СПЕКТРАЛЬНЫЙ АНАЛИЗ ЛИНЕЙНОГО УРАВНЕНИЯ РЕГРЕССИИ

Клюжев Н.А.[1], Алексеева А.М.[2]

[1]Клюжев Николай Алексеевич - кандидат технических наук, доцент

[2]Алексеева Алла Михайловна - кандидат экономических наук, доцент

Псков, Российская Федерация

Аннотация: *рассмотрено применение спектрального анализа для идентификации "линейного смещения" в оценках коэффициентов уравнения многомерной и линейной регрессии, которое влияет на интерпретацию параметров регрессии в практике применения этой математической модели.*

Ключевые слова: *билинейная метрика, полином, спектр, матрица, пример.*

SPECTRAL ANALYSIS OF THE LINEAR REGRESSION EQUATION

Klyuzhev N.A.[1], Alekseeva A.M.[2]

[1] Klyuzhev Nikolay Alexeyevich - Candidate of Technical Sciences, Associate Professor

[2]Alekseeva Alla Mikhailovna - Candidate of Economic Sciences, Associate Professor

Pskov, Russian Federation

Abstract: *the application of spectral analysis to identify the "linear bias" in the estimates of the coefficients of the multidimensional and linear regression equation, which affects the interpretation of regression parameters in the practice of using this mathematical model, is considered.*

Keywords: *bilinear metric, polynomial, parameters, spectrum, matrix, examples.*

УДК 330.43

Математический метод в анализе протекающих в экономических объектах процессов вскрывает в них логическую основу, что способствует структурированию и построению статистических моделей. Значительную роль играют простые внешне, но логически и статистически фундаментально обоснованные модели. Они позволяют «анатомировать» системную зависимость в данных, а гипотеза о линейности значительно упрощает расчёты.

Интерпретация коэффициентов в модели линейной по параметрам регрессии как «чистых» коэффициентов пропорциональности ошибочно в случае наличия «тесной» статистической связанности факторов. Этот делает актуальным вопрос о новых подходах к оцениванию параметров многомерной регрессии [1].

Известно, что разложение функции в ряд по системе ортогональных функций называют рядом Фурье. Обнаружение такой системы для построения ряда Фурье называют *спектральным анализом*. Применяют его во всех случаях, когда даны массив данных и оценка (estimation) наличия определённого объекта или свойства в данных.

Актуально для практики, с этой точки зрения*,* решение проблем структурирования взаимосвязи факторов линейной регрессии, постановки и решения прямой и обратной задач прогнозирования приращений средних значений факторов при теоретически обоснованной интерпретации параметров [2].

Цель исследования: разработать алгоритм конструирования спектральной модели многомерной регрессии, линейной по параметрам.

Метод исследования: построение ортогональной системы полиномов по эмпирическим значениям факторов регрессии с использованием ППП *Excel.*

Рассмотрим детально метод исследования, пояснив его простыми примерами с выполнением вычислений в ППП *Excel*.

Уравнение регрессии с оценёнными по МНК параметрами, например,

$$\hat{y}_i = a_1 x_{1i} + a_2 x_{2i} + a_3 x_{3i}\ ,\ \ i = \overline{1, n}. \qquad (1)$$

можно записать в форме:

$$\hat{y}_i = c_1 f_1(x_{1i}) + c_2 f_2(x_{1i}, x_{2i}) + c_3 f_3(x_{1i}, x_{2i}, x_{3i}),\ \ i = \overline{1, n} \qquad (2)$$

где $\hat{y}_i -$ зависимая переменная (результативный признак); $x_1, x_2, x_3 -$ факторы, имеющие нулевые средние значения.

Назовём уравнение (1) классической формой линейной регрессии, уравнение (2) – его спектральной формой, уравнения (1) и (2) дуальными, функции $f -$ спектральные функции, ортогональные со скалярным произведением $(f_i, f_j) = 0,\quad i \neq j,$ имеют нулевые средние значения и их выражения представлены:

$$f_1(x_1) = x_1; f_2(x_1, x_2) = v_{12}x_1 + x_2; f_3(x_1, x_2, x_3) = v_{13}x_1 + v_{23}x_2 + x_3. \quad (3)$$

Из равенства значений дуальных уравнений регрессии следует равенство

$$a_1x_1 + a_2x_2 + a_3x_3 = c_1x_1 + c_2(v_{12}x_1 + x_2) + c_3(v_{13}x_1 + v_{23}x_2 + x_3), \quad (4)$$

где c_k –к-й спектральный коэффициент, v_{ij} –коэффициенты.

