Лекція 11

Аналіз вимірювання ПЗ: регресійний аналіз

1. Лінійний та нелінійний регресійний аналіз.
2. Побудова лінії регресії.
3. Багатомірний регресійний аналіз.

Регресійний аналіз

- Найпростішою формою оцінки стохастичного зв'язку є одновимірний регресійний аналіз, за яким формуються обчислювальні процедури відтворення лінії регресії.
- Припускається, що дві нормально розподілені випадкові величини η та ξ пов'язані між собою лінійною регресійною з $\eta = \theta_1 + \theta_2 \xi + \varepsilon$,
- де ε - похибка, яка має нормальний розподіл

Регресія

- Регресією називають таку криву, вздовж якої розсіювання результатів спостереження мінімальне.
- Лінійну регресію визначають записом у формі

$$\overline{y}(x) = a + bx$$

Відтворення функції регресії

- ідентифікації вигляду регресійної залежності;
- вибору типу функції регресії $\hat{y}(x) = \varphi(x; \hat{\Theta})$
- оцінювання нелінійних регресійних залежностей, якщо вони мають місце;
- оцінювання точності оцінок параметрів $\hat{\Theta}$;
- перевірки адекватності відтворення регресійної залежності.

Початкові умови регресійного аналізу

- Сумісний розподіл випадкових величин η, ξ має бути нормальним.
- Дисперсія залежної змінної *y* залишається постійною при зміні значення аргументу *x*, отже,

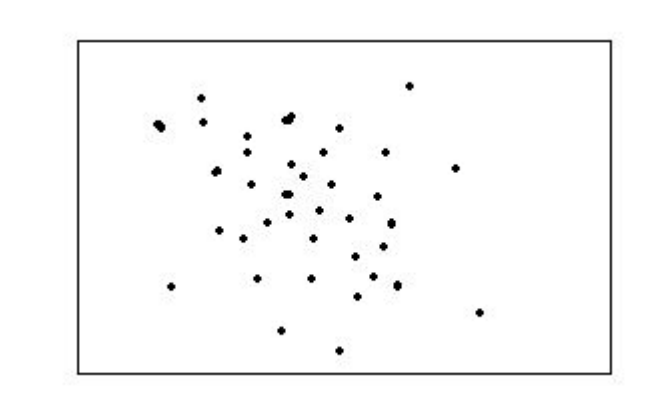
$$D\{y/x\} = \sigma^2 = \text{const}$$

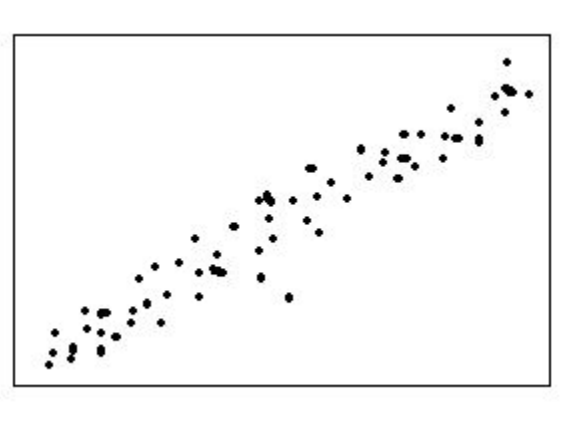
- Підсумки спостережень x_i, y_i — стохастично незалежні, отже, результати, одержані на *i* -му кроці експерименту, не пов'язані з попереднім (*i* - 1) -м кроком і не містять інформації для (*i* + 1) -го кроку.
- Якщо обсяг вибірок досить великий,

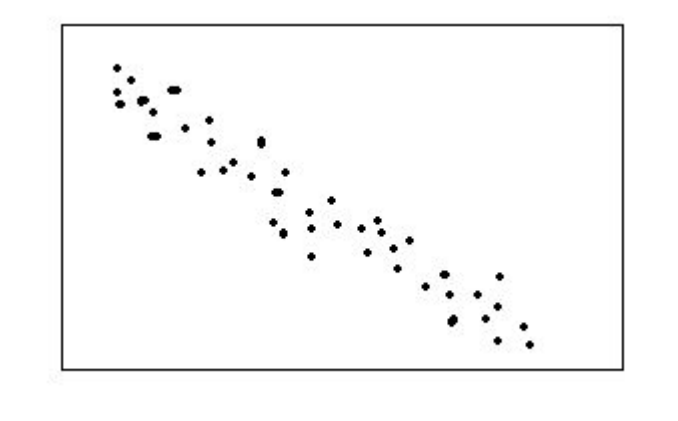
Ідентифікація регресії

- Метою процедури ідентифікації вигляду регресії є:
  - виявлення наявності зв'язку між *X* та *Y;*
  - якщо зв'язок виявлено, проведення класифікації на лінійність або нелінійність як відносно змінних *X* та *Y,* так і відносно вектора параметрів .
- Процедура ідентифікації зумовлює реалізацію як візуальної схеми, так і кількісної оцінки зв'язку. При візуалізації оцінюються початкові масиви, які відображаються на у вигляді кореляційного поля.

