Методические указания
к лабораторной работе №1

«Разработка программных средств решения задачи методом статистических гипотез»

1 Цель работы:

Изучение метода статических гипотез для разработки программного средства

2 Задача работы:

Объяснение сути метода

3 Содержание работы:

Описание метода статических гипотез для разработки программных средств

4 Требования к отчету:

Отчет о проделанной работе должен содержать:

- название работы, ее задачи и описание последовательности выполнения;
- решение задачи по вариантам, указанным преподавателем;

Отчет представить в распечатанном виде.

5 Общие положения и методика выполнения задания

Для того чтобы оценить уровень различия двух групп инженерно-технических работников на предприятиях легкой промышленности (мужчины и женщины) по двум признакам, было проведено выборочное обследование. В выборку попали 20 человек (10 мужчин и 10 женщин). Результаты наблюдения представлены в таблице 1.

Таблица 1

Мужчины		Женщины	
X_1	X_2	X_1	X_2
2	3,0	21	7,1
10	8,2	8	6,0
17	10,0	11	8,3
20	8,0	6	6,5
22	6,5	4	4,6
12	5,5	18	8,0
20	7,5	2	8,4
9	7,3	25	9,0
18	9,7	10	8,1
4	4,4	17	7,8

В таблице X_1 - стаж работы, лет; X_2 - средняя дневная заработная плата, ден. ед.

Проверьте гипотезу о сходстве двух групп работающих:

a) по каждому признаку отдельно;
b) по двум признакам вместе.

Рекомендуемая литература:

1. **Быковский, В. В.** Применение теории планирования эксперимента в научных и инженерных расчетах: учеб. пособие / В. В. Быковский, Л. В. Быковская, Ю. А. Дормидонов. - Оренбург : ОГУ, 2002. - 66 с - ISBN 5-7410-0442-3.

2. **Костин, В. Н.** Статистические методы и модели: учеб. пособие для вузов / В. Н. Костин, Н. А. Тишина . - Оренбург : ОГУ, 2004. - 138 с. - Библиогр.: с. 125. - ISBN 5-7410-0399-0.

Методические указания
к лабораторной работе №2

«Разработка программных средств решения задачи методом кластерного анализа»

1 Цель работы:

Изучение метода кластерного анализа для разработки программного средства

2 Задача работы:

Объяснение сути метода

3 Содержание работы:

Описание метода кластерного анализа для разработки программных средств

4 Требования к отчету:

Отчет о проделанной работе должен содержать:

- название работы, ее задачи и описание последовательности выполнения;
- решение задачи по вариантам, указанным преподавателем;

Отчет представить в распечатанном виде.

5 Общие положения и методика выполнения задания

На основании приведенных данных таблицы 1 необходимо провести классификацию пяти предприятий при помощи иерархического агломеративного кластерного анализа.

Таблица 1

Номер предприятия	X_1	X_2	X_3
1	220,0	94,0	264,0
2	185,0	75,0	192,0
3	245,0	80,0	220,0
4	178,0	75,2	96,0
Среднее значение ($\bar{x}_j$)	199,6	79,5	175,4

Среднее квадратическое	28,4	7,6	65,4

Здесь: X_1 – среднегодовая стоимость основных производственных фондов, млрд. р.; X_2 – материальные затраты на один рубль произведенной продукции, коп.; X_3 – объем произведенной продукции, млрд. р.

Рекомендуемая литература:

1. **Быковский, В. В.** Применение теории планирования эксперимента в научных и инженерных расчетах: учеб. пособие / В. В. Быковский, Л. В. Быковская, Ю. А. Дормидонов. - Оренбург : ОГУ, 2002. - 66 с - ISBN 5-7410-0442-3.

2. **Костин, В. Н.** Статистические методы и модели: учеб. пособие для вузов / В. Н. Костин, Н. А. Тишина . - Оренбург : ОГУ, 2004. - 138 с. - Библиогр.: с. 125. - ISBN 5-7410-0399-0.

Методические указания
к лабораторной работе №2

«Разработка программных средств решения задачи методом дискрименантного анализа»

1 Цель работы:

Изучение метода дискрименантного анализа для разработки программного средства

2 Задача работы:

Объяснение сути метода

3 Содержание работы:

Описание метода дискрименантного анализа для разработки программных средств

4 Требования к отчету:

Отчет о проделанной работе должен содержать:

– название работы, ее задачи и описание последовательности выполнения;
– решение задачи по вариантам, указанным преподавателем;

Отчет представить в распечатанном виде.

5 Общие положения и методика выполнения задания

Имеются следующие данные по двум группам промышленных предприятий (таблица 1).

Таблица 1

Первая группа (k_1)			Вторая группа (k_2)		
Номер предприятия	Удельный вес потерь от брака, % (X_1)	Фондоотдача активной части основных фондов,	Номер предприятия	Удельный вес потерь от брака, % (Y_1)	Фондоотдача активной части основных фондов, ден. ед. (Y)
1	0,15	1,91	1	0,48	0,88
2	0,34	1,68	2	0,41	0,62
3	0,09	1,89	3	0,62	1,09
4	0,21	2,30	4	0,50	1,32
			5	1,20	0,68
–	$\overline{X}_1$ =0,198	$\overline{X}_2$ = 1,945	—	$\overline{Y}_1$ = 0,642	$\overline{Y}_2$ =0,918

1. На основании приведенных данных следует найти оценки векторов средних $\overline{X}$, $\overline{Y}$ и ковариационных матриц (S_x, S_y) а также оценку суммарной ковариационной матрицы (S_*) и обратной к ней (S_*^{-1}).

2. Рассчитайте вектор оценок коэффициентов дискриминантной функции (A) и определите ее средние значения для каждого множества. Определите константу дискриминации (c).

Вычислите оценки значений дискриминантной функции для предприятия, у которого переменные принимают значения: удельный вес потерь от брака (Z_1) равен 0,2%; фондоотдача активной части основных фондов (Z_2) равна 0,75 ден. ед.

Определите, к какой из двух групп следует отнести данное предприятие.

Рекомендуемая литература:

1. **Быковский, В. В.** Применение теории планирования эксперимента в научных и инженерных расчетах: учеб. пособие / В. В. Быковский, Л. В. Быковская, Ю. А. Дормидонов. - Оренбург : ОГУ, 2002. - 66 с - ISBN 5-7410-0442-3.

2. **Костин, В. Н.** Статистические методы и модели: учеб. пособие для вузов / В. Н. Костин, Н. А. Тишина . - Оренбург : ОГУ, 2004. - 138 с. - Библиогр.: с. 125. - ISBN 5-7410-0399-0.

Методические указания
к лабораторной работе №3

«Разработка программных средств решения задачи обработки информации, методом главных компонент»

1 Цель работы:

Изучение метода главных компонент для разработки программного средства

2 Задача работы:

Объяснение сути метода

3 Содержание работы:

Описание метода главных компонент для разработки программных средств

4 Требования к отчету:

Отчет о проделанной работе должен содержать:

– название работы, ее задачи и описание последовательности выполнения;
– решение задачи по вариантам, указанным преподавателем;

Отчет представить в распечатанном виде.

5 Общие положения и методика выполнения задания

Пусть имеется матрица исходных данных, представляющая совокупность четырех промышленных предприятий, оцененных по трем признакам: X_1 – уровень выработки на одного среднегодового работника; X_2 – уровень рентабельности продукции, %; X_3 – уровень фондоотдачи основных фондов, ден. ед. Обратим внимание, что в факторном анализе исходная матрица (X)строится как транспонированная

$$X^T = \begin{pmatrix} 9{,}26 & 9{,}38 & 12{,}11 & 10{,}81 \\ 13{,}26 & 10{,}16 & 13{,}72 & 12{,}85 \\ 1{,}45 & 1{,}30 & 1{,}37 & 1{,}65 \end{pmatrix}.$$

Используем сначала метод главных компонент, его реализация предусматривает решение следующих формальных уравнений:

$R = \dfrac{1}{n} ZZ^T$; $|R - \lambda E| = 0$; $(R - \lambda E)U = 0$; $A = V\Lambda^{1/2}$, с $V = \|V_j\|$ и $V_j = \dfrac{U_j}{|U_j|}$; $F = A^{-1}Z^T$.

