ЭКОНОМЕТРИКА

Электронный контент

Аннотация

Данный электронный контент предназначен для студентов, изучающих дисциплину «Эконометрика (продвинутый уровень)». Контент впервые был разработан в 2012/2013 учебном году, в соответствии с ФГОС ВПО. В 2015 и 2016 годах был переработан и приведен в соответствие с ФГОС ВО.

Карпенко Н.В.

Тема 1. ОСНОВЫ ЭКОНОМЕТРИКИ

Лекция

План лекции

1. Понятие эконометрики: предмет, цель, задачи
2. Эконометрическая модель
3. Классы эконометрических моделей
4. Типы данных и виды переменных в эконометрических исследованиях
5. Этапы эконометрического моделирования

1. Понятие эконометрики: предмет, цель, задачи

Эконометрическое знание выделилось и сформировалось как закономерный результат развития и взаимодействия экономической теории, математической экономики, экономической статистики, математической статистики и теории вероятностей. Эконометрика формулирует собственные предмет, цель и задачи исследования.

Термин «эконометрика» впервые появился в исследованиях П. Цъемпы (1910), Й. Шумпетера (1923), Р. Фриша (1930) и является соединения двух слов: «экономика» и «метрика». В переводе с греческого *oikonomos* (экономист) – это управляющий домом, *metrihe, metron* (метрика) – мера, размер.

Ученые-эконометристы, признанные авторитеты в области эконометрических знаний, по-разному подходили к определению эконометрики.

Р. Фриш: «Эконометрика есть единство трех составляющих – статистики, экономической теории и математики».

С. Фишер – «Эконометрика занимается разработкой и применением статистических методов для измерения взаимосвязей между экономическими переменными».

С. Айвазян – «Эконометрика объединяет совокупность методов и моделей, позволяющих придавать количественные выражения качественным зависимостям».

В настоящее время не существует единого строгого определения эконометрики. Наиболее часто употребляется следующее понятие:

Эконометрика – это наука, предметом изучения которой является количественное выражение взаимосвязей экономических явлений и процессов

Содержание эконометрики, ее структура и область применения тесно связаны с другими науками.

Взаимосвязь эконометрики с другими науками

Эконометрика	Другие науки
Изучение экономических явлений с точки зрения количественных характеристик	***Экономическая теория*** Изучение качественных аспектов экономических явлений
Опытная проверка экономических законов	***Математическая экономика*** Получение выражений экономических законов в форме математических моделей

Применение инструментария экономической статистики для анализа и прогноза экономических взаимосвязей	***Экономическая статистика*** Сбор, обработка и представление экономических данных в наглядном виде
Применение аппарата математической статистики в силу случайного характера большей части экономических показателей	***Математическая статистика*** Разработка методов анализа данных в зависимости от целей исследования

Анализ подходов к определению эконометрики, состояние эконометрической науки позволяют сформулировать цель эконометрики, которая достигается решением определенных задач.

Цель эконометрики – это разработка способов моделирования и количественного анализа реальных экономических объектов и систем.

Задачи эконометрики можно разделить на три типа

По конечным прикладным целям	**По уровню иерархии**	**По области решения проблем изучаемой экономической системы**
Задачи прогноза экономических показателей, характеризующих состояние и развитие изучаемого явления ***Задачи моделирования*** возможных вариантов экономического развития системы для определения параметров, которые оказывают наиболее сильное влияние на состояние системы в целом	***Задачи макроуровня*** (страна в целом) ***Задачи мезоуровня*** (уровень отраслей, регионов) ***Задачи микроуровня*** (уровень фирмы, предприятия, семьи)	Задачи изучения ***рынка*** Задачи изучения ***инвестиционной***, ***социальной***, ***финансовой политики*** Задачи изучения ***ценообразования*** Задачи изучения ***спроса*** и ***потребления*** Задачи изучения ***отдельно выделенного комплекса проблем***

2. Эконометрическая модель

Основой эконометрического моделирования является эконометрическая модель. Экономический объект в такой модели описывается и изучается с помощью эмпирических (статистических) данных. Эконометрическая модель учитывает реальные условия существования изучаемого объекта и не противоречит общим законам экономики. Ошибка предсказаний по такой модели не превосходит заданной величины.

Общий вид эконометрической модели

$Y = f(x) + \varepsilon$, где

Y – наблюдаемое значение зависимой переменной (объясняемая переменная, результат);

f(x) – объясненная часть зависимой переменной Y, которая зависит от значений объясняющих переменных X (факторов);

ε – случайная составляющая (ошибка, возмущение).

Объясняемая переменная Y – случайная величина с некоторым распределением при заданных значениях объясняющих переменных X_i (i = 1, 2,…, n). Объясняющие переменные в модели - неслучайные (детерминированные) величины.

Задачи эконометрического моделирования
1. Определение объясненной части, на основе экспериментальных данных
2. Получение оценок параметров распределения случайной составляющей, рассматривая ее как случайную величину

Эконометрическая модель – главный инструмент эконометрики, предназначена для анализа и прогноза экономических объектов, явлений и процессов.

3. Классы эконометрических моделей

Эконометрические модели можно условно разделить на три класса.

Классы эконометрических моделей

1. Регрессионные модели с одним уравнением	**2. Системы одновременных уравнений**	**3. Модели временных рядов**
Результативный признак является функцией от факторных признаков $Y = f(X_1, X_2,..., X_m) + \varepsilon$. Объясненная составляющая $f(X_1, X_2,..., X_m)$ - это $M_x(Y)$, т.е. ожидаемое значение результата Y при заданных значениях факторов $X_1, X_2,..., X_m$. Уравнение регрессионной модели имеет вид $Y = M_x(Y) + \varepsilon$. По количеству факторных переменных регрессионные модели делятся на ***парные*** (с одной переменной) и ***множественные***. По виду функции $f(X_1, X_2,..., X_m)$ - регрессионные модели делятся на ***линейные*** и ***нелинейные***	Состоят из тождеств и регрессионных уравнений, в которых наряду с факторными признаками включены результативные признаки из других уравнений системы. Т.о., в системе уравнений одни и те же переменные рассматриваются как зависимые переменные в одни уравнениях и независимые – в других. В тождествах вид и значения параметров известны, в уравнениях – оцениваются.	Результативный признак является функцией переменной времени или переменных, относящихся к другим моментам времени. Модели, описывающие зависимость результативного признака от времени: - ***тренда*** (зависимость результативного признака от трендовой компоненты); - ***сезонности*** (зависимость результативного признака от сезонной компоненты); - ***тренда и сезонности.*** Модели, представляющие зависимость результата от переменных, датированных другими моментами времени: - ***с распределенным лагом*** (зависимость результативного признака от предыдущих значений факторных переменных); - ***авторегрессии*** (зависимость результативного признака от предыдущих значений результативного признака).
Примеры эконометрических моделей		
Модель ***цены*** от объема поставки. Модель ***спроса*** от цены на отдельный товар и от реальных доходов потребителей. Модель зависимости ***объема производства*** от производственных факторов.	Модель ***спроса и предложения***. ***Кейнсианская*** модель формирования доходов. ***Балансовая*** модель Леонтьева.	Модель ***потребления энергоресурсов*** с учетом сезонности Модель ***объема продаж*** Модель ***динамики товарооборота*** и доходов населения

Эконометрические модели отражают свойства изучаемых объектов, явлений и процессов:

- свойство времени двигаться вперед используется в моделях временных рядов (экономические явления происходят в пространстве и во времени);

- свойства динамического равновесия многих экономических явлений применяется в решении систем одновременных уравнений;

- свойство прошлых, настоящих и будущих значений переменных влиять на текущее состояние экономического явления реализуется в моделях авторегрессии и автокорреляции, в моделях адаптивного прогноза;

- свойство временной задержки (лага) между причиной и следствием экономического явления проявляется в модели с распределенным лагом;

- свойство цикличности большого количества экономических явлений находит место в моделях временных рядов с сезонной составляющей.

4. Типы данных и виды переменных в эконометрических исследованиях

В эконометрике применяются два основных типа ***выборочных данных***:

- ***пространственные***;

- ***временные***.

Типы данных

Пространственные	**Временные**
Совокупность экономической информации, характеризующей разные объекты за определенный период (момент) времени.	Совокупность экономической информации, характеризующей определенный объект за различные периоды (моменты) времени. Различают ***стационарные*** и ***нестационарные*** (***динамические***) временные ряды (рис.1.1, 1.2).
Примеры данных	
Объем производства предприятий региона Численность экономически активного населения Показатели экономической деятельности предприятий	Индекс потребительских цен Численность занятых за последние годы Объем произведенных предприятием товаров

Пространственная выборка

Номер объекта наблюдения	**Значения показателей**			
	$\mathbf{X_1}$	$\mathbf{X_2}$	**…**	$\mathbf{X_m}$
1	x_{11}	x_{12}	…	x_{1m}
2	x_{21}	x_{22}	…	x_{2m}
3	x_{31}	x_{32}	…	x_{3m}
…	…	…	…	…
n	x_{n1}	x_{n2}	…	x_{nm}

По числу показателей выделяют ***изолированные*** (***одномерные***) и ***компонентные*** (***многомерные***) временные ряды. Если проводится анализ во времени одного показателя, то временной ряд изолированный. Например, объем произведенных предприятием товаров (помесячно) представляет собой одномерный временной ряд.

Одномерный временной ряд

Момент времени (дата)	Январь 2004 г.	Февраль 2004 г.	Март 2004 г.	Апрель 2004 г.	…	…	…
Номер момента времени *t*	1	2	3	4	…	n - 1	n
Значение показателя y_t	y_1	y_2	y_3	y_4	…	y_{n-1}	y_n

В многомерном ряду представлена динамика нескольких показателей, характеризующих одно экономическое явление (процесс). Например, характеристики деятельности предприятия, наблюдаемые помесячно, можно представить в виде многомерного ряда. В качестве показателей производственной деятельности можно взять: Y - производительность труда, X_1 - удельный вес покупных изделий, X_2 - премии и вознаграждения на одного работника, X_3 - среднегодовая численность ППП, X_4 - среднегодовой фонд заработной платы ППП, X_5 - непроизводственные расходы.

Многомерный временной ряд

Момент времени (дата)	**Номер момента времени *t***	**Значения показателей**				
		Y	**X_1**	**X_2**	**…**	**X_5**
Январь 2004 г.	1	y_1	x_{11}	x_{12}	…	x_{15}
Февраль 2004 г.	2	y_2	x_{21}	x_{22}	…	x_{25}
Март 2004 г.	3	y_3	x_{31}	x_{32}	…	x_{35}
…	…	…	…	…	…	…
…	n	y_n	x_{n1}	x_{n2}	…	$x_{n\,5}$

Объект эконометрического моделирования характеризуется многими признаками. Признаки в модели взаимосвязаны и выступают в роли результата (объясняемой переменной), либо в роли фактора (объясняющей переменной). Переменные в эконометрической модели любого класса условно можно разделить на следующие виды.

Виды переменных

Переменные	**Характеристика**
Экзогенные (независимые, x)	Значения переменных задаются извне модели. В определенной степени являются управляемыми.
Эндогенные (зависимые, y)	Значения переменных определяются внутри модели.
Лаговые (экзогенные или эндогенные)	Значения переменных датируются предыдущими моментами времени и находятся в уравнении с текущими переменными.
Предопределенные (лаговые и текущие экзогенные переменные, лаговые эндогенные переменные)	

Количество независимых переменных (***m***), включенных в модель, не должно быть слишком большим и должно быть теоретически обоснованным.

Качество модели зависит от объема исходных данных (***n*** – объем выборки).

Объем выборки и число независимых переменных в модели связаны соотношением

$$\boldsymbol{n} \geq (6 - 8) \times \boldsymbol{m} .$$

Цель эконометрической модели каждого класса – объяснить значения текущих эндогенных (результативных) переменных в зависимости от значений предопределенных (независимых) переменных.

5. Этапы эконометрического моделирования

Эконометрическое моделирование заключается в решении целого класса задач, которое состоит из следующих этапов.

Этапы эконометрического моделирования

Этап	Содержание
1. Постановочный	Формулировка целей и задач исследования (анализ, прогноз, имитационное моделирование, выработка управленческих решений и т.д.) Определение экономических переменных модели (факторных и результативных)
2. Априорный	Теоретический анализ изучаемого экономического явления Формирование и формализация информации, известной до начала моделирования
3. Параметризация	Определение общего вида эконометрической модели Выражение в математической форме взаимосвязи между модельными переменными Формулировка исходных предпосылок и ограничений модели
4. Информационный	Сбор необходимой статистической информации Анализ качества собранных данных
5. Идентификация модели	Статистический анализ модели Оценка параметров модели
6. Верификация модели	Оценка качества параметров модели и всей модели в целом Проверка достоверности и адекватности модели реальному экономическому явлению
7. Интерпретация результатов моделирования	Формулировка итогов и выводов эконометрического исследования Разработка рекомендаций

Чем шире круг задач, решаемых в пределах одного исследования, тем меньше возможность получить эффективный результат.

Тема 2. КОРРЕЛЯЦИЯ

Лекция

План лекции

1. Понятие корреляционной зависимости
2. Понятие корреляционного анализа
3. Парная корреляция
4. Частная корреляция
5. Множественная корреляция
6. Коэффициент детерминации

1. Понятие корреляционной зависимости

Экономические явления, обладая большим разнообразием, характеризуются множеством признаков, отражающих те или иные их свойства. Эти признаки изменяются (варьируются) во времени и пространстве. Нередко изменения признаков взаимозависимы и взаимообусловлены. В одних случаях связь (зависимость) между признаками оказывается очень тесной (например, часовая выработка и заработная плата), а в других случаях связь между признаками вовсе не обнаруживается или выражается очень слабо (например, пол студентов и их успеваемость). Чем теснее связь между признаками, тем точнее принимаемые решения и легче управление системами.

Различают два типа зависимости между явлениями и их признаками: ***функциональную,*** или жестко детерминированную (например, зависимость выработки продукции на одного рабочего от объема выпущенной продукции и численности рабочих), и ***статистическую,*** или стохастически детерминированную (например, зависимость между производительностью труда и себестоимостью единицы продукции).

Понятие функциональной и статистической зависимости

Функциональная зависимость – это связь, при которой каждому значению независимой переменной x соответствует точно *определенное* (единственное) *значение* зависимой переменной y. Функциональная зависимость чаще всего встречается в естественных науках. Реже подобные связи наблюдаются в общественной жизни, в частности в экономических процессах.

Для социально-экономических явлений характерно то, что наряду с существенными факторами на них оказывают воздействие многие другие, в том числе случайные факторы. В связи с этим существующая зависимость не проявляется здесь в каждом отдельном случае, как при функциональных связях, а лишь «в общем и среднем» при большом числе наблюдений. В этом случае говорят о статистической зависимости.

Статистическая зависимость – это связь, при которой каждому значению независимой переменной x соответствует *множество значений* зависимой переменной *y,* причем неизвестно заранее, какое именно значение примет y.

Частным случаем статистической зависимости является *корреляционная зависимость.*

Корреляционная зависимость – это связь, при которой каждому значению независимой переменной *x* соответствует *определенное математическое ожидание (среднее значение)* зависимой переменной *y.*

Корреляционная связь является «неполной» зависимостью, которая проявляется не в каждом отдельном случае, а только в средних величинах при достаточно большом числе случаев.

Известно, например, что повышение квалификации работника ведет к росту производительности труда. Это положение подтверждается в массе явлений и не означает, что у двух или более рабочих одного разряда, занятых аналогичным процессом, будет одинаковая производительность труда. Уровни их выработки будут различаться, хотя и незначительно, так как у этих рабочих могут быть различными стаж работы, техническое состояние станка, состояние здоровья и т.д.

Из этого следует, что статистическая зависимость - свойство совокупности в целом, а не отдельных ее единиц.

Особенности зависимости

Функциональная	**Корреляционная**
Всегда выражается формулами, что в большей степени присуще точным наукам (математике, физике) С одинаковой силой проявляется у всех единиц совокупности. Является полной и точной, так как обычно известен перечень всех факторов и механизм их воздействия на переменную *в виде уравнения*	Разнообразие факторов, их взаимосвязи и противоречивые действия вызывают широкое варьирование переменной у. Обнаруживается не в единичных случаях, а в массе и требует для своего исследования массовых наблюдений Связь между переменными х и у *неполная и проявляется лишь в средних величинах*

Виды функциональной и корреляционной зависимости

Функциональная и корреляционная связь в зависимости от направления действия бывает ***прямая*** и ***обратная.***

Функциональная и корреляционная зависимость	
Прямая	**Обратная**
С *увеличением* (уменьшением) значений факторного признака происходит *увеличение* (уменьшение) результативного признака	С *увеличением* (уменьшением) значений факторного признака происходит *уменьшение* (увеличение) результативного признака

По аналитическому выражению зависимость может быть *прямолинейной (**линейной**) и криволинейной (**нелинейной**).*

Функциональная и корреляционная зависимость	
Прямолинейная	**Криволинейная**
С возрастанием величины факторного признака происходит *равномерное* возрастание (или убывание) величин результативного признака (выражаются уравнением прямой линии)	С возрастанием величины факторного признака возрастание (или убывание) результативного признака происходит *неравномерно* (выражаются уравнениями кривых линий)

В зависимости от количества признаков, включенных в модель, корреляционные связи делят на ***однофакторные*** и ***многофакторные.***

Корреляционные связи	
Однофакторные (*парные*)	**Многофакторные *(множественные)***
Связь между одним признаком-фактором и результативным признаком (при абстрагировании влияния других)	Связь между несколькими факторными признаками и результативным признаком (факторы действуют комплексно, т.е. одновременно и во взаимосвязи)

Корреляционная зависимость исследуется с помощью методов ***корреляционного*** и ***регрессионного*** анализа.

При построении корреляционных моделей исходят из условия нормальности многомерного закона распределения генеральной совокупности. Эти условия обеспечивают линейный характер связи между изучаемыми признаками, что делает правомерным использование в качестве показателей тесноты связи парного, частного коэффициентов корреляции и коэффициента множественной корреляции.

2. Понятие корреляционного анализа

Корреляционный анализ – это раздел математической статистики, посвященный изучению взаимосвязей между случайными величинами. Применяется тогда, когда данные наблюдений можно считать случайными и выбранными из генеральной совокупности, распределенной по многомерному нормальному закону.

Корреляционный анализ заключается в количественном определении тесноты связи между двумя признаками (при парной связи) и между результативным и множеством факторных признаков (при многофакторной связи).

Понятие корреляции

Корреляция – это статистическая зависимость между случайными величинами, при которой изменение одной из случайных величин приводит к изменению математического ожидания другой.

Варианты корреляции

Корреляция		
Парная	**Частная**	**Множественная**
Связь между двумя признаками (результативным и факторным или двумя факторными)	Зависимость между результативным и одним факторным признаками или двумя факторными признаками при фиксированном значении других факторных признаков	Зависимость между результативным признаком и двумя и более факторными признаками, включенными в исследование

3. Парная корреляция

Теснота связи количественно выражается величиной коэффициентов корреляции. Построение коэффициентов корреляции основано на сумме произведений отклонений индивидуальных значений признаков x_i и y_i от их средних значений $\bar{x}$ и $\bar{y}$: $\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$. Эта величина, деленная на число единиц совокупности (объем выборки) n, называется ковариацией. Она характеризует сопряженность вариации двух признаков и представляет собой статистическую меру взаимодействия двух случайных переменных.

Формула определения ковариации

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

где n - объем исследуемой совокупности (объем выборки);

x_i - i-е значение независимой переменной (i = 1, 2,…, n);

yi - i-е значение зависимой переменной (i = 1, 2,…, n);

$\bar{x}$ - среднее значение независимой переменной. Определяется по формуле

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

$\bar{y}$ - среднее значение зависимой переменной. Определяется по формуле

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i .$$

При наличии прямой связи большие значения *x* должны сочетаться с большими значениями *y,* следовательно, отклонения ($x_i - \bar{x}$) и ($y_i - \bar{y}$) будут положительными.

Для малых значений x и y эти отклонения будут отрицательными, а их произведения - положительными. Значит, при прямой связи ковариация будет величиной положительной.

При наличии обратной связи отклонения ($x_i - \bar{x}$) и ($y_i - \bar{y}$) будут иметь разные знаки (большие значения x сочетаются с меньшими значениями y и наоборот). Ковариация будет отрицательной величиной.

Наконец, при отсутствии связи сочетание знаков отклонений ($x_i - \bar{x}$) и ($y_i - \bar{y}$) будет беспорядочным, при суммировании отрицательные и положительные произведения ($x_i - \bar{x}$) и ($y_i - \bar{y}$) будут взаимно погашаться и ковариация будет близка к нулю.

Размер ковариации зависит от масштаба признаков x и y. Для получения относительной характеристики связи ковариацию делят на максимально возможное значение, равное произведению средних квадратических отклонений двух признаков s_x, s_y. В результате получают ***парный коэффициент корреляции r_{xy}.***

Формула парного коэффициента корреляции

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

где n - объем выборки;

$\bar{x}$, $\bar{y}$ - средние значения;

s_x^2, s_y^2 - исправленные дисперсии,

s_x, s_y - среднеквадратические (стандартные) отклонения признаков y, x ,

определяются по формулам:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i , \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i ;$$

$$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 , \quad s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2 ,$$

$$s_x = \sqrt{s_x^2} , \quad s_y = \sqrt{s_y^2} .$$

Коэффициент корреляции r_{yx} служит мерой **линейной корреляционной зависимости между величинами y и x** , при условии, что на формирование их значений оказывают влияние некоторые другие, неучтенные факторы.

Для расчета парного коэффициента корреляции можно воспользоваться также следующими формулами:

1) $$r_{xy} = \frac{\overline{y \cdot x} - \bar{y} \cdot \bar{x}}{s_x \cdot s_y},$$

где $\overline{y \cdot x}$ - средняя арифметическая произведения двух величин

$$\overline{y \cdot x} = \frac{1}{n} \sum_{i=1}^{n} (y_i \cdot x_i);$$

2) $$r_{xy} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (y_i - \bar{y})^2}}.$$

Коэффициент корреляции принимает значение от -1 до +1. Положительное значение коэффициента свидетельствует о наличии прямой связи, отрицательное - обратной. Если $r_{xy} = \pm 1$, корреляционная связь представляется линейной функциональной зависимостью. При $r_{xy} = 0$ **линейная** корреляционная связь отсутствует.

Качественные характеристики связи

Значение коэффициента корреляции $\lvert r_{yx} \rvert$	**Характер линейной корреляционной связи между признаками y и x**
0 - 0,2	Практически отсутствует
0,2 - 0,5	Слабая
0,5 - 0,7	Средняя (умеренная)
0,7 - 0,95	Сильная (тесная)
0,95 - 1	Практически функциональная

Основные свойства парного коэффициента корреляции

1. Если величины x и y независимы, то $r_{xy} = 0$;

2. $r_{xy} = r_{yx}$;

3. $-1 \le r_{xy} \le 1$;

4. При $r_{xy} > 0$ величины x и y одновременно возрастают (прямая зависимость). При $r_{xy} < 0$ с возрастанием величины x (y) величина y (x) убывает (обратная зависимость);

5. Из равенства $r_{xy} = 0$ не следует независимость величин x и y , отрицается лишь их линейная корреляционная зависимость (может существовать нелинейная зависимость);

6. Значения $r_{xy} = -1$, $r_{xy} = 1$ соответствуют практически функциональной линейной связи между величинами x и y .

Коэффициенты корреляции как статистические величины необходимо оценить на достоверность. Это объясняется тем, что любая совокупность наблюдений представляет собой некоторую выборку, следовательно, значение любого показателя, вычисленное на основе выборки, не может рассматриваться как истинное, а является только более или менее точной его оценкой. В связи с этим возникает необходимость проверки существенности (***статистической значимости***) показателей.

Оценка статистической значимости коэффициента корреляции

Для оценки значимости коэффициента корреляции используют ***t-критерий Стьюдента,*** который применяется при *t*-распределении, отличном от нормального, но приближающемся к нормальному при $n > 30$.

При этом выдвигается нулевая гипотеза H_0: $r_{xy} = 0$, т.е. коэффициент корреляции предполагается статистически незначимым, случайно отклоняющимся от 0, при конкурирующей гипотезе H_1: $r_{xy} \neq 0$ (коэффициент корреляции неслучайно отличается от нуля). Находится расчетное значение (статистика) критерия

$$t_{расч} = \frac{r_{xy}}{\sqrt{1 - r_{xy}^2}} \cdot \sqrt{n - m - 1} \ ,$$

где m – число факторных (объясняющих) признаков, включенных в модель

которое сравнивается с табличным (критическим) значением $t_{табл}$, найденным по таблице односторонних критических значений распределения Стьюдента по уровню значимости α (вероятность отвергнуть гипотезу H_0, при условии что она верна) и числу степеней свободы $df = n - m - 1$. В экономических расчетах значение α обычно принимается равным 0,05 или 0,01.

Если $|t_{расч}| > t_{табл}$, то гипотеза H_0 отвергается с вероятность ошибки α , т.е. **коэффициент корреляции** признается **статистически значимым**. Если $|t_{расч}| < t_{табл}$, то гипотеза H_0 принимается, коэффициент корреляции считается незначимым, случайно отличающимся от нуля.

Для значимого коэффициента парной корреляции находится его интервальная оценка, в качестве которой служит ***доверительный интервал***, определенный с надежностью (вероятностью) $\gamma = 1 - \alpha$:

$$r_{xy} - m_r \cdot t_{табл} < r < r_{xy} + m_r \cdot t_{табл}$$

или

$$\left(r_{xy} - m_r \cdot t_{табл} \ ; \ r_{xy} + m_r \cdot t_{табл} \right).$$

Здесь $t_{табл}$ - значение, найденное по таблице двусторонних критических точек распределения Стьюдента по уровню значимости $\alpha/2$ и числу степеней свободы $df = n - m - 1$; m_r - стандартная ошибка коэффициента корреляции

$$m_r = \sqrt{\frac{1 - r_{xy}^2}{n - m - 1}} \ .$$

С вероятностью γ доверительный интервал покрывает истинное значение коэффициента корреляции r. Если границы доверительного интервала имеют разные знаки, r_{xy} статистически незначим.

4. Частная корреляция

Частные коэффициенты корреляции рассчитываются в случае, когда необходимо исследовать взаимозависимость трех и более признаков.

Частный коэффициент корреляции характеризует тесноту связи между двумя признаками из совокупности признаков при условии, что все связи этих признаков с другими признаками элиминированы, т.е. закреплены на условно-постоянном (среднем) уровне.

Пусть имеются три признака x, y, z, для каждого из которых зафиксированы n значений. Для них вычислены парные коэффициенты корреляции, которые записаны в виде таблицы. Такая таблица называется ***корреляционной матрицей***.

Корреляционная матрица

	x	y	z
x	1	r_{xy}	r_{xz}
y	r_{yx}	1	r_{yz}
z	r_{zx}	r_{zx}	1

Корреляционная матрица обладает свойством симметричности относительно главной диагонали, т.е. $r_{xy} = r_{yx}$. Элементы, стоящие на главной диагонали матрицы равны единице, поскольку $r_{xx} = r_{yy} = r_{zz} = 1$.

Частный коэффициент корреляции между x и y, при фиксированном z (при исключении влияния z) находится по формуле

$$r_{xy/z} = \frac{r_{xy} - r_{xz} \cdot r_{yz}}{\sqrt{\left(1 - r_{xz}^2\right)\left(1 - r_{yz}^2\right)}},$$

и обладает всеми свойствами парного коэффициента корреляции. Знак частного коэффициента корреляции $r_{xy/z}$ совпадает со знаком соответствующего парного коэффициента корреляции r_{xy} . Он характеризует взаимосвязь факторных признаков при устранении влияния результативного признака:

Частные коэффициенты корреляции между y и z, при фиксированном x; между z и x, при фиксированном y определяются по формулам

$$r_{yz/x} = \frac{r_{yz} - r_{yx} \cdot r_{zx}}{\sqrt{\left(1 - r_{yx}^2\right)\left(1 - r_{zx}^2\right)}},$$

$$r_{zx/y} = \frac{r_{zx} - r_{zy} \cdot r_{xy}}{\sqrt{\left(1 - r_{zy}^2\right)\left(1 - r_{xy}^2\right)}}.$$

Формулы для вычисления частных коэффициентов корреляции для более чем трех переменных на основе корреляционной матрицы сложны и трудоемки. Для их расчета на практике используют пакеты прикладных программ (ППП).

Замечание

Если парный коэффициент корреляции между двумя случайными величинами оказался больше частного коэффициента между теми же случайными величинами, то третья фиксированная величина усиливает взаимосвязь между изучаемыми величи-

нами, т.е. более высокое значение парного коэффициента обусловлено присутствием третьей величины.

Более низкое значение парного коэффициента корреляции в сравнении с соответствующими частными свидетельствует об ослаблении связи между изучаемыми величинами действием фиксируемой величины.

Частный коэффициент корреляции изменяется в пределах от -1 до +1. Если частный коэффициент корреляции равен ±1, то линейная связь между двумя величинами функциональная, а равенство нулю свидетельствует о линейной независимости этих величин.

Частные коэффициенты корреляции как статистические величины необходимо проверять на достоверность.

Проверка статистической значимости частных коэффициентов корреляции проводится аналогично проверке значимости парного коэффициента корреляции с помощью t-критерия Стьюдента. Расчетное значение критерия, найденное по формуле

$$t_{расч} = \frac{r_{xy/z}}{\sqrt{1 - r^2_{xy/z}}} \cdot \sqrt{n - m - 1},$$

сравнивается с табличным значением $t_{табл}$, найденным по таблице односторонних критических значений распределения Стъюдента по уровню значимости α и числу степеней свободы df = n - m - 1.

Если $|t_{расч}| > t_{табл}$, то коэффициент частной корреляции $r_{xy/z}$ признается статистически значимым.

Частные коэффициенты детерминации

Частные коэффициенты детерминации находятся как квадраты частных коэффициентов корреляции:

$$R^2_{xy/z} = \left(r_{xy/z}\right)^2,$$
$$R^2_{yz/x} = \left(r_{yz/x}\right)^2,$$
$$R^2_{xz/y} = \left(r_{xz/y}\right)^2.$$

Частный коэффициент детерминации показывает долю вариации признака под действием одного из признаков при неизменном значении третьего признака.

5. Множественная корреляция

Коэффициент множественной корреляции рассчитывается в случае, когда исследуется взаимосвязь трех и более признаков. Для трех признаков x, y, z коэффициент множественной корреляции между величиной z и парой величин x, y вычисляется на основе корреляционной матрицы.

Формула определения коэффициента множественной корреляции

$$R_z = r_{z/xy} = \sqrt{\frac{r^2_{zx} + r^2_{zy} - 2 r_{xy} \cdot r_{zx} \cdot r_{zy}}{1 - r^2_{xy}}}.$$

Коэффициенты множественной корреляции R_x, R_y определяются аналогично:

$$R_x = r_{x/zy} = \sqrt{\frac{r_{xz}^2 + r_{xy}^2 - 2 r_{xy} \cdot r_{zx} \cdot r_{zy}}{1 - r_{zy}^2}} \; ;$$

$$R_y = r_{y/xz} = \sqrt{\frac{r_{yx}^2 + r_{yz}^2 - 2 r_{xy} \cdot r_{zx} \cdot r_{zy}}{1 - r_{xz}^2}} \; .$$

Коэффициент множественной корреляции является мерой **линейной** связи между одним признаком и двумя остальными.

Коэффициент множественной корреляции принимает значения от 0 до 1. При $R_z = 1$ связь между величинами z и парой x, y является функциональной, линейной. При $R_z = 0$ величина z не зависит от пары x, y . Теснота связи между признаками устанавливается по величине коэффициента множественной корреляции.

Качественные характеристики связи

Значение коэффициента множественной корреляции R_z	**Характер линейной корреляционной связи между признаком z и парой x, y**
0 - 0,1	Слабая
0,1 - 0,5	Средняя (умеренная)
0,5 – 1	Сильная (тесная)

Для коэффициента множественной корреляции справедливы следующие неравенства

$$R_z \geq |r_{zx}| \; , \; R_z \geq |r_{zy}| \; ,$$

$$R_z \geq |r_{zx/y}| \; , \; R_z \geq |r_{zy/x}| \; .$$

Из них следует, что коэффициент множественной корреляции может только увеличиться, если в модель включать дополнительные признаки, и не увеличиться, если из имеющегося набора признаков производить исключение.

