ЛЕКЦИЯ № 1. Понятие эконометрики и эконометрических моделей

Эконометрика — это наука, которая на базе статистических данных дает количественную характеристику взаимозависимым экономическим явлениям и процессам.

Слово «эконометрика» произошло от двух слов: «экономика» и «метрика» (от греч. «метрон» — «правило определения расстояния между двумя точками в пространстве», «метрия» — «измерение»). Эконометрика — это наука об экономических измерениях. Зарождение эконометрики является следствием междисциплинарного подхода к изучению экономики. Эконометрика представляет собой сочетание трех наук:

1) экономической теории;
2) математической и экономической статистики;
3) математики.

На современном этапе развития науки неотъемлемым фактором развития эконометрики является развитие компьютерных технологий и специальных пакетов прикладных программ.

Основным предметом исследования эконометрики являются массовые экономические явления и процессы. Предметы эконометрики и статистики очень схожи, так как статистика имеет дело с массовыми социально-экономическими явлениями.

Эконометрика ставит своей целью количественно охарактеризовать те экономические закономерности, которые экономическая теория выявляет и определяет лишь в общем.

Анализ экономических процессов и явлений в эконометрике осуществляется с помощью математических моделей, построенных на эмпирических данных.

Практически все эконометрические методы и приемы изучения экономических закономерностей позаимствованы из математической статистики. Специфика применения методов математической статистики в эконометрике заключается в том, что практически все экономические показатели являются величинами случайными, а не результатами контролируемого эксперимента.

Поэтому существуют определенные усовершенствования и дополнения методов, которые в математической статистике не используются.

Зачастую экономические данные содержат ошибки измерения. В эконометрике разрабатываются специальные методы анализа, позволяющие устранить или снизить влияние этих ошибок на результаты экспериментов.

Таким образом, эконометрика через математические и статистические методы анализирует экономические закономерности, доказанные экономической теорией.

С помощью эконометрики решается очень широкий круг задач. Их можно **классифицировать по трем признакам**:

1) по конечным прикладным целям:
   а) прогноз социально-экономических показателей, определяющих состояние и развитие изучаемой системы;
   б) моделирование возможных вариантов социально-экономического развития системы для определения тех параметров, которые оказывают наиболее мощное влияние на состояние системы в целом;
2) по уровню иерархии:
   а) задачи, решаемые на макроуровне (страна в целом);
   б) задачи, решаемые на мезоуровне (уровень отраслей, регионов);
   в) задачи, решаемые на микроуровне (уровень фирмы, семьи, предприятия);
3) по области решения проблем изучаемой экономической системы:
   а) рынок;
   б) инвестиционная, социальная, финансовая политика;
   в) ценообразование;
   г) распределительные отношения;
   д) спрос и потребление;
   е) отдельно выделенный комплекс проблем.

1. Основные виды эконометрических моделей

Выделяют три основных класса эконометрических моделей.

1. Модель временных рядов.

Модель представляет собой зависимость результативного признака от переменной времени или переменных, относящихся к другим моментам времени.

К моделям временных рядов, в которых результативный признак зависит от времени, относятся:

1) модель тренда (модель зависимости результативного признака от трендовой компоненты);

2) модель сезонности (модель зависимости результативного признака от сезонной компоненты);

3) модель тренда и сезонности.

К моделям временных рядов, в которых результативный признак зависит от переменных, датированных другими моментами времени, относятся:

1) модели с распределенным лагом, которые объясняют вариацию результативного признака в зависимости от предыдущих значений факторных переменных;

2) модели авторегрессии, которые объясняют вариацию результативного признака в зависимости от предыдущих значений результативных переменных;

3) модели ожидания, объясняющие вариацию результативного признака в зависимости от будущих значений факторных или результативных переменных.

Модели временных рядов делятся на модели, построенные по стационарным и нестационарным временным рядам.

Стационарные временные ряды характеризуются постоянными во времени средней, дисперсией и автокорреляцией, т. е. данный временной ряд не содержит трендового и сезонного компонента.

Если временной ряд не отвечает перечисленным условиям, то он является нестационарным (т. е. содержит трендовую и сезонную компоненты).

2. **Регрессионные модели с одним уравнением.**

В подобных моделях зависимая или результативная переменная, обозначаемая обычно , представляется в виде функции факторных или независимых признаков $x_1 ... x_n$:

$$y = f(x, \beta) = f(x_1, \ \ldots, \ x_n, \ \beta_1, \ \ldots, \ \beta_k),$$

где $\beta_1, ..., \beta_к$ — параметры регрессионного уравнения.

Регрессионные модели делятся на парные (с одним факторным признаком) и множественные регрессии.

В зависимости от вида функции $f(x, \beta)$ модели делятся на линейные и нелинейные регрессии.

3. **Системы одновременных уравнений.**

Данные модели описываются системами взаимозависимых регрессионных уравнений. Системы могут состоять из тождеств и регрессионных уравнений, каждое из которых может включать в себя не только факторные переменные, но и результативные переменные из других уравнений системы.

Для тождеств характерно то, что их вид и значения параметров известны.

Регрессионные уравнения, из которых состоит система, называются поведенческими уравнениями. В поведенческих уравнениях значения параметров являются неизвестными и подлежат оцениванию.

Примером системы одновременных уравнений может служить модель спроса и предложения, включающая три уравнения:

$Q^S_t = a_0 + a_1 \times P_t + a_2 \times P_{t-1}$ — уравнение предложения;

$Q^d_t = b_0 + b_1 \times P_t + b_2 \times I_t$ — уравнение спроса;

$Q^S_t = Q^d_t$ — тождество равновесия,

где Q^S_t — предложение товара в момент времени *t*;

Q^d_t — спрос на товар в момент времени *t*;

P_t — цена товара в момент времени *t*;

P_{t-1} — цена товара в предшествующий момент времени *t*;

I_t — доход потребителей в момент времени *t*.

2. Эконометрическое моделирование

Можно выделить несколько этапов эконометрического моделирования.

1. **Постановочный.** На данном этапе определяются конечные цели и задачи исследования и набор участвующих в модели факторных и результативных экономических переменных.

Можно выделить следующие цели эконометрического исследования:

1) анализ изучаемого экономического процесса (явления, объекта);

2) прогноз экономических показателей, характеризующих изучаемый процесс;

3) моделирование поведения процесса при различных значениях независимых (факторных) переменных;

4) выработка управленческих решений.

Включение в эконометрическую модель той или иной переменной должно быть теоретически обосновано. Число переменных не должно быть слишком большим. Факторные переменные не должны быть связаны функциональной или тесной корреляционной связью, присутствие в модели условия мультиколлинеарности может привести к негативным последствиям всего процесса моделирования.

2. **Априорный.** На этом этапе проводятся теоретический анализ сущности изучаемого процесса, а также формирование и формализация известной до моделирования (априорной) информации.

3. **Параметризация.** Осуществляется выбор общего вида модели и выявление состава и формы входящих в нее связей, т. е. происходит непосредственно моделирование.

Основная задача этапа моделирования заключается в выборе наиболее оптимального вида функции зависимости результативной переменной от факторных признаков. Если возникает возможность выбора между нелинейной и линейной формой зависимости, то предпочтение всегда отдается линейной форме как наиболее простой и надежной.

Помимо этого, на этапе моделирования решается задача спецификации модели путем:

1) аппроксимации математической формой выявленных связей и соотношений между переменными;

2) определения зависимых и независимых переменных;

3) формулировки исходных предпосылок и ограничений модели.

Успех эконометрического моделирования во многом зависит от правильного решения проблемы спецификации модели.

4. **Информационный.** Происходит сбор необходимой статистической базы данных, т. е. эмпирических (наблюдаемых) значений экономических переменных, анализ качества собранной информации.

5. **Идентификация модели.** На данном этапе осуществляются статистический анализ модели и оценка ее параметров.

6. **Оценка качества модели.** Проверяются достоверность и адекватность модели, т. е. определяется, насколько успешно решены

задачи спецификации и идентификации модели, какова точность расчетов, полученных на ее основе. Построенная модель должна быть адекватна реальному экономическому процессу. Если качество модели оказалось неудовлетворительным, то вновь возвращаются ко второму этапу моделирования.

7. **Интерпретация результатов моделирования.** Среди наиболее известных эконометрических моделей можно выделить:

1) модели потребительского и сберегательного потребления;
2) модели взаимосвязи риска и доходности ценных бумаг;
3) модели предложения труда;
4) макроэкономические модели (модель роста);
5) модели инвестиций;
6) маркетинговые модели;
7) модели валютных курсов и валютных кризисов и др.

3. Классификация видов эконометрических переменных и типов данных

В эконометрических исследованиях, как правило, используются **два типа выборочных данных**:

1) пространственные данные (cross-sectional data);
2) временные данные (time-series data).

Под пространственными данными понимается совокупность экономической информации, относящейся к разным объектам, полученной за один и тот же период или момент времени. Пространственные данные представляют собой выборочную совокупность из некоторой генеральной совокупности. В качестве примера пространственных данных можно привести совокупность различной информации по какому-либо предприятию (численность работников, объем производства, размер основных фондов), об объемах потребления продукции определенного вида и т. д.

Под временными данными понимается совокупность экономической информации, характеризующей один и тот же объект, но за разные периоды времени. По аналогии с пространственной выборкой отдельно взятый временной ряд можно считать выборкой из бесконечного ряда значений показателей во времени.

В качестве примера временных данных можно привести данные о динамике индекса потребительских цен, ежедневные обменные курсы валют. Временная информация естественным образом упорядочена во времени в отличие от пространственных данных.

Существуют определенные отличия временного ряда от пространственной выборки:

1) элементы динамического ряда не являются статистически независимыми, в отличие от элементов случайной пространственной выборки, т. е. они подвержены явлению автокорреляции (зависимости между прошлыми и текущими наблюдениями временного ряда);

2) элементы динамического ряда не являются одинаково распределенными величинами.

Совокупность экономической информации, которая характеризует изучаемый процесс или объект, представляет собой набор признаков. Данные признаки связаны между собой и в эконометрической модели могут выступать в одной из двух ролей;

3) в роли результативного или зависимого признака, который в эконометрическом моделировании называется объясняемой переменной;

4) в роли факторного или независимого признака, который называется объясняющей переменной.

Экономические переменные, участвующие в любой эконометрической модели, делятся на четыре вида:

1) экзогенные (независимые) — переменные, значения которых задаются извне. В определенной степени данные переменные являются управляемыми (x);

2) эндогенные (зависимые) — переменные, значения которых определяются внутри модели, или взаимозависимые (y);

3) лаговые — экзогенные или эндогенные переменные в эконометрической модели, относящиеся к предыдущим моментам времени и находящиеся в уравнении с переменными, относящимися к текущему моменту времени. Например, x_{i-1} — лаговая экзогенная переменная, y_{i-1} — лаговая эндогенная переменная;

4) предопределенные (объясняющие переменные) — лаговые (x_{i-1}) и текущие (x) экзогенные переменные, а также лаговые эндогенные переменные (y_{i-1}).

Любая эконометрическая модель предназначена для объяснения значений одной или нескольких текущих эндогенных переменных в зависимости от значений предопределенных переменных.

ЛЕКЦИЯ № 2. Общая и нормальная линейная модели парной регрессии

1. Общая модель парной регрессии

После того как в ходе экспериментов было доказано наличие взаимосвязи между изучаемыми переменными, встает задача определения точного вида выявленной зависимости с помощью регрессионного анализа.

Регрессионный анализ заключается в определении аналитического выражения связи (в определении функции), в котором изменение одной величины (результативного признака) обусловлено влиянием независимой величины (факторного признака). Количественно оценить данную взаимосвязь можно с помощью построения уравнения регрессии или регрессионной функции.

Базисной регрессионной моделью является модель парной (однофакторной) регрессии. Данная регрессионная функция называется полиномом первой степени и используется для описания равномерно развивающихся во времени процессов. Общий вид парного уравнения регрессии зависимости y от x:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \qquad (1)$$

где y_i — зависимые переменные, $i = \overline{1.n}$;

x_i — независимые переменные;

β_0, β_1 — параметры уравнения регрессии, подлежащие оцениванию;

ε_i — случайная ошибка модели регрессии, появление которой может быть обусловлено **следующими объективными предпосылками**:

1) нерепрезентативностью выборки. В модель парной регрессии включается одни фактор, неспособный полностью объяснить вариацию результативного признака, который может быть подвержен влиянию множества других факторов в гораздо большей степени;

2) вероятностью того, что переменные, участвующие в модели, могут быть измерены с ошибкой.

Аналитическая форма зависимости между изучаемой парой признаков (регрессионная функция) определяется **с помощью следующих методов**:

1) на основе визуальной оценки характера связи. На линейном графике по оси абсцисс откладываются значения факторного (независимого) признака x, по оси ординат — значения результативного признака y. На пересечении соответствующих значений отмечаются точки. Полученный точечный график в указанной системе координат называется корреляционным полем. При соединении полученных точек получается **эмпирическая линия**, по виду которой можно судить не только о наличии, но и о форме зависимости между изучаемыми переменными;

2) на основе теоретического и логического анализа природы изучаемых явлений, их социально-экономической сущности.

Параметр β_1 уравнения парной регрессии называется коэффициентом регрессии. Его величина показывает, на сколько в среднем изменится результативный признак y при изменении факторного признака x на единицу своего измерения. Знак параметра β_1 в уравнении парной регрессии указывает на направление связи. Если, $\beta_1 > 0$, то связь между изучаемыми показателями прямая, т. е. с увеличением факторного признака x увеличивается и результативный признак, и наоборот. Если $\beta_1 < 0$, то связь между изучаемыми показателями обратная, т. е. с увеличением фактора x результат уменьшается, и наоборот.

Значение параметра β_0 в уравнении парной регрессии трактуется как среднее значение результативного признака y при условии, что факторный признак x равен нулю. Такая трактовка параметра β_0 возможна только в том случае, если значение $x = 0$ имеет смысл.

2. Нормальная линейная модель парной регрессии

Нормальная, или классическая, линейная модель парной регрессии (регрессии с одной переменной) строится исходя **из следующих предположений**:

1) факторный признак x_i является неслучайной или детерми-

нированной величиной, не зависящей от распределения случайной ошибки уравнения регрессии ε_i;

2) математическое ожидание случайной ошибки уравнения регрессии равно нулю во всех наблюдениях:

$$\mathrm{E}(\varepsilon_i) = 0,$$

где $i = \overline{1,n}$;

3) дисперсия случайной ошибки уравнения регрессии является постоянной для всех наблюдений:

$$D(\varepsilon_i) = \mathrm{E}(\varepsilon_i^2) = G^2 = \mathrm{const};$$

4) случайные ошибки уравнения регрессии не коррелированы между собой, т. е. ковариация случайных ошибок любых двух разных наблюдений равна нулю:

$$Cov(\varepsilon_i, \varepsilon_j) = \mathrm{E}(\varepsilon_i \varepsilon_j) = 0,$$

где $i \neq j$.

Это предположение верно в том случае, если изучаемые данные не являются временными рядами;

5) основываясь на 3 и 4-м предположениях, добавляется условие о том, что случайная ошибка уравнения регрессии является случайной величиной, подчиняющейся нормальному закону распределения с нулевым математическим ожиданием и дисперсией

$$G^2 / \varepsilon_i \sim N(0, G^2).$$

Исходя из указанных предпосылок **нормальную линейную модель парной регрессии можно записать в следующем виде**:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \qquad (1)$$

где y_i — значения зависимой переменной, $i = \overline{1,n}$;

x_i — значения независимой переменной;

β_0, β_1 — коэффициенты уравнения регрессии, подлежащие оценке;

ε_i — случайная ошибка уравнения регрессии.

Матричная **форма нормальной линейной модели парной регрессии**:

$$Y = \beta X + \varepsilon, \tag{2}$$

где

$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$ — вектор значений зависимой переменной размер ности $n \times 1$;

$X = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \dots & \dots \\ 1 & x_n \end{pmatrix}$ — вектор значений независимой переменной размерности $n \times 2$. Первый столбец является единичным, так как в уравнении регрессии параметр β_0 умножается на 1;

$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$ — вектор коэффициентов уравнения регрессии размерности 2×1;

$\varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$ — вектор случайных ошибок уравнения регрессии размерности $n \times 1$.

Предположения о модели, записанные в матричном виде:

1) факторный признак x является неслучайной или детерминированной величиной, не зависящей от распределения случайной ошибки уравнения регрессии ε;

2) математическое ожидание случайной ошибки уравнения регрессии равно нулю во всех наблюдениях:

$$\mathrm{E}(\varepsilon) = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} = 0;$$

3) предположения о том, что дисперсия случайной ошибки уравнения регрессии является постоянной для всех наблюдений и ковариация случайных ошибок любых двух разных наб-

людений равна нулю, можно записать с помощью ковариационной матрицы случайных ошибок нормальной линейной модели парной регрессии:

$$\Sigma_{\varepsilon} = \begin{pmatrix} G^2 & 0 & \cdots & 0 \\ 0 & G^2 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & G^2 \end{pmatrix}. \qquad (3)$$

Данную **ковариационную матрицу можно преобразовать следующим образом**:

$$\Sigma_{\varepsilon} = G^2 \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & 1 \end{pmatrix} = G^2 \mathrm{I} n,$$

где G^2 — дисперсия случайной ошибки уравнения регрессии ε;

I_n — единичная матрица размерности $n \times n$.

Ковариация — это показатель тесноты связи между изучаемыми переменными, которая вычисляется по формуле:

$$Cov(x,y) = \overline{x\ y} - \bar{x}\ \bar{y},$$

где $\overline{xy}$ — среднее арифметическое значение произведения факторного и результативного признаков:

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

На диагонали ковариационной матрицы случайных ошибок нормальной линейной модели парной регрессии располагается дисперсия случайных ошибок, так как ковариация переменной с самой собой равна дисперсии переменной. Таким образом:

$$Cov(\varepsilon,\varepsilon) = G^2(\varepsilon);$$

4) случайная ошибка уравнения регрессии имеет нормальный закон распределения:

$$\varepsilon \sim N(0, G^2 \mathrm{I} n).$$

ЛЕКЦИЯ № 3. Методы оценивания и нахождения параметров уравнения регрессии. Классический метод наименьших квадратов (МНК)

На первом этапе проведения регрессионного анализа была выбрана функция $f(x)$, отражающая зависимость результативного признака y от факторного признака x. Необходимо оценить неизвестные параметры модели. В качестве методов оценки неизвестных параметров **уравнения регрессии** $\beta_0, ..., \beta_n$ **могут выступать**:

1) сумма квадратов отклонений наблюдаемых значений результативного признака y от теоретических значений $\tilde{y}$, рассчитанных на основании регрессионной функции, $f(x)$:

$$F = \sum_{i=1}^{n} (y_i - f(x_i, \beta))^2 \quad \text{или} \quad F = \sum_{i=1}^{n} (y_i - \tilde{y}_i)^2.$$

Этот метод оценивания неизвестных параметров уравнения регрессии называется **методом наименьших квадратов (МНК)**. Термин МНК был впервые использован в работе ***А. М. Лежандра*** в 1805 г. Можно выделить следующие достоинства метода:

а) расчеты сводятся к механической процедуре нахождения коэффициентов;

б) доступность полученных математических выводов.

Основным недостатком МНК является чувствительность оценок к резким выбросам, которые встречаются в исходных данных.

2) сумма модулей отклонений наблюдаемых значений результативного признака y от теоретических значений $\tilde{y}$ (рассчитанных на основании регрессионной функции) $f(x)$:

$$F = \sum_{i=1}^{n} \left| y_i - f(x_i, \beta) \right| \quad \text{или} \quad F = \sum_{i=1}^{n} \left| y_i - \tilde{y}_i \right|.$$

Основным достоинством метода является нечувствительность оценок к резким выбросам (в отличие от МНК). **Среди недостатков можно выделить следующие**:

а) сложности в ходе вычислительной процедуры;

б) зачастую большим отклонениям в исходных данных следует придавать больший вес для уравновешивания их в общей сумме наблюдений;

в) неодинаковым значениям оцениваемых параметров $\beta_0, ..., \beta_n$ могут соответствовать одинаковые суммы модулей отклонений;

$$F = \sum_{i=1}^{n} g(y_i - f(x_i, \beta)) \text{ или } F = \sum_{i=1}^{n} g(y_i - \tilde{y}_i),$$

где g — мера или вес, с которой отклонение $(y_i - f(x_i,\beta))$ входит в функционал F. Примером меры g является функция Хубера, которая при малых значениях переменной x является квадратичной, а при больших значениях x — линейной:

$$g(x) = \begin{cases} x^2, & |x| < c \\ 2cx - c^2, & x \geq c \\ -2cx - c^2, & x \leq -c, \end{cases}$$

где c — ограничения функции.

Третий метод оценки неизвестных параметров уравнения регрессии $\beta_0, ..., \beta_n$ — объединие достоинства предыдущих двух методов. Оценки неизвестных параметров, найденные с его помощью, являются менее чувствительными к случайным выбросам в исходных данных, чем оценки, полученные МНК. Этот метод применяют, когда выборка сильно «засорена».

Для нахождения оптимальных значений неизвестных параметров $\beta_0, ..., \beta_n$ **необходимо минимизировать функционал F по данным параметрам**:

1) $F = \sum_{i=1}^{n} (y_i - f(x_i, \beta))^2 \to \min$ — процесс минимизации функционала F состоит в отыскании таких параметров $\beta_0, ..., \beta_n$, при которых сумма квадратов отклонений наблюдаемых значений результативного признака y от теоретических значений $\tilde{y}$ была бы минимальной;

2) $F = \sum_{i=1}^{n} |y_i - f(x_i, \beta)| \to \min$ — процесс минимизации функционала F состоит в отыскании таких параметров $\beta_0, ..., \beta_n$, при которых сумма модулей отклонений наблюдаемых значений результативного признака y от теоретических значений $\tilde{y}$ была бы минимальной;

3) $F = \sum_{i=1}^{n} g(y_i - f(x_i, \beta)) \to \min$ — процесс минимизации функционала F состоит

в отыскании таких параметров β_0, ..., β_n, при которых сумма отклонений наблюдаемых значений результативного признака y от теоретических значений $\tilde{y}$ с учетом заданных весов g была бы минимальной.

Наиболее распространенным методом оценивания параметров уравнения регрессии является метод наименьших квадратов.

1. Классический метод наименьших квадратов для модели парной регрессии

Рассмотрим применение метода наименьших квадратов для нахождения неизвестных параметров уравнения регрессии на примере модели линейной парной регрессии.

Пусть подобрана эмпирическая линия, по виду которой можно судить о том, что связь между независимой переменной и зависимой переменной **линейна и описывается равенством**:

$$y_i = \beta_0 + \beta_1 x_i. \qquad (1)$$

Необходимо найти такие значения параметров $\tilde{\beta}_0$ и $\tilde{\beta}_1$, которые бы доставляли минимум функции (1), т. е. минимизировали бы сумму квадратов отклонений наблюдаемых значений результативного признака y от теоретических значений $\tilde{y}$ (значений, рассчитанных на основании уравнения регрессии):

$$F = \sum_{i=1}^{n} (y_i - \tilde{y}_i)^2 = \sum_{i=1}^{n} (y_i - \tilde{\beta}_0 - \tilde{\beta}_1 x_i) \to \min. \qquad (2)$$

При минимизации функции (1) неизвестными являются значения коэффициентов регрессии β_0 и β_1. Значения зависимой и независимой переменных известны из наблюдений.

Для того чтобы найти минимум функции двух переменных, нужно вычислить частные производные этой функции по каждому из оцениваемых параметров и приравнять их к нулю. **В результате получаем стационарную систему уравнений для функции (2):**

$$\begin{cases} \dfrac{\partial F}{\partial \beta_0} = -2\sum\limits_{i=1}^{n} (y_i - \tilde{\beta}_0 - \tilde{\beta}_1) = 0, \\ \dfrac{\partial F}{\partial \beta_1} = -2\sum\limits_{i=1}^{n} (y_i - \tilde{\beta}_0 - \tilde{\beta}_1) \times x_i = 0. \end{cases}$$

Если разделить обе части каждого уравнения системы на (−2), раскрыть скобки и привести подобные члены, **то получим систему**:

$$\begin{cases} \tilde{\beta}_1 \sum_{i=1}^{n} x_i^2 + \tilde{\beta}_0 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i \times y_i, \\ \tilde{\beta}_1 \sum_{i=1}^{n} x_i + \tilde{\beta}_0 \times n = \sum_{i=1}^{n} y_i. \end{cases}$$

Это система нормальных уравнений относительно коэффициентов β_0 и β_1 для зависимости $y_i = \beta_0 + \beta_1 x_i$.

Решением системы нормальных уравнений являются оценки неизвестных параметров уравнения регрессии β_0 и β_1:

$$\tilde{\beta}_1 = \frac{n\sum_{i=1}^{n} x_i \times y_i - \sum_{i=1}^{n} x_i \times \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} = \frac{\overline{xy} - \overline{x}\,\overline{y}}{\overline{x^2} - \overline{x}^2} = \frac{Cov(x,y)}{G^2(x)},$$

$$\tilde{\beta}_0 = \overline{y} - \tilde{\beta}_1 \times \overline{x},$$

где $\overline{y}$ — среднее значение зависимого признака;

$\overline{x}$ — среднее значение независимого признака;

$\overline{xy}$ — среднее арифметическое значение произведения зависимого и независимого признаков;

$G^2(x)$ — дисперсия независимого признака;

$Cov(x,y)$ — ковариация между зависимым и независимым признаками.

Рассмотрим применение МНК на конкретном примере.

Имеются данные о цене на нефть x (долларов за баррель) и индексе акций нефтяной компании y (в процентных пунктах). Требуется найти эмпирическую формулу, отражающую связь между ценой на нефть и индексом акций нефтяной компании исходя из предположения, что связь между указанными переменными линейна и описывается функцией вида $y_i = \beta_0 + \beta_1 x_i$. Зависимой переменной ($y$) в данной регрессионной модели будет являться индекс акций нефтяной компании, а независимой (x) — цена на нефть.

Для нахождения коэффициентов β_0 и β_1 построим вспомогательную таблицу 1.

Таблица 1

Таблица для нахождения коэффициентов β_0 и β_1

№ Наблюдения	Цена на нефть — x, ден. Ед.	Индекс нефтяной компании — процентные пункты	$x_i \times y_i$	x_i^2
1	17,28	537	9279,36	298,5984
2	17,05	534	9104,70	290,7025
3	18,30	550	10 065,00	334,8900
4	18,80	555	10 434,00	353,4400
5	19,20	560	10 752,00	368,6400
6	18,50	552	10 212,00	342,2500
Сумма по столбцу	110,13	3288	59 847,06	1988,52

Запишем систему нормальных уравнений исходя из данных таблицы:

$$\begin{cases} 1988{,}52\tilde{\beta}_1 + 110{,}13\tilde{\beta}_0 = 59\,847{,}06, \\ 110{,}13\beta_1 + 6\tilde{\beta}_0 = 3288. \end{cases}$$

Решением данной системы нормальных уравнений будут следующие числа: $\tilde{\beta}_1 = 15{,}317$; $\tilde{\beta}_0 = 266{,}86$.

Таким образом, уравнение регрессии, описывающее зависимость между ценой на нефть и индексом акций нефтяной компании, можно записать как: $\tilde{y} = 15{,}317x + 266{,}86$.

На основании полученного уравнения регрессии можно сделать вывод о том, что с изменением цены на нефть на 1 денежную единицу за баррель индекс акций нефтяной компании изменяется примерно на 15,317 процентных пункта.

2. Альтернативный метод нахождения параметров уравнения парной регрессии

Традиционно параметры уравнения парной регрессии β_0 и β_1 оцениваются с помощью МНК, однако в случае парной регрессионной модели возможен и другой подход к оценке параметров регрессионной функции. Запишем уравнение парной регрессии в следующем виде:

$$y = \bar{y} + \beta_{yx}\left(x - \bar{x}\right), \tag{1}$$

где y — значение зависимой переменной;

x — значение независимой переменной;

$\bar{y}$ — среднее значение зависимой переменной, вычисленное на основе выборочных данных. Чаще всего это значение вычисляется по формуле средней арифметической:

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}, \tag{2}$$

где y_i — значения зависимой переменной, $i = \overline{1,n}$;

n — объем выборки;

$\bar{x}$ — среднее значение независимой переменной, которое вычисляется аналогично среднему значению зависимой переменной;

β_{yx} — выборочный коэффициент регрессии y по x. Он характеризует, насколько в среднем изменится результативный показатель y при изменении факторного показателя x на единицу своего измерения.

Вычисляется выборочный коэффициент регрессии y по x с помощью следующей формулы:

$$\beta_{yx} = r_{yx} \times \frac{S_y}{S_x}, \tag{3}$$

где r_{yx} — выборочный парный коэффициент корреляции, определяемый как:

$$r_{yx} = \frac{\overline{yx} - \bar{y}\bar{x}}{S_y S_x}. \tag{4}$$

Выборочный парный коэффициент корреляции показывает тесноту связи между изучаемыми признаками. Он изменяется в пределах [—1; +1]. Если $r_{yx} \in [0; +1]$, то связь между признаками прямая. Если $r_{yx} \in [-1;0]$, то связь между признаками обратная.

Если $r_{yx} = 0$, то связь между признаками отсутствует. Если $r_{yx} = 1$ или $r_{yx} = -1$, то связь между изучаемыми признаками является функциональной, т. е. характеризуется полным соответствием между x и y. Примером функциональной зависимости могут служить математические и статистические формулы, например: $S = a^2$. При таком значении парного коэффициента корреляции регрессионный анализ между изучаемыми показателями не проводится. Данная связь не подлежит численной характеристике, так как на практике массовым социально-экономическим явлениям присущи иные виды связи (в частности, корреляционная связь);

$\overline{yx}$ — среднее арифметическое значение произведения факторного и результативного признаков;

S_y — выборочное среднеквадратическое отклонение зависимой переменной y. Этот показатель характеризует, на сколько единиц в среднем отклоняются значения зависимого признака y от его среднего значения $\overline{y}$. Он вычисляется по формуле:

$$S_y = \sqrt{\overline{y^2} - \overline{y}^2}; \qquad (5)$$

$\overline{y^2}$ — среднее значение из квадратов значений результативной переменной y:

$$\overline{y^2} = \frac{\sum_{i=1}^{n} y_i^2}{n}; \qquad (6)$$

$\overline{y}^2$ — квадрат средних значений результативной переменной y:

$$\overline{y}^2 = \left(\frac{\sum_{i=1}^{n} y_i}{n} \right)^2; \qquad (7)$$

S_x — выборочное среднеквадратическое отклонение независимой переменной x. Этот показатель характеризует, на сколько единиц в среднем отклоняются значения независимого признака x от его среднего значения $\overline{x}$. Он вычисляется аналогично среднеквадратическому отклонению зависимого показателя y.

При оценивании коэффициента β_{yx} в модели регрессионной зависимости результативного показателя y от факторного показателя x с помощью рассмотренного метода следует помнить о том, что $r_{yx} = r_{xy}$, но $\beta_{yx} \neq \beta_{xy}$.

ЛЕКЦИЯ № 4. Оценка дисперсии случайной ошибки регрессии. Состоятельность и несмещенность МНК-оценок. Теорема Гаусса — Маркова

В большинстве случаев генеральная дисперсия случайной ошибки — величина неизвестная, поэтому возникает необходимость в расчете ее несмещенной выборочной оценки.

Несмещенной оценкой дисперсии случайной ошибки линейного **уравнения парной регрессии** является величина:

$$\tilde{G}^2(\varepsilon) = \tilde{S}^2(\varepsilon) = \frac{\sum_{i=1}^{n} e_i^2}{n-2}, \qquad (1)$$

где n — объем выборки;

e_i — остатки регрессионной модели:

$$e_i = y_i - \tilde{y}_i = y_i - \tilde{\beta}_0 - \tilde{\beta}_1 x_i.$$

Оценка дисперсии, вычисляемая по формуле (1), также называется **исправленной дисперсией**.

В случае множественной линейной регрессии оценка дисперсии случайной ошибки вычисляется по формуле:

$$\tilde{S}^2(\varepsilon) = \frac{\sum_{i=1}^{n} e_i^2}{n-k-1},$$

где k — число оцениваемых параметров модели регрессии.

Оценкой матрицы ковариаций случайных ошибок $Cov(\varepsilon)$ будет являться оценочная матрица ковариаций:

$$\tilde{C}(\varepsilon) = \tilde{S}^2(\varepsilon) \times \mathrm{I}n, \qquad (2)$$

где In — единичная матрица.

Оценка дисперсии случайной ошибки уравнения регрессии подчиняется χ^2 (хи-квадрат) закону распределения с (n — k — 1) степенями свободы, где k — число оцениваемых параметров.

Докажем несмещенность оценки дисперсии, т. е. необходимо доказать, что $\mathrm{E}(\tilde{S}^2(\varepsilon)) = G^2(\varepsilon)$.

Примем без доказательства следующее выражения:

$$\mathrm{E}(\tilde{S}^2(\varepsilon)) = \frac{n-1}{n} \times G^2(\varepsilon),$$

$$\tilde{S}^2(\varepsilon) = \frac{n}{n-1} \times S^2(\varepsilon),$$

где $G^2(\varepsilon)$ — генеральная дисперсия случайной ошибки;

$S^2(\varepsilon)$ — выборочная дисперсия случайной ошибки;

$\tilde{S}^2(\varepsilon)$ — выборочная оценка дисперсии случайной ошибки.

Тогда:

$$\mathrm{E}(\tilde{S}^2(\varepsilon)) = \mathrm{E}\left(\frac{n}{n-1} \times S^2(\varepsilon)\right) = \frac{n}{n-1}\mathrm{E}(S^2(\varepsilon)) =$$

$$= \frac{n}{n-1} \times \frac{n-1}{n} \times G^2(\varepsilon) = G^2(\varepsilon),$$

что и требовалось доказать.

Таким образом, $\tilde{S}^2(\varepsilon)$ является несмещенной оценкой для $G^2(\varepsilon)$.

Теоретически можно предположить, что оценка любого параметра регрессии, полученная методом наименьших квадратов, состоит из **двух компонент**:

1) константы, т. е. истинного значения параметра;

2) случайной ошибки $Cov(x,\varepsilon)$, вызывающей вариацию параметра регрессии.

На практике такое разложение невозможно в связи с неизвестностью истинных значений параметров уравнения регрессии и значений случайной ошибки, но в теории оно может оказаться полезным при изучении статистических свойств МНК-оценок: состоятельности, несмещенности и эффективности.

Докажем, что значение МНК-оценки $\tilde{\beta}_1$ зависит от величины случайной ошибки ε.

МНК-оценка параметра регрессии β_1 **рассчитывается по формуле**:

$$\tilde{\beta}_1 = \frac{Cov(x,y)}{G^2(x)}.$$

Ковариация между зависимой переменной y и независимой переменной x **может быть представлена как**:

$$Cov(x,y) = Cov(x,(\beta_0 + \beta_1 x + \varepsilon)) = Cov(x,\beta_0) + Cov(x,\beta_1 x) + Cov(x,\varepsilon).$$

Дальнейшие преобразования полученного выражения проводятся исходя из свойств ковариации:

1) ковариация между переменной x и какой-либо константой A равна нулю:

$$Cov(x,A) = 0, \text{ где } A = \text{const};$$

2) ковариация переменной x с самой собой равна дисперсии этой переменной:

$$Cov(x,x) = G^2(x).$$

Следовательно, на основании свойств ковариации можно записать, что:

$$Cov(x, \beta_0) = 0, \text{ так как } \beta_0 = \text{const};$$

$$Cov(x, \beta_1 x) = \beta_1 \times Cov(x, x) = \beta_1 \times G^2(x).$$

Таким образом, **ковариация** между зависимой и независимой переменными $Cov(x, y)$ **может быть представлена в виде выражения**:

$$Cov(x,y) = \beta_1 G^2(x) + Cov(x, \varepsilon).$$

В результате несложных преобразований МНК-оценка параметра уравнения регрессии β_1 принимает вид:

$$\tilde{\beta}_1 = \frac{\beta_1 G^2(x) + Cov(x,\varepsilon)}{G^2(x)} = \beta_1 + \frac{Cov(x,\varepsilon)}{G^2(x)}. \qquad (3)$$

Из формулы (3) следует, что МНК-оценка $\tilde{\beta}_1$ действительно может быть представлена как сумма константы β_1 и случайной ошибки $Cov(x, \varepsilon)$, которая и вызывает вариацию данного параметра регрессии.

Аналогично доказывается, что и оценка параметра регрессии $\tilde{\beta}_0$, полученная методом наименьших квадратов, и несмещенная оценка дисперсии случайной ошибки $\tilde{S}^2(\varepsilon)$ могут быть представлены как сумма постоянной составляющей (константы) и случайной компоненты, которая зависит от ошибки уравнения регрессии ε.

1. Состоятельность и несмещенность МНК-оценок

Для того чтобы оценку $\tilde{\vartheta}_i$, полученную с помощью метода наименьших квадратов, можно было бы принять за оценку параметра ϑ_i, необходимо и достаточно, чтобы оценка $\tilde{\vartheta}_i$ удовлетворяла трем статистическим свойствам: несмещенности, состоятельности и эффективности.

1. $\tilde{\vartheta}_i$ называется **несмещенной оценкой** для параметра ϑ_i, если ее выборочное математическое ожидание равно оцениваемому параметру генеральной совокупности, т. е.

$$\mathrm{E}(\tilde{\vartheta}_i) = \vartheta_i, \qquad (3)$$

$$\mathrm{E}\left(\tilde{\vartheta}_i\right) - \vartheta_i = \varphi_i,$$

где φ_i — смещение оценки.

Докажем, что МНК-оценка $\tilde{\beta}_1$ является несмещенной оценкой параметра β_1 для нормальной линейной регрессионной модели. Исходя из предпосылок данной модели, можно записать:

1) x — неслучайная детерминированная величина;

2) $G^2(x) = \text{const}$ — дисперсия независимого признака является известной постоянной величиной;

3) $\mathrm{E}(Cov(x,\varepsilon)) = 0$ — случайная ошибка и независимый признак не коррелированы между собой;

4) $\mathrm{E}(\varepsilon_i) = 0$ — математическое ожидание случайной ошибки уравнения равно нулю во всех наблюдениях;

5) $Cov(\varepsilon_1, \varepsilon_2) = \mathrm{E}(\varepsilon_1, \varepsilon_2) = 0$ — случайные ошибки уравнения регрессии не коррелированы между собой, т. е. ковариация случайных ошибок любых двух разных наблюдений равна нулю.

Исходя из определения свойства несмещенности необходимо доказать, что $\mathrm{E}\left(\tilde{\beta}_1\right) = \beta_1$.

Доказательство через ковариационную матрицу:

$$\mathrm{E}\left(\tilde{\beta}_1\right) = \mathrm{E}\left(\beta_1 + \frac{Cov(x,\varepsilon)}{G^2(x)}\right) = \beta_1 + \mathrm{E}\left(\frac{Cov(x,\varepsilon)}{G^2(x)}\right) = \beta_1 + \frac{0}{G^2(x)} = \beta_1$$

или в развернутом виде

$$\mathrm{E}(\tilde{\beta}_1) = \mathrm{E}\left(\beta_1 + \sum \frac{(x_i - \bar{x})}{\sum (x_i - \bar{x})^2} \times \varepsilon_i\right) = \beta_1 + \mathrm{E}\left(\sum \frac{(x_i - \bar{x})}{\sum (x_i - \bar{x})^2} \times \varepsilon_i\right) =$$

$$= \beta_1 + \mathrm{E}\left(\sum \frac{(x_i - \bar{x})}{\sum \left(x_i - \bar{x}\right)^2}\right) \times \mathrm{E}\left(\varepsilon_i\right) = \beta_1.$$

Таким образом, МНК-оценка $\tilde{\beta}_1$ является несмещенной оценкой параметра β_1.

Несмещенность МНК-оценки $\tilde{\beta}_0$ доказывается аналогично.

Запишем доказательство несмещенности МНК-оценок параметров β_1 в матричной форме:

$$\mathrm{E}\left(\tilde{\beta}\right) = \mathrm{E}((X^T X)^{-1} X^T Y) = \mathrm{E}\left[(X^T X)^{-1} X^T (X\beta + \varepsilon)\right] =$$

$$= \mathrm{E}\left[(X^T X)^{-1} X^T X\beta + (X^T X)^{-1} X^T \varepsilon\right] =$$

$$= \beta + ((X^T X)^{-1} X^T \mathrm{E}(\varepsilon)) = \beta,$$

т. е. $\mathrm{E}\left(\tilde{\beta}\right) = \beta$, что доказывает несмещенность МНК-оценок параметров β_i.

2. $\tilde{\vartheta}_i$ является **состоятельной оценкой** для параметра ϑ_i, если она удовлетворяет закону больших чисел (ЗБЧ). Закон больших чисел гласит о том, что с увеличением выборки значение оценки $\tilde{\vartheta}_i$ стремится к значению параметра ϑ_i генеральной совокупности:

$$P\left(\left|\tilde{\vartheta}_i - \vartheta_i\right| < \theta\right) \to 1 \text{ при } n \to \infty. \qquad (4)$$

Это же условие можно записать с помощью теоремы Бернулли:

$$\tilde{\vartheta}_i \xrightarrow{P} \vartheta_i \text{ при } n \to \infty,$$

т. е. значение оценки $\tilde{\vartheta}_i$ сходится по вероятности к значению параметра ϑ_i генеральной совокупности при условии, что объем выборки стремится к бесконечности.

Для определения состоятельности оценки достаточно выполнения двух условий:

1) $\varphi_i = 0$ или $\varphi_i \to 0$ при $n \to \infty$ — смещение оценки равно нулю или стремится к нему при объеме выборки, стремящемся к бесконечности;

2) $G^2(\tilde{\vartheta}_i) \to 0$ при $n \to \infty$ — дисперсия оценки параметра стремится к нулю при объеме выборки, стремящемся к бесконечности.

Докажем первое условие состоятельности для МНК-оценки $\tilde{\beta}_1$:

$$\varphi_1 = \mathrm{E}\left(\tilde{\beta}_1\right) - \beta_1 = \beta_1 - \beta_1 = 0.$$

Докажем второе условие состоятельности для МНК-оценки:

$$G^2(\tilde{\beta}_1) = \mathrm{E}(\tilde{\beta}_1 - \beta_1)^2 = \mathrm{E}\left[\left(\sum \frac{(x_i - \bar{x})}{\sum (x_i - \bar{x})^2} \times \varepsilon_i\right)^2\right] =$$

$$= \mathrm{E}\left[\sum \frac{(x_i - \bar{x})}{\left[\sum (x_i - \bar{x})^2\right]^2} \times \varepsilon_i^2\right] =$$

$$= \sum \frac{(x_i - \bar{x})^2}{\left[\sum (x_i - \bar{x})^2\right]^2} \times \mathrm{E}\left(\varepsilon_i^2\right) = \frac{G^2(x)}{\sum (x_i - \bar{x})^2}.$$

Докажем состоятельность МНК-оценок параметров β_i в матричной форме:

$$Cov(\tilde{\beta}) = \mathrm{E}\left[(\tilde{\beta}-\beta)\times\left(\tilde{\beta}-\beta\right)^T\right] = \mathrm{E}((X^TX)^{-1}X^T\varepsilon\varepsilon^TX(X^TX)^{-1}) =$$
$$= (X^TX)^{-1}\mathrm{E}(\varepsilon\varepsilon^T)X(X^TX)^{-1} = G^2(\varepsilon)(X^TX)^{-1},$$

Таким образом, МНК-оценка $\tilde{\beta}_1$ подчиняется нормальному закону распределения с математическим ожиданием β_1 и дисперсией

$$(G^2(x)/\sum(x_i-\bar{x})^2)\ /\tilde{\beta}_1 \sim N\left(\beta_1;\ \frac{G^2(x)}{\sum(x_i-\bar{x})^2}\right)$$
$$\text{или } \tilde{\beta}_1 \sim N\left(\beta_1;\ G^2(\varepsilon)(X^TX)^{-1}_{22}\right),$$

где индекс $_{22}$ указывает на расположение дисперсии параметра β_1 в матрице ковариаций.

Состоятельность МНК-оценки $\tilde{\beta}_0$ доказывается аналогично.

Величины

$$S\left(\tilde{\beta}_1\right) = \sqrt{S^2(\varepsilon)\left(X^TX\right)^{-1}_{22}}$$

называются **оценками стандартных ошибок МНК-оценок** $\tilde{\beta}_1$ **и** $\tilde{\beta}_0$

Эффективность МНК—оценок доказывается с помощью теоремы Гаусса—Маркова.

Таким образом, оценки параметров уравнения регрессии и дисперсии случайной ошибки, полученные методом наименьших квадратов, являются оптимальными оценками, т. е. несмещенными, состоятельными и эффективными.

2. Эффективность МНК-оценок. Теорема Гаусса—Маркова

С помощью теоремы Гаусса — Маркова доказывается эффективность оценок неизвестных параметров уравнения регрессии, полученных с помощью МНК.

Нормальная, или классическая, линейная модель парной регрессии (регрессии с одной переменной) строится исходя из следующих предположений:

1) факторный признак x_i является неслучайной или детерминированной величиной, не зависящей от распределения случайной ошибки уравнения регрессии ε_i;

2) математическое ожидание случайной ошибки уравнения регрессии равно нулю во всех наблюдениях: $\mathrm{E}(\varepsilon_i) = 0$, где $i = 1,n$;

3) дисперсия случайной ошибки уравнения регрессии является постоянной для всех наблюдений: $D(\varepsilon_i) = \mathrm{E}(\varepsilon_i^2) = G^2 = \mathrm{const}$;

4) случайные ошибки уравнения регрессии не коррелированы между собой, т. е. ковариация случайных ошибок любых двух разных наблюдений равна нулю: $Cov(\varepsilon_i, \varepsilon_j) = \mathrm{E}(\varepsilon_i\ \varepsilon_j) = 0$, где $i \neq j$. Это верно тогда, когда изучаемые данные не являются временными рядами;

5) основываясь на 3 и 4-м предположениях, добавляется условие о том, что ошибка уравнения регрессии является случайной величиной, подчиняющейся нормальному закону распределения с нулевым математическим ожиданием и дисперсией $G^2 / \varepsilon_i \sim N(0, G^2)$.

Тогда оценки неизвестных параметров уравнения регрессии, полученные методом наименьших квадратов, имеют наименьшую дисперсию в классе всех линейных несмещенных оценок, т. е. оценки МНК являются эффективными оценками неизвестных параметров $\beta_0, ..., \beta_n$.

Для нормальной линейной модели множественной регрессии теорема Гаусса — Маркова звучит точно так же.

Дисперсии МНК-оценок неизвестных параметров записываются с помощью матрицы ковариаций. Матрица ковариаций МНК-оценок параметров линейной модели парной регрессии выглядит так:

$$Cov(\tilde{\beta}) = \begin{pmatrix} G^2(\tilde{\beta}_0) & 0 \\ 0 & G^2(\tilde{\beta}_1) \end{pmatrix},$$

где $G^2(\tilde{\beta}_0)$ — дисперсия МНК-оценки параметра уравнения регрессии;

$G^2(\tilde{\beta}_1)$ — дисперсия МНК-оценки параметра уравнения регрессии .

Общая формула для расчета матрицы ковариаций МНК-оценок коэффициентов регрессии: $Cov(\tilde{\beta}) = G^2(\varepsilon) \times (X^T X)^{-1}$,

где $G^2(\varepsilon)$ — дисперсия случайной ошибки уравнения регрессии.

Рассмотрим процесс определения дисперсий оценок коэффициентов линейной модели парной регрессии, полученных с помощью метода наименьших квадратов.

Дисперсия МНК-оценки коэффициента уравнения регрессии β_0:

$$G^2(\tilde{\beta}_0) = \frac{G^2(\varepsilon)}{n}\left(1 + \frac{\overline{x^2}}{G^2(x)}\right);$$

дисперсия МНК-оценки коэффициента уравнения регрессии β_1:

$$G^2(\tilde{\beta}_1) = \frac{G^2(\varepsilon)}{n \times G^2(x)},$$

где $G^2(\varepsilon)$ — дисперсия случайной ошибки уравнения регрессии ε;

$G^2(x)$ — дисперсия независимого признака уравнения регрессии;

n — объем выборочной совокупности.

На практике значение дисперсии случайной ошибки уравнения регрессии $G^2(\varepsilon)$ зачастую неизвестно, поэтому для определения матрицы ковариаций МНК-оценок применяют оценку дисперсии случайной ошибки уравнения регрессии $S^2(\varepsilon)$. В случае парной линейной регрессии оценка дисперсии случайной ошибки будет рассчитываться по формуле:

$$\tilde{G}^2(\varepsilon) = \tilde{S}^2(\varepsilon) = \frac{\sum_{i=1}^{n} e_i^2}{n-2},$$

где $e_i^2 = y_i - \tilde{y}_i$ — остатки регрессионной модели.