В приведенных формулах $Z = \|z_{ij}\|$, $z_{ij} = \dfrac{x_{ij} - \bar{x}_j}{\sigma_j}$ и $X = X^T$ – транспонированная матрица исходных данных, размерности ($n \times m$).

Схематично алгоритм метода главных компонент имеет вид

$$\begin{tikzcd}
 & & & \Lambda \arrow[drr] & & & \\
X \arrow[r] & Z \arrow[r] & R(S) \arrow[ur] \arrow[dr] & & & A \arrow[r] & F \\
 & & & U \arrow[r] & V \arrow[ur] & &
\end{tikzcd}$$

1 Стандартизируем значения изучаемых признаков

$z_{ij} = \dfrac{x_{ij} - \bar{x}_j}{\sigma_j}$, получим $Z' = \begin{pmatrix} -0{,}971 & -0{,}868 & 1{,}478 & 0{,}361 \\ 0{,}549 & -1{,}684 & 0{,}882 & 0{,}253 \\ 0{,}076 & -1{,}069 & -0{,}534 & 1{,}527 \end{pmatrix}$.

2 Найдем матрицу парных корреляций $R = \frac{1}{n} Z^T Z$

$$R = \begin{pmatrix} 1 & 0{,}581 & 0{,}154 \\ 0{,}581 & 1 & 0{,}439 \\ 0{,}154 & 0{,}439 & 1 \end{pmatrix}.$$

3 Поиск матрицы собственных чисел в конечном счете означает решение характеристического уравнения $(R - \lambda E)V = 0$, откуда $|R - \lambda E| = 0$.

Существует множество различных подходов для вычисления элементов матрицы λ, остановимся на одном из них, широко применяемом на практике и использующем рекуррентные соотношения Фаддеева: пусть A – некоторая симметрическая матрица, размерностью $m \times m$, тогда ее определитель находим по следу матриц, производных от A:

$A_1 = A$	$P_1 = trA_1$	$B_1 = A_1 - P_1 E$
$A_2 = AB_1$	$P_2 = 1/2\, trA_2$	$B_2 = A_2 - P_2 E$
…	…	…
$A_{m-1} = AB_{m-2}$	$P_{m-1} = 1/(m-1)\, trA_{m-1}$	$B_{m-1} = A_{m-1} - P_{m-1} E$
$A_m = AB_{m-1}$	$P_m = 1/m \cdot trA_m$	$B_m = A_m - P_m E;\ B_m = 0$

На заключительном этапе расчетов $P_m = |A|$ можем записать характеристический многочлен

$$P_m(\lambda) = \lambda^m - P_1 \lambda^{m-1} - P_2 \lambda^{m-2} - \ldots - P_m.$$

Приравнивая характеристический многочлен нулю, можем получить его корни, т.е. множество значений λ.

В нашем случае: $R = A,\ \ A = A_1,\ \ P_1 = trA_1 = 1+1+1 = 3,$

$$B_1 = A_1 - P_1E = \begin{pmatrix} -2 & 0{,}581 & 0{,}154 \\ 0{,}581 & 2 & 0{,}439 \\ 0{,}154 & 0{,}439 & 2 \end{pmatrix};$$

$$A_2 = AB_1 = \begin{pmatrix} 1 & 0{,}581 & 0{,}154 \\ 0{,}581 & 1 & 0{,}439 \\ 0{,}154 & 0{,}439 & 1 \end{pmatrix}\begin{pmatrix} -2 & 0{,}581 & 0{,}154 \\ 0{,}581 & -2 & 0{,}439 \\ 0{,}154 & 0{,}439 & -2 \end{pmatrix} =$$

$$= \begin{pmatrix} -1{,}638 & -0{,}513 & 0{,}101 \\ -0{,}513 & -1{,}469 & -0{,}350 \\ 0{,}101 & -0{,}350 & -1{,}783 \end{pmatrix};$$

$$P_2 = 1/2\, trA_2 = 1/2\left[(-1{,}638) + (-1{,}469) + (-1{,}783)\right] = -2{,}445;$$

$$B_2 = A_2 - P_2E = \begin{pmatrix} 0{,}807 & -0{,}513 & 0{,}101 \\ -0{,}513 & 0{,}976 & -0{,}350 \\ 0{,}101 & -0{,}350 & 0{,}662 \end{pmatrix};$$

$$A_3 = AB_2 = \begin{pmatrix} 1 & 0{,}581 & 0{,}154 \\ 0{,}581 & 1 & 0{,}439 \\ 0{,}154 & 0{,}439 & 1 \end{pmatrix}\begin{pmatrix} 0{,}807 & -0{,}513 & 0{,}101 \\ -0{,}513 & 0{,}976 & -0{,}350 \\ 0{,}101 & -0{,}350 & 0{,}662 \end{pmatrix} =$$

$$= \begin{pmatrix} 0{,}524 & 0 & 0 \\ 0 & 0{,}524 & 0 \\ 0 & 0 & 0{,}524 \end{pmatrix};$$

$$P_3 = 1/3\,(3 \times 0{,}524) = 0{,}524\,; \quad B_3 = A_3 - P_3 E = 0\,.$$

Таким образом, $|R| = 0{,}524$ и характеристический многочлен будет $\lambda^3 - 3\lambda^2 + 2{,}445\lambda - 0{,}524 = 0$, откуда λ_1=1,798, λ_2=0,875, λ_3=0,327.

$$\lambda = \begin{pmatrix} 1{,}798 & 0 & 0 \\ 0 & 0{,}875 & 0 \\ 0 & 0 & 0{,}327 \end{pmatrix}.$$

4 Матрицу собственных векторов определим, решая системы линейных уравнений, каждому из собственных чисел (λ_j) соответственно. Так как подобные системы уравнений имеют бесконечное множество решений, каждый раз одному из неизвестных признаков будем задавать произвольное значение, например единицу:

для λ_1= 1,798 имеем

$$\left.\begin{array}{l} (1-1{,}798)u_{11} + 0{,}581u_{22} + 0{,}154u_{33} = 0 \\ 0{,}581u_{11} + (1-1{,}798)u_{21} + 0{,}439u_{31} = 0 \\ 0{,}154u_{11} + 0{,}439u_{21} + (1-1{,}798)u_{31} = 0 \end{array}\right| \quad \begin{array}{l} u_{11} = 1{,}262 \\ u_{21} = 1{,}469 \\ u_{31} = 1{,}000 \end{array}$$

для λ_2= 0,875

$$\left.\begin{array}{l} (1-0{,}875)u_{12} + 0{,}581u_{22} + 0{,}154u_{32} = 0 \\ 0{,}581u_{12} + (1-0{,}875)u_{22} + 0{,}439u_{32} = 0 \\ 0{,}154u_{12} + 0{,}439u_{22} + (1-0{,}875)u_{32} = 0 \end{array}\right| \quad \begin{array}{l} u_{12} = 1{,}307 \\ u_{22} = 1{,}779 \\ u_{32} = 1{,}000 \end{array}$$

для λ_3= 0,327

$$\left.\begin{array}{l} (1-0{,}327)u_{13} + 0{,}581u_{23} + 0{,}154u_{33} = 0 \end{array}\right| \quad \begin{array}{l} u_{13} = 1{,}307 \end{array}$$

$$0{,}581u_{13} + (1-0{,}327)u_{23} + 0{,}439u_{33} = 0 \qquad u_{23} = 1{,}779$$

$$0{,}154u_{13} + 0{,}39u_{23} + (1-0{,}327)u_{33} = 0 \qquad u_{33} = 1{,}000$$

Матрица собственных векторов принимает вид

$$\mathrm{U} = \begin{pmatrix} 1{,}262 & -0{,}144 & 1{,}307 \\ 1{,}469 & -0{,}234 & -1{,}779 \\ 1{,}000 & 1{,}000 & 1{,}000 \end{pmatrix}.$$

5 Поскольку в ходе расчетов собственные векторы приобретают различные шкалы измерения, следует привести их к нормируемому виду $\mathrm{V_j} = U_j / |U_j|$

$$\mathrm{V} = \begin{pmatrix} 0{,}579 & -0{,}139 & 0{,}539 \\ 0{,}674 & -0{,}225 & -0{,}734 \\ 0{,}459 & 0{,}964 & 0{,}413 \end{pmatrix}.$$

6 Теперь можем построить матрицу факторного отображения (A), $A = \mathrm{V}\lambda^{1/2}$

$$\mathrm{A} = \begin{pmatrix} 0{,}579 & -0{,}139 & 0{,}539 \\ 0{,}674 & -0{,}225 & -0{,}734 \\ 0{,}459 & 0{,}964 & 0{,}413 \end{pmatrix} \begin{pmatrix} \sqrt{1{,}798} & 0 & 0 \\ 0 & \sqrt{0{,}875} & 0 \\ 0 & 0 & \sqrt{0{,}327} \end{pmatrix} =$$

$$= \begin{array}{c|ccc} & \mathrm{F_1} & \mathrm{F_2} & \mathrm{F_3} \\ \hline X_1 & 0{,}776 & -0{,}130 & 0{,}308 \\ X_2 & 0{,}904 & -0{,}210 & -0{,}420 \\ X_3 & 0{,}616 & 0{,}902 & 0{,}236 \end{array}$$

Элементы матрицы A – это коэффициенты частной корреляции, характеризующие тесноту связей X_j – элементарных признаков с F_r – латентным факторами (главными компонентами).