Коэффициент множественной корреляции R для произвольного числа признаков $x_1, x_2, \ldots, x_m$ ($m > 3$) находятся по рекурентным формулам, что требует большого опыта вычислений. Для них справедливы все изложенные выше свойства и выводы. Для расчетов обычно применяют ППП.

Проверка статистической значимости коэффициента множественной корреляции

Проверка статистической значимости коэффициента множественной корреляции производится с помощью ***F-критерия Фишера***. Выдвигается гипотеза H_0: $R_z = 0$, т.е. коэффициент множественной корреляции случайно отличается от 0 (статистически незначим), при конкурирующей гипотезе H_1: $R_z \neq 0$, т.е. коэффициент множественной корреляции неслучайно отличается от 0 (статистически значим). Находится расчетное значение (статистика) критерия

$$F_{расч} = \frac{R_z^2}{1 - R_z^2} \cdot \frac{n - m - 1}{m} \; ,$$

где n - объем выборки, m - число независимых (факторных) переменных.

Значение $F_{расч}$ сравнивается с табличным (критическим) значением $F_{табл}$, найденным по таблице критических значений распределения Фишера-Снедекора (F-распределения) по уровню значимости $\boldsymbol{\alpha}$ и двум числам степеней свободы $df_1 = m$ и $df_2 = n - m - 1$.

Если $F_{расч} > F_{табл}$, то гипотеза H_0 отвергается с вероятность ошибки α , т.е. коэффициент множественной корреляции признается статистически значимым. В противном случае ($F_{расч} < F_{табл}$) R статистически незначим.

Интервальная оценка (доверительный интервал) коэффициента множественной корреляции R находится с помощью ***z-преобразования Фишера***

$$z(R) = \frac{1}{2}\ln\left(\frac{1+R}{1-R}\right) = \frac{1}{2}\left(\ln(1+R) - \ln(1-R)\right).$$

Функция ***z*** нечетная, т.е.

$$z(-R) = -z(R).$$

Предварительно устанавливается интервальная оценка для ***z*** в виде

$$z' - t_\gamma\sqrt{\frac{1}{n-m-3}} \le z \le z' + t_\gamma\sqrt{\frac{1}{n-m-3}},$$

где m - порядок коэффициента корреляции (число факторных признаков, т.е. независимых переменных); n - объем выборки, γ - вероятность выполнения неравенства; t_γ вычисляется по таблице значений интегральной функции Лапласа из условия

$$\Phi(t_\gamma) = \gamma .$$

Значение z' определяется по таблице z-преобразования по найденному значению R . Обратный переход от z к R осуществляется также по таблице z-преобразования, после использования которой получают интервальную оценку R с надежностью $\gamma = 1 - \alpha$:

$$R_{\min} \le R \le R_{\max} .$$

6. Коэффициент детерминации

На основе коэффициента корреляции рассчитывается ***коэффициент детерминации R^2*** .

В случае парной корреляции

$$\boxed{R^2 = (r_{yx})^2}.$$

Коэффициент детерминации оценивает долю вариации признака y , обусловленную изменением значений признака x. Чем ближе значение R^2 к единице, тем больше признак x участвует в формировании значений y.

В случае множественной корреляции

$$R^2 = (R)^2.$$

Коэффициент детерминации оценивает долю вариации зависимого признака y , обусловленную изменением значений признаков $x_1, x_2, \ldots, x_m$. Чем ближе значение R^2 к единице, тем большее признаки $x_1, x_2, \ldots, x_m$ участвуют в формировании значений y.

Для большого числа переменных коэффициент детерминации рассчитывается с помощью корреляционной матрицы

Корреляционная матрица

	y	x_1	x_2	…	x_m
y	1	r_{yx1}	r_{yx2}	…	r_{yxm}
x_1	r_{x1y}	1	r_{x1x2}	…	r_{x1xm}
x_2	r_{x2y}	r_{x2x1}	1	…	r_{x2xm}
…	…	…	…	…	…
x_m	r_{xmy}	r_{xmx1}	r_{xmx2}	…	1

По формуле

$$R^2 = 1 - \frac{\Delta}{\Delta_y},$$

где Δ - определитель корреляционной матрицы, Δ_y - определитель матрицы межфакторных корреляций (корреляций между независимыми переменными) (матрица выделена заливкой).

На практике при большом числе переменных сначала вычисляют коэффициент детерминации, а затем - коэффициент множественной корреляции по формуле

$$R = \sqrt{R^2}\ .$$

Проверка статистической значимости коэффициента детерминации (как в случае парной, так и множественной корреляции) производится с помощью F-критерия Фишера аналогично проверке статистической значимости коэффициента множественной корреляции.

Выдвигается гипотеза H_0: $R^2 = 0$, т.е. множественный коэффициент детерминации случайно отличается от 0 (статистически незначим), при конкурирующей гипотезе H_1: $R^2 \neq 0$, т.е. множественный коэффициент детерминации неслучайно отличается от 0 (статистически значим). Находится расчетное значение (статистика) критерия

$$F_{расч} = \frac{R^2}{1 - R^2} \cdot \frac{n - m - 1}{m},$$

где n - объем выборки, m - число независимых переменных.

Значение $F_{расч}$ сравнивается с табличным (критическим) значением $F_{табл}$, найденным по таблице критических значений распределения Фишера-Снедекора (F-распределения) по уровню значимости α и двум числам степеней свободы $df_1 = m$ и $df_2 = n - m - 1$.

Если $F_{расч} > F_{табл}$, то гипотеза H_0 отвергается с вероятность ошибки α , т.е. коэффициент детерминации признается статистически значимым. В противном случае ($F_{расч} < F_{табл}$) R^2 статистически незначим.

Статистическая значимость коэффициента множественной корреляции совпадает со статистической значимостью коэффициента детерминации.

Замечание

Наиболее достоверные результаты в корреляционном анализе можно получить, когда число объектов наблюдения, т.е. объем выборки (n), превышает число анализируемых признаков в 6÷8 раз.

Тема 3. ПАРНАЯ ЛИНЕЙНАЯ РЕГРЕССИЯ

Лекция

План лекции

1. Понятие регрессионного анализа
2. Уравнение парной линейной регрессии. Оценка параметров
3. Показатели тесноты связи
4. Оценка качества уравнения парной линейной регрессии
5. Анализ остатков
6. Практические приложения регрессионных моделей

1. Понятие регрессионного анализа

Среди многих форм связей между признаками важнейшую роль играет причинная, определяющая все другие формы. Сущность причинности состоит в порождении одного явления другим. В любой конкретной связи одни признаки выступают в качестве факторов, воздействующих на другие и обусловливающие их изменение, другие - в качестве результатов действия этих факторов. Иными словами, одни представляют собой причину, другие - следствие. Признаки, характеризующие следствие, называются ***результативными*** (***зависимыми***, ***объясняемыми переменными***, y), признаки, характеризующие причины - ***факторными*** (***независимыми***, ***объясняющими переменными***, x).

После выявления с помощью корреляционного анализа наличия статистически значимых связей между признаками и оценки степени тесноты их связи обычно переходят к математическому описанию конкретного вида зависимостей с использованием регрессионного анализа.

Регрессионным анализом называется метод статистического анализа зависимости ***случайной величины*** y от переменных x_j (j = 1, 2,…, m), рассматриваемых как ***неслучайные величины***, независимо от истинного закона распределения y.

Функция $f(x_1, x_2,\ldots, x_m)$, описывающая зависимость условного среднего значения $\tilde{y}$ ***результативного*** признака y от заданных значений ***объясняющих*** переменных $x_1, x_2,\ldots, x_m$, называется ***функцией регрессии***. Для точного описания уравнения регрессии необходимо знать условный закон распределения y.

Доказано, что если многомерная случайная величина $(y, x_1, x_2,\ldots, x_m)$ подчиняется (m + 1)-мерному нормальному закону распределения, то уравнение регрессии переменной y имеет линейный вид по объясняющим переменным $x_1, x_2,\ldots, x_m$.

На практике получить информацию о законе распределения обычно не удается, поэтому ограничиваются поиском подходящих аппроксимаций (аналитических выражений) для функции $f(x_1, x_2,\ldots, x_m)$, основанных на исходных статистических данных.

В общем случае ***модель регрессионного анализа*** (эконометрическая модель) имеет вид

$$y = f(x_1, x_2,\ldots, x_m) + \varepsilon$$

$y = \hat{y} + \varepsilon$, где $\hat{y} = f(x_1, x_2,\ldots, x_m)$ - ***уравнение регрессии***, **ε** - ***остаток*** (***отклонение***).

Теоретическая обоснованность моделей

Теоретическая обоснованность регрессионных моделей обеспечивается соблюдением определенных условий.

Условия теоретической обоснованности моделей

1	2	3
Все признаки и их совместные распределения должны починяться нормальному закону распределения	Дисперсия моделируемого признака должна все время оставаться постоянной при изменении величины и значений факторных признаков	Отдельные наблюдения должны быть независимыми, т.е. результаты, полученные в i-м наблюдении, не должны быть связаны с предыдущими и содержать информацию о последующих наблюдениях, а также влиять на них

Сущность регрессионного анализа

Регрессионный анализ заключается в определении аналитической формы связи (уравнения регрессии), в которой изменение результативного признака обусловлено влиянием одного или нескольких факторных (объясняющих) признаков, а множество всех прочих факторов, также оказывающих влияние на результативный признак, принимается за постоянные и средние значения и учитывается в остатках (ε) .

Основная цель регрессионного анализа - построение модели зависимости результативного признака от факторных признаков, а также определение степени воздействия факторных признаков на результативный.

Основной предпосылкой регрессионного анализа является то, что только результативный признак подчиняется нормальному закону распределения, а факторные признаки - произвольному закону распределения. При этом в регрессионном анализе заранее подразумевается наличие причинно-следственных связей между результативным и факторными признаками.

Теоретические положения и расчеты регрессионного анализа неразрывно связаны с корреляционным анализом. Поэтому часто его называют ***корреляционно-регрессионным***.

Корреляционно-регрессионный анализ проводится поэтапно в определенной логической последовательности.

Этапы проведения комплексного корреляционно-регрессионного анализа

1. Предварительный анализ явлений и выявление причин возникновения взаимосвязей между признаками, характеризующими эти явления.
2. Разделение признаков на факторные и результативные, выбор наиболее существенных признаков для их исследования на предмет включения в корреляционно-регрессионные модели.
3. Построение матрицы коэффициентов парной корреляции и оценка возможных вариантов группировки признаков корреляционно-регрессионных моделей.
4. Предварительная оценка формы уравнения регрессии.
5. Построение уравнения регрессии, вычисление коэффициентов регрессии и их смысловая интерпретация.

6. Расчет теоретически ожидаемых (воспроизведенных по уравнению регрессии) значений результативного признака.
7. Определение и сравнительный анализ дисперсий: общей, факторной и остаточной; оценка тесноты связи между признаками, включенными в регрессионную модель.
8. Общая оценка качества модели, отсев несущественных (или включение дополнительных) факторов, построение модели, т.е. повторение п. 1-7.
9. Статистическая оценка достоверности параметров уравнения регрессии, построение доверительных границ для теоретически ожидаемых по уравнению регрессии значений функции.
10. Практические выводы из анализа.

Этапы 2, 3, 4 называются ***спецификацией*** модели.

2. Уравнение парной линейной регрессии. Оценка параметров

Форма связи $f(x_1, x_2, ..., x_m)$ может быть выражена как линейной функцией (уравнение прямой), так и нелинейными функциями (полиномы разных порядков, гипербола, степенная функция и др.).

Уравнение регрессии выражается функцией

Парная регрессия	Множественная регрессия
$\hat{y} = f(x)$	$\hat{y} = f(x_1, x_2, ..., x_m)$, где m - число факторных признаков
Характеризует связь между двумя признаками: результативным и факторным	Характеризует связь между результативным признаком и двумя и более факторными признаками

Подбор функции $f(x_1, x_2, ..., x_m)$ для выражения формы связи между признаками проходит несколько этапов: графический, логический, экономический, а также математическую проверку близости эмпирических данных к теоретическим.

Часто для выражения формы регрессионной связи подходит одновременно несколько функций, поэтому желательно дать окончательное обоснование выбора функции для выражения формы связи на альтернативной основе.

Наиболее разработанной в эконометрике и наиболее простой с точки зрения понимания, интерпретации и техники расчетов является ***линейная форма*** регрессии, рассматривающая влияние вариации переменной x (m = 1) на переменную y и представляющая собой ***однофакторный регрессионный анализ.***

Уравнение парной линейной регрессии

$$\hat{y} = b_0 + b_1 \cdot x ,$$

где b_0, b_1 – параметры уравнения

Содержание параметров уравнения парной линейной регрессии

Параметр	Содержание параметра
b_0	***Свободный член*** регрессионного уравнения. Не имеет экономического смысла и показывает значение результативного признака y, если факторный признак x = 0
b_1	***Коэффициент регрессии*** показывает, на какую величину в среднем изменится результативный признак y, если переменную x увеличить на единицу измерения. Знак при коэффициенте регрессии показывает направление связи: при $b_1 > 0$ - связь прямая; при $b_1 < 0$ - связь обратная

Нахождение значений параметров b_0, b_1 производится на основе совокупности наблюдений (выборки, ***матрицы наблюдений***). Для разных выборок (даже одного объема) из генеральной совокупности будут найдены разные значения b_0, b_1 . Поэтому их рассматривают как приближенные значения (***оценки***) истинных параметров регрессионного уравнения. Сама процедура нахождения приближенных значений также называется ***оценкой параметров***.

Для получения оценок параметров линейного регрессионного уравнения можно использовать метод наибольшего правдоподобия, метод наименьших модулей, метод минимакса и др., однако, согласно ***теореме Гаусса-Маркова***, наилучшие результаты дает ***метод наименьших квадратов*** (**МНК**).

Требования теоремы Гаусса-Маркова

A - «***истинная» зависимость y*** от ***x имеет вид***

$$y = \beta_0 + \beta_1 \cdot x + \varepsilon ;$$

B - ***x - неслучайна переменная*** (детерминированная);

C - ***столбцы матрицы наблюдений, с добавлением единичного столбца, линейно независимы*** (ранг матрицы равен 2);

D - ***остатки ε имеют нулевое математическое ожидание***

$$M(\varepsilon_i) = 0$$

и постоянную дисперсию σ_ε^2, не зависящую от номера наблюдения (свойство ***гомоскедастичности***), т.е.

$$D(\varepsilon_i) = \sigma_\varepsilon^2 = const ;$$

E - для разных наблюдений ***остатки ε некоррелированы (независимы)***, т.е. выполняется условие

$$\mathrm{cov}(\varepsilon_i, \varepsilon_l) = 0 \text{ , при } i \neq l, \quad i, l = 1, 2, \ldots, n .$$

Часто вместо **D** добавляют условие

F - ***остатки ε подчиняются нормальному закону распределения***.

Теорема Гаусса-Маркова

В предположениях **A - E** оценки, полученные методом наименьших квадратов, являются несмещенными и обладают наименьшей дисперсией среди всех линейных несмещенных оценок параметров β_0 , β_1 .

Сущность метода наименьших квадратов

Отыскиваются такие значения параметров, при которых сумма квадратов отклонений фактических значений результативного признака y_i от вычисленных по уравнению регрессии $\hat{y}_i$, будет наименьшей из всех возможных:

$$\sum_{i=1}^{n}(y_i-\hat{y}_i)^2=\sum_{i=1}^{n}(y_i-b_0-b_1x_i)^2\to\min\ .$$

Путем преобразований требование минимума суммы квадратов отклонений сводится к системе нормальных уравнений.

Система нормальных уравнений для нахождения параметров парной линейной регрессии

$$\begin{cases} b_0 n + b_1\sum_{i=1}^{n}x_i=\sum_{i=1}^{n}y_i \\ b_0\sum_{i=1}^{n}x_i+b_1\sum_{i=1}^{n}x_i^2=\sum_{i=1}^{n}x_iy_i \end{cases}$$

Формулы для определения значения параметров парной линейной регрессии

$$b_1=\frac{\sum_{i=1}^{n}x_iy_i-\frac{1}{n}\sum_{i=1}^{n}x_i\sum_{i=1}^{n}y_i}{\sum_{i=1}^{n}x_i^2-\frac{1}{n}\left(\sum_{i=1}^{n}x_i\right)^2}\ , \qquad b_0=\bar{y}-b_1\bar{x}$$

После нахождения параметров можно записать парную регрессионную модель.

Модель парной линейной регрессии

$$y=b_0+b_1\cdot x+\varepsilon \ \text{ или}$$

$$y_i=b_0+b_1\cdot x_i+\varepsilon_i \ (\text{i} = 1, 2,\ldots, \text{n})$$

После построения регрессионной модели необходимо провести оценку ее качества, которая заключается в проверке уравнения регрессии и в анализе остатков.

Построенное уравнение парной линейной регрессии графически изображается прямой линией. На рисунке наблюдения изображены точками. Совокупность точек называется ***корреляционным полем*** (диаграммой рассеивания) (см. рис. 1). Точки корреляционного поля не лежат на линии регрессии, т.е. исходные данные при подстановке в уравнение регрессии не обращают его в тождество

$$y_i\neq b_0+b_1\cdot x_i\cdot$$

Уравнение регрессии справедливо только для средних величин

$$\bar{y} = b_0 + b_1 \cdot \bar{x} \ .$$

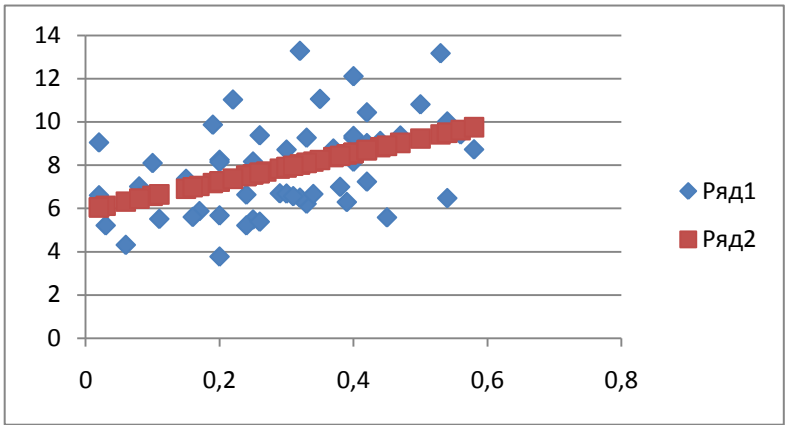


Рис. 1. Прямая линия регрессии на корреляционном поле

3. Показатели тесноты связи

Основные величины дисперсионного анализа

Полная (общая) сумма квадратов (TSS) определяет разброс (дисперсию) зависимой переменной относительно ее среднего значения

$$S^2_{общ} = \sum_{i=1}^{n} \left(y_i - \bar{y} \right)^2 .$$

Общая дисперсия

$$D_{общ} = \frac{S^2_{общ}}{n-1} \ .$$

Факторная (объясненная) сумма квадратов (RSS) определяет разброс значений $\hat{y}_i$ относительно среднего $\bar{y}$, объясненный регрессией,

$$S^2_{факт} = \sum_{i=1}^{n} \left(\hat{y}_i - \bar{y} \right)^2 .$$

Объясненная (факторная) дисперсия

$$D_{факт} = \frac{S^2_{факт}}{1} \ .$$

Остаточная сума квадратов (ESS), т.е. сумма квадратов остатков представляет часть дисперсии y , не объясненную регрессией

$$S^2_{ост} = \sum_{i=1}^{n} \left(y_i - \hat{y}_i \right)^2 = \sum_{i=1}^{n} \varepsilon_i^2 .$$

Остаточная дисперсия

$$D_{ост} = \frac{S^2_{ост}}{n-2} \ .$$

Здесь y_i - наблюдаемые значения зависимой переменной;

$\hat{y}_i$ - теоретические значения зависимой переменной, вычисленные по уравне-

нию регрессии;

$\bar{y}$ -среднее значение наблюдаемых значений.

Основное тождество дисперсионного анализа

$$S^2_{общ} = S^2_{факт} + S^2_{ост}$$

Построение уравнения парной линейной регрессии дополняется оценкой тесноты связи между зависимой и независимой переменными.

Коэффициент корреляции r_{yx} :

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{s_x \cdot s_y} \text{ или } r_{xy} = \frac{\overline{y \cdot x} - \bar{y} \cdot \bar{x}}{s_x \cdot s_y}$$

служит мерой **линейной корреляционно-регресионной зависимости между величинами y и x** , при условии, что на формирование их значений оказывают влияние некоторые другие, неучтенные факторы.

Коэффициент детерминации:

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}, \text{ или } R^2 = \frac{S^2_{факт}}{S^2_{общ}} = 1 - \frac{S^2_{ост}}{S^2_{общ}}, \text{ или } R^2 = (r_{xy})^2$$

оценивает долю вариации признака y , обусловленную изменением значений признака x. Чем ближе значение R^2 к единице, тем больше признак x участвует в формировании значений y.

4. Оценка качества уравнения парной линейной регрессии

Регрессионная модель представляет сумму уравнения регрессии и остатков. Проверяется качество обоих слагаемых.

Оценка качества уравнения линейной регрессии состоит из следующих этапов.

1. Оценка математической точности уравнения. Для этого рассчитывается ***средняя относительная ошибка аппроксимации***

$$\overline{A} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \cdot 100\%,$$

где y_i - фактические значения переменной y , $\hat{y}_i$ - теоретические значения y , найденные по уравнению регрессии. Для принятия решения о точности уравнения можно воспользоваться таблицей

Значение $\overline{A}$, %	**Точность уравнения**
менее 10	высокая
10 - 20	хорошая
20 - 50	удовлетворительная
более 50	неудовлетворительная

В случае, когда уравнение имеет неудовлетворительную точность, необходимо увеличить объем наблюдений (объем выборки) n, либо взять другое уравнение регрессии (нелинейное).

2. Проверка ***статистической значимости уравнения регрессии в целом*** с помощью ***F-критерия Фишера***.

Выдвигается гипотеза H_0: уравнение регрессии статистически незначимо, при конкурирующей гипотезе H_1: уравнение регрессии статистически значимо. Находится ***расчетное значение*** (***статистика***) ***критерия***

$$F_{расч} = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \cdot \frac{n - m - 1}{m},$$

или $F_{расч} = \frac{S^2_{факт}}{S^2_{ост}} \cdot \frac{n - m - 1}{m}$,

или $F_{расч} = \frac{R^2}{1 - R^2} \cdot \frac{n - m - 1}{m}$,

где y_i, $\hat{y}_i$, $\bar{y}$ - соответственно фактическое (наблюдаемое), теоретическое и среднее значение y ; n - объем выборки, m - число параметров уравнения регрессии при независимых переменных (***в случае парной линейной регрессии m = 1***), R^2 - коэффициент детерминации ($R^2 = (r)^2$) .

Табличное (***критическое***) значение $F_{табл}$, находится по таблице критических значений распределения Фишера-Снедекора (F-распределения) по уровню значимости α и двум числам степеней свободы $df_1 = m$ и $df_2 = n - m - 1$.

Если $F_{расч} > F_{табл}$, то гипотеза H_0 отвергается с вероятность ошибки α , т.е. уравнение регрессии признается в целом статистически значимым (***адекватно описывающим исходные данные***).

В противном случае ($F_{расч} < F_{табл}$) уравнение считается незначимым.

3. Проверка ***статистической значимости оценок параметров b_0, b_1*** с помощью ***t-критерия Стьюдента***.

Критерий Стьдента проверяется
только для линейного уравнения!

Выдвигается гипотеза H_0: параметр $b_j = 0$ (j = 0, 1) (статистически незначим, случайно отличается от 0), при конкурирующей гипотезе H_1: параметр $b_j \neq 0$ (статистически значим, неслучайно отличается от 0). Находится ***расчетное значение критерия***

$$t_{bj} = \left| \frac{b_j}{m_{bj}} \right|,$$

где средние квадратические ошибки параметров:

$$m_{b0} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{(n-2)} \cdot \frac{\sum_{i=1}^{n} x_i^2}{n\sum_{i=1}^{n}(x_i - \bar{x})^2}},$$

$$m_{b1} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{(n-2)\cdot\sum_{i=1}^{n}(x_i - \bar{x})^2}}.$$

Теоретическое значение критерия $t_{табл}$ находится по таблице критических значений распределения Стъюдента по уровню значимости α и числу степеней свободы df = n - m - 1 .

Если $t_{bj} > t_{табл}$, то гипотеза H_0 отвергается с вероятность ошибки α , т.е. оценка коэффициента регрессии b_j признается статистически значимой, т.е. не является результатом действия внешних случайных факторов.

В противном случае ($t_{bj} < t_{табл}$) - b_j статистически незначим.

4. Построение ***интервальных оценок*** (***доверительных интервалов***) параметров регрессии в виде

$$b_j - m_{bj} \cdot t_{табл} < \beta_j < b_j + m_{bj} \cdot t_{табл}$$

которые с ***надежностью*** (вероятностью) $\gamma = 1 - \alpha$ покрывают истинные параметры β_j. Здесь $t_{табл}$ - значение, найденное по таблице критических точек распределения Стъюдента по уровню значимости $\alpha/2$ и числу степеней свободы $df = n - m - 1$.

Если границы некоторого доверительного интервала имеют разные знаки, соответствующий параметр уравнения регрессии статистически незначим.

Если уравнение регрессии имеет хорошую математическую точность, статистически значимо в целом и по всем параметрам, оно признается качественным.

5. Анализ остатков

Согласно эконометрической модели

$$y_i = \hat{y}_i + \varepsilon_i$$

остатки ε_i находятся как разность между фактическими (наблюдаемыми) и теоретическими значениями зависимой переменной

Формула остатков

$$\varepsilon_i = y_i - \hat{y}_i$$

Остатки должны удовлетворять требованиям D, E теоремы Гаусса-Маркова.

1. Проверка требования D

D - ***остатки*** ε_i ***имеют нулевое математическое ожидание***

$$M(\varepsilon_i) = 0$$

и постоянную дисперсию σ_ε^2, не зависящую от номера наблюдения (свойство ***гомоскедастичности***), т.е.

$$D(\varepsilon_i) = \sigma_\varepsilon^2 = const;$$

Числовой оценкой математического ожидания является среднее значение. Необходимо, чтобы $\bar{\varepsilon} \approx 0$.

$$\bar{\varepsilon} = \frac{\sum_{i=1}^{n} \varepsilon_i}{n} \approx 0 .$$

Дисперсия остатков ε_i должна быть одинаковой для всех значений x_i (свойство ***гомоскедастичности***), т.е. $\sigma_\varepsilon^2 = const$.

Если это условие не соблюдается, то имеет место ***гетероскедастичность***, при которой разброс точек корреляционного поля по вертикале относительно линии регрессии меняется с ростом номера наблюдения. При гетероскедастичности оценки коэффициентов уравнения регрессии b_j ($j = 0, 1$) будут несмещенными, но неэффективными, вследствие чего окажутся завышенными расчетные значения ***t-критерия***, и будут сделаны неверные выводы о значимости коэффициентов регрессии.

Для обнаружения эффекта гетероскедастичности используют тесты Уайта, Голдфелда-Квандта, Глейзера и др., но они достаточно трудоемки, и на практике наиболее часто применяется графический метод, при котором строится и анализируется график зависимости остатков ε_i от номера наблюдения i.

В случае гомоскедастичности точки (остатки) равномерно располагаются внутри горизонтальной полосы, симметричной оси абсцисс (см. рис. 2).

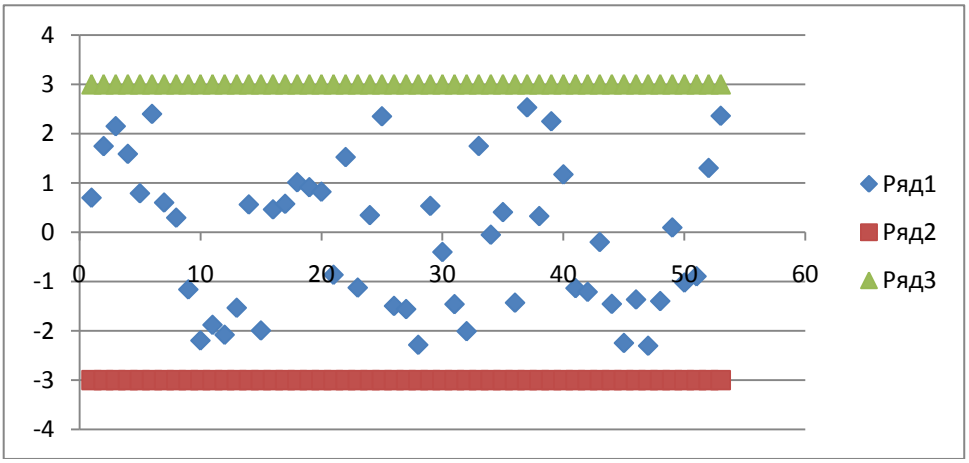


Рис. 2. Гомоскедастичность остатков

В случае гетероскедастичности точки (остатки), начиная с некоторого номера наблюдения, устойчиво выходят из горизонтальной полосы, симметричной оси абсцисс (см. рис. 3).

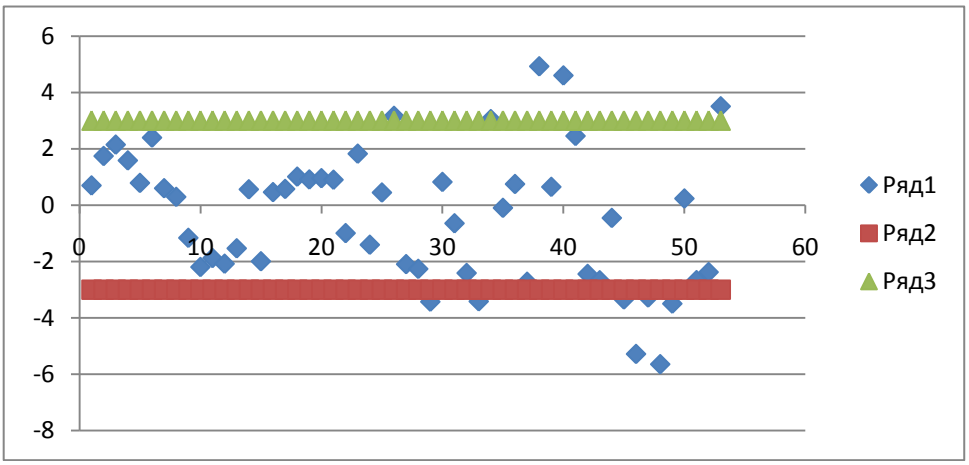


Рис. 3. Гетероскедастичность остатков

2. Проверка требования E

E - для разных наблюдений ***остатки*** ε_i ***некоррелированы*** **(*независимы*)**

Наиболее распространенный метод проверки требования о независимости остатков - ***критерий Дарбина-Уотсона*** (о наличии в остатках автокорреляции первого порядка), в котором рассчитывается статистика

$$d_{расч} = \frac{\sum_{i=2}^{n} (\varepsilon_i - \varepsilon_{i-1})^2}{\sum_{i=1}^{n} \varepsilon_i^2}.$$

Если в остатках существует полная положительная автокорреляция, то $d_{расч} = 0$. Если в остатках полная отрицательная автокорреляция, то $d_{расч} = 4$. Если автокорреляция остатков отсутствует, то $d_{расч} = 2$. Таким образом,

$$0 \le d_{расч} \le 4 .$$

Выдвигается гипотеза H_0 об отсутствии автокорреляции остатков. Конкурирующие гипотезы H_1 и H_1^* состоят в наличии в остатках положительной или отрицательной автокорреляции. Теоретические значения критерия Дарбина-Уотсона d_U и d_L находятся по таблице критических значений по объему выборки n, числу степеней свободы $df = m$ (m - число параметров уравнения при объясняющих переменных) и уровню значимости α. С помощью критических значений числовой промежуток (0; 4) разбивается на пять отрезков.

Решающее правило принятия или отклонения каждой из гипотез

Есть положительная автокорреляция остатков. H_0 отклоняется, принимается H_1	Зона неопределенности	Нет оснований отклонять H_0 (автокорреляция остатков отсутствует)	Зона неопределенности	Есть отрицательная автокорреляция остатков. H_0 отклоняется, принимается H_1^*
0 — d_L	d_L — d_U	d_U — 2	2 — $4 - d_U$	$4 - d_U$ — $4 - d_L$ — 4

В том случае, когда расчетное значение критерия попадает в зону неопределенности, на практике обычно признается наличие автокорреляции в остатках.

