ПРИМЕНЕНИЕ ПАКЕТА MATHCAD В МНОГОМЕРНОМ АНАЛИЗЕ ВЗАИМОСВЯЗЕЙ МИРОВЫХ ФОНДОВЫХ РЫНКОВ

УДК 519.2

Марина Анатольевна Скорик,
к.э.н., доцент, доцент кафедры Математической статистики и эконометрики Российского экономического университета им. Г.В. Плеханова
Тел.: (495) 442-71-77
Эл. почта: Skorik.MA@rea.ru

Андрей Геннадиевич Нефедов,
экономист, независимый эксперт, консультант
Эл. почта: andrew-n@yandex.ru

Компьютерные программы для аналитических исследований и прогнозирования экономических явлений и процессов по праву являются повседневным рабочим инструментом современного специалиста, связанного с обработкой статистической информации. Широкие возможности для изучения алгоритмов их работы предоставляет программа «SEMESTR» для пакета Mathcad, предназначенная для автоматизации обработки данных на основе методов многомерного анализа. Ее использование в учебном процессе позволяет сформировать устойчивые навыки применения математико-статистических методов анализа данных.

Ключевые слова: *компонентный анализ, многомерный анализ, корреляционный анализ, регрессионный анализ, фондовые индексы, Mathcad.*

Marina A. Skorik,
Ph.D. of Economics, Associate Professor of the Department of Mathematical Statistics and Econometrics, Plekhanov Russian University of Economics
Tel.: (495) 442-71-77
E-mail: Skorik.MA@rea.ru

Andrey G. Nefedov,
economist, independent expert, consultant
E-mail: andrew-n@yandex.ru

MATHCAD APPLICATION FOR THE MULTIDIMENSIONAL ANALYSIS OF THE INTERDEPENDENCES ON WORLD STOCK MARKETS

Computer programs for analytical studies and forecasting of economic phenomena and processes are the everyday operating tool of a modern expert associated with the processing of statistical information. The program "SEMESTR" created on the base of Mathcad provides the wide range of possibilities for the comprehension of these algorithms due to automation of the multivariate data processing. Its usage in the educational process allows to develop stable skills in applying of mathematical and statistical methods of data analysis.

Keywords: *factor analysis, multivariate analysis, correlation analysis, regression analysis, stock indexes, dimensionality of factors space, Mathcad.*

1. Введение

Владение компьютерными методами анализа данных многопризнаковой природы – один из базовых элементов подготовки современного специалиста. Программа «SEMESTR» предназначена для автоматизации обработки данных методами корреляционно-регрессионного и компонентного анализа и рекомендуется к использованию в учебном процессе по дисциплине «Многомерный статистический анализ». Представляется целесообразным, чтобы студенты ознакомились с ней до решения задачи большей размерности на базе современных статистических пакетов анализа данных, представляющих в распоряжение исследователя лишь конечные результаты машинной обработки, без промежуточных расчетов.

Это позволит проиллюстрировать алгоритмы реализации основных методов многомерного статистического анализа, их особенности и возможности содержательной интерпретации результатов [5]. Косвенно это поможет достигнуть еще одной цели: побудить студентов и молодых ученых к продолжению работы в данном направлении – созданию аналогичного программного продукта, например, для решения задачи классификации объектов, т.е. реализации кластерного и дискриминантного анализа. Именно поэтому так подробно расписана инструкция работы с программой и приведен текст самой программы.

2. Исходные данные для исследования зависимостей

В качестве исходной статистической информации для исследования зависимости российского индекса РТС от 12 фондовых индексов развитых и развивающихся стран были взяты их ежедневные значения на момент закрытия бирж за 2005–2014 гг., т.е. за 3347 дней (см. табл. 1) [6, 7].

Таблица 1

Условные обозначения мировых фондовых индексов, используемые в исследовании

Страна	Фондовый индекс	Переменная
Россия	*RTSI*	*Y*
Бразилия	*BOVESPA*	*X1*
Франция	*CAC 40*	*X2*
Германия	*DAX 30*	*X3*
Великобритания	*FTSE 100*	*X4*
Китай	*Hang Seng*	*X5*
Испания	*IBEX 35*	*X6*
США	*D&J*	*X7*
Мексика	*IPC*	*X8*
Индия	*BSE*	*X9*
Турция	*XU100*	*X10*
Южно-Африканская Республика	*JSE*	*X11*
Чехия	*PX50*	*X12*

Для автоматизации статистической обработки данных с подробным выводом промежуточных результатов в пакете Mathcad 15.0 PRO, на встроенном языке программирования, авторами была написана программа «SEMESTR». Она представляет собой файл с расширением *.xmcd, после открытия которого, появляется рабочий лист пакета Mathcad, разделённый на две части: в левой половине

Y :=

	1
3336	1343.62
3337	1306.00
3338	1303.15
3339	1315.54
3340	1315.54
3341	1315.54
3342	1318.98
3343	1306.37
3344	1286.07
3345	1260.95
3346	1267.27
3347	...

X :=

	9	10	11	12
3336	20634.21	64125.72	47112.43	1036.87
3337	20722.97	63382.13	47438.31	1038.36
3338	20536.64	63681.01	47149.35	1036.05
3339	20700.75	63885.62	47452.24	1046.06
3340	20700.75	63885.62	47452.24	1046.06
3341	20700.75	63885.62	47452.24	1046.06
3342	20811.44	64185.96	47394.00	1039.29
3343	20852.47	62123.84	46956.75	1029.56
3344	20986.99	61503.04	47017.19	1019.77
3345	20986.99	61779.24	47049.79	1014.43
3346	21120.12	62553.32	47328.92	1015.11
3347	21120.12	62553.32	47328.92	...

