибірці $\xi : x_1,\dots,x_n$ 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 $N(\bar{x}(n), s^2(n))$.

Спочатку будуємо графік щільності з вибірки $\xi : x_1,\dots,x_n$
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 *Ox* – гіпотеза про нормальність вибірки приймається.

4.Підвішена коренеграма

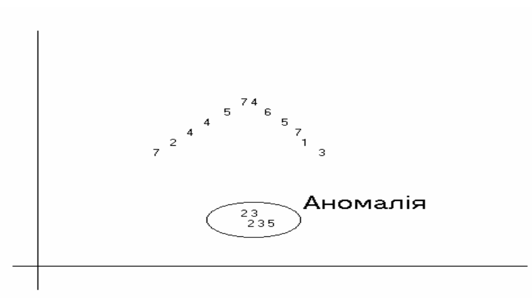


Для кожного інтервалу групування даних визначають ν_e - емпірична частота потрапляння в інтервал, а також теоретичне значення частоти ν_T згідно гіпотези про найбільш узгоджений нормальний розподіл. Потім на графіку відкладають такі різниці: $\sqrt{\nu_e} - \sqrt{\nu_T}$. І якщо ці значення не значно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

Спостереження за двома змінними

Використовуються
1) Діаграма розсіювання.
2) Таблиця спряженості.

1.Діаграми розсіювання.



Маємо дві вибірки $\xi : x_1, x_2,\dots,x_n$ та $\zeta : y_1, y_2,\dots,y_n$. Використовуються для з'ясування класу залежності між парою кількісних змінних, а також для з'ясування наявності аномальних спостережень у вибірці.

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}			$n_{\bullet 1}$
2	n_{21}				
…		…	n_{ij}		
r_2			…		
$\sum$	$n_{\bullet 1}$	$n_{\bullet 1}$	$n_{\bullet 1}$		$n_{\bullet 1}$ n

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

Попередня обробка

До попередньої обробки відносять:

- розвідувальний аналіз;
- обчислення основних характеристик спостережуваних величин;
- видалення аномалій;
- перевірка основних гіпотез;
- перевірка на стохастичність вибірки.

Квантиль 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 $\hat{u}_q(F)$ 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.

<u>Приклади квантилей:</u>

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 $p = 0.25$ (Половина інтерквантильної широти $p = 0.125$,називається імовірнісним відхиленням.)

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 $\xi : P\{\xi \le 0\} = 0$. $\hat{G}_\xi(n) = \sqrt[n]{\prod_{i=1}^{n} x_i}$.
3. Середнє гармонічне $H = M^{-1}\left(\frac{1}{\xi}\right)$ для $\xi : P\{\xi \le 0\} = 0$. $\hat{H}_\xi = \left(\frac{1}{n}\sum_{i=1}^{n}\frac{1}{x_i}\right)^{-1}$
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 $x_{med} = u_{0.5}$

@Лекція 3

Характеристики розсіювання значень

Нехай маємо вибірку об'єму *n* спостережень $x_1, x_2, \dots, x_n$ над випадковою величиною ξ.

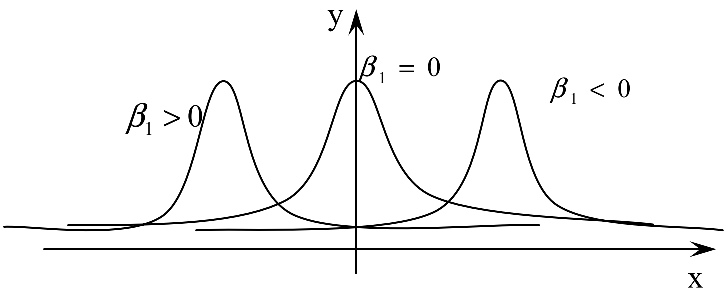
1. Дисперсія $D\xi = M(\xi - M\xi)^2$. Вибіркове значення
$$S^2(n) = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}(n)\right)^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - n\bar{x}^2(n)\right].$$
2. Стандартне (середньоквадратичне) відхилення $\sqrt{D\xi}$. Вибіркове значення $S(n)$.
3. Коефіцієнт варіацій $V_\xi = \frac{\sqrt{D\xi}}{M\xi}$ 100%, $M\xi \ne 0$. Вибіркове значення $\hat{V}_\xi(n) = \frac{S(n)}{\bar{x}(n)}$ 100%.
4. Стохастичне розсіювання (імовірнісне відхилення) – це половина інтерквартильної широти: $\frac{U_{0,75} - U_{0,25}}{2}$. Вибіркове значення $\frac{\hat{U}_{0,75} - \hat{U}_{0,25}}{2}$.
5. Розмах (широта) вибірки: $x_{\max} - x_{\min}$, де $x_{\max}, x_{\min}$ –найбільше та найменше значення у вибірці.
6. Інтервал концентрації $\left(M\xi - 3\sqrt{D\xi},\quad M\xi + 3\sqrt{D\xi}\right)$.

Вибіркове значення $\left(\bar{x}(n)-3S(n),\ \bar{x}(n)+3S(n)\right)$.

Характеристики скошеності та гостроверхості розподілу

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

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



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

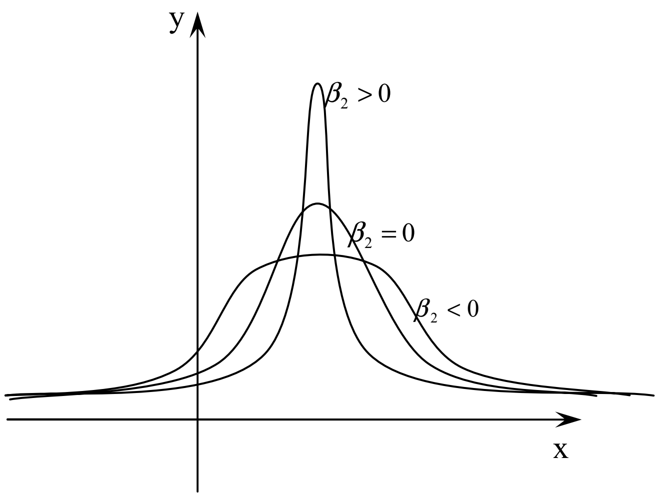
Вибіркове значення

$$\hat{\beta}_1(n) = \frac{\frac{1}{n}\sum_{k=1}^{n}\left(x_k - \bar{x}(n)\right)^3}{S^3(n)}.$$

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):



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

Вибіркове значення

$$\hat{\beta}_2(n) = \frac{\frac{1}{n}\sum_{k=1}^{n}\left(x_k - \bar{x}(n)\right)^4}{S^4(n)} - 3.$$

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.

Характеристики векторних величин

Аналіз q- вимірних векторних величин, отримано n спостережень над вектором $\bar{\xi}$

$x_1, x_2, \ldots, x_n, \quad x_i \in R^q, i = \overline{1,n}$.

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.

Характеристики розсіювання значень

1. Коваріаційна матриця $\Sigma = M(\xi - M\xi)(\xi - M\xi)^T$. Вибіркове значення $\hat{\Sigma}(n) = \frac{1}{n-1}\sum_{k=1}^{n}\left(x_k - \bar{x}(n)\right)\left(x_k - \bar{x}(n)\right)^T$.
2. Узагальнена дисперсія – визначник коваріаційної матриці: $\det\Sigma$. Вибіркове значення $\det\hat{\Sigma}$.
3. Слід коваріаційної матриці $tr\Sigma$. Вибіркове значення $tr\left(\hat{\Sigma}(n)\right)$.

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 $x_1, x_2, \ldots, x_n$ – 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 $\hat{x}_{med}$. Потім під кожним членом вибірки ставимо відповідно $\begin{cases} +, & x_i > \hat{x}_{med} \\ \ , & x_i = \hat{x}_{med} \\ -, & x_i < \hat{x}_{med} \end{cases}$ 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 $\nu(n)$, довжину найдовшої серії $\tau(n)$. Запишемо область прийняття нашої гіпотези: $\begin{cases} \nu(n) > \nu_\beta(n), \\ \tau(n) < \tau_{1-\beta}(n), \end{cases}$, де $\nu_\beta(n), \tau_\beta(n)$ – квантилі рівня β статистик $\nu(n)$, $\tau(n)$ 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 $x_1', x_2', \ldots, x_{n'}'$. 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 $\nu(n)$, довжину найдовшої серії $\tau(n)$. Запишемо область прийняття нашої гіпотези: $\begin{cases}\nu(n)>\nu_\beta(n),\\ \tau(n)<\tau_{1-\beta}(n),\end{cases}$, де $\nu_\beta(n),\tau_\beta(n)$ – квантилі рівня β статистик $\nu(n),\ \tau(n)$ відповідно. Гіпотеза приймається тільки у випадку справедливості обох нерівностей, причому при фіксованому значенні β рівень значимості α лежить у межах $\beta<\alpha<2\beta-\beta^2$.

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

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 $n\le 60$ визначається з таблиць, а протилежному випадку потрібно скористатися формулою $$\gamma_\alpha(n)=1+\frac{u_\alpha}{\sqrt{n+0.5\left(1+u_\alpha^2\right)}}.$$
u_α – квантиль рівня α нормального розподілу з параметрами 0 та 1.

@Лекція 4

Видалення аномальних спостережень

I. Випадок скалярних спостережень

1.Метод Грабса

Переіряємо гіпотезу H_0 - найбільш підозрюваний на аномальний вимір є аномальним з рівнем значимості $\alpha>0$ (1).