Пусть нас интересует приращение $\Delta\hat{y}$, обусловленное только приращением Δx_1. С учётом статистической взаимосвязи факторов следует положить, что должны выполняться равенства:

$$\Delta\hat{y} = c_1\Delta x_1;$$
$$f_2(\Delta x_1, \Delta x_2) = v_{12}\Delta x_1 + \Delta x_2 = 0; \quad (5)$$
$$f_3(\Delta x_1, \Delta x_2, \Delta x_3) = v_{13}\Delta x_1 + v_{23}\Delta x_2 + \Delta x_3 = 0.$$

Считая приращение Δx_1 заданным, решаем систему уравнений (5):

$$\Delta\hat{y} = c_1\Delta x_1;\ \Delta x_2 = -v_{12}\Delta x_1;\ \Delta x_3 = -(v_{13} - v_{12} \cdot v_{23})\Delta x_1. \quad (5а)$$

Решения (5а) являются оценками приращений результата $\Delta\hat{y}$ и факторов Δx_2, Δx_3. Коэффициенты в равенствах (5) и (5а) отражают структуру связи приращений и образуют матрицу

$$\boldsymbol{V} = \begin{bmatrix} 1 & v_{12} & v_{13} \\ 0 & 1 & v_{23} \\ 0 & 0 & 1 \end{bmatrix}. \quad (6)$$

Матрица $\boldsymbol{V}$ формируется в процессе перехода от ортонормированного единичного базиса $\langle \boldsymbol{e}_1, \boldsymbol{e}_2, \boldsymbol{e}_3 \rangle$ в линейном пространстве к базису $\langle \boldsymbol{f}_1, \boldsymbol{f}_2, \boldsymbol{f}_3 \rangle$ по алгоритму Грама-Шмидта с билинейной метрикой $(\boldsymbol{f}_i, \boldsymbol{f}_j) = (\mathbf{v}_i, \boldsymbol{COV} \cdot \mathbf{v}_j)$, $i \neq j$, где $\boldsymbol{COV}$ – симметрическая и положительно определённая матрица ковариаций факторов, адекватная матрице системы нормальных уравнений в МНК, представляет информацию о системности переменных. В итоге получим формулу, которая фиксирует связь между параметрами дуальных уравнений (1) и (2) [1]:

$$\boldsymbol{a} = \boldsymbol{V} \cdot \boldsymbol{c} \quad \text{или} \quad \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} = \begin{bmatrix} 1 & v_{12} & v_{13} \\ 0 & 1 & v_{23} \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \\ c_3 \end{bmatrix}. \quad (7)$$

Применение формулы (7) к уравнению регрессии в форме (1) преобразует его в уравнение в спектральной форме

$$\widehat{\boldsymbol{Y}} = \boldsymbol{X} \cdot \boldsymbol{a} = (\boldsymbol{XV})\boldsymbol{c} = \boldsymbol{F} \cdot \boldsymbol{c}. \quad (7a)$$

Применение формул (7) и (7а) требует построения матриц $\boldsymbol{V}$ и $\boldsymbol{F}$.

Рассмотрим поясняющий простой алгоритм построения этих матриц с помощью ППП *Excel*. Выполним произведение матриц в формуле (7):

$$a_1 = c_1 + v_{12}c_2 + v_{13}c_3; \; a_2 = c_2 + v_{23} \cdot c_3; \; a_3 = c_3. \quad (8)$$

Равенства (8) показывают, как формируются «отклонения» в параметрах уравнения (1), а также на структуру алгоритма, поскольку коэффициент регрессии при последнем факторе, вводимом в модель (1) регрессии, равен спектральному коэффициенту дуальной модели (2).