Тогда общую формулу для расчета матрицы ковариаций МНК-оценок коэффициентов регрессии на основе оценки дисперсии случайной ошибки уравнения регрессии можно записать следующим образом:

$$\tilde{G}(\tilde{\beta}) = S^2(\varepsilon) \times (X^T X)^{-1}.$$

В случае линейной модели парной регрессии оценка дисперсии МНК-оценки коэффициента уравнения регрессии β_0:

$$S^2(\tilde{\beta}_0) = \frac{\sum_{i=1}^{n} e_i^2 \times \sum_{i=1}^{n} x_i^2}{n \times (n-2) \times \sum_{i=1}^{n} (x_i - \bar{x})^2};$$

оценка дисперсии МНК-оценки коэффициента уравнения регрессии β_1:

$$S^2(\tilde{\beta}_1) = \frac{\sum_{i=1}^{n} e_i^2}{(n-2) \times \sum_{i=1}^{n} (x_i - \bar{x})^2}.$$

ЛЕКЦИЯ № 5. Определение качества модели регрессии. Проверка гипотез о значимости коэффициентов регрессии, корреляции и уравнения парной регрессии

Качество модели регрессии — адекватность построенной модели исходным (наблюдаемым) данным.

Качество парной линейной регрессии определяется с помощью парного **линейного коэффициента корреляции**:

$$r_{yx} = \frac{\overline{xy} - \bar{x}\bar{y}}{G(x)G(y)} = \frac{Cov(x, y)}{G(x)G(y)},$$

где $G(x)$ — среднеквадратическое отклонение независимого признака;

$G(y)$ — среднеквадратическое отклонение зависимого признака.

Коэффициент парной линейной корреляции можно рассчитать через МНК-оценку параметра уравнения регрессии $\tilde{\beta}$:

$$r_{yx} = \tilde{\beta}\frac{G(x)}{G(y)}.$$

Парный коэффициент корреляции показывает тесноту связи между изучаемыми признаками. Он изменяется в пределах $[-1; +1]$. Если $r_{yx} \in [0; +1]$ то связь между признаками прямая. Если $r_{yx} \in [-1; 0]$, то связь между признаками обратная. Если $r_{yx} = 0$, то связь между признаками отсутствует. Если $r_{yx} = 1$ или $C = -1$,то связь между изучаемыми признаками является функциональной, т. е. характеризуется полным соответствием между x и y. Чем ближе $|r_{xy}|$ к 1, тем более тесной считается связь между изучаемыми признаками.

Парный коэффициент корреляции определяется для количественных переменных.

Если парный линейный коэффициент корреляции r_{yx} возвести в квадрат, то получим коэффициент детерминации r^2_{yx}. Данный коэффициент показывает, на сколько процентов вариация результативного признака объясняется вариацией факторного признака в общем объеме вариации.

Чтобы оценить качество линейной множественной модели регрессии, необходимо воспользоваться теоремой о разложении дисперсий.

Общая дисперсия зависимой переменной может быть разложена на две составляющие — объясненную и необъясненную построенным уравнением регрессии дисперсии:

$$G^2(y) = \sigma^2(y) + \delta^2(y),$$

где $\sigma^2(y) = \dfrac{\sum_{i=1}^{n}(\tilde{y}_i - y_i)^2}{n}$ — объясненная с помощью построенного уравнения регрессии дисперсия переменной *y*;

$\delta^2(y) = \dfrac{\sum_{i=1}^{n} e_i^2}{n}$ — необъясненная или остаточная дисперсия переменной *y*.

С помощью данной теоремы можно рассчитать множественный коэффициент корреляции между результативным признаком *y* и несколькими **факторными признаками** *x*:

$$R_y = \sqrt{\frac{\sigma^2(y)}{G^2(y)}}.$$

Множественный коэффициент корреляции показывает тесноту связи между результативным и факторными признаками. Трактовка его значений аналогична трактовке значений парного линейного коэффициента корреляции.

Квадрат множественного линейного коэффициента корреляции называется теоретическим **коэффициентом детерминации**:

$$R_y^2 = \frac{\sigma^2(y)}{G^2(y)}.$$

Этот коэффициент показывает, на сколько процентов вариация результативного признака объясняется вариацией факторных признаков *x*. Величина $1 - R_y^2$ показывает ту долю вариации результативного признака, которую модель регрессии учесть не смогла.

Среднеквадратическая ошибка (Mean square error — MSE) уравнения регрессии схожа по построению с показателем **среднеквадратического отклонения**:

$$MSE = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n-h}},$$

где *h* — число параметров уравнения регрессии.

Если *MSE* окажется меньше $\sigma(y)$, то построенную модель можно считать качественной. Показатель среднеквадратического отклонения наблюдаемых значений зависимой переменной от модельных значений, рассчитанных по уравнению регрессии, определяется как:

$$\sigma(y) = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n}}.$$

Показатель средней ошибки аппроксимации рассчитывается по формуле:

$$\bar{A}_y = \frac{1}{n}\sum_{i=1}^{n}\frac{|y - \tilde{y}_i|}{y_i}.$$

Максимально допустимым значением данного показателя считается 12—15%. Если средняя ошибка аппроксимации составляет менее 6—7%, то качество модели считается хорошим.

1. Проверка гипотезы о значимости коэффициентов регрессии

Чтобы построенную модель можно было использовать для дальнейших экономических расчетов, например для построения прогноза зависимой переменной, проверки качества построенной модели недостаточно. Необходимо также проверить значимость полученных с помощью метода наименьших квадратов оценок коэффициентов регрессии, значимость парного линейного коэффициента корреляции и уравнения регрессии в целом с помощью статистических гипотез.

При проверке значимости (предположения того, что параметры отличаются от нуля) коэффициентов регрессии выдвигается основная гипотеза H_0 о незначимости полученных оценок, например:

$$H_0 / \ \beta_1 = 0;$$

в качестве альтернативной (или обратной) выдвигается гипотеза о значимости коэффициентов регрессии, например:

$$H_1 / \beta_1 \neq 0.$$

Для проверки выдвинутых гипотез используется t-критерий (t-статистика) Стьюдента. Наблюдаемое значение t-критерия, вычисленное на основе выборочных данных, сравнивают со значением t-критерия, определяемого по таблице распределения Стьюдента. Значение t-статистики, найденное по таблице, называется критическим. Критическое значение t-критерия $t_{крит}(\alpha; n-h)$ зависит от двух параметров: уровня значимости и числа степеней свободы.

Уровень значимости α — величина, определяемая по формуле:

$\alpha = 1 - \gamma$,

где γ — доверительная вероятность попадания оцениваемого параметра в доверительный интервал.

Данную величину необходимо брать близкую к единице (0,95—0,99). Таким образом, α — это вероятность того, что оцениваемый параметр не попадет в доверительный интервал, равный 0,05 или 0,01.

Число степеней свободы — показатель, который определяется как разность между объемом выборки (n) и числом оцениваемых параметров по данной выборке (h). Для модели парной линейной регрессии число степеней свободы рассчитывается как ($n - 2$), так как по выборке оцениваются два параметра: β_0 и β_1.

Выдвинутые гипотезы проверяются следующим образом:

1) если модуль наблюдаемого значения t-критерия больше критического значения t-критерия, т. е. $|t_{набл}| > t_{крит}$, то с вероятностью $(1 - \alpha)$ или γ основную гипотезу о незначимости параметров регрессии отвергают, т. е. параметры регрессии не равны нулю;

2) если модуль наблюдаемого значения t-критерия меньше или равен критическому значению t-критерия, т. е. $|t_{набл}| \le t_{крит}$, то с вероятностью α или $(1 - \gamma)$ основная гипотеза о незначимости параметров регрессии принимается, т. е. параметры регрессии почти не отличаются от нуля или равны нулю.

Формула наблюдаемого значения t-критерия Стьюдента для проверки гипотезы $H_0 / \beta_1 = 0$ имеет вид:

$$t_{набл} = \frac{\tilde{\beta}_1}{\omega(\beta_1)},$$

где $\tilde{\beta}_1$ — оценка параметра регрессии β_1;

$\omega(\beta_1)$ — величина стандартной ошибки параметра регрессии β_1.

В случае парной линейной модели регрессии показатель вычисляется следующим образом:

$$\omega(\beta_1) = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{(n-2) \times \sum_{i=1}^{n} (x_i - \bar{x})^2}}.$$

Числитель стандартной ошибки может быть рассчитан через парный коэффициент **детерминации как**:

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(y_i - \tilde{y}_i\right)^2 = n \times G^2(y) \times (1 - r_{yx}^2),$$

где $G^2(y)$ — общая дисперсия зависимого признака;

r_{yx}^2 — парный коэффициент детерминации между зависимым и независимым признаками.

Формула наблюдаемого значения t-критерия Стьюдента для проверки гипотезы $H_0 / \beta_0 = 0$ имеет вид:

$$t_{набл} = \frac{\tilde{\beta}_0}{\omega(\beta_0)},$$

где $\tilde{\beta}_0$ — оценка параметра регрессии;

$\omega(\beta_0)$ — величина стандартной ошибки параметра регрессии β_0.

В случае парной линейной модели регрессии показатель $\omega(\beta_0)$ вычисляется та

$$\omega(\beta_0) = \sqrt{\frac{\sum_{i=1}^{n} e_i^2 \times \sum_{i=1}^{n} x_i^2}{n \times (n-2) \times \sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2}}.$$

2. Проверка гипотезы о значимости парного линейного коэффициента корреляции

При проверке значимости коэффициента корреляции между независимым признаком *x* и зависимым признаком *y* (предположения того, что изучаемый параметр отличается от нуля), выдви-

гается основная гипотеза H_0 о его незначимости: $H_0 / r_{yx} = 0$; в качестве альтернативной (или обратной) выдвигается гипотеза H_1 о **значимости коэффициента корреляции**: $H_1 / r_{yx} \neq 0$.

Для проверки выдвинутых гипотез используется t-критерий (t-статистику) Стьюдента.

Гипотезы проверяются таким образом:

1) если модуль наблюдаемого значения t-критерия больше критического значения t-критерия, т. е. $|t_{набл}| > t_{крит}$, то с вероятностью $(1 - \alpha)$ или γ основную гипотезу о незначимости парного линейного коэффициента корреляции отвергают, между изучаемыми признаками и существует корреляционная связь, которую аналитически можно оценить с помощью построения уравнения парной регрессии;

2) если модуль наблюдаемого значения t-критерия меньше или равен критическому значению t-критерия, т. е. $|t_{набл}| \leq t_{крит}$,то с вероятностью α или $(1 - \gamma)$ основная гипотеза о незначимости коэффициента корреляции принимается, т. е. между изучаемыми признаками x и y корреляционная связь отсутствует, построение уравнения регрессии в данном случае нецелесообразно.

Критическое значение t-критерия $t_{крит}(\alpha; n - h)$, где α — уровень значимости, $(n - h)$ — число степеней свободы, определяется по таблице распределений t-критерия Стьюдента.

Формула значения t-критерия Стьюдента для проверки гипотезы $H_0 / r_{yx} = 0$ имеет вид:

$$t_{набл} = \frac{r_{yx}}{\omega(r_{yx})},$$

где r_{yx} — выборочный парный коэффициент корреляции между переменными x и y, вычисляемый по формуле:

$$r_{yx} = \frac{\overline{yx} - \bar{x}\bar{y}}{S_y S_x};$$

$\omega(r_{yx})$ — величина стандартной ошибки парного выборочного коэффициента корреляции.

При линейной парной модели регрессии эта величина рассчитывается как:

$$\omega(r_{yx}) = \sqrt{\frac{\left(1 - r_{yx}^2\right)}{(n-2)}}.$$

Подставим данную формулу в выражение для расчета наблюдаемого значения t-критерия Стьюдента для проверки гипотезы $H_0 / r_{yx} = 0$, получим:

$$t_{набл} = \frac{r_{yx}}{\sqrt{1 - r_{yx}^2}} \times (n - 2).$$

t-статистика Стьюдента применяется для проверки значимости парного выборочного коэффициента корреляции в случае, если объем выборки достаточно велик ($n \geq 30$) и коэффициент корреляции по модулю значительно меньше 1 $0{,}45 \leq |r_{xy}| \leq 0{,}75$.

Если модуль парного выборочного коэффициента корреляции близок к 1, то гипотеза $H_0 / r_{yx} = 0$ может быть проверена (помимо t-критерия) с помощью z-статистики. Этот метод оценки значимости коэффициента корреляции был предложен ***Р. Фишером***.

Величина z связана с парным выборочным коэффициентом корреляции определенным отношением:

$$z = 0{,}5 \; ln\left(\frac{1 + r_{yx}}{1 - r_{yx}}\right).$$

Величина z подчиняется нормальному закону распределения, поэтому проверка основной гипотезы о незначимости парного коэффициента корреляции $H_0 / r_{yx} = 0$ отождествляется с проверкой гипотезы о незначимости величины z $H_0 / z = 0$ по формуле:

$$t_{набл} = \frac{z}{\omega(z)},$$

где $\omega(z)$ — величина стандартной ошибки величины z, определяемая как:

$$\omega(z) = \frac{1}{\sqrt{n - 3}}.$$

Критическое значение этого критерия $t_{крит}$ определяют по таблице нормального распределения (z-распределения) с доверительной вероятностью γ или $(1 - \alpha)$.

Проверка гипотез осуществляется аналогично проверке гипотез по t-критерию Стьюдента:

1) при $|t_{набл}| > t_{крит}$ основная гипотеза $H_0 / r_{yx} = 0$ или $H_0 / r_{yx} = 0$ отвергается и выборочный парный коэффициент корреляции считается значимым;

2) при $|t_{набл}| \le t_{крит}$ основная гипотеза $H_0 / r_{yx} = 0$ или $H_0 / r_{yx} = 0$ принимается и выборочный парный коэффициент корреляции считается незначимым.

3. Проверка гипотезы о значимости уравнения парной регрессии. Теорема о разложении сумм квадратов

Проверка гипотезы значимости парного линейного уравнения регрессии сводится к проверке гипотез о значимости коэффициентов регрессии β_0 и β_1 или парного коэффициента детерминации r^2_{yx}.

В этом случае могут быть выдвинуты следующие **основные гипотезы**:

1) $H_0 / \beta_0 = 0$ и $H_0 / \beta_1 = 0$ — коэффициенты регрессии являются незначимыми и уравнение регрессии также является незначимым;

2) $H_0 / r^2_{yx} = 0$ — парный коэффициент детерминации незначим и уравнение регрессии также является незначимым.

Альтернативной (или обратных к основным) выступает гипотезы;

3) $H_1 / \beta_0 \neq 0$ и $H_1 / \beta_1 \neq 0$ — коэффициенты регрессии значимо отличаются от нуля и построенное уравнение регрессии является значимым;

4) $H_1 / r^2_{yx} \neq 0$ — парный коэффициент детерминации значимо отличается от нуля, следовательно, построенное уравнение регрессии является значимым.

Для проверки гипотезы значимости уравнения регрессии в целом используется F-критерий Фишера—Снедекора.

Гипотеза проверяется следующим образом:

1) если наблюдаемое значение F-критерия больше критического значения данного критерия, т. е. $F_{набл} > F_{крит}$, то с вероятностью α основная гипотеза о незначимости коэффициентов уравнения регрессии или парного коэффициента детерминации отвергается, и уравнение регрессии признается значимым;

2) если наблюдаемое значение F-критерия меньше критического значения данного критерия, т. е. $F_{набл} < F_{крит}$, то с вероятностью $(1 - \alpha)$ основная гипотеза о незначимости коэффициентов уравнения регрессии или парного коэффициента детерминации принимается, и построенное уравнение регрессии признается незначимым.

Критическое значение F-критерия находится по таблице распределения Фишера—Снедекора в зависимости от следующих параметров: уровня значимости α и числа степеней свободы: $k_1 = h - 1$ и $k_2 = n - h$, где n — это объем выборки, а h — число оцениваемых по выборке параметров. В случае проверки значимости уравнения парной регрессии критическое значение F-статистики вычисляется как $F_{крит}(\alpha; 1; n - 2)$.

Формула наблюдаемого значения F-критерия для проверки гипотезы о незначимости уравнения регрессии в целом имеет вид:

$$F_{набл} = \frac{r_{yx}^2}{1 - r_{yx}^2} \times \frac{n - h}{h - 1};$$

в случае парной регрессии наблюдаемое значение F-критерия преобразуется в вид:

$$F_{набл} = \frac{r_{yx}^2}{1 - r_{yx}^2} \times (n - 2).$$

Данный критерий имеет распределение Фишера—Снедекора.

Коэффициент детерминации можно определить не только как квадрат парного линейного коэффициента корреляции или через теорему о разложении общей дисперсии результативной переменной на составляющие, но и через теорему о разложении сумм квадратов результативной переменной.

Сумма квадратов разностей между значениями результативной переменной и ее средним значением **по выборке может быть представлена таким образом**:

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (y_i - \tilde{y}_i)^2 + \sum_{i=1}^{n} (\tilde{y}_i - \bar{y})^2,$$

где $\sum_{i=1}^{n} (y_i - \bar{y})^2$ — общая сумма квадратов (Total Sum Square — TSS);

$\sum_{i=1}^{n} (y_i - \tilde{y}_i)$ — сумма квадратов остатков (Error Sum Square — ESS);

$\sum_{i=1}^{n}(\tilde{y}_i - \bar{y})^2$ — сумма квадратов объясненной регрессии (Regression Sum Square — RSS).

В векторной форме данное равенство можно записать как:

$$|y - \bar{y}_i|^2 = |y - \tilde{y}_i|^2 + |\tilde{y}_i - \bar{y}_i|^2 .$$

Рассмотрим общую сумму квадратов:

$$\begin{aligned}(y-\bar{y}_i)^T \times (y-\bar{y}_i) &= \left([y-\tilde{y}]+[\tilde{y}-\bar{y}_i]\right)^T \times \left([y-\tilde{y}]+[\tilde{y}-\bar{y}_i]\right) = \\ &= \underbrace{(y-\tilde{y})^T \times (y-\tilde{y})}_{ESS} + (y-\tilde{y})^T \times (\tilde{y}-\bar{y}_i) + \\ &+ (\tilde{y}-\bar{y}_i)^T \times (y-\tilde{y}) + \underbrace{(\tilde{y}-\bar{y}_i)^T \times (\tilde{y}-\bar{y}_i)}_{RSS} = \\ &= (y-\tilde{y})^T \times (\tilde{y}-\bar{y}_i) = e^T(x\tilde{\beta}-\bar{y}_i) = e^T x\tilde{\beta} - \bar{y}_i e^T = 0.\end{aligned}$$

Если в уравнение регрессии не включается свободный член β_0, это разложение остается верным.

Парный коэффициент детерминации может быть вычислен по следующим формулам:

$$r_{yx}^2 = 1 - \frac{ESS}{TSS} \quad \text{или} \quad r_{yx}^2 = \frac{RSS}{TSS}.$$

ЛЕКЦИЯ № 6. Построение прогнозов для модели парной линейной регрессии. Примеры оценивания параметров парной регрессии и проверки гипотезы о значимости коэффициентов и уравнения регрессии

Целью построения регрессионной функции на основе эмпирических данных является не только аппроксимация исходных данных с заданной точностью, но и возможность дальнейшего применения в экономических расчетах полученного уравнения регрессии. В частности, на основе регрессионной модели можно рассчитать прогнозное значение результативного признака при заданном значении факторного признака.

Для модели парной линейной регрессии точечный прогноз зависимой переменной y при заданном значении независимой переменной x_m будет выглядеть следующим образом:

$$y_m = \beta_0 + \beta_1 x_m + \varepsilon_m.$$

С доверительной вероятностью γ или $(1 - \alpha)$точечная оценка прогноза результативного признака y_m попадет в интервал прогноза, который определяется по формуле:

$$y_m - t\omega(m) \le y_m \le y_m + t\omega(m),$$

где y_m — точечная оценка прогноза результативного признака;

t — t-критерий Стьюдента, который определяется в зависимости от заданного уровня значимости α и числа степеней свободы $(n - 2)$ (в случае парной регрессионной модели);

$\omega(m)$ — величина ошибки прогноза в точке m .

Величина ошибки прогноза рассчитывается по формуле:

$$\omega(m) = \sqrt{S^2(\varepsilon) \times \left(\frac{n+1}{n} + \frac{(x_m - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right)},$$

где $S^2(\varepsilon)$ — несмещенная оценка дисперсии случайной ошибки линейного уравнения парной регрессии.

Рассмотрим подробнее процесс определения величины ошибки прогноза.

Пусть задана парная линейная регрессионная модель следующего вида:

$$y_i = \tilde{\beta}_0 + \tilde{\beta}_1 (x_i - \bar{x}) + \varepsilon_i,$$

где независимая переменная *x* представлена в центрированном виде.

Необходимо построить прогноз зависимой переменной *y* при заданном значении независимой переменной x_m:

$$y_m = \tilde{\beta}_0 + \tilde{\beta}_1 (x_m - \bar{x}) + \varepsilon_m.$$

Математическое ожидание зависимой переменной *y* в точке *m* определяется как:

$$\mathrm{E}(y_m / x_m) = \tilde{\beta}_0 + \tilde{\beta}_1 (x_m - \bar{x}) + \varepsilon_m.$$

Дисперсия зависимой переменной *y* в точке *m* **определяется как**:

$$D(y_m / x_m - \bar{x}) = D(\tilde{\beta}_0 + \tilde{\beta}_1 (x_m - \bar{x}) + \varepsilon_m) =$$
$$= D(\tilde{\beta}_0) + D(\tilde{\beta}_1 (x_m - \bar{x})) + D(\varepsilon_m) =$$
$$= \frac{G^2}{n} + (x_m - \bar{x})^2 \times \frac{G^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} + G^2,$$

где $D(\beta_0)$ — дисперсия оценки параметра β_0 парной линейной регрессии, рассчитываемая по формуле:

$$D(\tilde{\beta}_0) = D\left(\beta_0 + \frac{\sum \varepsilon_i}{n}\right) = D\left(\frac{\sum \varepsilon_i}{n}\right) =$$
$$= \frac{1}{n^2} \sum D(\varepsilon_i) = \frac{nG^2}{n^2} = \frac{G^2}{n}.$$

Точечная оценка прогноза результативной переменной y_m имеет нормальный закон распределения с математическим ожиданием $(\tilde{\beta}_o + \tilde{\beta}_1 (x_m - \bar{x}))$ и дисперсией

$$G^2 \times \left(\frac{n + 1}{n} + \frac{(x_m - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right) :$$

$$y_m / x_m \sim N\left(\tilde{\beta}_0 + \tilde{\beta}_1 (x_m - \bar{x}); \; G^2 \left(\frac{n+1}{n} + \frac{(x_m - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right) \right).$$

Если в выражение для дисперсии зависимой переменной y в точке m вместо дисперсии G^2 подставить ее оценку выборочную оценку S^2, то можно построить доверительный интервал для прогноза зависимой переменной при заданном значении независимой переменной x_m:

$$y_m / x_m \in \left[\tilde{\beta}_0 + \tilde{\beta}_1 (x_m - \bar{x}) \pm \left[\sqrt{S^2 \left(\frac{n+1}{n} + \frac{(x_m - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right)} \right] \right],$$

где S^2 для модели парной линейной регрессии рассчитывается по следующей формуле:

$$S^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-2}.$$

Прогнозный интервал можно преобразовать к виду:

$$y_m / x_m \in \left[\tilde{\beta}_0 + \tilde{\beta}_1 (x_m - \bar{x}) \pm \left[\sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n-2} \left(\frac{n+1}{n} + \frac{(x_m - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right)} \right] \right],$$

что и требовалось доказать.

Рассчитаем точечный прогноз.

На основании данных о цене на нефть x (долларов за баррель) и индексе акций нефтяной компании y (в процентных пунктах) было построено уравнение регрессии:

$$\tilde{y} = 15{,}317x + 266{,}86.$$

Если цена на нефть подскочит в связи с нефтяным кризисом на Ближнем Востоке и преодолеет рубеж в 20 долларов за баррель, остановившись на отметке в 22,13 доллара за 1 баррель. Требуется определить, какое влияние окажет этот ценовой скачок на уровень индекса акций нефтяной компании.

Подставим новое значение независимой переменной в уравнение регрессии с целью получения прогноза:

$$\tilde{y} = 15{,}317 \times 22{,}13 + 266{,}86 = 605{,}825.$$

Нефтяной кризис благотворно сказался на финансовом положении нефтяной компании, повысив индекс ее акций с 550 процентных пункта до 605,825 процентных пункта.

1. Пример оценивания параметров парной регрессии с помощью альтернативного метода

Определим оценки неизвестных параметров парного линейного уравнения регрессии с помощью альтернативного метода (табл. 2). Имеются данные по двадцати банкам страны о размере прибыли в денежных единицах (результативная переменная) и объемах выданных кредитов в денежных единицах (факторная переменная).

Таблица 2

Пример определения оценок неизвестных параметров парного линейного уравнения регрессии

№ наблюдения	y — прибыль, ден. ед.	x — кредиты, ден. ед.
1	19	200
2	30	300
3	26	200
4	22	220
5	13	100
6	35	250
7	28	250
8	30	300
9	40	280
10	37	300
11	18	150
12	20	250
13	15	100
14	38	300
15	20	120
16	30	220

Окончание табл. 2

№ наблюдения	y — прибыль, ден. ед.	x — кредиты, ден. ед.
17	30	290
18	28	260
19	19	160
20	15	150
Сумма	513	4400

На первом этапе определим r_{yx} — выборочный парный коэффициент корреляции по формуле:

$$r_{yx} = \frac{\overline{yx} - \bar{y} \times \bar{x}}{S_y \times S_x}.$$

Рассчитаем вспомогательные характеристики.

$\overline{yx}$ — среднее арифметическое значение произведения факторного и результативного признаков:

$$\overline{yx} = \frac{\sum_{i=1}^{n} y_i \times x_i}{n} = \frac{122\,060}{20} = 6103;$$

$\bar{y}$ — среднее значение зависимой переменной:

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{513}{20} = 25{,}65;$$

$\bar{x}$ — среднее значение независимой переменной:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{4400}{20} = 220;$$

S_y — выборочное среднеквадратическое отклонение зависимой переменной y.

Этот показатель характеризует, на сколько единиц в среднем отклоняются значения зависимого признака y от его среднего значения $\bar{y}$.

Он вычисляется по формуле:

$$S_y = \sqrt{\overline{y^2} - \overline{y}^2},$$

где $\overline{y^2}$ — среднее значение из квадратов значений результативной переменной:

$$\overline{y^2} = \frac{\sum_{i=1}^{n} y_i^2}{n} = \frac{14\,431}{20} = 721{,}55;$$

$\overline{y}^2$ — квадрат средних значений результативной переменной:

$$\overline{y}^2 = \left(\frac{\sum_{i=1}^{n} y_i}{n} \right)^2 = (25{,}65)^2 = 657{,}92.$$

Тогда

$$S_y = \sqrt{\overline{y^2} - \overline{y}^2} = \sqrt{721{,}55 - 657{,}92} = 7{,}97.$$

S_x — выборочное среднеквадратическое отклонение независимой переменной *x*. Этот показатель характеризует, на сколько единиц в среднем отклоняются значения независимого признака от его среднего значения $\overline{x}$. Он вычисляется по формуле:

$$S_x = \sqrt{\overline{x^2} - \overline{x}^2},$$

где

$$\overline{x^2} = \frac{\sum_{i=1}^{n} x_i^2}{n} = \frac{1\,059\,400}{20} = 52\,970;$$

$$\overline{x}^2 = \left(\frac{\sum_{i=1}^{n} x_i}{n} \right)^2 = (220)^2 = 48\,400.$$

Тогда

$$S_x = \sqrt{\overline{x^2} - \overline{x}^2} = \sqrt{52\,970 - 48\,400} = 67{,}6.$$

Выборочный парный **коэффициент корреляции будет равен**:

$$r_{yx} = \frac{\overline{yx} - \overline{y}\,\overline{x}}{S_y S_x} = \frac{6103 - 25{,}65 \times 220}{7{,}97 \times 67{,}6} = \frac{460}{538{,}77} = 0{,}85.$$

На следующем этапе перед построением уравнения регрессии необходимо проверить значимость полученного коэффициента корреляции с помощью t-критерия Стьюдента.

Выдвигается гипотеза H_0 о незначимости парного **коэффициента корреляции**:

$$H_0 \;/\; r_{yx} = 0.$$

Альтернативной (или обратной) выдвигается гипотеза о **значимости парного коэффициента корреляции**:

$$H_1 \;/\; r_{yx} \neq 0.$$

Значение t-критерия Стьюдента для проверки гипотезы $H_0 \;/\; r_{yx} = 0.$ в случае парной линейной регрессионной модели рассчитывается по формуле:

$$t_{набл} = \frac{r_{yx}}{\sqrt{1 - r_{yx}^2}} \times (n-2).$$

Таким образом,

$$t_{набл} = \frac{0{,}85}{\sqrt{1 - 0{,}85^2}} \times (20-2) = 29{,}08.$$

Критическое значение t-критерия $t_{крит}(\alpha;\, n - h)$, где α — уровень значимости, $(n - h)$ — число степеней свободы, определяется по таблице распределений t-критерия Стьюдента.

В данном случае $t_{крит}(\alpha;\, n - h) = t_{крит}(0{,}05;\, 20 - 2) = 1{,}73$.

Получаем, что наблюдаемое значение t-критерия по модулю больше его критического значения, т. е. $|t_{набл}| > t_{крит}$. Основная гипотеза отклоняется, и парный коэффициент корреляции признается значимым. Построение линейного уравнения регрессии по исходным данным является обоснованным.

Запишем **уравнение парной регрессии** в виде:

$$y = \bar{y} + \beta_{yx}\left(x - \bar{x}\right),$$

где β_{yx} — выборочный коэффициент регрессии y по x.

Он характеризует, насколько в среднем изменится результативный показатель y при изменении факторного показателя x на единицу своего измерения. Вычисляется выборочный коэффициент регрессии y по x с помощью следующей формулы:

$$\beta_{yx} = r_{yx} \times \frac{S_y}{S_x}.$$

Рассчитаем выборочный коэффициент регрессии y по x на основе имеющихся данных:

$$\beta_{yx} = r_{yx} \times \frac{S_y}{S_x} = 0,85 \times \frac{7,97}{67,6} = 0,1.$$

Уравнение регрессии будет иметь вид:

$$y = 25,65 + 0,1 \times (x - 220).$$

Экономическая интерпретация данного уравнения выглядит так: если уставной капитал банка изменится на 1 денежную единицу, тогда прибыль в среднем изменится на 0,1 денежную единицу.

2. Пример проверки гипотезы о значимости коэффициентов парной регрессии и уравнения регрессии в целом

На основании исходных данных по двадцати банкам страны о размере прибыли в денежных единицах (результативная переменная) и объемах выданных кредитов в денежных единицах (факторная переменная) было построено уравнение парной регрессии вида:

$$y = 25,65 + 0,1 \times (x - 220).$$

При проверке значимости (предположения того, что параметры отличаются от нуля) коэффициента регрессии выдвигается основная гипотеза H_0 о незначимости полученной оценки: $H_0 / \beta_1 = 0$.

Альтернативной (или обратной) выдвигается гипотеза о **значимости коэффициента регрессии**: $H_1 / \beta_1 \neq 0$.

Для проверки выдвинутых гипотез используется t-критерий (t-статистика) Стьюдента.

Формула наблюдаемого значения t-критерия Стьюдента для проверки гипотезы $H_0 / \beta_1 = 0$ имеет вид:

$$t_{набл} = \frac{\tilde{\beta}_1}{\omega(\beta_1)},$$

где $\tilde{\beta}_1$ — оценка параметра регрессии β_1;

$\omega(\beta_1)$ — величина стандартной ошибки параметра регрессии β_1.

В случае парной **линейной модели регрессии** показатель $\omega(\beta_1)$ вычисляется таким образом:

$$\omega(\beta_1) = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{(n-2) \times \sum_{i=1}^{n} (x_i - \bar{x})^2}}.$$

Числитель стандартной ошибки может быть рассчитан через парный коэффициент детерминации как:

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(y_i - \tilde{y}_i\right)^2 = n \times G^2(y) \times (1 - r_{yx}^2),$$

где $G^2(y)$ — общая дисперсия зависимого признака;

r_{yx}^2 — парный коэффициент детерминации между зависимым и независимым признаками.

Рассчитаем общую дисперсию результативного признака по исходным данным:

$$G^2(y) = \overline{y^2} - \bar{y}^2 = 721{,}55 - 657{,}92 = 63{,}63.$$

Тогда стандартная ошибка будет равна:

$$\omega(\beta_1) = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{(n-2) \times \sum_{i=1}^{n} (x_i - \bar{x})^2}} = \sqrt{\frac{20 \times 63{,}63 \times (1 - 0{,}85)}{(20-2) \times 170}} =$$

$$= \sqrt{\frac{190{,}89}{3060}} = 0{,}053 \ .$$

Рассчитаем наблюдаемое значение t-критерия:

$$t_{набл} = \frac{\tilde{\beta}_1}{\omega(\beta_1)} = \frac{0,1}{0,053} = 1,88.$$

Критическое значение t-критерия $t_{крит}(\alpha; n - h)$, где α — уровень значимости, $(n - h)$ — число степеней свободы, определяется по таблице распределений t-критерия Стьюдента.

В данном случае $t_{крит}(\alpha; n - h) = t_{крит}(0,05; 20 - 2) = 1,73$.

Наблюдаемое значение t-критерия по модулю больше его критического значения, т. е. $|t_{набл}| > t_{крит}$. Таким образом, коэффициент парной регрессии оказался значимым.

Проверим значимость уравнения регрессии через проверку гипотезы о значимости парного коэффициента детерминации.

Основная гипотеза формулируется как $H_0 / r^2_{yx} = 0$ — парный коэффициент детерминации незначим, и, следовательно, уравнение регрессии также является незначимым.

Альтернативная ей гипотеза $H_1 / r^2_{yx} \neq 0$ — парный коэффициент детерминации значимо отличается от нуля, следовательно, построенное уравнение регрессии является значимым.

Рассчитаем коэффициент детерминации как квадрат парного коэффициента корреляции: $r^2_{yx} = 0,85^2 = 0,7225$.

Для проверки гипотезы о значимости уравнения регрессии в целом используется F-критерий Фишера.

Критическое значение F-критерия находится по таблице распределения Фишера — Снедекора в зависимости от уровня значимости α и числа степеней свободы: $k_1 = h - 1$ и $k_2 = n - h$. В случае проверки значимости уравнения парной регрессии критическое значение F-статистики вычисляется как $(\alpha; 1; n - 2)$. В нашем примере

$$F_{крит}(\alpha; 1; n-2) = F_{крит}(0,05; 1; 18) = 4,41.$$

Формула наблюдаемого значения F-критерия для проверки гипотезы о незначимости парного **уравнения регрессии имеет вид**:

$$F_{набл} = \frac{r^2_{yx}}{1-r^2_{yx}} \times (n-2) = \frac{0,7225}{1-0,7225} \times 18 = 46,84.$$

Наблюдаемое значение F-критерия оказалось больше его критического значения, следовательно, линейное уравнение парной регрессии является значимым.

Построенное уравнение регрессии между получаемой прибылью и объемом выдаваемых кредитов на 72,25% объясняет вариацию зависимой переменной в общем объеме ее вариации. 27,75% дисперсии зависимой переменной остались необъясненными.

ЛЕКЦИЯ № 7. Линейная модель множественной регрессии. Классический метод наименьших квадратов для модели множественной регрессии. Множественное линейное уравнение регрессии

Модель множественной регрессии является методом выявления аналитической формы связи между зависимым (или результативным) признаком и несколькими независимыми (или факторными) переменными. Ее построение целесообразно в том случае, если коэффициент множественной корреляции показал наличие связи между переменными.

Общий вид **линейного уравнения множественной регрессии**:

$$y_i = \beta_0 + \beta_1 x_{1k} + \ \dots \ + \beta_n x_{ik} + \varepsilon$$

где y_i — значение i-ой зависимой переменной, $i = \overline{1,n}$;

x_{1k}, ..., x_{ik} — значения независимых переменных;

β_0, ..., β_n — параметры уравнения регрессии, подлежащие оценке;

ε_t — случайные ошибки множественного уравнения регрессии.

Модель нормальной линейной множественной регрессии строится исходя из следующих предпосылок:

1) величины x_{1i}, ..., x_{ki} являются неслучайными и независимыми переменными;

2) математическое ожидание случайной ошибки уравнения регрессии равно нулю во всех наблюдениях: $E(\varepsilon_t) = 0$, где $i = \overline{1,n}$;

3) дисперсия случайной ошибки уравнения регрессии является постоянной для всех наблюдений: $D(\varepsilon_t) = E(\varepsilon^2_t) = \text{const}$;

4) случайные ошибки уравнения регрессии не коррелированы между собой, т. е. ковариация случайных ошибок любых двух разных наблюдений равна нулю: $Cov(\varepsilon_t, \varepsilon_j) = E(\varepsilon_t \varepsilon_j) = 0$. Это предположение верно в том случае, если изучаемые данные не являются временными рядами;

5) основываясь на 3 и 4-м предположениях, добавляется условие о том, что случайная ошибка уравнения регрессии является случайной величиной, подчиняющейся нормальному закону распределения с нулевым математическим ожиданием и дисперсией $G^2 / \varepsilon_t \sim N(0, G^2)$.

Уравнение множественной линейной регрессии в матричном виде:

$$Y = X\beta + \varepsilon,$$

где $Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$ — вектор значений зависимой переменной размерности $n \times 1$;

$X = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1k} \\ 1 & x_{21} & x_{22} & \dots & x_{2k} \\ \dots & \dots & \dots & \dots & \dots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nk} \end{pmatrix}$ — вектор значений независимой переменной размерности $n \times (k + 1)$. Первый столбец является единичным, так как в уравнении регрессии параметр β_0 умножается на 1.

$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}$ — вектор неизвестных параметров модели множественной регрессии размерности $(k + 1) \times 1$;

$\varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$ — вектор случайных ошибок уравнения регрессии размерности $n \times 1$.

Добавление в модель такого компонента, как вектор случайных ошибок, необходимо в связи с практической невозможностью оценить связь между переменными со стопроцентной точностью.

Нормальная линейная модель множественной регрессии в матричной форме строится исходя из **следующих предположений**:

1) факторные признаки x_{1k}, …, x_{ik} являются детерминированными неслучайными величинами. В терминах матричной записи X — это детерминированная матрица ранга $(k + 1)$, т. е. столбцы матрицы X линейно независимы между собой;

2) математическое ожидание случайной ошибки уравнения регрессии равно нулю во всех наблюдениях: $E(\varepsilon) = 0$;

3) предположения о том, что дисперсия случайной ошибки уравнения регрессии является постоянной для всех

наблюдений и ковариация случайных ошибок любых двух разных наблюдений равна нулю, можно записать с помощью ковариационной матрицы случайных ошибок нормальной линейной модели множественной регрессии:

$$\Sigma_{\varepsilon} = \begin{pmatrix} G^2 & 0 & \cdots & 0 \\ 0 & G^2 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & G^2 \end{pmatrix} = G^2 \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & 1 \end{pmatrix} = G^2 \mathrm{I}n,$$

где G^2 — дисперсия случайной ошибки уравнения регрессии ε;

$\mathrm{I}n$ — единичная матрица размерности $n \times n$;

4) ε — независимая и не зависящая от X случайная величина, подчиняющаяся многомерному нормальному закону распределения с нулевым математическим ожиданием и дисперсией G^2: $\varepsilon \to N(0;\ G^2\mathrm{I}n)$.

1. Классический метод наименьших квадратов для модели множественной регрессии

Общий вид линейного уравнения множественной регрессии:

$$y_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_n x_{ki} + \varepsilon_i,$$

где y_i — значение i-ой зависимой переменной, $i = \overline{1,n}$;

x_{1i}, ..., x_{ki} значения независимых переменных;

β_0, ..., β_n — параметры уравнения регрессии, подлежащие оценке;

ε_i — случайные ошибки множественного уравнения регрессии.

Чтобы найти оценки неизвестных параметров линейного уравнения множественной регрессии, используется обычный метод наименьших квадратов. Его суть состоит в нахождении вектора оценки β, который минимизировал бы сумму квадратов отклонений (остатков) наблюдаемых значений зависимой переменной y от модельных значений $\tilde{y}$. рассчитанных на основании построенного уравнения регрессии.

Рассмотрим матричную форму функционала F метода наименьших квадратов:

$$F = \sum_{i=1}^{n} \left(y_i - \tilde{y}_i\right) = (Y - X\tilde{\beta})^T \times (Y - X\tilde{\beta}) \to \min,$$

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$ — вектор значений зависимой переменной размерности $n \times 1$;

$$X = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1k} \\ 1 & x_{21} & x_{22} & \dots & x_{2k} \\ \dots & \dots & \dots & \dots & \dots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nk} \end{pmatrix}$$ — вектор значений независимой переменной размерности $n \times (k+1)$.

Первый столбец является единичным, так как в уравнении регрессии параметр β_0 умножается на 1.

Для того чтобы найти минимум функции (F), нужно вычислить частные производные этой функции по каждому из оцениваемых параметров и приравнять их к нулю. Полученная стационарная система уравнений может быть записана как:

$$\begin{cases} \dfrac{\partial F}{\partial \tilde{\beta}_0} = 0 \\ \dfrac{\partial F}{\partial \tilde{\beta}_1} = 0, \\ \vdots \end{cases}$$

где $$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}$$ — вектор оцениваемых параметров уравнения регрессии.

Общий вид стационарной системы уравнений можно записать как:

$$\frac{\partial F}{\partial \tilde{\beta}} = -2X^T Y + 2X^T X\tilde{\beta} = 0.$$

В результате решения системы нормальных уравнений получим следующие МНК-оценки неизвестных параметров уравнения регрессии:

$$\tilde{\beta} = (X^T X)^{-1} X^T Y.$$

Рассмотрим **применение метода наименьших квадратов на примере модели множественной линейной регрессии** с двумя переменными:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i,$$

где $i = \overline{1,n}$.

Для нахождения оценок неизвестных параметров данного уравнения регрессии минимизируем выражение:

$$F = \sum_{i=1}^{n} \left(y_i - \tilde{\beta}_0 - \tilde{\beta}_1 x_{1i} - \tilde{\beta}_2 x_{2i} \right) \xrightarrow{\beta_0, \beta_1, \beta_2} \min.$$

Стационарная система уравнений для модели множественной линейной регрессии с двумя переменными строится следующим образом:

$$\begin{cases} \dfrac{\partial F}{\partial \tilde{\beta}_0} = -2X^T Y + 2X^T X \beta_0, \\ \dfrac{\partial F}{\partial \tilde{\beta}_1} = -2X^T Y + 2X^T X \beta_1, \\ \dfrac{\partial F}{\partial \tilde{\beta}_2} = -2X^T Y + 2X^T X \beta_2. \end{cases}$$

После элементарных преобразований данной стационарной системы уравнений получим **систему нормальных уравнений**:

$$\begin{cases} n \times \tilde{\beta}_0 + \tilde{\beta}_1 \sum\limits_{i=1}^{n} x_{1i} + \tilde{\beta}_2 \sum\limits_{i=1}^{n} x_{2i} = \sum\limits_{i=1}^{n} y_i, \\ \tilde{\beta}_0 \sum\limits_{i=1}^{n} x_{1i} + \tilde{\beta}_1 \sum\limits_{i=1}^{n} x_{1i}^2 + \tilde{\beta}_2 \sum\limits_{i=1}^{n} x_{1i} \times x_{2i} = \sum\limits_{i=1}^{n} y_i \times x_{1i}, \\ \tilde{\beta}_0 \sum\limits_{i=1}^{n} x_{2i} + \tilde{\beta}_1 \sum\limits_{i=1}^{n} x_{1i} \times x_{2i} + \tilde{\beta}_2 \sum\limits_{i=1}^{n} x_{1i}^2 = \sum\limits_{i=1}^{n} y_i \times x_{2i}. \end{cases}$$

Данная система называется **системой нормальных уравнений относительно коэффициентов** $\tilde{\beta}_0$, $\tilde{\beta}_1$ и $\tilde{\beta}_2$ для зависимости

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i.$$

Система нормальных уравнений является квадратной, т. е. количество уравнений равняется количеству неизвестных переменных, поэтому коэффициенты $\tilde{\beta}_0, \tilde{\beta}_1$ и $\tilde{\beta}_2$ можно найти с помощью метода Крамера или метода Гаусса.

Метод Крамера заключается в следующем.

Единственное решение квадратной системы линейных уравнений **определяется по формуле**:

$$K_j = \frac{\Delta_j}{\Delta},\ j = \overline{1,n},$$

где Δ — основной определитель квадратной системы линейных уравнений;

Δ_j — определитель, полученный из основного определителя путем замены *j*-го столбца на столбец свободных членов.

Если основной определитель системы Δ равен нулю и все определители Δ_j также равны нулю, то данная система имеет бесконечное множество решений.

Если основной определитель системы Δ равен нулю и хотя бы один из определителей Δ_j также равен нулю, то система решений не имеет.

Метод Гаусса применяется в основном для решения систем линейных уравнений, когда количество неизвестных параметров не совпадает с количеством уравнений.

Однако его используют и для решения квадратных систем линейных уравнений.

2. Множественное линейное уравнение регрессии в стандартизированном масштабе. Решение квадратных систем линейных уравнений методом Гаусса

Оценки неизвестных параметров уравнения регрессии определяются с помощью метода наименьших квадратов. Однако существует и другой способ оценивания этих коэффициентов в случае множественной линейной регрессии. Для этого строится уравнение множественной регрессии в стандартизированном (нормированном) масштабе. Это означает, что все переменные, участвующие в регрессионной модели, стандартизируются с помощью специальных формул.

Процесс стандартизации позволяет установить точкой отсчета для каждой нормированной переменной ее среднее значение по выборке. При этом единицей измерения стандартизированной переменной становится ее среднеквадратическое отклонение.

Формула для перевода независимой переменной x в стандартизированный масштаб:

$$t\left(x_{ij}\right)=\frac{x_{ij}-\overline{x}_i}{G(x_i)}$$

где $i=\overline{1,n},\ j=\overline{1,k};$

$G(x_i)$ — среднеквадратическое отклонение независимой переменной.

Формула для перевода зависимой переменной y в стандартизированный масштаб:

$$t\left(y_i\right)=\frac{y_i-\overline{y}}{G\left(y\right)}.$$

В случае линейной зависимости между изучаемыми переменными процесс стандартизации не нарушает этой связи, поэтому справедливо следующее равенство:

$$\tilde{t}\left(y\right)=\sum_{i=1}^{n}\beta_i\times L\left(x_i\right).$$

Для того чтобы найти неизвестные коэффициенты данной функции, можно использовать классический метод наименьших квадратов для множественной регрессии, т. е. необходимо минимизировать **функционал вида**:

$$F=\left(\tilde{t}\left(y\right)-\sum_{i=1}^{n}\beta_i\times S\left(x_i\right)\right)\xrightarrow{\beta}\min.$$

При этом в качестве переменных в системе нормальных уравнений будут выступать парные коэффициенты корреляции. Такой подход основывается на следующем равенстве:

$$\sum_{j=1}^{m}t\left(x_{ij}\right)\times\ t\left(x_{kj}\right)=r_{L_iL_k}=r_{x_ix_k}.$$

Таким образом, система нормальных уравнений для стандартизированной модели множественной регрессии имеет вид:

$$\begin{cases} \tilde{\beta}_1 + r(x_1x_2)\tilde{\beta}_2 + \ldots + r(x_1x_n)\tilde{\beta}_n = r(x_1y), \\ r(x_2x_1)\tilde{\beta}_1 + \tilde{\beta}_2 + \ldots + r(x_2x_n)\tilde{\beta}_n = r(x_2y), \\ \vdots \\ r(x_nx_1)\tilde{\beta}_1 + r(x_nx_2)\tilde{\beta}_2 + \ldots + \tilde{\beta}_n = r(x_ny). \end{cases}$$

Данная система нормальных уравнений является квадратной, т. е. количество уравнений равняется количеству неизвестных переменных, поэтому оценки коэффициентов $\tilde{\beta}_0, \ldots, \tilde{\beta}_n$ можно найти с помощью метода Крамера, метода Гаусса или метода обратных матриц.

После того как параметры уравнения множественной регрессии в стандартизированном масштабе определены, необходимо перевести их в **масштаб исходных данных**:

$$\beta_i = \tilde{\beta}_i \times \frac{G(y)}{G(x)};$$

$$\beta_0 = \overline{y} - \sum_{i=1}^{n} \beta_i \times \overline{x}_i.$$

Основная идея решения квадратной системы линейных уравнений методом Гаусса заключается в том, что исходную квадратную систему из n линейных уравнений с n неизвестными переменными необходимо преобразовать к треугольному виду. С этой целью в одном из уравнений системы оставляют все неизвестные переменные. В другом уравнении сокращают одну из неизвестных переменных для того, чтобы число неизвестных стало $(n - 1)$.

В следующем уравнении сокращают две неизвестные переменные, чтобы число переменных стало $(n - 2)$. В конце данного процесса система примет треугольный вид, первое уравнение которой содержит все неизвестные, а последнее — только одну. В последнем уравнении системы остается $(n - (n - 1))$ неизвестных переменных, т. е. одна неизвестная переменная, которая называется базисной. Дальнейшее решение сводится к выражению свободных $(n - 1)$ неизвестных переменных через базисную переменную и получению общего решения квадратной системы линейных уравнений.