Для матрицы A выполняется условие $A^T A = \lambda$, т.е. $\sum_j a_{jr}^2 = \lambda_r$, или $\sum_j a_{j1}^2 = 0{,}776^2 + 0{,}904^2 + 0{,}616^2 = 1{,}798$; $\sum_j a_{j2}^2 = 0{,}875$; $\sum_j a_{j3}^2 = 0{,}327$.

Матрица *А*, представляющая признаковую структуру каждой из главных компонент, позволяет определять названия главных компонент:

F_1 – объясняет (1,798/3=0,599) около 60% общей дисперсии элементарных признаков – может быть названа эффективностью производства;

F_2 – объясняет (0,875/3 = 0,292) около 30% общей дисперсии признаков – назовем ее «эффективность использования основных средств»;

F_3 – объясняет (0,327/3 = 0,109) около 11% общей дисперсии. Ввиду низкой значимости этой главной компоненты в дальнейшем анализе она может не рассматриваться.

С целью более точной оценки структуры каждой из главных компонент может применяться специальный коэффициент уровня информативности составляющих ее элементарных признаков

$$\mathrm{K}_u = \frac{\sum_j a_{jr}^2}{\sum_{j=1}^{m} a_{jr}^2},$$

где $\sum_{j=1}^{m} a_{jr}^2$ – сумма квадратов всех значений нагрузок элементарных признаков для главной компоненты F_r; $\sum_{j=1} a_{jr}^2$ – сумма квадратов нагрузок тех элементарных признаков, которые наиболее значимы и в основном формируют название главной компоненты (F_r).

Исчисленная матрица A позволяет записать уравнения связи элементарных признаков с главными компонентами:

$$\mathrm{Z}_1 = 0{,}776F_1 - 0{,}130F_2 + 0{,}308F_3;$$

$$Z_2 = 0{,}904F_1 - 0{,}210F_2 - 0{,}420F_3;$$

$$Z_3 = 0{,}616F_1 + 0{,}902F_2 + 0{,}236F_3;$$

и наоборот – главных компонент с элементарными признаками:

$$F_1 = 1/1{,}798(0{,}776Z_1 + 0{,}904Z_2 + 0{,}616Z_3);$$

$$F_1 = 1/0{,}875(-0{,}130Z_1 - 0{,}210Z_2 + 0{,}236Z_3).$$

7 Расчет матрицы F позволяет получить значения главных компонент по каждому наблюдаемому объекту $F = A^{-1}Z^T$:

$$F = \begin{pmatrix} 0{,}542 & 0{,}507 & 0{,}196 \\ -0{,}776 & -0{,}010 & 0{,}994 \\ 1{,}554 & -1{,}283 & -0{,}075 \end{pmatrix} \begin{pmatrix} -0{,}971 & -0{,}868 & 1{,}478 & 0{,}361 \\ 0{,}549 & -1{,}684 & 0{,}882 & 0{,}253 \\ 0{,}076 & -1{,}069 & -0{,}534 & 1{,}527 \end{pmatrix} =$$

	n_1	n_2	n_3	n_4
F_1	-0,233	-1,533	1,143	0,623
$= F_2$	0,823	-0,372	-1,687	1,236
F_3	-2,218	0,892	1,205	0,121

Проверка: $\sum f_{ri} = 0$

Теперь для обработки тех же исходных данных используем метод главных факторов. Построим редуцированную матрицу корреляций. С этой целью можно применить один из следующих подходов.

1. *Метод наибольшей корреляции*. На главной диагонали с положительным знаком записывается наибольший по величине коэффициент корреляции столбца.

2. *Метод Барта*. По каждому столбцу матрицы R вначале находят среднее значение коэффициентов корреляции (r_{ij}), затем, если оно относительно велико,

за общность принимается значение несколько выше наибольшего в столбце коэффициента корреляции, а если сравнительно мало, несколько меньше наибольшего в столбце коэффициента корреляции.

3. *Метод триад*. Для каждого столбца общность вычисляют по формуле

$$h_j^2 = \frac{r_{kj} r_{ij}}{r_{kl}}$$

где r_{kj}, r_{ij} – коэффициенты корреляции наибольшие в столбце.

4. *Метод малого центроида.* Для каждой j-й переменной строится корреляционная матрица, размерностью 4х4. Включая саму переменную, в эту матрицу записывают оценки корреляции трех других переменных, наиболее тесно связанных с первой. По данным малой матрицы корреляций (малого центроида) рассчитываются общности

$$h_j^2 = \frac{\left(\sum r_{il}\right)^2}{\sum\limits_{ij} r_{ij}},$$

где $\sum r_{il}$ – сумма элементов первого столбца; $\sum\limits_{ij} r_{ij}$ — сумма всех элементов малого центроида.

Используя простой метод Барта, построим для нашего примера редуцированную матрицу корреляций (R_h)

$$R_h = \begin{pmatrix} 0{,}670 & 0{,}581 & 0{,}154 \\ 0{,}581 & 0{,}630 & 0{,}439 \\ 0{,}154 & 0{,}439 & 0{,}420 \end{pmatrix}.$$

В дальнейшем определение собственных чисел и собственных векторов может осуществляться выполнением уже продемонстрированных ранее шагов алгоритма метода главных компонент или применением подхода Хотеллинга, предусматривающего процедуру многократного возведения в квадрат исходной матрицы (R_h) Выберем подход Хотеллинга.

Для имеющейся у нас редуцированной матрицы выполним операцию возведения в степень $R^2 = R^T R$. Процедура должна повторяться до тех пор, пока некоторые величины (α-оценки матрицы *R)* до и после возведения в степень не перестанут существенно различаться, т.е. $d = (a_{(i)} - a_{(i-1)})$ должны быть минимальны, меньше некоторого заранее заданного порогового уровня. Оценки *a* – это приближения факторного отображения, $a = p_i / p_{\max}$, где p – скаляр ($p_i^{i+1} = R_h^1 S^1$), соответствующий величине суммы коэффициентов корреляции по каждой строке ($S = \sum_j r_{ij}$) значения величин p и S взаимно контролируются.

Матрица R_h будем возводить в степень, одновременно вычисляя оценки *a,* p и S, результаты расчетов сведем в таблицы 1.1 – 1.4.

После первого возведения в квадрат матрицы R_h значения α-характеристики еще достаточно велики. Продолжим операцию умножения матриц. Так как мы всякий раз имеем дело с симметрической матрицей, расчеты можно значительно сократить, вычисляя только элементы по главной диагонали и над главной диагональю. Кроме того, несколько шагов возведения в квадрат можно пропустить, используя матрицы исходных данных во второй, четвертой и т.д. степени (таблица 1.3).

