Министерство науки и высшего образования Российской Федерации

Федеральное государственное бюджетное образовательное учреждение высшего образования
«Новосибирский государственный технический университет»

Кафедра теоретической и прикладной информатики

Лабораторная работа №1
по дисциплине «Основы теории машинного обучения»

ОЦЕНКА ПАРАМЕТРОВ И ПРОВЕРКА ГИПОТЕЗ В ЛИНЕЙНЫХ МОДЕЛЯХ С КАЧЕСТВЕННЫМИ ФАКТОРАМИ

Факультет:	ПМИ
Группа:	ПМИМ-01
Студенты:	Ершов П.К. Малышкина Е.Д. Слободчикова А.Э.
Вариант:	4
Преподаватель:	Попов А.А.

Новосибирск

2021

1. Задание

1) По имеющимся данным (см. варианты заданий) сформировать матрицу наблюдений X, постулировать модель дисперсионного анализа с главными эффектами (без взаимодействий уровней факторов);
2) Провести редукцию модели к модели полного ранга, определить базис ФДО;
3) По методу МНК- оценивания провести оценивание ФДО в редуцированной модели. Проверить гипотезы о незначимости различий в эффектах уровней для каждого фактора и фактора в целом;
4) Отчет должен содержать постановочную часть, решения по редукции модели, компьютерный листинг, результаты расчетов по проверке гипотез, статистические выводы.

Данные варианта:

Уровни фактора 1	Уровни фактора 2			
	B1	*B2*	*B3*	*B4*
A1	7,14 6,9	4,11 3,95	8,13 7,97	4,07 3,99
A2	3,1 2,91	0,0 0,01	4,03 3,98	0,02 0,0
A3	7,09 6,92	4,11 3,96	8,12 8,01	4,07 4,0

2. Ход работы

В ячейках таблиц располагаются значения отклика, полученные в двух параллельных наблюдениях. Пересечение соответствующих строки и столбца определяет условия эксперимента. Таким образом, количество наблюдений – 2, количество уровней фактора 1 – 3, количество уровней фактора 2 – 4.

1) Для построения модели вида

$$y = X\theta + \varepsilon,$$

где $y - (n * 1)$ – наблюдаемый в эксперименте отклик, измеряемый в количественной шкале;

$X - \ (n * p)$ – матрица наблюдения, порождаемая факторами $x_1, \dots, x_k$,

$\theta - \ (p * 1)$ – вектор неизвестных параметров, подлежащих оцениванию из эксперимента;

ε – аддитивная случайная составляющая модели наблюдения;

необходимо сформировать матрицу наблюдений X:

$$X = \begin{bmatrix} 1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Затем вычислим μ_i – среднее значение измеряемой величины y по i-ой подгруппе, где $i = 1 \dots I$:

		Фактор 2					
	Фактор 1	B1	B2	B3	B4	Среднее значение по строкам A	
y1	A1	7,14	4,11	8,13	4,07	5,78	
y2		6,90	3,95	7,97	3,99		
y1	A2	3,10	0,00	4,03	0,02	1,76	
y2		2,91	0,01	3,98	0,00		
y1	A3	7,09	4,11	8,12	4,07	5,79	Общее сред. по A
y2		6,92	3,96	8,01	4,00		4,44
Среднее значение по столбцам B		5,68	2,69	6,71	2,69		
					4,44		
				Генеральное среднее		4,44	

Затем вычислим генеральное среднее по следующим формулам:

$$\mu = \frac{1}{n}\sum_{i=1}^{I} J_i \mu_i \text{ , где } n = \sum_{i=1}^{I} J_i ;$$

$$n = 4 + 4 + 4 + 3 + 3 + 3 + 3 = 24;$$

$\mu = \frac{1}{24}$ (4 * 5,78 + 4 * 1,76 + 4 * 5,79 + 3 * 5,68 + 3 * 2,69 + 3 * 6,71 + 3 * 2,69) = 4,44.

Таким образом, можем представить модель дисперсионного анализа в виде:

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

где $\alpha_i = \mu_i - \mu$ – эффект i-го уровня первого фактора ($i = \overline{1,3}$), $\beta_j = \mu_j - \mu$ – эффект j-го уровня второго фактора ($j = \overline{1,4}$), e_{ij} – ошибка эксперимента;

$$\alpha^T = [1.34, -2.69, 1.34], \ \beta^T = [1.24, -1.75, 2.27, -1.75].$$

2) Редуцирование построенной модели к модели полного ранга можно проводить через факторизацию матрицы $X = X_1 A$, где X_1 – матрица полного строчного ранга:

$$y = X\theta + \varepsilon = X_1 A\theta + \varepsilon = X_1\bar{\theta} + \varepsilon,$$

Где матрица A задает базис ФДО и имеет вид $A = (I_r, \tilde{A})$, I_r – единичная матрица размера r, $\tilde{A} = (X_1^T X_1)^{-1} X_1^T X_2$.

