3.1. КОРРЕЛЯЦИОННЫЙ АНАЛИЗ ПРИ ПОСТРОЕНИИ ОДНОФАКТОРНЫХ ПРОГНОЗНЫХ МОДЕЛЕЙ

Прогнозирование социально-экономической динамики с помощью трендов можно осуществлять только для весьма небольшой части инерционных социально-экономических процессов. Именно инерционность, проявляющаяся в постоянстве тенденций развития, позволяет описывать эти тенденции уравнениями регрессий в виде зависимостей показателя от времени. Но представление о том, что некоторый показатель y_t изменяется количественно сам по себе без связи с другими процессами и факторами, характерно только для первого и очень приближённого представления. В этом случае априорно предполагается, что на этот прогнозируемый показатель некоторого социально-экономического процесса либо вовсе не оказывается никакого влияния со стороны других факторов и явлений, либо их взаимное воздействие на процесс и его показатель нивелируется и он меняется «не взирая» на своё окружение.

Поэтому простота, присущая моделям тренда предопределяет и невысокую прогнозную точность этих моделей. В реальной экономике нет ни одного изолированного от других процесса и, соответственно, его показателя. Все они находятся в сложной взаимосвязи, а, значит, и во взаимозависимости друг от друга. А поскольку в экономике все процессы являются взаимосвязанными, то и прогнозирование их следует осуществлять, выявляя и описывая эти взаимосвязи. Таким образом достигается большая описательная точность прогнозных моделей по сравнению с моделями трендов, ведь здесь в качестве основного выступает важный принцип современной прогностики – выявления и моделирования причинно-следственной связи. Ведь ещё Г.В.Лейбниц писал: «Аксиома, что ничего не бывает без основания, должна считаться одной из самых важных и плодотворных аксиом во всём человеческом познании; на ней основывается большая часть метафизики, физики и нравственного учения, и без неё нельзя ни доказать ни существования Бога из творений, ни построить доказательство от причин к следствиям или от следствий к причинам, ни сделать какие-либо выводы в делах гражданских»[1]. Поэтому, включая в прогнозную модель причинно-следственную связь, мы существенно повышаем точность этой модели по сравнению с моделью тренда.

Очень часто возможны ситуации, когда из нескольких факторов, оказывающих влияние на прогнозируемый показатель, один из них оказывается наиболее весомым, почему оказывается целесообразным выявить его из множества имеющихся и использовать для прогнозирования однофакторную зависимость изменения прогнозируемого показателя y_t от основного фактора x_t.

Поскольку в данной главе мы рассматриваем процессы, представляющие собой реализацию некоторых случайных причин, а имеющиеся в распоряжении прогнозиста статистические данные об этих процессах – как выборочные зна-

[1] Лейбниц Г.В. Соч.: В 4 т. – Т.3. – М., 1984. – С.141.

чения из генеральной совокупности, то к ним в полном объёме применим инструментарий математической статистики. Из всего многообразия инструментария современной математической статистики к задачам социально-экономического прогнозирования чаще всего применяют аппарат корреляционно-регрессионного анализа, базирующийся на выборочном методе. Следует вновь напомнить, что все выводы выборочного метода являются выводом по индукции и носят предположительный характер.

Под корреляционным анализом, как известно, понимается раздел математической статистики, нацеленный на выявление взаимосвязи между случайными факторами и оценки степени этой взаимосвязи.

Под регрессионным анализом понимается раздел математической статистики, посвящённый выбору формы регрессионной зависимости, оценки её коэффициентов и достоверности полученных результатов.

Рассмотрим вначале корреляционный анализ применительно к задаче социально-экономического прогнозирования, по умолчанию понимая, что все приводимые расчётные коэффициенты будут являться выборочными значениями. Основной задачей корреляционного анализа является выявление из множества факторов, связанных с прогнозируемым показателем, тех из них, влияние которых оказывается наиболее значимым. В однофакторном случае – из множества факторов выбрать один из них, чьё влияние на прогнозируемый показатель оказывается решающим. В корреляционном анализе не предполагается определять направление взаимосвязи, то есть – получить ответ на вопрос: y_t влияет на x_t или наоборот x_t влияет на y_t? Задача - выявить наличие взаимосвязи между двумя (в случае однофакторной зависимости) случайными величинами и оценить тесноту этой взаимосвязи.

Корреляционный анализ оперирует довольно обширным арсеналом различных коэффициентов и показателей, но о наличии взаимосвязи в итоге предлагается судить в основном по трём расчётным коэффициентам:

- корреляционному отношению;
- коэффициенту парной корреляции;
- коэффициенту детерминации.

Корреляционное отношение используется в ситуации, когда рассматриваются зависимости нескольких групп одного фактора от нескольких групп других факторов. Например, когда рабочие разных разрядов получают зарплаты разного уровня, а результаты их работы в зависимости от этого уровня изменяются. Тогда для подобной задачи можно разбить рабочих в несколько групп по уровню зарплаты, соответствующих разряду квалификации рабочего, например:

- крайне маленькая зарплата (от 2 до 5 денежных единиц),
- низкий уровень(от 5,1 до 8 денежных единиц),
- ниже среднего (от 8,1 до 11 денежных единиц),
- средний уровень (от 11,1 до 14 денежных единиц),
- чуть выше среднего (от 14,1 до 17 денежных единиц),
- выше среднего (от 17,1 до 20 денежных единиц),
- значительно выше среднего (от 20,1 до 23 денежных единиц),

- очень высокий (от 23,1 денежных единиц и выше).

В каждой из этих i групп зарплата конкретных рабочих варьируется вокруг некоторой средней $\bar{x}_i$, соответствующей каждому из разрядов. То есть, внутри каждой i-й группы есть некоторая внутригрупповая дисперсия σ^2_{xi}. Точно также и производительность труда этих рабочих, измеренная в некоторых единицах, варьируется. Эту производительность разного уровня также можно сгруппировать. И вовсе не обязательно, что число этих j групп будет совпадать с числом i групп рабочих, разбитых по уровню их зарплаты. Внутри каждой группы, соответствующей разным уровням производительности труда, также есть некоторый средний уровень $\bar{y}_j$ и отклонения от него, то есть – имеется собственная внутригрупповая дисперсия σ^2_{yj}. Такая ситуация типична для применения к её анализу корреляционного отношения. В этом случае можно вычислить выборочное корреляционное отношение y к x. Оно будет рассчитываться так. Для всех значений y рассчитывается общая средняя арифметическая $\bar{y}$ и общее среднее квадратическое отклонение σ_y. Затем для каждой группы j считаются внутригрупповые средние квадратические отклонения σ_{yj}, и их средняя по всем группам $\bar{\sigma}_y$. Искомое корреляционное отношение η_{yx} будет иметь вид:

$$\eta_{yx} = 1 - \frac{\bar{\sigma}_y}{\sigma_y}. \qquad (3.1.1)$$

Если в какой-нибудь j-й группе будет всего лишь одно наблюдение, либо все наблюдения в группе будут равны одному и тому же значению, то очевидно, что внутригрупповая дисперсия в такой группе будет равна нулю. Если это наблюдается в каждой группе, то есть – одному x соответствует одно и только одно y, то числитель дроби (3.1.1) будет равен нулю и корреляционное отношение η_{yx} будет равно единице, характеризуя ситуацию, когда между y и x имеется функциональная зависимость. Другой крайний случай, когда переменные независимы друг от друга и вне зависимости от того, какие значения принимает один фактор, разброс точек в каждой группе другого фактора равномерный, а средние каждой группы одинаковы. Тогда средняя $\bar{\sigma}_y$ будет равна общему среднему квадратическому отклонению σ_y, а это значит, что числитель дроби (3.1.1) равен знаменателю этой дроби и что корреляционное отношение η_{yx} будет равно нулю, характеризуя отсутствие какой-либо зависимости.