ЛЕКЦИЯ № 8. Показатели тесноты связи, частной и множественной корреляции. Обычный и скорректированный показатели множественной детерминации

Соизмеримые показатели тесноты связи могут использоваться в случае, если факторные признаки имеют несопоставимые единицы измерения. Они позволяют определить тесноту связи между факторным и результативным признаками в модели множественной регрессии. К ним относятся **коэффициенты частной эластичности** и **стандартизированные частные коэффициенты регрессии.**

Для определения стандартизированных частных регрессионных коэффициентов строится уравнение множественной регрессии в стандартном масштабе. Все переменные, участвующие в регрессионной модели, стандартизируются с помощью специальных формул. Стандартизация позволяет установить точкой отсчета для каждой нормированной переменной ее среднее значение по выборке. Единицей измерения стандартизированной переменной становится ее среднеквадратическое отклонение. Их трактовка сводится к следующему: на какую долю среднеквадратического отклонения $G(y)$ изменится зависимый признак при условии изменения факторного признака $G(x)$ на величину своего среднеквадратического отклонения, если остальные факторы, участвующие в модели, будут зафиксированы.

Стандартизированный частный регрессионный коэффициент показывает степень непосредственной или прямой связи между результативным и факторным признаками. В модели множественной регрессии факторный признак оказывает на результативную переменную не только прямое, но и косвенное влияние, которое объясняется его связью с другими факторными модельными признаками. Для измерения косвенного влияния факторного признака на результативную переменную рассчитывается величина частного **коэффициента детерминации**:

$$d_i = \sum \beta_i \times r\left(x_i x_j\right),$$

где β_i — стандартизированный частный регрессионный коэффициент;

$r(x_i x_j)$ — коэффициент частной корреляции между факторными признаками x_i и x_j.

Частный коэффициент детерминации показывает, на сколько процентов вариация результативного признака объясняется вариацией *i*-го факторного признака, входящего в множественное уравнение регрессии, при фиксированных значениях остальных факторов.

Другим показателем тесноты связи является коэффициент частной эластичности. Он рассчитывается по формуле:

$$Э_i = \frac{\partial Y}{\partial X} \times \frac{\bar{X}_i}{\bar{Y}},$$

где $\bar{X}_i$ — среднее значение независимого признака по выборке $i = \overline{1,n}$;

$\bar{Y}$ — среднее значение результативного признака по выборке;

$\frac{\partial Y}{\partial X}$ — первая производная *y* по *x*.

Частный коэффициент эластичности отражает процентное изменение результативного признака при изменении на 1% от среднего уровня факторного признака x_i, если остальные переменные, участвующие в модели, зафиксированы.

Для линейной модели регрессии частный коэффициент эластичности рассчитывается:

$$Э_i = \beta_i \times \frac{\bar{X}_i}{\bar{Y}},$$

где β_i — коэффициент уравнения множественной регрессии.

Стандартизированные частные регрессионные коэффициенты и частные коэффициенты эластичности могут не совпадать по результатам. Это расхождение в выводах можно объяснить, например, тем, что величина среднеквадратического отклонения одного из факторных признаков слишком велика. Другой причиной расхождения может быть эффект неоднозначного воздействия одного из признаков на результативный показатель.

1. Показатели частной корреляции для модели линейной регрессии с двумя переменными

Применение частных коэффициентов корреляции вызвано необходимостью оценить взаимосвязь между результативным признаком и одним из факторных при условии фиксированности остальных переменных, участвующих в модели. Частный коэффициент корреляции позволяет элиминировать влияние на результат всех факторных модельных признаков, кроме одного.

Определим частные коэффициенты корреляции на примере модели линейной регрессии с двумя переменными. **Общий вид модели**:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \varepsilon_i,$$

где y_i — зависимая переменная, $i = \overline{1,n}$;

x_i — первый факторный признак;

z_i — второй факторный признак;

$\beta_0, \beta_1, \beta_2$ — неизвестные коэффициенты уравнения регрессии;

ε_i — случайная ошибка уравнения регрессии.

Определим взаимосвязь между результативным y_i и первым факторным признаком x_i при фиксированном значении второго факторного признака z_i, и наоборот, определим взаимосвязь между результативным и вторым факторным признаком при фиксированном значении первого факторного признака. Коэффициенты частной корреляции называются коэффициентами первого порядка, так как элиминируется влияние только одного фактора. Порядок частного коэффициента корреляции определяется количеством параметров, влияние которых устраняется. Порядок коэффициента парной корреляции в случае парной регрессионной модели равен нулю.

Коэффициент частной корреляции между y_i и x_i при фиксированном z_i:

$$r_{yx/z} = \frac{r_{yx} - r_{yz} \times r_{xz}}{\sqrt{\left(1 - r_{xz}^2\right) \times \left(1 - r_{yz}^2\right)}}.$$

Расчет ведется через обычные парные коэффициенты корреляции.

Рассчитаем коэффициент частной корреляции между y_i и z_i при фиксированном x_i:

$$r_{yz/x} = \frac{r_{yz} - r_{yx} \times r_{xz}}{\sqrt{\left(1 - r_{xz}^2\right) \times \left(1 - r_{yx}^2\right)}}.$$

Факторные признаки оказывают определенное влияние друг на друга. С помощью частного коэффициента корреляции можно оценить эту взаимосвязь при фиксированном признаке y_i:

$$r_{xz/y} = \frac{r_{xz} - r_{yx} \times r_{yz}}{\sqrt{\left(1 - r_{yz}^2\right) \times \left(1 - r_{yx}^2\right)}}.$$

Частные коэффициенты корреляции рассчитываются через коэффициент множественной детерминации, например коэффициент частной корреляции между y_i и x_i при фиксированном z_i:

$$r_{yx/z} = \sqrt{1 - \frac{1 - R_y^2}{1 - r_{yx}^2}} = \sqrt{\frac{R_y^2 - r_{yx}^2}{1 - r_{yx}}},$$

где R_y^2 — множественный коэффициент детерминации регрессионной модели с двумя переменными.

Коэффициент корреляции изменяется в пределах [0;1] в отличие от частных коэффициентов корреляции, рассчитанных через парную корреляцию, изменяющихся в пределах $[-1;+1)$.

На основании частных коэффициентов корреляции можно сделать вывод об обоснованности включения переменной в регрессионную модель. Если его значение мало или коэффициент незначим, следовательно, связь между данным фактором и результативной переменной либо очень слаба, либо вовсе отсутствует, поэтому фактор можно исключить из модели без ущерба для ее качества.

Значимость частных коэффициентов корреляции проверяют с помощью t-критерия Стьюдента. Критическое значение t-критерия $t_{крит}(\alpha; n - h)$ находится по таблице распределения Стьюдента, где α — уровень значимости, $n - h$ — число степеней свободы. Для модели множественной регрессии с двумя переменными число степеней свободы равняется $n - 3$.

Значение t-критерия рассчитывается по формуле (на примере частного коэффициента корреляции между y_i и x_i при фиксиро-

ванном z_i):

$$t_{набл} = \frac{r_{yx/z}}{\sqrt{1-r^2_{yx/z}}} \times \sqrt{n-k-2},$$

где k — порядок частного коэффициента корреляции (в случае модели регрессии с двумя переменными $k = 1$).

2. Показатели частной корреляции для модели множественной регрессии с тремя и более факторами

В случае модели множественной регрессии с тремя факторами можно рассчитать частные коэффициенты корреляции как первого, так и второго порядка. Выявляется взаимосвязь между результативной переменной и одним из факторов при фиксированных значениях двух других факторов.

Коэффициенты частной корреляции второго порядка для модели множественной регрессии вида:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \varepsilon_i$$

строятся так:

$$r\left(yx_1/x_2x_3\right) = \frac{r\left(yx_1/x_2\right) - r\left(yx_3/x_2\right) \times r\left(x_1x_3/x_2\right)}{\sqrt{\left(1-r^2\left(x_1x_3/x_2\right)\right) \times \left(1-r^2\left(yx_3/x_2\right)\right)}};$$

$$r\left(yx_2/x_1x_3\right) = \frac{r\left(yx_2/x_1\right) - r\left(yx_3/x_1\right) \times r\left(x_2x_3/x_1\right)}{\sqrt{\left(1-r^2\left(x_2x_3/x_1\right)\right) \times \left(1-r^2\left(yx_3/x_1\right)\right)}};$$

$$r\left(yx_3/x_1x_2\right) = \frac{r\left(yx_3/x_1\right) - r\left(yx_2/x_1\right) \times r\left(x_3x_2/x_1\right)}{\sqrt{\left(1-r^2\left(x_3x_2/x_1\right)\right) \times \left(1-r^2\left(yx_2/x_1\right)\right)}}.$$

Частные коэффициенты корреляции второго порядка построены через частные коэффициенты корреляции первого порядка.

Частный коэффициент корреляции порядка t может быть построен через частный коэффициент корреляции $(t - 1)$ порядка. Формулы, построенные через указанную взаимосвязь, называются **рекуррентными**.

В случае модели множественной регрессии, содержащей n факторных признаков, частный коэффициент $(n - 1)$ порядка

можно рассчитать по общей формуле:

$$r\left(yx_i/x_1,\ \ldots,\ x_n\right)=$$

$$=\frac{r\left(yx_i/x_1,\ \ldots,\ x_{n-1}\right)-r\left(yx_n/x_1,\ \ldots,\ x_{n-1}\right)\times r\left(x_ix_n/x_1,\ \ldots,\ x_{n-1}\right)}{\sqrt{\left(1-r^2\left(x_ix_n/x_1,\ \ldots,\ x_{n-1}\right)\right)}\times\sqrt{\left(1-r^2\left(yx_n/x_1,\ \ldots,\ x_{n-1}\right)\right)}}.$$

Рассмотрим построение частных коэффициентов корреляции через показатель остаточной дисперсии.

Для модели парной линейной регрессии остаточная дисперсия вычисляется как:

$$\delta^2\left(y,\ x_1\right)=\frac{\sum\left(y_i-\tilde{y}_i\left(x_1\right)\right)^2}{n},$$

где $\tilde{y}_i\left(x_1\right)$ — оценка уравнения парной регрессии с независимым фактором x_1.

Если в исходное уравнение парной регрессии добавить новый фактор x_2, остаточная дисперсия модели регрессии с двумя **факторными признаками** будет равна величине:

$$\delta^2\left(y,\ x_1,\ x_2\right)=\frac{\sum\left(y_i-\tilde{y}_i\left(x_1x_2\right)\right)^2}{n}$$

где $\tilde{y}_i\left(x_1x_2\right)$ — оценка уравнения регрессии с двумя независимыми факторами x_1 и x_2.

При любом качестве построенной модели двухфакторной регрессии будет выполняться неравенство: $\delta^2\left(y,\ x_1\right)\geq\delta^2\left(y,\ x_1,\ x_2\right)$.

Тогда величина

$$\frac{\delta^2\left(y,\ x_1\right)\ -\ \delta^2\left(y,\ x_1,\ x_2\right)}{\delta^2\left(y,\ x_1\right)}$$

будет означать долю сокращения остаточной дисперсии за счет включения в модель фактора x_2. Чем больше эта доля, тем сильнее дополнительный фактор x_2 влияет на результативный признак y, на качество модели регрессии в целом, тем, следовательно, сильнее связь между x_2 и y при фиксированном значении x_1.

Частный коэффициент корреляции между переменными x_2 и y при фиксированном значении x_1 через остаточную дисперсию

вычисляется:

$$r\left(yx_2/x_1\right)=\sqrt{\left(\frac{\delta^2\left(y,\,x_1\right)-\delta^2\left(y,\,x_1,\,x_2\right)}{\delta^2\left(y,\,x_1\right)}\right)}.$$

Для модели множественной регрессии с n факторными признаками частный коэффициент корреляции ($n-1$) порядка между результативным признаком y и факторным признаком x_1 при фиксированном значении остальных признаков можно рассчитать по формуле:

$$r\left(yx_1/x_2,\,\ldots,\,x_2\right)=\sqrt{\frac{\delta^2\left(y,\,x_2,\,\ldots,\,x_n\right)-\delta^2\left(y,\,x_1,\,x_2,\,\ldots,\,x_n\right)}{\delta^2\left(y,\,x_2,\,\ldots,\,x_n\right)}}.$$

Остаточная регрессия результативного признака и коэффициент множественной детерминации связаны отношением:

$$\delta^2(y)=1-R^2(y).$$

Если в формуле частного коэффициента корреляции выразить остаточную дисперсию результативного признака через коэффициент множественной детерминации, то для модели множественной регрессии с n факторными признаками частный **коэффициент корреляции в общем виде можно определить по формуле**:

$$r\left(yx_i/x_1,\,\ldots,\,x_n\right)=\sqrt{1-\frac{1-R^2\left(y,\,x_i\right)}{1-R^2\left(y,\,x_{i-1}\right)}}.$$

Частные коэффициенты корреляции, вычисленные по рекуррентным формулам, изменяются в пределах $[-1;\ +1]$. Частные коэффициенты корреляции, вычисленные через остаточную дисперсию или коэффициент множественной детерминации, изменяются в пределах $[0;\ +1]$.

Частный коэффициент корреляции для модели множественной регрессии показывает степень тесноты связи между результативным признаком и одним из факторных признаков при фиксированном или постоянном значении остальных переменных, входящих в модель.

3. Показатель множественной корреляции. Обычный и скорректированный показатели множественной детерминации

Построение множественного коэффициента корреляции целесообразно только в том случае, когда частные коэффициенты

корреляции оказались значимыми и связь между результативным признаком и факторами, включенными в модель, действительно существует. Множественный коэффициент корреляции позволяет оценить общее влияние всех факторных переменных на результативный признак в модели множественной регрессии.

В случае линейной модели множественной регрессии с n факторными признаками коэффициент множественной корреляции рассчитывается через стандартизированные частные коэффициенты регрессии и парные коэффициенты корреляции следующим образом:

$$R(y, x_1, \ldots, x_n) = \sqrt{\sum_{i=1}^{n} \beta_i^{станд} \times r(yx_i)},$$

где $r(yx_i)$ — парный (не частный) коэффициент корреляции между результативным признаком y и факторным признаком x_i, $i = \overline{1,n}$.

Коэффициент множественной корреляции изменяется в пределах [0; +1] и поэтому не предназначен для определения направ-ления связи между результативным и факторными признаками. Чем ближе множественный коэффициент корреляции к единице, тем сильнее взаимосвязь между зависимой и независимыми переменными, и, наоборот, чем ближе множественный коэффициент корреляции к нулю, тем слабее взаимосвязь между изучаемыми переменными.

Если возвести множественный коэффициент корреляции в квадрат, то получим **коэффициент множественной детерминации**:

$$R^2(y, x_1, \ldots, x_i) = \sum_{i=1}^{n} \beta_i^{станд} \times r(yx_i).$$

Множественный коэффициент детерминации показывает, на сколько процентов построенная модель регрессии объясняет разброс значений зависимой переменной относительно среднего значения, т. е. какая доля общей дисперсии результативного признака объясняется вариацией факторных модельных признаков. Множественный коэффициент детерминации можно назвать количественной характеристикой, объясненной построенным уравнением регрессии дисперсии результативного признака. Чем больше значение данного показателя, тем лучше уравнение регрессии описывает выявленную взаимосвязь.

Для множественного коэффициента детерминации всегда **справедливо неравенство**:

$$R^2(y, x_1, \ldots, x_{n-1}) \le R^2(y, x_1, \ldots, x_n),$$

т. е. включение в линейную регрессионную модель нового фак-

торного признака x_n не снижает значения коэффициента множественной детерминации.

Коэффициент множественной детерминации можно рассчитать на основании теоремы о разложении сумм квадратов:

$$R^2\left(y, x_1, \ldots, x_n\right) = 1 - \frac{ESS}{TSS},$$

где *ESS* (Error Sum Square) — сумма квадратов остатков множественного уравнения регрессии с *n* переменными:

$$\sum_{i=1}^{n}\left(y_i - \tilde{y}\left(y, x_1, \ldots, x_n\right)\right)^2;$$

TSS (Total Sum Square) — общая сумма квадратов множественного уравнения регрессии:

$$\sum_{i=1}^{n}\left(y_i - \overline{y}\right)^2.$$

Влияние на качество модели дополнительно включенного в регрессионное уравнение фактора не всегда можно выявить с помощью обычного множественного коэффициента детерминации. Поэтому рассчитывают также и скорректированный коэффициент множественной детерминации, в котором учитывается количество **факторных признаков в модели**:

$$R^2_{Adj} = 1 - \frac{\frac{ESS}{(n-h)}}{\frac{TSS}{(n-1)}} = 1 - \left(1 - R^2\right) \times \frac{(n-1)}{(n-h)},$$

где *n* — количество наблюдений в выборке;

h — число параметров в регрессионной модели.

При большом объеме выборки обычный и скорректированный (adjusted) коэффициенты множественной детерминации отличаться практически не будут.

ЛЕКЦИЯ № 9. Проверка гипотез о значимости частного и множественного коэффициентов корреляции, регрессионных коэффициентов и уравнения множественной регрессии в целом

После расчета всех частных коэффициентов корреляции множественного уравнения регрессии необходимо проверить их значимость.

Выдвигается **гипотеза** H_0 о незначимости частных **коэффициентов корреляции**.

Альтернативной гипотезой является утверждение о значимости частного **коэффициента корреляции**: $H_0 / r(y x_i / x_1, ..., x_{n-1}) \neq 0$.

Гипотеза о значимости частных коэффициентов корреляции проверяется с помощью t — критерия Стьюдента.

Наблюдаемое значение t-критерия $t_{набл}$ вычисляется по формуле:

$$t_{набл} = \frac{r\left(yx_i / x_1, \ \ldots, \ x_{n-1}\right)}{1 - r\left(yx / x_1, \ \ldots, \ x_{n-1}\right)} \times \sqrt{n - l - 1},$$

где n — объем выборочной совокупности (число наблюдений);

l — число оцениваемых по выборке параметров.

Критическое значение t-критерия $t_{крит}$ находится по таблице распределения Стьюдента с уровнем значимости $\alpha/2$ и степенью свободы $(n - l - 1)$ / $t_{крит}(\alpha/2; n - l - 1)$.

Если модуль наблюдаемого значения t-критерия больше критического значения t-критерия, т. е. $|t_{набл}| > t_{крит}$, то с вероятностью α основную гипотезу о незначимости частного коэффициента корреляции отвергают, т. е. между изучаемыми признаками x_i и y существует корреляционная связь при фиксированных значениях остальных переменных, участвующих в модели.

Если модуль наблюдаемого значения t-критерия меньше или равен критическому значению t-критерия, т. е. $|t_{набл}| \leq t_{крит}$,то основная гипотеза H_0 о незначимости частного коэффициента корреляции принимается, т. е. между изучаемыми признаками x_i и y при фиксированных значениях остальных переменных, участвующих в модели, корреляционная связь отсутствует и включение данно-

го фактора в регрессионную модель в данном случае нецелесообразно.

После проверки значимости всех частных коэффициентов корреляции осуществляется проверка значимости множественного коэффициента корреляции.

Основной гипотезой H_0 является утверждение о статистической незначимости множественного коэффициента корреляции.

$$H_0 \ / \ R(y, x_i) = 0, \quad i = \overline{1, n}.$$

Обратной к основной является гипотеза H_1 о значимости множественного коэффициента корреляции, т. е. о его значимом отличии от нуля:

$$H_1 \ / \ R(y, x_i) \neq 0.$$

Проверка гипотезы о незначимости множественного коэффициента корреляции осуществляется с помощью F-критерия Фишера через коэффициент множественной детерминации.

Наблюдаемое (фактическое) значение F-критерия $F_{набл}$ вычисляется по формуле:

$$F_{набл} = \frac{R^2(y, x_i)}{1 - R^2(y, x_i)} \times \frac{n - l}{l - 1},$$

где $R^2(y, x_i)$ — коэффициент множественный детерминации.

Критическое значение F-критерия $F_{крит}$ вычисляется по таблице распределения Фишера—Снедекора в зависимости от следующих параметров: уровня значимости α и числа степеней свободы:

$$k_1 = l - 1 \text{ и } k_2 = n - l \ / \ F_{крит}(\alpha; k_1; k_2) \text{ и } k_2 = n - l \ / \ F_{крит}(\alpha; k_1; k_2).$$

Гипотезы проверяются следующим образом.

Если наблюдаемое значение F-критерия больше критического значения данного критерия, т. е. $F_{набл} > F_{крит}$, то с вероятностью α основная гипотеза о незначимости коэффициента множественной регрессии отклоняется, а он признается значимым. В этом случае построение модели множественной регрессии на основании изучаемого набора переменных является обоснованным.

Если наблюдаемое значение F-критерия меньше критического значения данного критерия, т. е. $F_{набл} < F_{крит}$, то с вероятностью $(1 - \alpha)$ основная гипотеза о незначимости коэффициента множественной корреляции принимается, а он признается незначимым.

Построение модели множественной регрессии является нецелесообразным.

Чтобы построенную модель множественной регрессии можно было использовать при изучении экономических связей между модельными переменными, необходимо проверить значимость регрессионных коэффициентов.

При проверке значимости (предположения того, что параметры значимо отличаются от нуля) коэффициентов уравнения множественной регрессии выдвигается основная гипотеза H_0 о незначимости полученных оценок:

$$H_1 \,/\, \tilde{\beta}_0 = \tilde{\beta}_1 = \ldots = \tilde{\beta}_k = 0.$$

Альтернативной (или обратной) выдвигается гипотеза о **значимости коэффициентов множественной регрессии**:

$$H_1 \,/\, \tilde{\beta}_0 \neq \tilde{\beta}_1 \neq \ldots \neq \tilde{\beta}_k \neq 0.$$

Проверка этих гипотез осуществляется с помощью t-критерия Стьюдента, который, в свою очередь, вычисляется через частный F-критерий Фишера.

Между частным F-критерием и t-критерием Стьюдента существует взаимосвязь, которая используется при проверке значимости коэффициентов модели множественной регрессии:

$$t_{набл} = \sqrt{F_{набл}}.$$

Наблюдаемое значение частного F-критерия $F_{набл}$ рассчитывается по формуле:

$$F_{набл}(x_k) = \frac{R^2(y, x_1, \ldots, x_n) - R^2(y, x_1, \ldots, x_{n-1})}{1 - R^2(y, x_1, \ldots, x_n)} \times (n - l),$$

где n — объем выборки;

l — число оцениваемых по выборке параметров.

Критическое значение t-критерия $t_{крит}(\alpha; n - l - 1)$ определяется по таблице распределения Стьюдента.

Если наблюдаемое значение t-критерия больше или равно критическому значению данного критерия, т. е. $t_{набл} \geq t_{крит}$, то коэффициент $\beta_к$ уравнения множественной регрессии является значимым.

Если наблюдаемое значение t-критерия меньше, чем критическое значение данного критерия, т. е. $t_{набл} < t_{крит}$, то коэффициент уравнения множественной регрессии является незначи-

мым. Модель множественной регрессии необходимо оценить на адек-ватность в отношении реальных данных, т. е. проверить ее значимость в целом. Проверка гипотезы о значимости множественного уравнения регрессии сводится к проверке гипотезы о значимости множественного коэффициента корреляции или значимости параметров уравнения регрессии.

В качестве основной гипотезы (о незначимости уравнения множественной регрессии) может выступать: $H_0 / r(yx_i / x_1, ..., x_{n-1}) = 0$ — гипотеза о незначимости коэффициента множественной корреляции;

Основной гипотезе противостоит альтернативная **гипотеза вида**: $H_1 / r(yx_i / x_1, ..., x_{n-1}) \neq 0$ — гипотеза о значимости коэффициента множественной корреляции.

Чаще значимость уравнения множественной регрессии проверяется через значимость коэффициента множественной корреляции с помощью F-критерия Фишера.

Наблюдаемое значение F-критерия $F_{набл}$ вычисляется по формуле:

$$F_{набл} = \frac{R^2(y, x_i)}{1 - R^2(y, x_i)} \times \frac{n - l}{l - 1},$$

где $R^2(y, x_i)$ — коэффициент множественный детерминации.

Критическое значение F-критерия $F_{крит}$ вычисляется по таблице распределения Фишера—Снедекора в зависимости от уровня значимости α и числа степеней свободы: $k_1 = l - 1$ и $n - l$.

Если наблюдаемое значение F−критерия больше критического значения данного критерия, т. е. $F_{набл} > F_{крит}$, то с вероятностью α основная гипотеза о незначимости коэффициента множественной регрессии отклоняется, а уравнение множественной регрессии является значимым.

Если наблюдаемое значение F-критерия меньше критического значения данного критерия, т. е. $F_{набл} < F_{крит}$, то с вероятностью $(1 - \alpha)$ основная гипотеза о незначимости коэффициента множественной корреляции принимается.

Уравнение множественной регрессии признается незначимым.

ЛЕКЦИЯ № 10. Пример применения МНК к трехмерной модели регрессии. Пример расчета коэффициентов корреляции и проверки гипотез для трехмерной регрессионной модели

Имеются данные по двадцати банкам страны о размере прибыли в дененежных единицах (результативная переменная *y*), объемах выданных кредитов в денежных единицах (факторная переменная *x*) и размере уставного капитала в денежных единицах (факторная переменная *z*) (табл. 3).

Таблица 3

Данные по двадцати банкам страны о размере прибыли, объемах выданных кредитов и размере уставного капитала в денежных единицах

Σ_y — прибыль, ден. ед.	Σ_x — кредиты, ден. ед.	Σ_z — *уставный капитал, ден. ед*	Σ_{yx}	Σ_{yz}	Σx^2	Σz^2	Σxz
513	4 400	576	122 060	15 098	1 059 400	16 844	129 660

Исходя из предположения о линейной зависимости между переменными составим систему нормальных уравнений для определения параметров **уравнения множественной регрессии**:

$$\begin{cases} \tilde{\beta}_0 \times 20 + \tilde{\beta}_1 \times 4400 + \tilde{\beta}_2 \times 576 = 513; \\ \tilde{\beta}_0 \times 4400 + \tilde{\beta}_1 \times 1059400 + \tilde{\beta}_2 \times 129660 = 122\,060; \\ \tilde{\beta}_0 \times 576 + \tilde{\beta}_1 \times 129\,660 + \tilde{\beta}_2 \times 16\,844 = 15\,098. \end{cases}$$

Для решения данной квадратной системы линейных уравнений используем метод Крамера.

Для этого вычислим общий определитель системы:

$$\Delta = \begin{vmatrix} 20 & 4400 & 576 \\ 4400 & 1\,059\,400 & 129\,660 \\ 576 & 129660 & 16\,844 \end{vmatrix} =$$

$$= (20 \times 1\,059\,400 \times 16\,844 + 4400 \times 129\,660 \times 576 + 4400 \times 129\,660 \times 576) -$$

$$-(576 \times 1\,059\,400 \times 576 + 129\,660 \times 129\,660 \times 20 + 4400 \times 4400 \times 16\,844) =$$

$$= 293\,633\,600.$$

Аналогично вычисляем частные определители, заменяя при этом соответствующий столбец столбцом свободных членов:

$$\Delta_1 = \begin{vmatrix} 513 & 4400 & 576 \\ 122\,060 & 1\,059\,400 & 129\,660 \\ 15\,098 & 129\,660 & 16\,844 \end{vmatrix} =$$

$$= (513 \times 1\,059\,400 \times 16\,844 + 122\,060 \times 129\,660 \times 576 + 4400 \times 129\,660 \times 1\,5098) -$$

$$-(576 \times 1\,059\,400 \times 15\,098 + 129\,660 \times 129\,660 \times 513 + 4400 \times 122\,060 \times 16\,844) =$$

$$= -105\,920\,751\,440.$$

$$\Delta_2 = \begin{vmatrix} 20 & 513 & 576 \\ 4\,400 & 122\,060 & 129\,660 \\ 576 & 15\,098 & 16\,844 \end{vmatrix} =$$

$$= (20 \times 122\,060 \times 16\,844 + 4400 \times 15\,098 \times 576 + 513 \times 129\,660 \times 576) -$$

$$-(576 \times 122\,060 \times 576 + 15\,098 \times 129\,660 \times 20 + 4400 \times 513 \times 16\,844) =$$

$$= 2\,792\,9120.$$

$$\Delta_3 = \begin{vmatrix} 20 & 4400 & 513 \\ 4\,400 & 1\,059\,400 & 122\,060 \\ 576 & 129\,660 & 15\,098 \end{vmatrix} =$$

$$= (513 \times 1\,059\,400 \times 576 + 4400 \times 129\,660 \times 513 + 122\,060 \times 4400 \times 576) -$$

$$-(20\times 1\,059\,400\times 15\,098+129\,660\times 122\,060\times 20+4400\times 4400\times 15\,09\,8)=$$

$$=1\,366\,229\,280.$$

Определим коэффициенты регрессионного уравнения по формулам:

$$\tilde{\beta}_0=\frac{\Delta_1}{\Delta}\approx -0,236;$$

$$\tilde{\beta}_1=\frac{\Delta_2}{\Delta}\approx 0,09;$$

$$\tilde{\beta}_2=\frac{\Delta_3}{\Delta}\approx 0,17.$$

Таким образом, уравнение регрессии, описывающее зависимость прибыли банка от объема выдаваемых кредитов и размера уставного капитала, выглядит следующим образом:

$$y=-0,236+0,09x-\ 0,17z.$$

Параметр регрессии $\tilde{\beta}_1$ показывает, что при изменении переменной x на 1 дененежную единицу результативная переменная изменится на 0,09 денежных единиц при фиксированном значении переменной z.

Параметр регрессии $\tilde{\beta}_2$ показывает, что при изменении переменной z на 1 денежную единицу результативная переменная изменится на 0,17 денежных единиц при фиксированном значении переменной x.

Рассчитаем по имеющимся данным уравнение регрессии в стандартизированном масштабе:

$$\tilde{t}_y=\beta_1\times t_{x_1}+\beta_2\times t_{x_2}+\ldots+\beta_p\times t_{x_p},$$

где t_y, t_{x_1}, …, t_{x_p} — стандартизованные переменные:

$$t(x_{ij})=\frac{x_{ij}-\bar{x}_i}{G(x_i)};\ t(y_i)=\frac{y_i-\bar{y}}{G(y)}.$$

Система нормальных уравнений для стандартизированной модели множественной регрессии имеет вид:

$$\begin{cases} \tilde{\beta}_1 + r(x_1x_2)\tilde{\beta}_2 + \ldots + r(x_1x_n)\tilde{\beta}_n = r(x_1y), \\ r(x_2x_1)\tilde{\beta}_1 + \tilde{\beta}_2 + \ldots + r(x_2x_n)\tilde{\beta}_n = r(x_2y), \\ \vdots \\ r(x_nx_1)\tilde{\beta}_1 + r(x_nx_2)\tilde{\beta}_2 + \ldots + \tilde{\beta}_n = r(x_ny), \end{cases}$$

где $r(x_i x_j)$, $r(x_i y)$ — парные коэффициенты корреляции между переменными:

$$r_{yx_i} = \frac{\overline{yx_i} - \bar{y}\bar{x}_i}{S_y S_{x_i}}, \quad r_{x_ix_j} = \frac{\overline{x_ix_j} - \bar{x}_i\bar{x}_j}{S_{x_i} S_{x_j}}.$$

Рассчитаем вспомогательные характеристики для определения стандартизированных коэффициентов:

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{513}{20} = 25{,}65\sqrt{b^2 - 4ac}; \quad \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{4400}{20} = 220; \quad \bar{z} = \frac{\sum_{i=1}^{n} z_i}{n} = 28{,}8;$$

$$\overline{yx} = \frac{\sum_{i=1}^{n} y_i x_i}{n} = \frac{122\,060}{20} = 6103; \quad \overline{yz} = \frac{\sum_{i=1}^{n} y_i z_i}{n} = 754{,}9; \quad \overline{xz} = \frac{\sum_{i=1}^{n} x_i z_i}{n} = 6483;$$

$$S_y = \sqrt{\overline{y^2} - \bar{y}^2} = 7{,}97; \quad S_x = \sqrt{\overline{x^2} - \bar{x}^2} = 67{,}6; \quad S_z = \sqrt{\overline{z^2} - \bar{z}^2} = 3{,}66$$

$$r_{yx} = \frac{\overline{yx} - \bar{y}\,\bar{x}}{S_y S_x} = 0{,}85; \quad r_{yz} = \frac{\overline{yz} - \bar{y}\,\bar{z}}{S_y S_z} = 0{,}57; \quad r_{xz} = \frac{\overline{xz} - \bar{x}\,\bar{z}}{S_x S_z} = 0{,}61.$$

Таким образом, **система нормальных уравнение** будет иметь вид:

$$\begin{cases} \tilde{\beta}_1 + 0{,}61\tilde{\beta}_2 = 0{,}85; \\ \tilde{\beta}_2 + 0{,}61\tilde{\beta}_1 = 0{,}57. \end{cases}$$

Отсюда найдем стандартизированные регрессионные коэффициенты:

$$\tilde{\beta}_1 = 0{,}8012;$$

$$\tilde{\beta}_2 = 0{,}08.$$

Уравнение регрессии в стандартизированном масштабе можно записать следующим образом:

$$\tilde{t}_y = 0{,}8012 \times t_x + 0{,}08 \times t_z.$$

После того как параметры уравнения множественной регрессии в стандартизированном масштабе определены, необходимо перевести их в масштаб исходных данных по формулам:

$$\beta_i = \tilde{\beta}_i \times \frac{G(y)}{G(x)}; \quad \beta_0 = \overline{y} - \sum_{i=1}^{n} \beta_i \times \overline{x}_i.$$

Таким образом:

$$\beta_1 = 0{,}8012 \times \frac{7{,}97}{67{,}6} = 0{,}09;$$

$$\beta_2 = 0{,}08 \times \frac{7{,}97}{3{,}66} = 0{,}17;$$

$$\beta_0 = 25{,}65 - 0{,}8012 \times\ 220 - 0{,}08 \times\ 28{,}8 = -152{,}9.$$

Оцененное уравнение регрессии имеет вид:

$$y = -152{,}9 + 0{,}09x + \ 0{,}17z.$$

Как видим, оценки данного уравнения практически не отличаются от оценок, полученных с помощью МНК, кроме оценки свободного члена β_0. Выяснить, какое уравнение является более точным, можно с помощью **сравнения показателей остатков регрессии**:

$$\sum_{i=1}^{n} e_i = \sum_{i=1}^{n} (y_i - \tilde{y}_i).$$

Для первого уравнения регрессии сумма остатков равна нулю, поэтому оно является наилучшим.

Пример расчета коэффициентов корреляции и проверки гипотез для трехмерной регрессионной модели

На основе данных таблицы 2 рассчитаем частные коэффициенты корреляции для модели трехмерной регрессии.

Определим коэффициент частной корреляции между получаемой прибылью и объемом выданных кредитов при фиксированной величине уставного капитала банка z по формуле:

$$r_{yx/z} = \frac{r_{yx} - r_{yz} \times r_{xz}}{\sqrt{\left(1 - r_{xz}^2\right) \times \left(1 - r_{yz}^2\right)}}.$$

В качестве вспомогательных величин рассчитаем парные **коэффициенты корреляции**:

$$r_{yx} = \frac{\overline{yx} - \bar{y} \times x}{S_y S_x} = 0{,}85; \quad r_{yz} = \frac{\overline{yz} - \overline{yz}}{S_y S_z} = 0{,}57; \quad r_{xz} = \frac{\overline{xz} - \overline{xz}}{S_x S_z} = 0{,}61.$$

Тогда:

$$r_{yx/z} = \frac{r_{yx} - r_{yz} \times r_{xz}}{\sqrt{\left(1 - r_{xz}^2\right) \times \left(1 - r_{yz}^2\right)}} = \frac{0{,}85 - 0{,}57 \times 0{,}61}{\sqrt{\left(1 - 0{,}57^2\right) \times \left(1 - 0{,}61^2\right)}} = 0{,}77.$$

Рассчитаем коэффициент частной корреляции между получаемой прибылью y и размером уставного капитала z при фиксированной величине выдаваемых кредитов x по формуле:

$$r_{yz/x} = \frac{r_{yz} - r_{yx} \times r_{xz}}{\sqrt{\left(1 - r_{xz}^2\right) \times \left(1 - r_{yx}^2\right)}} = \frac{0{,}57 - 0{,}85 \times 0{,}61}{\sqrt{\left(1 - 0{,}61^2\right) \times \left(1 - 0{,}85^2\right)}} = 0{,}12.$$

Факторные признаки оказывают определенное влияние друг на друга. С помощью частного коэффициента корреляции можно оценить эту взаимосвязь при фиксированном значении прибыли по формуле:

$$r_{xz/y} = \frac{r_{xz} - r_{yx} \times r_{yz}}{\sqrt{\left(1 - r_{yz}^2\right) \times \left(1 - r_{yx}^2\right)}} = \frac{0{,}61 - 0{,}85 \times 0{,}57}{\sqrt{\left(1 - 0{,}57^2\right) \times (1 - 0{,}85)}} = 0{,}29.$$

После расчета всех частных коэффициентов корреляции множественного уравнения регрессии необходимо проверить их значимость.

Выдвигается основная гипотеза H_0 о незначимости частных коэффициентов корреляции:

$$H_0 \ / \ r\left(yx_i / x_1, \ \ldots, \ x_{n-1}\right) = 0.$$

Альтернативной гипотезой H_1 является утверждение о значимости частного коэффициента корреляции:

$$H_1 \ / \ r\left(yx_i / x_1, \ \ldots, \ x_{n-1}\right) \neq 0.$$

Гипотеза значимости частных коэффициентов корреляции проверяется с помощью t-критерия Стьюдента.

Наблюдаемое значение t-критерия $t_{набл}$ вычисляется по формуле:

$$t_{набл} = \frac{r\left(yx_i / x_1, \ \ldots, \ x_{n-1}\right)}{1 - r\left(yx / x_1, \ \ldots, \ x_{n-1}\right)} \times \sqrt{n - l - 1},$$

где n — объем выборочной совокупности (число наблюдений);

l — число оцениваемых по выборке параметров.

Критическое значение t-критерия $t_{крит}$ находится по таблице распределения Стьюдента с уровнем значимости $\alpha/2$ и степенью свободы $(n - l - 1)$: $t_{крит}(\alpha/2; n - l - 1)$.

Проверим значимость частного коэффициента корреляции $r_{yx/z}$. Наблюдаемое значение t-критерия равно:

$$t_{набл} = \frac{r_{yx/z}}{1 - r_{yx/z}} \times \sqrt{n - l - 1} = \frac{0,77}{1 - 0,77} \times \sqrt{20 - 3} = 13,8.$$

Критическое значение t-критерия

$$t_{крит}\left(\frac{\alpha}{2}; n - l - 1\right) = t_{крит}\left(0,025; 17\right) = 2,1.$$

$|t_{набл}| > t_{крит}$, следовательно, между изучаемыми признаками x и y существует корреляционная связь при фиксированном значении переменной z.

Проверим значимость частного коэффициента корреляции $r_{yz/x}$. Наблюдаемое значение t-критерия равно:

$$t_{набл} = \frac{r_{yz/x}}{1 - r_{yz/x}} \times \sqrt{n - l - 1} = \frac{0,2}{1 - 0,12} \times \sqrt{20 - 3} = 0,56.$$

Так как $|t_{набл}| < t_{крит}$, то данный коэффициент корреляции является незначимым, и переменную уставного капитала z можно вывести из модели без потери для ее качества.

Проверим значимость частного коэффициента корреляции $r_{yz/x}$. Наблюдаемое значение t-критерия равно:

$$t_{набл} = \frac{r_{xz/y}}{1 - r_{xz/y}} \times \sqrt{n - l - 1} = \frac{0,29}{1 - 0,29} \times \sqrt{20 - 3} = 1,68.$$

Так как $|t_{набл}| < t_{крит}$, то данный коэффициент корреляции является незначимым.

Рассчитаем множественный коэффициент корреляции для трехмерной модели регрессии по формуле:

$$R\left(y, x_1, \ldots, x_n\right) = \sqrt{\sum_{i=1}^{n} \beta_i^{станд} \times r\left(yx_i\right)} = 0,85.$$

Добавление в модель новой переменной не изменило коэффициента корреляции.

Рассчитаем множественный коэффициент детерминации как квадрат множественного коэффициента корреляции:

$$R^2\left(y,\, x_1,\, \ldots,\, x_i\right) = \sum_{i=1}^{n} \beta_i^{станд} \times r\left(yx_i\right) = 0{,}73.$$

Коэффициент детерминации для парной модели регрессии, включающей в качестве факторной переменной только объем выдаваемых кредитов, составил 0,72, т. е. включение в модель новой переменной не увеличило долю объясненной дисперсии.

Рассчитаем скорректированный коэффициент детерминации:

$$R^2_{Adj} = 1 - \frac{\dfrac{ESS}{(n-h)}}{\dfrac{TSS}{(n-1)}} = 1 - \left(1 - R^2\right) \times \frac{(n-1)}{(n-h)} = 0{,}7.$$

После расчета всех коэффициентов корреляции и детерминации можно сделать окончательный вывод о том, что парная модель регрессии между переменной прибыли и объемом выдаваемых кредитов является более предпочтительной по сравнению с трехмерной моделью регрессии, так как включение в уравнение нового фактора ощутимых результатов не принесло, а лишь сделало его более сложным.

ЛЕКЦИЯ № 11. Причины возникновения и последствия мультиколлинеарности. Устранение мультиколлинеарности

Явление мультиколлинеарности в случае линейной модели множественной регрессии — это нарушение одной из ее предпосылок, т. е. наличие линейной зависимости между факторами, участвующими в модели. В матричном виде мультиколлинеарность определяется как зависимость между столбцами матрицы факторных переменных *X*. Размерность матрицы факторных признаков $X - n \times n$ (без единичного вектора). Если ранг матрицы *X* меньше *n*, то говорят о полной, или строгой, мультиколлинеарности. Однако на практике полная мультиколлинеарность почти не встречается. Проблема простой мультиколлинеарности (нестрогой) характерна для временных рядов.

Таким образом, основной причиной мультиколлинеарности является плохая матрица независимых переменных *X*.

Среди **основных последствий, к которым может привести мультиколлинеарность**, можно выделить следующие:

1) при проверке основной гипотезы о незначимости коэффициентов множественной регрессии с помощью t-критерия в большинстве случаев она принимается, однако само уравнение регрессии при проверке с помощью F-критерия оказывается значимым, что говорит о завышенной величине коэффициента множественной корреляции;

2) полученные оценки коэффициентов уравнения множественной регрессии в основном неоправданно завышены или имеют неправильные знаки;

3) добавление или исключение из исходных данных одного-двух наблюдений оказывает сильное влияние на оценки коэффициентов модели;

4) наличие мультиколлинеарности в модели множественной регрессии может сделать ее непригодной для дальнейшего применения (например, для построения прогнозов).

Для обнаружения мультиколлинеарности не существует никаких точных критериев, а применяется ряд эмпирических методов.

Обычно при этом анализируется корреляционная матрица независимых переменных R либо матрица $(X^T X)$. Множественный регрессионный анализ всегда должен начинаться с рассмотрения этой матрицы.

Корреляционной матрицей независимых переменных называется симметричная относительно главной диагонали матрица линейных парных коэффициентов корреляции независимых переменных:

$$R = \begin{bmatrix} r_{11} & r_{12} & \dots & r_{1n} \\ r_{21} & r_{22} & \dots & r_{2n} \\ \dots & \dots & \dots & \dots \\ r_{n1} & r_{n2} & \cdots & r_{nn} \end{bmatrix} = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1n} \\ r_{21} & 1 & \cdots & r_{2n} \\ \dots & \dots & \dots & \dots \\ r_{n1} & r_{n2} & \cdots & 1 \end{bmatrix},$$

где r_{ij} — коэффициент парной линейной корреляции между i-м и j-ым независимыми признаками, $i, j = 1, n.$

На диагонали корреляционной матрицы находятся единицы, так как коэффициент корреляции признака с самим собой равен единице.

Если в корреляционной матрице независимых переменных есть парный коэффициент корреляции между i-м и j-ым признаками, то в данной модели множественной регрессии существует мультиколлинеарность.

Другим способом обнаружения мультиколлинеарности является вычисление собственных чисел корреляционной матрицы λ_{min} и λ_{max}. Если $\lambda_{min} < 10^{-5}$, то в модели присутствует мультиколлинеарность. Если отношение $\lambda_{min} / \lambda_{max} < 10^{-5}$, то также делают вывод о наличии мультиколлинеарности.

Вывод о присутствие мультиколлинеарности в модели множественной регрессии можно сделать также после вычисления определителя корреляционной матрицы независимых переменных. Если его величина очень мала, то мультиколлинеарность существует.

Устранение мультиколлинеарности

Устранение проблемы мультиколлинеарности является обязательным в том случае, если построенную модель множественной регрессии предполагается использовать с целью изучения экономических связей.

При этом весьма важными являются знаки при коэффициентах уравнения регрессии и их значение.

При удовлетворительной величине ошибки прогноза данное уравнение можно использовать и при наличии мультиколлинеарности. Если же прогноз получается неудовлетворительным, то с мультиколлинеарностью нужно бороться.

Одним из самых элементарных способов устранения мультиколлинеарности является сбор дополнительных данных. Но на практике это не всегда возможно.

Другим способом является **преобразование переменных**, например вместо значений всех переменных, участвующих в модели (и результативной в том числе), можно взять их логарифмы:

$$\ln y = \beta_0 + \beta_1 \ln x_1 + \beta_2 \ln x_2 + \varepsilon.$$

Но и это не гарантирует избавления от мультиколлинеарности.

Если ни один из вышеназванных способов не помог, то необходимо использовать либо смещенные методы оценки неизвестных параметров модели, либо методы исключения переменных из модели множественной регрессии.

К смещенным методам оценки коэффициентов регрессии можно отнести **гребневую регрессию** или **ридж** (ridge). Ее используют в том случае, когда ни одну из переменных, участвующих в модели, удалить нельзя.

Суть гребневой регрессии заключается в том, что ко всем диагональным элементам матрицы (X^TX) добавляется небольшое число τ (тау): $10^{-6} < \tau < 0{,}1$. Неизвестные параметры множественного уравнения регрессии в данном случае определяются по формуле:

$$\tilde{\beta}_{ridge} = \left(X^T X + \tau \mathrm{I}n\right)^{-1} X^T Y,$$

где *In* — единичная матрица.

В результате применения риджа оценки коэффициентов уравнения множественной регрессии стабилизируются к определенному числу, и их стандартные ошибки уменьшаются.

Основным методом исключения переменных из модели является **метод главных компонент**. В этом случае от матрицы факторных переменных *X* переходят к матрице главных компонент *F*, и уже на ее основе строят модель множественной регрессии.

Метод пошагового включения переменных в модель позволяет выбрать из возможного набора переменных именно те, которые усилят качество модели регрессии.

Алгоритм метода пошагового включения:

1) из числа всех переменных в модель регрессии включаются те, которые имеют наибольший модуль парного линейного коэффициента корреляции с результативной переменной;

2) при добавлении в модель новых факторов необходимо проверять их значимость с помощью F-критерия Фишера.

Основная гипотеза формулируется как нецелесообразность включения фактора x_k в модель множественной регрессии. Альтернативная гипотеза исходит из обратного утверждения.

$$F_{набл}(x_k) = \frac{R^2(y, x_1, \ldots, x_{n+1}) - R^2(y, x_1, \ldots, x_n)}{R^2(y, x_1, \ldots, x_n)} \times (n - q - 1),$$

где q — число уже включенных в модель переменных.

Критическое значение F-критерия вычисляется по таблице распределения Фишера—Снедекора с уровнем значимости α и числом степеней свободы: $k_1 = 1$ и $k_2 = n - l$: $F_{крит}(\alpha; k_1; k_2)$.

Если $F_{набл} > F_{крит}$, то включение переменной в модель множественной регрессии является обоснованным.

Проверка факторов на значимость осуществляется до тех пор, пока не найдется хотя бы одна переменная, для которой не выполняется условие $F_{набл} > F_{крит}$.

ЛЕКЦИЯ № 12. Нелинейные по переменным, по параметрам регрессионные модели. Регрессионные модели с точками разрыва

Помимо линейных регрессионных моделей, при изучении социально-экономических связей между различными явлениями применяются также модели нелинейной регрессионной зависимости. Их можно разделить на два класса: модели, нелинейные по переменным, входящим в уравнение, и модели, нелинейные по оцениваемым параметрам.

К нелинейным по переменным регрессионным моделям (но линейным по оцениваемым параметрам) относятся полиномиальные функции различных порядков (начиная со второго) и гиперболическая функция.

Общий вид полиномиальной функции *n*-го порядка или *n*-ой степени:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \ldots + \beta_n x_i^n + \varepsilon_i .$$

Полиномиальные функции используются для характеристики процессов с монотонным развитием и отсутствием пределов роста. Поставленному условию отвечают большинство экономических показателей, например натуральные показатели промышленного производства.

Наиболее часто из полиномиальных функций используются полином второго порядка или **параболическая функция**:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i .$$

Он характеризует равноускоренное развитие процесса (равноускоренный рост или снижение уровней).

Регрессионные модели, нелинейные по переменным, отличаются тем, что зависимая переменная линейно связана с оценивае-мыми параметрами β_0, ..., β_n.

Полиномы высоких степеней (более четвертой степени) использовать при изучении социально-экономических связей между переменными не рекомендуется. Это ограничение основано на том, что полиномы высоких порядков имеют больше изгибов

и отразить реальную зависимость результативного признака от факторных переменных практически неспособны.

Характерной особенностью полиномиальных функций является отсутствие явной зависимости приростов факторных переменных от значений результативного признака y_i.

Гиперболическая функция вида:

$$y_i = \beta_0 + \frac{\beta_1}{x_i} + \varepsilon_i$$

также отражает линейную связь между зависимой переменной y_i и параметрами β_0 и β_1, но является нелинейной по факторной переменной x_i. Эта гиперболическая функция является равносторонней.

Гиперболоид применяют при изучении зависимости затрат на единицу продукции от объема производства.