Таблица 1.1

Исходная редуцированная корреляционная матрица

Признак	R_h				
	X_1	X_2	X_3	$S_i^{(1)} = \sum_j r_{ij}$	$a^{(1)} = \frac{S_i^{(1)}}{S_{\max}}$
X_1	0,670	0,581	0,154	1,405	0,851
X_2	0,581	0,630	0,439	1,650	1,000
X_3	0,154	0,439	0,420	1,013	0,614

Таблица 1.2

Первый цикл итерации – возведение в квадрат корреляционной матрицы

Признак	$R_h{}^2 = R_h{}' R_h$			$S^{(2)} = \sum_j r_{ij}$	$p_i^{(2)} = R_h S^{(1)}$	$a_i^{(2)} = p_i / p_{\max}$	$d = \left\| a^{(2)} - a^{(1)} \right\|$
	X_1	X_2	X_3				
X_1	0,811	0,823	0,423	2,057	2,056	0,894	0,043
X_2	0,823	0,928	0,550	2,301	2,300	1,000	0,000
X_3	0,423	0,550	0,393	1,366	1,365	0,593	0,021

Таблица 1.3

Второй цикл итерации – корреляционная матрица в четвертой степени

Признак	$R_h{}^4 = R_h{}^2 R_h{}^2$			$S_i^{(3)}$	$p_i^{(3)} = R_h^2 S^{(2)}$	$a^{(3)} = p_i / p_{\max}$	$d = \left\| a^{(3)} - a^{(2)} \right\|$
	X_1	X_2	X_3				
X_1	1,514	1,664	0,962	4,140	4,140	0,904	0,010
X_2	1,664	1,836	1,074	4,574	4,579	1,000	0,000
X_3	0,962	1,074	0,635	2,671	2,637	0,585	0,008

После возведения корреляционной матрицы в четвертую степень α –разности резко уменьшились. Вычислим $R_h{}^8$ и завершим итерацию.

Таблица 1.3

Второй цикл итерации – корреляционная матрица в четвертой степени

Признак	$R_h^{\ 8} = R_h^{\ 4} R_h^{\ 4}$			$S_i^{(4)}$	$p_i^{(4)} = R_h^4 S^{(3)}$	$a^{(4)} = p_i / p_{\max}$	$d = \left\lvert a^{(4)} - a^{(3)} \right\rvert$
	X_1	X_2	X_3				
X_1	5,986	6,607	3,854	16,447	16,448	0,906	0,002
X_2	6,607	7,293	4,255	18,155	18,156	1,000	0,000
X_3	3,854	4,255	2,481	10,590	10,591	0,583	0,002

Оценки p и S подтверждают правильность проведенных вычислений, их максимальное расхождение после выполнения трех циклов итерации не превысило пяти тысячных. Таким образом, оценки компонент первого собственного вектора можно считать достоверными. Собственный вектор — это ненормированные значения $a^{(4)}$, т.е. $a_1^{(4)} = U_1$.

Перейдем к определению нагрузок первого главного фактора (таблица 1.5).

Таблица 1.5

Вычисление нагрузок первого главного фактора

Признак	$a_1^{(4)} = U^*{}_1$	$\beta_1 = R_h a_1^{(4)}$	$A = \dfrac{U_1 \sqrt{\lambda_1}}{\left(\sum U_{1i}^2\right)^{1/2}}$
X_1	0,906	1,278	0,732
X_2	1,000	1,412	0,808
X_3	0,583	0,823	0,471

* Данные представлены в таблице 1.4

В таблице 1.5 собственное число λ_1 – это наибольшая величина вектора β_1. Компоненты a_{j1} легко рассчитываются по известному λ_1 =1,412

$$\frac{U_1\sqrt{\lambda_1}}{\left(\sum U_{1i}^2\right)^{1/2}} = \frac{\sqrt{1{,}412}}{(0{,}821+1{,}000+0{,}340)^{1/2}} = 0{,}808.$$

Итогами первой итерации будут первое собственное число λ_1 =1,412 и вектор факторных нагрузок $A_1' = (0{,}732 \quad 0{,}808 \quad 0{,}471)$. Проверим выполнение требования

$\sum a_{j1}^2 = \lambda_1$, или $0{,}732^2 + 0{,}808^2 + 0{,}471^2 = 1{,}412$.

Остается определить воспроизведенную матрицу парных корреляций (R_h^+) и решить вопрос о необходимости выполнения второй итерации с поиском второго собственного числа (λ_2) и вектора факторных нагрузок A_2.

Воспроизведенная корреляционная матрица только по одному, первому, вектору факторного отображения будет $R_h^+ = AA^T$

$$R_h^+ = \begin{pmatrix} 0{,}732 \\ 0{,}808 \\ 0{,}471 \end{pmatrix} \times \begin{pmatrix} 0{,}732 & 0{,}808 & 0{,}471 \end{pmatrix} = \begin{pmatrix} 0{,}536 & 0{,}591 & 0{,}345 \\ 0{,}591 & 0{,}653 & 0{,}381 \\ 0{,}354 & 0{,}381 & 0{,}222 \end{pmatrix}.$$

Разность матриц $R_h - R_h^+$ покажет остаточную, не объясненную первым главным фактором, корреляцию и поможет ответить на вопрос о целесообразности выделения второго главного фактора

$$R_1 = R_h - R_h^+ = \begin{pmatrix} 0{,}670 & 0{,}581 & 0{,}154 \\ 0{,}581 & 0{,}630 & 0{,}439 \\ 0{,}154 & 0{,}439 & 0{,}420 \end{pmatrix} - \begin{pmatrix} 0{,}536 & 0{,}591 & 0{,}345 \\ 0{,}591 & 0{,}653 & 0{,}381 \\ 0{,}354 & 0{,}381 & 0{,}222 \end{pmatrix} =$$

$$= \begin{pmatrix} 0{,}134 & -0{,}010 & -0{,}191 \\ -0{,}010 & -0{,}023 & 0{,}058 \\ -0{,}191 & 0{,}058 & 0{,}198 \end{pmatrix}$$

Матрица первых остаточных коэффициентов корреляции содержит еще достаточно большие величины и вполне допускает оценку второго главного фактора. Последующее выполнение второй итерации аналогично первой, только вычисления производятся на данных матрицы остатков R_1.

В таблице 1.6 получены более грубые оценки элементов собственного вектора, чем в первом случае (таблица 1.5). Тем не менее, итерация прервана с учетом того, что элементы квадрируемой матрицы резко снижают свою значимость и теряют наглядность, в то же время заданная условность примера допускает различную степень приближения аналитических результатов, вычисляемых по уже знакомому алгоритму.

Таблица 1.6

Матрица первых остаточных коэффициентов корреляции

Признак	R_1			$S^{(1)}$	$a^{(1)}$
	X_1	X_2	X_3		
X_1	0,134	-0,010	-0,191	-0,067	-1,000
X_2	-0,010	-0,023	0,058	0,025	0,373
X_3	-0,191	0,058	0,198	0,065	0,970

По данным таблицы 1.7 определим нагрузки второго главного фактора (таблица 1.8)

Таблица 1.7

Первый и второй циклы итерации для матрицы первых остаточных коэффициентов корреляции

Признак	R_1			$S^{(3)}$	$p^{(3)}$	$a^{(3)}$	$d=\left\|a^{(3)}-a^{(2)}\right\|$
	X_1	X_2	X_3				
X_1	0,00723	-0,00149	-0,00870	-0,00296	-0,00297	-0,830	0,036
X_2	-0,00149	0,00032	0,00179	0,00061	0,00061	0,170	0,027
X_3	-0,00870	0,00179	0,01051	0,00360	0,00358	1,000	0,000

Таблица 1.8

Вычисление нагрузок второго главного фактора

Признак	$a_2^{(3)}=U_2$	$\beta_2=R_1a_2^{(3)}$	$A_2=\dfrac{U_2\sqrt{\lambda_{22}}}{\left(\sum U_{2i}^2\right)^{1/2}}$
X_1	-0,830	-0,304	-0,383
X_2	0,170	0,062	0,079
X_3	1,000	0,366	0,462
–	$\sum_i U_{2i}^2$ =1,718	λ_2=0,366	$\sum_j a_{jr}^2$ =0,366

Матрицу вторых остаточных коэффициентов корреляции (R_2) находят из разности $R_1-A_2\times A_2^T$.

Элементы матрицы R_2 имеют малые значения, и выделение третьего главного фактора не целесообразно. Такой же вывод следует и по данным собственных чисел λ, а именно: при помощи первого и второго главных факторов полностью воспроизведена общность (trR_h =1,720) и почти на 60% удалось объяснить вариацию элементарных признаков X_1, X_2, X_3 $\left(\frac{1{,}412+0{,}366}{3}\right)$

$$R_2=\begin{pmatrix}0{,}134 & -0{,}010 & -0{,}191\\ -0{,}010 & -0{,}023 & 0{,}058\\ -0{,}191 & 0{,}058 & 0{,}198\end{pmatrix}-\begin{pmatrix}-0{,}383\\ 0{,}079\\ 0{,}462\end{pmatrix}\cdot\begin{pmatrix}-0{,}383 & 0{,}079 & 0{,}462\end{pmatrix}=\begin{pmatrix}-0{,}013 & 0{,}020 & -0{,}014\\ 0{,}020 & -0{,}029 & 0{,}022\\ 0{,}014 & 0{,}022 & -0{,}015\end{pmatrix}.$$

В заключение обобщим итоги решения задачи методом главных факторов, выделяя общности и характерности для каждого из элементарных X_{j} признаков (таблица 1.9).