Спостерігається скалярна величина $\xi: x_1,\dots,x_n$ (2)

1.По вибірці визначаються характеристики:

середнє вибіркове $\bar{x}(n)$,

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

2.Будується послідовність $z_i=\left|x_i-\bar{x}(n)\right|$ - абсолютне значення відхилень. По цій послідовності будується варіаційний ряд.:

3. $z_{(1)},\dots,z_{(n)}$ - (3).Нехай є s-й член варіаційного ряду, то $z_{(s)}=\left|x_{i(s)}-\bar{x}(n)\right|, s=\overline{1,n}$.

Перевіряємо на аномальність $x_{i(n)}$:

4.Далі розглядається така статистика $$T(n)=\frac{x_{i(n)}-\bar{x}(n)}{S(n)}$$, в якості H_0 отримаємо:

$\left|T(n)\right|<T_{\frac{\alpha}{2}}(n)$ (4), де $T_{\frac{\alpha}{2}}(n)$ це $100\frac{\alpha}{2}\%$ точка статистики $T(n)$. Якщо (4) несправедливе, то $x_{i(n)}$ - аномальне і об’єм вибірки стає на 1 менше.

Далі повторюємо алгоритм, поки не буде виконуватись (4).

2.Метод Томпсона Модифікація методу Грабса

Приймаємо гіпотезу H_0 - найбільш підозрюваний на аномальний вимір є аномальним з рівнем значимості $\alpha > 0$ (1).

Спостерігаються скалярні величина $\xi : x_1, ..., x_n$ (2)

1.По вибірці визначаються характеристики:

середнє вибіркове $\bar{x}(n)$,

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

2.Будується послідовність $z_i = |x_i - \bar{x}(n)|$ - абсолютне значення відхилень. По цій послідовності будується варіаційний ряд.

3. $z_{(1)}, ..., z_{(n)}$ - (3).Якщо є s-й член варіаційного ряду, то $z_{(s)} = |x_{i(s)} - \bar{x}(n)|, s = \overline{1,n}$.

Перевіряємо на аномальність $x_{i(n)}$

4.Далі розглядається така статистика $T(n) = \frac{x_{i(n)} - \bar{x}(n)}{S(n)}$, в якості H_0 отримаємо:

$|T(n)| < T_{\frac{\alpha}{2}}(n)$ (4), де $T_{\frac{\alpha}{2}}(n)$ це $100\frac{\alpha}{2}\%$ точка статистики $T(n)$.

5.Візьмемо статистику $t_{(n-2)} = \frac{\sqrt{n-2}\,T(n)}{\sqrt{n-1-T^2(n)}}$ ця статистика має асимптотичний розподіл, t-розподіл Стьюдента з параметром $n-2$

Якщо нерівність несправедлива, то $x_{i(n)}$ видаляємо. Далі повторюємо алгоритм до тих пір поки нерівність (4) не стане вірною.

3.Метод Тітьєна-Мура - Дозволяє з вибірки викидати декілька вимірів

H_0 - найбільш підозрювані виміри є виміри, вказані в послідовності (*), не є аномальними,

$\alpha > 0$ - рівень значимості на базі вибірки (2)

1. Визначимо $\bar{x}(n)$.

2. Будуємо послідовність z_i: $z_i = |x_i - \bar{x}(n)|, i = \overline{1,n}$

На базі послідовності будуємо варіаційний ряд.: $z_{(1)}, ..., z_{(n-k)}, z_{(n-k+1)}, ..., z_{(n)}$ (5)

Найбільш підозрювані виміри це ті виміри, які фігурують в останніх k членах варіаційного ряду. $x_{i(n-k+1)}, ..., x_{i(n)}$ (*)

3. Розглянемо наступну статистику $E(n,k) = \frac{\sum_{i=1}^{n-k} \left(z_{(i)} - \bar{z}_{(n-k)}\right)^2}{\sum_{i=1}^{n} \left(z_{(i)} - \bar{z}_{(n)}\right)^2}$ де $z(m) = \frac{1}{m}\sum_{i=1}^{m} z_{(i)}$.

4. Таким чином область відхилення (критична область)- область малих значень, тобто область прийняття гіпотези $E(n,k) > E_\alpha(n,k)$, де $E_\alpha(n,k)$ - квантиль рівня α статистики $E(n,k)$.

Цей критерій чутливий до:

- нормальності вибору,
- питання вибору k залишається відкритим,
- не існує алгоритму, що дозволив би вірно вибрати довжину вибірки.

4.Графічні методи: Розвідувальний аналіз.

II. Векторний випадок

Нехай спостерігається $\vec{\xi} : \vec{x}_1,...,\vec{x}_n \in R^q$

1.Критерій на базі F-статистики.

1.Спочатку підраховують $\bar{x}_i = \frac{1}{n-1}\sum_{i\neq j} x_j, \forall i$, далі підраховуємо значення коваріаційної матриці по всій вибірці, крім i-го виміру.

$$\Sigma_i = \frac{1}{n-2}\sum_{j\neq i}(x_j - \bar{x}_i)(x_j - \bar{x}_i)^{\mathrm{T}}$$

2.Підраховуємо $D_i^2 = (x_i - \bar{x}_i)^{\mathrm{T}} \Sigma_i^{-1}(x_i - \bar{x}_i), \forall i$ - відстань Махаланобіса.

$\|x\|_Q^2 = x^T Qx$ зважена норма. Q - додатньовизначена матриця.

3.Підраховуються такі статистики: $F_i = \frac{(n-1)(n-1-q)}{n(n-2)q} D_i^2, \forall i$

4.Визначимо $i_* = \arg\max_i F_i$ F_{i_*} - індекс виміру для якого відстань Махаланобіса максимальна

5. Якщо x_{i_*} не є аномальним, тоді область прийняття гіпотези має вигляд: $F_{i_*} < F_\alpha(q, n-1-q)$ (*) де $F_\alpha(q, n-1-q)$, це $100\alpha\%$ точка F розподілу з параметрами $(q, n-1-q)$.

Якщо (*) виконується на деякому кроці, то останній з вибірки не видаляється і STOP.

2.Графічні методи :діаграма розсіювання

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.

I.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:

$f(x) = M(\eta / \xi = x)$, $g(x) = D(\eta / \xi = x)$ - *умовна дисперсія*. $D\eta = Df(\xi) + Mg(\xi)$

індексом кореляції для змінних $\eta \, ma \, \xi$ називається $I_{\eta\xi} = \sqrt{\frac{Df(\xi)}{D_\eta}} = \sqrt{1 - \frac{Mg(\xi)}{D_\eta}}$

Властивості

1. $0 \le I_{\eta\xi} \le 1$
2. якщо $I_{\eta\xi} = 0$, то зв'язку між $\eta \, ma \, \xi$ немає.

3. якщо $I_{\eta\xi}=1$, то є функціональний зв'язок між ними

Коефіцієнт детермінації $I_{\eta\xi}^2=\dfrac{Df(\xi)}{D_\eta}$ вказує яка частина варіації η визначаються варіацією функцій регресії в точці ξ

@Лекція 5

Коефіцієнт кореляції. Характеристика парного статистичного зв'язку.

Розглянемо нормальний випадок. Є дві величини ξ та η.

$\xi \sim N(m_\xi,\sigma_\xi^2)$, $x_1,\dots,x_n$

$\eta \sim N(m_\eta,\sigma_\eta^2)$, $y_1,\dots,y_n$

$$r_{\eta\xi}=\frac{M(\xi-M\xi)(\eta-M\eta)}{\sqrt{D\eta D\xi}}$$, вибіркове значення: $$\hat r_{\eta\xi}=\frac{\sum_{i=1}^{n}(x_i-\bar x(n))(y_i-\bar y(n))}{\sqrt{\sum_{i=1}^{n}(x_i-\bar x(n))^2\sum_{i=1}^{n}(y_i-\bar y(n))^2}}$$

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

Властивості.

1. $|r_{\eta\xi}|\le 1$.
2. якщо $r_{\eta\xi}=0\Rightarrow$ зв'язок між η і ξ відсутній.

   Якщо $r_{\eta\xi}=\pm1\Rightarrow$ зв'язок між η і ξ лінійний, причому формула зв'язку:

   $\eta=m_\eta+r_{\eta\xi}\sigma_\eta\dfrac{\xi-m_\xi}{\sigma_\xi}$.
3. Нехай $r_{\eta\xi}>0$. Якщо $\xi\uparrow$, то і $\eta\uparrow$.

   $r_{\eta\xi}<0$. При $\xi\uparrow,\eta\downarrow$.
4.

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: $t(n-2)=\dfrac{\sqrt{n-2}\,\hat r_{\eta\xi}}{\sqrt{1-r_{\eta\xi}^2}}$.

Ця статистика має асимптотичний t-розподіл Стьюдента з степенями $(n-2)$ свободи. Тоді логічно вважати, що H_0 гіпотеза несправедлива, коли статистика приймає екстремальні значення.

$|t(n-2)|<t_{\frac{\alpha}{2}}(n-2)\Rightarrow$ область прийняття гіпотези H_0, де $t_\alpha(\nu)-100*\alpha\%$ - точки t-розподілу Стьюдента з ν степенями свободи.

Характеристика парного статистичного зв'язку в загальному випадку.

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

- випадок групованих даних;

- випадок не згрупованих даних.

