

ՀՏԴ 004(519:628)

Информатика

Сандрян С.Н.

к. ф.-м.н. доцент, Прикладная математика и информатика

E-mail: - sandrun@yandex.ru, com: 097745598

ЛИНЕЙНАЯ ОЦЕНКА ПО МЕТОДУ НАИМЕНЬШИХ КВАДРАТОВ С ИСПЫТАНИЕМ ГИПОТЕЗЫ ЛИНЕЙНОЙ СВЯЗИ В ГЕНЕРАЛЬНОЙ СОВОКУПНОСТИ

**Ս.ՍԱՆԴՐՅԱՆ
ՓՈՔՐԱԳՈՒՅՆ ԲԱՌԱԿՈՒՄԻՆԵՐԻ ՄԵԹՈԴՈՎ
ԳԼԽԱՎՈՐ ՀԱՎԱՔՐՈՒՅԹԻ ԳԾԱՅԻՆ ԿԱԳԻ
ԳՆԱՀԱՏՈՒՄ և ՆՐԱ ԱՌԿԱՅՈՒԹՅԱՆ ՎԱՐԿԱԾԻ
ՍՏՈՒԳՈՒՄ**

Աշխատանքում փոքրագույն քառակուսիների մեթոդով, երկու փոփոխական մեծությունների փորձնական տվյալների հիման վրա կառուցված է գծային կապ: Առաջարկված է համաձայնության հայտանիշ, վարկածային գծային կապի առկայության ստուգման համար:

Բանալի բառեր՝ Փոքրագույն քառակուսիների մեթոդ, գծային կախվածություն, վիճակագրական վարկած, ընտրանքային հավասարում, համաձայնության հայտանիշ՝ գծային կապ, փորձնական տվյալներ, ընտրանք, գլխավոր հավաքույթ, փտտահեղի հավանականություն, կոռելյացիայի գործակից, Ստյուդենտի բաշխում, Պիրսոնի կոռելյացիայի գործակից, գրոյալկան վարկած:

В работе на основе экспериментальных данных с помощью метода наименьших квадратов, устанавливается линейная связь между двумя переменными величинами. Предлагаются критерии согласия для проверки о наличии гипотетической линейной связи.

Ключевые слова: метод наименьших квадратов, линейная зависимость. статистическая гипотеза. выборочное уравнение. критерии согласия. линейная связь. экспериментальные данные, выборка, генеральная совокупность, доверительная вероятность. коэффициент корреляции, распределение Стьюдента. коэффициент корреляции Пирсона. нулевая гипотеза.

S.SANDRYAN**LINEAR ESTIMATION BY THE METHOD OF LEAST SQUARES WITH TESTING THE LINEAR CONNECTION HYPOTHESIS IN THE GENERAL SET**

In the work, based on experimental data using the least squares method, a linear relationship is established between two variables. Goodness-of-fit criteria for testing the presence of a hypothetical linear relationship are proposed.

Key words: least square method, linear dependence. statistical hypothesis. sample equation. acceptance criteria. linear connection. experimental data, sample, general population, confidence level. correlation coefficient, Student's distribution. Pearson's correlation coefficient. null hypothesis.

В разных экономических проблемах часто исследуются задачи выявления связи между переменными величинами.

Пусть в результате эксперимента получены n парные значения между изучаемыми величинами x и y , для значений $x_1, x_2, \dots, x_n$ зарегистрированы значения $y_1, y_2, \dots, y_n$ соответственно. Требуется разработать математическую модель для описания связи между величинами x и y . На плоскости Oxy согласно экспериментальным данным из таблицы 1 отметим точки с координатами

$(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$.

