- **Що таке регресійний аналіз?**
- Це один з розділів аналізу даних який займається побудовою математичної моделі істотних зв'язків між кількісними змінними (залежними і незалежними)
- **Постановка задачі регресійного аналізу.**
- У нас є кількісна змінна η(Ета), вона є залежною. І є кількісний вектор незалежних змінних ξ(Ксі). Кінцева мета – побудова математичної моделі істотного зв'язку.
- **Який вигляд має ця математична модель?**
- $\eta = f(\bar{\xi}) + \varepsilon$**,** де ε – залишкова похибка апроксимації, а $f(\bar{x}) = M(\eta/\bar{\xi} = \bar{x})$ – функція регресії η щодо $\bar{\xi}$, причому $f(\bar{\xi})$ буде найкращою у середньоквадратичному розумінні апроксимацією η на класі борелівських функцій на множині R^q.
- **В якості функції f ми можемо взяти яку функцію?**
- Функцію регресії η(Ета) щодо ξ(Ксі).
- **Етапи розв'язання задачі регресійного аналізу.**
- Перший етап – параметричний вибір класу апроксимуючих функцій.

1. вибір класу апроксимуючих функцій $\tilde{F}$ для $f(\vec{x})$, тобто для функції регресії η щодо $\vec{\xi}$:

$$\tilde{f}(\vec{x},\alpha)\in\tilde{F},\quad \vec{x}\in\mathbb{R}^q,\alpha\in\mathbb{R}^p,$$

де α - вектор невідомих параметрів;

Другий етап – визначення оптимальної точкової оцінки

2. визначення оптимальної, згідно деякого критерію якості, точкової оцінки $\hat{\alpha}$ для α та її характеристики розсіювання $M(\hat{\alpha}-\alpha)(\hat{\alpha}-\alpha)^T$, або множинної оцінки для α, тобто довірчої області для α;

Третій етап – перевірка на значимість параметрів моделі у випадку лінійності моделі по α, а саме перевіряємо гіпотези:

$$H_0:\alpha=\theta,\gamma>0,\quad \text{або}$$
$$|H_0:\alpha_i=0,\gamma>0.$$

Четвертий етап – перевірка на адекватність отриманої моделі

Зауваження. В якості критерію якості на другому етапі при обчисленні точкової оцінки $\hat{\alpha}$ для α найчастіше використовують такий функціонал:

$$M\left(\eta-\tilde{f}\left(\vec{\xi},\alpha\right)\right)^2.$$

Вибіркове представлення його має такий вигляд:

$$\frac{1}{N}\sum_{k=1}^{n}\left(y(k)-\tilde{f}\left(\vec{x}'(k),\alpha\right)\right)^2. \qquad (1.1)$$

Саме його і буде використано у подальшому.

- **В класичному регресійному аналізі в якості апроксимуючої функції для цієї функції регресії η(Ета) щодо ξ(Ксі) ми вибирали яку функцію?**

- Функцію лінійну по параметрам у вигляді лінійної комбінації:

$$\eta = \sum_{i=1}^{p} \alpha_i \varphi_i\left(\vec{\xi}\right) + \varepsilon_{\mathbb{P}}$$

  Де φ_i – обраний нами набір функцій(беруться декілька перших членів з повного набору або незалежних функцій, або ортогоналних(що краще)), ε –похибка моделі.

- **Як називається оце $\varphi_i\left(\vec{\xi}\right)$?**
- і-тий регрессор
- **А і-та компонента вектора Ксі це і-та незалежна змінна.**
- $$y(k) = \sum_{i=1}^{p} \alpha_i \varphi_i\left(\vec{x}'(k)\right) + e(k), k = \overline{1, N} \quad \overbrace{\varphi(x'(k))}^{x(k)} \; y(k) = x^T(k)\alpha + e(k), k = \overline{1, N}.$$