1. Спостереження над залежною змінною η: $y_{11},\dots,y_{1m_1};y_{21},\dots,y_{2m_2};\dots;y_{s1},\dots,y_{sm_s}$
s- інтервалів групування, в і –му інтервалі не більш ніж m_i спостережень.
$\bar{y}_i=\frac{1}{m_i}\sum_{j=1}^{m_i}y_{ij}$ спостережень по m_i групі $n=\sum_{i=1}^{s}m_i$, $\bar{y}=\frac{1}{n}\sum_{i=1}^{s}m_i\bar{y}_i$
S_y^2- вибіркове значення дисперсії η.
$$S_y^2=\frac{1}{n}\sum_{i=1}^{S}\sum_{j=1}^{m_i}\left(y_{ij}-\bar{y}\right)^2$$
$$S_{y(x)}^2=\frac{1}{n}\sum_{i=1}^{S}m_i\left(\bar{y}_i-\bar{y}\right)^2$$
Запишемо оцінку для індексу кореляції (кореляційне відношення):
$\hat{p}_{\eta\xi}=\sqrt{\frac{S_{y(x)}{}^2}{S_y^2}}$.Властивості такі ж, як і в індексу кореляції. З'ясувалося, що
$F=\frac{\hat{p}_{\eta\xi}^2}{1-\hat{p}_{\eta\xi}^2}*\frac{n-s}{s-1}$ має асимптотичний розподіл, який тотожньо рівний $F\left(s-1,n-s\right)$.
$H_0:I_{\eta\xi}=0,\alpha>0$. Припускаємо, що спостереження нормальні.
Область прийняття гіпотези: $F<F_\alpha\left(S-1,n-S\right)$, де $F_\alpha-100*\alpha\%$ точка з параметрами $s-1,n-s$.

2. Функцію регресії f апроксимують на деякому класі параметричних функцій з точністю до вектор – параметру θ. $f\left(x,\theta\right),\theta\in R^p$.
По спостереженням досліджуваних змінних:
$\xi:x_1,\dots,x_n$
$\eta:y_1,\dots,y_n$
Методом найменших квадратів визначаємо $\hat{\theta}$, далі отримуємо деяку апроксимацію функції регресії $f\left(x,\theta\right)$.
Апроксимація індексу кореляції даних у вигляді:
$$\hat{I}_{\eta\xi}=\sqrt{1-\frac{\frac{1}{n-p}\sum_{i=1}^{n}\left(y_i-f\left(x_i,\hat{\theta}\right)\right)^2}{\frac{1}{n-1}\sum_{i=1}^{n}\left(y_i-\bar{y}(n)\right)^2}}$$

<u>Приклад</u>: $f\left(x,\theta\right)=\sum_{i=1}^{\mathrm{N}}\theta_i f_i\left(x\right)$

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:

$$R = \begin{pmatrix} 1 & r_{01} & \dots & r_{0q} \\ r_0 & 1 & \dots & r_{1q} \\ \dots & \dots & \dots & \dots \\ r_{q0} & r_{q1} & \dots & 1 \end{pmatrix}$$, r_{ij} - звичайний коефіцієнт кореляції.

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

$$r_{ij}^* = -\frac{\hat{R}_{ij}}{\sqrt{\hat{R}_{ii}\hat{R}_{jj}}}, \quad \hat{R} = \begin{pmatrix} 1 & \hat{r}_{01} & \dots & \hat{r}_{0q} \\ \hat{r}_0 & 1 & \dots & \hat{r}_{1q} \\ \dots & \dots & \dots & \dots \\ \hat{r}_{q0} & \hat{r}_{q1} & \dots & 1 \end{pmatrix}.$$

При $r_{ij}^* = 0$ зв'язку не існує.

При $r_{ij}^* = \pm 1$ зв'язок функціональний.

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

$$t(n-m-2) = \frac{\sqrt{n-m-2} * \hat{r}_{ij}^*}{\sqrt{1-\left(\hat{r}_{ij}^*\right)^2}}$$, де m кількість третіх змінних зафіксованих на певному рівні.

Вона має t - розподіл Стьюдента з $n-m-2$ степенями свободи.
Критична область – область великих і малих значень.
Область прийняття має вигляд:
$\left|t(n-m-2)\right| < t_{\frac{\alpha}{2}}(n-m-2)$, де $t_{\frac{\alpha}{2}}(n-m-2) - 100\%$ точка $t-$ розподілу Стьюдента з $n-m-2$ степенями свободи.

Множинний коефіцієнт кореляції.

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

$$R_{\eta\xi} = \sqrt{\frac{D(f(\xi))}{D\eta}} = \sqrt{1 - \frac{Mq(\xi)}{D\eta}}$$, де $f(x) = M(\eta / \vec{\xi} = \vec{x})$, $g(x) = D(\eta / \vec{\xi} = \vec{x})$.

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

$\vec{\xi} : \vec{x}_1, \dots, \vec{x}_n$, $\eta : y_1, \dots, y_n$.

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

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

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: $F = \frac{\hat{R}^2_{\eta\xi}}{1-\hat{R}^2_{\eta\xi}}\frac{n-p}{p-1}$

Вона має асимптотичний розподіл, який співпадає з F– розоділом з параметрами (p-1,n-p).

Тоді область прийняття – це область невеликих значень: $F < F_\alpha(p-1, n-p)$

Кореляційний аналіз порядкових змінних.

Нехай $\exists \eta -$ залежна порядкова змінна і $\vec{\xi} = \begin{pmatrix} \xi_1 \\ \dots \\ \xi_q \end{pmatrix}$.

$\eta, \xi_1, \dots, \xi_q$

$x^{(0)}, x^{(1)}, \dots, x^{(q)}$.

Нехай відбуваються спостереження над $x^{(i)}$

. В результаті отримаємо вектор:

$x^{(i)}_{\cdot} = \begin{pmatrix} x_1^{(i)} \\ \dots \\ x_n^{(i)} \end{pmatrix}$ - ранжировка, де $x_k^{(i)}$ - ранг к-го об'єкту по і –й змінній, який вказує степінь прояву і –ї властивості для к-го об'єкту. Сама ранжировка – перестановка чисел від 1 до n.

@Лекція 6

Якщо всі прояви об'єктів різні, то маємо $x^{(0)}, x^{(1)}, \dots, x^{(q)}$ - спостереження, $x_\bullet^{(0)}, x_\bullet^{(1)}, \dots, x_\bullet^{(q)}$ - 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.

Змінні / № об'єктів	$x^{(1)}$	$x^{(2)}$	$\cdots$	$x^{(q)}$
1	$x_1^{(1)}$	$x_1^{(2)}$	$\cdots$	$x_1^{(q)}$
2	$x_1^{(1)}$	$x_2^{(2)}$	$\cdots$	$x_2^{(q)}$

…	…	…	…	…
n	$x_n^{(1)}$	$x_n^{(2)}$	…	$x_n^{(q)}$

Характеристики парного статистичного зв'язку.

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

В якості характеристики парного зв'язку між змінними $x^{(i)}$ та $x^{(j)}$ можемо використати *коефіцієнт Спірмана*, який визначається таким чином: $\hat{\tau}_{ij}^{(s)} = 1 - \dfrac{\left\|x_\bullet^{(i)} - x_\bullet^{(j)}\right\|^2}{(n^3-n)/6}$, де $\|\bullet\|$ - норма Евкліда.

Властивості рангу коефіцієнта Спірмана:

1. $-1 \le \hat{\tau}_{ij}^{(s)} \le 1$;
2. якщо $\hat{\tau}_{ij}^{(s)} = 0$, тоді зв'язок відсутній;
3. якщо $\hat{\tau}_{ij}^{(s)} = 1$, то ранжировки по змінним співпадають, $x_\bullet^{(i)} = x_\bullet^{(j)}$;

якщо $\hat{\tau}_{ij}^{(s)} = -1$, тоді ранжировки протилежні, тобто $\forall\, l : x_l^{(i)} = n - x_l^{(j)} + 1$.

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

$$\hat{\hat{\tau}}_{ij}^{(s)} = \frac{\frac{1}{6}(n^3-n) - \left\|x_\bullet^{(i)} - x_\bullet^{(j)}\right\|^2 - \mathrm{T}^{(i)} - \mathrm{T}^{(j)}}{\sqrt{\left(\frac{1}{6}(n^3-n) - 2\mathrm{T}^{(i)}\right)\left(\frac{1}{6}(n^3-n) - 2\mathrm{T}^{(j)}\right)}}$$, де $\mathrm{T}^{(i)} = \dfrac{1}{12}\sum_{g=1}^{m^{(i)}}\left((n_g^{(i)})^3 - n_g^{(i)}\right)$ - корегуючий коефіцієнт.

$m^{(i)}$ 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 $H_0 : \tau_{ij}^{(s)} = 0, \alpha > 0$.

Якщо об'єм вибірки невеликий, то перевіряємо по таблиці, при $n = 4 \div 10$.

Якщо ж $n > 10$, то розглядаємо статистику $\dfrac{\sqrt{n-2}\,\hat{\tau}_{ij}^{(s)}}{\sqrt{1-(\hat{\tau}_{ij}^{(s)})^2}}$, що має t - розподіл Стьюдента з $(n-2)$ степенями свободи.

Область прийняття гіпотези: $\left|\dfrac{\sqrt{n-2}\,\hat{\tau}_{ij}^{(s)}}{\sqrt{1-(\hat{\tau}_{ij}^{(s)})^2}}\right| < t_{\alpha/2}(n-2)$.

- Розглянемо іншу характеристику: коефіцієнт Кендала

Ранговим коефіцієнтом Кендала для змінних $x^{(i)}$ та $x^{(j)}$ називається величина $\hat{\tau}_{ij}^{(k)} = \dfrac{4\nu(x_\bullet^{(i)}, x_\bullet^{(j)})}{n(n-1)}$, де ν - кількість перестановок сусідніх елементів у ранжировці $x_0^{(i)}$, яка приводить її до ражировки $x_0^{(j)}$.

