Модуль 2 Основы эконометрики

Тема 2.1. Методологические основы курса

Лекция 2.1.1. Предмет эконометрики. Классы эконометрических моделей. Основные этапы эконометрического моделирования

В повседневной жизни, бизнесе, иной профессиональной деятельности, научных исследованиях нам приходится принимать решения в неопределенных, связанных со многими случайностями ситуациях. При этом решения должны приниматься на основе тщательного анализа имеющейся информации, быть обоснованными и доказуемыми. Для решения подобных задач существует достаточно мощный набор методов анализа данных, основанных на аппарате математической статистики. Эконометрика же как наука расположена между экономикой, статистикой и математикой. Предмет эконометрики определяется как **исследование и установление количественных закономерностей и количественных взаимозависимостей в экономической жизни при помощи математических и математико-статистических методов**.

На процессы в экономике оказывает влияние множество факторов, причем некоторые из них являются существенными, а влияние других случайно. Для выявления существенности воздействия необходимо проанализировать большую группу наблюдений, при этом случайные воздействия несущественных факторов гасятся, и обнаруживается общая для всей совокупности закономерность.

Приведем ряд примеров применения методов анализа данных.

1. Предположим, было внедрено важное нововведение — изменена система оплаты труда, освоен выпуск нового вида продукции, введена новая технология и др. Является ли полученный в производстве эффект результатом нововведения или определяется естественной случайностью и уже завтра может быть получен прямо противоположный эффект? Статистические критерии

сравнения двух выборок покажут, случайны или неслучайны различия двух рядов чисел.

2. Предположим, для заключения коммерческой сделки необходимо знать поведение некоторого временного ряда — курса доллара, цен и спроса на продукцию или сырье и др. Для такого временного ряда строят регрессионное уравнение, включая в него набор существенных факторов (проверив существенность этого влияния), затем осуществляют прогноз и указывают его точность.

3. Для того, чтобы в технологическом процессе систематически контролировать его состояние и вовремя вмешаться при отклонениях от нормального режима, предотвратить выпуск некачественной продукции. Для этого используются статистические методы контроля качества (строятся контрольные карты изменения показателей качества с зоной допустимых пределов изменений и др.).

4. Например, необходимо определить надежность клиента, претендующего на кредит в банке. Для этого используются методы классификации объектов по некоторому набору показателей (размер основных фондов, валюта баланса, вид деятельности, объем реализации и др.). Имеющиеся объекты удается собрать в несколько групп (кластеров), и тогда можно будет увидеть, принадлежит ли запрашивающая кредит фирма к группе неплательщиков.

По этим примерам можно видеть, что сфера применения эконометрических моделей обширна: производство, бизнес, финансы, инвестиционная сфера и др. Возможно построение эконометрических моделей разных уровней: отдельного предприятия, отрасли, региона и даже отдельной страны.

К сожалению, статистические закономерности обладают относительной устойчивостью, которая определяется стабильностью условий, при которых она сформировалась. Значительные изменения в условиях повлекут за собой изменение самой статистической закономерности.

Можно выделить три основных класса моделей, которые используются для анализа и прогноза.

Модели временных рядов. К этому классу относятся модели:

— *тренда*:

$$y\,(t) = T\,(t) + \varepsilon_t,$$

где $T(t)$ — временной ряд заданного параметрического вида (например, линейный $T\,(t) = a + bt$),

ε_t — случайная (стохастическая) компонента;

— *сезонности*:

$$y\,(t) = S\,(t) + \varepsilon_t,$$

где $S(t)$ — периодическая (сезонная) компонента,

ε_t — случайная (стохастическая) компонента;

— *тренда и сезонности*:

$$y\,(t) = T\,(t) + S\,(t) + \varepsilon_t \quad \text{(аддитивная)},$$
$$y\,(t) = T\,(t) \cdot S\,(t) + \varepsilon_t \quad \text{(мультипликативная)};$$

где $T(t)$ — временной тренд заданного параметрического вида,

$S\,(t)$ — периодическая (сезонная) компонента,

ε_t — случайная (стохастическая) компонента.

К моделям временных рядов относится множество более сложных моделей, таких как модели адаптивного прогноза, модели авторегрессии, скользящего среднего и др. Общей чертой этих моделей является то, что они объясняют поведение временного ряда, исходя только из его предыдущих значений.