Рис. 1. Ввод исходных данных в пакете Mathcad

	sred =		SCO =
$\overline{Y}$	1462.388	$\hat{S}_y$	436.129
$\overline{X}_1$	52026.461	$\hat{S}_{x_1}$	13049.143
$\overline{X}_2$	4161.146	$\hat{S}_{x_2}$	832.073
$\overline{X}_3$	6413.407	$\hat{S}_{x_3}$	1267.204
$\overline{X}_4$	5662.039	$\hat{S}_{x_4}$	675.808
$\overline{X}_5$	19971.952	$\hat{S}_{x_5}$	3639.531
$\overline{X}_6$	10548.005	...	2252.282
$\overline{X}_7$	11839.621	$\hat{S}_{x_{12}}$	1911.382
...	29953.405		8922.088
$\overline{X}_{12}$	15186.661		4147.163
	50763.307		16192.036
	28253.548		8067.565
	1221.797		294.875

```
p := cols(X)   n := rows(Y)   YX := augment(Y, X)
Variation := | for j ∈ 1..p + 1
             |    | sred_j ← mean(YX<j>)
             |    | SCO_j ← stdev(YX<j>)
             | Out_1 ← sred
             | Out_2 ← SCO
             | Out
sred := Variation_1   SCO := Variation_2
```

Рис. 2. Вычисление выборочных характеристик показателей

CORR(YX) =

	1	2	3	4	5
1	1.00	0.65	0.52	0.64	0.69
2	0.65	1.00	-0.17	0.50	0.29
3	0.52	-0.17	1.00	0.31	0.60
4	0.64	0.50	0.31	1.00	0.87
5	0.69	0.29	0.60	0.87	1.00
6	0.81	0.82	0.23	0.80	0.66
7	0.58	0.06	0.88	0.18	0.38
8	0.52	0.27	0.33	0.92	0.91
9	0.49	0.76	-0.25	0.79	0.54
10	0.60	0.85	-0.12	0.80	0.56
11	0.42	0.64	-0.17	0.79	0.62
12	0.46	0.64	-0.15	0.88	0.63
13	0.61	-0.02	0.93	0.18	0.47

```
CORR(YX) := | for k1 ∈ 1..p + 1
            |   for k2 ∈ 1..p + 1
            |      CORR_{k1,k2} ← corr(YX<k1>, YX<k2>)
            | CORR
```

Рис. 3. Расчет парных коэффициентов корреляции между признаками

T =

	1	2	3	4	5
1	1.00	49.35	35.13	48.41	55.67
2	49.35	1.00	10.15	33.72	17.46
3	35.13	10.15	1.00	18.61	43.46
4	48.41	33.72	18.61	1.00	103.87
5	55.67	17.46	43.46	103.87	1.00
6	78.52	81.72	13.96	76.38	50.48
7	41.19	3.75	108.83	10.33	23.86
8	35.27	16.22	19.88	134.23	123.57
9	32.08	68.12	14.84	73.59	37.28
10	43.74	93.88	7.17	76.57	38.77
11	26.80	48.53	10.03	74.14	46.21
12	30.27	48.08	8.51	105.74	46.82
13	44.24	1.04	149.56	10.68	30.56

$\alpha := 0.01 \quad \nu := n - 2 \quad \nu = 3345$

```
T := | for i ∈ 1..p + 1
     |   for j ∈ 1..p + 1
     |      | CORR_{i,j} ← corr(YX<i>, YX<j>)
     |      | CORR_{i,i} ← 0
     |      | T_{i,j} ← |CORR_{i,j}| / sqrt(1 − (CORR_{i,j})^2) · sqrt(n − 2)
     |      | T_{i,i} ← 1
     | T
```

$tkr := \left| qt\left(\frac{\alpha}{2}, \nu\right) \right| \quad tkr = 2.577$

Рис. 4. Расчет *t*-статистики Стьюдента

(видимая часть) пользователь может ввести свои исходные данные и тут же получить результат, а в правой части представлен сам алгоритм программы (для того чтобы его увидеть нужно прокрутить нижний бегунок вправо).

Данная программа может работать с одним результативным признаком и любым числом объясняющих переменных, число значений каждого показателя также может быть сколь угодно большим. Вводить новые данные необходимо в векторе Y (результативный признак) и в матрице X (факторные признаки) (см. рис. 1).

Помимо ручного ввода данных, через буфера обмена, Mathcad также позволяет импортировать большие массивы чисел из таких форматов как: Delimited Text (*.dat, *.txt); Comma Separated Values (*.csv); Excel Files (*.xls); Lotus 1-2-3 Files (*.wks, *.wk1, *.wk*); MATLAB Files (*.mat); dbase Files (*.dbf) [4].

3. Корреляционный анализ

После того как все данные введены, Mathcad начинает вычислять такие характеристики как средние (sred) и среднеквадратические отклонения (SCO) для каждого признака. Для такой описательной статистики был написан следующий код с использованием встроенных функций пакета и осуществлен вывод полученных результатов (см. рис. 2).

Далее рассчитывается матрица парных коэффициентов корреляции R (CORR(YX) порядка $[k \times k]$, где $k = \overline{1, p+1}$), которая является симметрической и положительно определённой. В нашем случае k – общее число признаков, а $p = 12$ – число факторных показателей (см. рис. 3).

Далее проверяется значимость полученных коэффициентов корреляции с помощью t-статистики Стьюдента (см. рис. 4).

Тестируемый коэффициент корреляции считается значимым, т.е. гипотеза $H_0: \rho = 0$ отвергается с вероятностью ошибки α, если $t_{набл.}$ будет больше, чем $t_{кр.}$, определяемому по таблице t-распределения для заданного α и $\nu = n - 2$.

4. Регрессионный анализ

Далее находится оценка уравнения регрессии вида: $\hat{y} = b_0 + b_1 \cdot x_1 + b_2 \cdot x_2 + ... + b_{12} \cdot x_{12}$. Согласно методу наименьших квадратов вектор оценок коэффициентов регрессии b получается по формуле: $b = (X^T \cdot X)^{-1} \cdot (X^T \cdot Y)$. Алгоритм программы создан так, что пользователю не нужно вводить в матрице X единичный вектор-столбец (см. рис. 5, 6).