Властивості:

1. $-1 \le \hat{\tau}_{ij}^{(k)} \le 1$;
2. якщо $\hat{\tau}_{ij}^{(k)} = 0$, тоді зв'язок відсутній;
3. якщо $\hat{\tau}_{ij}^{(k)} = 1$, то ранжировки по змінним співпадають, $x_0^{(i)} = x_0^{(j)}$;

якщо $\hat{\tau}_{ij}^{(k)} = -1$, тоді ранжировки протилежні, тобто $\forall\, l : x_l^{(i)} = n - x_l^{(j)} + 1$.

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

$$\hat{\hat{\tau}}_{ij}^{(k)} = \frac{\hat{\tau}_{ij}^{(k)} - \dfrac{u^{(i)} - u^{(j)}}{n(n-1)}}{\sqrt{\left(1 - \dfrac{U^{(i)}}{n(n-1)}\right)\left(1 - \dfrac{U^{(j)}}{n(n-1)}\right)}}, \text{ де } U^{(i)} = \sum_{g=1}^{m^{(i)}} n_g^{(i)}(n_g^{(i)} - 1)$$

. $m^{(i)}$ 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.

Як і в коефіцієнті Спірмана, якщо $\hat{\tau}_{ij}^{(k)}$ приймають протилежне значення, то перевіряємо його на значимість

. Перевіряємо гіпотезу $H_0 : \hat{\tau}_{ij}^{(k)} = 0, \alpha > 0$. Якщо $n = 4 \div 10$, то перевіряємо по таблиці. Якщо $n > 10$:використовуємо $\left|\hat{\tau}_{ij}^{(k)}\right| \le U_{\alpha/2}\sqrt{\dfrac{2(2n+5)}{9n(n-1)}}$.

Зауваження. При великих n існує простий зв'язок : $\hat{\tau}_{ij}^{(s)} = \dfrac{3}{2}\hat{\tau}_{ij}^{(k)}$.

Характеристика множинних рангових статистичних зв'язків.

Нехай аналізується m змінних $\zeta = (x^{(k_1)}, x^{(k_2)}, \ldots, x^{(k_m)})^{\mathrm{T}}$.
В якості характеристики використовується коефіцієнт конкордації

Коефіцієнтом конкордації для змінної $\zeta = (x^{(k_1)}, x^{(k_2)}, \ldots, x^{(k_m)})^{\mathrm{T}}$ називають величину

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

Властивості:

1. $-1 \le \hat{w}_\zeta \le 1$;
2. якщо $\hat{w}_\zeta = 1$, то ранжировки по змінним співпадають: $x_0^{(k_1)} = x_0^{(k_2)} = \ldots = x_0^{(k_m)}$;
3. якщо $\hat{w}_\zeta = 0$, тоді відсутній зв'язок між ранжировками.

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

$$\hat{\hat{w}}_\zeta = \frac{\sum_{i=1}^{n}\left[\left(\sum_{j=1}^{m} x_i^{(k_i)}\right) - \dfrac{m(n+1)}{2}\right]^2}{\dfrac{m^2(n^3+n)}{2} - m\sum_{j=1}^{m} \mathrm{T}^{(k_i)}}, \text{ де } \mathrm{T}^{(k_j)} = \frac{1}{12}\sum_{g=1}^{m_j}\left((n_g^{(k_j)})^s - n_g^{(k_j)}\right).$$

Якщо $\hat{w}_\zeta$ приймають проміжне значення, то робимо перевірку на значимість $H_0 : \hat{w}_\zeta = 0, \alpha > 0$. Коли $n = \overline{3{,}7}$, $m = \overline{2{,}20}$, то за таблицею. Якщо $n > 7, m > 20$, то розглядаємо статистику $\hat{w}_\zeta$: $\hat{w}_\zeta < \dfrac{\chi^2_\alpha (n-1)}{m(n-1)}$, має χ^2 - розподіл з $(n-1)$ степенем свободи.

Кореляційний аналіз номінальних змінних.

Нехай η, ξ - змінні, які мають відповідні градації. r_1, r_2
Результат спостережень заноситься в таблицю спряженості. *(див. "Розвідувальний аналіз"), потім переходимо до характеристики парного статистичного зв'язку для номінальних змінних.

$\eta \quad \backslash \quad \xi$	$1 \ldots r_2$	$\sum$
1 $\vdots$ r_1	$\ldots$ n_{ij} $\ldots$	n_1 $\vdots$ n_{r_1}
$\sum$	$n_1 \ldots n_{r_2}$	n

Вводимо статистику яка називається квадратичне спряження і позначається

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

Коефіцієнти:

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

$\dfrac{n_{ij}}{n} \cong \dfrac{n_{i\bullet}}{n} \dfrac{n_{\bullet j}}{n}$ - якщо вони незалежні, то вони рівні, (або майже рівні).

@Лекція 7

Ентропією для змінної ξ називають величину $H_\xi = -\sum_i p(x_i) \ln p(x_i) = -M \ln p(\xi)$

Ймовірність, з якою приймається пара значень (x_i, y_j) дорівнює $p(x_i, y_j)$.

Ентропією для пари (ξ,η) називається величина $H_{\eta\xi} = -\sum_{i,j} p(x_i, y_j)\ln p(x_i, y_j)$.

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

Властивості інформаційної міри зв'язку

$I_{\eta\xi} \ge 0$

$I_{\eta\xi} = 0 \Rightarrow$ зв'язок між ξ та η відсутній.

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

$$\hat{H}_\eta = -\sum_i \frac{n_{i\cdot}}{n}\cdot\ln\frac{n_{i\cdot}}{n},$$

$$\hat{H}_\xi = -\sum_j \frac{n_{\cdot j}}{n}\cdot\ln\frac{n_{\cdot j}}{n}; \Rightarrow \quad \hat{H}_{\eta\xi} = -\sum_{i,j} \frac{n_{ij}}{n}\cdot\ln\frac{n_{ij}}{n}$$

$$\hat{I}_{\eta\xi} = \frac{1}{n}\left[\sum_{i,j} n_{ij}\ln n_{ij} - \sum_i n_{i\cdot}\ln n_{i\cdot} - \sum_j n_{\cdot j}\ln n_{\cdot j} + n\ln n\right]$$

При перевірці характеристики на значимість можливі два випадки:

Перший випадок
якщо ми вибрали з 1 до 4, то перевірку на значимість роблять шляхом перевірки гіпотези $H_0 : \chi^2_{\eta\xi} = 0,\ \alpha > 0$,
з'ясувалося, що $\chi^2_{\eta\xi}$ має хі-квадрат розподіл $\chi^2((r_1-1)(r_2-1))$ з $(r_1-1)(r_2-1)$ степенями свободи.
Тоді критична область, область великих значень та область прийняття гіпотези:

$$\chi^2_{\eta\xi} < \chi^2((r_1-1)(r_2-1)).$$

Другий випадок(з використанням інформаційної міри зв'язку)

$H_0 : \hat{I}_{\eta\xi} = 0, \quad \alpha > 0$

$\hat{I}_{\eta\xi} = 2n\hat{I}_{\eta\xi} - n_0$,де n_0 – кількість нульових елементів у таблиці спряженості.
Виявилось, що така перетворена статистика:

$$\hat{\hat{I}}_{\eta\xi} < \chi^2((r_1-1)(r_2-1)).$$

Оскільки ця статистика невід'ємна, то область прийняття гіпотези матиме такий вигляд (тобто, $100\cdot\alpha$ процентна точка хі-квадрат розподілу з $(r_1-1)(r_2-1)$ степенями свободи).

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.

y_{ijk} – спостерігаю i-й сорт зернових та j-й тип міндобрив на k-му полі.

α_i – вплив на залежну змінну, β_j – вплив на якісну змінну.

$y_{ijk} = \mu + \alpha_i + \beta_j + c_{ij} + e_{ijk}$, де $\mu, \alpha_i, \beta_j, c_{ij}$ – невідомі параметри, e_{ijk} – помилка моделі,

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).

Перевірка лінійних гіпотез для регресійної моделі

Лінійну регресій ну модель в матричному вигляді запишемо так:

$y = X\alpha + e$, де y розмірності N, X розмірності $N \times p$, α має розмірність p, e розмірність N.

Нехай ранг X дорівнює p (тобто матриця має повний ранг по стовпчикам) (1)

А сама оцінка знаходилась з критерію:

$$Q(\alpha) = \|y - X\alpha\|^2 \quad (2)$$

І оцінка має вигляд:

$$\hat{\alpha} = \left(X^T X\right)^{-1} X^T y.$$

Розглянемо таку множину $L = \{\alpha : A\alpha = b,\ rang\ A = q\}$ (3),

де A розмірності $q \times p$ (тобто матриця має повний ранг по рядкам).

Розглянемо задачу оцінки α для (1) методом найменших квадратів при наявності лінійних обмежень (3).

$$\hat{\alpha}_L = \hat{\alpha} + \left(X^T X\right)^{-1} A^T \left[A\left(X^T X\right)^{-1} A^T\right]^{-1} \left(b - A\hat{\alpha}\right).$$

Теорема

При справедливості гіпотези $H_0 :\ A\alpha = b,\ rang\ A = q;\ \gamma > 0$ наступна статистика

$$F = \frac{[Q(\hat{\alpha}_L) - Q(\hat{\alpha})]\big/ q}{Q(\hat{\alpha})\big/(N-p)}$$

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 $100 \cdot \gamma$ % точка F-розподілу.