Точно также можно рассчитать и корреляционное отношение x к y:

$$\eta_{xy} = 1 - \frac{\bar{\sigma}_x}{\sigma_x}. \qquad (3.1.2)$$

Однако чаще всего встречаются на практике ситуации, когда имеются пары случайных значений $\{y_i;x_i\}$, то есть каждому y_i соответствует x_i и наоборот – каждому x_i соответствует y_i. Но поскольку каждая из переменных является случайной величиной, то переменной y одного уровня может соответствовать несколько переменных x, отличающихся друг от друга на случайную величину. Здесь нет групп и соответственно нет внутригрупповых дисперсий. Если формально применить к этой ситуации корреляционное отношение, то будет получена единица. Но это ни в коем случае не говорит о том, что между переменными имеется строгая функциональная зависимость - ведь по n точкам можно построить функциональную зависимость – полином *(n-1)*-й степени. Просто для данного случая корреляционное отношение неприменимо. Надо использовать другой инструмент измерения случайной взаимосвязи между факторами, или, иначе говоря, корреляционной зависимости между факторами.

Здесь основным инструментом выступает коэффициент парной корреляции. Рассмотрим его свойства. Чаще всего в математической статистике коэффициент корреляции случайных величин y и x рассматривают как отношение корреляционного момента μ_{xy} к произведению средних квадратических отклонений этих величин:

$$r_{xy} = \frac{\mu_{xy}}{\sigma_x \sigma_y}. \qquad (3.1.3)$$

Корреляционный момент случайных величин определяют как математическое ожидание произведения отклонений этих величин:

$$\mu_{xy} = M\{[x - M(x)][y - M(y)]\}. \qquad (3.1.4)$$

Зная свойства математических ожиданий, а именно, что математическое ожидание произведения независимых случайных величин равно произведению математических ожиданий этих сомножителей, получим:

$$\mu_{xy} = M\{[x - M(x)][y - M(y)]\} = M[x - M(x)]M[y - M(y)] = 0. \qquad (3.1.5)$$

Из чего делается вывод о том, что для двух независимых случайных величин коэффициент корреляции будет равен нулю, поскольку из (3.1.5) следует, что числитель формулы (3.1.3) равен нулю, а значит и сам коэффициент равен нулю. Необходимо обратить внимание на одностороннюю направленность этого утверждения, а именно - *для независимых случайных величин коэффициент корреляции равен нулю, но вовсе не наоборот.* Обратное утверждение о том, что если коэффициент корреляции равен нулю, то случайные величины являются независимыми, не является верным.

Действительно, первый правильный вывод имеет схему классического дедуктивного вывода и если нам известно, что две случайные переменные являются независимыми друг от друга, то мы обязательно будем утверждать, что для них коэффициент парной корреляции равен нулю. Вот схема этого вывода:

Все случайные переменные (все A), являющиеся независимыми (при условии B), имеют коэффициент парной корреляции равный нулю (имеют r=0). Поэтому если перед нами некоторые другие случайные переменные (D принадлежит к множеству A), которые являются независимыми (удовлетворяют условию B), то для них коэффициент парной корреляции равен нулю (имеют r=0).

Или: для всех A, удовлетворяющих условию B, выполняется r=0. D принадлежит к A и удовлетворяет условием B, следовательно, для него выполняется r=0.

Вывод дедуктивный и является истинным.

А теперь рассмотрим структуру обратного вывода.

Две случайные переменные имеют коэффициент парной корреляции, равный нулю. Поскольку этот коэффициент равен нулю для независимых переменных, то эти две переменные не зависят друг от друга. Покажем, что этот вывод не является истинным.

Вот логическая структура этого вывода.

Для всех случайных переменных (для всего множества A), являющихся независимыми (которые удовлетворяют условию B), коэффициент парной корреляции равен нулю (обладают r=0). Случайные переменные (некоторая часть D множества A) имеют коэффициент парной корреляции, равный нулю (имеет свойство r=0). Следовательно, они также являются независимыми (они удовлетворяют условию B).

Или: для всех A, удовлетворяющих условию B, выполняется r=0. D принадлежит к A и имеет r=0. Значит, для D выполняется условие B.

Вывод индуктивный, поскольку свойством *r=0* могут обладать и другие элементы. Это значит, что данный вывод является предположительным. Действительно в этой ситуации возможно наличие параллельного индуктивного вывода, а именно – *для всех A, удовлетворяющих условию C, также выполняется равенство r=0. Из чего следует, что если D принадлежит A и для него r=0, то D может удовлетворять и условию C.*

Для того чтобы был понятен смысл этого предупреждения, приведём простой пример первого, дедуктивного, и второго, индуктивного, логических выводов.

Первый вывод: все члены партии «Единая Россия» являются гражданами России. Этот человек – член партии «Единая Россия», следовательно, он гражданин России. Это – дедуктивный вывод и он является истинным выводом. А вот другой вывод: все члены партии «Единая Россия» являются гражданами России. Авторы этих строк – граждане России, следовательно – они члены партии «Единая Россия». Это вывод – индуктивный, носит вероятностный характер, а потому и не является истинным. Тем более что авторы этих строк не являются членами партии «Единая Россия», а очень даже наоборот, считают её источником основных проблем и угроз реальной демократии в России.

Точно также и индуктивный вывод о том, что равенство нулю коэффициента парной корреляции свидетельствует о независимости двух случайных переменных, не является истинным, он даёт некоторую вероятностную оценку, которая может оказаться и ложной. Надо всегда иметь это в виду.

