

Тема 1. Обзор понятий и формулы вычисления: ковариации, дисперсии и корреляции

Эконометрика: основные понятия и определения

Эконометрика – это наука, изучающая методами математической статистики количественные закономерности и связи в экономике, выражаемые в виде математических моделей.

Целевое назначение эконометрики – эмпирический вывод экономических закономерностей.

Основные задачи эконометрики состоят в построении моделей, выражающей выводимые закономерности, оценка их параметров и проверка гипотез о закономерностях изменения и связях экономических показателей.

Процессы эконометрического анализа могут характеризоваться двумя типами обрабатываемых данных: пространственными данными и временными рядами.

Пространственные данные – это относящиеся к одному и тому же моменту времени данные о каком-либо экономическом показателе, характеризующем однотипные объекты. Например, данные об объеме производства на разных промышленных предприятиях за один и тот же период времени или о количестве работников разных промышленных предприятий в один и тот же момент времени.

Временные ряды – это данные о каких-либо показателях, характеризующих одни и те же объекты в различные моменты времени. К такому типу данных относятся ежемесячные статистические данные за ряд лет по стране в целом или по отдельным регионам. Например, по объему промышленного производства или о количестве безработных. Особенность временных данных состоит в том, что они упорядочены во времени.

Наиболее распространены три основных класса **эконометрических моделей**: регрессионные модели с одним уравнением, системы одновременных уравнений и модели временных рядов.

Регрессионная модель – это уравнение, в котором объясняемая переменная представляется в виде функции от объясняющих переменных (например, модель спроса на некоторый товар в зависимости от его цены и дохода покупателей). По виду функции различают **линейные** и **нелинейные** регрессионные модели. Наиболее детально изучены и потому наиболее часто встречается в эконометрическом анализе методы оценки и анализа линейных регрессионных моделей.

Системы одновременных уравнений представляют собой системы уравнений, состоящие из регрессионных уравнений и тождеств, в каждом из которых помимо объясняющих – независимых – переменных содержатся объясняемые переменные из других уравнений системы. Пример: система, включающая уравнение спроса, уравнение предложения и тождество –

уравнение равенства спроса и предложения, характеризующее рыночное равновесие.

К простейшим **моделям временных рядов** относятся **модели тренда** и **модели сезонности**. Тренд представляет собой устойчивое изменение уровня показателя в течение длительного времени. Сезонность характеризует устойчивые внутригодовые колебания уровня показателя. К более сложным моделям временных рядов относятся, например, модель адаптивного прогноза и авторегрессионная модель. Основная особенность моделей этого класса состоит в том, что они объясняют поведение временного ряда исходя из его предыдущих значений.

Основные задачи эконометрических исследований

Эконометрическая модель, как правило, основана на теоретическом предположении о круге взаимосвязанных переменных и характере связи между ними. При всем стремлении к «наилучшему» описанию связей приоритет отдается качественному анализу. Поэтому в качестве этапов эконометрического исследования можно указать:

- постановку проблемы;
- получение данных, анализ их качества;
- спецификацию модели;
- оценку параметров;
- интерпретацию результатов.

На начальном этапе решения любой эконометрической задачи необходимо сформулировать эконометрическую модель, т.е. представить модель в виде уравнений, характеризующих связи между экономическими показателями. Например, уравнение связи между доходами семей (x) и сбережениями семей (y), которое необходимо получить путем обработки результатов опроса нескольких сотен случайно отобранных семей:

$$y = \alpha + \beta \cdot x + \varepsilon,$$

где:

- x – объясняющая (независимая) переменная (доходы семей);
- y – объясняемая (зависимая) переменная (сбережения семей);
- ε – случайная составляющая (ошибка);
- α и β – параметры уравнения, заранее не известные и подлежащие

определению в результате эконометрического анализа задачи.

При решении любой задачи эконометрики необходима проверка соответствия полученной модели реальным экономическим данным. Если модель соответствует реальным данным, то возникает задача определения (оценки) параметров модели. Различают два уровня анализа: теоретический и эмпирический.

На **теоретическом** уровне предполагается, что известны все возможные реализации экономических показателей (т.е. имеется вся генеральная совокупность в целом). Теоретически параметры модели можно

оценить, если известны (или предполагаются заданными) статистические свойства генеральной совокупности. Как правило, все возможные исходы (т.е. возможные значения показателей) заранее неизвестны; на практике можно наблюдать только выбранные значения интересующих показателей, т.е. выборочную совокупность.

На *эмпирическом* уровне на основе выборочной совокупности нельзя точно определить значения параметров модели, можно лишь получить их оценки, являющиеся случайными величинами. Таким образом, цель оценивания параметров состоит в получении как можно более точных значений неизвестных параметров модели, которые характерны для всей генеральной совокупности.

Одной из основных задач экономических исследований является анализ зависимости между переменными (показателями), которая может быть функциональной (встречается очень редко) или статистической (в экономике, как правило, является преобладающей).

Функциональная зависимость (иначе ее называют детерминированной) задается в виде формулы, которая каждому значению одной переменной ставит в соответствие строго определенное значение другой переменной, при этом воздействием случайных факторов пренебрегают.