Зауваження 1. При справедливості гіпотези H_0 наступна величина

$$Q(\hat{\alpha}_L) - Q(\hat{\alpha}) = \|X\hat{\alpha}_L - X\hat{\alpha}\|^2.$$

Зауваження 2. Розглянемо наступну гіпотезу:

$H:\ \alpha_i = 0,\ \gamma > 0$ (перевіряємо, чи суттєво відхиляється від нуля відносний вплив *i*-ої градації), тоді статистика, побудована по умовам теореми F_i називається *частинною статистикою по і-й змінній,* а відповідний критерій, побудований на цій статистиці, для перевірки гіпотези H, називається *частинним F- критерієм по і-й змінній.*

Однофакторний дисперсійний аналіз

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

$$y_{ij} = \mu + \mu_i + e_{ij},\ i = \overline{1, I},\ j = \overline{1, J_i} \qquad (4)$$

Припустимо, що помилки моделі:

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

2)незалежні.

Запишемо модель (4) в матричному вигляді:

$$\begin{pmatrix} y_{11} \\ y_{12} \\ \dots \\ y_{1J_1} \\ y_{21} \\ y_{22} \\ \dots \\ y_{2J_2} \\ \dots \\ \dots \\ y_{I1} \\ y_{I2} \\ \dots \\ y_{IJ_I} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 & \cdots & 0 \\ 1 & 1 & 0 & & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & 1 & \vdots & & 0 \\ \vdots & 0 & 1 & \cdots & 0 \\ & 0 & 1 & & \vdots \\ \vdots & \vdots & \vdots & \cdots & \\ & 0 & 1 & & \vdots \\ \vdots & \vdots & 0 & & \\ 1 & & \vdots & \ddots & 0 \\ \vdots & \vdots & \vdots & & 1 \\ & & & & \vdots \\ \vdots & \vdots & \vdots & \cdots & \\ 1 & 0 & 0 & & 1 \end{pmatrix} \begin{pmatrix} \mu \\ \mu_1 \\ \mu_2 \\ \\ \mu_I \end{pmatrix} + \begin{pmatrix} e_{11} \\ e_{12} \\ \vdots \\ e_{1J_i} \\ e_{21} \\ e_{22} \\ \vdots \\ e_{2J_2} \\ \vdots \\ e_{I1} \\ e_{I2} \\ \vdots \\ e_{IJ_I} \end{pmatrix} \qquad (5)$$

Перепишемо у матричному вигляді

$$y = X\alpha + e \qquad (6),$$

де матриця X розмірності $N \times (I+1)$.

@Лекція 8

Будемо вважати, що помилки моделі незалежні і нормально розподілені. Позначимо через $N = \sum_{i=1}^{I} J_i$ - загальну кількість вимірів, тоді для нашої моделі маємо розмірності векторів $y \in R^N,\quad X \in R^{N\times(I+1)},\quad \alpha \in R^{I+1},\quad e \in R^N$.

Щоб оцінка існувала потрібно, щоб $rangX = I$, тобто

$$\exists w_i: \quad \sum_{i=1}^{I} w_i \mu_i = 0; \quad \sum_{i=1}^{I} w_i = 1 \quad \forall i: w_i > 0 \quad (7)$$

З (6) та (7) можемо отримати оцінку параметрів $\mu, \quad \mu_i$. Принциповим моментом є з'ясування суттєвості впливу однієї градації на іншу.

Запишемо це математично: $H: \quad \mu_1 = \mu_2 = ... = \mu_I = 0, \quad \gamma > 0$. (8)

Розв'яжемо задачу перевірки цієї гіпотези та паралельно знайдемо параметри $\mu, \quad \mu_i$.

Ця гіпотеза є лінійною, запишемо її у стандартному вигляді:

$$H: \quad \underbrace{\begin{pmatrix} 0 & 1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & 0 \end{pmatrix}}_{\text{матриця} \quad A} \underbrace{\begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_I \end{pmatrix}}_{\text{вектор} \quad \alpha} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \qquad \gamma > 0 .$$

Позначимо Θ - нульову матрицю, θ - нульовий вектор, тоді перепишемо нашу гіпотезу згідно цих позначень: $H: \quad A\alpha = \theta, \qquad \gamma > 0, \quad rangA = I - 1$. (9)

Фактично потрібно перевірити гіпотезу (9) для моделі (6) при наявності умов (7).

Згідно теореми область прийняття гіпотези матиме вигляд:

$$F = \frac{\left(Q(\hat{\alpha}_L) - Q(\hat{\alpha})\right) / (I-1)}{Q(\hat{\alpha}) / (N-I)} < F_\gamma (I-1, N-I)$$

,де $Q(\alpha) = \|y - X\alpha\|^2$ (*)

$$I = \{\alpha : A\alpha = \theta, \quad rangA = I - 1\}.$$

Зауваження. Якщо I-1 параметр є нульовими, то і I-й параметр теж нульовий, згідно (7).

Знайдемо $Q(\hat{\alpha})$. Відомо, що оцінка методом найменших квадратів є розв'язком

системи нормальних рівнянь $X^T X \hat{\alpha} = X^T y$. (10)

Перепишемо систему в матричному вигляді:

$$\begin{pmatrix} N & J_1 & J_2 & \cdots & J_{Ii} \\ J_1 & J_1 & 0 & \cdots & 0 \\ J_2 & 0 & J_2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ J_I & 0 & 0 & \cdots & J_I \end{pmatrix} \begin{pmatrix} \hat{\mu} \\ \hat{\mu}_1 \\ \hat{\mu}_2 \\ \vdots \\ \hat{\mu}_I \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{I} \sum_{j=1}^{J_i} y_{ij} \\ J_1 \bar{y}_1 \\ J_2 \bar{y}_2 \\ \vdots \\ J_I \bar{y}_I \end{pmatrix}, \quad \text{де} \quad \bar{y}_i = \frac{1}{J_i} \sum_{j=1}^{J_i} y_{ij} .$$

Розглянемо окреме рівняння системи, отримаємо

$$\forall i: \quad J_i \hat{\mu} + J_i \hat{\mu}_i = J_i \bar{y}_i \quad \Rightarrow \quad \bar{y}_i = \hat{\mu} + \hat{\mu}_i, \quad (11)$$

тобто оцінкою абсолютного впливу і-тої градації є $\bar{y}_i, i = \overline{1, I}$.

Звідси маємо $\hat{\mu}_i = \bar{y}_i - \hat{\mu}$. (12)

Скористаємось (7): $\sum_{i=1}^{I} w_i \hat{\mu}_i = 0$, згідно (12) отримаємо:

$$0 = \sum_{i=1}^{I} w_i (\bar{y}_i - \hat{\mu}_i) = \sum_{i=1}^{I} w_i \bar{y}_i - \sum_{i=1}^{I} w_i \hat{\mu}_i = \sum_{i=1}^{I} w_i \bar{y}_i - \hat{\mu},$$

$$\Rightarrow \quad \hat{\mu} = \sum_{i=1}^{I} w_i \bar{y}_i, \qquad (13)$$

$$\Rightarrow \quad \hat{\mu}_i = \bar{y}_i - \sum_{i=1}^{I} w_i \bar{y}_i. \qquad (14)$$

Згідно викладеного вище отримаємо $Q(\hat{\alpha}) \overset{(*)}{=} \sum_{i=1}^{I} \sum_{j=1}^{J_i} (y_{ij} - \bar{y}_i)^2$. (15)

L – лінійне обмеження, еквівалентне (8), тому $\hat{\alpha}_L$: $\tilde{X}^T \tilde{X} \hat{\alpha}_L = \tilde{X}^T y$,

де $\tilde{X} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$ - вектор стовпчик при обмеженні L (8), $N\hat{\mu}_L = \sum_{i=1}^{I} \sum_{j=1}^{J_i} y_{ij} \quad \Rightarrow$

$\hat{\mu}_L = \frac{1}{N} \sum_{i=1}^{I} \sum_{j=1}^{J_i} y_{ij} = \bar{y}$.- загальне середнє по всім вимірам (16)

Наслідок (Зауваження.1)

$$Q(\hat{\alpha}_L) - Q(\hat{\alpha}) = \left\| X\hat{\alpha} - X\hat{\alpha}_L \right\|^2 = \sum_{i=1}^{I} \sum_{j=1}^{J_i} (\bar{y}_i - \bar{y})^2 = \sum_{i=1}^{I} J_i (\bar{y}_i - \bar{y})^2 \qquad (17)$$

(Дивись структуру матриці X.)

Підставимо (15) та (17) в F і отримаємо

$$F = \frac{\left. \sum_{i=1}^{I} J_i (y_i - \bar{y})^2 \middle/ (I-1) \right.}{\left. \sum_{i=1}^{I} \sum_{j=1}^{J_i} (y_{ij} - \bar{y}_i)^2 \middle/ (N-I) \right.} < F_\gamma (I-1, N-I). \qquad (18)$$

Таким чином, для однофакторної моделі дисперсійного аналізу оцінки її параметрів визначаються згідно (13), (14), а область прийняття гіпотези (8) для об'єкту (4) при наявності лінійних обмежень (7) має вигляд (18).

Впевнимось в справедливості тотожності:

$$\underbrace{\sum_{i=1}^{I} \sum_{j=1}^{J_i} (y_{ij} - \bar{y})^2}_{S_n - \text{сума квадратів відхилення від середнього}} = \underbrace{\sum_{i=1}^{I} \sum_{j=1}^{J_i} (y_{ij} - \bar{y}_i)^2}_{S_E - \text{сума квадратів відхилення середнього градацій}} + \underbrace{\sum_{i=1}^{I} J_i (y_i - \bar{y})^2}_{S_A - \text{сума квадратыв выдхилень мыж градацыями}}.$$

Таблиця результатів однофакторного дисперсійного аналізу.