Точечная (числовая) оценка дисперсии остатков σ_ε^2 :

$$s_\varepsilon^2 = \frac{1}{(n-2)} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \approx \sigma_\varepsilon^2 .$$

Интервальная оценка дисперсии остатков - доверительный интервал:

$$\left(\frac{s_\varepsilon^2 \cdot (n-2)}{\chi^2(k, \alpha_1)}; \frac{s_\varepsilon^2 \cdot (n-2)}{\chi^2(k, \alpha_2)} \right), \quad \text{или} \quad \frac{s_\varepsilon^2 \cdot (n-2)}{\chi^2(k, \alpha_1)} < \sigma_\varepsilon^2 < \frac{s_\varepsilon^2 \cdot (n-2)}{\chi^2(k, \alpha_2)}$$

который с вероятностью (надежностью) $\gamma = 1 - \alpha$ покрывает σ_ε^2.

Здесь χ^2 - критические значения распределения Пирсона, найденные по таблице по числу степеней свободы $df = n - 2$ и уровням значимости $\alpha_1 = 1 - \alpha/2$, $\alpha_2 = \alpha/2$.

Если уравнение регрессии признано качественным, а остатки удовлетворяют требованиям D, E теоремы Гаусса-Маркова, то регрессионная модель считается качественной, т.е. она адекватно описывает исходные данные.

Для получения качественной парной линейной регрессионной модели необходима выборка объема не меньше $n = (6 \div 8) \times 2$.

6. Практические приложения регрессионных моделей

Наиболее часто встречающиеся приложения регрессионной модели - оценка влияния объясняющих переменных на результативный признак и построение прогноза.

1. Эластичность

Параметр b_1 нельзя использовать для непосредственной оценки влияния факторного признака x на результативный признак y из-за различия единиц измерения исследуемых показателей. Для этих целей вычисляют ***средний*** и ***частные коэффициенты эластичности*** и ***бета-коэффициент***.

Формула определения среднего коэффициента эластичности

$$\overline{Э}_x = b_1 \cdot \frac{\bar{x}}{\bar{y}}$$

Средний по совокупности коэффициент эластичности показывает, на сколько процентов изменяется $\bar{y}$ при изменении $\bar{x}$ на один процент.

Средний коэффициент эластичности позволяет выявить общегрупповые закономерности.

Формула определения частного коэффициента эластичности

$$Э_{xi} = b_1 \cdot \frac{x_i}{b_0 + b_1 \cdot x_i} \cdot$$

Здесь b_0 ,b_1 – параметры уравнения регрессии, x_i - i-е наблюдение независимой переменной (i = 1, 2,…, n).

Частный коэффициент эластичности показывает, на сколько процентов изменится y_i , если x_i увеличится на 1%.

Частные коэффициенты эластичности позволяют выявить особенности, присущие отдельным объектам наблюдения.

Формула определения бета-коэффициента

$$\beta_x = b_1 \frac{S_x}{S_y}$$

Бета-коэффициент показывает, на какую часть своего среднего квадратического отклонения изменится $\bar{y}$ при изменении $\bar{x}$ на величину своего среднего квадратического отклонения.

2. Прогнозирование

В прогнозных расчетах предсказываемое значение $y_{пр}$ определяется как точечный прогноз $\hat{y}_{пр}$ путем подстановки в уравнение регрессии соответствующих значений $x_{пр}$. Точечный прогноз дополняется расчетом средней ошибки, а также построением интервальной оценки (доверительного интервала).

Для парного линейного уравнения регрессии

$$y_{пр} = \hat{y}_{пр} = b_0 + b_1 \cdot x_{пр},$$

Формула средней стандартной ошибки прогноза

$$m_{\hat{y}пр} = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n-2} \cdot \left(1 + \frac{1}{n} + \frac{(x_{пр} - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}\right)}$$

Формула доверительного интервала прогноза

$$\hat{y}_{пр} - t_{табл} \cdot m_{\hat{y}пр} \le y_{пр} \le \hat{y}_{пр} + t_{табл} \cdot m_{\hat{y}пр}$$

Здесь $t_{табл}$ - критическое значение распределения Стьюдента, найденное по уровню значимости α и числу степеней свободы df = n - 2.

Наибольшая точность прогноза достигается в тех случаях, когда $x_{пр}$ находится в центре области наблюдений x , используемых при построении уравнения регрессии (незначительно отличается от $\bar{x}$). При удалении $x_{пр}$ от $\bar{x}$ средняя ошибка прогноза растет, а ширина интервальной оценки увеличивается. В случае, когда $x_{пр}$ оказывается за пределами области наблюдаемых значений, нельзя говорить о надежности $\hat{y}_{пр}$.

Тема 4. МНОЖЕСТВЕННАЯ ЛИНЕЙНАЯ РЕГРЕССИЯ

Лекция

План лекции

1. Множественный корреляционно-регрессионый анализ
2. Уравнение множественной линейной регрессии. Теорема Гаусса-Маркова
3. Оценка тесноты множественной линейной корреляционной связи
4. Оценка качества уравнения множественной линейной регрессии
5. Анализ остатков
6. Отбор факторных признаков в модель
7. Практические приложения регрессионной модели

1. Множественный корреляционно-регрессионый анализ

Экономические явления, как правило, определяются большим числом одновременно и совокупно действующих факторов. В связи с этим часто возникает задача исследования зависимости переменной y от нескольких объясняющих переменных $x_1, x_2, \ldots, x_m$, т.е. задача построения эконометрической модели

$$y = f(x_1, x_2, \ldots, x_m) + \varepsilon \ ,$$

где $\hat{y} = f(x_1, x_2, \ldots, x_m)$ - ***уравнение регрессии***, **ε** - ***остаток*** (***отклонение***),

которая может быть решена с помощью ***множественного корреляционно-регрессионного анализа***.

Функциональные задачи
множественного корреляционно-регрессионного анализа

Измерение тесноты связи между признаками
Отбор факторных признаков в модель
Установление неизвестных причин связей
Определение вида уравнения регрессии
Построение регрессионной модели и оценка ее параметров
Проверка значимости параметров связи
Интервальное оценивание параметров связи

При исследовании зависимости методами множественной регрессии задача формулируется так же, как и при использование парной регрессии: требуется определить аналитическое выражение формы связи между результативным признаком y и факторными признаками $x_1, x_2, \ldots, x_m$, т.е. найти функцию

$$\hat{y} = f(x_1, x_2, \ldots, x_m),$$

где m - число факторных признаков (независимых переменных).

2. Уравнение множественной линейной регрессии. Теорема Гаусса-Маркова

Из-за особенностей метода наименьших квадратов во множественной регрессии, как и в парной, применяются только линейные уравнения и уравнения, приводимые к линейному виду путем преобразования переменных. Причем из-за трудности обоснования формы связи чаще всего используется линейное уравнение.

Ввиду четкой интерпретации параметров наиболее часто используется

Уравнение множественной линейной регрессии в натуральном масштабе:

$$\hat{y} = b_0 + b_1 \cdot x_1 + b_2 \cdot x_2 + \ldots + b_m \cdot x_m \ .$$

Параметры $b_1, b_2, \ldots, b_m$ называются ***коэффициентами регрессии*** (***коэффициентами чистой регрессии***).

Содержание параметров уравнения множественной линейной регрессии

Параметр b_j (j = 1, 2,…, m) (коэффициент регрессии) показывает, на сколько единиц изменится $\bar{y}$ при увеличении независимой переменной x_j на единицу своего измерения, при неизменном значении других независимых переменных, закрепленных на среднем уровне.

Параметр b_0 (свободный член), как правило, не имеет экономического смысла, однако если независимые переменные одновременно могут принимать нулевое значение, он интерпретируется как начальный (нулевой) уровень значений y.

Нахождение значений параметров $b_0, b_1, b_2, \ldots, b_m$ производится на основе совокупности наблюдений (выборки, ***матрицы наблюдений***).

Для получения оценок параметров линейного регрессионного уравнения можно использовать метод наибольшего правдоподобия, метод наименьших модулей, метод минимакса и др., однако, согласно ***теореме Гаусса-Маркова***, наилучшие результаты дает ***метод наименьших квадратов*** (**МНК**).

Требования теоремы Гаусса-Маркова

A - «***истинная» зависимость*** y от $x_1, x_2, \ldots, x_m$ ***имеет вид***

$$y = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \ldots + \beta_m \cdot x_m + \varepsilon \ ;$$

B - $x_1, x_2, \ldots, x_m$ - ***неслучайные переменные*** (детерминированые);

C - ***столбцы матрицы наблюдений, с добавлением единичного столбца, линейно независимы*** (ранг матрицы равен m + 1);

D - ***остатки ε имеют нулевое математическое ожидание***, т.е.

$$M(\varepsilon_i) = 0 \ ,$$

и постоянную дисперсию σ_ε^2, не зависящую от номера наблюдения (свойство ***гомоскедастичности***), т.е.

$$D(\varepsilon_i) = \sigma_\varepsilon^2 = const \ ;$$

E - для разных наблюдений ***остатки ε некоррелированы*** (***независимы***), т.е. выполняется условие

$$\mathrm{cov}(\varepsilon_i, \varepsilon_l) = 0 \ , \text{ при } i \neq l, \quad i, l = 1, 2, \ldots, n \ .$$

Часто вместо **D** добавляют условие

F - ***остатки ε подчиняются нормальному закону распределения***.

Теорема Гаусса-Маркова

В предположениях **A - E** оценки, полученные методом наименьших квадратов, являются несмещенными и обладают наименьшей дисперсией среди всех линейных несмещенных оценок параметров β_j .

Сущность метода наименьших квадратов заключается в том, что отыскиваются такие значения параметров уравнения регрессии, при которых сумма квадратов отклонений фактических значений результативного признака y_i от вычисленных по уравнению регрессии ***теоретических значений*** $\hat{y}_i$, будет наименьшей из всех возможных:

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - (b_0 + b_1 \cdot x_1 + b_2 \cdot x_2 + \ldots + b_m \cdot x_m))^2 \to \min .$$

Путем преобразований требование минимума суммы квадратов отклонений сводится к системе нормальных уравнений.

В случае двух независимых переменных (m = 2) уравнение множественной линейной регрессии имеет вид

$$\hat{y} = b_0 + b_1 \cdot x_1 + b_2 \cdot x_2 .$$

Система нормальных уравнений для нахождения параметров множественной линейной регрессии (m = 2)

$$\begin{cases} b_0 n + b_1 \sum_{i=1}^{n} x_{1i} + b_2 \sum_{i=1}^{n} x_{2i} = \sum_{i=1}^{n} y_i \\ b_0 \sum_{i=1}^{n} x_{1i} + b_1 \sum_{i=1}^{n} x_{1i}^2 + b_2 \sum_{i=1}^{n} x_{1i} x_{2i} = \sum_{i=1}^{n} x_{1i} y_i \\ b_0 \sum_{i=1}^{n} x_{2i} + b_1 \sum_{i=1}^{n} x_{1i} x_{2i} + b_2 \sum_{i=1}^{n} x_{2i}^2 = \sum_{i=1}^{n} x_{2i} y_i \end{cases}$$

Доказано, что данная система совместна и имеет единственное решение.

Решение системы (значения параметров b_0, b_1, b_2) можно найти с помощью ***метода Крамера*** или ***метода Гаусса***.

! На практике для нахождения параметров уравнения множественной линейной регрессии обычно пользуются встроенным пакетом прикладных программ (ППП) MS Excel «Анализ данных - Регрессия».

После нахождения параметров можно записать регрессионную модель.

Модель множественной линейной регрессии

$$y = b_0 + b_1 \cdot x_1 + b_2 \cdot x_2 + \ldots + b_m \cdot x_m + \varepsilon$$

или

$$y_i = b_0 + b_1 \cdot x_{1i} + b_2 \cdot x_{2i} + \ldots + b_m \cdot x_{mi} + \varepsilon_i \quad (i = 1, 2, \ldots, n)$$

Для построенной регрессионной модели необходимо:

- оценить тесноту множественной линейной регрессионно-корреляционной связи;

- провести оценку качества модели, которая заключается в проверке уравнения

регрессии и в анализе остатков.

3. Оценка тесноты множественной линейной корреляционной связи

Построение уравнения парной линейной регрессии дополняется оценкой тесноты связи между зависимой и независимыми переменными.

Коэффициент детерминации:

$$R^2 = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}, \text{ или } R^2 = \frac{S^2_{факт}}{S^2_{общ}} = 1 - \frac{S^2_{ост}}{S^2_{общ}}$$

оценивает ***долю вариации*** признака y , обусловленную изменением значений переменных $x_1, x_2,\ldots, x_m$.

Коэффициент детерминации принимает значения от 0 до 1.

Чем ближе значение R^2 к единице, тем больше факторы $x_1, x_2,\ldots, x_m$ участвует в формировании значений y.

Коэффициент множественной корреляции

$$R = \sqrt{R^2}$$

служит ***мерой линейной корреляционно-регресионной зависимости*** между величинами y и $x_1, x_2,\ldots, x_m$, при условии, что на формирование их значений оказывают влияние некоторые другие, неучтенные факторы (см. Тему 2).

Коэффициент множественной корреляции принимает значения от 0 до 1.

Теснота связи между признаками устанавливается с помощью таблицы:

Качественные характеристики множественной корреляционной связи

Значение коэффициента множественной корреляции R	**Характер линейной корреляционной связи между признаками y и $x_1, x_2,\ldots, x_m$**
0 - 0,1	Слабая
0,1 - 0,5	Средняя (умеренная)
0,5 – 1	Сильная (тесная)

4. Оценка качества уравнения множественной линейной регрессии

Регрессионная модель представляет сумму уравнения регрессии и остатков. Проверяется качество обоих слагаемых.

Оценка качества уравнения линейной регрессии состоит из следующих этапов.

1. Оценка математической точности уравнения. Для этого рассчитывается ***средняя относительная ошибка аппроксимации***

$$\overline{A} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \cdot 100\%,$$

где y_i и $\hat{y}_i$ - соответственно фактические и теоретические значения переменной y .

Для принятия решения о точности уравнения можно воспользоваться таблицей

Значение $\bar{A}$, %	**Точность уравнения**
менее 10	высокая
10 - 20	хорошая
20 - 50	удовлетворительная
более 50	неудовлетворительная

В случае, когда уравнение имеет неудовлетворительную точность, необходимо увеличить объем наблюдений (объем выборки) n, скорректировать набор объясняющих переменных $x_1, x_2,\ldots, x_m$, либо взять другое уравнение регрессии (нелинейное).

2. Проверка ***статистической значимости уравнения регрессии в целом*** с помощью ***F-критерия Фишера***.

Выдвигается гипотеза $\boldsymbol{H_0}$: уравнение регрессии статистически незначимо, при конкурирующей гипотезе $\boldsymbol{H_1}$: уравнение регрессии статистически значимо. Находится ***расчетное значение*** (***статистика***) ***критерия***

$$F_{расч} = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \cdot \frac{n - m - 1}{m},$$

или $$F_{расч} = \frac{S^2_{факт}}{S^2_{ост}} \cdot \frac{n - m - 1}{m},$$

или $$F_{расч} = \frac{R^2}{1 - R^2} \cdot \frac{n - m - 1}{m},$$

где y_i, $\hat{y}_i$, $\bar{y}$ - соответственно фактическое (наблюдаемое), теоретическое и среднее значение y ; $\boldsymbol{n}$ - объем выборки, $\boldsymbol{m}$ - число параметров уравнения регрессии при независимых переменных, R^2 - коэффициент детерминации.

Табличное (***критическое***) значение $\boldsymbol{F_{табл}}$, находится по таблице критических значений распределения Фишера-Снедекора (***F-распределения***) по уровню значимости $\boldsymbol{\alpha}$ и двум числам степеней свободы $\boldsymbol{df_1 = m}$ и $\boldsymbol{df_2 = n - m - 1}$.

Если $\boldsymbol{F_{расч} > F_{табл}}$, то гипотеза $\boldsymbol{H_0}$ отвергается с вероятность ошибки $\boldsymbol{\alpha}$, т.е. уравнение регрессии признается в целом статистически значимым (***адекватно описывающим исходные данные***).

В противном случае ($\boldsymbol{F_{расч} < F_{табл}}$) уравнение считается незначимым.

3. Проверка ***статистической значимости оценок параметров $b_0, b_1, b_2,\ldots, b_m$*** с помощью ***t-критерия Стьюдента***.

! **Критерий Стьюдента проверяется только для линейного уравнения**

Выдвигается гипотеза $\boldsymbol{H_0}$: параметр $\boldsymbol{b_j} = 0$ (j = 0, 1, 2,…, m) (статистически незначим, случайно отличается от 0), при конкурирующей гипотезе $\boldsymbol{H_1}$: параметр $\boldsymbol{b_j} \neq 0$

(статистически значим, неслучайно отличается от 0). Находятся ***расчетные значения критерия***

$$t_{bj} = \left| \frac{b_j}{S_\varepsilon \sqrt{a_j}} \right|,$$

где S_ε - стандартное (среднее квадратическое) отклонение (ошибка) уравнения регрессии, определяется по формуле

$$S_\varepsilon = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n-m-1}},$$

a_j - диагональные элементы матрицы $(X'X)^{-1}$

X – матрица значений независимых переменных $x_1, x_2, \ldots, x_m$, размерность матрицы равна $n \times (m+1)$. Первый столбец матрицы является единичным.

$$X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1m} \\ 1 & x_{21} & \cdots & x_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ 1 & x_{n1} & \cdots & x_{nm} \end{pmatrix}.$$

X' - транспонированная матрица X ,

$$X' = \begin{pmatrix} 1 & 1 & \ldots & 1 \\ x_{11} & x_{21} & \cdots & x_{n1} \\ \ldots & \ldots & \ldots & \ldots \\ x_{1m} & x_{2m} & \cdots & x_{nm} \end{pmatrix}.$$

$(X'X)^{-1}$ - обратная матрица.

Теоретическое значение критерия $t_{табл}$ находится по таблице критических значений распределения Стъюдента по уровню значимости α и числу степеней свободы $df = n - m - 1$.

Если $t_{bj} > t_{табл}$, то гипотеза H_0 отвергается с вероятность ошибки α , т.е. оценка коэффициента регрессии b_j признается статистически значимой, т.е. не является результатом действия внешних случайных факторов.

В противном случае ($t_{bj} < t_{табл}$) - незначимой.

Если статистическая значимость коэффициента регрессии b_j не подтверждается, то следует вывод о несущественности в модели факторного признака x_j и необходимости его устранения из модели или замены на другой факторный признак.

4. Построение ***интервальных оценок*** (***доверительных интервалов***) параметров регрессии.

Интервальные оценки параметров регрессии

$$b_j - m_{bj} \cdot t_{табл} < \beta_j < b_j + m_{bj} \cdot t_{табл} \qquad (j = 0, 1, 2, \ldots, m)$$

которые с ***надежностью*** (вероятностью) $\gamma = 1 - \alpha$ покрывают истинные параметры β_j. Если границы некоторого доверительного интервала имеют разные знаки, соответствующий параметр уравнения регрессии статистически незначим.

Здесь $t_{табл}$ - значение, найденное по таблице критических точек распределения Стьюдента по уровню значимости $1-\alpha$ и числу степеней свободы $df = n - m - 1$, m_{bj} - стандартные ошибки коэффициентов регрессии.

В случае уравнения регрессии с двумя независимыми переменными (m = 2) $\hat{y} = b_0 + b_1 \cdot x_1 + b_2 \cdot x_2$ m_{bj} находятся по формулам:

$$m_{b0} = \sqrt{S_\varepsilon^2 \cdot \frac{\frac{1}{n} + \bar{x}_1^2 \sum_{i=1}^{n} (x_{i2} - \bar{x}_2)^2 + \bar{x}_2^2 \sum_{i=1}^{n} (x_{i1} - \bar{x}_1)^2 - 2\bar{x}_1\bar{x}_2 \sum_{i=1}^{n} (x_{i1} - \bar{x}_1)(x_{i2} - \bar{x}_2)}{\sum_{i=1}^{n} (x_{i2} - \bar{x}_2)^2 \cdot \sum_{i=1}^{n} (x_{i1} - \bar{x}_1)^2 - \left(\sum_{i=1}^{n} (x_{i1} - \bar{x}_1)(x_{i2} - \bar{x}_2)\right)^2}};$$

$$m_{b1} = \sqrt{S_\varepsilon^2 \cdot \frac{1}{\sum_{i=1}^{n} (x_{i1} - \bar{x}_1)^2 \cdot (1 - r_{x1x2}^2)^2}};$$

$$m_{b2} = \sqrt{S_\varepsilon^2 \cdot \frac{1}{\sum_{i=1}^{n} (x_{i2} - \bar{x}_2)^2 \cdot (1 - r_{x1x2}^2)^2}}.$$

В случае большего числа факторов эти формулы более громоздки и трудоемки.

! **На практике стандартные ошибки коэффициентов регрессии и расчетные значения критерия Стьюдента находятся с помощью пакета MS Excel «Анализ данных – Регрессия»**

Если уравнение регрессии имеет хорошую математическую точность, статистически значимо в целом и по отдельным параметрам, оно признается качественным.

5. Анализ остатков

Согласно эконометрической модели

$$y_i = \hat{y}_i + \varepsilon_i$$

остатки ε_i находятся как разность между фактическими (наблюдаемыми) и теоретическими значениями зависимой переменной

Формула остатков

$$\varepsilon_i = y_i - \hat{y}_i$$

Остатки должны удовлетворять требованиям D, E теоремы Гаусса-Маркова.

1. Проверка требования D

D - ***остатки*** ε_i ***имеют нулевое математическое ожидание***

$$M(\varepsilon_i) = 0$$

и постоянную дисперсию σ_ε^2**,** не зависящую от номера наблюдения (свойство ***гомоскедастичности***), т.е.

$$D(\varepsilon_i) = \sigma_\varepsilon^2 = const\ ;$$

Числовой оценкой математического ожидания является среднее значение. Необходимо, чтобы $\bar{\varepsilon} \approx 0$.

$$\bar{\varepsilon} = \frac{\sum_{i=1}^{n} \varepsilon_i}{n} \approx 0 \ .$$

Дисперсия остатков ε_i должна быть одинаковой для всех значений x_i (свойство ***гомоскедастичности***), т.е. $\sigma_\varepsilon^2 = const$.

Если это условие не соблюдается, то имеет место ***гетероскедастичность.*** При гетероскедастичности оценки коэффициентов уравнения регрессии $\boldsymbol{b}_j$ ($\boldsymbol{j} = 0, 1, \ldots, \boldsymbol{m}$) будут несмещенными, но неэффективными, вследствие чего окажутся завышенными расчетные значения ***t***-критерия, и будут сделаны неверные выводы о значимости коэффициентов регрессии.

Для обнаружения эффекта гетероскедастичности используют тесты Уайта, Голдфелда-Квандта, Глейзера и др., но на практике наиболее часто применяется графический метод, при котором строится и анализируется график зависимости остатков $\boldsymbol{\varepsilon_i}$ от номера наблюдения ***i*** (см. Тему 3).

2. Проверка требования E

E - для разных наблюдений ***остатки*** ε_i ***некоррелированы*** **(*независимы*)**

Наиболее распространенный метод проверки требования о независимости остатков - ***критерий Дарбина-Уотсона*** (о наличии в остатках автокорреляции первого порядка), в котором рассчитывается статистика

$$d_{расч} = \frac{\sum_{i=2}^{n} (\varepsilon_i - \varepsilon_{i-1})^2}{\sum_{i=1}^{n} \varepsilon_i^2} \ .$$

Если в остатках существует полная положительная автокорреляция, то $\boldsymbol{d_{расч}} = 0$. Если в остатках полная отрицательная автокорреляция, то $\boldsymbol{d_{расч}} = 4$. Если автокорреляция остатков отсутствует, то $\boldsymbol{d_{расч}} = 2$. Таким образом,

$0 \le d_{расч} \le 4$.

Выдвигается гипотеза H_0 об отсутствии автокорреляции остатков. Конкурирующие гипотезы H_1 и H_1^* состоят в наличии в остатках положительной или отрицательной автокорреляции. Теоретические значения критерия Дарбина-Уотсона d_U и d_L находятся по таблице критических значений по объему выборки n, числу степеней свободы $df = m$ (m - число параметров уравнения при объясняющих переменных) и уровню значимости α. С помощью критических значений числовой промежуток (0; 4) разбивается на пять отрезков.

Решающее правило принятия или отклонения с вероятностью (1 – α) каждой из гипотез

Есть положительная автокорреляция остатков. H_0 отклоняется, принимается H_1	Зона неопределенности	Нет оснований отклонять H_0 (автокорреляция остатков отсутствует)	Зона неопределенности	Есть отрицательная автокорреляция остатков. H_0 отклоняется, принимается H_1^*
0 — d_L	d_L — d_U	d_U — 2	2 — $4 - d_U$	$4 - d_U$ — $4 - d_L$ — 4

В том случае, когда расчетное значение критерия попадает в зону неопределенности, на практике обычно признается наличие автокорреляции в остатках.

Точечная (числовая) оценка дисперсии остатков σ_ε^2

$$s_\varepsilon^2 = \frac{1}{(n-2)} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \approx \sigma_\varepsilon^2 .$$

Интервальная оценка дисперсии остатков - доверительный интервал,

$$\left(\frac{s_\varepsilon^2 \cdot (n-2)}{\chi^2(k, \alpha_1)}; \frac{s_\varepsilon^2 \cdot (n-2)}{\chi^2(k, \alpha_2)} \right), \quad \text{или} \quad \frac{s_\varepsilon^2 \cdot (n-2)}{\chi^2(k, \alpha_1)} < \sigma_\varepsilon^2 < \frac{s_\varepsilon^2 \cdot (n-2)}{\chi^2(k, \alpha_2)}$$

который с вероятностью (надежностью) $\gamma = 1 - \alpha$ покрывает σ_ε^2.

Здесь χ^2 - критические значения ***распределения Пирсона***, найденные по таблице по числу степеней свободы $df = n - m - 1$ и уровням значимости $\alpha_1 = \alpha/2$ и $\alpha_2 = 1 - \alpha/2$.

Если уравнение регрессии признано качественным, а остатки удовлетворяют требованиям D, E теоремы Гаусса-Маркова, то регрессионная модель считается качественной, т.е. она адекватно описывает исходные данные.

! **Для получения качественной множественной линейной регрессионной модели с m факторами необходима выборка объема не менее n = (6÷8)×(m+1).**

6. Отбор факторных признаков в модель

Формирование оптимального набора объясняющих переменных - важное условие построения качественной регрессионной модели. Доказано, что для получения статистически значимого значения одного коэффициента уравнения регрессии необходимо, как минимум 6÷8 наблюдений. Если уравнение строится по m объясняющим переменным, для его статистической значимости необходима выборка объема $n = (6÷8)×m$. На практике не всегда возможно получить выборку большого объема. Поэтому модель должна содержать как можно меньше независимых переменных, оптимальных с точки зрения их влияния на формирование значений y и объяснения сущности исследуемого процесса. При этом необходимо помнить, что ошибки спецификации могут не только значительно ухудшить качество модели, но и сделать ее ошибочной, неадекватно описывающей моделируемый экономический процесс.

Наиболее важные и часто встречающиеся ошибки спецификации

Неверный выбор вида уравнения регрессии.
Неправомерное исключение объясняющих переменных, приводящее к нарушению свойства несмещенности оценок коэффициентов регрессии.
Неправомерное включение объясняющих переменных, приводящее к мультиколлинеарности переменных и неразрешимости системы нормальных уравнений метода наименьших квадратов.

С формальной точки зрения, объясняющие переменные в эконометрической модели должны быть сильно коррелированны с зависимой переменной; быть слабо коррелированны между собой; быть сильно коррелированны с представляемыми ими другими переменными, не используемыми в качестве объясняющих, т.е. хорошо представлять те переменные, которые не были включены в модель.

Отбор факторов проводится на основе качественного и количественного анализа социально-экономических явлений с использованием статистических и математических критериев. Проводят три стадии отбора факторов.

Стадии отбора факторов в модель множественной регрессии
1. Предварительное определение перечня факторов исходя из экономической сущности решаемой проблемы, оказывающих влияние на переменную y.
2. Сравнительная оценка и отсев факторов.
3. Окончательный отбор факторов в процессе построения разных вариантов моделей и оценки значимости их параметров.

Требования к объясняющим переменным включаемым в модель.

1. Переменные x1, x2,…, x_m , должны быть ***количественно измеримы***. Если необходимо включение неколичественных данных, они должны быть оцифрованы.

2. Переменные должны иметь ***высокую вариабельность*** (иметь достаточно большой разброс значений). В качестве меры вариабельности переменной x_j ($j = 1, 2,…, m$) используется коэффициент вариации

$$\boxed{v_j = \frac{S_{xj}}{\overline{X}_j}},$$

где $\overline{X}_j$, S_{xj} -среднее значение и среднеквадратическое отклонение переменной x_j.

Принять решение о степени вариабельности переменной можно с помощью таблицы, либо с помощью экспертного задания некоторого критического значения коэффициента вариации v^*, например $v^* = 0,2$. Переменные, удовлетворяющие нера-

венству $v_j \le v^*$, признаются квазинеизменными (условно неизменными) и исключаются из множества потенциальных объясняющих переменных, поскольку не несут значимой информации.

Значение коэффициента вариации *v*	**Характер распределения переменной**
0 - 0,2	весьма равномерный
0,2 - 0,4	равномерный
0,4 - 1	неравномерный
1 - 1,5	весьма неравномерный
более 1,5	крайне неравномерный

3. Переменные должны ***объяснять вариацию (изменчивость) зависимой переменной*** y. Вклад независимых переменных в формирование значений y оценивается с помощью коэффициента детерминации R^2 (см. Тему 2). Чем ближе значение R^2 к единице, тем лучше включенные в модель факторы описывают изменчивость y.

При формировании набора объясняющих переменных желательно, что бы коэффициент детерминации принимал как можно большее значение. В тоже время, при добавлении в модель новых объясняющих переменных коэффициент детерминации R^2 не может уменьшиться. Поэтому при сравнении наборов с разным числом переменных не всегда ясно за счет чего возрос коэффициент детерминации: за счет реального влияния дополнительно включенного фактора на результат y или просто из-за возрастания числа переменных. Для возможности сравнения моделей с различным числом переменных вводится скорректированный коэффициент детерминации $R^2_{скорр}$

$$R^2_{скорр} = 1 - (1 - R^2)\frac{n-1}{n-m-1},$$

здесь n - число наблюдений, m - число объясняющих переменных. Чем больше m , тем меньше различия R^2 и $R^2_{скорр}$.

4. Переменные ***не должны быть интеркоррелированными***, тем более не должны находиться в функциональной зависимости. Говорят, что переменные x_1 и x_2 интеркоррелированы, если выполняются условия:

$$|r_{x1x2}| > 0{,}75 \quad \text{и} \quad \begin{cases} |r_{yx1}| < |r_{x1x2}| \\ |r_{yx2}| < |r_{x1x2}| \end{cases}.$$

Другими словами, интеркоррелированные независимые переменные связаны между собой теснее, чем с y. Одна из пары переменных, обладающих интеркорреляцией, должна быть исключена из набора, поскольку она оказывает на y дублирующее влияние.

5. Переменные ***не должны быть мультиколлинеарными***.

7. Практические приложения регрессионной модели

Наиболее часто встречающиеся приложения регрессионной модели - оценка влияния объясняющих переменных на результативный признак и построение прогноза.

1. Оценка влияния фактора на результат. Эластичность

Важную роль при оценке влияния факторов играют коэффициенты регрессионной модели. Однако непосредственно с их помощью нельзя сопоставлять факторные признаки по степени их влияния на зависимую переменную из-за различия единиц измерения и разной степени колеблемости. Для устранения таких различий применяются ***средние коэффициенты эластичности*** $Э_j$. , ***бета-коэффициенты*** β_j (стандартизованные коэффициенты регрессии) и дельта-коэффициенты Δ_j.

Формула определения средних коэффициентов эластичности

$$\overline{Э}_j = b_j \cdot \frac{\overline{x}_j}{\overline{y}} ,$$

где b_j - коэффициент регрессии фактора x_j , $\overline{y}$ - среднее значение результативного признака y, $\overline{x}_j$ среднее значение признака x_j .

Средний по совокупности наблюдений коэффициент эластичности показывает, на сколько процентов изменяется зависимая переменная y при изменении фактора x_j на 1%, при условии, что остальные факторы постоянны и закреплены на среднем уровне. Средний коэффициент эластичности позволяет выявить общегрупповые закономерности.

