МАТЕМАТИЧЕСКИЕ МЕТОДЫ В ЕСТЕСТВЕННЫХ, ТЕХНИЧЕСКИХ И СОЦИАЛЬНЫХ НАУКАХ

УДК 311

В.П. Колпашников, Д.Е. Красильников

О ПОСТРОЕНИИ ИНТЕРВАЛА РАЗБРОСА ЗНАЧЕНИЙ ЭКОНОМИЧЕСКИХ ПОКАЗАТЕЛЕЙ, ПОЛУЧЕННЫХ ПО ЛИНИИ РЕГРЕССИИ

ООО «Смоуквент», г. Нижний Новгород

В статье рассмотрена методика построения доверительного интервала для системы внешне несвязанных уравнений на основе теоремы Гаусса-Маркова.

Ключевые слова: технически сложные системы, системы регрессионных уравнений, стоимость эксплуатации.

В экономической литературе часто решается задача о нахождении экономических величин по линии регрессии. В качестве примера можно привести работы [1]; [4, с. 228-243]; [7].

Тем не менее, эту задачу можно еще более обобщить, построив доверительный интервал для этой величины, поскольку маловероятно, что ее значения будут в точности описаны найденным уравнением. Сделать это можно на основании теоремы Гаусса-Маркова[1] [10, с. 41-46, 69-72].

В качестве примера рассмотрим систему регрессионных уравнений, полученную в статье [7], в которой рассмотрена методика оценки стоимости эксплуатации технически сложной системы на основе линии регрессии.

$$\begin{cases} \hat{x}_2 = 18472{,}73 + 0{,}186x_3 \\ \hat{x}_4 = 5976{,}632 + 1{,}016x_1 + 1{,}017x_3 \end{cases} \quad (1)$$

Таблица 1

Переменные

X(1)	Входящие материалы
X(2)	Входящие работы
X(3)	Работы
X(4)	Выходящие материалы

[1] Although known as the Gauss-Markov theorem the least squares approach of Gauss antedates (1821) the minimum-variance approach of Markov (1900). [12, глава 3.4] «Несмотря на то, что эта теорема называется теоремой Гаусса-Маркова, подход наименьших квадратов был рассмотрен Гауссом в 1821 году, а подход наименьшей вариации – Марковым в 1900 году» (перевод Красильникова Д.Е.).

Таблица 2

Исходная информация по проектам в рублях

Входящие материалы	Входящие работы	Работы	Выходящие материалы
1 095 000,00	40 000,00	128 140,00	1 157 913,41
1 117 138,00	97 000,00	499 800,00	1 999 200,00
294 762,92	22 000,00	69 456,40	303 462,80
1 171 619,00	150 000,00	667 663,51	1 620 931,00
422 501,20	70 000,00	175 388,20	682 706,00

Такие системы уравнений (1) в экономической теории и математической статистике называется системами регрессионных уравнений. Они традиционно входят в перечень тем, включенных в большинство учебников по эконометрике. Для ознакомления с этой проблемой можно порекомендовать статью [6].

Рассматриваемая система (1) относится к типу внешне не связанных между собой уравнений (Seemingly Unrelated Regression, SUR) [10, с. 220-223]. Они связаны между собой лишь благодаря наличию слабой корреляции (коэффициент корреляции Пирсона менее 0,5) [4, с. 205] между остатками[2] в разных уравнениях. Убедимся в этом.

Для этого составим табл. 3 и табл. 4. В первых столбцах этих таблиц даны значения независимых переменных: для регрессии Входящих работ на Работы – Работы (табл. 3); для регрессии Выходящих материалов на Входящие материалы и Работы - Входящие материалы и Работы (табл. 4). В следующем столбце даются значения зависимых переменных по выборке: для регрессии Входящих работ на Работы – Входящие работы (табл. 3); для регрессии Выходящих материалов на Входящие материалы и Работы – Выходящие материалы (табл. 4). В следующий столбец (X(2) – табл. 3 и X(4) – табл. 4) заносятся значения зависимых переменных, полученные по системе уравнений. Последний столбец (остаток – e) представляет собой разность между значениями, полученными по выборке и регрессии.