Далее введем обозначения для коэффициентов регрессии классической модели: $\mathrm{a}_k^{(p)}$, где $k-$ номер коэффициента регрессии, m число факторов в модели, $k = 1, 2, \dots, m$. Тогда имеем равенства:

$$\begin{aligned} &\text{для } m = 1: \quad \mathrm{a}_1^{(1)} = c_1, \\ &\text{для } m = 2: \quad \mathrm{a}_1^{(2)} = c_1 + v_{12}c_2; \;\; \mathrm{a}_2^{(2)} = c_2, \qquad (9) \\ &\text{для } m = 3: \quad \mathrm{a}_1^{(3)} = c_1 + v_{12}c_2 + v_{13}c_3; \;\; \mathrm{a}_2^{(3)} = c_2 + v_{23} \cdot c_3, a_3 = c_3. \end{aligned}$$

Из уравнений (9) следует, что, оценив по МНК коэффициент регрессии однофакторного уравнения регрессии, получаем оценку первого спектрального коэффициента, т.е. $c_1 = \mathrm{a}_1^{(1)}$. Далее вводим в модель второй фактор и оцениваем его коэффициенты регрессии: $\mathrm{a}_1^{(2)}$и $\mathrm{a}_2^{(2)}$. Теперь получаем оценку второго спектрального коэффициента, т. е.$c_2 = \mathrm{a}_2^{(2)}$. Используя полученные оценки, находим элемент матрицы $\boldsymbol{V}$:

$$v_{12} = (\mathrm{a}_1^{(2)} - \mathrm{a}_1^{(1)})/\mathrm{a}_2^{(2)} \,. \quad (10)$$

Далее строим трехфакторное уравнение регрессии и оцениваем его коэффициенты регрессии. В результате получаем оценки:

$$c_3 = \mathrm{a}_3^{(3)}; \; v_{13} = (\mathrm{a}_1^{(3)} - \mathrm{a}_1^{(2)})/\mathrm{a}_3^{(3)}; \; v_{23} = (\mathrm{a}_2^{(3)} - \mathrm{a}_2^{(2)})/\mathrm{a}_3^{(3)}. \quad (11)$$

Обобщая полученные формулы, получаем общую формулу

$$v_{kp} = (\mathrm{a}_k^{(p)} - \mathrm{a}_k^{(p-1)})/\mathrm{a}_p^{(p)}, \quad (12)$$

где k номер строки, p номер столбца матрицы $\boldsymbol{V}$.

Пример 1. В ковариационной матрице $\boldsymbol{COV}$ представлена информация о центрированных значениях темпов роста цен (в %) производителей промышленных товаров за 15 лет: $\boldsymbol{y}_0$ — промышленные товары (результат); $\boldsymbol{x}_1$ — добыча полезных ископаемых, $\boldsymbol{x}_2$ — обрабатывающие производства, $\boldsymbol{x}_3$ —производство и распределение электроэнергии, газа и воды (факторы):

$$\mathbf{COV} = \begin{array}{c} x_1 \\ x_2 \\ x_3 \\ y_0 \end{array} \begin{bmatrix} 1332{,}773 & 416{,}517 & 61{,}761 & 556{,}123 \\ 416{,}517 & 245{,}558 & 8{,}240 & 245{,}136 \\ 61{,}761 & 8{,}240 & 90{,}625 & 29{,}872 \\ 556{,}123 & 245{,}136 & 29{,}872 & 278{,}562 \end{bmatrix}.$$

Уравнение трёхфакторной регрессии темпа роста $\boldsymbol{y}_0$ (промышленные товары) представлено уравнением

$$\widehat{\boldsymbol{y}}_0 = a_1\boldsymbol{x}_1 + a_2\boldsymbol{x}_2 + a_3\boldsymbol{x}_3, \qquad (13)$$

где оценки коэффициентов a_1, a_2, a_3 получаем, решая систему уравнений:

$$\begin{bmatrix} 1332{,}773 & 416{,}517 & 61{,}76 \\ 416{,}517 & 245{,}558 & 8{,}240 \\ 61{,}76 & 8{,}240 & 90{,}625 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} = \begin{bmatrix} 556{,}123 \\ 245{,}136 \\ 29{,}872 \end{bmatrix}. \qquad (13)$$

Вычислены оценки: $a_1 = 0{,}2144$; $a_2 = 0{,}6303$; $a_3 =$0,1262.

