КАЗАНСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ
ИНСТИТУТ УПРАВЛЕНИЯ, ЭКОНОМИКИ И ФИНАНСОВ
Кафедра экономико-математического моделирования

И. И. ИСМАГИЛОВ, Е. И. КАДОЧНИКОВА,
А. В. КОСТРОМИН

ЭКОНОМЕТРИКА

Конспект лекций

Казань – 2014

УДК 330.43

ББК Ув631я73-1

Принято на заседании кафедры статистики, эконометрики и естествознания

Протокол № 10 от 24 июня 2014 года

Рецензенты:
кандидат экономических наук,
доцент кафедры статистики, эконометрики и
естествознания КФУ **Е. Л. Фесина;**
кандидат технических наук,
доцент кафедры статистики, эконометрики и
естествознания КФУ **Д. М. Мухаметгалеев**

Исмагилов И. И., Кадочникова Е. И., Костромин А. В.
Эконометрика / И. И. Исмагилов, Е. И. Кадочникова, А. В. Костромин. – Казань: Казан. ун-т, 2014. – 235 с.

В современное время особенно актуальным является обучение студентов теоретическим основам эконометрической методологии и практическим навыкам применения эконометрических методов для исследования экономических закономерностей и взаимосвязей между экономическими переменными. В круг основных задач дисциплины «Эконометрика» входят: получение теоретических знаний об эконометрических методах эмпирического анализа экономических процессов с целью имитации альтернативных сценариев развития анализируемой системы; формирование умения и навыков выбирать и применять необходимые инструменты эконометрического анализа для обоснования управленческих решений.

Настоящий конспект лекций адресован студентам для самостоятельного изучения, с выполнением предлагаемых заданий и самоконтролем усвоения материала.

Оглавление

Лекция 1

Тема 1. Эконометрика как научная дисциплина

Вопросы для изучения:

1. Цели, предмет, задачи эконометрики.
2. Инструментарий эконометрики. Типы моделей и переменных.
3. Этапы эконометрического моделирования

Аннотация. Данная тема раскрывает основные понятия эконометрики.

Ключевые слова. Модели, переменные, типы данных, этапы моделирования.

Методические рекомендации по изучению темы

- Изучить лекционную часть, где даются общие представления по теме.
- Для закрепления теоретического материала ознакомиться с решениями типовых задач и ответить на вопросы для самоконтроля.
- Для проверки усвоения темы выполнить тест для самоконтроля.

Рекомендуемые информационные ресурсы:

1. http://tulpar.kfu.ru/course/view.php?id=2213

2. Эконометрика: [Электронный ресурс] Учеб.пособие / А.И. Новиков. - 2-е изд., испр. и доп. - М.: ИНФРА-М, 2011. - 144 с.: (http://znanium.com/catalog.php?item=booksearch&code=%D1%8D%D0%BA%D0%BE%D0%BD%D0%BE%D0%BC%D0%B5%D1%82%D1%80%D0%B8%D0%BA%D0%B0&page=1#none)С. 4-7.

3. Эконометрика: учебник / И. И. Елисеева. – М.: Проспект, 2010. – 288 с. С. 6-11.

Цели, предмет, задачи эконометрики

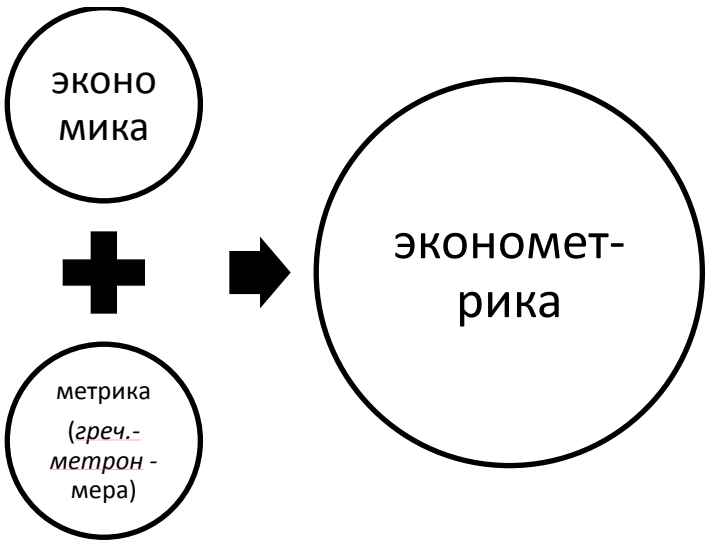


Рис. 1.1. Термин «эконометрика»

Эконометрика – это наука, в которой на базе реальных статистических данных строятся, анализируются и совершенствуются математические модели реальных экономических явлений. Эконометрика позволяет найти количественное подтверждение либо опровержение того или иного экономического закона либо гипотезы.

Термин «эконометрика» впервые был использован бухгалтером П. Цьемпой в Австро-Венгрии, в 1910 году.

1912 г. – И. Фишер сделал безуспешную попытку создать группу ученых для стимулирования развития экономической теории путем ее связи со статистикой и математикой.

1930 г., 29 декабря – на заседании Американской ассоциации развития науки по инициативе И. Фишера, Й. Шумпетера, О. Андерсона , Я. Тинбергена создано эконометрическое общество, на котором норвежский ученый Р. Фриш дал новой науке название «эконометрика».

1933 г. – стал издаваться журнал «Econometrica».

1941 г. – издан первый учебник по эконометрике, автор Я. Тинберген.

1970 – е гг. – противоречия между кейнсианцами, монетаристами и марксистами привели к тому, что методы эконометрики стали применяться не только для оценки теоретических моделей, но и для доказательства причинности при вы-

боре теоретических концепций. Появление компьютеров, создание ARIMA-моделей, VAR-моделей, развитие анализа временных рядов.

Определения эконометрики:

«Эконометрика – это наука, которая дает количественное выражение взаимосвязей экономических явлений и процессов, которые раскрыты и обоснованы экономической теорией» (И.И. Елисеева, см. 1, стр. 16).

«Эконометрика – это наука, которая на базе экономической теории, экономической статистики, экономических измерений и математико-статистического инструментария придает количественное выражение качественным закономерностям, обусловленным экономической теорией» (С. А. Айвазян, см. 3, стр. 12).

«Эконометрика – это наука, связанная с эмпирическим выводом экономических законов» (Магнус Я. Р., см. 6., стр. 13).

«Эконометрика – это наука, в которой на базе реальных статистических данных строятся, анализируются и совершенствуются математические модели реальных экономических явлений» (С. А. Бородич, см. 2, стр. 7).

Зарождение эконометрики является следствием междисциплинарного подхода к изучению экономики. Эконометрика как научная дисциплина зародилась и получила развитие на основе слияния экономической теории, математической экономики и экономической и математической статистики.

«Эконометрика – это не то же самое, что экономическая статистика. Она не идентична и тому, что мы называем экономической теорией. Эконометрика не является синонимом приложений математики к экономике. Каждая из трех отправных точек – статистика, экономическая теория и математика – необходимое, но не достаточное условие для понимания количественных соотношений в современной экономической жизни. Это единство всех трех составляющих. И это единство образует эконометрику» (Р. Фриш, 1933 г., см. 1, стр. 16).

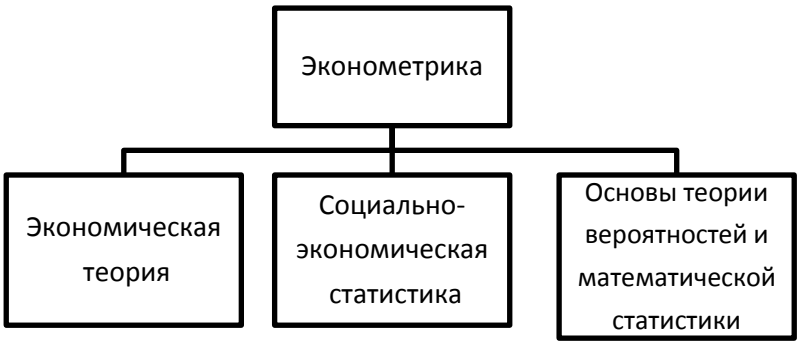


Рис. 1.2. Три составляющие эконометрики

Предметом эконометрики являются количественные закономерности между экономическими явлениями. Однако, в отличие от экономической теории, эконометрика делает упор на количественные, а не на качественные аспекты этих явлений. Например, известно, что спрос на товар с ростом его цены падает. Однако, как быстро и по какому закону это происходит, в экономической теории не определяется. Это в каждом конкретном случае делает эконометрика. С другой стороны, математическая экономика строит и анализирует модели экономических процессов без использования реальных числовых значений. Эконометрика же изучает модели на базе эмпирических данных. Наконец, в эконометрике широко используется аппарат математической статистики, особенно при установлении связей между экономическими показателями. В то же время в экономике невозможно проведение управляемого эксперимента, и эконометристы используют свои собственные приемы анализа, которые в математической статистике не встречаются.

Основными целями эконометрики являются:

1. Прогноз экономических и социально-экономических показателей, характеризующих состояние и развитие анализируемой системы.

2. Имитация различных возможных сценариев социально-экономического развития.

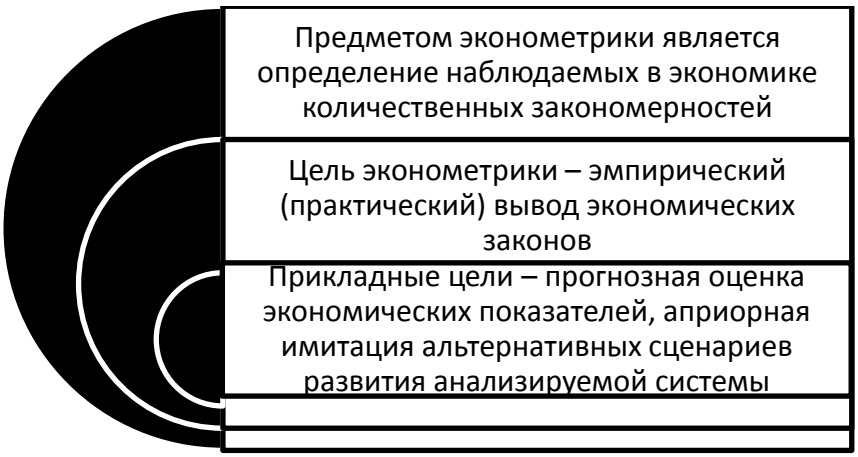


Рис. 1.3. Предмет и цели эконометрики

Основные задачи эконометрики:

- построение эконометрической модели;
- оценка параметров построенной модели, делающих выбранную модель наиболее адекватной реальным данным;
- проверка качества найденных параметров модели и самой модели в целом;
- использование построенных моделей для объяснения поведения исследуемых экономических показателей, прогнозирования, осмысленного проведения экономической политики (С. А. Бородич).

Инструментарий эконометрики. Типы моделей и переменных. Инструментарий эконометрики включает четыре основных раздела: линейная модель регрессии и МНК; обобщенная линейная модель регрессии и ОМНК; статистический анализ временных рядов; анализ систем одновременных уравнений.

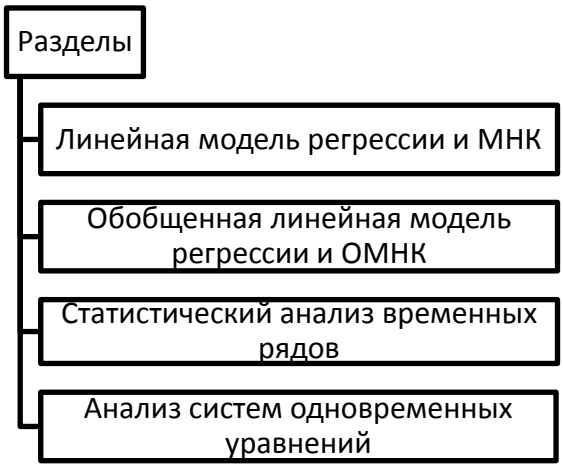


Рис. 1.4. Разделы эконометрики

Особенности эконометрического метода заключаются в следующем:

- исследование статистических зависимостей, а не функциональных;
- отражение особенностей экономических переменных и связей между ними (оптимальность и взаимодействие переменных);
- содержательное обоснование уравнений;
- изучение всей совокупности связей между переменными, а не изолированно взятого уравнения регрессии;
- развитие анализа временных рядов через решение проблем ложной корреляции, лага и других.

Для моделирования эконометрических взаимосвязей между экономическими явлениями чаще всего применяется три типа моделей и три типа переменных.

Рис. 1.5. Типы моделей

Рис. 1.6. Типы переменных

Рис. 1.7. Различия переменных

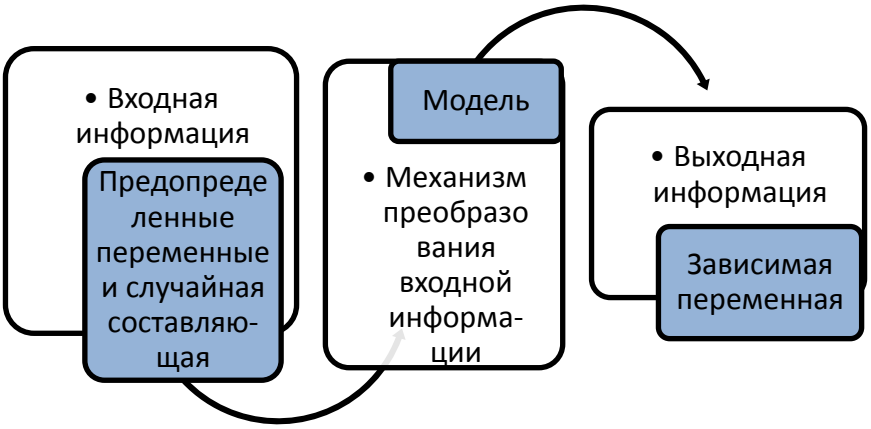


Рис. 1.8. Взаимодействие переменных

Приведем примеры эконометрических моделей.

Модели временных рядов:

- модель тренда: $Y_t = T_t + \varepsilon_t$

- модель сезонности: $Y_t = S_t + \varepsilon_t$

- модель тренда и сезонности: $Y_t = T_t + S_t + \varepsilon_t;$ $Y_t = T_t \cdot S_t \cdot \varepsilon_t$

Модель регрессии:

$$Y_x = f(x, \beta) = f(x_1, ..x_k, \beta_1 ... \beta_k)$$

Линейная модель множественной регрессии:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \varepsilon$$

Система одновременных уравнений:

$$\begin{cases} y_1 = b_{12} y_2 + a_{11} x_1 + a_{12} x_2 + a_{13} x_3 + \varepsilon_1, \\ y_2 = b_{21} y_1 + a_{21} x_1 + a_{22} x_2 + a_{23} x_3 + \varepsilon_2 \end{cases}$$

В эконометрике применяют три типа исходных данных.

Множество данных, состоящих из наблюдений за несколькими однотипными статистическими объектами в течение одного периода или за один момент времени, называется перекрестными данными.

Множество данных, состоящих из наблюдений за одним статистическим объектом в течение нескольких периодов или за несколько моментов времени, называется временным рядом.

Множество данных, состоящих из наблюдений за несколькими однотипными статистическими объектами в течение нескольких временных периодов, называется панельными, или пространственными, данными.

Этапы эконометрического моделирования. Этапы эконометрического моделирования:

1. постановочный – определение целей и задач модели;
2. априорный – предварительный анализ ситуации;
3. спецификация модели – выбор типа модели, состава переменных и формы математической связи между ними;
4. информационный – сбор первичной информации;
5. идентификация модели – оценивание параметров модели;
6. верификация модели– проверка качества модели в целом и ее параметров;
7. интерпретация результатов – формулирование выводов и рекомендаций на основе построенной модели.

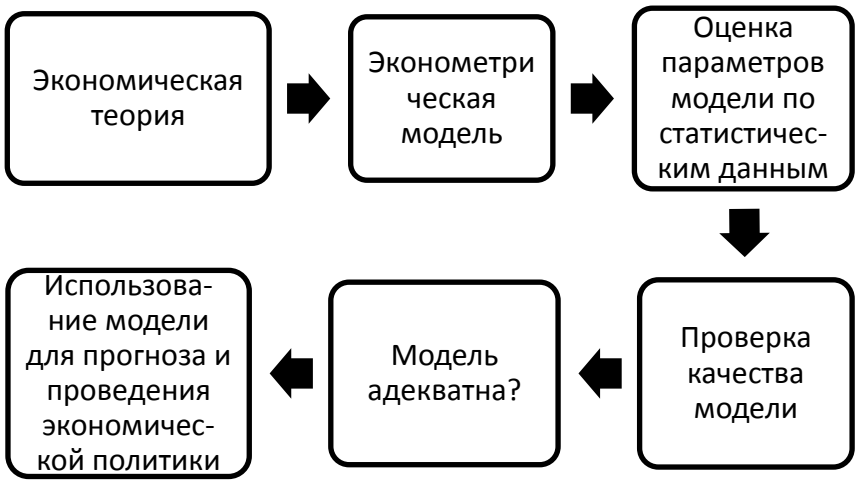


Рис. 1.9. Этапы эконометрического моделирования

Вопросы для самоконтроля

1. Что измеряет эконометрика?
2. Каковы основныецелиэконометрики?
3. В чем состоят предмет и задачи эконометрики?
4. Каковы типы моделей и переменных, применяемых в эконометрике?
5. В чем особенности перекрестных и панельных данных?
6. В чем особенности временных рядов?
7. Что понимается под спецификацией модели?
8. Что такое параметризация?
9. Что понимается под верификацией модели?
10. В чем основное отличие эконометрической модели от математической?

Лекция 2

Тема 2. Основные понятия теории вероятностей и статистики, применяемые в эконометрике

Вопросы для изучения

1. Основные понятия теории вероятностей. Нормальное распределение и связанные с ним $\chi 2$ - распределение, распределение Стьюдента и Фишера.
2. Генеральная совокупность и выборка. Свойства статистических оценок.
3. Статистические выводы и проверка гипотез.

Аннотация. Данная тема раскрывает основные понятия теории вероятностей и статистики, применяемые в эконометрике.

Ключевые слова. Генеральная совокупность, выборка, статистическая оценка, гипотеза.

Методические рекомендации по изучению темы

•Изучить лекционную часть, где даются общие представления по данной теме.

• Для закрепления теоретического материала ознакомиться с решениями типовых задач и ответить на вопросы для самоконтроля.

• Для проверки усвоения темы выполнить практические задания и тест для самоконтроля.

•Для подготовки к экзамену выполнить итоговый тест и итоговые практические задания.

Рекомендуемые информационные ресурсы:

1. http://tulpar.kfu.ru/course/view.php?id=2213.

2. Эконометрика: [Электронный ресурс] Учеб. пособие / А.И. Новиков. - 3-е изд., испр. и доп. - М.: ИНФРА-М, 2014. - 272 с.: (http://znanium.com/catalog.php?item=booksearch&code=%D1%8D%D0%BA%D0%BE%D0%BD%D0%BE%D0%BC%D0%B5%D1%82%D1%80%D0%B8%D0%BA%D0%B0&page=1#none) С. 7-21.

3.Уткин, В. Б. Эконометрика [Электронный ресурс] : Учебник / В. Б. Уткин; Под ред. проф. В. Б. Уткина. - 2-е изд. - М.: Издательско-торговая корпорация «Дашков и К°», 2012. - 564 с.

(http://znanium.com/catalog.php?item=booksearch&code=%D1%8D%D0%BA%D0%BE%D0%BD%D0%BE%D0%BC%D0%B5%D1%82%D1%80%D0%B8%D0%BA%D0%B0&page=4#none) С.11-226.

4. Эконометрика. Практикум: [Электронный ресурс] Учебное пособие / С.А. Бородич. - М.: НИЦ ИНФРА-М; Мн.: Нов. знание, 2014. - 329 с. (http://znanium.com/catalog.php?item=booksearch&code=%D1%8D%D0%BA%D0%BE%D0%BD%D0%BE%D0%BC%D0%B5%D1%82%D1%80%D0%B8%D0%BA%D0%B0&page=4#none) С.8-82.

Основные понятия теории вероятностей. Нормальное распределение и связанные с ним χ2 - распределение, распределение Стьюдента и Фишера. Вероятностью события А – *P(A)* – называется отношение числа *m* элементарных событий (исходов), благоприятствующих появлению события *A,* к числу *n* всех элементарных событий в условиях данного вероятностного эксперимента.

$$P(A) = \frac{m}{n} \qquad (1)$$

Из определения вытекают следующие свойства вероятности:

1.Вероятность случайного события есть положительное число, заключенное между 0 и 1:

$$0 \le P(A) \le 1. \qquad (2)$$

2. Вероятность достоверного события A равна 1: $P(A) = 1$ (3)

3. Если событие невозможное, то его вероятность равна

$$0 : P(A) = 0. \qquad (4)$$

4. Если события A и B несовместны, то

$$P(A + B) = P(A) + P(B). \qquad (5)$$

5. Если события А и В совместны, то вероятность их суммы равна сумме вероятностей этих событий без вероятности их совместного наступления:

$$P(A+B) = P(A) + P(B) - P(AB) \qquad (6)$$

6. Если A и $\overline{A}$ - противоположные события, то

$$P(\overline{A}) = 1 - P(A). \qquad (7)$$

7. Сумма вероятностей событий $A_1, A_2, \ldots, A_n$, образующих полную группу, равна 1:

$$P(A_1) + P(A_2) + \ldots + P(A_n) = 1. \qquad (8)$$

В экономических исследованиях значения m и n в формуле могут интерпретироваться по-другому. При статистическом определении вероятности события A под n понимается количество наблюдений результатов эксперимента, в которых событие A встречалось ровно m раз. В этом случае отношение $\frac{m}{n}$ называется относительной частотой (частостью) события A.

Случайной величиной (СВ) называют величину, которая в результате наблюдения (испытания) принимает то или иное значение, заранее не известное и зависящее от случайных обстоятельств.

Дискретной называют такую СВ, которая принимает отдельные, изолированные (конечные или счетные) значения с определенными вероятностями.

Непрерывной называют такую СВ, которая может принимать любое значение из некоторого конечного или бесконечного числового промежутка (т.е. количество возможных значений непрерывной СВ бесконечно и несчетно).

Законом распределения случайной величины называется всякое соотношение, устанавливающее связь между возможными значениями случайной величины и соответствующими им вероятностями.

Его можно задать таблично, аналитически (т.е. в виде формулы) и графически. Для любой дискретной случайной величины

$$\sum_{i=1}^{n} P(X = x_i) = \sum_{i=1}^{n} p_i = 1 \qquad (9)$$

Функцией распределения СВ *X* называют функцию $F(x)$, определяющую вероятность того, что СВ *X* принимает значение меньшее, чем x, т.е.

$$F(x) = P(X < x) \qquad (10)$$

Плотностью вероятности (плотностью распределения вероятностей) непрерывной СВ X называются производная ее функции распределения:

$$f(x) = \lim_{\Delta x \to 0} \frac{P(x \le X < x + \Delta x)}{\Delta x} = \lim_{\Delta x \to 0} \frac{F(x + \Delta x) - F(x)}{\Delta x} = F'(x) \qquad (11)$$

Плотность вероятности *f(x)*, как и функция распределения *F(x)*, является одной из форм закона распределения и существует только для непрерывных случайных величин.

Числовые характеристики СВ условно подразделяют на:

- характеристики положения (математическое ожидание, мода, медиана, начальные моменты различных порядков);
- характеристики рассеивания (дисперсия, среднее квадратическое отклонение, центральные моменты различных порядков).

Математическое ожидание характеризует среднее ожидаемое значение СВ, т.е. приближенно равно ее среднему значению. Для дискретной СВ:

$$M(x) = \sum_{i=1}^{k} x_i p_i, \qquad (12)$$

где k – число всех возможных значений СВ x.

Для непрерывной СВ:

$$M(x) = \int_{-\infty}^{+\infty} x f(x) dx \qquad (13)$$

Дисперсией $D(X)$ (иногда она обозначается σ_x^2) СВ X называется математическое ожидание квадрата отклонения СВ от ее математического ожидания. Она рассчитывается по формуле:

$$D(X) = M(X - M(X))^2 = M(X^2) - M^2(X) \qquad (14)$$

При этом для дискретной СВ:

$$D(X) = \sum_{i=1}^{k} (x_i - M(X))^2 p_i = \sum_{i=1}^{k} x_i^2 p_i - M^2(X) \qquad (15)$$

Для непрерывной СВ:

$$D(X) = \int_{-\infty}^{+\infty} (x - M(X))^2 f(x)dx = \int_{-\infty}^{+\infty} x^2 f(x)dx - M^2(X) \qquad (16)$$

Свойства дисперсии:

1. $D(C) = 0$, где $C -$ константа; (17)

2. $D(CX) = C^2 D(X)$; (18)

3. $D(X \pm Y) = D(X) + D(Y)$, где X и $Y -$ независимые СВ; (19)

4. $D(aX + b) = a^2 D(X)$, где a и $b -$ константа. (20)

Средним квадратическим отклонением $\sigma(x)$ СВ X называется квадратичный корень из дисперсии $D(X)$: $\sigma = \sqrt{D(x)}$ (21)

Чтобы оценить разброс значений СВ в процентах относительно ее среднего значения, вводится коэффициент вариации $V(x)$, рассчитываемый по формуле:

$$V(x) = \frac{\sigma(x)}{|M(x)|} \cdot 100\% \qquad (22)$$

Большинство СВ подчиняется определенному закону распределения, зная который можно предвидеть вероятности попадания исследуемой СВ в опреде-

ленные интервалы. К числу наиболее активно использующихся в эконометрическом анализе относятся:

- нормальное распределение (распределение Гаусса);
- распределение χ^2;
- распределение Стьюдента;
- распределение Фишера.

Нормальное распределение. СВ X имеет нормальное распределение, если ее плотность вероятности имеет вид:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma}} \cdot e^{\frac{-(x-m)^2}{2\sigma^2}} \tag{23}$$

Это равносильно тому, что

$$F(x) = \frac{1}{\sqrt{2\pi\sigma}} \cdot \int_{-\infty}^{x} e^{-\frac{(t-m)^2}{2\sigma^2}} dt \tag{24}$$

СВ, имеющая нормальное распределение, называется нормально распределенной или нормальной. Графики плотности вероятности и функции распределения нормальной СВ изображены на рис.1 и 2.

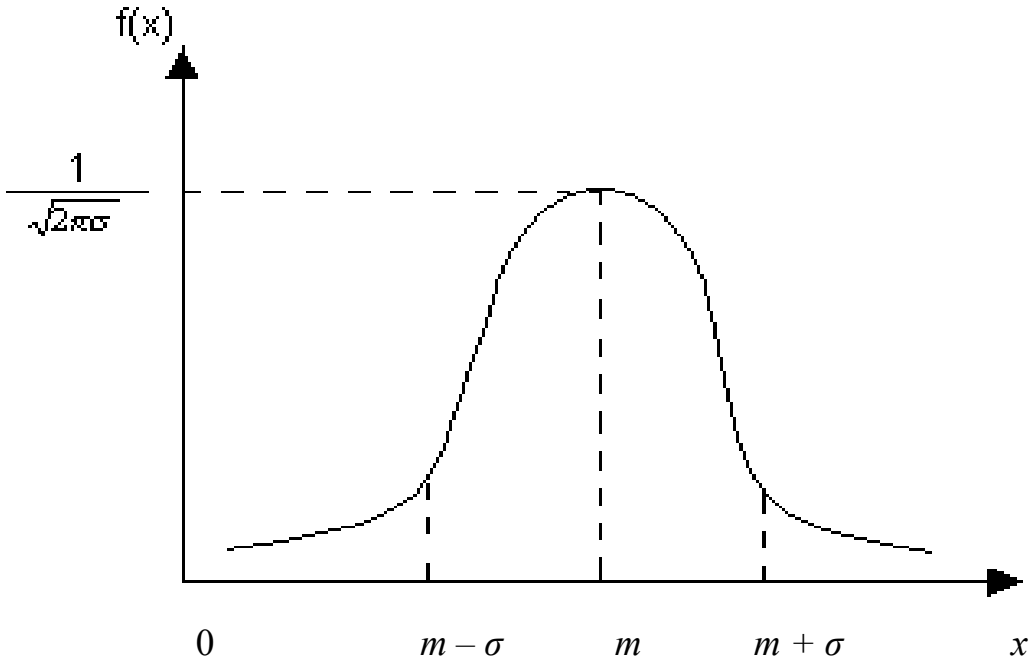


Рис. 2. 1. График плотности вероятности нормального распределения СВ X

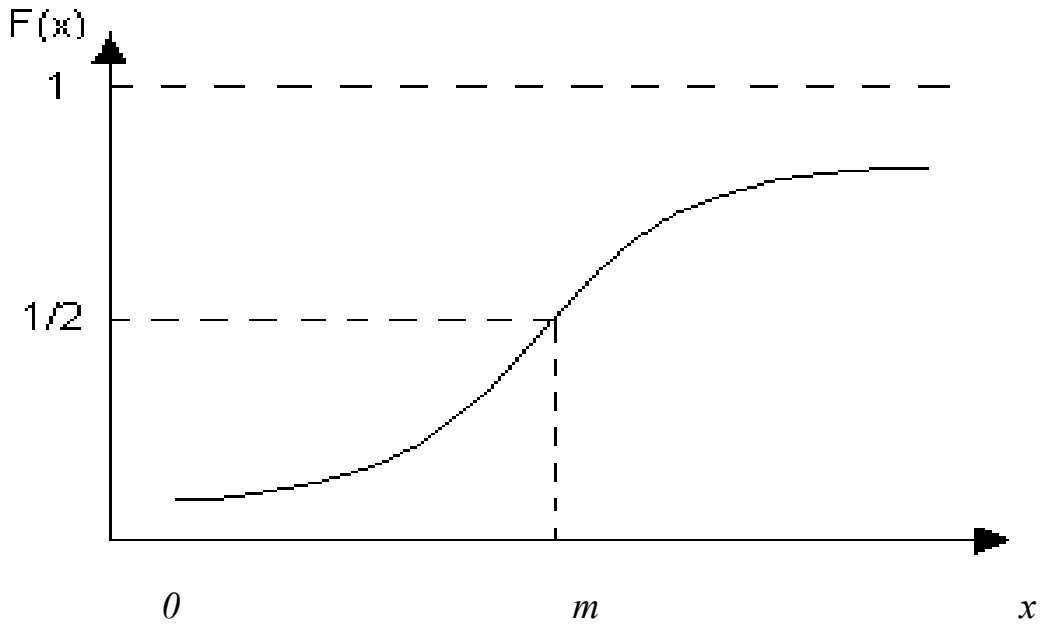


Рис. 2.2. Функция распределения нормальной СВ.

Как видно из формул (1) и (2), нормальное распределение зависит от параметров m и σ и полностью определяется ими. При этом $m = M(X)$, $\sigma = \sigma(X)$, т.е. $D(X) = \sigma^2$, $\pi = 3{,}14159\ldots$, $e = 2{,}71828\ldots$.

Если СВ X имеет нормальное распределение с параметрами $M(X) = m$ и $\sigma(X) = \sigma$, то символически это можно записать так:

$X \sim N(m, \sigma)$ *или* $X \sim N(m, \sigma^2)$.

Очень важным частным случаем нормального распределения является ситуация, когда $m = 0$ и $\sigma = 1$. В этом случае говорят о стандартизированном (стандартном) нормальном распределении.

Стандартизированную нормальную СВ обозначают через U ($U \sim N(0,1)$), учитывая при этом, что

$$f(u) = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{u^2}{2}}; \quad F(u) = \frac{1}{\sqrt{2\pi}} \cdot \int_{-\infty}^{u} e^{-\frac{t^2}{2}} dt \qquad (25)$$

Для практических расчетов специально разработаны таблицы функций $f(u)$, $F(u)$ стандартизированного нормального распределения, но чаще используется так называемая таблица значений Лапласа $\Phi(u)$. Функция Лапласа имеет вид:

$$\Phi(u) = \frac{1}{\sqrt{2\pi}} \cdot \int_0^u e^{-\frac{t^2}{2}} dt = F(u) - 0{,}5 \qquad (26)$$

Эту таблицу можно использовать для любой нормальной СВ X ($X \sim N(m, \sigma)$) при расчете соответствующих вероятностей:

$$P(a \le x \le b) = F\left(\frac{b-m}{\sigma}\right) - F\left(\frac{a-m}{\sigma}\right) = \Phi\left(\frac{b-m}{\sigma}\right) - \Phi\left(\frac{a-m}{\sigma}\right) \qquad (27)$$

Заметим, что если $X \sim N(m, \sigma)$, то $U = \frac{X-m}{\sigma} \sim N(0,1)$.

Распределение $\chi 2$ (хи – квадрат). Пусть X_i, $i = 1, 2, \ldots, n$ – независимые нормально распределенные СВ с математическими ожиданиями m_i и средними квадратическими отклонениями σ_i соответственно, т.е. $X_i \sim N(m_i, \sigma_i)$.

Тогда СВ $U_i = \frac{(x_i - m_i)}{\sigma_i}$, $i = 1, 2, \ldots, n$, являются независимыми СВ, имеющими стандартизированное нормальное распределение, $U_i \sim N(0,1)$.

СВ χ^2 имеет хи – квадрат распределение с **n** степенями свободы ($\chi^2 \sim \chi_n^2$), если

$$\chi^2 = \sum_{i=1}^{n} U_i^2 = U_1^2 + U_2^2 + \ldots + U_n^2 \qquad (28)$$

Отметим, что число степеней свободы (это число обозначается ***v***) исследуемой СВ определяется числом СВ, ее составляющих, уменьшенным на число линейных связей между ними.

Например, число степеней свободы СВ, являющейся композицией n случайных величин, которые в свою очередь связаны m линейными уравнениями, определяется числом $v = m - n$. Таким образом, $U^2 \sim \chi_1^2$.

Из определения (20) следует, что распределение χ^2 определяется одним параметром – числом степеней свободы v.

График плотности вероятности СВ, имеющий χ^2 – распределение, лежит только в первой четверти декартовой системы координат и имеет асимметрич-

ный вид с вытянутым правым «хвостом» (рис.3). Но с увеличением числа степеней свободы распределение χ^2 постепенно приближается к нормальному:

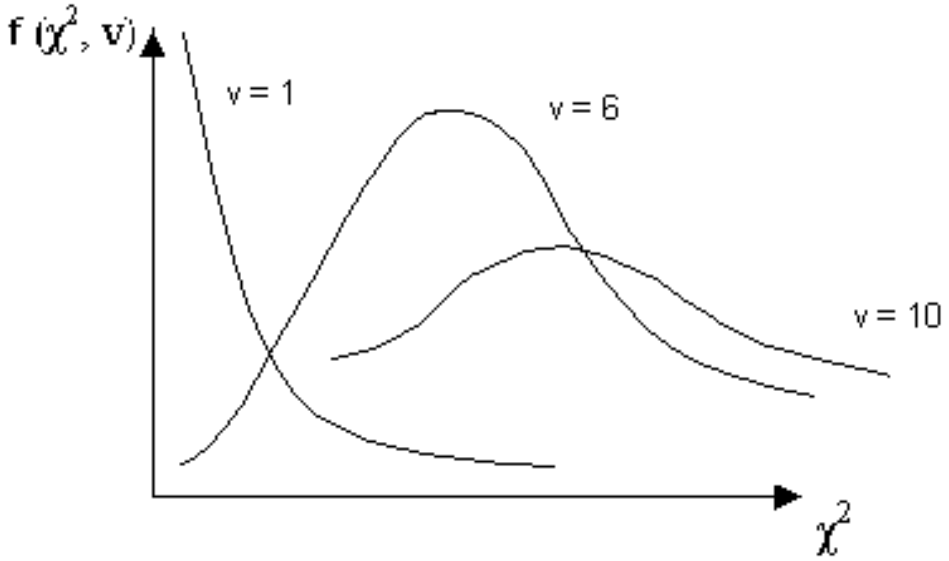


Рис. 2.3. График плотности вероятности СВ X, имеюший χ^2 – распределение.

$M(\chi^2) = v = n - m,$

$D(\chi^2) = 2v = 2(n - m).$

Если X и Y – две независимые χ^2 – распределенные СВ с числами степеней свободы n и k соответственно $(X \sim \chi_n^2,\ Y \sim \chi_k^2)$, то их сумма $(X + Y)$ также является χ^2 – распределенной СВ с числом степеней свободы $v = n + k$.

Распределение χ^2 применяется для нахождения интервальных оценок и проверки статистических гипотез. При этом используется таблица критических точек χ^2 – распределения.

Распределение Стьюдента. Пусть СВ $U \sim N(0,1)$, СВ V – независимая от U величина, распределенная по закону χ^2 *с* ***n*** степенями свободы. Тогда величина

$$T = \frac{U}{\sqrt{\frac{V}{n}}} \qquad (29)$$

имеет распределение Стьюдента (t – распределение) с n степенями свободы ($T \sim T_n$).

Из формулы (21) видно, что распределение Стьюдента определяется только одним параметром n – числом степеней свободы. График функции плотности вероятности СВ, имеющей распределение Стьюдента, является симметричной кривой (линия симметрии – ось ординат) (рис.4)

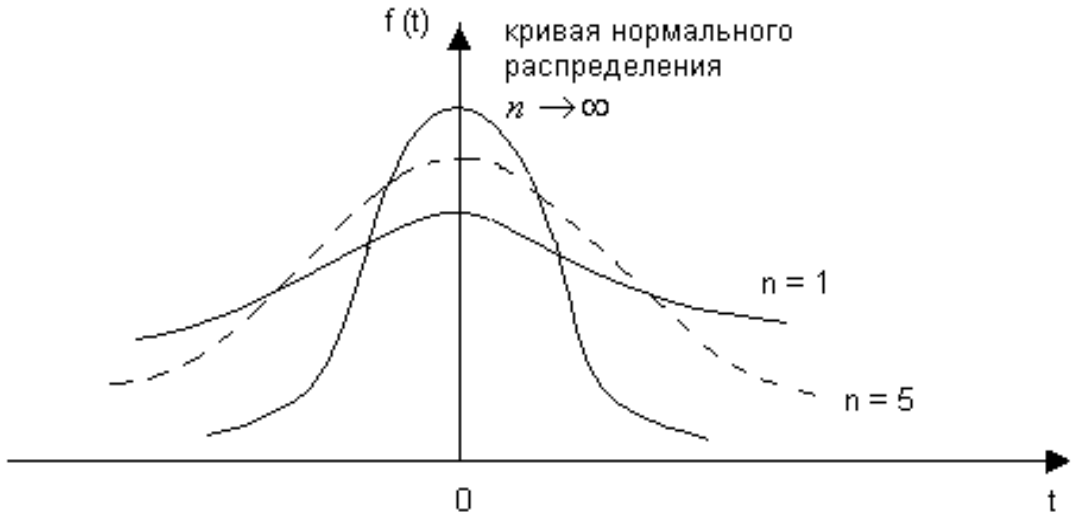


Рис. 2.4. График функции плотности вероятности СВ X, имеющий распределение Стьюдента

$M(T) = 0,$

$D(T) = n / (n - 2).$

При этом с увеличением числа степеней свободы распределение Стьюдента приближается к стандартизированному нормальному, причем при $n > 30$ распределение Стьюдента практически можно заменить нормальным распределением.

Распределение Стьюдента применяется для нахождения интервальных оценок, а также при проверке статистических гипотез. При этом активно используется таблица критических точек распределения Стьюдента.

Распределение Фишера. Пусть V и W – независимые СВ, распределенные по закону χ^2 со степенями свободы $v_1 = m$ и $v_2 = n$ соответственно. Тогда величина

$$F = \frac{V/m}{W/n} \qquad (30)$$

имеет распределение Фишера со степенями свободы $v_1 = m$ *и* $v_2 = n$ $(F \sim F_{m,n})$. Таким образом, распределение Фишера F определяется двумя параметрами – числами степеней свободы m и n.

При больших m и n это распределение приближается к нормальному (рис.5). Нетрудно заметить, что $T_n^2 = F_{1,n}$, где T_n – СВ, имеющая распределение Стьюдента с числом степеней свободы $v = n$, $F_{1,n}$ – СВ, имеющая распределение Фишера с числами степеней свободы $v_1 = 1$ *и* $v_2 = n$.

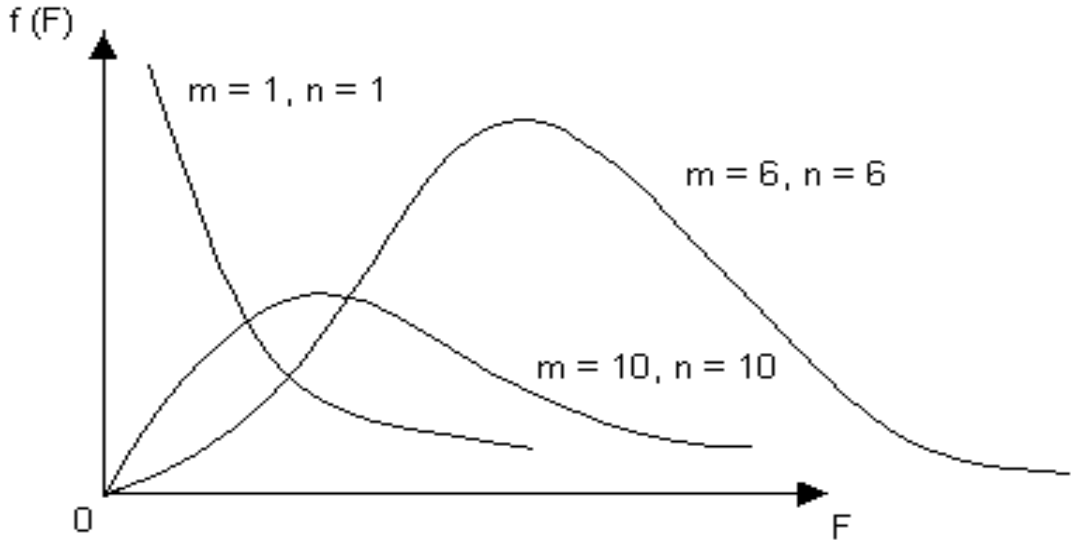


Рис.2.5. График функции плотности вероятности СВ X, имеющий распределение Фишера

$$\left.\begin{aligned} M(F) &= \frac{n}{n-2}\cdot(n-2), \\ D(F) &= \frac{2n^2(m+n-2)}{m(n-2)^2(n-4)}, (n>4) \end{aligned}\right\}$$

Распределение Фишера используется при проверке статистических гипотез в дисперсионном и регрессионном анализах. При этом активно используется таблица критических точек распределения Фишера.

Генеральная совокупность и выборка. Свойства статистических оценок. Статистические выводы – это заключения о генеральной совокупности (т.е. законе распределения исследуемой СВ и его параметрах либо о наличии и силе связи между исследуемыми переменными) на основе выборки, случайно отобранной из генеральной совокупности. Или обобщение результатов, полученных по выборке, на генеральную совокупность и есть суть статистических выводов. Процесс нахождения оценок по определенному правилу (формуле) называется оцениванием. В качестве оценок параметров распределения генеральной совокупности берутся их выборочные оценки. При этом различают 2 вида оценок: точечные, интервальные.

Точечной оценкой θ^* параметра θ называется числовое значение этого параметра, полученное по выборке объема n.

Так как выборка носит случайный характер, то оценка θ^* является СВ, принимающей различные значения для различных выборок. Любую оценку $\theta^* = \theta^*(x_{1,}x_{2...}x_n)$ называют статистикой или статистической оценкой параметра θ.

Точностью оценки называют такое число ε, что $\left|\theta - \theta^*\right| \le \varepsilon$.Качество оценок характеризуется следующими основными свойствами.

Оценка θ^* называется несмещенной оценкой параметра θ, если ее математическое ожидание равно оцениваемому параметру: $M(\theta^*) = \theta.$ В противном случае – оценка называется смещенной.

Разность $M(\theta^*) - \theta$ - называется смещением или систематической ошибкой оценивания. Для несмещенных оценок систематическая ошибка равна нулю. Если $M(\theta^*) > \theta$, то θ^* завышает среднее значение θ. Нетрудно заметить, что в этом случае она будет иметь наименьшую среди других оценок дисперсию.

Оценка θ^* называется эффективной оценкой параметра θ, если ее дисперсия $D(\theta^*)$ меньше дисперсии любой другой альтернативной несмещенной оценки при фиксированном объеме выборки n, т.е. $D(\theta^*) = D_{\min}$.

Оценка называется асимптотически эффективной, если с увеличением объема выборки ее дисперсия стремится к нулю, т.е. $D(\theta_n^*) \to 0$ при $n \to \infty$ (индекс n в оценке θ_n^* применяется для подчеркивания объема выборки).

Оценка θ_n^* называется состоятельной оценкой параметра θ, если θ_n^* сходится по вероятности к оцениваемому параметру θ при $n \to \infty$. Другими словами, состоятельной называется такая оценка, которая дает истинное значение при достаточно большом объеме выборки вне зависимости от значений входящих в нее конкретных наблюдений.

Справедливо следующее утверждение: если $M(\theta_n^*) \to \theta$ и $D(\theta_n^*) \to 0$ при $n \to \infty$, то θ_n^* – состоятельная оценка параметра θ.

Оценки, являющиеся линейными функциями от выборочных наблюдений, называются линейными. Точечная оценка может быть дополнена интервальной оценкой – интервалом (θ_1, θ_2), внутри которого с наперед заданной вероятностью γ находится точное значение оцениваемого параметра θ. Определение такого интервала называют интервальным оцениванием, а сам интервал – доверительным интервалом. При этом γ называют доверительной вероятностью – или надежностью, с которой оцениваемый параметр θ попадает в интервал (θ_1, θ_2). Для определения доверительного интервала заранее выбирают число $\alpha = 1 - \gamma, \quad 0 < \alpha < 1$, называемое уровнем значимости, и находят два числа θ_1 и θ_2, зависящих от точечной оценки θ^*, такие, что

$$P(\theta_1 < \theta < \theta_2) = 1 - \alpha = \gamma \qquad (31)$$

В этом случае говорят, что интервал (θ_1, θ_2) накрывает неизвестный параметр θ с вероятностью $(1-\alpha)$. Границы интервала θ_1 и θ_2 называются доверительными, и они обычно находятся из условия $P(\theta > \theta_2) = \alpha/2$. Длина доверительного интервала, характеризующая точность интервальной оценки, зависит от объема выборки n и надежности γ (уровня значимости $\alpha = 1 - \gamma$). При увеличении величины n длина доверительного интервала уменьшается, а с приближением надежности γ к единице – увеличивается. Выбор α (или $\gamma = 1 - \alpha$) определяется конкретными условиями. Обычно используется $\alpha = 0{,}1; 0{,}05; 0{,}01$, что соответствует 90, 95, 99%-м доверительным интервалам.

Общая схема построения доверительного интервала:

1. Из генеральной совокупности с известным распределением $f(x, \theta)$ CB X извлекается выборка объема n, по которой находится точечная оценка θ^* параметра θ.

2. Строится CB $Y(\theta)$, связанная с параметром θ и имеющая известную плотность вероятности $f(y, \theta)$.

3. Задается уровень значимости α.

4. Используя плотность вероятности CB Y, определяют два числа l_1 и l_2 такие, что $P(l_1 < Y(\theta) < l_2) = \int_{l_1}^{l_2} f(y, \theta) dy = 1 - \alpha$ (32)

5. Выбираются значения l_1 и l_2 из условий

$$P(Y(\theta) < l_1) = \alpha/2; \qquad P(Y(\theta) > l_2) = \alpha/2.$$

Неравенство $l_1 < Y(\theta) < l_2$ преобразуется в равносильное $\theta^* - \delta < \theta < \theta^* + \delta$ такое, что $P(\theta^* - \delta < \theta < \theta^* + \delta) = 1 - \alpha$ (33)

Полученный интервал $(\theta^* - \delta, \theta^* + \delta)$, накрывающий неизвестный параметр θ с вероятностью $1 - \alpha$, и является интервальной оценкой параметра θ.

Доверительный интервал для математического ожидания нормальной СВ при известной дисперсии

$$(\bar{x} - u_{\alpha/2}\frac{\sigma}{\sqrt{n}};\quad \bar{x} + u_{\alpha/2}\ \frac{\sigma}{\sqrt{n}})$$

$$\Phi(u_{\alpha/2}) = \frac{1-\alpha}{2} = \frac{\gamma}{2}.$$

Доверительный интервал для математического ожидания нормальной СВ при неизвестной дисперсии.

$$(\bar{x} - t_{\alpha/2,n-1}\frac{S}{\sqrt{n}}; \bar{x} + t_{\alpha/2,n-1}\frac{S}{\sqrt{n}})$$

Доверительный интервал для дисперсии нормальной СВ

$$\frac{S^2(n-1)}{\chi^2_{\frac{a}{2},n-1}} < \sigma^2 < \frac{S^2(n-1)}{\chi^1_{1-\frac{a}{2},n-1}}$$

Статистические выводы и проверка гипотез. Статистической гипотезой называется любое предположение о виде закона распределения или о параметрах неизвестного закона распределения. В первом случае гипотеза называется непараметрической, а во втором параметрической.

Гипотеза $H_{0,}$ подлежащая проверке, называется нулевой. Наряду с нулевой рассматривают гипотезу $H_{1,}$ которая будет приниматься, если отклоняется $H_{0.}$. Такая гипотеза называется альтернативной (конкурирующей).

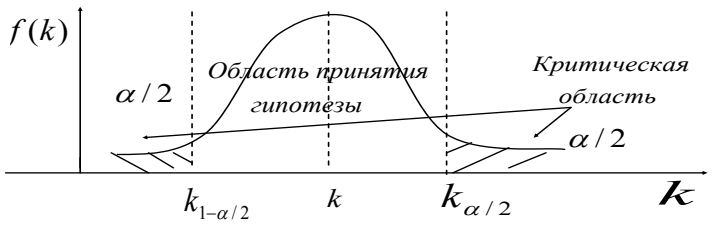
Статистическим критерием или статистическим тестом называют СВ K, которая служит для проверки нулевой гипотезы. Совокупность значений критерия, при которых нулевую гипотезу отклоняют, называют критической областью.

Совокупность значений критерия, при которых нулевую гипотезу не отклоняют, называют областью принятия гипотезы.

Основной принцип проверки статистических гипотез можно сформулировать так: если наблюдаемое значение критерия K (вычисленное по выборке) принадлежит критической области, то нулевую гипотезу отклоняют. Если же наблюдаемое значение критерия K принадлежит области принятия гипотезы, то нулевую гипотезу не отклоняют (принимают).

Точки, разделяющие критическую область и область принятия гипотезы, называют критическими.

Критическая область $(-\infty; k_{1-a/2}) \cup (k_{a/2,} + \infty)$ называется двусторонней критической областью. Она определяется в случае, когда альтернативная гипотеза имеет вид $H_1 : \theta \neq \theta_0$.



$H_1 : \theta \neq \theta_0$

Рис. 2.6. Двусторонняя критическая область

Кроме двусторонней, рассматривают также односторонние критические области – правостороннюю и левостороннюю.

Правосторонней называют критическую область $(k_{\alpha,} + \infty)$, определяющуюся из соотношения $P(K > k_\alpha) = \alpha.$

Она используется в случае, когда альтернативная гипотеза имеет вид: $H_1: \theta > \theta_0.$

Левосторонней называют критическую область $(-\infty, k_{1-\alpha})$, определяющую из соотношения $P(K < k_{1-\alpha}) = \alpha.$ Она используется в случае, когда альтернативная гипотеза имеет вид $H_1: \theta < \theta_0.$

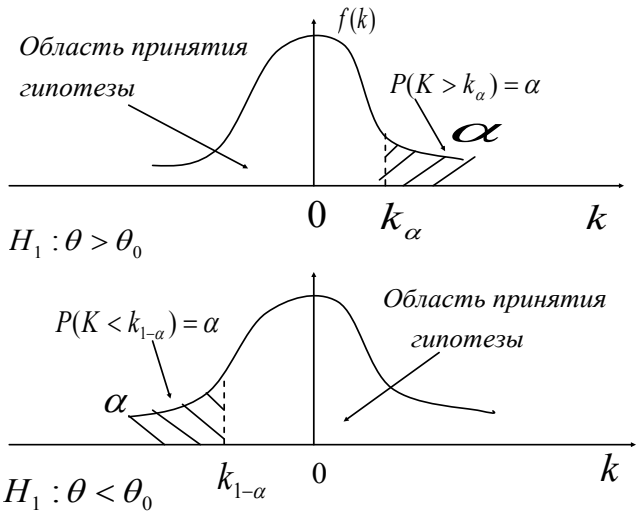


Рис. 2.7. Правосторонняя и левосторонняя критическая область

Общая схема проверки гипотез:

1. Формулировка проверяемой нулевой (H_0) и альтернативной (H_1) гипотез.

2. Выбор соответствующего уровня значимости α.

3. Определение объема выборки $n.$

4. Выбор критерия K для проверки $H_0.$

5. Определение критической области и области принятия гипотезы.

6. Вычисление наблюдаемого значения критерия $K_{набл.}$

7. Принятие статистического решения.

I. Схема проверки гипотезы о математическом ожидании нормальной СВ при известной дисперсии

$$H_0: m = m_0$$

$$H_1^{(1)}: m \neq m_0 \,(H_1^{(2)}: m > m_0; H_1^{(3)}: m < m_0)$$

$$U = \frac{\bar{x} - m_0}{\sigma / \sqrt{n}} \qquad (34)$$

где $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$, $\sigma = \sqrt{\sigma^2}$.

1. Пусть в качестве альтернативной рассматривается гипотеза: $H_1^{(1)}: m \neq m_0$. Тогда критические точки $u_{\alpha/2}$ и $u_{1-\alpha/2} = -u_{\alpha/2}$ будут определяться по таблице значений функции Лапласа из условия

$$\Phi(u_{\alpha/2}) = \frac{1-\alpha}{2}$$

Если $|U_{набл.}| = \left|\frac{\bar{x} - m_0}{\sigma / \sqrt{n}}\right| < u_{\alpha/2}$ - нет оснований для отклонения H_0.

Если $|U_{набл.}| \geq u_{\alpha/2}$ - гипотеза H_0 отклоняется в пользу альтернативной гипотезы $H_1^{(1)}$.

2. При $H_1^{(2)}: m > m_0$ критическую точку u_α правосторонней критической области находят из равенства

$$\Phi(И_\alpha) = \frac{1-2\alpha}{2}$$

Если $U_{набл.} < u_\alpha$ - нет оснований для отклонения H_0.

Если $U_{набл.} \geq u_\alpha$ - H_0 отклоняют в пользу $H_1^{(2)}$.