Статистическая зависимость – это связь переменных, на которую накладывается воздействие случайных факторов, при этом изменение одной переменной приводит к изменению математического ожидания (т.е. наиболее вероятного ожидаемого значения) другой переменной. Наиболее распространенной формулой статистической связи между переменными является уравнение регрессии. Если эта формула линейная (нелинейная), то регрессию называют линейной (нелинейной). Многие нелинейные модели можно преобразовать в линейные.

Основные понятия теории вероятностей и математической статистики

Случайная величина характеризуется тем, что под воздействием случайных факторов она может с определенными вероятностями принимать те или иные значения из некоторого множества чисел. Случайная величина называется дискретной, если она принимает отдельные, изолированные друг от друга значения, и непрерывной, если множество ее значений непрерывно заполняет некоторый числовой промежуток.

Дискретную случайную величину, число возможных значений которой конечно, обычно представляют в виде ряда распределения, состоящего из пары чисел, одно из которых – значение величины, другое – вероятность появления этого значения, при этом сумма вероятностей появления всех значений равна 1.

Характеристикой непрерывной случайной величины является функция распределения, указывающая вероятность того, что эта случайная величина принимает значение, меньше заданной величины. Всему диапазону

изменения случайной величины соответствует единичное значение функции распределения.

К основным числовым характеристикам случайных величин относятся математическое ожидание (наиболее вероятное ожидаемое значение), дисперсия (вариация) и среднее квадратическое отклонение.

Математическим ожиданием дискретной случайной величины называется сумма произведений всех ее значений на соответствующие вероятности:

$$M(x) = \sum_{i=1}^n x_i \cdot p_i ,$$

где:

$M(x)$ – математическое ожидание случайной величины x ;

$x_i - i$ – значение случайной величины x ;

p_i – вероятность появления i - го значения случайной величины x ;

i – порядковый номер дискретного значения случайной величины x ;

n – общее число дискретных значений случайной величины x .

Математическим ожиданием непрерывной случайной величины называется интеграл:

$$M(x) = \int x \cdot f(x) \cdot dx ,$$

где:

$f(x)$ – плотность распределения случайной величины x , представляющая собой производную по x функции распределения случайной величины x ;

\int – интеграл, который берется на всем интервале, в котором определена случайная величина x ;

dx – дифференциал случайной величины x .

Для большого числа случайных величин, с которыми имеют дело в эконометрике, предполагается нормальное или близкое к нему распределение. Для случайной величины (m_x), имеющей нормальное распределение, математическое ожидание равно среднему значению генеральной совокупности.

Теоретическая (генеральная) дисперсия случайной величины определяется как математическое ожидание квадрата отклонения случайной величины x относительно ее математического ожидания:

$$D(x) = M(x - m_x)^2 .$$

Среднее квадратическое отклонение случайной величины x , характеризующее степень отклонения в среднем случайной величины в совокупности от своего среднего значения, представляет собой корень квадратный из ее дисперсии:

$$\sigma_x = \sqrt{D(x)} .$$

Данные о случайных величинах, которые используются в эконометрическом анализе, обычно представляются ограниченной выборкой, математическое ожидание которой оценивается выборочной средней, т.е. средним арифметическим значений случайной величины в выборке:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i ,$$

где:

\bar{x} – выборочная средняя,

x_i – i -е значение случайной величины x ,

i – порядковый номер выборочного значения случайной величины x ,

n – общее число данных в выборке.

Выборочная дисперсия (вариация) представляет собой среднее арифметическое квадратов отклонений случайной величины от среднего значения:

$$var(x) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 .$$

Выборочное среднеквадратическое отклонение случайной величины x представляет собой корень квадратный из выборочной дисперсии:

$$\sigma_x = \sqrt{var(x)} .$$

Характеристики генеральной совокупности (т.е. всего возможного набора показателей) обычно неизвестны, поэтому они оцениваются на основе характеристик выборочной совокупности (т.е. ограниченного числа значений показателей). Характеристики генеральной совокупности принято называть параметрами, а выборочной совокупности – оценками. Чтобы выборочная оценка давала хорошее приближение оцениваемого параметра, она должна удовлетворять требованиям несмещенности, эффективности и состоятельности.

Несмещенность является желательным свойством, так как только в этом случае они могут иметь практическую значимость. Оценка называется **несмещенной**, если ее математическое ожидание равно оцениваемому параметру при любом объеме выборки, т.е. математическое ожидание остатков равно нулю. Например, выборочное среднее является несмещенной оценкой математического ожидания \bar{x} генеральной совокупности – генеральной средней m_x :

$$m_x = \bar{x} .$$

Итак, если несмещенность имеет место, то при большом числе полученных выборочных оценок искомого параметра остатки не будут накапливаться, и потому найденный параметр регрессии можно рассматривать как среднее значение из возможно большого количества несмещенных оценок. Если оценки обладают свойством несмещенности, то их можно сравнивать по разным выборкам.