Чтобы оценить неизвестные параметры β_0, ..., β_n нелинейной регрессионной модели, необходимо привести ее к линейному виду. Суть линеаризации нелинейных по факторным переменным регрессионных моделей заключается в замене нелинейных факторных переменных на линейные переменные. В общем случае полиномиальной регрессии замена нелинейных переменных функции n-го порядка выглядит таким образом:

$$x = c_1;\ x^2 = c_2;\ x^3 = c_3;\ \ldots\ x^n = c_n.$$

Уравнение множественной регрессии можно записать в виде:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \ldots + \beta_n x_i^n + \varepsilon_i \Rightarrow$$
$$\Rightarrow y_i = \beta_0 + \beta_1 c_{1i} + \beta_2 c_{2i} + \ldots + \beta_n c_{ni} + \varepsilon_i .$$

Гиперболическую функцию также можно привести к линейному виду с помощью метода замены нелинейной факторной переменной на линейную. Пусть $1/x = c$. Тогда исходное уравнение гиперболической функции можно записать в преобразованном виде:

$$y_i = \beta_0 + \frac{\beta_1}{x_i} + \varepsilon_i \Rightarrow y_i = \beta_0 + \beta_1 c_i + \varepsilon_i .$$

Полиномиальную функцию любой степени и гиперболоид можно свести к модели линейной регрессии, что позволяет применять к преобразованной линейной модели традиционные методы нахождения неизвестных параметров уравнения регрессии (например, классический метод наименьших квадратов) и методы проверки различных гипотез.

1. Нелинейные по параметрам регрессионные модели

К нелинейным моделям относятся регрессионные модели, в которых результативная переменная нелинейно связана с параметрами уравнения $\beta_0, ..., \beta_n$. К такому типу регрессионных моделей относятся:

1) степенная функция $y_i = \beta_0 \times x_i^{\beta_1} \times \varepsilon_i$;

2) показательная функция (простая экспоненциальная)

$$y_i = \beta_0 \times \beta_1^{x_i} \times \varepsilon_i;$$

3) логарифмическая парабола

$$y_i = \beta_0 \times \beta_1^{x_i} \times \beta_2^{x_i^2} \times \varepsilon_i;$$

4) экспоненциальная функция

$$y_i = e^{\beta_0 + \beta_1 x_i} \times \varepsilon_i;$$

5) обратная функция

$$y_i = \frac{1}{\beta_0 + \beta_1 x_i + \varepsilon_i};$$

6) кривая Гомперца

$$y_i = k \times \beta_0^{\beta_1^x};$$

7) логистическая функция (кривая Перла—Рида)

$$y_i = \frac{k}{1 + \beta_1 e^{-\beta_0 x} + \varepsilon_i}.$$

Показательная, логарифмическая и экспоненциальная функции называются кривыми насыщения, потому что дальнейший прирост результативной переменной зависит от уже достигнутого уровня функции. Они используются для описания процессов, имеющие предел роста в изучаемом периоде, например в демографии.

Кривая Гомперца и кривая Перла—Рида относятся к так называемым S-образным кривым. Это кривые насыщения, которые имеют точку перегиба. Эти кривые описывают два последовательных процесса — один с ускорением развития, другой с замедлением достигнутого развития. Этот тип кривых применяется в демографии, в страховании, при решении задач о спросе на новый товар.

Нелинейные по параметрам регрессионные модели, в свою очередь, делятся на модели, подлежащие линеаризации, и модели, которые невозможно свести к линейным.

Для примера моделей, которых можно свести к линейной форме, рассмотрим показательную функцию вида:

$$y_i = \beta_0 \times \beta_1^{x_i} \times \varepsilon_i,$$

где случайная ошибка ε_i мультипликативно связана с факторным признаком x_i. Эта модель является нелинейной по параметру β_1. Для ее линеаризации применим вначале процесс логарифмирования:

$$\log y_i = \log \beta_0 + x_i \times \log \beta_1 + \log \varepsilon_i.$$

После логарифмирования воспользуемся методом замен. Пусть $\log y_i = Y_i$; $\log \beta_0 = A$; $\log \beta_1 = B$; $\log \varepsilon_i = E$.

Преобразованный вид показательной функции можно записать следующим образом:

$$Y_i = A + Bx_i + E_i.$$

Показательная функция является внутренне линейной, и оценки ее параметров могут быть найдены с помощью классического метода наименьших квадратов.

Однако если взять показательную функцию, включающую случайную ошибку ε_i аддитивно, т. е.

$$y_i = \beta_0 \times \beta_1^{x_i} + \varepsilon_i,$$

то данную модель уже невозможно привести к линейному виду процессом логарифмирования. Она является внутренне нелинейной.

Таким же образом можно рассмотреть и степенную функцию, которая является очень популярной в эконометрических исследованиях. Степенными функциями являются кривые Энгеля, кривые спроса и предложения, производственные функции (ПФ).

Пусть задана степенная функция вида:

$$y_i = \beta_0 \times x_i^{\beta_1} \times \varepsilon_i.$$

П

Заменим следующие показатели в полученном уравнении:

$$\ln y_i = Y_i; \ln \beta_0 = A; \ln x_i = X_i; \ln \varepsilon_i = E.$$

Тогда преобразованный вид степенной функции можно записать как:

$$Y_i = A + \beta_1 X_i + \mathrm{E}_i.$$

Степенная функция также является внутренне линейной, и ее оценки можно найти с помощью классического метода наименьших квадратов. Но если взять ее в виде уравнения

$$y_i = \beta_0 \times x_i^{\beta_1} \times \varepsilon_i,$$

где случайная ошибка аддитивно связана с факторной переменной, то модель становится внутренне нелинейной.

К оценке параметров регрессионных моделей, которые нельзя свести к линейным, применяются итеративные процедуры оценивания. Это могут быть квазиньютоновский метод, симплекс-метод, метод Хука—Дживса, метод Розенброка и др.

2. Регрессионные модели с точками разрыва

К регрессионным моделям, являющимися внутренне нелинейными, относятся **регрессионные модели с точками разрыва**, которые, в свою очередь, делятся на **кусочно-линейные модели регрессии** и **собственно модели регрессии с точками разрыва**.

Существование кусочно-линейной регрессии вызвано тем, что нередко вид зависимости между зависимой переменной и независимыми факторами неодинаков в различных областях значений независимых переменных. Можно рассматривать регрессионную зависимость себестоимости единицы какого-либо продукта от объема произведенной продукции за месяц. Данная зависимость носит линейный характер, т. е. с увеличением объема производства себестоимость единицы товара снижается. В некоторых случаях себестоимость может меняться резко, скачкообразно. Если в производстве используются устаревшие модели станков, то с увеличением объема производства себестоимость может также увеличиваться. Если старые станки используются в производстве до того момента, когда объем производства достигнет определенного, заранее заданного значения (например, 300 единиц продукции), то данную зависимость можно аппроксимировать уравнением регрессии вида:

$$y = \beta_0 + \beta_1 \times x \times (x \le 300) + \beta_2 \times x \times (x > 300),$$

где y — себестоимость единицы продукции;

x — объем произведенной за месяц продукции;

$(x \le 300)$ и $(x > 300)$ — логические выражения, принимающие значения единице, если они истинны, или нуля, если они ложны.

Эта регрессионная модель зависит от общего свободного члена β_0 и углового коэффициента, равного β_1 (если выражение $(x \le 300)$ истинно, т. е. равно единице) или β_2 (если выражение $(x > 300)$ истинно, т. е. равно единице).

Если точка разрыва регрессионной кривой (в приведенном примере равная 300 единицам) точно не определена, то можно оценить значение данной точки.

В уравнение регрессии необходимо ввести дополнительный параметр β_3 вместо логических выражений:

$$y = \beta_0 + \beta_1 \times x \times (x \le \beta_3) + \beta_2 \times x \times (x > \beta_3).$$

Это регрессионное уравнение можно легко преобразовать в собственно регрессию с точками разрыва, которая характеризуется скачкообразными изменениями зависимой переменной в некоторых точках кривой. Например, с началом использования старых машин в производстве себестоимость единицы продукции резко подскочила, а затем продолжила медленно снижаться при условии увеличения объемов производства данной продукции. В этом случае регрессионная зависимость примет вид:

$$y = (\beta_0 + \beta_1 \times x) \times (x \le 300) + (\beta_3 + \beta_2 \times x) \times (x > 300).$$

Оценивание параметров регрессии с точками разрыва осуществляется с помощью метода максимального правдоподобия или итерационных методов нелинейного оценивания.

Если стоит выбор между аппроксимацией исходных данных одной из рассмотренных моделей или другой регрессионной моделью (например, линейной или нелинейной, но сводящейся к линейной), то предпочтение отдается более простой форме моделей.

Регрессионные модели могут быть использованы в анализе принадлежности элементов различным группам. Например, если в предыдущий пример добавить условие, что продукцию производят не один, а три завода, то можно сгруппировать переменные по принадлежности к определенному заводу и отразить это в уравнении регрессии. Функция правдоподобия данной модели может быть намного больше, чем у обычной регрессии.

ЛЕКЦИЯ № 13. МНК для нелинейных моделей, методы нелинейного оценивания регрессионных параметров. Показатели корреляции и детерминации для нелинейной регрессии

Метод наименьших квадратов можно применять к нелинейным регрессионным моделям только в том случае, если они являются нелинейными по факторным переменным или нелинейными по параметрам, но внутренне линейными, т. е. возможна линеаризация этих моделей.

Рассмотрим применение метода наименьших квадратов для **определения неизвестных параметров уравнения** параболической зависимости вида:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i.$$

Данный полином второго порядка (или второй степени) является нелинейным по факторным переменным x_i.

Для нахождения неизвестных параметров уравнения регрессии β_0, β_1 и β_2 необходимо минимизировать функционал F:

$$F = \sum_{i=1}^{n} \left(y_i - \tilde{y}_i\right)^2 = \sum_{i=1}^{n} \left(y_i - \tilde{\beta}_0 - \tilde{\beta}_1 x_i - \tilde{\beta}_2 x_i^2\right)^2 \to \min.$$

Процесс минимизации функционала сводится к вычислению частных производных этой функции по каждому из оцениваемых параметров.

Составим стационарную систему уравнений для данного функционала F, не пользуясь при этом методом замен:

$$\begin{cases} \dfrac{\partial F}{\partial \tilde{\beta}_0} = -2\sum\limits_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 x_i - \beta_2 x_i^2\right) = 0, \\ \dfrac{\partial F}{\partial \tilde{\beta}_1} = -2\sum\limits_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 x_i - \beta_2 x_i^2\right) \times x_i = 0, \\ \dfrac{\partial F}{\partial \tilde{\beta}_2} = -2\sum\limits_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 x_i - \beta_2 x_i^2\right) \times x_i^2 = 0. \end{cases}$$

Проведя элементарные преобразования стационарной системы уравнений, получим систему нормальных уравнений для определения коэффициентов параболической зависимости:

$$\begin{cases} n\times\tilde{\beta}_0+\tilde{\beta}_1\sum\limits_{i=1}^{n}x_i+\tilde{\beta}_2\sum\limits_{i=1}^{n}x_i^2=\sum\limits_{i=1}^{n}y_i, \\ \tilde{\beta}_0\sum\limits_{i=1}^{n}x_i+\tilde{\beta}_1\sum\limits_{i=1}^{n}x_i^2+\tilde{\beta}_2\sum\limits_{i=1}^{n}x_i^3=\sum\limits_{i=1}^{n}x_i\times y_i, \\ \tilde{\beta}_0\sum\limits_{i=1}^{n}x_i^2+\tilde{\beta}_1\sum\limits_{i=1}^{n}x_i^3+\tilde{\beta}_2\sum\limits_{i=1}^{n}x_i^4=\sum\limits_{i=1}^{n}x_i^2\times y_i. \end{cases}$$

Данная система является системой нормальных уравнений относительно параметров $\tilde{\beta}_0,\tilde{\beta}_1$ и $\tilde{\beta}_2$ для параболической зависимости

$$y_i=\beta_0+\beta_1x_i+\beta_2x_i^2+\varepsilon_i.$$

Система нормальных уравнений является квадратной, т. е. количество уравнений равняется количеству неизвестных переменных, поэтому коэффициенты $\tilde{\beta}_0,\tilde{\beta}_1$ и $\tilde{\beta}_2$, и можно найти с помощью метода Крамера**,** метода ***Гаусса*** или метода обратных матриц.

В общем случае полинома n-ой степени

$$y_i=\beta_0+\beta_1x_i+\beta_2x_i^2+\ldots+\beta_nx_i^n+\varepsilon_i,$$

для нахождения неизвестных коэффициентов уравнения регрессии методом наименьших квадратов необходимо минимизировать функционал F вида:

$$F=\sum_{i=1}^{n}\left(y_i-\tilde{y}_i\right)^2=\sum_{i=1}^{n}\left(y_i-\tilde{\beta}_0-\tilde{\beta}_1x_i-\tilde{\beta}_2x_i^2-\ldots-\tilde{\beta}_nx_i^n\right)^2\to\min.$$

Тогда система нормальных уравнений будет выглядеть следующим образом:

$$\begin{cases} \sum y_i=\tilde{\beta}_0\times n+\tilde{\beta}_1\sum x_i+\tilde{\beta}_2\sum x_i^2+\ldots+\tilde{\beta}_n\sum x_i^n, \\ \sum y_i\times x_i=\tilde{\beta}_0\sum x_i+\tilde{\beta}_1\sum x_i^2+\tilde{\beta}_2\sum x_i^3+\ldots+\tilde{\beta}_n\sum x_i^{n+1}, \\ \ldots \\ \sum y_i\times x_i^{n-1}=\tilde{\beta}_0\sum x_i^{n-1}+\tilde{\beta}_1\sum x_i^n+\tilde{\beta}_2\sum x_i^{n+1}+\ldots+\tilde{\beta}_n\sum x_i^{2n-1}, \\ \sum y_i\times x_i^n=\tilde{\beta}_0\sum x_i^n+\tilde{\beta}_1\sum x_i^{n+1}+\tilde{\beta}_2\sum x_i^{n+2}+\ldots+\tilde{\beta}_n\sum x_i^{2n}. \end{cases}$$

Решение данной системы позволит найти оценки коэффициентов полинома n-го порядка.

Рассмотрим применение метода наименьших квадратов для нахождения оценок коэффициентов нелинейного по параметрам уравнения регрессии (но внутренне линейного) на примере показательной функции вида:

$$y_i = \beta_0 \times \beta_1^{x_i} \times \varepsilon_i,$$

где случайная ошибка ε_i мультипликативно связана с факторным признаком x_i.

Данная модель является нелинейной по параметру β_1. Для ее линеаризации применим вначале процесс логарифмирования: $\log y_i = \log \beta_0 + x_i \times \log \beta_1 + \log \varepsilon_i$.

После логарифмирования исходного регрессионного уравнения воспользуемся методом замен. Обозначим $\log y_i = Y_i$; $\log \beta_0 = A$; $\log \beta_1 = B$; $\log \varepsilon_i = E$.

Тогда **преобразованный вид показательной функции** можно записать следующим образом:

$$Y_i = A + Bx_i + \mathrm{E}_i.$$

Метод наименьших квадратов применяется не к исходному нелинейному уравнению, а к его линеаризованной форме.

Таким образом, в отличие от линейных регрессионных моделей минимизируется сумма квадратов отклонений логарифмов наблюдаемых значений результативного признака y от теоретических значений $\tilde{y}$ (значений, рассчитанных на основании уравнения регрессии), т. е. минимизируется функционал МНК вида:

$$F = \sum (\log y - \log \tilde{y})^2 \to \min.$$

Для нахождения оценок неизвестных параметров линеаризованного уравнения регрессии A и B решается система нормальных уравнений:

$$\begin{cases} n \times A + B \times \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} Y_i = \sum_{i=1}^{n} \log y_i, \\ A \times \sum_{i=1}^{n} x_i + B \times \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i \times Y_i = \sum_{i=1}^{n} x_i \times \log y_i. \end{cases}$$

Данная система является системой нормальных уравнений относительно коэффициентов A и B для зависимости $Y_i = A + Bx_i + E$.

Оценки параметров для нелинейных регрессионных моделей, сводимых к линейному виду, являются смещенными.

1. Методы нелинейного оценивания регрессионных параметров

Функционал $F = \sum_{i=1}^{n} \left(y_i - \tilde{y}_i\right)^2$ называется функционалом ошибок или функцией потерь, так как любые отклонения наблюдаемых величин от теоретических (т. е. рассчитанных с помощью уравнения регрессии) являются потерями в точности аппроксимации исходных данных. В качестве функции потерь может быть использована сумма модулей отклонений наблюдаемых значений результативного признака *y* от теоретических значений

$$\tilde{y} / F = \sum_{i=1}^{n} \left| y_i - f(x_i, \beta) \right| \quad \text{или} \quad F = \sum_{i=1}^{n} \left| y_i - \tilde{y}_i \right|.$$

Для минимизации функционала ошибок применяются различные методы. **Основной проблемой всех методов** являются локальные минимумы. При небольшом изменении оцениваемого параметра функция потерь практически не изменится, существует вероятность того, что ошибочное значение оцениваемого параметра уравнения регрессии даст в результате существенное умень-шение функции ошибок. Это явление называется локальным минимумом. Локальные минимумы приводят к неправдоподобно завышенным или заниженным оценкам регрессионных параметров. Выходом из ситуации является повторение процедуры оценивания с измененными начальными условиями (шагом, ограничением оцениваемых параметров и т. д.). **Оптимальные оценки коэффициентов** получаются тогда, когда функция ошибок достигает глобального минимума.

Одним из основных методов минимизации функции ошибок является метод Ньютона. Основной шаг в направлении глобального минимума метода Ньютона определяется по формуле:

$$\beta_{k+1} = \beta_k - H_k^{-1} g_k,$$

где $\beta_к$ — вектор значений оцениваемых параметров на *k*-ой итерации;

H — матрица вторых частных производных, или матрица Гессе;

g_k — вектор градиента на k-ой итерации.

Пусть дана скалярная функция y от переменных x_i, $i = \overline{1,n} - f(x)$.

Независимые переменные представлены в виде вектора:

$$x = \left[x_1, x_2, \ldots, x_n\right]^T.$$

Тогда по определению производной:

$$\frac{\partial y}{\partial x} = \frac{\partial f(x)}{\partial x} = \left[\frac{\partial f(x)}{\partial x_1}\frac{\partial f(x)}{\partial x_2}\cdots\frac{\partial f(x)}{\partial x_n}\right].$$

Вектор-столбец

$$\nabla f(x) = \left[\frac{\partial f(x)}{\partial x}\right]^T$$

называется градиентом функции $y = f(x)$ в точке x.

Чтобы избежать громоздких вычислений матрицы Гессе, существуют различные способы ее замены приближенными выражениями, что легло в основу квазиньютоновых методов. Сущность квазиньютоновых методов заключается в том, что вычисляются значения функции ошибок в различных точках для определения первой и второй производной. Первая производная функции в заданной точке равна тангенсу угла наклона графика функции, а вторая производная функции в заданной точке равна скорости его изменения. Эти данные используются для определения направления изменения параметров, а соответственно, и для минимизации функции ошибок.

Методом, не использующим производные функции ошибок, является симплекс-метод. На каждом шаге или на каждой итерации функция ошибок оценивается в $n + 1$ точках n-мерного пространства, образуя фигуру, называемую симплексом. В многомерном пространстве симплекс будет постепенно менять параметры, смещаясь в сторону минимизации функции потерь.

Преимущество симплекс-метода заключается в том, что при слишком большом шаге для точного определения направления минимизации функции потерь (т. е. при слишком большом симплексе), алгоритм автоматически уменьшает симплекс, и вычислительная процедура продолжается.

При обнаружении минимума симплекс снова увеличивается для проверки минимума на локальность.

2. Показатели корреляции и детерминации для нелинейной регрессии. Проверка значимости уравнения нелинейной регрессии

Качество нелинейной регрессионной модели определяется с помощью нелинейного показателя корреляции, который называется индексом корреляции для нелинейных форм связи. Он вычисляется через теорему о разложении дисперсий следующим образом:

$$R = \sqrt{\frac{\sigma^2(y)}{G^2(y)}} = \sqrt{1 - \frac{\delta^2(y)}{G^2(y)}},$$

где $G^2(y)$ — общая дисперсия результативного признака;

$\sigma^2(y)$ — объясненная с помощью построенного уравнения регрессии дисперсия зависимой переменной;

$\delta^2(y)$ — необъясненная или остаточная дисперсия зависимой переменной.

Также индекс корреляции можно вычислить через теорему о разложении сумм квадратов:

$$R = \sqrt{\frac{RSS}{TSS}} = \sqrt{1 - \frac{ESS}{TSS}} = \sqrt{1 - \frac{\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}}.$$

Индекс корреляции для нелинейных форм связи изменяется в пределах [0; +1]. Чем ближе его значение к единице, тем сильнее взаимосвязь между изучаемыми переменными.

Если возвести индекс корреляции в квадрат, то полученная величина будет называться индексом детерминации:

$$R^2 = \frac{\sigma^2(y)}{G^2(y)} = \frac{RSS}{TSS}.$$

Индекс детерминации для нелинейных форм связи по характеристикам аналогичен обычному коэффициенту детерминации.

Если нелинейное по факторным переменным уравнение регрессии с помощью метода замен можно свести к парному линей-

ному уравнению регрессии, то на это уравнение будут распространяться все методы проверки гипотез для парной линейной зависимости.

Проверка гипотезы о значимости индекса корреляции аналогична проверке гипотезы о значимости множественного коэффициента корреляции через F-критерий.

Проверка гипотезы о значимости нелинейной регрессионной модели в целом осуществляется через F-критерий Фишера.

Выдвигается основная гипотеза H_0 о незначимости полученного уравнения регрессии:

$$H_0 \ / \ R^2 = 0.$$

Альтернативной является обратная гипотеза H_1 о значимости построенного уравнения регрессии:

$$H_1 \ / \ R^2 \neq 0.$$

Значение F-критерия вычисляется по формуле:

$$F_{набл} = \frac{R^2 (n-l)}{(1-R^2) \times (l-1)},$$

где n — объем выборочной совокупности;

l — число оцениваемых параметров по выборочной совокупности.

Значение F-критерия $F_{крит}$ вычисляется по таблице распределения Фишера—Снедекора в зависимости от уровня значимости α и числа степеней свободы: $k_1 = l - 1$ и $k_2 = n - l$.

Если $F_{набл} > F_{крит}$, то основная гипотеза отклоняется, и уравнение нелинейной регрессии является значимым.

Если $F_{набл} < F_{крит}$, то основная гипотеза принимается, и уравнение нелинейной регрессии признается незначимым.

Если есть возможность выбора между линейной и нелинейной регрессионными моделями, то предпочтение всегда отдается более простой линейной форме. Проверить предположение о вероятной линейной зависимости между изучаемыми переменными можно с помощью коэффициента детерминации r^2 и индекса детерминации R^2.

Выдвигается гипотеза о линейной зависимости между переменными.

Альтернативной является гипотеза о нелинейной зависимости между переменными. Проверка осуществляется с помощью t-критерия Стьюдента.

Наблюдаемое значение t-критерия находится по формуле:

$$t_{набл} = \frac{R^2 - r^2}{\nu_{R-r}},$$

где ν_{R-r} — величина ошибки разности ($R^2 - r^2$), вычисляемая по формуле:

$$\nu_{R-r} = \sqrt{\frac{\left(R^2 - r^2\right) - \left(R^2 - r^2\right) \times \left(2 - \left(R^2 + r^2\right)\right)}{n}}.$$

Критическое значение t—критерия $t_{крит}(\alpha; n - l - 1)$ определяется по таблице распределения Стьюдента.

Если $t_{набл} > t_{крит}$, то основная гипотеза отклоняется, и между изучаемыми переменными существует нелинейная взаимосвязь.

Если $t_{набл} < t_{крит}$, то зависимость между переменными может быть аппроксимирована линейным регрессионным уравнением.

ЛЕКЦИЯ № 14. Тесты Бокса—Кокса. Средние и точечные коэффициенты эластичности

Если существует выбор между построением линейной и нелинейной регрессионных моделей для изучаемых данных, то предпочтение всегда отдается более простой форме зависимости. Регрессионные модели, имеющие разную функциональную форму, не подлежат сравнению по стандартным критериям (например, сравнению по множественному коэффициенту детерминации или суммам квадратов отклонений), позволяющим выбрать наиболее подходящее уравнение.

При сравнении линейной и логарифмической регрессий оказывается, что общая сумма квадратов отклонений для логарифмической модели намного меньше, чем для линейной модели. Но значение логарифма результативной переменной log *y* намного меньше, чем соответствующее значение *y*, поэтому сравнение сумм квадратов отклонений моделей дает неадекватные результаты.

Коэффициент множественной детерминации для линейной регрессии характеризует объясненную регрессией долю дисперсии результативной переменной *y*. Коэффициент множественной детерминации для логарифмической модели характеризует объясненную регрессией долю дисперсии переменной log *y*. Если значения коэффициентов множественной детерминации примерно равны, то сделать выбор между моделями на основе данного критерия также не представляется возможным.

Использоваться метод проверки гипотезы о линейной зависимости между переменными с помощью коэффициента и индекса детерминации. **Другим эффективным методом выбора функциональной зависимости является тест Бокса—Кокса.** Рассмотрим эту процедуру на примере выбора между линейной и логарифмической регрессионными моделями.

В основе теста Бокса—Кокса лежит утверждение о том, что (y−1) и log y являются частными случаями **функции**

$$F = \frac{y^{\lambda} - 1}{\lambda}.$$

Если λ равен единице, то функция равна $F = y - 1$.
Если λ стремится к нулю, то функция равна $F = \log y$.
Для определения оптимального значения параметра λ проводятся эксперименты с множеством его значений. Эта процедура позволит найти то значение λ, которое дает минимальную величину суммы квадратов отклонений. Метод поиска оптимального значения параметра — **поиск на решетке (или на сетке) значений**.

Один из вариантов теста для линейной и логарифмической моделей разработан ***П. Зарембеки***. Его идея заключается в применении процедуры масштабирования к зависимой переменной, что в дальнейшем позволит сравнивать величины сумм квадратов отклонений регрессий.

Тест Зарембеки состоит из следующих этапов:

1) определяется среднее геометрическое значений y в выборке по формуле:

$$\bar{y} = \sqrt[n]{y_1 \times y_2 \times \ldots \times y_n} = \sqrt[n]{\prod y_i},\ i = \overline{1}$$

2) наблюдения пересчитываются по формуле:

$$\tilde{y}_i = \frac{y_i}{\bar{y}},$$

где $\tilde{y}_i$ — пересчитанное (масштабированное) значение переменной для i-го наблюдения;

3) на последнем этапе оценивается регрессионная зависимость для линейной модели с использованием масштабированных значений $\tilde{y}_i$ вместо y и для логарифмической модели с использованием $\tilde{y}_i$ вместо $\log y$.

Все факторные переменные и регрессионные коэффициенты остаются при этом неизменными. После масштабирования зависимых переменных значения сумм квадратов отклонений для данных регрессионных моделей являются величинами сопоставимыми. Выбор падает на ту модель, для которой данный показатель оказался наименьшим.

Средние и точечные коэффициенты эластичности

Помимо индексов корреляции и детерминации для нелинейных форм связи, для изучения зависимости между результатив-

ной переменной и факторными признаками используются также коэффициенты эластичности, которые позволяют оценить степень связи между *x* и *y*.

Коэффициент эластичности показывает, на сколько процентов приблизительно изменится результативный показатель *y* при изменении величины факторного признака на 1%.

Общая формула коэффициента эластичности:

$$Э = y'_x \times \frac{x}{y} = \frac{\partial y}{\partial x} \times \frac{x}{y} = \frac{\partial y}{\partial x} / \frac{y}{x},$$

где y'_x — первая производная результативной переменной по факторному признаку.

Коэффициент эластичности может быть рассчитан для среднего значения факторного признака по общей формуле:

$$Э(\bar{x}) = \frac{\partial y}{\partial x} \times \frac{\bar{x}}{y(\bar{x})},$$

где $y(\bar{x})$ — значение функции при среднем значении факторного признака.

Средний коэффициент эластичности характеризует процентное изменение результативного признака *y* относительно своего среднего значения при изменении факторного признака на 1% относительного $\bar{x}$.

Средние коэффициенты эластичности рассчитываются по индивидуальным формулам для каждой разновидности функции.

Для наиболее простой линейной зависимости вида $y_i = \beta_0 + \beta_1 x_i$ средний **коэффициент эластичности рассчитывается** как:

$$Э(\bar{x}) = \frac{\beta_1 \bar{x}}{y(\bar{x})}.$$

Для полинома второго порядка (параболической функции) $y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i$ средний коэффициент эластичности рассчитывается по формуле:

$$Э(\bar{x}) = \frac{(2\beta_2 \bar{x} + \beta_1) \times \bar{x}}{y(\bar{x})}.$$

Для показательной функции вида $y_i = \beta_0 \times \beta_1^{x_i} \times \varepsilon_i$ средний коэффициент эластичности определяется как:

$$Э(\bar{x}) = \ln \beta_1 \times \bar{x}.$$

Основным достоинством степенной функции вида

$$y_i = \beta_0 \times x_i^{\beta_1} \times \varepsilon_i$$

является то, что средний коэффициент эластичности $Э(\bar{x})$ равен коэффициенту регрессии β_1:

$$Э(\bar{x}) = \beta_1.$$

Это единственная функция подобного рода.

Помимо средних коэффициентов эластичности, могут быть также рассчитаны точечные коэффициенты эластичности. Общая формула их расчета:

$$Э(x_1) = \frac{\partial y}{\partial x} \times \frac{x_1}{y(x_1)},$$

т. е. эластичность зависит от конкретного заданного значения факторного признака x_1. Точечный коэффициент эластичности характеризует процентное изменение результативной переменной *y* относительно уровня функции $y(x_1)$ при изменении факторного признака на 1% относительно заданного уровня x_1.

Для линейной зависимости точечный коэффициент эластичности будет рассчитываться по формуле:

$$Э(x_1) = \frac{\beta_1 x_1}{\beta_0 + \beta_1 x_1}.$$

Знаменателем данного показателя является значение линейной функции в точке.

Для параболической функции **точечный коэффициент эластичности находится** как:

$$Э(x_1) = \frac{(2\beta_2 x_1 + \beta_1) \times x_1}{\beta_0 + \beta_1 x_1 + \beta_2 x_1^2}.$$

Знаменателем данного показателя также является значение параболической функции в точке.

Для показательной функции точечный коэффициент эластичности определяется по формуле:

$$Э(x_1) = \ln \beta_1 \times x_1.$$

В случае степенной функции точечный коэффициент эластичности $Э(x_1)$ будет равен коэффициенту регрессии β_1.

Докажем предыдущее утверждение.

Запишем точечный коэффициент эластичности для степенной функции вида $y_i = \beta_0 \times x_i^{\beta_1} \times \varepsilon_i$ через первую производную результативной переменной по заданной факторной переменной x_1:

$$Э(x_1) = \frac{\beta_0 \beta_1 x_1^{\beta_1 - 1}}{\frac{\beta_0 x_1^{\beta_1}}{x_1}} = \frac{\beta_0 \beta_1 x_1^{\beta_1 - 1}}{\beta_0 x_1^{\beta_1 - 1}} = \beta_1,$$

таким образом, $Э(x_1) = \beta_1$, что и требовалось доказать.

Коэффициенты эластичности имеют очень большое значение в анализе производственных функций. Однако их расчет не всегда имеет смысл. В некоторых случаях интерпретация факторных переменных в процентном отношении невозможна или бессмысленна.

ЛЕКЦИЯ № 15. Производственные функции. Эффект от масштаба производства

Производственная функция — экономико-математическая модель, позволяющая аппроксимировать зависимость результатов производственной деятельности предприятия, отрасли или национальной экономики в целом от повлиявших на эти результаты факторов.

В качестве **факторов производственной функции** могут выступать следующие переменные:

1) объем выпущенной продукции (в стоимостном или натуральном выражении);
2) объем основного капитала или основных фондов;
3) объем трудовых ресурсов или трудовых затрат (измеряемое количеством рабочих или количеством человекодней);
4) затраты электроэнергии;
5) количество станков, используемых в производстве, и др.

Простой **разновидностью производственных функций** являются однофакторные производственные функции (ОПФ). Зависимой переменной в данных функциях является объем производства y, который зависит от единственной независимой переменной x. В качестве независимой переменной выступать показатель общих производственных затрат.

К **однофакторным производственным функциям относятся**:

1) линейная ОПФ $y = \beta_0 + \beta_1 x$. Данная функция выражает зависимость объема производимой продукции от величины затрат определенного ресурса. **Линейная ОПФ характеризуется следующими особенностями:**

   а)если величина независимого признака x равна нулю, то объем производства не будет нулевым, так как $y = \beta_0$ ($\beta_0 > 0$);

   б) объем произведенной продукции неограниченно увеличивается с ростом затрат определенного ресурса x на постоянную величину β_1 ($\beta_1 > 0$). Это свойство линейной ОПФ выполняется только на практике;

2) параболическая ОПФ при $y_i = \beta_0 + \beta_1 x + \beta_2 x^2$ при $\beta_0 > 0, \beta_1 > 0, \beta_2 > 0$.

Особенностью данной функции является то, что с увеличением затрат ресурса x объем произведенной продукции y вначале возрастает до некоторой максимальной величины, а затем снижается до нуля;

3) степенная ОПФ $y = \beta_0 \times x^{\beta_1}$ при $\beta_0 > 0, \beta_1 > 0$. Функция характеризуется тем, что с увеличением затрат ресурса x объем производства возрастает неограниченно;

4) показательная ОПФ вида $y = \beta_0 - k \times \beta_1^x$ при $0 < \beta_1 < 0$. С возрастанием затрат ресурса x объем произведенной продукции также растет, стремясь при этом к значению β_0.

Гиперболическая ОПФ $y = \beta_0 + \beta_1 / x$ практически не используется при изучении зависимости объема производства от затрат какого-либо ресурса, так как нет необходимости в изучении ресурсов, увеличение которых приводит к уменьшению объема производства.

Двухфакторные производственные функции характеризуют зависимость объема производства от каких-либо двух факторов.

Чаще всего это факторы объема основного капитала и трудовых ресурсов. К наиболее известным двухфакторным производственным функциям относятся функции Кобба—Дугласа и Солоу.

Для графического изображения двухфакторных производственных функций строят семейство кривых, основанных на различном сочетании двух факторов, но в результате они дают один и тот же объем выпуска продукции. Кривые, построенные на основании равенства $f(x_1, x_2) = \text{const}$, называются изоквантами. Изокванта — сочетание минимально необходимых ресурсных затрат для заданного уровня объема производства.

Многофакторная производственная функция (МПФ) имеет вид $f(x_i)$, где $i = \overline{1,n}$. МПФ характеризует зависимость объема производства от n-го количества факторов производства.

1. Двухфакторная производственная функция Кобба—Дугласа

Основоположниками теории производственных функций считают американских ученых ***Д. Кобба*** и ***П. Дугласа***, опубликовавших в 1928 г. работу «Теория производства».

Учеными предложена производственная функция, которая носит название функции Кобба—Дугласа. В общем виде ее можно записать:

$$f(x_i) = a\prod_{i=1}^{n} x_i^{\alpha_i},$$

где a — числовой параметр функции;

x_i — i-тый аргумент или i-ый фактор производственной функции;

α — показатель степени i-го аргумента.

Часто используемой формой функции Кобба—Дугласа является ее двухфакторный вариант $f(K, L)$:

$$Q = A \times K^{\alpha} \times L^{\beta},$$

где Q— объем выпущенной продукции (в стоимостном или натуральном выражении);

K — объем основного капитала или основных фондов;

L — объем трудовых ресурсов или трудовых затрат (измеряемый количеством рабочих или количеством человеко-дней);

A, α, β — неизвестные числовые параметры функции, которые подчиняются следующим условиям:

$$0 \le \alpha \le 1,\ 0 \le \beta \le 1,\ A > 0,\ \alpha + \beta = 1.$$

Величина A зависит от единиц измерения результативной и факторных переменных.

Исходя из условия $\alpha + \beta = 1$ **функцию Кобба—Дугласа можно записать как**:

$$Q = A \times K^{\alpha} \times L^{1-\alpha}.$$

На двухфакторную функцию Кобба—Дугласа накладываются **определенные ограничения**:

$$1)\ Q'_K = \frac{\partial Q}{\partial K} > 0;$$

$$2)\ Q'_L = \frac{\partial Q}{\partial L} > 0.$$

Первое и второе ограничения означают, что объем выпускаемой продукции увеличивается при постоянном значении одного из факторов и росте другого фактора;

3) $Q''_{KK} = \frac{\partial^2 Q}{\partial^2 K} < 0;$

4) $Q''_{LL} = \frac{\partial^2 Q}{\partial^2 L} < 0.$

Ограничения 3 и 4 означают, что при фиксированном значении одного из факторов последовательное увеличение другого фактора будет приводить к сокращению прироста значения y;

5) $K > 0$;
6) $L > 0$.

Для функции Кобба—Дугласа можно рассчитать **частные коэффициенты эластичности**.

Частный коэффициент эластичности функции Кобба—Дугласа по переменной K:

$$Э_K(Q) = Q'_K \times \frac{K}{Q} = \frac{K}{A \times K^\alpha \times L^\beta} \times \alpha \times A \times K^{\alpha-1} \times L^\beta = \alpha.$$

Частный коэффициент эластичности функции Кобба—Дугласа $Э_k(y) = \alpha$, т. е. является независимым от переменных K и L.

Частный коэффициент эластичности функции Кобба—Дугласа по переменной L:

$$Э_L(Q) = Q'_L \times \frac{L}{Q} = \frac{L \times \beta \times A \times K^\alpha \times L^{\beta-1}}{A \times K^\alpha \times L^\beta} = \beta.$$

Частный коэффициент эластичности по фактору трудовых ресурсов также является независимым от переменных K и L.

Средними показателями двухфакторной функции Кобба—Дугласа являются средняя производительность труда и средняя фондоотдача:

$b = \frac{Q}{L} = \frac{A \times K^\alpha \times L^\beta}{L} = A \times K^\alpha \times L^{\beta-1}$ — коэффициент средней производительности труда;

$z = \frac{Q}{K} = \frac{A \times K^\alpha \times L^\beta}{K} = A \times K^{\alpha-1} \times L^\beta$ — коэффициент средней фондоотдачи.

Предельными показателями двухфакторной функции Кобба—Дугласа являются предельная производительность труда и предельная фондоотдача:

$$V = Q'_L = \left(A \times K^\alpha \times L^\beta\right)'_L = \beta \times A \times K^\alpha \times L^{\beta-1} = \frac{\beta \times Q}{L} = \beta \times b$$ —

коэффициент предельной производительности труда, который характеризует величину эффекта от каждой дополнительной единицы затраченного труда.

Показатель предельной производительности пропорционален показателю средней производительности, но всегда меньше этой величины, так как $0 \le \beta \le 1$.

$W = Q'_K = \alpha \times A \times K^{\alpha-1} \times L^{\beta} = \frac{\alpha \times Q}{K} = \alpha \times z$ — показатель предельной фондоотдачи, характеризующий величину эффекта от каждой дополнительной единицы основных фондов, использованной в производстве.

Показатель предельной фондоотдачи пропорционален показателю средней производительности, но всегда меньше этой величины, так как $0 \le \alpha \le 1$.

$T = \frac{V}{W} = \frac{Q'_L}{Q'_K}$ — показатель предельной нормы технической замены факторов, т. е. замены труда капиталом. Он показывает, на сколько единиц можно уменьшить объем используемого капитала при увеличении объема трудовых затрат на единицу и фиксированном объеме выпуска продукции.

2. Эффект от масштаба производства. Двухфакторная производственная функция Солоу

Эффект от масштаба — изменение объема произведенной продукции при пропорциональном изменении затрат труда и капитала (для двухфакторной производственной функции).

Пусть изменение объема основного капитала составило nK, а увеличение объема трудовых затрат составило nL. Определим изменение объема производства для функции Кобба-Дугласа

$$Q = A \times K^{\alpha} \times L^{\beta}:$$

$$Q(n) = A \times (nK^{\alpha}) \times (nL^{\beta}) = A \times K^{\alpha} \times L^{\beta} \times n^{\alpha+\beta} = Q \times n^{\alpha+\beta}.$$

Функция имеет возрастающий эффект от масштабов производства, если $(\alpha + \beta) > 1$, т. е. с увеличением факторов K и L в n раз объем производства возрастает в Q раз.

Функция имеет фиксированный эффект от масштабов производства если $(\alpha + \beta) = 1$, т. е. с увеличением K и L в n раз объем производства также возрастает в n раз.

Функция имеет убывающий эффект от масштабов производства, если $(\alpha + \beta) < 1$, т. е. с увеличением K и L в n раз объем производства возрастает меньшими, чем n, темпами.

Американским ученым Р. Солоу в 1956 г. предложена двухфакторная производственная функция вида:

$$Q = \varphi(K, L) = A \times \left[\alpha \times K^{-\rho} + (1-\alpha) \times L^{-\rho}\right]^{-\frac{1}{\rho}},$$

которая получила широкое применение.

Параметры A, ρ, α являются технологическими характеристиками функции Солоу и удовлетворяют условиям: $A > 0$, $0 \le \alpha \le 1$, $\rho > 0$. Параметр α имеет ту же размерность, что и факторные переменные.

Производственная функция Солоу имеет много преимуществ по сравнению с производственной функцией Кобба-Дугласа.

Функция Солоу является **однородной** относительно переменных, т. е. для нее также выполняется правило эффекта от масштаба производства:

$$Q(n) = A \times \left(\alpha \times (nK)^{-\rho} + (1-\alpha) \times (nL)^{-\rho}\right)^{-\frac{1}{\rho}} =$$

$$= A \times n \times \left(\alpha \times K^{-\rho} + (1-\alpha) \times L^{-\rho}\right)^{-\frac{1}{\rho}} = n \times Q.$$

Данное равенство означает, что при увеличении факторов K и L в n раз объем произведенной продукции Q увеличивается также в n раз (если). При уменьшении факторов K и L в n раз объем произведенной продукции Q также уменьшается в n раз (если $0 < n < 1$).

Если один из факторов производственной функции Солоу равен нулю, например $K = 0$, то изменение объема производства будет линейно зависеть от изменения объема второго фактора, т. е. затрат труда. И, наоборот, если $L = 0$, то изменение Q линейно зависит от изменения затрат основного капитала.

Если зафиксировать факторную переменную K на уровне K_0, то объем произведенной продукции Q будет возрастать с увеличением фактора L. Аналогично, если зафиксировать переменную L на уровне L_0, то объем произведенной продукции Q будет возрастать с увеличением фактора K. Для доказательства этого утверждения рассчитаем предельную производительность факторной переменной L:

$$\alpha \times K^{-\rho} + (1-\alpha) \times L^{-\rho\ -\frac{1}{\rho}-1} \times (1-\alpha) \times (-\rho) \times \left(L^{-\rho-1}\right) =$$

$$= \left[\alpha \times K^{-\rho} + (1-\alpha) L^{-\rho}\right]^{-\frac{1}{\rho}-1} \times (1-\alpha)^{-\rho-1} > 0.$$

Предельная производительность ресурса L всегда больше нуля.

Предельная производительность второго ресурса (объема основных фондов) также больше нуля, что говорит о возрастании объема произведенной продукции с увеличением второго фактора K и при фиксированном значении фактора L.

Функция Солоу характеризуется тремя технологическими параметрами A, ρ, α, для определения которых достаточно всего трех измерений переменных функции (основного капитала, трудовых затрат и объема производства).

Изоквантой для производственной функции Солоу является кривая, которая определяется равенством $\varphi(L, K) = \text{const}$.

Частный коэффициент эластичности функции Солоу по переменной K определяется по формуле:

$$Э_K(Q) = Q'_K \times \frac{K}{Q} = \frac{\alpha}{A} \times \left(\frac{Q}{K}\right)^{\rho}.$$

Частный коэффициент эластичности функции Солоу по переменной :

$$Э_K(Q) = Q'_K \times \frac{K}{Q} = \frac{\alpha}{A} \times \left(\frac{Q}{K}\right)^{\rho}.$$

3. МНК для функции Кобба-Дугласа. Многофакторные производственные функции

Функция Кобба-Дугласа — нелинейный по параметрам класс функций, которые являются внутренне линейными. Оценки параметров данной функции можно найти с помощью метода наименьших квадратов.

Применим процесс логарифмирования к двухфакторной функции Кобба= Дугласа для приведения ее к линейному виду:

$$\ln Q - \ln L = \ln \alpha + \beta(\ln K - \ln L),$$

или

$$\ln Q_j - \ln L_j = \ln \alpha + \beta\left(\ln K_j - \ln L_j\right) + \varepsilon_j; \quad j = \overline{1,24},$$

где ε_j — является случайной ошибкой функции.

Для дальнейших действий воспользуемся **методом замен**. Введем обозначения: $y_j = \ln Q_j - \ln L_j$; $b_0 = \ln \alpha$; $b_1 = \beta$; $b = [b_0\, b_1]^{\mathrm{T}}$; $x_j = \ln K_j - \ln L_j$; $\delta^{\mathrm{T}}(x_j) = [0\, x_j]$.

Тогда функцию Кобба-Дугласа можно записать в виде:

$$y_j = \delta^T\left(x_j\right) \times b + \varepsilon_j, \quad j = \overline{1,n}.$$

Методом наименьших квадратов определяется оценка вектора b неизвестных коэффициентов данного уравнения по формулам:

$$\tilde{b}_0 = \bar{y} - \bar{x}\tilde{b}_1;\ \ \tilde{b}_1 = \frac{\overline{xy} - \bar{x} \times \bar{y}}{\overline{x^2} - \bar{x}^2},$$

где $\bar{x} = \dfrac{\sum_{j=1}^{n} x_j}{n}$ — среднее арифметическое значение переменной x;

$\bar{y} = \dfrac{\sum_{j=1}^{n} y_j}{n}$ — среднее арифметическое значение переменной y;

$\overline{x^2} = \dfrac{\sum_{j=1}^{n} x_j^2}{n}$ — среднее значение квадрата переменной x;

$\overline{xy} = \dfrac{\sum_{j=1}^{n} x_j y_j}{n}$ — среднее значение произведения переменных x и y.

Определив оценки b_0 и b_1 можно без труда найти оценки параметров A, α, β исходного регрессионного уравнения, т. е. собственно функции Кобба—Дугласа.

Многофакторная производственная функция (МПФ) имеет вид $y = f(x_i)$, где $i = \overline{1,n}$. МПФ характеризует зависимость объема производства от n-го количества факторов производства. При изучении МПФ можно получить целый ряд важных расчетных экономических показателей.

Показатель средней производительности (эффективности, отдачи) i-го фактора при условии фиксированности всех остальных факторов **определяется по формуле**:

$$\frac{y}{x} = \frac{f\left(x_1,\ x_2,\ \ldots,\ x_n\right)}{x_i}.$$

Предельная производительность (эффективность, отдача) i-го фактора рассчитывается как частная производная по фактору x_i:

$$y'_{x_i} = f'_{x_i}\left(x_1,\ \ldots,\ x_n\right).$$

Определение **характера изменения предельной производительности** с изменением объема *i*-го фактора при фиксированном объеме остальных факторов рассчитывается частная производная второго порядка по фактору x_i:

$$y''_{x_i x_i} = f''_{x_i x_i}(x_1, \ldots, x_n).$$

Если показатель $y''_{x_i x_i} > 0,$ то предельная производительность увеличивается с увеличением объема *i*-го фактора.

Если $y''_{x_i x_i} = 0,$ то можно найти такое значение объема *i*-го фактора, при котором предельная производительность будет или минимальной, или максимальной.

Показатель частной эластичности *i*-го ресурса для многофакторной производственной функции дает характеристику относительного изменения результата производства на единицу относительного изменения *i*-го фактора:

$$Э_i = y'_{x_i} \times \frac{x_i}{y}.$$

Потребность производства в *i*-том факторе может быть выражена через функциональную зависимость вида:

$$x_i = \varphi(y, x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n).$$

Для любой пары факторов производства *i* и *j* можно рассчитать предельную норму замещения *j*-го фактора *i*-тым фактором. Эта норма равна взятому со знаком минус отношению показателей предельной производительности *i*-го и *j*-го ресурсов:

$$h_{ij} = -\frac{y'_{x_i}}{y'_{x_j}}.$$

При выборе конкретного вида производственной функции необходимо учитывать закономерности изменения всех вышеперечисленных показателей. Иногда выбранную форму ПФ приходится отвергать, так как соответствующая ей система показателей противоречит результатам качественного анализа или эмпирическим данным. Предварительные заключения о характере изменений рассмотренных показателей могут стать основным доводом в пользу выбора той или иной формы ПФ.

ЛЕКЦИЯ № 16. Модели бинарного выбора. Метод максимума правдоподобия

В нормальной линейной регрессионной модели вида:

$$y_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_k x_{ki} + \varepsilon_i$$

зависимая переменная y_i является непрерывной величиной, которая может принимать любые значения. Существуют регрессионные зависимости, в которых переменная y_i должна принимать определенный узкий круг заранее заданных значений. Эти зависимости называются моделями бинарного выбора. Примерами такой переменной могут служить:

$$y_i = \begin{cases} 1, \textit{занятые} \\ 0, \textit{безработные} \end{cases} \text{ или } y_i = \begin{cases} 1, \textit{выздоровление} \\ 0, \textit{болезнь}. \end{cases}$$

Рассмотренные бинарные переменные являются величинами дискретными. Бинарная непрерывная величина задается как:

$$y_i = \begin{cases} 0 \\ y_i \end{cases}.$$

Прогнозные значения $y_i^{прогноз}$ будут выходить за пределы интервала [0; +1], поэтому их нельзя будет интерпретировать.