На основе линейного уравнения множественной регрессии

$$\hat{y} = b_0 + b_1 \cdot x_1 + b_2 \cdot x_2 + \ldots + b_p \cdot x_p$$

можно построить ***частные уравнения регрессии***, которые характеризуют изолированное влияние отдельных объясняющих переменных на результат, поскольку остальные независимые переменные закреплены на среднем уровне. Эти уравнения имеют вид

$$\hat{y}_{x1 \cdot x2,x3,\ldots,xp} = b_0 + b_1 \cdot x_1 + b_2 \cdot \overline{x}_2 + b_3 \cdot \overline{x}_3 + \ldots + b_p \cdot \overline{x}_p ;$$

$$\hat{y}_{x2 \cdot x1,x3,\ldots,xp} = b_0 + b_1 \cdot \overline{x}_1 + b_2 \cdot x_2 + b_3 \cdot \overline{x}_3 + \ldots + b_p \cdot \overline{x}_p ;$$

……..

$$\hat{y}_{xp \cdot x1,x2,\ldots,xp-1} = b_0 + b_1 \cdot \overline{x}_1 + b_2 \cdot \overline{x}_2 + \ldots + b_{p-1} \cdot \overline{x}_{p-1} + b_p \cdot x_p .$$

После приведения подобных частные уравнения регрессии принимают вид парных уравнений линейной регрессии

$$\hat{y}_{x1 \cdot x2,x3,\ldots,xp} = A_1 + b_1 \cdot x_1 ;$$

$$\hat{y}_{x2 \cdot x1,x3,\ldots,xp} = A_2 + b_2 \cdot x_2 ;$$

……

$$\hat{y}_{xp \cdot x1,x2,\ldots,xp-1} = A_p + b_p \cdot x_p .$$

Здесь

$$A_1 = b_0 + b_2 \cdot \overline{x}_2 + b_3 \cdot \overline{x}_3 + \ldots + b_p \cdot \overline{x}_p ;$$

$$A_2 = b_0 + b_1 \cdot \bar{x}_1 + b_3 \cdot \bar{x}_3 + \ldots + b_p \cdot \bar{x}_p \ ;$$

……

$$A_p = b_0 + b_1 \cdot \bar{x}_1 + b_2 \cdot \bar{x}_2 + \ldots + b_{p-1} \cdot \bar{x}_{p-1} \ .$$

В частных уравнениях регрессии эффекты влияния остальных независимых переменных присоединены к свободному члену. Это позволяет определять ***частные коэффициенты эластичности***

$$Э_{xij} = b_j \cdot \frac{x_{ij}}{\hat{y}_{xj \cdot x1, x2, \ldots, xj-1, xj+1, \ldots, xp}} \quad \text{или}$$

$$Э_{xij} = b_j \cdot \frac{x_{ij}}{A_j + b_j \cdot x_{ij}} \ ,$$

здесь $\boldsymbol{b}_j$ - коэффициент регрессии, $\boldsymbol{x}_{ij}$ - ***i***-е наблюдение ***j***-й независимой переменной ($\boldsymbol{i}$ = 1, 2,…, $\boldsymbol{n}$, $\boldsymbol{j}$ = 1, 2,…, $\boldsymbol{p}$), $\hat{y}_{xj \cdot x1, x2, \ldots, xj-1, xj+1, \ldots, xp}$ - частное уравнение регрессии.

Частный коэффициент эластичности показывает, на сколько процентов изменится $\boldsymbol{y}_i$, если $\boldsymbol{x}_{ij}$ увеличится на 1%, при условии, что остальные переменные $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_{j-1}, \boldsymbol{x}_{j+1}, \ldots, \boldsymbol{x}_p$ закреплены на среднем уровне. Частные коэффициенты эластичности позволяют выявить особенности, присущие отдельным объектам наблюдения.

Формула определения бета-коэффициентов

$$\beta_j = b_j \frac{s_{xj}}{s_y} \ ,$$

где s_{xj} - среднее квадратическое отклонение фактора x_j , s_y – среднее квадратическое отклонение фактора y.

Бета-коэффициент показывает, на какую часть своего среднего квадратического отклонения s_y изменится $\bar{y}$ при изменении $\bar{x}_j$ на величину своего среднего квадратического отклонения s_{xj} при фиксированном значении остальных независимых переменных.

Указанные коэффициенты позволяют проранжировать факторы по степени их влияния на зависимую переменную.

Долю влияния фактора в суммарном влиянии всех факторов можно оценить по величине дельта-коэффициентов.

Формула определения дельта-коэффициентов

$$\Delta_j = r_{yj} \frac{\beta_j}{R^2}$$

где r_{yj} - коэффициент парной корреляции между фактором x_j и зависимой переменной y; β_j - бета-коэффициент фактора x_j , R^2 - множественный коэффициент детерминации.

2. Прогнозирование

В прогнозных расчетах предсказываемое значение $y_{\text{пр}}$ определяется как точечный прогноз $\hat{y}_{пр}$ путем подстановки в уравнение регрессии соответствующих значений факторных переменных $x_{1\text{пр}}, x_{2\text{пр}},\ldots, x_{m\text{пр}}$.

Наибольшая точность прогноза достигается в тех случаях, когда $x_{j\text{пр}}$ находятся в центре области наблюдений x_j , используемых при построении уравнения регрессии (незначительно отличаются от $\bar{x}_j$). При удалении $x_{j\text{пр}}$ от $\bar{x}_j$ средняя ошибка прогноза растет. В случае, когда $x_{j\text{пр}}$ оказывается за пределами области наблюдаемых значений, нельзя говорить о надежности $\hat{y}_{пр}$.

Тема 5. НЕЛИНЕЙНАЯ РЕГРЕССИЯ

Лекция

План лекции

1. Нелинейные модели
2. Форма нелинейной регрессионной зависимости
3. Оценка параметров уравнения нелинейной регрессии
4. Характеристики нелинейной регрессионной модели
5. Оценка качества уравнения нелинейной регрессии
6. Приложения нелинейных регрессионных моделей

1. Нелинейные модели

Во многих практических случаях моделирование экономических явлений и процессов линейными уравнениями дает вполне удовлетворительный результат и может использоваться для целей анализа и прогнозирования. Однако в силу многообразия и сложности экономических процессов многие экономические зависимости не являются линейными по своей сути и не могут быть описаны линейными моделями.

Если между экономическими явлениями существуют нелинейные соотношения, то они выражаются с помощью соответствующих нелинейных функций.

Например, при рассмотрении спроса y на некоторый товар от цены x данного товара в ряде случаев можно ограничиться линейным уравнением регрессии. Если необходимо проанализировать эластичность спроса по цене необходимо рассматривать ***логарифмическую модель***.

При анализе издержек y от объема выпуска x наиболее обоснованной является ***полиномиальная*** (а именно, кубическая) ***модель***.

Нелинейными являются ***производственные функции*** (зависимости между объемом произведенной продукции и основными факторами производства - трудом, капиталом и т.д.), ***функции спроса*** (зависимость между спросом на товары, услуги и их ценами или доходом) и др.

При рассмотрении производственных функций обычно применяют ***степенные модели***. Наиболее широкую известность имеет ***производственная функция Кобба-Дугласа*** $y = A \cdot K^{\alpha} \cdot L^{\beta}$, здесь y - объем выпуска; K и L - затраты капитала и труда соответственно; A, α, β - параметры модели.

В эконометрическом анализе достаточно широко применяются ***обратная*** и ***экспоненциальная модели***.

2. Форма нелинейной регрессионной зависимости

Под ***формой нелинейной регрессионной зависимости*** (видом связи) понимают вид уравнения регрессии.

Наиболее часто используются следующие виды нелинейных зависимостей:

1. ***степенное*** уравнение $\hat{y} = b_0 \cdot x_1^{b1} \cdot x_2^{b2} \cdot \ldots \cdot x_m^{bm}$

Параметры $b_1, b_2, \ldots, b_m$ степенного уравнения являются коэффициентами эластичности. Параметр b_j ($j = 1, 2, \ldots, m$) показывает, на сколько процентов изменится $\bar{y}$

при увеличении независимой переменной $\boldsymbol{x}_j$ на один процент, при неизменном значении других независимых переменных, закрепленных на среднем уровне.

2. ***показательное*** уравнение $\hat{y} = b_0 \cdot b_1^{x1} \cdot b_2^{x2} \cdot \ldots \cdot b_m^{xm}$;

3. ***экспоненциальное*** уравнение $\hat{y} = b_0 \cdot e^{b1x1} \cdot e^{b2x2} \cdot \ldots \cdot e^{bmxm}$;

4. ***гиперболическое*** уравнение $\hat{y} = b_0 + b_1 \frac{1}{x_1} + b_2 \frac{1}{x_2} + \ldots + b_m \frac{1}{x_m}$;

5. различные ***полиномиальные*** уравнения, например

$$\hat{y} = b_0 + b_1 \cdot x_1 + b_2 \cdot x_2^2 + \ldots + b_m \cdot x_m^m \cdot$$

Здесь y - зависимая переменная, x_1, x_2,…, x_m - независимые (объясняющие) переменные, b_0, b_1, b_2,…, b_m - параметры уравнения.

Для заданной матрицы наблюдений (таблицы исходных данных) можно построить несколько статистически значимых регрессионных уравнений. Выбор уравнения, наилучшим образом описывающего исходные данные, осуществляется по наибольшему значению ***скорректированного коэффициента детерминации*** $R^2_{скорр}$

$$\boxed{R^2_{скорр} = 1 - (1 - R^2)\frac{n-1}{n-m-1}} \cdot$$

Скорректированный коэффициент детерминации обладает свойством

$$0 < R^2_{скорр} < 1 \cdot$$

Однако на практике, если построенное уравнение не подходит для описания существующей в исходных данных нелинейной зависимости, коэффициент детерминации принимает значение $R^2 < 0{,}1$, а скорректированный коэффициент детерминации становится отрицательным ($R^2_{скорр} < 0$) .

Уравнение нелинейной регрессионной связи должно быть по возможности более простым, чтобы сущность изучаемой зависимости между переменными проявлялась достаточно четко, а параметры уравнения поддавались определенному экономическому толкованию. Вопрос выбора соответствующего уравнения связи решается в каждом случае отдельно.

Построение и анализ нелинейных моделей имеют свою специфику.

3. Оценка параметров уравнения нелинейной регрессии

Для получения оценок параметров нелинейных регрессионных уравнений можно использовать метод наибольшего правдоподобия, метод наименьших модулей, метод минимакса и др. Эти методы сложны и трудоемки, требуют хороших знаний многомерной математической статистики.

Оценка параметров нелинейной регрессии может осуществляться также с помощью метода наименьших квадратов (МНК) путем решения системы нормальных уравнений с предварительной линеаризацией уравнения регрессии.

Линеаризация уравнения – приведение уравнения к линейному виду.

Нелинейность может проявляться как относительно переменных, так и относительно входящих в функцию коэффициентов (параметров). Различают два класса нелинейных регрессий.

Классы нелинейных регрессий

Регрессии, ***нелинейные по переменным***, включенным в анализ, но линейные по оцениваемым параметрам (различные полиномы, гипербола)	Регрессии, ***нелинейные по оцениваемым параметрам*** (степенная, показательная, экспоненциальная функции)

В рамках данного курса ограничимся рассмотрение нелинейных моделей, допускающих их сведение к линейным.

Рассмотрим приемы линеаризации и нахождение параметров на примере нелинейных регрессионных уравнений с одной независимой переменной m = 1 (парных нелинейных регрессий).

Регрессии, нелинейные по переменным, но линейные по оцениваемым параметрам

Наименование регрессии	Уравнение регрессии	Нормальные уравнения
Полином второго порядка	$\hat{y} = b_0 + b_1 \cdot x + b_2 \cdot x^2$	$\begin{cases} b_0 n + b_1 \sum_{i=1}^{n} x_i + b_2 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} y_i \\ b_0 \sum_{i=1}^{n} x_i + b_1 \sum_{i=1}^{n} x_i^2 + b_2 \sum_{i=1}^{n} x_i^3 = \sum_{i=1}^{n} x_i y_i \\ b_0 \sum_{i=1}^{n} x_i^2 + b_1 \sum_{i=1}^{n} x_i^3 + b_2 \sum_{i=1}^{n} x_i^4 = \sum_{i=1}^{n} x_i^2 y_i \end{cases}$
Гипербола	$\hat{y} = b_0 + b_1 \frac{1}{x}$	$\begin{cases} b_0 n + b_1 \sum_{i=1}^{n} \frac{1}{x_i} = \sum_{i=1}^{n} y_i \\ b_0 \sum_{i=1}^{n} \frac{1}{x_i} + b_1 \sum_{i=1}^{n} \frac{1}{x_i^2} = \sum_{i=1}^{n} \frac{1}{x_i} y_i \end{cases}$ Или заменим 1/x на новую переменную X. В результате получим линейное уравнение $\hat{y} = b_0 + b_1 \cdot X$, параметры которого определяются по формулам: $b_1 = \frac{\sum_{i=1}^{n} X_i y_i - \frac{1}{n} \sum_{i=1}^{n} X_i \sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} X_i^2 - \frac{1}{n} \left(\sum_{i=1}^{n} X_i \right)^2}$, $b_0 = \bar{y} - b_1 \bar{X}$.

Регрессии, нелинейные по оцениваемым параметрам

Наименование регрессии	Уравнение регрессии	Линеаризация
Степенная функция	$\hat{y} = b_0 \cdot x^{b1}$	Для определения параметров степенной функции с помощью МНК необходимо привести ее к линейному виду путем логарифмирования обеих частей уравнения, после чего получим $\ln \hat{y} = \ln b_0 + b_1 \ln x \cdot$ Это уравнение представляет собой прямую линию на графике, по осям которого откладываются не сами числа, а их логарифмы (так называемая логарифмическая шкала или логарифмическая сетка). Пусть $Y = \ln \hat{y}$, $X = \ln x$, $B_0 = \ln b_0 \cdot$ Тогда уравнение примет вид $Y = B_0 + b_1 \cdot X \cdot$ Параметры уравнения определяются по формулам: $b_1 = \dfrac{\sum_{i=1}^{n} X_i Y_i - \frac{1}{n} \sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{\sum_{i=1}^{n} X_i^2 - \frac{1}{n} \left(\sum_{i=1}^{n} X_i \right)^2}$, $B_0 = \overline{Y} - b_1 \overline{X} \cdot$ Обратный переход к параметру b_0 осуществляется по формуле $b_0 = e^{B0} \cdot$
Показательная функция	$\hat{y} = b_0 \cdot b_1^x$	Линеаризацию переменных проведем путем логарифмирования обеих частей уравнения, после чего получим: $\ln \hat{y} = \ln b_0 + x \ln b_1 \cdot$ Уравнение изображается прямой линией на полулогарифмической сетке, которая получается как сочетание натуральной шкалы для значений независимой переменной x и логарифмической шкалы – для значений зависимой переменной y. Пусть $Y = \ln \hat{y}$, $B_0 = \ln b_0$, $B_1 = \ln b_1$, тогда уравнение примет вид

		$Y = B_0 + B_1 \cdot x$. Параметры модели определяются по формулам $B_1 = \frac{\sum_{i=1}^{n} x_i Y_i - \frac{1}{n}\sum_{i=1}^{n} x_i \sum_{i=1}^{n} Y_i}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2}$; $B_0 = \bar{Y} - B_1 \bar{x}$. Обратный переход к параметрам b_0 , b_1 осуществляется по формулам $b_0 = e^{B0}$, $b_1 = e^{B1}$.

4. Характеристики нелинейной регрессионной модели

Теснота нелинейной регрессионной связи между переменными может быть измерена с помощью ***индекса корреляции*** (***корреляционного отношения***)

$$\rho = \sqrt{1 - \frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}} .$$

Индекс корреляции принимает значение от 0 до 1. Чем больше значение индекса корреляции, тем ближе расчетные значения результативного признака $\hat{y}_i$ к фактическим y_i , тем теснее (сильнее) нелинейная связь.

Индекс корреляции используется при любой форме связи любого количества независимых переменных; при парной линейной регрессии он равен парному коэффициенту корреляции

Вклад независимых переменных в формирование значений у оценивается с помощью коэффициента детерминации R^2

$$R^2 = \frac{\sum_{i=1}^{n}\left(\hat{y}_i - \bar{y}\right)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2} .$$

Чем ближе значение R^2 к единице, тем лучше включенные в модель факторы описывают изменчивость у.

5. Оценка качества уравнения нелинейной регрессии

Оценка качества уравнения нелинейной регрессии состоит из следующих этапов.

1. Оценка математической точности уравнения. Для этого рассчитывается ***средняя относительная ошибка аппроксимации***

$$\overline{A} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \cdot 100\%,$$

где y_i и $\hat{y}_i$ - соответственно фактические и теоретические значения переменной y .
Для принятия решения о точности уравнения можно воспользоваться таблицей

Значение $\overline{A}$, %	**Точность уравнения**
менее 10	высокая
10 - 20	хорошая
20 - 50	удовлетворительная
более 50	неудовлетворительная

В случае, когда уравнение имеет неудовлетворительную точность, необходимо увеличить объем наблюдений (объем выборки) n, скорректировать набор объясняющих переменных x1, x2,…, xm, либо взять другое уравнение регрессии (нелинейное).

2. Проверка ***статистической значимости уравнения регрессии в целом*** с помощью ***F-критерия Фишера***.

Выдвигается гипотеза $\boldsymbol{H_0}$: уравнение регрессии статистически незначимо, при конкурирующей гипотезе $\boldsymbol{H_1}$: уравнение регрессии статистически значимо. Находится ***расчетное значение*** (***статистика***) ***критерия***

$$F_{расч} = \frac{\sum_{i=1}^{n}\left(\hat{y}_i - \overline{y}\right)^2}{\sum_{i=1}^{n}\left(\hat{y}_i - y_i\right)^2} \cdot \frac{n - m - 1}{m},$$

где y_i, $\hat{y}_i$, $\overline{y}$ - соответственно фактическое (наблюдаемое), теоретическое и среднее значение y ; $\boldsymbol{n}$ - объем выборки, $\boldsymbol{m}$ - число параметров уравнения регрессии при независимых переменных.

Табличное (***критическое***) значение $\boldsymbol{F_{табл}}$, находится по таблице критических значений распределения Фишера-Снедекора (***F-распределения***) по уровню значимости $\boldsymbol{\alpha}$ и двум числам степеней свободы $\boldsymbol{df_1} = \boldsymbol{m}$ и $\boldsymbol{df_2} = \boldsymbol{n} - \boldsymbol{m} - 1$.

Если $\boldsymbol{F_{расч}} > \boldsymbol{F_{табл}}$, то гипотеза $\boldsymbol{H_0}$ отвергается с вероятность ошибки $\boldsymbol{\alpha}$, т.е. уравнение регрессии признается в целом статистически значимым (***адекватно описывающим исходные данные***).

В противном случае ($\boldsymbol{F_{расч}} < \boldsymbol{F_{табл}}$) уравнение считается незначимым.

! **Для нелинейных регрессионных уравнений критерий Стьюдента не проверяется**

! **Для нелинейных регрессионных уравнений анализ остатков не производится. Проводится анализ остатков соответствующих линеаризованных уравнений (не входит в программу курса)**

**Если уравнение нелинейной регрессии признано качественным то регрессионная модель считается качественной,
т.е. она адекватно описывает исходные данные.**

! **Для получения качественной нелинейной регрессионной модели с m факторами необходима выборка объема не менее n = (6÷8)×(m+1).**

6. Приложения нелинейных регрессионных моделей

Наиболее часто встречающиеся приложения регрессионной модели - оценка влияния объясняющих переменных на результативный признак и построение прогноза.

1. Эластичность

Важную роль при оценке влияния факторов играют коэффициенты регрессионной модели. Однако непосредственно с их помощью нельзя сопоставлять факторные признаки по степени их влияния на зависимую переменную из-за различия единиц измерения и разной степени колеблемости. Для устранения таких различий применяются ***средние*** $\overline{Э}$ и ***частные*** $Э_i$ ***коэффициенты эластичности***.

В общем случае коэффициент эластичности для парного уравнения регрессии $\hat{y} = f(x)$ находится по формуле

$$Э = f'(x) \cdot \frac{x}{y}.$$

После дифференцирования парных нелинейных уравнений регрессии получим следующие формулы .

Формулы определения коэффициентов эластичности

Вид уравнения регрессии	**Средний коэффициент эластичности** $\overline{Э}$	**Частные коэффициенты эластичности** $Э_i$
Степенная функция $\hat{y} = b_0 \cdot x^{b1}$	$\overline{Э} = b_1$	$Э_i = b_1$
Показательная функция $\hat{y} = b_0 \cdot b_1^x$	$\overline{Э} = \overline{x} \cdot \ln b_1$	$Э_i = x_i \cdot \ln b_1$
Гипербола $\hat{y} = b_0 + b_1 \frac{1}{x}$	$\overline{Э} = -\frac{b_1}{b_0 \cdot \overline{x} + b_1}$	$Э_i = -\frac{b_1}{b_0 \cdot x_i + b_1}$
Полином второго порядка $\hat{y} = b_0 + b_1 \cdot x + b_2 \cdot x^2$	$\overline{Э} = \frac{(b_1 + 2b_2 \overline{x}) \cdot \overline{x}}{b_0 + b_1 \overline{x} + b_2 \overline{x}^2}$	$Э_i = \frac{(b_1 + 2b_2 x_i) \cdot x_i}{b_0 + b_1 x_i + b_2 x_i^2}$

Средний по совокупности наблюдений коэффициент эластичности показывает, на сколько процентов изменится среднее значение зависимой переменной $\bar{y}$ при увеличении среднего значения $\bar{x}$ на 1%. Средний коэффициент эластичности позволяет выявить общегрупповые закономерности.

Частный коэффициент эластичности показывает, на сколько процентов изменится значение зависимой переменной для i-го наблюдения (y_i) при увеличении значения фактора для i-го наблюдения (x_i) на 1%.Частные коэффициенты эластичности выявляют особенности объектов наблюдений.

2. Прогнозирование

В прогнозных расчетах предсказываемое значение $y_{пр}$ определяется как точечный прогноз $\hat{y}_{пр}$ путем подстановки в уравнение регрессии соответствующего значения факторной переменной $x_{пр}$.

Наибольшая точность прогноза достигается в тех случаях, когда $x_{jпр}$ находятся в центре области наблюдений x , используемых при построении уравнения регрессии (незначительно отличаются от $\bar{x}$). При удалении $x_{jпр}$ от $\bar{x}$ средняя ошибка прогноза растет. В случае, когда $x_{пр}$ оказывается за пределами области наблюдаемых значений, нельзя говорить о надежности $\hat{y}_{пр}$.

Тема 6. ВРЕМЕННЫЕ РЯДЫ

Лекция

План лекции

1. Этапы моделирование динамики социально-экономических процессов
2. Понятие и классификация временных рядов
3. Проверка гипотезы о наличии тренда
4. Сглаживание временных рядов методом скользящих средних
5. Построение уравнения тренда
6. Определение степени полиномиального тренда
7. Исследование структуры ряда. Автокорреляция уровней ряда
8. Моделирование тенденции ряда при наличии структурных изменений
9. Моделирование сезонных и циклических колебаний
10. Прогнозирование временных рядов

1. Этапы моделирование динамики социально-экономических процессов

Важную роль в экономике играет анализ, моделирование и прогнозирование на базе данных по одному объекту, собранных в последовательные моменты или периоды времени, т.е. на основе временных рядов.

Первым этапом такого исследования является построение ряда, отвечающего требованиям статистической науки: уровни ряда должны быть получены из надежных источников информации и сопоставимы друг с другом по качественному содержанию.

Второй этап - определение типа основной тенденции динамики (либо констатация отсутствия надежной тенденции). Методика этого этапа излагается в курсах общей теории статистики и обычно включает:
- содержательный подход - на основе закономерностей экономики, технологии, общего состояния изучаемого объекта;
- графическое изображение временного ряда и подбор подходящей линии тренда;
- математико-статистическую оценку наличия достаточно надежного абсолютного прироста уровней, ускорения этого прироста, темпов роста, наличия автокорреляции уровней.

Третий этап - вычисление уравнения тренда, т.е. уравнения такой линии, которая оптимально выражает фактическую тенденцию изменения уровней ряда.
На этом этапе применяются методы математической статистики: метод наименьших квадратов (МНК), критерии оценки надежности - Фишера, Стьюдента, Дарбина-Уотсона и др. Используются программы STATISTICA, Statgraphics, Excel и др.

Четвертый этап заключается в исследовании отклонений фактических значений уровней ряда от расчетных уровней тренда, т.е. изучении колеблемости, а также в измерении и моделировании сезонных (циклических) колебаний.

Пятым этапом является расчет прогнозируемых значений временного ряда для будущих периодов, вероятных интервалов этих прогнозов, а иногда и страхового запаса.

В данной теме более подробно рассмотрены только те вопросы моделирования временных рядов, которые не были изложены в рамках регрессионного анализа.

2. Понятие и классификация временных рядов

Развитие экономического явления или процесса во времени в статистике называется динамикой. Для отображения динамики строятся временные ряды (ряды динамики), которые представляют собой ряды изменяющихся во времени значений статистического показателя, расположенных в хронологическом порядке.

Элементами временного ряда являются показатели уровней ряда и периоды (годы, кварталы, месяцы, сутки) или моменты (даты) времени. Уровни (члены) ряда обычно обозначаются y_t , а соответствующие им моменты (периоды) времени - t .

Классифицировать различные виды временных рядов можно по следующим признакам.

1. В зависимости от способа выражения уровней временные ряды подразделяются на ***ряды абсолютных, относительных и средних величин***.

2. В зависимости от того, как выражают уровни ряда состояние экономического явления на определенные моменты времени (начало месяца, квартала, года и т.п.) или его величину за определенные интервалы времени (например, за месяц, квартал, год и т.п.) различают соответственно ***моментные и интервальные временные ряды***.

3. В зависимости от расстояния между уровнями временные ряды разделяются на ***ряды с равноотстоящими уровнями и неравноотстоящими уровнями*** во времени.

4. В зависимости от наличия основной тенденции изучаемого процесса временные ряды разделяются на ***стационарные и нестационарные (динамические)***.

Если основные характеристики случайного процесса (математическое ожидание и дисперсия) постоянны и не зависят от времени, то процесс считается стационарным и временной ряд также считается стационарным. График стационарного ряда располагается относительно прямой $y = y_{ср}$, параллельной горизонтальной оси. На рис. 1 среднее значение ряда $\overline{y}_t = 4$.

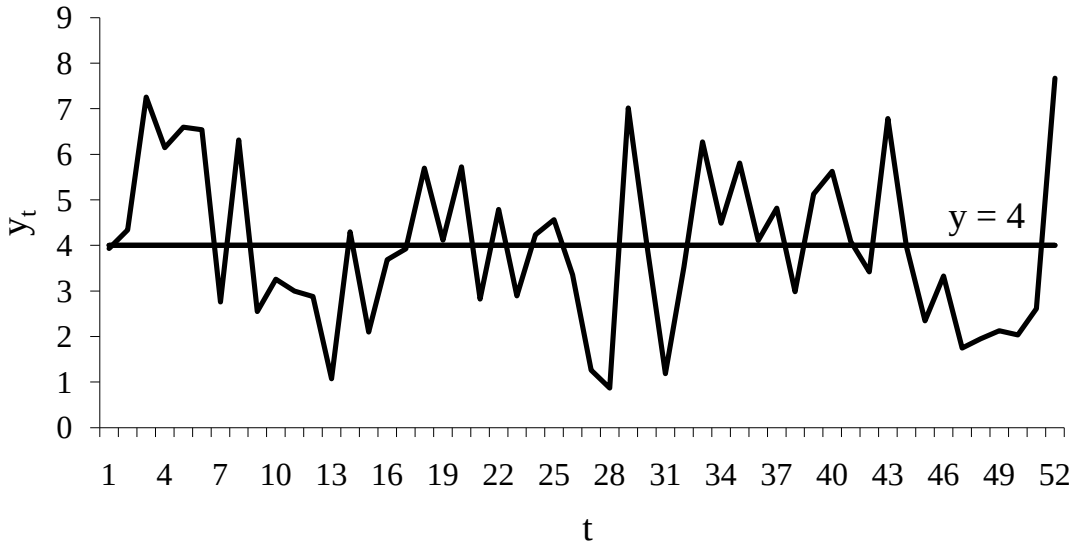


Рис. 1. График стационарного временного ряда

Экономические процессы во времени обычно не являются стационарными, так как содержат основную тенденцию развития, и описываются динамическими временными рядами. На рис. 2 график динамического ряда располагается относительно наклонной прямой (тренда), характеризующей тенденцию.

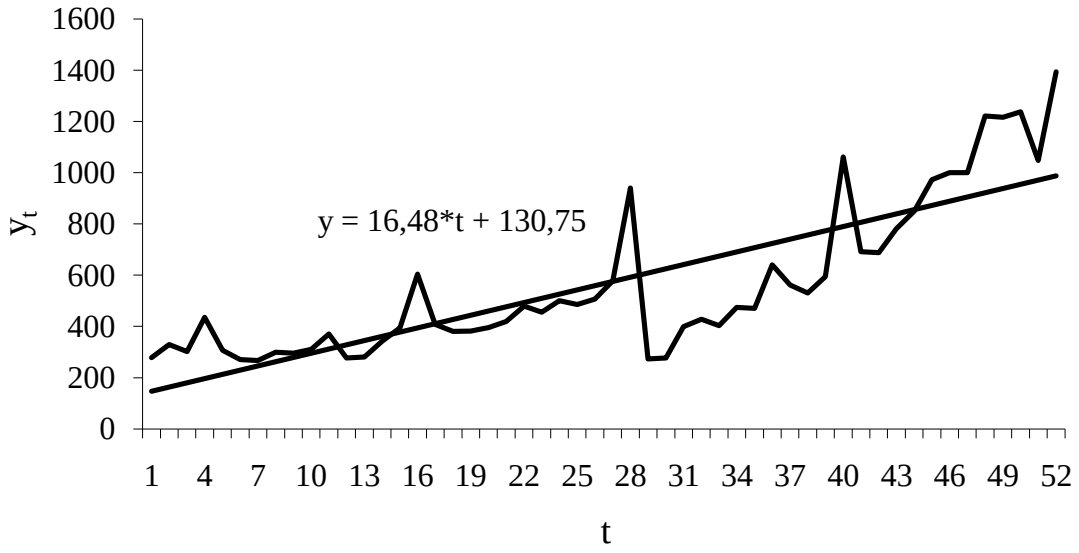


Рис. 2. График динамического временного ряда

5. По числу показателей выделяют ***изолированные (одномерные) и компонентные (многомерные) временные ряды***. Если проводится анализ во времени одного показателя, то временной ряд изолированный. Например, объем произведенных предприятием товаров (помесячно) представляет собой одномерный временной ряд (см. табл. 1).

Таблица 1

Одномерный временной ряд

Момент времени (дата)	Январь 2004 г.	Февраль 2004 г.	Март 2004 г.	Апрель 2004 г.	…	…	…
Номер момента времени *t*	1	2	3	4	…	n - 1	n
Значение показателя y_t	y_1	y_2	y_3	y_4	…	y_{n-1}	y_n

В многомерном ряду представлена динамика нескольких показателей, характеризующих одно экономическое явление (процесс). Например, характеристики деятельности предприятия, наблюдаемые помесячно, можно представить в виде многомерного ряда (см. табл. 2). В качестве показателей производственной деятельности можно взять: Y - производительность труда, X_1 - удельный вес покупных изделий, X_2 - премии и вознаграждения на одного работника, X_3 - среднегодовая численность ППП, X_4 - среднегодовой фонд заработной платы ППП, X5 - непроизводственные расходы.

Таблица 2

Многомерный временной ряд

Момент времени (дата)	**Номер момента времени *t***	**Значения показателей**				
		Y	**X_1**	**X_2**	**…**	**X_5**
Январь 2004 г.	1	y_1	x_{11}	x_{12}	…	x_{15}
Февраль 2004 г.	2	y_2	x_{21}	x_{22}	…	x_{25}
Март 2004 г.	3	y_3	x_{31}	x_{32}	…	x_{35}
…	…	…	…	…	…	…
…	n	y_n	x_{n1}	x_{n2}	…	x_{n5}

В данной лекции рассмотрены вопросы теории абсолютных моментных одномерных временных рядов с равноотстоящими уровнями, которые наиболее часто встречаются в экономике.