Оценку, не являющуюся несмещенной, называют **смещенной**. Например, выборочная дисперсия $var(x)$ является смещенной оценкой генеральной дисперсии. В качестве несмещенной оценки этой дисперсии используется уточненная величина (исправленная дисперсия):

$$S^2 = \frac{n}{n-1} var(x) = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2,$$

где:

S^2 – несмещенная оценка дисперсии генеральной совокупности;
 S – несмещенная оценка стандартного отклонения генеральной совокупности;

n – число измерений в выборке;

$x_i - i$ – значение измеренного показателя в выборке;

i – порядковый номер измерения.

Для практических целей важна не только несмещенность, но и эффективность оценок. Несмещенная оценка называется **эффективной**, если она имеет минимальную дисперсию по сравнению с другими выборочными оценками. Поэтому несмещенность оценки должна дополняться минимальной дисперсией. В практических исследованиях это означает возможность перехода от точечного оценивания к интервальному. Пример: выборочная средняя \bar{x} является эффективной оценкой генеральной средней, так как она имеет наименьшую дисперсию в классе несмещенных оценок.

Степень реалистичности доверительных интервалов параметров регрессии обеспечивается, если оценки будут не только несмещенными и эффективными, но и состоятельными.

Оценка называется **состоятельной**, если при увеличении объема выборки (т.е. если $n \rightarrow \infty$) она стремится к оцениваемому параметру. Примером состоятельной оценки математического ожидания генеральной совокупности (генеральной средней m_x) является выборочное среднее \bar{x} .

Состоятельность оценок характеризует увеличение их точности с увеличением объема выборки. Большой практический интерес представляют те результаты регрессии, для которых доверительный интервал ожидаемого значения параметра регрессии имеет предел значений вероятности, равный единице. Иными словами, вероятность получения оценки на заданном расстоянии от истинного значения параметра близка к единице.

Меру связи между двумя случайными величинами x и y характеризуют выборочная ковариация и коэффициент корреляции. **Выборочной ковариацией двух случайных величин x и y** называется среднее арифметическое произведений отклонений значений этих величин от своих выборочных средних:

$$cov(x, y) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x}) \cdot (y_i - \bar{y}),$$

где:

$cov(x, y)$ – ковариация случайных величин x и y ;
 x_i и y_i – i -е значения случайных величин x и y ;
 \bar{x} и \bar{y} – средние значения случайных величин x и y ;
 i – порядковый номер дискретного значения пар случайных величин x и y ;

n – общее число дискретных значений пар случайных величин x и y .

Коэффициент корреляции определяется выражением:

$$r_{xy} = \frac{cov(x, y)}{\sqrt{var(x) \cdot var(y)}} = \frac{cov(x, y)}{\sigma_x \cdot \sigma_y},$$

где:

$cov(x, y)$ – ковариация случайных величин x и y ;

$var(x)$ и $var(y)$ – вариации случайных величин x и y ;

σ_x и σ_y – стандартные отклонения случайных величин x и y .

Коэффициент корреляции является безразмерной величиной и показывает степень линейной связи двух переменных:

$r > 0$ при положительной связи и $r = 1$ при строгой положительной линейной связи;

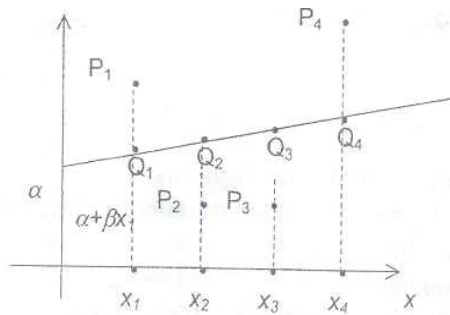
$r < 0$ при отрицательной связи и $r = -1$ при строгой отрицательной линейной связи;

$r = 0$ при отсутствии линейной связи.

Случайные величины x и y называются **некоррелированными**, если $r = 0$, и коррелированными, если $r \neq 0$. Независимые случайные величины x и y всегда некоррелированные (т.е. $r = 0$), но из некоррелированности случайных величин x и y не следует их независимость. Некоррелированность указывает лишь на отсутствие линейной связи между переменными, но не на отсутствие связи между ними вообще.

Тема 2. Парный регрессионный анализ

Рассмотрим некоторый экономический объект (процесс, явление, систему) и выдели только две переменные, характеризующие объект. Обозначим переменные буквами X и Y . Будем предполагать, что независимая (объясняющая) переменная X оказывает воздействие на значения переменной Y , которая, таким образом, является зависимой переменной, т.е. имеет место зависимость: $Y = f(x)$.



В модели парной линейной регрессии зависимость между переменными в генеральной совокупности представляется в виде:

$$y = \alpha + \beta * x + u$$

где x – неслучайная величина, а y и u – случайные величины.

Величина y называется объясняемой (зависимой) переменной, а x – объясняющей (независимой) переменной. Постоянные α и β – параметры уравнения. Наличие случайного члена u (ошибки регрессии) связано с воздействием на зависимую переменную других неучтенных в уравнении факторов, с возможной нелинейностью модели и ошибками измерения.

На основе выборочного наблюдения оценивается выборочное уравнение регрессии (линия регрессии):

$$\hat{Y} = a + b * x$$

Где (a, b) – оценка параметра (α, β) .

Неизвестные значения (a, b) определяются методом наименьших квадратов (МНК). Суть МНК заключается в минимизации суммы квадратов остатков.