Задачу регрессии можно сформулировать не как предсказание конкретных значений бинарной переменной, а как предсказание непрерывной переменной, значения которой заключаются в интервале [0; +1].

Для аппроксимации данной регрессионной зависимости необходимо подобрать кривую, которая отвечала бы **следующим свойствам**: $F(-\infty) = 0$; $F(+\infty) = 1$; при $x_1 > x_2$ – $F(x_1) > F(x_2)$.

Указанным свойствам удовлетворяет функция распределения вероятности. С помощью данной функции парную регрессионную модель с зависимой бинарной переменной можно представить в виде:

$$prob(y_i = 1) = F(\beta_0 + \beta_1 x_i),$$

где $prob(y_i = 1)$ — это вероятность того, что зависимая переменная y_i примет значение, равное единице.

Достоинством применения функции распределения вероятности является то, что прогнозные значения $y_i^{прогноз}$ будут лежать в пределах интервала [0; +1].

Модель бинарного выбора можно записать через скрытую (латентную) переменную:

$$y_i^* = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_k x_{ki} + \varepsilon_i$$

или в векторном виде:

$$y_i^* = x_i^T \beta + \varepsilon_i,$$

где зависимая бинарная переменная y_i принимает следующие значения в зависимости от латентной y_i^*:

$$y_i = \begin{cases} 1, y_i^* \geq 0 \\ 0, y_i^* < 0. \end{cases}$$

Если предположить, что остатки регрессионной модели бинарного выбора ε_i являются случайными нормально распределенными величинами, а функция распределения вероятностей является нормальной вероятностной функцией, то модель бинарного выбора будет называться пробит-моделью или пробит-регрессией (probit regression).

Пробит-регрессия может быть выражена уравнением вида:

$$NP(y_i) = NP(\beta_0 + \beta_1 x_{1i} + \ldots + \beta_k x_{ki}),$$

где NP — это нормальная вероятность (normal probability).

Если же предположить, что случайные остатки ε_i подчиняются логистическому закону распределения, то модель бинарного выбора называется логит-моделью или логит-регрессией (logit regression).

Логит-регрессию можно записать с помощью следующего уравнения:

$$y_i = \frac{\exp(\beta_0 + \beta_1 x_{1i} + \ldots + \beta_k x_{ki})}{(1 + \exp(\beta_0 + \beta_1 x_{1i} + \ldots + \beta_k x_{ki}))}.$$

Основное достоинство данного уравнения заключается в том, что при любых значениях факторных переменных и регрессионных коэффициентов значения зависимой переменной y_i будут всегда лежать в интервале [0; +1].

Помимо рассмотренной логит-модели, существует также обобщенная логит-модель, которая может выражаться уравнением:

$$y_i = \frac{\beta_0}{1 + \beta_1 \times \exp(\beta_2 \times x_i)},$$

которая позволяет зависимой переменной произвольно меняться внутри фиксированного интервала (не только [0; +1]).

Логит - модель может быть сведена к линейной с помощью преобразования, носящего название логистического, или логит-преобразования, которое можно записать на примере преобразования обычной вероятности *p*:

$$p^* = \log_e\left(\frac{p}{1-p}\right).$$

Показателем качества построенной пробит- или логит-регрессии является псевдокоэффициент детерминации:

$$psevdoR^2 = 1 - \frac{1}{\frac{1 + 2(l_1 - l_0)}{N}}.$$

Если его значение близко к единице, то модель считается адекватной реальным данным.

Метод максимума правдоподобия.

Термин «метод максимума правдоподобия» (maximum likelihood function) был впервые использован в работе Р. А. Фишера в 1922 г.

Этот метод — альтернатива методу наименьших квадратов и состоит в максимизации функции правдоподобия или ее логарифма.

Общий вид функции правдоподобия:

$$L(X, \beta) = \prod_{i=1}^{n} \{p(y_i, x_i)\},$$

где $\prod$ — это геометрическая сумма, означающая перемножение вероятностей по всем возможным случаям внутри скобок.

Построена регрессионная модель бинарного выбора, где зависимая переменная представлена через скрытую (латентную) переменную:

$$y_i = \begin{cases} 1, y_i^* \geq 0 \\ 0, y_i^* < 0, \end{cases}$$

где $y_i^* = x_i^T \beta + \varepsilon_i$.

Вероятность того, что переменная y_i примет значение единицы, можно выразить следующим образом:

$$p(y_i = 1) = p(y_i^* \geq 0) = p(x_i^T \beta + \varepsilon \geq 0) = p(\varepsilon \leq -x_i^T \beta) = F(x_i^T \beta).$$

Вероятность того, что переменная y_i примет значение нуль, будет равно:

$$p(y_i = 0) = 1 - F(x_i^T \beta).$$

Для вероятностей выполняется следующее равенство:

$$p(y_1 = 1, y_2 = 0) = p(y_1 = 1) \times p(y_2 = 0).$$

С учетом данного равенства **функцию правдоподобия можно записать как геометрическую сумму вероятностей наблюдений**:

$$L(\beta, X) = p(y_1 = 1, y_2 = 0 \ldots) = \prod_{y_i = 1} F(x_i^T \beta) \prod_{y_i = 0} (1 - F(x_i^T \beta))$$

Функция правдоподобия для регрессионных логит- и пробит-моделей строится через сумму натуральных логарифмов правдоподобия:

$$l(\beta, X) = \ln L(\beta, X) = \sum \ln F(x_i^T \beta) + \sum \ln (1 - F(x_i^T \beta))$$

Для нахождения оценок неизвестных коэффициентов логит- и пробит-регрессии метод наименьших квадратов применять не оптимально. Оценки β определяются с помощью максимизации функции правдоподобия для логит- и пробит-регрессий:

$$l(\beta, X) \xrightarrow{\beta} \max.$$

Для нахождения максимума функции $l(\beta, X)$ вычислим частные производные по каждому из оцениваемых параметров и приравняем их к нулю:

$$\begin{cases} \dfrac{\partial l}{\partial \beta_1} = 0, \\ \dfrac{\partial l}{\partial \beta_2} = 0, \\ \ldots \\ \dfrac{\partial l}{\partial \beta_k} = 0. \end{cases}$$

Путем преобразований исходной системы уравнений находим стационарную систему уравнений, а затем систему нормальных уравнений.

Решениями системы нормальных уравнений будут оценки максимального правдоподобия $\tilde{\beta}_{ML}$.

Проверка значимости вычисленных коэффициентов пробит- и логит-регрессии и уравнения регрессии определяется с помощью величины $(l_1 - l_0)$, где l_1 соответствует максимально правдоподобной оценке основного уравнения регрессии;
l_0 — оценка нулевой модели регрессии, т. е. $y_i = \beta_0$.

Выдвигается основная **гипотеза о незначимости коэффициентов** пробит- или логит-регрессии:

$$H_0 \,/\, \beta_1 = \beta_2 = \ldots = \beta_k = 0.$$

Для проверки гипотезы вычисляется величина $H = -2\left(l_1 - l_0\right)$, которая подчиняется распределению с k степенями свободы.

Величина H сравнивается с критическим значением χ^2-критерия, которое зависит от заданного значения вероятности α и степени свободы k.

Если $H > \chi^2$, то основная гипотеза отвергается, коэффициенты регрессионной зависимости являются значимыми, следовательно, значимым является само уравнение логит- или пробит-регрессии.

Пусть ω — это элемент, принадлежащий заданному пространству A. Если A является открытым интервалом, а функция $L(\omega)$ дифференцируема и достигает максимума в заданном интервале A, то оценки максимального правдоподобия удовлетворяют равенству

$$\frac{\partial L(\omega)}{\partial \omega} = 0.$$

Докажем высказанное утверждение на примере логит-регрессии.

Функция максимального правдоподобия для логит-модели имеет вид:

$$l(\beta, X) = \ln L(\beta, X) = \sum_{y_i=1} \ln F(x_i^T \beta) + \sum_{y_i=0} \ln(1 - F(x_i^T \beta)) =$$

$$= \sum_{y_i=1} \ln \frac{e^{x_i^T \beta}}{1 + e^{x_i^T \beta}} + \sum_{y_i=0} \ln(\frac{1}{1 + e^{x_i^T \beta}}) = \sum_{y_i=1} x_i^T \beta - \sum_{i=1}^{n} \ln(1 + e^{x_i^T \beta}).$$

Полученную функцию продифференцируем по β:

$$\frac{\partial l}{\partial \beta} = \sum_{i=1}^{n} x_i y - \sum_{i=1}^{n} \frac{x_i \times e^{x_i^T \beta}}{1 + e^{x_i^T \beta}} = \sum x_i \left(y_i - \frac{e^{x_i^T \beta}}{1 + e^{x_i^T \beta}} \right) =$$

$$= \sum x_i (y_i - \tilde{p}_i) = 0.$$

Утверждение доказано.

Если регрессионная модель удовлетворяет предпосылкам нормальной линейной регрессионной модели, то оценки коэффициентов, полученные с помощью метода наименьших квадратов, и оценки, полученные с помощью метода максимума правдоподобия, будут одинаковыми.

ЛЕКЦИЯ № 17. Гетероскедастичность остатков регрессионной модели. Обнаружение и устранение гетероскедастичности

Термин «**гетероскедастичность**» в широком смысле означает предположение о дисперсии случайных ошибок регрессионной модели. Случайная ошибка — отклонение в модели линейной множественной регрессии:

$$\varepsilon = y_{\text{i}} - \beta_0 - \beta_1 x_{1R} - \ldots - \beta_n x_{ik}.$$

Величина случайной регрессионной ошибки является неизвестной, поэтому вычисляется выборочная оценка случайной ошибки регрессионной модели по формуле:

$$e_i = y_i - \tilde{y}_i = y_i - \tilde{\beta}_0 - \tilde{\beta}_1 x_{1k} - \ldots - \tilde{\beta}_n x_{ik},$$

где e_i — остатки регрессионной модели.

Нормальная линейная регрессионная модель строится на основании следующих предпосылок о случайной ошибке:

1) математическое ожидание случайной ошибки уравнения регрессии равно нулю во всех наблюдениях: $E(\varepsilon_i) = 0$, где $i = \overline{1,n}$;

2) дисперсия случайной ошибки уравнения регрессии является постоянной для всех наблюдений: $D(e_i) = E(e_i^2) = G^2 = \text{const}$;

3) случайные ошибки уравнения регрессии не коррелированы между собой, т. е. ковариация случайных ошибок любых двух разных наблюдений равна нулю: $Cov(\varepsilon_i, \varepsilon_j) = E(\varepsilon_i, \varepsilon_j) = 0$, где $i \neq j$.

Условие $D(\varepsilon_i) = E(\varepsilon_i^2) = G^2 = \text{const}$ трактуется как гомоскедастичность (homoscedasticity — «однородный разброс») дисперсий случайных ошибок регрессионной модели. Гомоскедастичность — это предположение о том, что дисперсия случайной ошибки ε_i является известной постоянной величиной для всех i наблюдений регрессионной модели.

На практике предположение о гомоскедастичности случайной ошибки ε_i или остатков регрессионной модели e_i далеко не всегда оказывается верным.

Предположение о том, что дисперсии случайных ошибок являются разными величинами для всех наблюдений, называется гетероскедастичностью (heteroscedasticity — неоднородный разброс):

$$D(\varepsilon_i) \neq D(\varepsilon_j) \neq G^2 \neq \text{const},$$

где $i \neq j$.

Условие гетероскедастичности можно записать через ковариационную матрицу случайных ошибок регрессионной модели:

$$Cov(\varepsilon_i) = \Omega = \begin{pmatrix} G_1^2 & 0 & \dots & 0 \\ 0 & G_2^2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & G_n^2 \end{pmatrix}.$$

где $G_1{}^2 \neq G_2{}^2 \neq \dots \neq G_n{}^2$.

Тогда ε_t подчиняется нормальному закону распределения с параметрами: $\varepsilon_t \sim N(0; G^2\,\Omega)$, где Ω — матрица ковариаций случайной ошибки.

Наличие гетероскедастичности в регрессионной модели может привести к негативным последствиям:

1) оценки уравнения нормальной линейной регрессии остаются несмещенными и состоятельными, но при этом теряется эффективность;

2) появляется большая вероятность того, что оценки стандартных ошибок коэффициентов регрессионной модели будут рассчитаны неверно, что конечном итоге может привести к утверждению неверной гипотезы о значимости регрессионных коэффициентов и значимости уравнения регрессии в целом.

Если дисперсии случайных ошибок регрессионной модели G_i^2 известны заранее, то от проблемы гетероскедастичности можно было бы легко избавится. Но на практике, как правило, неизвестна даже точная функция зависимости $y = f(x)$ между изучаемыми переменными, которую предстоит построить и оценить. Чтобы в подобранной регрессионной модели обнаружить гетероскедастичность, необходимо провести анализ остатков регрессионной модели. Проверяются следующие гипотезы.

Основная гипотеза H_0, утверждающая о постоянстве дисперсий случайных ошибок регрессии, т. е. о присутствии в модели условия гомоскедастичности:

$$H_0 / G_1^2 = G_2^2 = \dots = G_n^2.$$

Альтернативной гипотезой H_1 является предположение о неодинаковых дисперсиях случайных ошибок в различных наблюдениях, т. е. о присутствии в модели условия гетероскедастичности:

$$H_0 / G_1^2 \neq G_2^2 \neq \ldots \neq G_n^2.$$

1. Обнаружение гетероскедастичности

Существует несколько тестов на обнаружение гетероскедастичности в регрессионной модели.

Тест Глейзера

На первом этапе строится обычная **регрессионная модель**:

$$y_i = \beta_0 + \beta_1 x_i.$$

Методом наименьших квадратов вычисляются **оценки коэффициентов построенной модели**:

$$\tilde{y}_i = \tilde{\beta}_0 + \tilde{\beta}_1 \tilde{x}_i.$$

На следующем этапе вычисляются **остатки регрессионной модели**:

$$e_i = y_i - \tilde{y}_i = y_i - \tilde{\beta}_0 - \tilde{\beta}_1 x_i.$$

Полученные регрессионные остатки возводятся в квадрат: e_i^2.

С целью обнаружения гетероскедастичности определяется **коэффициент Спирмена** между регрессионными остатками e_i^2 и независимой переменной x_i.

Коэффициент Спирмена является аналогом парного коэффициента корреляции, но позволяет выявить взаимосвязь между качественным и количественным признаками. Зависимой переменной выступает e_i^2, в качестве независимой — x_i. Переменная x_i ранжируется и располагается по возрастанию. Ранги обозначаются как R_x. Далее проставляются ранги переменной, обозначаемые как R_e.

Коэффициент Спирмена рассчитывается по формуле:

$$K_{Спир} = 1 - \frac{6\sum d}{n(n^2 - 1)},$$

где d — ранговая разность ($R_x - R_e$);

n — количество пар вариантов.

Значимость коэффициента Спирмена проверяется с помощью t-критерия Стьюдента при основной гипотезе об отсутствии связи между переменными.

Значение t-критерия определяется как:

$$t_{набл} = \frac{K_{Спир}}{\sqrt{1 - K_{Спир}^2}} \times (n-2),$$

а критическое значение—по таблице распределения Стьюдента: $t_{крит}(\alpha; n-2)$.

Если $|t_{набл}| > t_{крит}$, то основная гипотеза отклоняется, и между переменной x_i и остатками регрессионной модели e_i^2 существует взаимосвязь, т. е. в модели присутствует гетероскедастичность.

Если $|t_{набл}| \le t_{крит}$, то основная гипотеза принимается, и в модели парной регрессии гетероскедастичность отсутствует.

Для модели множественной регрессии вывод может быть следующий: гетероскедастичность не зависит от выбранной переменной x_{ik}.

Тест Голдфелда—Квандта

Этот тест исходит из предположения о нормальном законе распределения случайной ошибки ε_i.

В модели множественной регрессии выбирается переменная x_{ik}, от которой могут зависеть остатки модели e_i. Значения x_{ik} ранжируются ($i = \overline{1,n}$), располагаются по возрастанию и делятся на три части.

Для первой и третьей частей строятся две независимые регрессионные модели:

$$y_i^1 = \beta_0^1 + \beta_2^1 x_i^1,$$

где $i = \overline{1,n^I}$;

$$y_i^3 = \beta_0^3 + \beta_2^3 x_i^3,$$

где $i = \overline{n^{II}+1,n}$.

По каждой из построенных регрессий **находятся суммы квадратов остатков**:

$$ESS^I = \sum_{i=1}^{n^I} e_i^2 = \sum_{i=1}^{n^I} (y_i - \tilde{y}_i)^2;$$

$$ESS^{III} = \sum_{n^{II}+1}^{n} e_i^2 = \sum_{n^{II}+1}^{n} (y_i - \tilde{y}_i)^2.$$

Далее проверяется основная гипотеза об отсутствии гетероскедастичности в регрессионной модели через F-критерий Фишера.

Значение F-критерия находят по формуле:

$$F_{набл} = \frac{ESS^{III}}{ESS^{I}}, \text{ если } ESS^{III} > ESS^{I},$$

или

$$F_{набл} = \frac{ESS^{I}}{ESS^{III}}, \text{ если } ESS^{I} > ESS^{III}.$$

Критическое значение F-критерия находят по таблице распределения Фишера - Снедекора с уровнем значимости α и двумя степенями свободы: $k_1 = n^I - l$ и $k_2 = n^I - l$, где l — количество оцениваемых параметров в регрессионной модели.

Если $F_{набл} > F_{крит}$, то основная гипотеза отклоняется, в регрессионной модели присутствует гетероскедастичность, зависящая от переменной x_{ik}.

Если $F_{набл} < F_{крит}$, то основная гипотеза принимается, и гетероскедастичность в модели множественной регрессии не зависит от переменной x_{ik}. Необходимо проверить и другие независимые переменные, если есть предположение об их тесной связи с $G^2(\varepsilon_i)$. Для модели парной регрессии данный вывод означает, что модель гомоскедастична.

Кроме этих тестов на гетероскедастичность, существуют также тесты Бреуша—Пагана, Уайта и др.

2. Устранение гетероскедастичности

Наиболее простым методом устранения гетероскедастичности является взвешивание параметров регрессионной модели. **Суть метода** состоит в том, что отдельным наблюдениям независимой переменой с максимальным среднеквадратическим отклонением случайной ошибки придается больший вес, а остальным наблюдениям с минимальным среднеквадратическим отклонением случайной ошибки придается меньший вес. Благодаря этому оценки коэффициентов уравнения регрессии остаются эффективными.

Модель регрессии при таком подходе называется взвешенной регрессией с весами $\frac{1}{G(\varepsilon_i)}$.

Рассмотрим процесс взвешивания для линейной модели парной регрессии, в которой доказано наличие гетероскедастичности:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i.$$

$$G^2(\varepsilon_i) \neq G^2(\varepsilon_j),$$

где $i \neq j$.

Разделим регрессионное уравнение на среднеквадратическое отклонение случайной ошибки $G(\varepsilon_i)$:

$$\frac{y_1}{G(\varepsilon_1)} = \frac{\beta_0}{G(\varepsilon_1)} + \frac{\beta_1 x_1}{G(\varepsilon_1)} + \frac{\varepsilon_1}{G(\varepsilon_1)};$$

$$\frac{y_2}{G(\varepsilon_2)} = \frac{\beta_0}{G(\varepsilon_2)} + \frac{\beta_1 x_2}{G(\varepsilon_2)} + \frac{\varepsilon_2}{G(\varepsilon_2)}$$

и т. д.

Процесс взвешивания для модели парной регрессии можно записать так:

$$\frac{y_i}{G(\varepsilon_i)} = \frac{\beta_0}{G(\varepsilon_i)} + \frac{\beta_1 x_i}{G(\varepsilon_i)} + \frac{\varepsilon_i}{G(\varepsilon_i)}, \quad i = \overline{1, n}.$$

Данное уравнение записывают в линейном виде с помощью метода замен. Введем обозначения:

$$w_i = \frac{y_i}{G(\varepsilon_i)}; \; z_i = \frac{x_i}{G(\varepsilon_i)}; \; v_i = \frac{1}{G(\varepsilon_i)}; \; V_i = \frac{\varepsilon_i}{G(\varepsilon_i)}.$$

Уравнение регрессии записывают в преобразованном виде:

$$w_i = \beta_0 \times v_i + \beta_1 \times z_i + V_i.$$

Эта регрессионная модель является моделью с двумя факторными переменными — v_i и z_i.

Дисперсию случайной ошибки взвешенной регрессионной модели можно рассчитать следующим образом:

$$D(V_i) = D\left(\frac{\varepsilon_i}{G(\varepsilon_i)}\right) = \frac{D(\varepsilon_i)}{G^2(\varepsilon_i)} = 1,$$

что говорит о постоянстве дисперсий случайных ошибок преобразованной регрессионной модели, т. е. о присутствии условия гомоскедастичности.

Основной проблемой рассмотренного подхода к устранению гетероскедастичности является необходимость априорного знания среднеквадратических отклонений случайных ошибок регрессионной модели. Такое условие в реальности практически невыполнимо, приходится прибегать к другим методам коррекции гетероскедастичности.

Методы коррекции гетероскедастичности сводятся к нахождению оценки ковариационной матрицы случайных ошибок регрессионной модели:

$$Cov(\varepsilon_i) = \Omega = \begin{pmatrix} G_1^2 & 0 & \cdots & 0 \\ 0 & G_2^2 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & G_n^2 \end{pmatrix},$$

где $G_1{}^2 \neq G_2{}^2 \neq ... \neq G_n{}^2$.

Оценки $\tilde{G}^2(e_i)$ находят с помощью метода Бреуша—Пагана:

1) на основании уравнения регрессии находятся остатки e_i и сумма квадратов остатков

$$\sum_{i=1}^{n} e_i^2;$$

2) оценкой дисперсии остатков регрессионной модели будет величина:

$$\tilde{G}^2(e_i) = \frac{1}{n}\sum_{i=1}^{n} e_i^2;$$

3) строится взвешенная регрессия, где весами является оценка дисперсии остатков регрессионной модели

$$\frac{1}{\tilde{G}^2(e_i)}\$;$$

4) если взвешенное уравнение регрессии получается незначимым, то и оценки матрицы ковариаций Ω являются неточными.

После нахождения оценок дисперсий остатков можно воспользоваться доступным обобщенным или взвешенным методом наименьших квадратов для вычисления оценок коэффициентов уравнения регрессии, которые различаются лишь оценкой $\tilde{\Omega}$.

Если нельзя выполнить коррекцию гетероскедастичности, то вполне возможно вычислить оценки коэффициентов уравнения регрессии по обычному МНК, но корректировать ковариационную матрицу оценок коэффициентов $Cov(\tilde{\beta}_i)$, так как условие гетероскедастичности приводит к увеличению данной матрицы. Корректировка $Cov(\tilde{\beta}_i)$ методом Уайта:

$$Cov\left(\tilde{\beta}\right) = N\left(X^T X\right)^{-1}\left(\frac{1}{N}\times\sum_{I=1}^{N} e_i^2 x_i x_i^T\right)\left(X^T X\right)^{-1},$$

где N — количество наблюдений;

X — матрица независимых переменных;

e_i^2 — квадрат остатков регрессионной модели;

x_i^T — транспонированная i-тая строка матрицы данных X.

Корректировка приводит к изменению t-статистики и доверительных интервалов для коэффициентов регрессии.

ЛЕКЦИЯ № 18. Автокорреляция остатков регрессионной модели, ее устранение. Критерий Дарбина—Уотсона. Метод Кохрана—Оркутта и Хилдрета—Лу

Корреляция, возникающая между уровнями изучаемой переменной, называется **автокорреляцией**. Явление автокорреляции в основном присуще данным, представленным в виде временных рядов.

Автокорреляция остатков регрессионной модели e_i (или случайных ошибок уравнения регрессии ε_t) — корреляционная зависимость между настоящими и прошлыми значениями остатков.

Величина сдвига между рядами остатков называется временным лагом. Значение временного лага определяет порядок коэффициента автокорреляции. Если существует корреляционная зависимость между остатками e_n и e_{n-1}, то величина временного лага равняется. Данную зависимость будет характеризовать коэффициент автокорреляции первого порядка между рядами остатков $e_1, ..., e_{n-1}$ и $e_2, ..., e_n$.

Нормальная линейная модель регрессии строится исходя из предположения, что случайные ошибки уравнения регрессии не коррелированы между собой, т. е. ковариация случайных ошибок любых двух разных наблюдений равна нулю: $Cov(e_i, e_j) = E(e_i e_j) = 0$, где $i \neq j$. Явление автокорреляции остатков регрессионной модели нарушает эту предпосылку, что приводит к необходимости устранения корреляционной зависимости между случайными ошибками модели.

Наличие процесса автокорреляции остатков в регрессионной модели приводит практически к тем же последствиям, что и проблема гетероскедастичности остатков:

1) оценки уравнения нормальной линейной регрессии остаются несмещенными и состоятельными, но при этом теряется эффективность;

2) появляется большая вероятность того, что оценки стандартных ошибок коэффициентов регрессионной модели будут рассчитаны неверно, что конечном итоге может привести

к утверждению неверной гипотезы о значимости регрессионных коэффициентов и значимости уравнения регрессии в целом.

Наиболее простым и распространенным методом обнаружения автокорреляции случайных остатков регрессионной модели является графический метод, сутью которого является построение графиков **автокорреляционной и частной автокорреляционной функций (АКФ И ЧАКФ)**.

АКФ — это функция оценки коэффициента автокорреляции в зависимости от величины временного лага между исследуемыми рядами. **Графиком АКФ** является коррелограмма. Коррелограмма отражает численно и графически АКФ или коэффициенты корреляции (и их стандартные ошибки) для последовательности лагов из определенного диапазона (например, от 1 до 15). По оси откладываются значения τ (тау) — величины сдвига между рядами остатков. Значение τ совпадает с порядком автокорреляционного коэффициента. На коррелограмме отмечается диапазон в размере двух стандартных ошибок на каждом лаге. **ЧАКФ** представляет собой углубленное понятие обычной АКФ. В ЧАКФ устраняется корреляционная зависимость между наблюдениями внутри лагов. Частная автокорреляция на данном лаге отличается от обычной автокорреляции на величину удаленных автокорреляций с меньшими временными лагами. ЧАКФ дает более точную картину автокорреляционных зависимостей внутри временного ряда.

1. Критерий Дарбина-Уотсона

Критерий Дарбина—Уотсона является одним из методов обнаружения автокорреляции остатков регрессионной модели. Этот критерий применяется только для обнаружения автокорреляции первого порядка между соседними рядами случайных остатков.

Для модели множественной регрессии вида:

$$Y = X\beta + \varepsilon_t,$$

ошибка, порожденная автокорреляцией первого порядка, определяется как:

$$\varepsilon_t = \rho\varepsilon_{t-1} + v_t,$$

где ρ — коэффициент автокорреляции, $|\rho| < 1$;

ν_t — независимые, одинаково распределенные случайные величины с нулевым математическим ожиданием и дисперсией $G^2(\nu_t)$.

При проверке значимости автокорреляционного коэффициента первого порядка выдвигается основная гипотеза:

$$H_0 \,/\, \rho_1 = 0.$$

Альтернативной гипотезой является утверждение о значимости коэффициента автокорреляции:

$$H_1 \,/\, \rho_1 = 0.$$

Проверка значимости коэффициента автокорреляции осуществляется с помощью критерия Дарбина-Уотсона.

Наблюдаемое значение $d_{набл}$ критерия Дарбина-Уотсона вычисляется по формуле:

$$d_{набл} = \frac{\sum_{t=2}^{T} \left(e_t - e_{t-1}\right)^2}{\sum_{t=1}^{T} e_t^2},$$

где e_t — остатки регрессионной модели в наблюдении t, определяемые из уравнения регрессии:

$$e_t = y_t - \tilde{y}_t = y_t - \tilde{\beta}_0 - \tilde{\beta}_1 x_{1t} - \ldots - \beta_n x_{nt};$$

e_{t-1} — остатки регрессионной модели в наблюдении $t - 1$, определяемые как:

$$e_{t-1} = y_{t-1} - \tilde{y}_{t-1} = y_{t-1} - \tilde{\beta}_0 - \tilde{\beta}_1 x_{1t-1} - \ldots \tilde{\beta}_n x_{nt-1}.$$

Для определения критического значения критерия Дарбина-Уотсона существуют специальные таблицы, в которых указаны значения верхней d_1 и нижней d_2 границы критерия. Такие границы рассчитываются на основании объема выборки и числа степеней свободы $(h - 1)$, где h — количество оцениваемых по выборке параметров.

Приближенное значение величины критерия Дарбина-Уотсона можно вычислить по формуле:

$$d_{набл} \approx 2\left(1 - r_1\right),$$

где r_1 — выборочный автокорреляционный коэффициент первого порядка.

В зависимости от его величины наблюдаемое значение критерия Дарбина—Уотсона определяется следующим образом:

1) если $r_1 = 0$, то $d_{набл} = 2$;

2) если $r_1 = +1$, то $d_{набл} = 0$;

3) если $r_1 = -1$, то $d_{набл} = 4$.

Если наблюдаемое значение критерия Дарбина-Уотсона меньше критического значения его нижней границы, т. е. $d_{набл} < d_1$, то основная гипотеза об отсутствии автокорреляции остатков отклоняется.

Если наблюдаемое значение критерия Дарбина-Уотсона больше критического значения его верхней границы, т. е. $d_{набл} > d_2$, то основная гипотеза о независимости регрессионных остатков принимается.

Если наблюдаемое значение критерия Дарбина-Уотсона находится между верхней и нижней критическими границами, т. е. $d_1 < d_{набл} < d_2$,

то достаточных оснований для принятия единственно правильного решения нет, необходимы дополнительные исследования.

Рассмотренные варианты используются в случае, когда коэффициент автокорреляции остатков является положительной величиной. Если коэффициент автокорреляции — величина отрицательная, то пользуются следующими правилами.

Если наблюдаемое значение критерия Дарбина-Уотсона больше критической величины $4 - d_1$, т. е. $d_{набл} > 4 - d_1$, то основная гипотеза о независимости регрессионных остатков отклоняется.

Если наблюдаемое значение критерия Дарбина-Уотсона меньше критической величины $4 - d_2$, т. е. $d_{набл} < 4 - d_2$, то основная гипотеза принимается, автокорреляция остатков регрессионной модели отсутствует.

Если наблюдаемое значение критерия Дарбина-Уотсона находится в критическом интервале между величинами $4 - d_1$ и $4 - d_2$, т. е. $4 - d_1 < d_{набл} < 4 - d_2$, то достаточных оснований для принятия единственно правильного решения нет, необходимы дополнительные исследования.

2. Устранение автокорреляции остатков регрессионной модели

Устранение проблемы автокорреляции остатков является необходимым этапом в оценивании регрессионной модели в связи

с теми негативными последствиями, к которым может привести корреляционная зависимость между значениями случайных ошибок.

Один из наиболее простых методов борьбы с автокорреляцией остатков — это **включение в регрессионную модель автокорреляционного параметра,** но на практике данный подход реализовать невозможно, так как оценка коэффициента автокорреляции — величина заранее неизвестная.

Иным методом устранения автокорреляции первого порядка между соседними членами остаточного ряда в линейных регрессионных моделях либо моделях, сводящихся к линейным, является **авторегрессионная схема первого порядка**. Однако для ее применения необходимо знать величину коэффициента автокорреляции. Так как величина данного коэффициента на практике неизвестна, то в качестве его оценки используется выборочный автокорреляционный коэффициент остатков первого порядка ρ_1, который определяется по формуле:

$$\rho_1 = r_1 = \frac{\sum_{t=2}^{T} e_t e_{t-1}}{\sum_{t=2}^{T} e_t^2}.$$

Коэффициент автокорреляции порядка *l* вычисляется по общей формуле:

$$r_l = \frac{\frac{1}{T-l} \times \sum_{t=1}^{T-l} (x_t - \bar{x}) \times (x_{t+l} - \bar{x})}{\frac{1}{T} \times \sum_{t=1}^{T} (x_t - \bar{x})^2},$$

где l — временной лаг;

T — число наблюдений;

t — момент времени, в который осуществлялось наблюдение;

$\bar{x}$ — среднее значение фактического динамического ряда.

Рассмотрим применение авторегрессионной модели первого порядка на примере парной регрессионной модели вида:

$$y_t = \beta_0 + \beta_1 x_t + \varepsilon_t. \quad (1)$$

С учетом процесса автокорреляции остатков первого порядка данную регрессионную модель можно представить следующим

образом:

$$y_t = \beta_0 + \beta_1 x_t + \rho\varepsilon_{t-1} + v_t,$$

$$\varepsilon_t = \rho\varepsilon_{t-1} + v_t$$

где ρ — коэффициент автокорреляции, $|\rho| < 1$;

v_t — независимые, одинаково распределенные случайные величины с нулевым математическим ожиданием и дисперсией $G^2(v_i)$.

Регрессионное уравнение (1) в предыдущий момент времени $(t-1)$определялось как:

$$y_{t-1} = \beta_0 + \beta_1 x_{t-1} + \varepsilon_{t-1}. \quad (2)$$

Если второе регрессионное уравнение в момент времени $(t-1)$ умножить на величину ρ и вычесть его из первого регрессионного уравнения, то получим преобразованное регрессионное **уравнение с учетом автокорреляции первого порядка**:

$$y_t = \beta_0 + \beta_1 x_t + \rho\varepsilon_{t-1} + v_t - \underline{\rho y_{t-1} = \rho\beta_0 + \rho\beta_1 x_{t-1} + \rho\varepsilon_{t-1}},$$

$$y_t - \rho y_{t-1} = \beta_0(1-\rho) + \beta_1(x_t - \rho x_{t-1}) + v_t. \quad (3)$$

К преобразованному регрессионному уравнению применим метод замен. Пусть

$$Y_t = y_t - \rho y_{t-1};$$
$$X_t = x_t - \rho x_{t-1};$$
$$Z_t = 1 - \rho.$$

С учетом замен регрессионное уравнение (3) может быть записано следующим образом:

$$Y_t = Z_t \times \beta_0 + \beta_1 \times X_t + v_t. \quad (4)$$

Случайная ошибка v_t преобразованной формы не подвержена процессу автокорреляции, поэтому в регрессионном уравнении (4) корреляционная зависимость остатков устранена.

Подобное преобразование методом разностей можно применить ко всем строкам матрицы данных X, кроме первого наблюдения.

Если не вычислять Y_1 и X_1, то подобная потеря в небольшой выборке может привести к неэффективности оценок коэффициентов регрессии преобразованного уравнения.

Для решения данной проблемы применяется поправка Прайса-Уинстена:

$$Y_1 = \sqrt{1-\rho^2} \times y_1;$$
$$X_1 = \sqrt{1-\rho^2} \times x_1;$$
$$Z_1 = \sqrt{1-\rho^2}.$$

Оценки неизвестных коэффициентов регрессионного уравнения (4) вычисляются с помощью классического метода наименьших квадратов:

$$\tilde{Y}_t = \tilde{b}_0 + \tilde{b}_1 \times X_t.$$

Оценки коэффициентов регрессии исходного уравнения (1) определяются следующим образом:

$$\tilde{\beta}_0 = \frac{\tilde{b}_0}{(1-\rho)};$$
$$\tilde{\beta}_1 = \tilde{b}_1.$$

Тогда конечное регрессионное уравнение можно записать как:

$$\tilde{y}_t = \tilde{\beta}_0 + \tilde{\beta}_1 \times x_t.$$

3. Метод Кохрана-Оркутта. Метод Хилдрета-Лу

Оценку автокорреляционного коэффициента ρ можно найти через формулу выборочного коэффициента автокорреляции остатков. Однако есть и другой способ оценки величины ρ, который носит название **метода Кохрана-Оркутта**. Рассмотрим применение данного метода на основе парной регрессионной модели вида:

$$y_t = \beta_0 + \beta_1 x_t + \varepsilon_t.$$

Алгоритм метода Кохрана-Оркутта реализуется в несколько этапов.

1.На первом этапе исходное регрессионное уравнение $y_t = \beta_0 + \beta_1 x_t + \varepsilon_t$ (1) оценивается традиционным методом наименьших квадратов: $\tilde{y}_t^{(1)} = \tilde{\beta}_0 + \tilde{\beta}_1 \times x_t$ (2).

2. На основании исходного (1) и оцененного (2) уравнений регрессии вычисляются остатки модели: $e_t = y_t - \tilde{y}_t$, где $t = \overline{1,T}$.

3. На третьем этапе вычисляется выборочный автокорреляционный коэффициент первого порядка по формуле:

$$r_1 = \frac{\sum_{t=2}^{T} e_t e_{t-1}}{\sum_{t=2}^{T} e_t^2}.$$

Данный коэффициент позволяет оценить авторегрессионную зависимость остатков: $\tilde{e}_t = \rho_1 \times e_{t-1}$, где $\rho_1 = r_1$.

4. Строится преобразованное уравнение регрессии. Исходное регрессионное уравнение в предыдущий момент времени $(t - 1)$ определялось как: $y_{t-1} = \beta_0 + \beta_1 x_{t-1} + \varepsilon_{t-1}$ (3). Если регрессионное уравнение (3) в момент времени $(t - 1)$ умножить на величину ρ и вычесть его из регрессионного уравнения (1), то получим преобразованное регрессионное уравнение с учетом автокорреляции первого порядка:

$$y_t - \rho y_{t-1} = \beta_0 (1 - \rho) + \beta_1 (x_t - \rho x_{t-1}) + v_t.$$

Воспользовавшись методом замен, приведем преобразованное уравнение к виду

$$Y_t = Z_t \times \beta_0 + \beta_1 \times X_t + v_t, \qquad (4)$$

где $Y_t = y_t - \rho y_{t-1}$;
$X_t = x_t - \rho x_{t-1}$;
$Z_t = 1 - \rho$.

5. Оценки неизвестных коэффициентов преобразованного уравнения регрессии вычисляются традиционным методом наименьших квадратов:

$$\tilde{Y}_t = \tilde{b}_0 + \tilde{b}_1 \times X_t. \qquad (5)$$

Далее определяются оценки коэффициентов регрессии исходного уравнения следующим образом:

$$\tilde{\beta}_0 = \frac{\tilde{b}_0}{(1 - \rho)}; \quad \tilde{\beta}_1 = \tilde{b}_1.$$

Конечное регрессионное уравнение можно записать как:

$$\tilde{y}_t^{(2)} = \tilde{\beta}_0 + \tilde{\beta}_1 \times x_t. \qquad (6)$$

6. На последнем этапе вновь вычисляются регрессионные остатки e_t между исходным (1) и преобразованным оцененным (6) уравнениями регрессии, и процесс повторяется с третьего этапа.

Метод Кохрана-Оркутта относится к итеративным методам оценивания. Процесс итеративного оценивания исходного регрессионного уравнения сходится или останавливается при условии, если вновь вычисленное значение оценки автокорреляционного коэффициента первого порядка ρ_1 почти не отличается от своего предыдущего значения. Метод Кохрана-Оркутта характеризуется достаточно быстрой сходимостью.

Более трудоемким по сравнению с предыдущими приемами оценки автокорреляционного коэффициента является **метод Хилдрета-Лу**.

Коэффициент автокорреляции задается в данном случае двумя параметрами: диапазоном и величиной шага. Например, ρ_1 заключается в пределах $[-1;+1]$. Его значения определяются исходя из шага 0,05.

Для каждого из значений коэффициента автокорреляции с помощью метода разностей строится преобразованное регрессионное уравнение вида:

$$Y_t = Z_t \times \beta_0 + \beta_1 \times X_t + v_t,$$

где $Y_t = y_t - ry_{t-1}$;

$X_t = x_t - rx_{t-1}$;

$Z_t = 1 - r$.

Оценки неизвестных коэффициентов преобразованного уравнения регрессии вычисляются традиционным методом наименьших квадратов:

$$\tilde{Y}_t = \tilde{b}_0 + \tilde{b}_1 \times X_t.$$

То из значений коэффициента автокорреляции, с помощью которого вычисляется минимальная сумма квадратов отклонений теоретических значений от расчетных значений (на основании преобразованного регрессионного уравнения), является оценкой коэффициента автокорреляции ρ_1. Далее вычисляются **оценки конечного уравнения регрессии** по формулам:

$$\tilde{\beta}_0 = \frac{\tilde{b}_0}{(1-\rho)}; \quad \tilde{\beta}_1 = \tilde{b}_1.$$

ЛЕКЦИЯ № 19. Обобщенный метод наименьших квадратов. Регрессионные модели с переменной структурой. Фиктивные переменные. Метод Чоу

Если в линейной регрессионной модели случайные ошибки подвержены явлениям гетероскедастичности или автокорреляции, то оценки коэффициентов регрессионного уравнения, полученные с помощью традиционного метода наименьших квадратов, не будут удовлетворять основным статистическим свойствам.

Характеристики состоятельности и несмещенности оценки сохранятся, однако свойство эффективности в этом случае утрачивается, т. е. не выполняется теорема Гаусса-Маркова.

Состоятельные, несмещенные и эффективные оценки коэффициентов регрессионной модели с гетероскедастичными или коррелированными случайными ошибками определяются с помощью обобщенного метода наименьших квадратов (ОМНК).

Нормальная линейная регрессионная модель строится на основании следующих предпосылок о случайных ошибках:

1) дисперсия случайной ошибки уравнения регрессии является величиной, постоянной для всех наблюдений:

$$D(\varepsilon_i) = \mathrm{E}(\varepsilon_i^2) = G^2 = \mathrm{const};$$

2) случайные ошибки уравнения регрессии не коррелированы между собой, т. е. ковариация случайных ошибок любых двух разных наблюдений равна нулю:

$$Cov(\varepsilon_i, \varepsilon_j) = \mathrm{E}(\varepsilon_i \varepsilon_j) = 0, \text{ где } i \neq j.$$

В случае гетероскедастичности остатков нарушается первое из перечисленных свойств $D(\varepsilon_i) \neq D(\varepsilon_j) \neq G^2 \neq \mathrm{const}$, где $i \neq j$, а в случае автокорреляции остатков нарушается второе свойство $Cov(\varepsilon_i, \varepsilon_j) \neq E(\varepsilon_i \varepsilon_j) \neq 0$. Регрессионная модель, для которой не выполняются указанные свойства, называется обобщенной линейной регрессионной моделью.

В матричном виде обобщенную линейную регрессию можно записать как:

$$Y = \beta \times X + \varepsilon,$$

где X — неслучайная матрица факторных переменных;

ε — случайная ошибка регрессионной модели с нулевым математическим ожиданием $E(\varepsilon) = 0$ и дисперсией $G^2(\varepsilon)\Omega$, $\varepsilon \sim N(0; G^2\Omega)$;

Ω — ковариационная матрица случайных ошибок обобщенного регрессионного уравнения.

Для нормальной линейной регрессионной модели дисперсия случайной ошибки определялась из условия постоянства дисперсий случайных ошибок:

$$\Sigma_{\varepsilon} = \begin{pmatrix} G^2 & 0 & \cdots & 0 \\ 0 & G^2 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & G^2 \end{pmatrix} = G^2 \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & 1 \end{pmatrix} = G^2 \mathrm{I}n,$$

где $G^2 = const$ — дисперсия случайной ошибки уравнения регрессии ε;

In — единичная матрица размерности $n \times n$.

В обобщенной регрессионной модели ковариационная матрица случайных ошибок строится исходя из условия непостоянства дисперсий регрессионных остатков $D(\varepsilon_i) \neq D(\varepsilon_j) \neq G^2 \neq \text{const}$:

$$Cov(\varepsilon_i) = \Omega = \begin{pmatrix} G_1^2 & 0 & \cdots & 0 \\ 0 & G_2^2 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & G_n^2 \end{pmatrix}.$$

Теорема Айткена. В классе линейных несмещенных оценок неизвестных коэффициентов обобщенной регрессионной модели оценка

$$\tilde{\beta}_{ОМНК} = \left(X^T \Omega^{-1} X\right)^{-1} X^T \Omega^{-1} Y$$

будет иметь наименьшую ковариационную матрицу.

Формула для расчета матрицы ковариаций ОМНК-оценок **коэффициентов обобщенной регрессии**:

$$Cov(\tilde{\beta}) = G^2(\varepsilon)\left(X^T \Omega^{-1} X\right)^{-1}$$

Величину $G^2(\varepsilon)$ необходимо оценить для определения матрицы ковариаций ОМНК-оценок по формуле:

$$S_\varepsilon^2 = \frac{1}{n-h}\left(Y - X \times \tilde{\beta}_{ОМНК}\right)^T \times \Omega^{-1}\left(Y - X \times \tilde{\beta}_{ОМНК}\right).$$

Значение $G^2(\varepsilon)$ не является дисперсией случайной ошибки регрессионного уравнения.

В оценке качества обобщенной регрессионной линейной модели коэффициент детерминации использовать нельзя, так как он не отвечает требованиям, предъявляемым к обычному множественному коэффициенту детерминации.

Для проверки гипотез значимости коэффициентов обобщенного нормального уравнения регрессии и регрессионной модели применяются те же статистические критерии, что в случае нормальной линейной регрессионной модели.

1. Доступный обобщенный метод наименьших квадратов

Главное отличие доступного обобщенного метода наименьших квадратов от обобщенного метода состоит в оценке ковариационной матрицы случайных ошибок Ω обобщенной регрессионной модели. В случае автокоррелированности остатков регрессионной модели для определения оценок неизвестных коэффициентов используется именно доступный обобщенный метод наименьших квадратов (ДОМНК или FGLS).

Оценки неизвестных коэффициентов обобщенной регрессионной модели находятся с помощью FGLS по формуле:

$$\tilde{\beta}_{FGLS} = \left(X^T \tilde{\Omega}^{-1} X\right)^{-1} X^T \tilde{\Omega}^{-1} Y,$$

где $\tilde{\Omega}$ — оценка матрицы ковариаций случайных ошибок обобщенной регрессии.

Оценивание матрицы ковариаций случайных ошибок в модели с автокоррелированными, но гомоскедастичными остатками рассмотрим на примере модели парной регрессии:

$$y_t = \beta_0 + \beta_1 x_t + \varepsilon_t.$$

Исходя из предположения, что остатки данной регрессионной модели подчиняются авторегрессионному процессу первого порядка, исходную модель можно представить следующим образом:

$$y_t = \beta_0 + \beta_1 x_t + \rho\varepsilon_{t-1} + v_t,$$
$$\varepsilon_t = \rho\varepsilon_{t-1} + v_t,$$

где ρ — коэффициент автокорреляции, $|\rho| < 1$;

v_t — независимые, одинаково распределенные случайные величины с нулевым математическим ожиданием и дисперсией $G^2(v_t)$.

Математическое ожидание случайной ошибки регрессионного уравнения равно нулю:

$$E(\varepsilon_t) = E(\rho\varepsilon_{t-1} + v_t) = \rho E(\varepsilon_{t-1}) + E(v_t) = 0.$$

Предположим, что дисперсия случайной ошибки регрессии определяется как:

$$D(\varepsilon_t) = \frac{G^2(v_t)}{1-\rho^2}.$$

Рассчитаем ковариацию между двумя случайными регрессионными ошибками ε_2 и ε_1.

$$\operatorname{cov}(\varepsilon_2\varepsilon_1) = \mathrm{E}(\varepsilon_2 \times \varepsilon_1) = \mathrm{E}((\rho\varepsilon_2 + v_2) \times \varepsilon_1) =$$
$$= \mathrm{E}(\rho\varepsilon_1^2) + \mathrm{E}(v_2\varepsilon_1) = \rho\frac{G^2(v_t)}{1-\rho^2}.$$

Рассчитаем ковариацию между следующими случайными регрессионными ошибками ε_3 и ε_1:

$$\operatorname{cov}(\varepsilon_3\varepsilon_1) = \mathrm{E}(\varepsilon_3 \times \varepsilon_1) = \mathrm{E}((\rho\varepsilon_2 + v_3) \times \varepsilon_1) =$$
$$= \mathrm{E}((\rho(\rho\varepsilon_1 + v_2) + v_3) \times \varepsilon_1) = \rho^2\mathrm{E}(\varepsilon_1^2) + \rho\mathrm{E}(v_2\varepsilon_1) + \mathrm{E}(v_3\varepsilon_1) =$$
$$= \rho^2\frac{G^2(v_t)}{1-\rho^2}.$$

Дальнейший процесс расчета ковариаций продолжается для всех случайных ошибок обобщенного регрессионного уравнения по тому же принципу.

Тогда корреляционную матрицу остатков обобщенной линейной регрессионной модели можно представить:

$$\Omega = \frac{G^2(v_t)}{1-\rho^2}\begin{pmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{T-1} \\ \rho & 1 & \rho & \cdots & \rho^{T-2} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \rho^{T-1} & \rho^{T-2} & \rho^{T-3} & \cdots & 1 \end{pmatrix}.$$

Величина $G^2(v_t)$ является дисперсией случайной ошибки регрессионного уравнения. Ее выборочную оценку вычисляют по формуле:

$$S^2(v_t) = \frac{\sum_{t=1}^{T} e_t^2}{T-h} = \frac{\sum_{t=1}^{T} (y_t - \tilde{y}_t)^2}{T-h},$$

где T — объем выборочной совокупности;

h — число оцениваемых по выборке параметров.

