Міністерство освіти і науки України

Національний лісотехнічний університет України

Кафедра інформаційних технологій

Методичні вказівки

до лабораторної роботи №4

“ Кореляційно-регресійний аналіз ”

з дисципліни

“Інтелектуальний аналіз даних”

для студентів з напрямку підготовки 6.050101 **“Комп’ютерні науки”**

Львів-2017

Лабораторна робота №4

Кореляційно-регресійний аналіз

Теоретичні відомості

Алгоритм кореляційно-регресійного аналізу

Даний алгоритм складається з шести частин, кожна з який або їх комбінації можуть складати окремий алгоритм, який буде вирішувати конкретну задачу.

Для реалізації алгоритму слід врахувати, що дані спостережень об'єднані в кореляційну таблицю.

I. Розрахунок вибіркового кореляційного відношення

1. За формулами (3.1) та (3.2) знаходимо суму частот по кожному рядку та по кожному стовпцю (m_j) та загальне число спостережень (n_i):

$$m_i = \sum_{i=1}^{p} m_{ji},$$

$$n_i = \sum_{j=1}^{q} m_{ji},$$

$$n = \sum_{i=1}^{p} n_i = \sum_{j=1}^{q} m_j$$

2. Розраховуємо групові середні за формулами (3.3) :

$$\overline{Y}^{(i)} = \frac{1}{n_i} \sum_{j=1}^{q} y_j m_{ji}, \quad i = 1,2,..., p$$

3. Розраховуємо групові дисперсії за формулами (3.4):

$$\hat{\sigma}_i^2 = \frac{1}{n_i} \sum_{j=1}^{q} y_j^2 m_{ji} - (\overline{Y}^{(i)})^2, \; i = 1,2,..., p$$

4. Розраховуємо середнє значення спостережуваної ознаки $\overline{Y}$ за формулою (3.5):

$$\overline{Y} = \frac{1}{n} \sum_{j=1}^{q} y_j m_j$$

5. Розраховуємо вибіркову дисперсію $\hat{\sigma}_y^2$ величини *Y* за формулою (3.6)

$$\hat{\sigma}_y^2 = \frac{1}{n} \sum_{j=1}^{q} y_j^2 m_{ji} - (\overline{Y})^2$$

6. Розраховуємо вибіркову фактичну дисперсію (вибіркову дисперсію групових середніх) за формулою (3.7)

$$\hat{\sigma}_{\hat{o}}^2 = \frac{1}{n}\sum_{j=1}^{q}(\overline{Y}^{(i)})^2 n_i - (\overline{Y})^2$$

7. Розраховуємо вибіркову залишкову дисперсію $\hat{\sigma}_y^2$ величини Y за формулою (3.8)

$$\hat{\sigma}_0^2 = \frac{1}{n}\sum_{i=1}^{p}(\hat{\sigma}_i^2 n_i)$$

8. Здійснюємо перевірку за формулою (3.9)

$$\hat{\sigma}_y^2 = \hat{\sigma}_{\hat{o}}^2 + \hat{\sigma}_0^2$$

9. Знаходимо вибіркові кореляційні відношення $\hat{\rho}_{xy}$ і вибірковий коефіцієнт детермінації $\hat{\rho}_{xy}^2$ у відповідності з формулами (3.10)

$$\hat{\rho}_{xy} = \sqrt{\hat{\sigma}_{\hat{o}}^2 / \hat{\sigma}_y^2}\ ,\ \hat{\rho}_{xy}^2 = \hat{\sigma}_{\hat{o}}^2 / \hat{\sigma}_y^2$$

II. Перевірка гіпотези про значущість вибраного кореляційного відношення.

Перевіримо гіпотезу: $H_0 : \rho_{xy} = 0$

1. Розраховуємо не зміщення оцінки генеральної дисперсії σ_0^2:

А) $S_{\hat{o}}^2 = \hat{\sigma}_{\hat{o}}^2 \dfrac{n}{(p-1)}$;

Б) $S_0^2 = \hat{\sigma}_0^2 \dfrac{n}{(n-p)}$;

В) $S_y^2 = \hat{\sigma}_y^2 \dfrac{n}{(p-1)}$.

