Лекция 23

ҚАЗАҚСТАННЫҢ АШЫҚ УНИВЕРСИТЕТІ

ОСНОВЫ СТАТИСТИКИ

Регрессионно-корреляционный анализ в MS Excel

Темой нашей лекции будет являться изучение возможностей MS Excel для проведения анализа уравнения парной регрессии, в том числе показателей корреляция и ковариация, а также возможность построения линии тренда предварительного анализа уравнения парной регрессии.

Регрессионный анализ является одним из самых востребованных методов статистического исследования. С его помощью можно установить степень влияния независимых величин на зависимую переменную. В функционале Microsoft Excel имеются инструменты, предназначенные для проведения подобного вида анализа. Давайте разберем, что они собой представляют и как ими пользоваться. Как подключается Пакет, вы уже знаете.

Виды регрессионного анализа:

Существует несколько видов регрессий:

- параболическая;
- степенная;
- логарифмическая;
- линейная;
- экспоненциальная;
- показательная;
- гиперболическая;

Прежде чем переходить к инструменту Регрессия рассмотрим две основные функции, используемые при расчетах параметров модели, это функция вычисления корреляция и ковариация.

Количественная характеристика взаимосвязи может быть получена при вычислении коэффициента корреляции. Этот популярный в статистических анализах коэффициент показывает, связаны ли какие-либо параметры друг с другом (например, рост и вес; уровень интеллекта и успеваемость; количество травм и продолжительность работы).

Использование корреляции.

Вычисление корреляции особенно широко используется в экономике, социологических исследованиях, медицине и биометрии – везде, где можно получить два массива данных, между которыми может обнаружиться связь.

Как выполнить корреляцию в MS Excel?

Самым трудоемким этапом определения корреляции является набор массива данных. Сравниваемые данные располагаются обычно в двух колонках или строчках. Таблицу следует делать без пропусков в ячейках. Необходимые манипуляции можно сделать в разделе формул:

Выбрать пустую ячейку, в которую будет выведен результат расчетов.

Нажать в главном меню MS Excel пункт «Формулы». Среди кнопок, сгруппированных в «Библиотеку функций», выбрать «Другие функции». В выпадающих списках выбрать функцию расчета корреляции (Статистические — КОРРЕЛ).

В MS Excel откроется панель «Аргументы функции». «Массив 1» и «Массив 2» — это диапазоны сравниваемых данных. Для автоматического заполнения этих полей можно просто выделить нужные ячейки таблицы.

Кликнуть «ОК», закрыв окно аргументов функции. В ячейке появится подсчитанный коэффициент корреляции.

Напомним, что Корреляция может быть прямая (если коэффициент больше нуля) и обратная (от -1 до 0). Первая означает, что при росте одного параметра растет и другой. Обратная (отрицательная) корреляция отражает факт, что при росте одной переменной другая уменьшается. Корреляция может быть близка к нулю. Это обычно свидетельствует, что исследуемые параметры не связаны друг с другом. Но иногда нулевая корреляция возникает, если сделана неудачная выборка, которая не отразила связь, либо связь имеет сложный нелинейный характер.

Расчет коэффициента корреляции.

Теперь давайте попробуем посчитать коэффициент корреляции на конкретном примере. Имеем таблицу, в которой расписана в отдельных колонках рассматриваемые года с 1991 по 2017 г. и количество экономически активного населения г. Алматы. Нам предстоит выяснить степень зависимости количества экономически активного населения г. Алматы от общей численности населения.

Одним из способов, с помощью которого можно провести корреляционный анализ, является использование функции КОРРЕЛ. Синтаксис функции имеет общий вид КОРРЕЛ(массив1; массив2).

Для этого в диапазоне $B$5:$B$31 набираем данные экономически активного населения г. Алматы по годам. Затем диапазону ячеек $B$5:$B$31 присваиваем имя «Выборка». Выделяем ячейку, в которой должен выводиться результат расчета. Кликаем по кнопке «Вставить функцию», которая размещается слева от строки формул (все процедуры делали ранее).