Если остатки регрессионной модели являются величинами независимыми (неавтокоррелированными), но гетероскедастичными, имеет смысл применение взвешенного метода наименьших квадратов (ВМНК или WLS).

Суть взвешенного метода наименьших квадратов состоит в том, что остаткам обобщенной регрессионной модели придаются определенные веса, которые равны обратным величинам соответствующих дисперсий $G^2(\varepsilon_t)$. На практике значения дисперсий являются величинами неизвестными, но существует предположение, что они пропорциональны значениям факторных переменных x_t. Это свойство используется для вычисления наиболее подходящих весов.

Ковариационная матрица случайных ошибок может определяться исходя из предположения о пропорциональности величины $G^2(\varepsilon_t)$ факторному признаку x_t:

$$x_t = \gamma \times G(\varepsilon_t),$$

где γ — ошибка высказанного предположения или некоторая поправка.

Матрицу ковариаций случайных ошибок регрессии можно представить в виде:

$$\Omega = \begin{pmatrix} \frac{x_t^2}{\gamma^2} & 0 & \cdots & 0 \\ 0 & \frac{x_t^2}{\gamma^2} & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \frac{x_t^2}{\gamma^2} \end{pmatrix} = \frac{1}{\gamma^2} \begin{pmatrix} x_t^2 & 0 & \cdots & 0 \\ 0 & x_t^2 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & x_t^2 \end{pmatrix}.$$

Существует вероятность, что оценки неизвестных коэффициентов, полученные доступным обобщенным или взвешенным методом наименьших квадратов, не будут удовлетворять основным статистическим свойствам: несмещенности, состоятельности и эффективности. Все зависит от точности оценки матрицы ковариаций случайных ошибок регрессии Ω.

2. Регрессионные модели с переменной структурой. Фиктивные переменные

Помимо количественных переменных, включаемых в регрессионную модель, может возникнуть ситуация, когда в модель необходимо включить качественную переменную (например, возраст, образование, пол, расовую принадлежность и др.).

Атрибутивный «или качественный фактор», представленный с помощью определенного цифрового кода, называется **фиктивной переменной** (dummy variable).

Самый распространенный пример применения фиктивных переменных — это проблема разрыва в заработной плате у мужчин и женщин. Допустим, что построена регрессионная зависимость заработной платы рабочих *y* от их возраста *x*:

$$y = \beta_0 + \beta_1 x.$$

Однако данная регрессионная зависимость неспособна в полной мере отразить вариацию результативного признака. Поэтому в модель необходимо ввести дополнительный фактор, например пол.

Это предположение основывается на том, что у мужчин в среднем заработная плата выше, чем у женщин. Так как пол является качественным признаком, то нужно представить данную переменную в виде фиктивной:

$$D = \begin{cases} 1, \textit{муж}, \\ 0, \textit{жен}. \end{cases}$$

Тогда регрессионную модель можно записать с учетом нового фактора:

$$y = \beta_0 + \beta_1 x + \beta_2 \times D,$$

где параметр β_2 будет отражать в среднем разницу в заработной плате у мужчин и женщин.

Регрессионная модель, включающая в качестве фактора (факторов) фиктивную переменную, называется регрессионной моделью с переменной структурой.

Рассмотрим регрессионную зависимость размера заработной платы (y) от стажа работников (x) с различным образованием. Качественная переменная «образование» может принимать три значения: среднее, среднее специальное и высшее. Для того чтобы данный фактор включить в регрессионную модель, необходимо ввести только две фиктивные переменные, потому что их количество должно быть на единицу меньше, чем значения качественной переменной.

Таким образом, переменную «образование» можно представить в виде:

$$D_1 = \begin{cases} 0, \textit{среднее}, \\ 1, \textit{ср. спец.}, \\ 0, \textit{высшее}; \end{cases} \qquad D_2 = \begin{cases} 0, \textit{среднее}, \\ 0, \textit{ср. спец.}, \\ 1, \textit{высшее}. \end{cases}$$

Тогда уравнение регрессии с переменной структурой можно записать как:

$$y = \beta_0 + \beta_1 x + \beta_2 \times D_1 + \beta_3 \times D_2. \qquad \text{(A)}$$

Уравнение регрессии (А) называется моделью регрессии без ограничений (unrestricted regression).

Если в уравнении регрессии все значения фиктивных переменных равны нулю, т. е. $D_1 = D_2 = 0$, то регрессия вида $y = \beta_0 + \beta_1 x_2$

называется базисной моделью или регрессией с ограничениями (restricted regression). В рассматриваемом примере базисная модель регрессии соответствует регрессионной зависимости заработной платы рабочих со средним образованием от стажа работы.

Для модели регрессии без ограничений можно также выделить частные регрессии.

Например, частная регрессионная зависимость заработной платы работников со средним специальным образованием от стажа:

$$y = \beta_0 + \beta_1 x + \beta_2 \times D_1.$$

В этом случае коэффициент β_2 показывает, на сколько большую заработную плату получают рабочие со средним специальным образованием по сравнению с работниками со средним образованием при одинаковом стаже работы.

Частная регрессионная зависимость заработной платы работников с высшим образованием от стажа:

$$y = \beta_0 + \beta_1 x + \beta_3 \times D_2.$$

Коэффициент β_2 показывает, на сколько большую заработную плату получают рабочие с высшим образованием по сравнению с рабочими со средним образованием при одинаковом стаже работы.

Коэффициенты регрессионных моделей с фиктивными переменными оцениваются традиционным МНК.

3. Метод Чоу

Метод Чоу применяется в случае, когда основную выборку можно разделить на части или подвыборки. Регрессии для подвыборок могут оказаться более эффективными, чем общая регрессионная модель.

Будем считать, что общая регрессионная модель — это регрессионная модель без ограничений, которую обозначим через. Отдельными подвыборками будем считать частные (private) случаи регрессионной модели без ограничений:

1) PR_1 — первая подвыборка;
2) PR_2 — вторая подвыборка;
3) $ESS(PR_1)$ — сумма квадратов остатков для первой подвыборки;
4) $ESS(PR_2)$ — сумма квадратов остатков для второй подвыборки;

5) $ESS(UN)$ — сумма квадратов остатков для общей регрессии;

6) $ESS_{PR_1}^{UN}$ — сумма квадратов остатков для наблюдений первой подвыборки в общей регрессионной модели;

7) $ESS_{PR_2}^{UN}$ — сумма квадратов остатков для наблюдений второй подвыборки в общей регрессионной модели.

Для частных регрессионных моделей должны выполняться следующие условия:

$$ESS(PR_1) < ESS_{PR_1}^{UN}; \quad ESS(PR_2) < ESS_{PR_2}^{UN} \text{ или}$$
$$(ESS(PR_1) + ESS(PR_2)) < ESS(UN).$$

Для определения значимости частных регрессионных моделей используется F-критерий Фишера.

Выдвигается гипотеза о том, что качество общей регрессионной модели без ограничений лучше качества частных регрессионных моделей или подвыборок.

Значение F-критерия определяется по формуле:

$$F_{набл} = \frac{(ESS(UN) - ESS(PR_1) - ESS(PR_2))}{m+1} / \frac{(ESS(PR_1) + ESS(PR_2))}{n-2m-2},$$

где $ESS(UN) - ESS(PR_1) - ESS(PR_2)$ — величина, характеризующая улучшение качества модели регрессии после разделения ее на подвыборки;

m — количество факторных переменных (в том числе фиктивных);

n — объем общей выборочной совокупности.

Критическое значение F-критерия Фишера определяется по таблице распределения Фишера-Снедекора в зависимости от уровня значимости α и двух степеней свободы: $k_1 = m + 1$ и $k_2 = n - 2m - 2$.

Если наблюдаемое значение F-критерия больше критического его значения, т. е. $F_{набл} > F_{крит}$, то основная гипотеза отклоняется, и качество частных регрессионных моделей превосходит качество общей модели регрессии.

Если наблюдаемое значение F-критерия меньше критического его значения, т. е. $F_{набл} < F_{крит}$, то основная гипотеза принимается, и разбивать общую регрессию на подвыборки не имеет смысла.

Если проверяется значимость базисной регрессии или регрессии с ограничениями (restricted regression), то выдвигается основная гипотеза вида: $H_0 / \beta_{m+1} = \beta_{m+2} = , ..., = \beta_k = 0$, где

$$y = \underbrace{\beta_0 + \beta_1 x}_{k} + \underbrace{\beta_2 \times D_1 + \beta_3 \times D_2}_{m}.$$

Тогда наблюдаемое значение F-критерия преобразуется к виду:

$$F_{набл} = \frac{ESS_R - ESS_{UR}}{m} / \frac{TSS - ESS_{UR}}{n - k - 1},$$

Критическое значение F-критерия Фишера определяется в зависимости от уровня значимости α и двух степеней свободы: $k_1 = m$ и $k_2 = n - k - 1$.

Если наблюдаемое значение F-критерия больше его критического значения, то основная гипотеза H_0 отклоняется, и в регрессионное уравнение необходимо вводить дополнительные фиктивные переменные, так как качество регрессионной модели с ограничениями выше качества базисной или ограниченной регрессионной модели.

Если наблюдаемое значение F-критерия Фишера меньше его критического значения, то основная гипотеза H_0 принимается, и базисная регрессионная модель является удовлетворительной для изучаемой зависимости между переменными, вводить в уравнение дополнительные фиктивные переменные не имеет смысла.

Условие $(ESS(PR_1) + ESS(PR_2)) = ESS(UN)$ возможно только в том случае, если коэффициенты частных регрессионных моделей и коэффициенты общей модели без ограничений будут одинаковы, но на практике такое совпадение встречается очень редко.

4. Спецификация переменных

Проблема спецификации переменных заключается в отборе наиболее важных факторов при построении регрессионной зависимости.

Неправильная спецификация может привести к следующим результатам:

1) исключение существенных переменных;
2) включение несущественных переменных.

Предположим, что построена некоторая нормальная множественная регрессионная зависимость вида:

$$Y = X\beta + \varepsilon, \qquad (1)$$

которая является базисной или ограниченной (restricted) моделью изучаемой регрессионной зависимости.

Существует и неограниченная модель изучаемой регрессионной зависимости (unrestricted model):

$$Y = X\beta + Z\lambda + \varepsilon, \qquad (2)$$

где Y — вектор зависимых переменных;

X — вектор количественных факторных переменных;

Z — некоторая фиктивная переменная;

β, λ — вектор неизвестных регрессионных коэффициентов, подлежащих оцениванию.

Рассмотрим случай исключения существенных переменных из модели.

Рассчитаем оценку коэффициента β, полученную методом наименьших квадратов, в оцениваемой регрессионной модели с ограничениями (1), при условии, что регрессионная зависимость (2) является значимой.

$$\tilde{\beta} = \left(X^T X\right)^{-1} X^T Y.$$

Подставим в данную формулу вместо Y выражение $X\beta + Z\lambda + \varepsilon$:

$$\tilde{\beta} = \left(X^T X\right)^{-1} X^T \left(X\beta + Z\lambda + \varepsilon\right) = \beta + \left(X^T X\right)^{-1} X^T Z\lambda + \left(X^T X\right)^{-1} X^T \varepsilon.$$

Выясним, обладает ли полученная оценка коэффициента β базисной или ограниченной регрессионной модели свойством несмещенности. С этой **целью найдем математическое ожидание** $\tilde{\beta}$:

$$E\left(\tilde{\beta}\right) = \beta + \underbrace{\left(X^T X\right)^{-1} X^T Z\lambda}_{BIAS},$$

где $BIAS$ — это смещение оценки.

Устранить несмещенность данной оценки коэффициента β невозможно даже при условии увеличения объема выборки.

Оценка коэффициента β базисной регрессионной модели (1) будет являться несмещенной в двух случаях:

1) если коэффициент при фиктивной переменной Z будет ра-

вен нулю:

$$\lambda = 0 \Rightarrow E\left(\tilde{\beta}\right) = \beta;$$

2) при условии, что пропущенные переменные будут ортогонально включены в модель, т. е. $X^T Z = 0$.

Рассмотрим ковариацию оценки коэффициента β для базисной регрессионной модели (1):

$$Cov\left(\tilde{\beta}\right) = G^2 \left(X^T X\right)^{-1}.$$

Ковариационная матрица МНК-оценок принимает такой вид только в том случае, если модель (1) является истинной или значимой.

Рассмотрим случай включения несущественных переменных в модель.

Оценим коэффициенты уравнения регрессии без ограничений (2) при условии, что базовая регрессионная модель (1) является значимой.

Представим регрессионную модель **с ограничениями** (1) в следующем виде:

$$Y = (X, Z) \times \begin{pmatrix} \beta \\ \lambda \end{pmatrix} + \varepsilon.$$

Обозначим через W переменные регрессионной модели (X, Z). Тогда оценку коэффициента регрессионной модели **без ограничений можно записать как**:

$$\tilde{\beta}_\lambda = \left(W^T W\right)^{-1} WY = \left[\begin{pmatrix} X^T \\ Z^T \end{pmatrix} \times (X\,Z)\right]^{-1} \begin{bmatrix} X^T \\ Z^T \end{bmatrix} \times Y =$$

$$= \begin{bmatrix} X^T X & X^T Z \\ Z^T X & Z^T Z \end{bmatrix} \times \begin{bmatrix} X^T Y \\ Z^T Y \end{bmatrix}.$$

Выясним, обладает ли полученная оценка коэффициента β регрессионной модели без ограничений свойством несмещенности.

С этой целью найдем математическое ожидание $\tilde{\beta}_\lambda$:

$$E\left(\tilde{\beta}_\lambda\right) = E\begin{pmatrix}\beta\\ \lambda\end{pmatrix} = \begin{bmatrix}\beta\\ 0\end{bmatrix}.$$

Таким образом, оценка $\tilde{\beta}_\lambda$ является несмещенной оценкой регрессионного коэффициента β модели (2). Если в данную модель включить один дополнительный фактор, то оценки уже включенных факторов свойства несмещенности не утратят.

Однако если в модель включить много лишних параметров, то точность оценок будет падать.

Рассмотрим ковариационную матрицу МНК-оценок регрессионной модели без ограничений:

$$Cov\begin{pmatrix}\tilde{\beta}\\ \lambda\end{pmatrix} = G^2 \times \begin{bmatrix}X^T X & X^T Z\\ Z^T X & Z^T Z\end{bmatrix}.$$

Ковариационная матрица будет иметь такой вид только в случае истинности или значимости регрессионного уравнения без ограничений.

ЛЕКЦИЯ № 20. Основные компоненты временного ряда. Проверка гипотез о существовании тренда во временном ряду. Метод Чоу проверки стабильности тенденции

Временной ряд — это ряд наблюдаемых значений изучаемого показателя, расположенных в хронологическом порядке или в порядке возрастания времени. Наблюдения $y_t \quad t = \overline{1,n}$, из которых состоит временной ряд, называются уровнями этого ряда. Отдельно взятый временной ряд можно считать выборкой из бесконечного ряда значений показателей во времени.

Если уровень временного ряда фиксирует значение изучаемого показателя на определенный момент времени, то такой ряд называется моментным. Если уровень временного ряда характеризует значение показателя за определенный период времени, то такой ряд является интервальным. Если уровни ряда представлены в виде производных величин (средних или относительных показателей), то такие ряды называются производными.

При изучении временных рядов выделяют **две основные цели**:

1) характеристику структуры ряда;

2) прогнозирование будущих уровней временного ряда на основании прошлых и настоящих уровней.

Для достижения целей необходимо идентифицировать модель временного ряда и описать ее. Идентификация модели предполагает выявление основных компонент, содержащийся в изучаемом временном ряду.

Данные, представленные в виде временных рядов, могут содержать два вида компонент — это систематическая и случайная составляющие.

Систематическая составляющая является результатом воздействия постоянно действующих факторов. Выделяют **три основные систематические компоненты временного ряда**: тренд, сезонность, цикличность.

Тренд представляет собой систематическую линейную или нелинейную компоненту, изменяющуюся во времени.

Сезонность — это периодические колебания уровней временного ряда внутри года.

Цикличность — это периодические колебания, выходящие за рамки 1 года. Промежуток времени между двумя соседними вершинами или впадинами в масштабах года считается длиной цикла.

Систематические составляющие могут одновременно присутствовать во временном ряду.

Случайной составляющей называется случайный шум или ошибка, воздействующая на временной ряд нерегулярно. Основными причинами случайного шума могут выступать факторы резкого и внезапного действия, а также действия текущих факторов.

Шум, порожденный факторами резкого и внезапного действия, называется катастрофическими колебаниями, потому что оказывает наиболее сильное влияние на основную тенденцию временного ряда.

Шум, вызванный действиями текущих факторов, может быть связан также с ошибками наблюдений.

Отдельный уровень временного ряда y_t можно представить в виде функции от основных компонент: $f(T,S,C\,\varepsilon)$, где T — это трендовая компонента, S — сезонная компонента, C — циклическая компонента, ε — случайный шум.

Модель временного ряда может быть представлена в нескольких вариантах.

Аддитивная модель временного ряда, состоящая из слагаемых:

$$y_t = T_t + S_t + C_t + \varepsilon_t.$$

Мультипликативная модель временного ряда, состоящая из сомножителей:

$$y_t = T_t \times S_t \times C_t \times \varepsilon_t.$$

Смешанная модель временного ряда:

$$y_t = T_t \times S_t \times C_t + \varepsilon_t.$$

Большинство методов изучения структуры временного ряда (фильтрации уровней ряда) направлено на выявление и описание систематических или регулярных компонент ряда.

1. Проверка гипотез о существовании тренда во временном ряду

Основная тенденция временного ряда не всегда может быть определена визуально, поэтому применяют специальные критерии проверки гипотезы о существовании тренда в ряду динамики.

2. Гипотеза, основанная на сравнении средних уровней ряда

Временной ряд из N наблюдений разбивают на две равные части y_i объемом n_i $i=\overline{1,n}$ и y_j объемом n_j $\left(j=\overline{n+1;\ N}\right)$. Каждая из этих частей рассматривается как самостоятельная выборочная совокупность с нормальным законом распределения. Для y_i и y_j определяются выборочные характеристики:

$\overline{y}_i=\dfrac{\sum\limits_{i=1}^{n} y_i}{n_i}$ и $\overline{y}_i=\dfrac{\sum\limits_{i=1}^{n} y_i}{n_i}$ — средние арифметические значения;

$S^2{}_i=\dfrac{\sum\limits_{i=1}^{n}\left(y_i-\overline{y}_i\right)}{n_i}$ и $S^2{}_j=\dfrac{\sum\limits_{j=n+1}^{N}\left(y_j-\overline{y}_j\right)}{n_j}$ — выборочные дисперсии.

Проверяется основная гипотеза о равенстве генеральных средних для двух полученных совокупностей:

$$H_0 \,/\, \mu_i=\mu_j;$$
$$H_1 \,/\, \mu_i\neq\mu_j.$$

Основная гипотеза проверяется при справедливости предположения о равенстве генеральных дисперсий:

$$H_0 \,/\, G_i^2=G_j^2;$$
$$H_1 \,/\, G_i^2\neq G_j^2.$$

Гипотеза о равенстве дисперсий проверяется с помощью F-критерия Фишера. Значение F-критерия определяется по формуле:

$$F_{набл}=\frac{S_i^2}{S_j^2}$$ при условии, что $S_i^2>S_j^2$.

Критическое значение F-критерия определяется для уровня значимости α и двух степеней свободы: $k_1 = n - 1$ и $k_2 = N - n - 1$.

Гипотеза принимается, если $F_{\text{набл}} > F_{\text{крит}}$.

Гипотеза о равенстве генеральных средних проверяется с помощью t-критерия Стьюдента. Наблюдаемое значение t-критерия определяется по формуле:

$$t_{набл} = \frac{\bar{y}_i - \bar{y}_j}{\sqrt{S_i^2 (n_i - 1) + S_j^2 (n_j - 1)}} \times \sqrt{\frac{n_i n_j (N-2)}{N}}.$$

Критическое значение t-критерия определяется для уровня значимости α и степени свободы $(N - 2)$.

Если основная гипотеза H_0 отклоняется, то во временном ряду присутствует тренд.

4. Критерий «восходящих и нисходящих» серий

Против каждого из уровней временного ряда объемом N ставится знак «+», если данный уровень больше предыдущего, или знак «−», если уровень меньше предыдущего. Получаем совокупность знаков объемом $(N - 1)$. Последовательность из знаков «+» или «−» называется серией. Обозначим через γ общее количество серий данного временного ряда. Самую длинную серию из каких-либо знаков обозначим через ϕ.

Гипотеза об отсутствии тренда проверяется при уровне значимости $\alpha = 0{,}05$. Если хотя бы одно из следующих неравенств не выполняется, то гипотеза об отсутствии тренда отклоняется:

1) $\gamma > \left[\frac{1}{3}(\alpha N - 1) - 1{,}96 \times \frac{\sqrt{16N - 29}}{90}\right]$;

2) $\phi_{набл} \le \phi_0$,

где $\phi_0 = 5$, если $N < 26$;

$\phi_0 = 6$, если $26 < N < 153$;

$\phi_0 = 7$, если $153 < N < 170$.

5. Критерий серий, основанный на медиане выборки

Временной ряд объемом N ранжируется, т. е. наблюдения упорядочиваются по возрастанию. Определяется медиана ранжированного ряда.

Медиана — наблюдение, делящее ранжированный ряд на две равные части. При нечетном числе наблюдений во временном ряду в качестве медианы принимается значение, стоящее в середине данного ряда. При четном числе наблюдений в качестве медианы берется среднее арифметическое значение двух наблюдений, находящихся посередине временного ряда.

Начальные уровни временного ряда сравниваются с медианой. Если уровень ряда больше медианы, то ему приписывается знак «+», если уровень меньше медианы — знак «–». Определяем γ — общее количество серий исследуемого временного ряда и ϕ — самую длинную серию из знаков «+» или «–».

Основная гипотеза H_0 об отсутствии тренда проверяется при уровне значимости $\alpha = 0{,}05$. Если хотя бы одно из неравенств не выполняется, то гипотеза H_0 об отсутствии тренда в изучаемом динамическом ряде отклоняется.

1. $\phi < \left[3{,}3 \times (\lg N + 1)\right]$;

2. $\gamma > \left[\frac{1}{2}\left(N + 1 - 1{,}96 \times \sqrt{N-1}\right)\right]$.

3. Метод Форстера-Стьюарта проверки гипотез о наличии или отсутствии тренда. Метод Чоу проверки стабильности тенденции

Метод Форстера-Стьюарта одним из наиболее простых и распространенных приемов выявления тренда в одномерном временном ряду.

На первом этапе каждый уровень временного ряда y_t ($t = \overline{1,\ N}$) сравнивается со всеми предыдущими уровнями. На основании результатов сравнений определяются вспомогательные величины:

$$m_t = \begin{cases} 1, y_t > y_{t-1} > \ldots > y_1, \\ 0, \textit{в противном случае}; \end{cases}$$

$$l_t = \begin{cases} 1, y_t < y_{t-1} < \ldots < y_1, \\ 0, \textit{в противном случае}; \end{cases}$$

$$m_t - l_t.$$

Число вспомогательных величин будет $N - 1$. d_t может принимать значения +1, 0, −1.

На втором этапе все значения d_t суммируются, и определяется величина D:

$$D = \sum_{t=2}^{N} d_t.$$

Основная гипотеза об отсутствии тренда в изучаемом временном ряде проверяется с помощью t-критерия Стьюдента. Наблюдаемое значение t-критерия определяется по формуле:

$$t_{набл} = \frac{D}{S_D},$$

где S_D — стандартное отклонение величины D. Значения S_D для временных рядов, длиной от 10 до 100 наблюдений, представлены в специальной таблице.

Критическое значение t-критерия $t_{крит}(\alpha, N - 1)$ определяется по таблице распределения Стьюдента в зависимости от уровня значимости α и числа степеней свободы N-1.

Если $|t_{набл}| \geq t_{крит}$, то основная гипотеза об отсутствии тенденции в исследуемом временном ряде отвергается.

Если $|t_{набл}| < t_{крит}$, то в изучаемом временном ряде тренд отсутствует.

Метод или тест Чоу применяется для проверки гипотезы о стабильности временного ряда. Если ряд имеет нестабильную тенденцию, то с определенного момента времени t^* происходит изменение характера динамики анализируемого показателя под влиянием ряда внешних факторов (например, экономических кризисов, смены экономической политики и др.). Это приводит к изменению параметров уравнения тренда, описывающего данную динамику.

Весь временной ряд можно представить в виде двух подвыборок — до переломного момента t^* и после этого момента.

Введем следующие обозначения.

Будем считать, что весь временной ряд — это регрессионная модель без ограничений (UN). Отдельными подвыборками будем считать частные (private) случаи общей регрессионной модели:
PR_1 — первая подвыборка;
PR_2 — вторая подвыборка;
$ESS(PR_1)$ — сумма квадратов остатков для первой подвыборки;
$ESS(PR_2)$ — сумма квадратов остатков для второй подвыборки;

$ESS(UN)$ — сумма квадратов остатков для общей регрессии;

$ESS_{PR_1}^{UN}$ — сумма квадратов остатков для наблюдений первой подвыборки в общей регрессионной модели;

$ESS_{PR_2}^{UN}$ — сумма квадратов остатков для наблюдений второй подвыборки в общей регрессионной модели.

Для частных регрессионных моделей должны выполняться следующие условия:

$$ESS\left(PR_1\right) < ESS_{PR_1}^{UN};\ \ ESS\left(PR_2\right) < ESS_{PR_2}^{UN},$$

или

$$\left(ESS\left(PR_1\right) + ESS\left(PR_2\right)\right) < ESS\left(UN\right).$$

Выдвигается основная гипотеза о структурной стабильности тенденции общего временного ряда. Для проверки гипотезы используется F-критерий Фишера.

Наблюдаемое **значение F-критерия определяется по формуле**:

$$F_{набл} = \frac{\left(ESS\left(UN\right) - ESS\left(PR_1\right) - ESS\left(PR_2\right)\right)}{m+1} \Big/ \Big/ \frac{\left(ESS\left(PR_1\right) + ESS\left(PR_2\right)\right)}{n-2m-2},$$

где $ESS(UN) - ESS(PR_1) - ESS(PR_2)$ — величина, характеризующая улучшение качества временного ряда после разделения его на две части;

m — число факторных переменных;

n — объем общей выборочной совокупности.

Критическое значение F-критерия определяется по таблице распределения Фишера-Снедекора в зависимости от уровня значимости α и двух степеней свободы: $k_1 = m + 1$ и $k_2 = n - 2m - 2$.

Если $F_{набл} \geq F_{крит}$, то основная гипотеза отклоняется, и временной ряд не имеет общей стабильной тенденции. Иначе временной ряд может быть описан одним трендовым уравнением.

ЛЕКЦИЯ № 21. Представление тренда в аналитическом виде. Проверка адекватности трендовой модели

Основным способом представления тренда в аналитическом виде, используемом в эконометрике, является метод аналитического выравнивания с помощью функций времени или кривых роста. Суть данного метода заключается в аппроксимации временного ряда, определенной формой регрессионной зависимости. При аналитическом выравнивании динамического ряда наиболее проблематичным является вопрос о выборе **функции тренда**.

Выбор выравнивающей кривой может осуществляться на основании заранее заданных критериев. Например, хорошей оценкой качества подобранной формы тренда является множественный коэффициент детерминации. Помимо этого, можно рассчитать сумму квадратов отклонений наблюдаемых значений временного ряда от теоретических значений, рассчитанных с помощью функции тренда.

Если временной ряд содержит равностоящие друг от друга уровни, то одним из методов, позволяющих подобрать подходящую форму кривой, является метод конечных разностей.

Конечной разностью первого порядка (или разностным оператором первого порядка) называется разность между соседними уровнями динамического ряда:

$$\nabla_1^t = y_{t+1} - y_t, \ \left(t = \overline{2,n}\right).$$

Разностным оператором (конечной разностью) второго порядка является разность между соседними разностными операторами первого порядка:

$$\nabla_2^t = \nabla_1^{t+1} - \nabla_1^t = y_{t+1} - y_t - y_t + y_{t-1} = y_{t+1} - 2y_t + y_{t+1}, \ \left(t = \overline{3,n}\right).$$

В общем случае разностный оператор i-го порядка может быть рассчитан как разность между соседними разностными операторами $(i-1)$-го порядка:

$$\nabla_i^t = \nabla_{i-1}^t - \nabla_{i-1}^{t-1}, \ \left(t = \overline{i+n,n}\right).$$

Если разностные операторы первого порядка постоянны и равны между собой $\nabla_1^2 = \nabla_1^3 = , ..., = \nabla_1^n$, а разностные операторы второго порядка равны нулю

$$\nabla_2^3 = \nabla_2^4 = \ldots = \nabla_2^n = 0,$$

то тренд изучаемого динамического ряда можно аппроксимировать линейной зависимостью вида: $y = a + \beta \times t + \varepsilon$.

Если разностные операторы второго порядка постоянны и равны между собой $\nabla_2^3 = \nabla_2^4 = , ..., = \nabla_2^n$, а разностные операторы третьего порядка равны нулю $\nabla_3^4 = \nabla_3^5 = ... = \nabla_3^n = 0$, то тренд изучаемого динамического ряда можно аппроксимироватьпараболической зависимостью второго порядка вида

$$y = a + \beta_1 t + \beta_2 t^2.$$

Порядок разностных операторов, являющихся постоянными для изучаемой временной зависимости, определяет степень уравнения тренда:

$$y = \sum \beta_i \times t^i.$$

Коэффициенты уравнения тренда определяются традиционным методом наименьших квадратов. Если временной ряд содержит линейную тенденцию, то коэффициенты уравнения тренда можно найти также с помощью метода моментов. При этом в уравнение вводится новая переменная времени T, которая началом координат имеет середину динамического ряда. Таким образом, ее сумма по всем элементам равняется нулю.

Для динамического ряда с нечетным количеством уровней переменная $T = 0$ соответствует середине данного ряда. Выше нулевого уровня проставляются числа -1, -2, -3, ..., а ниже данного уровня — числа $+1$, $+2$, $+3$, т. д.

Для динамического ряда с четным количеством уровней числа -1, -2, -3 и далее проставляются до середины ряда, а числа $+1$, $+2$, $+3$ ставятся после середины ряда.

Линейное уравнение регрессии с учетом новой переменной принимает вид:

$$y_t = a + \beta \times T_t + \varepsilon_t.$$

Для нахождения оценок неизвестных коэффициентов данного уравнения составим систему нормальных уравнений:

$$\begin{cases} a \times n = \sum_{t=1}^{n} y_t, \\ \beta \sum_{t=1}^{n} T_t^2 = \sum_{t=1}^{n} y_t \times T_t. \end{cases}$$

Решением системы нормальных уравнений будут являться оценки коэффициентов уравнения тренда:

$$a = \frac{\sum_{t=1}^{n} y_t}{n}; \quad \beta = \frac{\sum_{t=1}^{n} y_t \times T_t}{\sum_{t=1}^{n} T_t^2}.$$

Проверка адекватности трендовой модели

Модель адекватна описываемому процессу, если значения случайной остаточной компоненты ε_t являются случайными центрированными некоррелированными нормально распределенными величинами. Проверка адекватности модели сводится к проверке указанных свойств ряда остатков.

Проверка случайности остатков модели может осуществляться с помощью критериев исследования ряда на предмет наличия в нем тренда (вместо исходных уровней ряда y_1, y_2, ..., y_t используются элементы остаточного ряда e_1, e_2, ..., e_t). Проверка случайности может осуществляться с помощью критерия поворотных точек.

При использовании критерия поворотных точек сравнивают остаток модели e_t с двумя соседними элементами ряда. Если он окажется меньше или больше их, то данная точка является поворотной. В конце сравнений подсчитывается количество m всех поворотных точек. Ряд остатков модели считается случайным, если выполняется условие:

$$m > \left[\frac{2(N-2)}{3 - 2\left(\sqrt{\frac{16N-29}{90}} \right)} \right],$$

где N — объем выборочной совокупности.

Проверка центрированности остатков ряда осуществляется с помощью t-критерия Стьюдента. Наблюдаемое значение t-критерия определяется по формуле:

$$t_{набл} = \frac{|\bar{e}|\sqrt{N}}{G(e)},$$

где $\bar{e} = \frac{\sum_{t=1}^{N} e_t}{N}$ —среднее арифметическое значение остаточного ряда

$G(e) = \sqrt{\frac{\sum_{t=1}^{n} (e_t - \bar{e})^2}{N-1}}$ — среднеквадратическое отклонение остаточного ряда.

Критическое значение t-критерия определяется для уровня значимости $\alpha/2$ и числа степеней свободы $(N - 1)$ по таблице распределения Стьюдента: $t_{крит}(\alpha/2; N - 1)$.

Если $t_{набл} > t_{крит}$, то гипотеза о центрированности ряда остатков отвергается с вероятностью ошибки α. Если $t_{набл} < t_{крит}$, то ряд остатков признается центрированным с вероятностью ошибки $(1 - \alpha)$.

Проверка независимости остатков модели проводится для того, чтобы выявить возможную систематическую составляющую в составе остаточного ряда. Если модель подобрана неудачно, то остатки будут подвержены автокорреляционной зависимости.

Для проверки независимости остатков используется критерий Дарбина-Уотсона, связанный с гипотезой о наличии в ряду остатков автокорреляции первого порядка, т. е. о корреляционной зависимости соседних остатков.

Проверка ряда остатков на нормальность осуществляется с помощью показателей асимметрии и эксцесса (если объем выборочной совокупности не превышает пятидесяти значений). При нормальном распределении показатели асимметрии и эксцесса равны нулю.

На основании выборочных данных строятся эмпирические коэффициенты асимметрии и эксцесса по формулам:

$$K_A = \frac{\frac{1}{N}\sum_{t=1}^{N} e_t^3}{\sqrt{\left(\frac{1}{N}\sum_{t=1}^{N} e_t^2\right)^3}}; \quad K_Э = \frac{\frac{1}{N}\sum_{t=1}^{N} e_t^4}{\left(\frac{1}{N}\sum_{t=1}^{N} e_t^2\right)^2} - 3.$$

Если вычисленные коэффициенты близки к нулю, то имеются основания считать ряд остатков нормально распределенным.

В дополнение к выборочным коэффициентам асимметрии и эксцесса определяют среднеквадратические отклонения коэффициентов:

$$G(A)=\sqrt{\frac{6(N-1)}{(N+1)(N+3)}};\ \ G(A)=\sqrt{\frac{6(N-1)}{(N+1)(N+3)}}.$$

Если одновременно выполняются следующие неравенства:

$$|K_A|\le 1{,}5G(A),$$
$$|K_Э|\le 1{,}5G(Э),$$

то гипотеза о нормальном характере распределения случайной компоненты принимается. Если хотя бы одно из указанных неравенств нарушается, то гипотеза о нормальном распределении остатков отвергается.

Помимо адекватности выбранной модели, необходимо охарактеризовать ее точность. Наиболее простым **критерием точности модели является относительная ошибка**:

$$\omega_{отн}=\frac{1}{N}\sum_{t=1}^{N}\frac{|e_t|}{|y_t|}\times 100\%.$$

Если относительная ошибка составляет менее 13%, то точность подобранной модели признается удовлетворительной.

ЛЕКЦИЯ № 22. Определение сезонной компоненты временного ряда. Сезонные фиктивные переменные. Одномерный анализ Фурье

Существует несколько **методов моделирования сезонных и циклических колебаний**.

К ним относятся:

1) расчет сезонной компоненты и построение аддитивной или мультипликативной модели временного ряда;
2) применение сезонных фиктивных переменных;
3) анализ сезонности с помощью автокорреляционной функции;
4) использование рядов Фурье.

Рассмотрим первый из указанных подходов на примере моделирования сезонных колебаний, так как циклические колебания моделируются аналогично.

Если амплитуда сезонных колебаний не меняется во времени, то применяют аддитивную модель временного ряда:

$$y_t = T_t + S_t + \varepsilon_t,$$

где T — это трендовая компонента;

S — сезонная компонента;

ε — случайный шум.

Если амплитуда сезонных колебаний со временем изменяется, то применяется мультипликативная модель временного ряда:

$$y_t = T_t \times S_t + \varepsilon_t.$$

Рассмотрим временной ряд X_{ij},

где i — это номер сезона (периода времени внутри года, например месяц или квартал), $i = \overline{1, L}$ (L — число сезонов в году);

j — номер года, $j = \overline{1, m}$ (m — общее количество лет).

Количество уровней исходного ряда равно $Lm = n$.

При построении модели временного ряда сезонная компонента рассчитывается первой, а затем определяется трендовая составляющая.

В качестве сезонной составляющей для аддитивной модели временного ряда применяют абсолютное отклонение, обозначаемое как $S\alpha_i$. Для мультипликативной модели временного ряда в качестве сезонной компоненты применяют индекс сезонности — Is_i. Данные сезонные компоненты **должны удовлетворять следующим требованиям**:

1) в случае аддитивной модели сумма всех сезонных компонент (абсолютных отклонений $S\alpha_i$) должна быть равна нулю;

2) в случае мультипликативной модели произведение всех сезонных компонент (индексов сезонности Is_i) должно быть равно единице.

Перед расчетом сезонной составляющей исходный временной ряд подвергают процедуре выравнивания. Чаще всего используются методы механического выравнивания (метод скользящих средних, экспоненциальное сглаживание, медианное сглаживание и др.). В результате получают ряд выровненных значений $\tilde{X}_{ij}$, который не содержит сезонной компоненты.

Абсолютное отклонение в i-том сезоне рассчитывается как среднее арифметическое из отклонений фактического и выровненного уровней ряда:

$$Sa_i = \frac{\sum_{j=1}^{m}\left(X_{ij} - \tilde{X}_{ij}\right)}{m}.$$

Индекс сезонности в i-том сезоне рассчитывается как среднее арифметическое из отношений фактического уровня ряда к выровненному:

$$Is_i = \frac{1}{m}\sum_{j=1}^{m}\frac{X_{ij}}{\tilde{X}_{ij}}.$$

При расчете трендовой составляющей временного ряда используются метод аналитического выравнивания с помощью функций времени или кривых роста. Этот метод выравнивания применяют не к исходному фактическому динамическому ряду, а к ряду, из которого удалена сезонная компонента. Начальные уровни ряда корректируются на величину сезонной компоненты. В случае аддитивной модели из исходных уровней вычитают абсолютные отклонения $S\alpha_i$. В случае мультипликативной модели начальные уровни временного ряда делятся на индексы сезонности Is_i.

Если при построении аддитивной модели сумма всех абсолютных отклонений не равна нулю, рассчитывают **скорректированные значения сезонных компонент** по формуле:

$$Sa_i^{скоррект} = Sa_i - \frac{\sum_{i=1}^{L} Sa_i}{L},$$

где L — общее количество сезонных компонент.

1. Сезонные фиктивные переменные

Использование сезонных фиктивных переменных является одним из методов моделирования сезонных составляющих временного ряда. При этом подходе строится регрессионная модель, которая, помимо фактора времени, включает сезонные фиктивные переменные.

Фиктивной переменной (dummy variable) называется атрибутивный (или качественный) фактор, представленный с помощью определенного цифрового кода.

Регрессионная модель, включающая в качестве фактора (факторов) фиктивную переменную, называется **регрессионной моделью с переменной структурой**.

Рассмотрим временной ряд X_{ij},

где i — это номер сезона (периода времени внутри года, например, месяца или квартала);

$i = \overline{1, L}$ (L — число сезонов в году);

j — номер года, $j = \overline{1, m}$ (m — общее количество лет).

Количество уровней исходного ряда равно $L \times m = n$.

Число сезонных фиктивных переменных в регрессионной модели всегда должно быть на единицу меньше сезонов внутри года, т. е. должно быть равно величине $L - 1$. При моделировании годовых данных регрессионная модель, помимо фактора времени, должна содержать одиннадцать фиктивных компонент $(12 - 1)$.

При моделировании поквартальных данных регрессионная модель должна содержать три фиктивные компоненты $(4 - 1)$ и т. д.

Каждому из сезонов соответствует определенное сочетание фиктивных переменных. Сезон, для которого значения всех фиктивных переменных равны нулю, принимается за базу сравнения. Для остальных сезонов одна из фиктивных переменных принимает значение, равное единице.

Если имеются поквартальные данные, то значения фиктивных переменных D_1, D_2, D_3 будут принимать следующие значения для каждого из кварталов (табл. 4).

Таблица 4

Значения для каждого из кварталов для значения фиктивных переменных D_1, D_2, D_3

Квартал	D_2	D_3	D_4
1	0	0	0
2	1	0	0
3	0	1	0
4	0	0	1

Общий вид регрессионной модели с переменной структурой в данном случае будет иметь вид:

$$y_t = \beta_0 + \beta_1 \times t + \delta_2 \times D_2 + \delta_3 \times D_3 + \delta_4 \times D_4 + \varepsilon_t.$$

Построенная модель регрессии является разновидностью аддитивной модели временного ряда.

Базисным уравнением исследуемой регрессионной зависимости будет являться уравнение тренда для первого квартала:

$$y_t = \beta_0 + \beta_1 \times t + \varepsilon_t.$$

Частными случаями регрессионной зависимости будут являться уравнения трендов для остальных кварталов:

$y_t = \beta_0 + \beta_1 \times t + \delta_2 + \varepsilon_t$ — для второго квартала;

$y_t = \beta_0 + \beta_1 \times t + \delta_3 + \varepsilon_t$ — для третьего квартала;

$y_t = \beta_0 + \beta_1 \times t + \delta_4 + \varepsilon_t$ — для четвертого квартала.

Частные регрессионные уравнения отличаются друг от друга только на величину свободного члена уравнения регрессии δ_t. Коэффициент β_1 в данной регрессионной зависимости характеризует среднее абсолютное изменение уровней динамического ряда под влиянием основной тенденции.

Оценку сезонной компоненты для каждого сезона можно рассчитать как разность между средним значением свободных членов всех частных регрессионных уравнений и значением постоянного члена одного из уравнений.

Среднее значение свободных членов всех регрессионных уравнений рассчитывается по формуле:

$$\bar{\beta}_0 = \frac{\beta_0 + (\beta_0 + \delta_2) + (\beta_0 + \delta_3) + (\beta_0 + \delta_4)}{4}.$$

Оценки сезонных отклонений в случае поквартальных данных могут быть рассчитаны следующим образом:

1) ($\beta_0 - \bar{\beta}_0$) — для первого квартала;
2) ($\beta_0 + \delta_2 - \bar{\beta}_0$) — для второго квартала;
3) ($\beta_0 + \delta_3 - \bar{\beta}_0$) — для третьего квартала;
4) ($\beta_0 + \delta_4 - \bar{\beta}_0$) — для четвертого квартала.

В рассмотренной аддитивной модели сумма сезонных отклонений также должна равняться нулю.

2. Одномерный анализ Фурье

Использование рядов Фурье является одной из разновидностей спектрального анализа. Спектральный анализ временных рядов имеет весьма большое практическое значение, так как при моделировании и прогнозировании динамических рядов изучается вопрос о наличии и периоде сезонной компоненты в ряду. С помощью спектрального анализа можно определить в структуре временного ряда пик отклонений от тренда, что и позволит рассчитать длительность периодической компоненты ряда.

Суть спектрального анализа в том, что случайный стационарный процесс может быть представлен в виде суммы гармонических колебаний различных частот, которые называются гармониками. Функция, описывающая распределение амплитуд этого процесса по различным частотам, называется спектром.

Сезонную составляющую можно представить в виде модели разложения в ряд Фурье, где сезонные колебания представляют собой сумму нескольких синусоидальных и косинусоидальных гармоник с различными периодами:

$$y_t = \sum_{k=1}^{\infty} (u_k \cos \omega_k t + v_k \sin \omega_k t),$$

где u_k, v_k — некоррелированные случайные величины с нулевым математическим ожиданием и одинаковыми дисперсиями, т. е.

$$D(u_k) = D(v_k) = D_k;$$

ω_k — длина волны функции синуса или косинуса, выражаемая числом циклов (периодов) в единицу времени, т. е. частота.

Цель спектрального анализа временных рядов — оценивание спектра ряда. Спектр временного ряда — разложение дисперсии ряда по частотам для определения значимых гармонических составляющих. Значение спектра рассчитывается:

$$f(\omega_j) = \frac{1}{2\pi}\left[\lambda_o c_o + 2\sum_{k=1}^{m}\lambda_k c_k \cos\omega_j k\right],$$

где ω_j — частоты, для которых оцениваются спектры: $\omega_j = \frac{\pi_j}{m}$;

c_k — автокорреляционная функция, значения которой определяются так:

$$c_k = \frac{1}{n-k}\left[\sum_{t=i}^{n-k} z_t z_{t+k} - \frac{1}{n-k}\sum_{t=i-1}^{n} z_t + \sum_{t=i}^{n-k} z_t\right];$$

λ_κ — специально подобранные веса значений ковариационной функции, зависящие от частоты m, которые называются корреляционным окном.

Корреляционные окна — преобразования взвешенного скользящего среднего с шириной окна m.

Дисперсия ряда Фурье будет определяться по формуле:

$$D(y_t) = D\left[\sum_{k=0}^{\infty}\left(u_k \cos\omega_k t + v_k \sin\omega_k t\right)\right] =$$

$$= \sum_{k=0}^{\infty}\left(\cos^2\omega_k t + \sin^2\omega_k t\right)D_k = \sum_{k=0}^{\infty} D_k.$$

Дисперсия ряда Фурье равна сумме всех гармоник ее спектрального разложения. Можно сделать вывод, что дисперсия $D(y_t)$ распределена по различным частотам. Графически данное распределение можно изобразить с помощью периодограммы. Значения периодограммы обычно строятся в зависимости от частот или периодов. Период — величина, обратная частоте. Сущность анализа периодограммы сводится к тому, что необходимо найти частоты или периоды с большими спектральными плотностями, которые вносят наибольший вклад в периодические колебания

временного ряда, определив тем самым его основной период колебания.

Ряд Фурье:

$$y_t = \sum_{k=1}^{\infty} \left(u_k \cos \omega_k t + v_k \sin \omega_k t \right)$$

можно представить как модель множественной линейной регрессии, где зависимой переменной является временной ряд, а независимыми переменными — функции синусов всех возможных частот. Коэффициенты u_k при косинусах и v_k при синусах — это коэффициенты регрессии, которые показывают степень, с которой соответствующие функции коррелируют с данными. Если найденная корреляция (коэффициент при определенном синусе или косинусе) велика, то на соответствующей частоте в данных существует строгая периодичность.

Перед применением спектрального анализа временной ряд необходимо привести к стационарному виду.

3. Фильтрация временного ряда (исключение тренда и сезонной компоненты)

При изучении взаимосвязи между двумя и более временными рядами могут возникнуть такие **проблемы**:

1) ошибочность показателей тесноты и силы связи:
   а) если ряды содержат циклическую или сезонную компоненту одинаковой периодичности, то это приведет к завышению показателей тесноты связи;
   б) если один из временных рядов содержит циклическую или трендовую компоненту либо периодичность совместных колебаний различна, то это приведет к занижению показателей тесноты связи;
2) проблема «ложной корреляции»:
   а) если ряды содержат тренды одинаковой направленности, то между уровнями этих рядов всегда будет существовать положительная корреляция;
   б) если ряды содержат тренды противоположной направленности, то корреляция всегда будет отрицательной.

Методы фильтрации временного ряда направлены на устранение данных проблем путем исключения из них трендовой и сезонной компонент.

Сезонную компоненту в случае аддитивной модели устраняют, рассчитав абсолютные разности $S\alpha_i$ и вычтя их из исходных уровней ряда. В случае мультипликативной модели рассчитывают индексы сезонности Is_i и разделить исходные уровни ряда на них.

«Ложная корреляция» устраняется с помощью исключения тренда из временного ряда.

Пусть на основе двух временных рядов построено регрессионное уравнение вида:

$$Y_t = \beta_0 + \beta_1 \times X_t + \varepsilon_t.$$

Анализ остатков этой модели позволит определить наличие «ложной» автокорреляции (если есть обычная автокорреляция остатков, значит, существует и «ложная» автокорреляция).

Трендовую компоненту исключают методом отклонений от тренда. Необходимо определить отклонения уровней Y_t и X_t от их значений, рассчитанных на основании трендовых уравнений:

$$e(x_t) = x_t - \tilde{x}_t; \quad e(y_t) = y_t - \tilde{y}_t.$$

Определяют степень связи между этими отклонениями, например, с помощью коэффициента корреляции:

$$r(e(x_t), e(y_t)) = \frac{\sum_{i=1}^{N} e(x_t) \times e(y_t)}{\sqrt{\sum_{i=1}^{N} e^2(x_t) \times \sum_{i=1}^{N} e^2(y_t)}}.$$

В случае линейной регрессии можно построить уравнение зависимости отклонения $e(y_t)$ от $e(x_t)$:

$$e(y_t) = \alpha_0 + \alpha_1 \times e(x_t).$$

Параметры данного уравнения могут быть определены с помощью МНК по формулам:

$$\alpha_1 = \frac{\sum_{i=1}^{N} e(y_t) \times e(x_t)}{\sum_{i=1}^{N} e^2(x_t)}; \quad \alpha_0 = \overline{e(x_t)} - \alpha_1 \times \overline{e(x_t)} = 0.$$

Уравнение можно записать в виде:

$$e(y_t) = \alpha_1 \times e(x_t).$$

Трендовую компоненту можно исключить методом последовательных разностей. Рассчитываются разности между текущим и предыдущим уровнями временного ряда:

$$\nabla y_t = Y_t - Y_{t-1};$$
$$\nabla x_t = X_t - X_{t-1}.$$

Данные величины — абсолютные цепные приросты. Показатель линейной корреляции абсолютных цепных приростов можно рассчитать так:

$$r(\nabla x_t, \nabla y_t) = \frac{\sum_{i=1}^{N} \nabla x_t \times \nabla y_t}{\sqrt{\sum_{i=1}^{N} \nabla^2 x_t \times \sum_{i=1}^{N} \nabla^2 y_t}}.$$

Линейное уравнение регрессии по абсолютным приростам имеет вид:

$$\nabla y_t = \alpha_0 + \alpha_1 \times \nabla x_t.$$

Параметр α_1 в данном уравнении характеризует в среднем прирост Y при изменении прироста X на единицу своего измерения. Параметр α_0 характеризует прирост Y при нулевом приросте X.