Джерело Варіацій	Сума Квадратів	КСС	ССК	F	γ_*
Між Градаціями	S_A	*I-1*	$\bar{S}_A = \frac{S_A}{(I-1)}$	$F = \frac{\bar{S}_A}{\bar{S}_E}$	γ_*
Всередині Градацій	S_E	*N-I*	$\bar{S}_E = \frac{S_E}{(N-I)}$		γ_*
..„	$S_n = S_A + S_E$	*N*-1			

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

Перевірка контрастів

Довірчі інтервали для контрастів.

Якщо гіпотеза (8) 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* –ї градації, це $\theta_i = \mu - \mu_i$, та $\hat{\theta} = \bar{y}_i$.

@Лекція 9

Означення.

Контрастами будемо називати статистики вище наведеного вигляду для коефіцієнтів яких справедлива умова. Нас цікавить перевірка гіпотези:

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

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.

Довірчі інтервали для контрастів.

1.Якщо c_i - наперед задані, то $\left|\sum_i c_i\bar{y} - \sum_i c_i\theta_i\right| < \bar{S}_e \sum_i \frac{c_i^2}{J_i} t_{\alpha/2}(N-I)$,

де $\left|\sum_i c_i\bar{y} - \sum_i c_i\theta_i\right|$ - усереднена залишкова сума квадратів.

2. S метод Шофе.

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

3.Т метод Кьюні.

Метод орієнтований на побудову довірчих інтервалів для контрастів статистики $\theta_i - \theta_j$, крім того припускаємо, що кількість вимірів при кожній градації однакова, тобто

$$\forall i: \quad J_i = J \quad \left|\bar{y}_i - \bar{y}_j - (\theta_i - \theta_j) < \bar{S}_e q_\alpha(I, N-I)\right|,$$

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

Зауваження.

Нехай є η_i, $i = \overline{1, I}$ - нормально розподілені величини з параметрами 0 та1;

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.

$$\begin{matrix} \textit{Якісні} \\ \textit{змінні} \end{matrix} \begin{cases} T_A - \textit{приймає} \quad \textit{знчення} \quad \textit{з} \quad I - \textit{тої} \quad \textit{градації}; \\ T_B - \textit{приймає} \quad \textit{значення} \quad \textit{з} \quad J - \textit{тої} \quad \textit{градації}. \end{cases}$$

Якщо фактор А приймає значення з i-тої градації, фактор В 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:

$$y_{ijk} - \mu + \alpha_i + \beta_j + \gamma_{ij} + e_{ijk}, \qquad i = \overline{1, I} \quad j = \overline{1, J} \quad k = \overline{1, K}, \qquad (1)$$

де y_{ijk} - спостереження за нзалежною змінною η;

μ - загально середнє;

α_i - кількісний вираз відносно впливу і- тої градації ξ_A на η (головний ефект і-того рівня фактору А);

β_j - кількісний вираз відносно впливу j-тої градації фактору В;

γ_{ij} - кількісний вираз відносно впливу і-тої градації фактору А, j-ї градації фактору В 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 модель (1), відомі тільки e_{ijk} та які градації чому відповідають.

Припускаємо: $\forall v_i > 0, \quad \exists w_j > 0:$

$$\begin{cases} \sum_{i=1}^{I} v_i \alpha_i = 0 \\ \sum_{j=1}^{J} w_i \gamma_{ij} = 0, \\ \sum_{j=1}^{J} w_j p_j = 0 \\ \sum_{i=1}^{I} v_j \gamma_{ij} = 0 \end{cases} \qquad \forall i, \quad \forall j$$

Наявність таких лінійних обмежень дозволяє стверджувати, що методом МНК ми зможемо знайти оцінки вектора невідомих параметрів з моделі (1).

Зауваження.

З метою спрощення виразів розглянемо випадок

$$v_i = \frac{1}{I}, \quad \forall i$$
$$w_j = \frac{1}{J}, \quad \forall j$$
$$k_{ij} = k, \quad \forall i, j$$
$$N = I \cdot J \cdot K.$$

Перевіремо гіпотези
$$H_A: \quad \alpha_1 = \alpha_2 = \ldots \alpha_I, \quad \gamma > 0;$$
$$H_B: \quad \beta_1 = \beta_2 = \ldots \beta_J, \quad \gamma > 0;$$
$$H_{AB}: \quad \gamma_{ij} = 0, \quad \forall i, j, \quad \gamma > 0.$$

Оцінки МНК об'єкту (1) при наявності обмежень (2) обчислюються за формулами:
(Крапочка замість індексу означає, що по цьому індексу береться усереднене.)

$$\hat{\mu} = \bar{y};$$
$$\hat{\alpha}_i = y_{i\bullet\bullet} - \bar{y};$$
$$\hat{\beta}_j = y_{\bullet j\bullet} - \bar{y};$$
$$\hat{\gamma}_{ij} = (y_{ij\bullet} - \bar{y}) - \hat{\alpha}_i - \hat{\beta}_j = y_{ij\bullet} - y_{i\bullet\bullet} - y_{\bullet j\bullet} + \bar{y};$$

$$y_{ij\bullet} = \frac{1}{K} y_{ijk}; \qquad y_{i\bullet\bullet} = \frac{1}{JK}\sum_{j=1}^{J}\sum_{k=1}^{K} y_{ijk}; \qquad y_{\bullet j\bullet} = \frac{1}{IK}\sum_{i=1}^{I}\sum_{k=1}^{K} y_{ijk}.$$

Розглянемо вираз:

$$(y_{ijk} - \bar{y}) = (y_{ijk} - y_{ij\bullet}) + (y_{ij\bullet} - y_{i\bullet\bullet} - y_{\bullet j\bullet} + \bar{y}) + (y_{i\bullet\bullet} - \bar{y}) + (y_{ij\bullet} - \bar{y}), \quad \Rightarrow$$

$$\Rightarrow \sum_{i,j,k} (y_{ijk} - \bar{y})^2 = \sum_{i,j,k} (y_{ijk} - y_{ij\bullet})^2 + K\sum_{i,j} (y_{ij\bullet} - y_{i\bullet\bullet} - y_{\bullet j\bullet} + \bar{y}) +$$
$$+ JK\sum_{i} (y_{i\bullet\bullet} - \bar{y})^2 + IK\sum_{j} (y_{ij\bullet} - \bar{y})^2;$$

Розглянемо таблицю результатів.

Джерело Варіації	СК	КСС
Головний ефект фактору А	S_A	$(I-1)$
Головний ефект фактору В	S_B	$(J-1)$
Головний ефект взаємодії	S_A	$(I-1)(J-1)$
Загальна сума квадратів	S_e	$IJ(K-1)$
	$S_П$	$(N-1)$

$$H_A: \quad F_A = \frac{S_A / (I-1)}{S_e / IJ(K-1)} < F_\gamma(I-1, IJ(K-1));$$

$$H_B: \quad F_B = \frac{S_B / (I-1)}{S_e / IJ(K-1)} < F_\gamma(J-1, IJ(K-1));$$

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

Зауваження.

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

@Лекція 10

Регресійний аналіз

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

Нехай маємо η – залежну кількісну скалярну змінну.

$\xi \in R^p$ – 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) : $f(x) = M\left(\eta/\xi = x\right)$.

Лема

Нехай $M\eta^2 < \infty$. Позначимо $\Re : R^q \to R^1$, тоді $f(\bullet) = \arg\min\limits_{g(\bullet)\in\Re} M[\eta - g(\xi)]^2$.

$\lhd$ Припустимо, що $Mg^2(\xi) < \infty$. Розглянемо $M[\eta - g(\xi)]^2 = M[\eta - f(\xi) + f(\xi) - g(\xi)]^2 =$
$= M[\eta - f(\xi)]^2 + 2M(\eta - f(\xi))(f(\xi) - g(\xi)) + M[f(\xi) - g(\xi)]^2 \geq$
$\geq M[\eta - f(\xi)]^2 + 2M\left\{\left(M\left(\eta/\xi\right) - f(\xi)\right)(f(\xi) - g(\xi))\right\} = M[\eta - f(\xi)]^2$. $\rhd$

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

$y(k) = f(x(k)) + e(k)$, $k = \overline{1, N}$

Основні етапи розв'язку задачі регресійного аналізу

1. Вибір класу апроксимуючих функцій $g(x, \alpha) \in \Im$.
2. Отримання точеної чи множинної оцінки для вектора невідомих параметрів, а також її характеристики розсіяності. $M(\hat{\alpha} - \alpha)(\hat{\alpha} - \alpha)^T$, де α –точне значення , $\hat{\alpha}$ – оцінка для α.
3. Перевірка на значимість відхилення параметрів моделі від нуля: $H_0 : \alpha_i = 0$ з рівнем значимості $\gamma > 0$. $H_0 : \alpha = 0, \gamma > 0$.
4. Перевірка на адекватність отриманої моделі: $M[\eta - g(\xi, \alpha)]^2, \frac{1}{N}\sum\limits_{k=1}^{N}[y(k) - g(x(k), \alpha)]^2$.