Кореляційні поля







Лінійний регресійний аналіз

- Лінійний зв’язок визначається

$$\bar{y}(x) = a' + bx,$$

$$\bar{y}(x) = a + b(x - \bar{x}),$$

$$\frac{\bar{y}(x) - \bar{y}}{S_y} = \hat{r}\frac{x - \bar{x}}{S_x},$$

$$\hat{a} = \bar{y}, \quad \hat{b} = \hat{r}\frac{S_y}{S_x}, \quad \hat{a}' = \bar{y} - \hat{r}\frac{S_y}{S_x}\bar{x}.$$

Довірче оцінювання емпіричної лінії регресії

- 1. Обчислення коефіцієнта детермінації R^2.
- 2. Побудова довірчого інтервалу для лінії регресії $\hat{y}(x)=\hat{a}+\hat{b}(x-\bar{x})$ з урахуванням помилки оцінки $\hat{a}$.
- 3. Оцінка відхилень окремих значень $y_j \quad i=\overline{1,n}$ залежної змінної від емпіричної регресії $\hat{y}(x_i)$.
- 4. Дослідження значущості і точності оцінок параметрів $\hat{a}$, $\hat{b}$.

Нелінійний регресійний аналіз

- У процесі ідентифікації кореляційного поля виявляється, що у багатьох випадках треба відтворювати нелінійну регресійну залежність. При цьому підбір кривої може бути здійснено на підставі:
- поліноміальної регресії другого
- $\bar{y}(x) = a + bx + cx^2$
- або більш високого порядку
- $\bar{y}(x) = a_0 + a_1 x + a_2 x^2 + \ldots + a_k x^k$, $k \geq 3$;
- нелінійних залежностей як відносно параметрів, так і відносно аргументів лінії регресії.

Метод найменших квадратів

- Використовується для відтворення параболічної регресії

$$\bar{y}(x) = a_1 + b_1\varphi_1(x) + c_1\varphi_2(x),$$

- Де $\varphi_1(x) = x - \bar{x}$;
- МНК: $\hat{a} = \bar{y} - \hat{b}\bar{x} - \hat{c}\overline{x^2}$, , де $\hat{b}$ і $\hat{c}$ отримують з

$$\begin{cases} \hat{b}\sum_{i=1}^{n}(x_i - \bar{x})^2 + \hat{c}\sum_{i=1}^{n}\left(x_i^2 - \overline{x^2}\right)(x_i - \bar{x}) = \sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x}), \\ \hat{b}\sum_{i=1}^{n}\left(x_i^2 - \overline{x^2}\right)(x_i - \bar{x}) + \hat{c}\sum_{i=1}^{n}\left(x_i^2 - \overline{x^2}\right)^2 = \sum_{i=1}^{n}(y_i - \bar{y})\left(x_i^2 - \overline{x^2}\right). \end{cases}$$

Метод найменших квадратів

- Після розв’язку системи отримуємо

$$\hat{b} = \frac{\left(\overline{x^4} - \left(\overline{x^2}\right)^2\right)\hat{r}S_x S_y - \left(\overline{x^3} - \overline{x^2}\,\overline{x}\right)\overline{(y-\bar{y})\left(x^2 - \overline{x^2}\right)}}{S_x^2\left(\overline{x^4} - \left(\overline{x^2}\right)^2\right) - \left(\overline{x^3} - \overline{x^2}\,\overline{x}\right)^2},$$

$$\hat{c} = \frac{S_x^2\overline{(y-\bar{y})\left(x^2 - \overline{x^2}\right)} - \left(\overline{x^3} - \overline{x^2}\,\overline{x}\right)\hat{r}S_x S_y}{S_x^2\left(\overline{x^4} - \left(\overline{x^2}\right)^2\right) - \left(\overline{x^3} - \overline{x^2}\,\overline{x}\right)^2}.$$

Ортогональні поліноми Чебишева

- Найпростіша обчислювальна схема відтворення поліноми регресії основана на ортогональних поліномах Чебишева
- З умови

$$\min_{\hat{a}_1,\hat{b}_1,\hat{c}_1} S^2_{\text{зал}(2)} = \min_{\hat{a}_1,\hat{b}_1,\hat{c}_1} \frac{1}{n-3}\sum_{i=1}^{n}\left(y_i - \hat{a}_1 - \hat{b}_1\varphi_1(x_i) - \hat{c}_1\varphi_2(x_i)\right)^2$$

