

Лекция 4. Корреляционный анализ.

Вопросы лекции:

- 1. Сущность корреляции**
- 2. Определение формы связи при парной зависимости**
- 3. Автокорреляционная функция и мультиколлинеарность**

Вопрос 1. Сущность корреляции

Корреляционный анализ (корреляционная модель) — метод применяемый тогда, когда данные наблюдения или эксперимента можно считать случайными и выбранными из совокупности распределенной по многомерному нормальному закону.

Основная задача корреляционного анализа — выявление связи между случайными переменными путем точечной и интервальной оценки различных (парных, множественных, частных) коэффициентов корреляции. Дополнительная задача корреляционного анализа (являющаяся основной в регрессионном анализе) состоит в оценке уравнений регрессии одной переменной по другой.

Понятие корреляция появилось в середине XIX века в работах английских статистиков Ф. Гальтона и К.Пирсона. Этот термин произошел от латинского «correlatio» - соотношение, взаимосвязь.

Методы теории корреляции позволяют определять количественную зависимость между различными техническими, технологическими, организационными, экономическими и другими факторами, т. е. строить экономико-статистические модели.

Различают функциональную и корреляционную зависимости.

Под функциональной понимается такая зависимость, когда с изменением одного фактора изменяется другой, одному значению независимого фактора обычно соответствует только одно значение зависимого фактора. Корреляционная зависимость - это такая зависимость, при которой изменение одной случайной величины вызывает изменение среднего значения другой. Конкретных же значений зависимого переменного, соответствующих одному значению независимого, может быть несколько. На рисунках приведены примеры функциональной и корреляционной зависимостей.