В списке, который представлен в окне Мастера функций, ищем и выделяем функцию КОРРЕЛ. Кликаем на кнопку «ОК».

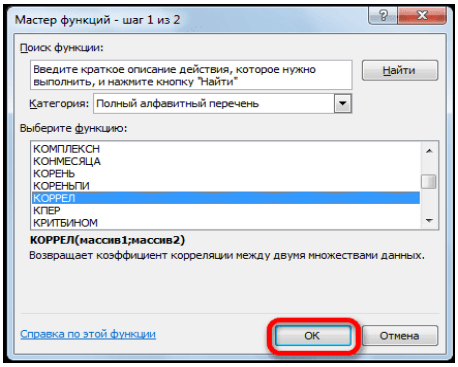


Рисунок 23.1 Оконо запуска функции КОРРЕЛ.

Открывается окно аргументов функции. В поле «Массив1» вводим координаты диапазона ячеек одного из значений, зависимость которого следует определить. В нашем случае это будут значения в колонке «Численность населения». Для того, чтобы внести адрес массива в поле, просто выделяем все ячейки с данными в вышеуказанном столбце.

В поле «Массив2» нужно внести координаты второго столбца. У нас количество экономически активного населения г. Алматы. Точно так же, как и в предыдущем случае, заносим данные в поле. Кликаем на кнопку «ОК».

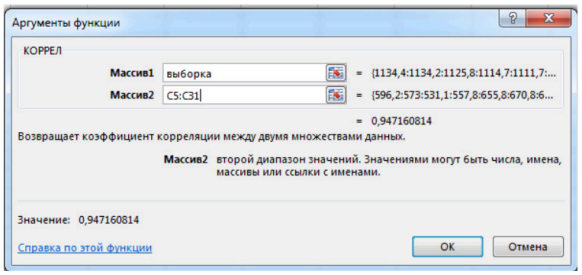


Рисунок 23.2 Диалогвое окно функции КОРРЕЛ

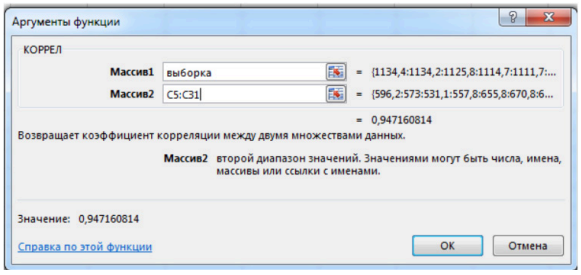
Как видим, коэффициент корреляции в виде числа появляется в заранее выбранной нами ячейке. В данном случае он равен 0,947, что является очень высоким признаком зависимости одной величины от другой.

Использование MS EXCEL для расчета ковариации. Ковариация близка по смыслу с дисперсией (также является мерой разброса) с тем отличием, что она определена для 2-х переменных, а дисперсия - для одной. Поэтому для вычисления ковариации в MS EXCEL используются функции КОВАРИАЦИЯ.Г() и КОВАРИАЦИЯ.В(). В первом случае формула для вычисления аналогична вышеуказанной (окончание .Г – обозначает Генеральная совокупность), во втором – вместо множителя 1/n используется 1/(n-1), т.е.

окончание. В - обозначает Выборка.

Как рассчитать ковариацию в MS Excel?

Все процедуры аналогичны процессу вычисления корреляции, только вызываем функцию КОВАР. Сама функция имеет общий вид КОВАР(массив1; массив2). В окне аргументов функции в поле «Массив1» вводим координаты диапазона ячеек одного из значений, зависимость которого следует определить, в поле «Массив2» нужно внести координаты второго столбца. Точно так же, как и в предыдущем случае, заносим данные в поле.



Кликаем на кнопку «ОК».