Угловым коэффициентом связи (параметр b) т.н. коэффициент наклона, показывает на сколько единиц в среднем изменяется переменная y при увеличении независимой переменной x на единицу. Вычисляется по формуле:

$$b = \frac{Cov(x, y)}{D_b(x)}$$

Свободный член уравнения (параметр a) – это постоянная, которая дает прогнозируемое значение зависимой переменной при $x = 0$. Это может иметь смысл в зависимости от того, как далеко находится $x = 0$ от выборочного значения x . Определяется по формуле:

$$a = \bar{y} - b\bar{x}$$

Тема 3. Свойства коэффициентов регрессии и проверка гипотез

Коэффициенты регрессии **a** и **b**, полученные по любой выборке, представляются в виде суммы двух слагаемых: 1) постоянной величин, равной истинному значению коэффициентов **α** и **β**; 2) случайной

составляющей, зависящей от $\text{Cov}(\mathbf{x}, \mathbf{u})$, которой обусловлены отклонения коэффициентов \mathbf{a} и \mathbf{b} от констант α и β .

Поэтому свойства коэффициентов регрессии существенно зависят от случайной составляющей \mathbf{u} .

Эксперимент по методу Монте – Карло

Для решения вероятностных задач, в которых не удаётся установить формальную зависимость конечного результата от исходных данных, т.е. получить аналитическое решение задачи, используется метод Монте - Карло (метод статистических испытаний).

Основное понятие будет объяснено посредством следующей аналогии. Предположим, что свинья обучена находить трюфели. Это дикорастущие земляные грибы, встречающиеся во Франции и Италии и считающиеся деликатесом. Они дороги, так как их трудно найти, и хорошая свинья, обученная поиску трюфелей, стоит дорого. Проблема состоит в том, чтобы узнать, насколько хорошо свинья ищет трюфели. Для выяснения этого вы могли бы закопать трюфели в нескольких местах, отпустить свинью и посмотреть, сколько грибов она обнаружит. Посредством такого контролируемого эксперимента можно было бы непосредственно оценить степень успешности поиска.

Какое это отношение имеет к регрессионному анализу? Проблема в том, что мы никогда не знаем истинных значений α и β (иначе зачем бы мы использовали регрессионный анализ для их оценки?). Поэтому мы не можем сказать хорошие или плохие оценки даёт наш метод. Эксперимент по методу **Монте – Карло** – это искусственный контролируемый эксперимент, дающий возможность такой проверки. Простейший возможный эксперимент по этому методу состоит из трёх частей. Во- первых:

- 1) выбираются «истинные» значения α и β ;
- 2) в каждом наблюдении выбирается значения \mathbf{x} ;
- 3) используется некоторый процесс генерации случайных чисел (или берётся последовательность из таблицы случайных чисел) для получения значений случайного фактора \mathbf{u} в каждом из наблюдений.

Во -вторых, в каждом наблюдении рассчитывается значение \mathbf{y} с использованием соотношения $\mathbf{y} = \alpha + \beta\mathbf{x} + \mathbf{u}$ и значений $\alpha, \beta, \mathbf{x}, \mathbf{u}$.

В- третьих, применяется регрессионный анализ для оценивания параметров α и β с использованием только полученных указанным образом значений \mathbf{y} и данных \mathbf{x} . При этом вы можете увидеть, являются ли α и β хорошими оценками α и β , и это позволит почувствовать пригодность метода для построения регрессии.

На первых двух шагах проводится подготовка к применению регрессионного метода. Мы полностью контролируем модель, которую создаём, и знаем «истинные» значения параметров, потому что сами определили. На третьем этапе мы определяем, может ли поставленная нами задача решаться с помощью метода регрессии, т.е. могут ли быть получены хорошие оценки при использовании только данных об \mathbf{y} и \mathbf{x} . Заметим, что

проблема возникает вследствие включения случайного фактора в процессе получения y .

Итак, очевидно, что свойства коэффициентов регрессии существенным образом зависят от свойств случайной составляющей. Для того чтобы регрессионный анализ, основанный на обычном **методе наименьших квадратов**, давал наилучшие результаты, случайный член должен удовлетворять четырём условиям, известным как **условия Гаусса – Маркова**.

1. Случайный член регрессии в каждом наблюдении должен иметь нулевое математическое ожидание $M(u_i) = 0$ для любого i .
2. Дисперсия случайного члена регрессии не зависит от номера наблюдения: $\sigma^2(u_i)$ – не зависит от i .
3. Случайные члены регрессии в разных наблюдениях не зависят друг от друга

$Cov(u_i, u_j) = 0$, если i не равно j .

4. Случайный член регрессии и объясняющая переменная в каждом наблюдении независимы друг от друга $Cov(u_k, x_k) = 0$ для любого k .
- Дадим следующую формулировку теоремы Гаусса – Маркова: **Если выполнены все условия Гаусса – Маркова для модели парной регрессии, то «метод наименьших квадратов» даёт несмещённые, эффективные и состоятельные оценки параметров регрессии α и β .**

Наряду с этими условиями обычно также предполагается нормальность распределение случайного члена.