Таблица 1.9

Факторные нагрузки, общности и характерности,

вычисленные методом главных факторов

<table>
<tr><td rowspan="2">Признак</td><td colspan="2">Главный фактор (факторные нагрузки)</td><td rowspan="2">Общность</td><td rowspan="2">Характерность</td></tr>
<tr><td>F_1</td><td>F_2</td></tr>
<tr><td>X_1</td><td>0,732</td><td>-0,383</td><td>0,683 (0,670)</td><td>0,317</td></tr>
<tr><td>X_2</td><td>0,808</td><td>0,079</td><td>0,660 (0,630)</td><td>0,340</td></tr>
<tr><td>X_3</td><td>0,471</td><td>0,462</td><td>0,435 (0,420)</td><td>0,565</td></tr>
<tr><td>–</td><td>$\sum_j a_{jr}^2$ =1,412</td><td>$\sum_j a_{j2}^2$ =0,366</td><td>$\sum h_j^2$ =1,778</td><td>$\sum d_j^2$ =1,222</td></tr>
</table>

В таблице 1.9 уровни общностей несколько отличаются от данных исходной матрицы R_h (приведенных в скобках). Это могло произойти из-за допускаемой грубости итеративных решений и округлений. В скобках приведены первоначально принятые величины общностей для каждого признака. Общности и характерности совместно полностью представляют дисперсии признаков, равные единице

$$\sum h_j^2 + \sum d_j^2 = 3 .$$

В дальнейшем может быть найдена матрица значений главных факторов

$$F = (A^T A)^{-1} A^T Z' .$$

Рекомендуемая литература:

1. **Быковский, В. В.** Применение теории планирования эксперимента в научных и инженерных расчетах: учеб. пособие / В. В. Быковский, Л. В. Быковская, Ю. А. Дормидонов. - Оренбург : ОГУ, 2002. - 66 с - ISBN 5-7410-0442-3.

2. **Костин, В. Н.** Статистические методы и модели: учеб. пособие для вузов / В. Н. Костин, Н. А. Тишина . - Оренбург : ОГУ, 2004. - 138 с. - Библиогр.: с. 125. - ISBN 5-7410-0399-0.

Методические указания
к лабораторной работе №4

«Разработка программных средств решения задачи обработки информации, методом факторного анализа»

1 Цель работы:

Изучение метода факторного анализа для разработки программного средства

2 Задача работы:

Объяснение сути метода

3 Содержание работы:

Описание метода факторного анализа для разработки программных средств

4 Требования к отчету:

Отчет о проделанной работе должен содержать:

– название работы, ее задачи и описание последовательности выполнения;
– решение задачи по вариантам, указанным преподавателем;

Отчет представить в распечатанном виде.

5 Общие положения и методика выполнения задания

Пусть имеется матрица исходных данных, представляющая совокупность четырех промышленных предприятий, оцененных по трем признакам: X_1 – уровень выработки на одного среднегодового работника; X_2 – уровень рентабельности продукции, %; X_3 – уровень фондоотдачи основных фондов, ден. ед. Обратим внимание, что в факторном анализе исходная матрица (X)строится как транспонированная

$$X^T = \begin{pmatrix} 9{,}26 & 9{,}38 & 12{,}11 & 10{,}81 \\ 13{,}26 & 10{,}16 & 13{,}72 & 12{,}85 \\ 1{,}45 & 1{,}30 & 1{,}37 & 1{,}65 \end{pmatrix}.$$

Используем сначала метод главных компонент, его реализация предусматривает решение следующих формальных уравнений:

$$R = \frac{1}{n} ZZ^T; \quad |R - \lambda E| = 0; \quad (R - \lambda E)U = 0; \qquad A = V\Lambda^{1/2}, \text{ с } V = \|V_j\| \text{ и } V_j = \frac{U_j}{|U_j|}; \quad F = A^{-1}Z^T.$$

В приведенных формулах $Z = \|z_{ij}\|$, $z_{ij} = \dfrac{x_{ij} - \bar{x}_j}{\sigma_j}$ и $X = X^T$ – транспонированная матрица исходных данных, размерности ($n \times m$).

Схематично алгоритм метода главных компонент имеет вид

$$\begin{tikzcd}
 & & & \Lambda \arrow[drr] & & & \\
X \arrow[r] & Z \arrow[r] & R(S) \arrow[ur] \arrow[dr] & & & A \arrow[r] & F \\
 & & & U \arrow[r] & V \arrow[ur] & &
\end{tikzcd}$$

1 Стандартизируем значения изучаемых признаков

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{\sigma_j}, \text{ получим } Z' = \begin{pmatrix} -0{,}971 & -0{,}868 & 1{,}478 & 0{,}361 \\ 0{,}549 & -1{,}684 & 0{,}882 & 0{,}253 \\ 0{,}076 & -1{,}069 & -0{,}534 & 1{,}527 \end{pmatrix}.$$

2 Найдем матрицу парных корреляций $R = \frac{1}{n} Z^T Z$

$$R = \begin{pmatrix} 1 & 0{,}581 & 0{,}154 \\ 0{,}581 & 1 & 0{,}439 \\ 0{,}154 & 0{,}439 & 1 \end{pmatrix}.$$

3 Поиск матрицы собственных чисел в конечном счете означает решение характеристического уравнения $(R - \lambda E)V = 0$, откуда $|R - \lambda E| = 0$.

Существует множество различных подходов для вычисления элементов матрицы λ, остановимся на одном из них, широко применяемом на практике и использующем рекуррентные соотношения Фаддеева: пусть A – некоторая симметрическая матрица, размерностью $m \times m$, тогда ее определитель находим по следу матриц, производных от A:

$A_1 = A$	$P_1 = trA_1$	$B_1 = A_1 - P_1 E$
$A_2 = AB_1$	$P_2 = 1/2\, trA_2$	$B_2 = A_2 - P_2 E$
…	…	…
$A_{m-1} = AB_{m-2}$	$P_{m-1} = 1/(m-1)\, trA_{m-1}$	$B_{m-1} = A_{m-1} - P_{m-1} E$
$A_m = AB_{m-1}$	$P_m = 1/m \cdot trA_m$	$B_m = A_m - P_m E;\ B_m = 0$

На заключительном этапе расчетов $P_m = |A|$ можем записать характеристический многочлен

$$P_m(\lambda) = \lambda^m - P_1 \lambda^{m-1} - P_2 \lambda^{m-2} - \ldots - P_m .$$

Приравнивая характеристический многочлен нулю, можем получить его корни, т.е. множество значений λ.

В нашем случае: $R = A,\ \ A = A_1,\ \ P_1 = trA_1 = 1+1+1 = 3,$

$$B_1 = A_1 - P_1E = \begin{pmatrix} -2 & 0{,}581 & 0{,}154 \\ 0{,}581 & 2 & 0{,}439 \\ 0{,}154 & 0{,}439 & 2 \end{pmatrix};$$

$$A_2 = AB_1 = \begin{pmatrix} 1 & 0{,}581 & 0{,}154 \\ 0{,}581 & 1 & 0{,}439 \\ 0{,}154 & 0{,}439 & 1 \end{pmatrix}\begin{pmatrix} -2 & 0{,}581 & 0{,}154 \\ 0{,}581 & -2 & 0{,}439 \\ 0{,}154 & 0{,}439 & -2 \end{pmatrix} =$$

$$= \begin{pmatrix} -1{,}638 & -0{,}513 & 0{,}101 \\ -0{,}513 & -1{,}469 & -0{,}350 \\ 0{,}101 & -0{,}350 & -1{,}783 \end{pmatrix};$$

$$P_2 = 1/2\,trA_2 = 1/2\left[(-1{,}638) + (-1{,}469) + (-1{,}783)\right] = -2{,}445;$$