Таблица 3

Расчет остатка для регрессии Входящих работ на Работы

X(3)	X(2)	X(2)=18472,73+0,186*X(3)	e
128 140,00	40 000,00	42306,77	-2 306,77
499 800,00	97 000,00	111435,53	-14 435,53
69 456,40	22 000,00	31391,6204	-9 391,62
667 663,51	150 000,00	142658,1429	7 341,86
175 388,20	70 000,00	51094,9352	18 905,06

Таблица 4

Расчет остатка для регрессии Выходящих материалов на Входящие материалы и Работы

X(1)	X(3)	X(4)	X(4)=5976,632+1,016*X(1)+1,017*X(3)	e
1 095 000,00	128 140,00	1 157 913,41	1248815,01	90 901,60
1 117 138,00	499 800,00	1 999 200,00	1649285,44	-349 914,56
294 762,92	69 456,40	303 462,80	376092,92	72 630,12
1 171 619,00	667 663,51	1 620 931,00	1875355,33	254 424,33
422 501,20	175 388,20	682 706,00	613607,65	-69 098,35

[2] Под остатком понимается разность между значением, полученным по выборке, и значением, полученным по регрессии.

Вычислив коэффициент корреляции Пирсона между остатками из табл. 3 и табл. 4, получили 0,39, что свидетельствует о слабой корреляции остатков, а, следовательно, внешней не связанности уравнений. Другими словами, каждое отдельное уравнение в (1) удовлетворяет условиям классической регрессионной модели и может быть оценено обычным методом наименьших квадратов.

С точки зрения экономической теории, это означает, что мы имеем двухуровневый процесс производства: часть работ отдается на аутсорсинг нижней фирме, а другая производится организацией, которую мы изучаем, - верхней фирмой[3].

Если бы это было не так, то мы бы получили не систему внешне не связанных между собой уравнений (Seemingly Unrelated Regressions, SUR), а систему одновременных уравнений (Simultaneous Equations), которую не так просто идентифицировать, поскольку не ясно в какое уравнение, какая независимая переменная входит. Для того чтобы ее составить, необходимо сначала идентифицировать уравнения – выяснить какая независимая переменная к какому уравнению относится. Наиболее известным правилом для идентификации систем одновременных уравнений является так называемое порядковое условие идентификации (Order Condition for Identification):

«Необходимым условием для идентификации уравнения является наличие ограничений на переменные, в него входящие, в количестве, как минимум, на единицу меньше, чем число самих уравнений в системе. Эти ограничения могут быть очень простыми. Например, некоторые переменные в системе уравнений не встречаются в изучаемом нами уравнении» (перевод Красильникова Д.Е.)[4].

Следует отметить, что системы одновременных уравнений являются малоизученной областью регрессионного анализа. По этой причине исследователь, столкнувшийся с такой системой, неизбежно будет вынужден использовать нестандартные методы оценивания, что приведет к усложнению математических расчетов.

Убедившись в том, что уравнения, описывающие стоимость эксплуатации технически сложных систем внешне не связаны, можно использовать классические методы прогнозирования, основанные на теореме Гаусса-Маркова.

Итак, у нас есть проект, для которого стоимость Работ (X(3)) оценивается в 420 000 рублей, а стоимость Входящих материалов в 615 000. Для того чтобы сказать можно ли рассчитать прогнозную стоимость Выходящих материалов (стоимость эксплуатации технически сложной системы), необходимо сначала проверить, входят ли данные значения в интервалы для наименьшего и наибольшего значения соответствующих переменных (табл. 5). Если они в эти интервалы не входят, то прогноз по системе регрессионных уравнений сделать нельзя, поскольку данные относятся к другой статистической совокупности (массиву данных). В нашем случае значения независимых переменных входят в интервалы для них.

Таблица 5

Размах вариации статистической совокупности (рубли)

Переменная	Название	Минимальное значение	Среднее (x_j)	Максимальное значение
X(1)	Входящие материалы	294 762,92	820 204,2	1 117 138
X(2)	Входящие работы	22 000	75 800	150 000
X(3)	Работы	69 456,4	308 089,6	667 663,51

[3] Понятия верхняя и нижняя фирма берут свое начало в Средневековье. Тогда под нижней фирмой понимали фирму, рубящую лес (обычно она располагалась ниже по течению), а под верхней – фирму, использующей этот лес в своей деятельности (она располагалась выше по течению).