3. При $H_1^{(3)}: m < m_0$ критическая точка $u_{1-\alpha} = -u_\alpha$.

Если $U_{набл.} > u_{1-\alpha}$ - нет оснований для отклонения H_0.

Если $U_{набл.} \le u_{1-\alpha}$ - H_0 отклоняют в пользу $H_1^{(3)}$.

II. Схема проверки гипотезы о математическом ожидании нормальной СВ при неизвестной дисперсии

$$H_0: m = m_0$$

$$H_1^{(1)}: m \neq m_0 \quad (H_1^{(2)}: m > m_0; H_1^{(3)}: m < m_0).$$

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

Исправленная выборочная дисперсия $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$.

Стандартное отклонение $S = \sqrt{S^2}$. Далее строится t - статистика:

$$T = \frac{\bar{x} - m_0}{S/\sqrt{n}}, \qquad (35)$$

имеющая при справедливости H_0 распределение Стьюдента с $\nu = n-1$ степенями свободы. Критическая область строится в зависимости от вида альтернативной гипотезы так же, как и в предыдущем разделе.

1. При $H_1^{(1)}: m \neq m_0$ по таблице критических точек распределения значимости α и числу степеней свободы $\nu = n-1$ находятся критические точки:

$t_{\alpha/2,n-1}$ и $t_{1-\alpha/2,n-1} = -t_{\alpha/2,n-1}$.

Если $|T_{набл.}| = \left|\frac{\bar{x} - m_0}{S/\sqrt{n}}\right| < t_{\alpha/2,n-1}$ - нет оснований для отклонения H_0.

Если $|T_{набл.}| \geq t_{\alpha/2,n-1}$ - H_0 отклоняют в пользу $H_1^{(1)}$.

2. При $H_1^{(2)} : m > m_0$ определяют критическую точку $t_{\alpha,n-1}$ правосторонней критической области.

Если $T_{набл.} < t_{\alpha,n-1}$ - нет оснований для отклонения H_0.

Если $T_{набл.} \geq t_{\alpha,n-1}$ - H_0 отклоняется в пользу $H_1^{(2)}$.

3. При $H_1^{(3)} : m < m_0$ определяют критическую точку $t_{1-\alpha,n-1} = -t_{\alpha,n-1}$ левосторонней критической области.

Если $T_{набл.} > -t_{\alpha,n-1}$ - нет оснований для отклонения H_0.

Если $T_{набл.} \leq -t_{\alpha,n-1}$ - H_0 отклоняется в пользу $H_1^{(3)}$.

III. Схема проверки гипотезы о величине дисперсии нормальной СВ

$$H_0 : \sigma^2 = \sigma_0^2$$

$$H_1^{(1)} : \sigma^2 \neq \sigma_0^2 \left(H_1^{(2)} : \sigma^2 > \sigma_0^2 ; H_1^{(3)} : \sigma^2 < \sigma_0^2 \right).$$

Для проверки H_0 извлекается выборка объема $n : x_1, x_2 \ldots x_n$, вычисляются выборочное среднее $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$, исправленная выборочная дисперсия

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

Тогда критерий проверки H_0 имеет вид:

$$\chi^2 = \frac{(n-1) \cdot S^2}{\sigma_0^2} \qquad (36)$$

При справедливости H_0 построенная статистика χ^2 имеет χ^2- распределение с $v = n-1$ степенями свободы.

1. При $H_1^{(1)} : \sigma^2 \neq \sigma_0^2$ по таблице критических точек $\chi 2$- распределения по заданному уровню значимости α и числу степеней свободы $v = n-1$ находят критические точки $\chi^2_{1-\alpha/2,n-1}$ и $\chi^2_{\alpha/2,n-1}$ двусторонней критической области.

Если $\chi^2_{1-\alpha/2,n-1} < \chi^2_{набл.} < \chi^2_{\alpha/2,n-1}$ - нет оснований для отклонения H_0.

Если $\chi^2_{набл.} < \chi^2_{1-\alpha/2,n-1}$ или $\chi^2 набл. \geq \chi^2 \alpha/2,n-1$ - H_0 отклоняется в пользу $H_1^{(1)}$.

2. При $H_1^{(2)} : \sigma^2 > \sigma_0^2$ определяют критическую точку $\chi^2_{\alpha,n-1}$ правосторонней критической области.

Если $\chi^2_{набл.} < \chi^2_{\alpha,n-1}$- нет оснований для отклонения H_0.

Если $\chi^2_{набл.} \geq \chi^2_{\alpha,n-1}$- H_0 отклоняется в пользу $H_1^{(2)}$.

3. При $H_1^{(3)} : \sigma^2 < \sigma_0^2$ находят критическую точку $\chi^2_{1-\alpha,n-1}$ левосторонней критической области.

Если $\chi^2_{набл.} > \chi^2_{1-\alpha,n-1}$- нет оснований для отклонения H_0.

Если $\chi^2_{набл.} \leq \chi^2_{1-\alpha,n-1}$- H_0 отклоняется в пользу $H_1^{(3)}$.

IV. Схема проверки гипотезы о равенстве $M(X)$ *двух нормальных СВ при известных дисперсиях*

$H_0 : M(X) = M(Y)$
$H_1^{(1)} : M(X) \neq M(Y)$ $\quad \left(H_1^{(2)} : M(X) > M(Y)\right)$ $\left(H_1^{(3)} : M(X) < M(Y)\right)$

В качестве критерия проверки H_0 принимается СВ U:

$$U = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{k}}} \qquad (37)$$

При справедливости H_0 СВ $U \sim N(0, 1)$.

1. При $H_1^{(1)} : M(X) \neq M(Y)$ по таблице функции Лапласа определяют 2 критические точки $u_{1-\alpha}$ и $u_{\alpha/2}$ из условий:

$$\Phi_{(u_{\alpha/2})} = \frac{1-\alpha}{2},\ u_{1-\alpha/2} = u_{\alpha/2}.$$

Если $|U_{набл.}| < u_{\alpha/2}$ - нет оснований для отклонения H_0.

Если $\left|U_{набл.}\right| \geq u_{\alpha/2}$ - H_0 отклоняется в пользу $H_1^{(1)}$.

2. При $H_1^{(2)} : M(X) > M(Y)$ критическую точку u_α правосторонней критической области находят их равенства: $\Phi_{(u_\alpha)} = \frac{1-2\alpha}{2}$.

Если $U_{набл.} < u_\alpha$ - нет оснований для отклонения H_0.

Если $U_{набл.} \geq u_\alpha$ - H_0 отклоняется в пользу $H_1^{(2)}$.

3. При $H_1^{(3)} : M(X) < M(Y)$ критическая точка $u_{1-\alpha}$ левосторонней критической области определяется из соотношения $u_{1-\alpha} = -u_{\alpha}$.

Если $U_{набл.} > u_{1-\alpha}$ - нет оснований для отклонения H_0.

Если $U_{набл.} \le u_{1-\alpha}$ - H_0 отклоняется в пользу $H_1^{(3)}$.

V. Схема проверки гипотезы о равенстве математических ожиданий двух нормальных СВ при неизвестных дисперсиях

$H_0 : M(X) = M(Y)$

$H_1 : M(X) \neq M(Y) \quad \left(H_1^{(2)} : M(X) > M(Y)\right); \quad \left(H_1^{(3)} : M(X) < M(Y)\right).$

При этих условиях в качестве критерия проверки H_0 принимают СВ T :

$$T = \frac{\bar{x} - \bar{y}}{\sqrt{(n-1)S_x^2 + (k-1)S_y^2}} \cdot \sqrt{\frac{nk(n+k-2)}{n+k}} \qquad (38)$$

где n, k - объемы выборок $x_1, x_2 \ldots x_n$ и $y_1, y_2 \ldots y_k$ соответственно

$$\bar{x} = \frac{1}{n}\Sigma x_i; \;\; S_x^2 = \frac{1}{n-1}\Sigma (x_i - x)^2,$$

$$\bar{y} = \frac{1}{k}\Sigma y_i; \qquad S_y^2 = \frac{1}{k-1}\Sigma (y_i - \bar{y})^2.$$

1. При $H_1{}^{(1)} : M(X) \neq M(Y)$ с помощью таблицы критических точек распределения Стьюдента по заданному уровню значимости α и числу степеней свободы $\nu = n + k - 2$ определяются критические точки $t_{1-\alpha/2, n+k-2}$ и $t_{\alpha/2, n+k-2}$ ($t_{1-\alpha/2, n+k-2} = -t_{\alpha/2, n+k-2}$) двусторонней критической области.

Если $|T_{набл.}| < t_{\alpha/2, n+k-2}$ - нет оснований для отклонения H_0.

Если $|T_{набл.}| \geq t_{\alpha/2,n+k-2}$ - H_0 отклоняется в пользу $H_1^{(1)}$.

2. При $H_1^{(2)} : M(X) > M(Y)$ находят критическую точку $t_{\alpha,n+k-2}$ правосторонней критической области.

Если $T_{набл.} < t_{\alpha,n+k-2}$ - нет оснований для отклонения H_0.

Если $T_{набл.} \geq t_{\alpha,n+k-2}$ - H_0 отклоняется в пользу $H_1^{(2)}$.

3. При $H_1^{(3)} : M(X) < M(Y)$ находят критическую точку левосторонней критической области $t_{1-\alpha,n+k-2} = -t_{\alpha,n+k-2}$.

Если $T_{набл.} > -t_{\alpha,n+k-2}$ - нет оснований для отклонения H_0.

Если $T_{набл.} \leq -t_{\alpha,n+k-2}$ - H_0 отклоняется в пользу $H_1^{(3)}$.

VI. Схема проверки гипотезы о равенстве дисперсий двух нормальных СВ

При сравнении двух экономических показателей иногда, в первую очередь, проводят анализ разброса значений рассматриваемых СВ. Например, при решении инвестирования в одну из отраслей остро стоит проблема риска вложений. При сравнивании уровня жизни двух стран среднедушевые доходы могут быть примерно одинаковы. Необходимо сопоставить разброс в доходах.

Анализ проводится путем сравнения дисперсий исследуемых СВ.

Пусть $X \approx N(m_x, \sigma_x^2)$ и $Y \approx N(m_y, \sigma_y^2)$, причем их дисперсии σ_x^2 и σ_y^2 неизвестны. Выдвигается гипотеза о равенстве дисперсий σ_x^2 и σ_y^2.

$$H_0 : \sigma_x^2 = \sigma_y^2$$

$$H_1^{(1)} : \sigma_x^2 \neq \sigma_y^2 \quad \left(H_1^{(2)} : \sigma_x^2 > \sigma_y^2\right).$$

По независимым выборкам $x_1, x_2 \ldots x_n$ и $y_1, y_2 \ldots y_k$ объемов n и k соответственно определяется:

$\bar{x}, \bar{y}, S_x^2$ и S_y^2 (для определенности пусть $S_x^2 \geq S_y^2$, в противном случае эти величины можно переобозначить).

В качестве критерия проверки H_0 принимают СВ

$$F = \frac{S_x^2}{S_y^2} \quad , \tag{39}$$

определяемую отношением большей исправленной выборочной дисперсии к меньшей.

Если H_0 верна, то данная статистика F имеет F - распределение Фишера с $v_1 = n-1$ и $v_2 = k-1$ степенями свободы.

1. При $H_1^{(1)} : \sigma_x^2 \neq \sigma_y^2$ по таблицам критических точек распределения Фишера по уровню значимости α и числам степеней свободы v_1 и v_2 определяется критическая точка $F_{\alpha/2, v_1, v_2}$.

Если $F_{набл.} < F_{\alpha/2, v_1, v_2}$ - нет оснований для отклонения H_0.

Если $F_{набл.} \geq F_{\alpha/2, v_1, v_2}$ - H_0 отклоняется в пользу $H_1^{(1)}$.

2. При $H_1^{(2)} : \sigma_x^2 > \sigma_y^2$ определяется критическая точка F_{α, v_1, v_2}.

Если $F_{набл.} < F_{\alpha, v_1, v_2}$ - нет оснований для отклонения H_0.

Если $F_{набл.} \geq F_{\alpha, v_1, v_2}$ - H_0 отклоняется в пользу $H_1^{(2)}$.

В основном, при проверке гипотезы о равенстве дисперсий в качестве альтернативной гипотезы в большинстве случаев используется гипотеза $H_1^{(2)}$.

VII. Схема проверки гипотезы о значимости коэффициента корреляции

$H_0 : \rho_{xy} = 0$

$H_1^{(1)} : \rho_{xy} \neq 0.$

Для проверки H_0 по выборке $(x_1, y_1),(x_1, y_2)...(x_n, y_n)$ объема n строится статистика:

$$T = \frac{r_{xy} \cdot \sqrt{n-2}}{\sqrt{1-r_{xy}^2}} \qquad (40)$$

где r_{xy} - выборочный коэффициент корреляции.

При справедливости H_0 статистика T имеет распределение Стьюдента с $v = n-2$ степенями свободы.

По таблице критических точек распределения Стьюдента по заданному уровню значимости α и числу степеней свободы $n-2$ определяем критическую точку $t_{\alpha/2,n-2}$.

Если $|T_{набл.}| < t_{\alpha/2,n-2}$ - то нет оснований для отклонения H_0.

Если $|T_{набл.}| \geq t_{\alpha/2,n-2}$ - то H_0 отклоняется в пользу альтернативной гипотезы $H_1^{(1)}$.

Если H_0 отклоняется, то фактически это означает, что коэффициент корреляции статистически значим (существенно отличен от нуля). Следовательно, X и Y - коррелированны, т.е. между ними существует линейная связь.

Вопросы и задания для самоконтроля

1. Как связаны между собой случайные величины, имеющие стандартизованное нормальное распределение, распределение Стьюдента, χ^2 и Фишера?

2. В чем заключаются несмещенность, эффективность и состоятельность статистических оценок?

3. Что такое точечная и интервальная оценка?

4. Что такое нулевая и альтернативная гипотезы?

5. Что такое статистический критерий, уровень значимости?

6. Какая случайная величина применяется в качестве критерия проверки гипотезы о величине дисперсии нормальной случайной величины?

7. Какая случайная величина применяется в качестве критерия проверки гипотезы о равенстве дисперсий двух нормальных случайных величин?

8. К проверке каких гипотез сводятся исследования среднего дохода населения и анализ разброса в уровне дохода?

Задание 1. В университете проведен анализ успеваемости среди студентов и студенток за последние 25 лет. Случайные величины X, Y, представляющие их суммарный балл за время учебы соответственно, имеют нормальный закон распределения. Получены следующие данные: $\bar{x} = 400$, $\bar{y} = 420$, $S_x^2=300$, $S_y^2=150$. Проверить, можно ли на уровне значимости $\alpha = 0{,}05$ утверждать, что девушки в среднем учатся лучше ребят.

Задание 2. Точность работы станка-автомата, заполняющего пакеты со стиральным порошком, определяется совпадением веса пакетов. Дисперсия веса не должна превышать 25. По выборке из 20 пакетов определена исправленная дисперсия $S^2=30$. Определить на уровне значимости $\alpha = 0{,}05$ требуется ли переналадка станка.

Задание 3. По двум независимым выборкам, объемы которых $n_1 = 9$ и $n_2 = 6$, найдены выборочные дисперсии $S_x^2=14{,}4$ и $S_y^2=20{,}5$ годовых дивидендов от вложений в отрасли А и В соответственно. Проверить при уровне значимости $\alpha = 0{,}05$ гипотезу о равенстве рисков при вложении денег в обе отрасли.

Лекция 3

Тема 3. Линейная модель парной регрессии и метод наименьших квадратов

Вопросы для изучения

1. Спецификация линейной модели парной регрессии.

2. Метод наименьших квадратов (МНК) – идентификация линейной модели парной регрессии.

3. Предпосылки МНК и свойства МНК-оценок.

Аннотация. Данная тема раскрывает суть регрессионного анализа в эконометрике.

Ключевые слова. Модель регрессии, метод наименьших квадратов, остатки регрессии.

Методические рекомендации по изучению темы

• Изучить лекционную часть, где даются общие представления по данной теме.

• Для закрепления теоретического материала ознакомиться с решениями типовых задач и ответить на вопросы для самоконтроля.

• Для проверки усвоения темы выполнить практические задания и тест для самоконтроля.

Рекомендуемые информационные ресурсы:

1. http://tulpar.kfu.ru/course/view.php?id=2213

2. Эконометрика: [Электронный ресурс] Учеб.пособие / А.И. Новиков. - 2-е изд., испр. и доп. - М.: ИНФРА-М, 2011. - 144 с.: с. (http://znanium.com/catalog.php?item=booksearch&code=%D1%8D%D0%BA%D0%BE%D0%BD%D0%BE%D0%BC%D0%B5%D1%82%D1%80%D0%B8%D0%BA%D0%B0&page=1#none) С. 28-46.

3.Уткин, В. Б. Эконометрика [Электронный ресурс] : Учебник / В. Б. Уткин; Под ред. проф. В. Б. Уткина. - 2-е изд. - М.: Издательско-торговая корпорация «Дашков и К°», 2012. - 564 с.

(http://znanium.com/catalog.php?item=booksearch&code=%D1%8D%D0%BA%D0%BE%D0%BD%D0%BE%D0%BC%D0%B5%D1%82%D1%80%D0%B8%D0%BA%D0%B0&page=4#none) С. 323-338.

4. Валентинов, В. А. Эконометрика [Электронный ресурс]: Практикум / В. А. Валентинов. - 3-е изд. - М.: Дашков и К, 2010. - 436 с.

(http://znanium.com/catalog.php?item=booksearch&code=%D1%8D%D0%BA%D0%BE%D0%BD%D0%BE%D0%BC%D0%B5%D1%82%D1%80%D0%B8%D0%BA%D0%B0&page=3#none) С. 38-99.

5.Эконометрика. Практикум: [Электронный ресурс] Учебное пособие / С.А. Бородич. - М.: НИЦ ИНФРА-М; Мн.: Нов.знание, 2014. - 329 с. (http://znanium.com/catalog.php?item=booksearch&code=%D1%8D%D0%BA%D0%BE%D0%BD%D0%BE%D0%BC%D0%B5%D1%82%D1%80%D0%B8%D0%BA%D0%B0&page=4#none) С. 82-99.

6. Электронный курс “Econometrics and Public Policy (Advanced)”, Princeton University, URL: https://blackboard.princeton.edu/webapps/portal/frameset.jsp?tab_group=courses&url=%2Fwebapps%2Fblackboard%2Fexecute%2FcourseMain%3Fcourse_id%3D_214206_1

Спецификация линейной модели парной регрессии. В зависимости от количества факторов, включенных в уравнение регрессии, принято различать простую (парную) и множественную регрессии. Простая регрессия представляет собой регрессию между двумя переменными – y и x, т.е. модель вида:

$$y = \hat{f}(x) \qquad (1)$$

где y – зависимая переменная (результативный признак); x - независимая, или объясняющая переменная (признак – фактор, или регрессор).

Множественная регрессия представляет собой регрессию результативного признака с двумя и большим числом факторов, т.е. модель вида:

$$y = \hat{f}(x_1, x_2, ..x_k) \qquad (2)$$

Любое эконометрическое исследование начинается со спецификации модели, т.е. с формулировки вида модели, исходя из соответствующей теории связи между переменными. Из всего круга факторов, влияющих на результативный признак, необходимо выделить наиболее существенно влияющие факторы. Парная регрессия достаточна, если имеется доминирующий фактор, который и используется в качестве объясняющей переменной. Например, выдвигается ги-

потеза о том, что величина спроса у на товар находится в обратной зависимости от цены х, т.е. $\hat{y}_x = a - b \cdot x.$

Уравнение простой регрессии характеризует связь между двумя переменными, которая проявляется как закономерность лишь в среднем по совокупности наблюдений. (Например, если зависимость спроса «у» от цены «х» имеет вид: у=5000-2х. Это означает, что с ростом цены на 1 д.е. спрос в среднем уменьшается на 2 д.е.). В уравнении регрессии корреляционная по сути связь признаков представляется в виде функциональной связи. В каждом отдельном случае величина у складывается из двух слагаемых:

$y_j = \hat{y}_{x_j} + \varepsilon_j$,

где y_j – фактическое значение результативного признака; $\hat{y}_{x_j}$ - значение признака, найденное из математической функции связи у и х, т.е. из уравнения регрессии; ε_j - случайная величина, характеризующая отклонение реального значения признака от найденного по уравнению регрессии.

Случайная величина ε называется также возмущением. Она включает влияние не учтенных в модели факторов, случайных ошибок и особенностей измерения. Ее порождают 3 источника: спецификация модели, выборочный характер исходных данных и ошибки измерения. Например, зависимость спроса от цены точнее следует записывать так: у=5000-2х+ε. В данном случае слева записано просто у, что означает фактическое значение, а не $\hat{y}$, отвечающее значению, рассчитанному по уравнению регрессии.

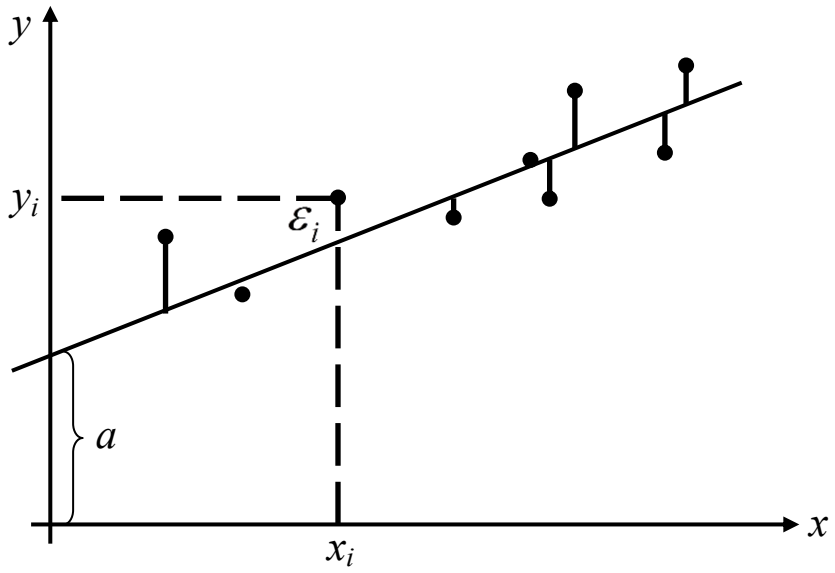
Метод наименьших квадратов (МНК) – идентификация линейной модели парной регрессии. Линейная регрессия сводится к нахождению уравнения вида

$$\hat{y}_x = a + bx \text{ (или } y = a + bx + \varepsilon) \qquad (3)$$

Первое выражение позволяет по заданным значениям фактора *x* рассчитать теоретические значения результативного признака, подставляя в него фак-

тические значения фактора *x*. На графике теоретические значения лежат на прямой, которая представляют собой линию регрессии.

Построение линейной регрессии сводится к оценке ее параметров- *a* и *b*. Классический подход к оцениванию параметров линейной регрессии основан на методе наименьших квадратов (МНК).



МНК позволяет получить такие оценки параметров *a* и *b*, при которых сумма квадратов отклонений фактических значений y от теоретических $\hat{y}_x$ минимальна:

$$\sum_i \left(y_i - \hat{y}_{x_i}\right)^2 \to \min, \text{ или } \sum_{i=1}^{n} \varepsilon_i^2 \to \min \qquad (4)$$

Для нахождения минимума надо вычислить частные производные суммы (4) по каждому из параметров - *a* и *b* - и приравнять их к нулю.

$$\Sigma\varepsilon_i^2 = S^2 = \Sigma(y - \hat{y}_x)^2 = \Sigma(y - a - bx)^2$$

$$\begin{cases} \dfrac{\partial s}{\partial a} = -2\sum y + 2na + 2\bar{b}\sum x = 0; \\ \dfrac{\partial s}{\partial b} = -2\sum y \cdot x + 2a\sum x + 2b\sum x^2 = 0 \end{cases} \qquad (5)$$

Преобразуем, получаем систему нормальных уравнений:

$$\begin{cases} n \cdot a + b\sum x = \sum y, \\ a\sum x + b\sum x^2 = \sum yx \end{cases} \quad (6)$$

В этой системе *n*- объем выборки, суммы легко рассчитываются из исходных данных. Решаем систему относительно *a* и *b*, получаем:

$$b = \frac{n\sum yx - \left(\sum y\right)\left(\sum x\right)}{n\sum x^2 - \left(\sum x\right)^2}, \quad (7)$$

$$a = \frac{1}{n}\sum y - \frac{b}{n}\sum x. \quad (8)$$

Выражение (7) можно записать в другом виде:

$$b = \frac{\overline{yx} - \bar{y} \cdot \bar{x}}{\overline{x^2} - \bar{x}^2} = \frac{cov(x, y)}{\sigma_x^2}, \quad (9)$$

где $cov(x, y)$ – ковариация признаков, σ_x^2 – дисперсия фактора *x*.

Параметр *b* называется коэффициентом регрессии. Его величина показывает среднее изменение результата с изменением фактора на одну единицу. Возможность четкой экономической интерпретации коэффициента регрессии сделала линейное уравнение регрессии достаточно распространенным в эконометрических исследованиях.

Формально *a*- значение *y* при *x=0.* Если *x* не имеет и не может иметь нулевого значения, то такая трактовка свободного члена *a* не имеет смысла. Параметр *a* может не иметь экономического содержания. Попытки экономически интерпретировать его могут привести к абсурду, особенно при *a*< 0. Интерпретировать можно лишь знак при параметре *a.* Если *a*> 0, то относительное изменение результата происходит медленнее, чем изменение фактора. Сравним эти относительные изменения:

$$bx < a + bx \text{ при } a > 0,\ x > 0 \Rightarrow b < \frac{a + bx}{x}$$

$$\Rightarrow \frac{bdx}{dx.} < \frac{a + bx}{x} \Rightarrow \frac{dy}{dx} < \frac{y}{x} \Rightarrow \frac{dy}{y} < \frac{dx}{x}.$$

Иногда линейное уравнение парной регрессии записывают для отклонений от средних значений:

$$y' = b \cdot x', \qquad (10)$$

где $y' = y - \bar{y}$, $x' = x - \bar{x}$. При этом свободный член равен нулю, что и отражено в выражении (10). Этот факт следует из геометрических соображений: уравнению регрессии отвечает та же прямая (3), но при оценке регрессии в отклонениях начало координат перемещается в точку с координатами $(\bar{x}, \bar{y})$. При этом в выражении (8) обе суммы будут равны нулю, что и повлечет равенство нулю свободного члена.

Предпосылки МНК и свойства МНК-оценок. Как было сказано выше, связь между *y* и *x* в парной регрессии является не функциональной, а корреляционной. Поэтому оценки параметров *a* и *b* являются случайными величинами, свойства которых существенно зависят от свойств случайной составляющей ε. Для получения по МНК наилучших результатов необходимо выполнение следующих предпосылок относительно случайного отклонения (условия Гаусса – Маркова):

1^0. Математическое ожидание случайного отклонения равно нулю для всех наблюдений: $M(\varepsilon_i) = 0$.

2^0. Дисперсия случайных отклонений постоянна: $D(\varepsilon_i) = D(\varepsilon_j) = \sigma^2$.

Выполнимость данной предпосылки называется гомоскедастичностью (постоянством дисперсии отклонений). Невыполнимость данной предпосылки называется гетероскедастичностью (непостоянством дисперсии отклонений)

3^0. Случайные отклонения ε_i и ε_j являются независимыми друг от друга для $i \neq j$:

$$\operatorname{cov}(\varepsilon_i, \varepsilon_j) = \begin{cases} 0, & i \neq j \\ \sigma^2, & i = j. \end{cases}$$

Выполнимость этого условия называется отсутствием автокорреляции.

4^0. Случайное отклонение должно быть независимо от объясняющих переменных. Обычно это условие выполняется автоматически, если объясняющие переменные в данной модели не являются случайными. Кроме того, выполнимость данной предпосылки для эконометрических моделей не столь критична по сравнению с первыми тремя.

При выполнимости указанных предпосылок имеет место теорема Гаусса-Маркова: оценки (7) и (8), полученные по МНК, имеют наименьшую дисперсию в классе всех линейных несмещенных оценок.

Таким образом, при выполнении условий Гаусса-Маркова оценки (7) и (8) являются не только несмещенными оценками коэффициентов регрессии, но и наиболее эффективными, т.е. имеют наименьшую дисперсию по сравнению с любыми другими оценками данных параметров, линейными относительно величин y_i.

Вопросы и задания для самоконтроля

1. Что такое функция регрессии?
2. Чем регрессионная модель отличается от функции регрессии?
3. Каковы основные причины наличия в регрессионной модели случайного отклонения?
4. Как осуществляется спецификация модели?
5. В чем состоит различие между теоретическим и эмпирическим уравнениями регрессии?
6. В чем суть метода наименьших квадратов?
7. Каковы формулы расчета коэффициентов эмпирического парного

линейного уравнения регрессии по МНК?

8. Каковы предпосылки МНК? Каковы последствия их выполнимости или невыполнимости?

9. Действительно ли оценки коэффициентов регрессии будут иметь нормальное распределение, если случайные отклонения распределены нормально?

10. Действительно ли в любой линейной регрессионной модели, построенной по МНК, сумма случайных отклонений равна нулю?

Задание 1. При исследовании корреляционной зависимости между ценой на нефть X и индексом нефтяных компаний Y получены следующие данные: $\bar{x} = 16{,}2;\ \bar{y} = 4000;\ \sigma_x^2 = 4;\ \mathrm{cov}(x, y) = 40.$

Задание: построить линейное уравнение регрессии Y на X.

Задание 2. По выборке объема $n = 10$ получены следующие данные:

$$\sum x_i = 100;\ \sum y_i = 200;\ \sum x_i y_i = 21000;\ \sum x_i^2 = 12000;\ \sum y_i^2 = 45000.$$

Задание: оценить с помощью МНК параметры линейного уравнения регрессии, найти выборочный коэффициент корреляции r_{xy} .

Лекция 4,5

Тема 4. Экономическая и статистическая интерпретация линейной модели парной регрессии

Вопросы для изучения:

1. Экономическая интерпретация параметров модели.

2. Коэффициенты корреляции и детерминации в линейной модели парной регрессии.

3. Проверка качества линейной модели парной регрессии (верификация модели).

4. Интервалы прогноза по линейному уравнению регрессии.

Аннотация. Данная тема раскрывает прикладное содержание регрессионного анализа.

Ключевые слова. Коэффициент регрессии, статистическая значимость, метод наименьших квадратов, коэффициент детерминации.

Методические рекомендации по изучению темы

•Изучить лекционную часть, где даются общие представления по данной теме.

• Для закрепления теоретического материала ознакомиться с решениями типовых задач и ответить на вопросы для самоконтроля.

• Для проверки усвоения темы выполнить практические задания и тест для самоконтроля.

•Для подготовки к экзамену выполнить итоговый тест и итоговые практические задания.

Рекомендуемые информационные ресурсы:

1. http://tulpar.kfu.ru/course/view.php?id=2213.

2. Эконометрика: [Электронный ресурс] Учеб.пособие / А.И. Новиков. - 2-е изд., испр. и доп. - М.: ИНФРА-М, 2011. - 144 с.: с. (http://znanium.com/catalog.php?item=booksearch&code=%D1%8D%D0%BA%D0%BE%D0%BD%D0%BE%D0%BC%D0%B5%D1%82%D1%80%D0%B8%D0%BA%D0%B0&page=1#none) С. 46-56.

3. Уткин, В. Б. Эконометрика [Электронный ресурс] : Учебник / В. Б. Уткин; Под ред. проф. В. Б. Уткина. - 2-е изд. - М.: Издательско-торговая корпорация «Дашков и К°», 2012. - 564 с.

(http://znanium.com/catalog.php?item=booksearch&code=%D1%8D%D0%BA%D0%BE%D0%BD%D0%BE%D0%BC%D0%B5%D1%82%D1%80%D0%B8%D0%BA%D0%B0&page=4#none) С. 338-365.

4. Валентинов, В. А. Эконометрика [Электронный ресурс]: Практикум / В. А. Валентинов. - 3-е изд. - М.: Дашков и К, 2010. - 436 с.

(http://znanium.com/catalog.php?item=booksearch&code=%D1%8D%D0%BA%D0%BE%D0%BD%D0%BE%D0%BC%D0%B5%D1%82%D1%80%D0%B8%D0%BA%D0%B0&page=3#none) С. 67-99.

5.Эконометрика. Практикум: [Электронный ресурс] Учебное пособие / С.А. Бородич. - М.: НИЦ ИНФРА-М; Мн.: Нов.знание, 2014. - 329 с. (http://znanium.com/catalog.php?item=booksearch&code=%D1%8D%D0%BA%D0%BE%D0%BD%D0%BE%D0%BC%D0%B5%D1%82%D1%80%D0%B8%D0%BA%D0%B0&page=4#none) С. 99-133.

Экономическая интерпретация параметров модели

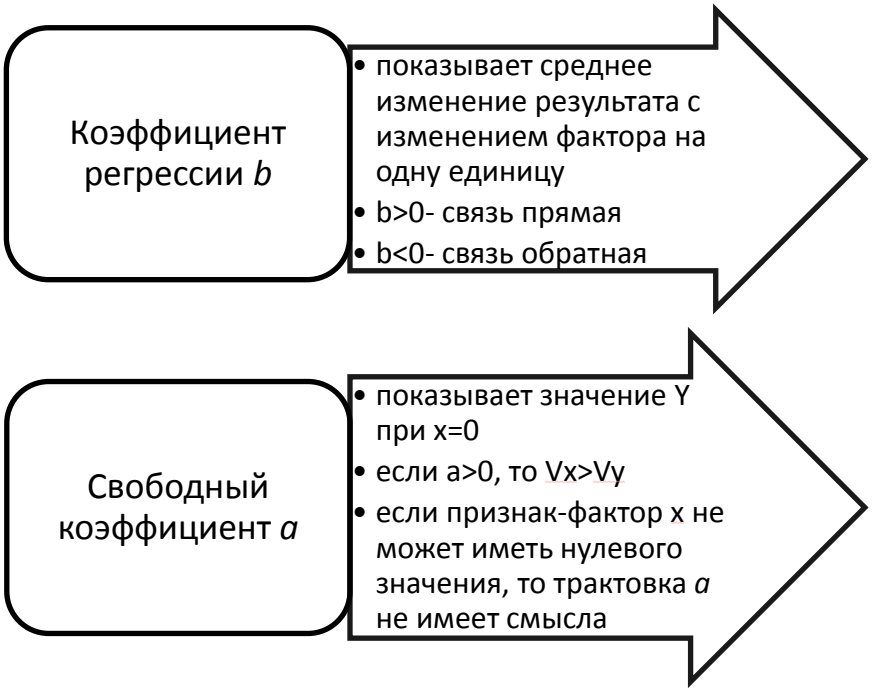


Рис. 4.1. Интерпретация параметров модели

Коэффициенты корреляции и детерминации в линейной модели парной регрессии. Если все точки лежат на построенной прямой, то регрессия Y на X «идеально» объясняет поведение зависимой переменной. Обычно поведение Y лишь частично объясняется влиянием переменной X.

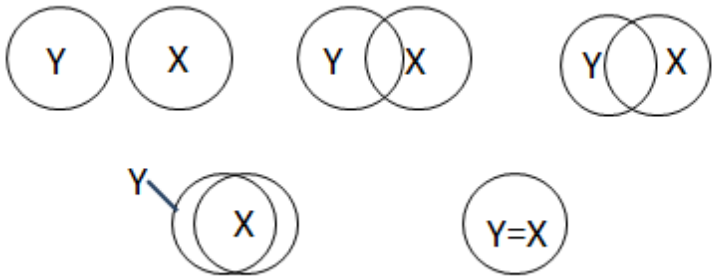


Рис. 4.2. Диаграмма Венна

Линейный коэффициент парной корреляции:

$$r_{yx} = b\frac{\sigma_x}{\sigma_y} = \frac{\mathrm{cov}(x,y)}{\sigma_x\sigma_y} = \frac{\overline{yx} - \bar{y}\cdot\bar{x}}{\sigma_x\cdot\sigma_y}$$

$$-1 \le r_{yx} \le 1$$

Если b>0, то $r_{yx}>0$; если b<0, то $r_{yx}<0$.

По абсолютной величине, чем ближе значение r_{xy} к единице, тем теснее связь, чем ближе значение r_{xy} к нулю, тем слабее связь между *y* и *x*.

$$|r_{yx}| < 0{,}3 - \textit{слабая}$$

$$0{,}3 \le |r_{yx}| \le 0{,}7 - \textit{средняя}$$

$$|r_{yx}| > 0{,}7 - \textit{сильная, тесная}$$

Суммы квадратов отклонений:

- общая (TSS): $\sum(y_i - \bar{y})^2$

- регрессионная (ESS): $\sum(y_x - \bar{y})^2$

- остаточная (RSS): $\sum(y_i - y_x)^2$

$$\sum(y_i - \bar{y})^2 = \sum(y_x - \bar{y})^2 + \sum(y_i - y_x)^2$$

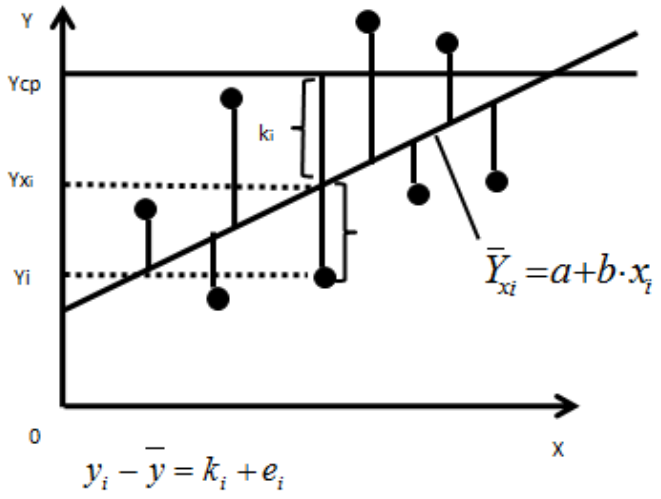


Рис.4.3. Геометрическая интерпретация

Число степеней свободы (df-degrees of freedom) - это число независимо варьируемых значений признака. Для общей СКО требуется *(n-1)* независимых отклонений, т.к. $\sum(y - \bar{y}) = 0,$ что позволяет свободно варьировать *(n-1)* значений, а последнее n-е отклонение определяется из общей суммы, равной нулю. Поэтому $df_{общ.} = n - 1.$

Факторную СКО можно выразить так:

$$\sum(\hat{y}_x - \bar{y})^2 = \sum[(a + bx) - (a + b\bar{x})]^2$$
$$= \sum(bx - b\bar{x})^2 = b^2 \sum(x - \bar{x})^2$$

Эта СКО зависит только от одного параметра - *b,*поскольку выражение под знаком суммы к значениям результативного признака не относится. Следовательно, факторная СКО имеет одну степень свободы, и $df_{факт.} = 1.$

Для определения $df_{остат.}$ воспользуемся аналогией с балансовым равенством (11). Можно записать равенство и между числами степеней свободы:

$$df_{общ.} = df_{факт.} + df_{остат.}$$

Таким образом, можем записать: $(n-1) = 1 + (n-2)$

Из этого баланса определяем, что $df_{\text{остат.}}=n-2$.

Разделив каждую СКО на свое число степеней свободы, получим средний квадрат отклонений, или дисперсию на одну степень свободы.

Выборочные оценки дисперсий:

- общая дисперсия: $S^2_{TSS}=\dfrac{\sum(y_i-\bar{y})^2}{n-1}$

- регрессионная дисперсия: $S^2_{ESS}=\dfrac{\sum(y_x-\bar{y})^2}{m}$

- остаточная дисперсия: $S^2_{RSS}=\dfrac{\sum(y_i-y_x)^2}{n-m-1}$

Коэффициент детерминации:

$$R^2=\frac{\sum(y_x-\bar{y})^2}{\sum(y_i-\bar{y})^2}=1-\frac{\sum(y_i-y_x)^2}{\sum(y_i-\bar{y})^2}$$

$$R^2=r^2_{yx};0\le R^2\le 1$$

Коэффициент детерминации определяет долю разброса зависимой переменной Y, объясняемую регрессией Y на X.

Проверка качества модели линейной парной регрессии (верификация модели)

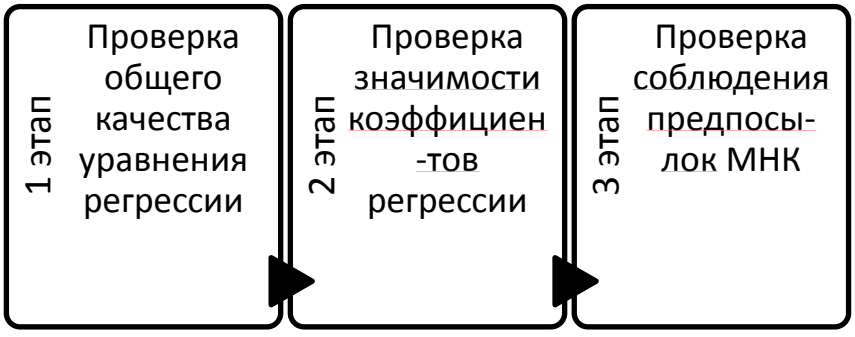


Рис. 4.5. Этапы проверки качества модели

1 этап: F-тест состоит в проверке гипотезы H_0 о статистической незначимости уравнения регрессии и показателя тесноты связи.

$$H_0 : D^2{}_{ESS} = D^2{}_{RSS}$$
$$H_1 : D^2{}_{ESS} > D^2{}_{RSS}$$
$$F = \frac{\sum (y_x - \overline{y})^2 / m}{\sum (y_i - y_x)^2 / (n-m-1)} = \frac{r^2{}_{xy}}{1 - r^2{}_{xy}} \cdot (n-2)$$
$$F > F_{\alpha, v_1 = m, v_2 = n-m-1} \Rightarrow H_1$$
$$F < F_{\alpha, v_1 = m, v_2 = n-m-1} \Rightarrow H_0$$

2 этап: Т-тест состоит в проверке гипотезы H_0 о статистической незначимости коэффициентов регрессии и корреляции.

$$H_0 : \beta = 0$$
$$H_1 : \beta \neq 0$$
$$|t_b| > t_{\alpha/2, n-2} \Rightarrow H_1$$
$$|t_b| < t_{\alpha/2, n-2} \Rightarrow H_0$$

$$_b = \frac{b}{m_b}; t_a = \frac{a}{m_a}; t_r = \frac{r}{m_r} = \frac{r}{\sqrt{1 - r^2}} \cdot \sqrt{n-2}$$
$$m_b = \sqrt{\frac{\sum (y - y_x)^2 / (n-2)}{\sum (x - \overline{x})^2}} = \sqrt{\frac{S^2{}_{RSS}}{\sum (x - \overline{x})^2}} = \frac{S_{RSS}}{\sigma_x \sqrt{n}}$$

$$m_a = \sqrt{\frac{\sum (y - y_x)^2}{(n-2)} \cdot \frac{\sum x^2}{n \sum (x - \overline{x})^2}}$$
$$m_r = \sqrt{\frac{1 - r^2}{n-2}}; t^2{}_r = t^2{}_b = F$$

3 этап: проведение тестов на гетероскедастичность и автокорреляцию остатков.

Доверительные интервалы для коэффициентов теоретического уравнения регрессии:

$$t = \frac{b - \beta}{m_b}$$

$$\Delta b = t_{\alpha/2,n-2} \cdot m_b;$$

$$b - \Delta b < \beta < b + \Delta b$$

$$\Delta a = t_{\alpha/2,n-2} \cdot m_a;$$

$$a - \Delta a < \alpha < a + \Delta a$$

Интервалы прогноза по линейному уравнению регрессии. Прогнозирование по уравнению регрессии представляет собой подстановку в уравнение регрессии соответственного значения *x*. Такой прогноз $\hat{y}_x$ называется точечным. Он не является точным, поэтому дополняется расчетом стандартной ошибки $\hat{y}_x$; получается интервальная оценка прогнозного значения y^*:

$$\hat{y}_x - m_{\hat{y}_x} \le y^* \le \hat{y} + m_{\hat{y}_x}$$

Преобразуем уравнение регрессии:

$$\hat{y}_x = a + bx = (\bar{y} - b\bar{x}) + bx = \bar{y} + b(x - \bar{x})$$

ошибка $m_{\hat{y}_x}$ зависит от ошибки $\bar{y}$ и ошибки коэффициента регрессии b, т.е. $m_{\hat{y}_x}^2 = m_{\bar{y}}^2 + m_b^2 (x - \bar{x})^2$.

Из теории выборки известно, что $m_{\bar{y}}^2 = \frac{\sigma^2}{n}$

Используем в качестве оценки σ^2 остаточную дисперсию на одну степень свободы S^2, получаем: $m_{\bar{y}}^2 = \frac{S^2}{n}$

Ошибка коэффициента регрессии :

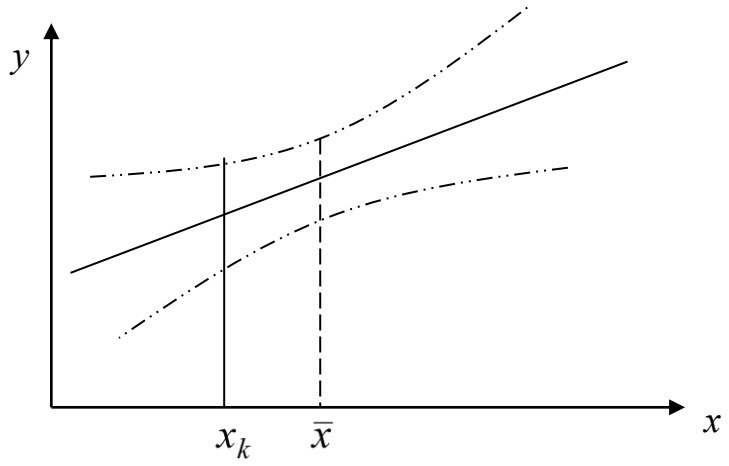
$$m_b^2 = \frac{S^2}{\sum (x - \bar{x})^2}$$

Таким образом, при $x = x_к$ получаем:

$$m_{\hat{y}_x}^2 = \frac{S^2}{n} + \frac{S^2}{\sum (x - \bar{x})^2}(x_{\hat{e}} - \bar{\delta})^2 = S^2\left(\frac{1}{n} + \frac{(x_{\hat{e}} - \bar{\delta})^2}{\sum (\tilde{\delta} - \bar{\delta})^2}\right)$$

$$m_{\hat{y}_x} = S\sqrt{\frac{1}{n} + \frac{(x_{\hat{e}} - \bar{\delta})^2}{\sum (\tilde{\delta} - \bar{\delta})^2}} \qquad (11)$$

Как видно из формулы (11), величина $m_{\hat{y}_x}$ достигает минимума при $x_к = \bar{x}$ и возрастает по мере удаления $x_к$ от $\bar{x}$ в любом направлении.



Для нашего примера эта величина составит:

$$m_{\hat{y}_x} = \sqrt{53\left(\frac{1}{7} + \frac{(x_{\hat{e}} - 3{,}143)^2}{10{,}857}\right)}$$

При $x_к = \bar{x}$ $m_{\hat{y}_x} = \sqrt{53:7} = 2{,}75$. При $x_к = 4$

$$m_{\hat{y}_x} = \sqrt{53\left(\frac{1}{7} + \frac{(4 - 3{,}143)^2}{10{,}857}\right)} = 3{,}34$$

Для прогнозируемого значения $\tilde{o}_{\tilde{o}}$ 95% - ные доверительные интервалы при заданном $x_к$ определены выражением:

$$\tilde{o}_{\tilde{o}_{\hat{e}}} \pm t_\alpha \cdot m_{\hat{y}_{x,}} \qquad (12)$$

т.е. при $x_к = 4$ $\hat{y}_{x_ê} \pm 2{,}57 \cdot 3{,}34$ или $\hat{y}_{x_ê} \pm 8{,}58$. При $x_к = 4$ прогнозное значение составит $y_p = -5{,}79 + 36{,}84 \cdot 4 = 141{,}57$ - это точечный прогноз.

Прогноз линии регрессии лежит в интервале:

$132{,}99 \le \tilde{o}_{\tilde{o}_ê} \le 150{,}15$

Мы рассмотрели доверительные интервалы для среднего значения y при заданном x.Однако фактические значения y варьируются около среднего значения $\tilde{o}_{\tilde{o}}$, они могут отклоняться на величину случайной ошибки ε, дисперсия которой оценивается как остаточная дисперсия на одну степень свободы S^2. Поэтому ошибка прогноза отдельного значения y должна включать не только стандартную ошибку $m_{\hat{y}_x}$, но и случайную ошибку *S*. Таким образом, средняя ошибка прогноза индивидуального значения y составит:

$$m_{y_{i(x_к)}} = S\sqrt{1 + \frac{1}{n} + \frac{(x_к - \bar{x})^2}{\sum (x - \bar{x})^2}} \qquad (13)$$

Для примера:

$$m_{y_{i(x_к=4)}} = \sqrt{53 \cdot \left(1 + \frac{1}{7} + \frac{(4 - 3{,}143)^2}{10{,}857}\right)} = 8{,}01$$

Доверительный интервал прогноза индивидуальных значений y при $x_к = 4$ с вероятностью 0,95 составит: $141{,}57 \pm 2{,}57 \cdot 8{,}01,$ или $120{,}98 \le y_p \le 162{,}16.$

Пусть в примере с функцией издержек выдвигается предположение, что в предстоящем году в связи со стабилизацией экономики затраты на производство 8 тыс. ед. продукции не превысят 250 млн. руб. Означает ли это изменение найденной закономерности или затраты соответствуют регрессионной модели?

Точечный прогноз: $\hat{y}_{x=8} = -5,79 + 36,84 \cdot 8 = 288,93.$

Предполагаемое значение - 250. Средняя ошибка прогнозного индивидуального значения:

$$m_{y_{i(x_i)}} = S\sqrt{1 + \frac{1}{n} + \frac{(x_K - \bar{x})}{\sum(x - \bar{x})}} =$$

$$= \sqrt{53 \cdot \left(1 + \frac{1}{7} + \frac{(8 - 3,143)^2}{10,857}\right)} = 13,26$$

Сравним ее с предполагаемым снижением издержек производства, т.е. 250-288,93=-38,93:

$$t = \frac{-38,93}{12,26} = -2,93.$$

Поскольку оценивается только значимость уменьшения затрат, то используется односторонний *t*- критерий Стьюдента. При ошибке в 5 % с $n - 2 = 5 \quad t_{таб.} = 2,015$, поэтому предполагаемое уменьшение затрат значимо отличается от прогнозируемого значения при 95 % - ном уровне доверия. Однако, если увеличить вероятность до 99%, при ошибке 1 % фактическое значение *t* – критерия оказывается ниже табличного 3,365, и различие в затратах статистически не значимо, т.е. затраты соответствуют предложенной регрессионной модели.

Вопросы и задания для самоконтроля

1. Каков экономический смысл коэффициента регрессии?
2. Какой смысл может иметь свободный коэффициент уравнения регрессии?
3. Какова связь между линейным коэффициентом корреляции и коэффициентом регрессии в линейной модели парной регрессии?
4. Каков статистический смысл коэффициента детерминации?

5. Как записывается баланс для сумм квадратов отклонений результативного признака?

6. Что происходит, когда общая СКО равна остаточной? В каком случае общая СКО равна факторной?

7. Что такое число степеней свободы? Чему равны числа степеней свободы для различных СКО в парной регрессии?

8. Как используется F-статистика в регрессионном анализе?

9. Как F-статистика связана с коэффициентом детерминации в парной регрессии?

10. Как рассчитать критерий Стьюдента для коэффициента регрессии в линейной модели парной регрессии?

11. В чем суть предсказания индивидуальных значений зависимой переменной?

Задание 1. Пусть имеется следующая модель парной регрессии, построенная по 20 наблюдениям: $\tilde{y} = 8 - 7x$. При этом r_{xy} = - 0,5.

Задание: построить доверительный интервал для коэффициента регрессии в этой модели с вероятностями 0,9 и 0,95.

Задание 2. Анализируется зависимость между доходами горожан (X), имеющими индивидуальные домовладения, и рыночной стоимостью их домов (Y). По случайной выборке из 120 горожан данной категории получены результаты:

$$\sum x_i = 27343;\ \sum y_i = 115870;\ \sum (x_i - \bar{x})^2 = 75200;$$

$$\sum (y_i - \bar{y})^2 = 1620340;\ \sum (x_i - \bar{x})(y_i - \bar{y}) = 250431.$$

Задание: найти оценку коэффициента регрессии b_1 и построить 95% доверительный интервал для коэффициента регрессии.

Лекция 6

Тема 5. Линейная модель множественной регрессии, оценка ее параметров

Вопросы для изучения

1. Линейная модель множественной регрессии. Эмпирическая форма записи.
2. Оценка параметров модели с помощью МНК.

Аннотация. Данная тема раскрывает особенности линейной модели множественной регрессии.

Ключевые слова. Стандартизованный коэффициент регрессии, метод наименьших квадратов, МНК-оценки.

Методические рекомендации по изучению темы

•Изучить лекционную часть, где даются общие представления по данной теме.

• Для закрепления теоретического материала ознакомиться с решениями типовых задач и ответить на вопросы для самоконтроля.

• Для проверки усвоения темы выполнить практические задания и тест для самоконтроля.

Рекомендуемые информационные ресурсы:

1. http://tulpar.kfu.ru/course/view.php?id=2213.

2. Эконометрика: [Электронный ресурс] Учеб.пособие / А.И. Новиков. - 2-е изд., испр. и доп. - М.: ИНФРА-М, 2011. - 144 с.: с. (http://znanium.com/catalog.php?item=booksearch&code=%D1%8D%D0%BA%D0%BE%D0%BD%D0%BE%D0%BC%D0%B5%D1%82%D1%80%D0%B8%D0%BA%D0%B0&page=1#none) С.50-58.

3. Уткин, В. Б. Эконометрика [Электронный ресурс] : Учебник / В. Б. Уткин; Под ред. проф. В. Б. Уткина. - 2-е изд. - М.: Издательско-торговая корпорация «Дашков и К°», 2012. - 564 с.

(http://znanium.com/catalog.php?item=booksearch&code=%D1%8D%D0%BA%D0%BE%D0%BD%D0%BE%D0%BC%D0%B5%D1%82%D1%80%D0%B8%D0%BA%D0%B0&page=4#none) С. 323-369.

4. Валентинов, В. А. Эконометрика [Электронный ресурс]: Практикум / В. А. Валентинов. - 3-е изд. - М.: Дашков и К, 2010. - 436 с.