Для построения матрицы парных коэффициентов регрессии разделили элементы каждой строки матрицы $\mathbf{COV}$ на её диагональный элемент. В результате получим матрицу, составленную из коэффициентов парных регрессий b_{ij}:

$$\boldsymbol{B} = \begin{array}{c} x_1 \\ x_2 \\ x_3 \\ y_0 \end{array} \begin{bmatrix} b_{11} & b_{12} & b_{13} & b_{10} \\ b_{21} & b_{22} & b_{23} & b_{20} \\ b_{31} & b_{32} & b_{33} & b_{30} \\ b_{01} & b_{02} & b_{03} & b_{00} \end{bmatrix} = \begin{bmatrix} 1 & 0{,}3125 & 0{,}0463 & 0{,}4173 \\ 1{,}6962 & 1 & 0{,}0336 & 0{,}9983 \\ 0{,}6815 & 0{,}0909 & 1 & 0{,}3296 \\ 1{,}9964 & 0{,}8800 & 0{,}1072 & 1 \end{bmatrix}. \qquad (15)$$

Коэффициент b_{ij} равен коэффициенту линейной регрессии (проекции) переменной с индексом j на переменную с индексом i: $\hat{x}_2 = b_{12} * x_1 = 0{,}3125 * x_1$.

Элементы последнего столбца равны коэффициентам регрессии (проекции) показателя y_0 на факторы, например, если $\hat{y}_{10} = b_{10}x_1$, то $\Delta\hat{y}_{10} = 0{,}4173 \cdot \Delta x_1$, т.е. по приращению Δx_1 можно статистически оценить приращения всех остальных факторов и приращение результативного признака. Подстановка в уравнение вида (1) произвольного набора приращений факторов без учёта их статистических связей есть «грубая оплошность» на практике.

Рассмотрим две задачи полезные на практике.

i). *Прямая задача*. Матрица $\boldsymbol{B}$ содержит решение задачи, например, когда выбран доминирующий фактор x_2 - обрабатывающие производства. Он имеет наибольший коэффициент регрессии с результатом y_0. Решение непосредственно считывается как наибольшее значение приращения результата, равное 0,9983 при $\Delta x_2 = 1$ и совпадающее с парным коэффициентом b_{20} регрессии. В этой же строке для остальных факторов даны приращения, вызванные приращением фактора $\Delta x_2 = 1$. Проверка:

0,2144$*$ 1,6962+0,6303 $*$ 1 +0,1262$*$0,0326=0,9983.

Линейность модели позволяет утверждать, что увеличение приращения фактора x_2, например, в 2 раза, во столько же раз изменит приращения результата и остальных факторов относительно их средних значений.

ii). *Обратная задача*. Решение обратной задачи, когда по прогнозному значению приращения целевого показателя $y_0 = 1$ надо оценить получаемые приращения факторов, находится в последней строке матрицы $\boldsymbol{B}$: $x_1 = 1{,}9964$; $x_2 = 0{,}8800$; $x_3 = 0{,}1072$. Вычислили с погрешностью 0,37% по уравнению (1):

0,2144 $*$ 1,9964+0,6303 $*$ 0,88 +0,1262$*$ 0,1072 = 0,9963.

iii) Для построения матриц $\boldsymbol{V}$ и $\boldsymbol{F}$ используем данные и, применяя функцию «*РЕГРЕССИЯ*», строим уравнения регрессии последовательно с одним фактором, далее с двумя факторами и тремя факторами. Данные заданы табл. 1 и они центрированы.

По формулам (10 – 11) и расчётными данными составлена таблица 3.