В математической статистике доказывается, что корреляционный момент μ_{xy} лежит в пределах:

$$-\sigma_x\sigma_y \le \mu_{xy} \le \sigma_x\sigma_y . \qquad (3.1.6)$$

Это двойное неравенство, как легко заметить, может быть заменено на такое:

$$\left|\mu_{xy}\right| \le \sigma_x\sigma_y . \qquad (3.1.7)$$

Если теперь обратиться к формуле коэффициента парной корреляции (3.1.3), то, воспользовавшись неравенством (3.1.7) получаем ещё одно свойство этого коэффициента:

$$\frac{\left|\mu_{xy}\right|}{\sigma_x\sigma_y} = \left|r\right| \le 1 , \qquad (3.1.8)$$

причём строгое равенство достигается в том случае, когда между исходными переменными имеется строго функциональная линейная зависимость.

Откуда следует дедуктивный вывод о том, что для двух случайных переменных, между которыми имеется линейная зависимость, коэффициент парной корреляции по модулю будет равен единице. Обратный вывод о том, что если между двумя переменными коэффициент парной корреляции равен единице, то между ними имеется линейная функциональная зависимость, является индуктивным, и поэтому не является истинным.

Простой пример, подтверждающий это - если в некоторой больнице одному из пациентов поставлена капельница и количество влитых в его организм капель меняется строго линейно функционально относительно времени:

$x_t=at$,

а в некотором банке кассир, пересчитывающий пачку 100-рублёвых денег, записывает сумму так, что она меняется строго линейно функционально относительно времени как

$y_t=bt$,

то выборочное значение коэффициента корреляции между этими двумя рядами $\{x_t\}$ и $\{y_t\}$ будет равно единице. Но коэффициент парной корреляции, равный единице для этого случая, очевидно, вовсе не означает, что между этими двумя рядами имеется строгая линейная функциональная зависимость. Между ними, как это ясно из сути процессов, вообще никакой взаимосвязи нет и быть не может.

Итак, по значениям коэффициента парной корреляции нельзя судить о степени корреляционной взаимосвязи между факторами, поскольку такой вывод – индуктивный и носит предположительный характер.

Для корректного применения этого коэффициента необходимо предварительно обосновать гипотезу о наличии корреляции между двумя случайными переменными, а только после этого подтвердить или опровергнуть эту гипотезу с помощью коэффициента корреляции.

Поскольку с генеральной совокупностью никто из экономистов не работает, а работает только с выборочными значениями из генеральных совокупностей, то непосредственное применение вышеприведённой формулы для расчёта коэффициента парной корреляции невозможно. Необходимо использовать выборочное значение корреляционного момента и выборочные значения средних квадратических.

Центральный выборочный корреляционный момент рассчитывается по формуле:

$$\mu_{xy}=\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y}). \qquad (3.1.9)$$

Выборочные значения средних квадратических рассчитываются так:

$$\sigma_x=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2}, \qquad (3.1.10)$$

$$\sigma_y=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\overline{y})^2}. \qquad (3.1.11)$$

Сделав замену показателей в формуле (3.1.3) на их выборочные значения (3.1.9) – (3.1.11), получим выборочное значение коэффициента парной корреляции:

$$r = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_i (x_i - \overline{x})^2 \sum_i (y_i - \overline{y})^2}}. \qquad (3.1.12)$$

Любой коэффициент тем и хорош, что с его помощью можно с той или иной степенью успешности диагностировать некоторую анализируемую ситуацию. Для того чтобы понять свойства выборочного коэффициента парной корреляции, и уметь правильно его применять в практике прогнозирования, рассмотрим другой способ получения формулы (3.1.9)[1].

Запишем уравнение линейной регрессии:

$$y_i = a_0 + a_1 x_i. \qquad (3.1.13)$$

Коэффициенты этой модели можно найти с помощью МНК. Но для упрощения процесса нахождения выборочных значений этих коэффициентов произведём центрирование исходных переменных относительно их средних:

$$y_i - \overline{y}, \quad x_i - \overline{x}.$$

Тогда легко найти значение коэффициента регрессии с помощью формулы Маркова:

$$a_1 = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{\sum_i (x_i - \overline{x})^2}. \qquad (3.1.14)$$

Находить значение свободного члена не будем. Вспомним, что коэффициент a_1 характеризует тангенс угла α наклона линии регрессии к оси *0x* (рис. 3.1).

Поскольку на этом этапе исследования нам не известно направление зависимости, а мы изучаем сам факт её существования, то вполне возможно существование и другого направления зависимости, а именно:

$$x_i = b_0 + b_1 y_i. \qquad (3.1.15)$$

Для центрированных значений исходных переменных относительно их средних арифметических легко найти значение коэффициента регрессии с помощью МНК. Формула Маркова для этого случая будет записана так:

$$b_1 = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{\sum_i (y_i - \overline{y})^2}. \qquad (3.1.16)$$

Этот коэффициент, как известно, характеризует угол β наклона прямой линии к оси *0y*, а именно – он равен тангенсу этого угла (рис. 3.1.1).

Из рисунка легко заметить, что сумма углов α и β меньше π/2:

$$\alpha + \beta \le \pi/2. \qquad (3.1.17)$$

Откуда со всей очевидностью имеем:

---

[1] Венецкий И.Г., Венецкая В.И. Основные математико-статистические понятия и формулы в экономическом анализе: Справочник. - М.: Статистика, 1979. - 447 с.

$tg\alpha \cdot tg\beta \le 1$. (3.1.18)

Причём строгое равенство достигается только в том случае, когда сумма углов будет равна π/2, то есть, когда две линии регрессии совпадают. Чем дальше эти две линии регрессии друг от друга, тем дальше сумма углов от π/2 и дальше от единицы произведение тангенсов этих углов (3.1.18).

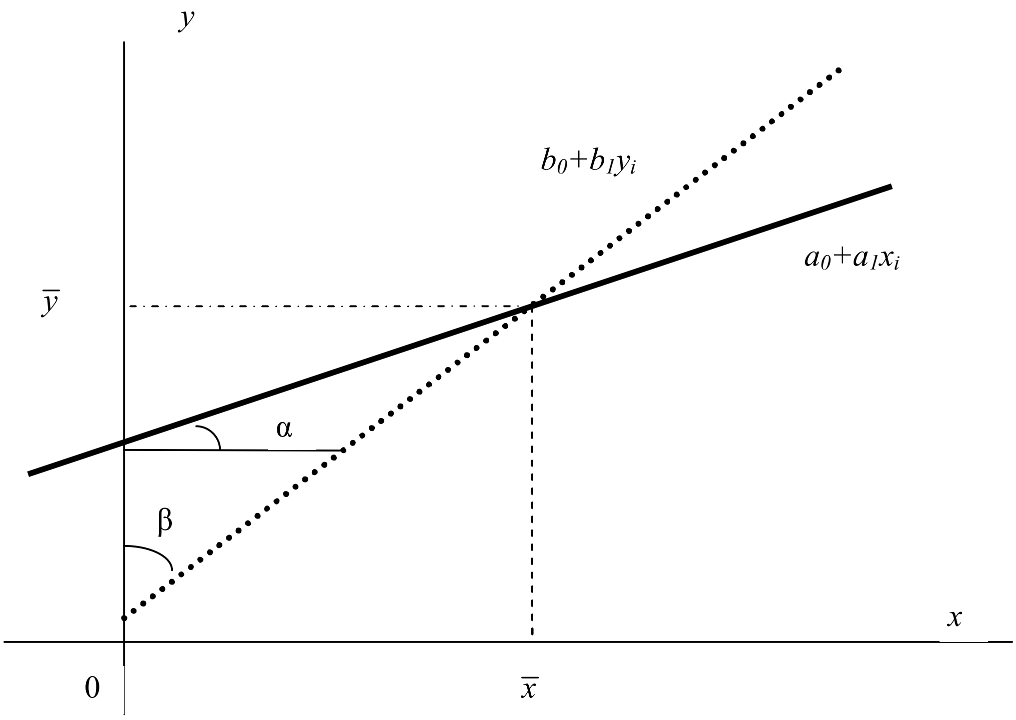

Рис. 3.1.1. Графическое изображение линий регрессии *x* на *y* и *y* на *x*.