(http://znanium.com/catalog.php?item=booksearch&code=%D1%8D%D0%BA%D0%BE%D0%BD%D0%BE%D0%BC%D0%B5%D1%82%D1%80%D0%B8%D0%BA%D0%B0&page=3#none) С. 142-181.

5.Эконометрика. Практикум: [Электронный ресурс] Учебное пособие / С.А. Бородич. - М.: НИЦ ИНФРА-М; Мн.: Нов.знание, 2014. - 329 с. (http://znanium.com/catalog.php?item=booksearch&code=%D1%8D%D0%BA%D0%BE%D0%BD%D0%BE%D0%BC%D0%B5%D1%82%D1%80%D0%B8%D0%BA%D0%B0&page=4#none) С. 133-140.

6. Электронный курс “Econometrics and Public Policy (Advanced)”, Princeton University, URL: https://blackboard.princeton.edu/webapps/portal/frameset.jsp?tab_group=courses&url=%2Fwebapps%2Fblackboard%2Fexecute%2FcourseMain%3Fcourse_id%3D_214206_1

Линейная модель множественной регрессии. Эмпирическая форма записи. На любой экономический показатель чаще всего оказывает влияние не один, а несколько факторов. Например, спрос на некоторое благо определяется не только ценой данного блага, но и ценами на замещающие и дополняющие блага, доходом потребителей и многими другими факторами. В этом случае вместо парной регрессии рассматривается множественная регрессия $\hat{y} = f\left(x_1, x_2, \ldots, x_p\right)$.

Множественная регрессия широко используется в решении проблем спроса, доходности акций, при изучении функции издержек производства, в макроэкономических расчетах и в ряде других вопросов экономики. В настоящее время множественная регрессия – один из наиболее распространенных ме-

тодов в эконометрике. Основной целью множественной регрессии является построение модели с большим числом факторов, а также определение влияния каждого фактора в отдельности и совокупного их воздействия на моделируемый показатель.

Множественный регрессионный анализ является развитием парного регрессионного анализа в случаях, когда зависимая переменная связана более чем с одной независимой переменной. Большая часть анализа является непосредственным расширением парной регрессионной модели, но здесь также появляются и некоторые новые проблемы, из которых следует выделить две. Первая проблема касается исследования влияния конкретной независимой переменной на зависимую переменную, а также разграничения еѐ воздействия и воздействий других независимых переменных. Второй важной проблемой является спецификация модели, которая состоит в том, что необходимо ответить на вопрос, какие факторы следует включить в регрессию (1), а какие – исключить из неѐ. В дальнейшем изложение общих вопросов множественного регрессионного анализа будем вести, разграничивая эти проблемы. Поэтому вначале будем полагать, что спецификация модели правильна.

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

$$y = \alpha' + \beta_1' x_1 + \beta_2' x_2 + \ldots + \beta_p' x_p + \varepsilon \qquad (1)$$

Оценка параметров модели с помощью МНК. По математическому смыслу коэффициенты β_j' в уравнении (1) равны частным производным результативного признака y по соответствующим факторам:

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

Параметр α называется свободным членом и определяет значение y в случае, когда все объясняющие переменные равны нулю. Однако, как и в случае парной регрессии, факторы по своему экономическому содержанию часто не могут принимать нулевых значений, и значение свободного члена не имеет экономи-

ческого смысла. При этом, в отличие от парной регрессии, значение каждого регрессионного коэффициента β'_j равно среднему изменению y при увеличении x_j на одну единицу лишь при условии, что все остальные факторы остались неизменными. Величина ε представляет собой случайную ошибку регрессионной зависимости. Поскольку параметры $\alpha', \beta_1', \beta_2' ..., \beta_p'$ являются случайными величинами, определить их истинные значения по выборке невозможно. Поэтому вместо теоретического уравнения регрессии оценивается так называемое эмпирическое уравнение множественной регрессии, которое можно представить в виде:

$$y = a + b_1 x_1 + b_2 x_2 + ... + b_p x_p + e \qquad (2)$$

Здесь $a, b_1, b_2, ..., b_p$ - оценки теоретических значений $\alpha', \beta_1', \beta_2' ..., \beta_p'$, или эмпирические коэффициенты регрессии, e – оценка отклонения ε. Тогда расчетное выражение имеет вид: $\hat{y} = a + b_1 x_1 + b_2 x_2 + ... + b_p x_p$

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

$$\left(x_{i1}, x_{i2}, ..., x_{ip}, y_i\right), \quad i = \overline{1, n}$$

Для однозначного определения значений параметров эмпирического уравнения множественной регрессии объем выборки n должен быть не меньше количества параметров, т.е. $n \geq p + 1$. В противном случае значения параметров не могут быть определены однозначно. Для получения надежных оценок параметров уравнения объѐм выборки должен значительно превышать количество определяемых по нему параметров. Практически, как было сказано ранее, объѐм выборки должен превышать количество параметров при x_j в уравнении (4) в 6-7 раз.

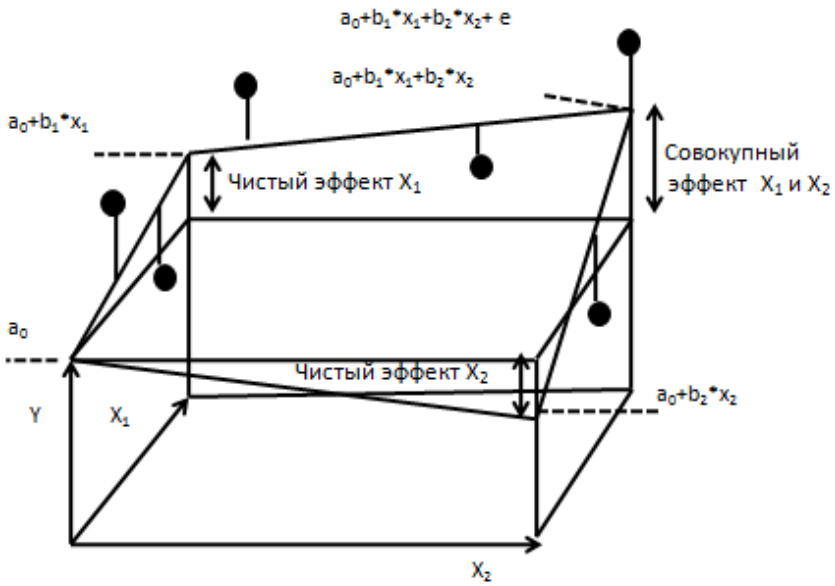


Рис.5.1. Геометрическая интерпретация линейной модели множественной регрессии

$Y = a_0 + b_1 \cdot x_1 + b_2 \cdot x_2 + e$, где

Y-общая величина расходов на питание;

X_1- располагаемый личный доход;

X_2- цена продуктов питания.

Экономическая интерпретация: При каждом увеличении располагаемого личного дохода X_1 на 1 единицу собственного измерения, расходы на питание (Y) увеличиваются на b_1 единиц измерения при сохранении постоянных цен. На каждую единицу индекса цен X_2 эти расходы уменьшаются на b_2 единиц измерения при сохранении постоянных доходов. Если $a_0>0$, то вариация расходов меньше вариации факторов; если $a_0<0$, то вариация расходов больше вариации факторов.

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

5^0.Спецификация модели имеет вид: $y = \alpha' + \beta_1' x_1 + \beta_2' x_2 + \ldots + \beta_p' x_p + \varepsilon$.

6^0.Отсутствие мультиколлинеарности: между объясняющими переменными: отсутствует строгая линейная зависимость, что играет важную роль в отборе факторов при решении проблемы спецификации модели.

7^0.Ошибки $\varepsilon_i, i = \overline{1,n}$ имеют нормальное распределение $(\varepsilon_i \sim N(0,\sigma))$. Выполнимость этого условия нужна для проверки статистических гипотез и построения интервальных оценок.

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

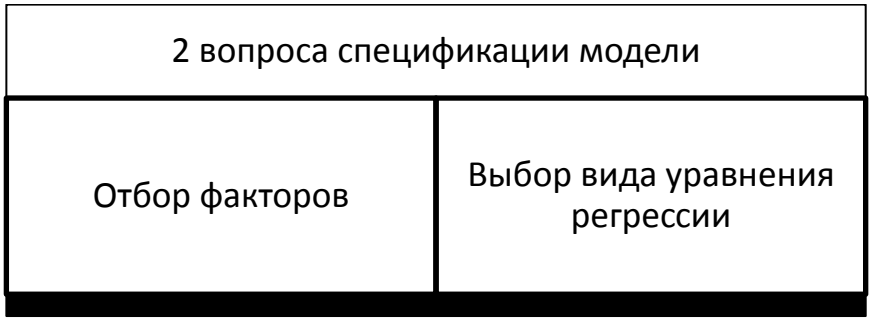


Рис.5.2. Проблемы спецификации модели

Кроме этого, факторы, включаемые во множественную регрессию, должны быть количественно измеримы.

Рассмотрим три метода расчета параметров множественной линейной регрессии.

1. *Матричный метод.* Представим данные наблюдений и параметры модели в матричной форме.

$Y = [y_1, y_2, ..., y_n]'$ - n – мерный вектор – столбец наблюдений зависимой переменной;

$B = [a, b_1, b_2, ..., b_p]'$ - (p+1) – мерный вектор – столбец параметров уравнения регрессии $y = a + b_1 x_1 + b_2 x_2 + ... + b_p x_p + e$;

$Y = [y_1, y_2, ..., y_n]'$ - n – мерный вектор – столбец отклонений выборочных значений y_i от значений $\hat{y}_i$, получаемых по уравнению $\hat{y} = a + b_1 x_1 + b_2 x_2 + ... + b_p x_p$.

Для удобства записи столбцы записаны как строки и поэтому снабжены штрихом для обозначения операции транспонирования.

Наконец, значения независимых переменных запишем в виде прямоугольной матрицы размерности $n \times (p+1)$:

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \dots & \dots & \dots & \dots \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix}$$

Каждому столбцу этой матрицы отвечает набор из n значений одного из факторов, а первый столбец состоит из единиц, которые соответствуют значениям переменной при свободном члене.

В этих обозначениях эмпирическое уравнение регрессии выглядит так: $Y = XB + e$. Отсюда вектор остатков регрессии можно выразить таким образом: $e = Y - XB$.

Таким образом, функционал $Q = \sum e_i^2$, который, собственно, и минимизируется по МНК, можно записать как произведение вектора – строки ***e'*** на вектор – столбец ***e***: $Q = e'e = (Y - XB)'(Y - XB)$.

В соответствии с МНК дифференцирование ***Q*** по вектору ***B*** приводит к выражению: $\frac{\partial Q}{\partial B} = -2X'Y + 2(X'X)B$, которое для нахождения экстремума следует приравнять к нулю. В результате преобразований получаем выражение для вектора параметров регрессии: $B = (X'X)^{-1} X'Y$. Здесь $(X'X)^{-1}$ - матрица, обратная к $X'X$.

Пример. Бюджетное обследование пяти случайно выбранных семей дало следующие результаты (в тыс. руб.):

Семья	Накопления, ***S***	Доход, ***Y***	Имущество, ***W***
1	3	40	60
2	6	55	36
3	5	45	36
4	3,5	30	15
5	1,5	30	90

Оценим регрессию ***S*** на ***Y*** и ***W***. Введем обозначения:

S=[3;6;5;3,5;1,5]’ – вектор наблюдений зависимой переменной;

B=[$a;b_1;b_2$]’ – вектор параметров уравнения регрессии;

$$X = \begin{bmatrix} 1 & 40 & 60 \\ 1 & 55 & 36 \\ 1 & 45 & 36 \\ 1 & 30 & 15 \\ 1 & 30 & 90 \end{bmatrix}$$

- матрица значений независимых переменных.

Далее с помощью матричных операций вычисляем (используем табличный процессор MS Excel и функции ТРАНСП, МУМНОЖ и МОБР в нем):

$$X'X = \begin{bmatrix} 5 & 200 & 237 \\ 200 & 8450 & 9150 \\ 237 & 9150 & 14517 \end{bmatrix}; \quad (X'X)^{-1} = \begin{bmatrix} 5{,}6916 & -0{,}1074 & -0{,}0252 \\ -0{,}1074 & 0{,}0024 & 0{,}00024 \\ -0{,}0252 & 0{,}00024 & 0{,}00033 \end{bmatrix}$$

$$B = (X'X)^{-1}X'Y = (0{,}2787 \quad 0{,}1229 \quad -0{,}0294)$$

Регрессионная модель в скалярном виде:

$$\hat{S} = 0{,}2787 + 0{,}1229Y - 0{,}0294W$$

2. Скалярный метод. При его применении строится система нормальных уравнений, решение которой и позволяет получить оценки параметров регрессии:

$$\begin{cases} an & +b_1\sum x_1 & +b_2\sum x_2 & +\ldots+b_p\sum x_p & =\sum y \\ a\sum x_1 & +b_1\sum x_1^2 & +b_2\sum x_2x_1 & .+\ldots+b_p\sum x_px_1 & =\sum yx_1 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ a\sum x_p & +b_1\sum x_1x_p & +b_2\sum x_2x_p & +\ldots+b_p\sum x_p^2 & =\sum yx_p \end{cases}$$

Решить эту систему можно любым подходящим способом, например, методом определителей или методом Гаусса. При небольшом количестве определяемых параметров использование определителей предпочтительнее.

Рассмотрим пример, приведенный выше. Здесь для двух факторов, *Y* и *W*, система нормальных уравнений запишется так:

$$\begin{cases} an & +b1\sum Y & +b2\sum W & =\sum S \\ a\sum Y & +b1\sum Y^2 & +b2\sum WY & =\sum SY \\ a\sum W & +b1\sum YW & +b2\sum W^2 & =\sum SW \end{cases}$$

Рассчитываем значения сумм, получаем:

$$\begin{cases} 5a & +200b1 & +237b2 & =19 \\ 200a & +8450b1 & +9150b2 & =825 \\ 237a & +9150b1 & +14517b2 & =863{,}5 \end{cases}$$

Рассчитаем значения определителей этой системы, используем функцию МОПРЕД в Excel:

$$\Delta = 6842700;\quad \Delta_a = 1903325;\quad \Delta_{b_1} = 840825;\quad \Delta_{b_2} = -201225.$$

Отсюда получим оценки параметров модели:

$$a = \Delta / \Delta_a = 1903325 / 6842700 = 0{,}2787;$$
$$b_1 = \Delta_{b_1} / \Delta = 840825 / 6842700 = 0{,}1229;$$
$$b_2 = \Delta_{b_2} / \Delta = -201205 / 6842700 = -0{,}0294.$$

Обратите внимание, что коэффициенты в левой части системы нормальных уравнений совпадают с соответствующими элементами матрицы $X'X$.

3. Регрессионная модель в стандартизованном масштабе. Уравнение регрессии в стандартизованном масштабе имеет вид:

$$t_y = \beta_1 t_{x_1} + \beta_2 t_{x_2} + \ldots + \beta_p t_{x_p} + \varepsilon$$

где $t_y, t_{x_1}, t_{x_2}, ..., t_{x_p}$ - стандартизованные переменные:

$$t_y = \frac{y - \bar{y}}{\sigma_y}; \quad t_{x_j} = \frac{x_j - \bar{x}_j}{\sigma_{x_j}}, \quad j = \overline{1,n}$$

для которых среднее значение равно нулю: $\bar{t}_y = \bar{t}_{x_1} = \bar{t}_{x_2} = ... = \bar{t}_{x_p} = 0$, а среднее квадратическое отклонение равно единице: $\sigma_y = \sigma_{t_{x_j}} = 1,\ j = \overline{1,n}$; β_j – стандартизованные коэффициенты регрессии, или β – коэффициенты (не следует путать их с параметрами уравнения $y = \alpha' + \beta_1' x_1 + \beta_2' x_2 + ... + \beta_p' x_p + \varepsilon$).

Применяя МНК к уравнению $t_y = \beta_1 t_{x_1} + \beta_2 t_{x_2} + \ldots + \beta_p t_{x_p} + \varepsilon$, после соответствующих преобразований получим систему нормальных уравнений:

$$\begin{cases} \beta_1 & + \beta_2 r_{x_2x_1} & + \beta_3 r_{x_3x_1} & + \beta_p r_{x_px_1} & = r_{yx_1} \\ \beta_1 r_{x_1x_2} & + \beta_2 & + \beta_3 r_{x_3x_2} & + \beta_p r_{x_px_2} & = r_{yx_2} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \beta_1 r_{x_1x_p} & + \beta_2 r_{x_2x_p} & + \beta_3 r_{x_3x_p} & + \beta_p & = r_{yx_p} \end{cases}$$

В этой системе r_{yx_j}, $r_{x_ix_j}$, $j,k = \overline{1,p}$ - элементы расширенной матрицы парных коэффициентов корреляции или, другими словами, коэффициенты парной корреляции между различными факторами или между факторами и результативным признаком. Имея измеренные значения всех переменных, вычислить матрицу парных коэффициентов корреляции на компьютере не составляет большого труда, используя, например, табличный процессор MS Excel или программу Statistica.

Решением данной системы определяются β – коэффициенты. Эти коэффициенты показывают, на сколько значений с.к.о. изменится в среднем результат, если соответствующий фактор x_j изменится на одну с.к.о. при неизменном среднем уровне других факторов. Поскольку все переменные заданы как центрированные и нормированные, β – коэффициенты сравнимы между собой. Сравнивая их друг с другом, можно ранжировать факторы по силе их воздей-

ствия на результат. В этом основное достоинство стандартизованных коэффициентов регрессии, в отличие от коэффициентов обычной регрессии, которые несравнимы между собой.

Пусть функция издержек производства y (тыс. руб.) характеризуется уравнением вида: $y = 200 + 1{,}2x_1 + 1{,}1x_2 + \varepsilon$, где факторами являются основные производственные фонды (тыс. руб.) и численность занятых в производстве (чел.). Отсюда видно, что при постоянной занятости рост стоимости основных производственных фондов на 1 тыс. руб. влечет за собой увеличение затрат в среднем на 1,2 тыс. руб., а увеличение числа занятых на одного человека при неизменной технической оснащенности приводит к росту затрат в среднем на 1,1 тыс. руб.. Однако это не означает, что первый фактор сильнее влияет на издержки производства по сравнению со вторым. Такое сравнение возможно, если обратиться к уравнению регрессии в стандартизованном масштабе. Пусть оно выглядит так: $\hat{t}_y = 0{,}5t_{x_1} + 0{,}8t_{x_2}$. Это означает, что с ростом первого фактора на одно с.к.о. при неизменном числе занятых затраты на продукцию увеличиваются в среднем на 0,5 с.к.о. Так как $\beta_1 < \beta_2$ (0,5<0,8), то можно заключить, что большее влияние на производство продукции оказывает второй фактор, а не первый, как кажется из уравнения регрессии в натуральном масштабе.

В парной зависимости стандартизованный коэффициент регрессии есть не что иное, как линейный коэффициент корреляции r. Подобно тому, как в парной зависимости коэффициенты регрессии и корреляции связаны между собой, так и во множественной регрессии коэффициенты «чистой» регрессии b_j связаны с β – коэффициентами: $b_j = \beta_j \dfrac{\sigma_y}{\sigma_{x_j}}$.

Это позволяет от уравнения регрессии в стандартизованном масштабе: $\hat{t}_y = \beta_1 t_{x_1} + \beta_2 t_{x_2} + \ldots + \beta_p t_{x_p}$ переходить к уравнению регрессии в натуральном мас-

штабе $\hat{y} = a + b_1 x_1 + b_2 x_2 + \ldots + b_p x_p$. Параметр a определяется так: $a = \bar{y} - b_1 \bar{x}_1 - b_2 \bar{x}_2 - \ldots - b_p \bar{x}_p$.

Свободный член в уравнении $\hat{t}_y = \beta_1 t_{x_1} + \beta_2 t_{x_2} + \ldots + \beta_p t_{x_p}$ отсутствует, поскольку все стандартизованные переменные имеют нулевое среднее значение.

Рассмотренный смысл стандартизованных коэффициентов регрессии позволяет использовать их при отсеве факторов – из модели исключаются факторы с наименьшим значением β_j.

Компьютерные программы построения уравнения множественной регрессии в зависимости от использованного в них алгоритма решения позволяют получить либо только уравнение регрессии для исходных данных, либо, кроме того, уравнение регрессии в стандартизованном масштабе.

В заключение приведем расчет стандартизованного уравнения регрессии по данным рассмотренного выше числового примера. Используя функцию КОРРЕЛ в Excel, рассчитаем расширенную матрицу парных коэффициентов корреляции:

$$R = \begin{bmatrix} 1 & -0{,}27149 & 0{,}873684 \\ -0{,}27149 & 1 & -0{,}68224 \end{bmatrix},$$

в которой последний столбец состоит из элементов $r_{yx_1}\,(r_{SY})$ и $r_{SW}\,(r_{yx_2})$ соответственно, а неединичные элементы в первых двух столбцах соответствуют $r_{YW}\,(r_{x_1x_2})$. Эта матрица является расширенной матрицей системы уравнений для определения β – коэффициентов:

$$\begin{cases} \beta_1 + 0{,}27149\beta_2 = 0{,}873684, \\ -0{,}27149\beta_1 + \beta_2 = -0{,}68224 \end{cases}$$

Решаем систему методом определителей, получаем:

Δ=0,926291; Δ_1=0,688461; Δ_2=-0,44504;

β_1=0,688461/0,926291=0,743245;

β_2=-0,44504/0,926291=-0,48045;

Тогда стандартизованное уравнение регрессии запишется так:

$$\hat{t}_y = 0{,}743245 t_Y - 0{,}48045 t_W$$

Отсюда видно, что первый фактор оказывает большее воздействие на результат, чем второй ($|\beta_1|>|\beta_2|$), однако эта разница не так велика, как для коэффициентов в натуральном масштабе (0,1229 и –0,0294). От этого уравнения можно перейти к уравнению в натуральном масштабе. Для этого с помощью функции СТАНДОТКЛОН в Excel определим стандартные отклонения всех переменных: $\sigma_S = 1{,}75357$; $\sigma_Y = 10{,}6066$; $\sigma_W = 28{,}6496$,

а с помощью функции СРЗНАЧ – средние значения: $\overline{S} = 3{,}8$; $\overline{Y} = 40$; $\overline{W} = 47{,}4$.

Далее определяем оценки параметров:

$$b_1 = \beta_1 \frac{\sigma_y}{\sigma_{x_1}} = 0{,}743245 \cdot \frac{1{,}75357}{10{,}6066} = 0{,}1229;$$

$$b_2 = \beta_2 \frac{\sigma_y}{\sigma_{x_2}} = -0{,}48045 \cdot \frac{1{,}75357}{28{,}6496} = -0{,}0294;$$

$$a = \bar{s} - b_1 \overline{Y} - b_2 \overline{W} = 3{,}8 - 0{,}1229 \cdot 40 + 0{,}0294 \cdot 47{,}4 = 0{,}2787.$$

Эти значения оценок совпадают с оценками, полученными ранее.

Вопросы и задания для самоконтроля

1. Как записывается эмпирическое уравнение линейной модели множественной регрессии?

2. Что измеряют коэффициенты регрессии линейной модели множественной регрессии?

3. Какие этапы включает алгоритм определения коэффициентов множественной линейной регрессии по МНК в матричной форме?

4. Какие требования предъявляются к факторам для их включения их в модель множественной регрессии?

5. Как интерпретируются коэффициенты регрессии линейной модели потребления?

6. Какой смысл приобретает сумма коэффициентов регрессии в производственных фукнциях?

7. Как в линейной модели множественной регрессии, записанной в стандартизованном виде, сравнить факторы по силе их воздействия на результат?

8. Как связаны стандартизованные коэффициенты регрессии с натуральными?

Задание 1. Получены следующие величины:

$\bar{y} = 15{,}0;\ \bar{x}_1 = 6{,}5;\ \bar{x}_2 = 12{,}0;\ \sigma_y = 4{,}0;\ \sigma_{x_1} = 2{,}5;\ \sigma_{x_2} = 3{,}5; r_{yx_1} = 0{,}63;\ r_{yx_2} = 0{,}78;\ r_{x_1x_2} = 0{,}52$. Записать регрессию y на x_1 и x_2 в стандартизованной и естественной формах.

Задание 2. Уравнение регрессии, построенное по 15 наблюдениям, имеет вид:

$\tilde{y} = 12{,}4 - 9{,}6x_1 + ?x_2 - 6{,}3x_3$

m_b (?) (3,2) (0,12) (?)

t_b (1,55) (?) (4,0) (–3,15).

Определить пропущенные значения и построить доверительный интервал для β_3 с вероятностью 0,99.

Задание 3. Уравнение регрессии в стандартизованной форме имеет вид $t_y = 0{,}37t_{x_1} - 0{,}52t_{x_2} + 0{,}43t_{x_3}$. При этом коэффициенты вариации равны: $V_y = 18\%,\ V_{x_1} = 25\%,\ V_{x_2} = 38\%,\ V_{x_3} = 30\%$. Определить частные коэффициенты эластичности.

Лекция 7

Тема 6. Оценка качества модели множественной регрессии

Вопросы для изучения:

1. Показатели качества множественной регрессии: индекс множественной корреляции и коэффициент детерминации. Скорректированный коэффициент

детерминации.

2. Оценка значимости уравнения в целом и каждого параметра в отдельности.

3. Сравнение двух регрессий при включении и при исключении отдельных наборов переменных. Частные F-критерии.

Аннотация. Данная тема раскрывает особенности оценки качества линейной модели множественной регрессии.

Ключевые слова. Индекс множественной корреляции, коэффициент детерминации, частные F-критерии.

Методические рекомендации по изучению темы

• Изучить лекционную часть, где даются общие представления по данной теме.

• Для закрепления теоретического материала ознакомиться с решениями типовых задач и ответить на вопросы для самоконтроля.

• Для проверки усвоения темы выполнить практические задания и тест для самоконтроля.

Рекомендуемые информационные ресурсы:

1. http://tulpar.kfu.ru/course/view.php?id=2213.

2 Эконометрика: [Электронный ресурс] Учеб.пособие / А.И. Новиков. - 2-е изд., испр. и доп. - М.: ИНФРА-М, 2011. - 144 с.: с. (http://znanium.com/catalog.php?item=booksearch&code=%D1%8D%D0%BA%D0%BE%D0%BD%D0%BE%D0%BC%D0%B5%D1%82%D1%80%D0%B8%D0%BA%D0%B0&page=1#none) С.50-58.

3. Уткин, В. Б. Эконометрика [Электронный ресурс] : Учебник / В. Б. Уткин; Под ред. проф. В. Б. Уткина. - 2-е изд. - М.: Издательско-торговая корпорация «Дашков и К°», 2012. - 564 с.

(http://znanium.com/catalog.php?item=booksearch&code=%D1%8D%D0%BA%D0%BE%D0%BD%D0%BE%D0%BC%D0%B5%D1%82%D1%80%D0%B8%D0%BA%D0%B0&page=4#none) С. 323-369.

4. Валентинов, В. А. Эконометрика [Электронный ресурс]: Практикум / В. А. Валентинов. - 3-е изд. - М.: Дашков и К, 2010. - 436 с.

(http://znanium.com/catalog.php?item=booksearch&code=%D1%8D%D0%BA%D0%BE%D0%BD%D0%BE%D0%BC%D0%B5%D1%82%D1%80%D0%B8%D0%BA%D0%B0&page=3#none) С. 142-181.

5.Эконометрика. Практикум: [Электронный ресурс] Учебное пособие / С.А. Бородич. - М.: НИЦ ИНФРА-М; Мн.: Нов.знание, 2014. - 329 с. (http://znanium.com/catalog.php?item=booksearch&code=%D1%8D%D0%BA%D0%BE%D0%BD%D0%BE%D0%BC%D0%B5%D1%82%D1%80%D0%B8%D0%BA%D0%B0&page=4#none) С. 133-140.

6. Электронный курс “Econometrics and Public Policy (Advanced)”, Princeton University, URL: https://blackboard.princeton.edu/webapps/portal/frameset.jsp?tab_group=courses&url=%2Fwebapps%2Fblackboard%2Fexecute%2FcourseMain%3Fcourse_id%3D_214206_1

Показатели качества множественной регрессии: индекс множественной корреляции и коэффициент детерминации. Скорректированный коэффициент детерминации. Практическая значимость уравнения множественной регрессии оценивается с помощью показателя множественной корреляции и его квадрата – коэффициента детерминации. Показатель множественной корреляции характеризует тесноту связи рассматриваемого набора факторов с исследуемым признаком, или оценивает тесноту совместного влияния факторов на результат. Независимо от формы связи показатель множественной корреляции может быть найден как индекс множественной корреляции:

$$R_{yx_1x_2\ldots x_p} = \sqrt{1 - \frac{\sigma^2{}_{îñò}}{\sigma^2{}_y}} \ .$$

Методика построения индекса множественной корреляции аналогична построению индекса корреляции для парной зависимости. Границы его изменения те же: от 0 до 1. Чем ближе его значение к 1, тем теснее связь результатив-

ного признака со всем набором исследуемых факторов.

Если обратиться к линейному уравнению множественной регрессии в стандартизованном масштабе, то для расчета индекса множественной корреляции можно использовать формулу следующего вида:

$$R_{yx_1x_2\ldots x_p} = \sqrt{\sum \beta_{x_i} \cdot r_{yx_i}}$$

Формула индекса множественной корреляции для линейной регрессии получила название линейного коэффициента множественной корреляции, или совокупного коэффициента корреляции.

При линейной зависимости определение совокупного коэффициента корреляции возможно без построения регрессии и оценки еѐ параметров, а с использованием только матрицы парных коэффициентов корреляции:

$$R_{yx_1x_2\ldots x_p} = \sqrt{1 - \frac{\Delta r}{\Delta r_{11}}},$$

где Δr – определитель матрицы парных коэффициентов корреляции:

$$\Delta r = \begin{vmatrix} 1 & r_{yx_1} & r_{yx_2} & \ldots & r_{yx_p} \\ r_{x_1y} & 1 & r_{x_1x_2} & \ldots & r_{x_1x_p} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ r_{x_py} & r_{x_px_1} & r_{x_px_2} & \ldots & 1 \end{vmatrix}$$

а Δr_{11} – определитель матрицы межфакторной корреляции:

$$\Delta r_{11} = \begin{vmatrix} 1 & r_{x_1x_2} & r_{x_1x_3} & \ldots & r_{x_1x_p} \\ r_{x_2x_1} & 1 & r_{x_2x_3} & \ldots & r_{x_2x_p} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ r_{x_px_1} & r_{x_px_2} & r_{x_px_3} & \ldots & 1 \end{vmatrix}$$

Определитель матрицы межфакторной корреляции остаѐтся после вычеркивания из матрицы коэффициентов парной корреляции первого столбца и первой строки, что и соответствует матрице коэффициентов парной корреляции между факторами.

Проверка статистического качества оцененного уравнения регрессии проводится, с одной стороны, по статистической значимости параметров уравнения, а с другой стороны, по общему качеству уравнения регрессии. Кроме этого, проверяется выполнимость предпосылок МНК.

Сначала рассмотрим первые два вида проверок и связанные с ними вопросы. Некоторые предпосылки МНК и проверки их выполнимости будем рассматривать отдельно.

Для проверки общего качества уравнения регрессии используется коэффициент детерминации R^2, который в общем случае рассчитывается по формуле: $R^2 = 1 - \dfrac{\sum e_i^2}{\sum (y_i - \bar{y})^2}$. Он показывает, как и в парной регрессии, долю общей дисперсии y, объясненную уравнением регрессии. Его значения находятся между нулем и единицей. Чем ближе этот коэффициент к единице, тем больше уравнение регрессии объясняет поведение y.

Для множественной регрессии R^2 является неубывающей функцией числа объясняющих переменных. Добавление новой объясняющей переменной никогда не уменьшает значение R^2. Действительно, каждая следующая объясняющая переменная может лишь дополнить, но никак не сократить информацию, объясняющую поведение зависимой переменной.

В формуле расчета коэффициента детерминации используется остаточная дисперсия, которая имеет систематическую ошибку в сторону уменьшения, тем более значительную, чем больше параметров определяется в уравнении регрессии при заданном объѐме наблюдений n. Если число параметров (p+1) приближается к n, то остаточная дисперсия будет близка к нулю и коэффициент детерминации приблизится к единице даже при слабой связи факторов с результатом.

Поэтому в числителе и знаменателе делается поправка на число степеней свободы остаточной и общей дисперсии соответственно и рассчитывается скорректированный коэффициент детерминации:

$$\overline{R}^2 = 1 - \frac{\sum e_i^2 / (n - p - 1)}{\sum (y_i - \bar{y})^2 / (n - 1)}$$

Поскольку величина обычного коэффициента детерминации, как правило, увеличивается при добавлении объясняющей переменной к уравнению регрессии даже без достаточных на то оснований, скорректированный коэффициент детерминации компенсирует это увеличение путем наложения «штрафа» за увеличение числа независимых переменных. Перепишем формулу скорректированного коэффициента детерминации следующим образом:

$$\overline{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-p-1} = \frac{n-1}{n-p-1}R^2 - \frac{p}{n-p-1} = R^2 - \frac{p}{n-p-1}(1 - R^2)$$

По мере роста p увеличивается отношение $p/(n\text{-}p\text{-}1)$ и, следовательно, возрастает размер корректировки коэффициента R^2 в сторону уменьшения.

Очевидно, что $\overline{R}^2 < R^2$ при $p>1$. С ростом p $\overline{R}^2$ растет медленнее, чем R^2. Другими словами, он корректируется в сторону уменьшения с ростом числа объясняющих переменных. При этом $\overline{R}^2 = R^2$ только при $R^2=1$. $\overline{R}^2$ может даже принимать отрицательные значения (например, при $R^2=0$). Поэтому для корректировки формулы скорректированного коэффициента детерминации нет строгого математического обоснования.

Доказано, что $\overline{R}^2$ увеличивается при добавлении новой объясняющей переменной тогда и только тогда, когда t – статистика для этой переменной по модулю больше единицы. Из этого отнюдь не следует, как можно было бы предположить, что увеличение $\overline{R}^2$ означает улучшение спецификации уравнения. Тем не менее, добавление в модель новых факторов осуществляется до тех пор, пока растет скорректированный коэффициент детерминации.

Обычно приводятся данные как по R^2 , так и по $\overline{R}^2$, являющиеся суммарными мерами общего качества уравнения регрессии. Однако не следует абсолютизировать значимость коэффициентов детерминации. Существует немало примеров неправильно построенных моделей, имеющих высокие коэффициенты детерминации. Поэтому коэффициент детерминации в настоящее время рассматривается лишь как один из ряда показателей, которые нужно проанализировать, чтобы уточнить строящуюся модель.

Оценка значимости уравнения в целом и каждого параметра в отдельности. Анализ статистической значимости коэффициента детерминации проводится на основе проверки нуль – гипотезы H_0: R^2=0 против альтернативной гипотезы H_1: R^2>0. Для проверки данной гипотезы используется следующая F – статистика:

$$F = \frac{R^2}{1-R^2} \cdot \frac{n-p-1}{p}$$

Величина F при выполнении предпосылок МНК и при справедливости нуль – гипотезы имеет распределение Фишера. Из формулы расчета F-статистики видно, что показатели F и R^2 равны или не равны нулю одновременно. Если F=0, то R^2=0, и линия регрессии $y = \bar{y}$ является наилучшей по МНК, и, следовательно, величина y линейно не зависит от $x_1, x_2, ..., x_p$. Для проверки нуль – гипотезы при заданном уровне значимости α по таблицам критических точек распределения Фишера находится критическое значение $F_{табл}(\alpha;\ p;\ n\text{-}p\text{-}1)$. Если F>$F_{\text{табл}}$, нуль – гипотеза отклоняется, что равносильно статистической значимости R^2, т.е. R^2>1.

Эквивалентный анализ может быть предложен рассмотрением другой нуль – гипотезы, которая формулируется как $H_0 : \beta_1' = \beta_2' = ... = \beta_p' = 0$. Эту гипотезу можно назвать гипотезой об общей значимости уравнения регрессии. Если данная гипотеза не отклоняется, то делается вывод о том, что совокупное

влияние всех p объясняющих переменных $x_1, x_2, ..., x_p$ на зависимую переменную y можно считать статистически несущественным, а общее качество уравнения регрессии невысоким.

Проверка такой гипотезы осуществляется на основе дисперсионного анализа сравнения объясненной и остаточной дисперсий, т.е. нуль – гипотеза формулируется как $H_0: D_{факт}=D_{ост}$ против альтернативной гипотезы $H_1: D_{факт}>D_{ост}$. При этом строится F – статистика:

$$F = \frac{\sum(\hat{y}_i - \bar{y})^2 / p}{\sum(y_i - \hat{y}_i)^2 / (n-p-1)}$$

Здесь в числителе – объясненная (факторная) дисперсия в расчете на одну степень свободы (число степеней свободы равно числу факторов, т.е. p). В знаменателе – остаточная дисперсия на одну степень свободы. Еѐ число степеней свободы равно (n-p-1). Потеря (p+1) степени свободы связана с необходимостью решения системы (p+1) линейных уравнений при определении параметров эмпирического уравнения регрессии. Если учесть, что число степеней свободы общей дисперсии равно (n-1), то число степеней свободы объясненной дисперсии равна разности (n-1) – (n-p-1), т.е. p. Следует отметить, что выражение $F = \frac{\sum(\hat{y}_i - \bar{y})^2 / p}{\sum(y_i - \hat{y}_i)^2 / (n-p-1)}$ эквивалентно выражению $F = \frac{R^2}{1-R^2} \cdot \frac{n-p-1}{p}$. Это становится ясно, если числитель и знаменатель $F = \frac{\sum(\hat{y}_i - \bar{y})^2 / p}{\sum(y_i - \hat{y}_i)^2 / (n-p-1)}$ разделить на общую СКО:

$$F = \frac{\sum(\hat{y}_i - \bar{y})^2 / \sum(y_i - \bar{y})^2}{\sum(y_i - \hat{y}_i)^2 / \sum(y_i - \bar{y})^2} \cdot \frac{n-p-1}{p} = \frac{R^2}{1-R^2} \cdot \frac{n-p-1}{p}$$

Поэтому методика принятия или отклонения нуль – гипотезы для статистики $F = \dfrac{\sum(\hat{y}_i - \bar{y})^2 / p}{\sum(y_i - \hat{y}_i)^2 / (n-p-1)}$ ничем не отличается от таковой для статистики $F = \dfrac{R^2}{1-R^2} \cdot \dfrac{n-p-1}{p}$.

Анализ статистики F позволяет сделать вывод о том, что для принятия гипотезы об одновременном равенстве нулю всех коэффициентов линейной регрессии коэффициент детерминации R^2 должен существенно отличаться от нуля. Его критическое значение уменьшается при росте числа наблюдений и может стать сколь угодно малым.

Например, пусть при оценке регрессии с двумя объясняющими переменными по 30 наблюдениям R^2 =0,65. Тогда $F = \dfrac{0{,}65}{0{,}35} \cdot \dfrac{30-2-1}{2} \approx 25{,}07$

По таблицам критических точек распределения Фишера найдем F(0,05; 2; 27)=3,36; F(0,01; 2; 27)=5,49. Поскольку $F_{\text{набл}}$=25,05>$F_{\text{кр}}$ как при 5% - ном, так и при 1% - ном уровне значимости, то нулевая гипотеза в обоих случаях отклоняется. Если в той же ситуации R^2=0,4, то $F = \dfrac{0{,}65}{0{,}35} \cdot \dfrac{30-2-1}{2} \approx 25{,}07$. Предположение о незначимости связи отвергается и здесь.

Как и в случае парной регрессии, статистическая значимость параметров множественной линейной регрессии с p факторами проверяется на основе t – статистики: $t_{b_j} = \dfrac{b_j}{m_{b_j}}$ $\left(\text{или } t_a = \dfrac{a}{m_a}\right)$, где величина $m_{b_j}\left(m_a\right)$ называется стандартной ошибкой параметра $b_j(a)$. Она определяется так. Обозначим матрицу: $Z^{-1} = \left(X'X\right)^{-1}$, и в этой матрице обозначим j – й диагональный элемент как z_{jj}'. Тогда выборочная дисперсия эмпирического параметра регрессии равна: $m_{b_j}^2 = s^2 z_{jj}'$, $j = \overline{1,p}$, а для свободного члена выражение имеет вид:

$m_a^2 = s^2 z_{00}'$, если считать, что в матрице Z^{-1} индексы изменяются от 0 до p.

Здесь S^2 – несмещенная оценка дисперсии случайной ошибки ε: $s^2 = \frac{\sum e_i^2}{n-p-1}$.

Стандартные ошибки параметров регрессии равны: $m_{b_j} = \sqrt{m_{b_j}^2}$ $\left(\text{или } m_a = \sqrt{m_a^2}\right)$.

Полученная по выражению $t_{b_j} = \frac{b_j}{m_{b_j}}$ $\left(\text{или } t_a = \frac{a}{m_a}\right) t$ – статистика для соответствующего параметра имеет распределение Стьюдента с числом степеней свободы (n-p-1). При требуемом уровне значимости α эта статистика сравнивается с критической точкой распределения Стьюдента $t(\alpha; n\text{-}p\text{-}1)$ (двухсторонней). Если $|t|>t(\alpha; n\text{-}p\text{-}1)$, то соответствующий параметр считается статистически значимым, и нуль – гипотеза в виде $H_0: b_j=0$ или $H_0: a=0$ отвергается. В противном случае ($|t|<t(\alpha; n\text{-}p\text{-}1)$) параметр считается статистически незначимым, и нуль – гипотеза не может быть отвергнута. Поскольку b_j не отличается значимо от нуля, фактор x_j линейно не связан с результатом. Его наличие среди объясняющих переменных не оправдано со статистической точки зрения. Не оказывая какого – либо серьѐзного влияния на зависимую переменную, он лишь искажает реальную картину взаимосвязи. Поэтому после установления того факта, что коэффициент b_j статистически незначим, переменную x_j рекомендуется исключить из уравнения регрессии. Это не приведет к существенной потере качества модели, но сделает еѐ более конкретной.

Строгую проверку значимости параметров можно заменить простым сравнительным анализом.

Если $|t| \le 1$, т.е. $b_j < m_{b_j}$, то коэффициент статистически незначим.

Если $1 < |t| \le 2$, т.е. $b_j < 2m_{b_j}$, то коэффициент относительно значим. В данном случае рекомендуется воспользоваться таблицей критических точек распределения Стьюдента.

Если $2 < |t| \le 3$, то коэффициент значим. Это утверждение является гарантированным при (n-p-1)>20 и $\alpha \ge 0{,}05$.

Если $|t|>3$, то коэффициент считается сильно значимым. Вероятность ошибки в данном случае при достаточном числе наблюдений не превосходит 0,001.

К анализу значимости коэффициента b_j можно подойти по – другому. Для этого строится интервальная оценка соответствующего коэффициента. Если задать уровень значимости α, то доверительный интервал, в который с вероятностью (1-α) попадает неизвестное значение параметра $\beta_j'(\alpha')$, определяется неравенством:

$$b_j - t(\alpha; n-p-1)\cdot m_{b_j} < \beta_j' < b_j + t(\alpha; n-p-1)\cdot m_{b_j}$$

или

$$a - t(\alpha; n-p-1)\cdot m_a < \alpha' < a + t(\alpha; n-p-1)\cdot m_a$$

Если доверительный интервал не содержит нулевого значения, то соответствующий параметр является статистически значимым, в противном случае гипотезу о нулевом значении параметра отвергать нельзя.

Сравнение двух регрессий при включении и при исключении отдельных наборов переменных. Частные F-критерии. Другим важным направлением использования статистики Фишера является проверка гипотезы о равенстве нулю не всех коэффициентов регрессии одновременно, а только некоторой части этих коэффициентов. Это позволяет оценить обоснованность исключения или добавления в уравнение регрессии некоторых наборов факторов, что особенно важно при совершенствовании линейной регрессионной модели.

Пусть первоначально построенное по n наблюдениям уравнение регрессии имеет вид $\hat{y} = a + b_1 x_1 + b_2 x_2 + \ldots + b_p x_p$, и коэффициент детерминации для этой модели равен R_1^2. Исключим из рассмотрения k объясняющих переменных. Не нарушая общности, предположим, что это будут k последних переменных. По первоначальным n наблюдениям для оставшихся факторов построим другое уравнение регрессии: $\hat{y} = c + d_1 x_1 + d_2 x_2 + \ldots + d_{p-k} x_{p-k}$, для которого коэффициент детерминации равен R_2^2. Очевидно, $R_2^2 \le R_1^2$, т.к. каждая дополнительная переменная объясняет часть рассеивания зависимой переменной. Проверяя гипотезу $H_0 : R_1^2 - R_2^2 = 0$, можно определить, существенно ли ухудшилось качество описания поведения зависимой переменной. Для этого используют статистику:

$$F = \frac{R_1^2 - R_2^2}{1 - R_1^2} \cdot \frac{n - p - 1}{k}$$

В случае справедливости H_0 приведенная статистика имеет распределение Фишера с числом степеней свободы p и (n-p-1). Здесь $R_2^2 \le R_1^2$ - потеря качества уравнения в результате отбрасывания k факторов; k – число дополнительно появившихся степеней свободы; $\left(1 - R_1^2\right)/\left(n - p - 1\right)$ - необъясненная дисперсия первоначального уравнения.

Если величина $F = \frac{R_1^2 - R_2^2}{1 - R_1^2} \cdot \frac{n - p - 1}{k}$ превосходит критическое $F_{кр} = F\left(\alpha; k; n - p - 1\right)$ на требуемом уровне значимости α, то нуль – гипотеза должна быть отклонена. В этом случае одновременное исключение из рассмотрения k объясняющих переменных некорректно, т.к. R_1^2 существенно превышает R_2^2. Это означает, что общее качество первоначального уравнения регрессии существенно лучше качества уравнения регрессии с отброшенными пере-

менными, т.к. первоначальное уравнение объясняет гораздо большую долю разброса зависимой переменной. Если же, наоборот, $F_{\text{набл}}<F_{\text{кр}}$, это означает что разность $R_1^2 - R_2^2$ незначительна и можно сделать вывод о целесообразности одновременного отбрасывания k факторов, поскольку это не привело к существенному ухудшению общего качества уравнения регрессии. Тогда нуль – гипотеза не может быть отброшена.

Аналогичные рассуждения можно использовать и для проверки обоснованности включения новых k факторов. В этом случае рассматривается следующая статистика:

$$F = \frac{R_2^2 - R_1^2}{1 - R_2^2} \cdot \frac{n - p - 1}{k}$$

Если она превышает критическое значение $F_{\text{кр}}$, то включение новых факторов объясняет существенную часть не объясненной ранее дисперсии зависимой переменной. Поэтому такое добавление оправдано. Добавлять переменные, как правило, целесообразно по одной. Кроме того, при добавлении факторов логично использовать скорректированный коэффициент детерминации, т.к. обычный R^2 всегда растет при добавлении новой переменной, а в скорректированном $\overline{R}^2$ одновременно растет величина p, уменьшающая его. Если увеличения доли объясненной дисперсии при добавлении новой переменной незначительно, то $\overline{R}^2$ может уменьшиться. В этом случае добавление указанного фактора нецелесообразно.

Вопросы и задания для самоконтроля

1. Как определяется статистическая значимость коэффициентов регрессии в линейной модели множественной регрессии?
2. В чем недостаток использования коэффициента детерминации при оценке общего качества линейной модели множественной регрессии?
3. Как корректируется коэффициент детерминации?

4. Как проверяется адекватность линейной модели множественной регрессии в целом?

5. Как определяется индекс множественной корреляции и какой он имеет смысл?

6. Как проверить обоснованность исключения части переменных из уравнения регрессии?

7. Как проверить обоснованность включения группы новых переменных в уравнение регрессии?

8. Что такое частный F-критерий и чем он отличается от последовательного F-критерия?

Задание 1. На основе статистических данных за 10 лет оценены параметры и их стандартные ошибки для линейной модели, описывающей зависимость объемов производства y от количества работающих x_1 и установочной мощности оборудования x_2:

$$\tilde{y} = 54 + 23{,}41x_1 + 6{,}44x_2$$

$$(6{,}5)\ (5{,}1) \qquad (0{,}83)$$

Установить для уровня значимости $\alpha = 0{,}05$, оказывают ли объясняющие переменные x_1, x_2 существенное влияние на объясняемую переменную y.

Задание 2. Имеются данные регрессионного анализа чистого дохода в зависимости от стоимости капитала и численности служащих по 20 фирмам:

Множественный R	?
R-квадрат	?
Нормированный R-квадрат	?
Стандартная ошибка	1,249
Наблюдения	20

	df	*SS*	*MS*	*F*
Регрессия	?	30,821	?	?
Остаток	?	26,537	?	
Итого	?	57,358		

	Коэффициенты	*Стандартная ошибка*	*t-статистика*	*P-Значение*

Y-пересечение	1,706	0,463	?	0,002
X1	0,072	0,016	?	0,0003
X2	-0,002	0,002	?	0,202

1) записать линейное уравнение множественной регрессии и пояснить экономический смысл его параметров;

2) оценить качество уравнения и проверить значимость коэффициентов регрессии и R^2 при $\alpha=0,05$.

Лекция 8

Тема 7. Мультиколлинеарность

Вопросы для изучения

1. Понятие мультиколлинеарности, ее причины и последствия.

2. Обнаружение мультиколлинеарности и способы ее устранения или снижения.

Аннотация. Данная тема раскрывает понятие мультиколлинеарности, ее причины и последствия, способы обнаружения и устранения.

Ключевые слова. Мультиколлинеарность, совершенная мультиколлинеарность.

Методические рекомендации по изучению темы

•Изучить лекционную часть, где даются общие представления по данной теме.

• Для закрепления теоретического материала ознакомиться с решениями типовых задач и ответить на вопросы для самоконтроля.

• Для проверки усвоения темы выполнить практические задания и тест для самоконтроля.

Рекомендуемые информационные ресурсы:

1. http://tulpar.kfu.ru/course/view.php?id=2213

2. Эконометрика: [Электронный ресурс] Учеб. пособие / А.И. Новиков. - 2-е изд., испр. и доп. - М.: ИНФРА-М, 2011. - 144 с.: (http://znanium.com/catalog.php?item=booksearch&code=%D1%8D%D0%BA%D0

%BE%D0%BD%D0%BE%D0%BC%D0%B5%D1%82%D1%80%D0%B8%D0%BA%D0%B0&page=1#none) С. 69-70.

4. Валентинов, В. А. Эконометрика [Электронный ресурс]: Практикум / В. А. Валентинов. - 3-е изд. - М.: Дашков и К, 2010. - 436 с.

(http://znanium.com/catalog.php?item=booksearch&code=%D1%8D%D0%BA%D0%BE%D0%BD%D0%BE%D0%BC%D0%B5%D1%82%D1%80%D0%B8%D0%BA%D0%B0&page=3#none) С. 142-181.

5.Эконометрика. Практикум: [Электронный ресурс] Учебное пособие / С.А. Бородич. - М.: НИЦ ИНФРА-М; Мн.: Нов.знание, 2014. - 329 с. (http://znanium.com/catalog.php?item=booksearch&code=%D1%8D%D0%BA%D0%BE%D0%BD%D0%BE%D0%BC%D0%B5%D1%82%D1%80%D0%B8%D0%BA%D0%B0&page=4#none) С. 244-253.

6. Электронный курс “Econometrics and Public Policy (Advanced)”, Princeton University, URL: https://blackboard.princeton.edu/webapps/portal/frameset.jsp?tab_group=courses&url=%2Fwebapps%2Fblackboard%2Fexecute%2FcourseMain%3Fcourse_id%3D_214206_1

Понятие мультиколлинеарности, ее причины и последствия. Мультиколлинеарность - это линейная взаимосвязь двух или нескольких объясняющих переменных (x_1, x_2, … x_m). Если объясняющие переменные связаны строгой функциональной зависимостью, то говорят о совершенной мультиколлинеарности. Мультиколлинеарность не позволяет однозначно разделить вклады объясняющих переменных x1,x2,…xm в их влияние на зависимую переменную Y.

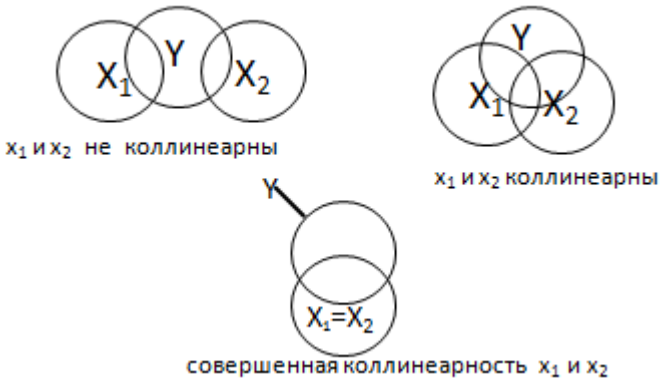


Рис.5.3. Диаграмма Венна

Коррелированность регрессоров обусловливает существенное усложнение процедуры поиска наилучшего уравнения регрессии, так как любое изменение состава регрессоров приводит к необходимости пересчитывать все параметры заново.Если факторы связаны строгой функциональной зависимостью, то это свидетельствует о полной (совершенной, строгой) мультиколлинеарности. Полная мультиколлинеарность не позволяет однозначно оценить параметры исходной модели и разделить вклады регрессоров в зависимую переменную Y. Наличие линейно связанных регрессоров относят к ошибкам спецификации. Такие ошибки при двух линейно связанных регрессорах встречаются крайне редко и легко могут быть выявлены при анализе матрицы парных коэффициентов корреляции.

Чаще возникают ошибки, обусловленные включением в модель факторов, один из которых является линейной комбинацией нескольких других. Так, при использовании количественных показателей, характеризующих часть какой-либо величины, в число объясняющих переменных нельзя включать все составляющие этой величины, так как при этом одну из них можно определить путем вычитания из этой величины значений остальных факторов. Например, в линейной регрессионной модели оборота банка (Y) недопустимым является одновременное использование в модели следующих независимых переменных: сумма кредитов, выданных юридическим лицам (X1), сумма кредитов, выданных физическим лицам (X2), общая сумма кредитов, выданных банком (X3=X1+X2). В регрессионной модели $\hat{Y} = a + b_1X_1 + b_2X_2 + b_3X_3$ увеличение значений коэффициентов при первых двух регрессорах на произвольную константу c и уменьшение на эту же константу значения коэффициента при третьем регрессоре не приведет к изменению значения зависимой переменной. Это означает, что при одних и тех же значениях регрессоров и зависимой переменной существует множество различных значений параметров уравнения.