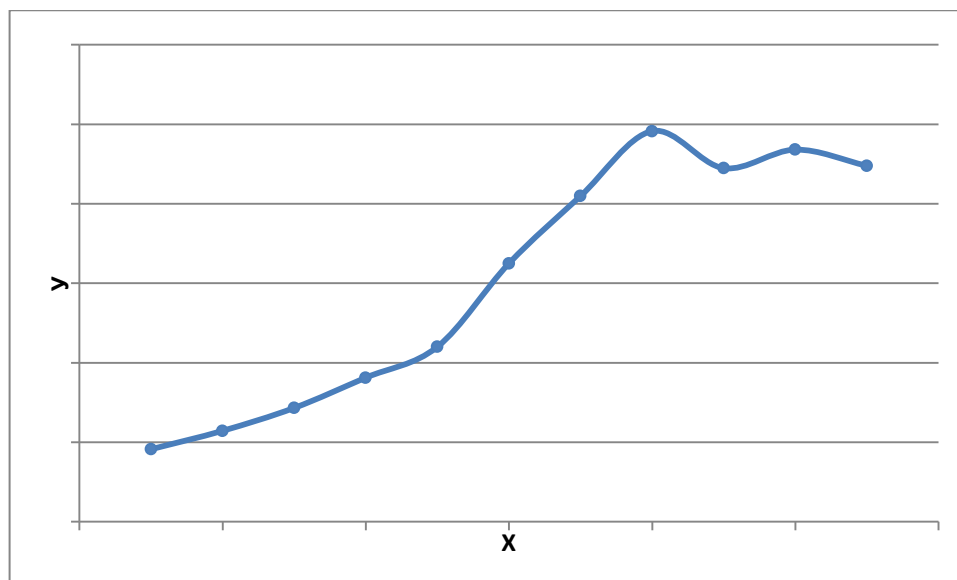


Рис.4.1. График функциональной зависимости

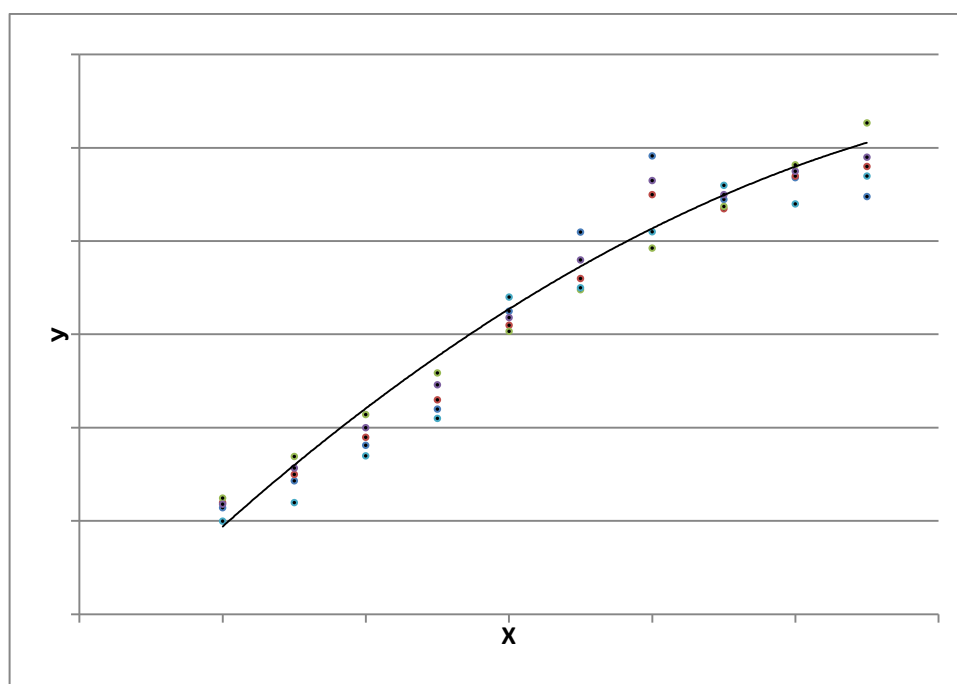


Рис.4.2. График корреляционной зависимости

Корреляционные зависимости могут быть установлены только при обработке большого количества наблюдений.

При корреляционном анализе решаются следующие задачи:

1. Устанавливается наличие корреляции (связи) между величинами.
1. Устанавливается форма линии связи (линии регрессии).
2. Определяются параметры линии регрессии.
3. Определяются достоверность установленной зависимости и достоверность отдельных параметров.

Вопрос 2. Определение формы связи при парной зависимости

В простейшем случае корреляционного анализа исследуется связь меж-

ду двумя показателями, из которых один рассматривается как независимый показатель-фактор (его величина обозначается через x), а второй – как зависимая переменная (её величина обозначается через y).

Наличие самой зависимости между этими показателями устанавливается, конечно, не математическим путём, а в результате качественного анализа, позволяющего вскрыть внутреннюю сущность изучаемого явления и порождающих его причин. Сам же корреляционный анализ предназначен для количественного измерения выявленной связи, хотя он нередко способствует и уточнению выводов самого качественного анализа.

Таким образом, ещё до математического расчёта считается установленным, что связь между независимым показателем-фактором x и зависимой переменной y существует (по меньшей мере, может существовать) и характеризуется функцией $y=f(x)$.

Одной из первых задач корреляционного анализа является установление вида этой функции, то есть отыскание такого корреляционного уравнения (иначе оно называется уравнением регрессии), которое наилучшим образом соответствует характеру изучаемой связи.

Коэффициент корреляции

Тесноту связи между двумя величинами можно определить при помощи коэффициента корреляции, который определяется по формуле:

$$r = \frac{N \sum xy - \sum x \sum y}{\sqrt{N \sum x^2 - (\sum x)^2} \times \sqrt{N \sum y^2 - (\sum y)^2}} \quad (4.1)$$

где x и y —текущие значения наблюдаемых величин; N — число наблюдений. Существует несколько модификаций данной формулы, наиболее простая из которых имеет вид:

$$r = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\sigma_x \cdot \sigma_y} \quad (4.2)$$

где $\overline{xy}$ —среднее значение произведения двух коррелируемых величин; $\bar{x}$ и $\bar{y}$ —средние значения этих величин; σ_x , σ_y —среднеквадратичные (стандартные) отклонения соответствующих величин.

Отметим основные свойства коэффициента корреляции (при достаточно большом объеме выборки n).

1. Коэффициент корреляции принимает значения на отрезке $[-1,1]$, т.е. $-1 \leq r \leq 1$.