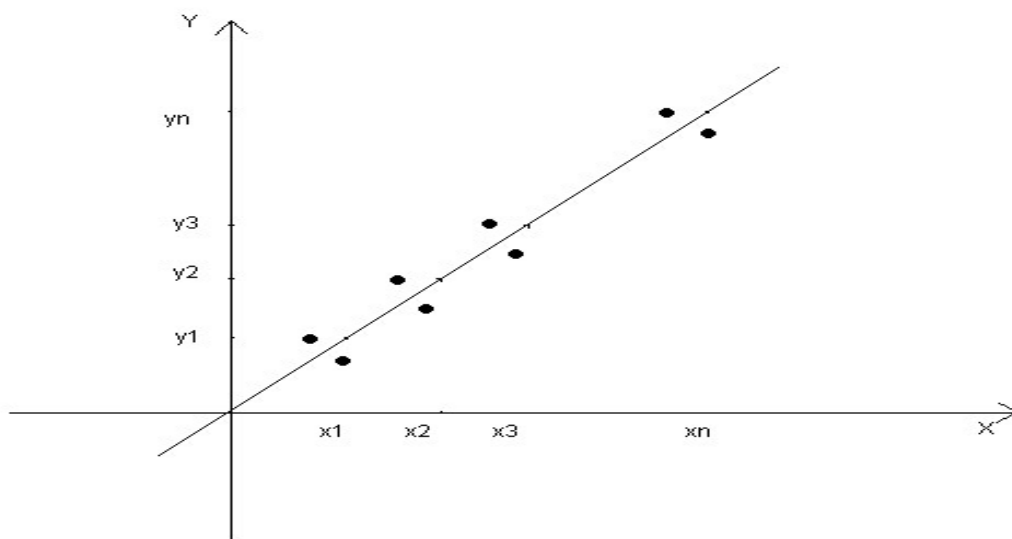


Рис.1

Вид связи $y=f(x)$ устанавливается из накопленных теоретических соображений, или из основания характера расположенных на координатной плоскости Оху точек, соответствующих экспериментальным результатам. В частности, если экспериментальные точки расположены на координатной плоскости так, как показано на рисунке 1, естественно предположить искомую связь в виде линейной функции $f(x)=a+bx$.

При выбранном виде функции $y=f(x,a,b)$ остается подобрать входящие в нее параметры a,b , так, чтобы она в каком-то смысле наилучшим образом описывала реальную взаимосвязь между переменными x и y .

Выбор подходящих коэффициентов можно построить методом наименьших квадратов [4]. Суть этого метода заключается в следующем [1]:

а) построим функцию нескольких переменных $\Phi(a,b)$, задавая сумму квадратных разностей значению y_i даваемых экспериментов и к предлагаемой гипотетической функции $f(x,a,b)$ в точках x_i : $\hat{y} = f(x_i, a, b)$

$$\Phi(a,b) = \sum_{i=1}^n (y_i - f(x_i, a, b))^2.$$

б) подбираются параметры a,b так, чтобы эта сумма имела наименьшее значение

$$\min_{a,b} F(a,b).$$

Таким образом задача выбора неизвестных параметров свелась к нахождению экстремума функции нескольких переменных $\Phi(a,b)$. Для случая линейной связи $y=a+bx$ согласно методу наименьших квадратов, оценки для параметров a и b , как экстремум для функции $\Phi(a,b)$ определяются по формулам [2]:

$$b = \frac{n \cdot \sum_{i=1}^n x_i \cdot y_i - \sum_{i=1}^n x_i \cdot \sum_{i=1}^n y_i}{n \cdot \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2}, \quad a = \frac{\sum_{i=1}^n y_i - b \cdot \sum_{i=1}^n x_i}{n}. \quad (1)$$

Рассмотрим полученное уравнение линейной связи

$$\tilde{y} = a + bx \quad (2)$$

Если вместо x подставить в уравнение (2) экспериментальные значения x_i , то для величин y будут получены значения $\tilde{y}_i$ соответственно. Множество точек $(x_i, \tilde{y}_i)$ находятся на прямой (2), а множество чисто экспериментальных точек (x_i, y_i) могут не находиться точно на этой прямой линии. Это объясняется тем, что на поведение величины y кроме переменной величины x оказывают влияние и другие случайные факторы, которые мы не учитывали в получении линейной связи (2).

Обозначим через $\bar{y} = \frac{1}{n} \cdot \sum_{i=1}^n y_i$ и $\bar{x} = \frac{1}{n} \cdot \sum_{i=1}^n x_i$ средние значения переменных x и y соответственно и $\varepsilon_i = y_i - \hat{y}_i$.

Разница между экспериментальной и теоретической значениями называется ошибкой (отклонением). Влияние не учитываемых факторов скрыто в ошибке ε_i . Если бы связь была строго линейной, то $\varepsilon_i = 0$ для каждой точки. Расположение точек представим на плоскости Oxy .