[4] «A necessary condition for an equation to be identified is that the total number of restrictions placed on its parameters should be at least as great as the number of equations in the model less one. The restrictions referred to in the rule may be of a very simple kind, merely implying some variable in the model does not appear in the equation of interest…» [14, с. 218].

X(4)	Выходящие материалы	303 462,8	1 152 843	1 999 200

Сначала выясним, какую сумму денег придется отдать сторонним организациям для выполнения подряда. Для этого в первое уравнение (1) подставим значение переменной Работы (X(3)):

$$\hat{x}_2 = 18472{,}73 + 0{,}186 x_3 = 18472{,}73 + 0{,}186 * 420000 = 96592{,}73.$$

В качестве проверки точности расчетов можно посмотреть входит ли значение зависимой переменной в интервал между наибольшим и наименьшим значением для нее, имеющимся в статистической совокупности. Если полученное значение не входит в него, то расчет произведен не верно.

Поскольку прогнозное значение считается по ограниченной совокупности, то необходимо построить доверительный интервал для него. Методика построения доверительных интервалов для зависимой переменной в парной регрессии в отечественной литературе рассматривается весьма поверхностно. В большинстве источников приводится лишь формула (часто с ошибками) для его построения. Для более детального знакомства с проблемой следует читать английские работы, например [13, с. 85-88, 101-103]; [14, с. 151-153] и другие. Из работ на русском языке наиболее полно эта проблема освещена в книге [9]. Из современных изданий можно порекомендовать [3, гл. 4.3].

Формула для вычисления длины доверительного интервала для парной регрессии имеет вид:

$$z = t_{n-2;\frac{\alpha}{2}} \sigma \sqrt{1 + \frac{1}{n} + \frac{(x_3 - \bar{x}_3)^2}{\sigma_{x_3}^2 n}} \qquad (2)$$

где $t_{n-2;\frac{\alpha}{2}}$ – величина, получаемая по таблице процентных точек распределения Стьюдента с n–2 степенями свободы (на две меньше, чем число элементов в выборке) и двусторонней критической областью ($\frac{\alpha}{2}$). «За величину α принято брать 5% в качестве вероятности для отклонения гипотезы. Тем не менее, можно использовать и другие значения этого параметра. Теория тестирования гипотез с общепринятым уровнем значимости 0,05 и 0,01 была создана известным английским статистиком сэром Р.А. Фишером (1890-1962). Он считается отцом основателем современных статистических методов и числа 0,05 и 0,01, предложенные им, используются по всему миру» (перевод Красильникова Д.Е.)[5].

Другой причиной, по которой следует использовать именно 5%, является факт, что при переходе от 0,05 к 0,01 оценка вероятности становится крайне ненадежной [11, с. 96].

В нашем случае $t_{n-2;\frac{\alpha}{2}} = t_{5-2;\frac{5\%}{2}} = 3{,}1824$ - [2, табл. 3.2].

n – число элементов в выборке. В нашем случае n=5

$\sigma = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n-2}}$ – ошибка функции. Вычисляется как корень квадратный из суммы квадратов остатков, деленный на число элементов в выборке минус два. На основе данных табл. 3 сумма квадратов остатков равна 713 212 589,31 ($\sum_{i=1}^{n} e_i^2 = 713\,212\,589{,}31$). Тогда ошибка функ-

[5] Although it is customary to use the 5% probability level for rejection of the suggested hypothesis, there is nothing sacred about this number. The theory of significance tests with the commonly used significance levels of 0,05 and 0,01 owes its origins to the famous British statistician Sir R. A. Fischer (1890-1962). He is considered the father of modern statistical methods and the numbers 0,05 and 0,01 suggested by him have been adopted universally [13, с. 81].

ции составит

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n-2}} = \sqrt{\frac{713212589{,}31}{5-2}} \approx 15418$$

X(3) – значение переменной Работы равное 420 000

$\bar{x}_3 = \frac{\sum_{i=1}^{n} x_{3i}}{n}$ – среднее значение переменной работы (определяется как сумма элементов в столбце Работы табл. 2, деленная на их число). Из таблицы 5 видно, что оно составляет 308 089,6.