Регрессионные модели. В таких моделях зависимая (объясняемая) переменная y представляется в виде функции

$$f\,(x,\,\beta) = f\,(x_1,\,\ldots,\,x_k,\,\beta_1,\,\ldots,\,\beta_p),$$

где $x_1,\,\ldots,\,x_k$ — независимые (объясняющие) переменные, а $\beta_1,\,\ldots,\,\beta_p$ — параметры.

В зависимости от вида функции $f(x, \beta)$ модели делятся на *линейные* и *нелинейные*. Область применения таких моделей значительно шире, чем моделей временных рядов. Поэтому данная тема является основной в эконометрике.

Системы одновременных уравнений. Эти модели описываются системами уравнений. Системы могут состоять из тождеств и регрессионных уравнений, каждое из которых может, кроме объясняющих переменных, включать в себя также объясняемые переменные из других уравнений системы. Таким образом, мы имеем здесь набор объясняемых переменных, связанных через уравнения системы.

Пример — модель спроса и предложения (см. пример 1, С. 90).

При моделировании экономических процессов встречаются два типа данных: пространственные данные и временные ряды. Примером пространственных данных являются объемы производства, количество работников и др. по разным фирмам в один и тот же момент времени. Примерами временных данных являются ежемесячные данные по средней заработной плате, ежедневный курс доллара и др. Отличительной чертой временных данных является то, что они естественным образом упорядочены во времени. Кроме того, наблюдения в близкие моменты времени часто бывают зависимыми.

Этапы эконометрического моделирования

К основным этапам эконометрического моделирования относятся:

1. изучение объекта,
2. сбор и предварительная обработка информации,
3. построение модели,
4. статистический анализ модели,
5. проверка модели на адекватность,
6. практическое использование модели.

Первый этап включает качественный анализ объектов, изучение взаимосвязей отдельных показателей, определение конечных целей моделирования. Анализ опирается на теоретические представления о процессе функционирования данного объекта. Результатом первого этапа является формирование концепции эконометрической модели.

Основной целью **второго** этапа является предварительная обработка полученных данных, которая заключается в статистическом описании выборки методами математической статистики. Проверяются гипотезы относительно однородности выборок, независимости наблюдений и стационарности исследуемых процессов. Выясняют причины возникновения аномальных наблюдений и возможность их отсечения без нанесения содержательного вреда модели, восстанавливаются пропуски в данных. Осуществляется проверка соответствия распределения результатов измерения закону нормального распределения. Если эта гипотеза неприемлема, то определяют, какому закону подчиняется распределение данных и возможно ли преобразование данного распределения к нормальному.

Задача **третьего** этапа заключается в определении общего вида модельных соотношений. Устанавливается общий вид модельных соотношений, связывающих входные и выходные показатели, формируют структуру модели и ее символическую запись. На данном этапе существенное значение имеет использование корреляционного анализа, который дает возможность установить наличие и тесноту взаимосвязи количественных случайных величин.

На **четвертом** этапе по выборочным данным проводят статистическое оценивание неизвестных параметров модели. Здесь используются процедуры регрессионного анализа и анализа временных рядов, которые позволяют представить зависимости в аналитическом виде. Конечным результатом данного этапа является эконометрическая модель.

Пятый этап заключается в проверке построенной модели на адекватность, которая осуществляется путем сравнения реальных результатов и результатов, полученных с помощью данной модели.

Последний (**шестой**) этап связан с практическим использованием полученной модели.

Для анализа данных применяются также такие методы, которые относятся к разряду специальных — это методы оптимального планирования эксперимента, ковариационный анализ, модели с использованием цепей Маркова и др.

Лекция 2.1.2. Основные этапы предварительной обработки данных

Предварительная обработка результатов измерений необходима для того, чтобы в дальнейшем с наибольшей эффективностью и корректно использовать для построения эмпирических зависимостей статистические методы. К основным этапам предварительной обработки статистических данных относятся:

а) вычисление выборочных характеристик;

б) отсев грубых погрешностей;

в) проверка нормальности распределения;

г) преобразование распределения к нормальному (если требуется).

Рассмотрим, каким образом осуществляются эти процедуры.