Таблица 1. Центрированные данные примера 1 (%)					Таблица 2. Ортогональный базис		
x_1	x_2	x_3	y_0	z_0	f_1	f_2	f_3
-29,147	14,347	-13,826	0,806	18,2	-29,147	23,456	-10,227
97,353	50,477	1,704	52,066	21,8	97,353	20,053	-0,885
21,253	7,687	25,574	13,336	7,4	21,253	1,045	24,689

-23,847	-10,803	11,384	-10,254	35	-23,847	-3,350	12,168
-2,047	-3,883	10,004	-0,934	-7,7	-2,047	-3,243	9,788
-26,047	-1,283	-1,556	-6,044	-10,5	-26,047	6,857	0,308
36,853	4,377	-3,556	10,246	-3,5	36,853	-7,140	-5,948
3,153	-8,933	-3,416	-5,234	-1,3	3,153	-9,918	-4,513
-26,247	-3,733	-5,706	-8,214	-1,4	-26,247	4,470	-4,061
24,453	0,867	-2,746	6,486	-8,6	24,453	-6,775	-4,529
-66,247	-15,153	1,994	-25,584	1,6	-66,247	5,551	5,596
21,353	-11,193	2,284	-4,734	-10,8	21,353	-17,866	-0,418
-10,747	-0,173	-2,186	-1,924	-18	-10,747	3,186	-1,383
-1,547	-8,753	-10,916	-6,574	-9,4	-1,547	-8,269	-11,637
-18,547	-13,853	-9,036	-13,444	-12,8	-18,547	-8,056	-8,949

Таблица 3. Параметры регрессии при последовательном введении факторов в уравнение регрессии		
m=1	*коэф-ты.*	*t-критерий*
x1 a11=c1	0,4173	8,053578
m=2		
x1 a12	0,2241	12,51609
x2 a22=c2	0,6182	14,82358
m=3		
x1 a13	0,2144	17,195451
x2 a23	0,6303	22,0136
x3 a33=c3	0,12618	3,8431142

По формулам (10-11) и данным табл. 3 вычислены элементы матрицы $\boldsymbol{V}$:

$$\boldsymbol{V} = \begin{bmatrix} 1 & v_{12} & v_{13} \\ 0 & 1 & v_{23} \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & -0{,}31252 & -0{,}07630 \\ 0 & 1 & 0{,}09586 \\ 0 & 0 & 1 \end{bmatrix}.$$

По формуле $\boldsymbol{F} = \boldsymbol{X} \cdot \boldsymbol{V}$, где матрица $\boldsymbol{X}$ представлена первыми тремя столбцами в табл. 2, вычислена матрица $\boldsymbol{F}$, составленная из трёх столбцов, которые образуют ортогональный базис в пространстве факторных переменных.

Спектральные коэффициенты (спектр) уравнения регрессии (2) табл. 3 вычисляются по формуле:

$$c_i = \frac{(\boldsymbol{f}_i, \mathbf{Y}_0)}{\|\boldsymbol{f}_i\|^2} = \frac{1}{\|\boldsymbol{f}_i\|^2} \cdot (\boldsymbol{f}_i, \mathbf{Y}_0), \ (i = 1, 2, \dots, m), \qquad (16)$$

где $\|\boldsymbol{f}_i\| = \sqrt{\frac{1}{n}\sum_{i=1}^{n} f_i^2}$ − евклидова норма (длина) вектора $\boldsymbol{f}_i$.

Спектральный коэффициент интерпретируем как коэффициент пропорциональности в формуле для коэффициента корреляции между $\boldsymbol{f}_i$ и $\mathbf{Y}_0$:

$$c_i \frac{\|\mathbf{Y}_0\|}{\|\boldsymbol{f}_i\|} = r_{f_i y_0} = \cos(\varphi_i), \ (i = 1, 2, \dots, m), \qquad (17)$$

где φ_i −угол между векторами $\boldsymbol{f}_i$ и $\mathbf{Y}_0$. Если все данные нормированы, то $c_i = r_{f_i y_0} = \cos(\varphi_i)$.