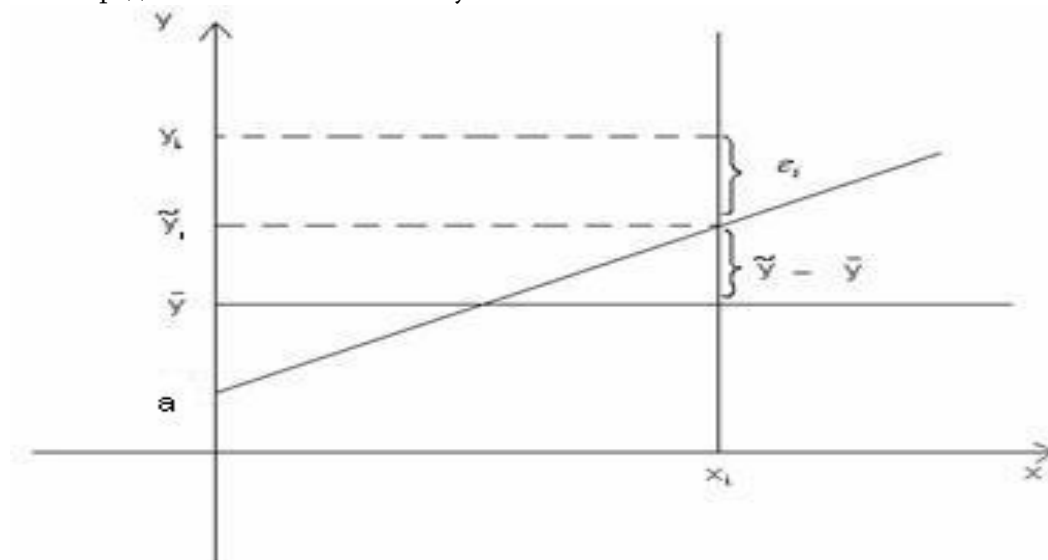


Рис.2

Точность приближения линейной модели экспериментальным данным зависит от разности $y_i - \bar{y}$ - отклонение y_i от среднего значения. Эту разность, согласно рисунку 2, можно представить в виде суммы: $y_i - \bar{y} = (\hat{y}_i - y_i) + \varepsilon_i$, где первое слагаемое (детерминированная часть) не зависит от случайных факторов, а ε_i - непредсказуемая величина. Таким образом, между переменными x и y выявлена гипотетическая линейная связь $y = \alpha + \beta x + \varepsilon_i$, где ε_i - отклонение значения y от линии $y = \alpha + \beta x$. Оценки для α и β методом наименьших квадратов (2), получено соответственно a и b . Очевидно, что для другой выборки оценки a и b будут другими. Дальнейшее изучение поставленной проблемы заключается в том, что как можно на основе оценки a и b согласно экспериментальным результатам убедиться в том, что между переменными x и y в генеральной совокупности действительно существует линейная связь.

Показатель наличия линейной связи в генеральной совокупности - это коэффициент корреляции - K , который не известен. По данным выборки построим выборочный коэффициент корреляции - K^* , оценка для K .

Как оценка K^* выбирается коэффициент корреляции Пирсона [3]:

$$K = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2},$$

для практического вычисления можно преобразовать его к удобной форме:

$$K^* = \frac{n \cdot \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \cdot \sum_{i=1}^n y_i}{\sqrt{(n \cdot \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2) \cdot (n \cdot \sum_{i=1}^n y_i^2 - (\sum_{i=1}^n y_i)^2)}}, \quad |K^*| \leq 1. \quad (3)$$

В случае строгой линейной зависимости между x и y следует, что $K^* = 1$, если линейная зависимость отсутствует, то $K^* = 0$.

Для установления линейной связи в генеральной совокупности для величин x и y на основе выборочного коэффициента корреляции K^* проводим испытание гипотезы.

$H_0: K = 0$, между переменными x и y отсутствует линейная связь в генеральной совокупности.

$H_1: K \neq 0$, между переменными x и y есть линейная связь в генеральной совокупности.

Выводы о наличии линейной связи между переменными x и y , в генеральной совокупности строго зависят от объема выборки. Чем больше объем выборки, тем надежнее полученный результат. Далее, задается доверительная вероятность p и для n -объема парной выборки устанавливается двусторонняя проверка уровня значимости $\alpha = (1 - p)/2$

По таблице t -распределения Стьюдента находим $t_{\alpha, n-2}$.

Устанавливается двусторонняя граница $\pm t_{\alpha, n-2}$.

Согласно статистике $t = \frac{(K^*)^2 (i - 2)}{1 - (K^*)^2}$ принимается или опровергается

нулевая гипотеза H_0 ; т.е. на основе экспериментальных данных согласуются гипотезы о наличии линейной связи между переменными x и y в генеральной совокупности.