$\sigma_{x_3}^2 = \frac{\sum_{i=1}^{n} (x_{3i} - \bar{x}_3)^2}{n}$ - дисперсия элементов в столбце Работы табл. 2. Определяется как сумма квадратов отклонений от средней, деленное на число элементов в столбце Работы. Исходя из данных таблицы 2, дисперсия равна 68 245 899 616.

Таким образом, длина доверительного интервала будет равна:

$$z = t_{n-2;\frac{\alpha}{2}} \sigma \sqrt{1 + \frac{1}{n} + \frac{(x_3 - \bar{x}_3)^2}{\sigma_{x_3}^2 n}} =$$

$$= 3{,}1824 * 15418 *$$

$$* \sqrt{1 + \frac{1}{5} + \frac{(420000 - 308089{,}6)^2}{68245899616 * 5}} \approx$$

$$\approx 54565.$$

Чтобы получить сам доверительный интервал для переменной Входящие работы, необходимо вычесть его длину из полученного по регрессии значения (нижняя граница доверительного интервала) и прибавить его (верхняя граница доверительного интервала).

Нижняя граница: 96592,73-54565=42027,73 рублей.

Верхняя граница: 96592,73+54565=151157,73 рублей.

Поскольку верхняя граница выходит за максимальное значение переменной Входящие работы, то за нее следует считать ее максимальное значение, то есть 150 000.

Таким образом, проект, затраты работ на который оцениваются в 420 000 рублей, потребует передачи суммы от 42 027,73 до 150 000 рублей сторонним организациям. Наиболее вероятно, что эта сумма составит 96 592,73 рубля.

В принципе длину доверительного интервала можно значительно уменьшить. Так, если посмотреть (2), то можно заметить, что величина $\sqrt{1 + \frac{1}{n} + \frac{(x_3 - \bar{x}_3)^2}{\sigma_{x_3}^2 n}} \approx 1$, если стоимость Работ, изучаемого проекта достаточно близка к средней (наиболее типичной стоимости). Значение t-статистики ($t_{n-2;\frac{\alpha}{2}}$) падает с увеличением числа элементов в выборке. Так для 7 элементов оно составит 2,5706; 12 – 2,2281; 57 – 2,0003 и т. д. Придел t-статистики – 1,96. Таким образом, длина доверительного интервала в пределе равна:

$$z = t_{n-2;\frac{\alpha}{2}}\sigma\sqrt{1+\frac{1}{n}+\frac{(x_3-\bar{x}_3)^2}{\sigma_{x_3}^2 n}} =$$

$$= 1{,}96 * \sigma \approx 1{,}96 * 15418 \approx 30000.$$

Другими словами, почти в два раза меньше, чем получено по исходным данным, но, к сожалению, на предприятии ООО «Смоуквент», на котором были собраны данные, в полугодие выполняется лишь 5-8 проектов. По этой причине длина доверительного интервала получается сильно завышенной.

После того как мы рассчитали сумму, которую придется отдать сторонним организациям, мы можем рассчитать стоимость Выходящих материалов (стоимость обслуживания технически сложной системы). Для этого, подставим исходные данные о проекте во второе уравнение (1):

$$\hat{x}_4 = 5976{,}632 + 1{,}016x_1 + 1{,}017x_3 =$$

$$= 5976{,}632 + 1{,}016 * 615000 + 1{,}017 * 420000 = 1057956{,}63$$

Это число входит в интервал между наибольшим и наименьшим значением переменной Выходящие материалы, то есть считается верным. Для того чтобы получить доверительный интервал для него, нужно составить расширенную матрицу объясняющих переменных (X) [10, стр. глава 7] – она получается добавлением столбца единиц к значениям в столбцах Входящие материалы (X(1)) и Работы (X(3)) табл. 2:

$$X = \begin{bmatrix} 1 & 1095000 & 128140 \\ 1 & 1117138 & 499800 \\ 1 & 294762{,}92 & 69456{,}4 \\ 1 & 1171619 & 667663{,}51 \\ 1 & 422501{,}2 & 175388{,}2 \end{bmatrix}$$