$$B_2 = A_2 - P_2E = \begin{pmatrix} 0{,}807 & -0{,}513 & 0{,}101 \\ -0{,}513 & 0{,}976 & -0{,}350 \\ 0{,}101 & -0{,}350 & 0{,}662 \end{pmatrix};$$

$$A_3 = AB_2 = \begin{pmatrix} 1 & 0{,}581 & 0{,}154 \\ 0{,}581 & 1 & 0{,}439 \\ 0{,}154 & 0{,}439 & 1 \end{pmatrix}\begin{pmatrix} 0{,}807 & -0{,}513 & 0{,}101 \\ -0{,}513 & 0{,}976 & -0{,}350 \\ 0{,}101 & -0{,}350 & 0{,}662 \end{pmatrix} =$$

$$= \begin{pmatrix} 0{,}524 & 0 & 0 \\ 0 & 0{,}524 & 0 \\ 0 & 0 & 0{,}524 \end{pmatrix};$$

$$P_3 = 1/3\,(3 \times 0{,}524) = 0{,}524\,; \quad B_3 = A_3 - P_3 E = 0\,.$$

Таким образом, $|R| = 0{,}524$ и характеристический многочлен будет $\lambda^3 - 3\lambda^2 + 2{,}445\lambda - 0{,}524 = 0$, откуда λ_1=1,798, λ_2=0,875, λ_3=0,327.

$$\lambda = \begin{pmatrix} 1{,}798 & 0 & 0 \\ 0 & 0{,}875 & 0 \\ 0 & 0 & 0{,}327 \end{pmatrix}.$$

4 Матрицу собственных векторов определим, решая системы линейных уравнений, каждому из собственных чисел (λ_j) соответственно. Так как подобные системы уравнений имеют бесконечное множество решений, каждый раз одному из неизвестных признаков будем задавать произвольное значение, например единицу:

для λ_1 = 1,798 имеем

$(1-1{,}798)u_{11} + 0{,}581u_{22} + 0{,}154u_{33} = 0$	$u_{11} = 1{,}262$
$0{,}581u_{11} + (1-1{,}798)u_{21} + 0{,}439u_{31} = 0$	$u_{21} = 1{,}469$
$0{,}154u_{11} + 0{,}439u_{21} + (1-1{,}798)u_{31} = 0$	$u_{31} = 1{,}000$

для λ_2 = 0,875

$(1-0{,}875)u_{12} + 0{,}581u_{22} + 0{,}154u_{32} = 0$	$u_{12} = 1{,}307$
$0{,}581u_{12} + (1-0{,}875)u_{22} + 0{,}439u_{32} = 0$	$u_{22} = 1{,}779$
$0{,}154u_{12} + 0{,}439u_{22} + (1-0{,}875)u_{32} = 0$	$u_{32} = 1{,}000$

для λ_3 = 0,327

$(1-0{,}327)u_{13} + 0{,}581u_{23} + 0{,}154u_{33} = 0$	$u_{13} = 1{,}307$

$$0{,}581u_{13}+(1-0{,}327)u_{23}+0{,}439u_{33}=0 \qquad u_{23}=1{,}779$$

$$0{,}154u_{13}+0{,}39u_{23}+(1-0{,}327)u_{33}=0 \qquad u_{33}=1{,}000$$

Матрица собственных векторов принимает вид

$$\mathrm{U}=\begin{pmatrix}1{,}262 & -0{,}144 & 1{,}307\\ 1{,}469 & -0{,}234 & -1{,}779\\ 1{,}000 & 1{,}000 & 1{,}000\end{pmatrix}.$$

5 Поскольку в ходе расчетов собственные векторы приобретают различные шкалы измерения, следует привести их к нормируемому виду $\mathrm{V_j}=U_j/|U_j|$

$$\mathrm{V}=\begin{pmatrix}0{,}579 & -0{,}139 & 0{,}539\\ 0{,}674 & -0{,}225 & -0{,}734\\ 0{,}459 & 0{,}964 & 0{,}413\end{pmatrix}.$$

6 Теперь можем построить матрицу факторного отображения (A), $A=\mathrm{V}\lambda^{1/2}$

$$\mathrm{A}=\begin{pmatrix}0{,}579 & -0{,}139 & 0{,}539\\ 0{,}674 & -0{,}225 & -0{,}734\\ 0{,}459 & 0{,}964 & 0{,}413\end{pmatrix}\begin{pmatrix}\sqrt{1{,}798} & 0 & 0\\ 0 & \sqrt{0{,}875} & 0\\ 0 & 0 & \sqrt{0{,}327}\end{pmatrix}=$$

$$=\begin{array}{c|ccc} & \mathrm{F_1} & \mathrm{F_2} & \mathrm{F_3}\\ \hline X_1 & 0{,}776 & -0{,}130 & 0{,}308\\ X_2 & 0{,}904 & -0{,}210 & -0{,}420\\ X_3 & 0{,}616 & 0{,}902 & 0{,}236\end{array}$$

Элементы матрицы A – это коэффициенты частной корреляции, характеризующие тесноту связей X_j – элементарных признаков с F_r – латентным факторами (главными компонентами).

Для матрицы A выполняется условие $A^T A = \lambda$, т.е. $\sum_j a_{jr}^2 = \lambda_r$, или $\sum_j a_{j1}^2 = 0{,}776^2 + 0{,}904^2 + 0{,}616^2 = 1{,}798$; $\sum_j a_{j2}^2 = 0{,}875$; $\sum_j a_{j3}^2 = 0{,}327$.

Матрица *А*, представляющая признаковую структуру каждой из главных компонент, позволяет определять названия главных компонент:

F_1 – объясняет (1,798/3=0,599) около 60% общей дисперсии элементарных признаков – может быть названа эффективностью производства;

F_2 – объясняет (0,875/3 = 0,292) около 30% общей дисперсии признаков – назовем ее «эффективность использования основных средств»;

F_3 – объясняет (0,327/3 = 0,109) около 11% общей дисперсии. Ввиду низкой значимости этой главной компоненты в дальнейшем анализе она может не рассматриваться.

С целью более точной оценки структуры каждой из главных компонент может применяться специальный коэффициент уровня информативности составляющих ее элементарных признаков

$$K_u = \frac{\sum_j a_{jr}^2}{\sum_{j=1}^{m} a_{jr}^2},$$

где $\sum_{j=1}^{m} a_{jr}^2$ – сумма квадратов всех значений нагрузок элементарных признаков для главной компоненты F_r; $\sum_{j=1} a_{jr}^2$ – сумма квадратов нагрузок тех элементарных признаков, которые наиболее значимы и в основном формируют название главной компоненты (F_r).

Исчисленная матрица A позволяет записать уравнения связи элементарных признаков с главными компонентами:

$$Z_1 = 0{,}776F_1 - 0{,}130F_2 + 0{,}308F_3;$$

$$Z_2 = 0{,}904F_1 - 0{,}210F_2 - 0{,}420F_3;$$

$$Z_3 = 0{,}616F_1 + 0{,}902F_2 + 0{,}236F_3;$$

и наоборот – главных компонент с элементарными признаками:

$$F_1 = 1/1{,}798(0{,}776Z_1 + 0{,}904Z_2 + 0{,}616Z_3);$$

$$F_1 = 1/0{,}875(-0{,}130Z_1 - 0{,}210Z_2 + 0{,}236Z_3).$$

7 Расчет матрицы F позволяет получить значения главных компонент по каждому наблюдаемому объекту $F = A^{-1}Z^T$:

$$F = \begin{pmatrix} 0{,}542 & 0{,}507 & 0{,}196 \\ -0{,}776 & -0{,}010 & 0{,}994 \\ 1{,}554 & -1{,}283 & -0{,}075 \end{pmatrix} \begin{pmatrix} -0{,}971 & -0{,}868 & 1{,}478 & 0{,}361 \\ 0{,}549 & -1{,}684 & 0{,}882 & 0{,}253 \\ 0{,}076 & -1{,}069 & -0{,}534 & 1{,}527 \end{pmatrix} =$$

	n_1	n_2	n_3	n_4
F_1	-0,233	-1,533	1,143	0,623
$= F_2$	0,823	-0,372	-1,687	1,236
F_3	-2,218	0,892	1,205	0,121

Проверка: $\sum f_{ri} = 0$

Теперь для обработки тех же исходных данных используем метод главных факторов. Построим редуцированную матрицу корреляций. С этой целью можно применить один из следующих подходов.