Последствия мультиколлинеарности: увеличиваются стандартные ошибки оценок; уменьшаются t-статистики МНК-оценок регрессии; МНК-оценки чувствительны к изменениям данных; возможность неверного знака МНК-оценок;

трудность в определении вклада независимых переменных в дисперсию зависимой переменной. В реальных эконометрических исследованиях мультиколлинеарность чаще проявляется в стохастической форме, когда между хотя бы двумя объясняющими переменными существует тесная корреляционная связь. Иногда такой вид мультиколлинеарности называют частичной (несовершенной, реальной, скрытой, неполной). Матрица $X'X$ в этом случае является неособенной (близкой к вырожденной), имеет полный ранг, но ее определитель очень мал, т.е. близок к нулю. Такие матрицы ещѐ называют плохо обусловленными.

Частичная мультиколлинеарность приводит к следующим последствиям: Увеличение дисперсий оценок параметров. Это расширяет интервальные оценки и ухудшает их точность. Уменьшение t-статистик коэффициентов, что приводит к неоправданному выводу о значимости регрессоров. Неустойчивость МНК – оценок параметров и их дисперсий: небольшое изменение исходных данных (добавление или исключение одного – двух наблюдений) будет приводить к значительному изменению этих оценок. Возможность получения неверного (с точки зрения теории) знака у параметра регрессии или неоправданно большого значения этого параметра. В результате получаются значительные средние квадраты отклонения коэффициентов регрессии a, b1, b2, b3 … bp и оценка их значимости по t-критерию Стьюдента не имеет смысла, хотя в целом регрессионная модель может оказаться значимой по F-критерию.

Обнаружение мультиколлинеарности и способы ее устранения или снижения. Наиболее простой формой сильной взаимосвязи факторов является высокая парная корреляция регрессоров. Она может быть выявлена при анализе матрицы парных коэффициентов корреляции. Обычно факторы считаются тесно связанными, если значения выборочных парных коэффициентов корреляции $\left|r_{x_i x_j}\right| > (0{,}7 \ldots 0{,}8)$. При наличии такой тесной связи для какой-либо пары признаков обычно рекомендуется не включать в модель один из них, если это допустимо с точки зрения корректности модели.

Действительная мультиколлинеарность в полном смысле слова возникает при наличии тесной взаимосвязи множества независимых переменных. Она может и не обнаруживаться по матрице парных коэффициентов корреляции. В отсутствие тесной корреляционной связи одного из признаков с каждым из остальных может наблюдаться тесная связь с их совокупностью. Такую связь можно выявить путем углубленного корреляционного анализа. Он состоит в том, что при значениях множественного коэффициента корреляции какого-либо j – го независимого фактора с остальными регрессорами модели Rj ≥ (0,7…0,8) можно говорить о наличии проблемы мультиколлинеарности. Основная проблема заключается в том, что расчет множественных коэффициентов корреляции каждого из регрессоров с совокупностью остальных факторов модели может не дать нужного результата, поскольку наличие мультиколлинеарности в этой совокупности искажает и результат оценки степени взаимосвязи независимых переменных.Поскольку заранее корреляционная структура данных, как правило, неизвестна, это приводит к необходимости рассчитывать большое число множественных коэффициентов корреляции, начиная с анализа взаимосвязи одного признака со всеми возможными парами из остальных, затем с тройками признаков и т.д. Такой анализ становится очень трудоемким и редко используется на практике.

Признаки мультиколлинеарности: высокий R2; близкая к 1 парная корреляция между малозначимыми независимыми переменными; высокие частные коэффициенты корреляции; сильная дополнительная регрессия между независимыми переменными.

Методы устранения мультиколлинеарности: исключение из модели коррелированных переменных (при отборе факторов); сбор дополнительных данных или новая выборка; изменение спецификации модели; использование предварительной информации о параметрах; преобразование переменных.

Мультиколлинеарность чаще всего обнаруживает себя в ходе регрессионного анализа. К ее признакам можно отнести следующие:

1) значительные изменения коэффициентов при регрессорах при изменениях состава регрессоров и объектов, входящих в выборку;

2) незначимость большинства или всех коэффициентов при значимости уравнения в целом;

3) чрезмерно высокие или противоречащие экономической теории значения коэффициентов регрессионной модели.

Таким образом, точных количественных критериев для определения наличия или отсутствия мультиколлинеарности не существует. Тем не менее, ее наличие можно обнаружить с помощью:

1. Анализа корреляционной матрицы между объясняющими переменными и выявлении пар переменных, имеющих высокие коэффициенты корреляции.

2. Расчета множественных коэффициентов корреляции (коэффициентов детерминации) между одной из объясняющих переменных и некоторой группы из них. Наличие высокого множественного коэффициента детерминации свидетельствует о мультиколлинеарности.

3. Проверки чувствительности (устойчивости) оценок коэффициентов к небольшим изменениям исходных данных.

4. Исследования матрицы $(X'X)$. Если определитель матрицы $(X'X)$ либо ее минимальное собственное значение λ_{min} близки к нулю, то это говорит о наличии мультиколлинеарности. Об этом же может свидетельствовать и значительное отклонение максимального собственного значения $\lambda_{\max}$ матрицы $(X'X)$ от ее минимального собственного значения λ_{min}.

Одним из способов устранения мультиколлинеарности является исключение переменных из модели. Самым простым, но далеко не всегда возможным является способ, когда из двух объясняющих переменных, имеющих высокий коэффициент корреляции (обычно больше 0,8), одну переменную исключают из рассмотрения. При этом в первую очередь на основании экономических соображений решают, какую переменную оставить, а какую удалить из анализа.

Если с экономической точки зрения ни одной из переменных нельзя отдать предпочтение, то оставляют ту из двух переменных, которая имеет больший коэффициент корреляции с зависимой переменной.

Более углубленный анализ регрессоров можно получить, используя метод дополнительной регрессии. Его суть заключается в том, что для выявления списка зависимых регрессоров проводится дополнительная регрессия – регрессия каждого независимого фактора X_j, $j=1,2,...p$ на оставшиеся независимые факторы. Стандартным способом, на основе F-статистики, проверяется статистическая значимость коэффициентов детерминации R_j^2 дополнительных регрессий:

$$F_j = \frac{R_j^2}{1-R_j^2} \cdot \frac{n-p}{p-1}$$

где n – число наблюдений, p – число независимых переменных в первоначальной спецификации регрессионной модели. Статистика F_j имеет распределение Фишера с параметрами: $\nu_1 = p-1$, $\nu_2 = n-p$. Если коэффициент R_j^2 статистически не значим, то регрессор X_j не приводит к мультиколлинеарности и его оставляют в списке переменных модели. В противном случае рекомендуется исключить его из списка.

В ряде случаев можно попытаться изменить спецификацию модели: либо изменить форму модели, либо добавить объясняющие переменные, не учтенные в первоначальной модели, но существенно влияющие на зависимую переменную. В результате уменьшается сумма квадратов отклонений, а, следовательно, сокращается стандартная ошибка регрессии. В свою очередь это приводит к уменьшению стандартных ошибок параметров модели.

Вопросы и задания для самоконтроля

1. В чем различие терминов "коллинеарность" и "мультиколлинеарность"?
2. Каковы причины и последствия мультиколлинеарности?
3. Как можно обнаружить мультиколлинеарность?

4. Каковы основные методы устранения мультиколлинеарности?

5. Каковы основные типы процедур пошагового отбора переменных в регрессионную модель?

6. Действительно ли, что при наличии высокой мультиколлинеарности невозможно оценить статистическую значимость коэффициентов регрессии при коррелированных переменных?

Задание 1. По выборке n=50 для X1, X2, X3 построена следующая корреляционная матрица

$$R = \begin{bmatrix} 1{,}0 & 0{,}45 & -0{,}35 \\ 0{,}45 & 1{,}0 & 0{,}52 \\ -0{,}35 & 0{,}52 & 1{,}0 \end{bmatrix}$$

1) оценить статистическую значимость следующих частных коэффициентов корреляции r_{12*3}, r_{23*1}, r_{13*2}.

2) ответить на вопрос: при рассмотрении какой регрессии будет иметь место мультиколлинеарность?

Задание 2. Имеется выборка из10 наблюдений за переменными X1,X2,Y:

X1	1	2	3	4	5	6	7	8	9	10
X2	1	1,6	2,2	2,8	3,4	4	4,6	5,2	5,6	6,2
Y	0	3	6	9	12	15	18	21	24	27

1) ответить на вопрос: можно ли по этим данным по МНК оценить коэффициенты регрессии с двумя объясняющими переменными?

2) предложить преобразования, которые позволят оценить коэффициенты регрессии в случае отрицательного ответа на вопрос.

Лекция 9

Тема 8. Гетероскедастичность

Вопросы для изучения:

1. Понятие и последствия гетероскедастичности.

2. Методы обнаружения гетероскедастичности.

3. Коррекция на гетероскедастичность.

Аннотация. Данная тема раскрывает способы проверки соблюдения второй предпосылки МНК в остатках регрессии.

Ключевые слова. Гетероскедастичность, гомоскедастичность, остатки регрессии, метод взвешенных наименьших квадратов.

Методические рекомендации по изучению темы

•Изучить лекционную часть, где даются общие представления по данной теме.

• Для закрепления теоретического материала ознакомиться с решениями типовых задач и ответить на вопросы для самоконтроля.

• Для проверки усвоения темы выполнить практические задания и тест для самоконтроля.

Рекомендуемые информационные ресурсы:

1. http://tulpar.kfu.ru/course/view.php?id=2213.

2. Эконометрика: [Электронный ресурс] Учеб.пособие / А.И. Новиков. - 2-е изд., испр. и доп. - М.: ИНФРА-М, 2011. - 144 с.: с. (http://znanium.com/catalog.php?item=booksearch&code=%D1%8D%D0%BA%D0%BE%D0%BD%D0%BE%D0%BC%D0%B5%D1%82%D1%80%D0%B8%D0%BA%D0%B0&page=1#none) С. 92-106.

3. Валентинов, В. А. Эконометрика [Электронный ресурс]: Практикум / В. А. Валентинов. - 3-е изд. - М.: Дашков и К, 2010. - 436 с.

(http://znanium.com/catalog.php?item=booksearch&code=%D1%8D%D0%BA%D0%BE%D0%BD%D0%BE%D0%BC%D0%B5%D1%82%D1%80%D0%B8%D0%BA%D0%B0&page=3#none) С. 202-229.

4. Уткин, В. Б. Эконометрика [Электронный ресурс] : Учебник / В. Б. Уткин; Под ред. проф. В. Б. Уткина. - 2-е изд. - М.: Издательско-торговая корпорация «Дашков и К°», 2012. - 564 с.

(http://znanium.com/catalog.php?item=booksearch&code=%D1%8D%D0%BA%D0%BE%D0%BD%D0%BE%D0%BC%D0%B5%D1%82%D1%80%D0%B8%D0%BA%D0%B0&page=4#none) С. 369-383.

5. Эконометрика. Практикум: [Электронный ресурс] Учебное пособие / С.А. Бородич. - М.: НИЦ ИНФРА-М; Мн.: Нов.знание, 2014. - 329 с. (http://znanium.com/catalog.php?item=booksearch&code=%D1%8D%D0%BA%D0%BE%D0%BD%D0%BE%D0%BC%D0%B5%D1%82%D1%80%D0%B8%D0%BA%D0%B0&page=4#none) С. 197-244.

Понятие и последствия гетероскедастичности. Гетероскедастичностью остатков называется нарушение 2 предпосылки МНК о постоянстве дисперсий случайных отклонений. Если предпосылка МНК о том, что $D(\varepsilon_i)=D(\varepsilon_j)=\sigma^2$ соблюдена, то имеет место гомоскедастичность случайных отклонений. Последствия гетероскедастичности: МНК-оценки сохраняют свойства несмещенности и линейности, но теряют свойство эффективности; дисперсии МНК-оценок смещены; t-статистика и F-статистика завышены. В качестве примера реальной гетероскедастичности можно привести то, что люди с большим доходом не только тратят в среднем больше, чем люди с меньшим доходом, но и разброс в их потреблении также больше, поскольку они имеют больше простора для распределения дохода.

В ряде случаев, зная характер исходных данных, можно предвидеть гетероскедастичность и попытаться устранить еѐ ещѐ на стадии спецификации. Однако значительно чаще эту проблему приходится решать после построения уравнения регрессии.

Методы обнаружения гетероскедастичности. Графическое построение отклонений от эмпирического уравнения регрессии позволяет визуально определить наличие гетероскедастичности. В этом случае по оси абсцисс откладываются значения объясняющей переменной x_i (для парной регрессии) либо линейную комбинацию объясняющих переменных:

$$\hat{y}_i = a + b_1 x_{i1} + ... + b_p x_{ip}, \quad i = \overline{1,n}$$

(для множественной регрессии), а по оси ординат либо отклонения e_i, либо их квадраты $e_i^2, \quad i = \overline{1,n}$.

Если все отклонения e_i^2 находятся внутри горизонтальной полосы постоянной ширины, это говорит о независимости дисперсий e_i^2 от значений объясняющей переменной и выполнимости условия гомоскедастичности.

В других случаях наблюдаются систематические изменения в соотношениях между значениями $\hat{y}_i$ и квадратами отклонений e_i^2. Такие ситуации отражают большую вероятность наличия гетероскедастичности для рассматриваемых статистических данных. В настоящее время для определения гетероскедастичности разработаны специальные тесты и критерии для них.

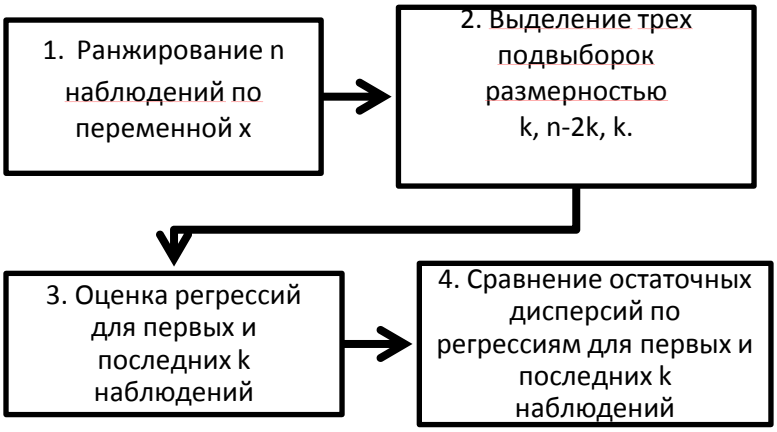


Рис. 8.1. Тест Голдфелда-Квандта

F-статистика для сравнения дисперсий:

$$S^2{}_1 = \sum_{i=1}^{k} e_i{}^2; S^2{}_3 = \sum_{i=n-k+1}^{n} e_i{}^2,$$

$$H_0 : S^2{}_3 = S^2{}_1 (\textit{гомоскедастичность})$$

$$H_1 : S^2{}_3 > S^2{}_1 (\textit{гетероскедастичность})$$

$$F = \frac{S^2{}_3 / (k-m-1)}{S^2{}_1 / (k-m-1)},$$

$$F > F_{\alpha,m,k-m-1} \Rightarrow H_1$$

Тест ранговой корреляции Спирмена. При использовании данного теста предполагается, что дисперсия отклонений будет либо увеличиваться, либо уменьшаться с увеличением значений *x*. Поэтому для регрессии, построенной по МНК, абсолютные величины отклонений $|e_i|$ и значения x_i будут коррелированы. t- статистика для проверки значимости $r_{x,e}$:

$$r_{x,e} = 1 - 6 \cdot (\sum d_i^2 / n(n^2 - 1))$$
$$H_0 : r_{x,e} = 0 (\textit{гомоскедастичность})$$
$$H_1 : r_{x,e} \neq 0 (\textit{гетероскедастичность})$$
$$t = \frac{r_{x,e} \cdot \sqrt{n-2}}{\sqrt{1 - r^2_{x,e}}},$$
$$t > t_{\alpha, n-2} \Rightarrow H_1$$

Коррекция на гетероскедастичность. Для устранения гетероскедастичности в случае, если дисперсии отклонений известны для каждого наблюдения, применяется метод взвешенных наименьших квадратов (ВНК). Гетероскедастичность устраняется, если разделить каждое наблюдаемое значение на соответствующее ему значение дисперсии:

$$y = \alpha + \beta \cdot x + \varepsilon$$
$$\frac{y}{\sigma} = \alpha \cdot \frac{1}{\sigma} + \beta \cdot \frac{x}{\sigma} + \frac{\varepsilon}{\sigma}$$
$$y^* = \alpha \cdot z + \beta \cdot x^* + v$$

Если дисперсии отклонений неизвестны для каждого наблюдения, то предполагается, что дисперсии σ^2_e пропорциональны x_i

$$\sigma^2{}_i = \sigma^2 x_i$$
$$\frac{y_i}{\sqrt{x_i}} = \alpha \cdot \frac{1}{\sqrt{x_i}} + \beta \cdot \frac{x_i}{\sqrt{x_i}} + \frac{\varepsilon_i}{\sqrt{x_i}}$$
$$\frac{y_i}{\sqrt{x_i}} = \alpha \cdot \frac{1}{\sqrt{x_i}} + \beta \cdot \sqrt{x_i} + v_i$$
$$y^* = \alpha \cdot z + \beta \cdot x^* + v_i$$

Дисперсии σ^2_e пропорциональны x^2_i

$$\sigma^2{}_i = \sigma^2 \cdot x^2{}_i$$
$$\frac{y_i}{x_i} = \alpha \cdot \frac{1}{x_i} + \beta \cdot \frac{x_i}{x_i} + \frac{\varepsilon_i}{x_i}$$
$$\frac{y_i}{x_i} = \alpha \cdot \frac{1}{x_i} + \beta + v_i$$
$$y^* = \alpha \cdot z + \beta + v_i$$

Таким образом, наблюдения с наименьшими дисперсиями получают наибольшие «веса», а наблюдения с наибольшими дисперсиями – наименьшие «веса». Поэтому наблюдения с меньшими дисперсиями отклонений будут более значимыми при оценке параметров регрессии, чем наблюдения с большими

дисперсиями. При этом повышается вероятность получения более точных оценок. В этом заключается смысл метода взвешенных наименьших квадратов. Полученные по методу взвешенных наименьших квадратов оценки параметров модели можно использовать в первоначальной модели.

Для применения метода взвешенных наименьших квадратов необходимо знать фактические значения дисперсий отклонений σ_i^2. На практике такие значения известны крайне редко. Поэтому, чтобы применить ВНК, необходимо сделать реалистические предположения о значениях σ_i^2. Чаще всего предполагается, что дисперсии отклонений пропорциональны или значениям x_i, или значениям x_i^2. Если в уравнении регрессии присутствует несколько объясняющих переменных, вместо конкретной переменной x_j используется исходное уравнение множественной регрессии

$$\hat{y} = a + b_1 x_1 + b_2 x_2 + ... + b_p x_p \ ,$$

т.е. фактически линейная комбинация факторов. В этом случае получают следующую регрессию: $\frac{y_i}{\sqrt{\hat{y}_i}} = a\frac{1}{\sqrt{\hat{y}_i}} + b_1\frac{x_{i1}}{\sqrt{\hat{y}_i}} + ... + b_p\frac{x_{ip}}{\sqrt{\hat{y}_i}} + \frac{\varepsilon_i}{\sqrt{\hat{y}_i}}$.

Вопросы и задания для самоконтроля

1. Действительно ли, вследствие гетероскедастичности оценки перестают быть эффективными и состоятельными?
2. Какие критерии могут быть использованы для проверки гипотезы о гомоскедастичности регрессионных остатков?
3. В чем заключается тест Спирмена?
4. Какова схема теста Голдфелда-Квандта?
5. В чем суть метода взвешенных наименьших квадратов?
6. Какие типы преобразований применяются для устранения гетероскедастичности?

Задание 1. Заданы следующие значения остатков линейной модели:

Ранг x_i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18

e_i	-1	2	-3	2	0	-3	3	1	-2	-4	5	-11	8	-20	12	-21	18	14

Установить, имеется ли гетероскедастичность по тесту ранговой корреляции Спирмена на уровне значимости $\alpha = 0{,}05$.

Задание 2. Для линейной модели переменной y относительно переменной x получены следующие остатки, соотнесенные последовательным наблюдениям переменной x_i.

x_i	1,3	0,9	0,8	0,7	1,1	1,0	1,5	1,0	0,8	1,4	1,2	1,1
e_i	-5	1	2	-6	4	-4	1	4	5	-6	-1	6
x_i	1,5	1,8	1,2	0,8	1,3	1,1	1,2	1,0	0,9	1,3	1,2	1,0
e_i	-4	9	-5	-2	8	-5	6	-4	5	7	-8	5

На уровне значимости $\alpha = 0{,}05$ с помощью F – теста проверить гипотезу о равенстве дисперсий случайных ошибок.

Лекция 10

Тема 9. Автокорреляция

Вопросы для изучения

1. Понятие и последствия автокорреляции.
2. Обнаружение автокорреляции.
3. Коррекция на автокорреляцию.

Аннотация. Данная тема раскрывает способы проверки соблюдения третьей предпосылки МНК в остатках регрессии.

Ключевые слова. Автокорреляция, остатки регрессии, авторегрессионное преобразование.

Методические рекомендации по изучению темы

•Изучить лекционную часть, где даются общие представления по данной теме.

• Для закрепления теоретического материала ознакомиться с решениями типовых задач и ответить на вопросы для самоконтроля.

• Для проверки усвоения темы выполнить практические задания и тест для самоконтроля.

Рекомендуемые информационные ресурсы:

1. http://tulpar.kfu.ru/course/view.php?id=2213.

2. Эконометрика: [Электронный ресурс] Учеб.пособие / А.И. Новиков. - 2-е изд., испр. и доп. - М.: ИНФРА-М, 2011. - 144 с.: с. (http://znanium.com/catalog.php?item=booksearch&code=%D1%8D%D0%BA%D0%BE%D0%BD%D0%BE%D0%BC%D0%B5%D1%82%D1%80%D0%B8%D0%BA%D0%B0&page=1#none) С. 92-106.

3. Валентинов, В. А. Эконометрика [Электронный ресурс]: Практикум / В. А. Валентинов. - 3-е изд. - М.: Дашков и К, 2010. - 436 с.

(http://znanium.com/catalog.php?item=booksearch&code=%D1%8D%D0%BA%D0%BE%D0%BD%D0%BE%D0%BC%D0%B5%D1%82%D1%80%D0%B8%D0%BA%D0%B0&page=3#none) С. 202-229.

5.Эконометрика. Практикум: [Электронный ресурс] Учебное пособие / С.А. Бородич. - М.: НИЦ ИНФРА-М; Мн.: Нов.знание, 2014. - 329 с. (http://znanium.com/catalog.php?item=booksearch&code=%D1%8D%D0%BA%D0%BE%D0%BD%D0%BE%D0%BC%D0%B5%D1%82%D1%80%D0%B8%D0%BA%D0%B0&page=4#none) С. 197-244.

Понятие и последствия автокорреляции. Автокорреляцией остатков называется нарушение третьей предпосылки МНК о независимости случайного отклонения ε i от отклонений во всех других наблюдениях. Если предпосылка МНК о том, что cov(εi,εj)=0, соблюдена, то автокорреляция случайных отклонений отсутствует. Чаще всего положительная автокорреляция вызывается направленным постоянным воздействием некоторых не учтенных в регрессии факторов. Например, при исследовании спроса *y* на прохладительные напитки в зависимости от дохода *x* на трендовую зависимость накладываются изменения спроса в летние и зимние периоды. Аналогичная картина может иметь место в макроэкономическом анализе с учетом циклов деловой активности. Автокорреляция остатков обычно встречается при использовании данных временных ря-

дов. В перекрестных данных наличие автокорреляции бывает редко. Положительная автокорреляция имеет место, когда rεi,εj>0. Отрицательная автокорреляция имеет место, когда rεi,εj<0. Отрицательная автокорреляция фактически означает, что за положительным отклонением следует отрицательное и наоборот. Такая ситуация может иметь место, если ту же зависимость между спросом на прохладительные напитки и доходами рассматривать не ежемесячно, а раз в сезон (зима–лето).

Последствия автокорреляции: МНК-оценки сохраняют свойства несмещенности и линейности, но теряют свойство эффективности; дисперсии МНК-оценок смещены в сторону занижения; t-статистика и F-статистика завышены.

Обнаружение автокорреляции. Методы обнаружения автокорреляции: графический анализ остатков; критерий Дарбина-Уотсона; метод рядов.

Метод рядов. По этому методу последовательно определяются знаки отклонений $e_t, t = \overline{1,n}$ от регрессионной зависимости. Например, имеем при 20 наблюдениях

(-----)(++++++++)(---)(++++)(-) .

Ряд определяется как непрерывная последовательность одинаковых знаков. Количество знаков в ряду называется длиной ряда. Если рядов слишком мало по сравнению с количеством наблюдений n, то вполне вероятна положительная автокорреляция. Если же рядов слишком много, то вероятна отрицательная автокорреляция.

Пусть n – объѐм выборки, n_1 – общее количество положительных отклонений; n_2 – общее количество отрицательных отклонений; k – количество рядов. В приведенном примере n=20, n_1=11, n_2=5.

При достаточно большом количестве наблюдений (n_1>10, n_2>10) и отсутствии автокорреляции СВ k имеет асимптотически нормальное распределение, в котором

$$M(k)=\frac{2n_1n_2}{n_1+n_2}+1;$$

$$D(k)=\frac{2n_1n_2(2n_1n_2-n_1-n_2)}{(n_1+n_2)^2(n_1+n_2-1)}$$

Тогда, если

$$M(k)-u_{\alpha/2}\cdot\sqrt{D(k)}<k<M(k)+u_{\alpha/2}\cdot\sqrt{D(k)},$$

то гипотеза об отсутствии автокорреляции не отклоняется. Если $k\le M(k)-u_{\alpha/2}\cdot\sqrt{D(k)}$, то констатируется положительная автокорреляция; в случае $k\ge M(k)+u_{\alpha/2}\cdot\sqrt{D(k)}$ признается наличие отрицательной автокорреляции.

Для небольшого числа наблюдений (n_1<20, n_2<20) были разработаны таблицы критических значений количества рядов при n наблюдениях. В одной таблице в зависимости от n_1 и n_2 определяется нижняя граница k_1 количества рядов, в другой – верхняя граница k_2. Если $k_1<k<k_2$, то говорят об отсутствии автокорреляции. Если $k\le k_1$, то говорят о положительной автокорреляции. Если $k\ge k_2$,то говорят об отрицательной автокорреляции. Например, для приведенных выше данных k_1=6, k_2=16 при уровне значимости 0,05. Поскольку $k=5<k_1=6$, определяем положительную автокорреляцию.

Критерий Дарбина-Уотсона:

$$DW=\frac{\sum_{n=2}^{N}(e_i-e_{i-1})^2}{\sum_{n=1}^{N}e_i^{\,2}}$$

$$DW\approx 2\cdot(1-r_{e_i,e_{i-1}});0\le DW\le 4$$

$$r_{e_i,e_{i-1}}\approx 0\Rightarrow DW\approx 2$$

$$r_{e_i,e_{i-1}}\approx 1\Rightarrow DW\approx 0(\text{"+"}\textit{автокорреляция})$$

$$r_{e_i,e_{i-1}}\approx -1\Rightarrow DW\approx 4(\text{"–"}\textit{автокорреляция})$$

d_1, d_2 – критические точки распределения Дарбина-Уотсона (табличные значения)

Рис. 9.1. Проверка гипотезы об автокорреляции остатков по DW-критерию

Можно показать, что статистика *DW* тесно связана с коэффициентом автокорреляции первого порядка: $r_{e_{t-1}e_t} = \dfrac{\sum_{t=2}^{n} e_{t-1}e_t}{\sqrt{\sum_{t=1}^{n-1} e_t^2 \sum_{t=2}^{n} e_{t-1}^2}}$.

Связь выражается формулой: $DW \approx 2\left(1 - r_{e_{t-1}e_t}\right)$.

Отсюда вытекает смысл статистического анализа автокорреляции. Поскольку значения *r* изменяются от –1 до +1, *DW* изменяется от 0 до 4. Когда автокорреляция отсутствует, коэффициент автокорреляции равен нулю, и статистика *DW* равна 2. *DW*=0 соответствует положительной автокорреляции, когда выражение в скобках равно нулю $\left(r = +1\right)$. При отрицательной автокорреляции $\left(r = -1\right)$. *DW*=4, и выражение в скобках равно двум.

Ограничения критерия Дарбина-Уотсона:

1. Критерий *DW* применяется лишь для тех моделей, которые содержат свободный член.

2. Предполагается, что случайные отклонения определяются по итерационной схеме $e_t = \rho e_{t-1} + v_t$, называемой авторегрессионной схемой первого порядка *AR*(1). Здесь v_t – случайный член.

3. Статистические данные должны иметь одинаковую периодичность (не должно быть пропусков в наблюдениях).

4. Критерий Дарбина – Уотсона не применим к авторегрессионным моделям вида: $y_t = a + b_1 x_{t1} + ... + b_p x_{tp} + c y_{t-1} + e_t$, которые содержат в числе факторов также зависимую переменную с временным лагом (запаздыванием) в один период.

Коррекция на автокорреляцию. Автокорреляция чаще всего вызывается неправильной спецификацией модели. Поэтому следует попытаться скорректировать саму модель, в частности, ввести какой – нибудь неучтенный фактор или изменить форму модели (например, с линейной на полулогарифмическую или гиперболическую). Если все эти способы не помогают и автокорреляция вызвана какими – то внутренними свойствами ряда {et}, можно воспользоваться преобразованием, которое называется авторегрессионной схемой первого порядка AR(1).

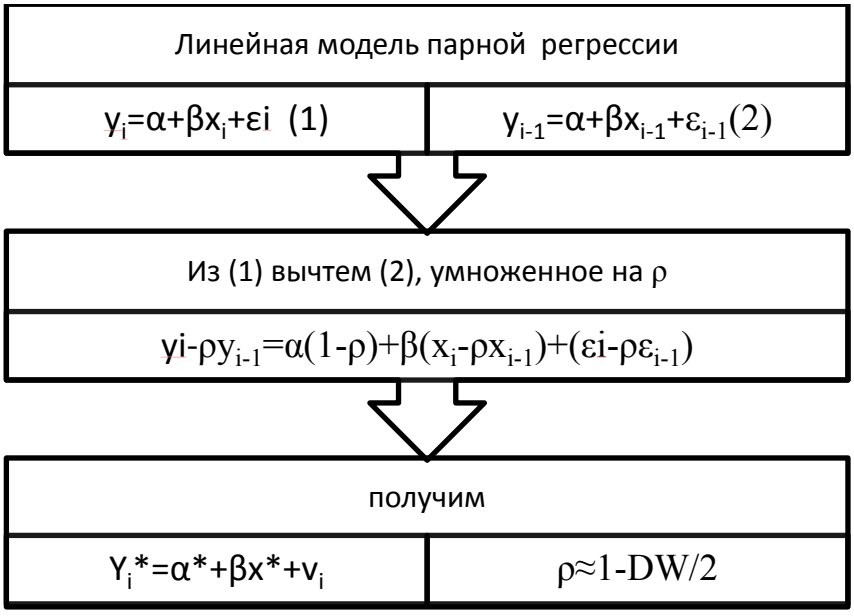


Рис.9.2. Авторегрессионное преобразование

Поскольку случайные отклонения v_t удовлетворяют предпосылкам МНК, оценки a^* и b будут обладать свойствами наилучших линейных несмещенных оценок. По преобразованным значениям всех переменных с помощью обычного

МНК вычисляются оценки параметров a^* и b, которые затем можно использовать в регрессии (71).

Однако способ вычисления преобразованных переменных (75) приводит к потере первого наблюдения, если нет информации о предшествующих наблюдениях. Это уменьшает на единицу число степеней свободы, что при больших выборках не очень существенно, однако при малых выборках приводит к потере эффективности. Тогда первое наблюдение восстанавливается с помощью поправки Прайса – Уинстена:

$$x_1^* = \sqrt{1-\rho^2} \cdot x_1,$$
$$y_1^* = \sqrt{1-\rho^2} \cdot y_1$$

Авторегрессионное преобразование может быть обобщено на произвольное число объясняющих переменных, т.е. использовано для уравнения множественной регрессии.

Вопросы и задания для самоконтроля

1. Каковы основные причины и последствия автокорреляции?
2. Что такое автокорреляционная функция?
3. Какова основная идея метода рядов при обнаружении автокорреляции?
4. Как проводится тест Дарбина-Уотсона?
5. В чем состоит авторегрессионная схема 1-го порядка?

Задание 1. По статистическим данным за 20 лет построено уравнение регрессии между ценой бензина и объемом продаж бензина, $d = DW = 0{,}71$. Ответить на вопросы: будет ли иметь место автокорреляция остатков? Что могло послужить причиной автокорреляции?

Задание 2. Для модели $\tilde{y} = 32 + 0{,}35x_1 - 0{,}46x_2$, параметры которой оценены по МНК, получена следующая последовательность остатков:

Номер i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
e_i	-2	3	-1	2	-4	2	0	1	-1	0	-4	3	-2	3	0

Рассчитать коэффициент автокорреляции первого порядка. При уровне значимости $\alpha = 0{,}05$ исследовать с помощью теста Дарбина-Уотсона наличие автокорреляции между ошибками ε_i и ε_{i-1}.

Лекция 11

Тема 10. Фиктивные переменные

Вопросы для изучения

1. Регрессионные модели с переменной структурой (фиктивные переменные).

2. Правило использования фиктивных переменных.

3. ANOVA–модели и ANCOVA–модели. Тест Чоу на наличие структурной перестройки.

Аннотация. Данная тема раскрывает особенности построения регрессионных моделей с переменной структурой.

Ключевые слова. Фиктивные переменные, Anova – модели, Ancova – модели, тест Чоу.

Методические рекомендации по изучению темы

•Изучить лекционную часть, где даются общие представления по данной теме.

• Для закрепления теоретического материала ознакомиться с решениями типовых задач и ответить на вопросы для самоконтроля.

• Для проверки усвоения темы выполнить практические задания и тест для самоконтроля.

Рекомендуемые информационные ресурсы:

1. http://tulpar.kfu.ru/course/view.php?id=2213.

2. Эконометрика: [Электронный ресурс] Учеб.пособие / А.И. Новиков. - 3-е изд., испр. и доп. - М.: ИНФРА-М, 2014. - 272 с.: (http://znanium.com/catalog.php?item=booksearch&code=%D1%8D%D0%BA%D0

%BE%D0%BD%D0%BE%D0%BC%D0%B5%D1%82%D1%80%D0%B8%D0%BA%D0%B0&page=1#none) С. 63-65.

3. Валентинов, В. А. Эконометрика [Электронный ресурс]: Практикум / В. А. Валентинов. - 3-е изд. - М.: Дашков и К, 2010. - 436 с.

(http://znanium.com/catalog.php?item=booksearch&code=%D1%8D%D0%BA%D0%BE%D0%BD%D0%BE%D0%BC%D0%B5%D1%82%D1%80%D0%B8%D0%BA%D0%B0&page=3#none) С. 229-242.

4. Эконометрика. Практикум: [Электронный ресурс] Учебное пособие / С.А. Бородич. - М.: НИЦ ИНФРА-М; Мн.: Нов.знание, 2014. - 329 с. (http://znanium.com/catalog.php?item=booksearch&code=%D1%8D%D0%BA%D0%BE%D0%BD%D0%BE%D0%BC%D0%B5%D1%82%D1%80%D0%B8%D0%BA%D0%B0&page=4#none) С. 253-273.

Регрессионные модели с переменной структурой (фиктивные переменные). Исходные статистические данные называют однородными, если все они зарегистрированы при одних и тех же условиях (время года, регион, образование, пол человека). Если же данные объединяют в себе наблюдения, зарегистрированные при различных условиях, то они могут быть неоднородными. В этом случае в модель включается фактор, имеющий два или более качественных уровней.

Влияние качественных факторов иногда приводит к изменению структуры линейных связей в модели (то есть значений коэффициентов a и b_i). Построение регрессионной модели по неоднородным данным проводится по одной из двух схем:

- по каждой регрессионно однородной подвыборке;
- по объединенной регрессионно неоднородной выборке путем введения в модель фиктивных переменных (полезно в условиях дефицита исходных данных).

Фиктивные (dummy variables, искусственные, двоичные, структурные) переменные отражают в модели влияние качественного фактора, содержащего атрибутивные признаки двух и более уровней.

Для того, чтобы ввести такие переменные в регрессионную модель, им должны быть присвоены те или иные цифровые метки, то есть качественные переменные необходимо преобразовать в количественные.

Правило использования фиктивных переменных. В случае, когда качественная переменная принимает не два, а большее число значений, может возникнуть ситуация, которая называется ловушкой фиктивной переменной. Она возникает, когда для моделирования k значений качественного признака используется ровно k бинарных (фиктивных) переменных. В этом случае одна из таких переменных линейно выражается через все остальные, и матрица $(X'X)$ становится вырожденной. Тогда исследователь попадает в ситуацию совершенной мультиколлинеарности. Избежать подобной ловушки позволяет правило: если качественная переменная имеет k альтернативных значений, то при моделировании используется только (k-1) фиктивных переменных.

Например, если качественная переменная имеет 3 уровня, то для моделирования достаточно двух фиктивных переменных D_1 и D_2. Тогда для обозначения третьего уровня достаточно принять, например, обе переменные равными нулю: $D_1=D_2=0$. В частности, для обозначения уровня экономического развития страны (развитая, развивающаяся или страна «третьего мира») можно использовать обозначения:

$$D_1 = \begin{cases} 0, & \textit{страна не является развитой} \\ 1, & \textit{страна развитая} \end{cases}$$

$$D_2 = \begin{cases} 0, & \textit{страна не является развивающейся} \\ 1, & \textit{страна развивающаяся} \end{cases}$$

Тогда $D_1=D_2=0$ означает страну «третьего мира». Нулевой уровень качественной переменной называется базовым или сравнительным.

Кроме того, значения фиктивных переменных можно изменять на противоположные. Суть модели от этого не изменится. Изменится только знак коэффициента g в модели (80).

Коэффициент g в модели (80) называется дифференциальным свободным членом, т.к. он показывает, на какую величину изменится свободный член модели при изменении значения фиктивной переменной.

Возможны модели, в которых используются несколько фиктивных переменных, не связанных между собой по смыслу. Например, переменная D_1 означает пол работника, а D_2 – наличие или отсутствие у него высшего образования. Тогда возможны все комбинации значений различных качественных переменных, в которых регрессии отличаются лишь свободными членами.

Подобные схемы можно распространить на произвольное число количественных или качественных факторов. При этом не следует забывать, что если качественный фактор имеет k альтернативных состояний, то для его описания можно использовать только k различных сочетаний значений (k-1) фиктивных переменных. Например, если качественная переменная имеет 4 уровня, то для еѐ описания следует использовать 3 фиктивные (бинарные) переменные. Такой случай на практике применяется при моделировании сезонности по кварталам, где 3 переменные будут индикаторами первых трех кварталов, а четвертый квартал обозначается базовым уровнем всех трех переменных. Максимально возможное число сочетаний их значений равно восьми (два в третьей степени), однако в регрессии можно реально использовать только четыре из них (поскольку нельзя одновременно использовать больше одной единицы – это будет означать, к примеру, первый и второй кварталы сразу) .

Влияние качественного фактора может сказываться не только на значении свободного члена, но и на угловом коэффициенте линейной регрессионной модели. Обычно это характерно для временных рядов экономических данных при изменении институциональных условий, введении новых правовых или налоговых ограничений. Тогда зависимость может быть выражена так:

$$y = a + bx + g_1 D + g_2 Dx + e, \qquad (81)$$

где

$$D = \begin{cases} 0, & \textit{до изменения условий}, \\ 1, & \textit{после изменения условий}. \end{cases}$$

В этой ситуации ожидаемое значение зависимой переменной определяется следующим образом:

$$\hat{y} = a + bx, \qquad D = 0$$
$$\hat{y} = (a + g_1) + (b + g_2)x \qquad , D = 1$$

Коэффициенты g_1 и g_2 называются соответственно дифференциальным свободным членом и дифференциальным угловым коэффициентом. Последний по своему смыслу показывает, на какую величину изменится угловой коэффициент при изменении фиктивной переменной. Здесь фиктивная переменная разбивает зависимость на две части – до и после внесения изменений в условия еѐ действия.

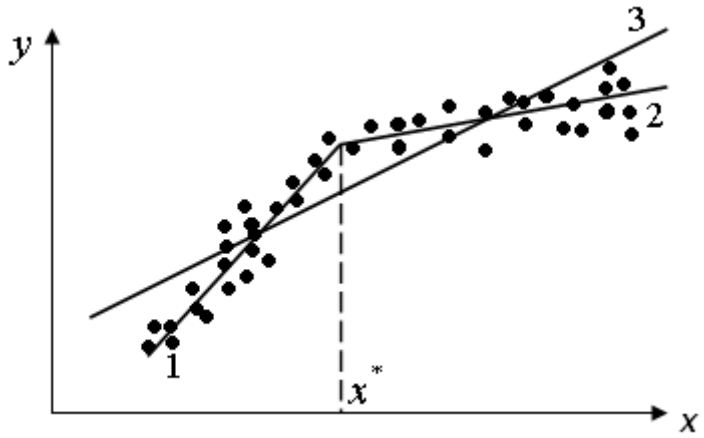


Рис. 10.1 Кусочно-линейная функция

Общая зависимость имеет вид кусочно – линейной функции, а изменения условий отображаются изменением угла наклона прямой к оси абсцисс (линии 1 – 2). Здесь исследователь должен принять решение, стоит ли разбивать выборку на части и строить для каждой из них уравнение регрессии (прямые 1 и

2) или ограничиться одной общей линией регрессии (линия 3). Для этого используют тест Чоу.

ANOVA–модели и ANCOVA–модели. Тест Чоу на наличие структурной перестройки. Регрессионные модели, содержащие лишь качественные объясняющие переменные, называются ANOVA-моделями (моделями дисперсионного анализа). Например, зависимость начальной заработной платы от образования может быть записана так:

$$y = a + gD + e,$$

где D=0, если претендент на рабочее место не имеет высшего образования, D=1, если имеет. Тогда при отсутствии высшего образования начальная заработная плата равна:

$$\hat{y} = a + g \cdot 0 = a,$$

а при его наличии:

$$\hat{y} = a + g \cdot 1 = a + g.$$

При этом параметр a определяет среднюю начальную заработную плату при отсутствии высшего образования. Коэффициент g показывает, на какую величину отличаются средние начальные заработные платы при наличии и при отсутствии высшего образования у претендента. Проверяя статистическую значимость коэффициента g с помощью t – статистики, можно определить, влияет или нет наличие высшего образования на начальную заработную плату.

Нетрудно заметить, что ANOVA – модели представляют собой кусочно – постоянные функции. Такие модели в экономике крайне редки. Гораздо чаще встречаются модели, содержащие как количественные, так и качественные переменные. Регрессионные модели, в которых объясняющие переменные носят как количественный, так и качественный характер, называются ANCOVA-моделями (моделями ковариационного анализа).

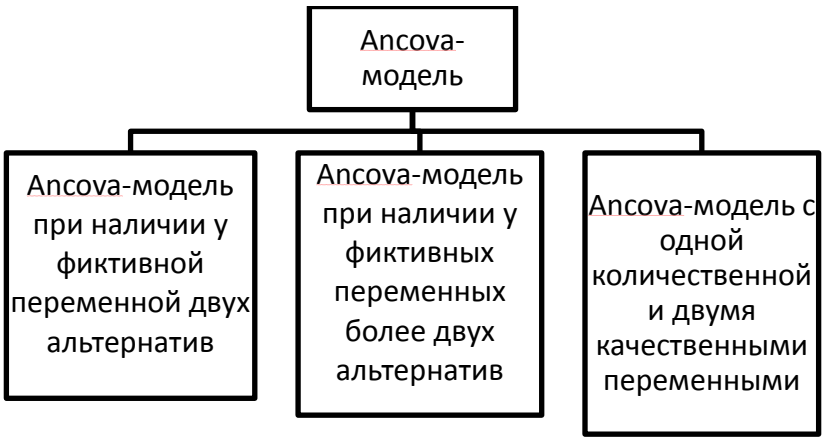


Рис.10.2. Виды Ancova-моделей

Ancova-модель при наличии у фиктивной переменной двух альтернатив:

y=a+b*x+γ*D+ε, D=1- лица мужского пола, D=0 – лица женского пола. Ожидаемое потребление кофе при цене *x* будет:

y=a+b*x+ε для женщины;

y=a+b*x+γ*D+ε=(a+ γ)+b*x +ε – для мужчины.

Если γ будет статистически значим по t-статистике, то пол влияет на потребление кофе. При $\gamma>0$ - в пользу мужчин, при $\gamma<0$ – в пользу женщин.

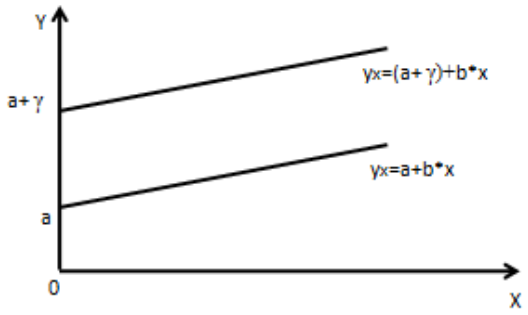


Рис. 10.3. Ancova-модель для фиктивной переменной с двумя альтернативами

Ancova-модель при наличии у фиктивной переменной более двух альтернатив:

$y=a+b*x+\gamma_1*D_1+\gamma_2*D_2+\varepsilon$,

Y- расходы на содержание ребенка, X- доходы домохозяйств, D_1=0- дошкольник, D_1=1- в противоположном случае, D_2=0 – дошкольник или младший школьник, D_2=1 – в противоположном случае.

Ожидаемые средние расходы при доходах ***x*** будут:

y=a+b*x - на дошкольника;

y=(a+ γ_1)+b*x –на младшего школьника;

y=(a+ γ_1+ γ_2)+b*x – на старшего школьника.

Если γ_1, γ_2 – дифференциальные свободные члены, будут статистически значимы по t-статистике, то возраст ребенка влияет на расходы по его содержанию.

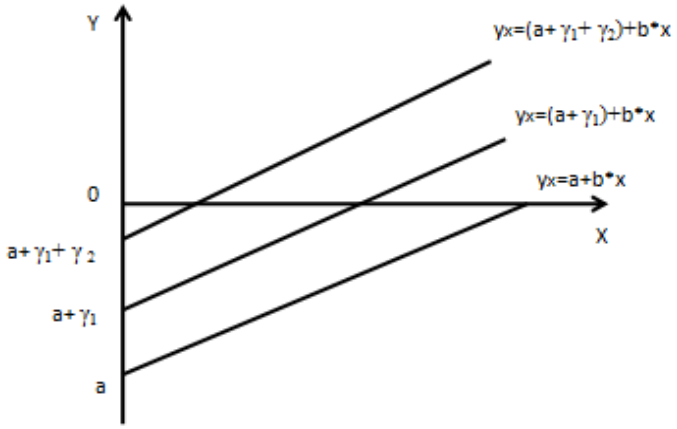


Рис. 10.4. Ancova-модель при наличии у фиктивной переменной более двух альтернатив

Регрессия с одной количественной и двумя качественными переменными:

y=a+b*x+γ_1*D_1+ γ_2*D_2+ε,

y - заработная плата сотрудников фирмы, x- стаж работы, D_1=0- женщина, D_1=1- мужчина, D_2=0 – нет высшего образования, D_2=1 – есть высшее образование.

Ожидаемая средняя заработная плата при стаже ***x*** будет:

y=a+b*x- для женщины без высшего образования;

y=(a+ γ_2)+b*x –для женщины с высшим образованием;

y=(a+ γ_1)+b*x – для мужчины без высшего образования;

y=(a+ γ_1+ γ_2)+b*x – для мужчины с высшим образованием.

Если γ_1, γ_2 – дифференциальные свободные члены, будут статистически значимы по t-статистике, то пол сотрудника и его образование влияют на среднюю заработную плату.

Тест Чоу: Вся выборка объѐма n разбивается на две подвыборки объѐмами n_1 и n_2 (n_1+n_2=n), и для каждой строится уравнение регрессии. Обозначим через s_1 и s_2 остаточные СКО для каждой из регрессий. Кроме того, строится общая регрессия для всех наблюдений (линия 3), и для неѐ определяется остаточная СКО, которую обозначим s_3. Равенство s_3=s_1+s_2 возможно лишь при совпадении коэффициентов регрессии для всех трѐх уравнений. Если сумма s_1+s_2 будет значительно меньше, чем s_3, то можно считать разбиение общей выборки на две подвыборки обоснованным. В этом смысле разность (s_3-(s_1+s_2)) можно считать мерой улучшения качества модели при разбиении выборки на две части. Однако при разбиении уменьшается число степеней свободы каждой из подвыборок. Эта альтернатива между числом степеней свободы и уменьшением остаточной СКО выражается через статистику

$$F = \frac{s_3 - (s_1 + s_2)}{s_1 + s_2} \cdot \frac{n - 2p - 2}{p + 1}\ , \qquad (82)$$

где p – число факторов. Выражение (82) равно отношению уменьшения необъясненной дисперсии к необъясненной дисперсии кусочно – линейной модели.

Если уменьшение дисперсии статистически незначимо, статистика (82) имеет распределение Фишера с (p+1, n-2p-2) степенями свободы. Если на заданном уровне значимости α $F_{набл} < F(\alpha; p+1; n-2p-2)$, то нет смысла разбивать уравнение регрессии на части. В противном случае разбиение на подвыборки целесообразно с точки зрения улучшения качества модели.

Если гипотеза о структурной стабильности выборки отклоняется, то исследуется вопрос о причинах структурных различий в подвыборках. Пусть данные в подвыборках описываются двумя уравнениями регрессии:

$$\hat{y} = a_1 + b_1 x,$$
$$\hat{y} = a_2 + b_2 x.$$

Тогда возможны следующие варианты:

1. Различие между a_1 и a_2 является статистически значимым, а коэффициенты b_1 и b_2 статистически не различаются. При этом наблюдается скачкообразное изменение зависимости при сохранении наклона линии регрессии:

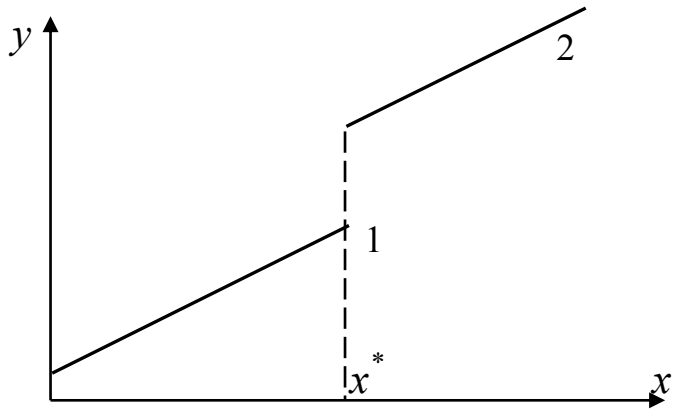


Рис. 10.5. Структурная нестабильность выборки, 1 вариант

2. Различие между b_1 и b_2 статистически значимо, а различие между a_1 и a_2 статистически не значимо:

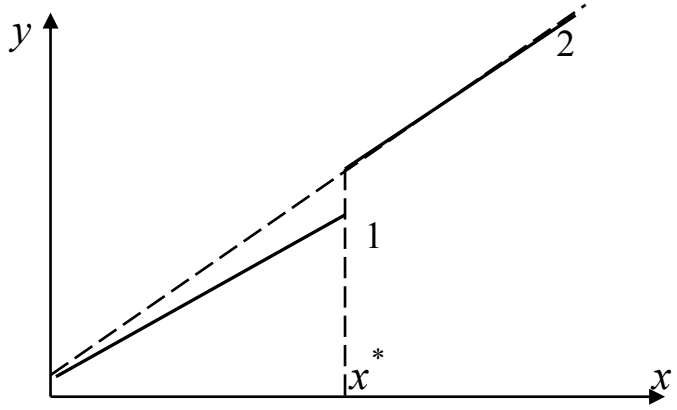


Рис. 10.6. Структурная нестабильность выборки, 2 вариант

3. Статистически значимыми являются и различия между a_1 и a_2, и различия между b_1 и b_2:

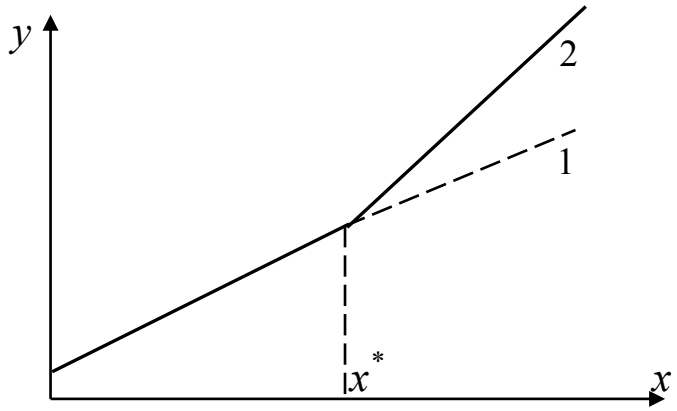


Рис. 10.7. Структурная нестабильность выборки, 3 вариант

Для тестирования всех этих ситуаций применяется следующая методика, предложенная Гуйарати. Она основана на включении в модель регрессии фиктивной переменной D, которая равна 1 для всех $x<x^*$ и равна 0 для всех $x>x^*$. Далее определяются параметры следующего уравнения регрессии:

$$y = a + bD + cx + dDx + e. \qquad (83)$$

Отсюда видно, что

$a_1=(a+b);\quad b_1=(c+d)\quad (D=1)$,

$a_2=a;\ b_2=c;\ (D=0)$.

Следовательно, параметр b есть разница между a_1 и a_2, параметр d – разница между b_1 и b_2. Если в уравнении (83) b является статистически значимым, а d – нет, то имеем первый вариант структурной перестройки. Если, наоборот, статистически значимым является d, а b – незначим, имеем второй вариант структурных изменений. Наконец, третий вариант имеем в случае, если оба коэффициента b и d являются статистически значимыми.

В заключение следует отметить, что преимущество метода Гуйарати перед тестом Чоу состоит в том, что нужно построить только одно, а не три уравнения регрессии.

