СИБИРСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ
SIBERIAN FEDERAL UNIVERSITY

В. А. Середа, А. В. Литаврин, Н. Л. Собачкина

ЭКОНОМЕТРИКА

Учебное пособие

**ИНСТИТУТ МАТЕМАТИКИ
И ФУНДАМЕНТАЛЬНОЙ ИНФОРМАТИКИ**

УДК 338:519.862.6(07)
ББК 65в631я73
С325

Р е ц е н з е н т ы:

С. И. Сенашов, доктор физико-математических наук, профессор Сибирского государственного университета науки и технологий имени академика М. Ф. Решетнёва;

К. А. Филиппов, доктор физико-математических наук, профессор Красноярского государственного аграрного университета

Середа, В. А.

С325 Эконометрика : учеб. пособие / В. А. Середа, А. В. Литаврин, Н. Л. Собачкина. – Красноярск : Сиб. федер. ун-т, 2018. – 148 с.

ISBN 978-5-7638-3996-8

Рассмотрены линейная модель множественной регрессии, нелинейные модели, временные ряды, системы эконометрических уравнений. Приведены примеры решения задач и варианты индивидуальных заданий, а также тесты и вопросы к экзамену.

Предназначено для студентов укрупненной группы направлений подготовки бакалавров 38.03.01 «Экономика», 38.03.02 «Менеджмент», а также для преподавателей, применяющих методы эконометрики в социально-экономических исследованиях.

Электронный вариант издания см.:
http://catalog.sfu-kras.ru

УДК 338:519.862.6(07)
ББК 65в631я73

ISBN 978-5-7638-3996-8

ОГЛАВЛЕНИЕ

ПРЕДИСЛОВИЕ

Переход к рыночной экономике повышает требования к качеству подготовки экономистов, которые, чтобы быть конкурентоспособными и востребованными на рынке труда, должны владеть количественными методами анализа в экономике. При этом высокий динамизм происходящих в стране социально-экономических процессов приводит к тому, что знания об экономике отстают от потребностей управления. Сегодня нужны специалисты, не только владеющие опытом и знаниями предыдущих поколений, но и готовые к встрече с новыми постановками задач, обусловленными спецификой России.

В связи с этим дисциплина «Эконометрика» входит в учебные планы подготовки экономистов всех специальностей и направлений в качестве базовой, обязательной дисциплины и преподается как во всех ведущих университетах мира, так и в отечественных вузах.

В данном курсе рассматриваются задачи эконометрики – науки о связях экономических явлений, условия и методы построения экономических регрессионных моделей по статистическим данным и временным рядам. Изучение этих прикладных разделов математики занимает важное место в формировании экономистов высокой квалификации и служит основой для описания, анализа и прогнозирования реальных экономических процессов. Курс эконометрики призван научить различным способам выражения связей и закономерностей через эконометрические модели, основанные на данных статистических наблюдений.

Значительное внимание в настоящем издании уделяется логическому анализу исходной информации и экономической интерпретации получаемых результатов. В нем даются необходимый теоретический материал и большое количество примеров и задач для самостоятельного решения, а также тесты и вопросы для закрепления материала.

Учебное пособие предназначено для студентов экономических специальностей, направлений подготовки бакалавров, а также для преподавателей, применяющих методы эконометрики в социально-экономических исследованиях.

Авторы выражают искреннюю благодарность доктору физико-математических наук, профессору С. И. Сенашову, доктору физико-математических наук, профессору К. А. Филиппову за рецензирование и редактору Л. Г. Семухиной за целый ряд полезных замечаний и поправок, способствовавших улучшению пособия.

ВВЕДЕНИЕ В КОРРЕЛЯЦИОННО-РЕГРЕССИОННЫЙ АНАЛИЗ

Одна из наиболее общих задач в экономических исследованиях состоит в оценивании степени зависимости изучаемой величины Y от одной или нескольких случайных (или неслучайных) величин X, называемых *факторами*. Зависимость может быть *функциональной*, *статистической* либо отсутствовать вовсе.