В зависимости от того, насколько $|r|$ приближается к 1, различают связь слабую, умеренную, заметную, достаточно тесную, тесную и весьма

тесную. В случае если $r=0$, то линейной связи нет. Если $r=1$, то между двумя величинами существует функциональная связь. При положительном r наблюдается прямая связь, т.е. с увеличением независимой переменной x увеличивается зависимая - y . При отрицательном коэффициенте существует обратная связь - с увеличением независимой переменной зависимая переменная уменьшается.

2. Если все значения переменных увеличить (уменьшить) на одно и то же число или в одно и то же число раз, то величина коэффициента корреляции не изменится.

3. При $r = \pm 1$ корреляционная связь представляет линейную функциональную зависимость. При этом линии регрессии Y по X и X по Y совпадают и все наблюдаемые значения переменных располагаются на общей прямой.

4. При $r = 0$ линейная корреляционная связь отсутствует. При этом групповые средние переменных совпадают с их общими средними, а -линии регрессии Y по X и X по Y параллельны осям координат.

Количественно тесноту связи при множественной корреляции можно оценить с помощью множественного (совокупного) коэффициента корреляции R . Для расчета совокупного коэффициента корреляции необходимо определить парные коэффициенты корреляции r_{0i} между всеми факторами x_i , входящими в модель, и результирующим показателем y и все парные коэффициенты корреляции между факторами. Все коэффициенты корреляции записываются в квадратную симметричную матрицу.

Множественный коэффициент корреляции определяется по формуле:

$$R = \sqrt{1 - \frac{D}{D_{11}}} \quad (4.3)$$

где D —определитель матрицы парных коэффициентов корреляции; D_{11} —определитель той же матрицы с вычеркнутыми первой строкой и первым столбцом, т. е. определитель матрицы парных коэффициентов корреляции между факторами.

Для случая зависимости от двух факторов:

$$R_{y/x_1x_2} = \sqrt{\frac{r_{yx_1}^2 + r_{yx_2}^2 + 2 \times r_{yx_1} \times r_{yx_2} \times s_{x_1x_2}}{1 - r_{x_1x_2}^2}} \quad (4.4)$$

Для определения влияния только одного i -го фактора на результирующий показатель с исключением влияния других факторов используется частный коэффициент корреляции:

$$r_{y/x_i} = \frac{D_{1i}}{\sqrt{D_{11}D_{ii}}} \quad (4.5)$$

где D_{1i} —определитель матрицы с вычеркнутой первой строкой и i -м столбцом; D_{ii} —определитель матрицы с вычеркнутой i -й строкой и i -м столбцом.

При множественной корреляции от двух факторов коэффициент частной корреляции первого фактора:

$$r_{y/x_1x_2} = \frac{r_{yx_1} - r_{yx_2} \times r_{x_1x_2}}{\sqrt{(1-r_{yx_2}^2)(1-r_{x_1x_2}^2)}} \quad (4.6)$$

а коэффициент частной корреляции для второго фактора:

$$r_{y/x_1x_2} = \frac{r_{yx_2} - r_{yx_1} \times r_{x_1x_2}}{\sqrt{(1-r_{yx_1}^2)(1-r_{x_1x_2}^2)}} \quad (4.7)$$

Частный коэффициент корреляции отражает «чистое» влияние фактора на результирующий показатель.

Пример применения корреляционного анализа к решению практических задач.

Для определения факторов, которые оказывают или могут оказать влияние на состояние страховой компании, был проведен корреляционный анализ формализуемых показателей деятельности ста ведущих российских страховых компаний.

В качестве результирующего показателя соответствия страховой организации запросам страхователей был взят объем страховых взносов за соответствующий период.

Рассчитывался парный коэффициент корреляции показателей. Результаты расчетов сведены в таблицу.

Таблица 4.1 - Количественная оценка влияния факторов качества страховых компаний за 2012г.

<i>Показатели, влияющие на страховые взносы</i>	<i>Коэффициент корреляции</i>
<i>Размер Уставного капитала</i>	<i>0,65</i>
<i>Совокупные активы</i>	<i>0,67</i>

<i>Страховые резервы</i>	<i>0,74</i>
<i>Число лет работы на страховом рынке</i>	<i>0,46</i>
<i>Количество работников</i>	<i>0,49</i>
<i>Количество филиалов</i>	<i>0,55</i>
<i>Доля перестрахования в резервах</i>	<i>-0,38</i>
<i>Количество лицензированных видов страхования</i>	<i>0,63</i>

Оценка тесноты связи проводилась по граничным значениям коэффициента корреляции . На основе представленных данных были сделаны следующие выводы:

1. Максимальная теснота связи между показателями взносов и резервов свидетельствует о пропорциональности внутреннего и внешнего развития страховых организаций, так как увеличение взятых на себя обязательств привело к росту страховых резервов. О том же свидетельствует корреляция показателей взносов и размера совокупных активов.