Рисунок 23.3 Диалоговое окно функции КОВАР.

Как видим, ковариации в виде числа появляется в заранее выбранной нами ячейке. В данном случае он равен 22306,19, что является очень высоким признаком зависимости одной величины от другой.

Линейная регрессия в программе MS Excel. Общее уравнение регрессии линейного вида выглядит следующим образом:

Формула 23.1

$$Y= a + b1X1 + b2X2 + \ldots + bkXk + \varepsilon$$

В этой формуле Y означает переменную, влияние факторов на которую мы пытаемся изучить. В нашем случае, это количество покупателей. Значение x – это различные факторы, влияющие на переменную. Параметры a и b являются коэффициентами регрессии. То есть, именно они определяют значимость того или иного фактора. Индекс k обозначает общее количество этих самых факторов.

Кликаем по кнопке «Анализ данных». Она размещена во вкладке «Главная» в блоке инструментов «Анализ данных».

Открывается небольшое окошко. В нём выбираем пункт «Регрессия». Кликаем на кнопку «ОК».

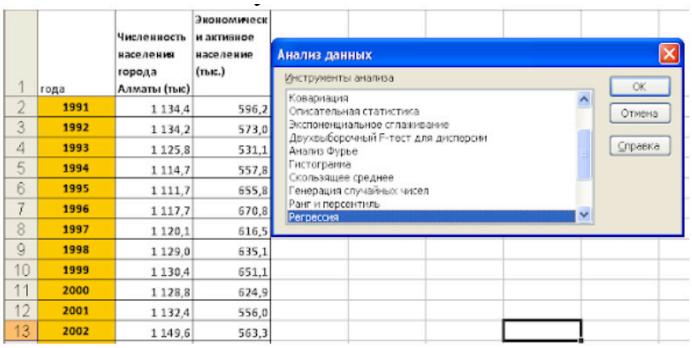


Рисунок 23.4 Запуск инструмента Регрессия

Открывается окно настроек регрессии. В нём обязательными для заполнения полями являются «Входной интервал Y» и «Входной интервал X». Все остальные настройки можно оставить по умолчанию.

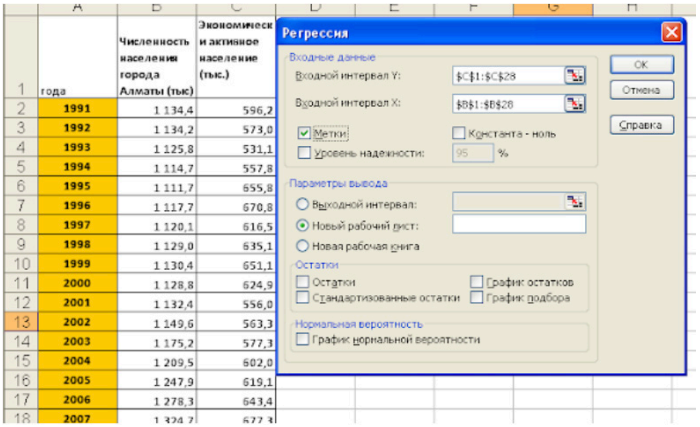


Рисунок 23.5. Диалоговое окно Регрессия.

В поле «Входной интервал Y» указываем адрес диапазона ячеек, где расположены переменные данные, влияние факторов на которые мы пытаемся установить. В нашем случае это будут ячейки столбца «Количество экономически активного населения». Адрес можно вписать вручную с клавиатуры, а можно, просто выделить требуемый столбец. Последний вариант намного проще и удобнее.

В поле «Входной интервал X» вводим адрес диапазона ячеек, где находятся данные того фактора, влияние которого на переменную мы хотим установить. Как говорилось выше, нам нужно установить влияние численности населения на численность экономически активного населения, поэтому вводим соответствующий адрес. Это можно сделать теми же способами, что и в поле «Экономически активное население».