3. Проверка гипотезы о наличии тренда

Исследование временного ряда начинается с решения вопроса о его стационарности. Одним из возможных вариантов для этого служит проверка гипотезы $\boldsymbol{H}_0$ о случайности (отсутствии) временного тренда при конкурирующей гипотезе $\boldsymbol{H}_1$ о неслучайности (наличии) временного тренда, основанная на сравнении средних значений первой и второй половины ряда по статистике (расчетному значению критерия)

$$t_{расч} = \frac{\overline{y}_n - \overline{y}_в}{\sqrt{(n_n - 1)s_n^2 + (n_в - 1)s_в^2}} \cdot \sqrt{\frac{n_n \cdot n_в \cdot (n_n + n_в - 2)}{n_n + n_в}}, \qquad (1)$$

где $\overline{y}_n$, $\overline{y}_в$ - средние значения первой и второй половины ряда, имеющих длины $\boldsymbol{n}_n$, $\boldsymbol{n}_в$ ($\boldsymbol{n}_n + \boldsymbol{n}_в = \boldsymbol{n}$); $\boldsymbol{s}_n^2$, $\boldsymbol{s}_в^2$ – дисперсии первой и второй половины ряда.

Средние значения и дисперсии определяются по формулам

$$\overline{y}_n = \frac{1}{n_n}\sum_{t=1}^{n_n} y_t \quad \text{и} \quad \overline{y}_в = \frac{1}{n_в}\sum_{t=n_{n1}+1}^{n} y_t, \qquad (2)$$

$$s_n^2 = \frac{1}{n_n - 1}\sum_{t=1}^{n_n} (y_t - \overline{y}_n)^2 \quad \text{и} \quad s_в^2 = \frac{1}{n_в - 1}\sum_{t=n_n+1}^{n} (y_t - \overline{y}_в)^2. \qquad (3)$$

Значение $|\boldsymbol{t}_{расч}|$ сравнивается с критическим значением распределения Стьюдента $\boldsymbol{t}_{кр}(\boldsymbol{\alpha}, \boldsymbol{k})$ с $\boldsymbol{k} = \boldsymbol{n}_n + \boldsymbol{n}_в - 2$ степенями свободы и уровнем значимости $\boldsymbol{\alpha}$. В случае, если $|\boldsymbol{t}_{расч}| < \boldsymbol{t}_{кр}$ гипотеза $\boldsymbol{H}_0$ о случайности временного тренда принимается и ряд считается стационарным. В противном случае ($|\boldsymbol{t}_{расч}| > \boldsymbol{t}_{кр}$) - гипотеза $\boldsymbol{H}_0$ отвергается, что свидетельствует о значимости различия средних первой и второй половины ряда и неслучайности (наличии) временного тренда. Другими словами, ряд является динамическим.

Для проверки гипотезы о наличии тренда также можно воспользоваться критерием серий, методом Фостера-Стюарта и др. [1, 13].

Каждый член динамического ряда можно представить в виде

$$y_t = \hat{y}_t + \varepsilon_t., \qquad (4)$$

где $\hat{y}_t$ - теоретические значения уровней, вычисленные по уравнению тренда, ε_t - остатки (отклонения фактических уровней ряда от теоретических). Выражение (4) носит название ***модели временного ряда.***

Моделирование временного ряда заключается в анализе его структуры; подборе, построении и проверке качества уравнения тренда; анализе остатков.

Построение тренда в теории временных рядов называется ***сглаживанием***. Существующие методы выделения тренда можно разделить на две группы: методы численного сглаживания, когда тренд задается численными значениями сглаженных уровней, вычисленных по значениям уровней исходного ряда (метод скользящих средних, экспоненциальное сглаживание и др.), а также методы аналитического выравнивания (построение уравнения тренда).

4. Сглаживание временных рядов методом скользящих средних

Сглаживание временного ряда выполняется для выделения тренда по выбранному числу членов ряда ***m***. Для этого используется метод наименьших квадратов, с помощью которого по ***m*** членам ряда строятся полиномы выбранной степени ***p*** , начиная с первого и т.д. членов ряда. Степень полинома и число точек сглаживания выбираются из общих соображений, включая существо решаемой задачи, и подбора степени по пробным кратковременным прогнозам.

Новое сглаженное значение временного ряда в средней точке из ***m*** заданных находится как линейная комбинация старых ***m*** значений ряда с коэффициентами, зависящими от степени полинома.

Если сглаживание ряда осуществляется по ***m*** = 5 точкам, то для ***p*** = 1 (линейное сглаживание) новое сглаженное значение уровня ряда вычисляется по формуле

$$\tilde{y}_t = \frac{1}{5}\left(y_{t-2} + y_{t-1} + y_t + y_{t+1} + y_{t+2}\right), \qquad (5)$$

где y_t , $\tilde{y}_t$ - заданное и новое сглаженное значения уровня ряда (***t*** = 3, 4, …, ***n*** - 2).

Для ***p*** = 2, (квадратичное сглаживание) новое сглаженное значение уровня ряда вычисляется по формуле

$$\tilde{y}_t = \frac{1}{35}\left(-3y_{t-2} + 12y_{t-1} + 17y_t + 12y_{t+1} - 3y_{t+2}\right). \qquad (6)$$

Для вычисления сглаженных первых и последних (***m*** – 1)/2 значений ряда (при ***m*** = 5 вычисляются два первых и два последних члена) используются следующие формулы:

при ***p*** = 1

$$\begin{aligned}
\tilde{y}_1 &= \frac{1}{5}\left(3y_1 + 2y_2 + y_3 - y_5\right),\\
\tilde{y}_2 &= \frac{1}{10}\left(4y_1 + 3y_2 + 2y_3 + y_5\right),\\
\tilde{y}_{n-1} &= \frac{1}{10}\left(4y_n + 3y_{n-1} + 2y_{n-2} + y_{n-4}\right),\\
\tilde{y}_n &= \frac{1}{5}\left(3y_n + 2y_{n-1} + y_{n-2} - y_{n-4}\right);
\end{aligned} \qquad (7)$$

при ***p*** = 2

$$\begin{aligned}
\tilde{y}_1 &= \frac{1}{35}\left(31y_1 + 9y_2 - 3y_3 - 5y_4 + 3y_5\right),\\
\tilde{y}_2 &= \frac{1}{35}\left(9y_1 + 13y_2 + 12y_3 + 6y_4 - 5y_5\right),\\
\tilde{y}_{n-1} &= \frac{1}{35}\left(9y_n + 13y_{n-1} + 12y_{n-2} + 6y_{n-3} - 5y_{n-4}\right),\\
\tilde{y}_n &= \frac{1}{35}\left(31y_n + 9y_{n-1} - 3y_{n-2} - 5y_{n-3} + 3y_{n-4}\right).
\end{aligned} \qquad (8)$$

Процедура численного сглаживания может применяться последовательно несколько раз. После вычисления сглаженных значений ряда строится его графическое изображение (рис. 3).

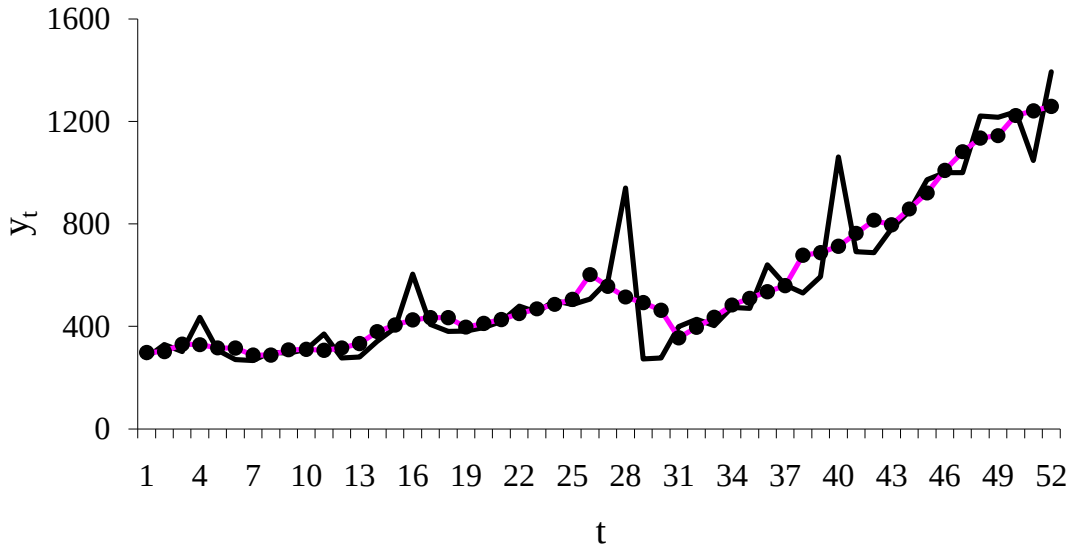


Рис. 3. Графики исходного и сглаженного ряда

Аналогичным способом производится сглаживание ряда по любому нечетному числу ***m*** членов ряда. В случае отсутствия необходимых формул, вычисление первых и последних членов сглаженного ряда не производится и в соответствующих строках расчетной таблицы ставится прочерк.

При сглаживании по четному числу членов ряда сначала вычисляются средние значения m уровней ряда, которые затем центрируются, т.е. в качестве сглаженных значений принимаются средние значения двух рядом стоящих средних.

В общем случае (линейное сглаживание, p = 1) расчет сглаженных значений $\tilde{y}_t$ в случае нечетной длины интервала сглаживания m = 2g + 1 осуществляется по формуле

$$\tilde{y}_t = \frac{y_{t-g} + \ldots + y_{t-1} + y_t + y_{t+1} + \ldots + y_{t+g}}{2g+1}, \; t = g+1,\ldots, n-g \qquad (9)$$

(для m = 3 значение g = 1, для m = 5 значение g = 2).

Расчет сглаженных значений $\tilde{y}_t$ в случае четной длины интервала сглаживания m = 2g производится по формуле

$$\tilde{y}_t = \frac{\frac{1}{2}y_{t-g} + \ldots + y_{t-1} + y_t + y_{t+1} + \ldots + \frac{1}{2}y_{t+g}}{2g+1}, \; t = g+1,\ldots, n-g \qquad (10)$$

(для m = 4 значение g = 2).

Расчет сглаженных значений для незаполненных уровней ряда, в котором отсутствует цикличность, можно произвести на основании средних абсолютных приростов. Сглаженные значения в начале временного ряда рассчитываются путем последовательного вычитания среднего прироста

$$\Delta y_{нач} = \frac{y_{2g+1} - y_1}{2g}$$

на первом активном участке из первого доступного сглаженного значения. Сглаженные значения в конце временного ряда рассчитываются путем последовательного прибавления среднего прироста

$$\Delta y_{кон} = \frac{y_n - y_{n-2g}}{2g}$$

на последнем активном участке к последнему доступному сглаженному значению. Под первым и последним активными интервалами понимаются интервалы, значения которых используются при расчетах первого и последнего усредненного значения.

В случае m = 3 (g = 1) первый активный участок включает в себя три первых уровней ряда, средний прирост $\Delta y_{нач} = \frac{y_3 - y_1}{2}$, последний активный участок - три последних уровня ряда, $\Delta y_{кон} = \frac{y_n - y_{n-2}}{2}$.

В случае m = 5 (g = 2) первый активный участок включает в себя пять первых уровней ряда, средний прирост $\Delta y_{нач} = \frac{y_5 - y_1}{2 \cdot 2}$, а последний активный участок – пять последних уровней ряда и $\Delta y_{кон} = \frac{y_n - y_{n-4}}{2 \cdot 2}$.

Поскольку расчет сглаженных значений для m = 5 и m = 4 производится по значениям одних и тех же интервалов, то средние приросты для этих значений параметра m также одинаковы.

В случае m = 3 необходимо рассчитать сглаженные значения только для первого и последнего уровней ряда:

$$\tilde{y}_1 = \tilde{y}_2 - \Delta y_{нач}\ ;\ \tilde{y}_n = \tilde{y}_{n-1} + \Delta y_{кон}\ . \qquad (11)$$

В случае m = 5 необходимо рассчитать сглаженные значения для двух первых и двух последних уровней ряда:

$$\tilde{y}_2 = \tilde{y}_3 - \Delta y_{нач}\ ;\ \tilde{y}_1 = \tilde{y}_2 - \Delta y_{нач}; \qquad (12)$$

$$\tilde{y}_{n-1} = \tilde{y}_{n-2} + \Delta y_{кон}\ ;\ \tilde{y}_n = \tilde{y}_{n-1} + \Delta y_{кон}\ . \qquad (13)$$

Аналогично в случае m = 4 необходимо рассчитать сглаженные значения для двух первых и двух последних уровней ряда по формулам (12, 13).

5. Построение уравнения тренда

В статистике построение аналитической функции (уравнения тренда) для моделирования тенденции временного ряда называют ***аналитическим выравниванием* (*сглаживанием*) *временного ряда***. Для этого чаще всего применяются следующие функции:

1. полиномиальная $\hat{y}_t = b_0 + b_1 \cdot t + b_2 \cdot t^2 + \ldots + b_k \cdot t^k$ (***k*** - степень полинома);

2. линейная $\hat{y}_t = b_0 + b_1 \cdot t$ (частный случай полиномиальной при ***k*** = 1), используется для описания процессов, развитие которых протекает во времени равномерно. Параметр b_0 интерпретируется как параметр начальных условий (значение переменной y_t в нулевой момент времени $t = 0$), b_1 - как скорость роста переменной y_t;

3. параболическая $\hat{y}_t = b_0 + b_1 \cdot t + b_2 \cdot t^2$ (частный случай полиномиальной при ***k*** = 2), используется для описания процессов, развитие которых характеризуется равноускоренным ростом (снижением). Параметр b_0 интерпретируется как параметр начальных условий (значение переменной y_t в нулевой момент времени $t = 0$), b_1 - как скорость роста и b_2 - как ускорение роста переменной y_t;

4. показательная $\hat{y}_t = b_0 \cdot b_1^t$;

5. экспоненциальная $\hat{y}_t = b_0 \cdot e^{b1t}$;

6. степенная $\hat{y}_t = b_0 \cdot t^{b1}$;

7. гиперболическая $\hat{y}_t = b_0 + \frac{b_1}{t}$.

Параметры уравнений трендов (уравнение тренда представляет собой уравнение регрессии), как правило, определяются методом наименьших квадратов. В качестве независимой переменной выступает время ***t*** = 1, 2, …, ***n*** , а в качестве зависимой переменной - фактические уровни временного ряда y_t . Критериями отбора наилучшей формы тренда являются наибольшее значение скорректированного коэффициента детерминации $R^2_{скорр}$ и наименьшее значение средней относительной ошибки аппроксимации $\overline{A}$.

На практике при выборе формы тренда обычно используют положения и выводы экономической теории, визуальный анализ графика ряда, а также результаты исследования структуры ряда (автокорреляция уровней ряда (см. пункт 6), определение степени полиномиального тренда (см. пункт 5)). На рис. 2 изображен линейный тренд, на рис. 4 - полиномиальный тренд.

Оценка качества уравнения тренда производится аналогично оценке качества уравнения регрессии с помощью средней относительной ошибки аппроксимации, критериев Фишера, Стьюдента и Дарбина-Уотсона.

6. Определение степени полиномиального тренда

В случае, когда исследуемый экономический процесс носит колебательный характер, его динамику нужно описывать полиномиальным трендом (на рис. 4 динамика ряда описывается полиномом пятой степени).

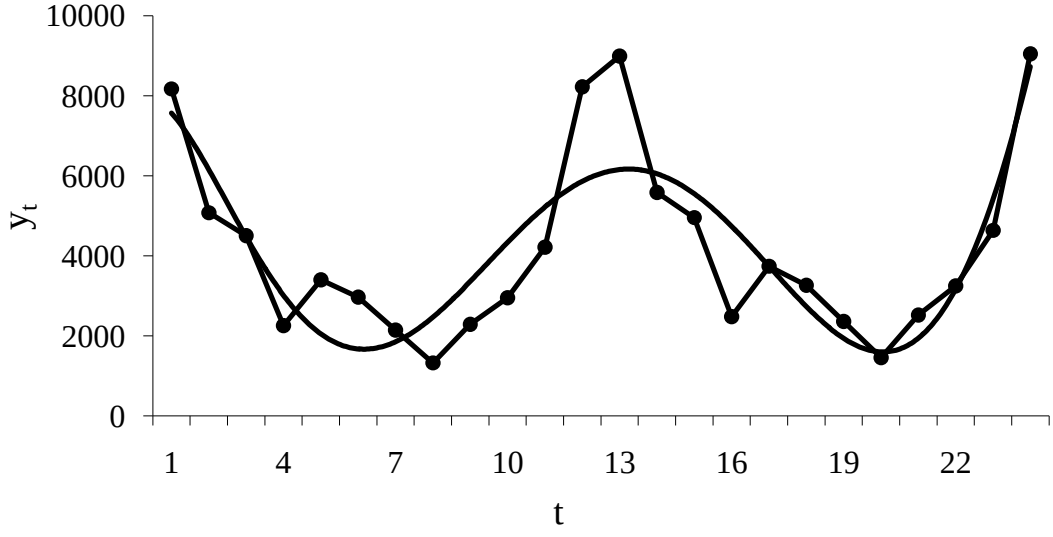


Рис. 4. Графики ряда и полиномиального тренда

Для аппроксимации ряда полиномом степени ***p*** предварительно находится значение этой степени по следующей процедуре, которая аналогична дифференцированию полинома. Очевидно, что вторая производная полинома первой степени, третья производная полинома второй степени и т.д. равны нулю. В случае временного ряда

операция дифференцирования заменяется вычислением переменных разностей, а условие равенства нулю - проверкой гипотезы о равенстве дисперсий предыдущих и последующих разностей.

Сначала вычисляются первые разности

$$\Delta^1 y_t = y_{t+1} - y_t ,$$

где $t = 1, 2, \ldots, n - 1$.

Затем по первым разностям вычисляются вторые разности

$$\Delta^2 y_t = \Delta^1 y_{t+1} - \Delta^1 y_t ,$$

где $t = 1, 2, \ldots, n - 2$.

И далее последовательно разности 3-го, …, k-го порядков

$$\Delta^k y_t = \Delta^{k-1} y_{t+1} - \Delta^{k-1} y_t ,$$

где $t = 1, 2, \ldots, n - k$.

Под разностями нулевого порядка понимается сам временной ряд.

На каждом шаге, начиная с $k = 0$, вычисляются дисперсии разностей k-го порядка по формуле

$$S_k^2 = \frac{\sum_{t=1}^{n-k} \left(\Delta^k y_t - \overline{\Delta^k y_t} \right)^2}{(n-k-1) \cdot (k!)^2} . \qquad (14)$$

При $k = 0$ по формуле (14) вычисляется дисперсия заданного временного ряда y_t .

На каждом шаге с помощью критерия Фишера проверяется гипотеза о равенстве предыдущей и последующей дисперсий.

Для этого вычисляется расчетное значение критерия

$$F^k_{расч} = \frac{S^2_{k-1}}{S^2_k} .$$

Критическое значение критерия $F^k_{крит}(\alpha, k_1, k_2)$ находится в таблице критических значений распределения Фишера по уровню значимости α и числам степеней свободы $k_1 = n - k$, $k_2 = n - k - 1$.

Если $F^k_{расч} > F^k_{крит}$ дисперсии отличаются значимо. В этом случае процедура вычислений дисперсий и их разностей продолжается. Доказано, что последовательность дисперсий (14) убывает с ростом k, и при некотором значении $p = k - 1$ выполнится неравенство $F^k_{расч} < F^k_{крит}$ (различие дисперсий становится незначимым). Полученное значение p и является степенью полиномиального тренда. Дисперсия s_k^2 называется дисперсией случайностей, а разности порядка p являются случайной компонентой временного ряда.

7. Исследование структуры ряда. Автокорреляция уровней ряда

При наличии во временном ряде тенденции и циклических колебаний значение каждого последующего уровня ряда зависят от предыдущих. В этом случае говорят, что ряд имеет автокорреляцию.

Автокорреляцией называется корреляция уровней временного ряда друг с другом, со сдвигом во времени на l тактов (лагом l). Количественно ее можно измерить с помощью парного коэффициента корреляции между уровнями исходного временного ряда и уровнями этого ряда, сдвинутыми на l шагов во времени.

Коэффициент автокорреляции первого порядка (лаг $l = 1$) равен

$$r_1 = \frac{\sum_{t=2}^{n}\left(y_t - \overline{y}_t\right)\left(y_{t-1} - \overline{y}_{t-1}\right)}{\sqrt{\sum_{t=2}^{n}\left(y_t - \overline{y}_t\right)^2 \cdot \sum_{t=2}^{n}\left(y_{t-1} - \overline{y}_{t-1}\right)^2}}, \qquad (15)$$

где

$$\overline{y}_t = \frac{1}{n-1}\sum_{t=2}^{n} y_t ; \qquad \overline{y}_{t-1} = \frac{1}{n-1}\sum_{t=2}^{n} y_{t-1} .$$

Коэффициент автокорреляции ***l***-го порядка находится по формуле

$$r_l = \frac{\sum_{t=l+1}^{n}\left(y_t - \overline{y}_t\right)\left(y_{t-l} - \overline{y}_{t-l}\right)}{\sqrt{\sum_{t=l+1}^{n}\left(y_t - \overline{y}_t\right)^2 \cdot \sum_{t=l+1}^{n}\left(y_{t-l} - \overline{y}_{t-l}\right)^2}}, \qquad (16)$$

где

$$\overline{y}_t = \frac{1}{n-l}\sum_{t=l+1}^{n} y_t ; \qquad \overline{y}_{t-l} = \frac{1}{n-l}\sum_{t=l+1}^{n} y_{t-l} .$$

После вычисления коэффициентов автокорреляции необходимо проверить их статистическую значимость сравнением с критическими значениями коэффициента корреляции $r^l_{крит}(\alpha, k)$. Критические значения берутся из таблицы критических значений корреляции по уровню значимости α и числу степеней свободы $k = n - l - 2$. Если $|r_l| < r^l_{крит}$ то коэффициент корреляции r_l статистически незначим и выводы, сделанные по его значению, имеют вероятность ошибки, равную $1 - \alpha$.

Последовательность коэффициентов автокорреляции называют ***автокорреляционной функцией*** временного ряда. График зависимости ее значений от величины лага (порядка коэффициента автокорреляции) называется ***коррелограммой***. Поскольку знаки коэффициентов автокорреляции при анализе не учитываются, коррелограмма обычно строится по их абсолютным значениям (см. рис. 5).

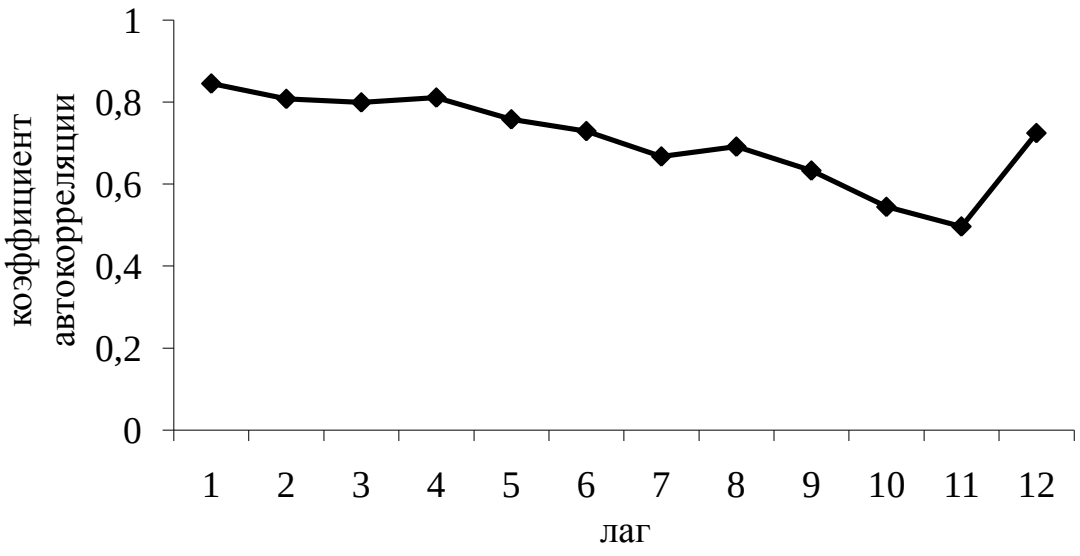


Рис. 5. Коррелограмма

Анализ автокорреляционной функции и коррелограммы позволяет определить лаг, при котором связь между текущими и предыдущими уровнями ряда наиболее тесная (сильная).

Если абсолютное значение коэффициента автокорреляции первого порядка $|r_1| > 0{,}7$, ряд содержит линейную тенденцию, если $|r_1| < 0{,}7$ - ряд содержит нелинейную тенденцию.

В случае, когда наибольшее абсолютное значение имеет коэффициент автокорреляции порядка $l = \tau$ и при этом, $| r_\tau | > 0{,}7$, ряд содержит циклические колебания с периодом в τ моментов времени.

Возникают ситуации, когда $| r_l | > 0{,}7$ и $l \geq 2$, но сущность изучаемого процесса, а также вид графика ряда не позволяют сделать вывод о наличии цикличности. В таких случаях динамика описывается авторегрессионным уравнением либо уравнением с распределенным лагом.

8. Моделирование тенденции ряда при наличии структурных изменений

При протекании экономических процессов иногда возникает ситуация единовременного изменения характера динамики, вызванная структурными изменениями в экономике или иными факторами. В этом случае, начиная с некоторого момента t^* , происходит изменение тенденции временного ряда, т.е. изменение параметров тренда, описывающего эту динамику, а иногда и самого вида уравнения тренда. Момент времени t^* обычно называют точкой смены тенденции. Например, на рис. 6 изображен график ряда, в котором произошла смена тенденции в момент времени, близкий к $t^* = 32$.

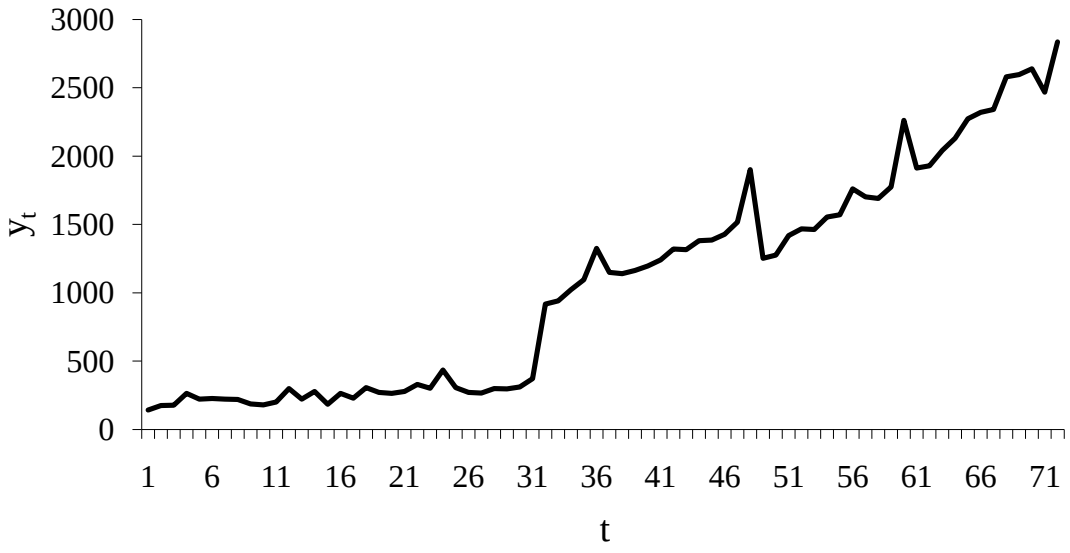


Рис. 6. Смена тенденции временного ряда.

Основная задача исследования временного ряда, включающего точку смены тенденции, - выяснить, насколько значимо повлияли структурные изменения на характер тенденции. Если это влияние значимо, для моделирования тенденции необходимо разделить ряд на две части (до момента времени t^* и после момента t^*) и построить отдельно по каждой части ряда уравнения тренда (кусочно-линейная модель). Если изменения незначительно повлияли на характер тенденции, то ее можно описать единым для всего ряда уравнением.

Значимость структурных изменений можно оценить с помощью статистического критерия Грегори Чоу. Система обозначений для его проверки приведена в табл. 3.

Таблица 3

№ уравнения	Вид уравнения тренда*	Длина ряда	Остаточная сумма квадратов	Число параметров в уравнении*	Число степеней свободы остаточной дисперсии
Кусочно-линейная модель тренда					
(I)	$y^{(1)} = a_1 + b_1 \cdot t$	n_1	$S^1_{\text{ост}}$	k_1	$n_1 - k_1$
(II)	$y^{(2)} = a_2 + b_2 \cdot t$	n_2	$S^2_{\text{ост}}$	k_2	$n_2 - k_2$
Уравнение тренда по всему ряду					
(III)	$y^{(3)} = a_3 + b_3 \cdot t$	$n = n_1 + n_2$	$S^3_{\text{ост}}$	k_3	$n - k_3 = (n_1 + n_2) - k_3$
* В рассматриваемом примере все уравнения тренда линейные, число параметров всех уравнений $k_1 = k_2 = k_3 = 2$. В общем случае вид уравнений и число параметров в каждом уравнении могут различаться.					

Выдвинем гипотезу $\boldsymbol{H}_0$ о структурной стабильности тенденции изучаемого временного ряда $\boldsymbol{y_t}$.

Найдем остаточные суммы квадратов

$$S^1_{ост} = \sum_{t=1}^{n1} \left(y_t - \hat{y}_t^{(1)} \right)^2 , \qquad (17)$$

$$S^2_{ост} = \sum_{t=n1+1}^{n} \left(y_t - \hat{y}_t^{(2)} \right)^2 , \qquad (18)$$

$$S^3_{ост} = \sum_{t=1}^{n} \left(y_t - \hat{y}_t^{(3)} \right)^2 , \qquad (19)$$

где $\hat{y}_t^{(1)}$, $\hat{y}_t^{(2)}$, $\hat{y}_t^{(3)}$ - теоретические значения уровней, найденные соответственно по уравнениям (I), (II) и (III).

Остаточная сумма квадратов кусочно-линейной модели равна

$$S^{кл}_{ост} = S^1_{ост} + S^2_{ост} .$$

Соответствующее ей число степеней свободы составит

$$(n_1 - k_1) + (n_2 - k_2) = n - k_1 - k_2 .$$

Изменение остаточной дисперсии при переходе от единого уравнения тренда к кусочно-линейной модели определяется как разность

$$\Delta S_{ост} = S^3_{ост} - S^{кл}_{ост} ,$$

соответствующее ему число степеней свободы равно

$$n - k_3 - (n - k_1 - k_2) = k_1 + k_2 - k_3 .$$

Расчетное значение $\boldsymbol{F}$-критерия

$$F_{расч} = \frac{\Delta S_{ост}}{S^{кл}_{ост}} \cdot \frac{n - k_1 - k_2}{k_1 + k_2 - k_3} . \qquad (20)$$

сравнивается с табличным $\boldsymbol{F}_{табл}(\boldsymbol{\alpha}, \boldsymbol{df}_1, \boldsymbol{df}_2)$, найденным по таблице критических точек распределения Фишера для уровня значимости $\boldsymbol{\alpha}$ и числа степеней свободы $\boldsymbol{df}_1 = \boldsymbol{k}_1 + \boldsymbol{k}_2 - \boldsymbol{k}_3$, $\boldsymbol{df}_2 = \boldsymbol{n} - \boldsymbol{k}_1 - \boldsymbol{k}_2$.

Если $\boldsymbol{F}_{расч} > \boldsymbol{F}_{табл}$, то гипотеза $\boldsymbol{H}_0$ о структурной стабильности отклоняется, и влияние структурных изменений на динамику изучаемого показателя считается значимым. В этом случае моделирование тенденции временного ряда необходимо про-

водить с помощью кусочно-линейной модели. В противном случае, когда $F_{расч} < F_{табл}$, нет оснований отклонять гипотезу H_0, и моделирование тенденции следует осуществлять с помощью единого для всего ряда уравнения тренда.

9. Моделирование сезонных и циклических колебаний

Сезонными колебаниями называют изменения уровней ряда, связанные со сменой времени года или с регулярно повторяющимися из года в год событиями, например, связь изменения температуры воздуха с потребительским спросом, объемом товарооборота, энергопотреблением и др. ***Циклические колебания*** уровней обусловлены социальными, юридическими, экономическими, технологическими факторами (изменение тарифов, повышение заработной платы и пенсий и т.п.).

Сезонные колебания, как правило, имеют характер плавных циклов без скачкообразных изменений уровней. Циклические колебания могут иметь резкие скачки уровней, несколько максимумов и минимумов за год.