Вопросы и задания для самоконтроля

1. В чем преимущества фиктивных переменных?
2. Как фиктивные переменные включаются в модель регрессии?
3. В чем суть ANOVA-моделей?
4. В чем суть ANCOVA-моделей?

5. В чем состоит правило применения фиктивных переменных?

6. Какой смысл имеет дифференциальный свободный член?

7. Какой смысл имеет дифференциальный угловой коэффициент?

8. Какова идея теста Чоу?

Задание 1. Исследуется зависимость заработной платы y от возраста рабочего x для мужчин и женщин. Оценивание объединенной регрессии ($n = 20$) и отдельных регрессий для рабочих-мужчин ($n_1 = 13$) и рабочих-женщин ($n_2 = 7$) дали следующие результаты:

Выборка	Оцененное уравнение	R^2	Сумма квадратов остатков
Объединенная	$\tilde{y} = 62{,}27 + 7{,}23x$	0,728	24888
Мужчины	$\tilde{y} = 55 + 7{,}39x$	0,735	18619
Женщины	$\tilde{y} = 59{,}43 + 7{,}3x$	0,712	5658

Проверить на уровне значимости $\alpha = 0{,}05$ с использованием критерия Чоу, улучшилось ли качество регрессии после разделения выборки на части.

Лекция 12

Тема 11. Нелинейные регрессии и их линеаризация

Вопросы для изучения

1. Классы и виды нелинейных регрессий.

2. Линеаризация нелинейных моделей. Выбор формы модели.

3. Индекс корреляции. Подбор линеаризующего преобразования (подход Бокса-Кокса).

Аннотация. Данная тема раскрывает особенности построения нелинейных моделей регрессии.

Ключевые слова. Нелинейная регрессия, индекс корреляции, коэффициент эластичности, подход Бокса-Кокса.

Методические рекомендации по изучению темы

•Изучить лекционную часть, где даются общие представления по данной теме.

• Для закрепления теоретического материала ответить на вопросы для самоконтроля.

• Для проверки усвоения темы выполнить практические задания и тест для самоконтроля.

Рекомендуемые информационные ресурсы:

1. http://tulpar.kfu.ru/course/view.php?id=2213.

2. Эконометрика: [Электронный ресурс] Учеб. пособие / А.И. Новиков. - 3-е изд., испр. и доп. - М.: ИНФРА-М, 2014. - 272 с.: (http://znanium.com/catalog.php?item=booksearch&code=%D1%8D%D0%BA%D0%BE%D0%BD%D0%BE%D0%BC%D0%B5%D1%82%D1%80%D0%B8%D0%BA%D0%B0&page=1#none) С. 41-45.

3.Уткин, В. Б. Эконометрика [Электронный ресурс] : Учебник / В. Б. Уткин; Под ред. проф. В. Б. Уткина. - 2-е изд. - М.: Издательско-торговая корпорация «Дашков и К°», 2012. - 564 с.

(http://znanium.com/catalog.php?item=booksearch&code=%D1%8D%D0%BA%D0%BE%D0%BD%D0%BE%D0%BC%D0%B5%D1%82%D1%80%D0%B8%D0%BA%D0%B0&page=4#none) С. 383-399.

4. Эконометрика. Практикум: [Электронный ресурс] Учебное пособие / С.А. Бородич. - М.: НИЦ ИНФРА-М; Мн.: Нов. знание, 2014. - 329 с. (http://znanium.com/catalog.php?item=booksearch&code=%D1%8D%D0%BA%D0%BE%D0%BD%D0%BE%D0%BC%D0%B5%D1%82%D1%80%D0%B8%D0%BA%D0%B0&page=4#none) С.172-174.

Классы и виды нелинейных регрессий. Если между экономическими явлениями существуют нелинейные соотношения, то они выражаются с помощью нелинейных функций. Различают два класса нелинейных регрессий:

- регрессии, нелинейные относительно включенных в анализ объясняющих переменных;
- регрессии, нелинейные по оцениваемым параметрам.

К первому классу моделей относятся полиномы разных степеней и равносторонняя гипербола. Ко второму классу относятся степенная и показательная

(экспоненциальная) функции. Например, полиномиальная модель произвольной степени: $y = a_0 + a_1 x + a_2 x^2 + \ldots + a_k x^k + \varepsilon$ подвергается МНК без всякой предварительной линеаризации.

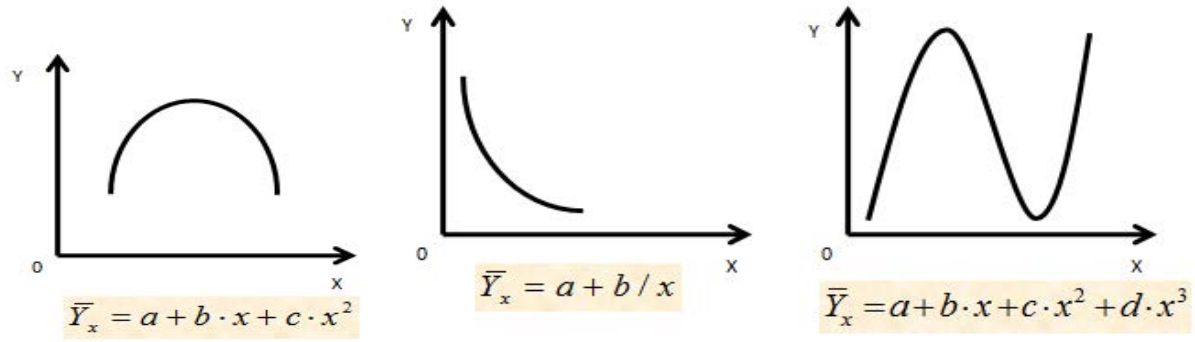


Рис. 11. 1. Регрессии, нелинейные относительно переменных

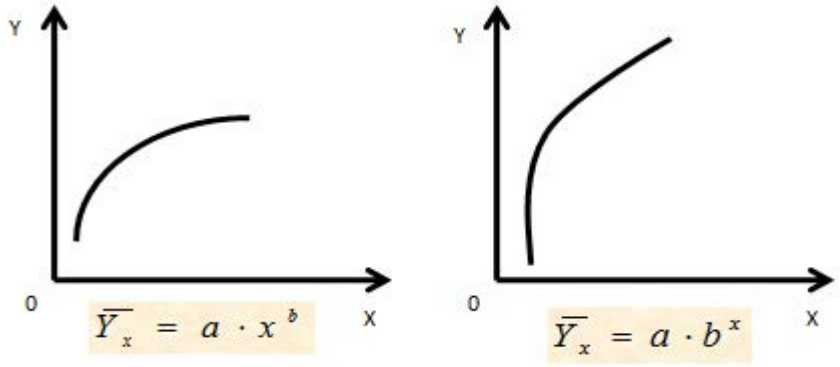


Рис.11.2. Регрессии, нелинейные по оцениваемым параметрам

Нелинейная модель, внутренне линейная, с помощью преобразований может быть приведена к линейному виду. Нелинейная модель, внутренне нелинейная, не может быть сведена к линейной функции.

При анализе нелинейных регрессионных зависимостей наиболее важным вопросом применения классического МНК является способ их линеаризации.

Линеаризация нелинейных моделей. Рассмотрим процедуру линеаризации применительно к параболе второй степени:

$$y = a + bx + cx^2 + \varepsilon .$$

Такая зависимость целесообразна в случае, если для некоторого интервала значений фактора возрастающая зависимость меняется на убывающую или

наоборот. В этом случае можно определить значение фактора, при котором достигается максимальное или минимальное значение результативного признака. Если исходные данные не обнаруживают изменение направленности связи, параметры параболы становятся трудно интерпретируемыми, и форму связи лучше заменить другими нелинейными моделями.

Применение МНК для оценки параметров параболы второй степени сводится к дифференцированию суммы квадратов остатков регрессии по каждому из оцениваемых параметров и приравниванию полученных выражений нулю. Получается система нормальных уравнений, число которых равно числу оцениваемых параметров, т.е. трем:

$$\begin{cases} a \cdot n & + b\sum x & + c\sum x^2 & = \sum y, \\ a\sum x & + b\sum x^2 & + c\sum x^3 & = \sum yx, \\ a\sum x^2 & + b\sum x^3 & + c\sum x^4 & = \sum yx^2. \end{cases}$$

Решать эту систему можно любым способом, в частности, методом определителей.

Экстремальное значение функции наблюдается при значении фактора, равном: $x = -\frac{b}{2a}$.

Если $b>0$, $c<0$, имеет место максимум, т.е. зависимость сначала растет, а затем падает. Такого рода зависимости наблюдаются в экономике труда при изучении заработной платы работников физического труда, когда в роли фактора выступает возраст. При $b<0$, $c>0$ парабола имеет минимум, что обычно проявляется в удельных затратах на производство в зависимости от объема выпускаемой продукции.

В нелинейных зависимостях, не являющихся классическими полиномами, обязательно проводится предварительная линеаризация, которая заключается в преобразовании или переменных, или параметров модели, или в комбинации этих преобразований. Рассмотрим некоторые классы таких зависимостей.

Рис. 11.3. Способы линеаризации

Замена переменных заключается в замене нелинейных объясняющих переменных новыми линейными переменными и сведении нелинейной регрессии к линейной. Например, полиномиальная модель:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \ldots + \beta_k x^k + \varepsilon$$
$$y = \beta_0 + \beta_1 z_1 + \beta_2 z_2 + \ldots + \beta_k z_k + \varepsilon$$

Например, гиперболическая модель:

$$y = \beta_0 + \frac{\beta_1}{x} + \varepsilon$$
$$y = \beta_0 + \beta_1 z + \varepsilon$$
$$z_1 = \frac{1}{x_1}; z_2 = \frac{1}{x_2}; \ldots z_n = \frac{1}{x_n}$$

Например, кривая Филлипса (равносторонняя гипербола), где

x- норма безработицы, y – процент прироста заработной платы:

$$y = \frac{1}{\beta_0 + \beta_1 x + \varepsilon}, \left(-\frac{\beta_0}{\beta_1} < x < \infty \right)$$
$$z = \beta_0 + \beta_1 x + \varepsilon$$
$$z_1 = \frac{1}{y_1}, z_2 = \frac{1}{y_2}, \ldots z_3 = \frac{1}{y_n}$$

Например, кривая Энгеля, где x- доход потребителей, y – спрос на определенный вид товаров или услуг

$$y = \frac{x}{\beta_0 x + \beta_1 + \varepsilon x}, \left(-\frac{\beta_0}{\beta_1} < x < \infty \right)$$
$$z = \beta_0 + \beta_1 z' + \varepsilon$$
$$z_1 = \frac{1}{y_1}, z_2 = \frac{1}{y_2}, \ldots z_3 = \frac{1}{y_n}$$
$$z'_1 = \frac{1}{x_1}, z'_2 = \frac{1}{x_2}, \ldots z'_3 = \frac{1}{x_n}$$

Например, полулогарифмические модели:

1) $\ln y = \beta_0 + \beta_1 x + \varepsilon$

Такие модели обычно используются в тех случаях, когда необходимо исследовать зависимость темпа роста или прироста экономических показателей:

- прирост объема выпуска от процентного увеличения затрат ресурсов;
- прирост бюджетного дефицита от темпа роста ВНП;
- темп роста инфляции от объема денежной массы.

2) $y = \beta_0 + \beta_1 \ln x + \varepsilon \qquad z = \ln y; z = \beta_0 + \beta_1 x + \varepsilon$

Используется обычно в тех случаях, когда необходимо исследовать, как процентное изменение независимой переменной влияет на абсолютное изменение зависимой переменной: влияние относительного (процентного) увеличения денежной массы на абсолютное изменение ВНП.

$z = \ln x; y = \beta_0 + \beta_1 z + \varepsilon$

Логарифмирование обеих частей уравнения применяется обычно, когда мультипликативную модель необходимо привести к линейному виду.

Например, степенные модели:

$y = \beta_0 x_1^{\beta_1} \cdot x_2^{\beta_2} \cdot \ldots \cdot x_p^{\beta_k} \cdot \varepsilon$

$y' = \ln y; x'_j = \ln x_j; \varepsilon' = \ln \varepsilon (j = 1,2,..k)$

$y' = \beta'_0 + \beta_1 x'_1 + \ldots + \beta_k x'_k + \varepsilon, \tilde{a}\ddot{a}\mathring{a}\beta'_0 = \ln \beta_0$

К классу степенных функций относятся: кривые спроса и предложения, производственная функция Кобба-Дугласа, кривые освоения для характеристики связи между трудоемкостью продукции и масштабами производства в период освоения и выпуска нового вида изделий, зависимость валового национального дохода от уровня занятости.

Например, показательные (экспоненциальные) модели. Широкий класс экономических показателей характеризуется приблизительно постоянным темпом относительного прироста во времени. Этому соответствует следующая форма зависимости показателя Y от времени X:

$$y = \beta_0 e^{\beta_1 x+\varepsilon}, (e = 2{,}7182818).$$
$$y' = \ln y$$
$$y' = \beta_0' + \beta_1 x + \varepsilon, \beta_0' = \ln \beta_0$$
$$b_0 = e^{\beta_0'}$$

$$y = \beta_0 e^{\beta_1 \frac{1}{x}+\varepsilon}, (e = 2{,}7182818).$$
$$y' = \ln y, x' = \frac{1}{x}$$
$$y' = \beta_0' + \beta_1 x' + \varepsilon, \beta_0' = \ln \beta_0$$
$$b_0 = e^{\beta_0'}$$

Например, логистическая кривая. Применяется для описания поведения показателей, имеющих определенные «уровни насыщения»: зависимость спроса на товар Y от дохода X, развитие производства новых товаров, рост численности населения (впервые применил А. Кетле (1796-1874)).

$$y = \frac{1}{\beta_0 + \beta_1 e^{-x} + \varepsilon}, (-\infty < x < \infty).$$
$$y' = \frac{1}{y}, x' = e^{-x}$$
$$y' = \beta_0 + \beta_1 x' + \varepsilon$$

Например, логлинейная модель. Используется в банковском и финансовом анализе. Y_0 – начальная величина переменной Y (первоначальная сумма вклада), r – сложный темп прироста величины Y (процентная ставка); Y_t – значение величины Y в момент времени t (вклад в банке в момент времени t).

$$\ln y_t = \ln y_0 + t \ln(1+r) + \ln \varepsilon$$
$$\ln y_0 = \beta_0, \ln(1+r) = \beta_1, \ln \varepsilon_t = \varepsilon_t'$$
$$\ln y_t = \beta_0 + \beta_1 t + \varepsilon_t'$$

Индекс корреляции. Подбор линеаризующего преобразования (подход Бокса-Кокса). Выбор модели не всегда осуществляется однозначно, и в дальнейшем требуется сравнивать модель как с теоретическими, так и с эмпирическими данными, совершенствовать ее. При определении качества модели обычно анализируются следующие параметры: скорректированный коэффициент детерминации; t – статистики; статистика Дарбина-Уотсона (DW); согласованность знаков коэффициентов с экономической теорией; прогнозные качества (ошибки) модели.

Любое уравнение нелинейной регрессии, как и линейной зависимости, дополняется показателем корреляции, который в данном случае называется индексом корреляции:

$$R = \sqrt{1 - \frac{\sigma_{ост}^2}{\sigma_y^2}}$$

Здесь σ_y^2 - общая дисперсия результативного признака у, $\sigma_{ост}^2$ - остаточная дисперсия, определяемая по уравнению нелинейной регрессии $\hat{y}_x = f(x)$. По-другому можно записать так:

$$R = \sqrt{1 - \frac{\sum (y - \hat{y}_x)^2}{\sum (y - \bar{y})^2}}$$

Следует обратить внимание на то, что разности в соответствующих суммах $\sum (y - \bar{y})^2$ и $\sum (y - \hat{y}_x)^2$ берутся не в преобразованных, а в исходных значениях результативного признака. Иначе говоря, при вычислении этих сумм следует использовать не преобразованные (линеаризованные) зависимости, а именно исходные нелинейные уравнения регрессии.

Величина R находится в границах $0 \le R \le 1$, и чем ближе она к единице, тем теснее связь рассматриваемых признаков, тем более надежно найденное уравнение регрессии. При этом индекс корреляции совпадает с линейным коэффициентом корреляции в случае, когда преобразование переменных с целью линеаризации уравнения регрессии не проводится с величинами результативного признака. Так обстоит дело с полулогарифмической и полиномиальной регрессиями, а также с равносторонней гиперболой (37). Определив линейный коэффициент корреляции для линеаризованных уравнений, например, в пакете Excel с помощью функции ЛИНЕЙН, можно использовать его и для нелинейной зависимости.

Иначе обстоит дело в случае, когда преобразование проводится также с величиной у, например, взятие обратной величины или логарифмирование. Тогда значение R, вычисленное той же функцией ЛИНЕЙН, будет относиться к линеаризованному уравнению регрессии, а не к исходному нелинейному уравнению, и величины разностей под суммами в (54) будут относиться к преобразо-

ванным величинам, а не к исходным, что не одно и то же. При этом, как было сказано выше, для расчета R следует воспользоваться выражением (54), вычисленным по исходному нелинейному уравнению.

Поскольку в расчете индекса корреляции используется соотношение остаточной и общей СКО, то R^2 имеет тот же смысл, что и коэффициент детерминации. В специальных исследованиях величину R^2 для нелинейных связей называют индексом детерминации.

Оценка существенности индекса корреляции проводится так же, как и оценка надежности коэффициента корреляции. Индекс детерминации используется для проверки адекватности уравнения нелинейной регрессии в целом по F-критерию Фишера:

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

где n – число наблюдений, m – число параметров при переменных x. Во всех рассмотренных нами случаях, кроме полиномиальной регрессии, m=1, для полиномов (34) m=k, т.е. степени полинома. Величина m характеризует число степеней свободы для факторной СКО, а (n-m-1) – число степеней свободы для остаточной СКО.

Индекс детерминации R^2 можно сравнивать с коэффициентом детерминации r^2 для обоснования возможности применения линейной функции. Чем больше кривизна линии регрессии, тем больше разница между R^2 и r^2. Близость этих показателей означает, что усложнять форму уравнения регрессии не следует и можно использовать линейную функцию. Практически, если величина (R^2-r^2) не превышает 0,1, то линейная зависимость считается оправданной. В противном случае проводится оценка существенности различия показателей детерминации, вычисленных по одним и тем же данным, через t-критерий Стьюдента:

$$t = \frac{R^2 - r^2}{m_{|R-r|}}.$$

Здесь в знаменателе находится ошибка разности (R^2-r^2), определяемая по формуле:

$$m_{|R-r|} = 2 \cdot \sqrt{\frac{\left(R^2 - r^2\right) - \left(R^2 - r^2\right)^2 \cdot \left(2 - \left(R^2 - r^2\right)\right)}{n}}$$

.

Если $t > t_{табл}(\alpha; n - m - 1)$, то различия между показателями корреляции существенны и замена нелинейной регрессии линейной нецелесообразна.

Если разные модели используют разные функциональные формы для зависимой переменной, то проблема выбора модели становится более сложной, так как нельзя непосредственно сравнивать коэффициенты R^2 или суммы квадратов отклонений. Например, нельзя сравнивать эти статистики для линейного и логарифмического вариантов. Пусть в линейной модели в качестве зависимой переменной используется заработок, а в нелинейной – логарифм заработка. Тогда R^2 в одном уравнении измеряет объясненную регрессией долю дисперсии заработка, а в другом - объясненную регрессией долю дисперсии логарифма заработка. В случае, если значения R^2 для двух моделей близки друг к другу, проблема выбора усложняется.

Здесь следует использовать *тест Бокса – Кокса.* При сравнении моделей с использованием в качестве зависимой переменной y и lny проводится такое преобразование масштаба наблюдений y, при котором можно непосредственно сравнивать СКО в линейной и логарифмической моделях. Здесь выполняются следующие шаги. Вычисляется среднее геометрическое значений y в выборке. Оно совпадает с экспонентой среднего арифметического логарифмов y. Все значения y пересчитываются делением на среднее геометрическое, получаем значения y*. Оцениваются две регрессии: для линейной модели с использованием y* в качестве зависимой переменной и для логарифмической модели с использованием ln y* вместо ln y. Во всех других отношениях модели должны оставаться неизменными. Теперь значения СКО для двух регрессий сравнимы, и модель с меньшей остаточной СКО обеспечивает лучшее соответствие исходным данным. Для проверки, обеспечивает ли одна из моделей значимо луч-

шее соответствие, можно вычислить величину (n/2)lnz, где z – отношение значений остаточной СКО в перечисленных регрессиях. Эта статистика имеет распределение хи – квадрат с одной степенью свободы. Если она превышает критическое значение при выбранном уровне значимости α, то делается вывод о наличии значимой разницы в качестве оценивания.

Величина коэффициента эластичности показывает, на сколько процентов изменится результативный признак Y, если факторный признак изменится на 1 %: $\acute{Y} = f'(x)\frac{x}{y}$

В заключение приведем формулы расчета коэффициентов эластичности для наиболее распространенных уравнений регрессии:

Вид уравнения регрессии	Коэффициент эластичности
$y = a + b \cdot x + \varepsilon$	$\frac{b \cdot x}{a + bx}$
$y = a + bx + cx^2 + \varepsilon$	$\frac{(b + 2cx) \cdot x}{a + bx + cx^2}$
$y = a + \frac{b}{x} + \varepsilon$	$\frac{-b}{ax + b}$
$y = a \cdot b^x \cdot \varepsilon$	$x \ln b$
$y = a \cdot x^b$	b
$y = a + b \ln x + \varepsilon$	$\frac{b}{a + b \ln x}$
$y = \frac{1}{a + bx + \varepsilon}$	$\frac{-bx}{a + bx}$
$y = \frac{a}{1 + b \cdot e^{-cx + \varepsilon}}$	$\frac{c \cdot x}{\frac{1}{b} \cdot e^{cx} + 1}$

Вопросы и задания для самоконтроля

1. Какие модели являются нелинейными относительно: а) включаемых переменных; б) оцениваемых параметров?

2. Какие преобразования используются для линеаризации нелинейных моделей?

3. Чем отличается применение МНК к моделям, нелинейным относительно включаемых переменных, от применения к моделям, нелинейным по оцениваемым параметрам?

4. Как определяются коэффициенты эластичности по разным видам регрессионных моделей?

5. Какие показатели корреляции используются при нелинейных соотношениях рассматриваемых признаков?

6. В каких случаях используют обратные и степенные модели?

Задание 1. По группе предприятий, производящих однородную продукцию известно, как зависит себестоимость единицы продукции (Y) от факторов, приведенных в таблице:

Признак-фактор	Уравнение парной регрессии	Среднее значение фактора
Объем производства, x_1, млн. руб.	$\tilde{y}_{x_1} = 0{,}62 + \frac{58{,}74}{x_1}$	$\bar{x}_1 = 2{,}64$
Трудоемкость единицы продукции, x_2, чел/час	$\tilde{y}_{x_2} = 9{,}30 + 9{,}83x_2$	$\bar{x}_2 = 1{,}38$
Оптовая цена за 1т энергоносителя, x_3, млн. руб.	$\tilde{y}_{x_3} = 11{,}45 + x_3^{1{,}6281}$	$\bar{x}_3 = 1{,}503$
Доля прибыли, изымаемая государством, x_4 ,%	$\tilde{y}_{x_4} = 14{,}87 \cdot 1{,}016^{x_4}$	$\bar{x}_4 = 26{,}3$

1) определить с помощью коэффициентов эластичности силу влияния каждого фактора на результат;

2) ранжировать факторы по силе влияния на результат.

Задание 2. По группе из 10 заводов, производящих однородную продукцию, получено уравнение регрессии себестоимости единицы продукции y (тыс. руб) от уровня технической оснащенности x (тыс. руб.)

$$\tilde{y} = 20 + \frac{700}{x}.$$

Доля остаточной дисперсии в общей составила 0,19.

1) определить коэффициент эластичности, предполагая, что стоимость активных производственных фондов составляет 200 тыс. руб.;

2) вычислить индекс корреляции;

3) оценить значимость уравнения регрессии с помощью F – критерия.

Лекция 13,14

Тема 12. Модели с дискретной зависимой переменной

Вопросы для изучения

1. Модели бинарного выбора.
2. Оценивание параметров моделей бинарного выбора.
3. Модели множественного выбора с упорядоченными альтернативами.

Аннотация. Данная тема раскрывает особенности построения моделей регрессии с дискретной зависимой переменной.

Ключевые слова. Логит-модель, пробит-модель, метод максимального правдоподобия, тест Вальда.

Методические рекомендации по изучению темы

•Изучить лекционную часть, где даются общие представления по данной теме.

• Для закрепления теоретического материала ответить на вопросы для самоконтроля.

• Для проверки усвоения темы выполнить практические задания и тест для самоконтроля.

Рекомендуемые информационные ресурсы:

1. http://tulpar.kfu.ru/course/view.php?id=2213.

2. Эконометрика: учеб. / под ред. В. С. Мхитаряна.- М.: Проспект, 2008. – 384 с. С. 216-256.

4. Эконометрика. Практикум: [Электронный ресурс] Учебное пособие / С.А. Бородич. - М.: НИЦ ИНФРА-М; Мн.: Нов.знание, 2014. - 329 с. (http://znanium.com/catalog.php?item=booksearch&code=%D1%8D%D0%BA%D0%BE%D0%BD%D0%BE%D0%BC%D0%B5%D1%82%D1%80%D0%B8%D0%BA%D0%B0&page=4#none) С.257-273.

Модели бинарного выбора. Зависимую переменную, которая принимает несколько значений, называют дискретной. Например, наличие собственного жилья: да – 1, нет – 2. В зависимости от числа альтернатив выделяют модели бинарного и множественного выбора. Чаще применяются модели бинарного выбора. Бинарная переменная принимает лишь два значения: 0 и 1. Например, 1 – занятый, 0 – безработный; 1 – есть мобильный телефон, 0 – нет мобильного телефона. Следовательно, вектор $Y=(y_1, y_2, \ldots, y_n)$ исходных данных будет содержать только дихотомические (бинарные) признаки 0 и 1.

$$P(y_i = 1) = F(x_i'\beta)$$
$$P(y_i = 0) = 1 - F(x_i'\beta)$$
$$0 \le F(\bullet) \le 1$$

Выбор функции $F(\bullet)$ определяет тип бинарной модели. Если используют функцию стандартного нормального распределения,

$$F(u) = \hat{O}(u) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{u} e^{-\frac{z2}{2}} dz$$

то модель бинарного выбора называют пробит-моделью (probit model). Если используют функцию логистического распределения,

$$F(u) = \Lambda(u) = \frac{e^u}{1+e^u}$$

то модель бинарного выбора называют логит-моделью (logit model).

$$P(y_i = 1) = p_i = \frac{1}{1+e^{-(\beta_0+\beta_1 x_i)}} = \frac{1}{1+e^{-z}}$$
$$P(y_i = 0) = 1 - p_i = \frac{1}{1+e^{z_i}}$$
$$\frac{p_i}{1-p_i} = \frac{1+e^{z_i}}{1+e^{-z}} = e^{z_i}$$
$$\ln \frac{p_i}{1-p_i} = z_i = \beta_0 + \beta_1 x_i$$

Оценивание параметров моделей бинарного выбора. Для оценивания параметров β в моделях бинарного выбора обычно используют метод максимального правдоподобия. Общее уравнение правдоподобия:

$$P(y_i = 1 | x_i; \beta) = F(x_i'\beta),$$

Подставив $L(\beta) = \prod_{i=1}^{N} p(y_i = 1 | x_i; \beta)^{y_i} P(y_i = 0 | x_i; \beta)^{1-y_i}$

получим, $\log L(\beta) = \sum_{i=1}^{N} y_i \log F(x_i'\beta) + \sum_{i=1}^{N} (1 - y_i) \log(1 - F(x_i'\beta)).$

Дифференцируя равенство по β, получим уравнение правдоподобия:

$$\frac{\partial \log L(\beta)}{\partial \beta} = \sum_{i=1}^{N} \left[\frac{y_i - F(x_i'\beta)}{F(x_i'\beta)(1 - F(x_i'\beta))} f(x_i'\beta) \right] x_i = 0, f = F'$$

Для логит-модели уравнение упрощается:

$$\frac{\partial \log L(\beta)}{\partial \beta} = \sum_{i=1}^{N} \left[y_i - \frac{\exp(x_i'\beta)}{1 + \exp(x_i'\beta)} \right] x_i = 0$$

Отсюда мы можем найти вероятность того, что $y_i=1$:

$$\hat{p}_i = \frac{\exp(x_i'\hat{\beta})}{1 + \exp(x_i'\hat{\beta})}$$

Уравнение правдоподобия является системой нелинейных (относительно β) уравнений и решается обычно итерационными методами. Для пробит- и логит-моделей данная функция является вогнутой по β, следовательно, решение уравнения правдоподобия дает оценку максимального правдоподобия параметров β_i. Унифицированного, как в линейной регрессии R^2, показателя качества «подгонки» модели не существует. Пусть, $\log L_f$ - значение функции правдоподобия исходной модели, $\log L_c$ -значение функции правдоподобия той же модели с нулевыми параметрами, но с константой. Чем больше их разность, тем лучше должна быть модель. На этой идее основаны нижеследующие показатели качества модели:

$$EpseudoR^2 = 1 - \frac{1}{1 + 2(\log L_f - \log L_c)/n}$$

$$McFaddenR^2 = 1 - \frac{\log L_f}{\log L_c}$$

Чем больше значение этих показателей, тем лучше модель. Данные показатели редко достигают значений, превышающих 0,5. Для проверки гипотезы о значимости коэффициентов моделей бинарного выбора применяют:

- тест Вальда (Wald test);
- тест множителей Лагранжа (Lagrange multiplier (LM) test);
- отношения правдоподобия (Likelihood ratio (LR) test).

Статистика Вальда имеет распределение χ^2 с числом степеней свободы, равным количеству ограничений в модели. Если наблюдаемое значение превышает критическое для заданного уровня значимости, то нулевая гипотеза о равенстве коэффициентов нулю отклоняется. В качестве аналога F-теста в линейной регрессии о совместной незначимости всех коэффициентов в бинарных моделях используют LR – тест.

$$LR = -2(\log \hat{L}_c - \log \hat{L}_f)$$

LR – тест имеет χ^2 распределение с числом степеней свободы, равным количеству независимых переменных в модели. Если наблюдаемое значение превышает критическое, то нулевая гипотеза о незначимости коэффициентов отклоняется в пользу альтернативной.

Модели множественного выбора с упорядоченными альтернативами. Модели множественного выбора (multinomial, multi-response models) используются в тех случаях, когда имеется более чем две альтернативы.

Различают: модели с упорядоченными альтернативами (ordered response models); модели с неупорядоченными альтернативами (unordered response models).

Если существует логическое упорядочивание M альтернатив, то может использоваться дискретная модель с упорядоченными альтернативами. Эта мо-

дель основывается на предположении о существовании одной ненаблюдаемой латентной переменной Y_i^*: $y^*_i = x_i'\beta + \varepsilon_i$

Стандартное нормальное распределение остатков дает упорядоченную probit-модель (ordered probit model). Логистическое распределение остатков дает упорядоченную logit-модель (ordered logit model).

Для случая трех альтернатив:

$$P(y_i = 1|x_i) = P(y_i^* \le 0|x_i) = \hat{O}(-x_i'\beta)$$
$$P(y_i = 3|x_i) = P(y_i^* > \gamma|x_i) = 1 - \hat{O}(\gamma - x_i'\beta)$$
$$P(y_i = 2|x_i) = \hat{O}(\gamma - x_i'\beta) - \hat{O}(-x_i'\beta)$$

Для случая M вариантов выбора:

$$P(y_i = 0|x_i) = P(y_i^* \le 0|x_i) = \hat{O}(-x_i'\beta)$$
$$P(y_i = 1|x_i) = \hat{O}(\gamma_1 - x_i'\beta) - \hat{O}(-x_i'\beta)$$
$$P(y_i = 2|x_i) = \hat{O}(\gamma_2 - x_i'\beta) - \hat{O}(\gamma_1 - x_i'\beta)$$
$$\ldots$$
$$P(y_i = M|x_i) = P(y_i^* > \gamma|x_i) = 1 - \hat{O}(\gamma_{M-1} - x_i'\beta)$$

Оценивание осуществляется при помощи метода максимального правдоподобия, где перечисленные вероятности включены в функцию правдоподобия.

$$\log L(\beta,\gamma) = \sum_{i:y_i=0} \log(\Pr(y_i = 0|x_i,\beta,\gamma)) + \sum_{i:y_i=1} \log(\Pr(y_i = 1|x_i,\beta,\gamma)) +$$
$$+ \ldots + \sum_{i:y_i=M} \log(\Pr(y_i = M|x_i,\beta,\gamma)).$$

Модели множественного выбора с неупорядоченными альтернативами. В некоторых случаях не существует естественного упорядочивания между альтернативами. Например, при моделировании способа передвижения (автобус, поезд, машина, велосипед, пешком). Предполагается существование случайной полезности, которая влияет на выбор альтернатив. Случайные полезности являются линейными функциями от наблюдаемых характеристик и имеют аддитивно-разделяемую структуру.

Полезность:

$$U_{ij} = \mu_{ij} + \varepsilon_{ij}$$

μ_{ij} - неслучайная функция наблюдаемых неизвестных параметров;

ε_{ij} – ненаблюдаемый остаточный член.

$$P(y_i = j) = P(U_{ij} = \max(U_{i1},..U_{iM})) =$$
$$= P(\mu_{ij} + \varepsilon_{ij} > \max_{k=1,...J,k\neq j} (\mu_{ik} + \varepsilon_{ik}).$$

Предположим, что все ε_{ij} взаимно независимы и распределены по закону распределения Вейбулла.

$$P(y_i = j) = \frac{\exp(\mu_{ij})}{\exp(\mu_{i1}) + \exp(\mu_{i2}) + ... + \exp(\mu_{iM})}$$
$$0 \le P(y_i = j) \le 1$$
$$\sum_{j=1}^{M} P(y_i = j) = 1$$

Один из уровней полезности принимают равным нулю (μ_{i1}=0) и полагают, что μ_{ij} является линейной функцией от наблюдаемых переменных:

$$\mu_{ij} = x'_{ij}\beta$$

Д$P(y_i = j) = \frac{\exp(x'_{ij}b)}{1 + \exp(x'_{i2}b) + ... + \exp(x'_{iM}b)}, j = 1,2,..M.$t – модель с множественными альтернативами (Multinomial Logit Model (MNL) или Independet Logit Model).

Вопросы для самоконтроля

1. В каких ситуациях фиктивная переменная используется в качестве зависимой переменной?

2. Какие законы распределения чаще всего используются в моделях бинарного выбора?

2. В чем суть логит-модели?

3. В чем суть пробит-модели?

4. Какова интерпретация коэффициентов моделей бинарного выбора?

5. Как осуществляется проверка значимости коэффициентов в модели бинарного выбора?

6. Как получить прогноз вероятности по логит-модели?

7. Как получить прогноз вероятности по пробит-модели?

8. Можно ли рассчитать по логит-модели коэффициент детерминации?

9. В чем отличие моделей упорядоченного и неупорядоченного выбора?

Лекция 15

Тема 13. Модели панельных данных

Вопросы для изучения

1. Основные понятия и характеристики панельных данных.

2. Модель сквозной регрессии и модель регрессии со случайным индивидуальным эффектом. Оценивание модели со случайным индивидуальным эффектом.

Аннотация. Данная тема раскрывает особенности построения моделей регрессии по панельным данным.

Ключевые слова. Панельные данные, сбалансированная панель, фиксированные эффекты, случайные эффекты.

Методические рекомендации по изучению темы

•Изучить лекционную часть, где даются общие представления по данной теме.

• Для закрепления теоретического материала ответить на вопросы для самоконтроля.

• Для проверки усвоения темы выполнить тест для самоконтроля.

Рекомендуемые информационные ресурсы:

1. http://tulpar.kfu.ru/course/view.php?id=2213.

2. Эконометрика: учебник / Под ред. И. И. Елисеевой. 2-е изд. -М.: Финансы и статистика, 2008. – 576 с. С.495-556.

3. Эконометрика: учеб. / под ред. В. С. Мхитаряна.- М.: Проспект, 2008. – 384 с. С. 133-178.

Основные понятия и характеристики панельных данных. Множество данных, состоящих из наблюдений за однотипными статистическими объектами, в течение нескольких временных периодов, называется панельными, или

пространственными, данными. Когда периодов времени больше числа наблюдаемых объектов, панельные данные называют также объединенным временным рядом (pooled time series). Например, имеется выборка по n=10 однотипным объектам с наблюдениями в периоды t=1, t=2.

Сведения по студентам (панельные данные за два периода)

Студент	Семестр 1		Семестр 2	
	время	баллы	время	баллы
1	60	81	60	84
2	100	75	120	87
3	30	60	60	79
4	45	82	30	78
5	120	78	150	87
6	180	95	150	92
7	100	79	100	84
8	60	92	80	97
9	90	78	90	75
10	90	67	60	66

По этим данным можно построить пять осмысленных регрессий:

- отдельно по первому семестру;
- отдельно по второму семестру;
- по объединению первого и второго семестров;
- по разности между первым и вторым семестрами с константой;
- по разности между первым и вторым семестрами без константы.

Результаты оценивания регрессий на основе сведений по студентам

(в скобках дано значение стандартного отклонения)

Модель	Константа β_0	Наклон β_1	R^2
Семестр 1	68,570 (7,203)	0,116 (0,075)	0,231
Семестр 2	72,605 (6,485)	0,114 (0,066)	0,271
Объединение	70,444 (4,725)	0,117 (0,049)	0,242
Разности с константой	3,512 (1,417)	0,275 (0,066)	0,685
Разности без константы	-	0,294 (0,082)	0,588

Допустим, что $y_{it} = \beta_0 + \beta_1 x_{it} + u_{it}$. Если вычтем уравнение для t=1 из уравнения для t=2, то получим $(y_{i2} - y_{i1}) = \beta_1(x_{i2} - x_{i1}) + (u_{i2} - u_{i1})$. Панельные данные позволяют учитывать ненаблюдаемую гетерогенность. Чтобы смоделировать разницу между измерениями в два разных периода времени вводят фиктивную переменную (d2=0 при t=1, d2=1 при t=2): $y_{it} = \alpha d2_t + \beta_{0i} + \beta_1 x_{it} + u_{it}$

Для d2=0 при t=1: $(y_{2t} - y_{1t}) = \alpha + \beta_1(x_{2t} - x_{1t}) + (u_{2t} - u_{1t})$.

Свойства панельных данных:

- позволяют учесть в модели ненаблюдаемую гетерогенность;

- позволяют идентифицировать потоки или перемещения между различными состояниями наблюдаемых объектов.

Сбалансированной панелью называют панельные данные, в которых нет пропущенных наблюдений. Сокращение объектов в выборке называют панельным истощением. Ротационной панелью называют панельные данные, в которых в обследуемую выборку периодически добавляется новый объект. В микроэконометрических панелях объекты наблюдения – индивиды, домохо-

зяйства, предприятия. В макроэконометрических панелях объектами наблюдения служат страны, регионы, города.

Описательный анализ данных включает ряд этапов:

- проверка множества данных на несоответствия, пропущенные значения, ошибки форматирования;

- анализ амплитуды разброса, наличие возможных выбросов или кластеров;

- проверка на коллинеарность переменных;

- графическая визуализация панельных данных.

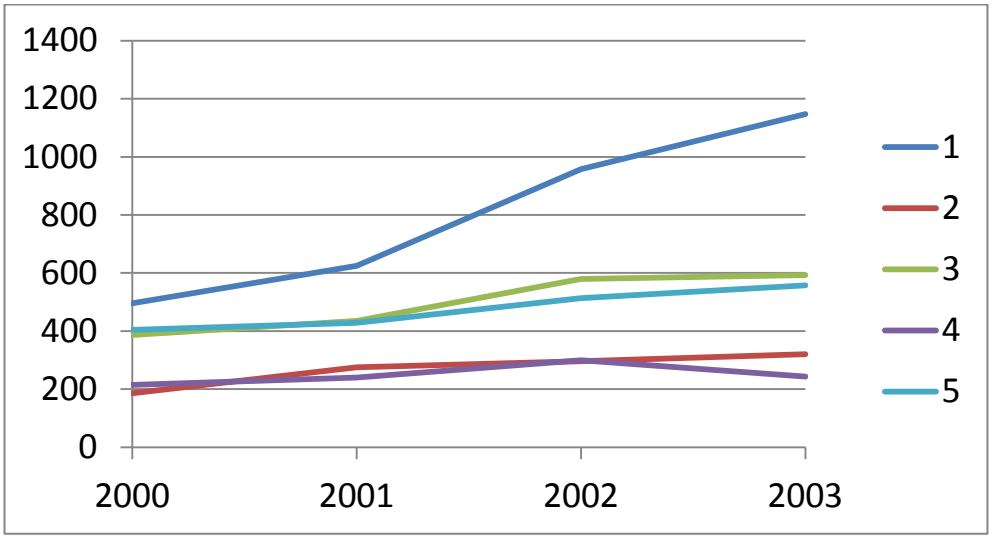


Рис. 13.1. Динамика рыночной стоимости

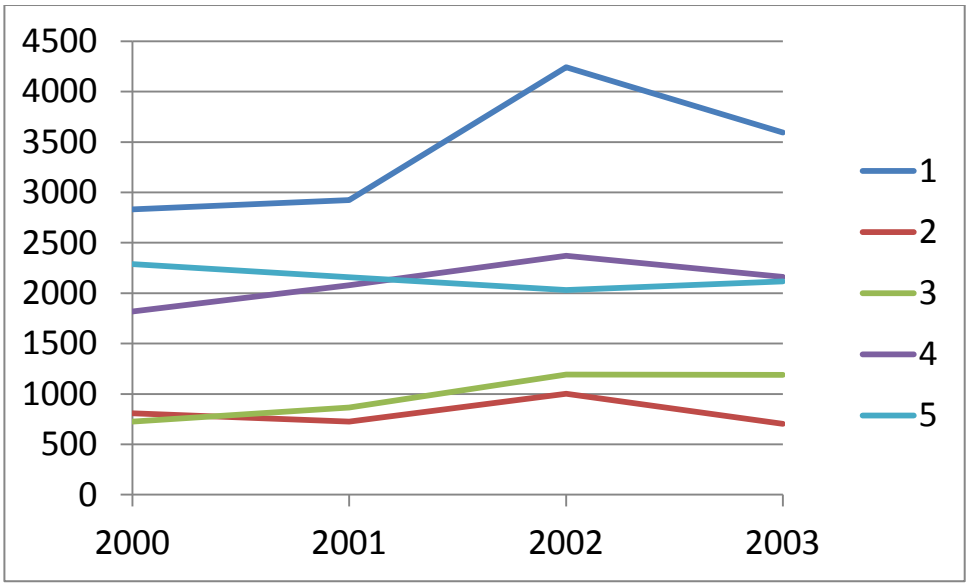


Рис. 13.2. Динамика валового оборота

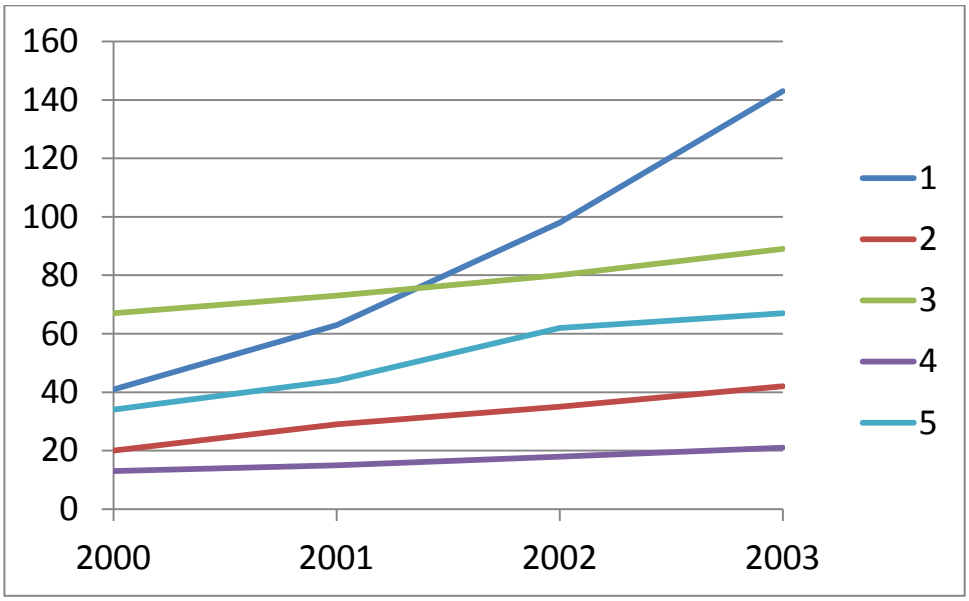


Рис. 13.3. Динамика прибыли

Поскольку панельные данные имеют временное и пространственное измерения, можно записать их в виде матрицы.

$$y = \begin{pmatrix} y_{11} \\ \vdots \\ y_{1T} \\ \vdots \\ y_{n1} \\ \vdots \\ y_{nT} \end{pmatrix}, X = \begin{pmatrix} & & \\ x_{1,11} & \cdots & x_{d,11} \\ \vdots & \ddots & \vdots \\ x_{1,1T} & \cdots & x_{d,1T} \\ \vdots & & \vdots \\ x_{1,n1} & \cdots & x_{d,n1} \\ \vdots & \ddots & \vdots \\ x_{1,nT} & \cdots & x_{d,nT} \end{pmatrix}$$

(«быстрый» индекс: время; «медленный» индекс: объекты наблюдения)

Фирма (индекс)	Год (период времени)	Рыночная стоимость (Y)	Оборот (X1)	Прибыль (X2)
1	2000	496	2833	41
1	2001	625	2925	63
1	2002	958	4242	98
1	2003	1147	3594	143

2	2000	186	809	20
2	2001	275	727	29
2	2002	296	1002	35
2	2003	320	703	42
3	2000	387	724	67
3	2001	435	864	73
3	2002	580	1194	80
3	2003	593	1189	89
4	2000	215	1819	13
4	2001	240	2080	15
4	2002	300	2372	18
4	2003	243	2160	21
5	2000	404	2290	34
5	2001	429	2159	44
5	2002	513	2031	62
5	2003	557	2116	67

В модели с фиксированными эффектами моделируется эффект гетерогенности между объектами наблюдения с инвариантным по отношению ко времени, но специфическим для каждого объекта наблюдения параметром местоположения μ_i. Это в точности модель с фиктивными переменными.

$$y_{it} = \mu_i + x'_{it}\beta + u_{it}$$

Модель с фиксированными эффектами - это простая регрессионная модель, оценки параметров тестируют с помощью обычных t- и F – тестов.

Проверка на наличие фиксированных эффектов выполняется с помощью распределения Фишера.

$$F = \frac{RSS_{pool} - RSS_{FE}}{RSS_{FE}} \cdot \frac{nT - n - d}{n - 1} =$$
$$= \frac{R^2_{FE} - R^2_{pool}}{1 - R^2_{FE}} \cdot \frac{nT - n - d}{n - 1}$$

$$F > F(\alpha, n-1, nT - n - d) \Rightarrow H_1 : \mu_i \neq 0$$

Панельные данные применяются в эмпирических исследованиях с 60-х годов XX века. Впервые сбор панельных данных осуществлен в США.

Базы панельных данных в США:

- National Longitudinal Surveys of Labor Market Experience;
- University of Michigan's Panel Study of Income Dynamics.

В России сбор панельных данных начался в 90-е годы XX века.

Базы панельных данных в России:

- Rassia Longitudinal Monitoring Survay – РМЭЗ – Российский мониторинг экономического положения и здоровья населения (доступно в Интернет);
- Российский экономический тренд (доступно в Интернет);
- Российский экономический барометр (платный доступ).

Предположения простейших моделей панельных данных:

- Статические модели, без лаговых значений зависимых переменных;
- Сбалансированные панели с одинаковым числом временных тактов;
- Панели с короткими временными рядами;
- Включение аддитивных фиктивных переменных для отражения

временного эффекта;

- Учет ненаблюдаемых и неизменяемых во времени характеристик объектов выборки – индивидуального эффекта.

Модель сквозной регрессии и модель регрессии со случайным индивидуальным эффектом. Оценивание модели со случайным индивидуальным эффектом. В модели со случайными эффектами моделируется эффект гетерогенности объектов наблюдения путем введения неизменного во времени, но специфического для каждого объекта наблюдения слагаемого ошибки m_i, которое предполагается независимым от оставшейся части ошибки $\upsilon_{it.}$

$$y_{it} = \mu_i + x'_{it}\beta + u_{it}$$
$$u_{it} = m_i + \upsilon_{it}$$

Эффекты m_i, описывающие гетерогенность, являются случайными переменными в смысле случайности выборки из генеральной совокупности, поскольку каждый объект наблюдения имеет специфический, не зависящий от времени, эффект. МНК – оценки в модели со случайными эффектами неэффективны из-за присутствия автокорреляции в слагаемом ошибки m_i. Применяется двухшаговая процедура обобщенного метода наименьших квадратов – ВОМНК – выполнимый обобщенный метод наименьших квадратов.

Вводится вспомогательная переменная:

$$\theta = 1 - \frac{\sigma^2{}_V}{\sqrt{\sigma_V^2 + T\sigma_M^2}} = 1 - \left(\frac{1}{1 + T\left(\frac{\sigma_M^2}{\sigma_V^2}\right)}\right)^{\frac{1}{2}}$$

Поскольку на практике дисперсии не известны, заменяем их на состоятельные оценки:

$$\theta = 1 - \frac{\hat{\sigma}^2{}_V}{\sqrt{\hat{\sigma}_V^2 + T\hat{\sigma}_M^2}}$$

ВОМНК- оценка модели со случайным эффектом:

$$(y_{it} - \theta\bar{y}_i) = \mu(1-\theta) + (x_{it} - \theta\bar{x}_i)'\beta + (u_{it} - \theta\bar{u}_i)$$

$$\hat{\sigma}_V^2 = \frac{1}{nT-n-d}\sum_{i=1}^{n}\sum_{t=1}^{T}(\hat{u}_{it}-\hat{u}_i)^2$$

$$\hat{\sigma}_A^2 = \frac{1}{n-d-1}\sum_{i=1}^{N}\hat{u}_i{}^2$$

$$\hat{\sigma}_M^2 = \hat{\sigma}_A^2 - \frac{1}{T}\hat{\sigma}_V^2$$

Проверка на наличие случайных эффектов выполняется с помощью теста множителей Лагранжа:

$$LM = \frac{nT}{2(T-1)}\left(1-\frac{\sum_{i=1}^{n}(\sum_{t=1}^{T}u_{it})^2}{\sum_{i=1}^{n}\sum_{t=1}^{T}u_{it}^2}\right)^2$$

$$P\left[\chi_1^2 > LM\right] \Rightarrow H_1 : \sigma_M^2 > 0$$

Гетерогенность присутствует, можно предположить наличие случайных эффектов. Фиксированные и случайные эффекты – это случайные переменные. Оба эффекта моделируют ненаблюдаемые различия в объектах наблюдения. Фиксированные эффекты – параметры. Случайные эффекты – слагаемые ошибок. Фиксированные эффекты могут коррелировать с регрессорами. Случайные эффекты предполагаются некоррелированными с регрессорами.

Какие эффекты моделировать?

Тест Хаусмана:

$$H = (\hat{\beta}_{FE} - \hat{\beta}_{RE})'\hat{O}^{-1}(\hat{\beta}_{FE} - \hat{\beta}_{RE})$$

$$P\left[\chi_2^2 > H\right] \Rightarrow H_0$$

Вопросы для самоконтроля

1. Какие данные называют панельными?

2. Назовите преимущества использования панельных данных.

3. В чем отличия моделей с фиксированными и случайными эффектами для панельных данных?

4. Можно ли модель с фиксированными эффектами для панельных данных рассматривать как частный случай использования фиктивных переменных?

5. Охарактеризуйте роль инструментальных переменных в оценивании моделей по панельным данным.

6. Для проверки какой гипотезы применяется тест Хаусмана?

7. Как проверить значимость фиксированных эффектов и случайных эффектов?

8. Каковы достоинства и недостатки моделей фиксированных и случайных эффектов?

Лекция 16

Тема 14. Ошибки спецификации

Вопросы для изучения

1. Спецификация регрессионной модели.

2. Исключение существенных переменных и включение несущественных переменных.

3. Замещающие переменные в регрессионных моделях.

Аннотация. Данная тема раскрывает типы ошибок спецификации, последствия исключения существенных переменных и включения несущественных переменных, использования замещающих переменных.

Ключевые слова. Спецификация модели, ошибки спецификации, замещающие переменные.

Методические рекомендации по изучению темы

• Изучить лекционную часть, где даются общие представления по данной теме.

• Для закрепления теоретического материала ответить на вопросы для самоконтроля.

• Для проверки усвоения темы выполнить тест для самоконтроля и практические задания.

• Для подготовки к экзамену выполнить итоговый тест и итоговые практические задания.

Рекомендуемые информационные ресурсы:

1. http://tulpar.kfu.ru/course/view.php?id=2213.

2. Эконометрика: учебник / Под ред. И. И. Елисеевой. 2-е изд. -М.: Финансы и статистика, 2008.- 576 с. С. 109-125.

3. Эконометрика. Практикум: [Электронный ресурс] Учебное пособие / С.А. Бородич. - М.: НИЦ ИНФРА-М; Мн.: Нов. знание, 2014. - 329 с. (http://znanium.com/catalog.php?item=booksearch&code=%D1%8D%D0%BA%D0%BE%D0%BD%D0%BE%D0%BC%D0%B5%D1%82%D1%80%D0%B8%D0%BA%D0%B0&page=4#none) С.174-197.

4. Эконометрика: [Электронный ресурс] Учеб. пособие / А.И. Новиков. - 3-е изд., испр. и доп. - М.: ИНФРА-М, 2014. - 272 с.: (http://znanium.com/catalog.php?item=booksearch&code=%D1%8D%D0%BA%D0%BE%D0%BD%D0%BE%D0%BC%D0%B5%D1%82%D1%80%D0%B8%D0%BA%D0%B0&page=1#none) С. 62-63.

Спецификация модели. Любое эконометрическое исследование начинается со спецификации модели, т.е. с формулировки вида модели, исходя из соответствующей теории связи между переменными. Из всего круга факторов, влияющих на результативный признак, необходимо выделить наиболее существенно влияющие факторы. С другой стороны, выбрать максимально адекватную форму модели. Для случая парной регрессии подбор модели обычно осуществляется по виду расположения наблюдаемых точек на корреляционном поле. Однако нередки ситуации, когда расположение точек приблизительно соответствует нескольким функциям. Например, криволинейные зависимости могут аппроксимироваться полиномиальной, показательной, степенной, логарифмической функциями. Еще более неоднозначна ситуация для множественной регрессии. Чтобы выбрать качественную модель, необходимо ответить на ряд вопросов, возникающих при ее анализе:

1. Каковы признаки «хорошей» модели?