Значимость уравнения регрессии, т.е. гипотеза $H_0 : \beta_0 = \beta_1 = \beta_2 = \beta_3 = 0$, проверяется по F-критерию, наблюдаемое значение которого определяется по формуле:

$$F_{набл.} = \frac{Q_R/(p+1)}{Q_{ост.}/(n-p-1)},$$

где:

$$Q_R = (X \cdot b)^T \cdot (X \cdot b),$$
$$Q_{ост.} = (Y - X \cdot b)^T \cdot (Y - X \cdot b),.$$

По таблице F-распределения для заданных α, $\nu_1 = p + 1$, $\nu_2 = n - p - 1$ находят $F_{кр}$. Гипотеза H_0 отклоняется с вероятностью ошибки α, если $F_{набл.} > F_{кр}$. Из этого следует, что уравнение является значимым, т.е. хотя бы один из коэффициентов регрессии отличен от нуля.

Для проверки значимости отдельных коэффициентов регрессии, т.е. гипотез $H_0 : \beta_j = 0$, где $j = \overline{1,k}$, используется t-критерий Стьюдента и вычисляют: $t_{набл}(b_j) = \frac{b_j}{\hat{S}_{b_j}}$. По таблице t-распределения для заданного уровня значимости α и числа степеней свободы $\nu = n - k - 1$ находят $t_{кр}$ (см. рис.7).

Гипотеза H_0 отвергается с вероятностью ошибки α если $|t_{набл}| > t_{кр}$. Из этого следует, что соответствующий коэффициент регрессии β_j значим, т.е. $\beta_j \neq 0$. В противном случае коэффициент регрессии незначим, и, соответственно, переменная в регрессионную модель не включается. В алгоритме программы для проверки значимости отдельных коэффициентов регрессии дополнительно рассчитывается вероятность p, используя распределение Стьюдента:

$$p\_t_j = 2 \cdot \left[1 - P\left(t_{набл}(b_j), \nu = n - p - 1\right)\right]$$

если $t_{набл}(b_j) > 0$ и

X =

	1	2	3	4	5
3341	1.00	47380.24	4381.06	9656.95	6838.06
3342	1.00	47500.87	4419.13	9708.94	6865.86
3343	1.00	46728.19	4414.55	9699.35	6830.50
3344	1.00	46574.31	4396.91	9661.73	6799.15
3345	1.00	47590.03	4396.39	9588.33	6810.27
3346	1.00	47094.59	4408.08	9692.08	6809.70
3347	1.00	47094.40	4408.08	9692.08	6809.70

$$X := \left| \begin{array}{l} \text{for } i \in 1..n \\ \quad X_{i,1} \leftarrow 1 \\ \text{for } k \in 2..p+1 \\ \quad X^{\langle k \rangle} \leftarrow YX^{\langle k \rangle} \\ X \end{array} \right.$$

Рис. 5. Корректировка матрицы факторных признаков X

Ykr =

	1
3332	1512.543
3333	1513.197
3334	1513.197
3335	1546.446
3336	1596.098
3337	1615.730
3338	1600.215
3339	1626.588
3340	1628.308
3341	1628.308
3342	1608.106
3343	1591.806
3344	1582.830
3345	1603.189
3346	1584.327
3347	...

b =

	1
1	-1570.13
2	0.01
3	-0.07
4	-0.09
5	0.39
6	0.01
7	-0.06
8	-0.04
9	0.01
10	0.03
11	-0.02
12	0.01
13	1.21

$$\mathrm{Regress}(YX) := \left| \begin{array}{l} y \leftarrow YX^{\langle 1 \rangle} \\ sumy \leftarrow \sum_{i=1}^{n} y_i \\ b \leftarrow (X^T \cdot X)^{-1} \cdot (X^T \cdot y) \\ Ykr \leftarrow X \cdot b \\ sumYkr \leftarrow \sum_{i=1}^{n} Ykr_i \\ e \leftarrow (y - Ykr) \\ \text{for } i \in 1..n \\ \quad \delta_i \leftarrow \frac{e_i}{Ykr_i} \\ DF\_Qr \leftarrow p + 1 \\ DF\_Qost \leftarrow [n - (p + 1)] \\ Qr \leftarrow Ykr^T \cdot Ykr \\ Qost \leftarrow e^T \cdot e \\ Q\_sum \leftarrow Qr + Qost \\ SR \leftarrow \frac{Qr}{DF\_Qr} \\ SS \leftarrow \frac{Qost}{DF\_Qost} \\ F \leftarrow \frac{SR}{SS} \\ p\_F \leftarrow 1 - pF(F, DF\_Qr, DF\_Qost) \\ Out_1 \leftarrow \begin{pmatrix} Ykr \\ e \\ \delta \end{pmatrix} \\ Out_2 \leftarrow \begin{pmatrix} DF\_Qr & Qr & SR & F \\ DF\_Qost & Qost & SS & p\_F \end{pmatrix} \\ Out_3 \leftarrow b \\ Out_4 \leftarrow \begin{pmatrix} sumy \\ sumYkr \end{pmatrix} \\ Out_5 \leftarrow Q\_sum \\ Out \end{array} \right.$$

$$Ykr := \left(\mathrm{Regress}(YX)_1\right)_{1,1} \quad e := \left(\mathrm{Regress}(YX)_1\right)_{2,1} \quad \delta := \left(\mathrm{Regress}(YX)_1\right)_{3,1}$$

$$b := \mathrm{Regress}(YX)_3 \qquad \begin{pmatrix} sumy \\ sumYkr \end{pmatrix} := \mathrm{Regress}(YX)_4$$

$$\begin{pmatrix} DF\_Qr & Qr & SR & F \\ DF\_Qost & Qost & SS & p\_F \end{pmatrix} := \mathrm{Regress}(YX)_2 \quad Q\_sum := \mathrm{Regress}(YX)_5$$

$$DF\_Qsum := DF\_Qr + DF\_Qost$$

Рис. 6. Алгоритм оценки линейного уравнения регрессии

DF-степени свободы	Q- дисперсии	SS=Q/DF_Q	F-значение	p-вероятность
DF_Qr = 13	Qr = 7730381230.681	SR = 594644710.052	F = 30943.868	p_F = 0.000
DF_Qost = 3334	Qost = 64069089.643	SS = 19216.884	Fkr = 2.665	α = 0.001
DF_Qsum = 3347	Q_sum = 7794450320.32	$\sqrt{SS}$ = 138.625		