Класичний регресій ний аналіз

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

$y(k) = \sum\limits_{i=1}^{p} \varphi_i(x'(k))\alpha_i + e(k), \ k = \overline{1, N}$ (1)

$$x(k)=\begin{pmatrix}\square\varphi_1(x'(k))\\ \varphi_2(x'(k))\\ \ldots\\ \varphi_p(x'(k))\end{pmatrix},\ \alpha=\begin{pmatrix}\alpha_1\\ \alpha_2\\ \ldots\\ \alpha_p\end{pmatrix}.$$

$$y(k)=\sum_{i=1}^{p}x_i(k)\alpha_i+e(k)=x^T(k)\alpha+e(k),\ k=\overline{1,N}.$$

$x_i(k)$– регресори. Деяка функція від векторів незалежних змінних

$$y=\begin{pmatrix}y(1)\\ ..\\ y(N)\end{pmatrix},\ e=\begin{pmatrix}e(1)\\ \ldots\\ e(N)\end{pmatrix},\ X=\begin{pmatrix}x^T(1)\\ \ldots\\ x^T(N)\end{pmatrix}.$$

Тоді (2) можна переписати у вигляді: $y=X\alpha+e$ (3)

Основні припущення класичного регресійного аналізу

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

Будемо шукати оцінку мінімізуючи функціонал:

$$\sum_{k=1}^{N}e^2(k)=\|y-X\alpha\|^2=\sum_{k=1}^{N}\left[y(k)-X^T(k)\alpha\right]^2\to\min_{\alpha}$$

Точкою мінімуму буде $\hat{\alpha}=(X^TX)^{-1}X^Ty$.

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

Розпишемо функціонал $\|y-X\alpha\|^2=\alpha^TX^TX\alpha-2\alpha^TX^Ty+\|y\|^2$

$$grad_\alpha\left\{\|y-X\alpha\|^2\right\}\Big|_{\alpha=\hat{\alpha}}=\left\{2X^TX\alpha-2X^Ty\right\}\Big|_{\alpha=\hat{\alpha}}=0$$

Отже, в системі $X^TX\hat{\alpha}=X^Ty$ матриця X^TX – не вироджена, тому

$\hat{\alpha}=(X^TX)^{-1}X^Ty$ (4)

Таким чином $\hat{y}(k)=X^T(k)\hat{\alpha}$ (5)

$\hat{y}=X\hat{\alpha}$ (6)

$\hat{\sigma}^2=\dfrac{\|y-X\hat{\alpha}\|}{N-p}$ (7) – незміщена оцінка максимальної правдоподібності.

Властивості оцінок

1.а) $\hat{\alpha}\sim N\left(\alpha,\sigma^2(X^TX)^{-1}\right)$

б) $\hat{\alpha}_i\sim N(\alpha,\sigma^2d_i),\ d_i=\left\{(X^TX)^{-1}\right\}_i$

2.статистика $\dfrac{(\hat{\alpha}-\alpha)X^TX(\hat{\alpha}-\alpha)}{\sigma^2}\sim\chi^2(p)$

3.а) $\hat{y}\sim N\left(X\alpha,\sigma^2X(X^tX)^{-1}X^T\right)$

б) $\hat{y}(k)\sim N\left(X^T(k)\alpha,\sigma^2X^T(k)(X^TX)^{-1}X(k)\right)$

4.а) $M\hat{\sigma}^2 = \sigma^2$ незміщена

б) $(N-p)\dfrac{\hat{\sigma}^2}{\sigma^2} \sim \chi^2(N-p)$

5. $\hat{\alpha}, \hat{\sigma}^2$ незалежні

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

Позначимо U_α

$$M(\hat{\alpha}-\alpha)(\hat{\alpha}-\alpha)^T = \inf_{\tilde{\alpha}\in U_\alpha} M(\tilde{\alpha}-\alpha)(\tilde{\alpha}-\alpha)^T$$

Є дві симетричні матриці P та Q тоді кажуть, що $P < Q$, якщо для довільного вектора $l \neq \theta$ виконується $l^T P l \le l^T Q l$

7. $y_N = x_N\alpha + e_N$ (8)

Оцінка по об'єму спостережень $\hat{\alpha}(N)$

$\hat{\alpha}(N)"\to" \Leftrightarrow \left(X_N^T X_N\right)^{-1} \underset{N\to\infty}{\to} \Theta$ - нульова матриця

Доведемо деякі властивості

$\triangleleft$ 1 а)

Згідно (4) $\hat{\alpha} = \left(X^T X\right)^{-1} X^T (X\alpha + e) = \left(X^T X\right)^{-1} X^T X\alpha + \left(X^T X\right)^{-1} X^T e = \alpha + \left(X^T X\right) X^T e$ (9)

$$M\hat{\alpha} = \alpha$$

$$M(\hat{\alpha}-\alpha)(\hat{\alpha}-\alpha)^T = M\left\{\left(X^T X\right)^{-1} ee^T X\left(X^T X\right)^{-1}\right\} = \sigma^2\left(X^T X\right)^{-1}\left(X^T X\right)\left(X^T X\right)^{-1} = \sigma^2\left(X^T X\right)^{-1}$$

1 б) Потрібно визначити розмір компоненти

$$\hat{\alpha}_i = l_i^T\hat{\alpha}$$

$$\hat{\alpha}_i \sim N\left(l_i^T\hat{\alpha}, \sigma^2 l_i^T \left(X^T X\right)^{-1} l_i\right)$$

$$\hat{\alpha}_i \sim N\left(\alpha_i, \sigma^2 d_i\right)$$

2

<u>Лема1</u>

Нехай $\xi \in R^1$, $\xi \sim N(m,R)$, $R>0$, тоді виконується (*) $(\xi-m)^T R^{-1}(\xi-m) \sim \chi^2(q)$

$$\triangleleft \quad M(\xi-m)=0 \quad R^{-\frac{1}{2}}(\xi-m) \sim N\left(\theta, E_q\right) \quad R = T\begin{pmatrix} \mu_1 & & 0 \\ & \ddots & \\ 0 & & \mu_n \end{pmatrix} T^T$$

$\left\|R^{-\frac{1}{2}}(\xi-m)\right\|^2 \sim \chi^2(q)$ співпадає з квадратичною формою (*) $\triangleright$

Застосуємо лему 1 до $\hat{\alpha}$

$$(\hat{\alpha}-\alpha)^T\left\{\sigma^2\left(X^T X\right)^{-1}\right\}^{-1}(\hat{\alpha}-\alpha) = \frac{(\hat{\alpha}-\alpha)^T\left(X^T X\right)(\hat{\alpha}-\alpha)}{\sigma^2} \sim \chi^2(p)$$

3 а) $\hat{y} = X\hat{\alpha}$, $\hat{y} \sim N\left(X\alpha, \sigma^2 X\left(X^T X\right)^{-1} X^T\right)$

б) $\hat{y}(k) = l_k^T\hat{y}$, $\hat{y}(k) \sim N\left(l_k^T X\alpha, \sigma^2 l_k^T X\left(X^T X\right)^{-1} X^T l_k\right)$, де $l_k^T X = X^T(k)$, а $X^T l_k = X(k)$ $\triangleright$

Теорема Андерсона-Тейлора

Довірчі області та інтервали для невідомих параметрів моделі.

I. Довірча область для α з рівнем значимості $(1-\gamma), \gamma>0$

З властивості II маємо $\chi^2(p)=\dfrac{(\hat{\alpha}-\alpha)^T X^T X(\hat{\alpha}-\alpha)}{\sigma^2}\overset{(9)}{=}\dfrac{e^T X\left(X^T X\right)^{-1}\left(X^T X\right)^{-1} X^T e}{N-p}=$

$=\dfrac{e^T X\left(X^T X\right)^{-1} X^T e}{\sigma^2}$

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

маємо $\dfrac{\left\|y-X\left(X^T X\right)^{-1} X^T y\right\|^2}{N-p}=\dfrac{\left\|\left(E-X\left(X^T X\right)^{-1} X^T\right)(X\alpha+e)\right\|^2}{N-p}=$

$=\dfrac{\|PX\alpha+Pe\|^2}{N-p}=\dfrac{e^T P^T Pe}{N-p}=\dfrac{e^T P^2 e}{N-p}=\dfrac{e^T Pe}{N-p}$, де

$P=E-X\left(X^T X\right)^{-1} X^T$ тому $PX=\Theta$ (10)

Покажемо, що $P^2=P$:

$P^2=\left(E-X\left(X^T X\right)^{-1} X^T\right)^2=E-2X\left(X^T X\right)^{-1} X^T+X\left(X^T X\right)^{-1} X^T X\left(X^T X\right)^{-1} X^T=$

$=E-X\left(X^T X\right)^{-1} X^T=P$,

отже маємо $\hat{\sigma}^2=\dfrac{e^T Pe}{(N-p)}$ (11)

і отже $\dfrac{(N-p)(\hat{\sigma})^2}{\sigma^2}=\dfrac{(N-p)e^T Pe}{(N-p)\sigma^2}=\dfrac{e^T Pe}{\sigma^2}$ (**),

візьмемо $B=\dfrac{P}{\sigma^2}$

<u>Лема2</u>

Нехай $\xi\in R^n$, $\xi\sim N\left(m,\sigma^2 E\right)$, A,B - матриці розмірності $N\times N$, тоді квадратичні форми $\xi^T A\xi, \xi^T B\xi$ будуть незалежні тоді і тільки тоді, коли $AB=\Theta$

Впевнимось, що (*) та (**) незалежні $\dfrac{X\left(X^T X\right)^{-1} X^T}{\sigma^2}\dfrac{P}{\sigma^2}=\Theta$

$$\frac{\dfrac{(\hat{\alpha}-\alpha)^T X^T X(\hat{\alpha}-\alpha)}{\sigma^2 P}}{\dfrac{(N-p)\hat{\sigma}^2}{\sigma^2(N-p)}}=\frac{(\hat{\alpha}-\alpha)^T X^T X(\hat{\alpha}-\alpha)}{\hat{\sigma}^2 P}\sim F(p,N-p)$$

$\dfrac{(\hat{\alpha}-\alpha)^T X^T X(\hat{\alpha}-\alpha)}{\hat{\sigma}^2 P}\sim F_\gamma(p,N-p)$ - довірча область (12)

II. Довірчий інтервал для α_i

За властивістю I маємо $N(0,1)\sim\dfrac{\hat{\alpha}_i-\alpha_i}{\sigma\sqrt{d_i}}\overset{(9)}{=}\dfrac{l_i^T\left(X^T X\right)^{-1} X^T e}{\sigma\sqrt{d_i}}$ (***)

Позначимо $\dfrac{l_i^T\left(X^T X\right)^{-1} X^T}{\sigma\sqrt{d_i}}=A$

отже $\dfrac{l_i^T\left(X^T X\right)^{-1} X^T}{\sigma\sqrt{d_i}}\dfrac{P}{\sigma^2}=AB=\Theta$ - статистично незалежні.