Значит, по величине (3.1.18) можно судить, насколько линии регрессии близки друг к другу и насколько, в конечном итоге, зависимость между двумя случайными величинами приближается к линейной.

Подставив в (3.1.18) вместо тангенсов углов их числовые выражения – a_1 и b_1, найденные ранее (3.1.14), (3.1.16) и, исчислив квадратный корень из этого произведения, получим:

$$\sqrt{tg\alpha \cdot tg\beta} = \sqrt{a_1 b_1} = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_i (x_i - \overline{x})^2 \sum_i (y_i - \overline{y})^2}}. \qquad (3.1.19)$$

Сравнив с (3.1.12) можно сделать вывод о том, что мы получили формулу выборочного значения коэффициента парной корреляции. Поэтому можно рассматривать *выборочное значение коэффициента парной корреляции как среднегеометрическое из коэффициентов регрессии, оценённых с помощью МНК.*

А теперь можно сделать вывод о том, что же характеризует выборочное значение коэффициента парной корреляции. Он показывает, насколько зависимость между случайными переменными, если она есть, приближается к линейной. И только! Как было показано выше, из тех или иных значений коэффициента парной корреляции вовсе не следует вывод о наличии или отсутствии зависимости между случайными переменными. Первична гипотеза о наличии такой взаимосвязи, а затем – инструментальное подтверждение этой гипотезы с помощью коэффициента парной корреляции.

Если поступать наоборот, то можно столкнуться с явлением *ложной корреляции – когда два совершенно независимых случайных показателя имеют высокое значение коэффициента парной корреляции.* Теперь ясно, что если два независимых показателя имеют монотонную тенденцию своего изменения, Например, один возрастает, а другой убывает, то, нанеся эти точки на график рис. 3.1, можно убедится в том, что они лягут на линию, с той или иной степенью приближающуюся к прямой, а коэффициент парной корреляции между ними будет показывать эту степень.

Таким образом, для того, чтобы избежать явления ложной корреляции, следует вначале исследования с помощью аналитического анализа обосновать наличие зависимости между случайными переменными, а затем, с помощью корреляционного анализа подтвердить наличие этой зависимости.

Если по модулю коэффициент парной корреляции близок к единице, то это подтверждает гипотезу о том, что между переменными может существовать линейная взаимосвязь.

Если по модулю коэффициент парной корреляции меньше, чем 0,8, но всё же выше, чем 0,6, то можно говорить о том, что зависимость между переменными может быть представлена как линейная, но с очень большой погрешностью.

Если модуль коэффициента парной корреляции меньше, чем 0,6, то следует искать другую нелинейную форму зависимости.

Ни в коем случае низкое значение коэффициента корреляции не свидетельствует об отсутствии взаимосвязи вообще. Оно свидетельствует только об отсутствии линейной взаимосвязи. Например, для нелинейной функциональной зависимости $y_i = \sin x_i$ коэффициент парной корреляции между x и y будет близок, почти равен нулю. Но равенство нулю коэффициента парной корреляции в этом случае, очевидно, не говорит об отсутствии взаимосвязи между y_t и x_t вообще, а свидетельствует именно об отсутствии *линейной* взаимосвязи между ними. поскольку функциональная нелинейная взаимосвязь очевидна.

Но что делать исследователю, если он убеждён в наличии взаимосвязи между случайными факторами, а коэффициент парной корреляции близок по модулю к нулю? Как подтвердить наличие или отсутствие взаимосвязи вообще? Для этого можно использовать графический анализ этой взаимосвязи. Вполне возможно, что точки на графике лягут так, что исследователь сможет предполагать наличие, например, экспоненциальной зависимости:

$$y_i = a_0 e^{a_1 x_i}.$$

Тогда для подтверждения этой гипотезы следует прологарифмировать левую и правую части равенства. Получим:

$$\lg y_i = \lg a_0 + a_1 x_i$$

и коэффициент парной корреляции необходимо находить между $\lg y_i$ и x_i. Близость его к единице по модулю будет подтверждать высказанную гипотезу.

Таким точно образом можно поступать и при выдвижении гипотез о наличии нелинейных зависимостей другого вида – показательной зависимости, степенной зависимости, квадратичной зависимости, логарифмической зависимости и т.п.

Коэффициент детерминации, который также предлагает аппарат корреляционного анализа, представляет собой ни что иное, как квадрат коэффициента парной корреляции:

$$R = r^2 . \qquad (3.1.20)$$

Он имеет простой и ясный смысл – его значения показывают, насколько процентов линейная динамика одного показателя соответствует линейной динамике другого показателя. Например, если коэффициент парной корреляции равен 0,9, то коэффициент детерминации будет равен в соответствии с (3.1.20) квадрату коэффициента парной корреляции, то есть – 0,81. Это говорит, что изменение каждого из показателей на 81% может быть объяснено линейным влиянием другого показателя, а 19% этого изменения может быть отнесено к влиянию иных факторов. Опять же следует указать на то, что данное утверждение носит предположительный характер, поскольку он носит индуктивный характер. Для того чтобы оценить, насколько можно доверять подобному высказыванию, необходимо оценить доверительные границы для расчётных выборочных значений как коэффициента парной корреляции, так и коэффициента детерминации.

Основной вывод, который следует из материалов данного параграфа, заключается в том, что корреляционный анализ так и не даёт ответ на главный вопрос – наличие или отсутствие взаимосвязи между случайными факторами. Его основные коэффициенты – парной корреляции и коэффициента детерминации, - позволяют лишь оценить степень приближения случайной зависимости к линейной, если эта зависимость обоснована аналитически.

В этой связи интересные перспективы открывает совместное использование коэффициента парной корреляции и коэффициента согласия в динамике[1].