Останню систему рівнянь можна записати у матричному вигляді:

$$y = X\alpha + e \qquad (1.5)$$

Де 'y' – вектор-стовпчик з компонентами 'y_k', 'e' – вектор-стовпчик з компонентами e_k, α -- вектор-стовпчик з компонентами α_i, матриця X – це V-матриця k-тий рядок якої дорівнює $x^T(k)$

$$y = \begin{pmatrix} y(1) \\ y(2) \\ \vdots \\ y(N) \end{pmatrix}, X = \begin{pmatrix} x^T(1) \\ x^T(2) \\ \vdots \\ x^T(N) \end{pmatrix} \in M_{N,p}(\mathbb{R}), e = \begin{pmatrix} e(1) \\ e(2) \\ \vdots \\ e(N) \end{pmatrix}. \quad \underset{N}{y} = \underset{N \times p}{X} \underset{p}{\alpha} + \underset{N}{e}$$

- **Припущення класичного регресійного аналізу.**
- Перше припущення – е має нормальний розподіл з параметрами ТЕТА розмірності N, СІГМА^2 * Е(одиничну матрицю) розмірності N. З цього випливає що усі компоненти вектора е – незалежні і однаково розподілені.

  Друге припущення – ранг матриці X дорівнює p. Дивлячись на розмірність бачимо що він повний по стовпчикам

  Третє припущення – немає ніяких обмежень на α

Припущення класичного регресійного аналізу:

I. $e \sim \mathcal{N}\left(\theta_N, \sigma^2 E_N\right), \sigma^2 \in \mathbb{R}_+,$

II. $rank(X) = p,$

III. немає ніяких обмежень на α, тобто $\alpha \in \mathbb{R}^p$.

Наша кінцева мета – пошук оцінки α.

- **Яким методом будемо шукати?**
- Методом найменших квадратів
- **Як визначається ця оцінка?**

- $\hat{\alpha} = \arg\min\limits_{\alpha} \|e\|^2,$

Означення. Оцінка $\hat{\alpha}$ для вектора невідомих параметрів α моделі (1.5), яка розв'язком задачі

$$\hat{\alpha} = \arg\min_{\alpha} \|e\|^2,$$

називається **оцінкою методу найменших квадратів** (МНК).

Аргумент мінімуму евклідової норми у квадраті по всім α.

- **Якщо тут не квадрат евклідової норми, а квадрат зваженої норми, як називається ця оцінка?**
- Оцінка зваженого методу найменших квадратів.

Означення. Оцінка $\hat{\alpha}$ для вектора невідомих параметрів α моделі (1.5), яка розв'язком задачі

$$\hat{\alpha}_W = \arg\min_{\alpha} \|e\|_W^2, \; W > 0,$$

називається **оцінкою зваженого методу найменших квадратів** (ЗМНК), де $\|e\|_W^2 = e^T W e$.

- **Остаточний результат розв'язку цієї задачі.**
- $\hat{\alpha} = \arg\min\limits_{\alpha} \|e\|^2 = (X^T X)^{-1} X^T y$
- **Який буде мати вигляд оцінка вектору y?**
- X на оцінку $\hat{\alpha}$: $\hat{y} = X\hat{\alpha}$
- **Який вигляд буде мати незміщена оцінка методу максимально правдоподібного для СІГМА^2**
- $\hat{\sigma}^2 = \frac{1}{N-p} \|y - X\hat{\alpha}\|^2$ (квадрат евклідової норми y – X*α з кришечкою)

<u>РОЗІДЛ ОЦІНКИ ПАРАМЕТРІВ ПРИ НАЯВНОСТІ ЛІНІЙНИХ ОБМЕЖЕНЬ</u>

*(Початок на сторінці 19 в файлі *Регресійний аналіз(лекції).pdf*)*

III'. $\alpha \in \mathcal{L}$, де $\mathcal{L} = \{\alpha : A\alpha = b, rank(A) = q\}, A \in M_{q.p}(\mathbb{R}), b \in \mathbb{R}^q$.

Тоді оцінка МНК для об'єкту (0.15) при справедливості припущень I, II, III', тобто при наявності лінійних обмежень, є розв'язком такої оптимізаційної задачі

$$\hat{\alpha}_{\mathcal{L}} = \arg\min_{\alpha \in \mathcal{L}} Q(\alpha).$$

А її вигляд задає таке твердження.