Проверка:Q_sum= $\sum_{i=1}^{n} (Y_i)^2 = 7794450320.33$ коэффициент детерминации R^2 Rsq = 0.8994

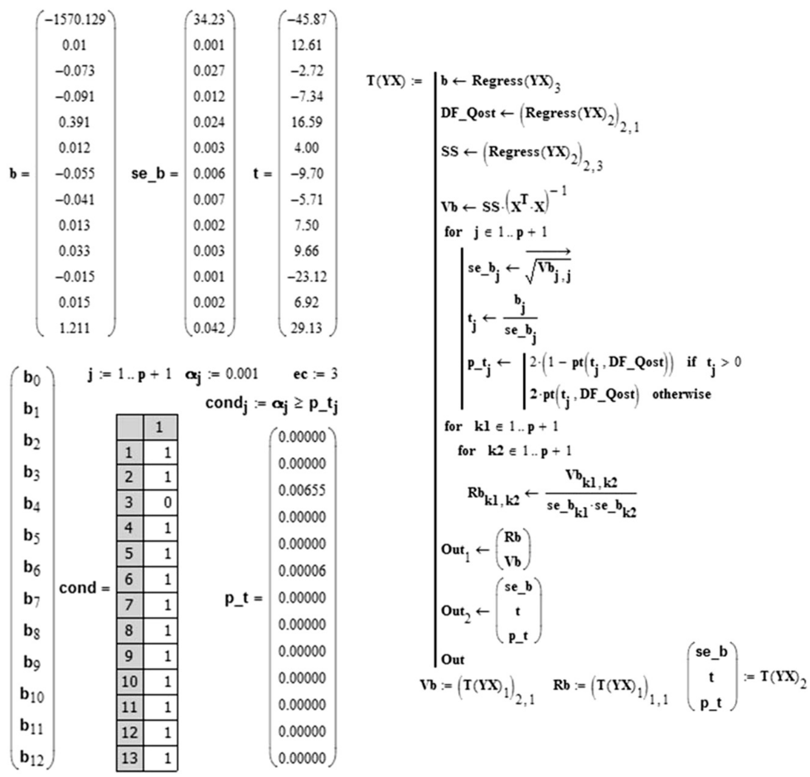

Рис. 7. Проверка значимости коэффициентов регрессии

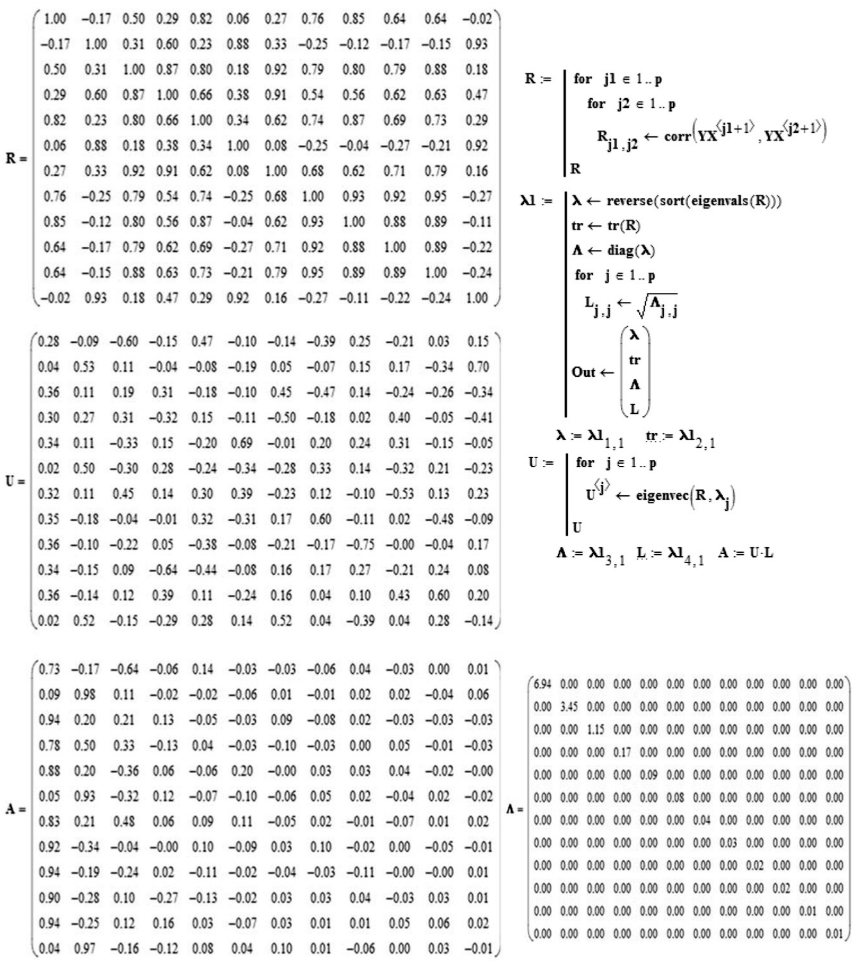

Рис. 8. Расчет собственных значений корреляционной матрицы и нагрузок

$$p\_t_j = 2 \cdot P\left(t_{набл}(b_j), \nu = n - p - 1\right)$$

если $t_{набл}(b_j) \le 0$.