В модели в связи с ее внутренним дефектом ранга несмещенно будут оцениваться только $r = p - k = rg(X)$ линейно независимых функций, допускающих оценку (ФДО), образующих базис ФДО.

Вычислим r: $r = 8 - 2 = 6$.

В матрицу X_1 будут входить μ и следующие уровни 1 фактора: A1, A2, 2 фактора: B1, B2, B3. В матрицу X_2 – уровни A3 и B4 факторов 1 и 2 соответственно.

$$X_1 = \begin{bmatrix} 4{,}44 & 1 & 0 & 1 & 0 & 0 \\ 4{,}44 & 1 & 0 & 0 & 1 & 0 \\ 4{,}44 & 1 & 0 & 0 & 0 & 1 \\ 4{,}44 & 1 & 0 & 0 & 0 & 0 \\ 4{,}44 & 0 & 1 & 1 & 0 & 0 \\ 4{,}44 & 0 & 1 & 0 & 1 & 0 \\ 4{,}44 & 0 & 1 & 0 & 0 & 1 \\ 4{,}44 & 0 & 1 & 0 & 0 & 0 \\ 4{,}44 & 0 & 0 & 1 & 0 & 0 \\ 4{,}44 & 0 & 0 & 0 & 1 & 0 \\ 4{,}44 & 0 & 0 & 0 & 0 & 1 \\ 4{,}44 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \qquad X_2 = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 1 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 1 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 1 \end{bmatrix}$$

Таким образом:

$$\tilde{A} = \begin{bmatrix} 0.225 & 0.225 \\ -1 & 0 \\ -1 & 0 \\ 0 & -1 \\ 0 & -1 \\ 0 & -1 \end{bmatrix}$$

Значит, A примет следующий вид:

$$A = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0.225 & 0.225 \\ 0 & 1 & 0 & 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & -1 \end{bmatrix}$$

Известно, что матрица $H = (X^T X)^{-} X^T X$ имеет ранг r и ненулевые ее строки образуют базис ФДО, в нашем случае:

$$H = \begin{bmatrix} I_r & \tilde{A} \\ 0 & 0 \end{bmatrix}.$$

Таким образом, матрица A задает вид базиса ФДО. Базис ФДО составят следующие параметрические функции:

$$\bar{\mu} = \mu + \alpha_3 + \beta_4 = 4.44 + 1.34 - 1.75 = 4.03;$$

$$\overline{\alpha_1} = \alpha_1 - \alpha_3 = 1.34 - 1.34 = 0;$$

$$\overline{\alpha_2} = \alpha_2 - \alpha_3 = -2.69 - 1.34 = 4.3;$$

$$\overline{\beta_1} = \beta_1 - \beta_4 = 1.24 + 1.75 = 2.99;$$

$$\overline{\beta_2} = \beta_2 - \beta_4 = -1.75 + 1.75 = 0;$$

$$\overline{\beta_3} = \beta_3 - \beta_4 = 2.27 + 1.75 = 4.02.$$

3) Чтобы найти оценки для ФДО $\bar{\theta} = A\theta$ достаточно применить теорему Гаусса-Маркова для построенной модели: $\hat{\bar{\theta}} = (X_1^T X_1)^{-1} X_1^T y$.

Тогда:
$y_1 = [7.14, 4.11, 8.13, 4.07, 3.10, \ \ 0, 4.03, 0.02, 7.09, 4.11, 8.12, 4.07]$
$y_2 = [6.90, 3.95, 7.97, 3.99, 2.91, 0.01, 3.98, 0, 6.92, 3.96, 8.01, 4.00]$
$y_3 = [7.02, 4.03, 8.05, 4.03, 3.01, 0.01, 4.01, 0.01, 7.01, 4.04, 8.07, 4.04]$

Найдем оценки для 2 параллельных исследований:

$$\widehat{\overline{\theta 1}} = [0.92, 0.02, -4.06, 3.06, 0.02, 4.04]$$

$$\widehat{\overline{\theta 2}} = [0.90, -0.02, -4.00, 2.91, -0.02, 3.99]$$

Таким образом,

$$\hat{\bar{\theta}} = [0.91, -0.01, -4.03, 2.99, 0, 4.02].$$

Были получены оценки для μ, уровней A1, A2 и B1, B2, B3. Оценки параметров, соответствующие исключенным из модели линейно зависимым столбцам (регрессорам) – A3 и B4, автоматически считаются равными нулю.