Разностные операторы первого порядка позволяют исключать автокорреляцию только в тех временных рядах, в которых тенденция выражена прямой линией.

Разности второго порядка позволяют исключать автокорреляцию в тех рядах динамики, основная тенденция которых выражена параболой второго порядка.

Основным недостатком данного метода является то, что происходит потеря информации за счет сокращения числа наблюдений.

Если ряды динамики имеют различные виды трендов, то можно коррелировать соответствующие им цепные показатели.

4. Автокорреляция уровней временного ряда

Если временной ряд является нестационарным, т. е. содержит тренд и цикличность, то значения каждого последующего уровня ряда корреляционно зависят от предыдущих значений.

Автокорреляцией уровней временного ряда называется корреляционная зависимость между настоящими и прошлыми значениями уровней данного ряда.

Величина сдвига между рядами наблюдений — временной лаг (l).

Значение временного лага определяет порядок коэффициента автокорреляции. Если существует корреляционная зависимость между уровнями x_n и x_{n-1}, то величина временного лага равняется единице.

Данную зависимость будет характеризовать коэффициент автокорреляции первого порядка между рядами наблюдений x_1, ..., $x_{n\text{-}1}$ и x_2, ..., x_n. Если лаг $l = 2$, то корреляционная зависимость будет характеризоваться коэффициентом автокорреляции второго порядка и т. д. С увеличением величины лага на единицу число пар значений, по которым рассчитывается коэффициент автокорреляции, уменьшается на единицу. Поэтому рекомендуется брать максимальный порядок коэффициента автокорреляции равным $n/4$, где n — число уровней временного ряда.

Автокорреляция определяется с помощью выборочного коэффициента автокорреляции:

$$r_l(x_t) = \frac{\overline{x_t \times x_{t-l}} - \overline{x_t} \times \overline{x_{t-l}}}{G(x_t) \times G(x_{t-l})},$$

где $\overline{x_t \times x_{t-l}}$ — среднее арифметическое произведения двух рядов, взятых с лагом l:

$$\overline{x_t \times x_{t-l}} = \frac{\sum_{i=1+l}^{n} x_i \times x_{i-l}}{n-l};$$

$\overline{x_t}$ — значение среднего уровня ряда x_{1+l}, x_{2+l}, ..., x_n:

$$\overline{x_t} = \frac{\sum_{i=1+l}^{n} x_i}{n-l};$$

$\overline{x_{t-l}}$ — значение среднего уровня ряда x_1, x_2, ..., x_{n-l}:

$$\overline{x_{t-l}} = \frac{\sum_{i=1+l}^{n} x_{i-l}}{n-l};$$

$G(x_i)$, $G(x_{i\text{-}l})$ — средние квадратические отклонения, рассчитанные для рядов x_{1+l}, x_{2+l}, ..., x_n и x_1, x_2, ..., x_{n-l}, соответственно.

Определив несколько последовательных коэффициентов автокорреляции для исследуемого ряда, можно выявить лаг l, при котором автокорреляция r_l наиболее высока, рассчитав таким образом структуру временного ряда.

Выделить следующие **основные положения анализа структуры временного** ряда на основании автокорреляционных коэффициентов:

1) если наиболее высоким окажется значения коэффициента автокорреляции первого порядка $r_{l=1}$, то изучаемый ряд содержит только трендовую компоненту;

2) если наиболее высоким окажется коэффициент автокорреляции порядка l, то, кроме трендовой компоненты, исследуемый временной ряд содержит колебания периодом. Это могут быть как циклические, так и сезонные колебания;

3) если же ни один из коэффициентов автокорреляции r_l, где $l = \overline{1, L}$, не окажется значимым, то можно сделать один из двух возможных выводов:

   а) ряд не содержит трендовой и циклической компонент, а его колебания вызваны воздействием случайной компоненты, т. е. мы имеем дело с моделью случайного тренда;

   б) ряд содержит сильную нелинейную тенденцию, для выявления которой необходимо провести дополнительный анализ временного ряда.

Наиболее простым и распространенным методом определения структуры временного ряда является построение графиков автокорреляционной и частной автокорреляционной функций (АКФ и ЧАКФ).

АКФ — это функция оценки коэффициента автокорреляции в зависимости от величины временного лага между исследуемыми рядами. Графиком АКФ является коррелограмма. ЧАКФ отличается от АКФ тем, что при ее построении устраняется корреляционная зависимость между наблюдениями внутри лагов.

ЛЕКЦИЯ № 23. Стационарные ряды. Модель авторегрессии и проинтегрированного скользящего среднего (arima). Показатели качества модели АРПСС. Критерий Дики-Фуллера

Если значения уровней временного ряда точно определены какой-либо математической функцией, то данный временной ряд называется **детерминированным**, а сама функция называется **реализацией** исследуемого процесса. Если же уровни временного ряда могут быть описаны с помощью функции распределения вероятностей, то такой временной ряд называется **случайным**.

Уровни временного ряда могут быть детерминированными и случайными величинами.

Уровни случайного временного ряда могут представлять собой **непрерывные** и **дискретные случайные величины**. Случайная величина называется дискретной, если множество ее возможных значений является конечным или счетным. Случайная величина называется непрерывной, если она может принимать любое значение из конечного или бесконечного интервала. Примером случайного временного ряда с дискретными уровнями может служить фиксация значений ежемесячной выдачи стипендий студентам. Пример случайного временного ряда с непрерывными уровнями — регистрация с определенной периодичностью температуры воздуха, которая во времени изменяется непрерывно.

Процесс, развивающийся во времени в соответствии с законами теории вероятностей, называется **стохастическим процессом**. К стохастическим процессам относится класс процессов, называемых **стационарными**.

Стохастический процесс называется стационарным, если его основные свойства остаются неизменными во времени.

Обозначим через x_t уровень временного ряда. Тогда стационарный процесс в широком смысле будет характеризоваться такими свойствами**,** как:

1) постоянное математическое ожидание стационарного ряда $E(y_t)$, т. е. среднее значение временного ряда, вокруг которо-

го изменяются уровни, является величиной неизменной: $E(y_t) = \bar{y} = \text{const}$;

2) постоянная дисперсия стационарного ряда, определяющая размах его колебаний относительно среднего значения $\bar{x}$: $D(y) = E(y_t - \bar{y})^2 = G^2(y) = \text{const}$;

3) постоянная автоковариация стационарного ряда с лагом l, т. е. ковариация между значениями x_t и x_{t+1}, отделенными интервалом в l единиц времени, определяемая по формуле:

$$R_l(y_t) = \text{cov}(y_t, y_{t+l}) = \text{E}\left[(y_t - \bar{y})(y_{t+l} - \bar{y})\right].$$

Для стационарных рядов автоковариация зависит только от величины лага l, поэтому $R_{l=0}(y_t) = G^2(y)$;

4) постоянство коэффициентов автокорреляция стационарного ряда с лагом l, т. е. автокорреляция является нормированной автоковариацией:

$$\rho_l = \frac{\text{E}\left[(y_t - \bar{y})(y_{t+l} - \bar{y})\right]}{\sqrt{\text{E}(y_t - \bar{y})^2 \, \text{E}(y_{t+l} - \bar{y})^2}} = \frac{\text{E}\left[(y_t - \bar{y})(y_{t+l} - \bar{y})\right]}{G^2(y)},$$

так как для стационарного процесса $G^2(y) = \text{const}$. Коэффициент автокорреляции порядка l равен:

$$\rho_l = \frac{R_l(y_t)}{R_{l=0}(y_t)}.$$

Временной ряд, не удовлетворяющий вышеперечисленным свойствам, называется **нестационарным** временным рядом.

Частным случаем стационарных временных рядов является случайный процесс, называемый **белым шумом**.

Случайная последовательность значений $y_1, y_2, \ldots, y_N$ называется белым шумом, если ее математическое ожидание равно нулю, т. е. $E(Y_t) = 0$, где $t = \overline{1, N}$, ее элементы являются некоррелированными (независимыми друг от друга) одинаково распределенными величинами, а дисперсия постоянна — $D(Y_t) = G^2 = \text{const}$.

Белый шум является абсолютно теоретическим процессом, который реально не существует, однако он представляет собой очень важную математическую модель, которая широко применяется при решении множества практических задач.

1. Линейные модели стационарного временного ряда

Модели авторегрессии и скользящего среднего являются основными линейными моделями стационарных временных рядов. Стохастический временной ряд называется стационарным, если его математическое ожидание, дисперсия, автоковариация и автокорреляция остаются неизменными во времени. Исходя из данного определения можно ввести для временного ряда модель авторегрессии порядка *p*. В этом случае уровень временного ряда представляется в виде:

$$y_t = \delta_1 y_{t-1} + \delta_2 y_{t-2} + \ldots + \delta_p y_{t-p} + v_t,$$

где p — это порядок авторегрессии;

δ_t — коэффициенты авторегрессии, подлежащие оцениванию;

v_t— белый шум (т. е. случайная величина с нулевым математическим ожиданием).

На практике чаще всего используются модели первого, второго, максимум третьего порядков. Авторегрессионная модель порядка p обозначается как $AP(p)$ или $AR(p)$.

Модель авторегрессии первого порядка $AP(1)$ может быть записана в виде равенства:

$$y_t = \delta y_{t-1} + v_t.$$

Данная модель называется **марковским процессом**, так как значения переменной y в текущий момент времени i зависят только от значений переменной y в предыдущий момент времени $(t-1)$. Для модели $AP(1)$ действует ограничение $|\delta| < 1$.

Модель авторегрессии второго порядка $AP(2)$ можно записать как:

$$y_t = \delta_1 y_{t-1} + \delta_2 y_{t-2} + v_t.$$

Данная модель имеет название **процесс юла**. На коэффициенты авторегрессионной модели второго порядка накладываются следующие дополнительные ограничения:

$$\begin{aligned} &(\delta_1 + \delta_2) < 1; \\ &(\delta_2 - \delta_1) < 1; \\ &|\delta_2| < 1. \end{aligned}$$

Другой разновидностью линейных моделей стационарных временных рядов является модель скользящего среднего.

Простой класс моделей временных рядов с конечным числом параметров можно получить, представив уровень временного ряда в виде алгебраической суммы членов ряда белого шума с числом слагаемых q.

Общая модель скользящего среднего порядка q имеет вид:

$$y_t = v_t - \varphi_1 \times v_{t-1} - \varphi_2 \times v_{t-2} - \ldots - \varphi_q \times v_{t-q},$$

где q — это порядок скользящего среднего;

φ_i — неизвестные коэффициенты, подлежащие оцениванию;

v_t — белый шум.

Модель скользящего среднего порядка q обозначается как $CC(q)$ или $MA(q)$. Как и в случае авторегрессионных моделей, наиболее широко применяются модели скользящего среднего первого и второго порядков — $CC(1)$ и $CC(2)$.

Коэффициенты модели скользящего среднего порядка q не обязательно должны в сумме давать единицу и не обязательно должны быть положительными.

Для достижения большей гибкости при построении модели изучаемых временных рядов полезно включать в нее и авторегрессионные члены, и члены скользящего среднего. Такая модель называется смешанной моделью авторегрессии скользящего среднего. Она обозначается как $APCC(p,q)$ или $ARMA(p,q)$. Модель $APCC(p,q)$ также является линейной моделью стационарных временных рядов.

Наибольшее применение получила смешанная модель с одним параметром авторегрессии $p = 1$ и одним параметром скользящего среднего $q = 1$. Она записывается как $APCC(1, 1)$:

$$y_t = \delta \times y_{t-1} + v_t - \varphi \times v_{t-1},$$

где δ — параметр авторегрессии;

φ — параметр скользящего среднего;

v_t — белый шум.

Условие, обеспечивающее стационарность данной модели, — это ограничение $|\delta| < 1$, а условие, обеспечивающее обратимость, — это ограничение $|\varphi| < 1$.

Модель $APCC(p,q)$ отвечает свойству обратимости, т. е. описанное выше уравнение процесса скользящего среднего можно обратить (переписать) в виде уравнения авторегрессии неограниченного порядка, и наоборот.

2. Модель авторегрессии и проинтегрированного скользящего среднего (ARIMA)

Модель, используемая при моделировании нестационарных временных рядов, называется **моделью авторегрессии и проинтегрированного скользящего среднего — АРПСС** или **ARIMA**.

В основе данной модели лежат два процесса:

1) **процесс авторегрессии**:

$$x_t = \alpha + \delta_1 \times x_{t-1} + \delta_2 \times x_{t-2} + \ldots + \varepsilon,$$

где α — константа (свободный член);

δ_1, δ_2 — параметры авторегрессии;

ε — случайное воздействие (ошибка).

2) **процесс скользящего среднего**:

$$x_t = \mu + \varepsilon_t - \theta_1 \times \varepsilon_{t-1} - \theta_2 \times \varepsilon_{t-2} - \ldots,$$

где μ — константа;

$\theta_1, \theta_2 \ldots$ — параметры скользящего среднего;

ε — случайное воздействие (ошибка).

Общий вид модели среднего значения однофакторного динамического процесса описывается следующей формулой:

$$y_t = C + \sum_{i=1}^{R} AR_i y_{t-i} + \sum_{j=1}^{M} MA_j \varepsilon_{t-j} + \varepsilon_t,$$

где C — константа;

ε_t — некомпенсированный моделью случайный остаток.

В оригинальных обозначениях Бокса и Дженкинса модель ARIMA записывается как ARIMA (p, d, q),

где p — параметры авторегрессии;

d — порядок разностного оператора;

q — параметры скользящего среднего.

Для рядов с периодической сезонной компонентой применяется АРПСС с сезонностью, которая в обозначениях Бокса и Дженкинса записывается как АРПСС (p, d, q) (ps, ds, qs), где параметры во второй скобке соответственно называются «сезонная авторегрессия», «сезонная разность» и «сезонное скользящее среднее».

Аппроксимация временного ряда с помощью модели АРПСС происходит в несколько этапов.

1**. Проверка ряда на стационарность.** Ряд является стационарным, если он имеет постоянные по времени средние, дисперсию и автокорреляцию с удаленными сезонной, циклической и трендовой составляющей. Такой ряд называется стационарным рядом случайных остатков.

Проверить исходной ряд на стационарность можно с помощью *автокорреляционной функции (АКФ) и частной автокорреляционной функции (ЧАКФ) остатков.* Остатки представляют собой разности наблюдаемого временного ряда и значений, вычисленных с помощью модели.

Применение модели АРПСС предполагает обязательную стационарность исследуемого ряда. Временной ряд к стационарному виду приводят с помощью применения разностных операторов, порядок которых определяется параметром (*d*).

2. **Идентификация порядка модели и оценивание ее параметров.** На этом этапе необходимо решить, как много параметров *p* и *q* должно присутствовать в модели процесса. Основными инструментами идентификации модели АРПСС являются АКФ и ЧАКФ.

Во время оценивания порядка модели используется квазиньютоновский алгоритм максимизации правдоподобия наблюдения значений ряда по значениям параметров. Во время итераций минимизируется (условная) сумма квадратов остатков модели.

Для оценки параметров используется t-статистика Стьюдента. Если значения вычисляемой t-статистики незначимы, соответствующие параметры в большинстве случаев удаляются из модели без ущерба подгонки.

3. **Прогноз**. Полученные оценки параметров используются на последнем этапе для вычисления нового значения ряда и построения доверительного интервала для прогноза.

Показателем качества построенной модели АРПСС является анализ остатков. Модель АРПСС адекватно описывает исходные данные, если остатки модели являются некоррелированными нормально распределенными случайными величинами.

3. Показатели качества модели АРПСС

Общим показателями качества построенной АРПСС модели являются критерий Акайке и байесовский критерий Шварца.

По своей сути они аналогичны критерию максимума скорректированного множественного коэффициента детерминации R^2 или минимума дисперсии случайной ошибки модели G^2. На современном этапе развития науки с помощью этих критериев находятся оптимальные значения порядков параметров p и q модели.

Информационный критерий Акайке (Akaike information criterion— AIC) предназначен для выбора наилучшей модели для временного ряда y_t из некоторого множества моделей.

Обозначим через $l_n(\tilde{\varphi})$ максимальное значение логарифмической функции правдоподобия эконометрической модели,

где $\tilde{\varphi}$ — оценка максимального правдоподобия вектора φ параметров модели.

Критерий Акайке имеет вид:

$$AIC_l = l_n(\tilde{\varphi}) - h,$$

где h — размерность вектора φ, т. е. число оцениваемых коэффициентов регрессионной модели.

В случае линейной или нелинейной модели регрессии, состоящей из одного уравнения, критерий Акайка может быть преобразован к виду:

$$AIC_G = \ln\left(\tilde{G}^2\right) + \frac{2p}{n},$$

где n — объем выборки;

$\tilde{G}^2$ — оценка максимального правдоподобия дисперсии остатков регрессионной модели e_t.

Оба варианта критерия Акайке дают одинаковый результат, но в первом случае выбирается модель с наибольшим значением критерия, а во втором случае — с наименьшим значением критерия.

Байесовский критерий Шварца (Schwarz Bayesian criterion — SBC) также предназначен для выбора наилучшей модели временного ряда из некоторого множества моделей. Он определяется как:

$$SBC_l = l_n(\tilde{\varphi}) - \frac{1}{2} p \ln n.$$

Для регрессионных моделей существует альтернативный вариант данного критерия:

$$SBC_G = \ln\left(\tilde{G}^2\right) + \left(\frac{\ln n}{n}\right) p.$$

По первому варианту расчета критерия *SBC* выбирается та модель, для которой значение SBC_l является наибольшим. При втором варианте выбирается та модель, для которой значение SBC_G является наибольшим. Оба варианта критериев дают одинаковый результат при выборе моделей. Результаты критериев Акайке и Шварца могут отличаться.

Качество построенной модели осуществляется через проверку автокорреляции ее остатков. В этом случае применяют **общий критерий множителей Лагранжа (LM test)**, позволяющий обнаруживать в регрессионных остатках автокорреляцию более высоких порядков, чем первый. Но он применим только для больших выборок.

Рассмотрим регрессионную модель вида:

$$y_t = \sum_{i=1}^{k} x_{it}\beta_i + \varepsilon_i, \quad t = \overline{1,n},$$

где ε_t — случайная ошибка модели:

$$\varepsilon_t = \rho_1\varepsilon_{t-1} + \rho_2\varepsilon_{t-2} + \ldots + \rho_p\varepsilon_{t-p} + u_t;$$

ρ — коэффициент автокорреляции порядка 1, ..., p;

u_t — нормально распределенная случайная величина с нулевым математическим ожиданием и дисперсией G^2: $u_t \sim N(0, G^2)$.

Данная регрессионная модель может включать лаговые значения зависимой переменной.

Необходимо проверить основную гипотезу H_0 о незначимости коэффициентов автокорреляции:

$$H_0 \;/\; \rho_1 = \rho_2 = \ldots = \rho_p = 0.$$

Метод Лагранжа состоит из нескольких этапов:

1) с помощью метода наименьших квадратов оценивается регрессия

$$y_t = \sum_{i=1}^{k} x_{it}\beta_i + \varepsilon_i$$

и рассчитываются остатки модели e_t: $e_t = y_t - \tilde{y}_t$;

2) оценивается регрессия вида:

$$e_t = \sum_{i=1}^{k} x_{it}a_i + \sum_{i=1}^{p} e_{t-i}\rho_i + v_t,$$

для которой проверяется значимость коэффициентов ρ_t при лаговых значениях остатков.

Для этого вычисляется F-статистика, которая подчиняется χ^2-распределению с p степенями свободы. Если $\chi^2_{\text{набл}} > \chi^2_{крит}$, то основная гипотеза об отсутствии автокорреляции в остатках отвергается. Если $\chi^2_{\text{набл}} < \chi^2_{крит}$, то гипотеза об отсутствии автокорреляции принимается.

4. Критерий Дики-Фуллера

Задача проверки основной гипотезы вида $\rho = 0$ в авторегрессионном уравнении первого порядка

$$y_t = a + \rho\, y_{t-1} + \varepsilon_t$$

называется **проверкой наличия единичных корней**. Временной ряд y_t является стационарным, если $-1 < \rho < 1$. Если $\rho = 1$, то временной ряд y_t является нестационарным и представляет собой модель со случайным трендом. Если $\rho > 1$, то временной ряд y_t также является нестационарным. Гипотеза о стационарности ряда сводится к проверке гипотезы вида $\rho = 1$.

Наиболее распространенным критерием проверки наличия единичных корней является **критерий Дики-Фуллера**. Выдвигается основная гипотеза $\rho = 1$ для модели:

$$y_t = a + \rho\, y_{t-1} + \varepsilon_t.$$

Происходит оценивание не этого авторегрессионного уравнения, а модели, получаемой после перехода к первым разностям:

$$\Delta y_t = \delta\, y_{t-1} + \varepsilon_t,$$

где $\delta = \rho - 1$.

Проверка основной гипотезы $\rho = 1$ аналогична проверке гипотезы $\delta = 0$. Проверка данной гипотезы может осуществляться для трех типов регрессионных уравнений:

$$\Delta y_t = \delta y_{t-1} + \varepsilon_t; \qquad (1)$$

$$\Delta y_t = a + \delta y_{t-1} + \varepsilon_t; \qquad (2)$$

$$\Delta y_t = a + \delta y_{t-1} + \beta_t + \varepsilon_t. \qquad (3)$$

Данные модели отличаются только наличием членов уравнения α и β_t.

Первая модель является моделью случайного тренда, во вторую модель включается свободный член α, являющийся коэффициентом случайного тренда.

В третью модель включены и случайный тренд, и линейный временной тренд β_t.

Процедура проверки гипотезы $\delta = 0$ сводится к оцениванию МНК одной или нескольких из указанных регрессионных моделей для получения оценки $\tilde{\delta}$ и ее стандартной ошибки. Наблюдаемое значение t-статистики определяют по формуле:

$$t_{набл} = \frac{\tilde{\delta}}{\omega(\tilde{\delta})},$$

где $\omega(\tilde{\delta})$ — стандартная ошибка оценки $\tilde{\delta}$.

Однако t-статистика в данном случае не подчиняется распределению Стьюдента. В результате исследования Дики-Фуллера были найдены критические значения t-критерия для гипотезы $\delta = 0$ в зависимости от вида регрессионного уравнения и объема выборочной совокупности. Эти статистики обозначаются как τ, τ_μ и τ_τ соответственно. Они приведены в таблице критических значений статистик Дики-Фуллера для различных уровней значимости.

Расширенный критерий Дики-Фуллера (Augmented Dickey-Fuller Test ADF) используется при проверке гипотезы о наличии авторегрессии порядков выше первого.

Авторегрессионный процесс порядка p может быть записан в виде:

$$\Delta y_t = a + \delta y_{t-1} + \beta t + \sum_{i=2}^{p} \varphi_i \Delta y_{t-i+1} + \varepsilon_t.$$

Как и при проверке гипотезы о наличии единичного корня, основная гипотеза формулируется как $\delta = 0$. Если данная гипотеза верна, то исследуемое уравнение имеет единичный корень, т. е. подчиняется процессу авторегрессии первого порядка.

Проверка гипотезы $\delta = 0$ осуществляется для различных типов регрессионных уравнений:

$$\Delta y_t = \delta y_{t-1} + \sum_{i=2}^{p} \varphi_i \Delta y_{t-i+1} + \varepsilon_t; \qquad (1)$$

$$\Delta y_t = a + \delta y_{t-1} + \sum_{i=2}^{p} \varphi_i \Delta y_{t-i+1} + \varepsilon_t; \qquad (2)$$

$$\Delta y_t = a + \delta y_{t-1} + \beta t + \sum_{i=2}^{p} \varphi_i \Delta y_{t-i+1} + \varepsilon_t. \qquad (3)$$

Для первой модели регрессии используется статистика τ (при отсутствии свободного члена и временного тренда); для второй модели регрессии, включающей свободный член, используется статистика τ_μ; для третьей модели регрессии, включающей свободный член и временной линейный тренд, используется статистика τ_τ.

Если сумма коэффициентов регрессионной модели вида:

$$\Delta y_t = a + \delta y_{t-1} + \beta t + \sum_{i=2}^{p} \varphi_i \Delta y_{t-i+1} + \varepsilon_t$$

равна единице, т. е. $\sum_{i=1}^{p} a_i = 1,$ то параметр $\delta = 0$, т. е. в данном уравнении имеется единичный корень.

ЛЕКЦИЯ № 24. Цензурированные и стохастические объясняющие переменные

Объясняющая переменная называется цензурированной в том случае, если она представляет собой момент наступления интересующего нас события, а продолжительность исследования ограничена во времени.

Концепция цензурирования переменных или наблюдений впервые возникла в исследованиях, связанных с биологией и медициной. Однако на современном этапе развития науки метод цензурирования применяется во многих областях, например в социологии, демографии и др. В экономике с помощью цензурирования изучается время «выживания» новых предприятий или новой продукции, поступившей на рынок.

Существует понятие правого и левого **цензурирования**, которое определяется в зависимости от направления процесса цензурирования. Например, осуществляется проверка 100 однотипных предприятий по определенным параметрам, которая заканчивается через некоторое время. В данном случае применяется правое цензурирование, так как известно в какой момент процесс был начат и в какой конкретно момент времени, расположенный справа от точки начала проверки, он закончится. Левое цензурирование может применяться в биомедицинских исследованиях.

Цензурирование может быть однократное (наступающее в один момент времени) и многократное (наступающее в различные моменты времени). Например, проверка 100 предприятий может закончиться спустя фиксированный отрезок времени. Если процесс проверки завершился в определенный момент времени, то использовалось однократное цензурирование, а исследуемые данные были цензурированы один раз. Многократное цензурирование используется в биомедицинских исследованиях.

Существует цензурирование I и II типа. Цензурирование I типа применяется в тех ситуациях, когда процесс тестирования завершается в заранее известный момент времени. При проверке 100 предприятий процесс заканчивается через фиксированный отрезок времени.

Количество предприятий, не прошедших проверку по установленным критериям, является случайной величиной, а время эксперимента — величиной заранее известной. При цензурировании II типа процесс тестирования продолжается до тех пор, пока не будет достигнут предел по браку, т. е. при проверке 100 предприятий процесс будет закончен тогда, когда 15 предприятий не будут удовлетворять заданным критериям. Число забракованных элементов известно, а время эксперимента является величиной случайной.

Линейную модель регрессии с цензурированной зависимой переменной можно записать следующим образом:

$$y_i^* = \beta_0 + \beta_1 x_i + \varepsilon_i,$$

где $y_i = y_i^*, y_i^* > 0.$

При цензурировании зависимой переменной пользуются методом усечения:

$$y_i = \begin{cases} y_i^*, y_i^* > 0, \\ 0,\ y_i^* \le 0. \end{cases}$$

Оценки неизвестных коэффициентов модели регрессии с цензурированными переменными находятся с помощью метода максимума правдоподобия. В данном случае минимизируется логарифм функционала максимального правдоподобия вида:

$$l = \ln L = \sum_{i \in I_1} \left[-\frac{1}{2} \ln 2 \prod_G - \frac{1}{2} \left(\frac{y_i^* - x_i^T \beta}{G} \right) \right] + \sum_{i \in I_2} \ln \Phi \left(-\frac{x_i^T \beta}{G} \right).$$

Дифференцируя данный функционал по вектору неизвестных коэффициентов, получим оценки максимального правдоподобия $\tilde{\beta}_{ML}$.

Данные оценки могут получиться смещенными. Смещение устраняется с помощью изменения функционала максимального правдоподобия и приведения его к виду:

$$l = \ln L = \sum_{i=1}^{n} \ln \left(\frac{1}{\sqrt{2\Pi_G}} e^{-\frac{1}{2}\left(\frac{y_i^* - X_i^T \beta}{G} \right)} \right) - \sum_{i=1}^{n} \ln \left(1 - \Phi \left(-\frac{x_i^T \beta}{G} \right) \right).$$

Стохастические объясняющие переменные

Первая предпосылка нормальной линейной регрессионной модели заключается в том, что независимые (объясняющие) переменные x_i $\left(i=\overline{1,n}\right)$ являются детерминированными (нестохастическими), т. е. неслучайными величинами. Однако на практике данная предпосылка не выполняется, так как объясняющие переменные сами были определены из других экономических моделей, а также могли быть ошибочно измерены.

Выделяют три вида моделей со **стохастическими объясняющими переменными**:

1) величины x и ε распределены независимо друг от друга;

2) величины x и ε одномоментно не коррелированы, т. е. объясняющие переменные зависят от случайной ошибки регрессии, но их значения на текущий момент времени не коррелированы;

3) величины x и ε одномоментно коррелированы, т. е. значения объясняющих переменных и случайной ошибки регрессии коррелируют в каждый момент времени.

Рассмотрим применение обычного метода наименьших квадратов для перечисленных моделей исходя из предположения о том, что средние значения и дисперсия объясняющей переменной в генеральной совокупности конечны. Однако следует учитывать, что данное предположение не распространяется на модели временных рядов.

Пусть распределение x **имеет конечное математическое ожидание** и конечную дисперсию:

1) величины x и ε распределены независимо друг от друга. Данное условие характерно для нормальной линейной регрессионной модели, например:

$$y_i = \beta_0 + \beta_1 \times x_i + \varepsilon_i.$$

В данном случае МНК-оценки коэффициентов регрессии будут удовлетворять свойствам несмещенности, состоятельности и эффективности;

2) величины x и ε одномоментно не коррелированы. В качестве примера подобной модели можно привести модель регрес-

сии, включающую лаговую зависимую переменную как одну из объясняющих переменных.

Рассмотрим свойства МНК-оценок для авторегрессионной модели первого порядка:

$$y_t = \beta_0 + \beta_1 x_t + \beta_2 y_{t-1} + \varepsilon_i.$$

Текущая зависимая переменная y_t и случайная ошибка ε_t не коррелированны, однако регрессор y_{t-1} коррелирован со всеми предыдущими ошибками модели. В этом случае возникает вопрос о несмещенности МНК-оценки $\tilde{\beta}_2$.

Приведем первоначальное уравнение регрессии к виду:

$$y_t = \beta_0 + \beta_2 y_{t-1} + \varepsilon_i.$$

Тогда

$$\tilde{\beta}_2 = \beta_2 + \frac{S_{y\varepsilon}}{S_{yy}},$$

где $\frac{S_{y\varepsilon}}{S_{yy}} \neq 0$, так как переменная коррелирована со всеми предыдущими ошибками.

Таким образом, МНК-оценка $\tilde{\beta}_2$ является смещенной в конечных выборках при условии неодномоментного коррелирования величин x и ε. Однако если объем выборки стремится к бесконечности, то МНК оценка $\tilde{\beta}_2$ будет состоятельной:

$$p\lim \tilde{\beta}_2 = \beta_2 \text{ и } \frac{S_{y\varepsilon}}{S_{yy}} \xrightarrow{P} 0,$$

где p lim означает предел по вероятности;

3) величины x и ε одномоментно коррелированы. В данном случае даже при большом объеме выборочной совокупности МНК оценки будут смещенными и несостоятельными.

При условии, что дисперсия x неограниченно возрастает, т. е. является бесконечной, оценить свойства МНК-оценок весьма затруднительно, так как $Cov(x, \varepsilon)$ не имеет предела.

Выявить неслучайный характер объясняющих переменных можно с помощью графика остатков регрессионного уравнения: $e_i = y_i - \tilde{y}_i$. Если остатки расположены в виде горизонтальной полосы, то регрессионные остатки e_i являются случайными величинами.

ЛЕКЦИЯ № 25. Системы эконометрических и одновременных уравнений. Проблема и условия идентификации модели

Необходимость использования систем эконометрических уравнений вызвана тем, что многие экономические процессы не могут быть реально описаны с помощью одного уравнения. В таких случаях прибегают к построению нескольких эконометрических уравнений, которые образуют систему.

Системы эконометрических уравнений включают множество зависимых или эндогенных переменных и множество предопределенных переменных (лаговые и текущие независимые переменные, а также лаговые эндогенные переменные). Как и эконометрические модели с одним уравнением, системы эконометрических уравнений **направлены на объяснение текущих значений эндогенных переменных** в зависимости от значений предопределенных переменных. В эконометрическом моделировании выделяют три вида систем уравнений.

1. Система независимых уравнений определяется тем, что каждая эндогенная переменная *y* является функцией только от одних и тех же переменных *x*:

$$\begin{cases} y_1 = a_{11}x_1 + a_{12}x_2 + \ldots + a_{1m}x_m + \varepsilon_1, \\ y_2 = a_{21}x_1 + a_{22}x_2 + \ldots + a_{2m}x_m + \varepsilon_2, \\ \ldots \\ y_n = a_{n1}x_1 + a_{n2}x_2 + \ldots + a_{nm}x_m + \varepsilon_n. \end{cases}$$

2. Система рекурсивных уравнений определяется тем, что в каждом последующем уравнении эндогенная переменная выступает в качестве экзогенной переменной:

$$\begin{cases} y_1 = a_{11}x_1 + a_{12}x_2 + \ldots + a_{1m}x_m + \varepsilon_1, \\ y_2 = b_{21}y_1 + a_{21}x_1 + a_{22}x_2 + \ldots + a_{2m}x_m + \varepsilon_2, \\ y_3 = b_{31}y_1 + b_{32}y_2 + a_{31}x_1 + a_{32}x_2 + \ldots + a_{3m}x_m + \varepsilon_3, \\ \ldots \\ y_n = b_{n1}y_1 + b_{n2}y_2 + \ldots + b_{nn-1}y_{n-1} + a_{n1}x_1 + a_{n2}x_2 + \ldots + a_{mn}x_m + \varepsilon_n. \end{cases}$$

В таких системах каждое уравнение может рассматриваться самостоятельно, и неизвестные коэффициенты таких уравнений можно найти с помощью классического метода наименьших квадратов.

3. Система взаимозависимых уравнений определяется тем, что эндогенные переменные в одних уравнениях входят в левую часть (т. е. являются результативными признаками), а в других уравнениях — в правую часть (т. е. являются факторными признаками):

$$\begin{cases} y_1 = b_{12}y_2 + b_{13}y_3 + \ldots + b_{1n}y_n + a_{11}x_1 + a_{12}x_2 + \ldots + a_{1m}x_m + \varepsilon_1, \\ y_2 = b_{21}y_1 + b_{23}y_3 + \ldots + b_{2n}y_n + a_{21}x_1 + a_{22}x_2 + \ldots + a_{2m}x_m + \varepsilon_2, \\ y_3 = b_{31}y_1 + b_{32}y_2 + \ldots + b_{3n}y_n + a_{31}x_1 + a_{32}x_2 + \ldots + a_{3m}x_m + \varepsilon_3, \\ \ldots \\ y_n = b_{n1}y_1 + b_{n2}y_2 + \ldots + b_{nn-1}y_{n-1} + a_{n1}x_1 + a_{n2}x_2 + \ldots + a_{nm}x_m + \varepsilon_n. \end{cases}$$

В системе взаимозависимых уравнений значения результативных и факторных переменных формируются **одновременно** под влиянием внешних факторов. Эта система — система одновременных, или совместных, уравнений.

Каждое уравнение системы одновременных уравнений не может рассматриваться как самостоятельная часть системы, вследствие чего применение традиционного метода наименьших квадратов для определения его параметров невозможно, так как нарушаются условия МНК:

1) одновременная зависимость между переменными модели, т. е. в первом уравнении y_1 — это функция от y_2, а во втором уравнении y_2 — это функция от y_1;

2) проблема мультиколлинеарности, т. е. во втором уравнении системы y_2 зависит от x_1, а в других уравнениях обе переменные выступают в качестве факторных;

3) случайные ошибки уравнения коррелируют с результативными переменными.

В результате применения обычного МНК к оцениванию одновременных уравнений оценки неизвестных параметров получаются смещенными и несостоятельными.

Часто приводимым примером системы взаимозависимых уравнений является моделирование одновременного формирования спроса Q^d и предложения Q^s товара в зависимости от его цены P в момент времени t:

$Q^S{}_t = a_0 + a_1 \times P_1 + a_2 \times P_{t-1}$ — уравнение предложения;

$Q^d{}_t = b_0 + b_1 \times P_t + b_2 \times I_t$ — уравнение спроса;

где Q_t^s — предложение товара в момент времени t;

Q_t^d — спрос на товар в момент времени t;

P_t — цена товара в момент времени t;

P_{t-1} — цена товара в предшествующий момент времени $(t - 1)$;

I_t — доход потребителей в момент времени t.

Если рынок находится в состоянии равновесия, то имеет место следующее тождество равновесия: $Q_t^s = Q_t^d$.

1. Структурная и приведенная формы системы одновременных уравнений. Проблема идентификации модели

Уравнения, из которых состоит исходная система одновременных уравнений, называются **структурными уравнениями**, а модель в данном случае имеет структурную форму. С помощью структурной формы модели непосредственно отражают реальный экономический процесс. Коэффициенты уравнений структурной формы называются структурными коэффициентами, или параметрами.

Структурные уравнения могут быть представлены либо поведенческими уравнениями, либо уравнениями—тождествами.

Поведенческие уравнения характеризуют все типы взаимодействия между эндогенными и экзогенными переменными. В поведенческих уравнениях значения параметров являются неизвестными и подлежат оцениванию. Примером поведенческого уравнения является уравнения спроса или предложения:
$Q^S{}_t = a_0 + a_1 \times P_t + a_2 \times P_{t-1}$ или $Q^d{}_t = b_0 + b_1 \times P_t + b_2 \times I_t$.

Тождествами называют равенства, выполняющиеся во всех случаях. Для них характерно, что их вид и значения параметров известны и они не содержат случайной компоненты. Примером уравнения-тождества является тождество равновесия в модели «спрос — предложение»: $Q_t^s = Q_t^d$.

Для определения неизвестных структурных параметров системы одновременных уравнений переходят к приведенной форме модели.

Приведенной формой модели называется система независимых уравнений, в которой все эндогенные переменные выражены только через экзогенные или предопределенные переменные и случайные компоненты, например:

$$\begin{cases} y_1 = C_{11}x_1 + C_{12}x_2 + \ldots + C_{1m}x_m + \varepsilon_1, \\ y_2 = C_{21}x_1 + C_{22}x_2 + \ldots + C_{2m}x_m + \varepsilon_2, \\ \ldots \\ y_n = C_{n1}x_1 + C_{n2}x_2 + \ldots + C_{nm}x_m + \varepsilon_n. \end{cases}$$

Коэффициенты приведенной формы называются приведенными коэффициентами, или параметрами, которые можно оценить традиционным методом наименьших квадратов. С помощью МНК-оценок приведенных коэффициентов определяются оценки структурных коэффициентов.

При переходе от структурной формы модели к приведенной форме возникает проблема идентификации модели.

Проблема идентификации заключается в возможности численной оценки неизвестных коэффициентов структурных уравнений по МНК-оценкам коэффициентов приведенных уравнений.

Исходная система одновременных уравнений является идентифицированной, если все ее уравнения точно идентифицированы. Уравнение является точно идентифицированным, если по оценкам коэффициентов приведенной модели можно однозначно найти оценки коэффициентов структурной модели. Признаком идентифицированности системы является равенство между количеством уравнений, определяющих структурные коэффициенты, и количеством этих коэффициентов, т. е. когда структурная система уравнений является квадратной.

Исходная система одновременных уравнений является сверхидентифицированной, если среди уравнений модели есть хотя бы одно сверхидентифицированное. Уравнение является сверхидентифицированным, если по оценкам коэффициентов приведенной модели можно получить более одного значения для коэффициентов структурной модели.

Исходная система одновременных уравнений является неидентифицированной, если среди уравнений модели есть хотя бы одно неидентифицированное. Уравнение является неидентифицированным, если по оценкам коэффициентов приведенной модели невозможно рассчитать оценки коэффициентов структурной модели.

2. Необходимые и достаточные условия идентификации модели

Необходимые и достаточные условия идентификации применяются только к структурной форме модели.

Введем обозначения:

1) N — количество предопределенных переменных в модели;
2) n — количество предопределенных переменных в уравнении, проверяемом на идентифицируемость;
3) M — количество эндогенных переменных в модели;
4) m — количество эндогенных переменных в уравнении, проверяемом на идентифицируемость;
5) K — матрица коэффициентов при переменных, не входящих в уравнение, проверяемое на идентифицируемость.

Первое необходимое условие идентифицируемости уравнения модели.

Уравнение модели идентифицируемо в том случае, если оно исключает хотя бы $N-1$ предопределенную переменную модели, т. е.:

$$(N-n)+(M-m)\geq N-1.$$

Второе необходимое условие идентифицируемости уравнения модели.

Уравнение модели идентифицируемо в случае, если количество предопределенных переменных, не входящих в данное уравнение, будет не меньше числа эндогенных переменных этого уравнения минус единица, т. е.:

$$N-n\geq m-1.$$

Достаточное условие идентифицируемости уравнения модели.

Уравнение модели идентифицируемо в случае, если ранг матрицы K равен $N-1$.

Ранг матрицы — размер наибольшей ее квадратной подматрицы, определитель которой не равен нулю.

Исходя из перечисленных условий идентификации можно сформулировать необходимые и достаточные условия идентифицируемости уравнения модели:

1) если $M-m>n-1$ и ранг матрицы K равен $N-1$, то уравнение модели считается сверхидентифицированным;
2) если $M-m=n-1$ и ранг матрицы K равен $N-1$, то уравнение модели считается точно идентифицированным;
3) если $M-m\geq n-1$ и ранг матрицы K меньше $N-1$, то уравнение модели считается неидентифицированным;
4) если $M-m<n-1$, то уравнение считается неидентифицированным, так как ранг матрицы K будет меньше $N-1$.

Рассмотрим пример идентификации на основе структурной модели спроса и предложения.

$Q^S{}_t = a_0 + a_1 \times P_t + a_2 \times P_{t-1}$ — уравнение предложения;

$Q^d{}_t = b_0 + b_1 \times P_t + b_2 \times I_t$ — уравнение спроса;

$Q_t^s = Q_t^d$ — тождество равновесия.

Учитывая тождество, систему можно записать:

$$\begin{cases} Q_t = a_0 + a_1 \times P_t + a_2 \times P_{t-1} + \varepsilon_{1t}, \\ Q_t = b_0 + b_1 \times P_t + b_2 \times I_t + \varepsilon_{2t}. \end{cases}$$

Количество эндогенных переменных модели равно $M = 2$ (P_t и Q_t), количество предопределенных переменных модели равно $N = 2$ ($P_t - 1$ и I_t).

Проверим выполнение первого условия идентифицируемости.
Для функции спроса — $m = 2$, $n = 1$.
Тогда $(N-n)+(M-m)=(2-1)+(2-2)=1=(N-1)=1$, уравнение точно идентифицировано.
Для функции предложения — $m = 2$, $n = 1$.
Тогда $(N-n)+(M-m)=(2-1)+(2-2)=1=(N-1)=1$, уравнение точно идентифицировано.

Проверим выполнение второго необходимого условия идентифицируемости.
Для функции спроса — $m = 2$, $n = 1$.
Тогда $N - n = 2 - 1 = 1 = m - 1 = 2 - 1 = 1$, уравнение точно идентифицировано.
Для функции предложения — $m = 2$, $n = 1$.
Тогда $N - n = 2 - 1 = 1 = m - 1 = 2 - 1 = 1$, уравнение точно идентифицировано.

Проверим выполнение достаточного условия идентифицируемости. В данном случае достаточно, чтобы хотя бы один из коэффициентов матрицы K не был равен нулю, так как $M - 1 = 1$.

В первом уравнении исключена переменная I_t. Матрица $K = [b_2]$. Определитель данной матрицы не равен нулю, следовательно, $rank = 1 = M - 1$ и уравнение идентифицировано.

Во втором уравнении исключена переменная P_{t-1}. Матрица $K = [\alpha_2]$. Определитель данной матрицы не равен нулю, следовательно, $rank = 1 = M - 1$, и уравнение идентифицировано.

Уравнения спроса и предложения точно идентифицированы, следовательно, система уравнений в целом точно идентифицирована.

Составим приведенную форму данной системы:

$$\begin{cases} Q = A_1 + A_2 \times I_t + A_3 \times P_{t-1} + v_1; \\ P_t = B_1 + B_2 \times I_t + B_3 \times P_{t-1} + v_2. \end{cases}$$

ЛЕКЦИЯ № 26. Косвенный и двухшаговый метод наименьших квадратов. Примеры их применения. Инструментальные переменные

Каждое уравнение системы одновременных уравнений не может рассматриваться как самостоятельная часть системы, поэтому применение традиционного метода наименьших квадратов для определения его параметров невозможно, так как нарушаются условия МНК:

1) одновременная зависимость между переменными модели, т. е. в первом уравнении y_1 — это функция от y_2, а во втором уравнении y_2 — это функция от y_1;

2) проблема мультиколлинеарности, т. е. во втором уравнении системы y_2 зависит от x_1, а в других уравнениях обе переменные выступают в качестве факторных;

3) случайные ошибки уравнения коррелируют с результативными переменными.

Применение МНК к оцениванию параметров одновременных уравнений дает смещенные и несостоятельные оценки.

Для получения оценок параметров системы одновременных уравнений, удовлетворяющих свойствам эффективности, несмещенности и состоятельности, применяется косвенный метод наименьших квадратов (КМНК). КМНК пользуются в случае, если структурная форма модели является точно идентифицированной.

Алгоритм КМНК включает в себя следующие шаги:

1) на основе структурной формы модели составляется ее приведенная форма, все параметры которой выражены через структурные коэффициенты;

2) приведенные коэффициенты каждого уравнения оцениваются обычным методом наименьших квадратов;

3) на основе оценок приведенных коэффициентов определяются оценки структурных коэффициентов через приведенные уравнения.

Рассмотрим применение косвенного метода на примере структурной **модели спроса и предложения вида**:

$Q_t = a_0 + a_1 \times P_t + a_2 \times P_{t-1} + \varepsilon_{1t}$ — уравнение предложения;

$Q_t = b_0 + b_1 \times P_t + b_2 \times I_t + \varepsilon_{2t}$ — уравнение спроса.

Эндогенными переменными в данной модели являются Q_t — объем товара и P_t — цена товара, а предопределенными переменными являются I_t — доход потребителей и P_{t-1} — цена товара в предыдущий момент времени. ε_{1t} и ε_{2t} являются случайными компонентами модели, а a_0, a_1, a_2, b_0, b_1, b_2 — структурные параметры модели.

Структурная модель спроса и предложения является точно идентифицированной, поэтому для оценивания ее параметров можно применить косвенный метод наименьших квадратов.

1. Запишем модель спроса и предложения в приведенной форме:

$$\begin{cases} Q = A_1 + A_2 \times I_t + A_3 \times P_{t-1} + v_1; \\ P_t = B_1 + B_2 \times I_t + B_3 \times P_{t-1} + v_2. \end{cases}$$

2. Оценки коэффициентов приведенной формы определяются с помощью обычного метода наименьших квадратов. Систему нормальных уравнений для определения коэффициентов первого уравнения приведенной системы можно записать в виде:

$$\begin{cases} n \times A_1 + A_2 \times \sum I_t + A_3 \times \sum P_{t-1} = \sum Q, \\ A_1 \times \sum I_t + A_2 \times \sum I_t^2 + A_3 \times \sum I_t \times P_{t-1} = \sum Q \times I_t, \\ A_1 \times \sum P_{t-1} + A_2 \times \sum P_t \times I_t + A_3 \times \sum P_{t-1}^2 = \sum Q \times P_{t-1}. \end{cases}$$

Система нормальных уравнений для определения коэффициентов второго уравнения приведенной системы записывается аналогично. Решением данных систем нормальных уравнений будут являться численные оценки приведенных коэффициентов A_1, A_2, A_3 и B_1, B_2, B_3.

3. Чтобы по оценкам приведенных коэффициентов получить оценки структурных коэффициентов первого уравнения, необходимо из второго приведенного уравнения выразить переменную I_t и подставить полученное выражение в первое уравнение приведенной формы. Чтобы получить оценки структурных коэффициентов второго уравнения, необходимо из второго приведенно-

го уравнения выразить переменную P_{t-1} и подставить полученное выражение в первое уравнение приведенной формы.

1. Двухшаговый метод наименьших квадратов

Если уравнение сверхидентифицировано, то оценки его параметров нельзя определить косвенным методом наименьших квадратов. Обычный МНК также применять нельзя в связи с нарушением основных предпосылок его применения. В данном случае могут использоваться различные методы оценивания неизвестных параметров, однако наиболее простым и распространенным является двухшаговый метод наименьших квадратов (ДМНК).