2. Размер Уставного капитала страховщика оказывает влияние на выбор страхователя, так как при катастрофических убытках страховщик имеет право совершать выплаты за счет средств Уставного капитала.

3. Опыт работы компании на страховом рынке не оказывает сильного влияния на предпочтения страхователей, что может являться следствием недоверия к страховым компаниям после развала Госстраха.

4. С увеличением количества филиалов и страховых агентов взносы компаний растут.

5. Слишком большая доля перестрахования, либо вообще созданные псевдостраховые компании не являются привлекательными для клиентов, так как возрастает риск неплатежеспособности страховщика за счет сомнений в надежности перестраховщиков.

Вопрос 3. Автокорреляционная функция и мультиколлинеарность

Автокорреляционная функция состоит из значений корреляции между значениями исходного ряда и значениями этого же ряда, сдвинутыми на определенный шаг, который называют лаг (время запаздывания).



Рис.4.3. Автокорреляционная функция

Автокорреляционная функция показывает степень влияния прошлых значений исследуемой величины на текущие значения.

Глубина влияния (лаг) и сила влияния (значение корреляции) прошлого значения на текущее часто называют «память рынка», которая характеризует цикличность экономических явлений и процессов.



Рис.4.4. Изменение темпов роста инвестиций в основной капитал и ввода в действие объектов строительства

Автокорреляционная функция может охарактеризовать непосредственно динамику ряда.

Анализ:

1. Если первый член АКФ оказался самым максимальным, то изучаемый ценовой ряд является трендовым (то есть имеется зависимость изменения показателя от времени)

2. Если самым большим оказался коэффициент k -го порядка, то значит, изучаемый ценовой ряд содержит циклические колебания (с циклом, период которого составит k временных единиц).

3. Если ни один из коэффициентов не является значимым, то значит, ценовой ряд содержит сильную нелинейную тенденцию или это случайный ряд.