Если активизировать переключатель Метки, то во входные интервалы для X и Y можно добавить ячейки с названиями, и соответствующие метки появятся в итоговой таблице, что значительно облегчит её понимание, можно установить уровень надёжности, константу-ноль, отобразить график нормальной вероятности, и выполнить другие действия. Но, в большинстве случаев, эти настройки изменять не нужно. Единственное на что следует обратить внимание, так это на параметры вывода.

1	ВЫВОД ИТОГОВ								
2									
3	*Регрессионная статистика*								
4	Множественный R	0,947160814							
5	R-квадрат	0,897113607							
6	Нормированный R-квадрат	0,892998152							
7	Стандартная ошибка	36,62894808							
8	Наблюдения	27							
9									
10	Дисперсионный анализ								
11		*df*	*SS*	*MS*	*F*	*Значимость F*			
12	Регрессия	1	292468,0337	292468,0337	217,986457	7,49638E-14			
13	Остаток	25	33541,99594	1341,679838					
14	Итого	26	326010,0296						
15									
16		*Коэффициенты*	*Стандартная ошибка*	*t-статистика*	*P-Значение*	*Нижние 95%*	*Верхние 95%*	*Нижние 95,0%*	*Верхние 95,0%*
17	Y-пересечение	45,04823645	43,52588438	1,034975787	0,31058849	-44,59499973	134,6914726	-44,59499973	134,6914726
18	Численность населения города Алматы (тыс)	0,485611816	0,032890804	14,76436442	7,4964E-14	0,417871937	0,553351694	0,417871937	0,553351694

Рисунок 23.4 Результат работы пакета Регрессия.

По умолчанию вывод результатов анализа осуществляется на другом листе, но переставив переключатель, вы можете установить вывод в указанном диапазоне на том, же листе, где расположена таблица с исходными данными, или в отдельной книге, то есть в новом файле. После того, как все настройки установлены, кликаем на кнопку «ОК».

Разбор результатов анализа.

Данная таблица содержит большое количество информации, поэтому будем изучать её содержимое постепенно, в нескольких последующих работах. Представленные в этой таблице данные можно условно разделить на три раздела:

- регрессионная статистика;

- дисперсионный анализ;
- коэффициенты.

Весь раздел регрессионная статистика посвящен описанию коэффициента детерминации и его различным характеристикам. В пунктах множественный R и R-квадрат выводится значение коэффициента детерминации и его квадрата. Пункты меню нормированный R-квадрат, и стандартная ошибка будут нами рассмотрены позднее, при изучении множественной регрессии. Кроме этого выдается общее количество наблюдений.

Рассмотрим раздел дисперсионный анализ.

В столбце SS выдаются все виды сумм квадратов отклонений. В данном случае в первой строке, которая соответствует надписи Регрессия, выдается объясненная сумма квадратов отклонений RSS.

В строке – Остаток – выдается необъясненная (остаточная) сумма квадратов отклонений ESS.

В строке – Итого – выдается общая сумма квадратов отклонений TSS.

В последнем разделе, который не имеет названия, будет интерпретироваться как раздел – коэффициенты, содержится полная информация по коэффициентам. Рассмотрим значения, полученные в столбце Коэффициенты. Пункт Y-пересечение выдает значение коэффициента a. Пункт «Численность населения» выдает значение коэффициента b.

Представленные в таблице значения полностью совпадают с данными, полученными посредством статистических функций и линий тренда на точечной диаграмме.