Двухшаговый метод наименьших квадратов реализуется в несколько этапов:

1) на основе структурной формы модели составляется ее приведенная форма;

2) с помощью обычного метода наименьших квадратов определяются оценки коэффициентов приведенных уравнений;

3) рассчитываются значения тех эндогенных переменных, которые выступают в качестве факторных в сверхидентифицированном уравнении;

4) с помощью обычного метода наименьших квадратов определяются все структурные параметры уравнений системы через предопределенные переменные, входящие в это уравнение в качестве факторов, и значения эндогенных переменных, полученных на предыдущем шаге.

Данный метод наименьших квадратов называется двухшаговым, потому что МНК используется дважды: первый раз для определения оценок эндогенных переменных приведенной формы и второй раз для определения оценок структурных параметров уравнений системы.

Сверхидентифицированная структурная модель может быть двух видов:

1) помимо сверхидентифицированного уравнения, в модели также содержатся точно идентифицированные уравнения;

2) все уравнения модели являются сверхидентифицированными.

В первом случае оценки структурных коэффициентов точно идентифицированного уравнения определяются на основании системы приведенных уравнений. Во втором случае оценки структурных коэффициентов системы определяются с помощью двухшагового метода наименьших квадратов.

Если все уравнения системы являются точно идентифицированными, то оценки структурных коэффициентов, полученные КМНК, будут совпадать с оценками, полученными ДМНК.

Применение обычного МНК к оценке параметров сверхидентифицированного уравнения невозможно, так как в уравнении в качестве факторной переменной выступает эндогенная переменная y_t. В данном случае можно перейти к такой переменной, которая удовлетворяла бы условиям нормальной линейной регрессионной модели. Делается это с помощью **метода инструментальных переменных (IV – Instrumental variables)**. Его суть состоит в следующем.

Переменная y_t из правой части уравнения, для которой нарушается предпосылка МНК, заменяется на новую переменную y_t^*, удовлетворяющую следующим двум требованиям:

1) она должна тесно коррелировать с переменной y_t: $cov(y_t, y_t^*) \neq 0$;

2) она не должна коррелировать со случайной компонентой ε_t: $cov(y_t^*, \varepsilon)$.

Переменные, удовлетворяющие данным требованиям, называются инструментальными переменными.

Далее оценивают уравнение регрессии с новой инструментальной переменной y_t^* с помощью обычного МНК.

Обычно метод инструментальных переменных используется в приведенной форме сверхидентифицированного уравнения. Поэтому ДМНК также называется **обобщенным методом инструментальных переменных**.

ДМНК может применяться и для оценки точно идентифицированных уравнений.

2. Пример применения косвенного метода наименьших квадратов для оценки параметров точно идентифицированного уравнения

Рассмотрим применение косвенного метода на примере структурной модели спроса и предложения вида:

$Q_t = a_0 + a_1 \times P_t + a_2 \times P_{t-1} + \varepsilon_{1t}$ — уравнение предложения;

$Q_t = b_0 + b_1 \times P_t + b_2 \times I_t + \varepsilon_{2t}$ — уравнение спроса.

Как было показано раньше, структурная модель спроса и предложения является точно идентифицированной, поэтому

для оценивания ее параметров можно применить косвенный метод наименьших квадратов.

Эндогенными переменными в данной модели являются Q_t — объем товара и P_t — цена товара, а предопределенными переменными являются I_t — доход потребителей и P_{t-1} — цена товара в предыдущий момент времени. ε_{1t} и ε_{2t} являются случайными компонентами, а a_0, a_1, a_2, b_0, b_1, b_2 — структурными параметрами модели.

С помощью косвенного МНК найдем оценки структурных параметров уравнений системы.

1. Запишем исходную модель спроса и предложения в приведенной форме:

$$\begin{cases} Q = A_1 + A_2 \times I_t + A_3 \times P_{t-1} + v_1, \\ P_t = B_1 + B_2 \times I_t + B_3 \times P_{t-1} + v_2. \end{cases}$$

2. Оценки коэффициентов приведенной формы определяются с помощью обычного метода наименьших квадратов.

В этом случае систему нормальных уравнений для определения коэффициентов первого уравнения приведенной системы можно записать в виде:

$$\begin{cases} n \times A_1 + A_2 \times \sum I_t + A_3 \times \sum P_{t-1} = \sum Q, \\ A_1 \times \sum I_t + A_2 \times \sum I_t^2 + A_3 \times \sum I_t \times P_{t-1} = \sum Q \times I_t, \\ A_1 \times \sum P_{t-1} + A_2 \times \sum P_t \times I_t + A_3 \times \sum P_{t-1}^2 = \sum Q \times P_{t-1}. \end{cases}$$

Подставим значения переменных, определенных по таблице 4, в уравнения системы:

$$\begin{cases} 10A_1 + 129A_2 + 62A_3 = 655, \\ 129A_1 + 16\,641A_2 + 7\,998A_3 = 84\,495, \\ 62A_1 + 8\,127A_2 + 3\,844A_3 = 40\,610. \end{cases}$$

Решив данную систему нормальных уравнений методом Крамера, получаем оценки неизвестных коэффициентов первого уравнения приведенной формы:

$$A_1 = 44{,}7,$$
$$A_2 = -1{,}12,$$
$$A_3 = 2{,}15.$$

Система нормальных уравнений для определения коэффициентов второго уравнения приведенной системы записывается аналогично.

Ее решением будут следующие числа:

$$B_1 = 89{,}6,$$
$$B_2 = -7{,}2,$$
$$B_3 = 1{,}67.$$

Окончательно оцененные уравнения приведенной формы модели можно записать в виде:

$$\begin{cases} Q_t = 44{,}7 - 1{,}12 \times I_t + 2{,}15 \times P_{t-1}, \\ P_t = 89{,}6 - 7{,}2 \times I_t + 1{,}67 \times P_{t-1}. \end{cases}$$

3. На основании полученных оценок параметров приведенных уравнений можно определить оценки структурных параметров модели. Рассчитаем первое уравнение структурной формы модели, т. е. зависимость Q_t от P_t и P_{t-1}. Необходимо выразить из второго уравнения приведенной формы переменную I_t:

$$I_t = \frac{89{,}6}{7{,}2} + \frac{1{,}67}{7{,}2} P_{t-1} - \frac{1}{7{,}2} P_t = 12{,}4 + 0{,}23 P_{t-1} - 0{,}14 P_t.$$

Подставим данное выражение в первое уравнение приведенной формы модели:

$$Q_t = 44{,}7 + 1{,}12 \times (12{,}4 + 0{,}23 P_{t-1} - 0{,}14 P_t) + 2{,}15 P_{t-1} =$$
$$= 172{,}2 + 2{,}4 P_{t-1} - 0{,}16 P_t$$

Оценки первого уравнения структурной формы записывают:

$$a_0 = 172{,}2,$$
$$a_1 = -0{,}16,$$
$$a_2 = 2{,}4.$$

Для определения параметров второго уравнения структурной формы, представляющего собой зависимость переменной Q_t от P_t и I_t, необходимо выразить из второго уравнения приведенной формы модели переменную P_{t-1}:

$$P_{t-1} = -\frac{89{,}6}{1{,}67} + \frac{7{,}2}{1{,}67} I_t + \frac{1}{1{,}67} P_t = -53{,}6 + 4{,}3 I_t + 0{,}6 P_t.$$

Далее подставим выражение в первое уравнение приведенной формы модели:

$$Q_t = 44{,}7 + 1{,}12 I_t + 2{,}15 \times (-53{,}6 + 4{,}3 I_t + 0{,}6 P_t) =$$
$$= -70{,}5 + 10{,}4 I_t + 1{,}3 P_t$$

Оценки второго уравнения структурной формы модели записывают:

$$b_0 = -70{,}5;\ b_1 = 1{,}3;\ b_2 = 10{,}4.$$

В результате проведенных вычислений **структурную форму модели** спроса и предложения можно записать в следующем виде:

$$\begin{cases} Q_t = 172{,}2 - 0{,}16\,P_t + 2{,}4\,P_{t-1}, \\ Q_t = -70{,}5 + 1{,}3\,P_t + 10{,}4\,I_t. \end{cases}$$

Рассчитаем коэффициенты множественной детерминации для каждого структурного уравнения как показатели качества построенной модели.

Для уравнения предложения: $R^2 = 0{,}1$, т. е. построенное уравнение на 81% объясняет дисперсию зависимой переменной в общем объеме ее дисперсии.

Для уравнения спроса: $R^2 = 0{,}76$, т. е. построенное уравнение на 76% объясняет дисперсию зависимой переменной в общем объеме ее дисперсии.

3. Пример применения двухшагового метода наименьших квадратов к модели, включающей сверхидентифицированное уравнение

Если система содержит хотя бы одно сверхидентифицированное уравнение, то ни обычный, ни косвенный метод наименьших квадратов применять для оценки коэффициентов структурной формы модели нельзя. В данном случае необходимо применять двухшаговый метод наименьших квадратов, который реализуется в несколько этапов.

Рассмотрим конкретный пример оценивания сверхидентифицированного уравнения системы с помощью ДМНК. Для этого в модель спроса и предложения введем новую независимую переменную R_t, которая характеризует благосостояние потребителей:

$$\begin{cases} Q_t = a_0 + a_1 \times P_t + a_2 \times P_{t-1} + \varepsilon_{1t}, \\ Q_t = b_0 + b_1 \times P_t + b_2 \times I_t + a_3 \times R_t + \varepsilon_{2t}. \end{cases}$$

Таким образом, первое уравнение данной модели является сверхидентифицированным, что не позволяет применять к его оценке МНК и КМНК.

Определение оценок функции предложения будет проходить в несколько этапов:

1) запишем исходную модель спроса и предложения в приве-

денной форме:

$$\begin{cases} Q = A_1 + A_2 \times I_t + A_3 \times P_{t-1} + A_4 \times R_t + v_1, \\ P_t = B_1 + B_2 \times I_t + B_3 \times P_{t-1} + B_4 \times R_t + v_2. \end{cases}$$

Из второго уравнения приведенной формы модели можно найти расчетные значения $\tilde{P}_t$:

$$\tilde{P}_t = \tilde{B}_1 + \tilde{B}_2 \times I_t + \tilde{B}_3 \times P_{t-1} + \tilde{B}_4 \times R_t.$$

Тогда второе уравнение приведенной формы можно записать в виде:

$$P_t = \tilde{P}_t + v_2.$$

С учетом расчетных значений $\tilde{P}_t$ сверхидентифицированное уравнение предложения можно записать в виде:

$$Q_t = a_0 + a_1 \times \left(\tilde{P}_t + v_2\right) + a_2 \times P_{t-1} + \varepsilon_{1t}$$

или

$$Q_t = a_0 + a_1 \times \tilde{P}_t + a_2 \times P_{t-1} + u_{1t},$$

где $u_{1t} = \varepsilon_{1t} + a_1 \times v_2$.

Следовательно, переменную $\tilde{P}_t$ можно считать инструментальной переменной, так как:

а) $\tilde{P}_t$ тесно коррелирует с P_t, потому что является линейной комбинацией независимых переменных I_t, P_{t-1} и R_t;

б) $\tilde{P}_t$ не коррелирует со случайной составляющей u_{1t};

2) с помощью обычного МНК определим коэффициенты уравнений приведенной формы модели:

$$\tilde{Q}_t = 43{,}08 + 2{,}15 I_t - 1{,}1 P_{t-1} + 0{,}13 R_t,$$

$$\tilde{P}_t = 6{,}02 - 0{,}004 I_t - 0{,}15 P_{t-1} + 0{,}1 R_t;$$

3) определим расчетные значения $\tilde{P}_t$, подставив во второе уравнение приведенной формы фактические значения переменных I_t, P_{t-1} и R_t и добавим их к данным таблицы 5;

4) применим обычный метод наименьших квадратов к урав-

нению предложения с инструментальной переменной:

$$Q_t = a_0 + a_1 \times \tilde{P}_t + a_2 \times P_{t-1} + u_{1t}.$$

В результате получим оцененное структурное уравнение предложения:

$$Q_t = 81 - 2{,}93 \times \tilde{P}_t + 0{,}45 \times P_{t-1}.$$

Рассчитаем **коэффициент множественной детерминации** для данного уравнения: $R^2 = 0{,}944$.

Таким образом, полученное уравнение предложения на 94,4% объясняет дисперсию зависимой переменной в общем объеме ее дисперсии.

Неизвестные коэффициенты точно идентифицированного уравнения спроса можно найти и с помощью двухшагового МНК, и с помощью косвенного МНК.

Найдем оценки структурного уравнения спроса с помощью двухшагового МНК:

$$Q_t = b_0 + b_1 \times \tilde{P}_t + b_2 \times I_t + a_3 \times R_t + \varepsilon_{2t}.$$

Тогда

$$Q_t = -1{,}44 + 7{,}4 \times \tilde{P}_t + 2{,}18 \times I_t - 0{,}6 \times R_t + \varepsilon_{2t}.$$

Рассчитаем коэффициент множественной детерминации для данного уравнения: $R^2 = 0{,}872$.

Таким образом, полученное с помощью ДМНК уравнение спроса на 87,2% объясняет дисперсию зависимой переменной в общем объеме ее дисперсии.

Найдем оценки структурного уравнения спроса с помощью косвенного метода наименьших квадратов.

Выразим из второго уравнения приведенной формы модели переменную P_{t-1}:

$$P_{t-1} = \frac{6{,}02}{0{,}5} - \frac{0{,}004}{0{,}15} I_t + \frac{0{,}1}{0{,}15} R_t + \frac{1}{0{,}15} P_t =$$
$$= 40 - 0{,}026 I_t + 0{,}67 R_t + 6{,}7 P_t.$$

Подставим данное выражение в первое уравнение приведенной формы модели вместо P_{t-1}:

$$Q_t = -1{,}44 + 2{,}18 I_t - 0{,}607 R_t - 7{,}37 P_t.$$

Таким образом, оценки структурного уравнения спроса, полученные разными методами, абсолютно одинаковы.

Запишем оцененную структурную форму модели:

$$Q_t = 81 - 2{,}93 \times \tilde{P}_t + 0{,}45 \times P_{t-1},$$
$$Q_t = -1{,}44 + 2{,}18 I_t - 0{,}607 R_t - 7{,}37 P_t.$$

4. Инструментальные переменные

В основе возникновения метода инструментальных переменных лежат критические замечания Милтона Фридмана об оценивании кейнсианской функции потребления.

Функцию потребления в общем виде можно записать следующим образом:

$$C_{it} = \alpha + \beta y_{it} + \varepsilon_{it}, \qquad (1)$$

где C_{it} — объем потребления i-го домашнего хозяйства в t-ом году;

y_{it} — объем доходов i-го домашнего хозяйства в t-ом году;

β — коэффициент предельной склонности к потреблению ($0 < \beta < 1$);

α — коэффициент автономного потребления.

В соответствии с кейнсианской трактовкой данной модели потребления коэффициент автономного потребления α равен нулю.

Выделим основные **недостатки модели** (1):

1) оценки параметров уравнения регрессии, полученные обычным методом наименьших квадратов, меняются год от года;

2) в ходе экспериментов было доказано, что оценка коэффициента β для фермеров ниже, чем для городского населения.

Рассмотрим объяснение невозможности применения метода наименьших квадратов к оцениванию параметров модели (1) на основе теории постоянных доходов Фридмана.

Пусть

$$y_{it} = y_{it}^P + y_{it}^T,$$

$$C_{it} = C_{it}^P + C_{it}^T,$$

где индекс P означает постоянство (permanent), а индекс T означает непостоянство (transitory) переменных.

Предположим, что доход y_{it} и потребление C_{it} являются случайными величинами с нулевым математическим ожиданием

и дисперсиями $G^2_{y^T}$ и $G^2_{C^T}$ соответственно, т. е.

$$y_{it} \sim (0 : G^2_{y^T}) \text{ и } C_{it} \sim (0 : G^2_{c^T}).$$

Данные величины связаны соотношением по Фридману:

$$C^P_{it} = \alpha + \beta\, y^P_{it}. \qquad (2)$$

Возникает вопрос: верна ли функция потребления (2) при существовании функции (1).

Представим функцию потребления (2) в виде следующего равенства:

$$\underbrace{C^P_{it} + C^T_{it}}_{C_{it}} - C^T_{it} = \alpha + \beta \left(\underbrace{y^P_{it} + y^T_{it}}_{y_{it}} - y^T_{it} \right).$$

Тогда

$$C_{it} = \alpha + \beta\, y_{it} + C^T_{it} - \beta\, y^T_{it}.$$

Обозначим $C^T_{it} - \beta\, y^T_{it}$ через u_{it}. Таким образом, уравнение (2) после преобразований примет вид:

$$C_{it} = \alpha + \beta\, y_{it} + u_{it}.$$

В модели потребления вида (1) ε_{it} является независимой случайной составляющей, а в уравнении вида (2) нарушается первая предпосылка нормальной регрессионной модели, так как u_{it} коррелирован с $\beta\, y_{it}$.

Рассмотрим ковариацию между переменной y_{it} и u_{it}:

$$cov\left(y_{it}, u_{it}\right) = cov\left(y^P_{it} + y^T_{it}; C^T_{it} - \beta\, y^T_{it}\right) =$$
$$= cov\left(y^P_{it}; C^T_{it}\right) + cov\left(y^P_{it}; -\beta\, y^T_{it}\right) + cov\left(y^T_{it}; C^T_{it}\right) -$$
$$- \; cov\left(y^T_{it}; \beta\, y^T_{it}\right) = -\beta\, G^2_{y^T}.$$

Запишем МНК-оценку параметра β модели регрессии (1):

$$\tilde{\beta} = \beta - \frac{\beta G_{y^T}^2}{D(y_{it})} = \beta - \frac{\beta G_{y^T}^2}{G_{y^P}^2 + G_{y^T}^2}.$$

Таким образом, метод наименьших квадратов в данном случае будет всегда давать заниженные оценки параметров, поэтому им пользоваться нельзя.

В 1950-е гг. М. Фридман предложил новый метод для оценивания параметров подобных функций. Он назвал его **методом инструментальных переменных** (Instrumental Variables — IV).

Его суть состоит в следующем. Переменная y_{it} из правой части уравнения, для которой нарушается первая предпосылка нормальной регрессионной модели, заменяется на новую переменную, называемую инструментом:

$$y'_{it} = y_{it}^{P'} + y_{it}^{T'},$$

$$u_{it} = C_{it}^{T} - \beta\, y_{it}^{T}.$$

В данном случае случайная ошибка u_{it} и y_{it} не коррелированны между собой, но коррелированы с переменной y'_{it}, которая называется инструментом. Индекс y' означает, что переменная дохода относится к следующему году.

Оценка, полученная с помощью метода инструментальных переменных, выглядит следующим образом:

$$\tilde{\beta}_{IV} = \frac{\sum (C_{it} - \bar{C})}{(y'_{it} - \bar{y}'_{it})} : \frac{\sum (y_{it} - \bar{y})}{(y'_{it} - \bar{y}'_{it})}.$$

Данная оценка по своим свойствам превосходит обычную МНК-оценку.

В общем случае инструментальная переменная z должна **удовлетворять свойствам**:

1) она должна тесно коррелировать с переменной y: $cov(y, z) \neq 0$;

2) она не должна коррелировать со случайной ошибкой ε_t: $cov(z, \varepsilon) = 0$.

Оценка коэффициента регрессии определяется по формуле:

$$\tilde{\beta}_{IV} = (Z^T Y)^{-1} Z^T Y.$$

ЛЕКЦИЯ № 27. Динамические эконометрические модели (ДЭМ). Модель авторегрессии. Характеристика моделей с распределенным лагом

К динамическим эконометрическим моделям относят те модели, которые в настоящий момент времени учитывают значения входящих в них переменных, относящихся не только к текущему, но и к предыдущему моментам времени. Регрессионные модели вида:

$$y_t = f\left(x_t, x_{t-l}\right),$$

$$y_t = f\left(x_t, y_{t-l}\right)$$

являются динамическими эконометрическими моделями, а регрессия вида:

$$y_t = f\left(x_1 \ldots x_n\right) = f\left(x_i\right)$$

не является ДЭМ.

Выделяют два основных типа ДЭМ:

1) модели, в которых значения переменных, относящихся к прошлым моментам времени (лаговые значения), включены в модель с текущими значениями этих переменных. К таким моделям относятся:

а) **модель авторегрессии.** Это динамическая эконометрическая модель, в которой в качестве факторных переменных содержатся лаговые значения результативной переменной. Примером модели авторегрессии является модель:

$$y_t = \beta_0 + \beta_1 \times x_t + \delta_1 \times y_{t-1} + \varepsilon_t;$$

б) **модель с распределенным лагом.** Это динамическая эконометрическая модель, включающая текущие и лаговые значения факторных переменных. Примером модели с распределенным лагом является:

$$y_t = \beta_0 + \beta_1 \times x_t + \beta_2 \times x_{t-1} + \ldots + \beta_L \times x_{t-L} + \varepsilon_t,$$

где L — это величина временного лага (запаздывания) между рядами;

2) модели, включающие переменные, отражающие предполагаемый или желаемый уровень результативной переменной или одного из факторных признаков в определенный момент времени ($t + 1$). Этот уровень является неизвестным и определяется на основании той информации, которая имеется в наличии на предшествующий момент времени t. Предполагаемые значения переменных рассчитываются различными способами. В зависимости от способа расчета данных переменных различают следующие виды моделей:

а) **модель адаптивных ожиданий (МАО)**, учитывающая предполагаемое (или желаемое) значение факторной переменной x^*_{t+1}. В общем виде модель адаптивных ожиданий записывают так:

$$y_t = \beta_0 + \beta_1 \times x^*_{t+1} + \varepsilon_t.$$

Примером МАО служит влияние предполагаемой в будущем периоде ($t + 1$) индексации заработных плат и пенсий на текущие цены;

б) **модель частичной (неполной) корректировки (МЧК)**, учитывающая предполагаемое (или желаемое) значение результативной переменной y_t^*. В общем виде модель частичной корректировки можно записать так:

$$y^*_t = \beta_0 + \beta_1 \times x_t + \varepsilon_t.$$

Примером модели частичной корректировки является зависимость желаемого объема дивидендов y_t^* от фактического текущего объема прибыли x_t. Данная МЧК более известна как модель Литнера.

Особенность динамических эконометрических моделей состоит в том, что для оценивания их неизвестных параметров обычный метод наименьших квадратов неприменим по различным причинам.

Для оценивания коэффициентов модели авторегрессии применяется метод инструментальных переменных, который позволяет получить наиболее оптимальные в данных условиях оценки.

Для моделей с распределенным лагом в зависимости от структуры лага для оценивания параметров применяются **метод Алмона** и **метод Койка**.

Суть данных методов состоит в том, чтобы преобразовать исходную модель с распределенным лагом в модель авторегрессии, которую можно оценить с помощью метода инструментальных переменных.

Модель адаптивных ожиданий и модель частичной корректировки также с целью нахождения неизвестных параметров преобразуются в вид модели авторегрессии.

1. Модель авторегрессии и оценивание ее параметров

Авторегрессионная модель — это динамическая эконометрическая модель, в которой в качестве факторных переменных содержатся лаговые значения результативной переменной. Примером модели авторегрессии является:

$$y_t = \beta_0 + \beta_1 \times x_t + \delta_1 \times y_{t-1} + \varepsilon_t.$$

В авторегрессионной модели коэффициент β_1 характеризует краткосрочное изменение переменной y под влиянием изменения переменной x на единицу своего измерения.

Коэффициент δ_1 характеризует изменение переменной y под влиянием своего изменения в предыдущий момент времени $(t-1)$. Произведение регрессионных коэффициентов $(\beta_1 \times \delta_1)$ называется промежуточным мультипликатором. Этот показатель характеризует общее абсолютное изменение результативной переменной y в момент времени $(t+1)$.

Показатель

$$\beta = \beta_1 + \beta_1 \times \delta_1 + \beta_1 \times \delta_1^2 + \beta_1 \times \delta_1^3 + \ldots$$

называется долгосрочным мультипликатором. Он характеризует общее абсолютное изменение результативной переменной y в долгосрочном периоде.

В большинство моделей авторегрессии вводится условие стабильности, которое состоит в том, что $|\delta_1| < 1$. При наличии бесконечного лага будет выполняться следующее равенство:

$$\beta = \beta_1 \times \left(\delta_1 + \delta_1^2 + \delta_1^3 + \ldots\right) = \frac{\beta_1}{1-\delta_1}.$$

Нормальная линейная регрессионная модель строится исходя из предпосылки о том, что все факторные переменные являются величинами независимыми от случайной ошибки модели.

В случае авторегрессионных моделей данное условие нарушается, так как переменная y_{t-1} частично зависит от случайной ошибки модели ε_t. Применение метода наименьших квадратов для оценивания неизвестных параметров авторегрессионного уравнения невозможно, так как это приводит к получению смещенной оценки коэффициента при переменной y_{t-1}.

Для оценивания параметров авторегрессионного уравнения применяется **метод инструментальных переменных (IV – Instrumental variables)**. Его суть состоит в следующем.

Переменная y_{t-1} из правой части уравнения, для которой нарушается предпосылка МНК, заменяется на новую переменную z, удовлетворяющую следующим требованиям:

1) она должна тесно коррелировать с переменной y_{t-1}: $cov(y_{t-1}, z) \neq 0$;

2) она не должна коррелировать со случайной ошибкой ε_t: $cov(z, \varepsilon) = 0$.

Далее оценивают регрессию с новой инструментальной переменной z с помощью обычного метода наименьших квадратов.

Оценка коэффициента регрессии определяется так:

$$\tilde{\beta}_{IV} = \left(Z^T Y\right)^{-1} Z^T Y.$$

Рассмотрим пример применения метода инструментальных переменных для модели авторегрессии вида:

$$y_t = \beta_0 + \beta_1 \times x_t + \delta_1 \times y_{t-1} + \varepsilon_t.$$

В данной модели переменная y_t зависит от переменной x_t, из чего можно сделать вывод, что переменная y_{t-1} зависит от переменной x_{t-1}. Выразим эту зависимость через обычную парную регрессионную модель:

$$y_{t-1} = k_0 + k_1 \times x_{t-1} + u_t,$$

где k_0, k_1 — неизвестные коэффициенты регрессии;

u_t — случайная ошибка регрессионного уравнения.

Обозначим выражение $k_0 + k_1 \times x_{t-1}$ через переменную z_{t-1}.

Регрессия для y_{t-1} записывается:

$$y_{t-1} = z_{t-1} + u_t.$$

Новая переменная z_{t-1} удовлетворяет свойствам, предъявляемым к инструментальным переменным: она тесно коррелирует

с переменной y_{t-1}, т. е. $cov(z_{t-1}, y_{t-1}) \neq 0$, и не коррелирует со случайной ошибкой исходной авторегрессионной модели ε_t, т. е. $cov(\varepsilon_t, z_{t-1})$.

Исходная модель авторегрессии может быть записана так:

$$y_t = \beta_0 + \beta_1 \times x_t + \delta_1 \times \left(k_0 + k_1 x_{t-1} + u_t\right) + \varepsilon_t = \\ = \beta_0 + \beta_1 \times x_t + \delta_1 \times z_{t-1} + v_t,$$

где $v_t = \delta_1 \times u_t + \varepsilon_t$.

Оценки неизвестных коэффициентов преобразованной модели находятся с помощью обычного метода наименьших квадратов. Они являются оценками неизвестных коэффициентов исходной авторегрессионной модели.

2. Характеристика моделей с распределенным лагом

Модель с распределенным лагом — динамическая эконометрическая модель, включающая текущие и лаговые значения факторных переменных. Примером модели с распределенным лагом является:

$$y_t = \beta_0 + \beta_1 \times x_t + \beta_2 \times x_{t-1} + \ldots + \beta_L \times x_{t-L} + \varepsilon_t.$$

Модели с распределенным лагом позволяют определить влияние изменения факторной переменной x на результативную переменную x, т. е. изменение x в момент времени t будет оказывать влияние на значение переменной y в течение L следующих моментов времени.

Параметр регрессии β_1 называется краткосрочным мультипликатором. Он показывает среднее абсолютное изменение y_t при изменении x_t на единицу своего измерения в конкретный момент времени t при исключении влияния лаговых значений фактора x.

Параметр регрессии β_2 характеризует среднее абсолютное изменение переменной y_t в результате изменения переменной x_t на единицу своего измерения в момент времени $t - 1$.

Сумма параметров $(\beta_1 + \beta_2)$ называется промежуточным мультипликатором. Он отражает совокупное влияние фактора x на переменную y в момент времени $t + 1$, т. е. изменение x на единицу в момент времени t вызывает изменение y на β_1 единиц в момент времени t и изменение y на β_2 в момент времени $t + 1$.

Сумма параметров $\beta = \beta_1 + \beta_2 + \ldots + \beta_L$ называется долгосрочным мультипликатором. Он характеризует общее изменение переменной *y* в момент времени $(t + L)$ под воздействием изменения переменной *x* на единицу своего измерения в момент времени *t*.

Средним лагом называется средний период времени, в течение которого будет происходить изменение результативной переменной под влиянием изменения фактора *x* в момент *t*:

$$\bar{L} = \sum_{i=0}^{L} i \times \frac{\beta_i}{\beta}.$$

Если величина среднего лага небольшая, то *y* достаточно быстро реагирует на изменение фактора *x*. Если величина среднего лага большая, то факторная переменная *x* медленно воздействует на результативную переменную *y*.

Медианный лаг — период времени, в течение которого с момента начала изменения факторного признака *x* будет реализована половина его общего воздействия на результативный признак.

Оценивание неизвестных коэффициентов моделей с распределенным лагом МНК в большинстве случаев невозможно по следующим причинам:

1) нарушается первая предпосылка нормальной линейной регрессионной модели, так как текущие и лаговые значения факторной переменной коррелированы друг с другом;

2) при большой величине лага *L* уменьшается количество наблюдений, по которым строится модель регрессии, и увеличивается число факторных признаков $(x_t, x_{t-1}, x_{t-2} \ldots)$, что в результате ведет к потере числа степеней свободы в модели;

3) в подобных моделях возникает проблема автокорреляции остатков.

Эти причины ведут к нестабильности оценок коэффициентов регрессии, т. е. с изменением спецификации модели ее параметры значительно меняются, теряя точность и эффективность.

На практике параметры моделей с распределенным лагом оценивают с помощью специальных методов, к которым, в частности, можно отнести метод Алмона и метод Койка.

Основная трудность в выявлении структуры временного лага заключается в получении оценок параметров β_i.

Предположения о структуре лага основаны либо на априорной информации о модели, либо на общих положениях экономической теории.

3. Метод Алмона

Метод Алмона или **лаги Алмона** используются для описания моделей с распределенным лагом, имеющим полиномиальную структуру лага и конечную величину лага L:

$$y_t = \beta_0 + \beta_1 \times x_t + \beta_2 \times x_{t-1} + \ldots + \beta_L \times x_{t-L} + \varepsilon_t. \qquad (1)$$

Структура лага определяется при помощи графика зависимости параметров при факторных переменных от величины лага.

Суть метода Алмона состоит в следующем:

1) зависимость коэффициентов при факторных переменных β_i от величины лага i аппроксимируется полиномиальной функцией:

а) первого порядка $\beta_i = c_0 + c_1 \times i$;

б) второго порядка $\beta_i = c_0 + c_1 \times i + c_2 \times i^2$;

в) третьего порядка $\beta_i = c_0 + c_1 \times i + c_2 \times i^2 + c_3 \times i^3$;

г) или в общем случае порядка P:

$$\beta_i = c_0 + c_1 \times i + c_2 \times i^2 + \ldots + c_P \times i^P.$$

Алмон доказал, что во многих случаях легче оценить коэффициенты c_i, $i = \overline{0,P}$, чем непосредственно коэффициенты β_i. Этот метод оценивания коэффициентов β_i называется полиномиальной аппроксимацией;

2) каждый коэффициент модели (1) можно выразить так:

$$\begin{aligned}
&\beta_1 = c_0;\\
&\beta_2 = c_0 + c_1 + \ldots + c_P;\\
&\beta_3 = c_0 + 2c_1 + 4c_2 + \ldots + 2^P c_P;\\
&\beta_4 = c_0 + 3c_1 + 9c_2 + \ldots + 3^P c_P;\\
&\beta_L = c_0 + Lc_1 + L^2 c_2 + \ldots + L^P c_P.
\end{aligned}$$

Подставим полученные соотношения для коэффициентов β_i в модель (1):

$$y_t = \beta_0 + c_0 \times x_t + \left(c_0 + c_1 + \ldots + c_P\right) \times x_{t-1} +$$
$$+ \left(c_0 + 2c_1 + 4c_2 + \ldots + 2^P c_P\right) \times x_{t-2} + \ldots +$$
$$+ \left(c_0 + Lc_1 + L^2 c_2 + \ldots + L^P c_P\right) \times x_{t-L} + \varepsilon_t;$$

3) применим процедуру перегруппировки слагаемых к полученному выражению:

$$y_t = \beta_0 + c_0 \times \left(x_t + x_{t-1} + x_{t-2} + \ldots + x_{t-L}\right) +$$
$$+ c_1 \times \left(x_{t-1} + 2x_{t-2} + 3x_{t-3} + \ldots + Lx_{t-L}\right) +$$
$$+ c_2 \times \left(x_{t-1} + 4x_{t-2} + 9x_{t-3} + \ldots + L^2 x_{t-L}\right) + \ldots +$$
$$+ c_P \times \left(x_{t-1} + 2^P x_{t-2} + 3^P x_{t-3} + \ldots + L^P x_{t-L}\right) + \varepsilon_t.$$

Обозначим слагаемые в скобках при коэффициентах c_i, $i = \overline{0, P}$ как новые переменные:

$$z_0 = x_1 + x_{t-1} + x_{t-2} + \ldots + x_{t-L} = \sum_{i=0}^{L} x_{t-i};$$
$$z_1 = x_{t-1} + 2x_{t-2} + 3x_{t-3} + \ldots + Lx_{t-L} = \sum_{i=0}^{L} i \times x_{t-i};$$
$$z_2 = x_{t-1} + 4x_{t-2} + 9x_{t-3} + \ldots + L^2 x_{t-L} = \sum_{i=0}^{L} i^2 \times x_{t-i};$$
$$z_P = x_{t-1} + 2^P x_{t-2} + 3^P x_{t-3} + \ldots + L^P x_{t-L} = \sum_{i=0}^{L} i^P \times x_{t-i}.$$

С учетом новых переменных модель примет вид:

$$y_t = \beta_0 + c_0 z_0 + c_1 z_1 + \ldots + c_P z_P + \varepsilon_t; \qquad (2)$$

4) коэффициенты новой модели (2) определим с помощью обычного МНК. На основе полученных оценок коэффициентов c_i $\left(i = \overline{0, L}\right)$ найдем оценки параметров исходной модели (1) — β_i $\left(i = \overline{1, L}\right)$, используя соотношения, полученные на первом шаге.

Недостатки метода Алмона:

1) величина максимального временного лага L должна быть известна заранее, что на практике почти не встречается.

Одним из способов определения величины лага L является построение показателей тесноты связи, например линейных парных коэффициентов корреляции, между результативной переменной y и лаговым значением факторного признака x: $r(y, x_{t-1})$, $r(y, x_{t-2})$ и т. д. Если показатель тесноты связи значим, то данную переменную следует включить в модель с распределенным лагом. Порядок максимального значимого показателя тесноты связи принимается в качестве максимальной величины лага L;

2) неизвестен порядок полинома P. При выборе полиномиальной функции обычно исходят из того, что на практике не используются полиномы более второго порядка, а выбранная степень полинома должна быть на единицу меньше числа экстремумов в структуре лага;

3) если между факторными признаками существует сильная связь, то новые переменные z $\left(i = \overline{0, L}\right)$, которые определяются как линейная комбинация исходных факторов x, будут также коррелировать между собой. Проблема мультиколлинеарности в преобразованной модели регрессии (2) устранена не полностью. Но тем не менее мультиколлинеарность новых переменных z_i в меньшей степени отражается на оценках параметров исходной модели (1) β_i $\left(i = \overline{1, L}\right)$, чем в случае применения обычного МНК к данной модели.

Преимущества метода Алмона:

1) в случае небольшого количества переменных в преобразованной регрессионной модели (2) ($P = 2,3$), не приводящего к значительной потере числа степеней свободы, с помощью метода Алмона можно построить модели с распределенным лагом вида (1) любой длины, т. е. максимальный лаг L может быть достаточно большим;

2) метод Алмона является универсальным и может быть применен для моделирования процессов, характеризующихся различными структурами лагов.

ЛЕКЦИЯ № 28. Нелинейный метод наименьших квадратов. Метод Койка. Модель адаптивных ожиданий (МАО) и частичной (неполной) корректировки

При оценивании параметров моделей с распределенным лагом, в которых величина максимального лага L бесконечна, используются **метод нелинейного МНК** и **метод Койка**. Исходят из предположения о геометрической структуре лага, т. е. влияние лаговых значений факторного признака на результативный уменьшается с увеличением величины лага в геометрической прогрессии.

Если имеется одна объясняющая переменная, то модель можно представить:

$$y_t = \beta_0 + \beta_1 \times x_t + \beta_1 \times \lambda \times x_{t-1} + \beta_1 \times \\ \times \lambda^2 \times x_{t-2} + \beta_1 \times \lambda^3 \times x_{t-3} + \ldots + \varepsilon_t, \qquad (1)$$

или

$$y_t = \beta_0 + \beta_1 \times x_t + \beta_2 \times x_{t-1} + \beta_3 \times x_{t-3} + \ldots + \varepsilon_t,$$

где $\beta_i = \beta_1 \times \lambda^i$;

$i = \overline{1,\infty}$;

$\lambda \in [-1; +1]$.

В модели с распределенным лагом (1) неизвестными являются три параметра: β_0, β_1 и λ. Применение обычного МНК для их оценивания невозможно, так как:

1) возникает проблема мультиколлинеарности;

2) полученные МНК-оценки не помогли бы в определении значений параметров β_1 и λ. В данном случае можно получить одно значение оценки β_1 на основе коэффициента при переменной x_t, и совершенно иное при возведении в квадрат коэффициента при переменной x_{t-1} и деления его на коэффициент при переменной x_{i-2}.

Эти проблемы решаются с помощью нелинейного МНК и **метода Койка**.

1. Суть нелинейного МНК

Для параметра λ задаются значения в интервале $[-1;+1]$ с определенным шагом, например 0,05 (чем меньше шаг, тем точнее будет результат).

Для каждого значения λ рассчитывается переменная:

$$z_t = x_t + \lambda \times x_{t-1} + \lambda^2 \times x_{t-2} + \lambda^3 \times x_{t-3} + \ldots + \lambda^L \times x_{t-L},$$

с таким значением лага L, при котором дальнейшие лаговые значения переменной x не оказывают существенного влияния на z.

С помощью обычного МНК оценивается регрессия:

$$y_t = \beta_0 + \beta_1 \times z_t + \varepsilon_t \qquad (2)$$

и определяется коэффициент детерминации R^2. Подобная процедура повторяется для всех значений λ из интервала $[-1;+1]$.

Окончательными оценками β_0, β_1 и λ являются те, которые обеспечивают наибольшее значение R^2 для регрессии (2).

Суть метода Койка (преобразования Койка)

Если регрессия (1) справедлива для момента времени t, то она справедлива и для момента времени $t-1$:

$$y_{t-1} = \beta_0 + \beta_1 \times x_{t-1} + \beta_1 \times \lambda \times x_{t-2} + \beta_1 \times \lambda^2 \times x_{t-3} + \\ + \beta_1 \times \lambda^3 \times x_{t-4} + \ldots + \varepsilon_{t-1},$$

Умножим обе части данного уравнения на λ и вычтем их из уравнения (1):

$$y_t - \lambda \times y_{t-1} = \beta_0 \times (1-\lambda) + \beta_1 \times x_t + \varepsilon_t - \lambda \times \varepsilon_{t-1}$$

или

$$y_t = \beta_0 \times (1-\lambda) + \beta_1 \times x_t + \lambda \times y_{t-1} + v_t,$$

где $v_t = \varepsilon_t - \lambda \times \varepsilon_{t-1}$.

Эта модель является моделью авторегрессии.

Полученная форма модели позволяет анализировать ее краткосрочные и долгосрочные динамические свойства.

В краткосрочном периоде (в текущем периоде) значение y_{t-1} рассматривается как фиксированное, а воздействие x на y характеризует коэффициент β_1.

В долгосрочном периоде (без учета случайной компоненты уравнения), если x_t стремится к некоторому равновесному значению $\bar{x}$, то y_t и y_{t-1} будут стремиться к своему равновесному значению, которое определяется так:

$$\bar{y} = \beta_0(1-\lambda) + \beta_1 \times \bar{x} + \lambda\bar{y},$$

из которой следует:

$$\bar{y} = \beta_0 + \frac{\beta_1}{1-\lambda} \times \bar{x}.$$

Долгосрочное влияние *x* на *y* характеризуется коэффициентом

$$\frac{\beta_1}{1-\lambda}.$$

Если параметр $\lambda \in [0;+1]$, то он превысит значение β_1, т. е. долгосрочное воздействие окажется сильнее краткосрочного.
Модель преобразований Койка весьма удобна на практике, потому что оценки параметров β_0, β_1 и λ можно получить с помощью оценивания обычным МНК модели парной регрессии. Данные МНК-оценки получаются смещенными и несостоятельными, так как нарушается первая предпосылка нормальной линейной регрессионной модели (зависимая переменная *y* частично зависит от ε_{t-1} и поэтому коррелирует с одной из случайных ошибок ($\lambda \times \varepsilon_{t-1}$). Нелинейный метод наименьших квадратов требует больше вычислительных затрат по сравнению с методом Койка.

2. Модель адаптивных ожиданий (МАО)

Модель адаптивных ожиданий (МАО) учитывает предполагаемое (или желаемое) значение факторной переменной x^*_{t+1} в момент времени (t + 1). Данные модели относятся ко второму виду динамических эконометрических моделей. В общем виде модель адаптивных ожиданий можно записать следующим образом:

$$y_t = \beta_0 + \beta_1 \times x^*_{t+1} + \varepsilon_t. \qquad (1)$$

Предполагаемое (ожидаемое) значение переменной x^*_{t+1} в момент времени t + 1 определяется по значению фактических (реальных) переменных в предшествующий момент времени *t*.

В качестве примера модели адаптивных ожиданий можно привести влияние размера предполагаемой в будущем периоде t+1 индексации заработных плат и пенсий на текущие цены или зависимость объема текущих инвестиций в момент времени t от ожидаемого курса валюты в момент времени (t+1).

Механизм формирования ожиданий в модели адаптивных ожиданий выглядит следующим образом:

$$x^*_{t+1} - x^*_t = \lambda \times \left(x_t - x^*_t\right), \quad \text{где} \quad 0 \le \lambda \le 1,$$

или

$$x^*_{t+1} = \lambda \times x_t + (1-\lambda) \times x^*_t. \qquad (2)$$

Таким образом, ожидаемое значение переменной x_t в следующий момент времени (t + 1) является средним арифметическим взвешенным значением ее фактического x_t и ожидаемого x_t^* значений в текущем периоде t. Величина λ называется параметром адаптации (как и в модели экспоненциального сглаживания).

Чем больше его величина, тем быстрее ожидаемое значение адаптируется к предыдущим фактическим событиям x_t. Чем меньше его величина, тем ближе ожидаемое в будущем значение x^*_{t+1} к ожидаемому значению предшествующего периода x_t^*, что характеризует сохранение тенденций в ожиданиях.

Применение традиционного метода наименьших квадратов к оцениванию параметров модели адаптивных ожиданий невозможно, так как модель содержит предполагаемые значения факторной переменной, которые нельзя получить эмпирическим путем. В связи с этим исходную модель адаптивных ожиданий вида (1) преобразуют.

Подставим выражение (2) в исходную модель (1)

$$y_t = \beta_0 + \beta_1 \times \left(\lambda \times x_t + (1-\lambda) \times x^*_t\right) + \varepsilon_t =$$
$$= \beta_0 + \lambda \times \beta_1 \times x_t + (1-\lambda) \times \beta_1 x^*_t + \varepsilon_t. \qquad (3)$$

Если модель адаптивных ожиданий вида (1) верна для момента времени t, то она будет верна и для момента времени (t − 1). Исходя из этого предположения запишем модель адаптивных ожиданий для периода (t − 1):

$$y_{t-1} = \beta_0 + \beta_1 \times x^*_t + \varepsilon_{t-1}.$$

Умножим данное выражение на (1 − λ) и получим:

$$(1-\lambda) \times y_{t-1} = (1-\lambda) \times \beta_0 + (1-\lambda) \times \beta_1 \times x_t^* + (1-\lambda) \times \varepsilon_{t-1}.$$

Далее вычтем почленно полученное выражение из модели (3):

$$y_t - (1-\lambda) \times y_{t-1} = \beta_0 - (1-\lambda) \times \beta_0 + \lambda \times \beta_1 \times x_t + \varepsilon_t - (1-\lambda) \times \varepsilon_{t-1}$$

или

$$y_t = \lambda \times \beta_0 + \lambda \times \beta_1 \times x_t + (1-\lambda) \times y_{t-1} + \varepsilon_t^*, \qquad (4)$$

где $\varepsilon_t^* = \varepsilon_t - (1-\lambda) \times \varepsilon_{t-1}$.

Преобразованная модель (4) является обычной моделью авторегрессии. Определить ее параметры можно с помощью традиционных статистических процедур, так как модель (4) включает только фактические значения факторных переменных. После расчета оценок модели авторегрессии можно элементарно перейти к оценке параметров исходной модели адаптивных ожиданий (1).

Модель адаптивных ожиданий вида (1) характеризует зависимость результативной переменной от предполагаемых значений факторной переменной и называется долгосрочной функцией модели адаптивных ожиданий.

Модель вида (4), полученная в результате преобразований, характеризует зависимость результативной переменной от фактических значений факторной переменной и называется краткосрочной функцией модели адаптивных ожиданий.

3. Модель частичной (неполной) корректировки

Модель частичной (неполной) корректировки (МЧК) учитывает предполагаемое (или желаемое) значение результативной переменной y_t^*. Эти модели относятся ко второму виду динамических эконометрических моделей. В общем виде модель частичной корректировки можно записать следующим образом:

$$y_t^* = \beta_0 + \beta_1 \times x_t + \varepsilon_t. \qquad (1)$$

Предполагаемое значение переменной y_t^* в момент времени t определяется по значению фактических (реальных) переменных в предшествующий момент времени $(t - 1)$.

В качестве примера модели частичной корректировки можно привести зависимость желаемого объема дивидендов y_t^* от факти-

ческого текущего объема прибыли x_t. Данная МЧК более известна как **модель Литнера**.

При построении моделей частичной корректировки исходят из того предположения, что величина фактического приращения результативной переменной в текущем периоде по сравнению с предшествующим периодом $y_t - y_{t-1}$ пропорциональна разности между ее ожидаемым уровнем и фактическим значением в предшествующий момент времени $y_t^* - y_{t-1}$:

$$y_t - y_{t-1} = \lambda \times \left(y_t^* - y_{t-1}\right) + v_t, \text{ где } 0 \le \lambda \le 1,$$

или

$$y_t = \lambda \times y_t^* + (1-\lambda) \times y_{t-1} + v_t. \qquad (2)$$

Фактическое значение результативной переменной в момент времени t (y_t) является средним арифметическим взвешенным значением предполагаемого уровня результативной переменной в тот же самый момент времени (y_t^*) и фактического значения этой переменной в предшествующий момент времени $t-1$ (y_{t-1}).

Величина λ называется параметром корректировки. Чем больше его значение, тем быстрее происходит процесс корректировки результативной переменной y_t. Если параметр корректировки равен единице, то фактическое значение результативной переменной равно ее ожидаемому значению, т. е. $y_t = y_t^*$, и процесс полной корректировки происходит всего за один период. Если параметр корректировки равен нулю, то корректировка результативной переменной y_t не происходит вовсе.

Применение традиционного МНК к оцениванию параметров модели частичной корректировки невозможно, так как модель содержит предполагаемые значения результативной переменной, которые нельзя получить эмпирическим путем. В связи с этим исходную модель вида (1) преобразуют.

С этой целью подставим исходную модель (1) в выражение (2):

$$\begin{aligned} y_t &= \beta_0 \times \lambda + \beta_1 \times \lambda \times x_t + (1-\lambda) \times y_{t-1} + v_t + \lambda \times \varepsilon_t = \\ &= \beta_0 \times \lambda + \beta_1 \times \lambda \times x_t + (1-\lambda) \times y_{t-1} + w_t, \end{aligned} \qquad (3)$$

где $w_t = v_t + \lambda \times \varepsilon_t$.

Неизвестные параметры $\beta_0, \beta_1, \lambda$ преобразованного уравнения регрессии могут быть найдены с помощью обычного метода наименьших квадратов.

Преобразованная модель вида (3) содержит стохастическую объясняющую переменную y_{t-1}.Данная переменная не коррелирует с текущим значением совокупной случайной ошибки уравнения регрессии w_t, так как ошибки ε_t и v_t определяются только после расчета значения результативной переменной y_{t-1}.

С учетом таких предпосылок обычный метод наименьших квадратов позволяет получить асимптотически несмещенные и эффективные оценки неизвестных параметров. Исключение может составлять МНК-оценки в подобных условиях на малых выборках.

Исходная МЧК вида (1), которая содержит предполагаемые значения результативной переменной, называется долгосрочной функцией модели частичной корректировки.

Преобразованная модель вида (3), которая содержит только фактические значения переменных, называется краткосрочной функцией модели частичной корректировки.

Содержание