<u>ДОВІРЧІ ІНТЕРВАЛИ ТА ОБЛАСТІ ДЛЯ ПАРАМЕТРІВ РЕГРЕСІЙНОЇ МОДЕЛІ</u>

*(Початок на сторінці 24 в файлі *Регресійний аналіз(лекції).pdf*)*

- **Що таке інтервали Бонфероні?**
- Це коли процес побудови довірчої області у вигляді гіперпаралелепіпеда

- **Навіщо ми вводимо оцінку α з кришечкою?(Запитання від студента)**
- По-перше вона знадобиться при перевірці лінійних гіпотез. По-друге на практиці дуже часто виникають ситуації коли на вектор невідомих параметрів можуть накладатись саме обмеження такого плану.

<u>ПЕРЕВІРКА ЛІНІЙНИХ ГІПОТЕЗ ДЛЯ ЛІНІЙНОЇ РЕГРЕСІЙНОЇ МОДЕЛІ</u>

*(Початок на сторінці 29 в файлі *Регресійний аналіз(лекції).pdf*)*

- **Що таке і-та частинна F-статистика?**
- **Означення.** *i-тою частинною F-статистикою* називається F-статистика F_i для перевірки гіпотези
$$H_0 : \alpha_i = 0$$
а відповідний критерій для перевірки гіпотези H_0 називається *i-тим частинним F-критерієм.*
- **Який розподіл буде мати ця F-статистика?**
- **Зауваження.** З останньої теореми отримуємо вираз i-тої частинної F-статистики та вигляд її розподілу
$$F_i \sim F(1, N-p)$$
- **За допомогою чого ми можемо перевірити цю гіпотезу на значимість α_i=0 окрім і-тої частини F-статистики?**
- За допомогою статистики $t_i = \dfrac{\hat{\alpha}_i}{\hat{\sigma}\sqrt{d_i}} \sim t(N-p).$
- **Який зв'язок між t_i та F_i статистиками?**
- **Зауваження.** З (0.26) випливає, що $F_i = t_i^2$.
- **Якій гіпотезі еквівалентна ця гіпотеза $H_0 : \alpha_i = 0, i = \overline{2,p}$?**
- Гіпотезі яка зазначає що величина η та вектор змінних некорельовані.
- **Якщо всі α_i=0 крім вільного члена, то що це означає?**
- Це означає що відсутній вплив усієї множини регресорів на залежну змінну.

<u>ПЕРЕВІРКА НА АДЕКВАТНІСТЬ ЛІНІЙНОЇ РЕГРЕСІЙНОЇ МОДЕЛІ</u>

*(Початок на сторінці 37 в файлі *Регресійний аналіз(лекції).pdf*)*

- **Як буде здійснюватись перевірка на адекватність.**
- Ми будемо використовувати принцип економічності моделі, будемо перевіряти кількість регресорів у моделі.
- **Які ситуації неадекватності можливі?**
- Коли недобір регресорів і коли їх перебір
- **Якщо недобір регресорів, то чим це погано?**
- Оцінки АЛЬФА з кришечкою та СІГМА^2 з кришечкою – є зміщені та будуть неефективними, та деякі регресори також втрачені.

- **А якщо перебір?**
- На зміщеність оцінок це ніяк не вплине, проте характеристика розсіювання АЛЬФА з кришечкою може зрости. Тоді втрачається точність оцінювання (<u>за рахунок розбухання коваріаційної матриці похибки оцінювання</u>).
- **Як ви будете розв’язувати цю задачу?**