Гипотеза H_0 отвергается с вероятностью ошибки α если $\alpha > p\_t_j$ (в программе, в векторе **cond**, напротив соответствующего коэффициента регрессии при переменной x_j, будет стоять единица, т.е. условие выполнилось), что говорит о значимости коэффициента регрессии β_j. Иначе – коэффициент регрессии незначим (в векторе **cond** – 0).

В случае если хотя бы один коэффициент регрессии незначим, то переходят к пошаговой схеме исключения несущественных переменных. Для этого пользователю необходимо ввести номер исключаемой переменной, согласно порядковому номеру в векторе **cond**, т.е. присвоить переменной “**ec**” число, например **ec: = 3**. Это значит, что исключается факторный признак X_2, стоящий на 3-м месте.

Но непосредственно перед переходом к следующему шагу регрессионного анализа рассчитываются ещё доверительные границы для каждого коэффициента регрессии по формуле:

$$\beta_j \in \left[b_j \pm t_{кр} \cdot \sqrt{\hat{S}^2 \cdot \left(\left(X^T \cdot X\right)^{-1}\right)_{jj}} \right].$$

Далее находится интервальная оценка результативного признака в некоторой точке, определяемой вектором начальных значений X_0. Доверительные границы для результативного признака $\widetilde{y}_n$ находятся по формуле, вида:

$$\widetilde{y}_n \in \left[\begin{array}{c} X_0^T b \pm t_{кр} \cdot \sqrt{\hat{S}^2} \cdot \\ \cdot \sqrt{X_0^T \cdot \left(X^T \cdot X\right)^{-1} \cdot X_0} \end{array} \right].$$

Доверительная оценка для интервала предсказания условного математического ожидания $\widetilde{y}_{n+1}$ в точке, не принадлежащей выборке, определяется по следующей формуле:

$$\widetilde{y}_{n+1} \in \left[\begin{array}{c} X_0^T b \pm t_{кр} \cdot \sqrt{\hat{S}^2} \cdot \\ \cdot \sqrt{X_0^T \cdot \left(X^T \cdot X\right)^{-1} \cdot X_0 + 1} \end{array} \right].$$

Следует заметить, что пошаговую процедуру исключения пользователь может прервать на любой стадии пошагового регрессионного анализа, но только, когда все коэффициенты регрессии значимы [3]. Тогда можно переходить к следующему виду статистического анализа, а именно к компонентному анализу.

5. Компонентный анализ

Здесь на основании матрицы факторных признаков X размерности $[n \times p]$, рассчитывается матрица парных коэффициентов корреляции R размерности $[p \times p]$ с элементами

$$r_{jl} = \frac{\frac{1}{n} \cdot \sum_{i=1}^{n} \left(x_{ij} - \bar{x}_j\right) \cdot \left(x_{il} - \bar{x}_l\right)}{S_j \cdot S_l};$$

$$j, l = 1, 2, ..., p.$$

Далее матрица R преобразуется в диагональную матрицу Λ собственных значений характеристического многочлена $|\lambda E - R|$, где E – единичная матрица. Для этого решается уравнение вида: $|\lambda E - R| = 0$, корнями которого являются p собственных значений $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_{12} > 0$. Матрица собственных значений Λ имеет вид главной диагонали. Собственные значения характеризуют вклады соответствующих главных компонент в суммарную дисперсию исходных признаков, равную p, т.е. $\sum_{\nu=1}^{p} \lambda_\nu = p$. Таким образом, первая главная компонента оказывает наибольшее влияние на общую вариацию, а последняя p-ая – наименьшее.

На основе найденных собственных чисел определяются собственные вектора. V_ν – собственный вектор, соответствующий собственному значению λ_ν корреляционной матрицы R, определяется как отличное от нуля решение уравнения $(\lambda_\nu E - R) \cdot V_\nu = 0$, откуда нормированный собственный вектор U_ν равен: $U_\nu = \frac{V_\nu}{\sqrt{V_\nu^T \cdot V_\nu}}$. Так получаем матрицу U, составленную из нормированных собственных векторов: $U = \left(\vec{U}_1 \middle| \vec{U}_2 \middle| \vec{U}_3\right)$. Затем находим матрицу факторных нагрузок по формуле: $A = U \cdot \Lambda^{1/2}$, (в программе: $A = UL$, где $L = \Lambda^{1/2}$) (см. рис. 8). Элемент матрицы $a_{j\nu}$, где: $j, \nu = \overline{1, p}$, характери-

Таблица 2

Вклад в дисперсию ν-ой главной компоненты

f_ν	f_1	f_2	…	f_{12}
λ_ν	$\lambda_1 = 6{,}938$	$\lambda_2 = 3{,}454$	…	$\lambda_{12} = 0{,}007$
вклад, %	$\frac{\lambda_1}{\mathbf{tr}} \cdot 100 = 57{,}82$	$\frac{\lambda_2}{\mathbf{tr}} \cdot 100 = 28{,}78$	…	$\frac{\lambda_{12}}{\mathbf{tr}} \cdot 100 = 0{,}06$
Сумма, %	$\frac{\lambda_1}{\mathbf{tr}} \cdot 100 = 57{,}82$	$\left[\frac{(\lambda_1 + \lambda_2)}{\mathbf{tr}}\right] \cdot 100 = 86{,}60$	…	$\left(\frac{\sum_{j=1}^{p} \lambda_j}{\mathbf{tr}}\right) \cdot 100 = 100$