Стандартные отклонения и стандартные ошибки коэффициентов регрессии

Теоретические дисперсии оценок a и b задаются следующими выражениями:

$$\text{var}(a) = \frac{\sigma_u^2}{n} \left\{ 1 + \frac{x_{cp}^2}{\text{Var}(x)} \right\} \quad (1.33)$$

$$\text{var}(b) = \frac{\sigma_u^2}{n \text{Var}(x)} \quad (1.34)$$

Из этих уравнений можно сделать три очевидных заключения. Во первых, дисперсия a прямо пропорциональна дисперсии остаточного члена $\sigma^2 u$. Во вторых, чем больше число наблюдений, тем меньше дисперсия оценок. В третьих, чем больше дисперсия x , тем меньше будет дисперсия коэффициентов регрессии.

Несмещённая оценка $\sigma^2 u$ получается из выражения:

$$s_u^2 = \frac{n}{n-2} \text{Var}(e) \quad (1.35)$$

Теперь вспомним следующие определения:

Стандартное отклонения случайной величины (s.d.) – это корень квадратный из теоретической дисперсии случайной величины; среднее ожидаемое расстояние между наблюдениями этой случайной величины и её математическим ожиданием.

Стандартная ошибка случайной величины (с.о.) – это оценка случайной величины, полученная по данным выборки.

Используя уравнения (1.33), (1.34), можно получить оценки теоретических дисперсий для a и b и после извлечения квадратного корня – оценки их стандартных отклонений.

$$s.d.(a) = \sqrt{\frac{\sigma_u^2}{n} \times \left\{ 1 + \frac{x_{cp.}^2}{Var(x)} \right\}} \quad (1.36)$$

$$s.d.(b) = \sqrt{\frac{\sigma_u^2}{nVar(x)}} \quad (1.37)$$

Подводя итог вышесказанному о точности коэффициентов регрессии, сделаем следующие выводы:

1. Оценка a для параметра α имеет нормальное распределение с

математическим ожиданием α и стандартным отклонением $\sqrt{\frac{\sigma_u^2}{n} \times \left\{ 1 + \frac{x_{cp.}^2}{Var(x)} \right\}}$

, оценка b для параметра β имеет нормальное распределение с

математическим ожиданием β и стандартным отклонением $\sqrt{\frac{\sigma_u^2}{nVar(x)}}$.

2. Для улучшения точности оценок по **МНК** можно увеличивать количество наблюдений в выборке n , увеличивать диапазон наблюдений $Var(x)$ или уменьшить σ_u^2 (например, увеличить точность измерений).

3. Стандартная **ошибка** оценки b считается по формуле:

$$c.o.(a) = \sqrt{\frac{s_u^2}{n} \times \left\{ 1 + \frac{x_{cp.}^2}{Var(x)} \right\}} \quad (1.38)$$

4. Стандартная **ошибка** оценки a :

$$c.o.(b) = \sqrt{\frac{s_u^2}{nVar(x)}} \quad (1.39)$$

Проверка гипотез, относящихся к коэффициентам регрессии

В статистическом исследовании теория и практика взаимно обогащают друг друга и вопрос о том, первично теоретическое построение гипотез или эмпирический анализ не стоит.

Вопрос о проверке гипотез мы будем рассматривать с двух точек зрения. С одной стороны, мы можем предположить, что с начала **формируется гипотеза**, и цель эксперимента в **выяснении её применимости**. Это

приведёт к проверке гипотезы о **уровне значимости**. **Уровень значимости** – это достаточно малая вероятность при которой (в данной определённой задаче) событие можно считать практически невозможным. На практике обычно применяют уровни значимости между 0,1 и 0,05 или между 1 и 5 процентами. С другой стороны сначала **проводим эксперимент**, а затем определяем, какие из теоретических гипотез соответствуют **результатам эксперимента**. Это приводит к **построению доверительных интервалов**. **Доверительным** интервалом называют интервал $(Q^* - \delta)$, $(Q^* + \delta)$, который содержит в себе (покрывает) неизвестный параметр с заданной надёжностью γ .

Надёжностью (доверительной вероятностью) оценки какого нибудь параметра Q по Q^* называют вероятность γ , с которой осуществляется неравенство $|Q - Q^*| < \delta$, где $Q - \delta = Q^*$ характеризует **точность оценки**. Чем меньше δ тем точнее оценка. Обычно надёжность оценки задаётся наперёд, причём в качестве γ берут число близкое единице. Наиболее часто задают надёжность, равную 0,95; 0,99 и 0,999.

Формулирование нулевой гипотезы

Начнём с допущения о том, что *формулирование* гипотезы предшествует эксперименту и что уже имеется некоторая гипотетическая связь или зависимость. Например, можно считать, что темпы общей инфляции в экономике (p^* , в процентах) зависят от темпов инфляции, вызванной ростом заработной платы (w^* , в процентах), и что эта зависимость описывается линейным уравнением:

$$p^* = \alpha + \beta w^* + u, \quad (1.40)$$

где α и β – параметры, а u – случайный член.