Строгая функциональная зависимость между экономическими показателями (наличие всегда выполняющегося равенства $Y = f(X)$) реализуется редко, так как они подвержены влиянию случайных факторов. При статистической зависимости изменение одной из величин влечет изменение распределения другой (в частности, среднего значения; в этом случае статистическую зависимость называют *корреляционной*).

Причем всегда есть несколько величин, которые определяют главные тенденции изменения рассматриваемой величины, и в экономической теории и практике ограничиваются тем или иным кругом таких величин (*объясняющих переменных*). Однако всегда существует и воздействие большого числа других менее важных или трудно идентифицируемых факторов, приводящее к отклонению значений *объясняемой* (зависимой) *переменной* от конкретной формулы ее связи с объясняющими переменными, сколь бы точной эта формула ни была. Нахождение, оценка и анализ таких связей, идентификация объясняющих переменных, построение формул зависимости и оценка их параметров и составляют предмет ***корреляционно-регрессионного анализа***, при этом *корреляционный анализ* занимается исследованием взаимозависимости случайных величин, тогда как *регрессионный анализ* на базе выборочных данных исследует зависимость случайной величины от ряда неслучайных и случайных величин.

Примерами корреляционно, но не функционально связанных величин являются объемы производства и себестоимость продукции, объемы продаж и прибыль, урожай зерна и количество внесенных удобрений. Действительно, в последнем примере с одинаковых по площади участков земли при равных количествах внесенных удобрений снимают различный урожай, т. е. отсутствует функциональная связь. Это объясняется влиянием случайных факторов (осадки, температура, качество семян и др.). Вместе с тем, как показывает опыт, *средний* урожай меняется с изменением количества удобрений, т. е. прослеживается корреляционная зависимость.

Рассмотрим сначала *однофакторную регрессионную модель*.

В этом случае имеется n пар наблюдений (x_i, y_i), $i = 1, 2, \ldots, n$, над некоторыми случайными величинами $X = \{x_i\}$ и $Y = \{y_i\}$. Эти наблюдения можно представить точками на плоскости с координатами (x_i, y_i) в виде так называемой *диаграммы рассеяния*. Задача построения регрессионной модели заключается в том, что необходимо подобрать некоторую кривую (график соответствующей функции) таким образом, чтобы она располагалась как можно «ближе» к этим точкам. Такого рода кривую называют *эмпирической* или *аппроксимирующей* кривой. Весьма часто тип эмпирической кривой определяется экспериментальными или теоретическими соображениями (исходя из законов экономической теории), в противном случае выбор кривой осуществить довольно трудно. Иногда точки на диаграмме рассеяния располагаются таким образом, что не наблюдается никакого их группирования и, соответственно, нет никаких оснований предполагать наличие в наблюдениях какой-либо взаимозависимости.

Результатом исследования статистической взаимозависимости на основе выборочных данных является построение *уравнений регрессии* вида $y = f(x)$.

В самом простом случае предполагается, что f задает уравнение прямой $f(x) = \alpha_0 + \alpha_1 x$. Модель в этом случае имеет вид

$$y_i = \alpha_0 + \alpha_1 x_i + \varepsilon_i \quad (i = 1, 2, \ldots, n). \tag{1}$$

Здесь ε_i являются вертикальными уклонениями точек (x_i, y_i) от аппроксимирующей прямой. Вопрос о нахождении формулы зависимости можно ставить после положительного ответа на вопрос о существовании такой зависимости.

О наличии или отсутствии линейной связи можно судить по величине коэффициента корреляции.

Угловой коэффициент α_1 прямой линии регрессии Y на X называют *коэффициентом регрессии* Y на X и обозначают ρ_{yx}.

Выражение $\sigma_x^2 = \overline{x^2} - (\overline{x})^2$ есть *выборочная дисперсия* X (или *квадрат выборочного среднего квадратического отклонения*).

Выборочный коэффициент корреляции определяется равенством

$$r_{yx} = (\overline{xy} - \overline{x}\,\overline{y})/(\sigma_x \sigma_y), \tag{2}$$

где σ_y есть *выборочное среднее квадратическое отклонение* Y.

Замечание 1. Верхняя черта, как это принято в теории вероятностей и математической статистике, означает среднее значение выборочной совокупности, в данном случае

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}, \ \overline{y} = \frac{\sum_{i=1}^{n} y_i}{n}, \ \overline{x^2} = \frac{\sum_{i=1}^{n} x_i^2}{n}, \ \overline{xy} = \frac{\sum_{i=1}^{n} x_i y_i}{n}.$$

Коэффициент корреляции измеряет силу (тесноту) *линейной* связи между Y и X. Он является безразмерной величиной, не зависит от выбора единиц измерения обеих переменных. Для него всегда выполняется $0 \le |r_{yx}| \le 1$, и чем ближе его значение к ± 1, тем сильнее линейная связь. Коэффициент корреляции будет положительным, если зависимость переменных X и Y прямо пропорциональная, и отрицательным – если обратно пропорциональная.

1. ЛИНЕЙНАЯ МОДЕЛЬ МНОЖЕСТВЕННОЙ РЕГРЕССИИ

Под ***линейной моделью множественной регрессии*** понимается зависимость вида

$$y = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \ldots + \alpha_m x_m + \varepsilon. \tag{3}$$

В случае парной линейной регрессии имеется только один объясняющий фактор x и линейная регрессионная модель записывается следующим образом:

$$y = \alpha_0 + \alpha_1 x + \varepsilon, \tag{4}$$

где ε – случайная составляющая с независимыми значениями,

$$M\varepsilon = 0,\ D\varepsilon = \sigma^2. \tag{5}$$

Оценка параметров регрессии α_0 и α_1 производится по наблюденным значениям зависимой и объясняющей переменных (x_i, y_i), $i = 1, 2, \ldots, n$, где n – число пар наблюдений (объем выборки). После выполнения элементарных преобразований, которые называются методом наименьших квадратов (МНК), получают так называемую *систему нормальных уравнений*, из которой и находят искомые параметры. Для парной линейной регрессии записывают

$$\alpha_1 = \frac{\overline{xy} - \bar{x}\,\bar{y}}{\overline{x^2} - (\bar{x})^2}, \qquad \alpha_0 = \bar{y} - \alpha_1 \bar{x} = \frac{(\overline{x^2})\bar{y} - \bar{x}(\overline{xy})}{\overline{x^2} - (\bar{x})^2}, \tag{6}$$

где $\overline{xy} = \dfrac{\sum x_i y_i}{n},\ \bar{x} = \dfrac{\sum x_i}{n},\ \bar{y} = \dfrac{\sum y_i}{n},\ \overline{x^2} = \dfrac{\sum x_i^2}{n}$.

Из выражений (2) и (6) следует, что

$$\rho_{yx} = r_{yx}\,\sigma_y / \sigma_x. \tag{7}$$

Если выборка имеет достаточно большой объем и хорошо представляет генеральную совокупность (*репрезентативна*), то заключение о тесноте линейной зависимости между признаками, полученными по данным *выборки*, в известной степени может быть распространено и на *генеральную совокупность*, т. е. можно выдвинуть гипотезу об имеющейся линейной связи во всей генеральной совокупности вида $y = \alpha_0 + \alpha_1 x$.

Оценки, сделанные с помощью МНК, обладают следующими свойствами:

- оценки являются несмещенными, т. е. математическое ожидание оценки каждого параметра равно его истинному значению. Это вытекает из того, что $M\varepsilon = 0$, и свидетельствует об отсутствии систематической ошибки в определении положения линии регрессии;
- оценки состоятельны, так как дисперсия оценок параметров при возрастании числа наблюдений стремится к нулю. Иначе говоря, надежность оценки при увеличении выборки растет;
- оценки эффективны, они имеют наименьшую дисперсию по сравнению с любыми другими оценками данного параметра (при линейной аппроксимации).

1.1. Показатели качества регрессии

Надежность получаемых оценок α_0 и α_1 зависит от дисперсии отклонений переменной y от оцененной линии регрессии $\varepsilon_i = y_i - \alpha_0 - \alpha_1 x_i$. Несмещенная оценка дисперсии случайной составляющей вычисляется по формуле

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n} \varepsilon_i^2 , \qquad (8)$$

и является мерой разброса зависимой переменной вокруг линии регрессии (необъясненная дисперсия).

В качестве меры того, насколько хорошо регрессия описывает данную систему наблюдений, служит коэффициент детерминации, при этом вычисляются следующие суммы квадратов отклонений:

$S^2 = \sum_i (y_i - \overline{y})^2$ – фактических значений от их среднего арифметического;

$\hat{S}^2 = \sum_i (\hat{y}_i - \overline{y})^2$ – выровненных значений от среднего арифметического фактических значений;

$\check{S}^2 = \sum_i (y_i - \hat{y}_i)^2$ – фактических от выровненных значений.

Имеет место равенство $S^2 = \hat{S}^2 + \check{S}^2$.

Коэффициент детерминации есть отношение объясненной части вариации ко всей вариации в целом:

$$R^2 = \hat{S}^2/S^2 = 1 - \check{S}^2/S^2. \qquad (9)$$

Таким образом, чем «ближе» этот коэффициент к единице, тем лучше описание, разумеется, если при этом модель методически правильна.

П р и м е р 1. Исследовать зависимость розничного товарооборота, млрд руб., магазинов от среднесписочного числа работников. В табл. 1 в столбцах 2 и 3 приведены значения, соответственно, объемов розничного товарооборота *y* и среднесписочного числа работников *x*, а в следующих столбцах – значения необходимых расчетных величин.

Таблица 1

№ п/п	y	x	$(x-\bar{x})^2$	$(y-\bar{y})^2$	x^2	xy	$\hat{y}$	ε	ε^2
1	2	3	4	5	6	7	8	9	10
1	0,5	73	1 600	0,49	5 329	36,5	0,43	0,07	0,0049
2	0,7	85	784	0,25	7 225	59,5	0,661	0,039	0,0015
3	0,9	102	121	0,09	10 404	91,8	0,998	–0,098	0,0096
4	1,1	115	4	0,01	13 225	126,5	1,239	–0,139	0,0193
5	1,4	122	81	0,04	14 884	170,8	1,373	0,027	0,0007
6	1,4	126	169	0,04	15 876	176,4	1,45	–0,05	0,0025
7	1,7	134	441	0,25	17 956	227,8	1,604	0,096	0,0092
8	1,9	147	1 156	0,49	21 609	279,3	1,854	0,046	0,0021
Сумма	9,6	904	4 356	1,66	106 508	1 168,6	9,592	0,001	0,0479
Среднее	1,2	113	544,5	0,2075	13 313,5	146,075	1,199	0,0001	0,0060

В соответствии с выражением (6)

$$\alpha_1 = (\overline{xy} - \bar{x}\cdot\bar{y})/(\overline{x^2} - (\bar{x})^2) = (146{,}075 - 113\cdot 1{,}2)/544{,}5 = 0{,}019\ 24;$$

$$\alpha_0 = ((\overline{x^2})\cdot\bar{y} - \bar{x}\cdot\overline{xy})/(\overline{x^2} - (\bar{x})^2) = (13\ 313{,}5\cdot 1{,}2 - 113\cdot 146{,}075)/544{,}5 = \\ = -0{,}974.$$

Таким образом, получено уравнение регрессии

$$\hat{y} = -0{,}974 + 0{,}019\ 24x.$$

Коэффициент регрессии α_1 = 0,019 24 показывает, что увеличение среднесписочной численности на одного человека приводит к повышению объема товарооборота в среднем на 19,24 млн руб. Это своего рода эмпирический норматив приростной эффективности использования работников данной группы магазинов. Если увеличение численности на одного работника приводит к меньшему росту объема товарооборота, то прием его на работу необоснован.

Выборочный коэффициент корреляции можно определить, используя равенство (7):

$$r_{yx} = 0{,}019\ 24\cdot\sqrt{544{,}5/0{,}2075} = 0{,}9856,$$

что свидетельствует о наличии тесной линейной связи между численностью работников и розничным товарооборотом.

В столбцах 8 и 10 табл. 1 вычислены выровненные значения эмпирической функции регрессии и квадраты их отклонений от наблюденных значений.

В соответствии с выражением (8) получаем оценку дисперсии случайной составляющей:

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}\varepsilon_i^2 = 0{,}0479/6 = 0{,}008.$$

Согласно формуле (9) значение коэффициента детерминации

$$R^2 = 1 - \check{S}^2/S^2 = 1 - 0{,}0479/1{,}66 = 0{,}971$$

показывает, что 97,1 % общей вариабельности розничного товарооборота объясняется изменениями числа работников, в то время как на все остальные факторы приходится лишь 2,9 % вариабельности.

Найденные отклонения фактических значений от выровненных (столбец 9) позволяют провести сравнительный анализ работы различных магазинов. Прежде всего необходимо обратить внимание на магазины с отрицательным отклонением (3, 4, 6). Особенно велико отклонение у 4-го магазина. Нужно внимательно обследовать эти магазины и установить причины отклонений. Это может быть расположение магазина в стороне от основных потоков покупателей, плохое обслуживание, неудовлетворительный кадровый состав и т. п. Здесь, по-видимому, имеются резервы в организации труда работников. Напротив, в магазинах 1, 2, 5, 7 и 8 работники используются эффективнее статистического «норматива», но может оказаться, что эти магазины объективно находятся в лучших условиях.

Обозначим $S_x = \sum_i (x_i - \bar{x})^2$, тогда дисперсия параметра α_1 вычисляется по формуле $D(\alpha_1) = \sigma^2/ S_x$.

Значимость оцененного коэффициента регрессии α_1 может быть проверена с помощью анализа его отношения к своему стандартному отклонению:

$$t = \alpha_1/\sqrt{D(\alpha_1)}. \qquad (10)$$

Эта величина имеет распределение Стьюдента с $(n - 2)$ степенями свободы и называется *t-статистикой* (см. прил. 1, табл. П1.2). Можно использовать следующее грубое правило для оценки значимости коэффициента линейной регрессии:

- если $t < 1$, то он не может быть признан значимым, поскольку доверительная вероятность здесь составляет менее 0,7;
- если $1 < t < 2$, то сделанная оценка может рассматриваться как более или менее значимая (доверительная вероятность здесь примерно от 0,70 до 0,95);

● если $2 < t < 3$, то это свидетельствует о весьма значимой связи (доверительная вероятность от 0,95 до 0,99);

● если $t > 3$, то это есть практически стопроцентное свидетельство ее наличия.

Сформулированными правилами можно надежно пользоваться при $n \geq 10$.

При большом размере выборки повторяющиеся пары наблюдений группируются в виде корреляционной таблицы. Если n_{yx} – количество наблюдений одинаковых пар (x, y), то для вычисления коэффициента корреляции в формуле (2) необходимо брать $\overline{xy} = \dfrac{\sum x_i y_i}{n}$.

Для оценки тесноты любой корреляционной связи вводится ***корреляционное отношение*** Y к X как корень квадратный из

$$R^2 = 1 - \frac{\sum(\hat{y}_i - y_i)^2}{\sum(y_i - \overline{y})^2}, \qquad (11)$$

где $\hat{y}$ – вычисленное значение признака; y_i – фактическое значение признака Y; $\overline{y}$ – общее среднее признака Y.

Чем ближе корреляционное отношение к единице, тем теснее связь между признаками, однако, оно не задает вида этой связи и не позволяет судить о степени близости наблюдений к какой-либо кривой.

Пример 2. Пусть имеется распределение 50 га пахотной земли по количеству внесенных удобрений x (центнеров на 1 га) и по урожайности y (центнеров с 1 га), приведенное в табл. 2. В этой таблице, например, число 4, стоящее в 1-й строке 2-го столбца, показывает, что на 4 га из 50 было внесено по 10 ц удобрений и при этом получена урожайность по 15 ц с 1 га. Найти уравнение прямой линии регрессии y на x, коэффициент корреляции и корреляционное отношение по данным корреляционной табл. 2.