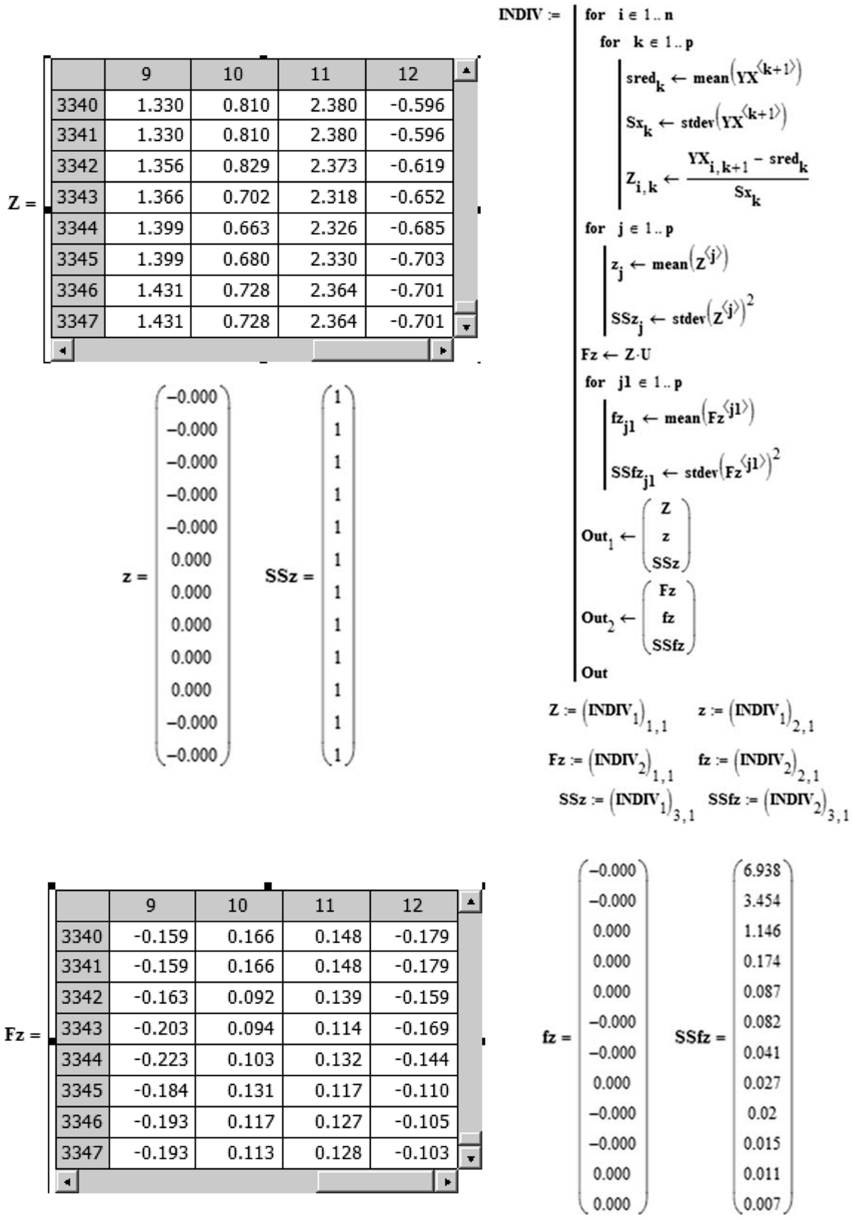

Рис. 9. Расчёт матриц стандартизованных значений признаков и индивидуальных значений главных компонент

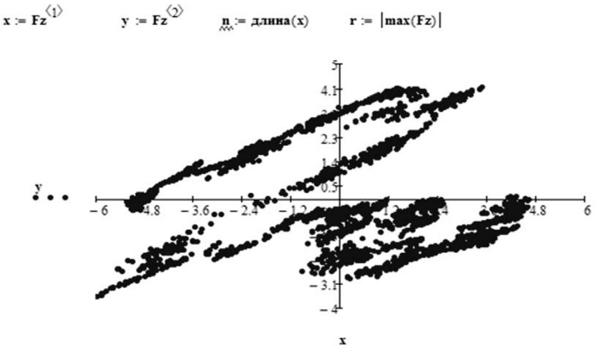

Рис. 10. Диаграмма рассеяния наблюдений в пространстве ГК

Y_t =

	1
3338	-0.37
3339	-0.34
3340	-0.34
3341	-0.34
3342	-0.33
3343	-0.36
3344	-0.40
3345	-0.46
3346	-0.45
3347	-0.45

Y_r =

	1
3335	0.15
3336	0.25
3337	0.30
3338	0.27
3339	0.33
3340	0.33
3341	0.33
3342	0.29
3343	0.25
3344	0.24
3345	0.29
3346	0.25
3347	0.25

$$b = \begin{pmatrix} 0.000 \\ 0.257 \\ 0.273 \\ -0.314 \\ -0.163 \\ 0.746 \\ 0.037 \\ 0.006 \\ -0.134 \\ -0.685 \\ 0.708 \\ 0.120 \end{pmatrix}$$

```
RegressF := | for i ∈ 1..n
            |    Y_t[i,1] ← (Y[i,1] − mean(Y)) / stdev(Y)
            | for i ∈ 1..n
            |    F[i,1] ← 1
            | for k ∈ 2..p
            |    F⟨k⟩ ← Fz⟨k−1⟩
            | b ← (Fᵀ·F)⁻¹·(Fᵀ·Y_t)
            | Y_r ← F·b
            | e ← (Y_t − Y_r)
            | for i ∈ 1..n
            |    δ[i] ← e[i] / Y_r[i]
            | DF_Qr ← p
            | DF_Qost ← (n − p)
            | Qr ← Y_rᵀ·Y_r
            | Qost ← eᵀ·e
            | Q_sum ← Qr + Qost
            | SR ← Qr / DF_Qr
            | SS ← Qost / DF_Qost
            | Fn ← SR / SS
            | p_F ← 1 − pF(Fn, DF_Qr, DF_Qost)
            | Out[1] ← (F, e, δ)
            | Out[2] ← ( DF_Qr    Qr    SR  Fn
            |            DF_Qost  Qost  SS  p_F )
            | Out[3] ← b
            | Out[4] ← (Y_t, Y_r)
            | Out[5] ← Q_sum
            | Out

Y_t := (RegressF[4])[1,1]     F := (RegressF[1])[1,1]
Y_r := (RegressF[4])[2,1]     b := RegressF[3]
e := (RegressF[1])[2,1]       δ := (RegressF[1])[3,1]
( DF_Qr    Qr    SR  Fn
  DF_Qost  Qost  SS  p_F ) := RegressF[2]
DF_Qsum := DF_Qr + DF_Qost
Q_sum := RegressF[5]
```