Далее можно построить гипотезу, что без учёта случайного члена, общая инфляция равна инфляции вызванной ростом зарплаты. В этих условиях можно сказать, что гипотеза, которую вы собираетесь проверить, считается **нулевой**, и обозначается H_0 состоит в том, что $\beta = 1$ при $\alpha = 0$. Мы также определяем **альтернативную** гипотезу, которая обозначается как H_1 , и представляет собой заключение, даваемое в том случае, если экспериментальная проверка указала на ложность H_0 . В данном случае эта гипотеза состоит в том, что $\beta \neq 1$. Две гипотезы сформулированы с использованием следующих обозначений:

$$H_0 : \beta = 1.$$

$$H_1 : \beta \neq 1.$$

На практике более обычным является построение **нулевой** гипотезы, которая затем проверяется с помощью **альтернативной** гипотезы, которая предполагается верной.

Обобщая сказанное, введем определения.

Нулевая гипотеза (H_0) – это утверждение о том, что неизвестный параметр модели принадлежит заданному множеству A .

Альтернативная гипотеза(H_1) – это утверждение о том, что неизвестный параметр модели принадлежит другому множеству **B**, при этом множества **A** и **B** не пересекаются.

Последующее рассмотрение будет касаться парной регрессии:

$$y = \alpha + \beta x + u \quad (1.41)$$

Оно будет относиться только к коэффициенту **β** , но точно такие же процедуры применимы и к постоянному коэффициенту **α** . Возмём общий случай, в котором утверждается, что **β** равно некоторому конкретному значению скажем **β_0** ; и альтернативная гипотеза состоит в том что **β** не равно этому значению **$H_0 : \beta = \beta_0$** и **$H_1 : \beta \neq \beta_0$** . Мы можем предпринять попытку отклонить или подтвердить нулевую гипотезу в зависимости от того, что нам необходимо в данном случае. Будем предполагать, что все 4 е условия Гаусса-Маркова выполняются.

Если **H_0** верна, то оценки **β** , полученные в ходе регрессивного анализа, будут иметь распределение с матожиданием **β_0** и дисперсией **$\sigma^2(u) / n \text{Var} (x)$** , предположим, что остаточный член **u** имеет нормальное распределение. Если это так, то величина **b** будет также нормально распределена

Тема 4. Множественная линейная регрессия

На любой экономический показатель чаще всего оказывает влияние не один, а несколько факторов. Например, спрос на некоторое благо определяется не только ценой данного блага, но и ценами на замещающие и дополняющие блага, доходом потребителей и многими другими факторами. В этом случае вместо парной регрессии рассматривается множественная регрессия

$$\hat{y} = f(x_1, x_2, \dots, x_p) \quad (1)$$

Множественная регрессия широко используется в решении проблем спроса, доходности акций, при изучении функции издержек производства, в макроэкономических расчетах и в ряде других вопросов экономики. В настоящее время множественная регрессия – один из наиболее распространенных методов в эконометрике. Основной целью множественной регрессии является построение модели с большим числом факторов, а также определение влияния каждого фактора в отдельности и совокупного их воздействия на моделируемый показатель.

Множественный регрессионный анализ является развитием парного регрессионного анализа в случаях, когда зависимая переменная связана более чем с одной независимой переменной. Большая часть анализа является непосредственным расширением парной регрессионной модели, но здесь также появляются и некоторые новые проблемы, из которых следует выделить две. Первая проблема касается исследования влияния конкретной независимой переменной на зависимую переменную, а также разграничения её воздействия и воздействий других независимых переменных. Второй

важной проблемой является спецификация модели, которая состоит в том, что необходимо ответить на вопрос, какие факторы следует включить в регрессию (1), а какие – исключить из неё. В дальнейшем изложение общих вопросов множественного регрессионного анализа будем вести, разграничивая эти проблемы. Поэтому вначале будем полагать, что спецификация модели правильна.

Самой употребляемой и наиболее простой из моделей множественной регрессии является линейная модель множественной регрессии:

$$y = \alpha' + \beta_1' x_1 + \beta_2' x_2 + \dots + \beta_p' x_p + \varepsilon \quad (2)$$

По математическому смыслу коэффициенты β_j' в уравнении (2) равны частным производным результативного признака y по соответствующим факторам:

$$\beta_1' = \frac{\partial y}{\partial x_1}, \beta_2' = \frac{\partial y}{\partial x_2}, \dots, \beta_p' = \frac{\partial y}{\partial x_p}.$$

Параметр α называется свободным членом и определяет значение y в случае, когда все объясняющие переменные равны нулю. Однако, как и в случае парной регрессии, факторы по своему экономическому содержанию часто не могут принимать нулевых значений, и значение свободного члена не имеет экономического смысла. При этом, в отличие от парной регрессии, значение каждого регрессионного коэффициента β_j' равно среднему изменению y при увеличении x_j на одну единицу лишь при условии, что все остальные факторы остались неизменными. Величина ε представляет собой случайную ошибку регрессионной зависимости.

Попутно отметим, что наиболее просто можно определять оценки параметров β_j' , изменяя только один фактор x_j , оставляя при этом значения других факторов неизменными. Тогда задача оценки параметров сводилась бы к последовательности задач парного регрессионного анализа по каждому фактору. Однако такой подход, широко используемый в естественнонаучных исследованиях, (физических, химических, биологических), в экономике является неприемлемым. Экономист, в отличие от экспериментатора – естественника, лишен возможности регулировать отдельные факторы, поскольку не удаётся обеспечить равенство всех прочих условий для оценки влияния одного исследуемого фактора.