5. *Метод наибольшей корреляции*. На главной диагонали с положительным знаком записывается наибольший по величине коэффициент корреляции столбца.

6. *Метод Барта*. По каждому столбцу матрицы R вначале находят среднее значение коэффициентов корреляции (r_{ij}), затем, если оно относительно велико,

за общность принимается значение несколько выше наибольшего в столбце коэффициента корреляции, а если сравнительно мало, несколько меньше наибольшего в столбце коэффициента корреляции.

7. *Метод триад*. Для каждого столбца общность вычисляют по формуле

$$h_j^2 = \frac{r_{kj} r_{ij}}{r_{kl}}$$

где r_{kj}, r_{ij} – коэффициенты корреляции наибольшие в столбце.

8. *Метод малого центроида.* Для каждой j-й переменной строится корреляционная матрица, размерностью 4х4. Включая саму переменную, в эту матрицу записывают оценки корреляции трех других переменных, наиболее тесно связанных с первой. По данным малой матрицы корреляций (малого центроида) рассчитываются общности

$$h_j^2 = \frac{\left(\sum r_{il}\right)^2}{\sum\limits_{ij} r_{ij}},$$

где $\sum r_{il}$ – сумма элементов первого столбца; $\sum\limits_{ij} r_{ij}$ — сумма всех элементов малого центроида.

Используя простой метод Барта, построим для нашего примера редуцированную матрицу корреляций (R_h)

$$R_h = \begin{pmatrix} 0{,}670 & 0{,}581 & 0{,}154 \\ 0{,}581 & 0{,}630 & 0{,}439 \\ 0{,}154 & 0{,}439 & 0{,}420 \end{pmatrix}.$$

В дальнейшем определение собственных чисел и собственных векторов может осуществляться выполнением уже продемонстрированных ранее шагов алгоритма метода главных компонент или применением подхода Хотеллинга, предусматривающего процедуру многократного возведения в квадрат исходной матрицы (R_h) Выберем подход Хотеллинга.

Для имеющейся у нас редуцированной матрицы выполним операцию возведения в степень $R^2 = R^T R$. Процедура должна повторяться до тех пор, пока некоторые величины (α-оценки матрицы *R)* до и после возведения в степень не перестанут существенно различаться, т.е. $d = (a_{(i)} - a_{(i-1)})$ должны быть минимальны, меньше некоторого заранее заданного порогового уровня. Оценки *a* – это приближения факторного отображения, $a = p_i / p_{\max}$, где p – скаляр ($p_i^{i+1} = R_h^1 S^1$), соответствующий величине суммы коэффициентов корреляции по каждой строке ($S = \sum_j r_{ij}$) значения величин p и S взаимно контролируются.

Матрица R_h будем возводить в степень, одновременно вычисляя оценки *a,* p и S, результаты расчетов сведем в таблицы 1.1 – 1.4.

После первого возведения в квадрат матрицы R_h значения α-характеристики еще достаточно велики. Продолжим операцию умножения матриц. Так как мы всякий раз имеем дело с симметрической матрицей, расчеты можно значительно сократить, вычисляя только элементы по главной диагонали и над главной диагональю. Кроме того, несколько шагов возведения в квадрат можно пропустить, используя матрицы исходных данных во второй, четвертой и т.д. степени (таблица 1.3).

Таблица 1.1

Исходная редуцированная корреляционная матрица

Признак	R_h				
	X_1	X_2	X_3	$S_i^{(1)} = \sum_j r_{ij}$	$a^{(1)} = \frac{S_i^{(1)}}{S_{\max}}$
X_1	0,670	0,581	0,154	1,405	0,851
X_2	0,581	0,630	0,439	1,650	1,000
X_3	0,154	0,439	0,420	1,013	0,614

Таблица 1.2

Первый цикл итерации – возведение в квадрат корреляционной матрицы

Признак	$R_h{}^2 = R_h{}' R_h$			$S^{(2)} = \sum_j r_{ij}$	$p_i^{(2)} = R_h S^{(1)}$	$a_i^{(2)} = p_i / p_{\max}$	$d = \lvert a^{(2)} - a^{(1)} \rvert$
	X_1	X_2	X_3				
X_1	0,811	0,823	0,423	2,057	2,056	0,894	0,043
X_2	0,823	0,928	0,550	2,301	2,300	1,000	0,000
X_3	0,423	0,550	0,393	1,366	1,365	0,593	0,021

Таблица 1.3

Второй цикл итерации – корреляционная матрица в четвертой степени

Признак	$R_h{}^4 = R_h{}^2 R_h{}^2$			$S_i^{(3)}$	$p_i^{(3)} = R_h^2 S^{(2)}$	$a^{(3)} = p_i / p_{\max}$	$d = \lvert a^{(3)} - a^{(2)} \rvert$
	X_1	X_2	X_3				
X_1	1,514	1,664	0,962	4,140	4,140	0,904	0,010
X_2	1,664	1,836	1,074	4,574	4,579	1,000	0,000
X_3	0,962	1,074	0,635	2,671	2,637	0,585	0,008

После возведения корреляционной матрицы в четвертую степень α –разности резко уменьшились. Вычислим $R_h{}^8$ и завершим итерацию.

Таблица 1.3

Второй цикл итерации – корреляционная матрица в четвертой степени

Признак	$R_h^{\ 8} = R_h^{\ 4} R_h^{\ 4}$			$S_i^{(4)}$	$p_i^{(4)} = R_h^4 S^{(3)}$	$a^{(4)} = p_i / p_{\max}$	$d = \left\|a^{(4)} - a^{(3)}\right\|$
	X_1	X_2	X_3				
X_1	5,986	6,607	3,854	16,447	16,448	0,906	0,002
X_2	6,607	7,293	4,255	18,155	18,156	1,000	0,000
X_3	3,854	4,255	2,481	10,590	10,591	0,583	0,002

Оценки p и S подтверждают правильность проведенных вычислений, их максимальное расхождение после выполнения трех циклов итерации не превысило пяти тысячных. Таким образом, оценки компонент первого собственного вектора можно считать достоверными. Собственный вектор — это ненормированные значения $a^{(4)}$, т.е. $a_1^{(4)} = U_1$.

Перейдем к определению нагрузок первого главного фактора (таблица 1.5).

Таблица 1.5

Вычисление нагрузок первого главного фактора

Признак	$a_1^{(4)} = U^*{}_1$	$\beta_1 = R_h a_1^{(4)}$	$A = \dfrac{U_1 \sqrt{\lambda_1}}{\left(\sum U_{1i}^2\right)^{1/2}}$
X_1	0,906	1,278	0,732
X_2	1,000	1,412	0,808
X_3	0,583	0,823	0,471

* Данные представлены в таблице 1.4

В таблице 1.5 собственное число λ_1 – это наибольшая величина вектора β_1. Компоненты a_{j1} легко рассчитываются по известному λ_1=1,412

$$\frac{U_1\sqrt{\lambda_1}}{\left(\sum U_{1i}^2\right)^{1/2}} = \frac{\sqrt{1,412}}{(0,821+1,000+0,340)^{1/2}} = 0,808.$$

Итогами первой итерации будут первое собственное число λ_1=1,412 и вектор факторных нагрузок $A_1' = (0,732 \quad 0,808 \quad 0,471)$. Проверим выполнение требования

$\sum a_{j1}^2 = \lambda_1$, или $0,732^2 + 0,808^2 + 0,471^2 = 1,412$.

Остается определить воспроизведенную матрицу парных корреляций (R_h^+) и решить вопрос о необходимости выполнения второй итерации с поиском второго собственного числа (λ_2) и вектора факторных нагрузок A_2.