Рис. 11. Алгоритм оценки уравнения регрессии на ГК

зует тесноту связи между факторным показателем X_j и f_v-ой главной компонентой, причём: $-1 \le a_{jv} \le 1$. При этом справедливы соотношения: $\sum_{j=1}^{p} a_{j\nu}^2 = \lambda_\nu$ и $\sum_{\nu=1}^{p} a_{j\nu}^2 = 1$.

Матрица факторных нагрузок A используется для экономической интерпретации главных компонент, представляющих собой линейные функции исходных признаков [2]. Для экономической интерпретации f_v используются лишь те факторы X_j, для которых $\left|a_{j\nu}\right| > 0{,}6$.

Затем рассчитывается вклад v-ой главной компоненты в суммарную дисперсию, а также суммарный вклад m первых главных компонент (см. табл. 2). Для дальнейшего анализа обычно используют m первых главных компонент, суммарный вклад которых превышает 70%.

Далее рассчитывается матрица нормированных значений Z с элементами $z_{ij} = \dfrac{x_{ij} - \bar{x}_j}{\hat{S}_{x_j}}$ (см. рис. 9).

Причём для матрицы Z должны выполняться следующие условия:

$$\bar{z}_j = \frac{1}{n} \cdot \sum_{i=1}^{n} z_{ij} = 0 \text{ для } j = \overline{1,p} \text{ и}$$

$$S_{z_j}^2 = \frac{1}{n} \cdot \sum_{i=1}^{n} \left(z_{ij} - \bar{z}_j\right)^2 = 1 \text{ для } j = \overline{1,p}.$$

Значения главных компонент для каждого i-ого объекта $\left(i = \overline{1,n}\right)$ задаются матрицей F. Матрица значений главных компонент находится по следующей формуле: $F = Z \cdot U$. Для матрицы F справедливы следующие соотношения:

$$\bar{f}_j = \frac{1}{n} \cdot \sum_{i=1}^{n} f_{ij} = 0 \text{ для } j = \overline{1,p} \text{ и}$$

$$S_{f_j}^2 = \frac{1}{n} \cdot \sum_{i=1}^{n} \left(f_{if} - \bar{f}_j\right)^2 = \frac{1}{n} \cdot \sum_{i=1}^{n} f_{ij}^2 = \lambda_\nu \quad \text{для } j,\nu = \overline{1,p}.$$

На основании матрицы индивидуальных значений главных компонент на объектах F размерности $[n \times p]$, рассчитывается матрица парных коэффициентов корреляции между главными компонентами CORR(Fz). Поскольку главные компоненты должны быть ортогональны, то парный коэффициент корреляции между ними стремится к нулю.

DF-степени свободы	Q- дисперсии	SS=Q/DF_Q	F-значение	p-вероятность
DF_Qr = 12	Qr = 3008.581	SR = 250.715	Fn = 2470.705	p_F = 0.000
DF_Qost = 3335	Qost = 338.419	SS = 0.101	Fkr = 2.752	α = 0.001
DF_Qsum = 3347	Q_sum = 3347.00	$\sqrt{SS}$ = 0.319	коэф-т детерминации	Rsq = 0.8989

$j := 1..p$ $\alpha_j := 0.05$ ***Проверка на значимость:*** $cond1_j := \alpha_j \geq p\_t_j$

	b =	se_b =	p_t =	cond1 =
	1	1	1	1
1	0.000	0.0055	1.00000	0
2	0.257	0.0021	0.00000	1
3	0.273	0.0030	0.00000	1
4	-0.314	0.0051	0.00000	1
5	-0.163	0.0132	0.00000	1
6	0.746	0.0187	0.00000	1
7	0.037	0.0193	0.05712	0
8	0.006	0.0274	0.83671	0
9	-0.134	0.0338	0.00007	1
10	-0.685	0.0387	0.00000	1
11	0.708	0.0447	0.00000	1
12	0.120	0.0521	0.02102	1

Рис. 12. Проверка значимости коэффициентов регрессии при ГК

Далее строится диаграмма рассеяния в пространстве двух первых главных компонент, обеспечивающий суммарный вклад в дисперсию процесса не менее 60–70%, т.е. в прямоугольной системе координат наносятся объекты (точки) и производится их классификация (см. рис. 10).

6. Регрессия на главные компоненты

Уравнение регрессии на главных компонентах строится по стандартному алгоритму регрессионного анализа [1], где в качестве аргументов используются наиболее весомые главные компоненты, а не исходные показатели (см. рис.11).

Сначала находится вектор стандартизированных значений результативного признака Y с элементами $\dot{y}_i = \dfrac{y_i - \overline{y}}{S_y}$. Причём должны выполняться следующие условия: $\sum_{i=1}^{n} \dot{y}_i = 0$ и $\sum_{i=1}^{n} \dot{y}_i^2 = n$ ($n = 3347$).

Затем, согласно, методу наименьших квадратов вектор оценок коэффициентов регрессии b получается по формуле:

$$b = \left(F^T \cdot F\right)^{-1} \cdot \left(F^T \cdot \dot{Y}\right),$$

где: F_T – транспонированная матрица F; $(F^T F)^{-1}$ – матрица, обратная к $F^T F$.