2. Какие ошибки спецификации встречаются, и каковы последствия данных ошибок?
3. Как обнаружить ошибку спецификации?
4. Каким образом можно исправить ошибку спецификации и перейти к лучшей (качественной) модели?

Для построения «хорошей» модели и сравнения ее с другими возможными моделями необходимо учитывать следующие свойства (критерии). *Скупость (простота).* Модель должна быть максимально простой, поскольку она является упрощением действительности и не отражает ее идеально. Поэтому из двух моделей, приблизительно одинаково отражающих реальность, предпочтение отдается модели, содержащей меньшее число объясняющих переменных. *Единственность.* Для любого набора статистических данных определяемые коэффициенты должны вычисляться однозначно. *Максимальное соответствие.* Уравнение тем лучше, чем большую часть разброса зависимой переменной оно может объяснить. Поэтому стремятся построить уравнение с максимально возможным скорректированным коэффициентом детерминации. *Согласованность с теорией.* Никакое уравнение не может быть признано качественным, если оно не соответствует известным теоретическим предпосылкам. Например, если в функции спроса коэффициент при цене положителен, то даже значительная величина коэффициента детерминации не позволит признать уравнение удовлетворительным. *Прогнозные качества.* Модель может быть признана качественной, если полученные на ее основе прогнозы подтверждаются реальностью. Критерием прогнозных качеств оцененной модели регрессии может служить следующее отношение:

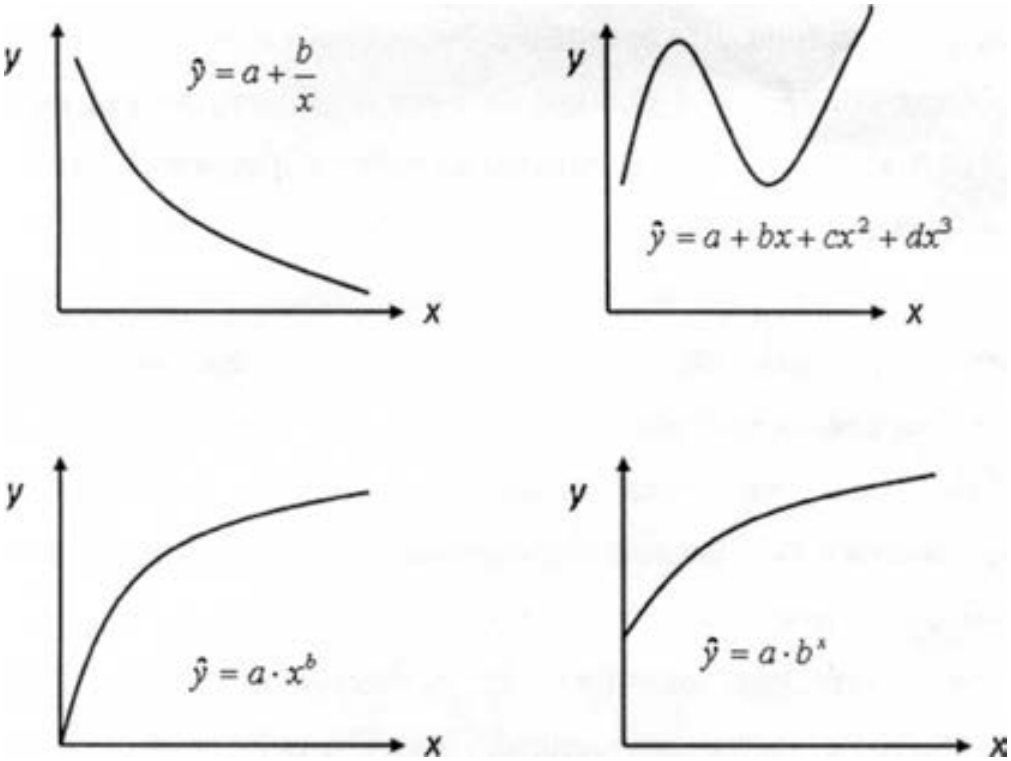
$$V = \frac{S}{\bar{y}}, S = \sqrt{\frac{\sum e_i^2}{n-m-1}}$$

Если величина V мала и отсутствует автокорреляция остатков, то прогнозные качества модели высоки. Неправильный выбор функциональной формы или набора объясняющих переменных называется ошибками спецификации. Рассмотрим основные типы ошибок спецификации. Это прежде всего неправильно выбранная форма модели. В частности, зависимость спроса от цены может

быть выражена линейно $\hat{y}_x = a - b \cdot x$, но возможны и другие соотношения, например

$$\hat{y}_x = ax^{-b}, \hat{y}_x = a + \frac{b}{x}, \hat{y}_x = \frac{1}{a + bx}.$$

Ошибки спецификации тем меньше, чем в большей мере теоретические значения признака подходят к фактическим данным Y. К ошибкам спецификации относится также недоучет в уравнении регрессии какого-либо существенного фактора, т.е. использование парной регрессии вместо множественной. Например, спрос на конкретный товар может определяться не только ценой, но и доходом на душу населения. *Ошибки выборки:* Исследователь при установлении связи между признаками имеет дело с выборочными данными. При изучении экономических процессов данные в исходной совокупности часто являются неоднородными. В этом случае уравнение регрессии не имеет практического смысла. Поэтому для получения хорошего результата из выборки исключают наблюдения с аномальными значениями исследуемых признаков. *Ошибки измерения:* Представляют наибольшую опасность в практическом использовании методов регрессии. Ошибки спецификации можно уменьшить, изменяя форму модели, ошибки выборки - увеличивая объем исходных данных, ошибки измерения сводят на нет все усилия по количественной оценке связи между признаками. Например, статистическое измерение дохода на душу населения может иметь ошибку в результате наличия сокрытых доходов. Другой пример: органы государственной статистики получают балансы предприятий, достоверность которых никто не подтверждает. В эконометрических исследованиях предполагается, что ошибки измерения сведены к минимуму. Поэтому основное внимание уделяется ошибкам спецификации модели. В парной регрессии выбор вида математической функции (1) может быть осуществлен тремя методами: графическим, аналитическим и экспериментальным. Графический метод достаточно нагляден. Он основан на поле корреляции. Рассмотрим типы кривых.



Используются и другие типы кривых:

$$\hat{y}_x = \frac{1}{a + b \cdot x}, \hat{y}_x = a + bx + \frac{c}{x}, \hat{y}_x = a + b \cdot \lg x, \hat{y}_x = \frac{1}{a + bx + cx^2}, \hat{y}_x = \frac{a}{1 + be^{-cx}}, \lg \hat{y}_x = a + bx + cx^2$$

Аналитический метод выбора типа уравнения регрессии основан на изучении материальной природы связи исследуемых признаков. Пусть, например, изучается потребность предприятия в электроэнергии *y* в зависимости от объема выпускаемой продукции x. Все потребление электроэнергии можно подразделить на 2 части:

- не связанное с производством продукции *a;*
- непосредственно связанное с объемом выпускаемой продукции, пропорционально возрастающее с увеличением объема выпуска *bx;*

Тогда зависимость потребления электроэнергии от объема продукции можно выразить уравнением регрессии вида:

$$\hat{y}_x = a + bx$$

Разделив на x, получим удельный расход электроэнергии на единицу продукции: $z_x = y / x$

$$\hat{z}_x = b + \frac{a}{x}$$

Это равносторонняя гипербола.

Аналогично затраты предприятия могут быть условно-переменные, изменяющиеся пропорционально изменению объема продукции (расход материала, оплата труда и др.) и условно-постоянные, не изменяющиеся с изменением объема производства (арендная плата, содержание администрации и др.). Соответствующая зависимость затрат на производство y от объема продукции x характеризуется линейной функцией

$$y = a + bx,$$

а зависимость себестоимости единицы продукции z_x от объема продукции - равносторонней гиперболой:

$$\hat{z}_x = b + \frac{a}{x}$$

Экспериментальный метод используется при обработке информации на компьютере путем сравнения величины остаточной дисперсии $D_{ост}$, рассчитанной на разных моделях. В практических исследованиях, как правило, имеет место некоторое рассеяние точек относительно линии регрессии. Оно обусловле-

$$D_{ост} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_{x_i})^2$$

но влиянием прочих, не учитываемых в уравнении регрессии факторов:

Чем меньше $D_{ост}$, тем меньше наблюдается влияние прочих факторов, тем лучше уравнение регрессии подходит к исходным данным. При обработке данных на компьютере разные математические функции перебираются в автоматическом режиме, и из них выбирается та, для которой $D_{ост}$ является наименьшей.

Если $D_{ост}$ примерно одинакова для нескольких функций, то на практике выбирают более простую, так как она в большей степени поддается интерпретации и требует меньшего объема наблюдений. Результаты многих исследований подтверждают, что число наблюдений должно в 6-7 раз превышать число рассчитываемых параметров при переменной x. Это означает, что искать линейную регрессию, имея менее 7 наблюдений, вообще не имеет смысла. Если вид функции усложняется, то требуется увеличение объема наблюдений. Для

рядов динамики, ограниченных по протяженности - 10, 20, 30 лет, - предпочтительна модель с меньшим числом параметров при x.

Исключение существенных переменных и включение несущественных переменных. Рассмотрим последствия отбрасывания значимой переменной. Пусть теоретическая модель, отражающая рассматриваемую экономическую зависимость, имеет вид:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon \qquad (1)$$

Данной модели соответствует следующее эмпирическое уравнение регрессии:

$$Y = b_0 + b_1 X_1 + b_2 X_2 + e \qquad (2)$$

Исследователь по каким-то причинам (недостаток информации, поверхностное знание о предмете исследования и т. п.) считает, что на переменную Y реально воздействует лишь переменная X_1. Он ограничивается рассмотрением модели

$$Y = \gamma_0 + \gamma_1 X_1 + v \qquad (3)$$

При этом он не рассматривает в качестве объясняющей переменную X_2, совершая ошибку отбрасывания существенной переменной.

Пусть эмпирическое уравнение регрессии, соответствующее теоретическому уравнению, имеет вид

$$Y = g_0 + g_1 X_1 + v \qquad (4)$$

Последствия данной ошибки достаточно серьезны. МНК-оценки (4) являются смещенными ($M(g_0) \neq \beta_0, M(g_1) \neq \beta_1$) и несостоятельными даже при бесконечно большом числе испытаний. Следовательно, возможные интервальные оценки и результаты проверки соответствующих гипотез будут ненадежными. При положительном β_2 и положительной коррелированности между X_1 и X_2 оценка g_1 будет завышать истинное значение β_1. Коэффициенты b_1 и b_2 (2) от-

ражают степень индивидуального воздействия на Y каждой из объясняющих переменных X_1 и X_2. В уравнении (4) через коэффициент g_1 отражается, кроме прямого воздействия переменной X_1, воздействие коррелированной с ней и не учтенной переменной X_2. Таким образом, косвенная роль переменной X_2 в уравнении (4) отражается на оценке параметра β_1, изменяя ее в среднем на величину β_2. Единственно возможным условием получения несмещенной оценки для коэффициента β_1 является некоррелированность Xi и X_2.

Ошибка данного рода существенно отражается и на коэффициенте детерминации. Его значение будет завышать роль переменной X_1 в объяснении дисперсии переменной Y. Это связано с косвенным присутствием в уравнении через коэффициент g_1 переменной X_2, что повышает объясняющую способность уравнения в целом.

Рассмотрим последствия ошибки добавления незначимой переменной. Пусть теоретическая модель имеет следующий вид:

$$Y = \beta_0 + \beta_1 X_1 + \varepsilon \qquad (5)$$

Пусть исследователь подменяет ее более сложной моделью:

$$Y = \gamma_0 + \gamma_1 X_1 + \gamma_2 X_2 + \varepsilon, \qquad (6)$$

добавляя при этом не оказывающую реального воздействия на Y объясняющую переменную X_2. В этом случае совершается ошибка добавления несущественной переменной. Последствия данной ошибки будут не столь серьезными, как в предыдущем случае. Оценки коэффициентов, найденные для модели (6) остаются, как правило, несмещенными и состоятельными. Однако их точность уменьшится, увеличивая при этом стандартные ошибки, т. е. оценки становятся неэффективными, что отразится на их устойчивости. Увеличение дисперсии оценок может привести к ошибочным результатам проверки гипотез относительно значений коэффициентов регрессии, расширению интервальных оценок.

Рассмотрим последствия выбора неправильной функциональной формы модели. Пусть правильная регрессионная модель имеет вид:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

Любое эмпирическое уравнение регрессии с теми же переменными, но имеющее другой функциональный вид, приводит к искажению истинной зависимости. Например, в следующих уравнениях:

$$\ln Y = a_0 + a_1 X_1 + a_2 X_2 + e,$$
$$Y = c_0 + c_1 \ln X_1 + c_2 \ln X_2 + u$$

совершена ошибка выбора неправильной функциональной формы уравнения регрессии. Последствия данной ошибки будут серьезными. Обычно такая ошибка приводит либо к получению смещенных оценок, либо к ухудшению статистических свойств оценок коэффициентов регрессии и других показателей качества уравнения. Это вызвано нарушением условий Гаусса-Маркова для отклонений. Прогнозные качества модели в этом случае очень низкие.

При определении качества модели обычно анализируются следующие параметры:

1) скорректированный коэффициент детерминации;
2) t-статистики;
3) статистика Дарбина -Уотсона;
4) согласованность знаков коэффициентов с теорией;
5) прогнозные качества (ошибки) модели.

Если все эти показатели удовлетворительны, то данная модель может быть предложена для описания исследуемого реального процесса. Если же какая-либо из описанных выше характеристик не является удовлетворительной, то есть основания сомневаться в качестве данной модели (неправильно выбрана функциональная форма уравнения, не учтена важная объясняющая переменная, имеется объясняющая переменная, не оказывающая значимого влияния на зависимую переменную.

Замещающие переменные. Замещающие (proxy) переменные применяются вместо отсутствующих переменных. Причины их использования: отсутствие данных, трудность измерения, неточные данные. Отбрасывание существенной переменной приведет к смещенным и несостоятельным МНК-

оценкам. Замещающая переменная может дать косвенную информацию о той самой существенной переменной, z – замещающая переменная:

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_k X_k + u$$
$$X_2 = \lambda + \mu Z$$
$$Y = \beta_1 + \beta_2(\lambda + \mu Z) + \beta_3 X_3 + \ldots + \beta_k X_k + u =$$
$$= \beta_1 + \beta_2\lambda + \beta_2\mu Z + \beta_3 X_3 + \ldots + \beta_k X_k + u$$

Например, исследуется вопрос об «утечке» мозгов из страны А в страну В по показателю относительного уровня миграции трудовых ресурсов (М). Полагается, что при более высокой относительной разнице в заработной плате будет более высокой и миграция. Однако исследователь располагает данными только по ВВП на душу населения, а не по заработной плате. Поэтому вводится замещающая переменная G, которая является отношением ВВП страны В к ВВП страны А (строгая линейная зависимость):

$$M = \beta_1 + \beta_2 W + u$$
$$W = \lambda + \mu G$$
$$M = \beta_1 + \beta_2(\lambda + \mu G) + u$$
$$M = \beta_1 + \beta_2\lambda + \beta_2\mu G + u$$
$$\lambda = 0, \mu = 1 \Rightarrow M = \beta_1 + \beta_2 G + u$$

На практике обычно невозможно найти замещающую переменную, имеющую строгую линейную зависимость с недостающей переменной. Но если зависимость близка к линейной, то результаты приблизительно сохраняются. Основной проблемой является то, что не существует средств для проверки выполнения указанного условия.

Вопросы и задания для самоконтроля

1. Что понимается под спецификацией модели?
2. Каковы основные виды ошибок спецификации?
3. Каковы признаки «хорошей» модели?
4. Во сколько раз число наблюдений должно превышать число рассчитываемых параметров при переменной *x*?
5. Как можно обнаружить ошибки спецификации?
6. Каковы последствия исключения существенных переменных?

7. Каковы последствия включения несущественных переменных?

8. В чем состоит смысл замещающих переменных?

9. В чем суть теста Рамсея?

10. Как можно исправить ошибку спецификации?

Задание 1. При построении регрессионной зависимости некоторого результативного признака на 8 факторов по 25 измерениям коэффициент множественной детерминации составил 0,736. После исключения 3 факторов коэффициент детерминации уменьшился до 0,584. Проверить, обосновано ли было принятое решение на уровнях значимости 0,1; 0,05; 0,01?

Задание 2. При построении регрессионной зависимости некоторого результативного признака на 10 факторов по 45 наблюдениям коэффициент множественной детерминации составил 0,347. После добавления 3 факторов коэффициент детерминации увеличился до 0,536. Проверить, обосновано ли было принятое решение на уровнях значимости 0,1; 0,05; 0,01?

Лекция 17

Тема 15. Модели одномерных временных рядов

Вопросы для изучения

1. Понятие временного ряда и его основные компоненты.

2. Построение аддитивной модели.

3. Построение мультипликативной модели.

Аннотация. Данная тема раскрывает порядок построения аддитивных и мультипликативных моделей одномерных временных рядов.

Ключевые слова. Тренд, сезонные и случайные колебания, аддитивная модель, мультипликативная модель.

Методические рекомендации по изучению темы

•Изучить лекционную часть, где даются общие представления по данной теме.

• Для закрепления теоретического материала ознакомиться с решениями типовых задач и ответить на вопросы для самоконтроля.

• Для проверки усвоения темы выполнить практические задания и тест для самоконтроля.

Рекомендуемые информационные ресурсы:

1. http://tulpar.kfu.ru/course/view.php?id=2213.

2. Валентинов В. А. Эконометрика [Электронный ресурс]: Практикум / В. А. Валентинов. - 3-е изд. - М.: Дашков и К, 2010. - 436 с.

(http://znanium.com/catalog.php?item=booksearch&code=%D1%8D%D0%BA%D0%BE%D0%BD%D0%BE%D0%BC%D0%B5%D1%82%D1%80%D0%B8%D0%BA%D0%B0&page=3#none) С. 242-261.

3. Эконометрика: учебник / И. И. Елисеева. – М.: Проспект, 2010. – 288 с. С.128-183.

4. Электронный курс “Time Series Econometrics”, Princeton University, URL: http://sims.princeton.edu/yftp/Times05/;https://blackboard.princeton.edu/webapps/portal/frameset.jsp?tab_group=courses&url=%2Fwebapps%2Fblackboard%2Fexecute%2FcourseMain%3Fcourse_id%3D_52968_1.

Понятие временного ряда и его основные компоненты. Временнóй ряд – это совокупность значений какого – либо показателя за несколько последовательных моментов или периодов времени. Каждое значение (уровень) временного ряда формируется под воздействием большого числа факторов, которые можно условно разделить на три группы: факторы, формирующие тенденцию ряда; факторы, формирующие циклические колебания ряда; случайные факторы.

Тенденция характеризует долговременное воздействие факторов на динамику показателя. Тенденция может быть возрастающей или убывающей.

Циклические колебания могут носить сезонный характер или отражать динамику конъюнктуры рынка, а также фазу бизнес – цикла.

Реальные данные часто содержат все три компоненты. В большинстве случаев временной ряд можно представить как сумму или произведение трендовой (T), циклической (S) и случайной (E) компонент. В случае суммы имеет место аддитивная модель временного ряда:

$$y = T + S + E, \qquad (1)$$

в случае произведения – мультипликативная модель:

$$y = T \cdot S \cdot E. \qquad (2)$$

Основная задача эконометрического исследования отдельного временного ряда – получение количественного выражения каждой из компонент и использование этой информации для прогноза будущих значений ряда или построение модели взаимосвязи двух или более временных рядов.

Сначала рассмотрим основные подходы к анализу отдельного временного ряда. Такой ряд может содержать, помимо случайной составляющей, либо только тенденцию, либо только сезонную (циклическую) компоненту, либо все компоненты вместе. Для того, чтобы выявить наличие той или иной неслучайной компоненты, исследуется корреляционная зависимость между последовательными уровнями временного ряда, или автокорреляция уровней ряда. Основная идея такого анализа заключается в том, что при наличии во временном ряде тенденции и циклических колебаний значения каждого последующего уровня ряда зависят от предыдущих.

Количественно автокорреляцию можно измерить с помощью линейного коэффициента корреляции между уровнями исходного временного ряда и уровнями этого ряда, сдвинутыми на несколько шагов во времени.

Коэффициент автокорреляции уровней ряда первого порядка измеряет зависимость между соседними уровнями ряда t и $t-1,$ т.е. при лаге 1.

Он вычисляется по следующей формуле:

$$r_1 = \frac{\sum_{t=2}^{n}(y_t - \bar{y}_1)(y_{t-1} - \bar{y}_2)}{\sqrt{\sum_{t=2}^{n}(y_t - \bar{y}_1)^2 \sum_{t=2}^{n}(y_{t-1} - \bar{y}_2)^2}}, \quad (3)$$

где в качестве средних величин берутся значения:

$$\bar{y}_1 = \frac{\sum_{t=2}^{n} y_t}{n-1}; \qquad \bar{y}_2 = \frac{\sum_{t=2}^{n} y_{t-1}}{n-1}. \quad (4)$$

В первом случае усредняются значения ряда, начиная со второго до последнего, во втором случае - значения ряда с первого до предпоследнего.

Формулу (3) можно представить как формулу выборочного коэффициента корреляции:

$$r_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}, \quad (5)$$

где в качестве переменной x берется ряд $y_2, y_3, \ldots, y_n$, а в качестве переменной $y-$ ряд $y_1, y_2, \ldots, y_{n-1}$.

Если значение коэффициента (3) близко к единице, это указывает на очень тесную зависимость между соседними уровнями временного ряда и о наличии во временном ряде сильной линейной тенденции.

Аналогично определяются коэффициенты автокорреляции более высоких порядков. Так, коэффициент автокорреляции второго порядка характеризует тесноту связи между уровнями y_t и y_{t-2} и определяется по формуле:

$$r_2 = \frac{\sum_{t=3}^{n}(y_t - \bar{y}_3)(y_{t-2} - \bar{y}_4)}{\sqrt{\sum_{t=3}^{n}(y_t - \bar{y}_3)^2 \sum_{t=3}^{n}(y_{t-2} - \bar{y}_4)^2}}, \quad (6)$$

где в качестве одной средней величины берут среднюю уровней ряда с третьего до последнего, а в качестве другой - среднюю всех уровней ряда, кроме последних двух:

$$\bar{y}_3 = \frac{\sum_{t=3}^{n} y_t}{n-2}; \qquad \bar{y}_4 = \frac{\sum_{t=3}^{n} y_{t-2}}{n-2}. \tag{7}$$

Число периодов, по которым рассчитывается коэффициент автокорреляции, называют лагом. С увеличением лага число пар значений, по которым рассчитывается коэффициент автокорреляции, уменьшается. Для обеспечения статистической достоверности максимальный лаг, как считают некоторые известные эконометристы, не должен превышать четверти общего объема выборки.

Коэффициент автокорреляции строится по аналогии с линейным коэффициентом корреляции, и поэтому он характеризует тесноту только линейной связи текущего и предыдущего уровней ряда. По нему можно судить о наличии линейной или близкой к линейной тенденции. Однако для некоторых временных рядов с сильной нелинейной тенденцией (например, параболической или экспоненциальной), коэффициент автокорреляции уровней ряда может приближаться к нулю.

Кроме того, по знаку коэффициента автокорреляции нельзя делать вывод о возрастающей или убывающей тенденции в уровнях ряда. Большинство временных рядов экономических данных имеют положительную автокорреляцию уровней, однако при этом не исключается убывающая тенденция.

Последовательность коэффициентов автокорреляции уровней различных порядков, начиная с первого, называется автокорреляционной функцией временного ряда. График зависимости ее значений от величины лага называется коррелограммой. Анализ автокорреляционной функции и коррелограммы помогает выявить структуру ряда. Здесь уместно привести следующие качественные рассуждения.

Если наиболее высоким является коэффициент автокорреляции первого порядка, очевидно, исследуемый ряд содержит только тенденцию. Если наиболее высоким оказался коэффициент автокорреляции порядка τ, ряд содержит циклические колебания с периодичностью в τ моментов времени. Если ни один из коэффициентов автокорреляции не является значимым, то либо ряд не содержит тенденции и циклических колебаний и имеет только случайную составляющую, либо ряд содержит сильную нелинейную тенденцию, для исследования которой нужно провести дополнительный анализ.

В случае, если при анализе структуры временного ряда обнаружена только тенденция и отсутствуют циклические колебания (случайная составляющая присутствует всегда), следует приступать к моделированию тенденции. Если же во временном ряде имеют место и циклические колебания, прежде всего следует исключить именно циклическую составляющую, и лишь затем приступать к моделированию тенденции. Выявление тенденции состоит в построении аналитической функции, характеризующей зависимость уровней ряда от времени, или тренда. Этот способ называют аналитическим выравниванием временного ряда.

Зависимость от времени может принимать разные формы, поэтому для еѐ формализации используют различные виды функций:

линейный тренд: $\hat{y}_t = a + b \cdot t$;

гипербола: $\hat{y}_t = a + b / t$;

экспоненциальный тренд: $\hat{y}_t = e^{a+b \cdot t}$ (или $\hat{y}_t = a \cdot b^t$);

степенной тренд: $\hat{y}_t = a \cdot t^b$;

параболический тренд второго и более высоких порядков:

$\hat{y}_t = a + b_1 \cdot t + b_2 \cdot t^2 + ... + b_k \cdot t^k$.

Параметры каждого из трендов можно определить обычным МНК, используя в качестве независимой переменной время $t = 1,2,...,n$, а в качестве зависимой переменной – фактические уровни временного ряда y_t (или уровни за

вычетом циклической составляющей, если таковая была обнаружена). Для нелинейных трендов предварительно проводят стандартную процедуру их линеаризации.

Существует несколько способов определения типа тенденции. Чаще всего используют качественный анализ изучаемого процесса, построение и визуальный анализ графика зависимости уровней ряда от времени, расчет некоторых основных показателей динамики. В этих же целях можно использовать и коэффициенты автокорреляции уровней ряда. Тип тенденции можно определить путем сравнения коэффициентов автокорреляции первого порядка, рассчитанных по исходным и преобразованным уровням ряда. Если временной ряд имеет линейную тенденцию, то его соседние уровни yt и yt-1 тесно коррелируют. В этом случае коэффициент автокорреляции первого порядка уровней исходного ряда должен быть высоким. Если временной ряд содержит нелинейную тенденцию, например, в форме экспоненты, то коэффициент автокорреляции первого порядка по логарифмам уровней исходного ряда будет выше, чем соответствующий коэффициент, рассчитанный по уровням ряда. Чем сильнее выражена нелинейная тенденция в изучаемом временном ряде, тем в большей степени будут различаться значения указанных коэффициентов.

Выбор наилучшего уравнения в случае, если ряд содержит нелинейную тенденцию, можно осуществить путем перебора основных форм тренда, расчета по каждому уравнению скорректированного коэффициента детерминации $\overline{R}^2$ и выбора уравнения тренда с максимальным значением этого коэффициента. Реализация этого метода относительно проста при компьютерной обработке данных.

При анализе временных рядов, содержащих сезонные или циклические колебания, наиболее простым подходом является расчет значений сезонной компоненты методом скользящей средней и построение аддитивной или мультипликативной модели временно́го ряда в форме (1) или (2).

Если амплитуда колебаний приблизительно постоянна, строят аддитивную модель (1), в которой значения сезонной компоненты предполагаются постоянными для различных циклов. Если амплитуда сезонных колебаний возрастает или уменьшается, строят мультипликативную модель (2), которая ставит уровни ряда в зависимость от значений сезонной компоненты.

Построение аддитивной модели.

1 шаг. *Выравнивание уровней ряда.* Просуммируем уровни ряда за каждые четыре квартала со сдвигом на один момент времени. Разделив полученные суммы на 4, найдем скользящие средние. Найдем центрированные скользящие средние как средние значения из двух последовательных скользящих средних.

2 шаг. *Расчет сезонной компоненты S.* Найдем разность между уровнями и центрированными скользящими средними. Расчет средней оценки сезонной компоненты для каждого квартала за все годы. Расчет скорректированной сезонной компоненты.Моделирование сезонных колебаний:Аддитивная модель: $Y_t = T_t + S_t + e_t$.

Оценка сезонной компоненты за каждый квартал: $s_t = y_t - \bar{y}_t$. Средняя оценка сезонной компоненты для квартала за все годы: $\bar{S}_t = \frac{\sum s_t}{n}$. Скорректированная сезонная компонента: $S_t = \bar{S}_t - k; k = \frac{\sum_{t=1}^{4} \bar{S}_t}{4}$

3 шаг. *Устранение сезонной компоненты S.*Вычтем скорректированное значение сезонной компоненты из каждого уровня исходного временного ряда. Получим: T+E=Y-S.

4 шаг. *Расчет значений тренда.* Проведем аналитическое выравнивание ряда (T+E) с помощью линейного тренда. Рассчитаем значения T для каждого момента времени по уравнению тренда.

5 шаг. *Расчет значений T+S.* Прибавим к уровням T значения сезонной компоненты (S) для соответствующих кварталов.

6 шаг. *Расчет абсолютной ошибки.* Выполним расчет ошибки для каждого уровня ряда по формуле: E=Y-(T+S). Расчет суммы квадратов абсолютных ошибок и ее сравнение с общей суммой квадратов отклонений уровней ряда.

Построение мультипликативной модели.

1 шаг. *Выравнивание уровней ряда.* Просуммируем уровни ряда за каждые четыре квартала со сдвигом на один момент времени. Разделив полученные суммы на 4, найдем скользящие средние. Найдем центрированные скользящие средние как средние значения из двух последовательных скользящих средних.

2 шаг. *Расчет сезонной компоненты S.* Найдем оценки сезонной компоненты как частное от деления уровней на центрированные скользящие средние. Расчет средней оценки сезонной компоненты для каждого квартала за все годы. Расчет скорректированной сезонной компоненты. Моделирование сезонных колебаний: Мультипликативная модель: $Y_t = T_t \cdot S_t \cdot e_t$.

Оценка сезонной компоненты за каждый квартал: $s_t = \frac{y_t}{\bar{y}_t}$. Средняя оценка сезонной компоненты для квартала за все годы: $\bar{S}_t = \frac{\sum S_t}{n}$. Скорректированная сезонная компонента: $S_t = \bar{S}_t \cdot k ; k = \frac{4}{\sum_{t=1}^{4} \bar{S}_t}$.

3 шаг. *Устранение сезонной компоненты S.*Разделим каждый уровень исходного временного ряда на скорректированное значение сезонной компоненты. Получим: T*E=Y/S.

4 шаг. *Расчет значений тренда.*Проведем аналитическое выравнивание ряда (T*E) с помощью линейного тренда. Рассчитаем значения T для каждого момента времени по уравнению тренда.

5 шаг. *Расчет значений T+S.* Умножим уровни T на значения сезонной компоненты (S) для соответствующих кварталов.

6 шаг. *Расчет абсолютной ошибки.*Выполним расчет ошибки для каждого уровня ряда по формуле: E=Y/(T*S). Расчет суммы квадратов абсолютных ошибок и ее сравнение с общей суммой квадратов отклонений уровней ряда.

Вопросы и задания для самоконтроля

1. В чем особенность временного ряда?
2. Каковы основные компоненты уровней временного ряда?
3. В чем состоит основная задача эконометрического исследования временного ряда?
4. Как определяется автокорреляция остатков во временных рядах?
5. Какие свойства имеет коэффициент автокорреляции?
6. Как определяется автокорреляционная функция?
7. Что такое коррелограмма? Что выявляют при помощи анализа коррелограммы?
8. Как сформулировать вывод о структуре временного ряда?
9. Какие методы применяются для выявления основной тенденции ряда?
10. В чем суть сглаживания временных рядов?

Задание 1. Имеются следующие данные об урожайности пшеницы y за 12 лет:

y_t	16,3	20,2	17,1	9,7	15,3	16,3	19,9	14,4	18,7	20,7	19,5	21,1
t	1	2	3	4	5	6	7	8	9	10	11	12

1) определить среднее значение, среднее квадратическое отклонение и коэффициенты автокорреляции (для лагов $\tau = 1,2$) временного ряда;

2) провести сглаживание исходного временного ряда методом скользящих средних, используя среднюю арифметическую с интервалом сглаживания:

а) $m = 3$;

б) $m = 4$;

3) записать уравнение тренда ряда, полагая, что он линейный, и проверить его значимость на уровне $\alpha = 0,05$.

Задание 2. Данные, отражающие динамику роста доходов y_t на душу населения за восемь лет, приведены в таблице:

Год, t	1	2	3	4	5	6	7	8

y_t	1130	1220	1350	1390	1340	1380	1490	1680

Определить точечный прогноз дохода населения по линейному тренду на 9 год.

Лекция 18

Тема 16. Адаптивные модели временных рядов

Вопросы для изучения

1. Адаптация в моделях временных рядов. Построение адаптивных моделей линейного роста.

2. Адаптивные модели с учетом аддитивных и мультипликативных сезонных составляющих.

3. Процедуры подбора параметров адаптивных моделей временных рядов.

Аннотация. Данная тема раскрывает особенности адаптивных моделей временных рядов.

Ключевые слова. Адаптивная модель, экспоненциальное сглаживание, параметр адаптации.

Методические рекомендации по изучению темы

•Изучить лекционную часть, где даются общие представления по данной теме.

• Для закрепления теоретического материала ответить на вопросы для самоконтроля.

• Для проверки усвоения темы выполнить практические задания и тест для самоконтроля.

Рекомендуемые информационные ресурсы:

1. http://tulpar.kfu.ru/course/view.php?id=2213.

2. Эконометрика: учебник / И. И. Елисеева. – М.: Проспект, 2010. – 288 с. С.200-205.

3.Эконометрика: [Электронный ресурс] Учеб. пособие / А.И. Новиков. - 3-е изд., испр. и доп. - М.: ИНФРА-М, 2014. - 272 с.:

(http://znanium.com/catalog.php?item=booksearch&code=%D1%8D%D0%BA%D0%BE%D0%BD%D0%BE%D0%BC%D0%B5%D1%82%D1%80%D0%B8%D0%BA%D0%B0&page=1#none) С. 96-113.

4. Эконометрика. Практикум: [Электронный ресурс] Учебное пособие / С.А. Бородич. - М.: НИЦ ИНФРА-М; Мн.: Нов. знание, 2014. - 329 с. (http://znanium.com/catalog.php?item=booksearch&code=%D1%8D%D0%BA%D0%BE%D0%BD%D0%BE%D0%BC%D0%B5%D1%82%D1%80%D0%B8%D0%BA%D0%B0&page=4#none) С.276-278.

5. Эконометрика: учебник / под ред. В. С. Мхитаряна. - М.: Проспект, 2008. -384 с. С.297-325.

6. Электронный курс “Time Series Econometrics”, Princeton University, URL: http://sims.princeton.edu/yftp/Times05/;https://blackboard.princeton.edu/webapps/portal/frameset.jsp?tab_group=courses&url=%2Fwebapps%2Fblackboard%2Fexecute%2FcourseMain%3Fcourse_id%3D_52968_1.

Адаптация в моделях временных рядов. Построение адаптивных моделей линейного роста. Развитие аппарата адаптивного прогнозирования экономических процессов в основном осуществлялось по двум направлениям. Первое направление связано с усложнением структуры адаптивных моделей до уровня, обеспечивающего адекватное отражение закономерностей реальных явлений, а второе – с совершенствованием самого адаптивного механизма этих моделей. Развитием простейшей модели $x_t=a_t+\varepsilon_t$ в рамках первого из определенных выше направлений можно считать полином первого порядка

$$x_{t+\tau}=a_{1t}+ a_{2t}\ \tau+ \varepsilon_{t+\tau} \qquad (1)$$

где a_{1t}, a_{2t} – текущие значения коэффициентов модели;

τ - период упреждения;

$\varepsilon_{t+\tau}$ - случайные независимые отклонения расчетных от фактических, имеющие нулевое математическое ожидание и конечную дисперсию σ_2.

Его структура, в отличие от полинома нулевой степени, способна адекватно отражать тенденцию линейного роста исследуемого процесса. Это позволяет избавиться от систематической ошибки, которая имеет место при использовании экспоненциальной средней в качестве прогнозной модели подобных процессов.

Одновременно с изменением структуры модели, как правило, претерпевает соответствующие изменения и ее адаптивный механизм. Неизменным может оставаться только принцип его построения. Причем, для одной и той же модели на основе одного и того же механизма можно строить различные варианты адаптивных механизмов. Примером модели, для которой можно построить различные варианты адаптивных механизмов, как раз и является адаптивный полином первой степени (1). Ее адаптивный механизм предусматривает расчет оценок текущих (т. е. на данный момент времени) коэффициентов модели по двум рекуррентным соотношениям:

$$\begin{aligned} \hat{a}_{1t} &= \alpha_1 x_t + (1-\alpha_1)(\hat{a}_{1t-1} + \hat{a}_{2t-1}), \\ \hat{a}_{2t} &= \alpha_2(\hat{a}_{1t} - \hat{a}_{1t-1}) + (1-\alpha_2)\hat{a}_{2t-1}, \end{aligned} \qquad (2\text{-}3)$$

где α_1,α_2 – параметры экспоненциального сглаживания ($0<\alpha_1, \alpha_2<1$).

Если через $\varepsilon_t = x_t - \hat{x}_t$ обозначить ошибку прогноза, то эти соотношения можно переписать в следующем виде:

$$\begin{aligned} \hat{a}_{1t} &= \hat{a}_{1t-1} + \hat{a}_{2t-1} + \alpha_1\varepsilon_t, \\ \hat{a}_{2t} &= \hat{a}_{2t-1} + \alpha_1\alpha_2\varepsilon_t. \end{aligned} \qquad (4\text{-}5)$$

Полученное представление показывает, что используемая в рекуррентных соотношениях (2-3) процедура экспоненциального сглаживания приводит, как и в случае полинома нулевой степени, к адаптивному механизму, построенному на принципе регулятора с обратной связью.

Адаптивные модели с учетом аддитивных и мультипликативных сезонных составляющих. Для прогнозирования сезонных процессов разработан специальный класс адаптивных моделей, отличительной особенностью которых является наличие в их структуре коэффициентов сезонности. В зависимости от способа включения этого коэффициента различают два типа этих моделей.

К первому типу относятся модели с мультипликативным коэффициентом сезонности:

$$x_t = a_{1t} f_t + \varepsilon_t, \qquad (6)$$

где a_{1t} – изменяющийся во времени коэффициент, динамика которого характеризует тенденцию развития процесса;

f_t, f_{t-1}, …, f_{t-l+1} – коэффициенты сезонности;

l – количество фаз в полном сезонном цикле (при месячных наблюдениях l=12, при квартальных – l=4).

Ко второму типу относятся модели с аддитивным коэффициентом сезонности:

$$x_t = a_{1t} + g_t + \varepsilon_t, \qquad (7)$$

где g_t, g_{t-1}, …, g_{t-l+1} – адаптивные коэффициенты сезонности.

Если моделируемый процесс имеет тенденцию линейного роста, то в моделях (6), (7) член, соответствующий полиному нулевого порядка, заменяется полиномом первого порядка, и тогда модели записываются в следующем виде:

$$\begin{aligned} x_t &= (a_{1t} + \tau a_{2t}) f_t + \varepsilon_t, \\ x_t &= a_{1t} + \tau a_{2t} + g_t + \varepsilon_t. \end{aligned} \qquad (8\text{-}9)$$

Расчет текущих оценок коэффициентов всех этих моделей осуществляется с использованием принципа экспоненциального сглаживания.

Возможно комбинирование различных типов тенденций с коэффициентами сезонности мультипликативного и аддитивного видов. В зависимости от характера динамики моделируемого процесса рекомендуется выбирать одну из девяти моделей, объединенных в три группы. Первая группа включает модели, отражающие:

1) отсутствие закономерностей роста (модель без тренда):

$$x_{t+\tau} = a_{1t} + \varepsilon_{t+\tau}, \qquad (10)$$

2) тенденцию линейного роста (модель с аддитивным линейным трендом)

$$x_{t+\tau} = a_{1t} + \tau a_{2t} + \varepsilon_{t+\tau}, \qquad (11)$$

3) тенденцию экспоненциального роста (модель с мультипликативным трендом)

$$x_{t+\tau} = a_{1t} b_t^{\tau} + \varepsilon_{t+\tau}. \qquad (12)$$

Во второй класс входят модели, получаемые из первого путем включения в их структуру аддитивных коэффициентов сезонности. Это включение трансформирует (10-12) в модели следующего вида:

$$\begin{aligned} x_{t+\tau} &= a_{1t} + g_{t-l+\tau} + \varepsilon_{t+\tau}, \\ x_{t+\tau} &= a_{1t} + \tau a_{2t} + g_{t-l+\tau} + \varepsilon_{t+\tau}, \\ x_{t+\tau} &= a_{1t} b_t^{\tau} + g_{t-l+\tau} + \varepsilon_{t+\tau} \end{aligned} \qquad (13\text{-}15)$$

Третий класс, в отличие от второго, в своей структуре содержит не аддитивный, а мультипликативный коэффициент сезонности:

$$\begin{aligned} x_{t+\tau} &= a_{1t} f_{t-l+\tau} + \varepsilon_{t+\tau}, \\ x_{t+\tau} &= (a_{1t} + \tau a_{2t}) f_{t-l+\tau} + \varepsilon_{t+\tau}, \\ x_{t+\tau} &= a_{1t} f_{t-l+\tau} b_t^{\tau} + \varepsilon_{t+\tau} \end{aligned} \qquad (16\text{-}18)$$

Для каждой из этих моделей оценка параметра a_{1t} осуществляется по формуле

$$\hat{a}_{1t} = \alpha_1 d_1 + (1-\alpha_1) d_2 \qquad (19)$$

где α_1 – параметр сглаживания ($0<\alpha<1$),

$d_1 = x_t$ - для моделей первой группы;

$d_1 = x_t - \hat{g}_{t-1}$ - для моделей второй группы;

$d_1 = x_t / \hat{f}_{t-l}$ - для моделей третьей группы;

$d_2 = \hat{a}_{1t-1}$ - для моделей, не отражающих рост;

$d_2 = \hat{a}_{1t-1} + \hat{a}_{2t-1}$ - для моделей, отражающих тенденцию линейного роста;

$d_2 = \hat{a}_{1t-1}\hat{b}_{t-1}$ - для моделей, отражающих тенденцию экспоненциального роста.

В свою очередь, оценка коэффициентов линейного роста a_{2t} осуществляется с помощью выражения

$$\hat{a}_{2t} = \alpha_2(\hat{a}_{1t} - \hat{a}_{1t-1}) + (1-\alpha_2)\hat{a}_{2t-1}, 0 < \alpha 2 < 1, \qquad (20)$$

а коэффициентов экспоненциального роста b_t по формуле

$$b_t = \alpha_b \frac{\hat{a}_{1t}}{\hat{a}_{1t-1}} + (1-\alpha_b)\hat{b}_{t-1}, 0 < \alpha_b < 1. \qquad (21)$$

Оценки коэффициентов сезонности g_t и f_t рассчитываются по формулам

$$\begin{aligned} \hat{g}_t &= \alpha_g(x_t - \hat{a}_{1t}) + (1-\alpha_g)\hat{g}_{t-l}, 0 < \alpha_g < 1, \\ \hat{f}_t &= \alpha_f \frac{x_t}{\hat{a}_{1t}} + (1-\alpha_f)\hat{f}_{t-1}, 0 < \alpha_f < 1. \end{aligned} \qquad (22)$$

Такое комбинирование позволяет строить адаптивные модели с целенаправленно выбранным набором свойств. Правильно выбранные свойства, гарантируя требуемую адекватность модели, обеспечивают тем самым повышение достоверности прогнозных расчетов. Достигнутое в настоящее время до-

вольно высокое совершенство моделей этого типа практически не расширило область их применения. По преимуществу, она ограничена рамками краткосрочного прогнозирования, так как в моделях не учитывается имеющее место в реальной действительности взаимодействие процессов с другими экономическими явлениями. Изолированное моделирование без учета причин, формирующих динамику и определяющих ее характер, является слабым местом всех методов, ориентированных на прогноз по одному временному ряду. Устранение этого недостатка в рамках моделей данного типа в принципе невозможно, так как требует введения в их структуру элементов, явно учитывающих взаимодействие, т. е. фактического перехода к моделям многофакторного типа с адаптивным механизмом. Но для построения многофакторных моделей расширяется набор необходимых для этого данных, а это не всегда осуществимо. Поэтому в практических ситуациях модели прогнозирования по одному временному ряду часто оказываются единственно приемлемым аппаратом для расчета прогнозных вариантов с позиций информационной обеспеченности и требуемой надежности.

Процедуры подбора параметров адаптивных моделей временных рядов. Основное внимание уделяется выбору такой величины параметра сглаживания α, которая минимизировала бы ошибку предсказания. Выбор величины этого параметра в зависимости от количества наблюдений m, входящих в интервал сглаживания, по формуле: $\alpha=2/(m+1)$ малопригоден для практического использования.

Наиболее часто используемой процедурой подбора оптимальной величины α является метод проб. Общая схема этой процедуры предусматривает деление временного ряда на две части: обучающую и контрольную. Затем по обучающей части при различных α строятся прогнозные модели и делаются расчеты на период, отведенный под контрольную часть. Для каждого α расчетные значения сравниваются с фактическими значениями контрольной части, и определяется среднеквадратическая ошибка прогноза. Оптимальным считается

то α*, для которого эта ошибка оказалась минимальной. Все прогнозные расчеты осуществляются с использованием оптимального значения сглаживающего параметра. В тех случаях, когда оптимальный уровень параметра α с течением времени подтвержден изменениями, эффективность этого подхода снижается, так как оптимум по обучающейся части может не совпадать с оптимумом по всему временному ряду.

Непрерывная настройка параметра α* может выполняться на основе «трекинг-сигнала» *Kt,* определяемого как отношение суммы ошибок прогнозирования $\sum e_i$ к величине их сглаженного абсолютного значения

$$Kt = \frac{\sum_{i=1}^{t} e_i}{\gamma |e_t| + (1-\gamma)\bar{e}_{t-1}}$$

где $\bar{e}_{t-1}$ - сглаженная в предшествующий момент времени величина абсолютного значения прогнозной ошибки e_t;

γ - константа сглаживания абсолютных ошибок прогнозирования.

Решение о перенастройке параметра α принимается в случае, когда значение трекинг-сигнала превысит установленный контрольный уровень.

Еще одной интересной процедурой оптимальной настройки изменяющегося во времени параметра сглаживания является метод адаптивного градиентного экспоненциального сглаживания, так как в его основе лежит применение градиентной оптимизации для поиска оптимальных значений параметра α_t в адаптивных моделях, использующих принцип экспоненциального сглаживания. Метод достаточно универсальный, его можно применять и к полиномам нулевой степени, к моделям линейного роста и к моделям с сезонным эффектом. Суть этого метода заключается в том, что сначала схема расчета прогнозных значений по исходной модели преобразуется к схеме разностного вида, а затем

к преобразованной схеме применяется градиентная процедура локальной оптимизации. Запишем градиент ошибки

$$\nabla = \frac{\partial e_t^2}{\partial \hat{\beta}_t} = 2e_t \frac{\partial e_t}{\partial \beta_t}$$

и введем в рассмотрение функцию чувствительности, представляющую собой производную по β_t=1-α_t от ошибки предсказания $e_t = x_t - \hat{x}_t$

$$S_t = \frac{\partial e_t}{\partial \hat{\beta}_t}$$

Используя градиент и так введенную функцию чувствительности, рекуррентную схему прогнозных расчетов по адаптивной градиентной модели экспоненциального сглаживания можно записать в виде

$$\hat{x}_{t+1} = x_t - \hat{\beta}_t e_t,$$
$$S_{t+1} = \beta_t S_t + e_t,$$
$$\beta_{t+1} = \beta_t - 2\mu e_t S_t.$$

Локальная сходимость гарантируется только в том случае, если μ удовлетворяет неравенству

$$\mu < (1 - \beta_t^2)/2\sigma^2,$$

в котором через σ^2 обозначено математическое ожидание квадрата прогнозной ошибки.

Идея построения моделей с изменяющимся параметром сглаживания актуальна в ситуациях, когда возникает необходимость в разработке прогнозных систем, которые должны использоваться длительное время в автоматическом режиме. Необходимость в подобного рода разработках чаще возникает в технических системах, нежели в экономических, поэтому эта идея при разработке экономических прогнозных моделей дальнейшего развития не получила.

Вопросы и задания для самоконтроля

1. В чем заключаются сущность, механизмы и формы адаптации в социально-экономических системах?

2. В чем заключается специфика экспоненциального сглаживания?

3. В чем состоит особенность модели с мультипликативным коэффициентом сезонности?

4. Какова особенность модели с аддитивным коэффициентом сезонности?

5. Как оценивается коэффициент сезонности для модели, учитывающей тенденцию линейного роста?

6. Какие модели включает группа адаптивных моделей с сезонными составляющими?

7. Какие особенности включает процедура подбора сглаживающего параметра методом проб?

8. В чем заключаются особенности процедуры подбора сглаживающего параметра методом градиентной оптимизации?

Задание 1. Имеются данные о потреблении мороженого, тыс. руб.:

Сезон	Год				
	2008	2009	2010	2011	2012
Зима	253,1	265,5	277,9	290,3	301,3
Весна	331,2	343,6	356,0	368,4	375,4
Лето	364,3	376,7	389,1	401,5	412,4
Осень	292,4	304,8	317,2	343,2	337,5

Постройте адаптивную модель с линейным трендом и аддитивной сезонной компонентой для прогнозирования потребления мороженого.

Лекция 19

Тема 17. Модели стационарных и нестационарных временных рядов

Вопросы для изучения

1. Модели стационарных и нестационарных временных рядов, их идентификация.

2.Модель авторегрессии–скользящего среднего (модель ARMA).

3. Авторегрессионная модель проинтегрированного скользящего среднего (модель ARIMA).

Аннотация. Данная тема раскрывает особенности моделей стационарных и нестационарных временных рядов и методы их оценивания.

Ключевые слова. Стационарный процесс, модель авторегрессии, модель Бокса-Дженкинса.

Методические рекомендации по изучению темы

•Изучить лекционную часть, где даются общие представления по данной теме.

• Для закрепления теоретического материала ознакомиться с решениями типовых задач и ответить на вопросы для самоконтроля.

• Для проверки усвоения темы выполнить практические задания и тест для самоконтроля.

Рекомендуемые информационные ресурсы:

1. http://tulpar.kfu.ru/course/view.php?id=2213.

2. Валентинов, В. А. Эконометрика [Электронный ресурс]: Практикум / В. А. Валентинов. - 3-е изд. - М.: Дашков и К, 2010. - 436 с.

(http://znanium.com/catalog.php?item=booksearch&code=%D1%8D%D0%BA%D0%BE%D0%BD%D0%BE%D0%BC%D0%B5%D1%82%D1%80%D0%B8%D0%BA%D0%B0&page=3#none) С. 328-338.

3.Тихомиров Н. П. Эконометрика: учебник. - М.: Экзамен, серия «Учебник Плехановской академии», 2007, -512 с. С.211-222.

4. Эконометрика: учебник / под ред. В. С. Мхитаряна. - М.: Проспект, 2008. -384 с. С. 325-336.

5. Электронный курс “Time Series Econometrics”, PrincetonUniversity, URL:
http://sims.princeton.edu/yftp/Times05/;https://blackboard.princeton.edu/webap

ps/portal/frameset.jsp?tab_group=courses&url=%2Fwebapps%2Fblackboard%2Fexecute%2FcourseMain%3Fcourse_id%3D_52968_1.

Модели стационарных и нестационарных временных рядов, их идентификация. Набор случайных переменных X(t) называется стохастическим процессом. Стохастический процесс Xt называется стационарным в сильном смысле, если совместное распределение вероятностей всех переменных Xt1, Xt2,…, Xtn, такое же, что и для переменных Xt1+τ, Xt2+τ,…, Xtn+τ. Для стационарного процесса в слабом смысле среднее и дисперсия независимо от рассматриваемого периода времени имеют постоянное значение, а автоковариация зависит только от длины лага между рассматриваемыми переменными.

Временной ряд x1,x2,...,xt, т.е. конкретная реализация стационарного стохастического процесса Xt, также называется стационарным.

Стационарность означает отсутствие:			
тренда	систематических изменений дисперсии	строго периодичных колебаний	систематически изменяющихся взаимозависимостей между элементами временного ряда

Рис. 17.1. Признаки стационарности временного ряда

Линейные модели временных рядов применяются, как правило, для описания стационарных процессов. Чаще всего это стационарные процессы второго порядка, то есть процессы, имеющие постоянные значения всех своих моментов до второго порядка включительно на всех временных отрезках, входящих в интервал $t = 1,2,...,T$. Следовательно, для любых двух интервалов времени (T_1, T_2) и (T_3, T_4) в таком процессе y_t выполняются условия равенства математических ожиданий, дисперсий и коэффициентов автокорреляции одинаковых порядков. На практике для оценок этих показателей должны выполняться соот-

ношения:

$$\bar{y}_1 = \frac{1}{T_2 - T_1}\sum_{t=T_1}^{T_2} y_t = \frac{1}{T_4 - T_3}\sum_{t=T_3}^{T_4} y_t = \bar{y}_2 \qquad (1)$$

$$D_1(y) = \frac{1}{T_2 - T_1}\sum_{t=T_1}^{T_2} (y_t - \bar{y})^2 = \frac{1}{T_4 - T_3}\sum_{t=T_3}^{T_4} (y_t - \bar{y})^2 = D_2(y) \qquad (2)$$

$$r_i^{(1)} = \sum_{t=T_1}^{T_2 - i} \frac{(y_t - \bar{y})(y_{t+i} - \bar{y})}{(T_2 - T_1 - i)\cdot D(y)} = \sum_{t=T_3}^{T_4 - i} \frac{(y_t - \bar{y})(y_{t+i} - \bar{y})}{(T_4 - T_3 - i)\cdot D(y)} = r_i^{(2)}, i = 1,2,... \qquad (3)$$

Здесь $\bar{y}^{(1)}$, $\bar{y}^{(2)}$, $D^{(1)}(y)$, $D^2(y)$, $r_i^{(1)}$, $r_i^{(2)}$ - оценки математических ожиданий, дисперсий и коэффициентов автокорреляции i - го порядка процесса yt на первом и втором интервалах; $\bar{y}$ и $D(y)$ - оценки среднего значения и дисперсии процесса на интервале $(1,T)$.