Суть коэффициента согласия в динамике заключается в следующем. Хорошо известен факт, что по всем N наблюдениям о показателе y_i ($i=1,2,3,\dots,N$),имеющимся в распоряжении исследователя, можно построить степенную функцию (N-l)-й степени:

$$\hat{y}_i = a_0 + a_1 i + a_2 i^2 + a_3 i^3 + \dots + a_{N-1} i^{N-1} \qquad (3.1.21)$$

Причём этот полином так опишет исходный ряд наблюдений $\{y_i\}$, что его расчётные значения $\hat{y}_i$ в каждой i-й точке в точности будут соответствовать фактическим значениям y_i:

$$\varepsilon_i = y_i - \hat{y}_i = 0 \text{ для каждого } i \ . \qquad (3.1.22)$$

Коэффициенты полинома (3.1.21) определяют динамические свойства анализируемого ряда, так как ошибка аппроксимации при этом будет равна

[1] Светуньков С.Г. Количественные методы прогнозирования эволюционных процессов. – Ульяновск: Изд-во УлГУ, 1997. – С.62 – 74.

нулю. Поэтому по их значениям можно судить об исходном ряде и его свойствах. Очевидно, что каждый ряд, отличающийся в своей динамике от другого ряда, будет отличаться от него и своими коэффициентами, если оба ряда представить в виде модели (3.1.21). А если ряды в динамике меняются во многом одинаково, то и коэффициенты модели их разложения в степенную функцию будут во многом совпадать. Поэтому сравнительный анализ коэффициентов при одинаковых степенях может позволить получить дополнительную информацию о взаимосвязи факторов.

Но для того, чтобы осуществить такой сравнительный анализ, необходимо вычислить коэффициенты модели разложения функции в степенной ряд (3.1.21). Для этого есть несколько возможных вариантов:

1) построить степенные функции (N-1)-й степени (3.1.21), решая системы из N уравнений с N неизвестными коэффициентами a,

2) воспользоваться разложением функций в степенные ряды (ряды Тейлора) и анализировать соответствующие члены степенного ряда,

3) оценить параметры степенных функций с помощью аппарата конечных разностей,

4) найти различные производные степенной функции (3.1.21) с помощью метода конечных разностей, поскольку они вполне характеризуют соответствующие коэффициенты.

Первый вариант предполагает решение системы N уравнений с N неизвестными, что в ряде случаев при большом числе наблюдений N представляет собой непростую вычислительную задачу, которую рекомендовать к широкому практическому использованию неразумно. Полиномы высоких степеней очень неустойчивы к ошибкам округлений.

Второй вариант предполагает представление самой функции в явном виде, разложение которой и необходимо осуществить, а она не известна. Впрочем, существует большое количество различных методов, позволяющих найти приближенные коэффициенты рядов Тейлора. Однако указанное множество методов вызвано именно различной степенью неточности такого приближения и поэтому этот вариант не может быть решён с удовлетворительной точностью.

Третий вариант - оценить параметры степенных функций с помощью аппарата конечных разностей - значительно проще, чем все рассмотренные выше варианты, однако и он достаточно громоздок. Рассмотрим эту возможность более подробно, так как на основе этого рассмотрения можно будет увидеть особенности простого использования четвёртого варианта.

Пусть нам дана линейная однофакторная модель

$$\hat{y}_i = a_0 + a_1 i, \qquad (3.1.22)$$

параметры которой a_0 и a_1 следует найти на множестве наблюдений N.

Метод конечных разностей исходит из того факта, что первая производная линейной функции (3.1.22) будет равна a_1, а одной из оценок этой производной может быть средняя первая разность.

Обозначим через $\Delta_i^1 y_i$ первую разность ряда в точке i. Эта разность вычисляется элементарно:

$$\Delta^1 y_i = y_i - y_{i-1}. \qquad (3.1.23)$$

Эта разность характеризует первую производную функции в точке i. Общей характеристикой первой производной на всем множестве значений может служить средняя этих разностей. Но поскольку первых разностей на единицу меньше, чем количество исходных членов ряда, то эта средняя будет вычисляться со второго члена ряда до последнего N-го:

$$\bar{\Delta}^1 y = \frac{1}{N-1}\sum_{i=2}^{N}\Delta^1 y_i . \qquad (3.1.24)$$

Подставляя в (3.1.24) значения разностей, полученные с помощью (3.1.23), и осуществив элементарные сокращения, получим окончательно:

$$\bar{\Delta}^1 y = \frac{1}{N-1}(y_N - y_1) \qquad (3.1.25)$$

Эта средняя разность характеризует первую производную функции и дает некоторую оценку коэффициента a_1 линейной функции:

$$a_1 \approx \bar{\Delta}^1 y . \qquad (3.1.26)$$

Для нахождения параметра a_0 модели (3.1.22) по известному значению коэффициента a_1 необходимо найти средние значения Y и i, подставить их в (3.1.22) и, решив уравнение, определить значение a_0.

Такую же процедуру можно сделать с другим рядом, который обозначим через X_i:

$$X_i = b_0 + b_1 i,$$

после чего можно сравнить друг с другом коэффициенты a_0 и b_0, a_1 и b_1.

Если теперь предположить наличие тенденции роста, описываемой с помощью параболы второй степени:

$$\hat{y}_i = a_0 + a_1 i + a_2 i^2, \qquad (3.1.27)$$

то для определения параметра a_2 этой модели необходимо рассчитать вторые разности:

$$\Delta^2 y_i = \Delta^1 y_i - \Delta^1 y_{i-1} = (y_i - y_{i-1}) - (y_{i-1} - y_{i-2}) = y_i - 2y_{i-1} + y_{i-2}, \qquad (3.1.28)$$

а затем найти среднюю вторых разностей. Поскольку вторых разностей на единицу меньше, чем первых, средняя вторых разностей вычисляется так:

$$\bar{\Delta}^2 y = \frac{1}{N-2}\sum_{i=3}^{N}\Delta^2 y_i . \qquad (3.1.29)$$

Известно, что вторая разность характеризует вторую производную любой функции, а применительно к функции (3.1.27) она будет характеризовать значение коэффициента a_2, поскольку вторая производная указанной функции равна $2a_2$.

Подставляя в (3.1.29) значения вторых разностей из (3.1.28), которые в свою очередь определяются через первые разности (3.1.23), получим после ряда элементарных сокращений:

$$\bar{\Delta}^2 y = \frac{1}{N-2}(y_N - y_{N-1} - y_2 + y_1) = \frac{1}{N-2}(\Delta^1 y_N - \Delta^1 y_2) \qquad (3.1.30)$$

Откуда легко найти оценку коэффициента a_2:

$$a_2 = \frac{\bar{\Delta}^2 y}{2}. \qquad (3.1.31)$$

Теперь для нахождения неизвестных значений коэффициентов a_0 и a_1 следует воспользоваться вышеизложенной для линейной модели процедурой и последовательно найти a_1, а затем a_0..