2. Розраховуємо спостережуване значення критерію перевірки нульової гіпотези:

$$F_{íàáë} = S_{\hat{o}}^2 / S_0^2$$

3. Розраховуємо степені свободи k_1 і k_2 F-розподілу Фішера:

$$k_1 = (p-2),\ k_2 = (n-p)$$

4. За таблицею критичних точок F-розподілу Фішера при заданому рівні значимості α знаходимо критичну точку $F_{êð}(\alpha, k_1, k_2)$.

5. Порівнюємо значення величин $F_{íàáë}$ та $F_{êð}(\alpha, k_1, k_2)$

А) якщо $F_{íàáë} > F_{êð}$, то гіпотеза відхиляється, тобто вважаємо, що $\rho_{xy} \neq 0$; це означає, що існує кореляційна залежність між величинами X та Y;

Б) якщо $F_{íàáë} < F_{êð}$, то приймаємо гіпотезу H_0, тобто вважаємо, що між величинами X та Y немає кореляційної залежності.

III. Формування гіпотези про вибір функції регресії. Розрахунок вибіркового коефіцієнта кореляції.

1. Розраховуємо $\overline{XY}, \overline{X}, \overline{Y}, \hat{\sigma}_x, \hat{\sigma}_y$ за формулами:

   А) $\overline{XY} = \frac{1}{n}\sum_{i=1}^{p} x_i \sum_{j=1}^{q} y_j$;

   Б) $\overline{X} = \frac{1}{n}\sum_{i=1}^{p} x_i n_i$, $\overline{Y} = \frac{1}{n}\sum_{j=1}^{q} y_j m_j$;

   В) $\hat{\sigma}_x = \sqrt{\frac{1}{n}\sum_{i=1}^{p} x_i^2 - (\overline{X})^2}$, $\hat{\sigma}_y = \sqrt{\frac{1}{n}\sum_{j=1}^{q} y_j^2 - (\overline{Y})^2}$

2. Розраховуємо вибірковий коефіцієнт $\hat{r}_{xy}$ кореляції за формулою:

   $\hat{r}_{xy} = \frac{\overline{XY} - \overline{X}\,\overline{Y}}{\hat{\sigma}_x \hat{\sigma}_y}$ - Формулювання гіпотези про вигляд функції регресії.

   Порівняємо значення величин $|\hat{r}_{xy}|$ та ρ_{xy}. Якщо значення $|\hat{r}_{xy}|$ близьке до ρ_{xy}, то це дає основи висунути гіпотезу про те, що функція регресії лінійна. Якщо функція регресії лінійна, то модуль генерального коефіцієнта кореляції рівний генеральному кореляційному відношенню, тобто $|\hat{r}_{xy}| = \rho_{xy}$.

IV. Побудова вибіркового лінійного рівняння регресії.

1. За розрахунками значень $\overline{XY}, \overline{X}, \overline{Y}, \hat{\sigma}_x, \hat{\sigma}_y$

   Запишемо рівняння регресії

   $\hat{Y} = \overline{Y} + \hat{r}_{xy}\frac{\hat{\sigma}_y}{\hat{\sigma}_x}(x - \overline{X})$

2. Знайдемо середню квадратичну похибку рівняння регресії

   $\hat{\sigma}_0^{\ddot{e}\ni\acute{i}} = +\sqrt{(1-\hat{r}_{xy}^2)\hat{\sigma}_y^2}$

V. Перевірка гіпотези про значущість вибіркового коефіцієнта кореляції.

Нехай згідно даним спостереження ми встановимо, що функція регресії лінійна, тобто, $|\hat{r}_{xy}|$ та ρ_{xy} майже співпадають.

Перевіримо гіпотезу: $H_0 : r_{xy} = 0$

1. Розрахуємо факторну дисперсію згідно формули

   $\hat{\sigma}_{\hat{o}}^2 = \hat{\sigma}_{\tilde{o}}^{2\ddot{e}\ni\acute{i}} = \hat{r}_{xy}^2\hat{\sigma}_y^2$

2. Розрахуємо остаточну дисперсію згідно формули

$\hat{\sigma}_0^2 = \hat{\sigma}_0^{2\text{ëçí}} = (1-\hat{r}_{xy}^2)\hat{\sigma}_y^2$