Равенства (1) – (3) понимаются в статистическом смысле. Это значит, что каждое из этих равенств может в точности не выполняться, однако разница между оценками укладывается в границы соответствующего критерия.

Такие критерии реализуются через различные тесты, которые мы сейчас рассмотрим.

Параметрические тесты стационарности применяются при достаточно строгих предположениях о законе распределения временного ряда и его параметрах. Они оценивают степень близости эмпирических характеристик временного ряда к их теоретическим аналогам. Для выражений (1) – (3) параметрическими критериями стационарности являются критерии Стьюдента и Фишера. Здесь предполагается нормальный закон распределения значений временного ряда и его выборочных характеристик, что справедливо для многих реальных процессов.

Тестирование математического ожидания по статистике Стьюдента требует разбить временной ряд (1,T) на две части, не обязательно одинаковые, H_0 – гипотеза о постоянстве математического ожидания:

$$\tau = \frac{|\bar{y}_1 - \bar{y}_2|}{\sqrt{\frac{s_1^2}{T_1} + \sqrt{\frac{s_2^2}{T_2}}}}, \sigma_1^2 \neq \sigma_2^2$$

$$\tau = \frac{|\bar{y}_1 - \bar{y}_2|}{s^2} \cdot \sqrt{\frac{T_1 \cdot T_2}{T_1 + T_2}}, \sigma_1^2 = \sigma_2^2 = \sigma$$

$$\tau < \tau(p, v = T_1 + T - 2) \Rightarrow H_0$$

Тестирование математического ожидания по статистике Фишера (если количество наблюдений достаточно велико), H_0 – гипотеза о постоянстве математического ожидания временного ряда. Интервал наблюдений делится на несколько частей.

$$F = \frac{\frac{1}{n-1} \cdot \sum_{j=1}^{n} T_j (\bar{y}_j - \bar{y})^2}{\bar{s}^2(n)}$$

$$\bar{s}^2(n) = \frac{1}{T-n} \cdot \sum_{j=1}^{n} (T_j - 1) \cdot \bar{s}_j^2$$

$$F < F(p, v_1 = n-1, v_2 = T_1 + T_2 + \ldots + T_n - n \Rightarrow H_0$$

где, n – число частей разбиения интервала (1,T); T_j- число измерений переменной y_t на j-ой части; j=1,2,…,n;

$\bar{y}$ - среднее значение временного ряда;

$\bar{s}^2(n)$ - средняя дисперсия.

Тестирование дисперсии временного ряда на постоянство ее значения проводится разбиением исходного интервала на две части с использованием двухстороннего критерия Фишера, который рассчитывается по формуле:

$$F = \frac{s_1^2}{s_2^2},$$

где s_1^2 и s_2^2 - оценки дисперсии ряда на первой и второй подвыборок соответственно и числом измерений Т1 и Т2.

Этот тест аналогичен проверке гипотезы о равенстве дисперсий двух нормальных CB с той разницей, что здесь сравнивается дисперсия в разных частях одного временного ряда.

Рассчитанное значение F - статистики сравнивается с критическими на

уровнях α/2 и 1-α/2 и числами степеней свободы (T1-1) и (T2-1). Если $F_{кр.}(1-\alpha/2;T_1-1;T_2-1) \le F \le F_{кр.}(\alpha/2;T_1-1;T_2-1)$, то гипотеза $\sigma_1^2 = \sigma_2^2 = \sigma^2$ принимается на уровне α.

Поскольку критические значения удовлетворяют соотношению

$$F(\alpha/2;v_1;v_2) = \frac{1}{F(1-\alpha/2;v_1;v_2)},$$

на практике проверяется только соотношение $F \le F_{кр.}(\alpha/2;T_1-1;T_2-1)$

при условии, что $s_1^2 \ge s_2^2$.

При достаточно больших объемах наблюдений временного ряда $(T \ge 40)$ вместо критерия Фишера рекомендуют использовать стандартизированное нормальное распределение. При выборках $40 \le T \le 100$ закону $N(0,1)$ подчиняется случайная величина

$$\Phi = \frac{\frac{1}{2}\ln\frac{s_1^2}{s_2^2} + \frac{1}{2}\left(\frac{1}{v_1} - \frac{1}{v_2}\right)}{\sqrt{\frac{1}{2}\left(\frac{1}{v_1} + \frac{1}{v_2}\right)}}$$

При больших выборках расчетное значение случайной величины определяется так:

$$\Phi = (s_1 - s_2)\sqrt{\frac{s_1^2}{2T_1} + \frac{s_2^2}{2T_2}}$$

В любом случае, если $\Phi < \Phi_{\alpha/2}$, то гипотеза о постоянстве дисперсии принимается.

Если временной ряд разбивается на большее число частей (n>2), гипотеза о постоянстве дисперсии может быть проверена критерием Кокрена, основанном на распределении Фишера. При этом обычно объемы этих частей принимаются равными между собой, то есть T1=T2=…=Tn. Рассчитывается критерий по формуле:

$$K = \frac{s_{\max}^2}{s_1^2 + ... + s_n^2}$$

Здесь $s_{\max}^2 = \max_j \left(s_j^2\right), \quad j = \overline{1,n}$.

Табличное значение критерия Кокрена определяется по формуле:

$$K(\alpha; n; N-1) = \frac{F(1-\alpha/n; N-1; (n-1)(N-1))}{(n-1) + F(1-\alpha/n; N-1; (n-1)(N-1))}$$

Если $K < K(\alpha; n; N-1)$, то гипотеза о постоянстве дисперсии временного ряда принимается на уровне α.

Критерий Бартлетта также используется при проверке гипотезы о постоянстве дисперсии. Он является более мощным, чем критерий Кокрена, но и более чувствительным к отклонениям значений временного ряда от нормального закона. Здесь временной ряд также разбивается на несколько частей, причем не обязательно одинаковых по величине.

Согласно критерию, следующая величина:

$$\lambda = -\frac{1}{c}\sum_{i=1}^{n}(T_i - 1)\ln\frac{s_i^2}{s^2}$$

распределена приблизительно по закону λ^2 с $(n-1)$ степенями свободы. Здесь

$$s^2 = \frac{\sum_{i=1}^{n}(T_i - 1)s_i^2}{\sum_{i=1}^{n}(T_i - 1)}$$

-

средняя дисперсия на n интервалах;

$$c = 1 + \frac{1}{3(n-1)} \cdot \sum_{i=1}^{n}\frac{1}{T_i - 1} - \frac{1}{\sum_{i=1}^{n}(T_i - 1)}$$

При больших Ti $c \approx 1$.

Для одинаковых размеров подвыборок $v_1 = v_2 = ... = v_n = v$, тогда $\sum_{i=1}^{n}(T_i - 1) = T - n$, поэтому $\lambda = \frac{1}{c} n v \left(\ln \bar{s}^2 - \frac{1}{n}\sum_{i=1}^{n} s_i^2 \right)$, где $c = 1 + \frac{n+1}{3 \cdot n \cdot v}$.

Если $\lambda < \chi^2(\alpha, n-1)$, то гипотеза о равенстве дисперсий на рассматриваемых частях временного интервала принимается.

По рассмотренным параметрическим критериям следует отметить их ограниченность в применении вследствие достаточно жестких предположений нормальности закона распределения временного ряда. Кроме того, они требуют значительных вычислений. Однако реальные временные ряды могут быть распределены по закону, отличному от нормального. Поэтому на практике при проверке стационарности процессов часто используют непараметрические критерии, не имеющие ограничений по закону распределения и не столь сложные по вычислениям.

Тест Манна – Уитни используется вместо критерия Стьюдента для проверки идентичности распределений двух совокупностей, то есть временных последовательностей одного временного ряда, определенных на разных временных частях интервала 1,2,…,T. Он тестирует постоянство математического ожидания.

Пусть первая подвыборка образована T1 последовательными значениями yt, а вторая – T2 его последовательными значениями, и эти последовательности не пересекаются.

Обозначим элементы первой подвыборки символом $y^{(1)}$, второй – символом $y^{(2)}$. Затем объединим эти подвыборки в одну совокупность объемом (T1+T2), расположив все элементы в порядке возрастания их значений. При этом элементы подвыборок оказываются перемешанными между собой.

Если ряд стационарный, то элементы разных подвыборок довольно равномерно перемешаны друг с другом. В противном случае общая последовательность оказывается разделенной на массивы, состоящие в основном из единиц одной из совокупностей. Например, для возрастающего или убывающего временного ряда элементы подвыборок скапливаются на разных концах общей последовательности.

В тесте Манна – Уитни проверяется гипотеза о стационарности времен-

ного ряда на основе критерия u^*, равного числу случаев, когда элементы из первой подвыборки предшествуют элементам из второй подвыборки. Значение u^* рассчитывается по формулам: $u^* = R_1 - \frac{T_1(T_1+1)}{2}$ или

$$u^* = T_1 \cdot T_2 - \frac{T_1(T_1+1)}{2} - R_2,$$

где R_1 и R_2 - суммы рангов элементов первой и второй подвыборок соответственно, определяемых по их общей последовательности.

Для достаточно больших последовательностей (T>50) случайная величина u^* распределена по нормальному закону с математическим ожиданием $M[u^*] \approx \frac{T_1 \cdot T_2}{2}$ и дисперсией $D[u^*] \approx \frac{T_1 \cdot T_2 \cdot (T_1 + T_2 + 1)}{12}$.

Таким образом, случайная величина $z = \frac{u^* - \frac{T_1 \cdot T_2}{2} \pm \frac{1}{2}}{\sigma(u^*)}$ распределена по закону $N(0,1)$. Поправка 1/2 вводится для обеспечения непрерывности z. Она прибавляется, если z<0. Она прибавляется, если z<0, и вычитается, если z>0

Если обе подвыборки идентичны, их элементы перемешаны между собой, тогда значение u^* будет находиться вблизи своего среднего значения, а значение z - около нуля. Поэтому гипотеза о стационарности процесса y_t принимается на уровне значимости α, если выполняется неравенство:

$$u_{1-\alpha/2} \le z \le u_{\alpha/2}$$

Непараметрический тест Сигела – Тьюки используют вместо параметрического критерия Фишера для проверки гипотезы о постоянстве дисперсии временного ряда. Он также основан на сопоставлении рангов элементов двух подвыборок из данного интервала.

Сначала исходный временный ряд центрируется, то есть каждое значение заменяется отклонением от среднего согласно выражению $y_{\text{å}} = y_t - \bar{y}$, где $\bar{y}$ -

среднее значение ряда yt.

Далее интервал (1,T) делится на две, желательно равные, части, где элементы обозначаются соответственно $y^{(1)}$ и $y^{(2)}$. Эти элементы в объединенной совокупности сортируются в порядке возрастания их значений. Затем каждому значению присваивается его ранг по следующему правилу: все нечетные номера получают отрицательные элементы в порядке возрастания их значений, а все четные номера – положительные элементы, но в порядке убывания их значений. Другими словами, ранг 1 получает наименьшее отрицательное значение, а ранг 2 – наибольшее положительное.

Если обозначить R_1 сумму рангов элементов первой подвыборки, то случайная величина

$$z = \frac{R_1 - \frac{T_1 \cdot (T_1 + T_2 + 1)}{2} \pm \frac{1}{2}}{\sqrt{\frac{T_1 \cdot T_2 \cdot (T_1 + T_2 + 1)}{12}}}$$

оказывается распределенной по закону $N(0,1)$. Здесь также поправка 1/2 вводится для обеспечения непрерывности z.

Отдельную группу непараметрических тестов стационарности составляют тесты, основанные на так называемых сериальных критериях. Они анализируют закономерности серий измеренных значений временного ряда. Для их применения необходим достаточно большой объем данных, чтобы считать обнаруженные закономерности устойчивыми.

Серией называют последовательность значений временного ряда, отклоняющихся от значения некоторого признака в одну и ту же сторону. Например, при тестировании автокорреляции в остатках по методу рядов таким признаком было расчетное значение результативного признака. Во временных рядах в роли этого признака часто выступает медиана значений ряда. Тогда элементы, по значению превышающие медиану, образуют серии с положительным знаком, а элементы, не превосходящие по значению медиану – серии с отрицательным знаком.

Критерий Вальда – Вольфовица основан на подсчете общего числа серий. Среднее число серий рассчитывается по выражению:

$$M[N_s]=\frac{2N_1N_2}{N_1+N_2}+1$$

, а дисперсия – по формуле:

$$D[N_s]=\frac{2N_1N_2\cdot[2N_1\cdot N_2-(N_1+N_2)]}{(N_1+N_2)^2\cdot(N_1+N_2-1)}$$

Здесь N1 и N2 - количества элементов соответственно с положительным и с отрицательным знаком; (N1+N2)=T; N_s - число серий.

Как видим, эти формулы в точности повторяют формулы метода рядов.

При большом объеме временного ряда случайная величина

$$z=\frac{N_s-M[N_s]\pm\frac{1}{2}}{\sigma(N_s)}$$

распределена по закону $N(0;1)$.

Если реальный временной ряд не представляет стационарный процесс второго порядка, его нужно привести к стационарному процессу. Это делается с помощью соответствующих преобразований: взятия конечных разностей, логарифмирования цепных индексов, расчета темпов прироста и др.

Например, когда закон изменения y_t близок к линейному, преобразование заключается во взятии первых разностей:

$$y'_t=\Delta y_t=y_t-y_{t-1}$$

Разности второго порядка:

$$y''_t=\Delta y'_t=y'_t-y'_{t-1}=(y_t-y_{t-1})-(y_{t-1}-y_{t-2})=y_t-2y_{t-1}+y_{t-2}$$

применяются при законе изменения y_t, близком к квадратической параболе.

При экспоненциальном росте y_t логарифмируются цепные индексы:

$$y'_t=\ln\frac{y_t}{y_{t-1}}=\ln y_t-\ln y_{t-1}$$

Расчет темпов прироста выполняется по формуле

$$y'_t = \frac{y_t - y_{t-1}}{y_{t-1}} = \frac{y_t}{y_{t-1}} - 1$$

Для трансформации исходного нестационарного ряда в стационарный можно использовать и другие преобразования. В каждом конкретном случае надо исходить из примерной формы временного графика y_t. Подходящее преобразование должно обеспечивать приблизительное выполнение условия $y'_t = f(y_t) = const$.

Особенности конкретного стационарного процесса второго порядка полностью определяются характером его автокорреляционной функции, представляющей собой последовательность коэффициентов автокорреляции r_0, r_1, r_2, … Здесь $r_0 = 1$, остальные значения располагаются на отрезке $[-1;1]$.

Аналогично формируется автокорреляционная функция как последовательность значений автокорреляций γ_0, γ_1, γ_2,… в зависимости от сдвига. Между значениями двух функций существует взаимосвязь $\gamma_i = r_i\sigma_y^2$, i=0,1,2,…; $\gamma_0 = \sigma_y^2$.

Модель авторегрессии–скользящего среднего (модель ARMA). Построение модели AP(k) сводится к решению двух задач:

- определение рационального порядка модели (величины k);
- оценивание параметров модели на основе уравнений Юла-Уокера.

$$y_t = \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + \ldots + \alpha_k y_{t-k} + \varepsilon_t$$

Система уравнений Юла-Уокера:

$$r_1 = a_1 + a_2 r_1 + \ldots + a_k r_{k-1};$$
$$r_2 = a_1 r_1 + a_2 + \ldots + a_k r_{k-2};$$
$$\ldots\ldots\ldots$$
$$r_k = a_1 r_{k-1} + a_2 r_{k-2} + \ldots + a_k;$$

$r_1, r_2, \ldots r_k$ – известные оценки коэффициентов автокорреляции;

$a_1, a_2, \ldots a_k$ - неизвестные оценки коэффициентов модели.

Модель авторегрессии первого порядка AP(1):

$$y_t = \alpha_1 y_{t-1} + \varepsilon_t,$$
$$a_1 = r_1$$

Модель авторегрессии второго порядка AP(2):

$$y_t = \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + \varepsilon_t,$$
$$r_1 = a_1 + a_2 r_1;$$
$$r_2 = a_1 r_1 + a_2.$$
$$a_1 = \frac{r_1(1-r_2)}{1-r_1^2};$$
$$a_2 = \frac{r_2 - r_1^2}{1-r_1^2}$$

Модель скользящего среднего первого порядка CC(1):

$$y_t = \varepsilon_t - \beta_1 \varepsilon_{t-1},$$
$$\sigma_y^2 = (1+\beta_1^2)\sigma_\varepsilon^2,$$
$$\rho_1 = \frac{-\beta_1}{1+\beta_1^2}$$

Модель скользящего среднего второго порядка CC(2):

$$y_t = \varepsilon_t - \beta_1 \varepsilon_{t-1} - \beta_2 \varepsilon_{t-2},$$
$$\sigma_y^2 = (1+\beta_1^2+\beta_2^2)\sigma_\varepsilon^2,$$
$$\rho_1 = \frac{-\beta_1(1-\beta_1)}{1+\beta_1^2+\beta_2^2}; \rho_2 = \frac{-\beta_2}{1+\beta_1^2+\beta_2^2}; \rho_i = 0, i \geq 3.$$

Простейшая модель авторегрессии - скользящего среднего APCC(k,m) - **(A**uto**R**egressive-**M**oving**A**verage (ARMA (k,m)) :

$$y_t = \alpha y_{t-1} + \varepsilon_t - \beta \varepsilon_{t-1}$$
$$y_t - \alpha y_{t-1} = \varepsilon_t - \beta \varepsilon_{t-1}, |\alpha| < 1, |\beta| < 1$$

Значения автокорреляционной функции для ARMA (1,1) будут иметь вид:

$$\rho(1) = \frac{(1-\alpha\beta)(\alpha-\beta)}{1+\beta^2-2\alpha\beta}, \tau = 1$$
$$\rho(\tau) = \alpha\rho(\tau-1) = \alpha^{\tau-1}\rho(1), \tau > 1$$

Авторегрессионная модель проинтегрированного скользящего среднего (модель ARIMA).Для описания нестационарных однородных временных рядов применяется модель Бокса-Дженкинса (ARIMA –модель). Наиболее распространены ARIMA (k,m,q) – модели, со значениями параметров, не превышающими 2, q – порядок разности (дискретной производной).

Этапы методологии Бокса-Дженкинса:

1. Тестирование исходного ряда на стационарность. Анализ автокорреляционной функции. Переход к стационарному ряду путем взятия последовательных разностей (дискретные производные). Определение параметра q.

2. Исследование характера автокорреляционной функции и предположение о значениях параметров k (порядок авторегрессии) и m (порядок скользящего среднего).

3. Оценивание параметров ARIMA (k,m,q) – модели.

4. Проверка пробной модели на адекватность путем анализа ряда остатков.

Для обнаружения «белого шума» в остатках применяют Q-статистику Бокса-Пирса, H_0 об отсутствии автокорреляции в остатках:

$$Q = n\sum_{p=1}^{\tau} r_p^2,$$

$$Q < \chi^2(\alpha, v = \tau - k - m) \Rightarrow H_0 : \rho = 0$$

Критерии качества подгонки модели Бокса-Дженкинса:

КритерийАкайка (Akaike information criterion, AIC):

$$AIC = \frac{k+m}{n} + \ln\left(\frac{\sum_{t=1}^{n} e_t^2}{n}\right)$$

Выбор следует сделать в пользу модели с меньшим значением AIC.

Критерий Шварца (Swarzcriterion):

$$SIK = \frac{(p+q)\ln n}{n} + \ln\left(\frac{\sum_{t=1}^{n} e_t^2}{n}\right)$$

Вопросы для самоконтроля

1. Какая модель временного ряда называется статической?
2. Когда модель временного ряда называется динамической?
3. Как определяются авторегрессионные модели?
4. Как определяется модель ARMA?
5. Как интерпретируют параметры моделей авторегрессии?
6. Что означает стационарность временного ряда?
7. Какой стационарный процесс называется «белым шумом»?
8. Какие типы включают модели стационарных временных рядов?
9. Какие типы включают модели нестационарных временных рядов?
10. Как определяется ARIMA-модель?

Лекция 20

Тема 18. Модели с лаговыми переменными

Вопросы для изучения

1. Статические и динамические модели.

2. Модели с распределенным лагом.

3. Модель частичной корректировки и модель адаптивных ожиданий.

Аннотация. Данная тема раскрывает особенности моделей с лаговыми переменными и методы их оценивания.

Ключевые слова. Модель с распределенным лагом, метод Алмон, метод Койка, модель частичной корректировки, модель адаптивных ожиданий.

Методические рекомендации по изучению темы

•Изучить лекционную часть, где даются общие представления по данной теме.

• Для закрепления теоретического материала ответить на вопросы для самоконтроля.

• Для проверки усвоения темы выполнить практические задания и тест для самоконтроля.

•Для подготовки к экзамену выполнить итоговый тест и итоговые практические задания.

Рекомендуемые информационные ресурсы:

1. http://tulpar.kfu.ru/course/view.php?id=2213.

2. Эконометрика: [Электронный ресурс] Учеб. пособие / А.И. Новиков. - 3-е изд., испр. и доп. - М.: ИНФРА-М, 2014. - 272 с.: (http://znanium.com/catalog.php?item=booksearch&code=%D1%8D%D0%BA%D0%BE%D0%BD%D0%BE%D0%BC%D0%B5%D1%82%D1%80%D0%B8%D0%BA%D0%B0&page=1#none) С. 138-146.

3. Эконометрика. Практикум: [Электронный ресурс] Учебное пособие / С.А. Бородич. - М.: НИЦ ИНФРА-М; Мн.: Нов. знание, 2014. - 329 с. (http://znanium.com/catalog.php?item=booksearch&code=%D1%8D%D0%BA%D0%BE%D0%BD%D0%BE%D0%BC%D0%B5%D1%82%D1%80%D0%B8%D0%BA%D0%B0&page=4#none) С.273-276.

4. Электронный курс “Time Series Econometrics”, Princeton University, URL: http://sims.princeton.edu/yftp/Times05/;https://blackboard.princeton.edu/webapps/portal/frameset.jsp?tab_group=courses&url=%2Fwebapps%2Fblackboard%2Fexecute%2FcourseMain%3Fcourse_id%3D_52968_1.

Статические и динамические модели. Эконометрическая модель, построенная по данным временного ряда, является статической, если она не содержит лаговые значения экзогенных и (или) эндогенных переменных. Эконометрическая модель является динамической, если в данный момент времени t она учитывает значения входящих в нее переменных, относящиеся как к текущему, так и к предыдущим моментам времени, то есть отражает динамику переменных в каждый момент времени. Различают два типа динамических моделей.

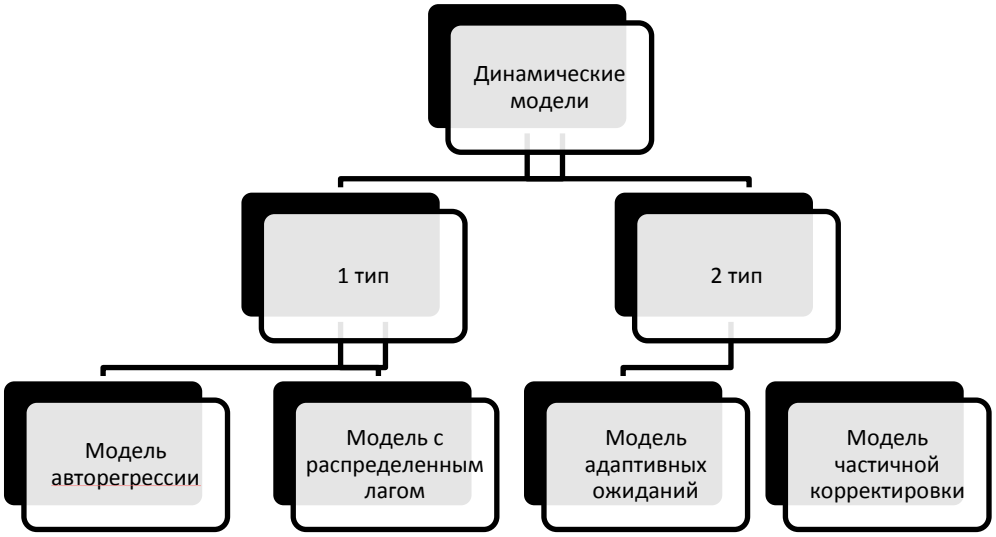


Рис. 18.1. Типы динамических моделей

В моделях первого типа значения переменной за прошлые периоды времени (лаговые переменные) непосредственно включены в модель. В моделях второго типа динамическая информация учитывается в неявном виде. Модели включают переменные, отражающие ожидаемый или желаемый уровень результата, или один из факторов в момент времени ***t***. Этот уровень является неизвестным и определяется на основании той информации, которая имеется в наличии на предшествующий момент времени ***t-1.***

Модели с распределенным лагом. Переменные, влияние которых характеризуется определенным запаздыванием, называются лаговыми переменными. Классифицируются динамические модели по-разному. Один из вариантов классификации следующий. Модели с распределенными лагами, которые содержат в качестве лаговых переменных лишь независимые (объясняющие) переменные, например:

$$y = a + b_0 x_t + b_1 x_{t-1} + \ldots + b_p x_{t-p} + \varepsilon_t . \qquad (1)$$

Авторегрессионные модели, уравнения которых включают в качестве объясняющих переменных лаговые значения зависимых переменных, например:

$$y = a + b x_t + c_1 y_{t-1} + c_2 y_{t-2} + \varepsilon_t . \qquad (2)$$

Рассмотрим модель (1), приняв, что p – конечное число. Модель говорит о том, что, если в некоторый момент времени t происходит изменение x, это изменение будет влиять на значение y в течение p последующих моментов времени. Коэффициент b0 называется краткосрочным мультипликатором, т.к. он характеризует изменение среднего значения y при единичном изменении x в тот же самый момент времени. Сумма $\sum_{j=0}^{p} b_j$ называется долгосрочным мультипликатором; он характеризует изменение y под воздействием единичного изменения x в каждом из моментов времени. Любая сумма $\sum_{j=0}^{k} b_j \ (k < p)$ называется промежуточным мультипликатором.

Относительные коэффициенты модели (1) с распределенным лагом определяются выражениями:

$$\beta_j = \frac{b_j}{\sum_{j=0}^{p} b_j}; \quad \sum_{j=0}^{p} \beta_j = 1 \qquad (3)$$

(условие нормировки имеет место, только если все bj имеют одинаковые знаки). Значения βj являются весами для соответствующих коэффициентов bj. Каждый из них измеряет долю общего изменения y, приходящегося на момент (t+j).

Средний лаг определяется по формуле средней арифметической взвешенной:

$$\bar{l} = \sum_{j=0}^{p} j \cdot \beta_j \ . \qquad (4)$$

Он означает период, в течение которого происходит изменение результата от изменения x в момент t. Небольшая величина (4) означает быструю реак-

цию y на изменение x, высокое значение говорит о том, что воздействие фактора y будет сказываться в течение длительного времени.

Медианный лаг – это величина лага, для которого

$$\sum_{j=0}^{l_{Me}} \beta_j \approx 0{,}5 \quad . \qquad (5)$$

Это время, в течение которого с момента t будет реализована половина общего воздействия фактора на результат.

Рассмотрим условный пример. Предположим, модель зависимости объѐмов продаж компании от расходов на рекламу имеет вид:

$$\hat{y}_t = -0{,}67 + 4{,}5x_t + 3x_{t-1} + 1{,}5x_{t-2} + 0{,}5x_{t-3} \, .$$

Краткосрочный мультипликатор равен 4,5: увеличение расходов на рекламу на 1 млн. руб. приводит к среднему росту продаж компании на 4,5 млн. руб. в том же периоде.

В момент (t+1) такой рост составит 4,5+3,0=7,5 млн. руб., в момент (t+2) - 7,5+1,5=9 млн. руб. и т.д. долгосрочный мультипликатор равен 9,5. В долгосрочной перспективе (в течение 3 месяцев) увеличение расходов на 1 млн. руб. приведет к общему росту продаж на 9,5 млн. руб.

Относительные коэффициенты:

$$\beta_0 = \frac{4{,}5}{9{,}5} = 0{,}474; \quad \beta_1 = \frac{3}{9{,}5} = 0{,}316; \quad \beta_2 = \frac{1{,}5}{9{,}5} = 0{,}158; \quad \beta_3 = \frac{0{,}5}{9{,}5} = 0{,}053 \, .$$

- 47,4% общего увеличения объѐма продаж от роста затрат на рекламу происходит в текущем месяце, 31,6% - в следующем месяце и т.д.

Средний лаг равен:

$$\bar{l} = 0 \cdot 0{,}474 + 1 \cdot 0{,}316 + 2 \cdot 0{,}158 + 3 \cdot 0{,}053 = 0{,}791 \text{ (мес.)}$$

- небольшая величина, поскольку большая часть эффекта роста затрат на рекламу проявляется сразу же. Медианный лаг в данном примере составляет чуть более 1 месяца.

Модель (1) можно свести к уравнению множественной регрессии через замены переменных:

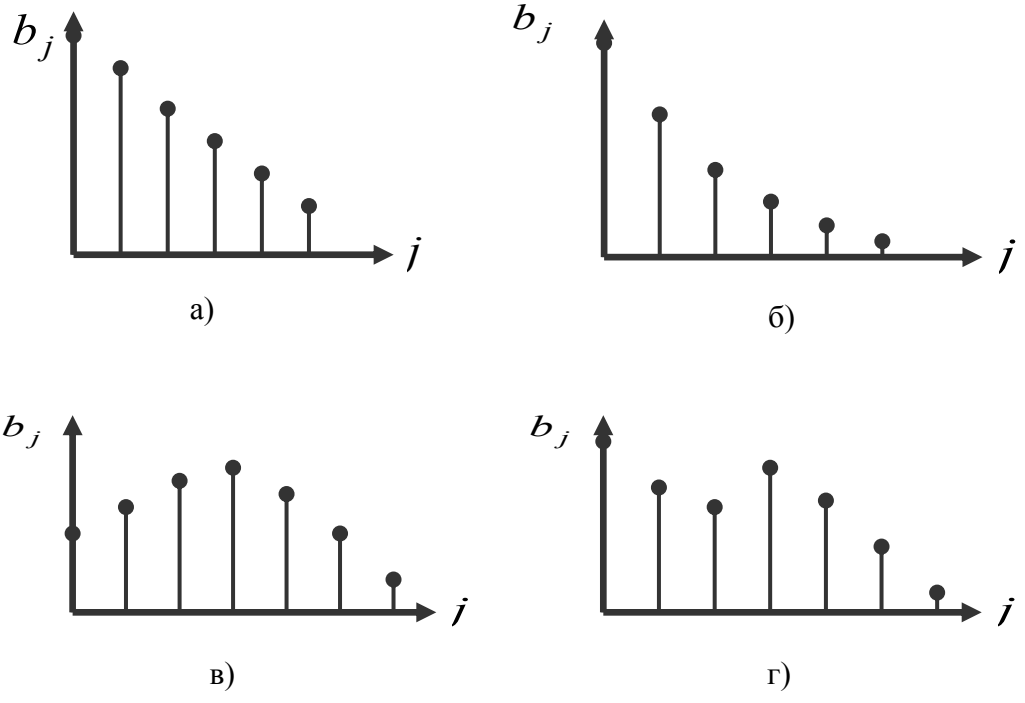
$$x_0^* = x_t, \quad x_1^* = x_{t-1}, \ldots, x_p^* = x_{t-p}, \tag{6}$$

в результате получаем:

$$y = a + b_0 x_1^* + b_1 x_2^* + \ldots + b_p x_p^* + \varepsilon_t. \tag{7}$$

Однако применение обычного МНК затруднительно по следующим причинам:

Текущие и лаговые значения x тесно связаны между собой, что приводит к высокой мультиколлинеарности факторов. При большой величине лага велико число параметров, что приводит к уменьшению числа степеней свободы. Часто возникает проблема автокорреляции остатков. Поэтому оценки параметров становятся неточными и неэффективными. Для получения более обоснованных оценок нужна информация о структуре лага. Эта структура может быть различной. На следующем рисунке представлены некоторые еѐ формы:



Если с ростом величины лага коэффициенты при лаговых переменных убывают, то имеет место линейная (или треугольная) структура лага (а), а так-

же геометрическая структура (б). Возможны и другие структуры лага (в или г). Рассмотрим некоторые подходы к расчету лагов.

Лаги Алмон. Предполагается, что в модели (1) с конечной максимальной величиной лага p значения коэффициентов bj описываются полиномом к – й степени:

$$b_j = c_0 + c_1 j + c_2 j^2 + ... + c_k j^k . \quad (8)$$

Каждый коэффициент, таким образом, запишется так:

$$b_0 = c_0,$$
$$b_1 = c_0 + c_1 + ... + c_k,$$
$$b_2 = c_0 + 2c_1 + ... + 2^k c_k,$$
$$\dots$$
$$b_p = c_0 + p \cdot c_1 + ... + p^k c_k . \quad (9)$$

Подставим эти соотношения в (1) и перегруппируем слагаемые, получим:

$$y_t = a + c_0 \sum_{j=0}^{p} x_{t-j} + c_1 \sum_{j=0}^{p} j \cdot x_{t-j} + c_2 \sum_{j=0}^{p} j^2 \cdot x_{t-j} + ... + c_k \sum_{j=0}^{p} j^k \cdot x_{t-j} + \varepsilon_t \quad (10)$$

Обозначим суммы соответственно как новые переменные $z_0, z_1, ..., z_k$, перепишем (10) в виде:

$$y_t = a + c_0 z_0 + c_1 z_1 + c_2 z_2 + ... + c_k z_k + \varepsilon_t . \quad (11)$$

Параметры cj определяются по МНК.

Достоинства метода:

Универсальность, применимость для моделирования процессов с разнообразными структурами лагов.

При малых k (2 или 3) можно построить модели с распределенным лагом любой длины.

Ограничения метода:

Величина p должна быть известна заранее. При этом приходится задавать максимально возможную величину лага. Выбор меньшего лага, чем его реаль-

ное значение, приведет к неверной спецификации модели, невозможности обеспечить случайность остатков, поскольку влияние значимых факторов будет выражено в остатках. Оценки параметров при этом окажутся неэффективными и смещенными. Включение в модель большей величины лага, чем его реальное значение, снижает эффективность оценок из-за наличия статистически незначимых факторов.

Необходимость установить степень полинома. Обычно принимают k=2 или 3 по правилу: степень полинома k должна быть на единицу больше числа экстремумов в структуре лага. В крайнем случае k определяется из сравнения моделей для различных k.

Возможна мультиколлинеарность факторов zj, однако она сказывается здесь в меньшей степени, чем в модели (1).

Метод Койка. Этот метод применяется в модели с бесконечным лагом:

$$y_t = a + b_0 x_t + b_1 x_{t-1} + b_2 x_{t-2} + ... + \varepsilon_t \,. \qquad (12)$$

Здесь обычный МНК применить нельзя. Для идентификации модели (12) предполагается, что параметры с увеличением лага убывают в геометрической прогрессии, т.е. с постоянным темпом $\lambda \in (0,1)$:

$$y_t = a + b_0 x_t + b_0 \lambda x_{t-1} + b_0 \lambda^2 x_{t-2} + ... + \beta_0 \lambda^m x_{t-m} + ... + \varepsilon_t \,. \qquad (13)$$

Запишем выражение (13) для момента (t-1):

$$y_{t-1} = a + b_0 x_{t-1} + b_0 \lambda x_{t-2} + b_0 \lambda^2 x_{t-3} + ... + \varepsilon_t \,. \qquad (14)$$

Умножим (14) на λ и вычтем из (13):

$$y_t = a(1-\lambda) + b_0 x_t + \lambda y_{t-1} + \varepsilon_t - \lambda \varepsilon_{t-1}$$

или

$$y_t = a(1-\lambda) + b_0 x_t + \lambda y_{t-1} + u_t \,. \qquad (15)$$

Это модель авторегрессии. Определив еѐ параметры, находим λ, a, b_0 исходной модели, а затем и параметры $b_j = \lambda^j b_0,\ j = 1,2,3,...$. Данная модель

позволяет определить долгосрочный мультипликатор $\sum_{j=0}^{\infty} b_j = b_0 \frac{1}{1-\lambda}$ и средний лаг $\bar{l} = \frac{\lambda}{1-\lambda}$.

Модель частичной корректировки и модель адаптивных ожиданий. Моделирование ожиданий является сложной задачей, поскольку фактор ожидания имеет качественную специфику. Например, инвестиции связаны не только с нормой процента, но и с ожиданиями инвесторов. Если в стране существенная безработица, то действия правительства в направлении стимулирования могут рассматриваться как позитивные, и это способствует инвестициям. Если экономика близка к полной занятости, то та же самая политика будет рассматриваться как ведущая к росту инфляции и приведет к падению инвестиционной активности.

Модель адаптивных ожиданий заключается в простой процедуре корректировки ожиданий, когда в каждый момент времени реальное значение переменной сравнивается с еѐ ожидаемым значением. Если реальное значение оказывается больше, то ожидаемое в следующий момент значение корректируется в сторону его увеличения, если меньше – то в сторону уменьшения. Предполагается, что размер корректировки пропорционален разности между реальным и ожидаемым значениями переменной. Таким образом, основную идею можно записать формулой:

$$x_{t+1}^e - x_t^e = \lambda\left(x_t - x_t^e\right) \quad (0 \le \lambda \le 1), \qquad (16)$$

где x_t^e - значение x, ожидаемое в момент t (expected). Это выражение можно переписать в форме взвешенного среднего:

$$x_{t+1}^e = \lambda x_t + (1-\lambda) x_t^e . \qquad (17)$$

Модель (16) и является моделью адаптивных ожиданий. Это выражение иногда называют моделью обучения на ошибках, т.к. ожидания экономических

объектов складываются из прошлых ожиданий, поправленных на величину ошибки в ожиданиях, допущенных ранее.

При λ=0 ожидания являются статичными, неизменными, т.е. $x_{t+1}^e = x_t^e$.

При λ=1 ожидания реализуются мгновенно, т.е. $x_{t+1}^e = x_t$.

Чем больше λ, тем быстрее ожидаемое значение адаптируется к предыдущим реальным значениям переменной.

Долгосрочная функция модели адаптивных ожиданий записывается в виде:

$$y_t = a + bx_{t+1}^e + \varepsilon_t . \qquad (18)$$

Подставим сюда выражение (17), получим:

$$y_t = a + \lambda b x_t + (1-\lambda) b x_t^e + \varepsilon_t . \qquad (19)$$

Запишем его для (t-1):

$$y_{t-1} = a + \lambda b x_{t-1} + (1-\lambda) b x_{t-1}^e + \varepsilon_{t-1} . \qquad (20)$$

Умножим (20) на (1-λ) и вычтем почленно из (18):

$$y_t = \lambda a + \lambda b x_t + (1-\lambda) b y_{t-1} + u_t , \qquad (21)$$

где $u_t = \varepsilon_t - (1-\lambda)\varepsilon_{t-1}$.

Это модель авторегрессии, в которой все переменные имеют фактические, а не ожидаемые значения. Модель в форме (21) называется краткосрочной функцией модели адаптивных ожиданий.

Модель неполной (частичной) корректировки. Здесь поведенческое уравнение определяет не фактическое значение yt, а еѐ желаемый (целевой) уровень y_t^*:

$$y_t^* = a + bx_t + u_t . \qquad (22)$$

Примером такой модели служит политика компаний относительно распределения дивидендов: прибыль расходуется частично на уплату дивидендов, частью на инвестиции. Когда прибыль увеличивается, дивиденды тоже растут,

но не в той же пропорции (это объясняется желанием руководства фирмы в любом случае не уменьшать дивиденды, т.к. это ударяет по репутации фирмы).

В модели предполагается, что фактическое приращение зависимой переменной пропорционально разнице между еѐ желаемым уровнем и значением в предыдущий период:

$$y_t - y_{t-1} = \lambda\left(y_t^* - y_{t-1}\right) + \nu_t \,, \qquad (23)$$

(νt – случайный член). Это выражение можно переписать так:

$$y_t = \lambda y_t^* + (1-\lambda) y_{t-1} + \nu_t \,, \qquad (24)$$

т.е. в форме взвешенного среднего. Чем больше λ, тем быстрее идет корректировка. При λ =1 полная корректировка происходит за один период. При λ =0 корректировка не происходит совсем.

Подставим (22) в (24), получим:

$$y_t = a\lambda + b\lambda x_t + (1-\lambda) y_{t-1} + \nu_t + \lambda u_t \,. \qquad (25)$$

Это и есть модель частичной корректировки, которая также является моделью авторегрессии.

Несколько слов об оценке параметров уравнений авторегрессии. Рассмотрим уравнение:

$$y_t = a + b x_t + c y_{t-1} + \varepsilon_t \,. \qquad (26)$$

Во всех рассмотренных выше моделях стоит проблема оценивания параметров. Обычный МНК чаще всего даѐт смещенные и несостоятельные оценки, вследствие автокорреляции между случайными отклонениями εt и εt-1 и корреляции между yt-1 и εt.

Один из возможных методов расчета параметров – метод инструментальных переменных, состоящий в замене yt-1 на новую переменную, которая тесно коррелирует с yt-1, но не коррелирует с остатками. Это можно сделать двумя способами.

Провести регрессию

$$y_{t-1} = \delta_0 + \delta_1 x_{t-1} + u_t \,, \qquad (27)$$

или

$$y_{t-1} = \hat{y}_{t-1} + u_t$$

и подставить $\hat{y}_{t-1}$ в уравнение авторегрессии, получаем:

$$y_t = a + bx_t + c\hat{y}_{t-1} + \nu_t, \qquad (28)$$

и далее применяем обычный МНК.

Подставим (27) в (26), получим модель с распределенным лагом:

$$y_t = (a + c\delta_0) + bx_t + c\delta_1 x_{t-1} + (cu_t + \varepsilon_t), \qquad (29)$$

для которой не нарушаются предпосылки обычного МНК.

Вопросы и задания для самоконтроля

1. Какая модель временного ряда называется динамической?
2. Какие типы включают динамические модели?
3. Как определяются модели с распределенными лагами?
4. Как интерпретируют параметры модели с распределенным лагом?
5. Как определяются авторегрессионные модели?
6. Как интерпретируют параметры моделей авторегрессии?
7. В чем основная идея метода Алмон и к каким моделям он применяется?
8. Когда применяется преобразование Койка?
9. Как оценить параметры моделей авторегрессии?
10. В чем суть метода инструментальных переменных?
11. Для чего применяется модель адаптивных ожиданий?
12. Для чего применяется модель частичной корректировки?

Задание 1. Модель зависимости объемов продаж компании в среднем за месяц от расходов на рекламу была следующая (млн. руб):

$$\tilde{y}_t = -0{,}73 + 4{,}3x_t + 3{,}5x_{t-1} + 1{,}2x_{t-2} + 0{,}8x_{t-3}.$$

Найти краткосрочный, долгосрочный мультипликатор и средний лаг.

Задание 2. Дана таблица следующих данных:

Момент времени	$t-3$	$t-2$	$t-1$	t	$t+1$
x^*	80				
x	90	95	110	120	-

Здесь x^*, x - ожидаемый и действительный спрос на некоторый товар соответственно. В соответствии с моделью адаптивных ожиданий

$x_t^* = \lambda x_{t-1} + (1-\lambda)x_{t-1}^*$, где λ =0,40 найти остальные значения x^*.

Лекция 21

Тема 19. Понятие о системах эконометрических уравнений

Вопросы для изучения

1. Понятие о системах уравнений. Системы независимых уравнений и системы взаимозависимых уравнений.

2. Структурная и приведенная формы модели.

3. Идентификация модели.

Аннотация. Данная тема излагает типы систем эконометрических уравнений.

Ключевые слова. Система взаимозависимых уравнений, идентификация системы взаимозависимых уравнений, структурная и приведенная формы модели.

Методические рекомендации по изучению темы

•Изучить лекционную часть, где даются общие представления по данной теме.

• Для закрепления теоретического материала ознакомиться с решениями типовых задач и ответить на вопросы для самоконтроля.

• Для проверки усвоения темы выполнить практические задания и тест для самоконтроля.

Рекомендуемые информационные ресурсы:

1. http://tulpar.kfu.ru/course/view.php?id=2213.

2. Эконометрика: [Электронный ресурс] Учеб.пособие / А.И. Новиков. - 2-е изд., испр. и доп. - М.: ИНФРА-М, 2011. - 144 с.: с.

(http://znanium.com/catalog.php?item=booksearch&code=%D1%8D%D0%BA%D0%BE%D0%BD%D0%BE%D0%BC%D0%B5%D1%82%D1%80%D0%B8%D0%BA%D0%B0&page=1#none)С. 117-136.

3. Валентинов В. А. Эконометрика [Электронный ресурс]: Практикум / В. А. Валентинов. - 3-е изд. - М.: Дашков и К, 2010. - 436 с.

(http://znanium.com/catalog.php?item=booksearch&code=%D1%8D%D0%BA%D0%BE%D0%BD%D0%BE%D0%BC%D0%B5%D1%82%D1%80%D0%B8%D0%BA%D0%B0&page=3#none) С. 338-356.

4.Эконометрика. Практикум: [Электронный ресурс] Учебное пособие / С.А. Бородич. - М.: НИЦ ИНФРА-М; Мн.: Нов.знание, 2014. - 329 с. (http://znanium.com/catalog.php?item=booksearch&code=%D1%8D%D0%BA%D0%BE%D0%BD%D0%BE%D0%BC%D0%B5%D1%82%D1%80%D0%B8%D0%BA%D0%B0&page=4#none) С. 286-313.

Понятие о системах уравнений. Системы независимых уравнений и системы взаимозависимых уравнений. Объектом статистического изучения в социальных науках являются сложные системы. Построение изолированных уравнений регрессии недостаточно для описания таких систем и объяснения механизма их функционирования. Изменение одной переменной, как правило, не может происходить без изменения других. Поэтому важное место занимает проблема описания структуры связей между переменными системой так называемых одновременных уравнений. Так, если изучается модель спроса как отношение цен и количества потребляемых товаров, то одновременно для прогнозирования спроса необходима модель предложения товаров, в которой рассматривается также взаимосвязь между количеством и ценой предлагаемых благ. Это позволяет достичь равновесия между спросом и предложением.

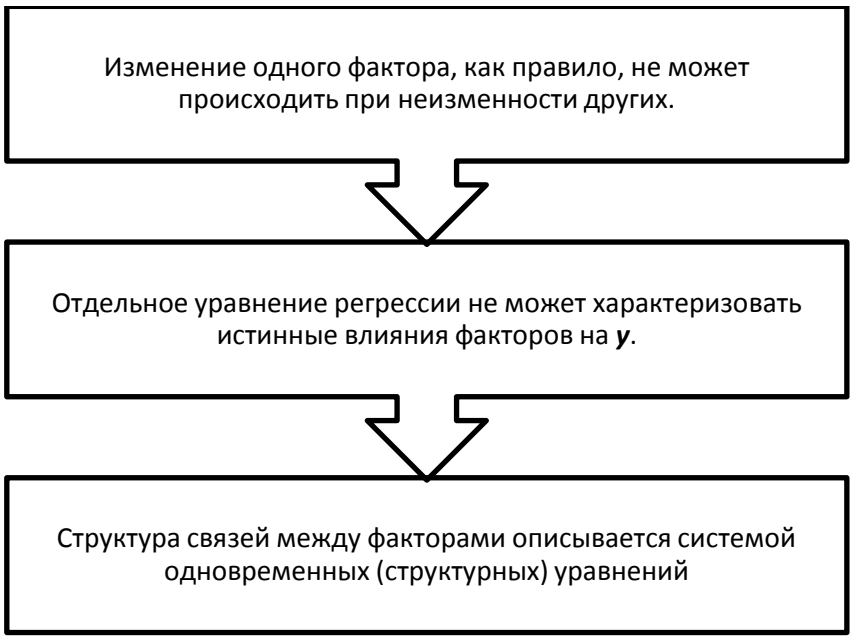


Рис. 19.1. Необходимость систем уравнений

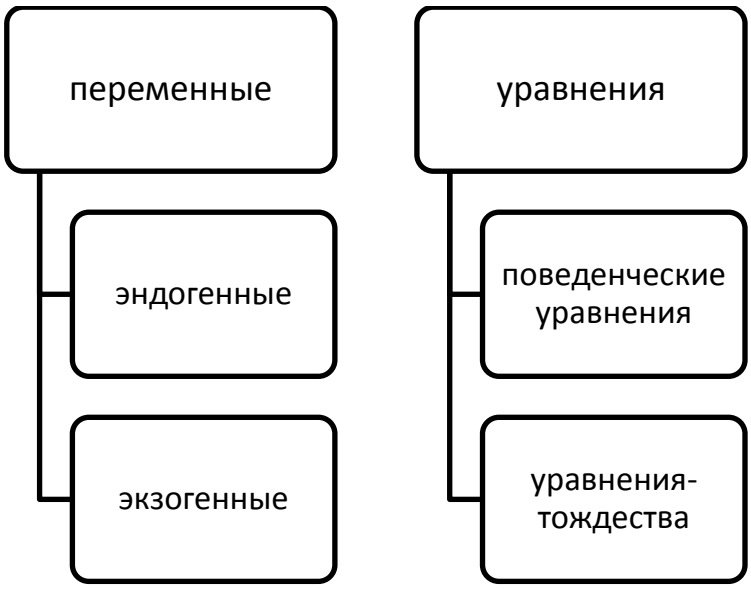


Рис. 19.2. Составляющие систем уравнений

Эндогенные переменные обычно обозначаются как y. Это зависимые переменные, значения которых определяются внутри модели. Их число равно числу уравнений в системе.

Экзогенные переменные обычно обозначаются как x. Это внешние по отношению к модели переменные. Они влияют на эндогенные переменные, но не зависят от них.

Лаговые переменные – это значения эндогенных переменных за предшествующий период времени (yt-1). В модели участвуют в качестве экзогенных переменных.

В поведенческих уравнениях описываются взаимодействия между переменными.

В уравнениях-тождествах описываются соотношения, которые должны выполняться во всех случаях. Тождества не содержат подлежащие оценке параметры a и b, а также случайное отклонение ε.

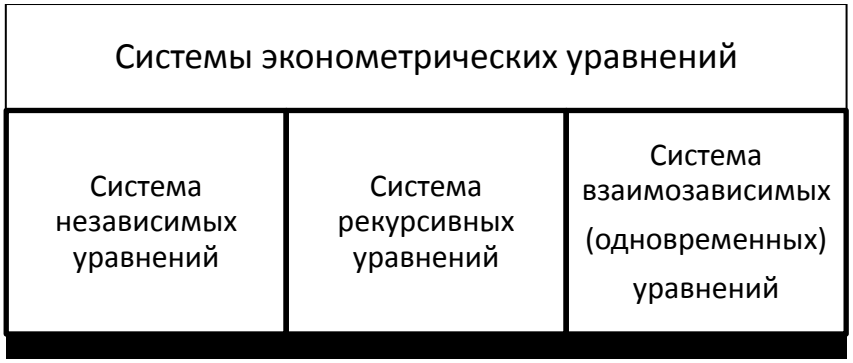


Рис. 19.3. Виды систем уравнений

В системе независимых уравнений каждая зависимая переменная y рассматривается как функция одного и того же набора факторов x:

$$\begin{cases} y_1 = a_{11}x_1 + a_{12}x_2 + \ldots + a_{1m}x_m + \varepsilon_1, \\ y_2 = a_{21}x_1 + a_{22}x_2 + \ldots + a_{2m}x_m + \varepsilon_2, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ y_n = a_{n1}x_1 + a_{n2}x_2 + \ldots + a_{nm}x_m + \varepsilon_n. \end{cases}$$

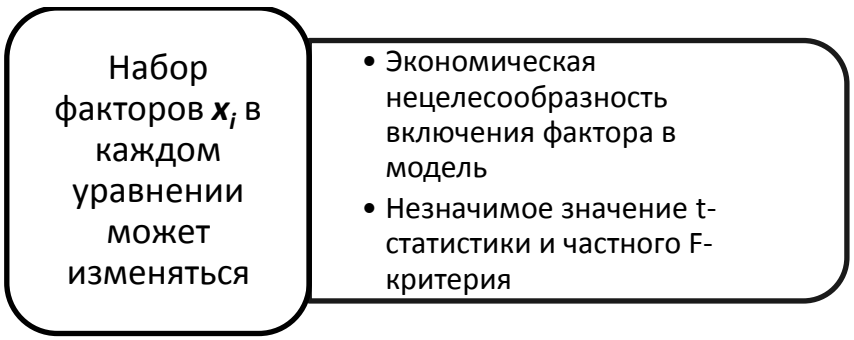


Рис. 19.4. Включение факторов в модель

Система независимых уравнений с различным набором факторов:

$$\begin{cases} y_1 = f(x_1, x_2, x_3, x_4, x_5), \\ y_2 = f(x_1, x_3, x_4, x_5), \\ y_3 = f(x_2, x_3, x_5), \\ y_4 = f(x_3, x_4, x_5). \end{cases}$$

В системе рекурсивных уравнений каждое последующее уравнение включает в качестве факторов все зависимые переменные у предшествующих уравнений наряду с набором собственно факторов х:

$$\begin{cases} y_1 = a_{11}x_1 + a_{12}x_2 + \ldots + a_{1m}x_m + \varepsilon_1, \\ y_2 = b_{21}y_1 + a_{21}x_1 + a_{22}x_2 + \ldots + a_{2m}x_m + \varepsilon_2, \\ y_3 = b_{31}y_1 + b_{32}y_2 + a_{31}x_1 + a_{32}x_2 + \ldots + a_{3m}x_m + \varepsilon_3, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ y_n = b_{n1}y_1 + b_{n2}y_2 + b_{n3}y_3 + \ldots + b_{nn-1}y_{n-1} + \\ + a_{n1}x_1 + a_{n2}x_2 + \ldots + a_{nm}x_m + \varepsilon_n. \end{cases}$$

В системе взаимозависимых уравнений одни и те же зависимые переменные у в одних уравнениях входят в левую часть, а в других уравнениях – в правую часть системы:

$$\begin{cases} y_1 = b_{12}y_2 + b_{13}y_3 + \ldots + b_{1n}y_n + a_{11}x_1 + a_{12}x_2 + \ldots + a_{1m}x_m + \varepsilon_1, \\ y_2 = b_{21}y_1 + b_{23}y_3 + \ldots + b_{2n}y_n + a_{21}x_1 + a_{22}x_2 + \ldots + a_{2m}x_m + \varepsilon_2, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ y_n = b_{n1}y_1 + b_{n2}y_2 + \ldots + b_{nn-1}y_{n-1} + a_{n1}x_1 + a_{n2}x_2 + \ldots + a_{nm}x_m + \varepsilon_n. \end{cases}$$

Этот вид систем уравнений получил наибольшее распространение в эконометрических исследованиях. В эконометрике эта система уравнений называется также структурной формой модели (СФМ). Для нахождения параметров каждого уравнения традиционный МНК неприменим, здесь используются специальные методы оценивания. В этом случае каждое из уравнений не может рассматриваться самостоятельно.