Наиболее простым подходом к анализу временных рядов, содержащих сезонные или циклические колебания, является расчет значений сезонной компоненты методом скользящей средней и построение модели временного ряда.

Если амплитуда колебаний уровней приблизительно постоянна (см. рис. 7), строят ***аддитивную*** модель, в которой значения сезонной компоненты полагаются постоянными для различных циклов,

$$y_t = \hat{y}_t + s_t + \varepsilon_t \ . \qquad (21)$$

Здесь и далее $\hat{y}_t$ - трендовая, s_t - циклическая (сезонная), ε_t - случайная составляющие (компоненты) уровней ряда.

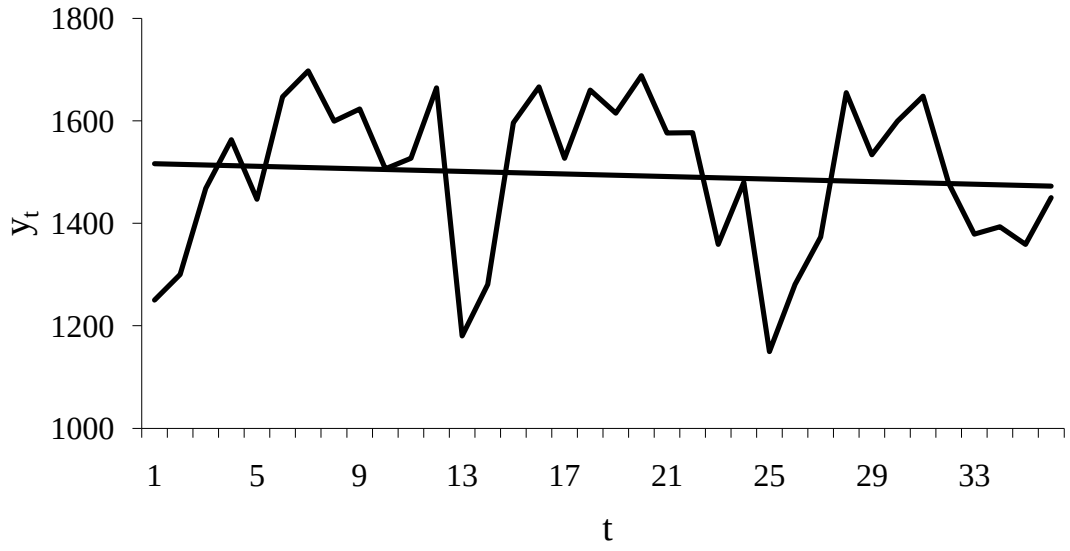


Рис. 7. Аддитивные циклические колебания

Если амплитуда сезонных колебаний возрастает (уменьшается) (см. рис. 8), строят ***мультипликативную*** модель

$$y_t = \hat{y}_t \cdot s_t \cdot \varepsilon_t \ . \qquad (22)$$

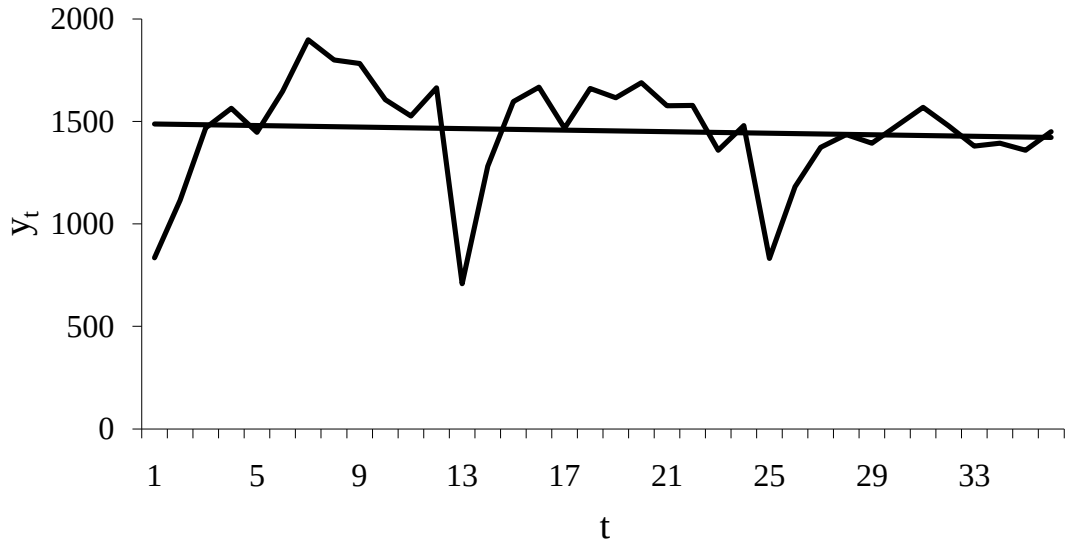


Рис. 8. Мультипликативные циклические колебания

Построение этих моделей сводится к расчету значений $\hat{y}_t$, s_t, ε_t для каждого уровня ряда и производится в следующем порядке.

1. Выравнивание исходного ряда методом скользящей средней.

2. Расчет значений сезонной компоненты s_t. Необходимо помнить, что в аддитивной модели сумма скорректированных сезонных компонент внутри цикла равна нулю. В мультипликативной модели произведение скорректированных сезонных компонент равно единице.

3. Устранение сезонной компоненты s_t из исходных уровней ряда и получение выравненных данных ($\hat{y}_t + e_t$) или ($\hat{y}_t \cdot e_t$).

4. Аналитическое выравнивание уровней ($\hat{y}_t + E$) или ($\hat{y}_t \cdot E$) и расчет теоретических значений $\hat{y}_t$ по полученному уравнению тренда.

5. Расчет полученных по модели значений ($\hat{y}_t + s_t$) или ($\hat{y}_t \cdot s_t$).

6. Расчет случайных составляющих (ошибок) e_t.

10. Прогнозирование временных рядов

Исследование динамики экономических явлений и процессов, выявление и характеристика основного тренда развития и моделей взаимосвязи дают основания для прогнозирования, т.е. определения будущих размеров уровней показателей, описывающих процессы.

Прогнозирование основывается на предположении, что закономерность развития, действующая в прошлом внутри ряда динамики, сохранится и в прогнозируемом будущем. Такой прогноз основан на перспективной интерполяции.

Теоретической основой распространения тенденции является инертность социально-экономических процессов. Инертность позволяет выявить взаимосвязь между уровнями динамического ряда или между группой связных временных рядов. Надежные результаты прогнозирования временных рядов получают, если уровни ряда динамики сопоставимы и синтезированы на основе единого методологического подхода.

Применение перспективной экстраполяции в практическом прогнозировании основывается на следующих предпосылках:

1) тенденция развития изучаемого явления графически описывается плавной линией;
2) общая тенденция развития явления в будущем существенно не меняется.

Надежность прогноза зависит от того, как точны эти предположения в действительности, а также как точно охарактеризована выявленная закономерность.

Экстраполяция - это начальная стадия построения прогноза. Механическое, без учета условий, предпосылок и содержательного экономического анализа, применение экстраполяции может стать причиной неадекватности выводов.

Чем шире временной горизонт прогноза (чем более долгосрочен прогноз), тем очевиднее недостаточность простого метода экстраполяции в результате изменения тенденции, влияния новых факторов и т.д. В этом случае динамичность экономических процессов противоречит инертности их развития.

Так как исследуемые временные ряды часто имеют недостаточно большую длину ***n***, то временной горизонт прогнозирования ограничен. Поэтому, чем короче срок прогнозирования (период упреждения), тем более надежны результаты прогноза. За относительно короткий период условия развития процесса не успевают измениться, что сохраняет характер его динамики.

В зависимости от принципов построения и эмпирических данных ряда выделяются следующие методы прогнозирования:

1) по среднему абсолютному приросту;
2) по среднему темпу роста;
3) на основе численного сглаживания временного ряда;
4) на основе аналитического выравнивания временного ряда.

Прогнозирование по ***среднему абсолютному приросту*** может быть выполнено в случае линейной тенденции развития. Этот метод основывается на предположении о стабильности (равномерности) изменения уровней ряда.

Для прогнозирования по среднему абсолютному приросту необходимо определить средний абсолютный прирост и последовательно увеличивать конечный уровень ряда на его величину на требуемое число периодов:

$$\hat{y}_{n+T} = y_n + \overline{\Delta} \cdot T \quad , \qquad (23)$$

где $\hat{y}_{n+T}$ - прогнозируемый уровень ряда; ***T*** - срок прогноза (период упреждения); y_n - последний уровень ряда, за который рассчитан средний абсолютный прирост $\overline{\Delta}$.

Следует иметь в виду, что использование среднего абсолютного прироста для прогноза возможно только при условии

$$S_{ост}^2 \le p^2 \quad ,$$

где $$p^2 = \frac{1}{2n}\sum_{t=2}^{n}\left(y_t - y_{t-1}\right)^2 ,$$

$$S_{ост}^2 = \frac{1}{n}\sum_{t=1}^{n}\left(y_t - \hat{y}_{n+T}\right)^2 .$$

Прогнозирование по ***среднему темпу роста*** осуществляется, когда общая тенденция ряда динамики характеризуется показательной (экспоненциальной) кривой. В этом случае для прогнозирования определяют средний коэффициент роста $\overline{K}_p$ и возводят его в степень, соответствующую периоду прогнозирования:

$$\hat{y}_{n+T} = y_n \cdot \overline{K}_p^T \ . \qquad (24)$$

Если временному ряду соответствует другая закономерность развития, то прогнозные значения, полученные по среднему темпу роста, будут отличаться от рассчитанных другими способами прогнозирования.

Рассмотренные два способа прогнозирования являются самыми простейшими и приближенными.

Хорошие результаты при краткосрочном прогнозировании значений временного ряда на один шаг вперед по времени (для $T = 1$) дает применение ***численного сглаживания***, например, метода скользящих средних по m последним уровням ряда. Для $m = 5$ можно воспользоваться формулой

$$\hat{y}_{n+1} = \frac{1}{10}\left(y_n + 5y_{n-1} + 2y_{n-2} - y_{n-3} - 4y_{n-4}\right). \qquad (25)$$

На практике наибольшее распространение получил метод прогнозирования на основе ***аналитического выравнивания*** временного ряда (на основе уравнения тренда). При этом для получения прогнозного значения продолжают значения независимой переменной времени за границы исследуемого периода. На рис. 9 прогнозная часть графика ряда изображена пунктиром.

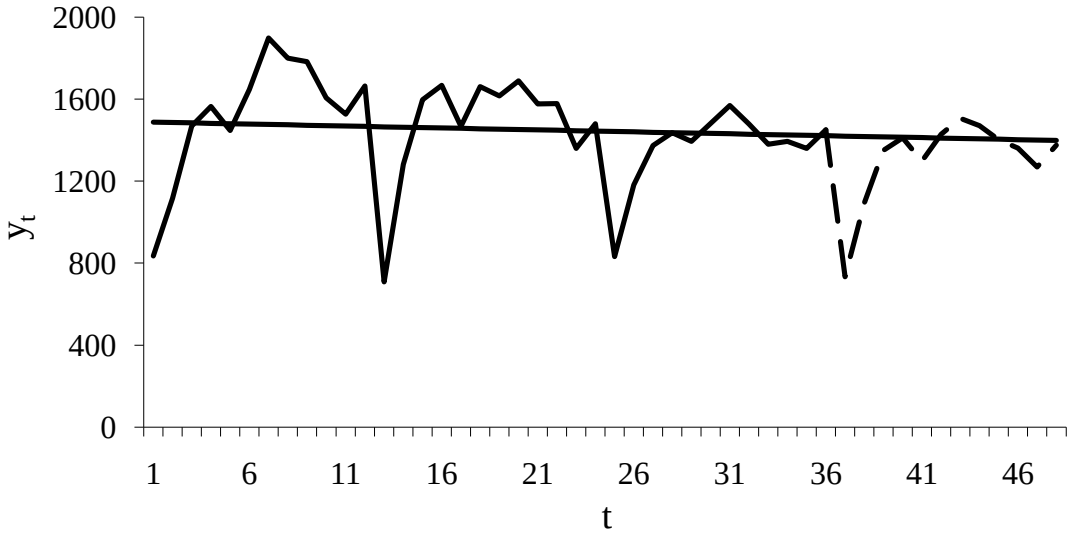


Рис. 9. Прогноз уровней ряда по мультипликативной модели

Этот подход прогнозирования предполагает, что уровень ряда динамики формируется под воздействием множества факторов, но при этом отдельно влияние каждого из них не выделяется. Следовательно, тенденция развития связана не с каким-либо фактором, а с течением времени.

Путем подстановки срока прогноза T в уравнение тренда получают точечную оценку прогноза $\hat{y}_{n+T}$ (в случае цикличности добавляется сезонная компонента). При этом полное совпадение фактического предсказываемого значения y_{n+T} и прогнозной оценки маловероятно. Возникновение отклонений фактических уровней временного ряда от выравненных по уравнению тренда связано со следующими причинами:

1) всегда существует тренд, который дает более точные результаты при описании тенденции, по сравнению с выбранным;

2) тренд, выбранный для прогнозирования, содержит случайную компоненту, так как каждый уровень исходных данных обладает случайной компонентой;

3) выявленная тенденция характеризует движение среднего уровня временного ряда, следовательно, возможны отклонения от него.

При прогнозировании, ввиду приближенного характера прогнозного значения, рекомендуется построение доверительного интервала прогноза. В случае несложных функциональных форм тренда он имеет вид

$$\left(\hat{y}_{n+T} - t_\alpha \cdot S_{\hat{y}} \cdot k(T);\ \hat{y}_{n+T} + t_\alpha \cdot S_{\hat{y}} \cdot k(T)\right), \qquad (26)$$

и с вероятностью $\gamma = 1 - \alpha$ покрывает предсказываемое значение y_{n+T}. Здесь t_α – критическое значение распределения Стьюдента, найденное по таблице по уровню значимости α и числу степеней свободы $k = n - m - 1$, n - длина ряда, m - число параметров уравнения тренда, T - период упреждения. Остаточная средняя квадратическая ошибка прогноза находится по формуле

$$S_{\hat{y}} = \sqrt{\frac{\sum_{t=1}^{n}(\hat{y}_{n+T} - y_t)^2}{n-m-1}}\ . \qquad (27)$$

В случае линейной модели коэффициент $k(T)$ определяется по формуле

$$k(T) = \sqrt{\frac{n+1}{n} + \frac{(t_к + T - \bar{t})^2}{\sum_{t=t_н}^{t_к}(t - \bar{t})^2}}, \qquad (28)$$

где $t_н$, $t_к$ - первая и последняя точки временного интервала, на котором проводилось оценивание параметров модели, $\bar{t}$ - середина временного интервала, на котором проводилось оценивание параметров тренда:

$$\bar{t} = \frac{t_н + t_к}{2}.$$

Если тренд строился по всему ряду, то $t_н = 1$, $t_к = n$, $\bar{t} = \frac{1+n}{2}$.

Для параболического тренда $\hat{y}_t = b_0 + b_1 \cdot t + b_2 \cdot t^2$

$$k(T) = \sqrt{1 + \frac{(t_k + T)^2}{\sum_{t=t_н}^{t_к} t^2} + \frac{\sum_{t=t_н}^{t_к} t^4 - 2(t_k + T)^2 \cdot \sum_{t=t_н}^{t_к} t^2 + n \cdot (t_к + T)^4}{n \cdot \sum_{t=t_н}^{t_к} t^4 - \left(\sum_{t=t_н}^{t_к} t^2\right)^2}}\ . \qquad (29)$$

В случае отсутствия формулы для расчета коэффициента $k(T)$, полагают, что он равен 1.

Относительная ошибка прогноза

$$\delta_{n+T} = \left|\frac{\hat{y}_{n+T} - y_{n+T}}{y_{n+T}}\right| \cdot 100\% \qquad (30)$$

может быть определена только при достижении точки прогноза (момента времени $t = n + T$) , и обычно используется на этапе разработки методики прогнозирования.

Тема 7. РЕГРЕССИЯ С ПЕРЕМЕННОЙ СТРУКТУРОЙ, ФИКТИВНЫЕ ПЕРЕМЕННЫЕ

Лекция

План лекции

1. Фиктивные переменные во множественной регрессии
2. Взаимодействие факторов в регрессионной модели
3. Моделирование сезонности с помощью фиктивных переменных
4. Нелинейная регрессия с фиктивными переменными
5. Регрессия только на фиктивных переменных
6. Регрессия с фиктивной зависимой переменной

1. Фиктивные переменные во множественной регрессии

До сих пор в качестве факторов рассматривались экономические переменные, принимающие количественные значения в некотором интервале. Вместе с тем может оказаться необходимым включить в модель фактор, имеющий два или более качественных уровней. Это могут быть разного рода атрибутивные признаки, такие, например, как профессия, пол, образование, климатические условия, принадлежность к определенному региону. Для того чтобы ввести такие переменные в регрессионную модель, им должны быть присвоены те или иные ***цифровые метки,*** т.е. качественные переменные необходимо преобразовать в количественные. Такого вида сконструированные переменные в эконометрике принято называть фиктивными переменными. В отечественной литературе можно встретить термин «структурные переменные».

Фиктивной переменной (dummy variable) называется атрибутивный или качественный фактор, представленный посредством определённого цифрового кода.

Моделью регрессии с переменной структурой называется модель регрессии, которая включает в качестве факторной переменной фиктивную переменную.

Моделью регрессии без ограничений (unrestricted regression) называется модель регрессии, в которую включены все фиктивные переменные.

Базисной моделью или регрессией с ограничениями (restricted regression) называется модель регрессии, в которой все значения фиктивных переменных равны нулю.

Качественные признаки могут приводить к неоднородности исследуемой совокупности, что может быть учтено при моделировании двумя путями:

• регрессия строится для каждой качественно отличной группы единиц совокупности, т.е. для каждой группы в отдельности, чтобы преодолеть неоднородность единиц общей совокупности;

• общая регрессионная модель строится для совокупности в целом, учитывающей неоднородность данных. В этом случае в регрессионную модель вводятся фиктивные переменные, т.е. строится регрессионная модель с переменной структурой, отражающей неоднородность данных.

Рассмотрим применение фиктивных переменных для функции спроса. Предположим, что по группе лиц мужского и женского пола изучается линейная зависимость потребления кофе y от цены x. В общем виде для совокупности обследуемых уравнение регрессии имеет вид:

$$y=a+bx+\varepsilon.$$

Аналогичные уравнения могут быть найдены отдельно для лиц мужского пола: $y_1=a_1+b_1x_1+\varepsilon$, и женского пола: $y_2=a_2+b_2x_2+\varepsilon_2$.

Различия в потреблении кофе проявятся в различии средних y_1 и y_2. Вместе с тем сила влияния x на y может быть одинаковой, т.е. $b\approx b_1\approx b_2$. В этом случае возможно построение общего уравнения регрессии с включением в него фактора «пол» в виде фиктивной переменной. Объединив уравнения y_1 и y_2 и введя фиктивные переменные, придем к следующему выражению:

$$y_1=a_1z_1+a_2z_2+bx+\varepsilon,$$

где z_1 и z_2 – фиктивные переменные, принимающие значения:

$$z_1=\begin{cases}1-\text{мужской пол}\\0-\text{женский пол}\end{cases};\qquad z_2=\begin{cases}0-\text{мужской пол}\\1-\text{женский пол}\end{cases}$$

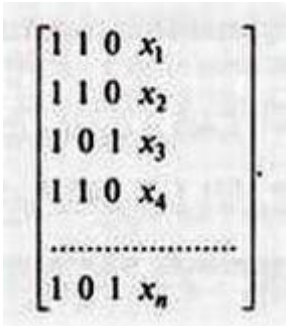
В общем уравнении регрессии зависимая переменная y рассматривается как функция не только цены x, но и пола (z_1, z_2). Переменная z рассматривается как дихотомическая переменная, принимающая всего два значения: 1 и 0. При этом когда $z_1=1$, то $z_2=0$ и, наоборот, при $z_1=0$ переменная $z_2=1$.

Для лиц мужского пола, когда $z_1=1$ и $z_2=0$, объединенное уравнение регрессии составит: $y=a_1+bx$, а для лиц женского пола, когда $z_1=0$ и $z_2=1$ $\hat{y}=a_2+bx$. Иными словами, различия в потреблении для лиц мужского и женского пола вызваны различиями свободных членов уравнения регрессии: $a_1\neq a_2$. Параметр b является общим для всей совокупности лиц как для мужчин, так и для женщин.

Следует иметь в виду, что при введении фиктивных переменных z_1 и z_2 в модель $y=a_xz_1+a_2z_2+bx+\varepsilon$ применение МНК для оценивания параметров a, и a_2 приведет к вырожденной матрице исходных данных, а следовательно, и к невозможности получения их оценок. Объясняется это тем, что при использовании МНК в данном уравнении появляется свободный член, т.е. уравнение примет вид

$$\hat{y}=a_1z_1+a_2z_2+bx+A.$$

Предполагая при параметре A независимую переменную, равную 1, имеем матрицу исходных данных:



В рассматриваемой матрице существует линейная зависимость между первым, вторым и третьим столбцами: первый равен сумме второго и третьего столбцов. Поэтому матрица исходных факторов вырождена. Выходом из создавшегося затруднения может явиться переход к уравнению

$$y=A+A_1z_1+bx+\varepsilon$$

или

$$y=A+A_2z_2+bx+\varepsilon$$

т.е. каждое уравнение включает только одну фиктивную переменную z_1 или z_2.

Предположим, что определено уравнение

$$y=A+A_1z_1+bx+\varepsilon$$

где z_1 – принимает значения 1 для мужчин и 0 для женщин.

Теоретические значения уровня потребления кофе для мужчин будут получены из уравнения

$$\hat{y}=A+A_1+bx.$$

Для женщин соответствующие значения получим из уравнения

$$\hat{y}=A+bx.$$

Сопоставив эти значения, видим, что различия в уровне потребления мужчин и женщин состоят в различии свободных членов данных уравнений: А – для женщин и $A+A_1$ – для мужчин.

Пример. Проанализируем с использованием фиктивных переменных зависимость урожайности пшеницы y от вида вспашки z и количества внесенного органического удобрения x.

По 25 наблюдениям уравнение парной регрессии (без учета вида вспашки) составило:

$$\hat{y}=11{,}463+0{,}326x;$$

$$F=8{,}7;\ t_A=1{,}9;\ t_B=2{,}95;\ r_{yx}=0{,}5246.$$

Для его расчета использовалась следующая система нормальных уравнений:

Уравнение регрессии статистически значимо - F, t_b, r_{yx} превышают табличные значения (на 5%-ном уровне значимости и числе степеней свободы 23 F=4,28; t_b=2,069; r_{yx}=0,398; на 1%-ном уровне значимости F=7,88; t_b=2,807; r_{yx}=0,507).

По виду вспашки поля характеризовались двумя категориями: зяблевая и весенняя. Вид вспашки не влияет на количество внесенных удобрений, но обуслов-

ливает различия в урожайности. Для того чтобы убедиться в этом, введем в уравнение регрессии фиктивную переменную *z* для отражения эффекта вида вспашки, а именно *z*=1 для зяблевой вспашки и *z*=0 для весенней вспашки. Уравнение регрессии примет вид:

$$y=a+bx+cz+\varepsilon$$

Используя метод наименьших квадратов для оценки параметров данного уравнения, получим следующую систему нормальных уравнений:

В рассматриваемом примере вся совокупность из 25 единиц разделена на две подгруппы: с зяблевой вспашкой – 13 полей и с весенней - 12 полей, т.е. n_1=13 и n_2=12. Уравнение регрессии имеет вид

$$\hat{y}=9{,}908+0{,}331x+2{,}908z.$$

Уравнение регрессии статистически значимо: F=15,6; R=0,766; R=0,741; t_a=11,8; t_h=3,9; t_c=4,1.

Как видим, добавление в регрессию фиктивной переменной существенно улучшило результат модели: доля объясненной вариации выросла с 27,5% (r^2_{yx}=0,2752) до 58,7% (R^2_{yxz}=0,5867). При этом сила влияния количества внесенных органических удобрений на урожайность осталась практически неизменной: коэффициенты регрессии, по существу, одинаковы (0,326 в парном уравнении и 0,331 во множественном). Корреляция между видом вспашки и количеством внесенного удобрения на 1 га практически отсутствует: r_{xz}=0,016.

Применение зяблевой вспашки способствует росту урожайности в среднем на 2,9 ц с 1 га при одном и том же количестве внесенного удобрения на 1 га, что в целом соответствует и различию средней урожайности по видам вспашки (15,3 ц с 1 га для зяблевой вспашки и 12,5 ц с 1 га для весенней вспашки). Частный F-критерий для фактора *z* составил 16,58, что выше табличного значения при числе степеней свободы 1 и 22 (4,30 при *a*=0,05 и 7,94 при *a*=0,01). Это подтверждает целесообразность включения фиктивной переменной в уравнение регрессии.

Уравнения парной регрессии по отдельным видам вспашки показывают практически единую меру влияния количества внесенного удобрения на урожайность:

$\hat{y}$=12,678+0,349*x, R*=0,638 - при зяблевой вспашке;

$\hat{y}$=10,148+0,300*x, R*=0,643 – при весенней вспашке.

Поэтому можно предположить единую меру влияния данного фактора, равную значению коэффициента регрессии, в уравнении регрессии с фиктивной переменной (0,331). Включив фиктивную переменную, удалось измерить ее влияние на изменение урожайности: частный коэффициент корреляции r_{yzx}, оценивающий в чистом виде влияние данного фактора, составил 0,6555, что несколько выше, чем аналогичный показатель для фактора *x,* т.е. $r_{yx\cdot z}$=0,6385.

Частные уравнения регрессии по отдельным видам вспашки составили:

$\hat{y}_{(z=1)}$=12,816+0,331*x* -для зяблевой вспашки;

$\hat{y}_{(z=1)}$=9,908+0,331x - для весенней вспашки.

Как видим, функция урожайности для первой группы (при z–1) параллельна функции для второй группы, но сдвинута вверх.

В примере качественный фактор имел только два состояния, которым и соответствовали обозначения 1 и 0. Если же число градаций качественного признака-фактора превышает два, то в модель вводится несколько фиктивных переменных, число которых должно быть меньше числа качественных градаций. Только при соблюдении этого положения матрица исходных фиктивных переменных не будет линейно зависима и возможна оценка параметров модели.

Пример. Проанализируем зависимость цены двухкомнатной квартиры от ее полезной площади. При этом в модель могут быть введены фиктивные переменные, отражающие тип дома: «хрущевка», панельный, кирпичный.

При использовании трех категорий домов вводятся две фиктивные переменные: z_1 и z_2. Пусть переменная z_1 принимает значение 1 для панельного дома и 0 для всех остальных типов домов; переменная z_2 принимает значение 1 для кирпичных домов и 0 для остальных; тогда переменные z_1 и z_2 принимают значения 0 для домов типа «хрущевки».

Предположим, что уравнение регрессии с фиктивными переменными составило:

$$y=320+500x+2200z_1+1600z_2$$

Частные уравнения регрессии для отдельных типов домов, свидетельствуя о наиболее высоких ценах квартир в панельных домах, будут иметь следующий вид:

• «хрущевки» $\hat{y}=320+500x$;

• панельные $\hat{y}=2520+500x$;

• кирпичные $\hat{y}=1920+500x$;

Параметры при фиктивных переменных z_1 и z_2 представляют собой разность между средним уровнем результативного признака для соответствующей группы и базовой группы. В рассматриваемом примере за базу сравнения цены взяты дома «хрущевки», для которых $z_1=z_2=0$. Параметр при $z_1=2200$ означает, что при одной и той же полезной площади квартиры цена ее в панельных домах в среднем на 2200 долл. США выше, чем в «хрущевках». Соответственно параметр при z_2 показывает, что в кирпичных домах цена выше в среднем на 1600 долл. при неизменной величине полезной площади по сравнению с указанным типом домов.

2. Взаимодействие факторов в регрессионной модели

Рассмотренная трактовка параметров регрессии при фиктивных переменных справедлива, если сила влияния на y фактора x действительно не меняется в разных структурных частях совокупности. Иными словами, отсутствует взаимодействие факторов x_j и фиктивных переменных z, т.е. для каждого значения z влияние факторах на y одинаково (рис.).

При отсутствии взаимодействия целесообразно построение модели:

$$\hat{y}=a+bx+cz.$$

При наличии взаимодействия факторов x и z модель с фиктивной переменной принимает вид:

$\hat{y}=a+bx+cz+d(xz),$

что соответствует графическому изображению (рис.):

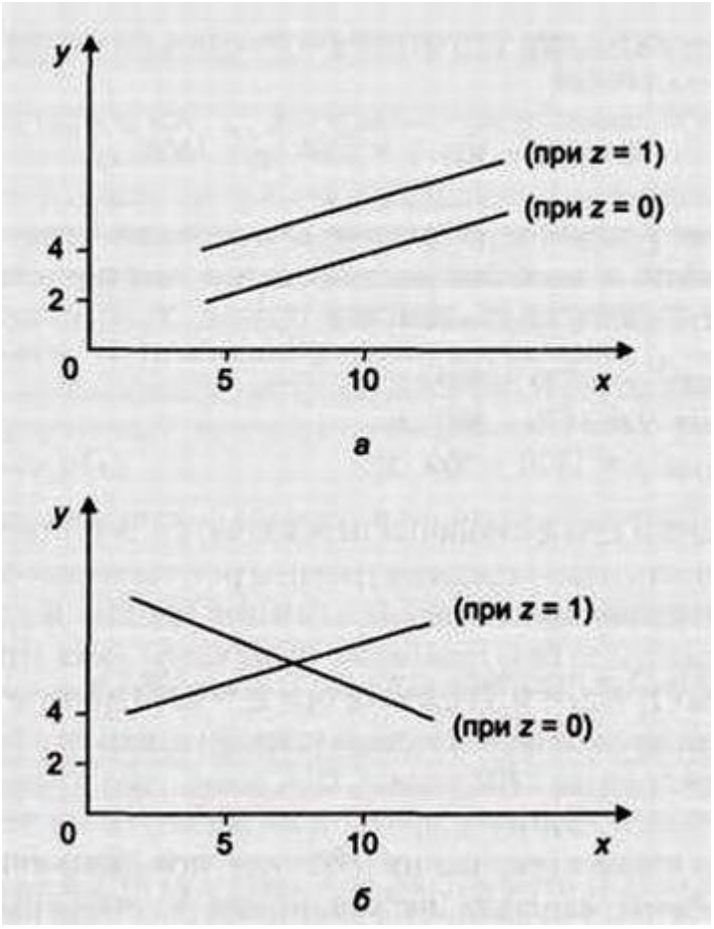


Рис. Графическая иллюстрация взаимодействия факторов:
а – без взаимодействия; *б* – с взаимодействием

Предположим, рассматриваются две группы наблюдений, для каждой из которых имеет место функциональная зависимость *y* от фактора *x:*

$$y_I = 20 + 5x; \quad \tau_{yx} = 1; \quad \overline{y_I} = 35;$$
$$y_{II} = 70 - 3x; \quad \tau_{yx} = 1; \quad \overline{y_{II}} = 61.$$

Поскольку налицо четкое взаимодействие факторов, попытка построить общую регрессионную модель вида $y=a+вx+cz$ приведет к ухудшению результатов аппроксимации модели

$$y=58+1x-26z; \; R^2=0{,}842.$$

Верной в ней будет лишь трактовка коэффициента регрессии при фиктивной переменной *z.* Поскольку в модели $z=1$ для I группы наблюдений, когда $y_1=35$, а $z=0$ для II группы наблюдений, когда $y_{II}=61$, то параметр при *z,* равный – 26, означает, что $y_I-y_{II}=-26$.

Модель с учетом взаимодействия факторов составит:

$$\hat{y}=70-3x-50z+8(zx); \; R^2=1,$$

т.е. функциональная зависимость, заложенная в информацию для каждой группы, продолжает действовать. При $z=0$ мы получим уравнение связи для второй группы,

т.е. y_{II}=70-*3x*. Параметр с при *z* показывает различие в параметрах *a* для двух сравниваемых групп: $c=a_I-a_{II}=-50$. Параметр *d* при совмещенной переменной *(zx)* фиксирует различие в силе связи *y* и *x в* группах:

$$d=b_I-b_{II}=5-(-3)=8.$$

3. Моделирование сезонности с помощью фиктивных переменных

Метод сезонных фиктивных переменных относится к методам моделирования сезонных компонент временного ряда. Суть данного метода заключается в построении модели регрессии, которая наряду с фактором времени включает сезонные фиктивные переменные.