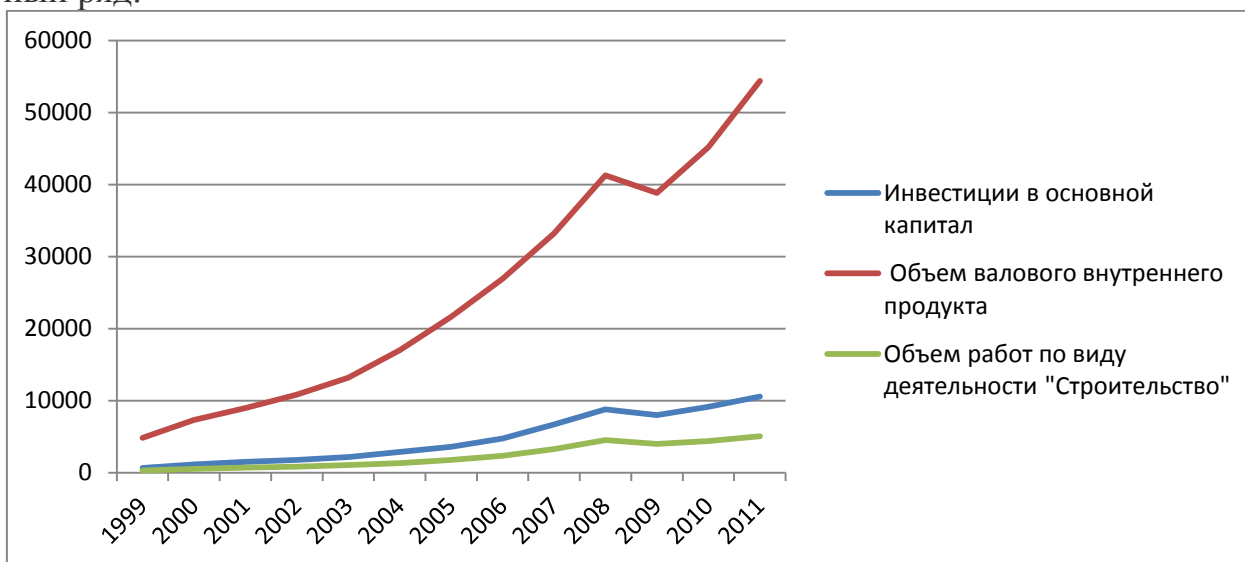


Рис. 4.5. Динамика показателей ВВП, инвестиций в основной капитал и объема работ по виду деятельности «Строительство» в 1999-2011 гг., млрд.руб.

Из графика динамики объема инвестиций, ВВП и объема работ по виду деятельности «Строительство» видно наличие линейного тренда. Это подтверждается и на коррелограмме.

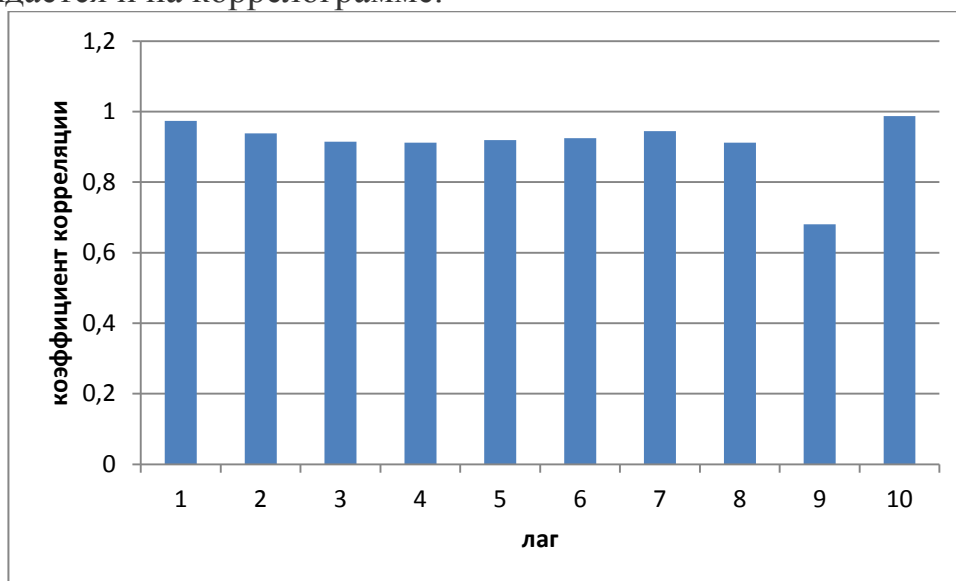


Рис. 4.6. Автокорреляционная функция объема инвестиций в основной капитал

Из коррелограммы видно, что данные обладают цикличностью с периодом 10 лет. Аналогичную автокорреляционную функцию имеет динамика ВВП и объема СМР. Причем, имеется резкий рост коэффициента автокорреляции, что подтверждает начало кризисной динамики в 2008 году.

Под **мультиколлинеарностью** понимают сильное взаимовлияние нескольких независимых переменных. Введение в модель таких факторов приводит к возникновению мультипликативного эффекта при оценке ошибки модели. Для предотвращения эффекта мультиколлинеарности оценивают значение парной корреляции между факторами модели, выявляя те, которые обладают сильной взаимосвязью. Далее из нескольких взаимозависимых выбирают один фактор, наиболее тесно связанный с зависимой переменной.

Пример применения корреляционного анализа к решению практических задач.

Необходимо выявить факторы, влияющие на величину потока денежных средств предприятия для построения эконометрической модели. Результаты корреляционного анализа сведены в таблицу.