Продолжая подобным образом, можно найти коэффициенты полинома третей, четвёртой и вообще любой степени N. Так, например, для полинома третьей степени третья разность будет находиться так:

$$\Delta^3 y_i = \Delta^2 y_i - \Delta^2 y_{i-1} = (y_i - 2y_{i-1} + y_{i-2}) - (y_{i-1} - 2y_{i-2} + y_{i-3}) = y_i - 3y_{i-1} + 3y_{i-2} - y_{i-3}. \quad (3.1.32)$$

Средняя третьих разностей равна:

$$\bar{\Delta}^3 y = \frac{1}{N-3}(y_N - 2y_{N-1} + y_{N-2} - y_3 + 2y_2 - y_1) = \frac{1}{N-3}(\Delta^2 y_N - \Delta^2 y_3). \quad (3.1.33)$$

Эта разность в свою очередь характеризует третью производную функции, равную для параболы третьей степени $6a_3$. Тогда коэффициент a_3 можно найти так:

$$a_3 = \frac{\bar{\Delta}^3 y}{6}. \quad (3.1.34)$$

Очевидно, что трудоемкость нахождения параметров параболы (N-1)-й степени меньше, чем трудоёмкость построения степенных функций (N-1)-й степени (решая системы из N уравнений с N неизвестными коэффициентами) или трудоёмкость разложения функций в степенные ряды Тейлора. Однако и его можно упростить, вычисляя лишь конечные разности, которые характеризуют конечные производные функций.

Как видно из вышеприведённых рассуждений, для оценки коэффициентов полиномов a_1, a_2 и a_3 найденные оценки производных с помощью конечных разностей делились для коэффициента a_1 на 1 (3.1.26),для коэффициента a_2 - на 2 (3.1.31), для коэффициента a_3 – на 6 (3.1.34) и т.п. В случае расчета оценок производных надобность в подобной процедуре умножения коэффициентов при показателях, возводимых в одинаковую степень, отпадает, что значительно облегчает расчеты. Кроме того, при расчете средних разностей сумму разностей было необходимо делить на число этих разностей. Без этой процедуры также можно обойтись. Действительно, если необходимо относить вторую производную данного ряда со второй производной другого ряда, то от умножения или деления каждой производной на одну и ту же константу, не равную нулю, результат отношения не изменится. Поэтому процедура еще более упрощается - следует находить не среднюю разностей, а просто их сумму. Сравнив друг с другом формулы для вычисления конечных разностей (3.1.23), (3.1.28) и (3.1.32), легко обнаружить закономерность, с помощью которой можно сгенерировать формулу для вычисления конечной разности любого порядка. Но поскольку подобная задача уже давно решена применительно к интерполяционной формуле Ньютона, воспользовавшись этим, запишем формулу вычисления конечной разности n-го порядка в точке i:

$$\Delta^n y_i = \sum_{k=0}^{n} (-1)^k C_n^k y_{i-k}, \quad (3.1.34)$$

где k - порядковый номер члена разности (3.1.34), k=1,2,3,…,n;

C - число всех сочетаний из n различных элементов по k, которое, как известно, находится по формуле:

$$C_n^k = \frac{n!}{k!(n-k)!}, \qquad (3.1.35)$$

в которой знак «!»- это знак факториала. Напомним, что по определению 0!=1.

Сумма n-х конечных разностей, т.е. неусредненная оценка n-й производной данного ряда, определяется по формуле:

$$\tilde{\Delta}^n y = (\Delta^{n-1} y_N - \Delta^{n-1} y_n), \qquad (3.1.36)$$

где $\Delta^{n-1} y_N$ – последняя n-я конечная разность в ряду вычисляемых n-х конечных разностей для исходных значений y_i, то есть n-я разность при i=N,

$\Delta^{n-1} y_n$ – первая в ряду вычисляемых n-х конечных разностей, когда наблюдение i=n.

Поскольку специалист в области математических методов в экономике без труда может вычислить конечные разности (3.1.34), а на их основе сумму этих конечных разностей (3.1.36), то для любых двух рядов $\{y_i\}$ и $\{x_i\}$, состоящих из N наблюдений, можно получить (N-1) пар чисел:

$$(\overline{\Delta}^1 y;\overline{\Delta}^1 x),(\overline{\Delta}^2 y;\overline{\Delta}^2 x),(\overline{\Delta}^3 y;\overline{\Delta}^3 x),...,(\overline{\Delta}^{N-1} y;\overline{\Delta}^{N-1} x). \qquad (3.1.37)$$

которые характеризуют первые, вторые, третьи и т.д. до (N-1)-й производные указанных рядов. С учетом того, что производные наилучшим образом характеризуют динамику процессов, сравнение указанных пар значений может характеризовать степень согласия в динамике рядов $\{y_i\}$ и $\{x_i\}$. Очевидно, что сравнение пар значений производных следует производить так, чтобы итогом данного сравнения стал некоторый расчетный коэффициент, имеющий ясное смысловое толкование. Сам коэффициент будет характеризовать согласие двух исходных рядов в динамике, поэтому его имеет смысл назвать именно так.

Для того чтобы предложить формулу вычисления коэффициента согласия в динамике воспользуемся неравенством Коши, которое для конечных сумм имеет вид:

$$(\sum_i y_i x_i)^2 \le \sum_i y_i^2 \sum_i x_i^2 \; . \qquad (3.1.38)$$

Если левую часть указанного неравенства разделить на его правую часть, будет получено выражение, которое никогда не будет больше единицы.

$$\frac{(\sum_i y_i x_i)^2}{\sum_i y_i^2 \sum_i x_i^2} \le 1. \qquad (3.1.39)$$

Равенство единице может быть только в том случае, когда ряд значений $\{y_i\}$ представляет собой линейное преобразование ряда $\{x_i\}$.

Если вместо пар значений y_i и x_i подставить в предлагаемую формулу пары значений сумм конечных разностей (3.1.37) и извлечь квадратный корень из полученного выражения, то получим формулу, внешне напоминающую формулу для расчета коэффициента парной корреляции, который и назовём «коэффициент согласия в динамике»:

$$k_s = \frac{\sum_i \bar{\Delta}^i y \bar{\Delta}^i x}{\sqrt{\sum_i (\bar{\Delta}^i y)^2 \sum_i (\bar{\Delta}^i x)^2}}. \qquad (3.1.40)$$

Тогда к этому коэффициенту можно будет применить толкование, аналогичное толкованию значений коэффициента парной корреляции.