Предположим, что задача состоит в исследовании временного ряда *Xij*, 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 временного ряда равно *n=L*m*.

При построении модели регрессии с переменной структурой необходимо учитывать, что число сезонных фиктивных переменных всегда должно быть на единицу меньше сезонов внутри года, т. е. должно быть равно величине *L-1*. Например, при моделировании годовых данных модель регрессии помимо фактора времени должна содержать одиннадцать фиктивных компонент (12-1). При моделировании поквартальных данных модель регрессии должна содержать три фиктивные компоненты (4-1) и т. д.

Каждому из сезонов соответствует определённое сочетание фиктивных переменных. Сезон, для которого значения всех фиктивных переменных равны нулю, является базой сравнения. Для остальных сезонов одна из фиктивных переменных принимает значение, равное единице. Например, если имеются поквартальные данные, то значения фиктивных переменных *D2,D3,D4* будут принимать следующие значения для каждого из кварталов:

Квартал	D_2	D_3	D_4
1	0	0	0
2	1	0	0
3	0	1	0
4	0	0	1

Тогда общий вид модели регрессии с переменной структурой будет иметь вид:

*yt=?0+ ?1*t+?2*D2+?3*D3+?4*D4+?t.*

Данная модель регрессии представляет собой одну из разновидностей аддитивной модели временного ряда.

На основе общей модели регрессии с переменной структурой можно составить базисную модель или модель тренда для первого квартала:

*yt=?0+ ?1*t+?t.*

Также на основе общей модели регрессии с переменной структурой можно составить частные модели регрессии*:*

1) частная модель регрессии для второго квартала:

*yt=?0+ ?1*t+?2+?t;*

2) частная модель регрессии для третьего квартала:

*yt=?0+ ?1*t+?3+?t;*

3) частная модель регрессии для четвёртого квартала:

*yt=?0+ ?1*t+?4+?t.*

Данные частные модели регрессии отличаются друг от друга только на величину свободного члена ?i.

Коэффициент ?1 характеризует среднее абсолютное изменение уровней временного ряда под влиянием основной тенденции.

Сезонная компонента для каждого сезона рассчитывается как разность между средним значением свободных членов всех частных моделей регрессий и значением постоянного члена одной из моделей.

Среднее значение свободных членов всех частных моделей регрессий рассчитывается по формуле:

$$\overline{\beta}_0 = \frac{\beta_0 + (\beta_0 + \delta_2) + (\beta_0 + \delta_3) + (\beta_0 + \delta_4)}{4}$$

Для поквартальных данных оценка сезонных отклонений осуществляется по формулам:

1) оценка сезонного отклонения для первого квартала:

$$(\beta_0 - \overline{\beta}_0);$$

2) оценка сезонного отклонения для второго квартала:

$$(\beta_0 + \delta_2 - \overline{\beta}_0);$$

3) оценка сезонного отклонения для третьего квартала:

$$(\beta_0 + \delta_3 - \overline{\beta}_0);$$

4) оценка сезонного отклонения для четвёртого квартала:

$$\left(\beta_0 + \delta_4 - \overline{\beta}_0\right).$$

Сумма сезонных отклонений должна равняться нулю.

4. Нелинейная регрессия с фиктивными переменными

Фиктивные переменные могут вводиться не только в линейные, но и в нелинейные модели, приводимые путем преобразований к линейному виду. Так, модель с фиктивными переменными может иметь вид:

$$\ln y = a + b_1 x_1 + \ldots + b_p x_p + cz + \varepsilon,$$

где z – фиктивная переменная.

Целесообразность такого вида модели диктуется характером связи между экономическими переменными:

$$y = a \cdot b_1^{x_1} \cdot b_2^{x_2} \ldots b_p^{x_p} \cdot \varepsilon.$$

Фиктивная переменная вводится в эту модель как очередной сомножитель:

$$y = a \cdot b_1^{x_1} \cdot b_2^{x_2} \ldots b_p^{x_p} \cdot c^z \cdot \varepsilon.$$

Логарифмируя данное выражение, получим модель вида

$$\ln y = \ln a + x_1 \ln b_1 + x_2 \ln b_2 + \ldots + x_p \ln b_p + z \ln c + \ln \varepsilon,$$

которая равносильна приведенной ранее

$$\ln y = a + b_1 x_1 + \ldots + b_p x_p + cz + \varepsilon,$$

где параметры и случайная составляющая представлены в логарифмах.

5. Регрессия только на фиктивных переменных

До сих пор мы рассматривали фиктивные переменные как факторы, которые используются в регрессионной модели наряду с количественными переменными. Вместе с тем возможна регрессия только на фиктивных переменных. Например, изучается дифференциация заработной платы рабочих высокой квалификации по регионам страны. Модель заработной платы может иметь вид:

$$\hat{y} = a + b_1 z_1 + b_2 i_2 + \ldots b_k z_k,$$

где y – средняя заработная плата рабочих высокой квалификации по отдельным предприятиям;

$$z_1 = \begin{cases} 1 - \text{если предприятие находится в Северо-Западном районе,} \\ 0 - \text{если предприятие находится в остальных районах;} \end{cases}$$

$$z_2 = \begin{cases} 1 - \text{если предприятие находится в Волго-Вятском районе,} \\ 0 - \text{если предприятие находится в остальных районах;} \end{cases}$$

$$z_4 = \begin{cases} 1 - \text{если предприятие находится в Дальневосточном районе,} \\ 0 - \text{если предприятие находится в остальных районах.} \end{cases}$$

Поскольку последний район, указанный в модели, обозначен z_k, в исследование включен $\kappa+1$ район.

Ввиду того что факторы данной регрессионной модели выражены как дихотомические признаки, параметры модели имеют свою специфику по сравнению с их традиционной интерпретацией. Параметр *a* представляет собой среднее значение результативного признака для базовой группы y_0. Параметр b_1 характеризует разность средних уровней результативного признака для группы 1 и базовой группы 0. Соответственно параметр b_i представляет собой разность между y_i и y_0. Иными словами, коэффициенты при *z* отражают величину эффекта соответствующей группы фактора *z*.

6. Регрессия с фиктивной зависимой переменной

Мы рассмотрели модели с фиктивными переменными, в которых последние выступают факторами. Может возникнуть необходимость построить модель, в которой дихотомический признак играет роль результата. Подобного вида модели применяются, например, при обработке данных социологических опросов. В качестве зависимой переменной *y* рассматриваются ответы на вопросы, данные в альтернативной форме: «да» или «нет». Поэтому зависимая переменная принимает два значения: 1, что значит ответ «да», и 0 - во всех остальных случаях. Модель такой зависимой переменной имеет вид:

$$y=a+b_1x_1+...+b_px_p+\varepsilon.$$

Модель является вероятностной линейной моделью. В ней *y* принимает значения 1 и 0, которым соответствуют вероятности *p* и 1–*p*. Поэтому при решении модели находят оценку условной вероятности события *y* при фиксированных значениях *x*. Для оценки параметров линейно-вероятностной модели применяются методы Tobit-, Logit- и Probit-анализа.

Модели такого рода используют при работе с неколичественными переменными. Как правило, это модели выбора из заданного набора альтернатив. Зависимая переменная *y* представлена дискретными значениями (набор альтернатив), объясняющие переменные x_j – характеристики альтернатив (время, цена), z_j – характеристики индивидов (возраст, доход, уровень образования). Модель такого рода позволяет предсказать долю индивидов в генеральной совокупности, которые выбирают данную альтернативу.

Включение в модель фиктивных переменных может иметь цель отразить в модели неоднородность совокупности. Однако нельзя рассматривать фиктивные переменные как панацею при применении методов регрессии к неоднородным данным.

Тема 8. СТРУКТУРНАЯ СТАБИЛЬНОСТЬ ДАННЫХ

Лекция

План лекции

1. Понятие структуры данных статистических наблюдений
2. Аномальные значения и грубые ошибки
3. Статистический анализ структурной однородности двух выборок
4. Проблема объединения двух выборок. Критерий ГрегориЧоу.

1. Понятие структуры данных статистических наблюдений

Изучаемые статистикой процессы и явления в финансово-экономической сфере, в демографии, в социальной и политической областях, как правило, характеризуются внутренней структурой, которая с течением времени может изменяться. Динамика структуры вызывает изменение внутреннего содержания исследуемых объектов и их экономической интерпретации, приводит к изменению установившихся причинно-следственных связей. Именно поэтому изучение структуры и структурных сдвигов занимает важное место в экономико-статистическом анализе.

В статистике под ***структурой*** понимают совокупность элементов социально-экономических явлений, обладающих определенной устойчивостью внутригрупповых связей при сохранении основных свойств, характеризующих эту совокупность как целое. В качестве примеров можно привести структуру населения региона по возрасту или уровню доходов, структуру предприятий отрасли по численности промышленно-производственного персонала или стоимости основных фондов и другие.

Основные направления статистического изучения структуры включают: а) характеристику структурных сдвигов отдельных частей совокупности за два и более периодов; б) обобщающую характеристику структурных сдвигов в целом по совокупности; в) оценку степени концентрации и централизации.

Анализ структуры и ее изменений базируется на относительных показателях структуры - долях или удельных весах, представляющих собой соотношения размеров частей и целого. При этом как частные, таки обобщающие показатели структурных сдвигов могут отражать либо «абсолютное» изменение структуры в процентных пунктах или долях единицы (кавычки показывают, что данные показатели являются абсолютными по методологии расчета, но не по единицам измерения), либо ее относительное изменение в процентах или коэффициентах.

Подробно эти вопросы рассматриваются в курсе «Общая теория статистики».

2. Аномальные значения и грубые ошибки

При обработке экономической информации, как правило, считают, что экономические показатели подчиняются нормальному закону распределения. Однако на практике это предположение зачастую не выполняется. Кроме того, исходная информация может содержать пропуски значений, а также аномальные значения (выбросы) и ***«грубые ошибки»***, которые могут появиться при сборе исходной

информации, а также в результате ее искажения в каналах передачи.

Предварительная обработка исходной информации предполагает устранение этих нарушений. Простейший способ восстановления отсутствующих значений - применение методов интерполяции (в первом приближении - линейных).

Методы, учитывающие наличие грубых ошибок и позволяющие получать достаточно точные оценки параметров без удаления аномальных значений, называются ***робастными*** или устойчивыми.

Наибольшее распространение получили методы выявления аномальных значений и их удаления из совокупности (метод выявления грубых ошибок Смирнова-Граббса, критерий Граббса, критерий обнаружения экстремальных наблюдений, критерий исключения нескольких грубых ошибок и др.). Эти методы достаточно просты в использовании на практике.

Необходимо отметить, что при обработке многомерной статистической информации такой подход может привести к отбрасыванию слишком большого количества точек наблюдения. А если объектом наблюдения является предприятие отрасли, то такое отбрасывание может привести к анализу узкой группы предприятий вместо изучения экономических закономерностей во всей отрасли.

По этим причинам в последнее время получила широкое развитие вторая группа методов устойчивого оценивания показателей, подчиняющихся различным законам распределения, без исключения из совокупности аномальных наблюдений (методы Тьюки, Хубера и др.). Применение этих методов требует от исследователя достаточно глубокого знания математической статистики.

При определении структуры неоднородных совокупностей возникают две задачи. Первая задача заключается в разбиении общей неоднородной совокупности на некоторое число однородных совокупностей, вторая - в оценке параметров совокупностей, содержащих грубые ошибки.

При решении первой задачи необходимо: классифицировать элементы по однородным совокупностям; оценить параметры распределения однородных составляющих, входящих в общую неоднородную совокупность.

При решении второй задачи чаще используются методы непосредственного выявления грубых ошибок и методы, сводящие к минимуму искажения, создаваемые грубыми ошибками, а также комбинированные методы.

Метод Смирнова-Граббса обнаружения грубых ошибок

Метод обнаружения грубых ошибок Смирнова-Граббса наиболее прост в применении. Он позволяет ответить на вопрос - является ли максимальное (минимальное) наблюдение грубой ошибкой (засорением).

Проверка максимального наблюдения. Пусть $x_1, x_2, \ldots, x_n$ - результаты наблюдения. Построим по ним вариационный ряд (упорядочим наблюдения по возрастанию)

$$X_{(1)} \le X_{(2)} \le \cdots \le X_{(N)} \ . \qquad (1)$$

Рассчитаем статистику (расчетное значение критерия)

$$T_{(N)} = \frac{x_{(N)} - \bar{x}}{s}, \qquad (2)$$

где среднее значение и среднеквадратическое отклонение вариационного ряда соответственно равны

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_{(i)} \ , \ S = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_{(i)} - \bar{x})^2} \ .$$

Значение $T_{(N)}$ сравнивается с критическим значением C_α (α - уровень значимости, вероятность отвергнуть нулевую гипотезу H_0), найденным по таблице Граббса (табл. 1).

Таблица 1

Таблица Граббса. Критические значения C_α , α = 0,05

Число наблюдений	1	2	3	4	5	6	7	8	9
C_α	-	-	1,412	1,689	1,869	1,996	2,093	2,172	2,237
Число наблюдений	10	11	12	13	14	15	16	17	18
C_α	2,294	2,343	2,387	2,426	2,461	2,493	2,523	2,551	2,577
Число наблюдений	19	20	21	22	23	24	25	26	
C_α	2,600	2,623	2,644	2,664	2,683	2,701	2,717	2,734	

Если $T_{(N)} < C_\alpha$, то верна гипотеза H_0 о том, что $x_{(N)}$ не является грубой ошибкой, при конкурирующей гипотезе H_1 - $x_{(N)}$ является грубой ошибкой.

При $T_{(N)} > C_\alpha$ значение $x_{(N)}$ значимо отклоняется от $\bar{x}$, следовательно, является грубой ошибкой (принимается гипотеза H_1).

Проверка минимального наблюдения производится аналогично. Находится расчетное значение критерия для минимального наблюдения

$$T_{(1)} = \frac{\bar{x} - x_{(1)}}{s} \ , \qquad (3)$$

если $T_{(1)} < C_\alpha$, то $x_{(1)}$ не является грубой ошибкой.

Пример. На основе данных средней производительности труда (выработка на одного работающего 10 предприятий отрасли (табл.2)) проверить наличие грубых ошибок.

Таблица 2

№ предприятия	1	2	3	4	5	6	7	8	9	10
Выработка	12	11	13	12	14	12	18	15	14	13

Составим вариационный ряд: 11, 12, 12, 12, 13, 13, 14, 14, 15, 18. Наибольшее значение равно $x_{(10)} = 18$. Вычислим среднее значение $\bar{x} = 13{,}4$ и среднее квадратическое отклонение s = 1,91. Для наибольшего значения найдем $T_{(10)} = (18 - 13{,}4)/1{,}91 = 2{,}41$. По табл. 2.1 для $N = 10$ и $\alpha = 0{,}05$ определим $C_\alpha = 2{,}294$. В данном случае $T_{(10)} > C_{0,05}$, т.е. 2,41 > 2,294. Следовательно, гипотеза об однородности ряда наблюдений отвергается.

Значение выработки на одного работающего для предприятия №7 является нетипичным. Это значение можно считать грубой ошибкой (аномальным наблюдением) при уровне значимости α = 0,05.

3. Статистический анализ структурной однородности двух выборок

Пусть необходимо проверить структурную однородность двух выборок, полученных с помощью независимых статистических наблюдений. Для этого можно использовать статистические критерии.

3.1. Сравнение двух средних

Пусть необходимо проверить гипотезу о том, что две независимые выборки, полученные из генеральных совокупностей X и Y, имеют равные средние значения.

1 случай.

Генеральные совокупности X и Y распределены нормально, причем известны их дисперсии. Из этих генеральных совокупностей извлечены выборки объемов соответственно m и n, для которых найдены выборочные средние $\overline{x}_в$ и $\overline{y}_в$. При заданном уровне значимости α проверяется нулевая гипотеза о равенстве математических ожиданий генеральных совокупностей: H_0: M (X) = M (Y).

Статистическим критерием для проверки этой гипотезы является нормированная нормально распределенная случайная величина

$$Z = \frac{M(X) - M(Y)}{\sqrt{\frac{D(X)}{m} + \frac{D(Y)}{n}}}.$$

Наблюдаемое значение критерия

$$Z_{набл} = \frac{\overline{x}_в - \overline{y}_в}{\sqrt{\frac{D(X)}{m} + \frac{D(Y)}{n}}}.$$

Вид критической области зависит от типа конкурирующей гипотезы:

a) $H_1 : M(X) \neq M(Y)$ – критическая область двусторонняя, $z_{кр}$ определяется как аргумент функции Лапласа, при котором

$$\Phi(z_{кр}) = \frac{1-\alpha}{2},$$

и критическая область задается неравенством $|Z_{набл}| > z_{кр}$.

b) $H_1 : M(X) > M(Y)$ – критическая область правосторонняя, $z_{кр}$ определяется как аргумент функции Лапласа, при котором

$$\Phi(z_{кр}) = \frac{1-2\alpha}{2},$$

и критическая область определяется неравенством $Z > z_{кр}$.

с) $H_1 : M(X) < M(Y)$ – критическая область левосторонняя, заданная неравенством $Z < -z_{кр}$, где $z_{кр}$ вычисляется так же, как в предыдущем случае.

2 случай.

Имеются две независимые выборки большого объема, извлеченные из генеральных совокупностей, законы распределения и дисперсии которых неизвестны. При этом для объема выборки, не меньшего 30, можно считать, что выборочные средние распределены приближенно нормально, а выборочные дисперсии являются достаточно хорошими оценками генеральных дисперсий (следовательно, считаем известными приближенные значения генеральных дисперсий). Тогда задача сводится к предыдущей, и статистический критерий имеет вид:

$$Z' = Z = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{D_в(X)}{m} + \frac{D_в(Y)}{n}}}.$$

Наблюдаемое значение критерия вычисляется по формуле:

$$Z'_{набл} = \frac{\overline{x}_в - \overline{y}_в}{\sqrt{\frac{D_в(X)}{m} + \frac{D_в(Y)}{n}}}.$$

При этом выбор вида критической области и определение критических точек проводятся так же, как в случае 1.

3 случай.

Генеральные совокупности распределены нормально, причем их дисперсии неизвестны, а объем выборок т и п мал (следовательно, нельзя получить хорошие оценки генеральных дисперсий). Если предположить, что генеральные дисперсии равны, то в качестве критерия для проверки нулевой гипотезы $H_0 : M(X) = M(Y)$ служит случайная величина

$$T = \frac{\overline{X} - \overline{Y}}{\sqrt{(m-1)\cdot S_X^2 + (n-1)\cdot S_Y^2}} \cdot \sqrt{\frac{mn\cdot(n+m-2)}{n+m}},$$

имеющая при справедливости нулевой гипотезы распределение Стьюдента с k = n + m – 2 степенями свободы. Наблюдаемое значение критерия вычисляется по формуле

$$T_{набл} = \frac{\overline{x}_в - \overline{y}_в}{\sqrt{(m-1)\cdot S_X^2 + (n-1)\cdot S_Y^2}} \cdot \sqrt{\frac{mn\cdot(n+m-2)}{n+m}}.$$

Критическая область строится в зависимости от вида конкурирующей гипотезы.

а) $H_1 : M(X) \neq M(Y)$ – критическая область двусторонняя, задаваемая неравенством $|T_{набл}| > t_{двусторкр}$., где $t_{двусторкр}(\alpha; k)$ находится из таблицы критических точек распределения Стьюдента.

b) $H_1 : M(X) > M(Y)$ – критическая область правосторонняя, определяемая условием $T_{наб} > t_{крав кр.}$. Критическая точка вновь находится по таблице критических точек распределения Стьюдента.

c) $H_1 : M(X) < M(Y)$ – критическая область левосторонняя, $T_{наб} < -t_{крав кр.}$

Пример. Имеются независимые выборки значений нормально распределенных случайных величин

X: 2, 2, 3, 3, 4, 4, 4, 5, 5, 6 и

Y: 3, 3, 3, 4, 4, 4, 4, 5, 5, 5, 5, 5, 7, 8, 9.

Требуется проверить для уровня значимости $\alpha = 0{,}1$ при условии равенства генеральных дисперсий нулевую гипотезу $H_0 : M(X) = M(Y)$ при конкурирующей гипотезе $H_1 : M(X) \neq M(Y)$.

Решение. Объемы выборок т = 10, п = 15. Вычислим выборочные средние и исправленные выборочные дисперсии: $\bar{x}_в = 3{,}8;\ \bar{y}_в = 4{,}93;\ S_x^2 = 1{,}73;\ S_y^2 = 3{,}21$. Вычислим наблюдаемое значение критерия:

$$T_{наб} = \frac{3{,}8 - 4{,}93}{\sqrt{9 \cdot 1{,}73 + 14 \cdot 3{,}21}} \sqrt{\frac{10 \cdot 15 \cdot 23}{25}} = -1{,}706 .$$

Критическая область – двусторонняя, tдвуст.кр.(0,1; 23) = 1,71 (из таблицы критических точек распределения Стьюдента). Итак, $|T_{наб}| < t_{двустор кр}$, следовательно, нет оснований отвергнуть нулевую гипотезу – можно считать, что математические ожидания генеральных совокупностей равны.

3.2. Сравнение двух дисперсий

Пусть генеральные совокупности X и Y распределены нормально. По независимым выборкам объемов n_1 и n_2, извлеченным из этих совокупностей, найдены исправленные выборочные дисперсии S^2_x и S^2_y. Требуется по исправленным дисперсиям при заданном уровне значимости α проверить нулевую гипотезу, состоящую в том, что генеральные дисперсии рассматриваемых совокупностей равны между собой: H_0: D(X) = D(Y).

Учитывая, что исправленные дисперсии являются несмещенными оценками генеральных дисперсий, т. е. M (S^2_x) =D(X), M (S^2_y) = D (Y), нулевую гипотезу можно записать так: H_0: M (S^2_x) = M (S^2_y).

Таким образом, требуется проверить, что математические ожидания исправленных выборочных дисперсий равны между собой. Такая задача ставится потому, что обычно исправленные дисперсии оказываются различными. Возникает вопрос: значимо (существенно) или незначимо различаются исправленные дисперсии?

Если окажется, что нулевая гипотеза справедлива, т. е. генеральные дисперсии одинаковы, то различие исправленных дисперсий незначимо и объясняется случайными причинами, в частности, случайным отбором объектов выборки.

Если нулевая гипотеза будет отвергнута, т.е. генеральные дисперсии неодинаковы, то различие исправленных дисперсий значимо и не может быть объяснено случайными причинами, а является следствием того, что сами генеральные дисперсии различны.

В качестве критерия проверки нулевой гипотезы о равенстве генеральных дисперсий, причем отношение большей исправленной дисперсии к меньшей, т. е. случайную величину:

$$F = \frac{S_б^2}{S_м^2}.$$

Величина F при условии справедливости нулевой гипотезы имеет распределение Фишера-Снедекора со степенями свободы $k_1 = n_1 - 1$ и $k_2 = n_2 - 1$, где n_1 – объем выборки, по которой вычислена большая исправленная дисперсия.

Критическая область строится в зависимости от вида конкурирующей гипотезы.

1 случай.

Нулевая гипотеза H_0: $D(X) = D(Y)$. Конкурирующая гипотеза H_1: $D(X) > D(Y)$. В этом случае строят одностороннюю, а именно правостороннюю критическую область, исходя из требования, чтобы вероятность попадания критерия F в эту область в предположении справедливости нулевой гипотезы была равна принятому уровню значимости:

$$P(F > F_{кр}(\alpha, k_1, k_2)) = \alpha.$$

Критическую точку $F_{кр}(\alpha, k_1, k_2)$ находят по таблице критических точек распределения Фишера-Снедекора, и тогда правосторонняя критическая область определяется неравенством $F > F_{кр}$, а область принятия нулевой гипотезы – неравенством $F < F_{кр}$.

Обозначим отношение большей исправленной дисперсии к меньшей, вычисленное по данным наблюдений, через Fнабл и сформулируем правило проверки нулевой гипотезы.

Правило.

Вычислить отношение большей исправленной дисперсии к меньшей, т.е.

$$F_{набл} = \frac{S_б^2}{S_м^2}.$$

По таблице критических точек распределения Фишера- Снедекора, по заданному уровню значимости и числам степеней свободы k_1 и k_2 (k_1 – число степеней свободы большей исправленной дисперсии) найти критическую точку $F_{кр}(\alpha, k_1, k_2)$.

Если $F_{набл} < F_{кр}$ – нет оснований отвергнуть нулевую гипотезу.

Если $F_{набл} > F_{кр}$ – нулевую гипотезу отвергают.

2 случай.

Нулевая гипотеза H_0: $D(X) = D(Y)$. Конкурирующая гипотеза H_1: $D(X) \neq D(Y)$.

В этом случае строят двустороннюю критическую область исходя из требования, чтобы вероятность попадания критерия в эту область, в предположении справедливости нулевой гипотезы, была равна этому уровню значимости.

Как выбрать границы критической области? Оказывается, что наибольшая мощность (вероятность попадания критерия в критическую область при справедливости конкурирующей гипотезы) достигается тогда, когда вероятность попадания критерия в каждый из двух интервалов критической области равна $\alpha/2$.

Таким образом, если обозначить через F_1 левую границу критической области и через F_2 – правую, то должны иметь место соотношения: $P(F<F_1)=\alpha/2$, $P(F>F_1)=\alpha/2$.

Достаточно найти критические точки, чтобы найти саму критическую область: $F < F_1$, $F > F_2$, а также область принятия нулевой гипотезы: $F_1 < F < F_2$.

Как практически отыскать критические точки?

Правую критическую точку $F_2 = F_{кр}(\alpha/2, k_1, k_2)$ находят непосредственно по таблице критических точек распределения Фишера-Снедекора по уровню значимости и степеням свободы k_1 и k_2.

Однако левых критических точек эта таблица не содержит, и поэтому найти F_1 непосредственно по таблице невозможно. Но можно левую критическую точку и не отыскивать.

Оказывается, достаточно найти правую критическую точку F_2 при уровне значимости, вдвое меньшем заданного. Тогда не только вероятность попадания критерия в «правую часть» критической области (т.е. правее F_2) равна $\alpha/2$, но и вероятность попадания этого критерия в «левую часть» критической области (т.е. левее F_1) будет также равна $\alpha/2$. Так как эти события несовместимы, то вероятность попадания рассматриваемого критерия во всю двустороннюю критическую область будет равна $\alpha/2$. Таким образом, в случае конкурирующей гипотезы H_1: $D(X) \neq D(Y)$ достаточно найти критическую точку $F_2=F_{кр}(\alpha/2, k_1, k_2)$.

Правило.

Вычислить отношение большей исправленной дисперсии к меньшей, т.е.

$F = \frac{S_б^2}{S_м^2}$.

По таблице критических точек распределения Фишера-Снедекора по уровню значимости $\alpha/2$ (вдвое меньше заданного) и числам степеней свободы k_1 и k_2 (k_1 - число степеней свободы большей дисперсии) найти критическую точку $F_{кр}(\alpha/2, k_1, k_2)$.

Если $F_{набл} < F_{кр}$ – нет оснований отвергать нулевую гипотезу.

Если $F_{набл} > F_{кр}$ – нулевую гипотезу отвергают.

4. Проблема объединения двух выборок. Критерий Грегори Чоу.

При исследовании экономических процессов иногда возникает возможность или необходимость объединения двух выборок типа cross-section. При этом необходимым условием объединения является полное совпадение показателей

$\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m$. Кроме того необходима структурная однородность этих выборок. Статистическую значимость структурных изменений можно оценить с помощью критерия Грегори Чоу. Система обозначений для его проверки приведена в табл. 3.

Таблица 3

№ выборки	Вид уравнения регрессии*	Объем выборки	Остаточная сумма квадратов	Число параметров в уравнении*
Две выборки				
(1)	$y^{(1)} = b_{10} + b_{11} \cdot x_1 + \ldots + b_{1m} \cdot x_m$	n_1	$S^1_{ост}$	k_1
(2)	$y^{(2)} = b_{20} + b_{21} \cdot x_1 + \ldots + b_{2m} \cdot x_m$	n_2	$S^2_{ост}$	k_2
Объединенная выборка				
(3)	$y^{(3)} = b_{30} + b_{31} \cdot x_1 + \ldots + b_{3m} \cdot x_m$	$n = n_1 + n_2$	$S^3_{ост}$	k_3
* В рассматриваемом примере все уравнения регрессии линейные, число параметров всех уравнений $k_1 = k_2 = k_3 = m + 1$. В общем случае вид уравнений и число параметров в каждом уравнении могут различаться.				

Основная гипотеза $\mathbf{H}_0$ формулируется как утверждение о том, что качество общей модели регрессии для объединенной выборки лучше качества частных моделей регрессии для отдельных выборок.

Альтернативная гипотеза $\mathbf{H}_1$ утверждает, что качество общей модели регрессии хуже качества частных моделей регрессии.

Найдем остаточные суммы квадратов

$$S^1_{ост} = \sum_{i=1}^{n1} \left(y_i - \hat{y}_i^{(1)} \right)^2, \qquad (4)$$

$$S^2_{ост} = \sum_{i=1}^{n2} \left(y_i - \hat{y}_i^{(2)} \right)^2, \qquad (5)$$

$$S^3_{ост} = \sum_{t=1}^{n} \left(y_i - \hat{y}_i^{(3)} \right)^2, \qquad (6)$$

где $\hat{y}_i^{(1)}$, $\hat{y}_i^{(2)}$, $\hat{y}_i^{(3)}$ - теоретические значения переменной $\boldsymbol{y}$, найденные соответственно по уравнениям (1), (2) и (3).

Сумма остаточных сумм квадратов отдельных выборок равна

$$S^{кл}_{ост} = S^1_{ост} + S^2_{ост} \ .$$

Соответствующее ей число степеней свободы составит

$$(n_1 - k_1) + (n_2 - k_2) = (n - k_1 - k_2)$$

Изменение остаточной дисперсии при переходе от уравнения регрессии для объединенной выборки к двум уравнениям необъединенных выборок определяется как разность

$$\Delta S_{ост} = S^{3}_{ост} - S^{кл}_{ост} ,$$

соответствующее ему число степеней свободы равно

$$n - k_3 - (n - k_1 - k_2) = k_1 + k_2 - k_3 .$$

Расчетное значение ***F***-критерия

$$F_{расч} = \frac{\Delta S_{ост}}{S^{кл}_{ост}} \cdot \frac{n - k_1 - k_2}{k_1 + k_2 - k_3}. \qquad (7)$$

сравнивается с табличным $F_{табл}(\alpha, df_1, df_2)$, полученным по таблице критических точек распределения Фишера для уровня значимости α и числа степеней свободы $df_1 = k_1 + k_2 - k_3$, $df_2 = n - k_1 - k_2$.

Если $F_{расч} > F_{табл}$, то гипотеза H_0 о структурной стабильности выборок отклоняется, и влияние структурных различий в значениях показателей считается значимым, выборки объединять нельзя. В противном случае, когда $F_{расч} < F_{табл}$, нет оснований отклонять гипотезу H_0 , выборки можно объединить.

Тема 9. СИСТЕМЫ ОДНОВРЕМЕНЫХ УРАВНЕНИЙ

Лекция

План лекции

1. Виды систем уравнений в эконометрике. Структурная и приведенная формы модели
2. Применение систем одновременных уравнений
3. Задача идентификации уравнений системы. Необходимое и достаточное условие идентифицируемости
4. Косвенный метод наименьших квадратов
5. Двухшаговый метод наименьших квадратов

1. Виды систем уравнений в эконометрике. Структурная и приведенная формы модели

Системы уравнений в эконометрике

Для изучения комплексных экономических явлений средствами эконометрики, как правило, применяют не отдельные уравнения регрессии, а системы уравнений. Это объясняется следующим.

Во-первых, описывая явление с помощью взаимосвязанных переменных, приходится учитывать, что изменение одной переменной влечет за собой изменение других. При рассмотрении же отдельного регрессионного уравнения часто предполагают, что объясняющие переменные можно изменять независимо одну от другой.

Во-вторых, взаимодействие переменных нередко затрудняет однозначную их классификацию при построении модели: одну и ту же переменную можно определить как объясняющую (фактор) и как объясняемую (результат).

Системы уравнений в эконометрике подразделяют на виды: независимых уравнений, рекурсивных уравнений и взаимозависимых (совместных) уравнений.