- знаходять оцінки параметрів:

$$\hat{a}_1 = \frac{1}{n}\sum_{i=1}^{n} y_i = \bar{y}, \quad \hat{b}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\overline{(x-\bar{x})y}}{S_x^2},$$

$$\hat{c}_1 = \frac{\sum_{i=1}^{n}\varphi_2(x_i)\cdot y_i}{\sum_{i=1}^{n}\varphi_2^2(x_i)} = \frac{\overline{\varphi_2(x)\,y}}{\overline{\varphi_2^2(x)}}.$$

Ортогональні поліноми Чебишева

Підвищуючи ступінь полінома, для кожної приєднаної функції $\varphi_к(x)$ обчислюють коефіцієнт регресії, зберігаючи одержані раніше параметри.

Оцінку якості відхилення емпіричної регресії (x) від теоретичної [illegible] статистичної ха

$$t(x)=\frac{\hat{\bar{y}}(x)-\bar{y}(x)}{S_{зал(2)}\sqrt{1+\frac{(x-\bar{x})^2}{S_x^2}+\frac{\varphi_2^2(x)}{\varphi_2^2(x)}}}.$$

чим вище порядок регресійної кривої, тим більшим є розбіг довірчих меж при віддаленні від середнього

Зведення нелінійний залежностей до лінійних

- Умовно всі квазілінійні регресійні залежності поділено на чотири групи:
  - До першої групи віднесені залежності з двома невідомими параметрами, які приводяться до лінійного вигляду після відповідного перетворення координат без додаткових обчислень.
  - До другої групи віднесено залежності з трьома невідомими параметрами. Параметр у них визначається за формулою

$$c = \frac{y_1 y_3 - y_2^2}{y_1 + y_3 - 2y_2},$$

Зведення нелінійний залежностей до лінійних

- До третьої групи віднесені теоретичні залежності, які після першого перетворення координат приводяться до вигляду параболічної регресії. Тоді параметри визначаються за вищенаведеними процедурами.
- До четвертої групи відносяться теоретичні залежності, які після перетворення приводяться до лінійного рівняння з трьома невідомими параметрами.

- Подальший аналіз проводять для перетвореної в лінійну форму залежності, після чого, при необхідності, виконують зворотне перетворення.

Підбір оптимальної лінії регресії

- Задача в загальному випадку зводиться до побудови декількох ліній регресії та порівнянні оптимальних значень регресій з фактичними.
- Вибирається лінія регресії, у якої відхилення фактичних значень найменше

Множинний аналіз

- Множинна кореляція
- Множинний регресійний аналіз

Множинна кореляція

- Множинний коефіцієнт кореляції $r_{x_i \square x_1,\ldots,x_{i-1},x_{i+1},\ldots,x_k}$ є мірою лінійної залежності між змінною X_i та набором X_1, ..., X_{i-1}, X_{i+1}, ..., X_k, причому
- $0 < r_{x_i \square x_1,\ldots,x_{i-1},x_{i+1},\ldots,x_k} < 1$
- Якщо $r_{x_i \square x_1,\ldots,x_{i-1},x_{i+1},\ldots,x_k} = 0$,
- то говорять про відсутність залежності X_i від інших змінних з множини X.

Множинна кореляція

- У випадку, коли $r_{x_i \mid x_1,\ldots,x_{i-1},x_{i+1},\ldots,x_k} = 1$
- має місце лінійна залежність, при якій змінна X_t визначається лінійною комбінацією змінних X_1, …, X_{i-1}, X_{i+1}, …, X_k:
-
- $\boldsymbol{X_i = \beta_0 + \beta_1 X_1 + \beta_{i-1} X_{i-1} + \beta_{i+1} X_{i+1} + \ldots + \beta_k X_k}$.
- Квадрат коефіцієнта множинної кореляції $r^2_{x_i \mid x_1,\ldots,x_{i-1},x_{i+1},\ldots,x_k}$ оцінює частку дисперсії X_i, яка пояснюється лінійною регресією X_1, …, X_{i-1}, X_{i+1}, …, X_k.

Багатовимірний регресійний аналіз

- Багатовимірний статистичний аналіз визначає причинно-наслідкові зв’язки об’єкта дослідження і його показників (вхідних та вихідних характеристик)

Багатовимірний регресійний аналіз

- Задачею регресійного аналізу є дослідження зв’язку між залежними та незалежними величинами
- Для вирішення поставленої задачі початковий масив даних переформуються у матриці спостережень

Висновки

- Задачею регресійного аналізу є виявлення виду закономірностей у залежностях між метриками програмного забезпечення