$$y = X\alpha + e$$

$\underline{\mathrm{I}}'$ $e \sim N(\theta, V)$, $V > 0$

$\underline{\mathrm{II}}$

$\underline{\mathrm{III}}$

- Наша оцінка МНК буде приймати трохи змінений вид. Це буде АЛЬФА з кришечкою для V^{-1}.
- **Якою оцінкою тут краще скористатись?**
- Більше підійде ЗМНК з ваговою матрице V^{-1} (МАРКОВСЬКА ОЦІНКА)
- **Чим саме приваблива Марковьска оцінка?**
- Вона ефективна(оцінка у векторному випадку є ефективною якщо характеристика розсіювання в неї, тобто коваріаційна матриця найменша, а отже вона буде мати найвищу точність на класі усіх незміщених оцінок)
- **Якщо виконується умова** $\underline{\mathrm{II}}'\ rank(X) = p - r,\ r \geq 1 \Leftrightarrow \exists\, a_i \quad X a_i = \theta \quad (*)$ **то ми кажем що ми знаходимся в яких умовах?**
- Строгої мультиколінеарності

Випадок строгої мультиколінеарності. У цьому випадку оцінка МНК існує і вона буде не єдина. А множина цих усіх оцінок МНК задається як множина усіх розв`язків *системи нормальних рівнянь для оцінки МНК*, а саме:

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

Зауважимо, що в умовах строгої мультиколінеарності оцінка МНК у вигляді

$$\hat{\alpha} = \left(X^T X\right)^{-1} X^T y \qquad (0.41)$$

не існує.

Випадок мультиколінеарності. В умовах мультиколінеарності оцінка МНК у вигляді (0.41) теоретично існує, бо матриця $\left(X^T X\right)$ є не виродженою, але її практичне використання буде проблематичним. Дійсно справедливість умов мультиколінеарності

- **Для подолання мультиколінеарності в деяких випадках можна використовувати що?**
- Стандартизацію незалежних змінних. Або перехід до ортогональних поліномів
- **Чим відрізняється гребнева оцінка від цієї оцінки? Як потрібно видозмінити формулу?**
- $\hat{\alpha}(\varepsilon) = \left(X^T X + \varepsilon E_p\right)^{-1} X^T y, \varepsilon > 0,$, де ЕПСІЛОН – мале позитивне дійсне число.

- **Який недолік цієї оцінки?**
- Вона є зміщеною.
- **Чому ми її продовжили використовувати?**
- Вона допомагає досягти більшої точності в середньо-квадратичному розумінні.
- **При якій умові?**
- При умові коректного вибору параметру ЕПСІЛОН.
- **Якщо ми коректно виберемо ЕПСІЛОН ми можемо досягти середньо-квадратичної похибки ще меншої ніж у якої оцінки?**
- Ніж у МНК.
- **Як ми вибираємо ЕПСІЛОН на практиці?**
- Візьмемо гребневу оцінку і-того компоненту вектора АЛЬФА. Вибираєм найменше ЕПСІЛОН при якому відбулась стабілізація усіх графіків.
- **Якщо намальовані графіки, а стабілізації нема, що це означає?**
- Якщо немає стабілізації то немає і мультиколінеарності, тому користуємось звичайною оцінкою МНК.

<u>МЕТОДИ ВИБОРУ СТРУКТУРИ РЕГРЕСІЙНОЇ МОДЕЛІ</u>

*(Початок на сторінці 49 в файлі *Регресійний аналіз(лекції).pdf*)*

- **Які ви знаєте методи вибору структури регресійної моделі?**
- Ми маємо даний набір регресорів

  $\varphi_i(\vec{\xi}) : x_i(1), \ldots, x_i(N), \; i = \overline{1,p}$

  I. Метод включення

  II. Метод виключення

  III. Пряма крокова регресія

  IV. Обернена крокова регресія

  Зауваження. Залежна змінна η матиме індекс 0. $\eta \equiv \xi_0$

  Позн.: $I = \{0, 1, 2, \ldots, p\}$
- **Розкажіть ідею методу включення.**
- У нас є початкова модель $y(k) = \alpha_0 + e_+(k), \; k = \overline{1,N}$
- **Скільки регресорів включається в праву частину даної початкової моделі?**
- Спочатку жодного регресора не включаємо.
- **Як вибираєм претендента на першочергове включення в праву частину?**
- $i_+ = \underset{i \in I_-}{\arg\max} \left| \hat{r}_{0i(I_i)} \right|$

=====================**Стрим оборвался**=====================