3. Розрахуємо оцінку результативного признака:
   $S_{\hat{o}}^2 = \hat{\sigma}_{\hat{o}}^2 \cdot n;\ S_0^2 = \hat{\sigma}_0^2 \cdot n/(n-2)$
4. Розраховуємо наближене значення критерію:
   $F_{набл} = S_{\phi}^2 / S_0^2$
5. Розрахуємо число степенем свободи k_1 і k_2 F-розподілу:
   $k_1 = (2-1) = 1,\ k_2 = (2-n)$
6. За таблицею критичних точок F-розподілу при заданому рівні значимості α знаходимо критичну точку $F_{кр}(\alpha, k_1, k_2)$
7. Порівнюємо $F_{набл}$ та $F_{кр}(\alpha, k_1, k_2)$:

   А) якщо $F_{набл} > F_{кр}(\alpha, k_1, k_2)$, то гіпотезу H_0 відхиляємо;

   Б) якщо $F_{набл} < F_{кр}(\alpha, k_1, k_2)$, то гіпотезу H_0 приймаємо і вважаємо, що вибірковий коефіцієнт кореляції незначний.

VI. Перевірка гіпотези про лінійність функції регресії.

Перевіримо гіпотезу про те, що функція регресії лінійна:

$$H_0 = M(Y/X = x) = M[Y] + r_{xy}\frac{\sigma_y}{\sigma_x}(x - M[X])$$

В пунктах I та III даного алгоритму були встановлені $\hat{\sigma}_y^2, \hat{\rho}_{yx}^2, \hat{r}_{xy}^2$. Використовуючи дисперсійну таблицю (Табл.1) виконаємо наступні дії:

1. Знайдемо дисперсію $\hat{\sigma}_B^2$, обумовлену нелінійною залежністю Y та X згідно формули:
   $\hat{\sigma}_B^2 = \hat{\sigma}_y^2(\hat{\rho}_{yx}^2 - \hat{r}_{xy}^2)$
2. Знаходимо дисперсію $\hat{\sigma}_0^2$, обумовлену залишковим фактором згідно формули:
   $\hat{\sigma}_0^2 = \hat{\sigma}_y^2(1-\hat{\rho}_{yx}^2)$
3. Знаходимо незміщені оцінки S_0^2 та S_B^2 генеральної дисперсійної величини Y:
   $S_B^2 = \hat{\sigma}_B^2 n/(p-2),\ S_0^2 = \hat{\sigma}_0^2 n/(n-p)$
4. Розраховуємо спостережуване значення критерію:
5. За таблицею критичних точок F-розподілу при заданому рівні значимості α знаходимо критичну точку $F_{кр}(\alpha, k_1, k_2)$
6. Порівнюємо $F_{набл}$ та $F_{кр}(\alpha, k_1, k_2)$:

   А) якщо $F_{набл} > F_{кр}(\alpha, k_1, k_2)$, то гіпотезу H_0 відхиляємо;

Б) якщо $F_{набл} < F_{кр}(\alpha, k_1, k_2)$, то гіпотезу H_0 про лінійність функції регресії не відхиляємо.

Завдання до лабораторної роботи

1. Запрограмувати алгоритм кореляційно-регресійного аналізу аналізу.
2. Вирішити завдання кореляційно-регресійного аналізу відповідно до номеру варіанту.
3. Сформулювати висновки в термінах предметної області.
4. Оформити звіт до лабораторної роботи

За даними, наведеними в таблиці, виконати кореляційно-регресійний аналіз згідно описаного вище алгоритму.

Завдання передбачає 30 варіантів В кожному варіанті всі можливі значення Y_i величини Y змінюються на постійну величину, рівну номеру варіанта + дані з першої лабораторної.

Накладні витрати (млн..грн.) (X)	1-2	2-3	3-4	4-5	5-6	6-7	7-8	8-9
Об'єм виконаних робіт (млн..грн.) (Y)	(1,5)	(2,5)	(3,5)	(4,5)	(5,5)	(6,5)	(7,5)	(8,5)
10-20 (15)	4	5						
20-30 (25)	1	3	1					
30-40 (35)	2	3	6	5	3	1		
40-50 (45)		5	9	19	8	7	2	1
50-60 (55)		1	2	7	16	9	4	2
60-70 (65)			1	5	6	4	2	2
70-80 (75)							1	