Поскольку сумма разностей может быть как положительной, так и отрицательной, то и итоговое значение коэффициента согласия в динамике может быть как положительным, так и отрицательным. Для того чтобы понять, что именно отражает коэффициент согласия в динамике, представим, что ряд $\{y_i\}$ *и ряд* $\{x_i\}$ *находятся в линейно функциональной зависимости друг от друга, но под воздействием внешних по отношению к ним факторов они меняются от наблюдения к наблюдению по сложному нелинейному закону. В этом случае поскольку друг с другом они связаны линейной функциональной зависимостью, то их производные будут также линейно зависеть друг от друга, а неравенство Коши превращается в таком случае равенство, поэтому коэффициент (3.1.40) будет равен единице.*

Тогда можно утверждать, что в случае если значение коэффициента (3.1.40) близко по модулю к единице, динамика двух исходных рядов $\{y_i\}$ *и* $\{x_i\}$ *находится в очень сильной степени согласия друг с другом, поскольку значения их производных разного порядка соответствуют друг другу. Согласованность их динамики свидетельствует о том, что между ними весьма вероятна взаимосвязь.*

Если же вычисленное значение модуля коэффициента (3.1.40) близко к нулю, следует признать, что динамика двух рядов различна и наличие взаимосвязи между ними вряд ли возможно – ведь производные высших порядков рассогласованы, а это значит, что динамика одного ряда $\{y_i\}$ *не похожа на динамику другого ряда* $\{x_i\}$ *и его изменение вряд ли вызвано действием фактора, отражаемого рядом значений* $\{x_i\}$*. То есть, в отличие от коэффициента парной корреляции, значение коэффициента согласия в динамике, близкое к нулю, однозначно указывает на то, что между двумя рядами никакой взаимосвязи нет.*

Так о чём же свидетельствуют значения коэффициента согласия в динамике, лежащие по модулю в пределах от нуля до единицы? Для ответа на этот вопрос продемонстрируем использование данного коэффициента на условных данных.

Если ряд $\{y_t\}$ *представляет собой синусоиду, изменяющуюся во времени t, а ряд* $\{x_t\}$ *представляет собой косинусоиду с той же амплитудой и фазой:*

$y_t = \sin t;\ x_t = \cos t,$

то для 100 точек, снятых в течение одного периода этих гармонических функций, коэффициент парной корреляции составил величину, равную

$r=-0.0013674,$

что свидетельствует об очевидном отсутствии линейной взаимосвязи. При этом, кстати, очень многие практикующие экономисты сделают вывод об *отсутствии взаимосвязи между y_t, и x_t вообще. Коэффициент согласия в динамике дает величину, равную*

k_s*=0.9796672,*

что свидетельствует о наличии очень сильного соответствия в динамике показателей. А именно так оно и есть. Это означает, что у исследователя, получившего такое значение коэффициента согласия в динамике, есть все основания предполагать наличие взаимосвязи между факторами.

Вывод, который вытекает из этого примера – если для изучаемых рядов коэффициент парной корреляции по модулю близок к нулю, а модуль коэффициента согласия при этом близок к единице, это свидетельствует о наличии тесной нелинейной взаимосвязи между исследуемыми факторами.

Для подтверждения хороших информационных свойств коэффициента согласия в динамике для выявления возможной взаимосвязи между факторами проведем расчеты еще на одном примере, условные данные которого приведены в таблице 3.1.1.

Таблица 3.1.1 Условные данные для расчета

t	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
y_t	*0*	*1*	*2.5*	*4*	*7*	*7*	*6*	*7*	*8.5*
x_t	*1*	*1.5*	*2*	*3*	*2*	*1.5*	*2*	*2.5*	*3.0*

Данные подобраны таким образом, чтобы резкий рост ряда x_t, на четвертом наблюдении и его последующее падение вплоть до седьмого наблюдения могли бы быть причиной аналогичной динамики показателя y_t со сдвижкой во времени – показатель y_t растет на пятом наблюдении и начинает падение на следующем шаге вплоть до восьмого наблюдения (иначе говоря, имеется лаг в единицу между ними). Это наглядно видно из графика рисунка 3.1.2.

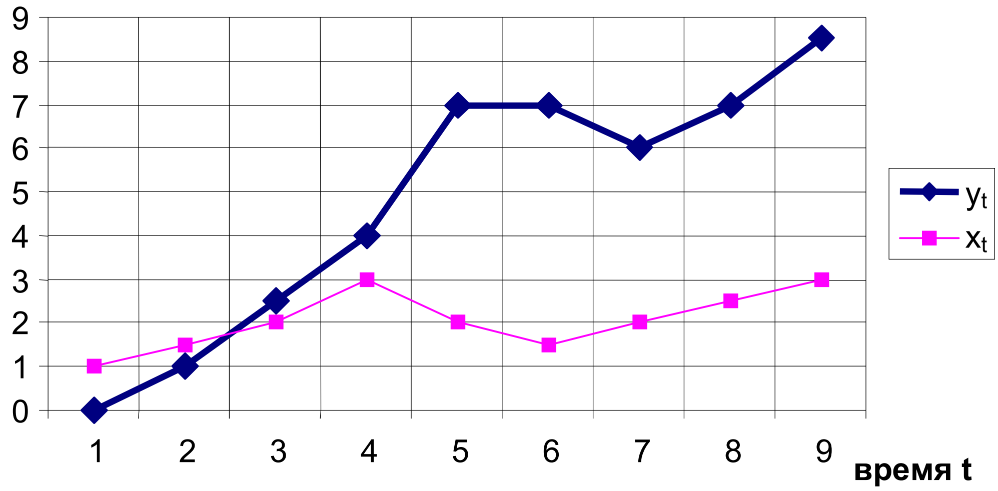

Рис. 3.1.2. Динамика рядов табл.3.1.1

Расчеты коэффициента парной корреляции и коэффициента согласия в динамике для этих двух рядов показывают следующее. Коэффициент парной корреляции между анализируемыми рядами равен

$r=0.58283525705,$

а коэффициент согласия в динамике оказался равным

$k_s=0.9918264168.$

Различие между этими коэффициентами довольно велико. Для того чтобы понять причину такого различия, посмотрим на график зависимости y_t *от* x_t *(рис. 3.1.3).*