В работе выше изложенный теоретический подход для анализа линейной связи методом наименьших квадратов и испытание гипотезы о линейной связи в генеральной совокупности проводится над экспериментальными данными, полученными с помощью социологического опроса для $n=10$ парных значений

(x_i, y_i) $i=1$, где x_i - вес студента, а y_i - рост студента. Результаты представлены в виде таблицы.

Табл. 1.

x	60кг	54 кг	56 кг	65 кг	50 кг	54 кг	63 кг	60 кг	55 кг	59 кг
y	1.70м	1.68м	1.71м	1.73м	1.62м	1.65м	1.69м	1.71м	1.63м	1.68м

Задаем доверительной вероятности $p=0.95$ и $n=10$. Полагая, что между переменными x и y имеет место гипотетическая линейная зависимость, определим выборочное уравнение линейной связи. Заполним таблицу:

Табл. 2.

Номер	x_i	y_i	x^2	y^2	$x \cdot y$
1	60кг	1.70м	3600	28900	10200
2	54 кг	1.68 м	2916	28224	9072
3	56 кг	1.71 м	3136	29241	9576
4	65 кг	1.73 м	4225	29929	11245
5	50 кг	1.62 м	2500	26244	8100
6	54 кг	1.65 м	2916	27225	8910
7	63 кг	1.69 м	3969	28561	10647
8	60 кг	1.71 м	3600	29241	10260
9	55 кг	1.63 м	3025	26569	8965
10	59 кг	1.68 м	3481	28224	9912
Итого	576кг	16.80м	33368	28.2358	968.87

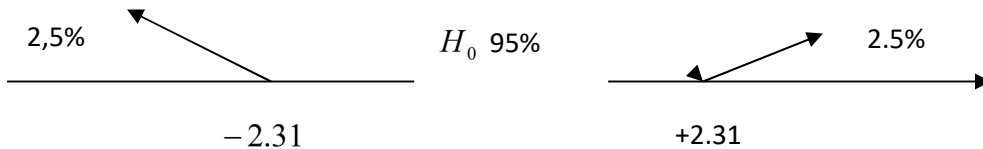
Находим оценки для линейной связи по формулам (1) $b=0.0063$, $a=1.32$, тогда получаем линейное уравнение $y = 1.32 + 0.0063x$.

Определим выборочный коэффициент корреляции Пирсона по формуле (3), проводя вычисления согласно таблице 2, получаем $K^*=0.79$. Проведем двустороннюю проверку: $\alpha = \frac{1-p}{2} = 0.025$.

По таблице t -распределения находим $t_{\alpha, i-2} = 2,31$. Установим граничные точки $t_{\alpha, i-2} = \pm 2.31$. По результатам выборки определим значение статистики:

$$t = \sqrt{\frac{(K^*)^2 \cdot (i - 2)}{(1 - (K^*)^2)}} = 3,64.$$

Отметим граничные точки на числовой оси



Согласно результатам выборки (табл.2) значение статистики находится в области ($3,64 > 2,31$) опровержения нулевой гипотезы, т. е. в генеральной совокупности существует линейная связь между переменными x и y -роста и веса студентов, которая задается линейным уравнением $y = 1.32 + 0.0063x$.

Согласно таблице 2 проводим предсказание переменной y для экспериментальных значений x_i . Результат предсказания представим в таблице 3.

Табл. 3.

x	60	54	56	65	50	54	63	60	55	59
$\hat{y}$	1.698	1.66	1.673	1.7295	1.635	1.635	1.717	1.698	1.66	1.691

Сравнивая данные таблицы 1 и таблицы 3, сделаем вывод о том, что ошибки в предсказаниях незначительные. По линейной формуле (4) также можно прогнозировать значение переменных.

Предложенная авторами методика анализа двумерной линейной связи представляет самостоятельный интерес и его можно применять для решения различных экономических задач.

Գրականություն

1. Пискунов И.С. Дифференциальные и интегральные исчисления. Учебник для вузов. В 2-х томах. Т.1:-М.: Интеграл-Пресс, 2001
2. Крамер Г. Математические методы статистики/Пер.с англ. 2-е изд.-М.: Мир, 1975.
3. Гнеденко Б.В. Курс теории вероятностей.- 5-е изд.-М.: Наука, 1971.
4. Большев Л.Н., Смиронов Н.В. Таблицы математической статистики.-М.: Наука, 1965.

Հոդվածը տպագրության է երաշխավորել խմբագրական խորհրդի անդամ, ֆ.մ.գ.թ. Գ. Հ.Սահակյանը: