Эконометрические методы.

Любая экономическая политика заключается в регулировании экономических переменных, и она должна основываться на знании того, как эти переменные влияют на другие переменные, являющиеся ключевыми для принимающего решения политика. Так, в рыночной экономике нельзя непосредственно регулировать темп инфляции, но на него можно воздействовать средствами бюджетно-налоговой и кредитно-денежной политики. Поэтому, в частности, должна быть изучена зависимость между предложением денег и уровнем цен. Невозможно строить, проверять или улучшать экономические модели без статистического анализа их переменных с использованием реальных статистических данных. Эконометрические методы позволяют изучать взаимосвязи экономических показателей – одну из важнейших проблем экономического анализа.

Предположим, что имеются ряды значений переменных, соответствующие им точки нанесены на график и соединены линией. Если это реальные статистические данные, то мы никогда не получим простую линию – линейную, квадратичную, экспоненциальную и т.д. Всегда будут присутствовать отклонения зависимой переменной, вызванные ошибками измерения, влиянием неучтенных величин или случайных факторов. Но если мы не получили, например, точную прямую линию, это еще не значит, что в основе рассматриваемой зависимости лежит нелинейная функция. Возможно, зависимость переменных линейна, и лишь случайные факторы приводят к некоторым отклонениям от нее. То же самое можно сказать и про любой другой вид функции. Связь переменных, на которую накладываются воздействия случайных факторов, называется статистической связью. Наличие такой связи заключается в том, что изменение одной переменной приводят к изменению математического ожидания другой переменной.

Далее будут использованы следующие обозначения:

x, y – переменные;

$\bar{x}, \bar{y}$ – средние значения переменных;

n – размерность выборки.

Корреляционный анализ

Задачей эконометрического исследования является количественная и качественная оценка взаимосвязи наборов данных, представленных в безразмерном виде. Можно указать два типа взаимосвязей между двумя переменными x и y. В одном случае может быть неизвестно, какая из двух переменных является независимой, и какая – зависимой. В этом случае переменные равноправны, и имеет смысл говорить о статистической взаимосвязи корреляционного типа. **Корреляционный анализ** дает возможность установить, ассоциированы ли наборы данных по величине, то есть, большие значения из одного набора данных связаны с большими значениями другого набора (положительная корреляция), или, наоборот, малые значения одного набора связаны с большими значениями другого (отрицательная корреляция), или данные двух диапазонов никак не связаны (корреляция близка к нулю).

В качестве меры для степени линейной связи двух переменных используется коэффициент их корреляции.

$$r[x,y]=\frac{\frac{1}{n}\sum_{k=1}^{n}(x_k-\bar{x})(y_k-\bar{y})}{\sqrt{\frac{1}{n}\sum_{j=1}^{n}(x_j-\bar{x})^2}\sqrt{\frac{1}{n}\sum_{m=1}^{n}(y_m-\bar{y})^2}}.$$

Коэффициент корреляции является безразмерной величиной (так как размерности числителя и знаменателя есть размерности произведения $X\cdot Y$); его величина не зависит от выбора единиц измерения обеих переменных. Величина коэффициента корреляции меняется от –1 в случае строгой линейной отрицательной связи до +1 в случае строгой линейной положительной связи. Близкая к нулю величина коэффициента корреляции говорит об отсутствии линейной связи переменных, но не об отсутствии связи между ними вообще.

Для оценки значимости коэффициента корреляции можно воспользоваться следующей грубой оценкой:

$\left|r_{xy}\right| < 0,3$ – линейная связь отсутствует;

$0,3 \le \left|r_{xy}\right| < 0,7$ – имеется слабая линейная связь;

$\left|r_{xy}\right| \ge 0,7$ – имеется сильная линейная связь.

Для вычисления коэффициента корреляции между двумя наборами данных на листе используется статистическая функция **КОРРЕЛ()** или метод **Корреляция** из Пакета анализа.

Задача 3. Предположим, что застройщик оценивает стоимость группы небольших офисных зданий в традиционном деловом районе. Застройщик может использовать корреляционный анализ для установления связи между выбранными переменными.

Переменная	Смысл переменной
y	Оценочная цена здания под офис, тыс. $;
x1	Общая площадь в квадратных метрах;

x2	Количество офисов;
x3	Количество входов;
x4	Время эксплуатации здания в годах.

В этом примере предполагается, что существует линейная зависимость между каждой независимой переменной (x1, x2, x3 и x4) и зависимой переменной (y), то есть ценой здания под офис в данном районе.

Застройщик наугад выбирает 11 зданий из имеющихся 1500 и получает следующие данные.

x1	x2	x3	x4	y
2310	2	2	20	142
2333	2	2	12	144
2356	3	1,5	33	151
2379	3	2	43	150
2402	2	3	53	139
2425	4	2	23	169
2448	2	1,5	99	126
2471	2	2	34	142
2494	3	3	23	163
2517	4	4	55	169
2540	2	3	22	149

«Пол-входа» (1/2) означает вход только для доставки корреспонденции.

Необходимо установить степень тесноты связи между объясняющими переменными и объясняемыми.

Выполнение.

Заполним данными диапазон A1:E12.

1. Для нахождения парной регрессии (например, между площадью и ценой) используем функцию КОРРЕЛ(), указав в окне диалога диапазоны A2:A12 и E2:E12. Полученное значение 0,32 свидетельствует о наличии слабой линейной связи между выбранными переменными.
2. Чтобы найти коэффициенты корреляции между всеми парами переменных воспользуемся средством **Корреляция** из Анализа данных.

	x1	*x2*	*x3*	*x4*	*y*
x1	1				
x2	0,22	1			
x3	0,62	0,31	1		
x4	0,22	-0,05	-0,05	1	
y	0,32	0,88	0,51	-0,45	1

После выполнения анализа из отчета можно увидеть, что в наибольшей степени цена дома определяется количеством офисов в нем (коэффициент корреляции 0,88).

Отрицательно на цене сказывается возраст дома, – чем он больше, тем дом дешевле (коэффициент корреляции -0,45). Можно также сделать вывод о существующей линейной зависимости площади дома и количества входов в него – коэффициент корреляции 0,62.

Линейная регрессия.

Регрессионный анализ используется, если две исследуемые переменные не равноправны, то есть изменение одной из переменных служит причиной для изменения другой. Например, рост дохода ведет к увеличению потребления; снижение процентной ставки увеличивает инвестиции; увеличение валютного курса сокращает чистый экспорт. Это тот случай, когда должно быть оценено уравнение регрессии $y = f(x)$. Уравнение регрессии – это формула статистической связи между переменными. Если эта формула линейна, то речь идет о линейной регрессии. Формула статистической связи двух переменных называется парной регрессией, зависимость от нескольких переменных – множественной регрессией.

Выбор формулы связи переменных называется **спецификацией** уравнения регрессии; в данном случае выбрана линейная формула. Однако до тех пор, пока не оценены количественные значения параметров, не проверена надежность сделанных оценок, эта формула остается лишь гипотезой. Оценка значений параметром выбранной формулы статистической связи переменных называется **параметризацией** уравнения регрессии.

Любую прямую можно задать ее наклоном и Y-пересечением. Обозначим наклон через α_1, а Y-пересечение – через α_0. Тогда уравнение парной регрессии примет вид $y = \alpha_1 \cdot x + \alpha_0$

Если известны значения α_1 и α_0, то можно вычислить любую точку на прямой, подставляя значения y или x в уравнение.

Оценки коэффициентов α в случае парной регрессии рассчитываются по формуле:

$$\alpha_1 = \frac{\sum_i y_i x_i - n\bar{x}}{\sum_i x_i^2 - n\bar{x}^2}; \; \alpha_0 = \bar{y} - \alpha_1 \bar{x}.$$

Проверка общего качества уравнения регрессии. Коэффициент детерминации R^2

Для анализа общего качества оцененной линейной регрессии используют обычно коэффициент детерминации $\boldsymbol{R^2}$, называемый также квадратом коэффициента множественной корреляции. Для случая парной регрессии это квадрат коэффициента корреляции переменных x и y. Коэффициент детерминации рассчитывается по формуле

$$R^2 = 1 - \frac{\sum_i e_i^2}{\sum_i (y_i - \bar{y})^2},$$

где $e_i = y_i - \alpha_1 x_i - \alpha_0$ – разница между теоретическим и реальным значением y_i.

Он характеризует долю вариации (разброса) зависимой переменной, объясненной с помощью данного уравнения. В качестве меры разброса зависимой переменной обычно используется ее дисперсия, а остаточная вариация может быть измерена как дисперсия отклонений вокруг линии регрессии. Если числитель и знаменатель вычитаемой из единицы дроби разделить на число наблюдений n, то получим, соответственно, выборочные оценки остаточной дисперсии и дисперсии зависимой переменной y. Отношение остаточной и общей дисперсий представляет собой долю необъясненной дисперсии. Если существует статистически значимая линейная связь величин x и y, то коэффициент $\boldsymbol{R}^2$ близок к единице[1].

Множественная линейная регрессия

Значения экономических переменных определяются обычно влиянием не одного, а нескольких объясняющих факторов. В таком случае зависимость $y = f(x)$ означает, что x – вектор, содержащий m компонентов: $x = (x_1, x_2, ..., x_m)$. Задача оценки статистической взаимосвязи переменных y и $x = (x_1, x_2, ..., x_m)$ формулируется аналогично случаю парной регрессии. Записывается функция $y = f(\alpha, x) + \varepsilon$, где α – вектор параметров, ε – случайная ошибка. По данным наблюдений выборки размерности n требуется оценить значения параметров α, то есть провести параметризацию выбранной формулы (спецификации) зависимости.

Мы будем говорить о линейной зависимости y от x, то есть о множественной линейной регрессии. Теоретическое уравнение регрессии имеет вид:

$$y = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + ... + \alpha_m x_m + \varepsilon.$$

Здесь α – вектор неизвестных параметров размерности $(m + 1)$. Пусть имеется n наблюдений вектора x и зависимой переменной y. Для того, чтобы формально можно было решить задачу, то есть найти некоторый наилучший вектор параметров, должно быть $n \geq m + 1$. Если это условие не выполняется, то можно найти бесконечно много разных

[1] Однако он может быть близким к единице просто в силу того, что обе эти величины имеют выраженный временной тренд, не связанный с их причинно-следственной взаимозависимостью. Поэтому будем считать данные задач перекрестными, то есть относящимися к одному временному периоду.

векторов коэффициентов, при которых линейная формула связывает между собой x и y для имеющихся наблюдений абсолютно точно. Если, в частном случае, $n = m + 1$ (например, при двух объясняющих переменных в уравнении $y = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2$ и трех наблюдениях), то оценки коэффициентов α рассчитываются единственным образом – путем решения системы линейных уравнений.

Функция ЛИНЕЙН()

Чтобы вычислить прямую линию, которая наилучшим образом аппроксимирует имеющиеся данные, используется статистическая функция ЛИНЕЙН(). Функция возвращает массив, который описывает полученную прямую. ЛИНЕЙН может также возвращать дополнительную регрессионную статистику. Отчет по регрессии располагается в заранее выделенном диапазоне ячеек следующим образом:

α_n	...	α_2	α_1	α_0
σ_n	…	σ_2	σ_1	σ_0
R^2	se_y			
F	k			
SSreg	*SSresid*			

se_y – стандартная ошибка для оценки y;

F – F-статистика, или F-наблюдаемое значение. F-статистика используется для определения того, является ли наблюдаемая взаимосвязь между зависимой и независимой переменными случайной или нет;

k – степени свободы. Степени свободы полезны для нахождения F-критических значений в статистической таблице. Для определения уровня надежности модели нужно сравнить значения в таблице с F-статистикой, возвращаемой функцией ЛИНЕЙН;

SSreg – регрессионная сумма квадратов;

SSresid – остаточная сумма квадратов.

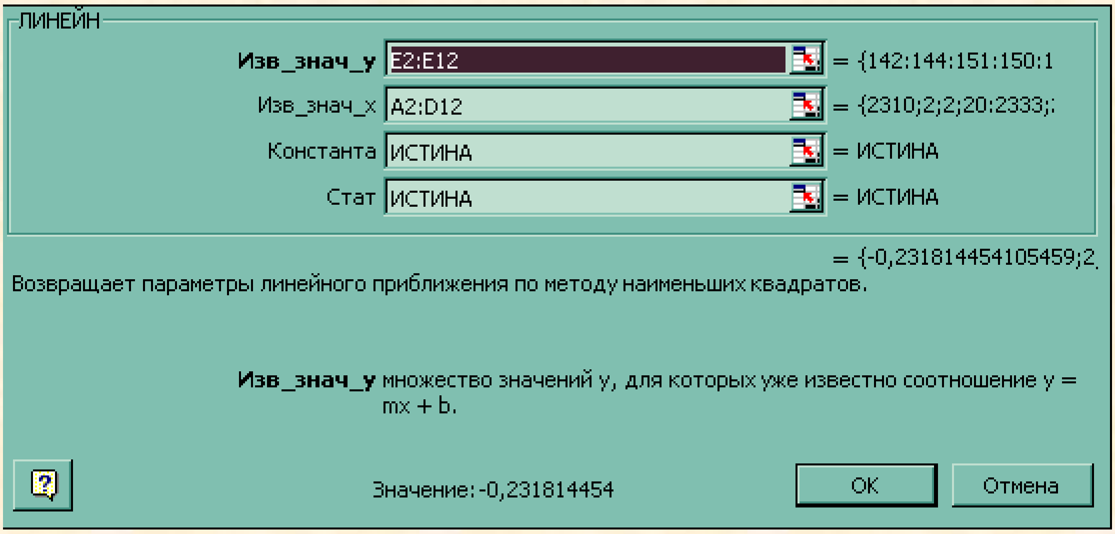
Проводя регрессионный анализ, Microsoft Excel вычисляет для каждой точки квадрат разности между прогнозируемым значением y и фактическим значением y. Сумма этих квадратов разностей называется остаточной суммой квадратов. Затем Microsoft Excel подсчитывает сумму квадратов разностей между фактическими значениями y и средним значением y, которая называется общей суммой квадратов (регрессионная сумма квадратов + остаточная сумма квадратов). Чем меньше остаточная сумма квадратов по сравнению с общей суммой квадратов, тем больше значение коэффициента детерминации R^2, который показывает, насколько хорошо уравнение, полученное с помощью регрессионного анализа, объясняет взаимосвязи между переменными.

Задача 2. Используя множественный регрессионный анализ, оценить цену офисного здания в заданном районе на основе описанных переменных (см. задачу 1).

Выполнение.

Для нахождения уравнения регрессии с помощью функции ЛИНЕЙН() необходимо выполнить следующие шаги.

1. Озаглавьте область будущего отчета, введя в клетку G1: Регрессионный анализ. Выделите диапазон G2:K9. Размер диапазона выбирается следующим образом. Число строк – одна, если требуется получить только коэффициенты регрессии, пять, если необходимо получить дополнительную статистику. Число столбцов соответствует числу переменных.
2. Выберите функцию ЛИНЕЙН() из раздела статистические и заполните окно диалога согласно рисунку.



Известные_значения_y – это множество значений *y*, которые уже известны (E2:E12).

Известные_значения_x – это множество значений *x*, которые уже известны для соотношения (A2:D12). Массив известные_значения_x может содержать одно или несколько множеств переменных при условии, что они имеют одинаковую размерность с массивом Известные_значения_y.

Конст – это логическое значение, которое указывает, требуется ли, чтобы константа α_0 была равна 0. Если конст имеет значение ИСТИНА или опущено, то α_0 вычисляется обычным образом. Если конст имеет значение ЛОЖЬ, то α_0 полагается равным 0 и значения α_1 подбираются так, чтобы выполнялось соотношение $y = \alpha_1 x$.

Статистика – это логическое значение, которое указывает, требуется ли вернуть дополнительную статистику по регрессии. Если статистика имеет значение ИСТИНА, то функция ЛИНЕЙН возвращает дополнительную регрессионную статистику. Если статистика имеет значение ЛОЖЬ или опущена, то функция ЛИНЕЙН возвращает только коэффициенты α_i.

3. Функция ЛИНЕЙН() является формулой массива. Поэтому после заполнения окна диалога и нажатия ОК, необходимо комбинацией Shift-Ctrl-Enter завершить заполнение выделенного диапазона.

ЛИНЕЙН(E2:E12;A2:D12;ИСТИНА;ИСТИНА) возвращает следующие результаты.

-0,23181	2,7092	12,6183	0,02556	56,5870
0,01372	0,54907	0,41393	0,00561	12,6616
0,99654	1,00423	#Н/Д	#Н/Д	#Н/Д
432,499	6	#Н/Д	#Н/Д	#Н/Д
1744,67	6,05090	#Н/Д	#Н/Д	#Н/Д

Теперь может быть получено уравнение множественной регрессии $y = \alpha_1 \cdot x_1 + \alpha_2 \cdot x_2 + \alpha_3 \cdot x_3 + \alpha_4 \cdot x_4 + \alpha_0$:

$$y = 0{,}03x_1 + 12{,}62x_2 + 2{,}71x_3 - 0{,}23x_4 + 56{,}59$$

По этому уравнению застройщик может определить оценочную стоимость здания под офис в том же районе, которое имеет площадь 2500 квадратных метров, три офиса, два входа, зданию 25 лет, используя следующее уравнение:

y = *0,03**2500 + *12,62**3 + *2,71**2 – *0,02**25 + *56,59* = 175,37 тыс.$.