а) **Вычисление выборочных характеристик**

Наиболее часто употребляемыми характеристиками случайной величины (и соответствующего распределения вероятностей) служат *моменты* и *квантили*.

Пусть имеется ограниченный ряд наблюдений $x_1, \ldots, x_n$ случайной величины ξ. *Среднее значение наблюдаемого признака* можно определить по формуле

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Далее вычисляется *дисперсия* или *второй центральный момент эмпирического распределения*

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

причем $S^2 = m_2$.

В случае одномерного эмпирического распределения *произвольным моментом порядка* k называется сумма k-ых степеней отклонений результатов наблюдений от произвольного числа с, деленная на объем выборки n:

$$m_k = \frac{1}{n}\sum_{i=1}^{n}(x_i - c)^k,$$

где k может принимать любые значения натурального ряда чисел. Если $c = 0$, то момент называют *начальным*. *Начальным моментом первого порядка* является выборочное среднее $\overline{x}$. При $c = \overline{x}$ момент называют центральным. *Первый центральный момент*

$$m_1 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x}) = 0.$$

Второй центральный момент

$$m_1 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2.$$

представляет собой дисперсию S^2 эмпирического распределения. Однако в статистике чаще в качестве выборочной дисперсии используют

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2,$$

поскольку математическое ожидание величины s^2 равно дисперсии. Из других моментов чаще всего используют центральные моменты третьего и четвертого порядка.

Если необходимо, чтобы показатель разброса случайной величины выражался в тех же единицах, что и значение этой случайной величины, то используют величину *выборочного среднеквадратического отклонения* $S = \sqrt{S^2}$.

Выборочное значение коэффициента вариации v, являющееся мерой относительной изменчивости наблюдаемой случайной величины, вычисляют по формуле

$$\mathrm{v} = \frac{S}{\overline{x}}.$$

Коэффициент вариации может быть выражен и в процентах:

$$\mathrm{v} = \frac{S}{\overline{x}} \cdot 100\%.$$

Коэффициент вариации имеет смысл абсолютной меры рассеяния, который применяется для сравнения меры рассеяния в разных числовых совокупностях, поскольку остальные рассмотренные меры рассеяния измеряются в тех же единицах, что и сами признаки.

Выборочной квантилью называется решение уравнения

$$F_n(x) = p,$$

в частности, *выборочная медиана* есть решение уравнения

$$F_n(x) = 0,5.$$

Содержательно *медиана* — срединное (центральное) значение в упорядоченном ряду значений признака, или величина, обладающая тем свойством, что число единиц совокупности с большими значениями признака и число единиц с меньшими значениями его одинаково. Применительно к кривой распределения медиану можно определить как такое значение признака на оси абсцисс, что ордината, проходящая через него, делит площадь кривой на две равные части. Однако это определение не всегда однозначно. Если имеется нечетное число различных наблюдений, например $2n+1$, то $n+1$-е значение по порядку нарастания значения будет единственным, отвечающим понятию медианы. Если же число наблюдений $2n$, то любое число между n-м и $n+1$-м значением удовлетворяет нашему требованию. В таких случаях за медиану принимают среднюю арифметическую из n-го и $n+1$-го значения.

Мода — значение признака, которое соответствует максимальной точке теоретической кривой, наилучше подобранной к действительному распределению. Она представляет наиболее часто встречающееся или типичное значение.

В симметричном распределении среднее арифметическое, мода и медиана равны. Для умеренно асимметричных распределений существует соотношение

$$X_{\text{мода}} = \overline{X} - 3\,(\overline{X} - X_{\text{медиана}}).$$

Все виды средних характеризуют уровень числовой совокупности, т. е. то значение признака, вокруг которого концентрируются прочие значения. К характеристикам меры рассеяния (амплитуды рассеяния) относятся уже перечисленные дисперсия, среднеквадратическое отклонение и коэффициент вариации. Сюда также относится простейшая мера рассеяния — вариационный размах

$$R = X_{max} - X_{min}.$$

б) **Отсев грубых погрешностей.**

Для практического использования целесообразно использовать простейшие методы отсева грубых погрешностей. Например, для выборок небольшого объема ($n \leqslant 25$) можно воспользоваться методом вычисления максимального относительного отклонения:

$$\frac{|x_i - \overline{x}|}{S} \leqslant t_{\alpha,n},$$

где x_i — крайний (наибольший или наименьший) элемент выборки, по которой подсчитывались $\overline{x}$ и S;

$t_{\alpha,n}$ — табличное значение статистики t, вычисленной при доверительной вероятности $p = 1 - \alpha$.