В диалоговом окне Регрессия имеется целый раздел переключателей для получения дополнительной информации по остаткам. Например, указав опцию Остатки, наряду со стандартной таблицей регрессии будет выдана дополнительная таблица (Рисунок 23.5) следующего вида:

23	ВЫВОД ОСТАТКА			
24	*Наблюдение*	*Предсказанное Экономически активное население (тыс.)*	*Остатки*	*Стандартные остатки*
25	1	595,92628	0,27371999	0,007620768
26	2	595,8291576	-22,82915765	-0,635597397
27	3	591,7500184	-60,6500184	-1,688585905
28	4	586,3597272	-28,55972724	-0,795144901
29	5	584,9028918	70,8971082	1,973879989
30	6	587,8165627	82,98343731	2,310381206
31	7	588,982031	27,51796895	0,766140815
32	8	593,2991201	41,80087991	1,163798108
33	9	594,0032572	57,09674278	1,589657475
34	10	593,1874294	31,71257063	0,882924708
35	11	594,9647686	-38,96476861	-1,084836589
36	12	603,3270041	-40,02700408	-1,11441079
37	13	615,7440982	-38,4440982	-1,070340357
38	14	632,3908712	-30,39087124	-0,846126649
39	15	651,038365	-31,93836496	-0,889211155
40	16	665,7815397	-22,38153968	-0,623135053
41	17	688,357633	-11,05763298	-0,3078608
42	18	706,3932558	-0,89325581	-0,024869558
43	19	720,3400272	-13,54002715	-0,376974313
44	20	731,3148542	-8,914854184	-0,24820268
45	21	748,8940019	-7,994001907	-0,222564796
46	22	761,228542	13,47145798	0,375065246
47	23	776,8166813	10,6833187	0,297439339
48	24	842,1314705	-32,93147049	-0,916860677
49	25	871,8994748	13,20052521	0,367522078
50	26	895,500209	20,79979098	0,579096838
51	27	920,1207281	19,07927193	0,531195052

Рисунок 23.5 Дополнительная таблица.

В этой таблице получены результаты предсказанных значений и значения остатков отдельно для каждого наблюдения. Указав опции График подбора, График остатков и График нормального распределения можно получить множество дополнительной информации и некоторые диаграммы.

Используя этот инструмент можно получить полную информацию относительно регрессионной модели. Таблица достаточно громоздкая, могут появиться затруднения с интерпретацией полученных результатов. Поэтому рекомендуется начинать исследование модели с использования статистических функций и линии тренда на точечной диаграмме.

Вернемся к нашей задаче. Одним из основных показателей является R-квадрат. В нем указывается качество модели. В нашем случае данный коэффициент равен 0,897 или около 89,7%. Это приемлемый уровень качества. Зависимость менее 0,5 является плохой.

Ещё один важный показатель расположен в ячейке на пересечении строки «Y-пересечение» и столбца «Коэффициенты». Тут указывается, какое значение будет у Y, а в нашем случае, это экономически активное население 45,048, при всех остальных факторах равных нулю. В этой таблице данное значение равно 58,04.

Значение на пересечении граф «Переменная X1» и «Коэффициенты» показывает уровень зависимости Y от X. В нашем случае — это уровень зависимости численности экономически активного населения от общей численности населения. Коэффициент 0, 485 считается довольно высоким показателем влияния для данного типа задачи.

Можно также использовать инструмент так называемого построения линий тренда. Для начала рассмотрим, что это такое и как данный инструмент можно использовать в анализе уравнения парной регрессии и корреляции.

Сущность и основные формы трендов.

Трендом называется выражение тенденции в форме достаточно простого и удобного уравнения, наилучшим образом аппроксимирующего (приближающего) истинную тенденцию динамического ряда.

По форме тренды могут быть линейными, параболическими, экспоненциальными, логарифмическими, степенными, гиперболическими, полиномиальными, логистическими и другими. Excel предоставляет инструменты построения линейного, экспоненциального, логарифмического, степенного и полиномиального (до полинома 6-й степени) трендов, а также скользящую среднюю.

Линейная форма тренда:

$$Y = a + bt,$$

где: Y– уровни показателя, освобожденные от колебаний и выравненные по прямой;

a – начальный уровень тренда в момент или за период, принятый за начало отсчета времени t;

b – среднее изменение за единицу времени, т. е. константа тренда, скорость изменения. Это может быть, например, среднедневной, среднемесячный и среднегодовой прирост какого-либо показателя.