Далее по F-критерию проверяется значимость уравнения регрессии, т.е. гипотеза $H_0 : \beta_0 = \beta_1 = \beta_2 = 0$, наблюдаемое значение которого определяется по формуле:

$$F_{набл.} = \frac{Q_R / (p+1)}{Q_{ост.} / (n-p-1)},$$

где:

$$Q_R = (F \cdot b)^T \cdot (F \cdot b),$$

$$Q_{ост.} = (\dot{Y} - F \cdot b)^T \cdot (\dot{Y} - F \cdot b).$$

По таблице F-распределения для заданных α, $\nu_1 = p + 1$, $\nu_2 = n - p - 1$ находят $F_{кр}$.

Гипотеза H_0 отклоняется с вероятностью α, если $F_{набл.} > F_{кр}$. Из этого следует, что уравнение является значимым, т.е. хотя бы один из коэффициентов регрессии отличен от нуля.

Для проверки значимости отдельных коэффициентов регрессии, т.е. гипотез $H_0 : \beta_j = 0$, где $j = \overline{1,p}$, используется t-критерий Стьюдента. По таблице t-распределения для заданного уровня значимости α и числа степеней свободы $\nu = n - k - 1$ находят $t_{кр}$.

Гипотеза H_0 отвергается с вероятностью ошибки α если $\left|t_{набл}\right| > t_{кр}$. Из этого следует, что соответствующий коэффициент регрессии β_j значим, т.е. $\beta_j \neq 0$, иначе – коэффициент регрессии незначим.

В алгоритме проверки значимости отдельных коэффициентов регрессии также рассчитывается вероятность p, используя распределение Стьюдента (см. рис. 12):

$$p\_t_j = 2 \cdot \left[1 - P\left(t_{набл}\left(b_j\right), \nu = n - p - 1\right)\right]$$

если $t_{набл}(b_j) > 0$ и

$$p\_t_j = 2 \cdot P\left(t_{набл}\left(b_j\right), \nu = n - p - 1\right)$$

если $t_{набл}(b_j) \leq 0$.

Гипотеза H_0 отвергается с вероятностью ошибки α если $\alpha > p\_t_j$ (в программе, в векторе **cond1**, напротив соответствующего коэффициента регрессии при переменной x_j, будет стоять единица, т.е. условие выполнилось), что говорит о значимости коэффициента регрессии β_j. Иначе – коэффициент регрессии незначим (в векторе **cond1** – 0).

7. Заключение

Таким образом, применение компонентного анализа позволяет исследователю несколько иначе взглянуть на изучаемый процесс и получить сведения, недоступные при использовании аппарата классического корреляционно-регрессионного анализа [5]. Если результативный признак стандартизован, то исследователь получает интересные дополнительные возможности. Компоненты не коррелированы, свободный коэффициент равен нулю. Поэтому можно отобрать только информативные главные компоненты (имеющие значимые коэффициенты корреляции с результативным признаком) и построить простые собственно линейные уравнения на каждую такую компоненту отдельно, а затем все эти уравнения сложить. В этом случае удается добиться значительного снижения объема вычислений.

Наконец, можно попытаться восстановить исходные показатели по наиболее информативным главным компонентам. Полученные в ходе этой процедуры стандартизованные признаки используют для восстановления исходных характеристик и оценки их гипотетических значений, которые можно сравнить с фактическими значениями Xij и проанализировать характер отклонений с целью определения «узких мест» в наблюдаемой совокупности и выработки рекомендаций по их устранению.

В заключение следует отметить, что данная программа обработки исходных данных является учебной и составлена в целях иллюстрации не только конечных, но и промежуточных результатов. В случае реальных исследований используют более серьезное математическое обеспечение.

Литература

1. Айвазян С.А. Методы эконометрики. – М.: ИНФРА-М, 2010.

2. Дубров А.М. Компонентный анализ и эффективность в экономике. – М.: Финансы и статистика, 2002.

3. Ниворожкина Л.И., Арженовский С.Б. Многомерные статистические методы в экономике. – М.: Дашков и Ко, 2009.

4. Плис А.И., Сливина Н.А. MATHCAD: математический практикум для инженеров и экономистов. – М.: Финансы и статистика, 2003.

5. Скорик М.А., Нефедов А.Г. Методические подходы к исследованию многомерных зависимостей на примере фондового рынка. // Экономика, статистика и информатика. Вестник УМО. – 2015. – №4. – С.158–163.

6. Официальный сайт Инвестиционного холдинга «Финам». [Электронный ресурс] – Код доступа: http://finam.ru/

7. Официальный сайт Московской биржи ММВБ-РТС [Электронный ресурс] – Код доступа: http://moex.com/

References

1. Aivazian S.A. Metody ekonometriki [Methods of econometrics]. – M.: INFRA-M, 2010

2. Dubrov A.M. Kjmponentny analiz i effektivnost’ v ekonomike [Component analysis and efficiency of economics]. – M.: Finance and statistics, 2002

3. Nivorozhkina L.I., Arzhenovsky S.B. Mnogomernye statisticheskie metody v ekonomike [Multivariate statistical methods in economics]. – M.: Finance and statistics, 2009

4. Plis A.I., Slivina N.A. MATHCAD: praktikum dlya inzhenerov i ekonomistov [MATHCAD: mathematical practical work for engineers and economists]. – M.: Finance and statistics, 2003

5. Skorik M.A., Nefedov A.G. Metodicheskie podhody k issledovaniyu mnogomernyh zavisimostey na primere fondovogo rynka [Approach to the research of the multivariate relationship by the example of the Russian stock market]. // Economics, statistics and informatics. UMO – 2015. – №4. – pp.158–163

6. Investment holding “FINAM” [Electronic resource] – Mode of access: http://finam.ru/

7. Moscow stock exchange MMVB-RTS [Electronic resource] – Mode of access: http://moex.com/