<u>Пример 2.</u> В табл. 1 (см. пример1) в столбце z_0 представлены центрированные данные о темпах роста производства, распределения и потребления электроэнергии. Две спектральные модели регрессии записываются в стандартизованном масштабе, когда все переменные нормированы и спектры оценены относительно одного и того же нормированного базиса (см. табл. 3):

$$t_{y0} = c_{1y}t_1 + c_{2y}t_2 + c_{3y}t_3, \quad t_{z0} = c_{1z}t_1 + c_{2z}t_2 + c_{3z}t_3, \qquad (18)$$

где $t_{y0} = \frac{y_0}{\|\mathbf{Y}_0\|}$, $t_{z0} = \frac{z_0}{\|\mathbf{Z}_0\|}$, $c_{iy} = r_{f_i y_0}$, $c_{iz} = r_{f_i z_0}$, $t_i = \frac{x_i}{\|\boldsymbol{X}_i\|}$, $i = \overline{1, n}$.

Теснота совместного влияния факторов на результаты оценивают коэффициентами множественной корреляции:

$$R_{y_0 f_1 f_2 f_3} = \sqrt{\sum r_{f_i y_0}^2}, \quad R_{z_0 f_1 f_2 f_3} = \sqrt{\sum r_{f_i z_0}^2}\,. \qquad (19)$$

Для оценки этих коэффициентов достаточно оценить матрицу корреляций нормированных переменных:

	f_1	f_2	f_3	y_0	z_0
f_1	1				
f_2	0	1			
f_3	0	0	1		
y_0	0,912707	0,397895	0,070394	1	
z_0	0,129003	0,489370	0,330946	0,326967	1

Качество в целом построенных уравнений спектральной регрессии оценивают коэффициентами детерминации: $R^2_{y_0 f_1 f_2 f_3} = 0{,}99631$; $R^2_{z_0 f_1 f_2 f_3} = 0{,}36565$.

Косинус угла φ между векторами спектров переменных y_0 и z_0 может служить оценкой их «коллинеарности»:

$$cos(\varphi) = \frac{(\boldsymbol{c}_y, \boldsymbol{c}_z)}{\|\boldsymbol{c}_y\| \cdot \|\boldsymbol{c}_z\|} = \frac{0{,}335756}{0{,}9982 * 0{,}6047} = 0{,}556245;\ \varphi = 56{,}2^{\circ}.$$

где $\|\boldsymbol{c}_y\| = R_{y_0 f_1 f_2 f_3} = 0{,}9982$; $\|\boldsymbol{c}_z\| = R_{z_0 f_1 f_2 f_3} = 0{,}6047$;

$(\boldsymbol{c}_y, \boldsymbol{c}_z)$ =0,912707*0,129003+0,397895*0,489370+0,070394*0,330946=0,335756.

Таким образом, две различные спектральные модели регрессии, спектры которых оценены относительно одного и того же ортонормированного базиса пространстве факторов, решают задачу выявления максимальных связей между двумя характеристиками, например, эффективности производства, относительно одной и той же группы факторов производства. При этом у каждой индивидуальной спектральной модели спектральные коэффициенты имеют свои равные стандартные ошибки и различные значения t-статистик, что дополняет аналитическое сравнение двух результативных переменных.

Выводы.

Разработан алгоритм конструирования дуальных уравнений многомерной регрессии, линейной по параметрам.

На примерах продемонстрирован с применением квадратичной метрики синтез ортогональных систем полиномов порядка не выше первого.

Наиболее простую форму имеют уравнения регрессии, если все переменные нормируются и спектральные коэффициенты равны коэффициентам корреляции в стандартизированной форме уравнения регрессии.

Задача о выявлении максимальных связей между двумя характеристиками относительно одной и той же группы факторов совпадает с задачей поиска канонических корреляций и не требует вычисления собственных значений положительно определённой матрицы и её собственных векторов.

Методика спектрального анализа реализуется с применением ППП *Excel* и рекомендуется в программы по дисциплине «Эконометрика».

Список литературы

1. Клюжев Н.А. Спектральный анализ регрессионных эконометрических моделей.// ВЕСТНИК ИНЖЭКОНА. 2007. Вып.4(17). Серия «ЭКОНОМИКА». - с. 219-226
2. Алексеева А.М., Клюжев Н А. Системно-обобщённая модель линейной регрессии в экономическом прогнозировании.//Наука и образование: проблемы и тенденции развития: материалы V Международной научно-практической конференции (Уфа, 29-30 декабря 2017 г.) – Уфа: РИО ИЦИПТ, 2017.- с. 138-144