Третий вид, а именно системы совместных уравнений, представляет наибольший практический интерес. Такие системы эффективны в эконометрических исследованиях и наиболее широко применяются в макроэкономике. В силу этого под системой эконометрических уравнений обычно понимают систему совместных уравнений. Систему совместных (взаимозависимых) уравнений по-другому называют ***системой одновременных уравнений,*** указывая на то, что одни и те же переменные системы рассматриваются одновременно как объясняемые в одном уравнении и как объясняющие - в остальных уравнениях.

Виды систем эконометрических уравнений

Система независимых уравнений	***Система рекурсивных уравнений***	***Система одновременных уравнений***
Каждый результативный признак (объясняемая переменная) y_j, где $j = 1, 2, ..., n$, является функцией одной и той же совокупности факторов (объясняющих переменных) x_i, где $i = 1, 2, ..., m$. Набор факторов в каждом уравнении системы может варьировать в зависимости от изучаемого явления	Результативный признак (объясняемая переменная) y_j, где $j = 1, 2, ..., n$, одного уравнения системы в каждом последующем уравнении является фактором наряду с одной и той же совокупностью факторов x_i, где $i = 1, 2, ..., m$.	Результативный признак (объясняемая переменная) y_j, где $j = 1, 2, ..., n$, одного уравнения системы входит во все другие уравнения системы в качестве фактора наряду с одной и той же совокупностью факторов x_i, где $i = 1, 2, ..., m$

Систему *независимых* или *рекурсивных уравнений* решают с помощью *метода наименьших квадратов* (МНК). Для решения системы одновременных уравнений требуются другие, отличные от МНК методы. Их применение обусловливается тем, что результативный признак одного уравнения системы в другом уравнении этой системы используется в качестве фактора, и будет коррелировать с соответствующей ошибкой.

Модели системы одновременных уравнений и их составляющие

Система одновременных уравнений	
В виде ***структурной формы*** модели	В виде ***приведенной формы*** модели

Основными составляющими обеих форм записи являются ***эндогенные*** и ***экзогенные переменные.*** Эндогенные переменные (y) определяются внутри модели и являются зависимыми переменными. Экзогенные переменные (x) определяются вне системы и являются независимыми переменными. Предполагается, что экзогенные переменные не коррелируют с ошибкой в соответствующем уравнении.

Структурная форма модели

$$\begin{cases} y_1 = c_{10} + b_{12} y_2 + b_{13} y_3 + \ldots + b_{1n} y_n + a_{11} x_1 + \ldots + a_{1m} x_m + \varepsilon_1 \\ y_2 = c_{20} + b_{21} y_1 + b_{23} y_3 + \ldots + b_{2n} y_n + a_{21} x_1 + \ldots + a_{2m} x_m + \varepsilon_2 \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ y_n = c_{n0} + b_{n1} y_1 + b_{n2} y_2 + \ldots + b_{nn-1} y_{n-1} + a_{n1} x_1 + \ldots + a_{nm} x_m + \varepsilon_n \end{cases}$$

Содержание параметров структурной формы модели

Параметр (структурный коэффициент модели)	**Содержание параметра**
c_{i0}, i = 1, 2,…, n	Свободный член уравнения модели
b_{ij}, i = 1, 2,…, n, j = 1, 2,…, m	Коэффициент при эндогенной переменной
a_{ij}, i = 1, 2,…, n, j = 1, 2,…, m	Коэффициент при экзогенной переменной

ε_i (i = 1, 2,…, n) является случайной составляющей (ошибкой) i-го уравнения структурной формы модели.

Если в структурной форме модели переменные y_i и x_j (i = 1, 2,…, n , j = 1, 2,…, m) являются отклонения ми от среднего уровня $\bar{y}$ и $\bar{x}$ соответственно, то в каждом уравнении системы свободный член не записывается.

Структурная форма модели отражает реальный экономический объект или явление и показывает, как изменение любой экзогенной переменной определяет значения эндогенных переменных

Наряду с регрессионными уравнениями в модели могут быть записаны и тождества. Таким образом, структурные уравнения модели разделятся на два класса.

Классы структурных уравнений модели

Структурные уравнения модели	
Поведенческие уравнения	**Тождества**
Описывают взаимодействие между экзогенными и эндогенными переменными	Устанавливают соотношение между эндогенными переменными, не содержат случайных составляющих и структурных коэффициентов модели

Структурная форма модели может быть преобразована в ***приведенную форму***.

$$\begin{cases} y_1 = \alpha_{10} + \alpha_{11} x_1 + \alpha_{12} x_2 + \dots + \alpha_{1m} x_m + \eta_1 \\ y_2 = \alpha_{20} + \alpha_{21} x_1 + \alpha_{22} x_2 + \dots + \alpha_{2m} x_m + \eta_2 \\ \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots \\ y_n = \alpha_{n0} + \alpha_{n1} x_1 + \alpha_{n2} x_2 + \dots + \alpha_{nm} x_m + \eta_n \end{cases}$$

Содержание параметров приведенной формы модели

Параметр (коэффициент приведенной формы модели)	**Содержание параметра**
α_{i0} , i = 1, 2,…, n	Свободный член уравнения системы
α_{ij} , i = 1, 2,…, n , j = 1, 2,…, m	Коэффициент при предопределенной переменной является функцией коэффициентов структурной формы модели

η_i (i = 1, 2,…, n) - случайная составляющая (ошибка) i-го уравнения приведенной формы модели.

Под ***предопределенной переменной*** системы одновременных уравнений понимают экзогенные и лаговые (за предыдущие моменты времени) эндогенные переменные этой системы.

Причины, по которым наряду со структурной формой модели строят ее приведенную форму
Оценки параметров структурной формы модели, найденные с помощью МНК, являются смещенными и несостоятельными (нарушаются предпосылки МНК) в силу того, что эндогенные переменные, как правило, коррелируют со случайным отклонением
Независимость уравнений в приведенной форме модели позволяет определять состоятельные оценки ее параметров с помощью МНК
Параметры (коэффициенты) приведенной формы модели связаны с параметрами ее структурной формы

2. Применение систем одновременных уравнений

Применение систем одновременных уравнений проходит во взаимосвязи с экономической теорией. В настоящее время наиболее разработаны следующие направления: исследование спроса и предложения, макроэкономическое моделирование механизмов функционирования экономики на примере конкретной страны, анализ функций издержек и производственных функций.

Модели спроса и предложения - классические примеры систем одновременных уравнений. Выбор переменных, оценка параметров уравнений системы отображают степень взаимного влияния признаков модели.

Модель 1. Предложение и спрос на рынке.

$$\begin{cases} y_{1t} = c_{10} + b_{13} y_{3t} + \varepsilon_1 & (1) \\ y_{2t} = c_{20} + b_{23} y_{3t} + \varepsilon_2 & (2) \\ y_{1t} = y_{2t} & (3) \end{cases},$$

где y_{1t} - спрос на товар в момент времени t;

y_{2t} - предложение количества товара в момент t;

y_{3t} - цена, по которой заключаются сделки в момент t.

Все переменные системы y_{1t} , y_{2t} , y_{3t} эндогенные в силу экономического содержания модели. Величины спрос-предложение и цена определяются одновременно. Так как $y_{1t} = y_{2t}$, то (1) = (2).

Выполним преобразования: $c_{10} + b_{13} y_{3t} + \varepsilon_1 = c_{20} + b_{23} y_{3t} + \varepsilon_2$,

$(b_{13} - b_{23}) y_{3t} = c_{20} - c_{10} + \varepsilon_2 - \varepsilon_1$, $y_{3t} = \dfrac{c_{20} - c_{10} + \varepsilon_2 - \varepsilon_1}{(b_{13} - b_{23})}$. Последнее равенство подтверждает зависимый характер цены.

Для того чтобы модель 1 была эконометрически значимой, необходимо ввести предопределенные (экзогенные и лаговые эндогенные) переменные. Например, в ***уравнение спроса*** (1) ввести экзогенную переменную x_{1t} - доход на душу населения в момент времени t, и в ***уравнение предложения*** (2) - эндогенную лаговую переменную $y_{3,t-1}$ - цену товара в момент (t - 1). Получим модель 2 спроса и предложения ***кейнсианского типа***.

Модель 2. Предложение и спрос кейнсианского типа.

$$\begin{cases} y_{1t} = c_{10} + b_{13} y_{3t} + a_{11} x_{1t} + \varepsilon_1 & (1) \\ y_{2t} = c_{20} + b_{23} y_{3t} + a_{23} y_{3,t-1} + \varepsilon_2 & (2) \\ y_{1t} = y_{2t} & (3) \end{cases},$$

где y_{1t} - спрос на товар в момент времени t ;

y_{2t} - предложение товара в момент t ;

y_{3t} - цена товара в момент t ;

$y_{3,t-1}$ - цена товара в момент t-1 ;

x_{1t} - доход в момент t ;

t - текущий период;

(t-1) - предыдущий период.

В модели три эндогенные переменные (y_{1t}, y_{2t}, y_{3t}) и две предопределенные переменные (x_{1t}, $y_{3,t-1}$).

3. Задача идентификации уравнений системы. Необходимое и достаточное условие идентифицируемости

Решение проблемы идентификации

Переход от приведенной формы модели к ее структурной форме связан с решением проблемы идентификации.

Понятие идентификации

Установление соответствия между приведенной и структурной формами модели

Единственность соответствия между приведенной и структурной формами модели и составляет задачу идентификации. В зависимости от условий определения структурных коэффициентов модели по приведенным коэффициентам любая структурная модель может быть отнесена к одному из трех классов; идентифицируемая, неидентифицируемая и сверхидентифицируемая.

Классы структурных моделей с точки зрения задачи идентификации

Идентифицируемая	Неидентифицируемая	Сверхидентифицируемая
Все структурные коэффициенты однозначно определяются через приведенные коэффициенты	Структурные коэффициенты невозможно найти по приведенным коэффициентам	Структурные коэффициенты, выраженные через приведенные коэффициенты, имеют два и более числовых значений

Установление неидентифицируемости (сверхидентифицируемости) модели

В идентифицируемой модели количество структурных и приведенных коэффициентов одинаково. Если структурных коэффициентов больше (меньше), чем приведенных, то модель соответственно неидентифицируема (сверхидентифицируема).

Идентифицируемая модель	$\leftrightarrow$	***Идентифицируемо каждое уравнение системы***
Сверхидентифицируемая модель	$\leftrightarrow$	***Сверхидентифицируемо хотя бы одно уравнение системы***

Проверка структурной модели на идентифицируемость позволяет установить степень возможности оценивания коэффициентов структурных уравнений по коэффициентам приведенных уравнений.

Необходимое и достаточное условие идентифицируемости уравнения системы

Необходимое условие $n = p + 1$	**Достаточное условие** $\Delta^* \neq 0$, rang $M^* = n - 1$
Уравнение модели идентифицируемо, если количество (n) эндогенных переменных этого уравнения на единицу больше количества (p) предопределенных переменных системы, не входящих в данное уравнение	Если определитель (Δ^*) матрицы коэффициентов (M^*) при переменных системы, не входящих в данное уравнение, не равен нулю и количество эндогенных переменных системы без единицы равно рангу этой матрицы, то уравнение модели идентифицируемо

Если выполнимо условие:

$n < p + 1$, то уравнение сверхидеитифицируемо;

n > p + 1, то уравнение неидентифицируемо.

Проверка структурной модели на идентифицируемость позволяет установить степень возможности оценки коэффициентов структурных уравнений по коэффициентам приведенных уравнений.

Применив соответствующие статистические данные, можно с помощью косвенного МНК найти несмещенные и состоятельные оценки структурной формы, тем самым смоделировав реальную экономическую ситуацию изучения спроса-предложения с учетом дохода в текущий период и цены товара в предыдущий период.

Каждое уравнение системы оценивают тогда и только тогда, когда установлена его идентифицируемость.

Отметим, что идентификация не применяется к тождествам модели.

Рассмотрим модель 3 предложения и спроса на деньги с очки зрения ее экономической разрешимости.

Модель 3. Предложение денег и спрос на деньги.

$$\begin{cases} y_{1t} = c_{10} + b_{12} y_{2t} + a_{11} x_{1t} + \varepsilon_1 & (1) \\ y_{2t} = c_{20} + b_{21} y_{1t} + \varepsilon_2 & (2) \end{cases},$$

где y_{1t} – процентные ставки в период t ;

y_{2t} – ВВП в период t ;

x_{1t} – денежная масса в период t .

Уравнения	**Переменные**		
	эндогенные		**предопределенные**
	y_{1t}	y_{2t}	x_{1t}
(1)	-1	b_{12}	a_{11}
(2)	b_{21}	-1	0

***Уравнение* (1)**:

а) n = 2 , p = 0 , n > p + 1

Уравнение неидентифицируемое, следовательно, неидентифицируема вся система.

Методы решения систем одновременных уравнений

Для получения качественных оценок параметров системы одновременных уравнений пользуются специальными методами. Выбор метода определяется условиями системы. Существенна также относительная простота алгоритма самого метода. В настоящее время классическими для решения систем одновременных уравнений являются ***косвенный МНК*** и ***двухшаговый МНК.***

4. Косвенный метод наименьших квадратов.

Косвенный МНК основан на получении состоятельных и несмещенных оценок параметров структурной формы модели по оценкам параметров приведенной формы. Последние являются состоятельными и несмещенными в силу применения к каждому уравнению приведенной формы МНК.

Алгоритм применения косвенного метода наименьших квадратов

Оценить параметры системы одновременных уравнений, которая задана структурной формой модели
1. Структурная форма модели преобразуется в приведенную форму
2. С помощью МНК оцениваются параметры приведенной формы
3. Приведенная форма преобразуется в структурную форму
Несмещенные и состоятельные оценки параметров структурной формы получены

Область применения косвенного МНК ограничивается идентифицируемыми системами одновременных уравнений

5. Двухшаговый метод наименьших квадратов

Двухшаговый МНК применяется как для идентифицируемых, так и для сверхидентифицируемых систем одновременных уравнений. В этом смысле метод является общим по отношению к косвенному МНК.

Алгоритм применения двухшагового метода наименьших квадратов

Оценить параметры сверхидентифицируемой системы одновременных уравнений, которая задана структурной формой модели
1. Структурная форма модели преобразуется в приведенную форму
2. С помощью МНК оцениваются параметры приведенной формы
3. В правой части сверхидентифиируемого уравнения структурной модели выбираются эндогенные переменные и рассчитываются их теоретические значения по соответствующим приведенным уравнениям
4. С помощью МНК на основе фактических значений предопределенных и теоретических значений эндогенных переменных оцениваются параметры сверхидентифицируемого уравнения структурной модели
Несмещенные и состоятельные оценки параметров структурной формы получены

Двухшаговый МНК обладает свойствами, благодаря которым его практическая эффективность остается достаточно высокой. Сформулируем эти свойства:

1) для двухшагового МНК достаточно оперировать экзогенными и предопределенными переменными модели;

2) эффективность двухшагового МНК определяется высоким коэффициентом детерминации R^2 приведенных уравнений модели. В том случае, когда R^2 низкий, расчетные значения эндогенной переменной слабо аппроксимируют ее фактические значения.

Тема 10. ГЕТЕРОСКЕДАСТИЧНОСТЬ ОСТАТКОВ

Лекция

План лекции

1. Понятие гетероскедастичности
2. Обнаружение гетероскедастичности
3. Методы смягчения гетероскедастичности

1. Понятие гетероскедастичности

Для получения качественных оценок параметров уравнения регрессии необходимо следить за выполнимостью предпосылок МНК. Применяя МНК мы предполагаем, что остатки ε_i подчиняются условиям Гаусса-Маркова, данное предположение необходимо проверить, после построения уравнения регрессии.

Допущение о постоянстве дисперсии остатков $\left(D(\varepsilon_i) = \sigma^2\right)$ известно как допущение о ***гомоскедастичности***. Если это допущение нарушено, и дисперсия остатков не является постоянной, то говорят, что оценки ***гетероскедастичны***.

На практике, для каждого i-го наблюдения определяется единственное значение ε_i, но мы говорим об определении дисперсии остатков, т.е. о множестве ε_i для каждого i-го наблюдения. Это объясняется тем, что мы имеем дело с выборочной совокупностью, а априори ε_i могли принимать любые значения на основе некоторых вероятностных распределений.

Гетероскедастичность приводит к тому, что коэффициенты регрессии не являются оценками с минимальной дисперсией, следовательно, они больше не являются наиболее эффективными коэффициентами. Вследствие, выводы, получаемые на основе t и F-статистик, а также интервальные оценки будут ненадежными. Дисперсии и, следовательно, стандартные ошибки этих коэффициентов будут смещенными.

Если смещение отрицательно, то оценочные стандартные ошибки будут меньше, чем они должны быть, а критерий проверки - больше чем в реальности. Таким образом, можно сделать вывод, что коэффициент значим, когда он таковым не является.

И наоборот если смещение положительно, то оценочные ошибки будут больше чем они должны быть, а критерии проверки - меньше. Значит, возможно, ошибочное принятие нулевой гипотезы.

2. Обнаружение гетероскедастичности

Существует несколько формальных тестов, позволяющих обнаружить гетероскедастичность (графический анализ остатков, тест ранговой корреляции Спирмена, тест Парка, тест Голфелда-Квандта, тест Уайта).

2.1. Графический анализ остатков

Использование графического представления отклонений позволяет определиться с наличием гетероскедастичности. В этом случае по оси абсцисс откладываются значения x_i объясняющей переменной X (либо линейной комбинации объясняющих переменных

$Y = b_0 + b_1 X_1 + ... + b_m X_m,)$,

а по оси ординат - либо отклонения ε_i либо их квадраты ε_i^2 , i = 1, 2, ..., *n*.

Если все отклонения ε_i^2 находятся внутри полуполосы постоянной ширины, параллельной оси абсцисс, это говорит о независимости дисперсий u_i^2 от значений переменной X и их постоянстве, т.е. в этом случае выполняются условия гомоскедастичности.

Графический анализ отклонений является удобным и достаточно надежным в случае парной регрессии.

Обычно не ограничиваются визуальной проверкой гетероскедастичности, а проводят ее эмпирическое подтверждение.

2.2. Тест ранговой корреляции Спирмена

При использовании данного теста предполагается, что дисперсия отклонения будет либо увеличиваться, либо уменьшаться с увеличением значений X. Поэтому для регрессии, построенной по МНК, абсолютные величины отклоненийεiи значения x_i будут коррелированны.

Значения x_i и ε_i ранжируются (упорядочиваются по величинам). Затем определяется коэффициент ранговой корреляции:

$$r_{x,\varepsilon} = 1 - 6 \cdot \frac{\sum d_i^2}{n(n^2 - 1)}$$

где d_i - разность между рангами x_i и ε_i, i = 1, 2, ..., *n; n* - число наблюдений.

Например, если x_{20} является 15 по величине среди всех наблюдений, а ε_{20} - 21, то d_{20} = 15-21 = - 6.

Если коэффициент корреляции $\rho_{x,\varepsilon}$ для генеральной совокупности равен нулю, то статистика

$$t = \frac{r_{x,\varepsilon}\sqrt{n-2}}{\sqrt{1 - r_{x,\varepsilon}^2}}$$

имеет распределение Стьюдента с числом степеней свободы v = n-2.
Следовательно, если наблюдаемое значение t-статистики превышает табличное, то необходимо отклонить гипотезу о равенстве нулю коэффициента корреляции $\rho_{x,u}$, а, следовательно, и об отсутствии гетероскедастичности.

Если в модели регрессии больше чем одна объясняющая переменная, то проверка гипотезы может осуществляться с помощью t-статистики для каждой из них отдельно.

2.3. Тест Голдфелда-Квандта

Самым популярным тестом обнаружения гетероскедастичности является тест, предложенный С. Голдфелдом и Р. Квандтом.

В данном случае также предполагается, что стандартное отклонение $\sigma_i = \sigma(\varepsilon_i)$ пропорционально значению x_i переменной X в этом наблюдении, т, е.

$$\sigma_i^2 = \sigma^2 x_i^2,\ i = 1,2,...,n.$$

Предполагается, что ε_i имеет нормальное распределение и отсутствует автокорреляция остатков.

Тест Голдфелда-Квандта состоит в следующем:

1. Все *n* наблюдений упорядочиваются по величине X.

2. Вся упорядоченная выборка после этого разбивается на три подвыборки размерностей k, (*n*-2k), k соответственно.

3. Оцениваются отдельные регрессии для первой подвыборки (*k* первых наблюдений) и для третьей подвыборки (*k* последних наблюдений). Если предположение о пропорциональности дисперсий отклонений значениям X верно, то дисперсия регрессии по первой подвыборке. Сумма квадратов отклонений

$$S_1 = \sum_{i=1}^{k} \varepsilon_i^2$$

будет существенно меньше дисперсии регрессии по третьей подвыборке. Суммы квадратов отклонений

$$S_3 = \sum_{i=n-k-1}^{n} \varepsilon_i^2 .$$

4. Для сравнения соответствующих дисперсий строится следующая F-статистика:

$$F = \frac{S_3/(k-m-1)}{S_1/(k-m-1)} = \frac{S_3}{S_1}.$$

При сделанных предположениях относительно случайных отклонений, построенная F-статистика имеет распределение Фишера с числами степеней свободы $v_1 = v_2 = k\text{-}m\text{-}l$.

Если

$$F_{набл} = \frac{S_3}{S_1} > F_{кр} = F_{\alpha;v_1;v_2},\ если\ \ S_3 > S_1$$

то гипотеза об отсутствии гетероскедастичности отклоняется (α - выбранный уровень значимости).

2.4. Тест Уайта (White test, 1980).

Если в модели присутствует гетероскедастичность, то очень часто это связано с тем, что дисперсии ошибок некоторым образом зависят от регрессоров, а гетероскедастичность отражается в остатках обычной регрессии исходной модели.

Проводится этот тест следующим образом:

1) допустим, исходная модель имеет вид:

$$y_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + \varepsilon_i.$$

МНК оцениваются ее параметры и получают регрессионные остатки ε_i;

2) оценивается вспомогательная регрессия квадратов остатков на все регрессоры, их квадраты, попарные произведения и константу:

$$\overset{\cup}{\varepsilon}^2 = \alpha_0 + \alpha_1 x_{1i} + \alpha_2 x_{2i} + \alpha_3 x_{1i}^2 + \alpha_4 x_{2i}^2 + \alpha_5 x_{1i} x_{2i} + \nu_i,$$

где ν_i - нормально распределенная ошибка, независимая от ε_i.

Напомним, что $D(\varepsilon_i) = M(\varepsilon_i - M(\varepsilon_i))^2$. Однако поскольку предполагается, что $M(\varepsilon) = 0$, то $D(\varepsilon_i) = M(\varepsilon_i^2)$.Так как нам неизвестна истинная величина квадратов остатков ε_i^2, то вопрос о наличии гетероскедастичности решается на основе их выборочных аналогов, $\overset{\cup}{\varepsilon}^2$.

Вспомогательная регрессия имеет именно такую форму, потому что необходимо исследовать, существует ли систематическая зависимость между изменениями $\overset{\cup}{\varepsilon}^2$ и какой-либо релевантной переменной модели (чтобы увидеть, что релевантными являются именно переменные, включенные во вспомогательную регрессию, следует представить ошибку в виде $\varepsilon_i = y_i - b_0 - b_1 x_{1i} - b_2 x_{2i}$ и возвести данное выражение в квадрат).

3) Проверяется нулевая гипотеза:

H_0: $\alpha_2 = 0$ и $\alpha_3 = 0$ и $\alpha_4 = 0$ и $\alpha_5 = 0$ и $\alpha_6 = 0$.

с помощью F– критерия Фишера.

Если фактические значения статистики превышают критические величины распределения $F_{расч} > F_{кр}(\alpha, v_1=p, v_2=n-p-1)$ то нулевая гипотеза о гомоскедастичности остатков отвергается, то есть делается вывод о присутствии гетероскедастичности.

3. Методы смягчения гетероскедастичности

Если известна причина (и, соответственно, форма гетероскедастичности), то для ее устранения можно воспользоваться ***обобщенным методом наименьших квадратов*** (ОМНК).

Предположим, что дисперсия регрессионных остатков связана с некоторой переменной z_t зависимостью вида

$$D(\varepsilon_i) = \sigma^2 z_i^2.$$

В качестве переменной может быть:

1) среднее квадратическое отклонение σ_i (если она известна), в этом случае получают взвешенный метод наименьших квадратов (ВМНК);

2) x_i или $\sqrt{x_i}$, т.е. дисперсия остатков пропорциональна либо x_i либо x_i^2 .

Для того чтобы избавиться от гетероскедастичности, необходимо разделить каждый член регрессионного уравнения

$$\frac{y_i}{z_i} = b_0 \frac{1}{z_i} + b_1 \frac{x_{1i}}{z_i} + b_2 \frac{x_{2i}}{z_i} + \nu_i, \quad (*)$$

где $\nu_i = \frac{\varepsilon_i}{z_i}$ случайная ошибка.

Поскольку $D(\varepsilon_i) = \sigma^2 z_i^2$, то $D(\nu_i) = D\left(\frac{\varepsilon_t}{z_i}\right) = \frac{D(\varepsilon_i)}{z_i^2} = \frac{\sigma^2 z_i^2}{z_i^2} = \sigma^2.$

Таким образом, ошибки в уравнении (*) будут гомоскедастичными.

Однако на практике часто не удается с уверенностью определить причину и форму гетероскедастичности. В этом случае можно либо перевести все переменные в логарифмическую форму (однако необходимо помнить, что этот прием неприменим, если переменные модели могут принимать нулевые или отрицательные значения), либо воспользоваться специальными робастными методами оценки.

Тема 11. МУЛЬТИКОЛЛИНЕАРНОСТЬ

Лекция

План лекции

1. Понятие мультиколлинеарности
2. Последствия мультиколлинеарности
3. Установление наличия мультиколлинеарности
4. Причины возникновения мультиколлинеарности
5. Методы устранения мультиколлинеарности

1. Понятие мультиколлинеарности

Мультиколлинеарность – это нарушение требования **С** теоремы Гаусса-Маркова о линейной независимости факторных переменных $x_1, x_2, \ldots, x_m$, включенных в модель. В этом случае столбцы матрицы наблюдений линейно зависимы.

Наличие линейно зависимых объясняющих переменных в модели множественной регрессии вызывает слабую обусловленность системы нормальных уравнений, что приводит к отрицательным последствиям.

2. Последствия мультиколлинеарности

1. t-критерий Стьюдента о незначимости коэффициентов регрессии подтверждается, однако само уравнение регрессии при проверке F-критерия Фишера оказывается значимым
2. Полученные оценки коэффициентов регрессии неоправданно завышены или имеют неправильные знаки
3. Добавление или исключение из исходных данных одного-двух наблюдений оказывает сильное влияние на оценки коэффициентов регрессии
4. Регрессионная модель становится непригодной для дальнейшего применения: изменяется смысл экономической интерпретации коэффициентов регрессии, построенный по уравнению регрессии прогноз имеет большую ошибку, осложняется определение наиболее существенных факторных признаков

В решении проблемы мультиколлинеарности можно выделить три этапа.

Этапы решения проблемы мультиколлинерности		
1	**2**	**3**
Установление наличия мультиколлинеарности	Определение причин возникновения мультиколлинеарности	Устранение мультиколлинеарности

3. Установление наличия мультиколлинеарности

1. Способ. Анализ корреляционной матрицы. Независимые переменные x_i, и x_j могут быть признаны коллинеарными, если парный коэффициент корреляции $r_{x1,x2} > 0{,}75$.

Если собственное число корреляционной матрицы факторных переменных $\lambda_{\min} < 10^{-5}$, то в модели присутствует мультиколлинеарность.

Если отношение собственных чисел корреляционной матрицы факторных переменных $\frac{\lambda_{\min}}{\lambda_{\max}} < 10^{-5}$, то в модели присутствует мультиколлинеарность.

2. Способ. Если определитель матрицы X'X близок к нулю, то это свидетельствует о наличии мультиколлинеарности.

4. Причины возникновения мультиколлинеарности

Основная причина мультиколлинеарности заключается в неправильном подборе факторных переменных x_1, x_2,…, x_m , включенных в модель.

Изучаемые факторные признаки характеризуют одну и ту же сторону явления или процесса (например, показатели объема произведенной продукции и среднегодовой стоимости основных фондов одновременно включать в модель не рекомендуется, так как оба характеризуют размер предприятия)
Использование в качестве факторных признаков, суммарное значение которых представляет собой постоянную величину (например, коэффициент годности и коэффициент износа основных фондов)
Факторные признаки являются элементами друг друга (например, затраты на производство продукции и себестоимость единицы продукции)
Факторные признаки по экономическому смыслу дублируют друг друга (например, прибыль и рентабельность продукции)

Если при прогнозировании результативной переменной величина ошибки прогноза является удовлетворительной, то модель множественной регрессии можно использовать и при наличии мультиколлинеарности. Если же прогноз получается неудовлетворительным, то мультиколлинеарность необходимо устранять.

5. Методы устранения мультиколлинеарности

Самый простой способ устранения мультиколлинеарности - **сбор дополнительных данных**, однако на практике это не всегда возможно.

Устранение (уменьшение) мультиколлинеарности возможно посредством исключения из регрессионной модели одного или нескольких линейно связанных факторных признаков или преобразование исходных факторных признаков в новые, укрупненные факторы. Вопрос о том, какой из факторов следует отбросить, решается на основе качественного и логического анализа изучаемого явления.

Описание методов устранения или уменьшения мультиколлинеарности

Метод	Суть метода
Методы преобразования переменных	Замена всех переменных, включенных в модель. Например, вместо значений зависимой и независимых переменных можно взять их логарифмы. Тогда модель множественной регрессии примет вид $\ln y = \beta_0 + \beta_1 \ln x_1 + \beta_2 \ln x_2 + ... + \beta_m \ln x_m + \varepsilon$.
	Если переменные x_k, и x_j коллинеарны, и ни одну из них нельзя исключить из модели, вместо них вводят новую переменную $x_{kj} = x_k \cdot x_j$

Анализ корреляционной матрицы	При отборе факторов предпочтение отдается тому фактору, который более тесно, чем другие факторы, связан с результативным признаком, причем желательно, чтобы связь данного факторного признака с у была выше, чем его связь с другим факторным признаком, т.е. $r_{yxj} > r_{xjxk}$, $r_{yxk} > r_{xjxk}$ и $r_{xjxk} < 0{,}75$
	На основе анализа матрицы межфакторных корреляций: - выявляются группы объясняющих переменных, для которых парные коэффициенты корреляции по абсолютной величине превосходят 0,75; - в нутрии группы поочередно каждую переменную объявляют зависимой и для нее рассчитывают коэффициент детерминации; - переменная, имеющая наибольший коэффициент детерминации, отвечает за мультиколлинеарность и должна быть исключена из набора объясняющих переменных.
Пошаговое включение факторов	Метод заключается в том, что в модель включаются факторы по одному в определенной последовательности. На первом шаге в модель вводится тот фактор, который имеет наибольший коэффициент корреляции с зависимой переменной. На втором и последующих шагах в модель включается фактор, который имеет наибольший коэффициент корреляции с остатками модели. После включения каждого фактора в модель рассчитывают ее характеристики и модель проверяют на достоверность. Построение модели заканчивается, если модель перестает удовлетворять определенным условиям, например: 1. $k < n/3$, $S_{\varepsilon,k-1} - S_{\varepsilon,k} > l$, где n - число наблюдений; k - число факторных признаков, включаемых в модель; l - некоторое заданное малое число; $S_{\varepsilon,k}$ - среднеквадратическая ошибка уравнения регрессии; $S_{\varepsilon,k-1}$, - среднеквадратическая ошибка, полученная на предыдущем шаге и включающая k - 1 переменных). или 2. При добавлении в модель регрессии новых факторных переменных их значимость проверяется с помощью F-критерия Фишера. Если $F_{набл} > F_{крит}$, то включение факторной переменной в модель множественной регрессии является обоснованным. Проверка факторных переменных на значимость осуществляется до тех пор, пока не найдется хотя бы одна переменная, для которой не выполняется условие $F_{набл} < F_{крит}$.
Пошаговое исключение факторов	Метод состоит в том, что в модель включаются все факторы. Затем после построения уравнения регрессии из модели исключают фактор, коэффициент при котором незначим и имеет наименьшее значение t-критерия. После этого получают новое уравнение регрессии и снова проводят оценку значимости всех оставшихся коэффициентов регрессии. Процесс исключения факторов продолжается до тех пор, пока модель не станет удовлетворять определенным условиям и все коэффициенты регрессии не будут значимы.

Существуют другие методы устранения мультиколлинеарности, более трудоемкие и требующие хорошего знания многомерного статистического анализа: гребневая регрессия (ридж), метод главных компонент и т.п.