Таблица 4.2 – Результаты корреляционного анализа факторов, влияющих на величину потока денежных средств предприятия

	Поток денежных средств (CF)	Эффективность работы аппарата управления	К-т обновления основных средств	Рентабельность продукции	Рентабельность капитала	Оборачиваемость оборотных активов	Оборачиваемость кредиторской задолженности	Фондоотдача	К-т ликвидности	Реальный собственный капитал	К-т обеспеченности собственными средствами	Платежеспособность	Обеспеченность заемными средствами	Чистая прибыль
Эффективность работы аппарата управления	0,33	1,00												
К-т обновления основных средств	0,45	-0,24	1,00											
Рентабельность продукции	0,42	0,37	-0,70	1,00										
Рентабельность капитала	0,51	0,28	-0,20	0,36	1,00									
Оборачиваемость оборотных активов	0,51	0,23	-0,35	0,31	0,60	1,00								
Оборачиваемость кредиторской задолженности	-0,50	0,49	-0,41	0,34	0,31	0,83	1,00							
Фондоотдача	0,67	-0,11	-0,09	-0,01	-0,11	-0,01	-0,08	1,00						

К-т ликвидности	0,80	0,20	-0,25	0,43	0,58	0,61	0,49	-0,53	1,00					
Реальный собственный капитал	0,66	0,24	-0,16	0,41	0,49	0,49	0,52	-0,46	0,85	1,00				
К-т обеспеченности собственными средствами	-0,26	-0,05	-0,12	0,20	0,20	0,34	0,27	-0,21	0,44	0,62	1,00			
Платежеспособность	0,34	-0,14	-0,29	0,21	0,09	0,40	0,26	-0,03	0,34	0,16	0,44	1,00		
Обеспеченность заемными средствами	-0,64	0,35	-0,40	0,43	0,31	0,53	0,64	-0,42	0,79	0,62	0,08	0,45	1,00	
Чистая прибыль	0,81	0,01	-0,20	0,55	0,68	0,22	0,02	0,18	0,35	0,46	0,29	0,10	0,10	1,00
Уровень использования мощности	0,44	0,71	-0,21	-0,09	-0,10	0,53	0,54	0,46	-0,19	-0,15	0,12	0,37	0,02	-0,21

Выбрать факторы, влияющие на поток денежных средств предприятия. Определить мультиколлинеарные зависимости между экзогенными (независимыми) переменными. Оценить характер связи между факторами.

Экспертно принимается (значения по модулю):

Коэффициент корреляции меньше 0,4 – связь слабая

От 0,4 до 0,6 – достаточная (средняя)

От 0,6 до 0,8 – достаточно сильная связь

Выше 0,8 – сильная связь.

Коэффициент корреляции фактора с самим собой=1.

Решение:

Смотрим столбец «поток денежных средств предприятия». Выбираем факторы, значения коэффициента корреляции которых с потоком денежных средств ближе к 1. Это чистая прибыль и ликвидность предприятия. Оценивают характер связи по знаку коэффициента корреляции: меньше нуля – обратная связь (с увеличением фактора другой уменьшается); больше нуля – прямая (с ростом одного показателя растет и другой). Например, коэффициент корреляции эффективности работы аппарата управления и потока средств предприятия=0,33 - связь прямая. Чем лучше управление, тем больше поток денежных средств. Но коэффициент корреляции низкий, значение меньше 0,4, значит, есть множество других факторов, влияющих на поток средств.

Связь между обеспеченностью заемными средствами и потоком средств предприятия обратная (коэффициент корреляции -0,64), то есть чем больше заемных средств, тем меньше поток средств предприятия. Связь достаточно сильная.

При мультиколлинеарности исключаются один из факторов, связь между которыми больше 0,85. (кроме связи с исследуемым показателем – потоком денежных средств). Смотрим столбцы, кроме «поток денежных средств предприятия». Коэффициент корреляции оборачиваемости обо-

ротных активов и оборачиваемости кредиторской задолженности равен 0,83. Следует исключить один из факторов, так как между ними есть сильная корреляция. Использование их вместе в модели влияния на поток денежных средств предприятия приведет к мультипликативному росту ошибки модели. Оставляем фактор, коэффициент корреляции которого с потоком средств больше. Для оборачиваемости активов коэффициент корреляции с потоком средств =0,51; для оборачиваемости кредиторской задолженности =0,5. Исключаем из модели оборачиваемость кредиторской задолженности. Другие мультиколлинеарные зависимости выделены зеленой заливкой.

Контрольные вопросы по теме 4.

1. Что такое поле корреляции?
2. Поясните смысл коэффициента корреляции, как оценить его значимость?
3. В чем смысл средней ошибки корреляции и как она определяется?