Воспроизведенная корреляционная матрица только по одному, первому, вектору факторного отображения будет $R_h^+ = AA^T$

$$R_h^+ = \begin{pmatrix} 0,732 \\ 0,808 \\ 0,471 \end{pmatrix} \times \begin{pmatrix} 0,732 & 0,808 & 0,471 \end{pmatrix} = \begin{pmatrix} 0,536 & 0,591 & 0,345 \\ 0,591 & 0,653 & 0,381 \\ 0,354 & 0,381 & 0,222 \end{pmatrix}.$$

Разность матриц $R_h - R_h^+$ покажет остаточную, не объясненную первым главным фактором, корреляцию и поможет ответить на вопрос о целесообразности выделения второго главного фактора

$$R_1 = R_h - R_h^+ = \begin{pmatrix} 0,670 & 0,581 & 0,154 \\ 0,581 & 0,630 & 0,439 \\ 0,154 & 0,439 & 0,420 \end{pmatrix} - \begin{pmatrix} 0,536 & 0,591 & 0,345 \\ 0,591 & 0,653 & 0,381 \\ 0,354 & 0,381 & 0,222 \end{pmatrix} =$$

$$= \begin{pmatrix} 0{,}134 & -0{,}010 & -0{,}191 \\ -0{,}010 & -0{,}023 & 0{,}058 \\ -0{,}191 & 0{,}058 & 0{,}198 \end{pmatrix}$$

Матрица первых остаточных коэффициентов корреляции содержит еще достаточно большие величины и вполне допускает оценку второго главного фактора. Последующее выполнение второй итерации аналогично первой, только вычисления производятся на данных матрицы остатков R_1.

В таблице 1.6 получены более грубые оценки элементов собственного вектора, чем в первом случае (таблица 1.5). Тем не менее, итерация прервана с учетом того, что элементы квадрируемой матрицы резко снижают свою значимость и теряют наглядность, в то же время заданная условность примера допускает различную степень приближения аналитических результатов, вычисляемых по уже знакомому алгоритму.

Таблица 1.6

Матрица первых остаточных коэффициентов корреляции

Признак	R_1			$S^{(1)}$	$a^{(1)}$
	X_1	X_2	X_3		
X_1	0,134	-0,010	-0,191	-0,067	-1,000
X_2	-0,010	-0,023	0,058	0,025	0,373
X_3	-0,191	0,058	0,198	0,065	0,970

По данным таблицы 1.7 определим нагрузки второго главного фактора (таблица 1.8)

Таблица 1.7

Первый и второй циклы итерации для матрицы первых остаточных коэффициентов корреляции

Признак	R_1			$S^{(3)}$	$p^{(3)}$	$a^{(3)}$	$d=\left\|a^{(3)}-a^{(2)}\right\|$
	X_1	X_2	X_3				
X_1	0,00723	-0,00149	-0,00870	-0,00296	-0,00297	-0,830	0,036
X_2	-0,00149	0,00032	0,00179	0,00061	0,00061	0,170	0,027
X_3	-0,00870	0,00179	0,01051	0,00360	0,00358	1,000	0,000

Таблица 1.8

Вычисление нагрузок второго главного фактора

Признак	$a_2^{(3)}=U_2$	$\beta_2=R_1a_2^{(3)}$	$A_2=\dfrac{U_2\sqrt{\lambda_{22}}}{\left(\sum U_{2i}^2\right)^{1/2}}$
X_1	-0,830	-0,304	-0,383
X_2	0,170	0,062	0,079
X_3	1,000	0,366	0,462
–	$\sum_i U_{2i}^2$ =1,718	λ_2=0,366	$\sum_j a_{jr}^2$ =0,366

Матрицу вторых остаточных коэффициентов корреляции (R_2) находят из разности $R_1-A_2\times A_2^T$.

Элементы матрицы R_2 имеют малые значения, и выделение третьего главного фактора не целесообразно. Такой же вывод следует и по данным собственных чисел λ, а именно: при помощи первого и второго главных факторов полностью воспроизведена общность (trR_h =1,720) и почти на 60% удалось объяснить вариацию элементарных признаков X_1, X_2, X_3 $\left(\frac{1,412+0,366}{3}\right)$

$$R_2 = \begin{pmatrix} 0,134 & -0,010 & -0,191 \\ -0,010 & -0,023 & 0,058 \\ -0,191 & 0,058 & 0,198 \end{pmatrix} - \begin{pmatrix} -0,383 \\ 0,079 \\ 0,462 \end{pmatrix} \cdot \begin{pmatrix} -0,383 & 0,079 & 0,462 \end{pmatrix} = \begin{pmatrix} -0,013 & 0,020 & -0,014 \\ 0,020 & -0,029 & 0,022 \\ 0,014 & 0,022 & -0,015 \end{pmatrix}.$$

В заключение обобщим итоги решения задачи методом главных факторов, выделяя общности и характерности для каждого из элементарных X_j признаков (таблица 1.9).

Таблица 1.9

Факторные нагрузки, общности и характерности, вычисленные методом главных факторов

Признак	Главный фактор (факторные нагрузки)		Общность	Характерность
	F_1	F_2		
X_1	0,732	-0,383	0,683 (0,670)	0,317
X_2	0,808	0,079	0,660 (0,630)	0,340
X_3	0,471	0,462	0,435 (0,420)	0,565
–	$\sum_j a_{jr}^2$ =1,412	$\sum_j a_{j2}^2$ =0,366	$\sum h_j^2$ =1,778	$\sum d_j^2$ =1,222

В таблице 1.9 уровни общностей несколько отличаются от данных исходной матрицы R_h (приведенных в скобках). Это могло произойти из-за допускаемой грубости итеративных решений и округлений. В скобках приведены первоначально принятые величины общностей для каждого признака. Общности и характерности совместно полностью представляют дисперсии признаков, равные единице

$$\sum h_j^2 + \sum d_j^2 = 3\,.$$

В дальнейшем может быть найдена матрица значений главных факторов

$$F = (A^T A)^{-1} A^T Z'\,.$$

Рекомендуемая литература:

1. **Быковский, В. В.** Применение теории планирования эксперимента в научных и инженерных расчетах: учеб. пособие / В. В. Быковский, Л. В. Быковская, Ю. А. Дормидонов. - Оренбург : ОГУ, 2002. - 66 с - ISBN 5-7410-0442-3.

2. **Костин, В. Н.** Статистические методы и модели: учеб. пособие для вузов / В. Н. Костин, Н. А. Тишина . - Оренбург : ОГУ, 2004. - 138 с. - Библиогр.: с. 125. - ISBN 5-7410-0399-0.

Методические указания
к лабораторной работе №5

«Разработка программных средств решения задачи обработки информации, методом регрессионного анализа»

1 Цель работы:

Изучение метода регрессионного анализа для разработки программного средства

2 Задача работы:

Объяснение сути метода

3 Содержание работы:

Описание метода регрессионного анализа для разработки программных средств

4 Требования к отчету:

Отчет о проделанной работе должен содержать:

– название работы, ее задачи и описание последовательности выполнения;
– решение задачи по вариантам, указанным преподавателем;

Отчет представить в распечатанном виде.

5 Общие положения и методика выполнения задания

Регрессионный анализ — это статистический метод исследования зависимости случайной величины Y-отклик от переменной (X) или переменных X_j ($j=1, 2, \dots k$) – предикторы.

Пример 1. По пяти промышленным предприятиям имеются следующие данные о фондовооруженности труда рабочих (X_1), уровне производительности труда (X_2), удельном весе потерь от брака (X_3) (таблица 2.1).

Таблица 2.1

Номер предприятия	Фондовооруженность труда рабочего, тыс. ден. ед.	Месячная производительность труда рабочего, тыс. ден. ед.	Удельный вес потерь от брака, %
	X_1	X_2	X_3
1	3,9	7,0	2,4
2	1,1	11,1	5,9
3	1,8	10,2	6,2
4	6,0	12,0	6,0
5	5,4	10,0	11,0
	$\bar{x}_1 = 3{,}64$	$\bar{x}_2 = 10{,}06$	$\bar{x}_3 = 6{,}3$
	$\sigma_1 = 1{,}93$	$\sigma_2 = 1{,}69$	$\sigma_3 = 2{,}74$

Определить:

1) матрицы парных и частных коэффициентов корреляции;

2) множественный коэффициент детерминации и множественный коэффициент корреляции при условии, что X_2 – зависимая переменная;

3) матрицу ковариаций.

Рекомендуемая литература:

1. **Быковский, В. В.** Применение теории планирования эксперимента в научных и инженерных расчетах: учеб. пособие / В. В. Быковский, Л. В. Быковская, Ю. А. Дормидонов. - Оренбург : ОГУ, 2002. - 66 с - ISBN 5-7410-0442-3.

2. **Костин, В. Н.** Статистические методы и модели: учеб. пособие для вузов / В. Н. Костин, Н. А. Тишина . - Оренбург : ОГУ, 2004. - 138 с. - Библиогр.: с. 125. - ISBN 5-7410-0399-0.