Получение оценок параметров $\alpha', \beta_1', \beta_2', \dots, \beta_p'$ уравнения регрессии (2) – одна из важнейших задач множественного регрессионного анализа. Самым распространенным методом решения этой задачи является метод наименьших квадратов (МНК). Его суть состоит в минимизации суммы квадратов отклонений наблюдаемых значений зависимой переменной y от её значений \hat{y} , получаемых по уравнению регрессии. Поскольку параметры $\alpha', \beta_1', \beta_2', \dots, \beta_p'$ являются случайными величинами, определить их

истинные значения по выборке невозможно. Поэтому вместо теоретического уравнения регрессии (2) оценивается так называемое **эмпирическое уравнение регрессии**, которое можно представить в виде:

$$y = a + b_1x_1 + b_2x_2 + \dots + b_px_p + e \quad (3)$$

Здесь a, b_1, b_2, \dots, b_p - оценки теоретических значений $\alpha', \beta_1', \beta_2', \dots, \beta_p'$, или эмпирические коэффициенты регрессии, e – оценка отклонения ε . Тогда расчетное выражение имеет вид:

$$\hat{y} = a + b_1x_1 + b_2x_2 + \dots + b_px_p \quad (4)$$

Пусть имеется n наблюдений объясняющих переменных и соответствующих им значений результативного признака:

$$(x_{i1}, x_{i2}, \dots, x_{ip}, y_i), \quad i = \overline{1, n} \quad (5)$$

Для однозначного определения значений параметров уравнения (4) объем выборки n должен быть не меньше количества параметров, т.е. $n \geq p + 1$. В противном случае значения параметров не могут быть определены однозначно. Если $n = p + 1$, оценки параметров рассчитываются единственным образом без МНК простой подстановкой значений (5) в выражение (4). Получается система $(p+1)$ уравнений с таким же количеством неизвестных, которая решается любым способом, применяемым к системам линейных алгебраических уравнений (СЛАУ). Однако с точки зрения статистического подхода такое решение задачи является ненадежным, поскольку измеренные значения переменных (5) содержат различные виды погрешностей. Поэтому для получения надежных оценок параметров уравнения (4) объем выборки должен значительно превышать количество определяемых по нему параметров. Практически, как было сказано ранее, объем выборки должен превышать количество параметров при x_j в уравнении (4) в 6-7 раз.

Для проведения анализа в рамках линейной модели множественной регрессии необходимо выполнение ряда предпосылок МНК. В основном это те же предпосылки, что и для парной регрессии, однако здесь нужно добавить предположения, специфичные для множественной регрессии:

5⁰. Спецификация модели имеет вид (2).

6⁰. Отсутствие мультиколлинеарности: между объясняющими переменными отсутствует строгая линейная зависимость, что играет важную роль в отборе факторов при решении проблемы спецификации модели.

7⁰. Ошибки $\varepsilon_i, i = \overline{1, n}$ имеют нормальное распределение ($\varepsilon_i \sim N(0, \sigma)$). Выполнимость этого условия нужна для проверки статистических гипотез и построения интервальных оценок.

При выполнении всех этих предпосылок имеет место многомерный аналог теоремы Гаусса – Маркова: оценки a, b_1, b_2, \dots, b_p , полученные по МНК, являются наиболее эффективными (в смысле наименьшей дисперсии) в классе линейных несмещенных оценок.

Тема 5. Фиктивные переменные в регрессионных моделях

При исследовании влияния качественных признаков на объясняемую (зависимую) переменную y в модель множественной линейной регрессии следует вводить фиктивные переменные, принимающие, как правило, два значения: 1, если данный признак присутствует в наблюдении; 0 – при его отсутствии.

Если включаемый в рассмотрение качественный признак имеет не два, а несколько значений, то используют несколько фиктивных переменных, число которых должно быть на единицу меньше числа значений признака. При назначении фиктивных переменных исследуемая совокупность по числу значений качественного признака разбивается на группы. Одну из групп выбирают как эталонную и определяют фиктивные переменные для остальных.

Если качественный признак имеет два значения, то достаточно ввести одну фиктивную переменную. Например, строится модель, характеризующая показатели предприятий двух отраслей промышленности: электроэнергетики и газовой промышленности. Вводится фиктивная переменная, которой присваивается значение 0, если данные относятся к предприятиям электроэнергетики, и значение 1, если данные относятся к предприятиям газовой промышленности.

При трех значениях качественного признака следует вводить две фиктивные переменные. Например, строится модель, характеризующая показатели предприятий трех регионов. Вводится одна фиктивная переменная, которой присваивается значение 0, если данные относятся к предприятиям первого региона, и значение 1, если данные относятся к предприятиям двух других регионов. Второй фиктивной переменной присваивается значение 0, если данные относятся ко второму региону, и значение 1, если данные относятся к первому и третьему регионам.