Таким образом, для выделения аномального значения вычисляют

$$t = \frac{|x_i - \overline{x}|}{S},$$

которое затем сравнивают с табличным значением $t_{\alpha,n}$: если

$$t \leqslant t_{\alpha,n},$$

то наблюдение не отсеивают, в противном случае наблюдение отсеивают, — после чего характеристики эмпирического распределения пересчитывают по данным сокращенной выборки.

Для больших выборок отсев грубых погрешностей проводят с использованием таблиц распределения Стьюдента[1].

[1]Стьюдент (англ. *Student*) — псевдоним английского математика и статистика Уи́льяма Си́ли Го́ссета (англ. *William Sealy Gosset*; 1876—1937).

в) **Проверка распределения на нормальность.**

Если большое число значений количественного признака зарегистрировано в той последовательности, в какой они встретились в действительности, то трудно охватить подлинный смысл наблюденного. Для того, чтобы выявить характерные черты явления, нужно сжато выразить данные, для чего и служат *группировка и анализ распределения численностей*.

Разбиение на классы проводится либо по правилу Штюргеса, когда число классов k определяется как

$$k = 1 + 3,32 \lg n,$$

либо число классов определяется произвольно, причем тогда при выборе интервала руководствуются двумя условиями:

1) возможностью без большой ошибки приравнять все значения признака, отнесенные к какой-либо группе, срединному значению интервала;
2) для удобства и краткости делать интервал достаточно большим.

Поскольку эти два условия противоречивы, то в каждом случае интервал выбирается в зависимости от количества наблюдений, но не более 25. Интервал, выбранный для группировки, называется *групповым интервалом*, а численность в пределах отдельного интервала — *численностью группы*. После группировки данных их можно представить в виде полигона (многоугольника, стороны которого являются отрезками, соединяющими центры интервалов на вертикальных отрезках) численностей или гистограммы (столбчатой диаграммы).

Графическое представление позволяет примерно представить характер распределения числовых данных. Поскольку для целей эконометрического моделирования желательно, чтобы это распределение приближенно соответствовало нормальному закону. К преимуществам нормального распределения относят следующие:

— нормальное распределение полностью определяется величинами μ и σ, причем математическое ожидание определяет положение кривой относительно оси абсцисс, а среднеквадратическое отклонение определяет форму кривой (чем больше σ, тем кривая становится более пологой, основание более широким);

— кривая нормального распределения симметрична относительно среднего значения;

— очень большие и очень малые значения переменной маловероятны;

— примерно 2/3 всех наблюдений лежит в площади, отсекаемой перпендикулярами к оси $(\mu \pm \sigma)$. Для нормального распределения мода, среднее и медиана совпадают.

Некоторое представление о близости эмпирического распределения к нормальному может дать анализ показателей асимметрии и эксцесса.

Показатель асимметрии определяется по формуле

$$g_1 = \frac{m_3}{m_2^{3/2}}.$$

Для симметричных распределений $m_3 = 0$ и $g_1 = 0$.

Для нормального распределения

$$\frac{m_4}{m_2^2} = 3.$$

Для удобства сравнения эмпирического распределения и нормального в качестве *показателя эксцесса* принимают величину

$$g_2 = \frac{m_4}{m_2^2} - 3.$$

г) **Преобразование распределения к нормальному.**

Если выяснено, что гипотеза нормальности распределения не может быть принята, то возможно преобразование исходных данных таким образом, что их распределение будет подчиняться нормальному закону. Причем, после получения окончательного результата надо выполнить обратное преобразование.

Для распределений, имеющих крутую правую ветвь гистограммы и пологую левую, выполняются преобразования матрицы исходных данных по формулам:

$$x' = \lg(x \pm a) \cdot 10^b,$$
$$x' = \frac{1}{x},$$
$$x' = \frac{1}{\sqrt{x}}.$$

Для распределений, смещенных влево, матрицу исходных данных преобразуют по формуле $x' = x^a$ (при $a = 1,5;\ 2$).