Структурная и приведенная формы модели. Система взаимозависимых (одновременных) уравнений, описывающая структуру связей между переменными, называется структурной формой модели. Коэффициенты b_i и a_j называются структурными коэффициентами модели. Приведенная форма модели представляет собой систему линейных функций эндогенных переменных от экзогенных. В каждое приведенное уравнение включаются все экзогенные пере-

менные структурной модели. Система одновременных уравнений (т.е. структурная форма модели) обычно содержит эндогенные и экзогенные переменные. Эндогенные переменные – это зависимые переменные, число которых равно числу уравнений в системе. Они обозначаются через у. Экзогенные переменные – это предопределенные переменные, влияющие на эндогенные переменные, но не зависящие от них. Они обозначаются через х.

Простейшая структурная форма модели имеет вид:

$$\begin{cases} y_1 = b_{12}y_2 + a_{11}x_1 + \varepsilon_{1,} \\ y_2 = b_{21}y_1 + a_{22}x_2 + \varepsilon_{2,} \end{cases}$$

где y_1, y_2 – эндогенные переменные, x_1, x_2 – экзогенные.

Классификация переменных на эндогенные и экзогенные зависит от теоретической концепции принятой модели. Экономические переменные могут выступать в одних моделях как эндогенные, а в других – как экзогенные переменные. Внеэкономические переменные (например, климатические условия) входят в систему как экзогенные переменные. В качестве экзогенных переменных можно рассматривать значения эндогенных переменных за предшествующий период времени (лаговые переменные). Например, потребление текущего года yt может зависеть также и от уровня потребления в предыдущем году y_{t-1}.

Структурная форма модели позволяет увидеть влияние изменений любой экзогенной переменной на значения эндогенной переменной. Целесообразно в качестве экзогенных переменных выбирать такие переменные, которые могут быть объектом регулирования. Меняя их и управляя ими, можно заранее иметь целевые значения эндогенных переменных.

Коэффициенты b_i при эндогенных и a_j – при экзогенных переменных называются структурными коэффициентами модели. Все переменные в модели могут быть выражены в отклонениях $(x - \bar{x})$ и $(y - \bar{y})$ от среднего уровня, и тогда свободный член в каждом уравнении отсутствует.

Использование МНК для оценивания структурных коэффициентов модели дает смещенные и несостоятельные оценки. Поэтому обычно для определения структурных коэффициентов модели структурная форма преобразуется в приведенную.

Приведенная форма модели (ПФМ) представляет собой систему линейных функций эндогенных переменных от экзогенных:

$$\begin{cases} \hat{y}_1 = \delta_{11}x_1 + \ldots + \delta_{1m}x_{m,} \\ \hat{y}_2 = \delta_{21}x_1 + \ldots + \delta_{2m}x_{m,} \\ \ldots \\ \hat{y}_n = \delta_{n1}x_1 + \ldots + \delta_{nm}x_{m.} \end{cases}$$

δ_{ij} – коэффициенты приведенной формы модели.

По своему виду приведенная форма модели ничем не отличается от системы независимых уравнений. Применяя МНК, можно оценить δ_{ij}, а затем оценить значения эндогенных переменных через экзогенные.

Приведенная форма позволяет выразить значения эндогенных переменных через экзогенные, однако аналитически уступает структурной форме модели, т.к. в ней отсутствуют оценки взаимосвязи между эндогенными переменными.

Идентификация модели. При переходе от приведенной формы модели к структурной исследователь сталкивается с проблемой идентификации. Идентификация – это единственность соответствия между приведенной и структурной формами модели.

Структурная модель в полном виде, состоящая в каждом уравнении системы из n эндогенных и m экзогенных переменных, содержит n(n-1+m) параметров. Приведенная модель в полном виде содержит nm параметров. Таким образом, в полном виде структурная модель содержит большее число параметров, чем приведенная форма модели. Поэтому n(n-1+m) параметров структур-

ной модели не могут быть однозначно определены через nm параметров приведенной формы модели.

Чтобы получить единственно возможное решение для структурной модели, необходимо предположить, что некоторые из структурных коэффициентов модели равны нулю. Тем самым уменьшится число структурных коэффициентов.

С позиции идентифицируемости структурные модели можно подразделить на три вида: идентифицируемые; неидентифицируемые; сверхидентифицируемые.

Модель идентифицируема, если все структурные ее коэффициенты определяются однозначно, единственным образом по коэффициентам приведенной формы модели, т.е. число параметров структурной модели равно числу параметров приведенной формы модели.

Модель неидентифицируема, если число приведенных коэффициентов меньше числа структурных коэффициентов, и в результате структурные коэффициенты не могут быть оценены через коэффициенты приведенной формы модели.

Модель сверхидентифицируема, если число приведенных коэффициентов больше числа структурных коэффициентов. В этом случае на основе приведенных коэффициентов можно получить два или более значений одного структурного коэффициента. Сверхидентифицируемая модель, в отличие от неидентифицируемой, практически решаема, но требует для этого специальных методов исчисления параметров.

Структурная модель всегда представляет собой систему совместных уравнений, каждое из которых требуется проверять на идентификацию. Модель считается идентифицируемой, если каждое уравнение системы идентифицируемо. Если хотя бы одно из уравнений системы неидентифицируемо, то и вся модель считается неидентифицируемой. Если же в системе нет неидентифицируемых уравнений и имеется хотя бы одно сверхидентифицируемое, то модель будет сверхидентифицируемой.

Обозначим H – число эндогенных переменных в i- ом уравнении системы, D – число экзогенных переменных, которые содержатся в системе, но не входят в данное уравнение. Тогда условие идентифицируемости уравнения может быть записано в виде следующего счетного правила:

D+1 = H – уравнение идентифицируемо;

D+1 < H – уравнение неидентифицируемо;

D+1 > H – уравнение сверхидентифицируемо.

Это счетное правило отражает необходимое, но не достаточное условие идентификации. Достаточное условие идентификации отдельного уравнения состоит в том, чтобы матрица из коэффициентов при переменных, которые в данном уравнении отсутствуют (то есть коэффициенты берутся из всех остальных уравнений системы), имела ранг не меньший, чем количество эндогенных переменных в системе минус единица.

Следует помнить, что на идентификацию проверяется каждое уравнение модели.

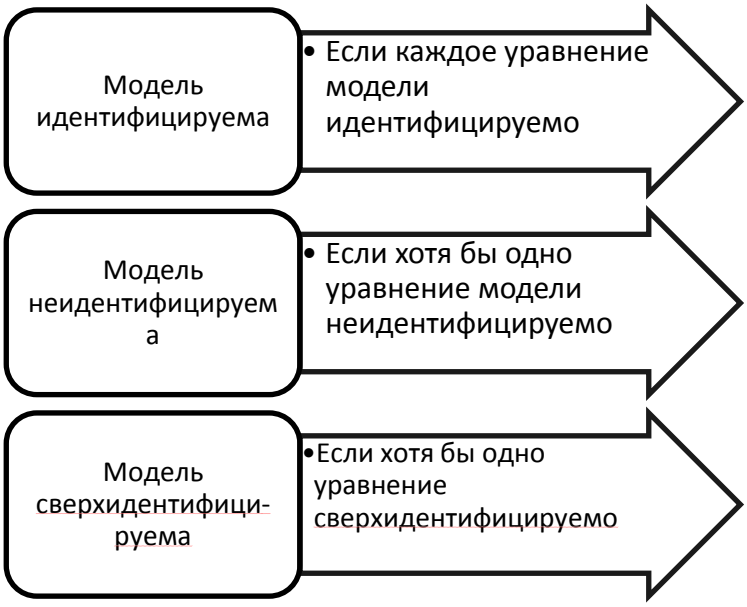


Рис. 19.5. Вывод об идентификации

Вопросы и задания для самоконтроля

1. В чем преимущество систем эконометрических уравнений?
2. Какие переменные называют предопределенными?

3. Что такое структурная форма модели?

4. Что такое приведенная форма модели?

5. Почему нужна приведенная форма модели?

6. Когда структурная модель является идентифицируемой?

7. Когда структурная модель является неидентифицируемой?

8. В каком случае модель является сверхидентифицируемой?

9. Как идентифицируется отдельное уравнение в системе по счетному правилу?

10. В чем состоит достаточное условие идентификации отдельного уравнения?

Задание 1. Дана модель Менгеса:

$$Y_t = \alpha_1 + b_{11}Y_{t-1} + b_{12}I_t + \varepsilon_1,$$

$$I_t = \alpha_2 + b_{21}Y_t + b_{22}Q_t + \varepsilon_2,$$

$$C_t = \alpha_3 + b_{13}Y_t + b_{32}C_{t-1} + b_{33}P_t + \varepsilon_3,$$

$$Q_t = \alpha_4 + b_{41}Q_{t-1} + b_{42}R_t + \varepsilon_4.$$

где Y - национальный доход; C - расходы на личное потребление; I - чистые инвестиции; Q - валовая прибыль экономики; P - индекс стоимости жизни; R - объем продукции промышленности; t - текущий период; $(t-1)$ - предыдущий период. Проверить идентифицируемость каждого уравнения с использованием необходимого и достаточного условий идентификации.

Задание 2. Имеется модель денежного и товарного рынков:

$$R_t = \alpha_1 + b_{12}Y_t + b_{14}M_t + \varepsilon_1,$$

$$Y_t = \alpha_2 + b_{21}R_t + b_{23}I_t + b_{25}G_t + \varepsilon_2,$$

$$I_t = \alpha_3 + b_{31}R_t + \varepsilon_3,$$

где R - процентные ставки; Y - реальный ВВП; M - денежная масса; I - внутренние инвестиции; G - реальные государственные расходы; t - текущий период.

Проверить идентифицируемость каждого уравнения с использованием необходимого и достаточного условий идентифицируемости и записать приведенную форму модели.

Лекция 22

Тема 20. Методы оценки систем одновременных уравнений

Вопросы для изучения

1. Косвенный, двухшаговый и трехшаговый МНК.

2. Применение систем уравнений для построения макроэкономических моделей и моделей спроса – предложения.

Аннотация. Данная тема раскрывает методы оценки систем эконометрических уравнений.

Ключевые слова. Косвенный метод наименьших квадратов, двухшаговый метод наименьших квадратов, модели спроса-предложения.

Методические рекомендации по изучению темы

•Изучить лекционную часть, где даются общие представления по данной теме.

• Для закрепления теоретического материала ознакомиться с решениями типовых задач и ответить на вопросы для самоконтроля.

• Для проверки усвоения темы выполнить практические задания и тест для самоконтроля.

•Для подготовки к экзамену выполнить итоговый контрольный тест и итоговые практические задания.

Рекомендуемые информационные ресурсы:

1. http://tulpar.kfu.ru/course/view.php?id=2213.

2. Эконометрика: [Электронный ресурс] Учеб.пособие / А.И. Новиков. - 2-е изд., испр. и доп. - М.: ИНФРА-М, 2011. - 144 с.: с. (http://znanium.com/catalog.php?item=booksearch&code=%D1%8D%D0%BA%D0

%BE%D0%BD%D0%BE%D0%BC%D0%B5%D1%82%D1%80%D0%B8%D0%BA%D0%B0&page=1#none) С. 117-136.

3. Валентинов В. А. Эконометрика [Электронный ресурс]: Практикум / В. А. Валентинов. - 3-е изд. - М.: Дашков и К, 2010. - 436 с.

(http://znanium.com/catalog.php?item=booksearch&code=%D1%8D%D0%BA%D0%BE%D0%BD%D0%BE%D0%BC%D0%B5%D1%82%D1%80%D0%B8%D0%BA%D0%B0&page=3#none) С. 338-356.

4.Эконометрика. Практикум: [Электронный ресурс] Учебное пособие / С.А. Бородич. - М.: НИЦ ИНФРА-М; Мн.: Нов.знание, 2014. - 329 с. (http://znanium.com/catalog.php?item=booksearch&code=%D1%8D%D0%BA%D0%BE%D0%BD%D0%BE%D0%BC%D0%B5%D1%82%D1%80%D0%B8%D0%BA%D0%B0&page=4#none) С. 286-313.

Косвенный, двухшаговый и трехшаговый МНК. Косвенный МНК применяется для оценивания идентифицируемых систем одновременных уравнений.

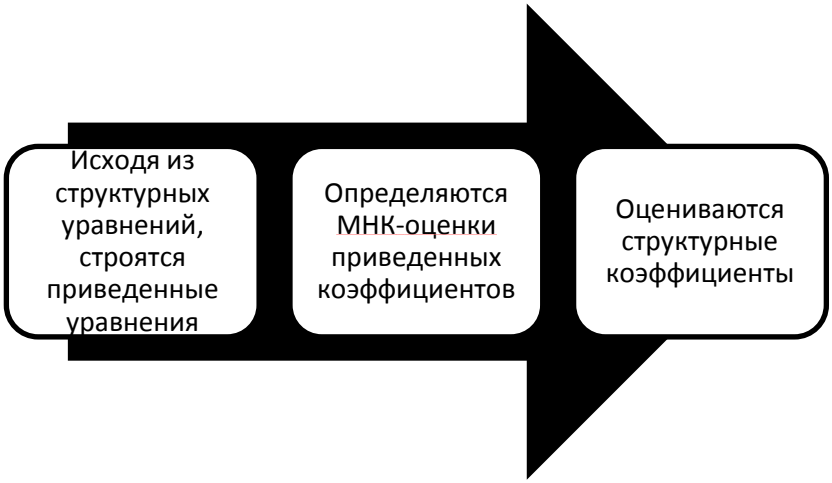


Рис. 20.1. Этапы косвенного МНК

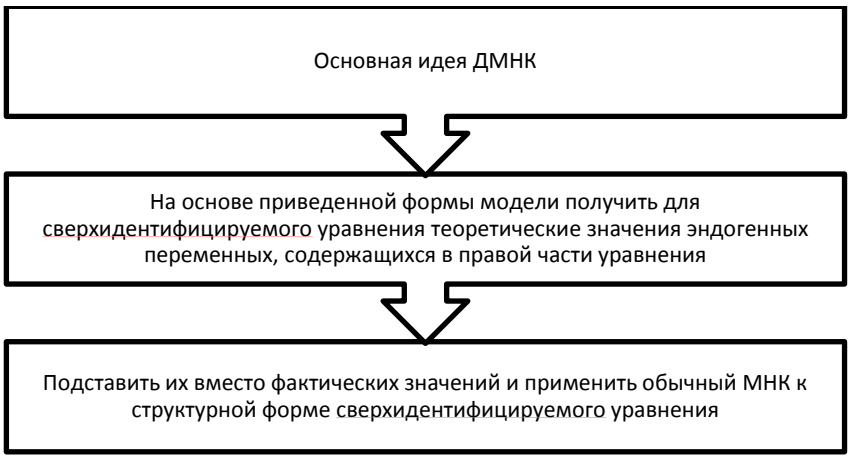


Рис. 20.2. Основная идея двухшагового метода наименьших квадратов

Таким образом, метод наименьших квадратов используется дважды:

1) При определении приведенной формы модели и нахождении на ее основе оценок эндогенной переменной $\hat{y}_i = \delta_{i1}x_1 + \delta_{i2}x_2 + \ldots + \delta_{ij}x_j$

2) При определении структурных коэффициентов структурного сверхидентифицируемого уравнения на основе оценок эндогенных переменных.

Двухшаговый метод наименьших квадратов является наиболее общим и широко распространенным методом решения системы одновременных уравнений. Для точно идентифицируемых уравнений ДМНК дает тот же результат, что и КМНК.

Трехшаговый МНК разработан для оценки одновременно всех уравнений структурной формы модели с учетом возможной взаимной коррелированности регрессионных остатков различных уравнений системы. Этот метод оказывается более эффективным, чем ДМНК, если случайные остатки различных уравнений системы взаимно коррелированы, т.е. если их взаимная ковариационная матрица отлична от диагональной. Однако и в этой ситуации ДМНК – оценки структурных параметров системы остаются состоятельными.

В трехшаговом МНК сохранены первые два шага ДМНК. Однако полученные в результате этих двух шагов, отдельно для каждого уравнения, оценки структурных параметров не являются окончательными, а пересчитываются на 3 – м шаге следующим образом. Оценки структурных коэффициентов используются для подсчета выборочной ковариационной матрицы случайных остатков.

Последняя, в свою очередь, используется для одновременного вычисления оценок всех структурных параметров системы с помощью обобщенного МНК в рамках соответствующим образом построенной обобщенной линейной модели множественной регрессии.

Применение систем уравнений для построения макроэкономических моделей и моделей спроса – предложения

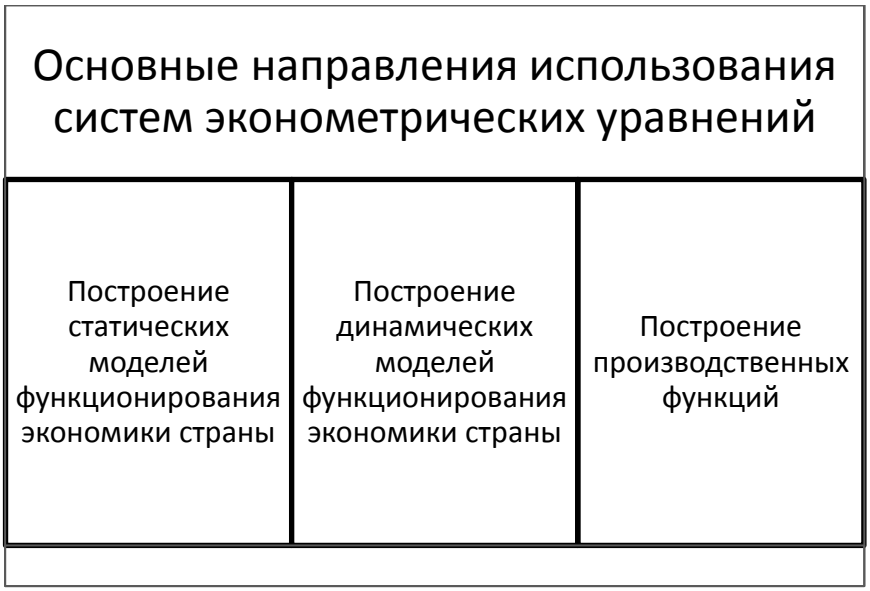


Рис. 20. 3. Основные направления использования систем эконометрических уравнений

Наиболее широко системы одновременных уравнений используются при построении макроэкономических моделей экономики страны. В большинстве случаев это мультипликаторные модели кейнсианского типа. Статическая модель Кейнса в самом простом виде следующая:

$$\begin{cases} C = a + by + \varepsilon, \\ y = C + I, \end{cases}$$

где C – личное потребление;

y – национальный доход в постоянных ценах;

I – инвестиции в постоянных ценах.

В силу наличия тождества в модели (второе уравнение системы) $b \le 1.$ Он характеризует предельную склонность к потреблению. Если $b = 0{,}65,$ из каждой дополнительной тысячи рублей дохода на потребление расходуется в среднем 650 рублей и 350 рублей инвестируется. Если b>1 то y<C+I, и на потребление расходуются не только доходы, но и сбережения. Параметр a Кейнс истолковывал как прирост потребления за счет других факторов.

Структурный коэффициент b используется для расчета мультипликаторов. По данной функции потребления можно определить два мультипликатора – инвестиционный мультипликатор потребления Мс и национального дохода Му :

$$M_c = \frac{b}{1-b},$$ т.е. при $b = 0{,}65;$ $M_c = \frac{0{,}65}{1-0{,}65} = 1{,}857$.

Это означает, что дополнительные вложения 1 тыс. руб. приведут при прочих равных условиях к дополнительному увеличению потребления на 1,857 тыс. руб.

$$M_y = \frac{1}{1-b},$$ т.е. при $b = 0{,}65$ $M_y = \frac{1}{1-0{,}65} = 2{,}857$,

т.е. дополнительные вложения 1 тыс. руб. на длительный срок приведут при прочих равных условиях к дополнительному доходу 2,857 тыс. руб.

Эта модель точно идентифицируема, и для получения b применяется КМНК. Строится система приведенных уравнений:

$$\begin{cases} C = A + B \cdot I + U_1 \\ y = A' + B' \cdot I + U_2, \end{cases}$$

в которой $A = A',$ а параметры B и B' являются мультипликаторами, т.е. $B = M_c$ и $B' = M_y$. Для проверки подставим балансовое равенство в первое уравнение структурной модели:

$$C = a + by + \varepsilon = a + b(C + I) + \varepsilon = a + bc + bI + \varepsilon \Rightarrow$$

$$\Rightarrow C(1-b) = a + bI + \varepsilon \Rightarrow C = \underset{A}{\frac{a}{1-b}} + \underset{M_c}{\frac{b}{1-b}} I + \varepsilon \underset{U_1}{\frac{1}{1-b}}$$

Аналогично поступим и со вторым уравнением структурной модели:

$$y = C + I \Rightarrow y = a + by + \varepsilon + I \Rightarrow y(1-b) = a + I + \varepsilon \Rightarrow$$

$$\Rightarrow y = \underset{A' = A}{\frac{a}{1-b}} + \underset{M_y}{\frac{1}{1-b}} I + \underset{U_2}{\frac{1}{1-b}} \varepsilon$$

Таким образом, приведенная форма содержит мультипликаторы, интерпретируемые как коэффициенты множественной регрессии, отвечающие на вопрос, на сколько единиц изменится значение эндогенной переменной, если экзогенная изменится на 1 единицу. Это делает модель удобной для прогнозирования.

В более поздних исследованиях статическая модель Кейнса включала уже не только функцию потребления, но и функцию сбережений:

$$\begin{cases} C = a + by + \varepsilon_1, \\ r = T + K(C + I) + \varepsilon_2, \\ y = C + I - r, \end{cases}$$

где r — сбережения.

Здесь три эндогенные переменные – C, r и y и одна экзогенная –I. Система идентифицируема: в первом уравнении H=2 и D=2, во втором H=1, D=0; $C + I$ рассматривается как предопределенная переменная.

Наряду со статическими широкое распространение получили динамические модели экономики. Они содержат в правой части лаговые переменные, а также учитывают тенденцию. Например, модель Кейнса экономики США 1950-1960 гг. в упрощенном варианте:

$$\begin{cases} C_t = b_1 S_t + b_2 P_t + b_3 + \varepsilon_{1,} \\ I_t = b_4 P_t + b_5 P_{t-1} + b_6 + \varepsilon_{2,} \\ S_t = b_7 R_t + b_8 R_{t-1} + b_9 t + b_{10} + \varepsilon_{3,} \\ R_t = S_t + P_t + T_{t,} \\ R_t = C_t + I_t + G_t \end{cases}$$

где $T_t -$ чистые трансферты в пользу администрации;

$I_t -$ кап. вложения;

$G_t -$ правительственные расходы;

$S_t -$ заработная плата в период t;

$P_t -$ прибыль;

$P_{t-1} -$ прибыль в период $t-1$;

$R_t -$ общий доход.

Модель содержит 5 эндогенных переменных – C_t, I_t, S_t, R_t (в левой части системы) и P_t (зависимая переменная, определяемая по первому тождеству), три экзогенные переменные – T_t, G_t, t и две лаговые предопределенные переменные P_{t-1} и $R_{t-1.}$ Данная модель сверхидентифицируема и решается ДМНК. Для прогнозных целей используется приведенная форма модели:

$$\begin{cases} C_t = d_1 T + d_2 G + d_3 t + d_4 P_{t-1} + d_5 R_{t-1} + U_{1,} \\ I_t = d_6 T + d_7 G + d_8 t + d_9 P_{t-1} + d_{10} R_{t-1} + U_{2,} \\ S_t = d_{11} T + d_{12} G + d_{13} t + d_{14} P_{t-1} + d_{15} R_{t-1} + U_{3,} \\ R_t = d_{16} T + d_{17} G + d_{18} t + d_{19} P_{t-1} + d_{20} R_{t-1} + U_{4,} \\ P_t = d_{21} T + d_{22} G + d_{23} t + d_{24} P_{t-1} + d_{25} R_{t-1} + U_5 \end{cases}$$

Здесь мультипликаторами являются коэффициенты при экзогенных переменных. Они отражают влияние экзогенной переменной на эндогенную переменную.

Система одновременных уравнений нашла применение в исследованиях спроса и предложения. Линейная модель спроса и предложения имеет вид:

$$\begin{cases} Q^d = a_0 + a_1 P + \varepsilon_{1,} \text{ - объѐм спроса,} \\ Q^s = b_0 + b_1 P + \varepsilon_{2,} \text{ - объѐм предложения,} \\ Q^d = Q^s \end{cases}$$

Здесь 3 эндогенные переменные: Q^d, Q^s и P. При этом, если Q^d и Q^s представляют собой эндогенные переменные, исходя из структуры самой системы, то P является эндогенной по экономическому содержанию (цена зависит от спроса и предложения), а также в результате наличия тождества $Q^d = Q^s$. Приравняем уравнения, получим:

$$a_0 + a_1 P + \varepsilon_1 = b_0 + b_1 P + \varepsilon_2,$$

$$P = \frac{b_0 - a_0}{a_1 - b_1} + \frac{\varepsilon_2 - \varepsilon_1}{a_1 - b_1}.$$

Модель не содержит экзогенной переменной. Однако, чтобы модель имела статистическое решение и можно было убедиться в ее справедливости, в модель вводятся экзогенные переменные.

Например, модель вида:

$$\begin{cases} Q^d = a_0 + a_1 P + a_2 R + \varepsilon_{1,} \\ Q^s = b_0 + b_1 P + b_2 W + \varepsilon_{2,} \\ Q^d = Q^s, \end{cases}$$

где R – доход на душу населения; W – климатические условия (при спросе и предложении зерна).

Переменные R и W экзогенные. Введя их в модель, получаем идентифицированную структурную модель.

Динамическая модель Клейна:

$$\begin{cases} C_t = b_1 \cdot S_t + b_2 \cdot P_t + b_3 + \varepsilon_1, \\ I_t = b_4 \cdot P_t + b_5 \cdot P_{t-1} + b_6 + \varepsilon_2, \\ S_t = b_7 \cdot R_t + b_8 \cdot R_{t-1} + b_9 \cdot t + b_{10} + \varepsilon_3, \\ R_t = S_t + P_t + T_t, \\ R_t = C_t + I_t + G_t. \end{cases}$$

Ct – функция потребления в период t; St – заработная плата в период t;Pt – прибыль в период t; Pt-1 – прибыль в период t-1; Rt – общий доход в период t; Rt-1 – общий доход в предыдущий период;t – время;Tt – чистые трансферты в пользу администрации в период t; It – капиталовложения в период t;Gt – спрос административного аппарата, правительственные расходы в период t.

Динамическая модель Клейна сверхидентифицируема и решаема ДМНК. Для прогнозных целей используется приведенная форма модели:

$$\begin{cases} C_t = d_1 T + d_2 G + d_3 t + d_4 P_{t-1} + d_5 R_{t-1} + u_1, \\ I_t = d_6 T + d_7 G + d_8 t + d_9 P_{t-1} + d_{10} R_{t-1} + u_2, \\ S_t = d_{11} T + d_{12} G + d_{13} t + d_{14} P_{t-1} + d_{15} R_{t-1} + u_3, \\ R_t = d_{16} T + d_{17} G + d_{18} t + d_{19} P_{t-1} + d_{20} R_{t-1} + u_4, \\ P_t = d_{21} T + d_{22} G + d_{23} t + d_{24} P_{t-1} + d_{25} R_{t-1} + u_5, \end{cases}$$

В данной модели коэффициенты при экзогенных переменных T и G являются мультипликаторами, отвечающими на вопрос: На сколько единиц изменится значение эндогенной переменной, если экзогенная переменная изменится на одну единицу своего измерения. Коэффициенты d1, d6, d11, d16, d21 – мультипликаторы чистых трансфертов в пользу администрации относительно личного потребления d1, инвестиций d6, заработной платы d11, дохода d16 и прибыли d21. Соответственно коэффициенты d2, d7, d12, d17, d22 являются мультипликаторами правительственных расходов относительно соответствующих эндогенных переменных.

Динамическая модель Кейнса:

$$\begin{cases} C_t = a + b_1 Y_t + b_2 Y_{t-1} + \varepsilon_1, \\ Y_t = C_t + G_t + I_t + L_t, \\ P_t = Y_t + Z_t. \end{cases}$$

Yt – имеющийся в распоряжении доход в период времени t; Ct – частное потребление в период времени t; Pt – валовой национальный продукт в период времени t; Yt-1 – доход предыдущего года; Gt - общественное потребление; It – валовые капиталовложения; Lt – изменение складских запасов; Zt – сальдо платежного баланса.

Первое уравнение динамической модели Кейнса является сверхидентифицируемым, а второе и третье – тождествами, доход является лаговой переменной.

Линейная модель спроса и предложения:

$$\begin{cases} Q^d = a_0 + a_1 P + \varepsilon_1, \\ Q^s = b_0 + b_1 P + \varepsilon_2, \\ Q^d = Q^s. \end{cases}$$

Qd – спрашиваемое количество благ (объем спроса);

Qs – предлагаемое количество благ (объем предложения).

В этой системе три эндогенные переменные – Qd, Qs и P. При этом если Qd и Qs представляют собой эндогенные переменные исходя из структуры самой системы (они расположены в левой части), то P является эндогенной по экономическому содержанию (цена зависит от предлагаемого и испрашиваемого количества благ), а также в результате наличия тождества Qd=Qs. Линейная модель спроса и предложения не содержит экзогенной переменной. Однако для того, чтобы модель имела статистическое решение и можно было убедиться в ее справедливости, в модель вводятся экзогенные переменные: R и W, после чего модель становится идентифицируемой и может быть оценена КМНК.

$$\begin{cases} Q_d = a_0 + a_1 P + a_2 R + \varepsilon_1, \\ Q_s = b_0 + b_1 P + b_2 W + \varepsilon_2, \\ Q_d = Q_s. \end{cases}$$

R – доход на душу населения;

W – климатические условия (например, спрос и предложение на зерно).

Перечень информационных ресурсов

1. http://tulpar.kfu.ru/course/view.php?id=2213.

2. Эконометрика: [Электронный ресурс] Учеб.пособие / А.И. Новиков. - 3-е изд., испр. и доп. - М.: ИНФРА-М, 2014. - 272 с.: (http://znanium.com/catalog.php?item=booksearch&code=%D1%8D%D0%BA%D0%BE%D0%BD%D0%BE%D0%BC%D0%B5%D1%82%D1%80%D0%B8%D0%BA%D0%B0&page=1#none)

3. Эконометрика: учебник / И. И. Елисеева. – М.: Проспект, 2010. – 288 с.

4. Уткин, В. Б. Эконометрика [Электронный ресурс] : Учебник / В. Б. Уткин; Под ред. проф. В. Б. Уткина. - 2-е изд. - М.: Издательско-торговая корпорация «Дашков и К°», 2012. - 564 с.

(http://znanium.com/catalog.php?item=booksearch&code=%D1%8D%D0%BA%D0%BE%D0%BD%D0%BE%D0%BC%D0%B5%D1%82%D1%80%D0%B8%D0%BA%D0%B0&page=4#none)

5. Валентинов, В. А. Эконометрика [Электронный ресурс]: Практикум / В. А. Валентинов. - 3-е изд. - М.: Дашков и К, 2010. - 436 с.

(http://znanium.com/catalog.php?item=booksearch&code=%D1%8D%D0%BA%D0%BE%D0%BD%D0%BE%D0%BC%D0%B5%D1%82%D1%80%D0%B8%D0%BA%D0%B0&page=3#none)

6. Эконометрика. Практикум: [Электронный ресурс] Учебное пособие / С.А. Бородич. - М.: НИЦ ИНФРА-М; Мн.: Нов.знание, 2014. - 329 с. (http://znanium.com/catalog.php?item=booksearch&code=%D1%8D%D0%BA%D0%BE%D0%BD%D0%BE%D0%BC%D0%B5%D1%82%D1%80%D0%B8%D0%BA%D0%B0&page=4#none)

7.Тихомиров Н. П. Эконометрика: учебник. - М.: Экзамен, серия «Учебник Плехановской академии», 2007, -512 с.

8. Эконометрика: учебник / под ред. В. С. Мхитаряна. - М.: Проспект, 2008. -384 с.

9. Электронный курс “Econometrics and Public Policy (Advanced)”, Princeton University, URL: https://blackboard.princeton.edu/webapps/portal/frameset.jsp?tab_group=courses&url=%2Fwebapps%2Fblackboard%2Fexecute%2FcourseMain%3Fcourse_id%3D_214206_1

10. Электронный курс “Time Series Econometrics”, Princeton University, URL: http://sims.princeton.edu/yftp/Times05/; https://blackboard.princeton.edu/webapps/portal/frameset.jsp?tab_group=courses&url=%2Fwebapps%2Fblackboard%2Fexecute%2FcourseMain%3Fcourse_id%3D_52968_1.

Вопросы и задания для экзамена

1. Типы моделей и переменных, применяемых в эконометрике. Чем регрессионная модель отличается от функции регрессии?
2. Этапы эконометрического моделирования. Каковы основные причины наличия в регрессионной модели случайного отклонения?
3. Основные понятия теории вероятностей. Нормальное распределение и связанные с ним $\chi 2$ - распределение, распределение Стьюдента и Фишера.
4. Генеральная совокупность и выборка. Свойства статистических оценок.
5. Суть метода наименьших квадратов. Предпосылки МНК. Каковы последствия их выполнимости или невыполнимости?
6. Экономическая интерпретация параметров линейной модели парной регрессии. Какой смысл может иметь свободный коэффициент?
7. Статистический смысл коэффициента детерминации. Какова связь между линейным коэффициентом корреляции и коэффициентом регрессии в линейной модели парной регрессии?
8. Баланс для сумм квадратов отклонений результативного признака. В каком случае общая СКО равна факторной? Что происходит, когда общая СКО равна остаточной?

9. Число степеней свободы. Чему равны числа степеней свободы для различных СКО в парной регрессии?

10. Проверка нулевой гипотезы о статистической незначимости уравнения регрессии в целом. Как используется F-статистика в регрессионном анализе?

11. Проверка нулевой гипотезы о статистической незначимости параметров уравнения регрессии. Как рассчитать критерий Стьюдента для коэффициента регрессии в линейной модели парной регрессии?

12. "Грубое" правило анализа статистической значимости коэффициентов регрессии. Какая связь между t_b- и F- статистиками в парной линейной регрессии?

13. Схема определения интервальных оценок коэффициентов регрессии.

14. Схема предсказания индивидуальных значений зависимой переменной. В каком месте доверительный интервал прогноза по парной модели является наименьшим?

15. Спецификация эмпирического уравнения линейной модели множественной регрессии. Что измеряют коэффициенты регрессии линейной модели множественной регрессии?

16. Требования к факторам для включения их в модель множественной регрессии. Мультиколлинеарность.

17. Способы обнаружения мультиколлинеарности.

18. Способы оценивания параметров регрессии в условиях мультиколлинеарности.

19. Стандартизованный вид линейной модели множественной регрессии: форма записи и практическое применение. Как связаны стандартизованные коэффициенты регрессии с натуральными?

20. Скорректированный коэффициент детерминации. В чем недостаток использования коэффициента детерминации при оценке общего качества линейной модели множественной регрессии?

21. Назначение частной корреляции при построении модели множественной регрессии.

22. Смысл и определение индекса множественной корреляции.

23. Способы отбора факторов для включения в линейную модель множественной регрессии.

24. Проверка обоснованности исключения части переменных из уравнения регрессии.

25. Проверка обоснованности включения группы новых переменных в уравнение регрессии.

26. Частный F-критерий. Чем он отличается от последовательногоF-критерия?

27. Гомоскедастичности и гетероскедастичности остатков регрессии. Каковы последствия гетероскедастичности остатков регрессии?

28. Способы обнаружения гетероскедастичности остатков регрессии. Какие критерии могут быть использованы для проверки гипотезы о гомоскедастичности регрессионных остатков?

29. Способы устранения гетероскедастичности остатков регрессии. Метод взвешенных наименьших квадратов.

30. Автокорреляция случайных отклонений. Каковы основные причины и последствия автокорреляции?

31. Основные методы обнаружения автокорреляции.

32. Способы устранения автокорреляции остатков регрессии. Авторегрессионное преобразование.

33. Суть ANOVA-моделей и ANCOVA-моделей.

34. Правило применения фиктивных переменных. Ловушка фиктивных переменных.

35. Смысл дифференциального свободного члена и дифференциального углового коэффициента в моделях с фиктивными переменными.

36. Тест Чоу в моделях с фиктивными переменными.

37. Классы и виды нелинейных регрессий.

38. Линеаризация нелинейных моделей. Выбор формы модели.

39. Индекс корреляции. Подбор линеаризующего преобразования (подход Бокса-Кокса).

40. Коэффициенты эластичности в нелинейных регрессионных моделях.

41. Показатели корреляции при нелинейных соотношениях рассматриваемых признаков. Смысл средней ошибки аппроксимации.

42. Исключение существенных переменных и включение несущественных переменных.

43. Замещающие переменные в регрессионных моделях.

44. Логит-модели и пробит–модели. Какова интерпретация коэффициентов моделей бинарного выбора?

45. Проверка значимости коэффициентов в модели бинарного выбора?

46. Прогноз вероятности по логит-модели. Прогноз вероятности по пробит-модели.

47. Основные понятия и характеристики панельных данных.

48. Модель регрессии с фиксированным эффектом и модель регрессии со случайным индивидуальным эффектом. Оценивание модели со случайным индивидуальным эффектом.

49. Этапы построения тренд-сезонных моделей временных рядов. В чем отличие аддитивной и мультипликативной моделей временных рядов?

50. Прогнозирование на основе трендовой и тренд-сезонной моделей временных рядов. Чему равна сумма сезонных компонент в аддитивной модели временного ряда?

51. Модель ARMA. Как интерпретируют параметры моделей авторегрессии?

52. Стационарность временного ряда. Какой стационарный процесс называется «белым шумом»?

53. Типы моделей стационарных временных рядов.

54. Типы моделей нестационарных временных рядов.

55. ARIMA-модель.

56. Типы систем одновременных уравнений. В чем особенность системы рекурсивных уравнений?

57. Структурная и приведенная формы модели в системах одновременных уравнений.

58. Идентификация модели в системах одновременных уравнений.

59. Косвенный МНК. Всегда ли можно применить косвенный МНК?

60. Двухшаговый МНК. Всегда ли можно применить двухшаговый МНК?

Задание 1. Пусть X,Y – годовые дивиденды от вложений денежных средств в акции компаний А и В соответственно. Риск от вложений характеризуется дисперсиями D(X)=25, D(Y)=16. Коэффициент корреляции σ =+0,8. Куда менее рискованно вкладывать денежные средства: в отрасль В, в отрасль А, в обе отрасли в соотношении 30% на 70%?

Задание 2. Доход X населения имеет нормальный закон распределения со средним значением 5000 руб. и средним квадратическим отклонением 1000 руб. Обследуется 1000 человек. Каково наиболее вероятное количество человек, имеющих доход более 6000 руб.?

Задание 3. Статистика по годовым темпам инфляции в стране за последние 10 лет составила (%) : 2,6; 3,0; 5,2; 1,7; -0,5; 0,6; 2,2; 2,9; 4,2; 3,8. Определите ресмещенные оценки среднего темпа инфляции, дисперсии и среднего квадратического отклонения.

Задание 4. Предполагается, что месячный доход граждан страны имеет нормальное распределение с математическим ожиданием m=500 $ и дисперсией σ2=22500. По выборке из 500 человек определен выборочный средний доход $\bar{x}$ =450 $. Определите доверительный интервал для среднедушевого дохода в стране при уровне значимости 0,05.

Задание 5. При анализе зависимости между двумя показателями X и Y по 30 наблюдениям получены следующие данные: $\bar{x}$ = 105; $\bar{y}$ =80; $\sum_{ш=1}^{30}(x_i-\bar{x})^2$

=900; $\sum_{i=1}^{30} x_i y_i$ =252600; $\sum_{i=1}^{30} (y_i - \bar{y})^2$ =635. Оцените наличие линейной зависимости между X и Y и статистическую значимость коэффициента корреляции ρxy. Задача 6. Предполагается, что месячная зарплата сотрудников фирмы составляет 500 $ при стандартном отклонении σ = 50 $. Выборка из 49 человек дала следующие результаты : $\bar{x}$ =450$ и S = 60$. На основании результатов проведенных наблюдений можно утверждать, что средняя зарплата сотрудников меньше рекламируемой на всех уровнях значимости, а разброс в зарплатах больше на уровне значимости α=0,05 и α=0,1.

Задание 6. Имеется три вида акций A, B и C каждая стоимостью 20 у.е., дивиденды по которым являются независимыми СВ со средним значением 8 % и дисперсией 25. Формируются два портфеля инвестиций. Портфель z1 состоит из 60 акций A. Портфель z2 включает в себя по 20 акций A, B и C. Коэффициент корреляции между дивидендами по акциям A и C равен -0,5, но обе величины не коррелируют с дивидендами по акциям B. Рассчитать риски от вложений средств в данные портфели инвестиций.

Задание 7. Зависимость спроса на кухонные комбайны y от цены x по 20 торговым точкам компании имеет вид: $\ln y = 6{,}8 - 0{,}6 \ln x + \varepsilon$, $(2{,}7)\ (-2{,}8)$ В скобках – фактическое значение t – критерия. Ранее предполагалось, что увеличение цены на 1 % приводит к уменьшению спроса на 1,2 %. Можно ли утверждать, что приведенное уравнение регрессии подтверждает это предположение?

Задание 8. Для двух видов продукции А, Б зависимость удельных постоянных расходов от объема выпускаемой продукции выглядят следующим образом

$$y_A = 15 + 8 \ln x,$$

$$y_Б = 25 x^{0,3}$$

Сравнить эластичности затрат по каждому виду продукции при x=50 и определить объем выпускаемой продукции обоих видов, при котором их эластичность будут одинаковы.

Задание 9. Пусть имеется уравнение парной регрессии: $y = 5 - 6x + \varepsilon$, построенное по 15 наблюдениям. При этом r=-0.7. Определите доверительный интервал с вероятностью 0,99 для коэффициента регрессии в этой модели.

Задание 10. Уравнение регрессии потребления материалов y от объема производства x, построенное по 15 наблюдениям, имеет вид:

$y = 5 + 5x + \varepsilon$, $(4,0)$. В скобках – фактическое значение t – критерия. Определите коэффициент детерминации для этого уравнения.

Задание 11. Уравнение регрессии имеет вид : ln y = 4,5 + 0,003x + ln e. При значении фактора, равном 85, определите коэффициент эластичности Y по X.

Задание 12. По совокупности 15 предприятий торговли изучается зависимость между ценой x на товар А и прибылью y торгового предприятия. При оценке линейной регрессионной модели были получены следующие результаты

$\sum (y - \hat{y})^2 = 32000$ $\sum (y - \bar{y})^2 = 40000$

Определите индекс корреляции, фактическое значение F- критерия, значимость уравнения регрессии.

Задание 13. Изучалась зависимость вида y=a*xb. Для преобразованных в логарифмах переменных (X, Y) получены следующие данные

$\sum XY = 4,2087$
$\sum X = 8,2370$
$\sum X^2 = 9,2334$
$\sum Y = 3,9310, n = 10$

Определите значение параметра b.

Задание 14. Изучалась зависимость вида y=a+b*x+e. Получены следующие данные

$\sum xy = 42{,}087$
$\sum x = 82{,}370$
$\sum x^2 = 92{,}334$
$\sum y = 39{,}310, n = 100$

Определите значение параметра b.

Задание 15. Зависимость объема продаж Y от расходов на рекламу X характеризуется по 12 предприятиям концерна следующим образом

$y = 10{,}6 + 0{,}6 \cdot x$

$\sigma_x = 4{,}7$

$\sigma_y = 3{,}4$

Определите t-статистику коэффициента регрессии.

Задание 16. По совокупности 15 предприятий торговли изучается зависимость между ценой x на товар А и прибылью y торгового предприятия. При оценке квадратической регрессионной модели были получены следующие результаты: $\sum (y - \hat{y})^2 = 32000$, $\Sigma (y - \bar{y})^2 = 40000$. Определите фактическое значение F- критерия, значимость уравнения регрессии.

Задание 17. Уравнение регрессии в стандартизированном виде имеет вид:

$\hat{t}_y = 0{,}37 t_{x_1} - 0{,}52 t_{x_2} + 0{,}43 t_{x_3},$

$V_y = 18\%; \quad V_{x_1} = 25\%; \quad V_{x_2} = 38\%; \quad V_{x_3} = 30\%.$

Определите частные коэффициенты эластичности.

Задание 18. По 18 наблюдениям получены следующие данные: $\hat{y} = a + 0{,}36x_1 - 0{,}255x_2 + 2{,}86x_3, \; R^2 = 0{,}65; \quad \bar{y} = 70;$

$\bar{x}_1 = 110; \quad \bar{x}_2 = 150; \quad \bar{x}_3 = 85$. Определите значения скорректированного коэффициента детерминации, частных коэффициентов эластичности и параметра a.

Задание 19. Уравнение регрессии в стандартизованном виде имеет вид:

$$\hat{t}_y = -0{,}82t_{x_1} + 0{,}65t_{x_2} - 0{,}43t_{x3,}$$

$$V_y = 32\%; V_{x1} = 38\%; V_{x2} = 43\%; V_{x3} = 35\%$$

Как влияют факторы на результат и каковы значения частных коэффициентов эластичности?

Задание 20. По следующим данным: $\bar{y} = 15{,}0;\quad \bar{x}_1 = 6{,}5; \bar{x}_2 = 12{,}0;$ $\sigma_y = 4{,}0;\ \sigma_{x1} = 2{,}5;\ \ \sigma_{x_2} = 3{,}5;\ \ r_{yx1} = 0{,}63;\ \ r_{yx2} = 0{,}78;\ \ r_{x1x2} = 0{,}52$, запишите уравнения регрессии y на x_1 и x_2 в стандартизованном и натуральном масштабе.

Задание 21. При построении регрессионной зависимости некоторого результативного признака на 8 факторов по 25 измерениям коэффициент детерминации составил 0,736. После исключения 3 факторов коэффициент детерминации уменьшился до 0,584. Обоснованно ли было принятое решение на уровнях значимости 0,1, 0,05 и 0,01?

Задание 22. По данным 150 наблюдений о доходе индивидуума Y, уровне его образования X1, и возрасте X2 определите, можно ли считать на уровне значимости 5 % линейную регрессионную модель Y на X1 и X2 гетероскедастичной, если суммы квадратов остатков после упорядочения данных по уровню образования следующие RSS1 (для 50 значений с наименьшим уровнем образования) = 894,1; RSS2 (для 50 значений с наибольшим уровнем образования) = 3918,2.

Задание 23. При построении регрессионной зависимости $y = f\left(x_1 x_2, \ldots x_9\right)$ по 40 измерениям коэффициент детерминации составил 0,618. После исключения факторов x_4 и x_5 коэффициент детерминации уменьшился до 0,547. Обоснованно ли было принятое решение на уровнях значимости 0,1; 0,05 и 0,01?

Задание 24. При анализе данных на гетероскедастичность вся выборка была после упорядочения разбита на три подвыборки. Затем по результатам пар-

ных регрессий остаточная СКО в первой подвыборке составила 6450, в третьей – 3480. Подтверждается ли наличие гетероскедастичности на уровнях 0,1; 0,05 и 0,01, если объем данных в каждой подвыборке равен 25?

Задание 25. Уравнение регрессии, построенное по 12 наблюдениям, имеет вид:

$$\begin{array}{llll} y = & 12 - & 0{,}24x_1 + & 6{,}4x_2 + \quad ?x_3 \\ m_b & (8) & (\ ) & (3{,}2) \quad (4{,}0) \\ t_b & (\ ) & (-2{,}4) & (\ ) \quad (-3{,}1) \end{array}$$

Определите пропущенные значения и доверительный интервал для b_3 с вероятностью 0,99.

Задание 26. На основе помесячных данных за последние 4 года была построена аддитивная модель временного потребления тепла. Скорректированные значения сезонной компоненты приведены в таблице:

Январь	+ 30	май	- 20	сентябрь	- 10
февраль	+ 25	июнь	- 34	октябрь	?
март	+ 15	июль	- 42	ноябрь	+22
апрель	- 2	август	- 18	декабрь	+27

Уравнение тренда выглядит так $T = 350 + 1{,}3t$. Определите значение сезонной компоненты за октябрь, а также точечный прогноз потребления тепла на 1 квартал следующего года.

Задание 27. На основе поквартальных данных построена мультипликативная модель некоторого временного ряда. Уравнение тренда имеет вид:

$T = 11{,}6 - 0{,}1 \cdot t \ \ (t = \overline{1{,}48})$. Скорректированные значения сезонной компоненты равны:

—I квартал – 1,6

—II квартал – 0,8

—III квартал – 0,7

—IV квартал - ?

Определите значение сезонной компоненты за IV квартал и прогноз на II и III кварталы следующего года .

Задание 28. На основе квартальных данных объемов продаж 2008 – 2013гг. была построена аддитивная модель временного ряда. Трендовая компонента имеет вид $T = 260 + 3 \cdot t \ \ (t = 1,2,...)$. Показатели за 2014 г. приведены в таблице:

Квартал	Фактический объем продаж	Компонента аддитивной модели		
		трендовая	сезонная	случайная
1	270	T_1	S_1	-9
2	y_2	T_2	10	+4
3	310	T_3	40	E_3
4	y_4	T_4	S_4	E_4
ИТОГО	2000			

Определите отдельные недостающие данные в таблице.

Задание 29. На основе квартальных данных с 2000 г. по 2004 г. получено уравнение y = - 0,67 + 0,0098 x t1 – 5,62 x t2 + 0,044 x t3.

ESS =110,3, RSS = 21,4 (ESS – объясненная сумма квадратов, RSS – остаточная сумма квадратов). В уравнение были добавлены три фиктивные переменные, соответствующие трем первым кварталам года, величина ESS увеличилась до 120,2. Проверьте гипотезу о сезонности (α =0,05)

Задание 30. Модель зависимости объемов продаж компании от расходов на рекламу имеет вид y = - 0,67 + 4,5 x t + 3 x t-1 + 1,5 x t-2 + 0,5 x t-3. Определите краткосрочный, долгосрочный мультипликатор и средний лаг.

Задание 31. На основе квартальных данных получено уравнение множественной регрессии и ESS = 120,32, RSS = 41,4. (ESS – объясненная сумма квадратов, RSS – остаточная сумма квадратов). Для этой же модели были раздельно проведены регрессии на основе данных: 1-й квартал 1991 г. - 1-й квартал 1995 г. и 2-й квартал 1995 г. – 4 квартал 1996 г., соответственно получены следующие значения сумм квадратов остатков RSS1 = 22,25, RSS2=12,32. Про-

верьте гипотезу о том, что произошли структурные изменения на уровне α =0,05.

Задание 32. На основе квартальных данных с 1991 года по 1996 год с помощью МНК получено следующее уравнение

Y t = 1,12 – 0, 0098 x t1 – 5, 62 x t2 + 0, 044 x t3

(2,14) (0,0034) (3,42) (0,009)

В скобках указаны стандартные ошибки, ESS (объясненная сумма квадратов) = 116, 32; RSS (остаточная сумма квадратов) = 31, 43

Проверьте значимости коэффициентов и модели в целом при уровне значимости α = 0,05.

Задание 33.Дана таблица

Момент времени	$t-3$	$t-2$	$t-1$	t	$t+1$
S^*	70				
S	85	100	120	135	—

где $S^*, S-$ ожидаемый и действительный объемы предложения. В соответствии с моделью адаптивных ожиданий, где $\lambda = 0{,}45$, определите значение S^*_{t+1}

Задание 34. Модель зависимости объемов продаж компании от расходов на рекламу имеет вид $y = 0{,}67 + 4{,}5x_t + 3x_{\text{t-1}} + 1{,}5x_{\text{t}-2} + 0{,}5x_{t-3}$.

Определите средний лаг

Задание 35. Имеется следующая структурная модель:

$$\begin{cases} y_1 = b_{12}y_2 + a_{11}x_1 + a_{12}x_2, \\ y_2 = b_{21}y_1 + b_{23}y_3 + a_{22}x_2, \\ y_3 = b_{32}y_2 + a_{31}x_1 + a_{33}x_3. \end{cases}$$

Соответствующая ей приведенная форма модели имеет вид

$$\begin{cases} y_1 = 3x_1 - 4x_2 + 2x_3, \\ y_2 = 2x_1 + 4x_2 + 5x_3, \\ y_3 = -5x_1 + 6x_2 + 5x_3. \end{cases}$$

Определите параметры первого уравнения структурной формы.

Задание 36. Имеется следующая структурная модель

$$\begin{cases} y_1 = b_{12}y_2 + a_{11}x_1 + a_{12}x_2, \\ y_2 = b_{21}y_1 + b_{23}y_3 + a_{22}x_2, \\ y_3 = b_{32}y_2 + a_{31}x_1 + a_{33}x_3. \end{cases}$$

Ей соответствует приведенная форма:

$$\begin{cases} y_1 = 3x_1 - 4x_2 + 2x_3, \\ y_2 = 2x_1 + 4x_2 + 5x_3, \\ y_3 = -5x_1 + 6x_2 + 5x_3. \end{cases}$$

Определите параметры третьего уравнения структурной формы.

Задание 37. Имеется следующая модель

$$\begin{cases} R_t = a_1 + b_{11}Mt + b_{12}Yt + \varepsilon_1, \\ Y_t = a_2 + b_{21}R_t + b_{22}I_t + \varepsilon_2, \\ I_t = a_3 + b_{33}Rt + \varepsilon_3. \end{cases}$$

Проверьте модель на идентификацию.

Задание 38. Имеется следующая модель

$$\begin{cases} C_t = a_1 + b_{11}D_t + \varepsilon_{1t}, \\ I_t = a_2 + b_{22}Y_t + b_{23}Y_{t-1} + \varepsilon_{2t}, \\ Y_t = D_t + T_t, \\ D_t = C_t + I_t + G_t. \end{cases}$$

Проверьте модель на идентификацию.