Таблица 2

y	x			
	10	20	30	n_y
15	4	28	6	38
25	6	–	6	12
n_x	10	28	12	$n = 50$
$\overline{y}_x$	21	15	20	

Вычислим сначала все средние и дисперсии:

$$\overline{y} = (38 \cdot 15 + 12 \cdot 25)/50 = 17{,}4,$$

$$\overline{x} = (10 \cdot 10 + 28 \cdot 20 + 12 \cdot 30)/50 = 20{,}4,$$

$$\overline{x^2} = (10 \cdot 100 + 28 \cdot 400 + 12 \cdot 900)/50 = 460,$$

$$\overline{xy} = (4 \cdot 10 \cdot 15 + 28 \cdot 20 \cdot 15 + 6 \cdot 30 \cdot 15 + 6 \cdot 10 \cdot 25 + 6 \cdot 30 \cdot 25)/50 = 354,$$

$$\sigma_x = \sqrt{460 - 20{,}4^2} = 6{,}62,$$

$$\sigma_y = \sqrt{(38 \cdot (15 - 17{,}4)^2 + 12 \cdot (25 - 17{,}4)^2 + 12 \cdot (25 - 17{,}4)^2)/50} = 4{,}27,$$

$$\sigma_{\overline{y}y} = \sqrt{(10 \cdot (21 - 17{,}4)^2 + 28 \cdot (15 - 17{,}4)^2 + 12 \cdot (20 - 17{,}4)^2)/50} = 2{,}73.$$

Тогда коэффициент корреляции из выражения (2)

$$r_{yx} = (354 - 20{,}4 \cdot 17{,}4)/(6{,}62 \cdot 4{,}27) = -0{,}034,$$

коэффициент регрессии из равенства (7)

$$\rho_{yx} = -0{,}034 \cdot 4{,}27/6{,}62 = -0{,}022,$$

уравнение прямой регрессии имеет вид

$$\overline{y}_x - 17{,}4 = -0{,}022 \cdot (x - 20{,}4) \quad \text{или} \quad \overline{y}_x = -0{,}022x + 17{,}85,$$

и корреляционное отношение из выражения (11)

$$\eta_{yx} = 2{,}73/4{,}27 = 0{,}64.$$

Диаграмма рассеяния и прямая линия регрессии представлены на рис. 1: в кружках указаны частоты распределения выборки n_{yx}.

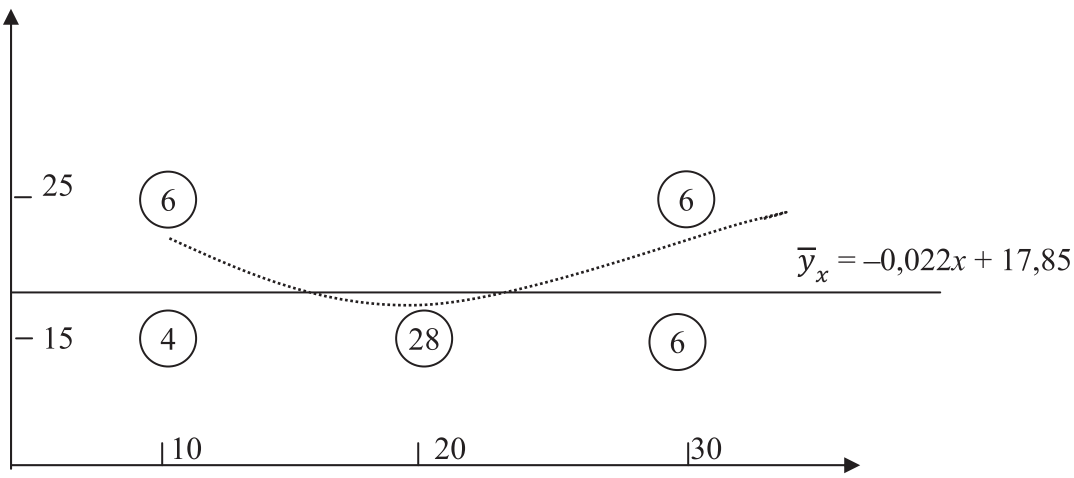

Рис. 1

Из вычисленных показателей можно сделать следующий вывод: линейной связи между признаками нет, но какая-то связь есть, причем весьма существенная.