Данные о проекте преобразуются в векторный вид:

$$\vec{x}_0 \{1; 615000; 420000\}$$

После этого нужно решить матричное уравнение, описывающее длину доверительного интервала для зависимой переменной в многомерной регрессии:

$$z = t_{\frac{\alpha}{2};n-k-1}\sigma\sqrt{1+\vec{x}_0(x^T x)^{-1}\vec{x}_0^T} \qquad (3)$$

В этой формуле:

$t_{\frac{\alpha}{2};n-k-1}$ – величина, получаемая по таблице процентных точек распределения Стьюдента с n-k-1 степенями свободы, на k+1 (число оцениваемых параметров регрессии) меньше, чем число элементов в выборке, и двусторонней критической областью ($\frac{\alpha}{2}$). В нашем случае k+1=3. Таким образом, значение t-статистики составит: $t_{\frac{\alpha}{2};n-k-1} = t_{\frac{5\%}{2};5-3} = 4{,}3027$ - [2, табл. 3.2].

$\sigma = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n-k-1}}$ – ошибка функции. Вычисляется как корень квадратный из суммы квадратов остатков, деленный на число элементов в выборке минус число оцениваемых параметров регрессии. На основе данных таблицы 4 сумма квадратов остатков равна 205 484 753

899,57 ($\sum_{i=1}^{n} e_i^2 = 205484753899{,}57$).

Тогда ошибка функции составит:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n-k-1}} = \sqrt{\frac{205484753899{,}57}{5-2-1}} = 320534{,}52 .$$

Уравнение для *z* проще всего решать в программе Matlab [8].

Решив это уравнение, получаем длину доверительного интервала (*z*) равную 1 710 163,197. Чтобы получить сам доверительный интервал для переменной Выходящие материалы, необходимо вычесть его длину из полученного по регрессии значения (нижняя граница доверительного интервала) и прибавить его (верхняя граница доверительного интервала).

Нижняя граница: 1057956,63-1710163,197=-652206,567

Верхняя граница: 1057956,63+1710163,197=2768119,827

Очевидно, что в обоих случаях доверительный интервал превышает минимальное и максимальное значение переменной Выходящие материалы. По этой причине за нижнюю границу доверительного интервала берут минимальное значение, а за верхнее – максимальное (табл. 5).

Нижняя граница: 303 462,8 рублей.

Верхняя граница: 1 999 200 рублей.

Таким образом, прогнозная стоимость эксплуатации проекта, затраты Входящих материалов на который оцениваются в 615 000 рублей, а Работ в 420 000 находится в интервале от 303 462,8 до 1 999 200 рублей. Наиболее вероятно, что его стоимость составит 1 057 956,63 рубля.

Как и в предыдущем случае, длина доверительного интервала сильно завышена из-за небольшого числа наблюдений, и, следовательно, высокого значения t-статистики даже по сравнению с первым регрессионным уравнением (1).

Следует отметить, что для многомерной регрессии длина доверительного интервала не обязательно будет уменьшаться, если показатели проекта будут стремиться к своим средним величинам, поскольку значительное воздействие на него оказывает совместное влияние переменных:

«В случае парной регрессии было отмечено, что длина доверительного интервала увеличивается вместе с удалением от средней. В случае многомерной регрессии нельзя сказать, что длина доверительного интервала увеличивается вместе с Евклидовым расстоянием $\sqrt{(x_{ij} - \bar{x}_j)^2 + (x_{ij+1} - \bar{x}_{j+1})^2}$. Это объясняется наличием ковариации… Если x_1 и x_2 сильно коррелированны, то все равно длина доверительного интервала для изучаемого проекта (x_j) будет довольно большой, несмотря на то, что Евклидово расстояние параметров этого проекта от внутригрупповых средних одно и то же. Таким образом, простое соотношение найденное для парной регрессии не действует в случае многомерной[6]» (перевод Красильникова Д.Е.).

Библиографический список

[6] In the case of simple regression we said that the variance of the prediction error increase as we increase the distance of the point x_j from $\bar{x}$. In the case of multiple regression we cannot say that the variance of the prediction error increase with the Euclidean distance $\sqrt{(x_{ij} - \bar{x}_j)^2 + (x_{ij+1} - \bar{x}_{j+1})^2}$. This is because there is the covariance term as well… If x_1 and x_2 are highly correlated , we will observe wide discrepancies in the variance of the prediction error for the same Euclidean distance of the value of x_j from the sample mean. Thus the simple relationship we found in the case of simple regression does not hold in multiple regression [13, c. 155-156].

1. **Баранова, С.В.** К вопросу о степени влияния кредита на финансовые результаты деятельности организаций АПК / С.В. Баранова, Е.С. Филонова // Инновационный путь развития РФ как важнейшее условие преодоления мирового финансово-экономического кризиса: мат. международной научно-практической конференции, 21-22 апреля 2009 г., – М. Т. 2.
2. **Большев, Л.Н.** Таблицы математической статистики / Л.Н. Большев, Н.В. Смирнов**. –** М.: Наука, 1983.
3. **Дубров, А.М.** Многомерные статистически методы / В.С.Мхитарян, Л.В. Трошин. **–** М.: Финансы и статистика, 2003.
4. **Елисеева, И. И.** Общая теория статистики **/** И. И. Елисеева, М. М. Юзбашев. – М.: Финансы и статистика, 1995.
5. **Иванов, С.И.** Основы экономической теории / С.И. Иванов. – М.: Вита Пресс, 2001.
6. **Красильников, Д.Е.** Обзор литературы по корреляционно-регрессионному анализу с момента возникновения по настоящее время // Математика в высшем образовании. 2010. №8.
7. **Красильников, Д.Е.** Оценка стоимости обслуживания технически сложных систем / Инновационный путь развития РФ как важнейшее условие преодоления мирового финансово-экономического кризиса: мат. международной научно-практической конференции, 21-22 апреля 2009 г. – М. Т. 2.
8. **Красильников, Д.Е.** Программное обеспечение эконометрического исследования // Вестник Нижегородского университета им. Н. И. Лобачевского. 2011. №3(2).
9. **Линник, Ю.В.** Метод наименьших квадратов и основы теории обработки наблюдений. 2-е изд., исп., доп. – М.: Физматгиз, 1962.
10. **Магнус, Я.Р.** Эконометрика: начальный курс / Я.Р. Магнус, П.К. Катышев, А.А. Пересецкий. – 6-е изд. перераб. и доп. – М.: Дело, 2004.
11. **Тунтубалин, В.Н.** Теория вероятностей / В.Н. Тунтубалин. – М.: Изд-во Московского университета, 1972.
12. **Gujarati, D. N.** Basic Econometrics. 3d ed. - McGrawHill, 1995.
13. **Maddala, G. S.** Introduction to Econometrics. 2nd ed. – Willey, 1992.
14. **Thomas, R. L.** Modern Econometrics: an Introduction. – Longman, 1997.

Дата поступления
в редакцию01.02.2014

V.P. Kolpashnikov, D.E. Krasilnikov

ON DISPERSION RANGE CONSTRUCTION FOR ECONOMIC VALUES OBTAINED FROM LINEAR REGRESSION

SMOKEVENT, Limited Liability Company

Purpose: The article proposes the method for range construction for economic values obtained from Seemingly Unrelated Regression (SUR) based on Gauss-Markov theorem.
Design/methodology/approach: A theoretical framework is described for the case of two equations: one of them includes two independent variables and another – one. Therefore two different types of calculations are illustrated.
Findings: The results of research are applicable in economic analysis, may be contained in academic courses on econometrics and industrial organization. Besides, the framework can be adapted by accountants for cost assessment of manufacturing products.
Research limitations/implications: The article considers only the case of Seemingly Unrelated Regression (SUR) and cannot be implemented in case of Systems of Simultaneous Equations which are apparently similar. The criterion for emplacement of the framework is Pearson coefficient value of correlation between errors of two equations less then 0,5.
Originality/value: In practice seemingly unrelated regression models are described only on theoretical economic models like overall equilibrium, IS-LM framework and others. This article suggests another approach to this mathematical model as a “tool” for cost assessment.

Key words: technically compound systems, econometrics analysis, maintenance cost, Seemingly Unrelated Regression (SUR).