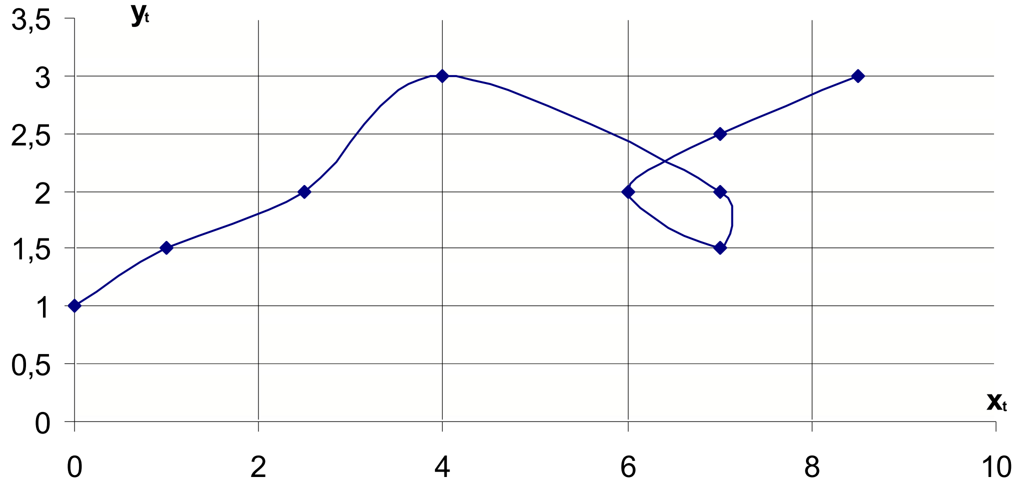

Рис. 3.1.3. График совместного изменения рядов y_t и x_t табл. 3.1.1

Как видно из графика, предположить наличие зависимости между этими двумя рядами сложно. Кроме того, если рассматривать их так, как будто они засорены некоторыми случайными ошибками, и попытаться построить прямую линию, проходящую через них, можно убедиться в том, что ошибка аппроксимации такой линии будет очень высокой, что и демонстрирует значение коэффициента парной корреляции, равное 0.58283525705. Коэффициент парной корреляции не «уловил» возможную взаимосвязь между двумя рядами. Знание того, как ведут себя эти два ряда, которые мы наглядно получили из рисунков 3.1.2 и 3.1.1, показывает, как надо было бы действовать в этом случае. Для того чтобы «уловить» взаимосвязь между факторами с помощью классических методов корреляционного анализа, необходимо осуществить расчет лагов построением коррелограмм, что является трудоемкой работой и в экономической практике никогда не делается. Высокое значение коэффициента согласия в динамике выявляет наличие этой сложной взаимосвязи между рядами, которое формальный корреляционный анализ не подтверждает. Коэффициент согласия в динамике, напротив, достаточно высок и свидетельст-

вует о возможности взаимосвязи между факторами, так как характеры динамики факторов близки друг к другу.

Вновь следует вывод - если для изучаемых рядов коэффициент парной корреляции мал, а модуль коэффициента согласия при этом близок к единице, это свидетельствует о наличии тесной взаимосвязи между исследуемыми факторами, возможно – нелинейной, а возможно – с лагами.

Указанные примеры показывают, что предлагаемый коэффициент согласия в динамике действительно дает исследователю новую информацию, весьма важную для того, чтобы решить задачу определения степени взаимосвязи между факторами. Для того чтобы окончательно убедиться в этом, воспользуемся следующим изменением характера рядов, которое описывается функциональной зависимостью между y_t *и* x_t*:*

$$x_t=t,\ y_t=x_t^{\,b}\ . \qquad (3.1.41)$$

Здесь коэффициент b *задается на множестве значений от -4 до +4. Количество наблюдений возьмём равным* $N=40$ $(t=1,2,3,\dots,40)$*.*

Очевидно, что при разных значениях показателя степени b *между* x_t *и* y_t*, формируется различной степени нелинейные зависимости. Исключением являются два значения показателя степени* b*. Первый случай, когда* $b=1$*. В этом случае, очевидно, между* x_t *и* y_t*, наблюдается линейная функциональная зависимость. Второй случай, когда* $b=0$*. В этом случае один фактор возрастает (*x_t*), а другой показатель (*y_t*) не меняется, поскольку равен единице (любое число в нулевой степени равно единице).*

Результаты расчетов как коэффициента парной корреляции, так и коэффициента согласия в динамике, приведены в таблице 3.1.2.

Приведенные в таблице результаты расчетов позволяют сделать ряд интересных выводов.

Во-первых, характеры динамики зависимостей двух сравниваемых коэффициентов в зависимости от изменения показателя степени b *в заданных пределах, в общем, совпадают. Если показатель степени* b *лежит в пределах:*

$$-\infty < b < 0,$$

то и коэффициент согласия в динамике, и коэффициент парной корреляции являются отрицательными, демонстрируя то, что между рядами обратная зависимость – рост x_t *ведёт к уменьшению* y_t*. А если показатель степени положителен:*

$$b > 0,$$

то оба коэффициента также становятся положительными. К тому же, при

$$b=1$$

оба коэффициента также равны единице.

Во-вторых, динамика коэффициента парной корреляции распадается на два участка, точкой раздела которых является значение показателя степени b*, равное нулю, при котором функция не определена и переход через нуль показателя степени в этой точке характеризует наличие в динамике коэффициента парной корреляции «проколотую точку». Динамика же коэффициента согласия не имеет никакого разрыва, что говорит о его большей устойчивости, чем у коэффициента парной корреляции.*

Таблица 3.1.2.

Результаты расчетов коэффициентов для зависимости (3.1.41)

b	**коэффициент согласия в динамике, k_s**	**коэффициент парной корреляции, r**
-4.0	-0.053262	-0.122896
-3.0	-0.110767	-0.439260
-2.0	-0.160509	-0.522314
-1.0	-0.501095	-0.707623
-0.70	-0.706631	-0.781567
-0.50	-0.804549	-0.830307
-0.30	-0.865595	-0.875419
-0.20	-0.892118	-0.895841
-0.15	-0.919842	-0.905411
-0.10	-0.943193	-0.914523
-0.05	-0.672157	-0.923160
-0.01	-0.540874	-0.929718
-0.005	-0.097326	-0.930516
0	0	0/0 – неопределённость
0.005	0.059462	0.932096
0.01	0.157010	0.932879
0.05	0.523571	0.938959
0.10	0.948079	0.946109
0.15	0.867250	0.952755
0.20	0.911257	0.958901
0.30	0.986885	0.969719
0.50	0.987293	0.985793
0.70	0.995242	0.995323
1.0	1	1
2.0	0.995121	0.971348
3.0	0.984165	0.922050
4.0	0.979655	0.873017

В третьих, в широком диапазоне изменения показателя степени b, а именно:

$-0{,}3 \le b \le 4{,}0$

коэффициент парной корреляции убедительно свидетельствует о возможной и весьма тесной линейной взаимосвязи между факторами, хотя она таковой является только при b=1.

Коэффициент согласия в динамике показывает высокую степень согласованности динамики факторов в таких диапазонах:

$-0{,}3 \le b \le -0{,}1$ *и*

$0{,}1 \le b \le 4{,}0.$

Таким образом, можно сделать вывод о том, что коэффициент согласия в динамике можно использовать в качестве дополнительного инструмента, позволяющего уточнить наличие взаимосвязи между факторами.

Совместное использование сравниваемых коэффициентов, дает прогнозисту значительно больше информации, чем простое использование каждого из них в отдельности. Коэффициент парной корреляции показывает исследователю - насколько тенденции двух изучаемых рядов совпадают в их линейном росте, а коэффициент согласия в динамике показывает исследователю - насколько вариации одного показателя соответствуют вариациям другого. В любом случае необходимо твердо помнить, что окончательное решение о том являются ли ряды взаимосвязанными друг с другом или нет, следует принимать, опираясь на результаты экономического анализа изучаемой взаимосвязи.