Через скорость изменения линейный тренд хорошо отражает результирующее влияние многих других факторов, одновременно действовавших в единицу времени (цен месяц, год, и т. д.). Тренд можно рассматривать в качестве обобщенного выражения действий комплекса факторов, т. е. их равнодействующей. При этом, в отличие от уравнения множественной регрессии, сами факторы здесь не показываются и влияние каждого из них не выделяется. «От имени» всех факторов в тренде выступает единым – результирующим фактором - время. Например, так в конечном счете в макроэкономике выражаются тенденции изменения важнейших показателей: национального дохода, заработной платы, урожайности и др.

Параболическая форма тренда имеет вид $Y = a + bt + ct2$,

где: Y, a, b, t определены при описании линейного тренда;

c – это константа параболического тренда, его квадратический параметр, равный, половине ускорения.

Параболическая форма тренда достаточно хорошо отражает ускорение или замедление развития при наличии постоянного ускорения, которое обеспечивается влиянием важных факторов (снятием ограничений в распределении дохода, уменьшением налогов, прогрессирующим внедрением нового оборудования и т. п.). При $c<0$, т. е. при отрицательном ускорении, тренд отражает замедление роста со все большей скоростью, что характерно, например, для производства устаревшего товара или оборудования.

Экспоненциальная форма тренда имеет вид $Y = akt$, где константа тренда k выражает темп изменения в количестве раз.

При $k>1$ экспоненциальный тренд показывает тенденцию все более ускоряющегося развития (рост населения в эпоху «демографического взрыва» в XX столетии). Такой рост может продолжаться лишь на небольшом историческом отрезке времени, поскольку он неизбежно приходит в противоречие с имеющимися ресурсами. При $k<l$ экспоненциальный тренд показывает тенденцию все более замедляющегося процесса (трудоемкость продукции, удельные затраты топлива).

Логарифмическая форма тренда $Y = a + b \ln t$ пригодна для отражения тенденция замедляющегося роста при отсутствии предельного возможного значения. При достаточно большом t логарифмическая кривая становится мало отличимой от прямой линии. Такая форма характерна для развития показателей, которые все труднее улучшить (спортивные рекорды, рост производительности процесса при отсутствии

качественного его улучшения).

Степенная форма тренда Y = atb , где b - это константа тренда. При b=1 степенной тренд превращается в линейный, а при b=2 мы имеем параболический тренд. Степенной тренд хорошо подходит для отображения процессов с разной мерой пропорциональности изменений во времени. Линия степенного тренда обязательно должна проходить через начало координат.

Гиперболическая форма тренда Y = a + при b > 0 выражает тенденцию замедляющегося снижения уровня, стремящегося к пределу a, однако при b<0 тренд выражает тенденцию замедляющегося роста уровней, стремящихся в пределе к a. В целом же, гиперболический тренд подходит для отображения тенденций процессов, ограниченных предельным значением уровня (грамотность населения, КПД двигателя и т. п.).

Логистическая форма тренда подходит для отображения развития во всех его фазах в течение длительного периода (вначале медленное насыщение потребителей товарами, затем ускорение, равномерность, замедление). Логистический тренд имеет форму:

$$Y = \frac{Y_{max} - Y_{min}}{e^{a+bt} + 1} + Y_{min}$$

где e - основание натурального логарифма; Ymax, Ymin – максимальное и минимальное значения уровня; a, b - параметры тренда.

Алгоритм построения линии тренда. Приложение MS Excel предоставляет возможность построение линии тренда при помощи графика. При этом, исходные данные для его формирования берутся из заранее подготовленной таблицы.

Подведем итоги практико-ориентированной лекции, сегодня мы рассмотрели два ключевых инструмента оценки уравнения парной регрессии, которые не только позволяют оценивать параметры модели, но и строить прогнозные значения.