Тоді

$$\frac{\dfrac{\hat{\alpha}_i-\alpha_i}{\sigma\sqrt{d_i}}}{\sqrt{\dfrac{(N-p)\hat{\sigma}^2}{\sigma^2(N-p)}}}=\frac{\hat{\alpha}_i-\alpha_i}{\hat{\sigma}\sqrt{d_i}}\sim t(N-p)$$

довірчий інтервал для α_i $\left|\dfrac{\hat{\alpha}_i-\alpha_i}{\hat{\sigma}\sqrt{d_i}}\right|<t_{\frac{\gamma}{2}}(N-p)$ (13)

$\hat{\alpha}_i-\hat{\sigma}\sqrt{d_i}t_{\frac{\gamma}{2}}(N-p)<\alpha_i<\hat{\alpha}_i+\hat{\sigma}\sqrt{d_i}t_{\frac{\gamma}{2}}(N-p)$ - інтервал

III. Інтервали Бонфероні

Якщо для кожної компоненти побудувати довірчий інтервал з рівнями довіри $1-\dfrac{\gamma}{p}$, $p\dim\alpha$.

A_i - ймовірність того, що і-та компонента $\in$ своєму довірчому інтервалу, який побудований за (13).

$\bigcap\limits_{i=1}^{p}A_i$ - всі компоненти $\in$ своїм довірчим інтервалам.

Треба знайти $P\left\{\bigcap\limits_{i=1}^{p}A_i\right\}=1-P\left\{\overline{\bigcap\limits_{i=1}^{p}A_i}\right\}=1-P\left\{\bigcup\limits_{i=1}^{p}\overline{A_i}\right\}\geq 1-\sum\limits_{i=1}^{p}\underbrace{P\{\overline{A_i}\}}_{\frac{\gamma}{p}}=1-\gamma$.

(Знак $\geq$ в ланцюжку з'являється внаслідок того, що $P\left\{\bigcup\limits_{i=1}^{p}\overline{A_i}\right\}\leq\sum\limits_{i=1}^{p}P\{\overline{A_i}\}$).

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

Перевірка на значимість параметрів моделі

1). Перевіряємо гіпотезу:

$H_0:\alpha=0$ - вектор параметрів, $\gamma>0$ - рівень довіри.

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

Критична область – область великих значень.

Область прийняття гіпотези: $\dfrac{\|X\hat{\alpha}\|^2}{p\hat{\sigma}^2}<F_\gamma(p,N-p)$ - $100\gamma\%$ точка F - розподілу з параметрами $(p,N-p)$.

2). Перевіряємо гіпотезу:

$H_0:\alpha_i=0,\gamma>0$.

$\dfrac{\hat{\alpha}_i-\alpha_i}{\hat{\sigma}\sqrt{d_i}}\sim t(N-p)$. При справедливості H_0, статистика: $t_i=\dfrac{\hat{\alpha}_i}{\hat{\sigma}\sqrt{d_i}}\sim t(N-p)$.

d_i - і-й діагональний елемент матриці $\left(X^T X\right)^{-1}$.

Критична область – область дуже малих і дуже великих значень.

Область прийняття: $\dfrac{|\hat{\alpha}_i|}{\hat{\sigma}\sqrt{d_i}}<t_{\frac{\gamma}{2}}(N-p)$.

3). Нехай в моделі перший регресор $x_1(k)\equiv 1$. Тоді $y(k)=\alpha_1+\sum\limits_{i=2}^{p}\alpha_i x_i(k)+e(k)$, $k=\overline{1,N}$.

Потрібно перевірити на значимість всі $\alpha_i : i=\overline{2,p}$.

$H_0 : \alpha_2 = \alpha_3 = ... = \alpha_p = 0$, $\gamma > 0$.

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, та множина незалежної змінної суттєво відхиляються від 0.

Тобто $\hat{\alpha}_1 = \hat{y}(k) = \bar{y} = \frac{1}{N}\sum_{k=1}^{N} y(k)$.

$$H_0 : \underbrace{\begin{pmatrix} 0 & 1 & .. & 0 \\ 0 & 0 & \ddots & 0 \\ .. & .. & .. & 1 \\ 0 & 0 & .. & 0 \end{pmatrix}}_{A}, \ \alpha = \begin{pmatrix} 0 \\ .. \\ 0 \end{pmatrix}, \ \gamma > 0, \ rangA = p-1.$$

Тоді за теоремою про перевірку лінійної гіпотези:

$$F = \frac{(Q(\hat{\alpha}_Z) - Q(\hat{\alpha}))/(p-1)}{\hat{\sigma}^2} = \text{за зауваженням до теореми} = \frac{\|X\hat{\alpha} - X\hat{\alpha}_Z\|^2/(p-1)}{\hat{\sigma}^2} =$$

$$= \frac{\sum_{i=1}^{N}\left(X^T(k)\hat{\alpha} - \bar{y}\right)^2/(p-1)}{\hat{\sigma}^2} \Rightarrow F < F_\gamma(p-1, N-p).$$

Зауваження до пункту 2):

Знайдемо зв'язок між частинною F - статистикою і статистикою t_i.

За теоремою для перевірки гіпотези $H_0 : \alpha_1 = 0, \gamma > 0$.

$$F_i = \frac{\hat{\alpha}_i^{\,2}[d_i]^{-1}}{\hat{\sigma}^2} = \frac{\hat{\alpha}_i^{\,2}}{\hat{\sigma}^2 d_i} = t_i^{\,2}$$

$$A = (0 ... \underset{i-me}{1} ... 0), \ b = 0$$

!!!!!! $\dfrac{(b - A\hat{\alpha})^T\left[A\left(X^TX\right)^{-1}A^T\right]^{-1}(b - A\hat{\alpha})}{\hat{\sigma}^2}$.

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 $X\alpha$.

Згідно (12) можна записати ліву частину у вигляді:

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

2). Довірча область для $x^T(k)\alpha$.

За властивістю <u>3) б)</u> :

$$\hat{y}(k) \sim N\left(x^T(k)\alpha, \sigma^2 x^T(k)\left(X^TX\right)^{-1}x(k)\right).$$

Тоді наступна статистика: $\dfrac{\hat{y}(k) - x^T(k)\alpha}{\sigma\sqrt{x^T(k)\left(X^TX\right)^{-1}x(k)}} \sim N(0,1)$.

За властивістю <u>4) б)</u> :

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

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

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

Довірчий інтервал:

$$\hat{y}(k)-\hat{\sigma}\sqrt{x^T(k)\left(X^TX\right)^{-1}x(k)}<x^T(k)\alpha<\hat{y}(k)+\hat{\sigma}\sqrt{x^T(k)\left(X^TX\right)^{-1}x(k)}.$$

Перевірка на адекватність

Перевірка на адекватність здійснюється шляхом перевірки спільномірності оцінки $\hat{\sigma}$, 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. $\hat{\alpha}=\left(X^TX\right)^{-1}X^Ty$.

Знаходимо оцінки $\hat{\alpha},\hat{\sigma}$: $\hat{\sigma}^2=\left\|y-X\hat{\alpha}\right\|^2/(N-p)$.

Оцінка $\hat{\alpha}$ зміщена, тобто $M\hat{\alpha}=\alpha+\Delta\alpha=\alpha+\left(X^TX\right)^{-1}X^TX_1\alpha_1$.

$\hat{\alpha}$ - неслушна оцінка, $\hat{\sigma}$ - зміщена оцінка, тобто: $M\hat{\sigma}^2=\sigma^2+\Delta\sigma^2$.

2). Нехай в істинній моделі менше параметрів, ніж у вибраній, у якій є зайві.

Тобто: $My = X\alpha$ істинна, а $My = X\alpha + X_1\alpha_1 = \overline{X}\overline{\alpha}$ - вибрана.

$$\overline{X}=\left(X \vdots X_1\right),\ \overline{\alpha}=\begin{pmatrix}\alpha\\ \alpha_1\end{pmatrix}.$$

В цьому випадку:

$\hat{\overline{\alpha}}$ – незміщена, і $M\hat{\overline{\alpha}}=\begin{pmatrix}\alpha\\ \Theta\end{pmatrix}$

$\hat{\overline{\alpha}}$ – слушна оцінка.

Оцінка $\hat{\sigma}^2$ - незміщена, $M\hat{\sigma}^2=\sigma^2$.

Ми втратили точності оцінки у вигляді:

$M\left(\hat{\overline{\alpha}}-M\hat{\overline{\alpha}}\right)\left(\hat{\overline{\alpha}}-M\hat{\overline{\alpha}}\right)^T=\sigma^2\left(X^TX\right)^{-1}+\Delta u$, $\Delta u\geq 0$ Точність оцінки може збільшуватись.