Проверим гипотезы о незначимости различий в эффектах уровней для каждого фактора:

$$H: \overline{\alpha_i} = 0;$$

$$H: \overline{\alpha_1} - \overline{\alpha_2} = 0$$

$\overline{\alpha_1} = 0$ – гипотеза не отвергается,

$\overline{\alpha_2} = 4.3$ – гипотеза отвергается,

$\overline{\alpha_1} - \overline{\alpha_2} = -4.3$ – гипотеза отвергается.

$$H: \overline{\beta_j} = 0;$$

$$H: \overline{\beta_1} - \overline{\beta_2} = 0;$$

$$H: \overline{\beta_2} - \overline{\beta_3} = 0;$$

$\overline{\beta_1} = 2.99$ – гипотеза отвергается,

$\overline{\beta_2} = 0$ – гипотеза не отвергается,

$\overline{\beta_3} = 4.02$ – гипотеза отвергается,

$\overline{\beta_1} - \overline{\beta_2} = 2.99$ – гипотеза отвергается,

$\overline{\beta_2} - \overline{\beta_3} = -4.02$ – гипотеза отвергается.

Исходя из результатов проверок гипотез, проверим гипотезы о незначимости факторов в целом:

$$H: \begin{cases} \overline{\alpha_1} = 0 \\ \overline{\alpha_2} = 0 \end{cases};$$

$$H: \begin{cases} \overline{\beta_1} = 0 \\ \overline{\beta_2} = 0 \\ \overline{\beta_3} = 0 \end{cases}$$

$\begin{cases} \overline{\alpha_1} = 0 \\ \overline{\alpha_2} = 4.3 \end{cases}$ – гипотеза отвергается,

$$\begin{cases} \overline{\beta_1} = 2.99 \\ \overline{\beta_2} = 0 \\ \overline{\beta_3} = 4.02 \end{cases}$$ – гипотеза отвергается.

3. Статистические выводы

В ходе проведённой работы было установлено, что оба фактора являются значимыми. В тоже время, была установленная не значимость эффектов первого и третьего уровня у первого фактора и не значимость эффектов второго и четвёртого уровня у второго фактора.

4. Код

```
import numpy as p
from numpy import matrix
from numpy import linalg

X1 = matrix([[4.44, 1, 0, 1, 0, 0],
             [4.44, 1, 0, 0, 1, 0],
             [4.44, 1, 0, 0, 0, 1],
             [4.44, 1, 0, 0, 0, 0],
             [4.44, 0, 1, 1, 0, 0],
             [4.44, 0, 1, 0, 1, 0],
             [4.44, 0, 1, 0, 0, 1],
             [4.44, 0, 1, 0, 0, 0],
             [4.44, 0, 0, 1, 0, 0],
             [4.44, 0, 0, 0, 1, 0],
             [4.44, 0, 0, 0, 0, 1],
             [4.44, 0, 0, 0, 0, 0]])

print(X1)
X2 = matrix([[0, 0],
             [0, 0],
             [0, 0],
             [0, 1],
             [0, 0],
             [0, 0],
             [0, 0],
             [0, 1],
             [1, 0],
             [1, 0],
             [1, 0],
             [1, 1]])
print(X2)

y1 = matrix([7.14, 4.11, 8.13, 4.07, 3.10, 0, 4.03, 0.02, 7.09, 4.11, 8.12, 4.07])
print(y1)
```

```
y2 = matrix([6.90, 3.95, 7.97, 3.99, 2.91, 0.01, 3.98, 0, 6.92, 3.96, 8.01, 4.00])
print(y2)
y3 = matrix([7.02, 4.03, 8.05, 4.03, 3.01, 0.01, 4.01, 0.01, 7.01, 4.04, 8.07, 4.04])

A_ = (X1.T * X1).I * X1.T * X2
print("A_:")
A_l = A_.shape
for i in range(A_l[0]):
    for j in range(A_l[1]):
        print("%.4f" % A_[i, j], end=' ')
    print("\n")

print("Проверка:")
Pr = X1 * A_
print(Pr)

print("Оценки тетта1:")
theta1 = (X1.T * X1).I * X1.T * y1.T
for i in range(theta1.shape[0]):
    print("%.2f" % theta1[i])

print("Оценки тетта2:")
theta2 = (X1.T * X1).I * X1.T * y2.T
for i in range(theta2.shape[0]):
    print("%.2f" % theta2[i])

print("Оценки тетта:")
theta = (X1.T * X1).I * X1.T * y3.T
for i in range(theta.shape[0]):
    print("%.2f" % theta[i])
```