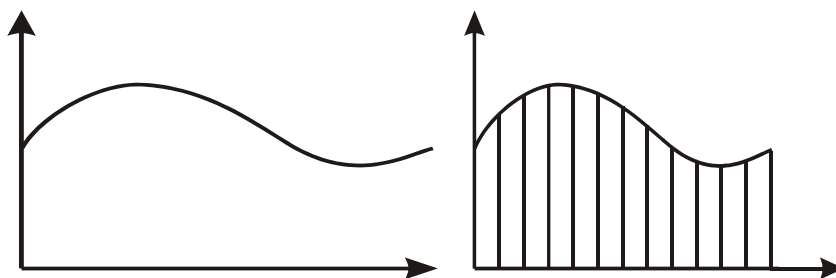
Введение в регрессию фиктивных переменных существенно улучшает качество оценивания.

Тема 6. Динамические модели

Примеры пространственной динамики: спрос в функции цены; объемы продаж в функции рекламного бюджета, объем привлекаемых банком средств в функции нормы затрат на привлекаемые средства и т.д.

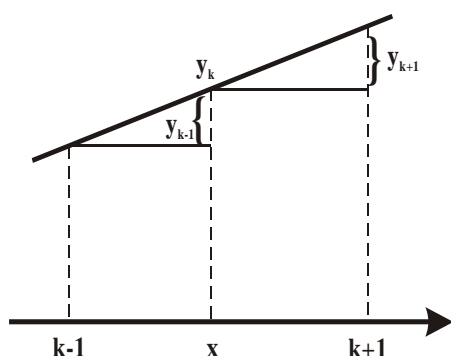
Примеры временной динамики: ежегодные, ежеквартальные, ежемесячные или ежедневные данные по прибыли, остаткам на счетах банка, данные на кв. метр жилья, объемы перевозок авиакомпании, показатели качества добываемой бокситовой руды, данные по ВВП, ВВП, объему экспорта, инфляции и т.д.

Равномерная дискретизация \Rightarrow Решетчатые функции



$$Y_{t-1}, Y_t, Y_{t+1}, Y_{t+2}$$

$$x_{t-1}, x_t, x_{t+1}, x_{t+2}$$



Левая первая разность

$$\Delta Y_k = Y_k - Y_{k-1}$$

Правая первая разность

$$\Delta Y_{k+1} = Y_{k+1} - Y_k$$

Симметричная первая разность

$$\Delta Y_k = Y_{k+1} - Y_{k-1}$$

$$Y' = \lim_{\Delta t \rightarrow 0} \frac{\Delta Y}{\Delta t}$$

$$\Delta^2 Y_k = \Delta Y_k - \Delta Y_{k-1} = Y_k - Y_{k-1} - Y_{k-1} + Y_{k-2} = Y_k - 2Y_{k-1} + Y_{k-2}$$

Динамические модели - используют не только текущие значения переменных, но и некоторые предыдущие по времени значения, а также само время.

$t = k\Delta$ - временные ряды или

$t = kh$ - пространственные ряды динамики $\Delta \leq \frac{1}{2f_{\max}}$.

Тема 7. Системы одновременных уравнений

Коэффициенты структурной модели могут быть оценены разными способами в зависимости от вида системы одновременных уравнений. Наибольшее распространение в литературе получили следующие методы оценивания коэффициентов структурной модели:

- 1) косвенный метод наименьших квадратов;

- 2) двухшаговый метод наименьших квадратов;
- 3) трехшаговый метод наименьших квадратов;
- 4) метод максимального правдоподобия с полной информацией;
- 5) метод максимального правдоподобия при ограниченной информации.

Эконометрические модели обычно включают в систему не только уравнения, отражающие взаимосвязи между отдельными переменными, но и выражения тенденции развития явления и тождества различного рода. Для определения структурных коэффициентов (коэффициенты при эндогенных и экзогенных переменных в правой части уравнения) структурная форма преобразуется в приведенную форму. Тогда, применяя МНК к приведенной форме модели, оценивают коэффициенты приведенной формы модели, а уже затем оценивают значения эндогенных переменных посредством получения значений экзогенных переменных.

Исходная система одновременных уравнений называется идентифицированной, если все её уравнения точно идентифицированы.

Уравнение называется точно идентифицированным, если по оценкам коэффициентов приведённой формы системы одновременных уравнений можно однозначно найти оценки коэффициентов структурной формы системы одновременных уравнений.

Признаком идентифицированности системы одновременных уравнений является равенство между количеством уравнений, определяющих структурные коэффициенты, и количеством этих коэффициентов, т. е. квадратная форма структурной системы уравнений.

Исходная система одновременных уравнений называется сверхидентифицированной, если среди уравнений модели есть хотя бы одно сверхидентифицированное.

Уравнение называется сверхидентифицированным, если по оценкам коэффициентов приведённой формы системы одновременных уравнений можно получить более одного значения для коэффициентов структурной формы системы одновременных уравнений.

Исходная система одновременных уравнений называется неидентифицированной, если среди уравнений системы есть хотя бы одно неидентифицированное.

Уравнение называется неидентифицированным, если по оценкам коэффициентов приведённой формы системы одновременных уравнений невозможно рассчитать оценки коэффициентов структурной формы системы одновременных уравнений.

Параметры рекурсивной системы можно оценивать обычным МНК и нет необходимости исследования уравнения на идентификацию.